Generalization Bounds for Inductive Matrix Completion in Low-noise Settings

Antoine Ledent, Rodrigo Alves, Yunwen Lei, Yann Guermeur, Marius Kloft

Generalization Bounds for Inductive Matrix Completion in Low-noise Settings*

Antoine Ledent[1†], Rodrigo Alves[2], Yunwen Lei[3], Yann Guermeur[4], and Marius Kloft[5]

[1] Singapore Management University (SMU)
aledent@smu.edu.sg
[2] Czech Technical University in Prague (CTU)
rodrigo.alves@fit.cvut.cz
[3] Hong Kong Baptist University (HKBU)
yunwen.lei@hotmail.com
[4] Centre National de la Recherche Scientique (CNRS)
yann.guermeur@loria.fr
[5] Technische Universität Kaiserslautern (TUK)
kloft@cs.uni-kl.de

Abstract

We study inductive matrix completion (matrix completion with side information) under an i.i.d. subgaussian noise assumption at a low noise regime, with uniform sampling of the entries. We obtain for the first time generalization bounds with the following three properties: (1) they scale like the standard deviation of the noise and in particular approach zero in the exact recovery case; (2) even in the presence of noise, they converge to zero when the sample size approaches infinity; and (3) for a fixed dimension of the side information, they only have a logarithmic dependence on the size of the matrix. Differently from many works in approximate recovery, we present results both for bounded Lipschitz losses and for the absolute loss, with the latter relying on Talagrand-type inequalities. The proofs create a bridge between two approaches to the theoretical analysis of matrix completion, since they consist in a combination of techniques from both the exact recovery literature and the approximate recovery literature.

Introduction

Matrix Completion (MC), the problem which consists in predicting the unseen entries of a matrix based on a small number of observations, presents the rare combination of (1) a rich mathematical playground rife with fundamental unsolved problems, and (2) a wealth of unexpected applications in lucrative and meaningful fields, from Recommender Systems (Yao and Kwok 2019; Chen and Li 2017; Aggarwal 2016) to the prediction of drug interaction (Li et al. 2015).

One of the most celebrated algorithms for standard matrix completion is the Softimpute algorithm (Mazumder, Hastie, and Tibshirani 2010), which solves the following optimization problem:

$$\min_{Z\in\mathbb{R}^{m\times n}} \frac{1}{2}\|P_\Omega(Z-R)\|_{\mathrm{Fr}}^2 + \lambda\|Z\|_*, \tag{1}$$

where P_Ω denotes the projection on the set Ω of observed entries, R is the ground truth matrix, $\|\cdot\|_*$ denotes the *nuclear norm* (the sum of the matrix's singular values) and $\|\cdot\|_{\mathrm{Fr}}$ denotes the Frobenius norm. The idea of the algorithm is to encourage *low-rank* solutions in a similar way to how L^1 regularization encourages component sparsity. The parameter λ must be tuned with cross-validation.

Inductive matrix completion (IMC) (Herbster, Pasteris, and Tse 2019; Zhang, Du, and Gu 2018; Menon and Elkan 2011; Chen et al. 2012) is another closely related model which assumes that additional information is available in the form of feature vectors for each user (row) and item (column). It assumes that the side information is summarized in matrices $X \in \mathbb{R}^{m\times a}$ and $Y \in \mathbb{R}^{n\times b}$. IMC then optimizes the following objective function

$$\min_{M\in\mathbb{R}^{a\times b}} \frac{1}{N}\|P_\Omega(XMY^\top - R)\|_{\mathrm{Fr}}^2 + \lambda\|M\|_*. \tag{2}$$

An interesting question is whether one can provide sample complexity guarantees for the optimization problem above. Typically, doing so requires minor modification to the problem for technical convenience. There are several such analogues optimization problems (1) and (2), depending on the type of statistical guarantee expected and the assumptions: in exact recovery (with the assumption of perfectly noiseless observations), the Frobenius norm is replaced by a hard equality constraint, whilst in approximate (noisy) recovery, the nuclear norm regrulariser is replaced by a hard contraint.

More precisely, *exact recovery* results study the following hard version of the optimization problem:

$$\min_{Z\in\mathbb{R}^{m\times n}} \|Z\|_* \quad \text{subject to}$$
$$Z_{i,j} = R_{i,j} \quad \forall (i,j)\in\Omega. \tag{3}$$

In the case of IMC, the equivalent hard version is:

$$\min_{M\in\mathbb{R}^{a\times b}} \|M\|_* \quad \text{subject to}$$
$$[XMY^\top]_{i,j} = R_{i,j} \quad \forall (i,j)\in\Omega. \tag{4}$$

The study of problem (3) is the earliest branch of the related literature: it was shown in a series of papers (Candès and Tao

*Accepted for presentation at AAAI 2023
†Corresponding author

2010; Candès and Recht 2009; Recht 2011, to name but a few) that if the number of samples is $\geqslant \widetilde{O}(nr)$ (where r is the rank and n is the size of the matrix, i.e. the number of rows or columns, which ever is larger), then it is possible to recover the whole matrix exactly with high probability as long as the entries are sampled uniformly at random. There has also been some more recent interest in the problem (4): it was shown in (Xu, Jin, and Zhou 2013) that *assuming the side information X, Y is made up of orthonormal columns*, exact recovery is possible as long as the number of samples $N = |\Omega|$ satisfies $\widetilde{O}(ar) \leqslant N \leqslant \widetilde{O}(abr)$. Here, the $\widetilde{O}$ notation hides logarithmic factors in all relevant quantities (including the size $m \times n$ of the matrix).

Approximate recovery results typically study modified problems such as the problem below, for which Equation (1) can be interpreted as a Lagranian form):

$$\min_{Z} \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \ell(R_{i,j}, Z_{i,j}) \quad \text{subject to} \quad \|Z\|_* \leqslant \mathcal{M}, \tag{5}$$

for some loss function ℓ which is typically assumed to be bounded and Lipschitz, and some constant $\mathcal{M}$ which must be tuned through cross-validation in a way analogous to the tuning of λ in equation (1) in real-life applications. In the case of IMC, the equivalent problem is:

$$\min_{M} \frac{1}{|\Omega|} \sum_{(i,j)\in\Omega} \ell([XMY^\top]_{i,j}, Z_{i,j}) \quad \text{subject to} \quad \|M\|_* \leqslant \mathcal{M}. \tag{6}$$

Approaching the problem this way allows one to deploy the machinery of Rademacher complexities from traditional statistical learning theory to obtain uniform bounds on the generalization gap of any predictor in the given class. Using such techniques, bounds of $\widetilde{O}\left(\sqrt{\frac{nr}{N}}\right)$ (resp. $\widetilde{O}(a^2 r/\sqrt{N})$, more recently $\widetilde{O}\left(\sqrt{\frac{ar}{N}}\right)$) were shown for approximate recovery MC (resp. IMC) under uniform sampling (MC: see (Shamir and Shalev-Shwartz 2011, 2014), IMC, see (Chiang, Dhillon, and Hsieh 2018; Ledent et al. 2021), cf. also related works). In the distribution-free case, the corresponding rates are $\widetilde{O}\left(\sqrt{\frac{n^{3/2}r^{1/2}}{N}}\right)$ and $\widetilde{O}\left(\sqrt{\frac{a^{3/2}r^{1/2}}{N}}\right)$.

The above rates do not make any assumptions on the noise whatsoever, and depend only on explicit dimensional quantities: they are classified as "uniform convergence" bounds in the classic paradigm of statistical learning theory. In particular, while they do also apply to the noiseless case, they are subsumed by the exact recovery results in this case provided the exact recovery threshold is reached.

Thus the most striking hole in the existing theory is the chasm between exact recovery and approximate recovery in Inductive Matrix Completion: on the one hand, we know that if the entries are observed exactly, solving problem (4) will eventually recover the whole matrix exactly with high probability given enough entries. On the other hand, we know from the approximate recovery literature that regardless of the noise distribution, solving a properly cross-validated version of problem (6) will allow us to approach the Bayes error at speed at least $1/\sqrt{N}$ as we observe more entries. It seems reasonable to expect that in real life, neither of these approaches fully explains the statistical generalization landscape of the problem: we never expect to observe the entries exactly, and the ground truth is probably not exactly low-rank either, but we still do not expect convergence to the Bayes error to be as slow as in the worst case. What would be more reasonable to expect is a sharp decline of the error around a threshold value before which no method can work even if the entries are observed exactly, followed by a slower decline as the model refines its predictions and evens out the noise in the observations. This can be observed practically as well, as can be seen from Figure 1 in the experiments section: the decay of the error as the number of samples increases is neither convex (unlike the functions $1/\sqrt{N}$ and $1/N$), nor completely abrupt (as exact recovery results suggest), which indicates the presence of a threshold phenomenon.

In this paper, we theoretically capture this phenomenon through generalization error bounds for the solutions to problem (2) when the ground truth matrix is observed with some subgaussian noise of subgaussianity constant σ. In addition, our results completely remove the orthogonality assumptions on the side information matrices X, Y which are present in the related work (Xu, Jin, and Zhou 2013), thus improving the state of the art even in the exact recovery case.

In summary, we make the following important contributions:

1. We prove (cf. Theorem 1) that *exact recovery* is possible for IMC (when the entries are observed exactly) with probability $1 - \Delta$ given $\widetilde{O}\left(\mu^5 r^2 (a+b)\sigma_0^{-4} \log\left(\frac{mn}{\Delta}\right)\right)$ samples or more. This is a significant extension of the results in (Xu, Jin, and Zhou 2013) in that we remove most many of their assumptions. In the formula above, μ is a measure of incoherence, and σ_0 denotes the smallest singular value of X or Y assuming they are normalized so that the largest singular value is 1 in each case. This means that after suitable scaling, σ_0 can be replaced by the ratio between the largest and smallest singular values of either X or Y. The presence of this factor underpins one of the main differences between (Xu, Jin, and Zhou 2013) and our work. Indeed, the most limiting assumption in (Xu, Jin, and Zhou 2013) is that the columns of the side information matrices X and Y are *orthonormal*, which is equivalent to assuming that $\sigma_0 = 1$.
2. We experimentally observe the two-phase phenomenon described above via synthetic data experiments.
3. We prove generalization bounds (cf. Theorem 2) which capture this phenomenon in the case of bounded loss functions such as the truncated L^2 loss. Indeed, we show that as long as N exceeds the threshold from the exact recovery result, the expected loss scales as $\widetilde{O}\left(\sigma_0^{-2}\,\sigma\,\mu\frac{\sqrt{a^3 b}}{\sqrt{N}}\log^3(N/\Delta)\right)$, where σ is the subgaussianity constant of the noise. If σ is very small, this implies that before the exact recovery threshold (ERT) is reached, the best available bounds are the uniform convergence bounds (which are vacuous at that regime), whereas as soon as the ERT is crossed, our bounds become valid

and already have a small value, which continues to drop further as the number of samples increases. This partially explains the sharp drop in the reconstruction error around the ERT even in the noisy case.

4. Using Talagrand-type inequalities, we further prove a similar result (cf. Theorem 3) which applies to the absolute loss $\ell(x, y) = |x-y|$, despite the fact that it is unbounded.

Note as a side benefit that both of the last two results apply to the Lagrangian formulation of the IMC problem, unlike most of the existing literature on approximate recovery.

Our second result creates a bridge between the approximate recovery literature and the exact recovery literature: as the subgaussianity constant of the noise σ converges to zero, so does the error: the result then reduces to our exact recovery result. Furthermore, our proof techniques also marry both approaches: we rely *both* on the geometry of dual certificates (the tool of choice in the exact recovery literature) *and* Rademacher complexities to reach our result. Beyond our current preliminary results, we believe that the direction we initiate here will prove fertile and that many improved results can be proved, bringing us closer to a complete understanding of the sample complexity landscape of nuclear norm based Inductive Matrix Completion.

Related work

Perturbed exact recovery with the nuclear norm: For Matrix Completion without side information, bounds which capture the two-phase phenomenon by incorporating a multiplicative factor of the the variance of the noise have been shown: in (Candès and Plan 2010), a bound of order $O\left(\sqrt{\frac{n^3}{N}}\,\sigma + \sigma\right)$ is shown for the L^2 generalisation error of matrix completion with noise of variance σ (Cf. equation III.3 on page 7). The proof relies on a perturbed version of the exact recovery arguments presented in (Candès and Tao 2010). The result considers a different loss function and does not consider side information and the proof is purely based on directly computing various norms without relying on Rademacher complexities. In a recent and very impressive contribution (Chen et al. 2020) provided some bounds in the same setting with a finer multiplicative dependency on the size of the matrix n that matches the order of magnitude of the exact recovery threshold (when expressed in terms of sample complexity). The proof is very involved and contrary to our work, the results do not apply to inductive matrix completion.

Exact recovery with the nuclear norm: In (Recht 2011), extending and simplifying earlier work of (Candès and Tao 2010; Candès and Recht 2009), the author proves that exact recovery is possible for matrix completion with the nuclear norm with $\widetilde{O}(nr)$ entries. The result is extended to the case where side information is present in (Xu, Jin, and Zhou 2013) where it is shown exact recovery is possible with $\widetilde{O}((a+b)r)$ observations, where a, b are the sizes of the side information. However, the result only applies as long as this side information consists of orthonormal columns, significantly reducing the applicability. Other variations of the results exist with improved dependence on certain parameters such as the incoherence constants (Chen 2015).

Perturbed exact recovery for other algorithms in learning settings other than nuclear norm minimization, there is some work with low-noise regimes where the bounds also approach zero as the noise approaches zero (for large enough N). For instance, some work on max norm regularisation has this property (Cai and Zhou 2016). Some results of order $\widetilde{O}\left(\sigma\sqrt{\frac{nr}{N}}\right)$ were also obtained for matrix completion with a special algorithm that *requires explicit rank restriction* (Keshavan, Montanari, and Oh 2009; Wang et al. 2021).

Approximate recovery results: There is a wide body of works proving uniform-convergence type generalization bounds for various matrix completion settings. the vast majority are of order $\widetilde{O}(1/\sqrt{N})$, with most bounds differing from each other in their dependence on other quantities such as m, n, r, μ, σ and (in IMC) a, b. For matrix completion, (Shamir and Shalev-Shwartz 2011, 2014) proves bounds of order $\widetilde{O}\left(\sqrt{\frac{n^{3/2}r^{1/2}}{N}}\right)$ in the *distribution-free setting* with replacement, as well as $\widetilde{O}\left(\frac{nr\log(n)}{N} + \sqrt{\frac{\log(1/\delta)}{N}}\right)$ in the transductive setting (i.e. for *uniform* sampling *without* replacement). In the case of inductive matrix completion, rates of $\widetilde{O}\left(\sqrt{\frac{rab}{N}}\right)$ were shown in (Chiang, Dhillon, and Hsieh 2018; Chiang, Hsieh, and Dhillon 2015; Giménez-Febrer, Pagès-Zamora, and Giannakis 2020) in a distribution-free situation, whilst (Ledent et al. 2021) provides rates of order $\widetilde{O}\left(\sqrt{\frac{ra}{N}}\right)$ and $\widetilde{O}\left(\sqrt{\frac{r^{1/2}a^{3/2}}{N}}\right)$ in the uniform sampling and distribution-free cases respectively. Similar rates were implicitly proved in the more algorithmic contribution (Ledent, Alves, and Kloft 2021) under very strict assumptions on the side information X, Y. It is also worth noting that although the component of our result which involves the subgaussianity of the noise is vacuous when the size of the side information approaches that of the matrix, that is also the case of every approximate recovery result for IMC to date except the very recent paper (Ledent et al. 2021), whose results are also uniform convergence bounds. Our bounds are far tighter those in all of those works when the noise is small.

Matrix sensing: Matrix sensing is a learning setting with some similarities to inductive matrix completion where rank-one measurements $\langle vw^\top, R\rangle$ of an unknown matrix R are taken, and the matrix R is estimated. There are a wide variety of results depending on the assumptions on the matrix and the sampling distribution (Gross et al. 2010; Kueng, Rauhut, and Terstiege 2017; Tanner, Thompson, and Vary 2019; Zhong, Jain, and Dhillon 2015). In most cases, the measurements are sampled i.i.d. from some distribution, which introduces some substantial technical differences to the IMC setting. Often, the underlying measurements need to satisfy the restricted isometry property, which is not directly comparable to the joint incoherence assumptions on the side information matrices made in this paper and in the IMC literature. In addition, most results relate to pure exact recovery rather than a low-noise model such as the one studied here.

Notation and setting

We assume there is an unknown ground truth matrix $R \in \mathbb{R}^{m\times n}$ that we observe noisily. To draw a sample from the distribution, we first sample an entry $\xi = (\xi_1, \xi_2) = (i, j)$ from the uniform distribution over $[m]\times[n]$. We then observe the quantity $R_{(i,j)} + \zeta_{(i,j)}$ where $\zeta_{(i,j)}$ is the noise, whose distribution can depend on the entry (i, j). The samples are drawn i.i.d.

We suppose we have a training set of N samples and we write Ω for the set of sampled entries $\xi^1, \xi^2, \ldots, \xi^N$. It is possible to sample the same entry several times (which results in potentially different observations due to the i.i.d. nature of the noise). However, for simplicity of notation we will sometimes write $\sum_{(i,j)\in\Omega} f(R_{(i,j)})$ instead of $\sum_{\xi\in\Omega} f(R_{\xi_1,\xi_2}, \xi)$ as long as no ambiguity is possible. We are given two side information matrices $X \in \mathbb{R}^{m\times a}$ and $Y \in \mathbb{R}^{n\times b}$. Throughout this paper, $\|\cdot\|$ denotes the spectral norm, $\|\cdot\|_{\mathrm{Fr}}$ denotes the Frobenius norm, $\|\cdot\|_*$ denotes the nuclear norm, and for any integer l, $[l] = \{1, 2, \ldots, l\}$.

We make the following assumptions throughout the paper:

Assumption 1 (Realizability)**.** There exists a matrix $M_* \in \mathbb{R}^{a\times b}$ such that $R = XM_*Y^\top$.

Assumption 2 (Assumptions on the subgaussian noise)**.** We assume the noise is σ subgaussian: $\mathbb{E}(\zeta) = 0$ and $\mathbb{P}(|\zeta| \geqslant t) \leqslant 2\exp(-t^2/(2\sigma^2))$ for all t.

We will write $\bar{X}$ and $\bar{Y}$ for the matrices obtained by normalizing the columns of X, Y and we will write Σ_1, Σ_2 for the diagonal matrices containing the singular values of X, Y. Similarly we will also write $\overline{\bar{X}} = \bar{X}\Sigma_1$ etc.

We also make the following incoherence assumption.

Assumption 3. There exists a constant μ such that the following inequalities hold.

$$\|\bar{X}\|_\infty \leqslant \sqrt{\frac{\mu}{m}}, \qquad \|\bar{Y}\|_\infty \leqslant \sqrt{\frac{\mu}{n}},$$
$$\|A\|_\infty \leqslant \sqrt{\frac{\mu}{a}}, \qquad \|B\|_\infty \leqslant \sqrt{\frac{\mu}{b}}, \tag{7}$$

Here the matrices A, B are from the SVD decomposition of the ground truth core matrix $M_* = ADB^\top$ for some diagonal D.

