# Tensor vs Matrix Methods:
# Robust Tensor Decomposition under Block Sparse Perturbations

Animashree Anandkumar[*], Prateek Jain[†], Yang Shi[‡], U. N. Niranjan[§]

April 26, 2016

## Abstract

Robust tensor CP decomposition involves decomposing a tensor into low rank and sparse components. We propose a novel non-convex iterative algorithm with guaranteed recovery. It alternates between low-rank CP decomposition through gradient ascent (a variant of the tensor power method), and hard thresholding of the residual. We prove convergence to the globally optimal solution under natural incoherence conditions on the low rank component, and bounded level of sparse perturbations. We compare our method with natural baselines, viz., which apply robust matrix PCA either to the *flattened* tensor, or to the matrix slices of the tensor. Our method can provably handle a far greater level of perturbation when the sparse tensor is block-structured. Thus, we establish that tensor methods can tolerate a higher level of gross corruptions compared to matrix methods.

## 1 Introduction

In this paper, we develop a robust tensor decomposition method, which recovers a low rank tensor subject to gross corruptions. Given an input tensor $T = L^* + S^*$, we aim to recover both $L^*$ and $S^*$, where $L^*$ is a low rank tensor and $S^*$ is a sparse tensor

$$T = L^* + S^*, \quad L^* = \sum_{i=1}^{r} \sigma_i^* u_i \otimes u_i \otimes u_i \tag{1}$$

and $T, L^*, S^* \in \mathbb{R}^{n \times n \times n}$. The above form of $L^*$ is known as the Candecomp/Parafac or the CP-form. We assume that $L^*$ is a rank-$r$ orthogonal tensor, i.e., $\langle u_i, u_j \rangle = 1$ if $i = j$ and 0 otherwise. The above problem arises in numerous applications such as image and video denoising [17], multi-task learning, and robust learning of latent variable models (LVMs) with grossly-corrupted moments, for details see Section 1.2.

The matrix version of (1), viz., decomposing a matrix into sparse and low rank matrices, is known as robust principal component analysis (PCA). It has been studied extensively [6, 8, 15, 22]. Both convex [6, 8] as well as non-convex [22] methods have been proposed with provable recovery.

One can attempt to solve the robust tensor problem in (1) using matrix methods. In other words, robust matrix PCA can be applied either to each matrix slice of the tensor, or to the matrix obtained by flattening the tensor. However, such matrix methods ignore the tensor algebraic constraints or the CP rank constraints, which differ from the matrix rank constraints. There are however a number of challenges to incorporating the tensor CP rank constraints. Enforcing a given tensor rank is NP-hard [14], unlike the matrix case, where low rank projections can be computed efficiently. Moreover, finding the best convex relaxation of the tensor CP rank is also NP-hard [14], unlike the matrix case, where the convex relaxation of the rank, viz., the nuclear norm, can be computed efficiently.

---

[*]University of California, Irvine. Email: a.anandkumar@uci.edu
[†]Microsoft Research, India. Email: prajain@microsoft.com
[‡]University of California, Irvine. Email: shiy4@uci.edu
[§]University of California, Irvine. Email: un.niranjan@uci.edu

## 1.1 Summary of Results

**Proposed method:** We propose a non-convex iterative method, termed RTD, that maintains low rank and sparse estimates $\hat{L}, \hat{S}$, which are alternately updated. The low rank estimate $\hat{L}$ is updated through the eigenvector computation of $T - \hat{S}$, and the sparse estimate is updated through (hard) thresholding of the residual $T - \hat{L}$.

**Tensor Eigenvector Computation:** Computing eigenvectors of $T - \hat{S}$ is challenging as the tensor can have arbitrary "noise" added to an orthogonal tensor and hence the techniques of [2] do not apply as they only guarantee an approximation to the eigenvectors up to the "noise" level. Results similar to the shifted power method of [19] should apply for our problem, but their results hold in an arbitrarily small centered at the true eigenvectors, and the size of the ball is typically not well-defined. In this work, we provide a simple variant of the tensor power method based on gradient ascent of a regularized variational form of the eigenvalue problem of a tensor. We show that our method converges to the true eigenvectors at a linear rate when initialized within a reasonably small ball around eigenvectors. See Theorem 2 for details.

**Guaranteed recovery:** As a main result, we prove convergence to the global optimum $\{L^*, S^*\}$ for RTD under an incoherence assumption on $L^*$, and a bounded sparsity level for $S^*$. These conditions are similar to the conditions required for the success of matrix robust PCA. We also prove fast linear convergence rate for RTD, i.e. we obtain an additive $\epsilon$-approximation in $O(\log(1/\epsilon))$ iterations.

**Superiority over matrix robust PCA:** We compare our RTD method with matrix robust PCA, carried out either on matrix slices of the tensor, or on the *flattened* tensor. We prove our RTD method is superior and can handle higher sparsity levels in the noise tensor $S^*$, when it is block structured. Intuitively, each block of noise represents correlated noise which persists for a subset of slices in the tensor. For example, in a video if there is an occlusion then the occlusion remains fixed in a small number of frames. In the scenario of moment-based estimation, $S^*$ represents gross corruptions of the moments of some multivariate distribution, and we can assume that it occurs over a small subset of variables.

We prove that our tensor methods can handle a much higher level of block sparse perturbations, when the overlap between the blocks is controlled (e.g. random block sparsity). For example, for a rank-1 tensor, our method can handle $O(n^{17/12})$ corrupted entries per fiber of the tensor (i.e. row/column of a slice of the tensor). In contrast, matrix robust PCA methods only allows for $O(n)$ corrupted entries, and this bound is tight [22]. We prove that even better gains are obtained for RTD when the rank $r$ of $L^*$ increases, and we provide precise results in this paper. Thus, our RTD achieves best of both the worlds: better accuracy and faster running times.

We conduct extensive simulations to empirically validate the performance of our method and compare it to various matrix robust PCA methods. Our synthetic experiments show that our tensor method is 2-3 times more accurate, and about 8-14 times faster, compared to matrix decomposition methods. On the real-world *Curtain* dataset, for the activity detection, our tensor method obtains better recovery with a 10% speedup.

**Overview of techniques:** At a high level, the proposed method RTD is a tensor analogue of the non-convex matrix robust PCA method in [22]. However, both the algorithm (RTD) and the analysis of RTD is significantly challenging due to two key reasons: a) there can be significantly more structure in the tensor problem that needs to be exploited carefully using structure in the noise,b) unlike matrices, tensors can have an exponential number of eigenvectors [7].

We would like to stress that we need to establish convergence to the globally optimal solution $\{L^*, S^*\}$, and not just to a local optimum, despite the non-convexity of the decomposition problem. Intuitively, if we are in the basin of attraction of the global optimum, it is natural to expect that the estimates $\{\hat{L}, \hat{S}\}$ under RTD are progressively refined, and get closer to the true solution $\{L^*, S^*\}$. However, characterizing this basin, and the conditions needed to ensure we "land" in this basin is non-trivial and novel.

As mentioned above, our method alternates between finding low rank estimate $\hat{L}$ on the residual $T - \hat{S}$ and viceversa. The main steps in our proof are as follows: *(i)* For updating the low rank estimate, we propose a modified tensor power method, and prove that it converges to one of the eigenvectors of $T - \hat{S}$. In addition, the recovered eigenvectors are "close" to the components of $L^*$. *(ii)* When the sparse estimate $\hat{S}$ is updated through hard thresholding, we prove that the support of $\hat{S}$ is contained within that of $S^*$. *(iii)* We make strict progress in each epoch, where $\hat{L}$ and $\hat{S}$ are alternately updated.

In order to prove the first part, we establish that the proposed method performs gradient ascent on a regularized variational form of the eigenvector problem. We then establish that the regularized objective satisfies local strong convexity and smoothness. We also establish that by having a polynomial number of initializations, we can recover vectors that are "reasonably" close to eigenvectors of $T - \hat{S}$. Using the above two facts, we establish a linear convergence to the true eigenvectors of $T - \hat{S}$, which are close to the components of $L^*$.

For step *(ii)* and *(iii)*, we show that using an intuitive block structure in the noise, we can bound the affect of noise on the eigenvectors of the true low-rank tensor $L^*$ and show that the proposed iterative scheme refines the estimates of $L^*$ and converge to $L^*$ at a linear rate.

## 1.2 Applications

In addition to the standard applications of robust tensor decomposition to video denoising [17], we propose two applications in probabilistic learning.

**Learning latent variable models using grossly corrupted moments:** Using tensor decomposition for learning LVMs has been intensely studied in the last few years, e.g. [1, 2, 18]. The idea is to learn the models through CP decomposition of higher order moment tensors.While the above works assume access to empirical moments, we can extend the framework to that of *robust* estimation, where the moments are subject to gross corruptions. In this case, gross corruptions on the moments can occur either due to adversarial manipulations or systematic bias in estimating moments of some subset of variables.

**Multi-task learning of linear Bayesian networks:** Let $z_{(i)} \sim \mathcal{N}(0, \Sigma_{(i)})$. The samples for the $i$th Bayesian network are generated as $x_{(i)} = U h_{(i)} + z_{(i)}$, and $\mathbb{E}[x_{(i)} x_{(i)}^\top] = U \operatorname{diag}(w_{(i)}) U^\top + \Sigma_{(i)}$, where $h_{(i)}$ is the hidden variable for the $i$th network. If the Bayesian networks are related, we can share parameters among them. In the above framework, we share parameters $U$, which map the hidden variables to observed ones. Assuming that all the covariances $\Sigma_{(i)}$ are sparse, when they are stacked together, they form a sparse tensor. Similarly $U \operatorname{diag}(w_{(i)}) U^\top$ stacked constitutes a low rank tensor. Thus, we can consider the samples jointly, and learn the parameters by performing robust tensor decomposition.

## 1.3 Related Work

**Robust matrix decomposition:** In the matrix setting, the above problem of decomposition into sparse and low rank parts is popularly known as *robust PCA*, and has been studied in a number of works ([8],[6]). The popular method is based on convex relaxation, where the low rank penalty is replaced by nuclear norm and the sparsity is replaced by the $\ell_1$ norm. However, this technique is not applicable in the tensor setting, when we consider the *CP rank*. There is no convex surrogate available for the CP rank.

Recently a non convex method for robust PCA is proposed in [22]. It involves alternating steps of PCA and thresholding of the residual. Our proposed tensor method can be seen as a tensor analogue of the method in [22]. However, the analysis is very different, since the optimization landscape for tensor decomposition differs significantly from that of the matrix.

**Convex Robust Tucker decomposition:** Previous works which employ convex surrogates for tensor problems employ a different notion of rank, known as the *Tucker* rank or the *multi*-rank, e.g. [10, 12, 16, 20, 25]. However, the notion of a multi-rank is based on ranks of the matricization or flattening of the tensor, and thus, this method does not exploit the tensor algebraic constraints. The problem of robust tensor PCA is specifically tackled in [11, 13]. In [11], convex and non-convex methods are proposed based on *Tucker* rank, but there are no guarantees on if it yields the original tensors $L^*$ and $S^*$ in (1). In [13], they prove success under restricted eigenvalue conditions. However, these conditions are opaque and it is not clear regarding the level of sparsity that can be handled.

**Sum of squares:** Barak et al [5] recently consider CP-tensor completion using algorithms based on the sum of squares hierarchy. However, these algorithms are expensive. In contrast, in this paper, we consider simple iterative methods based on the power method that are efficient and scalable for large datasets. It is however unclear if the sum of squares algorithm improves the result for the block sparse model considered here.

---

**Algorithm 1** $(\widehat{L}, \ \widehat{S})$ = RTD $(T, \delta, r, \beta)$: Tensor Robust PCA

---

1: **Input**: Tensor $T \in \mathbb{R}^{n \times n \times n}$, convergence criterion $\delta$, target rank $r$, thresholding scale parameter $\beta$. $P_l(A)$ denote estimated rank-$l$ approximation of tensor $A$, and let $\sigma_l(A)$ denote the estimated $l^{\text{th}}$ largest eigenvalue using Procedure 1. $HT_\zeta(A)$ denotes hard-thresholding, i.e. $\mathcal{H}_\zeta(A))_{ijk} = A_{ijk}$ if $|A_{ijk}| \geq \zeta$ and 0 otherwise.

2: Set initial threshold $\zeta_0 \leftarrow \beta\sigma_1(T)$ and estimates $S^{(0)} = \mathcal{H}_{\zeta_0}(T - L^{(0)})$.

