JMLR: Workshop and Conference Proceedings vol 40:1–77, 2015

Learning Overcomplete Latent Variable Models through Tensor Methods

Animashree Anandkumar A.ANANDKUMAR@UCI.EDU
Department of Electrical Engineering and Computer Science
University of California, Irvine
Engineering Hall, #4408
Irvine, CA 92697, USA

Rong Ge RONGGE@MICROSOFT.COM
Microsoft Research
One Memorial Drive
Cambridge, MA 02142, USA

Majid Janzamin MJANZAMI@UCI.EDU
Department of Electrical Engineering and Computer Science
University of California, Irvine
Engineering Hall, #4406
Irvine, CA 92697, USA

Abstract

We provide guarantees for learning latent variable models emphasizing on the overcomplete regime, where the dimensionality of the latent space exceeds the observed dimensionality. In particular, we consider multiview mixtures, ICA, and sparse coding models. Our main tool is a new algorithm for tensor decomposition that works in the overcomplete regime.

In the semi-supervised setting, we exploit label information to get a rough estimate of the model parameters, and then refine it using the tensor method on unlabeled samples. We establish learning guarantees when the number of components scales as $k = o(d^{p/2})$, where d is the observed dimension, and p is the order of the observed moment employed in the tensor method (usually $p = 3, 4$). In the unsupervised setting, a simple initialization algorithm based on SVD of the tensor slices is proposed, and the guarantees are provided under the stricter condition that $k \leq \beta d$ (where constant β can be larger than 1). For the learning applications, we provide tight sample complexity bounds through novel covering arguments.

Keywords: unsupervised and semi-supervised learning, latent variable models, overcomplete representations, tensor decomposition.

1. Introduction

Tensor decompositions have been popular for unsupervised learning of a wide range of latent variable models (LVMs) such as topic models, Gaussian mixtures, independent component analysis, network community models, and so on (Anandkumar et al., 2014a, 2013b; De Lathauwer et al., 2007). It involves decomposition of a certain low order multivariate moment tensor (typically up to fourth order), and is guaranteed to provide a consistent estimate of the model parameters. In practice, the tensor decomposition techniques have been effective in a number of applications such as

blind source separation (Comon, 2002), computer vision (Vasilescu and Terzopoulos, 2003), topic modeling (Zou et al., 2013), and community detection (Huang et al., 2013).

The state of art for guaranteed tensor decomposition involves two steps: converting the input tensor to an orthogonal symmetric form, and then solving the orthogonal decomposition through tensor eigen decomposition (Comon, 1994; Zhang and Golub, 2001; Anandkumar et al., 2014a). While having efficient guarantees, this approach is unable to learn *overcomplete representations*, where the latent dimensionality exceeds the observed dimensionality. This is especially limiting given the recent popularity of overcomplete feature learning in many domains, e.g., see Bengio et al. (2012); Lewicki and Sejnowski (2000). Overcomplete representations also provide flexible modeling, and are robust to noise (Lewicki and Sejnowski, 2000).

In this paper, we establish guarantees for tensor decomposition in learning overcomplete LVMs, such as multiview mixtures, independent component analysis (ICA), Gaussian mixtures and sparse coding models. Note that learning general overcomplete models is ill-posed since the latent dimensionality exceeds the observed dimensionality. We impose a natural incoherence condition on the components, which can be viewed as a *soft orthogonality* constraint, and limits the redundancy among the components. Incoherence constraints are natural in the overcomplete regime, and have been considered before, e.g., in compressed sensing (Donoho, 2006), independent component analysis (Le et al., 2011), and sparse coding (Arora et al., 2013; Agarwal et al., 2013).

1.1. Summary of Results

In this paper, we provide semi-supervised and unsupervised learning guarantees for LVMs such as multiview mixtures, ICA and sparse coding models. Our algorithm is based on method of moments, and employs a tensor decomposition algorithm for learning. One of our main contributions is the convergence analysis of a new tensor decomposition algorithm that works in the overcomplete regime. Under the semi-supervised setting, we establish that highly overcomplete models can be learned efficiently through the tensor decomposition method. The moment tensors are constructed using unlabeled samples, and the labeled samples are used to provide a rough initialization to the tensor decomposition algorithm. In the unsupervised setting, we propose a simple initialization strategy for the tensor method, and can handle mildly overcomplete models. In both settings we provide tight sample complexity bounds through novel covering arguments.

1.1.1. OVERCOMPLETE TENSOR DECOMPOSITION GUARANTEES

We employ a tensor decomposition algorithm for learning. Given rank-k tensor

$$T = \sum_{i\in[k]} w_i \cdot a_i \otimes b_i \otimes c_i, \quad w_i \in \mathbb{R},\ a_i, b_i, c_i \in \mathbb{R}^d, \tag{1}$$

the goal is to recover its rank-1 components $\{(a_i, b_i, c_i), i \in [k]\}$. Here, $\otimes$ denotes the tensor outer product; see Section 3.1 for the details of tensor notations and tensor rank. An overview of our tensor decomposition algorithm is provided in Figure 2. The main step of our tensor decomposition algorithm is *alternating asymmetric tensor power update*; see(11) for this update. We provide robust analysis of the algorithm leading to local and global convergence guarantees when the input tensor is noisy.[1] Our analysis emphasizes on the challenging *overcomplete* regime where the tensor rank is larger than the dimension, i.e., $k > d$.

1. Note that in the learning applications, we form the empirical moments as the input tensor which is noisy.

We require natural deterministic conditions on the tensor components to argue the convergence guarantees; see Appendix G for the details. All of these conditions are satisfied if the true rank-1 components of the tensor are uniformly i.i.d. drawn from the unit d-dimensional sphere $\mathcal{S}^{d-1}$. Among the deterministic assumptions, the most important one is the *incoherence* condition which imposes a soft-orthogonality constraint between different rank-1 components of the tensor.

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

Theorem 1 (Local convergence guarantee of the tensor decomposition algorithm) *Consider noisy rank-k tensor $\widehat{T} = T + \Psi$ as the input to the tensor decomposition algorithm, where $T = \sum_{i\in[k]} w_i \cdot a_i \otimes b_i \otimes c_i$, and $\psi := \|\Psi\| \leq w_{\min}/6$. Let the rank condition $k \leq o\left(d^{1.5}\right)$ is satisfied. Assuming we have good initialization vectors (which have constant error with the true components), then the algorithm outputs estimates $\widehat{A} := [\widehat{a}_1 \cdots \widehat{a}_k] \in \mathbb{R}^{d\times k}$ and $\widehat{w} := [\widehat{w}_1 \cdots \widehat{w}_k]^\top \in \mathbb{R}^k$, satisfying w.h.p.*

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

Same error bounds hold for other factor matrices $B := [b_1 \cdots b_k]$ and $C := [c_1 \cdots c_k]$. The number of iterations is $N = \Theta\left(\log(1/\hat{\epsilon}_R)\right)$ where $\hat{\epsilon}_R := \min\{\psi/w_{\min}, \tilde{O}\left(\sqrt{k}/d\right)\}$.

Thus, we can decompose the tensor in the highly overcomplete regime $k \leq o\left(d^{1.5}\right)$. The $\sqrt{k}$ factor in the bound is from the fact that the final recovery guarantee is on the Frobenius norm of the whole factor matrix A. In the following, we provide stronger column-wise guarantees (where there is no $\sqrt{k}$ factor) with the expense of having an additional residual error term. Our algorithm includes two main update steps including tensor power iteration in (11) and residual error removal in (12). The guarantee for the first step — tensor power iteration — is

Lemma 2 (Local convergence guarantee of the tensor power updates) *Consider the same settings as in Theorem 1. Then, the outputs of tensor power iteration steps in our algorithm satisfy w.h.p.*[2]

$$\min_{z\in\{-1,1\}} \|z\widehat{a}_j - a_j\| \leq \tilde{O}\left(\frac{\psi}{w_{\min}}\right) + \tilde{O}\left(\frac{\sqrt{k}}{d}\right), \quad |\widehat{w}_j - w_j| \leq \tilde{O}\left(\psi\right) + \tilde{O}\left(w_{\max}\frac{\sqrt{k}}{d}\right), \quad j \in [k].$$

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

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

For the *global convergence* guarantee, we obtain good initialization vectors by performing a rank-1 SVD on the random slices of the moment tensor.

2. Note that recovery of components is up to sign. This is because a third order tensor is unchanged if the sign along one of the modes is fixed and the signs along the other two modes are flipped.

Theorem 3 (Global convergence guarantee of the tensor decomposition algorithm) *Consider the same input tensor to the algorithm as in Theorem 1 with noise bound* $\psi := \|\Psi\| \leq \tilde{O}\big(w_{\min}/\sqrt{d}\big)$. *Let* $k \leq \beta d$ *(for arbitrary constant* $\beta > 1$*), and the initialization is performed by SVD-based method in Procedure 3 (in the Appendix) using a polynomial number of initializations scaled as* k^{β^2}. *Then, the same guarantees as in Theorem 1 hold.*

Note that the argument in Lemma 2 can be similarly adapted leading to global convergence guarantee of the tensor power iteration step.

For 4th and higher order tensors, same techniques can be exploited to argue similar results. These new tensor decomposition guarantees combined with the new tensor concentration bounds we derive are the key to prove our learning results, where the local convergence guarantee leads to the semi-supervised learning and the global convergence guarantee leads to the unsupervised learning results.

The above tensor decomposition results can be naturally applied to many settings like mulitiview linear mixtures, ICA, and so on. But, there is a caveat as the incoherence condition is not natural for models like topic models where there is a non-negativity constraint on the hidden components.

1.1.2. LEARNING MULTIVIEW MIXTURE MODEL

In the multiview mixtures model, given the hidden mixture component, each observation (view) is independently drawn with some unknown mean parameter and noise distribution around that mean; see (4) for the precise form. The goal is to estimate the conditional mean parameters. In this setting, we assume reasonable property on noise, and for brevity, we consider the "low" noise regime (where the norm of noise is of the same order as that of the component means).

In the *semi-supervised* setting, we use labeled samples to initialize the tensor decomposition algorithm, and provide the following recovery guarantee.

Theorem 4 (Semi-supervised learning of multiview mixtures model: informal) *Let* k *be the number of mixture components, and* d *be the observed dimensionality, and suppose* $k \leq o(d^{1.5})$. *We show that having* $\mathrm{polylog}(d, k)$ *number of labeled samples for each label, and* $n \geq \tilde{\Omega}(k)$ *number of unlabeled samples are sufficient to consistently estimate the model parameters.*

See Theorem 7 for the formal statement of this result. Thus, for recovering each rank-1 component, we need far less number of labeled samples compared to the number of unlabeled samples required. 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). Furthermore, note that the unlabeled sample complexity is the *minimax* bound up to $\mathrm{polylog}$ factors.

We also provide *unsupervised* learning guarantees when no label is available. Here, the initialization is performing by the SVD-based method stated in the previous section. This imposes additional conditions on rank and sample complexity as follows.

Theorem 5 (Unsupervised learning of multiview mixtures model: informal) *Suppose the number of unlabeled samples* n *satisfies* $n \geq \tilde{\Omega}\left(kd\right)$. *If* $k \leq \beta d$ *(for arbitrary constant* $\beta > 1$*), then the model parameters can be learned using a polynomial number of initializations scaled as* k^{β^2}.

See Theorem 10 for the formal statement of this result. This result is an improvement over existing results since we do not have dependence on the condition number of the component means and in addition, we can handle overcomplete models.

1.1.3. LEARNING ICA AND SPARSE ICA MODELS

We also provide semi-supervised and unsupervised learning guarantees for *Independent Component Analysis* (ICA). By semi-supervised setting in ICA, we mean some prior information is available which provides good initializations for the tensor decomposition algorithm. In the *semi-supervised* setting, we show that when the number of components $k = \Theta(d^2)/\operatorname{polylog}(d)$, the ICA model can be efficiently learned from fourth order moments with $n \geq \tilde{\Omega}(k^{2.5})$ number of unlabeled samples. In the *unsupervised* setting, we show that when $k \leq \beta d$ (for arbitrary constant $\beta > 1$), the ICA model can be learned with number of samples scaling as $n \geq \tilde{\Omega}(k^3)$ in k^{β^2} number of initializations.

We also provide learning results for the *sparse coding* model, when the mixing coefficients are independently drawn from a Bernoulli-Gaussian distribution and the dictionary satisfies some deterministic conditions (see Appendix G and (RIP) in Section 2.1). Notice this corresponds to a sparse ICA model since the hidden coefficients are independent.

Theorem 6 (Learning (sparse) ICA: informal) *We can efficiently estimate the dictionary in the (sparse) ICA model under the following conditions. Let s be the expected sparsity of the hidden coefficients. In the* semi-supervised *setting (where prior information provides us good initialization), we need the number of components to be bounded by $k = o(d^2)$, and unlabeled sample complexity satisfies $n \geq \tilde{\Omega}(\max\{sk, s^2k^2/d^3\})$. In the* unsupervised *setting, we need $k = \Theta(d)$, and $n \geq \tilde{\Omega}(k^2 s)$.*

In the special case when s is a constant, the sample complexity is akin to learning multiview models, and when $s = \Theta(k)$, it is akin to learning the "dense" ICA model. Thus, the sparse coding model bridges the range of models between multiview mixtures model and ICA. See Theorem 12 for the formal statement of above result on learning sparse ICA. Since dense ICA is a special case, its detailed results are provided in the Appendix in Theorems 53 and 54.

1.2. Related Works

Several latent variable models can be learned through tensor decomposition including independent component analysis (De Lathauwer et al., 2007), topic models, Gaussian mixtures, hidden Markov models (Anandkumar et al., 2014a) and network community models (Anandkumar et al., 2013b). In the undercomplete setting, Anandkumar et al. (2014a) analyze robust tensor power iteration for learning LVMs, and Song et al. (2013) extend analysis to the nonparametric setting. These works require the tensor factors to have full column rank, which rules out overcomplete models. Moreover, they require whitening the input data, and the sample complexity depends on the condition number of the factor matrices. For instance, when $k = d$, for random factor matrices, the previous tensor approaches in (Song et al., 2013; Anandkumar et al., 2013a) have a sample complexity of $\tilde{\Omega}(k^{6.5})$. Our result can be also extended to learning mixtures of spherical Gaussians, where we have better sample complexity than the work by Hsu and Kakade (2012) (we have $\tilde{\Omega}(d^2)$ instead of their $\tilde{\Omega}(d^3)$ when $k = d$). Note that this comparison is in the low noise regime (where the norm of noise is of the same order as that of the component means). Thus, we provide the best known sample bounds for semi-supervised and unsupervised learning of multiview mixtures model in the overcomplete setting, assuming incoherent components.

In general, learning overcomplete models is challenging, and they may not even be identifiable in general. The FOOBI procedure by De Lathauwer et al. (2007) shows that a polynomial-time procedure can recover the components of ICA model (with *generic* factors) when $k = O(d^2)$, where

the moment is fourth order. However, the procedure does not work for third-order overcomplete tensors. For the fifth order tensor, Goyal et al. (2013); Bhaskara et al. (2013) perform simultaneous diagonalization on the matricized versions of random slices of the tensor and provide careful perturbation analysis. But, this procedure cannot handle the same level of overcompleteness as FOOBI. In addition, Goyal et al. (2013) provide stronger results for ICA, where the tensor slices can be obtained in the Fourier domain. Given 4th order tensor, they need $\mathrm{poly}(k^4)$ number of unlabeled samples for learning ICA (where the poly factor is not explicitly characterized), while we only need $\tilde{\Omega}(k^{2.5})$ (when $k = \Theta(d^2)/\operatorname{polylog}(d)$).

More discussions on related works in provided in Appendix A.

Notations: Let $[n] := \{1, 2, \ldots, n\}$. Let $\|\cdot\|$ and $\|\cdot\|_F$ respectively denote the spectral and Frobenius norms. We also use $\tilde{O}$ and $\tilde{\Omega}$ to hide $\mathrm{polylog}$ factors in O and Ω notations, respectively.

2. Learning Latent Variable Models

In this section, we provide our main results on learning different LVMs including multiview mixtures, mixture of Gaussians, ICA, and sparse coding. We establish sample complexity bounds for all the LVMs.[3] The details of the learning algorithm is provided in Section 3.

We consider two learning settings: 1) semi-supervised setting where a small amount of label information is available, and 2) unsupervised setting where such information is not available. In the former setting, we can handle overcomplete mixtures with number of components $k = o(d^{p/2})$, where d is the observed dimension and p is the order of observed moment. In the latter case, our analysis works when $k \leq \beta d$ for any constant β.

Here, we review some of the assumptions and settings throughout the section. Consider tensor decomposition form in (1). Let $A := [a_1\ a_2\ \cdots\ a_k] \in \mathbb{R}^{d\times k}$ denote the *factor matrix*. Similar factor matrices are defined as B and C in the asymmetric cases, e.g., multiview mixtures model. Without loss of generality, we assume that the columns of factor matrices have unit ℓ_2 norm, since we can always rescale them, and adjust the weights appropriately. We also require natural deterministic conditions on the tensor components, but for simplicity we assume $a_i, b_i, c_i \in \mathbb{R}^d, i \in [k]$, are uniformly i.i.d. drawn from the unit d-dimensional sphere $\mathcal{S}^{d-1}$ (see Remark 8). For brevity, we also assume the ratio between largest and smallest weights to be a constant: $\frac{w_{\max}}{w_{\min}} \leq O(1)$. For the sake of saving the notations, we only provide the learning results in the challenging overcomplete regime where the number of components/mixtures is larger than observed dimension, i.e., $k \geq \Omega(d)$. But the results can be easily adapted to the easier highly undercomplete regime when $k \leq o(d)$.

