Robust Principal Component Analysis with Side Information

Kai-Yang Chiang[⋆] KYCHIANG@CS.UTEXAS.EDU
Cho-Jui Hsieh[†] CHOHSIEH@UCDAVIS.EDU
Inderjit S. Dhillon[⋆] INDERJIT@CS.UTEXAS.EDU

[⋆]Department of Computer Science, The University of Texas at Austin, Austin, TX 78712, USA
[†] Department of Statistics and Computer Science, University of California at Davis, Davis, CA 95616, USA

Abstract

The robust principal component analysis (robust PCA) problem has been considered in many machine learning applications, where the goal is to decompose the data matrix to a low rank part plus a sparse residual. While current approaches are developed by only considering the low rank plus sparse structure, in many applications, side information of row and/or column entities may also be given, and it is still unclear to what extent could such information help robust PCA. Thus, in this paper, we study the problem of robust PCA with side information, where both prior structure and features of entities are exploited for recovery. We propose a convex problem to incorporate side information in robust PCA and show that the low rank matrix can be exactly recovered via the proposed method under certain conditions. In particular, our guarantee suggests that a substantial amount of low rank matrices, which cannot be recovered by standard robust PCA, become recoverable by our proposed method. The result theoretically justifies the effectiveness of features in robust PCA. In addition, we conduct synthetic experiments as well as a real application on noisy image classification to show that our method also improves the performance in practice by exploiting side information.

1. Introduction

Robust principal component analysis (robust PCA) has received much attention in recent studies for its ability to recover the low rank model from sparse noise. Such sparse structure of noise is common in many real applications such as image processing and bioinformatics (Wright et al., 2009). Formally, assuming that the given observation $R \in \mathbb{R}^{n\times n}$ is in the form of:

$$R = L_0 + S_0,$$

where L_0 is a low rank matrix and S_0 is a sparse noise matrix with unknown support and magnitude, the goal of robust PCA is to recover L_0 given R. One state-of-the-art approach is to decompose R into a low rank and a sparse component via a simple convex program. This approach has been shown to be advantageous for several reasons. First, it overcomes the weakness of standard PCA where the solution could be extremely skewed even if a single entry is corrupted, and second, exact recovery of L_0 can be guaranteed under certain assumptions (Chandrasekaran et al., 2011; Candès et al., 2011).

Despite these strengths, one criticism of robust PCA is that it disregards *side information*, or *features*, in the recovery process even if it is provided. For example, imagine that data matrix R represents the gene-disease association where few entries are corrupted due to some contamination in experiments, and furthermore, some additional features of each gene (e.g. its gene-expression profile) and each disease (e.g. co-occurrence of diseases for a patient) are also given in advance. Then, instead of simply applying robust PCA to filter out the noise, one would expect to incorporate this side information in order to better recover the clean association. To the best of our knowledge, however, it is still unclear as to how much side information could help the recoverability of robust PCA.

With the above motivation, in this paper, we consider the problem of robust PCA with side information, where the goal is to recover the underlying matrix by utilizing both the "low rank plus sparse structure" and additional feature information. Our approach is to link feature information to the underlying matrix via an implicit bilinear function, which leads us to solve the problem via a feature-embedded convex program. Furthermore, to justify the usefulness of features theoretically, we show that under certain assumptions, exact recovery could be attained by the proposed method if the rank of underlying low rank matrix is $O(n^2/(d \log n \log d))$, with up to $O(n^2)$ corrupted entries,

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

where d is the dimensionality of features. Compared to the result of standard robust PCA (Candès et al., 2011) where the recovery could be guaranteed if rank is $O(n/(\log n)^2)$, our result shows that the boundary of rank could be significantly improved by the proposed method if $d \ll n$. In addition, we conduct several synthetic experiments and an application on noisy image classification to show that the proposed method achieves better performance than standard robust PCA. Our results thus conclude that side information is indeed useful in the robust PCA in both theoretical and practical aspects. Our contribution can be summarized as follows.

- We propose a feature-embedded objective for robust PCA which learns the underlying low rank matrix using both prior structure and side information simultaneously.
- We provide an exact recovery guarantee of our model under certain conditions. As a consequence, our result asymptotically improves the guarantee of standard robust PCA by taking side information into account.
- Experimental results show that the proposed method improves the performance on both synthetic datasets and a real application on noisy image classification.

Connection to prior work. Robust PCA is one of the most prominent examples to demonstrate the power of convex programs in matrix recovery. Researchers have investigated several approaches for providing theoretical guarantees of robust PCA (Wright et al., 2009; Chandrasekaran et al., 2011; Candès et al., 2011). Perhaps the most remarkable milestone is the strong guarantee provided by Candès et al. (2011) (see Corollary 1 for details). Compared to existing work where the recoverability completely relies on the separability between the low rank and sparse structure, our problem setting is more general since it aims to exploit the strength of both structure and side information for recovery, and therefore an improved guarantee could be derived with the aid of features. We will discuss the improvement in detail in Section 3.

On the other hand, side information has been shown to be useful in several related problems such as matrix completion (Jain & Dhillon, 2013; Chiang et al., 2015) and compressed sensing (Mota et al., 2014). In particular, since robust PCA shares several similarities with the matrix completion problem [1], it may appear a positive sign for the effect of side information in robust PCA. However, the robust PCA problem is still essentially different—in fact harder—from matrix completion. In matrix completion, it is often assumed that the observed entries are generated from a low rank matrix without substantial noise (Candès & Tao, 2010), while in robust PCA the observations are corrupted with unknown support and magnitude. This fundamental difference can be illustrated by comparing this work with inductive matrix completion and its extension (Jain & Dhillon, 2013; Chiang et al., 2015), in which features are used to improve matrix completion. In Jain & Dhillon (2013), the observed entries are generated from XHY^T where X, Y are row/column features. This is essentially a special case of our formulation (3). In Chiang et al. (2015), the observed entries are assumed to be $XHY^T + M$, where M is a low-rank part of the underlying matrix that cannot be explained by features. In this paper, we assume the model is $XHY^T + S$, where S represents the sparse noise. Also, we provide an exact recovery guarantee while Chiang et al. (2015) only show that the expected error decays as $n \to \infty$.

Our model also shares certain similarity with Low-Rank Representation (LRR) (Liu et al., 2010), which assumes that the clean data could be represented by a linear combination of a given dictionary. Interestingly, LRR could be thought of as a special case of our model where the dictionary is like one of our features X (or Y). Our model is more general as we incorporate both row and column features to help the recovery.

Organization of the paper. In section 2, we state the setup and assumptions of our problem. We then present the exact recovery guarantee and an algorithm in Section 3. In Section 4, we provide an overview of the proof and a follow-up discussion. We then show experimental results in Section 5 and state conclusions in Section 6.

2. Problem Setup

Let $L_0 \in \mathbb{R}^{n_1 \times n_2}$ be the underlying model matrix, where $\text{rank}(L_0) = r \ll \min(n_1, n_2)$ so L_0 is low rank. Let $S_0 \in \mathbb{R}^{n_1 \times n_2}$ be the sparse noise matrix whose support set Ω is unknown and values could be arbitrary. $R = L_0 + S_0$ will be the noisy data we observe in practice. In addition, let $X \in \mathbb{R}^{n_1 \times d_1}$, $Y \in \mathbb{R}^{n_2 \times d_2}$ be the feature matrix where $\mathbf{x}_i$ ($\mathbf{y}_i$) denotes the feature of i-th row (column) entity. Without loss of generality, we assume both X, Y are orthogonal. [2] For simplicity, throughout the analysis we will consider the case $n_1 = n_2 = n$ and $d_1 = d_2 = d$.

2.1. Robust PCA with Features

We begin with the standard setting of robust PCA which aims to recover the underlying matrix without using any feature information. The most popular approach in previous studies (Chandrasekaran et al., 2011; Candès et al., 2011) is to consider a matrix separation objective, in which the matrix R is decomposed to a low rank term L and a sparse term S, whose structures are forced by minimizing nuclear norm and ℓ-1 norm respectively. Specifically, they

[1] For example, both robust PCA and matrix completion try to recover a low rank matrix from imperfect observations.

[2] In practice, one could conduct QR factorization to orthogonize the given feature sets.

consider the following Principal Component Pursuit (PCP) objective (Candès et al., 2011):

$$\min_{L,S} \|L\|_* + \lambda\|S\|_1 \quad \text{s.t. } L + S = R, \tag{1}$$

where $\|L\|_* := \sum_{i=1}^{\text{rank}(L)} \sigma_i$ is the nuclear norm of L, and $\|S\|_1 := \sum_{i,j} |S_{ij}|$ is the elementwise one norm of S. Remarkable theoretical foundations have also been established for PCP beyond heuristics, which will be discussed in more detail in Section 3.

However, as stated in the motivation, side information may also be given in real-world problems. The question is thus how to incorporate useful feature information in robust PCA so that one could possibly learn L_0 more effectively. A natural approach is to assume that features X and Y reveal clean information via a bilinear mapping ϕ, i.e.

$$L_{ij} = \phi(\mathbf{x}_i, \mathbf{y}_j) = \mathbf{x}_i^T H \mathbf{y}_j \tag{2}$$

with some unknown ϕ (or equivalently H) that aims to be learned. Such a bilinear form is commonly considered for incorporating side information in recent matrix recovery literature, e.g. Jain & Dhillon (2013); Xu et al. (2013); Zhong et al. (2015), and it also enjoys several properties such as low model complexity in some settings. Therefore, by linking feature information using (2), we propose to solve the following convex objective PCP with Features (PCPF in brief):

$$\min_{H,S} \|H\|_* + \lambda\|S\|_1 \quad \text{s.t. } XHY^T + S = R. \tag{3}$$

Let (H^*, S^*) be the optimal solution of the problem (3), and the low rank matrix will be recovered by $L^* = XH^*Y^T$. Note that the proposed PCPF is more general beyond standard robust PCA, as X and/or Y could be set as identity when (one of) features are absent, and the problem reduces to standard PCP when both $X = Y = I$.

2.2. Assumptions

Observant readers may already notice that it is not always possible to recover the underlying low rank model from sparse noise even with the aid of side information in (3). Certain assumptions have to be made in order to make the problem well-posed.

