Regularized gradient descent: a non-convex recipe for fast joint blind deconvolution and demixing

Shuyang Ling[†]

Courant Institute of Mathematical Sciences, New York University
[†]Corresponding author: sling@cims.nyu.edu

and

Thomas Strohmer

Department of Mathematics, University of California at Davis
strohmer@math.ucdavis.edu

We study the question of extracting a sequence of functions $\{f_i, g_i\}_{i=1}^s$ from observing only the sum of their convolutions, i.e. from $y = \sum_{i=1}^s f_i * g_i$. While convex optimization techniques are able to solve this joint blind deconvolution–demixing problem provably and robustly under certain conditions, for medium-size or large-size problems we need computationally faster methods without sacrificing the benefits of mathematical rigor that come with convex methods. In this paper we present a non-convex algorithm which guarantees exact recovery under conditions that are competitive with convex optimization methods, with the additional advantage of being computationally much more efficient. Our two-step algorithm converges to the global minimum linearly and is also robust in the presence of additive noise. While the derived performance bounds are suboptimal in terms of the information-theoretic limit, numerical simulations show remarkable performance even if the number of measurements is close to the number of degrees of freedom. We discuss an application of the proposed framework in wireless communications in connection with the Internet-of-Things.

Keywords: Blind deconvolution; blind demixing; non-convex optimization; random matrix; signal processing; wireless communication.

1. Introduction

The goal of blind deconvolution is the task of estimating two unknown functions from their convolution. While it is a highly ill-posed bilinear inverse problem, blind deconvolution is also an extremely important problem in signal processing [1], communications engineering [39], imaging processing [5], audio processing [24], etc. In this paper, we deal with an even more difficult and more general variation of the blind deconvolution problem, in which we have to extract multiple convolved signals mixed together in one observation signal. This joint blind deconvolution–demixing problem arises in a range of applications such as acoustics [24], dictionary learning [2] and wireless communications [39].

We briefly discuss one such application in more detail. Blind deconvolution/demixing problems are expected to play a vital role in the future Internet-of-Things. The Internet-of-Things will connect billions of wireless devices, which is far more than the current wireless systems can technically and economically accommodate. One of the many challenges in the design of the Internet-of-Things will be its ability to manage the massive number of sporadic traffic generating devices which are most of the time inactive, but regularly access the network for minor updates with no human interaction [43]. This
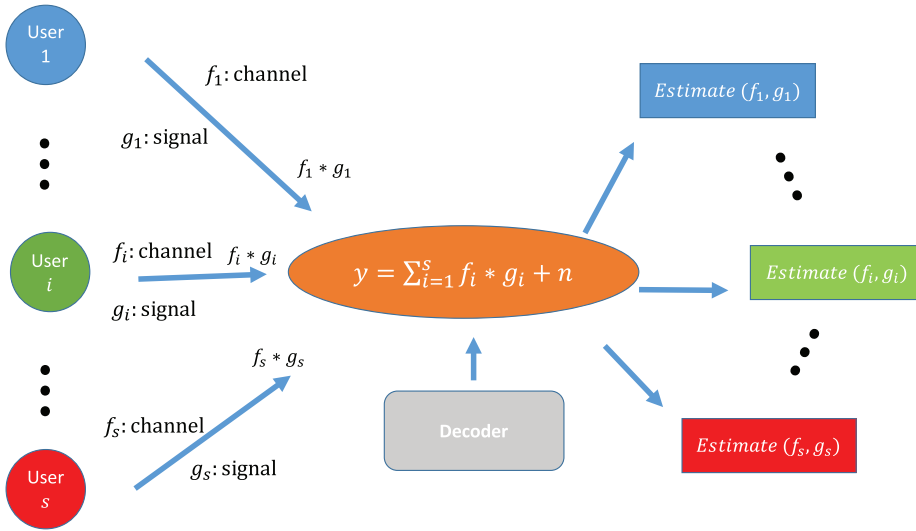
FIG. 1. Single-antenna multi-user communication scenario without explicit channel estimation: each of the s users sends a signal g_i through an unknown channel f_i to a common base station. The base station measures the superposition of all those signals, namely, $y = \sum_{i=1}^{s} f_i * g_i$ (plus noise). The goal is to extract all pairs of $\{(f_i, g_i)\}_{i=1}^{s}$ simultaneously from y.

means among others that the overhead caused by the exchange of certain types of information between transmitter and receiver, such as channel estimation, assignment of data slots, etc., has to be avoided as much as possible [27,36].

Focusing on the underlying mathematical challenges, we consider a *multi-user communication* scenario where many different users/devices communicate with a common base station, as illustrated in Fig. 1. Suppose we have s users and each of them sends a signal g_i through an unknown channel (which differs from user to user) to a common base station. We assume that the ith channel, represented by its impulse response f_i, does not change during the transmission of the signal g_i. Therefore f_i acts as convolution operator, i.e., the signal transmitted by the ith user arriving at the base station becomes $f_i * g_i$, where '$*$' denotes convolution.

The antenna at the base station, instead of receiving each individual component $f_i * g_i$, is only able to record the superposition of all those signals, namely,

$$y = \sum_{i=1}^{s} f_i * g_i + n, \tag{1.1}$$

where n represents noise. We aim to develop a fast algorithm to simultaneously extract all pairs $\{(f_i, g_i)\}_{i=1}^{s}$ from y (i.e., estimating the channel/impulse responses f_i and the signals g_i jointly) in a numerically efficient and robust way, while keeping the number of required measurements as small as possible.

1.1 *State of the art and contributions of this paper*

A thorough theoretical analysis concerning the solvability of demixing problems via convex optimization can be found in [26]. There, the authors derive explicit sharp bounds and phase transitions regarding

the number of measurements required to successfully demix structured signals (such as sparse signals or low-rank matrices) from a single measurement vector. In principle we could recast the blind deconvolution/demixing problem as the demixing of a sum of rank-one matrices, see (4). As such, it seems to fit into the framework analyzed by McCoy and Tropp. However, the setup in [26] differs from ours in a crucial manner. McCoy and Tropp consider as measurement matrices (see the matrices $\mathcal{A}_i$ in (4)) full-rank random matrices, while in our setting the measurement matrices are rank-one. This difference fundamentally changes the theoretical analysis. The findings in [26] are therefore not applicable to the problem of joint blind deconvolution/demixing. The compressive principal component analysis in [42] is also a form of demixing problem, but its setting is only vaguely related to ours. There is a large amount of literature on demixing problems, but the vast majority does not have a 'blind deconvolution component', therefore this body of work is only marginally related to the topic of our paper.

Blind deconvolution/demixing problems also appear in convolutional dictionary learning, see e.g. [2]. There, the aim is to factorize an ensemble of input vectors into a linear combination of overcomplete basis elements which are modeled as shift-invariant—the latter property is why the factorization turns into a convolution. The setup is similar to (1.1), but with an additional penalty term to enforce sparsity of the convolving filters. The existing literature on convolutional dictionary learning is mainly focused on empirical results, therefore there is little overlap with our work. But it is an interesting challenge for future research to see whether the approach in this paper can be modified to provide a fast and theoretically sound solver for the sparse convolutional coding problem.

There are numerous papers concerned with blind deconvolution/demixing problems in the area of wireless communications [20,31,37]. But the majority of these papers assumes the availability of multiple measurement vectors, which makes the problem significantly easier. Those methods however cannot be applied to the case of a single measurement vector, which is the focus of this paper. Thus there is essentially no overlap of those papers with our work.

Our previous paper [23] solves (1.1) under subspace conditions, i.e., assuming that both f_i and g_i belong to known linear subspaces. This contributes to generalizing the pioneering work by Ahmed *et al.* [1] from the 'single-user' scenario to the 'multi-user' scenario. Both [1] and [23] employ a two-step convex approach: first 'lifting' [9] is used and then the lifted version of the original bilinear inverse problems is relaxed into a semi-definite program. An improvement of the theoretical bounds in [23] was announced in [29].

While the convex approach is certainly effective and elegant, it can hardly handle large-scale problems. This motivates us to apply a non-convex optimization approach [8,21] to this blind-deconvolution-blind-demixing problem. The mathematical challenge, when using non-convex methods, is to derive a rigorous convergence framework with conditions that are competitive with those in a convex framework.

In the last few years several excellent articles have appeared on provably convergent non-convex optimization applied to various problems in signal processing and machine learning, e.g., matrix completion [16,17,34], phase retrieval [3,8,11,33], blind deconvolution [4,19,21], dictionary learning [32], super-resolution [12] and low-rank matrix recovery [35,41]. In this paper we derive the first non-convex optimization algorithm to solve (1.1) fast and with rigorous theoretical guarantees concerning exact recovery, convergence rates, as well as robustness for noisy data. Our work can be viewed as a generalization of blind deconvolution [21] ($s = 1$) to the multi-user scenario ($s > 1$).

The idea behind our approach is strongly motivated by the non-convex optimization algorithm for phase retrieval proposed in [8]. In this foundational paper, the authors use a two-step approach: (i) construct a good initial guess with a numerically efficient algorithm; (ii) starting with this initial guess, prove that simple gradient descent will converge to the true solution. Our paper follows a similar

two-step scheme. However, the techniques used here are quite different from [8]. Like the matrix completion problem [7], the performance of the algorithm relies heavily and inherently on how much the ground truth signals are aligned with the design matrix. Due to this so-called 'incoherence' issue, we need to impose extra constraints, which results in a different construction of the so-called *basin of attraction*. Therefore, influenced by [17,21,34], we add penalty terms to control the incoherence and this leads to the regularized gradient descent method, which forms the core of our proposed algorithm.

To the best of our knowledge, our algorithm is the first algorithm for the blind deconvolution/blind demixing problem that is numerically efficient, is robust against noise and comes with rigorous recovery guarantees.

1.2 *Notation*

For a matrix Z, $\|Z\|$ denotes its operator norm and $\|Z\|_F$ is its the Frobenius norm. For a vector z, $\|z\|$ is its Euclidean norm and $\|z\|_\infty$ is the ℓ_∞-norm. For both matrices and vectors, Z^* and z^* denote their complex conjugate transpose. $\bar{z}$ is the complex conjugate of z. We equip the matrix space $\mathbb{C}^{K \times N}$ with the inner product defined by $\langle U, V \rangle := \mathrm{Tr}(U^*V)$. For a given vector z, $\mathrm{diag}(z)$ represents the diagonal matrix whose diagonal entries are z. For any $z \in \mathbb{R}$, let $z_+ = \frac{z+|z|}{2}$.

2. Preliminaries

Obviously, without any further assumption, it is impossible to solve (1.1). Therefore, we impose the following subspace assumptions throughout our discussion [1,23].

- **Channel subspace assumption:** each finite impulse response $f_i \in \mathbb{C}^L$ is assumed to have *maximum delay spread K*, i.e.,

$$f_i = \begin{bmatrix} h_i \\ 0 \end{bmatrix}.$$

   Here $h_i \in \mathbb{C}^K$ is the non-zero part of f_i and $f_i(n) = 0$ for $n > K$.

- **Signal subspace assumption:** let $g_i := C_i \bar{x}_i$ be the outcome of the signal $\bar{x}_i \in \mathbb{C}^N$ encoded by a matrix $C_i \in \mathbb{C}^{L \times N}$ with $L > N$, where the encoding matrix C_i is known and assumed to have full rank.[1]

REMARK 2.1 Both subspace assumptions are common in various applications. For instance in wireless communications, the channel impulse response can always be modeled to have finite support (or maximum delay spread, as it is called in engineering jargon) due to the physical properties of wave propagation [14]; and the signal subspace assumption is a standard feature found in many current communication systems [14], including Code-division multiple access (CDMA) where C_i is known as spreading matrix and Orthogonal frequency-division multiplexing (OFDM) where C_i is known as precoding matrix.

The specific choice of the encoding matrices C_i depends on a variety of conditions. In this paper, we derive our theory by assuming that C_i is a complex Gaussian random matrix, i.e., each entry in C_i is i.i.d. $\mathcal{CN}(0,1)$. This assumption, while sometimes imposed in the wireless communications literature, is somewhat unrealistic in practice, due to the lack of a fast algorithm to apply C_i and due to storage

---

[1] Here we use the conjugate $\bar{x}_i$ instead of x_i because it will simplify our notation in later derivations.

requirements. In practice one would rather choose C_i to be something like the product of a Hadamard matrix and a diagonal matrix with random binary entries. We hope to address such more structured encoding matrices in our future research. Our numerical simulations (see Section 4) show no difference in the performance of our algorithm for either choice.

Under the two assumptions above, the model actually has a simpler form in the *frequency* domain. We assume throughout the paper that the convolution of finite sequences is circular convolution.[2] By applying the Discrete Fourier Transform (DFT) to (1.1) along with the two assumptions, we have

$$\frac{1}{\sqrt{L}}Fy = \sum_{i=1}^{s} \text{diag}(Fh_i)(FC_i\bar{x}_i) + \frac{1}{\sqrt{L}}Fn,$$

where F is the $L \times L$ normalized unitary DFT matrix with $F^*F = FF^* = I_L$. The noise is assumed to be additive white complex Gaussian noise with $n \sim \mathcal{CN}(0, \sigma^2 d_0^2 I_L)$ where $d_0 = \sqrt{\sum_{i=1}^{s} \|h_{i0}\|^2 \|x_{i0}\|^2}$, and $\{(h_{i0}, x_{i0})\}_{i=1}^{s}$ is the ground truth. We define $d_{i0} = \|h_{i0}x_{i0}^*\|_F$ and assume without loss of generality that $\|h_{i0}\|$ and $\|x_{i0}\|$ are of the same norm, i.e., $\|h_{i0}\| = \|x_{i0}\| = \sqrt{d_{i0}}$, which is due to the scaling ambiguity.[3] In that way, $\frac{1}{\sigma^2}$ actually is a measure of signal to noise ratio (SNR).

Let $h_i \in \mathbb{C}^K$ be the first K non-zero entries of f_i and $B \in \mathbb{C}^{L \times K}$ be a low-frequency DFT matrix (the first K columns of an $L \times L$ unitary DFT matrix). Then a simple relation holds,

$$Ff_i = Bh_i, \quad B^*B = I_K.$$

We also denote $A_i := \overline{FC_i}$ and $e := \frac{1}{\sqrt{L}}Fn$. Due to the Gaussianity, A_i also possesses complex Gaussian distribution and so does e. From now on, instead of focusing on the original model, we consider (with a slight abuse of notation) the following equivalent formulation throughout our discussion:

$$y = \sum_{i=1}^{s} \text{diag}(Bh_i)\overline{A_ix_i} + e, \tag{2.1}$$

where $e \sim \mathcal{CN}\left(0, \frac{\sigma^2 d_0^2}{L}I_L\right)$. Our goal here is to estimate all $\{h_i, x_i\}_{i=1}^{s}$ from y, B and $\{A_i\}_{i=1}^{s}$. Obviously, this is a bilinear inverse problem, i.e., if all $\{h_i\}_{i=1}^{s}$ are given, it is a linear inverse problem (the ordinary demixing problem) to recover all $\{x_i\}_{i=1}^{s}$, and vice versa. We note that there is a scaling ambiguity in all blind deconvolution problems that cannot be resolved by any reconstruction method without further information. Therefore, when we talk about exact recovery in the following, then this is understood modulo such a trivial scaling ambiguity.

Before proceeding to our proposed algorithm we introduce some notation to facilitate a more convenient presentation of our approach. Let b_l be the lth column of B^* and a_{il} be the lth column of A_i^*. Based on our assumptions the following properties hold:

$$\sum_{l=1}^{L} b_l b_l^* = I_K, \quad \|b_l\|^2 = \frac{K}{L}, \quad a_{il} \sim \mathcal{CN}(0, I_N).$$

---

[2] This circular convolution assumption can often be reinforced directly (e.g., in wireless communications the use of a cyclic prefix in OFDM renders the convolution circular) or indirectly (e.g., via zero-padding). In the first case replacing regular convolution by circular convolution does not introduce any errors at all. In the latter case one introduces an additional approximation error in the inversion which is negligible, since it decays exponentially for impulse responses of finite length [30].

[3] Namely, if the pair (h_i, x_i) is a solution, then so is $(\alpha h_i, \alpha^{-1}x_i)$ for any $\alpha \neq 0$.

Moreover, inspired by the well-known *lifting* idea [**1**,**6**,**9**,**22**], we define the useful matrix-valued linear operator $\mathcal{A}_i : \mathbb{C}^{K \times N} \to \mathbb{C}^L$ and its adjoint $\mathcal{A}_i^* : \mathbb{C}^L \to \mathbb{C}^{K \times N}$ by

$$\mathcal{A}_i(\boldsymbol{Z}) := \{\boldsymbol{b}_l^* \boldsymbol{Z} \boldsymbol{a}_{il}\}_{l=1}^L, \quad \mathcal{A}_i^*(\boldsymbol{z}) := \sum_{l=1}^L z_l \boldsymbol{b}_l \boldsymbol{a}_{il}^* = \boldsymbol{B}^* \operatorname{diag}(\boldsymbol{z}) \boldsymbol{A}_i \tag{2.2}$$

for each $1 \le i \le s$ under canonical inner product over $\mathbb{C}^{K \times N}$. Therefore, (2.1) can be written in the following equivalent form

$$\boldsymbol{y} = \sum_{i=1}^s \mathcal{A}_i(\boldsymbol{h}_i \boldsymbol{x}_i^*) + \boldsymbol{e}. \tag{2.3}$$

Hence, we can think of $\boldsymbol{y}$ as the observation vector obtained from taking *linear* measurements with respect to a set of rank-one matrices $\{\boldsymbol{h}_i \boldsymbol{x}_i^*\}_{i=1}^s$. In fact, with a bit of linear algebra (and ignoring the noise term for the moment), the lth entry of $\boldsymbol{y}$ in (2.3) equals the inner product of two block-diagonal matrices:

$$y_l = \left\langle \underbrace{\begin{bmatrix} \boldsymbol{h}_{1,0} \boldsymbol{x}_{1,0}^* & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{h}_{2,0} \boldsymbol{x}_{2,0}^* & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{h}_{s0} \boldsymbol{x}_{s0}^* \end{bmatrix}}_{\text{defined as } \boldsymbol{X}_0}, \begin{bmatrix} \boldsymbol{b}_l \boldsymbol{a}_{1l}^* & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{b}_l \boldsymbol{a}_{2l}^* & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{b}_l \boldsymbol{a}_{sl}^* \end{bmatrix} \right\rangle + e_l, \tag{2.4}$$

where $y_l = \sum_{i=1}^s \boldsymbol{b}_l^* \boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^* \boldsymbol{a}_{il} + e_l$, $1 \le l \le L$ and $\boldsymbol{X}_0$ is defined as the ground truth matrix. In other words, we aim to recover such a block-diagonal matrix $\boldsymbol{X}_0$ from L linear measurements with block structure if $\boldsymbol{e} = \boldsymbol{0}$.

By stacking all $\{\boldsymbol{h}_i\}_{i=1}^s$ (and $\{\boldsymbol{x}_i\}_{i=1}^s, \{\boldsymbol{h}_{i0}\}_{i=1}^s, \{\boldsymbol{x}_{i0}\}_{i=1}^s$) into a long column, we let

$$\boldsymbol{h} := \begin{bmatrix} \boldsymbol{h}_1 \\ \vdots \\ \boldsymbol{h}_s \end{bmatrix}, \quad \boldsymbol{h}_0 := \begin{bmatrix} \boldsymbol{h}_{1,0} \\ \vdots \\ \boldsymbol{h}_{s0} \end{bmatrix} \in \mathbb{C}^{Ks}, \quad \boldsymbol{x} := \begin{bmatrix} \boldsymbol{x}_1 \\ \vdots \\ \boldsymbol{x}_s \end{bmatrix}, \quad \boldsymbol{x}_0 := \begin{bmatrix} \boldsymbol{x}_{1,0} \\ \vdots \\ \boldsymbol{x}_{s0} \end{bmatrix} \in \mathbb{C}^{Ns}. \tag{2.5}$$

We define $\mathcal{H}$ as a bilinear operator which maps a pair $(\boldsymbol{h}, \boldsymbol{x}) \in \mathbb{C}^{Ks} \times \mathbb{C}^{Ns}$ into a block diagonal matrix in $\mathbb{C}^{Ks \times Ns}$, i.e.,

$$\mathcal{H}(\boldsymbol{h}, \boldsymbol{x}) := \begin{bmatrix} \boldsymbol{h}_1 \boldsymbol{x}_1^* & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{h}_2 \boldsymbol{x}_2^* & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{h}_s \boldsymbol{x}_s^* \end{bmatrix} \in \mathbb{C}^{Ks \times Ns}. \tag{2.6}$$

Let $\boldsymbol{X} := \mathcal{H}(\boldsymbol{h}, \boldsymbol{x})$ and $\boldsymbol{X}_0 := \mathcal{H}(\boldsymbol{h}_0, \boldsymbol{x}_0)$ where $\boldsymbol{X}_0$ is the ground truth as illustrated in (2.4). Define $\mathcal{A}(\boldsymbol{Z}) : \mathbb{C}^{Ks \times Ns} \to \mathbb{C}^L$ as

$$\mathcal{A}(\boldsymbol{Z}) := \sum_{i=1}^s \mathcal{A}_i(\boldsymbol{Z}_i), \tag{2.7}$$

where $\boldsymbol{Z} = \text{blkdiag}(\boldsymbol{Z}_1, \cdots, \boldsymbol{Z}_s)$ and blkdiag is the standard MATLAB function to construct block diagonal matrix. Therefore, $\mathcal{A}(\mathcal{H}(\boldsymbol{h}, \boldsymbol{x})) = \sum_{i=1}^{s} \mathcal{A}_i(\boldsymbol{h}_i \boldsymbol{x}_i^*)$ and $\boldsymbol{y} = \mathcal{A}(\mathcal{H}(\boldsymbol{h}_0, \boldsymbol{x}_0)) + \boldsymbol{e}$. The adjoint operator $\mathcal{A}^*$ is defined naturally as

$$\mathcal{A}^*(\boldsymbol{z}) := \begin{bmatrix} \mathcal{A}_1^*(\boldsymbol{z}) & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \mathcal{A}_2^*(\boldsymbol{z}) & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \mathcal{A}_s^*(\boldsymbol{z}) \end{bmatrix} \in \mathbb{C}^{Ks \times Ns}, \tag{2.8}$$

which is a linear map from $\mathbb{C}^L$ to $\mathbb{C}^{Ks \times Ns}$. To measure the approximation error of $\boldsymbol{X}_0$ given by $\boldsymbol{X}$, we define $\delta(\boldsymbol{h}, \boldsymbol{x})$ as the global relative error:

$$\delta(\boldsymbol{h}, \boldsymbol{x}) := \frac{\|\boldsymbol{X} - \boldsymbol{X}_0\|_F}{\|\boldsymbol{X}_0\|_F} = \frac{\sqrt{\sum_{i=1}^{s} \|\boldsymbol{h}_i \boldsymbol{x}_i^* - \boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^*\|_F^2}}{d_0} = \sqrt{\frac{\sum_{i=1}^{s} \delta_i^2 d_{i0}^2}{\sum_{i=1}^{s} d_{i0}^2}}, \tag{2.9}$$

where $\delta_i := \delta_i(\boldsymbol{h}_i, \boldsymbol{x}_i)$ is the relative error within each component:

$$\delta_i(\boldsymbol{h}_i, \boldsymbol{x}_i) := \frac{\|\boldsymbol{h}_i \boldsymbol{x}_i^* - \boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^*\|_F}{d_{i0}}.$$

Note that δ and δ_i are functions of $(\boldsymbol{h}, \boldsymbol{x})$ and $(\boldsymbol{h}_i, \boldsymbol{x}_i)$, respectively, and in most cases, we just simply use δ and δ_i if no possibility of confusion exists.

2.1 *Convex versus non-convex approaches*

As indicated in (2.4), joint blind deconvolution–demixing can be recast as the task to recover a rank-s block-diagonal matrix from linear measurements. In general, such a low-rank matrix recovery problem is NP-hard. In order to take advantage of the low-rank property of the ground truth, it is natural to adopt convex relaxation by solving a convenient nuclear norm minimization program, i.e.,

$$\min \sum_{i=1}^{s} \|\boldsymbol{Z}_i\|_*, \quad s.t. \quad \sum_{i=1}^{s} \mathcal{A}_i(\boldsymbol{Z}_i) = \boldsymbol{y}. \tag{2.10}$$

The question of when the solution of (2.10) yields exact recovery is first answered in our previous work [23]. Late, [15,29] have improved this result to the near-optimal bound $L \geq C_0 s(K + N)$ up to some log-factors where the main theoretical result is informally summarized in the following theorem.

THEOREM 2.2 (**Theorem 1.1 in [15]**). Suppose that $\boldsymbol{A}_i$ are $L \times N$ i.i.d. complex Gaussian matrices and $\boldsymbol{B}$ is an $L \times K$ partial DFT matrix with $\boldsymbol{B}^*\boldsymbol{B} = \boldsymbol{I}_K$. Then solving (2.10) gives exact recovery if the number of measurements L yields

$$L \geq C_\gamma s(K + N) \log^3 L$$

with probability at least $1 - L^{-\gamma}$ where C_γ is an absolute scalar only depending on γ linearly.

