On Local Linear Convergence Rate of Iterative Hard Thresholding for Matrix Completion

Trung Vu, *Graduate Student Member, IEEE,* Evgenia Chunikhina, and Raviv Raich, *Senior Member, IEEE*

***Abstract*—Iterative hard thresholding (IHT) has gained in popularity over the past decades for large-scale optimization. However, the convergence properties of this method have only been explored recently in non-convex settings. Existing works in matrix completion provide global convergence guarantees for IHT. To do so, they rely on standard assumptions such as incoherence property and uniform sampling. While such analysis provides a global upper bound on the linear convergence rate, it does not describe the actual performance of IHT in practice. In this paper, we provide a novel insight into the local convergence of IHT for matrix completion. We uncover the exact asymptotic linear rate of IHT in closed-form and identify the region in which the algorithm is guaranteed to converge. Furthermore, we utilize random matrix theory to study the linear rate of convergence of IHT for large-scale matrix completion. We find that asymptotically, the rate can be expressed explicitly in terms of the relative rank and the sampling rate. Finally, we present numerical results to verify our theoretical analysis.**

***Index Terms*—Matrix completion, iterative hard thresholding, local convergence analysis, random matrix theory.**

I. Introduction

Matrix completion is a fundamental problem that arises in many areas of signal processing and machine learning such as collaborative filtering [1]–[4], system identification [5]–[7] and dimension reduction [8], [9]. The problem can be explained as follows. Let $\boldsymbol{M} \in \mathbb{R}^{n_1 \times n_2}$ be the underlying matrix with rank r and Ω be the set of locations corresponding to the observed entries of $\boldsymbol{M}$, i.e., $(i,j) \in \Omega$ if M_{ij} is observed. The goal is to recover the unknown entries of $\boldsymbol{M}$, belonging to the complement set $\bar{\Omega}$.

To understand the feasibility of matrix completion, let us represent $\boldsymbol{M}$ using its singular value decomposition as

$$\boldsymbol{M} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top,$$

where σ_i is the i-th largest singular value of $\boldsymbol{M}$, $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$ are the corresponding left and right singular vectors. Since each set of the left and right singular vectors are orthonormal, the degrees of freedom (DoF) of matrix completion is given by

$$r + \sum_{i=1}^{r} (n_1 - i) + \sum_{j=1}^{r} (n_2 - j) = (n_1 + n_2 - r) r,$$

which is significantly less than the total number of entries in $\boldsymbol{M}$ when r is small. This implies the possibility of recovering the entire matrix even when only a few entries are observed. However, not every matrix with more than $(n_1 + n_2 - r)r$ observed entries can be completed. For instance, if an entire column (or row) of a rank-one matrix is missing, then the matrix cannot be recovered. Similarly, if a low-rank matrix contains too many zero entries, then the observed entries might end up being all zero, thereby not providing any clue about the missing entries.

The aforementioned argument motivates two well-known assumptions in matrix completion: the incoherence property of the underlying matrix $\boldsymbol{M}$ and the uniform sampling model for Ω [10]. Under these assumptions, there has been a long line of work on provable methods for globally solving matrix completion. Based on the formulation of the optimization problem, these methods fall into three major categories: linearly constrained nuclear norm minimization, low-rank factorization, and rank-constrained least squares (see Table I). The first approach, nuclear norm minimization, is a convex relaxation of the original rank-constraint problem and can be solved using proximal-type algorithms [11], [12], [29]. It is noted that such algorithms often come with sublinear convergence guarantees in the literature. The second approach, low-rank factorization, stems from the Burer-Monteiro factorization [20], whereby the low-rank matrix is viewed as a product of two low-rank components. The resulting least-squares problem is unconstrained albeit non-convex. Recent progress in this approach has shown that basic optimization procedures, such as gradient descent [19] and alternating minimization [16], converge linearly to the global solution at a rate at most 0.5, under the assumption that the number of known entries is sufficiently large. The third approach, rank-constrained least squares, is also a non-convex formulation of matrix completion. One of the most popular algorithms for solving this optimization problem is iterative hard thresholding (IHT) [22]. When converging to a low-rank solution, it is stated in [22] that hard-thresholding algorithms are more efficient than their soft-thresholding counterparts in both computational complexity per iteration and convergence speed. Interestingly, by assuming a sufficiently large number of known entries, Ding and Chen [23] have shown that IHT with a specific choice of step size converges linearly to the global minimum at a rate at most 0.5.

While the aforementioned analyses for matrix completion are powerful, they provide universal bounds on the convergence rate that are conservative. These bounds are primarily developed to prove the convergence to a global solution of the problem but may not offer a precise estimate of the

Trung Vu and Raviv Raich are with the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR 97331, USA.

Evgenia Chunikhina is with Department of Mathematics and Computer Science, Pacific University, Forest Grove, OR 97116, USA.

E-mails: vutru@oregonstate.edu; chunikhina@pacificu.edu; and raich@oregonstate.edu.

This work was partially supported by the National Science Foundation grant CCF-1254218.

Manuscript received 2022.

TABLE I: Three well-known formulations of the matrix completion problem.

Problem formulation	Description	Algorithms
Linearly constrained nuclear norm minimization	$\min_{\boldsymbol{X}\in\mathbb{R}^{n_1\times n_2}} \|\boldsymbol{X}\|_* \text{ s.t. } X_{ij}=M_{ij}, \quad (i,j)\in\Omega$	Semi-definite programming (SDP) [10], singular value thresholding (SVT) [11], accelerated proximal gradient (APG) [12], conditional gradient descent (CGD) [13]–[15]
Low-rank factorization	$\min_{\boldsymbol{Y}\in\mathbb{R}^{n_1\times r},\boldsymbol{Z}\in\mathbb{R}^{n_2\times r}} \sum_{(i,j)\in\Omega} ((\boldsymbol{Y}\boldsymbol{Z}^\top)_{ij}-M_{ij})^2$	Alternating minimization (AM) [16], [17], gradient descent (GD) [18], [19], projected gradient descent (PGD) [20], [21], stochastic gradient descent (SGD) [18]
Rank-constrained least squares	$\min_{\boldsymbol{X}\in\mathbb{R}^{n_1\times n_2}} \sum_{(i,j)\in\Omega} (X_{ij}-M_{ij})^2 \text{ s.t. } \operatorname{rank}(\boldsymbol{X})\le r$	Singular value projection (SVP) [22], [23], normalized IHT (NIHT) [24], conjugate gradient IHT (CGIHT) [25], iterative hard-thresholded SVD (IHTSVD) [26], accelerated IHT [27], [28]

optimization performance. Moreover, their fixed choice of the sample complexity raises the question of how would the convergence rate vary for arbitrary values of the number of observations s. Intuitively, as the rank of the matrix remains constant and the number of observed entries increases, one can expect the algorithm converges at a faster linear rate. Such dependence of the rate on the specifics of the problem is missing from existing results in the literature. To address this issue, our goal in this paper is to develop an exact convergence rate analysis that enables a more precise estimate of the optimization performance. We restrict our attention to the iterative hard-thresholded singular value decomposition (IHTSVD) algorithm for matrix completion. By exploiting the local structure of the problem, we establish a tight closed-form bound on the convergence rate and the number of iterations required to achieve a given accuracy.

The contribution of this work is three-fold. First, we propose a novel analysis of the local convergence of IHTSVD for matrix completion. The proposed analysis establishes the region of convergence that is proportional to the least non-zero singular value of the matrix $\boldsymbol{M}$. It also provides a closed-form bound on the linear convergence rate in terms of $\boldsymbol{M}$ and Ω. Recognizing IHTSVD as a special case of IHT with a unit step size, our analysis can also be extended to the general case of IHT with a step size other than 1 (see Section III-C). To the best of our knowledge, our proposed analysis is exact and is the tightest possible bound on the convergence rate of IHT. Second, using random matrix theory, we obtain a simplified asymptotic expression for the exact linear rate of IHTSVD in the large-scale regime. As the size of $\boldsymbol{M}$ grows to infinity, we show that the linear rate of IHTSVD converges to a deterministic constant that can be expressed in closed-form in terms of the relative rank and the sampling rate. Finally, we present numerical results to verify our proposed exact rate of convergence as well as the asymptotic rate of IHTSVD in large-scale settings.[1]

The rest of this paper is organized as follows. Section II presents background and related work on the matrix completion problem and the IHTSVD algorithm. Next, our main results on the convergence rate analysis of IHTSVD and the asymptotic behavior of the rate in large-scale matrix completion are given in Sections III and IV, respectively. Then, Section V verifies the correctness of our analysis through numerical simulations. Finally, we summarize our results and discuss some of the possible extensions in Section VI.

II. Preliminaries

A. Notation

Throughout the paper, we use the notations $\|\cdot\|_F$, $\|\cdot\|_2$, and $\|\cdot\|_{2,\infty}$ to denote the Frobenius norm, the spectral norm and the l_2/l_∞ norm (i.e., the largest l_2 norm of the rows) of a matrix, respectively. Occasionally, $\|\cdot\|_2$ is used on a vector to denote the Euclidean norm. The notation $[n]$ refers to the set $\{1,2,\ldots,n\}$. Boldfaced symbols are reserved for vectors and matrices. In addition, let $\boldsymbol{I}_n$ denote the $n\times n$ identity matrix. We also use $\otimes$ to denote the Kronecker product between two matrices.

For a matrix $\boldsymbol{X}\in\mathbb{R}^{n_1\times n_2}$, X_{ij} refers to the (i,j) element of $\boldsymbol{X}$. We denote $\sigma_{\max}(\boldsymbol{X})$ and $\sigma_{\min}(\boldsymbol{X})$ as the largest and smallest singular values of $\boldsymbol{X}$, respectively, and denote $\kappa(\boldsymbol{X})=\sigma_{\max}(\boldsymbol{X})/\sigma_{\min}(\boldsymbol{X})$ as the condition number of $\boldsymbol{X}$. Similarly, $\lambda_{\max}(\boldsymbol{X})$ and $\lambda_{\min}(\boldsymbol{X})$ are used to denote the maximum and minimum eigenvalues of $\boldsymbol{X}$, respectively. The notation $\operatorname{vec}(\boldsymbol{X})$ denotes the vectorization of $\boldsymbol{X}$ by stacking its columns on top of one another. Let $\boldsymbol{F}(\boldsymbol{X})$ be a matrix-valued function of $\boldsymbol{X}$. Then, for some $k>0$, we use $\boldsymbol{F}(\boldsymbol{X})=\boldsymbol{\mathcal{O}}(\|\boldsymbol{X}\|_F^k)$ to imply

$$\lim_{\delta\to 0}\sup_{\|\boldsymbol{X}\|_F=\delta}\frac{\|\boldsymbol{F}(\boldsymbol{X})\|_F}{\|\boldsymbol{X}\|_F^k}<\infty.$$

B. Background

Let us use $\boldsymbol{M}$ to denote the underlying $n_1\times n_2$ real matrix with rank

$$1\le r\le m=\min\{n_1,n_2\}. \tag{1}$$

The sampling set Ω is a subset of the Cartesian product $[n_1]\times[n_2]$, with cardinality $s=|\Omega|$ where $1\le s<n_1n_2$. Furthermore, the orthogonal projection associated with Ω is given in the following:

[1] Part of this work is leveraged on our conference paper [26]. In this paper, we provide complete proof of all the results. Moreover, we present a novel analysis of the asymptotic convergence rate in the large-scale matrix completion setting and a comprehensive set of experiments to verify our theoretical results. A similar result on the local linear convergence of IHTSVD can be found in the unpublished work of Lai and Varghese [30]. However, we emphasize that our initial result is published before the time [30] appeared in arXiv. More importantly, our bound is tighter than that in [30] (see our discussion in the Supplementary Material - Section A).

Definition 1. *The orthogonal projection onto the set of matrices supported in Ω is defined as a linear operator $\mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ satisfying*

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

where $\bar{\Omega}$ denotes the complement set of Ω.

If we consider vector spaces instead of matrix spaces, the orthogonal projection $\mathcal{P}_\Omega$ can also be viewed as a selection matrix corresponding to Ω:

Definition 2. *The selection matrix $\boldsymbol{S}_\Omega \in \mathbb{R}^{n_1 n_2 \times s}$ comprises a subset of s columns of the identity matrix of dimension $n_1 n_2$ such that*

$$\begin{cases} \boldsymbol{S}_\Omega^\top \boldsymbol{S}_\Omega = \boldsymbol{I}_s, \\ \operatorname{vec}\big(\mathcal{P}_\Omega(\boldsymbol{X})\big) = \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \operatorname{vec}(\boldsymbol{X}). \end{cases}$$

Corresponding to the complement set $\bar{\Omega}$, we also define similar notations for $\mathcal{P}_{\bar{\Omega}} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ and $\boldsymbol{S}_{\bar{\Omega}} \in \mathbb{R}^{n_1 n_2 \times (n_1 n_2 - s)}$.

