Proceedings of Machine Learning Research vol 75:1–33, 2018 31st Annual Conference on Learning Theory

Non-Convex Matrix Completion Against a Semi-Random Adversary

Yu Cheng YUCHENG@CS.DUKE.EDU and **Rong Ge** RONGGE@CS.DUKE.EDU
Duke University

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

Matrix completion is a well-studied problem with many machine learning applications. In practice, the problem is often solved by non-convex optimization algorithms. However, the current theoretical analysis for non-convex algorithms relies crucially on the assumption that each entry of the matrix is observed with exactly the same probability p, which is not realistic in practice.

In this paper, we investigate a more realistic semi-random model, where the probability of observing each entry is *at least* p. Even with this mild semi-random perturbation, we can construct counter-examples where existing non-convex algorithms get stuck in bad local optima.

In light of the negative results, we propose a pre-processing step that tries to re-weight the semi-random input, so that it becomes "similar" to a random input. We give a nearly-linear time algorithm for this problem, and show that after our pre-processing, all the local minima of the non-convex objective can be used to approximately recover the underlying ground-truth matrix.

Keywords: Matrix Completion, Non-Convex Optimization, Semi-Random Model.

1. Introduction

Non-convex optimization techniques are now very popular for machine learning applications, especially in learning neural networks (Bengio et al., 2013; Schmidhuber, 2015). These methods are easier to implement and extremely efficient in practice. Even though optimizing a non-convex function is hard in general, recent works proved convergence guarantees for problems including matrix completion, dictionary learning and tensor decomposition (Jain et al., 2013; Sun and Luo, 2015; Arora et al., 2015; Ge et al., 2015).

Unlike convex optimization, non-convex problems may have many bad local optima. Existing techniques rely heavily on model assumptions to get a strong initialization (e.g., (Jain et al., 2013; Sun and Luo, 2015)), or to prove that the objective function has no bad local optima (e.g., (Ge et al., 2016, 2017)). In this paper, we investigate the robustness of non-convex algorithms against model misspecification. In particular, we focus on the matrix completion problem – a well-known learning problem with applications to recommendation systems (Koren, 2009; Rennie and Srebro, 2005). Both its convex relaxations (Srebro and Shraibman, 2005; Recht, 2011) and non-convex approaches (Jain et al., 2013; Sun and Luo, 2015; Ge et al., 2016) were studied extensively before (see Section 1.3 for more related work).

In the matrix completion problem, there is an unknown low-rank matrix $M^\star$ that can be factored into $M^\star = U^\star(V^\star)^\top$, where $U^\star \in \mathbb{R}^{n_1\times r}$, $V^\star = \mathbb{R}^{n_2\times r}$, and r is the rank of the matrix. After observing a random set of entries, the goal is to recover the entire low-rank matrix $M^\star$.

Matrix completion arises naturally in the design of recommendation systems. For example, the rows of the matrix may correspond to users, and the columns of the matrix correspond to items. Each row of $U^\star$ is then a vector representing the preference of a user, and each row of $V^\star$ is a vector representing the properties of an item. Revealing an entry $M^\star_{i,j}$ can be interpreted as user i providing his rating for item j. If the matrix $M^\star$ can be recovered from few observations, the system can simply recommend the item with the highest rating in $M^\star$ (among all the items that the user has not rated before).

The standard assumption made by both convex and non-convex approaches for matrix completion is that the entries are observed *uniformly* at random. That is, every entry is observed with the same probability p. This is not realistic in recommendations systems, as different groups of users may have different probabilities of rating different items.

In this paper, we investigate whether the popular approaches for matrix completion can handle *model misspecification*. In particular, we show that if the observation probability is always *at least* p, then the popular non-convex approaches provably fail. In light of this, we give an efficient preprocessing algorithm that reweights the observed entries, and prove that non-convex optimization algorithms applied to the reweighted objective will always find the desired solution.

1.1. Non-Convex Matrix Completion

Traditionally, matrix completion can be solved by a convex relaxation (Candès and Tao, 2010).

$$\min \ \|M\|_* \qquad \text{s.t. } M_{i,j} = M^\star_{i,j}, \ \ \forall (i,j) \in \Omega.$$

Here $\|\cdot\|_*$ is the nuclear norm of a matrix and Ω is the set of observed entries. However, in practice the matrix $M^\star$ can have very high dimensions. Solving the convex relaxation can be very expensive. The most popular approaches in practice use non-convex heuristics: The algorithms represent the matrix M as $UV^\top$ where U, V have the same dimensions as $U^\star, V^\star$, and then try to minimize

$$\sum_{(i,j)\in\Omega} \left((UV^\top)_{i,j} - M^\star_{i,j}\right)^2,$$

using techniques such as alternating minimization (Koren, 2009) or gradient descent (Rennie and Srebro, 2005). Sometimes additional terms are added as regularizers. Recently, many results established strong convergence guarantees for these non-convex approaches (e.g., (Jain et al., 2013; Sun and Luo, 2015; Ge et al., 2016), see more discussions in Section 1.3).

In order to understand robustness of algorithms under model misspecification, we consider a natural setting where the probability $p_{i,j}$ of observing entry (i,j) can be different, but they are still all at least p. In fact, our algorithm can work with a slightly stronger semi-random model: each entry is first revealed with probability p (same as the standard model). After that, an adversary is allowed to examine the ground-truth matrix $M^\star$ and the set of currently observed entries. The adversary can choose to reveal additional entries of the matrix (adding elements to Ω). The setting where every entry is observed with probability $p_{i,j} \geq p$ is a special case of this semi-random model.

Intuitively, what the adversary does is beneficial for us, because we get to observe more entries of the matrix. Indeed, the convex relaxation will still work in this semi-random model.[1] Our work is motivated by the following questions: Are the non-convex approaches robust in this semi-random model? If not, is there a way to fix the non-convex algorithms to get an algorithm that is both robust to the semi-random adversary and efficient in practice?

[1] To see this, notice that the additional entries correspond to additional constraints, and the original optimal solution ($M^\star$) also satisfies all of these additional constraints. Therefore $M^\star$ must still be the optimal solution.

1.2. Our Results

We first give some examples where the non-convex approaches fail. When each entry is revealed with equal probability, Ge et al. (2017) showed that all local minima of a non-convex objective (see Equation (1) in Section 2.2) recovers the ground-truth matrix $M^\star$. We give an example where Equation (1) has a local minimum that is very far from $M^\star$ in the semi-random model. Another popular approach in non-convex approaches is to rely on SVD to do initialization. In the second example, we show that in the semi-random model, the SVD of the observed entries can be very far from the ground truth $M^\star$. These examples show that the current non-convex approaches rely heavily on the assumption that entries are revealed independently. We leave the details to Appendix A.

We then give an efficient pre-processing algorithm that can make non-convex algorithms robust to the semi-random adversary. Intuitively, we view the observed entries as edges of a bipartite graph (between rows and columns of the matrix): the initial random procedure generates a random bipartite graph with independent edges, and the adversary can add additional edges. We design an algorithm that will *re-weight* the edges, so that the resulting weighted graph "looks like" a random bipartite graph in terms of its spectral properties. In Section 4, we will show that such spectral properties are sufficient for non-convex approaches to approximately recover the ground-truth $M^\star$.

Problem 1 (Removing Noisy Edges) *Let $G = (V_1, V_2, E)$ be an unweighted bipartite graph. Assume there exists a subset $S \subseteq E$ of edges such that (V_1, V_2, S) is spectrally similar to a complete bipartite graph. Compute a set of weights $w_e \geq 0$ for every edge $e \in E$ such that the weighted graph is spectrally similar to a complete bipartite graph.*

The existence of weights w_e in Problem 1 is guaranteed by assigning weight 1 to the edges selected randomly, and weight 0 to the edges added by the adversary. Roughly speaking, the pre-processing algorithm will try to put large weights on the edges generated originally, and assign small weights to the edges added by the adversary. Note that these weights cannot be found by simple reweighting schemes based on the number of observed entries in each row/column, because our counter-examples have the same number of observed entries in every row/column.

We give a nearly-linear time algorithm for Problem 1 in Section 3. Our approach extends techniques from a line of work on linear-sized graph sparsification algorithms (Batson et al., 2012; Allen Zhu et al., 2015; Lee and Sun, 2017).

Theorem 1 (Pre-processing (Informal)) *Given an input graph with m edges, assume a subset of the edges forms a graph that is β-spectrally close to the complete bipartite graph. There is an algorithm that runs in time $\widetilde{O}(m/\epsilon^{O(1)})$ that outputs a set of weights for the edges, and the weighted graph is $O(\beta) + \epsilon$ spectrally close to the complete bipartite graph.* [2]

See Definition 5 for the definition of spectrally similarity, and Theorem 6 for the formal version of the theorem. After this pre-processing step, we get a weighted graph that is spectrally similar to the original random graph (before the adversary added edges). We can then use these weights to change the non-convex objective, and minimize:

$$\sum_{(i,j)\in\Omega} W_{i,j} \left((UV^\top)_{i,j} - M^\star_{i,j}\right)^2 .$$

We will show this objective (with some additional regularizers) has no bad local minimum.

[2] Throughout the paper, $\widetilde{O}(\cdot)$ hides polylogarithmic factors.

Theorem 2 (Main) *In the semi-random model, if the reveal probability* $p \geq \frac{C\mu^6 r^6 (\kappa^\star)^6 \log(n_1+n_2)}{n_1 \epsilon^2}$ *for a large enough constant* C*, with high probability, using weights produced by Algorithm 1, all local minima of Objective (1) satisfies* $\|UV^\top - M^\star\|_F^2 \leq \epsilon \|M^\star\|_F^2$*. The pre-processing time is* $\widetilde{O}(m \cdot \mathrm{poly}(\mu, r, \kappa^\star, \epsilon^{-1}))$ *where* m *is the total number of revealed entries.*

Here $\mu, r, \kappa^\star$ are the incoherence parameter, rank, and the condition number of $M^\star$ (see Section 2 for formal definitions). Previous analyses crucially rely on concentration bounds obtained from observing the entries of $M^\star$ uniformly at random. We prove Theorem 2 in Section 4, where we replace these concentration bounds with spectral properties guaranteed by our pre-processing algorithm (Theorem 1). See Section 4 and Appendix D for details.

1.3. Related Work

Matrix Completion. The first theoretical guarantees on matrix completion come from convex relaxation (Srebro and Shraibman, 2005; Recht, 2011; Candès and Tao, 2010; Candès and Recht, 2009; Negahban and Wainwright, 2012). In particular, Candès and Tao (2010) and Recht (2011) showed that if $\Omega((n_1 + n_2) r \mu^2 \log^2(n_1 + n_2))$ entries are observed randomly (where μ is the incoherence parameter, see Section 2.2), the nuclear norm convex relaxation recovers the exact underlying low-rank matrix. There have been many works trying to improve the running time (e.g., (Srebro et al., 2004; Mazumder et al., 2010; Hastie et al., 2014) and the references therein).

For the non-convex approaches, the first set of results require a good initialization. Keshavan et al. (2010a,b) showed that well-initialized gradient descent recovers $M^\star$. Later, it was shown that several other non-convex algorithms, including alternating minimization (Jain et al., 2013; Hardt and Wootters, 2014; Hardt, 2014) and gradient descent (Chen and Wainwright, 2015), converge to the correct solution with a careful initialization.

Recently, the work of Sun and Luo (2015) (and subsequent works (Zhao et al., 2015; Zheng and Lafferty, 2016; Tu et al., 2015)) established a common framework for matrix completion algorithms with a good initialization. In particular, they proved an analog of strong convexity in the neighborhood of the optimal solution. As a result, many different algorithms can converge to $M^\star$.

For guarantees without careful initialization, Sa et al. (2015) analyzed stochastic gradient descent from random initialization. More recently, Ge et al. (2016); Park et al. (2016); Ge et al. (2017); Chen and Li (2017) showed that the non-convex objective (with careful regularization) does not have any bad local minima.

All of the works above require uniformly random observations. There have also been works that try to solve matrix completion problem under deterministic assumptions (Bhojanapalli and Jain, 2014; Li et al., 2016). (Bhojanapalli and Jain, 2014) uses convex relaxations, and the conditions in (Li et al., 2016) does not apply to the semi-random model.

Graph Sparsification. The goal of graph sparsification is to use a few weighted edges to approximate a given graph. We focus on the notion of spectral similarity in this paper (see Definition 5). [3] The seminal work of Spielman and Teng (2011) showed that, for any undirected graph G, an ϵ-spectral sparsifier of G with $\widetilde{O}(n/\epsilon^2)$ edges can be constructed in nearly-linear time. In a beautiful piece of work, Batson et al. (2012) showed that an ϵ-spectral sparsifier with a linear number of $O(n/\epsilon^2)$ edges exist and can be computed in polynomial time. Since then, there have been many

[3] For an undirected graph G, we use m and n to denote the number of edges and vertices respectively.

efforts (Allen Zhu et al., 2015; Lee and Sun, 2015, 2017) on speeding up the construction of linear-sized sparsifiers. Recently, Lee and Sun (2017) gave an algorithm for this problem that runs in near-linear time $\widetilde{O}(m/\epsilon^{O(1)})$.

Semi-Random Model. The semi-random model was first proposed by Blum and Spencer (1995) as an intermediate model between average-case and worst-case. Algorithms for semi-random models were developed for many graph problems, including planted clique (Jerrum, 1992; Feige and Krauthgamer, 2000), community detection/stochastic block model (Feige and Kilian, 2001; Perry and Wein, 2017; Moitra et al., 2016), graph partitioning (Kucera, 1995; Makarychev et al., 2012) and correlation clustering (Mathieu and Schudy, 2010; Makarychev et al., 2015). Most of these works use convex relaxations, and non-convex approaches (including spectral algorithms) were known to fail (Feige and Kilian, 2001; Moitra et al., 2016) for some of these problems. To the best of our knowledge, our work is the first one that tries to fix the non-convex approach using a light-weight convex pre-processing step.

Non-Convex Optimization. Although non-convex optimization is NP-hard in general, under reasonable assumptions it is possible to find a local minimum efficiently (e.g., (Ge et al., 2015; Agarwal et al., 2016; Jin et al., 2017)). It follows from Theorem 2 that, by running these algorithms after our pre-processing step, we get non-convex algorithms that can recover the ground-truth matrix $M^\star$ approximately even in the semi-random model.

2. Preliminary

2.1. Notations

We use $[n]$ to denote the set $\{1, \ldots, n\}$. We use e_i or χ_i for the i-th standard basis vector. We write I for the identity matrix, and J for the all ones matrix.

For a vector x, we use $\|x\|_1$, $\|x\|$ and $\|x\|_\infty$ for the ℓ_1, ℓ_2, and ℓ_∞ norm of x respectively.

For a matrix A, we use A_i to denote i-th row of A. We use $\|A\|_1$, $\|A\|$, $\|A\|_\infty$, $\|A\|_F$, and $\|A\|_{\max}$ for the ℓ_1, spectral, ℓ_∞, Frobenius norm, and maximum absolute entry of A. Note that $\|A\|_\infty$ (and $\|A\|_1$) is just the maximum ℓ_1-norm of the rows (and columns) of A. Let $\lambda_{\min}(A)$ denote the minimum eigenvalue of A, and let $\sigma_i(A)$ denote the i-th largest singular value of A.

A symmetric $n \times n$ matrix A is said to be positive semidefinite (PSD) if $x^\top Ax \geq 0$ for all $x \in \mathbb{R}^n$, and positive definite if $x^\top Ax > 0$ for any $x \neq 0$. For two symmetric matrices A and B of the same dimensions, we write $A \preceq B$ (or equivalently $B \succeq A$) when $B - A$ is positive semidefinite, and $A \prec B$ when $B - A$ is positive definite.

