Fine-grained Generalisation Analysis of Inductive Matrix Completion

Antoine Ledent
TU Kaiserslautern
ledent@cs.uni-kl.de

Rodrigo Alves
TU Kaiserslautern
alves@cs.uni-kl.de

Yunwen Lei
University of Birmingham
y.lei@bham.ac.uk

Marius Kloft
TU Kaiserslautern
kloft@cs.uni-kl.de

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 sample complexity bounds improving the previously best rate 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_Ω denotes the projection on the set Ω 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].

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

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

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 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[1] 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 ϵ 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 $O(rd^2) = O(rn^2)$ 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[2] 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] with some orthogonality assumptions on the side information

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

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 typical sample complexity of standard MC under uniform sampling is $O(nr\log^2(n))$ for exact recovery (proved in the series of breakthrough papers [27, 28, 30]) and $O(nr\log(n))$ for approximate recovery [23]. In [31, 32], an improved rate of $nr\log(n)\log(r)$ (for exact recovery) was shown.

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, 33, 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). In the case of exact recovery, the rate of $rd\log(d)\log(n)$ was obtained only under the assumption that the side information matrices have orthonormal columns. Some bounds use a completely different regulariser (such as the max norm) to achieve better rates [34, 35] etc. These works also do not involve side information.

In Figures 1 and 2, we summarize state-of-the-art (s.o.t.a.) 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). In approximate recovery bounds, we omit a factor of $1/\epsilon^2$, where ϵ is the tolerance threshold in terms of expected loss), as this factor is present in all approximate recovery bounds[3]. In exact recovery bounds, the rate is the order of magnitude of the threshold past which exact recovery occurs with high probability.

[3] To our best knowledge, all results show a decline in population expected loss of the order of $\sqrt{1/N}$ where N is the sample size

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
	$nr\log(n)\log(r)$ ([31, 32])		
Approx.	$nr\log(n)$ ([23, 22])	$n^{3/2}\sqrt{r}$ ([22])	$rn\log(n)$ ([23])

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.a.)	rd^2 ([21, 33, 19])	rd^2 ([21, 33, 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:

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

where λ_1, λ_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 λ_1, λ_2 to 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, 33] together with a kernel formulation of IMC. In Subsection 4.2 we combine our framework with this strategy to reach competitive results.

In [36], the authors introduce a model consisting of a sum of mutually orthogonal IMC terms together with an explicit optimization strategy in the specific case where the available side information consists in partitions of the users and items into communities. In [37], the authors further extend the model to learn the community membership functions together with the ground truth matrix, based only on the sampled entries. The case of a single IMC term where the side information is in the form of a community partition is useful to develop intuition into the equivalent roles of d_1, d_2 in our bounds versus m, n in MC bounds. Whilst generalization bounds were proved in [36] with a similar scaling as our bound from Thm 3.1 (and in particular are better than the state-of-the-art IMC bounds if applied to this situation), they only apply to the specific case of community side information. In this work (Theorem 3.1) we achieve the first IMC bounds which cover the whole range of possible side information matrices X, Y, whilst providing the correct scaling (up to log terms) in the case of identity or community side information. Community side information has also been studied in other discrete contexts where individual behaviour is assumed to be a noisy realisation of community side information [38, 39].

Another 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. Of course, there are also many other works which propose modified optimization problems for the Recommender Systems task through other rank-sparsity inducing regularisers [35, 34, 40] and even exploiting other ground truth structure besides the low-rank property [41, 42].

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 $\xi^1, \xi^2, \ldots, \xi^N \in \{1, 2, \ldots, m\} \times$

$\{1, 2, \ldots, n\}$ form a multiset Ω: our setting allows for the observations to be noisy with a different noise distribution for each entry, but purely for notation convenience we often treat the issue as if there is no noise when no ambiguity is possible. When written explicitly, the noise is denoted by ζ. 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 [4]. 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 ℓ 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 K.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: $\mathcal{F}_\mathcal{M} = \left\{XMY^\top : \|M\|_* \leqslant \mathcal{M}\right\}$, which contains all predictors $XMY^\top$ where M has its spectral norm bounded by $\mathcal{S}$. 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. $\hat{l}_S(A)$ or even $l_S(A)$) for $\mathbb{E}_{(i,j)\sim p}(l(A_{i,j}, G_{i,j} + \zeta_{i,j}))$ (resp. $\sum_{(i,j)\in\Omega} l(A_{i,j}, G_{i,j} + \zeta_{i,j})/N$), the overall expected (resp. empirical) loss associated to matrix $A \in \mathbb{R}^{m\times n}$. In particular, in the noiseless setting, $\inf_{Z\in\mathcal{F}_\mathcal{M}} l(Z) = 0$ as long as $\|G\|_* \leqslant \mathcal{M}$.

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

$$l(\hat{Z}) \leqslant \inf_{Z\in\mathcal{F}_\mathcal{M}} l(Z) + C\left[\sqrt{\frac{\ell b\mathbf{x}\mathbf{y}\mathcal{M}\sqrt{d}}{N}}\Psi + \frac{b}{\sqrt{N}} + \frac{\mathbf{x}\mathbf{y}\ell\mathcal{M}+\ell}{N}\log(2d)\right] + 4b\sqrt{\frac{\log(2/\delta)}{2N}}, \quad (5)$$

where C is a universal constant, b is a bound on the loss, ℓ is the Lipschitz constant of the loss l, and $\Psi = \left[\sqrt{\log(2d)} + \sqrt{\log(N(20\mathcal{M}^2\ell\sqrt{d}[\mathbf{x}^2\mathbf{y}^2]/b+1)}\right]$ is a logarithmic quantity. Furthermore, in expectation over the training set we have:

$$l(\hat{Z}) \leqslant \inf_{Z\in\mathcal{F}_\mathcal{M}} l(Z) + C\left[\sqrt{\frac{\ell b\mathbf{x}\mathbf{y}\mathcal{M}\sqrt{d}}{N}}\Psi + \frac{b}{\sqrt{N}} + \frac{\mathbf{x}\mathbf{y}\ell\mathcal{M}+\ell}{N}\log(2d)\right] + 20b\sqrt{\frac{1}{N}}. \quad (6)$$

The proof is provided in Appendix A. Assuming that ℓ, b are treated as constants, the above bound on the generalisation gap $l(\hat{Z}) - \inf_{Z\in\mathcal{F}_\mathcal{M}} l(Z)$ scales like

$$O\left(\frac{\mathbf{x}\mathbf{y}\mathcal{M}}{N}\log(d) + \sqrt{\frac{\mathbf{x}\mathbf{y}\mathcal{M}\sqrt{d}}{N}}\left[\sqrt{\log(d)} + \sqrt{\log(N)} + \sqrt{\log(\mathbf{x}\mathbf{y}\mathcal{M})}\right]\right). \quad (7)$$

If we further think of the maximum entry of the core matrix M as bounded by a constant, $\mathcal{M}$ scales like $\sqrt{d_1 d_2 r}$ where r is the rank of M. Assuming the rescaling is also set so that $\mathbf{x}, \mathbf{y}$ are constants, the above yields a sample complexity of

$$O\left(\frac{\sqrt{d_1 d_2}\sqrt{dr}\log(d)}{\epsilon^2}\right),$$

[4] More rigorously the observations are i.i.d of the form $(\xi^o, \bar{\xi}^o)$ with $\xi^o \in \{1, 2, \ldots, m\} \times \{1, 2, \ldots, n\}$ 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 $\xi, \bar{\xi}$

where ϵ is the tolerance threshold. Indeed, the $\sqrt{\log(N)}$ term can be treated via the following simple observation: If $N \geqslant \Theta\log(\Theta)$ and Θ is sufficiently large then

$$\frac{N}{\log(N)} \geqslant \frac{\Theta\log(\Theta)}{\log(\Theta)+\log(\log(\Theta))} \geqslant \frac{\Theta\log(\Theta)}{2\log(\Theta)} \geqslant \Theta/2.$$

Remark on the proof technique: The proof of the result in [22] relies on a lemma of Latala (lemma A.1) from [43] 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 $(X_{\bullet,u}, Y_{\bullet,v})$ 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 [43]. Instead, we must rely again on the matrix Bernstein inequality F.4 Obtaining a covariance structure that is amenable to application of this result requires performing an iterative procedure involving series of *distribution 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 $M_S = \arg\min_{\|M\|_*\leqslant\mathcal{M}} \sum_{\xi\in\Omega} l((XMY^\top)_\xi, G_\xi+\zeta_\xi)$ be the trained matrix M and $M_* = \arg\min_{\|M\|_*\leqslant\mathcal{M}} \mathbb{E}l((XMY^\top)_\xi, G_\xi+\zeta_\xi)$ 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. $\geqslant 1-\delta$:

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

where $\sqrt{r} = \mathcal{M}/\sqrt{d_1d_2}$ 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"):

$$\forall i, \quad \sum_{i,j} p_{i,j}\|y_j\|^2 = \frac{\|Y\|_{\mathrm{Fr}}^2}{mn} \quad \textit{and} \quad \forall j, \quad \sum_{i,j} p_{i,j}\|x_i\|^2 = \frac{\|X\|_{\mathrm{Fr}}^2}{mn}. \tag{9}$$

Remarks: If $\|x_i\|$ and $\|y_j\|$ are constant over i and j, then the above conditions (9) reduce to a requirement of uniform marginal probabilities. Note that $\sqrt{r} = (\mathcal{M}/\sqrt{d_1d_2})$ 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 (9) 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 ρ, then this quantity will approach ρ. 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_is and κ_js are defined by

$$q_i = \sum_{j=1}^{n} p_{i,j} \|y_j\|^2 \quad \forall i \leqslant m \qquad\qquad \kappa_j = \sum_{i=1}^{m} p_{i,j} \|x_i\|^2 \quad \forall j \leqslant n. \tag{10}$$

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 $L := X^\top \operatorname{diag}(q) X \in \mathbb{R}^{d_1 \times d_1} = \sum_{i,j} p_{i,j} x_i x_i^\top \|y_j\|^2$ (resp. $R := Y^\top \operatorname{diag}(\kappa) Y = \sum_{i,j} p_{i,j} y_j y_j^\top \|x_i\|^2 \in \mathbb{R}^{d_2 \times d_2}$). We also write $\sigma_*^1 = \max(\sigma^1)$ and $\sigma_*^2 = \max(\sigma^2)$.

Proposition 3.2. *With the same notation as in Proposition 3.1, w.p.* $\geqslant 1 - \delta$ *over the draw of the training set* Ω*:*

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

Remarks: Note that both σ^1 and σ^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 $P^{-1} D P$ (resp. $Q^{-1} E Q$). 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

$$XMY^\top = XP^{-1} D^{\frac{1}{2}} [D^{-\frac{1}{2}} PMQ^{-1} E^{-\frac{1}{2}}] E^{\frac{1}{2}} Q Y^\top = X\widehat{P}^{-1} \widehat{D}^{\frac{1}{2}} [\widehat{D}^{-\frac{1}{2}} M \widehat{E}^{-\frac{1}{2}}] \widehat{E}^{\frac{1}{2}} \widehat{Q} Y^\top. \tag{11}$$

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 $\widetilde{D} = \alpha D + (1-\alpha)\Gamma I/d_1$ and $\widetilde{E} = \alpha E + (1-\alpha)\Gamma I/d_2$ where I is the identity matrix. Similarly, $\breve{D} = \alpha \widehat{D} + (1-\alpha)\widehat{\Gamma} I/d_1$ and $\breve{E} = \alpha \widehat{E} + (1-\alpha)\widehat{\Gamma} I/d_2$.

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 $\widetilde{X} = XP^{-1}\widetilde{D}^{-\frac{1}{2}}$, $X' = XP^{-1}D^{-\frac{1}{2}}$, $\widehat{X} = X\widehat{P}^{-1}\widehat{D}^{-\frac{1}{2}}$,$\breve{X} = X\widehat{P}^{-1}\breve{D}^{-\frac{1}{2}}$, $\widetilde{Y} = YP^{-1}\widetilde{E}^{-\frac{1}{2}}$, $Y' = YP^{-1}D^{-\frac{1}{2}}$, $\widehat{Y} = Y\widehat{Q}^{-1}\widehat{D}^{-\frac{1}{2}}$, $\breve{Y} = Y\widehat{Q}^{-1}\breve{D}^{-\frac{1}{2}}$. 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:

$$\min_M \frac{1}{N} \sum_{\xi \in \Omega} l((XMY^\top)_\xi, G_\xi + \zeta_\xi) + \lambda \|\breve{D}^{\frac{1}{2}} \widehat{P} M \widehat{Q}^{-1} \breve{E}^{\frac{1}{2}}\|_*. \tag{12}$$

Remark: Similarly to the matrix case the smoothing parameter α is set to $\frac{1}{2}$ in all theorem statements [5]. In the experiments, we vary α as indicated.

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

$$\widetilde{\mathcal{F}}_r := \left\{ XMY^\top : \|\widetilde{M}\|_* \leqslant \sqrt{r}\Gamma \right\} \qquad \breve{\mathcal{F}}_r := \left\{ XMY^\top : \|\breve{M}\|_* \leqslant \sqrt{r}\widehat{\Gamma} \right\}, \tag{13}$$

corresponding to the smoothed and smoothed empirical versions of our algorithm. Note that the factors of Γ 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

$$\|\widetilde{D}^{\frac{1}{2}}\|_{\mathrm{Fr}}^2 \leqslant d_1 \frac{\Gamma}{2d_1} + \frac{1}{2} \|\sqrt{\mathrm{diag}(q)} X\|_{\mathrm{Fr}}^2 = (1/2)\Gamma + (1/2) \sum_{i,u} X_{i,u}^2 \sum_j p_{i,j} \|y_j\|^2 = \Gamma, \tag{14}$$

and similarly $\|\widetilde{E}^{\frac{1}{2}}\|^2 \leqslant \Gamma$. Thus if $\|M\|_\infty \leqslant 1$ and $\mathrm{rank}(M) \leqslant \rho$, we have $\|\widetilde{M}\|_* \leqslant \sqrt{\rho}\|\widetilde{M}\|_{\mathrm{Fr}} \leqslant \sqrt{\rho}\sqrt{\sum_{u,v} [\widetilde{D}_u^{\frac{1}{2}}]^2 [\widetilde{E}_v^{\frac{1}{2}}]^2 M_{i,j}^2} \leqslant \sqrt{\rho}\|M\|_\infty \sqrt{\sum_{u,v} [\widetilde{D}_u^{\frac{1}{2}}]^2 [\widetilde{E}_v^{\frac{1}{2}}]^2} \leqslant \sqrt{\rho}\Gamma$. 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.

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

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

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 $\breve{\mathcal{F}}_r$ corresponding to the empirically smoothed regulariser.

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

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

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

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

$$l(\breve{Z}) \leqslant \inf_{Z \in \widetilde{\mathcal{F}}_r} l(Z) + 2C\left[\ell\sqrt{r}\gamma(\mathbf{x}+\mathbf{y})^2 + b\right]\sqrt{\frac{\gamma^2 d\log(d)}{N}}. \tag{16}$$

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).

3.6 Variations on the optimization problems

As in the related literature ([19, 22] etc.), we worked with a *bounded loss*, and expressed our results for the *loss minimizer* within a function class defined by *explicit norm constraints*. However, it is also possible to modify the results (under some boundedness assumptions) to make them apply to lagrangian formulations such as (18) (1), (2). In typical contexts where the entries are known to be bounded, this can even be done with the square loss. As an example, we consider the following immediate corollary of Proposition 3.3 and its global version C.1 (appendix):

Corollary 3.4. *Assume that all of the entries of the ground truth are bounded by a constant C, and that they are observed without noise. Let $Z_\# = XM_\#Y^\top$ be the solution to the following optimization problem:*

$$\min_M \quad \|\widetilde{D}^{\frac{1}{2}}PMQ^{-1}\widetilde{E}^{\frac{1}{2}}\|_* \qquad \textit{subject to} \qquad [XMY^\top]_\xi = G_\xi \quad \forall \xi \in \Omega. \tag{17}$$

Let $\Phi_C(x) = \operatorname{sign}(x)\min(|x|, C)$. For any ℓ-Lipschitz loss l, we have (with probability $\geqslant 1-\delta$)

$$l(\Phi_C(Z_\#)) \leqslant \frac{8\ell\sqrt{\Gamma}\sqrt{dr_G}(1+\sqrt{\log(2d)})}{\sqrt{N}} + \frac{12\ell\mathbf{x}\mathbf{y}\sqrt{d_1d_2r_G}(1+\log(2d))}{N} + 2C\ell\sqrt{\frac{\log(2/\delta)}{2N}}.$$

where r_G is the smallest r such that the ground truth G satisfies $G \in \widetilde{\mathcal{F}}_r$.

A further result which applies in the presence of noise is provided in Appendix H.

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 column spaces. The sampling distribution is a power-type law depending on Λ 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 α) (ATR)[6]; and finally (3) our smoothed empirically adjusted regulariser $\lambda\|\breve{M}\|$ (for several values of α) (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.

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

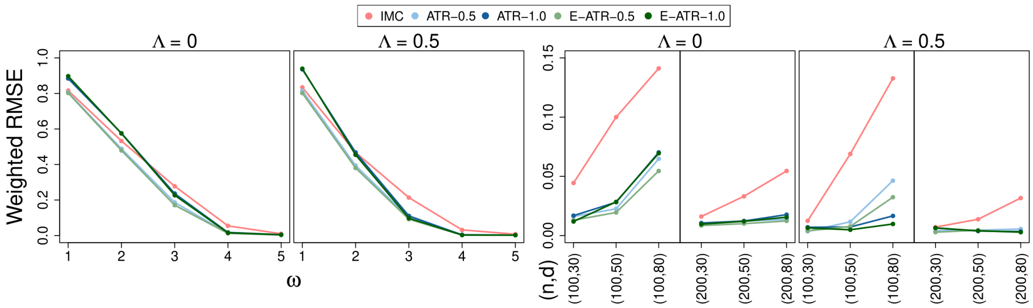

Figure 1: Left: performance as a function of the data sparsity parameter ω for $n, d, r = 200, 80, 10$. Right: Performance on different n, d, r combinations for $\omega = 4$. Legend: parameter to the right is α.

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]:

$$\min_{M,Z} \frac{1}{N} \sum_{(i,j)\in\Omega} l(XMY^\top + Z, G_{i,j} + \zeta_{i,j}) + \lambda_1 \|\breve{D}^{\frac{1}{2}} \widehat{P} M \widehat{Q}^{-1} \breve{E}^{\frac{1}{2}}\|_* + \lambda_2 \|\breve{D}_I^{\frac{1}{2}} Z \breve{E}_I^{\frac{1}{2}}\|_* \tag{18}$$

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 further 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.

Broader impact

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

Acknowledgements

The authors acknowledge support by the German Research Foundation (DFG) awards KL 2698/2-1 and KL 2698/5-1, as well as by the Federal Ministry of Science and Education (BMBF) awards 031L0023A, 01IS18051A, and 031B0770E.

References

[1] Zhao Kang, Chong Peng, and Qiang Cheng. Top-n recommender system via matrix completion. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 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 R Lyu, and Irwin King. Recommender systems with social regularization. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 287–296, 2011.

[6] Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 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. *Advances in Neural Information Processing Systems*, 33, 2020.

[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. *Journal of Machine Learning Ressearch*, 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 Transactions on Information Theory*, 56(5):2053–2080, May 2010.

[28] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 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] Yudong Chen. Incoherence-optimal matrix completion. *Information Theory, IEEE Transactions on*, 61, 10 2013.

[32] Lijun Ding and Yudong Chen. Leave-one-out approach for matrix completion: Primal and dual analysis. *IEEE Transactions on Information Theory*, 66:7274–7301, 2020.

[33] 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.

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

[35] 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.

[36] Antoine Ledent, Rodrigo Alves, and Marius Kloft. Orthogonal inductive matrix completion. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2021.

[37] Rodrigo Alves, Antoine Ledent, Renato Assunção, and Marius Kloft. An empirical study of the discreteness prior in low-rank matrix completion. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, volume 148 of *Proceedings of Machine Learning Research*, pages 111–125. PMLR, 11 Dec 2021.

[38] Qiaosheng Zhang, Vincent Tan, and Changho Suh. Community detection and matrix completion with social and item similarity graphs. *IEEE Transactions on Signal Processing*, 69:917–931, 01 2021.

[39] Qiaosheng Eric Zhang, Geewon Suh, Changho Suh, and Vincent Y. F. Tan. MC2G: an efficient algorithm for matrix completion with social and item similarity graphs. *CoRR*, abs/2006.04373, 2020.

[40] Rina Foygel, Nathan Srebro, and Russ R Salakhutdinov. Matrix reconstruction with the local max norm. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[41] Muhan Zhang and Yixin Chen. Inductive matrix completion based on graph neural networks. In *International Conference on Learning Representations*, 2020.

[42] Qitian Wu, Hengrui Zhang, Xiaofeng Gao, Junchi Yan, and Hongyuan Zha. Towards open-world recommendation: An inductive model-based collaborative filtering approach. In *International Conference on Machine Learning*, pages 11329–11339. PMLR, 2021.

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

Supplementary Material for “Fine-grained Generalisation Analysis of Inductive Matrix Completion”

Antoine Ledent
TU Kaiserslautern
ledent@cs.uni-kl.de

Rodrigo Alves
TU Kaiserslautern
alves@cs.uni-kl.de

Yunwen Lei
University of Birmingham
y.lei@bham.ac.uk

Marius Kloft
TU Kaiserslautern
kloft@cs.uni-kl.de

A Proof of Theorem 3.1

Preliminary discussion:
The following lemma from [1] was used in the proof in the non inductive case [2].