Note that we do not make the assumption that the matrices X, Y have orthonormal columns (and in particular constant spectrum) as in (Xu, Jin, and Zhou 2013). Therefore, to cope with such extra difficulty (7) is needed in the general non orthogonal case. Whilst that reference simply assumes that the column spaces of X, Y are μ incoherent, our assumption requires that *each individual eigenspace* corresponding to each singular value of X, Y and M *be* μ*-incoherent*. In the supplementary we explain to what extent this slightly stronger assumption is necessary in the non-orthogonal case.

Optimization problem: whether considering inductive matrix completion or matrix completion with the nuclear norm, it is common to assume that the entries are sampled exactly (without noise) and that the algorithm used to recover the ground truth is the following:

$$\arg\min \left(\|M\|_* \text{ s.t. } \forall (i, j) \in \Omega, [XMY^\top]_{i,j} = R_{i,j}\right). \tag{8}$$

This is also the optimization problem we study in the exact recovery portion of our results.

In real situations where there is some noise, some relaxation of the problem is necessary. From an optimization perspective, the most common strategy is to minimize the L^2 loss on the observed entries plus a nuclear norm regularisation term:

$$\min \frac{1}{N}\sum_{\xi\in\Omega} \left|[R_\xi + \zeta_\xi] - XMY^\top\right|^2 + \lambda\|M\|_*, \tag{9}$$

where λ is a regularization parameter. The problem we will consider in this paper is the one defined by equation (C.1). We will also need to impose the following conditions on λ:

$$\frac{\sigma\,\sigma_0^2}{C\sqrt{aN}} \leqslant \lambda \leqslant \frac{C\,\sigma\,\sigma_0^2}{\sqrt{aN}}. \tag{10}$$

for some constant C. It is assumed that λ has been cross-validated to reach a value which satisfies these conditions.

Main results

Exact recovery

We have the following extension of the main theorem in (Xu, Jin, and Zhou 2013):

Theorem 1. *Assume that the entries are observed without noise and that the strong incoherence assumption* (7) *is satisfied for a fixed* μ*. For any* $\Delta > 0$ *as long as*

$$N \geqslant \widetilde{O}\left(\mu^5 r^2 (a + b)\sigma_0^{-4} \log\left(\frac{mn}{\Delta}\right)\right),$$

with probability $\geqslant 1 - \Delta$ *we have that any solution* $M_{\min}$ *to the optimization problem below*

$$M_{\min} \in \arg\min \|M\|_* \quad \textit{s.t.}$$
$$\forall (i, j) \in \Omega, \quad [XMY^\top]_{i,j} = R_{i,j}, \tag{11}$$

satisfies

$$XM_{\min}Y^\top = R.$$

Here, as usual, the $\widetilde{O}$ *notation hides further log terms in the quantities* $m, n, \sigma_0^{-1}, \log(\frac{mn}{\Delta})$.

Remark: The above optimization problem can be seen as a limiting case of (C.1) with $\lambda \to 0$.

Remark: The above theorem has several advantages over the main theorem in (Xu, Jin, and Zhou 2013):

1. It is expressed entirely in terms of a fixed high probability $1 - \Delta$ (as opposed to relying on dimensional quantities in the expression for the high probability).
2. It works without assuming that the side information matrices have unit singular values. This is quite a significant improvement as the result in (Xu, Jin, and Zhou 2013) only holds when the side information matrices belong to a given set of measure zero. There is a quadratic dependence on σ_0^{-1} (the inverse of the smallest singular value of either X or Y), which matches the dependence in (Jain and Dhillon 2013) (although that paper works with a completely different optimization problem away from traditional nuclear norm regularization).

3. It holds for any value of N, whereas the result in (Xu, Jin, and Zhou 2013) required $N \leqslant \widetilde{O}(abr)$ and the result in (Recht 2011) (which concerns standard MC without side information) required $N \leqslant mn$.

Approximate recovery in a low-noise setting

Below we present theorems which provide generalization bounds for the IMC model (2) with the favourable property that they improve when the noise is reduced, and they reduce exactly to the exact recovery result when $\sigma = 0$.

The following theorem provides a generalization bound of order $\widetilde{O}\left(a^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\sqrt{\frac{1}{N}}\right)$ for a bounded Lipschitz loss.

Theorem 2. *Let ℓ be an L_ℓ-Lipschitz loss function bounded by B_ℓ. Assume that condition* (10) *on λ holds. For any $\Delta > 0$, with probability $1-\Delta$ as long as*

$$N \geqslant \widetilde{O}\left(\mu^5 r^2 (a+b)\sigma_0^{-4}\log\left(\frac{mn}{\Delta}\right)\right),$$

we have the following bound on the performance of the solution $\widehat{R}$ to the optimization problem (2)*:*

$$\mathbb{E}_{(i,j)\sim\mathcal{U}}(\ell(\widehat{R}_{(i,j)}, [R+\zeta]_{(i,j)})) \leqslant \tag{12}$$
$$O\left(a^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\mathrm{L}_\ell\log^3\left(\frac{Nmn}{\Delta}\right)\sqrt{\frac{1}{N}} + B_\ell\frac{\log(\frac{1}{\Delta})}{N}\right),$$

where $\mathcal{U}$ stands for the uniform distribution on the entries $[m]\times[n]$.

Next, our proof techniques also allow us to prove results which apply to the absolute value loss, despite the fact that it is unbounded. Indeed, a bound of order $\sqrt{N}$ on the nuclear norm of the difference between the solution and the ground truth is a byproduct of the approximations we perform before applying Rademacher arguments. It can also be used to provide a bound on the *effective* value of B_ℓ, still yielding an overall rate of $1/\sqrt{N}$ thanks to the fact that the last term in equation (13) has the strong decay $1/N$. This is a result of our use of the more fine-grained, talagrand-type results from (Bartlett, Bousquet, and Mendelson 2005) and would not have been possible if we had used standard results on Rademacher complexities such as (Bartlett and Mendelson 2001).

Theorem 3. *Assume that condition* (10) *on λ holds. For any $\Delta > 0$, with probability $1-\Delta$ as long as*

$$N \geqslant \widetilde{O}\left(\mu^5 r^2 (a+b)\sigma_0^{-4}\log\left(\frac{mn}{\Delta}\right)\right),$$

we have

$$\mathbb{E}_{(i,j)\sim\mathcal{U}}\left|\widehat{R}_{(i,j)} - [R+\zeta]_{(i,j)}\right| \leqslant$$
$$O\left(a^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\mathrm{L}_\ell\log^3\left(\frac{Nmn}{\Delta}\right)\sqrt{\frac{1}{N}}\right). \tag{13}$$

Here, $\mathcal{U}$ stands for the uniform distribution on the entries $[m]\times[n]$.

Proof strategy

The main ideas of our proof are (1) to redefine a norm on $\mathbb{R}^{m\times n}$ matrices that captures the effect of the side information matrices, and (2) to combine proof techniques from both the approximate recovery literature and the exact recovery literature: we perturb the analysis from the exact recovery literature to obtain a bound on the discrepancy between the ground truth and the recovered matrix, and then bootstrap the argument by exploiting the i.i.d. nature of the noise and results from traditional complexity analysis to yield a generalization bound.

In this informal description, we sometimes write formulae with such as $P_\Omega(\widehat{R}-R)$, denoting the projection of $\widehat{R}-R$ onto the set of matrices whose non zero entries are in Ω, which requires assuming that each entry was sampled only once. However, this assumption is made purely for simplicity of exposition and it is not made or needed in the formal proofs in the supplementary.

Background on existing techniques

The main strategy of the proof of the exact recovery results in both (Xu, Jin, and Zhou 2013) and (Recht 2011), which goes back to earlier work (Candès and Tao 2010; Candès and Recht 2009; Candès and Plan 2010) is to use the duality between the nuclear norm and the spectral norm to study the behavior of the nuclear norm around the ground truth.

It is easiest to explain the strategy in the case of standard matrix completion (as in Recht 2011; Candès and Plan 2010 etc.). For a given matrix R with singular value decomposition $EDF^\top$, if the columns and rows of W are orthogonal to those of R and it satisfies $\|W\| \leqslant 1$, the matrix $\mathcal{Y} := EF^\top + W$ is a *subgradient to the nuclear norm* at R, and a solution to the maximization problem

$$\max_{\mathcal{Y}} \langle \mathcal{Y}, R\rangle \quad \text{subject to}$$
$$\|\mathcal{Y}\| \leqslant 1.$$

The subgradients as above allow us to understand the local behavior of the nuclear norm around the ground truth, and one of the most important observations in the early exact recovery analysis of matrix completion is that exact recovery is guaranteed if there exists such a subgradient *whose non zero entries are all in the set of observed entries* and whose spectral norm is < 1. A subgradient with this property is referred to as a *dual certificate*. Indeed, we have the following result from (Candès and Plan 2010):

Lemma 4. *If there exists a dual certificate $\mathcal{Y}$, then for any Z with $Z_{i,j} = 0 \quad \forall (i,j) \in \Omega$ we have*

$$\|R+Z\|_* \geqslant \|R\|_* - (1 - P_{T^\top}(\mathcal{Y}))\|P_{T^\top}(Z)\|_*. \tag{14}$$

In particular, R is the unique solution to the optimization problem (8)*. Here $P_T(Z) = ZP_F + P_E Z - P_E Z P_F$ where P_E and P_F are the projection operators onto the column and row spaces of the ground truth respectively.*

The high-level intuition behind such a result is that if the set of "observable" matrices whose entries are constrained to lie in the set of observed entries is big enough to contain suitable subgradients, then it is big enough to make the solution to (8) unique.

Whilst most of the early works in the field (Candès and Tao 2010; Candès and Recht 2009) work with sampling without replacement and rely on complex combinatorial arguments to prove the existence of a dual certificate, the breakthrough in the work of (Recht 2011) is to sample with replacement (simplifying the concentration arguments) and to show that the existence of an *approximate* dual certificate is also enough to guarantee uniqueness. More precisely, let $Z \in \mathbb{R}^{\Omega^\top}$ be a matrix with zeros in all entries outside Ω, and let $U, U^\top$ be the canonical subgradients of R and $P_T(Z)$ respectively. Assume there is an approximate dual certificate $\mathcal{Y}$ with the property that $\|U - P_T(\mathcal{Y})\|_{\mathrm{Fr}}$ is very small and $P_{T^\top}(\mathcal{Y}) < 1/2$, then we have

$$\begin{aligned}
&\|R + Z\|_* \\
&\geqslant \left\langle U + U^\top, R + Z \right\rangle \\
&= \|R\|_* + \left\langle U + U^\top, Z \right\rangle \\
&= \|R\|_* + \langle U - P_T(\mathcal{Y}), P_T(Z) \rangle \\
&\qquad + \left\langle U^\top - P_{T^\top}(\mathcal{Y}), P_{T^\top}(Z) \right\rangle \\
&\geqslant \|R\|_* - \|U - P_T(\mathcal{Y})\|_{\mathrm{Fr}} \|P_T(Z)\|_{\mathrm{Fr}} \\
&\qquad + \|P_{T^\top}(Z)\|_* \left(1 - \|P_T(\mathcal{Y})\|\right). \qquad (15)
\end{aligned}$$

As long as $\|P_T(\mathcal{Y})\| < 1$, $\|U - P_T(\mathcal{Y})\|_{\mathrm{Fr}}$ is small enough and $\|P_T(Z)\|_*$ is not too large in relation to $\|P_{T^\top}(Z)\|_*$, the solution will thus be unique.

In (Xu, Jin, and Zhou 2013) these ideas are extended to the case where side information matrices X, Y *with orthonormal columns* is provided. The key here is that with this assumption on the columns, $\|XMY^\top\|_* = \|M\|_*$ for any matrix M, so that most of the arguments above still hold with minor modification, even after replacing the projection operator P_T by its inductive analogue $P_T(Z) = P_X Z P_F + P_E Z P_B - P_E Z P_F$.

Removing the homogeneity assumption: proof strategy

In our case, where X, Y are arbitrary (they can without loss of generality be assumed to have orthogonal columns, though not necessarily of norm 1), it is no longer true that $\|XMY^\top\|_* = \|M\|_*$ for any M. To tackle this issue, we define a norm $\|Z\|_{\mathcal{I},*}$ on the set of matrices $\mathbb{R}^{m \times n}$ which equals the minimum possible nuclear norm of a matrix M such that $XMY^\top = Z$:

$$\|Z\|_{\mathcal{I},*} = \min \left(\|M\|_* \quad : XMY^\top = Z \right). \qquad (16)$$

A key observation is that both this norm and *its dual* can be computed easily. Indeed, it is easy to see that $\|Z\|_{\mathcal{I},*} = \Sigma_1^2 X^\top Z Y \Sigma_2^2$ where Σ_1, Σ_2 are matrices containing the singular values of X, Y. Furthermore, we also show in the supplementary that in fact the dual norm $\|\cdot\|_{\mathcal{I},\sigma}$ is simply the spectral norm of the matrix $X^\top R Y$. These modifications mean that during the proof, we must manipulate 5 different norms ($\|\cdot\|, \|\cdot\|_*, \|\cdot\|_{\mathcal{I},\sigma}, \|\cdot\|_{\mathcal{I},*}$ and $\|\cdot\|_{\mathrm{Fr}}$), sometimes incurring factors of the smallest singular value σ_0 of X, Y.

We note that removing the homogeneity assumption has consequences in the proofs, including the need for a stronger incoherence assumption.

Fast decay in low-noise settings: proof strategy

In addition, we need to account for the noise, thus instead of perturbing the matrix R only by a matrix Z with $P_\Omega(Z) = 0$, we also perturb it by a matrix H with $P_{\Omega^\top}(H) = 0$ corresponding to the difference between the recovered matrix and the ground truth on the observed entries. Thus our recovered matrix, the solution to algorithm (2), $\widehat{R}$, can be written $\widehat{R} = R + H + Z$.

Our next step is to perform a perturbed version of the calculation in equation (15) taking into account the difference $H = P_\Omega(\widehat{R} - R)$. This is the calculation performed in the proof of Lemma C.1. As previously we write U for a subgradient of $\|R\|_{\mathcal{I},*}$ and $U^\top$ for a subgradient of $\|P_T(Z)\|_{\mathcal{I},*}$. We start by expressing $\|\widehat{R}\|_{\mathcal{I},*}$ as $\left\langle R + H + Z, U + U^\top \right\rangle$ and after some calculations we obtain the following conclusion:

$$\begin{aligned}
\|R\|_{\mathcal{I},*} &\geqslant \|\widehat{R}\|_{\mathcal{I},*} \\
&\geqslant \|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} + \frac{1}{4}\|P_{T^\top}(Z)\|_{\mathcal{I},*}, \qquad (17)
\end{aligned}$$

which holds as long as several concentration phenomena occur (which will happen with high probability as long as N is large enough).

Our next step is to bound $\|H\|_{\mathcal{I},*}$. With high probability, the noisily observed entries of R on Ω (the $R_\xi + \zeta_\xi$) are close to the actual entries R, which in turn implies that the entries of H will not be too large (see the beginning of the proof of Theorem D.2).

This yields a bound of order $\widetilde{O}(\sqrt{N}\nu)$ for $\|H\|_{\mathcal{I},*}$, and then via equation (17), on $\|P_T(Z)\|_*$. Together with further modifications, this eventually yields a bound on the nuclear norm of $Z + H = \widehat{R} - R$. This means that our perturbed version of the exact recovery results places the recovered matrix $\widehat{R}$ inside of the smaller function class of matrices within a bounded spectral norm of the ground truth matrix. At this point, we can leverage classical results on the Rademacher complexity of the function class of matrices with bounded nuclear norm (see Lemma A.6 below for the inductive version we use in practice) to further bound the generalization gap. Several further steps are needed to process the final result into an elegant formula that holds for any value of N. The details are in the supplementary material.

Lemma 5 (Chiang, Dhillon, and Hsieh 2018)**.** *The function class* $\left\{XMY^\top : \|M\|_* \leqslant \mathcal{M}\right\}$ *satisfies*

$$\mathfrak{R}(\mathcal{F}_\mathcal{M}) \leqslant \mathbf{x}\mathbf{y}\mathcal{M}\sqrt{\frac{1}{N}}, \qquad (18)$$

where $\mathbf{x} := \|X^\top\|_{2,\infty}$ *and* $\mathbf{y} := \|Y^\top\|_{2,\infty}$.

Proof. Follows directly from Theorem 1 in (Kakade, Sridharan, and Tewari 2009), together with the duality between the nuclear and spectral norms (Fazel, Hindi, and Boyd 2001). Cf. also (Chiang, Dhillon, and Hsieh 2018). □

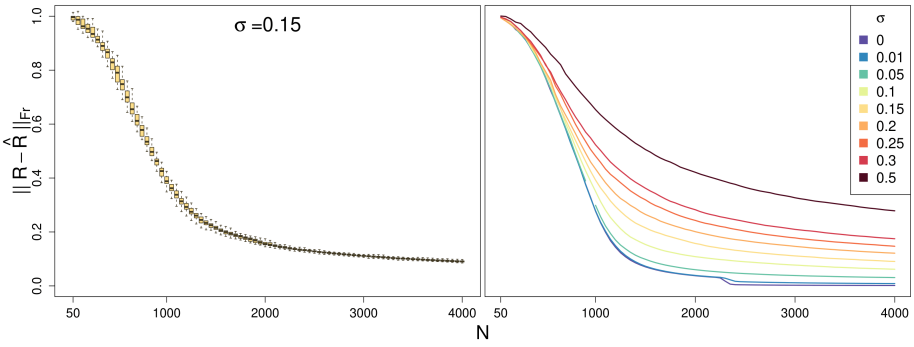

Figure 1: $\|\widehat{R} - R\|_{\mathrm{Fr}}$ as a function of N, σ

Experiments

In this paper, we have posited that an accurate understanding of the sample complexity landscape of inductive matrix completion requires treating the noise component differently from the ground truth entries for the purposes of complexity. In this section we present the experiments we ran to confirm that a two-phase phenomenon as suggested by our bounds does in fact occur in practice.

We considered random matrices of size 100×100 and of rank 10[1], and created random orthonormal side information of rank 40, ensuring that the singular vectors of the ground truth matrix are in the span of the relevant side information, but with the orientation being otherwise uniformly random. The ground truth matrices were normalized to have Frobenius norm 100, and we then added i.i.d. $N(0, \sigma^2)$ gaussian noise to each observation. We performed classic inductive matrix completion (with the square loss) on the resulting training set, cross-validating the parameter λ on a validation set, and evaluated the RMSE distance between the resulting trained matrix and the ground truth. We performed this whole procedure for a wide range of different values for the number of samples N. For each value of N we perform the procedure on 40 different random matrix and side information.

The results are presented in Figure 1 below. The graph on the left contains box plots for our simulation with $\sigma = 0.15$ whilst the graph on the right presents our results, averaged over the 40 simulations, for several values of σ.

As can be observed in the figure, the graph of the error as a function of N is not convex, despite the fact that traditional approximate recovery bounds $\widetilde{O}(1/\sqrt{N})$ are convex. Instead, the graph looks like a sigmoid: we can clearly observe a thresholding phenomenon where the performance is very poor initially, but very quickly improves past a minimum number of entries. Furthermore, as can be expected, after the threshold is crossed the error decreases slowly (at least as $\widetilde{O}(\frac{\sigma}{\sqrt{N}})$ as per the bounds in Theorem 2 above), confirming that inductive matrix completion in low-noise settings exhibits a two-phase phenomenon matching our theoretical results. Furthermore, the fact that the post threshold error curve scales as σ is also apparent from the graphs.

