# Fine-grained Generalisation Analysis of Inductive Matrix Completion

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

In this paper, we bridge the gap between the state-of-the-art theoretical results for matrix completion with the nuclear norm and their equivalent in *inductive matrix completion*: (1) In the distribution-free setting, we prove bounds improving the previously best scaling of $rd^2$ to $d^{3/2}\sqrt{r\log(d)}$, where $d$ is the dimension of the side information and $r$ is the rank. (2) We introduce the (smoothed) *adjusted trace-norm minimization* strategy, an inductive analogue of the weighted trace norm, for which we show guarantees of the order $O(dr\log(d))$ under arbitrary sampling. In the inductive case, a similar rate was previously achieved only under uniform sampling and for exact recovery. Both our results align with the state of the art in the particular case of standard (non-inductive) matrix completion, where they are known to be tight up to log terms. Experiments further confirm that our strategy outperforms standard inductive matrix completion on various synthetic datasets and real problems, justifying its place as an important tool in the arsenal of methods for matrix completion using side information.

## 1 Introduction

Matrix completion (MC) is the machine learning problem of recovering the missing entries of a partially observed matrix. It is the go-to approach in various application domains such as recommender systems [1, 2] and social network analysis [3, 4, 5]. The SoftImpute algorithm [6, 7] is among the most popular MC methods. It solves the following convex problem encouraging low-rank solutions:

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

where $P_\Omega$ denotes the projection on the set $\Omega$ of observed entries, $G$ is the ground truth matrix, and $\|\cdot\|_*$ denotes the *nuclear norm* (i.e., the sum of the matrix's singular values).

Besides the incomplete matrix, additional information may be available in applications such as movie recommendation or drug interaction prediction [8, 9, 10, 11]. For instance in movie recommendation, one may have access to the movies' genres, their synopsis, the gender and occupation of the users, or a friendship network between the users. *Inductive matrix completion* (IMC) [11, 12, 13, 14] exploits such side information. It assumes that the side information is summarized in matrices $X \in \mathbb{R}^{m\times d_1}$ and $Y \in \mathbb{R}^{n\times d_2}$, with the row vectors representing the users and items, respectively. IMC then optimizes the following objective function

$$\min_{M\in\mathbb{R}^{d_1\times d_2}} \frac{1}{2}\|P_\Omega(XMY^\top-G)\|_{\mathrm{Fr}}^2 + \lambda\|M\|_*. \tag{2}$$

This model has been used in many domains also beyond movie recommendation [8, 10, 15].

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

In this paper, we contribute to a better theoretical understanding of IMC and related methods in the approximate recovery case. In this setting we obtain guarantees in terms of a bound on the expected loss which decreases with the number of samples. Our best results concern the *distribution-free* case, meaning that our bounds are valid for any sampling distribution. This is in sharp contrast to the vast areas of literature where one assumes the distribution is uniform [16, 17, 18]. Our analysis leads to substantial gains compared to the the state of the art results [19, 20, 21], as we obtain near optimal bounds in situations where the state of the art bounds are vacuous, as is explained below.

Although for uniform sampling, near-tight exact recovery bounds of $O(rd\log(d)\log(n))$ exist for IMC [16, 17], the approximate recovery case (especially in a *distribution-free* setting) is far less understood. The state-of-the-art distribution-free results for IMC were proved in [19, 20] (and in [21] for a kernel formulation of IMC) and, expressed in terms of generalisation error bounds, scale as

$$O\left(\mathbf{x}\mathbf{y}\mathcal{M}\sqrt{1/N}\right), \tag{3}$$

where $\mathbf{x} := \|X^\top\|_{2,\infty} = \max_u \|X_{\bullet,u}\|^2$ is the maximum norm of a left side information vector (row of $X$), $N$ is the number of available samples, and $\mathbf{y} := \|Y^\top\|_{2,\infty} = \max_v \|Y_{\bullet,v}\|^2$ is the maximum norm of a right side information vector (row of $Y$). This implies that reaching a given loss threshold $\epsilon$ requires $O(\mathbf{x}^2\mathbf{y}^2\mathcal{M}^2/\epsilon^2)$ entries, where $\mathcal{M}$ is a bound on the nuclear norm of $M$. In this case, we say that the 'sample complexity' is $O(\mathbf{x}^2\mathbf{y}^2\mathcal{M}^2)$. To understand how those bounds scale with the matrix dimensions, consider the simple case where $X$ and $Y$ are made up of blocks of identity matrices. In that case, we have $\mathbf{x} = \mathbf{y} = 1$, yielding a sample complexity of $O(\mathcal{M}^2)$. Since $\|M\|_*^2 \sim d^2r$, this yields a bound of order $rd^2$.

Such bounds have a remarkable property: they do not depend on the size $n$ of the matrix and instead depend only on the size $d$ of the side information. This means that they capture the fact that valuable information can be extracted even for users and items for which no ratings are observed. On the other hand, these bounds have a strong dependence on the size $d$ of the side information. As an illustration, consider that they are vacuous when $X = I$ and $Y = I$, since the required number of entries $O(r^2) = O(rn^2)$ then grows faster than the total number of entries $n^2$. This is despite the fact that in that situation, distribution-free bounds for standard matrix completion yield a sample complexity of $O(n^{3/2}\sqrt{r})$ for the standard regulariser [22] and $O(nr\log(n))$ for a modified regulariser (the smoothed weighted trace norm from [23]). Thus, these existing distribution-free IMC bounds are very far from tight. In fact, they are only meaningful when the size of the side information is *negligible* compared to the general scale of the problem, which is a significant limitation in terms of the elegance of the theory (mismatch with MC bounds, separate proof techniques for separate regimes) and in practice (real-life side information could be very high-dimensional, especially if it is extracted from a neural network [24] or from a wide variety of different sources). To reinforce that point, note that any side information with a strong cluster structure[1] would exhibit similar failings to the above mentioned identity side information case.

In this work, we bridge the gap between the state-of-the art in matrix completion and inductive matrix completion with the trace norm by providing distribution-free bounds for IMC which combine both of the following advantages: (1) a lack of dependence in the size of the original matrices, and (2) a more refined dependence on the size of the side information: the dependence on $d$ in our bounds is almost the same as the dependence on $n$ (the size of the matrix) for the state-of-the-art MC results. More precisely, our first contribution is to provide a bound of order $O(d^{3/2}\sqrt{r\log(d)})$ for the standard regulariser (2). The proof builds on techniques from [22, 25], but is substantially more involved due to the complicated dependence structure generated by the side information. As our second contribution, we construct analogues of the ideas of [23, 26] for the IMC setting: we begin by showing a bound of order $O(rd\log(d))$ for a class of distributions with certain uniformity assumptions (our "uniform inductive marginals"), and then design a new "adjusted trace norm regulariser" for the problem (2) with similar properties to the weighted trace norm [26, 23] in MC. Instead of simply renormalising rows and columns of $M$ as in previous work, our method requires rescaling the core matrix $M$ along data-dependent orientations that capture interplay between the sampling distribution and the side information matrices $X, Y$.

