Low-rank Matrix Recovery from Local Coherence Perspective [1] [2]

Huishuai Zhang, Yi Zhou, Yingbin Liang [3]

Abstract

We investigate the robust PCA problem of decomposing an observed matrix into the sum of a low-rank and a sparse error matrices via convex programming Principal Component Pursuit (PCP). In contrast to previous studies that assume the support of the sparse error matrix is generated by uniform Bernoulli sampling, we allow non-uniform sampling, i.e., entries of the low-rank matrix are corrupted by errors with unequal probabilities. We characterize conditions on error corruption of each individual entry based on the local coherence of the low-rank matrix, under which correct matrix decomposition by PCP is guaranteed. Such a refined analysis of robust PCA captures how robust each entry of the low-rank matrix combats error corruption. Moreover, this result has several immediate implications on graph clustering problem, which have been partially studied in random graph clustering literatures. In order to deal with non-uniform error corruption, our technical proof introduces a new weighted norm and develops/exploits the concentration properties that such a norm satisfies. We also investigate the partial observation setting and establish the general theory of matrix recovery from both error corruption and partial observation based on local coherence.

1 Introduction

We consider the problem of robust Principal Component Analysis (PCA). Suppose a n-by-n[4] data matrix M can be decomposed into a low-rank matrix L and a sparse matrix S as

$$M = L + S. \tag{1}$$

Robust PCA aims to find L and S with M given. This problem has been extensively studied recently. In [1,2], *Principal Component Pursuit (PCP)* has been proposed to solve the robust PCA problem via the following convex programming

$$\text{PCP:} \quad \begin{aligned} &\underset{L,S}{\text{minimize}} \quad \|L\|_* + \lambda \|S\|_1 \\ &\text{subject to} \quad M = L + S, \end{aligned} \tag{2}$$

[1]The material in this paper was presented in part at the 29th Neural Information Processing Systems (NIPS), Montreal, Quebec, Canada, December 2015.

[2]The work of H. Zhang, Y. Zhou and Y. Liang was supported by a National Science Foundation CAREER Award under Grant CCF-10-26565 and by the National Science Foundation under Grant CNS-11-16932.

[3]H. Zhang, Y. Zhou and Y. Liang are with the Department of Electrical Engineering and Computer Science, Syracuse University, Syracuse, NY 13244 USA (email: {hzhan23, yzhou35, yliang06}@syr.edu).

[4]In this paper, we focus on square matrices for simplicity. Our results can be extended to rectangular matrices in a standard way.

where $\|\cdot\|_*$ denotes the nuclear norm, i.e., the sum of singular values, and $\|\cdot\|_1$ denotes the l_1 norm i.e., the sum of absolute values of all entries. It was shown in [1,2] that PCP successfully recovers L and S if the two matrices are distinguishable from each other in properties, i.e., L is not sparse and S is not low-rank. One important quantity that determines similarity of L to a sparse matrix is the *coherence* of L, which measures how column and row spaces of L are aligned with canonical basis and between themselves. Namely, suppose that L is a rank-r matrix with SVD $L = U\Sigma V^*$, where Σ is a $r \times r$ diagonal matrix with singular values as its diagonal entries, U is a $n \times r$ matrix with columns as the left singular vectors of L, V is a $n \times r$ matrix with columns as the right singular vectors of L, and V^* denotes the transpose of V. The *coherence* of L is measured by $\mu = \max\{\mu_0, \mu_1\}$, where μ_0 and μ_1 are defined as

$$\|U^* e_i\| \leq \sqrt{\frac{\mu_0 r}{n}}, \qquad \|V^* e_j\| \leq \sqrt{\frac{\mu_0 r}{n}}, \quad \text{for all } i, j = 1, \cdots, n \tag{3}$$

$$\|UV^*\|_\infty \leq \sqrt{\frac{\mu_1 r}{n^2}}. \tag{4}$$

Previous studies suggest that the coherent level of L and the sparse level of S jointly determine whether PCP could successfully recover L and S. For example, Theorem 2 in [3] explicitly shows that the matrix L with larger μ can tolerate only smaller error density to guarantee correct matrix decomposition by PCP. In all previous works on robust PCA, the coherence is defined to be the maximum over all column and row spaces of L as in (3) and (4), which can be viewed as the *global* parameter for the entire matrix L, and consequently, characterization of error density is based on such global (and in fact *the worst case*) coherence.

In fact, each (i, j) entry of the low-rank matrix L can be associated with a *local* coherence parameter μ_{ij}, which is less than or equal to the *global* parameter μ, and then the allowable entry-wise error density can be potentially higher than that characterized based on the global coherence. Thus, the total number of errors that the matrix can tolerate in robust PCA can be much higher than that characterized based on the global coherence when errors are distributed accordingly. Motivated by such an observation, this paper aims to characterize conditions on error corruption of each entry of the low-rank matrix based on the corresponding local coherence parameter, which guarantee success of PCP. Such conditions imply how robust each individual entry of L to resist error corruption. Naturally, the error corruption probability is allowed to be *non-uniform* over the matrix (i.e., locations of non-zero entries in S are sampled non-uniformly).

We note that the notion of local coherence was first introduced in [4] for studying the matrix completion problem, in which local coherence determines the local sampling density in order to guarantee correct matrix completion. Here, local coherence plays a similar role, and determines the maximum allowable error density at each entry to guarantee correct matrix decomposition. The difference lies in that local coherence here depends on both localized μ_0 and μ_1 rather than only on localized μ_0 in matrix completion. This difference is intrinsically due to further difficulty of robust PCA, in which locations of error corrupted entries are unknown [1, 3].

Our Contribution. In this paper, we investigate a more general robust PCA problem, in which entries of the low-rank matrix are corrupted by non-uniformly distributed Bernoulli errors. We characterize the conditions that guarantee correct matrix decomposition by PCP. Our results identify that the local coherence (defined by localized μ_0 and μ_1 for each entry of the low-rank matrix) plays a critical role in such characterization. Our results provide the following useful understanding of the robust PCA problem:

- Our characterization provides a localized (and hence more refined) view of robust PCA, and determines how robust each entry of the low-rank matrix combats error corruption.
- Our results suggest that the total number of errors that the low-rank matrix can tolerate depends on how errors are distributed over the matrix.
- Our results interpret several phenomenons in graph clustering: the minimum cluster size and the necessity of μ_1 in characterizing conditions for robust PCA.

In order to deal with non-uniform error corruption, our technical proof introduces a new weighted norm denoted by $l_{w(\infty)}$, which involves the information of both localized μ_0 and μ_1 and is hence different from the weighted norms introduced in [4] for matrix completion. Thus, our proof necessarily involves new technical developments associated with such a new norm.

We also generalize our theory to the partial observation setting. For the scenario with both error corruptions and missing entries, we establish a general condition that guarantees the success of PCP, which recovers the results in previous literatures as special cases.

Related Work. A closely related but different problem from robust PCA is *matrix completion*, in which a low-rank matrix is partially observed and is to be completed. Such a problem has been previously studied in [5–8], and it was shown that a rank-r n-by-n matrix can be provably recoverable by convex optimization with as few as $\Theta(\max\{\mu_0, \mu_1\}nr\log^2 n)$[5] observed entries. Later on, it was shown in [4] that μ_1 does not affect sample complexity for matrix completion and hence $\Theta(\mu_0 nr\log^2 n)$ observed entries are sufficient for guaranteeing correct matrix completion. It was further shown in [9] that a coherent low-rank matrix (i.e., with large μ_0) can be recovered with $\Theta(nr\log^2 n)$ observations as long as the sampling probability is proportional to the leverage score (i.e., localized μ_0). Our problem can be viewed as its counterpart in robust PCA, where the difference lies in the local coherence in our problem depends on both localized μ_0 and μ_1.

Robust PCA aims to decompose an observed matrix into the sum of a low-rank matrix and a sparse matrix. In [2,10], robust PCA with fixed error matrix was studied, and it was shown that the maximum number of errors in any row or column should be bounded from above in order to guarantee correct decomposition by PCP. Robust PCA with random error matrix was investigated in a number of studies. It has been shown in [1] that such decomposition can be exact with high probability if the percentage of corrupted entries is small enough, under the assumptions that the low-rank matrix is incoherent and the support set of the sparse matrix is uniformly distributed. It was further shown in [11] that if signs of nonzero entries in the sparse matrix are randomly chosen, then an adjusted convex optimization can produce

[5] $f(n) \in \Theta(g(n))$ means $k_1 \cdot g(n) \leq f(n) \leq k_2 \cdot g(n)$ for some positive k_1, k_2.

exact decomposition even when the percentage of corrupted entries goes to one (i.e., error is dense). The problem was further studied in [1, 3, 12] for the case with the error-corrupted low-rank matrix only partially observed. Our work provides a more refined (i.e. entry-wise) view of robust PCA with random error matrix, aiming at understanding how local coherence affects susceptibility of each matrix entry to error corruption.

2 Main Result

2.1 Problem Statement

We consider the robust PCA problem introduced in Section 1. Namely, suppose an n-by-n matrix M can be decomposed into two parts: $M = L + S$, where L is a low-rank matrix and S is a sparse (error) matrix. We assume that the rank of L is r, and the support of S is selected randomly but *non-uniformly*. More specifically, let Ω denote the support of S and then $\Omega \subseteq [n] \times [n]$, where $[n]$ denotes the set $\{1, 2, \ldots, n\}$. The event $\{(i, j) \in \Omega\}$ is independent across different pairs (i, j) and

$$\mathbb{P}\left((i, j) \in \Omega\right) = \rho_{ij}, \tag{5}$$

where ρ_{ij} represents the probability that the (i, j)-entry of L is corrupted by error. Hence, Ω is determined by Bernoulli sampling with non-uniform probabilities.

We study both the *random sign* and *fixed sign* models for S. For the fixed sign model, we assume signs of nonzero entries in S are arbitrary and fixed, whereas for the random sign model, we assume that signs of nonzero entries in S are independently distributed Bernoulli variables, randomly taking values $+1$ or -1 with probability $1/2$ as follows:

$$[\mathrm{sgn}(S)]_{ij} = \begin{cases} 1 & \text{with prob.} \quad \rho_{ij}/2 \\ 0 & \text{with prob.} \quad 1 - \rho_{ij} \\ -1 & \text{with prob.} \quad \rho_{ij}/2. \end{cases} \tag{6}$$

In this paper, our goal is to characterize conditions on ρ_{ij} that guarantees correct recovery of L and S with observation of M.

We provide some notations that are used throughout this paper. A matrix X is associated with five norms: $\|X\|_F$ denotes the Frobenius norm, $\|X\|_*$ denotes the nuclear norm (i.e., the sum of singular values), $\|X\|$ denotes the spectral norm (i.e., the largest singular value), and $\|X\|_1$ and $\|X\|_\infty$ represent respectively the l_1 and l_∞ norms of the long vector stacked by X. The inner product between two matrices is defined as $\langle X, Y \rangle := trace(X^* Y)$. For a linear operator $\mathcal{A}$ that acts on the space of matrices, $\|\mathcal{A}\|$ denotes the operator norm given by $\|\mathcal{A}\| = \sup_{\{\|X\|_F = 1\}} \|\mathcal{A}X\|_F$.

2.2 Main Theorems

We adopt the PCP to solve the robust PCA problem. We define the following *local* coherence parameters, which play an important role in our characterization of conditions on entry-wise

ρ_{ij}.

$$\mu_{0ij} := \frac{n}{2r}\left(\|U^* e_i\|^2 + \|V^* e_j\|^2\right), \quad \mu_{1ij} := \frac{n^2([UV^*]_{ij})^2}{r} \tag{7}$$

$$\mu_{ij} := \max\{\mu_{0ij}, \mu_{1ij}\}. \tag{8}$$

It is clear that $\mu_{0ij} \leq \mu_0$ and $\mu_{1ij} \leq \mu_1$ for all $i, j = 1, \cdots, n$. We note that although $\max_{i,j} \mu_{ij} > 1$, some μ_{ij} might take values as small as zero.

We first consider the robust PCA problem under the *random sign model* as introduced in Section 2.1. The following theorem characterizes the condition that guarantees correct recovery by PCP.

Theorem 1. *Consider the robust PCA problem under the random sign model. If*

$$1 - \rho_{ij} \geq \max\left\{C_0\sqrt{\frac{\mu_{ij} r}{n}}\log n, \frac{1}{n^3}\right\}$$

for some sufficiently large constant C_0 and for all $i, j \in [n]$, then PCP yields correct matrix recovery with $\lambda = \frac{1}{32\sqrt{n \log n}}$, with probability at least $1 - cn^{-10}$ for some constant c.

We note that the term $1/n^3$ is introduced to justify dual certificate conditions in the proof (see Appendix A.1). We further note that satisfying the condition in Theorem 1 implies $C_0\sqrt{\mu r/n}\log n \leq 1$, which is an essential bound required in our proof and coincides with the conditions in previous studies [1, 12]. Although we set $\lambda = \frac{1}{32\sqrt{n \log n}}$ for the sake of proof, in practice λ is often determined via cross validation.

