ON THE ASYMPTOTIC LINEAR CONVERGENCE OF GRADIENT DESCENT FOR NON-SYMMETRIC MATRIX COMPLETION

Trung Vu[1,2] *and Raviv Raich*[1]

[1] School of EECS, Oregon State University, Corvallis, OR 97331-5501
[2] Department of CSEE, University of Maryland, Baltimore County, MD 21250-0002
{vutru, raich}@oregonstate.edu

ABSTRACT

This paper studies a factorization-based gradient descent approach for non-symmetric matrix completion. We introduce an objective that includes an orthogonality regularization for one of the factors. Additionally, we introduce a scaling term to ensure that the two factors are of equal magnitude to improve the convergence speed. For the proposed objective, we analyze the exact linear convergence rate of gradient descent via the asymptotically linear update equation for the error matrix. Our proposed result is the first closed-form expression of the exact linear rate. To illustrate the correctness and tightness of our analysis, we compare the empirical convergence rate against the analytical rate. Additional numerical experiments are done to verify the efficacy of the scaling approach.

1. INTRODUCTION

Matrix completion has numerous applications in machine learning and signal processing, namely system identification [1], collaborative filtering [2], and dimension reduction [3]. This problem can be described as follows. Let $\boldsymbol{M} \in \mathbb{R}^{n\times m}$ be a rank r matrix with $1 \leq r \leq \min(m,n)$, and $\Omega = \{(i,j) \mid M_{ij} \text{ is observed}\}$ be an index subset of cardinality s such that $s \leq mn$. We wish to recover the unobserved entries of $\boldsymbol{M}$ by solving a rank minimization problem, subject to linear constraints consistent with the observations [4].

A body of work has been devoted to using first-order iterative methods for low-rank matrix completion that enforce the low-rank constraint using a factorization approach. By reparametrizing the $m \times n$ matrix as the product of two smaller matrices $\boldsymbol{M} = \boldsymbol{A}\boldsymbol{B}^\top$, for $\boldsymbol{A} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{B} \in \mathbb{R}^{m\times r}$, the resulting equivalent problem is unconstrained and more computationally efficient to solve [5]. Although this problem is non-convex, recent progress shows that for such a problem any local minimum is also a global minimum [6, 7]. Thus, basic optimization algorithms such as gradient descent [6, 8, 9] and alternating minimization [10–13] can provably solve matrix completion under a specific sampling regime. Existing convergence analyses of algorithms for low-rank matrix completion often rely on *standard assumptions*, such as the incoherence of the underlying matrix $\boldsymbol{M}$ and the uniform randomness of the sampling pattern Ω [4]. Under these assumptions and a sample complexity bound on the number of observed entries s, linear convergence to a global solution can be guaranteed (see [11] for alternating minimization, [9] for factorization-based gradient descent, and [14] for iterative hard thresholding), with an upper bound on the rate of convergence $\rho < 0.5$.

In this paper, we present a gradient descent (GD) approach for matrix completion that relies on the factorization of the matrix in the form of $\boldsymbol{A}\boldsymbol{B}^\top$. To ensure the linear convergence of gradient descent, we incorporate orthogonality constraints (up to a scaling factor) on the left factor $\boldsymbol{A}$. Our focus is on the theoretical analysis of the asymptotic convergence rate. By exploiting the *local* structure of the problem, we characterize the exact linear rate of local convergence of the algorithm. The closed-form expression we obtained can be used to determine sufficient conditions that ensure local linear convergence. Moreover, since our expression is exact, one can identify conditions that are potentially less stringent than existing conditions. Similar to the previous work [15], we derive the asymptotically linear update equation for the error matrix and obtain the convergence rate in terms of this matrix. As a sanity check, we conduct numerical experiments to illustrate the correctness of the formula. We apply the algorithm to solve the matrix completion problem in a few scenarios and compare the empirical convergence rate against the analytic expression for the rate we derived. The agreement confirms our derivation. While other upper bounds for the rate of convergence of GD for matrix completion are available in the literature [6], the expression for the rate provided in this paper is the *first* to provide an exact prediction of the asymptotic rate of convergence of GD for non-symmetric matrix completion.

Notation Throughout the paper, we use the notations $\|\cdot\|_F$ and $\|\cdot\|_2$ to denote the Frobenius norm and the spectral norm of a matrix, respectively. Occasionally, $\|\cdot\|_2$ is used on a vector to denote the Euclidean norm. Boldfaced symbols are reserved for vectors and matrices, while the elements of a vector/matrix are unbold. In addition, $\boldsymbol{I}_n$ denotes the $n \times n$ identity matrix, and $\boldsymbol{0}_{m\times n}$ denotes the $m \times n$ matrix of all zeros. We also use $\otimes$ to denote the Kronecker product between two matrices. For a matrix $\boldsymbol{X} \in \mathbb{R}^{m\times n}$, X_{ij} refers to the (i,j) element of $\boldsymbol{X}$. The spectral radius of $\boldsymbol{X}$ is the largest absolute value of the eigenvalues of $\boldsymbol{X}$, denoted by $\rho(\boldsymbol{X})$. The notation $\mathrm{vec}(\boldsymbol{X})$ denotes the vectorization of $\boldsymbol{X}$ by stacking its columns on top of one another.

