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

Cong Ma[1], Yuanxin Li[2], Yuejie Chi[2]
[1]Department of Operations Research and Financial Engineering, Princeton University
[2]Department of Electrical and Computer Engineering, Carnegie Mellon University

***Abstract*—Low-rank matrix estimation plays a central role in many 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 two matrix factors which in practice can be removed safely without hurting the performance when initialized via the spectral method. In this paper, we justify this theoretically 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 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 [1], [2], [3]. Broadly speaking, we are interested in estimating a low-rank matrix $\boldsymbol{M}_\natural = \boldsymbol{X}_\natural \boldsymbol{Y}_\natural^\top \in \mathbb{R}^{n_1 \times n_2}$ 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 the rank $r \ll n := \min\{n_1, n_2\}$ is much smaller than the dimensions of the matrix. To reduce computational complexity, a common approach, popularized by the work of Burer and Monteiro [4], is to factorize $\boldsymbol{M} = \boldsymbol{X}\boldsymbol{Y}^\top$ where $\boldsymbol{X} \in \mathbb{R}^{n_1 \times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{n_2 \times r}$, and rewrite the above problem into an unconstrained nonconvex optimization problem:

$$\min_{\boldsymbol{X}, \boldsymbol{Y}} \ f(\boldsymbol{X}, \boldsymbol{Y}) = \mathcal{L}(\boldsymbol{X}\boldsymbol{Y}^\top). \tag{2}$$

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

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

This work is supported in part by ONR under the grant N00014-18-1-2142, by ARO under the grant W911NF-18-1-0303, and by NSF under the grants CAREER ECCS-1818571 and CCF-1806154. Part of this work was performed with C. Ma was visiting CMU.

where η_t is the step size and $(\boldsymbol{X}_0, \boldsymbol{Y}_0)$ is some proper initialization.

Significant progress has been made recently in understanding the performance of gradient descent for nonconvex matrix factorization. Somewhat surprisingly, most of the existing guarantees are not directly applicable to the vanilla gradient descent rule (3). One challenge is associated with the identifiability of the factors, since they are indistinguishable as long as their product is the same – and if the norms of the factors become highly imbalanced, gradient descent might diverge easily. Consequently, it becomes a routine procedure to insert a regularizer that balances the two factors [5], [6], [7]:

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

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

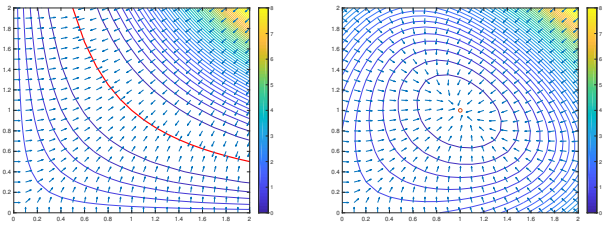
$$\min_{\boldsymbol{X}, \boldsymbol{Y}} \ f_{\mathrm{reg}}(\boldsymbol{X}, \boldsymbol{Y}) := f(\boldsymbol{X}, \boldsymbol{Y}) + g(\boldsymbol{X}, \boldsymbol{Y}). \tag{5}$$

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?

To handle such asymmetric factorization, it is common to stack the two factors into one augmented factor $\boldsymbol{W}_\natural = \begin{bmatrix} \boldsymbol{X}_\natural \\ \boldsymbol{Y}_\natural \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)r}$ and then seek to estimate $\boldsymbol{W}_\natural$ directly, by rewriting the loss function with respect to the lifted low-rank matrix: $\boldsymbol{W}_\natural \boldsymbol{W}_\natural^\top = \begin{bmatrix} \boldsymbol{X}_\natural \boldsymbol{X}_\natural^\top & \boldsymbol{X}_\natural \boldsymbol{Y}_\natural^\top \\ \boldsymbol{Y}_\natural \boldsymbol{X}_\natural^\top & \boldsymbol{Y}_\natural \boldsymbol{Y}_\natural^\top \end{bmatrix} \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$. It is obvious that the loss function originally with respect to the asymmetric matrix $\boldsymbol{X}_\natural \boldsymbol{Y}_\natural^\top$ only constrains the off-diagonal blocks of $\boldsymbol{W}_\natural \boldsymbol{W}_\natural^\top$ and not the diagonal ones; correspondingly, the loss function is not (restricted) strongly convex with respect to the augmented factor, unless we appropriately regularize the diagonal blocks, which gives rise to the adoption of the regularization term in (4).

