Non-convex Robust PCA

Praneeth Netrapalli[1*] U N Niranjan[2*] Sujay Sanghavi[3]
Animashree Anandkumar[2]
Prateek Jain[4]
[1]Microsoft Research, Cambridge MA. [2]The University of California at Irvine.
[3]The University of Texas at Austin. [4]Microsoft Research, India.

October 28, 2014

Abstract

We propose a new method for robust PCA – the task of recovering a low-rank matrix from sparse corruptions that are of unknown value and support. Our method involves alternating between projecting appropriate residuals onto the set of low-rank matrices, and the set of sparse matrices; each projection is *non-convex* but easy to compute. In spite of this non-convexity, we establish exact recovery of the low-rank matrix, under the same conditions that are required by existing methods (which are based on convex optimization). For an $m \times n$ input matrix ($m \leq n$), our method has a running time of $O\left(r^2mn\right)$ per iteration, and needs $O\left(\log(1/\epsilon)\right)$ iterations to reach an accuracy of ϵ. This is close to the running times of simple PCA via the power method, which requires $O\left(rmn\right)$ per iteration, and $O\left(\log(1/\epsilon)\right)$ iterations. In contrast, the existing methods for robust PCA, which are based on convex optimization, have $O\left(m^2n\right)$ complexity per iteration, and take $O\left(1/\epsilon\right)$ iterations, i.e., exponentially more iterations for the same accuracy.

Experiments on both synthetic and real data establishes the improved speed and accuracy of our method over existing convex implementations.

Keywords: Robust PCA, matrix decomposition, non-convex methods, alternating projections.

1 Introduction

Principal component analysis (PCA) is a common procedure for preprocessing and denoising, where a low rank approximation to the input matrix (such as the covariance matrix) is carried out. Although PCA is simple to implement via eigen-decomposition, it is sensitive to the presence of outliers, since it attempts to “force fit” the outliers to the low rank approximation. To overcome this, the notion of robust PCA is employed, where the goal is to remove sparse corruptions from an input matrix and obtain a low rank approximation. Robust PCA has been employed in a wide range of applications, including background modeling [LHGT04], 3d reconstruction [MZYM11], robust topic modeling [Shi13], and community detection [CSX12], and so on.

Concretely, robust PCA refers to the following problem: given an input matrix $M = L^* + S^*$, the goal is to decompose it into sparse S^* and low rank L^* matrices. The seminal works of [CSPW11, CLMW11] showed that this problem can be provably solved via convex relaxation methods, under some natural conditions on the low rank and sparse components. While the theory is elegant, in practice, convex techniques are expensive to run on a large scale and have poor convergence rates. Concretely, for decomposing an $m \times n$ matrix, say with $m \leq n$, the best specialized implementations (typically first-order methods) have a *per-iteration complexity* of $O\left(m^2n\right)$, and require $O(1/\epsilon)$ number of iterations to achieve an error of ϵ. In contrast, the usual PCA, which carries out a rank-r approximation of the input matrix, has $O(rmn)$ complexity per iteration – drastically smaller when r is much smaller than m, n. Moreover, PCA requires exponentially fewer iterations for convergence: an ϵ accuracy is achieved with only $O\left(\log(1/\epsilon)\right)$ iterations (assuming constant gap in singular values).

[*] Part of the work done while interning at Microsoft Research, India

In this paper, we design a non-convex algorithm which is "best of both the worlds" and bridges the gap between (the usual) PCA and convex methods for robust PCA. Our method has low computational complexity similar to PCA (i.e. scaling costs and convergence rates), and at the same time, has provable global convergence guarantees, similar to the convex methods. Proving global convergence for non-convex methods is an exciting recent development in machine learning. Non-convex alternating minimization techniques have recently shown success in many settings such as matrix completion [Kes12, JNS13, Har13], phase retrieval [NJS13], dictionary learning [AAJ$^+$13], tensor decompositions for unsupervised learning [AGH$^+$12], and so on. Our current work on the analysis of non-convex methods for robust PCA is an important addition to this growing list.

1.1 Summary of Contributions

We propose a simple intuitive algorithm for robust PCA with low per-iteration cost and a fast convergence rate. We prove tight guarantees for recovery of sparse and low rank components, which match those for the convex methods. In the process, we derive novel matrix perturbation bounds, when subject to sparse perturbations. Our experiments reveal significant gains in terms of speed-ups over the convex relaxation techniques, especially as we scale the size of the input matrices.

Our method consists of simple alternating (non-convex) projections onto low-rank and sparse matrices. For an $m \times n$ matrix, our method has a running time of $O(r^2 mn \log(1/\epsilon))$, where r is the rank of the low rank component. Thus, our method has a linear convergence rate, i.e. it requires $O(\log(1/\epsilon))$ iterations to achieve an error of ϵ, where r is the rank of the low rank component L^*. When the rank r is small, this nearly matches the complexity of PCA, (which is $O(rmn \log(1/\epsilon))$).

We prove recovery of the sparse and low rank components under a set of requirements which are tight and match those for the convex techniques (up to constant factors). In particular, under the deterministic sparsity model, where each row and each column of the sparse matrix S^* has at most α fraction of non-zeros, we require that $\alpha = O\left(1/(\mu^2 r)\right)$, where μ is the incoherence factor (see Section 3).

In addition to strong theoretical guarantees, in practice, our method enjoys significant advantages over the state-of-art solver for (1), viz., the inexact augmented Lagrange multiplier (IALM) method [CLMW11]. Our method outperforms IALM in all instances, as we vary the sparsity levels, incoherence, and rank, in terms of running time to achieve a fixed level of accuracy. In addition, on a real dataset involving the standard task of foreground-background separation [CLMW11], our method is significantly faster and provides visually better separation.

Overview of our techniques: Our proof technique involves establishing error contraction with each projection onto the sets of low rank and sparse matrices. We first describe the proof ideas when L^* is rank one. The first projection step is a hard thresholding procedure on the input matrix M to remove large entries and then we perform rank-1 projection of the residual to obtain $L^{(1)}$. Standard matrix perturbation results (such as Davis-Kahan) provide ℓ_2 error bounds between the singular vectors of $L^{(1)}$ and L^*. However, these bounds do not suffice for establishing the correctness of our method. Since the next step in our method involves hard thresholding of the residual $M - L^{(1)}$, we require element-wise error bounds on our low rank estimate. Inspired by the approach of Erdős et al. [EKYY13], where they obtain similar element-wise bounds for the eigenvectors of sparse Erdős–Rényi graphs, we derive these bounds by exploiting the fixed point characterization of the eigenvectors[1]. A Taylor's series expansion reveals that the perturbation between the estimated and the true eigenvectors consists of bounding the walks in a graph whose adjacency matrix corresponds to (a subgraph of) the sparse component S^*. We then show that if the graph is sparse enough, then this perturbation can be controlled, and thus, the next thresholding step results in further error contraction. We use an induction argument to show that the sparse estimate is always contained in the true support of S^*, and that there is an error contraction in each step. For the case, where L^* has rank $r > 1$, our algorithm proceeds in several stages, where we progressively compute higher rank projections which alternate with the hard thresholding steps. In stage $k = [1, 2, \ldots, r]$, we compute rank-k projections, and show that after a sufficient number of alternating projections, we reduce the error to the level of $(k+1)^{\text{th}}$ singular value of L^*, using similar arguments as in the

[1] If the input matrix M is not symmetric, we embed it in a symmetric matrix and consider the eigenvectors of the corresponding matrix.

rank-1 case. We then proceed to performing rank-$(k+1)$ projections which alternate with hard thresholding. This stage-wise procedure is needed for ill-conditioned matrices, since we cannot hope to recover lower eigenvectors in the beginning when there are large perturbations. Thus, we establish global convergence guarantees for our proposed non-convex robust PCA method.

1.2 Related Work

Guaranteed methods for robust PCA have received a lot of attention in the past few years, starting from the seminal works of [CSPW11, CLMW11], where they showed recovery of an incoherent low rank matrix L^* through the following convex relaxation method:

$$\text{Conv-RPCA}: \qquad \min_{L,S} \|L\|_* + \lambda\|S\|_1, \qquad \text{s.t.,} \quad M = L + S, \tag{1}$$

where $\|L\|_*$ denotes the nuclear norm of L (nuclear norm is the sum of singular values). A typical solver for this convex program involves projection on to ℓ_1 and nuclear norm balls (which are convex sets). Note that the convex method can be viewed as "soft" thresholding in the standard and spectral domains, while our method involves hard thresholding in these domains.

[CSPW11] and [CLMW11] consider two different models of sparsity for S^*. Chandrasekaran et al. [CSPW11] consider a deterministic sparsity model, where each row and column of the $m \times n$ matrix, S, has at most α fraction of non-zero entries. For guaranteed recovery, they require $\alpha = O\left(1/(\mu^2 r\sqrt{n})\right)$, where μ is the incoherence level of L^*, and r is its rank. Hsu et al. [HKZ11] improve upon this result to obtain guarantees for an optimal sparsity level of $\alpha = O\left(1/(\mu^2 r)\right)$. This *matches* the requirements of our non-convex method for exact recovery. Note that when the rank $r = O(1)$, this allows for a constant fraction of corrupted entries. Candès et al. [CLMW11] consider a different model with random sparsity and additional incoherence constraints, viz., they require $\|UV^\top\|_\infty < \mu\sqrt{r}/n$. Note that our assumption of incoherence, viz., $\|U^{(i)}\| < \mu\sqrt{r/n}$, only yields $\|UV^\top\|_\infty < \mu^2 r/n$. The additional assumption enables [CLMW11] to prove exact recovery with a constant fraction of corrupted entries, even when L^* is nearly full-rank. We note that removing the $\|UV^\top\|_\infty$ condition for robust PCA would imply solving the planted clique problem when the clique size is less than $\sqrt{n}$ [Che13]. Thus, our recovery guarantees are *tight* upto constants without these additional assumptions.

A number of works have considered modified models under the robust PCA framework, e.g. [ANW12, XCS12]. For instance, Agarwal et al. [ANW12] relax the incoherence assumption to a weaker "diffusivity" assumption, which bounds the magnitude of the entries in the low rank part, but incurs an additional approximation error. Xu et al.[XCS12] impose special sparsity structure where a column can either be non-zero or fully zero.

In terms of state-of-art specialized solvers, [CLMW11] implements the in-exact augmented Lagrangian multipliers (IALM) method and provides guidelines for parameter tuning. Other related methods such as multi-block alternating directions method of multipliers (ADMM) have also been considered for robust PCA, e.g. [WHML13]. Recently, a multi-step multi-block stochastic ADMM method was analyzed for this problem [SAJ14], and this requires $1/\epsilon$ iterations to achieve an error of ϵ. In addition, the convergence rate is tight in terms of scaling with respect to problem size (m, n) and sparsity and rank parameters, under random noise models.

There is only one other work which considers a non-convex method for robust PCA [KC12]. However, their result holds only for significantly more restrictive settings and does not cover the deterministic sparsity assumption that we study. Moreover, the projection step in their method can have an arbitrarily large rank, so the running time is still $O(m^2 n)$, which is the same as the convex methods. In contrast, we have an improved running time of $O(r^2 mn)$.

2 Algorithm

In this section, we present our algorithm for the robust PCA problem. The robust PCA problem can be formulated as the following optimization problem: find L, S s.t. $\|M - L - S\|_F \leq \epsilon$[2] and

[2] ϵ is the desired reconstruction error

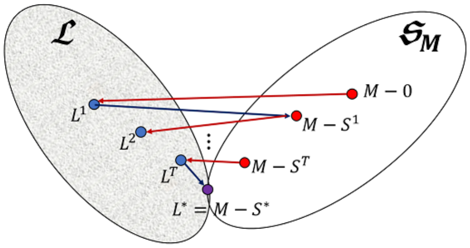


Figure 1: Illustration of alternating projections. The goal is to find a matrix L^* which lies in the intersection of two sets: $\mathcal{L} = \{$ set of rank-r matrices$\}$ and $\mathcal{S}_M = \{M - S,\ \text{where } S \text{ is a sparse matrix}\}$. Intuitively, our algorithm alternately projects onto the above two non-convex sets, while appropriately relaxing the rank and the sparsity levels.

1. L lies in the set of low-rank matrices,
2. S lies in the set of sparse matrices.

A natural algorithm for the above problem is to iteratively project $M - L$ onto the set of sparse matrices to update S, and then to project $M - S$ onto the set of low-rank matrices to update L. Alternatively, one can view the problem as that of finding a matrix L in the intersection of the following two sets: a) $\mathcal{L} = \{$ set of rank-r matrices$\}$, b) $\mathcal{S}_M = \{M - S,\ \text{where } S \text{ is a sparse matrix}\}$. Note that these projections can be done efficiently, even though the sets are non-convex. Hard thresholding (HT) is employed for projections on to sparse matrices, and singular value decomposition (SVD) is used for projections on to low rank matrices.

Rank-1 case: We first describe our algorithm for the special case when L^* is rank 1. Our algorithm performs an initial hard thresholding to remove very large entries from input M. Note that if we performed the projection on to rank-1 matrices without the initial hard thresholding, we would not make any progress since it is subject to large perturbations. We alternate between computing the rank-1 projection of $M - S$, and performing hard thresholding on $M - L$ to remove entries exceeding a certain threshold. This threshold is gradually decreased as the iterations proceed, and the algorithm is run for a certain number of iterations (which depends on the desired reconstruction error).

General rank case: When L^* has rank $r > 1$, a naive extension of our algorithm consists of alternating projections on to rank-r matrices and sparse matrices. However, such a method has poor performance on ill-conditioned matrices. This is because after the initial thresholding of the input matrix M, the sparse corruptions in the residual are of the order of the top singular value (with the choice of threshold as specified in the algorithm). When the lower singular values are much smaller, the corresponding singular vectors are subject to relatively large perturbations and thus, we cannot make progress in improving the reconstruction error. To alleviate the dependence on the condition number, we propose an algorithm that proceeds in stages. In the k^{th} stage, the algorithm alternates between rank-k projections and hard thresholding for a certain number of iterations. We run the algorithm for r stages, where r is the rank of L^*. Intuitively, through this procedure, we recover the lower singular values only after the input matrix is sufficiently denoised, i.e. sparse corruptions at the desired level have been removed. Figure 1 shows a pictorial representation of the alternating projections in different stages.

Parameters: As can be seen, the only real parameter to the algorithm is β, used in thresholding, which represents "spikiness" of L^*. That is if the user expects L^* to be "spiky" and the sparse part to be heavily diffused, then higher value of β can be provided. In our implementation, we found that selecting β aggressively helped speed up recovery of our algorithm. In particular, we selected $\beta = 1/\sqrt{n}$.