We use $\mathrm{tr}(A)$ for the trace of a square matrix A. Let $A * B$ be the Hadamard (entry-wise) product of two matrices (where $(A * B)_{i,j} = A_{i,j}B_{i,j}$). For $n_1 \times n_2$ matrices A and B, we write $A \bullet B$ or $\langle A, B\rangle$ for the entry-wise inner product of A and B: $\langle A, B\rangle = \mathrm{tr}(A^\top B) = \sum_{i,j}(A*B)_{i,j}$. For a subset of entries $\Omega \subseteq [n_1] \times [n_2]$, let $\langle A, B\rangle_\Omega = \sum_{(i,j)\in\Omega}(A * B)_{i,j}$. For a weight matrix W, let $\langle A, B\rangle_W = \sum_{i,j} W_{i,j}(A * B)_{i,j} = \langle W * A, B\rangle$. Let $\|A\|_\Omega^2 = \langle A, A\rangle_\Omega$ and $\|A\|_W^2 = \langle A, A\rangle_W$.

Let $A \otimes B$ denote the Kronecker product of A and B. If $X \in \mathbb{R}^{n\times r}$, define the Katri-Rao product $X \odot X$ be an $n \times r^2$ matrix whose i-th row is equal to $X_i \otimes X_i$. [4]

[4] This is the transpose of the traditional Katri-Rao product that works on columns instead of rows.

2.2. Matrix Completion and Non-convex Formulation

Throughout the paper, the ground-truth matrix is always denoted by $M^\star$. We assume the hidden low-rank matrix $M^\star \in \mathbb{R}^{n_1 \times n_2}$ is of rank r and can be decomposed as $M^\star = U^\star(V^\star)^\top$, where $U^\star \in \mathbb{R}^{n_1 \times r}, V^\star \in \mathbb{R}^{n_2 \times r}$ (or $M^\star = U^\star(U^\star)^\top$ in symmetric case). For the symmetric case, we assume $n_1 = n_2 = n$. For the asymmetric case we assume (w.l.o.g.) that $n_1 \le n_2$, and let $n = n_1 + n_2$. We also assume w.l.o.g. that $(U^\star)^\top U^\star = (V^\star)^\top V^\star$. [5] We use $\sigma_1^\star$, $\sigma_r^\star$ to denote the first and r-th singular values of $M^\star$ respectively. Let $\kappa^\star = \sigma_1^\star/\sigma_r^\star$ be the condition number of $M^\star$.

Definition 3 (Incoherence) *The matrix $M^\star$ with SVD $M^\star = XDY^\top$ is μ-incoherent, if for all $i \in [n_1], j \in [n_2]$, $\|X_i\| \le \sqrt{\mu r/n_1}$ and $\|Y_j\| \le \sqrt{\mu r/n_2}$.*

In the standard matrix completion setting, each entry of the matrix is observed with probability p. Let Ω be the set of observed entries. To make the notation more flexible, we define a weight matrix W such that $W_{i,j} = \frac{1}{p}$ if $(i,j) \in \Omega$ (and 0 otherwise). We use the objective function in (Ge et al., 2016, 2017) that includes additional regularizers:

$$\min\ f(U,V) = 2\|UV^\top - M^\star\|_W^2 + \frac{1}{2}\|U^\top U - V^\top V\|_F^2 + Q(U,V), \tag{1}$$

where

$$Q(U,V) = \lambda_1 \sum_{i=1}^{n_1} (\|U_i\|_2 - \alpha_1)_+^4 + \lambda_2 \sum_{i=1}^{n_2} (\|V_i\|_2 - \alpha_2)_+^4.$$

Here $\lambda_1, \lambda_2, \alpha_1, \alpha_2$ are parameters to choose. The notation x_+ represents $\max\{x, 0\}$. Intuitively, the regularizer $Q(U,V)$ ensures U, V have bounded row norms. The additional term $\frac{1}{2}\|U^\top U - V^\top V\|_F^2$ is a regularizer that is popular for asymmetric matrix completion (Park et al., 2016). Ge et al. (2017) showed the following result:

Theorem 4 (Ge et al. (2017), Informal) *Under appropriate settings of parameters λ, α, if $p \ge poly(r, \mu, \kappa^\star, \log n)/n$ for some fixed polynomial, then with high probability, all local minima of Equation (1) are globally optimal, and they satisfy $UV^\top = M^\star$.*

2.3. Graph Laplacians

For a weighted undirected graph $G = (V, E, w)$ with n vertices and edge weights $w_e \ge 0$, let D be a diagonal matrix containing the weighted degree of each vertex ($D_{i,i} = \sum_{(i,j)\in E} w_{(i,j)}$). Let A be the adjacency matrix of G ($A_{i,j} = A_{j,i} = w_{(i,j)}$). The Laplacian matrix of G is a symmetric $n \times n$ matrix L defined as $L = D - A$. In other words, if we orient every edge $e = (i,j) \in E$ arbitrarily and represent it by a vector $b_e \in \mathbb{R}^n$ with $b_e(i) = 1$ and $b_e(j) = -1$. The Laplacian matrix L is equal to $L = \sum_{e\in E} w_e b_e b_e^\top$.

We use $L^{1/2}$ to denote the principal square root of a PSD matrix L. Abusing notation, we use L^{-1} for the Moore-Penrose inverse of L, and $L^{-1/2}$ for $(L^{-1})^{1/2}$. The normalized Laplacian is the matrix $D^{-1/2}LD^{-1/2}$. The effective resistance of an edge e in a graph with Laplacian L is $b_e^\top L^{-1} b_e$. For any Laplacian L, we have $L^{-1}L = I_{\mathrm{im}(G)}$, where $I_{\mathrm{im}(G)} = I - \frac{1}{n}J$ is the identity matrix on the image space of L. We often abbreviate $I_{\mathrm{im}(G)}$ as I.

We say two graphs are *spectrally similar* if the following holds:

[5] This can be achieved by setting $U^\star = XD^{1/2}$ and $V^\star = YD^{1/2}$ if $M^\star = XDY^\top$ is the SVD of $M^\star$.

Definition 5 (Spectral Similarity (Spielman and Teng, 2011)) *Suppose L_1, L_2 are the Laplacians for graphs G_1, G_2 respectively, we say G_1 and G_2 are ϵ-spectrally similar if and only if*

$$(1-\epsilon)L_1 \preceq L_2 \preceq (1+\epsilon)L_1.$$

Graph sparsification is known to be a special case of sparsifying sum of rank-one PSD matrices (see, e.g., (Batson et al., 2012)). Similarly, we can reduce Problem 1 to Problem 2.

Problem 2 *For a set of m vectors $\{v_i\}_{i=1}^m$, assume there exists a subset S of vectors such that $(1-\beta)I \preceq \sum_{i\in S} v_i v_i^\top \preceq (1+\beta)I$. Compute a set of weights $w_i \geq 0$ such that*

$$(1 - O(\beta) - \epsilon) \preceq \textstyle\sum_i w_i v_i v_i^\top \preceq I.$$

To see why Problem 1 can be reduced to Problem 2, let L_H be the Laplacian for the complete bipartite graph. The reduction simply sets $v_e = (L_H)^{-1/2} b_e$ for each edge $e \in E$.

3. Pre-Processing: Reweighting the Entries

In this section, we present a nearly-linear time algorithm for Problem 1. As we discussed in Section 2.3, Problem 1 is equivalent to the problem of approximating the identity matrix (Problem 2). We prove the following theorem for Problem 2:

Theorem 6 (Our Preprocessing Algorithm) *Fix $0 \leq \epsilon, \beta \leq 1/10$, and a graph Laplacian $L \in \mathbb{R}^{n\times n}$. Given a set of m vectors $\{v_i\}_{i=1}^m$, where each $v_i = L^{-1/2} b_i$ for some b_i representing an edge (with only two non-zero entries, one $+1$ and one -1). Assume there exist weights $w_i \geq 0$ such that* [6]

$$(1-\beta)I \preceq \textstyle\sum_{i=1}^m w_i v_i v_i^\top \preceq (1+\beta)I.$$

We can find a set of weights $\widetilde{w}_i \geq 0$ in $\widetilde{O}(m/\epsilon^{O(1)})$ time, such that with high probability,

$$(1 - O(\beta) - \epsilon)I \preceq \textstyle\sum_{i=1}^m \widetilde{w}_i v_i v_i^\top \preceq I.$$

We adapt techniques from recent developments on linear-sized graph sparsification (Batson et al., 2012; Allen Zhu et al., 2015; Lee and Sun, 2017). The main difference between our problem and the graph sparsification problem is the following: instead of assuming $\sum_i v_i v_i^\top = I$, we only know the *existence* of an unknown set S such that $\sum_{i\in S} v_i v_i^\top = I$. This prevents us from using some of the well-known techniques in graph sparsification (e.g., sampling by effective resistance Spielman and Srivastava (2011); Lee and Sun (2015)). For the same reason, any simple reweighting algorithms that are oblivious to whether a good set S exists will not work.

One of our main contributions is to identify that the framework of Batson et al. (2012) is not only limited to graph sparsification. The fact that the algorithm picks edges *deterministically* makes it much more powerful, and the analysis only requires the *existence* of a "good" edge to add in each iteration. On the technical level, our work departs from previous works in two important ways: (1) our algorithm works even when the hidden set S has sum only *approximately* equal to I; and (2) our analysis is considerably simpler, partly because we do not require the output weights to be sparse.

[6] We assume $0 < \beta \leq \frac{1}{10}$ is given, because we can do a binary search by running our algorithm and see if it succeeds. It is worth mentioning that we never explicitly compute any $v_i = L^{-1/2} b_i$. See Appendix B for more details.

We first give an overview of the framework of Batson et al. (2012). We will maintain two barrier values $\ell < u$, and a weighted sum of the rank-one matrices $A = \sum_{i=1}^m w_i v_i v_i^\top$ such that $\ell I \prec A \prec uI$. The plan is to start with some constants $\ell < 0 < u$, $A = 0$, and gradually increase the weights $\{w_i\}_i$, u and ℓ, while making sure that A stays between the two barriers uI and ℓI. If we can increase u and ℓ at roughly the same rate, the condition number of A will become smaller.

Our approach in this section is most directly inspired by the recent work of Lee and Sun (2017). We use the following potential function proposed in (Lee and Sun, 2017) to measure how far A is from the barriers (both uI and ℓI):

$$\Phi_{u,\ell}(A) = \Phi_u(A) + \Phi_\ell(A),$$
$$\text{where } \Phi_u(A) = \operatorname{tr}\exp\left((uI - A)^{-1}\right), \text{ and } \Phi_\ell(A) = \operatorname{tr}\exp\left((A - \ell I)^{-1}\right).$$

If A is far from uI and ℓI, then all eigenvalues of $uI - A$ and $A - \ell I$ are large and $\Phi_{u,\ell}(A)$ is small. The potential function is going to guide us on how to increase the weights w_i so that A stays away from the barriers. The derivatives of the potential functions with respect to A are

$$\nabla\Phi_u(A) = \exp\left((uI - A)^{-1}\right)(uI - A)^{-2}, \text{ and } \nabla\Phi_\ell(A) = -\exp\left((A - \ell I)^{-1}\right)(A - \ell I)^{-2}.$$

For notational convenience, we define $C_- = \nabla\Phi_u(A)$, $C_+ = -\nabla\Phi_\ell(A)$, and $C = C_+ - C_-$. Note that when $\ell I \prec A \prec uI$, both C_+ and C_- are PSD matrices. The first order approximation of the potential function is $\Phi_{u,\ell}(A + \Delta) \approx \Phi_{u,\ell}(A) + \nabla\Phi_u(A) \bullet \Delta + \nabla\Phi_\ell(A) \bullet \Delta = \Phi_{u,\ell}(A) - C \bullet \Delta$.

We want $\Phi_{u,\ell}(A + \Delta)$ to be small, which guarantees that $A + \Delta$ is far away from ℓI and uI. Therefore, in each iteration, we seek a matrix Δ such that

(1) Δ is small enough for the first-order approximation of $\Phi_{u,\ell}(A + \Delta)$ to be accurate; and

(2) Δ maximizes $C \bullet \Delta$, the reduction to (first-order approximation of) the potential function.

Formally, let $\rho = (\lambda_{\min}\{uI - A, A - \ell I\})^2$. When $0 \preceq \Delta \preceq \epsilon\rho I$, the first-order approximation of $\Phi_{u,\ell}(A + \Delta)$ is accurate (see Lemma 18 in Appendix B). We are interested in the following SDP:

$$\begin{aligned} \text{maximize} \quad & C \bullet X \\ \text{subject to} \quad & X \preceq \epsilon\rho I, \quad X = \textstyle\sum_{i=1}^m x_i v_i v_i^\top \text{ (which implies } 0 \preceq X), \end{aligned} \tag{2}$$

We give a full description of our algorithm in Algorithm 1.

We will use the following lemmas (Lemmas 7 and 8) to analyze Algorithm 1 and prove Theorem 6. Lemma 7 shows that the SDP in (2) admits a good solution, and we can solve it approximately in nearly-linear time. Lemma 8 says that the potential function $\Phi_{u_j,\ell_j}(A_j)$ never increases, which guarantees that A_j is far away from both $u_j I$ and $\ell_j I$ for all j.

Lemma 7 *Fix $0 < \beta, \epsilon \le 1/10$. In any iteration j of Algorithm 1, given $A_j = \sum_{i=1}^m w_i v_i v_i^\top$ (implicitly by the weights $\{w_i \ge 0\}_{i=1}^m$) and corresponding barrier values $\ell = \ell_j$ and $u = u_j$,*

(1) We can compute $\rho \in [1 - \epsilon, 1] \cdot (\lambda_{\min}\{uI - A, A - \ell I\})^2$ w.h.p. in time $\widetilde{O}(m/\epsilon^{O(1)})$.

(2) Let C, C_+, C_- be defined as above. We can compute a set of weights $\{x_i \ge 0\}_{i=1}^m$ in time $\widetilde{O}(m/\epsilon^{O(1)})$, such that w.h.p. for $X = \sum_{i=1}^m x_i v_i v_i^\top$,

$$C \bullet X \ge \frac{\epsilon\rho}{2}\left((1 - \beta - \epsilon)\operatorname{tr}(C_+) - (1 + \beta + \epsilon)\operatorname{tr}(C_-)\right).$$

Algorithm 1: Find $A = \sum_i w_i v_i v_i^\top \approx I$.

```
**Input**: $\{v_i\}_{i=1}^m$, $\epsilon \le 1/10$.
$j = 0$, $A_0 = 0$, $\ell_0 = -\frac{1}{4}$, $u_0 = \frac{1}{4}$;
**while** $u_j - \ell_j \le 1$ **do**
    Let $\rho \in [1-\epsilon, 1] \cdot (\lambda_{\min}\{u_j I - A_j, A_j - \ell I_j\})^2$;
    Let $\Delta_j$ be an approximate solution to the SDP (2) with $C = -\left(\nabla\Phi_{u_j}(A_j) + \nabla\Phi_{\ell_j}(A_j)\right)$;
    $\delta_{u,j} = \frac{\epsilon\rho}{2} \cdot \frac{(1+\beta+5\epsilon)}{1-2\epsilon}$, $\delta_{\ell,j} = \frac{\epsilon\rho}{2} \cdot \frac{(1-\beta-5\epsilon)}{1+2\epsilon}$;
    $A_{j+1} = A_j + \Delta_j$, $u_{j+1} = u_j + \delta_{u,j}$, $\ell_{j+1} = \ell_j + \delta_{\ell,j}$;  $j = j+1$;
**return** $A_j/u_j$;
```

Lemma 8 *Fix $0 < \beta, \epsilon \le 1/10$. Let $A_{j+1} = A_j + \Delta_j$ denote the matrix in the j-th iteration of Algorithm 1. If Δ_j is an approximate solution to the SDP that satisfies Lemma 7, then we have $\Phi_{u_{j+1},\ell_{j+1}}(A_{j+1}) \le \Phi_{u_j,\ell_j}(A_j)$.*

We defer the proofs of these two lemmas to Appendix B. Now we are ready to prove Theorem 6.

