Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates

Anima Anandkumar* Rong Ge† Majid Janzamin‡

March 5, 2015

Abstract

In this paper, we provide local and global convergence guarantees for recovering CP (Candecomp/Parafac) tensor decomposition. The main step of the proposed algorithm is a simple alternating rank-1 update which is the alternating version of the tensor power iteration adapted for asymmetric tensors. Local convergence guarantees are established for third order tensors of rank k in d dimensions, when $k = o\big(d^{1.5}\big)$ and the tensor components are incoherent. Thus, we can recover overcomplete tensor decomposition. We also strengthen the results to global convergence guarantees under stricter rank condition $k \leq \beta d$ (for arbitrary constant $\beta > 1$) through a simple initialization procedure where the algorithm is initialized by top singular vectors of random tensor slices. Furthermore, the approximate local convergence guarantees for p-th order tensors are also provided under rank condition $k = o\big(d^{p/2}\big)$. The guarantees also include tight perturbation analysis given noisy tensor.

Keywords: Tensor decomposition, alternating minimization, overcomplete representation, latent variable models.

1 Introduction

Tensor decompositions have been recently popular for unsupervised learning of a wide range of latent variable models such as independent component analysis (De Lathauwer et al., 2007), topic models, Gaussian mixtures, hidden Markov models (Anandkumar et al., 2014a), network community models (Anandkumar et al., 2013a), and so on. The decomposition of a certain low order multivariate moment tensor (typically up to fourth order) in these models is guaranteed to provide a consistent estimate of the model parameters. Moreover, the sample and computational requirements are only a low order polynomial in the rank of the tensor (Anandkumar et al., 2014a; Song et al., 2013). In practice, the tensor decomposition techniques have been shown to be effective in a number of applications such as blind source separation (Comon, 2002), computer vision (Vasilescu and Terzopoulos, 2003), contrastive topic modeling (Zou et al., 2013), and community detection (Huang et al., 2013). In many cases, the tensor approach is shown to be orders of magnitude faster than existing techniques such as the stochastic variational approach.

*University of California, Irvine. Email: a.anandkumar@uci.edu
†Microsoft Research, New England. Email: rongge@microsoft.com
‡University of California, Irvine. Email: mjanzami@uci.edu

The state of art for guaranteed tensor decomposition involves two steps: converting the input tensor to an orthogonal symmetric form, and then solving the orthogonal decomposition through tensor eigen decomposition (Comon, 1994; Kolda and Mayo, 2011; Zhang and Golub, 2001; Anandkumar et al., 2014a). The first step of converting the input tensor to an orthogonal symmetric form is known as *whitening*. For the second step, the tensor eigen pairs can be found through a simple tensor *power* iteration procedure.

While having efficient guarantees, the above procedure suffers from a number of theoretical and practical limitations. For instance, in practice, the learning performance is especially sensitive to whitening (Le et al., 2011). Moreover, whitening is computationally the most expensive step in deployments (Huang et al., 2013), and it can suffer from numerical instability in high-dimensions due to ill-conditioning. Lastly, the above approach is unable to learn *overcomplete representations* (this is the case when number of features/components is much larger than the dimension) due to the orthogonality constraint, which is especially limiting, given the recent popularity of overcomplete feature learning in many domains (Bengio et al., 2012; Lewicki and Sejnowski, 2000).

The current practice for tensor decomposition is the *alternating least squares* (ALS) procedure, which has been described as the "workhorse" of tensor decomposition (Kolda and Bader, 2009). This involves solving the least squares problem on a *mode* of the tensor, while keeping the other modes fixed, and alternating between the tensor modes. The method is extremely fast since it involves calculating linear updates, but is not guaranteed to converge to the global optimum in general (Kolda and Bader, 2009).

In this paper, we provide local and global convergence guarantees for a modified alternating method, for which the main step is making rank-1 updates along different modes of the tensor. This update is basically a rank-1 ALS update. This method is extremely fast to deploy, trivially parallelizable, and does not suffer from ill-conditioning issues faced by both ALS (Kolda and Bader, 2009) and whitening approaches (Le et al., 2011). Our analysis assumes the presence of *incoherent* tensor components, which can be viewed as a *soft-orthogonality* constraint. Incoherent representations have been extensively considered in literature in a number of contexts, e.g., compressed sensing (Donoho, 2006) and sparse coding (Arora et al., 2013; Agarwal et al., 2013). Incoherent representations provide flexible modeling, can handle overcomplete signals, and are robust to noise (Lewicki and Sejnowski, 2000). Moreover, when the latent variable model parameters are *generic* or when we have randomly constructed (multiview) features (McWilliams et al., 2013), the moment tensors have incoherent components, as assumed here. In this work, we establish that incoherence leads to efficient guarantees for tensor decomposition. The guarantees also include a tight perturbation analysis. In a subsequent work (Anandkumar et al., 2014b), we apply the tensor decomposition guarantees of this paper to various learning settings, and derive sample complexity bounds through novel covering arguments.

1.1 Summary of results

In this paper, we propose and analyze an algorithm for non-orthogonal CP (Candecomp/Parafac) tensor decomposition; see Figure 1 for the details of the algorithm. The main step of the algorithm is a simple alternating rank-1 update which is the alternating version of the tensor power iteration adapted for asymmetric tensors. In each iteration, one of the tensor modes is updated by projecting the other modes along their estimated directions, and the process is alternated between all the modes of the tensor; see (5) for this update.

For the above update, we provide local convergence guarantees under incoherent tensor compo-

nents for a rank-k third order tensor in d dimensions. We prove a linear rate of convergence under appropriate initialization when $k = o(d^{3/2})$. Due to incoherence, the actual tensor components are not the stationary points of the update (even in the noiseless setting), and thus, there is an approximation error in the estimate after this update. The approximation error depends on the extent of overcompleteness, and scales as [1] $\tilde{O}(\sqrt{k}/d)$, which is small since $k = o(d^{3/2})$. The generalization to higher order tensors is also provided. To the best of our knowledge, we give the first guarantees for overcomplete tensor decomposition under mild incoherence conditions.

In order to remove the approximation error $\tilde{O}(\sqrt{k}/d)$ after the above rank-1 updates, we propose an additional update to the algorithm which is basically a type of coordinate descent update; see (9). We run this update after the main rank-1 updates and show that this removes the approximation error in a linear rate of convergence, and thus, we finally consistently recover the tensor decomposition.

In the undercomplete or mildly overcomplete settings ($k = O(d)$), a simple initialization procedure (see Procedure 2) based on rank-1 SVD of random tensor slices is provided. This initialization procedure lands the estimate in the basin of attraction for the alternating update procedure in polynomial number of trials (in the tensor rank k). This leads to global convergence guarantees for tensor decomposition.

We then extend the global convergence guarantees to settings where two modes of the tensor are (sufficiently) undercomplete (the dimension d_u is much larger than tensor rank k), and the third tensor mode is (highly) overcomplete (the dimension d_o is much smaller than tensor rank k). For instance, consider tensors arising from multi-view mixture models such as $\mathbb{E}[x_1 \otimes x_2 \otimes y]$, where x_i are multi-view high dimensional features and y is a low dimensional label. Previous procedures in (Anandkumar et al., 2014a) which rely on transforming the input tensor to an orthogonal symmetric form cannot handle this setting. Algorithms based on simultaneous diagonalization (Harshman and Lundy, 1994) can handle this case, but is not as robust to noise. We prove global convergence guarantees by considering rank-1 SVD of random tensor slices along the y-mode as initialization for the x_i-modes of the tensor, and then running the alternating update procedure.

Overview of techniques: Greedy or rank-1 updates are perhaps the most natural procedure for CP tensor decomposition. For orthogonal tensors, they lead to guaranteed recovery (Zhang and Golub, 2001). However, when the tensor is non-orthogonal, greedy procedure is not optimal in general (Kolda, 2001). Finding tensor decomposition in general is NP-hard (Hillar and Lim, 2009). We circumvent this obstacle by limiting ourselves to tensors with incoherent components. We exploit incoherence to prove error contraction under each step of the alternating update procedure with an approximation error, which is decaying, when $k = o(d^{1.5})$. To this end, we require tools from random matrix theory, bounds on $2 \to p$ norm for random matrices (Guédon and Rudelson, 2007; Adamczak et al., 2011) for some $p < 3$, and matrix perturbation results to provide tight bounds on error contraction.

1.2 Related work

CP tensor decomposition (Carroll and Chang, 1970), also known as PARAFAC decomposition (Harshman, 1970; Harshman and Lundy, 1994) is a classical definition for tensor decomposition with many applications. The most commonly used algorithm for CP decomposition is Alternating Least Squares

[1] $\tilde{O}$ is O up to polylog factors.

(ALS) (Comon et al., 2009), which has no convergence guarantees in general. Kolda (2001) and Zhang and Golub (2001) analyze the greedy or the rank-1 updates in the orthogonal setting. In the noisy setting, Anandkumar et al. (2014a) analyze deflation procedure for orthogonal decomposition, and Song et al. (2013) extend the analysis to the nonparametric setting. For the non-orthogonal tensors, a common strategy is to first apply a procedure called *whitening* to reduce it to the orthogonal case. But as discussed earlier, the whitening procedure can lead to poor performance and bad sample complexity. Moreover, it requires the tensor factors to have full column rank, which rules out overcomplete tensors.

Learning overcomplete tensors is challenging, and they may not even be identifiable in general. Kruskal (1976, 1977) provided an identifiability result based on the *Kruskal* rank of the factor matrices of the tensor. However, this result is limiting since it requires $k = O(d)$, where k is the tensor rank and d is the dimension. The FOOBI procedure by De Lathauwer et al. (2007) overcomes this limitation by assuming *generic* factors, and shows that a polynomial-time procedure can recover the tensor components when $k = O(d^2)$, and the tensor is fourth order. However, the procedure does not work for third-order overcomplete tensors, and has no polynomial sample complexity bounds. Simple procedures can recover overcomplete tensors for higher order tensors (five or higher). For instance, for the fifth order tensor, when $k = O(d^2)$, we can utilize random slices along a mode of the tensor, and perform simultaneous diagonalization on the matricized versions. Note that this procedure cannot handle the same level of overcompleteness as FOOBI, since an additional dimension is required for obtaining two (or more) fourth order tensor slices. The simultaneous diagonalization procedure entails careful perturbation analysis, carried out by (Goyal et al., 2013; Bhaskara et al., 2013). In addition, Goyal et al. (2013) provide stronger results for independent components analysis (ICA), where the tensor slices can be obtained in the Fourier domain.

There are other recent works which can learn overcomplete models, but under different settings than the ones considered in this paper. For instance, Arora et al. (2013); Agarwal et al. (2013) provide guarantees for the sparse coding problem. Anandkumar et al. (2013b) learn overcomplete sparse topic models, and provide guarantees for *Tucker* tensor decomposition under sparsity constraints. Specifically, the model is identifiable using $(2n)^{\text{th}}$ order moments when the latent dimension $k = O(d^n)$ and the sparsity level of the factor matrix is $O(d^{1/n})$, where d is the observed dimension. The Tucker decomposition is different from the CP decomposition considered here (it has weaker assumptions and guarantees), and the techniques in (Anandkumar et al., 2013b) differ significantly from the ones considered here.

The algorithm employed here falls under the general framework of alternating minimization. There are many recent works which provide guarantees on local/global convergence for alternating minimization, e.g., for matrix completion (Jain et al., 2013; Hardt, 2013), phase retrieval (Netrapalli et al., 2013) and sparse coding (Agarwal et al., 2013). However, the techniques in this paper are significantly different, since they involve tensors, while the previous works only required matrix analysis.

1.3 Notations and tensor preliminaries

Let $[n]$ denote the set $\{1, 2, \ldots, n\}$.

Notice that while the standard asymptotic notation is to write $f(d) = O(g(d))$ and $g(d) = \Omega(f(d))$, we sometimes use $f(d) \leq O(g(d))$ and $g(d) \geq \Omega(f(d))$ for additional clarity. We also use the asymptotic notation $f(d) = \tilde{O}(g(d))$ if and only if $f(d) \leq \alpha g(d)$ for all $d \geq d_0$, for some $d_0 > 0$ and $\alpha = \text{polylog}(d)$, i.e., $\tilde{O}$ hides polylog factors.

Tensor preliminaries

A real *p-th order tensor* $T \in \bigotimes_{i=1}^{p} \mathbb{R}^{d_i}$ is a member of the outer product of Euclidean spaces $\mathbb{R}^{d_i}$, $i \in [p]$. For convenience, we restrict to the case where $d_1 = d_2 = \cdots = d_p = d$, and simply write $T \in \bigotimes^{p} \mathbb{R}^{d}$. As is the case for vectors (where $p = 1$) and matrices (where $p = 2$), we may identify a p-th order tensor with the p-way array of real numbers $[T_{i_1,i_2,\ldots,i_p} : i_1, i_2, \ldots, i_p \in [d]]$, where $T_{i_1,i_2,\ldots,i_p}$ is the $(i_1, i_2, \ldots, i_p)$-th coordinate of T with respect to a canonical basis. For convenience, we limit to third order tensors ($p = 3$) in our analysis, while the results for higher order tensors are also provided.

The different dimensions of the tensor are referred to as *modes*. For instance, for a matrix, the first mode refers to columns and the second mode refers to rows. 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 (and is arranged as a column vector). For instance, for a matrix, its mode-1 fiber is any matrix column while a mode-2 fiber is any row. For a third order tensor $T \in \mathbb{R}^{d\times d\times d}$, the mode-1 fiber is given by $T(:, j, l)$, mode-2 by $T(i, :, l)$ and mode-3 by $T(i, j, :)$. 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 3rd mode are given by $T(:, :, l)$. For $r \in \{1, 2, 3\}$, the mode-r matricization of a third order tensor $T \in \mathbb{R}^{d\times d\times d}$, denoted by $\mathrm{mat}(T, r) \in \mathbb{R}^{d\times d^2}$, consists of all mode-$r$ fibers arranged as column vectors.

We view a tensor $T \in \mathbb{R}^{d\times d\times d}$ as a multilinear form. Consider matrices $M_r \in \mathbb{R}^{d\times d_r}, r \in \{1, 2, 3\}$. Then tensor $T(M_1, M_2, M_3) \in \mathbb{R}^{d_1} \otimes \mathbb{R}^{d_2} \otimes \mathbb{R}^{d_3}$ is defined as

$$T(M_1, M_2, M_3)_{i_1,i_2,i_3} := \sum_{j_1,j_2,j_3\in[d]} T_{j_1,j_2,j_3} \cdot M_1(j_1, i_1) \cdot M_2(j_2, i_2) \cdot M_3(j_3, i_3). \tag{1}$$

In particular, for vectors $u, v, w \in \mathbb{R}^d$, we have [2]

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

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, and $T(I, I, w) \in \mathbb{R}^{d\times d}$ is a linear combination of the tensor slices.

A 3rd order tensor $T \in \mathbb{R}^{d\times d\times d}$ is said to be rank-1 if it can be written in the form

$$T = w \cdot a \otimes b \otimes c \Leftrightarrow T(i, j, l) = w \cdot a(i) \cdot b(j) \cdot c(l), \tag{3}$$

where notation $\otimes$ represents the *outer product* and $a \in \mathbb{R}^d$, $b \in \mathbb{R}^d$, $c \in \mathbb{R}^d$ are unit vectors (without loss of generality). A tensor $T \in \mathbb{R}^{d\times d\times d}$ is said to have a CP rank $k \geq 1$ if it can be written as the sum of k rank-1 tensors

$$T = \sum_{i\in[k]} w_i a_i \otimes b_i \otimes c_i, \quad w_i \in \mathbb{R},\ a_i, b_i, c_i \in \mathbb{R}^d. \tag{4}$$

This decomposition is closely related to the multilinear form. In particular, for vectors $\hat{a}, \hat{b}, \hat{c} \in \mathbb{R}^d$, we have

$$T(\hat{a}, \hat{b}, \hat{c}) = \sum_{i\in[k]} w_i \langle a_i, \hat{a}\rangle \langle b_i, \hat{b}\rangle \langle c_i, \hat{c}\rangle.$$

[2] Compare with the matrix case where for $M \in \mathbb{R}^{d\times d}$, we have $M(I, u) = Mu := \sum_{j\in[d]} u_j M(:, j) \in \mathbb{R}^d$.

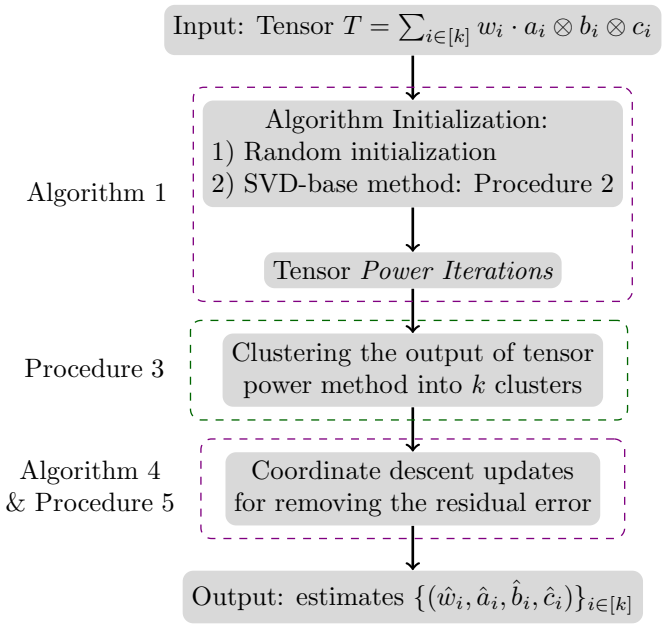

Figure 1: Overview of tensor decomposition algorithm.

Consider the decomposition in equation (4), denote matrix $A := [a_1\ a_2\ \cdots\ a_k] \in \mathbb{R}^{d\times k}$, and similarly B and C. Without loss of generality, we assume that the matrices have normalized columns (in 2-norm), since we can always rescale them, and adjust the weights w_i appropriately.

Throughout, $\|v\| := (\sum_i v_i^2)^{1/2}$ denotes the Euclidean (ℓ_2) norm of a vector v, and $\|M\|$ denotes the spectral (operator) norm of a matrix M. Furthermore, $\|T\|$ and $\|T\|_F$ denote the spectral (operator) norm and the Frobenius norm of a tensor, respectively. In particular, for a 3rd order tensor, we have

$$\|T\| := \sup_{\|u\|=\|v\|=\|w\|=1} |T(u,v,w)|, \quad \|T\|_F := \sqrt{\sum_{i,j,l\in[d]} T_{i,j,l}^2}.$$

2 Tensor Decomposition Algorithm

In this section, we introduce the alternating tensor decomposition algorithm, and the guarantees are provided in Section 3. The goal of tensor decomposition algorithm is to recover the rank-1 components of tensor; see (4) for the notion of tensor rank. Figure 1 depicts the overview of our tensor decomposition method where the corresponding algorithms and procedures are also specified. Our algorithm includes two main steps as 1) alternating tensor power iteration, and 2) coordinate descent iteration for removing the residual error. The former one is performed in Algorithm 1 (see equation (5), and the latter one is done in Algorithm 4 (see equation (9)). We now describe these steps of the algorithm in more details as well as providing the auxiliary procedures required to complete the algorithm.

2.1 Tensor power iteration in Algorithm 1

The main step of the algorithm is tensor power iteration which basically performs alternating *asymmetric power updates*[3] on different modes of the tensor as

$$\hat{a}^{(t+1)} = \frac{T\left(I, \hat{b}^{(t)}, \hat{c}^{(t)}\right)}{\left\|T\left(I, \hat{b}^{(t)}, \hat{c}^{(t)}\right)\right\|}, \quad \hat{b}^{(t+1)} = \frac{T\left(\hat{a}^{(t)}, I, \hat{c}^{(t)}\right)}{\left\|T\left(\hat{a}^{(t)}, I, \hat{c}^{(t)}\right)\right\|}, \quad \hat{c}^{(t+1)} = \frac{T\left(\hat{a}^{(t)}, \hat{b}^{(t)}, I\right)}{\left\|T\left(\hat{a}^{(t)}, \hat{b}^{(t)}, I\right)\right\|}, \tag{5}$$

where $\{\hat{a}^{(t)}, \hat{b}^{(t)}, \hat{c}^{(t)}\}$ denotes estimate in the t-th iteration. Recall that for vectors $v, w \in \mathbb{R}^d$, the multilinear form $T(I, v, w) \in \mathbb{R}^d$ used in the above update formula is defined in (2), where $T(I, v, w)$ is a multilinear combination of the tensor mode-1 fibers. Notice that the updates alternate among different modes of the tensor which can be viewed as a rank-1 form of the standard Alternating Least Squares (ALS) method. We later discuss this relation in more details.

