JMLR: Workshop and Conference Proceedings vol 40:1–28, 2015

Fast Exact Matrix Completion with Finite Samples

Prateek Jain PRAJAIN@MICROSOFT.COM
Microsoft Research India, Bangalore India.

Praneeth Netrapalli PRANEETH@MICROSOFT.COM
Microsoft Research New England, Cambridge MA 02142 USA.

Abstract

Matrix completion is the problem of recovering a low rank matrix by observing a small fraction of its entries. A series of recent works (Keshavan, 2012; Jain et al., 2013; Hardt, 2014) have proposed fast non-convex optimization based iterative algorithms to solve this problem. However, the sample complexity in all these results is sub-optimal in its dependence on the rank, condition number and the desired accuracy.

In this paper, we present a fast iterative algorithm that solves the matrix completion problem by observing $O\left(nr^5 \log^3 n\right)$ entries, which is independent of the condition number and the desired accuracy. The run time of our algorithm is $O\left(nr^7 \log^3 n \log 1/\epsilon\right)$ which is near linear in the dimension of the matrix. To the best of our knowledge, this is the first near linear time algorithm for exact matrix completion with finite sample complexity (i.e. independent of ϵ). Our algorithm is based on a well known projected gradient descent method, where the projection is onto the (non-convex) set of low rank matrices. There are two key ideas in our result: 1) our argument is based on a ℓ_∞ norm potential function (as opposed to the spectral norm) and provides a novel way to obtain perturbation bounds for it. 2) we prove and use a natural extension of the Davis-Kahan theorem to obtain perturbation bounds on the best low rank approximation of matrices with good eigen gap. Both of these ideas may be of independent interest.

Keywords: Matrix completion, Matrix perturbation theory, Non-convex Optimization

1. Introduction

In this paper, we study the problem of low-rank matrix completion (LRMC) where the goal is to recover a low-rank matrix by observing a tiny fraction of its entries. That is, given $\mathcal{M} = \{M_{ij}, (i,j) \in \Omega\}$, where $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ is an unknown rank-r matrix and $\Omega \subseteq [n_1] \times [n_2]$ is the set of observed indices, the goal is to recover $\mathbf{M}$. An optimization version of the problem can be posed as follows:

$$(LRMC): \qquad \min_{\boldsymbol{X}} \|P_\Omega\left(\boldsymbol{X} - \mathbf{M}\right)\|_F^2, \;\; s.t. \;\; rank(\boldsymbol{X}) \leq r, \tag{1}$$

where $P_\Omega\left(\boldsymbol{A}\right)$ is defined as:

$$P_\Omega\left(\boldsymbol{A}\right)_{ij} = \begin{cases} A_{ij}, & \text{if } (i,j) \in \Omega, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

LRMC is by now a well studied problem with applications in several machine learning tasks such as collaborative filtering (Bell and Koren, 2007), link analysis (Gleich and Lim, 2011), distance embedding (Candès and Recht, 2009) etc. Motivated by widespread applications, several practical

algorithms have been proposed to solve the problem (heuristically) (Recht and Ré, 2013; Hsieh et al., 2012).

On the theoretical front, the non-convex rank constraint implies NP-hardness in general (Hardt et al., 2014). However, under certain (by now) standard assumptions, a few algorithms have been shown to solve the problem efficiently. These approaches can be categorized into the following two broad groups:

a) The first approach relaxes the rank constraint in (1) to a trace norm constraint (sum of singular values of $\boldsymbol{X}$) and then solves the resulting convex optimization problem (Candès and Recht, 2009). Candès and Tao (2009); Recht (2009) showed that this approach has a near optimal sample complexity (i.e. the number of observed entries of $\mathbf{M}$) of $|\Omega| = O\left(rn\log^2 n\right)$, where we abbreviate $n = n_1 + n_2$. However, current iterative algorithms used to solve the trace-norm constrained optimization problem require $O\left(n^2\right)$ memory and $O\left(n^3\right)$ time per iteration, which is prohibitive for large-scale applications.

b) The second approach is based on an empirically popular iterative technique called Alternating Minimization (AltMin) that factorizes $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$ where $\boldsymbol{U}, \boldsymbol{V}$ have r columns, and the algorithm alternately optimizes over $\boldsymbol{U}$ and $\boldsymbol{V}$ holding the other fixed. Recently, Keshavan (2012); Jain et al. (2013); Hardt (2014); Hardt and Wootters (2014) showed convergence of variants of this algorithm. The best known sample complexity results for AltMin are the incomparable bounds $|\Omega| = O\left(r\kappa^8 n\log\frac{n}{\epsilon}\right)$ and $|\Omega| = O\left(\text{ poly }(r)\,(\log\kappa)\,n\log\frac{n}{\epsilon}\right)$ due to Keshavan (2012) and Hardt and Wootters (2014) respectively. Here, $\kappa = \sigma_1(\mathbf{M})/\sigma_r(\mathbf{M})$ is the condition number of $\mathbf{M}$ and ϵ is the desired accuracy. The computational cost of these methods is $O\left(|\Omega|r + nr^3\right)$ per iteration, making these methods very fast as long as the condition number κ is not too large.

Of the above two approaches AltMin is known to be the most practical and runs in near linear time. However, its sample complexity as well as computational complexity depend on the condition number of $\mathbf{M}$ which can be arbitrarily large. Moreover, for "exact" recovery of $\mathbf{M}$, i.e., with error $\epsilon = 0$, the method needs to observe the entire matrix. The dependence of sample complexity on the desired accuracy arises due to the use of independent samples in each iteration, which in turn is necessitated by the fact that using the same samples in each iteration leads to complex dependencies among iterates which are hard to analyze. Nevertheless, practitioners have been using AltMin with same samples in each iteration successfully in a wide range of applications.

Our results: In this paper, we address this issue by proposing a new algorithm called Stagewise-SVP (**St-SVP**) and showing that it solves the matrix completion problem *exactly* with a sample complexity $|\Omega| = O\left(nr^5\log^3 n\right)$, which is independent of both the condition number, and desired accuracy and time complexity per iteration $O\left(|\Omega|\,r^2\right)$, which is near linear in n.

The basic block of our algorithm is a simple projected gradient descent step, first proposed by Jain et al. (2010) in the context of this problem. More precisely, given the t^{th} iterate $\boldsymbol{X}_t$, Jain et al. (2010) proposed the following update rule, which they call singular value projection (SVP).

$$(SVP): \boldsymbol{X}_{t+1} = P_r\left(\boldsymbol{X}_t + \frac{n_1 n_2}{|\Omega|}P_\Omega\left(\mathbf{M} - \boldsymbol{X}_t\right)\right), \tag{3}$$

where P_r is the projection onto the set of rank-r matrices and can be efficiently computed using singular value decomposition (SVD). Note that the SVP step is just a projected gradient descent step where the projection is onto the (non-convex) set of low rank matrices. Jain et al. (2010) showed that despite involving projections onto a non-convex set, SVP solves the related problem of low-rank matrix sensing, where instead of observing elements of the unknown matrix, we observe dense

linear measurements of this matrix. However, their result does not extend to the matrix completion problem and the correctness of SVP for matrix completion was left as an open question.

Our preliminary result resolves this question by showing the correctness of SVP for the matrix completion problem, albeit with a sample complexity that depends on the condition number and desired accuracy. We then develop a stage-wise variant of this algorithm, where in the k^{th} stage, we try to recover $P_k(\mathbf{M})$, thereby getting rid of the dependence on the condition number. Interestingly, our stage-wise variant not only admits better theoretical bounds, it also improves the algorithm in practice (see Appendix E). Finally, in each stage, we use independent samples for $\log n$ iterations, but use same samples for the remaining iterations, thereby eliminating the dependence of sample complexity on ϵ.

Our techniques: Our analysis relies on two key novel techniques that enable us to understand SVP style projected gradient methods even though the projection is onto a *non-convex* set. Firstly, we use ℓ_∞ norm of the error $\boldsymbol{X}_t - \mathbf{M}$ as our potential function, instead of its spectral norm that most existing analysis of matrix completion use. In general, bounds on the ℓ_∞ norm are much harder to obtain as projection via SVD is optimal only in the spectral and Frobenius norms. We obtain ℓ_∞ norm bounds by writing down explicit eigenvector equations for the low rank projection and using this to control the ℓ_∞ norm of the error. Secondly, in order to analyze the SVP updates with same samples in each iteration, we prove and use a natural extension of the Davis-Kahan theorem. This extension bounds the perturbation in the best rank-k approximation of a matrix (with large enough eigen-gap) due to any additive perturbation; despite this being a very natural extension of the Davis-Kahan theorem, to the best of our knowledge, it has not been considered before. We believe both of the above techniques should be of independent interest.

Paper Organization: We first present the problem setup, our main result and an overview of our techniques in the next section. We then present a "warm-up" result for the basic SVP method in Section 3. We then present our main algorithm (St-SVP) and its analysis in Section 4. We conclude the discussion in Section 5. The proofs of all the technical lemmas will follow thereafter in the appendix.

Notation: We denote matrices with boldface capital letters ($\mathbf{M}$) and vectors with boldface letters ($\boldsymbol{x}$). $\mathbf{m}_i$ denotes the i^{th} column and M_{ij} denotes the $(i,j)^{\text{th}}$ entry respectively of $\mathbf{M}$. SVD and EVD stand for the singular value decomposition and eigenvalue decomposition respectively. $P_k(\boldsymbol{A})$ denotes the projection of $\boldsymbol{A}$ onto the set of rank-k matrices. That is, if $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the SVD of $\boldsymbol{A}$, then $P_k(\boldsymbol{A}) = \boldsymbol{U}_k\boldsymbol{\Sigma}_k\boldsymbol{V}_k^\top$ where $\boldsymbol{U}_k \in \mathbb{R}^{n_1\times k}$ and $\boldsymbol{V}_k \in \mathbb{R}^{n_2\times k}$ are the k left and right singular vectors respectively of $\boldsymbol{A}$ corresponding to the k largest singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k$. $\|\boldsymbol{u}\|_q$ denotes the ℓ_q norm of $\boldsymbol{u}$. We denote the operator norm of $\mathbf{M}$ by $\|\mathbf{M}\|_2 = \max_{\boldsymbol{u},\|\boldsymbol{u}\|_2=1} \|\mathbf{M}\boldsymbol{u}\|_2$. In general, $\|\alpha\|_2$ denotes the ℓ_2 norm of α if it is a vector and the operator norm of α if it is a matrix. $\|\mathbf{M}\|_F$ denotes the Frobenius norm of $\mathbf{M}$.

2. Our Results and Techniques

In this section, we will first describe the problem set up and then present our results as well as the main techniques we use.

2.1. Problem Setup

Let $\mathbf{M}$ be an $n_1 \times n_2$ matrix of rank-r. Let $\Omega \subseteq [n_1] \times [n_2]$ be a subset of the indices. Recall that $P_\Omega(\mathbf{M})$ (as defined in (2)) is the projection of $\mathbf{M}$ on to the indices in Ω. Given Ω, $P_\Omega(\mathbf{M})$ and r,

the goal is to recover $\mathbf{M}$. The problem is in general ill posed, so we make the following standard assumptions on $\mathbf{M}$ and Ω Candès and Recht (2009).

Assumption 1 (Incoherence) *$\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ is a rank-r, μ-incoherent matrix i.e., $\max_i \|\boldsymbol{e}_i^T \boldsymbol{U}^*\|_2 \leq \frac{\mu\sqrt{r}}{\sqrt{n_1}}$ and $\max_j \|\boldsymbol{e}_j^T \boldsymbol{V}^*\|_2 \leq \frac{\mu\sqrt{r}}{\sqrt{n_2}}$, where $\mathbf{M} = \boldsymbol{U}^* \boldsymbol{\Sigma} \boldsymbol{V}^{*\top}$ is the singular value decomposition of $\mathbf{M}$.*

Assumption 2 (Uniform sampling) *Ω is generated by sampling each element of $[n_1] \times [n_2]$ independently with probability p.*

The incoherence assumption ensures that the mass of the matrix is well spread out and a small fraction of uniformly random observations give enough information about the matrix. Both of the above assumptions are standard and are used by most of the existing results, for instance (Candès and Recht, 2009; Candès and Tao, 2009; Keshavan et al., 2010; Recht, 2009; Keshavan, 2012). A few exceptions include the works of Meka et al. (2009); Chen et al. (2014); Bhojanapalli and Jain (2014).

2.2. Main Result

The following theorem is the main result of this paper.

Theorem 1 *Suppose $\mathbf{M}$ and Ω satisfy Assumptions 1 and 2 respectively. Also, let*

$$\mathbb{E}[|\Omega|] \geq C\alpha\mu^4 r^5 n \log^3 n,$$

where $\alpha > 1$, $n := n_1 + n_2$ and $C > 0$ is a global constant. Then, the output $\widehat{\mathbf{M}}$ of Algorithm 2 satisfies: $\left\|\widehat{\mathbf{M}} - \mathbf{M}\right\|_F \leq \epsilon$, with probability greater than $1 - n^{-10-\log\alpha}$. Moreover, the run time of Algorithm 2 is $O\left(|\Omega|\, r^2 \log(1/\epsilon)\right)$.

Algorithm 2 is based on the projected gradient descent update (3) and proceeds in r stages where in the k-th stage, projections are performed onto the set of rank-k matrices. See Section 4 for a detailed description and the underlying intuition behind our algorithm.

Table 1 compares our result to that for nuclear norm minimization, which is the only other polynomial time method with finite sample complexity guarantees (i.e. no dependence on the desired accuracy ϵ). Note that St-SVP runs in time near linear in the ambient dimension of the matrix (n), where as nuclear norm minimization runs in time cubic in the ambient dimension. However, the sample complexity of St-SVP is suboptimal in its dependence on the incoherence parameter μ and rank r. We believe closing this gap between the sample complexity of St-SVP and that of nuclear norm minimization should be possible and leave it for future work.

	Sample complexity	Comp. complexity
Nuclear norm minimization Recht (2009)	$O\left(\mu^2 r n \log^2 n\right)$	$O\left(n^3 \log \frac{1}{\epsilon}\right)$
St-SVP (This paper)	$O\left(\mu^4 r^5 n \log^3 n\right)$	$O\left(\mu^4 r^7 n \log^3 n \log(1/\epsilon)\right)$

Table 1: Comparison of our result to that for nuclear norm minimization.

2.3. Overview of Techniques

In this section, we briefly present the key ideas and lemmas we use to prove Theorem 1. Our proof revolves around analyzing the basic SVP step (3): $\boldsymbol{X}_{t+1} = P_k\left(\boldsymbol{X}_t + \frac{1}{p}P_\Omega\left(\mathbf{M} - \boldsymbol{X}_t\right)\right) = P_k(\mathbf{M} + \widehat{\mathbf{H}})$ where p is the sampling probability, $\widehat{\mathbf{H}} := \boldsymbol{X}_t - \mathbf{M} - \frac{1}{p}P_\Omega\left(\boldsymbol{X}_t - \mathbf{M}\right) = \mathbf{E} - \frac{1}{p}P_\Omega(\mathbf{E})$ and $\mathbf{E} := \boldsymbol{X}_t - \mathbf{M}$ is the error matrix. Hence, $\boldsymbol{X}_{t+1}$ is given by a rank-k projection of $\mathbf{M} + \widehat{\mathbf{H}}$, which is a perturbation of the desired matrix $\mathbf{M}$.