The above theorem suggests that the local coherence parameter μ_{ij} is closely related to how robust each entry of L to error corruption in matrix recovery. An entry corresponding to smaller μ_{ij} tolerates larger error density ρ_{ij}. This is consistent with the result in [4] for matrix completion, in which smaller local coherence parameter requires lower local sampling rate. The difference lies in that here both μ_{0ij} and μ_{1ij} play roles in μ_{ij} whereas only μ_{0ij} matters in matrix completion. This is the critical difference between the proofs of robust PCA and matrix completion. The necessity of μ_{1ij} for robust PCA is further demonstrated in Section 2.3 via an example.

Theorem 1 also provides a more refined view for robust PCA in the dense error regime, in which the error corruption probability approaches one. Such an interesting regime was previously studied in [3, 11]. In [11], it is argued that PCP with adaptive λ yields exact recovery even when the error corruption probability approaches one if errors take random signs and the dimension n is sufficiently large. In [3], it is further shown that PCP with a fixed λ also yields exact recovery and the scaling behavior of the error corruption probability is characterized. The above Theorem 1 further provides the scaling behavior of the *local entry-wise* error corruption probability ρ_{ij} as it approaches one, and captures how such scaling behavior depends on local coherence parameters μ_{ij}. Such a result implies that robustness of PCP depends not only on the error density but also on how errors are distributed over the matrix with regard to μ_{ij}.

We next consider the robust PCA problem under the *fixed sign model* as introduced in Section 2.1. In this case, non-zero entries of the error matrix S can take arbitrary and fixed values, and only locations of non-zero entries are random.

Theorem 2. *Consider the robust PCA problem under the fixed sign model. If*

$$(1 - 2\rho_{ij}) \geq \max\left\{C_0\sqrt{\frac{\mu_{ij}r}{n}}\log n, \frac{1}{n^3}\right\}$$

for some sufficient large constant C_0 and for all $i, j \in [n]$, then PCP yields correct recovery with $\lambda = \frac{1}{32\sqrt{n\log n}}$, with probability at least $1 - cn^{-10}$ for some constant c.

Theorem 2 follows from Theorem 1 by adapting the elimination and derandomization arguments [1, Section 2.2] as follows. Let $\boldsymbol{\rho}$ be the matrix with each (i, j)-entry being ρ_{ij}. If PCP yields exact recovery with a certain probability for the random sign model with the parameter $2\boldsymbol{\rho}$, then it also yields exact recovery with at least the same probability for the fixed sign model with locations of non-zero entries sampled using Bernoulli model with the parameter $\boldsymbol{\rho}$. The detailed argument is provided in Appendix B.

We now compare Theorem 2 for robust PCA with *non-uniform* error corruption to Theorem 1.1 in [1] for robust PCA with *uniform* error corruption. It is clear that if we set $\rho_{i,j} = \rho$ for all $i, j \in [n]$, then the two models are the same. It can then be easily checked that conditions $\sqrt{\mu r/n}\log n \leq \rho_r$ and $\rho \leq \rho_s$ in Theorem 1.1 of [1] implies the conditions in Theorem 2. Thus, Theorem 2 provides a more relaxed condition than Theorem 1.1 in [1]. Such benefit of condition relaxation should be attributed to the new golfing scheme introduced in [3, 12], and this paper provides a more refined view of robust PCA by further taking advantage of such a new golfing scheme to analyze local coherence conditions.

More importantly, Theorem 2 characterizes relationship between local coherence parameters and local error corruption probabilities, which implies that different areas of the low-rank matrix have different levels of ability to resist errors: a more coherent area (i.e., with smaller μ_{ij}) can tolerate more errors. Thus, Theorem 2 illustrates the following interesting fact. Whether PCP yields correct recovery depends not only on the total number of errors but also on how errors are distributed. If more errors are distributed to less coherent areas (i.e, with smaller μ_{ij}), then more errors in total can be tolerated. However, if errors are distributed in an opposite manner, then only small number of errors can be tolerated.

2.3 Implication on Graph Clustering

In this subsection, we further illustrate our result when the low-rank matrix is a cluster matrix. Although robust PCA and even more sophisticated approaches have been applied to solve clustering problems, e.g., [13–15], our perspective here is to demonstrate how local coherence affects entry-wise robustness to error corruption, which has not been illustrated in previous studies.

Suppose there are n elements to be clustered. We use a cluster matrix L to represent the clustering relationship of these n elements with $L_{ij} = 1$ if elements i and j are in the

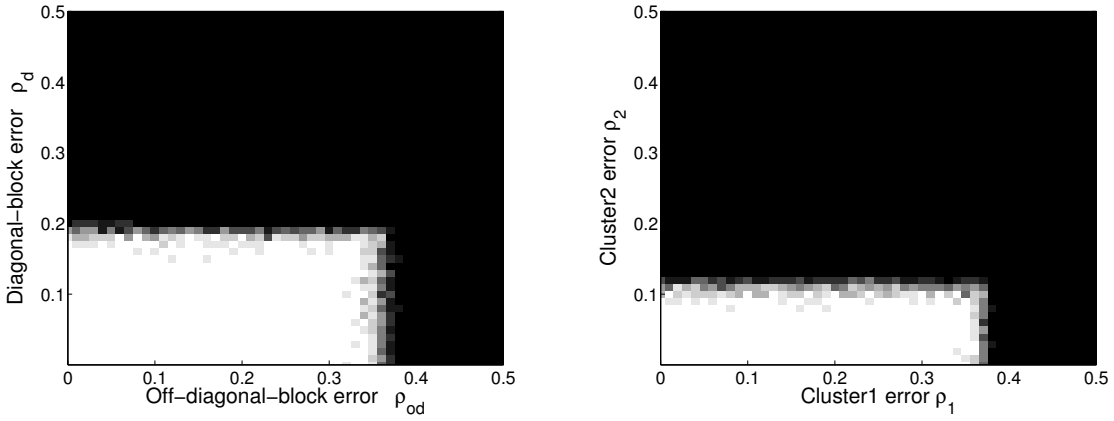
same cluster and $L_{ij} = 0$ otherwise. Thus, with appropriate ordering of the elements, L is a block diagonal matrix with all diagonal blocks containing all '1's and off-diagonal blocks containing all '0's. Hence, the rank r of L equals the number of clusters, which is typically small compared to n. Suppose these entries are corrupted by errors that flip entries from one to zero or from zero to one. This can be thought of as adding a (possibly sparse) error matrix S to L so that the observed matrix is $L + S$. Then PCP can be applied to recover the cluster matrix L.

We first consider an example with clusters having equal size n/r. We set $n = 600$ and $r = 4$ (i.e., four equal-size clusters). We apply errors to diagonal-block entries and off-diagonal-block entries respectively with the probabilities ρ_d and ρ_{od}. In Fig. 1a, we plot recovery accuracy of PCP for each pairs of (ρ_{od}, ρ_d). It is clear from the figure that failure occurs for larger ρ_{od} than ρ_d, which thus implies that off-diagonal blocks are more robust to errors than diagonal blocks. This can be explained by Theorem 2 as follows. For a cluster matrix with equal cluster size n/r, the local coherence parameters are given by

$$\mu_{0ij} = 1 \text{ for all } (i,j), \quad \text{and} \quad \mu_{1ij} = \begin{cases} r, & (i,j) \text{ is in diagonal blocks} \\ 0, & (i,j) \text{ is in off-diagonal blocks,} \end{cases}$$

and thus

$$\mu_{ij} = \max\{\mu_{0ij}, \mu_{1ij}\} = \begin{cases} r, & (i,j) \text{ is in diagonal blocks} \\ 1, & (i,j) \text{ is in off-diagonal blocks.} \end{cases}$$



(a) Diagonal-block error vs. off-diagonal-block error. $n = 600, r = 4$ with equal cluster sizes

(b) Error vulnerability with respect to cluster sizes 500 vs. 100

Figure 1: Error vulnerability on different parts for cluster matrix. In both cases, for each probability pair, we generate 10 trials of independent random error matrices and count the number of successes of PCP. We declare a trial to be successful if the recovered $\hat{L}$ satisfies $\|\hat{L} - L\|_F / \|L\|_F \leq 10^{-3}$. Color from white to black represents the number of successful trials changes from 10 to 0.

Based on Theorem 2, it is clear that diagonal-block entries are more locally coherent and hence are more vulnerable to errors, whereas off-diagonal-block entries are more locally coherent and hence are more robust to errors.

Moreover, this example also demonstrates the necessity of μ_1 in the robust PCA problem. [4] showed that μ_1 is not necessary for matrix completion and argued informally that μ_1 is necessary for robust PCA by connecting the robust PCA problem to hardness of finding a small clique in a large random graph. Here, the above example provides an evidence for such a fact. In the example, μ_{0ij} are the same over the entire matrix, and hence it is μ_{1ij} that differentiates coherence between diagonal blocks and off-diagonal blocks, and thus differentiates their robustness to errors.

We then consider the case with two clusters that have different sizes: cluster1 size 500 versus cluster2 size 100. Hence, $r = 2$. We apply errors to block diagonal entries corresponding to clusters 1 and 2 respectively with the probabilities ρ_1 and ρ_2. In Fig. 1b, we plot the recovery accuracy of PCP for each pair of (ρ_1, ρ_2). It is clear from the figure that failure occurs for larger ρ_1 than ρ_2, which thus implies that entries corresponding to the larger cluster are more robust to errors than entries corresponding to smaller clusters. This can be explained by Theorem 2 because the local coherence of a block diagonal entry is given by $\mu_{ij} = \frac{n^2}{rK^2}$, where K is the corresponding cluster size, and hence the error corruption probability should satisfy $1 - 2\rho_{ij} > C_0 \frac{\sqrt{n}}{K} \log n$ for correct recovery. Thus, a larger cluster can resist denser errors. This also coincides with the results on graph clustering in [13, 16].

2.4 Recovery from errors and partial observation

In this subsection, we generalize our main results to the case when the matrix is only partially observed. Such a study can also be viewed as a refined analysis of recent works [1, 3, 12], where the analysis are based on only global coherence parameters.

We assume that the low-rank matrix L is partially observed on the set $O \subset [n] \times [n]$ and each observed entry $L_{ij}, (i, j) \in O$ is corrupted by an arbitrary noise S_{ij} with probability ρ_{ij} independently. We denote the support set of S by Ω, and hence $\Omega \subseteq O$. We observe $M = \mathcal{P}_O(L) + S$. More specifically, this model can be described as follows.

- Define a random set O by Bernoulli sampling such that

$$\mathbb{P}((i, j) \in O) = \alpha_{ij}. \tag{9}$$

- Conditioning on $(i, j) \in O$, assume that $(i, j) \in \Omega$ for $i, j \in [n]$ are independent events with

$$\mathbb{P}((i, j) \in \Omega | (i, j) \in O) = \rho_{ij}, \tag{10}$$

  which implies that

$$\mathbb{P}((i, j) \in \Omega) = \alpha_{ij}\rho_{ij}. \tag{11}$$

- Define a set $\Gamma := O \backslash \Omega$, then

$$\mathbb{P}((i, j) \in \Gamma) = \alpha_{ij}(1 - \rho_{ij}). \tag{12}$$

- Let S be a matrix supported on Ω with random sign (i.e., the assumption (6)).

For the above model, PCP can be modified as follows:

$$\textit{Adaptive PCP:} \qquad \begin{array}{ll} \underset{L,S}{\text{minimize}} & \|L\|_* + \|\Lambda \circ S\|_1 \\ \text{subject to} & Y = \mathcal{P}_O(L) + S. \end{array} \tag{13}$$

We note that here the parameter Λ is a matrix instead of a scalar in previous results. The following theorem characterizes the conditions under which the *adaptive PCP* returns correct recovery.

Theorem 3. *Consider the robust PCA with partially observed entries under the random sign assumption. Suppose that the partial observation probabilities α_{ij} are known and $\alpha_{ij} > \log^2 n/n$. Then if*

$$\sqrt{\alpha_{ij}}(1-\rho_{ij}) \geq C_0\sqrt{\frac{\mu_{ij}r}{n}}\log n \quad \textit{and} \quad C_0\sqrt{\frac{\mu r}{n}}\log n \leq 1$$

for some sufficient large constant C_0, the adaptive PCP with $\Lambda = \left[\frac{1}{\sqrt{\alpha_{ij}n}\log n}\right]$ recovers the underlying low-rank matrix L with probability at least $1 - cn^{-10}$ for some constant c.

We note that here the base probability is $\frac{\log^2 n}{n}$, which is essentially from the proof of Lemma 4 and consistent with the result of [16]. We note that $\lambda_{\min} \leq \|\Lambda\|_\infty \leq \lambda_{\max}$, where $\lambda_{\min} = \frac{1}{\sqrt{n}\log n}$ and $\lambda_{\max} = \frac{1}{\log^2 n}$. This result generalizes Theorem 1 to robust PCA with only partial observations. It coincides with the results in [3, 12] when all ρ_{ij} are specialized to be the same parameter ρ and all α_{ij} are specialized to be the same parameter α.

