GLOBAL OPTIMALITY IN INDUCTIVE MATRIX COMPLETION

Mohsen Ghassemi[⋆] *Anand D. Sarwate*[⋆] *Naveen Goela*[†]

[⋆] Department of ECE, Rutgers, The State University of New Jersey
[†] Technicolor Research and Innovation Lab

ABSTRACT

Inductive matrix completion (IMC) is a model for incorporating side information in form of "features" of the row and column entities of an unknown matrix in the matrix completion problem. As side information, features can substantially reduce the number of observed entries required for reconstructing an unknown matrix from its given entries. The IMC problem can be formulated as a low-rank matrix recovery problem where the observed entries are seen as measurements of a smaller matrix that models the interaction between the column and row features. We take advantage of this property to study the optimization landscape of the factorized IMC problem. In particular, we show that the critical points of the objective function of this problem are either global minima that correspond to the true solution or are "escapable" saddle points. This result implies that any minimization algorithm with guaranteed convergence to a local minimum can be used for solving the factorized IMC problem.

Index Terms— inductive matrix completion, matrix recovery, saddle points, local minima

1. INTRODUCTION

Matrix completion [1,2] is an important technique in machine learning with applications in areas such as recommendation systems [3] or computer vision [4] where the task is to reconstruct a low-rank matrix $\mathbf{M}^* \in \mathbb{R}^{n_1 \times n_2}$ from a small number of given entries. Theoretical results in the literature show that the number of required samples for exact recovery is $O(rn \log^2 n)$ where $n = n_1 + n_2$ and $r = \mathrm{rank}(\mathbf{M}^*)$ [5,6]. In some applications, the algorithm may have acccess to *side information* that can be exploited to improve this sample complexity. For example, in many recommendation systems the system has additional information about both user profiles and items.

Among the many approaches to incorporate side information [7–13], *inductive matrix completion (IMC)* [7,8] models side information as knowledge of *feature spaces*. This is incorporated in the model by assuming that each entry of the unknown matrix of interest $\mathbf{M}^* \in \mathbb{R}^{n_1 \times n_2}$ is in form of $\mathbf{M}^*_{ij} = \mathbf{x}_i^T \mathbf{W}^* \mathbf{y}_j$, where $\mathbf{x}_i \in \mathbb{R}^{d_1}$ and $\mathbf{y}_j \in \mathbb{R}^{d_2}$ are known feature vectors of i-th row (user) and j-th column (item), respectively. The low-rank matrix completion problem in this case can be formulated as recovering a rank-r matrix $\mathbf{W}^* \in \mathbb{R}^{d_1 \times d_2}$ such that the observed entries are $\mathbf{M}_{ij} = \mathbf{x}_i^T \mathbf{W}^* \mathbf{y}_j$. In fact, the IMC problem translates to finding missing entries of $\mathbf{M}^*$ as recovering matrix $\mathbf{W}^*$ from its measurements in form of $\mathbf{x}_i \mathbf{W}^* \mathbf{y}_j = \left\langle \mathbf{x}_i \mathbf{y}_j^T, \mathbf{W}^* \right\rangle$ for $(i,j) \in \Omega$.

Using this model, the sample complexity decreases considerably if the size of matrix $\mathbf{M}$ is much larger than $\mathbf{W}^*$. Another advantage of this model is that rows/columns of the unknown matrix can be predicted without knowing even one of their entries using the corresponding feature vectors once we recover $\mathbf{W}^*$ using the given entries. This is not possible in standard matrix completion since a necessary condition for completing a rank-r matrix is that at least r entries of every row and every column are observed [1].

The nonconvex rank-r constraint makes the problem challenging. There are two main approaches in the matrix recovery literature to impose the low-rank structure in a tractable way. The first approach is using convex relaxations of the nonconvex rank-constrained problem [1, 6, 14–17]. In the IMC problem, at least $O(rd \log d \log n)$ samples are required for recovery of $\mathbf{W}^*$ using a trace-norm relaxation, where $d = d_1 + d_2$ [7, 8]. The trace-norm approach has also been proposed for the IMC problem with noisy features where the unknown matrix is modeled as $\mathbf{X}\mathbf{W}^*\mathbf{Y}^T + \mathbf{N}$ where the residual matrix N models imperfections and noise in the features [10].

Another approach uses matrix factorization, where the $d_1 \times d_2$ matrix $\mathbf{W}$ is expressed as $\mathbf{W} = \mathbf{U}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{d_1 \times r}$ and $\mathbf{V} \in \mathbb{R}^{d_2 \times r}$ [3, 18]. Jain et al. show that alternating minimization (AM) converges to the global solution of matrix sensing and matrix completion problems in linear time under standard conditions [18]. Inspired by this result, Zhong et al. [8] show that for the factorized IMC problem, $O(r^3 d \log d \max\{r, \log n\})$ samples are sufficient for ϵ-recovery of $\mathbf{W}^*$ using AM.

On the computational side, the per-iteration cost of the solvers of the convex matrix estimation problem is high since they require finding the SVD of a matrix in case of implementing singular value thresholding [19] or proximal gradient methods [20], or they involve solving a semi-definite program. On the other hand, both empirically and theoretically, stochastic gradient descent (SGD) and AM have been shown to find good local optima in many *nonconvex* matrix estimation problems and that suitable modifications to the objective function can find *global optima* [18, 21]. These simple local search algorithms have low memory requirement and per-iteration computational cost, due to the fact that in low-rank problems $r \ll d_1, d_2$. Although the IMC model reduces the dimensionality of the matrix estimation problem from $n_1 \times n_2$ to $d_1 \times d_2$, the lower complexity of the solvers of the factorized model is appealing [8].

