Inductive Matrix Completion: No Bad Local Minima and a Fast Algorithm

Pini Zilber [1] **Boaz Nadler** [1]

Abstract

The inductive matrix completion (IMC) problem is to recover a low rank matrix from few observed entries while incorporating prior knowledge about its row and column subspaces. In this work, we make three contributions to the IMC problem: (i) we prove that under suitable conditions, the IMC optimization landscape has no bad local minima; (ii) we derive a simple scheme with theoretical guarantees to estimate the rank of the unknown matrix; and (iii) we propose `GNIMC`, a simple Gauss-Newton based method to solve the IMC problem, analyze its runtime and derive for it strong recovery guarantees. The guarantees for `GNIMC` are sharper in several aspects than those available for other methods, including a quadratic convergence rate, fewer required observed entries and stability to errors or deviations from low-rank. Empirically, given entries observed uniformly at random, `GNIMC` recovers the underlying matrix substantially faster than several competing methods.

1. Introduction

In low rank matrix completion, a well-known problem that appears in various applications, the task is to recover a rank-r matrix $X^* \in \mathbb{R}^{n_1 \times n_2}$ given few of its entries, where $r \ll \min\{n_1, n_2\}$. In the problem of inductive matrix completion (IMC), beyond being low rank, X^* is assumed to have additional structure as follows: its columns belong to the range of a known matrix $A \in \mathbb{R}^{n_1 \times d_1}$ and its rows belong to the range of a known matrix $B \in \mathbb{R}^{n_2 \times d_2}$, where $r \leq d_1 \leq n_1$ and $r \leq d_2 \leq n_2$. Hence, X^* may be written as $X^* = AM^*B^\top$, and the task reduces to finding the smaller matrix $M^* \in \mathbb{R}^{d_1 \times d_2}$. In practice, the low rank and/or the additional structure assumptions may hold only approximately, and in addition, the observed entries may be corrupted by noise.

The side information matrices A, B may be viewed as feature representations. For example, in movies recommender systems, the task is to complete a matrix X^* of the ratings given by n_1 users to n_2 movies. The columns of A, B may correspond to viewers' demographic details (age, gender) and movies' properties (length, genre), respectively (Abernethy et al., 2009; Menon et al., 2011; Chen et al., 2012; Yao & Li, 2019). The underlying assumption in IMC is that uncovering the relations between the viewers and the movies in the feature space, as encoded in M^*, suffices to deduce the ratings $X^* = AM^*B^\top$. Other examples of IMC include multi-label learning (Xu et al., 2013; Si et al., 2016; Zhang et al., 2018), disease prediction from gene/miRNA/lncRNA data (Natarajan & Dhillon, 2014; Chen et al., 2018; Lu et al., 2018) and link prediction in networks (Menon & Elkan, 2011; Chiang et al., 2018).

If the side information matrices allow for a significant dimensionality reduction, namely $d \ll n$ where $d = \max\{d_1, d_2\}$ and $n = \max\{n_1, n_2\}$, recovering X^* is easier from both theoretical and computational perspectives. From the information limit aspect, the minimal number of observed entries required to complete a matrix of rank r with side information scales as $\mathcal{O}(dr)$, compared to $\mathcal{O}(nr)$ without side information. Similarly, the number of variables scales as d rather than as n, enabling more efficient computation and less memory. Finally, features also allow completion of rows and columns of X^* that do not have even a single observed entry. Unlike standard matrix completion which requires at least r observed entries in each row and column of X^*, in IMC the feature vector is sufficient to inductively predict the full corresponding row/column; hence the name 'Inductive Matrix Completion'.

Several IMC methods were devised in the past years. Perhaps the most popular ones are nuclear norm minimization (Xu et al., 2013; Lu et al., 2018) and alternating minimization (Jain & Dhillon, 2013; Natarajan & Dhillon, 2014; Zhong et al., 2015; Chen et al., 2018). A more recent method is multi-phase Procrustes flow (Zhang et al., 2018). While nuclear norm minimization enjoys strong recovery guarantees, it is computationally slow. Other methods are faster, but the number of observed entries for their recovery

[1] Faculty of Mathematics and Computer Science, Weizmann Institute of Science, Israel. Correspondence to: Pini Zilber <pini.zilber@weizmann.ac.il>, Boaz Nadler <boaz.nadler@weizmann.ac.il>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

Table 1. Recovery guarantees for the algorithms: `Maxide` (Xu et al., 2013), `AltMin` (Jain & Dhillon, 2013), `MPPF` (Zhang et al., 2018) and `GNIMC` (this work), for an $n \times n$ matrix X^* of rank r and condition number κ, and $d \times d$ side information matrices of incoherence μ, given a fixed target accuracy. Here $f(\kappa, \mu)$ is some function of κ and μ. For a more detailed comparison, see Section 5.3.

Algorithm	Sample complexity	Requires	Error	Time complexity
	$\lvert\Omega\rvert \gtrsim$ ...	incoherent X^*?	decay rate	$\sim \mathcal{O}(...)$
`Maxide`	$\mu^2 dr[1 + \log(d/r)] \log n$	yes	unspecified	unspecified
`AltMin`	$\kappa^2 \mu^4 d^2 r^3 \log n$	no	unspecified	unspecified
`MPPF`	$(\kappa r + d)\kappa^2 \mu^2 r^2 \log d \log n$	yes	linear	$f(\kappa, \mu) \cdot n^{3/2} d^2 r^3 \log d \log n$
`GNIMC` (ours)	$\mu^2 d^2 \log n$	no	quadratic	$\mu^2 d^3 r \log n$

guarantees to hold depends on the condition number of X^*.

In this work, we make three contributions to the IMC problem. First, by deriving an RIP (Restricted Isometry Property) guarantee for IMC, we prove that under certain conditions the optimization landscape of IMC is benign (Theorem 3.1). Compared to a similar result derived by Ghassemi et al. (2018), our guarantee requires significantly milder conditions, and in addition, addresses the vanilla IMC problem rather than a suitably regularized one.

Second, we propose a simple scheme to estimate the rank of X^* from its observed entries and the side information matrices A, B. We also provide a theoretical guarantee for the accuracy of the estimated rank (Theorem 4.1), which holds for either exactly or approximately low rank X^* and with noisy measurements.

Third, we propose a simple Gauss-Newton based method to solve the IMC problem, that is both fast and enjoys strong recovery guarantees. Our algorithm, named `GNIMC` (Gauss-Newton IMC), is an adaptation of the `GNMR` algorithm (Zilber & Nadler, 2022) to IMC. At each iteration, `GNIMC` solves a least squares problem; yet, its per-iteration complexity is of *the same order as gradient descent*. As a result, empirically, our tuning-free `GNIMC` implementation is 2 to 17 times faster than competing algorithms in a wide range of settings.

On the theoretical front, we prove that given a standard incoherence assumption on A, B and sufficiently many observed entries sampled uniformly at random, `GNIMC` recovers X^* at a *quadratic* convergence rate (Theorem 5.1). As far as we know, this is the only available quadratic convergence rate guarantee for any IMC algorithm. In addition, we prove that `GNIMC` is stable against small arbitrary additive error (Theorem 5.4), which may originate from (i) inaccurate measurements of X^*, (ii) inaccurate side information, and/or (iii) X^* being only approximately low rank.

Remarkably, our guarantees do not require X^* to be incoherent, and the required number of observations depends only on properties of A, B and not on those of X^*. Other guarantees have similar dependence on A, B, but in addition either depend on the condition number of X^* and/or require incoherence of X^*, see Table 1. Relaxing the incoherence assumption on X^* is important, since X^* is only partially observed and such an assumption cannot be verified. In contrast, the matrices A, B are known and their incoherence can be easily verified (see Definition 2.1 below).

Notation. The i-th largest singular value of a matrix X is denoted by $\sigma_i = \sigma_i(X)$. The condition number of a rank-r matrix is denoted by $\kappa = \sigma_1/\sigma_r$. The i-th standard basis vector is denoted by e_i, and the Euclidean norm of a vector x by $\|x\|$. The spectral norm of a matrix X is denoted by $\|X\|_2$, its Frobenius norm by $\|X\|_F$, its largest row norm by $\|X\|_{2,\infty} \equiv \max_i \|X^\top e_i\|$, its largest entry magnitude by $\|X\|_\infty \equiv \max_{i,j} |X_{ij}|$, and the set of its column vectors by $\text{col}(X)$. A matrix X is an isometry if $X^\top X = I$, where I is the identity matrix. Denote by $\mathcal{P}_{AB} : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ the projection operator onto the row and column spaces of A, B, respectively, such that $\mathcal{P}_{AB}(X) = AA^\top XBB^\top$ if A, B are isometries. Denote by $\mathcal{P}_\Omega : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{n_1 \times n_2}$ the sampling operator that projects a matrix in $\mathbb{R}^{n_1 \times n_2}$ onto an observation set $\Omega \subseteq [n_1] \times [n_2]$, such that $[\mathcal{P}_\Omega(X)]_{ij} = X_{ij}$ if $(i,j) \in \Omega$ and 0 otherwise. Denote by $\text{Vec}_\Omega(X) \in \mathbb{R}^{|\Omega|}$ the vector with the entries X_{ij} for all $(i,j) \in \Omega$. Finally, denote by $p = |\Omega|/(n_1 n_2)$ the sampling rate of Ω.

2. Problem Formulation

Let $X^* \in \mathbb{R}^{n_1 \times n_2}$ be a matrix of rank r. For now we assume r is known; in Section 4 we present a scheme to estimate r, and prove its accuracy. Assume $\Omega \subseteq [n_1] \times [n_2]$ is uniformly sampled and known, and let $Y = \mathcal{P}_\Omega(X^* + \mathcal{E})$ be the observed matrix where $\mathcal{E}$ is additive error. In the standard matrix completion problem, the goal is to solve

$$\min_X \|\mathcal{P}_\Omega(X) - Y\|_F^2 \quad \text{s.t. } \text{rank}(X) \leq r. \tag{MC}$$

In IMC, in addition to the observations Y we are given two side information matrices $A \in \mathbb{R}^{n_1 \times d_1}$ and $B \in \mathbb{R}^{n_2 \times d_2}$ with $r \leq d_i \leq n_i$ for $i = 1, 2$, such that

$$\text{col}(X^*) \subseteq \text{span col}(A), \quad \text{col}(X^{*\top}) \subseteq \text{span col}(B). \quad (1)$$

Note that w.l.o.g., we may assume that A and B are isometries, $A^\top A = I_{d_1}$ and $B^\top B = I_{d_2}$, as property (1) is invariant to orthonormalization of the columns of A and B. Standard matrix completion corresponds to $d_i = n_i$ with the trivial side information $A = I_{n_1}, B = I_{n_2}$. A common assumption in IMC is $d_i \ll n_i$, so that the side information is valuable. Note that beyond allowing for (potentially adversarial) inaccurate measurements, $\mathcal{E}$ may also capture violations of the low rank and the side information assumption (1), as we can view $X^* + \mathcal{E}$ as the true underlying matrix whose only first component, X^*, has exact low rank and satisfies (1).

Assumption (1) implies that $X^* = AM^*B^\top$ for some rank-r matrix $M^* \in \mathbb{R}^{d_1 \times d_2}$. The IMC problem thus reads

$$\min_M \|\mathcal{P}_\Omega(AMB^\top) - Y\|_F^2 \quad \text{s.t. } \text{rank}(M) \leq r. \quad \text{(IMC)}$$

Some works on IMC (Xu et al., 2013; Zhang et al., 2018) assume that both X^* and A, B are incoherent, namely have small incoherence, defined as follows (Candès & Recht, 2009; Keshavan et al., 2010).

Definition 2.1 (μ-incoherence)**.** A matrix $X \in \mathbb{R}^{n_1 \times n_2}$ of rank r is μ-incoherent if its Singular Value Decomposition (SVD), $U\Sigma V^\top$ with $U \in \mathbb{R}^{n_1 \times r}$ and $V \in \mathbb{R}^{n_2 \times r}$, satisfies

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

However, for IMC to be well-posed, X^* does not have to be incoherent, and it suffices for A, B to be incoherent (Jain & Dhillon, 2013). In case A and B are isometries, their incoherence assumption corresponds to bounded row norms, $\|A\|_{2,\infty} \leq \sqrt{\mu d_1 / n_1}$ and $\|B\|_{2,\infty} \leq \sqrt{\mu d_2 / n_2}$.

3. No Bad Local Minima Guarantee

In this section we present a novel characterization of the optimization landscape of IMC in the noiseless setting, $\mathcal{E} = 0$. Following the factorization approach to matrix recovery problems (see Chi et al. (2019) and references therein), we first incorporate the rank constraint into the objective by writing the unknown matrix as $M = UV^\top$ where $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$. Then, problem (IMC) reads

$$\min_{U,V} \|\mathcal{P}_\Omega(AUV^\top B^\top) - Y\|_F^2. \quad (2)$$

Clearly, any pair of matrices (U, V) which satisfy $UV^\top = M^*$ is a global minimizer of (2) with an objective value of zero. However, as (2) is non-convex, some of its first-order critical points, namely points at which the gradient vanishes, may be bad local minima. The next result, proven in Appendix C, states that if sufficiently many entries are observed, all critical points are either global minima or strict saddle points. At a strict saddle point the Hessian has at least one strictly negative eigenvalue, so that gradient descent does not get stuck there. Hence, under the conditions of Theorem 3.1, gradient descent recovers M^* from a random initialization.

Theorem 3.1. *Let $X^* \in \mathbb{R}^{n_1 \times n_2}$ be a rank r matrix which satisfies (1) with μ-incoherent matrices $A \in \mathbb{R}^{n_1 \times d_1}$ and $B \in \mathbb{R}^{n_2 \times d_2}$. Assume $\Omega \subseteq [n_1] \times [n_2]$ is uniformly sampled with $|\Omega| \gtrsim \mu^2 d_1 d_2 \log n$. Then w.p. at least $1 - 2n^{-2}$, any critical point (U, V) of problem (2) is either a global minimum with $UV^\top = M^*$, or a strict saddle point.*

To the best of our knowledge, Theorem 3.1 is the first guarantee for the geometry of vanilla IMC. A previous result by Ghassemi et al. (2018) addressed only a suitably balance-regularized version of (2). In addition, their guarantee requires $\mathcal{O}(\mu^2 r \max\{d_1, d_2\} \max\{d_1 d_2, \log^2 n\})$ observed entries with cubic scaling in d_1, d_2,[1] which is significantly larger than the quadratic scaling in our Theorem 3.1.

Theorem 3.1 guarantees exact recovery for a family of gradient-based algorithms beyond vanilla gradient descent. However, as illustrated in Section 6.1, solving the IMC problem can be done much faster than by gradient descent or variants thereof, e.g. by our proposed `GNIMC` method described in Section 5.