To understand a bit better why this regularization term (4) may help analysis, consider a toy example of factorizing a rank-one matrix $\boldsymbol{x}_\natural \boldsymbol{y}_\natural^\top$, where $f(\boldsymbol{x}, \boldsymbol{y})$ and $g(\boldsymbol{x}, \boldsymbol{y})$ respectively are $f(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{x}\boldsymbol{y}^\top - \boldsymbol{x}_\natural \boldsymbol{y}_\natural^\top\|_{\mathrm{F}}^2$ and $g(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{8}(\|\boldsymbol{x}\|_2^2 - \|\boldsymbol{y}\|_2^2)^2$. Figure 1 illustrates the landscape of the unregularized loss function $f(x, y)$ and the regularized loss function $f_{\mathrm{reg}}(x, y)$,



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

Fig. 1. The geometry for the scalar case $f(x,y) = (xy-1)^2$ and $g(x,y) = (x^2-y^2)/8$. 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).

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 whenever $xy = 1$, making the analysis less tractable.

B. This paper: balancing-free procedure?

This goal of this paper is to understand the effectiveness of vanilla gradient descent (3) when initialized with balanced factors via the spectral method. Indeed, Figure 2 plots the normalized error $\|\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\natural\|_{\mathrm{F}}/\|\boldsymbol{M}_\natural\|_{\mathrm{F}}$ for low-rank matrix completion 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 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 regularization parameters.

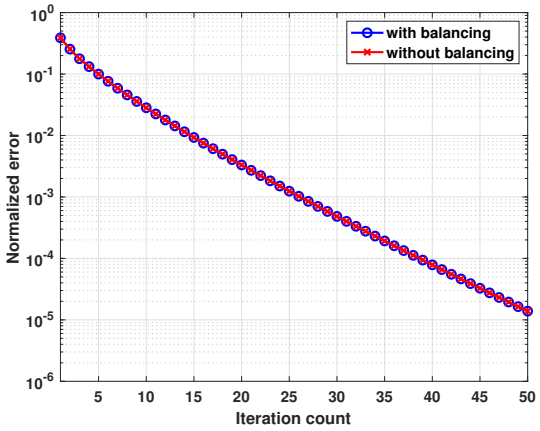


Fig. 2. The normalized reconstruction error $\|\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\natural\|_{\mathrm{F}}/\|\boldsymbol{M}_\natural\|_{\mathrm{F}}$ with respect to the iteration count, for completing a rank-10 1000×1000 matrix where each entry is observed i.i.d. with probability $p = 0.15$. The balancing regularizer is $g(\boldsymbol{X},\boldsymbol{Y}) = \frac{1}{64}\|\boldsymbol{X}^\top\boldsymbol{X} - \boldsymbol{Y}^\top\boldsymbol{Y}\|_{\mathrm{F}}^2$ as suggested in [8].

This paper justifies formally that even without explicit balancing in non-square matrix factorization, gradient descent converges linearly to the global optimum, as long as the initialization is balanced for low-rank matrix sensing, where the goal is to recover a low-rank matrix from a small number of linear measurements. 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 main message is this: as long as the factors are balanced at the initialization, they will stay approximately balanced throughout the trajectory of gradient descent, and therefore no additional regularization is necessary.

C. Notations and organization of this paper

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 $\mathrm{Tr}\left(\cdot\right)$ is the trace. Denote $\mathcal{O}^{r\times r}$ as 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 $\boldsymbol{M}_\natural \in \mathbb{R}^{n_1\times n_2}$ be a rank-r matrix with the Singular Value Decomposition (SVD) given as

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

where $\boldsymbol{U}_\natural \in \mathbb{R}^{n_1\times r}$, $\boldsymbol{V}_\natural \in \mathbb{R}^{n_2\times r}$ and $\boldsymbol{\Sigma}_\natural \in \mathbb{R}^{r\times r}$. Without loss of generality, we denote the ground truth factors as