Next, using the notation of $\mathcal{P}_\Omega$, we can formulate the matrix completion problem as follows:

$$\min_{\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}} \frac{1}{2} \left\| \mathcal{P}_\Omega(\boldsymbol{X} - \boldsymbol{M}) \right\|_F^2 \text{ s.t. } \operatorname{rank}(\boldsymbol{X}) \leq r. \tag{2}$$

One natural approach to the optimization problem (2) is projected gradient descent (PGD). Starting at some $\boldsymbol{X}^{(0)}$, we iteratively update the current matrix by *(i)* taking a step in the opposite direction of the gradient; and *(ii)* projecting the result back onto the set of matrices with rank less than or equal to r. It follows that

$$\boldsymbol{X}^{(k+1)} = \mathcal{P}_r\big(\boldsymbol{X}^{(k)} - \eta \mathcal{P}_\Omega(\boldsymbol{X}^{(k)} - \boldsymbol{M})\big), \tag{3}$$

where η is the step size and $\mathcal{P}_r$ is the rank-r projection (formally defined later in Definition 3). Due to the singular value truncating nature of the projection $\mathcal{P}_r$, PGD is often referred as the iterative hard thresholding (IHT) method for matrix completion [31]. In [22], IHT with step size $\eta = n_1 n_2 / s$ is also named as the Singular Value Projection (SVP) algorithm for matrix completion. It is interesting to note that under certain assumptions, [32] shows that the algorithm enjoys a fast global linear convergence with this choice of step size. Alternatively, setting the step size $\eta = 1$ yields the following update

$$\begin{aligned} \boldsymbol{X}^{(k+1)} &= \mathcal{P}_r\big(\boldsymbol{X}^{(k)} - \mathcal{P}_\Omega(\boldsymbol{X}^{(k)} - \boldsymbol{M})\big) \\ &= \mathcal{P}_r\big(\mathcal{P}_{\bar{\Omega}}(\boldsymbol{X}^{(k)}) + \mathcal{P}_\Omega(\boldsymbol{M})\big). \end{aligned} \tag{4}$$

Note that $\mathcal{P}_{\bar{\Omega}}(\boldsymbol{X}^{(k)}) + \mathcal{P}_\Omega(\boldsymbol{M})$ is a linear orthogonal projection of $\boldsymbol{X}^{(k)}$ onto the set of matrices with the same support as $\boldsymbol{M}$ in Ω. This motivates the IHTSVD algorithm [26] that alternates between two projection steps: the projection onto the set of low-rank matrices and the projection onto the set of matrices with the same support as $\boldsymbol{M}$ in Ω (see Algorithm 1). This paper, developed based on [26], focuses on local convergence properties of IHTSVD. Compared to the existing global convergence analysis for matrix completion, our setting does not require certain assumptions such as the incoherence of $\boldsymbol{M}$, the uniform randomness of Ω, and the low sample complexity, e.g., $s = \mathcal{O}(r^5 n \log n)$ in [32]. We also note that the proposed analysis can be extended to other variants of IHT with different step sizes see Section III-C.

Algorithm 1: IHTSVD

```
Input: $\mathcal{P}_\Omega(\boldsymbol{M})$, $r$, $K$, $\boldsymbol{X}^{(0)}$
Output: $\boldsymbol{X}^{(K)}$
$\boldsymbol{X}^{(0)} = \mathcal{P}_\Omega(\boldsymbol{M})$
for $k = 0, 1, \ldots, K-1$ do
    $\boldsymbol{X}^{(k+1/2)} = \mathcal{P}_r(\boldsymbol{X}^{(k)})$
    $\boldsymbol{X}^{(k+1)} = \mathcal{P}_{\bar{\Omega}}\big(\boldsymbol{X}^{(k+1/2)}\big) + \mathcal{P}_\Omega(\boldsymbol{M})$
```

Finally, we present a formal definition of the rank-r projection. Consider a matrix $\boldsymbol{X} \in \mathbb{R}^{n_1 \times n_2}$ with the singular value decomposition

$$\boldsymbol{X} = \sum_{i=1}^{m} \sigma_i(\boldsymbol{X}) \boldsymbol{u}_i(\boldsymbol{X}) \boldsymbol{v}_i^\top(\boldsymbol{X}),$$

where $\sigma_1(\boldsymbol{X}) \geq \ldots \geq \sigma_m(\boldsymbol{X}) \geq 0$ are the singular values of $\boldsymbol{X}$ and $\{\boldsymbol{u}_1(\boldsymbol{X}), \ldots, \boldsymbol{u}_m(\boldsymbol{X})\}$, $\{\boldsymbol{v}_1(\boldsymbol{X}), \ldots, \boldsymbol{v}_m(\boldsymbol{X})\}$ are the sets of left and right singular vectors of $\boldsymbol{X}$, respectively.

Definition 3. *The rank-r projection of $\boldsymbol{X}$ is defined as*

$$\mathcal{P}_r(\boldsymbol{X}) = \sum_{i=1}^{r} \sigma_i(\boldsymbol{X}) \boldsymbol{u}_i(\boldsymbol{X}) \boldsymbol{v}_i^\top(\boldsymbol{X}).$$

The rank-r projection of $\boldsymbol{X}$ is unique if and only if $\sigma_r(\boldsymbol{X}) > \sigma_{r+1}(\boldsymbol{X})$ or $\sigma_r(\boldsymbol{X}) = 0$ [33]. Since $\mathcal{P}_r(\boldsymbol{X})$ zeroes out all the small singular value of $\boldsymbol{X}$, it is often referred as the singular value hard-thresholding operator. Since $\boldsymbol{M}$ is a rank-r matrix, we have

$$\boldsymbol{M} = \mathcal{P}_r(\boldsymbol{M}) = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^\top,$$

where $\boldsymbol{\Sigma}_r = \operatorname{diag}(\sigma_1, \ldots, \sigma_r)$ contains the singular values of $\boldsymbol{M}$ and $\boldsymbol{U}_r = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_r] \in \mathbb{R}^{n_1 \times r}$, $\boldsymbol{V}_r = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_r] \in \mathbb{R}^{n_2 \times r}$ are comprised of the first r left and right singular vectors of $\boldsymbol{M}$, respectively.[2] Denote $\boldsymbol{U}_\perp = [\boldsymbol{u}_{r+1}, \ldots, \boldsymbol{u}_{n_1}] \in \mathbb{R}^{n_1 \times (n_1 - r)}$ and $\boldsymbol{V}_\perp = [\boldsymbol{v}_{r+1}, \ldots, \boldsymbol{v}_{n_2}] \in \mathbb{R}^{n_2 \times (n_2 - r)}$. The projections onto the left and right null spaces of $\boldsymbol{M}$ are uniquely defined as $\boldsymbol{P}_{\boldsymbol{U}_\perp} = \boldsymbol{U}_\perp \boldsymbol{U}_\perp^\top = \boldsymbol{I}_{n_1} - \sum_{i=1}^r \boldsymbol{u}_i \boldsymbol{u}_i^\top$ and $\boldsymbol{P}_{\boldsymbol{V}_\perp} = \boldsymbol{V}_\perp \boldsymbol{V}_\perp^\top = \boldsymbol{I}_{n_2} - \sum_{i=1}^r \boldsymbol{v}_i \boldsymbol{v}_i^\top$, respectively.

C. Related Work

This section discusses existing results in convergence analysis of IHT for matrix completion. As mentioned in Section I, these results focus on guarantees for convergence to a global solution of the problem, with a conservative bound on the linear rate. To better understand their assumptions, we begin by introducing a few key concepts and proceed to explain how such key concepts are used in assuring global linear convergence. First, the incoherence condition for matrix completion, introduced by Candès and Recht [10], is stated as:

[2] In the rest of this paper, we omit the parameter in the notation of the singular values and the singular vectors of $\boldsymbol{M}$ for simplicity.

Assumption 1 (Incoherence)**.** *The matrix $\boldsymbol{M} = \boldsymbol{U}_r \boldsymbol{\Sigma}_r \boldsymbol{V}_r^\top$ is μ-incoherent, i.e.,*

$$\|\boldsymbol{U}_r\|_{2,\infty} \leq \sqrt{\frac{\mu r}{n_1}} \text{ and } \|\boldsymbol{V}_r\|_{2,\infty} \leq \sqrt{\frac{\mu r}{n_2}}.$$

Intuitively, an incoherent matrix has well-spread singular vectors and is less likely in the null space of the sampling operator. A common setting that generates incoherent matrices is the random orthogonal model:

Definition 4 (Random orthogonal model)**.** *The Haar measure provides a uniform and translation-invariant distribution over the group of orthogonal matrices $\mathbb{O}(n)$. $\boldsymbol{M}$ is said to follow a random orthogonal model if $\boldsymbol{U}_r$ and $\boldsymbol{V}_r$ are sub-matrices of Haar-distributed matrices in $\mathbb{O}(n_1)$ and $\mathbb{O}(n_2)$, respectively.*

Second, to avoid adversarial patterns in the sampling set, it is common to assume that each entry in Ω is selected randomly:

Assumption 2 (Uniform sampling)**.** *The sampling set Ω is obtained by selecting s elements uniformly at random from the Cartesian product $[n_1] \times [n_2]$.*

We note that a similar but not equivalent assumption on the sampling set is the Bernoulli model in which each entry of $\boldsymbol{M}$ is observed independently with probability $s/n_1 n_2$ [18]. Under these two standard assumptions, Candès and Recht [10] showed that symmetric matrix completion of size n can be solved exactly provided that the number of observations is sufficiently large, i.e., $s = \mathcal{O}(n^{1.2} r \log n)$. Later on, global convergence guarantees for various matrix-completion algorithms have been actively developed, under similar assumptions on the sample complexity. Examples of these works include [16], [18], [19], [23]. It is noted that linear convergence has been shown in the aforementioned papers via a universal upper bound on the rate of convergence, often in the form of an exponential decay bound on the error through iterations. However, since such technique is generally developed for proving global convergence, it does not offer a tight bound on the convergence rate. Moreover, the matrix completion setting considered in these papers is restricted to a particular large-scale regime (n is very large) with the specific sample complexity $s \approx \mathcal{O}(nr \log n)$.

In this paper, we address two questions that arise from existing convergence analyses of IHT for matrix completion: *(i)* can we estimate the linear convergence rate of IHT more accurately? and *(ii)* are there other settings in which linear convergence occurs? By exploiting the local structure of the problem, we identify a deterministic condition on $\boldsymbol{M}$ and Ω such that the linear convergence of IHTSVD can be guaranteed. Compared to the aforementioned analyses on the linear convergence of IHT, our result guarantees local convergence rather than global convergence. However, our estimate of the linear rate is exact and is tighter than the existing global bounds in the literature. In addition, we do not make assumptions on the incoherence of $\boldsymbol{M}$ and the randomness of the sampling set Ω, as well as not require a specific choice of sample complexity. As a result, our analysis covers a larger set of matrix completion setting.[3] Our technique utilizes the recently developed error bound for the first-order Taylor expansion of the rank-r projection, proposed by Vu *et. al.* in [34]. The result is rephrased below.

Proposition 1 (Rephrased from [34])**.** *For any $\boldsymbol{\Delta} \in \mathbb{R}^{n_1 \times n_2}$, we have*

$$\mathcal{P}_r(\boldsymbol{M} + \boldsymbol{\Delta}) = \boldsymbol{M} + \boldsymbol{\Delta} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{\Delta} \boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{\Delta}), \tag{5}$$

where the residual $\boldsymbol{R} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ satisfies:

$$\|\boldsymbol{R}(\boldsymbol{\Delta})\|_F \leq \frac{c_1}{\sigma_r} \|\boldsymbol{\Delta}\|_F^2,$$

for some universal constant $1 + 1/\sqrt{2} \leq c_1 \leq 4(1 + \sqrt{2})$.

III. Local Convergence of IHTSVD

This section presents our analysis of the local convergence of IHTSVD. First, we leverage the results in perturbation analysis to identify the Taylor series expansion of the rank-r projection. Next, the approximation allows us to derive the nonlinear difference equation that describes the change in the distance to the local optimum through IHT iterations. Closed-form expressions of the asymptotic convergence rate and the region of convergence are also given as a result of our analysis.