Feasibility Condition. First of all, since PCPF aims to recover the low rank matrix L_0 by learning a matrix H_0 such that $XH_0Y^T = L_0$, a modest necessary condition is that the solution H_0 has to be *feasible*. In PCPF, the following condition has to be provided for feasibility:

$$\text{col}(X) \supseteq \text{col}(L_0), \quad \text{col}(Y) \supseteq \text{col}(L_0^T), \tag{4}$$

where $\text{col}(X)$ represents the column space of X. Such a condition is standard for matrix recovery with bilinear models, see e.g. Xu et al. (2013); Yi et al. (2013). Intuitively, the condition suggests that feature matrices X and Y have to be correlated to the underlying true low rank space (i.e. they have to be truly informative), so that one could utilize hidden information in X and Y by seeking a matrix spanned jointly by $\text{col}(X)$ and $\text{col}(Y)$.

Incoherence Condition. Even if the feasibility condition holds, recovery can still be naturally hard due to an identifiability issue. For example, consider the case where XH_0Y^T is also "sparse", then one cannot identify whether the solution XH_0Y^T is produced by sparse noise or not. Typically, an incoherence assumption on the underlying low rank space has to be made in order to avoid sparse solutions. In this work, we extend the incoherence condition to the given feature sets X, Y in the following sense. Let $H_0 = U\Sigma V^T$ be the reduced SVD of H_0. We assume that the feature matrix is incoherent w.r.t. the matrix H_0:

$$\max_i \|U^T\mathbf{x}_i\|_2 \le \sqrt{\frac{\mu_0 r}{n}}, \max_j \|V^T\mathbf{y}_j\|_2 \le \sqrt{\frac{\mu_0 r}{n}}, \tag{5}$$

$$\max_{i,j} |\mathbf{x}_i^T UV^T \mathbf{y}_j| \le \frac{\sqrt{\mu_0 r}}{n}. \tag{6}$$

Also, the feature matrices X, Y are self-incoherent as:

$$\max_i \|\mathbf{x}_i\|_2 \le \sqrt{\frac{\mu_1 d}{n}}, \quad \max_j \|\mathbf{y}_j\|_2 \le \sqrt{\frac{\mu_1 d}{n}}. \tag{7}$$

Incoherence conditions are quite standard in matrix recovery literature (e.g. Candès & Tao (2010); Candès & Recht (2012)). Intuitively, such conditions imply that a matrix cannot be too spiky, eliminating the possibility of underlying matrix being sparse in our problem.

On the other hand, we shall also avoid the case where the underlying sparse noise matrix is also low rank. This could happen when the noise appears only in few columns or rows of S_0. To avoid such cases, we assume that noise S_0 appears uniformly at random.

Finally, though both assumptions are presented for the analysis of exact recovery (stated in Theorem 1), it should be noted that in real applications, our algorithm may achieve good performance even when these conditions are not satisfied. We will see an example in Section 5.

3. Main Results and Algorithm

The core question we focus on is to what extent side information is able to help the recovery of robust PCA in theory. As noted, previous theoretical results have shown that PCP could surprisingly recover a large class of matrices given only limited information (Chandrasekaran et al., 2011; Candès et al., 2011). Roughly speaking, the main reason of such success is because the low rank and sparse subspace are naturally distinguishable under incoherence assumptions, which makes separation become possible even without any hint on how the subspace looks like. Upon this realization, one may doubt the effect of features since such information seems to be redundant. However, we have

found that side information is in fact powerful as it makes a broader class of low rank matrices become recoverable. The following main Theorem states the result:

Theorem 1 (Main result: Exact Recovery of PCPF)**.** *Let $L_0 \in \mathbb{R}^{n\times n}$ be a low rank matrix with rank r. Let S_0 be an arbitrary sparse matrix with cardinality m whose support set Ω is distributed uniformly at random but location is unknown. Suppose we are given orthogonal features $X, Y \in \mathbb{R}^{n\times d}$ which satisfy the feasibility and incoherence conditions (4) $\sim$ (7). Then there exists universal constants $\rho_r, \rho_s > 0$, such that if:*

$$r \le \rho_r(\mu_0\mu_1)^{-1}n^2(d\log n\log d)^{-1},$$
$$m \le \rho_s n^2,$$

then with probability at least $1 - O(d^{-10})$, the solution (H^*, S^*) of the convex problem (3) with $\lambda = 1/\sqrt{n}$ exactly recovers the underlying low rank matrix in the sense that $XH^*Y^T = L_0$ and $S^* = S_0$.

Several interesting results could be further inferred from Theorem 1. First, the Theorem indicates that the recoverability depends not only on the rank of underlying matrix r and sparsity ρ_s but also on the feature dimension d. In addition, lower d yields a better rank boundary. The reasoning behind the Theorem is quite intuitive: when d is small, H has much lower degree of freedom compared with L and thus is much easier to recover. Another interesting fact is that in the special case where $X = Y = I$, problem (3) reduces to the standard PCP and moreover, the guarantee of Theorem 1 coincides with the guarantee of PCP provided by Candès et al. (2011):

Corollary 1 (Exact Recovery of PCP)**.** *Suppose $X = Y = I$ and L_0, S_0 all follow the same assumptions of Theorem 1. Then with high probability, the solution (H^*, S^*) of PCPF with $\lambda = 1/\sqrt{n}$ is exact in the sense that $H^* = L_0$ and $S^* = S_0$, provided that*

$$r \le \rho_r{\mu_0}^{-1}n(\log n)^{-2}, \quad m \le \rho_s n^2.$$

More generally, Theorem 1 suggests that the rank boundary is on the same order of PCP when $d = O(n)$ *as the worse case*. However, for informative features, it is expected that $d \ll n$ since it should reveal the low rank structure of L_0, in which case Theorem 1 suggests that a substantial improvement of rank boundary of L_0 could be made. For example, if d is on the order of r, the rank could be approximately up to the order of $O(n/\sqrt{\log n})$, which significantly increases the rank constraint of low rank matrices (As an instance, for modern full-HD images where $n \approx 2000$, $n/(\log n)^2$ is on the order of 30, while $n/\sqrt{\log n}$ is on the order of 700). Therefore, Theorem 1 shows that the effect of features in robust PCA problem can be significant because it asymptotically improves the boundary of rank constraint, making a larger class of matrices to be recoverable.

Algorithm 1 ALM method for PCPF

Input: Observation R, feature X, Y, max iteration $t_{\max}$
$\lambda \leftarrow 1/\sqrt{n}, \quad \mu \leftarrow 1/\|R\|, \quad t \leftarrow 0$
$H \leftarrow 0, \quad S \leftarrow 0$
while not converged **and** $t < t_{\max}$ **do**
  $M \leftarrow \mathcal{D}_{\mu^{-1}}(R - S + \mu^{-1}Z)$
  $H \leftarrow X^T M Y$
  $S \leftarrow \mathcal{S}_{\lambda\mu^{-1}}(R - XHY^T + \mu^{-1}Z)$
  $Z \leftarrow Z + \mu(R - S - XHY^T)$
  $t \leftarrow t + 1, \quad \mu \leftarrow \mu/0.95$
end while
return $L^* \leftarrow XH^*Y^T$

Finally, as a remark, there is no parameter tuning required for λ in PCPF, since Theorem 1 proves that $\lambda = 1/\sqrt{n}$ always succeeds. This advantage is inherited from the result of PCP on top of uniformly random sparse noise, and detailed discussions can be found in Candès et al. (2011).

3.1. Solving PCPF Objective

Many algorithms have been proposed to solve the PCP objective (1), which is convex but non-smooth as it includes both ℓ-1 and nuclear norm regularization, e.g. SDP (Chandrasekaran et al., 2011), APG (Lin et al., 2009) and ALM (Yuan & Yang, 2009; Lin et al., 2010). Among many of them, ALM method is shown to be competitive for its stability and fast convergence in empirical studies (Candès et al., 2011). Thus, we consider to extend the ALM method to PCPF objective (3). ALM converts the equality constraint to a soft penalty term and a Lagrangian term, resulting in the following function $L(H, S, Z)$:

$$L(H,S,Z) = \|H\|_* + \lambda\|S\|_1 + \langle Z, R - S - XHY^T\rangle + \frac{\mu}{2}\|R - S - XMY^T\|_F^2.$$

It then iteratively updates H, S, Z until converged. In each iteration, ALM first updates variables H, S by alternatively solving single variable minimization problems, $\min_H L(H, S, Z)$ and $\min_S L(H, S, Z)$. It then updates the Lagrange multiplier Z by $Z + \mu(R - S - XHY^T)$.

We now briefly state how to solve each subproblem of updating H and S. Let $\mathcal{S}_x(M) := \text{sgn}(M)\circ\max(|M| - x, 0)$ be the soft thresholding operator on elements of M, where $\circ$ denotes the elementwise product. Similarly, let $\mathcal{D}_x(M)$ be the thresholding operator on singular values of M, i.e. $\mathcal{D}_x(M) := U_M S_x(\Sigma_M)V_M^T$ where $U_M\Sigma_M V_M^T$ is the SVD of M. Then, solving for H given fixed S, Z is equivalent to solving the following problem:

$$\min_H \|H\|_* + \frac{\mu}{2}\|M' - XHY^T\|_F^2$$

where $M' = R - S + \mu^{-1}Z$, and the solution is thus given by $X^T\mathcal{D}_{\mu^{-1}}(M')Y$. On the other hand, the update rule for S given fixed H, Z can be written as $\mathcal{S}_{\lambda\mu^{-1}}(R - XHY^T +$

$\mu^{-1}Z$). This means that each subproblem can be efficiently solved by a simple closed form solution.

Finally, as a heuristic, we also apply the continuation technique described in Lin et al. (2009) for faster convergence, where we set $\mu = 1/\|R\|$ at the beginning and increase μ a bit for each iteration. The convergence criterion is set to be $\|R - S - XHY^T\|_F/\|R\|_F < 10^{-7}$ as suggested in Candès et al. (2011). Our algorithm can be summarized as Algorithm 1 and will be used in our experiments.