3: **for** Stage $l = 1$ to $r$ **do**

4:     **for** $t = 0$ to $\tau = 10\log\left(n\beta\left\|T - S^{(0)}\right\|_2 / \delta\right)$ **do**

5:         $L^{(t+1)} = P_l(T - S^{(t)})$.

6:         $S^{(t+1)} = \mathcal{H}_\zeta(T - L^{(t+1)})$.

7:         $\zeta_{t+1} = \beta(\sigma_{l+1}(T - S^{(t+1)}) + \left(\frac{1}{2}\right)^t \sigma_l(T - S^{(t+1)}))$.

8:     If $\beta\sigma_{l+1}(L^{(t+1)}) < \frac{\delta}{2n}$, then return $L^{(\tau)}, S^{(\tau)}$, else reset $S^{(0)} = S^{(\tau)}$

9: **Return:** $\widehat{L} = L^{(\tau)}, \widehat{S} = S^{(\tau)}$

---

**Robust tensor decomposition:** Shah et al [23] consider robust tensor decomposition method using a randomized convex relaxation formulation. Under their random sparsity model, their algorithm provides guaranteed recovery as long as the number of non-zeros per fibre is $O(\sqrt{n})$. This is in contrast to our method which potentially tolerates upto $O(n^{17/12})$ non-zero sparse corruptions per fibre.

## 2   Proposed Algorithm

**Notations:** Let $[n] := \{1, 2, \ldots, n\}$, and $\|v\|$ denote the $\ell_2$ norm of vector $v$. For a matrix or a tensor $M$, $\|M\|$ refers to spectral norm and $\|M\|_\infty$ refers to maximum absolute entry.

**Tensor preliminaries:** A real third order tensor $T \in \mathbb{R}^{n \times n \times n}$ is a three-way array. The different dimensions of the tensor are referred to as *modes*. In addition, *fibers* are higher order analogues of matrix rows and columns. A fiber is obtained by fixing all but one of the indices of the tensor. For example, for a third order tensor $T \in \mathbb{R}^{n \times n \times n}$, the mode-1 fiber is given by $T(:, j, l)$. Similarly, *slices* are obtained by fixing all but two of the indices of the tensor. For example, for the third order tensor $T$, the slices along third mode are given by $T(:, :, l)$. A *flattening* of tensor $T$ along mode $k$ is a matrix $M$ whose columns correspond to mode-$k$ fibres of $T$.

We view a tensor $T \in \mathbb{R}^{n \times n \times n}$ as a multilinear form. In particular, for vectors $u, v, w \in \mathbb{R}^n$, let

$$T(I, v, w) := \sum_{j, l \in [n]} v_j w_l T(:, j, l) \ \in \mathbb{R}^n, \tag{2}$$

which is a multilinear combination of the tensor mode-1 fibers. Similarly $T(u, v, w) \in \mathbb{R}$ is a multilinear combination of the tensor entries.

A tensor $T \in \mathbb{R}^{n \times n \times n}$ has a CP *rank* at most $r$ if it can be written as the sum of $r$ rank-1 tensors as

$$T = \sum_{i \in [r]} \sigma_i^* u_i \otimes u_i \otimes u_i, \quad u_i \in \mathbb{R}^n, \|u_i\| = 1, \tag{3}$$

where notation $\otimes$ represents the *outer product*. We sometimes abbreviate $a \otimes a \otimes a$ as $a^{\otimes 3}$. Without loss of generality, $\sigma_i^* > 0$, since $-\sigma_i^* u_i^{\otimes 3} = \sigma_i^*(-u_i)^{\otimes 3}$.

**RTD method:** We propose non-convex algorithm RTD for robust tensor decomposition, described in Algorithm 1. The algorithm proceeds in stages, $l = 1, \ldots, r$, where $r$ is the target rank of the low rank estimate. In $l^{\text{th}}$ stage, we consider alternating steps of low rank projection $P_l(\cdot)$ and hard thresholding of the residual, $\mathcal{H}(\cdot)$. For computing $P_l(\widetilde{L})$, that denotes the $l$ leading eigenpairs of $\widetilde{L}$, we execute gradient ascent on a function $f(v) = \widetilde{L}(v, v, v) - \lambda\|v\|^4$ with multiple restarts and deflation (see Procedure 1).

---

**Procedure 1** $\{\hat{L}_l, (\hat{u}_j, \lambda_j)_{j \in [l]}\} = P_l(T)$: GradAscent (Gradient Ascent method)

---

1: **Input**: Symmetric tensor $T \in \mathbb{R}^{n \times n \times n}$, target rank $l$, exact rank $r$, $N_1$ number of initializations or restarts, $N_2$ number of power iterations for each initialization. Let $T_1 \leftarrow T$.

2: **for** $j = 1, \ldots, r$ **do**

3:     **for** $i = 1, \ldots, N_1$ **do**

4:         $\theta \sim \mathcal{N}(0, I_n)$. Compute top singular vector $u$ of $T_j(I, I, \theta)$. Initialize $v_i^{(1)} \leftarrow u$. Let $\lambda = T_j(u, u, u)$.

5:         **repeat**

6:             $v_i^{(t+1)} \leftarrow T_j(I, v_i^{(t)}, v_i^{(t)})/\|T_j(I, v_i^{(t)}, v_i^{(t)})\|_2$ {Run power method to land in spectral ball}

7:             $\lambda_i^{(t+1)} \leftarrow T_j(v_i^{(t+1)}, v_i^{(t+1)}, v_i^{(t+1)})$

8:         **until** $t = N_2$

9:         Pick the best: reset $i \leftarrow \arg\max_{i \in [N_1]} T_j(v_i^{(t+1)}, v_i^{(t+1)}, v_i^{(t+1)})$ and $\lambda_i = \lambda_i^{(t+1)}$ and $v_i = v_i^{(t+1)}$.

10:         Deflate: $T_j \leftarrow T_j - \lambda_i v_i \otimes v_i \otimes v_i$.

11: **for** $j = 1, \ldots, r$ **do**

12:     **repeat**

13:         Gradient Ascent iteration: $v_j^{(t+1)} \leftarrow v_j^{(t)} + \frac{1}{4\lambda(1+\lambda/\sqrt{n})} \cdot \left(T(I, v_j^{(t)}, v_j^{(t)}) - \lambda\|v_j^{(t)}\|^2 v_j^{(t)}\right)$.

14:     **until** convergence (linear rate, refer Lemma 3).

15:     Set $\hat{u}_j = v_j^{(t+1)}$, $\lambda_j = T(v_j^{(t+1)}, v_j^{(t+1)}, v_j^{(t+1)})$

16: **return** Estimated top $l$ out of all the top $r$ eigenpairs $(\hat{u}_j, \lambda_j)_{j \in [l]}$, and low rank estimate $\hat{L}_l = \sum_{i \in [l]} \lambda_i \hat{u}_j \otimes \hat{u}_j \otimes \hat{u}_j$.

---

**Computational complexity:** In RTD, at the $l^{\text{th}}$ stage, the $l$-eigenpairs are computed using Procedure 1, whose complexity is $O(n^3 l N_1 N_2)$. The hard thresholding operation $\mathcal{H}_\zeta(T - L^{(t+1)})$ requires $O(n^3)$ time. We have $O\left(\log\left(\frac{n\|T\|}{\delta}\right)\right)$ iterations for each stage of the RTD algorithm and there are $r$ stages. By Theorem 2, it suffices to have $N_1 = \tilde{O}\left(n^{1+c}\right)$ and $N_2 = \tilde{O}\left(1\right)$, and where $\tilde{O}(\cdot)$ represents $O(\cdot)$ up to polylogarithmic factors and $c$ is a small constant. Hence, the overall computational complexity of RTD is $\tilde{O}\left(n^{4+c}r^2\right)$.

## 3   Theoretical Guarantees

In this section, we provide guarantees for the RTD proposed in the previous section for RTD in (1). Even though we consider a symmetric $L^*$ and $S^*$ in (1), we can extend the results to asymmetric tensors, by embedding them into symmetric tensors, on lines of [24].

In general, it is impossible to have a unique recovery of low-rank and sparse components. Instead, we assume the following conditions to guarantee uniqueness:

**(L)** $L^*$ is a rank-$r$ orthogonal tensor in (1), i.e., $\langle u_i, u_j \rangle = \delta_{i,j}$, where $\delta_{i,j} = 1$ iff $i = j$ and 0 o.w. $L^*$ is $\mu$-incoherent, i.e., $\|u_i\|_\infty \leq \frac{\mu}{n^{1/2}}$ and $\sigma_i^* > 0$, $\forall 1 \leq i \leq r$.

The above conditions of having an incoherent low rank tensor $L^*$ are similar to the conditions for robust matrix PCA. For tensors, the assumption of an orthogonal decomposition is limiting, since there exist tensors whose CP decomposition is non-orthogonal [2]. We later discuss how our analysis can be extended to non-orthogonal tensors. We now list the conditions for sparse tensor $S^*$.

The tensor $S^*$ is block sparse, where each block has at most $d$ non-zero entries along any fibre and the number of blocks is $B$. Let $\Psi$ be the support tensor that has the same sparsity pattern as $S^*$, but with unit entries, i.e. $\Psi_{i,j,k} = 1$ iff. $S^*_{i,j,k} \neq 0$ for all $i, j, k \in [n]$. We now make assumptions on sparsity pattern of $\Psi$. Let $\eta$ be the maximum fraction of overlap between any two block fibres $\psi_i$ and $\psi_j$. In other words, $\max_{i \neq j} \langle \psi_i, \psi_j \rangle < \eta d$. **(S)** Let $d$ be the sparsity level along any fibre of a block and let $B$ be the number of

blocks. We require

$$\Psi = \sum_{i=1}^{B} \psi_i \otimes \psi_i \otimes \psi_i, \|\psi_i\|_0 \le d, \ \psi_i(j) = 0 \text{ or } 1$$

$$\forall i \in [B], \ j \in [n], \tag{4}$$

$$d = O(n/r\mu^3)^{2/3}, B = O(\min(n^{2/3}r^{1/3}, \eta^{-1.5})). \tag{5}$$

We assume a block sparsity model for $S^*$ above. Under this model, the support tensor $\Psi$ which encodes sparsity pattern, has rank $B$, but not the sparse tensor $S^*$ since the entries are allowed to be arbitrary. We also note that we set $d$ to be $n^{2/3}$ for ease of exposition and show one concrete example where our method significantly outperforms matrix robust PCA methods.

As discussed in the introduction, it may be advantageous to consider tensor methods for robust decomposition only when sparsity across the different matrix slices are aligned/structured in some manner, and a block sparse model is a natural structure to consider. We later demonstrate the precise nature of superiority of tensor methods under block sparse perturbations.

For the above mentioned sparsity structure, we set $\beta = 4\mu^3 r/n^{3/2}$ in our algorithm. Under the above conditions, our proposed algorithm RTD establishes convergence to the globally optimal solution.

**Theorem 1** (Convergence to global optimum for RTD)**.** *Let $L^*, S^*$ satisfy $(L)$ and $(S)$, and $\beta = 4\frac{\mu^3 r}{n^{3/2}}$. The outputs $\widehat{L}$ (and its parameters $\hat{u}_i$ and $\hat{\lambda}_i$) and $\widehat{S}$ of Algorithm 1 satisfy w.h.p.:*

$$\|\hat{u}_i - u_i\|_\infty \le \delta/\mu^2 r n^{1/2} \sigma^*_{\min},$$

$$|\hat{\lambda}_i - \sigma^*_i| \le \delta, \quad \forall i \in [n],$$

$$\left\|\widehat{L} - L^*\right\|_F \le \delta, \quad \|\widehat{S} - S^*\|_\infty \le \delta/n^{3/2},$$

$$and \quad \text{supp}\,\widehat{S} \subseteq \text{supp}\,S^*.$$

**Comparison with matrix methods:**    We now compare with the matrix methods for recovering the sparse and low rank tensor components in (1). Robust matrix PCA can be performed either on all the matrix slices of the input tensor $M_i := T(I, I, e_i)$, for $i \in [n]$, or on the *flattened* tensor $T$ (see the definition in Section 2). We can either use convex relaxation methods [6, 8, 15] or non-convex methods [22] for robust matrix PCA.

Recall that $\eta$ measures the fraction of overlapping entries between any two different block fibres, i.e. $\max_{i \ne j}\langle \psi_i, \psi_j \rangle < \eta d$, where $\psi_i$ are the fibres of the block components of tensor $\Psi$ in (4) which encodes the sparsity pattern of $S^*$ with 0-1 entries. A short proof is given in Appendix B.1.