2.1. Multiview Mixtures Model

Consider a multiview mixtures model with k components and $p \geq 3$ views; see Figure 1. Throughout the paper, we assume $p = 3$ for simplicity, while the results can be also extended to higher-order observations. Suppose that hidden variable $h \in [k]$ is a discrete categorical random variable with $\Pr[h = j] = w_j, j \in [k]$. The variables (views) $x_l \in \mathbb{R}^d$ are conditionally independent given the k-categorical latent variable $h \in [k]$, and the conditional means are

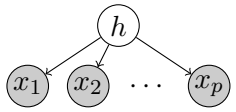

Figure 1: Multiview mixtures.

$$\mathbb{E}[x_1|h] = a_h, \quad \mathbb{E}[x_2|h] = b_h, \quad \mathbb{E}[x_3|h] = c_h, \tag{2}$$

3. The proof of theorems in this section are provided in Appendix K.

where $A := [a_1\ a_2\ \cdots\ a_k] \in \mathbb{R}^{d\times k}$ denotes the *factor matrix* and B, C are similarly defined. The goal of the learning problem is to recover the parameters of the model (factor matrices) A, B, and C given observations x_l's.

For this model, the third order observed moment can be written as (Anandkumar et al., 2014a)

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{j\in[k]} w_j a_j \otimes b_j \otimes c_j. \tag{3}$$

Hence, given third order observed moment, the unsupervised learning problem (recovering factor matrices A, B, and C) reduces to computing a tensor decomposition as in (3).

Given hidden state h, suppose the observed variables $x_l \in \mathbb{R}^d$ have conditional distributions as

$$x_1|h \sim a_h + \zeta\sqrt{d}\cdot\varepsilon_A, \quad x_2|h \sim b_h + \zeta\sqrt{d}\cdot\varepsilon_B, \quad x_3|h \sim c_h + \zeta\sqrt{d}\cdot\varepsilon_C, \tag{4}$$

where the noise vectors $\varepsilon_A, \varepsilon_B, \varepsilon_C \in \mathbb{R}^d$ are independent random vectors with zero mean and covariance $\frac{1}{d}I_d$, and ζ^2 is a scalar denoting the variance of each entry. We also assume that noise vectors $\varepsilon_A, \varepsilon_B, \varepsilon_C$ are independent of hidden vector h.

In addition, we assume the noise matrices satisfy RIP property as follows. Given n samples for the model, define noise matrix $E_A := [\varepsilon_A^{(1)}, \varepsilon_A^{(2)}, \ldots, \varepsilon_A^{(n)}] \in \mathbb{R}^{d\times n}$, where $\varepsilon_A^{(i)} \in \mathbb{R}^d$ is the i-th sample of noise vector ε_A. E_B and E_C are similarly defined. These matrices need to satisfy the following RIP property which is adapted from Candes and Tao (2006).

(RIP) Matrix $E \in \mathbb{R}^{d\times n}$ satisfies a weak RIP condition such that for any subset of $O\left(\frac{d}{\log^2 d}\right)$ number of columns, the spectral norm of E restricted to those columns is bounded by 2.

It is known that when $n = \mathrm{poly}(d)$, the above condition is satisfied w.h.p. for many random models such as when the entries are i.i.d. zero mean Gaussian or Bernoulli random variables.

For brevity, we consider the low noise regime where the expected norm of noise vector is bounded by a constant, i.e., $\zeta^2 d = O(1)$. Note that since model parameters $a_i, b_i, c_i, i \in [k]$, have unit norm, low noise regime imposes that the expected norm of noise is in the same order of norm of model parameters. The results for the general high noise regime is provided in Appendix L. In addition, since w_j's are the mixture probabilities, for simplicity we consider $w_j = \Theta(1/k), j \in [k]$.

2.1.1. SEMI-SUPERVISED LEARNING

In the semi-supervised setting, label information is exploited to build good initialization vectors for the tensor decomposition algorithm as follows. Let $x_{1,j}^{(l)}, x_{2,j}^{(l)}, x_{3,j}^{(l)} \in \mathbb{R}^d, j \in [k], l \in [m_j]$, denote $m = \sum_{j\in[k]} m_j$ labeled samples, where the samples with subscript j have label j, i.e., they are generated from hidden state $h = j$. Then, given conditional mean model in (2), we can compute the empirical estimate of mixture components as

$$\widehat{a}_j := \frac{1}{m_j}\sum_{l\in[m_j]} x_{1,j}^{(l)}, \quad \widehat{b}_j := \frac{1}{m_j}\sum_{l\in[m_j]} x_{2,j}^{(l)}, \quad \widehat{c}_j := \frac{1}{m_j}\sum_{l\in[m_j]} x_{3,j}^{(l)}, \quad \text{for any } j \in [k]. \tag{5}$$

We first provide the settings of learning algorithm which include input tensor T, number of iterations N and the initialization setting.

Settings of Algorithm in Theorem 7: Given n unlabeled samples $x_1^{(i)}, x_2^{(i)}, x_3^{(i)} \in \mathbb{R}^d, i \in [n]$, consider the empirical estimate of 3rd order moment in (3) as the input to the algorithm in Figure 2.

Let the number of iterations $N = \tilde{\Theta}\left(\log\left(1/\epsilon_R\right)\right)$ where $\epsilon_R := \min\{\sqrt{k/n}, \sqrt{k}/d\}$. Exploit the empirical estimates in (5) as initialization vectors.

Theorem 7 (Semi-supervised learning of multiview mixtures model) *Assume the Algorithm settings mentioned above hold. Suppose the number of labeled samples with label $j \in [k]$, denoted by m_j, and the number of unlabeled samples n satisfy $m_j \geq \tilde{\Omega}\left(1\right)$, $n \geq \tilde{\Omega}\left(k\right)$. If* rank condition *$\Omega(d) \leq k \leq o(d^{1.5})$ holds, then the algorithm outputs estimates $\widehat{A} := [\widehat{a}_1 \cdots \widehat{a}_k] \in \mathbb{R}^{d\times k}$ and $\widehat{w} := [\widehat{w}_1 \cdots \widehat{w}_k]^\top \in \mathbb{R}^k$, satisfying w.h.p.*

$$\left\|\widehat{A} - A\right\|_F \leq \tilde{O}\left(\frac{k}{\sqrt{n}}\right), \quad \|\widehat{w} - w\| \leq \tilde{O}\left(\frac{1}{\sqrt{n}}\right). \tag{6}$$

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

Thus, we provide efficient learning guarantees for overcomplete multiview mixtures in the semi-supervised setting given small number of labeled samples. It is also worth mentioning that there is no dependence on the condition numbers of moment matrices in the sample complexity result.

Column-wise error bounds: In Section 1.1.1, we explain that the algorithm analysis also provides column-wise error bounds with the expense of introducing an additional approximation error. More precisely, we provide stronger guarantees on the column-wise errors as

$$\|\widehat{a}_j - a_j\| \leq \tilde{O}\left(\sqrt{k/n}\right) + \tilde{O}\left(\sqrt{k}/d\right), \quad j \in [k], \tag{7}$$

where a $\sqrt{k}$ factor is removed in the first term of bound comparing with the bound in (6), but an additional approximation error $\tilde{O}\left(\sqrt{k}/d\right)$ is introduced. See Lemma 2 and the corresponding discussions for exact description.

Remark 8 (Random assumption) In the above learning result, we assume that the mixture components are uniformly i.i.d. drawn from unit d-dimensional sphere $\mathcal{S}^{d-1}$. This assumption is provided for simplicity, while the original conditions for the recovery guarantees are deterministic (provided in Appendix G). We show that random matrices satisfy these deterministic assumptions with high probability. Notice the random assumption is reasonable for continuous models including the multiview mixtures model described here. But, it is not appropriate for discrete models where the non-negativity assumptions on the entries of factor matrices are required.

Remark 9 (Spherical Gaussian mixtures) Similar learning results as in Theorem 7 hold for the spherical Gaussian mixtures. Note that in Appendix D, it is provided how learning this model can be reduced to the tensor decomposition problem.

2.1.2. UNSUPERVISED LEARNING

In the unsupervised setting, there is no label information available to build the initialization vectors. Here, the initialization is performed by doing rank-1 SVD on random slices of the moment tensor.

Settings of Algorithm in Theorem 10: Consider the same settings as in Theorem 7 with new initialization method as follows. The initialization in each run of Algorithm 1 is performed by SVD-based technique in Procedure 3, with the number of initializations as $L \geq k^{\Omega(k^2/d^2)}$.

Theorem 10 (Unsupervised learning of multiview mixtures model) *Assume the Algorithm settings mentioned above hold. Suppose the number of unlabeled samples n satisfies $n \geq \tilde{\Omega}(kd)$. If rank condition $k = \Theta(d)$ holds, then the same guarantees as in Theorem 7 are satisfied.*

We now compare the sample complexity in the above theorem with the previous result by Song et al. (2013), which employs whitening procedure followed by tensor power updates in the undercomplete setting. When $k \approx d$, the sample complexity in (Song et al., 2013) is scaled as $n \geq \tilde{\Omega}(k^{6.5})$. In comparison, the sample complexity for our method scales as $\tilde{\Omega}(k^2)$, which is far 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.

The above unsupervised learning result can be also adapted for learning mixture of spherical Gaussians. An algorithm for learning mixture of spherical Gaussians in the undercomplete setting is provided in (Hsu and Kakade, 2012), which is a moment-based technique combined with a whitening step. When $k = d$, the sample complexity in (Hsu and Kakade, 2012) scales as $n \geq \tilde{\Omega}(d^3)$. But, our tight tensor concentration analysis leads to the better sample complexity of $n \geq \tilde{\Omega}(d^2)$.

Remark 11 (Extension to $k \leq o(d^{1.5})$) We also argue that the SVD initialization can be slightly modified, and under some regime of noise we can extend the above unsupervised learning result to the highly overcomplete regime $k \leq o(d^{1.5})$. Suppose the expected norm of noise is constant (low noise regime), and the noise vectors are incoherent with the true mean components (which is satisfied for random mean components). Then if the SVD initialization is performed using samples[4] $x_3^{(i)}$, then the same guarantees as in Theorem 10 hold under highly overcomplete regime $k \leq o(d^{1.5})$.

2.2. Independent Component Analysis (ICA) and Sparse ICA

In this section, we propose the semi-supervised and unsupervised learning results for the ICA and sparse ICA models. By semi-supervised setting in ICA, we mean some prior information is available which provides good initializations for the components. Recall the standard ICA model (Comon, 1994), where *independent* source signals are linearly mixed to generate the observations. Let $h \in \mathbb{R}^k$ be a random latent signal where its coordinates are independent, and $A \in \mathbb{R}^{d \times k}$ be the mixing matrix. Then, the observed vector is

$$x = Ah \in \mathbb{R}^d.$$

For simplicity, we limit to noiseless setting. This is the standard setting, and is already challenging because samples in ICA are mixtures of many components, unlike the mixture models. It is discussed in Appendix D how estimating the parameters of ICA model can be formulated as a tensor decomposition problem where a modified version of 4th order observed moment (denoted by M_4) is characterized in a tensor decomposition form; see Lemma 15 in the appendix.

We now provide the learning results for the *sparse* ICA problem which is more general. This is the ICA setting with the assumption that hidden vector $h \in \mathbb{R}^k$ can be sparse with i.i.d. Bernoulli-subgaussian random entries. Assume the probability of each Bernoulli variable being 1 is s/k. Note that (dense) ICA is special case when $s = k$. For the sparse ICA model, we also assume that mixing matrix A satisfies the RIP property (see condition (RIP) in Section 2.1).

4. The SVD of $T(I, I, x_3^{(i)})$ is computed; see Procedure 3 in Appendix E for the details.

Settings of Algorithm in Theorem 12: Given n samples $x^i = Ah^i, i \in [n]$, consider the empirical estimate of 4th order (modified) moment M_4 (see (17) in the Appendix) as the input to the algorithm with symmetric 4th order updates; see Appendix F.1.1 for higher order extension of the algorithm. Let the number of iterations $N = \tilde{\Theta}\left(\log\left(1/\tilde{\epsilon}_R\right)\right)$, where $\tilde{\epsilon}_R := \min\{k^2/\min\{n, \sqrt{d^3 n}\}, \sqrt{k}/d^{1.5}\}$. The initialization is performed differently in different learning settings. In the *semi-supervised* setting, it is assumed that for any $j \in [k]$, an approximation of a_j denoted by $\widehat{a}_j^{(0)}$ is given satisfying $\|\widehat{a}_j^{(0)} - a_j\| \leq \alpha$ for some constant $\alpha < 1$. In the *unsupervised* setting, the initialization is performed by 4-th order generalization [5] of SVD-based technique in Procedure 3, with the number of initializations as $L \geq k^{\Omega(k^2/d^2)}$.

Theorem 12 (Semi-supervised and unsupervised learning of (sparse) ICA) *Assume the Algorithm settings mentioned above hold. In the* semi-supervised *setting, suppose*

$$n \geq \begin{cases} \tilde{\Omega}(sk), & sk \leq O(d^3)/\operatorname{polylog}(d), \\ \tilde{\Omega}\left(s^2k^2/d^3\right), & \text{o.w.}, \end{cases}$$

and rank condition $\Omega(d) \leq k \leq o(d^2)$ hold. In the unsupervised *setting, suppose $n \geq \tilde{\Omega}\left(k^2 s\right)$, and rank condition $\Omega(d) = \Theta(d)$ hold. Then the algorithm outputs estimates $\widehat{A}$ and $\widehat{w}$, satisfying w.h.p.*

$$\max\left\{\left\|\widehat{A} - A\right\|_F, \|\widehat{w} - w\|\right\} \leq \tilde{O}\left(\frac{s \cdot k^{1.5}}{\min\{n, \sqrt{d^3 n}\}}\right).$$

In one extreme when $s = \Theta(k)$, it is akin to learning the "dense" ICA model.[6] On the other extreme when s is a constant, it is akin to learning multiview models. Thus, the sparse coding model bridges the range of models between multiview mixtures model and ICA.

Similar to the multiview mixture model, we can also provide column-wise recovery guarantees with introducing additional approximation error $\tilde{O}\left(\sqrt{k}/d^{1.5}\right)$. Note that this error is different from multiview mixture since we exploit different tensor orders in the two models.

3. Algorithm

We first introduce tensor preliminaries, and then describe our tensor decomposition algorithm.

3.1. Tensor Preliminaries

A real *p-th order tensor* $T \in \bigotimes^p \mathbb{R}^d$ is a member of the outer product of Euclidean spaces $\mathbb{R}^d$. The different dimensions of the tensor are referred to as *modes*. For instance, for a matrix, the first mode refers to columns and the second mode refers to rows. In addition, *fibers* are higher order analogues of matrix rows and columns. A fiber is obtained by fixing all but one of the indices of the tensor (and is arranged as a column vector). For example, for a third order tensor $T \in \mathbb{R}^{d \times d \times d}$, the mode-1 fiber is given by $T(:, j, l)$. Similarly, *slices* are obtained by fixing all but two of the indices of the tensor. For example, for the third order tensor T, the slices along 3rd mode are given by $T(:, :, l)$.

5. In the 4th order case, the SVD is performed on $T(I, I, \theta, \theta) \in \mathbb{R}^{d \times d}$ for some random vector θ.
6. The complete results for learning ICA are provided in Appendix M which is a special case when $s = k$. Note that since we provide a different proof for the ICA model, it does not need the RIP condition on dictionary matrix.

We view a tensor $T \in \mathbb{R}^{d\times d\times d}$ as a *multilinear* operator. For vectors $u, v, w \in \mathbb{R}^d$, we have [7]

$$T(I, v, w) := \sum_{j,l\in[d]} v_j w_l T(:, j, l) \ \in \mathbb{R}^d, \tag{8}$$

which is a multilinear combination of the tensor mode-1 fibers. Similarly $T(u, v, w) \in \mathbb{R}$ is a multilinear combination of the tensor entries, and $T(I, I, w) \in \mathbb{R}^{d\times d}$ is a linear combination of the tensor slices. See (14) in the Appendix for the general definition of the multilinear form.

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

$$T = w \cdot a \otimes b \otimes c \Leftrightarrow T(i, j, l) = w \cdot a(i) \cdot b(j) \cdot c(l), \tag{9}$$

where $\otimes$ represents the *outer product* notation and $a, b, c \in \mathbb{R}^d$ are unit vectors. A tensor $T \in \mathbb{R}^{d\times d\times d}$ is said to have a CP *rank* k if it can be written as the (minimal) sum of k rank-1 tensors

$$T = \sum_{i\in[k]} w_i a_i \otimes b_i \otimes c_i, \quad w_i \in \mathbb{R},\ a_i, b_i, c_i \in \mathbb{R}^d. \tag{10}$$

3.2. Tensor Decomposition Algorithm

In this section, we introduce the tensor decomposition algorithm. The goal of tensor decomposition algorithm is to recover the rank-1 components of tensor; see (10) for the notion of tensor rank. Figure 2 depicts the overview of our tensor decomposition method where the corresponding algorithms and procedures are also specified. We now describe different steps of the algorithm as follows. For the complete details of algorithms and more discussions, see Appendix E.