4. The Sketch of Proof

We now give a high-level overview of the proof for Theorem 1. The roadmap of the proof consists of two main steps. The first step is to provide a sufficient condition that guarantees the optimal solution is exact if certain "dual certificates" exist. The second step is then showing that under conditions of Theorem 1, a valid set of dual certificates can be constructed with high probability, which concludes the proof. Such proof technique is popular in matrix recovery literatures (Candès & Tao, 2010; Candès et al., 2011; Xu et al., 2013). While our proof structure is mainly built on Candès et al. (2011), the proof is fundamentally different as side information X, Y now plays an important role. We will revisit the differences in the end of this section.

4.1. Reduction of Sampling Model

First, note that in Theorem 1 S_0 is assumed to be sampled from the set $\{\Omega \mid |\Omega| = m\}$ uniformly at random. However, in the proof we consider the support of S_0 to be sampled via Bernoulli model instead, i.e. $(i, j) \in \Omega$ with probability ρ, and $|\Omega| = m$ in expectation when $\rho = m/n^2$. Standard analysis showed that these two models are equivalent (see Candès & Tao (2010)), and thus guarantees on Bernoulli model will also hold on the uniform model.

Another useful sampling reduction lemma (first introduced by Candès et al. (2011)) is to further reduce the signs of nonzero entries (i, j) from fixed to Bernoulli random. Specifically, in Theorem 1, the values of S_0 is fixed, and therefore $\text{sgn}(S_0)$ is also fixed. However, it turns out to be easier to consider the model where each nonzero of S_0 is an independent symmetric Bernoulli variable that takes ± 1 with equal probability. The following theorem shows that proving recovery on this "random sign model" is sufficient.

Theorem 2. *Let L_0, X and Y be the model and feature matrices which satisfy conditions in Theorem 1, and S_0 is supported on Ω where $\Omega \sim Ber(2\rho_s)$. In addition, suppose the sign of each nonzero S_0 takes ± 1 with equal probability (and independent to its location). Then, if PCPF recovers L_0 from such S_0 with high probability, with at least the same probability PCPF will also recover the model where signs of S_0 are fixed and location of $S_0 \sim Ber(\rho_s)$.*

This theorem facilitates the analysis since we could now focus on random sign model, without being worried by arbitrary values that may appear in sparse noise.

4.2. Dual Certification

We now introduce our proposed dual certification, which is a sufficient condition for the solution of (3) to be exact. We first define some linear operators and projections. Recall $H_0 = U\Sigma V^T$ is the reduced SVD of H_0. Let the space T to be defined as:

$$T := \{UA^T + BV^T \mid A, B \in \mathbb{R}^{d\times r}\},$$

and $\mathcal{P}_T$ is the orthogonal projection onto T. Similarly, we extend the definition of Ω for the set representation, where Ω denotes the set of $n \times n$ matrices with the same support as S_0, and $\mathcal{P}_\Omega$ is the orthogonal projection onto Ω. We also define the following linear transformation:

$$\mathcal{T}_S(A) := X^T AY, \quad A \in \mathbb{R}^{n\times n}$$
$$\mathcal{T}_L(B) := XBY^T, \quad B \in \mathbb{R}^{d\times d}$$

which maps $n \times n$ matrices to $d \times d$ and vice versa. Note that since X and Y are orthogonal, $\mathcal{T}_S\mathcal{T}_L = \mathcal{I}$. Finally, we define the space Q as:

$$Q := \{XX^T AYY^T \mid A \in \mathbb{R}^{n\times n}\}.$$

The orthogonal projection $\mathcal{P}_Q$ onto Q is simply $\mathcal{T}_L\mathcal{T}_S$.

With these definitions, now we can present our dual certification lemma:

Lemma 1 (Dual Certification)**.** *Suppose $\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\| \le 1/2$ and $\lambda < 1$. Then, (H_0, S_0) is the unique solution of problem (3) if there exists W, F, M and D such that:*

$$\mathcal{T}_L(UV^T + W) + M = \lambda(sgn(S_0) + F + \mathcal{P}_\Omega D),$$

where $W \in T^\perp$, $\|W\| \le \frac{1}{2}$, $M \in Q^\perp$, $F \in \Omega^\perp$, $\|F\|_\infty \le \frac{1}{2}$, and $\|\mathcal{P}_\Omega D\|_F \le \frac{1}{4}$.

Therefore, from Lemma 1, it is sufficient to prove that the pair (H_0, S_0) is the unique solution of (3) by providing dual certificates (W, M) obeying:

$$\begin{cases} W \in T^\perp \\ M \in Q^\perp \\ \|W\| \le \frac{1}{2} \\ \|\mathcal{P}_\Omega(\mathcal{T}_L(UV^T + W) + M) - \lambda\text{sgn}(S_0)\|_F \le \frac{\lambda}{4} \\ \|\mathcal{P}_{\Omega^\perp}(\mathcal{T}_L(UV^T + W) + M)\|_\infty \le \frac{\lambda}{2} \end{cases}$$

4.3. Construction of Dual Certificates

Our proposed dual certificates are independently constructed from two parts: One is constructed using golfing scheme to handle low rank part and the other is constructed using inverse of operator to handle sparse noise part. Golfing scheme (Gross, 2011) is a clever technique to construct dual certificates in many recovery proofs. The

idea is as follows. Consider the noise set $\Omega \sim Ber(\rho)$, or equivalently $\Omega^C \sim Ber(1-\rho)$. The complement set Ω^C can also be viewed as being jointly sampled from j_0 i.i.d. Bernoulli procedures, each of which follows $Ber(q)$. Formally, $\Omega^C = \bigcup_{j=1}^{j_0} \Omega_j$, $\Omega_j \sim Ber(q)$ if:

$$(1-q)^{j_0} = \rho. \tag{8}$$

Therefore, a certificate can be additively constructed upon each Ω_j. Via such procedure, the norm of constructed certificates will exponentially decrease step by step for each Ω_j, and thus, it will be useful for proving that the constructed certificate has a small magnitude in certain norm.

Certificates of low rank part. Fix $j_0 \geq \lceil 2\log n \rceil$. Let $\Omega^C = \bigcup_{j=1}^{j_0} \Omega_j$, $\Omega_j \sim Ber(q)$ i.i.d. for each Ω_j where q satisfies (8). Let $Y_j, Z_j \in \mathbb{R}^{d\times d}$, $Y_0 = 0$, and define Y_j, Z_j recursively as:

$$\begin{aligned} Y_j &= Y_{j-1} + q^{-1}\mathcal{T}_S\mathcal{P}_{\Omega_j}\mathcal{T}_L(Z_j), \\ Z_j &= UV^T - \mathcal{P}_T Y_j. \end{aligned}$$

We then set

$$\begin{aligned} W^L &= \mathcal{P}_{T^\perp} Y_{j_0}, \\ M^L &= \mathcal{P}_{Q^\perp} \sum_j q^{-1}\mathcal{P}_{\Omega_j}\mathcal{T}_L Z_{j-1}. \end{aligned}$$

Certificates of sparse noise part. Again, assume $\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\| \leq 1/2$, then $\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega\| \leq 1/4$ and thus the operator $\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega$ mapping $\Omega \to \Omega$ is invertible. We then set:

$$\begin{aligned} W^S &= \lambda\mathcal{P}_{T^\perp}\mathcal{T}_S(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^{-1}\mathrm{sgn}(S_0), \\ M^S &= \lambda\mathcal{P}_{Q^\perp}(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^{-1}\mathrm{sgn}(S_0). \end{aligned}$$

From above construction, we propose to produce dual certificates $W = W^L + W^S$ and $M = M^L + M^S$. Note that each part consists of two components, which is different from certificates in PCP provided by Candès et al. (2011).

4.4. Proving Validity of Dual Certificates

Obviously, $W \in T^\perp$ and $M \in Q^\perp$ by construction. Furthermore, observe that:

$$\begin{aligned} &\mathcal{P}_\Omega\mathcal{T}_L W^S + \mathcal{P}_\Omega M^S \\ =&\lambda\mathcal{P}_\Omega\mathcal{T}_L(\mathcal{I} - \mathcal{P}_T)\mathcal{T}_S(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^{-1}\mathrm{sgn}(S_0) \\ &+ \lambda\mathcal{P}_\Omega\mathcal{P}_{Q^\perp}(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^{-1}\mathrm{sgn}(S_0) \\ =&\lambda\mathcal{P}_\Omega(\mathcal{I} - \mathcal{T}_L\mathcal{P}_T\mathcal{T}_S)(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^{-1}\mathrm{sgn}(S_0) \\ =&\lambda\mathrm{sgn}(S_0). \end{aligned}$$

Therefore, it is clear that (W, M) is a pair of dual certificates if W^L, W^S, M^L and M^S obey:

$$\begin{cases} \|W^L + W^S\| \leq \frac{1}{2} \\ \|\mathcal{P}_\Omega\mathcal{T}_L(UV^T + W^L) + \mathcal{P}_\Omega M^L\|_F \leq \frac{\lambda}{4} \\ \|\mathcal{P}_{\Omega^\perp}\mathcal{T}_L(UV^T + W^L + W^S) + \mathcal{P}_{\Omega^\perp}(M^S + M^L)\|_\infty \leq \frac{\lambda}{2} \end{cases}$$

under the condition $\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\| \leq 1/2$. However, we can further prove that such condition will naturally hold with large probability (see Lemma 5 in Appendix for details). Therefore, the Theorem can be concluded by proving the following two lemmas.

Lemma 2 (Validity of Certificates of Low Rank Part)**.** *Let $\Omega \sim Ber(\rho)$ where $0 < \rho \leq \rho_s$, and $j_0 = \lceil 2\log n \rceil$. Then under the conditions of Theorem 1, W^L and M^L obey:*

a. $\|W^L\| \leq \frac{1}{4}$,
b. $\|\mathcal{P}_\Omega\mathcal{T}_L(UV^T + W^L) + \mathcal{P}_\Omega M^L\|_F \leq \frac{\lambda}{4}$
c. $\|\mathcal{P}_{\Omega^\perp}\mathcal{T}_L(UV^T + W^L) + \mathcal{P}_{\Omega^\perp}M^L\|_\infty \leq \frac{\lambda}{4}$.

Lemma 3 (Validity of Certificates of Sparse Noise Part)**.** *Suppose Ω is sampled via Bernoulli model $Ber(\rho)$. Assume the sign of each nonzero in S_0 is i.i.d. symmetric whose randomness is independent to the location. Then under conditions of Theorem 1, W^S and M^S obey:*

a. $\|W^S\| \leq \frac{1}{4}$,
b. $\|\mathcal{P}_{\Omega^\perp}\mathcal{T}_L W^S + \mathcal{P}_{\Omega^\perp}M^S\|_\infty \leq \frac{\lambda}{4}$.