[1]To generate such a random matrix, we generate matrices $U, V \in \mathbb{R}^{100\times 10}$ with i.i.d. gaussian entries, then we form the matrix $UV^\top$ and we normalize it to have Frobenius norm 100.

Conclusion and future directions

In this paper, we have studied *Inductive Matrix Completion* with nuclear norm regularisation in low-noise regimes. Our first contribution is an exact recovery result which generalizes the existing ones to the case where the side information is no longer assumed to be orthonormal, and to an arbitrary sampling regime (previously, the number of samples was required to be bounded *above* by $\widetilde{O}(abr)$). Our second contribution consists in generalization bounds composed of two components: (1) the requirement that the number of samples should exceed a given threshold and (2) a term of order $\widetilde{O}(\sigma\, \sigma_0^{-2} a^{3/2}\sqrt{b}\log^3(N/\Delta)\sqrt{\frac{1}{N}})$ (ignoring incoherence constants and other constant quantities), which is directly proportional to the subgaussianity constant σ of the noise. In particular, the result forms a bridge between exact recovery results and approximate recovery results: at the regimes where exact recovery is possible, the error converges to zero when the noise converges to zero.

We believe our result and proof strategy open the door to a new and unexplored direction of research. Possible future directions include improving the dependence on N from $1/\sqrt{N}$ to $\frac{1}{N}$, extending the results to non-trivially non uniform distributions or providing analogues of our results for other low-rank learning problems such as density estimation (Song et al. 2014; Vandermeulen and Ledent 2021; Kargas and Sidiropoulos 2019; Anandkumar et al. 2014; Vandermeulen 2020; Amiridi, Kargas, and Sidiropoulos 2020, 2021) or more complex recommender systems models that involve implicit feedback or graph/cluster information (Zhang and Chen 2020; Alves et al. 2020; Wu et al. 2020; Steck 2019; Vančura et al. 2022; Lin et al. 2022; Shen et al. 2021). Improving the dependence on a, b to match the scaling of the ERT is also a very ambitious and interesting aim.

Acknowledgements

Rodrigo Alves thanks Recombee for supporting his research. Marius Kloft acknowledges support by the Carl-Zeiss Foundation, the DFG awards KL 2698/2-1, KL 2698/5-1, KL 2698/6-1, and KL 2698/7-1, and the BMBF awards 01|S18051A, 03|B0770E, and 01|S21010C.

References

Aggarwal, C. C. 2016. *Recommender Systems: The Textbook*. Springer Publishing Company, Incorporated, 1st edition. ISBN 3319296574.

Alves, R.; Ledent, A.; Assunção, R.; and Kloft, M. 2020. An Empirical Study of the Discreteness Prior in Low-Rank Matrix Completion. *Proceedings of Machine Learning Research (PMLR): NeurIPS 2020 Workshop on the Pre-registration Experiment: An Alternative Publication Model For Machine Learning Research*.

Amiridi, M.; Kargas, N.; and Sidiropoulos, N. D. 2020. Low-rank Characteristic Tensor Density Estimation Part I: Foundations. *arXiv e-prints*, arXiv:2008.12315.

Amiridi, M.; Kargas, N.; and Sidiropoulos, N. D. 2021. Low-rank Characteristic Tensor Density Estimation Part II: Compression and Latent Density Estimation. *arXiv e-prints*, arXiv:2106.10591.

Anandkumar, A.; Ge, R.; Hsu, D.; Kakade, S. M.; and Telgarsky, M. 2014. Tensor Decompositions for Learning Latent Variable Models. *Journal of Machine Learning Research*, 15: 2773–2832.

Bartlett, P. L.; Bousquet, O.; and Mendelson, S. 2005. Local Rademacher complexities. *The Annals of Statistics*, 33(4): 1497 – 1537.

Bartlett, P. L.; and Mendelson, S. 2001. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. In Helmbold, D.; and Williamson, B., eds., *Computational Learning Theory*, 224–240. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-44581-4.

Boucheron, S.; Lugosi, G.; and Bousquet, O. 2004. Concentration inequalities. *Lecture Notes in Computer Science*, 3176: 208–240.

Cai, T. T.; and Zhou, W.-X. 2016. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1): 1493 – 1525.

Candès, E. J.; and Recht, B. 2009. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6): 717.

Candès, E. J.; and Tao, T. 2010. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE Trans. Inf. Theor.*, 56(5): 2053–2080.

Candès, E.; and Plan, Y. 2010. Matrix Completion With Noise. *Proceedings of the IEEE*, 98: 925 – 936.

Chen, H.; and Li, J. 2017. Learning Multiple Similarities of Users and Items in Recommender Systems. In *2017 IEEE International Conference on Data Mining (ICDM)*, 811–816.

Chen, T.; Zhang, W.; lu, Q.; Chen, K.; Zheng, Z.; and Yu, Y. 2012. SVDFeature: A Toolkit for Feature-based Collaborative Filtering. *The Journal of Machine Learning Research*.

Chen, Y. 2015. Incoherence-Optimal Matrix Completion. *IEEE Transactions on Information Theory*, 61(5): 2909–2923.

Chen, Y.; Chi, Y.; Fan, J.; Ma, C.; and Yan, Y. 2020. Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization.

Chiang, K.-Y.; Dhillon, I. S.; and Hsieh, C.-J. 2018. Using Side Information to Reliably Learn Low-Rank Matrices from Missing and Corrupted Observations. *J. Mach. Learn. Res.*

Chiang, K.-Y.; Hsieh, C.-J.; and Dhillon, I. S. 2015. Matrix Completion with Noisy Side Information. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Fazel, M. 2002. Matrix Rank Minimization with Applications. *PhD Thesis*.

Fazel, M.; Hindi, H.; and Boyd, S. P. 2001. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference. (Cat. No.01CH37148)*, volume 6, 4734–4739 vol.6.

Giménez-Febrer, P.; Pagès-Zamora, A.; and Giannakis, G. B. 2020. Generalization Error Bounds for Kernel Matrix Completion and Extrapolation. *IEEE Signal Processing Letters*, 27: 326–330.

Gross, D.; Liu, Y.-K.; Flammia, S. T.; Becker, S.; and Eisert, J. 2010. Quantum State Tomography via Compressed Sensing. *Phys. Rev. Lett.*, 105: 150401.

Herbster, M.; Pasteris, S.; and Tse, L. 2019. Online Matrix Completion with Side Information. *CoRR*, abs/1906.07255.

Jain, P.; and Dhillon, I. S. 2013. Provable Inductive Matrix Completion. *CoRR*, abs/1306.0626.

Kakade, S. M.; Sridharan, K.; and Tewari, A. 2009. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. In Koller, D.; Schuurmans, D.; Bengio, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 21*, 793–800. Curran Associates, Inc.

Kargas, N.; and Sidiropoulos, N. D. 2019. Learning Mixtures of Smooth Product Distributions: Identifiability and Algorithm. In Chaudhuri, K.; and Sugiyama, M., eds., *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, 388–396. PMLR.

Keshavan, R.; Montanari, A.; and Oh, S. 2009. Matrix Completion from Noisy Entries. In Bengio, Y.; Schuurmans, D.; Lafferty, J. D.; Williams, C. K. I.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 22*, 952–960. Curran Associates, Inc.

Kueng, R.; Rauhut, H.; and Terstiege, U. 2017. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1): 88–116.

Ledent, A.; Alves, R.; and Kloft, M. 2021. Orthogonal Inductive Matrix Completion. *IEEE Transactions on Neural Networks and Learning Systems*, 1–12.

Ledent, A.; Alves, R.; Lei, Y.; and Kloft, M. 2021. Fine-grained Generalization Analysis of Inductive Matrix Completion. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 25540–25552. Curran Associates, Inc.

Li, R.; Dong, Y.; Kuang, Q.; Wu, Y.; Li, Y.; Zhu, M.; and Li, M. 2015. Inductive matrix completion for predicting adverse drug reactions (ADRs) integrating drug–target interactions. *Chemometrics and Intelligent Laboratory Systems*, 144: 71 – 79.

Lin, W.-Y.; Liu, S.; Ren, C.; Cheung, N.-M.; Li, H.; and Matsushita, Y. 2022. Shell Theory: A Statistical Model of Reality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6438–6453.

Mazumder, R.; Hastie, T.; and Tibshirani, R. 2010. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *J. Mach. Learn. Res.*, 11: 2287–2322.

Menon, A. K.; and Elkan, C. 2011. Link Prediction via Matrix Factorization. In *Machine Learning and Knowledge Discovery in Databases*, 437–452. Springer Berlin Heidelberg.

Recht, B. 2011. A Simpler Approach to Matrix Completion. *J. Mach. Learn. Res.*, 12(null): 3413–3430.

Shamir, O.; and Shalev-Shwartz, S. 2011. Collaborative Filtering with the Trace Norm: Learning, Bounding, and Transducing. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, 661–678. PMLR.

Shamir, O.; and Shalev-Shwartz, S. 2014. Matrix Completion with the Trace Norm: Learning, Bounding, and Transducing. *Journal of Machine Learning Research*, 15: 3401–3423.

Shen, W.; Zhang, C.; Tian, Y.; Zeng, L.; He, X.; Dou, W.; and Xu, X. 2021. Inductive Matrix Completion Using Graph Autoencoder. *CoRR*, abs/2108.11124.

Song, L.; Anandkumar, A.; Dai, B.; and Xie, B. 2014. Nonparametric Estimation of Multi-View Latent Variable Models. In Xing, E. P.; and Jebara, T., eds., *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, 640–648. Bejing, China: PMLR.

Steck, H. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In *The World Wide Web Conference*, WWW '19, 3251–3257. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.

Tanner, J.; Thompson, A.; and Vary, S. 2019. Matrix Rigidity and the Ill-Posedness of Robust PCA and Matrix Completion. *SIAM Journal on Mathematics of Data Science*, 1(3): 537–554.

Vančura, V.; Alves, R.; Kasalickỳ, P.; and Kordík, P. 2022. Scalable Linear Shallow Autoencoder for Collaborative Filtering. In *Proceedings of the 16th ACM Conference on Recommender Systems*, 604–609.

Vandermeulen, R. A. 2020. Improving Nonparametric Density Estimation with Tensor Decompositions.

Vandermeulen, R. A.; and Ledent, A. 2021. Beyond Smoothness: Incorporating Low-Rank Analysis into Nonparametric Density Estimation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 12180–12193. Curran Associates, Inc.

Vershynin, R. 2019. High-Dimensional Probability.

Wang, J.; Wong, R. K. W.; Mao, X.; and Chan, K. C. G. 2021. Matrix Completion with Model-free Weighting. arXiv:2106.05850.

Wu, Q.; Zhang, H.; Gao, X.; and Zha, H. 2020. Inductive Collaborative Filtering via Relation Graph Learning.

Xu, M.; Jin, R.; and Zhou, Z.-H. 2013. Speedup Matrix Completion with Side Information: Application to Multi-Label Learning. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, 2301–2309. Red Hook, NY, USA: Curran Associates Inc.

Yao, Q.; and Kwok, J. T. 2019. Accelerated and Inexact Soft-Impute for Large-Scale Matrix and Tensor Completion. *IEEE Transactions on Knowledge and Data Engineering*, 31(9): 1665–1679.

Zhang, M.; and Chen, Y. 2020. Inductive Matrix Completion Based on Graph Neural Networks. In *International Conference on Learning Representations*.

Zhang, X.; Du, S.; and Gu, Q. 2018. Fast and Sample Efficient Inductive Matrix Completion via Multi-Phase Procrustes Flow. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 5756–5765. Stockholmsmässan, Stockholm Sweden: PMLR.

Zhong, K.; Jain, P.; and Dhillon, I. S. 2015. Efficient Matrix Sensing Using Rank-1 Gaussian Measurements. In Chaudhuri, K.; GENTILE, C.; and Zilles, S., eds., *Algorithmic Learning Theory*, 3–18. Cham: Springer International Publishing. ISBN 978-3-319-24486-0.

A Some Concentration Inequalities and Classic Results

Proposition A.1. *Let $\zeta_1, \dots, \zeta_N$ be i.i.d. σ^2-subgaussian random variables, i.e. $\log(\mathbb{E}(\exp(\lambda\zeta))) \leqslant \frac{\lambda^2\sigma^2}{2}$ for all λ. For any $\delta > 0$ we have with probability $\geqslant 1-\delta$:*

$$\sum_{i=1}^{N} \zeta_i^2 \leqslant 32N\sigma^2 \log\left(\frac{1}{\delta}\right). \tag{A.1}$$

Proof. By Theorem 2.1 in (Boucheron, Lugosi, and Bousquet 2004) (page 25) we have

$$\mathbb{E}(\zeta^4) \leqslant 2 \times 2!(2\,\sigma^2)^2 \leqslant 16\,\sigma^4 \qquad \text{and} \qquad \forall q \tag{A.2}$$

$$\mathbb{E}([\zeta^2]^q)) = \mathbb{E}([\zeta^2]_+^q)) \leqslant q![4\,\sigma^2]^q. \tag{A.3}$$

We will now apply Theorem 2.10 (Bernstein's inequality, page 37) from (Boucheron, Lugosi, and Bousquet 2004) to the random variable $\sum_{i=1}^N \zeta_i^2$. By equation (A.2) we have

$$\sum_{i=1}^{N} \mathbb{E}([\zeta_i^2]_+^2) \leqslant 16N\,\sigma^4\,. \tag{A.4}$$

Furthermore, we also have by equation (A.3) (for all $q \geqslant 3$):

$$\sum_{i=1}^{N} \mathbb{E}([\zeta_i^2]_+^q) \leqslant Nq![4\,\sigma^2]^q = q![16N\,\sigma^4][4\,\sigma^2]^{q-2} \tag{A.5}$$

$$\leqslant \frac{q!}{2}[16N\,\sigma^4][8\,\sigma^2]^{q-2}. \tag{A.6}$$

Thus we can apply Theorem 2.10 from (Boucheron, Lugosi, and Bousquet 2004) with "ν" being $[16N\,\sigma^4]$ and "c" being $[8\,\sigma^2]$. We obtain that with probability $\geqslant 1-\delta$ we have

$$\left|\sum_{i=1}^{N} \zeta_i^2 - N\,\sigma^2\right| \leqslant \sqrt{2[16N\,\sigma^4]\log\left(\frac{1}{\delta}\right)} + [8\,\sigma^2]\log\left(\frac{1}{\delta}\right) \leqslant 16\,\sigma^2\log\left(\frac{1}{\delta}\right)\sqrt{N}, \tag{A.7}$$

as expected.

□

Proposition A.2. *Let ζ be a σ^2-subgaussian random variable, i.e. $\log(\mathbb{E}(\exp(\lambda\zeta))) \leqslant \frac{\lambda^2\sigma^2}{2}$ for all λ. We have*

$$\mathbb{V}\mathrm{ar}(\zeta) \leqslant \sigma^2. \tag{A.8}$$

Proof. Cf. Exercise 2.16 on page 49 of (Boucheron, Lugosi, and Bousquet 2004). □

Proposition A.3 (Theorem 2.1 on page 8 of (Bartlett, Bousquet, and Mendelson 2005))**.** *Let $\mathcal{F}$ be a class of functions that maps $\mathcal{X}$ to $[a,b]$. Assume that there is some $r > 0$ such that for all for all $f \in \mathcal{F}$, $\mathbb{V}\mathrm{ar}(f) \leqslant r$. Then for all $\delta > 0$ we have with probability $\geqslant 1-\delta$:*

$$\sup_{f\in\mathcal{F}}\left[\mathbb{E}(f) - \sum_{i=1}^{N} f(x_i)\right] \leqslant \inf_{\alpha>0}\left[2(1+\alpha)\mathbb{E}(\widehat{\mathfrak{R}}_n(\mathcal{F})) + \sqrt{\frac{2r\log\left(\frac{1}{\delta}\right)}{N}} + (b-a)\left(\frac{1}{3}+\frac{1}{\alpha}\right)\frac{\log\left(\frac{1}{\delta}\right)}{N}\right], \tag{A.9}$$

where $\widehat{\mathfrak{R}}_n(\mathcal{F})$ denotes the empirical rademacher complexity of $\mathcal{F}$. Furthermore, the same result holds for $\sup_{f\in\mathcal{F}}\left[\sum_{i=1}^N f(x_i) - \mathbb{E}(f)\right]$.

We recall the following matrix Bernstein inequality. This version is a combination of Lemmas 3 and 4 from (Xu, Jin, and Zhou 2013), but similar results are well known (Vershynin 2019).

Lemma A.4. *Let $X_1, \dots X_L$ be independent, zero mean random matrices with dimensions $m \times n$. Suppose also that for all $k \leqslant L$, $\rho_k^2 := \max\left(\mathbb{E}\left(\|X_k X_k^\top\|\right), \mathbb{E}\left(\|X_k^\top X_k\|\right)\right)$ and $\|X_k\| \leqslant M$ almost surely. Then for all $\delta > 0$ we have with probability $\geqslant 1/\delta$:*

$$\left\|\sum_{k=1}^{L} X_k\right\| \leqslant \max\left(\sqrt{\frac{8}{3}\log\left(\frac{m+n}{\delta}\right)\sum_{k=1}^{L}\rho_k^2},\ \frac{8}{3}M\log\left(\frac{m+n}{\delta}\right)\right). \tag{A.10}$$

Lemma A.5 (Cf also Proposition 3.3 in (Recht 2011))**.** *Assume we sample entries from $[m] \times [n]$ independently and uniformly at random. For any $\delta_5 > 0$, with probability $\geqslant 1 - \delta_5$ the number of repetitions of a single entry is bounded by*

$$\frac{N}{mn} + \max\left(\frac{8}{3}\log\left(\frac{2mn}{\delta_5}\right), \sqrt{\frac{8}{3}\log\left(\frac{2mn}{\delta_5}\right)\frac{N}{mn}}\right) \tag{A.11}$$

$$\leqslant \frac{N}{mn} + \frac{8}{3}\log\left(\frac{2mn}{\delta_5}\right)\sqrt{\frac{N}{mn}} =: \tau_5. \tag{A.12}$$

In particular, as long as $N \leqslant mn$, the number of repetitions is bounded by $\tilde{\tau_5} := 5\log\left(\frac{2mn}{\delta_5}\right) \geqslant \tau_5$.

Proof. Follows from Lemma A.4 applied to each (1×1)-dimensional entry of the matrix, together with a union bound over all $m \times n$ entries. □

Note that classic approximate recovery bounds for inductive matrix completion typically rely on the following result.

Lemma A.6 ((Chiang, Dhillon, and Hsieh 2018))**.** *The function class $\left\{XMY^\top \middle| \|M\|_* \leqslant \mathcal{M}\right\}$ satisfies*

$$\mathfrak{R}(\mathcal{F}_\mathcal{M}) \leqslant \mathbf{x}\mathbf{y}\mathcal{M}\sqrt{\frac{1}{N}}, \tag{A.13}$$

where $\mathbf{x} := \|X^\top\|_{2,\infty}$ and $\mathbf{y} := \|Y^\top\|_{2,\infty}$.

Proof. Follows directly from Theorem 1 in (Kakade, Sridharan, and Tewari 2009), together with the duality between the nuclear and spectral norms (Fazel, Hindi, and Boyd 2001). Cf also (Chiang, Dhillon, and Hsieh 2018; Ledent, Alves, and Kloft 2021). □

B Assumptions and first consequences

Running generic assumptions: recall the following assumptions from the main paper:

Assumption 4 (Realizability)**.** There exists a matrix $M_* \in \mathbb{R}^{a\times b}$ such that $R = XM_*Y^\top$.