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

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

Since the factors are identifiable up to invertible transforms since $(\boldsymbol{X}_\natural\boldsymbol{P})(\boldsymbol{Y}_\natural\boldsymbol{P}^{-\top})^\top = \boldsymbol{X}_\natural\boldsymbol{Y}_\natural^\top$ for any invertible matrix $\boldsymbol{P} \in \mathbb{R}^{r\times r}$, we measure the distance between two pairs of factors $\boldsymbol{Z} = (\boldsymbol{X},\boldsymbol{Y})$ and $\boldsymbol{Z}_\natural = (\boldsymbol{X}_\natural,\boldsymbol{Y}_\natural)$ as:

$$\mathrm{dist}\left(\boldsymbol{Z},\boldsymbol{Z}_\natural\right) = \min_{\boldsymbol{P}\in\mathbb{R}^{r\times r},\text{invertible}} \sqrt{\|\boldsymbol{X}\boldsymbol{P} - \boldsymbol{X}_\natural\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\boldsymbol{P}^{-\top} - \boldsymbol{Y}_\natural\|_{\mathrm{F}}^2}. \tag{7}$$

A. Low-rank matrix sensing

Suppose we are given a set of m measurements as follows

$$y_i = \langle \boldsymbol{A}_i, \boldsymbol{M}_\natural\rangle = \langle \boldsymbol{A}_i, \boldsymbol{X}_\natural\boldsymbol{Y}_\natural^\top\rangle, \qquad i = 1,\cdots,m, \tag{8}$$

where $\boldsymbol{A}_i \in \mathbb{R}^{n_1\times n_2}$ is the ith sensing matrix, $i = 1,\cdots,m$. For convenience, we define $\mathcal{A}: \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^m$ as an affine

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

Input: Measurements $\boldsymbol{y} = \{y_i\}_{i=1}^m$, and sensing matrices $\{\boldsymbol{A}_i\}_{i=1}^m$.
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 $\boldsymbol{K} = \frac{1}{m}\sum_{i=1}^m y_i \boldsymbol{A}_i$.
Gradient loop: For $t = 0 : 1 : T-1$, do

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t - \frac{\eta_t}{\|\boldsymbol{Y}_0\|^2} \cdot \left[\sum_{i=1}^m \left(\langle \boldsymbol{A}_i, \boldsymbol{X}_t\boldsymbol{Y}_t^\top\rangle - y_i\right)\boldsymbol{A}_i\boldsymbol{Y}_t\right]; \tag{10a}$$

$$\boldsymbol{Y}_{t+1} = \boldsymbol{Y}_t - \frac{\eta_t}{\|\boldsymbol{X}_0\|^2} \cdot \left[\sum_{i=1}^m \left(\langle \boldsymbol{A}_i, \boldsymbol{X}_t\boldsymbol{Y}_t^\top\rangle - y_i\right)\boldsymbol{A}_i^\top\boldsymbol{X}_t\right]. \tag{10b}$$

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

transformation from $\mathbb{R}^{n_1\times n_2}$ to $\mathbb{R}^m$, such that $\mathcal{A}(\boldsymbol{M}) = \{\langle \boldsymbol{A}_i, \boldsymbol{M}\rangle\}_{i=1}^m$. Consequently, one can write $\boldsymbol{y} = \mathcal{A}(\boldsymbol{M}_\natural)$. The adjoint operator $\mathcal{A}^*$ 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 squared loss function

$$f(\boldsymbol{X}, \boldsymbol{Y}) := \frac{1}{2}\left\|\boldsymbol{y} - \mathcal{A}(\boldsymbol{X}\boldsymbol{Y}^\top)\right\|_2^2. \tag{9}$$

Algorithm 1 describes the gradient descent algorithm initialized by the spectral method for minimizing (9). Compared to the Procrustes Flow (PF) algorithm [5], which minimizes the regularized loss function in (5), the new algorithm does not include the balancing regularizer $g(\boldsymbol{X}, \boldsymbol{Y})$.

B. Theoretical Guarantees

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

Definition 1 (RIP): The mapping operation $\mathcal{A}$ is said to satisfy the rank-r RIP with constant δ_r, 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 [9]. For example, if the entries of $\boldsymbol{A}_i$'s are composed of i.i.d. Gaussian entries $\mathcal{N}(0, 1/m)$, then the RIP is satisfied as long as m is on the order of $(n_1+n_2)r/\delta_r^2$.