3.1. IMC as a Special Case of Matrix Sensing

Similar to Ghassemi et al. (2018), our proof of Theorem 3.1 is based on an RIP (Restricted Isometry Property) result we derive for IMC. The RIP result forms a connection between IMC and the matrix sensing (MS) problem, as follows. Recall that in IMC, the goal is to recover $M^* \in \mathbb{R}^{d_1 \times d_2}$ from the observations $Y = \mathcal{P}_\Omega(AM^*B^\top + \mathcal{E})$. In MS, we observe a set of linear measurements $b \equiv \mathcal{A}(M^*) + \xi$ where $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ is a sensing operator and $\xi \in \mathbb{R}^m$ is additive error. Assuming a known or estimated rank r of M^*, the goal is to solve

$$\min_M \|\mathcal{A}(M) - b\|^2 \quad \text{s.t. } \text{rank}(M) \leq r. \quad \text{(MS)}$$

Problem (IMC) is in the form of (MS) with the operator

$$\mathcal{A}(M) = \text{Vec}_\Omega(AMB^\top)/\sqrt{p} \quad (3)$$

and the error vector $\xi = \text{Vec}_\Omega(\mathcal{E})/\sqrt{p}$. However, unlike IMC, in MS the operator $\mathcal{A}$ is assumed to satisfy a suitable RIP, defined as follows (Candes, 2008; Recht et al., 2010).

[1] The cited complexity is in our notation. The notation in Ghassemi et al. (2018) is slightly different from ours; see Appendix F for more details.

Definition 3.2. A linear map $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ satisfies a k-RIP with a constant $\delta \in [0, 1)$, if for all matrices $M \in \mathbb{R}^{d_1 \times d_2}$ of rank at most k,

$$(1 - \delta)\|M\|_F^2 \leq \|\mathcal{A}(M)\|^2 \leq (1 + \delta)\|M\|_F^2. \quad (4)$$

The following theorem, proven in Appendix A, states that if A, B are incoherent and $|\Omega|$ is sufficiently large, w.h.p. the IMC sensing operator (3) satisfies the RIP. This observation creates a bridge between IMC and MS: for a given MS method, its RIP-based theoretical guarantees can be directly transferred to IMC.

Theorem 3.3. *Let $A \in \mathbb{R}^{n_1 \times d_1}$, $B \in \mathbb{R}^{n_2 \times d_2}$ be two isometry matrices such that $\|A\|_{2,\infty} \leq \sqrt{\mu d_1 / n_1}$ and $\|B\|_{2,\infty} \leq \sqrt{\mu d_2 / n_2}$. Let $\delta \in [0, 1)$, and assume $\Omega \subseteq [n_1] \times [n_2]$ is uniformly sampled with $|\Omega| \equiv m \geq (8/\delta^2)\mu^2 d_1 d_2 \log n$. Then, w.p. at least $1 - 2n^{-2}$, the sensing operator $\mathcal{A}$ defined in (3) satisfies an RIP (4) with $k = \min\{d_1, d_2\}$ and with the constant δ.*

A similar result was derived by Ghassemi et al. (2018). Theorem 3.3 improves upon it both in terms of the required conditions and in terms of the RIP guarantee. First, as in their landscape guarantee, Ghassemi et al. (2018) require cubic scaling with d_1, d_2 rather than quadratic as in our result. Moreover, their sample complexity includes an additional factor of $r \log(1/\delta)$ (see Appendix F). Second, they proved only a $\min\{2r, d_1, d_2\}$-RIP, whereas Theorem 3.3 guarantees that $\mathcal{A}$ satisfies the RIP with the *maximal* possible rank $\min\{d_1, d_2\}$. In particular, this allows us to employ a recent result due to Li et al. (2020) to prove Theorem 3.1 for vanilla IMC, as detailed in Appendix C.

The technical reason behind our sharper results is that instead of applying the Bernstein matrix inequality to a fixed *matrix* and then proving an ϵ-net union bound for all matrices, we apply it to a cleverly designed *operator* which directly guarantees the result for all matrices; see Lemma A.1.

4. Rank Estimation Scheme

The factorization approach (2) requires knowing r in advance, although in practice it is often unknown. In this section we propose a simple scheme to estimate the underlying rank, and provide a theoretical guarantee for it. Importantly, our scheme does not assume X^* is exactly low rank, but rather the existence of a sufficiently large spectral gap between its r-th and $(r+1)$-th singular values.

Let $\hat{X} = \mathcal{P}_{AB}(Y)/p = AA^\top YBB^\top/p$ where Y is the observed matrix and $p \equiv |\Omega|/(n_1 n_2)$, and denote its singular values by $\hat{\sigma}_i$. Our estimator for the rank of X^* is

$$\hat{r} = \arg\max_i \, g_i(\hat{X}), \quad g_i(\hat{X}) = \frac{\hat{\sigma}_i}{\hat{\sigma}_{i+1} + D \cdot \hat{\sigma}_1 \sqrt{i}}, \quad (5)$$

for some constant $D \in [0, 1)$. In our simulations we set $D = (\sqrt{d_1 d_2}/|\Omega|)^{1/2}$. The function g_i measures the i-th spectral gap, with the second term in the denominator added for robustness of the estimate. For $D = 0$, g_i is simply the ratio between two consecutive singular values. A similar estimator was proposed by Keshavan and Oh (2009) for standard matrix completion. The main difference in our estimator is the incorporation of the side information matrices A, B. We present the following theoretical guarantee for our estimator, proven in Appendix B. Note that using the side information matrices A, B allows us to reduce the sample complexity from $\mathcal{O}(n)$, as necessary in standard matrix completion (Keshavan & Oh, 2009), to only $\mathcal{O}(\log(n))$.

Theorem 4.1. *There exists a sufficiently small constant c such that the following holds w.p. at least $1 - 2n^{-2}$. Let $X^* \in \mathbb{R}^{n_1 \times n_2}$ be a matrix which satisfies (1) with μ-incoherent A, B. Assume X^* is approximately rank r, in the sense that for all $i \neq r$, $g_r(X^*) > \min\{(11/10)g_i(X^*), 1/10\}$. Denote $\delta = \min_i\{\sigma_{i+1}(X^*) + D\sigma_1(X^*)\sqrt{i}\}$, and assume $\Omega \subseteq [n_1] \times [n_2]$ is uniformly sampled with $|\Omega| \geq 8\mu^2 d_1 d_2 \log(n)\|X\|_F^2/(c\delta)^2$. Further assume bounded error $\epsilon \equiv \|\mathcal{P}_{AB}\mathcal{P}_\Omega(\mathcal{E})\|_F/p \leq c\delta$. Then $\hat{r} = r$.*

To the best of our knowledge, Theorem 4.1 is the first guarantee in the literature for rank estimation in IMC. We remark that with a suitably modified δ, our guarantee holds for other choices of g_i as well (including $g_i = \sigma_i/\sigma_{i+1}$, corresponding to $D = 0$). An empirical demonstration of our scheme appears in Section 6.2.

5. GNIMC Algorithm

In this section, we describe `GNIMC`, an adaptation of the `GNMR` algorithm (Zilber & Nadler, 2022) to IMC, and present recovery guarantees for it.

5.1. Description of GNIMC

Consider the factorized objective (2). Given an estimate (U, V), the goal is to find an update $(\Delta U, \Delta V)$ such that $(U', V') = (U + \Delta U, V + \Delta V)$ minimizes (2). In terms of the variables $(\Delta U, \Delta V)$, problem (2) reads

$$\min_{\Delta U, \Delta V} \|\mathcal{P}_\Omega(AUV^\top B^\top + AU\Delta V^\top B^\top + A\Delta UV^\top B^\top + A\Delta U\Delta V^\top B^\top) - Y\|_F^2,$$

which is nonconvex due to the mixed term $\Delta U \Delta V^\top$. The Gauss-Newton approach is to neglect this term. This yields the key iterative step of `GNIMC`, which is solving the following sub-problem:

$$\min_{\Delta U, \Delta V} \|\mathcal{P}_\Omega(AUV^\top B^\top + AU\Delta V^\top B^\top + A\Delta UV^\top B^\top) - Y\|_F^2. \quad (6)$$

```
Algorithm 1 GNIMC
input sampling operator $\mathcal{P}_\Omega$, observed matrix $Y$, side information matrices $(A, B)$, maximal number of iterations $T$, initialization $(U_0, V_0)$
output rank-$r$ (approximate) solution to $\mathcal{P}_\Omega(\hat{X}) = Y$
for $t = 0, \ldots, T-1$ do
    set $\binom{U_{t+1}}{V_{t+1}} = \binom{U_t}{V_t} + \binom{\Delta U_{t+1}}{\Delta V_{t+1}}$, where $\binom{\Delta U_{t+1}}{\Delta V_{t+1}}$ is the minimal norm solution of
    $\arg\min_{\Delta U, \Delta V} \|\mathcal{P}_\Omega[A(U_t V_t^\top + U_t \Delta V^\top + \Delta U V_t^\top)B^\top] - Y\|_F^2$
end for
return $\hat{X} = A U_T V_T^\top B^\top$
```

Problem (6) is a linear least squares problem. Note, however, that it has an infinite number of solutions: for example, if $(\Delta U, \Delta V)$ is a solution, so is $(\Delta U + UR, \Delta V - VR^\top)$ for any $R \in \mathbb{R}^{r \times r}$. We choose the solution with minimal norm $\|\Delta U\|_F^2 + \|\Delta V\|_F^2$, see Algorithm 1. In practice, this solution can be computed using the standard LSQR algorithm (Paige & Saunders, 1982).

In general, the computational complexity of solving problem (6) scales with the condition number κ of X^*. To decouple the runtime of GNIMC from κ, we use the QR decompositions of U and V as was similarly done for alternating minimization by Jain et al. (2013). In Appendix D we describe the full procedure, and prove it is analytically equivalent to (6). Remarkably, despite the fact that GNIMC performs a non-local update at each iteration, its resulting per-iteration complexity is as low as a single gradient descent step, as proven in Appendix D.3. For a discussion on parallel and distributed computing considerations, see Appendix D.4.

GNIMC requires an initial guess (U_0, V_0). A suitable initialization procedure for our theoretical guarantees is discussed in Proposition 5.3. In practice, GNIMC works well also from a random initialization.

The proposed GNIMC algorithm is extremely simple, as it merely solves a least squares problem in each iteration. In contrast to several previous methods, it requires no parameter estimation such as the minimal and maximal singular values of X^*, or tuning of hyperparameters such as regularization coefficients. Altogether, this makes GNIMC easy to implement and use. Furthermore, GNIMC enjoys strong recovery guarantees and fast runtimes, as described below.

5.2. Recovery Guarantees for GNIMC

We first analyze the noiseless case, $\mathcal{E} = 0$. The following theorem, proven in Appendix C, states that starting from a sufficiently accurate initialization with small imbalance $\|U^\top U - V^\top V\|_F$, GNIMC exactly recovers the matrix at a quadratic rate. In fact, the balance condition can be eliminated by adding a single SVD step as discussed below.

Theorem 5.1. *There exists a constant $c > 1$ such that the following holds w.p. at least $1 - 2n^{-2}$. Let $X^* \in \mathbb{R}^{n_1 \times n_2}$ be a rank-r matrix which satisfies (1) with μ-incoherent side matrices $A \in \mathbb{R}^{n_1 \times d_1}$ and $B \in \mathbb{R}^{n_2 \times d_2}$. Denote $\gamma = c/(2\sigma_r^*)$ where $\sigma_r^* = \sigma_r(X^*)$. Assume $\Omega \subseteq [n_1] \times [n_2]$ is uniformly sampled with*

$$|\Omega| \geq 32\mu^2 d_1 d_2 \log n. \tag{7}$$

Then, for any initial iterate (U_0, V_0) that satisfies

$$\|AU_0V_0^\top B^\top - X^*\|_F \leq \frac{\sigma_r^*}{c}, \tag{8a}$$

$$\|U_0^\top U_0 - V_0^\top V_0\|_F \leq \frac{\sigma_r^*}{2c}, \tag{8b}$$

the estimates $X_t = AU_tV_t^\top B^\top$ of Algorithm 1 satisfy

$$\|X_{t+1} - X^*\|_F \leq \gamma \cdot \|X_t - X^*\|_F^2, \quad \forall t = 0, 1, \ldots. \tag{9}$$

Note that by assumption (8a), $\gamma \cdot \|X_0 - X^*\|_F \leq 1/2$. Hence, (9) implies that GNIMC achieves exact recovery, since $X_t \to X^*$ as $t \to \infty$. The computational complexity of GNIMC is provided in the following proposition, proven in Appendix D.

Proposition 5.2. *Under the conditions of Theorem 5.1, the time complexity of GNIMC (Algorithm 1) until recovery with a fixed accuracy (w.h.p.) is $\mathcal{O}(\mu^2(d_1 + d_2)d_1d_2 r \log n)$.*

To meet the initialization conditions of Theorem 5.1, we need to find a rank-r matrix M which satisfies $\|AMB^\top - X^*\| \leq \sigma_r^*/c$. By taking its SVD $M = U\Sigma V^\top$, we obtain that $(U\Sigma^{\frac{1}{2}}, V\Sigma^{\frac{1}{2}})$ satisfies conditions (8a-8b). Such a matrix M can be computed in polynomial time using the initialization procedure suggested by Tu et al. (2016) for matrix sensing. Starting from $M_0 = 0$, it iteratively performs a gradient descent step and projects the result into the rank-r manifold. Its adaptation to IMC reads

$$M_{\tau+1} = \mathcal{P}_r\left[M_\tau - A^\top(\mathcal{P}_\Omega(AM_\tau B^\top)/p - Y)B\right] \tag{10}$$

where $\mathcal{P}_r(M)$ is the rank-r truncated SVD of M. The following proposition, proven in Appendix E, states that $\mathcal{O}(\log(r\kappa))$ iterations suffice to meet the initialization conditions of Theorem 5.1 under a slightly larger sample size requirement.