Detailed proofs of these lemmas are provided in Appendix.

4.5. Discussions

Although our proof structure is based on the proof of PCP (Candès et al., 2011), the proof is essentially different as the side information comes in. We now highlight some high-level differences between our analysis and previous analysis on standard robust PCA.

The first major difference comes from the dual certification Lemma 1, in which we introduce a crucial term $M \in Q^\perp$. Compared to the previous dual certification lemmas (Chandrasekaran et al., 2011; Candès et al., 2011), the term M absorbs the part outside the feature space and enables us to build a bounded certificate W under a smaller $d \times d$ space in the later proof. However, to deal with the additional term M, we have to carefully develop a more sophisticated set of certificates W^L, W^S, M^L and M^S. This is different from Candès et al. (2011) in which breaking the certificate into W^L and W^S is sufficient.

Another major difference comes from the different dimensions of low rank space T and sparse support Ω. While the technique of handling this issue varies under different contexts, a major approach used in many steps in the proof is to apply linear transformation $\mathcal{T}_L$ and $\mathcal{T}_S$ to resolve the mismatch of dimensionality. However, such modification is far from trivial for at least two reasons. First, many arguments in the previous proof become implausible under that modification (e.g. key lemmas 4 and 6 for Golfing scheme), and second, some steps may even become incorrect by directly converting dimensions using $\mathcal{T}_S$ and $\mathcal{T}_L$, in which case alternative arguments are required. For instance, in Candès et al. (2011), a key property that makes low rank and sparse components distinguishable is to show that $\Omega \cap T = \{0\}$

where both Ω and T are a set of $n \times n$ matrices. However, in our analysis where matrices in T are $d \times d$, showing $\mathcal{T}_S(\Omega) \cap T = \{0\}$ becomes invalid since with large probability $\mathcal{T}_S(\Omega)$ will be rank d, which must span a non-trivial subspace of T. Thus, we have to prove a key lemma 5 to show that the opposite argument (i.e. $\Omega \cap \mathcal{T}_L(T) = \{0\}$) holds with high probability instead.

Finally, we emphasize that the above complication is a blessing rather than a curse to the result. In some sense, by considering more sophisticated dual certificates, The bounded norm requirement of W becomes easier to satisfy because its dimension is reduced from n to d (see Lemma 1). As a consequence, valid certificates become producible for higher rank matrices, and thus the rank boundary of Theorem 1 is improved.

5. Experimental Results

We now conduct experiments to show that side information is indeed useful in the robust PCA problem. In synthetic experiments, we show that PCPF is able to recover a set of low rank matrices which cannot be recovered by PCP as stated in Theorem 1. We then consider an application on noisy image classification. We will see that by incorporating features, images denoised by PCPF will be classified more accurately. The parameters λ in both PCP and PCPF are set as $1/\sqrt{n}$ by default as Theorem 1 suggested.

Synthetic experiments. We first examine the effect of features in robust PCA on synthetic datasets. We create a true low rank matrix $L_0 = UV^T$, where $U, V \in \mathbb{R}^{n\times r}$, $U_{ij}, V_{ij} \sim N(0, 1/n)$ with $n = 200$ and different r. We also generate a sparse error matrix S_0 whose support follows $Ber(\rho_s)$, with its values determined by either random sign model (i.e. each non-zero value takes ± 1 with equal probability) or coherent sign model (where $S_0 = \mathcal{P}_\Omega(\text{sgn}(L_0))$). In addition, we generate feature matrices $X, Y \in \mathbb{R}^{n\times d}$, $d = r + 10$ that both satisfy (4). We then take $R = L_0 + S_0$ as the observation, and input R to PCP [3] and R, X and Y to PCPF. We regard the recovery to be successful if the output low rank matrix L^* obeys:

$$\frac{\|L^* - L_0\|_F}{\|L_0\|_F} < 10^{-4}. \tag{9}$$

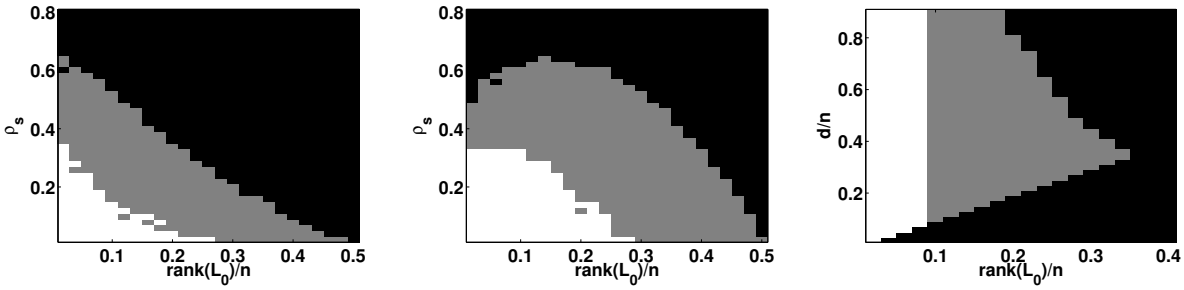
We first consider the recoverability of PCP and PCPF while varying rank of L_0 (r) and sparsity of S_0 (ρ_s) under both random and coherent sign model. For each pair of (r, ρ_s), we create three random problems, and deem the recovery of an algorithm to be attained if it successfully recovers all problems. We then mark the grid to be white if recovery is attained by both PCP and PCPF and black if both fail. We also observe that in several cases recovery cannot be attained by PCP, but can be attained by PCPF, and these grid points are marked as gray. The results are shown in Figure 1a and 1b. First, we see that for both PCP and PCPF, the recovery results under random or coherent sign model are of the same order. This supports the argument in Theorem 1 that only the location of support (and not signs) matters for both algorithms to succeed in the recovery. More importantly, there exists a substantial gray region where matrices in such region could be recovered only by PCPF. The result shows that PCPF is more effective as it recovers a larger class of matrices by leveraging feature information.

Furthermore, Theorem 1 suggests that the improvement of PCPF is also determined by the feature dimension d. To conduct a supporting experiment, we consider the same construction of L_0, and create S_0 under random sign model with $\rho_s = 0.2$. For each choice of rank(L_0), we construct several sets of features $X, Y \in \mathbb{R}^{n\times d}$ satisfying (4) with different d by varying d from r to n and applying PCP and PCPF to each (d, r). The results are shown in Figure 1c. We again observe that there exists a substantial gray region where matrices are not able to be recovered by PCP because of higher rank, but become recoverable by PCPF given feature information. Moreover, recovery of higher rank matrices could be achieved with a smaller d. The result matches the discussion in Section 3 that higher rank matrices would be recovered with a smaller d, and it also empirically supports Theorem 1.

Application: multiclass classification on noisy images. One application of robust PCA is (sparse) noise removal for images. In the problem, we are given a set of noisy yet correlated images, and the noise is known to be sparse. Since the underlying clean images are correlated and thus share an implicit low rank structure, standard robust PCA could be used to identify sparse noise. However, in certain cases, low-dimensional features of images may also be available from other sources. For example, suppose the set of images are human faces, then the principal components of general human faces—known as Eigenface (Turk & Pentland, 1991)—could be used as features, and such features could be helpful in the denoising process.

Motivated by the above realization, here we conduct an experiment on multiclass classification on a set of noisy images. We consider the digit recognition dataset MNIST, which includes $50,000$ training images and $10,000$ testing images, and each image is a handwriting digit described as a 784 dimensional vector. We first take the training image set to produce "Eigendigit" features $X \in \mathbb{R}^{784\times d}$ where $d = 300$. We then take testing image set to generate noisy images. Precisely, let $L_0 \in \mathbb{R}^{784\times 10000}$ be the set of (clean) testing images, and S_0 be the sparse noise matrix in which

[3] We use the ALM solver available at http://perception.csl.illinois.edu/matrix-rank/sample_code.html for PCP, which is an implementation of Lin et al. (2010).



(a) Random sign (b) Coherent sign (c) Rank versus feature dimension

Figure 1. Synthetic experiments on recovery of PCP and PCPF under different rank(L_0), sparsity of noise ρ_s and feature dimension d. Both algorithms succeed in recovery in white region and fail in black region. However, there exists a substantial region marked as gray where PCP fails yet PCPF succeeds to recover these low rank matrices, justifying the usefulness of features.

ρ_s	0.05	0.1	0.2	0.3
PCP	0.3643	0.3781	0.3913	0.4265
PCPF	**0.3424**	**0.3444**	**0.3607**	**0.4036**

Table 1. Relative error between clean images L_0 and recovered images L^*. Images recovered by PCPF achieve smaller relative error than PCP under various different sparsity of noise ρ_s.

ρ_s of entries are randomly picked to be corrupted (by setting the value to be 255). Then given a set of noisy images $R = \min(L_0 + S_0, 255)$ and Eigendigit features X, the goal is to denoise the noisy images for classification.

We again compare PCP and PCPF for noise removal in this experiment. For PCP, we directly input R to derive a denoised L^*_{pcp}. For PCPF, we take Eigendigit features X as row features in objective (3) and set $Y = I$ as there are no column features given in this problem. The denoised image from PCPF is given by $L^*_{pcpf} = XH^*$. Both L^*_{pcp} and L^*_{pcpf} will be low rank approximations of the clean image set. Note that though X will no longer satisfy (4), it could be used in PCPF in practice since X is still expected to contain much information on how does the low rank approximation of clean digits looks like. [4]

We compare the quality of denoised solutions from PCP and PCPF using two metrics. First, we could again directly evaluate the relative error between ground-truth images L_0 and denoised images L^* (equation (9)). The error is reported in Table 1. As shown, PCPF consistently achieves lower relative error than PCP under different ρ_s from 0.05 to 0.3, showing that numerically the low rank approximation derived from PCPF is closer to the ground-truth L_0.

