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

Anima Anandkumar[*] Rong Ge[†] Majid Janzamin[‡]

March 12, 2014

Abstract

A simple alternating rank-1 update procedure is considered for CP tensor decomposition. Local convergence guarantees are established for third order tensors of rank k in d dimensions, when $k = o(d^{1.5})$ and the tensor components are incoherent. We strengthen the results to global convergence guarantees when $k = O(d)$ through a simple initialization procedure based on rank-1 singular value decomposition of random tensor slices. Our tight perturbation analysis leads to efficient sample guarantees for unsupervised learning of discrete multi-view mixtures when $k = O(d)$, where k is the number of mixture components and d is the observed dimension. For learning overcomplete decompositions ($k = \omega(d)$), we prove that having an extremely small number of labeled samples, scaling as polylog(k) for each label, under the semi-supervised setting (where the label corresponds to the choice variable in the mixture model) leads to global convergence guarantees for learning mixture models.

Keywords: Tensor decomposition, alternating minimization, unsupervised and semi-supervised learning, 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., 2012a), network community models (Anandkumar et al., 2013b), and so on. The decomposition of a certain low order multivariate moment tensor (typically up to fourth order) in these models are guaranteed to provide a consistent estimate of the model parameters. Moreover, the sample and computational requirements are only a low order polynomial in the latent dimensionality for the tensor method (Anandkumar et al., 2012a; 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), where 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 current state of art for guaranteed tensor decomposition involves two steps, viz., 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., 2012a). 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* 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, which involves making rank-1 updates along different modes of the tensor. 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.

1.1 Summary of Results

In this paper, we analyze alternating rank-1 updates for CP tensor decomposition. This involves maintaining a rank-1 estimate of the tensor. 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.

We provide local convergence guarantees under incoherent tensor components 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 final estimate. The approximation error depends on the extent of overcompleteness, and scales as[1]

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

$\tilde{O}(\sqrt{k}/d)$, which is small since $k = o(d^{3/2})$.

In the undercomplete or mildly overcomplete settings ($k = O(d)$), a simple initialization procedure 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 k). This leads to global convergence guarantees for tensor decomposition: the algorithm returns a tensor whose components are $\tilde{O}(\sqrt{k}/d)$ close to the correct tensor. To the best of our knowledge, these are first guarantees for tensor decomposition under incoherent tensor components.

We then extend the global convergence guarantees to settings where two modes of the tensor are (sufficiently) undercomplete, and the third tensor mode is (highly) overcomplete. 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. (2012a) which rely on transforming the input tensor to an orthogonal symmetric form cannot handle this setting. 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.

Our convergence results for alternating update under perturbation can be translated into sample complexity bounds for different latent variable models. For convenience, we consider the discrete multi-view model with *generic* model parameters, and employ third order moment tensors for unsupervised learning. In this case, the sample complexity scales as $O(\text{polylog}(d,k)/w_{\min}^2)$, where $w_{\min}$ is the minimum weight of any tensor component. This is an improvement over existing results in (Song et al., 2013; Anandkumar et al., 2013a, 2012b), since we do not have dependence on the condition number of the factor matrices. This is especially relevant since generic factor matrices can be ill-conditioned, when $k \approx d$. Specifically, for a random $d \times d$ matrix, the condition number scales as $O(d)$ (Tao and Vu, 2010). On the other hand, the previous tensor approaches in (Song et al., 2013; Anandkumar et al., 2013a) have a sample complexity of $O(k^2 \cdot \text{polylog}(d,k)/w_{\min}^{1.5} \cdot \sigma_k(A)^3)$, where $\sigma_k(\cdot)$ is the k^{th} singular value and A is the factor matrix. The earlier work by Anandkumar et al. (2012b) considers a simultaneous diagonalization approach on two slices of the tensor. The sample bound in Anandkumar et al. (2012b) is even worse since it relies on the random tensor slices being well conditioned, which scales poorly in the tensor rank. For a detailed discussion on this aspect, see (Anandkumar et al., 2012a, Appendix D). Thus, we provide the best known sample bounds for unsupervised learning of discrete multiview mixtures assuming incoherent factor matrices.

For highly overcomplete tensors (up to $k = o(d^{3/2})$), we consider constructing a rough initial estimate using labeled samples under the semi-supervised setting. Here, the labels correspond to the choice variable of the discrete multi-view mixture model, and enables us to construct a rough estimate of the tensor components. We prove that having an extremely small number of labeled samples, scaling as $O\left(\text{polylog}(d,k)\right)$ for each label, is sufficient to provide global convergence guarantees for overcomplete tensors. Note that in most applications, labeled samples are expensive/hard to obtain, while many more unlabeled samples are easily available, e.g., see Le et al. (2011); Coates et al. (2011). To the best of our knowledge, we give the first guarantees for overcomplete tensor decomposition of third order tensors under mild incoherence conditions.

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 incoherence 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-p norm for random matrices (Guédon and Rudelson, 2007; Adamczak et al., 2011) 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 notion. The most commonly used algorithm for CP decomposition is Alternating Least Squares (ALS) (Comon et al., 2009), which has no convergence guarantees in general. A guaranteed approach for CP decomposition consists of two steps, viz., whitening the input tensor to obtain an orthogonal symmetric form, and then employing the tensor power update procedure to find the orthogonal decomposition. 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. (2012a) analyze deflation procedure for orthogonal decomposition, and Song et al. (2013) extend analysis to the nonparametric setting. 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. 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. Moreover, 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 one considered in this paper. For instance, Arora et al. (2013); Agarwal et al. (2013) provide guarantees for the sparse coding problem. Anandkumar et al. (2013c) 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 more general than the CP decomposition considered here, and the techniques in (Anandkumar et al., 2013c) 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

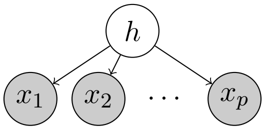

Figure 1: Multi-view mixture model.

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 Learning Mixture Models via Tensor Decomposition

As discussed in the introduction, the general tensor decomposition guarantees provided in this paper can be applied to unsupervised (and semi-supervised) learning problems. In this section, we briefly introduce the multi-view mixture model and discuss its connection with tensor decomposition problem.

Consider a multiview mixture model in Figure 1 with k components and $p \geq 3$ views. Here, the variables (views) $x_i \in \mathbb{R}^d$ are conditionally independent given the k-dimensional latent variable $h \in \mathbb{R}^k$. Let the conditional expectation of views x_l, denoted by $\mathbb{E}[x_l|h]$ be a linear map of hidden state h. Suppose the hidden variable h is a discrete k-dimensional random variable, with[2] $h = e_j \in \mathbb{R}^k$ if it takes the j-th value, then, the third order moment has the form (See Anandkumar et al. (2012a))

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{j \in [k]} w_j a_j \otimes b_j \otimes c_j. \tag{1}$$

Denote matrix $A := [a_1|a_2|\dots|a_k]$, and similarly B and C. These matrices correspond to the linear maps for each of the observed views in the model. 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. The decomposition in (3) is referred to as the CP decomposition, and k denotes the CP tensor rank.

Therefore, given third order moment of observed variables (views), the unsupervised learning problem reduces to computing a tensor decomposition in (3). Moreover, this framework can be extended to a number of other latent variable models by considering modified moment tensors. This includes Latent Dirichlet Allocation (LDA), Independent component analysis (ICA) and Gaussian mixtures (Anandkumar et al., 2012a). Thus, an efficient tensor decomposition procedure leads to efficient learning procedure for a wide range of latent variable models.