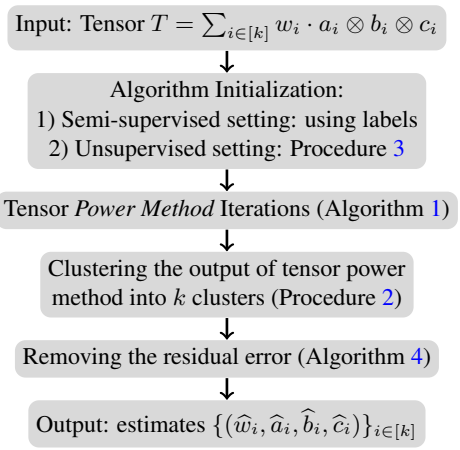

Figure 2: Overview of tensor decomposition algorithm. See Appendix E for algorithms and more discussions.

Tensor power iteration: The main step of the algorithm is tensor power iteration which basically performs alternating *asymmetric power updates* [8] on different modes of the tensor as (see (8) for the definition of multilinear form)

$$\widehat{a}^{(t+1)} = \frac{T\left(I, \widehat{b}^{(t)}, \widehat{c}^{(t)}\right)}{\left\|T\left(I, \widehat{b}^{(t)}, \widehat{c}^{(t)}\right)\right\|}, \ \widehat{b}^{(t+1)} = \frac{T\left(\widehat{a}^{(t)}, I, \widehat{c}^{(t)}\right)}{\left\|T\left(\widehat{a}^{(t)}, I, \widehat{c}^{(t)}\right)\right\|}, \ \widehat{c}^{(t+1)} = \frac{T\left(\widehat{a}^{(t)}, \widehat{b}^{(t)}, I\right)}{\left\|T\left(\widehat{a}^{(t)}, \widehat{b}^{(t)}, I\right)\right\|}, \tag{11}$$

where $\{\widehat{a}^{(t)}, \widehat{b}^{(t)}, \widehat{c}^{(t)}\}$ denotes estimate in the t-th iteration. Notice that the updates alternate among different modes of the tensor which can be viewed as a rank-1 form of the standard Alternating Least Squares (ALS) method.

We now provide an intuitive argument on the functionality of this step. Consider a rank-k tensor T as in (10), and suppose we start at the correct vectors $\widehat{a} = a_j$ and $\widehat{b} = b_j$, for some $j \in [k]$. Then, the power update in (11) leads to $T\big(\widehat{a}, \widehat{b}, I\big) = w_j c_j + \sum_{i\neq j} w_i \langle a_j, a_i\rangle\langle b_j, b_i\rangle c_i$, where the first

7. Compare with the matrix case where for $M \in \mathbb{R}^{d\times d}$, we have $M(I, u) = Mu := \sum_{j\in[d]} u_j M(:, j)$.
8. This is exactly the generalization of asymmetric matrix power update to 3rd order tensors.

term is along c_j and the second term is an error term due to non-orthogonality of the components. For orthogonal decomposition, the second term is zero, and thus the true vectors a_j, b_j and c_j are stationary points of the power update procedure. However, since we consider non-orthogonal tensors, this procedure cannot recover the decomposition exactly leading to a residual error after running this step. Under incoherence conditions which encourages soft-orthogonality[9] (and some other conditions), we show that the residual error is small (see Lemma 2 where the guarantees for the tensor power iteration step is provided). This enables us to remove this residual error with the following additional step.

Removing residual error: We propose Algorithm 4 to remove the additional residual error after tensor power iteration. This algorithm mainly runs a coordinate descent iteration as

$$\tilde{c}_i^{(t+1)} = \mathrm{Norm}\Big(T\left(\widehat{a}_i^{(t)}, \widehat{b}_i^{(t)}, I\right) - \sum_{j \neq i} \widehat{w}_j^{(t)} \langle \widehat{a}_i^{(t)}, \widehat{a}_j^{(t)} \rangle \langle \widehat{b}_i^{(t)}, \widehat{b}_j^{(t)} \rangle \cdot \widehat{c}_j^{(t)} \Big), \quad i \in [k], \tag{12}$$

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

Initialization and clustering procedures: It can be shown that the tensor power updates in (11) are the alternating iterations for the problem of rank-1 approximation of the tensor; see (19) in the appendix for the optimization viewpoint. This is a non-convex problem and has many local optima. Thus, the power update requires careful initialization to ensure convergence to the true rank-1 tensor components. In the semi-supervised setting, we exploit labeled samples for the initialization, and in the unsupervised setting, we propose an SVD-based technique stated in Procedure 3. In this procedure, we introduce the top singular vectors of matrix $T(I, I, \theta)$ (for some random Gaussian vector θ) as the initialization vectors. We establish this method initializes the non-convex power iteration with good initialization vectors when we try large enough number of initializations as characterized in Theorem 3.

In the unsupervised setting, we also need to identify which initializations are successful in recovering the true rank-1 components of the tensor which is performed by the clustering Procedure 2.

Tensor decomposition guarantees: Recall that we provided local and global convergence guarantees of the tensor decomposition algorithm in Section 1.1.1. Those guarantees are required for proving unsupervised and semi-supervised learning results proposed in Section 2.

Acknowledgments

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 26. 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.

9. See Assumption (A2) in Appendix G for precise description.

References

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

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

A. Anandkumar, 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.

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

Anima Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates. *arXiv preprint arXiv:1402.5180*, Feb. 2014b.

J. Anderson, M. Belkin, N. Goyal, L. Rademacher, and J. Voss. The More, the Merrier: the Blessing of Dimensionality for Learning Large Gaussian Mixtures. *arXiv preprint arXiv:1311.2891*, Nov. 2013.

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

Boaz Barak, Jonathan Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. *arXiv preprint arXiv:1407.1543*, 2014.

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

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

Emmanuel J Candes and Terence Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.

J. F. Cardoso and Pierre Comon. Independent component analysis, a survey of some algebraic methods. In *IEEE International Symposium on Circuits and Systems*, pages 93–96, 1996.

A. Coates, H. Lee, and A. 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.

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

P. Comon and C. Jutten. *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press. Elsevier, 2010.

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

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

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

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

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

D. Hsu and S. M. Kakade. Learning Mixtures of Spherical Gaussians: Moment Methods and Spectral Decompositions. *arXiv preprint arXiv:1206.5766*, 2012.

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

A. Hyvarinen and E. Oja. Independent component analysis: algorithms and applications. *Neural Networks*, 13(4-5):411–430, 2000.

R. Latala. Estimates of moments and tails of Gaussian chaoses. *Ann. Prob.*, 34(6):2315–2331, 2006.

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

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

N. H. Nguyen, P. Drineas, and T. D. Tran. Tensor sparsification via a bound on the spectral norm of random tensors. *arXiv preprint arXiv:1005.4732*, May 2010.

M. Rudelson and R. Vershynin. The smallest singular value of a random rectangular matrix. *Communications on Pure and Applied Mathematics*, 62(12):1707–1739, 2009.

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

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

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

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

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

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

Appendix Organization

We first provide more discussions on related works in Section A and propose some experiment results in Section B.

In Section C, we introduce more tensor notations, which are extensively used later in the appendix. Then in Section D, we review some previous approaches in applying tensor decomposition to learning latent variable models, especially how to compute the corresponding moment tensor.

Part I of the appendix contains materials related to the new tensor decomposition algorithm, including intuitive descriptions and detailed proofs.

Part II of the appendix contains the proofs for sample complexity bounds. The most important tools we develop to prove these bounds are the tensor concentration bounds in Section N.

Appendix A. Related work

Tensor decomposition for learning undercomplete models: Several latent variable models can be learned through tensor decomposition including independent component analysis (De Lathauwer et al., 2007), topic models, Gaussian mixtures, hidden Markov models (Anandkumar et al., 2014a) and network community models (Anandkumar et al., 2013b). In the undercomplete setting, Anandkumar et al. (2014a) analyze robust tensor power iteration for learning LVMs, and Song et al. (2013) extend analysis to the nonparametric setting. These works require the tensor factors to have full column rank, which rules out overcomplete models. Moreover, they require whitening the input data, and hence the sample complexity depends on the condition number of the factor matrices. For instance, when $k = d$, for random factor matrices, the previous tensor approaches in Song et al. (2013); Anandkumar et al. (2013a) have a sample complexity of $\tilde{\Omega}(k^{6.5})$, while our result provides improved sample complexity $\tilde{\Omega}(k^2)$ assuming incoherent components.

Learning overcomplete models: In general, learning overcomplete models is challenging, and they may not even be identifiable. The FOOBI procedure by De Lathauwer et al. (2007) shows that a polynomial-time procedure can recover the components of ICA model (with *generic* factors) when $k = O(d^2)$, where the moment is fourth order. However, the procedure does not work for third-order

overcomplete tensors. For the fifth order tensor, Goyal et al. (2013); Bhaskara et al. (2013) perform simultaneous diagonalization on the matricized versions of random slices of the tensor and provide careful perturbation analysis. But, 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. In addition, Goyal et al. (2013) provide stronger results for ICA, where the tensor slices can be obtained in the Fourier domain. Given 4th order tensor, they need $\mathrm{poly}(k^4)$ number of unlabeled samples for learning ICA (where the poly factor is not explicitly characterized), while we only need $\tilde{\Omega}(k^{2.5})$ (when $k = \Theta(d^2)/\operatorname{polylog}(d)$). Anderson et al. (2013) convert the problem of learning Gaussian mixtures to an ICA problem and exploit the Fourier PCA method in Goyal et al. (2013). More precisely, for a Gaussian mixtures model with known identical covariance matrices, when the number of components $k = \mathrm{poly}(d)$, the model can be learned in polynomial time (as long as a certain non-degeneracy condition is satisfied).

Arora et al. (2013); Agarwal et al. (2013); Barak et al. (2014) provide guarantees for the sparse coding model (also known as dictionary learning problem). Arora et al. (2013); Agarwal et al. (2013) provide clustering based approaches for approximately learning incoherent dictionaries and then refining them through alternating minimization to obtain exact recovery of both the dictionary and the coefficients. They can handle sparsity level up to $O(\sqrt{d})$ (per sample) and the size of the dictionary k can be arbitrary. Barak et al. (2014) consider tensor decomposition and dictionary learning using sum-of-squares (SOS) method. In contrast to simple iterative updates considered here, SOS involves solving semi-definite programs. They provide guaranteed recovery by a polynomial time complexity $k^{O(1/\delta)}$ for some $0 < \delta < 1$, when the size of the dictionary $k = \Theta(d)$, and the sparsity level is $k^{1-\delta}$. They also provide guarantees for higher sparsity levels up to (a small enough) constant fraction of k, but the computational complexity of the algorithm becomes quasi-polynomial: $k^{O(\log k)}$. They can also handle higher level of overcompleteness at the expense of reduced sparsity level. They do not require any incoherence conditions on the factor matrices and they can handle the signal to noise ratio being a constant. Thus, their work has strong guarantees, but at the expense of running a complicated algorithm. In contrast, we consider a simple alternating rank-1 updates algorithm, but require more stringent conditions on the model.

There are other recent works which can learn overcomplete models, but under different settings than the one considered in this paper. 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, since they incorporate sparsity, while we incorporate incoherence here.

Concentration Bounds: We obtain tight concentration bounds for empirical tensors in this paper. In contrast, applying matrix concentration bounds, e.g. (Tropp, 2012), leads to strictly worse bounds since they require matricizations of the tensor. Latala (2006) provides an upper bound on the moments of the Gaussian chaos, but they are limited to independent Gaussian distributions (and can be extended to other cases such as Rademacher distribution). The principle of entropy-concentration trade-off (Rudelson and Vershynin, 2009), employed in this paper, have been used in other contexts. For instance, Nguyen et al. (2010) provide a spectral norm bound for random tensors. They first apply a symmetrization argument which reduces the problem to bounding the spectral norm

of a random Gaussian tensor and then employ entropy-concentration trade-off to bound its spectral norm. They also exploit the bounds on the Lipschitz functions of Gaussian random variables. While Nguyen et al. (2010) employ a rough classification of vectors (to be covered) into dense and sparse vectors, we require a finer classification of vectors into different "buckets" (based on their inner products with given vectors) to obtain the tight concentration bounds in this paper. Moreover, we do not impose Gaussian assumption in this paper, and instead require more general conditions such as RIP or bounded 2-to-3 norms.

Appendix B. Experiments

In this Section, we run the algorithm for learning multiview Gaussian mixtures model. We consider model $\mathcal{S}$ described in Section 2.1. The mixture components are uniformly i.i.d. drawn from d-dimensional sphere $\mathcal{S}^{d-1}$. We assume low-noise regime such that $\zeta\sqrt{d} = 0.1$. In addition, let [10] $w_j = \Pr[h = j] = \frac{1}{k}, j \in [k]$. We consider $d = 100$ and $k = \{10, 20, 50, 100, 200, 500\}$. In order to see the effect of number of components k, we fix the number of samples $n = 1000$.

Notice that the empirical tensor $\widehat{T}$ in (22) is not explicitly computed, and the tensor power updates in the algorithm are computed through the multilinear form stated in (23). This leads to efficient computational complexity. See Section 3 for detailed discussion.

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

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

where t_{S} is the stopping threshold. According to the error bound provided in Theorem 7, we let

$$t_{\mathrm{S}} := t_1(\log d)^2\sqrt{\frac{k}{n}} + t_2(\log d)^2\frac{\sqrt{k}}{d}, \tag{13}$$

for some constants $t_1, t_2 > 0$. Here, we set $t_1 = 1e-08$, and $t_2 = 1e-07$.

A random initialization approach is used where $\widehat{a}^{(0)}$ and $\widehat{b}^{(0)}$ are uniformly i.i.d. drawn from sphere $\mathcal{S}^{d-1}$. Initialization vector $\widehat{c}^{(0)}$ is generated through update formula in (11). Figure 3 depicts the ratio of recovered components vs. the number of initializations. We observe that the algorithm is capable of recovering mixture components even in the overcomplete regime $k \geq d$. As suggested in the experimental results of Anandkumar et al. (2014b), we also observe that random initialization works efficiently in the experiments, while the theoretical results for random initialization appear to be highly pessimistic. This suggests additional room for improving the theoretical guarantees under random initialization.

Table 1 provides the average square error of the estimates, the average weight error and the average number of iterations for different values of k. The averages are over different initializations and random runs. The square error is computed as

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

10. In order to see the algorithm performance more easily, we generate n samples such that each mixture component is exactly appeared in $\frac{n}{k}$ observations. Note that this is basically imposing equal number of different mixture components in the observations.

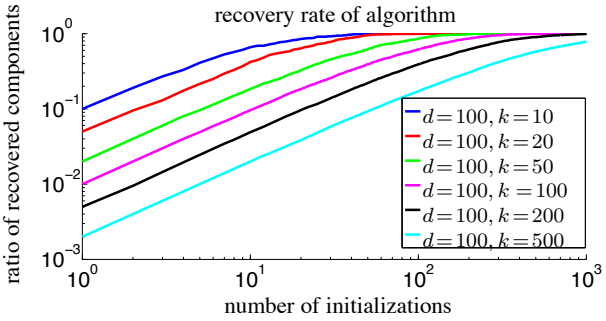

Figure 3: Ratio of recovered components vs. the number of initializations. The figure is an average over 10 random runs.

Table 1: Results for learning a multi-view mixture model. $d = 100$, $n = 1000$, $\zeta\sqrt{d} = 0.1$.

k	avg. square error	avg. weight error	avg. # of iterations	avg. square error $/k$	avg. weight error $/k$
10	1.24e-03	1.73e-05	9.81	1.24e-04	1.73e-06
20	2.94e-03	5.28e-05	10.98	1.41e-04	2.64e-06
50	7.21e-03	1.84e-04	12.74	1.44e-04	3.69e-06
100	1.47e-02	5.36e-04	14.86	1.47e-04	5.36e-06
200	3.03e-02	1.85e-03	18.34	1.51e-04	9.23e-06
500	8.26e-02	1.23e-02	30.02	1.65e-04	2.45e-05

for the corresponding recovered column j. The weight error is computed as square relative error $|\widehat{w} - w_j|^2/w_j^2$. The number of iterations performed before stopping the algorithm is mentioned in the fourth column. We observe that we can still get good error bounds even for overcomplete models with $d = 100$ and $k = 500$.

In the last two columns, the normalized values of errors are provided. The normalization is done by the number of mixtures k. Here, we observe that the normalized values (specially for the square error) are very close for different k. This complies with the theoretical error bound in (7) which claims that the square recovery error is bounded as $\tilde{O}(k)$ when d and n are fixed as here.

Appendix C. More Matrix and Tensor Notations

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

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 $T(:,j,l)$, mode-2 by $T(i,:,l)$ and mode-3 by $T(i,j,:)$. For $r \in \{1,2,3\}$, the mode-r matricization of a third order tensor $T \in \mathbb{R}^{d\times d\times d}$, denoted by $\mathrm{mat}(T,r) \in \mathbb{R}^{d\times d^2}$, consists of all mode-$r$ fibers arranged as column vectors.