Proposition 5.3 (Initialization guarantee)**.** *Let X^*, A, B be as in Theorem 5.1. Assume Ω is uniformly sampled with $|\Omega| \geq 50\mu^2 d_1 d_2 \log n$. Let M_τ be the result after $\tau \geq 5\log(c\sqrt{r}\kappa)$ iterations of (10), and denote its SVD by $U\Sigma V$. Then w.p. $1 - 2n^{-2}$, $\binom{U_0}{V_0} = \binom{U\Sigma^{\frac{1}{2}}}{V\Sigma^{\frac{1}{2}}}$ satisfies the initialization conditions (8a)-(8b) of Theorem 5.1.*

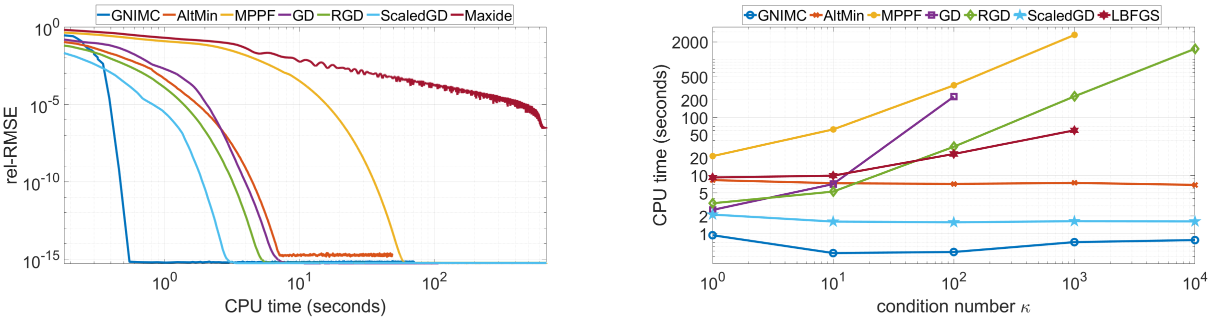

Figure 1. Left panel: `rel-RMSE` (17) as a function of CPU runtime for several IMC algorithms. Here X^* has a condition number $\kappa = 10$. Right panel: CPU runtime till successful recovery as a function of κ, where each point corresponds to the median of 50 independent realizations. In both panels $X^* \in \mathbb{R}^{1000\times 1000}$, $A, B \in \mathbb{R}^{20\times 20}$, $r = 10$ and oversampling ratio $\rho = 1.5$.

We conclude this subsection with a guarantee for `GNIMC` in the noisy setting. Suppose we observe $Y = \mathcal{P}_\Omega(X^* + \mathcal{E})$ where $\mathcal{E}$ is arbitrary additive error. To cope with the error, we slightly modify Algorithm 1, and add the following balancing step at the start of each iteration: calculate the SVD $\bar{U}\Sigma\bar{V}^\top$ of the current estimate $U_t V_t^\top$, and update

$$U_t \leftarrow \bar{U}\Sigma^{\frac{1}{2}}, \quad V_t \leftarrow \bar{V}\Sigma^{\frac{1}{2}}, \tag{11}$$

so that (U_t, V_t) are perfectly balanced with $U_t^\top U_t = V_t^\top V_t$. The following result holds for the modified algorithm.

Theorem 5.4. *Let X^*, A, B, Ω and c be defined as in Theorem 5.1, and suppose the error is bounded as*

$$\epsilon \equiv \tfrac{1}{\sqrt{p}} \|\mathcal{P}_\Omega(\mathcal{E})\|_F \le \frac{\sigma_r^*}{9c}. \tag{12}$$

Then for any initial iterate (U_0, V_0) that satisfies (8a), the estimates $X_t = AU_tV_t^\top B^\top$ of Algorithm 1 with the balancing step (11) satisfy

$$\|X_t - X^*\|_F \le \frac{\sigma_r^*}{4^{2^t-1}c} + 6\epsilon \xrightarrow{t\to\infty} 6\epsilon. \tag{13}$$

In the absence of errors, $\epsilon = 0$, this result reduces to the exact recovery guarantee with quadratic rate of Theorem 5.1.

5.3. Comparison to Prior Art

Here we describe recovery guarantees for three other algorithms. We compare them only to Theorem 5.1, as none of these works derived a stability to error result analogous to our Theorem 5.4. A summary appears in Table 1. In the following, let $n = \max\{n_1, n_2\}$ and $d = \max\{d_1, d_2\}$. For works which require incoherence condition on several matrices, we use for simplicity the same incoherence coefficient μ. All guarantees are w.p. at least $1 - \mathcal{O}(1/n)$.

Nuclear norm minimization (Maxide) (Xu et al., 2013). If (i) both X^* and A, B are μ-incoherent, (ii) $\|LR^\top\|_\infty \le \mu r/(n_1 n_2)$ where $L\Sigma R$ is the SVD of X^*, (iii) $d_1 d_2 + r^2 \ge 8[1 + \log_2(d/r)](d_1 + d_2)r$, and (iv)

$$|\Omega| \gtrsim \mu^2 r d[1 + \log(d/r)] \log n, \tag{14}$$

then `Maxide` exactly recovers X^*.

Alternating minimization (Jain & Dhillon, 2013). If A, B are μ-incoherent and

$$|\Omega| \gtrsim \kappa^2 \mu^4 r^3 d_1 d_2 \log n \log(1/\epsilon), \tag{15}$$

then `AltMin` recovers X^* up to error ϵ in spectral norm at a linear rate with a constant contraction factor.

Multi-phase Procrustes flow (Zhang et al., 2018). If both X^* and A, B are μ-incoherent and

$$|\Omega| \gtrsim \max\{\kappa r, d\}\kappa^2 \mu^2 r^2 \log d \log n, \tag{16}$$

then `MPPF` recovers X^* at a linear rate with a contraction factor smaller than $1 - \mathcal{O}(1/(r\kappa))$.[2] This guarantee implies a required number of iterations which may scale linearly with κ, as is indeed empirically demonstrated in Figure 1(right).

Notably, in terms of the dimension parameters n, d, r, the sample complexity for `Maxide` (14) is order optimal up to logarithmic factors. However, their guarantee requires additional assumptions, including incoherence of X^*. Also, from a practical point of view, `Maxide` is computationally slow and not easily scalable to large matrices (see Figure 1(left)). In contrast, `GNIMC` is computationally much faster and does not require X^* to be incoherent, a relaxation which can be important in practice as discussed in the introduction. Furthermore, our sample complexity requirement (7) is the only one independent of the condition number without requiring incoherent X^*. Compared to the other factorization-based methods, our sample complexity is strictly better than that of `AltMin`, and

[2] When the estimation error decreases below $\mathcal{O}(1/(\mu d))$, the contraction factor is improved to $1 - \mathcal{O}(1/\kappa)$.

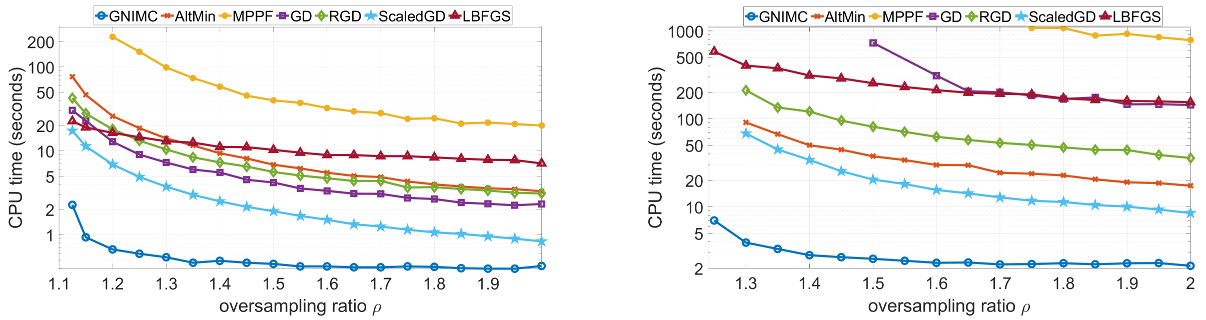

Figure 2. CPU runtime till successful recovery as a function of the oversampling ratio for several IMC algorithms. Left panel: $n_1 = n_2 = 1000$, $d_1 = d_2 = 20$ and $r = 10$. Right panel: $n_1 = 20000$, $n_2 = 1000$, $d_1 = 100$, $d_2 = 50$ and $r = 5$. In both panels $\kappa = 10$. Each point corresponds to the median of 50 independent realizations.

better than MPPF if $\min\{d_1, d_2\} \lesssim \kappa^2 r^2 \log d$. Since $\min\{d_1, d_2\} \leq r^2$ is a practical setting (see e.g. (Natarajan & Dhillon, 2014, Section 4.4) and (Zhang et al., 2018, Sections 6.1-6.2)), our complexity requirement is often smaller than that of MPPF even for well-conditioned matrices. In fact, if $\min\{d_1, d_2\} \leq 54r$, then our guarantee is the sharpest one, as condition (iii) of Maxide is violated. In addition, to the best of our knowledge, GNIMC is the only method with a quadratic convergence rate guarantee. Finally, its contraction factor is constant, and in particular independent of the rank r and the condition number κ.

We conclude this subsection with a computational complexity comparison. Among the above works, only the computational complexity of MPPF was analyzed, and it is given by $\mathcal{O}(f(\kappa, \mu) \cdot n^{3/2} d^2 r^3 \log d \log n)$ where $f(\kappa, \mu)$ is some function of κ and μ which was left unspecified by Zhang et al. (2018). The dependence on the large dimension factor $n^{3/2}$ implies that MPPF does not exploit the available side information in terms of computation time. Our complexity guarantee, Proposition 5.2, is fundamentally better. In particular, it depends on n only logarithmically, and is independent of the condition number κ. This independence is demonstrated empirically in Figure 1(right).

6. Simulation Results

In the following subsection we compare the performance of GNIMC to several other algorithms. Then, in the next subsection, we exemplify our rank estimation scheme.

6.1. Comparison Between Algorithms

We compare the performance of GNIMC to the following IMC algorithms, all implemented in MATLAB.[3] AltMin (Jain & Dhillon, 2013): our implementation of alternating minimization including the QR decomposition for reduced runtime; Maxide (Xu et al., 2013): nuclear norm minimization as implemented by the authors;[4] MPPF (Zhang et al., 2018): multi-phase Procrustes flow as implemented by the authors;[5] GD, RGD: our implementations of vanilla gradient descent (GD) and a variant regularized by an imbalance factor $\|U^\top U - V^\top V\|_F$ (RGD); ScaledGD (Tong et al., 2021): a preconditioned variant of gradient descent;[6] and L-BFGS: limited-memory quasi-Newton BFGS algorithm, as implemented in MATLAB R2022a. Details on initialization, early stopping criteria and a tuning scheme for the hyperparameters of Maxide, MPPF, RGD, ScaledGD and L-BFGS appear in Appendix G. GNIMC and AltMin require no tuning.

In each simulation we construct $U \in \mathbb{R}^{d_1 \times r}$, $V \in \mathbb{R}^{d_2 \times r}$, $A \in \mathbb{R}^{n_1 \times d_1}$ and $B \in \mathbb{R}^{n_2 \times d_2}$ with entries i.i.d. from the standard normal distribution, and orthonormalize their columns. We then set $X^* = AUDV^\top B^\top$ where $D \in \mathbb{R}^{r \times r}$ is diagonal with entries linearly interpolated between 1 and κ. A similar scheme was used by Zhang et al. (2018), with a key difference that we explicitly control the condition number of X^* to study how it affects the performance of the various methods. Next, we sample Ω of a given size $|\Omega|$ from the uniform distribution over $[n_1] \times [n_2]$. Since A and B are known, the $n_1 \times n_2$ matrix X^* has only $(d_1 + d_2 - r)r$ degrees of freedom. Denote the oversampling ratio by $\rho = \frac{|\Omega|}{(d_1+d_2-r)r}$. As ρ is closer to the information limit value of 1, the more challenging the problem becomes. Notably, our simulations cover a broad range of settings, including much fewer observed entries and higher condition numbers than

[3] MATLAB and Python code implementations of GNIMC, AltMin, GD and RGD are available at github.com/pizilber/IMC.

[4] www.lamda.nju.edu.cn/code_Maxide.ashx

[5] github.com/xiaozhanguva/Inductive-MC

[6] github.com/Titan-Tong/ScaledGD. We adapted the algorithm, originally designed for matrix completion, to the IMC problem. In addition, we implemented computations with sparse matrices to enhance its performance.

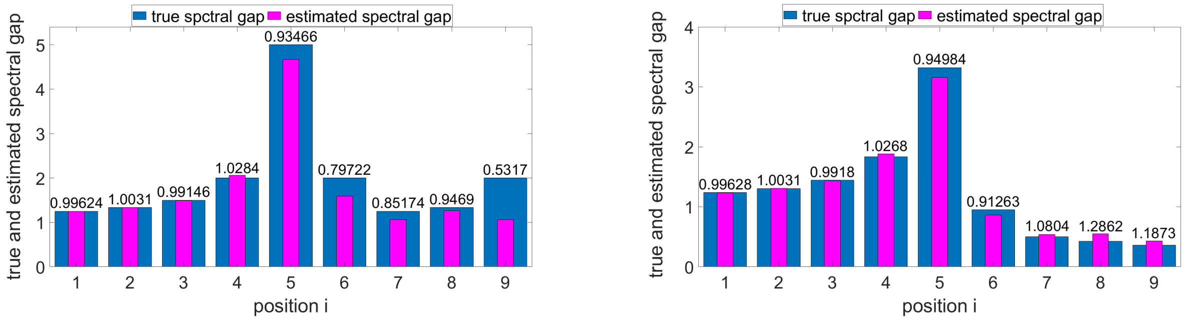

Figure 3. The estimated spectral gaps $\hat{g}_i$ (inner magenta) compared to the true ones g_i (outer blue) as defined in (5), for $X^* \in \mathbb{R}^{30000\times 10000}$ of approximate rank $r = 5$ with singular values $[5, 4, 3, 2, 1, 0.2, 0.1, 0.08, 0.06, 0.03]$, side information $d_1 = 30$, $d_2 = 20$, and sampling rate $p = 0.1\%$. The numbers above the bars indicate the ratio $\hat{g}_i/g_i$. Left panel: $D = 0$. Right panel: $D = (\sqrt{d_1 d_2}/|\Omega|)^{1/2} \approx 0.009$.

previous studies (Xu et al., 2013; Zhang et al., 2018).

We measure the quality of an estimate $\hat{X}$ by its relative RMSE,

$$\texttt{rel-RMSE} = \frac{\|X^* - \hat{X}\|_F}{\|X^*\|_F}. \tag{17}$$

First, we explore the convergence rate of the various algorithms, by comparing their relative RMSE as a function of runtime, in the setting $n_1 = n_2 = 1000$, $d_1 = d_2 = 20$, $r = \kappa = 10$ and $\rho = 1.5$ (sampling rate $p = 0.045\%$). Representative results of a single instance of the simulation, illustrating the behavior of the algorithms near convergence, are depicted in Figure 1(left). As shown in the figure, `GNIMC` converges much faster than the competing algorithms due to its quadratic convergence rate.