Bounding the ℓ_∞ norm of errors: As the SVP update is based on projection onto the set of rank-k matrices, a natural potential function to analyze would be $\|\mathbf{E}\|_2$ or $\|\mathbf{E}\|_F$. However, such a potential function requires bounding norms of $\mathbf{E} - \frac{1}{p}P_\Omega(\mathbf{E})$ which in turn would require us to show that $\mathbf{E}$ is incoherent. This is the approach taken by papers on AltMin (Keshavan, 2012; Jain et al., 2013; Hardt, 2014).

In contrast, in this paper, we consider $\|\mathbf{E}\|_\infty$ as the potential function. So the goal is to show that $\left\|P_k\left(\mathbf{M} + \widehat{\mathbf{H}}\right) - \mathbf{M}\right\|_\infty$ is much smaller than $\|\mathbf{E}\|_\infty$. Unfortunately, standard perturbation results such as the Davis-Kahan theorem provide bounds on spectral, Frobenius or other unitarily invariant norms and do not apply to the ℓ_∞ norm.

In order to carry out this argument, we write the singular vectors of $\mathbf{M} + \widehat{\mathbf{H}}$ as solutions to eigenvector equations and then use these to write $\boldsymbol{X}_{t+1}$ explicitly via Taylor series expansion. We use this technique to prove the following more general lemma.

Lemma 2 *Suppose $\mathbf{M} \in \mathbb{R}^{n\times n}$ is a symmetric matrix satisfying Assumption 1. Let $\sigma_1 \geq \cdots \geq \sigma_r$ denote its singular values. Let $\mathbf{H} \in \mathbb{R}^{n\times n}$ be a random symmetric matrix such that each H_{ij} is independent with $\mathbb{E}\left[H_{ij}\right] = 0$ and $\mathbb{E}\left[|H_{ij}|^a\right] \leq 1/n$ for $2 \leq a \leq \log n$. Then, for any $\alpha > 1$ and $|\beta| \leq \frac{\sigma_k}{200\sqrt{\alpha}\log n}$ we have:*

$$\|\mathbf{M} - P_k\left(\mathbf{M} + \beta\mathbf{H}\right)\|_\infty \leq \frac{\mu^2 r^2}{n}\left(\sigma_{k+1} + 15|\beta|\sqrt{\alpha}\log n\right),$$

with probability greater than $1 - n^{-10-\log\alpha}$.

Proceeding in stages: If we applied Lemma 2 with $k = r$, we would require $|\beta|$ to be much smaller than σ_r. Now, β can be thought of as $\beta \approx \sqrt{\frac{n}{p}}\|\mathbf{E}\|_\infty$. If we start with $\boldsymbol{X}_0 = 0$, we have $\mathbf{E} = -\mathbf{M}$, and so $\|\mathbf{E}\|_\infty = \|\mathbf{M}\|_\infty \leq \frac{\sigma_1\mu^2 r}{n}$. To make $\beta \leq \sigma_r$, we would need the sampling probability p to be quadratic in the condition number $\kappa = \sigma_1/\sigma_r$. In order to overcome this issue, we perform SVP in r stages with the k^{th} stage performing projections on to the set of rank-k matrices while maintaining the invariant that at the end of $(k-1)^{\text{th}}$ stage, $\|\mathbf{E}\|_\infty = O(\sigma_k/n)$. This lets us choose a p independent of κ while still ensuring $\beta \approx \sqrt{\frac{n}{p}}\|\mathbf{E}\|_\infty \leq \sigma_k$. Lemma 2 tells us that at the end of the k^{th} stage, the error $\|\mathbf{E}\|_\infty$ is $O\left(\frac{\sigma_{k+1}}{n}\right)$, thereby establishing the invariant for the $(k+1)^{\text{th}}$ stage.

Using same samples: In order to reduce the error from $O\left(\frac{\sigma_k}{n}\right)$ to $O\left(\frac{\sigma_{k+1}}{n}\right)$, the k^{th} stage would require $O\left(\log\frac{\sigma_k}{\sigma_{k+1}}\right)$ iterations. Since Lemma 2 requires the elements of $\mathbf{H}$ to be independent, in order to apply it, we need to use fresh samples in each iteration. This means that the sample complexity increases with $\frac{\sigma_k}{\sigma_{k+1}}$, or the desired accuracy ϵ if $\epsilon < \sigma_{k+1}$. This problem is faced by all the existing analysis for iterative algorithms for matrix completion (Keshavan, 2012; Jain et al., 2013; Hardt, 2014; Hardt and Wootters, 2014). We tackle this issue by observing that when $\mathbf{M}$ is

Algorithm 1 SVP for matrix completion

Input: $\Omega, P_\Omega(\mathbf{M}), r, \epsilon$
$T \leftarrow \log \frac{(n_1+n_2)\|\mathbf{M}\|_F}{\epsilon}$
Partition Ω randomly into T subsets $\{\Omega_t : t \in [T]\}$
$\boldsymbol{X}_t \leftarrow 0$
for $t \leftarrow 1, \cdots, T$ **do**
    $\boldsymbol{X}_t \leftarrow P_r\left(\boldsymbol{X}_{t-1} - \frac{n_1 n_2}{|\Omega_t|} P_{\Omega_t}\left(\boldsymbol{X}_{t-1} - \mathbf{M}\right)\right)$
end for
Output: $\boldsymbol{X}_T$

ill conditioned and $\|\mathbf{E}\|_F$ is very small, we can show a decay in $\|\mathbf{E}\|_F$ using the same samples for SVP iterations:

Lemma 3 *Let $\mathbf{M}$ and Ω be as in Theorem 1 with $\mathbf{M}$ being a symmetric matrix. Further, let $\mathbf{M}$ be ill conditioned in the sense that $\|\mathbf{M} - P_k(\mathbf{M})\|_F < \frac{\sigma_k}{n^2}$, where $\sigma_1 \geq \cdots \geq \sigma_r$ are the singular values of $\mathbf{M}$. Then, the following holds for all rank-k $\boldsymbol{X}$ s.t. $\|\boldsymbol{X} - P_k(\mathbf{M})\|_F < \frac{\sigma_k}{n^2}$ (w.p. $\geq 1 - n^{-10-\alpha}$):*

$$\|\boldsymbol{X}_+ - P_k(\mathbf{M})\|_F \leq \frac{1}{10}\|\boldsymbol{X} - P_k(\mathbf{M})\|_F + \frac{1}{p}\|\mathbf{M} - P_k(\mathbf{M})\|_F,$$

where $\boldsymbol{X}_+ := P_k\left(\boldsymbol{X} - \frac{1}{p}P_\Omega\left(\boldsymbol{X} - \mathbf{M}\right)\right)$ denotes the rank-k SVP update of $\boldsymbol{X}$ and $p = \mathbb{E}[|\Omega|]/n^2 = \frac{C\alpha\mu^4 r^5 \log^3 n}{n}$ is the sampling probability.

The following lemma plays a crucial role in proving Lemma 3. It is a natural extension of the Davis-Kahan theorem for singular vector subspace perturbation.

Lemma 4 *Suppose $\boldsymbol{A}$ is a matrix such that $\sigma_{k+1}(\boldsymbol{A}) \leq \frac{1}{4}\sigma_k(\boldsymbol{A})$. Then, for any matrix $\mathbf{E}$ such that $\|\mathbf{E}\|_F < \frac{1}{4}\sigma_k(\boldsymbol{A})$, we have:*

$$\|P_k\left(\boldsymbol{A} + \mathbf{E}\right) - P_k\left(\boldsymbol{A}\right)\|_F \leq c\left(\sqrt{k}\,\|\mathbf{E}\|_2 + \|\mathbf{E}\|_F\right),$$

for some absolute constant c.

In contrast to the Davis-Kahan theorem, which establishes a bound on the perturbation of the space of singular vectors, Lemma 4 establishes a bound on the perturbation of the best rank-k approximation of a matrix $\boldsymbol{A}$ with good eigen gap, under small perturbations. This is a very natural quantity while considering perturbations of low rank approximations, and we believe it may find applications in other scenarios as well. A naïve argument using Davis-Kahan theorem would only yield a weaker version of Lemma 4 with a bound that depends on $\sigma_1(\boldsymbol{A})/\sigma_k(\boldsymbol{A})$. Our proof of Lemma 4 is much more intricate where we keep track of perturbations in various subspaces simultaneously. A final remark regarding Lemma 4: we suspect it might be possible to tighten the right hand side of the result to $c\min\left(\sqrt{k}\,\|\mathbf{E}\|_2, \|\mathbf{E}\|_F\right)$, but have not been able to prove it.

3. Singular Value Projection

Before we go on to prove Theorem 1, in this section we will analyze the basic SVP algorithm (Algorithm 1), bounding its sample complexity and thereby resolving a question posed by Jain et al. (2010). This analysis also serves as a warm-up exercise for our main result and brings out the key ideas in analyzing the ℓ_∞ norm potential function while also highlighting some issues with Algorithm 1 that we will fix later on.

As is clear from the pseudocode in Algorithm 1, SVP is a simple projected gradient descent method for solving the matrix completion problem. Note that Algorithm 1 first splits the set Ω into T random subsets and updates iterate $\boldsymbol{X}_t$ using Ω_t. This step is critical for analysis as it ensures that Ω_t is independent of $\boldsymbol{X}_{t-1}$, allowing for the use of standard tail bounds. The following theorem is our main result for Algorithm 1:

Theorem 5 *Suppose* $\mathbf{M}$ *and* Ω *satisfy Assumptions 1 and 2 respectively with*

$$\mathbb{E}\left[|\Omega|\right] \geq C\alpha\mu^4\kappa^2 r^5 n \left(\log^2 n\right) T,$$

where $n = n_1 + n_2, \alpha > 1, \kappa = \left(\frac{\sigma_1}{\sigma_r}\right)$ *with* $\sigma_1 \geq \cdots \geq \sigma_r$ *denoting the singular values of* $\mathbf{M}$, $T = \log \frac{100\mu^2 r\|\mathbf{M}\|_F}{\epsilon}$ *and* $C > 0$ *is a large enough global constant. Then, the output of Algorithm 1 satisfies (w.p.* $\geq 1 - T\min(n_1, n_2)^{-10-\log\alpha}$*):* $\|\boldsymbol{X}_T - \mathbf{M}\|_F \leq \epsilon$

Proof Using a standard dilation argument (Lemma 6), it suffices to prove the result for symmetric matrices. Let $p = \frac{\mathbb{E}[|\Omega_t|]}{n^2} = \frac{\mathbb{E}[|\Omega|]}{n^2 T}$ be the probability of sampling in each iteration. Now, let $\mathbf{E} = \boldsymbol{X}_{t-1} - \mathbf{M}$ and $\widehat{\mathbf{H}} = \mathbf{E} - \frac{1}{p}P_{\Omega_t}(\mathbf{E})$. Then, the SVP update (line 6 of Algorithm 1) is given by: $\boldsymbol{X}_t = P_r(\mathbf{M} + \widehat{\mathbf{H}})$. Since Ω_t is sampled uniformly at random, it is easy to check that $\mathbb{E}[\widehat{H}_{ij}] = 0$ and $\mathbb{E}\left[\left|\widehat{H}_{ij}\right|^s\right] \leq \beta^s/n$ where $\beta = \frac{2\sqrt{n}\|E\|_\infty}{\sqrt{p}} \leq \frac{2\mu^2 r\sigma_1}{\sqrt{np}}$ (Lemma 8). By our choice of p, we have $\beta < \frac{\sigma_r}{200\sqrt{\alpha}}$. Applying Lemma 2 with $k = r$, we have $\|\boldsymbol{X}_t - \mathbf{M}\|_\infty \leq \frac{15\mu^2 r^2}{n}\beta\sqrt{\alpha}\log n \leq (1/\sqrt{30C})\|\mathbf{E}\|_\infty = \frac{1}{2}\|\boldsymbol{X}_{t-1} - \mathbf{M}\|_\infty$, where the last inequality is obtained by selecting C large enough. The theorem is immediate from this error decay in each step. ■

3.1. Detailed Proof of Lemma 2

We are now ready to present a proof of Lemma 2. Recall that $\boldsymbol{X}_+ = P_k(\mathbf{M} + \beta\mathbf{H})$, hence,

$$(\mathbf{M} + \beta\mathbf{H})\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i, \ \forall 1 \leq i \leq k, \tag{4}$$

where $(\boldsymbol{u}_i, \lambda_i)$ is the i^{th} $(i \leq k)$ top eigenvector-eigenvalue pair (in terms of magnitude).