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

Theorem 1: Suppose $\mathcal{A}$ satisfies the RIP with $\delta_{2r} \le c$ for some sufficiently small constant c. Let $\boldsymbol{Z}_0 = (\boldsymbol{X}_0, \boldsymbol{Y}_0)$ be an initialization which satisfies

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{Z}_0\boldsymbol{R} - \boldsymbol{Z}_\natural\|_{\mathrm{F}} \le c_0 \frac{1}{\kappa^{3/2}}\sigma_{\min}(\boldsymbol{X}_\natural), \tag{11}$$

for some small enough constant c_0. There exist some c_1 such that at as long as $\eta_t = \eta \le c_1$, the iterates of GD satisfy

$$\operatorname{dist}(\boldsymbol{Z}_t, \boldsymbol{Z}_\natural) \le \left(1 - \frac{\eta}{20\kappa}\right)^t \operatorname{dist}(\boldsymbol{Z}_0, \boldsymbol{Z}_\natural).$$

Theorem 1 says that if the initialization $\boldsymbol{Z}_0$ lands in a basin of attraction given by (11), then Algorithm 1 converges linearly with a constant step size. To reach ϵ-accuracy, i.e. $\operatorname{dist}(\boldsymbol{Z}_t, \boldsymbol{Z}_\natural) \le \epsilon$, it takes an order of $\kappa\log(1/\epsilon)$ iterations, which is order-wise equivalent to the regularized PF algorithm in [5]. Comparing to [5], 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 [5], which is $\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{Z}_0\boldsymbol{R} - \boldsymbol{Z}_\natural\|_{\mathrm{F}} \le c_0\sigma_{\min}(\boldsymbol{X}_\natural)$.

We still need to find a good initialization that satisfies (11). In general, one could initialize with the balanced factors of the output after running multiple iterations 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 $\mathcal{P}_r$ is the projection to the best rank-r approximation. 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 [10], [5], the iterates satisfy

$$\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{Z}_\tau\boldsymbol{R} - \boldsymbol{Z}_\natural\|_{\mathrm{F}} \le c_2(2\delta_{4r})^\tau \frac{\|\boldsymbol{M}_\natural\|_{\mathrm{F}}}{\sigma_{\min}(\boldsymbol{X}_\natural)} \tag{12}$$

for some constant c_2. Thus, to achieve the required initialization condition, if we use the spectral method specified in Algorithm 1, which corresponds to setting $\tau = 1$ in (12), we need

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

Alternatively, if we allow multiple iterations of (12) as suggested by [5], we can still set $\delta_{4r} \le \delta_c$ for a sufficiently small constant δ_c, by running at least

$$\tau \ge c_1\log\left(\kappa^{3/2}\frac{\|\boldsymbol{M}_\natural\|_{\mathrm{F}}}{\sigma_{\min}}\right)/\log\left(\delta_c^{-1}\right) = c_2\log(\kappa r)/\log\left(\delta_c^{-1}\right)$$

iterations of projected gradient descent for initialization, which matches the requirement in [5].

III. Related Work

Low-rank matrix estimation has been extensively studied in recent years [2], [3], 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 provable near-optimal statistical guarantees [1], [11], [12], [13], [14], [15], [16], [17], [18], 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 popularized by Burer and Monteiro [4] 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 [5], [19], [20], [21], [22], [23], [24], [25], [26]. The readers are referred to the recent overview [27] for additional references.

To the best of our knowledge, the balancing regularization term (4) was first introduced in [5] to deal with non-square matrix factorization, and has become a standard approach to deal with asymmetric low-rank matrix estimation [6], [7], [8], [28], [29], [30]. A major benefit of adding the regularization term is to reduce the ambiguity set from invertible transforms to orthonormal transforms, so that the distance defined in (7) is minimized over $\boldsymbol{P} \in \mathcal{O}^{r\times r}$. 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. [23] established that vanilla gradient descent works for blind deconvolution at a near-optimal sample complexity with spectral initialization. In [31], 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, we remark that a similar regularization term (4) is also adopted when analyzing the optimization landscape of low-rank matrix estimation, e.g. [32], [33], [34], [35]. Without such a regularization term, the landscape of matrix factorization no longer possesses the intriguing property "all saddle points are strict saddle" and therefore one cannot invoke theory such as [36] to argue the global convergence of gradient descent using an unregularized loss function. Our work partially bridges this gap and suggests the benign behavior of gradient descent even in the absence of local strong convexity.