The tensor-based approach can also be extended to supervised learning with latent variable modeling. For instance, consider the multi-view mixture model with $x_1, x_2 \in \mathbb{R}^{d_u}$ as high-dimensional features and $x_3 = y \in \mathbb{R}^{d_o}$ as a low dimensional label ($d_o \ll d_u$), and the views and the label are conditionally independent given the hidden variable h. The moment tensor $\mathbb{E}[x_1 \otimes x_2 \otimes y]$ satisfies the form in (3), but this is a more challenging setting, since the label dimension is typically $d_o \ll k$, where k is the dimension of the hidden variable h. In other words, $\mathbb{E}[x_1 \otimes x_2 \otimes y]$ tensor is overcomplete along the y-mode. We provide a guaranteed procedure for learning overcomplete tensors, and thus, we can learn the above model efficiently.

[2] $e_j \in \mathbb{R}^k$ denotes the j-the basis vector in k-dimensional space.

1.4 Notation

A real *p-th order tensor* $T \in \bigotimes_{i=1}^{p} \mathbb{R}^{d_i}$ is a member of the tensor product of Euclidean spaces $\mathbb{R}^{d_i}$, $i \in [p]$. We generally 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 $[T_{i_1,i_2,\ldots,i_p} \colon i_1, i_2, \ldots, i_p \in [d]]$, 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). We generally limit to third order tensors ($p = 3$) in our analysis.

We view a tensor $T \in \mathbb{R}^{d \times d \times d}$ as a multilinear form. Consider matrices $M_i \in \mathbb{R}^{d \times d_i}, i = \{1, 2, 3\}$. 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

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

In particular, if θ is a vector and T is a third order tensor, then $T(\theta, \theta, \theta)$ is a number, $T(\theta, \theta, I)$ is a vector and $T(\theta, I, I)$ is a matrix.
Throughout, $\|v\| = (\sum_i v_i^2)^{1/2}$ denotes the Euclidean norm of a vector v, and $\|M\|$ denotes the spectral (operator) norm of a matrix. Also, $\|T\|$ and $\|T\|_F$ denotes the operator norm and the Frobenius norm of a tensor, respectively. In particular, for a symmetric order-3 tensor,

$$\|T\| = \sup_{\|\theta\|=1} |T(\theta, \theta, \theta)|; \quad \|T\|_F = \sqrt{\sum_{i,j,k \in [d]} T_{i,j,k}^2}.$$

A third order tensor T is said to be rank-1 if it can be written in the form,

$$T = w a \otimes b \otimes c \Leftrightarrow T(i, j, k) = w a(i) b(j) c(k), \tag{2}$$

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

$$T = \sum_{i \in [k]} w_i a_i \otimes b_i \otimes c_i. \tag{3}$$

This decomposition is closely related to the multilinear form. In particular, for vectors $\hat{a}, \hat{b}, \hat{c}$, we know $T(\hat{a}, \hat{b}, \hat{c}) = \sum_{i \in [k]} w_i \langle a_i, \hat{a} \rangle \langle b_i, \hat{b} \rangle \langle c_i, \hat{c} \rangle$. Denote matrix $A := [a_1 | a_2 | \ldots | a_k]$, 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.

Fibers are higher order analogues of matrix rows and columns. A fiber is obtained by fixing all but one of the indices (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 $T(:, j, k)$, mode-2 by $T(i, :, k)$ and so on. 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.

Finally, we use the asymptotic notation $f(d) = \tilde{O}(g(d))$ if $f(d) \leq \alpha g(d)$ for $\alpha = \mathrm{polylog}(d)$, i.e., $\tilde{O}$ hides polylog factors.

Algorithm 1 Tensor decomposition via alternating asymmetric power method

```
Require: Tensor $T \in \mathbb{R}^{d\times d\times d}$, number of initializations $L$, number of iterations $N$.
  for $\tau = 1$ to $L$ do
    Initialize unit vectors $\hat{a}_\tau^{(0)} \in \mathbb{R}^d$, $\hat{b}_\tau^{(0)} \in \mathbb{R}^d$, and $\hat{c}_\tau^{(0)} \in \mathbb{R}^d$. Option 1: SVD-based initialization in Algorithm 2 for undercomplete regime. Option 2: random initialization.
    for $t = 0$ to $N-1$ do
      Updates:
      $$\hat{a}_\tau^{(t+1)} = \frac{T\left(I, \hat{b}_\tau^{(t)}, \hat{c}_\tau^{(t)}\right)}{\left\|T\left(I, \hat{b}_\tau^{(t)}, \hat{c}_\tau^{(t)}\right)\right\|}, \quad \hat{b}_\tau^{(t+1)} = \frac{T\left(\hat{a}_\tau^{(t)}, I, \hat{c}_\tau^{(t)}\right)}{\left\|T\left(\hat{a}_\tau^{(t)}, I, \hat{c}_\tau^{(t)}\right)\right\|}, \quad \hat{c}_\tau^{(t+1)} = \frac{T\left(\hat{a}_\tau^{(t)}, \hat{b}_\tau^{(t)}, I\right)}{\left\|T\left(\hat{a}_\tau^{(t)}, \hat{b}_\tau^{(t)}, I\right)\right\|}, \quad (4)$$
      $$\hat{w}_\tau^{(t+1)} = \sqrt[3]{\left\|T\left(I, \hat{b}_\tau^{(t)}, \hat{c}_\tau^{(t)}\right)\right\| \cdot \left\|T\left(\hat{a}_\tau^{(t)}, I, \hat{c}_\tau^{(t)}\right)\right\| \cdot \left\|T\left(\hat{a}_\tau^{(t)}, \hat{b}_\tau^{(t)}, I\right)\right\|}.$$
    end for
  end for
  Cluster set $\left\{\left(\hat{w}_\tau^{(N)}, \hat{a}_\tau^{(N)}, \hat{b}_\tau^{(N)}, \hat{c}_\tau^{(N)}\right), \tau \in [L]\right\}$ into $k$ clusters.
  return the center member of these $k$ clusters as estimations $(\hat{w}_j, \hat{a}_j, \hat{b}_j, \hat{c}_j), j \in [k]$.
```

Algorithm 2 SVD-based initialization in the undercomplete setting

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

2 Alternating Tensor Decomposition Algorithm

The rank-1 alternating update procedure is given in Algorithm 1. Given an initial estimate of the vectors, denoted by $\left(\hat{a}^{(0)}, \hat{b}^{(0)}, \hat{c}^{(0)}\right)$, the algorithm performs an *asymmetric* power update on the given tensor T in each iteration, by alternating between different modes of the tensor. The asymmetric power update procedure is then run for N iterations.

We now provide a simple intuition behind the power update procedure: consider a rank-k tensor T, as in (3), and initialization $\hat{a} = a_j$ and $\hat{b} = b_j$, for some $j \in [k]$. Then, we have

$$T\left(\hat{a}, \hat{b}, I\right) = T\left(a_j, b_j, I\right) = w_j c_j + \sum_{i \neq j} w_i \langle a_j, a_i \rangle \langle b_j, b_i \rangle c_i, \qquad (5)$$

where the first term is along c_j and the second term is arising due to non-orthogonality. For orthogonal decomposition, the second term is zero, leading to the result that the true vectors a_j, b_j and c_j are stationary points for the power update procedure, for $j \in [k]$. However, since we consider non-orthogonal tensors, this procedure cannot recover the decomposition exactly. Under incoherence, we establish that the second term in (5) is small, which leads to approximate recovery results.

For generating initialization vectors $\left(\hat{a}^{(0)}, \hat{b}^{(0)}, \hat{c}^{(0)}\right)$, we introduce two possibilities. The first 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 second option is SVD-based Algorithm 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 (4). We establish in Section 3.2 that the SVD procedure leads to global convergence guarantees when $k = O(d)$ under polynomial trials.