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

$$T(M_1,M_2,M_3)_{i_1,i_2,i_3} := \sum_{j_1,j_2,j_3\in[d]} T_{j_1,j_2,j_3}\cdot M_1(j_1,i_1)\cdot M_2(j_2,i_2)\cdot M_3(j_3,i_3). \tag{14}$$

In particular, if u, v and w are vectors and T is a 3rd order tensor, then $T(u,v,w)$ is a scalar, $T(I,v,w)$ is a vector, and $T(I,I,w)$ is a matrix. See (8) for $T(I,v,w)$.

The CP decomposition is closely related to the multilinear form in (14). In particular, consider rank-k tensor

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

Then, for vectors $\widehat{a},\widehat{b},\widehat{c}\in\mathbb{R}^d$, we have

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

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.

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

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

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

For two matrices $A\in\mathbb{R}^{d_1\times k}$ and $B\in\mathbb{R}^{d_2\times k}$, the *Khatri-Rao* product is denoted by $A\odot B\in\mathbb{R}^{d_1d_2\times k}$, and its $(\mathbf{i},j)^{\text{th}}$ entry is given by

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

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

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

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

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

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

Appendix D. Tensor Decomposition for Learning Latent Variable Models

In this section, we discuss that the problem of learning several latent variable models reduces to the tensor decomposition problem. We show that the observed moment of the latent variable models can be written in a tensor-structured form when appropriate modifications are performed. This is done for multiview linear mixture models (in the main text), spherical Gaussian mixtures and ICA (Independent Component Analysis). For a more detailed discussion on the connection between observed moments of LVMs and tensor decomposition, see Section 3 in Anandkumar et al. (2014a).

D.1. Spherical Gaussian mixtures

Consider a mixture of k different Gaussian distributions with spherical covariances. Let $w_j, j \in [k]$ denote the proportion for choosing each mixture. For each Gaussian component $j \in [k]$, $a_j \in \mathbb{R}^d$ is the mean, and $\zeta_i^2 I$ is the spherical covariance. For simplicity, we restrict to the case where all the components have the same spherical variance, i.e., $\zeta_1^2 = \zeta_2^2 = \cdots = \zeta_k^2 = \zeta^2$. The generalization is discussed in Hsu and Kakade (2012). In addition, in order to generalize the learning result to the overcomplete setting, we assume that variance parameter ζ^2 is known (see Remark 14 for more discussions). The following lemma shows that the problem of estimating parameters of this mixture model can be formulated as a tensor decomposition problem. This is a special case of Theorem 1 in Hsu and Kakade (2012) where we assume the variance parameter is known.

Lemma 13 (Hsu and Kakade 2012) *If*

$$M_3 := \mathbb{E}[x \otimes x \otimes x] - \zeta^2 \sum_{i\in[d]} \left(\mathbb{E}[x] \otimes e_i \otimes e_i + e_i \otimes \mathbb{E}[x] \otimes e_i + e_i \otimes e_i \otimes \mathbb{E}[x]\right), \tag{15}$$

then

$$M_3 = \sum_{j\in[k]} w_j a_j \otimes a_j \otimes a_j.$$

In order to provide the learning guarantee, we define the following empirical estimates. Let $\widehat{\mathcal{M}}_3$, $\widehat{\mathcal{M}}_2$, and $\widehat{\mathcal{M}}_1$ respectively denote the empirical estimates of the raw moments $\mathbb{E}[x \otimes x \otimes x]$, $\mathbb{E}[x \otimes x]$, and $\mathbb{E}[x]$. Then, the empirical estimate of the third order modified moment in (15) is

$$\widehat{M}_3 := \widehat{\mathcal{M}}_3 - \zeta^2 \sum_{i\in[d]} \left(\widehat{\mathcal{M}}_1 \otimes e_i \otimes e_i + e_i \otimes \widehat{\mathcal{M}}_1 \otimes e_i + e_i \otimes e_i \otimes \widehat{\mathcal{M}}_1\right). \tag{16}$$

Remark 14 (Variance parameter estimation) *Notice that we assume variance ζ^2 is known in order to generalize the learning result to the overcomplete setting. Since ζ is a scalar parameter, it is reasonable to try different values of ζ till we get a good reconstruction. On the other hand, in the undercomplete setting, variance ζ^2 can be also estimated as proposed in Hsu and Kakade (2012), where estimate $\hat{\zeta}^2$ is the k-th largest eigenvalue of the empirical covariance matrix $\widehat{\mathcal{M}}_2 - \widehat{\mathcal{M}}_1 \widehat{\mathcal{M}}_1^\top$.*

D.2. Independent component analysis (ICA)

In the standard ICA model (Comon, 1994; Cardoso and Comon, 1996; Hyvarinen and Oja, 2000; Comon and Jutten, 2010), random independent latent signals are linearly mixed and perturbed with noise to generate the observations. Let $h \in \mathbb{R}^k$ be a random latent signal, where its coordinates are independent, $A \in \mathbb{R}^{d \times k}$ be the mixing matrix, and $z \in \mathbb{R}^d$ be the Gaussian noise. In addition, h and z are also independent. Then, the observed random vector is

$$x = Ah + z.$$

The following lemma shows that the problem of estimating parameters of the ICA model can be formulated as a tensor decomposition problem.

Lemma 15 (Comon and Jutten 2010) *Define*

$$M_4 := \mathbb{E}[x \otimes x \otimes x \otimes x] - T, \tag{17}$$

where $T \in \mathbb{R}^{d \times d \times d \times d}$ is the fourth order tensor with

$$T_{i_1,i_2,i_3,i_4} := \mathbb{E}[x_{i_1} x_{i_2}]\mathbb{E}[x_{i_3} x_{i_4}] + \mathbb{E}[x_{i_1} x_{i_3}]\mathbb{E}[x_{i_2} x_{i_4}] + \mathbb{E}[x_{i_1} x_{i_4}]\mathbb{E}[x_{i_2} x_{i_3}], \quad i_1, i_2, i_3, i_4 \in [d].$$

Let $\kappa_j := \mathbb{E}[h_j^4] - 3\mathbb{E}[h_j^2]$, $j \in [k]$. Then, we have

$$M_4 = \sum_{j \in [k]} \kappa_j a_j \otimes a_j \otimes a_j \otimes a_j.$$

See Hsu and Kakade (2012) for a proof of this theorem in this form. Let $\widehat{M}_4$ be the empirical estimate of M_4 given n samples.

Part I: Algorithm Analysis

Appendix E. Tensor Decomposition Algorithm

Figure 2 in the main part depicts the flowchart of the tensor decomposition algorithm. In this section, we provide the details of all steps of that flowchart, and state more detailed discussions. More detailed guarantees for the algorithm (compared with what presented in Section 1.1.1) are also provided in Appendix F.

Recall that our algorithm includes two main update steps. First step is the tensor power iteration which is provided in Algorithm 1, and the second step is residual error removal which is provided in Algorithm 4.

Algorithm 1 Tensor decomposition via alternating asymmetric power updates

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

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

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

- Semi-supervised setting: label information is exploited; see equation (5).
- Unsupervised setting: SVD-based Procedure 3 when $k \leq \beta d$ (for arbitrary constant β).

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

4: Asymmetric power updates (see (8) for the definition of multilinear form):

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

5: weight estimation: $\widehat{w}_\tau = T\left(\widehat{a}_\tau^{(N)}, \widehat{b}_\tau^{(N)}, \widehat{c}_\tau^{(N)}\right)$.

6: Cluster set $\left\{\left(\widehat{w}_\tau, \widehat{a}_\tau^{(N)}, \widehat{b}_\tau^{(N)}, \widehat{c}_\tau^{(N)}\right), \tau \in [L]\right\}$ into k clusters as in Procedure 2.

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

Procedure 2 Clustering process

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

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

2: Among the remaining 4-tuples, choose $\widehat{a}, \widehat{b}, \widehat{c}$ which correspond to the largest $|T(\widehat{a}, \widehat{b}, \widehat{c})|$.

3: Do N more iterations of alternating updates in (11) starting from $\widehat{a}, \widehat{b}, \widehat{c}$.

4: Let the output of iterations denoted by $(\widehat{a}, \widehat{b}, \widehat{c})$ be the center of cluster i.

5: Remove all the tuples with $\max\{|\langle \widehat{a}_\tau, \widehat{a}\rangle|, |\langle \widehat{b}_\tau, \widehat{b}\rangle|, |\langle \widehat{c}_\tau, \widehat{c}\rangle|\} > \epsilon/2$.

6: **return** the k cluster centers.

We now provide more discussions about the tensor power iteration step which is the main step of our tensor decomposition algorithm. The rank-1 alternating update method for tensor decomposition is given in Algorithm 1. Given an initial estimate of the vectors denoted by $\left(\widehat{a}^{(0)}, \widehat{b}^{(0)}, \widehat{c}^{(0)}\right)$, an *asymmetric power update* in (11) on the input tensor T is performed in each iteration of the

Procedure 3 SVD-based initialization when $k \leq \beta d$ for arbitrary constant β

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

1: Draw a random standard Gaussian vector $\theta \sim \mathcal{N}(0, I_d)$.
2: Compute u_1 and v_1 as the top left and right singular vectors of $T(I, I, \theta) \in \mathbb{R}^{d\times d}$.
3: $\widehat{a}^{(0)} \leftarrow u_1$, $\widehat{b}^{(0)} \leftarrow v_1$, and initialize $\widehat{c}^{(0)}$ by update formula in (11).
4: **return** $\left(\widehat{a}^{(0)}, \widehat{b}^{(0)}, \widehat{c}^{(0)}\right)$.

Algorithm 4 Coordinate descent algorithm for removing the residual error

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

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

$$\widehat{A}^{(0)} := \arg\min_{\tilde{A}} \|\tilde{A}\| \quad \text{s. t. } \|\tilde{a}_i - \widehat{a}_i\| \leq \tilde{O}\left(\sqrt{k}/d\right), i \in [k]. \tag{18}$$

2: **for** $t = 0$ **to** $N - 1$ **do**
3: **for** $i = 1$ **to** k **do**
4:

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

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

5: Update $\widehat{C}^{(t+1)}$ by applying Procedure 5 with inputs $\tilde{C}^{(t+1)}$ and $\widehat{C}^{(t)}$.
6: Repeat the above steps (with appropriate changes) to update $\widehat{A}^{(t+1)}$ and $\widehat{B}^{(t+1)}$.
7: Update $\widehat{w}^{(t+1)}$:

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

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

algorithm. Notice that the update in (11) alternates among different modes of the tensor. Recall that for vectors $u \in \mathbb{R}^d$ and $v \in \mathbb{R}^d$, the multilinear form $T(I, u, v) \in \mathbb{R}^d$ which is used in the update formula (11) is defined in (14) as

$$T(I, u, v)_i := \sum_{j,l\in[d]} T_{i,j,l} u_j v_l.$$

$T(u, I, v)$ and $T(u, v, I)$ are also defined in a similar way.

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

$$\min_{\substack{a,b,c\in\mathcal{S}^{d-1} \\ w\in\mathbb{R}}} \|T - w \cdot a \otimes b \otimes c\|_F, \tag{19}$$

Procedure 5 Projection procedure

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

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

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

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

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

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

where $\mathcal{S}^{d-1}$ denotes the unit d-dimensional sphere. The optimization program is non-convex, and has multiple local optima. Three updates in (11) are the alternating optimization for this program where in each update, optimization over one vector is performed while the other two vectors are assumed fixed. This alternating minimization approach does not converge to the true components of tensor T in general, and we provide sufficient conditions for the decomposition guarantees.

We now provide a simple intuition behind the power update procedure. Consider a rank-k tensor T as in (10), and suppose we have exact initialization vectors $\widehat{a} = a_j$ and $\widehat{b} = b_j$, for some $j \in [k]$. Then, we have

$$T\left(\widehat{a}, \widehat{b}, I\right) = T\left(a_j, b_j, I\right) = w_j c_j + \sum_{i \neq j} w_i \langle a_j, a_i\rangle\langle b_j, b_i\rangle c_i, \tag{20}$$

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 that the true vectors a_j, b_j and c_j are stationary points for the power update procedure. However, since we consider non-orthogonal tensors, this procedure cannot recover the decomposition exactly. Under incoherence conditions which encourages soft-orthogonality constraints,[11] we establish that the second term in (20) is small, which leads to approximate recovery results.

Notice that the algorithm is run for L different initialization vectors for which we do not know the good ones in prior. In order to identify which initializations are successful at the end, we also need a *clustering* step proposed in Procedure 2 in the main text to obtain the final estimates of the vectors. The detailed analysis of clustering procedure is provided in Appendix J.

For generating initialization vectors $\left(\widehat{a}^{(0)}, \widehat{b}^{(0)}, \widehat{c}^{(0)}\right)$, we introduce two possibilities. One is the simple random initializations, where $\widehat{a}^{(0)}$ and $\widehat{b}^{(0)}$ are uniformly drawn from unit sphere $\mathcal{S}^{d-1}$. The other option is SVD-based Procedure 3 in the main text, where top left and right singular vectors of $T(I, I, \theta)$ (for some random $\theta \in \mathbb{R}^d$) are respectively introduced as $\widehat{a}^{(0)}$ and $\widehat{b}^{(0)}$. Under both initialization procedures, vector $\widehat{c}^{(0)}$ is generated through update formula in (11). We establish in Section F.1.2 that when $k = O(d)$, the SVD procedure leads to global convergence guarantees under polynomial trials.

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

11. See Assumption (A2) in Appendix G for precise description.

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

Comparison with Alternating Least Square(ALS): The updates in Algorithm 1 can be viewed as a rank-1 form of the standard alternating least squares (ALS) procedure. This is because the unnormalized update for c in (11) can be rewritten as

$$\tilde{c}_\tau^{(t+1)} := T\left(\widehat{a}_\tau^{(t)}, \widehat{b}_\tau^{(t)}, I\right) = \mathrm{mat}(T,3)\cdot\left(\widehat{b}_\tau^{(t)} \odot \widehat{a}_\tau^{(t)}\right), \tag{21}$$

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

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

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 $\widehat{T}$ of the form

$$\widehat{T} := \frac{1}{n}\sum_{l\in[n]} x_1^{(l)} \otimes x_2^{(l)} \otimes x_3^{(l)}, \tag{22}$$

where $x_r^{(l)}$ is the l^{th} sample from view $r \in [3]$. Applying the power update (11) in Algorithm 1 to $\widehat{T}$, we have

$$\tilde{c} := \widehat{T}(\widehat{a}, \widehat{b}, I) = \frac{1}{n} X_3(X_1^\top \widehat{a} * X_2^\top \widehat{b}), \tag{23}$$

where $*$ corresponds to the *Hadamard* product. Here, $X_r := \left[x_r^{(1)}\ x_r^{(2)}\ \cdots\ x_r^{(n)}\right] \in \mathbb{R}^{d\times n}$. Thus, the update can be computed efficiently using simple matrix and vector operations. It is easy to see that the above update in (23) is easily parallelizable, and moreover, the different initializations can be parallelized, making the algorithm scalable for large problems.

Appendix F. Guarantees for Tensor Decomposition

As described in the main text and previous section, our tensor decomposition algorithm includes two main update steps including tensor power iteration in (11) and residual error removal in (12). In this section, we provide the guarantees for these steps. The proofs are provided in Appendix H.

F.1. Guarantees for Tensor Power Iteration Step: Algorithm 1

In this section, we provide the local and global convergence guarantees for tensor decomposition Algorithm 1. Throughout the section, we assume tensor $\widehat{T} \in \mathbb{R}^{d\times d\times d}$ is of the form $\widehat{T} = T + \Psi$, 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 $\mathcal{S}^{d-1}$. We state the deterministic assumptions in Appendix G, and show that random matrices satisfy these assumptions. Notice that it is also reasonable to assume these assumptions hold for some non-random matrices.

BASIC DEFINITIONS

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

Definition 16 *For any two vectors $u, v \in \mathbb{R}^d$, the distance between them is defined as*

$$\mathrm{dist}(u, v) := \sup_{z\perp u} \frac{\langle z, v\rangle}{\|z\| \cdot \|v\|} = \sup_{z\perp v} \frac{\langle z, u\rangle}{\|z\| \cdot \|u\|}. \tag{24}$$

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

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

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

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

$$\epsilon_T := \frac{\psi}{w_{\min}} + \tilde{O}\left(\gamma\frac{\sqrt{k}}{d}\right), \tag{25}$$

denote the target error where $\gamma := \frac{w_{\max}}{w_{\min}}$.

F.1.1. LOCAL CONVERGENCE GUARANTEE

The local convergence result is provided in the following theorem which bounds the estimation error after t iterations of the Algorithm. Note that a good initialization is assumed in the local convergence guarantee and the behavior of asymmetric power update in the inner loop of Algorithm 1 is analyzed.

Conditions for Theorem 17:

- Rank-k true tensor with generic components: Let

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

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

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

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

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

- Initialization: The following initialization bound holds w.r.t. some $j \in [k]$ as

$$\epsilon_0 := \max\left\{\mathrm{dist}\left(\widehat{a}^{(0)}, a_j\right), \mathrm{dist}\left(\widehat{b}^{(0)}, b_j\right)\right\} = O(1/\gamma), \tag{26}$$

  where $\gamma := \frac{w_{\max}}{w_{\min}}$. In addition, given $\widehat{a}^{(0)}$ and $\widehat{b}^{(0)}$, suppose $\widehat{c}^{(0)}$ is also calculated by the update formula in (11).