Now, as $\mathbf{H}$ satisfies conditions of Definition 7, we can apply Lemma 10 to obtain:

$$|\beta|\,\|\mathbf{H}\|_2 \leq |\beta| \cdot 3\sqrt{\alpha} \leq \frac{|\sigma_k|}{5}. \tag{5}$$

Using Lemma 11 and (5), we have:

$$|\lambda_i| \geq |\sigma_i| - |\beta|\,\|\mathbf{H}\|_2 \geq \frac{4\,|\sigma_k|}{5} \ \forall\, i \in [k]. \tag{6}$$

Using (4), we have: $\left(\mathbf{I} - \frac{\beta}{\lambda_i}\mathbf{H}\right)\boldsymbol{u}_i = \frac{1}{\lambda_i}\mathbf{M}\boldsymbol{u}_i$. Moreover, using (6), $\mathbf{I} - \frac{\beta}{\lambda_i}\mathbf{H}$ is invertible. Hence, using Taylor series expansion, we have:

$$\boldsymbol{u}_i = \frac{1}{\lambda_i}\left(\mathbf{I} + \frac{\beta}{\lambda_i}\mathbf{H} + \frac{\beta^2}{\lambda_i^2}(\mathbf{H})^2 + \cdots\right)\mathbf{M}\boldsymbol{u}_i.$$

Letting $\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ denote the eigenvalue decomposition (EVD) of $\boldsymbol{X}_+$, we obtain:

$$\boldsymbol{X}_+ = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top = \sum_{a,b\geq 0}\beta^{a+b}(\mathbf{H})^a\,\mathbf{M}\boldsymbol{U}\Lambda^{-(a+b+1)}\boldsymbol{U}^\top\mathbf{M}\,(\mathbf{H})^b.$$

Using triangle inequality, we have:

$$\|\boldsymbol{X}_+ - \mathbf{M}\|_\infty \leq \left\|\mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\mathbf{M} - \mathbf{M}\right\|_\infty + \sum_{\substack{a,b\geq 0\\ a+b\geq 1}}|\beta|^{a+b}\left\|(\mathbf{H})^a\,\mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-(a+b+1)}\boldsymbol{U}^\top\mathbf{M}^\top(\mathbf{H})^b\right\|_\infty. \tag{7}$$

Using Lemma 12, we have the following bound for the first term above:

$$\left\|\mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\mathbf{M} - \mathbf{M}\right\|_\infty \leq \frac{\mu^2 r}{n}\left\|\mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\mathbf{M} - \mathbf{M}\right\|_2. \tag{8}$$

Furthermore, using Lemma 13 we have:

$$\left\|\mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\mathbf{M} - \mathbf{M}\right\|_2 \leq |\sigma_{k+1}| + 5\,|\beta|\,\|\mathbf{H}\|_2, \text{ and} \tag{9}$$

$$\left\|\mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-a}\boldsymbol{U}^\top\mathbf{M}\right\|_2 \leq 4\left(\frac{|\sigma_k|}{2}\right)^{-a+2}\quad\forall\, a\geq 2. \tag{10}$$

Plugging (9) into (8) gives us:

$$\left\|\mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\mathbf{M} - \mathbf{M}\right\|_\infty \leq \frac{\mu^2 r}{n}\left(|\sigma_{k+1}| + 5\,|\beta|\,\|\mathbf{H}\|_2\right). \tag{11}$$

Let $\mathbf{M} = \mathbf{U}^*\boldsymbol{\Sigma}\left(\mathbf{U}^*\right)^\top$ denote the EVD of $\mathbf{M}$. We now bound the terms in the summation in (7) for $1 \leq a+b < \log n$.

$$\begin{aligned}
&|\beta|^{a+b}\left\|(\mathbf{H})^a\,\mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-(a+b+1)}\boldsymbol{U}^\top\mathbf{M}\,(\mathbf{H})^b\right\|_\infty\\
&= |\beta|^{a+b}\max_{i,j}\boldsymbol{e}_i^\top(\mathbf{H})^a\,\mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-(a+b+1)}\boldsymbol{U}^\top\mathbf{M}\,(\mathbf{H})^b\,\boldsymbol{e}_j\\
&\leq |\beta|^{a+b}\left(\max_i\left\|\boldsymbol{e}_i^\top(\mathbf{H})^a\,\mathbf{U}^*\right\|_2\right)\left\|\boldsymbol{\Sigma}\left(\mathbf{U}^*\right)^\top\boldsymbol{U}\boldsymbol{\Lambda}^{-(a+b+1)}\boldsymbol{U}^\top\mathbf{U}^*\boldsymbol{\Sigma}\right\|_2\left(\max_j\left\|\left(\mathbf{U}^*\right)^\top(\mathbf{H})^b\,\boldsymbol{e}_j\right\|_2\right)\\
&\leq |\beta|^{a+b}\left(\sqrt{r}\max_i\|(\mathbf{H})^a\,\boldsymbol{u}_i^*\|_\infty\right)\left\|\mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-(a+b+1)}\boldsymbol{U}^\top\mathbf{M}\right\|_2\left(\sqrt{r}\max_j\left\|(\mathbf{H})^b\,\boldsymbol{u}_j^*\right\|_\infty\right)\\
&\overset{(\zeta_1)}{\leq}\frac{\mu^2 r^2}{n}|\beta|^{a+b}\left(10\sqrt{\alpha}\log n\right)^{a+b}\left\|\mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-(a+b+1)}\boldsymbol{U}^\top\mathbf{M}\right\|_2\\
&\overset{(\zeta_2)}{\leq}\frac{\mu^2 r^2}{n}|\beta|^{a+b}\left(10\sqrt{\alpha}\log n\right)^{a+b}\cdot 4\left(\frac{2}{|\sigma_k|}\right)^{a+b-1}\\
&\leq\frac{\mu^2 r^2}{n}\left(\frac{80\,|\beta|\sqrt{\alpha}\log n}{|\sigma_k|}\right)^{a+b-1}\left(10\,|\beta|\sqrt{\alpha}\log n\right)\leq\frac{\mu^2 r^2}{n}\left(\frac{1}{20}\right)^{a+b-1}\cdot 10\,|\beta|\sqrt{\alpha}\log n,
\end{aligned} \tag{12}$$

where (ζ_1) follows from Lemma 9 and (ζ_2) follows from (10).

For $a+b \geq \log n$, we have

$$\begin{aligned}
|\beta|^{a+b} \left\| (\mathbf{H})^a \, \mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-(a+b+1)}\boldsymbol{U}^\top \mathbf{M} \, (\mathbf{H})^b \right\|_\infty &\leq |\beta|^{a+b} \left\| (\mathbf{H})^a \, \mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-(a+b+1)}\boldsymbol{U}^\top \mathbf{M} \, (\mathbf{H})^b \right\|_2 \\
&\leq |\beta|^{a+b} \, \|\mathbf{H}\|_2^a \left\| \mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-(a+b+1)}\boldsymbol{U}^\top \mathbf{M} \right\|_2 \|\mathbf{H}\|_2^b \\
&\leq |\beta|^{a+b} \, \|\mathbf{H}\|_2^{a+b} \left(\frac{5}{4\,|\sigma_k|} \right)^{a+b-1} \\
&\leq \left(\frac{15\,|\beta|\,\sqrt{\alpha}}{4\,|\sigma_k|} \right)^{a+b-1} \cdot 3\,|\beta|\,\sqrt{\alpha} \\
&\leq \frac{\mu^2 r^2}{n} \left(\frac{1}{20} \right)^{a+b-1} \left(10\,|\beta|\,\sqrt{\alpha}\log n \right), \qquad (13)
\end{aligned}$$

where we used Lemma 13 to bound $\left\| \mathbf{M}\boldsymbol{U}\boldsymbol{\Lambda}^{-(a+b+1)}\boldsymbol{U}^\top \mathbf{M} \right\|_2$ and Lemma 10 to bound $\|\mathbf{H}\|_2$. The last inequality follows from using $(1/2)^{a+b} \leq 1/n \leq \frac{\mu^2 r^2}{n}$ as $a+b > \log n$.

Plugging (11), (12) and (13) in (7) gives us:

$$\begin{aligned}
\|\boldsymbol{X}_+ - \mathbf{M}\|_\infty &\leq \frac{\mu^2 r}{n} \left(|\sigma_{k+1}| + 5\,|\beta|\,\|\mathbf{H}\|_2 \right) + \frac{\mu^2 r^2}{n} \sum_{\substack{a,b\geq 0 \\ a+b\geq 1}} \left(\frac{1}{20} \right)^{a+b} \left(10\,|\beta|\,\sqrt{\alpha}\log n \right) \\
&\leq \frac{\mu^2 r^2}{n} \left(|\sigma_{k+1}| + 15\,|\beta|\,\sqrt{\alpha}\log n \right).
\end{aligned}$$

This proves the lemma.

4. Stagewise-SVP

Theorem 5 is suboptimal in its sample complexity dependence on the rank, condition number and desired accuracy. In this section, we will fix two of these issues – the dependence on condition number and desired accuracy – by designing a stagewise version of Algorithm 1 and proving Theorem 1.

Our algorithm, St-SVP (pseudocode presented in Algorithm 2) runs in r stages, where in the k^{th} stage, the projection is onto the set of *rank-k matrices*. In each stage, the goal is to obtain an approximation of $\mathbf{M}$ up to an error of σ_{k+1}. In order to do this, we use the basic SVP updates, but in a very specific way, so as to avoid the dependence on condition number and desired accuracy.

- **(Step I) Apply SVP update with fresh samples for $\log n$ iterations**: Run $\log n$ steps of SVP update (3), with fresh samples in each iteration. Using fresh samples allows us to use Lemma 2 ensuring that the ℓ_∞ norm of the error between our estimate and $\mathbf{M}$ decays to $\|\boldsymbol{X}_{k,\log n} - \mathbf{M}\|_\infty = O\left(\frac{1}{n}\left(\sigma_{k+1} + \frac{\sigma_k}{n^3}\right)\right)$.
- **(Step II) Determine if $\sigma_{k+1} > \frac{\sigma_k}{n^3}$**: Note that we can determine this, by using the $(k+1)^{\text{th}}$ singular value of the matrix obtained after the gradient step, i.e., $\sigma_{k+1}(\boldsymbol{X}_{k,\log n} - \frac{1}{p}P_{\Omega_{k,\log n}}(\boldsymbol{X}_{k,\log n} - \mathbf{M}))$. If true, the error $\|\boldsymbol{X}_{k,\log n} - \mathbf{M}\|_\infty = O\left(\frac{\sigma_{k+1}}{n}\right)$, and so the algorithm proceeds to the $(k+1)^{\text{th}}$ stage.

Algorithm 2 Stagewise SVP (St-SVP) for matrix completion

Input: $\Omega, P_\Omega(\mathbf{M}), \epsilon, r$
$T \leftarrow \log \frac{100\mu^2 r\|\mathbf{M}\|_F}{\epsilon}$
Partition Ω into $2r\log n$ subsets $\{\Omega_{k,t} : k \in [r], t \in [\log n] \cup \{T+\log n+1, \cdots, T+2\log n\}\}$ uniformly at random
$k \leftarrow 1$, $\boldsymbol{X}_{k,0} \leftarrow 0$
for $k \leftarrow 1, \cdots, r$ **do**
  / Stage-k */*
  for $t = 1, \cdots, \log n$ **do**
    $\boldsymbol{X}_{k,t} \leftarrow PGD(\boldsymbol{X}_{k,t-1}, P_{\Omega_{k,t}}(M), \Omega_{k,t}, k)$*/* SVP Step with re-sampling*/* } Step I
  end for
  if $\sigma_{k+1}\left(GD\left(\boldsymbol{X}_{k,\log n}, P_{\Omega_{k,\log n}}(\mathbf{M}), \Omega_{k,\log n}\right)\right) > \frac{\sigma_k\left(\boldsymbol{X}_{k,\log n}\right)}{n^2}$ **then**
    $\boldsymbol{X}_{k+1,0} \leftarrow \boldsymbol{X}_{k,\log n}$ */* Initialize for next stage and continue*/* } Step II
    continue
  end if
  for $t = \log n + 1, \cdots, \log n + T$ **do**
    $\boldsymbol{X}_{k,t} \leftarrow PGD(\boldsymbol{X}_{k,t-1}, P_\Omega(\mathbf{M}), \Omega, k)$ */* SVP Step without re-sampling */* } Step III
  end for
  for $t = \log n + T + 1, \cdots, \log n + T + \log n$ **do**
    $\boldsymbol{X}_{k,t} \leftarrow PGD(\boldsymbol{X}_{k,t-1}, P_{\Omega_{k,t}}(M), \Omega_{k,t}, k)$ */* SVP Step with re-sampling */* } Step IV
  end for
  $\boldsymbol{X}_{k+1,0} \leftarrow \boldsymbol{X}_{k,t}$ */* Initialization for next stage */*
  Output: $\boldsymbol{X}_{k,t}$ if $\sigma_{k+1}\left(GD(\boldsymbol{X}_{k,t-1}, P_{\Omega_{k,t}}(\mathbf{M}), \Omega_{k,t})\right) < \frac{\epsilon}{10\mu^2 r}$
end for

Sub-routine 3 Projected Gradient Descent (PGD)

Input: $X \in \mathbb{R}^{n_1 \times n_2}, P_\Omega(\mathbf{M}), \Omega, k$
Output: $X_{next} \leftarrow P_k(X - \frac{n_1 n_2}{|\Omega|} P_\Omega(X - M))$

Sub-routine 4 Gradient Descent (GD)

Input: $X \in \mathbb{R}^{n_1 \times n_2}, P_\Omega(\mathbf{M}), \Omega$
Output: $X_{next} \leftarrow X - \frac{n_1 n_2}{|\Omega|} P_\Omega(X - M)$

- **(Step III) If not (i.e., $\sigma_{k+1} \le \frac{\sigma_k}{n^3}$), apply SVP update for $T = \log\frac{1}{\epsilon}$ iterations with same samples**: If $\sigma_{k+1} \le \frac{\sigma_k}{n^3}$, we can use Lemma 3 to conclude that after $\log\frac{1}{\epsilon}$ iterations, the Frobenius norm of error is $\|\boldsymbol{X}_{k,\log n+T} - \mathbf{M}\|_F = O\left(n\sigma_{k+1} + \epsilon\right)$.
- **(Step IV) Apply SVP update with fresh samples for $\log n$ iterations**: To set up the invariant $\|\boldsymbol{X}_{k+1,0} - \mathbf{M}\|_\infty = O\left(\sigma_{k+1}/n\right)$ for the next stage, we wish to convert our Frobenius norm bound $\|\boldsymbol{X}_{k,\log n+T} - \mathbf{M}\|_F = O\left(n\sigma_{k+1}\right)$ to an ℓ_∞ bound $\|\boldsymbol{X}_{k,2\log n+T} - \mathbf{M}\|_\infty = O\left(\frac{\sigma_{k+1}}{n}\right)$. Since $\sigma_{k+1} < \frac{\sigma_k}{n^3}$, we can bound the initial Frobenius error by $O\left(\frac{1}{n}\left(\left(\frac{1}{2}\right)^{\widehat{T}}\sigma_k + \sigma_{k+1}\right)\right)$ for some $\widehat{T} = O\left(\log\frac{\sigma_k}{n^2\sigma_{k+1}}\right)$. As in Step I, after $\log n$ SVP updates with fresh samples,

Lemma 2 lets us conclude that $\|\boldsymbol{X}_{k,2\log n+T} - \mathbf{M}\|_\infty = O\left(\frac{\sigma_{k+1}}{n}\right)$, setting up the invariant for the next stage.

4.1. Analysis of St-SVP (Proof of Theorem 1)

We will now present a proof of Theorem 1.