Remark 1. *If the prior knowledge on α_{ij} is not available, the standard PCP can still recover the underlying low-rank matrix L with high probability given*

$$\alpha_{ij}(1-\rho_{ij}) \geq C_0\sqrt{\frac{\mu_{ij}r}{n}}\log n \quad \textit{and} \quad C_0\sqrt{\frac{\mu r}{n}}\log n \leq 1,$$

for some sufficient large constant C_0.

This implies the unobserved entries can be treated as corrupted ones and the standard PCP works anyway.

3 Proof Sketch of Theorem 1

The proof of Theorem 1 follows the idea established in [1] and further developed in [3, 12]. Our main technical development lies in analysis of non-uniform error corruption based on local coherence parameters, for which we introduce a new weighted norm $l_{w(\infty)}$, and establish concentration properties and bounds associated with this norm. As a generalization of matrix infinity norm, $l_{w(\infty)}$ incorporates both μ_{0ij} and μ_{1ij}, and is hence different from the weighted

norms $l_{\mu(\infty)}$ and $l_{\mu(\infty,2)}$ in [9]. We next provide the proof sketch here and proofs of related lemmas are regulated to Appendix.

We first introduce some notations. We define the subspace $T := \{UX^* + YV^* : X, Y \in \mathbb{R}^{n\times r}\}$, where U, V are left and right singular matrix of L. Then T induces a projection operator $\mathcal{P}_T$ given by $\mathcal{P}_T(M) = UU^*M + MVV^* - UU^*MVV^*$. Moreover, $T^\perp$, the complement subspace to T, induces an orthogonal projection operator $\mathcal{P}_{T^\perp}$ with $\mathcal{P}_{T^\perp}(M) = (I-UU^*)M(I-VV^*)$. We further define two operators associated with Bernoulli sampling. Let Ω_0 denote a generic subset of $[n] \times [n]$. We define a corresponding projection operator $\mathcal{P}_{\Omega_0}$ as $\mathcal{P}_{\Omega_0}(M) = \sum_{ij} \mathbb{I}_{\{(i,j)\in\Omega_0\}}\langle M, e_ie_j^*\rangle e_ie_j^*$, where $\mathbb{I}_{\{\cdot\}}$ is the indicator function. If Ω_0 is a random set generated by Bernoulli sampling with $\mathbb{P}((i,j) \in \Omega_0) = t_{ij}$ with $0 < t_{ij} \leq 1$ for all $i, j \in [n]$, we further define a linear operator $\mathcal{R}_{\Omega_0}$ as $\mathcal{R}_{\Omega_0}(M) = \sum_{ij} \frac{1}{t_{ij}}\mathbb{I}_{\{(i,j)\in\Omega_0\}}\langle M, e_ie_j^*\rangle e_ie_j^*$. For two variables a and b, $a \vee b = \max\{a, b\}$.

We introduce a new weighted norm. Suppose that μ_{ij}'s are local coherence parameters of L as defined in (8). Let $\hat{w}_{ij} = \sqrt{\frac{\mu_{ij}r}{n^2}}$ and $w_{ij} = \max\{\hat{w}_{ij}, \epsilon\}$, where ϵ is the smallest nonzero $\hat{w}_{ij}$. Here ϵ is introduced to avoid singularity. Then for any matrix Z, define

$$\|Z\|_{w(\infty)} = \max_{i,j} \frac{|Z_{ij}|}{w_{ij}}. \tag{14}$$

It is easy to verify $\|\cdot\|_{w(\infty)}$ is a well defined norm. We will establish several concentration inequalities on this weighted infinity norm, which help to simplify the proof of dual certificate construction.

We further note that throughout this paper "with high probability" means "with probability at least $1 - cn^{-10}$", where the constant c may be different in varying contexts.

Our proof includes two main steps: establishing that existence of a certain dual certificate is sufficient to guarantee correct recovery and constructing such a dual certificate. We first introduce some supporting lemmas.

3.1 Key Properties

We provide a number of concentration properties under non-uniform sampling. These properties are in parallel to those under uniform sampling used in [1, 3, 12]. More specifically, Lemma 1 is proven in [9], which readily implies Lemma 3. We develop the proofs for other lemmas based on local coherence, which are provided in Appendix D.

Lemma 1. *[9, Lemma 9] Suppose $\mathbb{P}((i,j) \in \Omega_0) = q_{ij}$ for all $i, j \in [n]$. If $q_{ij} \geq C_0(\mu_{0ij}r\log n)/n$ for some sufficiently large constant C_0 and for all $i, j \in [n]$, then with high probability*

$$\|\mathcal{P}_T - \mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T\| \leq \frac{1}{2}. \tag{15}$$

Lemma 2. *Suppose $\mathbb{P}((i,j) \in \Omega_0) = q_{ij}$ for all $i, j \in [n]$. If $\|\mathcal{P}_T - \mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T\| \leq \frac{1}{2}$ and $q_{ij} \geq p_0$ for all $i, j \in [n]$, then*

(a) $\|\mathcal{P}_T\mathcal{R}_{\Omega_0}\| \le \sqrt{\frac{3}{2p_0}}$;

(b) $\mathcal{P}_{\Omega_0}\mathcal{P}_T$ *is injective on* T.

Lemma 3. *Suppose* $\mathbb{P}((i,j) \in \Omega_0) = q_{ij}$ *for all* $i, j \in [n]$. *For a fixed matrix* $Z \in T$, *if* $q_{ij} \ge C_0(\mu_{ij} r \log n)/n$ *for some sufficiently large constant* C_0 *and for all* $i, j \in [n]$, *then with high probability*

$$\|Z - \mathcal{P}_T\mathcal{R}_{\Omega_0}(Z)\|_F \le \frac{1}{2}\|Z\|_F. \tag{16}$$

Lemma 4. *Suppose* $\mathbb{P}((i,j) \in \Omega_0) = q_{ij}$ *for all* $i, j \in [n]$. *For a fixed matrix* $Z \in T$, *if* $q_{ij} \ge C_0(\sqrt{\mu_{ij} r} \vee \mu_{ij} r)\frac{\log n}{n}$ *for some sufficiently large constant* C_0 *and for all* $i, j \in [n]$, *then with high probability*

$$\|(\mathcal{R}_{\Omega_0} - I)Z\| \le \frac{C}{C_0}\|Z\|_{w(\infty)} \tag{17}$$

for some constant C.

Lemma 5. *Suppose* $\mathbb{P}((i,j) \in \Omega_0) = q_{ij}$ *for all* $i, j \in [n]$. *Suppose* $\beta > 0$ *is a scaling factor. For a fixed matrix* $Z \in T$, *if* $q_{ij} \ge C_0\beta^{-2}(\mu_{ij} r)\frac{\log n}{n}$ *for some sufficiently large* C_0 *and for all* $i, j \in [n]$, *then with high probability*

$$\|(\mathcal{P}_T\mathcal{R}_{\Omega_0} - \mathcal{P}_T)Z\|_{w(\infty)} \le \frac{1}{2}\beta\|Z\|_{w(\infty)}. \tag{18}$$

Proposition 1. *Suppose* S *is the error matrix in the random sign model defined in Section 2.1. Then for any given index* (a, b) *with* $a, b \in [n]$, *with high probability*

$$|[\mathcal{P}_T \operatorname{sgn}(S)]_{ab}| \le C\sqrt{\frac{\mu_{ab} r \log n}{n}} \tag{19}$$

for some constant C.

3.2 Dual Certificate Condition

Proposition 2. *If* $1 - \rho_{ij} \ge \max\left\{C_0\frac{\mu_{ij} r}{n}\log n, \frac{1}{n^3}\right\}$, *PCP yields a unique solution which agrees with the correct* (L, S) *with high probability if there exists a dual certificate* Y *obeying*

$$\mathcal{P}_\Omega Y = 0, \tag{20}$$

$$\|Y\|_\infty \le \frac{\lambda}{4}, \tag{21}$$

$$\|\mathcal{P}_{T^\perp}(\lambda \operatorname{sgn}(S) + Y)\| \le \frac{1}{4}, \tag{22}$$

$$\|\mathcal{P}_T(Y + \lambda \operatorname{sgn}(S) - UV^*)\|_F \le \frac{\lambda}{n^2} \tag{23}$$

where $\lambda = \frac{1}{32\sqrt{n \log n}}$.

The proof of the above proposition adapts the idea in [1, 12] for uniform errors to non-uniform errors. In particular, the proof exploits the properties of $\mathcal{R}_\Omega$, which are presented as Lemma 1 (established in [9]) and Lemma 2.

3.3 Dual Certificate Construction

Proposition 2 suggests that it suffices to prove Theorem 1 if we find a dual certificate Y that satisfies the dual certificate conditions (20)-(23). Thus, the second step is to construct Y via the golfing scheme. Although we adapt the steps in [12] to construct the dual certificate Y, our analysis requires new technical development based on local coherence parameters. Recall the following definitions in Section 2.1: $\mathbb{P}((i,j) \in \Omega) = \rho_{ij}$ and $\mathbb{P}((i,j) \in \Gamma) = p_{ij}$, where $\Gamma = \Omega^c$ and $p_{ij} = 1 - \rho_{ij}$.

Consider the golfing scheme with nonuniform sizes as suggested in [12] to establish bounds with fewer log factors. Let $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \cdots \cup \Gamma_l$, where $\{\Gamma_k\}$ are independent random sets given by

$$\mathbb{P}((i,j) \in \Gamma_1) = \frac{p_{ij}}{6}, \qquad \mathbb{P}((i,j) \in \Gamma_k) = q_{ij}, \quad \text{for} \quad k = 2, \cdots, l.$$

Thus, if $\rho_{ij} = (1 - \frac{p_{ij}}{6})(1 - q_{ij})^{l-1}$, the two sampling strategies are equivalent. Due to the overlap between $\{\Gamma_k\}$, we have $q_{ij} \geq \frac{5}{6}\frac{p_{ij}}{l-1}$. We set $l = \lfloor 5 \log n + 1 \rfloor$ and construct a dual certificate Y in the following iterative way:

$$Z_0 = \mathcal{P}_T(UV^* - \lambda \operatorname{sgn}(S)) \tag{24}$$

$$Z_k = (\mathcal{P}_T - \mathcal{P}_T \mathcal{R}_{\Gamma_k} \mathcal{P}_T) Z_{k-1}, \quad \text{for} \quad k = 1, \cdots, l \tag{25}$$

$$Y = \sum_{k=1}^{l} \mathcal{R}_{\Gamma_k} Z_{k-1}. \tag{26}$$

It is then sufficient to show that such constructed Y satisfies the dual certificate conditions (20)-(23). Condition (20) is satisfied due to the construction of Y. Condition (23) can be shown by a concentration property of each iteration step (45) with $\|\cdot\|_F$ characterized in Lemma 3. Conditions (21) and (22) can be shown by concentration properties of each iteration step with regard to $\|\cdot\|$ and $\|\cdot\|_{w(\infty)}$ norms, which are characterized respectively in Lemmas 4 and 5.

More specifically, note that we require in the theorem $(1 - \rho_{ij}) \geq C_0 \sqrt{\mu_{ij} r / n} \log n$ and $C_0 \sqrt{\mu r / n} \log n \leq 1$, which imply

$$p_{ij} = (1 - \rho_{ij}) \geq C_0 \sqrt{\frac{\mu_{ij} r}{n}} \log n \geq C_0 \left(\frac{\mu_{ij} r}{n} \vee \frac{\sqrt{\mu_{ij} r}}{n} \right) \log^2 n. \tag{27}$$

Thus by ignoring the constant factors, we have

$$\mathbb{P}((i,j) \in \Gamma_1) \geq C_0 \sqrt{\frac{\mu_{ij} r}{n}} \log n \geq C_0 \left(\frac{\mu_{ij} r}{n} \vee \frac{\sqrt{\mu_{ij} r}}{n} \right) \log^2 n, \tag{28}$$

$$\mathbb{P}((i,j) \in \Gamma_k) \geq C_0 \sqrt{\frac{\mu_{ij} r}{n}} \geq C_0 \left(\frac{\mu_{ij} r}{n} \vee \frac{\sqrt{\mu_{ij} r}}{n} \right) \log n, \quad \text{for k=1, 2, ..., l.} \tag{29}$$

Then by the result of Lemma 4 and (29), we have with high probability

$$\|(I - \mathcal{R}_{\Gamma_k}) Z_{k-1}\| \leq \frac{C}{C_0} \|Z_{k-1}\|_{w(\infty)}, \quad \text{for k=1,2,..., l.} \tag{30}$$

By the result of Lemma 5 and (28), we have with high probability

$$\|Z_1\|_{w(\infty)} \leq \frac{1}{2\sqrt{\log n}}\|Z_0\|_{w(\infty)}, \tag{31}$$

Consequently, further by Lemma 5 and (29), we have

$$\|Z_k\|_{w(\infty)} \leq \frac{1}{2}\|Z_{k-1}\|_{w(\infty)} \leq \frac{1}{2^k\sqrt{\log n}}\|Z_0\|_{w(\infty)} \quad \text{for } k = 2, \cdots, l. \tag{32}$$

We further note that Lemma 3 and (29) imply with high probability

$$\|Z_k\|_F \leq \frac{1}{2}\|Z_{k-1}\|_F. \tag{33}$$

We next bound $\|Z_0\|_F$ and $\|Z_0\|_{w(\infty)}$. Observe that for an index pair (a, b), we have

$$|[Z_0]_{ab}| \leq |[UV^*]_{ab}| + \lambda|[\mathcal{P}_T \operatorname{sgn}(S)]_{ab}|.$$

Applying Proposition 1 and using the facts $|[UV^*]_{ab}| \leq \sqrt{\frac{\mu_{ab}r}{n^2}}$ and $\lambda = \frac{1}{32\sqrt{n\log n}}$, we obtain

$$\|Z_0\|_\infty \leq C\sqrt{\mu r}/n. \tag{34}$$

Consequently,

$$\|Z_0\|_F \leq \|UV^*\|_F + \lambda\|\mathcal{P}_T \operatorname{sgn}(S)\|_F \leq \sqrt{r} + C\sqrt{\mu r} \leq C'\sqrt{\mu r} \tag{35}$$

where we used $\|Z\|_F \leq n\|Z\|_\infty$ for any matrix Z, and

$$\|Z_0\|_{w(\infty)} \leq \|UV^*\|_{w(\infty)} + \lambda\|\mathcal{P}_T \operatorname{sgn}(S)\|_{w(\infty)} \leq 1 + \max_{a,b} \lambda\frac{|[\mathcal{P}_T \operatorname{sgn}(S)]_{ab}|}{w_{ab}} \leq C', \tag{36}$$

where we used the definition of w_{ab} and $\lambda = \frac{1}{32\sqrt{n\log n}}$. We note that for the sake of convenience, the constants C and C' may be different from line to line.