Assumption 5 (The noise is ν^2 subgaussian)**.** Recall that we assume the noise is subgaussian: $\mathbb{E}(\zeta) = 0$ and $\log(\mathbb{E}(\exp(\lambda\zeta))) \leqslant \frac{\lambda^2\sigma^2}{2}$ for all λ.

Incoherence assumptions: we will write $\bar{X}$ and $\bar{Y}$ for the matrices obtained by normalizing the columns of X, Y and we will write Σ_1, Σ_2 for the diagonal matrices containing the singular values of X, Y. Similarly we will also write $\bar{\bar{X}} = \bar{X}\Sigma_1$ etc.

We organize our incoherence assumptions slightly differently from the main paper to obtain the most general results possible:

Assumption 6. We make the following assumption on the coherence of the column spaces of X, Y:

$$\|\bar{X}_{i,\bullet}\| \leqslant \sqrt{\frac{\mu a}{m}} \quad (\forall i) \qquad \text{and} \tag{B.1}$$

$$\|\bar{Y}_{j,\bullet}\| \leqslant \sqrt{\frac{\mu b}{n}} \quad (\forall j). \tag{B.2}$$

Assumption 7. We assume we have the following bound on the coherence of the ground truth matrix R: let $ADB^\top$ be the SVD of the core matrix M^*, we have

$$\|\bar{X}\Sigma_1^{-1}AB^\top\Sigma_2^{-1}\bar{Y}\|_\infty \leqslant \sqrt{\frac{\mu_1 r}{mn}}\sigma_0^{-2}. \tag{B.3}$$

Remark: if the columns of X, Y are normalised (as is assumed in (Xu, Jin, and Zhou 2013)), we directly obtain the same assumption as in (Xu, Jin, and Zhou 2013). In the general case, it is not immediately clear how to deduce our assumption from theirs: our assumption requires some form of "joint" incoherence between the side information matrices X, Y and the ground truth core matrix M^*.

Nevertheless, such an assumption can be reasonably expected to hold for many matrices. Indeed, it can be deduced as long as X, Y and M satisfy the stricter notion of incoherence used in the main paper, i.e.

$$\|\bar{X}\|_\infty \leqslant \sqrt{\frac{\mu}{m}}, \qquad \|\bar{Y}\|_\infty \leqslant \sqrt{\frac{\mu}{n}}$$

$$\|A\|_\infty \leqslant \sqrt{\frac{\bar{\mu}}{a}}, \qquad \|B\|_\infty \leqslant \sqrt{\frac{\bar{\mu}}{b}}.$$

In that case

$$\|\bar{X}\Sigma_1^{-1}AB^\top\Sigma_2^{-1}\bar{Y}\|_\infty \leqslant \max_{i,j}\|\bar{X}\Sigma_1^{-1}A\|_{i,\cdot}\|\bar{Y}\Sigma_1^{-1}B\|_{j,\cdot} \tag{B.4}$$

$$\leqslant r\sqrt{a}\sigma_0^{-1}\|\bar{X}\|_\infty\|A\|_\infty\sqrt{b}\sigma_0^{-1}\|\bar{Y}\|_\infty\|B\|_\infty \tag{B.5}$$

$$\leqslant r\sigma_0^{-2}\sqrt{a}\sqrt{\mu/m}\sqrt{\bar{\mu}/a}\sqrt{b}\sqrt{\mu/n}\sqrt{\bar{\mu}/b} \tag{B.6}$$

$$\leqslant r\sigma_0^{-2}\sqrt{\bar{\mu}^2\mu^2\frac{1}{mn}}, \tag{B.7}$$

yielding

$$\mu_1 \leqslant \bar{\mu}^2\mu^2 r. \tag{B.8}$$

Using (B.8) in the lemmas and theorems in this supplementary allows one to obtain the results in the main paper, which assume the strict notion of in coherence above.

Further notation: in addition to the notation introduced in the main paper, we will, similarly to (Xu, Jin, and Zhou 2013), define the following operators: P_X, P_Y. Here $P_X : \mathbb{R}^m \to \mathbb{R}^m$ and $P_Y : \mathbb{R}^n \to \mathbb{R}^n$ are the projection operators onto the column subspaces of X and Y respectively. We assume without loss of generality that the columns of X, Y are orthogonal and ordered with decreasing norm with the norm of the first column being 1 and the norm of the last column being more than σ_0 .

Let $E = \bar{X}A$ and $F = \bar{Y}B$. We will also write (analogously to (Xu, Jin, and Zhou 2013)) P_T for the operator defined as follows

$$P_T(Z) = P_E Z P_Y + P_X Z P_F - P_X Z P_Y,$$

as well as denote the operator $P_{T^\top}$ by

$$P_{T^\top}(Z) = P_{X^\top} Z P_{Y^\top}.$$

We also write P_Ω for the operator from $\mathbb{R}^{m\times n}$ to itself defined by $[P_\Omega(Z)]_{i,j} = h_{i,j}Z_{i,j}$, where $h_{i,j} = \#(k \leqslant N : \xi^k = (i,j))$ denotes the number of times that entry (i,j) was sampled. Note that P_Ω is the sum of N i.i.d. samples from a uniform distribution over the operators $P_{\{(i,j)\}}$ for $(i,j) \in [m] \times [n]$, and it is not necessarily a projection operator (because the same entry can be sampled several times).

Recall the optimization problem considered is the following:

$$\min \frac{1}{N}\sum_{\xi\in\Omega}\left|[R_\xi + \zeta_\xi] - XMY^\top\right|^2 + \lambda\|M\|_*, \tag{B.9}$$

where λ is a regularization parameter.

The norms $\|\cdot\|_{\mathcal{I},*}$ and $\|\cdot\|_{\mathcal{I},\sigma}$: now, in order to better take into account the non homogeneous spectrum of X and Y, we define two norms $\|\cdot\|_{\mathcal{I},*}$ and $\|\cdot\|_{\mathcal{I},\sigma}$ on the space $\mathbb{R}^{m\times n}$:

$$\|Z\|_{\mathcal{I},*} = \min\left(\|M\|_* \quad \middle| XMY^\top = Z\right) = \left\|\overline{\overline{X}}^\top R\bar{\bar{Y}}\right\|_* \tag{B.10}$$

$$\|Z\|_{\mathcal{I},\sigma} = \|X^\top RY\|. \tag{B.11}$$

One of the key aspects of our proof is that the above norms are dual to each other, and therefore the Taylor decomposition around the ground truth still has similar properties as in the case studied in (Xu, Jin, and Zhou 2013).

Lemma B.1. *The norms $\|\cdot\|_{\mathcal{I},*}$ and $\|\cdot\|_{\mathcal{I},\sigma}$ are dual to each other (with respect to the Frobenius inner product on $\mathbb{R}^{m\times n}$).*

Proof. Define $\|\cdot\|_{\mathrm{d}}$ on $\mathbb{R}^{m\times n}$ to be the dual norm to $\|\cdot\|_{\mathcal{I},*}$. We will show that $\|\cdot\|_{\mathrm{d}} = \|\cdot\|_{\mathcal{I},\sigma}$.

Let $B \in \mathbb{R}^{m\times n}$, we have

$$\begin{aligned}
\|B\|_{\mathrm{d}} &:= \sup_{\|A\|_{\mathcal{I},*}=1}\langle A, B\rangle \\
&= \sup_{M\in\mathbb{R}^{a\times b},\|M\|_*=1}\left\langle [XMY^\top], B\right\rangle \\
&= \sup_{M\in\mathbb{R}^{a\times b},\|M\|_*=1}\left\langle M, X^\top BY\right\rangle_{a\times b} \\
&= \|X^\top BY\| \qquad \text{(B.12)} \\
&= \|B\|_{\mathcal{I},\sigma}, \qquad \text{(B.13)}
\end{aligned}$$

where the first line follows by the definition of $\|\cdot\|_{\mathrm{d}}$, the second line follows from the definition of $\|A\|_{\mathcal{I},*}$, the third line follows from direct calculation and the properties of the trace and inner products, the fourth line follows from the duality between the (ordinary) nuclear norm and the (ordinary) spectral norm on the space $\mathbb{R}^{a\times b}$ (cf. e.g. (Fazel 2002)), and the last identity follows from the definition of the norm $\|B\|_{\mathcal{I},\sigma}$. This concludes the proof.

□

Remark: It is worth making a few observations which are necessary to understand the differences between our proofs and the exact recovery proofs in (Xu, Jin, and Zhou 2013), which only apply to the case where the columns of X and Y are normalised.

Let $\bar{X}$ and $\bar{Y}$ be the matrices obtained from X and Y by normalizing the columns. For a matrix R, the canonical way of expressing it as $R = XMY^\top$, is by setting $M = \Sigma_1^{-1}\bar{X}^\top R\bar{Y}\Sigma_2^{-1}$ where Σ_X (resp. Σ_Y) denotes the diagonal matrix whose entries are the norms of the columns of X (resp. Y). We then have $\|R\|_{\mathcal{I},\sigma} = \|M\|_*$. In particular, writing P_X (resp. P_Y) for the projection operator on the space spanned by the columns of X (resp. Y), i.e. $P_X = \bar{X}^\top\bar{X}$ and $P_Y = \bar{Y}^\top\bar{Y}$, we have $\|M\|_* = \|R\|_{\mathcal{I},*} \geqslant \|P_X R P_Y\|_* = \|\bar{M}\|_*$ where $\bar{M}$ is defined by $\bar{X}^\top R\bar{Y}$, and the inequality can be strict (there is equality if the columns of X and Y are both normalized). On the other hand, by definition, we have $\|R\|_{\mathcal{I},\sigma} = \|X^\top RY\|_\sigma = \|\Sigma_1\bar{X}^\top R\bar{Y}\Sigma_2\|_\sigma = \|\Sigma_2^2[\Sigma_1^{-1}\bar{X}^\top R\bar{Y}\Sigma_1^{-1}]\Sigma_2^2\|_\sigma = \|\Sigma_1^2 M\Sigma_2^2\|_\sigma \leqslant \|M\|_\sigma$ (and also $\|R\|_{\mathcal{I},\sigma} = \|\Sigma_1\bar{X}^\top R\bar{Y}\Sigma_2\|_\sigma = \|\Sigma_X\bar{M}\Sigma_2\|_\sigma \leqslant \|\bar{M}\| = \|P_X R P_Y\|$). In other words, directions in R which correspond to columns of X or Y with small norms will make $\|R\|_{\mathcal{I},*}$ large, but $\|R\|_{\mathcal{I},\sigma}$ small, which is reasonable by duality.

It further makes sense intuitively: directions corresponding to small norms for X and Y correspond to a strong prior that the ground truth matrix R doesn't have a significant component in these directions. Thus, if a recovered matrix Z presents with a large component in these directions, then it must have a high $\|\cdot\|_{\mathcal{I},\sigma}$ norm to discourage it. Then, since the optimisation algorithm already discourages such solutions, the values of the components of the dual certificate in those directions is also less important.

C Main technical lemmas

As explained in the main document, we are especially interested in the following optimization problem:

$$\min \frac{1}{N}\sum_{\xi\in\Omega}\left|[R_\xi + \zeta_\xi] - XMY^\top\right|^2 + \lambda\|M\|_*. \tag{C.1}$$

Note first that this optimization problem is trivially equivalent to the following:

$$\min_M \frac{1}{N}\left\|\left[R + \zeta - XMY^\top\right]\circ\mathcal{H}\right\|_{\mathrm{Fr}}^2 + \lambda\|M\|_*, \tag{C.2}$$

where $\circ$ denotes the Hadamard (entry-wise) product between matrices and $\mathcal{H}\in\mathbb{N}^{m\times n}$ is the matrix containing the number of times that each entry was observed. By abuse of notation we write $\zeta\in\mathbb{R}^{m\times n}$ for the matrix whose (i,j) entry is the average of the noise of the observations corresponding to entry (i,j) (for unobserved entries we set $\zeta_{i,j}$ to zero).

Let $\widehat{R}$ be the solution to the constrained optimization problem (C.1) . We will write $\widehat{R} = Z + H + R\in\mathbb{R}^{m\times n}$ with R being the ground truth, $H\in P_\Omega(\mathbb{R}^{m\times n})$ and $Z\in P_{\Omega^\top}(\mathbb{R}^{m\times n})$. Our strategy is to show that if the noise matrix ζ is small enough and if the number of samples is large enough, the dual certificate of R with respect to the norm $\|\cdot\|_{\mathcal{I},*}$ can be well captured by a matrix in the span of the observed entries, which allows us to show that Z must have small nuclear norm, from which it later follows that H also has low nuclear norm.

Lemma C.1. *Let $\tau > 0$.*

Assume that the regularization parameter λ satisfies

$$\frac{\sigma\,\sigma_0^2}{C\sqrt{aN}} \leqslant \lambda \leqslant \frac{C\,\sigma\,\sigma_0^2}{\sqrt{aN}}, \tag{C.3}$$

for some constant C and that Z satisfies

$$\|P_T(Z)\|_{\mathrm{Fr}} \leqslant \sqrt{\frac{3\tau a}{r}}\|P_{T^\top}(Z)\|_{\mathrm{Fr}} \tag{C.4}$$

for some given $\tau > 0$. Assume also that

$$\|\zeta\|_\infty \leqslant B\,\sigma/\sqrt{\kappa} \tag{C.5}$$

for some $B > 1$.

Now, let U be a dual certificate of $R = P_T(R)$ with respect to the norm $\|\cdot\|_{\mathcal{I},*}$, which is to say that $\|\mathrm{U}\|_{\mathcal{I},\sigma} = 1$ and $\langle R, \mathrm{U}\rangle = \|R\|_{\mathcal{I},*}$. Similarly, let $\mathrm{U}^\top$ be a dual certificate for $P_{T^\top}(Z)$. Assume also that there exists a $\mathcal{Y}$ in the image of P_Ω such that

$$\|P_T(\mathcal{Y}) - \mathrm{U}\|_{\mathrm{Fr}} \leqslant \frac{1}{4}\sqrt{\frac{r}{3a\tau}}\frac{1}{\sigma_0^{-2}} \tag{C.6}$$

and

$$\|P_{T^\top}(\mathcal{Y})\|_{\mathcal{I},\sigma} \leqslant \frac{1}{2}. \tag{C.7}$$

(Here, as usual, σ_0 denotes the smallest singular value of X or Y.)

We have

$$\|Z\|_* \leqslant 32CaB^2\sigma_0^{-2}\,\sigma\sqrt{3\tau N/\kappa}. \tag{C.8}$$

Furthermore, we also have

$$\|H\|_* \leqslant 6C\sqrt{a}B\,\sigma\sqrt{N/\kappa}. \tag{C.9}$$

In particular, this implies that

$$\left\|\widehat{R}-R\right\|_* = \|H\|_* + \|Z\|_* \leqslant 70CaB^2\sigma_0^{-2}\,\sigma\sqrt{\tau N/\kappa}. \tag{C.10}$$

Proof. Since $\widehat{R}$ is the solution to the optimization problem (C.2), we have

$$\lambda\|\widehat{R}\|_{\mathcal{I},*} + \frac{1}{N}\|P_\Omega(\widehat{R}-R-\zeta)\|_{\mathrm{Fr}}^2 \leqslant \lambda\|R\|_{\mathcal{I},*} + \frac{1}{N}\|\zeta\circ\mathcal{H}\|_{\mathrm{Fr}}^2. \tag{C.11}$$

Note that below by abuse of notation we write $\|P_\Omega(\zeta)\|_{\mathrm{Fr}}^2$ for $\sum_{o=1}^N \zeta_o^2 = \|\zeta\circ\mathcal{H}\|_{\mathrm{Fr}}^2$ (even when some entries have been sampled several times).

Now note that we have

$$\begin{aligned}
&\|\widehat{R}\|_{\mathcal{I},*}\\
&\geqslant \left\langle R+H+Z, \mathrm{U}+\mathrm{U}^\top\right\rangle\\
&\geqslant \|R\|_{\mathcal{I},*} + \left\langle H, \mathrm{U}+\mathrm{U}^\top\right\rangle + \left\langle Z, \mathrm{U}+\mathrm{U}^\top\right\rangle\\
&\geqslant \|R\|_{\mathcal{I},*} + \left\langle H, \mathrm{U}+\mathrm{U}^\top\right\rangle + \left\langle Z, -P_T(\mathcal{Y}) - P_{T^\top}(\mathcal{Y}) + \mathrm{U}+\mathrm{U}^\top\right\rangle && \text{(C.12)}\\
&\geqslant \|R\|_{\mathcal{I},*} + \left\langle H, \mathrm{U}+\mathrm{U}^\top\right\rangle + \langle Z, \mathrm{U} - P_T(\mathcal{Y})\rangle + \left\langle Z, \mathrm{U}^\top - P_{T^\top}(\mathcal{Y})\right\rangle\\
&\geqslant \|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} + \langle Z, \mathrm{U} - P_T(\mathcal{Y})\rangle + \left\langle Z, \mathrm{U}^\top - P_{T^\top}(\mathcal{Y})\right\rangle && \text{(C.13)}\\
&\geqslant \|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} - \|P_T(Z)\|_{\mathrm{Fr}}\|\,\mathrm{U} - P_T(\mathcal{Y})\|_{\mathrm{Fr}} + \left\langle Z, \mathrm{U}^\top - P_{T^\top}(\mathcal{Y})\right\rangle && \text{(C.14)}\\
&\geqslant \|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} - \|P_T(Z)\|_{\mathrm{Fr}}\|\,\mathrm{U} - P_T(\mathcal{Y})\|_{\mathrm{Fr}} + \|P_{T^\top}(Z)\|_{\mathcal{I},*} - \|P_{T^\top}(Z)\|_{\mathcal{I},*}\|P_{T^\top}(\mathcal{Y})\|_{\mathcal{I},\sigma} && \text{(C.15)}\\
&\geqslant \|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} - \|P_T(Z)\|_{\mathrm{Fr}}\|\,\mathrm{U} - P_T(\mathcal{Y})\|_{\mathrm{Fr}} + \|P_{T^\top}(Z)\|_{\mathcal{I},*}\left[1 - \|P_{T^\top}(\mathcal{Y})\|_{\mathcal{I},\sigma}\right]\\
&> \|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} - \|P_T(Z)\|_{\mathrm{Fr}}\frac{1}{4}\sqrt{\frac{r}{3a\tau}}\frac{1}{\sigma_0^{-2}} + \frac{1}{2}\|P_{T^\top}(Z)\|_{\mathcal{I},*} && \text{(C.16)}\\
&> \|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} - \sqrt{\frac{3\tau a}{r}}\|P_{T^\top}(Z)\|_{\mathrm{Fr}}\frac{1}{4}\sqrt{\frac{r}{3a\tau}}\frac{1}{\sigma_0^{-2}} + \frac{1}{2}\|P_{T^\top}(Z)\|_{\mathcal{I},*} && \text{(C.17)}\\
&> \|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} - \frac{1}{4\sigma_0^{-2}}\|P_{T^\top}(Z)\|_{\mathrm{Fr}} + \frac{1}{2}\|P_{T^\top}(Z)\|_{\mathcal{I},*} && \text{(C.18)}\\
&\geqslant \|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} + \frac{1}{4}\|P_{T^\top}(Z)\|_{\mathcal{I},*}, && \text{(C.19)}
\end{aligned}$$

where at equation (C.12), we have used the fact that $\langle Z, \mathcal{Y}\rangle = 0$ (since $P_\Omega(Z) = 0$ and $P_\Omega(\mathcal{Y}) = \mathcal{Y}$), at equation (C.13) we have used the fact that $\|\,\mathrm{U}\,\|_{\mathcal{I},\sigma} = 1 = \|\,\mathrm{U}^\top\|_{\mathcal{I},\sigma} = 1$, at equation (C.14) we have used the duality between the norms $\|\cdot\|_{\mathcal{I},\sigma}$ and $\|\cdot\|_{\mathcal{I},*}$, at equation (C.15) we have used the duality and the definition of $\mathrm{U}^\top$, at equation (C.16) we have used the assumptions on $\mathcal{Y}$ from the Lemma statement, at equation (C.17) we have used equation (C.4), at equation (C.18) we have simply simplified the expression, and at equation (C.19) we have used the assumptions on Z as well as the fact $\|P_{T^\top}(Z)\|_{\mathrm{Fr}} \leqslant \sigma_0^{-2}\|P_{T^\top}(Z)\|_{\mathcal{I},*}$.