2. PROBLEM FORMULATION

In this paper, we consider a GD approach to solving the problem of low-rank matrix completion [6,9]. In the (noiseless) low-rank matrix completion problem, the goal is to find a rank-r matrix $\boldsymbol{X}$ such that

$$X_{ij} = M_{ij}, \qquad (i,j) \in \Omega. \tag{1}$$

Here, $\boldsymbol{X}$ and $\boldsymbol{M}$ are $m \times n$ real-valued matrices and Ω is the sampling pattern that is a subset of $\{1,\ldots,m\} \times \{1,\ldots,n\}$. A factorization-based approach can be used to solve the aforementioned problem. Consider parameterizing $\boldsymbol{X}$ as $\boldsymbol{X} = \boldsymbol{A}\boldsymbol{B}^\top$, where $\boldsymbol{A}$ is $m \times r$ and $\boldsymbol{B}$ is $n \times r$. This parameterization forces the rank of $\boldsymbol{X}$ to be no more than r. Using this parametrization, one can

rephrase the low-rank matrix completion problem (in its noiseless case) as the problem of finding an $m \times r$ $\boldsymbol{A}$ and an $n \times r$ $\boldsymbol{B}$ such that $(\boldsymbol{A}\boldsymbol{B}^\top)_{ij} = M_{ij}$ for $(i,j) \in \Omega$. If we define the projection P_Ω such that

$$[\mathcal{P}_\Omega(\boldsymbol{X})]_{ij} = \begin{cases} X_{ij} & (i,j) \in \Omega \\ 0 & (i,j) \notin \Omega \end{cases}, \tag{2}$$

then we can write the low-rank matrix completion problem as

$$\mathcal{P}_\Omega(\boldsymbol{A}\boldsymbol{B}^\top - \boldsymbol{M}) = \boldsymbol{0}. \tag{3}$$

A simultaneous solution of all the equations in (3) is non-trivial. It is common to construct an objective function that is minimized when (3) holds. In [15], we analyzed the convergence of factorization-based gradient descent for such an objective function in the case of symmetric matrix completion. Adapting the convergence results for symmetric matrix completion in [15], one needs to perform some modification since the proposed matrix is not positive semi-definite and using symmetric matrix completion of the form $\boldsymbol{G} = \boldsymbol{X}\boldsymbol{X}^\top$ will not allow reducing the fitting error to zero. To remedy the problem, semidefinite lifting [16] in which $\boldsymbol{G} = [\boldsymbol{A};\boldsymbol{B}][\boldsymbol{A};\boldsymbol{B}]^\top = [\boldsymbol{A}\boldsymbol{A}^\top, \boldsymbol{M}; \boldsymbol{M}^\top, \boldsymbol{B}\boldsymbol{B}^\top]$ can be used, which in turn is compatible with symmetric matrix completion. However, the challenge is this method is not guaranteed to converge linearly due to ambiguity in the solution $(\boldsymbol{A}, \boldsymbol{B})$. Moreover, a combination of semidefinite lifting and the analysis in [15] yields a rate of 1, i.e., no linear convergence guarantees. The following section presents a novel regularized objective for non-symmetric matrix completion and the details of the gradient descent approach for minimizing such an objective.

2.1. A Gradient Descent Approach

To solve this problem efficiently in a large-scale setting, we consider a gradient descent approach. Specifically, we consider the following objective

$$f(\boldsymbol{A}, \boldsymbol{B}) = \frac{1}{2}\|\mathcal{P}_\Omega(\boldsymbol{A}\boldsymbol{B}^\top - \boldsymbol{M})\|_F^2 + \frac{1}{4}\|\boldsymbol{A}^\top\boldsymbol{A} - c\boldsymbol{I}_r\|_F^2, \tag{4}$$

where[1]

$$c = \sqrt{\frac{mn}{r|\Omega|}}\|\mathcal{P}_\Omega(\boldsymbol{M})\|_F. \tag{5}$$

Note that the first term of the objective can be minimized to zero, by selecting $\boldsymbol{A}$ and $\boldsymbol{B}$ such that $P_\Omega(\boldsymbol{A}\boldsymbol{B}^\top - \boldsymbol{M}) = \boldsymbol{0}$, as in the low-rank matrix completion formulation in (3). The second term in the objective is used to alleviate the ambiguity of the solution. In particular, if the singular value decomposition of $\boldsymbol{M}$ is given by $\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ where $\boldsymbol{U}$ is an $m \times r$ semi-orthogonal matrix, $\boldsymbol{\Sigma}$ is a non-negative diagonal matrix of dimension $r \times r$, and $\boldsymbol{V}$ is an $n \times r$ semi-orthogonal matrix. Then, $\boldsymbol{A}$ and $\boldsymbol{B}$ that satisfy $\boldsymbol{A}\boldsymbol{B}^\top = \boldsymbol{M}$ are given by $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{G}$ and $\boldsymbol{B} = \boldsymbol{V}\boldsymbol{\Sigma}(\boldsymbol{G}^{-1})^\top$ for any invertible $r \times r$ matrix $\boldsymbol{G}$. Minimizing the second term of the objective to zero implies a solution for $\boldsymbol{A}$ of the form $\boldsymbol{A} = \sqrt{c}\boldsymbol{U}\boldsymbol{Q}$ where $\boldsymbol{Q}$ is an orthogonal $r \times r$ matrix satisfying $\boldsymbol{Q}\boldsymbol{Q}^\top = \boldsymbol{Q}^\top\boldsymbol{Q} = \boldsymbol{I}_r$. We will show later, that this reduction of the ambiguity in the solution can allow us to establish a linear rate convergence. Note that the corresponding solution for $\boldsymbol{B}$ is given by $\boldsymbol{B} = \frac{1}{\sqrt{c}}\boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{Q}$. Using the objective in (4), we can obtain the objective gradient with respect to $\boldsymbol{A}$ and $\boldsymbol{B}$ as follows:

$$\begin{aligned} \frac{df}{d\boldsymbol{A}} &= \mathcal{P}_\Omega(\boldsymbol{A}\boldsymbol{B}^\top - \boldsymbol{M})\boldsymbol{B} + \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A} - c\boldsymbol{I}_r) \\ \frac{df}{d\boldsymbol{B}} &= \mathcal{P}_\Omega(\boldsymbol{A}\boldsymbol{B}^\top - \boldsymbol{M})^\top\boldsymbol{A}. \end{aligned} \tag{6}$$