As a final note, in order to identify which initializations are successful, we need a "clustering" procedure to obtain the final estimates of the vectors, and this is described in Appendix G.

Comparison with symmetric tensor power method: This algorithm is similar to the symmetric tensor power method analyzed by Anandkumar et al. (2012a) 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 symmetric power method in (Anandkumar et al., 2012a) is only for symmetric tensors.
- Linearity: The updates in Algorithm 1 are linear in each variable, while the tensor power update is a quadratic operator given a third order tensor.
- Guarantees: In (Anandkumar et al., 2012a), 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 (4) can be rewritten as

$$\tilde{c}_{\tau}^{(t+1)} := T\left(\hat{a}_{\tau}^{(t)}, \hat{b}_{\tau}^{(t)}, I\right) = \mathrm{mat}(T, 3) \cdot \left(\hat{b}_{\tau}^{(t)} \odot \hat{a}_{\tau}^{(t)}\right). \tag{6}$$

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 $\tilde{C}$) are simultaneously updated. In contrast, our procedure updates only one vector (with the target of recovering a 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.

Efficient implementation given samples: In Algorithm 1, a given tensor T is input, and we then perform the updates. However, in many settings (especially machine learning applications), the tensor is not available before hand, and needs to be computed from samples. Computing and storing the tensor can be enormously expensive for high-dimensional problems. Here, we provide a simple observation on how we can manipulate the samples directly to carry out the update procedure in Algorithm 1 as *multi-linear* operations, leading to efficient computational complexity.

Consider the setting, where the goal is to decompose the empirical moment tensor $\hat{T}$ of the form

$$\hat{T} := \frac{1}{n} \sum_{l \in [n]} x_1^{(l)} \otimes x_2^{(l)} \otimes x_3^{(l)}, \tag{7}$$

where $x_i^{(l)}$ are the l^{th} samples from views $i \in [3]$. Applying the update in (4) in Algorithm 1 to $\hat{T}$, we have

$$\tilde{c} = \frac{1}{n} X_3 (X_1^\top \hat{a} * X_2^\top \hat{b}), \tag{8}$$

where $*$ corresponds to the Hadamard product. Here, $X_i := \left[x_i^{(1)} x_i^{(2)} \cdots x_i^{(n)} \right] \in \mathbb{R}^{d \times n}$. Thus, the update can be computed efficiently using simple matrix and vector operations. It is easy to see that the above update in (8) is easily parallelizable, and moreover, the different initializations can be parallelized, making the algorithm scalable for large problems.

3 Analysis

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

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

is a rank-k tensor such that $a_i, b_i, c_i \in \mathbb{R}^d, i \in [k]$, are unit vectors. Without loss of generality we assume $w_{\max} = w_1 \geq w_2 \geq \cdots \geq w_k = w_{\min} > 0$. Let $A := [a_1 \ a_2 \ \cdots \ a_k] \in \mathbb{R}^{d \times k}$, and B and C are similarly defined. Also, for simplicity we assume $a_i, b_i, c_i, i \in [k]$, are generated uniformly at random from the unit sphere S^{d-1}. We state the deterministic assumptions in Appendix B, and show that random matrices satisfy these assumptions. Notice that it is also reasonable to assume these assumptions hold for some non-random matrices.

3.1 Local Convergence Guarantee

The local convergence guarantee is 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*

$$\operatorname{dist}(u, v) := \sup_{z \perp u} \frac{\langle z, v \rangle}{\|z\| \cdot \|v\|} = \sup_{z \perp v} \frac{\langle z, u \rangle}{\|z\| \cdot \|u\|}. \tag{9}$$

Note that distance function $\operatorname{dist}(u, v)$ 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))

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

Incorporating distance notion resolves the sign ambiguity issue in recovering the components: 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.

Let $\psi := \|\Psi\|_F$ denote the Frobenius norm of error tensor Ψ, and

$$\epsilon_T := \frac{\psi}{w_{\min}} + \tilde{O}\left(\gamma \frac{\sqrt{k}}{d}\right), \tag{10}$$

denote the target error where $\gamma := \frac{w_{\max}}{w_{\min}}$. The estimation error after t iterations is bounded in the following theorem.

Theorem 1 (Local convergence guarantee of Algorithm 1)**.** *Consider $\hat{T} = T + \Psi$ as the input to Algorithm 1, where*

$$T = \sum_{i\in[k]} w_i a_i \otimes b_i \otimes c_i,$$
$$\psi := \|\Psi\|_F \le w_{\min}/ \operatorname{polylog} d.$$

Suppose the following initialization bound holds w.r.t. some $j \in [k]$ as

$$\epsilon_0 := \max\left\{\operatorname{dist}\left(\hat{a}^{(0)}, a_j\right), \operatorname{dist}\left(\hat{b}^{(0)}, b_j\right)\right\} = O(1/\gamma),$$

where $\gamma := \frac{w_{\max}}{w_{\min}}$. Given $\hat{a}^{(0)}$ and $\hat{b}^{(0)}$, suppose $\hat{c}^{(0)}$ is also calculated by the update formula in (4)*. Then, when*

$$\gamma = O(\min\{\sqrt{d}, d^{1.5}/k\}), \quad k = o\left(d^{1.5}\right),$$

the iterations of Algorithm 1 satisfy the following bound with high probability (whp)

$$\max\left\{\operatorname{dist}\left(\hat{a}^{(t)}, a_j\right), \operatorname{dist}\left(\hat{b}^{(t)}, b_j\right), \operatorname{dist}\left(\hat{c}^{(t)}, c_j\right)\right\} \le O(\epsilon_T) + q^t \epsilon_0. \tag{11}$$

Here $q < 1$ is a contraction factor and ϵ_T is defined in (10)*.*

Thus, we provide efficient recovery guarantees for alternating rank-1 updates under incoherent factors. Our recovery is in terms of distance between any true vector a_j (or b_j, c_j) and the estimate $\hat{a}^{(t)}$ (or $\hat{b}^{(t)}$, $\hat{c}^{(t)}$).

Note that the second term in (11) is decaying linearly with the number of iterations. The first term in (11) is fixed (even as $t \to \infty$), and arises due to perturbation tensor Ψ (given by $\frac{\psi}{w_{\min}}$) and non-orthogonality (given by $\tilde{O}\left(\gamma \frac{\sqrt{k}}{d}\right)$). Thus, there is an approximation error in recovery of the tensor components. As $t \to \infty$, (11) can be interpreted as an approximate local identifiability result for tensor decomposition under incoherent factors.

The result in (11) can be stated in the non-asymptotic form, and the contraction factor $q < 1$ can be characterized explicitly. See Appendix B for details.

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

Theorem 1 provides local convergence guarantee given good initialization vectors for different components. In this section, we exploit SVD-based initialization method in Algorithm 2 to provide good initialization vectors when $k = O(d)$. Combining the theoretical guarantees of this initialization method (provided in Appendix D) with the local convergence guarantee in Theorem 1, we provide the following global convergence result.

Theorem 2 (Global convergence guarantee of Algorithm 1 when $k = O(d)$)**.** *Consider $\hat{T} = T + \Psi$ as the input to Algorithm 1, where $T = \sum_{i\in[k]} w_i a_i \otimes b_i \otimes c_i$, and $\psi := \|\Psi\|_F \leq w_{\min}/\operatorname{poly}\log d$. Let the initialization in each run of Algorithm 1 is performed by SVD-based technique proposed in Algorithm 2. Suppose*

$$k = O(d), \quad N = \Theta(\log(\gamma/\epsilon_T)), \quad L \geq k^{\Omega\left(\gamma^4 (k/d)^2\right)},$$

where $\gamma := \frac{w_{\max}}{w_{\min}}$. Then, for any $j \in [k]$, the output of Algorithm 1 satisfies the following whp,

$$\max\left\{\operatorname{dist}\left(\hat{a}_j, a_j\right), \operatorname{dist}\left(\hat{b}_j, b_j\right), \operatorname{dist}\left(\hat{c}_j, c_j\right)\right\} = O(\epsilon_T).$$

Thus, we can efficiently recover the tensor decomposition up to an approximation error ϵ_T, when the tensor is undercomplete or mildly overcomplete (i.e. $k = O(d)$), using a simple SVD-based initialization and then running alternating rank-1 updates. The number of initialization trials L is polynomial when γ is a constant, and $k = O(d)$.

