Beyond Procrustes: Balancing-Free Gradient Descent for Asymmetric Low-Rank Matrix Sensing

Cong Ma, Yuanxin Li, and Yuejie Chi

***Abstract*—Low-rank matrix estimation plays a central role in various applications across science and engineering. Recently, nonconvex formulations based on matrix factorization are provably solved by simple gradient descent algorithms with strong computational and statistical guarantees. However, when the low-rank matrices are asymmetric, existing approaches rely on adding a regularization term to balance the scale of the two matrix factors which in practice can be removed safely without hurting the performance when initialized via the spectral method. In this paper, we provide a theoretical justification to this for the matrix sensing problem, which aims to recover a low-rank matrix from a small number of linear measurements. As long as the measurement ensemble satisfies the restricted isometry property, gradient descent—in conjunction with spectral initialization—converges linearly without the need of explicitly promoting balancedness of the factors; in fact, the factors stay balanced automatically throughout the execution of the algorithm. Our analysis is based on analyzing the evolution of a new distance metric that directly accounts for the ambiguity due to invertible transforms, and might be of independent interest.**

***Index Terms*—asymmetric low-rank matrix sensing, nonconvex optimization, gradient descent**

I. Introduction

Low-rank matrix estimation plays a central role in many applications [2], [3], [4]. Broadly speaking, we are interested in estimating a rank-r matrix $\boldsymbol{M}_\star = \boldsymbol{X}_\star \boldsymbol{Y}_\star^\top \in \mathbb{R}^{n_1 \times n_2}$ by solving a rank-constrained optimization problem:

$$\min_{\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}} \mathcal{L}(\boldsymbol{M}) \quad \text{subject to} \quad \text{rank}(\boldsymbol{M}) \le r, \tag{1}$$

where $\mathcal{L}(\cdot)$ denotes a certain loss function with the rank r typically much smaller than the dimension of the matrix. To reduce computational complexity, a common approach, popularized by the work of Burer and Monteiro [5], [6], [7], is to factorize $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{Y}^\top$ with $\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}$, and rewrite the above problem (1) into an unconstrained nonconvex optimization problem:

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}} f(\boldsymbol{X}, \boldsymbol{Y}) \triangleq \mathcal{L}(\boldsymbol{X}\boldsymbol{Y}^\top). \tag{2}$$

C. Ma is with Department of Electrical Engineering and Computer Science, UC Berkeley, Berkeley, CA 94720, USA; Email: congm@berkeley.edu.

Y. Li was with Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

Y. Chi is with Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213, USA; Email: yuejiechi@cmu.edu.

This work is supported in part by ONR under the grants N00014-18-1-2142 and N00014-19-1-2404, by ARO under the grant W911NF-18-1-0303, and by NSF under the grants CAREER ECCS-1818571, CCF-1806154 and CCF-1901199. A preliminary version of this paper was presented at the 2019 Asilomar Conference on Signals, Systems, and Computers [1].

Despite nonconvexity, one might be tempted to estimate the low-rank factors $(\boldsymbol{X}, \boldsymbol{Y})$ via gradient descent, which proceeds via the following update rule

$$\begin{bmatrix} \boldsymbol{X}_{t+1} \\ \boldsymbol{Y}_{t+1} \end{bmatrix} = \begin{bmatrix} \boldsymbol{X}_t \\ \boldsymbol{Y}_t \end{bmatrix} - \eta_t \begin{bmatrix} \nabla_{\boldsymbol{X}} f(\boldsymbol{X}_t, \boldsymbol{Y}_t) \\ \nabla_{\boldsymbol{Y}} f(\boldsymbol{X}_t, \boldsymbol{Y}_t) \end{bmatrix} \tag{3}$$

from $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ some proper initialization. Here, η_t is the step size, $\nabla_{\boldsymbol{X}} f$ and $\nabla_{\boldsymbol{Y}} f$ are the gradients of f w.r.t. $\boldsymbol{X}$ and $\boldsymbol{Y}$, respectively.

Significant progress has been made recently in understanding the performance of gradient descent for nonconvex matrix estimation. Somewhat surprisingly, most of the existing guarantees are not directly applicable to the vanilla gradient descent rule (3). One particular challenge is associated with the identifiability of the factors $(\boldsymbol{X}, \boldsymbol{Y})$—they are indistinguishable as long as their product $\boldsymbol{X}\boldsymbol{Y}^\top$ is the same. What is worse, if the norms of the factors become highly unbalanced, gradient descent might diverge easily. Consequently, it becomes a routine procedure to insert a regularizer $g(\boldsymbol{X}, \boldsymbol{Y})$ that balances the two factors [8], [9], [10]:

$$g(\boldsymbol{X}, \boldsymbol{Y}) \triangleq \lambda \|\boldsymbol{X}^\top \boldsymbol{X} - \boldsymbol{Y}^\top \boldsymbol{Y}\|_{\mathrm{F}}^2, \tag{4}$$

where $\lambda > 0$ is some regularization parameter, and apply gradient descent to the regularized loss function instead:

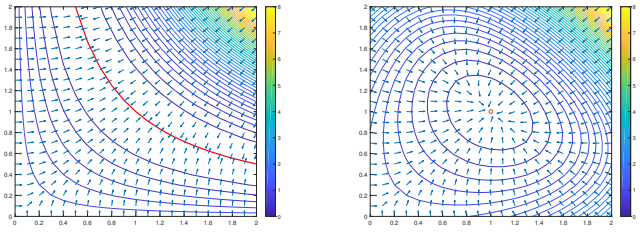
$$\min_{\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}, \boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}} f_{\mathrm{reg}}(\boldsymbol{X}, \boldsymbol{Y}) \triangleq f(\boldsymbol{X}, \boldsymbol{Y}) + g(\boldsymbol{X}, \boldsymbol{Y}). \tag{5}$$

For a variety of important problems such as low-rank matrix sensing and matrix completion, it has been established that gradient descent over the regularized loss function, when properly initialized, achieves compelling statistical and computational guarantees.

A. Why balancing is needed in prior work?

Before we investigate the possibility of a balancing-free procedure (i.e. vanilla gradient descent as in (3)), let us first explain using a heuristic argument why balancing is needed in the prior literature.

To handle the asymmetric factorization, it is common to stack the two factors into one augmented factor $\boldsymbol{Z}_\star \triangleq \begin{bmatrix} \boldsymbol{X}_\star \\ \boldsymbol{Y}_\star \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$ and then seek to estimate $\boldsymbol{Z}_\star$ directly, by rewriting the loss function with respect to the lifted low-rank matrix: $\boldsymbol{Z}_\star \boldsymbol{Z}_\star^\top = \begin{bmatrix} \boldsymbol{X}_\star \boldsymbol{X}_\star^\top & \boldsymbol{X}_\star \boldsymbol{Y}_\star^\top \\ \boldsymbol{Y}_\star \boldsymbol{X}_\star^\top & \boldsymbol{Y}_\star \boldsymbol{Y}_\star^\top \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$. It is obvious that the loss function originally with respect to the asymmetric matrix $\boldsymbol{X}_\star \boldsymbol{Y}_\star^\top$ only constrains the off-diagonal blocks of $\boldsymbol{Z}_\star \boldsymbol{Z}_\star^\top$ and not the diagonal ones; correspondingly,



(a) unregularized loss f (b) regularized loss f_{reg}

Fig. 1. The geometry for the scalar case $f(x,y)=(xy-1)^2$ and $g(x,y)=(x^2-y^2)^2/8$. The regularized loss function is locally strongly convex while the unregularized one is nonconvex; in particular, the Hessian of the unregularized loss function is rank deficient on the ambiguity set $xy=1$ (colored in red).

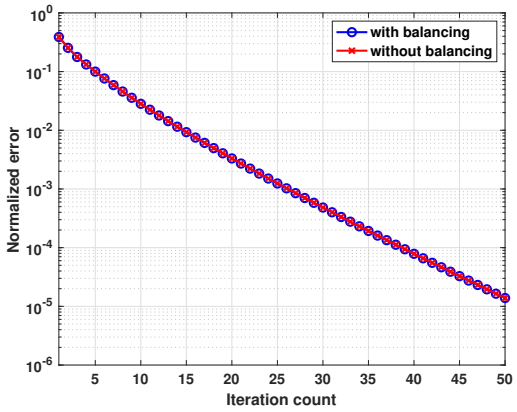


Fig. 2. The normalized reconstruction error $\|\boldsymbol{X}_t\boldsymbol{Y}_t^\top-\boldsymbol{M}_\star\|_{\mathrm{F}}/\|\boldsymbol{M}_\star\|_{\mathrm{F}}$ with respect to the iteration count, for completing a 1000×1000 matrix of rank-10 when each entry is observed independently with probability $p=0.15$. The balancing regularizer is set as $g(\boldsymbol{X},\boldsymbol{Y})=\frac{1}{64}\|\boldsymbol{X}^\top\boldsymbol{X}-\boldsymbol{Y}^\top\boldsymbol{Y}\|_{\mathrm{F}}^2$ following the suggestion in [12].

the loss function is not (restricted) strongly convex with respect to the augmented factor, unless we appropriately regularize the diagonal blocks. This gives rise to the adoption of the regularization term in (4).

To develop more intuitions regarding why this regularization term (4) may help analysis, consider a toy example of factorizing a rank-one matrix $\boldsymbol{x}_\star\boldsymbol{y}_\star^\top$, where $f(\boldsymbol{x},\boldsymbol{y})$ and $g(\boldsymbol{x},\boldsymbol{y})$ respectively are $f(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}\boldsymbol{y}^\top-\boldsymbol{x}_\star\boldsymbol{y}_\star^\top\|_{\mathrm{F}}^2/2$ and $g(\boldsymbol{x},\boldsymbol{y})=(\|\boldsymbol{x}\|_2^2-\|\boldsymbol{y}\|_2^2)^2/8$. Figure 1 illustrates the landscape of the unregularized loss function $f(x,y)$ and the regularized loss function $f_{\mathrm{reg}}(x,y)$, respectively, when the arguments are scalar-valued, i.e. $n_1=n_2=1$. One can clearly appreciate the value of the regularizer: $f_{\mathrm{reg}}(x,y)$ becomes strongly convex in the local neighborhood around the global optimum $(1,1)$. In contrast, the Hessian of the unregularized loss function $f_{\mathrm{reg}}(x,y)$ remains rank deficient along the ambiguity set $\{(x,y)\,|\,xy=1\}$, making the analysis less tractable.

B. This paper: balancing-free procedure?

The goal of this paper is to understand the effectiveness of vanilla gradient descent (3) when initialized with balanced factors. Indeed, Figure 2 plots the normalized error $\|\boldsymbol{X}_t\boldsymbol{Y}_t^\top-\boldsymbol{M}_\star\|_{\mathrm{F}}/\|\boldsymbol{M}_\star\|_{\mathrm{F}}$ for low-rank matrix completion, which aims to recover a low-rank matrix from a subset of its observations [11], with respect to the iteration count, using either a regularized loss function or an unregularized loss function when initialized by the spectral method. The two sequences of iterates converge in almost exactly the same trajectory, suggesting that gradient descent over the unregularized loss function converges almost in the same manner as its regularized counterpart, and perhaps is more natural to use in practice since it eliminates the tuning of the regularization parameters.

This paper justifies formally that even without explicit balancing in asymmetric low-rank matrix sensing, gradient descent converges linearly towards the global optimum, as long as the initialization is (nearly) balanced and close to the optimum. As will be detailed later, our analysis is simple and built on a novel distance metric that directly accounts for the ambiguity due to invertible transformations—in contrast, the ambiguity set reduces to orthonormal transforms when the balancing regularization is present. Our key message is this:

As long as the factors are (nearly) balanced in a basin of attraction at the initialization, they will stay approximately balanced throughout the trajectory of gradient descent, and therefore no additional regularization is necessary.