Optimization viewpoint: Consider the problem of best rank-1 approximation of tensor T as

$$\min_{\substack{a,b,c \in \mathcal{S}^{d-1} \\ w \in \mathbb{R}}} \|T - w \cdot a \otimes b \otimes c\|_F, \tag{6}$$

where $\mathcal{S}^{d-1}$ denotes the unit d-dimensional sphere. This optimization program is non-convex, and has multiple local optima. It can be shown that the updates in (5) are the alternating optimization for this program where in each update, optimization over one vector is performed while the other two vectors are assumed fixed. This alternating minimization approach does not converge to the true components of tensor T in general, and in this paper we provide sufficient conditions for the convergence guarantees.

Intuition: We now provide an intuitive argument on the functionality of power updates in (5). Consider a rank-k tensor T as in (4), and suppose we start at the correct vectors $\hat{a} = a_j$ and $\hat{b} = b_j$, for some $j \in [k]$. Then for the numerator of update formula (5), we have

$$T\left(\hat{a}, \hat{b}, I\right) = T\left(a_j, b_j, I\right) = w_j c_j + \sum_{i \neq j} w_i \langle a_j, a_i \rangle \langle b_j, b_i \rangle c_i, \tag{8}$$

where the first term is along c_j and the second term is an error term due to non-orthogonality. For orthogonal decomposition, the second term is zero, and the true vectors a_j, b_j and c_j are stationary points for the power update procedure. However, since we consider non-orthogonal tensors, this procedure cannot recover the decomposition exactly leading to a residual error after running this step. Under incoherence conditions which encourages soft-orthogonality constraints[4] (and some other conditions), we show that the residual error is small (see Lemma 1 where the guarantees for the tensor power iteration step is provided), and thus, with the additional step we propose in Section 2.2, we can also remove this residual error.

[3] This is exactly the generalization of asymmetric matrix power update to 3rd order tensors.

[4] See Assumption (A2) in Appendix A for precise description.

Algorithm 1 Tensor decomposition via alternating asymmetric power updates

Input: Tensor $T \in \mathbb{R}^{d\times d\times d}$, number of initializations L, number of iterations N.

1: **for** $\tau = 1$ **to** L **do**

2: **Initialize** unit vectors $\hat{a}_\tau^{(0)} \in \mathbb{R}^d$, $\hat{b}_\tau^{(0)} \in \mathbb{R}^d$, and $\hat{c}_\tau^{(0)} \in \mathbb{R}^d$ as

- Option 1: SVD-based method in Procedure 2 when $k \le \beta d$ for arbitrary constant β.
- Option 2: random initialization.

3: **for** $t = 0$ **to** $N - 1$ **do**

4: Asymmetric power updates (see (2) for the definition of the multilinear form):

$$\hat{a}_\tau^{(t+1)} = \frac{T\left(I, \hat{b}_\tau^{(t)}, \hat{c}_\tau^{(t)}\right)}{\left\|T\left(I, \hat{b}_\tau^{(t)}, \hat{c}_\tau^{(t)}\right)\right\|}, \quad \hat{b}_\tau^{(t+1)} = \frac{T\left(\hat{a}_\tau^{(t)}, I, \hat{c}_\tau^{(t)}\right)}{\left\|T\left(\hat{a}_\tau^{(t)}, I, \hat{c}_\tau^{(t)}\right)\right\|}, \quad \hat{c}_\tau^{(t+1)} = \frac{T\left(\hat{a}_\tau^{(t)}, \hat{b}_\tau^{(t)}, I\right)}{\left\|T\left(\hat{a}_\tau^{(t)}, \hat{b}_\tau^{(t)}, I\right)\right\|}.$$

5: **end for**

6: weight estimation:

$$\hat{w}_\tau = T\left(\hat{a}_\tau^{(N)}, \hat{b}_\tau^{(N)}, \hat{c}_\tau^{(N)}\right). \tag{7}$$

7: **end for**

8: Cluster set $\left\{\left(\hat{w}_\tau, \hat{a}_\tau^{(N)}, \hat{b}_\tau^{(N)}, \hat{c}_\tau^{(N)}\right), \tau \in [L]\right\}$ into k clusters as in Procedure 3.

9: **return** the center member of these k clusters as estimates $(\hat{w}_j, \hat{a}_j, \hat{b}_j, \hat{c}_j), j \in [k]$.

Initialization and clustering procedures: We discussed that the tensor power updates in (5) are the alternating iterations for the problem of rank-1 approximation of the tensor; see (6). This is a non-convex problem and has many local optima. Thus, the power update requires careful initialization to ensure convergence to the true rank-1 tensor components.

For generating initialization vectors $\left(\hat{a}^{(0)}, \hat{b}^{(0)}, \hat{c}^{(0)}\right)$, we introduce two possibilities. One is the simple random initializations, where $\hat{a}^{(0)}$ and $\hat{b}^{(0)}$ are uniformly drawn from unit sphere $\mathcal{S}^{d-1}$. The other option is SVD-based technique in Procedure 2 where top left and right singular vectors of $T(I, I, \theta)$ (for some random $\theta \in \mathbb{R}^d$) are respectively introduced as $\hat{a}^{(0)}$ and $\hat{b}^{(0)}$. Under both initialization procedures, vector $\hat{c}^{(0)}$ is generated through update formula in (5). We establish in Section 3.2 that when $k = O(d)$, the SVD procedure leads to global convergence guarantees under polynomial number of trials. In practice random initialization also works well, however the analysis is still an open problem.

Notice that the algorithm is run for L different initialization vectors for which we do not know the good ones in prior. In order to identify which initializations are successful at the end, we also need a *clustering* step proposed in Procedure 3 to obtain the final estimates of the vectors. The detailed analysis of clustering procedure is provided in Appendix D.

2.2 Coordinate descent iteration in Algorithm 4

We discussed in the previous section that the tensor power iteration recovers the tensor rank-1 components up to some residual error. We now propose Algorithm 4 to remove this additional

Procedure 2 SVD-based initialization when $k \leq \beta d$ for arbitrary constant β

Input: Tensor $T \in \mathbb{R}^{d \times d \times d}$.

1: Draw a random standard Gaussian vector $\theta \sim \mathcal{N}(0, I_d)$.
2: Compute u_1 and v_1 as the top left and right singular vectors of $T(I, I, \theta) \in \mathbb{R}^{d \times d}$.
3: $\hat{a}^{(0)} \leftarrow u_1$, $\hat{b}^{(0)} \leftarrow v_1$.
4: Initialize $\hat{c}^{(0)}$ by update formula in (5).
5: **return** $\left(\hat{a}^{(0)}, \hat{b}^{(0)}, \hat{c}^{(0)}\right)$.

Procedure 3 Clustering process

Input: Tensor $T \in \mathbb{R}^{d \times d \times d}$, set of 4-tuples $\left\{(\hat{w}_\tau, \hat{a}_\tau, \hat{b}_\tau, \hat{c}_\tau), \tau \in [L]\right\}$, parameter ν.

1: **for** $i = 1$ **to** k **do**
2: Among the remaining 4-tuples, choose $\hat{a}, \hat{b}, \hat{c}$ which correspond to the largest $|T(\hat{a}, \hat{b}, \hat{c})|$.
3: Do N more iterations of alternating updates in (5) starting from $\hat{a}, \hat{b}, \hat{c}$.
4: Let the output of iterations denoted by $(\hat{a}, \hat{b}, \hat{c})$ be the center of cluster i.
5: Remove all the tuples with $\max\{|\langle \hat{a}_\tau, \hat{a}\rangle|, |\langle \hat{b}_\tau, \hat{b}\rangle|, |\langle \hat{c}_\tau, \hat{c}\rangle|\} > \nu/2$.
6: **end for**
7: **return** the k cluster centers.

residual error. This algorithm mainly runs a coordinate descent iteration as

$$\tilde{c}_i^{(t+1)} = \mathrm{Norm}\left(T\left(\widehat{a}_i^{(t)}, \widehat{b}_i^{(t)}, I\right) - \sum_{j \neq i} \widehat{w}_j^{(t)} \langle \widehat{a}_i^{(t)}, \widehat{a}_j^{(t)}\rangle \langle \widehat{b}_i^{(t)}, \widehat{b}_j^{(t)}\rangle \cdot \widehat{c}_j^{(t)}\right), \quad i \in [k], \tag{9}$$

where for vector v, we have $\mathrm{Norm}(v) := v/\|v\|$, i.e., it normalizes the vector. The above is similarly applied for updating $\tilde{a}_i^{(t+1)}$ and $\tilde{b}_i^{(t+1)}$. Unlike the power iteration, it can be immediately seen that a_i, b_i and c_i are stationary points of the above update even if the components are not orthogonal to each other. Inspired by this intuition, we prove that when the residual error is small enough (as guaranteed in the analysis of tensor power iteration), this step removes it.

The analysis of this algorithm requires that the estimate matrices $\hat{A}, \hat{B}, \hat{C}$ satisfy some bound on the spectral norm and some column-wise error bounds; see Definition 2 in Appendix B.2 for the details. The optimization program in (10) (which is only run in the first iteration) and projection Procedure 5 ensure that these conditions are satisfied.

2.3 Discussions

We now provide some further discussions and comparisons about the algorithm.

Implicit tensor operations: In many applications, the input tensor T is not available in advance, and it is computed from samples. It is discussed in (Anandkumar et al., 2014b) that the tensor is not needed to be computed and stored explicitly, where the multilinear tensor updates (5) and (9) in the algorithm can be efficiently computed through multilinear operations on the samples directly.

Algorithm 4 Coordinate descent algorithm for removing the residual error

Input: Tensor $T \in \mathbb{R}^{d\times d\times d}$, initialization set $\left\{\widehat{A}, \widehat{B}, \widehat{C}, \widehat{w}^{(0)}\right\}$, number of iterations N.

Initialize $\widehat{A}^{(0)}$ as (similarly for $\widehat{B}^{(0)}, \widehat{C}^{(0)}$)

$$\widehat{A}^{(0)} := \arg\min_{\tilde{A}} \|\tilde{A}\| \quad \text{s. t. } \|\tilde{a}_i - \widehat{a}_i\| \leq \tilde{O}\left(\sqrt{k}/d\right), i \in [k]. \tag{10}$$

for $t = 0$ **to** $N-1$ **do**

  for $i = 1$ **to** k **do**

$$\tilde{w}_i^{(t+1)} = \left\| T\left(\widehat{a}_i^{(t)}, \widehat{b}_i^{(t)}, I\right) - \sum_{j\neq i} \widehat{w}_j^{(t)} \langle \widehat{a}_i^{(t)}, \widehat{a}_j^{(t)}\rangle \langle \widehat{b}_i^{(t)}, \widehat{b}_j^{(t)}\rangle \cdot \widehat{c}_j^{(t)} \right\|,$$

$$\tilde{c}_i^{(t+1)} = \frac{1}{\tilde{w}_i^{(t+1)}} \left(T\left(\widehat{a}_i^{(t)}, \widehat{b}_i^{(t)}, I\right) - \sum_{j\neq i} \widehat{w}_j^{(t)} \langle \widehat{a}_i^{(t)}, \widehat{a}_j^{(t)}\rangle \langle \widehat{b}_i^{(t)}, \widehat{b}_j^{(t)}\rangle \cdot \widehat{c}_j^{(t)} \right).$$

  end for

  Update $\widehat{C}^{(t+1)}$ by applying Procedure 5 with inputs $\tilde{C}^{(t+1)}$ and $\widehat{C}^{(t)}$.

  Repeat the above steps (with appropriate changes) to update $\widehat{A}^{(t+1)}$ and $\widehat{B}^{(t+1)}$.

  Update $\widehat{w}^{(t+1)}$:

$$\text{for any } i \in [k],\ \widehat{w}_i^{(t+1)} = \begin{cases} \tilde{w}_i^{(t+1)}, & \left|\tilde{w}_i^{(t+1)} - \widehat{w}_i^{(t)}\right| \leq \eta_0 \frac{\sqrt{k}}{d}, \\ \widehat{w}_i^{(t)} + \operatorname{sgn}\left(\tilde{w}_i^{(t+1)} - \widehat{w}_i^{(t)}\right) \cdot \eta_0 \frac{\sqrt{k}}{d}, & \text{o. w.} \end{cases}$$

end for

return $\left\{\widehat{A}^{(N)}, \widehat{B}^{(N)}, \widehat{C}^{(N)}, \widehat{w}^{(N)}\right\}$.

Comparison with symmetric orthogonal tensor power method: Algorithm 1 is similar to the symmetric tensor power method analyzed by Anandkumar et al. (2014a) with the following main differences, viz.,

- Symmetric and non-symmetric tensors: Our algorithm can be applied to both symmetric and non-symmetric tensors, while tensor power method in Anandkumar et al. (2014a) is only for symmetric tensors.
- Linearity: The updates in Algorithm 1 are linear in each variable, while the symmetric tensor power update is a quadratic operator given a third order tensor.
- Guarantees: In Anandkumar et al. (2014a), guarantees for the symmetric tensor power update under orthogonality are obtained, while here we consider non-orthogonal tensors under the alternating updates.

Comparison with Alternating Least Square(ALS): The updates in Algorithm 1 can be viewed as a rank-1 form of the standard alternating least squares (ALS) procedure. This is because the unnormalized update for c in (5) can be rewritten as

$$\tilde{c}_\tau^{(t+1)} := T\left(\hat{a}_\tau^{(t)}, \hat{b}_\tau^{(t)}, I\right) = \operatorname{mat}(T, 3) \cdot \left(\hat{b}_\tau^{(t)} \odot \hat{a}_\tau^{(t)}\right), \tag{11}$$

Procedure 5 Projection procedure

input Matrices $\tilde{C}^{(t+1)}$, $\widehat{C}^{(t)}$.

1: Compute the SVD of $\tilde{C}^{(t+1)} = UDV^\top$.

2: Let $\widehat{D}$ be the truncated version of D as $\widehat{D}_{i,i} := \min\left\{D_{i,i}, \eta_1\sqrt{\frac{k}{d}}\right\}$.

3: Let $Q := U\widehat{D}V^\top$.

4: Update $\widehat{C}^{(t+1)}$: for any $i \in [k]$, $\widehat{c}_i^{(t+1)} = \begin{cases} Q_i, & \left\|Q_i - \widehat{c}_i^{(t)}\right\| \le \eta_0\frac{\sqrt{k}}{d}, \\ \widehat{c}_i^{(t)} + \eta_0\frac{\sqrt{k}}{d}\frac{\left(Q_i - \widehat{c}_i^{(t)}\right)}{\left\|Q_i - \widehat{c}_i^{(t)}\right\|}, & \text{o. w.} \end{cases}$

5: **return** $\widehat{C}^{(t+1)}$.

where $\odot$ denotes the *Khatri-Rao* product, and $\mathrm{mat}(T, 3) \in \mathbb{R}^{d\times d^2}$ is the mode-3 matricization of tensor T. On the other hand, the ALS update has the form

$$\tilde{C}^{(t+1)} = \mathrm{mat}(T, 3) \cdot \left(\left(\hat{B}^{(t)} \odot \hat{A}^{(t)}\right)^\top\right)^\dagger,$$

where k vectors (all columns of $\tilde{C}^{(t+1)} \in \mathbb{R}^{d\times k}$) are simultaneously updated given the current estimates for the other two modes $\hat{A}^{(t)}$ and $\hat{B}^{(t)}$. In contrast, our procedure updates only one vector (with the target of recovering one column of C) in each iteration. In our update, we do not require finding matrix inverses. This leads to efficient computational complexity, and we also show that our update procedure is more robust to perturbations.

3 Analysis

In this section, we provide the local and global convergence guarantees for the tensor decomposition algorithm proposed in Section 2. Throughout the paper, we assume tensor $\hat{T} \in \mathbb{R}^{d\times d\times d}$ is of the form $\hat{T} = T + \Psi$, where Ψ is the error or perturbation tensor, and[5]

$$T = \sum_{i\in[k]} w_i \cdot a_i \otimes b_i \otimes c_i,$$

is a rank-k tensor such that $a_i, b_i, c_i \in \mathbb{R}^d, i \in [k]$, are unit vectors. Let $A := [a_1\ a_2\ \cdots\ a_k] \in \mathbb{R}^{d\times k}$, and B and C are similarly defined. The goal of robust tensor decomposition algorithm is to recover the rank-1 components $\{(a_i, b_i, c_i), i \in [k]\}$ given noisy tensor $\hat{T}$. Our analysis emphasizes on the challenging *overcomplete* regime where the tensor rank is larger than the dimension, i.e., $k > d$. Without loss of generality we also assume $w_{\max} = w_1 \ge w_2 \ge \cdots \ge w_k = w_{\min} > 0$.

We require natural deterministic conditions on the tensor components to argue the convergence guarantees; see Appendix A for the details. We show that all of these conditions are satisfied if the true rank-1 components of the tensor are uniformly i.i.d. drawn from the unit d-dimensional sphere $\mathcal{S}^{d-1}$. Thus, for simplicity we assume this random assumption in the main part, and state the

[5]For 4th and higher order tensors, same techniques we introduce in this paper, can be exploited to argue similar results.

deterministic assumptions in Appendix A. Notice that it is also reasonable to assume these deterministic assumptions hold for some non-random matrices. Among the deterministic assumptions, the most important one is the *incoherence* condition which imposes a soft-orthogonality constraint between different rank-1 components of the tensor.

The convergence guarantees are provided in terms of distance between the estimated and the true vectors, defined below.

Definition 1. *For any two vectors $u, v \in \mathbb{R}^d$, the distance between them is defined as*

$$\mathrm{dist}(u, v) := \sup_{z \perp u} \frac{\langle z, v \rangle}{\|z\| \cdot \|v\|} = \sup_{z \perp v} \frac{\langle z, u \rangle}{\|z\| \cdot \|u\|}. \tag{12}$$

Note that distance function $\mathrm{dist}(u, v)$ is invariant w.r.t. norm of input vectors u and v. Distance also provides an upper bound on the error between unit vectors u and v as (see Lemma A.1 of Agarwal et al. (2013))

$$\min_{z \in \{-1,1\}} \|zu - v\| \leq \sqrt{2}\, \mathrm{dist}(u, v).$$

Incorporating distance notion resolves the sign ambiguity issue in recovering the components: note that a third order tensor is unchanged if the sign of a vector along one of the modes is fixed and the signs of the corresponding vectors in the other two modes are flipped.

3.1 Local convergence guarantee

In the local convergence guarantee, we analyze the convergence properties of the algorithm assuming we have good initialization vectors for the non-convex tensor decomposition algorithm.

Settings of Algorithm in Theorem 1:

- Number of iterations: $N = \Theta\left(\log\left(\frac{1}{\gamma \epsilon_R}\right)\right)$, where $\gamma := \frac{w_{\max}}{w_{\min}}$ and $\epsilon_R := \min\left\{\frac{\psi}{w_{\min}}, \tilde{O}\left(\gamma \frac{\sqrt{k}}{d}\right)\right\}$.

Conditions for Theorem 1:

- Rank-k true tensor with random components: Let

$$T = \sum_{i \in [k]} w_i \cdot a_i \otimes b_i \otimes c_i, \quad w_i > 0, a_i, b_i, c_i \in \mathcal{S}^{d-1},$$

  where $a_i, b_i, c_i, i \in [k]$, are uniformly i.i.d. drawn from the unit d-dimensional sphere $\mathcal{S}^{d-1}$. We state the deterministic assumptions in Appendix A, and show that random matrices satisfy these assumptions.
- Rank condition: $k = o\left(d^{1.5}\right)$.
- Perturbation tensor Ψ satisfies the bound

$$\psi := \|\Psi\| \leq \frac{w_{\min}}{6}.$$

- Weight ratio: The maximum ratio of weights $\gamma := \frac{w_{\max}}{w_{\min}}$ satisfies the bound

$$\gamma = O\left(\min\left\{\sqrt{d}, \frac{d^{1.5}}{k}\right\}\right).$$

- Initialization: Assume we have good initialization vectors $\hat{a}_j^{(0)}, \hat{b}_j^{(0)}, j \in [k]$ satisfying

$$\epsilon_0 := \max\left\{\text{dist}\left(\hat{a}_j^{(0)}, a_j\right), \text{dist}\left(\hat{b}_j^{(0)}, b_j\right)\right\} = O(1/\gamma), \quad \forall j \in [k], \tag{13}$$

where $\gamma := \frac{w_{\max}}{w_{\min}}$. In addition, given $\hat{a}_j^{(0)}$ and $\hat{b}_j^{(0)}$, suppose $\hat{c}_j^{(0)}$ is also calculated by the update formula in (5).