A. Main Result

Our local convergence result is stated as follows:

Theorem 1. *Let $\{\boldsymbol{X}^{(k)}\}_{k=0}^\infty$ be the sequence of matrices generated by Algorithm 1, i.e.,*

$$\boldsymbol{X}^{(k+1)} = \mathcal{P}_{\bar{\Omega}}\big(\mathcal{P}_r(\boldsymbol{X}^{(k)})\big) + \mathcal{P}_\Omega(\boldsymbol{M}), \tag{6}$$

for all integer k. Assume that $\lambda_{\min}(\boldsymbol{H}) > 0$ and $\boldsymbol{X}^{(0)}$ satisfies

$$\left\|\boldsymbol{X}^{(0)} - \boldsymbol{M}\right\|_F < \frac{\lambda_{\min}(\boldsymbol{H})}{c_1} \sigma_r, \tag{7}$$

where $\boldsymbol{H}$ is an $(n_1 n_2 - s) \times (n_1 n_2 - s)$ matrix given by

$$\boldsymbol{H} = \boldsymbol{S}_{\bar{\Omega}}^\top (\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}) \boldsymbol{S}_{\bar{\Omega}}, \tag{8}$$

and c_1 is given in Proposition 1. Then, $\|\boldsymbol{X}^{(k)} - \boldsymbol{M}\|_F$ converges asymptotically at a the local linear rate

$$\rho = 1 - \lambda_{\min}(\boldsymbol{H}). \tag{9}$$

Specifically, for any $\epsilon > 0$, $\|\boldsymbol{X}^{(k)} - \boldsymbol{M}\|_F \leq \epsilon \|\boldsymbol{X}^{(0)} - \boldsymbol{M}\|_F$ for all integer k such that

$$k \geq K(\epsilon) = \frac{\log(1/\epsilon)}{\log(1/(1 - \lambda_{\min}(\boldsymbol{H})))} + C, \tag{10}$$

where $\tau = \frac{c_1 \|\boldsymbol{X}^{(0)} - \boldsymbol{M}\|_F}{\sigma_r \lambda_{\min}(\boldsymbol{H})}$ and

$$C = \frac{1}{\rho \log(1/\rho)} \left(E_1\Big(\log \frac{1}{\rho + \tau(1-\rho)}\Big) - E_1\Big(\log \frac{1}{\rho}\Big) + \frac{1}{2} \log\Big(\frac{\log(1/\rho)}{\log\big(1/(\rho + \tau(1-\rho))\big)}\Big) \right) + 1, \tag{11}$$

[3] For an intuitive comparison of the matrix completion setting in our paper versus the common setting in which $s \approx \mathcal{O}(nr \log n)$, we refer the readers to the Supplementary Material - Fig. 1.

with $E_1(t) = \int_t^\infty \frac{e^{-z}}{z} dz$ being the exponential integral [35].

Theorem 1 provides a closed-form expression of the local linear convergence rate of IHTSVD for matrix completion (see Eqn. (9)). As can be seen in (10), the speed of convergence depends strongly on how close the smallest eigenvalue of $\boldsymbol{H}$ is to zero: as $\lambda_{\min}(\boldsymbol{H})$ approaches 0, the number of iterations needed to reach a relative accuracy of ϵ, i.e., $K(\epsilon)$, grows to infinity. From (8), one can verify that all eigenvalues of $\boldsymbol{H}$ lie between 0 and 1 since the norm of either a projection matrix or a selection matrix is less than or equal to 1. This combined with the aforementioned condition that $\lambda_{\min}(\boldsymbol{H}) > 0$ ensures the linear convergence rate ρ in (9) belongs to $[0, 1)$.

Remark 1. *Theorem 1 does not guarantee linear convergence when $\lambda_{\min}(\boldsymbol{H}) = 0$. Interestingly, one such situation is when $\boldsymbol{H}$ is **rank-deficient**. Let us represent*

$$\begin{aligned}\boldsymbol{H} &= \boldsymbol{S}_{\bar{\Omega}}^\top (\boldsymbol{V}_\perp \otimes \boldsymbol{U}_\perp)(\boldsymbol{V}_\perp \otimes \boldsymbol{U}_\perp)^\top \boldsymbol{S}_{\bar{\Omega}} \\ &= \boldsymbol{W}\boldsymbol{W}^\top,\end{aligned}$$

where $\boldsymbol{W} = \boldsymbol{S}_{\bar{\Omega}}^\top (\boldsymbol{V}_\perp \otimes \boldsymbol{U}_\perp) \in \mathbb{R}^{(n_1 n_2 - s)\times(n_1 - r)(n_2 - r)}$. If $\boldsymbol{W}$ is a tall matrix, i.e.,

$$s < (n_1 + n_2 - r)r, \tag{12}$$

then it follows that $\boldsymbol{H}$ is rank-deficient and $\lambda_{\min}(\boldsymbol{H}) = 0$. We note that in this case, the number of sampled entries is less than the degrees of freedom of the problem.

Remark 2. *When $s \geq (n_1 + n_2 - r)r$, it is possible that $\lambda_{\min}(\boldsymbol{H}) = 0$ for certain (adversarial) sampling patterns. For example, consider a 3×2 rank-1 matrix*

$$\boldsymbol{M} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \end{bmatrix}^\top.$$

One choice of the matrices $\boldsymbol{U}_\perp$ and $\boldsymbol{V}_\perp$ is

$$\boldsymbol{U}_\perp = \begin{bmatrix} 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \boldsymbol{V}_\perp = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

If we observe $s = 4$ entries of the first two rows of $\boldsymbol{M}$, namely, $(1,1)$, $(1,2)$, $(2,1)$, and $(2,2)$, the selection matrix corresponding to the unobserved entries $(3,1)$ and $(3,2)$ is given by

$$\boldsymbol{S}_{\bar{\Omega}}^\top = \begin{bmatrix} 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

Then, we have

$$\boldsymbol{H} = \boldsymbol{S}_{\bar{\Omega}}^\top (\boldsymbol{V}_\perp \otimes \boldsymbol{U}_\perp)(\boldsymbol{V}_\perp \otimes \boldsymbol{U}_\perp)^\top \boldsymbol{S}_{\bar{\Omega}} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

and $\lambda_{\min}(\boldsymbol{H}) = 0$. While Theorem 1 does not guarantee linear convergence of IHTSVD, one may realize that it is impossible to recover the last row of $\boldsymbol{M}$ in this case.

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 [10]. 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 [16] for alternating minimization and [23] for IHT), with an upper bound on the rate of convergence $\rho < 0.5$. Our analysis, on the other hand, does not use the aforementioned assumptions but introduces a quantity that is fundamental to the problem in terms of optimization. By exploiting the local structure of the problem, we characterize the exact linear rate of local convergence of IHT. Similar to standard assumptions in prior works, the closed-form expression we obtained can be used to determine sufficient conditions that ensure linear convergence. However, since our expression is exact, one can identify conditions that are potentially less stringent than existing conditions. For a more comprehensive comparison between our work and prior rate analysis results, we refer interested readers to the Supplementary Material - Section I (Comparison to Prior Results).

B. Proof of Theorem 1

This section provides the proof of Theorem 1. We start by formulating the recursion on the error matrix from the update (6) and the linearization of the rank-r projection:

Lemma 1. *Let us define the error matrix and its economy-vectorized version, respectively, as*

$$\boldsymbol{E}^{(k)} = \boldsymbol{X}^{(k)} - \boldsymbol{M} \qquad \text{and} \qquad \boldsymbol{e}^{(k)} = \boldsymbol{S}_{\bar{\Omega}}^\top \operatorname{vec}(\boldsymbol{E}^{(k)}).$$

Then, we have

$$\boldsymbol{E}^{(k+1)} = \mathcal{P}_{\bar{\Omega}}\big(\boldsymbol{E}^{(k)} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{E}^{(k)} \boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{E}^{(k)})\big) \tag{13}$$

and

$$\boldsymbol{e}^{(k+1)} = \big(\boldsymbol{I} - \boldsymbol{S}_{\bar{\Omega}}^\top (\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}) \boldsymbol{S}_{\bar{\Omega}}\big)\boldsymbol{e}^{(k)} + \boldsymbol{r}\big(\boldsymbol{e}^{(k)}\big), \tag{14}$$

where $\boldsymbol{R}(\cdot)$ is the residual defined in Proposition 1 and

$$\boldsymbol{r}(\boldsymbol{e}) = \boldsymbol{S}_{\bar{\Omega}}^\top \operatorname{vec}\Big(\boldsymbol{R}\big(\operatorname{vec}^{-1}(\boldsymbol{S}_{\bar{\Omega}} \boldsymbol{e})\big)\Big) \quad \text{for } \boldsymbol{e} \in \mathbb{R}^{n_1 n_2 - s}.$$

Here we recall that $\operatorname{vec}^{-1}(\cdot)$ is the inverse vectorization operator such that $(\operatorname{vec}^{-1} \circ \operatorname{vec})$ is identity.

Equations (13)-(14) in Lemma 1 offer a recursion on the error that expresses the $k+1$-th error in terms of a linear transformation of the kth error and a residual term whose magnitude can be bounded and is asymptotically negligible. Note that $\boldsymbol{E}^{(k)}$ belongs to the set of matrices supported in Ω and hence, $\|\boldsymbol{E}^{(k)}\|_F = \|\boldsymbol{e}^{(k)}\|_2$. From (14), using the triangle inequality, the definition of the operator norm, and the fact that the error lies in the set of matrices supported in Ω, i.e., $\mathcal{P}_\Omega(\boldsymbol{E}^{(k)}) = \boldsymbol{E}^{(k)}$, one can obtain the following bound on the norm of the error matrix:

Lemma 2. *The Frobenius norm of the error matrix satisfies*

$$\left\|\boldsymbol{E}^{(k+1)}\right\|_F \leq \big(1 - \lambda_{\min}(\boldsymbol{H})\big) \left\|\boldsymbol{E}^{(k)}\right\|_F + \frac{c_1}{\sigma_r} \left\|\boldsymbol{E}^{(k)}\right\|_F^2. \tag{15}$$

The nonlinear difference equation (15) has been well-studied in the stability theory of difference equations [36]–[38]. In fact, our theorem follows directly on applying Theorem 1 in [38] to (15), with $a_0 = \|\boldsymbol{E}^{(0)}\|_F$, $\rho = 1 - \lambda_{\min}(\boldsymbol{H})$, and $q = c_1/\sigma_r$. The proofs of Lemmas 1 and 2 are given in Appendix A.

C. IHT with Step Sizes Different than 1

Recall from (3) that IHTSVD is a special case of IHT with a unit step size. This choice of step size helps our analysis to be simple and elegant. Thanks to the alternating-projection view in (4), the error $\boldsymbol{E}^{(k)} = \boldsymbol{X}^{(k)} - \boldsymbol{M}$ is guaranteed to be in the set of matrices supported in $\bar{\Omega}$, i.e., $\mathcal{P}_{\bar{\Omega}}(\boldsymbol{E}^{(k)}) = \boldsymbol{E}^{(k)}$. Hence, the error analysis reduces from the space $\mathbb{R}^{n_1 \times n_2}$ for $\boldsymbol{E}^{(k)}$ to the space $\mathbb{R}^{n_1 n_2 - s}$ for $\boldsymbol{e}^{(k)} = \boldsymbol{S}_{\bar{\Omega}}^\top \operatorname{vec}(\boldsymbol{E}^{(k)})$. For step sizes other than 1, this appeal no longer holds. Nonetheless, one can follow a similar track to obtain an exact rate analysis. Indeed, the linear convergence of IHT with a fixed step size different than 1 has been recently studied in [39]. In particular, Vu *et. al.* proved that for $0 < \eta < 2/\|\boldsymbol{K}\|_2$, where $\boldsymbol{K} = \boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp \in \mathbb{R}^{r(n_1+n_2-r) \times r(n_1+n_2-r)}$ and $\boldsymbol{Q}_\perp \in \mathbb{R}^{n_1 n_2 \times r(n_1+n_2-r)}$ satisfies $\boldsymbol{Q}_\perp^\top \boldsymbol{Q}_\perp = \boldsymbol{I}_{r(n_1+n_2-r)}$ and $\boldsymbol{Q}_\perp \boldsymbol{Q}_\perp^\top = \boldsymbol{I}_{n_1 n_2} - \boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}$, the local linear convergence rate of IHT with a fixed step size η is given by

$$\rho_\eta = \max\{|1 - \eta\lambda_{\max}(\boldsymbol{K})|, |1 - \eta\lambda_{\min}(\boldsymbol{K})|\}. \tag{16}$$

By comparing the two matrices $\boldsymbol{K} = \boldsymbol{Q}_\perp^\top \boldsymbol{S}_\Omega \boldsymbol{S}_\Omega^\top \boldsymbol{Q}_\perp \in \mathbb{R}^{r(n_1+n_2-r) \times r(n_1+n_2-r)}$ and $\boldsymbol{H} = \boldsymbol{S}_{\bar{\Omega}}^\top (\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}) \boldsymbol{S}_{\bar{\Omega}} \in \mathbb{R}^{(n_1 n_2 - s) \times (n_1 n_2 - s)}$, we recognize that they share the same set of eigenvalues in the interval $[0, 1)$ while may only differ by the eigenvalues at 1. Thus, substituting $\eta = 1$ into (16) yields the same expression of the rate in (9).