While the semidefinite programming (SDP) relaxation is definitely effective and has theoretic performance guarantees, the computational costs for solving an SDP already become too expensive for moderate size problems, let alone for large scale problems. Therefore, we try to look for a more efficient non-convex approach such as gradient descent, which hopefully is also reinforced by theory. It seems quite natural to achieve the goal by minimizing the following *nonlinear* least squares objective function with respect to $(\boldsymbol{h}, \boldsymbol{x})$

$$F(\boldsymbol{h}, \boldsymbol{x}) := \|\mathcal{A}\left(\mathcal{H}(\boldsymbol{h}, \boldsymbol{x})\right) - \boldsymbol{y}\|^2 = \left\| \sum_{i=1}^{s} \mathcal{A}_i(\boldsymbol{h}_i \boldsymbol{x}_i^*) - \boldsymbol{y} \right\|^2. \tag{2.11}$$

In particular, if $\boldsymbol{e} = \boldsymbol{0}$, we write

$$F_0(\boldsymbol{h}, \boldsymbol{x}) := \left\| \sum_{i=1}^{s} \mathcal{A}_i(\boldsymbol{h}_i \boldsymbol{x}_i^* - \boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^*) \right\|^2. \tag{2.12}$$

As also pointed out in [21], this is a highly non-convex optimization problem. Many of the commonly used algorithms, such as gradient descent or alternating minimization, may not necessarily yield convergence to the global minimum, so that we cannot always hope to obtain the desired solution. Often, those simple algorithms might get stuck in local minima.

2.2   *The basin of attraction*

Motivated by several excellent recent papers of non-convex optimization on various signal processing and machine learning problem, we propose our two-step algorithm: (i) compute an initial guess carefully; (ii) apply gradient descent to the objective function, starting with the carefully chosen initial guess. One difficulty of understanding non-convex optimization consists in how to construct the so-called *basin of attraction*, i.e., if the starting point is inside this basin of attraction, the iterates will always stay inside the region and converge to the global minimum. The construction of the basin of attraction varies for different problems [3,8,34]. For this problem, similar to [21], the construction follows from the following three observations. Each of these observations suggests the definition of a certain *neighborhood* and the basin of attraction is then defined as the intersection of these three neighborhood sets $\mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$.

1. **Ambiguity of solution**: in fact, we can only recover $(\boldsymbol{h}_i, \boldsymbol{x}_i)$ up to a scalar since $(\alpha \boldsymbol{h}_i, \alpha^{-1} \boldsymbol{x}_i)$ and $(\boldsymbol{h}_i, \boldsymbol{x}_i)$ are both solutions for $\alpha \neq 0$. From a numerical perspective, we want to avoid the scenario when $\|\boldsymbol{h}_i\| \to 0$ and $\|\boldsymbol{x}_i\| \to \infty$ while $\|\boldsymbol{h}_i\|\|\boldsymbol{x}_i\|$ is fixed, which potentially leads to numerical instability. To balance both the norm of $\|\boldsymbol{h}_i\|$ and $\|\boldsymbol{x}_i\|$ for all $1 \leq i \leq s$, we define

$$\mathcal{N}_d := \left\{ \{(\boldsymbol{h}_i, \boldsymbol{x}_i)\}_{i=1}^{s} : \|\boldsymbol{h}_i\| \leq 2\sqrt{d_{i0}}, \|\boldsymbol{x}_i\| \leq 2\sqrt{d_{i0}}, 1 \leq i \leq s \right\},$$

   which is a convex set.

2. **Incoherence**: the performance depends on how large/small the incoherence μ_h^2 is, where μ_h^2 is defined by

$$\mu_h^2 := \max_{1 \leq i \leq s} \frac{L\|\boldsymbol{B}\boldsymbol{h}_{i0}\|_\infty^2}{\|\boldsymbol{h}_{i0}\|^2}.$$

The idea is that: *the smaller the μ_h^2 is, the better the performance is.* Let us consider an extreme case: if $\boldsymbol{B}\boldsymbol{h}_{i0}$ is highly sparse or spiky, we lose much information on those zero/small entries and cannot hope to get satisfactory recovered signals. In other words, we need the ground truth $\boldsymbol{h}_{i0}$ has 'spectral flatness' and $\boldsymbol{h}_{i0}$ is not highly localized on the Fourier domain.

A similar quantity is also introduced in the matrix completion problem [7,34]. The larger μ_h^2 is, the more $\boldsymbol{h}_{i0}$ is aligned with one particular row of $\boldsymbol{B}$. To control the incoherence between $\boldsymbol{b}_l$ and $\boldsymbol{h}_i$, we define the second neighborhood,

$$\mathcal{N}_\mu := \left\{ \{\boldsymbol{h}_i\}_{i=1}^s : \sqrt{L}\|\boldsymbol{B}\boldsymbol{h}_i\|_\infty \leq 4\sqrt{d_{i0}}\mu, 1 \leq i \leq s \right\}, \tag{2.13}$$

where μ is a parameter and $\mu \geq \mu_h$. Note that $\mathcal{N}_\mu$ is also a convex set.

3. **Close to the ground truth**: we also want to construct an initial guess such that it is close to the ground truth, i.e.,

$$\mathcal{N}_\epsilon := \left\{ \{(\boldsymbol{h}_i, \boldsymbol{x}_i)\}_{i=1}^s : \delta_i = \frac{\|\boldsymbol{h}_i\boldsymbol{x}_i^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F}{d_{i0}} \leq \varepsilon, 1 \leq i \leq s \right\}, \tag{2.14}$$

where ε is a predetermined parameter in $(0, \frac{1}{15}]$.

REMARK 2.3 To ensure $\delta_i \leq \varepsilon$, it suffices to ensure $\delta \leq \frac{\varepsilon}{\sqrt{s\kappa}}$ where $\kappa := \frac{\max d_{i0}}{\min d_{i0}} \geq 1$. This is because

$$\frac{1}{s\kappa^2}\sum_{i=1}^s \delta_i^2 \leq \delta^2 \leq \frac{\varepsilon^2}{s\kappa^2}$$

which implies $\max_{1 \leq i \leq s} \delta_i \leq \varepsilon$.

REMARK 2.4 When we say $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_d, \mathcal{N}_\mu$ or $\mathcal{N}_\epsilon$, it means for all $i = 1, \ldots, s$ we have $(\boldsymbol{h}_i, \boldsymbol{x}_i) \in \mathcal{N}_d$, $\mathcal{N}_\mu$ or $\mathcal{N}_\epsilon$, respectively. In particular, $(\boldsymbol{h}_0, \boldsymbol{x}_0) \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$ where $\boldsymbol{h}_0$ and $\boldsymbol{x}_0$ are defined in (2.5).

2.3 *Objective function and Wirtinger derivative*

To implement the first two observations, we introduce the regularizer $G(\boldsymbol{h}, \boldsymbol{x})$, defined as the sum of s components

$$G(\boldsymbol{h}, \boldsymbol{x}) := \sum_{i=1}^s G_i(\boldsymbol{h}_i, \boldsymbol{x}_i). \tag{2.15}$$

For each component $G_i(\boldsymbol{h}_i, \boldsymbol{x}_i)$, we let $\rho \geq d^2 + 2\|\boldsymbol{e}\|^2$, $0.9d_0 \leq d \leq 1.1d_0$, $0.9d_{i0} \leq d_i \leq 1.1d_{i0}$ for all $1 \leq i \leq s$ and

$$G_i := \rho \left[\underbrace{G_0\left(\frac{\|\boldsymbol{h}_i\|^2}{2d_i}\right) + G_0\left(\frac{\|\boldsymbol{x}_i\|^2}{2d_i}\right)}_{\mathcal{N}_d} + \underbrace{\sum_{l=1}^L G_0\left(\frac{L|\boldsymbol{b}_l^*\boldsymbol{h}_i|^2}{8d_i\mu^2}\right)}_{\mathcal{N}_\mu} \right], \tag{2.16}$$

where $G_0(z) = \max\{z - 1, 0\}^2$. Here both d and $\{d_i\}_{i=1}^s$ are data-driven and well approximated by our spectral initialization procedure; and μ^2 is a tuning parameter which could be estimated if we assume a specific statistical model for the channel (for example, in the widely used Rayleigh fading model, the channel coefficients are assumed to be complex Gaussian). The idea behind G_i is quite straightforward though the formulation is complicated. For each G_i in (17), the first two terms try to force the iterates to lie in $\mathcal{N}_d$ and the third term tries to encourage the iterates to lie in $\mathcal{N}_\mu$. What about the neighborhood $\mathcal{N}_\epsilon$? A proper choice of the initialization followed by gradient descent which keeps the objective function decreasing will ensure that the iterates stay in $\mathcal{N}_\epsilon$.

Finally, we consider the objective function as the sum of nonlinear least squares objective function $F(h, x)$ in (2.11) and the regularizer $G(h, x)$,

$$\widetilde{F}(\boldsymbol{h}, \boldsymbol{x}) := F(\boldsymbol{h}, \boldsymbol{x}) + G(\boldsymbol{h}, \boldsymbol{x}). \tag{2.17}$$

Note that the input of the function $\widetilde{F}(\boldsymbol{h}, \boldsymbol{x})$ consists of complex variables, but the output is real-valued. As a result, the following simple relations hold

$$\frac{\partial \widetilde{F}}{\partial \bar{\boldsymbol{h}}_i} = \overline{\frac{\partial \widetilde{F}}{\partial \boldsymbol{h}_i}}, \quad \frac{\partial \widetilde{F}}{\partial \bar{\boldsymbol{x}}_i} = \overline{\frac{\partial \widetilde{F}}{\partial \boldsymbol{x}_i}}.$$

Similar properties also apply to both $F(\boldsymbol{h}, \boldsymbol{x})$ and $G(\boldsymbol{h}, \boldsymbol{x})$.

Therefore, to minimize this function, it suffices to consider only the gradient of $\widetilde{F}$ with respect to $\bar{\boldsymbol{h}}_i$ and $\bar{\boldsymbol{x}}_i$, which is also called Wirtinger derivative [8]. The Wirtinger derivatives of $F(\boldsymbol{h}, \boldsymbol{x})$ and $G(\boldsymbol{h}, \boldsymbol{x})$ w.r.t. $\bar{\boldsymbol{h}}_i$ and $\bar{\boldsymbol{x}}_i$ can be easily computed as follows

$$\nabla F_{\boldsymbol{h}_i} = \mathcal{A}_i^* \left(\mathcal{A}(X) - \boldsymbol{y} \right) \boldsymbol{x}_i = \mathcal{A}_i^* \left(\mathcal{A}(X - X_0) - \boldsymbol{e} \right) \boldsymbol{x}_i, \tag{2.18}$$

$$\nabla F_{\boldsymbol{x}_i} = \left(\mathcal{A}_i^* \left(\mathcal{A}(X) - \boldsymbol{y} \right) \right)^* \boldsymbol{h}_i = \left(\mathcal{A}_i^* \left(\mathcal{A}(X - X_0) - \boldsymbol{e} \right) \right)^* \boldsymbol{h}_i, \tag{2.19}$$

$$\nabla G_{\boldsymbol{h}_i} = \frac{\rho}{2d_i} \left[G_0' \left(\frac{\|\boldsymbol{h}_i\|^2}{2d_i} \right) \boldsymbol{h}_i + \frac{L}{4\mu^2} \sum_{l=1}^L G_0' \left(\frac{L|\boldsymbol{b}_l^* \boldsymbol{h}_i|^2}{8d_i \mu^2} \right) \boldsymbol{b}_l \boldsymbol{b}_l^* \boldsymbol{h}_i \right], \tag{2.20}$$

$$\nabla G_{\boldsymbol{x}_i} = \frac{\rho}{2d_i} G_0' \left(\frac{\|\boldsymbol{x}_i\|^2}{2d_i} \right) \boldsymbol{x}_i, \tag{2.21}$$

where $\mathcal{A}(X) = \sum_{i=1}^s \mathcal{A}_i(\boldsymbol{h}_i \boldsymbol{x}_i^*)$ and $\mathcal{A}^*$ is defined in (2.8). In short, we denote

$$\nabla \widetilde{F}_{\boldsymbol{h}} := \nabla F_{\boldsymbol{h}} + \nabla G_{\boldsymbol{h}}, \quad \nabla F_{\boldsymbol{h}} := \begin{bmatrix} \nabla F_{\boldsymbol{h}_1} \\ \vdots \\ \nabla F_{\boldsymbol{h}_s} \end{bmatrix}, \quad \nabla G_{\boldsymbol{h}} := \begin{bmatrix} \nabla G_{\boldsymbol{h}_1} \\ \vdots \\ \nabla G_{\boldsymbol{h}_s} \end{bmatrix}. \tag{2.22}$$

Similar definitions hold for $\nabla \widetilde{F}_{\boldsymbol{x}}$, $\nabla F_{\boldsymbol{x}}$ and $G_{\boldsymbol{x}}$. It is easy to see that $\nabla F_{\boldsymbol{h}} = \mathcal{A}^*(\mathcal{A}(X) - \boldsymbol{y})\boldsymbol{x}$ and $\nabla F_{\boldsymbol{x}} = (\mathcal{A}^*(\mathcal{A}(X) - \boldsymbol{y}))^* \boldsymbol{h}$.

3. Algorithm and theory

3.1 *Two-step algorithm*

As mentioned before, the first step is to find a good initial guess $(\boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}) \in \mathbb{C}^{Ks} \times \mathbb{C}^{Ns}$ such that it is inside the basin of attraction. The initialization follows from this key fact:

$$\mathbb{E}(\mathcal{A}_i^*(\boldsymbol{y})) = \mathbb{E}\left(\mathcal{A}_i^*\left(\sum_{j=1}^{s}\mathcal{A}_j(\boldsymbol{h}_{j0}\boldsymbol{x}_{j0}^*) + \boldsymbol{e}\right)\right) = \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*,$$

where we use $\boldsymbol{B}^*\boldsymbol{B} = \sum_{l=1}^{L}\boldsymbol{b}_l\boldsymbol{b}_l^* = \boldsymbol{I}_K$, $\mathbb{E}(\boldsymbol{a}_{il}\boldsymbol{a}_{il}^*) = \boldsymbol{I}_N$ and

$$\mathbb{E}(\mathcal{A}_i^*\mathcal{A}_i(\boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*)) = \sum_{l=1}^{L}\boldsymbol{b}_l\boldsymbol{b}_l^*\boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^* \, \mathbb{E}(\boldsymbol{a}_{il}\boldsymbol{a}_{il}^*) = \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*,$$

$$\mathbb{E}(\mathcal{A}_j^*\mathcal{A}_i(\boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*)) = \sum_{l=1}^{L}\boldsymbol{b}_l\boldsymbol{b}_l^*\boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^* \, \mathbb{E}(\boldsymbol{a}_{il}\boldsymbol{a}_{jl}^*) = \boldsymbol{0}, \quad \forall j \neq i.$$

Therefore, it is natural to extract the leading singular value and associated left and right singular vectors from each $\mathcal{A}_i^*(\boldsymbol{y})$ and use them as (a hopefully good) approximation to $(d_{i0}, \boldsymbol{h}_{i0}, \boldsymbol{x}_{i0})$. This idea leads to Algorithm 1, the theoretic guarantees of which are given in Section 6.5. The second step of the algorithm is just to apply gradient descent to $\widetilde{F}$ with the initial guess $\{(\boldsymbol{u}_i^{(0)}, \boldsymbol{v}_i^{(0)}, d_i)\}_{i=1}^{s}$ or $(\boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \{d_i\}_{i=1}^{s})$, where $\boldsymbol{u}^{(0)}$ stems from stacking all $\boldsymbol{u}_i^{(0)}$ into one long vector.[4]

1: **for** $i = 1, 2, \ldots, s$ **do**
2:     Compute $\mathcal{A}_i^*(\boldsymbol{y})$.
3:     Find the leading singular value, left and right singular vectors of $\mathcal{A}_i^*(\boldsymbol{y})$, denoted by $(d_i, \hat{\boldsymbol{h}}_{i0}, \hat{\boldsymbol{x}}_{i0})$.
4:     Solve the following optimization problem for $1 \leq i \leq s$:

$$\boldsymbol{u}_i^{(0)} := \operatorname{argmin}_{\boldsymbol{z} \in \mathbb{C}^K} \|\boldsymbol{z} - \sqrt{d_i}\hat{\boldsymbol{h}}_{i0}\|^2 \text{ s.t. } \sqrt{L}\|\boldsymbol{B}\boldsymbol{z}\|_\infty \leq 2\sqrt{d_i}\mu.$$

5:     Set $\boldsymbol{v}_i^{(0)} = \sqrt{d_i}\hat{\boldsymbol{x}}_{i0}$.
6: **end for**
7: Output: $\{(\boldsymbol{u}_i^{(0)}, \boldsymbol{v}_i^{(0)}, d_i)\}_{i=1}^{s}$ or $(\boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \{d_i\}_{i=1}^{s})$.

Algorithm 1: Initialization via spectral method and projection

REMARK 3.1 For Algorithm 2, we can rewrite each iteration into

$$\boldsymbol{u}^{(t)} = \boldsymbol{u}^{(t-1)} - \eta\nabla\widetilde{F}_{\boldsymbol{h}}(\boldsymbol{u}^{(t-1)}, \boldsymbol{v}^{(t-1)}), \quad \boldsymbol{v}^{(t)} = \boldsymbol{v}^{(t-1)} - \eta\nabla\widetilde{F}_{\boldsymbol{x}}(\boldsymbol{u}^{(t-1)}, \boldsymbol{v}^{(t-1)}),$$

---

[4] It is clear that instead of gradient descent one could also use a second-order method to achieve faster convergence at the tradeoff of increased computational cost per iteration. The theoretical convergence analysis for a second-order method will require a very different approach from the one developed in this paper.

1: **Initialization:** obtain $(\boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}, \{d_i\}_{i=1}^s)$ via Algorithm 1.
2: **for** $t = 1, 2, \ldots,$ **do**
3:      **for** $i = 1, 2, \ldots, s$ **do**
4:          $\boldsymbol{u}_i^{(t)} = \boldsymbol{u}_i^{(t-1)} - \eta \nabla \widetilde{F}_{\boldsymbol{h}_i}(\boldsymbol{u}_i^{(t-1)}, \boldsymbol{v}_i^{(t-1)}),$
5:          $\boldsymbol{v}_i^{(t)} = \boldsymbol{v}_i^{(t-1)} - \eta \nabla \widetilde{F}_{\boldsymbol{x}_i}(\boldsymbol{u}_i^{(t-1)}, \boldsymbol{v}_i^{(t-1)}),$
6:      **end for**
7: **end for**

Algorithm 2: Wirtinger gradient descent with constant stepsize η_{s}

where $\nabla \widetilde{F}_{\boldsymbol{h}}$ and $\nabla \widetilde{F}_{\boldsymbol{x}}$ are in (2.22), and

$$
\boldsymbol{u}^{(t)} := \begin{bmatrix} \boldsymbol{u}_1^{(t)} \\ \vdots \\ \boldsymbol{u}_s^{(t)} \end{bmatrix}, \quad \boldsymbol{v}^{(t)} := \begin{bmatrix} \boldsymbol{v}_1^{(t)} \\ \vdots \\ \boldsymbol{v}_s^{(t)} \end{bmatrix}.
$$

3.2 *Main results*

Our main findings are summarized as follows: Theorem 3.2 shows that the initial guess given by Algorithm 1 indeed belongs to the basin of attraction. Moreover, d_i also serves as a good approximation of d_{i0} for each i. Theorem 3.3 demonstrates that the regularized Wirtinger gradient descent will guarantee the linear convergence of the iterates and the recovery is exact in the noisefree case and stable in the presence of noise.

THEOREM 3.2 The initialization obtained via Algorithm 1 satisfies

$$
(\boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}) \in \frac{1}{\sqrt{3}} \mathcal{N}_d \bigcap \frac{1}{\sqrt{3}} \mathcal{N}_\mu \bigcap \mathcal{N}_{\frac{2\varepsilon}{5\sqrt{s\kappa}}} \tag{3.1}
$$

and

$$
0.9 d_{i0} \leq d_i \leq 1.1 d_{i0}, \quad 0.9 d_0 \leq d \leq 1.1 d_0, \tag{3.2}
$$

holds with probability at least $1 - L^{-\gamma+1}$ if the number of measurements satisfies

$$
L \geq C_{\gamma + \log(s)} (\mu_h^2 + \sigma^2) s^2 \kappa^4 \max\{K, N\} \log^2 L / \varepsilon^2. \tag{3.3}
$$

Here ε is any predetermined constant in $(0, \frac{1}{15}]$, and C_γ is a constant only linearly depending on γ with $\gamma \geq 1$.

THEOREM 3.3 Starting with the initial value $\boldsymbol{z}^{(0)} := (\boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)})$ satisfying (3.1), the Algorithm 2 creates a sequence of iterates $(\boldsymbol{u}^{(t)}, \boldsymbol{v}^{(t)})$ which converges to the global minimum linearly,

$$
\|\mathcal{H}(\boldsymbol{u}^{(t)}, \boldsymbol{v}^{(t)}) - \mathcal{H}(\boldsymbol{h}_0, \boldsymbol{x}_0)\|_F \leq \frac{\varepsilon d_0}{\sqrt{2s\kappa^2}} (1 - \eta\omega)^{t/2} + 60\sqrt{s} \|\mathcal{A}^*(\boldsymbol{e})\| \tag{3.4}
$$

with probability at least $1 - L^{-\gamma+1}$ where $\eta\omega = \mathcal{O}((s\kappa d_0(K+N)\log^2 L)^{-1})$ and

$$\|\mathcal{A}^*(e)\| \leq C_0\sigma d_0\sqrt{\frac{\gamma s(K+N)(\log^2 L)}{L}}$$

if the number of measurements L satisfies

$$L \geq C_{\gamma+\log(s)}(\mu^2 + \sigma^2)s^2\kappa^4 \max\{K,N\}\log^2 L/\varepsilon^2. \tag{3.5}$$

REMARK 3.4 Our previous work [23] shows that the convex approach via semidefinite programming (see (2.10)) requires $L \geq C_0s^2(K + \mu_h^2 N)\log^3(L)$ to ensure exact recovery. Later, [15] improves this result to the near-optimal bound $L \geq C_0s(K + \mu_h^2 N)$ up to some log-factors. The difference between non-convex and convex methods lies in the appearance of the condition number κ in (3.5). This is not just an artifact of the proof—empirically we also observe that the value of κ affects the convergence rate of our non-convex algorithm, see Fig. 5.

REMARK 3.5 Our theory suggests s^2-dependence for the number of measurements L, although numerically L in fact depends on s linearly, as shown in Section 4. The reason for s^2-dependence will be addressed in details in Section 5.2.

REMARK 3.6 In the theoretical analysis, we assume that A_i (or equivalently C_i) is a Gaussian random matrix. Numerical simulations suggest that this assumption is clearly not necessary. For example, C_i may be chosen to be a Hadamard-type matrix which is more appropriate and favorable for communications.

REMARK 3.7 If $e = 0$, (3.4) shows that $(u^{(t)}, v^{(t)})$ converges to the ground truth at a linear rate. On the other hand, if noise exists, $(u^{(t)}, v^{(t)})$ is guaranteed to converge to a point within a small neighborhood of (h_0, x_0). More importantly, if the number of measurements L gets larger, $\|\mathcal{A}^*(e)\|$ decays at the rate of $\mathcal{O}(L^{-1/2})$.

4. Numerical simulations

In this section we present a range of numerical simulations to illustrate and complement different aspects of our theoretical framework. We will empirically analyze the number of measurements needed for perfect joint deconvolution/demixing to see how this compares to our theoretical bounds. We will also study the robustness for noisy data. In our simulations we use Gaussian encoding matrices, as in our theorems. But we also try more realistic structured encoding matrices, that are more reminiscent of what one might come across in wireless communications.

    While Theorem 3.3 says that the number of measurements L depends *quadratically* on the number of sources s, numerical simulations suggest near-optimal performance. Figure 2 demonstrates that L actually depends linearly on s, i.e., the boundary between success (white) and failure (black) is approximately a linear function of s. In the experiment, $K = N = 50$ are fixed, all A_i are complex Gaussians and all (h_i, x_i) are standard complex Gaussian vectors. For each pair of (L, s), 25 experiments are performed and we treat the recovery as a success if $\frac{\|\hat{X}-X_0\|_F}{\|X_0\|_F} \leq 10^{-3}$. For our algorithm, we use backtracking to determine the stepsize and the iteration stops either if $\|\mathcal{A}(\mathcal{H}(h^{(t+1)}, x^{(t+1)}) - \mathcal{H}(h^{(t)}, x^{(t)}))\| < 10^{-6}\|y\|$ or if the number of iterations reaches 500. The backtracking is based on the
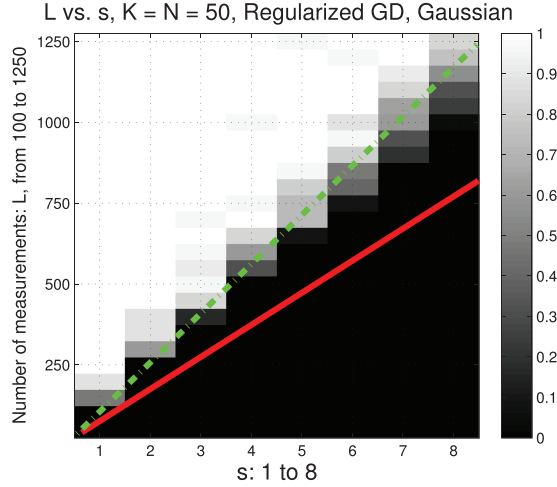
Fig. 2. Phase transition plot for empirical recovery performance under different choices of (L, s) where $K = N = 50$ are fixed. Black region: failure; white region: success. The red solid line depicts the number of degrees of freedom and the green dashed line shows the empirical phase transition bound for Algorithm 2.

Armijo–Goldstein condition [25]. The initial stepsize is chosen to be $\eta = \frac{1}{K+N}$. If $\widetilde{F}(z^{(t)} - \eta \nabla \widetilde{F}(z^{(t)})) > \widetilde{F}(z^{(t)})$, we just divide η by two and use a smaller stepsize.