Two undercomplete, and one overcomplete component

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

Since in the SVD initialization in Algorithm 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 (4), we can generalize the global convergence result in Theorem 2 to the setting where A, B are undercomplete, and C is relatively overcomplete.

Corollary 1. *Consider the same setting as in Theorem 2. In addition, suppose the regime of undercomplete components $A \in \mathbb{R}^{d_u \times k}$ and $B \in \mathbb{R}^{d_u \times k}$, 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*

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

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

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

Remark 1. *When the two undercomplete modes A and B have orthogonal columns, then the constraint $k = O(\sqrt{d_u d_o})$ in the above theorem can be further relaxed. It now suffices to have $k = O(d_u)$, for any d_o. This is because under orthogonality, the SVD initialization provides a much better initialization than under non-orthogonal components.*

3.3 Applications to learning of multiview mixture models

Recall that the tensor decomposition can be directly applied for learning multiview mixture models, introduced in Section 1.3, and the third order moment has the following form,

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{j \in [k]} w_j a_j \otimes b_j \otimes c_j.$$

We now apply the tensor recovery results in Theorems 1 and 2 to obtain sample complexity bounds for learning multiview mixture models. We consider two settings, viz., unsupervised setting where the information about label h is not available, and semi-supervised setting, where a small amount of label information is available. We see that in the former setting, we can handle mixtures with number of components $k = O(d)$, where d is the observed dimension, while in the latter case, we can go up to $k = o(d^{1.5})$.

3.3.1 Unsupervised learning, $k = O(d)$

In this section, we apply the global convergence guarantee to the multiview mixture model introduced in Section 1.3. The empirical tensor $\hat{T}$ is computed from samples, as given by (7). Then, the norm of error tensor $\Psi := \hat{T} - T$ can be bounded, given n samples. We have with probability at least $1 - \delta$ (See Lemma 7 of Song et al. (2013))

$$\|\hat{T} - T\|_F \leq C_1 \sqrt{\frac{\log \frac{1}{\delta}}{n}}, \tag{12}$$

for some constant $C_1 > 0$.

We now provide sample complexity guarantees, under the assumption that the conditional probability tables for the three views of the mixture model are generically drawn.

Corollary 2 (Unsupervised learning of multiview mixture model)**.** *Under the assumptions of Theorem 2, for a discrete multiview mixture model, given n samples such that*

$$n \geq \max \left\{ 6, \frac{\alpha}{\sqrt{\log k}} \right\} \cdot \frac{\alpha}{w_{\min}^2}, \tag{13}$$

for $\alpha = \operatorname{polylog}(d)$*, we have*

$$\max \left\{ \operatorname{dist}\left(\hat{a}_j, a_j\right), \operatorname{dist}\left(\hat{b}_j, b_j\right), \operatorname{dist}\left(\hat{c}_j, c_j\right) \right\} = O(\epsilon_T).$$

Thus, the simple alternating updates in Algorithm 1 efficiently learns the multiview mixture model with sample complexity given by (13). Note that the sample complexity (up to polylog factors) only has dependence on $w_{\min}$, the minimum weight among the tensor components. For the mixture model, under equal weights, we have $w_{\min} = \Theta(k^{-1} d^{-1.5})$, due to our convention of rescaling the columns of the conditional probability table to have unit 2-norm. Thus, in this scenario, the sample complexity scales as $n \geq \tilde{O}(k^2 d^3)$.

We now compare the sample complexity in (13) with the previous result by Song et al. (2013), which employs whitening procedure followed by tensor power updates. The sample complexity in (Song et al., 2013) is given by

$$n = \tilde{O}\left(\frac{k^2}{w_{\min}^{1.5} \sigma_{\min}^3} \right), \tag{14}$$

where $\sigma_{\min} := \min(\sigma_k(A), \sigma_k(B), \sigma_k(C))$. For generic matrices A, B and C, when $k \approx d$, the lowest singular value has poor scaling, and is given by $\Theta(1/d)$. In this case, (14) simplifies as $\tilde{O}(\frac{k^2 d^3}{w_{\min}^{1.5}})$. Further, if we assume the equal weights setting, we can substitute for $w_{\min}$, and we have the sample complexity for the whitening + power method scaling as $\tilde{O}(k^{3.5} d^{5.25})$. In comparison, the sample complexity for our method scales as $\tilde{O}(k^2 d^3)$, which is better. This is especially relevant in the high dimensional regime, where k and d are large, and our method requires fewer samples for learning than the previous approaches.

We can similarly provide global recovery guarantees when two views of the multiview mixture are high-dimensional and the third view is low-dimensional, by considering the relationship between perturbation error and number of samples in (12) and substituting it in Corollary 1. Again, we obtain the sample bound as $\tilde{O}(1/w_{\min}^2)$. Under the equal weights setting, this scales as

$$n \geq \tilde{O}(k^2 d_u^2 d_o),$$

and in Corollary 1, we require $k = O(\sqrt{d_u d_o})$ to obtain global recovery guarantees. Thus, we establish guaranteed learning for a wide range of multiview mixture models with (a low order) polynomial sample complexity.

3.3.2 Semi-supervised learning in the overcomplete setting

In the previous section, we provided efficient guarantees for unsupervised learning of multi-view mixtures when the number of mixture components $k = O(d)$. However, for overcomplete mixtures, these guarantees are not applicable, and we only have local convergence result in Theorem 1. To circumvent this problem, we consider the semi-supervised setting, where we access to a small number of labeled samples, where the labels correspond to the choice variable h for each sample. We now exploit the labeled samples to obtain an efficient initialization for the alternating procedure.

Let $x_{1,j}^{(l)} \in \mathbb{R}^d$, $x_{2,j}^{(l)} \in \mathbb{R}^d$, and $x_{3,j}^{(l)} \in \mathbb{R}^d$, $j \in [k], l \in [m_j]$, denote $m = \sum_j m_j$ samples of vectors corresponding to different labels where in the semi-supervised setting, the samples with subscript j have label j. Then, for any $j \in [k]$, we have the empirical estimate of each column of A as

$$\hat{a}_j = \frac{1}{m_j} \sum_{l \in [m_j]} x_{1,j}^{(l)}, \tag{15}$$

for which we have the deviation bound that with probability at least $1 - \delta$ as

$$\|\hat{a}_j - a_j\| \leq C_1 \sqrt{\frac{\log(1/\delta)}{m_j}},$$

for some $C_1 > 0$. Similar empirical estimates and deviation bounds hold for $b_j, c_j, j \in [k]$. Applying the initialization deviation bounds to the local convergence result in Theorem 1, we have the following result.