We are now ready to show that the constructed dual certificate Y obeys the conditions (20)-(23) in Proposition 2. Clearly, Y satisfies $\mathcal{P}_\Omega Y = 0$ due to the construction.

In order to show that Y satisfies (23), we derive

$$\begin{aligned}
&\|\mathcal{P}_T Y + \mathcal{P}_T(\lambda \operatorname{sgn}(S) - UV^*)\|_F \\
&\quad = \left\| Z_0 - \left(\sum_{k=1}^{l} \mathcal{P}_T\mathcal{R}_{\Gamma_k} Z_{k-1}\right)\right\|_F \\
&\quad = \left\| (\mathcal{P}_T - \mathcal{P}_T\mathcal{R}_{\Gamma_1})Z_0 - \left(\sum_{k=2}^{l} \mathcal{P}_T\mathcal{R}_{\Gamma_k} Z_{k-1}\right)\right\|_F \\
&\quad = \left\| \mathcal{P}_T Z_1 - \left(\sum_{k=1}^{l} \mathcal{P}_T\mathcal{R}_{\Gamma_k} Z_{k-1}\right)\right\|_F \\
&\quad = \cdots \\
&\quad = \|Z_l\|_F \overset{(a)}{\leq} \left(\frac{1}{2}\right)^l \cdot \|Z_0\|_F \overset{(b)}{\leq} C'\left(\frac{1}{2}\right)^l \sqrt{\mu r} \leq \frac{\lambda}{n^2},
\end{aligned}$$

where (a) follows from (33) and (b) follows from (35).

In order to show that Y satisfies (22), we respectively show that $\|\mathcal{P}_{T^\perp} Y\| \leq \frac{1}{8}$ and $\|\mathcal{P}_{T^\perp}(\lambda \operatorname{sgn}(S))\| \leq \frac{1}{8}$ as follows. Firstly,

$$\begin{aligned}
\|\mathcal{P}_{T^\perp} Y\| &= \left\| \mathcal{P}_{T^\perp} \sum_{k=1}^{l} \mathcal{R}_{\Gamma_k} Z_{k-1} \right\| \\
&\leq \sum_{k=1}^{l} \|\mathcal{P}_{T^\perp} \mathcal{R}_{\Gamma_k} Z_{k-1}\| \\
&\overset{(a)}{=} \sum_{k=1}^{l} \|\mathcal{P}_{T^\perp} (\mathcal{R}_{\Gamma_k} Z_{k-1} - Z_{k-1})\| \\
&\leq \sum_{k=1}^{l} \|\mathcal{R}_{\Gamma_k} Z_{k-1} - Z_{k-1}\| \\
&\overset{(b)}{\leq} \sum_{k=1}^{l} \frac{C}{C_0} \|Z_{k-1}\|_{w(\infty)} \\
&\overset{(c)}{\leq} \frac{C}{C_0} \left(1 + \sum_{k=2}^{l} \frac{1}{\sqrt{\log n}} \left(\frac{1}{2} \right)^{k-1} \right) \|Z_0\|_{w(\infty)} \\
&\leq \frac{2C}{C_0} \|Z_0\|_{w(\infty)} \overset{(d)}{\leq} \frac{1}{8},
\end{aligned}$$

where (a) follows because $Z_{k-1} \in T$, (b) follows from (30), (c) follows from (31) and (32), and (d) follows from (36) and C_0 is sufficiently large. Next, by applying the spectral norm bound on random matrix in [17], we have

$$\|\mathcal{P}_{T^\perp}(\lambda \operatorname{sgn}(S))\| \leq \lambda \|\operatorname{sgn}(S)\| \leq \lambda \cdot 4\sqrt{n}.$$

Since $\lambda = \frac{1}{32\sqrt{n \log n}}$, we have

$$\|\mathcal{P}_{T^\perp}(\lambda \operatorname{sgn}(S))\| \leq \frac{1}{8\sqrt{\log n}} \leq \frac{1}{8}.$$

In order to show that Y satisfies (21), we derive

$$\begin{aligned}
\|Y\|_\infty &= \left\| \sum_{k=1}^{l} \mathcal{R}_{\Gamma_k} Z_{k-1} \right\|_\infty \\
&\overset{(a)}{\leq} \left\| \sum_{i,j} \frac{\mathbb{I}_{\{(i,j)\in\Gamma_1\}}}{\mathbb{P}((i,j)\in\Gamma_1)} [Z_0]_{ij} e_i e_j^* \right\|_\infty + \sum_{k=2}^{l} \left\| \sum_{i,j} \frac{\mathbb{I}_{\{(i,j)\in\Gamma_k\}}}{\mathbb{P}((i,j)\in\Gamma_k)} [Z_{k-1}]_{ij} e_i e_j^* \right\|_\infty \\
&\overset{(b)}{\leq} \frac{1}{C_0\sqrt{n}\log n} \|Z_0\|_{w(\infty)} + \sum_{k=2}^{l} \frac{1}{C_0\sqrt{n}} \|Z_{k-1}\|_{w(\infty)} \\
&\overset{(c)}{\leq} \frac{1}{C_0\sqrt{n}\log n} \|Z_0\|_{w(\infty)} + \sum_{k=2}^{l} \frac{1}{C_0\sqrt{n\log n}} \left(\frac{1}{2}\right)^{k-1} \|Z_0\|_{w(\infty)} \\
&\leq \frac{2}{C_0\sqrt{n\log n}} \|Z_0\|_{w(\infty)} \\
&\overset{(d)}{\leq} \frac{64C}{C_0}\lambda \overset{(e)}{\leq} \frac{\lambda}{4},
\end{aligned}$$

where (a) is due to the golfing scheme with non-uniform partitions, (b) is due to the first inequalities of (28) and (29), (c) follows from (31) and (32), (d) follows from (36) and (e) follows because C_0 is sufficiently large.

4 Numerical Experiments

In this section, we provide numerical experiments to demonstrate our theoretical results. In these experiments, we adopt an augmented Lagrange multiplier algorithm in [18] to solve the PCP. We set $\lambda = 1/\sqrt{n\log n}$. A trial of PCP (for a given realization of error locations) is declared to be successful if $\hat{L}$ recovered by PCP satisfies $\|\hat{L} - L\|_F / \|L\|_F \leq 10^{-3}$.

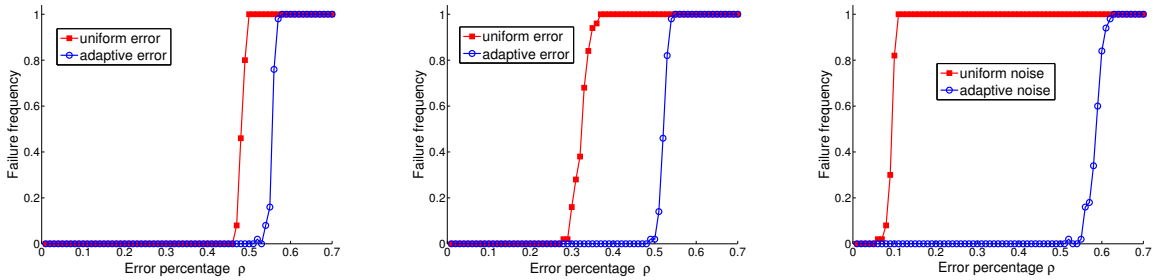
We apply the following three models to construct the low-rank matrix L.

- Bernoulli model: $L = XX^*$ where X is $n \times r$ matrix with entries independently taking values $+1/\sqrt{n}$ and $-1/\sqrt{n}$ equally likely.
- Gaussian model: $L = XX^*$, where X is $n \times r$ matrix with entries independently sampled from Gaussian distribution $\mathcal{N}(0, 1/n)$.
- Cluster model: L is a block diagonal matrix with r equal-size blocks containing all '1's.

In order to demonstrate that the local coherence parameter affects local robustness to error corruptions, we study the following two types of error corruption models.

- Uniform error corruption: $\mathrm{sgn}(S_{ij})$ is generated as (6) with $\rho_{ij} = \rho$ for all $i, j \in [n]$, and $S = \mathrm{sgn}(S)$.
- Adaptive error corruption: $\mathrm{sgn}(S_{ij})$ is generated as (6) with $\rho_{ij} = \rho \frac{n^2\sqrt{1/\mu_{ij}}}{\sum_{ij}\sqrt{1/\mu_{ij}}}$ for all $i, j \in [n]$, and $S = \mathrm{sgn}(S)$.

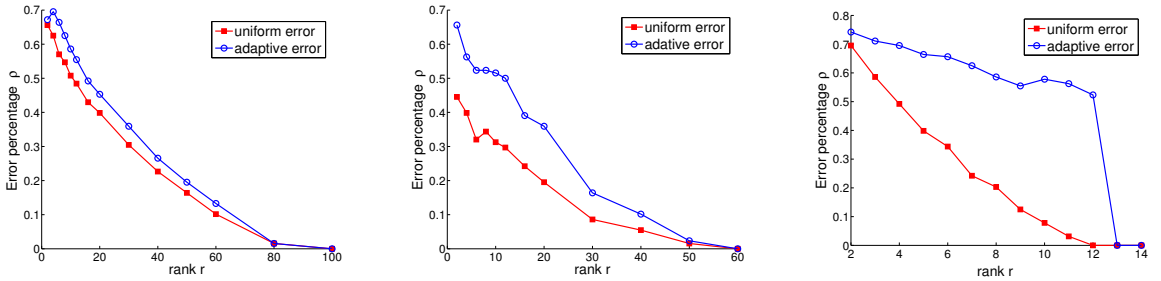
It is clear in both cases, the error matrix has the same average error corruption percentage ρ, but in adaptive error corruption, the local error corruption probability is adaptive to the local coherence.



(a) Bernoulli model (b) Gaussian model (c) Cluster model

Figure 2: Recovery failure of PCP versus error corruption percentage.

Our first experiment demonstrates that robustness of PCP to error corruption not only depends on the number of errors but also depends on how errors are distributed over the matrix. For all three low-rank matrix models, we set $n = 1200$ and rank $r = 10$. For each low-rank matrix model, we apply the uniform and adaptive error matrices, and plot the failure frequency of PCP versus the error corruption percentage ρ in Fig. 2. For each value of ρ, we perform 50 trials of independent error corruption and count the number of failures of PCP. Each plot of Fig. 2 compares robustness of PCP to uniform error corruption (the red square line) and adaptive error corruption (the blue circle line). We observe that PCP can tolerate more errors in the adaptive case. This is because the adaptive error matrix is distributed based on the local coherence parameter, where error density is higher in areas where matrices can tolerate more errors. Furthermore, comparison among the three plots in Fig. 2 illustrates that the gap between uniform and adaptive error matrices is the smallest for Bernoulli model and the largest for cluster model. Our theoretic results suggest that the gap is due to the variation of the local coherence parameter across the matrix, which can be measured by the variance of μ_{ij}. Larger variance of μ_{ij} should yield larger gap. Our numerical calculation of the variances for three models yield $\mathsf{Var}(\mu_{Bernoulli}) = 1.2109$, $\mathsf{Var}(\mu_{Gaussian}) = 2.1678$, and $\mathsf{Var}(\mu_{cluster}) = 7.29$, which confirms our explanation.



(a) Bernoulli model (b) Gaussian model (c) Cluster model

Figure 3: Largest allowable error corruption percentage versus rank of L so that PCP yields correct recovery.

We next study the phase transition in rank and error corruption probability. For the three low-rank matrix models, we set $n = 1200$. In Fig. 3, we plot the error corruption percentage versus the rank of L for both uniform and adaptive error corruption models.