**Corollary 1** (Superiority of tensor methods)**.** *The proposed tensor method* RTD *can handle perturbations $S^*$ at a higher sparsity level compared to performing matrix robust PCA on either matrix slices or the flattened tensor using guaranteed methods in [15, 22] when the (normalized) overlap between different blocks satisfies $\eta = O(r/n)^{2/9}$.*

**Simplifications under random block sparsity:**    We now obtain transparent results for a random block sparsity model, where the components $\psi_i$ in (4) for the support tensor $\Psi$ are drawn uniformly among all $d$-sparse vectors. In this case, the incoherence $\eta$ simplifies as $\eta \overset{w.h.p}{=} O(\frac{d}{n})$ when $d > \sqrt{n}$ and $\eta \overset{w.h.p}{=} O(1/\sqrt{n})$, o.w. Thus, the condition on $B$ in (5) simplifies as $B = O(\min(n^{2/3}r^{1/3}, (n/d)^{1.5}))$ when $d > \sqrt{n}$ and $B = O(\min(n^{2/3}r^{1/3}, n^{0.75}))$ o.w. Recall that as before, we require sparsity level of a fibre in any block as $d = O(n/r\mu^3)^{2/3}$. For simplicity, we assume that $\mu = O(1)$ for the remaining section.

We now explicitly compute the sparsity level of $S^*$ allowed by our method and compare it to the level allowed by matrix based robust PCA.
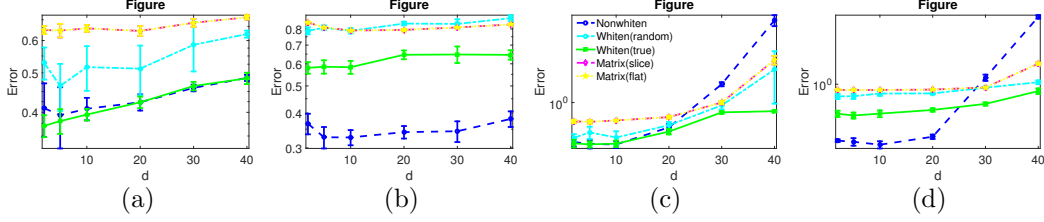
Figure 1: (a) Error comparison of different methods with deterministic sparsity, rank 5, varying $d$. (b) Error comparison of different methods with deterministic sparsity, rank 25, varying $d$. (c) Error comparison of different methods with block sparsity, rank 5, varying $d$. (d) Error comparison of different methods with block sparsity, rank 25, varying $d$. Error $= \|L^* - L\|_F / \|L^*\|_F$. The curve labeled 'T-RPCA-W(slice)' refers to considering recovered low rank part from a random slice of the tensor $T$ by using matrix non-convex RPCA method as the whiten matrix, 'T-RPCA-W(true)' is using true second order moment in whitening, 'M-RPCA(slice)' treats each slice of the input tensor as a non-convex matrix-RPCA(M-RPCA) problem, 'M-RPCA(flat)' reshapes the tensor along one mode and treat the resultant as a matrix RPCA problem. All four sub-figures share same curve descriptions.
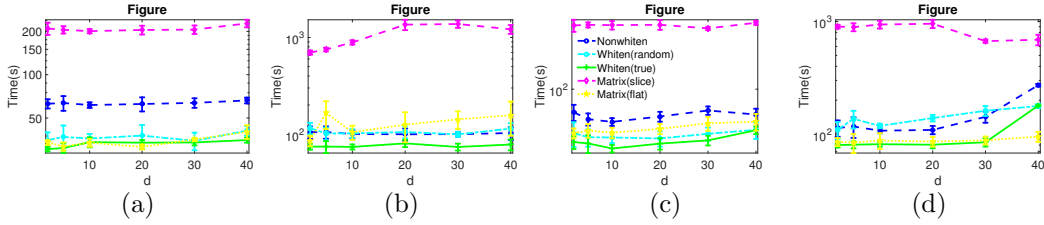


Figure 2: (a) Running time comparison of different methods with deterministic sparsity, rank 5, varying $d$. (b) Running time comparison of different methods with deterministic sparsity, rank 25, varying $d$. (c) Running time comparison of different methods with block sparsity, rank 5, varying $d$. (d) Running time comparison of different methods with block sparsity, rank 25, varying $d$. Curve descriptions are same as in Figure 1.

**Corollary 2** (Superiority of tensor methods under random sparsity). *Let $D_{\mathrm{RTD}}$ be the number of non-zeros in $S^*$ ($\|S^*\|_0$) allowed by RTD under the analysis of Theorem 1 and let $D_{\mathrm{matrix}}$ be the allowed $\|S^*\|_0$ using the standard matrix robust PCA analysis. Also, let $S^*$ be sampled from the random block sparse model. Then, the following holds:*

$$\frac{D_{\mathrm{RTD}}}{D_{\mathrm{matrix}}} = \begin{cases} \Omega\left(n^{1/6} r^{4/3}\right), & \text{for } r < n^{0.25}, \tag{6} \\ \Omega\left(n^{5/12} r^{1/3}\right), & o.w. \tag{7} \end{cases}$$

**Unstructured sparse perturbations** $S^*$: If we do not assume block sparsity in (S), but instead assume unstructured sparsity at level $D$, i.e. the number of non zeros along any fibre of $S^*$ is at most $D$, then the matrix methods are superior. In this case, for success of RTD, we require that $D = O\left(\frac{\sqrt{n}}{r\mu^3}\right)$ which is worse than the requirement of matrix methods $D = O(\frac{n}{r\mu^2})$. Our analysis suggests that if there is no structure in sparse patterns, then matrix methods are superior. This is possibly due to the fact that finding a low rank tensor decomposition requires more stringent conditions on the noise level. Meanwhile, when there is no block structure, the tensor algebraic constraints do not add significantly new information. However, in the experiments, we find that our tensor method RTD is superior to matrix methods even in this case, in terms of both accuracy and running times.

## 3.1 Analysis of Procedure 1

Our proof of Theorem 1 depends critically on an assumption that Procedure 1 indeed obtains the top-$r$ eigen-pairs. We now concretely prove this claim. Let $\widetilde{L}$ be a symmetric tensor which is a perturbed version of an orthogonal tensor $L^*$, $\widetilde{L} = L^* + E \in \mathbb{R}^{n \times n \times n}$, $L^* = \sum_{i \in [r]} \sigma_i^* u_i^{\otimes 3}$, where $\sigma_1^* \geq \sigma_2^* \ldots \sigma_r^* > 0$ and $\{u_1, u_2, \ldots, u_r\}$ form an orthonormal basis.

Recall that $N_1$ is the number of initializations to seed the power method, $N_2$ is the number of power iterations, and $\delta$ is the convergence criterion. For any $\xi \in (0,1)$, and $l \leq r$, assume the following

$$\|E\| \leq O(\sigma_l^*/\sqrt{n}),\ N_1 = O(n^{1+c}\log(1/\xi)),$$
$$N_2 \geq \Omega(\log(k) + \log\log(\sigma_{\max}^*/\|E\|)),$$

where $c$ is a small constant. We now state the main result for recovery of components of $L^*$ when Procedure 1 is applied to $\widetilde{L}$.

**Theorem 2** (Gradient Ascent method). *Let $\widetilde{L} = L^* + E$ be as defined above with $\|E\| \leq O(\sigma_r^*/\sqrt{n})$. Then, applying Procedure 1 with deflations on $\widetilde{L}$ with target rank $l \leq r$, yields $l$ eigen-pairs of $\widetilde{L}$, given by $(\lambda_1, \hat{u}_1), (\lambda_2, \hat{u}_2), \ldots, (\lambda_l, \hat{u}_l)$, up to arbitrary small error $\delta > 0$ and with probability at least $1 - \xi$. Moreover, there exists a permutation $\pi$ on $[l]$ such that: $\forall j \in [l]$,*

$$|\sigma_{\pi(j)}^* - \lambda_j| \leq O\left(\|E\| + \delta\right),$$
$$\|u_{\pi(j)} - \hat{u}_j\| \leq O((\|E\|/\sigma_{\pi(j)}^*) + \delta).$$

While [2, Thm. 5.1] considers power method, here we consider the power method followed by a gradient ascent procedure. With both methods, we obtain outputs $(\lambda_i, \hat{u}_i)$ which are "close" to the original eigen-pairs of $(\sigma_i^*, u_i)$ of $L^*$. However, the crucial difference is that Procedure 1 outputs $(\lambda_i, \hat{u}_i)$ correspond to specific eigen-pairs of input tensor $\tilde{L}$, while the outputs of the usual power method have no such property and only guarantees accuracy upto $O(\|E\|_2)$ error. We critically require the eigen property of the outputs in order to guarantee contraction of error in RTD between alternating steps of low rank decomposition and thresholding.

The analysis of Procedure 1 has two phases. In the first phase, we prove that with $N_1$ initializations and $N_2$ power iterations, we get close to true eigenpairs of $L^*$, i.e. $(\sigma_i^*, u_i)$ for $i \in [l]$. After this, in the second phase, we prove convergence to eigenpairs of $\widetilde{L}$.

The proof for the first phase is on lines of proof in [2], but with improved requirement on error tensor $E$ in (8). This is due to the use of SVD initializations rather than random initializations to seed the power method, and its analysis is given in [3].

Proof of the second phase follows using two observations: a) Procedure 1 is just a simple gradient ascent of the following program: $f(v) = \widetilde{L}(v,v,v) - \frac{3}{4}\lambda\|v\|_2^4$, b) with-in a small distance to the eigenvectors of $\widetilde{L}$, $f(v)$ is strongly concave and as well as strongly smooth with appropriate parameters. See below lemma for a detailed proof of the above claim. Hence, using our initialization guarantee from the phase-one, Procedure 1 converges to a $\delta$ approximation to eigen-pair of $\widetilde{L}$ in time $O(\log(1/\delta))$ and hence, Theorem 2 holds.

**Lemma 3.** *Let $f(v) = \widetilde{L}(v,v,v) - \frac{3}{4}\lambda\|v\|_2^4$. Then, $f$ is $\sigma_i^*(1 - \frac{300\sigma_r^*}{\sqrt{n}})$-strongly concave and $\sigma_i^*(1 + \frac{300\sigma_r^*}{\sqrt{n}})$ strongly smooth at all points $(v, \lambda)$ s.t. $\|v - u_i\| \leq \frac{10}{\sqrt{n}}$ and $|\lambda - \sigma_i^*| \leq \frac{10\sigma_r^*}{\sqrt{n}}$, for some $1 \leq i \leq r$. Procedure 1 converges to an eigenvector of $\widetilde{L}$ at a linear rate.*

*Proof.* Consider the gradient and Hessian of $f$ w.r.t. $v$:

$$\nabla f = 3\widetilde{L}(I, v, v) - 3\lambda\|v\|^2 v, \tag{8}$$
$$H = 6\widetilde{L}(I, I, v) - 6\lambda vv^\top - 3\lambda\|v\|^2 I. \tag{9}$$

We first note that the stationary points of $f$ indeed correspond to eigenvectors of $\widetilde{L}$ with eigenvalues $\lambda\|v\|^2$. Moreover, when $|\lambda - \sigma_i^*| \leq \frac{10\sigma_r^*}{\sqrt{n}}$ and $\|v - u_i\| \leq \frac{10}{\sqrt{n}}$, we have:

$$\|H - (-3\sigma_i^* I)\|_2 \leq 30\frac{\sigma_r^*}{\sqrt{n}} + 180\frac{\sigma_r^*}{\sqrt{n}}.$$

Recall that $\widetilde{L} = L^* + E$, where $L^*$ is an orthogonal tensor and $\|E\|_2 \leq \sigma_r^*/\sqrt{n}$. Hence, there exists one eigenvector in each of the above mentioned set, i.e., set of $(v, \lambda)$ s.t. $|\lambda - \sigma_i^*| \leq \frac{10\sigma_r^*}{\sqrt{n}}$ and $\|v - u_i\| \leq \frac{10}{\sqrt{n}}$. Hence, the standard gradient ascent procedure on $f$ would lead to convergence to an eigenvector of $\widetilde{L}$. ∎

8

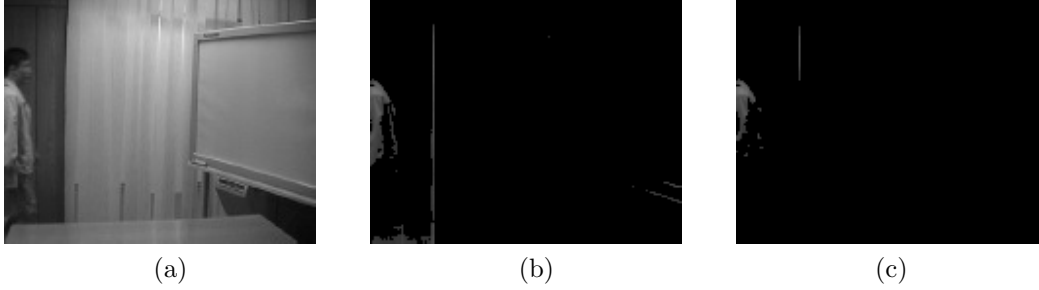(a)             (b)             (c)

Figure 3: Foreground filtering or activity detection in the *Curtain* video dataset. (a): Original image frame. (b): Foreground filtered (sparse part estimated) using tensor method; time taken is $5.1s$. (c): Foreground filtered (sparse part estimated) using matrix method; time taken is $5.7s$.