Lemma A.1 (Latała, 2005). *Let X be a random matrix with independent, zero mean entries, we have*

$$\mathbb{E}(\|X\|) \leqslant C_\ell \left(\max_i \sqrt{\sum_j \mathbb{E}(X_{i,j}^2)} + \max_j \sqrt{\sum_i \mathbb{E}(X_{i,j}^2)} + \sqrt[4]{\sum_{i,j} \mathbb{E}(X_{i,j}^4)} \right),$$

where C_ℓ is a universal constant.

The proof of the result in [2] relies on this Lemma, which applies to 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 (as the function is constant there), whilst on group 2, the entries are well spread out and Lemma A.1 limits the spectral norm similarly to the uniform case. The idea of the proof is to carefully tune those two contributions by adjusting the threshold involved in the split.

In our inductive situation, directly using a similar splitting strategy can only yield bounds with non-logarithmic dependence on n, or bounds of the type of equation (3) (which are well known and vacuous when the side information is of comparable size to the matrix). To understand the problem intuitively, it is helpful to think of the case of ’community side information’, where users and items can be divided into equally-sized groups (’communities’) by partition functions $c_U : \{1, 2, \ldots, m\} \to \{1, 2, \ldots, d_1\}$ and $c_I : \{1, 2, \ldots, n\} \to \{1, 2, \ldots, d_2\}$ respectively, with the rating of (i, j) depending only on the groups $c_U(i)$ and $c_I(j)$ to which i and j belong respectively. If the side information consists in indicator functions of the communities, simply applying known results for standard matrix completion yields distribution-free bounds of order $O(d^{3/2}\sqrt{r})$ (in this case $d = \max(d_1, d_2)$ will be equal to the max number of communities), whilst applying existing IMC results only yields bounds of order rd^2.

Comparing the proof techniques in the MC and "IMC with communities" cases with this example in mind, it becomes clear that the split should no longer be into frequently sampled entries and less frequently sampled entries, but into frequently sampled communities and less frequently sampled communities. To generalise this to arbitrary X, Y, we must define a concept of "frequently sampled" combinations $(X_{\bullet,u}, Y_{\bullet,v})$ of columns of the side information matrices. In practice this corresponds to a split between entries of $X^\top R_N Y$ (where $[R_N]_{i,j}$ contains the sum of the Rademacher variables corresponding to entry $i, j)$ by high or low variance: we use the rotational invariance of the trace operator and equivalently express the Rademacher averages in inductive space. However, the entries of the resulting matrix are certainly not independent, which makes it impossible to apply the

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

concentration results from [1]. Instead, we must rely again on the matrix Bernstein inequality F.4 Obtaining a covariance structure that is amenable to application of this result requires performing an iterative procedure involving series of *distribution dependent* rotational transformations of the side information and other estimates at each step.

Proof of Theorem 3.1. The theorem follows immediately from the classic result (Theorem F.1) as well as its variation F.2 together with the Rademacher complexity bound below (Theorem A.1). $\square$

Recall that for any $x_1, \ldots, x_N$ and any function class $\mathcal{F}$ we can define the (data dependent) Rademacher complexity $\widehat{\mathfrak{R}}_{(x_1,\ldots,x_n)}(\mathcal{F})$ as

$$\widehat{\mathfrak{R}}_{(x_1,\ldots,x_n)}(\mathcal{F}) := \mathbb{E}_\sigma \sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i f(x_i), \tag{A.1}$$

where the σ_i's are i.i.d. Rademacher random variables (i.e. $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 0.5$).

Theorem A.1. *Let $X \in \mathbb{R}^{m\times d_1}$ and $Y \in \mathbb{R}^{n\times d_2}$ be side information matrices. Consider the function class*

$$\mathcal{F}_{\mathcal{M}} := \left\{ XMY^\top \middle| \|M\|_* \leqslant \mathcal{M} \right\}$$

We have the following bound on the expected Rademacher complexity of $l \circ \mathcal{F}_{\mathcal{M}}$:

$$\mathbb{E}_{x_1,\ldots,x_N} \mathfrak{R}\left(l \circ \mathcal{F}_{\mathcal{M}}\right) \leqslant b\sqrt{\frac{2\pi}{N}} + \frac{16\mathbf{xy}\ell\mathcal{M} + \ell}{N}\log(2d) + \sqrt{\frac{10\ell b\mathbf{xy}\mathcal{M}\sqrt{d}}{N}}\Psi, \tag{A.2}$$

where $\Psi = \left[\sqrt{\log(2d)} + \sqrt{\log(N(20\mathcal{M}^2\ell\sqrt{d}[\mathbf{x}^2\mathbf{y}^2]/b + 1)}\right]$ *is a logarithmic quantity.*

In other words,

$$\mathbb{E}\,\mathfrak{R}\left(l \circ \mathcal{F}_{\mathcal{M}}\right) = \widetilde{O}\left(\sqrt{\frac{\ell b\mathbf{xy}\mathcal{M}\sqrt{d}}{N}} + \frac{\ell\mathbf{xy}\mathcal{M}}{N} + \frac{b}{\sqrt{N}}\right) \tag{A.3}$$

Before we proceed with the proof, we need to establish a few lemmas.

Lemma A.2 (Variation on Lemma 8 in [3]). *Let $r \in \mathbb{N}$ and suppose we are given r fixed matrices $E_1, E_2, \ldots, E_r \in \mathbb{R}^{m\times n}$ with the property that $|E_u|_{i,j} \leqslant B$ for all u, i, j. Now consider the following function class for a constant $C \in \mathbb{R}^+$:*

$$\mathcal{F}_C := \left\{ \sum_{u=1}^{r} \lambda_u E_u : |\lambda_u| \leqslant C \quad \forall u \leqslant r \right\}. \tag{A.4}$$

For any $\epsilon > 0$ there exists a cover $\mathcal{C}_\epsilon \subset \mathcal{F}$ with the following two properties:

1. *For any $Z \in \mathcal{F}$ there exists a $\widetilde{Z} \in \mathcal{C}_\epsilon$ such that for all $(i,j) \in [m] \times [n]$ we have $|Z_{i,j} - \widetilde{Z}_{i,j}| \leqslant \epsilon$*
2. $$|\mathcal{C}_\epsilon| \leqslant \left[\frac{2CBr}{\epsilon} + 1\right]^r \tag{A.5}$$

Proof. We consider the following discretised version of $\mathcal{F}$ for an ϵ' which will be determined later:

$$\mathcal{D}_{\epsilon'} := \left\{ \sum_{u=1}^{r} p_u\epsilon' E_u \quad \middle| (\forall u) \quad p_u \in \mathbb{Z} \quad \wedge \quad |p_u\epsilon'| \leqslant C \right\} \tag{A.6}$$

Let $Z \in \mathcal{F}$. We can write $Z = \sum_{u=1}^{r} \lambda_u E_u$ for some λ_us. Let $\widetilde{Z} = \sum_{u=1}^{r} \operatorname{sign}(\lambda_u)\left\lfloor \frac{|\lambda_u|}{\epsilon'} \right\rfloor \epsilon' E_u$.

Note that $\widetilde{Z} \in \mathcal{D}_{\epsilon'}$. Furthermore, for any i, j we have

$$\left|Z_{i,j} - \widetilde{Z}_{i,j}\right| = \left|\sum_{u=1}^{r} [E_u]_{i,j} \left[\lambda_u - \operatorname{sign}(\lambda_u) \left\lfloor \frac{|\lambda_u|}{\epsilon'} \right\rfloor \epsilon'\right]\right| \tag{A.7}$$

$$\leqslant B \sum_{u=1}^{r} \epsilon' = Br\epsilon'. \tag{A.8}$$

Thus, setting $\epsilon' = \frac{\epsilon}{Br}$, we obtain that $\mathcal{C}_\epsilon := \mathcal{D}_{\epsilon'}$ is indeed a uniform ϵ-cover of $\mathcal{F}$ w.r.t. to the L^∞ norm (over the whole sample space $[m] \times [n]$).

Finally, it is trivial to calculate that

$$|\mathcal{C}_{\epsilon'}| = \left[\frac{2C}{\epsilon'} + 1\right]^r = \left[\frac{2CBr}{\epsilon} + 1\right]^r, \tag{A.9}$$

as expected.

□

The following useful result is an immediate consequence of the McDiarmid inequality. A similar result was presented in [4] (cf. Theorem 11 page 469) for the expected Rademacher complexity.

Lemma A.3. *For any fixed $x_1, \ldots, x_N$ and any function class $\mathcal{F}$ mapping to $[-1, 1]$ we have with probability $\geqslant 1 - \delta$ over the draw of the Rademacher variables $\sigma_1, \ldots, \sigma_N$,*

$$\left|\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i f(x_i) - \widehat{\mathfrak{R}}_{(x_1,\ldots,x_n)}(\mathcal{F})\right| \leqslant \sqrt{\frac{2\log(2/\delta)}{N}}. \tag{A.10}$$

We now present the following result, of great importance to the proof of Theorem A.1, and which may be of independent interest. It may be viewed as a modification of Dudley's Entropy theorem F.6 entertwined with Talagrand's concentration Lemma.

Proposition A.4. *Let $\mathcal{F}_1, \mathcal{F}_2 : \mathcal{X} \to \mathbb{R}$ be two function classes, let $l : \mathbb{R}^2 \to [-1, 1]$ be a bounded loss function with Lipschitz constant ℓ.*

Assume that the function class $\mathcal{F}_1$ has the property for all ϵ, it has a uniform cover of size $\mathcal{N}(\mathcal{F}_1, \epsilon)$, where $\mathcal{N}(\mathcal{F}_1, \epsilon)$ is some function of ϵ. That is to say, there is a cover $\mathcal{C}(\epsilon)$ of size $\mathcal{N}(\mathcal{F}_1, \epsilon)$ such that for all $f_1 \in \mathcal{F}_1$ there exists $\tilde{f}_1 \in \mathcal{C}(\epsilon)$ such that for all $x \in \mathcal{X}$ we have

$$\left|f_1(x) - \tilde{f}_1(x)\right| \leqslant \epsilon. \tag{A.11}$$

Define the function class $\mathcal{F} = \{f_1 + f_2 | f_1 \in \mathcal{F}, f_2 \in \mathcal{F}_2\}$.

For all ϵ and for any training set $x_1, \ldots, x_N$, we have the following bound on the (expected) Rademacher complexity of the function class $l \circ \mathcal{F}$:

$$\widehat{\mathfrak{R}}(l \circ \mathcal{F}) \leqslant \ell\epsilon + 2\ell\widehat{\mathfrak{R}}(\mathcal{F}_2) + \sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_1, \epsilon))}{N}} + \sqrt{\frac{2\pi}{N}}. \tag{A.12}$$

In particular, the above also holds for the expected Rademacher complexity after taking expectations.

Remark: The requirement on the cover $\mathcal{C}(\epsilon)$ is quite strong: we require that one fixed cover be an ϵ-cover w.r.t. the l^∞ norm for *any* training set. However, this condition can be satisfied when the function class considered is parametric and globally Lipschitz, as is the case in our application of the result to the proof of Theorem A.1.

Proof. Fix an $\epsilon > 0$ and let $\mathcal{C}(\epsilon)$ be a uniform ϵ cover of $\mathcal{F}_1$. By the Lipschitz property we have for any $\sigma = (\sigma_1, \sigma_2 \ldots, \sigma_N)$:

$$\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i l(f(x_i), y_i) \tag{A.13}$$

$$= \sup_{f_1 \in \mathcal{F}} \sup_{f_2 \in \mathcal{F}_2} \frac{1}{N} \sum_{i=1}^{N} \sigma_i l(f(x_i), y_i) \tag{A.14}$$

$$= \sup_{f_1 \in \mathcal{F}} \sup_{f_2 \in \mathcal{F}_2} \frac{1}{N} \sum_{i=1}^{N} \sigma_i l(\tilde{f}_1(x_i) + [f_1 - \tilde{f}_1](x_i) + f_2(x_i), y_i) \tag{A.15}$$

$$\leqslant \sup_{f_1 \in \mathcal{F}} \sup_{f_2 \in \mathcal{F}_2} \frac{1}{N} \sum_{i=1}^{N} \sigma_i l(\tilde{f}_1(x_i) + f_2(x_i), y_i) \tag{A.16}$$

$$+ \sup_{f_1 \in \mathcal{F}} \sup_{f_2 \in \mathcal{F}_2} \frac{1}{N} \sum_{i=1}^{N} \sigma_i \left[l([f_1 - \tilde{f}_1](x_i) + \tilde{f}_1(x_i) + f_2(x_i), y_i) - l(\tilde{f}_1(x_i) + f_2(x_i), y_i) \right] \tag{A.17}$$

$$\leqslant \sup_{f_1 \in \mathcal{F}} \sup_{f_2 \in \mathcal{F}_2} \frac{1}{N} \sum_{i=1}^{N} \sigma_i l(\tilde{f}_1(x_i) + f_2(x_i), y_i) + \ell\epsilon, \tag{A.18}$$

where at the last line we have used the fact that l is ℓ-Lipschitz and that $\mathcal{C}(\epsilon)$ is an L^∞ cover for any dataset, including $x_1, x_2, \ldots, x_N$.

Now, observe that for all $\tilde{f}_1 \in \mathcal{C}(\epsilon)$ and for our fixed training set $x_1, \ldots, x_N$ we can apply Lemma A.3 to the function class

$$l_{\tilde{f}_1} \circ \mathcal{F}_2 := \left\{ [l(\tilde{f}_1(x_i) + f_2(x_i)), y_i]_{i=1}^N \quad | f_2 \in \mathcal{F}_2 \right\}.$$

Thus, for any $\delta > 0$, we have w.p. $\geqslant 1 - \delta$ over the draw of the Rademacher variables,

$$\left| \sup_{f_2 \in \mathcal{F}_2} \frac{1}{N} \sum_{i=1}^{N} \sigma_i l(\tilde{f}_1(x_i) + f_2(x_i), y_i) - \mathbb{E}_\sigma \sup_{f_2 \in \mathcal{F}_2} \frac{1}{N} \sum_{i=1}^{N} \sigma_i l(\tilde{f}_1(x_i) + f_2(x_i), y_i) \right| \tag{A.19}$$

$$= \left| \sup_{f_2 \in \mathcal{F}_2} \frac{1}{N} \sum_{i=1}^{N} \sigma_i l(\tilde{f}_1(x_i) + f_2(x_i), y_i) - \widehat{\mathfrak{R}}(l_{\tilde{f}_1} \circ \mathcal{F}_2) \right| \tag{A.20}$$

$$\leqslant \sqrt{\frac{2\log(2/\delta)}{N}} \tag{A.21}$$

where at the second line we have simply defined $\widehat{\mathfrak{R}}(l_{\tilde{f}_1} \circ \mathcal{F}_2) := \mathbb{E}_\sigma \sup_{f_2 \in \mathcal{F}_2} \frac{1}{N} \sum_{i=1}^N \sigma_i l(\tilde{f}_1(x_i) + f_2(x_i), y_i)$.

Now, composing inequality (A.21) with a union bound over all possible choices of $\tilde{f}_1 \in \mathcal{C}(\epsilon)$ we have that for all $\delta > 0$, w.p.$\geqslant 1 - \delta$, every $\tilde{f}_1 \in \mathcal{C}(\epsilon)$ satisfies

$$\left| \sup_{f_2 \in \mathcal{F}_2} \frac{1}{N} \sum_{i=1}^{N} \sigma_i l(\tilde{f}_1(x_i) + f_2(x_i), y_i) - \widehat{\mathfrak{R}}(l_{\tilde{f}_1} \circ \mathcal{F}_2) \right| \tag{A.22}$$

$$\leqslant \sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_1, \epsilon)) + 2\log(2/\delta)}{N}} \tag{A.23}$$

$$\leqslant \sqrt{\frac{2\log(2/\delta)}{N}} + \sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_1, \epsilon))}{N}}. \tag{A.24}$$

Now, note that for any choice of $\tilde{f}_1$, we can apply the Talagrand contraction Lemma(cf. [4] (Theorem 12 page 469), [5] (corollary 3.17) , [6](Lemma 8 page 1 of supplementary)) to the function class $\mathfrak{R}(l_{\tilde{f}_1} \circ \mathcal{F}_2)$ to obtain (for any $\tilde{f}_1$):

$$\widehat{\mathfrak{R}}(l_{\tilde{f}_1} \circ \mathcal{F}_2) \leqslant 2\ell\widehat{\mathfrak{R}}(\mathcal{F}_2). \tag{A.25}$$

Plugging Equations (A.24) and A.25 back into equation (A.18), we have that w.p. $\geqslant 1 - \delta$,

$$\sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \sigma_i l(f(x_i), y_i) \leqslant \ell\epsilon + 2\ell\widehat{\mathfrak{R}}(\mathcal{F}_2) + \sqrt{\frac{2\log(2/\delta)}{N}} + \sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_1, \epsilon))}{N}}. \tag{A.26}$$

The only thing left to do is a simple integration over δ: let X denote the random variable

$$X := \sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i l(f(x_i), y_i) - \ell\epsilon - 2\ell\widehat{\mathfrak{R}}(\mathcal{F}_2) - \sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_1,\epsilon))}{N}}. \tag{A.27}$$

By equation (A.26) we have for all $\varepsilon > 0$

$$\mathbb{P}(X \geqslant \varepsilon) \leqslant 2\exp\left(-\frac{\varepsilon^2 N}{2}\right). \tag{A.28}$$

Integrating over ε we obtain

$$\mathbb{E}(X) \leqslant \int_0^\infty 2\exp\left(-\frac{\varepsilon^2 N}{2}\right) d\varepsilon \tag{A.29}$$

$$= \frac{2\sqrt{2}}{\sqrt{N}}\int_0^\infty \exp(-\theta^2)d\theta = \sqrt{\frac{2\pi}{N}}. \tag{A.30}$$

Plugging this equation (A.29) back into the definition of X (eq. (A.27)) we obtain:

$$\widehat{\mathfrak{R}}(l\circ\mathcal{F}) \leqslant \ell\epsilon + 2\ell\widehat{\mathfrak{R}}(\mathcal{F}_2) + \sqrt{\frac{\log(\mathcal{N}(\mathcal{F}_1,\epsilon))}{N}} + \sqrt{\frac{2\pi}{N}}, \tag{A.31}$$

as expected.

□

Proof of Theorem A.1 Let $\xi_1,\ldots,\xi_n$ be sampled i.i.d from the sampling distribution $\mathcal{D}$ on $\{1,2,\ldots,m\}\times\{1,2,\ldots,n\}$. Let $s_1, s_2,\ldots,s_N$ be iid Rademacher random variables. For any element of $\xi \in \{1,2,\ldots,m\}\times\{1,2,\ldots,n\}$ we also write e_ξ for the matrix with all entries equal to 0 except the entry corresponding to ξ, which is set to 1.

Define the Rademacher matrix $R_N := \sum_{o=1}^{N} e_{\xi_o}s_o$. Define also $U = X^\top R_N Y$. This is a random variable.

We begin with the following easy observations:

$$\begin{aligned}\operatorname{Tr}\left(\mathbb{E}(UU^\top)\right) &= \operatorname{Tr}\left(\mathbb{E}(U^\top U)\right) = \sum_{u=1}^{d_1}\sum_{v=1}^{d_2}\mathbb{E}(U_{u,v}^2)\\ &= \sum_{u,v}\sum_{i,j} p_{i,j}(X_{i,u})^2(Y_{jv})^2 = N\sum_{i,j}p_{i,j}\|x_i\|^2\|y_j\|^2\\ &= N\Gamma.\end{aligned} \tag{A.32}$$

Note also that for any M, $\langle XMY^\top, R_N\rangle = \langle M, U\rangle$.

We will now need to iteratively define a sequence of matrices $U^k, \bar{U}^k, \bar{M}^k, \bar{V}^k \in \mathbb{R}^{d_1\times d_2}$ and $\mathcal{T}_k$ for $k = 0,1,\ldots,K$ for some stopping time K. The whole construction depends on a real parameter $p > 0$ which will be chosen later. It is important to note that although the construction of $U^k, \bar{U}^k, \bar{M}^k, \bar{V}^k$ also depends on the sampling distribution $\mathcal{D}$, it is a deterministic construction and does not depend on the data (the same is true of $\bar{M}^k$ for a given core matrix M).

$\mathcal{T}_k$ is a sequence of reals defined by $\mathcal{T}_k = \mathbb{E}(\|\bar{U}^k\|_{\mathrm{Fr}}^2)$

First, we set $\bar{U}^0 = \bar{V}^0 = U, \bar{M}^0 = M$ (and $\mathcal{T}_0 = N\Gamma$).

Assuming that $\bar{U}^k$ and $\bar{M}^k$ have been defined already, we define the next iteration as follows.

We first obtain an orthogonal matrix $A^k \in \mathbb{R}^{d_1\times d_1}$ (resp. $B^k \in \mathbb{R}^{d_2\times d_2}$) which diagonalises $\mathbb{E}(\bar{U}^k(\bar{U}^k)^\top)$ (resp. $\mathbb{E}((\bar{U}^k)^\top\bar{U}^k)$) so that $\mathbb{E}(\bar{U}^k(\bar{U}^k)^\top) = (A^k)^{-1}D_1A^k$ and $\mathbb{E}((U^k)^\top(U^k)) = (B^k)^{-1}D_2B^k$ for some diagonal matrices D_1, D_2.

Now, we define

$$\bar{V}^{k+1} = A^k\bar{U}^kB^k \tag{A.33}$$

$$\bar{M}^{k+1} = A^k \bar{M}^k B^k. \tag{A.34}$$

Now, by construction of the matrices A^{k+1} and B^{k+1}, the matrices $\mathbb{E}((\bar{V}^{k+1})[\bar{V}^{(k+1)}]^\top)$ and $\mathbb{E}([\bar{V}^{(k+1)}]^\top \bar{V}^{k+1})$ are both diagonal. We now split according to two cases: Case 1:

$$\mathbb{V}\mathrm{ar}(\bar{V}^{k+1}_{u,v}) \leqslant p \qquad \forall u, v \tag{A.35}$$

Case 2: equation (A.35) does not hold, i.e. there exists $u_{k+1}, v_{k+1} \in \mathbb{R}^{d_1 \times d_2}$ with $\mathbb{V}\mathrm{ar}(\bar{V}^{k+1}_{u_{k+1},v_{k+1}}) > p$.