IV. Proof Sketch of Theorem 1

In this section, we provide a proof sketch of Theorem 1. We first discuss some basic properties of aligning two low-rank factors via invertible transforms, then prove a similar result for a warm-up case of low-rank matrix factorization, of which our problem of interest can be regarded as a perturbed version.

A. Alignment via invertible transforms

For $\boldsymbol{Z} = [\boldsymbol{X}^\top, \boldsymbol{Y}^\top]^\top$ and $\boldsymbol{Z}_\natural = [\boldsymbol{X}_\natural^\top, \boldsymbol{Y}_\natural^\top]^\top$, we define the optimal alignment matrix $\boldsymbol{Q}$ as[1]

$$\boldsymbol{Q} := \operatorname*{argmin}_{\boldsymbol{P}\in\mathbb{R}^{r\times r}} \sqrt{\|\boldsymbol{X}\boldsymbol{P} - \boldsymbol{X}_\natural\|_{\mathrm{F}}^2 + \|\boldsymbol{Y}\boldsymbol{P}^{-\top} - \boldsymbol{Y}_\natural\|_{\mathrm{F}}^2}.$$

Furthermore, we call $\boldsymbol{Z}$ and $\boldsymbol{Z}_\natural$ are aligned if the corresponding optimal alignment matrix $\boldsymbol{Q} = \boldsymbol{I}$. Throughout the paper, we assume the optimal alignment matrix between the tth iterate $\boldsymbol{Z}_t = [\boldsymbol{X}_t^\top, \boldsymbol{Y}_t^\top]^\top$ and $\boldsymbol{Z}_\natural$ is denoted as $\boldsymbol{Q}_t$. Below we provide some basic understandings of this alignment operation.

Lemma 1: Given two matrices $\boldsymbol{Z}$ and $\boldsymbol{Z}_\natural$, and their optimal alignment matrix $\boldsymbol{Q}$, we have

$$\tilde{\boldsymbol{X}}^\top(\tilde{\boldsymbol{X}} - \boldsymbol{X}_\natural) = (\tilde{\boldsymbol{Y}} - \boldsymbol{Y}_\natural)^\top\tilde{\boldsymbol{Y}},$$

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

[1]It is guaranteed with high probability that the minimum is attained for $\boldsymbol{Z}_t$.

Lemma 2: Let $\boldsymbol{Q}$ be the optimal alignment matrix between $\boldsymbol{Z}$ and $\boldsymbol{Z}_\natural$. Suppose there exists a matrix $\boldsymbol{P}$ with $1/2 \le \sigma_{\min}(\boldsymbol{P}) \le \sigma_{\max}(\boldsymbol{P}) \le 3/2$ such that

$$\max\left\{\|\boldsymbol{X}\boldsymbol{P} - \boldsymbol{X}_\natural\|_{\mathrm{F}}, \left\|\boldsymbol{Y}\boldsymbol{P}^{-\top} - \boldsymbol{Y}_\natural\right\|_{\mathrm{F}}\right\} \le \delta \le \frac{1}{4}\sigma_{\min}(\boldsymbol{X}_\natural). \tag{13}$$

Then one has

$$\|\boldsymbol{P} - \boldsymbol{Q}\| \le \|\boldsymbol{P} - \boldsymbol{Q}\|_{\mathrm{F}} \le \frac{10\delta}{\sigma_{\min}(\boldsymbol{X}_\natural)}.$$

Both lemmas provide basic understandings on the solution of solving the alignment problem with invertible transformations, which can be regarded as a generalization of the classical orthogonal Procrustes problem which only considers orthonormal transforms. Clearly, this generalized problem is much more challenging and our work provides some first understandings into it, to the best of our knowledge. These lemmas provide the basis for the subsequent analyses.

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

We consider the following minimization problem

$$f_{\mathrm{MF}}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{1}{2}\left\|\boldsymbol{X}\boldsymbol{Y}^\top - \boldsymbol{M}_\natural\right\|_{\mathrm{F}}^2, \tag{14}$$

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

$$\begin{aligned}
\boldsymbol{X}_{t+1} &= \boldsymbol{X}_t - \frac{\eta}{\sigma_{\max}}\nabla_{\boldsymbol{X}} f_{\mathrm{MF}}(\boldsymbol{X}_t, \boldsymbol{Y}_t) \\
&= \boldsymbol{X}_t - \frac{\eta}{\sigma_{\max}}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\natural)\boldsymbol{Y}_t; \\
\boldsymbol{Y}_{t+1} &= \boldsymbol{Y}_t - \frac{\eta}{\sigma_{\max}}\nabla_{\boldsymbol{Y}} f_{\mathrm{MF}}(\boldsymbol{X}_t, \boldsymbol{Y}_t) \\
&= \boldsymbol{Y}_t - \frac{\eta}{\sigma_{\max}}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\natural)^\top\boldsymbol{X}_t.
\end{aligned} \tag{15}$$

We have the following theorem regarding the performance of (15), which parallels with Theorem 1.

