The Global Geometry of Centralized and Distributed Low-rank Matrix Recovery Without Regularization

Shuang Li, Qiuwei Li, Zhihui Zhu, Gongguo Tang, and Michael B. Wakin

***Abstract*—Low-rank matrix recovery is a fundamental problem in signal processing and machine learning. A recent very popular approach to recovering a low-rank matrix $\mathbf{X}$ is to factorize it as a product of two smaller matrices, i.e., $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, and then optimize over $\mathbf{U}$, $\mathbf{V}$ instead of $\mathbf{X}$. Despite the resulting non-convexity, recent results have shown that many factorized objective functions actually have benign global geometry—with no spurious local minima and satisfying the so-called strict saddle property—ensuring convergence to a global minimum for many local-search algorithms. Such results hold whenever the original objective function is restricted strongly convex and smooth. However, most of these results actually consider a modified cost function that includes a balancing regularizer. While useful for deriving theory, this balancing regularizer does not appear to be necessary in practice. In this work, we close this theory-practice gap by proving that the unaltered factorized non-convex problem, without the balancing regularizer, also has similar benign global geometry. Moreover, we also extend our theoretical results to the field of distributed optimization.**

***Index Terms*—Low-rank matrix recovery, non-convex optimization, geometric landscape, centralized optimization, distributed optimization.**

I. Introduction

In THE problem of low-rank matrix recovery, a great number of efforts have been made to minimize a loss function $f(\mathbf{X})$ over the non-convex rank constraint $\text{rank}(\mathbf{X}) \leq r$, where $\mathbf{X} \in \mathbb{R}^{n\times m}$ and $r \ll \min\{n, m\}$. Among which, a popular way is to replace the rank constraint with the Burer-Monteiro factorization, i.e., $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{n\times r}$ and $\mathbf{V} \in \mathbb{R}^{m\times r}$ [1], [2], changing the objective function from $f(\mathbf{X})$ to $g(\mathbf{U}, \mathbf{V}) = f(\mathbf{U}\mathbf{V}^\top)$. This factorization approach can often lead to lower computational and storage complexity, while raising new questions about whether an algorithm can converge to favorable solutions since the bilinear form $\mathbf{U}\mathbf{V}^\top$ naturally introduces non-convexity. Fortunately, it is observed that simple iterative algorithms find global optimal solutions in many low-rank matrix recovery problems [3]–[12].

Manuscript received March 24, 2020; revised June 3, 2020; accepted July 9, 2020. Date of publication July 15, 2020; date of current version August 20, 2020. This work was supported in part by the NSF under Grant CCF-1704204 and in part by the DARPA Lagrange Program ONR/SPAWAR under Contract N660011824020. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nelly Pustelnik. *(Corresponding author: Shuang Li.)*

Shuang Li, Qiuwei Li, Gongguo Tang, and Michael B. Wakin are with the Department of Electrical Engineering, Colorado School of Mines, Golden, CO 80401 USA (e-mail: lishuang0809@gmail.com; qiuli@mines.edu; gtang@mines.edu; mwakin@mines.edu).

Zhihui Zhu is with the Ritchie School of Engineering and Computer Science, University of Denver, Denver, CO 80208 USA (e-mail: zhihui.zhu@du.edu).

Digital Object Identifier 10.1109/LSP.2020.3008876

Recent years have seen a surge of interest in understanding these surprising phenomena by analyzing the landscape of the factorized cost function $g(\mathbf{U}, \mathbf{V})$. To accomplish this, many existing works [8]–[16] actually add a balancing regularizer

$$R(\mathbf{U}, \mathbf{V}) \doteq \|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2, \tag{1}$$

which implicitly forces $\mathbf{U}$ and $\mathbf{V}$ to have equal energy, to the objective function $g(\mathbf{U}, \mathbf{V})$. These works then show that, for broad classes of problems, the regularized cost functions have a benign geometry, where every local minimum is a global minimum and every first-order critical point is either a local minimum or a strict saddle [17], [18]. This favorable property ensures a convergence to a global minimum for many local search methods [18]–[24].

A. What Is The Role of The Balancing Regularizer?

If $(\mathbf{U}, \mathbf{V})$ is a critical point of $g(\mathbf{U}, \mathbf{V})$, then $(\mathbf{U}\mathbf{G}, \mathbf{V}\mathbf{G}^{-\top})$ is also a critical point for any invertible $\mathbf{G} \in \mathbb{R}^{r\times r}$. This scaling ambiguity in the critical points can result in an infinite number of connected critical points including those ill-conditioned points when $\|\mathbf{G}\|_F$ goes to 0 or ∞, which could bring new challenges in analyzing the geometric landscape as one must analyze the optimality of any critical point. In order to remove this ambiguity, many researchers [8]–[16] utilize the balancing regularizer (1). In particular, it has been shown that adding the regularizer (1) forces all critical points $(\mathbf{U}, \mathbf{V})$ to be balanced, i.e., $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V}$.

B. Is The Balancing Regularizer Really Necessary?

Most previous works add the balancing regularizer (1) to the cost function in order to simplify the landscape analysis. However, we have observed that one can achieve almost the same performance without adding the balancing regularizer in [10]. Also, in practice this additional regularizer is rarely utilized [25], which implies a gap between theory and practice. This naturally raises the main question that will be addressed in this work: Is the balancing regularizer (1) truly necessary? In other words, can we characterize the global geometry of the factorization approach without the balancing regularizer?