Complexity: The complexity of each iteration within a single stage is $O(kmn)$, since it involves calculating the rank-k approximation[3] of an $m \times n$ matrix (done e.g. via vanilla PCA). The number of iterations in each stage is $O(\log(1/\epsilon))$ and there are at most r stages. Thus the overall complexity of the entire algorithm is then $O(r^2 mn \log(1/\epsilon))$. This is drastically lower than the best known bound of $O\left(m^2 n/\epsilon\right)$ on the number of iterations required by convex methods, and just a factor r away from the complexity of vanilla PCA.

[3] Note that we only require a rank-k approximation of the matrix rather than the actual singular vectors. Thus, the computational complexity has no dependence on the gap between the singular values.

Algorithm 1 $(\widehat{L},\ \widehat{S}) = \text{AltProj}(M, \epsilon, r, \beta)$: Non-convex Alternating Projections based Robust PCA

```
Input: Matrix $M \in \mathbb{R}^{m \times n}$, convergence criterion $\epsilon$, target rank $r$, thresholding parameter $\beta$.
$P_k(A)$ denotes the best rank-$k$ approximation of matrix $A$. $HT_\zeta(A)$ denotes hard-thresholding, i.e. $(HT_\zeta(A))_{ij} = A_{ij}$ if $|A_{ij}| \geq \zeta$ and 0 otherwise.
Set initial threshold $\zeta_0 \leftarrow \beta\sigma_1(M)$.
$L^{(0)} = 0, S^{(0)} = HT_{\zeta_0}(M - L^{(0)})$
for Stage $k = 1$ to $r$ do
    for Iteration $t = 0$ to $T = 10\log\left(n\beta \left\|M - S^{(0)}\right\|_2 / \epsilon\right)$ do
        Set threshold $\zeta$ as
            $$\zeta \ = \ \beta\left(\sigma_{k+1}(M - S^{(t)}) + \left(\frac{1}{2}\right)^t \sigma_k(M - S^{(t)})\right) \qquad (2)$$
        $L^{(t+1)} = P_k(M - S^{(t)})$
        $S^{(t+1)} = HT_\zeta(M - L^{(t+1)})$
    end for
    if $\beta\sigma_{k+1}(L^{(t+1)}) < \frac{\epsilon}{2n}$ then
        Return: $L^{(T)}, S^{(T)}$   /* Return rank-k estimate if remaining part has small norm */
    else
        $S^{(0)} = S^{(T)}$   /* Continue to the next stage */
    end if
end for
Return: $L^{(T)}, S^{(T)}$
```

3 Analysis

In this section, we present our main result on the correctness of AltProj. We assume the following conditions:

(L1) Rank of L^* is at most r.

(L2) L^* is μ-incoherent, i.e., if $L^* = U^*\Sigma^*(V^*)^\top$ is the SVD of L^*, then $\|(U^*)^i\|_2 \leq \frac{\mu\sqrt{r}}{\sqrt{m}}$, $\forall 1 \leq i \leq m$ and $\|(V^*)^i\|_2 \leq \frac{\mu\sqrt{r}}{\sqrt{n}}$, $\forall 1 \leq i \leq n$, where $(U^*)^i$ and $(V^*)^i$ denote the i^{th} rows of U^* and V^* respectively.

(S1) Each row and column of S have at most α fraction of non-zero entries such that $\alpha \leq \frac{1}{512\mu^2 r}$.

Note that in general, it is not possible to have a unique recovery of low-rank and sparse components. For example, if the input matrix M is both sparse and low rank, then there is no unique decomposition (e.g. $M = \boldsymbol{e}_1\boldsymbol{e}_1^\top$). The above conditions ensure uniqueness of the matrix decomposition problem.

Additionally, we set the parameter β in Algorithm 1 be set as $\beta = \frac{4\mu^2 r}{\sqrt{mn}}$.

We now establish that our proposed algorithm recovers the low rank and sparse components under the above conditions.

Theorem 1 (Noiseless Recovery)**.** *Under conditions (L1), (L2) and S^*, and choice of β as above, the outputs $\widehat{L}$ and $\widehat{S}$ of Algorithm 1 satisfy:*

$$\left\|\widehat{L} - L^*\right\|_F \leq \epsilon, \left\|\widehat{S} - S^*\right\|_\infty \leq \frac{\epsilon}{\sqrt{mn}}, \ \textit{and} \quad \textit{Supp}\left(\widehat{S}\right) \subseteq \textit{Supp}\left(S^*\right).$$

Remark (tight recovery conditions): Our result is tight up to constants, in terms of allowable sparsity level under the deterministic sparsity model. In other words, if we exceed the sparsity limit imposed in S1, it is possible to construct instances where there is no unique decomposition[4]. Our conditions L1, L2 and S1 also

[4] For instance, consider the $n \times n$ matrix which has r copies of the all ones matrix, each of size $\frac{n}{r}$, placed across the diagonal. We see that this matrix has rank r and is incoherent with parameter $\mu = 1$. Note that a fraction of $\alpha = O\left(1/r\right)$ sparse perturbations suffice to erase one of these blocks making it impossible to recover the matrix.

match the conditions required by the convex method for recovery, as established in [HKZ11].
Remark (convergence rate): Our method has a linear rate of convergence, i.e. $O(\log(1/\epsilon))$ to achieve an error of ϵ, and hence we provide a strongly polynomial method for robust PCA. In contrast, the best known bound for convex methods for robust PCA is $O(1/\epsilon)$ iterations to converge to an ϵ-approximate solution.

Theorem 1 provides recovery guarantees assuming that L^* is exactly rank-r. However, in several real-world scenarios, L^* can be nearly rank-r. Our algorithm can handle such situations, where $M = L^* + N^* + S^*$, with N^* being an additive noise. Theorem 1 is a special case of the following theorem which provides recovery guarantees when N^* has small ℓ_∞ norm.
Theorem 2 (Noisy Recovery)**.** *Under conditions (L1), (L2) and S^*, and choice of β as in Theorem 1, when the noise $\|N^*\|_\infty \leq \frac{\sigma_r(L^*)}{100n}$,the outputs $\widehat{L}, \widehat{S}$ of Algorithm 1 satisfy:*

$$\left\|\widehat{L} - L^*\right\|_F \leq \epsilon + 2\mu^2 r \left(7\left\|N^*\right\|_2 + \frac{8\sqrt{mn}}{\sqrt{r}}\left\|N^*\right\|_\infty\right),$$

$$\left\|\widehat{S} - S^*\right\|_\infty \leq \frac{\epsilon}{\sqrt{mn}} + \frac{2\mu^2 r}{\sqrt{mn}}\left(7\left\|N^*\right\|_2 + \frac{8\sqrt{mn}}{\sqrt{r}}\left\|N^*\right\|_\infty\right), \text{ and } Supp\left(\widehat{S}\right) \subseteq Supp\left(S^*\right).$$

3.1 Proof Sketch

We now present the key steps in the proof of Theorem 1. A detailed proof is provided in the appendix.

Step I: Reduce to the symmetric case, while maintaining incoherence of L^* and sparsity of S^*. Using standard symmetrization arguments, we can reduce the problem to the symmetric case, where all the matrices involved are symmetric. See appendix for details on this step.

Step II: Show decay in $\|L - L^*\|_\infty$ after projection onto the set of rank-k matrices. The t-th iterate $L^{(t+1)}$ of the k-th stage is given by $L^{(t+1)} = P_k(L^* + S^* - S^{(t)})$. Hence, $L^{(t+1)}$ is obtained by using the top principal components of a perturbation of L^* given by $L^* + (S^* - S^{(t)})$. The key step in our analysis is to show that when an incoherent and low-rank L^* is perturbed by a sparse matrix $S^* - S^{(t)}$, then $\|L^{(t+1)} - L^*\|_\infty$ is small and is much smaller than $|S^* - S^{(t)}|_\infty$. The following lemma formalizes the intuition; see the appendix for a detailed proof.
Lemma 1. *Let L^*, S^* be symmetric and satisfy the assumptions of Theorem 1 and let $S^{(t)}$ and $L^{(t)}$ be the t^{th} iterates of the k^{th} stage of Algorithm 1. Let $\sigma_1^*, \ldots, \sigma_n^*$ be the eigenvalues of L^*, s.t., $|\sigma_1^*| \geq \cdots \geq |\sigma_r^*|$. Then, the following holds:*

$$\left\|L^{(t+1)} - L^*\right\|_\infty \leq \frac{2\mu^2 r}{n}\left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^t \left|\sigma_k^*\right|\right),$$

$$\left\|S^* - S^{(t+1)}\right\|_\infty \leq \frac{8\mu^2 r}{n}\left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^t \left|\sigma_k^*\right|\right), \text{ and } Supp\left(S^{(t+1)}\right) \subseteq Supp\left(S^*\right).$$

Moreover, the outputs $\widehat{L}$ and $\widehat{S}$ of Algorithm 1 satisfy:

$$\left\|\widehat{L} - L^*\right\|_F \leq \epsilon, \quad \left\|\widehat{S} - S^*\right\|_\infty \leq \frac{\epsilon}{n}, \text{ and } \quad Supp\left(\widehat{S}\right) \subseteq Supp\left(S^*\right).$$

Step III: Show decay in $\|S - S^*\|_\infty$ after projection onto the set of sparse matrices. We next show that if $\|L^{(t+1)} - L^*\|_\infty$ is much smaller than $\|S^{(t)} - S^*\|_\infty$ then the iterate $S^{(t+1)}$ also has a much smaller error (w.r.t. S^*) than $S^{(t)}$. The above given lemma formally provides the error bound.

Step IV: Recurse the argument. We have now reduced the ℓ_∞ norm of the sparse part by a factor of half, while maintaining its sparsity. We can now go back to steps II and III and repeat the arguments for subsequent iterations.

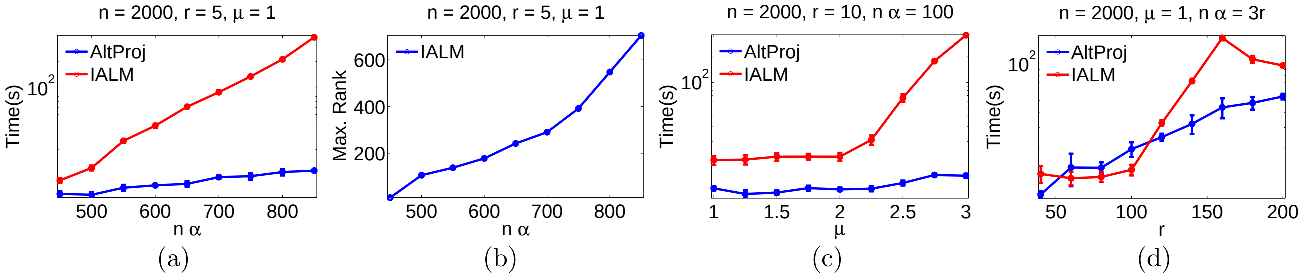


Figure 2: Comparison of AltProj and IALM on synthetic datasets. (a) Running time of AltProj and IALM with varying α. (b) Maximum rank of the intermediate iterates of IALM. (c) Running time of AltProj and IALM with varying μ. (d) Running time of AltProj and IALM with varying r.

4 Experiments

We now present an empirical study of our AltProj method. The goal of this study is two-fold: a) establish that our method indeed recovers the low-rank and sparse part exactly, without significant parameter tuning, b) demonstrate that AltProj is significantly faster than Conv-RPCA (see (1)); we solve Conv-RPCA using the IALM method [CLMW11], a state-of-the-art solver [LCM10]. We implemented our method in Matlab and used a Matlab implementation of the IALM method by [LCM10].

We consider both synthetic experiments and experiments on real data involving the problem of foreground-background separation in a video. Each of our results for synthetic datasets is averaged over 5 runs.

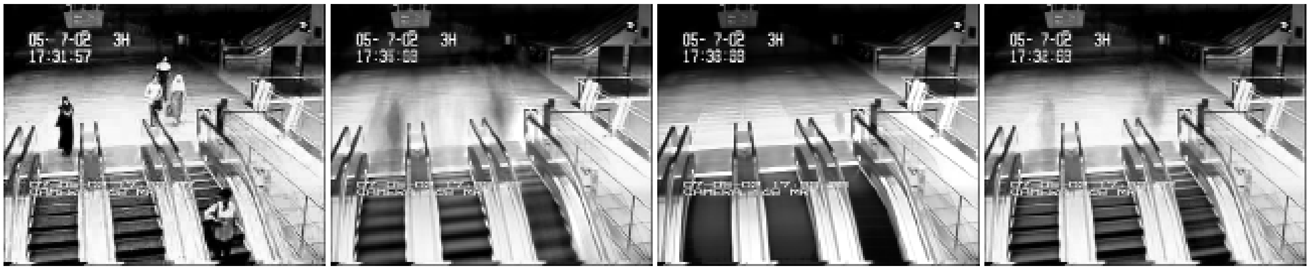
Parameter Setting: Our pseudo-code (Algorithm 1) prescribes the threshold ζ in Step 4, which depends on the knowledge of the singular values of the low rank component L^*. Instead, in the experiments, we set the threshold at the $(t+1)$-th step of k-th stage as $\zeta = \frac{\mu\sigma_{k+1}(M-S^{(t)})}{\sqrt{n}}$. For synthetic experiments, we employ the μ used for data generation, and for real-world datasets, we tune μ through cross-validation. We found that the above thresholding provides exact recovery while speeding up the computation significantly. We would also like to note that [CLMW11] sets the regularization parameter λ in Conv-RPCA (1) as $1/\sqrt{n}$ (assuming $m \leq n$). However, we found that for problems with large incoherence such a parameter setting *does not* provide exact recovery. Instead, we set $\lambda = \mu/\sqrt{n}$ in our experiments.

Synthetic datasets: Following the experimental setup of [CLMW11], the low-rank part $L^* = UV^T$ is generated using normally distributed $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$. Similarly, $supp(S^*)$ is generated by sampling a uniformly random subset of $[m]\times[n]$ with size $\|S^*\|_0$ and each non-zero S^*_{ij} is drawn i.i.d. from the uniform distribution over $[r/(2\sqrt{mn}), r/\sqrt{mn}]$. For increasing incoherence of L^*, we randomly zero-out rows of U, V and then renormalize them.

There are three key problem parameters for RPCA with a fixed matrix size: a) sparsity of S^*, b) incoherence of L^*, c) rank of L^*. We investigate performance of both AltProj and IALM by varying each of the three parameters while fixing the others. In our plots (see Figure 2), we report computational time required by each of the two methods for decomposing M into $L+S$ up to a relative error ($\|M-L-S\|_F/\|M\|_F$) of 10^{-3}. Figure 2 shows that AltProj scales significantly better than IALM for increasingly dense S^*. We attribute this observation to the fact that as $\|S^*\|_0$ increases, the problem is "harder" and the intermediate iterates of IALM have ranks significantly larger than r. Our intuition is confirmed by Figure 2 (b), which shows that when density (α) of S^* is 0.4 then the intermediate iterates of IALM can be of rank over 500 while the rank of L^* is only 5. We observe a similar trend for the other parameters, i.e., AltProj scales significantly better than IALM with increasing incoherence parameter μ (Figure 2 (c)) and increasing rank (Figure 2 (d)). See Appendix C for additional plots.