Our contributions are summarised as follows.

[1] where the users and items are approximately split into 'communities', see also Appendix A

Table 1 – Matrix completion results (trace norm-based only)

| MC | Unif.Sampling | Distr.-free | Weighted version |
|---|---|---|---|
| Exact | $nr\log^2(n)$ ([27, 28, 30]) | N/A | N/A |
| Approx. | $nr\log(n)$ ([23, 22]) | $n^{3/2}\sqrt{r}$ ([22]) | $rn\log(n)$ ([23]) |

1. We provide distribution-free generalisation bounds for the inductive matrix completion model (2) (assuming a fixed upper bound on the nuclear norm) which scale like $O(d^{3/2}\sqrt{r\log(d)})$ where $r$ is a soft relaxation of the rank.
2. In the case of uniform or approximately uniform sampling, we provide a bound of order $O(rd\log(d))$ for *approximate recovery*.
3. We introduce a modified version of the IMC objective (2), which we refer to as *adjusted trace norm regularsation* (ATR). An empirical version E-ATR is also introduced and both achieve bounds of order $O(rd\log(d))$ in the distribution-free setting.
4. We experimentally demonstrate on synthetic data that our adjusted regulariser outperforms the standard IMC objective (2) in many situations.
5. We incorporate our method into a model involving a non-inductive term and evaluate it on real-life datasets, demonstrating substantially improved performance.

This paper is organized as follows. In Section 2 we review some related work. In Section 3 we introduce our main results. Finally, in Section 4 we present our experimental results.

## 2 Related work

In both MC and IMC, the existing literature consists of several main branches differing in their main assumptions: *exact recovery* versus *approximate recovery* and *uniform sampling* versus *distribution-free bounds*. In *exact recovery*, the matrix is assumed deterministic, and we want to recover its missing entries exactly [17, 16, 27, 28]. In *approximate recovery*, the matrix is assumed noisy, and we want to recover its missing entries only approximately, within some interval around their expectation [19, 20, 21, 18, 29]. Approximate recovery theory is typically expressed in terms of uniform generalisation bounds over a function class using a matrix-norm constraint. Assuming that the entries are sampled from a specific distribution (e.g., uniform), one typically can achieve much faster rates than *distribution-free* theory regardless of the distribution. The sample complexity of standard MC under uniform sampling is $O(nr\log(n)^2)$ for exact recovery (proved in the series of breakthrough papers [27, 28, 30]) and $O(nr\log(n))$ for approximate recovery [23][2].

The most closely related papers to ours are [22] and [23], which both work only on standard matrix completion *without side information*. In [22], a bound of order $O(n^{3/2}\sqrt{r})$ was obtained in the distribution-free setting for matrix completion with the trace norm, whilst in [23], rates of $O(rn\log(n))$ are shown for sampling with uniform marginals and for a smoothed version of the weighted trace norm regulariser in the distribution-free case. We almost perfectly extend most of the results from both papers to the inductive case, which requires many technical modifications.

Within the IMC framework the closest works are those which also deal with approximate recovery in the non uniform sampling case: [21, 31, 19, 20]. Their bounds, presented in many different contexts, translate to sample complexities of type $O(rd^2)$. Other celebrated works in the theoretical study of IMC include: [16] and [17], which showed rates of order $d^2r^3\log(d)$ and $rd\log(d)\log(n)$ respectively for exact recovery with uniform sampling, together with other important contributions (see appendix). Some bounds use a completely different regulariser (such as the max norm) to achieve better rates [32, 33] etc. These works also do not involve side information.

In Figures 1 and 2, we summarize state-of-the-art (s.o.t.) results in both MC and IMC. Note the problem of exact recovery in the distribution-free case is ill-defined (hence the N/As in our table).

[2] Those rates are actually comparable since approximate recovery results are typically not posthoc.

Table 2 – Inductive matrix completion results (trace norm-based only)

| IMC | Unif.Sampling | Distr.-free | Weighted version |
|---|---|---|---|
| Exact | $rd\log(d)\log(n)$ ([17, 18]) | N/A | N/A |
| | $d^2r^3\log(d)$ ([16]) | | |
| Approx. (s.o.t.) | $rd^2$ ([21, 31, 19]) | $rd^2$ ([21, 31, 19]) | None |
| Approx. (ours) | $rd\log(d)$ (Ours) | $d^{3/2}\sqrt{r\log(d)}$ (Ours) | $rd\log(d)$ (Ours) |

Other related works include (IMCNF) [19, 20], which proposed the following model:

$$\min \frac{1}{2} \sum_{(i,j)\in\Omega} |G_{i,j} - (XMY^\top + Z)_{i,j}|^2 + \lambda_1\|M\|_* + \lambda_2\|Z\|_*, \tag{4}$$

where $\lambda_1, \lambda_2$ are regularisation parameters, $G_{i,j}$ denotes the observed entries and the predictors take the form $(XMY^\top + Z)$. This model relies on the cross-validated hyperparameters $\lambda_1, \lambda_2$ balance the importance of the side information. The authors also showed results based on a combination of a bound for the inductive term $XMY^\top$ and a bound for the non inductive term $Z$. The non inductive terms in the bounds are similar to [22], whilst the bounds for the inductive term are proved from scratch and have also later appeared in a different form in [21, 31] together with a kernel formulation of IMC. In Subsection 4.2 we combine our framework with this strategy can reach competitive results.

Another very interesting work is [18] which introduces a joint model that imposes a nuclear norm-based constraint on both $M$ and $XMY^\top$ through a modification of the objective. The authors prove bounds for their method which match the state of the art in IMC [17, 19] and MC [22] when the side information is perfect and useless respectively. The dependence on the side information is better in our case. Further discussion of that paper is included in the appendix.