C. Notation

We use boldface lowercase (resp. uppercase) letters to represent vectors (resp. matrices). We denote by $\|\boldsymbol{x}\|_2$ the ℓ_2 norm of a vector $\boldsymbol{x}$, and $\boldsymbol{X}^\top$, $\boldsymbol{X}^{-1}$, $\|\boldsymbol{X}\|$ and $\|\boldsymbol{X}\|_{\mathrm{F}}$ the transpose, the inverse, the spectral norm and the Frobenius norm of a matrix $\boldsymbol{X}$, respectively. Furthermore, we denote $\boldsymbol{X}^{-\top}=(\boldsymbol{X}^{-1})^\top=(\boldsymbol{X}^\top)^{-1}$ for an invertible matrix $\boldsymbol{X}$. The kth largest singular value of a matrix $\boldsymbol{X}$ is denoted by $\sigma_k(\boldsymbol{X})$. The inner product between two matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ is defined as $\langle\boldsymbol{X},\boldsymbol{Y}\rangle=\mathrm{Tr}\left(\boldsymbol{Y}^\top\boldsymbol{X}\right)$, where $\mathrm{Tr}\left(\cdot\right)$ denotes the trace operator. Denote by $\mathcal{O}^{r\times r}$ the set of $r\times r$ orthonormal matrices. In addition, we use c and C with different subscripts to represent positive numerical constants, whose values may change from line to line.

II. Main Results

Let the object of interest $\boldsymbol{M}_\star\in\mathbb{R}^{n_1\times n_2}$ be a rank-r matrix whose compact Singular Value Decomposition (SVD) is given by

$$\boldsymbol{M}_\star=\boldsymbol{U}_\star\boldsymbol{\Sigma}_\star\boldsymbol{V}_\star^\top,$$

where $\boldsymbol{U}_\star\in\mathbb{R}^{n_1\times r}$, $\boldsymbol{V}_\star\in\mathbb{R}^{n_2\times r}$ and $\boldsymbol{\Sigma}_\star\in\mathbb{R}^{r\times r}$ correspond to the left singular vectors, the right singular vectors and the singular values, respectively. Without loss of generality, we denote the ground truth factors as

$$\boldsymbol{X}_\star\triangleq\boldsymbol{U}_\star\boldsymbol{\Sigma}_\star^{1/2},\qquad\text{and}\qquad\boldsymbol{Y}_\star\triangleq\boldsymbol{V}_\star\boldsymbol{\Sigma}_\star^{1/2}.\tag{6}$$

Let $\sigma_{\max}\triangleq\sigma_1(\boldsymbol{M}_\star)$ (resp. $\sigma_{\min}\triangleq\sigma_r(\boldsymbol{M}_\star)$) be the largest (resp. smallest) nonzero singular value of $\boldsymbol{M}_\star$. The condition number of $\boldsymbol{M}_\star$ is therefore defined as $\kappa\triangleq\sigma_{\max}/\sigma_{\min}$.

Since the factors are identifiable up to invertible transforms, i.e. $(\boldsymbol{X}_\star\boldsymbol{P})(\boldsymbol{Y}_\star\boldsymbol{P}^{-\top})^\top=\boldsymbol{X}_\star\boldsymbol{Y}_\star^\top$ for any invertible matrix $\boldsymbol{P}\in\mathbb{R}^{r\times r}$, it is natural to measure the distance between

Algorithm 1 Gradient Descent with Spectral Initialization (unregularized Procrustes Flow)

Input: Measurements $\boldsymbol{y} = \{y_i\}_{1\le i\le m}$, and sensing matrices $\{\boldsymbol{A}_i\}_{1\le i\le m}$.
Parameters: Step size η_t, rank r, and number of iterations T.
Initialization: Initialize $\boldsymbol{X}_0 = \boldsymbol{U}\boldsymbol{\Sigma}^{1/2}$ and $\boldsymbol{Y}_0 = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$, where $\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the rank-r SVD of the surrogate matrix $\boldsymbol{K} = \frac{1}{m}\mathcal{A}^*(\boldsymbol{y}) = \frac{1}{m}\sum_{i=1}^m y_i\boldsymbol{A}_i$.
Gradient loop: For $t = 0, 1, \cdots, T-1$, do

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \frac{\eta_t}{\|\boldsymbol{Y}_0\|^2}\cdot\left[\sum_{i=1}^m\left(\langle\boldsymbol{A}_i, \boldsymbol{X}_t\boldsymbol{Y}_t^\top\rangle - y_i\right)\boldsymbol{A}_i\boldsymbol{Y}_t\right]; \tag{9a}$$

$$\boldsymbol{Y}_{t+1} = \boldsymbol{Y}_t - \frac{\eta_t}{\|\boldsymbol{X}_0\|^2}\cdot\left[\sum_{i=1}^m\left(\langle\boldsymbol{A}_i, \boldsymbol{X}_t\boldsymbol{Y}_t^\top\rangle - y_i\right)\boldsymbol{A}_i^\top\boldsymbol{X}_t\right]. \tag{9b}$$

Output: $\boldsymbol{X}_T$ and $\boldsymbol{Y}_T$.

two pairs of factors $\boldsymbol{Z} = \begin{bmatrix}\boldsymbol{X}\\ \boldsymbol{Y}\end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$ and $\boldsymbol{Z}_\star = \begin{bmatrix}\boldsymbol{X}_\star\\ \boldsymbol{Y}_\star\end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$ via the following function:[1]

$$\operatorname{dist}(\boldsymbol{Z}, \boldsymbol{Z}_\star) = \min_{\substack{\boldsymbol{P}\in\mathbb{R}^{r\times r}\\ \text{invertible}}}\sqrt{\|\boldsymbol{X}\boldsymbol{P} - \boldsymbol{X}_\star\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\boldsymbol{P}^{-\top} - \boldsymbol{Y}_\star\|_{\mathrm{F}}^2}.$$

A. Low-rank matrix sensing

Low-rank matrix sensing refers to the problem of recovering a low-rank matrix (i.e. $\boldsymbol{M}_\star$) from a small number of linear measurements. Specifically, we are given a set of m measurements as follows

$$y_i = \langle\boldsymbol{A}_i, \boldsymbol{M}_\star\rangle = \langle\boldsymbol{A}_i, \boldsymbol{X}_\star\boldsymbol{Y}_\star^\top\rangle, \qquad i = 1, \cdots, m, \tag{7}$$

where $\boldsymbol{A}_i \in \mathbb{R}^{n_1\times n_2}$ is the ith sensing matrix. For convenience, we define $\mathcal{A} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ as an affine transformation from $\mathbb{R}^{n_1\times n_2}$ to $\mathbb{R}^m$, such that $\mathcal{A}(\boldsymbol{M}) = \{\langle\boldsymbol{A}_i, \boldsymbol{M}\rangle\}_{1\le i\le m}$. Consequently, one can compactly write (7) as $\boldsymbol{y} = \mathcal{A}(\boldsymbol{M}_\star)$. The adjoint operator $\mathcal{A}^* : \mathbb{R}^m \to \mathbb{R}^{n_1\times n_2}$ is defined as $\mathcal{A}^*(\boldsymbol{y}) = \sum_{i=1}^m y_i\boldsymbol{A}_i$.

To recover the low-rank matrix, a natural choice is to minimize the least-squares loss function

$$f(\boldsymbol{X}, \boldsymbol{Y}) \triangleq \frac{1}{2}\left\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X}\boldsymbol{Y}^\top)\right\|_2^2. \tag{8}$$

Algorithm 1 describes the gradient descent algorithm initialized by the spectral method [13] for minimizing (8). Compared to the Procrustes Flow (PF) algorithm in [8], which minimizes the regularized loss function in (5), the new algorithm does not include the balancing regularizer $g(\boldsymbol{X}, \boldsymbol{Y})$.

[1]More rigorously, we should write inf instead of min in the definition of $\operatorname{dist}(\cdot,\cdot)$. However, as we will soon see, in the cases we care about, the minimum can always be achieved by some invertible matrix $\boldsymbol{P}$.

B. Theoretical guarantee for local linear convergence

To understand the performance of Algorithm 1, we adopt a standard assumption on the sensing operator $\mathcal{A}$, namely the Restricted Isometry Property (RIP).

Definition 1 (RIP)**.** *The operator $\mathcal{A}(\cdot)$ is said to satisfy the rank-r RIP with a constant $\delta_r \in [0, 1)$, if*

$$(1-\delta_r)\|\boldsymbol{M}\|_{\mathrm{F}}^2 \le \|\mathcal{A}(\boldsymbol{M})\|_2^2 \le (1+\delta_r)\|\boldsymbol{M}\|_{\mathrm{F}}^2$$

holds for all matrices $\boldsymbol{M} \in \mathbb{R}^{n_1\times n_2}$ of rank at most r.

It is well-known that many measurement ensembles satisfy the RIP property [14]. For example, under the Gaussian design where the entries of $\boldsymbol{A}_i$'s are composed of i.i.d. Gaussian entries $\mathcal{N}(0, 1/m)$, the RIP is satisfied as long as m is on the order of $(n_1+n_2)r/\delta_r^2$.

Armed with the RIP, we have the following theoretical guarantee for the local convergence of Algorithm 1.

Theorem 1. *Suppose that $\mathcal{A}(\cdot)$ satisfies the RIP with $\delta_{2r} \le c$ for some sufficiently small constant c. Let $\boldsymbol{Z}_0 \triangleq \begin{bmatrix}\boldsymbol{X}_0\\ \boldsymbol{Y}_0\end{bmatrix}$ be any initialization point that satisfies*

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\|\boldsymbol{Z}_0\boldsymbol{R} - \boldsymbol{Z}_\star\|_{\mathrm{F}} \le c_0\frac{1}{\kappa^{3/2}}\sigma_{\min}^{1/2}, \tag{10}$$

for some small enough constant $c_0 > 0$. Then there exist some constant $c_1 > 0$ such that at as long as $\eta_t = \eta = c_1$, the iterates of unregularized gradient descent (cf. (9)) satisfy

$$\operatorname{dist}(\boldsymbol{Z}_t, \boldsymbol{Z}_\star) \le \left(1 - \frac{\eta}{50\kappa}\right)^t\operatorname{dist}(\boldsymbol{Z}_0, \boldsymbol{Z}_\star).$$

In words, Theorem 1 reveals that if the initialization $\boldsymbol{Z}_0$ lands in a basin of attraction given by (10), then Algorithm 1 converges linearly with a constant step size. To reach ϵ-accuracy, i.e. $\operatorname{dist}(\boldsymbol{Z}_t, \boldsymbol{Z}_\star) \le \epsilon$, it takes an order of $\kappa\log(1/\epsilon)$ iterations, which is order-wise equivalent to the regularized PF algorithm proposed in [8]. Comparing to [8], which requires $\delta_{6r} \le c$, Theorem 1 only requires a weaker assumption $\delta_{2r} \le c$. However, the basin of attraction allowed by Theorem 1 is smaller than that in [8], which is specified by $\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}}\|\boldsymbol{Z}_0\boldsymbol{R} - \boldsymbol{Z}_\star\|_{\mathrm{F}} \le c_0\sigma_{\min}^{1/2}$. Compared with prior work that relies on local strong convexity to establish linear convergence, our result suggests the benign behavior of gradient descent even in the absence of local strong convexity.