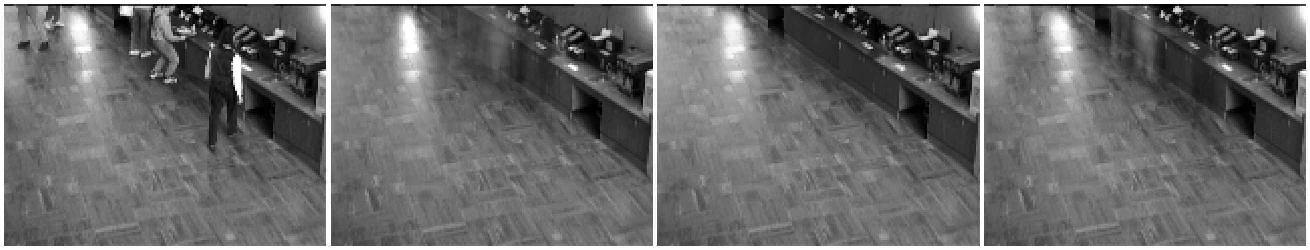
Real-world datasets: Next, we apply our method to the problem of foreground-background (F-B) separation in a video [LHGT04]. The observed matrix M is formed by vectorizing each frame and stacking them column-wise. Intuitively, the background in a video is the static part and hence forms a low-rank component while the foreground is a dynamic but sparse perturbation.

Here, we used two benchmark datasets named *Escalator* and *Restaurant* dataset. The *Escalator* dataset has 3417 frames at a resolution of 160×130. We first applied the standard PCA method for extracting low-rank part. Figure 3 (b) shows the extracted background from the video. There are several artifacts (shadows of



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

Figure 3: Foreground-background separation in the *Escalator* video. (a): Original image frame. (b): Best rank-10 approximation; time taken is $3.1s$. (c): Low-rank frame obtained using AltProj; time taken is $63.2s$. (d): Low-rank frame obtained using IALM; time taken is $1688.9s$.



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

Figure 4: Foreground-background separation in the *Restaurant* video. (a): Original frame from the video. (b): Best rank-10 approximation (using PCA) of the original frame; $2.8s$ were required to compute the solution (c): Low-rank part obtained using AltProj; computational time required by AltProj was $34.9s$. (d): Low-rank part obtained using IALM; $693.2s$ required by IALM to compute the low-rank+sparse decomposition.

people near the escalator) that are not desirable. In contrast, both IALM and AltProj obtain significantly better F-B separation (see Figure 3(c), (d)). Interestingly, AltProj removes the steps of the escalator which are moving and arguably are part of the dynamic foreground, while IALM keeps the steps in the background part. Also, our method is significantly faster, i.e., our method, which takes $63.2s$ is about 26 times faster than IALM, which takes $1688.9s$.

Restaurant dataset: Figure 4 shows the comparison of AltProj and IALM on a subset of the "Restaurant" dataset where we consider the last 2055 frames at a resolution of 120×160. AltProj was around 19 times faster than IALM. Moreover, visually, the background extraction seems to be of better quality (for example, notice the blur near top corner counter in the IALM solution). Plot(b) shows the PCA solution and that also suffers from a similar blur at the top corner of the image, while the background frame extracted by AltProj does not have any noticeable artifacts.

5 Conclusion

In this work, we proposed a non-convex method for robust PCA, which consists of alternating projections on to low rank and sparse matrices. We established global convergence of our method under conditions which match those for convex methods. At the same time, our method has much faster running times, and has superior experimental performance. This work opens up a number of interesting questions for future investigation. While we match the convex methods, under the deterministic sparsity model, studying the random sparsity model is of interest. Our noisy recovery results assume deterministic noise; improving the results under random noise needs to be investigated. There are many decomposition problems beyond the robust PCA setting, e.g. structured sparsity models, robust tensor PCA problem, and so on. It is interesting to see if we can establish global convergence for non-convex methods in these settings.

Acknowledgements

AA and UN would like to acknowledge NSF grant CCF-1219234, ONR N00014-14-1-0665, and Microsoft faculty fellowship. SS would like to acknowledge NSF grants 1302435, 0954059, 1017525 and DTRA grant HDTRA1-13-1-0024. PJ would like to acknowledge Nikhil Srivastava and Deeparnab Chakrabarty for several insightful discussions during the course of the project.

References

[AAJ+13] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization. *Available on arXiv:1310.7991*, Oct. 2013.

[AGH+12] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor Methods for Learning Latent Variable Models. *Available at arXiv:1210.7559*, Oct. 2012.

[ANW12] A. Agarwal, S. Negahban, and M. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *The Annals of Statistics*, 40(2):1171–1197, 2012.

[Bha97] Rajendra Bhatia. *Matrix Analysis*. Springer, 1997.

[Che13] Y. Chen. Incoherence-Optimal Matrix Completion. *ArXiv e-prints*, October 2013.

[CLMW11] Emmanuel J. Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *J. ACM*, 58(3):11, 2011.

[CSPW11] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A. Parrilo, and Alan S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011.

[CSX12] Yudong Chen, Sujay Sanghavi, and Huan Xu. Clustering sparse graphs. In *Advances in neural information processing systems*, pages 2204–2212, 2012.

[EKYY13] László Erdős, Antti Knowles, Horng-Tzer Yau, and Jun Yin. Spectral statistics of Erdős–Rényi graphs I: Local semicircle law. *The Annals of Probability*, 41(3B):2279–2375, 2013.

[Har13] Moritz Hardt. On the provable convergence of alternating minimization for matrix completion. *arXiv preprint arXiv:1312.0925*, 2013.

[HKZ11] Daniel Hsu, Sham M Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *Information Theory, IEEE Transactions on*, 57(11):7221–7234, 2011.

[JNS13] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In *Proceedings of the 45th annual ACM Symposium on theory of computing*, pages 665–674. ACM, 2013.

[KC12] Anastasios Kyrillidis and Volkan Cevher. Matrix alps: Accelerated low rank and sparse matrix reconstruction. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pages 185–188. Ieee, 2012.

[Kes12] Raghunandan H. Keshavan. Efficient algorithms for collaborative filtering. Phd Thesis, Stanford University, 2012.

[LCM10] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. *arXiv preprint arXiv:1009.5055*, 2010.

[LHGT04] Liyuan Li, Weimin Huang, IY-H Gu, and Qi Tian. Statistical modeling of complex backgrounds for foreground object detection. *Image Processing, IEEE Transactions on*, 13(11):1459–1472, 2004.

[MZYM11] Hossein Mobahi, Zihan Zhou, Allen Y. Yang, and Yi Ma. Holistic 3d reconstruction of urban structures from low-rank textures. In *ICCV Workshops*, pages 593–600, 2011.

[NJS13] Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. In *NIPS*, pages 2796–2804, 2013.

[SAJ14] H. Sedghi, A. Anandkumar, and E. Jonckheere. Guarantees for Stochastic ADMM in High Dimensions. *Preprint.*, Feb. 2014.

[Shi13] Lei Shi. Sparse additive text models with low rank background. In *Advances in Neural Information Processing Systems*, pages 172–180, 2013.

[WHML13] X. Wang, M. Hong, S. Ma, and Z. Luo. Solving multiple-block separable convex minimization problems using two-block alternating direction method of multipliers. *arXiv preprint arXiv:1308.5294*, 2013.

[XCS12] Huan Xu, Constantine Caramanis, and Sujay Sanghavi. Robust pca via outlier pursuit. *IEEE Transactions on Information Theory*, 58(5):3047–3064, 2012.

A Proof of Theorem 1

We will start with some preliminary lemmas. The first lemma is the well known Weyl's inequality in the matrix setting[Bha97].

Lemma 2. *Suppose $B = A + E$ be an $n \times n$ matrix. Let $\lambda_1, \cdots, \lambda_n$ and $\sigma_1, \cdots, \sigma_n$ be the eigenvalues of B and A respectively such that $\lambda_1 \geq \cdots \geq \lambda_n$ and $\sigma_1 \geq \cdots \geq \sigma_n$. Then we have:*

$$|\lambda_i - \sigma_i| \leq \|E\|_2 \ \forall\, i \in [n].$$

The following lemma is the Davis-Kahan theorem[Bha97], specialized for rank-1 matrices.

Lemma 3. *Suppose $B = A + E$. Let $A = \boldsymbol{u}^*(\boldsymbol{u}^*)^\top$ be a rank-1 matrix with unit spectral norm. Suppose further that $\|E\|_2 < \frac{1}{2}$. Then, we have:*

$$|\lambda - 1| < \|E\|_2\,, \ \textit{and}$$
$$\left| \langle \boldsymbol{u}, \boldsymbol{u}^* \rangle^2 - 1 \right| < 4\,\|E\|_2\,,$$

where λ and $\boldsymbol{u}$ are the top eigenvalue eigenvector pair of B.

As outlined in Section 3.1 (and formalized in the proof of Theorem 1), it is sufficient to prove the correctness of Algorithm 1 for the case of symmetric matrices. So, most of the lemmas we prove in this section assume that the matrices are symmetric.

Lemma 4. *Let $S \in \mathbb{R}^{n\times n}$ satisfy assumption $(S1)$. Then, $\|S\|_2 \leq \alpha n \|S\|_\infty$.*

Proof of Lemma 4. Let x, y be unit vectors such that $\|S\|_2 = x^T S y = \sum_{ij} x_i y_j S_{ij}$. Then, using $a \cdot b \leq (a^2 + b^2)/2$, we have:

$$\|S\|_2 \leq \frac{1}{2}\sum_{ij}(x_i^2 + y_j^2) S_{ij} \leq \frac{1}{2}(\alpha n \|S\|_\infty + \alpha n \|S\|_\infty), \tag{3}$$

where the last inequality follows from the fact that S has at most αn non-zeros per row and per column. □

Lemma 5. *Let $S \in \mathbb{R}^{n\times n}$ satisfy assumption $(S1)$. Also, let $U \in \mathbb{R}^{n\times r}$ be a μ-incoherent orthogonal matrix, i.e., $\max_i \left\|\boldsymbol{e}_i{}^\top U\right\|_2 \leq \frac{\mu\sqrt{r}}{\sqrt{n}}$, where $\boldsymbol{e}_i$ stands for the i^{th} standard basis vector. Then, $\forall p \geq 0$, the following holds:*

$$\max_i \left\|\boldsymbol{e}_i{}^\top S^p U\right\|_2 \leq \frac{\mu\sqrt{r}}{\sqrt{n}}(\alpha \cdot n \cdot \|S\|_\infty)^p.$$

Proof of Lemma 5. We prove the lemma using mathematical induction.

Base Case ($p = 0$): This is just a restatement of the incoherence of U.

Induction step: We have:

$$\begin{aligned}
\left\|\boldsymbol{e}_i{}^\top (S)^{p+1} U\right\|_2^2 &= \|\boldsymbol{e}_i{}^\top S(S^p U)\|_2^2 = \sum_\ell (\boldsymbol{e}_i{}^\top S (S^p U)\boldsymbol{e}_\ell)^2 = \sum_\ell (\sum_j S_{ij} \boldsymbol{e}_j{}^\top (S^p U)\boldsymbol{e}_\ell)^2 \\
&= \sum_{j_1 j_2} S_{ij_1} S_{ij_2} \sum_\ell (\boldsymbol{e}_{j_1}{}^\top (S^p U)\boldsymbol{e}_\ell)(\boldsymbol{e}_\ell{}^\top (S^p U)^\top \boldsymbol{e}_{j_2}) \\
&\overset{\zeta_1}{\leq} \sum_{j_1 j_2} S_{ij_1} S_{ij_2} (\boldsymbol{e}_{j_1}{}^\top (S^p U)(S^p U)^\top \boldsymbol{e}_{j_2}) \leq \sum_{j_1 j_2} S_{ij_1} S_{ij_2} \|\boldsymbol{e}_{j_1}^T (S^p U)\|_2 \|\boldsymbol{e}_{j_2}{}^\top (S^p U)\|_2 \\
&\overset{\zeta_2}{\leq} \frac{\mu^2 r}{n}(\alpha \cdot n \cdot \|S\|_\infty)^{2p},
\end{aligned}$$

where ζ_1 follows by $\sum_{\ell=1}^t \boldsymbol{e}_\ell \boldsymbol{e}_\ell{}^\top = I$, and ζ_2 follows from assumption $(S1)$ on S and from the inductive hypothesis on $\left\|\boldsymbol{e}_i{}^\top S^p U\right\|_2$. □

In what follows, we prove a number of lemmas concerning the structure of $L^{(t)}$ and $E^{(t)} := S^* - S^{(t)}$. The following lemma shows that the threshold in (2) is close to that with $M - S^{(t)}$ replaced by L^*.

Lemma 6. *Let L^*, S^* be symmetric and satisfy the assumptions of Theorem 1 and let $S^{(t)}$ be the t^{th} iterate of the k^{th} stage of Algorithm 1. Let $\sigma_1^*, \ldots, \sigma_r^*$ be the eigenvalues of L^*, such that $|\sigma_1^*| \geq \cdots \geq |\sigma_r^*|$ and $\lambda_1, \cdots, \lambda_n$ be the eigenvalues of $M - S^{(t)}$ such that $|\lambda_1| \geq \cdots \geq |\lambda_n|$. Recall that $E^{(t)} := S^* - S^{(t)}$. Suppose further that*

1. *$\left\|E^{(t)}\right\|_\infty \leq \frac{8\mu^2 r}{n}\left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^{t-1}|\sigma_k^*|\right)$, and*
2. *Supp $\left(E^{(t)}\right) \subseteq$ Supp (S^*).*

Then,

$$\frac{7}{8}\left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^t |\sigma_k^*|\right) \leq \left(|\lambda_{k+1}| + \left(\frac{1}{2}\right)^t |\lambda_k|\right) \leq \frac{9}{8}\left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^t |\sigma_k^*|\right). \tag{4}$$

Proof. Note that $M - S^{(t)} = L^* + E^{(t)}$. Now, using Lemmas 2 and 4, we have:

$$\left|\lambda_{k+1} - \sigma_{k+1}^*\right| \leq \left\|E^{(t)}\right\|_2 \leq \alpha n \left\|E^{(t)}\right\|_\infty \leq 8\mu^2 r\alpha\gamma_t,$$

where $\gamma_t := \left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^{t-1}|\sigma_k^*|\right)$. That is, $\left||\lambda_{k+1}| - |\sigma_{k+1}^*|\right| \leq 8\mu^2 r\alpha\gamma_t$. Similarly, $||\lambda_k| - |\sigma_k^*|| \leq 8\mu^2 r\alpha\gamma_t$. So we have:

$$\begin{aligned}\left|\left(|\lambda_{k+1}| + \left(\frac{1}{2}\right)^t |\lambda_k|\right) - \left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^t |\sigma_k^*|\right)\right| &\leq 8\mu^2 r\alpha\gamma_t\left(1 + \left(\frac{1}{2}\right)^t\right) \\ &\leq 16\mu^2 r\alpha\gamma_t \\ &\leq \frac{1}{8}\left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^t |\sigma_k^*|\right),\end{aligned}$$

where the last inequality follows from the bound $\alpha \leq \frac{1}{512\mu^2 r}$. □

The following lemma shows that under the same assumptions as in Lemma 6, we can obtain a bound on the ℓ_∞ norm of $L^{(t+1)} - L^*$. This is the most crucial step in our analysis since we bound ℓ_∞ norm of errors which are quite hard to obtain.