Next, we examine how the runtime of each algorithm is affected by the number of observations and by the condition number. The runtime is defined as the CPU time required for the algorithm to (i) converge, namely satisfy one of the stopping criteria (detailed in Appendix G), and (ii) achieve $\texttt{rel-RMSE} \leq 10^{-4}$. If the runtime exceeds 20 minutes without convergence, the run is stopped.

Figures 1(right) and 2(left) show the median recovery time on a log scale as a function of the condition number and of the oversampling ratio, respectively, in the same setting as above. Figure 2(right) corresponds to a larger matrix with $n_1 = 20000$, $n_2 = 1000$, $d_1 = 100$, $d_2 = 50$, $r = 5$ and $\kappa = 10$. Evidently, under a broad range of conditions, `GNIMC` is faster than the competing methods, in some cases by an order of magnitude. In general, the advantage of `GNIMC` with respect to the competing methods is more significant at low oversampling ratios. Finally, `GNIMC` outperforms the competing methods also in terms of the required number of iterations, see Figure 4 in Appendix H.1.

Remarkably, the runtime of `GNIMC`, `AltMin` and `ScaledGD` shows almost no sensitivity to the condition number, as illustrated in Figure 1(right). For `GNIMC`, this empirical observation is in agreement with Proposition 5.2, which states that the computational complexity of `GNIMC` does not depend on the condition number. In contrast, the runtime of the non-preconditioned gradient descent methods increases approximately linearly with the condition number.

Additional simulation results, including demonstration of the stability of `GNIMC` to noise, appear in Appendix H.

6.2. Demonstration of the Rank Estimation Scheme

In this subsection we demonstrate the accuracy of our proposed rank estimation scheme (5). Figure 3 compares the estimated singular gaps $\hat{g}_i$ with the true ones g_i for a matrix of approximate rank $r = 5$ and only $p = 0.1\%$ observed entries. We tested two values of D: $D = 0$ and $D = (\sqrt{d_1 d_2}/|\Omega|)^{1/2}$. The qualitative behavior depicted in the figure did not change in 50 independent realizations of the simulation. In particular, the estimated rank $\hat{r} = \max_i \hat{g}_i$ was always 5 for both values of D.

The figure also demonstrates the trade-off in the choice of the value of D: for larger D, $\hat{g}_i$ is a more accurate estimate of g_i, but g_i distorts the exact singular gaps $\sigma_i^*/\sigma_{i+1}^*$, especially at their tail (large values of i). Hence, in general, nonzero D is suitable in case the rank of X^* is expected to be relatively low compared to d_1, d_2.

7. Summary and Discussion

In this work, we presented three contributions to the IMC problem: benign optimization landscape guarantee; provable rank estimation scheme; and a simple Gauss-Newton based method, `GNIMC`, to solve the IMC problem. We derived recovery guarantees for `GNIMC`, and showed empirically that it is faster than several competing algorithms. A key theoretical contribution is a proof that under relatively mild conditions, IMC satisfies an RIP, similar to the matrix sensing problem.

Interestingly, in our simulations GNIMC recovers the matrix significantly faster than first-order methods, including a very recent one due to Tong et al. (2021). A possible explanation is that GNIMC makes large non-local updates, thus requires fewer iterations to converge; yet, the time complexity of each iteration is similar to a single local gradient descent step, leading to a shorter runtime. This raises the following intriguing questions: are there other non-convex problems for which non-local methods are faster than first-order ones? In particular, can these ideas be extended to faster training of deep neural networks?

Another possible future research is extending and analyzing our method under generalized frameworks of IMC. Interesting examples include recovering an unknown low rank X^* which lies in some known linear subspace instead of property (1) (Jawanpuria & Mishra, 2018) and non-linear IMC (Zhong et al., 2019). Another setting with practical importance is observations corrupted by outliers, as was extensively studied in other matrix recovery problems (Candès et al., 2011; Dutta et al., 2019).

8. Acknowledgements

The research of P.Z. was partially supported by a fellowship for data science from the Israeli Council for Higher Education (CHE). B.N. is the incumbent of the William Petschek Professorial Chair of Mathematics. Part of this research was funded by NIH grant R01GM135928. We thank Yoel Shkolnisky and Yariv Aizenbud for helpful discussions. We also thank the anonymous reviewers for their feedback.

References

Abernethy, J., Bach, F., Evgeniou, T., and Vert, J.-P. A new approach to collaborative filtering: Operator estimation with spectral regularization. *Journal of Machine Learning Research*, 10(Mar):803–826, 2009.

Balay et al. Petsc 2.0 users manual. Technical report, 1996.

Bi, Y., Zhang, H., and Lavaei, J. Local and global linear convergence of general low-rank matrix recovery problems. *arXiv preprint arXiv:2104.13348*, 2021.

Candes, E. J. The restricted isometry property and its implications for compressed sensing. *Comptes rendus mathematique*, 346(9-10):589–592, 2008.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3): 1–37, 2011.

Chen, T., Zhang, W., Lu, Q., Chen, K., Zheng, Z., and Yu, Y. Svdfeature: a toolkit for feature-based collaborative filtering. *The Journal of Machine Learning Research*, 13 (1):3619–3622, 2012.

Chen, X., Wang, L., Qu, J., Guan, N.-N., and Li, J.-Q. Predicting miRNA–disease association based on inductive matrix completion. *Bioinformatics*, 34(24):4256–4265, 2018.

Chi, Y., Lu, Y. M., and Chen, Y. Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

Chiang, K.-Y., Dhillon, I. S., and Hsieh, C.-J. Using side information to reliably learn low-rank matrices from missing and corrupted observations. *The Journal of Machine Learning Research*, 19(1):3005–3039, 2018.

Dutta, A., Hanzely, F., and Richtárik, P. A nonconvex projection method for robust pca. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pp. 1468–1476, 2019.

Dutta, A., Liang, J., and Li, X. A fast and adaptive svd-free algorithm for general weighted low-rank recovery. *arXiv preprint arXiv:2101.00749*, 2021.

Ghassemi, M., Sarwate, A., and Goela, N. Global optimality in inductive matrix completion. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2226–2230. IEEE, 2018.

Golub, G. and Pereyra, V. Separable nonlinear least squares: the variable projection method and its applications. *Inverse problems*, 19(2):R1, 2003.

Hayami, K. Convergence of the conjugate gradient method on singular systems. *arXiv preprint arXiv:1809.00793*, 2018.

Hubbard, C. and Hegde, C. Parallel computing heuristics for low-rank matrix completion. In *2017 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, pp. 764–768. IEEE, 2017.

Jain, P. and Dhillon, I. S. Provable inductive matrix completion. *arXiv preprint arXiv:1306.0626*, 2013.

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pp. 665–674. ACM, 2013.

Jawanpuria, P. and Mishra, B. A unified framework for structured low-rank matrix learning. In *International Conference on Machine Learning*, pp. 2254–2263. PMLR, 2018.

Keshavan, R. H. and Oh, S. A gradient descent algorithm on the grassman manifold for matrix completion. *arXiv preprint arXiv:0910.5260*, 2009.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from a few entries. *IEEE transactions on Information Theory*, 56(6):2980–2998, 2010.

Li, Q., Zhu, Z., and Tang, G. Geometry of factored nuclear norm regularization. *arXiv preprint arXiv:1704.01265*, 2017.

Li, S., Li, Q., Zhu, Z., Tang, G., and Wakin, M. B. The global geometry of centralized and distributed low-rank matrix recovery without regularization. *IEEE Signal Processing Letters*, 27:1400–1404, 2020.

Lu, C., Yang, M., Luo, F., Wu, F.-X., Li, M., Pan, Y., Li, Y., and Wang, J. Prediction of lncRNA–disease associations based on inductive matrix completion. *Bioinformatics*, 34(19):3357–3364, 2018.

Menon, A. K. and Elkan, C. Link prediction via matrix factorization. In *Proceedings of the 2011th European Conference on Machine Learning and Knowledge Discovery in Databases-Volume Part II*, pp. 437–452, 2011.

Menon, A. K., Chitrapura, K.-P., Garg, S., Agarwal, D., and Kota, N. Response prediction using collaborative filtering with hierarchies and side-information. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 141–149, 2011.

Natarajan, N. and Dhillon, I. S. Inductive matrix completion for predicting gene–disease associations. *Bioinformatics*, 30(12):i60–i68, 2014.

Paige, C. C. and Saunders, M. A. LSQR: An algorithm for sparse linear equations and sparse least squares. *ACM Transactions on Mathematical Software (TOMS)*, 8(1): 43–71, 1982.

Recht, B. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Si, S., Chiang, K.-Y., Hsieh, C.-J., Rao, N., and Dhillon, I. S. Goal-directed inductive matrix completion. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1165–1174, 2016.

Tong, T., Ma, C., and Chi, Y. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12 (4):389–434, 2012.

Tu, S., Boczar, R., Simchowitz, M., Soltanolkotabi, M., and Recht, B. Low-rank solutions of linear matrix equations via procrustes flow. In *International Conference on Machine Learning*, pp. 964–973. PMLR, 2016.

Xu, M., Jin, R., and Zhou, Z.-H. Speedup matrix completion with side information: Application to multi-label learning. In *Advances in neural information processing systems*, pp. 2301–2309, 2013.

Yao, K.-L. and Li, W.-J. Collaborative self-attention for recommender systems. *arXiv preprint arXiv:1905.13133*, 2019.

Zhang, X., Du, S., and Gu, Q. Fast and sample efficient inductive matrix completion via multi-phase procrustes flow. In *International Conference on Machine Learning*, pp. 5756–5765. PMLR, 2018.

Zhong, K., Jain, P., and Dhillon, I. S. Efficient matrix sensing using rank-1 gaussian measurements. In *International conference on algorithmic learning theory*, pp. 3–18. Springer, 2015.

Zhong, K., Song, Z., Jain, P., and Dhillon, I. S. Provable nonlinear inductive matrix completion. *Advances in Neural Information Processing Systems*, 32:11439–11449, 2019.

Zhu, Z., Li, Q., Tang, G., and Wakin, M. B. Global optimality in low-rank matrix optimization. *IEEE Transactions on Signal Processing*, 66(13):3614–3628, 2018.

Zilber, P. and Nadler, B. GNMR: A provable one-line algorithm for low rank matrix recovery. *arXiv preprint arXiv:2106.12933*, to appear in *SIAM Journal on Mathematics of Data Science*, 2022.

Additional notation. In the following appendices, the Frobenius inner product between two matrices is denoted by $\langle X, Y\rangle = \mathrm{Tr}(Y^\top X)$, where Tr denotes the matrix trace. The adjoint of an operator $\mathcal{P}$ is denoted by $\mathcal{P}^*$. The spectral norm of an operator $\mathcal{P}$ that acts on matrices is defined as $\|\mathcal{P}\| = \max_X \|\mathcal{P}(X)\|_F/\|X\|_F$.

A. Proof of Theorem 3.3 (RIP for IMC)

In the following subsection we state and prove a novel RIP guarantee that is key to the connection between IMC and matrix sensing. Then, in the next subsection, we use this result to prove Theorem 3.3.

A.1. An Auxiliary Lemma

To present our RIP result in the context of IMC, recall the definition of the linear operator $\mathcal{P}_{AB} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ which projects a matrix X into the row and column spaces of the isometry matrices A and B, respectively,

$$\mathcal{P}_{AB}(X) = AA^\top XBB^\top. \tag{18}$$

Note that since $\mathcal{P}_{AB}$ is a projection operator, $\|\mathcal{P}_{AB}\| = 1$.

Lemma A.1. *Let $A \in \mathbb{R}^{n_1\times d_1}$ and $B \in \mathbb{R}^{n_2\times d_2}$ be two isometry matrices such that $\|A\|_{2,\infty} \leq \sqrt{\mu d_1/n_1}$ and $\|B\|_{2,\infty} \leq \sqrt{\mu d_2/n_2}$. Let $\delta \in [0, 1)$, and assume $\Omega \subseteq [n_1] \times [n_2]$ is uniformly sampled with $|\Omega| \equiv n_1n_2p \geq (8/\delta^2)\mu^2 d_1d_2 \log n$ where $n = \max\{n_1, n_2\}$. Then w.p. at least $1 - 2n^{-2}$,*

$$\|\tfrac{1}{p}\mathcal{P}_{AB}\mathcal{P}_\Omega\mathcal{P}_{AB} - \mathcal{P}_{AB}\| \leq \delta. \tag{19}$$

Note that (19) is a bound on the norm of the operator $\frac{1}{p}\mathcal{P}_{AB}\mathcal{P}_\Omega\mathcal{P}_{AB} - \mathcal{P}_{AB}$, which acts on matrices. The numerical factor 8 in the bound on the sample complexity $|\Omega|$ of Lemma A.1 can be replaced by any other scalar β strictly greater than $8/3$, resulting in a modified probability guarantee $1 - 2n^{1-3\beta/8}$. We remark that $8/3$ is strict for our proof technique, which builds upon Recht's work (Recht, 2011). A lower value of β is possible by a more careful analysis, see the discussion after Proposition 5 in Recht (2011).

The proof of Lemma A.1 uses the following matrix Bernstein inequality (Tropp, 2012, Theorem 1.6).

Lemma A.2. *Consider a finite set $\{Z_k\}$ of independent, random matrices with dimensions $n_1 \times n_2$. Assume that each random matrix satisfies*

$$\mathbb{E}[Z_k] = 0 \quad \textit{and} \quad \|Z_k\|_2 \leq R \quad \textit{almost surely.} \tag{20}$$

Define $\sigma^2 = \max\{\|\sum_k \mathbb{E}[Z_kZ_k^\top]\|_2, \|\sum_k \mathbb{E}[Z_k^\top Z_k]\|_2\}$. Then, for all $t \geq 0$,

$$\mathbb{P}\left[\left\|\sum_k Z_k\right\|_2 \geq t\right] \leq (n_1 + n_2)\exp\left(-\frac{t^2/2}{\sigma^2 + Rt/3}\right). \tag{21}$$

Proof of Lemma A.1. The lemma assumes that Ω is uniformly sampled from the set of all collections of $m \equiv n_1n_2p$ entries of $[n_1] \times [n_2]$. Following Recht (2011), in the following proof we assume instead a different probabilistic model: sampling with replacement. Let $\Omega' = \{(i_k, j_k)\}_{k=1}^m$ be a collection of m elements, each i.i.d. from the uniform distribution over $[n_1] \times [n_2]$. Define also the corresponding operator

$$\mathcal{R}_{\Omega'}(X) = \sum_{k=1}^m \langle e_{i_k}e_{j_k}^\top, X\rangle\, e_{i_k}e_{j_k}^\top. \tag{22}$$

In contrast to $\mathcal{P}_\Omega$, the operator $\mathcal{R}_{\Omega'}$ is in general not a projection operator, since a pair of indices (i, j) may have been sampled more than once. In the following, rather than (19), we prove the following modified inequality that involves $\mathcal{R}_{\Omega'}$ in place of $\mathcal{P}_\Omega$,

$$\tfrac{1}{p}\|\mathcal{P}_{AB}\mathcal{R}_{\Omega'}\mathcal{P}_{AB} - p\mathcal{P}_{AB}\| \leq \delta. \tag{23}$$

This inequality implies the original (19), as $\mathcal{R}_{\Omega'}(X)$ reveals in general less information on X than $\mathcal{P}_\Omega(X)$ does due to possible duplicates in Ω'; see the proof of Proposition 3 in Recht (2011) for a rigorous formulation of this argument.