## 3 Main results

**Notation:** We observe $N$ entries of a ground truth matrix $G \in \mathbb{R}^{m\times n}$ which are sampled i.i.d (with replacement) through an arbitrary distribution $p$: we draw $(i,j) \in \{1,..,m\} \times \{1,...,n\}$ with probability $p_{i,j}$ where $\sum_{i,j} p_{i,j} = 1$. The sampled entries $\xi^1, \xi^2, \dots, \xi^N \in \{1,2,\dots,m\} \times \{1,2,\dots,n\}$ form a multiset $\Omega$: our setting allows for the observations to be noisy with a different noise distribution for each entry, but purely for notation convenience we treat the issue as if there is no noise. For a function $f : \mathbb{R} \to \mathbb{R}$ we will write $\sum_{(i,j)\in\Omega} f(G_{i,j})$ for the sum of the images of the observations, counted as many times as necessary [3]. We assume we are given side information matrices $X \in \mathbb{R}^{m\times d_1}$ and $Y \in \mathbb{R}^{n\times d_2}$. The maximum $L^2$ norm of a row of $X$ (resp. $Y$) is denoted by $\mathbf{x}$ (resp. $\mathbf{y}$). The minimums are denoted by $\underline{x}$ and $\underline{y}$ respectively. The row vectors of $X$ (resp. $Y$) are also written $x_i$ for $i \leqslant m$ (resp. $y_j$ for $j \leqslant n$ ). For any matrices $A, B$, $A \leqslant B$ means that $B - A$ is positive semi-definite, $\|A\|$ denotes the spectral norm of $A$ and $\|A\|_*$ denotes the nuclear norm of $A$. We have one fixed loss function $l$ used throughout the paper which is both *Lipschitz* with constant $\ell$ and bounded by $b$. For convenience we also frequently write $d$ instead of $\max(d_1, d_2)$. In the appendix, we provide a complete table of notations (Table I.1) that includes all notations introduced throughout the paper.

We now present our results, starting with the distribution-free bound for the standard regulariser, then moving on to the improved bounds under uniform sampling, and finally to our adjusted trace norm regulariser and the theoretical improvements it provides.

### 3.1 Distribution-free guarantees for the standard IMC objective

For a constant $\mathcal{M} \in \mathbb{R}$, we define the function class: $\mathcal{F}_{\mathcal{M},\mathcal{S}} = \{XMY^\top : \|M\|_* \leqslant \mathcal{M} \wedge \|M\| \leqslant \mathcal{S}\}$, which contains all predictors $XMY^\top$ where $M$ has its spectral norm bounded by $\mathcal{S}$ and nuclear norm

[3] More rigorously the observations are i.i.d of the form $(\xi^o, \bar{\xi}^o)$ with $\xi^o \in \{1,2,\dots,m\} \times \{1,2,\dots,n\}$ and $\bar{\xi}^o \in \mathbb{R}$ and write $\sum_{o=1}^N f(\bar{\xi}_o)$ instead of $\sum_{(i,j)\in\Omega} f(G_{i,j})$, and it should be assumed that the "ground truth" values $G$ (are defined so as to) minimize $\mathbb{E}(l(G_\xi, \bar{\xi}))$ for our loss function $l$ over the joint distribution of $\xi, \bar{\xi}$

bounded by $\mathcal{M}$. Our first main result is a uniform generalisation bound for the loss minimiser within this function class. Below we use the shorthand $l(A)$ (resp. $\hat{l}_S(A)$) for $\mathbb{E}_{(i,j)\sim p}(l(A_{i,j}, G_{i,j}))$ (resp. $\sum_{(i,j)\in\Omega} l(A_{i,j}, G_{i,j})/N$), the overall expected (resp. empirical) loss associated to matrix $A \in \mathbb{R}^{m\times n}$. In particular, in the noiseless setting, $\inf_{Z\in\mathcal{F}_{\mathcal{M},\mathcal{S}}} l(Z) = 0$ as long as $\|G\| \leqslant \mathcal{S}$ and $\|G\|_* \leqslant \mathcal{M}$.

**Theorem 3.1.** *Fix any target matrix $G$ and distribution $p$. Define $\hat{Z}_S = \arg\min(\hat{l}_S(Z) : Z \in \mathcal{F}_{\mathcal{M},\mathcal{S}})$. For any $\delta \in (0,1)$, with probability (w.p.) $\geqslant 1-\delta$ over the draw of the training set $\Omega$ we have*

$$l(\hat{Z}) \leqslant \inf_{Z\in\mathcal{F}_{\mathcal{M},\mathcal{S}}} l(Z) + C\ell[\mathbf{xy}\mathcal{S}]\sqrt{\frac{[\mathcal{M}/\mathcal{S}]\sqrt{d\log(d)}}{N}} + 3b\sqrt{\frac{\log(4/\delta)}{2N}} \tag{5}$$

*for some universal constant $C$. Furthermore, in expectation over the training set we have:*

$$l(\hat{Z}) \leqslant \inf_{Z\in\mathcal{F}_{\mathcal{M},\mathcal{S}}} l(Z) + C\ell[\sqrt{\Gamma}\mathcal{S}]\sqrt{\frac{[\mathcal{M}/\mathcal{S}]\sqrt{d\log(d)}}{N}} + 6b\sqrt{\frac{1}{N}}, \tag{6}$$

*where $\Gamma = \sum_{i,j}\|x_i\|^2\|y_j\|^2 p_{i,j}$, $b$ is a bound on the loss, $\ell$ is the Lipschitz constant of the loss $l$.*

The proof is provided in Appendix A. Note that writing $r$ for $\frac{\mathcal{M}}{\mathcal{S}d_1d_2}$, the main computational complexity term can be rewritten as $O\left[\ell\sqrt{\Gamma}\mathcal{S}\sqrt{\sqrt{d_1d_2dr\log(d)}/N}\right]$. In this context, the quantity $r$ acts as a soft relaxation of the rank. Indeed, note that if $\mathrm{rank}(M) \leqslant \rho$ and $\|M\| \leqslant \mathcal{S}$, then $\frac{\|M\|_*}{\mathcal{S}} \leqslant \sqrt{d_1d_2r}$. Furthermore, $\mathbf{xy}\mathcal{S}$ and $\sqrt{\Gamma}\mathcal{S}$ can be informally thought of as estimates of the 'maximum possible' or 'typical value' for an entry respectively.

Thus, our bound yields a sample complexity of order $\sqrt{d_1d_2dr\log(d)}$. In the case of identity side information, this bound matches the bound $n^{3/2}\sqrt{r}$ from [22] up to a logarithmic factor $\sqrt{\log(d)}$.

**Remark on the proof technique:** The proof of the result in [22] relies on a lemma of Latala (lemma A.1) from [34] for random matrices *with i.i.d. entries* and an elegant decomposition of the entries into two groups: (1) entries that have been sampled many times, and (2) entries that have not been sampled too often. On group 1, the partial sums of the Rademacher variables concentrate trivially, whilst on group 2, the entries are well spread out and Lemma A.1 limits the spectral norm similarly to the uniform case. The proof is about carefully balancing those two contributions.