Each point on the curve records the maximum allowable error corruption percentage under the corresponding rank such that PCP yields correction recovery. We count a (r, ρ) pair to be successful if nine trials out of ten are successful. We first observe that in each plot of Fig. 3, PCP is more robust in adaptive error corruption due to the same reason explained above. We further observe that the gap between the uniform and adaptive error corruption changes as the rank changes. In the low-rank regime, the gap is largely determined by the variance of coherence parameter μ_{ij} as we argued before. As the rank increases, the gap is more dominated by the rank and less affected by the local coherence. Eventually for large enough rank, no error can be tolerated no matter how errors are distributed.

5 Conclusion

We characterize refined conditions under which PCP succeeds to solve the robust PCA problem. Our result shows that the ability of PCP to correctly recover a low-rank matrix from errors is related not only to the total number of corrupted entries but also to locations of corrupted entries, more essentially to the local coherence of the low-rank matrix. Such result is well supported by our numerical experiments. Moreover, our result has rich implication when the low-rank matrix is a cluster matrix, and our result coincides with state-of-the-art studies on clustering problems via low-rank cluster matrix. Our result may motivate the development of weighted PCP to improve recovery performance similar to the weighted algorithms developed for matrix completion in [9, 19].

References

[1] E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

[2] V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[3] Y. Chen, A. Jalali, S. Sanghavi, and C. Caramanis. Low-rank matrix recovery from errors and erasures. *IEEE Transactions on Information Theory*, 59(7):4324–4337, 2013.

[4] Y. Chen. Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory*, 61(5):2909–2923, May 2015.

[5] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

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

[7] D. Gross. Recovering low-rank matrices from few coefficients in any basis. *IEEE Transactions on Information Theory*, 57(3):1548–1566, 2011.

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

[9] Y. Chen, S. Bhojanapalli, S. Sanghavi, and R. Ward. Completing any low-rank matrix, provably. *arXiv preprint arXiv:1306.2979*, 2013.

[10] D. Hsu, S. M. Kakade, and T. Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Transactions on Information Theory*, 57(11):7221–7234, 2011.

[11] A. Ganesh, J. Wright, X. Li, E. J. Candes, and Y. Ma. Dense error correction for low-rank matrices via principal component pursuit. In *IEEE International Symposium on Information Theory (ISIT)*, pages 1513–1517, Austin, TX, US, June 2010.

[12] X. Li. Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1):73–99, 2013.

[13] S. Oymak and B. Hassibi. Finding dense clusters via "low rank+ sparse" decomposition. *arXiv preprint arXiv:1104.5186*, 2011.

[14] Y. Chen, S. Sanghavi, and H. Xu. Clustering sparse graphs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2204–2212, Lake Tahoe, Nevada, US, December 2012.

[15] Y. Chen, S. Sanghavi, and H. Xu. Improved graph clustering. *IEEE Transactions on Information Theory*, 60(10):6440–6455, Oct 2014.

[16] Y. Chen, A. Jalali, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. *Journal of Machine Learning Research*, 15(1):2213–2238, 2014.

[17] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

[18] Z. Lin, M. Chen, and Y. Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[19] N. Srebro and R. R. Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2056–2064, Hyatt Regency, Vancouver, Canada, 2010. December.

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

Supplementary Materials

A Proofs in Theorem 1

A.1 Proof of Proposition 2 (Dual Certificate Conditions)

Due to the assumption of the proposition, $\Gamma = \Omega^c$ satisfies the conditions required in Lemma 1. Hence, due to Lemmas 1 and 2, we have $\|\mathcal{P}_T\mathcal{R}_\Gamma\| \leq \sqrt{\frac{3}{2p_0}}$ with $p_0 = 1/n^3$ and $\mathcal{P}_\Gamma\mathcal{P}_T$ is injective on T with high probability.

Suppose $\hat{L} = L + H$ and $\hat{S} = S - H$ satisfy

$$\|L + H\|_* + \lambda\|S - H\|_1 \leq \|L\|_* + \lambda\|S\|_1. \tag{37}$$

By the definition of subgradient, we have

$$\|L + H\|_* \geq \|L\|_* + \langle \mathcal{P}_T H, UV^* \rangle + \|\mathcal{P}_{T^\perp} H\|_*$$

where we use the fact that there exists $W \in T^\perp$ and $\|W\| \leq 1$ such that $\|\mathcal{P}_{T^\perp} H\|_* = \langle \mathcal{P}_{T^\perp} H, W \rangle$.

Thus, we have

$$\langle \mathcal{P}_T H, UV^* \rangle + \|\mathcal{P}_{T^\perp} H\|_* \leq \lambda\|S\|_1 - \lambda\|S - H\|_1.$$

Furthermore,

$$\begin{aligned} \|S - H\|_1 &= \|S - \mathcal{P}_\Omega H\|_1 + \|\mathcal{P}_\Gamma H\|_1 \\ &\geq \|S\|_1 + \langle \operatorname{sgn}(S), -H \rangle + \|\mathcal{P}_\Gamma H\|_1. \end{aligned}$$

Combining the last two inequalities, we have

$$\|\mathcal{P}_{T^\perp} H\|_* + \lambda\|\mathcal{P}_\Gamma H\|_1 \leq \langle H, \lambda \operatorname{sgn}(S) - UV^* \rangle.$$

For a matrix Y that obeys the conditions in the Proposition 2, we derive

$$\begin{aligned} &\langle H, \lambda \operatorname{sgn}(S) - UV^* \rangle \\ &\quad = \langle H, Y + \lambda \operatorname{sgn}(S) - UV^* \rangle - \langle H, Y \rangle \\ &\quad = \langle \mathcal{P}_T H, \mathcal{P}_T(Y + \lambda \operatorname{sgn}(S) - UV^*) \rangle + \langle \mathcal{P}_{T^\perp} H, \mathcal{P}_{T^\perp}(Y + \lambda \operatorname{sgn}(S)) \rangle \\ &\qquad - \langle \mathcal{P}_\Gamma H, \mathcal{P}_\Gamma Y \rangle - \langle \mathcal{P}_\Omega H, \mathcal{P}_\Omega Y \rangle \\ &\quad \leq \frac{\lambda}{n^2}\|\mathcal{P}_T H\|_F + \frac{1}{4}\|\mathcal{P}_{T^\perp} H\|_* + \frac{\lambda}{4}\|\mathcal{P}_\Gamma H\|_1. \end{aligned}$$

Combining the previous two inequalities, we obtain

$$\frac{3}{4}\|\mathcal{P}_{T^\perp} H\|_* + \frac{3}{4}\lambda\|\mathcal{P}_\Gamma H\|_1 \leq \frac{\lambda}{n^2}\|\mathcal{P}_T H\|_F.$$

We next bound $\|\mathcal{P}_T H\|_F$ as follows:

$$\begin{aligned}\|\mathcal{P}_T H\|_F &\leq 2\|\mathcal{P}_T \mathcal{R}_\Gamma \mathcal{P}_T(H)\|_F \\ &\leq 2\|\mathcal{P}_T \mathcal{R}_\Gamma \mathcal{P}_{T^\perp}(H)\|_F + 2\|\mathcal{P}_T \mathcal{R}_\Gamma(H)\|_F \\ &\leq \sqrt{\frac{6}{p_0}}\|\mathcal{P}_{T^\perp}(H)\|_F + \sqrt{\frac{6}{p_0}}\|\mathcal{P}_\Gamma(H)\|_F.\end{aligned}$$

We thus obtain

$$\left(\frac{3}{4} - \frac{\lambda}{n^2}\sqrt{\frac{6}{p_0}}\right)\|\mathcal{P}_{T^\perp}(H)\|_F + \left(\frac{3}{4}\lambda - \frac{\lambda}{n^2}\sqrt{\frac{6}{p_0}}\right)\|\mathcal{P}_\Gamma(H)\|_F \leq 0.$$

The above inequality implies that if $p_0 \geq 1/n^3$, then $\mathcal{P}_{T^\perp} H = \mathcal{P}_\Gamma H = 0$. This further implies $\mathcal{P}_\Gamma \mathcal{P}_T(H) = 0$. Since $\mathcal{P}_\Gamma \mathcal{P}_T$ is injective on T, we have $\mathcal{P}_T H = 0$. Consequently, $H = 0$.

B Proof of Theorem 2

The argument here adapts from Section 2.1 and 2.2 of [1]. For completeness, we present it again.

B.1 Elimination Procedure

We begin with a definition and then establish its property.

Definition 1. *S' is said to be a* trimmed version *of S if $supp(S') \subset supp(S)$ and $S'_{ij} = S_{ij}$ whenever $S'_{ij} \neq 0$.*

The following theorem claims that if PCP correctly recovers the low-rank and sparse components of $M_0 = L_0 + S_0$, it also correctly recovers the components of a matrix $M'_0 = L_0 + S'_0$ where S'_0 is a trimmed version of S_0.

Theorem 4 (Theorem 2.2 in [1]). *Suppose the solution to* (13) *with input data $M_0 = L_0 + S_0$ is unique and exact, and consider $M'_0 = L_0 + S'_0$, where S'_0 is a trimmed version of S_0. Then the solution to* (13) *with input M'_0 is exact as well.*

B.2 Derandomization Procedure

Let $\boldsymbol{\rho}$ be the matrix with each (i,j)-entry being ρ_{ij}. If PCP yields exact recovery with a certain probability for the random sign model with the parameter $2\boldsymbol{\rho}$, then it also yields exact recovery with at least the same probability for the fixed sign model with locations of non-zero entries sampled using Bernoulli model with the parameter $\boldsymbol{\rho}$.

Theorem 5 (Theorem 2.3 in [1]). *Suppose L_0 is a low-rank matrix with local coherence parameter $[\mu_{ij}]$ and S_0 follow the Bernoulli model with parameter $2\boldsymbol{\rho}$, and the signs of S_0 are independently distributed ± 1 as stated in* (6) *(and independent from the locations). Then, if the PCP solution is exact with high probability, then it is also exact with at least the same probability for the model in which the signs are fixed and the locations are sampled from the Bernoulli model with parameter $\boldsymbol{\rho}$.*

Proof. Consider the model with fixed signs assumption. We view S_0 as $\mathcal{P}_\Omega S$ for some fixed matrix S, where Ω is sampled from the Bernoulli model with parameter $\boldsymbol{\rho}$. Therefore, S_0 has following distribution

$$(S_0)_{ij} = \begin{cases} S_{ij}, & \text{w. p. } \rho_{ij} \\ 0, & \text{w. p. } 1-\rho_{ij}. \end{cases}$$

Now consider a random sign matrix with each entry distributed independently as follows

$$E_{ij} = \begin{cases} 1, & \text{w. p. } \rho_{ij}, \\ 0, & \text{w. p. } 1-2\rho_{ij}, \\ -1, & \text{w. p. } \rho_{ij}, \end{cases}$$

and an "elimination" matrix η with entries defined by

$$\eta_{ij} = \begin{cases} 0, & \text{if } E_{ij}[\operatorname{sgn}(S)]_{ij} = -1, \\ 1, & \text{otherwise.} \end{cases}$$

The entries of η are independent since they are functions of independent random variables.

Consider now $S_0' = \eta \circ (|S| \circ E)$, where $\circ$ denotes the componentwise product so that $[S_0']_{ij} = \eta_{ij} \circ (|S_{ij}| \circ E_{ij})$. Then, we claim that S_0' and S_0 have the same distribution. By independence of each entry, it suffices to check that their marginals match each other. For $S_{ij} \neq 0$, we have

$$\begin{aligned} \mathbb{P}([S_0']_{ij} = S_{ij}) &= \mathbb{P}(\eta_{ij} = 1 \text{ and } E_{ij} = [\operatorname{sgn}(S)]_{ij}) \\ &= \mathbb{P}(E_{ij}[\operatorname{sgn}(S)]_{ij} \neq -1 \text{ and } E_{ij} = [\operatorname{sgn}(S)]_{ij}) \\ &= \mathbb{P}(E_{ij} = [\operatorname{sgn}(S)]_{ij}) = \rho_{ij}, \end{aligned}$$

which establishes the claim.

Now, $|S| \circ E$ now obeys the random sign model, and by assumption, PCP recovers $|S| \circ E$ with high probability. By the elimination procedure, PCP also recovers $S_0' = \eta \circ (|S| \circ E)$. Since S_0' and S_0 have the same distribution, the theorem follows. □

C Proof of Theorem 3

C.1 Dual Certificate Condition

We follow the idea of [12] and introduce another model which is equivalent to the model defined in Section 2.4 and easy to deal with.

1. Define two independent random subsets of $[n] \times [n]$: Γ with $\mathbb{P}((i,j) \in \Gamma) = \alpha_{ij}(1-\rho_{ij})$ and Ω' with $\mathbb{P}((i,j) \in \Omega') = \frac{\alpha_{ij}\rho_{ij}}{1-\alpha_{ij}+\alpha_{ij}\rho_{ij}}$. Let $O = \Gamma \cup \Omega'$, then $\mathbb{P}((i,j) \in O) = \alpha_{ij}$.