On the theoretical side, the trace-norm based model is intriguing in that it allows for employing well-established tools to analyze the statistical performance of the convex program. Although the matrix factorization based models in general are theoretically less understood, recent works have studied the optimization landscape of some of these nonconvex problems and show that their objective functions are devoid of "poor" local minima. Problems such as matrix completion [22, 23], matrix sensing [24, 25], phase retrieval [26], deep (linear) neural networks [27, 28] are amenable to this approach. To the best of our knowledge, this work is the first to study the geometry and the statistical performance of IMC under the factorized model.

This paper is motivated by the recovery guarantees of AM for the (nonconvex) factorized IMC problem. Our key technical contribution is to use concentration inequalities to show that given a suf-

ficient number of measurements, the ensemble of sensing matrices $\mathbf{x}_i\mathbf{y}_j^T$ almost preverves the energy of all rank-$2r$ matrices, i.e. it satisfies *restricted isometry property* of order $2r$. This allows us to use the framework of Ge et al. [22] for matrix sensing problems. Our final result is that given at least $O(dr\max\{r^2, \log^2 n\})$ observations, in the (regularized) factorized IMC problem *i*) all local minima are globally optimal, *ii*) all local minima fulfill $\mathbf{U}\mathbf{V}^T = \mathbf{W}^*$, and *iii*) the saddle points are escapable in the sense that the Hessian at those points has at least one negative eigenvalue.

Our result implies that the success of AM in the nonconvex IMC problem is to some degree a result of the geometry of the problem and not solely due to the properties of the algorithm. In fact, any algorithm with guaranteed convergence to a local minimum, e.g. SGD [21], can be used for solving the factorized IMC problem.

2. PROBLEM MODEL

Notation and Definitions. Throughout this paper, vectors and matrices are, respectively, denoted by boldface lower case letters: $\mathbf{a}$ and boldface upper case letters: $\mathbf{A}$. We denote by $\mathbf{A}_{ij}$ the j-th element of the i-th row of $\mathbf{A}$. The smallest eigenvalue of $\mathbf{A}$ is denoted by $\lambda_{\min}(\mathbf{A})$. In matrix completion, the set of indices of the observed (given) entries of an incomplete matrix $\mathbf{A} \in \mathbb{R}^{n_1\times n_2}$ is denoted by Ω with size $m = |\Omega|$. Also, $\mathbf{A}_\Omega$ denotes the linear projection of $\mathbf{A}$ onto the space of $n_1 \times n_2$ matrices whose entries outside Ω are zero. The inner product of two matrices is defined as $\langle \mathbf{A}, \mathbf{B}\rangle = \mathrm{Tr}(\mathbf{A}^T\mathbf{B})$. In a noncovex optimization problem, a poor local minimum is a local minimum which is not globally optimum.

We repeatedly use the *(matrix) restricted isometry property (RIP)* [14] and the *strict saddle property* [21, 29] defined below.

Definition 1. *A linear operator $\mathcal{A}(\cdot) : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^m$ satisfies r-RIP with δ_r RIP constant if for every $\mathbf{W} \in \mathbb{R}^{d_1\times d_2}$ such that $rank(W) \le r$ it holds that*

$$(1-\delta_r)\|\mathbf{W}\|_F^2 \le \|\mathcal{A}(\mathbf{W})\|_2^2 \le (1+\delta_r)\|\mathbf{W}\|_F^2 .$$

Definition 2. *A twice differentiable function $f(\mathbf{x})$ is strict saddle if $\lambda_{\min}\left(\nabla^2 f(x)\right) < 0$ at its saddle points.*

Inductive Matrix Completion. Consider the nonconvex low-rank matrix completion problem

$$\min_{\mathbf{M}\in\mathbb{R}^{n_1\times n_2}} \|\mathbf{M}^*_\Omega - \mathbf{M}_\Omega\|_F^2 \quad \text{s.t.} \quad \mathrm{rank}(\mathbf{M}) \le r. \tag{1}$$

In an inductive matrix completion problem, the underlying matrix has the form $\mathbf{M}^* = \mathbf{X}\mathbf{W}^*\mathbf{Y}^T$ where the *side information* matrices $\mathbf{X} \in \mathbb{R}^{n_1\times d_1}$ and $\mathbf{Y} \in \mathbb{R}^{n_2\times d_2}$ are known and $\mathbf{W}^* = \mathbf{U}^*\mathbf{V}^{*T}$ with $\mathbf{U}^* \in \mathbb{R}^{d_1\times r}$, $\mathbf{V}^* \in \mathbb{R}^{d_2\times r}$ is unknown. Therefore, the problem can be written as

$$\min_{\mathbf{W}\in\mathbb{R}^{d_1\times d_2}} \left\|\left(\mathbf{M}^* - \mathbf{X}\mathbf{W}\mathbf{Y}^T\right)_\Omega\right\|_F^2 \text{ s.t. } \mathrm{rank}(\mathbf{W}) \le r. \tag{2}$$

This problem can be reformulated into an uncostrained nonconvex problem by expressing $\mathbf{W}$ as $\mathbf{U}\mathbf{V}^T$, where $\mathbf{U} \in \mathbb{R}^{d_1\times r}$, $\mathbf{V} \in \mathbb{R}^{d_2\times r}$:

$$\min_{\mathbf{U},\mathbf{V}} \left\|\left(\mathbf{M}^* - \mathbf{X}\mathbf{U}\mathbf{V}^T\mathbf{Y}^T\right)_\Omega\right\|_F^2 + R(\mathbf{U},\mathbf{V}) \tag{3}$$

The regularization term $R(\mathbf{U},\mathbf{V})$ is added to account for the invariance of the asymmetric factorized model to scaling of the factor matrices by reciprocal values. A common choice that suits our model is $R(\mathbf{U},\mathbf{V}) = \frac{1}{4}\|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\|_F^2$ [22, 25].