To further justify the quality of denoised images in terms of practical metrics in real application, we consider the classification accuracy achieved by denoised images as the second metric. We pre-train both multiclass linear and kernel SVM classifiers on 50000 clean training images to predict the digit from input vector space using LIBLINEAR (Fan et al., 2008) and LIBSVM (Chang & Lin, 2011). We then use these trained classifiers to classify the denoised images from PCP and PCPF. The results are reported in Table 2. The column "Clean" denotes the accuracy on clean testing images (i.e. L_0), and the column "Noisy" denotes the accuracy on noisy images (i.e. R) without any denoising process. The last two columns are accuracies on denoised images from PCP and PCPF respectively. Both methods are somehow effective for denoising sparse noise, since accuracy achieved by denoised images are much closer to the clean images. Furthermore, PCPF consistently achieves better accuracies than PCP, showing that incorporating side information as in PCPF is indeed helpful in denoising process in real-world applications.

Classification with trained linear SVM classifiers

ρ_s	Clean	Noisy	PCP	PCPF
0.05	91.96	77.93	86.94	**87.51**
0.1		59.63	86.33	**87.88**
0.2		38.16	85.94	**87.48**
0.3		25.63	78.52	**79.84**

Classification with trained kernel SVM classifiers

ρ_s	Clean	Noisy	PCP	PCPF
0.05	98.33	66.14	95.17	**95.74**
0.1		18.47	94.85	**95.89**
0.2		10.32	94.55	**95.48**
0.3		10.32	87.00	**87.78**

Table 2. Classification accuracy of denoised images on linear and kernel SVM under various sparsity of noise ρ_s. The column Clean shows the accuracy on L_0, and the column Noisy shows the accuracy on R. Denoised images from both PCP and PCPF achieve much higher accuracy than noisy images, and PCPF further outperforms PCP by utilizing Eigendigit features.

6. Conclusions

We propose a convex problem that incorporates side information to robust PCA. An improved exact recovery guarantee of the proposed method is provided, and the advantage of side information is discussed. The theoretical im-

[4] Rigorously speaking, the ground-truth image L_0 is not low rank, but only approximately low rank.

provement is further empirically supported by several experiments. These results conclude the usefulness of side information in robust PCA in both theory and practice.

Acknowledgement

This research was supported by NSF grants CCF-1320746 and IIS-1546459.

References

Candès, Emmanuel and Recht, Benjamin. Exact matrix completion via convex optimization. *Commun. ACM*, 55 (6):111–119, 2012.

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

Candès, Emmanuel J., Li, Xiaodong, Ma, Yi, and Wright, John. Robust principal component analysis? *Journal of ACM*, pp. 11:1–11:37, 2011.

Chandrasekaran, Venkat, Sanghavi, Sujay, Parrilo, Pablo A., and Willsky, Alan S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2), 2011.

Chang, Chih-Chung and Lin, Chih-Jen. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Chiang, Kai-Yang, Hsieh, Cho-Jui, and Dhillon, Inderjit S. Matrix completion with noisy side information. In *Advances in Neural Information Processing Systems (NIPS)*, 2015.

Fan, Rong-En, Chang, Kai-Wei, Hsieh, Cho-Jui, Wang, Xiang-Rui, and Lin, Chih-Jen. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

Gross, D. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theor.*, 57(3):1548–1566, 2011.

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

Lin, Z., Ganesh, A., Wright, J., Wu, L., Chen, M., and Ma, Y. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. In *Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP)*, 2009.

Lin, Zhouchen, Chen, Minming, and Ma, Yi. The augmented lagrange multiplier method for exact recovery of a corrupted low-rank matrices. *Mathematical Programming*, 2010.

Liu, Guangcan, Lin, Zhouchen, and Yu, Yong. Robust subspace segmentation by low-rank representation. In *ICML*, 2010.

Mota, Joao F. C., Deligiannis, Nikos, and Rodrigues, Miguel R. D. Compressed sensing with side information: Geometrical interpretation and performance bounds. In *IEEE Global Conf. Sig. and Inf. Proc*, pp. 675 – 679, 2014.

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

Turk, Matthew and Pentland, Alex. Eigenfaces for recognition. *J. Cognitive Neuroscience*, 3(1):71–86, 1991.

Vershynin, Roman. Introduction to the non-asymptotic analysis of random matrices. *CoRR abs/1011.3027*, 2010.

Wright, J., Ganesh, A., Rao, S., Peng, Y., and Ma, Y. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in Neural Information Processing Systems (NIPS)*, 2009.

Xu, Miao, Jin, Rong, and Zhou, Zhi-Hua. Speedup matrix completion with side information: Application to multi-label learning. In *NIPS*, 2013.

Yi, J., Zhang, L., Jin, R., Qian, Q., and Jain, A. K. Semi-supervised clustering by input pattern assisted pairwise similarity matrix completion. In *ICML*, 2013.

Yuan, X. and Yang, J. Sparse and low-rank matrix decomposition via alternating direction methods. *preprint*, 2009.

Zhong, K., Jain, P., and Dhillon, I. S. Efficient matrix sensing using rank-1 gaussian measurements. In *International Conference on Algorithmic Learning Theory(ALT)*, 2015.

7. Appendix

7.1. Preliminaries

We first revisit some basic properties of defined linear operators and projections. Recall that $H_0 = U\Sigma V^T$ is the reduced SVD of H_0, and the space T is defined as:

$$T := \{UA^T + BV^T \mid A, B \in \mathbb{R}^{d\times r}\},$$

and $\mathcal{P}_T$ is the orthogonal projection onto T. It is known that any subgradient of $\|H_0\|_*$ has the form $UV^T + W$, where $\mathcal{P}_T W = 0, \|W\| \le 1$.

Similarly, we have defined Ω to be the set of matrices whose entries supported as the same as S_0, and $\mathcal{P}_\Omega$ is the orthogonal projection onto Ω. It is also known that any subgradient of $\|S_0\|_1$ takes the form $\text{sgn}(S_0) + F$, where $\mathcal{P}_\Omega F = 0, \|F\|_\infty \le 1$.

Under the incoherence assumptions, we also introduce a norm inequality on rank-1 matrices which we will use frequently in the proof. Given any matrix with the form $\mathbf{x}_i\mathbf{y}_j^T \in \mathbb{R}^{d\times d}$, we have

$$\begin{aligned}\|\mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T)\|_F^2 &= \langle \mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T), \mathbf{x}_i\mathbf{y}_j^T\rangle \\ &\le \|U^T\mathbf{x}_i\|_2^2\|\mathbf{y}_j\|_2^2 + \|V^T\mathbf{y}_j\|_2^2\|\mathbf{x}_i\|_2^2 \\ &\le \frac{2\mu_0\mu_1 rd}{n^2}. \end{aligned} \tag{10}$$

In particular, if we let p be any probability that satisfies:

$$p \ge 2C\epsilon^{-2}\frac{\mu_0\mu_1 rd\log d}{n^2} \tag{11}$$

with a numerical constant $C > 0$, then the inequality becomes:

$$\|\mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T)\|_F^2 \le \epsilon^2\frac{p}{C\log d}. \tag{12}$$

7.2. Proof of Lemma 1

Here, we provide a proof of dual certification lemma (Lemma 1).

Proof. Consider any feasible perturbation $(H_0 + \Delta, S_0 - X\Delta Y^T)$ from the claimed optimum. We will prove the lemma by showing that such perturbed pair increases the objective (3) unless $\Delta = 0$. Let $UV^T + W_0$ be any subgradient of $\|H_0\|_*$ and $\text{sgn}(S_0) + F_0$ be any subgradient of $\|S_0\|_1$, then by the definition of subgradient, $W_0 \in T^\perp, \|W_0\|_2 \le 1, F_0 \in \Omega^\perp, \|F_0\|_\infty \le 1$, and

$$\begin{aligned}&\|H_0 + \Delta\|_* + \lambda\|S_0 - X\Delta Y^T\|_1 \\ \ge&\|H_0\|_* + \lambda\|S_0\|_1 + \langle UV^T + W_0, \Delta\rangle \\ &- \lambda\langle \text{sgn}(S_0) + F_0, X\Delta Y^T\rangle.\end{aligned}$$

Select W_0 and F_0 such that $\langle W_0, \Delta\rangle = \|\mathcal{P}_{T^\perp}\Delta\|_*$ and $\langle F_0, X\Delta Y^T\rangle = -\|\mathcal{P}_{\Omega^\perp}(X\Delta Y^T)\|_1$ [5], then we have:

$$\begin{aligned}&\|H_0 + \Delta\|_* + \lambda\|S_0 - X\Delta Y^T\|_1 \\ \ge&\|H_0\|_* + \lambda\|S_0\|_1 + \|\mathcal{P}_{T^\perp}\Delta\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}(X\Delta Y^T)\|_1 \\ &+ \langle UV^T, \Delta\rangle - \lambda\langle \text{sgn}(S_0), X\Delta Y^T\rangle.\end{aligned} \tag{13}$$

Now, since $\langle UV^T, \Delta\rangle = \langle \mathcal{T}_L(UV^T), \mathcal{T}_L\Delta\rangle$, we can bound the inner product terms by:

$$\begin{aligned}&|\langle UV^T, \Delta\rangle - \lambda\langle \text{sgn}(S_0), X\Delta Y^T\rangle| \\ =&|\langle \mathcal{T}_L(UV^T) - \lambda\text{sgn}(S_0), X\Delta Y^T\rangle| \\ \le&|\langle \mathcal{T}_L(W), X\Delta Y^T\rangle| + |\langle M, X\Delta Y^T\rangle| \\ &+ |\lambda\langle F, X\Delta Y^T\rangle| + |\lambda\langle \mathcal{P}_\Omega D, X\Delta Y^T\rangle| \\ \le&\frac{1}{2}\|\mathcal{P}_{T^\perp}\Delta\|_* + \frac{\lambda}{2}\|\mathcal{P}_{\Omega^\perp}X\Delta Y^T\|_1 + \frac{\lambda}{4}\|\mathcal{P}_\Omega X\Delta Y^T\|_F,\end{aligned}$$

where in the third inequality we use the fact that $\langle M, X\Delta Y^T\rangle = \langle M, \mathcal{P}_Q(X\Delta Y^T)\rangle = 0$. Thus, equation (13) can be reduced to:

$$\begin{aligned}&\|H_0 + \Delta\|_* + \lambda\|S_0 - X\Delta Y^T\|_1 \\ \ge&\|H_0\|_* + \lambda\|S_0\|_1 + \frac{1}{2}(\|\mathcal{P}_{T^\perp}\Delta\|_* + \lambda\|\mathcal{P}_{\Omega^\perp}(X\Delta Y^T)\|_1) \\ &- \frac{\lambda}{4}\|\mathcal{P}_\Omega X\Delta Y^T\|_F.\end{aligned} \tag{14}$$