Proof [Proof of Theorem 6] First, we show that $(1 - O(\beta+\epsilon))I \preceq A_j/u_j \preceq I$. The condition number of A_j is upper bounded by $\frac{u_j}{\ell_j} = \left(1 - \frac{u_j - \ell_j}{u_j}\right)^{-1}$, hence it suffices to show that $\frac{u_j-\ell_j}{u_j} = O(\beta+\epsilon)$. Since $u_j - \ell_j > 1$ when the algorithm terminates,

$$\frac{u_j - \ell_j}{u_j} < \frac{2(u_j - \ell_j) - 1}{u_j - \frac{1}{4}} = 2\frac{(u_j - u_0) - (\ell_j - \ell_0)}{u_j - u_0} \le 2\max_j \frac{\delta_{u,j} - \delta_{\ell,j}}{\delta_{u,j}} = O(\beta+\epsilon).$$

Next, we analyze the running time of Algorithm 1. The initial value of the potential function is $\Phi_{u_0,\ell_0}(0) = 2\operatorname{tr}\exp(I) = 2n$. By Lemma 8 and a union bound over j, we have that with high probability, $\Phi_{u_j,\ell_j}(A_j) \le 2n$ for all $j \le O(\log n/\epsilon^2)$. To see that A_j must be far away from the barriers, consider only the contribution of $\lambda_{\min}(u_j I - A_j)$ to the potential function:

$$2n \ge \Phi_{u_j}(A_j) = \operatorname{tr}\exp\left((u_j I - A_j)^{-1}\right) \ge \exp\left(\lambda_{\min}(u_j I - A_j)^{-1}\right).$$

It follows that $\lambda_{\min}(u_j I - A) = \Omega(\log^{-1} n)$, and similarly $\lambda_{\min}(A_j - \ell_j I) = \Omega(\log^{-1} n)$. Therefore, we know that $\rho \ge (1-\epsilon)(\lambda_{\min}\{u_j I - A_j, A_j - \ell I_j\})^2 = \Omega(\log^{-2} n)$, and $\delta_{u,j} - \delta_{\ell,j} = \Omega(\epsilon^2 \cdot \log^{-2} n)$ for all j. Since the algorithm starts with $u_0 > \ell_0$ and terminates when $u_j - \ell_j > 1$, the number of iterations is at most $1/(\min_j(\delta_{u,j} - \delta_{\ell,j})) = O\left(\epsilon^{-2}\log^2 n\right)$.

It remains to show that each iteration takes nearly-linear time. We maintain the matrices A_j and Δ_j implicitly by the corresponding sets of weights, and add their weights together to get A_{j+1}. The input and output of the SDP are also represented implicitly by the weights. By Lemma 7, we can compute ρ and find a near-optimal solution to the SDP in (2) in time $\widetilde{O}(m/\epsilon^{O(1)})$. The overall running time is $\widetilde{O}(\frac{\log^2 n}{\epsilon^2} \cdot \frac{m}{\epsilon^{O(1)}}) = \widetilde{O}(m/\epsilon^{O(1)})$. ■

3.1. From Graphs to Weight Matrices: Consequences of Spectral Similarity

In this section, we show that the weights computed by Algorithm 1 are useful for the semi-random matrix completion problem. More specifically, we prove the following corollary of Theorem 6. Corollary 9 provides the spectral property that is crucial to our analysis of non-convex matrix-completion algorithms in Section 4.

Corollary 9 *Fix $\beta > 0$. Consider the matrix completion problem with ground truth $M^\star \in \mathbb{R}^{n_1 \times n_2}$. There exists $p = O\left(\frac{\log n}{n_1 \beta^2}\right)$ such that if every entry of $M^\star$ is observed with probability at least p, then w.h.p., we can compute a weight matrix $W \in \mathbb{R}^{n_1 \times n_2}$ in time $\widetilde{O}(m/\beta^{O(1)})$, such that W is supported on the observed entries, $\|W\|_\infty \le n_2$, $\|W\|_1 \le n_1$, and $\|W - J\| = O(\beta\sqrt{n_1 n_2})$.*

Recall that J is the all ones matrix, $n = n_1 + n_2$, and we assume $n_1 \le n_2$.

Corollary 9 follows from Lemma 10, Theorem 6, and Lemma 11. Lemma 10 provides concentration bounds for random matrices, which implies that when p is large enough, the semi-random input contains a good subset of observations. We can then apply Theorem 6 to show that our preprocessing algorithm can recover a good set of weights (i.e., a weighted graph that is spectrally similar to the complete bipartite graph). Finally, Lemma 11 shows that the closeness in the Laplacians of two graphs implies the closeness in their (normalized) adjacency matrices.

Lemma 10 *Let G denote the $n_1 \times n_2$ complete bipartite graph. We write $n = n_1 + n_2$ for the number of vertices, and $m = n_1 n_2$ for the number of edges. Let H denote a random graph generated by including each edge of G independently with probability p. W.h.p, we can re-weight edges in H so that the Laplacian matrix L_H is ϵ-spectrally similar with L_G, where $\epsilon = O\left(\sqrt{\frac{n \log n}{pm}}\right)$.*

Lemma 11 *Let $L = D - A$ and $\widetilde{L} = \widetilde{D} - \widetilde{A}$ be two graph Laplacians, where D is the degree matrix and A is the adjacency matrix of the graph. If $(1-\epsilon)L \preceq \widetilde{L} \preceq L$, then we have:*

(1) $(1-\epsilon)D_{i,i} \le \widetilde{D}_{i,i} \le D_{i,i}$.

(2) $\|D^{-1/2}(\widetilde{A} - A)D^{-1/2}\| \le 3\epsilon$.

We defer the proofs of Lemmas 10 and 11 to Appendix C, and first prove Corollary 9.

Proof [Proof of Corollary 9] Let G denote the $n_1 \times n_2$ complete bipartite graph ($n_1 \le n_2$). Let H be the graph corresponds to the entries revealed randomly, and let H' denote the graph after the adversary added extra edges. By Lemma 10, for $p = O\left(\frac{\log n}{n_1 \beta^2}\right)$, with high probability, there exists edge weights for H such that $(1-\beta)L_G \preceq L_H \preceq (1+\beta)L_G$.

Because the edges of H' is a superset of the edges of H, there exist edge weights for H' such that the same condition holds. Since the vectors $\{L_G^{-1/2} b_e\}_{e \in H'}$ satisfy the condition in Problem 2, we can invoke Theorem 6 with $\epsilon = \beta$ to obtain a set of weights $\{\widetilde{w}_e\}_{e \in H'}$ such that $(1 - O(\beta))L_G \preceq \sum_{e \in H'} \widetilde{w}_e b_e b_e^\top \preceq L_G$ in time $\widetilde{O}(m/\beta^{O(1)})$, where m is the number of edges in H'.

Let A denote the adjacency matrix of G, and let A' denote the adjacency matrix of H' with weights $\widetilde{w}_e$. Since both A' and A_G include only edges in the complete bipartite graph, we can write

$$A = \begin{pmatrix} 0 & J \\ J^\top & 0 \end{pmatrix}, \qquad A' = \begin{pmatrix} 0 & W \\ W^\top & 0 \end{pmatrix},$$

where J is the all ones matrix and $W \in \mathbb{R}^{n_1 \times n_2}$ contains the edge weights $\widetilde{w}_e$ ($W_{i,j} = \widetilde{w}_e$ for every $(i,j) \in H'$, and $W_{i,j} = 0$ otherwise). By Lemma 11, the row sum of W is at most n_2 for every row, and the column sum of W is at most n_1 for every column. Again by Lemma 11, $\|D^{-1/2}(A' - A_G)D^{-1/2}\| \le O(\beta)$, which implies that $\|W - J\| = \|A - A'\| = O(\beta\sqrt{n_1 n_2})$. ■

4. Application to Matrix Completion

Most analysis of matrix completion relies on the fact that the observed entries are sampled uniformly at random. Let $W_{i,j} = 1/p$ if entry (i, j) is observed. This assumption is mostly used to prove *concentration inequalities* related to the norm of low-rank matrices $\|M\|_W^2$. In particular, the following two lemmas are used in most papers.

Lemma 12 shows that for an M that is in the "tangent space" (the linear space of $U^\star X^\top + Y(V^\star)^\top$), the norm of Z is preserved after we restrict to the observed entries. Lemma 13 shows that the norm is preserved for every incoherent matrix $XY^\top$.

Lemma 12 (Recht (2011)) *Suppose $M^\star = U^\star\Sigma^\star(V^\star)^\top$. Suppose entries are revealed with probability p independently, weight matrix $W_{i,j} = 1/p$ if (i, j) is revealed and 0 otherwise. For any $0 < \delta < 1$, when $p \geq \Omega(\frac{\mu r}{\delta^2 n}\log n)$, with high probability over the randomness of W we have*

$$|\|M\|_W - \|M\|_F| \leq \delta\|M\|_F,$$

for any matrix $M \in \mathbb{R}^{n_1\times n_2}$ of the form $M = U^\star X^\top + Y(V^\star)^\top$.

Lemma 13 *Let W be a random matrix where $W_{i,j} = 1/p$ with probability p, and $W_{i,j} = 0$ otherwise. There exist universal constants c_1 and c_2, so that for any $\delta > 0$, if $p \geq c_1\frac{\log n}{\delta^2 n_1}$, then with probability at least $1 - \frac{1}{2}n^{-4}$, we have for any matrices $X \in \mathbb{R}^{n_1\times r}, Y \in \mathbb{R}^{n_2\times r}$:*

$$\|XY^\top\|_W^2 \leq (1+\delta)\|X\|_F^2\|Y\|_F^2 + c_2\sqrt{\frac{n}{p}}\|X\|_F\|Y\|_F \cdot \max_i\|X_i\| \cdot \max_j\|Y_j\|.$$

However, neither of these lemmas is applicable in our semi-random setting, because the weight matrix W is no longer chosen randomly by nature.

Lemma 12 is used in both the convex analysis (e.g., (Recht, 2011)) and many local analyses for the non-convex methods (e.g., (Sun and Luo, 2015)). Deterministic versions of Lemma 12 include Assumption A2 in (Bhojanapalli and Jain, 2014) and Assumption A3 in (Li et al., 2016). Unfortunately, we do not know whether Assumption A2 in (Bhojanapalli and Jain, 2014) is true even for random matrices, and we cannot guarantee Assumption A3 in (Li et al., 2016) because it is a condition that depends on the (unknown) ground truth.

Since we do not know how to obtain a deterministic version of Lemma 12, we turn our attention to Lemma 13. Lemma 13 is only used in more recent non-convex analyses (e.g., (Sun and Luo, 2015; Ge et al., 2017)). We replace Lemma 13 with the following (stronger version of the) lemma, which states that if W is close to the all ones matrix J (which is guaranteed by our preprocessing algorithm), then the norm of $\|XY^\top\|_F$ is preserved after we weight the entries by W. Recently, (Chen and Li, 2017) has independently obtained a deterministic inequality similar to Lemma 14.

Lemma 14 (Preserving the Norm via Spectral Properties) *For any matrices $X \in \mathbb{R}^{n_1\times r}$, $Y \in \mathbb{R}^{n_2\times r}$, and $W \in \mathbb{R}^{n_1\times n_2}$, we have*

$$|\|XY^\top\|_W^2 - \|XY^\top\|_F^2| \leq \|W - J\|\|X\|_F\|Y\|_F \max_i \|X_i\| \max_i \|Y_i\|.$$

Proof Recall that X_i is the i-th row of X, and $X \odot X \in \mathbb{R}^{n_1\times r^2}$ is the Katri-Rao product. We have

$$\langle (X\odot X)_i, (Y\odot Y)_j\rangle = \langle X_i\otimes X_i, Y_j\otimes Y_j\rangle = (X_i^\top Y_j)^2 = (XY^\top)_{i,j}^2.$$

As a result, we know $\|XY^\top\|_W^2 = \langle XY^\top, XY^\top\rangle_W = \mathrm{tr}((X \odot X)^\top W(Y \odot Y))$, and $\|XY^\top\|_F^2 = \langle XY^\top, XY^\top\rangle = \mathrm{tr}((X \odot X)^\top J(Y \odot Y))$.

We can also bound the Frobenius norm of the two product by: $\|X \odot X\|_F \le \|X\|_F \max_i \|X_i\|$, $\|Y \odot Y\|_F \le \|Y\|_F \max_i \|Y_i\|$. Therefore,

$$\begin{aligned}|\langle XY^\top, XY^\top\rangle_W - \langle XY^\top, XY^\top\rangle| &= |\,\mathrm{tr}((X \odot X)^\top (W-J)(Y \odot Y))| \\ &\le \|X \odot X\|_F \|(W-J)(Y \odot Y)\|_F \\ &\le \|W-J\| \|X \odot X\|_F \|(Y \odot Y)\|_F \\ &\le \|W-J\| \|X\|_F \|Y\|_F \max_i \|X_i\| \max_i \|Y_i\|. \qquad \blacksquare\end{aligned}$$

Using this lemma, as well as techniques in (Ge et al., 2017), we can prove the following theorem (see Appendix D for its proof):

Theorem 15 *For matrix completion problem with ground truth $M^\star \in \mathbb{R}^{n_1 \times n_2}$, let $\mu, r, \sigma_1^\star, \kappa^\star$ be the incoherence parameter, rank, largest singular value and condition number of $M^\star$. Fix any error parameter $0 < \epsilon < 1$. Suppose the weight matrix $W \in \mathbb{R}^{n_1 \times n_2}$ satisfies $\|W\|_\infty \le n_2$, $\|W\|_1 \le n_1$, and $\|W - J\| \le \frac{\epsilon c \sqrt{n_1 n_2}}{\mu^3 r^3 (\kappa^\star)^3}$ for a small enough universal constant c. Let $\alpha_1^2 = \frac{C\mu r \sigma_1^\star}{n_1}$, $\alpha_2^2 = \frac{C\mu r \sigma_1^\star}{n_2}$, $\lambda_1 = \frac{C^2 n_1}{\mu r \kappa^\star}$, and $\lambda_2 = \frac{C^2 n_1}{\mu r \kappa^\star}$ for some large enough universal constant C. Then, any local minimum (U, V) of the asymmetric Objective (1) satisfies $\|UV^\top - M^\star\|_F^2 \le \epsilon \|M^\star\|_F^2$.*

Our main result (Theorem 2) follows immediately from Corollary 9 and Theorem 15.

We can choose $\beta = O(\epsilon/(\mu^3 r^3 (\kappa^\star)^3))$ in Corollary 9 so that our preprocessing algorithm produces a weight matrix W that satisfies the requirements of Theorem 15. By Theorem 15, the non-convex objective with weight matrix W has no bad local optima. The pre-processing time is $\widetilde{O}(m/\beta^{O(1)}) = \widetilde{O}(m \cdot \mathrm{poly}(\mu, r, \kappa^\star, \epsilon^{-1}))$.

5. Conclusions

In this paper, we showed that even though non-convex approaches for matrix completion are not robust in the semi-random model, but it is possible to fix them using a pre-processing step. The pre-processing step solves a few convex programs (packing SDPs) to ameliorate the influence of the semi-random adversary. Unlike the full convex relaxation for matrix completion, our pre-processing step runs in nearly-linear time. Combining our pre-processing step with non-convex optimization gives an algorithm that is robust in the semi-random model, and at the same time enjoys the efficiency of the non-convex approaches.

An immediate open problem is whether we can prove the output of the pre-processing step allows non-convex optimization to recover the ground truth *exactly*. More broadly, we hope this work will inspire new ideas that make non-convex optimization more robust.

Acknowledgments

This work is supported by NSF CCF-1704656. We thank Qingqing Huang, Andrej Risteski, Srinadh Bhojanapalli, Yin Tat Lee for discussions at various stages of the work. Yu Cheng is also supported in part by NSF CCF-1527084, CCF-1535972, CCF-1637397, IIS-1447554, and NSF CAREER Award CCF-1750140.

References

Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma. Finding approximate local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.

Rudolf Ahlswede and Andreas J. Winter. Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48(3):569–579, 2002.

Zeyuan Allen Zhu, Zhenyu Liao, and Lorenzo Orecchia. Spectral sparsification and regret minimization beyond matrix multiplicative updates. In *Proc. 46th ACM Symp. on Theory of Computing*, pages 237–245, 2015.

Zeyuan Allen Zhu, Yin Tat Lee, and Lorenzo Orecchia. Using optimization to obtain a width-independent, parallel, simpler, and faster positive SDP solver. In *Proc. 27th ACM-SIAM Symp. on Discrete Algorithms*, pages 1824–1831, 2016.

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient and neural algorithms for sparse coding. In *Proc. 28th Conference on Learning Theory*, page 113149, 2015.

Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. *SIAM Journal on Computing*, 41(6):1704–1721, 2012.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

Srinadh Bhojanapalli and Prateek Jain. Universal matrix completion. In *Proc. 31st Intl. Conf. on Machine Learning*, pages 1881–1889, 2014.

Avrim Blum and Joel Spencer. Coloring random and semi-random k-colorable graphs. *Journal of Algorithms*, 19(2):204–234, 1995.

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

Emmanuel J Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *Information Theory, IEEE Transactions on*, 56(5):2053–2080, 2010.

Ji Chen and Xiaodong Li. Memory-efficient kernel PCA via partial matrix sampling and nonconvex optimization: a model-free analysis of local minima. *arXiv preprint arXiv:1711.01742*, 2017.

Yudong Chen and Martin J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint*, 1509.03025, 2015.

Dehua Cheng, Yu Cheng, Yan Liu, Richard Peng, and Shang-Hua Teng. Efficient sampling for Gaussian graphical models via spectral sparsification. In *Proc. 28th Conference on Learning Theory*, pages 364–390, 2015.

Michael B. Cohen, Rasmus Kyng, Gary L. Miller, Jakub W. Pachocki, Richard Peng, Anup B. Rao, and Shen Chen Xu. Solving SDD linear systems in nearly $m \log^{1/2} n$ time. In *Proc. 45th ACM Symp. on Theory of Computing*, pages 343–352, 2014.

Uriel Feige and Joe Kilian. Heuristics for semirandom graph problems. *Journal of Computer and System Sciences*, 63(4):639–671, 2001.

Uriel Feige and Robert Krauthgamer. Finding and certifying a large hidden clique in a semirandom graph. *Random Structures and Algorithms*, 16(2):195–208, 2000.

Rong Ge, Furong Huang, Chi Jin, and Yang Yuan. Escaping from saddle points—online stochastic gradient for tensor decomposition. *arXiv preprint arXiv:1503.02101*, 2015.

Rong Ge, Jason D Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. In *Proc. 28th Advances in Neural Information Processing Systems*, pages 2973–2981, 2016.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proc. 34th Intl. Conf. on Machine Learning*, pages 1233–1242, 2017.

Moritz Hardt. Understanding alternating minimization for matrix completion. In *Proc. 55th IEEE Symp. on Foundations of Computer Science*. IEEE, 2014.

Moritz Hardt and Mary Wootters. Fast matrix completion without the condition number. In *Proc. 27th Conference on Learning Theory*, pages 638–678, 2014.

Trevor Hastie, Rahul Mazumder, Jason Lee, and Reza Zadeh. Matrix completion and low-rank SVD via fast alternating least squares. *Journal of Machine Learning Research*, 2014.

Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proc. 44th ACM Symp. on Theory of Computing*, pages 665–674. ACM, 2013.

Rahul Jain and Penghui Yao. A parallel approximation algorithm for positive semidefinite programming. In *Proc. 52nd IEEE Symp. on Foundations of Computer Science*, pages 463–471, 2011.

Mark Jerrum. Large cliques elude the metropolis process. *Random Struct. Algorithms*, 3(4):347–360, 1992.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M. Kakade, and Michael I. Jordan. How to escape saddle points efficiently. *arXiv preprint arXiv:1703.00887*, 2017.

Jonathan A. Kelner, Lorenzo Orecchia, Aaron Sidford, and Zeyuan Allen Zhu. A simple, combinatorial algorithm for solving SDD systems in nearly-linear time. In *Proc. 44th ACM Symp. on Theory of Computing*, pages 911–920, 2013.

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

Raghunandan H Keshavan, Andrea Montanari, and Sewoong Oh. Matrix completion from noisy entries. *The Journal of Machine Learning Research*, 11:2057–2078, 2010b.

Yehuda Koren. The Bellkor solution to the Netflix grand prize. *Netflix prize documentation*, 81, 2009.

Ioannis Koutis, Gary L. Miller, and Richard Peng. A nearly-$m \log n$ time solver for SDD linear systems. In *Proc. 52nd IEEE Symp. on Foundations of Computer Science*, pages 590–598, 2011.

Ludek Kucera. Expected complexity of graph partitioning problems. *Discrete Applied Mathematics*, 57(2-3):193–212, 1995.

Rasmus Kyng and Sushant Sachdeva. Approximate Gaussian elimination for Laplacians - fast, sparse, and simple. In *Proc. 57th IEEE Symp. on Foundations of Computer Science*, pages 573–582, 2016.

Yin Tat Lee and He Sun. Constructing linear-sized spectral sparsification in almost-linear time. In *Proc. 56th IEEE Symp. on Foundations of Computer Science*, pages 250–269, 2015.

Yin Tat Lee and He Sun. An SDP-based algorithm for linear-sized spectral sparsification. In *Proc. 48th ACM Symp. on Theory of Computing*, pages 678–687, 2017.

Yuanzhi Li, Yingyu Liang, and Andrej Risteski. Recovery guarantee of weighted low-rank approximation via alternating minimization. In *Proc. 33rd Intl. Conf. on Machine Learning*, pages 2358–2367, 2016.

Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Approximation algorithms for semi-random partitioning problems. In *Proc. 43rd ACM Symp. on Theory of Computing*, pages 367–384. ACM, 2012.

Konstantin Makarychev, Yury Makarychev, and Aravindan Vijayaraghavan. Correlation clustering with noisy partial information. In *Proc. 28th Conference on Learning Theory*, pages 1321–1342, 2015.

Claire Mathieu and Warren Schudy. Correlation clustering with noisy input. In *Proc. 21st ACM-SIAM Symp. on Discrete Algorithms*, pages 712–728, 2010.

Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11(Aug):2287–2322, 2010.

Ankur Moitra, William Perry, and Alexander S Wein. How robust are reconstruction thresholds for community detection? In *Proc. 47th ACM Symp. on Theory of Computing*, pages 828–841, 2016.

Sahand Negahban and Martin J. Wainwright. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13(May):1665–1697, 2012.

Dohyung Park, Anastasios Kyrillidis, Constantine Caramanis, and Sujay Sanghavi. Non-square matrix sensing without spurious local minima via the Burer-Monteiro approach. *arXiv preprint arXiv:1609.03240*, 2016.

Richard Peng and Daniel A. Spielman. An efficient parallel solver for SDD linear systems. In *Proc. 45th ACM Symp. on Theory of Computing*, pages 333–342, 2014.

Richard Peng, Kanat Tangwongsan, and Peng Zhang. Faster and simpler width-independent parallel algorithms for positive semidefinite programming. *arXiv preprint arXiv:1201.5135v3*, 2016.

Amelia Perry and Alexander S Wein. A semidefinite program for unbalanced multisection in the stochastic block model. In *International Conference on Sampling Theory and Applications (SampTA)*, pages 64–67. IEEE, 2017.

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

Jasson DM Rennie and Nathan Srebro. Fast maximum margin matrix factorization for collaborative prediction. In *Proc. 22nd Intl. Conf. on Machine Learning*, pages 713–719. ACM, 2005.

Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *Journal of the ACM*, 54(4):21, 2007.

Christopher De Sa, Christopher Ré, and Kunle Olukotun. Global convergence of stochastic gradient descent for some non-convex matrix problems. In *Proc. 32nd Intl. Conf. on Machine Learning*, pages 2332–2341, 2015.

J. Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, 2015.

Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 40(6):1913–1926, 2011.

Daniel A. Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *SIAM Journal on Computing*, 40(4):981–1025, 2011.

Daniel A. Spielman and Shang-Hua Teng. Nearly linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *SIAM J. Matrix Analysis Applications*, 35(3):835–885, 2014.

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *Proc. 18th Conference on Learning Theory*, pages 545–560, 2005.

Nathan Srebro, Jason D. M. Rennie, and Tommi S. Jaakkola. Maximum-margin matrix factorization. In *Proc. 16th Advances in Neural Information Processing Systems*, pages 1329–1336, 2004.

Ruoyu Sun and Zhi-Quan Luo. Guaranteed matrix completion via nonconvex factorization. In *Proc. 56th IEEE Symp. on Foundations of Computer Science*, pages 270–289. IEEE, 2015.

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

Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via Procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

Tuo Zhao, Zhaoran Wang, and Han Liu. A nonconvex optimization framework for low rank matrix estimation. In *Proc. 27th Advances in Neural Information Processing Systems*, pages 559–567, 2015.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

Appendix A. Counter Examples for Non-convex Approaches

In this section, we give counter-examples to some non-convex methods for matrix completion. For simplicity, we give examples for the *weighted* version of the non-convex objective. These examples can be translated to the semi-random adversary model using standard sampling techniques.

We will give the counter-examples in a simpler setting where $M^\star$ is known to be symmetric, we have $M^\star$ is an $n \times n$ matrix that can be decomposed as $M^\star = U^\star(U^\star)^\top$. In this case, we optimize:

$$\min\ f(U) = \frac{1}{2}\|UU^\top - M^\star\|_W^2 + Q(U). \tag{3}$$

Here $Q(U)$ is the regularizer $\lambda \sum_{i=1}^n (\|U_i\|_2 - \alpha)_+^4$ (where $x_+ = \max\{x, 0\}$). Parameters λ, α in the regularizer is specified later (see Lemma 29). Our examples also work for the asymmetric Objective (1).

Example Where Objective (3) Has Spurious Local Minimum. We first give an example where Objective (3) has a spurious local minimum. This is a simple rank 1 case where the intended solution $M^\star = u^\star(u^\star)^\top$ is the all ones matrix, and $u^\star = (1, 1, ..., 1)^\top$ is the all ones vector.

In this example, all vectors will be represented by two blocks of size $n/2 \times 1$, and the value within each block will be the same; similarly all matrices will be partitioned into 2×2 blocks (where each block has size $n/2 \times n/2$), and entries within each block are the same.

For this example, we choose

$$u = \begin{pmatrix} \beta \\ -\beta \end{pmatrix}, \quad W = \begin{pmatrix} \gamma J & J \\ J & \gamma J \end{pmatrix},$$

for any parameters $2^{-1/4} < \beta \leq \frac{9}{10}$ and $\gamma = \frac{1+\beta^2}{1-\beta^2} < 10$.

Lemma 16 *For the setting of $M^\star$, u, W stated above, the objective function (3) has a local minimum at u.*

We will prove this lemma by second-order sufficient condition: gradient is 0 and Hessian is positive definite. Clearly u is incoherent, so the incoherence regularizer does not matter. We first check the gradient of $f(u)$ is 0. This is due to our choice of γ satisfies $\gamma(\beta^2 - 1) + (\beta^2 + 1) = 0$:

$$\nabla f(u) = [(W + W^\top) * (uu^\top - M^\star)]u = 2[W * (uu^\top - M^\star)]u = 0.$$

Next, we consider the Hessian of $f(u)$. For any vector $\delta \in \mathbb{R}^n$, we know that

$$\frac{1}{2}\delta^\top[\nabla^2 f(u)]\delta = \|\delta u^\top\|_W^2 + \langle 2uu^\top - M^\star, \delta\delta^\top\rangle_W.$$

We show this is strictly greater than 0 for all $\delta \neq 0$. Let $\delta = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}$ for $\delta_1, \delta_2 \in \mathbb{R}^{n/2}$. Notice that $\|\delta u^\top\|_W^2$ is non-negative, therefore, we have

$$\begin{aligned}
&\|\delta u^\top\|_W^2 + \langle 2uu^\top - M^\star, \delta\delta^\top\rangle_W \\
&\geq \langle 2uu^\top - M^\star, \delta\delta^\top\rangle_W \\
&= 2\beta^2 \left(\gamma\|\delta_1\|_1^2 + \gamma\|\delta_2\|_1^2 - 2\|\delta_1\|_1\|\delta_2\|_1\right) - \left(\gamma\|\delta_1\|_1^2 + \gamma\|\delta_2\|_1^2 + 2\|\delta_1\|_1\|\delta_2\|_1\right) \\
&\geq 2\beta^2(\gamma - 1)\left(\|\delta_1\|_1^2 + \|\delta_2\|_1^2\right) - (\gamma + 1)\left(\|\delta_1\|_1^2 + \|\delta_2\|_1^2\right) \\
&= \left(2\beta^2(\gamma - 1) - (\gamma + 1)\right)\left(\|\delta_1\|_1^2 + \|\delta_2\|_1^2\right) > 0.
\end{aligned}$$

The last step is due to $\delta \neq 0$ and our choice of β and γ.

Example Where SVD Initialization Gives the Wrong Subspace. Many other non-convex methods depend on SVD to do initialization (e.g., (Jain et al., 2013; Hardt, 2014)). Now we give an example where SVD cannot find the subspace correctly.

This is a rank-2 example. For simplicity, all vectors are divided into blocks of size $n/4 \times 1$, and matrices are divided into blocks of size $n/4 \times n/4$. We write these as 4×2 or 4×4 matrices, and they should be interpreted as blocks (constant multiplied by $J_{n/4\times 1}$ or $J_{n/4\times n/4}$). The matrix $\mathrm{Diag}(4, 1)$ is a 2×2 diagonal matrix.

$$M^\star = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ 1 & -1 \\ 1 & -1 \end{pmatrix} \mathrm{Diag}(4, 1) \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \end{pmatrix}$$

$$W = \begin{pmatrix} 2 & 1 & 2 & 1 \\ 1 & 2 & 1 & 2 \\ 2 & 1 & 2 & 1 \\ 1 & 2 & 1 & 2 \end{pmatrix}; \qquad W * M^\star = \begin{pmatrix} 10 & 5 & 6 & 3 \\ 5 & 10 & 3 & 6 \\ 6 & 3 & 10 & 5 \\ 3 & 6 & 5 & 10 \end{pmatrix}$$

Lemma 17 *Under the above setting for $M^\star, W$, the top two singular vectors of $W * M^\star$ are (up to normalization) $(1, 1, 1, 1)$ and $(1, -1, 1, -1)$. The principled angle between this subspace and the true subspace span$(M^\star)$ is 90 degrees.*

This lemma is easy to verify numerically by computing the SVD of the 4×4 matrix. The SVD of the original matrix follows the same block structure.

Converting Weighted Examples to Semi-Random Examples. In order to get counter examples in the semi-random model, pick a probability of observation p (we need $p \geq \mathrm{poly}(r)\log(n)/n$, and in our examples p can be as large as $1/10$). The semi-random adversary will reveal entry (i, j) with probability $p_{i,j} = pW_{i,j} \geq p$ for the W given in the examples. This way, the expectation of the observed entries is exactly $W * M^\star$ (after scaling by $1/p$), and the expectation of the objective function is equal to $\|UU^\top - M^\star\|_W^2$.

For the first example, by standard concentration results, we know that the gradient and Hessian of the objective function are close to their expectations, so there is a local minimum near u. For the second example, by standard random matrix theory, we know when p is large enough the top singular space of the observed matrix is close to the top singular space of $W * M^\star$ (and thus far from the correct space).

Appendix B. Omitted Proofs from Section 3

In this section, we give more details about Section 3 and prove Lemmas 7 and 8.

Recall that we are given a set of input vectors $\{v_i\}_{i=1}^m$ where each $v_i = L^{-1/2}b_i$ for a fixed Laplacian L and some b_i representing an edge, and we assume that there exist weights w_i such that $(1-\beta)I \preceq \sum_{i=1}^m w_i v_i v_i^\top \preceq (1+\beta)I$. The goal is to compute a set of weights $\widetilde{w}_i$ in nearly-linear time so that $(1-O(\beta)-\epsilon)I \preceq \sum_{i=1}^m \widetilde{w}_i v_i v_i^\top \preceq I$.

We maintain barrier values u and ℓ, and a weighted sum $A = \sum_{i=1}^m \widetilde{w}_i v_i v_i^\top$ such that $\ell I \prec A \prec uI$. It is worth noting that we never explicitly compute the vectors $v_i = L^{-1/2}b_i$. We store A by keeping track of the weights w_i. When updating the weights, we approximate the quantities $v_i^\top C v_i$ (for some matrix $C = C(A)$) simultaneously for all i in nearly-linear time (see Lemma 20).