Using the gradient, we can define the gradient descent iterations as:

$$\begin{aligned} \boldsymbol{A}_{k+1} &= \boldsymbol{A}_k - \eta[\boldsymbol{P}_k\boldsymbol{B}_k + \boldsymbol{A}_k(\boldsymbol{A}_k^\top\boldsymbol{A}_k - c\boldsymbol{I}_r)] \\ \boldsymbol{B}_{k+1} &= \boldsymbol{B}_k - \eta\boldsymbol{P}_k^\top\boldsymbol{A}_k, \end{aligned} \tag{7}$$

where $\boldsymbol{P}_k = P_\Omega(\boldsymbol{A}_k\boldsymbol{B}_k^\top - \boldsymbol{M})$ and η is the step-size parameter satisfying $\eta > 0$ (see Algorithm 1).

Algorithm 1 Factorization-based Gradient Descent

Input: $\boldsymbol{A}_0, \boldsymbol{B}_0, \mathcal{P}_\Omega(\boldsymbol{M}), \eta$
Output: $\{\boldsymbol{A}_k, \boldsymbol{B}_k\}$
1: **for** $k = 0, 1, 2, \ldots$ **do**
2: $\quad\boldsymbol{P}_k = \mathcal{P}_\Omega(\boldsymbol{A}_k\boldsymbol{B}_k^\top - \boldsymbol{M})$
3: $\quad\boldsymbol{A}_{k+1} = \boldsymbol{A}_k - \eta\big(\boldsymbol{P}_k\boldsymbol{B}_k + \boldsymbol{A}_k(\boldsymbol{A}_k^\top\boldsymbol{A}_k - c\boldsymbol{I}_r)\big)$
4: $\quad\boldsymbol{B}_{k+1} = \boldsymbol{B}_k - \eta\boldsymbol{P}_k^\top\boldsymbol{A}_k$

3. CONVERGENCE ANALYSIS

Due to the ambiguity in the solution for $\boldsymbol{A}$ and $\boldsymbol{B}$, instead of considering convergence of $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ to $\boldsymbol{A}$ and $\boldsymbol{B}$ respectively, we consider the convergence of the product of matrices $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ with matrices $\boldsymbol{A}_k^\top$ and $\boldsymbol{B}_k^\top$. In particular, we define the concatenation of $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ as $\boldsymbol{G}_k = [\boldsymbol{A}_k^\top\boldsymbol{B}_k^\top]^\top$ and similarly the concatenation of $\boldsymbol{A}$ and $\boldsymbol{B}$ as $\boldsymbol{G} = [\boldsymbol{A}^\top\boldsymbol{B}^\top]^\top$. We consider the convergence of $\boldsymbol{G}_k\boldsymbol{G}_k^\top$ to $\boldsymbol{G}\boldsymbol{G}^\top$. Note that in this approach, the limit $\boldsymbol{G}\boldsymbol{G}^\top$ is well-defined since $\boldsymbol{A}\boldsymbol{A}^\top$, $\boldsymbol{A}\boldsymbol{B}^\top$, and $\boldsymbol{B}\boldsymbol{B}^\top$ are all uniquely defined. Specifically, we have $\boldsymbol{A}\boldsymbol{A}^\top = c\boldsymbol{U}\boldsymbol{U}^\top$, $\boldsymbol{A}\boldsymbol{B}^\top = \boldsymbol{M}$, and $\boldsymbol{B}\boldsymbol{B}^\top = \frac{1}{c}\boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^\top = \frac{1}{c}\boldsymbol{M}^\top\boldsymbol{M}$. Note that the ambiguous $\boldsymbol{Q}$ in the solution for $\boldsymbol{A}$ and $\boldsymbol{B}$ cancels out in $\boldsymbol{A}\boldsymbol{A}^\top$, $\boldsymbol{A}\boldsymbol{B}^\top$, and $\boldsymbol{B}\boldsymbol{B}^\top$ and hence in $\boldsymbol{G}\boldsymbol{G}^\top$. Additionally, it can be shown that $\boldsymbol{G}_{k+1}\boldsymbol{G}_{k+1}^\top$ can be expressed as a function of $\boldsymbol{G}_k\boldsymbol{G}_k^\top$. Hence, a recursion of the form $\boldsymbol{G}_{k+1}\boldsymbol{G}_{k+1}^\top = f(\boldsymbol{G}_k\boldsymbol{G}_k^\top)$ and a fixed point approach can be used to facilitate the analysis. We start by stating the update equation on the elements of $\boldsymbol{G}_k\boldsymbol{G}_k^\top$ and proceed with the structural constraints on $\boldsymbol{G}_k\boldsymbol{G}_k^\top$.

3.1. Update Equations

The update equations are obtained by taking the products of the matrices in (7). Since our goal is to analyze the asymptotic behavior of the GD iterations, we focus our analysis on the leading terms. Specifically, if we denote $\boldsymbol{G}_k\boldsymbol{G}_k^\top - \boldsymbol{G}\boldsymbol{G}^\top$ by $\boldsymbol{E}_k$, then asymptotically (when $\boldsymbol{E}_k \to \boldsymbol{0}$) terms that are $\mathcal{O}(\|\boldsymbol{E}_k\|^2)$ can be neglected. Hence, we consider the following update equations and omit the $\mathcal{O}(\|\boldsymbol{E}_k\|^2)$ from the RHS. Due to space limitations, intermediate calculations are omitted and only key steps are provided.