**Extending to non-orthogonal low rank tensors:** In (L), we assume that the low rank tensor $L^*$ in (1) is orthogonal. We can also extend to non-orthogonal tensors $L^*$, whose components $u_i$ are linearly independent. Here, there exists an invertible transformation $W$ known as *whitening* that orthogonalizes the tensors [2]. We can incorporate whitening in Procedure 1 to find low rank tensor decomposition, within the iterations of RTD.

# 4 Experiments

We now present an empirical study of our method. The goal of this study is three-fold: a) establish that our method indeed recovers the low-rank and sparse parts correctly, without significant parameter tuning, b) demonstrate that whitening during low rank decomposition gives computational advantages, c) show that our tensor methods are superior to matrix based RPCA methods in practice.

Our pseudo-code (Algorithm 1) prescribes the threshold $\zeta$ in Step 5, 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 $(t + 1)$ step of $l^{\text{th}}$ stage as $\zeta = \mu\sigma_{l+1}(T - S^{(t)})/n^{3/2}$. We found that the above thresholding, in the tensor case as well, provides exact recovery while speeding up the computation significantly.

**Synthetic datasets:** The low-rank part $L^* = \sum_{i\in[k]} \lambda_i u_i^{\otimes 3}$ is generated from a factor matrix $U \in \mathbb{R}^{n\times k}$ whose entries are i.i.d. samples from $\mathcal{N}(0, 1)$. For deterministic sparsity setting, $\text{supp}(S^*)$ is generated by setting each entry of $[n] \times [n] \times [n]$ to be non-zeros with probability $d/n$ and each non-zero value $S_{ijk}^*$ is drawn i.i.d. from the uniform distribution over $[r/(2n^{3/2}), r/n^{3/2}]$. For block sparsity setting, we randomly select $B$ numbers of $[n] \times [1]$ vectors $v_i, i = 1...B$ in which each entry is chosen to be non-zero with probability $d/n$. The value of non-zero entry is assigned similar to the one in deterministic sparsity case. Each of this vector will form a subtensor($v_i^{\otimes 3}$) and those subtensors form the whole $S$. For increasing incoherence of $L^*$, we randomly zero-out rows of $U$ and then re-normalize them. For the CP-decomposition, we use the alternating least squares (ALS) method available in the tensor toolbox [4]. Note that we use the ALS procedure in practice since we found that empirically, ALS performs quite well and is convenient to use. For whitening, we use two different whitening matrices: a) the true second order moment from the true low-rank part, b) the recovered low rank part from a random slice of the tensor $T$ by using matrix non-convex RPCA method. We compare our RTD with matrix non-convex RPCA in two ways: a) treat each slice of the input tensor as a matrix RPCA problem, b) reshape the tensor along one mode and treat the resultant as a matrix RPCA problem.

We vary sparsity of $S^*$ and rank of $L^*$ for RTD with a fixed tensor size. We investigate performance of our method, both with and without whitening, and compare with the state-of-the-art matrix non-convex RPCA algorithm. Our results for synthetic datasets is averaged over 5 runs. In Figure 1, we report relative error $(\|L^*-L\|_F/\|L^*\|_F)$ by each of the methods allowing maximum number of iterations up to 100. Comparing (a) and (b) in Figure 1, we can see that with block sparsity, RTD is better than matrix based non-convex RPCA method when $d$ is less than 20. If we use whitening, the advantage of RTD is more significant. But when rank becomes higher, the whitening method is not helpful. In Figure 2, we illustrate the computational time

9

of each methods. We can see that whitening simplifies the problem and give us computational advantage. Besides, the running time for the one using tensor method is similar to the one using matrix method when we reshape the tensor as one matrix. Doing matrix slices increases the running time.

**Real-world dataset:** To demonstrate the advantage of our method, we apply our method to *activity detection* or *foreground filtering* [21]. The goal of this task is to detect activities from a video coverage, which is a set of image frames that form a tensor. In our robust decomposition framework, the moving objects correspond sparse (foreground) perturbations which we wish to filter out. The static ambient background is of lesser interest since nothing changes.

We selected one of datasets, namely the *Curtain* video dataset wherein a person walks in and out of the room between the frame numbers 23731 and 23963. We compare our tensor method with the state-of-the-art matrix method in [22] on the set of 233 frames where the activity happens. In our tensor method, we preserve the multi-modal nature of videos and consider the set of image frames without vectorizing them. For the matrix method, we follow the setup of [22] by reshaping each image frame into a vector and stacking them together. We set the convergence criterion to $10^{-3}$ and run both the methods. Our tensor method yields a 10% speedup and obtains a better visual recovery for the same convergence accuracy as shown in Figure 3.

# 5   Conclusion

We proposed a non-convex alternating method for decomposing a tensor into low rank and sparse parts. We established convergence to the globally optimal solution under natural conditions such as incoherence of the low rank part and bounded sparsity levels for the sparse part. We prove that our proposed tensor method can handle perturbations at a much higher sparsity level compared to robust matrix methods. Our simulations show superior performance of our tensor method, both in terms of accuracy and computational time. Some future directions are analyzing: (1) our method with whitening (2) the setting where grossly corrupted samples arrive in streaming manner.

# References

[1] A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A Tensor Spectral Approach to Learning Mixed Membership Community Models. In *Conference on Learning Theory (COLT)*, June 2013.

[2] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *J. of Machine Learning Research*, 15:2773–2832, 2014.

[3] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *ArXiv 1402.5180*, 2014.

[4] Brett W. Bader, Tamara G. Kolda, et al. Matlab tensor toolbox version 2.6. Available online, February 2015.

[5] B. Barak and A. Moitra. Tensor Prediction, Rademacher Complexity and Random 3-XOR. *ArXiv e-prints 1501.06521*, January 2015.

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

[7] Dustin Cartwright and Bernd Sturmfels. The number of eigenvalues of a tensor. *Linear algebra and its applications*, 438(2):942–952, 2013.

[8] 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.

[9] Kung-Ching Chang, Kelly Pearson, Tan Zhang, et al. Perron-frobenius theorem for nonnegative tensors. *Commun. Math. Sci*, 6(2):507–520, 2008.

[10] Silvia Gandy, Benjamin Recht, and Isao Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2):025010, 2011.

[11] Donald Goldfarb and Zhiwei Qin. Robust low-rank tensor recovery: Models and algorithms. *SIAM Journal on Matrix Analysis and Applications*, 35(1):225–253, 2014.

[12] Quanquan Gu, Huan Gui, and Jiawei Han. Robust tensor decomposition with gross corruption. In Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 1422–1430. Curran Associates, Inc., 2014.

[13] Quanquan Gu, Huan Gui, and Jiawei Han. Robust tensor decomposition with gross corruption. In *Advances in Neural Information Processing Systems*, pages 1422–1430, 2014.

[14] Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45, 2013.

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

[16] Bo Huang, Cun Mu, Donald Goldfarb, and John Wright. Provable low-rank tensor recovery. *Preprint*, 2014.

[17] Heng Huang and Chris Ding. Robust tensor factorization using $r_1$ norm. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE, 2008.

[18] Prateek Jain and Sewoong Oh. Learning mixtures of discrete product distributions using spectral decompositions. 2014.

[19] Tamara G. Kolda and Jackson R. Mayo. Shifted power method for computing tensor eigenpairs. *SIAM J. Matrix Analysis Applications*, 32(4):1095–1124, 2011.

[20] Nadia Kreimer, Aaron Stanton, and Mauricio D Sacchi. Tensor completion based on nuclear norm minimization for 5d seismic data reconstruction. *Geophysics*, 78(6):V273–V284, 2013.

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

[22] Praneeth Netrapalli, UN Niranjan, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

[23] Parikshit Shah, Nikhil Rao, and Gongguo Tang. Sparse and low-rank tensor decomposition. In *Advances in Neural Information Processing Systems*, 2015.

[24] Stefan Ragnarsson and Charles F Van Loan. Block tensors and symmetric embeddings. *Linear Algebra and Its Applications*, 438(2):853–874, 2013.

[25] Ryota Tomioka, Kohei Hayashi, and Hisashi Kashima. Estimation of low-rank tensors via convex optimization. *arXiv preprint arXiv:1010.0789*, 2010.

# A  Bounds for block sparse tensors

One of the main bounds to control is the spectral norm of the sparse perturbation tensor $S$. The success of the power iterations and the improvement in accuracy of recovery over iterative steps of RTD requires this bound.

**Lemma 4** (Spectral norm bounds for block sparse tensors). *Let $M \in \mathbb{R}^{n \times n \times n}$ satisfy the block sparsity assumption (S). Then*

$$\|M\|_2 = O(d^{1.5} \|M\|_\infty). \tag{10}$$

*Proof.* Let $\Psi \in \mathbb{R}^{n \times n \times n}$ be a tensor that encodes the sparsity of $M$ i.e. $\Psi_{i,j,k} = 1$ iff $S^*_{i,j,k} \neq 0$ for all $i, j, k \in [n]$. We have that

$$\|M\| = \max_{u : \|u\| = 1} \sum_{i,j,k} M_{i,j,k} u(i) u(j) u(k)$$

$$= \max_{u : \|u\| = 1} \sum_{i,j,k} M_{i,j,k} \Psi_{i,j,k} u(i) u(j) u(k)$$

$$\leq \max_{u : \|u\| = 1} \sum_{i,j,k} |M_{i,j,k} \Psi_{i,j,k} u(i) u(j) u(k)|$$

$$\leq \|M\|_\infty \max_{u : \|u\| = 1} \sum_{i,j,k} |\Psi_{i,j,k} u(i) u(j) u(k)| = \|M\|_\infty \|\Psi\|,$$

where the last inequality is from Perron Frobenius theorem for non-negative tensors [9]. Note that $\Psi$ is non-negative by definition. Now we bound $\|\Psi\|$ on lines of [3, Lemma 4]. Recall that $\forall i \in [B]$, $j \in [n]$,

$$\Psi = \sum_{i=1}^{B} \psi_i \otimes \psi_i \otimes \psi_i, \quad \|\psi_i\|_0 \leq d, \, \psi_i(j) = 0 \text{ or } 1.$$

By definition $\|\psi_i\|_2 = \sqrt{d}$. Define normalized vectors $\tilde{\psi}_i := \psi_i / \|\psi_i\|$. We have

$$\Psi = d^{1.5} \sum_{i=1}^{B} \tilde{\psi}_i \otimes \tilde{\psi}_i \otimes \tilde{\psi}_i$$

Define matrix $\tilde{\psi} := [\tilde{\psi}_1 | \tilde{\psi}_2, \dots \tilde{\psi}_B]$. Note that $\tilde{\psi}^\top \tilde{\psi} \in \mathbb{R}^{B \times B}$ is a matrix with unit diagonal entries and absolute values of off-diagonal entries bounded by $\eta$, by assumption. From Gershgorin Disk Theorem, every subset of $L$ columns in $\tilde{\psi}$ has singular values within $1 \pm o(1)$, where $L < \frac{1}{\eta}$. Moreover, from Gershgorin Disk Theorem, $\|\tilde{\psi}\| < \sqrt{1 + B\eta}$.