The objective function $f(\mathbf{U},\mathbf{V})$ in problem (3) alternatively can be written as

$$\sum_{(i,j)\in\Omega}\left(\mathbf{M}^*_{ij} - \left\langle \mathbf{x}_i\mathbf{y}_j^T, \mathbf{U}\mathbf{V}^T\right\rangle\right)^2 + \frac{1}{4}\left\|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\right\|_F^2$$

where $\mathbf{x}_i^T$ and $\mathbf{y}_j^T$ respectively are the ith and jth rows of $\mathbf{X}$ and $\mathbf{Y}$. Observe that $\left\langle \mathbf{x}_i\mathbf{y}_j^T, \mathbf{U}\mathbf{V}^T\right\rangle = \mathbf{x}_i^T\mathbf{U}\mathbf{V}^T\mathbf{y}_j$.

This shows that the IMC problem (3) can be thought of as a matrix sensing problem where we are given linear measurements of the $d_1 \times d_2$ matrix $\mathbf{W}^*$ by sensing matrices $\mathbf{A}_{ij} = \mathbf{x}_i\mathbf{y}_j^T$. Define the linear operator $\mathcal{A}$ such that $\mathcal{A}(\mathbf{W})$ is a vector whose elements are the measurements $\frac{1}{\sqrt{m}}\langle \mathbf{A}_{ij}, \mathbf{W}\rangle$.

In this paper, we make the following assumptions regarding the side information matrices and the sampling model.

Assumption 1 (Side information)**.** *The side information matrices satisfy $\mathbf{X}^T\mathbf{X} = n_1 I_{d_1}$ and $\mathbf{Y}^T\mathbf{Y} = n_2 I_{d_2}$.*[1] *We also make the assumption that for any given matrices $\bar{\mathbf{U}}$ and $\bar{\mathbf{V}}$ with orthogonal columns, the rows of the side information matrices (feature vectors) satisfy $\left\|\bar{\mathbf{U}}\mathbf{x}_i\right\|_2^2 \le \mu\bar{r}$ and $\left\|\bar{\mathbf{V}}\mathbf{y}_j\right\|_2^2 \le \mu\bar{r}$, where $\bar{r} = \max(r, \log n_1, \log n_2)$ and μ is a positive constant. This assumption, for example, is satisfied with high probability when the side information matrices $\mathbf{X}$ and $\mathbf{Y}$ are instances generated from a random orthonormal matrix model (the first d_1 (respectively d_2) columns) and rescaled by $\sqrt{n_1}$ (respectively $\sqrt{n_2}$) [1, 8].*

Assumption 2 (Sampling model)**.** *Indices i and j are independent and uniformly distributed on $\{1, 2, \ldots, n\}$.*

3. GEOMETRIC ANALYSIS

We are interested in the geometric landscape of the objective function in the IMC problem (3). We will show that simple algorithms like AM can recover the true underlying matrix with arbitrary accuracy because given enough observations, the objective function in this problem *i*) has no poor local minima, *ii*) has only local minima which satisfy $\mathbf{U}\mathbf{V}^T = \mathbf{W}^*$, and *iii*) is *strict saddle*.

We employ the framework developed by Ge et al. [22] for matrix sensing to show that the objective function of the IMC problem (3) satisfies properties *i*), *ii*), and *iii*). Theorem 1 states the main result of this paper.

Theorem 1. *Consider the IMC problem* (3) *seen as a matrix recovery problem with sensing matrices $\mathbf{A}_{ij} = \mathbf{x}_i\mathbf{y}_j^T$ for $(i,j) \in \Omega$, such that Assumptions 1 and 2 hold. Let $\bar{r} = \max\{r, \log n_1, \log n_2\}$. If the number of measurements is $m = O\left(\mu^2 dr^2\bar{r}\right)$, then there exists a positive constant h such that with probability higher than $1 - 2\exp(-hm)$, the nonconvex objective function $f(\mathbf{U},\mathbf{V})$ has the following properties: i) all its local minima are globally optimal , ii) all its local minima satisfy $\mathbf{U}\mathbf{V}^T = \mathbf{M}^*$, and iii) it satisfies the strict saddle property.*

The proof strategy here is to show that at any stationary point of $f(\mathbf{U},\mathbf{V})$ (and its neighborhood), the "difference" $\mathbf{\Delta}$ between the point and and the true solution (which is basically the Euclidian distance between the point and its nearest global minimum) is a *descent direction*. This means that $(\mathbf{U},\mathbf{V})$ cannot be local minimum unless $\mathbf{\Delta} = \mathbf{0}$ (no poor local minima and exact recovery) and that the

[1] This is not a restrictive assumption since we can apply orthonormalization methods such as Gram-Schmidt process [30] and then rescale to ensure this assumption is satisfied.