$$\begin{aligned} \boldsymbol{A}_{k+1}\boldsymbol{A}_{k+1}^\top = \boldsymbol{A}_k\boldsymbol{A}_k^\top &- \eta[\boldsymbol{P}_k\boldsymbol{B}_k\boldsymbol{A}_k^\top + \boldsymbol{A}_k\boldsymbol{B}_k^\top\boldsymbol{P}_k^\top \\ &+2\boldsymbol{A}_k(\boldsymbol{A}_k^\top\boldsymbol{A}_k - c\boldsymbol{I}_r)\boldsymbol{A}_k^\top] \end{aligned} \tag{9}$$

$$\begin{aligned} \boldsymbol{A}_{k+1}\boldsymbol{B}_{k+1}^\top = \boldsymbol{A}_k\boldsymbol{B}_k^\top &- \eta[\boldsymbol{P}_k\boldsymbol{B}_k\boldsymbol{B}_k^\top + \boldsymbol{A}_kA_k^\top P_k \\ &+\boldsymbol{A}_k(\boldsymbol{A}_k^\top\boldsymbol{A}_k - c\boldsymbol{I}_r)\boldsymbol{B}_k^\top] \end{aligned} \tag{10}$$

$$\boldsymbol{B}_{k+1}^\top\boldsymbol{B}_{k+1}^\top = \boldsymbol{B}_k\boldsymbol{B}_k^\top - \eta[\boldsymbol{M}\boldsymbol{P}_k^\top\boldsymbol{A}_k\boldsymbol{B}_k^\top + \boldsymbol{B}_k\boldsymbol{A}_k^\top\boldsymbol{P}_k] \tag{11}$$

[1]This choice of c is aimed to ensure that $\|\boldsymbol{A}\|_F \approx \|\boldsymbol{B}\|_F$. Due to space limitations, we omit the proof.

$$\boldsymbol{H} = \left[\begin{array}{c|c|c|c} 2c\boldsymbol{P}_A \otimes \boldsymbol{P}_A & (\boldsymbol{I}_m \otimes \boldsymbol{M})\boldsymbol{T}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{T}^\top & (\boldsymbol{M} \otimes \boldsymbol{I}_m)\boldsymbol{S}\boldsymbol{S}^\top & \boldsymbol{0}_{m^2\times n^2} \\ \boldsymbol{I}_m \otimes \boldsymbol{M}^\top & (\boldsymbol{I}_m \otimes \boldsymbol{B}\boldsymbol{B}^\top + c\boldsymbol{P}_A \otimes \boldsymbol{I}_n)\boldsymbol{T}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{T}^\top - c\boldsymbol{P}_A^\perp \otimes \boldsymbol{I}_n & \boldsymbol{0}_{mn\times mn} & \boldsymbol{0}_{mn\times n^2} \\ \boldsymbol{M}^\top \otimes \boldsymbol{I}_m & \boldsymbol{0}_{mn\times mn} & (\boldsymbol{B}\boldsymbol{B}^\top \otimes \boldsymbol{I}_m) + c(\boldsymbol{I}_n \otimes \boldsymbol{P}_A)\boldsymbol{S}\boldsymbol{S}^\top - c\boldsymbol{I}_n \otimes \boldsymbol{P}_A^\perp & \boldsymbol{0}_{mn\times n^2} \\ \boldsymbol{0}_{n^2\times m^2} & (\boldsymbol{M}^\top \otimes \boldsymbol{I}_n)\boldsymbol{T}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{T}^\top & (\boldsymbol{I}_n \otimes \boldsymbol{M}^\top)\boldsymbol{S}\boldsymbol{S}^\top & \boldsymbol{0}_{n^2\times n^2} \end{array}\right]. \quad (8)$$

Let $\boldsymbol{E}_{AA}^k = \boldsymbol{A}_k\boldsymbol{A}_k^\top - \boldsymbol{A}\boldsymbol{A}^\top$, $\boldsymbol{E}_{AB}^k = \boldsymbol{A}_k\boldsymbol{B}_k^\top - \boldsymbol{A}\boldsymbol{B}^\top$, $\boldsymbol{E}_{BA}^k = \boldsymbol{B}_k\boldsymbol{A}_k^\top - \boldsymbol{B}\boldsymbol{A}^\top$, and $\boldsymbol{E}_{BB}^k = \boldsymbol{B}_k\boldsymbol{B}_k^\top - \boldsymbol{B}\boldsymbol{B}^\top$. We can substitute the definition of the different $\boldsymbol{E}_k$ terms into (9)-(11) and remove terms that are $\mathcal{O}(\|\boldsymbol{E}_k\|^2)$, simplify, and obtain update equations on the error terms:

$$\begin{aligned} \boldsymbol{E}_{AA}^{k+1} &= \boldsymbol{E}_{AA}^k - \eta[\boldsymbol{P}_k\boldsymbol{M}^\top + \boldsymbol{M}\boldsymbol{P}_k^\top + 2c\boldsymbol{P}_A\boldsymbol{E}_{AA}^k\boldsymbol{P}_A] \\ \boldsymbol{E}_{AB}^{k+1} &= \boldsymbol{E}_{AB}^k - \eta[\boldsymbol{P}_k\boldsymbol{B}\boldsymbol{B}^\top + c\boldsymbol{P}_A\boldsymbol{P}_k + \boldsymbol{E}_{AA}^k\boldsymbol{M} - c\boldsymbol{P}_A^\perp\boldsymbol{E}_{AB}^k] \\ \boldsymbol{E}_{BB}^{k+1} &= \boldsymbol{E}_{BB}^k - \eta[\boldsymbol{P}_k^\top\boldsymbol{M} + \boldsymbol{M}^\top\boldsymbol{P}_k]. \end{aligned} \quad (12)$$