2. Define $\Omega := \Omega' \backslash \Gamma = \{(i,j) : (i,j) \in \Omega' \text{ and } (i,j) \notin \Gamma\}$. Then $\mathbb{P}\{(i,j) \in \Omega\} = \alpha_{ij}\rho_{ij}$.

3. Define a matrix W with each entry W_{ij} being $+1$ or -1 equal randomly and independently. That is $\mathbb{P}\{W_{ij} = +1\} = \mathbb{P}\{W_{ij} = -1\} = 1/2$ for all $(i,j) \in [n] \times [n]$.

4. Let S be a matrix supported on Ω. The signs of S coincide with W on Ω. This means $[\mathrm{sgn}(S)]_{ij}$ are independent random variables with the following distribution

$$[\mathrm{sgn}(S)]_{ij} = \begin{cases} 1 & \text{with prob.} \quad \frac{\alpha_{ij}\rho_{ij}}{2}, \\ 0 & \text{with prob.} \quad 1-\alpha_{ij}\rho_{ij}, \\ -1 & \text{with prob.} \quad \frac{\alpha_{ij}\rho_{ij}}{2}. \end{cases} \tag{38}$$

Proposition 3. *Suppose $\|\mathcal{P}_T\mathcal{R}_\Gamma\| \le \sqrt{\frac{3}{2p_0}}$ and $\mathcal{P}_\Gamma\mathcal{P}_T$ is injective on T. The adaptive PCP program produces a unique solution if there exists a dual certificate Y obeying*

$$\mathcal{P}_{\Gamma^c} Y = 0, \tag{39}$$

$$\|\mathcal{P}_\Gamma Y \circ \frac{1}{\Lambda}\|_\infty \le \frac{1}{4}, \tag{40}$$

$$\|\mathcal{P}_{T^\perp}(\Lambda \circ \mathrm{sgn}(S) + Y)\| \le \frac{1}{4}, \tag{41}$$

$$\|\mathcal{P}_T(Y + \Lambda \circ \mathrm{sgn}(S) - UV^*)\|_F \le \frac{\lambda_{\min}}{n^2} \tag{42}$$

where Λ is a matrix with each entry $\Lambda_{ij} = \frac{1}{\sqrt{\alpha_{ij} n \log n}}$, $\lambda_{\min} = \frac{1}{\sqrt{n}\log n}$, and $\circ$ is entrywise product.

Proof. Suppose the PCP give us a solution $(\hat{L}, \hat{S})$. Assume $\hat{L} = L + H$. Then we have $\mathcal{P}_O(H) = S - \hat{S}$ because of the relation $\mathcal{P}_O(L) + S = \mathcal{P}_O(\hat{L}) + \hat{S}$. It is clear that $\hat{S}$ is supported on O because S is supported on $\Omega \subset O$. By the definition of $(\hat{L}, \hat{S})$, we have

$$\|\hat{L}\|_* + \|\Lambda \circ \hat{S}\|_1 \le \|L\|_* + \|\Lambda \circ S\|_1 \tag{43}$$

By the definition of subgradient, we know

$$\|L + H\|_* \ge \|L\|_* + \langle \mathcal{P}_T H, UV^* \rangle + \|\mathcal{P}_{T^\perp} H\|_*$$

because we can always find a $W \in T^\perp$ and $\|W\| \le 1$ such that $\|\mathcal{P}_{T^\perp} H\|_* = \langle \mathcal{P}_{T^\perp} H, W \rangle$.

Thus we have

$$\|\Lambda \circ S\|_1 - \|\Lambda \circ \hat{S}\|_1 \ge \langle \mathcal{P}_T H, UV^* \rangle + \|\mathcal{P}_{T^\perp} H\|_*,$$

which implies

$$\|\Lambda \circ S\|_1 - \|\Lambda \circ \mathcal{P}_\Omega(\hat{S})\|_1 \ge \langle H, UV^* \rangle + \|\mathcal{P}_{T^\perp} H\|_* + \|\Lambda \circ \mathcal{P}_\Gamma(\hat{S})\|_1,$$

because $\hat{S} = \mathcal{P}_\Omega(\hat{S}) + \mathcal{P}_\Gamma(\hat{S})$. Furthermore,

$$\begin{aligned} \|\Lambda \circ \mathcal{P}_\Omega(\hat{S})\|_1 &= \|\Lambda \circ (S + \mathcal{P}_\Omega(-H))\|_1 \\ &\ge \|\Lambda \circ S\|_1 + \langle \mathrm{sgn}(\Lambda \circ S), \Lambda \circ \mathcal{P}_\Omega(-H) \rangle \\ &= \|\Lambda \circ S\|_1 + \langle \Lambda \circ \mathrm{sgn}(S), -H \rangle . \end{aligned}$$

Combining the last two inequalities and using the fact $\mathcal{P}_\Gamma \hat{S} = \mathcal{P}_\Gamma(\hat{S} - S) = -\mathcal{P}_\Gamma H$, we have

$$\|\mathcal{P}_{T^\perp} H\|_* + \|\Lambda \circ \mathcal{P}_\Gamma H\|_1 \le \langle H, \Lambda \circ \mathrm{sgn}(S) - UV^* \rangle .$$

By introducing a matrix Y which obeys the conditions in the theorem, we can show

$$\begin{aligned}\langle H, \Lambda \circ \mathrm{sgn}(S) - UV^* \rangle &= \langle H, Y + \Lambda \circ \mathrm{sgn}(S) - UV^* \rangle - \langle H, Y \rangle \\ &= \langle \mathcal{P}_T H, \mathcal{P}_T(Y + \Lambda \circ \mathrm{sgn}(S) - UV^*) \rangle + \langle \mathcal{P}_{T^\perp} H, \mathcal{P}_{T^\perp}(Y + \Lambda \circ \mathrm{sgn}(S)) \rangle \\ &\quad - \langle \mathcal{P}_\Gamma H, \mathcal{P}_\Gamma Y \rangle - \langle \mathcal{P}_{\Gamma^c} H, \mathcal{P}_{\Gamma^c} Y \rangle \\ &\leq \frac{\lambda_{\min}}{n^2} \|\mathcal{P}_T H\|_F + \frac{1}{4} \|\mathcal{P}_{T^\perp} H\|_* + \frac{1}{4} \|\Lambda \circ \mathcal{P}_\Gamma H\|_1.\end{aligned}$$

That is

$$\frac{3}{4} \|\mathcal{P}_{T^\perp} H\|_* + \frac{3}{4} \|\Lambda \circ \mathcal{P}_\Gamma H\|_1 \leq \frac{\lambda_{\min}}{n^2} \|\mathcal{P}_T H\|_F.$$

Thus

$$\frac{3}{4} \|\mathcal{P}_{T^\perp} H\|_* + \frac{3\lambda_{\min}}{4} \|\mathcal{P}_\Gamma H\|_1 \leq \frac{\lambda_{\min}}{n^2} \|\mathcal{P}_T H\|_F.$$

Next we can bound $\|\mathcal{P}_T H\|_F$ as follows:

$$\begin{aligned}\|\mathcal{P}_T H\|_F &\leq 2\|\mathcal{P}_T \mathcal{R}_\Gamma \mathcal{P}_T(H)\|_F && \text{(Lemma 1)} \\ &\leq 2\|\mathcal{P}_T \mathcal{R}_\Gamma \mathcal{P}_{T^\perp}(H)\|_F + 2\|\mathcal{P}_T \mathcal{R}_\Gamma(H)\|_F && \text{(triangle inequality)} \\ &\leq \sqrt{\frac{6}{p_0}} \|\mathcal{P}_{T^\perp}(H)\|_F + \sqrt{\frac{6}{p_0}} \|\mathcal{P}_\Gamma(H)\|_F. && \text{(Lemma 2)}\end{aligned}$$

Since $\|\cdot\|_F \leq \|\cdot\|_*$ and $\|\cdot\|_F \leq \|\cdot\|_1$, we have

$$\left(\frac{3}{4} - \frac{\lambda_{\min}}{n^2}\sqrt{\frac{6}{p_0}}\right) \|\mathcal{P}_{T^\perp}(H)\|_F + \left(\frac{3}{4}\lambda_{\min} - \frac{\lambda_{\min}}{n^2}\sqrt{\frac{6}{p_0}}\right) \|\mathcal{P}_\Gamma(H)\|_F \leq 0.$$

Given $p_0 \geq \frac{\log^2 n}{n^4}$, this indicates $\mathcal{P}_{T^\perp} H = \mathcal{P}_\Gamma H = 0$, which implies $\mathcal{P}_\Gamma \mathcal{P}_T(H) = 0$. Since $\mathcal{P}_\Gamma \mathcal{P}_T$ is injective on T, we have $\mathcal{P}_T H = 0$. □

C.2 Dual Certificate Construction

In this part, we adapt the steps in [12], and construct the dual certificate Y only via golfing scheme. Note that $\mathbb{P}((i,j) \in \Gamma) = \alpha_{ij}(1 - \rho_{ij}) := p_{ij}$.

Suppose that $\Gamma = \Gamma_1 \cup \Gamma_2 \cup \cdots \cup \Gamma_l$, and Γ_k's are independent random sets given by

$$\begin{aligned}&\mathbb{P}((i,j) \in \Gamma_1) = \frac{p_{ij}}{6}, \\ &\mathbb{P}((i,j) \in \Gamma_2) = \frac{p_{ij}}{6}, \\ &\mathbb{P}((i,j) \in \Gamma_k) = q_{ij}, \quad \text{for} \quad k = 3, \cdots, l.\end{aligned}$$

Thus if $1 - p_{ij} = (1 - \frac{p_{ij}}{6})^2 (1 - q_{ij})^{l-2}$, the sampling strategies are equivalent. Because of the overlap between $\{\Gamma_k\}$, it is clear that $q_{ij} \geq \frac{2p_{ij}}{3(l-2)}$. For completing the proof, we set $l = \lfloor 5 \log n + 1 \rfloor$.

Construct Y in the following way:

$$Z_0 = \mathcal{P}_T(UV^* - \Lambda \circ \mathrm{sgn}(S)) \tag{44}$$

$$Z_k = (\mathcal{P}_T - \mathcal{P}_T\mathcal{R}_{\Gamma_k}\mathcal{P}_T)Z_{k-1}, \quad for \quad k = 1, \cdots, l \tag{45}$$

$$Y = \sum_{k=1}^{l} \mathcal{R}_{\Gamma_k} Z_{k-1}. \tag{46}$$

We next justify that such a Y satisfy the dual certificate conditions by bounding various norms of Z_0 and showing each iteration (45) reduce these norms at least by half.

We use results of previous lemmas here. Note that we require in the theorem $\sqrt{\alpha_{ij}}(1-\rho_{ij}) \geq C_0\sqrt{\frac{\mu_{ij}r}{n}}\log n$ and $\alpha_{ij} \geq \frac{\log^2 n}{n}$, which imply

$$p_{ij} = \alpha_{ij}(1-\rho_{ij}) \geq C_0\sqrt{\frac{\alpha_{ij}\mu_{ij}r}{n}}\log n \geq C_0\left(\frac{\mu_{ij}r}{n} \vee \frac{\sqrt{\mu_{ij}r}}{n}\right)\log^2 n. \tag{47}$$

Thus by ignoring the constant factor, we have

$$\mathbb{P}((i,j) \in \Gamma_1) = \mathbb{P}((i,j) \in \Gamma_2) \geq C_0\sqrt{\frac{\alpha_{ij}\mu_{ij}r}{n}}\log n, \tag{48}$$

$$\mathbb{P}((i,j) \in \Gamma_k) \geq C_0\sqrt{\frac{\alpha_{ij}\mu_{ij}r}{n}}, \quad \text{for k=3, 4, ..., l,} \tag{49}$$

$$\mathbb{P}((i,j) \in \Gamma_k) \geq C_0\left(\frac{\mu_{ij}r}{n} \vee \frac{\sqrt{\mu_{ij}r}}{n}\right)\log n, \quad \text{for k=1, 2, ..., l.} \tag{50}$$

Then by the result of Lemma 3 and (50), we have

$$\|Z_k\|_F \leq \frac{1}{2}\|Z_{k-1}\|_F, \quad \text{for k=1,2, ..., l.} \tag{51}$$

Recalling the definition of $\|\cdot\|_{w(\infty)}$ in Section 3, by the result of Lemma 4 and (50), we have with high probability

$$\|(I - \mathcal{R}_{\Gamma_k})Z_{k-1}\| \leq \frac{C}{C_0}\|Z_{k-1}\|_{w(\infty)}, \quad \text{for k=1,2,..., l.} \tag{52}$$

By noticing $\sqrt{\alpha_{ij}} \geq C_0\sqrt{\mu_{ij}r/n}\log n$ and (48), we further have

$$\mathbb{P}((i,j) \in \Gamma_1) = \mathbb{P}((i,j) \in \Gamma_2) \geq C_0(\sqrt{\log n})^2\left(\frac{\mu_{ij}r}{n}\log n\right). \tag{53}$$

Thus, by the result of Lemma 5, with high probability we have

$$\|Z_1\|_{w(\infty)} \leq \frac{1}{2\sqrt{\log n}}\|Z_0\|_{w(\infty)}, \tag{54}$$

$$\|Z_2\|_{w(\infty)} \leq \frac{1}{2\sqrt{\log n}}\|Z_1\|_{w(\infty)} \leq \frac{1}{2^2\log n}\|Z_0\|_{w(\infty)}. \tag{55}$$

Moreover, by Lemma 5 and (50), we have

$$\|Z_k\|_{w(\infty)} \leq \frac{1}{2}\|Z_{k-1}\|_{w(\infty)} \leq \frac{1}{2^k\log n}\|Z_0\|_{w(\infty)} \quad \text{for } k = 3, \cdots, l. \tag{56}$$

Next, we bound $\|Z_0\|_F$ and $\|Z_0\|_{w(\infty)}$. Observe that $|(Z_0)_{ab}| \leq |(UV^*)_{ab}| + |[\mathcal{P}_T(\Lambda \circ \mathrm{sgn}(S))]_{ab}|$ and $|(UV^*)_{ab}| = \sqrt{\frac{\mu_{ab}r}{n^2}}$. We need only to bound $|(\mathcal{P}_T\,\mathrm{sgn}(S))_{ab}|$.