Theorem 17 (Local convergence guarantee of Algorithm 1) *Consider $\widehat{T} = T + \Psi$ as the input to Algorithm 1, and assume the conditions and settings mentioned above hold. Given initialization vectors $(\widehat{a}^{(0)}, \widehat{b}^{(0)}, \widehat{c}^{(0)})$, then the asymmetric power iterations (in the inner loop) of Algorithm 1 satisfy the following bound with high probability (w.h.p.)*

$$\max\left\{\mathrm{dist}\left(\widehat{a}^{(t)}, a_j\right), \mathrm{dist}\left(\widehat{b}^{(t)}, b_j\right), \mathrm{dist}\left(\widehat{c}^{(t)}, c_j\right)\right\} \leq O(\epsilon_T) + q^t \epsilon_0. \tag{27}$$

Here $q < 1$ is a contraction factor and ϵ_T is defined in (25). Furthermore, the weight estimate $\widehat{w} = \widehat{T}\left(\widehat{a}^{(N)}, \widehat{b}^{(N)}, \widehat{c}^{(N)}\right)$ satisfies w.h.p.

$$|\widehat{w} - w_j| \leq O(w_{\min}\epsilon_T) + w_{\min} q^{N+1} \epsilon_0.$$

See the proof in Appendix H.

Thus, we provide efficient recovery guarantees for alternating rank-1 updates under incoherent factors. The incoherence property is precisely defined in Appendix G (see Assumption (A2)). We also show that the incoherence property is satisfied for random components which is assumed here for simplicity. Note that our recovery is in terms of distance between any true vector a_j (or b_j, c_j) and the estimate $\widehat{a}^{(t)}$ (or $\widehat{b}^{(t)}$, $\widehat{c}^{(t)}$).

Note that the second term in (27) is decaying linearly with the number of iterations. The first term in (27) 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 $t \to \infty$, (27) can be interpreted as an approximate local identifiability result for tensor decomposition under incoherent factors.

The result in (27) can be stated in the non-asymptotic form, and the contraction factor $q < 1$ can be characterized explicitly. See Appendix G for details.

Symmetric tensor decomposition: The above local convergence result also holds for recovering the components of a rank-k *symmetric* tensor. Consider symmetric tensor T with CP decomposition $T = \sum_{i \in [k]} w_i a_i \otimes a_i \otimes a_i$. Algorithm 1 is applied to recover the components $a_i, i \in [k]$, where we employ the symmetric update

$$\widehat{a}^{(t+1)} = \frac{T\left(\widehat{a}^{(t)}, \widehat{a}^{(t)}, I\right)}{\left\| T\left(\widehat{a}^{(t)}, \widehat{a}^{(t)}, I\right) \right\|}. \tag{28}$$

Then, the same local convergence result in Theorem 17 holds for this algorithm. The proof follows the lines of Theorem 17 proof with some slight modifications considering the symmetric structure.

Extension to higher order tensors: Algorithm 1 and local convergence guarantee in Theorem 17 are provided for a 3rd order tensor. The algorithm can be simply extended to higher order tensors to compute the corresponding CP decomposition. Consider p-th order tensor $T \in \bigotimes^p \mathbb{R}^d$ with CP decomposition

$$T = \sum_{i \in [k]} w_i a_{(1),i} \otimes a_{(2),i} \otimes \cdots \otimes a_{(p),i}, \tag{29}$$

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

$$\widehat{a}_{(p)}^{(t+1)} = \frac{T\left(\widehat{a}_{(1)}^{(t)}, \widehat{a}_{(2)}^{(t)}, \ldots, \widehat{a}_{(p-1)}^{(t)}, I\right)}{\left\| T\left(\widehat{a}_{(1)}^{(t)}, \widehat{a}_{(2)}^{(t)}, \ldots, \widehat{a}_{(p-1)}^{(t)}, I\right) \right\|}, \tag{30}$$

and similarly the other updates are changed. Define the target error as (generalization of 3rd order case in (25))

$$\tilde{\epsilon}_T := \frac{\psi}{w_{\min}} + \tilde{O}\left(\gamma \sqrt{\frac{k}{d^{p-1}}}\right). \tag{31}$$

Corollary 18 (Local convergence guarantee for p-th order tensor) *Consider the same conditions and settings as in Theorem 17, unless tensor T is p-th order with CP decomposition in (29) where $p \geq 3$ is a constant. In addition, the bounds on $\gamma := \frac{w_{\max}}{w_{\min}}$ and k are modified as*

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

Then, the asymmetric power iterations (in the inner loop) of Algorithm 1 satisfy the following bound with high probability (w.h.p.)

$$\operatorname{dist}\left(\widehat{a}_{(r)}^{(t)}, a_{(r),j}\right) \leq O\left(\tilde{\epsilon}_T\right) + \tilde{q}^t \epsilon_0, \quad \text{for } r \in [p].$$

Here $\tilde{q} < 1$ is a contraction factor and $\tilde{\epsilon}_T$ is defined in (31). Furthermore, the weight estimate $\widehat{w} = \widehat{T}\left(\widehat{a}_{(1)}^{(N)}, \widehat{a}_{(2)}^{(N)}, \ldots, \widehat{a}_{(p)}^{(N)}\right)$ satisfies w.h.p.

$$|\widehat{w} - w_j| \leq O(w_{\min}\tilde{\epsilon}_T) + w_{\min}\tilde{q}^{N+1}\epsilon_0.$$

F.1.2. GLOBAL CONVERGENCE GUARANTEE WHEN $k = O(d)$

Theorem 17 provides local convergence guarantee given good initialization vectors for different components. In this section, we exploit SVD-based initialization method in Procedure 3 to provide good initialization vectors when $k = O(d)$. Combining the theoretical guarantees of this initialization method (provided in Appendix I) with the local convergence guarantee in Theorem 17, we provide the following global convergence result.

Settings of Algorithm 1 in Theorem 19:

- Number of iterations: $N = \Theta\left(\log\left(\frac{1}{\gamma\epsilon_T}\right)\right)$, where $\gamma := \frac{w_{\max}}{w_{\min}}$.
- The initialization in each run of Algorithm 1 is performed by SVD-based technique proposed in Procedure 3, with the number of initializations as

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

Conditions for Theorem 19:

- Rank-k decomposition and perturbation conditions as [12]

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

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

12. Note that the perturbation condition is stricter than the corresponding condition in the local convergence guarantee (Theorem 17).

Theorem 19 (Global convergence guarantee of Algorithm 1 when $k = O(d)$) *Consider $\widehat{T} = T + \Psi$ as the input to Algorithm 1, and assume the conditions and settings mentioned above hold. Then, for any $j \in [k]$, the output of Algorithm 1 satisfies the following w.h.p.,*

$$\max\left\{\text{dist}\left(\widehat{a}_j, a_j\right), \text{dist}\left(\widehat{b}_j, b_j\right), \text{dist}\left(\widehat{c}_j, c_j\right)\right\} = O(\epsilon_T),$$
$$|\widehat{w}_j - w_j| = O(w_{\min}\epsilon_T),$$

where ϵ_T is defined in (25).

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. 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 multi-view high-dimensional features and $y \in \mathbb{R}^{d_o}$ is a low-dimensional label.

Since in the SVD initialization in Procedure 3, two components $\widehat{a}^{(0)}$ and $\widehat{b}^{(0)}$ are initialized through SVD, and the third component $\widehat{c}^{(0)}$ is initialized through update formula (11), we can generalize the global convergence result in Theorem 19 to the setting where A, B are undercomplete, and C is relatively overcomplete.

Corollary 20 *Consider the same setting as in Theorem 19. 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 19 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 21 *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.*

F.1.3. PROOF OUTLINE

The global convergence guarantee in Theorem 19 is established by combining the local convergence result in Theorem 17 and the SVD initialization result in Appendix I.

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 G. Under these conditions, with a good initialization (i.e. small enough $\mathrm{dist}(\widehat{a}, a_j)$ and $\mathrm{dist}(\widehat{b}, b_j)$), we show that the iterative update in (11) provides an estimate $\widehat{c}$ with

$$\mathrm{dist}(\widehat{c}, c_j) < O(\epsilon_T) + q\epsilon_0,$$

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

The initialization argument for SVD-based technique in Procedure 3 has two parts. The first part claims that by performing enough number of initializations (large enough L), a gap condition is satisfied, meaning that we obtain a vector θ which is relatively close to c_j compared to any $c_i, i \neq j$. This is a standard result for Gaussian vectors, e.g., see Lemma B.1 of Anandkumar et al. (2014a). In the second part of the argument, we analyze the dominant singular vectors of $T(I, I, \theta)$, for a vector θ with a good relative gap, to obtain an error bound on the initialization vectors. This is obtained through standard matrix perturbation results (Weyl and Wedin's theorems). See Appendix I for the complete proof.

F.2. Guarantees for Removing Residual Error: Algorithm 4

We first provide the following definition.

Definition 22 ((η_0, η_1)-nice) *Suppose*

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

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

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

and the weights satisfy

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

Note that the optimization program in (18) is proposed to ensure the above bound on $\|\widehat{A}\|$ is satisfied on the input of this part of algorithm. Given above conditions are satisfied, we prove the following guarantees for removing residual error, Algorithm 4.

Theorem 23 *Consider T as the input to Algorithm 4, where T is a rank-k tensor. Suppose Assumptions (A1)-(A5) and (A11) hold (which are satisfied whp when the components are uniformly i.i.d. drawn from unit d-dimensional sphere). Given initial solution $\left\{\widehat{A}^{(0)}, \widehat{B}^{(0)}, \widehat{C}^{(0)}, \widehat{w}^{(0)}\right\}$ which is (η_0, η_1)-nice, all the following iterations of Algorithm 4 are $(2\eta_0, 3\eta_1)$-nice. Furthermore, given the exact tensor T, the Frobenius norm error $\max\{\|\Delta A\|_F, \|\Delta B\|_F, \|\Delta C\|_F, \|\Delta w\|/w_{\min}\}$ shrinks by at least a factor of 2 in every iteration. In addition, if we have a noisy tensor $\widehat{T} = T + \Psi$ such that $\|\Psi\| \leq \psi$, then*

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

Appendix G. 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

$$T = \sum_{i\in[k]} w_i(a_i \otimes b_i \otimes c_i), \quad w_i > 0, a_i, b_i, c_i \in \mathcal{S}^{d-1}, \forall\, i \in [k], \tag{32}$$

where $\mathcal{S}^{d-1}$ denotes the unit d-dimensional sphere, i.e. all the vectors have unit [13] 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

$$\rho := \max_{i\neq j}\{|\langle a_i, a_j\rangle|, |\langle b_i, b_j\rangle|, |\langle c_i, c_j\rangle|\} \leq \frac{\alpha}{\sqrt{d}}, \tag{33}$$

for some $\alpha = \text{polylog}(d)$. In other words, $A^\top A = I + J_A$, $B^\top B = I + J_B$, and $C^\top C = I + J_C$, where J_A, J_B, and J_C, are incoherence matrices with zero diagonal entries. We have $\max\{\|J_A\|_\infty, \|J_B\|_\infty, \|J_C\|_\infty\} \leq \rho$ as in (33).

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

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

for some constant $\alpha_0 > 0$.

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

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

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

13. This normalization is for convenience and the results hold for general case.

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

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

$$\psi := \|\Psi\|. \tag{34}$$

Suppose ψ is bounded as [14]

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

where α_0 is a constant.

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

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

(A8) **Contraction factor:** The contraction factor q in Theorem 17 as

$$q := \frac{2w_{\max}}{w_{\min}}\left[\frac{2\alpha}{\sqrt{d}}\left(1 + \alpha_0\sqrt{\frac{k}{d}}\right)^2 + \beta'\right], \tag{35}$$

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

(A9) **Initialization:** Let

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

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

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

for some constants $\alpha_0, \beta' > 0$, $\alpha = \operatorname{polylog}(d)$, and $0 < q < 1$ which is defined in (35).

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

Remark 24 *Many of the assumptions are actually parameter choices. The only properties of random matrices required are (A2), (A3), (A4) and (A10),. See Appendix G.1 for detailed discussion.*

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

Let us provide a brief discussion about the above assumptions. Condition (A1) requires the presence of a rank-k decomposition for tensor T. We normalize the component vectors for convenience, and this removes the scaling indeterminacy issues which can lead to problems in convergence. Additionally, we impose incoherence constraint in (A2), which allows us to provide convergence guarantee in the overcomplete setting. Assumptions (A3) and (A4) impose bounds on the spectral norm of tensor T and its decomposition components. Note that assumptions (A2)-(A4) and (A10) are satisfied w.h.p. when the columns of A, B, and C are generically drawn from unit sphere $\mathcal{S}^{d-1}$ (see Lemma 25 and Guédon and Rudelson (2007)), all others are parameter choices. Assumption (A5) limits the overcompleteness of problem which is required for providing convergence guarantees. The first bound on perturbation in (A6) as $\psi \leq \frac{w_{\min}}{6}$ is required for local convergence guarantee and the second bound $\psi \leq \frac{w_{\min}\sqrt{\log k}}{\alpha_0\sqrt{d}}$ is needed for arguing initialization provided by Procedure 3. Assumption (A7) is required to ensure contraction happens in each iteration. Assumption (A8) defines contraction ratio q in each iteration, and Assumption (A9) is the initialization condition required for local convergence guarantee.

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

G.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 25 *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 result.

Lemma 26 *Consider a random matrix $A \in \mathbb{R}^{d\times k}$ whose columns are drawn uniformly at random from unit sphere. If $k < d^{p/2}/\operatorname{polylog}(d)$, then*

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

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

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

Lemma 27 *Suppose columns of matrix $A \in \mathbb{R}^{d\times k}$ has unit norm and satisfy the incoherence condition (A2), when $k \leq d^{1.25-\epsilon}$, the $2 \to p$ norm of $A^\top$ is bounded by $1+o(1)$ for any $p > 3-2\epsilon$.*

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

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

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

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

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

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

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

Consequently, we have the following corollary.

Corollary 28 *For vectors f, g, h, and weights $w_i \geq 0$, we have*

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

Proof: The proof applies Hölder's inequality twice as

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

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

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

Lemma 29 *Let A, B, and C be random matrices in $\mathbb{R}^{d\times k}$ whose columns are drawn uniformly at random from unit sphere. If $k < d^{3/2}/\operatorname{polylog}(d)$, and*

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

then

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

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

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

where Corollary 28 is exploited in the first inequality, and Lemma 26 is used in the last inequality. □

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

Lemma 30 *Let A, B, and $C \in \mathbb{R}^{d\times k}$ be independent, normalized (column) Gaussian matrices. Then for all $i \in [k]$, we have with high probability*

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

Proof: We have

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

where $\delta_j := w_j \langle A_i, A_j\rangle \langle B_i, B_j\rangle$ is independent of C_j. From Lemma 25, columns of A and B are incoherent, and therefore, for $j \neq i$, we have

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

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

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

The proof is completed by applying union bound. □

SPECTRAL NORM OF KHATRI-RAO PRODUCT

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

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

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

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

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

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

Lemma 31 *Let $M=\sum_{i=1}^n M_i$ be sum of independent symmetric $d\times d$ matrices with $\mathbb{E}[M_i]=0$, assume all matrices M_i's have spectral norm at most R almost surely, let $\sigma^2=\|\mathbb{E}[M_i^2]\|$, then for any τ*

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

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

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

Lemma 32 *Suppose A, $B\in\mathbb{R}^{d\times k}(k>\text{polylog}\, d)$ are independent random matrices with independent Gaussian entries, let $M=(A\odot B)^\top(A\odot B)=(A^\top A)*(B^\top B)$, then with high probability*

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

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

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

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

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

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

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

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

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

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

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

Appendix H. Proof of Algorithm Convergence Results

In this section, we prove the convergence guarantees of our algorithm provided in Appendix F.

H.1. Proof of Convergence Results in Theorems 17 and 19 of Appendix

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 $\widehat{T}$.

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

$$f(\epsilon; k, d) := \alpha \frac{\sqrt{k}}{d} + \frac{2\alpha}{\sqrt{d}} \left(1 + \alpha_0 \sqrt{\frac{k}{d}}\right)^2 \epsilon + \alpha_0 \epsilon^2, \tag{36}$$

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

Lemma 33 (Contraction result of Algorithm 1 in one update) *Consider $\widehat{T} = T + \Psi$ as the input to Algorithm 1, where T is a rank-k tensor, and Ψ is a perturbation tensor. Suppose Assumptions*

(A1)-(A5) hold, and estimates $\widehat{a}$ and $\widehat{b}$ satisfy distance bounds

$$\mathrm{dist}(\widehat{a}, a_j) \leq \epsilon_a,$$
$$\mathrm{dist}(\widehat{b}, b_j) \leq \epsilon_b,$$

for some $j \in [k]$, and $\epsilon_a, \epsilon_b > 0$. Let $\epsilon := \max\{\epsilon_a, \epsilon_b\}$, and suppose ψ defined in (34) be small enough such that [15]

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

where $f(\epsilon; k, d)$ is defined in (36). Then, update $\widehat{c}$ in (11) satisfies the following distance bound with high probability (w.h.p.)

$$\mathrm{dist}(\widehat{c}, c_j) \leq \frac{w_{\max} f(\epsilon; k, d) + \psi}{w_j - w_j\epsilon^2 - w_{\max} f(\epsilon; k, d) - \psi}. \tag{37}$$

Furthermore, if the bound in (37) is such that $\mathrm{dist}(\widehat{c}, c_j) \leq \epsilon$, then the update $\widehat{w} := \widehat{T}(\widehat{a}, \widehat{b}, \widehat{c})$ also satisfies w.h.p.