C. Achieving global convergence with a proper initialization

We are still in need of finding a good initialization that obeys (10). In general, one could initialize with the balanced factors of the output of projected gradient descent (over the low-rank matrix), i.e.

$$\boldsymbol{M}_{\tau+1} = \mathcal{P}_r\left(\boldsymbol{M}_\tau - \frac{1}{m}\sum_{i=1}^m\left(\langle\boldsymbol{A}_i, \boldsymbol{M}_\tau\rangle - y_i\right)\boldsymbol{A}_i\right),$$

where $\mathcal{P}_r(\cdot)$ is the Euclidean projection operator to the set of rank-r matrices. The spectral initialization specified in Algorithm 1 can be regarded as the output at the first iteration,

initialized at zero $\boldsymbol{M}_0 = \boldsymbol{0}$. Based on [15], [8], the balanced factorization of $\boldsymbol{M}_\tau$, denoted by $\widetilde{\boldsymbol{Z}}_\tau$, satisfy

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\widetilde{\boldsymbol{Z}}_\tau \boldsymbol{R} - \boldsymbol{Z}_\star\|_{\mathrm{F}} \le c_2(2\delta_{4r})^\tau \frac{\|\boldsymbol{M}_\star\|_{\mathrm{F}}}{\sigma_{\min}^{1/2}} \qquad (11)$$

for some constant c_2. Thus, to achieve the required initialization condition (10) using the spectral method specified in Algorithm 1 (which corresponds to setting $\boldsymbol{Z}_0 = \widetilde{\boldsymbol{Z}}_1$ with $\tau = 1$ in (11)), we need

$$\delta_{4r} \le c_2 \frac{1}{\kappa^{3/2}} \cdot \frac{\sigma_{\min}}{\|\boldsymbol{M}_\star\|_{\mathrm{F}}}.$$

In particular, under Gaussian design, where each measurement matrix $\boldsymbol{A}_i$ has i.i.d. $\mathcal{N}(0, 1/m)$ entries, a total of $m \gtrsim nr\kappa^3 \|\boldsymbol{M}_\star\|_{\mathrm{F}}^2 / \sigma_{\min}^2$ measurements suffice for the above requirement on δ_{4r}. This is worse by a factor of κ^3 compared with the sample complexity guarantee in [8] with the balancing regularizer, which is due to the restriction on the basin of attraction, as we have remarked earlier. Improving the dependence on κ is an interesting future direction.

In order to alleviate the dependency on the condition number κ, we can allow a few iterations of (11) and set the initialization as $\boldsymbol{Z}_0 = \widetilde{\boldsymbol{Z}}_\tau$, a procedure suggested by [8]. The advantage of this hybrid procedure is that the switch to factored gradient descent allows a smaller per-iteration memory and computation complexity after the iterates of projected gradient descent enter the basin of attraction. Consequently, the algorithm is still guaranteed to succeed when $\delta_{4r} \le \delta_c$ for a sufficiently small constant δ_c (which implies a near-optimal sample complexity of $m \gtrsim nr$ under Gaussian design), by running at least

$$\tau \ge c_1 \log\left(\kappa^{3/2} \frac{\|\boldsymbol{M}_\star\|_{\mathrm{F}}}{\sigma_{\min}}\right) / \log\left(\delta_c^{-1}\right)$$

iterations of projected gradient descent for initialization, which order-wise matches the requirement in [8].

III. Related Work

Low-rank matrix estimation has been extensively studied in recent years [3], [4], due to its broad applicability in collaborative filtering, imaging science, and machine learning, to name a few. Convex relaxation approaches based on nuclear norm minimization are among the first set of algorithms with near-optimal statistical guarantees, e. g. [2], [16], [17], [18], [19], [20], [21], [14], [22], however, their computational costs are often prohibitive in practice.

To cope with the computational challenges, a popular approach in practice is to invoke low-rank matrix factorization and then apply first-order methods such as gradient descent directly over the factors to recover the underlying low-rank structure. This approach is demonstrated to possess near-optimal statistical and computational guarantees in a variety of low-rank matrix recovery problems, including but not limited to [8], [23], [24], [25], [26], [27], [28], [29], [30], [31]. The readers are referred to the recent overview [32] for additional references.

To the best of our knowledge, the balancing regularization term (4) was first introduced in [8] to deal with asymmetric matrix factorization, and has become a standard approach to deal with asymmetric low-rank matrix estimation [9], [10], [12], [33], [34], [35]. A major benefit of adding the regularization term is to reduce the ambiguity set from invertible transforms to orthonormal transforms. For the special rank-one matrix recovery problem, there are some evidence in the prior literature that a balancing regularization is not needed, for example, Ma et al. [27] established that vanilla gradient descent works for blind deconvolution at a near-optimal sample complexity with spectral initialization. In [36], the trajectory of gradient descent is studied for asymmetric matrix factorization with an infinitesimal and diminishing step size; in contrast, we consider the case when the step size is constant for low-rank matrix estimation with incomplete observations. Finally, very recently, [37] also studied low-rank matrix sensing using a nonsmooth formulation without the balancing regularization via subgradient descent.

Complementary to the algorithmic analysis, we remark that a similar regularization term (4) is also adopted when analyzing the optimization landscape of low-rank matrix estimation, e.g. [38], [39], [40], [41], [42]. It is worth mentioning that when converting a nuclear-norm regularized problem into a nonconvex formulation, [43] demonstrated that the nonconvex problem has benign geometry without adding the balancing regularization, since the nuclear norm regularization induces a term $\frac{1}{2}(\|\boldsymbol{X}\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\|_{\mathrm{F}}^2)$ which ensures both factors have similar sizes. Very recently, [44] showed that the balancing regularizer is unnecessary from the landscape analysis perspective.

After the initial version of the current paper, several other works have further examined the balancing-free low-rank matrix optimization problem. In particular, Tian, Ma and Chi developed a scaled gradient descent algorithm [45] that achieves a faster convergence rate independent of the condition number κ without imposing the balancing regularization for a variety of low-rank matrix estimation problems, which are further extended in [46] to achieve robustness to adversarial outliers.

IV. Proof of Theorem 1

In this section, we provide the proof of Theorem 1. We first discuss some basic properties of aligning two low-rank factors via an invertible transformation. Then we prove a similar result for a warm-up case of low-rank matrix factorization. In the end, viewing matrix sensing as a perturbed version of low-rank matrix factorization helps us finish the proof of Theorem 1.

A. Alignment via invertible transformations

We begin with introducing the alignment matrix will play a key role in the subsequent analysis.

Definition 2. *Fix a matrix* $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$. *We define the optimal alignment matrix* $\boldsymbol{Q}$ *between* $\boldsymbol{Z}$ *and* $\boldsymbol{Z}_\star$ *as*

$$\boldsymbol{Q} \triangleq \operatorname*{argmin}_{\substack{\boldsymbol{P}\in\mathbb{R}^{r\times r} \\ \textit{invertible}}} \|\boldsymbol{X}\boldsymbol{P} - \boldsymbol{X}_\star\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}\boldsymbol{P}^{-\top} - \boldsymbol{Y}_\star\right\|_{\mathrm{F}}^2,$$

whenever the minimum is attained.

As we will soon see, for the iterates $\{\boldsymbol{Z}_t\}_{t\ge 0}$ generated by Algorithm 1, the optimal alignment matrix is always

well-defined. Furthermore, we call $\boldsymbol{Z}$ and $\boldsymbol{Z}_\star$ aligned if the corresponding optimal alignment matrix is just the identity matrix $\boldsymbol{I}_r$. Below we provide some basic understandings of this alignment matrix.

The following lemma provides a sufficient condition for the existence of the optimal alignment matrix.

Lemma 1. *Fix some matrix* $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$. *Suppose that there exists a matrix* $\boldsymbol{P} \in \mathbb{R}^{r\times r}$ *with* $1/2 \le \sigma_r(\boldsymbol{P}) \le \sigma_1(\boldsymbol{P}) \le 3/2$ *such that*

$$\max\left\{\|\boldsymbol{X}\boldsymbol{P}-\boldsymbol{X}_\star\|_{\mathrm{F}}, \left\|\boldsymbol{Y}\boldsymbol{P}^{-\top}-\boldsymbol{Y}_\star\right\|_{\mathrm{F}}\right\} \le \delta \le \frac{\sigma_r(\boldsymbol{X}_\star)}{80}. \tag{12}$$

Then the optimal alignment matrix $\boldsymbol{Q} \in \mathbb{R}^{r\times r}$ *between* $\boldsymbol{Z}$ *and* $\boldsymbol{Z}_\star$ *exists. In addition, the matrix* $\boldsymbol{Q}$ *satisfies*

$$\|\boldsymbol{P}-\boldsymbol{Q}\| \le \|\boldsymbol{P}-\boldsymbol{Q}\|_{\mathrm{F}} \le \frac{5\delta}{\sigma_r(\boldsymbol{X}_\star)}.$$

Next, the lemma below presents a necessary condition for $\boldsymbol{Q}$ to be the optimal alignment matrix between $\boldsymbol{Z}$ and $\boldsymbol{Z}_\star$.

Lemma 2. *Let* $\boldsymbol{Z}$ *and* $\boldsymbol{Z}_\star$ *be any two matrices. Suppose that the optimal alignment matrix* $\boldsymbol{Q}$ *between* $\boldsymbol{Z}$ *and* $\boldsymbol{Z}_\star$ *exists. Then we have*

$$\widetilde{\boldsymbol{X}}^\top(\widetilde{\boldsymbol{X}}-\boldsymbol{X}_\star) = (\widetilde{\boldsymbol{Y}}-\boldsymbol{Y}_\star)^\top\widetilde{\boldsymbol{Y}},$$

where $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\boldsymbol{Q}$ *and* $\widetilde{\boldsymbol{Y}} = \boldsymbol{Y}\boldsymbol{Q}^{-\top}$ *are two matrices after the alignment.*

Both lemmas provide basic understandings of the solution to the alignment problem with invertible transformations, which can be regarded as a generalization of the classical orthogonal Procrustes problem that only considers orthonormal transformations. Clearly, this generalized problem is more involved and our work provides some basic understandings.

B. A warm-up: low-rank matrix factorization

We consider the following minimization problem for low-rank matrix factorization

$$f_{\mathrm{MF}}(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}_\star\right\|_{\mathrm{F}}^2, \tag{13}$$

where $\boldsymbol{X} \in \mathbb{R}^{n_1\times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_2\times r}$. The gradient descent updates with an initialization $(\boldsymbol{X}_0,\boldsymbol{Y}_0)$ can be written as

$$\begin{aligned}
\boldsymbol{X}_{t+1} &= \boldsymbol{X}_t - \frac{\eta}{\sigma_{\max}}\nabla_{\boldsymbol{X}} f_{\mathrm{MF}}(\boldsymbol{X}_t,\boldsymbol{Y}_t) \\
&= \boldsymbol{X}_t - \frac{\eta}{\sigma_{\max}}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top-\boldsymbol{M}_\star)\boldsymbol{Y}_t; \\
\boldsymbol{Y}_{t+1} &= \boldsymbol{Y}_t - \frac{\eta}{\sigma_{\max}}\nabla_{\boldsymbol{Y}} f_{\mathrm{MF}}(\boldsymbol{X}_t,\boldsymbol{Y}_t) \\
&= \boldsymbol{Y}_t - \frac{\eta}{\sigma_{\max}}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top-\boldsymbol{M}_\star)^\top\boldsymbol{X}_t.
\end{aligned} \tag{14}$$

Here, $\eta > 0$ stands for the step size. We have the following theorem regarding the performance of (14), which parallels Theorem 1.

Theorem 2. *Let* $\boldsymbol{Z}_0 = \begin{bmatrix} \boldsymbol{X}_0 \\ \boldsymbol{Y}_0 \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$ *be any initialization point that satisfies*

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{Z}_0\boldsymbol{R}-\boldsymbol{Z}_\star\|_{\mathrm{F}} \le c_0\frac{1}{\kappa^{3/2}}\sigma_{\min}^{1/2} \tag{15}$$

for some sufficiently small constant $c_0 > 0$. *Then setting the step size* $\eta > 0$ *to be some sufficiently small constant, the iterates of GD (cf. (14)) satisfy*

$$\mathrm{dist}(\boldsymbol{Z}_t,\boldsymbol{Z}_\star) \le \left(1-\frac{\eta}{50\kappa}\right)^t \mathrm{dist}(\boldsymbol{Z}_0,\boldsymbol{Z}_\star).$$

To prove Theorem 2, we need the following properties regarding the gradients of $f_{\mathrm{MF}}(\boldsymbol{X},\boldsymbol{Y})$; the proofs are deferred to the appendix.