Hessian at the saddle points cannot be positive semidefinite (strict saddle property). To this end, following the proposed strategy by Ge, Jin, and Zheng [22], we construct $\mathbf{B} = \begin{pmatrix}\mathbf{U}\\ \mathbf{V}\end{pmatrix} \in \mathbb{R}^{(d_1+d_2)\times r}$, $\mathbf{W} = \mathbf{U}\mathbf{V}^T$, and $\mathbf{N} = \mathbf{B}\mathbf{B}^T$ and reformulate problem (3) as the positive semidefinite (PSD) low-rank matrix recovery problem

$$\min_{\mathbf{B}} f(\mathbf{B}) = \|\mathcal{T}(\mathbf{N}^*) - \mathcal{T}(\mathbf{B}\mathbf{B}^T)\|_2^2. \tag{4}$$

where $\mathbf{B}^* = \begin{pmatrix}\mathbf{U}^*\\ \mathbf{V}^*\end{pmatrix}$, $\mathbf{N}^* = \mathbf{B}^*\mathbf{B}^{*T}$, and $\mathcal{T}$ is a linear operator such that $\mathcal{T}(\mathbf{N})$ is an ensemble of m measurements $\langle \mathbf{T}_{ij}, \mathbf{N}\rangle$ such that $\langle \mathbf{T}_{ij}, \mathbf{N}\rangle^2 = \frac{1}{m}\left(4\langle \mathbf{A}_{ij}, \mathbf{W}\rangle^2 + \|\mathbf{U}\mathbf{U}^T - \mathbf{V}\mathbf{V}^T\|_F^2\right)$. The following definition captures the invariance of the solution of symmetric matrix recovery to rotation, negation, or column permutation.

Definition 3. *Given matrices $\mathbf{B}, \mathbf{B}^*$, define their difference $\mathbf{\Delta} = \mathbf{B} - \mathbf{B}^*\mathbf{D}$, where $\mathbf{D} = \underset{\mathbf{Z}:\mathbf{Z}\mathbf{Z}^T=\mathbf{Z}^T\mathbf{Z}=\mathbf{I}_{2r}}{\operatorname{argmin}} \|\mathbf{B} - \mathbf{B}^*\mathbf{Z}\|_F^2$.*

The second order term in the Taylor expansion of $f(\mathbf{B})$ becomes dominant in the neighborhood of stationary points. Therefore it suffices to show that $\boldsymbol{\delta}^T\nabla^2 f(\mathbf{B})\boldsymbol{\delta}$, where $\boldsymbol{\delta} = \operatorname{vec}(\mathbf{\Delta})$, is strictly negative for points in these regions, except when $\mathbf{\Delta} = \mathbf{0}$, to prove that $\mathbf{\Delta}$ is a descent direction. Theorem 2 states that if linear operator $\mathcal{B}$ is RIP, then we can show $\boldsymbol{\delta}^T\nabla^2 f(\mathbf{B})\boldsymbol{\delta}$ is strictly negative in the neighborhood of stationary points unless they correspond to $\mathbf{N}^*$ (and its submatrix $\mathbf{W}^*$) and consequently $\mathbf{M}^* = \mathbf{X}\mathbf{W}^*\mathbf{Y}^T$, the ground truth matrix in problem (3).

Theorem 2. *Consider the objective function of the PSD matrix recovery problem* (4)*. If the measurement operator $\mathcal{T}$ satisfies $(2r, \delta_{2r})$-RIP, then any point satisfying $\|\nabla f(\mathbf{B})\|_F \le \xi$, the quadratic form $\boldsymbol{\delta}^T\nabla^2 f(\mathbf{B})\boldsymbol{\delta}$ for $\boldsymbol{\delta} = \operatorname{vec}(\mathbf{\Delta})$ defined above is negative unless $\|\mathbf{\Delta}\|_F \le K\xi/(1-5\delta_{2r})$ for some positive constant K.*

Proof sketch. The proof is based on the following equality (Lemma 7 in [22]):

$$\begin{aligned}\boldsymbol{\delta}^T\nabla^2 f(\mathbf{B})\boldsymbol{\delta} = &\left\|\mathcal{T}\left(\mathbf{\Delta}\mathbf{\Delta}^T\right)\right\|_2^2 - 3\|\mathcal{T}(\mathbf{N} - \mathbf{N}^*)\|_2^2\\ &+ 4\langle \nabla f(\mathbf{B}), \mathbf{\Delta}\rangle.\end{aligned} \tag{5}$$

Using the RIP property of $\mathcal{T}$, which implies that the measuring operator captures the energy of the observed matrix with small deviation, and applying the bounds $\left\|\mathbf{\Delta}\mathbf{\Delta}^T\right\|_F^2 \le 2\|\mathbf{N} - \mathbf{N}^*\|_F^2$ and $k\|\mathbf{\Delta}\|_F^2 \le \|\mathbf{N} - \mathbf{N}^*\|_F^2$ (Lemma 6 in [22]) results in

$$\boldsymbol{\delta}^T\nabla^2 f(\mathbf{B})\boldsymbol{\delta} \le -k(1-5\delta_{2r})\|\mathbf{\Delta}\|_F^2 + 4\xi\|\mathbf{\Delta}\|_F. \tag{6}$$

Therefore the bilinear form on the left cannot be nonnegative unless $\|\mathbf{\Delta}\|_F^2 \le 4\xi/(k(1-5\delta_{2r}))$.

□

Now, we show that the linear operator $\mathcal{A}$ and consequently $\mathcal{T}$ are $2r$-RIP. Note that it is important that we show $2r$-RIP rather that r-RIP because in Theorem 2, $\mathcal{T}$ is applied to $\mathbf{B} - \mathbf{B}^*$ which can be of rank at most $2r$. It also guarantees that the null space of $\mathcal{T}$ does not include any matrices of rank $2r$ or less, which is a necessary and sufficient condition for unique recovery [31, 32].