Here, $\boldsymbol{P}_A = \boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top$ and $\boldsymbol{P}_A{}^\perp = \boldsymbol{I}_m - \boldsymbol{P}_A$. Using these updates, we can obtain an updated equation of the form

$$\boldsymbol{E}_{k+1} = f(\boldsymbol{E}_k) + \mathcal{O}(\|\boldsymbol{E}_k\|^2) \quad \text{where } \boldsymbol{E}_k = \left[\begin{array}{cc} \boldsymbol{E}_{AA}^k & \boldsymbol{E}_{AB}^k \\ \boldsymbol{E}_{BA}^k & \boldsymbol{E}_{BB}^k \end{array}\right]$$

and $f(\cdot)$ is a linear mapping $f : \mathcal{E} \to \mathcal{E}$, where

$$\mathcal{E} = \left\{ \boldsymbol{X}\boldsymbol{X}^\top - \left[\begin{array}{cc} \boldsymbol{A}\boldsymbol{A}^\top & \boldsymbol{A}\boldsymbol{B}^\top \\ \boldsymbol{B}\boldsymbol{A}^\top & \boldsymbol{B}\boldsymbol{B}^\top \end{array}\right] \mid \boldsymbol{X} \in \mathbb{R}^{(m+n)\times r} \right\}.$$

In other words, the space of all matrices can be written as the difference of two symmetric rank r $(m+n)\times(m+n)$ matrices. To obtain the linear mapping as a matrix, we consider using the vectorized version of $\boldsymbol{E}$. This involves the vectorized form of the error components, i.e., $\boldsymbol{e}_{aa}^k = \text{vec}(\boldsymbol{E}_{AA}^k)$, $\boldsymbol{e}_{ab}^k = \text{vec}(\boldsymbol{E}_{AB}^k)$, $\boldsymbol{e}_{ba}^k = \text{vec}(\boldsymbol{E}_{BA}^k)$, and $\boldsymbol{e}_{bb}^k = \text{vec}(\boldsymbol{E}_{BB}^k)$. Using this notation, we can write the update on the vectorized error terms by taking the vec operator on the corresponding matrix equations in (12). This yields:

$$\begin{aligned} \boldsymbol{e}_{aa}^{k+1} &= \boldsymbol{e}_{aa}^k - \eta[(\boldsymbol{M} \otimes \boldsymbol{I}_m)\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{e}_{ab}^k \\ &\quad +(\boldsymbol{I}_m \otimes \boldsymbol{M})\boldsymbol{T}_{mn}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{T}_{mn}^\top\boldsymbol{e}_{ba}^k + 2c(\boldsymbol{P}_A \otimes \boldsymbol{P}_A)\boldsymbol{e}_{aa}^k] \\ \boldsymbol{e}_{ab}^{k+1} &= \boldsymbol{e}_{ab}^k - \eta[(\boldsymbol{B}\boldsymbol{B}^\top \otimes \boldsymbol{I}_m)\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{e}_{ab}^k + c(\boldsymbol{I}_n \otimes \boldsymbol{P}_A)\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{e}_{ab}^k \\ &\quad +(\boldsymbol{M}^\top \otimes \boldsymbol{I}_m)\boldsymbol{e}_{aa}^k - c(\boldsymbol{I}_n \otimes \boldsymbol{P}_A^\perp)\boldsymbol{e}_{ab}^k] \\ \boldsymbol{e}_{ba}^{k+1} &= \boldsymbol{e}_{ba}^k - \eta[(\boldsymbol{I}_m \otimes \boldsymbol{B}\boldsymbol{B}^\top)\boldsymbol{T}_{mn}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{T}_{mn}^\top\boldsymbol{e}_{ba}^k \\ &\quad +c(\boldsymbol{P}_A \otimes \boldsymbol{I}_n)\boldsymbol{T}_{mn}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{T}_{mn}^\top\boldsymbol{e}_{ba}^k + (\boldsymbol{I}_m \otimes \boldsymbol{M}^\top)\boldsymbol{e}_{aa}^k \\ &\quad -c(\boldsymbol{P}_A^\perp \otimes \boldsymbol{I})\boldsymbol{e}_{ba}^k] \\ \boldsymbol{e}_{bb}^{k+1} &= \boldsymbol{e}_{bb}^k - \eta[(\boldsymbol{M}^\top \otimes \boldsymbol{I}_m)\boldsymbol{T}_{mn}\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{T}_{mn}^\top\boldsymbol{e}_{ab}^k \\ &\quad +(\boldsymbol{I}_m \otimes \boldsymbol{M}^\top)\boldsymbol{S}\boldsymbol{S}^\top\boldsymbol{e}_{ab}^k]. \end{aligned} \quad (13)$$

Here $\boldsymbol{T}_{mn}$ is the permutation matrix satisfying $\boldsymbol{T}_{mn}\text{vec}(\boldsymbol{X}) = \text{vec}(\boldsymbol{X}^\top)$ for all $\boldsymbol{X} \in \mathbb{R}^{m\times n}$ and $\boldsymbol{S} \in \mathbb{R}^{mn\times s}$. In a matrix form,

$$\left[\begin{array}{c} \boldsymbol{e}_{aa}^{k+1} \\ \boldsymbol{e}_{ba}^{k+1} \\ \boldsymbol{e}_{ab}^{k+1} \\ \boldsymbol{e}_{bb}^{k+1} \end{array}\right] = (\boldsymbol{I}_{(m+n)^2} - \eta\boldsymbol{H}) \left[\begin{array}{c} \boldsymbol{e}_{aa}^k \\ \boldsymbol{e}_{ba}^k \\ \boldsymbol{e}_{ab}^k \\ \boldsymbol{e}_{bb}^k \end{array}\right], \quad (14)$$

where $\boldsymbol{H}$ is given in (8). We can write the vec operator of the error matrix $\boldsymbol{E}_k$ in terms of the vectorized error terms as