Lemma 3 (Gradient dominance)**.** *Suppose that* $\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times r}$ *is aligned with* $\boldsymbol{Z}_\star$, *i.e.*

$$\boldsymbol{I}_r = \operatorname*{argmin}_{\substack{\boldsymbol{P}\in\mathbb{R}^{r\times r} \\ \text{invertible}}} \|\boldsymbol{X}\boldsymbol{P}-\boldsymbol{X}_\star\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}\boldsymbol{P}^{-\top}-\boldsymbol{X}_\star\right\|_{\mathrm{F}}^2.$$

Then we have

$$\begin{aligned}
&\left\langle \boldsymbol{X}-\boldsymbol{X}_\star, \left(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}_\star\right)\boldsymbol{Y}\right\rangle \\
&\qquad \ge \left\|\boldsymbol{Y}(\boldsymbol{X}-\boldsymbol{X}_\star)^\top\right\|_{\mathrm{F}}^2 - \frac{1}{4}\|\boldsymbol{X}-\boldsymbol{X}_\star\|_{\mathrm{F}}^4,
\end{aligned}$$

and similarly,

$$\begin{aligned}
&\left\langle \boldsymbol{Y}-\boldsymbol{Y}_\star, \left(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}_\star\right)^\top\boldsymbol{X}\right\rangle \\
&\qquad \ge \left\|\boldsymbol{X}(\boldsymbol{Y}-\boldsymbol{Y}_\star)^\top\right\|_{\mathrm{F}}^2 - \frac{1}{4}\|\boldsymbol{Y}-\boldsymbol{Y}_\star\|_{\mathrm{F}}^4.
\end{aligned}$$

Lemma 4 (Smoothness)**.** *Suppose that* $\|\boldsymbol{Y}-\boldsymbol{Y}_\star\| \le \sigma_1(\boldsymbol{Y}_\star)/4$, *then one has*

$$\begin{aligned}
&\left\|\left(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}_\star\right)\boldsymbol{Y}\right\|_{\mathrm{F}} \le \\
&\quad \frac{3}{2}\sigma_1(\boldsymbol{Y}_\star)\Big(\left\|(\boldsymbol{X}-\boldsymbol{X}_\star)\boldsymbol{Y}^\top\right\|_{\mathrm{F}} + \left\|\boldsymbol{X}(\boldsymbol{Y}-\boldsymbol{Y}_\star)^\top\right\|_{\mathrm{F}} \\
&\qquad\qquad + \|\boldsymbol{X}-\boldsymbol{X}_\star\|_{\mathrm{F}}\|\boldsymbol{Y}-\boldsymbol{Y}_\star\|_{\mathrm{F}}\Big).
\end{aligned}$$

Similarly, with the proviso that $\|\boldsymbol{X}-\boldsymbol{X}_\star\| \le \sigma_1(\boldsymbol{X}_\star)/4$, *one has*

$$\begin{aligned}
&\left\|\left(\boldsymbol{X}\boldsymbol{Y}^\top-\boldsymbol{M}_\star\right)^\top\boldsymbol{X}\right\|_{\mathrm{F}} \le \\
&\quad \frac{3}{2}\sigma_1(\boldsymbol{X}_\star)\Big(\left\|(\boldsymbol{X}-\boldsymbol{X}_\star)\boldsymbol{Y}^\top\right\|_{\mathrm{F}} + \left\|\boldsymbol{X}(\boldsymbol{Y}-\boldsymbol{Y}_\star)^\top\right\|_{\mathrm{F}} \\
&\qquad\qquad + \|\boldsymbol{X}-\boldsymbol{X}_\star\|_{\mathrm{F}}\|\boldsymbol{Y}-\boldsymbol{Y}_\star\|_{\mathrm{F}}\Big).
\end{aligned}$$

C. Proof of Theorem 2

With the help of Lemmas 1–4, we are in a position to establish Theorem 2. Denote by $\widehat{\boldsymbol{R}} \in \mathbb{R}^{r\times r}$ the best rotation matrix between $\boldsymbol{Z}_0$ and $\boldsymbol{Z}_\star$, that is

$$\widehat{\boldsymbol{R}} \triangleq \operatorname*{argmin}_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{Z}_0\boldsymbol{R}-\boldsymbol{Z}_\star\|_{\mathrm{F}}.$$

Combine the assumption of initialization (cf. (15)) and Lemma 1 to see that

$$\boldsymbol{Q}_0 \triangleq \operatorname*{argmin}_{\substack{\boldsymbol{P}\in\mathbb{R}^{r\times r} \\ \text{invertible}}} \|\boldsymbol{X}_0\boldsymbol{P}-\boldsymbol{X}_\star\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}_0\boldsymbol{P}^{-\top}-\boldsymbol{Y}_\star\right\|_{\mathrm{F}}^2$$

exists and in addition, one has

$$\|\boldsymbol{Q}_0-\widehat{\boldsymbol{R}}\| \le \frac{5c_0}{\kappa^{3/2}} \le \frac{1}{400\sqrt{\kappa}}$$

as long as $c_0 > 0$ is sufficiently small.

The remaining proof is inductive in nature. In particular, we aim at proving the following induction hypotheses.

1) The optimal alignment matrix $\boldsymbol{Q}_t$ between $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_\star$ exists.
2) The distance between $\boldsymbol{Z}_t$ and $\boldsymbol{Z}_\star$ obeys
$$\mathrm{dist}\left(\boldsymbol{Z}_t, \boldsymbol{Z}_\star\right) \le \left(1 - \frac{\eta}{50\kappa}\right)^t \mathrm{dist}\left(\boldsymbol{Z}_0, \boldsymbol{Z}_\star\right).$$
3) The optimal alignment matrix $\boldsymbol{Q}_t$ is nearly a rotation matrix in the sense that
$$\|\boldsymbol{Q}_t - \widehat{\boldsymbol{R}}\| \le \frac{1}{400\sqrt{\kappa}}.$$

It is straightforward to check that these three claims hold for $t = 0$. In what follows, we shall assume that the induction hypotheses hold for all iterations up to the tth iteration and intend to establish that they continue to hold for the $(t+1)$th iteration.

a) Verifying the first induction hypothesis: We begin with demonstrating the existence of $\boldsymbol{Q}_{t+1}$. In view of the gradient update rule (14), we have

$$\begin{aligned}
\boldsymbol{X}_{t+1}\boldsymbol{Q}_t &= \boldsymbol{X}_t\boldsymbol{Q}_t - \frac{\eta}{\sigma_{\max}}\left(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star\right)\boldsymbol{Y}_t\boldsymbol{Q}_t \\
&= \widetilde{\boldsymbol{X}}_t - \frac{\eta}{\sigma_{\max}}\left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\left(\boldsymbol{Q}_t^\top\boldsymbol{Q}_t\right), \\
\boldsymbol{Y}_{t+1}\boldsymbol{Q}_t^{-\top} &= \boldsymbol{Y}_t\boldsymbol{Q}_t^{-\top} - \frac{\eta}{\sigma_{\max}}\left(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star\right)^\top\boldsymbol{X}_t\boldsymbol{Q}_t^{-\top} \\
&= \widetilde{\boldsymbol{Y}}_t - \frac{\eta}{\sigma_{\max}}\left(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star\right)^\top\widetilde{\boldsymbol{X}}_t\left(\boldsymbol{Q}_t^\top\boldsymbol{Q}_t\right)^{-1},
\end{aligned}$$

where we denote

$$\widetilde{\boldsymbol{X}}_t \triangleq \boldsymbol{X}_t\boldsymbol{Q}_t \qquad \text{and} \qquad \widetilde{\boldsymbol{Y}}_t \triangleq \boldsymbol{Y}_t\boldsymbol{Q}_t^{-\top}.$$

As a result, one has the following equality

$$\begin{aligned}
&\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_t - \boldsymbol{X}_\star\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{Y}_\star\right\|_{\mathrm{F}}^2 \\
&= \underbrace{\left\|\widetilde{\boldsymbol{X}}_t - \boldsymbol{X}_\star - \frac{\eta}{\sigma_{\max}}\left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\|_{\mathrm{F}}^2}_{=:\alpha_1} \\
&\quad + \underbrace{\left\|\widetilde{\boldsymbol{Y}}_t - \boldsymbol{Y}_\star - \frac{\eta}{\sigma_{\max}}\left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)^\top\widetilde{\boldsymbol{X}}_t\boldsymbol{\Lambda}_t^{-1}\right\|_{\mathrm{F}}^2}_{=:\alpha_2},
\end{aligned}$$

where we have denoted $\boldsymbol{\Lambda}_t \triangleq \boldsymbol{Q}_t^\top\boldsymbol{Q}_t$. By virtue of the third induction hypothesis, namely $\|\boldsymbol{Q}_t - \widehat{\boldsymbol{R}}\| \le 1/(400\sqrt{\kappa})$, it is easy to check that $\|\boldsymbol{\Lambda}_t - \boldsymbol{I}_r\| \le 1/(180\sqrt{\kappa}) \triangleq \zeta$. Let

$$\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t} \triangleq \widetilde{\boldsymbol{X}}_t - \boldsymbol{X}_\star \qquad \text{and} \qquad \widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t} \triangleq \widetilde{\boldsymbol{Y}}_t - \boldsymbol{Y}_\star.$$

Expand α_1 to obtain

$$\begin{aligned}
\alpha_1 =& \|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}^2 + \left(\frac{\eta}{\sigma_{\max}}\right)^2 \underbrace{\left\|\left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\|_{\mathrm{F}}^2}_{=:\beta_1} \\
&- 2\frac{\eta}{\sigma_{\max}}\underbrace{\left\langle\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}, \left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\rangle}_{=:\gamma_1}.
\end{aligned}$$

Similarly, we can decompose α_2 into

$$\begin{aligned}
\alpha_2 =& \|\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}\|_{\mathrm{F}}^2 + \left(\frac{\eta}{\sigma_{\max}}\right)^2 \underbrace{\left\|\left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)^\top\widetilde{\boldsymbol{X}}_t\boldsymbol{\Lambda}_t^{-1}\right\|_{\mathrm{F}}^2}_{=:\beta_2} \\
&- 2\frac{\eta}{\sigma_{\max}}\underbrace{\left\langle\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}, \left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)^\top\widetilde{\boldsymbol{X}}_t\boldsymbol{\Lambda}_t^{-1}\right\rangle}_{=:\gamma_2}.
\end{aligned}$$

We intend to apply Lemma 3 to lower bound the terms γ_1 and γ_2 and apply Lemma 4 to upper bound β_1 and β_2. First, since $(\widetilde{\boldsymbol{X}}_t, \widetilde{\boldsymbol{Y}}_t)$ is aligned with $(\boldsymbol{X}_\star, \boldsymbol{Y}_\star)$, we can invoke Lemma 3 to see that

$$\begin{aligned}
\gamma_1 &\ge \left\langle\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}, \left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\right\rangle \\
&\quad - \left|\left\langle\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}, \left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\left(\boldsymbol{\Lambda}_t - \boldsymbol{I}_r\right)\right\rangle\right| \\
&\ge \|\widetilde{\boldsymbol{Y}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\|_{\mathrm{F}}^2 - \frac{1}{4}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}^4 \\
&\quad - \|\boldsymbol{\Lambda}_t - \boldsymbol{I}_r\|\cdot\left\|\left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\right\|_{\mathrm{F}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}} \\
&\ge \|\widetilde{\boldsymbol{Y}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\|_{\mathrm{F}}^2 - \frac{1}{400}\sigma_{\min}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}^2 \\
&\quad - \zeta\left\|\left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\right\|_{\mathrm{F}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}. \qquad (16)
\end{aligned}$$

Here the last line follows from the bound $\|\boldsymbol{\Lambda}_t - \boldsymbol{I}_r\| \le \zeta$ and the second induction hypothesis, i.e.

$$\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}^2 \le \mathrm{dist}^2\left(\boldsymbol{Z}_t, \boldsymbol{Z}_\star\right) \le \mathrm{dist}^2\left(\boldsymbol{Z}_0, \boldsymbol{Z}_\star\right) \le \frac{1}{100}\sigma_{\min}.$$