Lemma 7. *Assume the notation of Lemma 6. Also, let $L^{(t)}, S^{(t)}$ be the t^{th} iterates of k^{th} stage of Algorithm 1 and $L^{(t+1)}, S^{(t+1)}$ be the $(t+1)^{th}$ iterates of the same stage. Also, recall that $E^{(t)} := S^* - S^{(t)}$ and $E^{(t+1)} := S^* - S^{(t+1)}$. Suppose further that*

1. *$\left\|E^{(t)}\right\|_\infty \leq \frac{8\mu^2 r}{n}\left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^{t-1}|\sigma_k^*|\right)$, and*
2. *Supp $\left(E^{(t)}\right) \subseteq$ Supp (S^*).*

Then, we have:

$$\left\|L^{(t+1)} - L^*\right\|_\infty \leq \frac{2\mu^2 r}{n}\left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^t |\sigma_k^*|\right).$$

Proof. Let $L^{(t+1)} = P_k(M - S^{(t)}) = U\Lambda U^\top$ be the eigenvalue decomposition of $L^{(t+1)}$. Also, recall that

$M - S^{(t)} = L^* + E^{(t)}$. Then, for every eigenvector $\boldsymbol{u}_i$ of $L^{(t+1)}$, we have

$$\begin{aligned} \left(L^* + E^{(t)}\right)\boldsymbol{u}_i &= \lambda_i \boldsymbol{u}_i, \\ \left(I - \frac{E^{(t)}}{\lambda_i}\right)\boldsymbol{u}_i &= \frac{1}{\lambda_i} L^* \boldsymbol{u}_i, \\ \boldsymbol{u}_i &= \left(I - \frac{E^{(t)}}{\lambda_i}\right)^{-1} \frac{L^* \boldsymbol{u}_i}{\lambda_i} \\ &= \left(I + \frac{E^{(t)}}{\lambda_i} + \left(\frac{E^{(t)}}{\lambda_i}\right)^2 + \dots\right) \frac{L^* \boldsymbol{u}_i}{\lambda_i}. \end{aligned} \tag{5}$$

Note that we used Lemmas 2 and 4 to guarantee the existence of $\left(I - \frac{E^{(t)}}{\lambda_i}\right)^{-1}$. Hence,

$$U\Lambda U^\top - L^* = \left(L^* U \Lambda^{-1} U^\top L^* - L^*\right) + \sum_{p+q \geq 1} \left(E^{(t)}\right)^p L^* U \Lambda^{-(p+q+1)} U^\top L^* \left(E^{(t)}\right)^q.$$

By triangle inequality, we have

$$\begin{aligned} \left\|U\Lambda U^\top - L^*\right\|_\infty \leq \left\|L^* U \Lambda^{-1} U^\top L^* - L^*\right\|_\infty \\ + \sum_{p+q \geq 1} \left\|\left(E^{(t)}\right)^p L^* U \Lambda^{-(p+q+1)} U^\top L^* \left(E^{(t)}\right)^q\right\|_\infty. \end{aligned} \tag{6}$$

We now bound the two terms on the right hand side above.

We note that,

$$\begin{aligned} &\left\|L^* U \Lambda^{-1} U^\top L^* - L^*\right\|_\infty \\ &= \max_{ij} \boldsymbol{e}_i{}^\top \left(U^* \Sigma^* (U^*)^\top U \Lambda^{-1} U^\top U^* \Sigma^* (U^*)^\top - U^* \Sigma^* (U^*)^\top\right) \boldsymbol{e}_j \\ &= \max_{ij} \boldsymbol{e}_i{}^\top U^* \left(\Sigma^* (U^*)^\top U \Lambda^{-1} U^\top U^* \Sigma^* - \Sigma^*\right) (U^*)^\top \boldsymbol{e}_j \\ &\leq \max_{ij} \|\boldsymbol{e}_i{}^\top U^*\| \cdot \|\boldsymbol{e}_j{}^\top U^*\| \cdot \|U^* \Sigma^* (U^*)^\top U \Lambda^{-1} U^\top U^* \Sigma^* (U^*)^\top - U^* \Sigma^* (U^*)^\top\|_2 \\ &\leq \frac{\mu^2 r}{n} \|L^* U \Lambda^{-1} U^\top L^* - L^*\|_2, \end{aligned} \tag{7}$$

where we denote $U^* \Sigma^* (U^*)^\top$ to be the SVD of L^*. Let $L^* + E^{(t)} = U\Lambda U^\top + \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top$ be the eigenvalue decomposition of $L^* + E^{(t)}$. Note that $\widetilde{U}^\top U = 0$. Recall that, $U\Lambda U^\top = P_k(M - S^{(t)}) = P_k(L^* + E^{(t)}) = L^{(t+1)}$. Also note that,

$$\begin{aligned} &L^* U \Lambda^{-1} U^\top L^* - L^* \\ &= \left(U\Lambda U^\top + \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top - E^{(t)}\right) U \Lambda^{-1} U^\top \left(U\Lambda U^\top + \widetilde{U}\widetilde{\Lambda}\widetilde{U}^T - E^{(t)}\right) - L^*, \\ &= \left(UU^\top - \left(E^{(t)}\right) U \Lambda^{-1} U^\top\right) \left(U\Lambda U^\top + \widetilde{U}\widetilde{\Lambda}\widetilde{U}^T - E^{(t)}\right) - L^*, \\ &= -UU^\top E^{(t)} - E^{(t)} UU^\top - E^{(t)} U \Lambda^{-1} U^\top {E^{(t)}}^\top - \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top + E^{(t)}. \end{aligned} \tag{8}$$

Hence, using Lemma 8, we have:

$$\begin{aligned} \|L^* U \Lambda^{-1} U^\top L^* - L^*\|_2 &\leq 3\|E^{(t)}\|_2 + \frac{\|E^{(t)}\|_2^2}{|\lambda_k|} + |\lambda_{k+1}| \\ &\leq \left|\sigma_{k+1}^*\right| + 5\left\|E^{(t)}\right\|_2. \end{aligned} \tag{9}$$

Combining (7) and (9), we have:

$$\left\|L^*U\Lambda^{-1}U^\top L^* - L^*\right\|_\infty \le \frac{\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right| + 5\left\|E^{(t)}\right\|_2\right) \tag{10}$$

Now, we will bound the $(p,q)^{\text{th}}$ term of $\sum_{p+q\ge 1}\left\|\left(E^{(t)}\right)^p L^*U\Lambda^{-(p+q+1)}U^\top L^*\left(E^{(t)}\right)^q\right\|_\infty$:

$$\begin{aligned}
&\left\|(E^{(t)})^p L^*U\Lambda^{-(p+q+1)}U^\top L^*(E^{(t)})^q\right\|_\infty \\
&\quad= \max_{ij} \boldsymbol{e}_i{}^\top\left((E^{(t)})^p L^*U\Lambda^{-(p+q+1)}U^\top L^*(E^{(t)})^q\right)\boldsymbol{e}_j, \\
&\quad\le \max_{ij}\left\|\boldsymbol{e}_i{}^\top(E^{(t)})^pU^*\right\|_2\left\|\boldsymbol{e}_j{}^\top(E^{(t)})^qU^*\right\|_2\left\|L^*U\Lambda^{-(p+q+1)}U^\top L^*\right\|_2, \\
&\quad\overset{\zeta_1}{\le} \frac{\mu^2 r}{n}\left(\alpha n\left\|E^{(t)}\right\|_\infty\right)^p\left(\alpha n\left\|E^{(t)}\right\|_\infty\right)^q\left\|L^*U\Lambda^{-(p+q+1)}U^\top L^*\right\|_2,
\end{aligned} \tag{11}$$

where ζ_1 follows from Lemma 5 and the incoherence of L^*. Now, similar to (8), we have:

$$\begin{aligned}
&\left\|L^*U\Lambda^{-(p+q+1)}U^\top L^*\right\|_2 \\
&\quad= \left\|U\Lambda^{-(p+q-1)}U^\top - E^{(t)}U\Lambda^{-(p+q)}U^\top - U\Lambda^{-(p+q)}U^\top E^{(t)} + E^{(t)}U\Lambda^{-(p+q+1)}U^\top E^{(t)}\right\|_2, \\
&\quad\le \|\Lambda^{-(p+q-1)}\|_2 + 2\|E^{(t)}\|_2\|\Lambda^{-(p+q)}\|_2 + \|E^{(t)}\|_2^2\|\Lambda^{-(p+q+1)}\|_2, \\
&\quad\le |\lambda_k|^{-(p+q-1)}\left(1 + 2\frac{\|E^{(t)}\|_2}{|\lambda_k|} + \frac{\|E^{(t)}\|_2^2}{\lambda_k^2}\right) = |\lambda_k|^{-(p+q-1)}\left(1+\frac{\|E^{(t)}\|_2}{|\lambda_k|}\right)^2, \\
&\quad\le |\lambda_k|^{-(p+q-1)}\left(1+\frac{\left\|E^{(t)}\right\|_2}{|\lambda_k|}\right)^2, \\
&\quad\overset{\zeta_1}{\le} |\lambda_k|^{-(p+q-1)}\left(1+\frac{17\mu^2 r\alpha\left|\sigma^*_k\right|}{(1-17\mu^2 r\alpha)\left|\sigma^*_k\right|}\right)^2 \le 2\left|\lambda_k\right|^{-(p+q-1)},
\end{aligned} \tag{12}$$

where ζ_1 follows from Lemma 8.

Using (11), (12), we have:

$$\left\|(E^{(t)})^p L^*U\Lambda^{-(p+q+1)}U^\top L^*(E^{(t)})^q\right\|_\infty \le 2\alpha\mu^2 r\left\|E^{(t)}\right\|_\infty\left(\frac{\alpha n\left\|E^{(t)}\right\|_\infty}{|\lambda_k|}\right)^{p+q-1}. \tag{13}$$

Using the above bound, and the assumption on $\left\|E^{(t)}\right\|_\infty$:

$$\left\|E^{(t)}\right\|_\infty \le \frac{8\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right| + \left(\frac{1}{2}\right)^{t-1}\left|\sigma^*_k\right|\right) \le \frac{17\mu^2 r}{n}\left|\sigma^*_k\right|,$$

we have:

$$\begin{aligned}
&\sum_{p+q\ge 1}\left\|\left(E^{(t)}\right)^p L^*U\Lambda^{-(p+q+1)}U^\top L^*\left(E^{(t)}\right)^q\right\|_\infty \\
&\qquad\le 2\mu^2 r\alpha\left\|E^{(t)}\right\|_\infty\sum_{p+q\ge 1}\left(\frac{\alpha n\left\|E^{(t)}\right\|_\infty}{|\lambda_k|}\right)^{p+q-1} \\
&\qquad\le 2\mu^2 r\alpha\left\|E^{(t)}\right\|_\infty\left(\frac{1}{1-\frac{17\mu^2\alpha r}{1-17\mu^2\alpha\cdot r}}\right)^2 \\
&\qquad\le 2\mu^2 r\alpha\left\|E^{(t)}\right\|_\infty\left(\frac{1}{1-34\mu^2 r\alpha}\right)^2 \\
&\qquad\le 4\mu^2 r\alpha\left\|E^{(t)}\right\|_\infty.
\end{aligned} \tag{14}$$

Combining (6), (10), (14), we have:

$$\begin{aligned}\|U\Lambda U^\top - L^*\|_\infty &\le \frac{\mu^2 r}{n}\left(|\sigma^*_{k+1}| + 5\left\|E^{(t)}\right\|_2 + 4\mu^2 r\alpha n\left\|E^{(t)}\right\|_\infty\right)\\ &\le \frac{2\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^t |\sigma^*_k|\right),\end{aligned}$$

where we used Lemma 4 and the assumption on $\left\|E^{(t)}\right\|_\infty$. □

We used the following technical lemma in the proof of Lemma 7.

Lemma 8. *Assume the notation of Lemma 7. Suppose further that*

1. $\left\|E^{(t)}\right\|_\infty \le \frac{8\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^{t-1}|\sigma^*_k|\right)$, *and*
2. *Supp* $\left(E^{(t)}\right) \subseteq$ *Supp* (S^*).

Then we have:

$$\left\|E^{(t)}\right\|_2 \le 17\mu^2 r\alpha|\sigma^*_k|, \quad |\lambda_k| \ge |\sigma^*_k|\left(1 - 17\mu^2 r\alpha\right), \quad \textit{and} \quad |\lambda_{k+1}| \le |\sigma^*_{k+1}| + \left\|E^{(t)}\right\|_2.$$

Proof. Using Lemmas 4 and 2, we have:

$$|\lambda_i - \sigma^*_i| \le \|E^{(t)}\|_2 \le \alpha n\left\|E^{(t)}\right\|_\infty.$$

The result follows by using the bound on $\left\|E^{(t)}\right\|_\infty$. □

The following lemma bounds the support of $E^{(t+1)}$ and $\left\|E^{(t+1)}\right\|_\infty$, using an assumption on $\left\|L^{(t+1)} - L^*\right\|_\infty$.

Lemma 9. *Assume the notation of Lemma 7. Suppose*

$$\left\|L^{(t+1)} - L^*\right\|_\infty \le \frac{2\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^t |\sigma^*_k|\right).$$

Then, we have:

1. *Supp* $\left(E^{(t+1)}\right) \subseteq$ *Supp* (S^*).
2. $\left\|E^{(t+1)}\right\|_\infty \le \frac{7\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^t |\sigma^*_k|\right)$, *and*

Proof. We first prove the first conclusion. Recall that,

$$S^{(t+1)} = H_\zeta(M - L^{(t+1)}) = H_\zeta(L^* - L^{(t+1)} + S^*),$$

where $\zeta = \frac{4\mu^2 r}{n}\left(|\lambda_{k+1}| + \left(\frac{1}{2}\right)^t |\lambda_k|\right)$ is as defined in Algorithm 1 and $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of $M - S^{(t)}$ such that $|\lambda_1| \ge \cdots \ge |\lambda_n|$.

If $S^*_{ij} = 0$ then $E^{(t+1)}_{ij} = \mathbb{1}_{\left\{\left|L^*_{ij} - L^{(t+1)}_{ij}\right| > \zeta\right\}} \cdot (L^*_{ij} - L^{(t+1)}_{ij})$. The first part of the lemma now follows by using the assumption that $\left\|L^{(t+1)} - L^*\right\|_\infty \le \frac{2\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^t |\sigma^*_k|\right) \overset{(\zeta_1)}{\le} \frac{4\mu^2 r}{n}\left(|\lambda^*_{k+1}| + \left(\frac{1}{2}\right)^t |\lambda^*_k|\right) = \zeta$, where (ζ_1) follows from Lemma 6.