For any unit vector $u$, let $S$ be the set of $L$ indices that are largest in $\tilde{\psi}^\top u$. By the argument above we know $\|(\tilde{\psi}_S)^\top u\| \leq \|\tilde{\psi}_S\| \|u\| \leq 1 + o(1)$. In particular, the smallest entry in $\tilde{\psi}_S^\top u$ is at most $2/\sqrt{L}$. By construction of $S$ this implies for all $i$ not in $S$, $|\tilde{\psi}_i^\top u|$ is at most $2/\sqrt{L}$. Now we can write the $\ell_3$ norm of $\tilde{\psi}^\top u$ as

$$\|\tilde{\psi}^\top u\|_3^3 = \sum_{i \in S} |\tilde{\psi}_i^\top u|^3 + \sum_{i \notin S} |\tilde{\psi}_i^\top u|^3$$

$$\leq \sum_{i \in S} |\tilde{\psi}_i^\top u|^2 + (2/\sqrt{L})^{3-2} \sum_{i \notin S} |\tilde{\psi}_i^\top u|^2$$

$$\leq 1 + 2\sqrt{\eta} \|\tilde{\psi}\|^2 \leq 1 + 2B\eta^{1.5}.$$

Here the first inequality uses that every entry outside $S$ is small, and last inequality uses the bound argued on $\|(\tilde{\psi}_S)^\top u\|$, the spectral norm bound is assumed on $A_{S^c}$. Since $B = O(\eta^{-1.5})$, we have the result. ∎

Another important bound required is $\infty$-norm of certain contractions of the (normalized) sparse tensor and its powers, which we denote by $M$ below. We use a loose bound based on spectral norm and we require $\|M\| < 1/\sqrt{n}$. However, this constraint will also be needed for the power iterations to succeed and is not an additional requirement. Thus, the loose bound below will suffice for our results to hold.

**Lemma 5** (Infinity norm bounds). *Let $M \in \mathbb{R}^{n \times n \times n}$ satisfy the block sparsity assumption (S). Let $u, v$ satisfy the assumption $\|u\|_\infty, \|v\|_\infty \leq \frac{\mu}{n^{1/2}}$. Then, we have*

1. $\|M(u, v, I)\|_\infty \leq \frac{\kappa \mu}{n^{1/2}} \|M\|_\infty$, *where* $\kappa := \frac{Bd^2\mu}{\sqrt{n}}$.

2. $\|[M(u, v, I)]^p\|_\infty \leq \kappa \mu \|M\|_\infty \|M\|^{p-1}$ *for* $p > 1$.

3. $\sum_{p \geq 1} \|[M(u, I, I)]^p v\|_\infty \leq \frac{\kappa \mu}{\sqrt{n}} \|M\|_\infty \cdot \frac{\|M\|}{1 - \|M\|}$ *when* $\|M\| < 1/\sqrt{n}$.

*Proof.* We have from norm conversion

$$\|M(u, v, I)\|_\infty \leq \|u\|_\infty \cdot \|v\|_\infty \max_j \|M(I, I, e_j)\|_1 \tag{11}$$

$$\leq \frac{\mu^2}{n} \cdot Bd^2 \|M\|_\infty, \tag{12}$$

where $\ell_1$ norm (i.e. sum of absolute values of entries) of a slice $M(I, I, e_j)$ is $Bd^2$, since the number of non-zero entries in one block in a slice is $d^2$.

Let $Z = M(u, I, I) \in \mathbb{R}^{n \times n}$. Now, $\|M(u, I, I)^p v\|_\infty = \|Z^p v\|_\infty = \|Z^{p-1}a\|_\infty$ where $a = Zv$. Now,

$$\|Z^{p-1}a\|_\infty = \max_j |e_j^T Z^{p-1}a| \leq \|Z^{p-1}\|_2 \|a\|_2 \leq \|Z\|_2^{p-1} \|a\|_2 \leq \|M\|^{p-1} \sqrt{n} \|a\|_\infty \leq \kappa \mu \|M\|_\infty \|M\|^{p-1}.$$

Hence, $\sum_{p \geq 1} \|[M(u, I, I)]^p v\|_\infty \leq \kappa \mu \|M\|_\infty \cdot \frac{\|M\|_2}{1 - \|M\|_2}$. ∎

# B   Proof of Theorem 1

**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 $l^{th}$ stage of Algorithm 1. Let $\sigma_1^*, \ldots, \sigma_r^*$ be the eigenvalues of $L^*$, such that $\sigma_1^* \geq \cdots \geq \sigma_r^* \geq 0$ and $\lambda_1, \cdots, \lambda_r$ be the eigenvalues of $T - S^{(t)}$ such that $\lambda_1 \geq \cdots \geq \lambda_r \geq 0$. Recall that $E^{(t)} := S^* - S^{(t)}$. Suppose further that*

1. $\|E^{(t)}\|_\infty \leq \frac{8\mu^3 k}{n^{3/2}} \left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^{t-1} \sigma_l^*\right)$, *and*

2. $\operatorname{supp} E^{(t)} \subseteq \operatorname{supp} S^*$.

*Then, for some constant $c \in [0, 1)$, we have*

$$(1 - c)\left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^t \sigma_l^*\right) \leq \left(\lambda_{l+1} + \left(\frac{1}{2}\right)^t \lambda_l\right) \leq (1 + c)\left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^t \sigma_l^*\right). \tag{13}$$

*Proof.* Note that $T - S^{(t)} = L^* + E^{(t)}$. Now,

$$\left|\lambda_{l+1} - \sigma_{l+1}^*\right| \leq 8\left\|E^{(t)}\right\|_2 \leq 8d^{3/2} \|E^{(t)}\|_\infty \leq \frac{8\mu^3 r \gamma_t}{n^{3/2}} d^{3/2},$$

where $\gamma_t := \left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^{t-1} \sigma_l^*\right)$. That is, $\left|\lambda_{l+1} - \sigma_{l+1}^*\right| \leq 8\mu^3 r \left(\frac{d}{n}\right)^{3/2} \gamma_t$. Similarly, $\left|\lambda_l - \sigma_l^*\right| \leq 8\mu^3 r \left(\frac{d}{n}\right)^{3/2} \gamma_t$.
So we have:

$$\left|\left(\lambda_{l+1} + \left(\frac{1}{2}\right)^t \lambda_l\right) - \left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^t \sigma_l^*\right)\right| \leq 8\mu^3 r \left(\frac{d}{n}\right)^{3/2} \gamma_t \left(1 + \left(\frac{1}{2}\right)^t\right)$$

$$\leq 16\mu^3 r \left(\frac{d}{n}\right)^{3/2} \gamma_t$$

$$\leq c\left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^t \sigma_l^*\right),$$

where the last inequality follows from the bound $d \leq \left(\frac{n}{c'\mu^3 k}\right)^{2/3}$ for some constant $c'$. ∎

14

**Lemma 7.** *Assume the notation of Lemma 6. Also, let $L^{(t)}, S^{(t)}$ be the $t^{th}$ iterates of $r^{th}$ 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. *$\|E^{(t)}\|_\infty \leq \frac{8\mu^3 r}{n^{3/2}}\left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^{t-1}\sigma_l^*\right)$, and*

2. *$\operatorname{supp} E^{(t)} \subseteq \operatorname{supp} S^*$.*

3. *$\|E^{(t)}\|_2 < \frac{C\sigma_l^*}{\sqrt{n}}$, where $C < 1/2$ is a sufficiently small constant.*

*Then, we have:*

$$\|L^{(t+1)} - L^*\|_\infty \leq 2\frac{\mu^3 r}{n^{3/2}}\left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^t \sigma_l^*\right)$$

*Proof.* Let $L^{(t+1)} = \sum_{i=1}^l \lambda_i u_i^{(t+1)}$ be the eigen decomposition obtained using the tensor power method on $(T - S^{(t)})$ at the $(t+1)^{th}$ step of the $l^{th}$ stage. Also, recall that $T - S^{(t)} = L^* + E^{(t)}$ where $L^* = \sum_{j=1}^r \sigma_j^* u_j^{\otimes 3}$. Define $E^{(t)} := S^* - S^{(t)}$. Define $E^i := E^{(t)}(u_i^{(t+1)}, I, I)$. Let $\|E^{(t)}\|_2 := \epsilon$.

Consider the eigenvalue equation $(T - S^{(t)})(u_i^{(t+1)}, u_i^{(t+1)}, I) = \lambda_i u_i^{(t+1)}$:

$$L^*(u_i^{(t+1)}, u_i^{(t+1)}, I) + E^{(t)}(u_i^{(t+1)}, u_i^{(t+1)}, I) = \lambda_i u_i^{(t+1)}$$

$$\sum_{j=1}^r \sigma_i^* \left\langle u_i^{(t+1)}, u_j \right\rangle^2 u_j + E^{(t)}(u_i^{(t+1)}, u_i^{(t+1)}, I) = \lambda_i u_i^{(t+1)}$$

$$[\lambda_i I - E^{(t)}(u_i^{(t+1)}, I, I)]u_i^{(t+1)} = \sum_{j=1}^r \sigma_i^* \left\langle u_i^{(t+1)}, u_j \right\rangle^2 u_j$$

$$u_i^{(t+1)} = \left[I + \sum_{p\geq 1}\left(\frac{E^i}{\lambda_i}\right)^p\right]\sum_{j=1}^r \frac{\sigma_i^*}{\lambda_i}\left\langle u_i^{(t+1)}, u_j \right\rangle^2 u_j$$

Now,

$$\|L^{(t+1)} - L^*\|_\infty \leq \left\|\sum_{i\in[l]}\lambda_i(u_i^{(t+1)})^{\otimes 3} - \sum_{i\in[l]}\sigma_i^* u_i^{\otimes 3}\right\|_\infty + \left\|\sum_{i=l+1}^r \sigma_i^* u_i^{\otimes 3}\right\|_\infty$$

$$\leq \sum_{i\in[l]}\left\|\lambda_i(u_i^{(t+1)})^{\otimes 3} - \sigma_i^* u_i^{\otimes 3}\right\|_\infty + \sum_{i=l+1}^r \|\sigma_i^* u_i^{\otimes 3}\|_\infty$$

For a fixed $i$, using $\lambda_i \leq \sigma_i^* + \epsilon$ [2] and using Lemma 11, we obtain

$$\left\|\lambda_i(u_i^{(t+1)})^{\otimes 3} - \sigma_i^* u_i^{\otimes 3}\right\|_\infty \leq \left\|(\sigma_i^* + \epsilon)(u_i^{(t+1)})^{\otimes 3} - \sigma_i^* u_i^{\otimes 3}\right\|_\infty$$

$$\leq \left\|\sigma_i^*(u_i^{(t+1)})^{\otimes 3} - \sigma_i^* u_i^{\otimes 3}\right\|_\infty + \epsilon\left\|(u_i^{(t+1)})^{\otimes 3}\right\|_\infty$$

$$\leq \sigma_i^*\left\|(u_i^{(t+1)})^{\otimes 3} - u_i^{\otimes 3}\right\|_\infty + \epsilon\left\|(u_i^{(t+1)})^{\otimes 3}\right\|_\infty$$

$$\leq \sigma_i^*[3\|u_i^{(t+1)} - u_i\|_\infty\|u_i\|_\infty^2 + 3\|u_i^{(t+1)} - u_i\|_\infty^2\|u_i\|_\infty + \|u_i^{(t+1)} - u_i\|_\infty^3]$$

$$+ \epsilon\|(u_i^{(t+1)})^{\otimes 3}\|_\infty$$

$$\leq 7\sigma_i^*\|u_i^{(t+1)} - u_i\|_\infty\|u_i\|_\infty^2 + \epsilon\|(u_i^{(t+1)})\|_\infty^3$$

15

Now,

$$\left\| u_i^{(t+1)} - u_i \right\|_\infty = \left\| (\sum_{j=1}^r \frac{\sigma_i^*}{\lambda_i} \left\langle u_i^{(t+1)}, u_j \right\rangle^2 u_j - u_i) + \sum_{j=1, p\geq 1}^r \frac{\sigma_i^*}{\lambda_i} \left\langle u_i^{(t+1)}, u_j \right\rangle^2 (E^i)^p u_j \right\|_\infty$$

$$\leq \left\| (1 - \frac{\sigma_i^*}{\lambda_i} \left\langle u_i^{(t+1)}, u_i \right\rangle^2) u_i \right\|_\infty + \left\| \sum_{j\neq i} \frac{\sigma_i^*}{\lambda_i} \left\langle u_i^{(t+1)}, u_j \right\rangle^2 u_j \right\|_\infty$$

$$+ \left\| \sum_{p\geq 1} \frac{\sigma_i^*}{\lambda_i} \left\langle u_i^{(t+1)}, u_i \right\rangle^2 \left(\frac{E^i}{\lambda_i}\right)^p u_i \right\|_\infty + \left\| \sum_{p, j\neq i} \frac{\sigma_i^*}{\lambda_i} \left\langle u_i^{(t+1)}, u_j \right\rangle^2 \left(\frac{E^i}{\lambda_i}\right)^p u_j \right\|_\infty$$

For the first term, we have

$$\left\| (1 - \frac{\sigma_i^*}{\lambda_i} \left\langle u_i^{(t+1)}, u_i \right\rangle^2) u_i \right\|_\infty \leq \left( 1 - \frac{\sigma_i^*}{\sigma_i^* + \epsilon} \left( 1 - \left(\frac{\epsilon}{\sigma_i^*}\right)^2 \right) \right) \| u_i \|_\infty \leq \left( 1 - \left( 1 - \frac{\epsilon}{\sigma_i^*} \right) \right) \frac{\mu}{n^{1/2}}$$

$$\leq \frac{\mu}{\sigma_i^* n^{1/2}} \epsilon \leq \frac{C\mu\sigma_l^*}{\sigma_i^* n}$$

where we substitute for $\epsilon$ in the last step.