In case 1, we end the procedure and set $K = k$. In case 2, we set

$$\bar{U}^{k+1} = \bar{V}^{k+1} - e_{u_{k+1},v_{k+1}} \bar{V}^{k+1}_{u_{k+1},v_{k+1}} \tag{A.36}$$

(i.e. $\bar{U}^{k+1}$ is identical to $\bar{V}^{k+1}$ on all entries except (u_{k+1}, v_{k+1}) where it is set to zero.)

The procedure repeats until case 1 occurs. Note that since the only operations on $\bar{M}$ are from equation (A.34) we have that $\bar{M}^k = A^{k-1}A^{k-1}\ldots, A^0 M B^0 B^1 \ldots B^{k-1} = \bar{A}^{k-1} M \bar{B}^{k-1}$ where $\bar{A}^{k-1}$ (resp. $\bar{B}^{k-1}$) denotes the orthogonal matrix $A^{k-1}A^{k-1}\ldots A^0$ (resp. $B^0 B^1 \ldots B^{k-1}$). Finally, we define

$$U^k = \prod_{i=0}^{k-1} [A^i]^{-1} \bar{U}^k \prod_{i=k-1}^{0} [B^i]^{-1} = [\bar{A}^{k-1}]^{-1} \bar{U}^k [\bar{B}^{k-1}]^{-1}. \tag{A.37}$$

Now, observe that by the rotational invariance of the Frobenius norm and the nuclear norms:

$$\|\bar{M}^k\|_* = \|M\|_* \tag{A.38}$$
$$\mathbb{E}(\|\bar{V}^{k+1}\|^2_{\mathrm{Fr}}) = \mathbb{E}(\|\bar{U}^k\|^2_{\mathrm{Fr}}) = \mathbb{E}(\|U^k\|^2_{\mathrm{Fr}}) = \mathcal{T}_k \tag{A.39}$$

and therefore for all $k \leqslant K - 1$:

$$\mathcal{T}_{k+1} = \mathbb{E}(\|U^{k+1}\|^2_{\mathrm{Fr}}) = \mathcal{T}_k - \mathbb{V}\mathrm{ar}(V^{k+1}_{u_{k+1},v_{k+1}}) \leqslant \mathcal{T}_k - p. \tag{A.40}$$

In particular, since $\mathcal{T}_0 = \mathbb{E}(\|U\|^2_{\mathrm{Fr}}) = \Gamma N$ is finite, the procedure must finish in finite time K with

$$K \leqslant \frac{\Gamma N}{p}. \tag{A.41}$$

Now, U^k is of course only the reexpression of $\bar{U}^k$ in the original orthogonal basis: in particular by the rotational invariance of the Frobenius inner product we have

$$\langle M, U^k \rangle = \langle \bar{M}^k, \bar{U}^k \rangle.$$

Further, we can express the recurrence relations (A.36) and (A.33) directly in this original orthogonal basis in terms of transformations on the U^ks:

$$
\begin{aligned}
U^{k+1} &= [\bar{A}^k]^{-1} \bar{U}^{k+1} [\bar{B}^k]^{-1} && \text{(A.42)}\\
&= [\bar{A}^k]^{-1} \left[\bar{V}^{k+1} - e_{u_{k+1},v_{k+1}} \bar{V}^{k+1}_{u_{k+1},v_{k+1}}\right] [\bar{B}^k]^{-1} && \text{(A.43)}\\
&= [\bar{A}^k]^{-1} \left[A^k \bar{U}^k B^k - e_{u_{k+1},v_{k+1}} \langle A^k \bar{U}^k B^k, e_{u_{k+1},v_{k+1}} \rangle\right] [\bar{B}^k]^{-1} && \text{(A.44)}\\
&= [\bar{A}^k]^{-1} \left[A^k \bar{A}^{k-1} U^k \bar{B}^{k-1} B^k - e_{u_{k+1},v_{k+1}} \langle A^k \bar{U}^k B^k, e_{u_{k+1},v_{k+1}} \rangle\right] [\bar{B}^k]^{-1} && \text{(A.45)}\\
&= [\bar{A}^k]^{-1} \left[\bar{A}^k U^k \bar{B}^k - e_{u_{k+1},v_{k+1}} \langle \bar{A}^k U^k \bar{B}^k, e_{u_{k+1},v_{k+1}} \rangle\right] [\bar{B}^k]^{-1} && \text{(A.46)}\\
&= U^k - \langle \bar{A}^k U^k \bar{B}^k, e_{u_{k+1},v_{k+1}} \rangle [\bar{A}^k]^{-1} e_{u_{k+1},v_{k+1}} [\bar{B}^k]^{-1} && \text{(A.47)}\\
&= U^k - \langle U^k, [\bar{A}^k]^{-1} e_{u_{k+1},v_{k+1}} [\bar{B}^k]^{-1} \rangle [\bar{A}^k]^{-1} e_{u_{k+1},v_{k+1}} [\bar{B}^k]^{-1} && \text{(A.48)}\\
&= U^k - \langle U^k, E_k \rangle E_k, && \text{(A.49)}
\end{aligned}
$$

where at the second line (A.43) we have used equation (A.36), at the third line (A.44) we have used equation (A.33), at the fourth line (A.45) we have used equation (A.37), at the fifth line (A.46) we have used equation (A.37) again as well as a simplification via the definitions of $\bar{A}^k$ and $\bar{B}^k$, at the seventh line (A.48) we have used properties of the Frobenius inner product, and at the eighth and last line (A.49) we have defined $E_k = [\bar{A}^k]^{-1} e_{u_{k+1},v_{k+1}} [\bar{B}^k]^{-1}$. Note again crucially that the E_ks are *deterministic* matrices.

Now, we write $\mathcal{P}$ for the (projection) operator $\mathcal{P}_k : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^{d_1 \times d_2} : W \mapsto \langle W, E_k \rangle$. Then equation (A.49) can be written

$$U^{k+1} = (I - \mathcal{P}_k) U^k, \tag{A.50}$$

where I denotes the identity operator from $\mathbb{R}^{d_1 \times d_2}$ to itself. Iterating, we obtain for all k

$$U^k = \prod_{i=0}^{k-1} (I - \mathcal{P}_i) U. \tag{A.51}$$

Note that both $\mathcal{P}_k$ and $(I - \mathcal{P}_k)$ are self-adjoint. Hence, we can write

$$\begin{aligned}
\left\langle M, U^k \right\rangle &= \left\langle M, \prod_{i=0}^{k-1} (I - \mathcal{P}_i) U \right\rangle && \text{(A.52)}\\
&= \left\langle \prod_{i=k-1}^{0} (I - \mathcal{P}_i) M, U \right\rangle && \text{(A.53)}\\
&= \left\langle M^k, U \right\rangle, && \text{(A.54)}
\end{aligned}$$

where at the last line we have defined $M^k = \prod_{i=k-1}^{0} (I - \mathcal{P}_i) M$.

Now, note that we can write

$$\begin{aligned}
M^k &= \prod_{i=k-1}^{0} (I - \mathcal{P}_i) M && \text{(A.55)}\\
&= M - \sum_{u=0}^{k-1} \mathcal{P}_u \prod_{i=k-1}^{u+1} (I - \mathcal{P}_i) M && \text{(A.56)}\\
&= M - \sum_{u=0}^{k-1} E_u \left\langle E_u, \prod_{i=k-1}^{u+1} (I - \mathcal{P}_i) M \right\rangle && \text{(A.57)}\\
&= M - \sum_{u=0}^{k-1} E_u \lambda_u^k(M), && \text{(A.58)}
\end{aligned}$$

where we have defined $\lambda_u^k(M) := \left\langle E_u, \prod_{i=k-1}^{u+1} (I - \mathcal{P}_i) M \right\rangle$. Note that $\|E_u\|_{\mathrm{Fr}} = \|E_u\| = 1$ and since each operator $(I - \mathcal{P}_i)$ is a projection and in particular a contraction with respect to the Frobenius norm we have that $\| \prod_{i=k-1}^{u+1} (I - \mathcal{P}_i) M \|_{\mathrm{Fr}} \leqslant \|M\|_{\mathrm{Fr}} \leqslant \|M\|_*$. Hence for any M with $\|M\|_* \leqslant \mathcal{M}$ we have for any $u < k \leqslant K$:

$$\left| \lambda_u^k(M) \right| \leqslant \mathcal{M}. \tag{A.59}$$

We note that by construction, the matrix $\bar{V}^{K+1} = A^k \bar{U}^k B^k = A^k [\bar{A}^{k-1}] U^k \bar{B}^{k-1} B^k$, has the property that $\mathbb{E}((\bar{V}^{K+1})[\bar{V}^{(K+1)}]^\top)$ and $\mathbb{E}([\bar{V}^{(K+1)}]^\top \bar{V}^{K+1})$ are both diagonal, and

$$\mathbb{V}\mathrm{ar}(\bar{V}_{u,v}^{k+1}) \leqslant p \qquad \forall u, v \tag{A.60}$$

Thus, we have

$$\begin{aligned}
\left\| [U^K][U^K]^\top \right\| &= \left\| \mathbb{E}((\bar{V}^{K+1})[\bar{V}^{(K+1)}]^\top) \right\| \leqslant p d_2 \leqslant p d, && \text{(A.61)}\\
\left\| [U^K]^\top [U^K] \right\| &= \left\| \mathbb{E}([\bar{V}^{(K+1)}]^\top \bar{V}^{K+1}) \right\| \leqslant p d_1 \leqslant p d. && \text{(A.62)}
\end{aligned}$$

We now have the tools to proceed with the proof of the equation (A.2).

We define the following function classes:

$$\mathcal{F}_1 := \left\{ \sum_{k=0}^{K-1} \lambda_k X E_k Y^\top \middle| |\lambda_k| \leqslant \mathcal{M} \right\} \tag{A.63}$$

$$\mathcal{F}_2 := \left\{ X \left[\prod_{i=K-1}^{0} (I - \mathcal{P}_i) M \right] Y^\top \middle| \|M\|_* \leqslant \mathcal{M} \right\}. \tag{A.64}$$

By the constructions above and in particular equation (A.59) we have $\mathcal{F} \subset \mathcal{F}_1 + \mathcal{F}_2$. Furthermore, also by the construction of U^k etc., we can bound the Rademacher complexity of $\mathcal{F}_2$:

$$\begin{aligned}
\mathbb{E}_{\xi_1,\ldots,\xi_N}\left(\mathfrak{R}(\mathcal{F}_2)\right) &= \mathbb{E} \sup_{\|M\|_* \leqslant \mathcal{M}} \left\langle X \left[\prod_{i=K-1}^{0} (I - \mathcal{P}_i) M \right] Y^\top, R_N \right\rangle && \text{(A.65)} \\
&= \mathbb{E} \sup_{\|M\|_* \leqslant \mathcal{M}} \left\langle \left[\prod_{i=K-1}^{0} (I - \mathcal{P}_i) M \right], X^\top R_N Y \right\rangle && \text{(A.66)} \\
&= \mathbb{E} \sup_{\|M\|_* \leqslant \mathcal{M}} \left\langle \left[\prod_{i=K-1}^{0} (I - \mathcal{P}_i) M \right], U \right\rangle && \text{(A.67)} \\
&= \mathbb{E} \sup_{\|M\|_* \leqslant \mathcal{M}} \left\langle M, U^K \right\rangle && \text{(A.68)} \\
&\leqslant \mathcal{M}\mathbb{E}\left(\|U^K\|\right) && \text{(A.69)}
\end{aligned}$$

where as usual $\|\cdot\|$ denotes the spectral norm.

Now, observe that

$$\begin{aligned}
U^K &= \prod_{i=0}^{K-1} (I - \mathcal{P}_i) U = \sum_{o=1}^{N} \prod_{i=0}^{K-1} (I - \mathcal{P}_i) X^\top e_{\xi^o} Y && \text{(A.70)} \\
&= \sum_{o=1}^{N} s_o \prod_{i=0}^{K-1} (I - \mathcal{P}_i) x_{\xi_1^o} y_{\xi_2^o}^\top, && \text{(A.71)}
\end{aligned}$$

which is a sum of i.i.d centred random matrices. Thus we can apply Proposition (F.4) to it. The value of "M" in that proposition is clearly bounded by $\mathbf{xy}$ (indeed, for all i, j, $\|x_i y_j^\top\|_{\mathrm{Fr}} = \|x_i y_j^\top\| \leqslant \mathbf{xy}$, the operator $\prod_{i=0}^{K-1}(I - \mathcal{P}_i)$ is a contraction with respect to the Frobenius norm, and the spectral norm is certainly bounded by the Frobenius norm). A bound on the value of "σ" from Proposition (F.4) follows from our iterative construction and in particular from equations (A.61) which ensure that "σ" is bounded by $\sqrt{pd}$:

$$\sum_{o=1}^{N} \rho_o^2 \leqslant \sqrt{pd}. \tag{A.72}$$

It follows by an application of Proposition (F.4) to equation (A.69) that

$$\begin{aligned}
N\mathbb{E}_{\xi_1,\ldots,\xi_N}\left(\mathfrak{R}(\mathcal{F}_2)\right) &\leqslant \mathcal{M}\mathbb{E}\left(\|U^K\|\right) && \text{(A.73)} \\
&\leqslant \sqrt{8/3}(1 + \sqrt{\log(2d)})\mathcal{M}\sqrt{pd} + \mathcal{M}\frac{8\mathbf{xy}}{3}(1 + \log(2d))). && \text{(A.74)}
\end{aligned}$$

On the other hand, a simple application of Lemma A.2 tells us that $\mathcal{F}_1$ admits a uniform L^∞ cover $\mathcal{C}_{1/N}$ (w.r.t. the whole sample space), of granularity $1/N$ with

$$\mathcal{N}_\infty(\mathcal{F}_1, 1/N) = \left|\mathcal{C}_{1/N}\right| \leqslant \left[2N\mathcal{M}\mathbf{xy}K + 1\right]^K \leqslant \left[N(2\mathcal{M}\mathbf{xy}K + 1)\right]^K, \tag{A.75}$$

since the maximum entry of E_u is bounded by $\mathbf{xy}$ for any u.

By Proposition A.4 (rescaled taking into account the bound b on the loss function) we have for any training set

$$\widehat{\mathfrak{R}}_N(l\circ\mathcal{F})\leqslant \ell\epsilon+2\ell\widehat{\mathfrak{R}}_N(\mathcal{F}_2)+b\sqrt{\frac{\log(\mathcal{F}_\infty(\mathcal{F}_1,1/N))}{N}}+b\sqrt{\frac{2\pi}{N}}. \tag{A.76}$$

Taking expectations with respect to the training set on both sides and then applying equation (A.75) and (A.74) we obtain:

$$\begin{aligned}
\mathbb{E}[\widehat{\mathfrak{R}}_N(l\circ\mathcal{F})] &\leqslant \frac{\ell}{N}+2\ell\mathbb{E}(\widehat{\mathfrak{R}}_N(\mathcal{F}_2))+b\sqrt{\frac{\log(\mathcal{F}_\infty(\mathcal{F}_1,1/N))}{N}}+b\sqrt{\frac{2\pi}{N}} && \text{(A.77)}\\
&\leqslant \frac{\ell}{N}+\frac{2\ell\mathcal{M}}{N}\left[\sqrt{8/3}(1+\sqrt{\log(2d)})\sqrt{pd}+\frac{8\mathbf{xy}}{3}(1+\log(2d))\right] && \text{(A.78)}\\
&\quad+b\sqrt{\frac{K\log(N(2\mathcal{M}\mathbf{xy}K+1))}{N}}+b\sqrt{\frac{2\pi}{N}} && \text{(A.79)}\\
&\leqslant b\sqrt{\frac{2\pi}{N}}+\frac{\ell}{N}+\frac{10\ell\mathcal{M}}{N}\sqrt{\log(2d)}\sqrt{pd}+\frac{16\mathbf{xy}\ell\mathcal{M}}{N}\log(2d) && \text{(A.80)}\\
&\quad+b\sqrt{\frac{\Gamma\log(N(2\mathcal{M}\mathbf{xy}\Gamma N/p+1))}{p}} && \text{(A.81)}\\
&\leqslant b\sqrt{\frac{2\pi}{N}}+\frac{\ell}{N}+\frac{10\ell\mathcal{M}}{N}\sqrt{\log(2d)}\sqrt{pd}+\frac{16\mathbf{xy}\ell\mathcal{M}}{N}\log(2d) && \text{(A.82)}\\
&\quad+b\sqrt{\frac{\mathbf{x}^2\mathbf{y}^2\log(N(2\mathcal{M}\mathbf{xy}[\mathbf{x}^2\mathbf{y}^2]N/p+1))}{p}}, && \text{(A.83)}
\end{aligned}$$

where at line (A.81) we have plugged in the bound for K from equation (A.41) and at line (A.83) we have used the fact that $\Gamma\leqslant\mathbf{x}^2\mathbf{y}^2$.

We can finally set the value of p, to balance the two contributions in equation (A.81) above: we set

$$p:=\frac{\mathbf{xy}Nb}{10\mathcal{M}\ell\sqrt{d}}, \tag{A.84}$$

which plugged into equation (A.83) gives

$$\begin{aligned}
&\mathbb{E}[\mathfrak{R}_N(l\circ\mathcal{F})] && \text{(A.85)}\\
&\leqslant b\sqrt{\frac{2\pi}{N}}+\frac{16\mathbf{xy}\ell\mathcal{M}+\ell}{N}\log(2d)+\frac{10\ell\mathcal{M}}{N}\sqrt{\log(2d)}\sqrt{pd}+ && \text{(A.86)}\\
&\quad b\sqrt{\frac{\mathbf{x}^2\mathbf{y}^2\log(N(2\mathcal{M}[\mathbf{x}^3\mathbf{y}^3]N/p+1))}{p}} && \text{(A.87)}\\
&\leqslant b\sqrt{\frac{2\pi}{N}}+\frac{16\mathbf{xy}\ell\mathcal{M}+\ell}{N}\log(2d)+ && \text{(A.88)}\\
&\quad\sqrt{\frac{10\ell b\mathbf{xy}\mathcal{M}\sqrt{d}}{N}}\left[\sqrt{\log(2d)}+\sqrt{\log(N(20\mathcal{M}^2\ell\sqrt{d}[\mathbf{x}^2\mathbf{y}^2]/b+1)}\right], && \text{(A.89)}
\end{aligned}$$

as expected.

□

B Proof of Propositions 3.1 and 3.2

Proposition 3.2 is included in the wordier version B.1 and proved below.

Proposition B.1. *W.p.$\geqslant 1-\delta$ for all M with $\|M\| \leqslant \mathcal{M}$:*

$$\mathbb{E}\left[l((XMY^\top)_\xi, G_\xi)\right] - \frac{1}{N}\sum_{\xi\in\Omega} l((XMY^\top)_\xi, G_\xi) \tag{B.1}$$
$$\leqslant \frac{4\ell}{\sqrt{N}}\mathcal{M}\max(\sigma_*^1,\sigma_*^2)(1+\sqrt{\log(2d)}) + \frac{6\ell}{N}\mathcal{M}\mathbf{xy}(1+\log(2d)) + b\sqrt{\frac{\log(2/\delta)}{2N}},$$

thus as long as $N \geqslant 9[\mathbf{xy}/\max(\sigma_*^1,\sigma_*^2)]^2(1+\log(2d))$, we have with probability $\geqslant 1-\delta$ over the draw of the training set S

$$\mathbb{E}\left[l((XMY^\top)_\xi, G_\xi)\right] - \frac{1}{N}\sum_{\xi\in\Omega} l((XM_SY^\top)_\xi, G_\xi) \tag{B.2}$$
$$\leqslant \frac{6\ell\mathcal{M}\max(\sigma_*^1,\sigma_*^2)(1+\sqrt{\log(2d)})}{\sqrt{N}} + b\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Proof of Proposition B.1. We will show the following bound on the Rademacher complexity of the function class $\mathcal{F}_\mathcal{M} := \{XMY^\top : \|M\| \leqslant \mathcal{M}\}$

$$\mathbb{E}(\mathfrak{R}) \leqslant \frac{1}{\sqrt{N}}\mathcal{M}\sqrt{\frac{8}{3}}\max(\sigma_*^1,\sigma_*^2)(1+\sqrt{\log(2d)}) + \frac{1}{N}\mathcal{M}\frac{8}{3}\mathbf{xy}(1+\log(2d)) \tag{B.3}$$

and for $N \geqslant 9[\mathbf{xy}/\max(\sigma_*^1,\sigma_*^2)]^2(1+\log(2d))$:

$$\mathbb{E}(\mathfrak{R}) \leqslant \frac{3\mathcal{M}\max(\sigma_*^1,\sigma_*^2)(1+\sqrt{\log(2d)})}{\sqrt{N}}, \tag{B.4}$$

The claims then follow from Theorem F.1, together with Talagrand's contraction Lemma.

Now, by the circular properties of the trace and the duality between the nuclear and spectral norms, writing F for the matrix with $F_{i,j} := \sum_{o=1}^N \sigma_o 1_{\xi^o=(i,j)}$,

$$\left[\frac{1}{N}\langle XMY^\top, F\rangle\right] = \frac{1}{N}\operatorname{Tr}((XMY^\top)^\top F) = \frac{1}{N}\operatorname{Tr}(YM^\top X^\top F) = \frac{1}{N}\operatorname{Tr}(X^\top FYM^\top)$$
$$= \frac{1}{N}\langle X^\top FY, M\rangle \leqslant \|M\|_*\|X^\top FY\|. \tag{B.5}$$

$$\mathfrak{R}(\mathcal{F}_\mathcal{M}) = \mathbb{E}\sup_{\|M\|_*\leqslant\mathcal{M}}\left[\frac{1}{N}\langle XMY^\top, F\rangle\right]$$
$$\leqslant \frac{\mathcal{M}}{N}\mathbb{E}(\|X^\top FY\|). \tag{B.6}$$

The term $\mathbb{E}(\|X^\top FY\|)$ can be written as $\sum_{o=1}^N \sigma_o x_{\xi_1^o} y_{\xi_2^o}^\top = \sum_{o=1}^N \sigma_o x_{i_o} y_{j_o}^\top$, thus, we can prove concentration inequalities for it using the non commutative Bernstein inequality (Proposition (F.4)).