From the above equation together with equation (C.11) we obtain:

$$\begin{aligned}
\lambda\|R\|_{\mathcal{I},*} + \frac{1}{N}\|P_\Omega(\zeta)\|_{\mathrm{Fr}}^2 &\geqslant \lambda\|\widehat{R}\|_{\mathcal{I},*} + \frac{1}{N}\|P_\Omega(\widehat{R}-R-\zeta)\|_{\mathrm{Fr}}^2 && \text{(C.20)}\\
&\geqslant \lambda\left[\|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} + \frac{1}{4}\|P_{T^\top}(Z)\|_{\mathcal{I},*}\right] + \frac{1}{N}\|P_\Omega(\widehat{R}-R-\zeta)\|_{\mathrm{Fr}}^2, && \text{(C.21)}
\end{aligned}$$

from which it follows that

$$\|P_{T^\top}(Z)\|_{\mathcal{I},*} \leqslant 8\|H\|_{\mathcal{I},*} + \frac{4}{\lambda N}\left[\|P_\Omega(\zeta)\|_{\mathrm{Fr}}^2 - \|P_\Omega(H-\zeta)\|_{\mathrm{Fr}}^2\right]. \tag{C.22}$$

Note also that

$$\begin{aligned}
\|H\|_{\mathcal{I},*} &= \|(H-\zeta)+\zeta\|_{\mathcal{I},*} \leqslant \|(H-\zeta)\|_{\mathcal{I},*} + \|\zeta\|_{\mathcal{I},*} && \text{(C.23)}\\
&\leqslant \sigma_0^{-2}\sqrt{a}\|(H-\zeta)\|_{\mathrm{Fr}} + \|\zeta\|_{\mathcal{I},*} && \text{(C.24)}\\
&\leqslant \sigma_0^{-2}\sqrt{a/\kappa}\|(H-\zeta)\circ\mathcal{H}\|_{\mathrm{Fr}} + \|\zeta\|_{\mathcal{I},*}. && \text{(C.25)}
\end{aligned}$$

Thus we can further write

$$\begin{aligned}
\|P_{T^\top}(Z)\|_{\mathcal{I},*} &\leqslant 8\|H\|_{\mathcal{I},*} + \frac{4}{\lambda N}\left[\|P_\Omega(\zeta)\|_{\mathrm{Fr}}^2 - \|P_\Omega(H-\zeta)\|_{\mathrm{Fr}}^2\right] && \text{(C.26)}\\
&\leqslant 8\left[\sigma_0^{-2}\sqrt{a/\kappa}\|(H-\zeta)\circ\mathcal{H}\|_{\mathrm{Fr}} + \|\zeta\|_{\mathcal{I},*}\right] + \frac{4}{\lambda N}\left[\|P_\Omega(\zeta)\|_{\mathrm{Fr}}^2 - \|P_\Omega(H-\zeta)\|_{\mathrm{Fr}}^2\right] && \text{(C.27)}\\
&\leqslant 8\|\zeta\|_{\mathcal{I},*} + \upsilon\left[8\sigma_0^{-2}\sqrt{a/\kappa} - \frac{4}{\lambda N}\upsilon\right] + \frac{4}{\lambda N}\|\mathcal{H}\circ\zeta\|_{\mathrm{Fr}}^2, && \text{(C.28)}
\end{aligned}$$

where we have used the notation $\upsilon := \|[H-\zeta]\circ\mathcal{H}\|_{\mathrm{Fr}}$.

Now, note that the maximum of the function $\upsilon(b-a\upsilon)$ is $\frac{b^2}{4a}$ (attained at $\frac{b}{2a}$). Thus, optimising the above over υ we obtain:

$$\begin{aligned}
\|P_{T^\top}(Z)\|_{\mathcal{I},*} &\leqslant 8\|\zeta\|_{\mathcal{I},*} + \upsilon\left[8\sigma_0^{-2}\sqrt{a/\kappa} - \frac{4}{\lambda N}\upsilon\right] + \frac{4}{\lambda N}\|\mathcal{H}\circ\zeta\|_{\mathrm{Fr}}^2 && \text{(C.29)}\\
&\leqslant 8\|\zeta\|_{\mathcal{I},*} + 4a\sigma_0^{-4}\lambda N/\kappa + \frac{4}{\lambda N}\|\mathcal{H}\circ\zeta\|_{\mathrm{Fr}}^2 && \text{(C.30)}\\
&\leqslant 8\|\zeta\|_{\mathcal{I},*} + 4a\sigma_0^{-4}\lambda N/\kappa + \frac{4\,\sigma^2\,B^2}{\lambda\kappa} && \text{(C.31)}\\
&\leqslant 8\sigma_0^{-2}B\,\sigma\,\sqrt{aN/\kappa} + 4a\sigma_0^{-4}\lambda N/\kappa + \frac{4\,\sigma^2\,B^2}{\lambda\kappa} && \text{(C.32)}\\
&\leqslant 8\sigma_0^{-2}B\,\sigma\,\sqrt{aN/\kappa} + 4C\sigma_0^{-2}\,\sigma\,\sqrt{aN}/\kappa\left[1+B^2\right] \leqslant 16CB^2\sigma_0^{-2}\,\sigma\,\sqrt{aN/\kappa}, && \text{(C.33)}
\end{aligned}$$

where at lines (C.31) and (C.32) we have used the condition that $\|\zeta\|_\infty \leqslant B\,\sigma/\sqrt{\kappa}$, at line (C.32) we have used the fact that $\|\zeta\|_{\mathcal{I},*} \leqslant \sigma_0^{-2}\|\zeta\|_* \leqslant \sigma_0^{-2}\sqrt{a}\|\zeta\|_{\mathrm{Fr}} \leqslant \sigma_0^{-2}B\sqrt{a}\,\sigma\,\sqrt{N/\kappa}$ and at the last line (C.33) we have used the condition (C.3) on λ, the fact that $\kappa \geqslant 1$ and Lemma C.3.

From the above equation we obtain that

$$\|P_{T^\top}(Z)\|_* \leqslant \|P_{T^\top}(Z)\|_{\mathcal{I},*} \leqslant 16CB^2\sigma_0^{-2}\,\sigma\,\sqrt{aN/\kappa}. \tag{C.34}$$

With one more use of the assumptions on Z, we obtain

$$\|P_T(Z)\|_{\mathrm{Fr}} \leqslant \sqrt{\frac{3\tau a}{r}}\|P_{T^\top}(Z)\|_{\mathrm{Fr}} \leqslant \sqrt{\frac{3\tau a}{r}}\|P_{T^\top}(Z)\|_* \leqslant \sqrt{3a\tau/r}\|P_{T^\top}(Z)\|_{\mathcal{I},*} \leqslant 16Ca\sqrt{3\tau/r}B^2\sigma_0^{-2}\,\sigma\,\sqrt{N/\kappa}. \tag{C.35}$$

Then

$$\|P_T(Z)\|_* \leqslant \sqrt{r}\|P_T(Z)\|_{\mathrm{Fr}} \leqslant 16CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{3\tau N/\kappa}. \tag{C.36}$$

Putting equations (C.34) and (C.36) together, we obtain the result (C.8).

Regarding the bound on $\|H\|_{\mathcal{I},*}$, note that by equation (C.20)

$$\kappa\|H-\zeta\|_{\mathrm{Fr}}^2 \leqslant \|[H-\zeta]\circ\mathcal{H}\|_{\mathrm{Fr}}^2 \leqslant \|\zeta\circ\mathcal{H}\|_{\mathrm{Fr}}^2 + \lambda N\left[\|R\|_{\mathcal{I},*} - \|\widehat{R}\|_{\mathcal{I},*}\right] \leqslant \kappa\|\zeta\|_{\mathrm{Fr}}^2 + \lambda N\left[\|R\|_{\mathcal{I},*} - \|\widehat{R}\|_{\mathcal{I},*}\right]. \tag{C.37}$$

Hence

$$\kappa\|H-\zeta\|_{\mathrm{Fr}}^2 \tag{C.38}$$
$$\leqslant \kappa\|\zeta\|_{\mathrm{Fr}}^2+\lambda N\left[\|R\|_{\mathcal{I},*}-\|\widehat{R}\|_{\mathcal{I},*}\right] \tag{C.39}$$
$$\leqslant \kappa\|\zeta\|_{\mathrm{Fr}}^2+\lambda N\left[\|H\|_{\mathcal{I},*}+\|Z\|_{\mathcal{I},*}\right] \tag{C.40}$$
$$\leqslant \kappa\|\zeta\|_{\mathrm{Fr}}^2+\lambda N\left[\|\zeta\|_{\mathcal{I},*}+\|H-\zeta\|_{\mathcal{I},*}+\|Z\|_{\mathcal{I},*}\right] \tag{C.41}$$
$$\leqslant \kappa NB^2\,\sigma^2/\kappa+\lambda N\sigma_0^{-2}\sqrt{a/\kappa}\sqrt{NB^2\,\sigma^2}+\lambda N32Ca^{3/2}B^2\sigma_0^{-2}\,\sigma\sqrt{N\kappa}+\lambda N\sqrt{a}\sigma_0^{-2}\|H-\zeta\|_{\mathcal{I},*} \tag{C.42}$$
$$\leqslant NB^2\,\sigma^2+\frac{C\sqrt{N\kappa}\,\sigma\,\sigma_0^2}{\sqrt{a}}\sigma_0^{-2}\sqrt{a/\kappa}\sqrt{NB^2\,\sigma^2}+\frac{C\sqrt{N\kappa}\,\sigma\,\sigma_0^2}{\sqrt{a}}32Ca^{3/2}B^2\sigma_0^{-2}\,\sigma\sqrt{N\kappa}+\frac{C\sqrt{N\kappa}\,\sigma\,\sigma_0^2}{\sqrt{a}}\|H-\zeta\|_{\mathcal{I},*} \tag{C.43}$$
$$\leqslant NB^2\,\sigma^2+CNB\,\sigma^2+32C^2N\,\sigma^2\,aB^2+\frac{C\sqrt{N\kappa}\,\sigma\,\sigma_0^2}{\sqrt{a}}\|H-\zeta\|_{\mathcal{I},*} \tag{C.44}$$
$$\leqslant 65C^2NB^2\,\sigma^2+\frac{C\sqrt{N\kappa}\,\sigma\,\sigma_0^2}{\sqrt{a}}\|H-\zeta\|_{\mathcal{I},*} \tag{C.45}$$
$$\leqslant 65C^2NB^2\,\sigma^2+C\sqrt{N\kappa}\,\sigma\,\|H-\zeta\|_{\mathrm{Fr}}, \tag{C.46}$$

where at equation (C.42) we have used equation (C.5) as well as equation (C.8), at equation (C.43) we have used the conditions on λ (i.e. inequalities (C.3)), at line (C.45) we have used the fact that $B>1$, and at the last line we have used comparisons between different norms.

From this it follows that

$$\|H-\zeta\|_{\mathrm{Fr}} \leqslant C\sqrt{N\kappa}\,\sigma/2\kappa+\sqrt{C^2N\kappa\,\sigma^2+65C^2NB^2\,\sigma^2\,\kappa}/2\kappa \tag{C.47}$$
$$\leqslant 5CB\,\sigma\sqrt{N/\kappa}, \tag{C.48}$$

from which it follows that

$$\|H\|_*\leqslant 5C\sqrt{a}B\,\sigma\sqrt{N/\kappa}+\|\zeta\|_*\leqslant 6C\sqrt{a}B\,\sigma\sqrt{N/\kappa}, \tag{C.49}$$

as expected.

□

Lemma C.2. *For any $\delta>0$, with probability $\geqslant 1-(5q+1)\delta$ as long as $N\geqslant q\bar{T}$, the conditions* (C.4)*,* (C.6) *and* (C.7) *of Lemma C.1 hold (with $\tau=\frac{N}{mn}+\frac{8}{3}\log\left(\frac{2mn}{\delta}\right)\sqrt{\frac{N}{mn}}$).*

Here $\bar{T}=4\frac{\mu_1}{\mu}\sigma_0^{-4}T=\frac{128}{3}\mu\mu_1 r(a+b)\log\left(\frac{2mn}{\delta}\right)$, and $q=\log\left[e^6\sigma_0^8 a\log\left(\frac{mn}{\delta}\right)\right]$.

Proof. The proof will be provided below in Section H. □

Lemma C.3. *Let $a,b>0$ the minimum value of the function*

$$\lambda a+\frac{b}{\lambda} \tag{C.50}$$

over $\lambda>0$ is equal to $2\sqrt{ab}$, realised at $\lambda=\sqrt{\frac{b}{a}}$. Furthermore, as long as

$$\frac{1}{C}\sqrt{\frac{b}{a}}\leqslant\lambda\leqslant C\sqrt{\frac{b}{a}}, \tag{C.51}$$

we have

$$\lambda a+\frac{b}{\lambda}\leqslant 2C\sqrt{ab}. \tag{C.52}$$

Proof. The result is standard and a trivial application of standard calculus. The derivative of the expression (C.52) is $a-\frac{b}{\lambda^2}$ which only cancels at $\lambda=\sqrt{b/a}$ as expected. As for the second statement, each term in the expression (C.52) is bounded by twice its value for $\lambda=\sqrt{b/a}$.

□

Lemma C.4. *For any $\delta>0$, with probability greater than $1-\delta$ all $(i,j)\in\Omega$, $|\zeta_{(i,j)}|\leqslant\frac{\sigma}{\sqrt{\kappa}}\sqrt{2\log(2N/\delta)}$.*

Proof. This follows immediately from a simple union bound applied to our subgaussianity assumption on the noise. □

Lemma C.5. *With probability* $\geqslant 1 - Kmn\exp(-\frac{N}{2Kmn})$*, each entry is sampled at least* K *times.*

Proof. Let $N_1 = \left\lfloor \frac{N}{K} \right\rfloor$, and let us divide the first KN_1 samples into K different groups $\{g_1, \dots, g_K\}$.

The probability that any fixed entry (i, j) is not sampled in group g_k (for any k) is

$$\left(1 - \frac{1}{mn}\right)^{N_1} \leqslant \exp\left(-\frac{N_1}{mn}\right).$$

Thus the probability that there is at least one entry which is not sampled in group k is less than $mn\exp\left(-\frac{N_1}{mn}\right)$.

By a union bound over all the groups, we get that with probability $\geqslant 1 - Kmn\exp\left(-\frac{N_1}{mn}\right)$, each entry is sampled at least once in each group (and in particular is sampled at least K times over all). The result follows since $N_1 = \left\lfloor \frac{N}{K} \right\rfloor \geqslant \frac{1}{2}\frac{N}{K}$.

□

D Main results

Theorem D.1. *Assume that the entries are observed without noise. For any* $\delta > 0$ *as long as*

$$N \geqslant \log\left[e^6\sigma_0^8 a\log\left(\frac{mn}{\delta}\right)\right]\sigma_0^{-4}\frac{128}{3}\mu\mu_1 r(a+b)\log\left(\frac{2mn}{\delta}\right),$$

with probability

$$\geqslant 1 - \delta\left[1 + 5\log\left[e^6\sigma_0^{-8}a\log\left(\frac{mn}{\delta}\right)\right]\right],$$

we have that $XM_{\min}Y^\top = R$ *where* $M_{\min}$ *is a solution to the following optimization problem:*

$$M_{\min} \in \arg\min\left(\|M\|_* \quad \textit{s.t.} \quad \forall (i,j) \in \Omega, [XMY^\top]_{i,j} = R_{i,j}\right). \tag{D.1}$$

Proof. **Informal:** The condition (C.3) is trivially respected since $\sigma = 0$ and $\lambda = 0$. Taking the limit as $\lambda \to 0$, the theorem follows immediately from Lemma C.2 together with Lemma C.1 upon noting that in this case, the lack of noise implies that condition (C.5) holds with $B = 0$, which shows $Z = 0$ as expected.

Formal: By Lemma C.2, the conditions (C.4), (C.6) and (C.7) of Lemma C.1 hold (with $\tau = \frac{N}{mn} + \frac{8}{3}\log\left(\frac{2mn}{\delta}\right)\sqrt{\frac{N}{mn}}$). Thus we can write, as in the proof of lemma C.1 (equation (C.19)):

$$\|\widehat{R}\|_{\mathcal{I},*} \geqslant \|R\|_{\mathcal{I},*} - 2\|H\|_{\mathcal{I},*} + \frac{1}{4}\|P_{T^\top}(Z)\|_{\mathcal{I},*} \tag{D.2}$$

$$\geqslant \|R\|_{\mathcal{I},*} + \frac{1}{4}\|P_{T^\top}(Z)\|_{\mathcal{I},*}. \tag{D.3}$$

Indeed, the relevant calculation in lemma C.1 doesn't rely on the optimization problem or the value of λ, and at the second line we have simply used the fact that by definition of the optimization problem (D.1), $H = 0$.

Now, by definition of the optimization problem (D.1) we have $\|\widehat{R}\|_{\mathcal{I},*} \leqslant \|R\|_{\mathcal{I},*}$, which together with equation (D.2) implies that $\|P_{T^\top}(Z)\|_{\mathcal{I},*} = 0$, which implies $P_{T^\top}(Z) = 0$ and of course $\|P_{T^\top}(Z)\|_{\mathrm{Fr}} = 0$. Then by equation (C.4) (satisfied as stated above by Lemma C.2), we also have $\|P_T(Z)\|_{\mathrm{Fr}} = 0$, from which we finally obtain that $Z = 0$ as expected. □

Remarks:

1. There is a dependence on the conditioning number of X, Y both logarithmic in the high probability and non logarithmic in the threshold for N. The quadratic dependence matches that in (Jain and Dhillon 2013) (though they relied on a different optimization problem so the questions are different).
2. It is worth unpacking the log terms as well. We have a logarithmic term in $\log(1/\delta)$ (i.e. $\log\log(1/\delta)$) both in the failure probability and in the threshold for N. This comes from the definition of τ from Lemma A.5 (which controls the number of times the same entry can be sampled), which is necessary in lemma G.3. Note that a similar term is also present (implicitly) in the result of (Recht 2011) (cf. the logarithm in the term "$6\log(n_2)(n_1+n_2)^{2-2\beta}$" in the main theorem on page 2 (equation 1.3)). Similarly to our result, the log term comes from the use of proposition 3.3 on page 5 (which plays a similar role to our Lemma A.5), and is used later in the proof of the main Theorem (1.1) on page 8.
3. Contrary to (Recht 2011), we do not assume that $N \leqslant mn$. This results in extra logarithmic factors in the definition of τ from the use of Lemma C.5, which only show up as constants in the end after taking the logarithm of τ. This, together with our rather loose bounding of τ (aimed at a compact formula rather than the tightest bound) explains the higher implicit constant from the factor $\log(e^6\dots)$.