The last term in (16) can be further bounded via Lemma 4 as

$$\begin{aligned}
&\zeta\left\|\left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top - \boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\right\|_{\mathrm{F}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}} \le \frac{3\zeta}{2}\sqrt{\sigma_{\max}}\cdot \\
&\quad\left(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}\|_{\mathrm{F}}\right)\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}} \\
&= \frac{9\zeta\sqrt{\kappa}}{2}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}\cdot\frac{\sqrt{\sigma_{\min}}}{3}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}} \\
&\quad + \frac{9\zeta\sqrt{\kappa}}{2}\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}\cdot\frac{\sqrt{\sigma_{\min}}}{3}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}} \\
&\quad + \frac{3\zeta}{2}\sqrt{\sigma_{\max}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}\|_{\mathrm{F}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}^2 \\
&\le \frac{81\zeta^2\kappa}{8}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2 + \frac{81\zeta^2\kappa}{8}\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2 + \frac{\sigma_{\min}}{8}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}^2,
\end{aligned}$$

where the last inequality arises since $ab \le (a^2 + b^2)/2$ and

$$\begin{aligned}
\frac{3\zeta}{2}\sqrt{\sigma_{\max}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}\|_{\mathrm{F}} &\le \frac{3\zeta}{2}\sqrt{\sigma_{\max}}\,\mathrm{dist}\left(\boldsymbol{Z}_0, \boldsymbol{Z}_\star\right) \\
&\le \frac{3\zeta}{2}\sqrt{\sigma_{\max}}c_0\frac{1}{\kappa^{3/2}}\sqrt{\sigma_{\min}} \le \frac{\sigma_{\min}}{72}
\end{aligned}$$

as long as c_0 is sufficiently small. Combine the above two bounds to reach

$$\begin{aligned}
\gamma_1 \ge& \left(1 - \frac{81\zeta^2\kappa}{8}\right)\|\widetilde{\boldsymbol{Y}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\|_{\mathrm{F}}^2 \\
&- \frac{81\zeta^2\kappa}{8}\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2 - \frac{\sigma_{\min}}{7}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}^2.
\end{aligned}$$

Similarly, γ_2 can be lower bounded as

$$\gamma_2 \ge \left(1 - \frac{81\zeta^2\kappa}{8}\right)\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2$$

$$- \frac{81\zeta^2\kappa}{8}\|\widetilde{\boldsymbol{Y}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\|_{\mathrm{F}}^2 - \frac{\sigma_{\min}}{7}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}\|_{\mathrm{F}}^2,$$

which together with the bound on γ_1 implies

$$\begin{aligned}
&\gamma_1+\gamma_2\\
&\geq \left(1-\frac{81\zeta^2\kappa}{4}\right)\left(\|\widetilde{\boldsymbol{Y}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\|_{\mathrm{F}}^2+\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2\right)\\
&\quad-\frac{\sigma_{\min}}{7}\mathrm{dist}^2\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right)\\
&\geq \frac{3}{4}\left(\|\widetilde{\boldsymbol{Y}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\|_{\mathrm{F}}^2+\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2\right)-\frac{\sigma_{\min}}{7}\mathrm{dist}^2\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right),
\end{aligned}$$

where we plug in the definition of $\zeta = 1/(180\sqrt{\kappa})$.

Now we move on to controlling β_1 and β_2. Recognizing that $\|\boldsymbol{\Lambda}_t\|\leq 2$, one has

$$\begin{aligned}
\beta_1 &\leq 4\left\|\left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{Y}}_t^\top-\boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\right\|_{\mathrm{F}}^2\\
&\leq 9\sigma_{\max}\Big(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}+\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}+\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}\|_{\mathrm{F}}\Big)^2,
\end{aligned}\tag{17}$$

where the second line follows from Lemma 4. Apply the elementary inequality $(a+b+c)^2\leq 3(a^2+b^2+c^2)$ to see that

$$\begin{aligned}
\beta_1 &\leq 27\sigma_{\max}\Big(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2+\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2+\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}^2\|\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}\|_{\mathrm{F}}^2\Big)\\
&\leq 27\sigma_{\max}\Big(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2+\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2\Big)\\
&\qquad+27c_0^2\frac{\sigma_{\max}\sigma_{\min}}{\kappa^3}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}^2.
\end{aligned}$$

Here the second line relies on the fact that $\|\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}\|_{\mathrm{F}}^2\leq \mathrm{dist}^2(\boldsymbol{Z}_0,\boldsymbol{Z}_\star)\leq c_0^2\sigma_{\min}/\kappa^3$. Similarly, one can bound β_2 as

$$\begin{aligned}
\beta_2 &\leq 27\sigma_{\max}\left(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2+\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2\right)\\
&\qquad+27c_0^2\frac{\sigma_{\max}\sigma_{\min}}{\kappa^3}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}\|_{\mathrm{F}}^2,
\end{aligned}$$

which in conjunction with the bound on β_1 yields

$$\begin{aligned}
\beta_1+\beta_2 &\leq 54\sigma_{\max}\left(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2+\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2\right)\\
&\qquad+27c_0^2\frac{\sigma_{\max}\sigma_{\min}}{\kappa^3}\mathrm{dist}^2\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right).
\end{aligned}$$

Collect all the bounds on α_1 and α_2 to arrive at

$$\begin{aligned}
&\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_t-\boldsymbol{X}_\star\|_{\mathrm{F}}^2+\left\|\boldsymbol{Y}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{Y}_\star\right\|_{\mathrm{F}}^2\\
&\leq\left(1+\frac{27c_0^2\eta^2}{\kappa^4}\right)\mathrm{dist}^2\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right)\\
&\qquad+\left(\frac{54\eta^2}{\sigma_{\max}}\right)\left(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2+\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2\right)\\
&\qquad-2\frac{\eta}{\sigma_{\max}}\Big[\frac{3}{4}\Big(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2+\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2\Big)\\
&\qquad\qquad-\frac{\sigma_{\min}}{7}\mathrm{dist}^2\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right)\Big]\\
&=\left(1+\frac{27c_0^2\eta^2}{\kappa^4}+\frac{\eta}{3.5\kappa}\right)\mathrm{dist}^2\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right)\\
&+\left(\frac{54\eta^2}{\sigma_{\max}}-\frac{3\eta}{2\sigma_{\max}}\right)\cdot\left(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2+\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2\right)\\
&\leq\left(1+\frac{\eta}{3\kappa}\right)\mathrm{dist}^2\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right)\\
&\qquad-\frac{3\eta}{4\sigma_{\max}}\left(\sigma_r^2(\widetilde{\boldsymbol{Y}}_t)\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}}^2+\sigma_r^2(\widetilde{\boldsymbol{X}}_t)\|\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}\|_{\mathrm{F}}^2\right),
\end{aligned}$$

where the last line follows as long as $\eta\leq 1/24$. Furthermore, since $\sigma_r^2(\widetilde{\boldsymbol{Y}}_t)\geq\sigma_{\min}/2$ and $\sigma_r^2(\widetilde{\boldsymbol{X}}_t)\geq\sigma_{\min}/2$, we have

$$\begin{aligned}
&\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_t-\boldsymbol{X}_\star\|_{\mathrm{F}}^2+\left\|\boldsymbol{Y}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{Y}_\star\right\|_{\mathrm{F}}^2\\
&\qquad\leq\left(1-\frac{\eta}{24\kappa}\right)\mathrm{dist}^2\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right).
\end{aligned}\tag{18}$$

Lemma 1 then ensures the existence of $\boldsymbol{Q}_{t+1}$.

b) Verifying the second induction hypothesis: The second induction hypothesis for the $(t+1)$th iteration follows immediately from the above proof. Since $\boldsymbol{Q}_{t+1}$ exists, by definition, one has

$$\begin{aligned}
&\mathrm{dist}\left(\boldsymbol{Z}_{t+1},\boldsymbol{Z}_\star\right)\\
&\quad=\sqrt{\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_{t+1}-\boldsymbol{X}_\star\|_{\mathrm{F}}^2+\left\|\boldsymbol{Y}_{t+1}\boldsymbol{Q}_{t+1}^{-\top}-\boldsymbol{X}_\star\right\|_{\mathrm{F}}^2}\\
&\quad\leq\sqrt{\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_t-\boldsymbol{X}_\star\|_{\mathrm{F}}^2+\left\|\boldsymbol{Y}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{X}_\star\right\|_{\mathrm{F}}^2}\\
&\quad\leq\left(1-\frac{\eta}{50\kappa}\right)\mathrm{dist}\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right).
\end{aligned}$$

c) Verifying the third induction hypothesis: It remains to show the last induction hypothesis, namely $\|\boldsymbol{Q}_{t+1}-\widehat{\boldsymbol{R}}\|\leq 1/(400\sqrt{\kappa})$. In view of (18), one has $\max\{\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_t-\boldsymbol{X}_\star\|_{\mathrm{F}},\|\boldsymbol{Y}_{t+1}\boldsymbol{Q}_t^{-\top}-\boldsymbol{Y}_\star\|_{\mathrm{F}}\}\leq\mathrm{dist}\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right)$. Invoke Lemma 1 again to arrive at

$$\begin{aligned}
\|\boldsymbol{Q}_{t+1}-\boldsymbol{Q}_t\| &\leq\frac{5}{\sigma_r\left(\boldsymbol{X}_\star\right)}\mathrm{dist}\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right)\\
&\leq\frac{5}{\sigma_r\left(\boldsymbol{X}_\star\right)}\left(1-\frac{\eta}{50\kappa}\right)^t c_0\frac{1}{\kappa^{3/2}}\sigma_r\left(\boldsymbol{X}_\star\right)\\
&\leq 5c_0\left(1-\frac{\eta}{50\kappa}\right)^t\frac{1}{\kappa^{3/2}}.
\end{aligned}$$

Hence, by the triangle inequality and the telescoping sum, we obtain

$$\begin{aligned}
\|\boldsymbol{Q}_{t+1}-\widehat{\boldsymbol{R}}\| &\leq\sum_{s=0}^{t}\|\boldsymbol{Q}_{s+1}-\boldsymbol{Q}_s\|+\|\boldsymbol{Q}_0-\widehat{\boldsymbol{R}}\|\\
&\leq 5c_0\sum_{s=0}^{t}\left(1-\frac{\eta}{50\kappa}\right)^s\frac{1}{\kappa^{3/2}}+5c_0\frac{1}{\kappa^{3/2}}\\
&<5c_0\sum_{s=0}^{\infty}\left(1-\frac{\eta}{50\kappa}\right)^s\frac{1}{\kappa^{3/2}}+5c_0\frac{1}{\kappa^{3/2}}\\
&=5c_0\frac{50\kappa}{\eta}\frac{1}{\kappa^{3/2}}+5c_0\frac{1}{\kappa^{3/2}}\\
&\leq\frac{1}{400\sqrt{\kappa}},
\end{aligned}$$

as long as c_0 is small enough and η is some constant.

Putting everything together, we finish the induction step and the proof is then completed.

D. Analysis for matrix sensing

We now extend the techniques used in the proof of Theorem 2 to the matrix sensing case by leveraging the RIP. Suppose that the initialization $\boldsymbol{Z}_0$ satisfies the condition (10). By a standard

argument as in [8], [9], [33],[2] it is sufficient to consider the following update rule:

$$\begin{aligned}\boldsymbol{X}_{t+1} &= \boldsymbol{X}_t - \frac{\eta}{\sigma_{\max}}\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star)\right]\boldsymbol{Y}_t;\\ \boldsymbol{Y}_{t+1} &= \boldsymbol{Y}_t - \frac{\eta}{\sigma_{\max}}\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star)\right]^\top\boldsymbol{X}_t.\end{aligned} \tag{19}$$

Compared with the update rule (14) for low-rank matrix factorization, the update rule for matrix sensing differs by the operation of $\mathcal{A}^*\mathcal{A}$ when forming the gradient. Therefore, we expect that GD has similar behaviors as earlier as long as the operator $\mathcal{A}^*\mathcal{A}$ behaves as a near isometry on low-rank matrices. This can be supplied by the following consequence of the RIP.