Theorem 3. *Consider the IMC problem* (3) *seen as a matrix recovery problem with sensing matrices $\mathbf{A}_{ij} = \mathbf{x}_i\mathbf{y}_j^T$ for $(i,j) \in \Omega$, such that Assumptions 1 and 2 hold. If the number of measurements $m = O\left(\mu^2 d\bar{r}^2 r\log(36\sqrt{2}/\delta)/\delta^2\right)$, then there exists a positive constant h such that with probability higher than $1 - 2\exp(-hm)$, the linear operator $\mathcal{A}(\cdot)$, seen as an ensemble of m measurements $\frac{1}{\sqrt{m}}\langle \mathbf{A}_{ij}, \cdot\rangle$, is $2r$-RIP with RIP constant $\delta_{2r} = 2\delta$.*

Proof. We show that $\left\|\mathcal{A}(\widetilde{\mathbf{W}})\right\|_2^2$ is close to $\left\|\widetilde{\mathbf{W}}\right\|_F^2$ for all rank-$2r$ matrices $\widetilde{\mathbf{W}}$, i.e., $\left|\left\|\mathcal{A}(\widetilde{\mathbf{W}})\right\|_2^2 - \left\|\widetilde{\mathbf{W}}\right\|_F^2\right| \le \delta_{2r}\left\|\widetilde{\mathbf{W}}\right\|_F^2$. We use Bernstein's inequality to find a bound on the deviation of the sum of m random variables $\frac{1}{\sqrt{m}}\langle \mathbf{x}_i\mathbf{y}_j^T, \widetilde{\mathbf{W}}\rangle$ from their mean $\left\|\widetilde{\mathbf{W}}\right\|_F^2$ for a given rank-$2r$ matrix $\widetilde{\mathbf{W}}$. This is formally stated in Lemma 1. Then we find a similar bound for all rank-$2r$ (or less) matrices.

Lemma 1. *Consider the same setting as Theorem 3. For a given matrix $\widetilde{\mathbf{W}}$ of rank $2r$, with probability at least $1 - C\exp(-cm)$, for some positive constatnts C and c, we have*

$$(1-\delta_{2r})\left\|\widetilde{\mathbf{W}}\right\|_F^2 \le \left\|\mathcal{A}(\widetilde{\mathbf{W}})\right\|_2^2 \le (1+\delta_{2r})\left\|\widetilde{\mathbf{W}}\right\|_F^2.$$

Proof of Lemma 1. In order to show that the average random measurement $\|\mathcal{A}(\widetilde{\mathbf{W}})\|_2^2 = \frac{1}{m}\sum_{ij}\langle \mathbf{A}_{ij}, \widetilde{\mathbf{W}}\rangle^2$ is close to its expectation $\|\widetilde{\mathbf{W}}\|_F^2$, we use Bernstein's inequality [33]:

$$\mathbb{P}\left(\left|\bar{Z} - \eta_Z\right| > \epsilon\right) \le 2\exp\Big(\frac{-m\epsilon^2/2}{\frac{1}{m}\sum_{ij}\operatorname{Var}(Z_{ij}) + B_Z\epsilon/3}\Big),$$

where $\bar{Z} = \frac{1}{m}\sum_{ij} Z_{ij}$ and η_Z is the mean of the random variables. To apply Bernstein's inequality, we need to find the expectation, the variance (or an upper bound on the variance), and an upper bound on the absolute value of the random variables in the summand, denoted by $Z_{ij} = \mathbf{x}_i^T\widetilde{\mathbf{W}}\mathbf{y}_j\mathbf{y}_j^T\widetilde{\mathbf{W}}\mathbf{x}_i$. Note that $\mathbf{X}$ and $\mathbf{Y}$ are known orthogonal matrices and the only source of randomness is the choice of (i,j). First, we find the mean of the random variables:

$$\begin{aligned}\eta_Z &= \mathbb{E}\left[\mathbf{x}_i^T\widetilde{\mathbf{W}}\mathbf{y}_j\mathbf{y}_j^T\widetilde{\mathbf{W}}\mathbf{x}_i\right]\\ &= \mathbb{E}\left[e_i^T\mathbf{X}\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T\mathbf{Y}^T e_j e_j^T\mathbf{Y}\widetilde{\mathbf{W}}^T\mathbf{X}^T e_i\right]\\ &= \mathbb{E}\left[\operatorname{Tr}\left(\widetilde{\mathbf{V}}^T\mathbf{Y}^T e_j e_j^T\mathbf{Y}\widetilde{\mathbf{W}}^T\mathbf{X}^T e_i e_i^T\mathbf{X}\widetilde{\mathbf{U}}\right)\right]\\ &= \operatorname{Tr}\left(\widetilde{\mathbf{V}}^T\mathbf{Y}^T\mathbb{E}\left[e_j e_j^T\right]\mathbf{Y}\widetilde{\mathbf{W}}^T\mathbf{X}^T\mathbb{E}\left[e_i e_i^T\right]\mathbf{X}\widetilde{\mathbf{U}}\right)\\ &\overset{(a)}{=} \operatorname{Tr}\left(\widetilde{\mathbf{V}}^T\mathbf{Y}^T\mathbf{Y}\widetilde{\mathbf{W}}^T\mathbf{X}^T\mathbf{X}\widetilde{\mathbf{U}}\right)\\ &\overset{(b)}{=} \operatorname{Tr}\left(\widetilde{\mathbf{V}}^T\widetilde{\mathbf{W}}^T\widetilde{\mathbf{U}}\right)\\ &= \operatorname{Tr}\left(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T\cdot\widetilde{\mathbf{W}}^T\right)\\ &= \left\|\widetilde{\mathbf{W}}\right\|_F^2,\end{aligned} \tag{7}$$

where $\widetilde{\mathbf{W}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T$, equality (a) follows from $\mathbb{E}\left[e_i e_i^T\right] = \frac{1}{n_1}I_{n_1}$ and (b) follows from Assumption 1. Next we find an upper bound B_Z on $|Z_{ij}|$:

$$\begin{aligned}|Z_{ij}| &= \cdot\mathbf{x}_i^T\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^T\mathbf{y}_j\cdot\mathbf{y}_j^T\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^T\mathbf{x}_i\\ &\le \left(\left\|\mathbf{x}_i^T\widehat{\mathbf{U}}\right\|_2\left\|\widehat{\mathbf{\Sigma}}\right\|_2\left\|\widehat{\mathbf{V}}^T\mathbf{y}_j\right\|_2\right)^2\\ &= \sigma_1^2\left\|\widehat{\mathbf{V}}^T\mathbf{y}_j\right\|_2^2\cdot\left\|\widehat{\mathbf{U}}^T\mathbf{x}_i\right\|_2^2\\ &\le \bar{r}^2\mu^2\sigma_1^2,\end{aligned} \tag{8}$$

where $\widehat{\mathbf{U}}\widehat{\mathbf{\Sigma}}\widehat{\mathbf{V}}^T$ is the SVD of $\widetilde{\mathbf{W}}$, $\sigma_1 = \|\widetilde{\mathbf{W}}\|_2$, and the last inequality follows from Assumption 1. Finally, for the variance of the random variables we have

$$\begin{aligned}\frac{1}{m}\sum_{ij}\mathrm{Var}(Z_{ij}) &\le \frac{1}{m}\sum_{ij}\mathbb{E}\left[Z_{ij}^2\right]\\ &\overset{(a)}{\le} \frac{1}{m}B_z\sum_{ij}\mathbb{E}\left[Z_{ij}\right]\\ &\le \bar{r}^2\mu^2\sigma_1^2\|\widetilde{\mathbf{W}}\|_F^2, \end{aligned} \tag{9}$$

where inequality (a) is due to the fact that Z_{ij}'s are nonnegative random variables. Using Bernstein's inequality we get the following.

$$\begin{aligned}&\mathbb{P}\Big(\big|\|\mathcal{A}(\widetilde{\mathbf{W}})\|_2^2 - \|\widetilde{\mathbf{W}}\|_F^2\big| > \epsilon\Big)\\ &\quad\le 2\exp\Big(-\frac{m\epsilon^2/2}{\bar{r}^2\mu^2\sigma_1^2\|\widetilde{\mathbf{W}}\|_F^2 + \bar{r}^2\mu^2\sigma_1^2\epsilon/3}\Big).\end{aligned} \tag{10}$$

Set $\epsilon = \delta\|\mathbf{W}\|_F^2$. We have

$$\begin{aligned}&\mathbb{P}\Big(\big|\|\mathcal{A}(\widetilde{\mathbf{W}})\|_2^2 - \|\widetilde{\mathbf{W}}\|_F^2\big| > \delta\|\widetilde{\mathbf{W}}\|_F^2\Big)\\ &\qquad\le 2\exp\Big(-\frac{m\delta^2\|\widetilde{\mathbf{W}}\|_F^2/2}{\bar{r}^2\mu^2\sigma_1^2(1+\delta/3)}\Big)\\ &\qquad\le 2\exp\Big(-\frac{m\delta^2/2}{\mu^2\bar{r}^2(1+\delta/3)}\Big).\end{aligned} \tag{11}$$

Set $\delta = \sqrt{\frac{4\mu^2\bar{r}^2\log(2/\rho)}{m}}$. If $m > 4\mu^2\bar{r}^2\log(2/\rho)$ we have $\delta < 1$. Therefore,

$$\mathbb{P}\Big(\big|\|\mathcal{A}(\widetilde{\mathbf{W}})\|_2^2 - \|\widetilde{\mathbf{W}}\|_F^2\big| > \delta\|\widetilde{\mathbf{W}}\|_F^2\Big) \le 2\exp\Big(-\frac{m\delta^2}{4\mu^2\bar{r}^2}\Big).$$

This concludes the proof of Lemma 1. □

Now we return to the proof of Theorem 3. The rest of the proof is based on Theorem 2.3 in [34]. We showed in Lemma 1 that for a given matrix of rank at most $2r$,

$$\mathbb{P}\Big(\big|\|\mathcal{A}(\widetilde{\mathbf{W}})\|_2^2 - \|\widetilde{\mathbf{W}}\|_F^2\big| > \delta\|\widetilde{\mathbf{W}}\|_F^2\Big) \le C\exp(cm),$$

for positive constants C and c. In order to extend the result such that a similar result holds for all rank-$2r$ (or less) matrices, we use the union bound for an ϵ-net [35] of the space of such matrices with unit Frobenius norm. For the set $\mathbb{S}_{2r}^d = \{\widetilde{\mathbf{W}} \in \mathbb{R}^{d\times d} : \ \mathrm{rank}(\widetilde{\mathbf{W}}) \le 2r, \left\|\widetilde{\mathbf{W}}\right\|_F = 1\}$, there exists an ϵ'-net $\bar{\mathbb{S}}_{2r}^d \subset \mathbb{S}_{2r}^d$ such that $|\bar{\mathbb{S}}_{2r}^d| \le (9/\epsilon')^{(2d+1)2r}$ [31,34]. It follows from 12 and the union bound that

$$\mathbb{P}\Big(\max_{\widetilde{\mathbf{W}}\in\bar{\mathbb{S}}_{2r}^d}\Big|\left\|\mathcal{A}\left(\widetilde{\mathbf{W}}\right)\right\|_2^2 - 1\Big| > \delta\Big) \le |\hat{S}_{2r}^d|C\exp(-cm).$$