In our inductive situation, using the same split can only yield bounds of the type (3) which are well known and vacuous when the side information is of comparable size to the matrix. Our key idea to fix this issue is that instead of distinguishing frequently and less frequently sampled *entries*, we split between high and low energy *orientations* corresponding to pairs $(X_{\cdot,u}, Y_{\cdot,v})$ of *columns* of the side information matrices. To achieve this aim, we use the rotational invariance of the trace operator and equivalently express the Rademacher averages in inductive space ($\mathbb{R}^{d_1\times d_2}$). However, the entries of the resulting matrix are certainly not independent, which makes it impossible to apply the concentration results from [34]. Instead, we bound the Rademacher complexity by a Gaussian complexity, and leverage powerful concentration results (Lemma F.1) from [35, 36, 37] for the spectral norms of *Gaussian* matrices with *correlated* entries. Obtaining a covariance structure that allows application of this Lemma further requires performing an iterative procedure involving series of data-dependent rotational transformations of the side information and other estimates at each step.

### 3.2 Generalisation bounds for the trace norm regulariser under a uniformity assumption

We now move to our second main contribution, which is a broad generalisation of most of the results of [23] to the *inductive* case. In this direction, we begin with a result for approximate recovery in inductive matrix completion with the standard nuclear norm regulariser. Although this first result (proved in Appendix B) is original to the best of our knowledge, it is not surprising since a similar result is known in the exact recovery case. However, it is an excellent way to introduce notation which will be necessary in the rest of the paper.

**Proposition 3.1.** *Let us write $\mathcal{F}_{\mathcal{M}}$ for the function class corresponding to matrices of the form $XMY^\top$ with $\|M\|_* \leqslant \mathcal{M}$. Let $M_S = \arg\min_{\|M\|_*\leqslant\mathcal{M}} \sum_{\xi\in\Omega} l((XMY^\top)_\xi, G_\xi)$ be the trained matrix $M$ and $M_* = \arg\min_{\|M\|_*\leqslant\mathcal{M}} \mathbb{E}l((XMY^\top)_\xi, G_\xi)$ be the optimal $M$ when $M$ is restricted by $\|M\|_* \leqslant \mathcal{M}$. Write also $Z_S = XM_SY^\top$ and $Z_* = XM_*Y^\top$.*

*Write* $\mathcal{K} := \max\left[\sqrt{d_1 \frac{\|X^\top X\|}{m}\frac{\|Y\|_{\mathrm{Fr}}^2}{n}}, \sqrt{d_2 \frac{\|Y^\top Y\|}{n}\frac{\|X\|_{\mathrm{Fr}}^2}{m}}\right]$. *Under uniform sampling, w.p.* $\geqslant 1-\delta$*:*

$$l(Z_S) - l(Z_*) \leqslant \frac{8\ell\mathcal{K}\sqrt{rd}(1+\sqrt{\log(2d)})}{\sqrt{N}} + \frac{12\ell}{N}\mathcal{M}\mathbf{xy}(1+\log(2d)) + b\sqrt{\frac{\log(2/\delta)}{2N}}, \tag{7}$$

*where* $r = \mathcal{M}/\sqrt{d_1 d_2}$ *and* $b$ *is a bound on the loss. Furthermore, the above result holds under the following more general "uniform inductive marginals" condition (analogous to the "uniform marginals"):*

$$\forall i, \quad \sum_{i,j} p_{i,j}\|y_j\|^2 = \frac{\|Y\|_{\mathrm{Fr}}^2}{mn} \quad \textit{and} \quad \forall j, \quad \sum_{i,j} p_{i,j}\|x_i\|^2 = \frac{\|X\|_{\mathrm{Fr}}^2}{mn}. \tag{8}$$

**Remarks:** If $\|x_i\|$ and $\|y_j\|$ are constant over $i$ and $j$, then the above conditions (8) reduce to a requirement of uniform marginal probabilities. Note that $r = (\mathcal{M}/\sqrt{d_1 d_2})$ acts as a soft relaxation of the rank of $\mathcal{M}$ since if $M \in \mathcal{F}_{\mathcal{M}}$ and the entries of $M$ are bounded by 1 then $\mathrm{rank}(M) \leqslant r$. If $X = I$ and $Y = I$, then conditions (8) reduce to the uniform marginals condition from [23].

In particular, we see that in the case of identity side information, we require $O(dr\log(r))$ samples to reach a given accuracy. However, the result above is deeper when the side information is non trivial. Indeed, the quantity $\max(\sqrt{\|X^\top X\|\|Y\|_{\mathrm{Fr}}^2}, \sqrt{\|Y^\top Y\|\|X\|_{\mathrm{Fr}}^2})$, which equals $d = \max(d_1, d_2)$ in the case of identity (or equal-size community) side information, is sensitive to the relative orientation of the columns of $X$ and $Y$: if the side information $X$ and $Y$ are properly scaled and approximately of rank $\rho$, then this quantity will approach $\rho$. We discuss this in more details in the appendix.

To prove the above result, we will show a slightly more general result below (Prop 3.2). In order to capture the interaction between the side information and the data-distribution, we must define a distribution-dependent inner product $\langle\cdot,\cdot\rangle_l$ (resp. $\langle\cdot,\cdot\rangle_r$) on the column space of $X$ (resp. $Y$):

For two vectors $u^1, u^2 \in \mathbb{R}^m$ (resp. $v^1, v^2 \in \mathbb{R}^n$) we define $\langle u^1, u^2\rangle_l = \sum_{i=1}^m u_i^1 u_i^2 q_i$ (resp. $\langle v^1, v^2\rangle_r = \sum_{j=1}^n v_j^1 v_j^2 \kappa_j$) where the $q_i$s and $\kappa_j$s are defined by

$$q_i = \sum_{j=1}^n p_{i,j}\|y_j\|^2 \quad \forall i \leqslant m \qquad\qquad \kappa_j = \sum_{i=1}^m p_{i,j}\|x_i\|^2 \quad \forall j \leqslant n. \tag{9}$$

We now define the vector $\sigma^1 \in \mathbb{R}^{d_1}$ (resp. $\sigma^2 \in \mathbb{R}^{d_2}$) as the vector of singular values of the matrix $X$ (resp. $Y$) *with respect to (w.r.t) the inner product* $\langle\cdot,\cdot\rangle_l$ *(resp.* $\langle\cdot,\cdot\rangle_r$*).* In other words, the entries of $\sigma^1 \in \mathbb{R}^{d_1}$ (resp. $\sigma^2 \in \mathbb{R}^{d_2}$) are the square roots of the eigenvalues of the symmetric matrix $L := X^\top \mathrm{diag}(q) X \in \mathbb{R}^{d_1\times d_1} = \sum_{i,j} p_{i,j} x_i x_i^\top \|y_j\|^2$ (resp. $R := Y^\top \mathrm{diag}(\kappa) Y = \sum_{i,j} p_{i,j} y_j y_j^\top \|x_i\|^2 \in \mathbb{R}^{d_2\times d_2}$). We also write $\sigma_*^1 = \max(\sigma^1)$ and $\sigma_*^2 = \max(\sigma^2)$.