We use the following potential functions proposed in (Lee and Sun, 2017) to measure how far A is away from the barriers:

$$\begin{aligned}\Phi_{u,\ell}(A) &= \Phi_u(A) + \Phi_\ell(A), \text{ where}\\ \Phi_u(A) &= \operatorname{tr}\exp\left((uI-A)^{-1}\right),\\ \Phi_\ell(A) &= \operatorname{tr}\exp\left((A-\ell I)^{-1}\right).\end{aligned}$$

The derivatives of the potential functions with respect to A are

$$\begin{aligned}\nabla\Phi_u(A) &= \exp\left((uI-A)^{-1}\right)(uI-A)^{-2},\\ \nabla\Phi_\ell(A) &= -\exp\left((A-\ell I)^{-1}\right)(A-\ell I)^{-2}.\end{aligned}$$

By convexity we have

$$\Phi_{u,\ell}(A+\Delta) \geq \Phi_{u,\ell}(A) + \nabla\Phi_u(A)\bullet\Delta + \nabla\Phi_\ell(A)\bullet\Delta.$$

The following lemma from Lee and Sun (2017) shows that when Δ is small, the first-order approximation of $\Phi_{u,\ell}(A+\Delta)$ is a good estimation.

Lemma 18 (Lee and Sun (2017)) *Let A be a symmetric matrix. Let $\ell < u$ be barrier values such that $u-\ell \leq 1$ and $\ell I \prec A \prec uI$. Assume that $0 \preceq \Delta$, $\Delta \preceq \epsilon(uI-A)^2$, and $\Delta \preceq \epsilon(A-\ell I)^2$ for $\epsilon \leq 1/10$. Then,*

$$\begin{aligned}\Phi_u(A+\Delta) &\leq \Phi_u(A) + (1+2\epsilon)\nabla\Phi_u(A)\bullet\Delta = \Phi_u(A) + (1+2\epsilon)C_-\bullet\Delta,\\ \Phi_\ell(A+\Delta) &\leq \Phi_\ell(A) + (1-2\epsilon)\nabla\Phi_\ell(A)\bullet\Delta \;= \Phi_\ell(A) - (1-2\epsilon)C_+\bullet\Delta;\\ \Phi_u(A-\Delta) &\leq \Phi_u(A) - (1-2\epsilon)\nabla\Phi_u(A)\bullet\Delta = \Phi_u(A) - (1-2\epsilon)C_-\bullet\Delta,\\ \Phi_\ell(A-\Delta) &\leq \Phi_\ell(A) - (1+2\epsilon)\nabla\Phi_\ell(A)\bullet\Delta \;= \Phi_\ell(A) + (1+2\epsilon)C_+\bullet\Delta.\end{aligned}$$

Recall that for notational convenience, we write

$$\begin{aligned}C_+ &= -\nabla\Phi_\ell(A) = \exp\left((A-\ell I)^{-1}\right)(A-\ell I)^{-2},\\ C_- &= \nabla\Phi_u(A) \quad = \exp\left((uI-A)^{-1}\right)(uI-A)^{-2}.\end{aligned}$$

When $\ell I \prec A \prec uI$, both C_+ and C_- are PSD matrices. Recall that $\rho = (\lambda_{\min}\{uI - A, A - \ell I\})^2$. We are interested in the following packing SDP:

$$\begin{array}{ll} \text{maximize} & (C_+ - C_-) \bullet X \\ \text{subject to} & X \preceq \epsilon\rho I, \quad X = \sum_{i=1}^m x_i v_i v_i^\top \text{ (which implies } 0 \preceq X). \end{array} \tag{2}$$

The constraint $X \preceq \epsilon\rho I$ implies that $X \preceq \epsilon(uI - A)^2$ and $X \preceq \epsilon(A - \ell I)^2$. Thus, by Lemma 18, when $\epsilon \leq 1/10$, the first-order approximation of the potential function is accurate:

$$\Phi_{u,\ell}(X) - (C_+ - C_-) \bullet X \leq \Phi_{u,\ell}(A + X) \leq \Phi_{u,\ell}(A) - ((1-2\epsilon)C_+ - (1+2\epsilon)C_-) \bullet X.$$

Let $C = C_+ - C_-$. The SDP in (2) is naturally trying to find an X to maximize the $C \bullet X$, while making sure $C \bullet X$ is a good approximation to the reduction of the potential function.

Ideally, we would like to have $X = \epsilon\rho I$ so that A grows equally in each dimension, and the potential function stays the same:

$$\Phi_{u+\epsilon\rho,\ell+\epsilon\rho}(A + \epsilon\rho I) = \Phi_{u,\ell}(A).$$

When $X = \epsilon\rho I$, the objective value of the SDP is

$$(C_+ - C_-) \bullet X = \epsilon\rho(\operatorname{tr}(C_+) - \operatorname{tr}(C_-)).$$

While in general I may not be in the span of the rank-one matrices, we will show in Lemma 20 that we can compute an X with

$$(C_+ - C_-) \bullet X \geq \frac{\epsilon\rho}{2}\left((1 - \beta - 2\epsilon)\operatorname{tr}(C_+) - (1 + \beta + 2\epsilon)\operatorname{tr}(C_-)\right).$$

We first prove Lemma 8, which states that the potential function does not increase if Δ_j in each iteration satisfies Lemma 20.

Proof [Proof of Lemma 8] We want to show $\Phi_{u_{j+1},\ell_{j+1}}(A_{j+1}) \leq \Phi_{u_j,\ell_j}(A_j)$, where $A_{j+1} = A_j + \Delta_j$.

When we increase the weights $\{w_i\}$ and expand A, the lower barrier potential Φ_ℓ decreases (since A gets farther away from ℓI) and the upper barrier Φ_u increases (since A gets closer to uI). When we increase the barrier values u and ℓ, the opposite happens: Φ_ℓ increases and Φ_u decreases. Intuitively, the proof works by carefully increasing u and ℓ to cancel out the change due to adding Δ_j, while making sure both u and ℓ increase at roughly the same rate.

Recall that $C_+ = -\nabla\Phi_{\ell_j}(A_j)$ and $C_- = \nabla\Phi_{u_j}(A_j)$. Formally, we have

$$\begin{aligned} &\Phi_{u_{j+1},\ell_{j+1}}(A_{j+1}) \\ &= \Phi_{u_j}(A_{j+1} - \delta_{u,j}I) + \Phi_{\ell_j}(A_{j+1} - \delta_{\ell,j}I) \\ &\leq \Phi_{u_j,\ell_j}(A_{j+1}) - (1-2\epsilon)\nabla\Phi_{u_j}(A_{j+1}) \bullet (\delta_{u,j}I) - (1+2\epsilon)\nabla\Phi_{\ell_j}(A_{j+1}) \bullet (\delta_{\ell,j}I) \\ &\hfill \text{(Lemma 18)} \\ &\leq \Phi_{u_j,\ell_j}(A_{j+1}) - (1-2\epsilon)\nabla\Phi_{u_j}(A_j) \bullet (\delta_{u,j}I) - (1+2\epsilon)\nabla\Phi_{\ell_j}(A_j) \bullet (\delta_{\ell,j}I) && (4) \\ &= \Phi_{u_j,\ell_j}(A_{j+1}) - (1-2\epsilon)\delta_{u,j}\operatorname{tr}(C_-) + (1+2\epsilon)\delta_{\ell,j}\operatorname{tr}(C_+). && (5) \end{aligned}$$

Our choice of $\delta_{u,j}, \delta_{\ell,j}$ satisfies that $\delta_{u,j}, \delta_{\ell,j} \leq \epsilon\rho$ when $\epsilon, \beta \leq 1/10$, which allows us to apply Lemma 18. Step (4) uses $A_j \preceq A_{j+1}$ and the fact that, for any $\ell I \prec A_1 \preceq A_2 \prec uI$, we have $0 \preceq \nabla\Phi_u(A_1) \preceq \nabla\Phi_u(A_2)$ and $0 \preceq -\nabla\Phi_\ell(A_2) \preceq -\nabla\Phi_\ell(A_1)$.

We continue to bound the change of the potential function when we set $A_{j+1} = A_j + \Delta_j$ for any Δ_j that satisfies Lemma 20.

$$\begin{aligned}
&\Phi_{u_j,\ell_j}(A_{j+1}) = \Phi_{u_j,\ell_j}(A_j + \Delta_j) \\
&\leq \Phi_{u_j,\ell_j}(A_j) + (1+2\epsilon)C_- \bullet \Delta_j - (1-2\epsilon)C_+ \bullet \Delta_j && \text{(Lemma 18)} \\
&= \Phi_{u_j,\ell_j}(A_j) - (C_+ - C_-) \bullet \Delta_j + 2\epsilon(C_+ + C_-) \bullet \Delta_j \\
&= \Phi_{u_j,\ell_j}(A_j) - (C_+ - C_-) \bullet \Delta_j + 2\epsilon^2\rho \operatorname{tr}(C_+ + C_-) && (\Delta_j \leq \epsilon\rho I) \\
&\leq \Phi_{u_j,\ell_j}(A_j) - \frac{\epsilon\rho}{2}\left[(1-\beta-\epsilon)\operatorname{tr}(C_+) - (1+\beta+\epsilon)\operatorname{tr}(C_+)\right] + 2\epsilon^2\rho \operatorname{tr}(C_+ + C_-) \\
& && \text{(Lemma 20)} \\
&= \Phi_{u_j,\ell_j}(A_j) + \frac{\epsilon\rho(1+\beta+5\epsilon)}{2}\operatorname{tr}(C_-) - \frac{\epsilon\rho(1-\beta-5\epsilon)}{2}\operatorname{tr}(C^+). && (6)
\end{aligned}$$

We conclude the proof by comparing lines (5) and (6), and setting $\delta_{u,j} = \frac{\epsilon\rho(1+\beta+5\epsilon)}{2(1-2\epsilon)}$ and $\delta_{\ell,j} = \frac{\epsilon\rho(1-\beta-5\epsilon)}{2(1+2\epsilon)}$. The trace terms cancel out and we get $\Phi_{u_{j+1},\ell_{j+1}}(A_{j+1}) \leq \Phi_{u_j,\ell_j}(A_j)$ as needed. ■

We remark that the best we can hope for is to increase the upper and lower barriers at a rate of $1+\beta$ vs. $1-\beta$ (which is the case as $\epsilon \to 0$), because the ground-truth is a set of weights w_i with $(1-\beta)I \preceq \sum_i w_i v_i v_i^\top \preceq (1+\beta)I$. Our algorithm only achieves a ratio of $1+O(\beta+\epsilon)$ vs. $1-O(\beta+\epsilon)$ for several reasons: (1) the error in the first-order approximation of the potential function, (2) we solve the SDP approximately, and (3) we use Taylor expansion and Johnson-Lindenstrauss to speed up the computation.

We break Lemma 7 into two lemmas and prove them separately. Lemma 19 states that SDP (2) has a good solution. Lemma 20 shows that we can compute ρ and solve this packing SDP (2) approximately in nearly-linear time.

Lemma 19 *Let A be a symmetric matrix. Let $\ell < u$ be barrier values such that $\ell I \prec A \prec uI$. The SDP in (2) has a solution X with*

$$C \bullet X \geq \epsilon\rho\left(\frac{1-\beta}{1+\beta}\operatorname{tr}(C_+) - \operatorname{tr}(C_-)\right).$$

Proof When $X = \epsilon\rho I$, the objective value is $C \bullet X = \epsilon\rho \operatorname{tr}(C)$. Note that C is not PSD, so when $X \approx \epsilon\rho I$, we need to split C into the difference of two PSD matrices C_+ and C_- to bound the error.

Recall that there exists a set of weights w_i with $(1-\beta)I \preceq \sum_{i=1}^m w_i v_i v_i^\top \preceq (1+\beta)I$. We look at a solution of this SDP with $X = \frac{\rho}{1+\beta}\sum_{i=1}^m w_i v_i v_i^\top$. It follows directly that X is feasible since X is a weighted sum of $v_i v_i^\top$ and $X \preceq \epsilon\rho I$. For the objective value, since $\frac{1-\beta}{1+\beta}\epsilon\rho I \preceq X \preceq \epsilon\rho I$,

$$(C_+ - C_-) \bullet X \geq C_+ \bullet (\frac{1-\beta}{1+\beta}\epsilon\rho I) - C_- \bullet (\epsilon\rho I) = \epsilon\rho\left(\frac{1-\beta}{1+\beta}\operatorname{tr}(C_+) - \operatorname{tr}(C_-)\right). \quad ■$$

We now provide details about how to implement Algorithm 1 in nearly-linear time. We remark that similar implementations were shown in Allen Zhu et al. (2015); Lee and Sun (2015, 2017).

Lemma 20 *Fix $0 < \beta, \epsilon \le 1/10$. Given an $n \times n$ matrix $A = \sum_{i=1}^m w_i v_i v_i^\top$ represented by a set of weights $\{w_i \ge 0\}_{i=1}^m$, let $\ell < u$ be barrier values such that $u - \ell \le 1$ and $(\ell + g)I \prec A \prec (u - g)I$ for some gap $g = \Omega(\log^{-2} n)$.*

We can compute ρ and weights $\{\widetilde{w}_i\}_{i=1}^m$ in $\widetilde{O}(m/\epsilon^5)$ time, such that with high probability,

(1) $\rho \in [1 - \epsilon, 1] \cdot (\lambda_{\min}\{uI - A, A - \ell I\})^2$;

(2) $X = \sum_{i=1}^m \widetilde{x}_i v_i v_i^\top$ satisfies $X \preceq \epsilon \rho I$ and

$$(C_+ - C_-) \bullet X \ge \frac{\epsilon\rho}{2} \left((1 - \beta - \epsilon) \operatorname{tr}(C_+) - (1 + \beta + \epsilon) \operatorname{tr}(C_-) \right).$$

Proof Recall that $v_i = L^{-1/2} b_i$ for a fixed Laplacian L. Let $\widehat{L} = \sum_i w_i b_i b_i^\top$ be the Laplacian specified by the weights of A. In this proof, we will frequently use the following fact,

$$A = \sum_{i=1}^m w_i v_i v_i^\top = L^{-1/2} \sum_{i=1}^m w_i b_i b_i^\top L^{-1/2} = L^{-1/2} \widehat{L} L^{-1/2}.$$

(1) We show how to compute $\rho \in [1 - \epsilon, 1] \cdot \lambda_{\min}(uI - A)^2$. The approach is similar for $A - \ell I$. It is sufficient to compute $\rho \approx_{\epsilon/2} \lambda_{\min}(uI - A)^2$. [7] By Lemma 22, there exists a degree $\widetilde{O}\left(\frac{\log(1/\epsilon)}{g}\right) = \widetilde{O}(\log^2 n \log(1/\epsilon))$ polynomial $p(A)$ such that $p(A) \approx_{\epsilon/4} (uI - A)^{-2}$. Since $\lambda_{\max}(p(A)) \approx_{\epsilon/4} \lambda_{\max}((uI - A)^{-2}) = \left(\lambda_{\min}(uI - A)^2\right)^{-1}$, it is sufficient to approximate $\lambda_{\max}(p(A))$.

Observe that for any $n \times n$ PSD matrix M,

$$\lambda_{\max}(M) \le \left(\operatorname{tr}\left(M^{2k}\right)\right)^{1/2k} \le n^{1/2k} \lambda_{\max}(M).$$

In particular, for $k = O(\log n/\epsilon)$ we can get $\left(\operatorname{tr}(p(A)^{2k})\right)^{1/2k} \approx_{\epsilon/4} \lambda_{\max}(p(A))$, and thus, we can return $\rho = \left(\operatorname{tr}(p(A)^{2k})\right)^{-1/2k} \approx_{\epsilon/2} \lambda_{\min}(uI - A)^2$.