Several works [25]–[28] answer this question by analyzing the behavior of gradient descent on some particular optimization problems, and show that the iterates of gradient descent stay in the (approximately) balanced path from some specific initialization and finally converge to a global optimal solution. However, these results are restricted to gradient descent with a

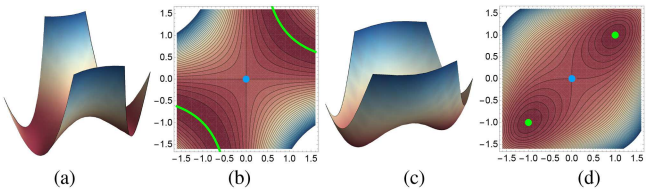


Fig. 1. (a, b) and (c, d) are the landscapes of the non-regularized function $g(u,v)$ and the regularized cost function $\widetilde{g}(u,v)$ with $\mu = \frac{1}{16}$, respectively. One can observe that although g has an infinite number of (connected) critical points while $\widetilde{g}$ has just three critical points at $\pm(1,1)$ and (0, 0), both cost functions have benign landscapes since any critical point is either a global minimizer or a strict saddle. The points marked with green and blue in (b, d) denote the global minimizers and saddle points, respectively.

specific initialization. There are also some works that analyze the geometric landscape of some specific optimization problems, such as matrix factorization [29], or linear neural network optimization [11], [30], [31].

In this work, we answer this question by directly analyzing the landscape of the unaltered factorized non-convex problem, *without* the balancing regularizer (1). In particular, over the general class of problems where the cost function f is restricted strongly convex and smooth (see Definition III.1), we show under mild conditions that any critical point of the factorized cost function g (including any unbalanced critical point) is either a global optimum or a strict saddle. This helps close the theory-practice gap and resolves the open problem in [10]. Moreover, we extend our results to the corresponding distributed setting and show that many global consensus problems inherit the benign geometry of their original centralized counterpart.

Before proceeding, we present a toy example to illustrate our main observation.

Example I.1 (Matrix factorization – the scalar case): Consider an asymmetric matrix factorization cost function $g(u,v) \doteq \frac{1}{2}(1-uv)^2$, whose critical points (u,v) satisfy $uv = 1$ or $(u,v) = (0,0)$. The critical points of the corresponding regularized function $\widetilde{g}(u,v) \doteq \frac{1}{2}(1-uv)^2 + \frac{\mu}{4}(u^2-v^2)^2$ with some $\mu > 0$ satisfy $(uv-1)v + \mu(u^2-v^2)u = 0$ and $(uv-1)u - \mu(u^2-v^2)v = 0$, which gives only three critical points $(1,1), (-1,-1)$ and (0,0). Therefore, for $\widetilde{g}(u,v)$, one only needs to check the Hessian evaluated at these three critical points: $\nabla^2\widetilde{g}(1,1) = \nabla^2\widetilde{g}(-1,-1) = \begin{bmatrix} 1+2\mu & 1-2\mu \\ 1-2\mu & 1+2\mu \end{bmatrix} \succ 0$, and $\nabla^2\widetilde{g}(0,0) = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$ which has a strictly negative eigenvalue -1. Thus any critical point of $\widetilde{g}(u,v)$ is either a global minimum or a strict saddle, which implies a favorable landscape of the regularized cost function $\widetilde{g}(u,v)$. As can be seen, adding the balancing regularizer can largely simplify the landscape analysis. However, this does not imply that the original function $g(u,v)$ does not have a benign geometry. Indeed, one can observe that any critical point of $g(u,v)$ either satisfies $uv = 1$ (globally optimal) or $(u,v) = (0,0)$ (strict saddle since $\nabla^2 g(0,0) = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$ has a negative eigenvalue -1). The landscapes of g and $\widetilde{g}$ are shown in Fig. 1.

The remainder of this letter is organized as follows. In Section II, we formulate the problems in both centralized and distributed settings. We present our main theorem and its proof in Section III. In Section IV, we conduct a series of experiments to further support our theory. Finally, we conclude our work in Section V.

II. Problem Formulation

We first consider the following problem of minimizing a general objective function over the set of low-rank matrices:

$$\underset{\mathbf{X}\in\mathbb{R}^{n\times m}}{\text{minimize}}\ f(\mathbf{X}) \quad \text{subject to}\ \ \text{rank}(\mathbf{X}) \le r, \tag{2}$$

which is a fundamental problem that often appears in the fields of signal processing and machine learning. Plugging the Burer-Monteiro type decomposition [1], [2], i.e., $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{n\times r}$ and $\mathbf{V} \in \mathbb{R}^{m\times r}$, into the above cost function, one can remove the low-rank constraint and get the following unconstrained optimization

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r},\mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}}\ g(\mathbf{U},\mathbf{V}) \doteq f(\mathbf{U}\mathbf{V}^\top), \tag{3}$$

which is a non-convex optimization problem we refer to as *centralized* low-rank matrix recovery. The above optimization appears in many applications including low-rank matrix approximation [8], matrix sensing [9], matrix completion [32], and linear neural network optimization [11], [30], [31]. Note that in centralized low-rank matrix recovery, all the computations happen at one "central" node that has full access, for example, to the data matrix or the measurements.