Since the elements of Ω' are uniformly sampled from the set $[n_1] \times [n_2]$ and $|\Omega'| = m \equiv pn_1n_2$, the expectation value of $\mathcal{R}_{\Omega'}$ over the random set Ω' is p times the identity operator. Hence,

$$\mathbb{E}[\mathcal{P}_{AB}\mathcal{R}_{\Omega'}\mathcal{P}_{AB}] = \mathcal{P}_{AB}\mathbb{E}[\mathcal{R}_{\Omega'}]\mathcal{P}_{AB} = p\mathcal{P}_{AB}^2 = p\mathcal{P}_{AB}, \tag{24}$$

where $\mathcal{P}_{AB}$ is defined in (18). We thus conclude that (23) is simply a concentration inequality, which we shall prove using Lemma A.2.

Let $X \in \mathbb{R}^{n_1\times n_2}$, and decompose it as $X = \sum_{i,j} \langle X, e_ie_j^\top \rangle \, e_ie_j^\top$. For future use, we define the linear operator $\mathcal{T}_{ij} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ as

$$\mathcal{T}_{ij}(X) = \langle X, \mathcal{P}_{AB}(e_ie_j^\top)\rangle \, \mathcal{P}_{AB}(e_ie_j^\top) = \langle \mathcal{P}_{AB}(X), e_ie_j^\top \rangle \, \mathcal{P}_{AB}(e_ie_j^\top), \tag{25}$$

and present some related equalities. By standard properties of the trace operator,

$$\mathcal{P}_{AB}\mathcal{R}_{\Omega'}\mathcal{P}_{AB} = \sum_{k=1}^m \mathcal{T}_{i_kj_k}. \tag{26}$$

Hence, taking the expectation over (i, j) uniformly sampled from $[n_1] \times [n_2]$ gives that

$$\mathbb{E}[\mathcal{T}_{ij}] = \frac{1}{m}\mathbb{E}\left[\sum_{k=1}^m \mathcal{T}_{i_kj_k}\right] = \frac{1}{pn_1n_2}\mathbb{E}[\mathcal{P}_{AB}\mathcal{R}_{\Omega'}\mathcal{P}_{AB}] = \frac{1}{n_1n_2}\mathcal{P}_{AB}. \tag{27}$$

In addition, by the definition (25) of $\mathcal{T}_{ij}$ and the fact that $\mathcal{P}_{AB}$ is a projection,

$$\mathcal{P}_{AB}\mathcal{T}_{ij} = \mathcal{T}_{ij}\mathcal{P}_{AB} = \mathcal{T}_{ij}. \tag{28}$$

Finally, by inserting (26) into inequality (23), we obtain that our goal is to bound $\| \sum_{k=1}^m \mathcal{T}_{i_kj_k} - p\mathcal{P}_{AB}\| = \| \sum_{k=1}^m \mathcal{D}_{i_kj_k}\|$, where the operator $\mathcal{D}_{ij} : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}^{n_1\times n_2}$ is given by

$$\mathcal{D}_{ij} = \mathcal{T}_{ij} - \frac{p}{m}\mathcal{P}_{AB} = \mathcal{T}_{ij} - \frac{1}{n_1n_2}\mathcal{P}_{AB}.$$

By (27), $\mathbb{E}[\mathcal{D}_{ij}] = 0$. To employ Lemma A.2 to the set $\{\mathcal{D}_{i_k,j_k}\}_{k=1}^m$, we first need to (i) find a scalar R such that $\|\mathcal{D}_{ij}\| \leq R$ almost surely, and (ii) bound $\max\{\| \sum_{k=1}^m \mathbb{E}[\mathcal{D}_{i_kj_k}\mathcal{D}_{i_kj_k}^*]\|, \| \sum_{k=1}^m \mathbb{E}[\mathcal{D}_{i_kj_k}^*\mathcal{D}_{i_kj_k}]\|\} = \| \sum_{k=1}^m \mathbb{E}[\mathcal{D}_{i_kj_k}^2]\|$, where the equality follows since $\mathcal{D}_{ij}$ is self-adjoint w.r.t. the Frobenius inner product.

We begin with bounding $\|\mathcal{D}_{ij}\| \equiv \max_X \|\mathcal{D}_{ij}(X)\|_F/\|X\|_F$. Recall that if X and Y are positive semidefinite matrices, then $\|X - Y\|_2 \leq \max\{\|X\|_2, \|Y\|_2\}$. Since any operator can be represented by a matrix, a similar result holds for operators with the spectral norm. As both $\mathcal{T}_{ij}$ and $\mathcal{P}_{AB}$ are positive semidefinite and $\mathcal{P}_{AB}$ is a projection, we have

$$\|\mathcal{D}_{ij}\| \leq \max\{\|\mathcal{T}_{ij}\|, \frac{1}{n_1n_2}\|\mathcal{P}_{AB}\|\} = \max\{\|\mathcal{T}_{ij}\|, \frac{1}{n_1n_2}\}. \tag{29}$$

Let us bound $\|\mathcal{T}_{ij}\|$. By the Cauchy-Schwarz inequality,

$$\|\mathcal{T}_{ij}(X)\| = | \langle X, \mathcal{P}_{AB}(e_ie_j^\top)\rangle | \cdot \|\mathcal{P}_{AB}(e_ie_j^\top)\|_F \leq \|\mathcal{P}_{AB}(e_ie_j^\top)\|_F^2\|X\|_F.$$

Inserting the definition of $\mathcal{P}_{AB}$ (18), the spectral norm of $\mathcal{T}_{ij}$ is bounded as

$$\begin{aligned}\|\mathcal{T}_{ij}\| &\leq \|\mathcal{P}_{AB}(e_ie_j^\top)\|_F^2 = \|AA^\top e_ie_j^\top BB^\top\|_F^2 \overset{(a)}{\leq} \|AA^\top e_i\|^2\|BB^\top e_j\|^2 \\ &\overset{(b)}{=} \|A^\top e_i\|^2\|B^\top e_j\|^2 \leq \|A\|_{2,\infty}^2\|B\|_{2,\infty}^2 \overset{(c)}{\leq} \frac{\mu^2 d_1d_2}{n_1n_2},\end{aligned} \tag{30}$$

where (a) follows from the Cauchy-Schwarz inequality, (b) from the isometry assumption, and (c) from the assumed bound on the row norms of A and B. Plugging (30) into (29) yields

$$\|\mathcal{D}_{ij}\| \leq \max\{\frac{\mu^2 d_1 d_2}{n_1 n_2}, \frac{1}{n_1 n_2}\} = \frac{\mu^2 d_1 d_2}{n_1 n_2} \equiv R, \tag{31}$$

where the equality follows since $\mu \geq 1$ by the definition of incoherence (Definition 2.1). Next, we bound $\| \sum_{k=1}^{m} \mathbb{E}[\mathcal{D}^2_{i_k j_k}]\|$. Combining (28), (27) and the fact that both $\mathcal{T}_{ij}^2$ and $\mathcal{P}_{AB}$ are positive semidefinite yields

$$\begin{aligned} \|\mathbb{E}[\mathcal{D}_{ij}^2]\| &= \|\mathbb{E}[\mathcal{T}_{ij}^2 - \frac{2}{n_1 n_2}\mathcal{T}_{ij} + \frac{1}{n_1^2 n_2^2}\mathcal{P}_{AB}]\| = \|\mathbb{E}[\mathcal{T}_{ij}^2] - \frac{1}{n_1^2 n_2^2}\mathcal{P}_{AB}\| \\ &\leq \max\{\|\mathbb{E}[\mathcal{T}_{ij}^2]\|, \frac{1}{n_1^2 n_2^2}\|\mathcal{P}_{AB}\|\} = \max\{\|\mathbb{E}[\mathcal{T}_{ij}^2]\|, \frac{1}{n_1^2 n_2^2}\}. \end{aligned}$$

Let us bound $\|\mathbb{E}[\mathcal{T}_{ij}^2]\|$. Since $\mathcal{T}_{ij}$ is positive semidefinite, we have $\mathcal{T}_{ij}^2 \preccurlyeq \|\mathcal{T}_{ij}\|\mathcal{T}_{ij}$. Thus $\mathbb{E}[\mathcal{T}_{ij}^2] \preccurlyeq \mathbb{E}[\|\mathcal{T}_{ij}\|\mathcal{T}_{ij}] \preccurlyeq \frac{\mu^2 d_1 d_2}{n_1 n_2}\mathbb{E}[\mathcal{T}_{ij}]$, where the last inequality follows from the deterministic bound (30). Together with (27) this implies

$$\|\mathbb{E}[\mathcal{T}_{ij}^2]\| \leq \frac{\mu^2 d_1 d_2}{n_1 n_2}\|\mathbb{E}[\mathcal{T}_{ij}]\| = \frac{\mu^2 d_1 d_2}{n_1^2 n_2^2}\|\mathcal{P}_{AB}\| = \frac{\mu^2 d_1 d_2}{n_1^2 n_2^2}.$$

We thus obtain the bound

$$\|\sum_{k=1}^{m} \mathbb{E}[\mathcal{D}^2_{i_k j_k}]\| = m \cdot \|\mathbb{E}[\mathcal{D}_{ij}^2]\| \leq m\frac{\mu^2 d_1 d_2}{n_1^2 n_2^2} = \frac{p\mu^2 d_1 d_2}{n_1 n_2} \equiv \sigma^2.$$

Plugging this together with the bound $\|\mathcal{D}_{ij}\| \leq R$ in (31) into Lemma A.2 yields

$$\mathbb{P}\left[\left\|\sum_{k=1}^{m}\mathcal{D}_{i_k j_k}\right\| > p\delta\right] \leq (n_1 + n_2)\exp\left(-\frac{p^2\delta^2/2}{\frac{p\mu^2 d_1 d_2}{n_1 n_2} + \frac{\mu^2 d_1 d_2}{n_1 n_2}p\delta/3}\right) \leq 2n\exp\left(-\frac{3\delta^2 m}{8\mu^2 d_1 d_2}\right).$$

Assuming that $m \geq (8/\delta^2)\mu^2 d_1 d_2 \log n$ gives

$$\mathbb{P}\left[\left\|\sum_{k=1}^{m}\mathcal{D}_{i_k j_k}\right\| > p\delta\right] \leq 2ne^{-3\log n} = 2n^{-2}.$$

This completes the proof of (23), and thus of (19). □

A.2. Proof of Theorem 3.3

Let $M \in \mathbb{R}^{d_1 \times d_2}$, and denote $X = AMB^\top$. By definition (3) of $\mathcal{A}$,

$$\frac{1}{p}\|\mathcal{P}_\Omega(X)\|_F^2 = \frac{1}{p}\|\mathcal{P}_\Omega(AMB^\top)\|_F^2 = \|\mathcal{A}(M)\|^2. \tag{32}$$

Next, observe that $\mathcal{P}_{AB}(X) = AA^\top AMB^\top BB^\top = AMB^\top = X$. Hence

$$\begin{aligned} \|\mathcal{P}_\Omega(X)\|_F^2 &= \langle \mathcal{P}_\Omega(X), \mathcal{P}_\Omega(X)\rangle = \langle X, \mathcal{P}_\Omega(X)\rangle = \langle X, pX\rangle + \langle X, \mathcal{P}_\Omega(X) - pX\rangle \\ &= p\|X\|_F^2 + \langle \mathcal{P}_{AB}(X), \mathcal{P}_\Omega\mathcal{P}_{AB}(X) - p\mathcal{P}_{AB}(X)\rangle \\ &= p\|X\|_F^2 + \langle X, \mathcal{P}_{AB}\mathcal{P}_\Omega\mathcal{P}_{AB}(X) - p\mathcal{P}_{AB}(X)\rangle\,. \end{aligned}$$

Applying the Cauchy-Schwarz inequality and (19) of Lemma A.1 yields

$$\begin{aligned} \left|\|\mathcal{P}_\Omega(X)\|_F^2 - p\|X\|_F^2\right| &= |\,\langle X, \mathcal{P}_{AB}\mathcal{P}_\Omega\mathcal{P}_{AB}(X) - p\mathcal{P}_{AB}(X)\rangle\,| \\ &\leq \|X\|_F\|\mathcal{P}_{AB}\mathcal{P}_\Omega\mathcal{P}_{AB}(X) - p\mathcal{P}_{AB}(X)\|_F \leq p\delta\|X\|_F^2. \end{aligned}$$

Hence

$$(1-\delta)\|X\|_F^2 \le \frac{1}{p}\|\mathcal{P}_\Omega(X)\|_F^2 \le (1+\delta)\|X\|_F^2. \tag{33}$$

Since A, B are isometries, $\|X\|_F = \|AMB^\top\|_F = \|M\|_F$. Plugging this together with (32) into (33) yields the RIP (4). □

In the following remark, we extend the connection between IMC and matrix sensing (MS) to another setting of the two problems, where the goal is to find the minimal rank matrix that agrees with the observations.

Remark A.3. An alternative setting of IMC, which does not assume a known rank but does assume noise-free observations, is to find a matrix with the lowest possible rank that is consistent with the data,

$$\min_M \ \text{rank}(M) \quad \text{s.t. } \mathcal{P}_\Omega(AMB^\top) = \mathcal{P}_\Omega(X^*). \tag{IMC*}$$

The analogous setting of MS is

$$\min_M \ \text{rank}(M) \quad \text{s.t. } \mathcal{A}(M) = \mathcal{A}(M^*). \tag{MS*}$$

With the sensing operator $\mathcal{A}$ defined in (3), (IMC*) is in the form of (MS*). Since this sensing operator satisfies the RIP under certain conditions as guaranteed by Theorem 3.3, the connection between IMC and MS holds in this setting as well.

B. Proof of Theorem 4.1 (Rank Estimation)

The proof of the theorem is based on the following lemma, which employs Lemma A.1 to bound the difference between the singular values of X^* and those of $\hat{X} = \mathcal{P}_{AB}(Y)/p$.