**Proposition 3.2.** *Under the same assumptions as Proposition 3.1, w.p.* $\geqslant 1-\delta$ *over the draw of the training set* $\Omega$*:*

$$l(Z_S) - l(Z_*) \leqslant \frac{8\ell}{\sqrt{N}}\mathcal{M}\max(\sigma_*^1, \sigma_*^2)(1+\sqrt{\log(2d)}) + \frac{12\ell}{N}\mathcal{M}\mathbf{xy}(1+\log(2d)) + b\sqrt{\frac{\log(2/\delta)}{2N}}.$$

**Remarks:** Note that both $\sigma^1$ and $\sigma^2$ scale as the product of the scaling of $X$ and $Y$. The above result shows that if the distribution is only approximately uniform (sampling probabilities within a given ratio), then the bound is only penalised proportionately to this ratio: for identity side information, $[\sigma_*^1]^2$ is the maximum user (marginal) probability which scales like $1/d_1$ for approximately uniform marginals. Similarly $\sigma_*^2 \sim 1/d_2$, yielding a sample complexity bound of order $dr\log(d)$ as expected.

### 3.3 Proposed adjusted regularisers and notation

In this section, we introduce our adjusted trace norm regulariser and its variants. We first recall that in standard (non-inductive) matrix completion, the weighted trace norm [26, 23] of a matrix $Z$ is defined as $\sqrt{D}Z\sqrt{E}$ where $D \in \mathbb{R}^{m\times m}$ (resp. $E \in \mathbb{R}^{n\times n}$) are diagonal matrices whose diagonal entries contain the marginal row (resp. column) sampling probabilities. Regularising the weighted trace norm instead of the standard trace norm increases performance [26] and leads to better theoretical guarantees. In this work we extend those advantages to the setting where side information is available.

**Notation**: Recall $\Gamma = \sum_{i,j} p_{i,j}\|x_i\|^2\|y_j\|^2$. Our method is based on a careful distribution-dependent rescaling of the matrix $M$. The idea is that we must look at the principal directions (singular vectors) of the side information matrices, but computed with respect to a distribution-sensitive inner product: when computing inner products of vectors in the column space of $x$, components corresponding to highly users which are more likely to be sampled must be weighted more. Accordingly, we diagonalise the matrix $L = X^\top \operatorname{diag}(q)X$ (resp. $L = Y^\top \operatorname{diag}(\kappa)Y$ from above to write it $P^{-1}DP$ (resp. $Q^{-1}EQ$). We also define empirical versions of those quantities: $\widehat{\Gamma} = \frac{1}{N}\sum_{i,j} h_{i,j}\|x_i\|^2\|y_j\|^2$ where $h_{i,j}$ is the number of times that entry $(i,j)$ was sampled: $h_{i,j} = \sum_{o=1}^{N} 1_{\xi_o=(i,j)} = \#(\Omega \cap \{(i,j)\})$; $\hat{q}_i = \sum_j \frac{h_{i,j}}{N}\|y_j\|^2$, $\hat{\kappa}_j = \sum_i \frac{h_{i,j}}{N}\|x_i\|^2$, $\widehat{L} = X^\top \operatorname{diag}(\hat{q}))X$, $\widehat{R} = Y^\top \operatorname{diag}(\hat{\kappa})Y$, and their diagonalisations $\widehat{P}^{-1}\widehat{D}\widehat{P}$ and $\widehat{Q}^{-1}\widehat{E}\widehat{Q}$. We can now write our predictors

$$XMY^\top = XP^{-1}D^{\frac{1}{2}}[D^{-\frac{1}{2}}PMQ^{-1}E^{-\frac{1}{2}}]E^{\frac{1}{2}}QY^\top = X\widehat{P}^{-1}\widehat{D}^{\frac{1}{2}}[\widehat{D}^{-\frac{1}{2}}M\widehat{E}^{-\frac{1}{2}}]\widehat{E}^{\frac{1}{2}}\widehat{Q}Y^\top. \quad (10)$$

The simplest version of our proposed algorithm is to regularise $[D^{-\frac{1}{2}}PMQ^{-1}E^{-\frac{1}{2}}]$ instead of $M$.

However, some extra technical modifications may be necessary: If some users or items have extremely small sampling probability, the corresponding entries of $D^{-\frac{1}{2}}$ and $E^{-\frac{1}{2}}$ will be very large. To obtain good bounds, we tackle this issue by forcing the entries of $D, \widehat{D}, E, \widehat{E}$ to be bounded below, which we achieve via smoothing: fixing a parameter $\alpha \in [0,1]$, we define $\widetilde{D} = \alpha D + (1-\alpha)\Gamma I$ and $\widetilde{E} = \alpha E + (1-\alpha)\Gamma I$ where $I$ is the identity matrix. Similarly, $\breve{D} = \alpha\widehat{D} + (1-\alpha)\widehat{\Gamma}I$ and $\breve{E} = \alpha\widehat{E} + (1-\alpha)\hat{\Gamma}I$.

We also define accordingly $M' = D^{\frac{1}{2}}PMQ^{-1}E^{\frac{1}{2}}$; $\widehat{M} = \widehat{D}^{\frac{1}{2}}\widehat{P}M\widehat{Q}^{-1}\widehat{E}^{\frac{1}{2}}$; $\widetilde{M} = \widetilde{D}^{\frac{1}{2}}PMQ^{-1}\widetilde{E}^{\frac{1}{2}}$; and $\breve{M} = \breve{D}^{\frac{1}{2}}\widehat{P}M\widehat{Q}^{-1}\breve{E}^{\frac{1}{2}}$; as well as similarly $\widetilde{X} = XP^{-1}\widetilde{D}^{\frac{1}{2}}$, $X' = XP^{-1}D^{\frac{1}{2}}$, $\widehat{X} = X\widehat{P}^{-1}\widehat{D}^{\frac{1}{2}}$,$\breve{X} = X\widehat{P}^{-1}\breve{D}^{\frac{1}{2}}$, $\widetilde{Y} = YP^{-1}\widetilde{E}^{\frac{1}{2}}$, $Y' = YP^{-1}D^{\frac{1}{2}}$, $\widehat{Y} = Y\widehat{Q}^{-\frac{1}{2}}\widehat{D}^{-\frac{1}{2}}$, $\breve{Y} = Y\widehat{Q}^{-\frac{1}{2}}\breve{D}^{-\frac{1}{2}}$. Thus $XMY^\top = X'M'[Y']^\top = \widetilde{X}\widetilde{M}\widetilde{Y}^\top = \widehat{X}\widehat{M}\widehat{Y}^\top = \breve{X}\breve{M}\breve{Y}^\top$.