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

Remark 34 *In the asymptotic regime, $f(\epsilon; k, d)$ is*

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

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

Remark 35 (Rate of convergence) *The local convergence result provided in Theorem 17 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 33 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 17, we approximated the almost-quadratic convergence rate in the beginning with linear convergence.*

Lemma 33 is proposed in the general form. In Lemma 36, we provide explicit contraction result by imposing additional perturbation, contraction and initialization Assumptions (A6), (A8) and (A9). We observe that under reasonable rank, perturbation and initialization conditions, the denominator in (37) 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 36 (Contraction result of Algorithm 1 in one update) *Consider $\widehat{T} = T + \Psi$ as the input to Algorithm 1, where T is a rank-k tensor, and Ψ is a perturbation tensor. Let Assumptions* [16] *(A1)-*

15. This is the denominator of bound provided in (37).
16. As mentioned in the assumptions, from perturbation bound in (A6), only the bound $\psi \leq \frac{w_{\min}}{6}$ is required here.

(A9) hold. Note that initialization bound in (A9) is satisfied for some $j \in [k]$. Then, update $\widehat{c}$ in (11) satisfies the following distance bound with high probability (w.h.p.)

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

where

$$\text{Const.} := \frac{2}{w_{\min}} \left(\psi + w_{\max} \alpha \frac{\sqrt{k}}{d} \right), \tag{38}$$

and contraction ratio $q < 1/2$ is defined in (35). Note that $\alpha = \operatorname{polylog}(d)$. The final estimation of weights satisfy

$$|\widehat{w} - w_j| \leq O(w_{max} \alpha \sqrt{k}/d + \psi).$$

Before proving Lemma 33 and 36, we first explain how they can be applied in proving the theorems.

Proof of Theorem 17: 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 36. □

Proof of Theorem 19: The error bounds on $w_i, a_i, b_i, c_i, i \in [k]$, are proved by combining the local convergence result in Theorem 17, and initialization result in Theorem 39. □

In order to prove Lemma 33 and 36, first recall 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 (33).

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 33: Let $z_a^* \perp a_j$ and $z_b^* \perp b_j$ denote the vectors that achieve supremum value in (24) corresponding to $\operatorname{dist}(\widehat{a}, a_j)$ and $\operatorname{dist}(\widehat{b}, b_j)$, respectively. Furthermore, without loss of generality, assume $\|z_a^*\| = \|z_b^*\| = 1$. Then, $\widehat{a}$ and $\widehat{b}$ are decomposed as

$$\widehat{a} = \langle a_j, \widehat{a} \rangle a_j + \operatorname{dist}(\widehat{a}, a_j) z_a^*, \tag{39a}$$

$$\widehat{b} = \langle b_j, \widehat{b} \rangle b_j + \operatorname{dist}(\widehat{b}, b_j) z_b^*. \tag{39b}$$

Let $\overline{C} := C \operatorname{Diag}(w)$ denote the unnormalized matrix C, and $\tilde{c} := \widehat{T}(\widehat{a}, \widehat{b}, I)$ denote the unnormalized update in (11). 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

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

Substituting $\widehat{a}$ and $\widehat{b}$ from (39*a*) and (39*b*), we have

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

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

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

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

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

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

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

Finally, for S_4, we have

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

for some $\alpha_0 = O(1)$. The bound on $\|T\|$ is from Assumption (A4). Note that for random components, we showed in Lemma 29 that this bound holds w.h.p. exploiting Assumption (A5) and results of Guédon and Rudelson (2007). For the error term $\Psi(\widehat{a},\widehat{b},z_c)$, we have

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

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

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

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

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

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

For $\tilde{c}$, we have

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

and therefore,

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

where inequality $\langle a_j, \widehat{a}\rangle\langle b_j, \widehat{b}\rangle \geq 1 - \epsilon^2$ 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

$$\operatorname{dist}(\tilde{c}, \overline{C}_j) \leq \frac{w_{\max} f(\epsilon; k, d) + \psi}{w_j - w_j\epsilon^2 - w_{\max} f(\epsilon; k, d) - \psi}. \tag{40}$$

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

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

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

The estimate $\widehat{w} = \widehat{T}(\widehat{a}, \widehat{b}, \widehat{c})$ can be expanded as

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

and therefore,

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

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

□

Proof of Lemma 36: The result is proved by applying Lemma 33, and incorporating additional conditions (A6), (A8), and (A9). $f(\epsilon_0; k, d)$ in (36) can be bounded as

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

where $\epsilon_0 \leq \frac{\beta'}{\alpha_0}$ from Assumption (A9) is exploited in the inequality. The last equality is concluded from definition of contracting factor q in (35). On the other hand, the denominator in (37) can be lower bounded as

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

where Assumptions (A9) and (A6) are used in the inequality. Applying Lemma 33, the result on $\mathrm{dist}(\widehat{c}, c_j)$ is proved.

From Lemma 33, we also have

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

where ϵ can be replaced by the final error $Const = 2(\psi + w_{max}\alpha\sqrt{k}/d)/w_{min}$ (see Equation 38). Substituting this, and notice the bounds on w_{max}/w_{min} and ψ, we know $|\widehat{w} - w_j| \leq O(w_{max}\alpha\sqrt{k}/d + \psi)$. □

H.2. Proof of Convergence Result in Theorem 23 of Appendix

To prove this theorem, we first observe that the algorithm update formula in (12) is (before normalization) $w_i\langle a_i, \widehat{a}_i\rangle\langle b_i, \widehat{b}_i\rangle c_i + \epsilon_i$ where

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

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

Lemma 37 *Before normalization* $\tilde{w}_i\tilde{c}_i = w_i\langle a_i, \widehat{a}_i\rangle\langle b_i, \widehat{b}_i\rangle c_i + \epsilon_i$ *where*

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

Proof: By the update formula in (12), we know

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

We expand it into several terms as follows.

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

The norm of three different types of terms mentioned above are bounded in Section H.2.1, which conclude the desired bound in the lemma. □

We are now ready to prove the main theorem.

Proof of Theorem 23: Since $\tilde{w}_i$ is the norm of $w_i\langle a_i, \widehat{a}_i\rangle\langle b_i, \widehat{b}_i\rangle c_i + \epsilon_i$, we know

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

and therefore

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

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

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

By Lemma 38, we know after the projection procedure, we get $\|\widehat{C} - C\|_F \leq 2\|\tilde{C} - C\|_F$. Therefore combining the two steps we know

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

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

□

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

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

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

H.2.1. CLAIMS FOR PROVING LEMMA 37

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

Claim 1 *We have*

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

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

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

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

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

□

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

Claim 2 *We have*

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

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

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

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

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

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

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

Notice that the proof works for both terms. □

Claim 3 *We have*

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

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

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

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

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

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

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

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

Here we used the assumption that $\|\Delta A_i\| \leq \tilde{O}(\sqrt{k}/d)$ and $\|\Delta A\| \leq O(\sqrt{k/d})$. Therefore $\|\widehat{A}^\top\|_{2\to 4} \leq \|A^\top\|_{2\to 4} + \|\Delta A^\top\|_{2\to 4} \leq O(1)$ (and similarly for $\widehat{B}^\top$).

Therefore

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

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

□

H.2.2. PROJECTION PROCEDURE 5

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

Lemma 38 *Suppose the initial solution $\{\widehat{A}^0, \widehat{B}^0, \widehat{C}^0, \widehat{w}^0\}$ is (η_0, η_1)-nice. For any solution $\{\tilde{A}, \tilde{B}, \tilde{C}, \tilde{w}\}$, let error $E = \max\{\|\tilde{A} - A\|_F, \|\tilde{B} - B\|_F, \|\tilde{C} - C\|_F, \|\tilde{w} - w\|/w_{\min}\}$. Then after the projection Procedure 5, the new solution is $(2\eta_0, 3\eta_1)$-nice and has error at most $2E$.*

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

Claim 4

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

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

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

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

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

Appendix I. 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 (32), we have

$$T(I, I, \theta) = \sum_{i\in[k]} w_i\langle\theta, c_i\rangle a_i b_i^\top = \sum_{i\in[k]} \lambda_i a_i b_i^\top = A\,\mathrm{Diag}(\lambda)B^\top, \tag{41}$$

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

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

Since A and B are not orthogonal matrices, the expansion in (41) is not the SVD [17] of $T(I, I, \theta)$. But, we show in the following theorem that if we draw enough number of random vectors θ in the regime of $k = O(d)$, we can eventually provide good initialization vectors through SVD of $T(I, I, \theta)$.

Define

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

Theorem 39 (SVD initialization when $k = O(d)$) *Consider tensor $\widehat{T} = T + \Psi$ 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 $\widehat{T}(I, I, \theta^{(j)})$. This is L random runs of Procedure 3. 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 (44), and some $0 < \tilde{\mu} < 1$. Note that $\rho \leq \frac{\alpha}{\sqrt{d}}$ is also defined as the incoherence parameter in Assumption (A2). Then, w.h.p., at least one of the pairs $(u_1^{(j)}, v_1^{(j)}), j \in [L]$, say j^*, satisfies

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

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

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

Proof: Let $\lambda^{(j)} := \mathrm{Diag}(w)C^\top\theta^{(j)} \in \mathbb{R}^k$ and $\tilde{\lambda}^{(j)} := C^\top\theta^{(j)} \in \mathbb{R}^k$. From Lemmata 40 and 41, there exists a $j^* \in [L]$ such that w.h.p., we have

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

From (42), with probability at least $1 - 2k^{-1}$, we have

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

From (43), with probability at least $1 - k^{-7}$, we have

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

where in the last inequality, we also applied upper bound on $\tilde{\lambda}_1^{(j^*)}$. Combining all above bounds and Lemma 45 finishes the proof. □

I.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 41 is satisfied under some number of random draws.

Lemma 40 (Gap condition) *Consider an arbitrary matrix $C \in \mathbb{R}^{d\times k}$ with unit-norm columns which also satisfies incoherence condition $\max_{i\neq j}|\langle c_i, c_j\rangle| \leq \rho$ for some $\rho > 0$. Let*

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

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

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

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

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

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

$$\Pr\left[\tilde{\lambda}_1^{(j^*)} \geq \sqrt{2\ln(L)} - \frac{\ln(\ln(L)) + c}{2\sqrt{2\ln(L)}} - \sqrt{2\ln(k)}\right] \geq 1 - \frac{2}{k}. \tag{42}$$

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 (42). From Lemma 43, 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 (42). By combining above two bounds, with probability at least $1 - k^{-7}$, we have

$$\tilde{\lambda}_{(2)}^{(j^*)} \leq \rho \tilde{\lambda}_1 + 4\sqrt{\log k}. \tag{43}$$

From the given bound on L in the lemma and inequalities (42) and (43), 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 $\widehat{T}(I, I, \theta)$ provide good initialization vectors for a_1 and b_1.

Before proposing the lemma, we define

$$\mu_E := \alpha \sqrt{\frac{k}{d}} \left(2 + 2\alpha_0 \sqrt{\frac{k}{d}} + \frac{\alpha}{\sqrt{d}}\right), \quad \mu_R := \left(1 + \alpha_0 \sqrt{\frac{k}{d}}\right)^2, \quad \mu_{\min} := \min\{\mu_E, \mu_R\}. \tag{44}$$

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

Lemma 41 *Consider $\widehat{T} = T + \Psi$, 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 $\widehat{T}(I, I, \theta)$. Let*

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

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

$$\lambda_1 \geq (1 + \mu)\lambda_{(2)}, \tag{45}$$

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 (44). Then, with high probability (w.h.p.),

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

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 (41), 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 $R = \mathcal{P}_\perp(R) + \mathcal{P}_\|(R)$. First term $\mathcal{P}_\perp(R)$ is the component with column space orthogonal to a_1 and row space orthogonal to b_1, and $\mathcal{P}_\|(R)$ is the component with either the column space equal to a_1 or the row space equal to b_1. We have

$$\mathcal{P}_\perp(R) = (I - P_{a_1})R(I - P_{b_1}),$$
$$\mathcal{P}_\|(R) = P_{a_1}R + RP_{b_1} - P_{a_1}RP_{b_1},$$

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 $\widehat{T} = T + \Psi$, we have

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

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 (45) assumed in the lemma and inequality (46), 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, $\widehat{T}(I,I,\theta)$ has the corresponding SVD form

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

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

$$\begin{aligned}
\tilde{\sigma}_2 = \|\mathcal{P}_\perp(R)\| &\leq \|R\| \\
&= \left\| \sum_{i=2}^{k} \lambda_i a_i b_i^\top \right\| \\
&\leq \lambda_{(2)} \left\| A_{\backslash 1} \right\| \left\| B_{\backslash 1}^\top \right\| \\
&\leq \lambda_{(2)} \|A\| \left\| B^\top \right\| \\
&\leq \left(1 + \alpha_0 \sqrt{\frac{k}{d}} \right)^2 \lambda_{(2)} =: \mu_R \lambda_{(2)},
\end{aligned} \tag{46}$$

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

$$\begin{aligned}
|\sigma_1 - \lambda_1| &\leq \|E\| + \|\Psi(I,I,\theta)\| \\
&\leq \lambda_{(2)} \alpha \sqrt{\frac{k}{d}} \left(2 + 2\alpha_0 \sqrt{\frac{k}{d}} + \frac{\alpha}{\sqrt{d}} \right) + \|\Psi(I,I,\theta)\| \\
&=: \mu_E \lambda_{(2)} + \|\Psi(I,I,\theta)\|,
\end{aligned} \tag{47}$$

where (48) 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 (46) and (47) are used in the first inequality, and the second inequality is concluded from the gap condition (45) 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 $\widehat{T}(I,I,\theta) = M + E + \Psi(I,I,\theta)$, which implies that

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

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

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

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

where $A_{(1)} := \big[\overbrace{a_1|a_1|\cdots|a_1}^{k-1 \text{ times}}\big] \in \mathbb{R}^{d\times(k-1)}$, $B_{\setminus 1} := [b_2|b_3|\cdots|b_k] \in \mathbb{R}^{d\times(k-1)}$, and $\lambda_{(a)} := [\lambda_i \rho_{1i}^{(a)}]_{i\neq 1} \in \mathbb{R}^{k-1}$. The other notations are similarly defined.
For E_1, we have

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

Where the first equality is concluded from Lemma 44, 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

$$\|E\| \leq \lambda_{(2)} \alpha \sqrt{\frac{k}{d}} \left(2 + 2\alpha_0 \sqrt{\frac{k}{d}} + \frac{\alpha}{\sqrt{d}}\right). \tag{48}$$

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 $\widehat{T}(I, I, \theta) = \lambda_1 a_1 b_1^\top + R + \Psi(I, I, \theta)$, 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 (46) and gap condition (45) are used in the second inequality. Since $\tilde{\sigma}_2 = 0$, by Wedin's theorem, we have

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

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

Lemma 42 *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 43 *Consider $r = [r_1, r_2, \ldots, r_k]^\top \in \mathbb{R}^k$ as a k-dimensional random Gaussian vector with zero mean and covariance Σ, i.e., $r \sim \mathcal{N}(0, \Sigma)$. For any $k \geq 2$, we have*

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

Proof: From Lemma 42, 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 44 *Given $h \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$, let $H = [h|h|\cdots|h]\,\mathrm{Diag}(v) \in \mathbb{R}^{m\times n}$. Then, $\|H\| = \|h\|\|v\|$.*

Proof: By definition

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

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

In the following lemma, we show that noise matrix $\Psi(I, I, \theta)$ has bounded norm with high probability which is useful for initialization argument in Theorem 39.

Lemma 45 *Let $\theta \in \mathbb{R}^d$ be standard multivariate Gaussian as $\mathcal{N}(0, I_d)$. Then, for any $\alpha_0 > 1$, we have*

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

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

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

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

where the last inequality is from the definition of tensor spectral norm. Applying the bound on $\|\theta\|$ in Lemma 46 finishes the proof. □

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

Lemma 46 (Lemma 15 of Dasgupta et al. (2006)) *Let the random vector θ is distributed as $\mathcal{N}(0, I_d)$. Then, for any $\alpha_0 > 1$, we have*

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

Appendix J. Clustering Process

In the last step of main algorithm, we need to cluster the generated 4-tuples into k clusters. Theoretically, we only have convergence guarantees when the initialization vectors are good enough, while the other initializations can potentially generate arbitrary 4-tuples. In the worst case, these arbitrary 4-tuples can make the clustering process hard, and therefore, we provide specific Procedure 2 for which the output properties are provided in Lemma 49.

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

For simplicity, we only prove this when the initialization procedure in Theorem 19 takes polynomial time, namely $k = O(d)$ and $w_{\max}/w_{\min} = O(1)$. Without loss of generality, we also assume $w_{\max} = w_1 \geq w_2 \geq \cdots \geq w_k = w_{\min}$. In this case, we choose the threshold ϵ in the following lemmata to be some small constant depending on k/d and $w_{\max}/w_{\min}$. Also, we work in the case when noise $\Psi = 0$, however the proof still works when the noise $\psi = \|\Psi\| = o(1)$.