Proof [Proof of Theorem 1] Just as in Theorem 5, it suffices to prove the result for when $\mathbf{M}$ is symmetric.For every stage, we will establish the following invariant:

$$\|\boldsymbol{X}_{k,0} - \mathbf{M}\|_\infty < \frac{4\mu^2 r^2}{n}\sigma_{k+1}, \text{ for } k < r. \tag{14}$$

We will use induction. (14) clearly holds for the base case $k = 1$. Now, suppose (14) holds for the k^{th} stage, we will prove that it holds for the $(k+1)^{\text{th}}$ stage. The analysis follows the four step outline in the previous section:

Step I: Here, we will show that for every iteration t, we have:

$$\|\boldsymbol{X}_{k,t} - \mathbf{M}\|_\infty < \frac{4\mu^2 r^2}{n}\gamma_{k,t}, \text{ where } \gamma_{k,t} := \sigma_{k+1} + \left(\frac{1}{2}\right)^{t-1}\sigma_k. \tag{15}$$

(15) holds for $t = 0$ by our induction hypothesis (14) for the k-th stage. Supposing it true for iteration t, we will show it for iteration $t+1$. The $(t+1)^{\text{th}}$ iterate is given by:

$$\boldsymbol{X}_{k,t+1} = P_k\left(\mathbf{M} + \beta\mathbf{H}\right), \text{ where } \mathbf{H} = \frac{1}{\beta}\left(\mathbf{E} - \frac{1}{p}P_{\Omega_{k,t}}(\mathbf{E})\right), \mathbf{E} = \boldsymbol{X}_{k,t} - \mathbf{M}, \tag{16}$$

$p = \frac{\mathbb{E}\left[|\Omega_{k,t}|\right]}{n^2} = \frac{C\alpha\mu^4 r^4 \log^2 n}{n}$, and $\beta = \frac{2\sqrt{n}\|\mathbf{E}\|_\infty}{\sqrt{p}} \leq \frac{8\mu^2 r^2 \gamma_{k,t}}{\sqrt{n \cdot p}}$. Our hypothesis on the sample size tells us that $\beta \leq \frac{8\gamma_{k,t}}{\sqrt{C\alpha}\log n}$ and Lemma 8 tells us that $\mathbf{H}$ satisfies the hypothesis of Lemma 2. So we have:

$$\|\boldsymbol{X}_{k,t+1} - \mathbf{M}\|_\infty < \frac{\mu^2 r^2}{n}\left(\sigma_{k+1} + 15\beta\sqrt{\alpha}\log n\right) < \frac{\mu^2 r^2}{n}\left(\sigma_{k+1} + \frac{1}{9}\gamma_{k,t}\right) \leq \frac{10\mu^2 r^2}{9n}\gamma_{k,t+1}.$$

This proves (15). Hence, after $O(\log n)$ steps, we have:

$$\|\boldsymbol{X}_{k,\log n} - \mathbf{M}\|_\infty < \frac{10\mu^2 r^2}{9n}\left(\frac{\sigma_k}{n^3} + \sigma_{k+1}\right). \tag{17}$$

Step II: Let $\boldsymbol{G} := \boldsymbol{X}_{k,\log n} - \frac{1}{p}P_{\Omega_{k,\log n}}\left(\boldsymbol{X}_{k,\log n} - \mathbf{M}\right) = \mathbf{M} + \beta\mathbf{H}$ be the gradient update with notation as above. A standard perturbation argument (Lemmas 10 and 11) tells us that:

$$\|\boldsymbol{G} - \mathbf{M}\|_2 < 3\beta\sqrt{\alpha} \leq \frac{24\mu^2 r^2\sqrt{\alpha}\gamma_{k,\log n}}{\sqrt{np}} < \frac{1}{100}\left(\frac{\sigma_k}{n^3} + \sigma_{k+1}\right).$$

So if $\sigma_{k+1}(\boldsymbol{G}) > \frac{\sigma_k(\boldsymbol{G})}{n^3}$, then we have $\sigma_{k+1} > \frac{9\sigma_k}{10n^3}$. Since we move on to the next stage with $\boldsymbol{X}_{k+1,0} = \boldsymbol{X}_{k,\log n}$, (17) tells us that:

$$\|\boldsymbol{X}_{k+1,0} - \mathbf{M}\|_\infty = \|\boldsymbol{X}_{k,\log n} - \mathbf{M}\|_\infty \leq \frac{10\mu^2 r^2}{9n}\left(\frac{\sigma_k}{n^3} + \sigma_{k+1}\right) \leq \frac{2\mu^2 r^2}{n}\left(2\sigma_{k+1}\right),$$

showing the invariant for the $(k+1)^{\text{th}}$ stage.

Step III: On the other hand, if $\sigma_{k+1}(\boldsymbol{G}) \leq \frac{\sigma_k(\boldsymbol{G})}{n^3}$, Lemmas 10 and 11 tell us that $\sigma_{k+1} \leq \frac{11\sigma_k}{10n^3}$. So, using Lemma 3 with $T = \log \frac{n\sigma_k}{\epsilon}$ iterations, we obtain:

$$\left\|\boldsymbol{X}_{k,T+\log n} - P_k\left(\mathbf{M}\right)\right\|_F \leq \max\left(\epsilon, \frac{2}{p}\left\|\mathbf{M} - P_k\left(\mathbf{M}\right)\right\|_F\right). \tag{18}$$

If $\epsilon > \frac{2}{p}\left\|\mathbf{M} - P_k\left(\mathbf{M}\right)\right\|_F$, then we have:

$$\left\|\boldsymbol{X}_{k,T+\log n} - \mathbf{M}\right\|_F \leq \left\|\boldsymbol{X}_{k,T+\log n} - P_k\left(\mathbf{M}\right)\right\|_F + \left\|\mathbf{M} - P_k\left(\mathbf{M}\right)\right\|_F \leq 2\epsilon.$$

On the other hand, if $\epsilon \leq \frac{2}{p}\left\|\mathbf{M} - P_k\left(\mathbf{M}\right)\right\|_F$, then we have:

$$\begin{aligned}\left\|\boldsymbol{X}_{k,T+\log n} - \mathbf{M}\right\|_\infty &\leq \left\|\boldsymbol{X}_{k,T+\log n} - P_k\left(\mathbf{M}\right)\right\|_F + \left\|\mathbf{M} - P_k\left(\mathbf{M}\right)\right\|_\infty \\ &\leq \frac{2}{p}\left\|\mathbf{M} - P_k\left(\mathbf{M}\right)\right\|_F + \frac{\mu^2 r^2 \sigma_{k+1}}{n} \leq \left(2\sqrt{r}n + \frac{\mu^2 r^2}{n}\right)\sigma_{k+1} \\ &\leq \frac{2\mu^2 r^2}{n}\left(\left(\frac{1}{2}\right)^{\log \frac{\sigma_k}{n^2\sigma_{k+1}}}\sigma_k + \sigma_{k+1}\right).\end{aligned} \tag{19}$$

Step IV: Using (19) and "fresh samples" analysis as in Step I (in particular (15)), we have:

$$\left\|\boldsymbol{X}_{k,T+2\log n} - \mathbf{M}\right\|_\infty \leq \frac{10\mu^2 r^2}{9n}\left(\left(\frac{1}{2}\right)^{\log \frac{\sigma_k}{\sigma_{k+1}}}\sigma_k + \sigma_{k+1}\right) \leq \frac{2\mu^2 r^2}{n}\left(2\sigma_{k+1}\right),$$

which establishes the invariant for the $(k+1)^{\text{th}}$ stage.

Combining the invariant (14) with the exit condition after Step IV, we have: $\|\widehat{\mathbf{M}} - \mathbf{M}\|_F \leq \epsilon$ where $\widehat{\mathbf{M}}$ is the output of the algorithm. As there are r stages, and in each stage, we need $2\log n$ sets of samples of size $O(pn^2)$. Hence, the total samplexity is $|\Omega| = O\left(\alpha\mu^4 r^5 n \log^3 n\right)$. Similarly, total computation complexity is $O\left(\alpha\mu^4 r^7 n \log^3 n \log(\|M\|_F/\epsilon)\right)$. ■

5. Discussion and Conclusions

In this paper, we proposed a fast projected gradient descent based algorithm for solving the matrix completion problem. The algorithm runs in time $O\left(nr^7 \log^3 n \log 1/\epsilon\right)$, with a sample complexity of $O\left(nr^5 \log^3 n\right)$. To the best our knowledge, this is the first near linear time algorithm for exact matrix completion with sample complexity independent of ϵ and condition number of $\mathbf{M}$.

The first key idea behind our result is to use the ℓ_∞ norm as a potential function which entails bounding all the terms of an explicit Taylor series expansion. The second key idea is an extension of the Davis-Kahan theorem, that provides perturbation bound for the best rank-k approximation of a matrix with good eigen-gap. We believe both these techniques may find applications in other contexts.

Designing an efficient algorithm with information-theoretic optimal sample complexity $|\Omega| = O\left(nr \log n\right)$ is still open; our result is suboptimal by a factor of $r^4 \log^2 n$ and nuclear norm approach is suboptimal by a factor of $\log n$. Another interesting direction in this area is to design optimal algorithms that can handle sampling distributions that are widely observed in practice, such as the power law distribution (Meka et al., 2009).

References

Robert Bell and Yehuda Koren. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *ICDM*, pages 43–52, 2007. doi: 10.1109/ICDM.2007.90.

Rajendra Bhatia. *Matrix Analysis*. Springer, 1997.

Srinadh Bhojanapalli and Prateek Jain. Universal matrix completion. In *ICML*, 2014.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, December 2009.

Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2009.

Yudong Chen, Srinadh Bhojanapalli, Sujay Sanghavi, and Rachel Ward. Coherent matrix completion. In *Proceedings of The 31st International Conference on Machine Learning*, pages 674–682, 2014.

László Erdos, Antti Knowles, Horng-Tzer Yau, and Jun Yin. Spectral statistics of Erdos–Rényi graphs I: Local semicircle law. *The Annals of Probability*, 41(3B):2279–2375, 2013.

David F. Gleich and Lek-Heng Lim. Rank aggregation via nuclear norm minimization. In *KDD*, pages 60–68, 2011.

Moritz Hardt. Understanding alternating minimization for matrix completion. In *FOCS*, 2014.

Moritz Hardt and May Wootters. Fast matrix completion without the condition number. In *COLT*, 2014.

Moritz Hardt, Raghu Meka, Prasad Raghavendra, and Benjamin Weitz. Computational limits for matrix completion. In *COLT*, pages 703–725, 2014.

Cho-Jui Hsieh, Kai-Yang Chiang, and Inderjit S. Dhillon. Low rank modeling of signed networks. In *KDD*, pages 507–515, 2012.

Prateek Jain, Raghu Meka, and Inderjit S. Dhillon. Guaranteed rank minimization via singular value projection. In *NIPS*, pages 937–945, 2010.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th annual ACM Symposium on theory of computing*, pages 665–674. ACM, 2013.

Raghunandan H. Keshavan. Efficient algorithms for collaborative filtering. Phd Thesis, Stanford University, 2012.

Raghunandan H. Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

Raghu Meka, Prateek Jain, and Inderjit S. Dhillon. Matrix completion from power-law distributed samples. In *NIPS*, 2009.

Benjamin Recht. A simple approach to matrix completion. *JMLR*, 2009.

Benjamin Recht and Christopher Ré. Parallel stochastic gradient algorithms for large-scale matrix completion. *Mathematical Programming Computation*, 5(2):201–226, 2013.

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

Appendix A. Preliminaries and Notations for Proofs

The following lemma shows that wlog we can assume $\mathbf{M}$ to be a symmetric matrix. A similar result is given in Section D of Hardt (2014).

Lemma 6 *Let $\mathbf{M} \in \mathbb{R}^{n_1 \times n_2}$ and $\Omega \subseteq [n_1] \times [n_2]$ satisfy Assumption 1 and 2, respectively. Then, there exists a symmetric $\widetilde{\mathbf{M}} \in \mathbb{R}^{n \times n}$, $n = n_1 + n_2$, s.t. $\widetilde{\mathbf{M}}$ is of rank-$2r$, incoherence of $\widetilde{\mathbf{M}}$ is twice the incoherence of $\mathbf{M}$. Moreover, there exists $|\widetilde{\Omega}| \subseteq [n] \times [n]$ that satisfy Assumption 2, $P_{\widetilde{\Omega}}(\widetilde{\mathbf{M}})$ is efficiently computable, and the output of a SVP update (3) with $P_\Omega(\mathbf{M})$ can also be obtained by the SVP update of $P_{\widetilde{\Omega}}(\widetilde{\mathbf{M}})$.*

Proof [Proof of Lemma 6] Define the following symmetric matrix from $\mathbf{M}$ using a dilation technique:

$$\widetilde{\mathbf{M}} = \begin{bmatrix} 0 & \mathbf{M} \\ \mathbf{M}^\top & 0 \end{bmatrix}.$$

Note that the rank of $\widetilde{\mathbf{M}}$ is $2 \cdot r$ and the incoherence of $\widetilde{\mathbf{M}}$ is bounded by $(n_1 + n_2)/n_2\mu$ (assume $n_1 \leq n_2$). Note that if $n_2 > n_1$, then we can split the columns of $\mathbf{M}$ in blocks of size n_1 and apply the argument separately to each block.

Now, we can split Ω to generate samples from $\mathbf{M}$ and $\mathbf{M}^T$, and then augment redundant samples from the 0 part above to obtain $\widetilde{\Omega} = [n] \times [n]$.

Moreover, if we run the SVP update (3) with input $\widetilde{\mathbf{M}}$, $\widetilde{\boldsymbol{X}}$ and $\widetilde{\Omega}$, an easy calculation shows that the iterates satisfy:

$$\widetilde{\boldsymbol{X}}_+ = \begin{bmatrix} 0 & \boldsymbol{X}_+ \\ \boldsymbol{X}_+{}^\top & 0 \end{bmatrix},$$

where $\boldsymbol{X}_+$ is the output of (3) with input $\mathbf{M}$, $\boldsymbol{X}$, and Ω. That is, a convergence result for $\widetilde{\boldsymbol{X}}_+$ would imply a convergence result for $\boldsymbol{X}_+$ as well. ■

For the remaining sections, we assume (wlog) that $\mathbf{M} \in \mathbb{R}^{n \times n}$ is symmetric and $\mathbf{M} = \boldsymbol{U}^* \Sigma \boldsymbol{U}^{*\top}$ is the eigenvalue decomposition (EVD) of $\mathbf{M}$. Also, unless specified, σ_i denotes the i-th eigenvalue of $\mathbf{M}$.

Appendix B. Proof of Lemma 2

Recall that we assume (wlog) that $\mathbf{M} \in \mathbb{R}^{n \times n}$ is symmetric and $\mathbf{M} = \boldsymbol{U}^* \boldsymbol{\Sigma} \boldsymbol{U}^{*\top}$ is the eigenvalue decomposition (EVD) of $\mathbf{M}$. Also, the goal is to bound $\|\boldsymbol{X}_+ - \mathbf{M}\|_\infty$, where $\boldsymbol{X}_+ = P_k(\mathbf{M} + \beta\mathbf{H})$ and $\mathbf{H}$ is such that it satisfies the following definition:

Definition 7 $\mathbf{H}$ *is a symmetric matrix with each of its elements drawn independently, satisfying the following moment conditions:*

$$\mathbb{E}[h_{ij}] = 0, \quad |h_{ij}| < 1, \quad \mathbb{E}\left[|h_{ij}|^k\right] \leq \tfrac{1}{n},$$

for $i, j \in [n]$ and $2 \leq k \leq 2 \log n$.

That is, we wish to understand $\|\boldsymbol{X}_+ - \mathbf{M}\|_\infty$ under perturbation $\mathbf{H}$. To this end, we first present a few lemmas that analyze how $\mathbf{H}$ is obtained in the context of our St-SVP algorithm and also bounds certain key quantities related to $\mathbf{H}$. We then present a few technical lemmas that are helpful for our proof of Lemma 2. The detailed proof of the lemma is given in Section 3.1. See Section B.3 for proofs of the technical lemmas.