It is also interesting to note that the optimal step size and the optimal convergence rate are given by [39]

$$\begin{aligned} \eta_{opt} &= \frac{2}{\lambda_{\max}(\boldsymbol{K}) + \lambda_{\min}(\boldsymbol{K})}, \\ \rho_{opt} &= 1 - \frac{2}{\kappa(\boldsymbol{K}) + 1}. \end{aligned} \tag{17}$$

Regardless of the step size, the approach with $\eta = 1$ and the approach with other values for η are both based on no more than the smallest and largest eigenvalues of the fundamental matrix $\boldsymbol{H}$. Thus, investigating the spectral properties of $\boldsymbol{H}$ in the context of matrix completion plays a pivotal role in understanding the linear convergence of IHT. In the rest of the paper, we focus on the case with a unit step size for simplicity and convenience. We will study the asymptotic behavior of IHTSVD in the large-scale matrix completion setting (Section IV) and provide further discussion on IHT with different step sizes in Supplementary Material - Section II.

IV. Local Convergence of IHTSVD for Large-Scale Matrix Completion

In this section, we study the local convergence of IHTSVD for large-scale matrix completion, a setting of practical interest in the rise of big data. Using recent results in random matrix theory, we show that, as its dimensions grow to infinity, the spectral distribution of $\boldsymbol{H}$ converges almost surely to a deterministic distribution with bounded support. Consequently, we propose a large-scale asymptotic estimate of the linear convergence rate of IHTSVD that is a closed-form expression of the relative rank and the sampling rate.

A. Overview

We are interested in the asymptotic setting in which the size of $\boldsymbol{M}$ grows to infinity, i.e., $m = \min\{n_1, n_2\} \to \infty$. Let us assume that the ratio n_1/n_2 remains to be a non-zero constant as $m \to \infty$. In addition, we introduce two concepts that are the normalization of the degrees of freedom and the number of measurements:

Definition 5 (Relative rank)**.** *The rank r increases as $m \to \infty$ such that the relative rank remains to be a constant*

$$\rho_r = 1 - \sqrt{\left(1 - \frac{r}{n_1}\right)\left(1 - \frac{r}{n_2}\right)} \in (0, 1]. \tag{18}$$

Definition 6 (Sampling rate)**.** *The number of observations increases as $m \to \infty$ such that the sampling rate remains to be a constant*

$$\rho_s = \frac{s}{n_1 n_2} \in (0, 1]. \tag{19}$$

When $\rho_s < 1 - (1 - \rho_r)^2$, we recover the case in Remark 1 where the number of measurements is less than the degrees of freedom. As far as the local linear rate of IHTSVD is concerned, we only consider the case $\rho_s \geq 1 - (1 - \rho_r)^2$.

Remark 3. *When $r = m$, we have $\rho_r = 1$. Moreover, when $n_1 = n_2 = m$, the relative rank is exactly the ratio r/m. As can be seen below, the proposed definition of the relative rank incorporates both dimensions of $\boldsymbol{M}$ to enable the compact representation of ρ in terms of ρ_r and ρ_s.*

We are in position to state our result on the asymptotic behavior of the linear rate ρ in large-scale matrix completion:

Theorem 2 (Informal)**.** *For $\rho_s > 1 - (1 - \rho_r)^2$, the linear convergence rate ρ of IHTSVD approaches*

$$\rho_\infty = 1 - \left(\sqrt{(1 - \rho_r)^2 \rho_s} - \sqrt{\rho_r (2 - \rho_r)(1 - \rho_s)}\right)^2, \tag{20}$$

as $m \to \infty$.

The formal statement of our result is given later in Theorem 3. Note that ρ_∞ is independent of the structure of the solution matrix $\boldsymbol{M}$ and the sampling set Ω. Moreover, it depends only on the relative rank and the sampling rate. Figure 1 depicts the contour plot of ρ_∞ as a function of ρ_r and ρ_s. It can be seen that for a fixed value of ρ_r, the asymptotic rate decreases towards 0 as the number of observed entries increases. This matches the intuition that more information leads to faster convergence. Conversely, for a fixed value of ρ_s, the algorithm converges slower as the rank of the matrix increases, due to the increasing uncertainty (i.e., more degrees of freedom) in the set $\bar{\Omega}$. On the boundary where $\rho_s = 1 - (1 - \rho_r)^2$, there is no linear convergence predicted by our theory since $\rho_\infty = 1$. In this case, we recall that the number of observed entries equals the degrees of freedom of the problem.

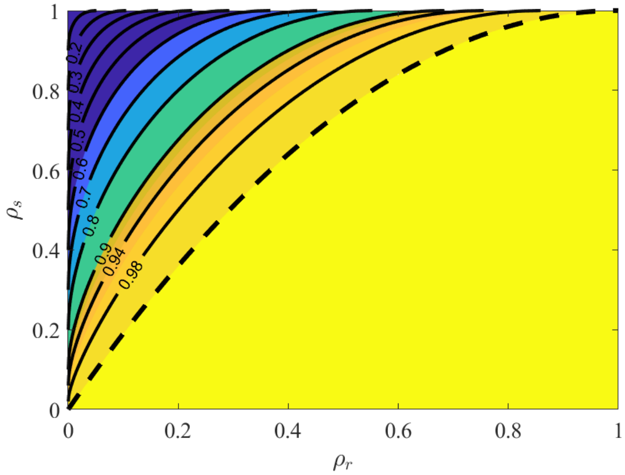


Fig. 1: Contour plot of ρ_∞ as a 2-D function of ρ_r and ρ_s given by (20). The isoline at which $\rho_\infty = 1$ is represented by the dashed line. The yellow region below this isoline corresponds to the under-determined setting $\rho_s < 1 - (1 - \rho_r)^2$.

Our technique relies on recent results in random matrix theory to exploit the special structure of $\boldsymbol{H}$. First, when n_1/n_2 remains constant, it holds that $n = n_1 n_2 \to \infty$ as $m \to \infty$. Then, $\boldsymbol{H}$ can be viewed as an element of a sequence of matrices of the form

$$\boldsymbol{H}_n = \boldsymbol{W}_{pq}^n (\boldsymbol{W}_{pq}^n)^\top, \tag{21}$$

where $\boldsymbol{W}_{pq}^n \in \mathbb{R}^{pn_1n_2 \times qn_1n_2}$ is a truncation of the orthogonal matrix $\boldsymbol{W}^n = \boldsymbol{V}^{n_2} \otimes \boldsymbol{U}^{n_1}$, for $\boldsymbol{U}^{n_1}$ and $\boldsymbol{V}_\perp^{n_2}$ orthogonal matrices of dimensions $n_1 \times n_1$ and $n_2 \times n_2$, respectively, and

$$p = \frac{n_1 n_2 - s}{n_1 n_2} = 1 - \rho_s,$$
$$q = \frac{(n_1 - r)(n_2 - r)}{n_1 n_2} = (1 - \rho_r)^2.$$

As n grows to infinity, we are interested in finding the limit (or even the limiting distribution) of the smallest eigenvalue of $\boldsymbol{H}_n$, which is a random truncation of the Kronecker product of two large dimensional semi-orthogonal matrices.

B. Truncations of Large Dimensional Orthogonal Matrices

Random matrix theory studies the asymptotic behavior of eigenvalues of matrices with entries drawn randomly from various matrix ensembles such as Gaussian orthogonal ensemble (GOE), Wishart ensemble, MANOVA ensemble [41]. The closest random matrix ensemble to our matrix ensemble $\{\boldsymbol{H}_n\}_{n\in\mathbb{N}^+}$ is the MANOVA ensemble in which truncations of large dimensional Haar orthogonal matrices are considered. Here we recall that the Haar measure provides a uniform distribution over the set of all $n \times n$ orthogonal matrices $\mathbb{O}(n)$. Indeed, it is a unique translation-invariant probability measure on $\mathbb{O}(n)$. If we assume that the matrix $\boldsymbol{M}$ follows a random orthogonal model [10], then $\boldsymbol{U}_\perp$ and $\boldsymbol{V}_\perp$ are essentially sub-matrices of Haar orthogonal matrices in $\mathbb{O}(n_1)$ and $\mathbb{O}(n_2)$, respectively, and $\{\boldsymbol{H}_n\}_{n\in\mathbb{N}^+}$ is a sequence of truncations of the Kronecker product of two Haar orthogonal matrices.

There have been certain theoretical works on truncations of Haar invariant matrices in the literature. In 1980, Wachter [42] established the limiting distribution of the eigenvalues in the MANOVA ensemble. Later on, the density function of the eigenvalues of such matrix has been shown to be the same as that of a Jacobi matrix [43]–[45]. Shortly afterward, Johnstone proved the Tracy-Widom behavior of the largest eigenvalue in [46]. More recently, Farrell and Nadakuditi relaxed the constraint on the uniform (Haar) distribution of the orthogonal matrix considered the Kronecker products of Haar-distributed orthogonal matrices, which is similar to our matrix completion setting in this paper. The authors showed that the limiting density of their truncations remains the same as the original case without Kronecker products. Further results on the eigenvalue distribution of truncations of Haar orthogonal matrices were also given in [47]–[49]. To the best of our knowledge, no result has been shown for the limiting behavior of the smallest eigenvalue of random MANOVA matrices.

In our context, we leverage the recent result in [40], which assumes the randomness on the truncation rather than the orthogonal matrix. This variant, while differs from the classic MANOVA ensemble in random matrix theory, is well-suited to the setting of matrix completion. Let us begin with the following definition of the empirical spectral distribution:

Definition 7. *Let $\boldsymbol{H}_n$ be an $n \times n$ real symmetric matrix with eigenvalues $\lambda_1, \dots, \lambda_n$. The **empirical spectral distribution (ESD)** of $\boldsymbol{H}_n$, denoted by $\mu_{\boldsymbol{H}_n}$, is the probability measure that puts equal mass at each of the eigenvalues of $\boldsymbol{H}_n$:*

$$\mu_{\boldsymbol{H}_n} \triangleq \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i},$$

where δ_λ is the Dirac mass at λ.

Next, we define the concepts of a sequence of row sub-sampled matrices and the concentration property:

Definition 8. *For each $n \in \mathbb{N}^+$, consider the $n \times qn$ matrix $\boldsymbol{W}_q^n = [\boldsymbol{w}_1^n, \dots, \boldsymbol{w}_n^n]^\top$, where $\boldsymbol{w}_i^n \in \mathbb{R}^{qn}$ and q is a constant in $(0, 1)$. Let P_n be a pn-permutation of $[n]$ selected uniformly at random, for p is a constant in $(0, 1)$, and $\boldsymbol{W}_{pq}^n \in \mathbb{R}^{pn \times qn}$ be the random matrix obtained by selecting the corresponding set of pn rows from $\boldsymbol{W}_q^n$. Then, the sequence $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$ is called **a sequence of q-tall matrices**, and the sequence $\{\boldsymbol{W}_{pq}^n\}_{n\in\mathbb{N}^+}$ is called **a sequence of row sub-sampled matrices** of $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$.*

Definition 9. *Given the setting in Definition 8, for each $j \in P_n$, denote $P_n^j = P_n \setminus \{j\}$. In addition, for $z \in \mathbb{C}$, define*

$$\boldsymbol{R}_j(z) = \Big(\sum_{i \in P_n^j} \boldsymbol{w}_i^n (\boldsymbol{w}_i^n)^\top - z\boldsymbol{I}_{qn} \Big)^{-1}.$$

*Then, the sequence $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$ is **concentrated** if and only if for any $j \in P_n$ and $z \in \mathbb{C}$, we have*

$$(\boldsymbol{w}_j^n)^\top \boldsymbol{R}_j(z) \boldsymbol{w}_j^n - \mathbb{E}_{j|P_n^j}\left[(\boldsymbol{w}_j^n)^\top \boldsymbol{R}_j(z) \boldsymbol{w}_j^n\right] \xrightarrow{p} 0. \tag{22}$$

In the following, we consider examples of sequences of matrices that are concentrated, as well as an example of the sequence of incoherent matrices that are **not** concentrated.