For the next theorem, we consider an arbitrary loss function ℓ which is assumed to be bounded by B_ℓ, and L_ℓ −Lipschitz continuous.

Theorem D.2. *Assume that condition* (C.3) *on* λ *holds. For any* $\delta_0, \delta_1, \delta_2, \delta > 0$*, with probability*

$$\geqslant 1 - \delta_1 - \delta_0 - \delta_2 - \delta \left[1 + 5 \log \left[e^6 \sigma_0^{-8} a \log\left(\frac{mn}{\delta}\right)\right]\right],$$

as long as

$$N \geqslant \log \left[e^6 \sigma_0^{-8} a \log\left(\frac{mn}{\delta}\right)\right] \frac{128}{3} \mu\mu_1 r(a+b) \log\left(\frac{2mn}{\delta}\right),$$

we have

$$\mathbb{E}_{(i,j)\sim\mathcal{U}}(\ell(\widehat{R}_{(i,j)}, [R+\zeta]_{(i,j)})) \leqslant \qquad 500C\,\mathrm{L}_\ell\, a^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\sqrt{\theta}\log\left(\frac{2N}{\delta_0}\right)\sqrt{\frac{\log(\frac{N}{2\delta_2})}{N}} + \mathrm{B}_\ell\,\frac{4\log(1/\delta_1)}{3N}, \tag{D.4}$$

where, as in the main paper, $\mathcal{U}$ *is the uniform distribution on the entries* $[m] \times [n]$ *and* θ *denotes the logarithmic quantity*

$$\theta = 2\log\left(\frac{N}{2\delta_2}\right) + \frac{8}{3}\log\left(\frac{2mn}{\delta_2}\right)\sqrt{2\log\left(\frac{N}{2\delta_2}\right)}. \tag{D.5}$$

Proof. If $N \leqslant 2$, the result holds trivially. If $N \geqslant 3$, let

$$K_{\delta_2} = K = \left\lfloor \frac{N}{2mn\log(\frac{N}{2\delta_2})} \right\rfloor. \tag{D.6}$$

(In particular, if $N \leqslant mn$, we have $K = 0$). Let also $\kappa := \max(K, 1)$. Note that we have

$$Kmn\exp\left(-\frac{N}{2Kmn}\right) \leqslant \frac{N}{2mn\log(\frac{N}{2\delta_2})} mn\exp\left(-\frac{N}{2Kmn}\right) \tag{D.7}$$

$$\leqslant \frac{N}{2\log(\frac{N}{2\delta_2})}\exp\left(-\frac{N}{\frac{N}{\log(\frac{N}{2\delta_2})}}\right) \tag{D.8}$$

$$\leqslant \frac{\delta_2}{\log(\frac{N}{2\delta_2})} \leqslant \delta_2. \tag{D.9}$$

Hence, by Lemma C.5, we have that each entry is sampled at least K times with probability $\geqslant 1 - \delta_2$, and below, we restrict ourselves to the high probability event where this occurs.

Then, by Lemma C.4 we have with probability $\geqslant 1 - \delta_0$ that

$$\max_{(i,j)\in\Omega} |\zeta_{(i,j)}| \leqslant \frac{\sigma}{\sqrt{\kappa}}\sqrt{2\log(2N/\delta_0)}, \tag{D.10}$$

which means that condition (C.5) holds with $B = \sqrt{2\log(2N/\delta_0)}$ (note that since $\delta_0 \leqslant 1$, $B > 1$ as required). The condition (C.3) also holds by assumption, and we can use Theorem D.1 with the value of κ defined as above ($\max(K, 1)$).

Furthermore by Lemma (C.2) we have with probability $\geqslant 1 - \delta\left[1 + 5\log\left[e^6\sigma_0^{-8} a \log\left(\frac{mn}{\delta}\right)\right]\right]$ that the conditions (C.4), (C.6) and (C.7) of Lemma C.1 hold (with $\tau = \frac{N}{mn} + \frac{8}{3}\log\left(\frac{2mn}{\delta}\right)\sqrt{\frac{N}{mn}}$). Hence by Lemma (C.1), on the same high-probability event (w.p. $\geqslant 1 - \delta_0 - \delta_2 - \delta\left[1 + 5\log\left[e^6\sigma_0^{-8} a\log\left(\frac{mn}{\delta}\right)\right]\right]$), we have

$$\left\|\widehat{R} - R\right\|_* \leqslant 70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\tau N/\kappa}. \tag{D.11}$$

Now, by Lemma A.6, the Rademacher complexity of the function class

$$\mathcal{F}_1 := \left\{R + \bar{X}M\bar{Y}^\top \middle| \|M\|_* \leqslant 70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\tau N/\kappa}\right\} = \left\{R + Z \middle| \|Z\|_* \leqslant\ 70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\tau N/\kappa}\right\}$$

is bounded as

$$\mathfrak{R}(\mathcal{F}_1) \leqslant \frac{1}{\sqrt{N}}\mu\sqrt{\frac{ab}{mn}}70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\tau N/\kappa} = \mu\sqrt{\frac{ab}{\kappa mn}}70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\tau}. \tag{D.12}$$

Thus, by proposition A.3, with probability $\geqslant 1-\delta_1$ we have that for any $Z \in \mathcal{F}_1$ (and for any $\alpha > 0$)

$$\mathbb{E}_{(i,j)\sim\mathcal{U}}(\ell(Z_{(i,j)}, [R+\zeta]_{(i,j)})) - \frac{1}{N}\sum_{(i,j)\in\Omega} \ell(Z_{(i,j)}, [R+\zeta]_{(i,j)}) \tag{D.13}$$

$$\leqslant 2(1+\alpha)\,\mathrm{L}_\ell\,\mu\sqrt{\frac{ab}{\kappa mn}}70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\tau} + \mathrm{L}_\ell\sqrt{\frac{1}{mn}}70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\tau N/\kappa}\sqrt{2\frac{\log(1/\delta_1)}{N}} \tag{D.14}$$

$$+ \mathrm{B}_\ell[1/3 + 1/\alpha]\frac{\log(1/\delta_1)}{N}, \tag{D.15}$$

where we have used the bound on the Lipschitz constant, Lemma A.6 and fact that the "variance" r is bounded as

$$r := \frac{1}{mn}\sum_{(i,j)\in[m]\times[n]} \ell^2(Z_{(i,j)}, [R+\zeta]_{(i,j)}))) \leqslant {\mathrm{L}_\ell}^2\,\frac{1}{mn}\|Z-R\|_{\mathrm{Fr}}^2 \leqslant \frac{1}{mn}\|Z-R\|_*^2, \tag{D.16}$$

together with another use of the result from Lemma C.1.

Now, by equation (D.11), $\widehat{R} \in \mathcal{F}_1$. Hence, (w.p. $\geqslant 1-\delta_1-\delta_2-\delta_0-5q\delta$) we can apply equation (D.13) to $\widehat{R}$. Furthermore, since obviously $R \in \mathcal{F}_1$ we certainly have

$$\frac{1}{N}\sum_{(i,j)\in\Omega} \ell(\widehat{R}_{(i,j)}, [R+\zeta]_{(i,j)}) \leqslant \frac{1}{N}\sum_{(i,j)\in\Omega} \ell(R_{i,j}, [R+\zeta]_{(i,j)}) \leqslant \mathrm{L}_\ell\,B\,\sigma\,.$$

Hence we can write (w.p. $\geqslant 1-\delta_1-\delta_2-\delta_0-5q\delta$), taking $\alpha = 1$,

$$\mathbb{E}_{(i,j)\sim\mathcal{U}}(\ell(\widehat{R}_{(i,j)}, [R+\zeta]_{(i,j)})) \leqslant 4\,\mathrm{L}_\ell\,\mu\sqrt{\frac{ab}{\kappa mn}}70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\tau} + \mathrm{L}_\ell\sqrt{\frac{1}{\kappa mn}}70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\tau}\sqrt{2\log(1/\delta_1)} \tag{D.17}$$

$$+ \mathrm{B}_\ell\,\frac{4}{3}\frac{\log(1/\delta_1)}{N} \tag{D.18}$$

$$= 500C\,\mathrm{L}_\ell\,a^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\sqrt{\tau}\log(2N/\delta_0)\sqrt{\frac{1}{\kappa mn}} + \mathrm{B}_\ell\,\frac{4\log(1/\delta_1)}{3N} \tag{D.19}$$

where at the last lines we have simply plugged in the definition of $B = \sqrt{2\log(2N/\delta_0)}$.

Now, note that by the definition of $\kappa = \max(K, 1)$, we have

$$\kappa \geqslant 2\frac{N}{2mn\log(\frac{N}{2\delta_2})}, \tag{D.20}$$

and hence

$$\kappa mn \geqslant \frac{N}{\log(\frac{N}{2\delta_2})}. \tag{D.21}$$

Plugging this back into equation (**??**), we have

$$\mathbb{E}_{(i,j)\sim\mathcal{U}}(\ell(\widehat{R}_{(i,j)}, [R+\zeta]_{(i,j)})) \leqslant 500C\,\mathrm{L}_\ell\,a^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\sqrt{\tau}\log\left(\frac{2N}{\delta_0}\right)\sqrt{\frac{\log(\frac{N}{2\delta_2})}{N}} + \mathrm{B}_\ell\,\frac{4\log(1/\delta_1)}{3N}, \tag{D.22}$$

where

$$\tau = \frac{N}{mn} + \frac{8}{3}\log\left(\frac{2mn}{\delta_2}\right)\sqrt{\frac{N}{mn}}. \tag{D.23}$$

Now, assume first that $N \leqslant 2mn\log(\frac{N}{2\delta_2})$. In this case, we can write

$$\tau = \frac{N}{mn} + \frac{8}{3}\log\left(\frac{2mn}{\delta_2}\right)\sqrt{\frac{N}{mn}} \tag{D.24}$$

$$\leqslant 2\log(\frac{N}{2\delta_2}) + \frac{8}{3}\log\left(\frac{2mn}{\delta_2}\right)\sqrt{2\log(\frac{N}{2\delta_2})} \tag{D.25}$$

$$:= \theta. \tag{D.26}$$

On the other hand, if $N \geqslant 2mn\log(\frac{N}{2\delta_2})$, the $K \geqslant 1$ and (on our event with probability $\geqslant 1-\delta_2$), we have that each entry is sampled at least once. In this case, using the notation of Lemma C.1, we have that

$$\|\widehat{R} - R\|_* = \|H\|_* \leqslant 6C\sqrt{a}B\,\sigma\sqrt{\frac{N}{\kappa}} \tag{D.27}$$

$$\leqslant 70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{N/\kappa}. \tag{D.28}$$

Note this is the same expression as (D.11) without the τ. Hence, by the same calculation as above, we have (w.p. $\geqslant 1-\delta_1-\delta_2-\delta_0-5q\delta$):

$$\mathbb{E}_{(i,j)\sim\mathcal{U}}(\ell(\widehat{R}_{(i,j)}, [R+\zeta]_{(i,j)})) \leqslant 500C\,\mathrm{L}_\ell\,a^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\log\left(\frac{2N}{\delta_0}\right)\sqrt{\frac{\log(\frac{N}{2\delta_2})}{N}} + \mathrm{B}_\ell\,\frac{4\log(1/\delta_1)}{3N}. \tag{D.29}$$

This is the same expression as equation (D.22) without the τ.

From this and the fact that $\theta \geqslant 1$ it follows that equation (E.1) also holds in this case. This concludes the proof.

□

E Result with the absolute loss

As an almost immediate consequence of Theorem D.2 we have the following corollary which gives a rate of

Corollary E.1. *Assume that condition* (C.3) *on λ holds. For any $\delta_0, \delta_1, \delta_2, \delta > 0$, with probability*

$$\geqslant 1-\delta_1-\delta_0-\delta_2-\delta\left[1+5\log\left[e^6\sigma_0^{-8}a\log\left(\frac{mn}{\delta}\right)\right]\right]$$

as long as

$$N \geqslant \log\left[e^6\sigma_0^{-8}a\log\left(\frac{mn}{\delta}\right)\right]\frac{128}{3}\mu\mu_1 r(a+b)\log\left(\frac{2mn}{\delta}\right)$$

we have

$$\mathbb{E}_{(i,j)\sim\mathcal{U}}\left|\widehat{R}_{(i,j)} - [R+\zeta]_{(i,j)}\right| \leqslant 700Ca^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\Theta\sqrt{\frac{1}{N}}, \tag{E.1}$$

where

$$\Theta := 2\log\left(\frac{Nmn}{\delta_2^2}\right)\log\left(\frac{2N}{\delta_0}\right)\log\left(\frac{1}{\delta_1}\right). \tag{E.2}$$

Proof. First note that by the same arguments as in the proof of Theorem D.2 we have

$$\|\widehat{R} - R\|_* \leqslant 70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\theta N/\kappa}. \tag{E.3}$$

In particular,

$$\|\widehat{R} - R\|_\infty \leqslant \|\widehat{R} - R\|_* \leqslant 70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\theta N/\kappa}. \tag{E.4}$$

Based on this we define the loss function ℓ by

$$\ell(x,y) := \min(|x-y|, 70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\theta N/\kappa}). \tag{E.5}$$

Applying Theorem D.2 together with equation (E.4) we obtain:

$$\mathbb{E}_{(i,j)\sim\mathcal{U}}\left|\widehat{R}_{(i,j)} - [R+\zeta]_{(i,j)}\right| = \mathbb{E}\ell\left(\widehat{R}_{(i,j)}, [R+\zeta]_{(i,j)}\right) \tag{E.6}$$

$$\leqslant 500C\,\mathrm{L}_\ell\,a^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\sqrt{\theta}\log\left(\frac{2N}{\delta_0}\right)\sqrt{\frac{\log(\frac{N}{2\delta_2})}{N}} + \mathrm{B}_\ell\,\frac{4\log(1/\delta_1)}{3N} \tag{E.7}$$

$$\leqslant 500Ca^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\sqrt{\theta}\log\left(\frac{2N}{\delta_0}\right)\sqrt{\frac{\log(\frac{N}{2\delta_2})}{N}} + [70CaB^2\sigma_0^{-2}\,\sigma\,\sqrt{\theta N/\kappa}]\frac{4\log(1/\delta_1)}{3N} \tag{E.8}$$

$$\leqslant 700Ca^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\sqrt{\theta}\log\left(\frac{2N}{\delta_0}\right)\log(1/\delta_1)\sqrt{\log(\frac{N}{2\delta_2})}\sqrt{\frac{1}{N}} \tag{E.9}$$

$$\leqslant 700Ca^{3/2}\sqrt{b}\mu\sigma_0^{-2}\,\sigma\,\Theta\sqrt{\frac{1}{N}}, \tag{E.10}$$

where we have replaced the value of $B = \sqrt{2\log(2N/\delta_0)}$. This concludes the proof.

□

F Proofs of results in O notation

Proof of Theorem 1. This follows immediately from Theorem D.1 together with the computational Lemma F.1 below. Indeed, recall that $\mu_1 \leqslant \mu^4 r$, then to tackle the different δ, Δs we can apply Lemma F.1 to see that we can set

$$\delta = \min\left(\frac{\frac{\Delta}{6K_1}}{\log\left(\frac{3K_1K_2}{\Delta}\right)}, \frac{\Delta}{3[1+5\log(e^6\sigma_0^{-8}a)]}\right) \tag{F.1}$$

with

$$K_2 = mn \tag{F.2}$$

$$K_1 = 5, \tag{F.3}$$

as this will ensure that $\delta\left[1+5\log\left[e^6\sigma_0^8 a\log\left(\frac{mn}{\delta}\right)\right]\right] \leqslant \frac{2}{3}\Delta$.

□

Proof of Theorem 2. Follows by the same arguments as the proof of Theorem 1 setting also $\delta_0 = \delta_1 = \delta_2 = \Delta/9$. □

Proof of Theorem 3. The proof follows directly from Lemma F.1 and corollary E.1 in exactly the same way as Theorem 2 above. □

F.1 Change of variables for δ

Lemma F.1. *Let $K_1, K_2 > 1$ and let $\Delta > 0$. As long as*

$$\delta \leqslant \frac{\frac{\Delta}{2K_1}}{\log\left(\frac{K_1K_2}{\Delta}\right)}, \tag{F.4}$$

then we have

$$K_1\delta\log\left(\frac{K_2}{\delta}\right) \leqslant \Delta. \tag{F.5}$$

Proof. Observe that the equation

$$\Delta \geqslant K_1\delta\log\left(\frac{K_2}{\delta}\right)$$

is equivalent to

$$\frac{\Delta}{K_1K_2} \geqslant \frac{\delta}{K_2}\log\left(\frac{K_2}{\delta}\right),$$

which is in turn equivalent to

$$\frac{K_1K_2}{\Delta} \leqslant \frac{\frac{K_2}{\delta}}{\log\left(\frac{K_2}{\delta}\right)}.$$

By Lemma F.2 this will certainly be satisfied as long as

$$\frac{K_2}{\delta} \geqslant 2\frac{K_1K_2}{\Delta}\log\left(\frac{K_1K_2}{\Delta}\right), \tag{F.6}$$

which can be equivalently rewritten

$$\delta \leqslant \frac{\frac{\Delta}{2K_1}}{\log\left(\frac{K_1K_2}{\Delta}\right)}, \tag{F.7}$$

as expected.

□

Lemma F.2. *If $x := 2y\log(y)$ then we have*

$$\frac{x}{\log(x)} \geqslant y. \tag{F.8}$$

Proof. We have

$$\frac{x}{\log(x)} = \frac{2y\log(y)}{\log(y)+\log(2\log(y))} \tag{F.9}$$

$$\geqslant \frac{2y\log(y)}{\log(y)+\log(y)} \tag{F.10}$$

$$= y, \tag{F.11}$$

where at the second line we have used the inequality $\log(y) \leqslant y/2$. □

G Concentration results for exact recovery

In this section we prove that the conditions of Lemma C.1 hold with high probability as long as the number of samples N is large enough. The proofs here are very similar to the analogues in (Xu, Jin, and Zhou 2013) (in some cases we quote the relevant result directly whenever this can be done without the need to modify the proof). However, some slight modifications are needed to make the results more general in the arbitrary tolerance thresholds δ_3 etc., and also to remove the condition $N = |\Omega| \leqslant |\Omega_1|$ from the reference in question as we are interested in compact results that hold for any value of N and in particular in the $N \to \infty$ limit.