Lemma 5. *Suppose that $\mathcal{A}$ satisfies $2r$-RIP with a constant δ_{2r}. Then, for all matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ of rank at most r, we have*

$$\left|\langle\mathcal{A}(\boldsymbol{M}_1),\mathcal{A}(\boldsymbol{M}_2)\rangle - \langle\boldsymbol{M}_1,\boldsymbol{M}_2\rangle\right| \le \delta_{2r}\left\|\boldsymbol{M}_1\right\|_{\mathrm{F}}\left\|\boldsymbol{M}_2\right\|_{\mathrm{F}}.$$

Equivalently, we can write this as

$$\left|\mathrm{Tr}\left[\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)(\boldsymbol{M}_1)\,\boldsymbol{M}_2^\top\right]\right| \le \delta_{2r}\left\|\boldsymbol{M}_1\right\|_{\mathrm{F}}\left\|\boldsymbol{M}_2\right\|_{\mathrm{F}}.$$

A simple consequence is that for any $\boldsymbol{A}\in\mathbb{R}^{n_2\times r}$

$$\left\|\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)(\boldsymbol{M}_1)\,\boldsymbol{A}\right\|_{\mathrm{F}} \le \delta_{2r}\left\|\boldsymbol{M}_1\right\|_{\mathrm{F}}\left\|\boldsymbol{A}\right\|.$$

Similar to before, we denote $\widetilde{\boldsymbol{X}}_t = \boldsymbol{X}_t\boldsymbol{Q}_t$ and $\widetilde{\boldsymbol{Y}}_t = \boldsymbol{Y}_t\boldsymbol{Q}_t^{-\top}$, which are aligned with $(\boldsymbol{X}_\star,\boldsymbol{Y}_\star)$. With this notation in place, we can rewrite the update rule as

$$\begin{aligned}\boldsymbol{X}_{t+1}\boldsymbol{Q}_t &= \widetilde{\boldsymbol{X}}_t - \frac{\eta}{\sigma_{\max}}\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t,\\ \boldsymbol{Y}_{t+1}\boldsymbol{Q}_t^{-\top} &= \widetilde{\boldsymbol{Y}}_t - \frac{\eta}{\sigma_{\max}}\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star)\right]^\top\widetilde{\boldsymbol{X}}_t\boldsymbol{\Lambda}_t^{-1}.\end{aligned}$$

where we recall $\boldsymbol{\Lambda}_t = \boldsymbol{Q}_t^\top\boldsymbol{Q}_t$. By the definition of the distance function, we further obtain

$$\begin{aligned}&\mathrm{dist}^2(\boldsymbol{Z}_{t+1},\boldsymbol{Z}_\star)\\ &\le \left\|\boldsymbol{X}_{t+1}\boldsymbol{Q}_t - \boldsymbol{X}_\star\right\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}_{t+1}\boldsymbol{Q}_t^{-\top} - \boldsymbol{Y}_\star\right\|_{\mathrm{F}}^2\\ &= \left\|\widetilde{\boldsymbol{X}}_t - \frac{\eta}{\sigma_{\max}}\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t - \boldsymbol{X}_\star\right\|_{\mathrm{F}}^2\\ &\quad + \left\|\widetilde{\boldsymbol{Y}}_t - \frac{\eta}{\sigma_{\max}}\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star)\right]^\top\widetilde{\boldsymbol{X}}_t\boldsymbol{\Lambda}_t^{-1} - \boldsymbol{Y}_\star\right\|_{\mathrm{F}}^2\\ &= \left\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\right\|_{\mathrm{F}}^2 + \left\|\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}\right\|_{\mathrm{F}}^2\\ &\quad + \left(\frac{\eta}{\sigma_{\max}}\right)^2\Bigg(\underbrace{\left\|\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\|_{\mathrm{F}}^2}_{=:\widetilde{\beta}_1}\\ &\qquad + \underbrace{\left\|\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star)\right]^\top\widetilde{\boldsymbol{X}}_t\boldsymbol{\Lambda}_t^{-1}\right\|_{\mathrm{F}}^2}_{=:\widetilde{\beta}_2}\Bigg)\\ &\quad - \frac{2\eta}{\sigma_{\max}}\Bigg(\underbrace{\left\langle\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t},\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\rangle}_{=:\widetilde{\gamma}_1}\\ &\qquad + \underbrace{\left\langle\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t},\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star)\right]^\top\widetilde{\boldsymbol{X}}_t\boldsymbol{\Lambda}_t^{-1}\right\rangle}_{=:\widetilde{\gamma}_2}\Bigg),\end{aligned} \tag{20}$$

where $\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t} \triangleq \widetilde{\boldsymbol{X}}_t - \boldsymbol{X}_\star$ and $\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t} \triangleq \widetilde{\boldsymbol{Y}}_t - \boldsymbol{Y}_\star$. From the high level, the four terms $\widetilde{\beta}_1,\widetilde{\beta}_2,\widetilde{\gamma}_1$ and $\widetilde{\gamma}_2$ are the perturbed versions of β_1,β_2,γ_1 and γ_2 in Section IV-C, respectively.

For the first term, we have

$$\begin{aligned}&\sqrt{\widetilde{\beta}_1} - \sqrt{\beta_1}\\ &\overset{\text{(i)}}{\le} \left\|\left[\mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star\right)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t - \left(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\|_{\mathrm{F}}\\ &= \left\|\left[\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star\right)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\|_{\mathrm{F}}\\ &\overset{\text{(ii)}}{\le} \left\|\left[\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\|_{\mathrm{F}}\\ &\quad + \left\|\left[\left(\mathcal{A}^*\mathcal{A}-\boldsymbol{I}\right)\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\|_{\mathrm{F}}\\ &\quad + \left\|\left[\left(\mathcal{A}^*\mathcal{A}-\boldsymbol{I}\right)\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\|_{\mathrm{F}}\\ &\overset{\text{(iii)}}{\le} \delta_{2r}\left(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}\right)\|\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\|\\ &\overset{\text{(iv)}}{\le} 4\delta_{2r}\sqrt{\sigma_{\max}}\left(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}\right).\end{aligned} \tag{21}$$

Here, the first (i) and second (ii) inequalities follow from the triangle inequality. The third one (iii) uses Lemma 5 and the last relation (iv) depends on $\|\boldsymbol{\Lambda}_t\|\le 2$ and $\|\widetilde{\boldsymbol{Y}}_t\|\le 2\sqrt{\sigma_{\max}}$. Comparing (21) with (17) reveals that $\widetilde{\beta}_1 - \beta_1$ constitutes a small perturbation to β_1 when δ_{2r} is small. Similar bounds hold for $\sqrt{\widetilde{\beta}_2} - \sqrt{\beta_2}$. As a result, when δ_{2r} is sufficiently small, we have

$$\begin{aligned}\widetilde{\beta}_1 + \widetilde{\beta}_2 \le{}& 108\sigma_{\max}\left(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2 + \|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2\right)\\ &+ 54c_0^2\frac{\sigma_{\max}\sigma_{\min}}{\kappa^3}\mathrm{dist}^2\left(\boldsymbol{Z}_t,\boldsymbol{Z}_\star\right).\end{aligned}$$

We now proceed to $\widetilde{\gamma}_1$, for which we have

$$\begin{aligned}&|\widetilde{\gamma}_1 - \gamma_1|\\ &\quad= \Big|\left\langle\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t},\left[\mathcal{A}^*\mathcal{A}\left(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star\right)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\rangle\\ &\qquad\qquad - \left\langle\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t},\left(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star\right)\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\rangle\Big|\\ &\quad= \left|\left\langle\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t},\left[\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star\right)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\right\rangle\right|\\ &\quad= \left|\mathrm{Tr}\left(\left[\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\star\right)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\right)\right|\\ &\quad\le \left|\mathrm{Tr}\left(\left[\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\right)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\right)\right|\\ &\qquad + \left|\mathrm{Tr}\left(\left[\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\right)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\right)\right|\\ &\qquad + \left|\mathrm{Tr}\left(\left[\left(\mathcal{A}^*\mathcal{A}-\mathcal{I}\right)\left(\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\right)\right]\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\right)\right|\\ &\quad\le \delta_{2r}\left(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}\right)\cdot\\ &\qquad\qquad \|\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\|_{\mathrm{F}}.\end{aligned}$$

Here once again, we utilize the triangle inequality and Lemma 5. Noticing that $\|\boldsymbol{\Lambda}_t - \boldsymbol{I}\|$ is small, we further have

$$\|\widetilde{\boldsymbol{Y}}_t\boldsymbol{\Lambda}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\|_{\mathrm{F}} \le \|\widetilde{\boldsymbol{Y}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\|_{\mathrm{F}} + \|\widetilde{\boldsymbol{Y}}_t\left(\boldsymbol{\Lambda}_t - \boldsymbol{I}_r\right)\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\|_{\mathrm{F}}$$

[2]Since (i) the initialization $\boldsymbol{Z}_0$ is close to the ground truth $\boldsymbol{Z}_\star$, (ii) $\boldsymbol{X}_0$ and $\boldsymbol{Y}_0$ are balanced, it is obvious that the operator norm $\|\boldsymbol{X}_0\|^2 = \|\boldsymbol{Y}_0\|^2$ is orderwise equivalent to $\sigma_{\max}$. Therefore, all the convergence claims on using $\sigma_{\max}$ can be translated to those on using $\|\boldsymbol{X}_0\|^2$ and $\|\boldsymbol{Y}_0\|^2$ by adjusting η up to some absolute constant.

$$\leq \|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}} + 2\sigma_{\max}^{1/2}\zeta\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}},$$

where we use $\|\boldsymbol{\Lambda}_t - \boldsymbol{I}_r\| \leq \zeta$ and $\|\widetilde{\boldsymbol{Y}}_t\| \leq 2\sqrt{\sigma_{\max}}$. Combine the previous two bounds and apply the basic inequality $2ab \leq a^2 + b^2$ to see

$$\begin{aligned}
&|\widetilde{\gamma}_1 - \gamma_1| \\
&\quad\leq \delta_{2r}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2 + 2\sigma_{\max}^{1/2}\zeta\delta_{2r}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}} \\
&\qquad + \delta_{2r}\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}} \\
&\qquad + 2\sigma_{\max}^{1/2}\zeta\delta_{2r}\|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}} \\
&\qquad + \delta_{2r}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}} \\
&\qquad + 2\sigma_{\max}^{1/2}\zeta\delta_{2r}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\|_{\mathrm{F}} \\
&\lesssim \delta_{2r}\left(\|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2 + \|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2 + \sigma_{\min}\mathrm{dist}\left(\boldsymbol{Z}_t, \boldsymbol{Z}_\star\right)\right), \\
&\ll \|\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}\widetilde{\boldsymbol{Y}}_t^\top\|_{\mathrm{F}}^2 + \|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2 + \sigma_{\min}\mathrm{dist}\left(\boldsymbol{Z}_t, \boldsymbol{Z}_\star\right)
\end{aligned}$$

as long as δ_{2r} is sufficiently small. The same bound applies to $|\widetilde{\gamma}_2 - \gamma_2|$. As a result, as long as δ_{2r} is small enough, $\widetilde{\gamma}_1 + \widetilde{\gamma}_2$ is lower bounded on the same order as $\gamma_1 + \gamma_2$, say

$$\begin{aligned}
\widetilde{\gamma}_1 + \widetilde{\gamma}_2 \geq \frac{1}{2}\left(\|\widetilde{\boldsymbol{Y}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{X}_t}^\top\|_{\mathrm{F}}^2 + \|\widetilde{\boldsymbol{X}}_t\widetilde{\boldsymbol{E}}_{\boldsymbol{Y}_t}^\top\|_{\mathrm{F}}^2\right) \\
- \frac{\sigma_{\min}}{6}\mathrm{dist}^2\left(\boldsymbol{Z}_t, \boldsymbol{Z}_\star\right).
\end{aligned}$$

One can then repeat the same arguments for the matrix factorization case to obtain the linear convergence. For the sake of space, we omit it.

V. Conclusions

This paper establishes the local linear convergence of gradient descent for asymmetric low-rank matrix sensing without explicit regularization of factor balancedness under the standard RIP assumption, as long as a balanced initialization is provided in the basin of attraction. Coupled with the standard spectral initialization, this leads to the global convergence guarantee of the balancing-free gradient descent algorithm for asymmetric low-rank matrix sensing. Different from previous work, we analyzed a new error metric that takes into account the ambiguity due to invertible transforms, and showed that it contracts linearly even without local restricted strong convexity. We believe that our technique can be used for other low-rank matrix estimation problems. To conclude, we outline a few future research directions.