In the second part of this work, we study the impact of *distributing* the centralized low-rank matrix recovery problem for general cost functions. Consider a separable cost function $f(\mathbf{U}\mathbf{V}^\top) = \sum_{j=1}^{J} f_j(\mathbf{U}\mathbf{V}_j^\top)$, where $\mathbf{U} \in \mathbb{R}^{n\times r}$ is the common variable in all of the objective functions $\{f_j\}_{j\in[J]}$ and $\mathbf{V}_j \in \mathbb{R}^{m_j\times r}$ as a submatrix of $\mathbf{V} = [\mathbf{V}_1^\top \cdots \mathbf{V}_J^\top]^\top \in \mathbb{R}^{m\times r}$ is the local variable only corresponding to objective function f_j. Then, the centralized optimization (3) becomes

$$\underset{\mathbf{U},\{\mathbf{V}_j\}}{\text{minimize}} \sum_{j=1}^{J} f_j(\mathbf{U}\mathbf{V}_j^\top). \tag{4}$$

In the *distributed* setting, one distributes (4) across a network of J agents and considers the following optimization

$$\underset{\{\mathbf{U}^j\},\{\mathbf{V}_j\}}{\text{minimize}} \sum_{j=1}^{J} f_j(\mathbf{U}^j\mathbf{V}_j^\top)\ \text{s.t.}\ \mathbf{U}^1 = \cdots = \mathbf{U}^J. \tag{5}$$

Here, $\mathbf{U}^j$ and $\mathbf{V}_j$ are the so-called consensus and local variables at node j. In this work, we consider the above equality-constrained distributed problem (5) by reformulating it as the following unconstrained optimization problem

$$\underset{\{\mathbf{U}^j\},\{\mathbf{V}_j\}}{\text{minimize}} \sum_{j=1}^{J} f_j(\mathbf{U}^j\mathbf{V}_j^\top) + \sum_{(i,j)\in\mathcal{G}} w_{i,j}\|\mathbf{U}^j - \mathbf{U}^i\|_F^2. \tag{6}$$

Here, $\mathcal{G}$ denotes any connected network over $[J]^2$ with $[J] \doteq \{1,\ldots,J\}$ and $[J]^2 = [J]\times[J]$ [29], and $\{w_{i,j}\}_{(i,j)\in\mathcal{G}}$ are symmetric positive weights, i.e., $w_{j,i} = w_{i,j} > 0$. The second term is added to the objective function for the purpose of promoting equality among the consensus variables $\mathbf{U}^j$.

In this work, our main goal is to characterize the global geometry of the non-convex *centralized* cost function (3) and non-convex *distributed* cost function (6). In particular, we show that under the same assumptions as required in the previous works, any critical point is either a global minimum or a strict saddle, where the Hessian has a strictly negative eigenvalue, without adding the balancing regularizer.

III. Main Results

A. Landscape of Centralized Low-rank Matrix Recovery

In this subsection, we present the geometric landscape of the *centralized* optimization (3). We start by introducing the restricted strongly convex and smooth property.

Definition III.1: ([5], [10]) A function $f(\mathbf{X})$ is said to be $(2r, 4r)$-restricted strongly convex and smooth if

$$\alpha\|\mathbf{D}\|_F^2 \le \nabla^2 f(\mathbf{X})[\mathbf{D},\mathbf{D}] \le \beta\|\mathbf{D}\|_F^2 \tag{7}$$

holds for any matrix $\mathbf{X}\in\mathbb{R}^{n\times m}$ with rank at most $2r$ and $\mathbf{D}\in\mathbb{R}^{n\times m}$ with rank at most $4r$. Here, α and β are some positive constants, and $\nabla^2 f(\mathbf{X})[\mathbf{D},\mathbf{D}]=\sum_{i,j,k,l}\frac{\partial^2 f(\mathbf{X})}{\partial \mathbf{X}_{ij}\partial \mathbf{X}_{kl}}\mathbf{D}_{ij}\mathbf{D}_{kl}$ denotes a bilinear form of the Hessian of $f(\mathbf{X})$.

Unlike the standard strongly convex and smooth condition which requires (7) to hold for any $\mathbf{X}$ and $\mathbf{D}$, the above restricted version only requires (7) to hold for low-rank matrices, making it amenable for low-rank matrix recovery problems. For example, in matrix sensing the goal is to recover a low-rank matrix $\mathbf{X}$ from linear measurements $\mathcal{A}(\mathbf{X})$. The linear operator $\mathcal{A}$ often satisfies the restricted isometry property (RIP), which can be interpreted as satisfying (7) for all low-rank matrices $\mathbf{D}$ and all $\mathbf{X}$; see [10] for details.

Theorem III.1: Assume that the cost function $f(\mathbf{X})$ in (2) satisfies the $(2r,4r)$-restricted strongly convex and smooth property with positive constants α and β satisfying $\beta/\alpha\le 3/2$. Also assume that $f(\mathbf{X})$ has a critical point $\mathbf{X}^\star$ with $\text{rank}(\mathbf{X}^\star)\le r$. Then, any critical point $(\mathbf{U},\mathbf{V})$ of $g(\mathbf{U},\mathbf{V})$ in (3) is either a global minimum (i.e., $\mathbf{U}\mathbf{V}^\top=\mathbf{X}^\star$) or a strict saddle (i.e., $\lambda_{\min}(\nabla^2 g(\mathbf{U},\mathbf{V}))<0$).