We can further bound the term $\|\mathcal{P}_\Omega X\Delta Y^T\|_F$ by:

$$\begin{aligned}&\|\mathcal{P}_\Omega X\Delta Y^T\|_F \\ \le&\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\Delta\|_F + \|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_{T^\perp}\Delta\|_F \\ \le&\frac{1}{2}\|\Delta\|_F + \|\mathcal{P}_{T^\perp}\Delta\|_F \\ \le&\frac{1}{2}(\|\mathcal{P}_\Omega\mathcal{T}_L\Delta\|_F + \|\mathcal{P}_{\Omega^\perp}\mathcal{T}_L\Delta\|_F) + \|\mathcal{P}_{T^\perp}\Delta\|_F.\end{aligned}$$

By definition, $\mathcal{P}_\Omega\mathcal{T}_L\Delta = \mathcal{P}_\Omega = X\Delta Y^T$, so

$$\begin{aligned}\|\mathcal{P}_\Omega\mathcal{T}_L\Delta\|_F &\le \|\mathcal{P}_{\Omega^\perp}\mathcal{T}_L\Delta\|_F + 2\|\mathcal{P}_{T^\perp}\Delta\|_F \\ &\le \|\mathcal{P}_{\Omega^\perp}\mathcal{T}_L\Delta\|_1 + 2\|\mathcal{P}_{T^\perp}\Delta\|_*.\end{aligned}$$

Therefore, equation (14) becomes:

$$\begin{aligned}&\|H_0 + \Delta\|_* + \lambda\|S_0 - X\Delta Y^T\|_1 \\ \ge&\|H_0\|_* + \lambda\|S_0\|_1 \\ &+ \frac{1}{2}\left((1-\lambda)\|\mathcal{P}_{T^\perp}\Delta\|_* + \frac{\lambda}{2}\|\mathcal{P}_{\Omega^\perp}(X\Delta Y^T)\|_1\right).\end{aligned} \tag{15}$$

However, by assumption, $\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\| \le \frac{1}{2} < 1$ implies that $\mathcal{T}_L(T)\cap\Omega = \{0\}$. Therefore, for any $\Delta \ne 0$, if $\Delta \notin T$ then $\|\mathcal{P}_{T^\perp}\Delta\| > 0$, and if $\Delta \in T$ then $\|\mathcal{P}_{\Omega^\perp}\mathcal{T}_L\Delta\|_1 > 0$. Thus the LHS of (15) will be strictly larger than RHS unless $\Delta = 0$, which concludes the proof. □

[5] Such W_0 and F_0 exist. See Candès et al. (2011) for an example of such matrices.

7.3. Preliminary Lemmas

We need several lemmas to prove the validity of constructed dual certificates introduced in Section 4.3. For following lemmas, when we say the equation holds with large probability, we mean that the event will hold with probability at least $1 - O(d^{-10})$.

Most of the probability bounds in our results are from the Bernstein inequality stated as below.

Proposition 1 (Noncommutative Matrix Bernstein Inequality (Recht, 2011)). *Let $X_1 \cdots X_k$ be k independent, zero-mean random matrices where each $X_i \in \mathbb{R}^{n_1 \times n_2}$. Suppose for each X_i, $\|X_i\| \leq R$, and the norm of the sum of covariance matrices is bounded by:*

$$\max\left\{\|\sum_{i=1}^{k}\mathbb{E}\left[X_i X_i^T\right]\|, \|\sum_{i=1}^{k}\mathbb{E}\left[X_i^T X_i\right]\|\right\} \leq \sigma^2.$$

Then for any $t > 0$:

$$\Pr(\|\sum_{i=1}^{k} X_i\| > t) \leq (n_1 + n_2)\exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right).$$

We begin with a core lemma which generalizes the result of Theorem 4.1 in Candès & Recht (2012).

Lemma 4. *Suppose $\Omega_0 \sim Ber(\rho)$. Then with large probability,*

$$\|\mathcal{P}_T - \rho^{-1}\mathcal{P}_T\mathcal{T}_S\mathcal{P}_{\Omega_0}\mathcal{T}_L\mathcal{P}_T\| \leq \epsilon$$

provided that $\rho \geq C_0\epsilon^{-2}(2\mu_0\mu_1 rd\log d)/n^2$ with some constant $C_0 > 0$.

Proof. First we decompose the matrix $(\mathcal{P}_T - \rho^{-1}\mathcal{P}_T\mathcal{T}_S\mathcal{P}_{\Omega_0}\mathcal{T}_L\mathcal{P}_T)Z$ as:

$$\begin{aligned}&(\mathcal{P}_T - \rho^{-1}\mathcal{P}_T\mathcal{T}_S\mathcal{P}_{\Omega_0}\mathcal{T}_L\mathcal{P}_T)Z\\ =&(\mathcal{P}_T\mathcal{T}_S(\mathcal{I} - \rho^{-1}\mathcal{P}_{\Omega_0})\mathcal{T}_L\mathcal{P}_T)Z\\ =&\sum_{(i,j)}(1-\rho^{-1}\delta_{ij})\langle Z, \mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T)\rangle\mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T).\end{aligned}$$

This yields us to define a linear operator $\mathcal{S}_{ij}$ as:

$$\mathcal{S}_{ij}(Z) = (1-\rho^{-1}\delta_{ij})\langle Z, \mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T)\rangle\mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T),$$

which maps any $Z \in \mathbb{R}^{d\times d}$ to $\mathbb{R}^{d\times d}$. The operator is symmetric, zero in expectation (i.e., $\mathbb{E}\left[\mathcal{S}_{ij}(Z)\right] = 0$) and its operator norm, by definition, is bounded by:

$$\sup_{Z\neq 0}\frac{\|\mathcal{S}_{ij}(Z)\|_F}{\|Z\|_F}.$$

Thus, the original operator $\mathcal{P}_T - \rho^{-1}\mathcal{P}_T\mathcal{T}_S\mathcal{P}_{\Omega_0}\mathcal{T}_L\mathcal{P}_T$ can be viewed as a sum of independent, zero-mean operators $\mathcal{S}_{ij}$, where each operator has a bounded operator norm as:

$$\begin{aligned}\|\mathcal{S}_{ij}(Z)\|_F &\leq \rho^{-1}|\langle Z, \mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T)\rangle|\|\mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T)\|_F\\ &\leq \rho^{-1}\|\mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T)\|_F^2\|Z\|_F\\ &\leq \frac{\epsilon^2}{C_0\log d}\|Z\|_F,\end{aligned}$$

where the last line is derived by applying (12). Also, we can bound the quantity $\|\sum_{(i,j)}\mathbb{E}\left[S_{ij}^2\right]\|$ similarly. Since

$$\begin{aligned}&\|\sum_{(i,j)}\mathbb{E}\left[\mathcal{S}_{ij}^2(Z)\right]\|_F\\ =&\Big\|\sum_{(i,j)}\mathbb{E}\left[(1-\rho^{-1}\delta_{ij})^2\right]\langle Z, \mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T)\rangle\\ &\qquad\|\mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T)\|_F^2\mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T)\Big\|_F,\end{aligned}$$

and $\mathbb{E}\left[(1-\rho^{-1}\delta_{ij})^2\right] = (1-\rho)/\rho \leq 1/\rho$, therefore,

$$\begin{aligned}&\|\sum_{(i,j)}\mathbb{E}\left[\mathcal{S}_{ij}^2(Z)\right]\|_F\\ \leq&\frac{\epsilon^2}{C_0\log d}\|\mathcal{P}_T\sum_{(i,j)}\langle\mathcal{P}_T Z, \mathbf{x}_i\mathbf{y}_j^T\rangle\mathbf{x}_i\mathbf{y}_j^T\|_F\\ =&\frac{\epsilon^2}{C_0\log d}\|\mathcal{P}_T\mathcal{T}_S\mathcal{T}_L\mathcal{P}_T(Z)\|_F\\ \leq&\frac{\epsilon^2}{C_0\log d}\|Z\|_F\end{aligned}$$

With above bounds, the claim follows by applying matrix Bernstein inequality. □

An important fact from this lemma is that it implies $\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\|$ will not be too large provided that $|\Omega|$ is not extremely large. More formally, we can prove the following Lemma:

Lemma 5. *Suppose $\Omega \sim Ber(\rho)$ where $1 - \rho \geq C_0\epsilon^{-2}(2\mu_0\mu_1 rd\log d)/n^2$. Then with high probability, we have $\|\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega\| \leq \sqrt{\rho+\epsilon}$.*

Proof. Suppose $1 - \rho \geq C_0\epsilon^{-2}(2\mu_0\mu_1 rd\log d)/n^2$, then from Lemma 4, we know that with high probability,

$$\|\mathcal{P}_T - (1-\rho)^{-1}\mathcal{P}_T\mathcal{T}_S\mathcal{P}_{\Omega^\perp}\mathcal{T}_L\mathcal{P}_T\| \leq \epsilon.$$

Now, by the fact that $\mathcal{P}_{\Omega^\perp} = \mathcal{I} - \mathcal{P}_\Omega$, we can rewrite the operator as:

$$\begin{aligned}&\mathcal{P}_T - (1-\rho)^{-1}\mathcal{P}_T\mathcal{T}_S\mathcal{P}_{\Omega^\perp}\mathcal{T}_L\mathcal{P}_T\\ =&(1-\rho)^{-1}(\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T - \rho\mathcal{P}_T),\end{aligned}$$

from which we can conclude that

$$\|\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\| \leq \epsilon(1-\rho) + \rho\|\mathcal{P}_T\| = \rho + \epsilon(1-\rho)$$

by the triangle inequality. The claim is thus proved by the fact that $\|\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\| \le \|\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega\|^2$. □

Lemma 4 implies that if $Z \in T$, then its Frobenius norm will decrease sufficiently large after applying the operator $\mathcal{I} - \mathcal{P}_T\mathcal{T}_S\mathcal{P}_{\Omega_0}\mathcal{T}_L$. The next lemma says that, after applying such operator, its "$\mathcal{T}_L$ infinity norm" will also decrease sufficiently large.