Lemma B.1. *Let $X^* \in \mathbb{R}^{n_1\times n_2}$ be a matrix which satisfies (1) with μ-incoherent matrices A, B. Let δ, ϵ and Ω be defined as in Theorem 4.1 with constant $c < 1/2$. Then w.p. at least $1-2n^{-2}$,*

$$|\hat{\sigma}_i - \sigma_i^*| \le 2c\delta, \quad \forall i, \tag{34}$$

where $\sigma_i^* = \sigma_i(X^*)$.

Proof. Since X^* satisfies the side information property (1), we have $\mathcal{P}_{AB}(X^*) = X^*$. Hence

$$\hat{X} = \frac{1}{p}\mathcal{P}_{AB}\mathcal{P}_\Omega(X^* + \mathcal{E}) = \frac{1}{p}\mathcal{P}_{AB}\mathcal{P}_\Omega\mathcal{P}_{AB}(X^*) + \frac{1}{p}\mathcal{P}_{AB}\mathcal{P}_\Omega(\mathcal{E}).$$

Using $\mathcal{P}_{AB}(X^*) = X^*$ again, we get

$$\hat{X} - X^* = \left(\frac{1}{p}\mathcal{P}_{AB}\mathcal{P}_\Omega\mathcal{P}_{AB} - \mathcal{P}_{AB}\right)(X^*) + \frac{1}{p}\mathcal{P}_{AB}\mathcal{P}_\Omega(\mathcal{E}).$$

Let $\delta' = c\delta/\|X^*\|_F$. By definition, $\delta \le \sigma_2^* + D\sigma_1^* < 2\sigma_1^*$. Hence $\delta' < 1$ for $c < 1/2$. Invoking Lemma A.1 with $|\Omega| \ge 8\mu^2 d_1 d_1 \log(n)\|X^*\|_F^2/(c\delta)^2 = 8\mu^2 d_1 d_1 \log(n)/\delta'^2$ and using the condition $\epsilon \le c\delta$ imply

$$\begin{aligned}\|\hat{X} - X^*\|_F &\le \left\|\left(\frac{1}{p}\mathcal{P}_{AB}\mathcal{P}_\Omega\mathcal{P}_{AB} - \mathcal{P}_{AB}\right)(X^*)\right\|_F + \frac{1}{p}\|\mathcal{P}_{AB}\mathcal{P}_\Omega(\mathcal{E})\|_F \le \delta'\|X^*\|_F + \epsilon \\ &\le 2c\delta. \end{aligned} \tag{35}$$

Hence also $\|\hat{X} - X^*\|_F \le 2c\delta$. Equation (34) of the lemma follows by Weyl's inequality. □

Proof of Theorem 4.1. Denote $\hat{g}_i = g_i(\hat{X})$ and $g_i^* = g_i(X^*)$. We need to show that $\arg\max_i \hat{g}_i = r$. Invoking Lemma B.1 implies

$$\hat{g}_r = \frac{\hat{\sigma}_r}{\hat{\sigma}_{r+1} + D\hat{\sigma}_1\sqrt{r}} \ge \frac{\sigma_r^* - 2c\delta}{\sigma_r^* + 2c\delta + D(\sigma_1^* + 2c\delta)\sqrt{r}}. \tag{36}$$

By the definition of δ, we have that $\delta \leq (1+D)\sigma_1^* < 2\sigma_1^*$ and also $\delta \leq \sigma_r^* + D\sigma_1^*\sqrt{r}$. Plugging this into (36) yields

$$\begin{aligned}\hat{g}_r &\geq \frac{\sigma_r^* - 2c\delta}{\sigma_r^* + 2c\delta + D(\sigma_1^* + 4c\sigma_1^*)\sqrt{r}} \geq \frac{\sigma_r^* - 2c(\sigma_r^* + D\sigma_1^*\sqrt{r})}{(1+2c)(\sigma_r^* + D\sigma_1^*\sqrt{r}) + 4cD\sigma_1^*\sqrt{r}} \\ &\geq \frac{\sigma_r^* - 2c(\sigma_r^* + D\sigma_1^*\sqrt{r})}{(1+6c)(\sigma_r^* + D\sigma_1^*\sqrt{r})} = \frac{1}{1+6c}g_r^* - \frac{2c}{1+6c}.\end{aligned}$$

Next, let $i \neq r$. Since $\delta \leq \sigma_{i+1}^* + D\sigma_1^*\sqrt{i}$, we similarly have

$$\begin{aligned}\hat{g}_i &= \frac{\hat{\sigma}_i}{\hat{\sigma}_{i+1} + D\hat{\sigma}_1\sqrt{i}} \leq \frac{\sigma_i^* + 2c\delta}{\sigma_{i+1}^* - 2c\delta + D(\sigma_1^* - 2c\delta)\sqrt{i}} \leq \frac{\sigma_i^* + 2c(\sigma_{i+1}^* + D\sigma_1^*\sqrt{i})}{(1-2c)(\sigma_{i+1}^* + D\sigma_1^*\sqrt{i}) - 4cD\sigma_{i+1}^*\sqrt{i}} \\ &\leq \frac{\sigma_i^* + 2c(\sigma_i^* + D\sigma_1^*\sqrt{i})}{(1-6c)(\sigma_{i+1}^* D\sigma_1^*\sqrt{i})} = \frac{1}{1-6c}g_i^* + \frac{2c}{1-6c}.\end{aligned}$$

By assumption, $g_r^* \geq \min\{(11/10)g_i^*, 1/10\}$. We thus obtain that $\hat{g}_r > \hat{g}_i$ for a sufficiently small constant c. Hence $\hat{r} = \arg\max_i \hat{g}_i = r$, as required. □

Remark B.2. One of the motivations for our proposed rank estimation scheme is that the rank is required as an input to most factorization-based matrix recovery algorithms, including GNIMC. Since our recovery guarantees for GNIMC, Theorems 5.1 and 5.4, are independent of the underlying matrix condition number, GNIMC provably (as well as empirically) recovers the matrix given an overestimated rank input. However, a tighter estimate of the true rank allows for a more efficient computation. An alternative to exact rank estimation is a rank continuation scheme, where one begins with an overestimated rank and adaptively adjusts it throughout the algorithm iterations (Dutta et al., 2021).

C. Proof of Theorems 3.1, 5.1 and 5.4

Our proof of Theorem 3.1 follows by combining Theorem 3.3 with a general result due to Li et al. (2020). In contrast to previous works (Li et al., 2017; Zhu et al., 2018; Bi et al., 2021), the result of Li et al. (2020) applies to a regularization-free objective. Consider the following general low rank optimization problem,

$$\min_{M \in \mathbb{R}^{d_1 \times d_2}} f(M), \quad \text{s.t. } \operatorname{rank}(M) \leq r. \tag{37}$$

By incorporating the rank constraint into the objective function, we obtain the factorized problem

$$\min_{U \in \mathbb{R}^{d_1 \times r}, V \in \mathbb{R}^{d_2 \times r}} g(U,V) \equiv f(UV^\top). \tag{38}$$

The following result provides a sufficient condition on $f(M)$ such that $g(U,V)$ has no bad local minima. The condition is on the bilinear form of the Hessian of $f(M)$, defined as $\nabla^2 f(M)[N,N] = \sum_{i,j,k,l} \frac{\partial^2 f(M)}{\partial M_{ij} \partial M_{kl}} N_{ij} N_{kl}$.

Lemma C.1. *Let α, β be two positive constants that satisfy $\beta/\alpha \leq 3/2$. Assume that f satisfies*

$$\alpha \|N\|_F^2 \leq \nabla^2 f(M)[N,N] \leq \beta \|N\|_F^2 \tag{39}$$

for all $M, N \in \mathbb{R}^{d_1 \times d_2}$. If $f(M)$ has a critical point M^* with rank$(M^*) \leq r$, then any critical point (U,V) of $g(U,V)$ in (38) is either a global minimum with $UV^\top = M^*$ or a strict saddle point.

To prove the lemma, we will use the following definition (Zhu et al., 2018).

Definition C.2. A function $f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ is (r_1, r_2)-restricted strongly convex and smooth if it satisfies (39) for any $M \in \mathbb{R}^{d_1 \times d_2}$ of rank at most r_1 and $N \in \mathbb{R}^{d_1 \times d_2}$ of rank at most r_2.

Proof of Lemma C.1. By Theorem III.1 in Li et al. (2020), the lemma follows if $f(X)$ satisfies the (r_1, r_2)-restricted strongly convex smoothness, where

$$r_1 = \min\{2r, d_1, d_2\} \quad \text{and} \quad r_2 = \min\{4r, d_1, d_2\}. \tag{40}$$

In fact, since Li et al. (2020) assume $r \ll \min\{d_1, d_2\}$ throughout their work, their Theorem III.1 is phrased with $r_1 = 2r$ and $r_2 = 4r$; however, it is straightforward to verify that in the general case, in which the rank of M, N is bounded by $\min\{d_1, d_2\}$, the theorem holds with r_1, r_2 as in (40). Our condition in the lemma is stronger, as it requires (39) to hold for all M, N, and thus implied by their Theorem III.1. □

Proof of Theorem 3.1. As discussed in the main text, the IMC problem can be written as a matrix sensing problem with the objective $f(M) = \|\mathcal{A}(M) - b\|^2$, where the sensing operator $\mathcal{A}$ is given in (3) and $b = \mathcal{A}(M^*) + \text{Vec}_\Omega(\mathcal{E})/\sqrt{p}$. Furthermore, by Theorem 3.3, for the assumed $|\Omega|$, the operator $\mathcal{A}$ satisfies a $\min\{d_1, d_2\}$-RIP (4) with a constant $\delta \leq 1/5$. Note that the $\min\{d_1, d_2\}$-RIP of $\mathcal{A} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2}$ in fact means that (4) holds for any $d_1 \times d_2$ matrix, since the rank of any such matrix is bounded by $\min\{d_1, d_2\}$.

Next, for any $M, N \in \mathbb{R}^{d_1 \times d_2}$, we have $\nabla f(M) = \mathcal{A}^*(\mathcal{A}(M) - b)$ and $\nabla^2 f(M)[N, N] = \|\mathcal{A}(N)\|^2$ (Zhu et al., 2018, Section C.1). Plugging the last equality into the RIP (4) of the sensing operator $\mathcal{A}$ yields

$$(1 - \delta)\|N\|_F^2 \leq \nabla^2 f(M)[N, N] \leq (1 + \delta)\|N\|_F^2.$$

Let $\alpha = 1 - \delta$ and $\beta = 1 + \delta$. Then f satisfies Equation (39) with the constants α, β. Further, since $\delta \leq 1/5$, we have $\beta/\alpha \leq 3/2$. The corollary thus follows by Lemma C.1. □

Finally, the proof of Theorems 5.1 and 5.4 is straightforward thanks to our Theorem 3.3.

Proof of Theorems 5.1 and 5.4. By Theorem 3.3, (IMC) is a special case of (MS) where the sensing operator $\mathcal{A}$ satisfies a rank $\min\{d_1, d_2\}$-RIP with a constant $\delta \leq 1/2$. Theorems 5.1 and 5.4 thus follow from the MS recovery guarantees for `GNMR` (Zilber & Nadler, 2022, Theorems 3.3-3.4). □

D. Computational Complexity Analysis

In Section 5.1 of the main text we briefly mentioned a way to use QR decompositions in order to efficiently find the minimal norm solution to the least squares problem (6). In the following subsection we describe the full procedure in detail. Then, in the next subsection, we prove Proposition 5.2 on the corresponding computational complexity. In both subsections we use the following simple result.

Lemma D.1. *Assume the conditions of Proposition 5.2. Then w.p. at least $1 - 2n^{-2}$, the factor matrices U_t, V_t of the iterates of `GNIMC` (Algorithm 1) have full column rank for all $t = 0, 1, ...$.*

Proof. We prove that if

$$\|AU_tV_t^\top B^\top - X^*\|_F < \sigma_r^*, \tag{41}$$

then U_t and V_t are full column rank. The lemma follows since (41) holds at $t = 0$ by assumption (8a) with $c > 1$, and at any $t > 0$ w.p. at least $1 - 2n^{-2}$ by the contraction principle (9).

By combining Weyl's inequality and (41),

$$|\sigma_r(AU_tV_t^\top B^\top) - \sigma_r^*| \leq \|AU_tV_t^\top B^\top - X^*\|_2 \leq \|AU_tV_t^\top B^\top - X^*\|_F < \sigma_r^*.$$

Since A and B are isometries, the above inequality implies that $|\sigma_r(U_tV_t^\top) - \sigma_r^*| < \sigma_r^*$. Hence

$$0 < \sigma_r(U_tV_t^\top) \leq \min\{\sigma_r(U_t)\|V_t\|_2, \sigma_r(V_t)\|U_t\|_2\},$$

which implies that both $\sigma_r(U_t)$ and $\sigma_r(V_t)$ are strictly positive, namely U_t, V_t have full column rank. □

D.1. A Computationally Efficient Way to Find The Minimal Norm Solution to (6)

At iteration t of `GNIMC` (Algorithm 1), our goal is to efficiently calculate the solution $(\Delta U_{t+1}, \Delta V_{t+1})$ to the rank deficient least squares problem (6) whose norm $\|\Delta U_{t+1}\|_F^2 + \|\Delta V_{t+1}\|_F^2$ is minimal. The least squares problem (6) at iteration t reads

$$\operatorname*{arg\,min}_{\Delta U, \Delta V} \|\mathcal{P}_\Omega[A(U_tV_t^\top + U_t\Delta V^\top + \Delta UV_t^\top)B^\top] - Y\|_F^2. \tag{42}$$

Denote the condition number of X^* by κ. If U_t, V_t are approximately balanced and their product $U_t V_t^\top$ is close to X^*, their condition number scales as $\sqrt{\kappa}$. Hence, the condition number of the least squares problem (namely, the condition number of the operator defined in (45) below) scales as $\sqrt{\kappa}$. As a result, directly solving (42) leads to a factor of $\sqrt{\kappa}$ in the computational complexity. In the following, we describe a procedure that gives the same solution to (42) but eliminates the dependency in $\sqrt{\kappa}$, as proven in the next subsection. The procedure consists of two phases. First, we efficiently compute a feasible solution to (42), not necessarily the minimal norm one. Second, we describe how, given a solution to (42), we can efficiently compute the one with minimal norm, $(\Delta U_{t+1}, \Delta V_{t+1})$. Algorithm 2 provides a sketch of this procedure.[7]

By Lemma D.1, the factor matrices of the current iterate U_t, V_t are full column rank. Let $Q_U R_U$ and $Q_V R_V$ be the QR decompositions of U_t and V_t, respectively, such that $Q_U \in \mathbb{R}^{d_1 \times r}$ and $Q_V \in \mathbb{R}^{d_2 \times r}$ are isometries, and $R_U, R_V \in \mathbb{R}^{r \times r}$ are invertible. Instead of (42), we solve the following modified least squares problem,

$$(\Delta U', \Delta V') = \operatorname*{arg\,min}_{\Delta U, \Delta V} \|\mathcal{P}_\Omega(A U_t V_t^\top B^\top + A Q_U \Delta V^\top B^\top + A \Delta U Q_V^\top B^\top) - Y\|_F^2. \tag{43}$$

Here, $(\Delta U', \Delta V')$ is any feasible solution to (43), not necessarily the minimal norm one. Next, let

$$\Delta U'' = \Delta U' (R_V^{-1})^\top \quad \text{and} \quad \Delta V'' = \Delta V' (R_U^{-1})^\top. \tag{44}$$

It is easy to verify that $(\Delta U'', \Delta V'')$ is a feasible solution to the original least squares problem (42). This concludes the first part of the procedure, which can be viewed as preconditioning: as we show below, (43) has a lower condition number than (42), and it hence faster to solve by iterative methods. The reason for the better conditioning is that Q_U, Q_V both have condition number one rather than $\sqrt{\kappa}$. The detailed computational complexity analysis is deferred to the next subsection.