Lemma G.1 (Adaptation of Lemma 5 in (Xu, Jin, and Zhou 2013))**.** *For any $\delta_3 > 0$ we have with probability $\geqslant 1 - \delta_3$:*

$$\left\| P_T - \frac{mn}{N} P_T P_\Omega P_T \right\| \leqslant \min\left(\sqrt{\frac{8}{3} \log\left(\frac{m+n}{\delta_3}\right) \frac{r\mu^2(a+b)}{N}},\ \frac{8}{3} \frac{r\mu^2(a+b)}{N} \log\left(\frac{m+n}{\delta_3}\right) \right). \tag{G.1}$$

In particular, as long as

$$N \geqslant T_3 := \frac{32}{3} r\mu^2(a+b) \log\left(\frac{m+n}{\delta_3}\right), \tag{G.2}$$

we have for any $Z \in \mathbb{R}^{m \times n}$:

$$\frac{mn}{N} \langle Z, P_T P_\Omega P_T(Z) \rangle \geqslant \frac{\|P_T(Z)\|_{\mathrm{Fr}}^2}{2}. \tag{G.3}$$

Proof. As computed in (Xu, Jin, and Zhou 2013) we have the following values for the $\sum_{k=1}^N \rho_k^2$ and M from Lemma A.4:

$$M := \frac{r\mu^2(a+b)}{N}, \tag{G.4}$$

$$\sum_{k=1}^N \rho_k^2 = \frac{r\mu^2(a+b)}{N}. \tag{G.5}$$

Plugging these values into Lemma A.4 we immediately obtain:

$$\left\| P_T - \frac{mn}{N} P_T P_\Omega P_T \right\| \leqslant \min\left(\sqrt{\frac{8}{3} \log\left(\frac{m+n}{\delta_3}\right) \frac{r\mu^2(a+b)}{N}},\ \frac{8}{3} \frac{r\mu^2(a+b)}{N} \log\left(\frac{m+n}{\delta_3}\right) \right) \tag{G.6}$$

as expected.

As for the second part of the theorem, not that if $N \geqslant T_3 := \frac{32}{3} r\mu^2(a+b) \log\left(\frac{m+n}{\delta_3}\right)$, then we clearly have

$$\frac{8}{3} \log\left(\frac{m+n}{\delta_3}\right) \frac{r\mu^2(a+b)}{N} \leqslant \frac{1}{4}, \tag{G.7}$$

and therefore by equation (G.6),

$$\left\| P_T - \frac{mn}{N} P_T P_\Omega P_T \right\| \leqslant \min\left(\sqrt{\frac{8}{3} \log\left(\frac{m+n}{\delta_3}\right) \frac{r\mu^2(a+b)}{N}},\ \frac{8}{3} \frac{r\mu^2(a+b)}{N} \log\left(\frac{m+n}{\delta_3}\right) \right) \leqslant \frac{1}{2}. \tag{G.8}$$

This in turn implies that

$$\frac{1}{2}\|P_T(Z)\|_{\mathrm{Fr}}^2 \geqslant \left\langle Z, P_T(Z) - \frac{mn}{N} P_T P_\Omega P_T(Z) \right\rangle = \langle Z, P_T(Z) \rangle - \left\langle Z, \frac{mn}{N} P_T P_\Omega P_T(Z) \right\rangle, \tag{G.9}$$

from which it follows that

$$\frac{1}{2}\|P_T(Z)\|_{\mathrm{Fr}}^2 \leqslant \frac{mn}{N} \langle Z, P_T P_\Omega P_T(Z) \rangle, \tag{G.10}$$

as expected.

□

Lemma G.2 (Substantially modified version of Lemma 6 in (Xu, Jin, and Zhou 2013))**.** *For any $\delta_4 > 0$ we have with probability $\geqslant 1 - \delta_4$:*

$$\left\| P_{T^\top} - \frac{mn}{N} P_{T^\top} P_\Omega P_{T^\top} \right\| \leqslant \left(\sqrt{\frac{8}{3} \log\left(\frac{m+n}{\delta_4}\right) \frac{\mu^2(ab+r^2)}{N}},\ \frac{8}{3} \frac{\mu^2(ab+r^2)}{N} \log\left(\frac{m+n}{\delta_4}\right) \right). \tag{G.11}$$

In particular, as long as

$$N \geqslant T_4 := \frac{32}{3} r\mu^2(a+b) \log\left(\frac{m+n}{\delta_4}\right) \tag{G.12}$$

we have

$$\left\langle Z, \frac{mn}{N} P_{T^\top} P_\Omega P_{T^\top}(Z) \right\rangle \leqslant \frac{3}{2} \frac{a}{r} \|P_{T^\top}(Z)\|_{\mathrm{Fr}}^2. \tag{G.13}$$

Proof. By the calculation in (Xu, Jin, and Zhou 2013) we can apply Lemma A.4 with the following values for M, ρ:

$$M := \frac{\mu^2(ab+r^2)}{N} \tag{G.14}$$

$$\sum_{k=1}^{N} \rho_k^2 = \frac{\mu^2(ab+r^2)}{N}. \tag{G.15}$$

From this, applying Lemma A.4 we immediately obtain:

$$\left\| P_{T^\top} - \frac{mn}{N} P_{T^\top} P_\Omega P_{T^\top} \right\| \leqslant \max\left(\sqrt{\frac{8}{3} \log\left(\frac{m+n}{\delta}\right) \sum_{k=1}^{L} \rho_k^2},\ \frac{8}{3} M \log\left(\frac{m+n}{\delta}\right) \right) \tag{G.16}$$

$$\leqslant \max\left(\sqrt{\frac{8}{3} \log\left(\frac{m+n}{\delta}\right) \frac{\mu^2(ab+r^2)}{N}},\ \frac{8}{3} \frac{\mu^2(ab+r^2)}{N} \log\left(\frac{m+n}{\delta}\right) \right), \tag{G.17}$$

as expected.

Note that

$$T_4 := \frac{32}{3} r\mu^2(a+b) \log\left(\frac{m+n}{\delta_4}\right) \tag{G.18}$$

$$\geqslant \frac{32}{3} \frac{r}{a} \mu^2(ab+r^2) \log\left(\frac{m+n}{\delta_4}\right). \tag{G.19}$$

(Indeed, $r^2/a \leqslant a$ (because $r \leqslant a$) and $rab/a = rb$.)

Thus we certainly have

$$\left\| P_{T^\top} - \frac{mn}{N} P_{T^\top} P_\Omega P_{T^\top} \right\| \leqslant \max\left(\sqrt{\frac{8}{3} \log\left(\frac{m+n}{\delta}\right) \frac{\mu^2(ab+r^2)}{N}},\ \frac{8}{3} \frac{\mu^2(ab+r^2)}{N} \log\left(\frac{m+n}{\delta}\right) \right) \tag{G.20}$$

$$\leqslant \max\left(\sqrt{\frac{a}{4r}}, \frac{a}{4r} \right) \leqslant \frac{a}{2r}. \tag{G.21}$$

This implies we have

$$\left\langle Z, \frac{mn}{N} P_{T^\top} P_\Omega P_{T^\top}(Z) \right\rangle \leqslant \left[1 + \frac{a}{2r} \right] \|P_{T^\top}(Z)\|_{\mathrm{Fr}}^2 \leqslant \frac{3a}{2r} \|P_{T^\top}(Z)\|_{\mathrm{Fr}}^2, \tag{G.22}$$

as expected.

□

Lemma G.3. *For all $\delta_3, \delta_4, \delta_5 > 0$, for any $Z \in \mathbb{R}^{m \times n}$ such that $P_\Omega(Z) = 0$ and $P_X Z P_Y = Z$, we have with probability $\geqslant 1 - \delta_3 - \delta_4 - \delta_5$:*

$$\|P_T(Z)\|_{\mathrm{Fr}} \leqslant \sqrt{3\tau_5} \sqrt{\frac{a}{r}} \|P_{T^\top}(Z)\|_{\mathrm{Fr}}, \tag{G.23}$$

as long as $N \geqslant \max(T_3, T_4)$ where T_3, T_4 are defined as in Lemmas G.1 and G.2.

Proof. First note that since $P_\Omega(Z) = 0$ and $P_X Z P_Y = Z$ we certainly have that

$$\langle P_\Omega P_T(Z), P_\Omega P_T(Z)\rangle = \langle P_\Omega P_{T^\top}(Z), P_\Omega P_{T^\top}(Z)\rangle, \tag{G.24}$$

which implies

$$\left\langle Z, P_T P_\Omega^2 P_T(Z)\right\rangle = \left\langle Z, P_{T^\top} P_\Omega^2 P_{T^\top}(Z)\right\rangle. \tag{G.25}$$

Next, observe also that

$$\begin{aligned}\langle Z, P_T P_\Omega P_T(Z)\rangle = \langle P_T Z, P_\Omega P_T(Z)\rangle &= \sum_{(i,j)} [P_T Z]_{i,j}^2 h_{i,j} \\ &\leqslant \sum_{(i,j)} [P_T Z]_{i,j}^2 h_{i,j}^2 = \left\langle P_T Z, P_\Omega^2 P_T(Z)\right\rangle = \left\langle Z, P_T P_\Omega^2 P_T(Z)\right\rangle,\end{aligned} \tag{G.26}$$

where $h_{i,j}$ denotes the number of times that entry (i,j) was sampled.

Now, by lemma A.5 we have that with probability $\geqslant 1-\delta_5$, $h_{i,j} \leqslant \tau_5$ for all i, j. Thus under the same condition we also have similarly to equation (G.26)

$$\left\langle Z, P_{T^\top} P_\Omega^2 P_{T^\top}(Z)\right\rangle \leqslant \tau_5 \langle Z, P_{T^\top} P_\Omega P_{T^\top}(Z)\rangle. \tag{G.27}$$

Now by Lemmas G.1 and G.2 together with the above, we have with probability $\geqslant 1-\delta_3-\delta_4-\delta_5$:

$$\begin{aligned}\frac{1}{2}\|P_T(Z)\|_{\mathrm{Fr}}^2 &\leqslant \frac{mn}{N}\langle Z, P_T P_\Omega P_T(Z)\rangle && \text{(G.28)}\\ &\leqslant \frac{mn}{N}\left\langle Z, P_T P_\Omega^2 P_T(Z)\right\rangle && \text{(G.29)}\\ &\leqslant \frac{mn}{N}\left\langle Z, P_{T^\top} P_\Omega^2 P_{T^\top}(Z)\right\rangle && \text{(G.30)}\\ &\leqslant \tau_5 \frac{mn}{N}\langle Z, P_{T^\top} P_\Omega P_{T^\top}(Z)\rangle && \text{(G.31)}\\ &\leqslant \tau_5 \frac{3}{2}\frac{a}{r}\|P_{T^\top}(Z)\|_{\mathrm{Fr}}^2, && \text{(G.32)}\end{aligned}$$

where at the first line (G.28) we have used Lemma G.1; at the second line (G.29) we have used equation (G.26); at the third line (G.30) we have used equation (G.25); at the fourth line (G.31) we have used equation (G.27); and at the fifth and last line (G.32) we have used Lemma G.2. The result follows.

□

Lemma G.4 (Variation of Lemma 8 in (Xu, Jin, and Zhou 2013))**.** *Let $Z \in \mathbb{R}^{m\times n}$, for any $\delta_6 > 0$ as long as $N \geqslant T_6 := \frac{8}{3}\mu^2 r(a+b)\log\left(\frac{m+n}{\delta_6}\right)$ we have w.p. $\geqslant 1-\delta_6$:*

$$\frac{mn}{N}\|P_{T^\top} P_\Omega P_T(Z)\| \leqslant \|P_T(Z)\|_\infty \sqrt{\frac{8\log\left(\frac{m+n}{\delta}\right) mn\mu\max(a,b)}{3N}}. \tag{G.33}$$

Proof. This again follows from an application of Bernstein's inequality A.4. As calculated in (Xu, Jin, and Zhou 2013) we note the following values for M, ρ:

$$M = \|P_T(Z)\|_\infty \sqrt{\frac{mn\mu^2(ab+r^2)}{N^2}} \tag{G.34}$$

$$\sum \rho_k^2 = \|P_T(Z)\|_\infty^2 \frac{\mu\max(a,b)mn}{N}. \tag{G.35}$$

Thus Lemma A.4 immediately implies that with probability $\geqslant 1-\delta_6$:

$$\frac{mn}{N}\|P_{T^\top} P_\Omega P_T(Z)\| \leqslant \|P_T(Z)\|_\infty \max\left(\sqrt{\frac{8\mu mn\max(a,b)}{3N}\log\left(\frac{m+n}{\delta_6}\right)}, \frac{8}{3}\sqrt{\frac{mn\mu^2(ab+r^2)}{N^2}}\log\left(\frac{m+n}{\delta_6}\right)\right). \tag{G.36}$$

Thus as long as $N \geqslant \frac{8}{3}\mu\frac{ab+r^2}{\max(a,b)}\log\left(\frac{m+n}{\delta_6}\right)$ we certainly have

$$\frac{mn}{N}\left\|P_{T^\top}P_\Omega P_T(Z)\right\| \leqslant \|P_T(Z)\|_\infty\sqrt{\frac{8\mu mn\max(a,b)}{3N}\log\left(\frac{m+n}{\delta_6}\right)}, \tag{G.37}$$

as expected. The result follows upon noting that $T_6 \geqslant \frac{8}{3}\mu\frac{ab+r^2}{\max(a,b)}\log\left(\frac{m+n}{\delta_6}\right)$.

□

Lemma G.5. *Let $Z \in \mathbb{R}^{m\times n}$ with probability $\geqslant 1-\delta_7$. Along as $N \geqslant T_7\frac{8}{3}\mu^2 r(a+b)\log\left(\frac{2mn}{\delta_7}\right)$ we have*

$$\left\|P_T(Z) - P_TP_\Omega P_T(Z)\right\|_\infty \leqslant \|P_T(Z)\|_\infty \max\sqrt{\frac{8}{3}\log\left(\frac{2mn}{\delta_7}\right)\frac{\mu^2 r(a+b)}{N}}. \tag{G.38}$$

In particular, as long as $N \geqslant T_7 := \frac{32}{3}\mu^2 r(a+b)\log\left(\frac{2mn}{\delta_7}\right)$, we have

$$\left\|P_T(Z) - P_TP_\Omega P_T(Z)\right\|_\infty \leqslant \frac{1}{2}\|P_T(Z)\|_\infty. \tag{G.39}$$

Proof. This follows from an application of the standard Bernstein inequality (i.e. Lemma A.4 with $m=n=1$) applied to each entry separately, together with a union bound over entries. As calculated in (Xu, Jin, and Zhou 2013) we have the following values for "M" and "ρ":

$$M = \frac{\mu^2 r(a+b)\|P_T(Z)\|_\infty}{N} \tag{G.40}$$

$$\sum\rho_k^2 = \frac{\mu^2 r(a+b)\|P_T(Z)\|_\infty^2}{N}. \tag{G.41}$$

Thus applying Lemma A.4 we see that for all i,j, the following holds with probability $\geqslant 1-\delta_7$:

$$\left|\left[P_T(Z) - P_TP_\Omega P_T(Z)\right]_{i,j}\right| \leqslant \|P_T(Z)\|_\infty \max\left(\sqrt{\frac{8}{3}\log\left(\frac{2}{\delta_7}\right)\frac{\mu^2 r(a+b)}{N}}, \frac{8}{3}\frac{\mu^2 r(a+b)\log\left(\frac{2}{\delta_7}\right)}{N}\right), \tag{G.42}$$

and as long as $N \geqslant \frac{8}{3}\mu^2 r(a+b)\log\left(\frac{2}{\delta_7}\right)$ we have

$$\left|\left[P_T(Z) - P_TP_\Omega P_T(Z)\right]_{i,j}\right| \leqslant \|P_T(Z)\|_\infty \max\sqrt{\frac{8}{3}\log\left(\frac{2}{\delta_7}\right)\frac{\mu^2 r(a+b)}{N}}. \tag{G.43}$$

Setting $\delta_7 \leftarrow \delta_7/(mn)$ and taking a union bound over entries yields the first result immediately. The second result follows directly from the first.

□

H Proof of Lemma C.2

First let us fix $\delta > 0$. We will set $\delta_3 = \delta_4 = \delta_5 = \delta_6 = \delta_7$ for the lemmas above.

Now, define

$$T = \frac{32}{3}\mu^2 r(a+b)\log\left(\frac{2mn}{\delta}\right). \tag{H.1}$$

Note that as long as $N \geqslant T$, we will have $N \geqslant \max(T_3, T_4, T_6, T_7)$ (with the same value δ used in all relevant theorems), which means the conditions of Lemmas G.1, G.2, A.5, G.3 G.4, and G.4 are all satisfied. Indeed, it is trivially the case that $N \geqslant \max(T_3, T_4, T_6)$. As for T_7, the inequality $N \geqslant T_7$ still follows upon noticing that $\log(2mn) \leqslant \log((m+n)^2) = 2\log(m+n)$.

Following the ideas from (Recht 2011) and (Xu, Jin, and Zhou 2013) we now construct a matrix $Y \in \mathbb{R}^{m\times n}$ with the properties from Lemma C.1.

We assume that $N \geqslant qT$ where q will be determined later. We randomly select q disjoint subsets of samples, each of size T, denoted by $\Omega_1, \Omega_2, \ldots, \Omega_q$, so we have

$$|\Omega_i| = T \qquad \forall i \leqslant q. \tag{H.2}$$

As in Lemma C.1 we define U to be the dual certificate of $R = XM^*Y^\top$ with respect to the norms $\|\cdot\|_{\mathcal{I},\sigma}$ and $\|\cdot\|_{\mathcal{I},*}$ and the standard Frobenius inner product. Note that because unlike (Xu, Jin, and Zhou 2013) we do not assume that the columns of X, Y are normed, we need to be a bit more careful about computing the relevant norms. By definition we certainly have $\|U\|_{\mathcal{I},\sigma} = 1$. However, to apply the above results, we will also need a bound on $\|U\|$.

Lemma H.1. *Let U be the dual certificate of $R = XM^*Y^\top$ with respect to the norms $\|\cdot\|_{\mathcal{I},\sigma}$ and $\|\cdot\|_{\mathcal{I},*}$ and the standard Frobenius inner product. Let σ_0 denote the smallest singular value of X and Y (after preprocessing of X, Y into matrices with orthogonal columns (ordered by decreasing norms), σ_0 denotes the minimum of the norm of the last column of X and that of Y). We have the following bound on the ordinary spectral norm of U:*

$$\|U\| \leqslant \sigma_0^{-2}. \tag{H.3}$$

Proof. By definition of U, $\|U\|_{\mathcal{I},\sigma} = 1$. By definition of the norm $\|\cdot\|_{\mathcal{I},\sigma}$ we have $\|U\|_{\mathcal{I},\sigma} = X^\top UY = \Sigma_1 \bar{X}^\top U \bar{Y} \Sigma_2$ where $\bar{X}, \bar{Y}$ are obtained by normalising the columns of X, Y and Σ_1, Σ_2 are diagonal matrices containing the singular values of X, Y. We can now write

$$\|U\| = \|\bar{X}^\top U \bar{Y}\| = \|\Sigma_1^{-1}[\Sigma_1 \bar{X}^\top U \bar{Y} \Sigma_2]\Sigma_2^{-1}\| \leqslant \sigma_0^{-2}\|\Sigma_1^{-1}[\Sigma_1 \bar{X}^\top U \bar{Y} \Sigma_2]\Sigma_2^{-1}\| = \sigma_0^{-2}\|U\|_{\mathcal{I},\sigma}, \sigma_0^2\|\bar{X}^\top U \bar{Y}\| = \sigma_0^2\|U\|, \tag{H.4}$$

as expected.

□

Armed with the above, we continue the construction of our approximation Y of U: we generate a sequence $Y_1, \ldots, Y_q$ as follows:

$$\mathcal{Y}_t = \frac{mn}{T} \sum_{i=1}^{t} P_{\Omega_i}(W_i) \tag{H.5}$$

where $W_1 = U$ and W_{t+1} is defined inductively as follows:

$$W_{t+1} = P_T(U - \mathcal{Y}_t) = W_t - \frac{mn}{T} P_T P_{\Omega_t}(W) \tag{H.6}$$

$$= \left(P_T - \frac{mn}{T} P_T P_{\Omega_t} P_T\right) W_t. \tag{H.7}$$

Finally, we set $\mathcal{Y} = \mathcal{Y}_q$.