Theorem 1 (Local convergence guarantee of the tensor decomposition algorithm)**.** *Consider noisy rank-k tensor $\hat{T} = T + \Psi$ as the input to the tensor decomposition algorithm, and assume the conditions and settings mentioned above hold. Then the algorithm outputs estimates $\hat{A} := [\hat{a}_1 \cdots \hat{a}_k] \in \mathbb{R}^{d\times k}$ and $\hat{w} := [\hat{w}_1 \cdots \hat{w}_k]^\top \in \mathbb{R}^k$, satisfying w.h.p.*

$$\left\|\widehat{A} - A\right\|_F \leq \tilde{O}\left(\frac{\sqrt{k}\cdot\psi}{w_{\min}}\right), \quad \|\hat{w} - w\| \leq \tilde{O}\left(\sqrt{k}\cdot\psi\right).$$

Same error bounds hold for other factor matrices $B := [b_1 \cdots b_k]$ and $C := [c_1 \cdots c_k]$.

See the proof in Appendix B.

Thus, we can efficiently decompose the tensor in the highly overcomplete regime $k \leq o\left(d^{1.5}\right)$ under incoherent factors and some other assumptions mentioned above. The deterministic version of assumptions are stated in Appendix A. We show that these assumptions are true for random components which is assumed here for simplicity. If k is significantly smaller than $d^{1.5}$ ($k \ll d^{1.25}$), then many of the assumptions can be derived from incoherence. See Appendix A for the details.

The above local convergence result can be also interpreted as a local identifiability result for tensor decomposition under incoherent factors.

The $\sqrt{k}$ factor in the above theorem error bound is from the fact that the final recovery guarantee is on the Frobenius norm of the whole factor matrix A. In the following, we provide stronger column-wise guarantees (where there is no $\sqrt{k}$ factor) with the expense of having an additional residual error term. Recall that our algorithm includes two main update steps including tensor power iteration in (5) and residual error removal in (9). The guarantee for the first step — tensor power iteration — is provided in the following lemma.

Lemma 1 (Local convergence guarantee of the tensor power updates, Algorithm 1)**.** *Consider the same settings as in Theorem 1. Then, the outputs of tensor power iteration steps (output of Algorithm 1) satisfy w.h.p.*

$$\text{dist}(\widehat{a}_j, a_j) \leq \tilde{O}\left(\frac{\psi}{w_{\min}}\right) + \tilde{O}\left(\gamma\frac{\sqrt{k}}{d}\right), \quad |\hat{w}_j - w_j| \leq \tilde{O}\left(\psi\right) + \tilde{O}\left(w_{\max}\frac{\sqrt{k}}{d}\right), \quad j \in [k].$$

Same error bounds hold for other factor matrices B and C.

The above result provides guarantees with the additional residual error $\tilde{O}\left(\gamma\frac{\sqrt{k}}{d}\right)$, but we believe this result also has independent importance for the following reasons. The above result provides column-wise guarantees which is stronger than the guarantees on the whole factor matrix in Theorem 1. Furthermore, we can only have recovery guarantees for a subset of rank-1 components of the tensor (the ones for which we have good initializations) without worrying about the rest of components. Finally, in the high-dimensional regime (large d), the residual error term goes to zero.

The result in the above lemma is actually stated in the non-asymptotic form, where the details of constants are explicitly provided in Appendix A.

Symmetric tensor decomposition: The above local convergence result also holds for recovering the components of a rank-k *symmetric* tensor. Consider symmetric tensor T with CP decomposition $T = \sum_{i\in[k]} w_i a_i \otimes a_i \otimes a_i$. The proposed algorithm can be also applied to recover the components $a_i, i \in [k]$, where the main updates are changed to adapt to the symmetric tensor. The tensor power iteration is changed to

$$\hat{a}^{(t+1)} = \frac{T\left(\hat{a}^{(t)}, \hat{a}^{(t)}, I\right)}{\left\|T\left(\hat{a}^{(t)}, \hat{a}^{(t)}, I\right)\right\|}, \tag{14}$$

and the coordinate descent update is changed to the form stated in (27). Then, the same local convergence result as in Theorem 1 holds for this algorithm. The proof is very similar to the proof of Theorem 1 with some slight modifications considering the symmetric structure.

Extension to higher order tensors: We also provide the generalization of the tensor decomposition guarantees to higher order tensors. We state and prove the result for the tensor power iteration part in details, while the generalization of coordinate descent part (for removing the residual error) to higher order tensors, can be argued by the same techniques we introduce in this paper

For brevity, Algorithm 1 and local convergence guarantee in Lemma 1 are provided for a 3rd order tensor. The algorithm can be simply extended to higher order tensors to compute the corresponding CP decomposition. Consider p-th order tensor $T \in \bigotimes^p \mathbb{R}^d$ with CP decomposition

$$T = \sum_{i\in[k]} w_i \cdot a_{(1),i} \otimes a_{(2),i} \otimes \cdots \otimes a_{(p),i}, \tag{15}$$

where $a_{(r),i} \in \mathbb{R}^d$ is the i-th column of r-th component $A_{(r)} := \left[a_{(r),1}\ a_{(r),2}\ \cdots\ a_{(r),k}\right] \in \mathbb{R}^{d\times k}$, for $r \in [p]$. Algorithm 1 can be extended to recover the components of above decomposition where update formula for the p-th mode is modified as

$$\hat{a}_{(p)}^{(t+1)} = \frac{T\left(\hat{a}_{(1)}^{(t)}, \hat{a}_{(2)}^{(t)}, \ldots, \hat{a}_{(p-1)}^{(t)}, I\right)}{\left\|T\left(\hat{a}_{(1)}^{(t)}, \hat{a}_{(2)}^{(t)}, \ldots, \hat{a}_{(p-1)}^{(t)}, I\right)\right\|}, \tag{16}$$

and similarly the other updates are changed. Then, we have the following generalization of Lemma 1 to higher order tensors.

Corollary 1 (Local convergence guarantee of the tensor power updates in Algorithm 1 for p-th order tensor)**.** *Consider the same conditions and settings as in Lemma 1, unless tensor T is p-th order with CP decomposition in* (15) *where $p \geq 3$ is a constant. In addition, the bounds on $\gamma := \frac{w_{\max}}{w_{\min}}$ and k are modified as*

$$\gamma = O\left(\min\left\{d^{\frac{p-2}{2}}, \frac{d^{p/2}}{k}\right\}\right), \quad k = o\left(d^{\frac{p}{2}}\right).$$

Then, the outputs of tensor power iteration steps (output of Algorithm 1) satisfy w.h.p.

$$\operatorname{dist}\left(\hat{a}_{(r),j}, a_{(r),j}\right) \leq \tilde{O}\left(\frac{\psi}{w_{\min}}\right) + \tilde{O}\left(\gamma\sqrt{\frac{k}{d^{p-1}}}\right), \quad |\hat{w}_j - w_j| \leq \tilde{O}\left(\psi\right) + \tilde{O}\left(w_{\max}\sqrt{\frac{k}{d^{p-1}}}\right),$$

for $j \in [k]$ and $r \in [p]$. The number of iterations is $N = \Theta\left(\log\left(\frac{1}{\gamma\tilde{\epsilon}_R}\right)\right)$, where $\gamma := \frac{w_{\max}}{w_{\min}}$ and $\tilde{\epsilon}_R := \min\left\{\frac{\psi}{w_{\min}}, \tilde{O}\left(\gamma\sqrt{k/d^{p-1}}\right)\right\}$.

3.2 Global convergence guarantee when $k = O(d)$

Theorem 1 provides local convergence guarantee given good initialization vectors. In this section, we exploit SVD-based initialization method in Procedure 2 to provide good initialization vectors when $k = O(d)$. This method proposes the top singular vectors of random slices of the moment tensor as the initialization. Combining the theoretical guarantees of this initialization method (provided in Appendix C) with the local convergence guarantee in Theorem 1, we provide the following global convergence result.

Settings of Algorithm in Theorem 2:

- Number of iterations: $N = \Theta\left(\log\left(\frac{1}{\gamma\epsilon_R}\right)\right)$, where $\gamma := \frac{w_{\max}}{w_{\min}}$ and $\epsilon_R := \min\left\{\frac{\psi}{w_{\min}}, \tilde{O}\left(\gamma\frac{\sqrt{k}}{d}\right)\right\}$.
- The initialization in each run of Algorithm 1 is performed by SVD-based technique proposed in Procedure 2, with the number of initializations as

$$L \geq k^{\Omega\left(\gamma^4 (k/d)^2\right)}.$$

Conditions for Theorem 2:

- Rank-k decomposition and perturbation conditions as [6]

$$T = \sum_{i\in[k]} w_i \cdot a_i \otimes b_i \otimes c_i, \quad \psi := \|\Psi\| \leq \frac{w_{\min}\sqrt{\log k}}{\alpha_0\sqrt{d}},$$

  where $a_i, b_i, c_i, i \in [k]$, are uniformly i.i.d. drawn from the unit d-dimensional sphere $\mathcal{S}^{d-1}$, and $\alpha_0 > 1$ is a constant.
- Rank condition: $k = O(d)$, i.e., $k \leq \beta d$ for arbitrary constant $\beta > 1$.

Theorem 2 (Global convergence guarantee of tensor decomposition algorithm when $k = O(d)$)**.** *Consider noisy rank-k tensor $\hat{T} = T + \Psi$ as the input to the tensor decomposition algorithm, and assume the conditions and settings mentioned above hold. Then, the same guarantees as in Theorem 1 hold.*

See the proof in Appendix B.

Thus, we can efficiently recover the tensor decomposition, when the tensor is undercomplete or mildly overcomplete (i.e., $k \leq \beta d$ for arbitrary constant $\beta > 1$), by initializing the algorithm with a simple SVD-based technique. The number of initialization trials L is polynomial when γ is a constant, and $k = O(d)$.

Note that the argument in Lemma 1 can be similarly adapted leading to global convergence guarantee of the tensor power iteration step.

[6] Note that the perturbation condition is stricter than the corresponding condition in the local convergence guarantee (Theorem 1).

Two undercomplete, and one overcomplete component

Here, we apply the global convergence result to the regime of two undercomplete and one overcomplete components. This arises in supervised learning problems under a multiview mixtures model and employing moment tensor $\mathbb{E}[x_1 \otimes x_2 \otimes y]$, where $x_i \in \mathbb{R}^{d_u}$ are multi-view high-dimensional features and $y \in \mathbb{R}^{d_o}$ is a low-dimensional label.

Since in the SVD initialization Procedure 2, two components $\hat{a}^{(0)}$ and $\hat{b}^{(0)}$ are initialized through SVD, and the third component $\hat{c}^{(0)}$ is initialized through update formula (5), we can generalize the global convergence result in Theorem 2 to the setting where A, B are undercomplete, and C is overcomplete.

Corollary 2. *Consider the same setting as in Theorem 2. In addition, suppose the regime of undercomplete components $A \in \mathbb{R}^{d_u \times k}$, $B \in \mathbb{R}^{d_u \times k}$, and overcomplete component $C \in \mathbb{R}^{d_o \times k}$ such that $d_u \geq k \geq d_o$. In addition, in this case the bound on $\gamma := \frac{w_{\max}}{w_{\min}}$ is*

$$\gamma = O\left(\min\left\{\sqrt{d_o}, \frac{d_u\sqrt{d_o}}{k}\right\}\right).$$

Then, if $k = O(d_u)$ and $d_o \geq \text{polylog}(k)$, the same convergence guarantee as in Theorem 2 holds.

See the proof in Appendix B.

We observe that given undercomplete modes A and B, mode C can be arbitrarily overcomplete, and we can still provide global recovery of A, B and C by employing SVD initialization procedure along modes A and B.

3.3 Proof outline

The global convergence guarantee in Theorem 2 is established by combining the local convergence result in Theorem 1 and the SVD initialization result in Appendix C.

The local convergence result in Theorem 1 is derived by establishing error contraction in each iteration of the tensor power iteration and the coordinate descent for removing the residual error. Note that these convergence properties are broken down in Lemmata 1 and 12, respectively.

Since we assume generic factor matrices A, B and C, we utilize many useful properties such as incoherence, bounded spectral norm of the matrices A, B and C, bounded tensor spectral norm and so on. We list the precise set of deterministic conditions required to establish the local convergence result in Appendix A. Under these conditions, with a good initialization (i.e., small enough $\max\{\text{dist}(\hat{a}, a_j), \text{dist}(\hat{b}, b_j)\} \leq \epsilon_0$), we show that the iterative update in (5) provides an estimate $\hat{c}$ with

$$\text{dist}(\hat{c}, c_j) < \tilde{O}\left(\frac{\psi}{w_{\min}}\right) + \tilde{O}\left(\gamma\frac{\sqrt{k}}{d}\right) + q\epsilon_0,$$

for some contraction factor $q < 1/2$. The incoherence condition is crucial for establishing this result. See Appendix B for the complete proof.

The initialization argument for SVD-based technique in Procedure 2 has two parts. The first part claims that by performing enough number of initializations (large enough L), a gap condition is satisfied, meaning that we obtain a vector θ which is relatively close to c_j compared to any $c_i, i \neq j$. This is a standard result for Gaussian vectors, e.g., see Lemma B.1 of Anandkumar et al. (2014a). In the second part of the argument, we analyze the dominant singular vectors of $T(I, I, \theta)$, for a vector

θ with a good relative gap, to obtain an error bound on the initialization vectors. This is obtained through standard matrix perturbation results (Weyl and Wedin's theorems). See Appendix C for the complete proof.

4 Experiments

In this section, we provide some synthetic experiments to evaluate the performance of Algorithm 1. Note that tensor power update in Algorithm 1 is the main step of our algorithm which is considered in this experiment. A random true tensor T is generated as follows. First, three components $A \in \mathbb{R}^{d\times k}$, $B \in \mathbb{R}^{d\times k}$, and $C \in \mathbb{R}^{d\times k}$ are randomly generated with i.i.d standard Gaussian entries. Then, the columns of these matrices are normalized where the normalization factors are aggregated as coefficients $w_j, j \in [k]$. From decomposition form in (4), tensor T is built through these random components. For each new initialization, $\hat{a}^{(0)}$ and $\hat{b}^{(0)}$ are randomly generated with i.i.d. standard Gaussian entries, and then normalized [7]. Initialization vector $\hat{c}^{(0)}$ is generated through update formula in (5).

For each initialization $\tau \in [L]$, an alternative option of running the algorithm with a fixed number of iterations N is to stop the iterations based on some stopping criteria. In this experiment, we stop the iterations when the improvement in subsequent steps is small as

$$\max\left(\left\|\hat{a}_\tau^{(t)} - \hat{a}_\tau^{(t-1)}\right\|^2, \left\|\hat{b}_\tau^{(t)} - \hat{b}_\tau^{(t-1)}\right\|^2, \left\|\hat{c}_\tau^{(t)} - \hat{c}_\tau^{(t-1)}\right\|^2\right) \le t_{\mathrm{S}},$$

where t_{S} is the stopping threshold. According to the bound in Theorem 1, we set

$$t_{\mathrm{S}} := t_1(\log d)^2\frac{\sqrt{k}}{d}, \tag{17}$$

for some constant $t_1 > 0$.

Effect of size d and k

Algorithm 1 is applied to random tensors with $d = 1000$ and $k = \{10, 50, 100, 200, 500, 1000, 2000\}$. The number of initializations is $L = 2000$. The parameter t_1 in (17) is fixed as $t_1 = 1e-08$. Figure 2 and Table 1 illustrate the outputs of running experiments which is the average of 10 random runs.

Figure 2 depicts the ratio of recovered columns versus the number of initializations. Both horizontal and vertical axes are plotted in log-scale. We observe that it is much easier to recover the columns in the undercomplete settings ($k \le d$), while it becomes harder when k increases. Linear start in Figure 2 suggests that recovering the first bunch of columns only needs polynomial number of initializations. For highly undercomplete settings like $d = 1000$ and $k = 10$, almost all columns are recovered in this linear phase. After this start, the concave part means that it needs many more initializations for recovering the next bunch of columns. As we go ahead, it becomes harder to recover true columns, which is intuitive.

Table 1 has the results from the experiments. Parameters k, stopping threshold t_{S}, and the average square error of the output, the average weight error and the average number of iterations

[7] Drawing i.i.d. standard Gaussian entries and normalizing them is equivalent to drawing vectors uniformly from the d-dimensional unit sphere.

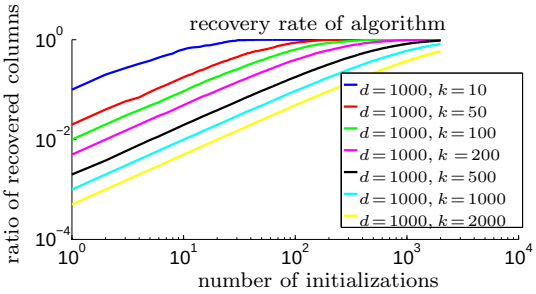

Figure 2: Ratio of recovered columns versus the number of initializations for $d = 1000$, and $k = \{10, 50, 100, 200, 500, 1000, 2000\}$. The number of initializations is $L = 2000$. The stopping parameter is set to $t_1 = 1e-08$. The figure is an average over 10 random runs.

Table 1: Parameters and more outputs related to results of Figure 2. Note that $d = 1000$.

Parameters		Outputs		
k	t_S	avg. square error	avg. weight error	avg. # of iterations
10	1.51e-08	1.03e-05	9.75e-09	7.71
50	3.37e-08	5.54e-05	6.69e-08	8.53
100	4.77e-08	1.08e-04	1.51e-07	8.81
200	6.75e-08	2.07e-04	3.41e-07	9.09
500	1.07e-07	5.09e-04	1.14e-06	9.52
1000	1.51e-07	1.01e-03	3.40e-06	10.01
2000	2.13e-07	2.00e-03	1.12e-05	10.69

are stated. The output averages are over several initializations and random runs. The square error is given by

$$\frac{1}{3}\left[\|a_j - \hat{a}\|^2 + \left\|b_j - \hat{b}\right\|^2 + \|c_j - \hat{c}\|^2\right],$$

for the corresponding recovered j. The error in estimating the weights is defined as $|\hat{w} - w_j|^2/w_j^2$ which is the square relative error of weight estimate. The number of iterations performed before stopping the algorithm is mentioned in the last column. We observe that by increasing k, all of these outputs are increased which means we get less accurate estimates with higher computation. This shows that recovering the overcomplete components is much harder. Note that by running the coordinate descent Algorithm 4, we can also remove this additional residual error left after the tensor power iteration step. Similar results and observations as above are seen when k is fixed and d is changed.

Running experiments with SVD initialization instead of random initialization yields nearly the same recovery rates, but with slightly smaller number of iterations. But, since the SVD computation is more expensive, in practice, it is desirable to initialize with random vectors. Our theoretical results for random initialization appear to be highly pessimistic compared to the efficient recovery results in our experiments. This suggests additional room for improving our theoretical guarantees under random initialization.

Acknowledgements

We acknowledge detailed discussions with Sham Kakade and Boaz Barak. We thank Praneeth Netrapalli for discussions on alternating minimization. We also thank Sham Kakade, Boaz Barak, Jonathan Kelner, Gregory Valiant and Daniel Hsu for earlier discussions on the $2 \to p$ norm bound for random matrices, used in Lemma 3. We also thank Niranjan U.N. for discussions on running experiments. A. Anandkumar is supported in part by Microsoft Faculty Fellowship, NSF Career award CCF-1254106, NSF Award CCF-1219234, and ARO YIP Award W911NF-13-1-0084. M. Janzamin is supported by NSF Award CCF-1219234, ARO Award W911NF-12-1-0404 and ARO YIP Award W911NF-13-1-0084.

Appendix

More Matrix Notations

Given vector $w \in \mathbb{R}^d$, let $\mathrm{Diag}(w) \in \mathbb{R}^{d \times d}$ denote the diagonal matrix with w on its main diagonal. Given matrix $A \in \mathbb{R}^{d \times k}$, the following notations are defined to refer to its sub-matrices. A_j denotes the j-th column and A^j denotes the j-th row of A. In addition, $A_{\setminus j} \in \mathbb{R}^{d \times (k-1)}$ is A with its j-th column removed, and $A^{\setminus j} \in \mathbb{R}^{(d-1) \times k}$ is A with its j-th row removed.

For two matrices $A \in \mathbb{R}^{d_1 \times k}$ and $B \in \mathbb{R}^{d_2 \times k}$, the *Khatri-Rao* product is denoted by $A \odot B \in \mathbb{R}^{d_1 d_2 \times k}$, and its $(\mathbf{i}, j)^{\text{th}}$ entry is given by

$$A \odot B(\mathbf{i}, j) := A_{i_1, j} B_{i_2, j}, \quad \mathbf{i} = (i_1, i_2) \in [d_1] \times [d_2], j \in [k].$$

For two matrices $A \in \mathbb{R}^{d \times k}$ and $B \in \mathbb{R}^{d \times k}$, the *Hadamard* product is defined as the entry-wise multiplication of the matrices,

$$A * B(i, j) := A(i, j) B(i, j), \quad i \in [d], j \in [k].$$

Let $\|u\|_p$ denote the ℓ_p norm of vector u. Let $\|A\|_\infty$ denote the ℓ_∞ element-wise norm of matrix A, and the induced $q \to p$ norm is defined as

$$\|A\|_{q \to p} := \sup_{\|u\|_q = 1} \|Au\|_p.$$

A Deterministic Assumptions

In the main text, we assume matrices A, B, and C are randomly generated. However, we are not using all the properties of randomness. In particular, we only need the following assumptions.