Proposition 4. *With the same assumptions on S and Λ as in the problem setup, for any given index (a, b), we have*

$$|[\mathcal{P}_T(\Lambda \circ \operatorname{sgn}(S))]_{ab}| \leq \frac{C}{\sqrt{\log n}} \cdot \frac{\sqrt{\mu_{ab} r}}{n} \tag{57}$$

with high probability for some constant C.

By applying triangle inequality, we have $\|Z_0\|_F \leq \|UV^*\|_F + \|\mathcal{P}_T(\Lambda \circ \operatorname{sgn}(S))\|_F \leq C\sqrt{\mu r}$ and $\|Z_0\|_{w(\infty)} \leq \|UV^*\|_{w(\infty)} + \|\mathcal{P}_T(\Lambda \circ \operatorname{sgn}(S))\|_{w(\infty)} \leq C$, for some constant C.

Next we verify the generated dual certificate Y obey the conditions in Proposition 3. Obviously, $\mathcal{P}_\Omega Y = 0$. Now we prove Y satisfies the following inequalities w.h.p.

$$\|\mathcal{P}_T(Y + \Lambda \circ \operatorname{sgn}(S) - UV^*)\|_F \leq \frac{\lambda_{\min}}{n^2} \tag{58}$$

$$\|\mathcal{P}_{T^\perp} Y\| \leq \frac{1}{8}, \tag{59}$$

$$\|\mathcal{P}_{T^\perp}(\Lambda \circ \operatorname{sgn}(S))\| \leq \frac{1}{8}, \tag{60}$$

$$\|Y \circ \frac{1}{\Lambda}\|_\infty \leq \frac{1}{4}. \tag{61}$$

The first two inequalities can be verified similarly to Section 3.3. Next, we bound $\|\mathcal{P}_{T^\perp}(\Lambda \circ \operatorname{sgn}(S))\| \leq \frac{1}{8}$. It suffices to bound $\|\Lambda \circ \mathcal{P}_\Omega W\|$. Using the way of proving Lemma 4, we first show if $\mathbb{P}((i,j) \in \Omega) = \alpha_{ij}\rho_{ij}$, then

$$\|\Lambda \circ \mathcal{P}_\Omega W - \Lambda \circ \Delta \circ W\| \leq C \frac{\|W\|_\infty}{\sqrt{\log n}}, \tag{62}$$

where $\Lambda = [\frac{1}{\sqrt{\alpha_{ij} n \log n}}]$ and $\Delta = [\alpha_{ij}\rho_{ij}]$. Then by applying the Latala's Theorem in [17], we can bound

$$\|\Delta \circ \Lambda \circ W\| \leq C' \frac{\|W\|_\infty}{\log n}. \tag{63}$$

Latala's Theorem: Let A be a random matrix whose entries a_{ij} are independent centered random variables with finite fourth moment. Then

$$\mathrm{E} s_{\max}(A) \leq C \left[\max_i \left(\sum_j \mathrm{E} a_{ij}^2 \right)^{1/2} + \max_j \left(\sum_i \mathrm{E} a_{ij}^2 \right)^{1/2} + \left(\sum_{i,j} \mathrm{E} a_{ij}^4 \right)^{1/4} \right]$$

Since $\|W\|_\infty = 1$,

$$\|\Lambda \circ \mathcal{P}_\Omega W\| \leq \frac{C}{\sqrt{\log n}} + \frac{C'}{\log n} \leq \frac{1}{8}, \tag{64}$$

provided n sufficiently large.

To show (62), for any matrix Z, we have

$$\Lambda \circ \mathcal{P}_\Omega Z - \Lambda \circ \Delta \circ Z \tag{65}$$
$$= \sum_{i,j} (\delta_{ij} - \Delta_{ij}) \Lambda_{ij} Z_{ij} e_i e_j^* \tag{66}$$
$$:= \sum_{i,j} X_{ij}. \tag{67}$$

X_{ij} are independent zero-mean random matrices. Moreover,

$$\begin{aligned}
\|X_{ij}\| &= \|(\delta_{ij} - \Delta_{ij}) \Lambda_{ij} Z_{ij} e_i e_j^*\| \\
&\leq \frac{\|Z\|_\infty}{\sqrt{n} \log n} \max_{i,j} \left(\frac{\delta_{ij}}{\sqrt{\alpha_{ij}}} - \sqrt{\alpha_{ij}} \rho_{ij} \right) \\
&\leq \frac{\|Z\|_\infty}{\sqrt{n} \log n} \max_{i,j} \frac{1}{\sqrt{\alpha_{ij}}} \\
&\leq \frac{1}{\log^2 n} \|Z\|_\infty,
\end{aligned}$$

where the last inequality is because of the assumption $\alpha_{ij} \geq \frac{\log^2 n}{n}$. And,

$$\begin{aligned}
\left\| \sum_{i,j} \mathrm{E} X_{ij} X_{ij}^* \right\| &= \left\| \sum_{i,j} \mathrm{E} (\delta_{ij} - \Delta_{ij})^2 \Lambda_{ij}^2 Z_{ij}^2 e_i e_i^* \right\| \\
&= \left\| \sum_{i,j} (1 - \Delta_{ij}) \Delta_{ij} \Lambda_{ij}^2 Z_{ij}^2 e_i e_i^* \right\| \\
&\leq \max_i \sum_j (1 - \Delta_{ij}) \Delta_{ij} \Lambda_{ij}^2 Z_{ij}^2 \\
&\leq \frac{\|Z\|_\infty^2}{n \log^2 n} \max_i \sum_j (1 - \alpha_{ij} \rho_{ij}) \rho_{ij} \\
&\leq \frac{\|Z\|_\infty^2}{\log^2 n}.
\end{aligned}$$

By noncommutative Bernstein Inequality, we have

$$\|\Lambda \circ \mathcal{P} Z - \Lambda \circ \Delta \circ Z\| \leq C \left(\sqrt{\frac{\|Z\|_\infty^2}{\log^2 n} \cdot \log n} + \frac{\|Z\|_\infty}{\log^2 n} \cdot \log n \right) \leq C \frac{\|Z\|_\infty}{\sqrt{\log n}}.$$

This finishes the proof of (62).

Finally,

$$\begin{aligned}
\left\|Y \circ \frac{1}{\Lambda}\right\|_\infty &= \left\|\frac{1}{\Lambda} \circ \sum_{k=1}^{l} \mathcal{R}_{\Gamma_k} Z_{k-1}\right\|_\infty \\
&\leq \left\|\frac{1}{\Lambda} \circ \mathcal{R}_{\Gamma_1} Z_0\right\|_\infty + \left\|\frac{1}{\Lambda} \circ \mathcal{R}_{\Gamma_2} Z_1\right\|_\infty + \sum_{k=3}^{l} \left\|\frac{1}{\Lambda} \circ \mathcal{R}_{\Gamma_k} Z_{k-1}\right\|_\infty \\
&\leq \frac{1}{C_0} \|Z_0\|_{w(\infty)} + \frac{1}{C_0} \|Z_1\|_{w(\infty)} + \sum_{k=3}^{l} \frac{\log n}{C_0} \|Z_{k-1}\|_{w(\infty)} \\
&\leq \frac{1}{C_0} \|Z_0\|_{w(\infty)} + \frac{1}{2C_0\sqrt{\log n}} \|Z_0\|_{w(\infty)} + \sum_{k=3}^{l} \frac{\log n}{C_0} \cdot (\frac{1}{2})^{k-1} \frac{1}{\log n} \|Z_0\|_{w(\infty)} \\
&\leq \frac{2}{C_0} \|Z_0\|_{w(\infty)} \\
&\leq \frac{2C}{C_0} \leq \frac{1}{4},
\end{aligned}$$

provided C_0 sufficiently large and the inequality is due to the fact (48)-(50) and (54)-(56).

D Proofs of Key Properties

In this section, we prove the key lemmas provided in Section 3.1 and Proposition 4. The central technique used here is non-communicative Bernstein inequality [20].

D.1 Proof of Lemma 1

For any matrix Z, we can write

$$\begin{aligned}
(\mathcal{P}_T \mathcal{R}_{\Omega_o} \mathcal{P}_T - \mathcal{P}_T)(Z) &= \sum_{ij} \left(\frac{1}{q_{ij}} \delta_{ij} - 1\right) \left\langle e_i e_j^*, \mathcal{P}_T(Z) \right\rangle \mathcal{P}_T(e_i e_j^*) \\
&=: \sum_{ij} \mathscr{A}_{ij}(Z),
\end{aligned}$$

where $\delta_{ij} = \mathbb{I}((i,j) \in \Omega_0)$. With the definition of operator $\mathscr{A}_{ij}$, we can write

$$\begin{aligned}
\mathscr{A}_{ij}^2(Z) &= \left(\frac{1}{q_{ij}} \delta_{ij} - 1\right) \left\langle e_i e_j^*, \mathcal{P}_T(\mathscr{A}_{ij}(Z)) \right\rangle \mathcal{P}_T(e_i e_j^*) \\
&= \left(\frac{1}{q_{ij}} \delta_{ij} - 1\right)^2 \left\langle e_i e_j^*, \mathcal{P}_T(Z) \right\rangle \left\langle e_i e_j^*, \mathcal{P}_T(e_i e_j^*) \right\rangle \mathcal{P}_T(e_i e_j^*).
\end{aligned}$$

Note that $\mathrm{E}(\mathscr{A}_{ij}) = 0$. Next we bound $\|\mathscr{A}_{ij}\|$ and $\left\|\sum_{i,j} \mathrm{E}\mathscr{A}_{ij}^2\right\|$. Since

$$\begin{aligned}
\|\mathscr{A}_{ij}(Z)\|_F &\le \frac{1}{q_{ij}} \left\|\mathcal{P}_T(e_i e_j^*)\right\|_F^2 \|Z\|_F \\
&\le \frac{1}{q_{ij}} \frac{2\mu r}{n} \|Z\|_F \\
&\le \frac{2}{C_0 \log n} \|Z\|_F,
\end{aligned}$$

we know $\|\mathscr{A}_{ij}\| \le 2/(C_0 \log n)$. Furthermore,

$$\begin{aligned}
\left\|\sum_{i,j} \mathrm{E}\mathscr{A}_{ij}^2(Z)\right\|_F &= \left\|\sum_{i,j} \mathrm{E}\left(\frac{1}{q_{ij}}\delta_{ij} - 1\right)^2 \left\langle e_i e_j^*, \mathcal{P}_T(Z)\right\rangle \left\langle e_i e_j^*, \mathcal{P}_T(e_i e_j^*)\right\rangle \mathcal{P}_T(e_i e_j^*)\right\|_F \\
&\le \left\|\sum_{i,j} \mathrm{E}\left(\frac{1}{q_{ij}}\delta_{ij} - 1\right)^2 \left\langle e_i e_j^*, \mathcal{P}_T(Z)\right\rangle \left\langle e_i e_j^*, \mathcal{P}_T(e_i e_j^*)\right\rangle e_i e_j^*\right\|_F \\
&\le \left\|\sum_{i,j} \frac{1}{q_{ij}} \|\mathcal{P}_T(e_i e_j^*)\|_F^2 \left\langle e_i e_j^*, \mathcal{P}_T(Z)\right\rangle e_i e_j^*\right\|_F \\
&\le \frac{2}{C_0 \log n} \left\|\sum_{i,j} \left\langle e_i e_j^*, \mathcal{P}_T(Z)\right\rangle e_i e_j^*\right\|_F \\
&\le \frac{2}{C_0 \log n} \left\|\mathcal{P}_T(Z)\right\|_F,
\end{aligned}$$

thus we have $\left\|\sum_{i,j} \mathrm{E}\mathscr{A}_{ij}^2\right\| \le 2/(C_0 \log n)$. Apply non-commutative Bernstein inequality, with high probability

$$\begin{aligned}
\|\mathcal{P}_T \mathcal{R}_{\Omega_0} \mathcal{P}_T - \mathcal{P}_T\| &= \left\|\sum_{i,j} \mathscr{A}_{ij}\right\| \\
&\le C\sqrt{\frac{2}{C_0 \log n} \log n} + C\frac{2}{C_0 \log n} \log n \\
&\le C\left(\sqrt{\frac{2}{C_0}} + \frac{2}{C_0}\right) \le \frac{1}{2},
\end{aligned}$$

provided that C_0 is sufficiently large.