B.1. Results for H

Recall that the SVP update (3) is given by: $\boldsymbol{X}_+ = P_k(\boldsymbol{X} - \frac{1}{p}P_\Omega(\boldsymbol{X} - \mathbf{M})) = P_k(\mathbf{M} + \mathbf{H})$ where $\mathbf{H} = \boldsymbol{E} - \frac{1}{p}P_\Omega(\boldsymbol{E})$ and $\boldsymbol{E} = \boldsymbol{X} - \mathbf{M}$. Our first lemma shows that matrices of the form $\boldsymbol{E} - \frac{1}{p}P_\Omega(\boldsymbol{E})$, scaled appropriately, satisfy Definition 7, i.e., satisfies the assumption of Lemma 2.

Lemma 8 *Let $\boldsymbol{A}$ be a symmetric $n \times n$ matrix. Suppose $\Omega \subseteq [n] \times [n]$ is obtained by sampling each element with probability $p \in \left[\frac{1}{4n}, 0.5\right]$. Then the matrix*

$$\boldsymbol{B} := \frac{\sqrt{p}}{2\sqrt{n}\,\|\boldsymbol{A}\|_\infty}\left(\boldsymbol{A} - \frac{1}{p}P_\Omega\left(\boldsymbol{A}\right)\right)$$

satisfies Definition 7.

We now present a critical lemma for our proof which bounds $\|H^a u\|_\infty$ for $2 \le a \le \log n$. Note that the entries of H^a can be dependent on each other, hence we cannot directly apply standard tail bounds. Our proof follows along very similar lines to Lemma 6.5 of Erdos et al. (2013); see Appendix D for a detailed proof.

Lemma 9 *Suppose $\widehat{\mathbf{H}}$ satisfies Definition 7. Fix $1 \le a \le \log n$. Let $\boldsymbol{e}_r$ denote the r^{th} standard basis vector. Then, for any fixed vector $\boldsymbol{u}$, we have:*

$$\left|\left\langle \boldsymbol{e}_r, \widehat{\mathbf{H}}^a \boldsymbol{u}\right\rangle\right| \le (c\log n)^a \|\boldsymbol{u}\|_\infty \;\; \forall\, r \in [n],$$

with probability greater than $1 - n^{1-2\log\frac{c}{4}}$.

Next, we bound $\|H\|_2$ using matrix Bernstein inequality by Tropp (2012); see Appendix B.3 for a proof.

Lemma 10 *Suppose $\mathbf{H}$ satisfies Definition 7. Then, w.p. $\ge 1 - 1/n^{10+\log\alpha}$, we have: $\|\mathbf{H}\|_2 \le 3\sqrt{\alpha}$.*

B.2. Technical Lemmas useful for Proof of Lemma 2

In this section, we present the technical lemmas used by our proof of Lemma 2.

First, we present the well known Weyl's perturbation inequality Bhatia (1997):

Lemma 11 *Suppose $\boldsymbol{B} = \boldsymbol{A} + \boldsymbol{N}$. Let $\lambda_1, \cdots, \lambda_n$ and $\sigma_1, \cdots, \sigma_n$ be the eigenvalues of $\boldsymbol{B}$ and $\boldsymbol{A}$ respectively. Then we have:*

$$|\lambda_i - \sigma_i| \le \|\boldsymbol{N}\|_2 \;\; \forall\, i \in [n].$$

The below given lemma bounds the ℓ_∞ norm of an appropriate incoherent matrix using its ℓ_2 norm.

Lemma 12 *Suppose* $\mathbf{M}$ *is a symmetric matrix with size* n *and satisfying Assumption 1. For any symmetric matrix* $\boldsymbol{B} \in \mathbb{R}^{n\times n}$*, we have:*

$$\|\mathbf{M}\boldsymbol{B}\mathbf{M} - \mathbf{M}\|_\infty \leq \frac{\mu^2 r}{n} \|\mathbf{M}\boldsymbol{B}\mathbf{M} - \mathbf{M}\|_2 .$$

Next, we present a natural perturbation lemma that bounds the spectral norm distance of A to $AB^{-1}A$ where $B = P_k(A+E)$ and E is a perturbation to A.

Lemma 13 *Let* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *be a symmetric matrix with eigenvalues* $\beta_1, \cdots, \beta_n$*, where* $|\beta_1| \geq \cdots \geq |\beta_n|$*. Let* $\boldsymbol{W} = \boldsymbol{A} + \boldsymbol{E}$ *be a perturbation of* $\boldsymbol{A}$*, where* $\boldsymbol{E}$ *is a symmetric matrix with* $\|\boldsymbol{E}\|_2 < \frac{|\beta_k|}{2}$*. Also, let* $P_k(\boldsymbol{W}) = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ *be the eigenvalue decomposition of the best rank-k approximation of* $\boldsymbol{W}$*. Then,* $\boldsymbol{\Lambda}^{-1}$ *exists. Furthermore, we have:*

$$\left\|\boldsymbol{A} - \boldsymbol{A}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\boldsymbol{A}\right\|_2 \leq |\beta_{k+1}| + 5\|\boldsymbol{E}\|_2 , \text{ and}$$
$$\left\|\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Lambda}^{-a}\boldsymbol{U}^\top\boldsymbol{A}\right\|_2 \leq 4\left(\frac{|\beta_k|}{2}\right)^{-a+2} \quad \forall\, a \geq 2.$$

B.3. Proofs of Technical Lemmas from Section B.1, Section B.2

Proof [Proof of Lemma 8] Since $(P_\Omega(\boldsymbol{A}))_{ij}$ is an unbiased estimate of $\boldsymbol{A}_{ij}$, we see that $\mathbb{E}[\boldsymbol{B}_{ij}] = 0$. For $k \geq 2$, we have:

$$\mathbb{E}\left[|\boldsymbol{B}_{ij}|^k\right] = \left(\frac{\sqrt{p}\boldsymbol{A}_{ij}}{2\sqrt{n}\|\boldsymbol{A}\|_\infty}\right)^k \left(p\left(\frac{1}{p}-1\right)^k + (1-p)\right) \leq \left(\frac{p}{2n}\right)^{\frac{k}{2}} \cdot \frac{2}{p^{k-1}} \leq \frac{1}{n(np)^{\frac{k}{2}-1}} \leq \frac{1}{n}.$$

■

Proof [Proof of Lemma 10]

Note that, $\mathbf{H} = \sum_{ij} h_{ij}\boldsymbol{e}_i\boldsymbol{e}_j{}^\top = \sum_{i\leq j} \boldsymbol{G}_{ij}$ where $\boldsymbol{G}_{ij} = h_{ij}\frac{\mathbb{1}_{\{i\neq j\}}+1}{2}\left(\boldsymbol{e}_i\boldsymbol{e}_j{}^\top + \boldsymbol{e}_j\boldsymbol{e}_i{}^\top\right)$. Now, $\mathbb{E}[\boldsymbol{G}_{ij}] = 0$, $\max_{ij}\|\boldsymbol{G}_{ij}\|_2 = 2$, and,

$$\left\|\mathbb{E}\left[\boldsymbol{G}_{ij}\boldsymbol{G}_{ij}^\top\right]\right\|_2 = \left\|\mathbb{E}\left[\sum_{ij} h_{ij}^2\boldsymbol{e}_i\boldsymbol{e}_i{}^\top\right]\right\|_2 = \max_i \sum_j \mathbb{E}\left[h_{ij}^2\right] \leq 1.$$

The lemma now follows using matrix Bernstein inequality (Lemma 21). ■

Proof [Proof of Lemma 12] Let $\mathbf{M} = \boldsymbol{U}^*\boldsymbol{\Sigma}\boldsymbol{U}^{*\top}$ be the eigenvalue decomposition $\mathbf{M}$. We have:

$$\begin{aligned}
\|\mathbf{M}\boldsymbol{B}\mathbf{M} - \mathbf{M}\|_\infty &= \max_{i,j} \boldsymbol{e}_i{}^\top(\mathbf{M}\boldsymbol{B}\mathbf{M} - \mathbf{M})\boldsymbol{e}_j \\
&= \max_{i,j} \boldsymbol{e}_i{}^\top\left(\boldsymbol{U}^*\boldsymbol{\Sigma}\boldsymbol{U}^\top\boldsymbol{B}\boldsymbol{U}^*\boldsymbol{\Sigma}\boldsymbol{U}^{*\top} - \boldsymbol{U}^*\boldsymbol{\Sigma}\boldsymbol{U}^{*\top}\right)\boldsymbol{e}_j \\
&\leq \left(\max_i \left\|\boldsymbol{e}_i{}^\top\boldsymbol{U}^*\right\|_2\right)\left\|\boldsymbol{\Sigma}\boldsymbol{U}^{*\top}\boldsymbol{B}\boldsymbol{U}^*\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\right\|_2\left(\max_j \boldsymbol{U}^{*\top}\boldsymbol{e}_j\right) \\
&\overset{(\zeta_1)}{\leq} \frac{\mu^2 r}{n}\left\|\boldsymbol{U}^*\left(\boldsymbol{\Sigma}\boldsymbol{U}^{*\top}\boldsymbol{B}\boldsymbol{U}^*\boldsymbol{\Sigma} - \boldsymbol{\Sigma}\right)\boldsymbol{U}^{*\top}\right\|_2 = \frac{\mu^2 r}{n}\|\mathbf{M}\boldsymbol{B}\mathbf{M} - \mathbf{M}\|_2 ,
\end{aligned}$$

where (ζ_1) follows from the incoherence of $\mathbf{M}$. ■

Proof [Proof of Lemma 13] Let $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top + \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{U}}^\top$ be the eigenvalue decomposition of $\boldsymbol{W}$ where $\widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{U}}^\top$ corresponds to the botton $n-k$ singular components. Since $P_k(\boldsymbol{W}) = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$, we see that $|\lambda_k| \geq \left|\widetilde{\lambda}_i\right|$.

From Lemma 11, we have:

$$|\lambda_i - \beta_i| \leq \|\boldsymbol{E}\|_2, \ \forall\, i \in [k], \quad \text{and,} \quad \left|\widetilde{\lambda}_i - \beta_{k+i}\right| \leq \|\boldsymbol{E}\|_2, \ \forall\, i \in [n-k]. \tag{20}$$

Since $\|\boldsymbol{E}\|_2 \leq \frac{\beta_k}{2}$, we see that

$$|\lambda_k| \geq |\beta_k|/2 > 0. \tag{21}$$

Hence, we conclude that $\boldsymbol{\Lambda} \in \mathbb{R}^{k\times k}$ is invertible proving the first claim of the lemma.

Using the eigenvalue decomposition of $\boldsymbol{W}$, we have the following expansion:

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\boldsymbol{A} - \boldsymbol{A} &= \left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top + \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{U}}^\top - \boldsymbol{E}\right)\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top + \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{U}}^\top - \boldsymbol{E}\right) - \boldsymbol{A} \\
&= \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top - \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{E} - \boldsymbol{E}\boldsymbol{U}\boldsymbol{U}^\top + \boldsymbol{E}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\boldsymbol{E} - \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top - \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{U}}^\top + \boldsymbol{E} \\
&= -\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{E} - \boldsymbol{E}\boldsymbol{U}\boldsymbol{U}^\top + \boldsymbol{E}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\boldsymbol{E} - \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{U}}^\top + \boldsymbol{E}.
\end{aligned} \tag{22}$$

Applying triangle inequality and using $\|\boldsymbol{B}\boldsymbol{C}\|_2 \leq \|\boldsymbol{B}\|_2 \|\boldsymbol{C}\|_2$, we get:

$$\left\|\boldsymbol{A} - \boldsymbol{A}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\boldsymbol{A}^\top\right\|_2 \leq 3\|\boldsymbol{E}\|_2 + \frac{\|\boldsymbol{E}\|_2^2}{|\lambda_k|} + \left|\widetilde{\lambda}_1\right|.$$

Using the above inequality with (21), we obtain:

$$\left\|\boldsymbol{A} - \boldsymbol{A}\boldsymbol{U}\boldsymbol{\Lambda}^{-1}\boldsymbol{U}^\top\boldsymbol{A}^\top\right\|_2 \leq |\beta_{k+1}| + 5\|\boldsymbol{E}\|_2.$$

This proves the second claim of the lemma.

Now, similar to (22), we have:

$$\begin{aligned}
\boldsymbol{A}\boldsymbol{U}\boldsymbol{\Lambda}^{-a}\boldsymbol{U}^\top\boldsymbol{A} &= \left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top + \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{U}}^\top - \boldsymbol{E}\right)\boldsymbol{U}\boldsymbol{\Lambda}^{-a}\boldsymbol{U}^\top\left(\boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top + \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Lambda}}\widetilde{\boldsymbol{U}}^\top - \boldsymbol{E}\right) \\
&= \boldsymbol{U}\boldsymbol{\Lambda}^{-a+2}\boldsymbol{U}^\top - \boldsymbol{U}\boldsymbol{\Lambda}^{-a+1}\boldsymbol{U}^\top\boldsymbol{E} - \boldsymbol{E}\boldsymbol{U}\boldsymbol{\Lambda}^{-a+1}\boldsymbol{U}^\top + \boldsymbol{E}\boldsymbol{U}\boldsymbol{\Lambda}^{-a}\boldsymbol{U}^\top\boldsymbol{E}.
\end{aligned}$$

The last claim of the lemma follows by using triangle inequality and (21) in the above equation. ■

Appendix C. Proof of Lemma 3

We now present a proof of Lemma 3 that show decrease in the Frobenius norm of the error matrix, despite using same samples in each iteration. In order to state our proof, we will first introduce certain notations and provide a few perturbation results that might be of independent interest. Then, in next subsection, we will present a detailed proof of Lemma 3. Finally, in Section C.3, we present proofs of the technical lemmas given below.

C.1. Notations and Technical Lemmas

Recall that we assume (wlog) that $\mathbf{M} \in \mathbb{R}^{n\times n}$ is symmetric and $\mathbf{M} = \boldsymbol{U}^*\boldsymbol{\Sigma}\boldsymbol{U}^{*\top}$ is the eigenvalue decomposition (EVD) of $\mathbf{M}$.

In order to state our first supporting lemma, we will introduce the concept of tangent spaces of matrices Bhatia (1997).

Definition 14 *Let $\boldsymbol{A}$ be a matrix with EVD (eigenvalue decomposition) $\boldsymbol{U}^*\boldsymbol{\Sigma}\boldsymbol{U}^{*\top}$. The following space of matrices is called the tangent space of $\boldsymbol{A}$:*

$$\mathcal{T}(\boldsymbol{A}) := \left\{\boldsymbol{U}^*\boldsymbol{\Lambda}_0\boldsymbol{U}^{*\top} + \boldsymbol{U}^*\boldsymbol{\Lambda}_1\boldsymbol{U}^{\top} + \boldsymbol{U}\boldsymbol{\Lambda}_2\boldsymbol{U}^{*\top}\right\},$$

where $\boldsymbol{U} \in \mathbb{R}^{n\times n}, \boldsymbol{U}^T\boldsymbol{U} = I$, and $\boldsymbol{\Lambda}_0, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2$ are all diagonal matrices.