Lemma 6. *Suppose $\Omega_0 \sim Ber(\rho)$ and $Z \in T$. Then with large probability,*

$$\|\mathcal{T}_L(Z - \rho^{-1}\mathcal{P}_T\mathcal{T}_S\mathcal{P}_{\Omega_0}\mathcal{T}_L Z)\|_\infty \le \epsilon\|\mathcal{T}_L Z\|_\infty$$

provided that $\rho \ge C_0\epsilon^{-2}(2\mu_0\mu_1 rd\log d)/n^2$ with some constant $C_0 > 0$.

Proof. Let $K = \mathcal{T}_L(Z - \rho^{-1}\mathcal{P}_T\mathcal{T}_S\mathcal{P}_{\Omega_0}\mathcal{T}_L Z)$. Observe that any element K_{ab} can be represented as a sum of independent variables, i.e. $K_{ab} = \sum_{(i,j)} s_{ij}$, where s_{ij} is defined as:

$$s_{ij} = (1 - \rho^{-1}\delta_{ij})\langle Z, \mathbf{x}_i\mathbf{y}_j^T\rangle\langle\mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T), \mathbf{x}_a\mathbf{y}_b^T\rangle.$$

Again, each s_{ij} has zero mean ($\mathbb{E}\big[s_{ij}\big] = 0$), and each $|s_{ij}|$ can be bounded by:

$$\begin{aligned}|s_{ij}| &\le \rho^{-1}|\mathbf{x}_i^T Z\mathbf{y}_j|\|\mathcal{P}_T(\mathbf{x}_i\mathbf{y}_j^T)\|_F\|\mathcal{P}_T(\mathbf{x}_a\mathbf{y}_b^T)\|_F \\ &\le \frac{\epsilon^2}{C_0\log d}\|\mathcal{T}_L Z\|_\infty.\end{aligned}$$

Also, $|\sum_{(i,j)}\mathbb{E}\big[s_{ij}^2\big]|$ can be bounded by:

$$\begin{aligned}|\sum_{(i,j)}\mathbb{E}\big[s_{ij}^2\big]| &\le |\sum_{(i,j)}\rho^{-1}(\mathbf{x}_i^T Z\mathbf{y}_j)^2\langle\mathbf{x}_i\mathbf{y}_j^T, \mathcal{P}_T(\mathbf{x}_a\mathbf{y}_b^T)\rangle^2| \\ &\le \rho^{-1}\|\mathcal{T}_L Z\|_\infty^2|\sum_{(i,j)}\langle\mathbf{x}_i\mathbf{y}_j^T, \mathcal{P}_T(\mathbf{x}_a\mathbf{y}_b^T)\rangle^2| \\ &\le \rho^{-1}\|\mathcal{T}_L Z\|_\infty^2\|\mathcal{T}_L\mathcal{P}_T(\mathbf{x}_a\mathbf{y}_b^T)\|_F^2 \\ &\le \frac{\epsilon^2}{C_0\log d}\|\mathcal{T}_L Z\|_\infty^2.\end{aligned}$$

Note that in both bounds we apply the inequality (12) because ρ obeys (11). Therefore, by Bernstein inequality, we have:

$$\Pr(|K_{ab}| > \epsilon\|\mathcal{T}_L Z\|_\infty) \le 2\exp\bigg(-\frac{3}{8}C_0\log d\bigg),$$

and the claim is proved by applying an union bound. □

Lemma 7. *For any fixed matrix $Z \in \mathbb{R}^{d\times d}$, with large probability,*

$$\|(I - \rho^{-1}\mathcal{T}_S\mathcal{P}_{\Omega_0}\mathcal{T}_L)Z\| \le C_0'\sqrt{\frac{d\log d}{\rho}}\|\mathcal{T}_L Z\|_\infty$$

with some constant $C_0' > 0$, provided that $\rho \ge C_0\mu_1^2 d\log d/n^2$ with some constant $C_0 > 0$.

Proof. Again we can decompose the matrix $(I - \rho^{-1}\mathcal{T}_S\mathcal{P}_{\Omega_0}\mathcal{T}_L)Z$ as $\sum_{(i,j)} S_{ij}$, where S_{ij} is defined as:

$$S_{ij} = (1 - \rho^{-1}\delta_{ij})\langle Z, \mathbf{x}_i\mathbf{y}_j^T\rangle\mathbf{x}_i\mathbf{y}_j^T.$$

Each S_{ij} is independent with zero means (i.e. $\mathbb{E}\big[S_{ij}\big] = 0$). Furthermore, we can bound $\|S_{ij}\|$ by:

$$\|S_{ij}\| \le \rho^{-1}|\mathbf{x}_i^T Z\mathbf{y}_j|\|\mathbf{x}_i\|_2\|\mathbf{y}_j\|_2 \le \rho^{-1}\frac{\mu_1 d}{n}\|\mathcal{T}_L Z\|_\infty,$$

and the term $\|\sum_{(i,j)}\mathbb{E}\big[S_{ij}^T S_{ij}\big]\|$ can be bounded by:

$$\begin{aligned}&\|\sum_{(i,j)}\mathbb{E}\big[S_{ij}^T S_{ij}\big]\| \\ =&\|\sum_{(i,j)}\mathbb{E}\big[(1 - \rho^{-1}\delta_{ij})^2\big](\mathbf{x}_i^T X\mathbf{y}_j)^2\mathbf{y}_j\mathbf{x}_i^T\mathbf{x}_i\mathbf{y}_j^T\| \\ \le&\rho^{-1}\|\mathcal{T}_L Z\|_\infty^2\|\sum_i\|\mathbf{x}_i\|_2^2\sum_j\mathbf{y}_j^T\mathbf{y}_j\| \\ =&\rho^{-1}d\|\mathcal{T}_L Z\|_\infty^2\|Y^T Y\| \\ =&\rho^{-1}d\|\mathcal{T}_L Z\|_\infty^2.\end{aligned}$$

Same bound on $\|\sum_{(i,j)}\mathbb{E}\big[S_{ij}S_{ij}^T\big]\|$ can be derived similarly. Thus, the lemma follows by applying matrix Bernstein inequality. □

Equipped with the above lemmas, now we are able to prove Lemma 2. For convenience, we will take $\epsilon \le e^{-1}$ in the proof.

7.4. Proof of Lemma 2

proof of 2a. Recall that by the definition of Y_j and Z_j, $Y_{j_0} = \sum_j q^{-1}\mathcal{T}_S\mathcal{P}_{\Omega_j}\mathcal{T}_L Z_{j-1}$. Thus,

$$\begin{aligned}\|W^L\| &= \|\mathcal{P}_{T^\perp}Y_{j_0}\| \\ &\le \sum_j\|q^{-1}\mathcal{P}_{T^\perp}\mathcal{T}_S\mathcal{P}_{\Omega_j}\mathcal{T}_L Z_{j-1}\| \\ &= \sum_j\|\mathcal{P}_{T^\perp}(q^{-1}\mathcal{T}_S\mathcal{P}_{\Omega_j}\mathcal{T}_L Z_{j-1} - Z_{j-1})\| \\ &\le \sum_j\|q^{-1}\mathcal{T}_S\mathcal{P}_{\Omega_j}\mathcal{T}_L Z_{j-1} - Z_{j-1}\|,\end{aligned}$$

where the second equality comes from $\mathcal{P}_{T^\perp}Z_{j-1} = 0$. As q is chosen to obey (11), we can apply Lemma 7 so that:

$$\begin{aligned}\|W^L\| &\le C_0'\sqrt{\frac{d\log d}{q}}\sum_j\|\mathcal{T}_L Z_{j-1}\|_\infty \\ &\le C_0'\sqrt{\frac{d\log d}{q}}\sum_j\epsilon^{j-1}\|\mathcal{T}_L(UV^T)\|_\infty \\ &\le C_0'(1-\epsilon)^{-1}\sqrt{\frac{d\log d}{q}}\frac{\sqrt{\mu_0 r}}{n}.\end{aligned}$$

From here we can conclude that

$$\|W^L\| \le C'\epsilon \le \frac{1}{4}$$

for some universal constant C', by choosing a small enough ϵ. □

proof of 2b. We have

$$\begin{aligned}
&\mathcal{P}_\Omega\mathcal{T}_L(UV^T + W^L) + \mathcal{P}_\Omega M^L \\
=&\mathcal{P}_\Omega\mathcal{T}_L(UV^T - \mathcal{P}_T Y_{j_0}) + \mathcal{P}_\Omega\mathcal{T}_L Y_{j_0} + \mathcal{P}_\Omega M^L \\
=&\mathcal{P}_\Omega\mathcal{T}_L(Z_{j_0}) + \mathcal{P}_\Omega(\mathcal{T}_L Y_{j_0} + M^L) \\
=&\mathcal{P}_\Omega\mathcal{T}_L(Z_{j_0}),
\end{aligned}$$

where the last equation holds because:

$$\begin{aligned}
&\mathcal{T}_L Y_{j_0} + M^L \\
=&\sum_j q^{-1}\mathcal{T}_L\mathcal{T}_S\mathcal{P}_{\Omega_j}\mathcal{T}_L Z_{j-1} + \mathcal{P}_{Q^\perp}\sum_j q^{-1}\mathcal{P}_{\Omega_j}\mathcal{T}_L Z_{j-1} \\
=&\sum_j q^{-1}\mathcal{P}_{\Omega_j}\mathcal{T}_L Z_{j-1} \qquad (16)
\end{aligned}$$

is a matrix only supported on Ω^C. Now, by applying Lemma 4, we have

$$\begin{aligned}
\|\mathcal{P}_\Omega\mathcal{T}_L Z_{j_0}\|_F \le&\|\mathcal{T}_L(Z_{j_0})\|_F = \|Z_{j_0}\|_F \\
\le&\epsilon^{j_0}\|UV^T\|_F = \epsilon^{j_0}\sqrt{r}.
\end{aligned}$$