We now prove the second conclusion. We consider the following two cases:

1. $\left|M_{ij} - L^{(t+1)}_{ij}\right| > \zeta$: Here, $S^{(t+1)}_{ij} = S^*_{ij} + L^*_{ij} - L^{(t+1)}_{ij}$. Hence, $|S^{(t+1)}_{ij} - S^*_{ij}| \le |L^*_{ij} - L^{(t+1)}_{ij}| \le \frac{2\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^t |\sigma^*_k|\right)$.

2. $\left|M_{ij} - L_{ij}^{(t+1)}\right| \leq \zeta$: In this case, $S_{ij}^{(t+1)} = 0$ and $\left|S_{ij}^* + L_{ij}^* - L_{ij}^{(t+1)}\right| \leq \zeta$. So we have, $\left|E_{ij}^{(t+1)}\right| = \left|S_{ij}^*\right| \leq \zeta + \left|L_{ij}^* - L_{ij}^{(t+1)}\right| \leq \frac{7\mu^2 r}{n}\left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^t \left|\sigma_k^*\right|\right)$. The last inequality above follows from Lemma 6.

This proves the lemma. □

We are now ready to prove Lemma 1. In fact, we prove the following stronger version.

Proof of Lemma 1. Recall that in the k^{th} stage, the update $L^{(t+1)}$ is given by: $L^{(t+1)} = P_k(M - S^{(t)})$ and $S^{(t+1)}$ is given by: $S^{(t+1)} = H_\zeta(M - L^{(t+1)})$. Also, recall that $E^{(t)} := S^* - S^{(t)}$ and $E^{(t+1)} := S^* - S^{(t+1)}$.

We prove the lemma by induction on both k and t. For the base case ($k = 1$ and $t = -1$), we first note that the first inequality on $\left\|L^{(0)} - L^*\right\|_\infty$ is trivially satisfied. Due to the thresholding step (step 3 in Algorithm 1) and the incoherence assumption on L^*, we have:

$$\begin{aligned}\left\|E^{(0)}\right\|_\infty &\leq \frac{8\mu^2 r}{n}\left(\sigma_2^* + 2\sigma_1^*\right), \text{ and} \\ \text{Supp}\left(E^{(0)}\right) &\subseteq \text{Supp}\left(S^*\right).\end{aligned}$$

So the base case of induction is satisfied.

We first do the inductive step over t (for a fixed k). By inductive hypothesis we assume that: a) $\left\|E^{(t)}\right\|_\infty \leq \frac{8\mu^2 r}{n}\left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^{t-1}\left|\sigma_k^*\right|\right)$, b) $\text{Supp}\left(E^{(t)}\right) \subseteq \text{Supp}\left(S^*\right)$. Then by Lemma 7, we have:

$$\left\|L^{(t+1)} - L^*\right\|_\infty \leq \frac{2\mu^2 r}{n}\left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^{t+1}\left|\sigma_k^*\right|\right).$$

Lemma 9 now tells us that

1. $\left\|E^{(t+1)}\right\|_\infty \leq \frac{8\mu^2 r}{n}\left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^t\left|\sigma_k^*\right|\right)$, and
2. $\text{Supp}\left(E^{(t+1)}\right) \subseteq \text{Supp}\left(S^*\right)$.

This finishes the induction over t. Note that we show a stronger bound than necessary on $\left\|E^{(t+1)}\right\|_\infty$.

We now do the induction over k. Suppose the hypothesis holds for stage k. Let T denote the number of iterations in each stage. We first obtain a lower bound on T. Since

$$\left\|M - S^{(0)}\right\|_2 \geq \left\|L^*\right\|_2 - \left\|E^{(0)}\right\|_2 \geq \left|\sigma_1^*\right| - \alpha n\left\|E^{(0)}\right\|_\infty \geq \frac{3}{4}\left|\sigma_1^*\right|,$$

we see that $T \geq 10\log\left(3\mu^2 r\left|\sigma_1^*\right|/\epsilon\right)$. So, at the end of stage k, we have:

1. $\left\|E^{(T)}\right\|_\infty \leq \frac{7\mu^2 r}{n}\left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^T\left|\sigma_k^*\right|\right) \leq \frac{7\mu^2 r\left|\sigma_{k+1}^*\right|}{n} + \frac{\epsilon}{10n}$, and
2. $\text{Supp}\left(E^{(T)}\right) \subseteq \text{Supp}\left(S^*\right)$.

Lemmas 4 and 2 tell us that $\left|\sigma_{k+1}\left(M - S^{(T)}\right) - \left|\sigma_{k+1}^*\right|\right| \leq \left\|E^{(T)}\right\|_2 \leq \alpha\left(7\mu^2 r\left|\sigma_{k+1}^*\right| + \epsilon\right)$. We will now consider two cases:

1. **Algorithm 1 terminates:** This means that $\beta\sigma_{k+1}\left(M - S^{(T)}\right) < \frac{\epsilon}{2n}$ which then implies that $\left|\sigma_{k+1}^*\right| < \frac{\epsilon}{6\mu^2 r}$. So we have:

$$\left\|\widehat{L} - L^*\right\|_\infty = \left\|L^{(T)} - L^*\right\|_\infty \leq \frac{2\mu^2 r}{n}\left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^T\left|\sigma_k^*\right|\right) \leq \frac{\epsilon}{5n}.$$

This proves the statement about $\widehat{L}$. A similar argument proves the claim on $\left\|\widehat{S} - S^*\right\|_\infty$. The claim on $\text{Supp}\left(\widehat{S}\right)$ follows since $\text{Supp}\left(E^{(T)}\right) \subseteq \text{Supp}\left(S^*\right)$.

2. **Algorithm 1 continues to stage $(k+1)$:** This means that $\beta\sigma_{k+1}\left(L^{(T)}\right) \geq \frac{\epsilon}{2n}$ which then implies that $\left|\sigma^*_{k+1}\right| > \frac{\epsilon}{8\mu^2 r}$. So we have:

$$\begin{aligned} \left\|E^{(T)}\right\|_\infty &\leq \frac{8\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right| + \left(\frac{1}{2}\right)^T \left|\sigma^*_k\right|\right) \\ &\leq \frac{8\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right| + \frac{\epsilon}{10\mu^2 rn}\right) \\ &\leq \frac{8\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right| + \frac{8\left|\sigma^*_{k+1}\right|}{10n}\right) \\ &\leq \frac{8\mu^2 r}{n}\left(\left|\sigma^*_{k+2}\right| + 2\left|\sigma^*_{k+1}\right|\right). \end{aligned}$$

   Similarly for $\left\|L^{(T)} - L^*\right\|_\infty$.

This finishes the proof. □

Proof of Theorem 1. Using Lemma 1, it suffices to show that the general case can be reduced to the case of symmetric matrices. We will now outline this reduction.

Recall that we are given an $m \times n$ matrix $M = L^* + S^*$ where L^* is the true low-rank matrix and S^* the sparse error matrix. Wlog, let $m \leq n$ and suppose $\beta m \leq n < (\beta+1)m$, for some $\beta \geq 1$. We then consider the symmetric matrices

$$\widetilde{M} = \left[\begin{array}{cc} \underbrace{\begin{matrix} 0 & 0 \\ \vdots \cdots & \vdots \\ 0 & 0 \\ M^\top \cdots & M^\top \end{matrix}}_{\beta \text{ times}} & \begin{matrix} M \\ \vdots \\ M \\ 0 \\ \end{matrix} \end{array}\right], \widetilde{L} = \left[\begin{array}{cc} \underbrace{\begin{matrix} 0 & 0 \\ \vdots \cdots & \vdots \\ 0 & 0 \\ \left(L^*\right)^\top \cdots & \left(L^*\right)^\top \end{matrix}}_{\beta \text{ times}} & \begin{matrix} L^* \\ \vdots \\ L^* \\ 0 \\ \end{matrix} \end{array}\right], \tag{15}$$

and $\widetilde{S} = \widetilde{M} - \widetilde{L}$. A simple calculation shows that $\widetilde{L}$ is incoherent with parameter $\sqrt{3}\mu$ and $\widetilde{S}$ satisfies the sparsity condition (S1) with parameter $\frac{\alpha}{\sqrt{2}}$. Moreover the iterates of AltProj with input $\widetilde{M}$ have similar expressions as in (15) in terms of the corresponding iterates with input M. This means that it suffices to obtain the same guarantees for Algorithm 1 for the symmetric case. Lemma 1 does precisely this, proving the theorem. □

B Proof of Theorem 2

In this section, we prove Theorem 2. The roadmap of the proofs in this section is essentially the same as that in Appendix A.

In what follows, we prove a number of lemmas concerning the structure of $L^{(t)}$ and $E^{(t)} := S^* - S^{(t)}$. The first lemma is a generalization of Lemma 6 and shows that the threshold in (2) is close to that with $M^{(t)}$ replaced by L^*.

Lemma 10. *Let L^*, S^*, N^* be symmetric and satisfy the assumptions of Theorem 2 and let $S^{(t)}$ be the t^{th} iterate of the k^{th} stage of Algorithm 1. Let $\sigma^*_1, \ldots, \sigma^*_r$ be the eigenvalues of L^*, such that $\left|\sigma^*_1\right| \geq \cdots \geq \left|\sigma^*_r\right|$ and $\lambda_1, \cdots, \lambda_n$ be the eigenvalues of $M - S^{(t)}$ such that $\left|\lambda_1\right| \geq \cdots \geq \left|\lambda_n\right|$. Recall that $E^{(t)} := S^* - S^{(t)}$. Suppose further that*

1. $\left\|E^{(t)}\right\|_\infty \leq \frac{8\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right| + \left(\frac{1}{2}\right)^{t-1}\left|\sigma^*_k\right| + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right)$, *and*
2. *Supp* $\left(S^{(t)}\right) \subseteq$ *Supp* $\left(S^*\right)$.

Then,

$$\frac{7}{8}\left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^t |\sigma_k^*|\right) \leq \left(|\lambda_{k+1}| + \left(\frac{1}{2}\right)^t |\lambda_k|\right) \leq \frac{9}{8}\left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^t |\sigma_k^*|\right). \tag{16}$$

Proof. Note that $M - S^{(t)} = L^* + N^* + E^{(t)}$. Now, using Lemmas 2 and 4, we have:

$$\left|\lambda_{k+1} - \sigma_{k+1}^*\right| \leq \left\|E^{(t)}\right\|_2 \leq \alpha n \left\|E^{(t)}\right\|_\infty \leq 8\mu^2 r \alpha \gamma_t,$$

where $\gamma_t = \left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^{t-1} |\sigma_k^*| + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right)$. That is, $\left||\lambda_{k+1}| - |\sigma_{k+1}^*|\right| \leq 8\mu^2 r\alpha\gamma_t$. Similarly, $||\lambda_k| - |\sigma_k^*|| \leq 8\mu^2 r\alpha\gamma_t$. So we have:

$$\begin{aligned}\left|\left(|\lambda_{k+1}| + \left(\frac{1}{2}\right)^t |\lambda_k|\right) - \left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^t |\sigma_k^*|\right)\right| &\leq 8\mu^2 r\alpha\gamma_t\left(1 + \left(\frac{1}{2}\right)^t\right) \\ &\leq 16\mu^2 r\alpha\gamma_t \\ &\leq \frac{1}{8}\left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^t |\sigma_k^*|\right),\end{aligned}$$

where the last inequality follows from the bound $\alpha \leq \frac{1}{512\mu^2 r}$ and the assumption on $\left\|N^*\right\|_\infty$. □

The following lemma shows that under the same assumptions as in Lemma 6, we can obtain a bound on the ℓ_∞ norm of $L^{(t+1)} - L^*$. This is the most crucial step in our analysis since we bound ℓ_∞ norm of errors which are quite hard to obtain.

Lemma 11. *Assume the notation of Lemma 6. Also, let $L^{(t)}, S^{(t)}$ be the t^{th} iterates of k^{th} stage of Algorithm 1 and $L^{(t+1)}, S^{(t+1)}$ be the $(t+1)^{th}$ iterates of the same stage. Also, recall that $E^{(t)} := S^* - S^{(t)}$ and $E^{(t+1)} := S^* - S^{(t+1)}$. Suppose further that*

1. $\left\|E^{(t)}\right\|_\infty \leq \frac{8\mu^2 r}{n}\left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^{t-1} |\sigma_k^*| + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right)$, *and*
2. *Supp* $\left(E^{(t)}\right) \subseteq$ *Supp* (S^*).

Then, we have:

$$\left\|L^{(t+1)} - L^*\right\|_\infty \leq \frac{2\mu^2 r}{n}\left(|\sigma_{k+1}^*| + \left(\frac{1}{2}\right)^t |\sigma_k^*| + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right).$$

Proof. Let $L^{(t+1)} = P_k(M - S^{(t)}) = U\Lambda U^\top$ be the eigenvalue decomposition of $L^{(t+1)}$. Also, recall that $M - S^{(t)} = L^* + N^* + E^{(t)}$. Then, for every eigenvector $\boldsymbol{u}_i$ of $L^{(t+1)}$, we have

$$\begin{aligned}\left(L^* + N^* + E^{(t)}\right)\boldsymbol{u}_i &= \lambda_i \boldsymbol{u}_i, \\ \left(I - \frac{E^{(t)}}{\lambda_i}\right)\boldsymbol{u}_i &= \frac{1}{\lambda_i}\left(L^* + N^*\right)\boldsymbol{u}_i, \\ \boldsymbol{u}_i &= \left(I - \frac{E^{(t)}}{\lambda_i}\right)^{-1} \frac{\left(L^* + N^*\right)\boldsymbol{u}_i}{\lambda_i} \\ &= \left(I + \frac{E^{(t)}}{\lambda_i} + \left(\frac{E^{(t)}}{\lambda_i}\right)^2 + \ldots\right)\frac{\left(L^* + N^*\right)\boldsymbol{u}_i}{\lambda_i}.\end{aligned}$$

Note that we used Lemmas 2 and 4 to guarantee the existence of $\left(I - \frac{E^{(t)}}{\lambda_i}\right)^{-1}$. Hence,

$$\begin{aligned}U\Lambda U^\top - L^* = &\left(\left(L^* + N^*\right)U\Lambda^{-1}U^\top\left(L^* + N^*\right) - L^*\right) \\ &+ \sum_{p+q\geq 1}(S^{(t)})^p\left(L^* + N^*\right)U\Lambda^{-(p+q+1)}U^\top\left(L^* + N^*\right)(S^{(t)})^q.\end{aligned}$$

By triangle inequality, we have

$$\begin{aligned}\left\|U\Lambda U^\top - L^*\right\|_\infty \le{}& \left\|\left(L^*+N^*\right)U\Lambda^{-1}U^\top\left(L^*+N^*\right) - L^*\right\|_\infty \\ &+ \sum_{p+q\ge 1}\left\|(S^{(t)})^p\left(L^*+N^*\right)U\Lambda^{-(p+q+1)}U^\top\left(L^*+N^*\right)(S^{(t)})^q\right\|_\infty .\end{aligned}\tag{17}$$

We now bound the two terms on the right hand side above.