We see from Fig. 2 that the number of measurements for the proposed algorithm to succeed not only seems to depend linearly on the number of sensors, but it is actually rather close to the information-theoretic limit $s(K + N)$. Indeed, the green dashed line in Fig. 2, which represents the empirical boundary for the phase transition between success and failure, corresponds to a line with slope about $\frac{3}{2} s(K+N)$. It is interesting to compare this empirical performance to the sharp theoretical phase transition bounds one would obtain via convex optimization [10,26]. Considering the convex approach based on lifting in [23], we can adapt the theoretical framework in [10] to the blind deconvolution/demixing setting, but with one modification. The bounds in [10] rely on Gaussian widths of tangent cones related to the measurement matrices $\mathcal{A}_i$. Since simply analytic formulas for these expressions seem to be out of reach for the structured rank-one measurement matrices used in our paper, we instead compute the bounds for full-rank Gaussian random matrices, which yields a sharp bound of about $3s(K + N)$ (the corresponding bounds for rank-one sensing matrices will likely have a constant larger than 3). Note that these sharp theoretical bounds predict quite accurately the empirical behavior of convex methods. Thus our empirical bound for using a non-convex methods compares rather favorably with that of the convex approach.

Similar conclusions can be drawn from Fig. 3; there all A_i are in the form of $A_i = FD_iH$ where F is the unitary $L \times L$ DFT matrix, all D_i are independent diagonal binary ± 1 matrices and H is an $L \times N$ fixed partial deterministic Hadamard matrix. The purpose of using D_i is to enhance the incoherence between each channel so that our algorithm is able to tell apart each individual signal and channel. As before we assume Gaussian channels, i.e., $h_i \sim \mathcal{CN}(\mathbf{0}, I_K)$. Therefore, our approach does not only work for Gaussian encoding matrices A_i, but also for the matrices that are interesting to real-world applications, although no satisfactory theory has been derived yet for that case. Moreover, due to the structure of A_i and B, fast transform algorithms are available, potentially allowing for real-time deployment.
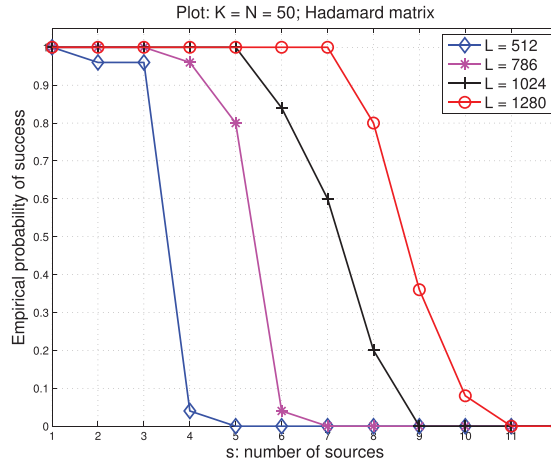
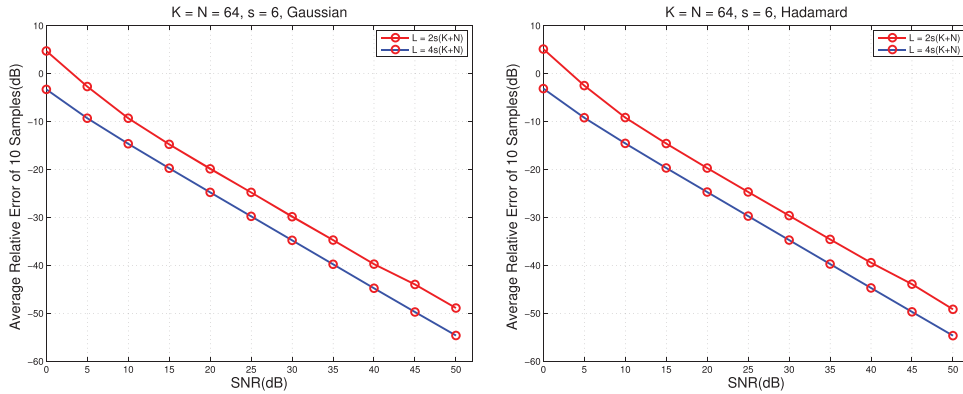FIG. 3. Empirical probability of successful recovery for different pairs of (L, s) when $K = N = 50$ are fixed.

FIG. 4. Relative error vs. SNR (dB): SNR $= 20 \log_{10} \left(\frac{\|\boldsymbol{y}\|}{\|\boldsymbol{e}\|} \right)$.

Figure 4 shows the robustness of our algorithm under different levels of noise. We also run 25 samples for each level of SNR and different L, and then compute the average relative error. It is easily seen that the relative error scales linearly with the SNR and one unit of increase in SNR (in dB) results in one unit of decrease in the relative error.

Theorem 3.3 suggests that the performance and convergence rate actually depend on the condition number of $\boldsymbol{X}_0 = \mathcal{H}(\boldsymbol{h}_0, \boldsymbol{x}_0)$, i.e., on $\kappa = \frac{\max d_{i0}}{\min d_{i0}}$ where $d_{i0} = \|\boldsymbol{h}_{i0}\| \|\boldsymbol{x}_{i0}\|$. Next we demonstrate that this dependence on the condition number is not an artifact of the proof, but is indeed also observed empirically. In this experiment, we let $s = 2$ and set for the first component $d_{1,0} = 1$ and for the second one $d_{2,0} = \kappa$ for $\kappa \in \{1, 2, 5\}$. Here, $\kappa = 1$ means that the received signals of both sensors have equal power, whereas $\kappa = 5$ means that the signal received from the second sensor is considerably stronger. The initial stepsize is chosen as $\eta = 1$, followed by the backtracking scheme. Figure 5 shows how the relative error decays with respect to the number of iterations t under different condition number κ and L.
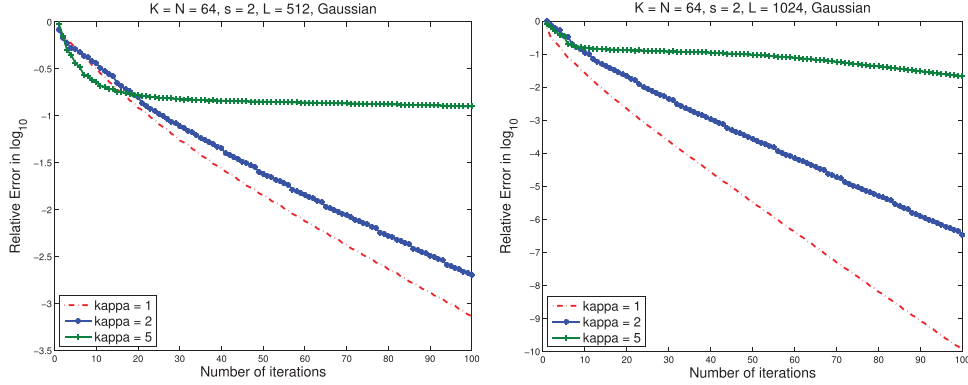
FIG. 5. Relative error vs. number of iterations t.

The larger κ is, the slower the convergence rate is, as we see from Fig. 5. This may result from two reasons: our spectral initialization may not be able to give a good initial guess for those weak components; moreover, during the gradient descent procedure, the gradient directions for the weak components could be totally dominated/polluted by the strong components. Currently, we still have no effective way of how to deal with this issue of slow convergence when κ is not small. We have to leave this topic for future investigations.

5. Convergence analysis

Our convergence analysis relies on the following four conditions where the first three of them are local properties. We will also briefly discuss how they contribute to the proof of our main theorem. Note that our previous work [21] on blind deconvolution is actually a special case ($s = 1$) of (2.1). The proof of Theorem 3.3 follows in part the main ideas in [21]. The readers may find the technical parts of [21] and this manuscript share many similarities. However, there are also important differences. After all, we are now dealing with a more complicated problem where the ground truth matrix X_0 and measurement matrices are both rank-s block-diagonal matrices, as shown in (2.4), instead of rank-one matrices in [21]. The key is to understand the properties of the linear operator $\mathcal{A}$ applying to different types of *block-diagonal* matrices. Therefore, many technical details are much more involved while on the other hand, some of results in [21] can be used directly. During the presentation, we will clearly point out both the similarities to and differences from [21].

5.1   *Four key conditions*

CONDITION 5.1 **Local regularity condition:** let $z := (h, x) \in \mathbb{C}^{s(K+N)}$ and $\nabla \widetilde{F}(z) := \begin{bmatrix} \nabla \widetilde{F}_h(z) \\ \nabla \widetilde{F}_x(z) \end{bmatrix} \in \mathbb{C}^{s(K+N)}$, then

$$\|\nabla \widetilde{F}(z)\|^2 \geq \omega[\widetilde{F}(z) - c]_+ \tag{5.1}$$

for $z \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$ where $\omega = \frac{d_0}{7000}$ and $c = \|e\|^2 + 2000s\|\mathcal{A}^*(e)\|^2$.

We will prove Condition 5.1 in Section 6.3. Condition 5.1 states that $\widetilde{F}(z) = 0$ if $\|\nabla \widetilde{F}(z)\| = 0$ and $e = 0$, i.e., all the stationary points inside the basin of attraction are global minima.

CONDITION 5.2 **Local smoothness condition:** let $z = (h, x)$ and $w = (u, v)$ and there holds

$$\|\widetilde{F}(z + w) - \widetilde{F}(z)\| \le C_L \|w\| \tag{5.2}$$

for $z + w$ and z inside $\mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$ where $C_L \approx \mathcal{O}(d_0 s\kappa (1 + \sigma^2)(K + N) \log^2 L)$ is the Lipschitz constant of $\widetilde{F}$ over $\mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$. The convergence rate is governed by C_L.

The proof of Condition 5.2 can be found in Section 6.4.

CONDITION 5.3 **Local restricted isometry property:** denote $X = \mathcal{H}(h, x)$ and $X_0 = \mathcal{H}(h_0, x_0)$. There holds

$$\frac{2}{3}\|X - X_0\|_F^2 \le \|\mathcal{A}(X - X_0)\|^2 \le \frac{3}{2}\|X - X_0\|_F^2 \tag{5.3}$$

uniformly all for $(h, x) \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$.

Condition 5.3 will be proven in Section 6.2. It says that the convergence of the objective function implies the convergence of the iterates.

REMARK 5.4 (**Necessity of inter-user incoherence**). Although Condition 5.3 is seemingly the same as the one in our previous work [21], it is indeed very different. Recall that $\mathcal{A}$ is a linear operator acting on block-diagonal matrices and its output is the sum of s different components involving $\mathcal{A}_i$. Therefore, the proof of Condition 5.3 heavily depends on the inter-user incoherence, whereas this notion of incoherence is not needed at all for the single-user scenario. At the beginning of Section 2, we discuss the choice of C_i (or A_i). In order to distinguish one user from another, it is essential to use sufficiently different[5] encoding matrices C_i (or A_i). Here the independence and Gaussianity of all C_i (or A_i) guarantee that $\|\mathcal{P}_{T_i} \mathcal{A}_i^* \mathcal{A}_j \mathcal{P}_{T_j}\|$ is sufficiently small for all $i \ne j$ where T_i is defined in (6.1). It is a key element to ensure the validity of Condition 5.3 which is also an important component to prove Condition 5.1. On the other hand, due to the recent progress on this joint deconvolution and demixing problem, one is also able to prove a local restricted isometry property with tools such as bounding the suprema of chaos processes [15] by assuming $\{A_i\}_{i=1}^s$ as Gaussian matrices.

CONDITION 5.5 **Robustness condition:** let $\varepsilon \le \frac{1}{15}$ be a predetermined constant. We have

$$\|\mathcal{A}^*(e)\| = \max_{1 \le i \le s} \|\mathcal{A}_i^*(e)\| \le \frac{\varepsilon d_0}{10\sqrt{2}s\kappa}, \tag{5.4}$$

where $e \sim \mathcal{CN}\left(0, \frac{\sigma^2 d_0^2}{L}\right)$ if $L \ge C_\gamma \kappa^2 s^2 (K + N)/\varepsilon^2$.

We will prove Condition 5.5 in Section 6.5. We now extract one useful result based on Conditions 5.3 and 5.5. From these two conditions, we are able to produce a good approximation of $F(h, x)$ for all $(h, x) \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$ in terms of δ in (2.9). For $(h, x) \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$, the following inequality holds

$$\frac{2}{3}\delta^2 d_0^2 - \frac{\varepsilon \delta d_0^2}{5\sqrt{s}\kappa} + \|e\|^2 \le F(h, x) \le \frac{3}{2}\delta^2 d_0^2 + \frac{\varepsilon \delta d_0^2}{5\sqrt{s}\kappa} + \|e\|^2. \tag{5.5}$$

---

[5] Suppose all C_i are the same, there is no hope to recover all pairs of $\{(h_i, x_i)\}_{i=1}^s$ simultaneously.

Note that (5.5) simply follows from

$$F(\boldsymbol{h},\boldsymbol{x}) = \|\mathcal{A}(\boldsymbol{X}-\boldsymbol{X}_0)\|_F^2 - 2\operatorname{Re}(\langle \boldsymbol{X}-\boldsymbol{X}_0, \mathcal{A}^*(\boldsymbol{e})\rangle) + \|\boldsymbol{e}\|^2.$$

Note that (5.3) implies $\frac{2}{3}\delta^2 d_0^2 \leq \|\mathcal{A}(\boldsymbol{X}-\boldsymbol{X}_0)\|_F^2 \leq \frac{3}{2}\delta^2 d_0^2$. Thus it suffices to estimate the cross-term,

$$
\begin{aligned}
|\operatorname{Re}(\langle \boldsymbol{X}-\boldsymbol{X}_0, \mathcal{A}^*(\boldsymbol{e})\rangle)| &\leq \|\mathcal{A}^*(\boldsymbol{e})\|\|\boldsymbol{X}-\boldsymbol{X}_0\|_* = \|\mathcal{A}^*(\boldsymbol{e})\| \sum_{i=1}^s \|\boldsymbol{h}_i\boldsymbol{x}_i^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_* \\
&\leq \sqrt{2}\|\mathcal{A}^*(\boldsymbol{e})\| \sum_{i=1}^s \|\boldsymbol{h}_i\boldsymbol{x}_i^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F \\
&\leq \sqrt{2s}\|\mathcal{A}^*(\boldsymbol{e})\|\|\boldsymbol{X}-\boldsymbol{X}_0\|_F \leq \frac{\varepsilon\delta d_0^2}{10\sqrt{s\kappa}},
\end{aligned}
\tag{5.6}
$$

where $\|\cdot\|_*$ and $\|\cdot\|$ are a pair of dual norms and $\|\mathcal{A}^*(\boldsymbol{e})\|$ comes from (5.4).

5.2 *Outline of the convergence analysis*

For the ease of proof, we introduce another neighborhood:

$$\mathcal{N}_{\widetilde{F}} = \left\{ (\boldsymbol{h},\boldsymbol{x}) : \widetilde{F}(\boldsymbol{h},\boldsymbol{x}) \leq \frac{\varepsilon^2 d_0^2}{3s\kappa^2} + \|\boldsymbol{e}\|^2 \right\}.$$

Moreover, another reason to consider $\mathcal{N}_{\widetilde{F}}$ is based on the fact that gradient descent *only* allows one to make the objective function decrease if the step size is chosen appropriately. In other words, all the iterates $\boldsymbol{z}^{(t)}$ generated by gradient descent are inside $\mathcal{N}_{\widetilde{F}}$ as long as $\boldsymbol{z}^{(0)} \in \mathcal{N}_{\widetilde{F}}$.

On the other hand, it is crucial to note that the decrease of the objective function does not necessarily imply the decrease of the relative error of the iterates. Therefore, we want to construct an initial guess in $\mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ so that $\boldsymbol{z}^{(0)}$ is sufficiently close to the ground truth and then analyze the behavior of $\boldsymbol{z}^{(t)}$.

In the rest of this section, we basically try to prove the following relation:

$$\underbrace{\frac{1}{\sqrt{3}}\mathcal{N}_d \cap \frac{1}{\sqrt{3}}\mathcal{N}_\mu \cap \mathcal{N}_{\frac{2\varepsilon}{5\sqrt{s\kappa}}}}_{\text{Initial guess}} \subset \underbrace{\mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}}_{\{\boldsymbol{z}^{(t)}\}_{t\geq 0}\ \text{in}\ \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}} \subset \underbrace{\mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon}_{\text{Key conditions hold over}\ \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon}.$$

Now we give a more detailed explanation of the relation above, which constitutes the main structure of the proof:

1. We will show $\frac{1}{\sqrt{3}}\mathcal{N}_d \cap \frac{1}{\sqrt{3}}\mathcal{N}_\mu \cap \mathcal{N}_{\frac{2\varepsilon}{5\sqrt{s\kappa}}} \subset \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ in the proof of Theorem 3.3 in Section 5.3, which is quite straightforward.

2. Lemma 5.6 explains why it holds that $\mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}} \subset \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$ and where the s^2-bottleneck comes from.

3. Lemma 5.8 implicitly shows that the iterates $z^{(t)}$ will remain in $\mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ if the initial guess $z^{(0)}$ is inside $\mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ and $\widetilde{F}(z^{(t)})$ is monotonically decreasing (simply by induction). Lemma 5.9 makes this observation explicit by showing that $z^{(t)} \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ implies $z^{(t+1)} :=$ $z^{(t)} - \eta \nabla \widetilde{F}(z^{(t)}) \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ if the stepsize η obeys $\eta \leq \frac{1}{C_L}$. Moreover, Lemma 5.9 guarantees sufficient decrease of $\widetilde{F}(z^{(t)})$ in each iteration, which paves the road toward the proof of linear convergence of $\widetilde{F}(z^{(t)})$ and thus $z^{(t)}$.

Remember that $\mathcal{N}_d$ and $\mathcal{N}_\mu$ are both convex sets, and the purpose of introducing regularizers $G_i(\boldsymbol{h}_i, \boldsymbol{x}_i)$ is to approximately project the iterates onto $\mathcal{N}_d \cap \mathcal{N}_\mu$. Moreover, we hope that once the iterates are inside $\mathcal{N}_\epsilon$ and inside a sublevel subset $\mathcal{N}_{\widetilde{F}}$, they will never escape from $\mathcal{N}_{\widetilde{F}} \cap \mathcal{N}_\epsilon$. Those ideas are fully reflected in the following lemma.

LEMMA 5.6 Assume $0.9d_{i0} \leq d_i \leq 1.1d_{i0}$ and $0.9d_0 \leq d \leq 1.1d_0$. There holds $\mathcal{N}_{\widetilde{F}} \subset \mathcal{N}_d \cap \mathcal{N}_\mu$; moreover, under Conditions 5.3 and 5.5, we have $\mathcal{N}_{\widetilde{F}} \cap \mathcal{N}_\epsilon \subset \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_{\frac{9}{10}\epsilon}$.

Proof. If $(\boldsymbol{h}, \boldsymbol{x}) \notin \mathcal{N}_d \cap \mathcal{N}_\mu$, by the definition of G in (2.15), at least one component in G exceeds $\rho G_0 \left(\frac{2d_{i0}}{d_i} \right)$. We have

$$\widetilde{F}(\boldsymbol{h}, \boldsymbol{x}) \geq \rho G_0 \left(\frac{2d_{i0}}{d_i} \right) \geq (d^2 + 2\|\boldsymbol{e}\|^2) \left(\frac{2d_{i0}}{d_i} - 1 \right)^2$$

$$\geq (2/1.1 - 1)^2 (d^2 + 2\|\boldsymbol{e}\|^2)$$

$$\geq \frac{1}{2}d_0^2 + \|\boldsymbol{e}\|^2 > \frac{\varepsilon^2 d_0^2}{3s\kappa^2} + \|\boldsymbol{e}\|^2,$$

where $\rho \geq d^2 + 2\|\boldsymbol{e}\|^2$, $0.9d_0 \leq d \leq 1.1d_0$ and $0.9d_{i0} \leq d_i \leq 1.1d_{i0}$. This implies $(\boldsymbol{h}, \boldsymbol{x}) \notin \mathcal{N}_{\widetilde{F}}$ and hence $\mathcal{N}_{\widetilde{F}} \subset \mathcal{N}_d \cap \mathcal{N}_\mu$.

Note that $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$ if $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_{\widetilde{F}} \cap \mathcal{N}_\epsilon$. Applying (5.5) gives

$$\frac{2}{3}\delta^2 d_0^2 - \frac{\varepsilon \delta d_0^2}{5\sqrt{s\kappa}} + \|\boldsymbol{e}\|^2 \leq F(\boldsymbol{h}, \boldsymbol{x}) \leq \widetilde{F}(\boldsymbol{h}, \boldsymbol{x}) \leq \frac{\varepsilon^2 d_0^2}{3s\kappa^2} + \|\boldsymbol{e}\|^2$$

which implies that $\delta \leq \frac{9}{10} \frac{\varepsilon}{\sqrt{s\kappa}}$. By definition of δ in (2.9), there holds

$$\frac{81\varepsilon^2}{100s\kappa^2} \geq \delta^2 = \frac{\sum_{i=1}^{s} \delta_i^2 d_{i0}^2}{\sum_{i=1}^{s} d_{i0}^2} \geq \frac{\sum_{i=1}^{s} \delta_i^2}{s\kappa^2} \geq \frac{1}{s\kappa^2} \max_{1 \leq i \leq s} \delta_i^2, \tag{5.7}$$

which gives $\delta_i \leq \frac{9}{10}\varepsilon$ and $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_{\frac{9}{10}\varepsilon}$. □

REMARK 5.7 The s^2-bottleneck comes from (5.7). If $\delta \leq \varepsilon$ is small, we cannot guarantee that each δ_i is also smaller than ε. Just consider the simplest case when all d_{i0} are the same: then $d_0^2 = \sum_{i=1}^{s} d_{i0}^2 = s d_{i0}^2$ and there holds

$$\varepsilon^2 \geq \delta^2 = \frac{1}{s} \sum_{i=1}^{s} \delta_i^2.$$

Obviously, we cannot conclude that $\max \delta_i \leq \varepsilon$, but only say that $\delta_i \leq \sqrt{s}\varepsilon$. This is why we require $\delta = \mathcal{O}\left(\frac{\varepsilon}{\sqrt{s}}\right)$ to ensure $\delta_i \leq \varepsilon$, which gives s^2-dependence in L.

LEMMA 5.8 Denote $z_1 = (\boldsymbol{h}_1, \boldsymbol{x}_1)$ and $z_2 = (\boldsymbol{h}_2, \boldsymbol{x}_2)$. Let $z(\lambda) := (1 - \lambda)z_1 + \lambda z_2$. If $z_1 \in \mathcal{N}_\epsilon$ and $z(\lambda) \in \mathcal{N}_{\widetilde{F}}$ for all $\lambda \in [0, 1]$, we have $z_2 \in \mathcal{N}_\epsilon$.

Proof. Note that for $z_1 \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$, we have $z_1 \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_{\frac{9}{10}\varepsilon}$ which follows from the second part of Lemma 5.6. Now we prove $z_2 \in \mathcal{N}_\epsilon$ by contradiction. Let us suppose that $z_2 \notin \mathcal{N}_\epsilon$ and $z_1 \in \mathcal{N}_\epsilon$. There exists $z(\lambda_0) := (\boldsymbol{h}(\lambda_0), \boldsymbol{x}(\lambda_0)) \in \mathcal{N}_\epsilon$ for some $\lambda_0 \in [0, 1]$ such that $\max_{1 \leq i \leq s} \frac{\|\boldsymbol{h}_i \boldsymbol{x}_i^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F}{d_{i0}} = \epsilon$. Therefore, $z(\lambda_0) \in \mathcal{N}_{\widetilde{F}} \cap \mathcal{N}_\epsilon$ and Lemma 5.6 implies $\max_{1 \leq i \leq s} \frac{\|\boldsymbol{h}_i \boldsymbol{x}_i^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F}{d_{i0}} \leq \frac{9}{10}\epsilon$, which contradicts $\max_{1 \leq i \leq s} \frac{\|\boldsymbol{h}_i \boldsymbol{x}_i^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F}{d_{i0}} = \epsilon$.   □

LEMMA 5.9 Let the stepsize $\eta \leq \frac{1}{C_L}$, $z^{(t)} := (\boldsymbol{u}^{(t)}, \boldsymbol{v}^{(t)}) \in \mathbb{C}^{s(K+N)}$ and C_L be the Lipschitz constant of $\nabla\widetilde{F}(z)$ over $\mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$ in (5.2). If $z^{(t)} \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$, we have $z^{(t+1)} \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ and

$$\widetilde{F}(z^{(t+1)}) \leq \widetilde{F}(z^{(t)}) - \eta\|\nabla\widetilde{F}(z^{(t)})\|^2, \tag{5.8}$$

where $z^{(t+1)} = z^{(t)} - \eta\nabla\widetilde{F}(z^{(t)})$.

REMARK 5.10 This lemma tells us that once $z^{(t)} \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$, the next iterate $z^{(t+1)} = z^{(t)} - \eta\nabla\widetilde{F}(z^{(t)})$ is also inside $\mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ as long as the stepsize $\eta \leq \frac{1}{C_L}$. In other words, $\mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ is in fact a stronger version of the basin of attraction. Moreover, the objective function will decay sufficiently in each step as long as we can control the lower bound of the $\nabla\widetilde{F}$, which is guaranteed by the Local Regularity Condition 5.3.

Proof. Let $\phi(\tau) := \widetilde{F}(z^{(t)} - \tau\nabla\widetilde{F}(z^{(t)}))$, $\phi(0) = \widetilde{F}(z^{(t)})$ and consider the following quantity:

$$\tau_{\max} := \max\{\mu : \phi(\tau) \leq \widetilde{F}(z^{(t)}), 0 \leq \tau \leq \mu\},$$

where $\tau_{\max}$ is the largest stepsize such that the objective function $\widetilde{F}(z)$ evaluated at any point over the whole line segment $\{z^{(t)} - \tau\widetilde{F}(z^{(t)}), 0 \leq \tau \leq \tau_{\max}\}$ is not greater than $\widetilde{F}(z^{(t)})$. Now we will show $\tau_{\max} \geq \frac{1}{C_L}$. Obviously, if $\|\nabla\widetilde{F}(z^{(t)})\| = 0$, it holds automatically.