Corollary 3 (Recovery in overcomplete semi-supervised setting)**.** *Consider the same setting as in Theorem 1 with $k = o\left(d^{1.5}\right)$. Suppose the number of labeled samples with label j, denoted by m_j, and the number of unlabeled samples n, satisfy*

$$n \geq \Omega\left(\text{poly}\log(d)/w_{\min}^2\right), \quad m_j \geq \Omega\left(\text{polylog}(d) \cdot \gamma^2\right), j \in [k]. \tag{16}$$

By employing the empirical estimates in (15) *as initializations, the output of Algorithm 1 satisfies the following bound with high probability*

$$\max\left\{\operatorname{dist}\left(\hat{a}_j, a_j\right), \operatorname{dist}\left(\hat{b}_j, b_j\right), \operatorname{dist}\left(\hat{c}_j, c_j\right)\right\} = O(\epsilon_T),$$

where ϵ_T *is defined in* (10) *and the Algorithm 1 is run for* N *iterations such that* $N = \Theta(\log(\gamma/\epsilon_T))$.

Thus, we provide guaranteed learning of overcomplete mixture models in the semi-supervised setting. Note that in (16), the requirement for number of labeled samples is far lower than that for the number of unlabeled samples: $m_j \ll n$, for $j \in [k]$. Specifically, when the ratio of weights γ is a constant, we require $m_j = O(\text{polylog}(d, k))$, while the requirement for $n = \text{poly}(d, k)$, since we have $w_{\min} = \text{poly}(d, k)$. Thus, we provide efficient guarantees for overcomplete models under an extremely small number of labeled samples.

3.4 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 D.

The local convergence result is derived by establishing error contraction in each iteration of Algorithm 1. 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 B. Under these conditions, with a good initialization (i.e. small enough $\text{dist}(\hat{a}, a_j)$ and $\text{dist}(\hat{b}, b_j)$), we show that the iterative update in (4) provides an estimate $\hat{c}$ with $\text{dist}(\hat{c}, c_j) < O(\epsilon_T) + q\epsilon_0$, for some contraction factor $q < 1$. The incoherence condition is crucial for establishing this result. See Appendix C for the complete proof.

The initialization argument for SVD-based technique in Algorithm 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. (2012a). 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 D for the complete proof.

4 Experiments

In this section, we provide some synthetic experiments to evaluate the performance of Algorithm 1. 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 (19), 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[3]. Initialization vector $\hat{c}^{(0)}$ is generated through update formula in (4).

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

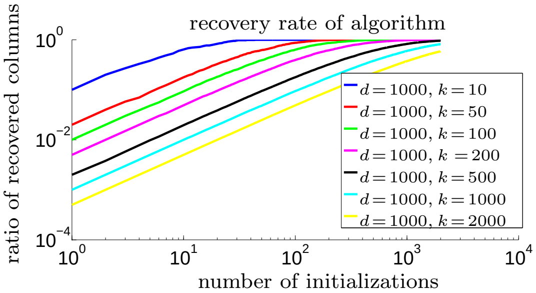

Figure 2: Ratio of recovered columns versus the number of initializations for $d = 1000$, and $k = \{10, 50, 100, 200, 500, 1000, 2000\}$. The number of initializations is $L = 2000$. The stopping parameter is set to $t_1 = 1e - 08$. The figure is an average over 10 random runs.

For each initialization $\tau \in [L]$, we run the algorithm with a fixed number of iterations N based on following stopping criterion

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

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 in t_1 (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 ($k \leq d$), 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 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 estimation. The number of iterations performed before

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

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. 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 2. 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

A More Matrix and Tensor Notations

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 $A \in \mathbb{R}^{p \times q}$ and $B \in \mathbb{R}^{m \times n}$, the *Kronecker* product [4] $A \otimes B \in \mathbb{R}^{pm \times qn}$ is defined as (Golub and Van Loan, 2012)

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \cdots & a_{1q}B \\ a_{21}B & a_{22}B & \cdots & a_{2q}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1}B & a_{p2}B & \cdots & a_{pq}B \end{bmatrix},$$

Thus, for vectors $a, b \in \mathbb{R}^d$, we have $a \otimes b \in \mathbb{R}^{d^2}$ such that

$$a \otimes b := \begin{bmatrix} a_1 b \\ a_2 b \\ a_3 b \\ \vdots \end{bmatrix}.$$

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

$$A \odot B(\mathbf{i}, j) := A_{i_1, j} B_{i_2, j}, \quad \mathbf{i} = (i_1, i_2) \in [d]^2, j \in [k].$$

In other words, we have

$$A \odot B := [a_1 \otimes b_1 \quad a_2 \otimes b_2 \quad \ldots a_k \otimes b_k],$$

where a_i, b_i are the i^{th} columns of A and B.

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

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

For any $A \in \mathbb{R}^{p \times k}, B \in \mathbb{R}^{q \times k}, C \in \mathbb{R}^{p \times n}$, and $D \in \mathbb{R}^{q \times n}$, we have the following property

$$(A \odot B)^\top (C \odot D) = (A^\top C) * (B^\top D). \tag{18}$$

Let $\|A\|_\infty$ denote the ℓ_∞ element-wise norm of matrix A, and

$$\|A\|_{2 \to p} := \sup_{\|\theta\|=1} \|A\theta\|_p.$$

B 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

$$T = \sum_{i \in [k]} w_i (a_i \otimes b_i \otimes c_i), \quad w_i > 0, a_i, b_i, c_i \in \mathcal{S}^{d-1}, \forall\, i \in [k], \tag{19}$$

[4] Throughout this paper, notation $\otimes$ is usually used for outer product introduced in (2). But, with slightly abuse of notation, we also use that for Kronecker product.

where $\mathcal{S}^{d-1}$ denotes the unit d-dimensional sphere, i.e. all the vectors have unit[5] 2-norm as $\|a_i\| = \|b_i\| = \|c_i\| = 1, i \in [k]$. Furthermore, define $w_{\min} := \min_{i\in[k]} w_i$ and $w_{\max} := \max_{i\in[k]} w_i$.

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

$$\rho := \max_{i\neq j}\{|\langle a_i, a_j\rangle|, |\langle b_i, b_j\rangle|, |\langle c_i, c_j\rangle|\} \leq \frac{\alpha}{\sqrt{d}}, \tag{20}$$

for some $\alpha = \operatorname{polylog}(d)$. 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 (20).

(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

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

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

(A5) **Size constraint:** Consider the size constraint where $k = o\left(d^{1.5}\right)$, i.e., $\frac{k}{d^{1.5}}$ converges to zero in the asymptotic regime.

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

$$\psi := \|\Psi\|_F. \tag{21}$$

Suppose ψ is bounded as

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

where $\alpha = \operatorname{polylog}(d)$.

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

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

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

(A8) **Contraction factor:** Define contraction factor q as

$$q := \frac{2w_{\max}}{w_{\min}} \left[\frac{2\alpha}{\sqrt{d}} \left(1 + \alpha_0 \sqrt{\frac{k}{d}} \right)^2 + \beta' \right], \tag{22}$$

for some constants $\alpha_0, \beta' > 0$, and $\alpha = \text{polylog}(d)$. Suppose parameters β', d, and k are such that $q < 1$.

(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

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

for some constants $\alpha_0, \beta' > 0$, $\alpha = \text{polylog}(d)$, and $0 < q < 1$ which is defined in (22).

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

Remark 2. *Most of the Assumptions are actually parameter choices. The only properties of random matrices are (A2), (A3), (A4), (A10), and (A10) is only used in the unsupervised learning setting. See Appendix B.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) are satisfied whp when the columns of A, B, and C are generically drawn from unit sphere $\mathcal{S}^{d-1}$ (see Lemma 1 and Guédon and Rudelson (2007)). Assumption (A5) limits the overcompleteness of problem which is required for providing convergence guarantees. The bound on perturbation in (A6) is required for local convergence analysis and arguing initialization bound for Algorithm 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.