For the first term, we again use triangle inequality to obtain

$$\begin{aligned}\left\|\left(L^*+N^*\right)U\Lambda^{-1}U^\top\left(L^*+N^*\right) - L^*\right\|_\infty \le{}& \left\|L^*U\Lambda^{-1}U^\top L^* - L^*\right\|_\infty + \left\|N^*U\Lambda^{-1}U^\top L^*\right\|_\infty \\ &+ \left\|L^*U\Lambda^{-1}U^\top N^*\right\|_\infty + \left\|N^*U\Lambda^{-1}U^\top N^*\right\|_\infty .\end{aligned}\tag{18}$$

We note that,

$$\begin{aligned}&\left\|L^*U\Lambda^{-1}U^\top L^* - L^*\right\|_\infty \\ &= \max_{ij}\boldsymbol{e}_i{}^\top\left(U^*\Sigma^*(U^*)^\top U\Lambda^{-1}U^\top U^*\Sigma^*(U^*)^\top - U^*\Sigma^*(U^*)^\top\right)\boldsymbol{e}_j \\ &= \max_{ij}\boldsymbol{e}_i{}^\top U^*\left(\Sigma^*(U^*)^\top U\Lambda^{-1}U^\top U^*\Sigma^* - \Sigma^*\right)(U^*)^\top\boldsymbol{e}_j \\ &\le \max_{ij}\|\boldsymbol{e}_i{}^\top U^*\|\cdot\|\boldsymbol{e}_j{}^\top U^*\|\cdot\|U^*\Sigma^*(U^*)^\top U\Lambda^{-1}U^\top U^*\Sigma^*(U^*)^\top - U^*\Sigma^*(U^*)^\top\|_2 \\ &\le \frac{\mu^2 r}{n}\|L^*U\Lambda^{-1}U^\top L^* - L^*\|_2,\end{aligned}\tag{19}$$

where we denote $U^*\Sigma^*(U^*)^\top$ to be the SVD of L^*. Let $L^*+N^*+E^{(t)} = U\Lambda U^\top + \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top$ be the eigenvalue decomposition of $L^*+N^*+E^{(t)}$. Note that $\widetilde{U}^\top U = 0$. Recall that, $U\Lambda U^\top = P_k(M^{(t)}) = L^{(t)}$. Also note that,

$$\begin{aligned}&L^*U\Lambda^{-1}U^\top L^* - L^* \\ &= (U\Lambda U^\top + \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top - N^* - E^{(t)})U\Lambda^{-1}U^\top(U\Lambda U^\top + \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top - N^* - E^{(t)}) - L^*, \\ &= (UU^\top - \left(N^*+E^{(t)}\right)U\Lambda^{-1}U^\top)(U\Lambda U^\top + \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top - N^* - E^{(t)}) - L^*, \\ &= U\Lambda U^\top - UU^\top\left(N^*+E^{(t)}\right) - \left(N^*+E^{(t)}\right)UU^\top \\ &- \left(N^*+E^{(t)}\right)U\Lambda^{-1}U^\top\left(N^*+E^{(t)}\right)^\top - U\Lambda U^\top - \widetilde{U}\widetilde{\Lambda}\widetilde{U}^\top + N^* + E^{(t)}.\end{aligned}\tag{20}$$

Hence, using Lemma 12, we have:

$$\begin{aligned}\left\|L^*U\Lambda^{-1}U^\top L^* - L^*\right\|_2 &\le 3\left\|N^*+E^{(t)}\right\|_2 + \frac{\left\|N^*+E^{(t)}\right\|_2^2}{|\lambda_k|} + |\lambda_{k+1}| \\ &\le \left|\sigma^*_{k+1}\right| + 4\left\|N^*+E^{(t)}\right\|_2 + \frac{\left\|N^*+E^{(t)}\right\|_2^2}{(1-17\mu^2 r\alpha)\left|\sigma^*_k\right|}.\end{aligned}\tag{21}$$

Using (19), (21), and Lemma 12:

$$\left\|L^*U\Lambda^{-1}U^\top L^* - L^*\right\|_\infty \le \frac{\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right| + 7\left\|N^*\right\|_2 + 5\left\|E^{(t)}\right\|_2\right)\tag{22}$$

Coming to the second term of (18), we have:

$$\begin{aligned}&\left\|N^*U\Lambda^{-1}U^\top L^*\right\|_\infty \\ &= \max_{i,j}\boldsymbol{e}_i{}^\top N^*U\Lambda^{-1}U^\top L^*\boldsymbol{e}_j \\ &\le \max_i\left\|\boldsymbol{e}_i{}^\top N^*U\right\|_2\left\|\Lambda^{-1}U^\top U^*\Sigma^*\right\|_2\left\|(U^*)^\top\boldsymbol{e}_j\right\|_2 \\ &\le \sqrt{n}\left\|N^*\right\|_\infty\left\|\Lambda^{-1}U^\top U^*\Sigma^*\right\|_2\frac{\mu\sqrt{r}}{\sqrt{n}} = \mu\sqrt{r}\left\|N^*\right\|_\infty\left\|U\Lambda^{-1}U^\top U^*\Sigma^*(U^*)^\top\right\|_2 .\end{aligned}\tag{23}$$

Using an expansion along the lines of (20), we see that

$$\left\|U\Lambda^{-1}U^\top U^*\Sigma^*(U^*)^\top\right\|_2 \leq 1 + \frac{\left\|N^* + E^{(t)}\right\|_2}{|\lambda_k|} \leq 1 + \frac{\|N^*\|_2 + \left\|E^{(t)}\right\|_2}{(1 - 17\mu^2 r \cdot \alpha)\left|\sigma_k^*\right|}$$
$$\leq 2 + \frac{\left\|E^{(t)}\right\|_2}{(1 - 17\mu^2 r\alpha)\left|\sigma_k^*\right|}.$$

Plugging this in (23) gives us

$$\left\|N^* U\Lambda^{-1}U^\top L^*\right\|_\infty \leq 3\mu\sqrt{r}\left\|N^*\right\|_\infty. \tag{24}$$

A similar argument as in (23) gives us the following bound on the last term in (18):

$$\left\|N^* U\Lambda^{-1}U^\top N^*\right\|_\infty \leq \frac{n\left\|N^*\right\|_\infty^2}{|\lambda_k|} \leq \left\|N^*\right\|_\infty. \tag{25}$$

Plugging (22), (24) and (25), we obtain:

$$\begin{aligned}&\left\|(L^* + N^*)\,U\Lambda^{-1}U^\top\,(L^* + N^*) - L^*\right\|_\infty \\ &\qquad\leq \frac{\mu^2 r}{n}\left(\left|\sigma_{k+1}^*\right| + 7\left\|N^*\right\|_2 + 7\left\|E^{(t)}\right\|_2 + \frac{7n}{\sqrt{r}}\left\|N^*\right\|_\infty\right).\end{aligned} \tag{26}$$

Next, we analyze $\sum_{p+q\geq 1}\left\|(E^{(t)})^p(L^* + N^*)U\Lambda^{-(p+q+1)}U^\top(L^* + N^*)(E^{(t)})^q\right\|_\infty$. This can again be bounded by four quantities:

$$\begin{aligned}&\left\|(E^{(t)})^p(L^* + N^*)U\Lambda^{-(p+q+1)}U^\top(L^* + N^*)(E^{(t)})^q\right\|_\infty \\ &\quad\leq \left\|(E^{(t)})^p L^* U\Lambda^{-(p+q+1)}U^\top L^*(E^{(t)})^q\right\|_\infty + \left\|(E^{(t)})^p N^* U\Lambda^{-(p+q+1)}U^\top L^*(E^{(t)})^q\right\|_\infty &(27)\\ &\quad + \left\|(E^{(t)})^p L^* U\Lambda^{-(p+q+1)}U^\top N^*(E^{(t)})^q\right\|_\infty + \left\|(E^{(t)})^p N^* U\Lambda^{-(p+q+1)}U^\top N^*(E^{(t)})^q\right\|_\infty. &(28)\end{aligned}$$

We bound the first term above:

$$\begin{aligned}&\left\|(E^{(t)})^p L^* U\Lambda^{-(p+q+1)}U^\top L^*(E^{(t)})^q\right\|_\infty \\ &\qquad= \max_{ij} \boldsymbol{e}_i{}^\top\left((E^{(t)})^p L^* U\Lambda^{-(p+q+1)}U^\top L^*(E^{(t)})^q\right)\boldsymbol{e}_j, \\ &\qquad\leq \max_{ij}\left\|\boldsymbol{e}_i{}^\top(E^{(t)})^p U^*\right\|_2\left\|\boldsymbol{e}_j{}^\top(E^{(t)})^q U^*\right\|_2\left\|L^* U\Lambda^{-(p+q+1)}U^\top L^*\right\|_2, \\ &\qquad\overset{(\zeta_1)}{\leq} \frac{\mu^2 r}{n}\left(\alpha n\left\|E^{(t)}\right\|_\infty\right)^p\left(\alpha n\left\|E^{(t)}\right\|_\infty\right)^q\left\|L^* U\Lambda^{-(p+q+1)}U^\top L^*\right\|_2,\end{aligned} \tag{29}$$

where (ζ_1) follows from Lemma 5 and the incoherence of L^*. Now, similar to (20), we have:

$$\begin{aligned}
&\left\|L^*U\Lambda^{-(p+q+1)}U^\top L^*\right\|_2 \\
&\quad= \Big\|U\Lambda^{-(p+q-1)}U^\top - \left(N^* + E^{(t)}\right)U\Lambda^{-(p+q)}U^\top - U\Lambda^{-(p+q)}U^\top\left(N^* + E^{(t)}\right) \\
&\qquad\quad + \left(N^* + E^{(t)}\right)U\Lambda^{-(p+q+1)}U^\top\left(N^* + E^{(t)}\right)\Big\|_2, \\
&\quad\le \|\Lambda^{-(p+q-1)}\|_2 + 2\|N^* + E^{(t)}\|_2\|\Lambda^{-(p+q)}\|_2 + \|N^* + E^{(t)}\|_2^2\|\Lambda^{-(p+q+1)}\|_2, \\
&\quad\le |\lambda_k|^{-(p+q-1)}\left(1 + 2\frac{\|N^* + E^{(t)}\|_2}{|\lambda_k|} + \frac{\|N^* + E^{(t)}\|_2^2}{\lambda_k^2}\right), \\
&\quad= |\lambda_k|^{-(p+q-1)}\left(1 + \frac{\|N^* + E^{(t)}\|_2}{|\lambda_k|}\right)^2, \\
&\quad\le |\lambda_k|^{-(p+q-1)}\left(1 + \frac{\|N^*\|_2 + \left\|E^{(t)}\right\|_2}{|\lambda_k|}\right)^2, \\
&\quad\overset{(\zeta_1)}{\le} |\lambda_k|^{-(p+q-1)}\left(1 + \frac{\|N^*\|_2 + 17\mu^2 r\alpha\left|\sigma_k^*\right|}{(1 - 17\mu^2 r\alpha)\left|\sigma_k^*\right|}\right)^2 \le 2\left|\lambda_k\right|^{-(p+q-1)},
\end{aligned} \tag{30}$$

where (ζ_1) follows from Lemma 12 and the bound on $\|N^*\|_\infty$.

Using (29), (30), we have:

$$\left\|(E^{(t)})^p L^*U\Lambda^{-(p+q+1)}U^\top L^*(E^{(t)})^q\right\|_\infty \le 2\alpha\mu^2 r\left\|E^{(t)}\right\|_\infty\left(\frac{\alpha n\left\|E^{(t)}\right\|_\infty}{|\lambda_k|}\right)^{p+q-1}. \tag{31}$$

Coming to the second term of (28), we have

$$\begin{aligned}
&\left\|(E^{(t)})^p N^*U\Lambda^{-(p+q+1)}U^\top L^*(E^{(t)})^q\right\|_\infty \\
&\qquad= \max_{i,j}\boldsymbol{e}_i{}^\top\left((E^{(t)})^p N^*U\Lambda^{-(p+q+1)}U^\top L^*(E^{(t)})^q\right)\boldsymbol{e}_j, \\
&\qquad\le \max_{ij}\left\|\boldsymbol{e}_i{}^\top(E^{(t)})^p N^*U\right\|_2\left\|\boldsymbol{e}_j{}^\top(E^{(t)})^q U^*\right\|_2\left\|\Lambda^{-(p+q+1)}U^\top L^*\right\|_2 \\
&\qquad\overset{(\zeta_1)}{\le} \frac{\mu\sqrt{r}}{\sqrt{n}}\left\|N^*U\right\|_\infty\left(\alpha n\left\|E^{(t)}\right\|_\infty\right)^p\left(\alpha n\left\|E^{(t)}\right\|_\infty\right)^q\left\|U\Lambda^{-(p+q+1)}U^\top L^*\right\|_2 \\
&\qquad\le \mu\sqrt{r}\left\|N^*\right\|_\infty\left(\alpha n\left\|E^{(t)}\right\|_\infty\right)^{p+q}\left\|U\Lambda^{-(p+q+1)}U^\top L^*\right\|_2,
\end{aligned} \tag{32}$$

where (ζ_1) follows from Lemma 5 and incoherence of U^*. Proceeding along the lines of (30), we obtain:

$$\left\|U\Lambda^{-(p+q+1)}U^\top L^*\right\|_2 \le |\lambda_k|^{-(p+q)}\left(1 + \frac{\|N^*\|_2 + \left\|E^{(t)}\right\|_2}{|\lambda_k|}\right) \le 2\left|\lambda_k\right|^{-(p+q)}.$$

Plugging the above in (32) gives us

$$\left\|(E^{(t)})^p N^*U\Lambda^{-(p+q+1)}U^\top L^*(E^{(t)})^q\right\|_\infty \le 2\mu\sqrt{r}\left(\frac{\alpha n\left\|E^{(t)}\right\|_\infty}{|\lambda_k|}\right)^{p+q}\left\|N^*\right\|_\infty. \tag{33}$$

A similar argument as in (32) gives us

$$\left\|(E^{(t)})^p N^*U\Lambda^{-(p+q+1)}U^\top N^*(E^{(t)})^q\right\|_\infty \le \frac{n\left\|N^*\right\|_\infty}{|\lambda_k|}\left(\frac{\alpha n\left\|E^{(t)}\right\|_\infty}{|\lambda_k|}\right)^{p+q}\left\|N^*\right\|_\infty.$$