(A1) **Rank-k decomposition:** The third order tensor T has a CP rank of $k \geq 1$ with decomposition

$$T = \sum_{i \in [k]} w_i (a_i \otimes b_i \otimes c_i), \quad w_i > 0, a_i, b_i, c_i \in \mathcal{S}^{d-1}, \forall\, i \in [k], \tag{18}$$

where $\mathcal{S}^{d-1}$ denotes the unit d-dimensional sphere, i.e. all the vectors have unit[8] 2-norm as $\|a_i\| = \|b_i\| = \|c_i\| = 1, i \in [k]$. Furthermore, define $w_{\min} := \min_{i \in [k]} w_i$ and $w_{\max} := \max_{i \in [k]} w_i$.

[8] This normalization is for convenience and the results hold for general case.

(A2) **Incoherence:** The components are incoherent, and let

$$\rho := \max_{i \neq j} \{|\langle a_i, a_j \rangle|, |\langle b_i, b_j \rangle|, |\langle c_i, c_j \rangle|\} \leq \frac{\alpha}{\sqrt{d}}, \tag{19}$$

for some $\alpha = \text{polylog}(d)$. In other words, $A^\top A = I + J_A$, $B^\top B = I + J_B$, and $C^\top C = I + J_C$, where J_A, J_B, and J_C, are incoherence matrices with zero diagonal entries. We have $\max\{\|J_A\|_\infty, \|J_B\|_\infty, \|J_C\|_\infty\} \leq \rho$ as in (19).

(A3) **Spectral norm conditions:** The components satisfy spectral norm bound

$$\max\{\|A\|, \|B\|, \|C\|\} \leq 1 + \alpha_0 \sqrt{\frac{k}{d}},$$

for some constant $\alpha_0 > 0$.

(A4) **Bounds on tensor norms:** Tensor T satisfies the bound

$$\|T\| \leq w_{\max} \alpha_0,$$
$$\left\|T_{\setminus j}(a_j, b_j, I)\right\| := \left\| \sum_{i \neq j} w_i \langle a_i, a_j \rangle \langle b_i, b_j \rangle c_j \right\| \leq \alpha w_{\max} \frac{\sqrt{k}}{d},$$

for some constant α_0 and $\alpha = \text{polylog}(d)$.

(A5) **Rank constraint:** The rank of the tensor is bounded by $k = o\left(d^{1.5}/\operatorname{polylog} d\right)$.

(A6) **Bounded perturbation:** Let ψ denote the spectral norm of perturbation tensor as

$$\psi := \|\Psi\|. \tag{20}$$

Suppose ψ is bounded as [9]

$$\psi \leq \min\left\{\frac{1}{6}, \frac{\sqrt{\log k}}{\alpha_0 \sqrt{d}}\right\} \cdot w_{\min},$$

where α_0 is a constant.

(A7) **Weights ratio:** The maximum ratio of weights $\gamma := \frac{w_{\max}}{w_{\min}}$ satisfies the bound

$$\gamma = O\left(\min\left\{\sqrt{d}, \frac{d^{1.5}}{k}\right\}\right).$$

(A8) **Contraction factor:** The contraction factor q in Theorem 1 is defined as

$$q := \frac{2 w_{\max}}{w_{\min}} \left[\frac{2\alpha}{\sqrt{d}} \left(1 + \alpha_0 \sqrt{\frac{k}{d}}\right)^2 + \beta'\right], \tag{21}$$

for some constants $\alpha_0, \beta' > 0$, and $\alpha = \text{polylog}(d)$. In particular, we need $\alpha\alpha_0\sqrt{k}/d + \beta' < w_{\max}/10 w_{\min}$ which ensures $q < 1/2$. This is satisfied when $\sqrt{k}/d < w_{\max}/w_{\min} \operatorname{poly}\log d$ and $\beta' < w_{\max}/20 w_{\min}$. The parameter β' is determined by the following assumption (initialization).

[9] Note that for the local convergence guarantee, only the first condition $\psi \leq \frac{w_{\min}}{6}$ is required.

(A9) **Initialization:** Let

$$\epsilon_0 := \max\left\{\text{dist}\left(\hat{a}^{(0)}, a_j\right), \text{dist}\left(\hat{b}^{(0)}, b_j\right)\right\},$$

denote the initialization error w.r.t. to some $j \in [k]$. Suppose it is bounded as

$$\epsilon_0 \leq \min\left\{\frac{\beta'}{\alpha_0}, \sqrt{\frac{w_{\min}}{6w_{\max}}}, \frac{w_{\min}q}{4w_{\max}}, \frac{2w_{\max}}{w_{\min}q}\left(\frac{w_{\min}}{6w_{\max}} - \alpha\frac{\sqrt{k}}{d}\right)\right\},$$

for some constants $\alpha_0, \beta' > 0$, $\alpha = \text{polylog}(d)$, and $0 < q < 1/2$ which is defined in (21).

(A10) $2 \to p$ **norm:** For some fixed constant $p < 3$, $\max\{\|A^\top\|_{2\to p}, \|B^\top\|_{2\to p}, \|C^\top\|_{2\to p}\} \leq 1 + o(1)$.

Remark 1. *Many of the assumptions are actually parameter choices. The only properties of random matrices required are (A2), (A3), (A4) and (A10),. See Appendix A.1 for detailed discussion.*

Let us provide a brief discussion about the above assumptions. Condition (A1) requires the presence of a rank-k decomposition for tensor T. We normalize the component vectors for convenience, and this removes the scaling indeterminacy issues which can lead to problems in convergence. Additionally, we impose incoherence constraint in (A2), which allows us to provide convergence guarantee in the overcomplete setting. Assumptions (A3) and (A4) impose bounds on the spectral norm of tensor T and its decomposition components. Note that assumptions (A2)-(A4) and (A10) are satisfied w.h.p. when the columns of A, B, and C are generically drawn from unit sphere $\mathcal{S}^{d-1}$ (see Lemma 2 and Guédon and Rudelson (2007)), all others are parameter choices. Assumption (A5) limits the overcompleteness of problem which is required for providing convergence guarantees. The first bound on perturbation in (A6) as $\psi \leq \frac{w_{\min}}{6}$ is required for local convergence guarantee and the second bound $\psi \leq \frac{w_{\min}\sqrt{\log k}}{\alpha_0\sqrt{d}}$ is needed for arguing initialization provided by Procedure 2. Assumption (A7) is required to ensure contraction happens in each iteration. Assumption (A8) defines contraction ratio q in each iteration, and Assumption (A9) is the initialization condition required for local convergence guarantee.

The tensor-spectral norm and $2 \to p$ norm assumptions (A4) and (A10) may seem strong as we cannot even verify them given the matrix. However, when $k < d^{1.25-\epsilon}$ for arbitrary constant $\epsilon > 0$, both conditions are implied by incoherence. See Lemma 4. We only need these assumptions to go to the very overcomplete setting.

A.1 Random matrices satisfy the deterministic assumptions

Here, we provide arguments that random matrices satisfy conditions (A2), (A3), (A4), and (A10). It is well known that random matrices are incoherent, and have small spectral norm (bound on spectral norm dates back to Wigner (1955)). See the following lemma.

Lemma 2. *Consider random matrix $X \in \mathbb{R}^{d\times k}$ where its columns are uniformly drawn at random from unit d-dimensional sphere $\mathcal{S}^{d-1}$. Then, it satisfies the following incoherence and spectral bounds with high probability as*

$$\max_{i,j\in[k], i\neq j} |\langle X_i, X_j\rangle| \leq \frac{\alpha}{\sqrt{d}},$$

$$\|X\| \leq 1 + \alpha_0\sqrt{\frac{k}{d}},$$

for some $\alpha = O(\sqrt{\log k})$ *and* $\alpha_0 = O(1)$.

The spectral norm of the tensor is less well-understood. However, it can be bounded by the $2 \to 3$ norm of matrices. Using tools from Guédon and Rudelson (2007); Adamczak et al. (2011), we have the following result.

Lemma 3. *Consider a random matrix* $A \in \mathbb{R}^{d\times k}$ *whose columns are drawn uniformly at random from unit sphere. If* $k < d^{p/2}/\operatorname{polylog}(d)$*, then*

$$\left\|A^\top\right\|_{2\to p} \le 1 + o(1).$$

This directly implies Assumption (A10). In particular, since we only apply Assumption (A10) to unsupervised setting ($k \le O(d)$) in Appendix D, for randomly generated tensor, Assumption (A10) holds for all $p > 2$ (notice that we only need it to hold for some $p < 3$).

We also give an alternative proof of $2 \to p$ norm which does not assume randomness and only relies on incoherence.

Lemma 4. *Suppose columns of matrix* $A \in \mathbb{R}^{d\times k}$ *have unit norm and satisfy the incoherence condition (A2). If* $k \le d^{1.25-\epsilon}$ *for arbitrary constant* $\epsilon > 0$*, then for any* $p > 3 - 2\epsilon$*, we have*

$$\left\|A^\top\right\|_{2\to p} \le 1 + o(1).$$

Proof: Let $L = \sqrt{d}/\operatorname{poly}\log d$. By incoherence assumption we know every subset of L columns in A has singular values within $1 \pm o(1)$ (by Gershgorin Disk Theorem).

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

$$\begin{aligned}
\|A^\top u\|_p^p &= \sum_{i\in S} |A_i^\top u|^p + \sum_{i\notin S} |A_i^\top u|^p \\
&\le \sum_{i\in S} |A_i^\top u|^2 + (2/\sqrt{L})^{p-2} \sum_{i\notin S} |A_i^\top u|^2 \\
&\le 1 + o(1).
\end{aligned}$$

Here the second inequality uses that every entry outside S is small, and last inequality uses the fact that $p > 3 - 2\epsilon$. □

The $2 \to 3$ norm implies a bound on the tensor spectral norm by Hölder's inequality.

Fact 1 (Hölder's Inequality)**.** *When* $1/p + 1/q = 1$*, for two sequence of numbers* $\{a_i\}, \{b_i\}$*, we have*

$$\sum_i a_i b_i \le \left(\sum_i |a_i|^p\right)^{1/p} \left(\sum_i |b_i|^q\right)^{1/q}.$$

Consequently, we have the following corollary.

Corollary 3. *For vectors f, g, h, and weights $w_i \geq 0$, we have*

$$\sum_i w_i f_i g_i h_i \leq w_{\max} \|f\|_3 \|g\|_3 \|h\|_3.$$

Proof: The proof applies Hölder's inequality twice as

$$\sum_i w_i f_i g_i h_i \leq w_{\max} \sum_i |f_i g_i h_i| \leq w_{\max} (\sum |f_i|^3)^{1/3} (\sum |g_i h_i|^{3/2})^{2/3} \leq w_{\max} \|f\|_3 \|g\|_3 \|h\|_3,$$

where in the first application, $p = 3$ and $q = 3/2$, and in the second application, $p = q = 2$ (which is the special case known as Cauchy-Schwartz). □

In the following lemma, it is shown that the first bound in Assumption (A4) holds for random matrices w.h.p.

Lemma 5. *Let A, B, and C be random matrices in $\mathbb{R}^{d \times k}$ whose columns are drawn uniformly at random from unit sphere. If $k < d^{3/2}/\operatorname{polylog}(d)$, and*

$$T = \sum_{i \in [k]} w_i a_i \otimes b_i \otimes c_i,$$

then

$$\|T\| \leq O(w_{\max}).$$

Proof: For any unit vectors $\hat{a}, \hat{b}, \hat{c}$, we have

$$\begin{aligned} T(\hat{a}, \hat{b}, \hat{c}) &= \sum_{i \in [k]} w_i (A^\top \hat{a})_i (B^\top \hat{b})_i (C^\top \hat{c})_i \\ &\leq w_{\max} \|A^\top \hat{a}\|_3 \|B^\top \hat{b}\|_3 \|C^\top \hat{c}\|_3 \\ &\leq w_{\max} \|A^\top\|_{2 \to 3} \|\hat{a}\| \cdot \|B^\top\|_{2 \to 3} \|\hat{b}\| \cdot \|C^\top\|_{2 \to 3} \|\hat{c}\| \\ &= O(w_{\max}), \end{aligned}$$

where Corollary 3 is exploited in the first inequality, and Lemma 3 is used in the last inequality. □

For the case with two undercomplete and one overcomplete dimensions (see Corollary 2), we can prove the tensor spectral norm using basic properties of the matrices A, B, C.

Lemma 6. *Let $A, B \in \mathbb{R}^{d_u \times k}$ be matrices with spectral norm bounded by $O(1)$, and $C \in \mathbb{R}^{d_o \times k}$ be a matrix whose columns have unit norm. Let*

$$T = \sum_{i=1}^{k} w_i a_i \otimes b_i \otimes c_i,$$

then we have

$$\|T\| \leq O(w_{\max}).$$

Proof: For any unit vectors $u, v \in \mathbb{R}^{d_u}$ and $w \in \mathbb{R}^{d_o}$, by assumptions we know $\|A^\top u\| \leq O(1), \|B^\top v\| \leq O(1)$ and $\|C^\top w\|_\infty \leq 1$. Now we have

$$\begin{aligned} T(u,v,w) &= \sum_{i=1}^{k} w_i \langle a_i, u\rangle \langle b_i, v\rangle \langle c_i, w\rangle \\ &\leq w_{\max} \sum_{i=1}^{k} |\langle a_i, u\rangle \langle b_i, v\rangle| \\ &\leq w_{\max} \|A^\top u\| \|B^\top v\| \\ &= O(w_{\max}). \end{aligned}$$

The first inequality uses triangle inequality and the fact that $|\langle c_i, w\rangle| \leq 1$. The Cauchy-Schwartz inequality is exploited in the second inequality. Therefore, the spectral norm of the tensor is bounded by $O(w_{\max})$. □

Finally, we show in the following lemma that the second bound in Assumption (A4) is satisfied for random matrices.

Lemma 7. *Let $A, B, C \in \mathbb{R}^{d\times k}$ be independent, normalized (column) Gaussian matrices. Then for all $i \in [k]$, we have with high probability*

$$\left\| C_{\setminus i} \operatorname{Diag}(w^{\setminus i}) (J_A * J_B)_i^{\setminus i} \right\| = \tilde{O}\left(w_{\max} \frac{\sqrt{k}}{d} \right).$$

Proof: We have

$$C_{\setminus i} \operatorname{Diag}(w^{\setminus i}) (J_A * J_B)_i^{\setminus i} = \sum_{j\neq i} C_j w_j \langle A_i, A_j\rangle \langle B_i, B_j\rangle = \sum_{j\neq i} C_j \delta_j,$$

where $\delta_j := w_j \langle A_i, A_j\rangle \langle B_i, B_j\rangle$ is independent of C_j. From Lemma 2, columns of A and B are incoherent, and therefore, for $j \neq i$, we have

$$|\delta_j| = \tilde{O}(w_{\max}/d).$$

Now since C_j's are independent, zero mean vectors, the sum $\sum_{j\neq i} \delta_j C_j$ is zero mean and its variance is bounded by $\tilde{O}(w_{\max}^2 k/d^2)$. Then, from vector Bernstein's bound we have with high probability

$$\left\| C_{\setminus i} \operatorname{Diag}(w^{\setminus i}) (J_A * J_B)_i^{\setminus i} \right\| = \tilde{O}\left(w_{\max} \frac{\sqrt{k}}{d} \right).$$

The proof is completed by applying union bound. □

Spectral norm of Khatri-Rao product

For the convergence guarantees of the second step of algorithm on removing residual error, we need the following additional bound on the spectral norm of Khatri-Rao product of random matrices.

(A11) **Spectral Norm Condition on Khatri-Rao Products:** The components satisfy the following spectral norm bound on the Khatri-Rao products as

$$\max\{\|A \odot B\|, \|B \odot C\|, \|A \odot C\|\} \leq 1 + \alpha_0 \frac{\sqrt{k}}{d},$$

for $\alpha_0 \leq \text{poly}\log d$.

We now prove that Assumption (A11) is satisfied with high probability, if the columns of A, B and C are uniformly i.i.d. drawn from unit d-dimensional sphere.

The key idea is to view $(A \odot B)^\top (A \odot B)$ as the sum of random matrices, and use the following Matrix Bernstein's inequality to prove concentration results.

Lemma 8. *Let $M = \sum_{i=1}^n M_i$ be sum of independent symmetric $d \times d$ matrices with $\mathbb{E}[M_i] = 0$, assume all matrices M_i's have spectral norm at most R almost surely, let $\sigma^2 = \|\mathbb{E}[M_i^2]\|$, then for any τ*

$$\Pr[\|M\| \geq \tau] \leq 2d \exp\left(\frac{-\tau^2/2}{\sigma^2 + R\tau/3}\right).$$

Remark: Although the lemma requires all M_i's to have spectral norm at most R almost surely, it suffices to have spectral norm bounded by R with high probability and bounded by $R^\infty = \text{poly}(d, k)$ almost surely. This is because we can always condition on the fact that $\|M_i\| \leq R$ for all i. Such conditioning can only change the expectations by a negligible amount, and does not affect independence between M_i's.

Random unit vectors are not easy to work with, as entries in the same column are not independent. Thus, we first prove the result for matrices A and B whose entries are independent Gaussian variables.

Lemma 9. *Suppose A, $B \in \mathbb{R}^{d \times k} (k > \text{polylog}\, d)$ are independent random matrices with independent Gaussian entries, let $M = (A \odot B)^\top (A \odot B) = (A^\top A) * (B^\top B)$, then with high probability*

$$\|M - \text{Diag}(M)\| \leq O(d\sqrt{k \log d})$$

Proof: Let $a_1, a_2, ..., a_d \in \mathbb{R}^k$ be the columns of $A^\top$ (the rows of A, but treated as column vectors). We can rewrite $M - \text{Diag}\, M$ as

$$M - \text{Diag}\, M = (\sum_{i \in [d]} a_i a_i^\top) * (B^\top B - \text{Diag}(B^\top B)) = \sum_{i \in [d]} (a_i a_i^\top) * (B^\top B - \text{Diag}(B^\top B)).$$

Now let $Q = B^\top B - \text{Diag}(B^\top B)$, and $M_i = (a_i a_i^\top) * Q$, we would like to bound the spectral norm of the sum $M = \sum_{i \in [d]} M_i$. Clearly these entries are independent, $\mathbb{E}[M_i] = \mathbb{E}[a_i a_i^\top] * Q = I * Q = 0$, so we can apply Matrix Bernstein bound.

Note that when $d < k$, by standard random matrix theory we know $\|Q\| \leq O(k)$. Also, every row of Q has norm smaller than the corresponding row of $B^\top B$, which is bounded by $\|B\| \|b_{(i)}\| \leq O(\sqrt{kd})$. When $d \geq k$, again by matrix concentration we know $\|Q\| \leq O(\sqrt{dk \log d})$. Every row of Q has norm bounded by $O(\sqrt{kd})$ (because entries in a row are independently random, with variance equal to d).

First let us bound the spectral norm for each of the M_i's. Notice that for any vector v, $v^\top[(a_i a_i^\top) * Q]v = (v * a_i)^\top Q(v * a_i)$ by definition of Hadamard product. On the other hand, $\|v * a_i\| \le \|v\|\|a_i\|_\infty$. With high probability $\|a_i\|_\infty \le O(\sqrt{\log k})$, hence $\|M_i\| \le \|a_i\|_\infty^2 \|Q\|$. This is bounded by $O(k \log d)$ when $d < k$ and $O(\sqrt{kd}\log^2 d)$ when $k \le d$.

Next we bound the variance $\|\mathbb{E}[\sum_{i\in[d]} M_i^2]\|$. Since all the M_i's are i.i.d., it suffices to analyze $\mathbb{E}[M_1^2]$. Let $T = \mathbb{E}[M_1^2] = \mathbb{E}[((a_1 a_1^\top) * Q)^2]$, by definition of Hadamard product, we know

$$T_{p,q} = \mathbb{E}[\sum_{r\in[k]} Q_{p,r}Q_{r,q}a_1(p)a_1(q)a_1(r)^2].$$

This number is 0 when $p \neq q$ by independence of entries of a_1. When $p = q$, this is bounded by $3\sum_{r\in[k]} Q_{p,r}^2$ because $\mathbb{E}[a_1(p)^2 a_1(r)^2]$ is 1 when $p \neq r$ and 3 when $p = r$. Therefore $T_{p,p} \le 3\sum_{r\in[k]} Q_{p,r}^2 = 3\|Q^{(p)}\|^2 \le O(dk)$. Since T is a diagonal matrix, we know $\|T\| \le O(dk)$, and $\sigma^2 = \|dT\| = O(d^2 k)$.