**Proposed models:** We then propose a variety of adjusted regularisation strategies as follows by replacing the regularisation of $M$ by that of $M'$, $\widetilde{M}$, $\widehat{M}$ or $\breve{M}$ depending on whether the ground truth distribution is known and whether smoothing is desired. For instance, in the smoothed, empirical case, we will solve the following optimization problem:

$$\min_M \frac{1}{N}\sum_{\xi\in\Omega} l((XM_SY^\top)_\xi, G_\xi) + \lambda\|\breve{D}^{\frac{1}{2}}\widehat{P}M\widehat{Q}^{-1}\breve{E}^{\frac{1}{2}}\|_*. \quad (11)$$

**Important remark:** Similarly to the matrix case the smoothing parameter $\alpha$ is set to $\frac{1}{2}$ in all theorem statements[4]. In the experiments, we vary $\alpha$ as indicated.

We will prove results for the empirical risk minimiser belonging to the following function classes:

$$\widetilde{\mathcal{F}}_r := \left\{XMY^\top : \|\widetilde{M}\|_* \leqslant \sqrt{r}\Gamma\right\} \qquad \breve{\mathcal{F}}_r := \left\{XMY^\top : \|\breve{M}\|_* \leqslant \sqrt{r}\widehat{\Gamma}\right\}, \quad (12)$$

corresponding to the smoothed and smoothed empirical versions of our algorithm. Note that the factors of $\Gamma$ are added purely for convenience in the final formula, so that we can understand the final formulae in terms of a soft concept of "rank". Indeed we have

$$\|\widetilde{D}^{\frac{1}{2}}\|_{\mathrm{Fr}}^2 \leqslant d_1\frac{\Gamma}{2d_1} + \frac{1}{2}\|\sqrt{\operatorname{diag}(q)}X\|_{\mathrm{Fr}}^2 = (1/2)\Gamma + (1/2)\sum_{i,u} X_{i,u}^2 \sum_j p_{i,j}\|y_j\|^2 = \Gamma, \quad (13)$$

and similarly $\|\widetilde{E}^{\frac{1}{2}}\|_{\mathrm{Fr}}^2 \leqslant \Gamma$. Thus if $\|M\|_\infty \leqslant 1$ and $\operatorname{rank}(M) \leqslant \rho$, we have $\|\widetilde{M}\|_* \leqslant \sqrt{\rho}\|\widetilde{M}\|_{\mathrm{Fr}} \leqslant \sqrt{\rho}\Gamma$. Similarly, $\|\breve{M}\|_* \leqslant \sqrt{\rho}\widehat{\Gamma}$ under the same condition.

### 3.4 Generalisation bounds for the smoothed adjusted trace norm

Although knowing the distribution is not realistic, it is instructive to see that one can obtain guarantees of order $O(dr\log(d))$ for the function class $\widetilde{\mathcal{F}}_r$ as a reasonably straightforward extension of the ideas developed for Proposition 3.2. The proof is provided in Appendix C.

[4] It is trivial to extend the proofs to arbitrary $\alpha$ at the cost of a factor of $1/\min(\alpha, 1-\alpha)$.

**Proposition 3.3.** *Let* $\widetilde{M}_S = \arg\min_{\|\widetilde{M}\| \leqslant \sqrt{r}\Gamma} \sum_{\xi \in \Omega} l((\widetilde{X}\widetilde{M}\widetilde{Y}^\top)_\xi, G_\xi)$ *be the trained matrix* $\widetilde{M}$ *and* $\widetilde{Z}_* = \arg\min_{Z \in \widetilde{\mathcal{F}}_r} \mathbb{E} l(Z_\xi, G_\xi)$ *be the optimal* $\widetilde{M}$ *when the predictors are restricted to the class* $\widetilde{\mathcal{F}}_r$*. Let also* $\widetilde{Z}_S = \widetilde{X}\widetilde{M}_S\widetilde{Y}^\top$*. We have w.p.* $\geqslant 1 - \delta$*:*

$$l(\widetilde{Z}_S) - l(\widetilde{Z}_*) \leqslant \frac{16\ell\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1 + \sqrt{\log(2d)})}{\sqrt{N}} + \frac{24\ell\mathbf{xy}\sqrt{d_1 d_2 r}(1 + \log(2d))}{N} + b\sqrt{\frac{\log(2/\delta)}{2N}}.$$

### 3.5 Generalisation bounds for the smoothed empirically adjusted trace norm

Below is a more challenging result (proof in Appendix D) which concerns the function class $\check{\mathcal{F}}_r$ corresponding to the empirically smoothed regulariser.

**Theorem 3.2.** *Fix any target matrix* $G$ *and distribution* $p$*. Define* $\check{Z}_S = \arg\min(\hat{l}_S(Z) : Z \in \check{\mathcal{F}}_r)$ *where* $\hat{l}_S(Z) = \frac{1}{N}\sum_{(i,j)\in\Omega} l(Z_{i,j}, G_{i,j})$*. For any* $\delta \in (0,1)$*, w.p.* $\geqslant 1 - \delta$

$$l(\check{Z}) \leqslant \inf_{Z \in \widetilde{\mathcal{F}}_r} l(Z) + C\left[\ell\sqrt{r}\gamma(\mathbf{x}+\mathbf{y})^2 + b\right]\sqrt{\frac{\gamma^2 d \log(\frac{d}{\delta})}{N}}, \tag{14}$$

*where* $\gamma = \frac{\mathbf{x}^2\mathbf{y}^2}{\underline{x}^2\underline{y}^2}$ *and* $C$ *is a universal constant. In particular, in expectation over the draw of the training set we have*

$$l(\check{Z}) \leqslant \inf_{Z \in \widetilde{\mathcal{F}}_r} l(Z) + 2C\left[\ell\sqrt{r}\gamma(\mathbf{x}+\mathbf{y})^2 + b\right]\sqrt{\frac{\gamma^2 d \log(d)}{N}}. \tag{15}$$

The significance of this result is that even in the case of an arbitrary distribution, minimizing the smoothed empirical adjusted nuclear norm $\|\widetilde{M}\|_*$ results in sample complexity bounds of order $dr\log(d)$, meaning that our distribution-dependent transformations have completely removed the negative effects of non-uniformity on the sample complexity. Note the proof requires careful technical variations compared to the proof of the comparable results in [23]. As an example, Lemma E.1 is the equivalent of Lemma 2 in page 8 of the supplementary in [23] (whose proof is far shorter).