Plugging the above inequality along with (31) and (33) into (28) gives us:

$$\begin{aligned}&\left\|(E^{(t)})^p(L^*+N^*)U\Lambda^{-(p+q+1)}U^\top(L^*+N^*)(E^{(t)})^q\right\|_\infty\\&\qquad\leq 2\mu^2 r\left(\alpha\left\|E^{(t)}\right\|_\infty+\frac{\|N^*\|_\infty}{\sqrt{r}}\right)\left(\frac{\alpha n\left\|E^{(t)}\right\|_\infty}{|\lambda_k|}\right)^{p+q-1}.\end{aligned}$$

Using the above bound, and the assumption on $\left\|E^{(t)}\right\|_\infty$:

$$\left\|E^{(t)}\right\|_\infty\leq\frac{8\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right|+\left(\frac{1}{2}\right)^{t-1}\left|\sigma^*_k\right|+7\left\|N^*\right\|_2+\frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right)\leq\frac{17\mu^2 r}{n}\left|\sigma^*_k\right|,$$

we have:

$$\begin{aligned}&\sum_{p+q\geq 1}\left\|(E^{(t)})^p\left(L^*+N^*\right)U\Lambda^{-(p+q+1)}U^\top\left(L^*+N^*\right)(E^{(t)})^q\right\|_\infty &\quad(34)\\&\leq 2\mu^2 r\left(\alpha\left\|E^{(t)}\right\|_\infty+\frac{\|N^*\|_\infty}{\sqrt{r}}\right)\sum_{p+q\geq 1}\left(\frac{\alpha n\left\|E^{(t)}\right\|_\infty}{|\lambda_k|}\right)^{p+q-1}\\&\leq 2\mu^2 r\left(\alpha\left\|E^{(t)}\right\|_\infty+\frac{\|N^*\|_\infty}{\sqrt{r}}\right)\left(\frac{1}{1-\frac{17\mu^2\alpha r}{1-17\mu^2\alpha\cdot r}}\right)^2\\&\leq 2\mu^2 r\left(\alpha\left\|E^{(t)}\right\|_\infty+\frac{\|N^*\|_\infty}{\sqrt{r}}\right)\left(\frac{1}{1-34\mu^2 r\alpha}\right)^2\\&\leq 4\mu^2 r\left(\alpha\left\|E^{(t)}\right\|_\infty+\frac{\|N^*\|_\infty}{\sqrt{r}}\right). &\quad(35)\end{aligned}$$

Combining (17), (26), (35), we have:

$$\begin{aligned}\left\|U\Lambda U^\top-L^*\right\|_\infty&\overset{(\zeta_1)}{\leq}\frac{\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right|+7\left\|N^*\right\|_2+11n\alpha\left\|E^{(t)}\right\|_\infty+\frac{11n}{\sqrt{r}}\left\|N^*\right\|_\infty\right)\\&\overset{(\zeta_2)}{\leq}\frac{2\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right|+\left(\frac{1}{2}\right)^{t}\left|\sigma^*_k\right|+7\left\|N^*\right\|_2+\frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right),\end{aligned}$$

where (ζ_1) follows from Lemma 4, and (ζ_2) follows from the assumption on $\left\|E^{(t)}\right\|_\infty$. □

We used the following technical lemma in the proof of Lemma 11.

Lemma 12. *Assume the notation of Lemma 11. Suppose further that*

1. $\left\|E^{(t)}\right\|_\infty\leq\frac{8\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right|+\left(\frac{1}{2}\right)^{t-1}\left|\sigma^*_k\right|+7\left\|N^*\right\|_2+\frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right)$, *and*
2. *Supp* $\left(E^{(t)}\right)\subseteq$ *Supp* (S^*).

Then we have:

$$\left\|E^{(t)}\right\|_2\leq 17\mu^2 r\alpha|\sigma^*_k|,\quad|\lambda_k|\geq|\sigma^*_k|\left(1-17\mu^2 r\alpha\right),\quad\textit{and}\quad|\lambda_{k+1}|\leq|\sigma^*_{k+1}|+\left\|E^{(t)}\right\|_2.$$

Proof. Using Lemmas 4 and 2, we have:

$$|\lambda_i-\sigma^*_i|\leq\left\|E^{(t)}\right\|_2\leq\alpha n\left\|E^{(t)}\right\|_\infty.$$

Using the bound on $\left\|E^{(t)}\right\|_\infty$ and recalling the assumption that

$$\left\|N^*\right\|_\infty\leq\frac{|\sigma^*_r|}{100}$$

finishes the proof. □

The following lemma bounds the support of $E^{(t+1)}$ and $\left\|E^{(t+1)}\right\|_\infty$, using an assumption on $\left\|L^{(t+1)} - L^*\right\|_\infty$.
Lemma 13. *Assume the notation of Lemma 11. Suppose*

$$\left\|L^{(t+1)} - L^*\right\|_\infty \leq \frac{2\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^t |\sigma^*_k| + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right).$$

Then, we have:

1. *Supp* $\left(E^{(t+1)}\right) \subseteq$ *Supp* (S^*).
2. $\left\|E^{(t+1)}\right\|_\infty \leq \frac{7\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^t |\sigma^*_k| + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right)$, *and*

Proof. We first prove the first conclusion. Recall that,

$$S^{(t+1)} = H_\zeta(M - L^{(t+1)}) = H_\zeta(L^* - L^{(t+1)} + N^* + S^*),$$

where $\zeta = \frac{4\mu^2 r}{n}\left(|\lambda_{k+1}| + \left(\frac{1}{2}\right)^t |\lambda_k|\right)$ is as defined in Algorithm 1 and $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of $M - S^{(t)}$ such that $|\lambda_1| \geq \cdots \geq |\lambda_n|$.

If $S^*_{ij} = 0$ then $E^{(t+1)}_{ij} = \mathbb{1}_{\left\{\left|L^*_{ij} - L^{(t+1)}_{ij} + N^*_{ij}\right| > \zeta\right\}} \cdot (L^*_{ij} - L^{(t+1)}_{ij} + N^*_{ij})$. The first part of the lemma now follows by using the assumption that $\left\|L^{(t+1)} - L^*\right\|_\infty \leq \frac{2\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^t |\sigma^*_k|\right) \overset{(\zeta_1)}{\leq} \frac{9\mu^2 r}{4n}\left(|\lambda^*_{k+1}| + \left(\frac{1}{2}\right)^t |\lambda^*_k|\right) = \zeta$, where (ζ_1) follows from Lemma 6, and the bound on $\left\|N^*\right\|_\infty$.

We now prove the second conclusion. We consider the following two cases:

1. $\left|M_{ij} - L^{(t+1)}_{ij}\right| > \zeta$: Here, $S^{(t+1)}_{ij} = S^*_{ij} + L^*_{ij} - L^{(t+1)}_{ij} + N^*_{ij}$. Hence, $|S^{(t+1)}_{ij} - S^*_{ij}| \leq |L^*_{ij} - L^{(t+1)}_{ij}| + |N^*_{ij}| \leq \frac{2\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^t |\sigma^*_k|\right) + \left\|N^*\right\|_\infty$.
2. $\left|M_{ij} - L^{(t+1)}_{ij}\right| \leq \zeta$: In this case, $S^{(t+1)}_{ij} = 0$ and $\left|S^*_{ij} + L^*_{ij} - L^{(t+1)}_{ij} + N^*_{ij}\right| \leq \zeta$. So we have, $\left|E^{(t+1)}_{ij}\right| = \left|S^*_{ij}\right| \leq \zeta + \left|L^*_{ij} - L^{(t+1)}_{ij}\right| + \left|N^*_{ij}\right| \leq \frac{7\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^t |\sigma^*_k|\right) + \left\|N^*\right\|_\infty$. The last inequality above follows from Lemma 6.

This proves the lemma. □

The following lemma is a generalization of Lemma 1.
Lemma 14. *Let L^*, S^*, N^* be symmetric and satisfy the assumptions of Theorem 2 and let $M^{(t)}$ and $L^{(t)}$ be the t^{th} iterates of the k^{th} stage of Algorithm 1. Let $\sigma^*_1, \ldots, \sigma^*_n$ be the eigenvalues of L^*, s.t., $|\sigma^*_1| \geq \cdots \geq |\sigma^*_r|$. Then, the following holds:*

$$\left\|L^{(t+1)} - L^*\right\|_\infty \leq \frac{2\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^t |\sigma^*_k| + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right),$$

$$\left\|E^{(t+1)}\right\|_\infty = \left\|S^* - S^{(t+1)}\right\|_\infty \leq \frac{8\mu^2 r}{n}\left(|\sigma^*_{k+1}| + \left(\frac{1}{2}\right)^{t-1} |\sigma^*_k| + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right), \text{ and}$$

$$\textit{Supp}\left(E^{(t+1)}\right) \subseteq \textit{Supp}\left(S^*\right).$$

Moreover, the outputs $\widehat{L}$ and $\widehat{S}$ of Algorithm 1 satisfy:

$$\left\|\widehat{L} - L^*\right\|_F \leq \epsilon + 2\mu^2 r\left(7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right),$$

$$\left\|\widehat{S} - S^*\right\|_\infty \leq \frac{\epsilon}{n} + \frac{8\mu^2 r}{n}\left(7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right), \text{ and}$$

$$\textit{Supp}\left(\widehat{S}\right) \subseteq \textit{Supp}\left(S^*\right).$$

Proof. Recall that in the k^{th} stage, the update $L^{(t+1)}$ is given by: $L^{(t+1)} = P_k(M - S^{(t)})$ and $S^{(t+1)}$ is given by: $S^{(t+1)} = H_{\zeta}(M - L^{(t+1)})$. Also, recall that $E^{(t)} := S^* - S^{(t)}$ and $E^{(t+1)} := S^* - S^{(t+1)}$.

We prove the lemma by induction on both k and t. For the base case ($k = 1$ and $t = -1$), we first note that the first inequality on $\left\| L^{(0)} - L^* \right\|_\infty$ is trivially satisfied. Due to the thresholding step (step 3 in Algorithm 1) and the incoherence assumption on L^*, we have:

$$\begin{aligned} \left\| E^{(0)} \right\|_\infty &\leq \frac{8\mu^2 r}{n} \left(\sigma_2^* + 2\sigma_1^* \right), \text{ and} \\ \text{Supp}\left(E^{(0)} \right) &\subseteq \text{Supp}\left(S^* \right). \end{aligned}$$

So the base case of induction is satisfied.

We first do the inductive step over t (for a fixed k). By inductive hypothesis we assume that: a) $\left\| E^{(t)} \right\|_\infty \leq \frac{8\mu^2 r}{n} \left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^{t-1} \left|\sigma_k^*\right| + 7 \left\| N^* \right\|_2 + \frac{8n}{\sqrt{r}} \left\| N^* \right\|_\infty \right)$, b) $\text{Supp}\left(E^{(t)} \right) \subseteq \text{Supp}\left(S^* \right)$. Then by Lemma 11, we have:

$$\left\| L^{(t+1)} - L^* \right\|_\infty \leq \frac{2\mu^2 r}{n} \left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^t \left|\sigma_k^*\right| + 7 \left\| N^* \right\|_2 + \frac{8n}{\sqrt{r}} \left\| N^* \right\|_\infty \right).$$

Lemma 13 now tells us that

1. $\left\| E^{(t+1)} \right\|_\infty \leq \frac{7\mu^2 r}{n} \left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^t \left|\sigma_k^*\right| + 7 \left\| N^* \right\|_2 + \frac{8n}{\sqrt{r}} \left\| N^* \right\|_\infty \right)$, and
2. $\text{Supp}\left(E^{(t+1)} \right) \subseteq \text{Supp}\left(S^* \right)$.

This finishes the induction over t. Note that we show a stronger bound than necessary on $\left\| E^{(t+1)} \right\|_\infty$.

We now do the induction over k. Suppose the hypothesis holds for stage k. Let T denote the number of iterations in each stage. We first obtain a lower bound on T. Since

$$\left\| M - S^{(0)} \right\|_2 \geq \left\| L^* + N^* \right\|_2 - \left\| E^{(0)} \right\|_2 \geq \left|\sigma_1^*\right| - \alpha n \left\| E^{(0)} \right\|_\infty \geq \frac{3}{4} \left|\sigma_1^*\right|,$$

we see that $T \geq 10 \log \left(3\mu^2 r \left|\sigma_1^*\right| / \epsilon \right)$. So, at the end of stage k, we have:

1. $\left\| E^{(T)} \right\|_\infty \leq \frac{7\mu^2 r}{n} \left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^T \left|\sigma_k^*\right| + 7 \left\| N^* \right\|_2 + \frac{8n}{\sqrt{r}} \left\| N^* \right\|_\infty \right) \leq \frac{7\mu^2 r \left|\sigma_{k+1}^*\right|}{n} + \frac{\epsilon}{10n}$, and
2. $\text{Supp}\left(E^{(T)} \right) \subseteq \text{Supp}\left(S^* \right)$.

Lemmas 4 and 2 tell us that $\left| \sigma_{k+1}\left(M - S^{(T)} \right) - \left|\sigma_{k+1}^*\right| \right| \leq \left\| E^{(T)} \right\|_2 \leq \alpha \left(7\mu^2 r \left|\sigma_{k+1}^*\right| + \epsilon \right)$. We will now consider two cases:

1. **Algorithm 1 terminates:** This means that $\beta \sigma_{k+1}\left(M - S^{(T)} \right) < \frac{\epsilon}{2n}$ which then implies that $\left|\sigma_{k+1}^*\right| < \frac{\epsilon}{6\mu^2 r}$. So we have:

$$\begin{aligned} \left\| \widehat{L} - L^* \right\|_\infty = \left\| L^{(T)} - L^* \right\|_\infty &\leq \frac{2\mu^2 r}{n} \left(\left|\sigma_{k+1}^*\right| + \left(\frac{1}{2}\right)^T \left|\sigma_k^*\right| + 7 \left\| N^* \right\|_2 + \frac{8n}{\sqrt{r}} \left\| N^* \right\|_\infty \right) \\ &\leq \frac{\epsilon}{5n} + \frac{2\mu^2 r}{n} \left(7 \left\| N^* \right\|_2 + \frac{8n}{\sqrt{r}} \left\| N^* \right\|_\infty \right). \end{aligned}$$

   This proves the statement about $\widehat{L}$. A similar argument proves the claim on $\left\| \widehat{S} - S^* \right\|_\infty$. The claim on $\text{Supp}\left(\widehat{S} \right)$ follows since $\text{Supp}\left(E^{(T)} \right) \subseteq \text{Supp}\left(S^* \right)$.