Since $\epsilon \le e^{-1}$ and $j_0 \ge 2\log n$, the above quantity is less than $\lambda/4$. □

proof of 2c. By construction, $\mathcal{T}_L(UV^T + W^L) + M^L = \mathcal{T}_L Z_{j_0} + \mathcal{T}_L Y_{j_0} + M^L$. From part b we have $\|\mathcal{T}_L Z_{j_0}\|_\infty \le \|\mathcal{T}_L Z_{j_0}\|_F \le \lambda/4$, and the matrix $\mathcal{T}_L Y_{j_0} + M^L$ is supported on Ω^C. Thus, the claim is proved if we can show:

$$\|\mathcal{T}_L Y_{j_0} + M^L\|_\infty \le \frac{\lambda}{8}.$$

Using (16), we have:

$$\begin{aligned}
\|\mathcal{T}_L Y_{j_0} + M^L\|_\infty \le& q^{-1}\sum_j \|\mathcal{P}_{\Omega_j}\mathcal{T}_L Z_{j-1}\|_\infty \\
\le& q^{-1}\sum_j \|\mathcal{T}_L Z_{j-1}\|_\infty \\
\le& q^{-1}\sum_j \epsilon^{j-1}\|\mathcal{T}_L(UV^T)\|_\infty \\
\le& q^{-1}(1-\epsilon)^{-1}\frac{\sqrt{\mu_0 r}}{n}.
\end{aligned}$$

For q obeys (11), we have:

$$\|\mathcal{T}_L Y_{j_0} + M^L\|_\infty \le C\epsilon^2\sqrt{\frac{n^2}{\mu_0^2\mu_1 r d^2(\log d)^2}},$$

which will be smaller than $\lambda/8$ if:

$$\epsilon \le C'\left(\frac{\mu_0^2\mu_1 r d^2(\log d)^2}{n^3}\right)^{1/4}.$$

□

In summary, the proof above shows that 2a ∼2c hold if q is chosen to obey (11) and ϵ is chosen to be sufficiently small. As we fix a $j_0 \ge \lceil 2\log n\rceil$ and a small enough ϵ, a well-defined q can always be set to obey $1 > q \ge 2C\epsilon^{-2}(\mu_0\mu_1 r d\log d)/n^2$. This concludes the proof.

7.5. Proof of Lemma 3

For convenience, define $E = \mathrm{sgn}(S_0)$ whose sign is randomly distributed as:

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

In the following two parts of proof, we will focus on the event $\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\| < \sigma$. Notice that by Lemma 5, for any $\sigma > 0$, the event holds with large probability given a small enough ρ.

Proof of 3a. By construction, we have:

$$\begin{aligned}
W^S &= \lambda\mathcal{P}_{T^\perp}\mathcal{T}_S(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^{-1}E \\
&= \mathcal{P}_{T^\perp}\mathcal{T}_S K^{(1)} + \mathcal{P}_{T^\perp}\mathcal{T}_S K^{(2)}, \qquad (17)
\end{aligned}$$

where $K^{(1)}$, $K^{(2)}$ is defined by

$$\begin{aligned}
K^{(1)} &= \lambda E, \\
K^{(2)} &= \lambda\sum_{k\ge 1}(\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^k E.
\end{aligned}$$

We first bound the first term of (17). Since $\|\mathcal{P}_{T^\perp}\mathcal{T}_S K^{(1)}\| \le \|K^{(1)}\| \le \|\lambda E\|$, thus, using the argument in both Vershynin (2010); Candès et al. (2011), with high probability,

$$\|E\| \le 4\sqrt{n\rho}.$$

As $\lambda = 1/\sqrt{n}$, it implies:

$$\|\mathcal{P}_{T^\perp}\mathcal{T}_S K^{(1)}\| \le \|\lambda E\| \le 4\sqrt{\rho}. \qquad (18)$$

Now consider the second term $\|\mathcal{P}_{T^\perp}\mathcal{T}_S K^{(2)}\|$. For convenience, set the operator $\mathcal{R} := \sum_{k\ge 1}(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^k$. Then, $\|\mathcal{P}_{T^\perp}\mathcal{T}_S K^{(2)}\| \le \|K^{(2)}\| \le \|\lambda\mathcal{R}(E)\|$, and a standard covering argument could bound this operator norm. By Lemma 5.2 in Vershynin (2010), There exists a $1/2-$net N for a hypersphere S^{n-1} with its

size $\leq 5^n$. Then, From Lemma 5.3 in Vershynin (2010), we have:

$$\|\mathcal{R}(E)\| = \sup_{\mathbf{a},\mathbf{b}\in S^{n-1}} \langle \mathbf{a}, \mathcal{R}(E)\mathbf{b}\rangle \leq 4 \sup_{\mathbf{a},\mathbf{b}\in N} \langle \mathbf{a}, \mathcal{R}(E)\mathbf{b}\rangle.$$

Thus, consider any arbitrary pair $(\mathbf{a}, \mathbf{b}) \in N \times N$ with $\|\mathbf{a}\|_2 = \|\mathbf{b}\|_2 = 1$, we can define a random variable $S(\mathbf{a}, \mathbf{b})$ as:

$$S(\mathbf{a}, \mathbf{b}) = \langle \mathbf{a}, \mathcal{R}(E)\mathbf{b}\rangle = \langle \mathcal{R}(\mathbf{a}\mathbf{b}^T), E\rangle$$

by the fact that $\mathcal{R}$ is self-adjoint. Moreover, observe that given position of Ω is fixed, only random part of E is its sign and since the distribution is i.i.d. symmetric, we could apply Hoeffding's inequality to bound the probability that:

$$\Pr(|S(\mathbf{a}, \mathbf{b})| > t) \leq 2\exp\left(-\frac{2t^2}{\|\mathcal{R}(\mathbf{a}\mathbf{b}^T)\|_F^2}\right).$$

Note that by definition of operator 2-norm, $\|\mathcal{R}\| = \sup_{\hat{\mathbf{a}},\hat{\mathbf{b}}} \|\mathcal{R}(\hat{\mathbf{a}}\hat{\mathbf{b}}^T)\|_F / \|\hat{\mathbf{a}}\hat{\mathbf{b}}^T\|_F \geq \|\mathcal{R}(\mathbf{a}\mathbf{b}^T)\|_F$. Therefore, by an union bound:

$$\Pr(\sup_{\mathbf{a},\mathbf{b}\in N} |S(\mathbf{a}, \mathbf{b})| > t) \leq 2|N|^2 \exp\left(-\frac{2t^2}{\|\mathcal{R}\|^2}\right),$$

which leads to:

$$\Pr(\|\mathcal{R}(E)\| > t) \leq 2|N|^2 \exp\left(-\frac{t^2}{8\|\mathcal{R}\|^2}\right).$$

Furthermore, on the event $\|\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega\| \leq \sigma$, we can bound the operator norm by:

$$\|\mathcal{R}\| \leq \sum_{k\geq 1} \sigma^{2k} = \frac{\sigma^2}{1-\sigma^2}.$$

Putting all together, we can upper bound the second term of (17) by:

$$\begin{aligned}\Pr(\lambda\|\mathcal{R}(E)\| > t) \leq\ & 2\times 5^{2n}\exp\left(\frac{\gamma^2 t^2}{2\lambda^2}\right) \\ & + \Pr(\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\| > \sigma)\end{aligned}$$

where $\gamma = (1-\sigma^2)/2\sigma^2$. Thus, combining this bound with (18), and set $\lambda = 1/\sqrt{n}$, we can conclude $\|W^S\| \leq \frac{1}{4}$ with high probability if ρ (and thus σ) is sufficiently small. □

Proof of 3b. Let K be the matrix $\mathcal{P}_{\Omega^\perp}\mathcal{T}_L W^S + \mathcal{P}_{\Omega^\perp} M^S$, and our goal is to bound $\|K\|_\infty$. We first note that

$$\begin{aligned}K =& \mathcal{P}_{\Omega^\perp}\mathcal{T}_L W^S + \mathcal{P}_{\Omega^\perp} M^S \\ =& \lambda\mathcal{P}_{\Omega^\perp}(\mathcal{P}_Q - \mathcal{T}_L\mathcal{P}_T\mathcal{T}_S + \mathcal{P}_{Q^\perp})(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^{-1}E \\ =& -\lambda\mathcal{P}_{\Omega^\perp}\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^{-1}E.\end{aligned}$$

Consider any $K_{ij} \neq 0$. It must be in support of Ω^C and the element can be expressed as:

$$K_{ij} = \langle K, \mathbf{e}_i\mathbf{e}_j^T\rangle = \lambda\langle S(i,j), E\rangle,$$

where $S(i,j)$ is an $n\times n$ matrix defined by:

$$S(i,j) = (\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^{-1}\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S(\mathbf{e}_i\mathbf{e}_j^T).$$

Now, conditional on Ω, the sign of E is i.i.d. symmetric and again, by Hoeffding's inequality, each K_{ij} could be bounded by:

$$\Pr(|K_{ij}| > t\lambda) \leq 2\exp\left(-\frac{2t^2}{\|S(i,j)\|_F^2}\right),$$

and thus, by an union bound, we have:

$$\Pr(\max_{(i,j)}|K_{ij}| > t\lambda) \leq 2n^2\exp\left(-\frac{2t^2}{\max_{(i,j)}\|S(i,j)\|_F^2}\right).$$

Furthermore, since (10) holds, we have:

$$\begin{aligned}&\|S(i,j)\|_F \\ \leq& \|(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^{-1}\|\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\|\frac{\sqrt{2\mu_0\mu_1 rd}}{n}.\end{aligned}$$

In addition, on the event $\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\| \leq \sigma$, we can also bound $\|(\mathcal{P}_\Omega - \mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\mathcal{T}_S\mathcal{P}_\Omega)^{-1}\| \leq 1/(1-\sigma^2)$, and therefore,

$$\begin{aligned}\Pr(\|K\|_\infty > t\lambda) \leq\ & 2n^2\exp\left(-\frac{n^2\gamma^2t^2}{\mu_0\mu_1 rd}\right) \\ & + \Pr(\|\mathcal{P}_\Omega\mathcal{T}_L\mathcal{P}_T\| > \sigma),\end{aligned}$$

where $\gamma = (1-\sigma^2)/\sigma$. The Lemma is thus proved provided that $r \leq \rho_r(\mu_0\mu_1)^{-1}n^2/(d\log n)$ with some small enough ρ_r.

□