For the second term, we have

$$\left\| \sum_{j\neq i} \frac{\sigma_i^*}{\lambda_i} \left\langle u_i^{(t+1)}, u_j \right\rangle^2 u_j \right\|_\infty \leq \frac{\sigma_i^*}{\sigma_i^* - \epsilon} \left(\frac{\epsilon}{\sigma_i^*}\right)^2 \| u_i \|_\infty \leq 2 \left(\frac{\epsilon}{\sigma_i^*}\right)^2 \frac{\mu}{n^{1/2}},$$

which is a lower order term.

Next,

$$\left\| \sum_{p\geq 1} \frac{\sigma_i^*}{\lambda_i} \left\langle u_i^{(t+1)}, u_i \right\rangle^2 \left(\frac{E^i}{\lambda_i}\right)^p u_i \right\|_\infty \leq \left\| \sum_{p\geq 1} \frac{\sigma_i^*}{\lambda_i} \left(\frac{E^i}{\lambda_i}\right)^p u_i \right\|_\infty \leq \sum_{p\geq 1} \frac{\sigma_i^*}{\lambda_i} \left\| \left(\frac{E^i}{\lambda_i}\right)^p u_i \right\|_\infty$$

$$\leq \frac{\sigma_i^*}{\lambda_i} \cdot \frac{\mu}{\sqrt{n}} \cdot \frac{\| E^{(t)} \|_2 \sqrt{n}/\lambda_i}{1 - \| E^{(t)} \|_2 \sqrt{n}/\lambda_i}$$

$$\leq \frac{2}{(1-C)} \frac{\kappa_t \mu}{\lambda_i \sqrt{n}} \| E^{(t)} \|_\infty$$

from Lemma 5, and the assumption on spectral norm of $\| E^{(t)} \|_2$, where

$$\kappa_t := \frac{Bd^2\mu}{\sqrt{n}}.$$

For the remaining terms, we have

$$\left\| \sum_{p, j\neq i} \frac{\sigma_i^*}{\lambda_i} \left\langle u_i^{(t+1)}, u_j \right\rangle^2 \left(\frac{E^i}{\lambda_i}\right)^p u_j \right\|_\infty \leq \sum_{j\neq i} \frac{\sigma_i^*}{\lambda_i} \left\langle u_i^{(t+1)}, u_j \right\rangle^2 \left\| \sum_{p\geq 1} \left(\frac{E^i}{\lambda_i}\right)^p u_1 \right\|_\infty \leq \frac{\sigma_i^*}{\lambda_i} \left\| \sum_{p\geq 1} \left(\frac{E^i}{\lambda_i}\right)^p u_1 \right\|_\infty \left(\frac{\epsilon}{\sigma_i^*}\right)^2,$$

which is a lower order term.

Combining the above and recalling $\epsilon \ll \sigma_i^*$, $\forall i \in [l]$

$$\left\| u_i^{(t+1)} - u_i \right\|_\infty \leq \frac{8}{(1-C)} \frac{\kappa_t \mu}{\lambda_i \sqrt{n}} \| E^{(t)} \|_\infty.$$

Also, from Lemma 1

$$|\lambda_i - \sigma_i^*| \leq 8 \| E^{(t)} \|_2 \leq 8\epsilon$$

16

Thus, from the above two equations, we obtain the bound for the parameters (eigenvectors and eigenvalues) of the low-rank tensor $\left\|u_i^{(t+1)} - u_i\right\|_\infty$ and $\|\lambda_i - \sigma_i^*\|_\infty$. We combine the individual parameter recovery bounds as:

$$\left\|\sum_{i\in[l]}\lambda_i(u_i^{(t+1)})^{\otimes 3} - \sum_{i\in[l]}\sigma_i^* u_i^{\otimes 3}\right\|_\infty \leq r[7\sigma_i^*\|u_i^{(t+1)} - u_i\|_\infty\|u_i\|_\infty^2 + \epsilon\|(u_i^{(t+1)})\|_\infty^3]$$

$$\leq \frac{224}{1-C}\frac{\kappa_t\mu^3 r}{n^{1.5}}\|E^{(t)}\|_\infty \tag{14}$$

and the other term

$$\sum_{i=l+1}^{r}\|\sigma_i^* u_i^{\otimes 3}\|_\infty \leq \sigma_{l+1}^*\frac{r\mu^3}{n^{1.5}}.$$

Combining bound in (14) with the above, we have

$$\|L^{(t+1)} - L^*\|_\infty \leq \frac{r\mu^3}{n^{1.5}}\left(\frac{224}{1-C}\kappa_t\|E^{(t)}\|_\infty + \sigma_{l+1}^*\right) < \frac{1}{4}\|E^{(t)}\|_\infty.$$

where the last inequality comes from the fact that $\frac{r\mu^3}{n^{1.5}}\sigma_{l+1}^* \leq \frac{\|E^{(t)}\|_\infty}{8}$ and the assumption that $C < 1/2$, and we can choose $B$ and $d$ s.t.

$$\frac{448r\mu^3}{n^{1.5}}\kappa_t < \frac{1}{8}.$$

This is possible from assumption (S). $\blacksquare$

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

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

$$\|L^{(t+1)} - L^*\|_\infty \leq 2\frac{\mu^3 r}{n^{3/2}}\left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^{t-1}\sigma_l^*\right).$$

*Then, we have:*

*1.* supp $E^{(t+1)} \subseteq$ supp $S^*$.

*2.* $\|E^{(t+1)}\|_\infty \leq 7\frac{\mu^3 r}{n^{3/2}}\left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^t\sigma_l^*\right)$, *and*

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

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

where $\zeta = 4\frac{\mu^3 r}{n^{3/2}}\left(\lambda_{l+1} + \left(\frac{1}{2}\right)^t\lambda_l\right)$ is as defined in Algorithm 1 and $\lambda_1, \cdots, \lambda_n$ are the eigenvalues of $T - S^{(t)}$ such that $\lambda_1 \geq \cdots \geq \lambda_n$.

If $S_{abc}^* = 0$ then $E_{ijk}^{(t+1)} = \mathbf{1}_{\left|L_{abc}^* - L_{abc}^{(t+1)}\right| > \zeta}\cdot(L_{abc}^* - L_{abc}^{(t+1)})$. The first part of the lemma now follows by using the assumption that $\|L^{(t+1)} - L^*\|_\infty \leq 2\frac{\mu^3 r}{n^{3/2}}\left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^t\sigma_l^*\right) \overset{(\zeta_1)}{\leq} 4\frac{\mu^3 r}{n^{3/2}}\left(\lambda_{l+1} + \left(\frac{1}{2}\right)^t\lambda_l\right) = \zeta$, where $(\zeta_1)$ follows from Lemma 6.

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

1. $\left|T_{abc} - L_{abc}^{(t+1)}\right| > \zeta$: Here, $S_{abc}^{(t+1)} = S_{abc}^* + L_{abc}^* - L_{abc}^{(t+1)}$. Hence, $|S_{abc}^{(t+1)} - S_{abc}^*| \leq |L_{abc}^* - L_{abc}^{(t+1)}| \leq 2\frac{\mu^3 r}{n^{3/2}}\left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^t\sigma_l^*\right)$.

2. $\left|T_{abc} - L_{abc}^{(t+1)}\right| \leq \zeta$: In this case, $S_{abc}^{(t+1)} = 0$ and $\left|S_{abc}^* + L_{abc}^* - L_{abc}^{(t+1)}\right| \leq \zeta$. So we have, $\left|E_{abc}^{(t+1)}\right| = |S_{abc}^*| \leq \zeta + \left|L_{abc}^* - L_{abc}^{(t+1)}\right| \leq 7\frac{\mu^3 r}{n^{3/2}}\left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^t\sigma_l^*\right)$. The last inequality above follows from Lemma 6.

This proves the lemma. ∎

**Theorem 9.** *Let $L^*, S^*$ be symmetric and satisfy (L) and (S), and $\beta = 4\frac{\mu^3 r}{n^{3/2}}$. The outputs $\widehat{L}$ (and its parameters $\hat{u}_i$ and $\hat{\lambda}_i$) and $\widehat{S}$ of Algorithm 1 satisfy w.h.p.:*

$$\|\hat{u}_i - u_i\|_\infty \leq \frac{\delta}{\mu^2 r n^{1/2} \sigma^*_{\min}}, \quad |\hat{\lambda}_i - \sigma_i^*| \leq \delta, \quad \forall i \in [n],$$

$$\left\|\widehat{L} - L^*\right\|_F \leq \delta, \quad \|\widehat{S} - S^*\|_\infty \leq \frac{\delta}{n^{3/2}}, \quad \text{and} \quad \text{supp}\,\widehat{S} \subseteq \text{supp}\,S^*.$$

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

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

$$\|E^{(0)}\|_\infty \leq \frac{8\mu^3 r}{n^{3/2}} \left(\sigma_2^* + 2\sigma_1^*\right), \text{ and}$$

$$\text{supp}\,E^{(0)} \subseteq \text{supp}\,S^*.$$

So the base case of induction is satisfied.

We first do the inductive step over $t$ (for a fixed $r$). By inductive hypothesis we assume that: a) $\|E^{(t)}\|_\infty \leq \frac{8\mu^3 r}{n^{3/2}} \left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^{t-1} \sigma_l^*\right)$, b) $\text{supp}\,E^{(t)} \subseteq \text{supp}\,S^*$. Then by Lemma 7, we have:

$$\|L^{(t+1)} - L^*\|_\infty \leq \frac{2\mu^3 r}{n^{3/2}} \left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^t \sigma_l^*\right).$$

Lemma 8 now tells us that

1. $\|E^{(t+1)}\|_\infty \leq \frac{8\mu^3 r}{n^{3/2}} \left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^t \sigma_l^*\right)$, and

2. $\text{supp}\,E^{(t+1)} \subseteq \text{supp}\,S^*$.

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

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

$$\left\|T - S^{(0)}\right\|_2 \geq \|L^*\|_2 - \left\|E^{(0)}\right\|_2 \geq \sigma_1^* - d^{3/2}\|E^{(0)}\|_\infty \geq \frac{3}{4}\sigma_1^*,$$

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

1. $\|E^{(T)}\|_\infty \leq \frac{7\mu^3 r}{n^{3/2}} \left(\sigma_{l+1}^* + \left(\frac{1}{2}\right)^T \sigma_l^*\right) \leq \frac{7\mu^3 r\sigma_{l+1}^*}{n^{3/2}} + \frac{\delta}{10n}$, and

2. $\text{supp}\,E^{(T)} \subseteq \text{supp}\,S^*$.

Recall, $\left|\sigma_{r+1}\left(T - S^{(T)}\right) - \sigma_{r+1}^*\right| \leq \left\|E^{(T)}\right\|_2 \leq \frac{d}{n}\left(\mu^3 r\left|\sigma_{r+1}^*\right| + \delta\right)$. We will now consider two cases:

1. **Algorithm 1 terminates:** This means that $\beta\sigma_{r+1}\left(T - S^{(T)}\right) < \frac{\delta}{2n^{3/2}}$ which then implies that $\sigma_{r+1}^* < \frac{\delta}{6\mu^3 r}$. So we have:

$$\|\widehat{L} - L^*\|_\infty = \|L^{(T)} - L^*\|_\infty \leq \frac{2\mu^3 r}{n^{3/2}} \left(\sigma_{r+1}^* + \left(\frac{1}{2}\right)^T \sigma_r^*\right) \leq \frac{\delta}{5n^{3/2}}.$$

This proves the statement about $\widehat{L}$ and its parameters (eigenvalues and eigenvectors). A similar argument proves the claim on $\|\widehat{S} - S^*\|_\infty$. The claim on $\text{supp}\,\widehat{S}$ follows since $\text{supp}\,E^{(T)} \subseteq \text{supp}\,S^*$.