It remains to show that we can approximate

$$\operatorname{tr}(p(A)^{2k}) = \operatorname{tr}(p(L^{-1/2} \widehat{L} L^{-1/2})^{2k}) = \operatorname{tr}(p(L^{-1} \widehat{L})^{2k}).$$

Let $M = p(L^{-1}\widehat{L})^k$ so $\operatorname{tr}(p(A)^{2k}) = \operatorname{tr}(M^2)$. We approximate each diagonal entry of M^2 by writing it as $\left(M^2\right)_{i,i} = \chi_i^\top M M \chi_i = \|M\chi_i\|_2^2$, where χ_i denote the i-th standard basis vector. By the Johnson-Lindenstrauss lemma, we can generate a random $O(\log n/\epsilon^2) \times n$ matrix Q, so that with high probability, for all $1 \le i \le n$,

$$\|M\chi_i\|_2^2 = \|QM\chi_i\|_2^2.$$

Note that $QM\chi_i$ is the i-th column of QM. We can compute (approximately) $QM = \left(MQ^\top\right)^\top$ by multiplying each column of $Q^\top$ through M.

[7] We write $a \approx_\epsilon b$ for $\exp(-\epsilon)a \le b \le \exp(\epsilon)a$. This extends naturally to PSD matrices, where $A \approx_\epsilon B$ means $\exp(-\epsilon)A \preceq B \preceq \exp(\epsilon)A$. It is sufficient to approximate ρ up to a factor of $1 \pm \exp(\epsilon/2)$ because $\exp(\epsilon) \le \frac{1}{1-\epsilon}$.

This can be done in time $\widetilde{O}(n/\epsilon^5)$, because $Q^\top$ has $O(\log n/\epsilon^2)$ columns, and matrix-vector multiplication with $M = p(L^{-1}\widehat{L})^k$ can be implemented using $k \cdot \deg(p) = \widetilde{O}(\log^3 n/\epsilon)$ matrix-vector multiplications with $L^{-1}\widehat{L}$. We will show that matrix-vector multiplication with $L^{-1}\widehat{L}$ can be done in time $\widetilde{O}(n/\epsilon^2)$, so the overall running time is $\widetilde{O}(n/\epsilon^5)$.

Recall that the number of edges in $\widehat{L}$ is at most m. Let m' denote the number of edges in L. W.l.o.g., we can assume both $m, m' = O(n/\epsilon^2)$ by sparsifying the input graphs first. Therefore, one matrix-vector multiplication with $L^{-1}\widehat{L}$ can be done in time $\widetilde{O}(n/\epsilon^2)$, by first multiplying the vector through $\widehat{L}$, and then solving a linear system in L in $\widetilde{O}(m' \log(1/\epsilon))$ time (Spielman and Teng, 2014; Koutis et al., 2011; Kelner et al., 2013; Peng and Spielman, 2014; Cheng et al., 2015; Cohen et al., 2014; Kyng and Sachdeva, 2016).

(2) Since we represent the variable X of the SDP by a set of weights $\{x_i\}_{i=1}^m$, the objective function is of the form $C \bullet X = C \bullet \left(\sum_{i=1}^m x_i v_i v_i^\top\right) = \sum_{i=1}^m x_i (v_i^\top C v_i)$. Let $c \in \mathbb{R}^m$ be a vector with $c_i = v_i^\top C v_i$. The SDP in (2) can be rewritten as

$$\begin{array}{ll} \text{maximize} & c^\top x \\ \text{subject to} & \sum_{i=1}^m x_i (v_i v_i^\top) \preceq \epsilon\rho I. \end{array}$$

This is a packing SDP that can be solved in polylogarithmic iterations (see, e.g., (Jain and Yao, 2011; Allen Zhu et al., 2016; Peng et al., 2016)). Formally, we use Lemma 21 from Allen Zhu et al. (2016). Because Lemma 21 returns a solution X with $\mathbb{E}\left[C \bullet X\right] \geq \frac{4}{5}\mathrm{OPT}$, it must return a $\frac{3}{5}$ approximation with probability at least $1/2$. Since $\frac{3}{5} > \frac{1+\beta}{2}$, we can invoke Lemma 21 $O(\log n)$ times so that we get $\frac{1+\beta}{2}$-approximation with high probability. We assume this event happens for the rest of the proof.

Let $c_i^+ = v_i^\top C_+ v_i$ and $c_i^- = v_i^\top C_- v_i$. If we can approximate c^+ and c^- by a (multiplicative) factor of $1 \pm \frac{\epsilon}{2}$, we have

$$\begin{aligned} (c^+ - c^-)^\top x &\geq \frac{1+\beta}{2}\mathrm{OPT} - \frac{\epsilon}{2}(c^+ + c^-)^\top x & \text{(Lemma 21)} \\ &\geq \frac{1+\beta}{2}\mathrm{OPT} - \frac{\epsilon^2\rho}{2}\operatorname{tr}(C^+ + C^-) & (X \preceq \epsilon\rho I) \\ &\geq \frac{1+\beta}{2}\epsilon\rho\left(\frac{1-\beta}{1+\beta}\operatorname{tr}(C_+) - \operatorname{tr}(C_-)\right) - \frac{\epsilon^2\rho}{2}\operatorname{tr}(C^+ + C^-). & \text{(Lemma 19)} \\ &= \frac{\epsilon\rho}{2}\left((1-\beta-\epsilon)\operatorname{tr}(C_+) - (1+\beta+\epsilon)\operatorname{tr}(C_-)\right). \end{aligned}$$

Finally, we will show how to approximate c_i^- by a factor of $1 \pm \frac{\epsilon}{2}$ for all $1 \leq i \leq m$ in time $\widetilde{O}(m/\epsilon^{O(1)})$. The algorithms for approximating c_i^+ and implementing the oracle required by Lemma 21 follow from a similar approach.[8]

Recall that $C_- = \exp((uI - A)^{-1})(uI - A)^{-2}$, where $A = L^{-1/2}\widehat{L}L^{-1/2}$. By Lemma 22 the assumption that $g = \Omega(\log^{-2} n)$, there exists a degree $\widetilde{O}\left(\frac{\log(1/\epsilon)}{g^2}\right) = \widetilde{O}(\log^4 n \log(1/\epsilon))$

[8] We remark that the problem of approximating these quantities is akin to that of approximating (relative) effective resistances (Spielman and Srivastava, 2011; Allen Zhu et al., 2015; Lee and Sun, 2015), and a nearly-linear time algorithm for computing the same quantities was shown in (Lee and Sun, 2017).

polynomial $q(A)$ such that

$$q(A) \approx_{\epsilon/6} \exp\left(\frac{1}{2}(uI - A)^{-1}\right)(uI - A)^{-1}.$$

Because both sides are matrix polynomials of A, we can diagonalize them simultaneously so that the approximation only happens to the eigenvalues. Therefore,

$$(q(A))^2 \approx_{\epsilon/3} \exp((uI - A)^{-1})(uI - A)^{-2},$$

which implies $v_i^\top q(A)^2 v_i \in [1 \pm \frac{\epsilon}{2}]v_i^\top C_- v_i$, since $\exp(\epsilon/3) \le 1 + \epsilon/2$ when $\epsilon \le \frac{1}{10}$.

Recall that m, m' denotes the number of edges in $\widehat{L}$ and L. Recall that $L = B^\top B$ where $B \in \mathbb{R}^{m' \times n}$ is the edge-vertex incident matrix of L. Fix some $1 \le i \le m$. Let $v_i = L^{-1/2} b_i$ where $b_i = \chi_u - \chi_{u'}$ for the i-th edge (u, u').

For any $1 \le i \le m$, we have

$$\begin{aligned} v_i^\top C_- v_i &\approx_{\epsilon/3} \|q(A)v_i\|_2^2 \\ &= \left\|q(L^{-1/2}\widehat{L}L^{-1/2})L^{-1/2}b_i\right\|_2^2 \\ &= \left\|L^{1/2}q(L^{-1}\widehat{L})L^{-1}b_i\right\|_2^2 \\ &= \left\|Bq(L^{-1}\widehat{L})L^{-1}b_i\right\|_2^2 \\ &= \left\|Bq(L^{-1}\widehat{L})L^{-1}(\chi_u - \chi_{u'})\right\|_2^2. \end{aligned}$$

So the quantities $\{c_i^-\}_{i=1}^m$ are just the squared distances between the m-dimensional points $\{Bq(L^{-1}\widehat{L})L^{-1}\chi_u\}_{u \in V}$. We invoke the Johnson-Lindenstrauss lemma and generate a random $O(\log n/\epsilon^2) \times m$ matrix Q, so that with high probability, for all $1 \le u, u' \le n$,

$$\|Bq(L^{-1}\widehat{L})L^{-1}(\chi_u - \chi_{u'})\|_2 \approx_{\epsilon/6} \|QBq(L^{-1}\widehat{L})L^{-1}(\chi_u - \chi_{u'})\|_2.$$

Recall that A_i is the i-th row of a matrix A. Let $Z = QBq(L^{-1}\widehat{L})L^{-1}$ and $Y = QBq(L^{-1}\widehat{L})$. Both Y and Z have $O(\log n/\epsilon^2)$ rows and n columns. We have $Z^\top = L^{-1}Y^\top$, which allows us to approximate each $(Z_i)^\top = L^{-1}(Y_i)^\top$ by solving a linear system in L. The time it takes to solve $O(\log n/\epsilon^2)$ linear systems in L is $\widetilde{O}(n/\epsilon^4)$, because we can assume $m, m' = O(n/\epsilon^2)$ by sparsifying the input graphs.

We can compute $Y^\top = q(L^{-1}\widehat{L})B^\top Q^\top$ in $\widetilde{O}(n/\epsilon^4)$ time, since we can perform matrix-vector multiplication with $q(L^{-1}\widehat{L})$ in time $\widetilde{O}(m/\epsilon^2)$, and $B^\top Q^\top$ can be computed in $\widetilde{O}(n/\epsilon^4)$ time because B has $2m' = O(n/\epsilon^2)$ non-zeros and Q has $O(\log n/\epsilon^2)$ rows. ■

The overall running time of Algorithm 1 is $\widetilde{O}(m/\epsilon^7)$, because there are at most $O(\log n/\epsilon^2)$ iterations (shown in Section 3), and each iteration can be implemented to run in time $\widetilde{O}(m/\epsilon^5)$ by Lemma 20.

One of our main contributions is conceptual: we show that the framework of Batson et al. (2012) can be applied to a much broader settings to obtain scalable algorithms. On a technical level, because there exists a hidden set S whose sum is only *approximately* equal to I, the optimal solution to the SDP will be worse, so we need to carefully control the error caused by this, and move the barriers at a slightly different rate. Our analysis is considerably simpler than that in (Lee and Sun, 2017), partly because we do not require the output weights to be sparse; We also take care of two minor issues with (Lee and Sun, 2017): They assumed ρ can be computed exactly for simplicity, and they proved Taylor expansion of C_- can be truncated (where it should be $C_-^{1/2}$ as in Lemma 22).

Lemma 21 (Allen Zhu et al. (2016)) *Consider the following SDP with $M_i \succeq 0$ and $c \in \mathbb{R}^m$.*

$$\begin{array}{ll} \textit{maximize} & c^\top x \\ \textit{subject to} & x \geq 0 \\ & \sum_{i=1}^m x_i M_i \preceq I. \end{array}$$

Suppose c is given explicitly and we have access to M_i via an oracle $\mathcal{O}_{\eta,\delta}$ which on input $x \in \mathbb{R}^m$ outputs a vector $y \in \mathbb{R}^m$ such that

$$y_i \in (1 \pm \frac{\delta}{2})\; M_i \bullet \exp\left(\eta\Big(\sum_i x_i M_i - I\Big)\right)$$

in time $T_{\eta,\delta}$ for any $x \in \mathbb{R}^m$ such that $x \geq 0$ and $\sum_{i=1}^m x_i M_i \preceq 2I$. Then, we can output an x in time $\widetilde{O}(T_{\eta,\delta} \log m/\delta^3)$ such that

$$\mathbb{E}\left[c^\top x\right] \geq (1 - O(\delta)) \cdot \mathrm{OPT} \textit{ with } \sum_{i=1}^m x_i M_i \preceq I.$$

Lemma 22 *Let A be a real symmetric matrix. When $(u-1)I \prec A \prec (u-g)I$ for some $0 < g < 1$, we can compute*

(1) A polynomial $p(A)$ of degree $O\left(\frac{\log(1/(\epsilon g))}{g}\right)$ such that $p(A) \approx_\epsilon (uI - A)^{-2}$.

(2) A polynomial $q(A)$ of degree $O\left(\frac{\log(1/(\epsilon g))}{g^2}\right)$ such that $q(A) \approx_\epsilon \exp\left(\frac{1}{2}(uI - A)^{-1}\right)(uI - A)^{-1}$.

Proof The lemma is proved by truncating Taylor expansions. Because A is symmetric, the matrix polynomials $p(A)$ and $q(A)$ can be diagonalized simultaneously with A. Therefore, it is sufficient to prove such polynomials exist for scalars.

(1) Let $f(x) = x^{-2}$. Let $p(a) = \widehat{p}(u - a)$, and we define $\widehat{p}(\cdot)$ to be the first d terms of the Taylor expansion of $f(x)$ at $x = 1$.

$$f(x) = \sum_{i=0}^d \frac{f^{(i)}(1)}{i!}(x-1)^n + \frac{1}{d!}\int_1^x f^{(d+1)}(t)(x-t)^d dt.$$

We know that all eigenvalues of $(uI - A)$ are in the interval $(g, 1)$. For any $x \in (g, 1)$, there exists some $d = \left(g^{-1}\log(1/(\epsilon g))\right)$ such that the remainder of the Taylor series satisfies

$$\begin{aligned} \left| f(x) - \sum_{i=0}^{d} \frac{f^{(i)}(1)}{i!}(x-1)^n \right| &= \left| \frac{1}{d!} \int_1^x f^{(d+1)}(t)(x-t)^d dt \right| \\ &= \frac{(1-x)^{d+1}(1+x+dx)}{x^2} \\ &\le (1-g)^{d+1}\frac{d+2}{g^2} \le \epsilon. \end{aligned}$$

(2) Let $h(x) = \exp\left(\frac{1}{2}x^{-1}\right)x^{-1}$. Let $q(a) = \widehat{q}(u-a)$, and we define $\widehat{q}(\cdot)$ to be the first d terms of the Taylor expansion of $h(x)$ at $x = 1$.

For any $x \in (g, 1)$ and $t \in [x, 1]$, h is holomorphic on a neighborhood of the ball $B := \{z \in \mathbb{R} : |z - t| \le r\}$ for $r = t - x/2$, so we can bound the coefficients of the Taylor expansion using Cauchy's estimates.

$$\frac{1}{(d+1)!}h^{(d+1)}(t) \le r^{-d-1}\sup_{z\in B}|h(z)| \le r^{-d-1}\cdot 2\exp(x^{-1})x^{-1}.$$

There exists some $d = O\left(g^{-2}\log(1/(\epsilon g))\right)$ such that the remainder at $x \in (g, 1)$ satisfies

$$\begin{aligned} |h(x) - \widehat{q}(x)| &= \left| \frac{1}{d!}\int_1^x h^{(d+1)}(t)(x-t)^d dt \right| \\ &\le 2(d+1)\exp(x^{-1})x^{-1}\int_x^1 \frac{(t-x)^d}{(t-x/2)^{d+1}}dt \\ &\le 4(d+1)\exp(x^{-1})x^{-2}\int_x^1 \left(1-\frac{x}{2}\right)^d dt \\ &\le 4(d+1)\exp(g^{-1})g^{-2}\left(1-\frac{g}{2}\right)^d \le \epsilon. \end{aligned}$$

■

Appendix C. Omitted Proof from Section 3.1

In this section, we restate Lemmas 10 and 11 and prove them.

Lemma 10 *Let G denote the $n_1 \times n_2$ complete bipartite graph. We write $n = n_1 + n_2$ for the number of vertices, and $m = n_1 n_2$ for the number of edges. Let H denote a random graph generated by including each edge of G independently with probability p. with high probability, we can re-weight edges of H so that the (weighted) Laplacian matrix L_H is ϵ-spectrally similar with L_G, where $\epsilon = O\left(\sqrt{\frac{n\log n}{pm}}\right)$.*

Proof For complete bipartite graphs, all edges have the same effective resistance, so uniform sampling among all the edges will produce a good spectral sparsifier.

Formally, we can use the main result of Spielman and Srivastava (2011): Fix any $0 < \epsilon < 1$. For sufficiently large n and all graphs on n vertices, there is a universal constant C so that sampling $Cn\log n/\epsilon^2$ edges independently (with sample probability p_e proportional to [edge weight × effective resistant]) produces an ϵ-spectral sparsifier with high probability. The lemma allows reweighting on H because when we include an edge we give it weight $1/p_e$.