D.2 Proof of Lemma 2

We note that the condition $\|\mathcal{P}_T - \mathcal{P}_T \mathcal{R}_{\Omega_0} \mathcal{P}_T\| \le \frac{1}{2}$ implies for any matrix Z

$$\frac{1}{2}\|\mathcal{P}_T Z\|_F \le \|\mathcal{P}_T \mathcal{R}_{\Omega_0} \mathcal{P}_T(Z)\|_F \le \frac{3}{2}\|\mathcal{P}_T Z\|_F.$$

Thus, for any matrix Z, we have

$$\begin{aligned}\left\|\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T(Z)\right\|_F^2 &= \langle \mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T(Z), \mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T(Z)\rangle \\ &= \langle Z, (\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T)^*\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T(Z)\rangle \\ &= \langle \mathcal{P}_T(Z), \mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T(Z)\rangle \\ &\le \|\mathcal{P}_T Z\|_F \|\mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T(Z)\|_F \\ &\le \frac{3}{2}\|\mathcal{P}_T Z\|_F^2.\end{aligned}$$

Thus, $\left\|\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T\right\| \le \sqrt{3/2}$ and hence $\left\|\mathcal{P}_T\mathcal{R}_{\Omega_0}^{1/2}\right\| \le \sqrt{3/2}$ because $\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T$ and $\mathcal{P}_T\mathcal{R}_{\Omega_0}^{1/2}$ are adjoint operators. On the other hand, we show $\left\|\mathcal{R}_{\Omega_0}^{1/2}\right\| \le 1/\sqrt{p_0}$. For any matrix Z,

$$\begin{aligned}\left\|\mathcal{R}_{\Omega_0}^{1/2}(Z)\right\|_F^2 &= \left\|\sum_{i,j}\frac{1}{\sqrt{q_{ij}}}\mathbb{I}_{\{(i,j)\in\Omega_0\}}Z_{ij}e_ie_j^*\right\|_F^2 \\ &\le \sum_{i,j}\frac{Z_{ij}^2}{q_{ij}} \le \frac{1}{p_0}\|Z\|_F^2.\end{aligned}$$

Thus, $\|\mathcal{R}_{\Omega_0}\mathcal{P}_T\| \le \left\|\mathcal{R}_{\Omega_0}^{1/2}\right\| \cdot \left\|\mathcal{R}_{\Omega_0}^{1/2}\mathcal{P}_T\right\| \le \sqrt{\frac{3}{2p_0}}$. Thus, $\|\mathcal{P}_T\mathcal{R}_{\Omega_0}\| \le \sqrt{\frac{3}{2p_0}}$.

Since we have $\frac{1}{2}\|\mathcal{P}_T Z\|_F \le \|\mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T(Z)\|_F \le \frac{3}{2}\|\mathcal{P}_T Z\|_F$ for any matrix $Z \in T$, the operator $\mathcal{P}_T\mathcal{R}_{\Omega_0}\mathcal{P}_T$ mapping T onto itself is well conditioned. Thus, $\mathcal{P}_{\Omega_0}\mathcal{P}_T$ is injective on T, i.e., for $Z \in T$, $\mathcal{P}_{\Omega_0}\mathcal{P}_T(Z) = 0$ if and only if $Z = 0$.

D.3 Proof of Lemma 3

This is a direct result of Lemma 1.

D.4 Proof of Lemma 4

Let δ_{ij} denote the Bernoulli random variable $\mathbb{I}((i,j) \in \Omega_0)$. We can derive

$$\begin{aligned}(\mathcal{R}_{\Omega_0} - I)Z &= \sum_{i,j}\left(\frac{1}{q_{ij}}\delta_{ij} - 1\right)\langle e_ie_j^*, Z\rangle e_ie_j^* \\ &=: \sum_{i,j} X_{ij}.\end{aligned}$$

We note that X_{ij} for all $i, j \in [n]$ are zero-mean independent random matrices. Furthermore,

$$\|X_{ij}\| \le \frac{1}{q_{ij}}|Z_{ij}| \le \frac{1}{C_0 \log n}\|Z\|_{w(\infty)}.$$

and

$$\begin{aligned}\left\|\sum_{i,j}\mathrm{E}X_{ij}X_{ij}^*\right\| &= \left\|\sum_{i,j}\mathrm{E}\left(\frac{1}{q_{ij}}\delta_{ij}-1\right)^2 Z_{ij}^2 e_i e_i^*\right\| \\ &= \left\|\sum_{i,j}\left(\frac{1}{q_{ij}}-1\right) Z_{ij}^2 e_i e_i^*\right\| \\ &\le \max_i \sum_j \frac{Z_{ij}^2}{q_{ij}} \\ &\le n\|Z\|_{w(\infty)}^2 \cdot \max_{i,j}\frac{w_{ij}^2}{q_{ij}} \\ &\le \frac{1}{C_0\log n}\|Z\|_{w(\infty)}^2\end{aligned}$$

Similarly, it can be shown that $\|\sum_{i,j}\mathrm{E}X_{ij}^*X_{ij}\| \le \frac{1}{C_0\log n}\|Z\|_{w(\infty)}^2$. Thus, applying the non-commutative Bernstein inequality, we obtain

$$\begin{aligned}\|(\mathcal{R}_{\Omega_0}-I)Z\| &= \left\|\sum_{i,j}X_{ij}\right\| \\ &\le C\left(\sqrt{\frac{1}{C_0\log n}\|Z\|_{w(\infty)}^2\cdot\log n}+\frac{1}{C_0\log n}\|Z\|_{w(\infty)}\cdot\log n\right) \\ &\le \frac{C}{\sqrt{C_0}}\|Z\|_{w(\infty)}\end{aligned}$$

with high probability.

D.5 Proof of Lemma 5

For any entry index pair (a,b), we have

$$\begin{aligned}&[(\mathcal{P}_T\mathcal{R}_{\Omega_0}-\mathcal{P}_T)Z]_{ab}\cdot\frac{1}{w_{ab}} \\ &\quad=\sum_{i,j}\left(\frac{1}{q_{ij}}\delta_{ij}-1\right)Z_{ij}\langle\mathcal{P}_T(e_ie_j^*),e_ae_b^*\rangle\cdot\frac{1}{w_{ab}} \\ &\quad=\sum_{i,j}\left(\frac{1}{q_{ij}}\delta_{ij}-1\right)Z_{ij}\langle e_ie_j^*,\mathcal{P}_T(e_ae_b^*)\rangle\cdot\frac{1}{w_{ab}} \\ &\quad=:\sum_{i,j}x_{ij}.\end{aligned}$$

We note that x_{ij} for $i,j \in [n]$ are independent random variables and $\mathrm{E}x_{ij} = 0$. Furthermore,

$$\begin{aligned}
|x_{ij}| &\le \frac{1}{q_{ij}} |Z_{ij}| \cdot |\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*)\rangle| \cdot \frac{1}{w_{ab}} \\
&\le |Z_{ij}| \cdot \frac{1}{C_0 \beta^{-2} (\frac{\mu_{ij} r}{n}) \log n} \cdot \sqrt{\frac{2\mu_{ij} r}{n}} \cdot \sqrt{\frac{2\mu_{ab} r}{n}} \cdot \frac{1}{\sqrt{\frac{\mu_{ab} r}{n^2}}} \\
&\le \frac{2\beta^2}{C_0 \log n} \frac{|Z_{ij}|}{w_{ij}} \le \frac{2\beta^2}{C_0 \log n} \|Z\|_{w_{(\infty)}},
\end{aligned}$$

and

$$\begin{aligned}
\left| \sum_{i,j} \mathrm{E} x_{ij}^2 \right| &\le \sum_{i,j} \mathrm{E} \left(\frac{1}{q_{ij}} \delta_{ij} - 1 \right)^2 Z_{ij}^2 \cdot |\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*)\rangle|^2 \cdot \frac{1}{w_{ab}^2} \\
&\le \sum_{i,j} \left(\frac{1}{q_{ij}} - 1 \right) \frac{Z_{ij}^2}{w_{ij}^2} \cdot \frac{w_{ij}^2}{w_{ab}^2} \cdot |\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*)\rangle|^2 \\
&\le \frac{1}{C_0 \beta^{-2} (\frac{\log n}{n})} \cdot \|Z\|_{w_{(\infty)}}^2 \cdot \frac{1}{\mu_{ab} r} \|\mathcal{P}_T(e_a e_b^*)\|_F^2 \\
&\le \frac{2\beta^2}{C_0 \log n} \cdot \|Z\|_{w_{(\infty)}}^2,
\end{aligned}$$

where we use the assumption $q_{ij} \ge C_0 \beta^{-2} (\frac{\mu_{ij} r}{n}) \log n$ and the fact $\|\mathcal{P}_T(e_a e_b^*)\|_F^2 \le \frac{2\mu_{ab} r}{n}$.

Thus, applying the non-commutative Bernstein inequality, we have

$$\begin{aligned}
\left| \sum_{i,j} x_{ij} \right| &\le C \left(\sqrt{\frac{2\beta^2}{C_0 \log n} \|Z\|_{w_{(\infty)}}^2 \cdot \log n} + \frac{2\beta^2}{C_0 \log n} \|Z\|_{w_{(\infty)}} \cdot \log n \right) \\
&= C \left(\sqrt{\frac{2}{C_0}} \beta + \frac{2}{C_0} \beta^2 \right) \|Z\|_{w_{(\infty)}} \\
&\le \frac{1}{2} \beta \|Z\|_{w_{(\infty)}},
\end{aligned}$$

with high probability, provided that C_0 is sufficiently large.

D.6 Proof of Proposition 1

The proof of Proposition 1 is obtained by taking $\Lambda_{ij} = 1/(\sqrt{n} \log n)$ in Proposition 4.

D.7 Proof of Proposition 4

$$\begin{aligned}
\langle e_a e_b^*, \mathcal{P}_T(\Lambda \circ \mathrm{sgn}(S))\rangle &= \langle \Lambda \circ \mathrm{sgn}(S), \mathcal{P}_T(e_a e_b^*)\rangle \\
&= \sum_{i,j} \frac{\delta_{ij}}{\sqrt{\alpha_{ij} n} \log n} \left\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*)\right\rangle \\
&=: \sum_{i,j} x_{ij}
\end{aligned}$$

where

$$\delta_{ij} = \begin{cases} 1 & \text{with prob.} \quad \alpha_{ij}\rho_{ij}/2 \\ 0 & \text{with prob.} \quad 1-\alpha_{ij}\rho_{ij} \\ -1 & \text{with prob.} \quad \alpha_{ij}\rho_{ij}/2. \end{cases}$$

Thus $\{x_{ij}\}$ are independent random variables and $\mathrm{E}x_{ij} = 0$. Furthermore,

$$|x_{ij}| \le \left| \frac{\delta_{ij}}{\sqrt{\alpha_{ij}n}\log n} \left\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*) \right\rangle \right| \le \frac{1}{\sqrt{\alpha_{ij}n}\log n} \sqrt{\frac{2\mu_{ij}r}{n}} \cdot \sqrt{\frac{2\mu_{ab}r}{n}} \le \frac{2\sqrt{\mu_{ab}r}}{C_0 n \log^2 n}$$

where the last inequality is due to the assumption $\sqrt{\alpha_{ij}} \ge C_0\sqrt{\mu_{ij}r/n}\log n$, and

$$\begin{aligned} \left| \sum_{i,j} \mathrm{E}x_{ij}^2 \right| &= \left| \sum_{i,j} \frac{\mathrm{E}\delta_{ij}^2}{\alpha_{ij} n \log^2 n} \left\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*) \right\rangle^2 \right| \\ &= \left| \sum_{i,j} \frac{\rho_{ij}}{n \log^2 n} \left\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*) \right\rangle^2 \right| \\ &\le \frac{1}{n \log^2 n} \left| \sum_{i,j} \left\langle e_i e_j^*, \mathcal{P}_T(e_a e_b^*) \right\rangle^2 \right| \\ &= \frac{1}{n \log^2 n} \|\mathcal{P}_T(e_a e_b^*)\|_F^2 \\ &\le \frac{2\mu_{ab}r}{n^2 \log^2 n}. \end{aligned}$$

Then apply non commutative Bernstein inequality, and we have

$$\begin{aligned} \left| \sum_{i,j} x_{ij} \right| &\le C \left(\sqrt{\frac{2\mu_{ab}r}{n^2 \log^2 n} \cdot \log n} + \frac{2\sqrt{\mu_{ab}r}}{C_0 n \log^2 n} \cdot \log n \right) \\ &\le \frac{C}{\sqrt{\log n}} \frac{\sqrt{\mu_{ab}r}}{n}. \end{aligned}$$