$$\text{vec}(\boldsymbol{E}_k) = \text{vec}\left(\left[\begin{array}{cc} \boldsymbol{E}_{AA}^k & \boldsymbol{E}_{AB}^k \\ \boldsymbol{E}_{BA}^k & \boldsymbol{E}_{BB}^k \end{array}\right]\right) = \boldsymbol{Z}\left[\begin{array}{c} \boldsymbol{e}_{aa}^k \\ \boldsymbol{e}_{ba}^k \\ \boldsymbol{e}_{ab}^k \\ \boldsymbol{e}_{bb}^k \end{array}\right], \quad (15)$$

where $\boldsymbol{Z}$ is a permutation matrix given by:

$$\begin{aligned} \boldsymbol{Z} = \text{diag}(&\left[\boldsymbol{I}_m \otimes \left[\begin{array}{c} \boldsymbol{I}_m \\ \boldsymbol{0}_{n\times m} \end{array}\right], \boldsymbol{I}_m \otimes \left[\begin{array}{c} \boldsymbol{0}_{m\times n} \\ \boldsymbol{I}_n \end{array}\right]\right], \\ &\left[\boldsymbol{I}_n \otimes \left[\begin{array}{c} \boldsymbol{I}_m \\ \boldsymbol{0}_{n\times m} \end{array}\right], \boldsymbol{I}_n \otimes \left[\begin{array}{c} \boldsymbol{0}_{m\times n} \\ \boldsymbol{I}_n \end{array}\right]\right]). \end{aligned} \quad (16)$$

Using equations (14) and (15), we can now write the update equation for $\boldsymbol{E}^k$ as follows:

$$\text{vec}(\boldsymbol{E}_{k+1}) = \boldsymbol{Z}(\boldsymbol{I}_{(m+n)^2} - \eta\boldsymbol{H})\boldsymbol{Z}^\top\text{vec}(\boldsymbol{E}_k). \quad (17)$$

As mentioned in [15], the integration of structural constraints is a necessary step in the analysis of the convergence behavior or the error matrix. We proceed by characterizing the structural constraints.

3.2. Integrating Structural Constraints

Recall the definition of $\mathcal{E}$ the set of all error matrices. It can be shown that a symmetry constraint and a rank r constraint can be used to characterize the set $\mathcal{E}$. The two constraints can be captured as follows. For symmetry, we have $\boldsymbol{E} = (\boldsymbol{E}+\boldsymbol{E}^\top)/2$. To satisfy the rank constraint, we can consider the manifold of all rank r matrices, $\mathcal{M}_r$. The first-order term of the error matrix resides in the tangent space $T_{\mathcal{M}_r}(\boldsymbol{G}\boldsymbol{G}^\top)$ and hence it satisfies:

$$\boldsymbol{E} = \boldsymbol{E} - \boldsymbol{P}_G^\perp\boldsymbol{E}\boldsymbol{P}_G^\perp + \mathcal{O}(\|\boldsymbol{E}\|^2).$$

See [15] for details. Vectorizing the two requirements, we can write them as

$$\begin{aligned} \text{vec}(\boldsymbol{E}) &= \frac{1}{2}(\boldsymbol{I}_{(m+n)^2} + \boldsymbol{T}_{(m+n)^2})\text{vec}(\boldsymbol{E}) \\ \text{vec}(\boldsymbol{E}) &= (\boldsymbol{I}_{(m+n)^2} - \boldsymbol{P}_G^\perp \otimes \boldsymbol{P}_G^\perp)\text{vec}(\boldsymbol{E}), \end{aligned} \quad (18)$$

where $\boldsymbol{T}$ is a commutation matrix such that $\boldsymbol{T}\,\text{vec}(\boldsymbol{E}) = \text{vec}(\boldsymbol{E}^\top)$. Letting $\boldsymbol{P}_1 = \frac{1}{2}(\boldsymbol{I}_{(m+n)^2} + \boldsymbol{T}_{(m+n)^2})$ and $\boldsymbol{P}_2 = (\boldsymbol{I}_{(m+n)^2} - \boldsymbol{P}_G^\perp \otimes \boldsymbol{P}_G^\perp)$, we can write the projection on the intersection of two requirements as

$$\boldsymbol{P} = \boldsymbol{P}_1\boldsymbol{P}_2 = \boldsymbol{P}_2\boldsymbol{P}_1. \quad (19)$$

Following [15], this product $\boldsymbol{P}$ is also a projection. Hence, once can locally characterize the tangent space of the symmetric rank r $(m+n)\times(m+n)$ matrices at $\boldsymbol{G}\boldsymbol{G}^\top$ using $\text{vec}(\boldsymbol{E}) = \boldsymbol{P}\text{vec}(\boldsymbol{E})$. The structural constraints can be incorporated into the update equation in (17) as follows:

$$\text{vec}(\boldsymbol{E}_{k+1}) = \boldsymbol{P}\boldsymbol{Z}(\boldsymbol{I}_{(m+n)^2} - \eta\boldsymbol{H})\boldsymbol{Z}^\top\boldsymbol{P}\text{vec}(\boldsymbol{E}_k). \quad (20)$$

In other words, the structural constraints on $\boldsymbol{E}_k$ are applied before and after the update. Since $\boldsymbol{P}$ is a projection matrix it can be written as $\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{Q}^\top$ where $\boldsymbol{Q}$ is an $(m+n)^2 \times d$ orthogonal matrix and $d = r(2(m+n)+1-r)/2$ is the rank of $\boldsymbol{P}$ that is also the dimension of the manifold of all $(m+n)\times(m+n)$ symmetric rank-r matrices. Using $\boldsymbol{Q}$, we can write an economy version of the update equation:

$$\boldsymbol{Q}^\top\text{vec}(\boldsymbol{E}_{k+1}) = \boldsymbol{Q}^\top\boldsymbol{Z}(\boldsymbol{I}_{(m+n)^2} - \eta\boldsymbol{H})\boldsymbol{Z}^\top\boldsymbol{Q}\boldsymbol{Q}^\top\text{vec}(\boldsymbol{E}_k). \quad (21)$$