Setting $\epsilon' = \delta/(4\sqrt{2})$ results in

$$\begin{aligned}&\mathbb{P}\Big(\max_{\widetilde{\mathbf{W}}\in\bar{\mathbb{S}}_{2r}^d}\Big|\left\|\mathcal{A}\left(\widetilde{\mathbf{W}}\right)\right\|_2^2 - 1\Big| > \delta\Big)\\ &\qquad\le C\exp\Big((2d+1)2r\log(36\sqrt{2}/\delta) - cm\Big)\\ &\qquad= C\exp\big(c'dr - cm\big)\\ &\qquad\le C\exp\left(-hm\right)\end{aligned} \tag{12}$$

where $c' = 6\log(36\sqrt{2}/\delta)$ and $h = c - c'/(K)$. We need $m > Kdr$ so that the last inequality above holds, and we need $K > c'/c$ so that h becomes positive. This means that $m > c'dr/c$. Plugging in the values for C, c, and c', we get that if with probability at least $1 - 2\exp\left(-hm\right)$,

$$\max_{\widetilde{\mathbf{W}}\in\bar{\mathbb{S}}_{2r}^d}\Big|\left\|\mathcal{A}(\widetilde{\mathbf{W}})\right\|_2^2 - 1\Big| \le \delta.$$

It follows from this bound that for all $\widetilde{\mathbf{W}}$ of rank at most $2r$ that with probability at least $1 - 2\exp\left(-hm\right)$ [34],

$$1 - 2\delta \le \Big\|\mathcal{A}\Big(\frac{\widetilde{\mathbf{W}}}{\|\widetilde{\mathbf{W}}\|_F^2}\Big)\Big\|_2^2 \le 1 + 2\delta.$$

Since $\mathcal{A}$ is a linear operator, for all $\widetilde{\mathbf{W}}$ with $\mathrm{rank}(\widetilde{\mathbf{W}}) \le 2r$,

$$(1-2\delta)\left\|\widetilde{\mathbf{W}}\right\|_F^2 \le \left\|\mathcal{A}(\widetilde{\mathbf{W}})\right\|_2^2 \le (1+2\delta)\left\|\widetilde{\mathbf{W}}\right\|_F^2.$$

This result means that $\mathcal{A}$ is $2r$-RIP with $\delta_{2r} = 2\delta$ when $m = O\left(\mu^2 d\bar{r}^2 r\log(36\sqrt{2}/\delta)/\delta^2\right)$.

□

Finally, we show that the sensing operator $\mathcal{T}$ is RIP on $(d_1 + d_2)\times(d_1+d_2)$ PSD matrices of rank at most $2r$. Any of these PSD matrices can be written in form of $\mathbf{N} = \begin{pmatrix}\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T & \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T\\ \widetilde{\mathbf{V}}^T\widetilde{\mathbf{U}} & \widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^T\end{pmatrix}$ where $\widetilde{\mathbf{U}} \in \mathbb{R}^{d_1\times 2r}$ and $\widetilde{\mathbf{V}} \in \mathbb{R}^{d_2\times 2r}$. We defined $\mathcal{T}$ such that $\mathcal{T}(\mathbf{N}) = 4\mathcal{A}(\widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T) + \left\|\widetilde{\mathbf{U}}\widetilde{\mathbf{U}}^T\right\|_F^2 + \left\|\widetilde{\mathbf{V}}\widetilde{\mathbf{V}}^T\right\|_F^2 - 2\left\|\widetilde{\mathbf{W}}\right\|_F^2$ where $\widetilde{\mathbf{W}} = \widetilde{\mathbf{U}}\widetilde{\mathbf{V}}^T$. Because $\|\mathbf{N}\|_F^2 = \left\|\mathbf{U}\mathbf{U}^T\right\|_F^2 + \left\|\mathbf{V}\mathbf{V}^T\right\|_F^2 + 2\left\|\widetilde{\mathbf{W}}\right\|_F^2$, if

$$(1-\delta)\left\|\widetilde{\mathbf{W}}\right\|_F^2 \le \left\|\mathcal{A}(\widetilde{\mathbf{W}})\right\|_2^2 - \left\|\widetilde{\mathbf{W}}\right\|_F^2 \le (1+\delta)\left\|\widetilde{\mathbf{W}}\right\|_F^2,$$

then

$$(1-2\delta)\left\|\mathbf{N}\right\|_F^2 \le \left\|\mathcal{T}(\mathbf{N})\right\|_2^2 - \left\|\mathbf{N}\right\|_F^2 \le (1+2\delta)\left\|\mathbf{N}\right\|_F^2.$$

Note that the deduction of the RIP of $\mathcal{T}$ from the RIP of $\mathcal{A}$ is thanks to the choice of the regularizer in (3).

4. CONCLUSION

In this paper, we discussed the geometric landscape of the inductive matrix completion (IMC) problem. The IMC model incorporates the side information in form of features of the row and column entities ($\mathbf{x}_i$'s and $\mathbf{y}_j$'s) and can be formulated as a low-rank matrix recovery problem where each observed entry of $\mathbf{M}^* = \mathbf{X}\mathbf{W}^*\mathbf{Y}$ is seen as a measurement of $\mathbf{W}^*$, that is $\mathbf{M}_{ij}^* = \mathbf{x}_i^T\mathbf{W}^*\mathbf{y}_j$. Motivated by the recovery guarantees of local search algorithms like AM for the factorized IMC problem [8], we study the optimization landscape of the factorized IMC problem. Using a framework developed by Ge et al. [22] for matrix sensing problems, we show that, given $O(\max\{r^2, \log^2 n\}rd)$ observations, for the (regularized) factorized IMC problem *i*) there are no poor local minima, *ii*) the global minima satisfy $\mathbf{U}\mathbf{V}^T = \mathbf{W}^*$, *iii*) The Hessian at the saddle point has at least one negative eigenvalue.