At the core of (Spielman and Srivastava, 2011) are matrix concentration inequalities (Rudelson and Vershynin, 2007; Ahlswede and Winter, 2002; Tropp, 2012). Note that the original proof in (Spielman and Srivastava, 2011) used sampling with replacement and holds only with constant probability, but the analysis can be adapted to show that sampling by effective resistance without replacement works with high probability. ■

Lemma 11 *Let $L = D - A$ and $\widetilde{L} = \widetilde{D} - \widetilde{A}$ be two graph Laplacians, where D is the degree matrix and A is the adjacency matrix of the graph. If $(1-\epsilon)L \preceq \widetilde{L} \preceq L$, then we have:*

(1) $(1-\epsilon)D_{i,i} \le \widetilde{D}_{i,i} \le D_{i,i}$.

(2) $\|D^{-1/2}(\widetilde{A} - A)D^{-1/2}\| \le 3\epsilon$.

Proof For (1), the spectral similarity between L and $\widetilde{L}$ implies that $(1-\epsilon)x^\top Lx \le x^\top \widetilde{L}x \le x^\top Lx$ for all $x \in \mathbb{R}^n$. In particular, this holds for all standard basis vectors, so $(1-\epsilon)D_{i,i} \le \widetilde{D}_{i,i} \le D_{i,i}$.

For (2), we know that $0 \preceq L - \widetilde{L} \preceq \epsilon L$ and similarly $0 \preceq D - \widetilde{D} \preceq \epsilon D$, and therefore

$$\begin{aligned}\|D^{-1/2}(\widetilde{A} - A)D^{-1/2}\| &= \|D^{-1/2}(\widetilde{D} - D + L - \widetilde{L})D^{-1/2}\| \\ &\le \|D^{-1/2}(D - \widetilde{D})D^{-1/2}\| + \|D^{-1/2}(L - \widetilde{L})D^{-1/2}\| \\ &\le \epsilon\|I\| + \epsilon\|D^{-1/2}LD^{-1/2}\| \le 3\epsilon.\end{aligned}$$

The last step uses the fact that eigenvalues of a normalized Laplacian matrix $D^{-1/2}LD^{-1/2}$ are always between 0 and 2. ■

Appendix D. Using Deterministic Conditions for Matrix Completion

In this section, we prove Theorem 15.

We use Lemma 14 and the techniques in (Ge et al., 2017) to show that all local minima of the non-convex objective functions are close to the ground truth. We first restate the objective functions: Equation (3) for the symmetric case and Equation (1) for the asymmetric case.

$$\min\ f(U) = \frac{1}{2}\|UU^\top - M^\star\|_W^2 + Q(U), \tag{3}$$

$$\min\ f(U,V) = 2\|UV^\top - M^\star\|_W^2 + \frac{1}{2}\|U^\top U - V^\top V\|_F^2 + Q(U,V), \tag{1}$$

where $x_+ = \max\{x, 0\}$, $Q(U) = \lambda \sum_{i=1}^n (\|U_i\|_2 - \alpha)_+^4$, and $Q(U,V) = \lambda_1 \sum_{i=1}^{n_1} (\|U_i\|_2 - \alpha_1)_+^4 + \lambda_2 \sum_{i=1}^{n_2} (\|V_i\|_2 - \alpha_2)_+^4$.

We start with an overview of the analysis in (Ge et al., 2017) in Appendix D.1. Because Lemma 12 is no longer true in the semi-random setting, we cannot use the proof of Ge et al. (2017) in a black-box way. We will handle symmetric (Appendix D.2) and asymmetric (Appendix D.3) cases separately.

D.1. Overview of the Analysis in (Ge et al., 2017)

We give a brief overview of the techniques in (Ge et al., 2017). The materials in this section are independent of the concentration bounds, so they remain valid in the semi-random model.

Measuring Distance between Matrices. The first problem in analyzing Objective (3) is that the optimal solution is not unique: given a matrix $M^\star = U^\star(U^\star)^\top$, for any orthonormal matrix R we also have $M^\star = (U^\star R)(U^\star R)^\top$. To take this symmetry into account, we define the distance between two matrices as follows:

Definition 23 *Given matrices $U, U^\star \in \mathbb{R}^{n\times r}$, their difference is defined to be $\Delta = U - U^\star R$, where $R \in \mathbb{R}^{r\times r}$ is an $r \times r$ orthonormal matrix that minimizes $\|U - U^\star R\|_F^2$.*

The benefit of this definition of distance is summarized in the following lemma:

Lemma 24 (Lemma 6 in (Ge et al., 2017)) *Given matrices $U, U^\star \in \mathbb{R}^{n\times r}$, let $M = UU^\top$ and $M^\star = U^\star(U^\star)^\top$, let Δ be the difference defined in Definition 23, then*

$$\|\Delta\Delta^\top\|_F^2 \le 2\|M - M^\star\|_F^2,$$

and

$$\sigma_r^\star\|\Delta\|_F^2 \le \frac{1}{2(\sqrt{2}-1)}\|M - M^\star\|_F^2.$$

The lemma states that when Δ is large, M is also far from $M^\star$. This would not be true if we simply defined $\Delta = U - U^\star$ without considering the best rotation of $U^\star$. From now on, we will always assume $U^\star$ is *aligned* with U in the sense that $R = I$ and $\Delta = U - U^\star$ (this can be guaranteed by choosing the appropriate global optimum that U is comparing to).

Main Proof for the Symmetric Case. First, we introduce notations for the Hessian. The Hessian of $f(U)$ is a 4-th order tensor (because the variable U is a matrix). For an $n \times r$ matrix X, we use $[\nabla^2 f(U)](X)$ to denote the quadratic form of the Hessian evaluated at X. The Hessian is positive semidefinite (PSD), iff $[\nabla^2 f(U)](X) \ge 0$ for every X.

The main idea of Ge et al. (2017) is to focus on the direction of Δ: To prove $UU^\top = M^\star$, instead of using $\nabla f(U) = 0$ and $\nabla^2 f(U)$ is PSD, it is sufficient to work with $\langle \nabla f(U), \Delta\rangle = 0$ and $[\nabla^2 f(U)](\Delta) \ge 0$. The next lemma, which is the main lemma in (Ge et al., 2017), derives a particular inequality that is very useful in proving convergence. Lemma 25 is proved by simplifying the second-order term $[\nabla^2 f(U)](\Delta)$ given that the first-order term $\langle \nabla f(U), \Delta\rangle$ is 0.

Lemma 25 (Lemma 7 in (Ge et al., 2017)) *Let $M = UU^\top$ and Δ is the difference of U and $U^\star$ as in Definition 23, if U is a local minimum of Objective (3), then*

$$0 \le [\nabla^2 f(U)](\Delta) = \|\Delta\Delta^\top\|_W^2 - 3\|M - M^\star\|_W^2 + ([\nabla^2 Q(U)](\Delta) - 4\langle \nabla Q(U), \Delta\rangle).$$

To see why this inequality is useful intuitively, assume the regularizer term is 0 (the current vector is incoherent so the incoherence regularizer is not active), and assume further the W-norms are very close to Frobenius norm (which is essentially guaranteed by Lemmas 12 and 13 when entries are observed randomly), then we have

$$\|\Delta\Delta^\top\|_F^2 - 3\|M - M^\star\|_F^2 \ge 0.$$

However, by Lemma 24 we know $\|\Delta\Delta^\top\|_F^2 \le 2\|M - M^\star\|_F^2$, so the only way this equation can hold is if $\|M - M^\star\|_F = 0$, and therefore, all local optima are global.

Finally, we state the lemma that shows the regularizer term is indeed small.[9]

[9] The constant in Lemma 26 is slightly different from that of (Ge et al., 2017), but it follows from the same proof by choosing a larger universal constant C.

Lemma 26 (Lemma 11 in (Ge et al., 2017)) *Let U and Δ be defined as above. Choose $\alpha^2 = \frac{C\mu r\sigma_1^\star}{n}$ and $\lambda = \frac{C^2 n}{\mu r\kappa^\star}$ where C is a large enough universal constant, then we have*

$$([\nabla^2 Q(U)](\Delta) - 4\langle \nabla Q(U), \Delta\rangle) \le 0.1\sigma_r^\star \|\Delta\|_F^2.$$

Reduction from Asymmetric Case to the Symmetric Case. To handle asymmetric matrices, (Ge et al., 2017) gives a way to essentially reduce asymmetric matrices to symmetric matrices.

For variables U, V and optimal solution $U^\star, V^\star$, we define the following matrices:

$$Z = \begin{pmatrix} U \\ V \end{pmatrix}; \; Z^\star = \begin{pmatrix} U^\star \\ V^\star \end{pmatrix}; \quad N = ZZ^\top; \; N^\star = (Z^\star)(Z^\star)^\top.$$

In the asymmetric setting, we consider $\Delta = \begin{pmatrix} \Delta_U \\ \Delta_V \end{pmatrix}$ as the difference between Z and $Z^\star$ as in Definition 23, and we also rotate $Z^\star$ so that $\Delta = Z - Z^\star$.

Roughly speaking, we want to design an objective function that reduces the asymmetric case to a symmetric matrix completion problem with variables Z and ground truth $N^\star$. This is impossible if we only focus on the term $2\|UV^\top - M^\star\|_W^2$, because it does not depend on the diagonal blocks of $(N - N^\star)$. Since we cannot observe the diagonal blocks of $N^\star$, we try to add a term so that the Hessian of $f(Z)$ acts like a block identity tensor on N. The additional term $\frac{1}{2}\|U^\top U - V^\top V\|_F^2$ is introduced for exactly this purpose.

Let $Q(Z) = Q(U, V)$ be the same regularizer as in Objective (1). Ge et al. (2017) proved the following lemma:

Lemma 27 (Essentially Lemma 16 in (Ge et al., 2017)) *Let Z, $Z^\star$, N, $N^\star$, and Δ be defined as above, if Z is a local minimum of Objective (1), then*

$$0 \le [\nabla^2 f(Z)](\Delta) \le \|\Delta\Delta^\top\|_{\overline{W}}^2 - 3\|N - N^\star\|_{\overline{W}}^2 + ([\nabla^2 Q(Z)](\Delta) - 4\langle \nabla Q(Z), \Delta\rangle).$$

where

$$\overline{W} = \begin{pmatrix} J & 2W - J \\ 2W^\top - J & J \end{pmatrix}.$$

Similar to the symmetric case, Lemma 27 is proved by simplifying the second-order term $[\nabla^2 f(Z)](\Delta)$ given that the first-order term $\langle \nabla f(Z), \Delta\rangle$ is 0.

Notice that we have $\|\overline{W} - J\| = 2\|W - J\|$. If our preprocessing algorithm guarantees that W is close to J, then $\overline{W}$ is close to J as well.

Finally, we have a corresponding lemma that shows the regularization term is small.

Lemma 28 (Lemma 22 in (Ge et al., 2017)) *Let Z and Δ be defined as above. Choose $\alpha_1^2 = \frac{C\mu r\sigma_1^\star}{n_1}, \alpha_2^2 = \frac{C\mu r\sigma_1^\star}{n_2}$ and $\lambda_1 = \frac{C^2 n_1}{\mu r\kappa^\star}, \lambda_2 = \frac{C^2 n_1}{\mu r\kappa^\star}$ where C is a large enough universal constant, then we have*

$$([\nabla^2 Q(Z)](\Delta) - 4\langle \nabla Q(Z), \Delta\rangle) \le 0.1\sigma_r^\star \|\Delta\|_F^2.$$

D.2. Proof of Our Symmetric Case

We first prove a variant of Lemma 9 in (Ge et al., 2017) in the semi-random model. Lemma 29 shows that any local minima of Objective (3) have bounded row norms.

Lemma 29 *When the weight matrix W satisfies $\|W\|_\infty \le n$, choose $\alpha^2 = \frac{C\mu r\sigma_1^\star}{n}$ and $\lambda = \frac{C^2 n}{\mu r \kappa^\star}$ where C is a large enough universal constant. For Objective (3), we have for any matrix U with $\nabla f(U) = 0$,*

$$\max_i \|U_i\|_2^2 = O\left(\frac{\mu^2 r^2 \kappa^\star \sigma_1^\star}{n}\right).$$

Proof Recall that $U_i \in \mathbb{R}^{1\times r}$ is the i-th row of $U \in \mathbb{R}^{n\times r}$ and $e_i \in \mathbb{R}^{r\times 1}$ is the i-th standard basis vector.

The gradient $\nabla f(U)$ is equal to $2(W * (M - M^\star))U + \nabla Q(U)$, where

$$\nabla Q(U) = 4\lambda \sum_{i=1}^n (\|U_i\|_2 - \alpha)_+^3 \frac{e_i U_i}{\|U_i\|_2^2}.$$

Let $i^\star$ be the row index with the maximum ℓ_2-norm, if $\|U_{i^\star}\|_2 \le 2\alpha$ then we are done. On the other hand, if $\|U_{i^\star}\|_2 > 2\alpha$, we will consider the gradient along $e_{i^\star}U_{i^\star}$. We have

$$\begin{aligned}
0 &= \langle \nabla f(U), e_{i^\star}U_{i^\star}\rangle \\
&= \langle e_{i^\star}^\top [2(W * (UU^\top - M^\star))U + \nabla Q(U)], U_{i^\star}\rangle \\
&\ge 4\lambda(\|U_{i^\star}\|_2 - \alpha)_+^3 \|U_{i^\star}\|_2 - 2\langle e_{i^\star}^\top M^\star, e_{i^\star}^\top UU^\top\rangle_W \\
&\ge \frac{\lambda}{2}\|U_{i^\star}\|_2^4 - 2n\|M^\star\|_{\max}\|UU^\top\|_{\max} \\
&\ge \frac{\lambda}{2}\|U_{i^\star}\|_2^4 - 2\mu r\sigma_1^\star \|U_{i^\star}\|_2^2.
\end{aligned}$$

The third step removes the term $\langle e_{i^\star}^\top UU^\top U, U_{i^\star}\rangle = \|e_{i^\star}^\top UU^\top\|_F^2 \ge 0$. The fourth step uses that $\|U_{i^\star}\|_2 > 2\alpha$ and $\|W\|_\infty \le n$ (every row of W has ℓ_1-norm at most n). The last step is due to $\|UU^\top\|_{\max} = \|U_{i^\star}\|_2^2$; and $\|M^\star\|_{\max} = \max_{i,j}\langle U_i, V_j\rangle \le \max_{i,j}\|U_i\|_2\|V_j\|_2 \le \frac{\sigma_1^\star \mu r}{n}$ because $M^\star$ is incoherent. As a result, we know that $\|U_{i^\star}\|_2^2 \le \frac{4\mu r\sigma_1^\star}{\lambda} = O\left(\frac{\mu^2 r^2\kappa^\star\sigma_1^\star}{n}\right)$ by our choice of λ.
■

Next, we will show that all local minima are close to the ground truth.

Lemma 30 *Fix any error parameter $0 < \epsilon < 1$. For a weight matrix W such that $\|W\|_\infty \le n$ and $\|W - J\| \le \frac{\epsilon c n}{\mu^2 r^2 (\kappa^\star)^2}$ for a small enough universal constant c, any local minimum U of Objective (3) satisfies $\|UU^\top - M^\star\|_F^2 \le \epsilon \|M^\star\|_F^2$.*

Proof By Lemma 25, we know that every local minimum of $f(U)$ satisfies

$$\|\Delta\Delta^\top\|_W^2 - 3\|UU^\top - M^\star\|_W^2 + \left([\nabla^2 Q(U)](\Delta) - 4\langle \nabla Q(U), \Delta\rangle\right) \ge 0.$$

We will bound these three terms. First, by Lemma 29, we know $\|\Delta_i\|_2^2 \le O\left(\frac{\mu^2 r^2 \kappa^\star \sigma_1^\star}{n}\right)$.