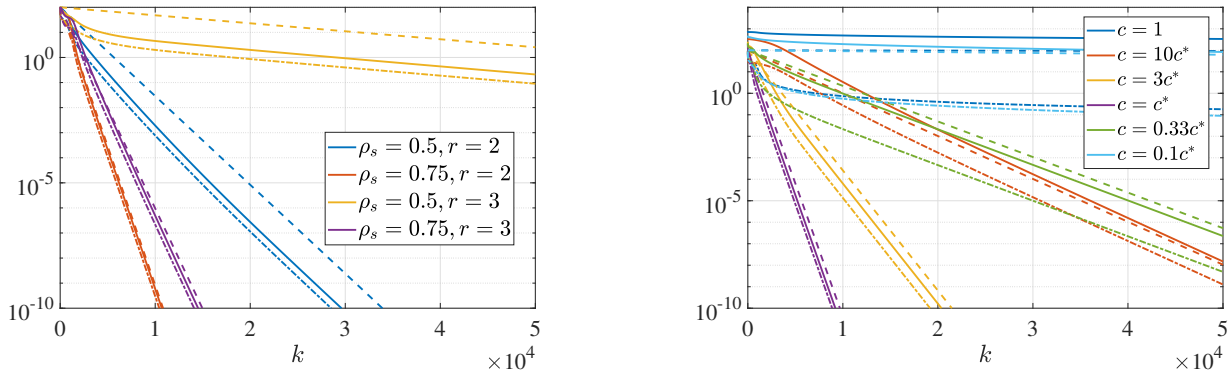

Fig. 1. (Left): a plot of the empirical error terms $\|\boldsymbol{E}_k\|$ (solid), $\|\boldsymbol{E}_{AB}^k\|$ (dash-dotted), and ρ^k (dashed) as a function of the number of iterations for the 4 settings described in Section 4 using a blue curve for $(\rho_s = 0.5, r = 2)$, red curves for $(0.75, 2)$, yellow curves for $(0.5, 3)$, and purple curves for $(0.75, 3)$. (Right): a plot of the empirical error terms $\|\boldsymbol{E}_k\|$ (solid), $\|\boldsymbol{E}_{AB}^k\|$ (dash-dotted), and ρ^k (dashed) as a function of the number of iterations for the setting of $n = 30$, $m = 20$, $r = 2$, and $\rho_s = 0.75$ for 6 different values of c in $\{1, 10c^*, 3c^*, c^*, c^*/3, c^*/10\}$ corresponding to the blue, red, yellow, purple, green, and cyan, respectively.

Note that $\|\boldsymbol{Q}^\top \text{vec}(\boldsymbol{E}^k)\| = \|\text{vec}(\boldsymbol{E}_k)\| + O(\|\text{vec}(\boldsymbol{E}_k)\|^2)$ and hence the asymptotic convergence rate of the error term $\boldsymbol{Q}^\top \text{vec}(\boldsymbol{E}_k)$ is the same the asymptotic convergence rate of the error term $\text{vec}(\boldsymbol{E}_k)$. Finally, the convergence rate can be obtained as the spectral radius of the matrix $\boldsymbol{Q}^\top \boldsymbol{Z}(\boldsymbol{I}_{(m+n)^2} - \eta \boldsymbol{H})\boldsymbol{Z}^\top \boldsymbol{Q} = \boldsymbol{I}_d - \eta \hat{\boldsymbol{H}}$, where $\hat{\boldsymbol{H}} = \boldsymbol{Q}^\top \boldsymbol{Z} \boldsymbol{H} \boldsymbol{Z}^\top \boldsymbol{Q} \in \mathbb{R}^{d \times d}$.

3.3. Main Result

We conclude this section by formally stating our convergence rate analysis in the following theorem.

Theorem 1. *Let matrices $\boldsymbol{P}$, $\boldsymbol{Z}$, and $\boldsymbol{H}$ be as in (19), (16), and (8), respectively. Additionally, the projection matrix $\boldsymbol{P}$ can be decomposed as $\boldsymbol{P} = \boldsymbol{Q}\boldsymbol{Q}^\top$ where $\boldsymbol{Q}$ is a $(m+n)^2 \times d$ semi-orthogonal matrix. Finally, define matrix $\hat{\boldsymbol{H}} = \boldsymbol{Q}^\top \boldsymbol{Z} \boldsymbol{H} \boldsymbol{Z}^\top \boldsymbol{Q}$. If $\hat{\boldsymbol{H}}$ is non-singular, then Algorithm 1 produces a sequence of matrices $\boldsymbol{A}_k \boldsymbol{B}_k^\top$ converging to $\boldsymbol{M}$ at an asymptotic linear rate $\rho(\boldsymbol{I}_d - \eta \hat{\boldsymbol{H}})$. Formally, there exists a neighborhood $\mathcal{N}(\boldsymbol{M})$ of $\boldsymbol{M}$ such that for any $\boldsymbol{A}_0 \boldsymbol{B}_0^\top \in \mathcal{N}(\boldsymbol{M})$,*

$$\|\boldsymbol{A}_k \boldsymbol{B}_k^\top - \boldsymbol{M}\|_F \leq C \|\boldsymbol{A}_0 \boldsymbol{B}_0^\top\|_F \rho(\boldsymbol{I}_d - \eta \hat{\boldsymbol{H}})^k, \qquad (22)$$

for some numerical constant $C > 0$.

4. NUMERICAL EXPERIMENTS

In this section, we conduct a numerical experiment to verify the validity of the theoretical analysis of the rate of convergence and to assess the provided choice of the constant c in the algorithm.