We first note that for all i,j, $\|x_i y_j^\top\| \leqslant \mathbf{xy}$. Furthermore, we have $\mathbb{E}_{(i,j)\sim p}\left(\|[x_iy_j^\top][x_iy_j^\top]^\top\|\right) = \|\sum_{i,j} p_{i,j} x_i y_j^\top y_j x_i^\top\| = \|\sum_{i,j} p_{i,j} x_i x_i^\top \|y_j\|^2\| = \|\sum_i x_i x_i^\top q_i\| = \|\widetilde{L}\| = (\sigma_*^1)^2$, and similarly, $\mathbb{E}_{(i,j)\sim p}\left(\|[x_iy_j^\top]^\top[x_iy_j^\top]\|\right) = (\sigma_*^2)^2$.

Using this together with Proposition (F.4) we obtain

$$\mathbb{E}(\|X^\top FY\|) \leqslant \sqrt{N}\sqrt{\frac{8}{3}}\max(\sigma_*^1,\sigma_*^2)(1+\sqrt{\log(2d)}) + \frac{8}{3}\mathbf{xy}(1+\log(2d)). \tag{B.7}$$

Plugging this back into equation (B.6), we obtain

$$\mathbb{E}(\mathfrak{R}) \leqslant \frac{1}{\sqrt{N}}\mathcal{M}\sqrt{\frac{8}{3}}\max(\sigma_*^1,\sigma_*^2)(1+\sqrt{\log(2d)}) + \frac{1}{N}\mathcal{M}\frac{8}{3}\mathbf{xy}(1+\log(2d)) \tag{B.8}$$

(which yields (B.3)) and as long as $N \geqslant 9[\mathbf{xy}/\max(\sigma_*^1, \sigma_*^2)]^2(1+\log(2d))$,

$$
\begin{aligned}
\mathbb{E}(\mathfrak{R}) &\leqslant \frac{1}{\sqrt{N}}\mathcal{M}\sqrt{\frac{8}{3}}\max(\sigma_*^1, \sigma_*^2)(1+\sqrt{\log(2d)}) + \frac{1}{\sqrt{N}}\mathcal{M}\max(\sigma_*^1, \sigma_*^2)\sqrt{1+\log(2d)} \\
&\leqslant \frac{3\mathcal{M}\max(\sigma_*^1, \sigma_*^2)(1+\sqrt{\log(2d)})}{\sqrt{N}},
\end{aligned}
\tag{B.9}
$$

as expected. This establishes equation (B.4) and the claim follows from Talagrand's concentration lemma and the Rademacher Theorem F.1. □

Proposition 3.1 follows from the more general result below.

Proposition B.2. *Let us write $\mathcal{F}_{\mathcal{M}}$ for the function class corresponding to matrices of the form $XMY^\top$ with $\|M\|_* \leqslant \mathcal{M}$. Assume uniform sampling and write $\mathcal{K} := \max\left[\sqrt{d_1\frac{\|X^\top X\|}{m}\frac{\|Y\|_{\mathrm{Fr}}^2}{n}}, \sqrt{d_2\frac{\|Y^\top Y\|}{n}\frac{\|X\|_{\mathrm{Fr}}^2}{m}}\right]$.*

We have with probability $\geqslant 1-\delta$, for all $M \in \mathcal{F}_{\mathcal{M}}$:

$$
\begin{aligned}
&\mathbb{E}\left[l((XMY^\top)_\xi, G_\xi+\zeta_\xi)\right] - \frac{1}{N}\sum_{\xi\in\Omega} l((XMY^\top)_\xi, G_\xi+\zeta_\xi) \\
&\leqslant \frac{4\ell\mathcal{K}\sqrt{rd}(1+\sqrt{\log(2d)})}{\sqrt{N}} + \frac{6\ell}{N}\mathcal{M}\mathbf{xy}(1+\log(2d)) + b\sqrt{\frac{\log(2/\delta)}{2N}},
\end{aligned}
\tag{B.10}
$$

where $\sqrt{r} = (\mathcal{M}/\sqrt{d_1 d_2})$ and b is a bound on the loss.

Similarly, as long as

$$
N \geqslant 9\left[\frac{\sqrt{d}\mathbf{xy}}{\mathcal{K}}\right]^2(1+\log(2d))
\tag{B.11}
$$

we have with probability $\geqslant 1-\delta$ over the draw of the training set S, for all a $M \in \mathcal{F}_{\mathcal{M}}$:

$$
\begin{aligned}
&\mathbb{E}\left[l((XMY^\top)_\xi, G_\xi+\zeta_\xi)\right] - \frac{1}{N}\sum_{\xi\in\Omega} l((XMY^\top)_\xi, G_\xi+\zeta_\xi) \\
&\leqslant \frac{6\ell(\mathcal{M}/\sqrt{mn})\max(\sqrt{\|X^\top X\|\|Y\|_{\mathrm{Fr}}^2}, \sqrt{\|Y^\top Y\|\|X\|_{\mathrm{Fr}}^2})(1+\sqrt{\log(2d)})}{\sqrt{N}} + b\sqrt{\frac{\log(2/\delta)}{2N}} \\
&= \frac{6\ell\mathcal{K}\sqrt{rd}(1+\sqrt{\log(2d)})}{\sqrt{N}} + b\sqrt{\frac{\log(2/\delta)}{2N}}.
\end{aligned}
\tag{B.12}
$$

Furthermore, the above result holds under the following more general "uniform inductive marginals" condition (analogous to the "uniform marginals"):

$$
\forall i, \quad \sum_{i,j} p_{i,j}\|y_j\|^2 = \frac{\|Y\|_{\mathrm{Fr}}^2}{mn} \quad \text{and} \quad \forall j, \quad \sum_{i,j} p_{i,j}\|x_i\|^2 = \frac{\|X\|_{\mathrm{Fr}}^2}{mn}.
\tag{B.13}
$$

Proof of Proposition B.2. In this case, let us simply compute the values of σ_*^1 and σ_*^2. We have, by definition, $q_i = \sum_j p_{i,j}\|y_j\|^2$, thus under conditions (B.13), $q_i = \frac{\|Y\|_{\mathrm{Fr}}^2}{mn}$ for all i, and therefore

$$
(\sigma_*^1)^2 = \|\widetilde{L}\| = \frac{\|X^\top X\|\|Y\|_{\mathrm{Fr}}^2}{mn}.
\tag{B.14}
$$

Similarly, we have $\kappa_j = \frac{\|X\|_{\mathrm{Fr}}^2}{mn}$ for all j and

$$
(\sigma_*^2)^2 = \|\widetilde{R}\| = \frac{\|Y^\top Y\|\|X\|_{\mathrm{Fr}}^2}{mn}.
\tag{B.15}
$$

Plugging equations (B.14) and (B.15) into the first result (B.2) yields inequality (B.12) as expected. □

Remark: The sample complexity provided by Proposition B.2 above scales like $O((1/\epsilon^2)[r\mathcal{K}^2 d\log(d)])$ where ϵ is the tolerance in terms of expected loss. In the case of identity side information we recover the result of $O([rd\log(d)]/\epsilon^2)$ from [7]. In the inductive case, the result is similar but with the correction term offered by $\mathcal{K}^2$, which makes the bound better when the side information has lower effective dimension.

For instance, suppose $d_1 = d_2$, $m = n$ and the dimensions of X and Y are both $k \ll d$, and the top left $k \times k$ entries of X and Y form an identity matrix, with all other entries of X and Y being zero. Suppose also we are in the uniform sampling scenario. We then have that $\mathcal{K}^2 = k^2/d^2$, yielding a sample complexity $O([drk^2/d^2\log(d)]/\epsilon^2) = O([kr\frac{k}{d}\log(d)]/\epsilon^2)$, which is counter-intuitively tight because of the extra factor of $\frac{k}{d}$. Indeed, it would appear the problem is similar to the uniform sampling case with identity side information and a $k \times k$ matrix, which should yield a bound of $O(kr\log(k))$, but not better.

However, this factor comes from the scale parameter ϵ. Indeed, recall that the expected error is computed with respect to the sampling distribution in both cases. In this example, every entry (i, j) where either $x_i = 0$ or $y_j = 0$ is known to be equal to zero. This means that we only need $\epsilon d^2/k^2$ accuracy on the non zero entries to reach ϵ accuracy overall. However, only k^2/d^2 entries are usable (corresponding to $x_i \neq 0$ and $x_j \neq 0$). This means if we were using an optimal strategy, we would actually have a sample complexity of $O(\frac{k^2}{d^2}k\log(k))$. Our own sample complexity is actually slightly worse than that due to the smoothing procedure, which ensures stability and theoretical guarantees, but deprives us of a small part of the advantages of the weighting and adjustment. It is worth noting that this slight limitation is similar to an analogous weakness in the results of [7]: indeed, even in the MC case treated in that reference, the smoothed weighted trace norm [1] (which requires knowledge of the distribution) yields bounds of order $O(rn\log(n))$. That is the case even if the (known) distribution happens to be supported on a subset of the matrix with size $\tilde{n} \times \tilde{n}$ where $\tilde{n} \ll n$, despite the fact that a direct application of the result to the smaller matrix would yield better bounds in this case. It is interesting but challenging to consider the possibility of extending both our results and those of [7] to cover for these effects.

C Proof of Proposition 3.3

Proposition 3.3 follows from the wordier result below:

Proposition C.1 (Long version of proposition 3.3)**.** *W.p. $\geqslant 1-\delta$, for all $M \in \widetilde{\mathcal{F}}_r$:*

$$\mathbb{E}\left[l((XMY^\top)_\xi, G_\xi + \zeta_\xi)\right] - \frac{1}{N}\sum_{\xi\in\Omega} l((XMY^\top)_\xi, G_\xi + \zeta_\xi) \tag{C.1}$$
$$\leqslant \frac{8\ell\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} + \frac{12\ell\mathbf{x}\mathbf{y}\sqrt{d_1 d_2 r}(1+\log(2d))}{N} + b\sqrt{\frac{\log(2/\delta)}{2N}}.$$

Further, as long as $N \geqslant \min(d_1, d_2)\frac{18\mathbf{x}^2\mathbf{y}^2}{\Gamma}(1+\log(2d))$, we have with probability $\geqslant 1-\delta$ over the draw of the training set S for all $M \in \widetilde{\mathcal{F}}_r$

$$\mathbb{E}\left[l((XMY^\top)_\xi, G_\xi)\right] - \frac{1}{N}\sum_{\xi\in\Omega} l((XMY^\top)_\xi, G_\xi)$$
$$\leqslant \frac{12\ell\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} + b\sqrt{\frac{\log(2/\delta)}{2N}}, \tag{C.2}$$

Proof. This follows from a careful application of the Proposition B.1 to a modified problem where the side information matrices X and Y are replaced by $XP^{-1}\widetilde{D}^{-\frac{1}{2}}$ and $YQ^{-1}\widetilde{E}^{-\frac{1}{2}}$.

Let $\theta(\mathbf{x})$, $\theta(\sigma_*^1)$ (etc.) denote the value taken by $\mathbf{x}$, σ_*^1 (etc.) after the substitution above. Thus, we only need to show that replacing the values of the quantities appearing in formula (B.2) by their new values (computed below gives the formula (C.2)).

[1] The exact, non-empirical version

We have $\theta(\mathbf{x}) = \|[XP^{-1}\widetilde{D}^{-\frac{1}{2}}]^\top\|_{2,\infty} \leqslant \mathbf{x}\|\widetilde{D}^{-\frac{1}{2}}\| \leqslant= \mathbf{x}\sqrt{2\frac{d_1}{\Gamma}}$. And similarly, $\theta(\mathbf{y}) \leqslant \mathbf{y}\sqrt{2\frac{d_2}{\Gamma}}$. We also have $\theta(\mathcal{M}) = \sqrt{r}\Gamma$.

One trickier computation is that of $\theta(\sigma_*^1)$ and $\theta(\sigma_*^2)$:

$\theta(\sigma_*^1)$ is the spectral norm of the matrix $\theta(X) = XP^{-1}\widetilde{D}^{-\frac{1}{2}}$ evaluated with respect to the post-substitution inner product $\langle,\rangle_{\theta(l)}$. Note that the new values $\theta(q_i)$ and $\theta(\kappa_i)$ for κ_j and q_j have the following properties:

$$\begin{aligned}
\theta(q_i) &= \sum_j p_{i,j}\|\theta(y_j)\|^2 \\
&= \sum_j p_{i,j}\|y_j Q^{-1}\widetilde{E}^{\frac{1}{2}}\|^2 \\
&\leqslant \sum_j p_{i,j}\|y_j\|^2\|\widetilde{E}^{\frac{1}{2}}\|^2 \\
&\leqslant \frac{2q_i d_2}{\Gamma},
\end{aligned} \tag{C.3}$$

and similarly

$$\theta(\kappa_j) \leqslant \frac{2\kappa_j d_1}{\Gamma}.$$

In particular, for any vector $v \in \mathbb{R}^m$ we have

$$\|v\|_{\theta(l)}^2 = \langle v, v\rangle_{\theta(l)} = v^\top \operatorname{diag}(\theta(q))v \leqslant v^\top \operatorname{diag}(q)v\frac{2d_2}{\Gamma} \leqslant \|v\|_l^2\frac{2d_2}{\Gamma}, \tag{C.4}$$

and similarly for vectors in $\mathbb{R}^n$ with a factor of $\frac{2d_1}{\Gamma}$.

As a result we can compute:

$$\begin{aligned}
\theta(\sigma_*^1)^2 &= \|\theta(X)^\top \operatorname{diag}(\theta(q))(\theta(X))\| \\
&= \|(XP^{-1}\widetilde{D}^{-\frac{1}{2}})^\top \operatorname{diag}(\theta(q))(XP^{-1}\widetilde{D}^{-\frac{1}{2}})\| \\
&\leqslant \frac{2d_2}{\Gamma}\|(XP^{-1}\widetilde{D}^{-\frac{1}{2}})^\top \operatorname{diag}(q)(XP^{-1}\widetilde{D}^{-\frac{1}{2}})\| \\
&= \frac{2d_2}{\Gamma}\|\widetilde{D}^{-\frac{1}{2}}P[P^{-1}DP]P^{-1}\widetilde{D}^{-\frac{1}{2}}\| = \frac{2d_2}{\Gamma}\|2I\| \\
&\leqslant \frac{4d_2}{\Gamma},
\end{aligned} \tag{C.5}$$

and similarly

$$\theta(\sigma_*^1)^2 \leqslant \frac{4d_1}{\Gamma}. \tag{C.6}$$

Plugging the post substitution values computed above into each of the relevant expressions in Proposition B.1, we obtain first that w.p. $\geqslant 1-\delta$:

$$\frac{4\ell}{\sqrt{N}}\theta(\mathcal{M})\max(\theta(\sigma_*^1),\theta(\sigma_*^2))(1+\sqrt{\log(2d)}) + \frac{6\ell}{N}\theta(\mathcal{M})\theta(\mathbf{xy})(1+\log(2d)) \tag{C.7}$$

$$\leqslant \frac{4\ell}{\sqrt{N}}\Gamma\sqrt{r}\max(\sqrt{\frac{4d_2}{\Gamma}},\sqrt{\frac{4d_1}{\Gamma}})(1+\sqrt{\log(2d)}) + +\frac{12\ell\sqrt{d_1d_2}/\Gamma}{N}\sqrt{r}\Gamma\mathbf{xy}(1+\log(2d)) \tag{C.8}$$

$$= \frac{8\ell\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} + \frac{12\ell\sqrt{rd_1d_2}}{N}\mathbf{xy}(1+\log(2d)) \tag{C.9}$$

$$\tag{C.10}$$

as expected.

And then also that (w.p. $\geqslant 1-\delta$) $\mathbb{E}\left[l((XMY^\top)_\xi, G_\xi)\right] - \frac{1}{N}\sum_{\xi\in\Omega} l((XMY^\top)_\xi, G_\xi) - b\sqrt{\frac{\log(2/\delta)}{2n}}$ is bounded above by

$$\begin{aligned}\frac{6\ell\theta(\mathcal{M})\max(\theta(\sigma_*^1),\theta(\sigma_*^2))(1+\sqrt{\log(2d)})}{\sqrt{N}} &= \frac{6\ell\sqrt{\Gamma}\sqrt{r}\max(\sqrt{\frac{4d_2}{\Gamma}},\sqrt{\frac{4d_1}{\Gamma}})(1+\sqrt{\log(2d)})}{\sqrt{N}} \\ &= \frac{12\ell\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}},\end{aligned}$$

with the condition that N needs to be larger than

$$\begin{aligned}&9\theta([\mathbf{xy}/\max(\sigma_*^1,\sigma_*^2)])^2(1+\log(2d)) \\ &= 9\left[\mathbf{xy}\sqrt{\frac{2d_1}{\Gamma}}\sqrt{\frac{2d_2}{\Gamma}}/\sqrt{\frac{2}{\Gamma}}\sqrt{d}\right]^2\sqrt{r}\Gamma(1+\log(2d)) \\ &= \min(d_1,d_2)\frac{18\mathbf{x}^2\mathbf{y}^2}{\Gamma}(1+\log(2d)),\end{aligned}\tag{C.11}$$

as expected. □

D Proof of Theorem 3.2

Theorem 3.2 follows from the longer version below.

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

$$l(\check{Z}) \leqslant \inf_{\check{\mathcal{F}}_r} l(Z) + \left[48\ell\sqrt{r}\gamma(\mathbf{x}+\mathbf{y})^2 + 2b\right]\sqrt{\frac{2\log(\frac{12d}{\delta})[\gamma(d+3)+\gamma^2]}{N}},\tag{D.1}$$

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

$$l(\check{Z}) \leqslant \inf_{\check{\mathcal{F}}_r} l(Z) + \left[96\ell\sqrt{r}\gamma(\mathbf{x}+\mathbf{y})^2 + 4b\right]\sqrt{\frac{2\log(12d)[\gamma(d+3)+\gamma^2]}{N}}.\tag{D.2}$$

Proof of Theorem D.1. The lemmas which are used are proved below.

We write Z^* for an element of $\arg\min_{\check{\mathcal{F}}_r} l(Z)$. First, by applying Proposition 3.3, we have that $N \geqslant \sqrt{\min(d_1,d_2)}18\gamma(1+\log(2d))$, we have with probabiltiy $\geqslant 1-\delta/3$:

$$l(\check{Z}) - \hat{l}_S(\check{Z}) \leqslant \frac{12\ell\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} + b\sqrt{\frac{\log(6/\delta)}{2N}}.\tag{D.3}$$

Define $C(S) = \max\left(0, \left\|\frac{1}{\sqrt{r_*\Gamma}}\widetilde{M_*}\right\|_* - 1\right)$. Note that $(1-C(S))Z^* \in \check{\mathcal{F}}_r$. Thus, using Lemma E.4 we also have similarly with probability $\geqslant 1-\delta/3$:

$$\begin{aligned}&\hat{l}_S((1-C(S))Z^*) - l((1-C(S))Z^*) \\ &\leqslant \frac{24\ell\sqrt{\Gamma}\gamma\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} + b\sqrt{\frac{\log(6/\delta)}{2N}},\end{aligned}\tag{D.4}$$

as long as $N \geqslant 8\gamma^2 + \gamma[8d+20][\log(2d)+\log(\frac{6}{\delta})]$. By definition, since $(1-C(S))Z^* \in \check{\mathcal{F}}_r$ we also have

$$\hat{l}_S(\check{Z}) - \hat{l}_S((1-C(S))Z^*) \leqslant 0.\tag{D.5}$$

Next, by Lemma E.3, as long as $N \geqslant 2\log(\frac{6d}{\delta})[\gamma(d+3)+\gamma^2]$, with probability $\geqslant 1-\delta/3$ over the draw of the training set:

$$l((1-C(S))Z_*) - l(Z_*)$$

$$\leqslant \ell \|\widetilde{M_*}\|_* \left[\frac{1}{\underline{x}^2} + \frac{1}{\underline{y}^2}\right] \sqrt{\frac{2\log(\frac{12d}{\delta})[\gamma(d+3)+\gamma^2]}{N}}$$

$$\leqslant \ell\sqrt{r}\Gamma \left[\frac{1}{\underline{x}^2} + \frac{1}{\underline{y}^2}\right] \sqrt{\frac{2\log(\frac{12d}{\delta})[\gamma(d+3)+\gamma^2]}{N}}. \tag{D.6}$$

Combining all of the above, we get that as long as $N \geqslant 2\log(\frac{6d}{\delta})[\gamma(d+3)+\gamma^2]$ and $N \geqslant \sqrt{\min(d_1,d_2)}18\gamma(1+\log(2d))$, we have

$$l(\breve{Z}) - l(Z_*) \tag{D.7}$$

$$\leqslant l(\breve{Z}) - \hat{l}_S(\breve{Z}) + \hat{l}_S(\breve{Z}) - \hat{l}_S((1-C(S))Z^*) + \tag{D.8}$$

$$\hat{l}_S((1-C(S))Z^*) - l((1-C(S))Z^*) + l((1-C(S))Z_*) - l(Z_*)$$

$$\leqslant \frac{12\ell\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} + b\sqrt{\frac{\log(6/\delta)}{2N}} \tag{D.9}$$

$$+ \frac{24\ell\gamma\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} + b\sqrt{\frac{\log(6/\delta)}{2N}} \tag{D.10}$$

$$+ \ell\sqrt{r}\Gamma \left[\frac{1}{\underline{x}^2} + \frac{1}{\underline{y}^2}\right] \sqrt{\frac{2\log(\frac{12d}{\delta})[\gamma(d+3)+\gamma^2]}{N}} \tag{D.11}$$

$$\leqslant \frac{48\ell\gamma\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} + 2b\sqrt{\frac{\log(6/\delta)}{2N}} \tag{D.12}$$

$$+ \ell\sqrt{r}\gamma(\mathbf{x}^2+\mathbf{y}^2)\sqrt{\frac{2\log(\frac{12d}{\delta})[\gamma(d+3)+\gamma^2]}{N}} \tag{D.13}$$

$$\leqslant \frac{48\ell\gamma\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} \tag{D.14}$$

$$+ \left[\ell\sqrt{r}\gamma(\mathbf{x}^2+\mathbf{y}^2) + 2b\right]\sqrt{\frac{2\log(\frac{12d}{\delta})[\gamma(d+3)+\gamma^2]}{N}} \tag{D.15}$$

$$\leqslant \left[48\ell\sqrt{r}\gamma(\mathbf{x}+\mathbf{y})^2 + 2b\right]\sqrt{\frac{2\log(\frac{12d}{\delta})[\gamma(d+3)+\gamma^2]}{N}}. \tag{D.16}$$

Furthermore, the conditions on N can now be dropped since the RHS is greater than b whenever N fails to satisfy either of them.