By Matrix Bernstein we know with high probability $\|M\| \le O(d\sqrt{k\log d})$. □

Using this lemma, it is easy to get a bound when columns of A, B are unit vectors. In this case, we just need to normalize the columns, the normalization factor is bounded between $d^2/2$ and $2d^2$ with high probability, and therefore, $\|(A^\top A)(B^\top B) - I\| \le O(\sqrt{k\log d}/d)$.

B Proof of Convergence Results in Theorems 1 and 2

The main part of the proof is to show that error contraction happens in each iteration of Algorithms 1 and 4 as the two main parts of the algorithm. Then, the contraction result after t iterations is directly argued.

In the following, we first provide a local contraction result for the tensor power iteration (5) in Algorithm 1 given noisy tensor $\hat{T}$. This leads to Lemma 1 which is the local convergence guarantee of the tensor power updates. Then, we provide a local contraction argument for the coordinate descent step (9) in Algorithm 4.

Combining the above convergence arguments for both updates conclude the overall local convergence guarantee in Theorem. 1. Then, combing this local convergence guarantee and the initialization result in Theorem 3 leads to the global convergence guarantee in Theorem 2. In addition, the result in Corollary 2 is similarly argued where the bound on the spectral norm of the tensor is argued in Lemma 6.

B.1 Convergence of tensor power iteration: Algorithm 1

In this section, we prove Lemma 1 which is the local convergence guarantee of the tensor power updates in Algorithm 1.

Define function $f(\epsilon; k, d)$ as

$$f(\epsilon; k, d) := \alpha\frac{\sqrt{k}}{d} + \frac{2\alpha}{\sqrt{d}}\left(1 + \alpha_0\sqrt{\frac{k}{d}}\right)^2 \epsilon + \alpha_0\epsilon^2, \tag{22}$$

where $\alpha = \text{polylog}(d)$ and $\alpha_0 = O(1)$. Notice that this function is a small constant when $k < d^{1.5}/\operatorname{poly}\log d$.

Lemma 10 (Contraction result of Algorithm 1 in one update)**.** *Consider $\hat{T} = T + \Psi$ as the input to Algorithm 1, where T is a rank-k tensor, and Ψ is a perturbation tensor. Suppose Assumptions (A1)-(A5) hold, and estimates $\hat{a}$ and $\hat{b}$ satisfy distance bounds*

$$\begin{aligned}\mathrm{dist}(\hat{a}, a_j) &\leq \epsilon_a,\\ \mathrm{dist}(\hat{b}, b_j) &\leq \epsilon_b,\end{aligned}$$

for some $j \in [k]$, and $\epsilon_a, \epsilon_b > 0$. Let $\epsilon := \max\{\epsilon_a, \epsilon_b\}$, and suppose ψ defined in (20) be small enough such that[10]

$$w_j - w_j\epsilon^2 - w_{\max} f(\epsilon; k, d) - \psi > 0,$$

where $f(\epsilon; k, d)$ is defined in (22). Then, update $\hat{c}$ in (5) satisfies the following distance bound with high probability (w.h.p.)

$$\mathrm{dist}(\hat{c}, c_j) \leq \frac{w_{\max} f(\epsilon; k, d) + \psi}{w_j - w_j\epsilon^2 - w_{\max} f(\epsilon; k, d) - \psi}. \tag{23}$$

Furthermore, if the bound in (23) is such that $\mathrm{dist}(\hat{c}, c_j) \leq \epsilon$, then the update $\hat{w} := \hat{T}(\hat{a}, \hat{b}, \hat{c})$ in (7) also satisfies w.h.p.

$$|\hat{w} - w_j| \leq 2w_j\epsilon^2 + w_{\max} f(\epsilon; k, d) + \psi.$$

Remark 2. *In the asymptotic regime, $f(\epsilon; k, d)$ is*

$$f(\epsilon; k, d) = \tilde{O}\left(\frac{\sqrt{k}}{d}\right) + \tilde{O}\left(\max\left\{\frac{1}{\sqrt{d}}, \frac{k}{d^{3/2}}\right\}\right)\epsilon + O(1)\epsilon^2.$$

Note that the last term is the only effective contracting term. The other terms include a constant term, and the term involving ϵ disappears in only one iteration as long as $k, d \to \infty$, and $\tilde{O}\left(\frac{k}{d^{3/2}}\right) \to 0$.

Remark 3 (Rate of convergence)**.** *The local convergence result provided in Theorem 1 has a linear convergence rate. But, Algorithm 1 actually provides an almost-quadratic convergence rate in the beginning, and linear convergence rate later on. It can be seen by referring to one-step contraction argument provided in Lemma 10 where the quadratic term $\alpha_0\epsilon^2$ exists. In the beginning, this term is dominant over linear term involving ϵ, and we have almost-quadratic convergence. Writing $\alpha_0\epsilon^2 = \alpha_0\epsilon^\zeta\epsilon^{2-\zeta}$, we observe that we get rate of convergence equal to $2-\zeta$ as long as we have initialization error bounded as $\epsilon_0^\zeta = O(1)$. Therefore, we can get arbitrarily close to quadratic convergence with appropriate initialization error. Note that when the model is more overcomplete, the algorithm more rapidly reaches to the linear convergence phase. For the sake of clarity, in proposing Theorem 1, we approximated the almost-quadratic convergence rate in the beginning with linear convergence.*

Lemma 10 is proposed in the general form. In Lemma 11, we provide explicit contraction result by imposing additional perturbation, contraction and initialization Assumptions (A6), (A8) and (A9). We observe that under reasonable rank, perturbation and initialization conditions, the denominator in (23) can be lower bounded by a constant, and the numerator is explicitly bounded by a term involving ϵ, and a constant non-contracting term.

[10] This is the denominator of bound provided in (23).

Lemma 11 (Contraction result of Algorithm 1 in one update)**.** *Consider $\hat{T} = T + \Psi$ as the input to Algorithm 1, where T is a rank-k tensor, and Ψ is a perturbation tensor. Let Assumptions*[11] *(A1)-(A9) hold. Note that initialization bound in (A9) is satisfied for some $j \in [k]$. Then, update $\hat{c}$ in* (5) *satisfies the following distance bound with high probability (w.h.p.)*

$$\operatorname{dist}(\hat{c}, c_j) \leq \underbrace{\text{Const.}}_{\text{non-contracting term}} + \underbrace{q\epsilon_0}_{\text{contracting term}},$$

where

$$\text{Const.} := \frac{2}{w_{\min}} \left(\psi + w_{\max} \alpha \frac{\sqrt{k}}{d} \right), \tag{24}$$

and contraction ratio $q < 1/2$ is defined in (21)*. Note that $\alpha = \operatorname{polylog}(d)$. In addition, if the above bound be such that $\operatorname{dist}(\hat{c}, c_j) \leq \epsilon_0$, then the update $\hat{w} := \hat{T}(\hat{a}, \hat{b}, \hat{c})$ in* (7) *also satisfies w.h.p.*

$$|\hat{w} - w_j| \leq \frac{w_{\min}}{2} \text{Const.} + w_{\min} q \epsilon_0.$$

Proof of Lemma 1: We incorporate condition (A7) to show that $q < 1/2$ in assumption (A8) is satisfied. In addition, (A7) implies that the bound on ϵ_0 in assumption (A9) holds where it can be shown that the bound in (A9) is bounded as $O(1/\gamma)$. Then, the result is directly proved by iteratively applying the result of Lemma 11. □

Proof of auxiliary lemmata: tensor power iteration in Algorithm 1

Before providing the proofs, we remind a few definitions and notations.

In Assumption (A2), matrices J_A, J_B, and J_C, are defined as incoherence matrices with zero diagonal entries such that $A^\top A = I + J_A$, $B^\top B = I + J_B$, and $C^\top C = I + J_C$. We have $\max\{\|J_A\|_\infty, \|J_B\|_\infty, \|J_C\|_\infty\} \leq \rho$ as in (19).

Given matrix $A \in \mathbb{R}^{d \times k}$, the following notations are defined to refer to its sub-matrices. A_j denotes the j-th column and A^j denotes the j-th row of A. Hence, we have $A_j = a_j, j \in [k]$. In addition, $A_{\setminus j} \in \mathbb{R}^{d \times (k-1)}$ is A with its j-th column removed, and $A^{\setminus j} \in \mathbb{R}^{(d-1) \times k}$ is A with its j-th row removed.

Proof of Lemma 10: Let $z_a^* \perp a_j$ and $z_b^* \perp b_j$ denote the vectors that achieve supremum value in (12) corresponding to $\operatorname{dist}(\hat{a}, a_j)$ and $\operatorname{dist}(\hat{b}, b_j)$, respectively. Furthermore, without loss of generality, assume $\|z_a^*\| = \|z_b^*\| = 1$. Then, $\hat{a}$ and $\hat{b}$ are decomposed as

$$\hat{a} = \langle a_j, \hat{a} \rangle a_j + \operatorname{dist}(\hat{a}, a_j) z_a^*, \tag{25a}$$

$$\hat{b} = \langle b_j, \hat{b} \rangle b_j + \operatorname{dist}(\hat{b}, b_j) z_b^*. \tag{25b}$$

Let $\overline{C} := C \operatorname{Diag}(w)$ denote the unnormalized matrix C, and $\tilde{c} := \hat{T}(\hat{a}, \hat{b}, I)$ denote the unnormalized update in (5). The goal is to bound $\operatorname{dist}\left(\tilde{c}, \overline{C}_j\right)$. Consider any $z_c \perp \overline{C}_j$ such that $\|z_c\| = 1$. Then, we have

$$\langle z_c, \tilde{c} \rangle = \hat{T}(\hat{a}, \hat{b}, z_c) = T(\hat{a}, \hat{b}, z_c) + \Psi(\hat{a}, \hat{b}, z_c).$$

[11] As mentioned in the assumptions, from perturbation bound in (A6), only the bound $\psi \leq \frac{w_{\min}}{6}$ is required here.

Substituting $\hat{a}$ and $\hat{b}$ from $(25a)$ and $(25b)$, we have

$$T(\hat{a},\hat{b},z_c) = \underbrace{\langle a_j,\hat{a}\rangle\langle b_j,\hat{b}\rangle T(a_j,b_j,z_c)}_{S_1} + \underbrace{\langle a_j,\hat{a}\rangle \operatorname{dist}(\hat{b},b_j)T(a_j,z_b^*,z_c)}_{S_2} + \underbrace{\operatorname{dist}(\hat{a},a_j)\langle b_j,\hat{b}\rangle T(z_a^*,b_j,z_c)}_{S_3} + \underbrace{\operatorname{dist}(\hat{a},a_j)\operatorname{dist}(\hat{b},b_j)T(z_a^*,z_b^*,z_c)}_{S_4}.$$

In the following derivations, we repeatedly use the equality that for any $u,v \in \mathbb{R}^d$, we have $T(u,v,I) = \overline{C}(A^\top u * B^\top v)$. For S_1, we have

$$\begin{aligned} S_1 \le |T(a_j,b_j,z_c)| &= |z_c^\top \overline{C}(A^\top a_j * B^\top b_j)| \\ &= \left| z_c^\top \overline{C}\left[e_j + (J_A * J_B)_j\right]\right| \\ &= \left| z_c^\top \overline{C}_{\setminus j}\,(J_A * J_B)_j^{\setminus j}\right| \\ &\le w_{\max}\alpha\frac{\sqrt{k}}{d}, \end{aligned}$$

where equalities $A^\top A = I + J_A$ and $B^\top B = I + J_B$ are exploited in the second equality, and the assumption that $z_c \perp \overline{C}_j$ is used in the last equality. The last inequality is from Assumption (A4). For S_2, we have

$$\begin{aligned} S_2 \le \epsilon_b|T(a_j,z_b^*,z_c)| &= \epsilon_b|z_c^\top \overline{C}(A^\top a_j * B^\top z_b^*)| \\ &= \epsilon_b\left| z_c^\top \overline{C}_{\setminus j}\left[(J_A)_j^{\setminus j} * \left(B_{\setminus j}\right)^\top z_b^*\right]\right| \\ &\le \epsilon_b \left\|\overline{C}_{\setminus j}\right\| \cdot \left\|(J_A)_j^{\setminus j}\right\|_\infty \cdot \left\|\left(B_{\setminus j}\right)^\top z_b^*\right\| \\ &\le w_{\max}\frac{\alpha}{\sqrt{d}}\left(1+\alpha_0\sqrt{\frac{k}{d}}\right)^2\epsilon_b, \end{aligned}$$

for some $\alpha = \operatorname{polylog}(d)$ and $\alpha_0 = O(1)$. Second inequality is concluded from $\|u*v\| \le \|u\|_\infty \cdot \|v\|$, and Assumptions (A2) and (A3) are exploited in the last inequality. Similarly, for S_3, we have

$$\begin{aligned} S_3 &\le \epsilon_a\left| z_c^\top \overline{C}_{\setminus j}\left[(J_B)_j^{\setminus j} * \left(A_{\setminus j}\right)^\top z_a^*\right]\right| \\ &\le w_{\max}\frac{\alpha}{\sqrt{d}}\left(1+\alpha_0\sqrt{\frac{k}{d}}\right)^2\epsilon_a. \end{aligned}$$

Finally, for S_4, we have

$$S_4 \le \epsilon_a\epsilon_b|T(z_a^*,z_b^*,z_c)| \le \epsilon_a\epsilon_b\|T\| \le w_{\max}\alpha_0\epsilon_a\epsilon_b,$$

for some $\alpha_0 = O(1)$. The bound on $\|T\|$ is from Assumption (A4). Note that for random components, we showed in Lemma 5 that this bound holds w.h.p. exploiting Assumption (A5) and results of Guédon and Rudelson (2007). For the error term $\Psi(\hat{a},\hat{b},z_c)$, we have

$$\Psi(\hat{a},\hat{b},z_c) \le \psi,$$

which is concluded from the definition of spectral norm of a tensor. Note that all vectors $\hat{a}$, $\hat{b}$, z_c have unit norm.

Let $\epsilon := \max\{\epsilon_a, \epsilon_b\}$. Then, combining all the above bounds, we have w.h.p.

$$\langle z_c, \tilde{c}\rangle \leq w_{\max} f(\epsilon; k, d) + \psi,$$

where $f(\epsilon; k, d)$ is

$$f(\epsilon; k, d) := \alpha\frac{\sqrt{k}}{d} + \frac{2\alpha}{\sqrt{d}}\left(1 + \alpha_0\sqrt{\frac{k}{d}}\right)^2\epsilon + \alpha_0\epsilon^2.$$

For $\tilde{c}$, we have

$$\begin{aligned}
\tilde{c} &= T(\hat{a}, \hat{b}, I) + \Psi(\hat{a}, \hat{b}, I) \\
&= \sum_i w_i\langle a_i, \hat{a}\rangle\langle b_i, \hat{b}\rangle c_i + \Psi(\hat{a}, \hat{b}, I) \\
&= w_j\langle a_j, \hat{a}\rangle\langle b_j, \hat{b}\rangle c_j + \sum_{i\neq j} w_i\langle a_i, \hat{a}\rangle\langle b_i, \hat{b}\rangle c_i + \Psi(\hat{a}, \hat{b}, I),
\end{aligned}$$

and therefore,

$$\begin{aligned}
\|\tilde{c}\| &\geq \left\|w_j\langle a_j, \hat{a}\rangle\langle b_j, \hat{b}\rangle c_j\right\| - \left\|\sum_{i\neq j} w_i\langle a_i, \hat{a}\rangle\langle b_i, \hat{b}\rangle c_i\right\| - \|\Psi(\hat{a}, \hat{b}, I)\| \\
&\geq w_j - w_j\epsilon^2 - w_{\max} f(\epsilon; k, d) - \psi,
\end{aligned}$$

where inequality $\langle a_j, \hat{a}\rangle\langle b_j, \hat{b}\rangle \geq 1 - \epsilon^2$ is exploited in the last inequality. Hence, as long as this lower bound on $\|\tilde{c}\|$ is positive (small enough ϵ and ψ), we have

$$\operatorname{dist}(\tilde{c}, \overline{C}_j) \leq \frac{w_{\max} f(\epsilon; k, d) + \psi}{w_j - w_j\epsilon^2 - w_{\max} f(\epsilon; k, d) - \psi}. \tag{26}$$

Since $\operatorname{dist}(\cdot,\cdot)$ function is invariant with respect to norm, we have $\operatorname{dist}(\hat{c}, c_j) = \operatorname{dist}\left(\tilde{c}, \overline{C}_j\right)$ which finishes the proof for bounding $\operatorname{dist}(\hat{c}, c_j)$. Note that $\tilde{c} = \|\tilde{c}\|\hat{c}$, and $\overline{C}_j = w_j c_j$ where $w_j > 0$.

Now, we provide the bound on $|w_j - \hat{w}|$. As assumed in the lemma, we have distance bounds

$$\max\left\{\operatorname{dist}(\hat{a}, a_j), \operatorname{dist}\left(\hat{b}, b_j\right), \operatorname{dist}(\hat{c}, c_j)\right\} \leq \epsilon.$$

The estimate $\hat{w} = \hat{T}(\hat{a}, \hat{b}, \hat{c})$ proposed in (7) can be expanded as

$$\begin{aligned}
\hat{w} &= T(\hat{a}, \hat{b}, \hat{c}) + \Psi(\hat{a}, \hat{b}, \hat{c}) \\
&= \sum_i w_i\langle a_i, \hat{a}\rangle\langle b_i, \hat{b}\rangle\langle c_i, \hat{c}\rangle + \Psi(\hat{a}, \hat{b}, \hat{c}) \\
&= w_j\langle a_j, \hat{a}\rangle\langle b_j, \hat{b}\rangle\langle c_j, \hat{c}\rangle + \sum_{i\neq j} w_i\langle a_i, \hat{a}\rangle\langle b_i, \hat{b}\rangle\langle c_i, \hat{c}\rangle + \Psi(\hat{a}, \hat{b}, \hat{c}),
\end{aligned}$$

and therefore,

$$\begin{aligned}|w_j - \hat{w}| &\leq \left|w_j\left(1 - \langle a_j, \hat{a}\rangle\langle b_j, \hat{b}\rangle\langle c_j, \hat{c}\rangle\right)\right| + \left|\sum_{i\neq j} w_i\langle a_i, \hat{a}\rangle\langle b_i, \hat{b}\rangle\langle c_i, \hat{c}\rangle\right| + \left|\Psi(\hat{a}, \hat{b}, \hat{c})\right| \\ &\leq w_j\left(1 - \left(1-\epsilon^2\right)^{1.5}\right) + w_{\max} f(\epsilon; k, d) + \psi \\ &\leq 2w_j\epsilon^2 + w_{\max} f(\epsilon; k, d) + \psi,\end{aligned}$$

where $\langle a_j, \hat{a}\rangle\langle b_j, \hat{b}\rangle\langle c_j, \hat{c}\rangle \geq \left(1-\epsilon^2\right)^{1.5}$ is exploited in the second inequality. Notice that this argument is similar to the argument provided earlier for lower bounding $\|\tilde{c}\|$.

□

Proof of Lemma 11: The result is proved by applying Lemma 10, and incorporating additional conditions (A6), (A8), and (A9). $f(\epsilon_0; k, d)$ in (22) can be bounded as

$$\begin{aligned}f(\epsilon_0; k, d) &= \alpha\frac{\sqrt{k}}{d} + \frac{2\alpha}{\sqrt{d}}\left(1 + \alpha_0\sqrt{\frac{k}{d}}\right)^2\epsilon_0 + \alpha_0\epsilon_0^2 \\ &\leq \alpha\frac{\sqrt{k}}{d} + \left[\frac{2\alpha}{\sqrt{d}}\left(1 + \alpha_0\sqrt{\frac{k}{d}}\right)^2 + \beta'\right]\epsilon_0 \\ &= \alpha\frac{\sqrt{k}}{d} + \frac{w_{\min}}{2w_{\max}}q\epsilon_0,\end{aligned}$$

where $\epsilon_0 \leq \frac{\beta'}{\alpha_0}$ from Assumption (A9) is exploited in the inequality. The last equality is concluded from definition of contracting factor q in (21). On the other hand, the denominator in (23) can be lower bounded as

$$w_{\min}\left[1 - \frac{w_{\max}}{w_{\min}}\epsilon_0^2 - \frac{w_{\max}}{w_{\min}}f(\epsilon_0; k, d) - \frac{\psi}{w_{\min}}\right] \geq w_{\min}\left[1 - \frac{1}{6} - \frac{1}{6} - \frac{1}{6}\right] = \frac{w_{\min}}{2},$$

where Assumptions (A9) and (A6) are used in the inequality. Applying Lemma 10, the result on $\text{dist}(\hat{c}, c_j)$ is proved.