This result shows that the recovery guarantees of AM in the IMC problem is not merely due to the algorithm and the geometry of the problem plays an important role. In fact, any algorithm, such as SGD, that can efficiently escape saddle points and find a local minimum can be used for solving the factorized IMC problem.

5. REFERENCES

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

[2] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from noisy entries," *J. Mach. Learn. Res.*, vol. 11, pp. 2057–2078, 2010.

[3] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *Computer*, vol. 42, no. 8, pp. 30–37, Aug 2009.

[4] M. Signoretto, R. Van de Plas, B. De Moor, and J. A. K. Suykens, "Tensor versus matrix completion: A comparison with application to spectral data," *IEEE Signal Process. Lett.*, vol. 18, no. 7, pp. 403–406, 2011.

[5] B. Recht, "A simpler approach to matrix completion," *J. Mach. Learn. Res.*, vol. 12, no. Dec, pp. 3413–3430, 2011.

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

[7] M. Xu, R. Jin, and Z.-H. Zhou, "Speedup matrix completion with side information: Application to multi-label learning," in *Proc. Int. Conf. Advances in Neural Inform. Process. Syst. (NIPS)*, 2013, pp. 2301–2309.

[8] K. Zhong, P. Jain, and I. S. Dhillon, "Efficient matrix sensing using rank-1 gaussian measurements," in *Proc. Int. Conf. Algorithmic Learn. Theory (ALT)*, 2015, pp. 3–18.

[9] J. Lu, G. Liang, J. Sun, and J. Bi, "A sparse interactive model for matrix completion with side information," in *Proc. Int. Conf. Advances in Neural Inform. Process. Syst. (NIPS)*, 2016, pp. 4071–4079.

[10] K.-Y. Chiang, C.-J. Hsieh, and I. S. Dhillon, "Matrix completion with noisy side information," in *Proc. Int. Conf. Advances in Neural Inform. Process. Syst. (NIPS)*, 2015, pp. 3447–3455.

[11] A. Soni, T. Chevalier, and S. Jain, "Noisy inductive matrix completion under sparse factor models," in *IEEE Int. Symp. Inf. Theory (ISIT)*, June 2017, pp. 2990–2994.

[12] A. Eftekhari, D. Yang, and M. B. Wakin, "Weighted matrix completion and recovery with prior subspace information," *arXiv preprint arXiv:1612.01720*, 2016.

[13] N. Rao, H. Yu, P. Ravikumar, and I. S. Dhillon, "Collaborative filtering with graph information: Consistency and scalable methods," in *Proc. Int. Conf. Advances in Neural Inform. Process. Syst. (NIPS)*, 2015, pp. 2107–2115.

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

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

[16] R. H. Keshavan, A. Montanari, and S. Oh, "Matrix completion from a few entries," *IEEE Trans. Inf. Theory*, vol. 56, no. 6, pp. 2980–2998, June 2010.

[17] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?," *J. ACM*, vol. 58, no. 3, pp. 11:1–11:37, June 2011.

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

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

[20] K. Toh and S. Yun, "An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems," *Pacific J. Optimization*, vol. 6, pp. 615–640, 2010.

[21] R. Ge, F. Huang, C. Jin, and Y. Yuan, "Escaping from saddle points — online stochastic gradient for tensor decomposition," in *Proc. Conf. Learn. Theory (COLT)*, 2015, pp. 797–842.

[22] R. Ge, C. Jin, and Y. Zheng, "No spurious local minima in nonconvex low rank problems: A unified geometric analysis," *arXiv preprint arXiv:1704.00708*, 2017.

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

[24] S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Global optimality of local search for low rank matrix recovery," in *Proc. Int. Conf. Advances in Neural Inform. Process. Syst. (NIPS)*, 2016, pp. 3873–3881.

[25] D. Park, A. Kyrillidis, C. Caramanis, and S. Sanghavi, "Non-square matrix sensing without spurious local minima via the burer-monteiro approach," *arXiv preprint arXiv:1609.03240*, 2016.

[26] P. Netrapalli, P. Jain, and S. Sanghavi, "Phase retrieval using alternating minimization," in *Proc. Int. Conf. Advances in Neural Inform. Process. Syst. (NIPS)*, 2013, pp. 2796–2804.

[27] K. Kawaguchi, "Deep learning without poor local minima," in *Proc. Int. Conf. Advances in Neural Inform. Process. Syst. (NIPS)*, 2016, pp. 586–594.

[28] C. Yun, S. Sra, and A. Jadbabaie, "Global optimality conditions for deep neural networks," *arXiv preprint arXiv:1707.02444*, 2017.

[29] C. Jin, R. Ge, P. Netrapalli, S. M. Kakade, and M. I. Jordan, "How to escape saddle points efficiently," *arXiv preprint arXiv:1703.00887*, 2017.

[30] W. Cheney and D. R. Kincaid, *Linear Algebra: Theory and Applications*, Jones and Bartlett Publishers, Inc., 2008.

[31] Y. C. Eldar, D. Needell, and Y. Plan, "Unicity conditions for low-rank matrix recovery. arxiv preprint," *arXiv preprint arXiv:1103.5479*, 2011.

[32] Y. Dai and H. Li, "Rank minimization or nuclear-norm minimization: Are we solving the right problem?," in *Proc. Int. Conf. Digit. Image Comput.: Techn. Applicat. (DICTA)*, Nov 2014, pp. 1–8.

[33] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*, Oxford university press, 2013.

[34] E. J. Candès and Y. Plan, "Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements," *IEEE Trans. Inf. Theory*, vol. 57, no. 4, pp. 2342–2359, April 2011.

[35] W. A. Sutherland, *Introduction to metric and topological spaces*, Oxford University Press, 2009.