That is, if $\boldsymbol{A} = \boldsymbol{U}^*\boldsymbol{\Sigma}\boldsymbol{U}^{*\top}$ is the EVD of $\boldsymbol{A}$, then any matrix $\boldsymbol{B}$ can be decomposed into four mutually orthogonal terms as

$$\boldsymbol{B} = \boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{B}\boldsymbol{U}^*\boldsymbol{U}^{*\top} + \boldsymbol{U}^*\boldsymbol{U}^{*\top}\boldsymbol{B}\boldsymbol{U}_\perp^*\boldsymbol{U}_\perp^{*\top} + \boldsymbol{U}_\perp^*\boldsymbol{U}_\perp^{*\top}\boldsymbol{B}\boldsymbol{U}^*\boldsymbol{U}^{*\top} + \boldsymbol{U}_\perp^*\boldsymbol{U}_\perp^{*\top}\boldsymbol{B}\boldsymbol{U}_\perp^*\boldsymbol{U}_\perp^{*\top}, \tag{23}$$

where $\boldsymbol{U}_\perp^*$ is a basis of the orthogonal space of $\boldsymbol{U}^*$. The first three terms above are in $\mathcal{T}(\boldsymbol{A})$ and the last term is in $\mathcal{T}(\boldsymbol{A})^\perp$. We let $\mathcal{P}_{\mathcal{T}(\boldsymbol{A})}$ and $\mathcal{P}_{\mathcal{T}(\boldsymbol{A})^\perp}$ denote the projection operators onto $\mathcal{T}(\boldsymbol{A})$ and $\mathcal{T}(\boldsymbol{A})^\perp$ respectively.

Lemma 15 *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be two symmetric matrices. Suppose further that $\boldsymbol{B}$ is rank-k. Then, we have:*

$$\left\|\mathcal{P}_{\mathcal{T}(\boldsymbol{A})^\perp}(\boldsymbol{B})\right\|_F \leq \frac{\|\boldsymbol{A}-\boldsymbol{B}\|_F^2}{\sigma_k(\boldsymbol{B})}.$$

Next, we present a few technical lemmas related to norm of $M - P_\Omega(M)$:

Lemma 16 *Let M, Ω be as given in Lemma 3 and let $p = |\Omega|/n^2$ be the sampling probability. Then, For every $r \times r$ matrix $\widehat{\boldsymbol{\Sigma}}$, we have $(w.p. \geq 1 - n^{-10-\alpha})$:*

$$\left\|\left(\boldsymbol{U}^*\widehat{\boldsymbol{\Sigma}}\boldsymbol{U}^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{U}^*\widehat{\boldsymbol{\Sigma}}\boldsymbol{U}^{*\top}\right)\right)\boldsymbol{U}^*\right\|_F \leq \frac{1}{40}\left\|\widehat{\boldsymbol{\Sigma}}\right\|_F.$$

Lemma 17 *Let M, Ω, p be as given in Lemma 3. Then, for every $i, j \in [r]$, we have $(w.p. \geq 1 - n^{-10-\alpha})$:*

$$\left\|\boldsymbol{u}_j^*\boldsymbol{u}_i^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{u}_j^*\boldsymbol{u}_i^{*\top}\right)\right\|_2 < \frac{1}{40r\sqrt{r}}.$$

Lemma 18 *Let M, Ω, p be as given in Lemma 3. Then, for every $i, j \in [r]$ and $s \in [n]$, we have $(w.p. \geq 1 - n^{-10-\alpha})$:*

$$\left|\left\langle \boldsymbol{u}_i^*, \boldsymbol{u}_j^*\right\rangle - \frac{1}{p}\sum_{(s,l)\in\Omega}\left(\boldsymbol{u}_i^*\right)_s\left(\boldsymbol{u}_j^*\right)_l\right| < \frac{1}{40r\sqrt{r}}.$$

C.2. Detailed Proof of Lemma 3

Let $\mathbf{E} := \boldsymbol{X} - P_k(\mathbf{M})$, $\mathbf{H} := \mathbf{E} - \frac{1}{p}P_\Omega(\mathbf{E})$ and $\boldsymbol{G} := \boldsymbol{X} - \frac{1}{p}P_\Omega(\boldsymbol{X} - \mathbf{M}) = P_k(\mathbf{M}) + \mathbf{H} - \frac{1}{p}P_\Omega(\mathbf{M} - P_k(\mathbf{M}))$. That is, $\boldsymbol{X}_+ = P_k(\boldsymbol{G})$.

For simplicity, *in this section*, we let $\mathbf{M} = \boldsymbol{U}^*\boldsymbol{\Sigma}\boldsymbol{U}^{*\top} + \boldsymbol{U}_\perp^*\underline{\boldsymbol{\Sigma}}\boldsymbol{U}_\perp^{*\top}$ denote the eigenvalue decomposition (EVD) of $\mathbf{M}$ with $P_k(\mathbf{M}) = \boldsymbol{U}^*\boldsymbol{\Sigma}\boldsymbol{U}^{*\top}$, and also let $\underline{\mathbf{M}} = \boldsymbol{U}_\perp^*\underline{\boldsymbol{\Sigma}}\boldsymbol{U}_\perp^{*\top}$. We also use the shorthand notation $\mathcal{T} := \mathcal{T}(P_k(\mathbf{M}))$.

Representing $\boldsymbol{X}$ in terms of its projection onto $\mathcal{T}$ and its complement, we have:

$$\boldsymbol{X} = \boldsymbol{U}^*\boldsymbol{\Lambda}_0\boldsymbol{U}^{*\top} + \boldsymbol{U}^*\boldsymbol{\Lambda}_1\boldsymbol{U}_\perp^{*\top} + \boldsymbol{U}_\perp^*\boldsymbol{\Lambda}_1{}^\top\boldsymbol{U}^{*\top} + \boldsymbol{U}_\perp^*\boldsymbol{\Lambda}_3\boldsymbol{U}_\perp^{*\top}, \tag{24}$$

and also conclude that:

$$\|\boldsymbol{\Sigma} - \boldsymbol{\Lambda}_0\|_F \le \|\boldsymbol{X} - P_k(\mathbf{M})\|_F, \quad \|\boldsymbol{\Lambda}_1\|_F \le \|\boldsymbol{X} - P_k(\mathbf{M})\|_F, \text{ and } \quad \|\boldsymbol{\Lambda}_3\|_F \le \frac{\|\boldsymbol{X} - P_k(\mathbf{M})\|_F}{n^2},$$

where the last conclusion follows from Lemma 15 and the hypothesis that $\|\boldsymbol{X} - P_k(\mathbf{M})\|_F < \frac{|\sigma_k|}{n^2}$. Using $\|\mathbf{E}\|_F \le \sigma_k/n^2$, we have:

$$\|\mathbf{H}\|_F \le \frac{2}{p}\|\mathbf{E}\|_F \le \frac{2}{p}\frac{\sigma_k}{n^2} \le \frac{\sigma_k}{8}, \text{ and,}$$
$$\left\|\frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right\|_F \le \frac{1}{p}\|\underline{\mathbf{M}}\|_F \le \frac{1}{p}\frac{\sigma_k}{n^2} \le \frac{\sigma_k}{8},$$

where we used the hypothesis that $\|\mathbf{M} - P_k(\mathbf{M})\|_F < \frac{\sigma_k}{n^2}$ in the second inequality.

The above bounds implies:

$$\left\|\mathcal{P}_\mathcal{T}\left(\mathbf{H} - \frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right)\right\|_F \le \left\|\mathbf{H} - \frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right\|_F \le \|\mathbf{H}\|_2 + \left\|\frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right\|_F \le \frac{\sigma_k}{4}. \tag{25}$$

Similarly,

$$\left\|\mathcal{P}_{\mathcal{T}^\perp}\left(\mathbf{H} - \frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right)\right\|_2 \le \frac{\sigma_k}{4}. \tag{26}$$

Since $\boldsymbol{X}_+ = P_k\left(P_k(\mathbf{M}) + \mathbf{H} - \frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right)$, using Lemma 4 with (25), (26), we have:

$$\begin{aligned}\|P_k(\mathbf{M}) - \boldsymbol{X}_+\|_F &= \left\|P_k\left(P_k(\mathbf{M}) + \mathcal{P}_{\mathcal{T}^\perp}\left(\mathbf{H} - \frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right)\right) - P_k\left(P_k(\mathbf{M}) + \mathbf{H} - \frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right)\right\|_F \\ &\le c\left\|\mathcal{P}_\mathcal{T}\left(\mathbf{H} - \frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right)\right\|_F.\end{aligned}$$

Now, using Claim 1, we have $\left\|\mathcal{P}_\mathcal{T}\left(\mathbf{H} - \frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right)\right\|_F < \frac{1}{10}\|P_k(\mathbf{M}) - \boldsymbol{X}\|_F + \frac{2}{\sqrt{p}}\|\underline{\mathbf{M}}\|_F$, which along with the above equation *establishes the lemma*. We now state and prove the claim bounding $\left\|\mathcal{P}_\mathcal{T}\left(\mathbf{H} - \frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right)\right\|_F$ that we used above to finish the proof.

Claim 1 *Assume notation defined in the section above. Then, we have:*

$$\left\|\mathcal{P}_\mathcal{T}\left(\mathbf{H} - \frac{1}{p}P_\Omega(\underline{\mathbf{M}})\right)\right\|_F < \frac{1}{10}\|P_k(\mathbf{M}) - \boldsymbol{X}\|_F + \frac{2}{\sqrt{p}}\|\underline{\mathbf{M}}\|_F.$$

Proof We first bound $\|\mathcal{P}_{\mathcal{T}}(\mathbf{H})\|_F$. Recalling that $P_k(\mathbf{M}) = \boldsymbol{U}^*\boldsymbol{\Sigma}\boldsymbol{U}^{*\top}$ is the EVD of $P_k(\mathbf{M})$, we have:

$$\|\mathcal{P}_{\mathcal{T}}(\mathbf{H})\|_F < 2\|\mathbf{H}\boldsymbol{U}^*\|_F.$$

Using (24), we have:

$$\begin{aligned}\mathbf{H}\boldsymbol{U}^* = &\left[\boldsymbol{U}^*(\boldsymbol{\Sigma}-\boldsymbol{\Lambda}_0)\boldsymbol{U}^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{U}^*(\boldsymbol{\Sigma}-\boldsymbol{\Lambda}_0)\boldsymbol{U}^{*\top}\right)\right]\boldsymbol{U}^* + \left[\boldsymbol{U}^*\boldsymbol{\Lambda}_1\boldsymbol{U}_\perp^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{U}^*\boldsymbol{\Lambda}_1\boldsymbol{U}_\perp^{*\top}\right)\right]\boldsymbol{U}^* \\ &+ \left[\boldsymbol{U}_\perp^*\boldsymbol{\Lambda}_2\boldsymbol{U}^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{U}_\perp^*\boldsymbol{\Lambda}_2\boldsymbol{U}^{*\top}\right)\right]\boldsymbol{U}^* + \left[\boldsymbol{U}_\perp^*\boldsymbol{\Lambda}_3\boldsymbol{U}_\perp^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{U}_\perp^*\boldsymbol{\Lambda}_3\boldsymbol{U}_\perp^{*\top}\right)\right]\boldsymbol{U}^*. \end{aligned} \tag{27}$$

Step I: To bound the first term in (27), we use Lemma 16 to obtain:

$$\left\|\left(\boldsymbol{U}^*(\boldsymbol{\Sigma}-\boldsymbol{\Lambda}_0)\boldsymbol{U}^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{U}^*(\boldsymbol{\Sigma}-\boldsymbol{\Lambda}_0)\boldsymbol{U}^{*\top}\right)\right)\boldsymbol{U}^*\right\|_F \leq \frac{1}{40}\|(\boldsymbol{\Sigma}-\boldsymbol{\Lambda}_0)\|_F \leq \frac{1}{40}\|\boldsymbol{X}-P_k(\mathbf{M})\|_F. \tag{28}$$

Step II: To bound the second term, we let $\boldsymbol{U} := \boldsymbol{U}_\perp^*\boldsymbol{\Lambda}_1{}^\top$, and proceed as follows:

$$\begin{aligned}\left\|\left[\boldsymbol{U}^*\boldsymbol{\Lambda}_1\boldsymbol{U}_\perp^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{U}^*\boldsymbol{\Lambda}_1\boldsymbol{U}_\perp^{*\top}\right)\right]\boldsymbol{u}_i^*\right\|_2 &= \left\|\sum_{j=1}^r\left(\boldsymbol{u}_j^*\boldsymbol{u}_j{}^\top - \frac{1}{p}P_\Omega\left(\boldsymbol{u}_j^*\boldsymbol{u}_j{}^\top\right)\right)\boldsymbol{u}_i^*\right\|_2 \\ &= \left\|\sum_{j=1}^r\left(\boldsymbol{u}_j^*\boldsymbol{u}_i^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{u}_j^*\boldsymbol{u}_i^{*\top}\right)\right)\boldsymbol{u}_j\right\|_2 \leq \sum_{j=1}^r\left\|\boldsymbol{u}_j^*\boldsymbol{u}_i^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{u}_j^*\boldsymbol{u}_i^{*\top}\right)\right\|_2\|\boldsymbol{u}_j\|_2 \\ &\overset{(\zeta_1)}{\leq} \frac{1}{40r\sqrt{r}}\sum_{j=1}^r\|\boldsymbol{u}_j\|_2 \leq \frac{1}{40\sqrt{r}}\left\|\boldsymbol{\Lambda}_1\boldsymbol{U}_\perp^{*\top}\right\|_F \leq \frac{1}{40\sqrt{r}}\|P_k(\mathbf{M})-\boldsymbol{X}\|_F,\end{aligned}$$

where (ζ_1) follows from Lemma 17. This means that we can bound the second term as:

$$\left\|\left[\boldsymbol{U}^*\boldsymbol{\Lambda}_1\boldsymbol{U}_\perp^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{U}^*\boldsymbol{\Lambda}_1\boldsymbol{U}_\perp^{*\top}\right)\right]\boldsymbol{U}^*\right\|_F \leq \frac{1}{40}\|P_k(\mathbf{M})-\boldsymbol{X}\|_F. \tag{29}$$