2. **Algorithm 1 continues to stage $(k+1)$:** This means that $\beta\sigma_{k+1}\left(L^{(T)}\right) \geq \frac{\epsilon}{2n}$ which then implies that $\left|\sigma^*_{k+1}\right| > \frac{\epsilon}{8\mu^2 r}$. So we have:

$$\begin{aligned}\left\|E^{(T)}\right\|_\infty &\leq \frac{7\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right| + \left(\frac{1}{2}\right)^T \left|\sigma^*_k\right| + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right)\\ &\leq \frac{7\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right| + \frac{\epsilon}{10\mu^2 rn} + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right)\\ &\leq \frac{7\mu^2 r}{n}\left(\left|\sigma^*_{k+1}\right| + \frac{8\left|\sigma^*_{k+1}\right|}{10n} + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right)\\ &\leq \frac{8\mu^2 r}{n}\left(\left|\sigma^*_{k+2}\right| + 2\left|\sigma^*_{k+1}\right| + 7\left\|N^*\right\|_2 + \frac{8n}{\sqrt{r}}\left\|N^*\right\|_\infty\right).\end{aligned}$$

   Similarly for $\left\|L^{(T)} - L^*\right\|_\infty$.

This finishes the proof. □

Proof of Theorem 2. Using Lemma 14, it suffices to show that the general case can be reduced to the case of symmetric matrices. We will now outline this reduction.

Recall that we are given an $m \times n$ matrix $M = L^* + N^* + S^*$ where L^* is the true low-rank matrix, N^* dense corruption matrix and S^* the sparse error matrix. Wlog, let $m \leq n$ and suppose $\beta m \leq n < (\beta+1)m$, for some $\beta \geq 1$. We then consider the symmetric matrices

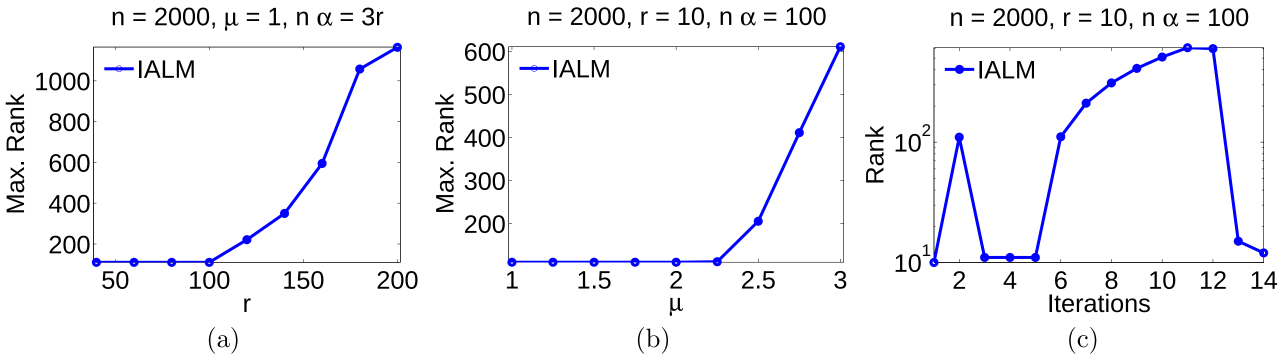
$$\widetilde{M} = \left[\begin{array}{cc} \underbrace{\begin{array}{ccc} 0 & & 0 \\ \vdots & \cdots & \vdots \\ 0 & & 0 \\ M^\top & \cdots & M^\top \end{array}}_{\beta \text{ times}} & \begin{array}{c} M \\ \vdots \\ M \\ 0 \end{array} \end{array}\right], \widetilde{L} = \left[\begin{array}{cc} \underbrace{\begin{array}{ccc} 0 & & 0 \\ \vdots & \cdots & \vdots \\ 0 & & 0 \\ \left(L^*\right)^\top & \cdots & \left(L^*\right)^\top \end{array}}_{\beta \text{ times}} & \begin{array}{c} L^* \\ \vdots \\ L^* \\ 0 \end{array} \end{array}\right],$$

$$\widetilde{N} = \left[\begin{array}{cc} \underbrace{\begin{array}{ccc} 0 & & 0 \\ \vdots & \cdots & \vdots \\ 0 & & 0 \\ \left(N^*\right)^\top & \cdots & \left(N^*\right)^\top \end{array}}_{\beta \text{ times}} & \begin{array}{c} L^* \\ \vdots \\ L^* \\ 0 \end{array} \end{array}\right], \tag{36}$$

and $\widetilde{S} = \widetilde{M} - \widetilde{L}$. A simple calculation shows that $\widetilde{L}$ is incoherent with parameter $\sqrt{3}\mu$, $\widetilde{N}$ satisfies the assumption of Theorem 2 and $\widetilde{S}$ satisfies the sparsity condition (S1) with parameter $\frac{\alpha}{\sqrt{2}}$. Moreover the iterates of AltProj with input $\widetilde{M}$ have similar expressions as in (36) in terms of the corresponding iterates with input M. This means that it suffices to obtain the same guarantees for Algorithm 1 for the symmetric case. Lemma 14 does precisely this, proving the theorem. □

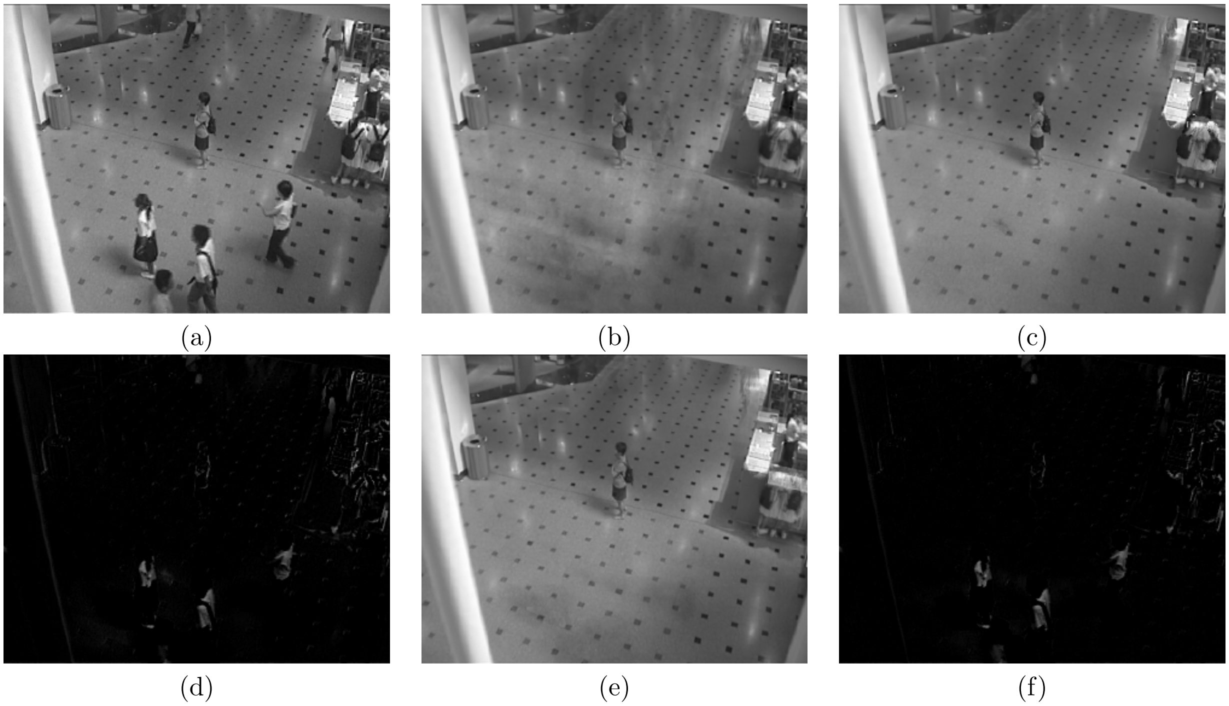
C Additional experimental results

Synthetic datasets: Extending Figure 2, the plots in Figure 5 illustrate the point that soft thresholding, i.e., the convex relaxation approach, leads to intermediate solutions with high ranks. Figures 5 (a)-(b) show the variation of the maximum rank of the intermediate low-rank solutions of IALM with rank and incoherence respectively; the results are averaged over 5 runs of the algorithm; we note that as the problem becomes harder, the maximum intermediate rank via soft thresholding (convex approach) increases, and this leads to higher running times. As an example of this phenomenon, Figure 5 (c) shows the rank of the intermediate iterates of IALM for a particular run with $n = 2000, r = 10, \alpha = 100/n, \mu = 3$; here, while the rank of the final output is 10, intermediate iterates have a rank as high as 800. We run our synthetic simulations on a machine with Intel Dual 8-core Xeon (E5-2650) 2.0GHz CPU with 192GB RAM.



(a) (b) (c)

Figure 5: (a): Variation of the maximum rank of the intermediate low-rank solutions of IALM with rank. (b): Variation of the maximum rank of the intermediate low-rank solutions of IALM with incoherence. (c): Rank of the intermediate iterates of IALM for a particular run with $n = 2000, r = 10, \alpha = 100/n, \mu = 3$. Note that while the rank of the final output is 10, intermediate iterates have rank as high as 800.



(a) (b) (c)

(d) (e) (f)

Figure 6: Foreground-background separation in the *Shopping Mall* video. (a): Original frame in the video given as a part of the input to NcRPCA and IALM. (b): Corresponding frame from the best rank-20 approximation obtained using vanilla PCA; time taken for computing the low-rank approximation is $8.8s$. (c): Corresponding frame from the low-rank part obtained using NcRPCA; time taken by NcRPCA to compute the low-rank and sparse solutions is $292.1s$. (d): Corresponding frame from the sparse part obtained using NcRPCA. (e): Corresponding frame from the low-rank part obtained using IALM; time taken by IALM to compute the low-rank and sparse solutions is $783.4s$. (f): Corresponding frame from the sparse part obtained using IALM.

Real-world datasets: We provide some additional results concerning foreground-background separation in videos [5]. We compare NcRPCA with IALM, and also with the low-rank solution obtained using vanilla PCA; we report the solutions obtained by NcRPCA and IALM methods for decomposing M into $L+S$ up to a relative error ($\|M - L - S\|_F / \|M\|_F$) of 10^{-3}. We report the rank and the sparsity of the solutions obtained by the two methods along with the computational time. As mentioned before, the observed matrix M is formed by

[5] The datasets are available at `http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html`

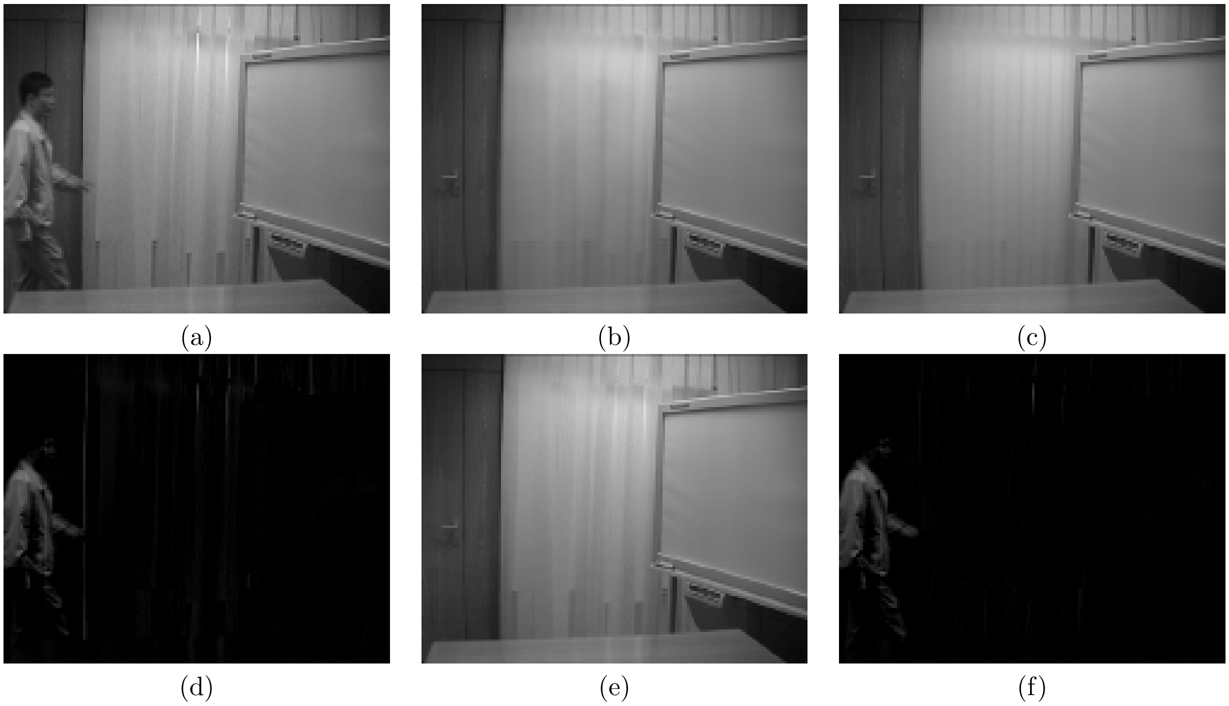


Figure 7: Foreground-background separation in the *Curtain* video. (a): Original image frame in the video given as a part of the input to NcRPCA and IALM. (b): Corresponding frame from the best rank-10 approximation obtained using vanilla PCA; time taken for computing the low-rank approximation is $2.8s$. (c): Corresponding frame from the low-rank part obtained using NcRPCA; time taken by NcRPCA to compute the low-rank and sparse solutions is $39.5s$. (d): Corresponding frame from the sparse part obtained using NcRPCA. (e): Corresponding frame from the low-rank part obtained using IALM; time taken by IALM to compute the low-rank and sparse solutions is $989.0s$. (f): Corresponding frame from the sparse part obtained using IALM.

vectorizing each frame and stacking them column-wise. For illustration purposes, we arbitrarily select one of the original frames in the sequence of image frames obtained from the video, i.e., one of the columns of M, and the corresponding columns in L and S obtained using NcRPCA and IALM. We run our real data experiments on a machine with Intel Dual 8-core Xeon (E5-2650) 2.0GHz CPU with 192GB RAM.

Shopping Mall dataset: Figure 6 shows the comparison of NcRPCA and IALM on the "Shopping Mall" dataset which has 1286 frames at a resolution of 256 × 320. NcRPCA achieves a solution of better visual quality (for example, unlike NcRPCA, notice the artifact of the low-rank solution from IALM in the top right corner of the image where the person is walking over the reflection of a light source; also notice the shadows of people in the low-rank part obtained by IALM which are not present in the low-rank solution obtained by NcRPCA), in $292.1s$, compared to IALM, which takes $783.4s$ until convergence. NcRPCA obtains a rank 20 solution for L with $\|S\|_0 = 95411896$ whereas IALM obtains a rank 286 solution for L with $\|S\|_0 = 86253965$.

Curtain dataset: We illustrate our recovery on one of the frames (frame 2773) wherein a person enters a room with a curtain on the background. Figure 7 shows the comparison of NcRPCA and IALM on the "Curtain" dataset which has 2964 frames at a resolution of 160 × 128. NcRPCA achieves a solution, in $39.5s$, which is of similar visual quality to that of IALM, which takes $989.0s$ until convergence. NcRPCA obtains a rank 1 solution for L with $\|S\|_0 = 53897769$ whereas IALM obtains a rank 701 solution for L with $\|S\|_0 = 42310582$.