Consider $\|\nabla\widetilde{F}(z^{(t)})\| \neq 0$ and assume $\tau_{\max} < \frac{1}{C_L}$. First note that,

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\phi(\tau) < 0 \Longrightarrow \tau_{\max} > 0.$$

By the definition of $\tau_{\max}$, there holds $\phi(\tau_{\max}) = \phi(0)$ since $\phi(\tau)$ is a continuous function w.r.t. τ. Lemma 5.8 implies

$$\{z^{(t)} - \tau\nabla\widetilde{F}(z^{(t)}), 0 \leq \tau \leq \tau_{\max}\} \subseteq \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}.$$

Now we apply Lemma 6.20, the modified descent lemma, and obtain

$$\widetilde{F}(z^{(t)} - \tau_{\max}\nabla\widetilde{F}(z^{(t)})) \leq \widetilde{F}(z^{(t)}) - (2\tau_{\max} - C_L\tau_{\max}^2)\|\widetilde{F}(z^{(t)})\|^2 \leq \widetilde{F}(z^{(t)}) - \tau_{\max}\|\widetilde{F}(z^{(t)})\|^2,$$

where $C_L \tau_{\max} \leq 1$. In other words, $\phi(\tau_{\max}) \leq \widetilde{F}(z^{(t)} - \tau_{\max} \nabla \widetilde{F}(z^{(t)})) < \widetilde{F}(z^{(t)}) = \phi(0)$ contradicts $\phi(\tau_{\max}) = \phi(0)$.

Therefore, we conclude that $\tau_{\max} \geq \frac{1}{C_L}$. For any $\eta \leq \frac{1}{C_L}$, Lemma 5.8 implies

$$\{z^{(t)} - \tau \nabla \widetilde{F}(z^{(t)}), 0 \leq \tau \leq \eta\} \subseteq \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$$

and applying Lemma 6.20 gives

$$\widetilde{F}(z^{(t)} - \eta \nabla \widetilde{F}(z^{(t)})) \leq \widetilde{F}(z^{(t)}) - (2\eta - C_L \eta^2) \|\widetilde{F}(z^{(t)})\|^2 \leq \widetilde{F}(z^{(t)}) - \eta \|\widetilde{F}(z^{(t)})\|^2.$$

$\square$

5.3 *Proof of Theorem* 3.3

Combining all the considerations above, we now prove Theorem 3.3 to conclude this section.

Proof. The proof consists of three parts:
Part I: proof of $z^{(0)} := (u^{(0)}, v^{(0)}) \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$. From the assumption of Theorem 3.3,

$$z^{(0)} \in \frac{1}{\sqrt{3}} \mathcal{N}_d \bigcap \frac{1}{\sqrt{3}} \mathcal{N}_\mu \cap \mathcal{N}_{\frac{2\varepsilon}{5\sqrt{s\kappa}}}.$$

First we show $G(u^{(0)}, v^{(0)}) = 0$: for $0 \leq i \leq s$ and the definition of $\mathcal{N}_d$ and $\mathcal{N}_\mu$,

$$\frac{\|u_i^{(0)}\|^2}{2d_i} \leq \frac{2d_{i0}}{3d_i} < 1, \quad \frac{L|b_l^* u_i^{(0)}|^2}{8d_i \mu^2} \leq \frac{L}{8d_i \mu^2} \cdot \frac{16d_{i0}\mu^2}{3L} \leq \frac{2d_{i0}}{3d_i} < 1,$$

where $\|u_i^{(0)}\| \leq \frac{2\sqrt{d_{i0}}}{\sqrt{3}}$, $\sqrt{L}\|Bu_i^{(0)}\|_\infty \leq \frac{4\sqrt{d_{i0}}\mu}{\sqrt{3}}$ and $\frac{9}{10}d_{i0} \leq d_i \leq \frac{11}{10}d_{i0}$. Therefore

$$G_0\left(\frac{\|u_i^{(0)}\|^2}{2d_i}\right) = G_0\left(\frac{\|v_i^{(0)}\|^2}{2d_i}\right) = G_0\left(\frac{L|b_l^* u_i^{(0)}|^2}{8d_i \mu^2}\right) = 0$$

for all $1 \leq l \leq L$ and $G(u^{(0)}, v^{(0)}) = 0$.

For $z^{(0)} = (u^{(0)}, v^{(0)}) \in \mathcal{N}_{\frac{2\varepsilon}{5\sqrt{s\kappa}}}$, we have $\delta(z^{(0)}) := \frac{\sqrt{\sum_{i=1}^s \delta_i^2 d_{i0}^2}}{d_0} \leq \frac{2\varepsilon}{5\sqrt{s\kappa}}$. By (5.5), there holds $\delta(z^{(0)}) \leq \frac{2\varepsilon}{5\sqrt{s\kappa}}$ and $G(u^{(0)}, v^{(0)}) = 0$,

$$\widetilde{F}(u^{(0)}, v^{(0)}) = F(u^{(0)}, v^{(0)}) \leq \|e\|^2 + \frac{3}{2}\delta^2(z^{(0)})d_0^2 + \frac{\varepsilon\delta(z^{(0)})d_0^2}{5\sqrt{s\kappa}} \leq \|e\|^2 + \frac{\varepsilon^2 d_0^2}{3s\kappa^2}$$

and hence $z^{(0)} = (u^{(0)}, v^{(0)}) \in \mathcal{N}_\epsilon \bigcap \mathcal{N}_{\widetilde{F}}$.

Part II: the linear convergence of the objective function $\widetilde{F}(z^{(t)})$. Denote $z^{(t)} := (u^{(t)}, v^{(t)})$. Note that $z^{(0)} \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$, Lemma 5.9 implies $z^{(t)} \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ for all $t \geq 0$ by induction if $\eta \leq \frac{1}{C_L}$. Moreover, combining Condition 5.1 with Lemma 5.9 leads to

$$\widetilde{F}(z^{(t)}) \leq \widetilde{F}(z^{(t-1)}) - \eta\omega \left[\widetilde{F}(z^{(t-1)}) - c \right]_+, \quad t \geq 1$$

with $c = \|e\|^2 + a\|\mathcal{A}^*(e)\|^2$ and $a = 2000s$. Therefore, by induction, we have

$$\left[\widetilde{F}(z^{(t)}) - c \right]_+ \leq (1 - \eta\omega) \left[\widetilde{F}(z^{(t-1)}) - c \right]_+ \leq (1 - \eta\omega)^t \left[\widetilde{F}(z^{(0)}) - c \right]_+ \leq \frac{\varepsilon^2 d_0^2}{3s\kappa^2} (1 - \eta\omega)^t,$$

where $\widetilde{F}(z^{(0)}) \leq \frac{\varepsilon^2 d_0^2}{3s\kappa^2} + \|e\|^2$ and $\left[\widetilde{F}(z^{(0)}) - c \right]_+ \leq \left[\frac{1}{3s\kappa^2}\varepsilon^2 d_0^2 - a\|\mathcal{A}^*(e)\|^2 \right]_+ \leq \frac{\varepsilon^2 d_0^2}{3s\kappa^2}$. Now we conclude that $\left[\widetilde{F}(z^{(t)}) - c \right]_+$ converges to 0 linearly.

Part III: the linear convergence of the iterates $(u^{(t)}, v^{(t)})$. Denote

$$\delta(z^{(t)}) := \frac{\|\mathcal{H}(u^{(t)}, v^{(t)}) - \mathcal{H}(h_0, x_0)\|_F}{d_0}.$$

Note that $z^{(t)} \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}} \subseteq \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$ and over $\mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$, there holds $F_0(z^{(t)}) \geq \frac{2}{3}\delta^2(z^{(t)})d_0^2$ which follows from Local restricted isometry property (RIP) Condition in (5.3) and $F_0(z^{(t)})$ defined in (2.12). Moreover

$$\widetilde{F}(z^{(t)}) - \|e\|^2 \geq F_0(z^{(t)}) - 2\operatorname{Re}\left(\langle \mathcal{A}^*(e), \mathcal{H}(u^{(0)}, v^{(0)}) - \mathcal{H}(h_0, x_0) \rangle \right)$$

$$\geq \frac{2}{3}\delta^2(z^{(t)})d_0^2 - 2\sqrt{2s}\|\mathcal{A}^*(e)\|\delta(z^{(t)})d_0,$$

where $G(z^{(t)}) \geq 0$ and the second inequality follows from (5.6). There holds

$$\frac{2}{3}\delta^2(z^{(t)})d_0^2 - 2\sqrt{2s}\|\mathcal{A}^*(e)\|\delta(z^{(t)})d_0 - a\|\mathcal{A}^*(e)\|^2 \leq \left[\widetilde{F}(z^{(t)}) - c \right]_+ \leq \frac{\varepsilon^2 d_0^2}{3s\kappa^2}(1 - \eta\omega)^t$$

and equivalently,

$$\left| \delta(z^{(t)})d_0 - \frac{3\sqrt{2}}{2}\|\mathcal{A}^*(e)\| \right|^2 \leq \frac{\varepsilon^2 d_0^2}{2s\kappa^2}(1 - \eta\omega)^t + \left(\frac{3}{2}a + \frac{9}{2} \right) \|\mathcal{A}^*(e)\|^2.$$

Solving the inequality above for $\delta(z^{(t)})$, we have

$$\delta(z^{(t)})d_0 \leq \frac{\varepsilon d_0}{\sqrt{2s\kappa^2}}(1-\eta\omega)^{t/2} + \left(\frac{3\sqrt{2}}{2} + \sqrt{\frac{3}{2}a + \frac{9}{2}}\right)\|\mathcal{A}^*(e)\|$$

$$\leq \frac{\varepsilon d_0}{\sqrt{2s\kappa^2}}(1-\eta\omega)^{t/2} + 60\sqrt{s}\|\mathcal{A}^*(e)\|, \tag{5.9}$$

where $a = 2000s$. Let $d^{(t)} := \sqrt{\sum_{i=1}^{s}\|u_i^{(t)}\|^2\|v_i^{(t)}\|^2}$ for $t \in \mathbb{Z}_{\geq 0}$. By (5.9) and triangle inequality, we immediately obtain $|d^{(t)} - d_0| \leq \frac{\varepsilon d_0}{\sqrt{2s\kappa^2}}(1-\eta\omega)^{t/2} + 60\sqrt{s}\|\mathcal{A}^*(e)\|$. □

6. Proof of the four conditions

This section is devoted to proving the four key conditions introduced in Section 5. The *local smoothness condition* and the *robustness condition* are relatively less challenging to deal with. The more difficult part is to show the *local regularity condition* and the *local isometry property*. The key to solve those problems is to understand how the vector-valued linear operator $\mathcal{A}$ in (2.7) behaves on block-diagonal matrices, such as $\mathcal{H}(h,x)$, $\mathcal{H}(h_0,x_0)$ and $\mathcal{H}(h,x) - \mathcal{H}(h_0,x_0)$. In particular, when $s = 1$, all those matrices become rank-one matrices, which have been well discussed in our previous work [21].

First of all, we define the linear subspace $T_i \subset \mathbb{C}^{K\times N}$ along with its orthogonal complement for $1 \leq i \leq s$ as

$$T_i := \{Z_i \in \mathbb{C}^{K\times N} : Z_i = h_{i0}v_i^* + u_i x_{i0}^*, \quad u_i \in \mathbb{C}^K, v_i \in \mathbb{C}^N\},$$

$$T_i^\perp := \left\{\left(I_K - \frac{h_{i0}h_{i0}^*}{d_{i0}}\right)Z_i\left(I_N - \frac{x_{i0}x_{i0}^*}{d_{i0}}\right) : Z_i \in \mathbb{C}^{K\times N}\right\}, \tag{6.1}$$

where $\|h_{i0}\| = \|x_{i0}\| = \sqrt{d_{i0}}$. In particular, $h_{i0}x_{i0}^* \in T_i$ for all $1 \leq i \leq s$.

The proof also requires us to consider block-diagonal matrices whose ith block belongs to T_i (or $T_i^\perp$). Let $Z = \text{blkdiag}(Z_1, \cdots, Z_s) \in \mathbb{C}^{Ks\times Ns}$ be a block-diagonal matrix and say $Z \in T$ if

$$T := \left\{\text{blkdiag}(\{Z_i\}_{i=1}^s)|Z_i \in T_i\right\}$$

and $Z \in T^\perp$ if

$$T^\perp := \left\{\text{blkdiag}(\{Z_i\}_{i=1}^s)|Z_i \in T_i^\perp\right\},$$

where both T and $T^\perp$ are subsets in $\mathbb{C}^{Ks\times Ns}$ and $\mathcal{H}(h_0,x_0) \in T$.

Now we take a closer look at a special case of block-diagonal matrices, i.e., $\mathcal{H}(h,x)$ and calculate its projection onto T and $T^\perp$, respectively, and it suffices to consider $\mathcal{P}_{T_i}(h_ix_i^*)$ and $\mathcal{P}_{T_i^\perp}(h_ix_i^*)$. For each block $h_ix_i^*$ and $1 \leq i \leq s$, there are unique orthogonal decompositions

$$h_i := \alpha_{i1}h_{i0} + \tilde{h}_i, \quad x := \alpha_{i2}x_{i0} + \tilde{x}_i, \tag{6.2}$$

where $\boldsymbol{h}_{i0} \perp \tilde{\boldsymbol{h}}_i$ and $\boldsymbol{x}_{i0} \perp \tilde{\boldsymbol{x}}_i$. It is important to note that $\alpha_{i1} = \alpha_{i1}(\boldsymbol{h}_i) = \frac{\langle \boldsymbol{h}_{i0}, \boldsymbol{h}_i \rangle}{d_{i0}}$ and $\alpha_{i2} = \alpha_{i2}(\boldsymbol{x}_i) = \frac{\langle \boldsymbol{x}_{i0}, \boldsymbol{x}_i \rangle}{d_{i0}}$, and thus α_{i1} and α_{i2} are functions of $\boldsymbol{h}_i$ and $\boldsymbol{x}_i$, respectively. Immediately, we have the following matrix orthogonal decomposition for $\boldsymbol{h}_i \boldsymbol{x}_i^*$ onto T_i and $T_i^\perp$,

$$\boldsymbol{h}_i \boldsymbol{x}_i^* - \boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^* = \underbrace{(\alpha_{i1} \overline{\alpha_{i2}} - 1) \boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^* + \overline{\alpha_{i2}} \tilde{\boldsymbol{h}}_i \boldsymbol{x}_{i0}^* + \alpha_{i1} \boldsymbol{h}_{i0} \tilde{\boldsymbol{x}}_i^*}_{\text{belong to } T_i} + \underbrace{\tilde{\boldsymbol{h}}_i \tilde{\boldsymbol{x}}_i^*}_{\text{belongs to } T_i^\perp}, \tag{6.3}$$

where the first three components are in T_i while $\tilde{\boldsymbol{h}}_i \tilde{\boldsymbol{x}}_i^* \in T_i^\perp$.

6.1 *Key lemmata*

From the decomposition in (6.2) and (6.3), we want to analyze how $\|\tilde{\boldsymbol{h}}_i\|$, $\|\tilde{\boldsymbol{x}}_i\|$, α_{i1} and α_{i2} depend on $\delta_i = \frac{\|\boldsymbol{h}_i \boldsymbol{x}_i^* - \boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^*\|_F}{d_{i0}}$ if $\delta_i < 1$. The following lemma answers this question, which can be viewed as an application of singular value/vector perturbation theory [40] applied to rank-one matrices. From the lemma below, we can see that if $\boldsymbol{h}_i \boldsymbol{x}_i^*$ is close to $\boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^*$, then $\mathcal{P}_{T_i^\perp}(\boldsymbol{h}_i \boldsymbol{x}_i^*)$ is in fact very small (of order $\mathcal{O}(\delta_i^2 d_{i0})$).

LEMMA 6.1 **(Lemma 5.9 in [21])** Recall that $\|\boldsymbol{h}_{i0}\| = \|\boldsymbol{x}_{i0}\| = \sqrt{d_{i0}}$. If $\delta_i := \frac{\|\boldsymbol{h}_i \boldsymbol{x}_i^* - \boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^*\|_F}{d_{i0}} < 1$, we have the following useful bounds

$$|\alpha_{i1}| \leq \frac{\|\boldsymbol{h}_i\|}{\|\boldsymbol{h}_{i0}\|}, \quad |\alpha_{i1} \overline{\alpha_{i2}} - 1| \leq \delta_i,$$

and

$$\|\tilde{\boldsymbol{h}}_i\| \leq \frac{\delta_i}{1 - \delta_i} \|\boldsymbol{h}_i\|, \quad \|\tilde{\boldsymbol{x}}_i\| \leq \frac{\delta_i}{1 - \delta_i} \|\boldsymbol{x}_i\|, \quad \|\tilde{\boldsymbol{h}}_i\| \|\tilde{\boldsymbol{x}}_i\| \leq \frac{\delta_i^2}{2(1 - \delta_i)} d_{i0}.$$

Moreover, if $\|\boldsymbol{h}_i\| \leq 2\sqrt{d_{i0}}$ and $\sqrt{L}\|\boldsymbol{B}\boldsymbol{h}_i\|_\infty \leq 4\mu\sqrt{d_{i0}}$, i.e., $\boldsymbol{h}_i \in \mathcal{N}_d \bigcap \mathcal{N}_\mu$, we have $\sqrt{L}\|\boldsymbol{B}_i \tilde{\boldsymbol{h}}_i\|_\infty \leq 6\mu\sqrt{d_{i0}}$.

Now we start to focus on several results related to the linear operator $\mathcal{A}$.

LEMMA 6.2 **(Operator norm of $\mathcal{A}$).** For $\mathcal{A}$ defined in (2.7), there holds

$$\|\mathcal{A}\| \leq \sqrt{s(N \log(NL/2) + (\gamma + \log s) \log L)} \tag{6.4}$$

with probability at least $1 - L^{-\gamma}$.

Proof. Note that $\mathcal{A}_i(\boldsymbol{Z}_i) := \{\boldsymbol{b}_l^* \boldsymbol{Z}_i \boldsymbol{a}_{il}\}_{l=1}^L$ in (2.2). Lemma 1 in [1] implies

$$\|\mathcal{A}_i\| \leq \sqrt{N \log(NL/2) + \gamma' \log L}$$

with probability at least $1 - L^{-\gamma'}$. By taking the union bound over $1 \leq i \leq s$,

$$\max \|\mathcal{A}_i\| \leq \sqrt{N \log(NL/2) + (\gamma + \log s) \log L}$$

with probability at least $1 - sL^{-\gamma - \log s} \geq 1 - L^{-\gamma}$.

For $\mathcal{A}$ defined in (2.7), applying the triangle inequality gives

$$\|\mathcal{A}(\mathbf{Z})\| = \left\| \sum_{i=1}^s \mathcal{A}_i(\mathbf{Z}_i) \right\| \leq \sum_{i=1}^s \|\mathcal{A}_i\| \|\mathbf{Z}_i\|_F \leq \max_{1 \leq i \leq s} \|\mathcal{A}_i\| \sqrt{s \sum_{i=1}^s \|\mathbf{Z}_i\|_F^2} = \sqrt{s} \max_{1 \leq i \leq s} \|\mathcal{A}_i\| \|\mathbf{Z}\|_F,$$

where $\mathbf{Z} = \text{blkdiag}(\mathbf{Z}_1, \cdots, \mathbf{Z}_s) \in \mathbb{C}^{Ks \times Ns}$. Therefore,

$$\|\mathcal{A}\| \leq \sqrt{s} \max_{1 \leq i \leq s} \|\mathcal{A}_i\| \leq \sqrt{s(N \log(NL/2) + (\gamma + \log s) \log L)}$$

with probability at least $1 - L^{-\gamma}$. □

LEMMA 6.3 (**Restricted isometry property for $\mathcal{A}$ on T**). The linear operator $\mathcal{A}$ restricted on T is well-conditioned, i.e.,

$$\|\mathcal{P}_T \mathcal{A}^* \mathcal{A} \mathcal{P}_T - \mathcal{P}_T\| \leq \frac{1}{10}, \tag{6.5}$$

where $\mathcal{P}_T$ is the projection operator from $\mathbb{C}^{Ks \times Ns}$ onto T, given $L \geq C_\gamma s^2 \max\{K, \mu_h^2 N\} \log^2 L$ with probability at least $1 - L^{-\gamma}$.

REMARK 6.4 Here $\mathcal{A}\mathcal{P}_T$ and $\mathcal{P}_T \mathcal{A}^*$ are defined as

$$\mathcal{A}\mathcal{P}_T(\mathbf{Z}) = \sum_{i=1}^s \mathcal{A}_i(\mathcal{P}_{T_i}(\mathbf{Z}_i)), \quad \mathcal{P}_T \mathcal{A}^*(z) = \text{blkdiag}(\mathcal{P}_{T_1}(\mathcal{A}_1^*(z)), \cdots, \mathcal{P}_{T_s}(\mathcal{A}_s^*(z))),$$

respectively, where $\mathbf{Z}$ is a block-diagonal matrix and $z \in \mathbb{C}^L$.

As shown in the remark above, the proof of Lemma 6.3 depends on the properties of both $\mathcal{P}_{T_i} \mathcal{A}_i^* \mathcal{A}_i \mathcal{P}_{T_i}$ and $\mathcal{P}_{T_i} \mathcal{A}_i^* \mathcal{A}_j \mathcal{P}_{T_j}$ for $i \neq j$. Fortunately, we have already proven related results in [23] which are written as follows:

LEMMA 6.5 (**Inter-user incoherence, Corollary 5.3 and 5.8 in [23]**). There hold

$$\|\mathcal{P}_{T_i} \mathcal{A}_i^* \mathcal{A}_j \mathcal{P}_{T_j}\| \leq \frac{1}{10s}, \quad \forall i \neq j; \qquad \|\mathcal{P}_{T_i} \mathcal{A}_i^* \mathcal{A}_i \mathcal{P}_{T_i} - \mathcal{P}_{T_i}\| \leq \frac{1}{10s}, \quad \forall 1 \leq i \leq s \tag{6.6}$$

with probability at least $1 - L^{-\gamma+1}$ if $L \geq C_\gamma s^2 \max\{K, \mu_h^2 N\} \log^2 L \log(s+1)$.

Note that $\|\mathcal{P}_{T_i} \mathcal{A}_i^* \mathcal{A}_j \mathcal{P}_{T_j}\| \leq \frac{1}{10s}$ holds because of independence between each individual random Gaussian matrix $\mathbf{A}_i$. In particular, if $s = 1$, the inter-user incoherence $\|\mathcal{P}_{T_i} \mathcal{A}_i^* \mathcal{A}_j \mathcal{P}_{T_j}\| \leq \frac{1}{10s}$ is not needed at all. With (6.6), it is easy to prove Lemma 6.3.

PROOF OF LEMMA 6.3   For any block diagonal matrix $\mathbf{Z} = \mathrm{blkdiag}(\mathbf{Z}_1, \cdots, \mathbf{Z}_s) \in \mathbb{C}^{Ks \times Ns}$ and $\mathbf{Z}_i \in \mathbb{C}^{K \times N}$,

$$
\begin{aligned}
\langle \mathbf{Z}, \mathcal{P}_T \mathcal{A}^* \mathcal{A} \mathcal{P}_T(\mathbf{Z}) - \mathcal{P}_T(\mathbf{Z}) \rangle &= \sum_{1 \le i,j \le s} \langle \mathcal{A}_i \mathcal{P}_{T_i}(\mathbf{Z}_i), \mathcal{A}_j \mathcal{P}_{T_j}(\mathbf{Z}_j) \rangle - \| \mathcal{P}_T(\mathbf{Z}) \|_F^2 \\
&= \sum_{i=1}^s \langle \mathbf{Z}_i, \mathcal{P}_{T_i} \mathcal{A}_i^* \mathcal{A}_i \mathcal{P}_{T_i}(\mathbf{Z}_i) - \mathcal{P}_{T_i}(\mathbf{Z}_i) \rangle + \sum_{i \ne j} \langle \mathcal{A}_i \mathcal{P}_{T_i}(\mathbf{Z}_i), \mathcal{A}_j \mathcal{P}_{T_j}(\mathbf{Z}_j) \rangle.
\end{aligned}
$$
(6.7)

Using (6.6), the following two inequalities hold,

$$
|\langle \mathbf{Z}_i, \mathcal{P}_{T_i} \mathcal{A}_i^* \mathcal{A}_i \mathcal{P}_{T_i}(\mathbf{Z}_i) - \mathcal{P}_{T_i}(\mathbf{Z}_i) \rangle| \le \| \mathcal{P}_{T_i} \mathcal{A}_i^* \mathcal{A}_i \mathcal{P}_{T_i} - \mathcal{P}_{T_i} \| \| \mathbf{Z}_i \|_F^2 \le \frac{\| \mathbf{Z}_i \|_F^2}{10s},
$$

$$
|\langle \mathcal{A}_i \mathcal{P}_{T_i}(\mathbf{Z}_i), \mathcal{A}_j \mathcal{P}_{T_j}(\mathbf{Z}_j) \rangle| \le \| \mathcal{P}_{T_i} \mathcal{A}_i^* \mathcal{A}_j \mathcal{P}_{T_j} \| \| \mathbf{Z}_i \|_F \| \mathbf{Z}_j \|_F \le \frac{\| \mathbf{Z}_i \|_F \| \mathbf{Z}_j \|_F}{10s}.
$$

After substituting both estimates into (6.7), we have

$$
|\langle \mathbf{Z}, \mathcal{P}_T \mathcal{A}^* \mathcal{A} \mathcal{P}_T(\mathbf{Z}) - \mathcal{P}_T(\mathbf{Z}) \rangle| \le \sum_{1 \le i,j \le s} \frac{\| \mathbf{Z}_i \|_F \| \mathbf{Z}_j \|_F}{10s} \le \frac{1}{10s} \left(\sum_{i=1}^s \| \mathbf{Z}_i \|_F \right)^2 \le \frac{\| \mathbf{Z} \|_F^2}{10}.
$$

$\square$

Finally, we show how $\mathcal{A}$ behaves when applied to block-diagonal matrices $\mathbf{X} = \mathcal{H}(\boldsymbol{h}, \boldsymbol{x})$. In particular, the calculations will be much simplified for the case $s = 1$.