- *Low-rank matrix completion*. We believe it is possible to extend our analysis to study rectangular matrix completion without regularization, by combining the leave-one-out technique in [27], [35] to carefully bound the incoherence of the iterates for both factors even without explicit balancing.
- *Improving dependence on κ and r*. The current paper does not try to optimize the dependence with respect to κ and r in terms of sample complexity and the size of the basin of attraction, which are slightly worse than their regularized counterparts. A finer analysis will likely lead to better dependencies, which we leave to the future work.

Appendix

A. Proof of Lemma 1

For notational convenience, we define the following function

$$g\left(\boldsymbol{Q}\right) \triangleq \|\boldsymbol{X}\boldsymbol{Q} - \boldsymbol{X}_\star\|_{\mathrm{F}}^2 + \left\|\boldsymbol{Y}\boldsymbol{Q}^{-\top} - \boldsymbol{Y}_\star\right\|_{\mathrm{F}}^2. \tag{22}$$

Clearly, the optimal alignment matrix, if exists, must be $\mathrm{argmin}\ g(\boldsymbol{P})$. With this notation in place, we consider the following constrained minimization problem:

$$\begin{aligned}
\min_{\boldsymbol{Q}\in\mathbb{R}^{r\times r}:\boldsymbol{Q}\text{ is invertible}} \quad & g\left(\boldsymbol{Q}\right) \\
\text{subject to} \quad & \|\boldsymbol{Q} - \boldsymbol{P}\|_{\mathrm{F}} \leq \frac{5\delta}{\sigma_{\min}\left(\boldsymbol{X}_\star\right)}.
\end{aligned}$$

In view of Weyl's inequality, we obtain that for any feasible $\boldsymbol{Q}$,

$$\sigma_{\min}\left(\boldsymbol{Q}\right) \geq \sigma_{\min}\left(\boldsymbol{P}\right) - \frac{5\delta}{\sigma_{\min}\left(\boldsymbol{X}_\star\right)} \geq \frac{1}{2} - \frac{1}{4} = \frac{1}{4}$$

as long as $\delta \leq \sigma_{\min}(\boldsymbol{X}_\star)/80$. As a result, one sees that $g(\boldsymbol{Q})$ is a continuous function over $\{\boldsymbol{Q} : \|\boldsymbol{Q} - \boldsymbol{P}\| \leq 5\delta/\sigma_{\min}(\boldsymbol{X}_\star)\}$, which is a compact set over invertible matrices. Applying the Weierstrass extreme value theorem yields the claim that the minimizer of the constrained problem exists. Denote this minimizer by $\boldsymbol{Q}_1$. In what follows, we intend to show that $\boldsymbol{Q}_1$ is also the minimizer of the unconstrained problem. Letting $\boldsymbol{Q}$ be an arbitrary matrix with $g(\boldsymbol{Q}) \leq 2\delta^2$ (the existence is assured since $g(\boldsymbol{Q}_1) \leq g(\boldsymbol{P}) \leq 2\delta^2$), we have

$$\sqrt{2}\delta \geq \|\boldsymbol{X}\boldsymbol{Q} - \boldsymbol{X}_\star\|_{\mathrm{F}} \geq \|\boldsymbol{X}\boldsymbol{Q} - \boldsymbol{X}\boldsymbol{P}\|_{\mathrm{F}} - \|\boldsymbol{X}\boldsymbol{P} - \boldsymbol{X}_\star\|_{\mathrm{F}},$$

which in conjunction with (12) implies

$$(1 + \sqrt{2})\delta \geq \|\boldsymbol{X}(\boldsymbol{Q} - \boldsymbol{P})\|_{\mathrm{F}} \geq \sigma_{\min}\left(\boldsymbol{X}\right)\|\boldsymbol{Q}_1 - \boldsymbol{P}\|_{\mathrm{F}}. \tag{23}$$

We now turn to investigating $\sigma_{\min}(\boldsymbol{X})$. Weyl's inequality tells us that

$$\begin{aligned}
|\sigma_{\min}\left(\boldsymbol{X}\boldsymbol{P}\right) - \sigma_{\min}\left(\boldsymbol{X}_\star\right)| &\leq \|\boldsymbol{X}\boldsymbol{P} - \boldsymbol{X}_\star\|_{\mathrm{F}} \\
&\leq \delta \leq \frac{1}{4}\sigma_{\min}\left(\boldsymbol{X}_\star\right),
\end{aligned}$$

which further implies

$$\begin{aligned}
\frac{3}{4}\sigma_{\min}\left(\boldsymbol{X}_\star\right) &\leq \sigma_{\min}\left(\boldsymbol{X}\boldsymbol{P}\right) \\
&\leq \sigma_{\min}\left(\boldsymbol{X}\right)\sigma_{\max}\left(\boldsymbol{P}\right) \leq \frac{3}{2}\sigma_{\min}\left(\boldsymbol{X}\right).
\end{aligned}$$

Therefore we arrive at $\sigma_{\min}(\boldsymbol{X}) \geq \sigma_{\min}(\boldsymbol{X}_\star)/2$. Putting this back to (23) yields which finally gives

$$\|\boldsymbol{Q} - \boldsymbol{P}\|_{\mathrm{F}} \leq 2(1 + \sqrt{2})\frac{\delta}{\sigma_{\min}(\boldsymbol{X}_\star)} < \frac{5\delta}{\sigma_{\min}(\boldsymbol{X}_\star)}.$$

In all, the above arguments reveal that any matrix $\boldsymbol{Q}$ such that $g(\boldsymbol{Q}) \leq 2\delta^2$ must obey the above bound. Therefore the minimizer of the constrained problem and that of the unconstrained one coincide with each other. This finished the proof.

B. Proof of Lemma 2

Recall the function $g\left(\boldsymbol{P}\right)$ defined in (22) as

$$\begin{aligned} g\left(\boldsymbol{P}\right) &= \left\|\boldsymbol{X}\boldsymbol{P}-\boldsymbol{X}_{\star}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{Y}\boldsymbol{P}^{-\top}-\boldsymbol{Y}_{\star}\right\|_{\mathrm{F}}^{2} \\ &= \mathrm{Tr}\left(\boldsymbol{X}\boldsymbol{P}\boldsymbol{P}^{\top}\boldsymbol{X}^{\top}\right)-2\mathrm{Tr}\left(\boldsymbol{P}^{\top}\boldsymbol{X}^{\top}\boldsymbol{X}_{\star}\right) \\ &\quad +\mathrm{Tr}\left(\boldsymbol{X}_{\star}^{\top}\boldsymbol{X}_{\star}\right)+\mathrm{Tr}\left(\boldsymbol{P}^{-1}\boldsymbol{Y}^{\top}\boldsymbol{Y}\boldsymbol{P}^{-\top}\right) \\ &\quad -2\mathrm{Tr}\left(\boldsymbol{P}^{-1}\boldsymbol{Y}^{\top}\boldsymbol{Y}_{\star}\right)+\mathrm{Tr}\left(\boldsymbol{Y}_{\star}^{\top}\boldsymbol{Y}_{\star}\right). \end{aligned}$$

The gradient is given by

$$\begin{aligned} \nabla g\left(\boldsymbol{P}\right) &= 2\boldsymbol{X}^{\top}\boldsymbol{X}\boldsymbol{P}-2\boldsymbol{X}^{\top}\boldsymbol{X}_{\star} \\ &\quad -2\left(\boldsymbol{P}\boldsymbol{P}^{\top}\right)^{-1}\boldsymbol{Y}^{\top}\boldsymbol{Y}\left(\boldsymbol{P}\boldsymbol{P}^{\top}\right)^{-1}\boldsymbol{P}+2\boldsymbol{P}^{-\top}\boldsymbol{Y}_{\star}^{\top}\boldsymbol{Y}\boldsymbol{P}^{-\top}. \end{aligned}$$

Since $\boldsymbol{Q}$ minimizes $g(\boldsymbol{P})$, it must satisfy the first-order optimality condition, i.e.

$$\nabla g\left(\boldsymbol{Q}\right)=\boldsymbol{0}.$$

Identify $\widetilde{\boldsymbol{X}}=\boldsymbol{X}\boldsymbol{Q}$ and $\widetilde{\boldsymbol{Y}}=\boldsymbol{Y}\boldsymbol{Q}^{-\top}$ to yield the condition

$$\widetilde{\boldsymbol{X}}^{\top}\widetilde{\boldsymbol{X}}-\widetilde{\boldsymbol{X}}^{\top}\boldsymbol{X}_{\star}=\widetilde{\boldsymbol{Y}}^{\top}\widetilde{\boldsymbol{Y}}-\boldsymbol{Y}_{\star}^{\top}\widetilde{\boldsymbol{Y}}.$$

C. Proof of Lemma 3

We prove the first part and the second part follows by symmetry. Denote $\boldsymbol{E}_{x}=\boldsymbol{X}-\boldsymbol{X}_{\star}$ and $\boldsymbol{E}_{y}=\boldsymbol{Y}-\boldsymbol{Y}_{\star}$. We have

$$\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{M}_{\star}=\boldsymbol{E}_{x}\boldsymbol{Y}^{\top}+\boldsymbol{X}_{\star}\boldsymbol{E}_{y}^{\top}.$$

Since $\boldsymbol{Z}$ is aligned with $\boldsymbol{Z}_{\star}$, Lemma 2 tells us that

$$\boldsymbol{X}^{\top}\boldsymbol{E}_{x}=\boldsymbol{E}_{y}^{\top}\boldsymbol{Y}.$$

As a result, one has

$$\begin{aligned} &\left\langle \boldsymbol{X}-\boldsymbol{X}_{\star},\left(\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{M}_{\star}\right)\boldsymbol{Y}\right\rangle \\ &\quad =\mathrm{Tr}\left(\boldsymbol{E}_{x}^{\top}\left(\boldsymbol{E}_{x}\boldsymbol{Y}^{\top}+\boldsymbol{X}_{\star}\boldsymbol{E}_{y}^{\top}\right)\boldsymbol{Y}\right) \\ &\quad =\mathrm{Tr}\left(\boldsymbol{E}_{x}^{\top}\boldsymbol{E}_{x}\boldsymbol{Y}^{\top}\boldsymbol{Y}\right)+\mathrm{Tr}\left(\boldsymbol{E}_{x}^{\top}\boldsymbol{X}_{\star}\boldsymbol{E}_{y}^{\top}\boldsymbol{Y}\right) \\ &\quad =\left\|\boldsymbol{Y}\boldsymbol{E}_{x}^{\top}\right\|_{\mathrm{F}}^{2}+\mathrm{Tr}\left(\boldsymbol{E}_{y}^{\top}\boldsymbol{Y}\boldsymbol{E}_{x}^{\top}\boldsymbol{X}\right)-\mathrm{Tr}\left(\boldsymbol{E}_{y}^{\top}\boldsymbol{Y}\boldsymbol{E}_{x}^{\top}\boldsymbol{E}_{x}\right) \\ &\quad =\left\|\boldsymbol{Y}\boldsymbol{E}_{x}^{\top}\right\|_{\mathrm{F}}^{2}+\left\|\boldsymbol{X}^{\top}\boldsymbol{E}_{x}\right\|_{\mathrm{F}}^{2}-\mathrm{Tr}\left(\boldsymbol{X}^{\top}\boldsymbol{E}_{x}\boldsymbol{E}_{x}^{\top}\boldsymbol{E}_{x}\right). \end{aligned} \tag{24}$$

Complete the squares to see that

$$\begin{aligned} &\left\|\boldsymbol{X}^{\top}\boldsymbol{E}_{x}\right\|_{\mathrm{F}}^{2}-\mathrm{Tr}\left(\boldsymbol{X}^{\top}\boldsymbol{E}_{x}\boldsymbol{E}_{x}^{\top}\boldsymbol{E}_{x}\right) \\ &\qquad =\left\|\boldsymbol{E}_{x}^{\top}\boldsymbol{X}-\frac{1}{2}\boldsymbol{E}_{x}^{\top}\boldsymbol{E}_{x}\right\|_{\mathrm{F}}^{2}-\frac{1}{4}\left\|\boldsymbol{E}_{x}\right\|_{\mathrm{F}}^{4}. \end{aligned}$$

Combine the previous two bounds to yield the desired result.