2. **Algorithm 1 continues to stage $(r + 1)$:** This means that $\beta\sigma_{r+1}\left(L^{(T)}\right) \geq \frac{\delta}{2n^{3/2}}$ which then implies that $\sigma^*_{r+1} > \frac{\delta}{8\mu^3 r}$. So we have:

$$\|E^{(T)}\|_\infty \leq \frac{8\mu^3 r}{n^{3/2}}\left(\sigma^*_{r+1} + \left(\frac{1}{2}\right)^T \sigma^*_r\right)$$

$$\leq \frac{8\mu^3 r}{n^{3/2}}\left(\sigma^*_{l+1} + \frac{\delta}{10\mu^3 r n^{3/2}}\right)$$

$$\leq \frac{8\mu^3 r}{n^{3/2}}\left(\sigma^*_{l+1} + \frac{8\sigma^*_{l+1}}{10n}\right)$$

$$\leq \frac{8\mu^3 r}{n^{3/2}}\left(\sigma^*_{l+2} + 2\sigma^*_{l+1}\right).$$

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

This finishes the proof. ∎

## B.1   Short proof of Corollary 1

The state of art guarantees for robust matrix PCA requires that the overall sparsity along any row or column of the input matrix be $D = O(\frac{n}{r\mu^2})$ (when the input matrix is $\mathbb{R}^{n\times n}$).

Under (S), the total sparsity along any row or column of $M_i$ is given by $D := dB$. Now, Theorem 1 holds when the sparsity condition in (5) is satisfied. That is, RTD succeeds when

$$D = O(d \cdot B) = O\left(\min\left(\frac{n^{4/3}}{r^{1/3}\mu^2}, \frac{n^{2/3}}{r^{2/3}\mu^2}(\frac{n}{r})^{1/3}\right)\right) = O\left(\frac{n}{r\mu^2}\right).$$

Hence, RTD can handle larger amount of corruption than the matrix methods and the gain becomes more significant for smaller $\eta$.

## B.2   Some auxiliary lemmas

We recall Theorem 5.1 from [2]. Let $\epsilon = 8\|E^{(t)}\|_2$ where $E^{(t)} := S^* - S^{(t)}$.

**Lemma 10.** *Let $L^{(t+1)} = \sum_{i=1}^k \lambda_i u_i^{(t+1)}$ be the eigen decomposition obtained using Algorithm 1 on $(T - S^{(t)})$. Then,*

1. *If $\|u_i^{(t+1)} - u_i\|_2 \leq \frac{\epsilon}{\sigma^*_{\min}}$, then $\mathrm{dist}(u_i^{(t+1)}, u_i) \leq \frac{\epsilon}{\sigma^*_{\min}}$.*

2. *$\sum_{j\neq i}\left\langle u_i^{(t+1)}, u_j\right\rangle^2 \leq \left(\frac{\epsilon}{\sigma^*_{\min}}\right)^2$.*

3. *$\|u_i^{(t+1)}\|_\infty \leq \frac{\mu}{n^{1/2}} + \frac{\epsilon}{\sigma^*_{\min}}$.*

4. *$|\sigma^*_i| - \epsilon \leq |\lambda_i| \leq |\sigma^*_i| + \epsilon$.*

*Proof.*   1. Let $z \perp u$ and $\|z\|_2 = 1$.

$$u_i^{(t+1)} = \left\langle u_i^{(t+1)}, u_i\right\rangle u_i + \mathrm{dist}(u_i^{(t+1)}, u_i)z$$

$$\|u_i^{(t+1)} - u_i\|_2^2 = (\left\langle u_i^{(t+1)}, u_i\right\rangle - 1)^2\|u_i\|_2^2 + \mathrm{dist}(u_i^{(t+1)}, u_i)\|z\|_2^2 + 0$$

$$\geq (\mathrm{dist}(u_i^{(t+1)}, u_i))^2$$

Then using Theorem 5.1 from [2], we obtain the result. Next, since $\langle u_i^{(t+1)}, u_i\rangle^2 + \mathrm{dist}(u_i^{(t+1)}, u_i)^2 = 1$, we have $\langle u_i^{(t+1)}, u_i\rangle^2 \geq 1 - \left(\frac{\epsilon}{\sigma^*_{\min}}\right)^2$.

2. Note that

$$u_i^{(t+1)} = \sum_{j=1}^{k} \left\langle u_i^{(t+1)}, u_j \right\rangle u_j + \mathrm{dist}(u_i^{(t+1)}, U)z$$

where $z \perp U$ such that $\|z\|_2 = 1$. Using $\|u_i^{(t+1)}\|_2 = 1$ and the Pythagoras theorem, we get

$$1 - \left\langle u_i^{(t+1)}, u_i \right\rangle^2 = \sum_{j \neq i} \left\langle u_i^{(t+1)}, u_j \right\rangle^2 + \mathrm{dist}(u_i^{(t+1)}, U)^2.1 \geq \sum_{j \neq i} \left\langle u_i^{(t+1)}, u_j \right\rangle^2$$

Using part 1 of Lemma 10, we get $\sum_{j \neq i} \left\langle u_i^{(t+1)}, u_j \right\rangle^2 \leq \left( \frac{\epsilon}{\sigma_{\min}^*} \right)^2$.

3. We have

$$u_i^{(t+1)} = \left\langle u_i^{(t+1)}, u_i \right\rangle u_i + \mathrm{dist}(u_i^{(t+1)}, u_i)z$$

$$\|u_i^{(t+1)}\|_\infty \leq | \left\langle u_i^{(t+1)}, u_i \right\rangle |\|u_i\|_\infty + | \mathrm{dist}(u_i^{(t+1)}, u_i)|\|z\|_\infty \leq 1.\frac{\mu}{n^{1/2}} + \frac{\epsilon}{\sigma_{\min}^*}$$

4. This follows from Theorem 5.1 from [2], i.e., $\forall i, \; ||\lambda_i| - |\sigma_i^*|| \leq \epsilon$.

$\blacksquare$

**Lemma 11.** *Let $a = b + \epsilon. \overrightarrow{1}$ where $a, b$ are any 2 vectors and $\epsilon > 0$. Then, $\|a^{\otimes 3} - b^{\otimes 3}\|_\infty \leq \|a - b\|_\infty.\|b\|_\infty^2 + O(\epsilon^2)$.*

*Proof.* We have

$$\|a^{\otimes 3} - b^{\otimes 3}\|_\infty = \|(b + \epsilon \overrightarrow{1})^{\otimes 3} - b^{\otimes 3}\|_\infty$$

Let $(i, j, k)$ be the maximum element. Therefore,

$$\|(b + \epsilon \overrightarrow{1})^{\otimes 3} - b^{\otimes 3}\|_\infty = (b_i + \epsilon)(b_j + \epsilon)(b_k + \epsilon) - b_i b_j b_k$$
$$= \epsilon(b_i b_j + b_j b_k + b_k b_i) + \epsilon^2(b_i + b_j + b_k) + \epsilon^3$$

With $b_i \leq c \; \forall i$ for some $c > 0$ and $\epsilon = \|a - b\|_\infty$, we have $\|a^{\otimes 3} - b^{\otimes 3}\|_\infty \leq 3\epsilon c^2 + O(\epsilon^2)$ $\blacksquare$

# C Symmetric embedding of an asymmetric tensor

We use the symmetric embedding $sym(L)$ of a tensor $L$ as defined in Section 2.3 of [24]. We focus on third order tensors which have low CP-rank. We have three properties to derive that is relevant to us:

1. *Symmetry:* From Lemma 2.2 of [24] we see that $sym(L)$ for any tensor is symmetric.

2. *CP-Rank:* From Equation 6.5 of [24] we see that CP-rank($sym(L)$) $\leq 6$.CP-rank($L$). Since this is a constant, we see that the symmetric embedding is also a low-rank tensor.

3. *Incoherece:* Theorem 4.7 of [24] says that if $u_1$, $u_2$ and $u_3$ are unit modal singular vectors of $T$, then the vector $\tilde{u} = 3^{-1/2}[u_1; u_2; u_3]$ is a unit eigenvector of $sym(T)$. Without loss of generality, assume that $T$ is of size $n_1 \times n_2 \times n_3$ with $n_1 \leq n_2 \leq n_3$. In this case, we have

$$\|\tilde{u}\|_\infty \leq \frac{\mu}{(3n_1)^{1/2}} \tag{15}$$

and

$$\|\tilde{u}\|_\infty \leq \frac{\tilde{\mu}}{(n_1 + n_2 + n_3)^{1/2}} \tag{16}$$

for $\tilde{\mu} = c\mu$ for some constant $c$ to be calculated. Equating the right hand sides of Equations (15) and (16), we obtain $c = [(n_1 + n_2 + n_3)/(3n_1)]^{1/2}$. When $\Theta(n_1) = \Theta(n_2) = \Theta(n_3)$, we see that the eigenvectors $\tilde{u}$ of $sym(T)$ as specified above have the incoherence-preserving property.

# D  Proof of Theorem 2

Let $\widetilde{L}$ be a symmetric tensor which is a perturbed version of an orthogonal tensor $L^*$, $\widetilde{L} = L^* + E \in \mathbb{R}^{n \times n \times n}$, $L^* = \sum_{i \in [r]} \sigma_i^* u_i^{\otimes 3}$, where $\sigma_1^* \geq \sigma_2^* \ldots \sigma_r^* > 0$ and $\{u_1, u_2, \ldots, u_r\}$ form an orthonormal basis.

The analysis proceeds iteratively. First, we prove convergence to eigenpair of $\widetilde{L}$, which is close to top eigenpair $(\sigma_1^*, u_1)$ of $L^*$. We then argue that the same holds on the deflated tensor, when the perturbation $E$ satisfies (8). from This finishes the proof of Theorem 2.

To prove convergence for the first stage, i.e. convergence to eigenpair of $\widetilde{L}$, which is close to top eigenpair $(\sigma_1^*, u_1)$ of $L^*$, we analyze two phases of the shifted power iteration. In the first phase, we prove that with $N_1$ initializations and $N_2$ power iterations, we get close to true top eigenpair of $L^*$, i.e. $(\sigma_1^*, u_1)$. After this, in the second phase, we prove convergence to an eigenpair of $\widetilde{L}$.

The proof of the second phase is outlined in the main text. Here, we now provide proof for the first phase.

## D.1  Analysis of first phase of shifted power iteration

In this section, we prove that the output of shifted power method is close to original eigenpairs of the (unperturbed) orthogonal tensor, i.e. Theorem 2 holds, except for the property that the output corresponds to the eigenpairs of the perturbed tensor. We adapt the proof of tensor power iteration from [2] but here, since we consider the shifted power method, we need to modify it. We adopt the notation of [2] in this section.

Recall the update rule used in the shifted power method. Let $\theta_t = \sum_{i=1}^k \theta_{i,t} v_i \in \mathbb{R}^k$ be the unit vector at time $t$. Then

$$\theta_{t+1} = \sum_{i=1}^k \theta_{i,t+1} v_i := (\tilde{T}(I, \theta_t, \theta_t) + \alpha \theta_t)/\|(\tilde{T}(I, \theta_t, \theta_t) + \alpha \theta_t)\|.$$

In this subsection, we assume that $\tilde{T}$ has the form

$$\tilde{T} = \sum_{i=1}^k \tilde{\lambda}_i v_i^{\otimes 3} + \tilde{E} \tag{17}$$

where $\{v_1, v_2, \ldots, v_k\}$ is an orthonormal basis, and, without loss of generality,

$$\tilde{\lambda}_1 |\theta_{1,t}| = \max_{i \in [k]} \tilde{\lambda}_i |\theta_{i,t}| > 0.$$

Also, define

$$\tilde{\lambda}_{\min} := \min\{\tilde{\lambda}_i : i \in [k],\ \tilde{\lambda}_i > 0\}, \quad \tilde{\lambda}_{\max} := \max\{\tilde{\lambda}_i : i \in [k]\}.$$

We assume the error $\tilde{E}$ is a symmetric tensor such that, for some constant $p > 1$,

$$\|\tilde{E}(I, u, u)\| \leq \tilde{\epsilon}, \quad \forall u \in S^{k-1}; \tag{18}$$

$$\|\tilde{E}(I, u, u)\| \leq \tilde{\epsilon}/p, \quad \forall u \in S^{k-1} \text{ s.t. } (u^\top v_1)^2 \geq 1 - (3\tilde{\epsilon}/\tilde{\lambda}_1)^2. \tag{19}$$

In the next two propositions (Propositions D.1 and D.2) and Lemmas D.1, we analyze the power method iterations using $\tilde{T}$ at some arbitrary iterate $\theta_t$ using only the property (18) of $\tilde{E}$. But throughout, the quantity $\tilde{\epsilon}$ can be replaced by $\tilde{\epsilon}/p$ if $\theta_t$ satisfies $(\theta_t^\top v_1)^2 \geq 1 - (3\tilde{\epsilon}/\tilde{\lambda}_1)^2$ as per property (19).