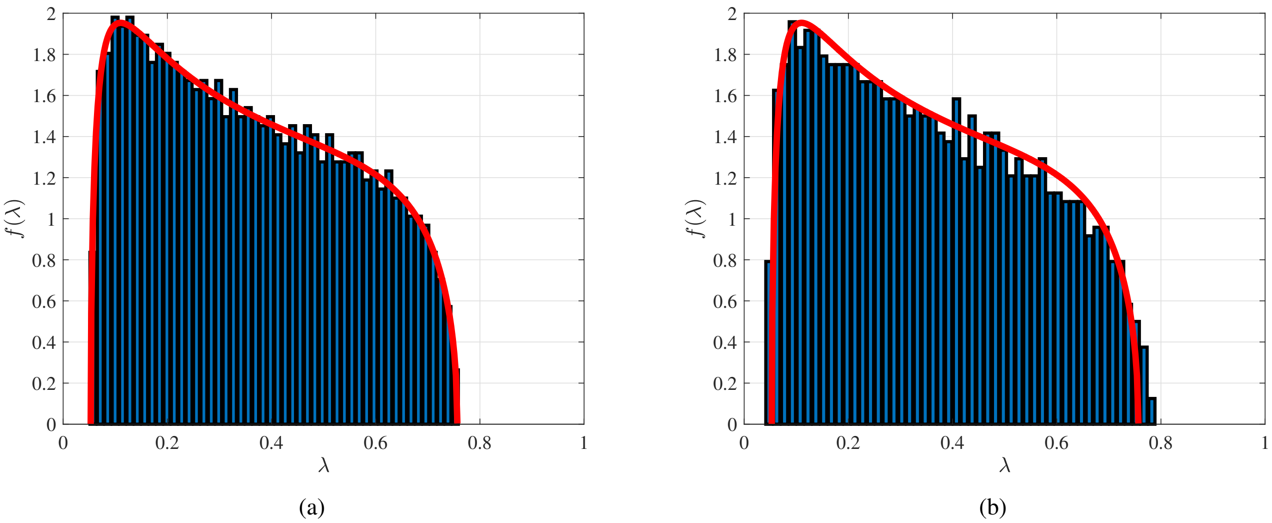


Fig. 2: (Motivated by Fig. 2 in [40]) Scaled histogram and the limiting ESD of $\boldsymbol{H}_n = \boldsymbol{W}_{pq}^n(\boldsymbol{W}_{pq}^n)^\top$, where $\boldsymbol{W}_{pq}^n$ is the $pn \times qn$ upper-left corner of an $n \times n$ orthogonal matrix $\boldsymbol{W}_n$, for $n = 10000$, $p = 0.16$, and $q = 0.36$. In (a), $\boldsymbol{W}_n$ is the orthogonal factor in the QR factorization of a 10000×10000 random matrix with $i.i.d$ standard normal entries. In (b), $\boldsymbol{W}_n = \boldsymbol{Q}_1 \otimes \boldsymbol{Q}_2$, where $\boldsymbol{Q}_1$ and $\boldsymbol{Q}_2$ are the orthogonal factors in the QR factorization of two independent 100×100 random matrices with $i.i.d$ standard normal entries. The histograms with 50 bins (blue) are scaled by a factor of $1/pnw$, where w is the bin width. The limiting ESD (red) is generated by (23). It can be seen that the histogram in (a) match the limiting ESD better than the histogram in (b).

Example 1. *Random settings:*[4]

1) *The sequence of q-tall matrices $\{\boldsymbol{A}_q^n\}_{n\in\mathbb{N}^+}$, where the entries of $\boldsymbol{A}_q^n$ are $i.i.d$ $\mathcal{N}(0, 1/n)$, is concentrated.*
2) *The sequence $\{\boldsymbol{B}_q^n \otimes \boldsymbol{C}_q^n\}_{n\in\mathbb{N}^+}$, where $\{\boldsymbol{B}_q^n\}_{n\in\mathbb{N}^+}$ and $\{\boldsymbol{C}_q^n\}_{n\in\mathbb{N}^+}$ are two sequences of q-tall matrices whose entries are $i.i.d$ $\mathcal{N}(0, 1/n)$, is also concentrated.*

Example 2. *Deterministic settings:*

1) *The sequence of q-tall matrices $\{\boldsymbol{D}_q^n\}_{n\in\mathbb{N}^+}$, where the entries of $\boldsymbol{D}_q^n$ are all 1, is concentrated.*
2) *The sequence of 1/2-tall matrices $\{\boldsymbol{E}_q^n\}_{n\in\mathbb{N}^+}$ where*

$$\boldsymbol{E}_q^n = \begin{bmatrix} 0.6\sqrt{\frac{2}{n}}\mathtt{H}_{n/2} \\ 0.8\sqrt{\frac{2}{n}}\mathtt{H}_{n/2} \end{bmatrix},$$

for $\mathtt{H}_{n/2}$ being a Hadamard matrix of order $n/2$ [50], is not concentrated. On the other hand, one can verify that $\boldsymbol{E}_q^n$ is μ-incoherent, for

$$\mu = \left\| 0.8\sqrt{2/n}\mathtt{H}_{n/2} \right\|_F^2 \frac{n}{n/2} = 1.28.$$

Thus, the concentration assumption in Definition 9 is stronger than the widely-used incoherence assumption.

With these definitions in place, we now state the result of the limiting ESD of truncation of orthogonal matrices. To fit our matrix completion setting in this paper, we rephrase the result in [40] to the case of row sub-sampled semi-orthogonal matrices (as opposed to column sub-sampled semi-orthogonal matrices in the aforementioned paper).

Proposition 2 (Rephrased from [40])**.** *Let $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$ be a sequence of q-tall matrices that is concentrated. In addition, assume that $\boldsymbol{W}_q^n$ is semi-orthogonal for all $n \in N^+$, i.e., $(\boldsymbol{W}_q^n)^\top \boldsymbol{W}_q^n = \boldsymbol{I}_{qn}$. Let $\{\boldsymbol{W}_{pq}^n\}_{n\in\mathbb{N}^+}$ be a sequence of row sub-sampled matrices of $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$. Then, as $n \to \infty$, the ESD of $\boldsymbol{H}_n = \boldsymbol{W}_{pq}^n(\boldsymbol{W}_{pq}^n)^\top$ converges almost surely to the deterministic distribution μ_{pq} such that*

$$d\mu_{pq} = \left(1 - \frac{q}{p}\right)_+ \delta(x)dx + \left(\frac{p+q-1}{p}\right)_+ \delta(x-1)dx + \frac{\sqrt{(\lambda^+ - x)(x - \lambda^-)}}{2\pi p x(1-x)} \mathbb{I}[\lambda^- \leq x \leq \lambda^+]dx, \quad (23)$$

where δ is the Dirac delta function and

$$\lambda^\pm = \left(\sqrt{q(1-p)} \pm \sqrt{p(1-q)}\right)^2.$$

The proposition asserts that the limiting ESD of $\boldsymbol{H}_n$ exists and depends only on the row ratio p and the column ratio q, provided that $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$ is concentrated. We note that the distribution μ_{pq} is exactly the same as the limiting distribution of the MANOVA ensemble. Indeed, one can show that the MANOVA ensemble is a concentrated matrix sequence:

Lemma 3. *Let $\boldsymbol{W}^n$ be a Haar-distributed orthogonal matrix in $\mathbb{O}(n)$ and $\boldsymbol{W}_q^n$ be the semi-orthogonal matrices obtained from any qn (for $q \in (0,1)$) columns of $\boldsymbol{W}^n$. Then the sequence $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$ is concentrated.*

Furthermore, the Kronecker product of two Haar-distributed orthogonal matrices also possesses the concentration property:

Lemma 4. *Let $\boldsymbol{U}^{n_1}$ and $\boldsymbol{V}^{n_2}$ be Haar-distributed orthogonal matrices in $\mathbb{O}(n_1)$ and $\mathbb{O}(n_2)$, respectively. Define $\boldsymbol{U}_{q_1}^{n_1}$ and*

[4] The detail of this example is provided in the Supplementary Material - Section III.

$\boldsymbol{V}_{q_2}^{n_2}$ as the semi-orthogonal matrices obtained from any q_1 and q_2 (for $q_1, q_2 \in (0,1)$) columns of $\boldsymbol{U}^{n_1}$ and $\boldsymbol{V}^{n_2}$, respectively. Then the sequence $\{\boldsymbol{W}_q^n = \boldsymbol{U}_{q_1}^{n_1} \otimes \boldsymbol{V}_{q_2}^{n_2}\}_{n\in\mathbb{N}^+}$ (with $q = q_1 q_2$) is concentrated.

Lemmas 3 and 4 are immediate consequences of Lemma 3.1 in [51], so we omit the proof of these lemmas here.

C. Proposed Estimation of the Linear Rate ρ

In order to apply Proposition 2 to our matrix completion setting, we recall that $\boldsymbol{W}_{pq}^n$ can be viewed as the n-th element of a sequence of row sub-sampled matrices of $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$, where $\boldsymbol{W}_q^n = \boldsymbol{V}_\perp^{n_2} \otimes \boldsymbol{U}_\perp^{n_1}$. If the sequence $\{\boldsymbol{W}_q^n\}_{n\in\mathbb{N}^+}$ is concentrated, then (23) holds for $p = 1-\rho_s$ and $q = (1-\rho_r)^2$. Therefore, one might expect that the smallest eigenvalue of $\boldsymbol{H}_n = \boldsymbol{W}_{pq}^n(\boldsymbol{W}_{pq}^n)^\top$ converges to

$$\lambda^- = \left(\sqrt{q(1-p)} - \sqrt{p(1-q)}\right)^2.$$

Thus, by Theorem 1, the convergence rate ρ converges to $1-\lambda^-$. The following theorem is an immediate application of Proposition 2 to our large-scale matrix completion setting:

Theorem 3. *As $m \to \infty$, assume that $\boldsymbol{M}$ is generated in a way that the Kronecker product $\boldsymbol{W}_q^n = \boldsymbol{V}_\perp^{n_2} \otimes \boldsymbol{U}_\perp^{n_1}$ forms a sequence of semi-orthogonal matrices that is concentrated. Then, provided $\rho_s \geq 1-(1-\rho_r)^2$, the ESD $\mu_{\boldsymbol{H}_n}$ converges almost surely to the deterministic distribution $\mu_{\rho_r\rho_s}$ such that*

$$\begin{aligned} d\mu_{\rho_r\rho_s} = &\left(\frac{(1-\rho_r)^2-\rho_s}{1-\rho_s}\right)_+ \delta(x-1)dx \\ &+ \frac{\sqrt{(\lambda^+-x)(x-\lambda^-)}}{2\pi(1-\rho_s)x(1-x)}\mathbb{I}[\lambda^- \leq x \leq \lambda^+]dx, \quad (24) \end{aligned}$$

where $\lambda^\pm = \left(\sqrt{(1-\rho_r)^2\rho_s} \pm \sqrt{\rho_r(2-\rho_r)(1-\rho_s)}\right)^2$.

Theorem 3 states the convergence of the spectral distribution of $\boldsymbol{H}$ as the dimensions grow to infinity. It is notable that the support of the distribution consists of the interval $[\lambda^-, \lambda^+]$ and a mass at 1. Based on this result, we conjecture that the smallest eigenvalue of $\boldsymbol{H}$ converges to λ^- and hence, the convergence rate ρ converges to ρ_∞:

Conjecture 1. *Assume the same setting as in Theorem 3. As $m \to \infty$, the linear rate ρ defined in (9) converges almost surely to $p_\infty = 1-\lambda^-$, given in (20).*

V. Numerical Results

In this section, we provide numerical results to verify the exact linear convergence rate of IHTSVD in (9) with the empirical rate observed in monitoring the error through iterations. Additionally, as supporting evidence for Theorem 3 and Conjecture 1, we demonstrate the increasing similarity between the empirical rate and the asymptotic rate in (20) as the dimensions of the matrix grow.