The expectation version of the theorem follows directly from Lemma F.5 □

E Lemmas for the proof of Theorem 3.2

Proposition E.1. *For any $\delta \in (0,1)$, with probability $\geqslant 1-\delta$, we have*

$$\frac{1}{\sqrt{2}} \leqslant \left\| \widetilde{D}^{\frac{1}{2}} P \widehat{P}^{-1} \breve{D}^{-\frac{1}{2}} \right\| \leqslant \sqrt{2}, \tag{E.1}$$

as long as $N \geqslant 8\gamma^2 + \gamma[8d_1+20][\log(2d_1)+\log(\frac{1}{\delta})]$.

Similarly, for any $\delta \in (0,1)$, with probability $\geqslant 1-\delta$, we have

$$\frac{1}{\sqrt{2}} \leqslant \left\| \widetilde{E}^{\frac{1}{2}} Q \widehat{Q}^{-1} \breve{E}^{-\frac{1}{2}} \right\| \leqslant \sqrt{2}, \tag{E.2}$$

as long as $N \geqslant [8\gamma^2 + \gamma[8d_2+20]][\log(2d_2)+\log(\frac{1}{\delta})]$.

Proof. We will write T for the matrix $\widetilde{D}^{\frac{1}{2}}P\widehat{P}^{-1}\breve{D}^{-\frac{1}{2}}$ whose spectral norm we want to bound.

We consider the matrix

$$\mathcal{T} := \widetilde{D}^{-\frac{1}{2}}P\widehat{P}^{-1}\breve{D}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}} = (T^{-1})^{\top}(T^{-1}). \tag{E.3}$$

We can write $\mathcal{T}$ as a sum of independent random matrices as follows:

$$\begin{aligned}\mathcal{T} :&= \frac{1}{N}\sum_{\xi\in\Omega}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2}x_{\xi_1}x_{\xi_1}^{\top}\|y_{\xi_2}\|^2 + \frac{1}{2d_1}\|x_{\xi_1}\|^2\|y_{\xi_2}\|^2 I\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\\ &= \frac{1}{N}\sum_{i,j}h_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2}x_i x_i^{\top}\|y_j\|^2 + \frac{1}{2d_1}\|x_i\|^2\|y_j\|^2 I\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\\ &= \frac{1}{N}\sum_{o=1}^{N}\Lambda_o,\end{aligned} \tag{E.4}$$

where Ω is the multi-set containing all the iid sampled entries and $\Lambda = \widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2}x_{\xi_1^o}x_{\xi_1^o}^{\top}\|y_{\xi_2^o}\|^2 + \frac{1}{2d_1}\|x_{\xi_1}\|^2\|y_{\xi_2}\|^2 I\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}$ and the ξ^o $(o = 1,\dots,N)$ are the sampled entries.

Now, we can compute the expectation of $\mathcal{T}$ and Λ as follows:

$$\mathbb{E}(\mathcal{T}) = \mathbb{E}(\Lambda_\xi) = \sum_{i,j}p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2}x_i x_i^{\top}\|y_j\|^2 + \frac{1}{2d_1}\|x_i\|^2\|y_j\|^2 I\right]P^{-1}\widetilde{D}^{-\frac{1}{2}} \tag{E.5}$$

$$= \widetilde{D}^{-\frac{1}{2}}PP^{-1}\widetilde{D}PP^{-1}\widetilde{D}^{-\frac{1}{2}} = I. \tag{E.6}$$

Now, note that for any $(i,j)\in\{1,2,\dots,m\}\times\{1,2,\dots,n\}$ we have

$$\begin{aligned}\|\Lambda_{(i,j)}\| &= \left\|\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2}x_i x_i^{\top}\|y_j\|^2 + \frac{1}{2d_1}\|x_i\|^2\|y_j\|^2 I\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\| \leqslant (\frac{1}{2}\mathbf{x}^2\mathbf{y}^2 + \frac{1}{2d_1}\widehat{\Gamma})\|\widetilde{D}\|^{-1}\\ &\leqslant (\frac{1}{2}\mathbf{x}^2\mathbf{y}^2 + \frac{1}{2d_1}\widehat{\Gamma})\frac{2d_1}{\widehat{\Gamma}} \leqslant \frac{\mathbf{x}^2\mathbf{y}^2}{\underline{x}^2\underline{y}^2} + 1 = \gamma + 1\end{aligned} \tag{E.7}$$

By abuse of notation, we write below Λ for the random variable Λ_ξ where $\xi\in\{1,2,\dots,m\}\times\{1,2,\dots,n\}$ is distributed according to p.

We now begin to bound $\|\mathbb{E}((\Lambda-\mathbb{E}(\Lambda))(\Lambda-\mathbb{E}(\Lambda))^{\top})\|$. We first note that

$$\begin{aligned}\left\|\mathbb{E}\left((\Lambda-\mathbb{E}(\Lambda))(\Lambda-\mathbb{E}(\Lambda))^{\top}\right)\right\| &= \left\|\mathbb{E}(\Lambda\Lambda^{\top}) - \mathbb{E}(\Lambda)\mathbb{E}(\Lambda)^{\top}\right\|\\ &= \left\|\mathbb{E}(\Lambda\Lambda^{\top}) - I\right\| \leqslant \left\|\mathbb{E}(\Lambda\Lambda^{\top})\right\|.\end{aligned} \tag{E.8}$$

Thus, we now note that by equation (E.4):

$$\mathbb{E}(\Lambda\Lambda^{\top}) = \tag{E.9}$$

$$\sum_{i,j}p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{\|y_j\|^2}{2}x_i x_i^{\top} + \frac{\|x_i\|^2\|y_j\|^2}{2d_1}I\right]P^{-1}\widetilde{D}^{-1}P\left[\frac{\|y_j\|^2}{2}x_i x_i^{\top} + \frac{\|x_i\|^2\|y_j\|^2}{2d_1}I\right]^{\top}P^{-1}\widetilde{D}^{-\frac{1}{2}}.$$

From this it follows that

$$\begin{aligned}\|\mathbb{E}(\Lambda\Lambda^{\top})\| \leqslant &\left\|\sum_{i,j}p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2}x_i x_i^{\top}\|y_j\|^2\right]P^{-1}\widetilde{D}^{-1}P\left[\frac{1}{2}x_i x_i^{\top}\|y_j\|^2\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\|\\ &+ \left\|\sum_{i,j}p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2d_1}\|x_i\|^2\|y_j\|^2 I\right]P^{-1}\widetilde{D}^{-1}P\left[\frac{1}{2}x_i x_i^{\top}\|y_j\|^2\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\|\end{aligned} \tag{E.10}$$

$$+\left\|\sum_{i,j} p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2}x_ix_i^\top\|y_j\|^2\right]P^{-1}\widetilde{D}^{-1}P\left[\frac{1}{2d_1}\|x_i\|^2\|y_j\|^2I\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\|$$
$$+\left\|\sum_{i,j} p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2d_1}\|x_i\|^2\|y_j\|^2I\right]P^{-1}\widetilde{D}^{-1}P\left[\frac{1}{2d_1}\|x_i\|^2\|y_j\|^2I\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\|.$$

We bound each of the four terms above separately:

For the first (and key) term, we have:

$$\begin{aligned}
&\frac{1}{4}\left\|\sum_{i,j} p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[x_ix_i^\top\|y_j\|^2\right]P^{-1}\widetilde{D}^{-1}P\left[x_ix_i^\top\|y_j\|^2\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\| \\
&\leqslant \frac{1}{4}\left\|\sum_{i,j} p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[x_ix_i^\top\|y_j\|^2\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\|\sup_{i,j}\left\|\widetilde{D}^{-\frac{1}{2}}P\left[x_ix_i^\top\|y_j\|^2\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\| \\
&\leqslant \frac{1}{4}\frac{2d_1}{\Gamma}\mathbf{x}^2\mathbf{y}^2\left\|\sum_{i,j} p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[x_ix_i^\top\|y_j\|^2\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\| \\
&= \frac{d_1}{2\Gamma}\mathbf{x}^2\mathbf{y}^2\left\|\widetilde{D}^{-\frac{1}{2}}PP^{-1}DPP^{-1}\widetilde{D}^{-\frac{1}{2}}\right\| = \frac{d_1}{2\Gamma}\mathbf{x}^2\mathbf{y}^2\|D\widetilde{D}^{-1}\| \leqslant \frac{d_1}{\underline{x}^2\underline{y}^2}\mathbf{x}^2\mathbf{y}^2 = d_1\gamma.
\end{aligned}\tag{E.11}$$

For the second term we have

$$\begin{aligned}
&\left\|\sum_{i,j} p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2d_1}\|x_i\|^2\|y_j\|^2I\right]P^{-1}\widetilde{D}^{-1}P\left[\frac{1}{2}x_ix_i^\top\|y_j\|^2\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\| \\
&\leqslant \frac{\mathbf{x}^2\mathbf{y}^2}{2d_1}\left\|\widetilde{D}^{-\frac{1}{2}}I\widetilde{D}^{-\frac{1}{2}}\right\|\left\|\sum_{i,j} p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2}x_ix_i^\top\|y_j\|^2\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\| \\
&\leqslant \frac{\mathbf{x}^2\mathbf{y}^2}{2d_1}\frac{2d_1}{\Gamma}\frac{1}{2}\left\|\widetilde{D}^{-\frac{1}{2}}PP^{-1}DPP^{-1}\widetilde{D}^{-\frac{1}{2}}\right\| \leqslant \frac{\mathbf{x}^2\mathbf{y}^2}{\underline{x}^2\underline{y}^2} = \gamma.
\end{aligned}\tag{E.12}$$

For the third term we obtain similarly:

$$\left\|\sum_{i,j} p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2}x_ix_i^\top\|y_j\|^2\right]P^{-1}\widetilde{D}^{-1}P\left[\frac{1}{2d_1}\|x_i\|^2\|y_j\|^2I\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\| \leqslant \frac{\mathbf{x}^2\mathbf{y}^2}{\underline{x}^2\underline{y}^2} = \gamma. \tag{E.13}$$

Finally for the fourth term we have:

$$\begin{aligned}
&\left\|\sum_{i,j} p_{i,j}\widetilde{D}^{-\frac{1}{2}}P\left[\frac{1}{2d_1}\|x_i\|^2\|y_j\|^2I\right]P^{-1}\widetilde{D}^{-1}P\left[\frac{1}{2d_1}\|x_i\|^2\|y_j\|^2I\right]P^{-1}\widetilde{D}^{-\frac{1}{2}}\right\| \\
&\leqslant \left\|\frac{\mathbf{x}^2\mathbf{y}^2\widetilde{D}^{-1}}{2d_1}\right\|^2 \leqslant \left\|\frac{\mathbf{x}^2\mathbf{y}^2 2d_1}{2d_1\Gamma}\right\|^2 \leqslant \gamma^2.
\end{aligned}\tag{E.14}$$

Plugging equations (E.11), (E.12), (E.13) and (E.14) into equations (E.10) and (E.8) we finally obtain:

$$\left\|\mathbb{E}\left((\Lambda-\mathbb{E}(\Lambda))(\Lambda-\mathbb{E}(\Lambda))^\top\right)\right\| \leqslant \left\|\mathbb{E}(\Lambda\Lambda^\top)\right\| \leqslant \gamma(d_1+2)+\gamma^2. \tag{E.15}$$

We now apply the non-communtative Bernstein inequality (F.3) to $\mathcal{T}-\mathbb{E}(\mathcal{T})$ which is the average of N i.i.d. instances of Λ. With the notation from Proposition F.3 we have $M=\gamma+1$ (from equation (E.7)), $\nu^2=\sum_{o=1}^N\frac{1}{N^2}[\gamma(d_1+2)+\gamma^2]=\frac{1}{N}[\gamma(d_1+2)+\gamma^2]$ (from equation (E.15)), $n=m=d_1$ and we obtain (for all τ):

$$\mathbb{P}\left(\|\mathcal{T}-\mathbb{E}(\mathcal{T})\| \geqslant \tau\right) \leqslant (2d_1)\exp\left(-\frac{\tau^2/2}{\nu^2+M\tau/3}\right)$$

$$\leqslant (2d_1) \exp\left(-\frac{N\tau^2/2}{[\gamma(d_1+2)+\gamma^2]+(\gamma+1)\tau/3}\right) \tag{E.16}$$

Setting $\tau = \frac{1}{2}$ we obtain, as long as $N \geqslant [8\gamma^2 + \gamma[8d_1 + 20]][\log(2d_1) + \log(\frac{1}{\delta})]$:

$$\begin{aligned}
\mathbb{P}\left(\|\mathcal{T} - \mathbb{E}(\mathcal{T})\| \geqslant \frac{1}{2}\right) &\leqslant (2d_1)\exp\left(-\frac{\tau^2/2}{\nu^2 + M\tau/3}\right) \\
&\leqslant (2d_1)\exp\left(-\frac{N}{8[\gamma(d_1+2)+\gamma^2]+2(\gamma+1)}\right) \\
&\leqslant (2d_1)\exp\left(-\frac{N}{8\gamma^2+\gamma[8d_1+20]}\right) \\
&\leqslant \delta. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (E.17)
\end{aligned}$$

Thus, we now know that as long as $N \geqslant 8\gamma^2 + \gamma[8d_1 + 20][\log(2d_1) + \log(\frac{1}{\delta})]$ we have with probability $\geqslant 1 - \delta$ that

$$\|\mathcal{T} - \mathbb{E}(\mathcal{T})\| \leqslant \frac{1}{2}. \tag{E.18}$$

This already implies that $\|\mathcal{T}\| \leqslant 1 + 0.5 \leqslant 2$ and therefore $\|T^{-1}\| \leqslant \sqrt{2}$, leaving us only the second inequality to prove.

We will show that inequality (E.18) actually implies inequality (E.1).

To that effect, recall from equation (E.3) that $\mathcal{T} = (T^{-1})^\top(T^{-1}) = G^{-1}$ where $G = TT^\top$. Thus we have $G = [I + (\mathcal{T} - I)]^{-1}$. Rewriting this as $G\,[I + (\mathcal{T} - I)] = I$ and taking spectral norms on both sides we obtain

$$\|G\|\sigma_{\inf}\left([I + (\mathcal{T} - I)]\right) \leqslant 1, \tag{E.19}$$

where for any symmetric matrix A, $\sigma_{\inf}(A)$ denotes the smallest eigenvalue of A.

Now note that by inequality (E.18), for any unit vector v, we have

$$v^\top\left[I + (\mathcal{T} - I)\right]v = 1 - v^\top\left(\mathcal{T} - I\right)v \geqslant 1 - \|\left(\mathcal{T} - I\right)\| \geqslant 1 - \frac{1}{2} = \frac{1}{2}. \tag{E.20}$$

Thus the smallest eigenvalue of $\|[I + (\mathcal{T} - I)]\|$ is bounded below by $\frac{1}{2}$, i.e.

$$\sigma_{\inf}\left([I + (\mathcal{T} - I)]\right) \geqslant \frac{1}{2}. \tag{E.21}$$

Plugging inequality (E.21) back into identity (E.19), we obtain:

$$\|G\| \leqslant 2. \tag{E.22}$$

Finally, recall that $G = TT^\top$ and thus $\|G\| = \|T\|^2$, which together with inequality (E.22) finally implies

$$\|T\| \leqslant \sqrt{2}, \tag{E.23}$$

as expected. □

Lemma E.2. *Let $\widetilde{M} \in \mathbb{R}^{d_1 \times d_2}$ be a fixed matrix with $\|M\|_* = 1$. For any $\delta \in (0,1)$ we have that w.p. $\geqslant 1 - \delta$, (as long as $N \geqslant 2\log(\frac{2d}{\delta})[\gamma(d+3)+\gamma^2]$):*

$$\|\breve{M}\|_* = \|\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}\widetilde{M}\widetilde{E}^{-\frac{1}{2}}Q\widehat{Q}^{-1}\breve{E}^{\frac{1}{2}}\|_* \leqslant \|\widetilde{M}\|_*\left[1 + \sqrt{\frac{2\log(\frac{4d}{\delta})[\gamma(d+3)+\gamma^2]}{N}}\right], \tag{E.24}$$

where $d := \max(d_1, d_2)$.

Proof. Writing $\breve{M}$ for the matrix $\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}\widetilde{M}\widetilde{E}^{-\frac{1}{2}}Q\widehat{Q}^{-1}\breve{E}^{\frac{1}{2}}$ we want to control, we have by the properties of the trace norm:

$$\|\breve{M}\|_* = \max_{A,B}\left(\frac{1}{2}\left[\|A\|_{\text{Fr}}^2 + \|B\|_{\text{Fr}}^2\right] : AB^\top = \breve{M}\right)$$

Let $\breve{A}$, $\breve{B}$ denote the matrices which realize the maximum above. Now note that we have $[\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}]^{-1}\breve{A}\breve{B}[\widetilde{E}^{-\frac{1}{2}}Q\widehat{Q}^{-1}\widehat{E}^{\frac{1}{2}}]^{-1} = [\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}]^{-1}\breve{M}[\widetilde{E}^{-\frac{1}{2}}Q\widehat{Q}^{-1}\widehat{E}^{\frac{1}{2}}]^{-1}$, i.e. $\widetilde{A}\widetilde{B} = \widetilde{M}$ where

$$\begin{aligned}\widetilde{A} &:= [\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}]^{-1}\breve{A} \quad \text{and} \\ \widetilde{B} &:= [\breve{E}^{\frac{1}{2}}\widehat{Q}Q^{-1}\widetilde{E}^{-\frac{1}{2}}]^{-1}\breve{B}.\end{aligned} \tag{E.25}$$

In particular, we have

$$\begin{aligned}\|\widetilde{M}\|_* &= \max_{A,B}\left(\frac{1}{2}\left[\|A\|_{\text{Fr}}^2 + \|B\|_{\text{Fr}}^2\right] : AB = \widetilde{M}\right) \\ &\geqslant \frac{1}{2}\left[\|\widetilde{A}\|_{\text{Fr}}^2 + \|\widetilde{B}\|_{\text{Fr}}^2\right].\end{aligned} \tag{E.26}$$

Now, we can express $\breve{A}$ and $\breve{B}$ as $[\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}]\widetilde{A}$ and $[\breve{E}^{\frac{1}{2}}\widehat{Q}Q^{-1}\widetilde{E}^{-\frac{1}{2}}]\widetilde{B}$ respectively, and thus we have

$$\begin{aligned}\|\breve{M}\|_* &= \frac{1}{2}\left[\|\breve{A}\|_{\text{Fr}}^2 + \|\breve{B}\|_{\text{Fr}}^2\right] \\ &= \frac{1}{2}\left[\|[\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}]]\widetilde{A}\|_{\text{Fr}}^2 + \|[\breve{E}^{\frac{1}{2}}\widehat{Q}Q^{-1}\widetilde{E}^{-\frac{1}{2}}]]\widetilde{B}\|_{\text{Fr}}^2\right] \\ &\leqslant \max(\|[\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}]\|, \|[\breve{E}^{\frac{1}{2}}\widehat{Q}Q^{-1}]\|)^2\frac{1}{2}\left[\|\widetilde{A}\|_{\text{Fr}}^2 + \|\widetilde{B}\|_{\text{Fr}}^2\right] \\ &\leqslant \max(\|[\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}]\|, \|[\breve{E}^{\frac{1}{2}}\widehat{Q}Q^{-1}\widetilde{E}^{-\frac{1}{2}}]]\|)^2\|\widetilde{M}\|_*.\end{aligned} \tag{E.27}$$

Hence, we need to bound the quantity $\max(\|[\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}]\|, \|[\breve{E}^{\frac{1}{2}}\widehat{Q}Q^{-1}\widetilde{E}^{-\frac{1}{2}}]]\|)$. Using similar notation to proposition E.1 we have $\mathcal{T}_1 = [\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}][\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}]^\top$ and $\mathcal{T}_2 = [\breve{E}^{\frac{1}{2}}\widehat{Q}Q^{-1}\widetilde{E}^{-\frac{1}{2}}]]$

Picking up the proof of proposition (E.1) at equation (E.16), we obtain (for all $\tau \leqslant 1$):

$$\begin{aligned}\mathbb{P}(\|[\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}]\|^2 \geqslant 1+\tau) &\leqslant \mathbb{P}\left(\|\mathcal{T}_1 - I\| \geqslant \tau\right) \\ &\leqslant (2d_1)\exp\left(-\frac{\tau^2/2}{\nu^2 + M\tau/3}\right) \\ &\leqslant (2d_1)\exp\left(-\frac{N\tau^2/2}{[\gamma(d_1+2)+\gamma^2]+(\gamma+1)\tau/3}\right) \\ &\leqslant (2d_1)\exp\left(-\frac{N\tau^2/2}{[\gamma(d_1+3)+\gamma^2]}\right).\end{aligned} \tag{E.28}$$

□

Rewriting, this implies that with probablity greater than $1-\delta$, we have

$$\|[\breve{D}^{\frac{1}{2}}\widehat{P}P^{-1}\widetilde{D}^{-\frac{1}{2}}]\|^2 \leqslant 1 + \sqrt{\frac{2\log(\frac{2d_1}{\delta})[\gamma(d_1+3)+\gamma^2]}{N}}, \tag{E.29}$$

as long as $N \geqslant 2\log(\frac{2d_1}{\delta})[\gamma(d_1+3)+\gamma^2]$.