B.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 1. *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 $2-3$ norm of matrices. Using tools from Guédon and Rudelson (2007); Adamczak et al. (2011), we have the following:

Lemma 2. *Consider a random matrix $A \in \mathbb{R}^{d\times k}$ whose columns are drawn uniformly at random from unit sphere. If $k < d^{p/2}/\operatorname{poly}\log d$, then $A^\top$ has $2 \to p$ norm at most $1 + o(1)$.*

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

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*

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

As a corollary

Corollary 4. *For vectors f, g, h, and weights $w_i \geq 0$, $\sum_i w_i f_i g_i h_i \leq w_{\max}\|f\|_3\|g\|_3\|h\|_3$.*

Proof: The proof applies Hölder's inequality twice. In the first application $p = 3$ and $q = 3/2$, in the second application $p = q = 2$ (which is the special case known as Cauchy-Schwartz)

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

□

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

Lemma 3. *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 $k < d^{3/2}/\operatorname{poly}\log d$, 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 $\hat{a}, \hat{b}, \hat{c}$, consider $T(\hat{a}, \hat{b}, \hat{c})$. It is equal to $\sum_{i\in[k]} w_i (A^\top \hat{a})_i (B^\top \hat{b})_i (C^\top \hat{c})_i$. By Corollary 4, $T(\hat{a}, \hat{b}, \hat{c})$ is upper bounded by $w_{\max}\|A^\top \hat{a}\|_3\|B^\top \hat{b}\|_3\|C^\top \hat{c}\|_3 \leq O(w_{\max})$, because $\|A^\top \hat{a}\|_3 \leq \|A^\top\|_{2\to 3}\|\hat{a}\| = O(1)$ (and similarly for b, c). □

Finally, in the following lemma we prove for random matrices that $\|C_{\setminus i} \operatorname{diag}(w)(J_A * J_B)_i^{\setminus i}\| = \tilde{O}(w_{\max}\sqrt{k}/d)$. This is the second bound in Assumption (A4).

Lemma 4. *If A, B, and $C \in \mathbb{R}^{d\times k}$ are independent, normalized Gaussian matrices, then for all i, we have with high probability, $\|C_{\setminus i} \operatorname{diag}(w)(J_A * J_B)_i^{\setminus i}\| = \tilde{O}(w_{\max}\sqrt{k}/d)$.*

Proof: We rewrite the vector $C_{\setminus i} \operatorname{diag}(w)(J_A * J_B)_i^{\setminus i}$ as $\sum_{j\neq i} C_j w_j \langle A_i, A_j\rangle\langle B_i, B_j\rangle = \sum_{j\neq i} C_j \delta_j$. Here $\delta_j = w_j\langle A_i, A_j\rangle\langle B_i, B_j\rangle$ is independent of C_j. Since A and B are random, in particular they are incoherent. Hence, for $j \neq i$, we have $|\delta_j| \leq \tilde{O}(w_{\max}/d)$. Now since C_j's are independent, mean 0 vectors, the sum $\sum_{j=1}^{k} \delta_j C_j$ is a sum with mean 0 and variance bounded by $\tilde{O}(w_{\max}^2 k/d^2)$. From vector Bernstein's bound we know $\|C_{\setminus i}(J_A * J_B)_i^{\setminus i}\| = \tilde{O}(w_{\max}\sqrt{k}/d)$ with high probability. Then, we can apply union bound for all i. □

C 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 Algorithm 1. Then, the contraction result after t iterations is directly argued. In the following two lemmata, we provide a local contraction result for one update (iteration) of Algorithm 1 given perturbed tensor $\hat{T}$.

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

$$f(\epsilon; k, d) := \alpha\frac{\sqrt{k}}{d} + \frac{2\alpha}{\sqrt{d}}\left(1 + \alpha_0\sqrt{\frac{k}{d}}\right)^2 \epsilon + \alpha_0\epsilon^2, \tag{23}$$

where $\alpha = \operatorname{polylog}(d)$ and $\alpha_0 = O(1)$.

Lemma 5 (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 (A1)-(A5) hold. Let estimates $\hat{a}$ and $\hat{b}$ satisfy distance bounds*

$$\operatorname{dist}(\hat{a}, a_j) \leq \epsilon_a,$$
$$\operatorname{dist}(\hat{b}, b_j) \leq \epsilon_b,$$

for some $j \in [k]$, and $\epsilon_a, \epsilon_b > 0$. Suppose $\epsilon := \max\{\epsilon_a, \epsilon_b\}$, and ψ defined in (21) be small enough such that[6]

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

where $f(\epsilon; k, d)$ is defined in (23). Then, update $\hat{c}$ in (4) satisfies the following distance bound with high probability (whp)

$$\operatorname{dist}(\hat{c}, c_j) \leq \frac{w_{\max} f(\epsilon; k, d) + \psi}{w_j - w_j\epsilon^2 - w_{\max} f(\epsilon; k, d) - \psi}. \tag{24}$$

Remark 3. *In the asymptotic regime, $f(\epsilon; k, d)$ is*

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

[6] This is the denominator of bound provided in (24).

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 4 (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 5 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 5 is proposed in the general form. In Lemma 6, we provide explicit contraction result by imposing additional perturbation, contraction and initialization Assumptions (A6), (A8) and (A9). We observe that under reasonable size, perturbation and initialization conditions, the denominator can be lower bounded by a constant, and the numerator is explicitly bounded by a term involving ϵ, and a constant non-contracting term.

Lemma 6 (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 (A1)-(A9) hold. Note that initialization bound in (A9) is satisfied for some $j \in [k]$. Then, update $\hat{c}$ in* (4) *satisfies the following distance bound with high probability (whp)*

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

where

$$\text{Const.} := \frac{2}{w_{\min}}\left(\psi + w_{\max}\alpha\frac{\sqrt{k}}{d}\right), \tag{25}$$

and contraction ratio $q < 1$ is defined in (22)*. Note that $\alpha = \operatorname{polylog}(d)$.*

Proof of Theorem 1: We incorporate condition (A7) to show that $q < 1$ 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 6. □

Proof of Theorem 2: The result is proved by combining the local convergence result in Theorem 1, and initialization result in Theorem 3. □

C.1 Proof of auxiliary lemmata

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 (20).

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 5: Let $z_a^* \perp a_j$ and $z_b^* \perp b_j$ denote the vectors that achieve supremum value in (9) 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

$$\hat{a} = \langle a_j, \hat{a}\rangle a_j + \operatorname{dist}(\hat{a}, a_j) z_a^*, \tag{26a}$$

$$\hat{b} = \langle b_j, \hat{b}\rangle b_j + \operatorname{dist}(\hat{b}, b_j) z_b^*. \tag{26b}$$

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 (4). 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

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

Substituting $\hat{a}$ and $\hat{b}$ from (26a) and (26b), we have

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

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

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

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), where it is shown in Lemma 4 that this condition holds for random matrices. For S_2, we have

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

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

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

Finally, for S_4, we have

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

for some $\alpha_0 = O(1)$. The bound on $\|T\|$ is from Assumption (A4). Note that for random components, this bound holds whp from Assumption (A5) and Guédon and Rudelson (2007). For $\Psi(\hat{a}, \hat{b}, z_c)$, we have

$$\Psi(\hat{a}, \hat{b}, z_c) = z_c^\top \operatorname{mat}(\Psi, 3)(\hat{b} \odot \hat{a}) \leq \|z_c^\top\| \cdot \|\operatorname{mat}(\Psi, 3)\| \cdot \left\|\hat{b} \odot \hat{a}\right\| = \|\operatorname{mat}(\Psi, 3)\| \leq \psi,$$

where equality $\|\hat{b} \odot \hat{a}\| = \|\hat{b}\| \cdot \|\hat{a}\| = 1$ is exploited in the last equality, and definition of ψ in (21) is used in the last inequality.
Let $\epsilon := \max\{\epsilon_a, \epsilon_b\}$. Then, we have whp