Proof: It follows from [10, Proposition 1] that the critical point $\mathbf{X}^\star$ of $f(\mathbf{X})$ with $\text{rank}(\mathbf{X}^\star)=r^\star\le r$ is its global minimum, namely, $f(\mathbf{X}^\star)\le f(\mathbf{X})$ holds for any $\mathbf{X}\in\mathbb{R}^{n\times m}$ with $\text{rank}(\mathbf{X})\le r$. Moreover, the equality holds only at $\mathbf{X}=\mathbf{X}^\star$. Then, for any critical point $(\mathbf{U},\mathbf{V})$ with $\mathbf{U}\mathbf{V}^\top=\mathbf{X}^\star$, we have $g(\mathbf{U},\mathbf{V})=f(\mathbf{X}^\star)$, and hence $(\mathbf{U},\mathbf{V})$ is a global minimum.

For any critical point $(\mathbf{U},\mathbf{V})$ with $\mathbf{U}\mathbf{V}^\top\neq\mathbf{X}^\star$, we next show that there exists a direction $\mathbf{D}\in\mathbb{R}^{(n+m)\times r}$ such that $\nabla^2 g(\mathbf{U},\mathbf{V})[\mathbf{D},\mathbf{D}]<0$, namely, $(\mathbf{U},\mathbf{V})$ is a strict saddle of $g(\mathbf{U},\mathbf{V})$. The remaining part of this proof is inspired by the proof of [29, Lemma 11.3] and [30, Theorem 8] and is split into two cases: 1) $\text{rank}(\mathbf{U}\mathbf{V}^\top)=r$, and 2) $\text{rank}(\mathbf{U}\mathbf{V}^\top)<r$.

Non-degenerate case: rank$(\mathbf{U}\mathbf{V}^\top)=r$

Let $\mathbf{U}\mathbf{V}^\top=\mathbf{P}\mathbf{\Sigma}\mathbf{Q}^\top$ be an SVD of $\mathbf{U}\mathbf{V}^\top$. It follows from $\text{rank}(\mathbf{U}\mathbf{V}^\top)=r$ that $\text{rank}(\mathbf{U})=\text{rank}(\mathbf{V})=r$, which further implies that $\mathbf{U}^\top\mathbf{U}$ and $\mathbf{V}^\top\mathbf{V}$ are invertible. Then, we define two matrices $\mathbf{G}_1\doteq(\mathbf{U}^\top\mathbf{U})^{-1}\mathbf{U}^\top\mathbf{P}\mathbf{\Sigma}^{1/2}$, and $\mathbf{G}_2\doteq(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top\mathbf{Q}\mathbf{\Sigma}^{1/2}$. It can be seen that $\mathbf{G}_1\mathbf{G}_2^\top=\mathbf{I}_r$. We also define balanced factors $\widetilde{\mathbf{U}}\doteq\mathbf{P}\mathbf{\Sigma}^{1/2}$ and $\widetilde{\mathbf{V}}\doteq\mathbf{Q}\mathbf{\Sigma}^{1/2}$.

It can be seen that the new matrix pair $(\widetilde{\mathbf{U}},\widetilde{\mathbf{V}})$ satisfies

$$\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top=\mathbf{U}\mathbf{V}^\top,\qquad \widetilde{\mathbf{U}}^\top\widetilde{\mathbf{U}}=\widetilde{\mathbf{V}}^\top\widetilde{\mathbf{V}}. \tag{8}$$

Recall that for any critical point $(\mathbf{U},\mathbf{V})$ of $g=f(\mathbf{U}\mathbf{V}^\top)$, we have $\nabla g(\mathbf{U},\mathbf{V})=\mathbf{0}$, i.e., $\nabla f(\mathbf{U}\mathbf{V}^\top)\mathbf{V}=\mathbf{0}$ and $(\nabla f(\mathbf{U}\mathbf{V}^\top))^\top\mathbf{U}=\mathbf{0}$. Together with the equalities in (8), we get $\nabla(g+\frac{\mu}{4}R)(\widetilde{\mathbf{U}},\widetilde{\mathbf{V}})=\begin{bmatrix}\nabla f(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top)\widetilde{\mathbf{V}}+\mu\widetilde{\mathbf{U}}(\widetilde{\mathbf{U}}^\top\widetilde{\mathbf{U}}-\widetilde{\mathbf{V}}^\top\widetilde{\mathbf{V}})\\(\nabla f(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top))^\top\widetilde{\mathbf{U}}-\mu\widetilde{\mathbf{V}}(\widetilde{\mathbf{U}}^\top\widetilde{\mathbf{U}}-\widetilde{\mathbf{V}}^\top\widetilde{\mathbf{V}})\end{bmatrix}=\mathbf{0}$, where $R(\cdot)$ is the balancing regularizer introduced in (1) and $\mu>0$ is a regularizer parameter. This immediately implies that the new matrix pair $(\widetilde{\mathbf{U}},\widetilde{\mathbf{V}})$ is a critical point of the regularized cost function $g(\mathbf{U},\mathbf{V})+\frac{\mu}{4}R(\mathbf{U},\mathbf{V})$.