Next, we describe how to transform a feasible solution, such as $(\Delta U'', \Delta V'')$, into the minimal norm one $(\Delta U_{t+1}, \Delta V_{t+1})$. To this end, we first express the least squares operator in terms of the sensing operator $\mathcal{A}$ defined in (3). In the matrix sensing formulation, the least squares problem (42) reads

$$\begin{aligned}
&\min_{(\Delta U, \Delta V)} \|\mathcal{P}_\Omega[A(U_t V_t^\top + U_t \Delta V^\top + \Delta U V_t^\top) B^\top] - Y\|_F \\
&= \min_{(\Delta U, \Delta V)} \|\text{Vec}_\Omega[A(U_t V_t + U_t \Delta V^\top + \Delta U V_t^\top) B^\top]/\sqrt{p} - \text{Vec}_\Omega(Y)/\sqrt{p}\| \\
&= \min_{(\Delta U, \Delta V)} \|\mathcal{A}(U_t V_t^\top + U_t \Delta V^\top + \Delta U V_t^\top) - b\| \\
&= \min_{(\Delta U, \Delta V)} \|\mathcal{A}(U_t \Delta V^\top + \Delta U V_t^\top) - b_t\|,
\end{aligned}$$

where $b = \text{Vec}_\Omega(Y)/\sqrt{p}$ and $b_t = b - \mathcal{A}(U_t V_t^\top)$. The least squares operator $\mathcal{L}_{(U_t, V_t)} : \mathbb{R}^{(d_1+d_2) \times r} \to \mathbb{R}^m$ is thus

$$\mathcal{L}_{(U_t, V_t)} \begin{pmatrix} U \\ V \end{pmatrix} = \mathcal{A}(U_t V^\top + U V_t^\top). \tag{45}$$

Let $\mathcal{K} = \ker \mathcal{L}_{(U_t, V_t)}$. According to the second part of Lemma 4.4 in Zilber and Nadler (2022), which holds due to our Theorem 3.3,

$$\begin{aligned}
\mathcal{K} &= \left\{ \left(\begin{smallmatrix} U \\ V \end{smallmatrix}\right) \in \mathbb{R}^{(d_1+d_2)r} \mid U^\top U_t = V_t^\top V \right\}^\perp \\
&= \left\{ \left(\begin{smallmatrix} U_t R \\ -V_t R^\top \end{smallmatrix}\right) \mid R \in \mathbb{R}^{r \times r} \right\} = \left\{ \left(\begin{smallmatrix} Q_U R \\ -Q_V R^\top \end{smallmatrix}\right) \mid R \in \mathbb{R}^{r \times r} \right\},
\end{aligned} \tag{46}$$

where the second equality follows by Eq. (26) in (Zilber & Nadler, 2022). Also, $\dim\{\mathcal{K}\} = r^2$ as Q_U, Q_V are isometries. By definition of the minimal norm solution $\left(\begin{smallmatrix} \Delta U_{t+1} \\ \Delta V_{t+1} \end{smallmatrix}\right)$, any other solution is of the form $\left(\begin{smallmatrix} \Delta U'' \\ \Delta V'' \end{smallmatrix}\right) = \left(\begin{smallmatrix} \Delta U_{t+1} \\ \Delta V_{t+1} \end{smallmatrix}\right) + \left(\begin{smallmatrix} K_U \\ K_V \end{smallmatrix}\right)$ where $\left(\begin{smallmatrix} \Delta U_{t+1} \\ \Delta V_{t+1} \end{smallmatrix}\right) \perp \mathcal{K}$ and $\left(\begin{smallmatrix} K_U \\ K_V \end{smallmatrix}\right) \in \mathcal{K}$. Hence, all we need to do is to subtract from $\left(\begin{smallmatrix} \Delta U'' \\ \Delta V'' \end{smallmatrix}\right)$ its component in $\mathcal{K}$. Denote the

[7] We remark that while the second phase works for any given feasible solution, in practice the feasible solution we find is also a minimal norm solution but of a different least squares problem. As this solution works well in practice, we did not implement the second phase in our simulations.

Algorithm 2 Efficient procedure to compute the minimal norm solution to (6)

```
input  sampling operator $\mathcal{P}_\Omega$, observed matrix $Y$, side information matrices $(A, B)$, current iterate $(U_t, V_t)$
output  the minimal norm solution to (6)
   {Phase I: compute a feasible solution to (6)}
1: compute $Q_U R_U$ and $Q_V R_V$, the QR decompositions of $U_t$ and $V_t$, respectively
2: compute $(\Delta U', \Delta V')$, any feasible solution to
       $$\operatorname*{arg\,min}_{(\Delta U, \Delta V)} \|\mathcal{P}_\Omega[A(U_t V_t^\top + Q_U \Delta V^\top + \Delta U Q_V^\top)B^\top] - Y\|_F^2$$
3: set $\Delta U'' = \Delta U'(R_V^{-1})^\top$, $\Delta V'' = \Delta V'(R_U^{-1})^\top$
   {Phase II: compute the minimal norm solution to (6)}
4: let $\mathcal{P}_\mathcal{K} : \mathbb{R}^{(d_1+d_2)\times r} \to R^{(d_1+d_2)\times r}$ be the projector onto $\mathcal{K}$, using its orthonormal basis given in (48)
5: set $\begin{pmatrix} \Delta U_{t+1} \\ \Delta V_{t+1} \end{pmatrix} = (\mathcal{I} - \mathcal{P}_\mathcal{K}) \begin{pmatrix} \Delta U'' \\ \Delta V'' \end{pmatrix}$
6: return $(\Delta U_{t+1}, \Delta V_{t+1})$
```

columns of Q_U, Q_V by u_i, v_i for $i \in [r]$, respectively, and let

$$K^{(ij)} = \frac{1}{\sqrt{2}} \begin{pmatrix} u_i e_j^\top \\ -v_j e_i^\top \end{pmatrix}, \quad \forall (i,j) \in [r] \times [r]. \tag{47}$$

Then the following set of r^2 matrices form an orthonormal basis for the kernel $\mathcal{K}$ of (46) under the Frobenius inner product $\langle C, D \rangle = \operatorname{Tr}(C^\top D)$:

$$\mathcal{K}_B = \left\{ K^{(ij)} \mid (i,j) \in [r] \times [r] \right\}. \tag{48}$$

Let $\mathcal{I}$ be the identity operator. By calculating the projector $\mathcal{P}_\mathcal{K}$ onto the span of $\mathcal{K}_B$, we obtain the minimal norm solution $\begin{pmatrix} \Delta U_{t+1} \\ \Delta V_{t+1} \end{pmatrix} = (\mathcal{I} - \mathcal{P}_\mathcal{K}) \begin{pmatrix} \Delta U'' \\ \Delta V'' \end{pmatrix}$.

The procedure described in this subsection is sketched in Algorithm 2.

D.2. Proof of Proposition 5.2

For the analysis of the computational complexity of GNIMC with the minimal norm solution computed via Algorithm 2, we first prove the following auxiliary lemma. Recall that the condition number of an operator $\mathcal{P} : \mathbb{R}^{(d_1+d_2)\times r} \to \mathbb{R}^m$ is defined as $\max_Z\{\|\mathcal{P}(Z)\|/\|Z\|_F\}/\min_Z\{\|\mathcal{P}(Z)\|/\|Z\|_F\}$.

Lemma D.2. *Let Ω, A, B be defined as in Proposition 5.2. Let $\mathcal{L}_{(Q_U,Q_V)}$ be the least squares operator of step 2 in Algorithm 2,*

$$\mathcal{L}_{(Q_U,Q_V)} \begin{pmatrix} U \\ V \end{pmatrix} = \mathcal{A}(Q_U V^\top + U Q_V^\top). \tag{49}$$

Denote its condition number by κ_L. Then

$$\kappa_L \leq \sqrt{6}. \tag{50}$$

We remark that the bound in (50) can be slightly improved (up to $\kappa_L \leq \sqrt{2}$) at the cost of increasing $|\Omega|$.

Proof of Lemma D.2. By combining assumption (7) and Theorem 3.3, the sensing operator $\mathcal{A}$ satisfies a $\min\{d_1, d_2\}$-RIP with a constant $\delta \leq 1/2$. Hence, as in the proof of Lemma 4.2 in Zilber and Nadler (2022), the minimal nonzero singular value of $\mathcal{L}_{(Q_U,Q_V)}$, $\sigma_{\min}(\mathcal{L}_{(Q_U,Q_V)})$, is bounded from below by $\sqrt{1-\delta}\min\{\sigma_r(Q_U), \sigma_r(Q_V)\}$. Next, by Lemma D.1,

U_t and V_t are of full column rank. Hence, Q_U and Q_V are isometries, and in particular $\sigma_r(Q_U) = \sigma_r(Q_V) = 1$. We thus obtain $\sigma_{\min}(\mathcal{L}_{(Q_U,Q_V)}) \geq \sqrt{1-\delta}$.

We similarly bound from above the maximal singular value, $\sigma_1(\mathcal{L}_{(Q_U,Q_V)})$. Let $U \in \mathbb{R}^{d_1 \times r}$, $V \in \mathbb{R}^{d_2 \times r}$. Then

$$\begin{aligned}\left\| \mathcal{L}_{(Q_U,Q_V)} \begin{pmatrix} U \\ V \end{pmatrix} \right\|^2 &\overset{(a)}{\leq} (1+\delta)\|UQ_V^\top + Q_U V^\top\|_F^2 \overset{(b)}{\leq} (1+\delta)(\|UQ_V^\top\|_F + \|Q_U V^\top\|_F)^2 \\ &\overset{(c)}{=} (1+\delta)(\|U\|_F^2 + 2\|U\|_F\|V\|_F + \|V\|_F^2) \overset{(d)}{\leq} 2(1+\delta)(\|U\|_F^2 + \|V\|_F^2),\end{aligned}$$

where (a) follows by the RIP of $\mathcal{A}$, (b) by the triangle inequality, (c) by the fact that Q_U, Q_V are isometries, and (d) by $ab \leq (a^2+b^2)/2$. Hence, the maximal singular value is bounded from above by $\sqrt{2(1+\delta)}$. The condition number of $\mathcal{L}_{(Q_U,Q_V)}$ is thus bounded as

$$\kappa_L \leq \sqrt{\frac{2(1+\delta)}{1-\delta}} \leq \sqrt{6}$$

where the second inequality follows since $\delta \leq 1/2$. □

We are now ready to prove Proposition 5.2.

Proof of Proposition 5.2. According to our quadratic convergence guarantee, the number of `GNIMC` iterations till recovery with a fixed accuracy is constant. Thus, up to a multiplicative constant, the complexity of `GNIMC` is the same as the complexity of a single iteration, which we shall now analyze according to its sketch in Algorithm 2.

The complexity of step 1, which consists of QR factorizations of $d \times r$ matrices, is $\mathcal{O}(dr^2)$ (Golub & Pereyra, 2003, Section 5.2.9). Step 2 is separately analyzed below. Step 3 is dominated by the calculation of the matrix product, which costs $\mathcal{O}(dr^2)$. Steps 4-5 are dominated by the calculation of the projection of a feasible solution $\left(\begin{smallmatrix}\Delta U'' \\ \Delta V''\end{smallmatrix}\right)$ onto the kernel $\mathcal{K}$ given in (46). To this end, we first construct a matrix $K \in \mathbb{R}^{(d_1+d_2)r \times r^2}$ whose columns are the vectorization of the elements of the orthonormal basis $\mathcal{K}_B$ given in (48). Then, to obtain the required projection, we calculate the product $KK^\top z$ where $z \equiv \text{Vec}\left(\begin{smallmatrix}\Delta U'' \\ \Delta V''\end{smallmatrix}\right) \in \mathbb{R}^{(d_1+d_2)r}$ is the vectorization of the feasible solution in hand. By first calculating $K^\top z$ and then $K(K^\top z)$ we obtain the complexity of $\mathcal{O}(dr^3)$.

Finally, we analyze the complexity of step 2. `GNIMC` solves the least squares problem using the standard LSQR algorithm (Paige & Saunders, 1982), which applies the conjugate gradient (CG) method to the normal equations. Each inner iteration of CG is dominated by the calculation of $AQ_U\Delta V^\top B^\top + A\Delta U Q_V^\top B^\top$ at the entries of Ω (Paige & Saunders, 1982, Section 7.7). To obtain a single entry of $AQ_U\Delta V^\top B^\top$, we calculate a single row of AQ_U, a single column of $\Delta V^\top B^\top$, and then take the product. Since $Q_U \in \mathbb{R}^{d_1 \times r}$ and $\Delta V^\top \in \mathbb{R}^{r \times d_2}$, this sums up to $\mathcal{O}(d_1 r + d_2 r + r^2) \sim \mathcal{O}(dr)$ operations. Similarly, calculating a single entry of $A\Delta U Q_V^\top B^\top$ takes $\mathcal{O}(dr)$ operations. The complexity of a single iteration of CG is thus $\mathcal{O}(dr|\Omega|)$.

Next, we analyze the required number of CG iterations. Let κ_L be the condition number of the least squares operator $\mathcal{L}_{(Q_U,Q_V)}$ as defined in Lemma D.2. The residual error of CG decays at least linearly with a contraction factor $\frac{\kappa_L - 1}{\kappa_L + 1}$ (Hayami, 2018, Section 4). By Lemma D.2, κ_L is bounded by a constant, and hence the required number of CG iterations is also a constant. We thus conclude that the total complexity of step 2 is $\mathcal{O}(dr|\Omega|)$.