D. Proof of Lemma 4

Again, we demonstrate the claim on $\boldsymbol{X}$ and the claim on $\boldsymbol{Y}$ follows by symmetry. Given the decomposition

$$\begin{aligned} \boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{M}_{\star} &= (\boldsymbol{X}-\boldsymbol{X}_{\star})\boldsymbol{Y}^{\top}+\boldsymbol{X}(\boldsymbol{Y}-\boldsymbol{Y}_{\star})^{\top} \\ &\quad +(\boldsymbol{X}_{\star}-\boldsymbol{X})(\boldsymbol{Y}-\boldsymbol{Y}_{\star})^{\top}, \end{aligned}$$

we obtain

$$\begin{aligned} &\left\|\left(\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{M}_{\star}\right)\boldsymbol{Y}\right\|_{\mathrm{F}}\leq\sigma_{1}(\boldsymbol{Y})\left\|\boldsymbol{X}\boldsymbol{Y}^{\top}-\boldsymbol{M}_{\star}\right\|_{\mathrm{F}} \\ &\quad \leq\frac{3}{2}\sigma_{1}(\boldsymbol{Y}_{\star})\Big(\left\|(\boldsymbol{X}-\boldsymbol{X}_{\star})\boldsymbol{Y}^{\top}\right\|_{\mathrm{F}} \\ &\qquad +\left\|\boldsymbol{X}(\boldsymbol{Y}-\boldsymbol{Y}_{\star})^{\top}\right\|_{\mathrm{F}}+\left\|(\boldsymbol{X}_{\star}-\boldsymbol{X})(\boldsymbol{Y}-\boldsymbol{Y}_{\star})^{\top}\right\|_{\mathrm{F}}\Big), \end{aligned}$$

where the last line combines the triangle inequality and Weyl's inequality

$$\sigma_{1}\left(\boldsymbol{Y}\right)\leq\sigma_{1}(\boldsymbol{Y}_{\star})+\left\|\boldsymbol{Y}-\boldsymbol{Y}_{\star}\right\|\leq\frac{3}{2}\sigma_{1}\left(\boldsymbol{Y}_{\star}\right).$$

The proof is then finished.

REFERENCES

[1] C. Ma, Y. Li, and Y. Chi, "Beyond Procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*. IEEE, 2019, pp. 721–725.

[2] E. Candès and T. Tao, "The power of convex relaxation: Near-optimal matrix completion," *IEEE Transactions on Information Theory*, vol. 56, no. 5, pp. 2053 –2080, May 2010.

[3] Y. Chen and Y. Chi, "Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization," *IEEE Signal Processing Magazine*, vol. 35, no. 4, pp. 14 – 31, 2018.

[4] M. A. Davenport and J. Romberg, "An overview of low-rank matrix recovery from incomplete observations," *IEEE Journal of Selected Topics in Signal Processing*, vol. 10, no. 4, pp. 608–622, 2016.

[5] S. Burer and R. Monteiro, "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization," *Mathematical Programming*, vol. 95, no. 2, pp. 329–357, 2003.

[6] S. Bhojanapalli, A. Kyrillidis, and S. Sanghavi, "Dropping convexity for faster semi-definite optimization," in *Conference on Learning Theory*, 2016, pp. 530–582.

[7] N. Boumal, V. Voroninski, and A. Bandeira, "The non-convex Burer-Monteiro approach works on smooth semidefinite programs," in *Advances in Neural Information Processing Systems*, 2016, pp. 2757–2765.

[8] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via procrustes flow," in *International Conference Machine Learning*, 2016, pp. 964–973.

[9] Q. Zheng and J. Lafferty, "Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent," *arXiv preprint arXiv:1605.07051*, 2016.

[10] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, "Finding low-rank solutions via nonconvex matrix factorization, efficiently and provably," *SIAM Journal on Imaging Sciences*, vol. 11, no. 4, pp. 2165–2204, 2018.

[11] Y. Chi, "Low-rank matrix completion," *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 178–181, 2018.

[12] X. Yi, D. Park, Y. Chen, and C. Caramanis, "Fast algorithms for robust PCA via gradient descent," in *Advances in neural information processing systems*, 2016, pp. 4152–4160.

[13] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Spectral methods for data science: A statistical perspective," *Foundations and Trends in Machine Learning*, 2020, preprint.

[14] B. Recht, M. Fazel, and P. A. Parrilo, "Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization," *SIAM review*, vol. 52, no. 3, pp. 471–501, 2010.

[15] S. Oymak, B. Recht, and M. Soltanolkotabi, "Sharp time–data tradeoffs for linear inverse problems," *IEEE Transactions on Information Theory*, vol. 64, no. 6, pp. 4129–4158, 2018.

[16] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Foundations of Computational Mathematics*, vol. 9, no. 6, pp. 717–772, 2009.

[17] E. J. Candès and Y. Plan, "Matrix completion with noise," *Proceedings of the IEEE*, vol. 98, no. 6, pp. 925 –936, June 2010.

[18] S. Negahban and M. Wainwright, "Restricted strong convexity and weighted matrix completion: Optimal bounds with noise," *The Journal of Machine Learning Research*, vol. 98888, pp. 1665–1697, May 2012.

[19] D. Gross, "Recovering low-rank matrices from few coefficients in any basis," *IEEE Transactions on Information Theory*, vol. 57, no. 3, pp. 1548–1566, March 2011.

[20] B. Recht, "A simpler approach to matrix completion," *Journal of Machine Learning Research*, vol. 12, pp. 3413–3430, Feburary 2011.

[21] Y. Chen and Y. Chi, "Robust spectral compressed sensing via structured matrix completion," *IEEE Transactions on Information Theory*, vol. 60, no. 10, pp. 6576–6601, 2014.

[22] S. Negahban and M. J. Wainwright, "Estimation of (near) low-rank matrices with noise and high-dimensional scaling," *The Annals of Statistics*, vol. 39, no. 2, pp. 1069–1097, 2011.

[23] Q. Zheng and J. Lafferty, "A convergent gradient descent algorithm for rank minimization and semidefinite programming from random linear measurements," in *Advances in Neural Information Processing Systems*, 2015, pp. 109–117.

[24] R. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Transactions on Information Theory*, vol. 56, no. 6, pp. 2980–2998, 2010.

[25] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via nonconvex factorization," in *Symposium on Foundations of Computer Science (FOCS)*. IEEE, 2015, pp. 270–289.

[26] P. Jain, P. Netrapalli, and S. Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*. ACM, 2013, pp. 665–674.

[27] C. Ma, K. Wang, Y. Chi, and Y. Chen, "Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution," *Foundations of Computational Mathematics*, pp. 1–182, 2019.

[28] Y. Chen and M. J. Wainwright, "Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees," *arXiv preprint arXiv:1509.03025*, 2015.

[29] Y. Li, C. Ma, Y. Chen, and Y. Chi, "Nonconvex matrix factorization from rank-one measurements," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 1496–1505.

[30] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Applied and computational harmonic analysis*, vol. 47, no. 3, pp. 893–934, 2019.

[31] Y. Chen, Y. Chi, J. Fan, and C. Ma, "Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval," *Mathematical Programming*, pp. 1–33, 2018.

[32] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *IEEE Transactions on Signal Processing*, vol. 67, no. 20, pp. 5239–5269, 2019.

[33] Y. Li, Y. Chi, H. Zhang, and Y. Liang, "Non-convex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent," *Information and Inference: A Journal of the IMA*, vol. 9, no. 2, pp. 289–325, 2020.

[34] X. Zhang, S. Du, and Q. Gu, "Fast and sample efficient inductive matrix completion via multi-phase procrustes flow," in *International Conference on Machine Learning*, 2018, pp. 5751–5760.

[35] J. Chen, D. Liu, and X. Li, "Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization," *IEEE Transactions on Information Theory*, vol. 66, no. 9, pp. 5806–5841, 2020.

[36] S. S. Du, W. Hu, and J. D. Lee, "Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced," in *Advances in Neural Information Processing Systems*, 2018, pp. 384–395.

[37] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy, "Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence," *arXiv preprint arXiv:1904.10020*, 2019.

[38] R. Ge, J. D. Lee, and T. Ma, "Matrix completion has no spurious local minimum," in *Advances in Neural Information Processing Systems*, 2016, pp. 2973–2981.

[39] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *International Conference on Machine Learning*, 2017, pp. 1233–1242.

[40] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "Global optimality in low-rank matrix optimization," *IEEE Transactions on Signal Processing*, vol. 66, no. 13, pp. 3614–3628, 2018.

[41] Z. Zhu, D. Soudry, Y. C. Eldar, and M. B. Wakin, "The global optimization geometry of shallow linear neural networks," *Journal of Mathematical Imaging and Vision*, pp. 1–14, 2019.

[42] X. Li, J. Lu, R. Arora, J. Haupt, H. Liu, Z. Wang, and T. Zhao, "Symmetry, saddle points, and global optimization landscape of nonconvex matrix factorization," *IEEE Transactions on Information Theory*, vol. 65, no. 6, pp. 3489–3514, 2019.

[43] Q. Li, Z. Zhu, and G. Tang, "The non-convex geometry of low-rank matrix optimization," *Information and Inference: A Journal of the IMA*, vol. 8, no. 1, pp. 51–96, 2019.

[44] S. Li, Q. Li, Z. Zhu, G. Tang, and M. B. Wakin, "The global geometry of centralized and distributed low-rank matrix recovery without regularization," *IEEE Signal Processing Letters*, vol. 27, pp. 1400–1404, 2020.

[45] T. Tong, C. Ma, and Y. Chi, "Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent," *arXiv preprint arXiv:2005.08898*, 2020.

[46] ——, "Low-rank matrix recovery with scaled subgradient methods: Fast and robust convergence without the condition number," *arXiv preprint arXiv:2010.13364*, 2020.

Cong Ma received the B.Eng. degree from Tsinghua University in 2015, and earned his Ph.D. degree from Princeton University in 2020. He is currently a postdoctoral researcher in the Department of Electrical Engineering and Computer Sciences at UC Berkeley, and will be joining the Department of Statistics at the University of Chicago as an assistant professor in July 2021. His research interests include mathematics of data science, machine learning, high-dimensional statistics, convex and nonconvex optimization as well as their applications to neuroscience. Dr. Ma has received the School of Engineering and Applied Science Award for Excellence from Princeton University in 2019, the AI Labs Fellowship from Hudson River Trading in 2019, and the Student Paper Award from International Chinese Statistical Association in 2017.

Yuanxin Li received the B.Eng. degree from Nanjing University of Posts and Telecommunications, Nanjing, China, in 2010, the M.Eng. degree from Tsinghua University, Beijing, China, in 2013, the M.S. degree as well as the Ph.D. degree from The Ohio State University, Columbus, OH, USA, in 2016 and 2018, respectively. He was a visiting student and subsequently a post-doctoral researcher in the Department of Electrical and Computer Engineering at Carnegie Mellon University in 2018. He is currently with Samsung Semiconductor, Inc. as Senior Engineer in San Diego, CA. His research interests include statistical signal processing, convex optimization, computer vision and data analysis.

Yuejie Chi (S'09–M'12–SM'17) received Ph.D. and M.A. in Electrical Engineering from Princeton University in 2012 and 2009, and B.E. (Hon.) in Electrical Engineering from Tsinghua University, Beijing, China, in 2007. She was with The Ohio State University from 2012 to 2017. Since 2018, she is an Associate Professor with the department of Electrical and Computer Engineering at Carnegie Mellon University, where she holds the Robert E. Doherty Early Career Development Professorship. Her research interests lie in the theoretical and algorithmic foundations of data science, signal processing, machine learning and inverse problems, with applications in sensing systems, broadly defined. Among others, she is a recipient of Presidential Early Career Award for Scientists and Engineers (PECASE), the inaugural IEEE Signal Processing Society Early Career Technical Achievement Award for contributions to high-dimensional structured signal processing, and named the 2021 Goldsmith Lecturer by IEEE Information Theory Society. She currently serves as an Associate Editor for IEEE Trans. on Signal Processing and IEEE Trans. on Pattern Recognition and Machine Intelligence.