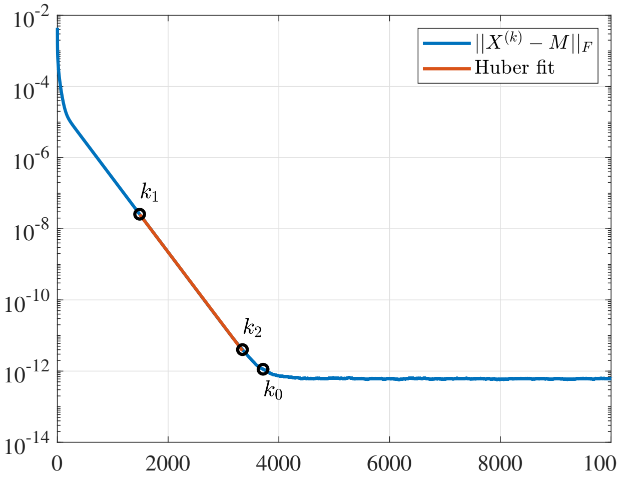


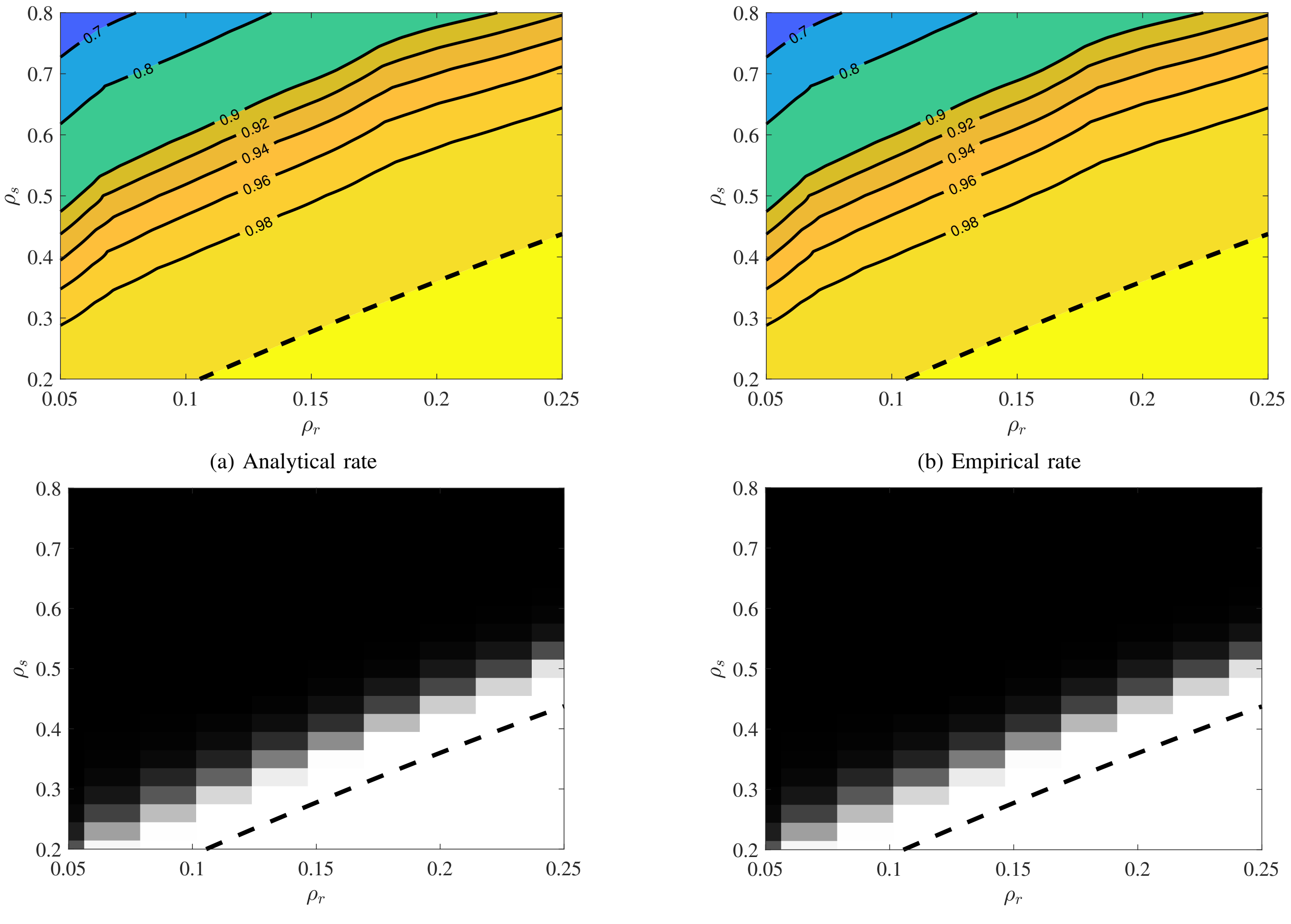
Fig. 3: Estimation of the empirical rate using the error sequence $\{\left\|\boldsymbol{X}^{(k)} - \boldsymbol{M}\right\|_F\}_{k=k_1}^{k_2}$. Due to the numerical error below 10^{-12}, we need to identify the 'turning point' at k_0 and then set $k_1 = \lfloor 0.4k_0 \rfloor$ and $k_2 = \lfloor 0.9k_0 \rfloor$.

A. Analytical Rate versus Empirical Rate

In this experiment, we verify the analytical expression of the linear convergence rate of IHTSVD by comparing it with the empirical rate obtained by measuring the decrease in the norm of the error matrix. Our goal is to demonstrate that they agree in various settings of ρ_r and ρ_s.

Data generation. We first set the dimensions $n_1 = 50$ and $n_2 = 40$. Next, for each r in $\{1, 2, \ldots, 12\}$, we generate the rank-r matrix $\boldsymbol{M}$ as follows. We construct the random orthogonal matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ by *(i)* generating a $n_1 \times n_2$ random matrix whose entries are $i.i.d$ normally distributed $\mathcal{N}(0,1)$ and *(ii)* performing the singular value decomposition of the resulting matrix. The matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are comprised of the corresponding left and right singular vectors. Then, the rank-r matrix $\boldsymbol{M}$ is generated by taking the product $\boldsymbol{U}_1\boldsymbol{\Sigma}_1\boldsymbol{V}_1^\top$, where $\boldsymbol{\Sigma}_1 = \mathrm{diag}(r, r-1, \ldots, 1)$ and $\boldsymbol{U}_1, \boldsymbol{V}_1$ are the first r columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Finally, for each s in the linearly spaced set $\{0.2n, 0.23n, 0.26n, \ldots, 0.8n\}$, we create the 1000 different sampling sets, each of them is obtained by generating a random permutation of the set $[n]$ and then selecting the first s elements of the permutation. Thus, we obtain a 12×21 grid based on the values of r and s such that *(i)* grid points corresponding to the same rank r share the same underlying matrix $\boldsymbol{M}$; and *(ii)* each point on the grid corresponds to 1000 different sampling sets.

Estimating Analytical Rate and Empirical Rate. We calculate the analytical rate for each aforementioned setting of $\boldsymbol{M}$ and Ω using (9). Due to numerical errors in computing small eigenvalues, we need to set all the resulting rates that are greater than 1 to 1, indicating there is no linear convergence in such cases. For the calculation of the empirical rate, we run Algorithm 1 in the same setting with $K = 10000$ iterations. The initial point $\boldsymbol{X}^{(0)}$ is obtained by adding $i.i.d.$ normally distributed noise with standard deviation $\sigma = 10^{-4}$ to the entries of $\boldsymbol{M}$. Here we note that σ is chosen to be small for



(a) Analytical rate

(b) Empirical rate

(c) Probability of linear convergence (analytical)

(d) Probability of linear convergence (empirical)

Fig. 4: The analytical rate and the empirical rate of convergence of IHTSVD as a function of the relative rank ρ_r and the sampling ratio ρ_s, with $n_1 = 50$ and $n_2 = 40$. (a) Contour plot of the analytical rate as a function of ρ_r and ρ_s. (b) Contour plot of the empirical rate as a function of ρ_r and ρ_s. (c) Empirical probability of linear convergence based on the analytical rate. (d) Empirical probability of linear convergence based on the empirical rate. In (c) and (d), the black color corresponds to linear convergence, whereas the white color corresponds to no linear convergence. The data is evaluated based on a 12×21 grid over ρ_r and ρ_s and the value of each point in the grid is averaged over 1000 runs. Additionally, a dashed line is included in each plot to indicate the line $1 - \rho_s = (1 - \rho_r)^2$. The similarity between the left column and the right column demonstrates the utility of the empirical rate in estimating/approximating the analytical rate.

two reasons: *(i)* for large matrices, even small σ for individual entry can add up to a large error on the entire matrix; and *(ii)* while the cost of computing $\lambda_{\min}$ (and hence, the region of convergence) is prohibitively expensive for large matrices, choosing small σ empirically guarantees the initialization is inside the region of convergence.

Next, we record the error sequence $\{\|\boldsymbol{X}^{(k)} - \boldsymbol{M}\|_F\}_{k=1}^{K}$ and determine if the algorithm converges linearly to $\boldsymbol{M}$ by checking whether there exists $\hat{K} \leq K$ such that $\|\boldsymbol{X}^{(\hat{K})} - \boldsymbol{M}\|_F < \epsilon\|\boldsymbol{X}^{(0)} - \boldsymbol{M}\|_F$, for $\epsilon = 10^{-8}$. If the relative error is above ϵ, we set the empirical rate to 1 to indicate that the algorithm does not converge linearly. However, it is important to note that this heuristic **does not perfectly** detect linear convergence since it overlooks the case in which the linear rate is extremely close to 1 and it requires more than $K = 10000$ iterations to reach a relative error below ϵ. As can be seen later, to compromise this computational limit, we resort to setting the analytical rate that is greater than 0.998 to 1 when making a comparison between the analytical rate and the empirical rate.[5] In case the relative error is less than ϵ, we terminate the algorithm at the $\hat{K}$-th iteration (early stop) and perform a simple fitting for an exponential decrease on the error sequence $\{\|\boldsymbol{X}^{(k)} - \boldsymbol{M}\|_F\}_{k=1}^{\hat{K}}$ to obtain the empirical rate.

After obtaining the analytical rate and the empirical rate over the 2-D grid, we report the result in the contour plots of the rate as a function of ρ_r and ρ_s in Fig. 4(a) and Fig. 4(b). Since our original grid is non-uniform, we perform a scattered data interpolation, which uses a Delaunay triangulation of the scattered sample points to perform interpolation [52], to

[5] Substituting $\epsilon = 10^{-8}$ and $K(\epsilon) = 10000$ into (10) and assuming the constant c is negligible, we obtain $\lambda_{\min}(\boldsymbol{H}) \approx 1.8 \times 10^{-3}$, which in turn implies $\rho = 1 - \lambda_{\min}(\boldsymbol{H}) = 0.998$.

evaluate the rate over a 1001×1001 uniform grid based on ρ_r and ρ_s. Due to the aforementioned limitation of estimating the empirical rate, we apply a threshold of 0.998 to both the interpolated data for the analytical rate and the empirical rate, setting any value above the threshold to 1.

Finally, at each point of the 12×21 grid, we calculate the probability of linear convergence over 1000 runs. For the analytical rate, the linear convergence is determined by checking whether $\lambda_{\min}(\boldsymbol{H}) < 1$. For the empirical rate, we use the aforementioned discussion on determining whether the algorithm converges linearly with $K = 10000$ and $\epsilon = 10^{-8}$. The results are visualized in Fig. 4(c) and Fig. 4(d).

Results. Given the values of the analytical rate and the empirical rate of 1000 matrix completion settings for each point on the 12×21 grid, the mean squared difference between the two rates in our experiment is 2.9659×10^{-5}. Figure 4 illustrates the similarity between the analytical rate and the empirical rate evaluated under various settings of matrix completion. In both Fig. 4(a) and Fig. 4(b), we observed a matching behavior as in Fig. 1: smaller rank and more observation result in faster linear convergence of IHTSVD. However, the contour lines in Fig. 4 are not as smooth as those with asymptotic behavior in Fig. 1 due to the resolution of the grid as well as the large variance of the convergence rate under different sampling patterns when n_1 and n_2 are relatively small. On the other hand, it can be seen from Fig. 4(c) and Fig. 4(d) that there is a linear-convergence area (black) above the boundary line at $1 - \rho_s = (1 - \rho_r)^2$ and a no-linear-convergence area (white) below the boundary line. The transition area (gray) near the boundary line corresponds to the settings in which some sampling sets yield $\lambda_{\min}(\boldsymbol{H}) = 0$ while some other sampling sets yield $\lambda_{\min}(\boldsymbol{H}) > 0$. We discuss this transition region further in the next experiment.

To conclude, note that in order to obtain the analytical rate, we need to compute the smallest eigenvalue of a $(n - s) \times (n - s)$ matrix, which is computationally expensive for large $n = n_1 n_2$. In particular, when $s = \mathcal{O}(n)$, the cost of computing the analytical rate is $\mathcal{O}(n^3)$. On the other hand, the empirical rate offers an alternative but more efficient way to estimate the convergence rate via running Algorithm 1 whose computational complexity per iteration is $\mathcal{O}(nr)$. As a by-product, our proposed empirical rate can be used to efficiently estimate the smallest eigenvalue of the large matrix $\boldsymbol{H}$.