Theorem 2: Let $\boldsymbol{Z}_0$ be an initialization which satisfies

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

for some small enough constant c_0. There exists some c_1 such at as long as $\eta \le c_1$, the iterates of GD satisfy

$$\operatorname{dist}(\boldsymbol{Z}_t, \boldsymbol{Z}_\natural) \le \left(1 - \frac{\eta}{20\kappa}\right)^t \operatorname{dist}(\boldsymbol{Z}_0, \boldsymbol{Z}_\natural).$$

C. Analysis for matrix sensing

We now extend the technique used in the proof of Theorem 2 to the matrix sensing case by leveraging the RIP. Suppose that the initialization $\boldsymbol{Z}_0$ satisfies (11). By a similar argument as in [5], it is sufficient to consider the following update rule:

$$\begin{aligned}
\boldsymbol{X}_{t+1} &= \boldsymbol{X}_t - \frac{\eta}{\sigma_{\max}}\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\natural)\right]\boldsymbol{Y}_t; \\
\boldsymbol{Y}_{t+1} &= \boldsymbol{Y}_t - \frac{\eta}{\sigma_{\max}}\left[\mathcal{A}^*\mathcal{A}(\boldsymbol{X}_t\boldsymbol{Y}_t^\top - \boldsymbol{M}_\natural)\right]^\top\boldsymbol{X}_t.
\end{aligned} \tag{16}$$

Compared with (15), the update rule for matrix sensing differs by the operation of $\mathcal{A}^*\mathcal{A}$ when forming the gradient. Therefore, we expect the GD has similar behaviors as earlier as long as

$\mathcal{A}$ behaves as a near isometry on low-rank matrices. This can be supplied by the following consequence of the RIP.

***Lemma** 3:* Suppose $\mathcal{A}$ satisfies $2r$-RIP with constant δ_{2r}. Then, for all matrices $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ of rank at most r, we have

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

V. Conclusions

This paper establishes the local linear convergence of gradient descent for rectangular 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, which can be found by the spectral method. 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 our technique can be used for other low-rank matrix estimation problems. To conclude, we outline a few exciting 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 [23], [30] 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.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] 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.

[9] 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.

[10] 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.

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

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

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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," *arXiv preprint arXiv:1711.10467*, 2017.

[24] 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.

[25] Y. Li, C. Ma, Y. Chen, and Y. Chi, "Nonconvex matrix factorization from rank-one measurements," *arXiv preprint arXiv:1802.06286*, 2018.

[26] X. Li, S. Ling, T. Strohmer, and K. Wei, "Rapid, robust, and reliable blind deconvolution via nonconvex optimization," *Applied and computational harmonic analysis*, 2018.

[27] Y. Chi, Y. M. Lu, and Y. Chen, "Nonconvex optimization meets low-rank matrix factorization: An overview," *arXiv preprint arXiv:1809.09573*, 2018.

[28] Y. Li, Y. Chi, H. Zhang, and Y. Liang, "Nonconvex low-rank matrix recovery with arbitrary outliers via median-truncated gradient descent," *arXiv preprint arXiv:1709.08114*, 2017.

[29] 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.

[30] J. Chen, D. Liu, and X. Li, "Nonconvex rectangular matrix completion via gradient descent without $\ell_{2,\infty}$ regularization," *arXiv preprint arXiv:1901.06116*, 2019.

[31] S. S. Du, W. Hu, and J. D. Lee, "Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced," *arXiv preprint arXiv:1806.00900*, 2018.

[32] 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.

[33] 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.

[34] 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.

[35] Z. Zhu, D. Soudry, Y. C. Eldar, and M. B. Wakin, "The global optimization geometry of shallow linear neural networks," *arXiv preprint arXiv:1805.04938*, 2018.

[36] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Conference on Learning Theory*, 2016, pp. 1246–1257.