Remark: the subsets Ω_i of the original sample are subsets of the observations rather than subsets of the entries. In particular, they can contain several obsevations of the same entry (and this is accounted for in the Lemmas above).

In the next two lemmas, we will now show that Y satisfies the conditions of Lemma C.1 with high probability.

Lemma H.2 (Improved version of Lemma 10 in (Xu, Jin, and Zhou 2013))**.** *Assume that $N \leqslant mn$.*

With probability $\geqslant 1 - 5q\delta$, as long as $N \geqslant Tq$ and

$$q \geqslant q_0 := 8\log(\sigma_0^{-1}) + 2\log(a) + 4 + \log(\tau) \tag{H.8}$$

$$= 8\log(\sigma_0^{-1}) + 2\log(a) + 4 + \log\left(5\log\left(\frac{2mn}{\delta}\right)\right) \tag{H.9}$$

where $\tau = \tilde{\tau}_5 = 5\log(\frac{2mn}{\delta})$, we have

$$\|P_T(\mathcal{Y}) - U\|_{\mathrm{Fr}} \leqslant \frac{1}{4}\sqrt{\frac{r}{3a\tau}}\frac{1}{\sigma_0^{-2}}. \tag{H.10}$$

Without the condition $N \geqslant mn$, the lemma still holds with $\tau = \tau_5 = \frac{N}{mn} + \frac{8}{3}\log\left(\frac{2mn}{\delta}\right)\sqrt{\frac{N}{mn}}$ (which depends logarithmically on N).

Proof. Setting $\delta_3 = \ldots = \delta_7 = \delta$ in all Lemmas above we have that (as long as $N \geqslant qT$) all the high probability events of Lemmas Lemmas G.1, G.2, A.5, G.3 G.4, and G.4 hold on *each of the groups of samples* Ω_i with probability $\geqslant 1 - 5q\delta$.

Since $W_{t+1} = \left(P_T - \frac{mn}{T} P_T P_{\Omega_t} P_T\right) W_t$, and $\|W_1\| = \|U\| \leqslant \sigma_0^{-2}$ (by Lemma H.1), we can apply Lemma G.1 iteratively and obtain:

$$\|W_{q+1}\| = \|P_T(\mathcal{Y}) - U\| \leqslant \sigma_0^{-2} \prod_{i=1}^{q} \left\|P_T - \frac{mn}{T} P_T P_{\Omega_i} P_T\right\| \leqslant \frac{\sigma_0^{-2}}{2^q}. \tag{H.11}$$

Now, from this we obtain:

$$\|P_T(\mathcal{Y}) - U\|_{\mathrm{Fr}} \leqslant \frac{\sigma_0^{-2}\sqrt{\min(a,b)}}{2^q}. \tag{H.12}$$

Thus, we see that the Lemma's statement will hold as long as we set

$$q \geqslant q_0 := 8\log(\sigma_0^{-1}) + 2\log(a) + 4 + \log(\tau) \tag{H.13}$$

$$\geqslant 8\log(\sigma_0^{-1}) + 2\log(a) + \log(48) + \log(\tau) \geqslant \log_2\left[\sigma_0^{-4}\sqrt{\min(a,b)}4\sqrt{\frac{3a\tau}{r}}\right]. \tag{H.14}$$

□

We will need the following additional Lemma.

Lemma H.3. *Let U be the dual certificate of R, we have the following bound on the maximum entry of U:*

$$\|U\|_\infty \leqslant \sqrt{\frac{r\mu_1}{mn}}\sigma_0^{-2}. \tag{H.15}$$

Proof. Let $M^* = A\Sigma B^\top$ be the singular value decomposition of the ground truth core matrix M^*. By definition of U and the relevant norms we have

$$X^\top U Y = AB^\top. \tag{H.16}$$

It follows that

$$\bar{X}^\top U \bar{Y} = \Sigma_1 AB^\top \Sigma_2,$$

where as usual, $\bar{X}, \bar{Y}$ are obtained from X, Y by normalizing the columns, and Σ_1, Σ_2 are diagonal matrices containing the singular values of X, Y. Next we have

$$U = \bar{X}\Sigma_1 AB^\top \Sigma_2 \bar{Y}, \tag{H.17}$$

the result follows by the incoherence assumption.

□

Lemma H.4 (Modification of Lemma 11 in (Xu, Jin, and Zhou 2013))**.** *With probability $\geqslant 1 - 5q_0\delta$ as long as $N \geqslant \bar{T}q_0$ we have*

$$\|P_{T^\top}(\mathcal{Y})\| \leqslant \frac{1}{2}, \tag{H.18}$$

where $\bar{T} := 4\frac{\mu_1}{\mu}\sigma_0^{-4}T$.

Proof. Similarly to Lemma H.2 under the condition $N \geqslant \bar{T}q_0$ if we randomly pick q groups of samples each of size $\bar{T} \geqslant T$ we have, with probability $\geqslant 1 - 5q_0\delta$, that all the high probability events of the previous lemmas hold for each of the sets of samples Ω_t.

Now by Lemma G.5 we have

$$\|W_{t+1}\|_\infty \leqslant \|(P_T - P_T P_{\Omega_t} P_T)\| \leqslant \frac{1}{2}\|W_t\|_\infty. \tag{H.19}$$

Next we also have

$$\|P_{T^\top}(Y)\| \leqslant \sum_{t=1}^{q} \frac{mn}{T}\|(P_T - P_T P_{\Omega_t} P_T)(W_t)\| \tag{H.20}$$

$$\leqslant \sqrt{\frac{8\log\left(\frac{m+n}{\delta}\right)mn\mu\max(a,b)}{3\bar{T}}}\sum_{t=1}^{q}\|W_t\|_\infty \tag{H.21}$$

$$\leqslant \sqrt{\frac{8\log\left(\frac{m+n}{\delta}\right)mn\mu\max(a,b)}{3\bar{T}}}\sum_{t=1}^{q}\frac{\sigma_0^{-2}}{2^{t-1}}\sqrt{\frac{r\mu_1}{mn}} \tag{H.22}$$

$$\leqslant 2\sqrt{\frac{8\log\left(\frac{m+n}{\delta}\right)\mu\mu_1 r\max(a,b)}{3\bar{T}}}\sigma_0^{-2} \tag{H.23}$$

$$\leqslant \frac{1}{2}, \tag{H.24}$$

where at the second line we have used Lemma G.4, at the third line we have used equation (H.19) as well as Lemma H.3.

□

We can now finally prove Lemma C.2.

Proof of Lemma C.2. **Case 1 :** $N \leqslant 2\log(\frac{mn}{\delta})mn$**.**

In this case, the lemma follows (even with probability $\geqslant 1 - 5q\delta$) immediately from Lemmas G.1, G.2, A.5, G.3, G.4, and G.5 upon noting that we then have:

$$\tau = \frac{N}{mn} + \frac{8}{3}\log\left(\frac{2mn}{\delta}\right)\sqrt{\frac{N}{mn}} \tag{H.25}$$

$$\leqslant 2\log(\frac{mn}{\delta}) + \frac{8}{3}\log\left(\frac{2mn}{\delta}\right)\sqrt{2\log\left(\frac{mn}{\delta}\right)} \tag{H.26}$$

$$\leqslant 5\log^{\frac{3}{2}}\left(\frac{mn}{\delta}\right), \tag{H.27}$$

and therefore

$$q = 8\log(\sigma_0^{-1}) + 2\log(a) + 4 + \log(\tau) \tag{H.28}$$

$$= 8\log(\sigma_0^{-1}) + 2\log(a) + 4 + \log(5\log^{\frac{3}{2}}\left(\frac{mn}{\delta}\right)) \tag{H.29}$$

$$= 4 + 8\log(\sigma_0^{-1}) + 2\log(a) + \log\left[5\log^{\frac{3}{2}}\left(\frac{mn}{\delta}\right)\right] \tag{H.30}$$

$$\leqslant 6 + 8\log(\sigma_0^{-1}) + 2\log(a) + \log\left[\log^{\frac{3}{2}}\left(\frac{mn}{\delta}\right)\right] \tag{H.31}$$

$$\leqslant 6 + 8\log(\sigma_0^{-1}) + 2\log(a) + \log\left[\log\left(\frac{mn}{\delta}\right)\right] \tag{H.32}$$

$$= \log\left[e^6\sigma_0^{-8}a\log\left(\frac{mn}{\delta}\right)\right]. \tag{H.33}$$

Case 2: $N \geqslant 2\log(\frac{mn}{\delta})mn$**.**

In this case, by Lemma C.5 we have with probabilty $\geqslant 1-\delta$ that each entry was sampled at least once. Hence, the dual certificate U itself is in the image of P_Ω and we can simply set $Y = U$. Note also that in this case $\|P_{T^\top}(Y)\| = \|P_{T^\top}(U)\| = 0 \leqslant \frac{1}{2}$, and of course $\|P_T(Y) - U\|_{\mathrm{Fr}} = \|P_T(U) - U\|_{\mathrm{Fr}} = \|0\|_{\mathrm{Fr}} = 0 \leqslant \frac{1}{4}\sqrt{\frac{r}{3a\tau}}\frac{1}{\sigma_0^2}$. □

References

Aggarwal, C. C. 2016. *Recommender Systems: The Textbook*. Springer Publishing Company, Incorporated, 1st edition. ISBN 3319296574.

Alves, R.; Ledent, A.; Assunção, R.; and Kloft, M. 2020. An Empirical Study of the Discreteness Prior in Low-Rank Matrix Completion. *Proceedings of Machine Learning Research (PMLR): NeurIPS 2020 Workshop on the Pre-registration Experiment: An Alternative Publication Model For Machine Learning Research*.

Amiridi, M.; Kargas, N.; and Sidiropoulos, N. D. 2020. Low-rank Characteristic Tensor Density Estimation Part I: Foundations. *arXiv e-prints*, arXiv:2008.12315.

Amiridi, M.; Kargas, N.; and Sidiropoulos, N. D. 2021. Low-rank Characteristic Tensor Density Estimation Part II: Compression and Latent Density Estimation. *arXiv e-prints*, arXiv:2106.10591.

Anandkumar, A.; Ge, R.; Hsu, D.; Kakade, S. M.; and Telgarsky, M. 2014. Tensor Decompositions for Learning Latent Variable Models. *Journal of Machine Learning Research*, 15: 2773–2832.

Bartlett, P. L.; Bousquet, O.; and Mendelson, S. 2005. Local Rademacher complexities. *The Annals of Statistics*, 33(4): 1497 – 1537.

Bartlett, P. L.; and Mendelson, S. 2001. Rademacher and Gaussian Complexities: Risk Bounds and Structural Results. In Helmbold, D.; and Williamson, B., eds., *Computational Learning Theory*, 224–240. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-540-44581-4.

Boucheron, S.; Lugosi, G.; and Bousquet, O. 2004. Concentration inequalities. *Lecture Notes in Computer Science*, 3176: 208–240.

Cai, T. T.; and Zhou, W.-X. 2016. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1): 1493 – 1525.

Candès, E. J.; and Recht, B. 2009. Exact Matrix Completion via Convex Optimization. *Foundations of Computational Mathematics*, 9(6): 717.

Candès, E. J.; and Tao, T. 2010. The Power of Convex Relaxation: Near-Optimal Matrix Completion. *IEEE Trans. Inf. Theor.*, 56(5): 2053–2080.

Candès, E.; and Plan, Y. 2010. Matrix Completion With Noise. *Proceedings of the IEEE*, 98: 925 – 936.

Chen, H.; and Li, J. 2017. Learning Multiple Similarities of Users and Items in Recommender Systems. In *2017 IEEE International Conference on Data Mining (ICDM)*, 811–816.

Chen, T.; Zhang, W.; lu, Q.; Chen, K.; Zheng, Z.; and Yu, Y. 2012. SVDFeature: A Toolkit for Feature-based Collaborative Filtering. *The Journal of Machine Learning Research*.

Chen, Y. 2015. Incoherence-Optimal Matrix Completion. *IEEE Transactions on Information Theory*, 61(5): 2909–2923.

Chen, Y.; Chi, Y.; Fan, J.; Ma, C.; and Yan, Y. 2020. Noisy Matrix Completion: Understanding Statistical Guarantees for Convex Relaxation via Nonconvex Optimization.

Chiang, K.-Y.; Dhillon, I. S.; and Hsieh, C.-J. 2018. Using Side Information to Reliably Learn Low-Rank Matrices from Missing and Corrupted Observations. *J. Mach. Learn. Res.*

Chiang, K.-Y.; Hsieh, C.-J.; and Dhillon, I. S. 2015. Matrix Completion with Noisy Side Information. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Fazel, M. 2002. Matrix Rank Minimization with Applications. *PhD Thesis*.

Fazel, M.; Hindi, H.; and Boyd, S. P. 2001. A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the 2001 American Control Conference. (Cat. No.01CH37148)*, volume 6, 4734–4739 vol.6.

Giménez-Febrer, P.; Pagès-Zamora, A.; and Giannakis, G. B. 2020. Generalization Error Bounds for Kernel Matrix Completion and Extrapolation. *IEEE Signal Processing Letters*, 27: 326–330.

Gross, D.; Liu, Y.-K.; Flammia, S. T.; Becker, S.; and Eisert, J. 2010. Quantum State Tomography via Compressed Sensing. *Phys. Rev. Lett.*, 105: 150401.

Herbster, M.; Pasteris, S.; and Tse, L. 2019. Online Matrix Completion with Side Information. *CoRR*, abs/1906.07255.

Jain, P.; and Dhillon, I. S. 2013. Provable Inductive Matrix Completion. *CoRR*, abs/1306.0626.

Kakade, S. M.; Sridharan, K.; and Tewari, A. 2009. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds, and Regularization. In Koller, D.; Schuurmans, D.; Bengio, Y.; and Bottou, L., eds., *Advances in Neural Information Processing Systems 21*, 793–800. Curran Associates, Inc.

Kargas, N.; and Sidiropoulos, N. D. 2019. Learning Mixtures of Smooth Product Distributions: Identifiability and Algorithm. In Chaudhuri, K.; and Sugiyama, M., eds., *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, 388–396. PMLR.

Keshavan, R.; Montanari, A.; and Oh, S. 2009. Matrix Completion from Noisy Entries. In Bengio, Y.; Schuurmans, D.; Lafferty, J. D.; Williams, C. K. I.; and Culotta, A., eds., *Advances in Neural Information Processing Systems 22*, 952–960. Curran Associates, Inc.

Kueng, R.; Rauhut, H.; and Terstiege, U. 2017. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1): 88–116.

Ledent, A.; Alves, R.; and Kloft, M. 2021. Orthogonal Inductive Matrix Completion. *IEEE Transactions on Neural Networks and Learning Systems*, 1–12.

Ledent, A.; Alves, R.; Lei, Y.; and Kloft, M. 2021. Fine-grained Generalization Analysis of Inductive Matrix Completion. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 25540–25552. Curran Associates, Inc.

Li, R.; Dong, Y.; Kuang, Q.; Wu, Y.; Li, Y.; Zhu, M.; and Li, M. 2015. Inductive matrix completion for predicting adverse drug reactions (ADRs) integrating drug–target interactions. *Chemometrics and Intelligent Laboratory Systems*, 144: 71 – 79.

Lin, W.-Y.; Liu, S.; Ren, C.; Cheung, N.-M.; Li, H.; and Matsushita, Y. 2022. Shell Theory: A Statistical Model of Reality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10): 6438–6453.

Mazumder, R.; Hastie, T.; and Tibshirani, R. 2010. Spectral Regularization Algorithms for Learning Large Incomplete Matrices. *J. Mach. Learn. Res.*, 11: 2287–2322.

Menon, A. K.; and Elkan, C. 2011. Link Prediction via Matrix Factorization. In *Machine Learning and Knowledge Discovery in Databases*, 437–452. Springer Berlin Heidelberg.

Recht, B. 2011. A Simpler Approach to Matrix Completion. *J. Mach. Learn. Res.*, 12(null): 3413–3430.

Shamir, O.; and Shalev-Shwartz, S. 2011. Collaborative Filtering with the Trace Norm: Learning, Bounding, and Transducing. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, 661–678. PMLR.

Shamir, O.; and Shalev-Shwartz, S. 2014. Matrix Completion with the Trace Norm: Learning, Bounding, and Transducing. *Journal of Machine Learning Research*, 15: 3401–3423.

Shen, W.; Zhang, C.; Tian, Y.; Zeng, L.; He, X.; Dou, W.; and Xu, X. 2021. Inductive Matrix Completion Using Graph Autoencoder. *CoRR*, abs/2108.11124.

Song, L.; Anandkumar, A.; Dai, B.; and Xie, B. 2014. Nonparametric Estimation of Multi-View Latent Variable Models. In Xing, E. P.; and Jebara, T., eds., *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, 640–648. Bejing, China: PMLR.

Steck, H. 2019. Embarrassingly Shallow Autoencoders for Sparse Data. In *The World Wide Web Conference*, WWW '19, 3251–3257. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.

Tanner, J.; Thompson, A.; and Vary, S. 2019. Matrix Rigidity and the Ill-Posedness of Robust PCA and Matrix Completion. *SIAM Journal on Mathematics of Data Science*, 1(3): 537–554.

Vančura, V.; Alves, R.; Kasalickỳ, P.; and Kordík, P. 2022. Scalable Linear Shallow Autoencoder for Collaborative Filtering. In *Proceedings of the 16th ACM Conference on Recommender Systems*, 604–609.

Vandermeulen, R. A. 2020. Improving Nonparametric Density Estimation with Tensor Decompositions.

Vandermeulen, R. A.; and Ledent, A. 2021. Beyond Smoothness: Incorporating Low-Rank Analysis into Nonparametric Density Estimation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 12180–12193. Curran Associates, Inc.

Vershynin, R. 2019. High-Dimensional Probability.

Wang, J.; Wong, R. K. W.; Mao, X.; and Chan, K. C. G. 2021. Matrix Completion with Model-free Weighting. arXiv:2106.05850.

Wu, Q.; Zhang, H.; Gao, X.; and Zha, H. 2020. Inductive Collaborative Filtering via Relation Graph Learning.

Xu, M.; Jin, R.; and Zhou, Z.-H. 2013. Speedup Matrix Completion with Side Information: Application to Multi-Label Learning. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, 2301–2309. Red Hook, NY, USA: Curran Associates Inc.

Yao, Q.; and Kwok, J. T. 2019. Accelerated and Inexact Soft-Impute for Large-Scale Matrix and Tensor Completion. *IEEE Transactions on Knowledge and Data Engineering*, 31(9): 1665–1679.

Zhang, M.; and Chen, Y. 2020. Inductive Matrix Completion Based on Graph Neural Networks. In *International Conference on Learning Representations*.

Zhang, X.; Du, S.; and Gu, Q. 2018. Fast and Sample Efficient Inductive Matrix Completion via Multi-Phase Procrustes Flow. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 5756–5765. Stockholmsmässan, Stockholm Sweden: PMLR.

Zhong, K.; Jain, P.; and Dhillon, I. S. 2015. Efficient Matrix Sensing Using Rank-1 Gaussian Measurements. In Chaudhuri, K.; GENTILE, C.; and Zilles, S., eds., *Algorithmic Learning Theory*, 3–18. Cham: Springer International Publishing. ISBN 978-3-319-24486-0.