## 4 Experimental verification

In this section, we experimentally validate the advantages of our adjusted regularisation strategies described in Subsection 3.3. In all experiments, we work with the square loss.

### 4.1 Experiments on synthetic data

We construct square data matrices in $\mathbb{R}^{n\times n}$ with a given rank $r \leqslant d$ for several combinations of $n, d, r$. We provide each model with $d$-dimensional side information spanning the row and columns spaces. The sampling distribution is a power-type law depending on $\Lambda$ such that $\Lambda = 0$ yields uniform sampling (details in appendix). We compare three approaches: (1) Standard inductive matrix completion with the side information matrices $X, Y$ (IMC) (2) Our smoothed adjusted regulariser $\lambda\|\widetilde{M}\|$ (for several values of $\alpha$) (ATR)[5]; and finally (3) our smoothed empirically adjusted regulariser $\lambda\|\widetilde{M}\|$ (for several values of $\alpha$) (E-ATR). For each $n \in \{100, 200\}$ we evaluate the following $d, r$ combinations: $(30, 4)$, $(50, 6)$ and $(80, 10)$. In order to study a meaningful data-sparsity regime, in each case we sampled $dr\omega$ entries where $\omega \in \{1, 2, 3, 4, 5\}$. We show the most representative results here. More comprehensive results are provided in the supplementary material.

We observe that our methods outperform standard inductive matrix completion by significant margins in many regimes, even in the case of uniform sampling. Furthermore, the empirical version of our model actually often performs better than the exact one, which matches the observations made in [23] in the case of standard matrix completion. More detailed results are reported in the appendix.

[5] Note that in this synthetic context, it is actually possible to compute $\widetilde{M}$ since the distribution is known.

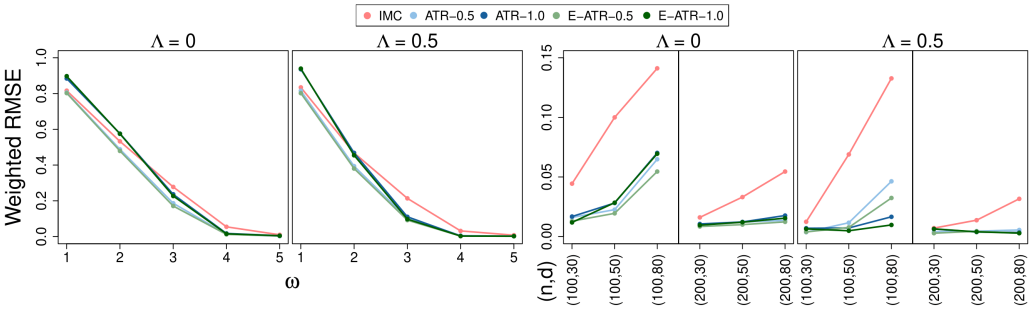


Figure 1 – Left: performance as a function of the data sparsity parameter $\omega$ for $n, d, r = 200, 80, 10$. Right: Performance on different $n, d, r$ combinations for $\omega = 4$. Legend: parameter to the right is $\alpha$.

Table 3 – Results of real-world datasets (RMSE)

| | **SoftImpute** [6] | **IMCNF** [19] | **E-ATR-0.5** | **E-ATR-0.75** | **E-ATR-1.0** |
|---|---|---|---|---|---|
| **Douban** | 0.9582 | 0.8197 | 0.7691 | **0.7614** | 0.8779 |
| **LastFM** | 2.4109 | 1.7612 | **1.6159** | 1.6943 | 2.3371 |
| **MovieLens** | 0.9280 | 0.9252 | **0.9056** | 0.9139 | 0.9262 |

### 4.2 Real data experiments

We evaluate the performance of our model on three real life datasets: Douban, LastFM and MovieLens (further described in the supplementary). In real data we work with the following adjusted version of the model in [19]:

$$\min_{M,Z} \frac{1}{N} \sum_{(i,j)\in\Omega} l(XMY^\top + Z, G_{i,j}) + \lambda_1 \|\breve{D}^{\frac{1}{2}} \widehat{P} M \widehat{Q}^{-1} \breve{E}^{\frac{1}{2}}\|_* + \lambda_2 \|\breve{D}_I^{\frac{1}{2}} Z \breve{E}_I^{\frac{1}{2}}\|_* \tag{16}$$

where $\breve{D}, \breve{E}$ are defined as above based on the side information matrices $X, Y$, and $\breve{D}_I, \breve{E}_I$ are defined as $\breve{D}, \breve{E}$ except based on the side information matrices $(I, I)$. In particular, $\|\breve{D}_I^{1/2} Z \breve{E}_I^{1/2}\|_* = \|\breve{Z}\|_*$ is the smoothed weighted trace norm of $Z$ in the sense of [23]. We report results in Table 3 and note our method outperforms both SoftImpute and IMCNF, especially with appropriate smoothing.

## 5 Conclusion

In this paper, we have provided the first distribution-free bounds for approximate recovery in inductive matrix completion with the trace norm with the following two desirable properties: (1) being non vacuous for identity or community side information and (2) being completely independent of the size of the matrix. We presented an adjusted regularisation strategy which relies on a careful rescaling along distribution-dependent directions that captures the interaction between the side information matrices and the sampling distribution. Our bounds, which concern both the standard regulariser (rate $O(d^{3/2}\sqrt{r\log(d)})$) and our adjusted version (rate $O(dr\log(d))$) are almost exactly what one would obtain by replacing the size of the matrix with the size of the side information in the standard matrix completion bound. Thus, we have bridged the large gap between the theoretical guarantees for matrix completion and inductive matrix completion. There are many potentially interesting problems for future work: one possibility would be to try to incorporate our distribution-sensitive renormalisation strategy into other models such as [18] or max-norm based models [33, 32].

## Broader impact

The work in this paper is theoretical and without any foreseeable significant societal impact.