On the other hand, it follows from [10] that there exists a matrix $\widetilde{\mathbf{D}}=[\widetilde{\mathbf{D}}_{\widetilde{\mathbf{U}}}^\top\ \widetilde{\mathbf{D}}_{\widetilde{\mathbf{V}}}^\top]^\top\in\mathbb{R}^{(n+m)\times r}$ such that

$$\nabla^2\left(g(\widetilde{\mathbf{U}},\widetilde{\mathbf{V}})+\frac{\mu}{4}R(\widetilde{\mathbf{U}},\widetilde{\mathbf{V}})\right)\left[\widetilde{\mathbf{D}},\widetilde{\mathbf{D}}\right]<0 \tag{9}$$

holds for any $\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top\neq\mathbf{X}^\star$.

Construct $\mathbf{D}=[\begin{smallmatrix}\mathbf{D}_{\mathbf{U}}\\\mathbf{D}_{\mathbf{V}}\end{smallmatrix}]=[\begin{smallmatrix}\widetilde{\mathbf{D}}_{\widetilde{\mathbf{U}}}\mathbf{G}_1^{-1}\\\widetilde{\mathbf{D}}_{\widetilde{\mathbf{V}}}\mathbf{G}_2^{-1}\end{smallmatrix}]$, and denote $\mathbf{\Pi}\doteq\mathbf{U}\mathbf{D}_{\mathbf{V}}^\top+\mathbf{D}_{\mathbf{U}}\mathbf{V}^\top$ and $\widetilde{\mathbf{\Pi}}\doteq\widetilde{\mathbf{U}}\widetilde{\mathbf{D}}_{\widetilde{\mathbf{V}}}^\top+\widetilde{\mathbf{D}}_{\widetilde{\mathbf{U}}}\widetilde{\mathbf{V}}^\top$. Note that $\mathbf{\Pi}=\widetilde{\mathbf{\Pi}}$. It follows from (9) that

$$\begin{aligned}0>&\left\langle 2\nabla f(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top),\widetilde{\mathbf{D}}_{\widetilde{\mathbf{U}}}\widetilde{\mathbf{D}}_{\widetilde{\mathbf{V}}}^\top\right\rangle+\nabla^2 f(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top)\left[\widetilde{\mathbf{\Pi}},\widetilde{\mathbf{\Pi}}\right]\\&+(\mu/2)\|\widetilde{\mathbf{D}}_{\widetilde{\mathbf{U}}}^\top\widetilde{\mathbf{U}}+\widetilde{\mathbf{U}}^\top\widetilde{\mathbf{D}}_{\widetilde{\mathbf{U}}}-\widetilde{\mathbf{D}}_{\widetilde{\mathbf{V}}}^\top\widetilde{\mathbf{V}}-\widetilde{\mathbf{V}}^\top\widetilde{\mathbf{D}}_{\widetilde{\mathbf{V}}}\|_F^2\\\ge&\left\langle 2\nabla f(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top),\widetilde{\mathbf{D}}_{\widetilde{\mathbf{U}}}\widetilde{\mathbf{D}}_{\widetilde{\mathbf{V}}}^\top\right\rangle+\nabla^2 f(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^\top)\left[\widetilde{\mathbf{\Pi}},\widetilde{\mathbf{\Pi}}\right]\\=&\nabla^2 g(\mathbf{U},\mathbf{V})[\mathbf{D},\mathbf{D}],\end{aligned}$$

which further implies that any non-degenerate critical point $(\mathbf{U},\mathbf{V})$ with $\mathbf{U}\mathbf{V}^\top\neq\mathbf{X}^\star$ is a strict saddle.

Degenerate case: rank$(\mathbf{U}\mathbf{V}^\top)<r$

Note that $\text{rank}(\mathbf{U}^\top\mathbf{U}\mathbf{V}^\top\mathbf{V})\le\text{rank}(\mathbf{U}\mathbf{V}^\top)<r$, which implies that $det(\mathbf{U}^\top\mathbf{U}\mathbf{V}^\top\mathbf{V})=det(\mathbf{U}^\top\mathbf{U})det(\mathbf{V}^\top\mathbf{V})=0$. Then, either $det(\mathbf{U}^\top\mathbf{U})=0$ or $det(\mathbf{V}^\top\mathbf{V})=0$. Or equivalently, either $\text{rank}(\mathbf{U}^\top\mathbf{U})<r$ or $\text{rank}(\mathbf{V}^\top\mathbf{V})<r$. Note that $\text{rank}(\mathbf{U})=\text{rank}(\mathbf{U}^\top\mathbf{U})$ and $\text{rank}(\mathbf{V})=\text{rank}(\mathbf{V}^\top\mathbf{V})$. Then, of the following two statements, at least one of them is true.

(i) $\exists\,\mathbf{b}\neq\mathbf{0}$ such that $\mathbf{b}\in null(\mathbf{U})$, i.e., $\mathbf{U}\mathbf{b}=\mathbf{0}$.

(ii) $\exists\,\mathbf{b}\neq\mathbf{0}$ such that $\mathbf{b}\in null(\mathbf{V})$, i.e., $\mathbf{V}\mathbf{b}=\mathbf{0}$.