LEMMA 6.6   ($\mathcal{A}$ **restricted on block-diagonal matrices with rank-one blocks**).
Consider $\mathbf{X} = \mathcal{H}(\boldsymbol{h}, \boldsymbol{x})$ and

$$
\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x}) := \max_{1 \le l \le L} \sum_{i=1}^s |\boldsymbol{b}_l^* \boldsymbol{h}_i|^2 \| \boldsymbol{x}_i \|^2.
$$
(6.8)

Conditioned on (6.4), we have

$$
\| \mathcal{A}(\mathbf{X}) \|^2 \le \frac{4}{3} \| \mathbf{X} \|_F^2 + 2\sqrt{2s \| \mathbf{X} \|_F^2 \sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})(K + N) \log L} + 8s\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})(K + N) \log L,
$$
(6.9)

uniformly for any $\boldsymbol{h} \in \mathbb{C}^{Ks}$ and $\boldsymbol{x} \in \mathbb{C}^{Ns}$ with probability at least $1 - \frac{1}{\gamma} \exp(-s(K + N))$ if $L \ge C_\gamma s(K + N) \log L$. Here $\| \mathbf{X} \|_F^2 = \| \mathcal{H}(\boldsymbol{h}, \boldsymbol{x}) \|_F^2 = \sum_{i=1}^s \| \boldsymbol{h}_i \|^2 \| \boldsymbol{x}_i \|^2$.

REMARK 6.7 Here are a few more explanations and facts about $\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})$. Note that $\|\mathcal{A}(\boldsymbol{X})\|^2$ is the sum of L sub-exponential[6] random variables, i.e.,

$$\|\mathcal{A}(\boldsymbol{X})\|^2 = \sum_{l=1}^{L} \left| \sum_{i=1}^{s} \boldsymbol{b}_l^* \boldsymbol{h}_i \boldsymbol{x}_i^* \boldsymbol{a}_{il} \right|^2. \tag{6.10}$$

Here $\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})$ corresponds to the largest expectation of all those components in $\|\mathcal{A}(\boldsymbol{X})\|^2$.

For $\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})$, without loss of generality, we assume $\|\boldsymbol{x}_i\| = 1$ for $1 \le i \le s$ and let $\boldsymbol{h} \in \mathbb{C}^{Ks}$ be a unit vector, i.e., $\|\boldsymbol{h}\|^2 = \sum_{i=1}^{s} \|\boldsymbol{h}_i\|^2 = 1$. The bound

$$\frac{1}{L} \le \sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x}) \le \frac{K}{L} \tag{6.11}$$

follows from $L\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x}) \ge \sum_{l=1}^{L} \sum_{i=1}^{s} |\boldsymbol{b}_l^* \boldsymbol{h}_i|^2 = \|\boldsymbol{h}\|^2 = 1$.

Moreover, $\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})$ and $\sigma_{\max}(\boldsymbol{h}, \boldsymbol{x})$ are both Lipschitz functions w.r.t. $\boldsymbol{h}$. Now we want to determine their Lipschitz constants. First note that for $\|\boldsymbol{x}_i\| = 1$, $\sigma_{\max}(\boldsymbol{h}, \boldsymbol{x})$ equals

$$\sigma_{\max}(\boldsymbol{h}, \boldsymbol{x}) = \max_{1 \le l \le L} \|(\boldsymbol{I}_s \otimes \boldsymbol{b}_l^*)\boldsymbol{h}\|,$$

where $\otimes$ denotes Kronecker product. Let $\boldsymbol{u} \in \mathbb{C}^{Ks}$ be another unit vector and we have

$$|\sigma_{\max}(\boldsymbol{h}, \boldsymbol{x}) - \sigma_{\max}(\boldsymbol{u}, \boldsymbol{x})| = \left| \max_{1 \le l \le L} \|(\boldsymbol{I}_s \otimes \boldsymbol{b}_l^*)\boldsymbol{h}\| - \max_{1 \le l \le L} \|(\boldsymbol{I}_s \otimes \boldsymbol{b}_l^*)\boldsymbol{u}\| \right|$$

$$= \max_{1 \le l \le L} \left| \|(\boldsymbol{I}_s \otimes \boldsymbol{b}_l^*)\boldsymbol{h}\| - \|(\boldsymbol{I}_s \otimes \boldsymbol{b}_l^*)\boldsymbol{u}\| \right|$$

$$\le \max_{1 \le l \le L} \|(\boldsymbol{I}_s \otimes \boldsymbol{b}_l^*)(\boldsymbol{h} - \boldsymbol{u})\| \le \|\boldsymbol{h} - \boldsymbol{u}\|, \tag{6.12}$$

where $\|\boldsymbol{I}_s \otimes \boldsymbol{b}_l^*\| = \|\boldsymbol{b}_l\|\sqrt{\frac{K}{L}} < 1$. For $\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})$,

$$|\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x}) - \sigma_{\max}^2(\boldsymbol{u}, \boldsymbol{x})| \le (\sigma_{\max}(\boldsymbol{h}, \boldsymbol{x}) + \sigma_{\max}(\boldsymbol{u}, \boldsymbol{x})) \cdot |\sigma_{\max}(\boldsymbol{h}, \boldsymbol{x}) - \sigma_{\max}(\boldsymbol{u}, \boldsymbol{x})|$$

$$\le \frac{2K}{L}\|\boldsymbol{h} - \boldsymbol{u}\| \le 2\|\boldsymbol{h} - \boldsymbol{u}\|. \tag{6.13}$$

PROOF OF LEMMA 6.6 Without loss of generality, let $\|\boldsymbol{x}_i\| = 1$ and $\sum_{i=1}^{s} \|\boldsymbol{h}_i\|^2 = 1$. It suffices to prove $f(\boldsymbol{h}, \boldsymbol{x}) \le \frac{4}{3}$ for all $(\boldsymbol{h}, \boldsymbol{x}) \in \mathbb{C}^{Ks} \times \mathbb{C}^{Ns}$ in (2.5) where $f(\boldsymbol{h}, \boldsymbol{x})$ is defined as

$$f(\boldsymbol{h}, \boldsymbol{x}) := \|\mathcal{A}(\boldsymbol{X})\|^2 - 2\sqrt{2s\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})(K + N)\log L} - 8s\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})(K + N)\log L.$$

---

[6] For the definition and properties of sub-exponential random variables, the readers can find all relevant information in [38].

Part I: bounds of $\|\mathcal{A}(X)\|^2$ for any fixed $(\boldsymbol{h}, \boldsymbol{x})$. From (47), we already know that $Y = \|\mathcal{A}(X)\|_F^2 = \sum_{i=1}^{2L} c_i \xi_i^2$ where $\{\xi_i\}$ are i.i.d. χ_1^2 random variables and $\boldsymbol{c} = (c_1, \cdots, c_{2L})^T \in \mathbb{R}^{2L}$. More precisely, we can determine $\{c_i\}_{i=1}^{2L}$ as

$$\left| \sum_{i=1}^{s} \boldsymbol{b}_l^* \boldsymbol{h}_i \boldsymbol{x}^* \boldsymbol{a}_{il} \right|^2 = c_{2l-1} \xi_{2l-1}^2 + c_{2l} \xi_{2l}^2, \quad c_{2l-1} = c_{2l} = \frac{1}{2} \sum_{i=1}^{s} |\boldsymbol{b}_l^* \boldsymbol{h}_i|^2$$

because $\sum_{i=1}^{s} \boldsymbol{b}_l^* \boldsymbol{h}_i \boldsymbol{x}_i^* \boldsymbol{a}_{il} \sim \mathcal{CN}\left(0, \sum_{i=1}^{s} |\boldsymbol{b}_l^* \boldsymbol{h}_i|^2\right)$.

By the Bernstein inequality, there holds

$$\mathbb{P}(Y - \mathbb{E}(Y) \geq t) \leq \exp\left(-\frac{t^2}{8\|\boldsymbol{c}\|^2}\right) \vee \exp\left(-\frac{t}{8\|\boldsymbol{c}\|_\infty}\right), \tag{6.14}$$

where $\mathbb{E}(Y) = \|X\|_F^2 = 1$. In order to apply the Bernstein inequality, we need to estimate $\|\boldsymbol{c}\|^2$ and $\|\boldsymbol{c}\|_\infty$ as follows,

$$\|\boldsymbol{c}\|_\infty = \frac{1}{2} \max_{1 \leq l \leq L} \sum_{i=1}^{s} |\boldsymbol{b}_l^* \boldsymbol{h}_i|^2 = \frac{1}{2} \sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x}),$$

$$\|\boldsymbol{c}\|_2^2 = \frac{1}{2} \sum_{l=1}^{L} \left| \sum_{i=1}^{s} |\boldsymbol{b}_l^* \boldsymbol{h}_i|^2 \right|^2 \leq \frac{1}{2} \left(\sum_{i=1}^{s} \sum_{l=1}^{L} |\boldsymbol{b}_l^* \boldsymbol{h}_i|^2 \right) \max_{1 \leq l \leq L} \sum_{i=1}^{s} |\boldsymbol{b}_l^* \boldsymbol{h}_i|^2 \leq \frac{1}{2} \sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x}).$$

Applying (6.14) gives

$$\mathbb{P}(\|\mathcal{A}(X)\|^2 \geq 1 + t) \leq \exp\left(-\frac{t^2}{4\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})}\right) \vee \exp\left(-\frac{t}{4\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})}\right).$$

In particular, by setting

$$t = g(\boldsymbol{h}, \boldsymbol{x}) := 2\sqrt{2s\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})(K+N)\log L} + 8s\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x})(K+N)\log L,$$

we have

$$\mathbb{P}\left(\|\mathcal{A}(X)\|^2 \geq 1 + g(\boldsymbol{h}, \boldsymbol{x})\right) \leq e^{-2s(K+N)(\log L)}.$$

So far, we have shown that $f(\boldsymbol{h}, \boldsymbol{x}) \leq 1$ with probability at least $1 - e^{-2s(K+N)(\log L)}$ for a fixed pair of $(\boldsymbol{h}, \boldsymbol{x})$.

Part II: covering argument. Now we will use a covering argument to extend this result for all $(\boldsymbol{h}, \boldsymbol{x})$ and thus prove that $f(\boldsymbol{h}, \boldsymbol{x}) \leq \frac{4}{3}$ uniformly for all $(\boldsymbol{h}, \boldsymbol{x})$.

We start with defining $\mathcal{K}$ and $\mathcal{N}_i$ as ϵ_0-nets of $\mathcal{S}^{Ks-1}$ and $\mathcal{S}^{N-1}$ for $\boldsymbol{h}$ and $\boldsymbol{x}_i$, $1 \leq i \leq s$, respectively. The bounds $|\mathcal{K}| \leq \left(1 + \frac{2}{\epsilon_0}\right)^{2sK}$ and $|\mathcal{N}_i| \leq \left(1 + \frac{2}{\epsilon_0}\right)^{2N}$ follow from the covering numbers of the sphere

(Lemma 5.2 in [38]). Here we let $\mathcal{N} := \mathcal{N}_1 \times \cdots \times \mathcal{N}_s$. By taking the union bound over $\mathcal{K} \times \mathcal{N}$, we have that $f(\boldsymbol{h}, \boldsymbol{x}) \leq 1$ holds uniformly for all $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{K} \times \mathcal{N}$ with probability at least

$$1 - (1 + 2/\epsilon_0)^{2s(K+N)} \, \mathrm{e}^{-2s(K+N)\log L} = 1 - \mathrm{e}^{-2s(K+N)(\log L - \log(1 + 2/\epsilon_0))}.$$

For any $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{S}^{Ks-1} \times \underbrace{\mathcal{S}^{N-1} \times \cdots \times \mathcal{S}^{N-1}}_{s \text{ times}}$, we can find a point $(\boldsymbol{u}, \boldsymbol{v}) \in \mathcal{K} \times \mathcal{N}$ satisfying $\|\boldsymbol{h} - \boldsymbol{u}\| \leq \varepsilon_0$ and $\|\boldsymbol{x}_i - \boldsymbol{v}_i\| \leq \varepsilon_0$ for all $1 \leq i \leq s$. Conditioned on (6.4), we know that

$$\|\mathcal{A}\|^2 \leq s(N \log(NL/2) + (\gamma + \log s)\log L) \leq s(N + \gamma + \log s)\log L.$$

Now we aim to evaluate $|f(\boldsymbol{h}, \boldsymbol{x}) - f(\boldsymbol{u}, \boldsymbol{v})|$. First we consider $|f(\boldsymbol{u}, \boldsymbol{x}) - f(\boldsymbol{u}, \boldsymbol{v})|$. Since $\sigma_{\max}^2(\boldsymbol{u}, \boldsymbol{x}) = \sigma_{\max}^2(\boldsymbol{u}, \boldsymbol{v})$ if $\|\boldsymbol{x}_i\| = \|\boldsymbol{v}_i\| = \|\boldsymbol{u}\| = 1$ for $1 \leq i \leq s$, we have

$$
\begin{aligned}
|f(\boldsymbol{u}, \boldsymbol{x}) - f(\boldsymbol{u}, \boldsymbol{v})| &= \left| \|\mathcal{A}(\mathcal{H}(\boldsymbol{u}, \boldsymbol{x}))\|_F^2 - \|\mathcal{A}(\mathcal{H}(\boldsymbol{u}, \boldsymbol{v}))\|_F^2 \right| \\
&\leq \|\mathcal{A}(\mathcal{H}(\boldsymbol{u}, \boldsymbol{x} - \boldsymbol{v}))\| \cdot \|\mathcal{A}(\mathcal{H}(\boldsymbol{u}, \boldsymbol{x} + \boldsymbol{v}))\| \\
&\leq \|\mathcal{A}\|^2 \sqrt{\sum_{i=1}^{s} \|\boldsymbol{u}_i\|^2 \|\boldsymbol{x}_i - \boldsymbol{v}_i\|^2} \sqrt{\sum_{i=1}^{s} \|\boldsymbol{u}_i\|^2 \|\boldsymbol{x}_i + \boldsymbol{v}_i\|^2} \\
&\leq 2\|\mathcal{A}\|^2 \varepsilon_0 \leq 2s(N + \gamma + \log s)(\log L)\varepsilon_0,
\end{aligned}
$$

where the first inequality is due to $||z_1|^2 - |z_2|^2| \leq |z_1 - z_2||z_1 + z_2|$ for any $z_1, z_2 \in \mathbb{C}$.

We proceed to estimate $|f(\boldsymbol{h}, \boldsymbol{x}) - f(\boldsymbol{u}, \boldsymbol{x})|$ by using (50) and (49),

$$
\begin{aligned}
|f(\boldsymbol{h}, \boldsymbol{x}) - f(\boldsymbol{u}, \boldsymbol{x})| &\leq J_1 + J_2 + J_3 \\
&\leq (2\|\mathcal{A}\|^2 + 2\sqrt{2s(K+N)\log L} + 16s(K+N)\log L)\varepsilon_0 \\
&\leq 25s(K + N + \gamma + \log s)(\log L)\varepsilon_0,
\end{aligned}
$$

where (6.13) and (6.12) give

$$J_1 = \left| \|\mathcal{A}(\mathcal{H}(\boldsymbol{h}, \boldsymbol{x}))\|_F^2 - \|\mathcal{A}(\mathcal{H}(\boldsymbol{u}, \boldsymbol{x}))\|_F^2 \right| \leq \|\mathcal{A}(\mathcal{H}(\boldsymbol{h} - \boldsymbol{u}, \boldsymbol{x}))\| \, \|\mathcal{A}(\mathcal{H}(\boldsymbol{h} + \boldsymbol{u}, \boldsymbol{x}))\| \leq 2\|\mathcal{A}\|^2 \varepsilon_0,$$

$$J_2 = 2\sqrt{2s(K+N)\log L} \cdot |\sigma_{\max}(\boldsymbol{h}, \boldsymbol{x}) - \sigma_{\max}(\boldsymbol{u}, \boldsymbol{x})| \leq 2\sqrt{2s(K+N)\log L}\varepsilon_0,$$

$$J_3 = 8s(K+N)(\log L) \cdot |\sigma_{\max}^2(\boldsymbol{h}, \boldsymbol{x}) - \sigma_{\max}^2(\boldsymbol{u}, \boldsymbol{x})| \leq 16s(K+N)(\log L)\varepsilon_0.$$

Therefore, if $\epsilon_0 = \frac{1}{81s(N+K+\gamma+\log s)\log L}$, there holds

$$f(\boldsymbol{h}, \boldsymbol{x}) \leq f(\boldsymbol{u}, \boldsymbol{v}) + \underbrace{|f(\boldsymbol{u}, \boldsymbol{x}) - f(\boldsymbol{u}, \boldsymbol{v})| + |f(\boldsymbol{h}, \boldsymbol{x}) - f(\boldsymbol{u}, \boldsymbol{x})|}_{\leq 27s(K+N+\gamma+\log s)(\log L)\varepsilon_0 \leq \frac{1}{3}} \leq \frac{4}{3}$$

for all $(\boldsymbol{h}, \boldsymbol{x})$ uniformly with probability at least $1 - \mathrm{e}^{-2s(K+N)(\log L - \log(1+2/\varepsilon_0))}$. By letting $L \geq C_\gamma s(K + N) \log L$ with C_γ reasonably large and $\gamma \geq 1$, we have $\log L - \log(1 + 2/\varepsilon_0) \geq \frac{1}{2}(1 + \log(\gamma))$ and with probability at least $1 - \frac{1}{\gamma} \exp(-s(K + N))$. $\qquad\square$

6.2 *Proof of the local restricted isometry property*

LEMMA 6.8 Conditioned on (6.5) and (6.9), the following RIP type of property holds:

$$\frac{2}{3} \|X - X_0\|_F^2 \leq \|\mathcal{A}(X - X_0)\|^2 \leq \frac{3}{2} \|X - X_0\|_F^2$$

uniformly for all $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$ with $\mu \geq \mu_h$ and $\epsilon \leq \frac{1}{15}$ if $L \geq C_\gamma \mu^2 s(K + N) \log^2 L$ for some numerical constant C_γ.

Proof. The main idea of the proof follows two steps: decompose $X - X_0$ onto T and $T^\perp$, then apply (6.5) and (6.9) to $\mathcal{P}_T(X - X_0)$ and $\mathcal{P}_{T^\perp}(X - X_0)$, respectively.

For any $X = \mathcal{H}(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_\epsilon$ with $\delta_i \leq \varepsilon \leq \frac{1}{15}$, we can decompose $X - X_0$ as the sum of two block diagonal matrices $U = \mathrm{blkdiag}(U_i, 1 \leq i \leq s)$ and $V = \mathrm{blkdiag}(V_i, 1 \leq i \leq s)$ where each pair of (U_i, V_i) corresponds to the orthogonal decomposition of $\boldsymbol{h}_i \boldsymbol{x}_i^* - \boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^*$,

$$\boldsymbol{h}_i \boldsymbol{x}_i^* - \boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^* := \underbrace{(\alpha_{i1} \overline{\alpha_{i2}} - 1) \boldsymbol{h}_{i0} \boldsymbol{x}_{i0}^* + \overline{\alpha_{i2}} \tilde{\boldsymbol{h}}_i \boldsymbol{x}_{i0}^* + \alpha_{i1} \boldsymbol{h}_{i0} \tilde{\boldsymbol{x}}_i^*}_{U_i \in T_i} + \underbrace{\tilde{\boldsymbol{h}}_i \tilde{\boldsymbol{x}}_i^*}_{V_i \in T_i^\perp} \qquad (6.15)$$

which has been briefly discussed in (6.2) and (6.3). Note that $\mathcal{A}(X - X_0) = \mathcal{A}(U + V)$ and

$$\|\mathcal{A}(U)\| - \|\mathcal{A}(V)\| \leq \|\mathcal{A}(U + V)\| \leq \|\mathcal{A}(U)\| + \|\mathcal{A}(V)\|.$$

Therefore, it suffices to have a two-side bound for $\|\mathcal{A}(U)\|$ and an upper bound for $\|\mathcal{A}(V)\|$ where $U \in T$ and $V \in T^\perp$ in order to establish the local isometry property.
Estimation of $\|\mathcal{A}(U)\|$: For $\|\mathcal{A}(U)\|$, we know from Lemma 6.3 that

$$\sqrt{\frac{9}{10}} \|U\|_F \leq \|\mathcal{A}(U)\| \leq \sqrt{\frac{11}{10}} \|U\|_F \qquad (6.16)$$

and hence we only need to compute $\|U\|_F$. By Lemma 6.1, there also hold $\|V_i\|_F \leq \frac{\delta_i^2}{2(1-\delta_i)} d_{i0}$ and $\delta_i - \|V_i\|_F \leq \|U_i\|_F \leq \delta_i + \|V_i\|_F$, i.e.,

$$\left(\delta_i - \frac{\delta_i^2}{2(1 - \delta_i)}\right) d_{i0} \leq \|U_i\|_F \leq \left(\delta_i + \frac{\delta_i^2}{2(1 - \delta_i)}\right) d_{i0}, \quad 1 \leq i \leq s.$$

With $\|U\|_F^2 = \sum_{i=1}^s \|U_i\|_F^2$, it is easy to get $\delta d_0 \left(1 - \frac{\varepsilon}{2(1-\varepsilon)}\right) \leq \|U\|_F \leq \delta d_0 \left(1 + \frac{\varepsilon}{2(1-\varepsilon)}\right)$. Combined with (6.16), we get

$$\sqrt{\frac{9}{10}} \left(1 - \frac{\varepsilon}{2(1 - \varepsilon)}\right) \delta d_0 \leq \|\mathcal{A}(U)\| \leq \sqrt{\frac{11}{10}} \left(1 + \frac{\varepsilon}{2(1 - \varepsilon)}\right) \delta d_0. \qquad (6.17)$$

Estimation of $\|\mathcal{A}(V)\|$: note that V is a block-diagonal matrix with rank-one block. So applying Lemma 6.6 gives us

$$\|\mathcal{A}(V)\|^2 \leq \frac{4}{3}\|V\|_F^2 + 2\sqrt{2s\|V\|_F^2 \sigma_{\max}^2(\tilde{h}, \tilde{x})(K+N)\log L} + 8s\sigma_{\max}^2(\tilde{h}, \tilde{x})(K+N)\log L, \qquad (6.18)$$

where $V = \mathcal{H}(\tilde{h}, \tilde{x})$ and $\tilde{h} = \begin{bmatrix} \tilde{h}_1 \\ \vdots \\ \tilde{h}_s \end{bmatrix}$. It suffices to get an estimation of $\|V\|_F$ and $\sigma_{\max}^2(\tilde{h}, \tilde{x})$ to bound $\|\mathcal{A}(V)\|$ in (6.18).

Lemma 6.1 says that $\|\tilde{h}_i\|\|\tilde{x}_i\| \leq \frac{\delta_i^2}{2(1-\delta_i)} d_{i0} \leq \frac{\varepsilon}{2(1-\varepsilon)} \delta_i d_{i0}$ if $\varepsilon < 1$. Moreover,

$$\|\tilde{x}_i\| \leq \frac{\delta_i}{1-\delta_i}\|x_i\| \leq \frac{2\delta_i}{1-\delta_i}\sqrt{d_{i0}}, \quad \sqrt{L}\|B\tilde{h}_i\|_\infty \leq 6\mu\sqrt{d_{i0}}, \quad 1 \leq i \leq s \qquad (6.19)$$

if (h, x) belongs to $\mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$. For $\|V\|_F$,

$$\|V\|_F = \sqrt{\sum_{i=1}^s \|V_i\|_F^2} = \sqrt{\sum_{i=1}^s \|\tilde{h}_i\|^2 \|\tilde{x}_i\|^2} \leq \frac{\varepsilon\delta d_0}{2(1-\varepsilon)}.$$

Now we aim to get an upper bound for $\sigma_{\max}^2(\tilde{h}, \tilde{x})$ by using (6.19),

$$\sigma_{\max}^2(\tilde{h}, \tilde{x}) = \max_{1 \leq l \leq L} \sum_{i=1}^s |b_l^* \tilde{h}_i|^2 \|\tilde{x}_i\|^2 \leq C_0 \frac{\mu^2 \sum_{i=1}^s \delta_i^2 d_{i0}^2}{L} = C_0 \frac{\mu^2 \delta^2 d_0^2}{L}.$$

By substituting the estimations of $\|V\|_F$ and $\sigma_{\max}^2(\tilde{h}, \tilde{x})$ into (6.18)

$$\|\mathcal{A}(V)\|^2 \leq \frac{\varepsilon^2\delta^2 d_0^2}{3(1-\varepsilon)^2} + \frac{\sqrt{2}\varepsilon\delta^2 d_0^2}{1-\varepsilon}\sqrt{\frac{C_0\mu^2 s(K+N)\log L}{L}} + \frac{8C_0\mu^2\delta^2 d_0^2 s(K+N)\log L}{L}. \qquad (6.20)$$

By letting $L \geq C_\gamma \mu^2 s(K+N)\log^2 L$ with C_γ sufficiently large and combining (6.20) and (6.17), we have

$$\sqrt{\frac{2}{3}}\delta d_0 \leq \|\mathcal{A}(U)\| - \|\mathcal{A}(V)\| \leq \|\mathcal{A}(U+V)\| \leq \|\mathcal{A}(U)\| + \|\mathcal{A}(V)\| \leq \sqrt{\frac{3}{2}}\delta d_0,$$

which gives $\frac{2}{3}\|X - X_0\|_F^2 \leq \|\mathcal{A}(X - X_0)\|^2 \leq \frac{3}{2}\|X - X_0\|_F^2$. $\qquad\square$

6.3 *Proof of the local regularity condition*

We first introduce a few notations: for all $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_d \cap \mathcal{N}_\epsilon$, consider $\alpha_{i1}, \alpha_{i2}, \tilde{\boldsymbol{h}}_i$ and $\tilde{\boldsymbol{x}}_i$ defined in (6.2) and define

$$\Delta \boldsymbol{h}_i = \boldsymbol{h}_i - \alpha_i \boldsymbol{h}_{i0}, \quad \Delta \boldsymbol{x}_i = \boldsymbol{x}_i - \overline{\alpha}_i^{-1} \boldsymbol{x}_{i0},$$

where

$$\alpha_i(\boldsymbol{h}_i, \boldsymbol{x}_i) = \begin{cases} (1 - \delta_0)\alpha_{i1}, & \text{if } \|\boldsymbol{h}_i\|_2 \geq \|\boldsymbol{x}_i\|_2 \\ \frac{1}{(1-\delta_0)\overline{\alpha}_{i2}}, & \text{if } \|\boldsymbol{h}_i\|_2 < \|\boldsymbol{x}_i\|_2 \end{cases}$$

with

$$\delta_0 := \frac{\delta}{10}. \tag{6.21}$$

The function $\alpha_i(\boldsymbol{h}_i, \boldsymbol{x}_i)$ is defined for each block of $\boldsymbol{X} = \mathcal{H}(\boldsymbol{h}, \boldsymbol{x})$. The particular form of $\alpha_i(\boldsymbol{h}, \boldsymbol{x})$ serves primarily for proving the Lemma 6.11, i.e., local regularity condition of $G(\boldsymbol{h}, \boldsymbol{x})$. We also define

$$\Delta \boldsymbol{h} := \begin{bmatrix} \boldsymbol{h}_1 - \alpha_1 \boldsymbol{h}_{1,0} \\ \vdots \\ \boldsymbol{h}_s - \alpha_s \boldsymbol{h}_{s0} \end{bmatrix} \in \mathbb{C}^{Ks}, \quad \Delta \boldsymbol{x} := \begin{bmatrix} \boldsymbol{x}_1 - \alpha_1 \boldsymbol{x}_{1,0} \\ \vdots \\ \boldsymbol{x}_s - \alpha_s \boldsymbol{x}_{s0} \end{bmatrix} \in \mathbb{C}^{Ns}.$$

The following lemma gives bounds of $\Delta \boldsymbol{x}_i$ and $\Delta \boldsymbol{h}_i$.