For the first term $\|\Delta\Delta^\top\|_W^2$, we can directly apply Lemma 14:

$$\begin{aligned}\|\Delta\Delta^\top\|_W^2 &\le \|\Delta\Delta^\top\|_F^2 + \|W-J\|\|\Delta\|_F^2 \max_i \|\Delta_i\|^2 \\ &\le \|\Delta\Delta^\top\|_F^2 + \|W-J\| \cdot O\left(\frac{\mu^2 r^2 \kappa^\star \sigma_1^\star}{n}\right) \cdot \|\Delta\|_F^2 \\ &\le \|\Delta\Delta^\top\|_F^2 + 0.1\sigma_r^\star \|\Delta\|_F^2.\end{aligned}$$

The last inequality uses the fact that $\|W-J\| \le \frac{cn}{\mu^2 r^2 (\kappa^\star)^2}$ for a small enough constant c.

For the second term $\|UU^\top - M^\star\|_W^2$, we invoke Lemma 14 with $X = (U, U^\star)$ and $Y = (U, -U^\star)$. Notice that $XY^\top = UU^\top - M^\star$. Moreover, we know that $\|X\|_F \le \|U\|_F + \|U^\star\|_F \le 2\|U^\star\|_F + \|\Delta\|_F$. Similarly, the row norms of X is still upper bounded by $O\left(\frac{\mu^2 r^2 \kappa^\star \sigma_1^\star}{n}\right)$. Therefore,

$$\begin{aligned}\|UU^\top - M^\star\|_W^2 &\ge \|UU^\top - M^\star\|_F^2 - \|W-J\|\|(U,U^\star)\|_F^2 \max_i \|(U,U^\star)_i\|_2^2 \\ &\ge \|UU^\top - M^\star\|_F^2 - \|W-J\| \cdot O\left(\frac{\mu^2 r^2 \kappa^\star \sigma_1^\star}{n}\right) \cdot (2\|U^\star\|_F + \|\Delta\|_F)^2 \\ &\ge \|UU^\top - M^\star\|_F^2 - 0.1\epsilon\sigma_r^\star \|U^\star\|_F^2 - 0.1\sigma_r^\star\|\Delta\|_F^2.\end{aligned}$$

Again, the last step uses the fact that $\|W-J\| \le \frac{\epsilon cn}{\mu^2 r^2 (\kappa^\star)^2}$ for a small enough constant c.

Finally, the third term is bounded by $0.1\sigma_r^\star\|\Delta\|_F^2$ by Lemma 26. Combining all these terms,

$$\begin{aligned}0 &\le \|\Delta\Delta^\top\|_W^2 - 3\|UU^\top - M^\star\|_W^2 + \left([\nabla^2 Q(U)](\Delta) - 4\langle \nabla Q(U), \Delta\rangle\right) \\ &\le \|\Delta\Delta^\top\|_F^2 + 0.2\sigma_r^\star\|\Delta\|_F^2 - 3\left(\|UU^\top - M^\star\|_F^2 - 0.1\epsilon\sigma_r^\star\|U^\star\|_F^2 - 0.1\sigma_r^\star\|\Delta\|_F^2\right) \\ &\le -\|UU^\top - M^\star\|_F^2 + 0.5\sigma_r^\star\|\Delta\|_F^2 + 0.3\epsilon\sigma_r^\star\|U^\star\|_F^2 \\ &\le -0.3\|UU^\top - M^\star\|_F^2 + 0.3\epsilon\sigma_r^\star\|U^\star\|_F^2.\end{aligned}$$

Here the calculations use the inequalities in Lemma 24.

As a result, $\|UU^\top - M^\star\|_F^2 \le \epsilon\sigma_r^\star\|U^\star\|_F^2$. We conclude the proof by noting that $\sigma_r^\star\|U^\star\|_F^2 = \sigma_r^\star \sum_{i=1}^r \sigma_i^\star \le \sum_{i=1}^r (\sigma_i^\star)^2 = \|M^\star\|_F^2$, because the singular values of $U^\star$ are $\sqrt{\sigma_i^\star}$'s. ■

D.3. Our Asymmetric Case: Proof of Theorem 15

The proof for the asymmetric case (Objective (1)) is very similar. Recall that $M^\star = U^\star(V^\star)^\top$, where $U^\star \in \mathbb{R}^{n_1\times r}$ and $V^\star = \mathbb{R}^{n_2 \times r}$. We again start by bounding the row norms of U and V.

Lemma 31 *Suppose* $\|W\|_\infty \le n_2$ *and* $\|W\|_1 \le n_1$. *Choose* $\alpha_1^2 = \frac{C\mu r\sigma_1^\star}{n_1}, \alpha_2^2 = \frac{C\mu r \sigma_1^\star}{n_2}$ *and* $\lambda_1 = \frac{C^2 n_1}{\mu r \kappa^\star}, \lambda_2 = \frac{C^2 n_1}{\mu r\kappa^\star}$ *where C is a large enough universal constant. For f as in Objective (1) and any matrix* $Z = \begin{pmatrix} U \\ V\end{pmatrix}$ *with* $\nabla f(Z) = 0$, *we have*

$$\max_i \|U_i\|_2^2 = O\left(\frac{\mu^3 r^3 (\kappa^\star)^2 \sigma_1^\star}{n_1}\right); \quad \max_i \|V_i\|_2^2 = O\left(\frac{\mu^3 r^3 (\kappa^\star)^2 \sigma_1^\star}{n_2}\right).$$

Proof Without loss of generality, we assume $\sqrt{n_1}\max_i\|U_i\|_2 \ge \sqrt{n_2}\max_i\|V_i\|_2$, so it is enough to upper bound $\max_i\|U_i\|_2$. The gradient can be computed as follows.

$$\nabla f(Z) = 4\begin{pmatrix}[W*(M-M^\star)]V\\ [W*(M-M^\star)]^\top U\end{pmatrix} + 2\begin{pmatrix}U(U^\top U - V^\top V)\\ V(V^\top V - U^\top U)\end{pmatrix} + \nabla Q(Z),$$

where

$$\nabla Q(Z) = 4\lambda_1\sum_{i=1}^{n_1}(\|Z_i\|_2-\alpha_1)_+^3\frac{e_iZ_i}{\|Z_i\|_2^2} + 4\lambda_2\sum_{i=n_1+1}^{n_1+n_2}(\|Z_i\|_2-\alpha_2)_+^3\frac{e_iZ_i}{\|Z_i\|_2^2}.$$

First, we observe that $\langle\nabla Q(Z), Z\rangle \ge 0$, therefore,

$$\begin{aligned}0 &= \langle\nabla f(Z), Z\rangle\\ &= 2\|U^\top U - V^\top V\|_F^2 + 8\langle M-M^\star, M\rangle_W + \langle\nabla Q(Z), Z\rangle\\ &\ge 2\|U^\top U - V^\top V\|_F^2 - 8\langle M^\star, M\rangle_W\\ &\ge 2\|U^\top U - V^\top V\|_F^2 - 8n_1n_2\|M^\star\|_{\max}\|M\|_{\max}.\\ &\ge 2\|U^\top U - V^\top V\|_F^2 - 8\sqrt{n_1n_2}\mu r\sigma_1^\star\|M\|_{\max}.\end{aligned}$$

Let $i^\star = \arg\max_i\|U_i\|_2$, if $\|U_{i^\star}\|_2 \le 2\alpha_1$ then we are done. On the other hand, if $\|U_{i^\star}\|_2 > 2\alpha_1$, we know $\|M\|_{\max} \le \max_{i,j}\|U_i\|_2\|V_j\|_2 \le \sqrt{n_1/n_2}\|U_{i^\star}\|_2^2$. As a result, we know

$$\|U^\top U - V^\top V\|_F^2 \le 4\sqrt{n_1n_2}\mu r\sigma_1^\star\|M\|_\infty \le 4n_1\mu r\sigma_1^\star\|U_{i^\star}\|_2^2.$$

Let $Q(Z) = Q(U,V) = Q_1(U) + Q_2(V)$. Consider the gradient of $f(Z)$ along the direction $e_{i^\star}Z_{i^\star}$, we have

$$\begin{aligned}0 &= \langle\nabla f(Z), e_{i^\star}Z_{i^\star}\rangle = \langle e_{i^\star}^\top\nabla f(Z), Z_{i^\star}\rangle\\ &= \langle e_{i^\star}^\top\left[4(W*(M-M^\star))V + 2U(U^\top U - V^\top V) + \nabla Q_1(U)\right], U_{i^\star}\rangle\\ &\ge 4\lambda_1(\|U_{i^\star}\|_2-\alpha_1)_+^3\|U_{i^\star}\|_2 - 4\langle e_{i^\star}^\top M^\star, e_{i^\star}^\top M\rangle_W - 2\|U^\top U - V^\top V\|_F\|U_{i^\star}\|_2^2\\ &\ge \frac{\lambda_1}{2}\|U_{i^\star}\|_2^4 - 4n_2\|M^\star\|_{\max}\|M\|_{\max} - 2\|U^\top U - V^\top V\|_F\|U_{i^\star}\|_2^2.\\ &\ge \frac{\lambda_1}{2}\|U_{i^\star}\|_2^4 - 4\mu r\sigma_1^\star\|U_{i^\star}\|_2^2 - 4\sqrt{n_1\mu r\sigma_1^\star}\|U_{i^\star}\|_2^3.\end{aligned}$$

This implies

$$\frac{\lambda_1}{2}\|U_{i^\star}\|_2^2 \le 4\mu r\sigma_1^\star + 4\sqrt{n_1\mu r\sigma_1^\star}\|U_{i^\star}\|_2.$$

Therefore, $\|U_{i^\star}\|_2^2 = O\big(\max\{\frac{\mu r\sigma_1^\star}{\lambda_1}, \frac{n_1\mu r\sigma_1^\star}{\lambda_1^2}\}\big) = O\left(\frac{\mu^3r^3(\kappa^\star)^2\sigma_1^\star}{n_1}\right)$ by our choices of α's and λ's. ■

Next, we will prove that all local minima are close to the ground truth.

Lemma 32 *Fix any error parameter $0 < \epsilon < 1$. Suppose the weight matrix W satisfies that $\|W\|_\infty \le n_2$, $\|W\|_1 \le n_1$, and $\|W - J\| \le \frac{\epsilon c\sqrt{n_1n_2}}{\mu^3r^3(\kappa^\star)^3}$ for a small enough universal constant c. Then, any local minimum (U, V) of Objective (1) has $\|UV^\top - M^\star\|_F^2 \le \epsilon\|M^\star\|_F^2$.*

Proof By Lemma 27 we know for every local minimum of $f(U,V)$ satisfies

$$\|\Delta\Delta^\top\|_{\overline{W}}^2 - 3\|N - N^\star\|_{\overline{W}}^2 + \left([\nabla^2 Q(Z)](\Delta) - 4\langle \nabla Q(Z), \Delta\rangle\right) \geq 0.$$

We will bound these three terms. First, by Lemma 31 we know the rows of Δ_U have squared ℓ_2-norm at most $O\left(\frac{\mu^3 r^3 (\kappa^\star)^2 \sigma_1^\star}{n_1}\right)$, and the rows of Δ_V have squared ℓ_2-norm at most $O\left(\frac{\mu^3 r^3 (\kappa^\star)^2 \sigma_1^\star}{n_2}\right)$.

For the first term $\|\Delta\Delta^\top\|_{\overline{W}}^2$, by the definition of $\overline{W}$,

$$\begin{aligned}\|\Delta\Delta^\top\|_{\overline{W}}^2 &= \|\Delta_U\Delta_U^\top\|_F^2 + \|\Delta_V\Delta_V^\top\|_F^2 - 2\|\Delta_U\Delta_V^\top\|_F^2 + 4\|\Delta_U\Delta_V^\top\|_W^2 \\ &= \|\Delta\Delta^\top\|_F^2 + 4(\|\Delta_U\Delta_V^\top\|_W^2 - \|\Delta_U\Delta_V^\top\|_F^2).\end{aligned}$$

We can directly apply Lemma 14 to $\|\Delta_U\Delta_V^\top\|_W^2$.

$$\begin{aligned}\|\Delta_U\Delta_V^\top\|_W^2 &\leq \|\Delta_U\Delta_V^\top\|_F^2 + \|W - J\| \cdot \|\Delta_U\|_F \cdot \|\Delta_V\|_F \cdot \max_{i=1}^{n_1} \|\Delta_i\|_2 \cdot \max_{j=n_1+1}^{n_1+n_2} \|\Delta_j\|_2 \\ &\leq \|\Delta_U\Delta_V^\top\|_F^2 + \|W - J\| \cdot O\left(\frac{\mu^3 r^3 (\kappa^\star)^2 \sigma_1^\star}{\sqrt{n_1 n_2}}\right) \cdot \|\Delta\|_F^2 \\ &\leq \|\Delta_U\Delta_V^\top\|_F^2 + 0.01\sigma_r^\star \|\Delta\|_F^2.\end{aligned}$$

Here the last inequality uses the fact that $\|W - J\| \leq \frac{c\sqrt{n_1 n_2}}{\mu^3 r^3 (\kappa^\star)^3}$ for a small enough constant c.

For the second term, we can relate $\overline{W}$-norm to W-norm similarly, which allows us to focus on the W-norm of the off-diagonal blocks, $\|UV^\top - M^\star\|_W^2$. We then invoke Lemma 14 with $X = (U, U^\star)$ and $Y = (V, -U^\star)$. We know $XY^\top = UV^\top - M^\star$ and $\|X\|_F \leq \|U\|_F + \|U^\star\|_F \leq 2\|U^\star\|_F + \|\Delta\|_F$ (the same upper bound holds for $\|Y\|_F$ because $\|U^\star\|_F = \|V^\star\|_F$). The row norms of X is still bounded by $O\left(\frac{\mu^2 r^2 \kappa^\star \sigma_1^\star}{n_1}\right)$ (and similarly for Y except the denominator is n_2).

$$\begin{aligned}&\|UV^\top - M^\star\|_W^2 \\ &\geq \|UV^\top - M^\star\|_F^2 - \|W - J\| \cdot \|(U, U^\star)\|_F \|(V, -V^\star)\|_F \cdot \max_i \|(U, U^\star)_i\|_2 \cdot \max_j \|(V, V^\star)_j\|_2 \\ &\geq \|UV^\top - M^\star\|_F^2 - \|W - J\| \cdot O\left(\frac{\mu^3 r^3 (\kappa^\star)^2 \sigma_1^\star}{\sqrt{n_1 n_2}}\right) \cdot (2\|U^\star\|_F + \|\Delta\|_F)^2 \\ &\geq \|UV^\top - M^\star\|_F^2 - 0.05\epsilon\sigma_r^\star \|U^\star\|_F^2 - 0.01\sigma_r^\star \|\Delta\|_F^2.\end{aligned}$$

Again, the last step uses the fact that $\|W - J\| \leq \frac{\epsilon c n}{\mu^3 r^3 (\kappa^\star)^3}$ for a small enough constant c.

Finally, the third term is bounded by $0.1\sigma_r^\star \|\Delta\|_F^2$ by Lemma 28. We combine all these terms and apply Lemma 24,

$$\begin{aligned}0 &\leq \|\Delta\Delta^\top\|_{\overline{W}}^2 - 3\|N - N^\star\|_{\overline{W}}^2 + \left([\nabla^2 Q(Z)](\Delta) - 4\langle \nabla Q(Z), \Delta\rangle\right) \\ &\leq \|\Delta\Delta^\top\|_F^2 + 0.14\sigma_r^\star \|\Delta\|_F^2 - 3\left(\|N - N^\star\|_F^2 - 0.2\epsilon\sigma_r^\star \|U^\star\|_F^2 - 0.01\sigma_r^\star \|\Delta\|_F^2\right) \\ &\leq -\|N - N^\star\|_F^2 + 0.17\sigma_r^\star \|\Delta\|_F^2 + 0.6\epsilon\sigma_r^\star \|U^\star\|_F^2 \\ &\leq -0.6\|N - N^\star\|_F^2 + 0.6\epsilon\sigma_r^\star \|U^\star\|_F^2.\end{aligned}$$

As a result, $\|M - M^\star\|_F^2 \leq \|N - N^\star\|_F^2 \leq \epsilon\sigma_r^\star \|U^\star\|_F^2 \leq \epsilon\|M^\star\|_F^2$. ■