From Lemma 10, we also have

$$\begin{aligned}|\hat{w} - w_j| &\leq 2w_j\epsilon_0^2 + w_{\max} f(\epsilon_0; k, d) + \psi \\ &\leq \frac{w_{\min}}{2}\,\text{Const.} + 2w_j\epsilon_0^2 + \frac{w_{\min}}{2}q\epsilon_0 \\ &\leq \frac{w_{\min}}{2}\,\text{Const.} + w_{\min}q\epsilon_0.\end{aligned}$$

where $\epsilon_0 \leq \frac{w_{\min}q}{4w_{\max}}$ from Assumption (A9) is used in the last inequality. □

B.2 Convergence of removing residual error: Algorithm 4

In this section, we provide convergence of the coordinate descent of Algorithm 4 for removing the residual error. We first provide the following definition.

Definition 2 ((η_0, η_1)-nice)**.** *Suppose*

$$\max\{\|A\|, \|B\|, \|C\|\} \le \eta_1 \sqrt{\frac{k}{d}}.$$

Given an approximate solution $\{\widehat{A}, \widehat{B}, \widehat{C}, \widehat{w}\}$, we call it (η_0, η_1)-nice if matrix $\widehat{A}$ (similarly $\widehat{B}$ and $\widehat{C}$) satisfies

$$\|\Delta A_i\| := \|\widehat{a}_i - a_i\| \le \eta_0 \frac{\sqrt{k}}{d}, \quad \forall i \in [k],$$
$$\|\widehat{A}\| \le \eta_1 \sqrt{\frac{k}{d}},$$

and the weights satisfy

$$|\widehat{w}_i - w_i| \le \eta_0 w_{\max} \frac{\sqrt{k}}{d}.$$

Given above conditions are satisfied, we prove the following guarantees for removing residual error, Algorithm 4.

Lemma 12 (Local convergence guarantee of the iterations for removing residual error, Algorithm 4)**.** *Consider T as the input to Algorithm 4, where T is a rank-k tensor. Suppose Assumptions (A1)-(A5) and (A11) hold (which are satisfied whp when the components are uniformly i.i.d. drawn from unit d-dimensional sphere). Given initial solution $\left\{\widehat{A}^{(0)}, \widehat{B}^{(0)}, \widehat{C}^{(0)}, \widehat{w}^{(0)}\right\}$ which is (η_0, η_1)-nice, all the following iterations of Algorithm 4 are $(2\eta_0, 3\eta_1)$-nice. Furthermore, given the exact tensor T, the Frobenius norm error $\max\{\|\Delta A\|_F, \|\Delta B\|_F, \|\Delta C\|_F, \|\Delta w\|/w_{\min}\}$ shrinks by at least a factor of 2 in every iteration. In addition, if we have a noisy tensor $\hat{T} = T + \Psi$ such that $\|\Psi\| \le \psi$, then*

$$\max\{\|\Delta A^{(t)}\|_F, \|\Delta B^{(t)}\|_F, \|\Delta C^{(t)}\|_F, \|\Delta w^{(t)}\|/w_{\min}\} \le 2^{-t}\eta_0 \frac{k}{d} + O\left(\frac{\psi\sqrt{k}}{w_{\min}}\right).$$

Proof: iteration for removing residual error in Algorithm 4

We now prove Lemma 12 as the local convergence guarantee of the iterations for removing residual error, Algorithm 4.

To prove this lemma, we first observe that the algorithm update formula in (9) is (before normalization) $w_i\langle a_i, \widehat{a}_i\rangle\langle b_i, \widehat{b}_i\rangle c_i + \epsilon_i$ where

$$\epsilon_i = \sum_{j \ne i} (w_i\langle a_j, \widehat{a}_i\rangle\langle b_j, \widehat{b}_i\rangle c_j - \widehat{w}_i\langle \widehat{a}_i, \widehat{a}_j\rangle\langle \widehat{b}_i, \widehat{b}_j\rangle \widehat{c}_j).$$

In the following lemma, we show that the error terms ϵ_i's are small.

Lemma 13. *Before normalization $\tilde{w}_i\tilde{c}_i = w_i\langle a_i, \widehat{a}_i\rangle\langle b_i, \widehat{b}_i\rangle c_i + \epsilon_i$ where*

$$\sum_{i=1}^{k} \|\epsilon_i\|^2 \le o(1)(w_{\max}(\|\Delta(A)\|_F^2 + \|\Delta(B)\|_F^2 + \|\Delta(C)\|_F^2) + \|\Delta w\|^2).$$

Proof: By the update formula in (9), we know

$$\epsilon_i = \sum_{j\neq i}(w_i\langle a_j,\widehat{a}_i\rangle\langle b_j,\widehat{b}_i\rangle c_j - \widehat{w}_i\langle\widehat{a}_i,\widehat{a}_j\rangle\langle\widehat{b}_i,\widehat{b}_j\rangle\widehat{c}_j).$$

We expand it into several terms as follows.

$$\begin{aligned}
\epsilon_i &= \sum_{j\neq i}(w_i\langle a_j,\widehat{a}_i\rangle\langle b_j,\widehat{b}_i\rangle c_j - \widehat{w}_i\langle\widehat{a}_i,\widehat{a}_j\rangle\langle\widehat{b}_i,\widehat{b}_j\rangle\widehat{c}_j)\\
&= \sum_{j\neq i}\langle a_i,a_j\rangle\langle b_i,b_j\rangle(w_jc_j - \widehat{w}_j\widehat{c}_j) \quad \text{(type 1)}\\
&\quad + \sum_{j\neq i} w_j\langle a_j,\Delta A_i\rangle\langle b_j,b_i\rangle c_j + \sum_{j\neq i} w_j\langle a_j,a_i\rangle\langle b_j,\Delta B_i\rangle c_j \quad \text{(type 2)}\\
&\quad - \sum_{j\neq i}\widehat{w}_j\langle a_j,a_i\rangle\langle b_j,\Delta B_i\rangle\widehat{c}_j - \sum_{j\neq i}\widehat{w}_j\langle a_j,a_i\rangle\langle\Delta B_j,\widehat{b}_i\rangle\widehat{c}_j\\
&\quad - \sum_{j\neq i}\widehat{w}_j\langle a_j,\Delta A_i\rangle\langle b_j,b_i\rangle\widehat{c}_j - \sum_{j\neq i}\widehat{w}_j\langle\Delta A_j,\widehat{a}_i\rangle\langle b_j,b_i\rangle\widehat{c}_j\\
&\quad + \sum_{j\neq i}\langle a_j,\Delta A_i\rangle\langle b_j,\Delta B_i\rangle c_j \quad \text{(type 3)}\\
&\quad - \sum_{j\neq i}\widehat{w}_j\langle a_j,\Delta A_i\rangle\langle b_j,\Delta B_i\rangle\widehat{c}_j - \sum_{j\neq i}\widehat{w}_j\langle\Delta A_j,\widehat{a}_i\rangle\langle b_j,\Delta B_i\rangle\widehat{c}_j\\
&\quad - \sum_{j\neq i}\widehat{w}_j\langle a_j,\Delta A_i\rangle\langle\Delta B_j,\widehat{b}_i\rangle\widehat{c}_j - \sum_{j\neq i}\widehat{w}_j\langle\Delta A_j,\widehat{a}_i\rangle\langle\Delta B_j,\widehat{b}_i\rangle\widehat{c}_j.
\end{aligned}$$

The norm of three different types of terms mentioned above are bounded in Section B.2, which conclude the desired bound in the lemma. □

We are now ready to prove main Lemma 12.

Proof of Lemma 12: Since $\tilde{w}_i$ is the norm of $w_i\langle a_i,\widehat{a}_i\rangle\langle b_i,\widehat{b}_i\rangle c_i + \epsilon_i$, we know

$$|\tilde{w}_i - w_i| \leq \|\epsilon_i\| + w_i(\Theta(\|\Delta A_i\|^2 + \|\Delta B_i\|^2)),$$

and therefore

$$\|\tilde{w} - w\| \leq o(1)(w_{\max}(\|\Delta(A)\|_F + \|\Delta(B)\|_F + \|\Delta(C)\|_F) + \|\Delta w\|).$$

On the other hand, since the coefficient $w_i\langle a_i,\widehat{a}_i\rangle\langle b_i,\widehat{b}_i\rangle$ is at least $1-o(1)$, we know $\|\tilde{c}_i - c_i\| \leq 4\|\epsilon_i\|/w_{\min}$. This implies

$$\|\tilde{C} - C\|_F \leq o(1)((\|\Delta(A)\|_F + \|\Delta(B)\|_F + \|\Delta(C)\|_F) + \|\Delta w\|/w_{\min}).$$

By Lemma 14, we know after the projection procedure, we get $\|\widehat{C} - C\|_F \leq 2\|\tilde{C} - C\|_F$. Therefore combining the two steps we know

$$\|\widehat{C} - C\|_F \leq 2\|\tilde{C} - C\|_F \leq o(1)(\|\Delta(A)\|_F + \|\Delta(B)\|_F + \|\Delta(C)\|_F + \|\Delta w\|/w_{\min}).$$

When we have noise, all the ϵ_i's have an additional term $\Psi(\widehat{a}_i,\widehat{b}_i,I)$ which is bounded by ψ, and thus, the second part of the lemma follows directly.

□

Handling Symmetric Tensors: For symmetric tensors we should change the algorithm as computing the following:

$$T(\widehat{a}_i, \widehat{b}_i, I) - \frac{1}{d}\sum_{i=1}^{d} T(e_i, e_i, I) - \sum_{j\neq i} \widehat{w}_j(\langle \widehat{a}_i, \widehat{a}_j\rangle\langle \widehat{b}_i, \widehat{b}_j\rangle - \frac{1}{d})\widehat{c}_j. \tag{27}$$

The result of this will be a change in the term of type 1. Now the Q matrix will be $(A\odot A)^T(A\odot A) - (1-\frac{1}{d})I - \frac{1}{d}J$ which has desired spectral norm for random matrices.

Claims for proving Lemma 13

The first term deals with the difference between C and $\widehat{C}$.

Claim 1. *We have*

$$\sqrt{\sum_{i=1}^{k} \| \sum_{j\neq i} \langle a_i, a_j\rangle\langle b_i, b_j\rangle(w_i c_i - \widehat{w}_i\widehat{c}_i)\|^2} \leq o(1)(w_{\max}\|\Delta C\|_F + \|\widehat{w} - w\|).$$

Proof: This sum is equal to the Frobenius norm of a matrix $M = QZ$. Here the matrix Q is a matrix such that is equal to $Q = (A\odot B)^\top(A\odot B) - I$:

$$Q_{i,j} = \begin{cases} \langle a_i, a_j\rangle\langle b_i, b_j\rangle, & i\neq j, \\ 0, & i = j, \end{cases}$$

The matrix Z has columns $Z_i = w_i c_i - \widehat{w}_i\widehat{c}_i$. By assumption we know $\|Q\| \leq o(1)$, and $\|Z\|_F \leq w_{\max}\|\Delta C\|_F + \|\widehat{w} - w\|$. Therefore we have

$$\|M\|_F = \|QZ\|_F \leq \|Q\|\|Z\|_F \leq o(1)(w_{\max}\|\Delta C\|_F + \|\widehat{w} - w\|).$$

□

Of course, in the error ϵ_i, we don't have $\sum_{j\neq i}\langle a_i, a_j\rangle\langle b_i, b_j\rangle w_i c_i$, instead we have terms like $\sum_{j\neq i}\langle \widehat{a}_i, a_j\rangle\langle \widehat{b}_i, b_j\rangle w_i c_i$. The next two lemmas show that these two terms are actually very close.

Claim 2. *We have*

$$\sqrt{\sum_{i=1}^{k} \| \sum_{j\neq i} \langle \Delta A_i, \widehat{a}_j\rangle\langle b_i, b_j\rangle\widehat{w}_i\widehat{c}_i\|^2} \leq o(w_{\max})\|\Delta A\|_F.$$

$$\sqrt{\sum_{i=1}^{k} \| \sum_{j\neq i} \langle \Delta A_j, \widehat{a}_i\rangle\langle b_i, b_j\rangle\widehat{w}_i\widehat{c}_i\|^2} \leq o(w_{\max})\|\Delta A\|_F.$$

Same is true if any $\widehat{\cdot}$ is replaced by the true value.

Proof: Similar as before, we treat the left hand side as the Frobenius norm of some matrix $M = QZ$. Here $Z_i = \widehat{w}_i\widehat{c}_i$, and Q is the following matrix:

$$Q_{i,j} = \begin{cases} \langle \Delta A_i, \widehat{a}_j\rangle\langle b_i, b_j\rangle, & i\neq j, \\ 0, & i = j, \end{cases}$$

We shall bound $\|M\|_F$ by $\|Z\|\|Q\|_F$. By assumption we know $\|Z\| \le w_{\max} \cdot 2\eta_1\sqrt{k/d} = O(w_{\max}\sqrt{k/d})$. On the other hand, we know $\langle b_i, b_j\rangle \le \tilde{O}(1/\sqrt{d})$ hence $\|Q\|_F \le \tilde{O}(1/\sqrt{d})\|\widehat{A}^T\Delta A\|_F \le \tilde{O}(1/\sqrt{d})\|\widehat{A}\|\|\Delta A\|_F = \tilde{O}(\sqrt{k}/d)\|\Delta A\|_F$. Therefore we have

$$\|M\|_F \le \|Z\|\|Q\|_F \le O(w_{\max}\sqrt{k/d})\cdot\tilde{O}(w_{\max}\sqrt{k}/d)\|\Delta A\|_F = \tilde{O}(k/d\sqrt{d})\|\Delta A\|_F = o(w_{\max})\|\Delta A\|_F.$$

Notice that the proof works for both terms. □

Claim 3. *We have*

$$\sqrt{\sum_{i=1}^{k}\|\sum_{j\neq i}\langle\Delta A_i, \widehat{a}_j\rangle\langle\Delta B_i, \widehat{b}_j\rangle\widehat{w}_i\widehat{c}_i\|^2} \le o(w_{\max})(\|\Delta A\|_F + \|\Delta B\|_F).$$

The same is true if the inner-products are between $\langle\Delta A_j, \widehat{a}_i\rangle$ or $\langle\Delta B_j, \widehat{b}_i\rangle$, or if any $\widehat{\cdot}$ is replaced by the true value.

Proof: Similar as before, we treat the left hand side as the Frobenius norm of some matrix $M = QZ$. Here $Z_i = \widehat{w}_i\widehat{c}_i$, and Q is the following matrix

$$Q_{i,j} = \begin{cases} \langle\Delta A_i, \widehat{a}_j\rangle\langle\Delta B_i, b_j\rangle, & i \neq j, \\ 0, & i = j, \end{cases}$$

Now using definition of $2 \to 4$ norm and $2ab \le a^2 + b^2$ we first bound the Frobenius norm of the matrix Q:

$$\sum_{i\neq j}(\langle\Delta A_i, \widehat{a}_j\rangle\langle\Delta B_i, \widehat{b}_j\rangle)^2 \le \sum_{i\neq j}(\langle\Delta A_i, \widehat{a}_j\rangle)^4 + (\langle\Delta B_i, \widehat{b}_j\rangle)^4 \le \sum_{i=1}^{k}\|\widehat{A}^\top\|_{2\to4}\|\Delta A_i\|^4 + \|\widehat{B}^\top\|_{2\to4}\|\Delta B_i\|^4$$

Now we first bound the $2 \to 4$ norm of the matrix $\widehat{A}^\top = A^\top + \Delta A^\top$. By assumption we already know $\|A^\top\|_{2\to4} \le O(1)$. On the other hand, for any unit vector u

$$\sum_{i=1}^{k}\langle\Delta A_i, u\rangle^4 \le \max_{i=1}^{k}\langle\Delta A_i, u\rangle^2 \sum_{i=1}^{k}\langle\Delta A_i, u\rangle^2 \le \tilde{O}(k^2/d^3) = o(1).$$

Here we used the assumption that $\|\Delta A_i\| \le \tilde{O}(\sqrt{k}/d)$ and $\|\Delta A\| \le O(\sqrt{k/d})$. Therefore $\|\widehat{A}^\top\|_{2\to4} \le \|A^\top\|_{2\to4} + \|\Delta A^\top\|_{2\to4} \le O(1)$ (and similarly for $\widehat{B}^\top$).

Therefore

$$\begin{aligned}
\|Q\|_F &\le \sqrt{\sum_{i=1}^{k}\|\widehat{A}^\top\|_{2\to4}\|\Delta A_i\|^4 + \|\widehat{B}^\top\|_{2\to4}\|\Delta B_i\|^4} \\
&\le O(1)\sqrt{\sum_{i=1}^{k}\|\Delta A_i\|^4 + \|\Delta B_i\|^4} \\
&\le O(1)\cdot\max_{i=1}^{k}(\|\Delta A\|_i + \|\Delta B\|_i)\sqrt{\sum_{i=1}^{k}\|\Delta A_i\|^2 + \|\Delta B_i\|^2} \\
&\le \tilde{O}(\sqrt{k}/d)(\|\Delta A\|_F + \|\Delta B\|_F).
\end{aligned}$$

On the other hand we know $\|Z\| \leq O(w_{\max}\sqrt{k/d})$, hence $\|M\|_F \leq \|Z\|\|Q\|_F \leq o(w_{\max})(\|\Delta A\|_F + \|\Delta B\|_F)$.

□

Projection Procedure 5

In this section, we describe the functionality of projection Procedure 5. Suppose the initial solution $\{\widehat{A}^0, \widehat{B}^0, \widehat{C}^0, \widehat{w}^0\}$ is (η_0, η_1)-nice. Then, given an arbitrary solution $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{w}\}$, we run projection Procedure 5 to get a $(2\eta_0, 4\eta_1)$-nice solution without losing too much in Frobenius norm error. This is shown in the following Lemma.

Lemma 14. *Suppose the initial solution $\{\widehat{A}^0, \widehat{B}^0, \widehat{C}^0, \widehat{w}^0\}$ is (η_0, η_1)-nice. For any solution $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{w}\}$, let error $E = \max\{\|\tilde{A} - A\|_F, \|\tilde{B} - B\|_F, \|\tilde{C} - C\|_F, \|\tilde{w} - w\|/w_{\min}\}$. Then after the projection Procedure 5, the new solution is $(2\eta_0, 3\eta_1)$-nice and has error at most $2E$.*

Proof: Intuitively, by truncating D the matrix we get is closest to $\tilde{A}$ among matrices with spectral norm $\eta_1\sqrt{k/d}$. We first prove this fact:

Claim 4.

$$\|Q - \tilde{A}\|_F = \min_{\|M\| \leq \eta_1\sqrt{k/d}} \|M - \tilde{A}\|_F.$$

Proof: By symmetric properties of Frobenius and spectral norm (both are invariant under rotation), we can rotate the matrices $Q, M, \tilde{A}$ simultaneously, so that $\tilde{A}$ becomes a diagonal matrix D. Since M has spectral norm bounded by $\eta_1\sqrt{k/d}$, in particular all its entries must be bounded by $\eta_1\sqrt{k/d}$. Also, we know $\|D - \widehat{D}\|_F = \min_{\forall(i,j) M_{i,j} \leq \eta_1\sqrt{k/d}} \|D - M\|_F$, therefore $\|D - \widehat{D}\|_F = \min_{\|M\| \leq \eta_1\sqrt{k/d}} \|D - M\|_F$. By the rotation invariant property this implies the claim. □

Since the optimal solution A has spectral norm bounded by $\eta_1\sqrt{k/d}$, in particular from above claim we know $\|Q - \tilde{A}\|_F \leq \|\tilde{A} - A\|_F$. By triangle inequality we get $\|Q - A\|_F \leq 2E$. In the next step we are essentially projecting the solution Q to a convex set that contains A (the set of matrices that are column-wise $\eta_1\sqrt{k}/d$ close to $\widehat{A}^0$), so the distance can only decrease. Similar arguments work for $\widehat{B}, \widehat{C}, \widehat{w}$, therefore the error of the new solution is bounded by $2E$.

By construction it is clear that the columns of the new solution is within $\eta_0\sqrt{k}/d$ to the columns of the initial solution, so they must be within $2\eta_0\sqrt{k}/d$ to the columns of the true solution. The only thing left to prove is that $\|\widehat{A}\| \leq 3\eta_1\sqrt{k/d}$.

First we observe that $\widehat{A} = \widehat{A}^0 + Z$ where Z is a matrix whose columns are multiples of $Q - \widehat{A}^0$, and the multiplier is never larger than 1. Therefore $\|\widehat{A}\| \leq \|hA^0\| + \|Z\| \leq \|\widehat{A}^0\| + \|Q - \widehat{A}^0\| \leq 2\|\widehat{A}^0\| + \|Q\| \leq 3\eta_1\sqrt{k/d}$. □

C SVD Initialization Result

In this section, we analyze the SVD-based initialization technique proposed in Procedure 2. The goal is to provide good initialization vectors close to the columns of true components A and B in the regime of $k = O(d)$.