Lemma 47 *Suppose*

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

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

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

Proof: Partition tensor $T = \sum_{i\in[k]} w_i a_i \otimes b_i \otimes c_i$ to $T_1 + T_2$, where T_1 contains all the terms indexed from 1 to $t-1$, and T_2 contains the remaining terms. From Corollary 28, we have

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

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

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

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

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

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

$$|T_2(\widehat{a},\widehat{b},\widehat{c})| \ge (1-2\delta)w_t. \tag{49}$$

On the other hand, from Corollary 28,

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

Since all the 3-norms are bounded by $1+o(1)$, each of them must be at least $1-O(\delta)$ to let inequality (49) hold. Now we have

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

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

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

when ϵ and δ are small enough.

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

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

Lemma 48 *Suppose there exists a good initialization (see initialization condition (26) in the local convergence theorem) for some column $t \in [k]$, and*

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

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

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

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

Proof: Similar to the proof of Lemma 47, partition tensor $T = \sum_{i\in[k]} w_i a_i \otimes b_i \otimes c_i$ to $T_2 = w_t a_t \otimes b_t \otimes c_t$ and $T_1 = T - T_2$. Since the initialization is good, by the local convergence result in Theorem 17, we have

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

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

Similar to Lemma 47, by using Corollary 28, we have $|T_1(\widehat{a}, \widehat{b}, \widehat{c})| \leq w_t\delta/2$. Applying these bounds, we have

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

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

Finally we prove the clustering process succeeds.

Lemma 49 *Procedure 2 outputs k cluster centers that are $\tilde{O}\left(\frac{w_{\max}}{w_{\min}}\frac{\sqrt{k}}{d}\right)$ close to the true components of the tensor.*

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

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

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

By Lemma 48 we know there must be a 4-tuple with $|T(\widehat{a}, \widehat{b}, \widehat{c})| > w_t(1 - \delta)$. On the other hand, by Lemma 47 we know the 4-tuple we found must satisfy $\max\{\mathrm{dist}(\widehat{a}, a_j), \mathrm{dist}(\widehat{b}, b_j)\} < w_{\min}/10w_{\max}$ for some j (and this cannot be some j that has already been found). This tuple then satisfies the conditions of the local convergence Theorem 17. 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.

□

Part II: Learning Guarantees and Concentration Bounds

Appendix K. Proof of Main Learning Guarantees

We first explain how different algorithm guarantees and tensor concentration bounds are combined leading to the main learning Theorems.

Proof of Theorem 7 (in the main text): The result is proved by applying the tensor concentration bound in Lemma 56 to the local convergence results of algorithm in Theorems 17 and 23 (all in the Supplementary materials). Note that in the low noise regime $\zeta^2 = O\left(\frac{1}{d}\right)$ considered here, the term $\zeta\sqrt{w_{\max}\frac{d}{n}}$ in Lemma 56 is dominant. □

Proof of Theorem 10 (in the main text): The result is proved by applying the tensor concentration bound in Lemma 56 to the global convergence results of algorithm in Theorems 19 and 23 (all in the Supplementary materials). □

Proof of Theorem 53 (in the appendix): The result is proved by applying the tensor concentration bound in Lemma 63 to the local convergence results of algorithm in Corollary 18 and Theorem 23 (all in the Supplementary materials). Note that for learning ICA, the 4th order generalization of the algorithm is applied. The details of generalization of coordinate descent Algorithm 4 to 4th order is omitted, but it can be argued with similar techniques we exploited for the 3rd order case. □

Proof of Theorem 54 (in the appendix): The result is proved by applying the tensor concentration bound in Lemma 63 to the global convergence results of algorithm in Theorems 19 and 23 (all in the Supplementary materials). Note that the SVD technique is applied to the 4-th order case as described in the settings. Therefore, the requirement on noise in global convergence result is changed as $\psi := \|\Psi\| \leq \frac{w_{\min}\sqrt{\log k}}{\alpha_0^2 d}$. □

Proof of Theorem 12 (in the main text): The result is proved similar to the ICA case with the difference that the concentration bound in Lemma 64 for sparse ICA is exploited. □

Appendix L. Learning Multiview linear Mixtures Model

Here, we state more general learning results for multiview linear mixtures model. In particular, the result in the high noise regime $\zeta^2 = \Theta(1)$ is also provided.

L.1. Semi-supervised learning of multiview linear mixtures model in the overcomplete setting

Suppose that the distribution of observed variables given hidden state is sub-Gaussian with covariance matrix $\zeta^2 I$ (as model $\mathcal{S}$ is described in Section 2.1 in the main text), we have the following concentration bound where with probability at least $1-\delta$, $\widehat{a}_j$ satisfies

$$\|\widehat{a}_j - a_j\| \leq C_1\sqrt{\frac{\zeta^2 d\log(1/\delta)}{m_j}}, \quad j \in [k],$$

for some constant $C_1 > 0$.

Settings of Algorithm 1 in Theorem 50:

- Number of iterations: $N = \Theta\left(\log\left(\frac{1}{\gamma\tilde{\epsilon}_T}\right)\right)$, where $\gamma := \frac{w_{\max}}{w_{\min}}$.
- Initialization: Exploit the empirical estimates in (5) in the main text as initialization vectors, and therefore the number of initializations $L = k$.

Conditions for Theorem 50:

- Rank condition: $k = o(d^{3/2})$.
- The columns of factor matrices are uniformly i.i.d. drawn from unit d-dimensional sphere $\mathcal{S}^{d-1}$ (see Remark 8 for more discussion).
- Suppose the distribution of observed variables given hidden state is sub-Gaussian, and the number of labeled samples with label j, denoted by m_j, satisfies [18]

$$m_j \geq \tilde{\Omega}\left(\gamma^2 \zeta^2 d\right),\ j \in [k], \tag{50}$$

  where $\gamma := \frac{w_{\max}}{w_{\min}}$.
- Sample complexity requirements and settings for different latent variable models:
  - ◇ *Multiview linear mixture model:* Consider the empirical estimate of 3rd order moment in (3) in the main text as the input of Algorithm 1. Furthermore, given n samples, noise matrices E_A, E_B and E_C satisfy the RIP condition in (RIP) which is satisfied with high probability for many random models (see Remark 51 for details on RIP condition). The number of samples n satisfies

$$n \geq \begin{cases} \Omega\left(\frac{d}{w_{\min}^2}\right), & \zeta^2 = \Theta(1), \\ \Omega\left(\frac{w_{\max}}{w_{\min}^2}\right), & \zeta^2 = \Theta\left(\frac{1}{d}\right), \end{cases} \tag{51}$$

    where ζ^2 is the variance of each entry of observation vectors.
  - ◇ *Spherical Gaussian mixtures:* Consider 3rd order empirical (modified) moment $\widehat{M}_3$ in (16) as the input of Algorithm 1 with symmetric updates. The number of samples n satisfies the same bounds as (51), where $\zeta^2 I$ is the spherical covariance matrix of observations.

Theorem 50 (Semi-supervised learning of multiview linear mixture models and spherical Gaussian mixtures) *Assume the conditions and settings mentioned above hold. Then, Algorithm 1 outputs $\widehat{a}_j, j \in [k]$ as the estimates of columns of true factor matrix A satisfying w.h.p.*

$$\mathrm{dist}\left(\widehat{a}_j, a_j\right) = O(\tilde{\epsilon}_T), \quad j \in [k],$$

where $\mathrm{dist}(\cdot,\cdot)$ function and $\tilde{\epsilon}_T$ are defined in (24) and (25), respectively. In the asymmetric cases, similar bounds hold for other factor matrices B and C. In addition, the weight estimates $\widehat{w}_j, j \in [k]$ satisfy w.h.p.

$$|\widehat{w}_j - w_j| = O(w_{\min} \tilde{\epsilon}_T).$$

Proof: The result is proved by applying the tensor concentration bound in Lemma 56 to the local convergence result of Algorithm 1 in Theorem 17. Note that in the high noise regime $\zeta^2 = \Theta(1)$, the term $\zeta^3\sqrt{\frac{d}{n}}$ in Lemma 56 is dominant, and in the low noise regime $\zeta^2 = \Theta\left(\frac{1}{d}\right)$, the term $\zeta\sqrt{w_{\max}\frac{d}{n}}$ in Lemma 56 is dominant □

18. In model $\mathcal{S}$, the columns of factor matrices are unit vectors, and therefore, the most reasonable regime of error is when the expected norm of error vector is constant, i.e., $\mathbb{E}\left[\|\zeta\sqrt{d}\varepsilon\|^2\right] = \zeta^2 d \leq O(1)$. But, note that the label complexity holds even if $\zeta^2 d \geq \omega(1)$.

Remark 51 (RIP property) *Given n samples for the model $\mathcal{S}$ proposed in Section 2.1, define noise matrix*

$$E_A := [\varepsilon_A^1, \varepsilon_A^2, \ldots, \varepsilon_A^n] \in \mathbb{R}^{d \times n},$$

where $\varepsilon_A^i \in \mathbb{R}^d$ is the i-th sample of noise vector ε_A. E_B and E_C are similarly defined. These matrices need to satisfy the RIP property as follows which is adapted from Candes and Tao (2006).

(RIP) Matrix $E \in \mathbb{R}^{d \times n}$ satisfies a weak RIP condition such that for any subset of $O\left(\frac{d}{\log^2 d}\right)$ number of columns, the spectral norm of E restricted to those columns is bounded by 2.

It is known that the above condition is satisfied with high probability for many random models such as when the entries are i.i.d. zero mean Gaussian or Bernoulli random variables.

L.2. Unsupervised learning of multiview linear mixtures models

The conditions and settings for unsupervised learning are stated as follows where comparing to the semi-supervised learning, the initialization setting, rank and sample complexity conditions are changed.

Settings of Algorithm 1 in Theorem 52:

- Number of iterations: $N = \Theta\left(\log\left(\frac{1}{\gamma \tilde{\epsilon}_T}\right)\right)$, where $\gamma := \frac{w_{\max}}{w_{\min}}$.
- Instead of initialization by exploiting label information (which is not available in the unsupervised setting), the initialization in each run of Algorithm 1 is performed by SVD-based technique proposed in Procedure 3, with the number of initializations as

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

Conditions for Theorem 52:

- Rank condition: $k = O(d)$.
- The columns of factor matrices are uniformly i.i.d. drawn from unit d-dimensional sphere $\mathcal{S}^{d-1}$.
- Sample complexity and input settings: Consider the same settings for input tensors (moments) as in semi-supervised setting, but the sample complexity for both multiview linear mixture models and spherical Gaussian mixtures are changed as

$$n \geq \begin{cases} \Omega\left(\frac{d^2}{w_{\min}^2}\right), & \zeta^2 = \Theta(1), \\ \Omega\left(\frac{w_{\max}}{w_{\min}^2} d\right), & \zeta^2 = \Theta\left(\frac{1}{d}\right). \end{cases}$$

Theorem 52 (Unsupervised learning of multiview linear mixture models and spherical Gaussian mixtures) *Assume the conditions and settings mentioned above hold. Then, Algorithm 1 outputs $\widehat{a}_j, j \in [k]$ as the estimates of columns of true factor matrix A satisfying w.h.p.*

$$\text{dist}\left(\widehat{a}_j, a_j\right) = O(\tilde{\epsilon}_T), \quad j \in [k],$$

where $\mathrm{dist}(\cdot,\cdot)$ *function and* $\tilde{\epsilon}_T$ *are defined in* (24) *and* (25)*, respectively. In the asymmetric cases, similar bounds hold for other factor matrices* B *and* C*. In addition, the weight estimates* $\widehat{w}_j, j \in [k]$ *satisfy w.h.p.*

$$|\widehat{w}_j - w_j| = O(w_{\min}\tilde{\epsilon}_T).$$

Proof: The result is proved by applying the tensor concentration bound in Lemma 56 to the global convergence result of Algorithm 1 in Theorem 19. The dominant error bounds in Lemma 56 are the same as what stated in the proof of Theorem 7. □

Appendix M. Learning ICA and Sparse ICA

In this section, we propose the semi-supervised and unsupervised learning results for ICA model. By semi-supervised setting in ICA, we mean some prior information is available which provides good initializations for the components. Recall the standard ICA model (Comon, 1994), where *random independent* latent signals are linearly mixed to generate the observations. Let $h \in \mathbb{R}^k$ be a random latent signal where its coordinates are independent, and $A \in \mathbb{R}^{d\times k}$ be the mixing matrix. Then, the observed vector is

$$x = Ah \in \mathbb{R}^d.$$

For simplicity, we limit to noiseless setting. This is the standard setting, and is already challenging because samples in ICA are mixtures of many components, unlike the mixture models. It is discussed in Appendix D how estimating the parameters of ICA model can be formulated as a tensor decomposition problem where a modified version of 4th order observed moment (denoted by M_4) is characterized in a tensor decomposition form; see Lemma 15 in the appendix.

Settings of Algorithm in Theorem 53: Given n samples $x^i = Ah^i, i \in [n]$, consider the empirical estimate of 4th order (modified) moment M_4 (see (17) in the Appendix) as the input to the algorithm with symmetric 4th order updates; see Appendix F.1.1 for higher order extension of the algorithm. Let the number of iterations $N = \tilde{\Theta}\left(\log\left(1/\tilde{\epsilon}_R\right)\right)$, where $\tilde{\epsilon}_R := \min\{k^2/\min\{n, \sqrt{d^3 n}\}, \sqrt{k}/d^{1.5}\}$. For initialization, it is assumed that for any $j \in [k]$, an approximation of a_j denoted by $\widehat{a}_j^{(0)}$ is given satisfying $\|\widehat{a}_j^{(0)} - a_j\| \le \alpha$ for some constant $\alpha < 1$.

Theorem 53 (Semi-supervised learning of ICA) *Assume the Algorithm settings mentioned above hold. Let the entries of* h *are independent subgaussian variables with* $\mathbb{E}[h_j^2] = 1$*, and constant nonzero* 4*th order cumulant. Suppose the rank condition* $\Omega(d) \le k \le o(d^2)$ *holds, and the number of unlabeled samples* n *satisfies*

$$n \ge \begin{cases} \tilde{\Omega}(k^2), & k \le O(d^{1.5})/\operatorname{polylog}(d), \\ \tilde{\Omega}\left(k^4/d^3\right), & \text{o.w.} \end{cases}$$

Then the algorithm outputs estimates $\widehat{A}$ *and* $\widehat{w}$*, satisfying w.h.p.*

$$\max\left\{\left\|\widehat{A} - A\right\|_F, \|\widehat{w} - w\|\right\} \le \tilde{O}\left(\frac{k^{2.5}}{\min\{n, \sqrt{d^3 n}\}}\right), \quad j \in [k].$$

We observe that for highly overcomplete regime $k = \Theta(d^2)/\operatorname{polylog}(d)$, the ICA model can be efficiently learned from fourth order moment with $n \geq \tilde{\Omega}(k^{2.5})$ number of unlabeled samples.

Similar to the multiview Gaussian mixture model, we can also provide column-wise recovery guarantees with introducing additional approximation error $\tilde{O}\left(\sqrt{k}/d^{1.5}\right)$. Note that this error is different from multiview Gaussian mixture since we exploit different tensor orders in the two models.

Settings of Algorithm in Theorem 54: Consider the same settings as in Theorem 53 for the input tensor and the number of iterations N. But, the initialization is performed by 4-th order generalization [19] of SVD-based technique in Procedure 3, with the number of initializations as $L \geq k^{\Omega(k^2/d^2)}$.

Theorem 54 (Unsupervised learning of ICA) *Assume the Algorithm settings mentioned above hold. Let the entries of h are independent subgaussian variables with $\mathbb{E}[h_j^2] = 1$, and constant nonzero 4th order cumulant. Suppose the rank condition $k = \Theta(d)$ holds, and the number of unlabeled samples satisfies $n \geq \tilde{\Omega}\left(k^3\right)$. Then the algorithm outputs satisfy the same guarantees as in Theorem 53.*

M.1. Sparse ICA

Finally, we discuss sparse ICA problem. This is the ICA setting with the additional assumption that hidden vector $h \in \mathbb{R}^k$ is sparse with i.i.d. Bernoulli-subgaussian random entries. Assume the probability of each Bernoulli variable being 1 is s/k. Here, we also assume that mixing matrix A satisfies the RIP property (see condition (RIP) in the Appendix).

Theorem 55 (Semi-supervised and unsupervised learning of sparse ICA) *Similar semi-supervised and unsupervised learning guarantees as in Theorems 53 and 54 hold for the sparse ICA model as*

$$\max\left\{\left\|\widehat{A} - A\right\|_F, \|\widehat{w} - w\|\right\} \leq \tilde{O}\left(\frac{s \cdot k^{1.5}}{\min\left\{n, \sqrt{d^3 n}\right\}}\right), \quad j \in [k].$$

The sample complexity requirements are changed as follows. For semi-supervised setting, we need

$$n \geq \begin{cases} \tilde{\Omega}(sk), & sk \leq O(d^3)/\operatorname{polylog}(d), \\ \tilde{\Omega}\left(s^2k^2/d^3\right), & \text{o.w.}, \end{cases}$$

and for unsupervised setting, we need $n \geq \tilde{\Omega}\left(k^2 s\right)$.