Note that for any critical point $(\mathbf{U},\mathbf{V})$, either

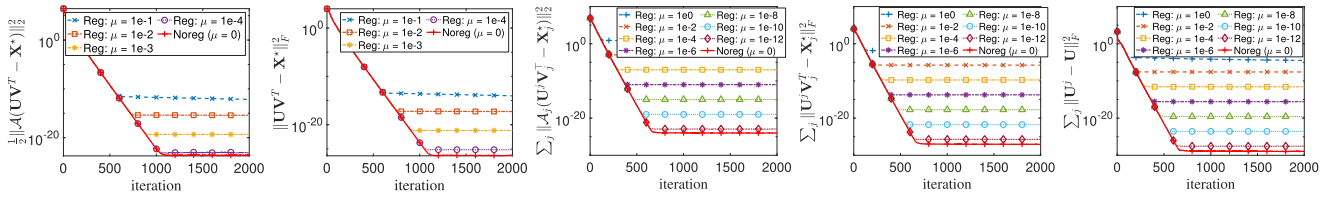
$$\nabla f(\mathbf{U}\mathbf{V}^\top)=\mathbf{0}\Rightarrow(\mathbf{U},\mathbf{V})\text{is a global minimum, } or$$

$$\nabla f(\mathbf{U}\mathbf{V}^\top)\neq\mathbf{0}\Rightarrow\exists\,(i,j),\ \left\langle\nabla f(\mathbf{U}\mathbf{V}^\top),\mathbf{e}_i\mathbf{e}_j^\top\right\rangle\neq 0.$$

Next, we focus on the second case and show that such kinds of critical points are strict saddles.

Assume that (i) is true. Construct $\mathbf{D}=[\mathbf{D}_{\mathbf{U}}^\top\ \mathbf{D}_{\mathbf{V}}^\top]^\top$ with $\mathbf{D}_{\mathbf{U}}^\top=\mathbf{b}\mathbf{e}_i^\top\in\mathbb{R}^{r\times n}$ and $\mathbf{D}_{\mathbf{V}}^\top=(\alpha\mathbf{b})\mathbf{e}_j^\top\in\mathbb{R}^{r\times m}$. Then, we have $\mathbf{D}_{\mathbf{U}}\mathbf{D}_{\mathbf{V}}^\top=\alpha\|\mathbf{b}\|_2^2\mathbf{e}_i\mathbf{e}_j^\top$, and $\mathbf{U}\mathbf{D}_{\mathbf{V}}^\top=\mathbf{U}(\alpha\mathbf{b})\mathbf{e}_j^\top=\mathbf{0}$. Plugging into the bilinear form of the Hessian, we get

$$\begin{aligned}\nabla^2 g(\mathbf{U},\mathbf{V})\,[\mathbf{D},\mathbf{D}]=&\,2\alpha\|\mathbf{b}\|_2^2\left\langle\nabla f(\mathbf{U}\mathbf{V}^\top),\mathbf{e}_i\mathbf{e}_j^\top\right\rangle\\&+\nabla^2 f(\mathbf{U}\mathbf{V})^\top[\mathbf{e}_i(\mathbf{V}\mathbf{b})^\top,\mathbf{e}_i(\mathbf{V}\mathbf{b})^\top]\end{aligned}$$



(a) Centralized (b) Centralized (c) Distributed (d) Distributed (e) Distributed

Fig. 2. Convergence of gradient descent and distributed gradient descent for solving the matrix sensing problems in terms of different optimality errors. Here, $m = 50$, $n = 40$, $r = 5$, and the number of measurements $p = 3\max\{m, n\}r$.

Now using the fact that $\langle \nabla f(\mathbf{U}\mathbf{V}^\top), \mathbf{e}_i\mathbf{e}_j^\top \rangle \neq 0$, $\|\mathbf{b}\|_2^2 \neq 0$ and that $\nabla^2 f(\mathbf{U}\mathbf{V}^\top)[\mathbf{e}_i(\mathbf{V}\mathbf{b})^\top, \mathbf{e}_i(\mathbf{V}\mathbf{b})^\top]$ is constant with respect to α, we can always choose α in order to let the first term $\alpha\|\mathbf{b}\|_2^2 \langle 2\nabla f(\mathbf{U}\mathbf{V}^\top), \mathbf{e}_i\mathbf{e}_j^\top \rangle$ be negative enough so that $\nabla^2 g(\mathbf{U}, \mathbf{V})[\mathbf{D}, \mathbf{D}]$ is negative. Therefore, we can conclude that such a critical point $(\mathbf{U}, \mathbf{V})$ is a strict saddle. Similarly, we can consider the case when (ii) is true and finish the proof. ■

We note that, in Theorem III.1, the requirement $\beta/\alpha \leq 3/2$ is the same as in [10] where matrix sensing and other low-rank matrix recovery problems are discussed.

B. Landscape of Distributed Low-rank Matrix Recovery

The following corollary extends the benign geometry to the *distributed* setting introduced in Section II.