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

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

$$\begin{aligned} \tilde{c} &= T(\hat{a}, \hat{b}, I) + \Psi(\hat{a}, \hat{b}, I) \\ &= \sum_i w_i \langle a_i, \hat{a} \rangle \langle b_i, \hat{b} \rangle c_i + \operatorname{mat}(\Psi, 3)(\hat{b} \odot \hat{a}) \\ &= w_j \langle a_j, \hat{a} \rangle \langle b_j, \hat{b} \rangle c_j + \sum_{i \neq j} w_i \langle a_i, \hat{a} \rangle \langle b_i, \hat{b} \rangle c_i + \operatorname{mat}(\Psi, 3)(\hat{b} \odot \hat{a}), \end{aligned}$$

and therefore,

$$\begin{aligned} \|\tilde{c}\| &\geq \left\| w_j \langle a_j, \hat{a} \rangle \langle b_j, \hat{b} \rangle c_j \right\| - \left\| \sum_{i \neq j} w_i \langle a_i, \hat{a} \rangle \langle b_i, \hat{b} \rangle c_i \right\| - \|\operatorname{mat}(\Psi, 3)(\hat{b} \odot \hat{a})\| \\ &\geq w_j - w_j \epsilon^2 - w_{\max} f(\epsilon; k, d) - \psi, \end{aligned}$$

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

$$\operatorname{dist}(\tilde{c}, \overline{C}_j) \leq \frac{w_{\max} f(\epsilon; k, d) + \psi}{w_j - w_j \epsilon^2 - w_{\max} f(\epsilon; k, d) - \psi}. \tag{27}$$

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. Note that $\tilde{c} = \|\tilde{c}\|\hat{c}$, and $\overline{C}_j = w_j c_j$ where $w_j > 0$.

□

Proof of Lemma 6: The result is proved by applying Lemma 5, and incorporating additional conditions (A6), (A8), and (A9). $f(\epsilon; k, d)$ in (23) can be bounded as

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

where $\epsilon \leq \frac{\beta'}{\alpha_0}$ from Assumption (A9) is exploited in the inequality. The last equality is concluded from definition of contracting factor q in (22). On the other hand, the denominator in (24) can be lower bounded as

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

where Assumptions (A9) and (A6) are used in the inequality. Applying Lemma 5, the result is proved. □

D SVD Initialization Result

In this section, we provide an SVD-based technique to propose 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 (19), we have

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

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

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

Since A and B are not orthogonal matrices, the expansion in (28) is not the SVD [7] 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

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

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

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 Algorithm 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 $0 < \tilde{\mu} < 1$. Then, whp, at least one of the pairs $(u_1^{(j)}, v_1^{(j)}), j \in [L]$, say j^*, satisfies*

$$\max\left\{\operatorname{dist}\left(u_1^{(j^*)}, a_1\right), \operatorname{dist}\left(v_1^{(j^*)}, b_1\right)\right\} \leq \frac{4 w_{\max}\mu_{\min}(1+\rho)\sqrt{\log k} + \alpha\|\Psi\|_F}{w_{\min}\tilde{\mu} g(L) - \alpha\|\Psi\|_F},$$

for $\alpha = \operatorname{polylog}(d)$.

Proof: Let $\lambda^{(j)} := \operatorname{Diag}(w) C^\top \theta^{(j)} \in \mathbb{R}^k$ and $\tilde{\lambda}^{(j)} := C^\top \theta^{(j)} \in \mathbb{R}^k$. From Lemmata 7 and 8, there exists a $j^* \in [L]$ such that whp, we have

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

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 12 finishes the proof. □

D.1 Auxiliary lemmata

In the following Lemma, we show that the gap condition between the maximum and the second maximum of vector λ required in Lemma 8 is satisfied under some number of random draws.

Lemma 7 (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*

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

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 $\lambda^{(j)} := \operatorname{Diag}(w) C^\top \theta^{(j)}$. 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. (2012a), 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

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

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 10, we have

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

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

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

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 8. *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 (whp),

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

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

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

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

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

where $A_{(1)} := \Big[\overbrace{a_1 | a_1 | \cdots | a_1}^{k-1 \text{ times}} \Big] \in \mathbb{R}^{d \times (k-1)}$, $B_{\backslash 1} := [b_2 | b_3 | \cdots | b_k] \in \mathbb{R}^{d \times (k-1)}$, 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

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

Where the first equality is concluded from Lemma 11, and Assumptions (A2) and (A3) are exploited in the last inequality. Similarly, for E_2 and E_3, we have

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

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 $\text{dist}^2(u_1,a_1) + \langle u_1,a_1\rangle^2 = 1$, the proof is complete for this case. □

E Auxiliary Lemmata

Lemma 9. *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 10. *Consider $r = [r_1, r_2, \dots, r_k]^\top \in \mathbb{R}^k$ 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 9, 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. □

Lemma 11. *Given $h \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$, let $H = [h|h|\cdots|h]\operatorname{Diag}(v) \in \mathbb{R}^{m\times n}$. 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. □

F Norm of Noise tensor

In the following lemma, we show that noise matrix $\Psi(I, I, \theta)$ has bounded norm with high probability which is useful for initialization.

Lemma 12. *Let $\theta \in \mathbb{R}^d$ be standard multivariate Gaussian. Then, with probability at least $1-\delta$, we have*

$$\|\Psi(I, I, \theta)\| \leq C_1 \|\Psi\|_F \sqrt{\log \frac{1}{\delta}},$$

for some constant $C_1 > 0$.

Proof: We have

$$\|\Psi(I, I, \theta)\| \leq \|\Psi(I, I, \theta)\|_F = \|\operatorname{mat}(\Psi, 3)^\top \theta\|,$$

where the equality is concluded from the fact that $\operatorname{vec}\left(\Psi(I, I, \theta)^\top\right) = \operatorname{mat}(\Psi, 3)^\top \theta$, where $\operatorname{vec}(\cdot)$ results the vectorized form of its argument. Then, by applying Lemma 13, the result can be proved. Let $M := \operatorname{mat}(\Psi, 3)^\top$, and $\Sigma = M^\top M$. Then,

$$\begin{aligned}
&\operatorname{Tr}(\Sigma) = \|M\|_F^2 = \|\Psi\|_F^2, \\
&\operatorname{Tr}(\Sigma^2) \leq \operatorname{Tr}(\Sigma)^2 = \|\Psi\|_F^4, \\
&\|\Sigma\| = \|M^\top M\| \leq \|M\|^2 \leq \|\Psi\|_F^2.
\end{aligned}$$

Finally, the result is proved by applying Lemma 13. □

For Gaussian random vectors, we know the following fact.

Lemma 13. *If $x \in \mathbb{R}^d$ is standard multivariate Gaussian, M be an arbitrary matrix, and $\Sigma = M^\top M$, then*

$$\Pr[\|Ax\|_2^2 > \operatorname{Tr}(\Sigma) + 2\sqrt{\operatorname{Tr}\Sigma^2 t} + 2\|\Sigma\|t] \leq e^{-t}.$$

G The Clustering Process

In the main algorithm, we need to cluster the 4-tuples into k clusters. Theoretically we only have convergence guarantees when the initialization vectors are good, and the failed initializations can potentially generate arbitrary 4-tuples. In the worst case these arbitrary 4-tuples can make the clustering hard, so we need to use a specifically designed algorithm.