Define

$$R_\tau := \left( \frac{\theta_{1,\tau}^2}{1 - \theta_{1,\tau}^2} \right)^{1/2}, \qquad r_{i,\tau} := \frac{\tilde{\lambda}_1 \theta_{1,\tau}}{\tilde{\lambda}_i |\theta_{i,\tau}|},$$

$$\gamma_\tau := 1 - \frac{1}{\min_{i \neq 1} |r_{i,\tau}|}, \qquad \delta_\tau := \frac{\tilde{\epsilon}}{\tilde{\lambda}_1 \theta_{1,\tau}^2}, \qquad \kappa := \frac{\tilde{\lambda}_{\max}}{\tilde{\lambda}_1} \tag{20}$$

for $\tau \in \{t, t+1\}$.

**Proposition D.1.**

$$\min_{i\neq 1}|r_{i,t}| \geq \frac{R_t}{\kappa}, \qquad\qquad \gamma_t \geq 1 - \frac{\kappa}{R_t}, \qquad\qquad \theta_{1,t}^2 = \frac{R_t^2}{1+R_t^2}.$$

**Proposition D.2.**

$$r_{i,t+1} \geq r_{i,t}^2 \cdot \frac{1 - \delta_t + \frac{\alpha}{\tilde{\lambda}_1 \theta_{1,t}}}{1 + \kappa \delta_t r_{i,t}^2 + \frac{\alpha}{\tilde{\lambda}_i \theta_{i,t}}}, \quad i \in [k], \tag{21}$$

$$R_{t+1} \geq = R_t \cdot \frac{1 - \delta_t + \frac{\alpha}{\tilde{\lambda}_1 |\theta_{1,t}|}}{1 - \gamma_t + \left(\delta_t + \frac{\alpha(1-\theta_{1,t})^{1/2}}{\tilde{\lambda}_1 \theta_{1,t}^2}\right) R_t} \geq \frac{1 - \delta_t + \frac{\alpha}{\tilde{\lambda}_1 |\theta_{1,t}|}}{\frac{\kappa}{R_t^2} + \delta_t + \frac{\alpha(1-\theta_{1,t})^{1/2}}{\tilde{\lambda}_1 \theta_{1,t}^2}}. \tag{22}$$

*Proof.* Let $\check{\theta}_{t+1} := \tilde{T}(I, \theta_t, \theta_t) + \alpha\theta_t$, so $\theta_{t+1} = \check{\theta}_{t+1}/\|\check{\theta}_{t+1}\|$. Since $\check{\theta}_{i,t+1} = \tilde{T}(v_i, \theta_t, \theta_t) = T(v_i, \theta_t, \theta_t) + \alpha\theta_t + E(v_i, \theta_t, \theta_t)$, we have

$$\check{\theta}_{i,t+1} = \tilde{\lambda}_i \theta_{i,t}^2 + E(v_i, \theta_t, \theta_t) + \alpha\theta_t^\top v_i, \quad i \in [k].$$

By definition, we have $\theta_{i,t} = \theta_t^\top v_i$. Using the triangle inequality and the fact $\|E(v_i, \theta_t, \theta_t)\| \leq \tilde{\epsilon}$, we have

$$\check{\theta}_{i,t+1} \geq \tilde{\lambda}_i \theta_{i,t}^2 - \tilde{\epsilon} + \alpha\theta_{i,t} \geq |\theta_{i,t}| \cdot \left(\tilde{\lambda}_i|\theta_{i,t}| - \tilde{\epsilon}/|\theta_{i,t}| + \alpha\right) \tag{23}$$

and

$$|\check{\theta}_{i,t+1}| \leq |\tilde{\lambda}_i \theta_{i,t}^2| + \tilde{\epsilon} + \alpha\theta_{i,t} \leq |\theta_{i,t}| \cdot \left(\tilde{\lambda}_i|\theta_{i,t}| + \tilde{\epsilon}/|\theta_{i,t}| + \alpha\right) \tag{24}$$

for all $i \in [k]$. Combining (23) and (24) gives

$$r_{i,t+1} = \frac{\tilde{\lambda}_1 \theta_{1,t+1}}{\tilde{\lambda}_i|\theta_{i,t+1}|} = \frac{\tilde{\lambda}_1 \check{\theta}_{1,t+1}}{\tilde{\lambda}_i|\check{\theta}_{i,t+1}|} \geq r_{i,t}^2 \cdot \frac{1 - \delta_t + \frac{\alpha}{\tilde{\lambda}_1\theta_{1,t}}}{1 + \frac{\tilde{\epsilon}}{\tilde{\lambda}_i\theta_{i,t}^2} + \frac{\alpha}{\tilde{\lambda}_i\theta_{i,t}}} = r_{i,t}^2 \cdot \frac{1 - \delta_t + \frac{\alpha}{\tilde{\lambda}_1\theta_{1,t}}}{1 + (\tilde{\lambda}_i/\tilde{\lambda}_1)\delta_t r_{i,t}^2 + \frac{\alpha}{\tilde{\lambda}_i\theta_{i,t}}} \geq r_{i,t}^2 \cdot \frac{1 - \delta_t + \frac{\alpha}{\tilde{\lambda}_1\theta_{1,t}}}{1 + \kappa\delta_t r_{i,t}^2 + \frac{\alpha}{\tilde{\lambda}_i\theta_{i,t}}}.$$

Moreover, by the triangle inequality and Hölder's inequality,

$$\left(\sum_{i=2}^n [\check{\theta}_{i,t+1}]^2\right)^{1/2} = \left(\sum_{i=2}^n \left(\tilde{\lambda}_i\theta_{i,t}^2 + E(v_i, \theta_t, \theta_t) + \alpha\theta_{i,t}\right)^2\right)^{1/2}$$

$$\leq \left(\sum_{i=2}^n \tilde{\lambda}_i^2\theta_{i,t}^4\right)^{1/2} + \left(\sum_{i=2}^n E(v_i, \theta_t, \theta_t)^2\right)^{1/2} + \left(\sum_{i=2}^k \alpha^2\theta_{i,t}^2\right)^{1/2}$$

$$\leq \max_{i\neq 1}\tilde{\lambda}_i|\theta_{i,t}|\left(\sum_{i=2}^n \theta_{i,t}^2\right)^{1/2} + \tilde{\epsilon} + \left(\alpha^2\sum_{i=2}^k \theta_{i,t}^2\right)^{1/2}$$

$$= (1 - \theta_{1,t}^2)^{1/2} \cdot \left(\max_{i\neq 1}\tilde{\lambda}_i|\theta_{i,t}| + \tilde{\epsilon}/(1-\theta_{1,t}^2)^{1/2} + \alpha\right). \tag{25}$$

Combining (23) and (25) gives

$$\frac{|\theta_{1,t+1}|}{(1-\theta_{1,t+1}^2)^{1/2}} = \frac{|\check{\theta}_{1,t+1}|}{\left(\sum_{i=2}^n [\check{\theta}_{i,t+1}]^2\right)^{1/2}} \geq \frac{|\theta_{1,t}|}{(1-\theta_{1,t}^2)^{1/2}} \cdot \frac{\tilde{\lambda}_1|\theta_{1,t}| - \tilde{\epsilon}/|\theta_{1,t}| + \alpha}{\max_{i\neq 1}\tilde{\lambda}_i|\theta_{i,t}| + \tilde{\epsilon}/(1-\theta_{1,t}^2)^{1/2} + \alpha}.$$

In terms of $R_{t+1}$, $R_t$, $\gamma_t$, and $\delta_t$, this reads

$$R_{t+1} \geq \frac{1 - \delta_t + \frac{\alpha}{\tilde{\lambda}_1|\theta_{1,t}|}}{(1-\gamma_t)\left(\frac{1-\theta_{1,t}^2}{\theta_{1,t}^2}\right)^{1/2} + \delta_t + \frac{\alpha(1-\theta_{1,t})^{1/2}}{\tilde{\lambda}_1\theta_{1,t}^2}} = R_t \cdot \frac{1 - \delta_t + \frac{\alpha}{\tilde{\lambda}_1|\theta_{1,t}|}}{1 - \gamma_t + \left(\delta_t + \frac{\alpha(1-\theta_{1,t}^2)^{1/2}}{\tilde{\lambda}_1\theta_{1,t}^2}\right) R_t}$$

$$= \frac{1 - \delta_t + \frac{\alpha}{\tilde{\lambda}_1|\theta_{1,t}|}}{\frac{1-\gamma_t}{R_t} + \left(\delta_t + \frac{\alpha(1-\theta_{1,t}^2)^{1/2}}{\tilde{\lambda}_1\theta_{1,t}^2}\right)} \geq \frac{1 - \delta_t + \frac{\alpha}{\tilde{\lambda}_1|\theta_{1,t}|}}{\frac{\kappa}{R_t^2} + \delta_t + \frac{\alpha(1-\theta_{1,t}^2)^{1/2}}{\tilde{\lambda}_1\theta_{1,t}^2}}$$

where the last inequality follows from Proposition D.1. ∎

**Lemma D.1.** *Fix any $\rho > 1$. Assume*

$$0 \le \delta_t < \min\left\{\frac{1}{2(1 + 2\kappa\rho^2)}, \ \frac{1 - 1/\rho}{1 + \kappa\rho}\right\}$$

*and $\gamma_t > 2(1 + 2\kappa\rho^2)\delta_t$.*

    *1. If $r_{i,t}^2 \le 2\rho^2$, then $r_{i,t+1} \ge |r_{i,t}|\left(1 + \frac{\gamma_t}{2}\right)$.*

*Proof.* By (21) from Proposition D.2,

$$r_{i,t+1} \ge r_{i,t}^2 \cdot \frac{1 - \delta_t + \frac{\alpha}{\bar{\lambda}_1 \theta_{1,t}}}{1 + \kappa\delta_t r_{i,t}^2 + \frac{\alpha}{\bar{\lambda}_i \theta_{i,t}}} \ge |r_{i,t}| \cdot \frac{1}{1 - \gamma_t} \cdot \frac{1 - \delta_t + \frac{\alpha}{\bar{\lambda}_1 \theta_{1,t}}}{1 + 2\kappa\rho^2\delta_t + \frac{\alpha}{\bar{\lambda}_i \theta_{i,t}}} \ge |r_{i,t}|\left(1 + \frac{\gamma_t}{2}\right)$$

where the last inequality is seen as follows: Let

$$\xi = 2 \cdot \frac{1 - \delta_t + \frac{\alpha}{\bar{\lambda}_1 \theta_{1,t}}}{1 + 2\kappa\rho^2\delta_t + \frac{\alpha}{\bar{\lambda}_i \theta_{i,t}}}$$

Then, we have $\gamma_t^2 + \gamma_t - 2 + \xi \ge 0$. The positive root is $\frac{-1 + (9 - 4\xi)^{1/2}}{2}$. Since $\gamma_t \ge 0$, we have $(9 - 4\xi)^{1/2} \ge 1$, so we assume $\xi \le 2$ for the inequality to hold, i.e., $\frac{\alpha}{\bar{\lambda}_1 \theta_{1,t}} - \frac{\alpha}{\bar{\lambda}_i \theta_{i,t}} \le (1 + 2\kappa\rho^2)\delta_t$. ∎

    The rest of the proof is along the similar lines of [2], except that we use SVD initialization instead of random initialization. The proof of SVD initialization is given in [3].