B. Non-asymptotic Rate versus Asymptotic Rate

In this experiment, we compare the asymptotic rate given in Theorem 3 with the convergence rate of IHTSVD for large-scale matrix completion. For convenience, we refer to the latter as the non-asymptotic rate. As mentioned, we use the empirical rate instead of the analytical rate to estimate the non-asymptotic rate due to the computational efficiency.

Data generation. We consider two settings of (n_1, n_2), i.e., $n_1 = 500, n_2 = 400$ and $n_1 = 1200, n_2 = 1000$. Similar to the previous experiment, we generate $\boldsymbol{M}$ and Ω based on a 2-D grid over r and s. While the values of s are still selected from the set $\{0.2n, 0.23n, 0.26n, \ldots, 0.8n\}$, the values of r are chosen differently for each setting of (n_1, n_2). In particular, for $n_1 = 500, n_2 = 400$, we select the values of r from the linearly spaced set $\{1, 4, 7, \ldots, 118\}$. For $n_1 = 1200, n_2 = 1000$, we select the values of r from the linearly spaced set $\{1, 9, 17, \ldots, 297\}$. Thus, in the former setting, the grid size is 40×21, while in the latter setting, the grid size is 38×21. We note that both grids are non-uniform in terms of ρ_r and ρ_s.

Implementation. The calculations of the empirical rate and the probability of linear convergence are the same as in the previous experiment. For computational efficiency, we omit the points on the grid that are below the boundary line $1 - \rho_s = (1 - \rho_r)^2$, i.e., $s < (n_1 + n_2 - r)r$, since it is evident that there is no linear convergence guaranteed at these points. No analytical rate is given in this experiment because calculating the smallest eigenvalue of a $(n_1 n_2 - s) \times (n_1 n_2 - s)$ matrix is computationally expensive for large n_1 and n_2. On the other hand, the contour plot of the asymptotic rate is straightforward to obtain using (20).

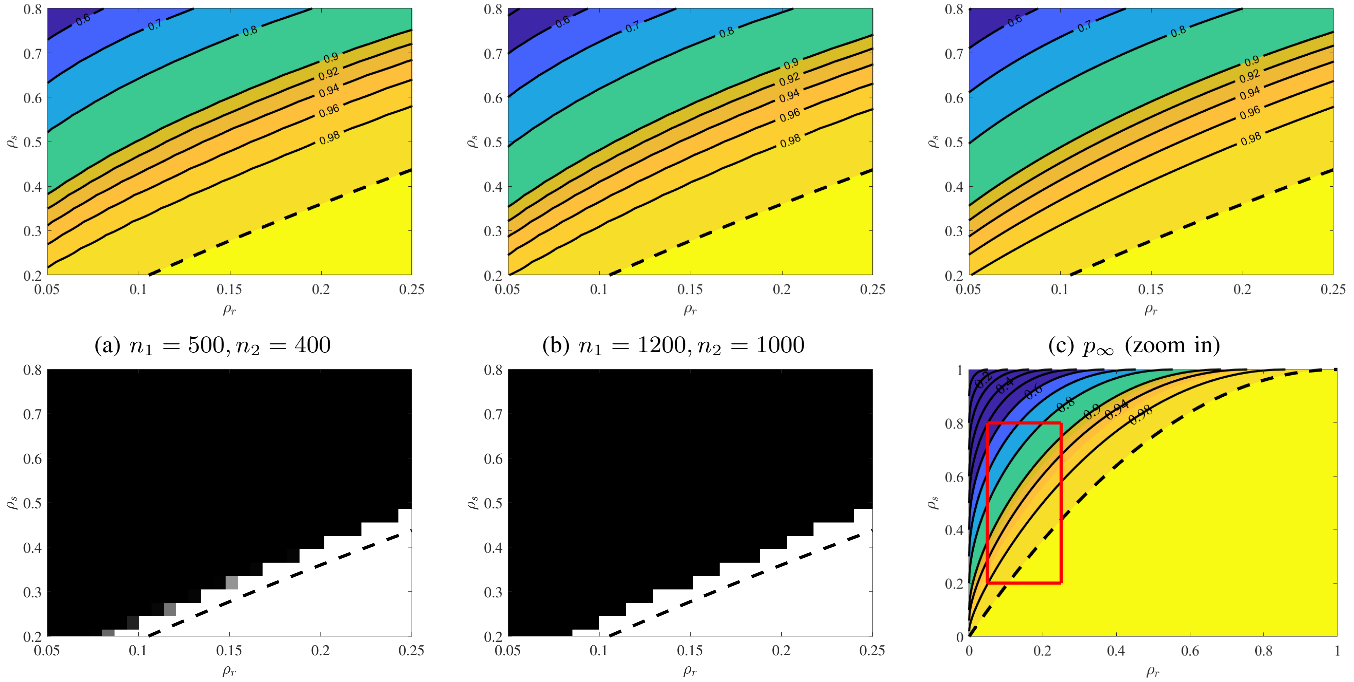
Results. In Fig. 5(a) and Fig. 5(b), we present the average empirical rate of linear convergence of IHTSVD as a function of the relative rank and the sampling rate in two large-scale settings. Observing the average empirical rate from Fig. 4(b) to Fig. 5(a) and to Fig. 5(b) as the dimensions increase, we note a shift of the contour lines towards the bottom-right corner, approaching those of the asymptotic rate in Fig. 5(c). This matches our intuition from Theorem 3 that as the dimensions grow to infinity, the linear rate of IHTSVD converges to the asymptotic rate p_∞. Additionally, from Fig. 4(d), Fig. 5(d), and Fig. 5(e), we observe that the linear-convergence area (black) becomes larger in larger matrix completion settings, indicating the isoline at 0.998 approaches closer to the line $1 - \rho_s = (1 - \rho_r)^2$ (dashed line). It is notable, however, that the transition between the linear-convergence area and the no-linear-convergence area is more abrupt as the dimensions increase. This phenomenon also matches our intuition in Conjecture 1, indicating that there is smaller variance in the empirical rate in large-scale settings, with respect to different random sampling patterns on the same underlying matrix.[7]

VI. Conclusions and Future Work

In this paper, we established a closed-form expression of the local linear convergence rate of an iterative hard thresholding method for solving matrix completion. We also identified the local region around the solution that guarantees the convergence of the algorithm. Our result holds for a wide range of matrix completion settings that do not necessarily require the assumptions on the incoherence of the underlying matrix and the specific choice of sample complexity. Furthermore, in large-scale settings, we leveraged the result from random matrix theory to offer a simple estimation of the asymptotic convergence rate in practice. Under certain assumptions, we

[6] In (d) and (e), the black color corresponds to linear convergence, whereas the white color corresponds to no linear convergence.

[7] Another evidence supporting this argument is the comparison of the coefficient of variation of the empirical rate in Fig. 5(a) and Fig. 5(b). We provide the detail in Fig. 3 in the Supplementary Material.



(a) $n_1 = 500, n_2 = 400$

(b) $n_1 = 1200, n_2 = 1000$

(c) p_∞ (zoom in)

(d) $n_1 = 500, n_2 = 400$

(e) $n_1 = 1200, n_2 = 1000$

(f) p_∞ (zoom out)

Fig. 5: The empirical rate and the asymptotic rate of convergence of IHTSVD as a function of the relative rank ρ_r and the sampling ratio ρ_s. (a) Contour plot of the empirical rate as a function of ρ_r and ρ_s for $n_1 = 500, n_2 = 400$. (b) Contour plot of the empirical rate as a function of ρ_r and ρ_s for $n_1 = 1200, n_2 = 1000$. (c) A zoom-in contour plot of the asymptotic rate as a function of ρ_r and ρ_s. (d) Empirical probability of linear convergence based on the empirical rate in (a). (e) Empirical probability of linear convergence based on the empirical rate in (b).[6] (f) A zoomed-out contour plot of the asymptotic rate as a function of ρ_r and ρ_s. The red solid rectangular corresponds to the zoomed-in region in (c). The data is evaluated based on 2-D grids over ρ_r and ρ_s and the value of each point in each grid is averaged over 100 runs. Additionally, a dashed line is included in each plot to indicate the line $1 - \rho_s = (1 - \rho_r)^2$. The striking similarity between plots (b) and (c) illustrates the utility of our convergence rate analysis in large-scale settings.

showed that the convergence rate of IHTSVD converges almost surely to our proposed estimate.

In future work, we would like to extend our local convergence analysis to other IHT methods such as accelerated IHT [27], [28]. In addition, it would also be interesting to study the non-asymptotic behavior of the convergence rate in large-scale settings. Finally, we believe the technique presented in this manuscript can be applied to study the local convergence of other non-convex methods such as alternating minimization [16] and gradient descent [18].

Appendix
Proof of Theorem 1

A. Proof of Lemma 1

By the definition of the error matrix, we have

$$\begin{aligned} \boldsymbol{E}^{(k+1)} &= \boldsymbol{X}^{(k+1)} - \boldsymbol{M} \\ &= \Big(\mathcal{P}_{\bar\Omega}\big(\mathcal{P}_r(\boldsymbol{X}^{(k)})\big) + \mathcal{P}_\Omega(\boldsymbol{M})\Big) - \big(\mathcal{P}_\Omega(\boldsymbol{M}) + \mathcal{P}_{\bar\Omega}(\boldsymbol{M})\big) \\ &= \mathcal{P}_{\bar\Omega}\big(\mathcal{P}_r(\boldsymbol{M} + \boldsymbol{E}^{(k)}) - \boldsymbol{M}\big). \end{aligned} \tag{25}$$

From Proposition 1, we can reorganize (5) to obtain

$$\mathcal{P}_r(\boldsymbol{M} + \boldsymbol{E}^{(k)}) - \boldsymbol{M} = \boldsymbol{E}^{(k)} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{E}^{(k)} \boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{E}^{(k)}).$$

Substituting the last equation back into (25) yields the recursion on the error matrix as in (13).

Next, let us denote $\boldsymbol{e}^{(k)} = \boldsymbol{S}_{\bar\Omega}^\top \mathrm{vec}(\boldsymbol{E}^{(k)})$, for $k = 1, 2, \ldots$. Vectorizing equation (13) and left-multiplying both sides with $\boldsymbol{S}_{\bar\Omega}$ yield

$$\boldsymbol{e}^{(k+1)} = \boldsymbol{S}_{\bar\Omega}^\top \mathrm{vec}\Big(\mathcal{P}_{\bar\Omega}\big(\boldsymbol{E}^{(k)} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{E}^{(k)} \boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{E}^{(k)})\big)\Big).$$

Using the property of selection matrices in Definition 2, we further have

$$\begin{aligned} \boldsymbol{e}^{(k+1)} &= \boldsymbol{S}_{\bar\Omega}^\top \boldsymbol{S}_{\bar\Omega} \boldsymbol{S}_{\bar\Omega}^\top \mathrm{vec}\big(\boldsymbol{E}^{(k)} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{E}^{(k)} \boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{E}^{(k)})\big) \\ &= \boldsymbol{S}_{\bar\Omega}^\top \mathrm{vec}\big(\boldsymbol{E}^{(k)} - \boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{E}^{(k)} \boldsymbol{P}_{\boldsymbol{V}_\perp} + \boldsymbol{R}(\boldsymbol{E}^{(k)})\big). \end{aligned}$$

Since $\mathrm{vec}(\boldsymbol{P}_{\boldsymbol{U}_\perp} \boldsymbol{E}^{(k)} \boldsymbol{P}_{\boldsymbol{V}_\perp}) = (\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}) \mathrm{vec}(\boldsymbol{E}^{(k)})$, the last equation can be represented as

$$\begin{aligned} \boldsymbol{e}^{(k+1)} = \boldsymbol{S}_{\bar\Omega}^\top \mathrm{vec}(\boldsymbol{E}^{(k)}) &- \boldsymbol{S}_{\bar\Omega}^\top (\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp}) \mathrm{vec}(\boldsymbol{E}^{(k)}) \\ &+ \boldsymbol{S}_{\bar\Omega}^\top \mathrm{vec}\big(\boldsymbol{R}(\boldsymbol{E}^{(k)})\big). \end{aligned} \tag{26}$$

On the other hand, (13) implies, for any $k \geq 1$, $\boldsymbol{E}^{(k)} = \mathcal{P}_{\bar\Omega}(\boldsymbol{E}^{(k)})$ and

$$\mathrm{vec}(\boldsymbol{E}^{(k)}) = \mathrm{vec}\big(\mathcal{P}_{\bar\Omega}(\boldsymbol{E}^{(k)})\big) = \boldsymbol{S}_{\bar\Omega} \boldsymbol{S}_{\bar\Omega}^\top \mathrm{vec}(\boldsymbol{E}^{(k)}) = \boldsymbol{S}_{\bar\Omega} \boldsymbol{e}^{(k)}.$$

Substituting the last equation into the RHS of (26) yields (14).