Given a vector $\theta \in \mathbb{R}^d$, matrix $T(I, I, \theta)$ results a linear combination of slices of tensor T. For tensor T in (18), we have

$$T(I, I, \theta) = \sum_{i \in [k]} w_i \langle \theta, c_i \rangle a_i b_i^\top = \sum_{i \in [k]} \lambda_i a_i b_i^\top = A \, \mathrm{Diag}(\lambda) B^\top, \tag{28}$$

where $\lambda_i := w_i \langle \theta, c_i \rangle, i \in [k]$, and $\lambda := [\lambda_1, \lambda_2, \ldots, \lambda_k]^\top \in \mathbb{R}^k$ is expressed as

$$\lambda = \mathrm{Diag}(w) C^\top \theta.$$

Since A and B are not orthogonal matrices, the expansion in (28) is not the SVD [12] of $T(I, I, \theta)$. But, we show in the following theorem that if we draw enough number of random vectors θ in the regime of $k = O(d)$, we can eventually provide good initialization vectors through SVD of $T(I, I, \theta)$. Define

$$g(L) := \sqrt{2 \ln(L)} - \frac{\ln(\ln(L)) + c}{2\sqrt{2 \ln(L)}} - \sqrt{2 \ln(k)}.$$

Theorem 3 (SVD initialization when $k = O(d)$)**.** *Consider tensor $\hat{T} = T + \Psi$ where T is a rank-k tensor, and Ψ is a perturbation tensor. Let Assumptions (A1)-(A3) hold and $k = O(d)$. Draw L i.i.d. random vectors $\theta^{(j)} \sim \mathcal{N}(0, I_d), j \in [L]$. Let $u_1^{(j)}$ and $v_1^{(j)}$ be the top left and right singular vectors of $\hat{T}(I, I, \theta^{(j)})$. This is L random runs of Procedure 2. Suppose L satisfies the bound*

$$g(L) \geq \frac{w_{\max}(1 + \mu)}{w_{\min} - \rho w_{\max}(1 + \mu)} 4\sqrt{\log k},$$

with $\mu = \frac{2\mu_R + \tilde{\mu} - 1}{1 - \tilde{\mu}} < \frac{w_{\min}}{w_{\max}\rho} - 1$, for μ_R and $\mu_{\min}$ defined in (31), and some $0 < \tilde{\mu} < 1$. Note that $\rho \leq \frac{\alpha}{\sqrt{d}}$ is also defined as the incoherence parameter in Assumption (A2). Then, w.h.p., at least one of the pairs $(u_1^{(j)}, v_1^{(j)}), j \in [L]$, say j^*, satisfies

$$\max\left\{\mathrm{dist}\left(u_1^{(j^*)}, a_1\right), \mathrm{dist}\left(v_1^{(j^*)}, b_1\right)\right\} \leq \frac{4 w_{\max} \mu_{\min}(1 + \rho)\sqrt{\log k} + \alpha_0 \sqrt{d} \psi}{w_{\min} \tilde{\mu} g(L) - \alpha_0 \sqrt{d} \psi},$$

where $\psi := \|\Psi\|$ is the spectral norm of perturbation tensor Ψ, and $\alpha_0 > 1$ is a constant.

Proof: Let $\lambda^{(j)} := \mathrm{Diag}(w) C^\top \theta^{(j)} \in \mathbb{R}^k$ and $\tilde{\lambda}^{(j)} := C^\top \theta^{(j)} \in \mathbb{R}^k$. From Lemmata 15 and 16, there exists a $j^* \in [L]$ such that w.h.p., we have

$$\max\left\{\mathrm{dist}\left(u_1^{(j^*)}, a_1\right), \mathrm{dist}\left(v_1^{(j^*)}, b_1\right)\right\} \leq \frac{\mu_{\min} \lambda_{(2)} + \|\Psi(I, I, \theta)\|}{\tilde{\mu} \lambda_1 - \|\Psi(I, I, \theta)\|}.$$

From (29), with probability at least $1 - 2k^{-1}$, we have

$$\lambda_1^{(j^*)} \geq w_{\min} g(L).$$

From (30), with probability at least $1 - k^{-7}$, we have

$$\lambda_{(2)}^{(j^*)} \leq w_{\max}\left(\rho \tilde{\lambda}_1^{(j^*)} + 4\sqrt{\log k}\right) \leq 4 w_{\max}(1 + \rho)\sqrt{\log k},$$

where in the last inequality, we also applied upper bound on $\tilde{\lambda}_1^{(j^*)}$. Combining all above bounds and Lemma 20 finishes the proof. □

[12] Note that if A and B are orthogonal matrices, columns of A and B are directly recovered by computing SVD of $T(I, I, \theta)$.

C.1 Auxiliary lemmata for initialization

In the following Lemma, we show that the gap condition between the maximum and the second maximum of vector λ required in Lemma 16 is satisfied under some number of random draws.

Lemma 15 (Gap condition)**.** *Consider an arbitrary matrix $C \in \mathbb{R}^{d\times k}$ with unit-norm columns which also satisfies incoherence condition $\max_{i\neq j} |\langle c_i, c_j\rangle| \leq \rho$ for some $\rho > 0$. Let*

$$\lambda := \mathrm{Diag}(w)C^\top\theta \in \mathbb{R}^k,$$

denote the vector that captures correlation of $\theta \in \mathbb{R}^d$ with columns of C. Without loss of generality, assume that $\lambda_1 = \max_i |\lambda_i|$, and let $\lambda_{(2)} := \max_{i\neq 1} |\lambda_i|$. Draw L i.i.d. random vectors $\theta^{(j)} \sim \mathcal{N}(0, I_d), j \in [L]$, and $\lambda^{(j)} := \mathrm{Diag}(w)C^\top\theta^{(j)}$. Suppose L satisfies the bound

$$\sqrt{\frac{\ln(L)}{8\ln(k)}}\left(1 - \frac{\ln(\ln(L)) + c}{4\ln(L)} - \sqrt{\frac{\ln(k)}{\ln(L)}}\right) \geq \frac{w_{\max}(1+\mu)}{w_{\min} - \rho w_{\max}(1+\mu)},$$

for some $0 < \mu < \frac{w_{\min}}{w_{\max}\rho} - 1$. Then, with probability at least $1 - 2k^{-1} - k^{-7}$, we have the following gap condition for at least one draw, say j^*,

$$\lambda_1^{(j^*)} \geq (1+\mu)\lambda_{(2)}^{(j^*)}.$$

Proof: Define $\tilde{\lambda} := \mathrm{Diag}(w)^{-1}\lambda = C^\top\theta$. We have $\lambda_j = w_j\tilde{\lambda}_j, j \in [k]$.
Each vector $\tilde{\lambda}^{(j)}$ is a random Gaussian vector $\tilde{\lambda}^{(j)} \sim \mathcal{N}(0, C^\top C)$. Let $j^* := \arg\max_{j\in[L]} \tilde{\lambda}_1^{(j)}$. Since $\max_{j\in[L]} \tilde{\lambda}_1^{(j)}$, is a 1-Lipschitz function of L independent $\mathcal{N}(0,1)$ random variables, similar to the analysis in Lemma B.1 of Anandkumar et al. (2014a), we have

$$\Pr\left[\tilde{\lambda}_1^{(j^*)} \geq \sqrt{2\ln(L)} - \frac{\ln(\ln(L)) + c}{2\sqrt{2\ln(L)}} - \sqrt{2\ln(k)}\right] \geq 1 - \frac{2}{k}. \tag{29}$$

Any vector $c_i, i \neq 1$, can be decomposed to two components parallel and perpendicular to c_1 as $c_i = \langle c_i, c_1\rangle c_1 + \mathcal{P}_{\perp_{c_1}}(c_i)$. Then, for any $\tilde{\lambda}_i, i \neq 1$, we have

$$\tilde{\lambda}_i := \langle\theta, c_i\rangle = \underbrace{\theta^\top\langle c_i, c_1\rangle c_1}_{=:\tilde{\lambda}_{i,\parallel}} + \underbrace{\theta^\top\mathcal{P}_{\perp_{c_1}}(c_i)}_{=:\tilde{\lambda}_{i,\perp}}.$$

Since $\mathcal{P}_{\perp_{c_1}}(c_i) \perp c_1, i \neq 1$, we have $\tilde{\lambda}_{i,\perp}, i \neq 1$, are independent of $\tilde{\lambda}_1 := \theta^\top c_1$, and therefore, the following bound can be argued independent of bound in (29). From Lemma 18, we have

$$\Pr\left[\max_{i\neq 1} \tilde{\lambda}_{i,\perp}^{(j^*)} \geq 4\sqrt{\log k}\right] \leq k^{-7}.$$

For $\tilde{\lambda}_{i,\parallel}$, we have

$$\tilde{\lambda}_{i,\parallel} = \theta^\top\langle c_i, c_1\rangle c_1 \leq \rho\theta^\top c_1 = \rho\tilde{\lambda}_1,$$

where we also assumed that $\tilde{\lambda}_1 := \theta^\top c_1 > 0$ which is true for large enough L, concluded from (29). By combining above two bounds, with probability at least $1 - k^{-7}$, we have

$$\tilde{\lambda}_{(2)}^{(j^*)} \leq \rho\tilde{\lambda}_1 + 4\sqrt{\log k}. \tag{30}$$

From the given bound on L in the lemma and inequalities (29) and (30), with probability at least $1 - 2k^{-1} - k^{-7}$, we have

$$\tilde{\lambda}_1^{(j^*)} \geq \frac{w_{\max}(1+\mu)}{w_{\min} - \rho w_{\max}(1+\mu)} \left(\tilde{\lambda}_{(2)}^{(j^*)} - \rho \tilde{\lambda}_1^{(j^*)} \right).$$

Simple calculations imply that

$$w_{\min} \tilde{\lambda}_1^{(j^*)} \geq (1+\mu) w_{\max} \tilde{\lambda}_{(2)}^{(j^*)}.$$

Incorporating inequalities $\lambda_1 \geq w_{\min}\tilde{\lambda}_1$ and $\lambda_{(2)} \leq w_{\max}\tilde{\lambda}_{(2)}$ finishes the proof saying that the result of lemma is valid for the j^*-th draw. □

In the following lemma, we show that if a vector $\theta \in \mathbb{R}^d$ is relatively more correlated with c_1 (comparing to $c_i, i \neq 1$), then dominant singular vectors of $\hat{T}(I, I, \theta)$ provide good initialization vectors for a_1 and b_1.

Before proposing the lemma, we define

$$\mu_E := \alpha \sqrt{\frac{k}{d}} \left(2 + 2\alpha_0 \sqrt{\frac{k}{d}} + \frac{\alpha}{\sqrt{d}} \right), \quad \mu_R := \left(1 + \alpha_0 \sqrt{\frac{k}{d}} \right)^2, \quad \mu_{\min} := \min\{\mu_E, \mu_R\}. \tag{31}$$

where $\alpha = \text{polylog}(d)$, and $\alpha_0 > 0$ is a constant.

Lemma 16. *Consider $\hat{T} = T + \Psi$, where T is a rank-k tensor, and Ψ is a perturbation tensor. Let assumptions (A1)-(A3) hold for T. Let u_1 and v_1 be the top left and right singular vectors of $\hat{T}(I, I, \theta)$. Let*

$$\lambda := \text{Diag}(w) C^\top \theta \in \mathbb{R}^k,$$

denote the vector that captures correlation of θ with different $c_i, i \in [k]$, weighted by $w_i, i \in [k]$. Without loss of generality, assume that $\lambda_1 = \max_i |\lambda_i|$, and let $\lambda_{(2)} := \max_{i \neq 1} |\lambda_i|$. Suppose the relative gap condition

$$\lambda_1 \geq (1+\mu)\lambda_{(2)}, \tag{32}$$

is satisfied for some $\mu > \frac{\lambda_1}{\lambda_1 - \|\Psi(I,I,\theta)\|} 2\mu_R - 1$, where μ_R and $\mu_{\min}$ are defined in (31). Then, with high probability (w.h.p.),

$$\max\{\text{dist}(u_1, a_1), \text{dist}(v_1, b_1)\} \leq \frac{\mu_{\min}\lambda_{(2)} + \|\Psi(I,I,\theta)\|}{\tilde{\mu}\lambda_1 - \|\Psi(I,I,\theta)\|},$$

for $\|\Psi(I,I,\theta)\|/\lambda_1 < \tilde{\mu} < 1$ defined as

$$\tilde{\mu} := \frac{1 + \mu - 2\mu_R}{1 + \mu}.$$

Proof: From Assumption (A1), $T(I, I, \theta)$ can be written as equation (28), Expanded as

$$T(I, I, \theta) = \lambda_1 a_1 b_1^\top + \underbrace{\sum_{i \neq 1} \lambda_i a_i b_i^\top}_{=:R}.$$

From here, we prove the result in two cases. First when $\mu_E < \mu_R$ and therefore $\mu_{\min} = \mu_E$, and second when $\mu_E \geq \mu_R$ and therefore $\mu_{\min} = \mu_R$.

Case 1 ($\mu_E < \mu_R$): According to the subspaces spanned by a_1 and b_1, we decompose matrix R to two components as $R = \mathcal{P}_\perp(R) + \mathcal{P}_\parallel(R)$. First term $\mathcal{P}_\perp(R)$ is the component with column space orthogonal to a_1 and row space orthogonal to b_1, and $\mathcal{P}_\parallel(R)$ is the component with either the column space equal to a_1 or the row space equal to b_1. We have

$$\begin{aligned}\mathcal{P}_\perp(R) &= (I - P_{a_1})R(I - P_{b_1}),\\ \mathcal{P}_\parallel(R) &= P_{a_1}R + RP_{b_1} - P_{a_1}RP_{b_1},\end{aligned}$$

where $P_{a_1} = a_1 a_1^\top$ is the projection operator on the subspace in $\mathbb{R}^d$ spanned by a_1, and similarly $P_{b_1} = b_1 b_1^\top$ is the projection operator on the subspace in $\mathbb{R}^d$ spanned by b_1. Thus, for $\hat{T} = T + \Psi$, we have

$$\hat{T}(I, I, \theta) = \underbrace{\lambda_1 a_1 b_1^\top + \mathcal{P}_\perp(R)}_{=:M} + \underbrace{\mathcal{P}_\parallel(R)}_{=:E} + \Psi(I, I, \theta).$$

Looking at M, it becomes more clear why we proposed the above decomposition for R. Since the column and row space of $\mathcal{P}_\perp(R)$ are orthogonal to a_1 and b_1, respectively, the SVD of M has a_1 and b_1 as its left and right singular vectors, respectively. Hence, M has the SVD form

$$M = [a_1\ \tilde{U}_2] \begin{bmatrix} \lambda_1 & 0 \\ 0 & \tilde{\Sigma}_2 \end{bmatrix} [b_1\ \tilde{V}_2]^\top,$$

where $\mathcal{P}_\perp(R) = \tilde{U}_2 \tilde{\Sigma}_2 \tilde{V}_2^\top$ is the SVD of $\mathcal{P}_\perp(R)$. Let $\tilde{\sigma}_2 := \max_i(\tilde{\Sigma}_2)_{ii}$. From gap condition (32) assumed in the lemma and inequality (33), we have $\lambda_1 \geq \tilde{\sigma}_2$, and therefore, a_1 and b_1 are the top left and right singular vectors of M. On the other hand, $\hat{T}(I, I, \theta)$ has the corresponding SVD form

$$\hat{T}(I, I, \theta) = [u_1\ U_2] \begin{bmatrix} \sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix} [v_1\ V_2]^\top,$$

where u_1 and v_1 are its top left and right singular vectors. We have

$$\begin{aligned}\tilde{\sigma}_2 = \|\mathcal{P}_\perp(R)\| &\leq \|R\| \\ &= \left\| \sum_{i=2}^{k} \lambda_i a_i b_i^\top \right\| \\ &\leq \lambda_{(2)} \left\| A_{\setminus 1} \right\| \left\| B_{\setminus 1}^\top \right\| \\ &\leq \lambda_{(2)} \|A\| \left\| B^\top \right\| \\ &\leq \left(1 + \alpha_0 \sqrt{\frac{k}{d}}\right)^2 \lambda_{(2)} =: \mu_R \lambda_{(2)}, \end{aligned} \tag{33}$$

where the sub-multiplicative property of spectral norm is used in the second inequality, and the last inequality is from Assumption (A3). From Weyl's theorem, we have

$$\begin{aligned}|\sigma_1 - \lambda_1| &\leq \|E\| + \|\Psi(I, I, \theta)\| \\ &\leq \lambda_{(2)} \alpha \sqrt{\frac{k}{d}} \left(2 + 2\alpha_0 \sqrt{\frac{k}{d}} + \frac{\alpha}{\sqrt{d}}\right) + \|\Psi(I, I, \theta)\| \\ &=: \mu_E \lambda_{(2)} + \|\Psi(I, I, \theta)\|, \end{aligned} \tag{34}$$

where (35) is used in the second inequality. Therefore, we have

$$\begin{aligned}\sigma_1 - \tilde{\sigma}_2 &= \sigma_1 - \lambda_1 + \lambda_1 - \tilde{\sigma}_2 \\ &\geq -\mu_E \lambda_{(2)} - \|\Psi(I,I,\theta)\| + \lambda_1 - \mu_R \lambda_{(2)} \\ &\geq \left(1 - \frac{\mu_E + \mu_R}{1+\mu}\right)\lambda_1 - \|\Psi(I,I,\theta)\|, \\ &=: \tilde{\mu}_1 \lambda_1 - \|\Psi(I,I,\theta)\| =: \nu,\end{aligned}$$

where bounds (33) and (34) are used in the first inequality, and the second inequality is concluded from the gap condition (32) assumed in the lemma. Therefore, since $\sigma_1 \geq \beta + \nu$ and $\tilde{\sigma}_2 \leq \beta$ for some $\beta > 0$, Wedin's theorem is applied to the equality $\hat{T}(I,I,\theta) = M + E + \Psi(I,I,\theta)$, which implies that

$$\begin{aligned}\max\left\{\sqrt{1-\langle u_1, a_1\rangle^2}, \sqrt{1-\langle v_1, b_1\rangle^2}\right\} &\leq \frac{\|E + \Psi(I,I,\theta)\|}{\nu} \\ &\leq \frac{\mu_E \lambda_{(2)} + \|\Psi(I,I,\theta)\|}{\tilde{\mu}_1 \lambda_1 - \|\Psi(I,I,\theta)\|} \\ &\leq \frac{\mu_{\min} \lambda_{(2)} + \|\Psi(I,I,\theta)\|}{\tilde{\mu} \lambda_1 - \|\Psi(I,I,\theta)\|},\end{aligned}$$

where we used $\mu_{\min} = \mu_E$ and $\tilde{\mu}_1 > \tilde{\mu}$ in the last inequality when $\mu_E < \mu_R$. Since $\mathrm{dist}^2(u_1, a_1) + \langle u_1, a_1\rangle^2 = 1$, the proof is complete for this case.

Bounding the spectral norm of E: For any $i \neq j$, let $\rho_{ij}^{(a)} := |\langle a_i, a_j\rangle|$ and $\rho_{ij}^{(b)} := |\langle b_i, b_j\rangle|$. We have

$$\begin{aligned}E := \mathcal{P}_{\|}(R) &= P_{a_1} R + R P_{b_1} - P_{a_1} R P_{b_1}, \\ &= a_1 a_1^\top R + R b_1 b_1^\top - a_1 a_1^\top R b_1 b_1^\top \\ &= \sum_{i\neq 1} \lambda_i a_1 a_1^\top a_i b_i^\top + \sum_{i\neq 1} \lambda_i a_i b_i^\top b_1 b_1^\top - \sum_{i\neq 1} \lambda_i a_1 a_1^\top a_i b_i^\top b_1 b_1^\top \\ &= \sum_{i\neq 1} \lambda_i \rho_{1i}^{(a)} a_1 b_i^\top + \sum_{i\neq 1} \lambda_i \rho_{1i}^{(b)} a_i b_1^\top - \sum_{i\neq 1} \lambda_i \rho_{1i}^{(a)} \rho_{1i}^{(b)} a_1 b_1^\top \\ &= \underbrace{A_{(1)} \operatorname{Diag}(\lambda_{(a)}) B_{\setminus 1}^\top}_{E_1} + \underbrace{A_{\setminus 1} \operatorname{Diag}(\lambda_{(b)}) B_{(1)}^\top}_{E_2} - \underbrace{A_{(1)} \operatorname{Diag}(\lambda_{(a,b)}) B_{(1)}^\top}_{E_3},\end{aligned}$$

where $A_{(1)} := \left[\overbrace{a_1|a_1|\cdots|a_1}^{k-1 \text{ times}}\right] \in \mathbb{R}^{d\times(k-1)}$, $B_{\setminus 1} := [b_2|b_3|\cdots|b_k] \in \mathbb{R}^{d\times(k-1)}$, and $\lambda_{(a)} := [\lambda_i \rho_{1i}^{(a)}]_{i\neq 1} \in \mathbb{R}^{k-1}$. The other notations are similarly defined.