Similarly, (as long as $N \geqslant 2\log(\frac{2d_2}{\delta})[\gamma(d_2+3)+\gamma^2]$) we have (for any δ) with probability $\geqslant 1-\delta$,

$$\|[\breve{E}^{\frac{1}{2}}\widehat{Q}Q^{-1}\widetilde{Q}^{-\frac{1}{2}}]\|^2 \leqslant 1+\sqrt{\frac{2\log(\frac{2d_2}{\delta})[\gamma(d_2+3)+\gamma^2]}{N}}. \tag{E.30}$$

Putting the above two results together and plugging them into equation (E.27), we obtain (as long as $N \geqslant 2\log(\frac{2d}{\delta})[\gamma(d+3)+\gamma^2]$) with probability greater than $1-\delta$:

$$\|\breve{M}\|_* \leqslant \|\widetilde{M}\|_* \left[1+\sqrt{\frac{2\log(\frac{4d}{\delta})[\gamma(d+3)+\gamma^2]}{N}}\right], \tag{E.31}$$

as expected.

Lemma E.3. *Fix M_* such that $\|\widetilde{M_*}\| = \|\widetilde{D}^{\frac{1}{2}}PMQ^{-1}\widetilde{E}^{\frac{1}{2}}\| = \sqrt{r_*}\Gamma \leqslant \sqrt{r}\Gamma$. Define*

$$C(S) = \max\left(0, \left\|\frac{1}{\sqrt{r_*}\Gamma}\breve{M_*}\right\|_* - 1\right), \tag{E.32}$$

where $\breve{M} = \breve{D}^{\frac{1}{2}}\widehat{P}M\widehat{Q}^{-1}\breve{E}^{\frac{1}{2}}$.

Writing $Z_* = XM_*Y^\top = \widetilde{X}\widetilde{M_*}\widetilde{Y}^\top = \breve{X}\breve{M_*}\breve{Y}^\top$, as long as $N \geqslant 2\log(\frac{2d}{\delta})[\gamma(d+3)+\gamma^2]$, with probability $\geqslant 1-\delta$ over the draw of the training set:

$$\begin{aligned}
&\mathbb{E}_{(i,j)\sim p}\left(l\left[(1-C(S))[Z_*]_{i,j}, G_{(i,j)}\right] - l\left[[Z_*]_{i,j}, G_{(i,j)}\right]\right) \\
&\leqslant \ell\|\widetilde{M_*}\|_* \left[\frac{1}{\underline{x}^2}+\frac{1}{\underline{y}^2}\right]\sqrt{\frac{2\log(\frac{4d}{\delta})[\gamma(d+3)+\gamma^2]}{N}}
\end{aligned} \tag{E.33}$$

Proof. We have, writing Θ for the matrix with $\Theta_{i,j} = p_{i,j}$ and using the notation $|A|$ for the matrix obtained from A by replacing each entry by its absolute value:

$$\begin{aligned}
&\mathbb{E}_{(i,j)\sim p}\left(l\left[(1-C(S))[Z_*]_{i,j}, G_{(i,j)}\right] - l\left[[Z_*]_{i,j}, G_{(i,j)}\right]\right) \\
&= \sum_{i,j} p_{i,j} l\left[(1-C(S))[Z_*]_{i,j}, G_{(i,j)}\right] - l\left[[Z_*]_{i,j}, G_{(i,j)}\right] \\
&\leqslant \ell\sum_{i,j} p_{i,j}\left|(1-C(S))[Z_*]_{i,j} - [Z_*]_{i,j}\right| \\
&\leqslant \ell C(S)\sum_{i,j} p_{i,j}\left|[Z_*]_{i,j}\right| = \ell C(S)\langle\Theta, |Z_*|\rangle = C(S)\left\langle\Theta, |XM_*Y^\top|\right\rangle \\
&= \ell C(S)\left\langle\widetilde{\Theta}, \widetilde{X}\widetilde{M_*}\widetilde{Y}^\top\right\rangle,
\end{aligned} \tag{E.34}$$

where we write $\widetilde{\Theta}$ for the matrix with $\widetilde{\Theta}_{i,j} = \Theta_{i,j}\,\mathrm{sign}([XM_*Y^\top]_{i,j})$ for all i,j.

Replacing the expressions $XP^{-1}\widetilde{D}^{-\frac{1}{2}}$ and $YQ^{-1}\widetilde{E}^{-\frac{1}{2}}$ for $\widetilde{X}$ and $\widetilde{Y}$ respectively and using the circular invariance of the trace we obtain:

$$\begin{aligned}
&\mathbb{E}_{(i,j)\sim p}\left(l\left[(1-C(S))[Z_*]_{i,j}, G_{(i,j)}\right] - l\left[[Z_*]_{i,j}, G_{(i,j)}\right]\right) \\
&\leqslant \ell C(S)\left\langle\widetilde{\Theta}, \widetilde{X}\widetilde{M_*}\widetilde{Y}^\top\right\rangle = C(S)\left\langle\widetilde{\Theta}, [XP^{-1}\widetilde{D}^{-\frac{1}{2}}]\widetilde{M_*}\widetilde{E}^{-\frac{1}{2}}QY^\top\right\rangle \\
&= \ell C(S)\left\langle\widetilde{D}^{-\frac{1}{2}}PX^\top\widetilde{\Theta}YQ^{-1}\widetilde{E}^{-\frac{1}{2}}, \widetilde{M_*}\right\rangle \\
&\leqslant \ell C(S)\|\widetilde{M_*}\|_*\|\widetilde{D}^{-\frac{1}{2}}PX^\top\widetilde{\Theta}YQ^{-1}\widetilde{E}^{-\frac{1}{2}}\| \\
&= \ell C(S)\|\widetilde{M_*}\|_*\left\|\left[\widetilde{D}^{-\frac{1}{2}}PX^\top A^{-1}\right]A\widetilde{\Theta}B\left[B^{-1}YQ^{-1}\widetilde{E}^{-\frac{1}{2}}\right]\right\|,
\end{aligned} \tag{E.35}$$

where A, B are arbitrary invertible matrices.

Now by Lemma E.5, setting $A = \operatorname{diag}(\|x_1\|^2, \ldots, \|x_m\|^2)$ and $B = \operatorname{diag}(\|y_1\|^2, \ldots, \|y_n\|^2)$, we obtain:

$$
\begin{aligned}
&\left\| \left[\widetilde{D}^{-\frac{1}{2}} P X^\top A^{-1}\right] A \widetilde{\Theta} B \left[B^{-1} Y Q^{-1} \widetilde{E}^{-\frac{1}{2}}\right] \right\| \\
&\leqslant \frac{1}{2} \left\| \left[\widetilde{D}^{-\frac{1}{2}} P X^\top A^{-1}\right] \operatorname{diag}(A\widetilde{\Theta} B 1_n) \left[\widetilde{D}^{-\frac{1}{2}} P X^\top A^{-1}\right]^\top \right\| \\
&\qquad + \frac{1}{2} \left\| \left[B^{-1} Y Q^{-1} \widetilde{E}^{-\frac{1}{2}}\right]^\top \operatorname{diag}(1_m^\top A \widetilde{\Theta} B) \left[B^{-1} Y Q^{-1} \widetilde{E}^{-\frac{1}{2}}\right] \right\| && \text{(E.36)} \\
&\leqslant \frac{1}{2\underline{x}^2} \left\| \left[\widetilde{D}^{-\frac{1}{2}} P X^\top\right] \operatorname{diag}(\widetilde{\Theta} B 1_n) \left[\widetilde{D}^{-\frac{1}{2}} P X^\top\right]^\top \right\| \\
&\qquad + \frac{1}{2\underline{y}^2} \left\| \left[Y Q^{-1} \widetilde{E}^{-\frac{1}{2}}\right]^\top \operatorname{diag}(1_m^\top A \widetilde{\Theta}) \left[Y Q^{-1} \widetilde{E}^{-\frac{1}{2}}\right] \right\| && \text{(E.37)} \\
&\leqslant \frac{1}{2\underline{x}^2} \left\| \left[\widetilde{D}^{-\frac{1}{2}} P X^\top\right] \operatorname{diag}(\Theta B 1_n) \left[\widetilde{D}^{-\frac{1}{2}} P X^\top\right]^\top \right\| \\
&\qquad + \frac{1}{2\underline{y}^2} \left\| \left[Y Q^{-1} \widetilde{E}^{-\frac{1}{2}}\right]^\top \operatorname{diag}(1_m^\top A \Theta) \left[Y Q^{-1} \widetilde{E}^{-\frac{1}{2}}\right] \right\| && \text{(E.38)} \\
&= \frac{1}{2\underline{x}^2} \left\| \left[\widetilde{D}^{-\frac{1}{2}} P X^\top\right] \operatorname{diag}(q) \left[\widetilde{D}^{-\frac{1}{2}} P X^\top\right]^\top \right\| \\
&\qquad + \frac{1}{2\underline{y}^2} \left\| \left[Y Q^{-1} \widetilde{E}^{-\frac{1}{2}}\right]^\top \operatorname{diag}(\kappa) \left[Y Q^{-1} \widetilde{E}^{-\frac{1}{2}}\right] \right\| && \text{(E.39)} \\
&= \frac{1}{2\underline{x}^2} \left\| \widetilde{D}^{-\frac{1}{2}} P P^{-1} D P P^{-1} \widetilde{D}^{-\frac{1}{2}} \right\| + \frac{1}{2\underline{y}^2} \left\| \widetilde{E}^{-\frac{1}{2}} Y Q^{-1} Q E Q^{-1} Q \widetilde{E}^{-\frac{1}{2}} \right\| \\
&\leqslant \frac{1}{\underline{x}^2} + \frac{1}{\underline{y}^2} && \text{(E.40)}
\end{aligned}
$$

where at line (E.36), we have used Lemma E.5 and at line (E.38) we have used that $\operatorname{diag}(\widetilde{\Theta} B 1_n) \leqslant \operatorname{diag}(\Theta B 1_n)$ (i.e. $\operatorname{diag}(\Theta B 1_n) - \operatorname{diag}(\widetilde{\Theta} B 1_n)$ is positive semi-definite).

Now, using Lemma E.2 together with equation (E.40) above plugged into equation (E.35), we finally obtain that as long as $N \geqslant 2\log(\frac{2d}{\delta})[\gamma(d+3)+\gamma^2]$, we have with probability $\geqslant 1-\delta$:

$$
\begin{aligned}
&\mathbb{E}_{(i,j)\sim p} \left(l\left[(1 - C(S))[Z_*]_{i,j}, G_{(i,j)}\right] - l\left[[Z_*]_{i,j}, G_{(i,j)}\right] \right) \\
&\leqslant \ell C(S) \|\widetilde{M_*}\|_* \left\| \left[\widetilde{D}^{-\frac{1}{2}} P X^\top A^{-1}\right] A \widetilde{\Theta} B \left[B^{-1} Y Q^{-1} \widetilde{E}^{-\frac{1}{2}}\right] \right\| \\
&\leqslant \ell C(S) \|\widetilde{M_*}\|_* \left[\frac{1}{\underline{x}^2} + \frac{1}{\underline{y}^2}\right] \\
&\leqslant \ell \|\widetilde{M_*}\|_* \left[\frac{1}{\underline{x}^2} + \frac{1}{\underline{y}^2}\right] \sqrt{\frac{2\log(\frac{4d}{\delta})[\gamma(d+3)+\gamma^2]}{N}}, && \text{(E.41)}
\end{aligned}
$$

as expected. ☐

Lemma E.4. *For any $r > 0$ and $\delta \in (0,1)$, as long as $N \geqslant 8\gamma^2 + \gamma[8d+20][\log(2d) + \log(\frac{2}{\delta})]$, we have with probability $\geqslant 1-\delta$ over the draw of the training set:*

$$
\sup_{Z \in \widetilde{\mathcal{F}}_r} \left[\left| l(Z) - \hat{l}_S(Z) \right| \right] \leqslant \sup_{Z \in \widetilde{\mathcal{F}}_{4r\gamma^2}} \left[\left| l(Z) - \hat{l}_S(Z) \right| \right] \tag{E.42}
$$

Proof. This follows from Lemma E.1 upon noticing that if $\left\|\widetilde{D}^{\frac{1}{2}} P\widehat{P}^{-1}\breve{D}^{-\frac{1}{2}}\right\| \leqslant \sqrt{2}$, and $XMY^\top \in \breve{\mathcal{F}}_r$ and $\left\|\widetilde{E}^{\frac{1}{2}} Q\widehat{Q}^{-1}\breve{E}^{-\frac{1}{2}}\right\| \leqslant \sqrt{2}$, and $XMY^\top \in \breve{\mathcal{F}}_r$:

$$\|\widetilde{M}\| = \|\breve{D}^{\frac{1}{2}} \widehat{P} P^{-1} \widetilde{D}^{-\frac{1}{2}} \widetilde{M} \widetilde{E}^{-\frac{1}{2}} Q \widehat{Q}^{-1} \breve{E}^{\frac{1}{2}}\|_* \leqslant 2\|\breve{M}\|. \tag{E.43}$$

Using this and the fact that $\Gamma/\widehat{\Gamma} \leqslant \gamma$ yields the result immediately. □

Lemma E.5. *Let $U \in \mathbb{R}^{d_1\times m}, K \in \mathbb{R}^{m\times n}, V \in \mathbb{R}^{n\times d_2}$ be matrices and let 1_m (resp. 1_n) denote a column vector in $\mathbb{R}^m$ (resp. $\mathbb{R}^n$) all of whose entries are equal to 1.*

We have the following bound on the spectral norm of UKV:

$$\|UKV\| \leqslant \frac{1}{2}\left[\|U \operatorname{diag}(K1_{d_1})U^\top\| + \|V^\top \operatorname{diag}(1_{d_2}^\top K)V\|\right]. \tag{E.44}$$

Proof. The result essentially follows from the Cauchy-Schwarz inequality. Indeed, let $u \in \mathbb{R}^{d_1}$ and $v \in \mathbb{R}^{d_2}$ be two unit vectors. We have, using Cauchy-Schwarz at the second line:

$$\begin{aligned}
u^\top UKVv &= \sum_{i=1}^m \sum_{j=1}^n [u^\top U]_i K_{i,j} [Vv]_j \\
&\leqslant \sum_{i=1}^m \sum_{j=1}^n \frac{1}{2}\left[[u^\top U]_i^2 K_{i,j} + [Vv]_j^2 K_{k,j}\right] \\
&= \frac{1}{2} u^\top U \operatorname{diag}(K1_{d_1})U^\top u + \frac{1}{2} v^\top V^\top \operatorname{diag}(1_{d_2}^\top K)Vv \\
&\leqslant \frac{1}{2}\left[\|U \operatorname{diag}(K1_{d_1})U^\top\| + \|V^\top \operatorname{diag}(1_{d_2}^\top K)V\|\right].
\end{aligned} \tag{E.45}$$

Since u and v were arbitrary unit vectors, the result follows. □

F Low-level lemmas

Here collect Lemmas from the literature that are useful for our proofs. Sometimes we need to prove them purely to obtain explicit constants, but everything in this section is known.

Lemma F.1 (Non commutative Khinchine inequality [8, 9, 10])**.** *Let $X \in \mathbb{R}^{d\times d}$ be a matrix with jointly Gaussian, centred real-valued entries. There exists a universal constant C_k such that the following bound holds on the expectation of the spectral norm of X:*

$$E(\|X\|) \leqslant C_k\sqrt{\log(d)}\left[\|E(X^\top X)\|^{\frac{1}{2}} + \|E(XX^\top)\|^{\frac{1}{2}}\right] \tag{F.1}$$

Recall the following classic theorem [11, 12, 4]:

Theorem F.1. *Let $Z, Z_1, \ldots, Z_n$ be i.i.d. random variables taking values in a set $\mathcal{Z}$, and let $a < b$. Consider a set of functions $\mathcal{F} \in [a, b]^{\mathcal{Z}}$. $\forall \delta \in (0, 1)$, we have with probability $\geqslant 1 - \delta$ over the draw of the sample S that*

$$\forall f \in \mathcal{F}, \quad \mathbb{E}(f(Z)) \leqslant \frac{1}{n}\sum_{i=1}^n f(z_i) + 2\mathbb{E}_S(\mathfrak{R}_S(\mathcal{F})) + (b-a)\sqrt{\frac{\log(2/\delta)}{2n}}.$$

We also have that with probability $\geqslant 1 - \delta$, the following data-dependent bound holds:

$$\forall f \in \mathcal{F}, \quad \mathbb{E}(f(Z)) \leqslant \frac{1}{n}\sum_{i=1}^n f(z_i) + 2\,\mathfrak{R}_S(\mathcal{F}) + 3(b-a)\sqrt{\frac{\log(4/\delta)}{2n}}.$$

Proposition F.2 (Bernstein inequality, cf. [13], Corollary 2.11)**.** *Let $X_1, X_2, \ldots, X_N$ be independent real valued random variables with the following properties for some real numbers ν, M*

- *$X_i \leqslant M$ almost surely*

- $\sum_{i=1}^{N} \mathbb{E}(X_i^2) \leqslant \nu^2$.

Let $S = \sum_{i=1}^{N} X_i - \mathbb{E}(X_i)$*, we have (for all* $t \geqslant 0$*)*

$$\mathbb{P}(S \geqslant t) \leqslant \exp\left(-\frac{t^2/2}{\nu^2 + Mt/3}\right). \tag{F.2}$$

The inequality can be extended to the matrix-wise case as follows:

Proposition F.3 (Non commutative Bernstein inequality, Cf. [14])**.** *Let* $X_1, \dots, X_S$ *be independent, zero mean random matrices of dimension* $m \times n$*. For all* k*, assume* $\|X_k\| \leqslant M$ *almost surely, and denote* $\rho_k^2 = \max(\|\mathbb{E}(X_k X_k^\top)\|, \|\mathbb{E}(X_k^\top X_k)\|)$ *and* $\nu^2 = \sum_k \rho_k^2$*. For any* $\tau > 0$*,*

$$\mathbb{P}\left(\left\|\sum_{k=1}^{S} X_k\right\| \geqslant \tau\right) \leqslant (m+n)\exp\left(-\frac{\tau^2/2}{\sum_{k=1}^{S}\rho_k^2 + M\tau/3}\right). \tag{F.3}$$

Proposition F.4. *Under the assumptions of Proposition F.3 writing* $\sigma^2 = \sum_{k=1}^{S} \rho_k^2$*, we have*

$$\mathbb{E}\left(\left\|\sum_{k=1}^{S} X_k\right\|\right) \leqslant \sqrt{8/3}\sigma(1 + \sqrt{\log(m+n)}) + \frac{8M}{3}(1 + \log(m+n)). \tag{F.4}$$

Proof. The result in O notation is an exercise from [15], and a similar result is also mentioned in both [7] and [16].

For completeness and to get the exact constants, we include a proof as follows.

Let $Y = \left\|\sum_{k=1}^{S} X_k\right\|$. By Proposition F.3, splitting into two cases depending on whether $\tau M \leqslant \sigma^2$ or $\tau M \geqslant \sigma^2$ we have

$$\mathbb{P}(Y \geqslant \tau) \leqslant \min\left(1, (m+n)\exp\left[-\frac{3\tau^2}{8\sigma^2}\right]\right) + \min\left(1, (m+n)\exp\left[-\frac{3\tau}{8M}\right]\right) \tag{F.5}$$

Now note that writing κ for $\log(m+n)8M/3$, we have

$$\int_0^\infty 1 \wedge (m+n)\exp\left(-\frac{3\tau}{8M}\right) d\tau \tag{F.6}$$

$$\leqslant \int_0^\kappa 1 \wedge (m+n)\exp\left(-\frac{3\tau}{8M}\right) d\tau + \int_\kappa^\infty (m+n)\exp\left(-\frac{3\tau}{8M}\right) d\tau$$

$$\leqslant \kappa + \left[\frac{-8M}{3}(m+n)\exp\left(-\frac{3\tau}{8M}\right)\right]_\kappa^\infty = \kappa + \frac{8M(m+n)}{3}\exp\left(-\frac{3\kappa}{8M}\right)$$

$$= \kappa + \frac{8M(m+n)}{3} = \frac{8M}{3}(1 + \log(m+n)). \tag{F.7}$$

We also have, writing ψ for $\sigma\sqrt{\log(m+n)8/3}$,

$$\int_0^\infty 1 \wedge (m+n)\exp\left(-\frac{3\tau^2}{8\sigma^2}\right) d\tau \leqslant \int_0^\psi 1 d\tau + \int_\psi^\infty (m+n)\exp\left(-\frac{3\tau^2}{8\sigma^2}\right) d\tau$$

$$\leqslant \psi + \int_\psi^\infty \exp\left(-\frac{3(\tau^2 - \psi^2)}{8\sigma^2}\right) d\tau \leqslant \psi + \int_\psi^\infty \exp\left(-\frac{3(\tau - \psi)^2}{8\sigma^2}\right) d\tau$$

$$\leqslant \psi + \sigma\sqrt{2\pi/3} = \sigma\left[\sqrt{\log(m+n)8/3} + \sqrt{2\pi/3}\right] \leqslant \sqrt{8/3}\sigma(1 + \sqrt{\log(m+n)}). \tag{F.8}$$

Plugging inequalities (F.6) and (F.8) into equation (F.5), we obtain:

$$\mathbb{E}(Y) \leqslant \int_0^\infty \mathbb{P}(Y \geqslant \tau)d\tau \leqslant \sqrt{8/3}\sigma(1 + \sqrt{\log(m+n)})\frac{8M}{3}(1 + \log(m+n)), \tag{F.9}$$

as expected. ∎

Lemma F.5. *Let F be a random variable that depends only on the draw of the training set. Assume that with probability $\geqslant 1-\delta$,*

$$\mathbb{E}(F) \leqslant f(\delta), \tag{F.10}$$

for some given monotone increasing function f. Then we have, in expectation over the training set:

$$\mathbb{E}(F) \leqslant \sum_{i=1}^{\infty} f(2^{-i})2^{1-i}, \tag{F.11}$$

In particular, if $f(\delta) = C_1\sqrt{\log(\frac{1}{\delta})} + C_2$, then we have in expectation over the draw of the training set:

$$\mathbb{E}(F) \leqslant \frac{C_1}{\sqrt{2}-1} + C_2. \tag{F.12}$$

Proof. By assumption we have for any δ:

$$\mathbb{P}\left(X \geqslant f(\delta)\right) \leqslant \delta \tag{F.13}$$

Let us write A_i for the event $A_i = \{F \leqslant f(\delta_i)\}$ where we set $\delta_i = 2^{-i}$ for $i = 1, 2, \ldots$. We also set $\widetilde{A}_i = A_i \backslash A_{i-1}$ for $i = 1, 2, \ldots$ with the convention that $A_0 = \varnothing$ so that $\tilde{A}_1 = A_1$.