Algorithm 3 Clustering algorithm for the main algorithm

```
Require: Tensor $T \in \mathbb{R}^{d\times d\times d}$, all the 4-tuple[8] $\left\{(\hat{w}_\tau, \hat{a}_\tau, \hat{b}_\tau, \hat{c}_\tau), \tau \in [L]\right\}$
  for $i = 1$ to $k$ do
    Among the remaining 4-tuples, let $\hat{a}, \hat{b}, \hat{c}$ has the largest $|T(\hat{a}, \hat{b}, \hat{c})|$ value.
    Do $N$ more iterations of the update from $\hat{a}, \hat{b}, \hat{c}$.
    Let $\hat{a}, \hat{b}, \hat{c}$ be the center of cluster $i$.
    Remove all the tuples with $\max\{|\langle \hat{a}_i, \hat{a}\rangle|, |\langle \hat{b}_i, \hat{b}\rangle|, |\langle \hat{c}_i, \hat{c}\rangle|\} > \epsilon/2$.
  end for
  return The $k$ cluster centers.
```

For simplicity we only prove this when the initialization procedure in Theorem 2 takes polynomial time, namely $k \le O(d)$ and $w_{\max}/w_{\min} = O(1)$. In this case, we choose the thresholds ϵ 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 $\psi = \|\Psi\| < o(1)$.

The key observation here is if $T(\hat{a}, \hat{b}, \hat{c})$ is large, the vectors must be close to some a_i, b_i, c_i.

Lemma 14. *Suppose $w_{\max} = w_1 \ge w_2 \ge \cdots \ge w_k = w_{\min}$. If $\max\{|\langle a_i, \hat{a}\rangle|, |\langle b_i, \hat{b}\rangle|, |\langle c_i, \hat{c}\rangle|\} \le \epsilon$ for all $i < t$, let $\delta = O(w_{\max}\epsilon^{9-3p}/w_{\min})$ and $|T(\hat{a}, \hat{b}, \hat{c})| \ge (1-\delta)w_t$, then there exists a column j such that $\max\{\mathrm{dist}(\hat{a}, a_j), \mathrm{dist}(\hat{b}, b_j)\} < w_{\min}/10w_{\max}$.*

Proof: Separate T as $T_1 + T_2$, where T_1 contains all the terms from 1 to $t-1$, and T_2 contains the rest of the terms.

Using Corollary 4 we know $|T_1(\hat{a}, \hat{b}, \hat{c})| \le w_{\max}\|A_{[t-1]}^\top \hat{a}\|_3 \|B_{[t-1]}^\top \hat{b}\|_3 \|C_{[t-1]}^\top \hat{c}\|_3$. On the other hand, $\|A_{[t-1]}^\top \hat{a}\|_3^3 \le \max_{i<t} |\langle a_i, \hat{a}\rangle|^{3-p} \|A_{[t-1]}^\top \hat{a}\|_p^p \le O(\epsilon)^{3-p}$ (similarly for b, c). Hence, $|T_1(\hat{a}, \hat{b}, \hat{c})| \le w_{\max} O(\epsilon)^{9-3p} \le w_t\delta$, and we must have $|T_2(\hat{a}, \hat{b}, \hat{c})| \ge (1-2\delta)w_t$.

Again we apply Corollary 4 to show $|T_2(\hat{a}, \hat{b}, \hat{c})| \le w_t\|A^\top \hat{a}\|_3 \|B^\top \hat{b}\|_3 \|C^\top \hat{c}\|_3$. Since all the 3-norms are bounded by $1 + o(1)$, each one of them must be at least $1 - O(\delta)$.

Now we have $1 - O(\delta) \le \sum_{j=1}^k |\langle a_j, \hat{a}\rangle|^3 \le \max\{|\langle a_j, \hat{a}\rangle|\}^{3-p} \sum_{t=1}^k |\langle a_j, \hat{a}\rangle|^p \le (1+o(1)) \max\{|\langle a_j, \hat{a}\rangle|\}^{3-p}$. This implies $\max\{|\langle a_j, \hat{a}\rangle|\} = 1 - O(\delta)$, which in turn implies there exists a j such that $\mathrm{dist}(\hat{a}, a_j) < w_{\min}/10w_{\max}$ when ϵ, δ are small enough.

By symmetry we know there is also a j' such that $\mathrm{dist}(\hat{b}, b_{j'}) < w_{\min}/10w_{\max}$. If $j \ne j'$, 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 successful initialization, the largest $T(\hat{a}, \hat{b}, \hat{c})$ must be large.

Lemma 15. *If there is a successful initialization for column t, then the corresponding $\hat{a}, \hat{b}, \hat{c}$ have $|T(\hat{a}, \hat{b}, \hat{c})| > w_t(1-\delta)$, and for any $i \ne t$, $\max\{|\langle \hat{a}, a_i\rangle|, |\langle \hat{b}, b_i\rangle|, |\langle \hat{c}, c_i\rangle|\} \ll \epsilon$.*

Proof: We again separate the tensor T as $T_1 = w_t(a_t \otimes b_t \otimes c_t)$ and $T_2 = T - T_1$. Since the initialization is good, by the local convergence result we know $\mathrm{dist}(\hat{a}, a_t) < \tilde{O}(w_{\max}\sqrt{k}/w_{\min}d) \ll \delta$. Therefore, $|T_2(\hat{a}, \hat{b}, \hat{c})| \ge w_t(1-\delta/2)$.

On the other hand, using Corollary 4 we know $|T_2(\hat{a}, \hat{b}, \hat{c})| \le w_t\delta/2$, so $|T(\hat{a}, \hat{b}, \hat{c})| \ge |T_1(\hat{a}, \hat{b}, \hat{c})| - |T_2(\hat{a}, \hat{b}, \hat{c})| \ge w_t(1-\delta)$.

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

Finally we prove the clustering process succeeds.

Lemma 16. *There are at least one successful initialization for every component (which happens whp). Algorithm 3 outputs k cluster centers that are $\tilde{O}(w_{\max}\sqrt{k}/w_{\min}d)$ 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 $\max\{|\langle a_i, \hat{a}\rangle|, |\langle b_i, \hat{b}\rangle|, |\langle c_i, \hat{c}\rangle|\} \le \epsilon$ for all $i < t$. By Lemma 15 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 14 we know the 4-tuple we found must satisfy $\max\{\mathrm{dist}(\hat{a}, a_j), \mathrm{dist}(\hat{b}, b_j)\} < w_{\min}/10w_{\max}$ for some j (and this cannot be some j that has already been found). This tuple then satisfies the conditions of Theorem 1 (the local convergence result). 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.

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization. *Available on arXiv:1310.7991*, Oct. 2013.

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

A. Anandkumar, D. Hsu, and S.M. Kakade. A Method of Moments for Mixture Models and Hidden Markov Models. In *Proc. of Conf. on Learning Theory*, June 2012b.

A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y. K. Liu. Two SVDs Suffice: Spectral Decompositions for Probabilistic Topic Modeling and Latent Dirichlet Allocation. *to appear in the special issue of Algorithmica on New Theoretical Challenges in Machine Learning*, July 2013a.

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 2013b.

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. 2013c.

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

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

Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan 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.

Adam Coates, Honglak Lee, and Andrew Y. Ng. An analysis of single-layer networks in unsupervised feature learning. *Journal of Machine Learning Research - Proceedings Track*, 15:215–223, 2011.

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

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

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

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.

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

Gene H. Golub and Charles F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

Navin Goyal, Santosh Vempala, and Ying 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. 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.

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

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

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.

Quoc V. Le, Alexandre Karpenko, Jiquan Ngiam, and Andrew Y. Ng. ICA with Reconstruction Cost for Efficient Overcomplete Feature Learning. In *NIPS*, pages 1017–1025, 2011.

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

Brian McWilliams, David Balduzzi, and Joachim 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.

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

Terence Tao and Van Vu. Random matrices: The distribution of the smallest singular values. *Geometric And Functional Analysis*, 20(1):260–297, 2010.

M Alex O Vasilescu and Demetri 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.

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