Putting everything together, the complexity of `GNIMC` is $\mathcal{O}(dr^3 + dr|\Omega|)$. One of the conditions of the proposition is the lower bound $|\Omega| \geq 32\mu^2 d_1 d_2 \log n$. W.l.o.g., we may assume $|\Omega| = 32\mu^2 d_1 d_2 \log n$ (if $|\Omega|$ is larger, we can ignore some of the observed entries). We thus obtain that the complexity of `GNIMC` is $\mathcal{O}(\mu^2 d d_1 d_2 r \log n)$.

□

D.3. Comparison to Gradient Descent

In this subsection we show that the per-iteration cost of `GNIMC`, as analyzed above, is of the same order as that of gradient descent. Denote $E_\Omega = \mathcal{P}_\Omega(AUV^\top B^\top) - Y$, and let $f(U,V) = \|E_\Omega\|_F^2$ be the objective of the factorized matrix completion problem (2). Its gradient is

$$\nabla_U f(U,V) = 2A^\top E_\Omega BV, \quad \nabla_V f(U,V) = 2B^\top E_\Omega^\top AU.$$

As explained in the analysis of step 2 above, calculating E_Ω costs $\mathcal{O}(dr|\Omega|)$. Since $A^\top E_\Omega$ and $B^\top E_\Omega^\top$ have at most $r|\Omega|$ nonzero entries, this is also the cost of calculating $(A^\top E_\Omega)B$ and $(B^\top E_\Omega^\top)A$. Finally, calculating $(A^\top E_\Omega B)V$ and $(B^\top E_\Omega^\top A)U$ is $\mathcal{O}(d_1 d_2 r)$. The per-iteration complexity of gradient descent is thus $\mathcal{O}(dr|\Omega| + d_1 d_2 r)$. Under assumption (7) on $|\Omega|$, this coincides with the per-iteration complexity of GNIMC.

We remark that empirically, the overall complexity (namely, from initialization to convergence) of gradient descent seems so be much larger than that of GNIMC, see Section 6.1. This observation is in agreement with the theoretical analysis of gradient descent in Zhang et al. (2018), see their Theorem 5.5.

D.4. Parallel and Distributed Computing Considerations

Large scale problems often raise the need for parallel computing as well as GPU implementations. However, as noted by Hubbard and Hedge (2017), parallelization is more challenging for inductive matrix completion than for standard matrix completion. In matrix completion, whose objective is $\min_{U,V} \|\mathcal{P}_\Omega(UV^\top) - Y\|_F$, a single row or column of the estimate $UV^\top$ is associated with a single row of the unknown factor U or V, respectively. In contrast, in IMC, with the objective given in (2), any single row or column of the estimate $AUV^\top B^\top$ is associated with the entire factors U, V, making parallelization harder. Yet, as each iteration of GNIMC solves a least squares problem, this allows for the following naive parallelization scheme: employ a parallelized version of the LSQR algorithm, see e.g. Balay et al. (1996).

Our description of the IMC problem assumes a centralized setting. Li et al. (2020) analyzed the global geometry of a family of low-rank matrix recovery problems in distributed setting. Building upon Li et al. (2020), our theoretical results can be directly extended to a distributed setting as well (cf. Appendix C). In particular, our Theorem 3.1 holds also in the distributed IMC setting.

E. Proof of Proposition 5.3 (Initialization Guarantee)

By its construction, $\left(\begin{smallmatrix} U_0 \\ V_0 \end{smallmatrix}\right)$ is perfectly balanced, $U_0^\top U_0 = V_0^\top V_0$, and thus satisfies (8b). Hence, we only need to prove (8a). Let $\mathcal{A}$ be the sensing operator that corresponds to (IMC) as defined in (3). By Theorem 3.3, w.p. at least $1 - 2n^{-2}$, the operator $\mathcal{A}$ satisfies a $\min\{d_1, d_2\}$-RIP with a constant $\delta = 2/5$. Hence, according to Lemmas 5.1-5.2 in Tu et al. (2016), or more explicitly Eq. (5.26) in their extended arXiv version, after $\tau \geq \log(c\sqrt{r}\kappa)/\log(1/(2\delta)) \geq 5\log(c\sqrt{r}\kappa)$ iterations of (10) we have $\|M_\tau - M^*\|_F \leq \sigma_r^*/c$. Since A, B are isometries, $\|AM_\tau B^\top - X^*\|_F = \|M_\tau - M^*\|_F \leq \sigma_r^*/c$. Hence $\left(\begin{smallmatrix} U_0 \\ V_0 \end{smallmatrix}\right)$ satisfies (8a) for any $\tau \geq 5\log(c\sqrt{r}\kappa)$. □

F. Comparison to Ghassemi et al. (2018)

Ghassemi et al. (2018) derived results analogous to our Theorems 3.1 and 3.3. However, there are three main differences between the claims. First, the sample complexity for the RIP result in Ghassemi et al. (2018) is

$$\mathcal{O}(\mu^2 r \max\{d_1, d_2\} \max\{d_1 d_2, \log^2 n\} \log(1/\delta)/\delta^2), \tag{51}$$

compared to our $\mathcal{O}(\mu^2 d_1 d_2 \log(n)/\delta^2)$. We remark that in their notation, their claimed sample complexity is $\mathcal{O}(\mu^2 \max\{d_1, d_2\}\bar{r}^2 r)$ where $\bar{r} = \max\{r, \log n\}$. However, there seems to be an error in their analysis. Their assumption is that $A/\sqrt{n_1}$ and $B/\sqrt{n_2}$ are isometries, and their corresponding incoherence assumption is $\|A\|_{2,\infty}^2 \leq \mu\bar{r}$ and $\|B\|_{2,\infty}^2 \leq \mu\bar{r}$ with a constant μ (Ghassemi et al., 2018, Assumption 1). But since $A \in \mathbb{R}^{n_1 \times d_1}$ and $B \in \mathbb{R}^{n_2 \times d_2}$, either the assumption should be $\|A\|_{2,\infty}^2 \leq \mu d_1$ and $\|B\|_{2,\infty}^2 \leq \mu d_2$, or the parameter μ is not a constant but rather scales with $\max\{d_1, d_2\}/\bar{r}$. In any case, in our notation their sample complexity is as given in (51).

Second, the RIP result in Ghassemi et al. (2018) is for rank-$\min\{2r, d_1, d_2\}$ matrices, compared to our stronger rank-$\min\{d_1, d_2\}$ RIP. We remark that while their RIP result is formulated as $2r$-RIP, this implicitly assumes $r \ll \min\{d_1, d_2\}$ (see also the discussion in Appendix C). In the general case, their guarantee is $\min\{2r, d_1, d_2\}$-RIP.

Third, Ghassemi et al. (2018) prove benign optimization landscape for the problem

$$\min_{U,V} \|\mathcal{P}_\Omega(AUV^\top B^\top) - Y\| + \frac{1}{4}\|UU^\top - VV^\top\|_F^2, \tag{52}$$

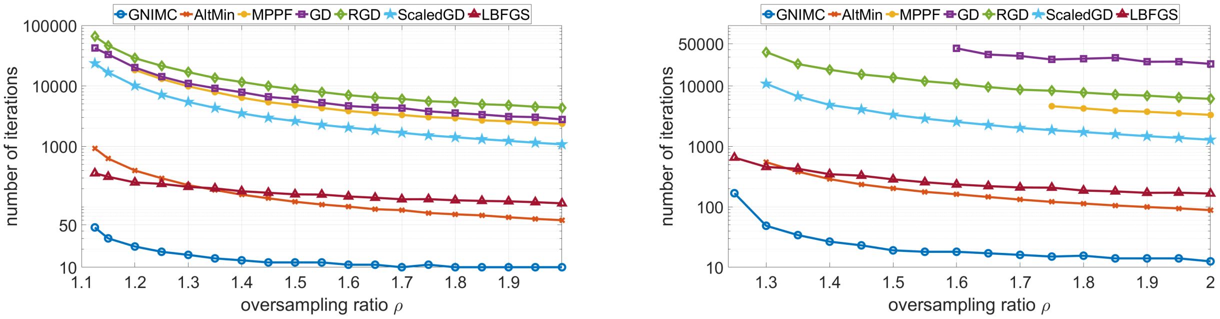

Figure 4. Same as Figure 2, but with Y-axis corresponds to number of iterations.

which is an imbalance regularized version of (2). Furthermore, it seems that their result cannot be readily extended to the vanilla IMC problem, as the regularization in (52) is essential in their proof.

G. Additional Simulation Details

All algorithms are initialized with the same procedure, which is the spectral initialization, except for `Maxide` which is not factorization based and is by default initialized with the zero matrix.

`Maxide` requires as input a regularization parameter; `MPPF`, `GD`, `RGD` and `ScaledGD` require a step size parameter; and `L-BFGS` requires a memory limit. For each simulation setting, we tuned the optimal parameter out of 10 logarithmically-scaled values. The permitted values for `Maxide` were $10^{-5}, ..., 10^{-14}$; for `MPPF`, `GD` and `RGD` were $10^{-2}/\kappa, ..., 10^{\frac{1}{2}}/\kappa$ where κ is the condition number of X^*; for `ScaledGD` were $10^{-2}, ..., 10^{\frac{1}{2}}$; and for `L-BFGS` were $1, ..., 10^3$. In all simulations, we verified that the selected value is an interior point of the permitted set, so that it is close to optimal. We remark that the regularization coefficient of `MPPF` and `RGD` can also be tuned, but we observed it has a very little effect. For `GNIMC`, in all simulations we identically set the maximal number of iterations for the inner least-squares solver to 10 if the observed error is low, $\frac{\|\mathcal{P}_\Omega(X_t)-Y\|_F}{\|Y\|_F} \leq 10^{-4}$, and 1000 otherwise. This scheme exhibits slightly better performance than setting a constant value of maximal inner iterations (but only marginally). While tuning this value for each simulation independently, as we did for the hyperparameters of the above algorithms, may enhance performance, we preferred to demonstrate the performance of a tuning-free version of `GNIMC`.

We used the following two stopping criteria for all methods: (i) small relative observed RMSE, $\frac{\|\mathcal{P}_\Omega(X_t)-Y\|_F}{\|Y\|_F} \leq \epsilon$, or (ii) small relative estimate change $\frac{\|\mathcal{P}_\Omega(X_t-X_{t-1})\|_F}{\|\mathcal{P}_\Omega(X_t)\|_F} \leq \epsilon$. In our simulations, we set $\epsilon = 10^{-14}$. For a fair comparison, we disabled all the other early stopping criteria defined in the algorithms.

H. Additional Simulation Results

In subsection H.1 we show the number of iterations required by each method to recover the matrix instead of the required CPU time as in the main text. In subsection H.2 we demonstrate the stability of `GNIMC` to Gaussian noise. In subsection H.3 we demonstrate the insensitivity of several algorithms to the condition number of the underlying matrix in terms of the number of observations required for recovery.

H.1. Number of Iterations Plots

To gain further insight into the convergence rate of each algorithm, Figure 4 shows the median number of iterations till convergence on a log scale as a function of the oversampling ratio, with the same settings as in Figure 2. In particular, all methods start from the same (spectral) initialization. Note that each iteration of `GNIMC` and `AltMin` includes solving an inner least-squares problem, so these methods are not directly comparable to the others in terms of the number of iterations.

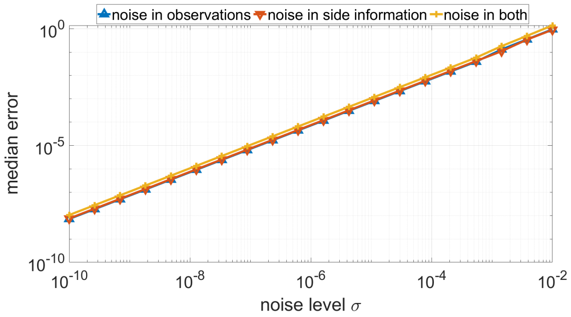

Figure 5. Stability of GNIMC to additive Gaussian noise, with the same settings as in Figure 1(left).

Table 2. Lowest oversampling ratio ρ from which the median of rel-RMSE (17) is lower than 10^{-4} as a function of the condition number κ, in the setting $n_1 = n_2 = 1000$, $d_1 = d_2 = 20$, $r = 10$. L-BFGS fails at $\kappa = 10^4$ for any oversampling ratio. The median rel-RMSE is taken over 50 independent realizations.

Alg. \ κ	1	10	10^2	10^3	10^4
GNIMC	1.2	1.1	1.1	1.1	1.1
AltMin	1.1	1.1	1.1	1.1	1.1
GD	1.1	1.1	1.1	1.1	1.1
RGD	1.2	1.2	1.1	1.1	1.1
ScaledGD	1.2	1.1	1.2	1.2	1.1
Maxide	1.4	1.4	1.4	1.4	1.4
L-BFGS	1.1	1.1	1.1	1.1	–

H.2. Stability of GNIMC to Noise

Figure 5 demonstrates the stability of GNIMC to noise. In this simulation, either the observed entries Y, the side information matrices A, B, or both, are corrupted by additive Gaussian noise of zero mean and standard deviation σ. As seen in the figure, the error of GNIMC scales linearly with the noise level σ.

H.3. Insensitivity of Several Algorithms to the Condition Number

In Figure 1(right) we addressed the sensitivity (or insensitivity) of several IMC algorithms to the condition number of X^* in terms of their runtime. In this subsection, we explore another aspect of sensitivity to the condition number: rather than runtime, we study how the condition number affects the number of observations required for a successful recovery given no time constraints.

In our simulations, we observed the following interesting phenomenon: For all algorithms, the number of observations $|\Omega|$ required for recovery is independent of the condition number κ. We demonstrate this in Table 2, which compares the minimal oversampling ratio, out of the values $\rho = 1.1, 1.2, ...$, required by several algorithms to reach relative RMSE of 10^{-4}. Since in this experiment our goal is to explore fundamental recovery abilities rather than speedy performance, the algorithms are given essentially unlimited runtime (in practice, the time limit was set to one CPU hour, and 3 hours for GD and RGD in the case of $\kappa = 10^4$). The table shows that the minimal oversampling ratio does not increase with κ; in fact, it sometimes slightly decreases when κ is small. We did not include MPPF in Table 2 due to its long runtime; however, a limited set of simulations suggests that the same conclusion also holds for it.

Beyond illustrating the abilities of the algorithms, this result demonstrates a basic property of the IMC problem: insensitivity to the condition number. This result corresponds well with our RIP guarantee for IMC, Theorem 3.3, as the RIP holds for all matrices of certain ranks regardless of their condition number.