B. Proof of Lemma 2

Applying the triangle inequality to the RHS of (14) yields

$$\left\|\boldsymbol{e}^{(k+1)}\right\|_2 \leq \left\|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{e}^{(k)}\right\|_2 + \left\|\boldsymbol{r}(\boldsymbol{e}^{(k)})\right\|_2, \tag{27}$$

where we recall $\boldsymbol{H} = \boldsymbol{S}_{\bar{\Omega}}^{\top}(\boldsymbol{P}_{\boldsymbol{V}_\perp} \otimes \boldsymbol{P}_{\boldsymbol{U}_\perp})\boldsymbol{S}_{\bar{\Omega}}$. By the definition of the operator norm, we have

$$\begin{aligned}\left\|(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{e}^{(k)}\right\|_2 &\leq \|\boldsymbol{I}-\boldsymbol{H}\|_2 \left\|\boldsymbol{e}^{(k)}\right\|_2 \\ &= \max_i \big\{|1-\lambda_i(\boldsymbol{H})|\big\} \cdot \left\|\boldsymbol{e}^{(k)}\right\|_2 \\ &= \big(1-\lambda_{\min}(\boldsymbol{H})\big)\left\|\boldsymbol{e}^{(k)}\right\|_2, \end{aligned} \tag{28}$$

where the last equality stems from the fact that all eigenvalues of $\boldsymbol{H}$ lie between 0 and 1. From (27) and (28), we obtain

$$\left\|\boldsymbol{e}^{(k+1)}\right\|_2 \leq \big(1-\lambda_{\min}(\boldsymbol{H})\big)\left\|\boldsymbol{e}^{(k)}\right\|_2 + \left\|\boldsymbol{r}(\boldsymbol{e}^{(k)})\right\|_2. \tag{29}$$

The conclusion of lemma follows from the fact that

$$\left\|\boldsymbol{e}^{(k)}\right\|_2 = \left\|\mathcal{P}_{\bar{\Omega}}(\boldsymbol{E}^{(k)})\right\|_F = \left\|\boldsymbol{E}^{(k)}\right\|_F$$

and

$$\left\|\boldsymbol{r}(\boldsymbol{e}^{(k)})\right\|_2 \leq \left\|\boldsymbol{R}(\boldsymbol{E}^{(k)})\right\|_F \leq \frac{c_1}{\sigma_r}\left\|\boldsymbol{E}^{(k)}\right\|_F^2.$$

Acknowledgment

We would like to thank Dr. Thinh Nguyen of Oregon State University for his feedback on the work at the early stages.

References

[1] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," in *Proc. Int. Conf. Mach. Learn.*, 2003, pp. 720–727.

[2] N. Srebro, J. Rennie, and T. S. Jaakkola, "Maximum-margin matrix factorization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1329–1336.

[3] J. D. Rennie and N. Srebro, "Fast maximum margin matrix factorization for collaborative prediction," in *Proc. Int. Conf. Mach. Learn.* ACM, 2005, pp. 713–719.

[4] G. Takács, I. Pilászy, B. Németh, and D. Tikk, "Investigation of various matrix factorization methods for large recommender systems," in *Proc. IEEE Int. Conf. Data Min. Workshops*. IEEE, 2008, pp. 553–562.

[5] Z. Liu and L. 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, 2010.

[6] K. Mohan and M. Fazel, "Reweighted nuclear norm minimization with application to system identification," in *Proc. Am. Control Conf.* IEEE, 2010, pp. 2953–2959.

[7] Z. Liu, A. Hansson, and L. Vandenberghe, "Nuclear norm system identification with missing inputs and outputs," *Syst. Control. Lett.*, vol. 62, no. 8, pp. 605–612, 2013.

[8] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2009, pp. 2080–2088.

[9] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, pp. 1–37, 2011.

[10] E. J. Candès and B. Recht, "Exact matrix completion via convex optimization," *Found. Comput. Math.*, vol. 9, no. 6, p. 717, 2009.

[11] J.-F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM J. Optim.*, vol. 20, no. 4, pp. 1956–1982, 2010.

[12] K.-C. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific J. Optim.*, vol. 6, no. 3, pp. 615–640, 2010.

[13] M. Jaggi, "Revisiting Frank-Wolfe: Projection-free sparse convex optimization," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2013, pp. 427–435.

[14] N. Rao, P. Shah, and S. Wright, "Forward-backward greedy algorithms for atomic norm regularization," *IEEE Trans. Signal Process.*, vol. 63, no. 21, pp. 5798–5811, 2015.

[15] N. Boyd, G. Schiebinger, and B. Recht, "The alternating descent conditional gradient method for sparse inverse problems," *SIAM J. Optim.*, vol. 27, no. 2, pp. 616–639, 2017.

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

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

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

[19] 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.* PMLR, 2018, pp. 3345–3354.

[20] 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.

[21] 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.

[22] P. Jain, R. Meka, and I. S. Dhillon, "Guaranteed rank minimization via singular value projection," in *Proc. Adv. Neural Inf. Process. Syst.*, 2010, pp. 937–945.

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

[24] J. Tanner and K. Wei, "Normalized iterative hard thresholding for matrix completion," *SIAM J. Sci. Comput.*, vol. 35, no. 5, pp. S104–S125, 2013.

[25] J. D. Blanchard, J. Tanner, and K. Wei, "CGIHT: conjugate gradient iterative hard thresholding for compressed sensing and matrix completion," *Inf. Inference: J. IMA*, vol. 4, no. 4, pp. 289–327, 2015.

[26] E. Chunikhina, R. Raich, and T. Nguyen, "Performance analysis for matrix completion via iterative hard-thresholded SVD," in *Proc. IEEE Stat. Signal Process. Workshop*. IEEE, 2014, pp. 392–395.

[27] T. Vu and R. Raich, "Accelerating iterative hard thresholding for low-rank matrix completion via adaptive restart," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2019, pp. 2917–2921.

[28] ——, "Local convergence of the Heavy Ball method in iterative hard thresholding for low-rank matrix completion," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2019, pp. 3417–3421.

[29] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proc. Int. Conf. Mach. Learn.*, 2009, pp. 457–464.

[30] M. J. Lai and A. Varghese, "On convergence of the alternating projection method for matrix completion and sparse recovery problems," *arXiv preprint arXiv:1711.02151*, 2017.

[31] D. Goldfarb and S. Ma, "Convergence of fixed-point continuation algorithms for matrix rank minimization," *Found. Comput. Math.*, vol. 11, no. 2, pp. 183–210, 2011.

[32] P. Jain and P. Netrapalli, "Fast exact matrix completion with finite samples," in *Proc. Conf. Learn. Theory*, 2015, pp. 1007–1034.

[33] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[34] T. Vu, E. Chunikhina, and R. Raich, "Perturbation expansions and error bounds for the truncated singular value decomposition," *Linear Algebra Appl.*, vol. 627, pp. 94–139, 2021.

[35] M. Abramowitz and I. A. Stegun, "Handbook of mathematical functions with formulas, graphs, and mathematical tables," *NBS Appl. Math. Ser.*, vol. 55, 1964.

[36] R. Bellman, *Stability theory of differential equations*. Courier Corporation, 2008.

[37] B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Comput. Math. Math. Phys.*, vol. 4, no. 5, pp. 1–17, 1964.

[38] T. Vu and R. Raich, "A closed-form bound on the asymptotic linear convergence of iterative methods via fixed point analysis," *Optim. Lett.*, vol. 1, pp. 1–14, 2022.

[39] ——, "On asymptotic linear convergence of projected gradient descent for constrained least squares," *IEEE Trans. Signal Process.*, vol. 4, pp. 4061–4076, 2022.

[40] R. Raich and J. Kim, "On the eigenvalue distribution of column sub-sampled semi-unitary matrices," in *Proc. IEEE Stat. Signal Process. Workshop*. IEEE, 2016, pp. 1–5.

[41] A. Edelman and N. R. Rao, “Random matrix theory,” *Acta Numer.*, vol. 14, p. 233, 2005.

[42] K. W. Wachter, “The limiting empirical measure of multiple discriminant ratios,” *Ann. Statist.*, vol. 8, pp. 937–957, 1980.

[43] M. Capitaine and M. Casalis, “Asymptotic freeness by generalized moments for gaussian and wishart matrices. Application to beta random matrices,” *Indiana Univ. Math. J.*, pp. 397–431, 2004.

[44] B. Collins, “Product of random projections, Jacobi ensembles and universality problems arising from free probability,” *Probab. Theory Relat. Fields*, vol. 133, no. 3, pp. 315–344, 2005.

[45] P. J. Forrester, “Quantum conductance problems and the Jacobi ensemble,” *J. Phys. A: Math. Gen.*, vol. 39, no. 22, p. 6861, 2006.

[46] I. M. Johnstone, “Multivariate analysis and Jacobi ensembles: Largest eigenvalue, Tracy–Widom limits and rates of convergence,” *Ann. Statist.*, vol. 36, no. 6, p. 2638, 2008.

[47] K. Zyczkowski and H.-J. Sommers, “Truncations of random unitary matrices,” *J. Phys. A: Math. Gen.*, vol. 33, no. 10, p. 2045, 2000.

[48] T. Jiang, “Approximation of haar distributed matrices and limiting distributions of eigenvalues of Jacobi ensembles,” *Probab. Theory Relat. Fields*, vol. 144, no. 1-2, pp. 221–246, 2009.

[49] Z. Dong, T. Jiang, and D. Li, “Circular law and arc law for truncation of random unitary matrix,” *J. Math. Phys.*, vol. 53, no. 1, 013301, pp. 1–14, 2012.

[50] A. Hedayat and W. D. Wallis, “Hadamard matrices and their applications,” *Ann. Statist.*, vol. 6, no. 6, pp. 1184–1238, 1978.

[51] B. Farrell and R. R. Nadakuditi, “Local spectrum of truncations of Kronecker products of Haar distributed unitary matrices,” *Random Matrices: Theory Appl.*, vol. 4, no. 1, 2013.

[52] I. Amidror, “Scattered data interpolation methods for electronic imaging systems: A survey,” *J. Electron. Imaging*, vol. 11, no. 2, pp. 157–176, 2002.

Trung Vu received the B.S. degree in Computer Science from Hanoi University of Science and Technology, Hanoi, Vietnam, in 2014. He received the Ph.D. degree in Computer Science at the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, Oregon, USA, in 2022. He is currently a postdoctoral research associate at the Department of Computer Science and Electrical Engineering, University of Maryland, Baltimore County, Baltimore, Maryland, USA. His research interests include optimization methods, independent component analysis, and matrix factorization with applications in machine learning and signal processing.

Evgenia Chunikhina (IEEE Member) received the B.S. degree in applied mathematics, computer science, and mechanics from Voronezh State University, Russia, and the M.S. degree in mathematics, the M.Eng. degree in computer science, and the Ph.D. degree in computer science from Oregon State University. She is currently an assistant professor of data science with Pacific University, Oregon, USA. She completed a one-year postdoc with the University of So Paulo, So Paulo, Brazil. Her research interests include machine learning, artificial intelligence, statistical signal processing, information theory, compressed sensing, networks, random graphs, and biostatistics.

Raviv Raich (S’98–M’04–SM’17) received the B.Sc. and M.Sc. degrees from Tel-Aviv University, Tel-Aviv, Israel, in 1994 and 1998, respectively, and the Ph.D. degree from the Georgia Institute of Technology, Atlanta, GA, USA, in 2004, all in electrical engineering. Between 1999 and 2000, he was a Researcher with the Communications Team, Industrial Research, Ltd., Wellington, New Zealand. From 2004 to 2007, he was a Postdoctoral Fellow with the University of Michigan, Ann Arbor. He has been with the School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, as an Assistant Professor (2007–2013) and is currently an Associate Professor (2013-present). He was an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING from 2011 to 2014. He served as a member (2011–2016) and then chair (2017–2018) of the Machine Learning for Signal Processing (MLSP) Technical Committee (TC) of the IEEE Signal Processing Society. Since 2019, he is a member of the Signal Processing Theory and Methods (SPTM), TC of the IEEE Signal Processing Society. His research interests include: probabilistic modeling and optimization in signal processing and machine learning.