4.1. Theoretical Rate Verification

To assess the validity of our analysis, we perform the following experiment. We consider matrix completion for a 20×30 matrix ($m = 20$ and $n = 30$) and four settings for (ρ_s, r), i.e., the pair (ratio of known entries, rank): $(0.5, 2)$, $(0.75, 2)$, $(0.5, 3)$, and $(0.75, 3)$. For each setting, we generate the factors $\boldsymbol{A}$ and $\boldsymbol{B}$ such that their entries are $i.i.d$ following the standard normal distribution. We then select uniformly at random $\lceil \rho_s mn \rceil$ of the matrix entries to be the known set of entries. For each of the four settings, we run Algorithm 1 with c as prescribed in (5) for $50,000$ iteration. We use a small step size $\eta = 0.0005$ to make sure the linear convergence occurs in all settings.[2] We initialize $\boldsymbol{A}_k$ and $\boldsymbol{B}_k$ (at $k = 0$) such that their entries are $i.i.d$ following the standard normal distribution. At each iteration, we compute $\|\boldsymbol{E}_{AB}^k\|$ and $\|\boldsymbol{E}_k\|$. In Fig. 1(Left), we plot for each of the aforementioned settings, $\|\boldsymbol{E}_k\|$ (solid), $\|\boldsymbol{E}_{AB}^k\|$ (dash-dotted), and ρ^k (dashed), where $\rho = \rho(\boldsymbol{I} - \eta \hat{\boldsymbol{H}})$ is the theoretical rate given by Theorem 1, as a function of the number of iterations k. We note that in each of the settings the rate at which $\|\boldsymbol{E}_{AB}^k\|$ and $\|\boldsymbol{E}_k\|$ decrease to zero matches the rate at which ρ^k decreases to zero.

4.2. The Impact of c

Here, we study the impact of the choice of c. We select the setting of $m = 20$, $n = 30$, $\rho_s = 0.5$, and $r = 2$. We generate a matrix using the aforementioned data generation process and apply the GD with the following values of c: $\{1, 10c^*, 3c^*, c^* c^*/3, c^*/10\}$ where c^* is given in (5). In Fig. 1(Right), we plot for each value of c, the value of $\|\boldsymbol{E}_k\|$ (solid), $\|\boldsymbol{E}_{AB}^k\|$ (dash-dotted), and ρ^k (dashed) as functions of k. We observe that the choice of $c = c^*$ provides the fastest rate among the 6 values of c. In particular, the case of $c = 1$ highlights the challenge of using a GD with the same step size for the $\boldsymbol{A}$ and $\boldsymbol{B}$ update when the two matrices are significantly different in terms of their norms.

5. CONCLUSION

In this paper, we presented a variant of GD for matrix completion that is based on parameterizing the matrix as a product of two factors, wherein the first factor is assumed to be a semi-orthogonal times a constant. The GD algorithm provides a simultaneous update to the two factors by minimizing an objective consisting of a fitting term and a regularization term that ensures that the first factor is a semi-orthogonal matrix (up to a scaling factor). A detailed analysis was provided to establish conditions for linear convergence of the

[2] In each specific setting, one can choose an optimal step size by maximizing the rate $\rho(\boldsymbol{I}_d - \eta \hat{\boldsymbol{H}})$ w.r.t. $\eta > 0$.

algorithm and a formula for the asymptotic linear rate was provided. Additionally, a formula was provided to determine the scaling constant for the semi-orthogonal factor. Numerical experiments were conducted to verify the correctness of the linear rate expression and to illustrate the merit of selecting the scaling constant. In contrast, the proposed c^* yields identical norms for $\boldsymbol{A}$ and $\boldsymbol{B}$ and appears to produce the fastest rate of convergence.

6. REFERENCES

[1] Zhang Liu and Lieven Vandenberghe, "Interior-point method for nuclear norm approximation with application to system identification," *SIAM J. Matrix Anal. Appl.*, vol. 31, no. 3, pp. 1235–1256, 2009.

[2] Jasson DM Rennie and Nathan Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proc. Int. Conf. Mach. Learn.*, 2005, pp. 713–719.

[3] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 11, 2011.

[4] Emmanuel J Candès and Benjamin Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, pp. 717, 2009.

[5] Samuel Burer and Renato DC Monteiro, "Local minima and convergence in low-rank semidefinite programming," *Math. Program.*, vol. 103, no. 3, pp. 427–444, 2005.

[6] Ruoyu Sun and Zhi-Quan Luo, "Guaranteed matrix completion via non-convex factorization," *IEEE Trans. Inf. Theory*, vol. 62, no. 11, pp. 6535–6579, 2016.

[7] Rong Ge, Jason D Lee, and Tengyu Ma, "Matrix completion has no spurious local minimum," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2973–2981.

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

[9] Cong Ma, Kaizheng Wang, Yuejie Chi, and Yuxin 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.

[10] Caihua Chen, Bingsheng He, and Xiaoming Yuan, "Matrix completion via an alternating direction method," *IMA J. Numer. Anal.*, vol. 32, no. 1, pp. 227–245, 2012.

[11] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi, "Low-rank matrix completion using alternating minimization," in *Proc. Annu. ACM Symp. Theory Comput.*, 2013, pp. 665–674.

[12] Moritz Hardt, "Understanding alternating minimization for matrix completion," in *Proc. Annu. IEEE Symp. Found. Comput.*, 2014, pp. 651–660.

[13] Moritz Hardt and Mary Wootters, "Fast matrix completion without the condition number," in *Proc. Conf. Learn. Theory*, 2014, pp. 638–678.

[14] Lijun Ding and Yudong Chen, "Leave-one-out approach for matrix completion: Primal and dual analysis," *IEEE Trans. Inf. Theory*, vol. 66, no. 11, pp. 7274–7301, 2020.

[15] Trung Vu and Raviv Raich, "Exact linear convergence rate analysis for low-rank symmetric matrix completion via gradient descent," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2021, pp. 3240–3244.

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