Corollary III.1: Under the assumptions in Theorem III.1, any critical point $(\{\mathbf{U}^j\}, \{\mathbf{V}_j\})$ of the distributed problem (6) satisfies $\mathbf{U}^1 = \mathbf{U}^2 = \cdots = \mathbf{U}^J = \mathbf{U}$ for some $\mathbf{U}$, and $(\mathbf{U}, \{\mathbf{V}_j\})$ is either a strict saddle or a global minimizer of (6).

This result follows from Theorem III.1 and Lemma III.1, which is a summary of [29, Proposition 2.3 and Theorem 2.7].

Lemma III.1: [29] Let $\{w_{i,j}\}_{(i,j)\in\mathcal{G}}$ be symmetric positive weights on any connected network $\mathcal{G}$ over $[J]^2$. Then, *(i)* any critical point $(\{\mathbf{U}^j\}, \{\mathbf{V}_j\})$ of the distributed problem (6) satisfies $\mathbf{U}^1 = \mathbf{U}^2 = \cdots = \mathbf{U}^J = \mathbf{U}$ for some $\mathbf{U}$, and *(ii)* $(\mathbf{U}, \{\mathbf{V}_j\})$ is a critical point of the centralized problem (4).

IV. Simulation Results

In this section, we conduct several experiments to further support our theory. In particular, we first consider the following *centralized* matrix sensing problem

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r}, \mathbf{V}\in\mathbb{R}^{m\times r}}{\text{minimize}} \; g(\mathbf{U}, \mathbf{V}) \doteq \frac{1}{2}\|\mathcal{A}(\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star)\|_2^2, \qquad (10)$$

where $\mathcal{A} : \mathbb{R}^{n\times m} \to \mathbb{R}^p$ is a linear sensing operator, and $\mathbf{X}^\star \in \mathbb{R}^{n\times m}$ is the true low-rank matrix with $\text{rank}(\mathbf{X}^\star) = r$. In order to compare the non-regularized setting with the regularized setting, we apply gradient descent with random initialization to minimize the following regularized cost function $\widetilde{g}(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\|\mathcal{A}(\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star)\|_2^2 + \frac{\mu}{4}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2$ with μ equal to 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4} and 0. Note that the regularized cost function $\widetilde{g}(\mathbf{U}, \mathbf{V})$ reduces to the non-regularized cost function $g(\mathbf{U}, \mathbf{V})$ when $\mu = 0$. To set up the experiment, we choose $m = 50$, $n = 40$, $r = 5$, and $p = 3\max\{m, n\}r$. The true data matrix $\mathbf{X}^\star$ is generated as $\mathbf{X}^\star = \mathbf{U}^\star\mathbf{V}^{\star\top}$, where $\mathbf{U}^\star$ and $\mathbf{V}^\star$ are two Gaussian random matrices with entries following $\mathcal{N}(0, 1)$. The linear sensing operator is generated as a $p \times nm$ Gaussian random matrix with entries following $\mathcal{N}(0, 1)$. We plot the fitting error $g(\mathbf{U}, \mathbf{V})$ and the optimality error $\|\mathbf{U}\mathbf{V}^\top - \mathbf{X}^\star\|_F^2$ as a function of the iteration number in Fig. 2 (a) and (b), respectively. One can observe that global optimality is achieved in all regularized and unregularized ($\mu = 0$) cases.[1] Therefore, in the case of *centralized* matrix sensing, the balancing regularizer $\mathrm{R}(\mathbf{U}, \mathbf{V})$ in (1) is not necessary to obtain a benign landscape.

Next, we repeat the above experiments on the corresponding *distributed* matrix sensing problem, namely, minimizing the following cost functions $\widetilde{g}_d(\mathbf{U}, \mathbf{V}) = g_d(\mathbf{U}, \mathbf{V}) + \frac{\mu}{4}\|\mathbf{U}^\top\mathbf{U} - \mathbf{V}^\top\mathbf{V}\|_F^2$, and $g_d(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\sum_{j=1}^J \|\mathcal{A}_j(\mathbf{U}^j\mathbf{V}_j^\top - \mathbf{X}_j^\star)\|_2^2 + \sum_{(i,j)\in\mathcal{G}} w_{i,j}\|\mathbf{U}^j - \mathbf{U}^i\|_F^2$. We set $\mu = 0$ and 10^α with $\alpha = 0 : -2 : -12$. We choose $J = 5$, $n = 50$, $m_j = 5$, $m = \sum_{j=1}^J m_j = 25$, $r = 5$, and $p = 3\max\{m, n\}r$. We generate $\{w_{i,j}\}_{\mathcal{G}}$ by performing hard thresholding on a random non-negative symmetric matrix with off-diagonal entries being uniformly distributed random numbers in the interval (0,1) and zero diagonal entries. Other parameters are set same as in the centralized framework. We again present the fitting error $\sum_{j=1}^J \|\mathcal{A}_j(\mathbf{U}^j\mathbf{V}_j^\top - \mathbf{X}_j^\star)\|_2^2$, the optimality error $\sum_{j=1}^J \|\mathbf{U}^j\mathbf{V}_j^\top - \mathbf{X}_j^\star\|_F^2$, and the *consensus* error $\sum_j \|\mathbf{U}^j - \mathbf{U}\|_F^2$ as a function of the iteration number in Fig. 2 (c), (d) and (e), respectively. While the existing literature does not guarantee a benign geometry for the distributed problem with regularization, we do see near-optimal convergence with near-consensus when the regularizer is sufficiently small. Our Corollary III.1 does apply to the distributed unregularized problem, and in this case we do indeed see global optimality and exact consensus.