For E_1, we have

$$\begin{aligned}\|E_1\| &\leq \|A_{(1)} \operatorname{Diag}(\lambda_{(a)})\| \|B_{\setminus 1}^\top\| \\ &= \|\lambda_{(a)}\| \|a_1\| \|B_{\setminus 1}^\top\| \\ &\leq \sqrt{k} \lambda_{(2)} \rho \|B^\top\| \\ &\leq \lambda_{(2)} \alpha \sqrt{\frac{k}{d}} \left(1 + \alpha_0 \sqrt{\frac{k}{d}}\right).\end{aligned}$$

Where the first equality is concluded from Lemma 19, and Assumptions (A2) and (A3) are exploited in the last inequality. Similarly, for E_2 and E_3, we have

$$\|E_2\| \leq \lambda_{(2)}\alpha\sqrt{\frac{k}{d}}\left(1+\alpha_0\sqrt{\frac{k}{d}}\right),$$
$$\|E_3\| \leq \lambda_{(2)}\alpha^2\frac{\sqrt{k}}{d}.$$

Therefore, we have

$$\|E\| \leq \lambda_{(2)}\alpha\sqrt{\frac{k}{d}}\left(2+2\alpha_0\sqrt{\frac{k}{d}}+\frac{\alpha}{\sqrt{d}}\right). \tag{35}$$

Case 2 ($\mu_R \leq \mu_E$): The result can be similarly achieved when $\mu_R \leq \mu_E$. Here we directly apply Wedin's theorem to $\hat{T}(I,I,\theta) = \lambda_1 a_1 b_1^\top + R + \Psi(I,I,\theta)$, treating $R+\Psi(I,I,\theta)$ as the error term. From Weyl's theorem, we have

$$\sigma_1 \geq \lambda_1 - \|R\| - \|\Psi(I,I,\theta)\| \geq \underbrace{\left(1-\frac{\mu_R}{1+\mu}\right)}_{=:\tilde{\mu}_2}\lambda_1 - \|\Psi(I,I,\theta)\|,$$

where (33) and gap condition (32) are used in the second inequality. Since $\tilde{\sigma}_2 = 0$, by Wedin's theorem, we have

$$\begin{aligned}\max\left\{\sqrt{1-\langle u_1,a_1\rangle^2}, \sqrt{1-\langle v_1,b_1\rangle^2}\right\} &\leq \frac{\mu_R\lambda_{(2)} + \|\Psi(I,I,\theta)\|}{\tilde{\mu}_2\lambda_1 - \|\Psi(I,I,\theta)\|} \\ &\leq \frac{\mu_{\min}\lambda_{(2)} + \|\Psi(I,I,\theta)\|}{\tilde{\mu}\lambda_1 - \|\Psi(I,I,\theta)\|},\end{aligned}$$

where we used $\mu_{\min} = \mu_R$ and $\tilde{\mu}_2 \geq \tilde{\mu}$ in the last inequality when $\mu_R \leq \mu_E$. Since $\mathrm{dist}^2(u_1,a_1) + \langle u_1,a_1\rangle^2 = 1$, the proof is complete for this case. □

The above lemma concludes the proof for initialization procedure, except for a few auxiliary lemmata that we prove next.

First we use Gaussian tail bounds to prove that the largest entry of a Gaussian vector can be quite large with inverse polynomial probability:

Lemma 17. *Let $x \sim \mathcal{N}(0,\sigma)$ be a Gaussian random variable with mean zero and variance σ^2. Then, for any $t > 0$, we have*

$$\left(\frac{\sigma}{t} - \frac{\sigma^3}{t^3}\right) f(t/\sigma) \leq \Pr[x \geq t] \leq \frac{\sigma}{t} f(t/\sigma),$$

where $f(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$.

Proof: Let $z = \frac{x}{\sigma}$, where $z \sim \mathcal{N}(0,1)$ is a standard Gaussian random variable. Then, we have $\Pr[x \geq t] = \Pr[z \geq t/\sigma]$, and therefore, the result is proved by using standard tail bounds for Gaussian random variable. □

Lemma 18. *Consider $r = [r_1, r_2, \ldots, r_k]^\top \in \mathbb{R}^k$ as a k-dimensional random Gaussian vector with zero mean and covariance Σ, i.e., $r \sim \mathcal{N}(0, \Sigma)$. For any $k \geq 2$, we have*

$$\Pr\left[r_{(1)} \geq 4\sigma_{\max}\sqrt{\log k}\right] \leq k^{-7}.$$

Proof: From Lemma 17, for any $i \in [k]$, we have

$$\Pr\left[|r_i| \geq 4\sigma_{\max}\sqrt{\log k}\right] \leq \frac{1}{2\sqrt{2\pi \log k}} k^{-8} \leq k^{-8},$$

where the last inequality is concluded from the fact that $k \geq 2$. The result is then proved by taking a union bound. □

Next we prove a basic fact about spectral norm that is used in the proof of Lemma 16.

Lemma 19. *Given $h \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$, let $H = [h|h|\cdots|h]\,\mathrm{Diag}(v) \in \mathbb{R}^{m\times n}$. Then, $\|H\| = \|h\|\|v\|$.*

Proof: By definition

$$\|H\| = \sup_{\|x\|=1} \|Hx\|.$$

We have $Hx = \langle v, x\rangle h$, and therefore, $\|Hx\| = |\langle v, x\rangle|\|h\|$. This is maximized by $x = v/\|v\|$, and this finishes the proof. □

Finally, we show that noise matrix $\Psi(I, I, \theta)$ has bounded norm with high probability which is useful for initialization argument in Theorem 3.

Lemma 20. *Let $\theta \in \mathbb{R}^d$ be standard multivariate Gaussian as $\mathcal{N}(0, I_d)$. Then, for any $\alpha_0 > 1$, we have*

$$\Pr\left[\|\Psi(I, I, \theta)\| \leq \alpha_0\sqrt{d}\psi\right] \geq 1 - e^{-(\alpha_0-1)^2 d/2},$$

where $\psi := \|\Psi\|$ is the spectral norm of error tensor Ψ.

Proof: Let $\theta_n := \frac{1}{\|\theta\|}\theta$ denote the normalized version of θ. Then, we have

$$\|\Psi(I, I, \theta)\| = \|\theta\| \cdot \|\Psi(I, I, \theta_n)\| \leq \|\theta\|\psi,$$

where the last inequality is from the definition of tensor spectral norm. Applying the bound on $\|\theta\|$ in Lemma 21 finishes the proof. □

The following lemma provides concentration bound for the norm of standard Gaussian vector which is basically a tail bound for the chi-squared random variable.

Lemma 21 (Lemma 15 of Dasgupta et al. (2006))**.** *Let the random vector θ is distributed as $\mathcal{N}(0, I_d)$. Then, for any $\alpha_0 > 1$, we have*

$$\Pr\left[\|\theta\| \geq \alpha_0\sqrt{d}\right] \leq e^{-(\alpha_0-1)^2 d/2}.$$

D Clustering Process

In the last step of main algorithm, we need to cluster the generated 4-tuples into k clusters. Theoretically, we only have convergence guarantees when the initialization vectors are good enough, while the other initializations can potentially generate arbitrary 4-tuples. In the worst case, these arbitrary 4-tuples can make the clustering process hard, and therefore, we provide specific Procedure 3 for which the output properties are provided in Lemma 24.

Note that the key observation for the algorithm is if $T(\hat{a}, \hat{b}, \hat{c})$ is large for some $(\hat{a}, \hat{b}, \hat{c})$, then these vectors are close to (a_i, b_i, c_i) for some $i \in [k]$.

For simplicity, we only prove this when the initialization procedure in Theorem 2 takes polynomial time, namely $k = O(d)$ and $w_{\max}/w_{\min} = O(1)$. Without loss of generality, we also assume $w_{\max} = w_1 \geq w_2 \geq \cdots \geq w_k = w_{\min}$. In this case, we choose the threshold ϵ in the following lemmata to be some small constant depending on k/d and $w_{\max}/w_{\min}$. Also, we work in the case when noise $\Psi = 0$, however the proof still works when the noise $\psi = \|\Psi\| = o(1)$.

Lemma 22. *Suppose*

$$\max\{|\langle a_i, \hat{a}\rangle|, |\langle b_i, \hat{b}\rangle|, |\langle c_i, \hat{c}\rangle|\} \leq \epsilon, \quad \forall i \in [t-1],$$

for some $t \in [k]$. Let $\delta := O\left(\frac{w_{\max}}{w_{\min}}\epsilon^{3-p}\right)$, and assume $|T(\hat{a}, \hat{b}, \hat{c})| \geq (1-\delta)w_t$. Then, there exists some j such that

$$\max\{\mathrm{dist}(\hat{a}, a_j), \mathrm{dist}(\hat{b}, b_j), \mathrm{dist}(\hat{c}, c_j)\} < \frac{w_{\min}}{10 w_{\max}}.$$

Proof: Partition tensor $T = \sum_{i\in[k]} w_i a_i \otimes b_i \otimes c_i$ to $T_1 + T_2$, where T_1 contains all the terms indexed from 1 to $t-1$, and T_2 contains the remaining terms. From Corollary 3, we have

$$|T_1(\hat{a}, \hat{b}, \hat{c})| \leq w_{\max} \left\|A_{[t-1]}^\top \hat{a}\right\|_3 \cdot \left\|B_{[t-1]}^\top \hat{b}\right\|_3 \cdot \left\|C_{[t-1]}^\top \hat{c}\right\|_3,$$

where $A_{[t-1]} \in \mathbb{R}^{d\times(t-1)}$ denotes the first $t-1$ columns of A, and similarly for $B_{[t-1]}$ and $C_{[t-1]}$. We also have

$$\left\|A_{[t-1]}^\top \hat{a}\right\|_3^3 \leq \left\|A_{[t-1]}^\top \hat{a}\right\|_p^p \cdot \max_{i\in[t-1]} |\langle a_i, \hat{a}\rangle|^{3-p} = O\left(\epsilon^{3-p}\right),$$

where Assumption (A10) and the assumption in the lemma are exploited in the last step. Similar arguments hold for b and c. Combining with the earliest inequality, we have

$$|T_1(\hat{a}, \hat{b}, \hat{c})| \leq w_{\max} O\left(\epsilon^{3-p}\right) \leq w_t \delta,$$

where the definition of δ is exploited in the last inequality. Applying assumption $|T(\hat{a}, \hat{b}, \hat{c})| \geq (1-\delta)w_t$ to the above bound, we have

$$|T_2(\hat{a}, \hat{b}, \hat{c})| \geq (1-2\delta)w_t. \tag{36}$$

On the other hand, from Corollary 3,

$$|T_2(\hat{a}, \hat{b}, \hat{c})| \leq w_t \|A^\top \hat{a}\|_3 \|B^\top \hat{b}\|_3 \|C^\top \hat{c}\|_3.$$

Since all the 3-norms are bounded by $1 + o(1)$, each of them must be at least $1 - O(\delta)$ to let inequality (36) hold. Now we have

$$1 - O(\delta) \leq \sum_{j=1}^{k} |\langle a_j, \hat{a}\rangle|^3 \leq \max\{|\langle a_j, \hat{a}\rangle|\}^{3-p} \sum_{t=1}^{k} |\langle a_j, \hat{a}\rangle|^p \leq (1 + o(1)) \max\{|\langle a_j, \hat{a}\rangle|\}^{3-p},$$

where the last inequality is from Assumption (A10). This implies $\max\{|\langle a_j, \hat{a}\rangle|\} = 1 - O(\delta)$, which in turn implies there exists a j such that

$$\text{dist}(\hat{a}, a_j) < w_{\min}/10w_{\max}$$

when ϵ and δ are small enough.

By symmetry we know there is also a j' such that $\text{dist}(\hat{b}, b_{j'}) < w_{\min}/10w_{\max}$. If $j \neq j'$, then it is easy to check $T_2(\hat{a}, \hat{b}, \hat{c})$ cannot be large. Hence, $j = j'$ and the Lemma is correct. □

On the other hand, we know if there is a good initialization, the largest $T(\hat{a}, \hat{b}, \hat{c})$ must be large.

Lemma 23. *Suppose there exists a good initialization (see initialization condition* (13) *in the local convergence theorem) for some column $t \in [k]$, and*

$$\max\{|\langle a_i, \hat{a}^{(0)}\rangle|, |\langle b_i, \hat{b}^{(0)}\rangle|, |\langle c_i, \hat{c}^{(0)}\rangle|\} \leq \epsilon, \quad \forall i \neq t.$$

Let $\delta := O\left(\frac{w_{\max}}{w_{\min}}\epsilon^{3-p}\right)$. Then the corresponding output of iterations in Algorithm 1 denoted by $(\hat{a}, \hat{b}, \hat{c})$ satisfy

$$|T(\hat{a}, \hat{b}, \hat{c})| > (1 - \delta)w_t.$$

Furthermore, for any $i \neq t$, $\max\{|\langle \hat{a}, a_i\rangle|, |\langle \hat{b}, b_i\rangle|, |\langle \hat{c}, c_i\rangle|\} \leq o(\epsilon)$.

Proof: Similar to the proof of Lemma 22, partition tensor $T = \sum_{i\in[k]} w_i a_i \otimes b_i \otimes c_i$ to $T_2 = w_t a_t \otimes b_t \otimes c_t$ and $T_1 = T - T_2$. Since the initialization is good, by the local convergence result in Theorem 1, we have

$$\text{dist}(\hat{a}, a_t) \leq \tilde{O}\left(\frac{w_{\max}}{w_{\min}}\frac{\sqrt{k}}{d}\right) \leq o(\delta),$$

where the incoherence condition and $p > 2$ are exploited in the last step. Therefore, $|T_2(\hat{a}, \hat{b}, \hat{c})| \geq (1 - \delta/2)w_t$.

Similar to Lemma 22, by using Corollary 3, we have $|T_1(\hat{a}, \hat{b}, \hat{c})| \leq w_t\delta/2$. Applying these bounds, we have

$$|T(\hat{a}, \hat{b}, \hat{c})| \geq |T_2(\hat{a}, \hat{b}, \hat{c})| - |T_1(\hat{a}, \hat{b}, \hat{c})| \geq (1 - \delta)w_t.$$

The last part of the Lemma is trivial because $\text{dist}(\hat{a}, a_t)$ is small and $\langle a_i, a_t\rangle$ is small by incoherence. □

Finally we prove the clustering process succeeds.

Lemma 24. *Procedure 3 outputs k cluster centers that are $\tilde{O}\left(\frac{w_{\max}}{w_{\min}}\frac{\sqrt{k}}{d}\right)$ close to the true components of the tensor.*

Proof: We prove by induction to show that every step of the algorithm correctly computes one component.

Suppose all previously found 4-tuples are $\tilde{O}(w_{\max}\sqrt{k}/w_{\min}d)$ close to some (a_i, b_i, c_i) (notice that this is true at the beginning when no components are found). Let t be the smallest index that has not been found. Then all the remaining 4-tuples satisfy

$$\max\{|\langle a_i, \hat{a}\rangle|, |\langle b_i, \hat{b}\rangle|, |\langle c_i, \hat{c}\rangle|\} \leq \epsilon, \quad \forall i < t.$$

By Lemma 23 we know there must be a 4-tuple with $|T(\hat{a}, \hat{b}, \hat{c})| > w_t(1-\delta)$. On the other hand, by Lemma 22 we know the 4-tuple we found must satisfy $\max\{\text{dist}(\hat{a}, a_j), \text{dist}(\hat{b}, b_j)\} < w_{\min}/10w_{\max}$ for some j (and this cannot be some j that has already been found). This tuple then satisfies the conditions of the local convergence Theorem 1. Hence, after N iterations it must have converged to (a_j, b_j, c_j). At this step the algorithm successfully found a new component of the tensor.

□

References

Radosław Adamczak, Rafał Latała, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Chevet type inequality and norms of submatrices. *arXiv preprint arXiv:1107.4066*, 2011.

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.

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 2013a.

A. Anandkumar, D. Hsu, M. Janzamin, and S. M. Kakade. When are Overcomplete Topic Models Identifiable? Uniqueness of Tensor Tucker Decompositions with Structured Sparsity. In *Neural Information Processing (NIPS)*, Dec. 2013b.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor Methods for Learning Latent Variable Models. *J. of Machine Learning Research*, 15:2773–2832, 2014a.

Anima Anandkumar, Rong Ge, and Majid Janzamin. Sample Complexity Analysis for Learning Overcomplete Latent Variable Models through Tensor Methods. *arXiv preprint arXiv:1408.0553*, Aug. 2014b.

S. Arora, R. Ge, and A. Moitra. New Algorithms for Learning Incoherent and Overcomplete Dictionaries. *ArXiv e-prints*, August 2013.

Y. Bengio, A. Courville, and P. Vincent. Unsupervised feature learning and deep learning: A review and new perspectives. *arXiv preprint arXiv:1206.5538*, 2012.

A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. *arXiv preprint arXiv:1311.3651*, 2013.

J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika*, 35(3):283–319, 1970.

P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

P. Comon. Tensor decompositions. *Mathematics in Signal Processing V*, pages 1–24, 2002.

P. Comon, X. Luciani, and A. De Almeida. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics*, 23(7-8):393–405, 2009.

Sanjoy Dasgupta, Daniel Hsu, and Nakul Verma. A concentration theorem for projections. In *Twenty-Second Conference on Uncertainty in Artificial Intelligence*, 2006.

L. De Lathauwer, J. Castaing, and J.-F. Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *Signal Processing, IEEE Transactions on*, 55(6):2965–2973, 2007.

D. Donoho. Compressed sensing. *Information Theory, IEEE Transactions on*, 52(4):1289–1306, 2006.

N. Goyal, S. Vempala, and Y. Xiao. Fourier pca. *arXiv preprint arXiv:1306.5825*, 2013.

Olivier Guédon and Mark Rudelson. Lp-moments of random vectors via majorizing measures. *Advances in Mathematics*, 208(2):798–823, 2007.

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

Richard A Harshman. Foundations of the parafac procedure: models and conditions for an" explanatory" multimodal factor analysis. 1970.

Richard A Harshman and Margaret E Lundy. Parafac: Parallel factor analysis. *Computational Statistics & Data Analysis*, 18(1):39–72, 1994.

Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are NP hard. *arXiv preprint arXiv:0911.1393*, 2009.

F. Huang, U. N. Niranjan, M. Hakeem, and A. Anandkumar. Fast Detection of Overlapping Communities via Online Tensor Methods. *ArXiv 1309.0787*, Sept. 2013.

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

T. Kolda. Orthogonal tensor decompositions. *SIAM Journal on Matrix Analysis and Applications*, 23(1):243–255, 2001.

T. Kolda and B. Bader. Tensor decompositions and applications. *SIREV*, 51(3):455–500, 2009.

T. G. Kolda and J. R. Mayo. Shifted power method for computing tensor eigenpairs. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1095–1124, October 2011.

J.B. Kruskal. More factors than subjects, tests and treatments: an indeterminacy theorem for canonical decomposition and individual differences scaling. *Psychometrika*, 41(3):281–293, 1976.

J.B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.

Q. V. Le, A. Karpenko, J. Ngiam, and A. Y. Ng. ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning. In *NIPS*, pages 1017–1025, 2011.

M. S. Lewicki and T. J. Sejnowski. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

B. McWilliams, D. Balduzzi, and J. Buhmann. Correlated random features for fast semi-supervised learning. In *Advances in Neural Information Processing Systems*, pages 440–448, 2013.

Praneeth Netrapalli, Prateek Jain, and Sujay Sanghavi. Phase retrieval using alternating minimization. *arXiv preprint arXiv:1306.0160*, 2013.

L. Song, A. Anandkumar, B. Dai, and B. Xie. Nonparametric estimation of multi-view latent variable models. *Available on arXiv:1311.3287*, Nov. 2013.

M. A. O. Vasilescu and D. Terzopoulos. Multilinear subspace analysis of image ensembles. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–93. IEEE, 2003.

Eugene P Wigner. Characteristic vectors of bordered matrices with infinite dimensions. *The Annals of Mathematics*, 62(3):548–564, 1955.

T. Zhang and G. Golub. Rank-one approximation to high order tensors. *SIAM Journal on Matrix Analysis and Applications*, 23:534–550, 2001.

J. Y. Zou, D. Hsu, D. C. Parkes, and R. P. Adams. Contrastive learning using spectral methods. In *Advances in Neural Information Processing Systems*, pages 2238–2246, 2013.