LEMMA 6.9 For all $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_d \cap \mathcal{N}_\epsilon$ with $\epsilon \leq \frac{1}{15}$, there hold

$$\max\{\|\Delta \boldsymbol{h}_i\|_2^2, \|\Delta \boldsymbol{x}_i\|_2^2\} \leq (7.5\delta_i^2 + 2.88\delta_0^2)d_{i0}, \quad \|\Delta \boldsymbol{h}_i\|_2^2 \|\Delta \boldsymbol{x}_i\|_2^2 \leq \frac{1}{26}(\delta_i^2 + \delta_0^2)d_{i0}^2.$$

Moreover, if we assume $(\boldsymbol{h}_i, \boldsymbol{x}_i) \in \mathcal{N}_\mu$ additionally, we have $\sqrt{L}\|\boldsymbol{B}(\Delta \boldsymbol{h}_i)\|_\infty \leq 6\mu\sqrt{d_{i0}}$.

Proof. We only consider $\|\boldsymbol{h}_i\|_2 \geq \|\boldsymbol{x}_i\|_2$ and $\alpha_i = (1 - \delta_0)\alpha_{1i}$, and the other case is exactly the same due to the symmetry. For both $\Delta \boldsymbol{h}_i$ and $\Delta \boldsymbol{x}_i$, by definition,

$$\Delta \boldsymbol{h}_i = \boldsymbol{h}_i - \alpha_i \boldsymbol{h}_{i0} = \delta_0 \alpha_{i1} \boldsymbol{h}_{i0} + \tilde{\boldsymbol{h}}_i, \tag{6.22}$$

$$\Delta \boldsymbol{x}_i = \boldsymbol{x}_i - \frac{1}{(1-\delta_0)\overline{\alpha}_{i1}} \boldsymbol{x}_{i0} = \left(\alpha_{i2} - \frac{1}{(1-\delta_0)\overline{\alpha}_{i1}}\right) \boldsymbol{x}_{i0} + \tilde{\boldsymbol{x}}_i, \tag{6.23}$$

where $\boldsymbol{h}_i = \alpha_{i1}\boldsymbol{h}_{i0} + \tilde{\boldsymbol{h}}_i$ and $\boldsymbol{x}_i = \alpha_{i2}\boldsymbol{x}_{i0} + \tilde{\boldsymbol{x}}_i$ come from the orthogonal decomposition in (6.2).

We start with estimating $\|\Delta \boldsymbol{h}_i\|^2$. Note that $\|\boldsymbol{h}_i\|_2^2 \leq 4d_{i0}$ and $\|\alpha_{i1}\boldsymbol{h}_{i0}\|_2^2 \leq \|\boldsymbol{h}_i\|_2^2$ since $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_d \cap \mathcal{N}_\mu$. By Lemma 6.1, we have

$$\|\Delta \boldsymbol{h}_i\|_2^2 = \|\tilde{\boldsymbol{h}}_i\|_2^2 + \delta_0^2 \|\alpha_{i1}\boldsymbol{h}_{i0}\|_2^2 \leq \left(\left(\frac{\delta_i}{1-\delta_i}\right)^2 + \delta_0^2\right)\|\boldsymbol{h}_i\|_2^2 \leq (4.6\delta_i^2 + 4\delta_0^2)d_{i0}. \tag{6.24}$$

Then we calculate $\|\Delta x_i\|$: from (6.23), we have

$$\|\Delta x_i\|^2 = \left| \alpha_{i2} - \frac{1}{(1-\delta_0)\overline{\alpha}_{i1}} \right|^2 d_{i0} + \|\tilde{x}_i\|^2 \leq \left| \alpha_{i2} - \frac{1}{(1-\delta_0)\overline{\alpha}_{i1}} \right|^2 d_{i0} + \frac{4\delta_i^2 d_{i0}}{(1-\delta_i)^2},$$

where Lemma 6.1 gives $\|\tilde{x}_i\|_2 \leq \frac{\delta_i}{1-\delta_i}\|x_i\|_2 \leq \frac{2\delta_i}{1-\delta_i}\sqrt{d_{i0}}$ for $(h, x) \in \mathcal{N}_d \cap \mathcal{N}_\epsilon$.

So it suffices to estimate $\left| \alpha_{i2} - \frac{1}{(1-\delta_0)\overline{\alpha}_{i1}} \right|$, which satisfies

$$\left| \alpha_{i2} - \frac{1}{(1-\delta_0)\overline{\alpha}_{i1}} \right| = \frac{1}{|\alpha_{i1}|} \left| \overline{\alpha_{i1}}\alpha_{i2} - 1 - \frac{\delta_0}{1-\delta_0} \right| \leq \frac{1}{|\alpha_{i1}|} \left(|(\overline{\alpha_{i1}}\alpha_{i2} - 1)| + \frac{\delta_0}{1-\delta_0} \right). \quad (6.25)$$

Lemma 6.1 implies that $|\overline{\alpha_{i1}}\alpha_{i2} - 1| \leq \delta_i$, and (6.2) gives

$$|\alpha_{i1}|^2 = \frac{1}{d_{i0}}(\|h_i\|^2 - \|\tilde{h}_i\|^2) \geq \frac{1}{d_{i0}}\left(1 - \frac{\delta_i^2}{(1-\delta_i)^2} \right)\|h_i\|^2 \geq \left(1 - \frac{\delta_i^2}{(1-\delta_i)^2} \right)(1-\varepsilon), \quad (6.26)$$

where $\|\tilde{h}_i\| \leq \frac{\delta_i}{1-\delta_i}\|h_i\|$ and $\|h_i\|^2 \geq \|h_i\|\|x_i\| \geq (1-\varepsilon)d_{i0}$ if $\|h_i\| \geq \|x_i\|$. Substituting (6.26) into (6.25) gives

$$\left| \alpha_{i2} - \frac{1}{(1-\delta_0)\overline{\alpha}_{i1}} \right| \leq \frac{1}{\sqrt{1-\varepsilon}}\left(1 - \frac{\delta_i^2}{(1-\delta_i)^2} \right)^{-1/2}\left(\delta_i + \frac{\delta_0}{1-\delta_0} \right) \leq 1.2(\delta_i + \delta_0).$$

Then we have

$$\|\Delta x_i\|_2^2 \leq \left(1.44(\delta_i + \delta_0)^2 + \frac{4\delta_i^2}{(1-\delta_i)^2} \right)d_{i0} \leq (7.5\delta_i^2 + 2.88\delta_0^2)d_{i0}. \quad (6.27)$$

Finally, we try to bound $\|\Delta h_i\|^2\|\Delta x_i\|^2$. Lemma 6.1 gives $\|\tilde{h}_i\|_2\|\tilde{x}_i\|_2 \leq \frac{\delta_i^2 d_{i0}}{2(1-\delta_i)}$ and $|\alpha_{i1}| \leq 2$. Combining them along with (6.22), (6.23), (6.24) and (6.27), we have

$$\|\Delta h_i\|_2^2\|\Delta x_i\|_2^2 \leq \|\tilde{h}_i\|_2^2\|\tilde{x}_i\|_2^2 + \delta_0^2|\alpha_{i1}|^2\|h_{i0}\|_2^2\|\Delta x_i\|_2^2 + \left| \alpha_{i2} - \frac{1}{(1-\delta_0)\overline{\alpha}_{i1}} \right|^2\|x_{i0}\|_2^2\|\Delta h_i\|_2^2$$

$$\leq \left(\frac{\delta_i^4}{4(1-\delta_i)^2} + 4\delta_0^2(7.5\delta_i^2 + 2.88\delta_0^2) + 1.44(\delta_i + \delta_0)^2(4.6\delta_i^2 + 4\delta_0^2) \right)d_{i0}^2$$

$$\leq \frac{(\delta_i^2 + \delta_0^2)d_{i0}^2}{26}.$$

By symmetry, similar results hold for the case $\|h_i\|_2 < \|x_i\|_2$ and $\max\{\|\Delta h_i\|, \|\Delta x_i\|\} \leq (7.5\delta_i^2 + 2.88\delta_0^2)d_{i0}$.

Next, under the additional assumption $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_\mu$, we now prove $\sqrt{L}\|\boldsymbol{B}(\Delta\boldsymbol{h}_i)\|_\infty \le 6\mu\sqrt{d_{i0}}$:

CASE 1 : $\|\boldsymbol{h}_i\|_2 \ge \|\boldsymbol{x}_i\|_2$ and $\alpha_i = (1 - \delta_0)\alpha_{i1}$. By Lemma 6.1 gives $|\alpha_{i1}| \le 2$, which implies

$$\sqrt{L}\|\boldsymbol{B}(\Delta\boldsymbol{h}_i)\|_\infty \le \sqrt{L}\|\boldsymbol{B}\boldsymbol{h}_i\|_\infty + (1 - \delta_0)|\alpha_{i1}|\sqrt{L}\|\boldsymbol{B}\boldsymbol{h}_{i0}\|_\infty$$
$$\le 4\mu\sqrt{d_{i0}} + 2(1 - \delta_0)\mu_h\sqrt{d_{i0}} \le 6\mu\sqrt{d_{i0}}.$$

CASE 2 : $\|\boldsymbol{h}_i\|_2 < \|\boldsymbol{x}_i\|_2$ and $\alpha_i = \frac{1}{(1-\delta_0)\overline{\alpha_{i2}}}$. Using the same argument as (63) gives

$$|\alpha_{i2}|^2 \ge \left(1 - \frac{\delta_i^2}{(1 - \delta_i)^2}\right)(1 - \varepsilon).$$

Therefore,

$$\sqrt{L}\|\boldsymbol{B}(\Delta\boldsymbol{h}_i)\|_\infty \le \sqrt{L}\|\boldsymbol{B}\boldsymbol{h}_i\|_\infty + \frac{1}{(1 - \delta_0)|\overline{\alpha_{i2}}|}\sqrt{L}\|\boldsymbol{B}\boldsymbol{h}_0\|_\infty$$
$$\le 4\mu\sqrt{d_0} + \left(1 - \frac{\delta_i^2}{(1 - \delta_i)^2}\right)^{-1/2}\frac{\mu_h\sqrt{d_0}}{(1 - \delta_0)\sqrt{1 - \varepsilon}} \le 6\mu\sqrt{d_0}.$$

$\square$

LEMMA 6.10 **(Local Regularity for $F(\boldsymbol{h}, \boldsymbol{x})$)** Conditioned on (31) and (46), the following inequality holds

$$\mathrm{Re}\left(\langle\nabla F_{\boldsymbol{h}}, \Delta\boldsymbol{h}\rangle + \langle\nabla F_{\boldsymbol{x}}, \Delta\boldsymbol{x}\rangle\right) \ge \frac{\delta^2 d_0^2}{8} - 2\sqrt{s}\delta d_0\|\mathcal{A}^*(\boldsymbol{e})\|,$$

uniformly for any $(\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$ with $\epsilon \le \frac{1}{15}$ if $L \ge C\mu^2 s(K + N)\log^2 L$ for some numerical constant C.

Proof. First note that for

$$I_0 = \langle\nabla F_{\boldsymbol{h}}, \Delta\boldsymbol{h}\rangle + \overline{\langle\nabla F_{\boldsymbol{x}}, \Delta\boldsymbol{x}\rangle} = \sum_{i=1}^{s}\langle\nabla F_{\boldsymbol{h}_i}, \Delta\boldsymbol{h}_i\rangle + \overline{\langle\nabla F_{\boldsymbol{x}_i}, \Delta\boldsymbol{x}_i\rangle}.$$

For each component, recall that (19) and (20), we have

$$\langle\nabla F_{\boldsymbol{h}_i}, \Delta\boldsymbol{h}_i\rangle + \overline{\langle\nabla F_{\boldsymbol{x}_i}, \Delta\boldsymbol{x}_i\rangle} = \langle\mathcal{A}_i^*(\mathcal{A}(X - X_0) - \boldsymbol{e})\boldsymbol{x}_i, \Delta\boldsymbol{h}_i\rangle + \overline{\langle(\mathcal{A}_i^*(\mathcal{A}(X - X_0) - \boldsymbol{e}))^*\boldsymbol{h}_i, \Delta\boldsymbol{x}_i\rangle}$$
$$= \left\langle\mathcal{A}(X - X_0) - \boldsymbol{e}, \mathcal{A}_i((\Delta\boldsymbol{h}_i)\boldsymbol{x}_i^* + \boldsymbol{h}_i(\Delta\boldsymbol{x}_i)^*)\right\rangle.$$

Define $\boldsymbol{U}_i$ and $\boldsymbol{V}_i$ as

$$\boldsymbol{U}_i := \alpha_i\boldsymbol{h}_{i0}(\Delta\boldsymbol{x}_i)^* + \overline{\alpha_i}^{-1}(\Delta\boldsymbol{h}_i)\boldsymbol{x}_{i0}^* \in T_i, \quad \boldsymbol{V}_i := \Delta\boldsymbol{h}_i(\Delta\boldsymbol{x}_i)^*. \tag{6.28}$$

Here V_i does not necessarily belong to $T_i^\perp$. From the way of how Δh_i, Δx_i, U_i and V_i are constructed, two simple relations hold:

$$h_i x_i^* - h_{i0} x_{i0}^* = U_i + V_i, \qquad (\Delta h_i) x_i^* + h_i (\Delta x_i)^* = U_i + 2V_i.$$

Define $U := \mathrm{blkdiag}(U_1, \cdots, U_s)$ and $V := \mathrm{blkdiag}(V_1, \cdots, V_s)$. I_0 can be simplified to

$$I_0 = \sum_{i=1}^{s} \langle \nabla F_{h_i}, \Delta h_i \rangle + \overline{\langle \nabla F_{x_i}, \Delta x_i \rangle} = \sum_{i=1}^{s} \langle \mathcal{A}(U+V) - e, \mathcal{A}_i(U_i + 2V_i) \rangle$$

$$= \underbrace{\langle \mathcal{A}(U+V), \mathcal{A}(U+2V) \rangle}_{I_{01}} - \underbrace{\langle e, \mathcal{A}(U+2V) \rangle}_{I_{02}}.$$

Now we will give a lower bound for $\mathrm{Re}(I_{01})$ and an upper bound for $\mathrm{Re}(I_{02})$ so that the lower bound of $\mathrm{Re}(I_0)$ is obtained. By the Cauchy–Schwarz inequality, $\mathrm{Re}(I_{01})$ has the lower bound

$$\mathrm{Re}(I_{01}) \geq (\|\mathcal{A}(U)\| - \|\mathcal{A}(V)\|)(\|\mathcal{A}(U)\| - 2\|\mathcal{A}(V)\|). \tag{6.29}$$

In the following, we will give an upper bound for $\|\mathcal{A}(V)\|$ and a lower bound for $\|\mathcal{A}(U)\|$.
Upper bound for $\|\mathcal{A}(V)\|$: note that V is a block-diagonal matrix with rank-one blocks, and applying Lemma 6.6 results in

$$\|\mathcal{A}(V)\|^2 \leq \frac{4}{3} \sum_{i=1}^{s} \|V\|_F^2 + 2\sigma_{\max}(\Delta h, \Delta x) \|V\|_F \sqrt{2s(K+N)\log L} + 8s\sigma_{\max}^2(\Delta h, \Delta x)(K+N)\log L.$$

By using Lemma 6.9, we have $\|\Delta h_i\|^2 \leq (7.5\delta_i^2 + 2.88\delta_0^2)d_{i0}$ and $\sqrt{L}\|B(\Delta h_i)\|_\infty \leq 6\mu\sqrt{d_{i0}}$. Substituting them into $\sigma_{\max}^2(\Delta h, \Delta x)$ gives

$$\sigma_{\max}^2(\Delta h, \Delta x) = \max_{1 \leq l \leq L} \left(\sum_{i=1}^{s} |b_l^* \Delta h_i|^2 \|\Delta x_i\|^2 \right) \leq \frac{36\mu^2}{L} \sum_{i=1}^{s} \left(7.5\delta_i^2 + 2.88\delta_0^2 \right) d_{i0}^2 \leq \frac{272\mu^2 \delta^2 d_0^2}{L}.$$

For $\|V\|_F$, note that $\|\Delta h_i\|^2 \|\Delta x_i\|^2 \leq \frac{1}{26}(\delta_i^2 + \delta_0^2)d_{i0}^2$ and thus

$$\|V\|_F^2 = \sum_{i=1}^{s} \|\Delta h_i\|^2 \|\Delta x_i\|^2 \leq \frac{1}{26} \sum_{i=1}^{s} \left(\delta_i^2 + \delta_0^2 \right) d_{i0}^2 \leq \frac{1}{26} \cdot 1.01 \delta^2 d_0^2 = \frac{\delta^2 d_0^2}{25}.$$

Then by $\delta \leq \varepsilon \leq \frac{1}{15}$ and letting $L \geq C\mu^2 s(K+N)\log^2 L$ for a sufficiently large numerical constant C, there holds

$$\|\mathcal{A}(V)\|^2 \leq \frac{\delta^2 d_0^2}{16} \implies \|\mathcal{A}(V)\| \leq \frac{\delta d_0}{4}. \tag{6.30}$$

Lower bound for $\|\mathcal{A}(U)\|$**:** by the triangle inequality, $\|U\|_F \geq \delta d_0 - \frac{1}{5}\delta d_0 \geq \frac{4}{5}\delta d_0$ if $\epsilon \leq \frac{1}{15}$ since $\|V\|_F \leq 0.2\delta d_0$. Since $U \in T$, by Lemma 6.3, there holds

$$\|\mathcal{A}(U)\| \geq \sqrt{\frac{9}{10}}\|U\|_F \geq \frac{3}{4}\delta d_0. \tag{6.31}$$

With the upper bound of $\mathcal{A}(V)$ in (6.30), the lower bound of $\mathcal{A}(U)$ in (6.31), and (6.29), we get $\mathrm{Re}(I_{01}) \geq \frac{\delta^2 d_0^2}{8}$.

Now let us give an upper bound for $\mathrm{Re}(I_{02})$,

$$\|I_{02}\| \leq \|\mathcal{A}^*(e)\|\|U + 2V\|_* = \|\mathcal{A}^*(e)\|\sum_{i=1}^{s}\|\underbrace{U_i + 2V_i}_{\text{rank-2}}\|_*$$

$$\leq \sqrt{2}\|\mathcal{A}^*(e)\|\sum_{i=1}^{s}\|U_i + 2V_i\|_F$$

$$\leq \sqrt{2s}\|\mathcal{A}^*(e)\|\|U + 2V\|_F \leq 2\sqrt{s}\delta d_0\|\mathcal{A}^*(e)\|,$$

where $\|\cdot\|$ and $\|\cdot\|_*$ are a pair of dual norms and

$$\|U + 2V\|_F \leq \|U + V\|_F + \|V\|_F \leq \delta d_0 + 0.2\delta d_0 \leq 1.2\delta d_0.$$

Combining the estimation of $\mathrm{Re}(I_{01})$ and $\mathrm{Re}(I_{02})$ above leads to

$$\mathrm{Re}(\langle \nabla F_h, \Delta h \rangle + \langle \nabla F_x, \Delta x \rangle) \geq \frac{\delta^2 d_0^2}{8} - 2\sqrt{s}\delta d_0\|\mathcal{A}^*(e)\|.$$

$\square$

LEMMA 6.11 **(Local Regularity for** $G(h, x)$**)** For any $(h, x) \in \mathcal{N}_d \bigcap \mathcal{N}_\epsilon$ with $\epsilon \leq \frac{1}{15}$ and $\frac{9}{10}d_0 \leq d \leq \frac{11}{10}d_0$, $\frac{9}{10}d_{i0} \leq d_i \leq \frac{11}{10}d_{i0}$, the following inequality holds uniformly

$$\mathrm{Re}\left(\langle \nabla G_{h_i}, \Delta h_i \rangle + \langle \nabla G_{x_i}, \Delta x_i \rangle\right) \geq 2\delta_0\sqrt{\rho G_i(h_i, x_i)} = \frac{\delta}{5}\sqrt{\rho G_i(h_i, x_i)}, \tag{6.32}$$

where $\rho \geq d^2 + 2\|e\|^2$. Immediately, we have

$$\mathrm{Re}\left(\langle \nabla G_h, \Delta h \rangle + \langle \nabla G_x, \Delta x \rangle\right) = \sum_{i=1}^{s}\mathrm{Re}\left(\langle \nabla G_{h_i}, \Delta h_i \rangle + \langle \nabla G_{x_i}, \Delta x_i \rangle\right) \geq \frac{\delta}{5}\sqrt{\rho G(h, x)}. \tag{6.33}$$

REMARK 6.12 For the local regularity condition for $G(h, x)$, we use the results from [21] when $s = 1$. This is because each component $G_i(h, x)$ only depends on (h_i, x_i) by definition and thus the lower bound of $\mathrm{Re}\left(\langle \nabla G_{h_i}, \Delta h_i \rangle + \langle \nabla G_{x_i}, \Delta x_i \rangle\right)$ is completely determined by (h_i, x_i) and δ_0, and is independent of s.

Proof. For each $i : 1 \leq i \leq s$, ∇G_{h_i} (or ∇G_{x_i}) only depends on h_i (or x_i) and there holds

$$\mathrm{Re}\left(\langle \nabla G_{h_i}, \Delta h_i \rangle + \langle \nabla G_{x_i}, \Delta x_i \rangle\right) \geq 2\delta_0 \sqrt{\rho G_i(h_i, x_i)} = \frac{\delta}{5}\sqrt{\rho G_i(h_i, x_i)},$$

which follows exactly from Lemma 5.17 in [**21**]. For (6.33), by definition of ∇G_h and ∇G_x in (2.22),

$$\mathrm{Re}\left(\langle \nabla G_h, \Delta h \rangle + \langle \nabla G_x, \Delta x \rangle\right) = \sum_{i=1}^{s} \mathrm{Re}\left(\langle \nabla G_{h_i}, \Delta h_i \rangle + \langle \nabla G_{x_i}, \Delta x_i \rangle\right)$$

$$\geq \frac{\delta}{5} \sum_{i=1}^{s} \sqrt{\rho G_i(h_i, x_i)} \geq \frac{\delta}{5}\sqrt{\rho G(h, x)},$$

where $G(h, x) = \sum_{i=1}^{s} G_i(h_i, x_i)$. $\qquad\qquad\square$

LEMMA 6.13 **(Proof of the Local Regularity Condition)** Conditioned on (5.3), for the objective function $\widetilde{F}(h, x)$ in (2.17), there exists a positive constant ω such that

$$\|\nabla \widetilde{F}(h, x)\|^2 \geq \omega \left[\widetilde{F}(h, x) - c\right]_+ \tag{6.34}$$

with $c = \|e\|^2 + 2000s\|\mathcal{A}^*(e)\|^2$ and $\omega = \frac{d_0}{7000}$ for all $(h, x) \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$. Here we set $\rho \geq d^2 + 2\|e\|^2$.