We have, for $i \geqslant 2$, $\mathbb{P}(\widetilde{A}_i) \leqslant \mathbb{P}(A^c_{i-1}) \leqslant \delta_{i-1}$, and for $i = 1$, $\mathbb{P}(\widetilde{A}_1) \leqslant 1 = \delta_{i-1}$. Thus we can write

$$\mathbb{E}(F) \leqslant \sum_{i=1}^{\infty} \mathbb{E}(X|\widetilde{A}_i)\mathbb{P}(\widetilde{A}_i) \leqslant \sum_{i=1}^{\infty} \mathbb{E}(X|\widetilde{A}_i)\delta_{i-1} \leqslant \sum_{i=1}^{\infty} f(\delta_i)\delta_{i-1}, \tag{F.14}$$

yielding identity (F.11) as expected.

Next, assuming $f(\delta) = C_1\sqrt{\log(\frac{1}{\delta})} + C_2$, we can continue as follows:

$$\mathbb{E}(F - C_2) \leqslant \sum_{i=1}^{\infty} f(\delta_i)\delta_{i-1} \leqslant \sum_{i=1}^{\infty} [C_1\sqrt{\log(2^i)}]2^{1-i} \tag{F.15}$$

$$\leqslant \sum_{i=1}^{\infty} [C_1\sqrt{i}]2^{1-i} \leqslant C_1 \sum_{i=1}^{\infty} \sqrt{2}^{1-i} = \frac{C_1}{\sqrt{2}-1} \tag{F.16}$$

where at the second line we have used the fact that for any natural number i, $\sqrt{i} \leqslant \sqrt{2}^{i-1}$. □

As an immediate consequence we obtain the following Rademacher type theorem in expectation:

Theorem F.2. *Let $Z, Z_1, \ldots, Z_N$ be i.i.d. random variables taking values in a set $\mathcal{Z}$, and let $a < b$. Consider a set of functions $\mathcal{F} \in [a,b]^{\mathcal{Z}}$. $\forall \delta \in (0,1)$, we have in expectation over the draw of the sample S that*

$$\inf_{f\in\mathcal{F}} \left(\mathbb{E}(f(Z)) - \frac{1}{N}\sum_{i=1}^{n} f(z_i)\right) \leqslant 2\mathbb{E}(\mathfrak{R}_S(\mathcal{F})) + 5(b-a)\sqrt{\frac{1}{N}}. \tag{F.17}$$

Proposition F.6 ([17, 18]). *Let $\mathcal{F}$ be a real-valued function class taking values in $[0,1]$, and assume that $0 \in \mathcal{F}$. Let S be a finite sample of size n. For any $2 \leqslant p \leqslant \infty$, we have the following relationship between the Rademacher complexity $\mathfrak{R}(\mathcal{F}|_S)$ and the covering number $\mathcal{N}(\mathcal{F}|S, \epsilon, \|\cdot\|_p)$.*

$$\mathfrak{R}(\mathcal{F}|_S) \leqslant \inf_{\alpha>0}\left(4\alpha + \frac{12}{\sqrt{n}}\int_{\alpha}^{1}\sqrt{\log\mathcal{N}(\mathcal{F}|S, \epsilon, \|\cdot\|_p)}d\epsilon\right),$$

where the norm $\|\cdot\|_p$ on $\mathbb{R}^m$ is defined by $\|x\|_p^p = \frac{1}{n}(\sum_{i=1}^m |x_i|^p)$.

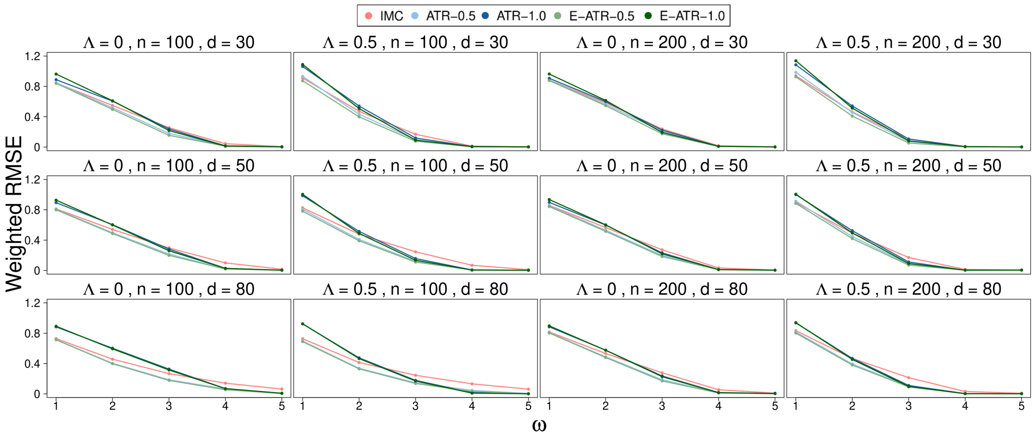

Figure G.1: Weighted RMSE as a function of ω .

G More detailed discussion of the experimental setting

G.1 Synthetic data

Generation and training procedure: First we sample matrices A and B in $\mathbb{R}^{m\times d}$ with i.i.d. $N(0,1)$ entries. We also sample K_1 and K_2 in $\mathbb{R}^{d\times r}$. We then compute $F = AK_1K_2^\top B^\top$ and set $G = m\,\text{normalize}(F)$, $X = \sqrt{md}\,\text{normalize}(A)$ and $Y = \sqrt{md}\,\text{normalize}(B)$ where the operator normalize normalises the matrix to have unit Frobenius norm. Regarding the sampling distribution, we set $p_{i,j} \propto \exp(\Lambda|G_{i,j}|)$ where Λ is a hyperparameter. In particular, when $\Lambda = 0$ we have uniform sampling. 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 $\omega \in \{1, 2, 3, 4, 5\}$. Each (n, d, r) configuration was tested on 50 matrices. Training details: the λs were chosen in the range $[10^{-6}, 2\times 10^2]$, each configuration was run to convergence without warm starts.

More detailed results: Below are detailed results of the syntehtic data experiments. The first graph G.1 shows the performance as a function of our data sparsity paramameter ω in different configurations, whilst Figure G.3 provides the corresponding boxplots documenting the variance with respect to the draw of the random matrix. Figure G.2 shows, in many different situations, the progression of performance as the size of the side information increases. Corresponding boxplots are provided in Figure G.4.

We observe that our methods (especially the smoothed version) generally outperform standard IMC in the meaningful sparsity regimes. Interestingly, when data is too sparse to make any meaningful prediction, standard IMC frequently outperforms our method (though our methods become better as more data becomes available), suggesting that α could be tuned depending on the sparsity of the observations.

G.2 Description of real-life datasets

- **Douban**[2] ($R \in \mathbb{R}^{4999\times 4577}$): Douban is a social network where users can produce content related to movies, music, and events. Douban users are members of the social network and Douban items are a subset of popular movies. The rating range is $\{1, 2, \ldots, 5\}$ and the entry (i, j) corresponds the rating of user i to movie j. To construct side information, we collected the following data from the Douban website: each movies' genres, its number of views, the number of people who rated the movie, and the number of reviews written.
- **LastFM** ($R \in \mathbb{R}^{1875\times 4354}$): Last.fm is a British music website that builds a detailed profile of each user's musical taste. Differently from the other datasets an entry (i, j) represents the number of views of user i to band/artist j. We expressed the number of views in a log scale.

[2]Rating matrix available in https://doi.org/10.7910/DVN/JGH1HA

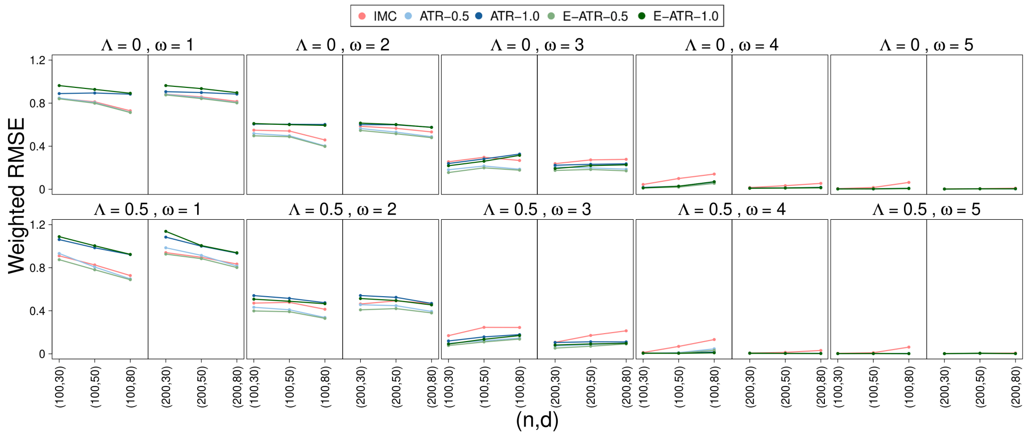

Figure G.2: Weighted RMSE as a function of the size of the side information.

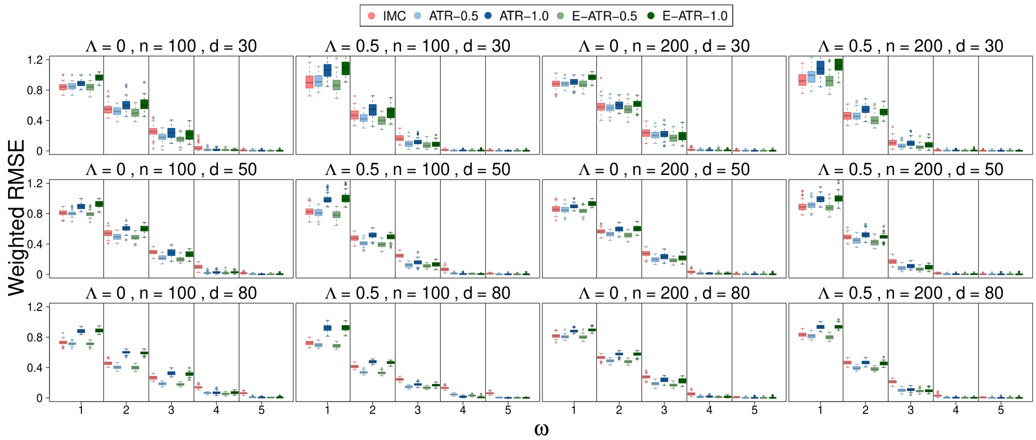

Figure G.3: Weighted RMSE as a function of ω, boxplots.

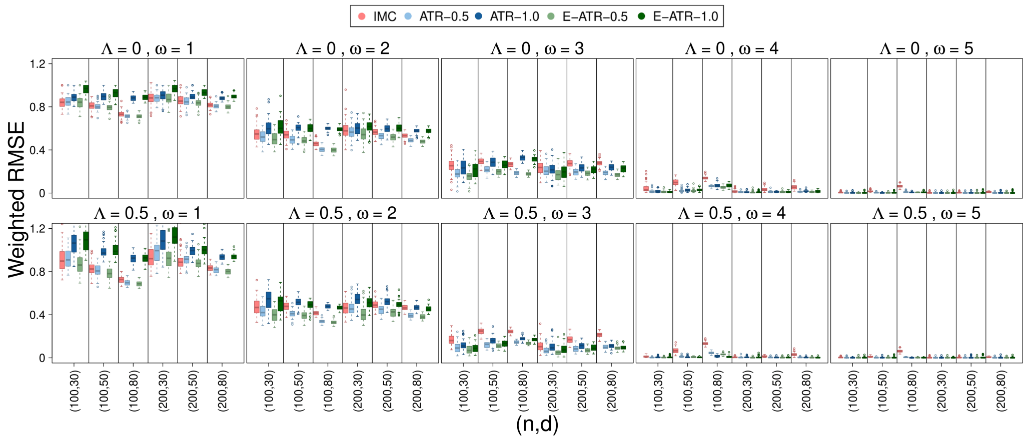

Figure G.4: Weighted RMSE as a function of the size of the side information, boxplots.

The website allows users to tag artists, which provides us with the opportunity to group the items (artists) by their associated tags.

- **MovieLens** ($R \in \mathbb{R}^{6040 \times 3382}$): We consider the MovieLens 1M dataset, which is a broadly used and stable benchmark dataset. MovieLens is a non-commercial website for movie recommendations. Just as in Douban, an entry (i, j) represents the rate of user i to movie j on a scale from 1 to 5. We used movies' genres and gender as item and user side information respectively.

Training details: In all real data experiments, we used 85% of the data for training, 10% for validation and 5% for the test set.

We optimized the model (18) via the accelerated subgradient method of [19], alternating the optimization between each term with only two iterations per term.

To choose a suitable hyper parameter range, the matrices $\check{X}$ and $\check{Y}$ were normalised to have Frobenius norm $\sqrt{m}$ and $\sqrt{n}$ respectively, and values in the range $[1, 200]$ were explored for both λ_1, λ_2. Initially, twenty alternations were run for each tested hyper parameter combination. We then ran the model to convergence for the final hyperparameter configuration. For the real data experiments, we used a rank-restricted version of the SVD's with rank 30.

We performed the experiments in a cluster with 72 CPUs (3GHz) and 750GB of RAM. We relied on warm starts to reach convergence faster. For a given X, Y, and given a solution $Z_1 + Z_2$ (with Z_1 (resp. Z_2) corresponding to the inductive (resp. non inductive) term), a warm start $XM_0Y^\top + Z_0$ can be constructed as follows: Set $Z_0 = Z_2$. Set $M_0 = (X^\top X)^{-1} X^\top Z_1 Y (Y^\top Y)^{-1}$. If X or Y is not full rank the above inverses can be replaces by pseudoinverses.

H Variations on the optimization problems and loss functions

Models involving a non-inductive term We first note that using the subadditivity of the Rademacher complexity, it is trivial to obtain results for a combined function class corresponding to the regulariser (18):

Proposition H.1. *Suppose for simplicity that $m = n$, $d_1 = d_2 = d$, and $\frac{\mathbf{x}^2\mathbf{y}^2}{\underline{x}^2\underline{y}^2} = \gamma \leqslant K$ for some constant $K = O(1)$ and define the function class $\widetilde{\mathcal{G}}_{r_1,r_2} := \{XMY^\top + Z : \|\widetilde{D}^{\frac{1}{2}} PMQ^{-1} \widetilde{E}^{\frac{1}{2}}\|_* \leqslant \Gamma\sqrt{r_1} \quad \wedge \|\widetilde{D}_I^{\frac{1}{2}} Z \widetilde{E}_I^{\frac{1}{2}}\|_* \leqslant \sqrt{r_2}\}$. As long as $N \geqslant T$ where T is $O(n)$, w.p. $\geqslant 1 - \delta$ we have for all $F \in \widetilde{\mathcal{G}}_{r_1,r_2}$:*

$$l(F) - \hat{l}_S(F) \leqslant \widetilde{O}\left((\ell + b)\frac{\sqrt{\Gamma r_1 d} + \sqrt{r_2 n}}{\sqrt{N}}\right). \tag{H.1}$$

Proof. Follows from the Rademacher complexity bound from Proposition 3.3 (cf. also Prop.B.1) applied to both side information pairs (X, Y) and (I, I), together with the subadditivity of the Rademacher complexity. Note that the condition on n is only necessary to get rid of $O(1/N)$ terms for cosmetic purposes. □

Lagrangian Formulation and Square Loss

Similarly to other work ([20, 2] etc.) we expressed our results in terms of bounds on the expected loss of the empirical risk minimizers subject to *explicit norm constraints*. However, it is easy to express similar results for the solution to a regularised optimization problem in "Lagrangian formulation" such as the ones we propose[3]. We have also relied on a *bounded* loss function. However, in most practical situations, the values of the entries are restricted by domain knowledge (for instance, in the Recommender Systems field, ratings are typically restricted to the range $[1, 5]$). This effectively renders any Lipschitz loss bounded, including the square loss, as long as one also truncates the output of the algorithm to fit the required range.

We begin by completing the (trivial) proof of Corollary 3.4.

[3] just as in the case of exact norm constraints, the hyperparameters must be assumed to have been properly tuned

Proof of Corollary 3.4. Since G satisfies the optimization constraints on the training set Ω, we must have $\|\widetilde{D}^{\frac{1}{2}} P M_{\#} Q^{-1} \widetilde{E}^{\frac{1}{2}}\|_* \leqslant \sqrt{r_G}\Gamma$, which allows us to apply proposition (C.1) to the loss function $\Phi_{2C} \circ l$, which coincides with l when applied to the matrix $\Phi_C(Z_{\#}) - G$. $\square$

We now show further how to adapt our result C.1 to make it apply to the solution to a Lagrangian formulation involving the square loss. Further similar manipulations can be applied to our other results.

Proposition H.2. *Assume that the noise ζ is bounded by a fixed constant C almost surely, and that so are all of the entries of the ground truth matrix G. Let*

$$M_\lambda = \arg\min_M \frac{1}{N} \sum_{(i,j)\in\Omega} [XMY^\top - G_{i,j} - \zeta_{i,j}]^2_{i,j} + \lambda \|\widetilde{D}^{\frac{1}{2}} P M Q^{-1} \widetilde{E}^{\frac{1}{2}}\|_*,$$

and $Z_\lambda := XM_\lambda Y^\top$ denote the solutions to a Lagrangian formulation of the problem with the square loss l (which is unbounded).

Furthermore, we also write $\Phi(x) = \Phi_{2C}(x) = \operatorname{sign}(x)\min(|x|, 2C)$, $\mathcal{E} = l(G)$ for the expected square loss at the ground truth (i.e. the variance of the noise) and $\Delta := C^2\sqrt{\frac{\log(4/\delta)}{2N}}$. We assume that λ is tuned so that $\frac{\mathcal{E}+\Delta}{2\sqrt{r_G}\Gamma} \leqslant \lambda \leqslant 2\frac{\mathcal{E}+\Delta}{\sqrt{r_G}\Gamma}$[4].

We have the following bound on the expected L^2 risk of $\Phi_{2C}(XM_\lambda Y^\top)$:

$$\mathbb{E}_\xi\left(\left|\Phi_{2C}[XM_\lambda Y^\top]_\xi - G_\xi - \zeta_\xi\right|^2\right) = l(\Phi_{2C}(XM_\lambda Y^\top)) \tag{H.2}$$
$$\leqslant 3\mathcal{E} + \frac{48\ell\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} + \frac{72\ell\mathbf{xy}\sqrt{d_1 d_2 r}(1+\log(2d))}{N} + 19C^2\sqrt{\frac{\log(4/\delta)}{2N}}.$$

Proof. Since $l(\zeta) \leqslant C^2$ for any $|\zeta| \leqslant C$ we have by Hoeffding's lemma that with probability $\geqslant 1-\delta/2$,

$$|l_S(G) - l(G)| \leqslant C^2\sqrt{\frac{\log(4/\delta)}{2N}}. \tag{H.3}$$

Then (with the same probability) we have

$$\begin{aligned} l_S(Z_\lambda) + \lambda\|\widetilde{D}^{\frac{1}{2}} P M_\lambda Q^{-1} \widetilde{E}^{\frac{1}{2}}\|_* &\leqslant l_S(G) + \lambda\sqrt{r_G}\Gamma \\ &\leqslant l(G) + \lambda\sqrt{r_G}\Gamma + C^2\sqrt{\frac{\log(4/\delta)}{2N}}. \\ &= l(G) + \Delta + \lambda\sqrt{r_G}\Gamma \\ &\leqslant 3[\mathcal{E}+\Delta], \end{aligned} \tag{H.4}$$

where at the last line, we have used the constraint on λ.

It follows that

$$\begin{aligned} \|\widetilde{D}^{\frac{1}{2}} P M_\lambda Q^{-1} \widetilde{E}^{\frac{1}{2}}\|_* &\leqslant \frac{3[\mathcal{E}+\Delta]}{\lambda} \\ &\leqslant 6\sqrt{r_G}\Gamma, \end{aligned} \tag{H.5}$$

where we have made another use of the constraint on λ.