Step III: We now let $\boldsymbol{U} := \boldsymbol{U}_\perp^*\boldsymbol{\Lambda}_2$ and turn to bound the third term in (27). We have:

$$\begin{aligned}\left\|\left[\boldsymbol{U}_\perp^*\boldsymbol{\Lambda}_2\boldsymbol{U}^{*\top} - \frac{1}{p}P_\Omega\left(\boldsymbol{U}_\perp^*\boldsymbol{\Lambda}_2\boldsymbol{U}^{*\top}\right)\right]\boldsymbol{u}_i^*\right\|_2 &= \left\|\sum_{j=1}^r\langle\boldsymbol{u}_j^*,\boldsymbol{u}_i^*\rangle\boldsymbol{u}_j - \boldsymbol{u}_j\odot\boldsymbol{e}_{\langle\boldsymbol{u}_j^*,\boldsymbol{u}_i^*\rangle_\Omega}\right\|_2 \\ &= \left\|\sum_{j=1}^r\boldsymbol{u}_j\odot\left(\langle\boldsymbol{u}_j^*,\boldsymbol{u}_i^*\rangle\mathbb{1}-\boldsymbol{e}_{\langle\boldsymbol{u}_j^*,\boldsymbol{u}_i^*\rangle_\Omega}\right)\right\|_2 \overset{(\zeta_1)}{\leq} \frac{1}{40r\sqrt{r}}\sum_{j=1}^r\|\boldsymbol{u}_j\|_2 \\ &\leq \frac{1}{40\sqrt{r}}\|\boldsymbol{U}_\perp^*\boldsymbol{\Lambda}_2\|_F \leq \frac{1}{40\sqrt{r}}\|P_k(\mathbf{M})-\boldsymbol{X}\|_F,\end{aligned}$$

where $\mathbb{1}$ denotes the all ones vector, and $\boldsymbol{e}_{\langle \boldsymbol{u}_j^*, \boldsymbol{u}_i^* \rangle_\Omega}$ denotes a vector whose s^{th} coordinate is given by $\frac{1}{p}\sum_{l:(s,l)\in\Omega}\left(\boldsymbol{u}_j^*\right)_l (\boldsymbol{u}_i^*)_l$. Note that (ζ_1) follows from Lemma 18. So, we again have:

$$\left\| \left[\boldsymbol{U}_\perp^* \boldsymbol{\Lambda}_2 \boldsymbol{U}^{*\top} - \frac{1}{p} P_\Omega \left(\boldsymbol{U}_\perp^* \boldsymbol{\Lambda}_2 \boldsymbol{U}^{*\top} \right) \right] \boldsymbol{U}^* \right\|_F \leq \frac{1}{40} \left\| P_k\left(\mathbf{M}\right) - \boldsymbol{X} \right\|_F . \tag{30}$$

Step IV: To bound the last term in (27), we use Lemma 15 to conclude

$$\begin{aligned} \left\| \left[\boldsymbol{U}_\perp^* \boldsymbol{\Lambda}_3 \boldsymbol{U}_\perp^{*\top} - \frac{1}{p} P_\Omega \left(\boldsymbol{U}_\perp^* \boldsymbol{\Lambda}_3 \boldsymbol{U}_\perp^{*\top} \right) \right] \boldsymbol{U}^* \right\|_F &\leq \left\| \boldsymbol{U}_\perp^* \boldsymbol{\Lambda}_3 \boldsymbol{U}_\perp^{*\top} - \frac{1}{p} P_\Omega \left(\boldsymbol{U}_\perp^* \boldsymbol{\Lambda}_3 \boldsymbol{U}_\perp^{*\top} \right) \right\|_F \\ \leq \frac{2}{p} \left\| \boldsymbol{U}_\perp^* \boldsymbol{\Lambda}_3 \boldsymbol{U}_\perp^{*\top} \right\|_F = \frac{2}{p} \left\| \mathcal{P}_{\mathcal{T}^\perp}\left(\boldsymbol{X}\right) \right\|_F &\leq \frac{2}{p} \frac{\left\| P_k\left(\mathbf{M}\right) - \boldsymbol{X} \right\|_F^2}{\sigma_k(\boldsymbol{X})} \leq \frac{1}{40n} \left\| P_k\left(\mathbf{M}\right) - \boldsymbol{X} \right\|_F . \end{aligned} \tag{31}$$

Combining (28), (29), (30) and (31), we have:

$$\left\| \mathcal{P}_{\mathcal{T}}\left(\mathbf{H}\right) \right\|_F \leq \frac{1}{10} \left\| P_k\left(\mathbf{M}\right) - \boldsymbol{X} \right\|_F . \tag{32}$$

On the other hand, we trivially have:

$$\left\| \frac{1}{p} P_\Omega\left(\underline{\mathbf{M}}\right) \right\|_F \leq \frac{2}{p} \left\| \underline{\mathbf{M}} \right\|_F . \tag{33}$$

Claim now follows by combining (32) and (33). ■

C.3. Proofs of Technical Lemmas from Section C.1

Proof [Proof of Lemma 15] Let $B = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\top$ be EVD of B. Then, we have:

$$\begin{aligned} \left\| \mathcal{P}_{\mathcal{T}(\boldsymbol{A})^\perp}\left(\boldsymbol{B}\right) \right\|_F &= \left\| \boldsymbol{U}_\perp^* \boldsymbol{U}_\perp^{*\top} \boldsymbol{B} \boldsymbol{U}_\perp^* \boldsymbol{U}_\perp^{*\top} \right\|_F = \left\| \boldsymbol{U}_\perp^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \right\|_F = \left\| \boldsymbol{U}_\perp^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{\Lambda}^{-1} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \right\|_F \\ &\leq \left\| \boldsymbol{U}_\perp^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \right\|_F \left\| \boldsymbol{\Lambda}^{-1} \right\|_2 \left\| \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \right\|_F = \left\| \boldsymbol{A} - \boldsymbol{B} \right\|_F \left\| \boldsymbol{\Lambda}^{-1} \right\|_2 \left\| \boldsymbol{A} - \boldsymbol{B} \right\|_F \\ &\leq \frac{\left\| \boldsymbol{A} - \boldsymbol{B} \right\|_F^2}{\sigma_k(\boldsymbol{B})}. \end{aligned}$$

Hence Proved. ■

We will now prove Lemma 4, which is a natural extension of the Davis-Kahan theorem. In order to do so, we will first recall the Davis-Kahan theorem:

Theorem 19 (Theorem VII.3.1 of Bhatia (1997)) *Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be symmetric matrices. Let $S_1, S_2 \subseteq \mathbb{R}$ be subsets separated by ν. Let $\mathbf{E} = P_{\boldsymbol{A}}(S_1)$ and $\mathbf{F} = P_{\boldsymbol{B}}(S_2)$ be an orthonormal basis of the eigenvectors of $\boldsymbol{A}$ with eigenvalues in S_1 and that of the eigenvectors of $\boldsymbol{B}$ with eigenvalues in S_2 respectively. Then, we have:*

$$\left\| \mathbf{E}\mathbf{F} \right\|_2 \leq \frac{1}{\nu} \left\| \boldsymbol{A} - \boldsymbol{B} \right\|_2 , \quad \left\| \mathbf{E}\mathbf{F} \right\|_F \leq \frac{1}{\nu} \left\| \boldsymbol{A} - \boldsymbol{B} \right\|_F .$$

Proof [Proof of Lemma 4] Let $\boldsymbol{A} = \boldsymbol{U}^* \boldsymbol{\Sigma} \boldsymbol{U}^{*\top} + \boldsymbol{U}_\perp^* \widehat{\boldsymbol{\Sigma}} \boldsymbol{U}_\perp^{*\top}$ be the EVD of $\boldsymbol{A}$ with $P_k(\boldsymbol{A}) = \boldsymbol{U}^* \boldsymbol{\Sigma} \boldsymbol{U}^{*\top}$. Similarly, let $\boldsymbol{A} + \mathbf{E} = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top + \boldsymbol{U}_\perp \widehat{\boldsymbol{\Lambda}} \boldsymbol{U}_\perp^\top$ denote the EVD of $\boldsymbol{A} + \mathbf{E}$ with $P_k(\boldsymbol{A} + \mathbf{E}) = \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top$. Expanding $P_k(\boldsymbol{A} + \mathbf{E})$ into components along $\boldsymbol{U}^*$ and orthogonal to it, we have:

$$\boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top = \boldsymbol{U}^* \boldsymbol{U}^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}^* \boldsymbol{U}^{*\top} + \boldsymbol{U}_\perp^* \boldsymbol{U}_\perp^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}^* \boldsymbol{U}^{*\top} + \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \boldsymbol{U}_\perp^{*\top}.$$

Now,

$$\begin{aligned}
&\|P_k(\boldsymbol{A} + \mathbf{E}) - P_k(\boldsymbol{A})\|_F \\
&= \left\| \boldsymbol{U}^* \boldsymbol{U}^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}^* \boldsymbol{U}^{*\top} + \boldsymbol{U}_\perp^* \boldsymbol{U}_\perp^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}^* \boldsymbol{U}^{*\top} + \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \boldsymbol{U}_\perp^{*\top} - \boldsymbol{U}^* \boldsymbol{\Sigma} \boldsymbol{U}^{*\top} \right\|_F \\
&\leq \left\| \boldsymbol{U}^* \boldsymbol{U}^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}^* \boldsymbol{U}^{*\top} - \boldsymbol{U}^* \boldsymbol{\Sigma} \boldsymbol{U}^{*\top} \right\|_F + \left\| \boldsymbol{U}_\perp^* \boldsymbol{U}_\perp^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}^* \boldsymbol{U}^{*\top} \right\|_F + \left\| \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \boldsymbol{U}_\perp^{*\top} \right\|_F \\
&\leq \left\| \boldsymbol{U}^* \boldsymbol{U}^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}^* \boldsymbol{U}^{*\top} - \boldsymbol{U}^* \boldsymbol{\Sigma} \boldsymbol{U}^{*\top} \right\|_F + \left\| \boldsymbol{U}_\perp^* \boldsymbol{U}_\perp^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \right\|_F + \left\| \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \boldsymbol{U}_\perp^{*\top} \right\|_F \\
&= \left\| \boldsymbol{U}^* \boldsymbol{U}^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}^* \boldsymbol{U}^{*\top} - \boldsymbol{U}^* \boldsymbol{\Sigma} \boldsymbol{U}^{*\top} \right\|_F + 2 \left\| \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \boldsymbol{U}_\perp^{*\top} \right\|_F \\
&\leq \left\| \boldsymbol{U}^* \boldsymbol{U}^{*\top} \boldsymbol{U} \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}^* \boldsymbol{U}^{*\top} + \boldsymbol{U}^* \boldsymbol{U}^{*\top} \boldsymbol{U}_\perp \widehat{\boldsymbol{\Lambda}} \boldsymbol{U}_\perp^\top \boldsymbol{U}^* \boldsymbol{U}^{*\top} - \boldsymbol{U}^* \boldsymbol{\Sigma} \boldsymbol{U}^{*\top} \right\|_F \\
&\quad + \left\| \boldsymbol{U}^* \boldsymbol{U}^{*\top} \boldsymbol{U}_\perp \widehat{\boldsymbol{\Lambda}} \boldsymbol{U}_\perp^\top \boldsymbol{U}^* \boldsymbol{U}^{*\top} \right\|_F + 2 \left\| \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \right\|_F \\
&= \left\| \boldsymbol{U}^* \boldsymbol{U}^{*\top} \mathbf{E} \boldsymbol{U}^* \boldsymbol{U}^{*\top} \right\|_F + \left\| \boldsymbol{U}^{*\top} \boldsymbol{U}_\perp \widehat{\boldsymbol{\Lambda}} \boldsymbol{U}_\perp^\top \boldsymbol{U}^* \right\|_F + 2 \left\| \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \right\|_F \\
&\leq \|\mathbf{E}\|_F + \left\| \boldsymbol{U}^{*\top} \boldsymbol{U}_\perp \widehat{\boldsymbol{\Lambda}} \boldsymbol{U}_\perp^\top \boldsymbol{U}^* \right\|_F + 2 \left\| \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \right\|_F
\end{aligned} \tag{34}$$

Before going on to bound the terms in (34), let us make some observations. We first use Lemma 11 to conclude that

$$\frac{3}{4} |\sigma_i| \leq |\lambda_i| \leq \frac{5}{4} |\sigma_i|, \text{ and } \left| \widehat{\lambda}_{k+i} \right| \leq \frac{|\sigma_k|}{2}.$$

Applying Theorem 19 with $S_1 = \left[\frac{-|\sigma_k|}{2}, \frac{|\sigma_k|}{2} \right]$ and $S_2 = \left(-\infty, \frac{-3|\sigma_i|}{4} \right] \cup \left[\frac{3|\sigma_i|}{4}, \infty \right)$, with separation parameter $\nu = \frac{|\sigma_i|}{4}$, we see that

$$\left\| \boldsymbol{u}_i^\top \boldsymbol{U}_\perp^* \right\|_2 \leq \frac{4}{|\sigma_i|} \|\mathbf{E}\|_2, \text{ and} \tag{35}$$

$$\left\| \boldsymbol{U}_\perp^\top \boldsymbol{U}^* \right\|_F \leq \frac{4}{|\sigma_k|} \|\mathbf{E}\|_F. \tag{36}$$

We are now ready to bound the last two terms in the right hand side of (34). Firstly, we have:

$$\left\| \boldsymbol{U}^{*\top} \boldsymbol{U}_\perp \widehat{\boldsymbol{\Lambda}} \boldsymbol{U}_\perp^\top \boldsymbol{U}^* \right\|_F \leq \left\| \widehat{\boldsymbol{\Lambda}} \right\|_2 \left\| \boldsymbol{U}^{*\top} \boldsymbol{U}_\perp \right\|_2 \left\| \boldsymbol{U}_\perp^\top \boldsymbol{U}^* \right\|_F \leq \left| \widehat{\lambda}_{k+1} \right| \left\| \boldsymbol{U}_\perp^\top \boldsymbol{U}^* \right\|_F^2 \leq 2 \|\mathbf{E}\|_F,$$

where the last step follows from (36) and the assumption on $\|\mathbf{E}\|_F$. For the other term, we have:

$$\left\| \boldsymbol{\Lambda} \boldsymbol{U}^\top \boldsymbol{U}_\perp^* \right\|_F^2 = \sum_i \lambda_i^2 \left\| \boldsymbol{u}_i^\top \boldsymbol{U}_\perp^* \right\|_2^2 \leq \frac{25}{16} \sum_i \sigma_i^2 \frac{16 \|\mathbf{E}\|_2^2}{\sigma_i^2} = 25k \|\mathbf{E}\|_2^2,$$

where we used (35). Combining the above two inequalities with (34) proves the lemma. ■

Finally, we present proofs for Lemma 16, Lemma 17, Lemma 18.