V. Conclusion

This work closes the theory-practice gap for the factorization approach in low-rank matrix optimization when the cost function is restricted strongly convex and smooth by showing that the balancing regularizer is not necessary in geometric analysis, in agreement with practical observations. We have proved that any critical point of the unaltered factorized objective function (without regularizer) is either a global minimum or a strict saddle in both *centralized* and *distributed* settings.

[1] All errors in the centralized experiments eventually decay below 10^{-20}.

REFERENCES

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

[2] S. Burer and R. D. Monteiro, "Local minima and convergence in low-rank semidefinite programming," *Math. Program.*, vol. 103, no. 3, pp. 427–444, 2005.

[3] Q. Li, Z. Zhu, G. Tang, and M. B. Wakin, "Provable bregman-divergence based methods for nonconvex and non-lipschitz problems," 2019, *arXiv:1904.09712*.

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

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

[6] S. Li, G. Tang, and M. B. Wakin, "The landscape of non-convex empirical risk with degenerate population risk," in *Proc. Advances Neural Inform. Process. Syst.*, 2019, pp. 3502–3512.

[7] Q. Li, Z. Zhu, and G. Tang, "Geometry of factored nuclear norm regularization," 2017, *arxiv:1704.01265*.

[8] Z. Zhu, Q. Li, G. Tang, and M. B. Wakin, "The global optimization geometry of low-rank matrix optimization," 2017, *arXiv:1703.01256*.

[9] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," in *Proc. 34th Int. Conf. Mach. Learn.-Vol. 70*, 2017, pp. 1233–1242.

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

[11] Z. Zhu, D. Soudry, Y. C. Eldar, and M. B. Wakin, "The global optimization geometry of shallow linear neural networks," *J. Math. Imag. Vision*, vol. 62, pp. 279–292, 2020.

[12] D. Park, A. Kyrillidis, C. Carmanis, and S. Sanghavi, "Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach," in *Proc. Artif. Intell. Statist.*, 2017, pp. 65–74.

[13] L. Wang, X. Zhang, and Q. Gu, "A unified computational and statistical framework for nonconvex low-rank matrix estimation," in *Proc. Artif. Intell. Statist.*, 2017, pp. 981–990.

[14] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht, "Low-rank solutions of linear matrix equations via Procrustes flow," in *Proc. 33rd Int. Conf. Int. Conf. Mach. Learn.-Vol. 48*, 2016, pp. 964–973.

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

[16] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, "Finding low-rank solutions to matrix problems, efficiently and provably," 2016, *arXiv:1606.03168*.

[17] J. Sun, Q. Qu, and J. Wright, "Complete dictionary recovery over the sphere II: Recovery by Riemannian trust-region method," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 885–914, Feb. 2016.

[18] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points — Online stochastic gradient for tensor decomposition," in *Proc. 28th Conf. Learn. Theory*, 2015, pp. 797–842.

[19] J. D. Lee, M. Simchowitz, M. I. Jordan, and B. Recht, "Gradient descent only converges to minimizers," in *Proc. 29th Annu. Conf. Learn. Theory*, 2016, pp. 1246–1257.

[20] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 1724–1732.

[21] Q. Li, Z. Zhu, and G. Tang, "Alternating minimizations converge to second-order optimal solutions," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 3935–3943.

[22] Y. Nesterov and B. T. Polyak, "Cubic regularization of Newton method and its global performance," *Math. Program.*, vol. 108, no. 1, pp. 177–205, 2006.

[23] J. J. Moré and D. C. Sorensen, "Computing a trust region step," *SIAM J. Sci. Statist. Comput.*, vol. 4, no. 3, pp. 553–572, 1983.

[24] S. Lu, M. Hong, and Z. Wang, "PA-GD: On the convergence of perturbed alternating gradient descent to second-order stationary points for structured nonconvex optimization," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 4134–4143.

[25] C. Ma, Y. Li, and Y. Chi, "Beyond Procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing," in *Proc. 53rd Asilomar Conf. Signals, Syst., Comput.*, 2019, pp. 721–725.

[26] S. S. Du, W. Hu, and J. D. Lee, "Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced," in *Proc. Advances Neural Inform. Process. Syst.*, 2018, pp. 384–395.

[27] S. Arora, N. Cohen, W. Hu, and Y. Luo, "Implicit regularization in deep matrix factorization," in *Proc. Advances Neural Inform. Process. Syst.*, 2019, pp. 7411–7422.

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

[29] Z. Zhu, Q. Li, X. Yang, G. Tang, and M. B. Wakin, "Distributed low-rank matrix factorization with exact consensus," in *Proc. Advances Neural Inform. Process. Syst.*, 2019, pp. 8420–8430.

[30] M. Nouiehed and M. Razaviyayn, "Learning deep models: Critical points and local openness," 2018, *arXiv:1803.02968*.

[31] K. Kawaguchi, "Deep learning without poor local minima," in *Proc. Advances Neural Inform. Process. Syst.*, 2016, pp. 586–594.

[32] R. Sun and Z.-Q. Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6535–6579, Nov. 2016.