It follows that $Z_\lambda \in \widetilde{\mathcal{F}}_{36r_G}$. Let $\tilde{l} = \Phi_{4C} \circ l$ be the truncated square loss: $\tilde{l}(a,b) = \min(|a-b|, 4C)^2$. By Proposition C.1 we now have with probability $\geqslant 1 - \frac{\delta}{2}$ over the draw of the training set:

$$\mathbb{E}\left[\tilde{l}((XM_\lambda Y^\top)_\xi, G_\xi + \zeta_\xi)\right] - \frac{1}{N}\sum_{\xi\in\Omega} \tilde{l}((XM_\lambda Y^\top)_\xi, G_\xi + \zeta_\xi) \tag{H.6}$$

[4]Although this tuning depends on the sample size N slightly, it converges as N tends to infinity and is there for purely cosmetic purposes (to avoid extra logarithmic terms in the final formula).

$$\leqslant \frac{48\ell\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} + \frac{72\ell\mathbf{x}\mathbf{y}\sqrt{d_1 d_2 r}(1+\log(2d))}{N} + 16C^2\sqrt{\frac{\log(4/\delta)}{2N}}.$$

Writing $\bar{\Delta}$ for the quantity

$$\frac{48\ell\sqrt{\Gamma}\sqrt{r}\sqrt{d}(1+\sqrt{\log(2d)})}{\sqrt{N}} + \frac{72\ell\mathbf{x}\mathbf{y}\sqrt{d_1 d_2 r}(1+\log(2d))}{N} + 16C^2\sqrt{\frac{\log(4/\delta)}{2N}},$$

it now follows that (w.p. $\geqslant 1-\delta$)

$$\begin{aligned}
&l(\Phi_{2C}(XM_\lambda Y^\top)) = \mathbb{E}_\xi\left[l((\Phi_{2C}(XM_\lambda Y^\top)_\xi, G_\xi+\zeta_\xi)\right] \\
&= \mathbb{E}_\xi\left[\tilde{l}((XM_\lambda Y^\top)_\xi, G_\xi+\zeta_\xi)\right] = \tilde{l}((XM_\lambda Y^\top)) \\
&\leqslant \tilde{l}_S(XM_\lambda Y^\top) + \bar{\Delta} && \text{(H.7)} \\
&\leqslant l_S(XM_\lambda Y^\top) + \bar{\Delta} \leqslant 3[\mathcal{E}+\Delta] + \bar{\Delta}, && \text{(H.8)}
\end{aligned}$$

where at equation (H.7) we have used equation (H.6) and at equation (H.8) we have used equation (H.4). The result follows.

□

I Further discussion

I.1 Deeper comparison to related works

Here we discuss some related works in more detail than in the main paper.

One very interesting other work is [21] which introduces a joint model that imposes a nulcear norm based constraint on both M and $XMY^\top$ through a modification of the objective: first, the matrices X and Y are augmented by columns of ones resulting in the matrices $\bar{X} = [X, \mathbf{1}]$ and $\bar{Y} = [Y, \mathbf{1}]$. Predictors then take the form $E = \bar{X}M(\bar{Y})^\top + \Delta$, with nuclear norm regularisation imposed on both E and M, and Frobenius norm regularization imposed on Δ, with the constraint that $P_\Omega(E) = R_\Omega$ where R_Ω denotes the observed entries. Thus the model achieves a similar aim as [20] through a different and more original approach. The authors then provide an efficient algorithm for their model and prove some theoretical guarantees: for exact recovery, they obtain a rate of $O(rd\log(d)\log(n))$ in the uniform sampling case. This is the same as [19], except that the assumptions on X and Y are weaker (no orthogonality assumption). Of course, both [21] and [19] require a realisability assumption for exact recovery to be possible. In addition to that, the authors of [21] also show distribution-free bounds for the approximate recovery case which scale as $O(\gamma^2\log(n))$ where γ is an upper bound on the ground truth spectral norm of the matrix M (G in their notation). That bound is comparable to the bounds of the form (3) from [20, 22, 23], though the precise results are different in formulation (and rely on a different optimizer). Note that in addition to pertaining to a completely different optimization problem, our results for approximate recovery lack any dependence on n, even logarithmic, and also do not have the implicit dependence on $d_1 d_2$ present in that paper. Note that although it is claimed in the paper that the rate is "$\log(n)$", this is because in that informal presentation of the results the authors are treating their "γ" (which scales at least as $\sqrt{d_1 d_2 r}$) as a constant, which amounts to treating the size of the side information as a constant. This type of formulation is standard and also used in [19], but corresponds to a different perspective as in this work we want to remove the dependence on d_1, d_2. Note also that although it is not explicitly stated in the paper that the exact recovery results rely on a uniform sampling assumption, such an assumption is implicit. Indeed, such an assumption is standard in all exact recovery results: there is no known exact recovery result for arbitrary distributions for either MC or IMC. Further, the results would be clearly wrong without such an assumption (assume for instance identity side information and a sampling distribution which only samples the top left quadrant, all of which is perfectly compatible with the coherence assumptions on X, Y and the ground truth matrix G (F in their notation)). The first obvious implicit use of the uniform sampling assumption is in line 70 of the supplementary material. As we explain later, even defining the concept of exact recovery in the non uniform sampling case has not been done explicitly to the best of our knowledge, and no results exist for this for either inductive matrix completion or matrix completion in general.

In [24], the authors explicitly study a disentangled version of [20] specifically tailored to the case of community side information. Whilst generalisation bounds are provided which scale similarly to ours in the case of community side information, those are obtained through a direct application of the matrix completion results from [2] to the auxiliary problem where each community is treated as a single user. In particular, the results are not applicable in a more general context and they did not introduce any of the novel proof techniques we rely on here.

[25] proves rates of $d^2r^3\log(d)$ in the case of *exact recovery*, as well as abstract conditions for the possibility of exact recovery in a more general context and results for other problems closely related to inductive matrix completion (such as matrix regression, see also [26, 27]); [28], which proved a similar sample complexity rate together with an efficient optimization strategy with favourable convergence rates; and of course [19], which *both* introduced the MaxIDE algorithm (an involved form of projected gradient method with an integrated line search over the step sizes) to solve problem (2), *and* proved sample complexity bounds of order $rd\log(d)\log(n)$ for exact (noiseless) recovery under the assumption of uniform sampling. Recently, convergence and generalisation guarantees were shown for an exciting model which functions as inductive matrix completion with unknown "side information matrices X, Y which must be learned by a two layer neural network from some raw user and item side information, jointly with the low rank problem [29]. We note that this applies to a fixed rank problem and does not rely on a nuclear norm regulariser.

Further remarks on related works: In Table 1 and Table 2, we are only concerned with sample complexity. It is worth noting that many important gains were also achieved in the direction of improving computational complexity through better algorithms [28, 30].

We also do not compare here with results obtained for other regularisation strategies including the max norm [31, 32, 33] etc., all of which apply exclusively to matrix completion without side information. We do note in passing that rates of $O(nr\log(n))$ were obtained very early for matrix completion with an *explicit low-rank assumption* [31]. In both MC and IMC, the relevance of the more recent branch of the literature is tied to the impractical nature of explicitly minimizing the rank and the fact that the low rank assumption is not satisfied *exactly*, justifying the use of nuclear norm based methods and the soft relaxations of the rank that they bring into the theoretical analysis.

J Discussion and future directions

J.1 On transductive Rademacher complexity:

Some results in [2] and [7] are formulated in the transductive [34] setting. In this context, we assume that the set of observed entries is sampled *without replacement*, and the training and test sets are divided uniformly. There is a parallel theory in this case with a concept of transductive Rademacher complexity at the key. In some cases the bounds can be better in some aspects. For instance, the transductive bound in [2] scales like $O(nr\log(n))$ in the case of a distribution where the probabilities of each entries are within a ratio of each other. Such a bound follows in our iid setting from Proposition 3.3, and indeed similar results had been otherwise obtained (for the non inductive case) in [31], as the authors of [2] mention. As another significant advantage, the *transductive* bounds in [7] involve a smaller power of the log term.

There are two reasons why we didn't prove transductive bounds in our setting: (1) The transductive Rademacher complexity is bounded above by the standard Rademacher complexity up to a constant of 4 [5]. *In particular, all of our results also hold up to a constant in a transductive setting.* [6]. (2) Contrary to the MC case, we do not believe that we would get better bounds in this context. Indeed, the main reason the transductive setting improves the bounds is because it prevents the oversampling of single entries (see how in the proof of the main theorem in [2], one must distinguish between the oversampled entries and the moderately sampled entries). It is easy to see by comparing to our proof of Theorem 3.1, especially consolidating the intuition via the example of community side information, that the benefits would *not* carry over to the inductive case: even if the *entries* are sampled without replacement, the *combinations of communities* can still be sampled many times. Thus we do not expect significant gains from this approach.

[5] See Footnote 1 on page 3407 of [2], and Lemma 1 in [34]

[6] This remark also applies to earlier work, they merely proved the transductive bounds because in the matrix case, this provides an actual improvement.

J.2 Open directions

There are many possible open problems related to this work and to distribution-free matrix completion in general:

- Is it possible to provide a rigorous theoretical explanation why the empirically weighted trace norm outperforms the exactly weighted version in the synthetic data experiments?
- Can we make the bounds even more sensitive to the alignement of the side information vectors?
- In what situations can one remove the $\sqrt{\log(d)}$ term in Proposition 3.1?

Regarding the extra log term in Theorem 3.1, we would like to note that although we do not see how to remove it in general, it is straightforward to remove it (at the cost of higher order dependence on the coherence of X and Y) in the specific case where the columns of X and Y each have distinct support (i.e. the columns of X^2 and Y^2, defined as matrices whose entries are the squares of those of X and Y respectively, are orthogonal), in which particular case a proof with more similarities to that in [2] still holds.

K Table of notations

Table K.1: Table of notations for quick reference

Notation	Meaning
$\|A\|$	spectral norm of matrix A
$A \preccurlyeq B$	$B - A$ is positive semi-definite
$\|A\|_*$	nuclear norm of matrix A
I	Identity matrix
$G \in \mathbb{R}^{m \times n}$	ground truth matrix
$\xi_1, \ldots, \xi_N$ $(\in \{1, \ldots, m\} \times \{1, \ldots, n\})$	sampled entries
ζ_ξ	Noise observed at sample ξ
$X \in \mathbb{R}^{m \times d}$ (resp. $Y \in \mathbb{R}^{n \times d}$)	Row (resp. column) side information matrix
M	matrix to optimize (predictors: $XMY^\top$)
$S = \Omega = \{\xi_1, \ldots, \xi_N\}$	(training) set of observed entries
$x_i = X_{i,\cdot}$	side information vector for ith user (row)
$y_j = X_{j,\cdot}$	side information vector for jth item (column)
$\mathbf{x}$ (resp. $\mathbf{y}$)	$\max_i \|x_i\|^2$ (resp. $\max_j \|x_j\|^2$)
$\underline{x}$ (resp. $\underline{y}$)	$\min_i \|x_i\|^2$ (resp. $\min_j \|x_j\|^2$)
γ	$\frac{\mathbf{x}^2 \mathbf{y}^2}{\underline{x}^2 \underline{y}^2}$
d	$\max(d_1, d_2)$
$p_{i,j}$	Probability of sampling (i, j) $=\mathbb{P}(\xi = (i, j))$
p	sampling distribution
$\mathcal{M}$	constraint on $\|M\|_*$
$h_{i,j} = \sum_{\xi \in \Omega} 1_{\xi = (i,j)}$	Number of times entry (i, j) was sampled
l	loss function
b	global upper bound on l
ℓ	Lipschitz constant of l
$l(Z)$	$\mathbb{E}_{(i,j) \sim p}(l([XMY^\top]_{i,j}, G_{i,j} + \zeta_{i,j}))$
(or more rigorously)	$\mathbb{E}_{\xi, \bar{\xi}} l([XMY^\top]_{\xi_1, \xi_2}, \bar{\xi}_o)$
$\hat{l}(Z)$	$\frac{1}{N} \sum_{(i,j) \in \Omega} l([XMY^\top]_{i,j}, G_{i,j} + \zeta_{i,j})$
(or more rigorously)	$\frac{1}{N} \sum_{o=1}^{N} l([XMY^\top]_{\xi_1, \xi_2}, \bar{\xi}_o)$
Γ	$\sum_{i,j} p_{i,j} \|x_i\|^2 \|y_j\|^2$
$\widehat{\Gamma}$	$\frac{1}{N} \sum_{i,j} h_{i,j} \|x_i\|^2 \|y_j\|^2$
q_i (resp. $\hat{q}_i$)	$\sum_{j=1}^{n} p_{i,j} \|y_j\|^2$ (resp. $\frac{1}{N} \sum_{j=1}^{n} h_{i,j} \|y_j\|^2$)
κ_j (resp. $\hat{q}_i$)	$\sum_{i=1}^{m} p_{i,j} \|x_i\|^2$ (resp. $\frac{1}{N} \sum_{i=1}^{m} h_{i,j} \|x_i\|^2$)
$\langle v, w \rangle_l$ (resp. $\langle v, w \rangle_r$)	$\sum_{i=1}^{m} v_i q_i w_i$ (resp. $\sum_{j=1}^{n} v_j h_j w_j$)
$\langle v, w \rangle_{\hat{l}}$ (resp. $\langle v, w \rangle_{\hat{r}}$)	$\sum_{i=1}^{m} v_i \hat{q}_i w_i \sum_{j=1}^{n} v_j \hat{\kappa}_j w_j$
L	$X^\top \operatorname{diag}(q) X = \sum_{i,j} p_{i,j} x_i x_i^\top \|y_j\|^2$
$\widehat{L}$	$X^\top \operatorname{diag}(\hat{q}) X = \sum_{i,j} \frac{h_{i,j}}{N} x_i x_i^\top \|y_j\|^2$
R	$Y^\top \operatorname{diag}(\kappa) Y = \sum_{i,j} p_{i,j} y_j y_j^\top \|x_i\|^2$
$\widehat{R}$	$Y^\top \operatorname{diag}(\hat{\kappa}) Y = \sum_{i,j} \frac{h_{i,j}}{N} y_j y_j^\top \|x_i\|^2$
D (resp. $\widehat{D}$)	Eigenvalues of L (resp. $\widehat{L}$)
E (resp. $\widehat{E}$)	Eigenvalues of R (resp. $\widehat{R}$)
P	orth. matrix diagonalising L so $L = P^{-1} D P$
Q	orth. matrix diagonalising R so $R = Q^{-1} E Q$
$\widetilde{D}$	$\alpha D + (1 - \alpha) \frac{\Gamma}{d_1} I$ (In theorems, $\alpha = \frac{1}{2}$)
$\widetilde{E}$	$\alpha E + (1 - \alpha) \frac{\Gamma}{d_2} I$
$\breve{D}$	$\alpha \widehat{D} + (1 - \alpha) \frac{\Gamma}{d_1} I$
$\breve{E}$	$\alpha \widehat{E} + (1 - \alpha) \frac{\Gamma}{d_2} I$

$\widetilde{X}$ (resp. $\widetilde{Y}$)	$XP^{-1}\widetilde{D}^{-\frac{1}{2}}$ (resp. $YQ^{-1}\widetilde{E}^{-\frac{1}{2}}$)
Y' (resp. Y')	$XP^{-1}D^{-\frac{1}{2}}$ (resp. $YQ^{-1}E^{-\frac{1}{2}}$)
$\widehat{X}$ (resp. $\widehat{Y}$)	$X\widehat{P}^{-1}\widehat{D}^{-\frac{1}{2}}$ (resp. $Y\widehat{Q}^{-1}\widehat{E}^{-\frac{1}{2}}$)
$\check{X}$ (resp. $\check{Y}$)	$X\widehat{P}^{-1}\check{D}^{-\frac{1}{2}}$ (resp. $Y\widehat{Q}^{-1}\check{E}^{-\frac{1}{2}}$)
M'	$D^{\frac{1}{2}}PMQ^{-1}E^{\frac{1}{2}}$
$\widehat{M}$	$\widehat{D}^{\frac{1}{2}}\widehat{P}M\widehat{Q}^{-1}\widehat{E}^{\frac{1}{2}}$
$\widetilde{M}$	$\widetilde{D}^{\frac{1}{2}}PMQ^{-1}\widetilde{E}^{\frac{1}{2}}$
$\breve{M}$	$\check{D}^{\frac{1}{2}}\widehat{P}M\widehat{Q}^{-1}\check{E}^{\frac{1}{2}}$
$\sigma^1 \in \mathbb{R}^{d_1}$ (resp. $\sigma^2 \in \mathbb{R}^{d_1}$)	singular values of X (resp. Y) wrt $\langle \cdot, \cdot \rangle_l$ (resp. $\langle \cdot, \cdot \rangle_r$)
equivalently:	$\sigma^1_u = \sqrt{D_{u,u}}$ ($\sigma^2_v = \sqrt{D_{v,v}}$) for all $u \leqslant d_1$ (resp. $v \leqslant d_2$)
σ^1_* (resp. σ^2_*)	$\max(\sigma^1)$ (resp. $\max(\sigma^2)$)
$c_U(i)$ (resp. $c_I(j)$))	community to which user i (resp. item j) belongs
$\check{D}_I$ (resp. $\check{E}_I$)	same as $\check{D}$ (resp. $\check{E}$)
	(with identity side info)
Hence: $[\check{D}_I]_{i,i} =$	$\alpha[\sum_{j=1}^{n} \frac{h_{i,j}}{N}] + (1-\alpha)\frac{1}{d_1}$
and: $[\check{E}_I]_{j,j} =$	$\alpha[\sum_{i=1}^{m} \frac{h_{i,j}}{N}] + (1-\alpha)\frac{1}{d_2}$
$\widetilde{\mathcal{F}}_r$	$\left\{XMY^\top : \lVert\widetilde{M}\rVert_* \leqslant \sqrt{r}\Gamma\right\}$
$\check{\mathcal{F}}_r$	$\left\{XMY^\top : \lVert\breve{M}\rVert_* \leqslant \sqrt{r}\widehat{\Gamma}\right\}$
$\check{Z}_*$	$\arg\min_{Z \in \check{\mathcal{F}}_r} \mathbb{E}l(Z_\xi, G_\xi + \zeta_\xi)$
$\check{Z}_S$	$\arg\min(\hat{l}_S(Z) : Z \in \check{\mathcal{F}}_r)$
$\widetilde{Z}_*$	$\arg\min_{Z \in \widetilde{\mathcal{F}}_r} \mathbb{E}l(Z_\xi, G_\xi + \zeta_\xi)$
$\widetilde{Z}_S$	$\arg\min_{Z \in \widetilde{\mathcal{F}}_r} \mathbb{E}l_S(Z)$
If $G \in \widetilde{F}_r$	$G = \widetilde{Z}_*$
If $G \in \check{F}_r$	$G = \check{Z}_*$
$\mathcal{E}$	$l(G) = \mathbb{E}_{\xi \sim p} l((XM_SY^\top)_\xi, G_\xi + \zeta_\xi)$
$\widetilde{\mathcal{G}}_{r_1,r_2}$	$\Big\{XMY^\top + Z$ s.t.
	$\lVert\widetilde{D}^{\frac{1}{2}}PMQ^{-1}\widetilde{E}^{\frac{1}{2}}\rVert_* \leqslant \Gamma\sqrt{r_1} \wedge \lVert\check{D}_I^{\frac{1}{2}}Z\check{E}_I^{\frac{1}{2}}\rVert_* \leqslant \sqrt{r_2}\Big\}$

References

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

[2] 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.

[3] Nathan Srebro and Tommi Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Advances In Neural Information Processing Systems 17*, pages 5–27. MIT Press, 2005.

[4] Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. In David Helmbold and Bob Williamson, editors, *Computational Learning Theory*, pages 224–240, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg.

[5] *Probability in Banach spaces : isoperimetry and processes / Michel Ledoux ; Michel Talagrand*. Springer, Berlin [u.a.], 1991.

[6] Mehryar Mohri and Andres Munoz Medina. Learning theory and algorithms for revenue optimization in second price auctions with reserve. In Eric P. Xing and Tony Jebara, editors, *Pro-*

ceedings of the 31st International Conference on Machine Learning, volume 32 of *Proceedings of Machine Learning Research*, pages 262–270, Bejing, China, 22–24 Jun 2014. PMLR.

[7] 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.

[8] Afonso S Bandeira and March T Boedihardjo. The spectral norm of gaussian matrices with correlated entries. *arXiv preprint arXiv:2104.02662*, 2021.

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

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

[11] Clayton Scott. Rademacher complexity. *Lecture Notes*, Statistical Learning Theory, 2014.

[12] Ron Meir and Tong . Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4(null):839–860, December 2003.

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

[14] Benjamin Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(null):3413–3430, December 2011.

[15] Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

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

[17] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6240–6249. Curran Associates, Inc., 2017.

[18] Antoine Ledent, Waleed Mustafa, Yunwen Lei, and Marius Kloft. Norm-based generalisation bounds for deep multi-class convolutional neural networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8279–8287, May 2021.

[19] 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.

[20] Kai-Yang Chiang, Inderjit S. Dhillon, and Cho-Jui Hsieh. Using side information to reliably learn low-rank matrices from missing and corrupted observations. *Journal of Machine Learning Ressearch*, 2018.

[21] 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.

[22] 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.

[23] 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.

[24] Antoine Ledent, Rodrigo Alves, and Marius Kloft. Orthogonal inductive matrix completion. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–12, 2021.

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

[26] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[27] Kai Zhong, Prateek Jain, and Inderjit S. Dhillon. Efficient matrix sensing using rank-1 gaussian measurements. In Kamalika Chaudhuri, CLAUDIO GENTILE, and Sandra Zilles, editors, *Algorithmic Learning Theory*, pages 3–18, Cham, 2015. Springer International Publishing.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[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] Rina Foygel, Nathan Srebro, and Russ R Salakhutdinov. Matrix reconstruction with the local max norm. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[34] Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.

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.

(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]