Proof. Following from Lemma 6.10 and Lemma 6.11, we have

$$\mathrm{Re}(\langle \nabla F_h, \Delta h \rangle + \langle \nabla F_x, \Delta x \rangle) \geq \frac{\delta^2 d_0^2}{8} - 2\sqrt{s}\delta d_0 \|\mathcal{A}^*(e)\|$$

$$\mathrm{Re}(\langle \nabla G_h, \Delta h \rangle + \langle \nabla G_x, \Delta x \rangle) \geq \frac{\delta d}{5}\sqrt{G(h, x)} \geq \frac{9\delta d_0}{50}\sqrt{G(h, x)}$$

for all $(h, x) \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$ where $\rho \geq d^2 + 2\|e\|^2 \geq d^2$ and $\frac{9}{10}d_0 \leq d \leq \frac{11}{10}d_0$. Adding them together gives $\mathrm{Re}\left(\langle \nabla \widetilde{F}_h, \Delta h \rangle + \langle \nabla \widetilde{F}_x, \Delta x \rangle\right)$ on the left side. Moreover, Cauchy–Schwarz inequality implies

$$\mathrm{Re}\left(\langle \nabla \widetilde{F}_h, \Delta h \rangle + \langle \nabla \widetilde{F}_x, \Delta x \rangle\right) \leq 4\delta\sqrt{d_0}\|\nabla \widetilde{F}(h, x)\|,$$

where both $\|\Delta h\|^2$ and $\|\Delta x\|^2$ are bounded by $8\delta^2 d_0$ in Lemma 6.9 since

$$\|\Delta h\|^2 = \sum_{i=1}^{s} \|\Delta h_i\|^2 \leq \sum_{i=1}^{s}(7.5\delta_i^2 + 2.88\delta_0^2)d_{i0} \leq 8\delta^2 d_0.$$

Therefore,

$$\frac{\delta^2 d_0^2}{8} + \frac{9\delta d_0 \sqrt{G(h, x)}}{50} - 2\sqrt{s}\delta d_0 \|\mathcal{A}^*(e)\| \leq 4\delta\sqrt{d_0}\|\nabla \widetilde{F}(h, x)\|. \tag{6.35}$$

Dividing both sides of (6.35) by δd_0, we obtain

$$\frac{4}{\sqrt{d_0}}\|\nabla\widetilde{F}(\boldsymbol{h},\boldsymbol{x})\| \geq \frac{\delta d_0}{12} + \frac{9}{50}\sqrt{G(\boldsymbol{h},\boldsymbol{x})} + \frac{\delta d_0}{24} - 2\sqrt{s}\|\mathcal{A}^*(\boldsymbol{e})\|$$

$$\geq \frac{1}{6\sqrt{6}}[\sqrt{F_0(\boldsymbol{h},\boldsymbol{x})} + \sqrt{G(\boldsymbol{h},\boldsymbol{x})}] + \frac{\delta d_0}{24} - 2\sqrt{s}\|\mathcal{A}^*(\boldsymbol{e})\|,$$

where the Local RIP condition (5.3) implies $F_0(\boldsymbol{h},\boldsymbol{x}) \leq \frac{3}{2}\delta^2 d_0^2$ and hence $\frac{\delta d_0}{12} \geq \frac{1}{6\sqrt{6}}\sqrt{F_0(\boldsymbol{h},\boldsymbol{x})}$, where $F_0(\boldsymbol{h},\boldsymbol{x})$ is defined in (2.12).

Note that (5.6) gives

$$\sqrt{2\left[\mathrm{Re}(\langle\mathcal{A}^*(\boldsymbol{e}),\boldsymbol{X}-\boldsymbol{X}_0\rangle)\right]_+} \leq \sqrt{2\sqrt{2s}\|\mathcal{A}^*(\boldsymbol{e})\|\delta d_0} \leq \frac{\sqrt{6}\delta d_0}{4} + \frac{4\sqrt{s}}{\sqrt{6}}\|\mathcal{A}^*(\boldsymbol{e})\|. \tag{6.36}$$

By (6.36) and $\widetilde{F}(\boldsymbol{h},\boldsymbol{x}) - \|\boldsymbol{e}\|^2 \leq F_0(\boldsymbol{h},\boldsymbol{x}) + 2[\mathrm{Re}(\langle\mathcal{A}^*(\boldsymbol{e}),\boldsymbol{X}-\boldsymbol{X}_0\rangle)]_+ + G(\boldsymbol{h},\boldsymbol{x})$, there holds

$$\frac{4}{\sqrt{d_0}}\|\nabla\widetilde{F}(\boldsymbol{h},\boldsymbol{x})\| \geq \frac{1}{6\sqrt{6}}\left[\left(\sqrt{F_0(\boldsymbol{h},\boldsymbol{x})} + \sqrt{2\left[\mathrm{Re}(\langle\mathcal{A}^*(\boldsymbol{e}),\boldsymbol{X}-\boldsymbol{X}_0\rangle)\right]_+} + \sqrt{G(\boldsymbol{h},\boldsymbol{x})}\right)\right.$$

$$+ \frac{\delta d_0}{24} - \frac{1}{6\sqrt{6}}\left(\frac{\sqrt{6}\delta d_0}{4} + \frac{4\sqrt{s}}{\sqrt{6}}\|\mathcal{A}^*(\boldsymbol{e})\|\right) - 2\sqrt{s}\|\mathcal{A}^*(\boldsymbol{e})\|$$

$$\geq \frac{1}{6\sqrt{6}}\left[\sqrt{\left[\widetilde{F}(\boldsymbol{h},\boldsymbol{x}) - \|\boldsymbol{e}\|^2\right]_+} - \sqrt{1000s}\|\mathcal{A}^*(\boldsymbol{e})\|\right].$$

For any non-negative real numbers a and b, we have $[\sqrt{(x-a)_+} - b]_+ + b \geq \sqrt{(x-a)_+}$ and it implies

$$(x-a)_+ \leq 2([\sqrt{(x-a)_+} - b]_+^2 + b^2) \Longrightarrow [\sqrt{(x-a)_+} - b]_+^2 \geq \frac{(x-a)_+}{2} - b^2.$$

Therefore, by setting $a = \|\boldsymbol{e}\|^2$ and $b = \sqrt{1000s}\|\mathcal{A}^*(\boldsymbol{e})\|$, there holds

$$\|\nabla\widetilde{F}(\boldsymbol{h},\boldsymbol{x})\|^2 \geq \frac{d_0}{3500}\left[\frac{\widetilde{F}(\boldsymbol{h},\boldsymbol{x}) - \|\boldsymbol{e}\|^2}{2} - 1000s\|\mathcal{A}^*(\boldsymbol{e})\|^2\right]_+$$

$$\geq \frac{d_0}{7000}\left[\widetilde{F}(\boldsymbol{h},\boldsymbol{x}) - (\|\boldsymbol{e}\|^2 + 2000s\|\mathcal{A}^*(\boldsymbol{e})\|^2)\right]_+.$$

$\square$

6.4 *Local smoothness*

LEMMA 6.14 Conditioned on (5.3), (5.4) and (6.4), for any $\boldsymbol{z} := (\boldsymbol{h},\boldsymbol{x}) \in \mathbb{C}^{(K+N)s}$ and $\boldsymbol{w} := (\boldsymbol{u},\boldsymbol{v}) \in \mathbb{C}^{(K+N)s}$ such that $\boldsymbol{z}$ and $\boldsymbol{z} + \boldsymbol{w} \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$, there holds

$$\|\nabla\widetilde{F}(\boldsymbol{z}+\boldsymbol{w}) - \nabla\widetilde{F}(\boldsymbol{z})\| \leq C_L\|\boldsymbol{w}\|,$$

with

$$C_L \leq \left(10\|\mathcal{A}\|^2 d_0 + \frac{2\rho}{\min d_{i0}} \left(5 + \frac{2L}{\mu^2} \right) \right),$$

where $\rho \geq d^2 + 2\|e\|^2$ and $\|\mathcal{A}\| \leq \sqrt{s(N \log(NL/2) + (\gamma + \log s) \log L)}$ holds with probability at least $1 - L^{-\gamma}$ from Lemma 6.2.

In particular, $L = \mathcal{O}((\mu^2 + \sigma^2)s(K+N)\log^2 L)$ and $\|e\|^2 = \mathcal{O}(\sigma^2 d_0^2)$ follows from $\|e\|^2 \sim \frac{\sigma^2 d_0^2}{2L} \chi_{2L}^2$ and (6.14). Therefore, C_L can be simplified to

$$C_L = \mathcal{O}(d_0 s\kappa(1 + \sigma^2)(K+N)\log^2 L)$$

by choosing $\rho \approx d^2 + 2\|e\|^2$.

Proof. By Lemma 5.6, we know that both $z = (h, x)$ and $z + w = (h + u, x + v) \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$. Note that

$$\nabla \widetilde{F} = (\nabla \widetilde{F}_h, \nabla \widetilde{F}_x) = (\nabla F_h + \nabla G_h, \nabla F_x + \nabla G_x),$$

where (2.18), (2.19), (2.20) and (2.21) give ∇F_h, ∇F_x, ∇G_h and ∇G_x. It suffices to find out the Lipschitz constants for all of those four functions.

Step 1: we first estimate the Lipschitz constant for ∇F_h and the result can be applied to ∇F_x due to symmetry.

$$\begin{aligned}
\nabla F_h(z + w) - \nabla F_h(z) &= \mathcal{A}^*\mathcal{A}(\mathcal{H}(h + u, x + v))(x + v) - \left[\mathcal{A}^*\mathcal{A}(\mathcal{H}(h, x))x + \mathcal{A}^*(y)v \right] \\
&= \mathcal{A}^*(\mathcal{A}(\mathcal{H}(h + u, x + v) - \mathcal{H}(h, x)))(x + v) \\
&\quad + \mathcal{A}^*\mathcal{A}(\mathcal{H}(h, x) - \mathcal{H}(h_0, x_0))v - \mathcal{A}^*(e)v \\
&= \mathcal{A}^*(\mathcal{A}(\mathcal{H}(h + u, v) + \mathcal{H}(u, x)))(x + v) \\
&\quad + \mathcal{A}^*\mathcal{A}(\mathcal{H}(h, x) - \mathcal{H}(h_0, x_0))v - \mathcal{A}^*(e)v.
\end{aligned}$$

Note that $\|\mathcal{H}(h, x)\|_F \leq \sqrt{\sum_{i=1}^{s} \|h_i\|^2 \|x_i\|^2} \leq \|h\|\|x\|$ and $z, z + w \in \mathcal{N}_d$ directly implies

$$\|\mathcal{H}(u, x) + \mathcal{H}(h + u, v)\|_F \leq \|u\|\|x\| + \|h + u\|\|v\| \leq 2\sqrt{d_0}(\|u\| + \|v\|),$$

where $\|h + u\| \leq 2\sqrt{d_0}$. Moreover, (31) implies

$$\|\mathcal{H}(h, x) - \mathcal{H}(h_0, x_0)\|_F \leq \epsilon d_0$$

since $z \in \mathcal{N}_d \cap \mathcal{N}_\mu \cap \mathcal{N}_\epsilon$. Combined with $\|\mathcal{A}^*(e)\| \leq \varepsilon d_0$ in (5.4) and $\|x + v\| \leq 2\sqrt{d_0}$, we have

$$\begin{aligned}
\|\nabla F_h(z + w) - \nabla F_h(z)\| &\leq 4d_0\|\mathcal{A}\|^2(\|u\| + \|v\|) + \varepsilon d_0\|\mathcal{A}\|^2\|v\| + \varepsilon d_0\|v\| \\
&\leq 5d_0\|\mathcal{A}\|^2(\|u\| + \|v\|). \quad (6.37)
\end{aligned}$$

Due to the symmetry between ∇F_h and ∇F_x, we have,

$$\|\nabla F_x(z + w) - \nabla F_x(z)\| \le 5d_0\|\mathcal{A}\|^2(\|u\| + \|v\|). \tag{6.38}$$

In other words,

$$\|\nabla F(z + w) - \nabla F(z)\| \le 5\sqrt{2}d_0\|\mathcal{A}\|^2(\|u\| + \|v\|) \le 10d_0\|\mathcal{A}\|^2\|w\|,$$

where $\|u\| + \|v\| \le \sqrt{2}\|w\|$.

Step 2: we estimate the upper bound of $\|\nabla G_{x_i}(z_i + w_i) - \nabla G_{x_i}(z_i)\|$. Implied by Lemma 5.19 in [21], we have

$$\|\nabla G_{x_i}(z_i + w_i) - \nabla G_{x_i}(z_i)\| \le \frac{5d_{i0}\rho}{d_i^2}\|v_i\|. \tag{6.39}$$

Step 3: we estimate the upper bound of $\|\nabla G_{h_i}(z + w) - \nabla G_{h_i}(z)\|$. Denote

$$\nabla G_{h_i}(z + w) - \nabla G_{h_i}(z) = \underbrace{\frac{\rho}{2d_i}\left[G_0'\left(\frac{\|h_i + u_i\|^2}{2d_i}\right)(h_i + u_i) - G_0'\left(\frac{\|h_i\|^2}{2d_i}\right)h_i\right]}_{j_1}$$

$$+ \underbrace{\frac{\rho L}{8d_i\mu^2}\sum_{l=1}^{L}\left[G_0'\left(\frac{L|b_l^*(h_i + u_i)|^2}{8d_i\mu^2}\right)b_l^*(h_i + u_i) - G_0'\left(\frac{L|b_l^*h_i|^2}{8d_i\mu^2}\right)b_l^*h_i\right]b_l}_{j_2}.$$

Following the same estimation of j_1 and j_2 in Lemma 5.19 of [21], we have

$$\|j_1\| \le \frac{5d_{i0}\rho}{d_i^2}\|u_i\|, \quad \|j_2\| \le \frac{3\rho L d_{i0}}{2d_i^2\mu^2}\|u_i\|. \tag{6.40}$$

Therefore, combining (6.39) and (6.40) gives

$$\|\nabla G(z + w) - \nabla G(z)\| = \sqrt{\sum_{i=1}^{s}\left(\|\nabla G_{h_i}(z + w) - \nabla G_{h_i}(z)\|^2 + \|\nabla G_{x_i}(z + w) - \nabla G_{x_i}(z)\|^2\right)}$$

$$\le \max\left\{\frac{5d_{i0}\rho}{d_i^2} + \frac{3\rho L d_{i0}}{2d_i^2\mu^2}\right\}\sqrt{\sum_{i=1}^{s}\|u_i\|^2} + \max\left\{\frac{5d_{i0}\rho}{d_i^2}\right\}\sqrt{\sum_{i=1}^{s}\|v_i\|^2}$$

$$\le \max\left\{\frac{5d_{i0}\rho}{d_i^2} + \frac{3\rho L d_{i0}}{2d_i^2\mu^2}\right\}\|u\| + \max\left\{\frac{5d_{i0}\rho}{d_i^2}\right\}\|v\|$$

$$\le \frac{2\rho}{\min d_{i0}}\left(5 + \frac{2L}{\mu^2}\right)\|w\|.$$

In summary, the Lipschitz constant C_L of $\widetilde{F}(z)$ has an upper bound as follows:

$$\|\nabla \widetilde{F}(z+w) - \nabla \widetilde{F}(z)\| \leq \|\nabla F(z+w) - \nabla F(z)\| + \|\nabla G(z+w) - \nabla G(z)\|$$

$$\leq \left(10\|\mathcal{A}\|^2 d_0 + \frac{2\rho}{\min d_{i0}} \left(5 + \frac{2L}{\mu^2} \right) \right) \|w\|.$$

$\square$

6.5 *Robustness condition and spectral initialization*

In this section, we will prove the robustness condition (5.4) and also Theorem 3.2. To prove (5.4), it suffices to show the following lemma, which is a more general version of (5.4).

LEMMA 6.15 Consider a sequence of Gaussian independent random variable $c = (c_1, \cdots, c_L) \in \mathbb{C}^L$ where $c_l \sim \mathcal{CN}\left(0, \frac{\lambda_i^2}{L}\right)$ with $\lambda_i \leq \lambda$. Moreover, we assume $\mathcal{A}_i$ in (2.2) is independent of c. Then there holds

$$\|\mathcal{A}^*(c) = \| \max_{1 \leq i \leq s} \|\mathcal{A}_i^*(c)\| \leq \xi$$

with probability at least $1 - L^{-\gamma}$ if $L \geq C_{\gamma + \log(s)}\left(\frac{\lambda}{\xi} + \frac{\lambda^2}{\xi^2}\right) \max\{K, N\} \log L / \xi^2$.

Proof. It suffices to show that $\max_{1 \leq i \leq s} \|\mathcal{A}_i^*(c)\| \leq \xi$. For each fixed $i : 1 \leq i \leq s$,

$$\mathcal{A}_i^*(c) = \sum_{l=1}^{L} c_l b_l a_{il}^*.$$

The key is to apply the matrix Bernstein inequality (6.53) and we need to estimate $\|\mathcal{Z}_l\|_{\psi_1}$, and the variance of $\sum_{l=1}^{L} \mathcal{Z}_l$. For each l, $\|c_l b_l a_{il}^*\|_{\psi_1} \leq \frac{\lambda \sqrt{KN}}{L}$ follows from (6.58). Moreover, the variance of $\mathcal{A}_i^*(c)$ is bounded by $\frac{\lambda^2 \max\{K,N\}}{L}$ since

$$\mathbb{E}[\mathcal{A}_i^*(c)(\mathcal{A}_i^*(c))^*] = \sum_{l=1}^{L} \mathbb{E}(|c_l|^2 \|a_{il}\|^2) b_l b_l^* = \frac{N}{L} \sum_{l=1}^{L} \lambda_i^2 b_l b_l^* \preceq \frac{\lambda^2 N}{L},$$

$$\mathbb{E}[(\mathcal{A}_i^*(c))^*(\mathcal{A}_i^*(c))] = \sum_{l=1}^{L} \|b_l\|^2 \mathbb{E}(|c_l|^2 a_{il} a_{il}^*) = \frac{K}{L^2} \sum_{l=1}^{L} \lambda_i^2 I_N \preceq \frac{\lambda^2 K}{L}.$$

Letting $t = \gamma \log L$ and applying (6.53) lead to

$$\|\mathcal{A}_i^*(c)\| \leq C_0 \max \left\{ \frac{\lambda \sqrt{KN} \log^2 L}{L}, \sqrt{\frac{C_\gamma \lambda^2 \max\{K, N\} \log L}{L}} \right\} \leq \xi.$$

Therefore, by taking the union bound over $1 \leq i \leq s$,

$$\|\mathcal{A}_i^*(c)\| \leq \xi$$

with probability at least $1 - L^{-\gamma}$ if $L \geq C_{\gamma + \log(s)}\left(\frac{\lambda}{\xi} + \frac{\lambda^2}{\xi^2}\right) \max\{K, N\} \log^2 L$. $\qquad \square$

The robustness condition is an immediate result of Lemma 6.15 by setting $\xi = \frac{\varepsilon d_0}{10\sqrt{2}s\kappa}$ and $\lambda = \sigma d_0$.

COROLLARY 6.16 **[Robustness Condition]** For $e \sim \mathcal{CN}\left(0, \frac{\sigma^2 d_0^2}{L} I_L\right)$

$$\|\mathcal{A}_i^*(e)\| \leq \frac{\varepsilon d_0}{10\sqrt{2}s\kappa}, \quad \forall 1 \leq i \leq s$$

with probability at least $1 - L^{-\gamma}$ if $L \geq C_\gamma\left(\frac{s^2\kappa^2\sigma^2}{\varepsilon^2} + \frac{s\kappa\sigma}{\varepsilon}\right) \max\{K, N\} \log L$.

LEMMA 6.17 For $e \sim \mathcal{CN}\left(0, \frac{\sigma^2 d_0^2}{L} I_L\right)$, there holds

$$\|\mathcal{A}_i^*(y) - h_{i0}x_{i0}^*\| \leq \xi d_{i0}, \quad \forall 1 \leq i \leq s \qquad (6.41)$$

with probability at least $1 - L^{-\gamma}$ if $L \geq C_{\gamma + \log(s)}s\kappa^2(\mu_h^2 + \sigma^2) \max\{K, N\} \log L / \xi^2$.

REMARK 6.18 The success of the initialization algorithm completely relies on the lemma above. As mentioned in Section 3, $\mathbb{E}(\mathcal{A}_i^*(y)) = h_{i0}x_{i0}^*$ and Lemma 6.41 confirms that $\mathcal{A}_i^*(y)$ is close to $h_{i0}x_{i0}^*$ in operator norm and hence the spectral method is able to give us a reliable initialization.

Proof. Note that

$$\mathcal{A}_i^*(y) = \mathcal{A}_i^*\mathcal{A}_i(h_{i0}x_{i0}^*) + \mathcal{A}_i^*(w_i),$$

where

$$w_i = y - \mathcal{A}_i(h_{i0}x_{i0}^*) = \sum_{j \neq i} \mathcal{A}_j(h_{j0}x_{j0}^*) + e \qquad (6.42)$$

is independent of $\mathcal{A}_i$. The proof consists of two parts: 1. show that $\|\mathcal{A}_i^*\mathcal{A}_i(h_{i0}x_{i0}^*) - h_{i0}x_{i0}^*\| \leq \frac{\xi d_{i0}}{2}$; 2. prove that $\|\mathcal{A}_i^*(w_i)\| \leq \frac{\xi d_{i0}}{2}$.
Part I: following from the definition of $\mathcal{A}_i$ and $\mathcal{A}_i^*$ in (2.2),

$$\mathcal{A}_i^*\mathcal{A}_i(h_{i0}x_{i0}^*) - h_{i0}x_{i0}^* = \sum_{l=1}^L \underbrace{b_l b_l^* h_{i0} x_{i0}^* (a_{il} a_{il}^* - I_N)}_{\text{defined as } \mathcal{Z}_l},$$

where $B^*B = I_K$. The sub-exponential norm of $\mathcal{Z}_l$ is bounded by

$$\|\mathcal{Z}_l\|_{\psi_1} \leq \max_{1 \leq l \leq L} \|b_l\|\|b_l^* h_{i0}\|\|(a_{il}a_{il}^* - I_N)x_{i0}\|_{\psi_1} \leq \frac{\mu\sqrt{KN}d_{i0}}{L},$$

where $\|b_l\| = \sqrt{\frac{K}{L}}$, $\max_l |b_l^* h_{i0}|^2 \leq \frac{\mu^2 d_{i0}}{L}$ and $\|(a_{il}a_{il}^* - I_N)x_{i0}\|_{\psi_1} \leq \sqrt{N d_{i0}}$ follows from (6.56).

We proceed to estimate the variance of $\sum_{l=1}^L \mathcal{Z}_l$ by using (6.55) and (6.57):

$$\left\|\sum_{l=1}^L \mathbb{E}(\mathcal{Z}_l \mathcal{Z}_l^*)\right\| = \left\|\sum |b_l^* h_{i0}|^2 x_{i0}^* \mathbb{E}(a_{il}a_{il}^* - I_N)^2 x_{i0} b_l b_l^*\right\| \leq \frac{\mu^2 N d_{i0}^2}{L},$$

$$\left\|\sum_{l=1}^L \mathbb{E}(\mathcal{Z}_l^* \mathcal{Z}_l)\right\| = \frac{K}{L}\left\|\sum_{l=1}^L |b_l^* h_{i0}|^2 \mathbb{E}\left[(a_{il}a_{il}^* - I_N)x_{i0}x_{i0}^*(a_{il}a_{il}^* - I_N)\right]\right\| \leq \frac{K d_{i0}^2}{L}.$$

Therefore, the variance of $\sum_{l=1}^L \mathcal{Z}_l$ is bounded by $\frac{\max\{K, \mu_h^2 N\}d_{i0}^2}{L}$. By applying matrix Bernstein inequality (6.53) and taking the union bound over all i, we prove that

$$\|\mathcal{A}_i^* \mathcal{A}_i(h_{i0}x_{i0}^*) - h_{i0}x_{i0}^*\| \leq \frac{\xi d_{i0}}{2}, \quad \forall 1 \leq i \leq s$$

holds with probability at least $1 - L^{-\gamma+1}$ if $L \geq C_{\gamma + \log(s)} \max\{K, \mu_h^2 N\} \log L / \xi^2$.

Part II: for each $1 \leq l \leq L$, the lth entry of w_i in (6.42), i.e., $(w_i)_l = \sum_{j \neq i} b_l^* h_{j0}x_{j0}^* a_{jl} + e_l$, is independent of $b_l^* h_{i0}x_{i0}^* a_{il}$ and obeys $\mathcal{CN}\left(0, \frac{\sigma_{il}^2}{L}\right)$. Here

$$\sigma_{il}^2 = L \mathbb{E} |(w_i)_l|^2 = L \sum_{j \neq i} |b_l^* h_{j0}|^2 \|x_{j0}\|^2 + \sigma^2 \|X_0\|_F^2$$

$$\leq \mu_h^2 \sum_{j \neq i} \|h_{j0}\|^2 \|x_{j0}\|^2 + \sigma^2 \|X_0\|_F^2 \leq (\mu_h^2 + \sigma^2)\|X_0\|_F^2.$$

This gives $\max_{i,l} \sigma_{il}^2 \leq (\mu_h^2 + \sigma^2)\|X_0\|_F^2$. Thanks to the independence between w_i and $\mathcal{A}_i$, applying Lemma 6.15 results in

$$\|\mathcal{A}_i^*(w_i)\| \leq \frac{\xi d_{i0}}{2} \tag{6.43}$$

with probability $1 - L^{-\gamma+1}$ if $L \geq C \max\left(\frac{(\mu_h^2 + \sigma^2)\|X_0\|_F^2}{\xi^2 d_{i0}^2}, \frac{\sqrt{\mu_h^2 + \sigma^2}\|X_0\|_F}{\xi d_{i0}}\right) \max\{K, N\} \log L$.