Proof [Proof of Lemma 16] Using Theorem 1 by Bhojanapalli and Jain (2014), the followings $\forall \widehat{\Sigma}$ (w.p. $\geq 1 - n^{-10-\alpha}$):

$$\left\| \boldsymbol{U}^* \widehat{\boldsymbol{\Sigma}} \boldsymbol{U}^{*\top} - \frac{1}{p} P_\Omega \left(\boldsymbol{U}^* \widehat{\boldsymbol{\Sigma}} \boldsymbol{U}^{*\top} \right) \right\|_2 \leq \frac{\mu^2 r}{\sqrt{np}} \|\widehat{\boldsymbol{\Sigma}}\|_2 \leq \frac{1}{\sqrt{r \cdot C \cdot \alpha \log n}} \|\widehat{\boldsymbol{\Sigma}}\|_2.$$

Lemma now follows by using the assumed value of p in the above bound along with the fact that $\left(\boldsymbol{U}^* \widehat{\boldsymbol{\Sigma}} \boldsymbol{U}^{*\top} - \frac{1}{p} P_\Omega \left(\boldsymbol{U}^* \widehat{\boldsymbol{\Sigma}} \boldsymbol{U}^{*\top} \right) \right) \boldsymbol{U}^*$ is a rank-r matrix. ■

Proof [Proof of Lemma 17] Let $H = \frac{1}{\beta} \left(\boldsymbol{u}_j^* \boldsymbol{u}_i^{*\top} - \frac{1}{p} P_\Omega \left(\boldsymbol{u}_j^* \boldsymbol{u}_i^{*\top} \right) \right)$, where $\beta = \frac{2\mu^2 r}{\sqrt{n \cdot p}}$. Then, using Lemma 8, H satisfies the conditions of Definition 7. Lemma now follows by using Lemma 10 and using p as given in the lemma statement. ■

Proof [Proof of Lemma 18] Let $\delta_{ij} = \mathbb{I}[(i,j) \in \Omega]$. Then,

$$\left\langle \boldsymbol{u}_i^*, \boldsymbol{u}_j^* \right\rangle - \frac{1}{p} \sum_{(s,l) \in \Omega} \left(\boldsymbol{u}_i^* \right)_l \left(\boldsymbol{u}_j^* \right)_l = \sum_l (1 - \frac{\delta_{sl}}{p}) \left(\boldsymbol{u}_i^* \right)_l \left(\boldsymbol{u}_j^* \right)_l = \sum_l B_l, \tag{37}$$

where $\mathbb{E}[B_l] = 0$, $|B_l| \leq \frac{2\mu^2 r}{n \cdot p}$, and $\sum_{\mathbb{E}} [B_l^2] = \frac{\mu^2 r}{n \cdot p}$. Lemma follows by using Bernstein inequality (given below) along with the sampling probability p specified in the lemma. ■

Lemma 20 (Bernstein Inequality) *Let b_i be a set of independent bounded random variables, then the following holds $\forall\, t > 0$:*

$$Pr \left(\left| \sum_{i=1}^n b_i - \mathbb{E}[\sum_{i=1}^n b_i] \right| \geq t \right) \leq \exp \left(- \frac{t^2}{\mathbb{E}[\sum_i b_i^2] + t \max_i |b_i| / 3} \right).$$

Lemma 21 (Matrix Bernstein Inequality (Theorem 1.4 of Tropp (2012))) *Let $B_i \in \mathbb{R}^{n \times n}$ be a set of independent bounded random matrices, then the following holds $\forall\, t > 0$:*

$$Pr \left(\left\| \sum_{i=1}^n B_i - \mathbb{E}[\sum_{i=1}^n B_i] \right\|_2 \geq t \right) \leq n \exp \left(- \frac{t^2}{\sigma^2 + tR/3} \right),$$

where $\sigma^2 = \mathbb{E} \left[\sum_i B_i^2 \right]$ and $R = \max_i \|B_i\|_2$.

Appendix D. Proof of Lemma 9

We will prove the statement for $r = 1$. The lemma can be proved by taking a union bound over all r. In order to prove the lemma, we will calculate a high order moment of the random variable

$$\widehat{X}_a := \left\langle \boldsymbol{e}_1, \widehat{\mathbf{H}}^a \boldsymbol{u} \right\rangle,$$

and then use Markov inequality. We use the following notation which is mostly consistent with Lemma 6.5 of Erdos et al. (2013). We abbreviate (i,j) as α and denote $\widehat{h}_{ij}$ by $\widehat{h}_\alpha$. We further let

$$B_{(i,j)(k,l)} := \delta_{jk}.$$

With this notation, we have:

$$\widehat{X}_a = \sum_{\substack{\alpha_1,\cdots,\alpha_a \\ \alpha_1(1)=1}} B_{\alpha_1\alpha_2}\cdots B_{\alpha_{a-1}\alpha_a}\widehat{h}_{\alpha_1}\cdots\widehat{h}_{\alpha_a}u_{\alpha_a(2)}.$$

We now split the matrix $\widehat{\mathbf{H}}$ into two parts $\mathbf{H}$ and $\mathbf{H}'$ which correspond to the upper triangular and lower triangular parts of $\widehat{\mathbf{H}}$. This means

$$\widehat{X}_a = \sum_{\substack{\alpha_1,\cdots,\alpha_a \\ \alpha_1(1)=1}} B_{\alpha_1\alpha_2}\cdots B_{\alpha_{a-1}\alpha_a}\left(h_{\alpha_1}+h'_{\alpha_1}\right)\cdots\left(h_{\alpha_a}+h'_{\alpha_a}\right)u_{\alpha_a(2)}. \tag{38}$$

The above summation has 2^a terms, of which we consider only

$$X_a := \sum_{\substack{\alpha_1,\cdots,\alpha_a \\ \alpha_1(1)=1}} B_{\alpha_1\alpha_2}\cdots B_{\alpha_{a-1}\alpha_a}h_{\alpha_1}\cdots h_{\alpha_a}u_{\alpha_a(2)}.$$

The resulting factor of 2^a does not change the result.

Abbreviating $\boldsymbol{\alpha} := (\alpha_1,\cdots,\alpha_a)$, and

$$\zeta_{\boldsymbol{\alpha}} := B_{\alpha_1\alpha_2}\cdots B_{\alpha_{a-1}\alpha_a}h_{\alpha_1}\cdots h_{\alpha_a}u_{\alpha_a(2)},$$

we can write

$$X_a = \sum_{\boldsymbol{\alpha}}\zeta_{\boldsymbol{\alpha}},$$

where the summation runs only over those $\boldsymbol{\alpha}$ such that $\alpha_1(1)=1$.

Calculating the k^{th} moment expansion of X_a for some even number k, we obtain:

$$\mathbb{E}\left[X_a^k\right] = \sum_{\boldsymbol{\alpha}^1,\cdots,\boldsymbol{\alpha}^k}\mathbb{E}\left[\zeta_{\boldsymbol{\alpha}^1}\cdots\zeta_{\boldsymbol{\alpha}^k}\right]. \tag{39}$$

For each valid $\boldsymbol{\alpha} = (\boldsymbol{\alpha}^s) = (\alpha_l^s)$, we define the partition $\Gamma(\boldsymbol{\alpha})$ of the index set $\{(s,l) : s\in[k]; l\in[a]\}$, where (s,l) and (s',l') are in the same equivalence class if $\alpha_l^s = \alpha_{l'}^{s'}$. We first bound the contribution of all $\boldsymbol{\alpha}$ corresponding to a partition Γ in the summation (39) and then bound the total number of partitions Γ possible. Since each h_α is centered, we can conclude that any partition Γ that has a non-zero contribution to the summation in (39) satisfies:

(*) each equivalence class of Γ contains at least two elements.

We further bound the summation in (39) by taking absolute values of the summands

$$\mathbb{E}\left[X_a^k\right] \le \sum_{\boldsymbol{\alpha}^1,\cdots,\boldsymbol{\alpha}^k}\mathbb{E}\left[|\zeta_{\boldsymbol{\alpha}^1}|\cdots|\zeta_{\boldsymbol{\alpha}^k}|\right], \tag{40}$$

where the summation runs over $(\boldsymbol{\alpha}^1, \cdots, \boldsymbol{\alpha}^k)$ that correspond to valid partitions Γ. Fixing one such partition Γ, we bound the contribution to (40) of all the terms $\boldsymbol{\alpha}$ such that $\Gamma(\boldsymbol{\alpha}) = \Gamma$.

We denote $G \equiv G(\Gamma)$ to be the graph constructed from Γ as follows. The vertex set $V(G)$ is given by the equivalence classes of Γ. For every (s, l), we have an edge between the equivalence class of (s, l) and the equivalence class of $(s, l+1)$.

Each term in (40) can be bounded as follows:

$$\begin{aligned}\mathbb{E}\left[|\zeta_{\boldsymbol{\alpha}^1}| \cdots |\zeta_{\boldsymbol{\alpha}^k}|\right] &\leq \|\boldsymbol{u}\|_\infty^k \left(\prod_{s=1}^{k}\prod_{l=1}^{a-1} B_{\alpha_l^s \alpha_{l+1}^s}\right) \mathbb{E}\left[\prod_{s=1}^{k}\left(\prod_{l=1}^{a}|h_{\alpha_l^s}|\right)\right] \\ &\leq \|\boldsymbol{u}\|_\infty^k \left(\prod_{s=1}^{k}\prod_{l=1}^{a-1} B_{\alpha_l^s \alpha_{l+1}^s}\right) \prod_{\gamma \in V(G)} \frac{1}{n},\end{aligned}$$

where the last step follows from property $(*)$ above and Definition 7.

Using the above, we can bound (40) as follows:

$$\mathbb{E}\left[X_a^k\right] \leq \frac{\|\boldsymbol{u}\|_\infty^k}{n^v} \sum_{\alpha_1, \cdots, \alpha_v} \left(\prod_{\{\gamma, \gamma'\} \in E(G)} B_{\alpha_\gamma \alpha_{\gamma'}}\right).$$

where $v := |V(G)|$ denotes the number of vertices in G.

Factorizing the above summation over different components of G, we obtain

$$\mathbb{E}\left[X_a^k\right] \leq \frac{\|\boldsymbol{u}\|_\infty^k}{n^v} \prod_{j=1}^{l} \sum_{\alpha_1, \cdots, \alpha_{v_j}} \left(\prod_{\{\gamma, \gamma'\} \in E(G_j)} B_{\alpha_\gamma \alpha_{\gamma'}}\right), \tag{41}$$

where l denotes the number of connected components of G, G_j denotes the j^{th} component of G, and v_j denotes the number of vertices in G_j. We will now bound terms corresponding to one connected component at a time. Pick a connected component G_j. Since $\alpha_1^s(1) = 1$ for every $s \in [a]$, we know that there exists a vertex $\alpha_\gamma \in G_j$ such that $\alpha_\gamma(1) = 1$. Pick one such vertex as a root vertex and create a spanning tree T_j of G_j. We use the bound $B_{\alpha_\gamma \alpha_{\gamma'}} \leq 1$ for every $\{\gamma, \gamma'\} \in E_j \setminus T_j$. The remaining summation $\sum_{\alpha_1, \cdots, \alpha_{v_j}} \left(\prod_{\{\gamma, \gamma'\} \in T_j} B_{\alpha_\gamma \alpha_{\gamma'}}\right)$ can be calculated bottom up from leaves to the root. Since

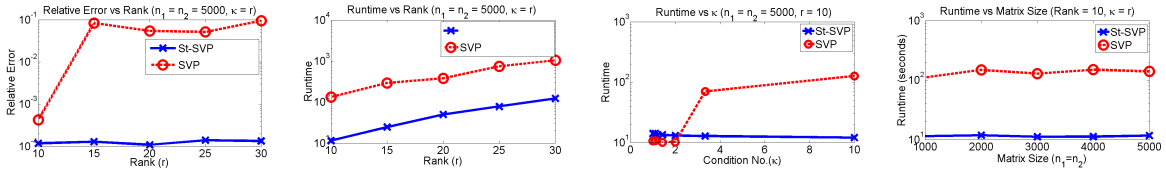
$$\sum_{\alpha_{\gamma'}} B_{\alpha_\gamma \alpha_{\gamma'}} = n, \ \forall\, \gamma,$$

we obtain

$$\sum_{\alpha_1, \cdots, \alpha_{v_j}} \left(\prod_{\{\gamma, \gamma'\} \in E(G_j)} B_{\alpha_\gamma \alpha_{\gamma'}}\right) \leq n^{v_j}.$$

Plugging the above in (41) gives us

$$\mathbb{E}\left[X_a^k\right] \leq \frac{\|\boldsymbol{u}\|_\infty^k}{n^v} n^{\sum_j v_j} = \|u\|_\infty^k.$$



(a) (b) (c) (d)

Figure 1: SVP vs St-SVP: simulations on synthetic datasets. a), b): recovery error and run time of the two methods for varying rank. c): run time required by St-SVP and SVP with varying condition number. d): run time of both the methods with varying matrix size.

Noting that the number of partitions Γ is at most $(ka)^{ka}$, we obtain the bound

$$\mathbb{E}\left[X_a^k\right] \leq \left(\|\boldsymbol{u}\|_\infty (ka)^a\right)^k.$$

Choosing $k = 2\lceil \frac{\log n}{a} \rceil$ and applying k^{th} moment Markov inequality, we obtain

$$\mathbf{Pr}\left[|X_a| > (c\log n)^a \|\boldsymbol{u}\|_\infty\right] \leq \mathbb{E}\left[|X_a|^k\right] \left(\frac{1}{(c\log n)^a \|\boldsymbol{u}\|_\infty}\right)^k \leq \left(\frac{ka}{c\log n}\right)^{ka} \leq n^{-2\log\frac{c}{2}}.$$

Going back to (38), we have:

$$\begin{aligned}
\mathbf{Pr}\left[\left|\widehat{X}_a\right| > (c\log n)^a \|\boldsymbol{u}\|_\infty\right] &\leq 2^a \mathbf{Pr}\left[|X_a| > \left(\frac{c}{2}\log n\right)^a \|\boldsymbol{u}\|_\infty\right] \\
&\leq 2^a \mathbb{E}\left[|X_a|^k\right] \left(\frac{1}{\left(\frac{c}{2}\log n\right)^a \|\boldsymbol{u}\|_\infty}\right)^k \\
&\leq 2^a \left(\frac{ka}{c\log n}\right)^{ka} \leq n^{-2\log\frac{c}{4}}.
\end{aligned}$$

Applying a union bound now gives us the result.

Appendix E. Empirical Results

In this section, we compare the performance of St-SVP with SVP on synthetic examples. We do not however include comparison to other matrix completion methods like nuclear norm minimization or alternating minimization; see Jain et al. (2010) for a comparison of SVP with those methods.

We implemented both the methods in Matlab and all the results are averaged over 5 random trials. In each trial we generate a random low rank matrix and observe $|\Omega| = 5(n_1+n_2)r\log(n_1+n_2)$ entries from it uniformly at random.

In the first experiment, we fix the matrix size ($n_1 = n_2 = 5000$) and generate random matrices with varying rank r. We choose the first singular value to be 1 and the remaining ones to be $1/r$, giving us a condition number of $\kappa = r$. Figure 1 (a) & (b) show the error in recovery and the run time of the two methods, where we define the recovery error as $\left\|\widehat{\mathbf{M}} - \mathbf{M}\right\|_2 / \|\mathbf{M}\|_2$. We see that St-SVP recovers the underlying matrix much more accurately as compared to SVP. Moreover, St-SVP is an order of magnitude faster than SVP.

In the next experiment, we vary the condition number of the generated matrices. Interestingly, for small κ, both SVP and St-SVP recover the underlying matrix in similar time. However, for larger κ, the running time of SVP increases significantly and is almost two orders of magnitude larger than that of St-SVP. Finally, we study the two methods with varying matrix sizes while keeping all the other parameters fixed ($r = 10$, $\kappa = 1/r$). Here again, St-SVP is much faster than SVP.