# References


[1] Zhao Kang, Chong Peng, and Qiang Cheng. Top-n recommender system via matrix completion. 2016.

[2] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[3] Cho-Jui Hsieh, Kai-Yang Chiang, and Inderjit S. Dhillon. Low rank modeling of signed networks. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 507–515, New York, NY, USA, 2012.

[4] Vassilis Kalofolias, Xavier Bresson, Michael Bronstein, and Pierre Vandergheynst. Matrix Completion on Graphs. *arXiv e-prints*, page arXiv:1408.1717, August 2014.

[5] Hao Ma, Dengyong Zhou, Chao Liu, Michael Lyu, and Irwin King. Recommender systems with social regularization. pages 287–296, 01 2011.

[6] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, 11:2287–2322, August 2010.

[7] Trevor Hastie, Rahul Mazumder, Jason D. Lee, and Reza Zadeh. Matrix completion and low-rank svd via fast alternating least squares. *Journal of Machine Learning Research*, 16(104):3367–3402, 2015.

[8] Rong Li, Yongcheng Dong, Qifan Kuang, Yiming Wu, Yizhou Li, Min Zhu, and Menglong Li. Inductive matrix completion for predicting adverse drug reactions (adrs) integrating drug–target interactions. *Chemometrics and Intelligent Laboratory Systems*, 144:71 – 79, 2015.

[9] H. Wang, Y. Wei, M. Cao, M. Xu, W. Wu, and E. P. Xing. Deep inductive matrix completion for biomedical interaction prediction. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 520–527, 2019.

[10] Nagarajan Natarajan and Inderjit S. Dhillon. Inductive matrix completion for predicting gene and disease associations. *Bioinformatics*, 30(12):i60–i68, 06 2014.

[11] Mark Herbster, Stephen Pasteris, and Lisa Tse. Online matrix completion with side information. *CoRR*, abs/1906.07255, 2019.

[12] Xiao Zhang, Simon Du, and Quanquan Gu. Fast and sample efficient inductive matrix completion via multi-phase procrustes flow. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5756–5765, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[13] Aditya Krishna Menon and Charles Elkan. Link prediction via matrix factorization. In *Machine Learning and Knowledge Discovery in Databases*, pages 437–452. Springer Berlin Heidelberg, 2011.

[14] Tianqi Chen, Weinan Zhang, Qiuxia lu, Kailong Chen, Zhao Zheng, and Yong Yu. Svdfeature: A toolkit for feature-based collaborative filtering. *The Journal of Machine Learning Research*, 2012.

[15] Fabian Jirasek, Rodrigo A. S. Alves, Julie Damay, Robert A. Vandermeulen, Robert Bamler, Michael Bortz, Stephan Mandt, Marius Kloft, and Hans Hasse. Machine learning in thermodynamics: Prediction of activity coefficients by matrix completion. *The Journal of Physical Chemistry Letters*, 11(3):981–985, 2020.

[16] Prateek Jain and Inderjit S. Dhillon. Provable inductive matrix completion. *CoRR*, abs/1306.0626, 2013.

[17] Miao Xu, Rong Jin, and Zhi-Hua Zhou. 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, page 2301–2309, Red Hook, NY, USA, 2013. Curran Associates Inc.

[18] Jin Lu, Guannan Liang, Jiangwen Sun, and Jinbo Bi. A sparse interactive model for matrix completion with side information. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

[19] Kai-Yang Chiang, Inderjit S. Dhillon, and Cho-Jui Hsieh. Using side information to reliably learn low-rank matrices from missing and corrupted observations. *J. Mach. Learn. Res.*, 2018.

[20] Kai-Yang Chiang, Cho-Jui Hsieh, and Inderjit S Dhillon. Matrix completion with noisy side information. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[21] P. Giménez-Febrer, A. Pagès-Zamora, and G. B. Giannakis. Generalization error bounds for kernel matrix completion and extrapolation. *IEEE Signal Processing Letters*, 27:326–330, 2020.

[22] Ohad Shamir and Shai Shalev-Shwartz. 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*, pages 661–678. PMLR, 2011.

[23] Rina Foygel, Ohad Shamir, Nati Srebro, and Russ R Salakhutdinov. Learning with the weighted trace-norm under arbitrary sampling distributions. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2133–2141. Curran Associates, Inc., 2011.

[24] Kai Zhong, Zhao Song, Prateek Jain, and Inderjit S Dhillon. Provable non-linear inductive matrix completion. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[25] Ohad Shamir and Shai Shalev-Shwartz. Matrix completion with the trace norm: Learning, bounding, and transducing. *Journal of Machine Learning Research*, 15:3401–3423, 2014.

[26] Nathan Srebro and Russ R Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems 23*, pages 2056–2064. Curran Associates, Inc., 2010.

[27] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor.*, 56(5):2053–2080, May 2010.

[28] Benjamin Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12(null):3413–3430, December 2011.

[29] Raghunandan Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 952–960. Curran Associates, Inc., 2009.

[30] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009.

[31] P. Giménez-Febrer, A. Pagès-Zamora, and G. B. Giannakis. Matrix completion and extrapolation via kernel regression. *IEEE Transactions on Signal Processing*, 67(19):5004–5017, 2019.

[32] T. Tony Cai and Wen-Xin Zhou. Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, 10(1):1493 – 1525, 2016.

[33] Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In Peter Auer and Ron Meir, editors, *Learning Theory*, pages 545–560, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg.

[34] Rafał Latała. Some estimates of norms of random matrices. *Proceedings of the American Mathematical Society*, 133(5):1273–1282, 2005.

[35] Afonso S. Bandeira and March T. Boedihardjo. The spectral norm of gaussian matrices with correlated entries, 2021.

[36] Roberto Oliveira. Sums of random Hermitian matrices and an inequality by Rudelson. *Electronic Communications in Probability*, 15(none):203 – 212, 2010.

[37] Artur Buchholz. Operator khintchine inequality in non-commutative probability. *Mathematische Annalen*, 319(1):1–16, 2001.

## Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] The conditions are clearly stated for each theorem, and we further discuss open questions and directions in the Appendix.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] The main part of the work is theoretical and the algorithmic part of it only improves the performance of existing methods on existing problems, rather than introducing a new machine learning problem.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the license to the code and datasets? [Yes] The datasets are freely available and cited in the appendix. The code will be made freely available for the camera ready version if accepted.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] This is discussed in the appendix
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Yes, we include box-plots in the supplementary.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] In the supplementary
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] The datasets and links are in the supplementary
   (b) Did you mention the license of the assets? [N/A] The datasets/assets are openly available
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]