Therefore, combining (6.42) with (6.43), we get

$$\|\mathcal{A}_i^*(y) - h_{i0}x_{i0}^*\| \leq \|\mathcal{A}_i^* \mathcal{A}_i(h_{i0}x_{i0}^*) - h_{i0}x_{i0}^*\| + \|\mathcal{A}_i^*(w_i)\| \leq \xi d_{i0}$$

for all $1 \leq i \leq s$ with probability at least $1 - L^{-\gamma+1}$ if

$$L \geq C_{\gamma+\log(s)}(\mu_h^2 + \sigma^2)s\kappa^2 \max\{K,N\}\log L/\xi^2,$$

where $\|\boldsymbol{X}_0\|_F/d_{i0} \leq \sqrt{s}\kappa$.   □

Before moving to the proof of Theorem 3.2, we introduce a property about the projection onto a closed convex set.

LEMMA 6.19 (Theorem 2.8 in [13]). Let $Q := \{\boldsymbol{w} \in \mathbb{C}^K | \sqrt{L}\|\boldsymbol{B}\boldsymbol{w}\|_\infty \leq 2\sqrt{d}\mu\}$ be a closed non-empty convex set. There holds

$$\operatorname{Re}(\langle z - \mathcal{P}_Q(z), \boldsymbol{w} - \mathcal{P}_Q(z)\rangle) \leq 0, \quad \forall \boldsymbol{w} \in Q, z \in \mathbb{C}^K,$$

where $\mathcal{P}_Q(z)$ is the projection of z onto Q.

With this lemma, we can easily see

$$\|z - \boldsymbol{w}\|^2 = \|z - \mathcal{P}_Q(z)\|^2 + \|\mathcal{P}_Q(z) - \boldsymbol{w}\|^2 + 2\operatorname{Re}(\langle z - \mathcal{P}_Q(z), \mathcal{P}_Q(z) - \boldsymbol{w}\rangle) \geq \|\mathcal{P}_Q(z) - \boldsymbol{w}\|^2 \quad (6.44)$$

for all $z \in \mathbb{C}^K$ and $\boldsymbol{w} \in Q$. It means that projection onto non-empty closed convex set is non-expansive. Now we present the proof of Theorem 3.2.

PROOF OF THEOREM 3.2   By choosing $L \geq C_{\gamma+\log(s)}(\mu_h^2 + \sigma^2)s^2\kappa^4 \max\{K,N\}\log L/\varepsilon^2$, we have

$$\|\mathcal{A}_i^*(\boldsymbol{y}) - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\| \leq \xi d_{i0}, \quad \forall 1 \leq i \leq s, \quad (6.45)$$

where $\xi = \frac{\varepsilon}{10\sqrt{2s}\kappa}$.

By applying the triangle inequality to (6.45), it is easy to see that

$$(1-\xi)d_{i0} \leq d_i \leq (1+\xi)d_{i0}, \quad |d_i - d_{i0}| \leq \xi d_{i0} \leq \frac{\varepsilon d_{i0}}{10\sqrt{2s}\kappa} < \frac{d_{i0}}{10}, \quad (6.46)$$

which gives $\frac{9}{10}d_{i0} \leq d_i \leq \frac{11}{10}d_{i0}$ where $d_i = \|\mathcal{A}_i^*(\boldsymbol{y})\|$ according to Algorithm 1.

Part I: proof of $(\boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}) \in \frac{1}{\sqrt{3}}\mathcal{N}_d \cap \frac{1}{\sqrt{3}}\mathcal{N}_\mu$ Note that $\boldsymbol{v}_i^{(0)} = \sqrt{d_i}\|\hat{\boldsymbol{x}}_{i0}\| = \sqrt{d_i}$ where $\hat{\boldsymbol{x}}_{i0}$ is the leading right singular vector of $\mathcal{A}_i^*(\boldsymbol{y})$. Therefore,

$$\|\boldsymbol{v}_i^{(0)}\| = \sqrt{d_i}\|\hat{\boldsymbol{x}}_{i0}\| = \sqrt{d_i} \leq \sqrt{(1+\xi)d_{i0}} \leq \frac{2}{\sqrt{3}}\sqrt{d_{i0}}, \quad \forall 1 \leq i \leq s$$

which implies $\{\boldsymbol{v}_i^{(0)}\}_{i=1}^s \in \frac{1}{\sqrt{3}}\mathcal{N}_d$.

Now we will prove that $\boldsymbol{u}_i^{(0)} \in \frac{1}{\sqrt{3}}\mathcal{N}_d \cap \frac{1}{\sqrt{3}}\mathcal{N}_\mu$ by Lemma 6.19. By Algorithm 1, $\boldsymbol{u}_i^{(0)}$ is the minimizer to the function $f(z) = \frac{1}{2}\|z - \sqrt{d_i}\hat{\boldsymbol{h}}_{i0}\|^2$ over $Q_i := \{z| \sqrt{L}\|\boldsymbol{B}z\|_\infty \leq 2\sqrt{d_i}\mu\}$. Obviously, by definition, $\boldsymbol{u}_i^{(0)}$ is the projection of $\sqrt{d_i}\hat{\boldsymbol{h}}_{i0}$ onto Q_i. Note that $\boldsymbol{u}_i^{(0)} \in Q_i$ implies $\sqrt{L}\|\boldsymbol{B}\boldsymbol{u}_i^{(0)}\|_\infty \leq 2\sqrt{d_i}\mu \leq 2\sqrt{(1+\xi)d_{i0}}\mu \leq \frac{4\sqrt{d_{i0}}\mu}{\sqrt{3}}$ and hence $\boldsymbol{u}_i^{(0)} \in \frac{1}{\sqrt{3}}\mathcal{N}_\mu$.

Moreover, due to (6.44), there holds

$$\|\sqrt{d_i}\hat{\boldsymbol{h}}_{i0} - \boldsymbol{w}\|^2 \geq \|\boldsymbol{u}_i^{(0)} - \boldsymbol{w}\|^2, \quad \forall \boldsymbol{w} \in Q_i. \tag{6.47}$$

In particular, let $\boldsymbol{w} = \boldsymbol{0} \in Q_i$ and immediately we have

$$\|\boldsymbol{u}_i^{(0)}\|^2 \leq d_i \leq \frac{4}{3} \Longrightarrow \boldsymbol{u}_i^{(0)} \in \frac{1}{\sqrt{3}}\mathcal{N}_\mu.$$

In other words, $\{(\boldsymbol{u}_i^{(0)}, \boldsymbol{v}_i^{(0)})\}_{i=1}^s \in \frac{1}{\sqrt{3}}\mathcal{N}_d \cap \frac{1}{\sqrt{3}}\mathcal{N}_\mu$.

Part II: proof of $(\boldsymbol{u}^{(0)}, \boldsymbol{v}^{(0)}) \in \mathcal{N}_{\frac{2\varepsilon}{5\sqrt{s\kappa}}}$ We will show $\|\boldsymbol{u}_i^{(0)}(\boldsymbol{v}_i^{(0)})^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F \leq 4\xi d_{i0}$ for $1 \leq i \leq s$ so that $\frac{\|\boldsymbol{u}_i^{(0)}(\boldsymbol{v}_i^{(0)})^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F}{d_{i0}} \leq \frac{2\varepsilon}{5\sqrt{s\kappa}}$.

First note that $\sigma_j(\mathcal{A}_i^*(\boldsymbol{y})) \leq \xi d_{i0}$ for all $j \geq 2$, which follows from Weyl's inequality [28] for singular values where $\sigma_j(\mathcal{A}_i^*(\boldsymbol{y}))$ denotes the jth largest singular value of $\mathcal{A}_i^*(\boldsymbol{y})$. Hence there holds

$$\|d_i\hat{\boldsymbol{h}}_{i0}\hat{\boldsymbol{x}}_{i0}^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\| \leq \|\mathcal{A}_i^*(\boldsymbol{y}) - d_i\hat{\boldsymbol{h}}_{i0}\hat{\boldsymbol{x}}_{i0}^*\| + \|\mathcal{A}_i^*(\boldsymbol{y}) - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\| \leq 2\xi d_{i0}. \tag{6.48}$$

On the other hand, for any i,

$$\begin{aligned}
\left\|\left(\boldsymbol{I}_K - \frac{\boldsymbol{h}_{i0}\boldsymbol{h}_{i0}^*}{d_{i0}}\right)\hat{\boldsymbol{h}}_{i0}\right\| &= \left\|\left(\boldsymbol{I}_K - \frac{\boldsymbol{h}_{i0}\boldsymbol{h}_{i0}^*}{d_{i0}}\right)\hat{\boldsymbol{h}}_{i0}\hat{\boldsymbol{x}}_{i0}^*\hat{\boldsymbol{x}}_{i0}\hat{\boldsymbol{h}}_{i0}^*\right\| \\
&= \left\|\left(\boldsymbol{I}_K - \frac{\boldsymbol{h}_{i0}\boldsymbol{h}_{i0}^*}{d_{i0}}\right)\left[\frac{1}{d_{i0}}((\mathcal{A}_i^*(\boldsymbol{y}) - d_i\hat{\boldsymbol{h}}_{i0}\hat{\boldsymbol{x}}_{i0}^*)) + \hat{\boldsymbol{h}}_{i0}\hat{\boldsymbol{x}}_{i0}^* - \frac{\boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*}{d_{i0}}\right]\hat{\boldsymbol{x}}_{i0}\hat{\boldsymbol{h}}_{i0}^*\right\| \\
&= \frac{1}{d_{i0}}\|\mathcal{A}_i^*(\boldsymbol{y}) - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\| + \left|\frac{d_i}{d_{i0}} - 1\right| \leq 2\xi,
\end{aligned}$$

where $\left(\boldsymbol{I}_K - \frac{\boldsymbol{h}_{i0}\boldsymbol{h}_{i0}^*}{d_{i0}}\right)\boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^* = \boldsymbol{0}$ and $(\mathcal{A}_i^*(\boldsymbol{y}) - d_i\hat{\boldsymbol{h}}_{i0}\hat{\boldsymbol{x}}_{i0}^*)\hat{\boldsymbol{x}}_{i0}\hat{\boldsymbol{h}}_{i0}^* = \boldsymbol{0}$. Therefore, we have

$$\left\|\hat{\boldsymbol{h}}_{i0} - \frac{\boldsymbol{h}_{i0}^*\hat{\boldsymbol{h}}_{i0}}{d_{i0}}\boldsymbol{h}_{i0}\right\| \leq 2\xi, \quad \|\sqrt{d_i}\hat{\boldsymbol{h}}_{i0} - t_{i0}\boldsymbol{h}_{i0}\| \leq 2\sqrt{d_i}\xi, \tag{6.49}$$

where $t_{i0} = \frac{\sqrt{d_i}\boldsymbol{h}_{i0}^*\hat{\boldsymbol{h}}_{i0}}{d_{i0}}$ and $|t_{i0}| \leq \sqrt{d_i/d_{i0}} < \sqrt{2}$. If we substitute $\boldsymbol{w}$ by $t_{i0}\boldsymbol{h}_{i0} \in Q_i$ into (6.47),

$$\|\sqrt{d_i}\hat{\boldsymbol{h}}_{i0} - t_{i0}\boldsymbol{h}_{i0}\| \geq \|\boldsymbol{u}_i^{(0)} - t_{i0}\boldsymbol{h}_{i0}\|, \tag{6.50}$$

where $t_{i0}\boldsymbol{h}_{i0} \in Q_i$ follows from $\sqrt{L}|t_{i0}|\|\boldsymbol{B}\boldsymbol{h}_{i0}\|_\infty \leq |t_{i0}|\sqrt{d_{i0}}\mu_h \leq \sqrt{2d_{i0}}\mu$.

Now we are ready to estimate $\|\boldsymbol{u}_i^{(0)}(\boldsymbol{v}_i^{(0)})^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F$ as follows,

$$\|\boldsymbol{u}_i^{(0)}(\boldsymbol{v}_i^{(0)})^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F \le \|\boldsymbol{u}_i^{(0)}(\boldsymbol{v}_i^{(0)})^* - t_{i0}\boldsymbol{h}_{i0}(\boldsymbol{v}_i^{(0)})^*\|_F + \|t_{i0}\boldsymbol{h}_{i0}(\boldsymbol{v}_i^{(0)})^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F$$

$$\le \underbrace{\|\boldsymbol{u}_i^{(0)} - t_{i0}\boldsymbol{h}_{i0}\|\|\boldsymbol{v}_i^{(0)}\|}_{I_1} + \underbrace{\left\|\frac{d_i}{d_{i0}}\boldsymbol{h}_{i0}\boldsymbol{h}_{i0}^*\hat{\boldsymbol{h}}_{i0}\hat{\boldsymbol{x}}_{i0}^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\right\|_F}_{I_2}.$$

Here $I_1 \le 2\xi d_i$ because $\|\boldsymbol{v}_i^{(0)}\| = \sqrt{d_i}$ and $\|\boldsymbol{u}_i^{(0)} - t_{i0}\boldsymbol{h}_{i0}\| \le 2\sqrt{d_i}\xi$ follows from (6.49) and (6.50). For I_2, there holds

$$I_2 = \left\|\frac{\boldsymbol{h}_{i0}\boldsymbol{h}_{i0}^*}{d_{i0}}\left(d_i\hat{\boldsymbol{h}}_{i0}\hat{\boldsymbol{x}}_{i0}^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\right)\right\|_F \le \|d_i\hat{\boldsymbol{h}}_{i0}\hat{\boldsymbol{x}}_{i0}^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F \le 2\sqrt{2}\xi d_{i0},$$

which follows from (6.48). Therefore,

$$\|\boldsymbol{u}_i^{(0)}(\boldsymbol{v}_i^{(0)})^* - \boldsymbol{h}_{i0}\boldsymbol{x}_{i0}^*\|_F \le 2\xi d_i + 2\sqrt{2}\xi d_{i0} \le 5\xi d_{i0} \le \frac{2\varepsilon d_{i0}}{5\sqrt{s\kappa}}.$$

$\square$

Acknowledgments

References

1. AHMED, A., RECHT, B. & ROMBERG, J. (2014) Blind deconvolution using convex programming. *IEEE Trans. Inf. Theory*, **60**, 1711–1732.
2. BRISTOW, H., ERIKSSON, A. & LUCEY, S. (2013) Fast convolutional sparse coding. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* pp. 391–398.
3. CAI, T. T., LI, X. & MA, Z. (2016) Optimal rates of convergence for noisy sparse phase retrieval via thresholded wirtinger flow. *Ann. Stat.*, **44**, 2221–2251.
4. CAMBARERI, V. & JACQUES, L. (2016) Through the haze: a non-convex approach to blind calibration for linear random sensing models. arXiv preprint arXiv: 1610.09028.
5. CAMPISI, P. & EGIAZARIAN, K. (2007) *Blind Image Deconvolution: Theory and Applications*. Boca Raton, FL: CRC press.
6. CANDÈS, E., ELDAR, Y., STROHMER, T. & VORONINSKI, V. (2013) Phase retrieval via matrix completion. *SIAM J. Imaging Sci.*, **6**, 199–225.
7. CANDÈS, E. & RECHT, B. (2009) Exact matrix completion via convex optimization. *Foundations Comput. Mathematics*, **9**, 717–772.
8. CANDES, E. J., LI, X. & SOLTANOLKOTABI, M. (2015) Phase retrieval via Wirtinger flow: theory and algorithms. *IEEE Trans. Inf. Theory*, **61**, 1985–2007.
9. CANDÈS, E. J., STROHMER, T. & VORONINSKI, V. (2013) Phaselift: exact and stable signal recovery from magnitude measurements via convex programming. *Commun. Pure Appl. Mathematics*, **66**, 1241–1274.

10. CHANDRASEKARAN, V., RECHT, B., PARRILO, P. & WILLSKY, A. (2012) The convex geometry of linear inverse problems. *Foundations Comput. Mathematics*, **12**, 805–849.

11. CHEN, Y. & CANDES, E. (2015) Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Communications on pure and applied mathematics*, **70**, 822–883.

12. EFTEKHARI, A. & WAKIN, M. B. (2015) Greed is super: a fast algorithm for super-resolution. arXiv preprint arXiv: 1511.03385.

13. ESCALANTE, R. & RAYDAN, M. (2011) *Alternating Projection Methods,* vol 8. SIAM.

14. GOLDSMITH, A. (2005) *Wireless Communications*. Cambridge University Press.

15. JUNG, P., KRAHMER, F. & STÖGER, D. (2017) Blind demixing and deconvolution at near-optimal rate. arXiv preprint arXiv: 1704.04178.

16. KESHAVAN, R., MONTANARI, A., MATRIX, S. & OH, S. (2009) Completion from noisy entries. *Advances in Neural Information Processing Systems*, pp.952–960.

17. KESHAVAN, R. H., MONTANARI, A. & OH, S. (2010) Matrix completion from a few entries. *IEEE Trans. Inf. Theory*, **56**, 2980–2998.

18. KOLTCHINSKII, V. (2011) Von Neumann entropy penalization and low-rank matrix estimation. *Ann. Stat*., **39**, 2936–2973.

19. LEE, K., LI, Y., JUNGE, M. & BRESLER, Y. (2017) Blind recovery of sparse signals from subsampled convolution. *IEEE Trans. Inf. Theory*, **63**, 802–821.

20. LI, X. & FAN, H. H. (2001) Direct blind multiuser detection for CDMA in multipath without channel estimation. *IEEE Trans. Signal Process*., **49**, 63–73.

21. LI, X., LING, S., STROHMER, T. & WEI, K. (2016) Rapid, robust, and reliable blind deconvolution via nonconvex optimization. arXiv preprint arXiv: 1606.04933.

22. LING, S. & STROHMER, T. (2015) Self-calibration and biconvex compressive sensing. *Inverse Probl*., **31**, 115002.

23. LING, S. & STROHMER, T. (2017) Blind deconvolution meets blind demixing: algorithms and performance bounds. *IEEE Trans. Inf. Theory*, **63**, 4497–4520.

24. LIU, J., XIN, J., QI, Y. & ZHENG, F.-G. (2009) A time domain algorithm for blind separation of convolutive sound mixtures and L_1 constrained minimization of cross correlations. *Commun. Math. Sci*., **7**, 109–128.

25. LUENBERGER, D. G. & YE, Y. *Linear and Nonlinear Programming*, vol. 228. US: Springer, 2015.

26. MCCOY, M. B. & TROPP, J. A. (2017) Achievable performance of convex demixing. Technical report, Caltech, Paper dated Feb. 2013. *ACM Technical Report* 2017–02.

27. SHAFI, M., MOLISCH, A. F., SMITH, P. J., HAUSTEIN, T., ZHU, P., DE SILVA, P., TUFVESSON, F., BENJEBBOUR, A. & WUNDER, G. (2017) 5G: a tutorial overview of standards, trials, challenges, deployment, and practice. *IEEE J. Selected Areas Commun*., **35**, 1201–1221.

28. STEWART, G. W. (1990) Perturbation theory for the singular value decomposition. *Technical Report CS-TR* 2539, University of Maryland.

29. STÖGER, D., JUNG, P. & KRAHMER, F. (2016) Blind deconvolution and compressed sensing. *Compressed Sensing Theory and its Applications to Radar, Sonar and Remote Sensing (CoSeRa), 2016 4th International Workshop on*, pp. 24–27. IEEE.

30. STROHMER, T. (2002) Four short stories about Toeplitz matrix calculations. *Linear Algebra Appl.*, 343/344: 321–344. *Special issue on structured and infinite systems of linear equations.*

31. SUDHAKAR, P., ARBERET, S. & GRIBONVAL, R. (2010) Double sparsity: towards blind estimation of multiple channels. *Latent Variable Analysis and Signal Separation*, pp. 571–578. Berlin Heidelberg: Springer.

32. SUN, J., QU, Q. & WRIGHT, J. (2017) Complete dictionary recovery over the sphere I: overview and the geometric picture. *IEEE Trans. Inf. Theory*, **63**, 853–884.

33. SUN, J., QU, Q. & WRIGHT, J. (2017) A geometric analysis of phase retrieval. *Foundation of Computational Mathematics*. Doi: 10.1007/s10208-017-9365-9, Issn: 1615-3383.

34. SUN, R. & LUO, Z.-Q. (2016) Guaranteed matrix completion via non-convex factorization. *IEEE Trans. Inf. Theory*, **62**, 6535–6579.

35. TU, S., BOCZAR, R., SIMCHOWITZ, M., SOLTANOLKOTABI, M. & RECHT, B. (2016) Low-rank solutions of linear matrix equations via procrustes flow. *Proceedings of The 33rd International Conference on Machine Learning*, pp. 964–973.

36. VANNITHAMBY, R. & TALWAR, S. (2017) *Towards 5G: Applications,Requirements and Candidate Technologies*. United Kingdom: John Wiley & Sons.

37. VERDU, S. (1998) *Multiuser Detection*. Cambridge University Press.

38. VERSHYNIN, R. (2012) Introduction to the non-asymptotic analysis of random matrices. *Compressed Sensing: Theory and Applications* (Y. C. Eldar & G. Kutyniok eds), chapter 5. Cambridge University Press.

39. WANG, X. & POOR, H. V. (1998) Blind equalization and multiuser detection in dispersive CDMA channels. *IEEE Trans. Commun.*, **46**, 91–103.

40. WEDIN, P.-Å. (1972) Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, **12**, 99–111.

41. WEI, K., CAI, J.-F., CHAN, T. F. & LEUNG, S. (2016) Guarantees of riemannian optimization for low rank matrix recovery. *SIAM J. Matrix Anal. Appl.*, **37**, 1198–1222.

42. WRIGHT, J., GANESH, A., MIN, K. & MA, Y. (2013) Compressive principal component pursuit. *Inf. Inference*, **2**, 32–68.

43. WUNDER, G., BOCHE, H., STROHMER, T. & JUNG, P. (2015) Sparse signal processing concepts for efficient 5G system design. *IEEE Access.*, **3**, 195–208.

Appendix

Descent Lemma

LEMMA 6.20 (Lemma 6.1 in [21]).   If $f(z,\bar{z})$ is a continuously differentiable real-valued function with two complex variables z and $\bar{z}$, (for simplicity, we just denote $f(z,\bar{z})$ by $f(z)$ and keep in the mind that $f(z)$ only assumes real values) for $z := (\boldsymbol{h}, \boldsymbol{x}) \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$. Suppose that there exists a constant C_L such that

$$\|\nabla f(z + t\Delta z) - \nabla f(z)\| \le C_L t \|\Delta z\|, \quad \forall 0 \le t \le 1,$$

for all $z \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ and Δz such that $z + t\Delta z \in \mathcal{N}_\epsilon \cap \mathcal{N}_{\widetilde{F}}$ and $0 \le t \le 1$. Then

$$f(z + \Delta z) \le f(z) + 2\operatorname{Re}((\Delta z)^T \overline{\nabla} f(z)) + C_L \|\Delta z\|^2,$$

where $\overline{\nabla} f(z) := \frac{\partial f(z,\bar{z})}{\partial z}$ is the complex conjugate of $\nabla f(z) = \frac{\partial f(z,\bar{z})}{\partial \bar{z}}$.

Concentration inequality

We define the matrix ψ_1-norm via

$$\|\boldsymbol{Z}\|_{\psi_1} := \inf_{u \ge 0}\{\mathbb{E}[\exp(\|\boldsymbol{Z}\|/u)] \le 2\}.$$

THEOREM 6.21 [18] Consider a finite sequence of $\mathcal{Z}_l$ of independent centered random matrices with dimension $M_1 \times M_2$. Assume that $R := \max_{1 \leq l \leq L} \|\mathcal{Z}_l\|_{\psi_1}$ and introduce the random matrix

$$\mathcal{S} = \sum_{l=1}^{L} \mathcal{Z}_l. \tag{A.1}$$

Compute the variance parameter

$$\sigma_0^2 = \max \left\{ \left\| \sum_{l=1}^{L} \mathbb{E}(\mathcal{Z}_l \mathcal{Z}_l^*) \right\|, \left\| \sum_{l=1}^{L} \mathbb{E}(\mathcal{Z}_l^* \mathcal{Z}_l) \right\| \right\}, \tag{A.2}$$

then for all $t \geq 0$

$$\|\mathcal{S}\| \leq C_0 \max \left\{ \sigma_0 \sqrt{t + \log(M_1 + M_2)}, R \log \left(\frac{\sqrt{L}R}{\sigma_0} \right) (t + \log(M_1 + M_2)) \right\} \tag{A.3}$$

with probability at least $1 - e^{-t}$ where C_0 is an absolute constant.

LEMMA 6.22 (Lemma 10–13 in [1], Lemma 12.4 in [23]). Let $\boldsymbol{u} \in \mathbb{C}^n \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_n)$, then $\|\boldsymbol{u}\|^2 \sim \frac{1}{2}\chi_{2n}^2$ and

$$\| \|\boldsymbol{u}\|^2 \|_{\psi_1} = \| \langle \boldsymbol{u}, \boldsymbol{u} \rangle \|_{\psi_1} \leq Cn \tag{A.4}$$

and

$$\mathbb{E}(\boldsymbol{u}\boldsymbol{u}^* - \boldsymbol{I}_n)^2 = n\boldsymbol{I}_n. \tag{A.5}$$

Let $\boldsymbol{q} \in \mathbb{C}^n$ be any deterministic vector, then the following properties hold

$$\|(\boldsymbol{u}\boldsymbol{u}^* - \boldsymbol{I})\boldsymbol{q}\|_{\psi_1} \leq C\sqrt{n}\|\boldsymbol{q}\|, \tag{A.6}$$

$$\mathbb{E}[(\boldsymbol{u}\boldsymbol{u}^* - \boldsymbol{I})\boldsymbol{q}\boldsymbol{q}^*(\boldsymbol{u}\boldsymbol{u}^* - \boldsymbol{I})] = \|\boldsymbol{q}\|^2\boldsymbol{I}_n. \tag{A.7}$$

Let $\boldsymbol{v} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{I}_m)$ be a complex Gaussian random vector in $\mathbb{C}^m$, independent of $\boldsymbol{u}$, then

$$\| \|\boldsymbol{u}\| \cdot \|\boldsymbol{v}\| \|_{\psi_1} \leq C\sqrt{mn}. \tag{A.8}$$