In terms of sparsity of latent vector h, the sparse ICA is between multiview Gaussian mixtures (where h has one nonzero entry in basis vector encoding), and ICA (where h is fully dense). Comparing the guarantees, we also observe that the sample complexity results for sparse ICA bridges the range of models between multiview mixtures model and ICA.

Appendix N. Tensor Concentration Bounds

In this section, we provide tensor concentration bounds for different latent variable models.

19. In the 4th order case, the SVD is performed on $T(I, I, \theta, \theta) \in \mathbb{R}^{d \times d}$ for some random vector θ.

N.1. Multiview linear mixture model

In this section, we provide a tensor concentration result for the multiview linear mixture model which bounds the spectral norm of error tensor given n samples. In order to get polynomial sample complexity bounds for unlabeled samples in semi-supervised and unsupervised learning results, it is usually enough to treat the tensor as a vector/matrix and apply appropriate vector/matrix concentartion bounds such as Bernstein bounds. However, these bounds can be significantly improved in many cases by considering the concentration property of the tensor spectral norm directly. Here we provide the tensor concentration bound for the multiview mixture models.

As introduced in Section 2.1, the conditional expectations of the three views are

$$\mathbb{E}[x_1|h] = a_h, \quad \mathbb{E}[x_2|h] = b_h, \quad \mathbb{E}[x_3|h] = c_h.$$

Let $A := [a_1\ a_2\ \cdots\ a_k] \in \mathbb{R}^{d\times k}$ and similarly B and C. Recall model $\mathcal{S}$ proposed in Section 2.1 saying the conditional distributions of observed variables given hidden states are

$$x_1|h \sim a_h + \zeta\sqrt{d}\cdot\varepsilon_A, \quad x_2|h \sim b_h + \zeta\sqrt{d}\cdot\varepsilon_B, \quad x_3|h \sim c_h + \zeta\sqrt{d}\cdot\varepsilon_C,$$

where $\varepsilon_A, \varepsilon_B, \varepsilon_C \in \mathbb{R}^d$ are independent random vectors with zero mean and variance $\frac{1}{d}I$, and ζ^2 is a scalar denoting the variance of each entry. We also assume that noise vectors $\varepsilon_A, \varepsilon_B, \varepsilon_C$ are independent of hidden vector h. In addition, let all the vectors $a_h, b_h, c_h, h \in [k]$, have unit ℓ_2 norm.

Let $x_1^i, x_2^i, x_3^i, i \in [n]$ denote n samples of views x_1, x_2, x_3. Since the main focus is on recovering the components, we bound the spectral norm of difference between the empirical tensor

$$\widehat{T} := \frac{1}{n}\sum_{i=1}^{n} x_1^i \otimes x_2^i \otimes x_3^i,$$

and

$$\tilde{T} := \mathbb{E}[x_1 \otimes x_2 \otimes x_3|h] = \frac{1}{n}\sum_{i=1}^{n}(a_{h_i}) \otimes (b_{h_i}) \otimes (c_{h_i}),$$

where the conditional expectation is over the choice of hidden states for n samples. Here, $h_i \in [k]$ denotes the hidden state for sample $i \in [n]$. Notice that this tensor has the same form as equation (3)

$$\tilde{T} = \sum_{j\in[k]} \tilde{w}_j a_j \otimes b_j \otimes c_j,$$

where $\tilde{w}_j, j \in [k]$ are the empirical frequencies of different hidden states $h \in [k]$. It is easy to see that if $n = \Omega\left(\frac{\log k}{w_{\min}}\right)$, then all the empirical frequencies $\tilde{w}_j$ are within $[w_j/2, 2w_j]$.

Given n samples, define noise matrix

$$E_A := \left[\varepsilon_A^1, \varepsilon_A^2, \ldots, \varepsilon_A^n\right] \in \mathbb{R}^{d\times n}.$$

E_B and E_C are similarly defined.

Lemma 56 (Tensor concentration bound for multiview linear mixture model) *Consider n samples $\{(x_1^i, x_2^i, x_3^i), i \in [n]\}$ from the multiview linear mixture model $\mathcal{S}$ with corresponding hidden states $\{h_i, i \in [n]\}$. Assume matrices $A^\top$, $B^\top$ and $C^\top$ have $2 \to 3$ norm bounded by $O(1)$,*

and noise matrices E_A, E_B and E_C satisfy the RIP condition in (RIP). For $\widehat{T}$ and $\tilde{T}$ as above, if $n = \text{poly}(d)$, we have with high probability (over the choice of h and the noise)

$$\|\widehat{T} - \tilde{T}\| \leq \tilde{O}\left(\zeta\left(\frac{\sqrt{d}}{n} + \sqrt{w_{\max}\frac{d}{n}}\right) + \zeta^2\left(\frac{d}{n} + \sqrt{w_{\max}\frac{d^{1.5}}{n}}\right) + \zeta^3\left(\frac{d^{1.5}}{n} + \sqrt{\frac{d}{n}}\right)\right).$$

Proof: Expanding the difference $\widehat{T} - \tilde{T}$, we have

$$\widehat{T} - \tilde{T} = \frac{1}{n}\zeta^3 d^{1.5}\sum_{i\in[n]} \varepsilon_A^i \otimes \varepsilon_B^i \otimes \varepsilon_C^i \tag{52a}$$

$$+ \frac{1}{n}\zeta^2 d\sum_{i\in[n]} a_{h_i}\otimes\varepsilon_B^i\otimes\varepsilon_C^i + \varepsilon_A^i\otimes b_{h_i}\otimes\varepsilon_C^i + \varepsilon_A^i\otimes\varepsilon_B^i\otimes c_{h_i} \tag{52b}$$

$$+ \frac{1}{n}\zeta\sqrt{d}\sum_{i\in[n]} a_{h_i}\otimes b_{h_i}\otimes\varepsilon_C^i + a_{h_i}\otimes\varepsilon_B^i\otimes c_{h_i} + \varepsilon_A^i\otimes b_{h_i}\otimes c_{h_i}. \tag{52c}$$

There are three types of terms in the above difference which are bounded separately in Claims 5-7 in Section N.1.2. Combining the results of claims, the lemma follows directly.

□

Remark 57 *In a more general setting, the above tensor concentration result is also valid for the following model. Let the hidden variable $h \in \mathbb{R}^k$ is a discrete categorical random variable taking value $e_j \in \mathbb{R}^k$ if the hidden variable is in the j-th state. The observed variables $x_l \in \mathbb{R}^d$ are conditionally independent given the k-dimensional latent variable h, and are represented as*

$$x_l = A_l h + \varepsilon_l,$$

where $\varepsilon_l \in \mathbb{R}^d$ is the noise vector. In addition, given n samples $x_l^i, i \in [n]$, suppose the noise matrices $E_l := \left[\varepsilon_l^1\ \varepsilon_l^2\ \cdots\ \varepsilon_l^n\right]$, $l \in [p]$, satisfy the RIP property in (RIP). Notice that the described multiview linear mixture model belongs to this class of models.

N.1.1. BASIC DEFINITIONS AND LEMMATA

In the proof of the claims in Section N.1.2, we extensively apply two different types of partitioning as follows.

Definition 58 (Small and large terms) *Consider matrices $E_A := [\varepsilon_A^1, \varepsilon_A^2, \ldots, \varepsilon_A^n] \in \mathbb{R}^{d\times n}$, and E_B and E_C which are similarly defined. For any set of vectors u, v and w, the set of columns $[n]$ are partitioned into 2 sets called sets of* small *and* large *terms according to the value of inner products $\langle u, \varepsilon_A^i\rangle$, $\langle v, \varepsilon_B^i\rangle$ and $\langle w, \varepsilon_C^i\rangle$ as follows. The set of small values denoted by $L^c \subseteq [n]$ is defined as*

$$L^c := \left\{i \in [n] : |\langle u, \varepsilon_A^i\rangle| < \frac{10\log d}{\sqrt{d}} \wedge |\langle v, \varepsilon_B^i\rangle| < \frac{10\log d}{\sqrt{d}} \wedge |\langle w, \varepsilon_C^i\rangle| < \frac{10\log d}{\sqrt{d}}\right\},$$

and the rest of columns belong to the set of large values denoted by $L \subseteq [n]$.

Note that when necessary, the above partitioning is similarly applied to one or two matrices.

Lemma 59 *Suppose matrix $E := [\varepsilon^1, \varepsilon^2, \dots, \varepsilon^n] \in \mathbb{R}^{d\times n}$ satisfies the RIP property (RIP). For a vector $u \in \mathbb{R}^d$, let set $L \subseteq [n]$ denote the set of columns of E corresponding to large inner products $\langle u, \varepsilon^i\rangle$ as defined in Definition 58, i.e.,*

$$L := \left\{ i \in [n] : |\langle u, \varepsilon^i\rangle| \geq \frac{10\log d}{\sqrt{d}} \right\}.$$

Then, the size of set L is bounded as

$$|L| \leq \frac{d}{25\log^2 d}. \tag{53}$$

Proof: It can be shown by a contradiction argument assuming $|L| > \frac{d}{25\log^2 d}$. Consider submatrix $E[L]$ (matrix E with columns restricted to set L). We have

$$\|E\|^2 \geq \left\|E[L]^\top u\right\|^2 = \sum_{i\in L}\langle u, \varepsilon^i\rangle^2 \geq |L|\frac{100\log^2 d}{d} > 4,$$

where the first inequality is from the definition of large terms for which $|\langle u, \varepsilon^i\rangle| > 10\log d/\sqrt{d}$, and the second inequality is from contradiction assumption on $|L|$. This contradicts with the RIP property that $\|E[L]\| \leq 2$, and therefore the bound in (53) holds. □

The above partitioning into small and large sets is not enough in part of the analysis, and in order to get a tight bound (specially in the low noise regime), we propose the following finer partitioning.

Definition 60 (Buckets and constrained vectors) *Consider matrix $C := [c_1, c_2, \dots, c_k] \in \mathbb{R}^{d\times k}$, and let $t := \left\lceil \log_2 \sqrt{d} \right\rceil$. For any vector w, the set of columns $[k]$ are partitioned into $t+1$ buckets according to the value of inner products $\langle c_j, w\rangle$ as*

$$K_0 := \left\{ j \in [k] : |\langle c_j, w\rangle| \leq \frac{1}{\sqrt{d}} \right\},$$
$$K_l := \left\{ j \in [k] : |\langle c_j, w\rangle| \in \left(\frac{2^{l-1}}{\sqrt{d}}, \frac{2^l}{\sqrt{d}}\right] \right\}, \quad l \in [t].$$

Furthermore, the constrained vector $z^l \in \mathbb{R}^k, l \in \{0, 1, 2, \dots, t\}$, corresponds to the inner products in bucket l as

$$z^l_j := \begin{cases} \langle c_j, w\rangle, & j \in K_l, \\ 0, & j \notin K_l. \end{cases}$$

One advantage of bucketing (which is not applicable to the small and large partitioning in the previous definition) is that buckets with large value has a smaller ε-net. This exploits the additional property of matrices with bounded $2 \to 3$ norm.

Lemma 61 *Consider matrix $C := [c_1, c_2, \dots, c_k] \in \mathbb{R}^{d\times k}$ where the columns have unit norm, and $\|C^\top\|_{2\to 3} = O(1)$. For a vector w with unit norm, consider the buckets on columns of matrix C defined in Definition 60. For constrained vector $z^l, l \in [t]$, let $p_l := 2^{l-1}$. Then, we have*

- *z^l has at most $O\left(\frac{d^{3/2}}{p_l^3}\right)$ nonzero entries.*

- *There is an ε-net of size* $\exp\left(O\left(\frac{d^{3/2}}{p_l^3}\left(\log k + \log\frac{1}{\varepsilon}\right)\right)\right)$ *for* z^l.

Proof: The first part can be proved by a contradiction argument assuming $|K_l| > O\left(\frac{d^{3/2}}{p_l^3}\right)$. Let $C[K_l]$ denote the restriction of matrix C to columns indexed by K_l. We have

$$\left\|C^\top w\right\|_3^3 \geq \left\|C[K_l]^\top w\right\|_3^3 = \sum_{j\in K_l} |\langle w, c_j\rangle|^3 > O\left(\frac{d^{3/2}}{p_l^3}\right)\cdot\left(\frac{p_l}{\sqrt{d}}\right)^3 = O(1),$$

where the second inequality is from the definition of terms in bucket K_l, and the assumption $|K_l| > O\left(\frac{d^{3/2}}{p_l^3}\right)$. This contradicts with the $2\to 3$ norm bound on $C^\top$, and therefore, we have $|K_l| \leq O\left(\frac{d^{3/2}}{p_l^3}\right)$. Since the number of nonzero entries in z^l is the same as $|K_l|$, the proof of first part is finished.

Let $q_l := \frac{d^{3/2}}{p_l^3}$. First enumerate the support of z^l. There are $\binom{k}{q_l}$ possibilities for the location of q_l nonzero entries in z^l which is bounded as

$$\binom{k}{q_l} \leq \left(e\frac{k}{q_l}\right)^{q_l} \leq e^{O(q_l\log k)}.$$

For a given support, take an ε-net for all vectors in that support which has size

$$e^{O\left(q_l\log\left(\frac{1}{\varepsilon}\right)\right)}.$$

The union of these ε-nets is a valid ε-net for z^l of the desired size. This finishes the proof of second claim.

□

A similar (but stronger) lemma can be proved for RIP matrices:

Lemma 62 *Consider matrix $E := [\varepsilon^1, \varepsilon^2, \ldots, \varepsilon^n] \in \mathbb{R}^{d\times n}$ where the columns have unit norm, and it satisfies RIP property (RIP). For a vector w with unit norm, consider the buckets on columns of matrix E defined in Definition 60. For constrained vector z^l, let $p_l := 2^{l-1}$. Then, for $l > 4\log\log d$ we have*

- *z^l has at most $O\left(\frac{d}{p_l^2}\right)$ nonzero entries.*
- *There is an ε-net of size* $\exp\left(O\left(\frac{d}{p_l^2}\left(\log n + \log\frac{1}{\varepsilon}\right)\right)\right)$ *for* z^l.

Proof: The first claim follows from the same argument as in Lemma 59. The ε-net is constructed in the same way as in the previous lemma. □

N.1.2. PROOF OF CLAIMS

In this section, we separately bound different error terms (52*a*)-(52*c*). Among all the terms, the terms like (52*c*) is most difficult to bound (intuitively because terms like b_{h_i} are not "as random" as terms like ε_A^i). In fact, the proof for the term (52*c*) can be adapted to bound all the other terms. Here for clarity we start from the simplest term (52*a*), and point out new ideas in the proofs of (52*b*) and (52*c*).

Claim 5 (Bounding norm of (52a)) *With high probability over $\varepsilon_A^i, \varepsilon_B^i, \varepsilon_C^i$'s and h_i's, we have*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_A^i \otimes \varepsilon_B^i \otimes \varepsilon_C^i \right\| \leq \tilde{O}\left(\frac{1}{n} + \frac{1}{d\sqrt{n}} \right).$$

Proof: Let

$$T_1 := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_A^i \otimes \varepsilon_B^i \otimes \varepsilon_C^i.$$

Rewrite the tensor as

$$T_1 = \frac{1}{n} \sum_{i=1}^{n} \eta_i \varepsilon_A^i \otimes \varepsilon_B^i \otimes \varepsilon_C^i, \tag{54}$$

where η_i's are independent random ± 1 variables with $\Pr[\eta_i = 1] = 1/2$. Clearly, T_1 has the same distribution as the original term, because of the symmetry in error vectors implying e.g. $\eta_i \varepsilon_A^i \sim \varepsilon_A^i$. We first sample the vectors $\varepsilon_A^i, \varepsilon_B^i, \varepsilon_C^i$, and therefore, the remaining random variables are just the η_i's.

The goal is to bound norm of T_1 in (54) which is defined as

$$\|T_1\| := \sup_{\|u\|=\|v\|=\|w\|=1} |T_1(u,v,w)| = \sup_{\|u\|=\|v\|=\|w\|=1} \left| \frac{1}{n} \sum_{i=1}^{n} \eta_i \langle u, \varepsilon_A^i \rangle \langle v, \varepsilon_B^i \rangle \langle w, \varepsilon_C^i \rangle \right|. \tag{55}$$

In order to bound the above, we provide an ε-net argument. Construct an ε-net for vectors u, v and w with $\varepsilon = 1/n^2$. By standard construction, size of the ε-net is $e^{O(d \log n)}$. First, for any fixed triple (u,v,w), we bound $|T_1(u,v,w)|$ where $T_1(u,v,w)$ is a sum of independent variables. As introduced in Definition 58, we partition the sum into *large* and *small* terms as

$$T_1(u,v,w) = \frac{1}{n} \sum_{i=1}^{n} \eta_i \langle u, \varepsilon_A^i \rangle \langle v, \varepsilon_B^i \rangle \langle w, \varepsilon_C^i \rangle := S_L + S_{L^c},$$

where S_{L^c} is the sum of *small* terms consisting of terms satisfying

$$\left\{ |\langle u, \varepsilon_A^i \rangle| < \frac{10 \log d}{\sqrt{d}} \wedge |\langle v, \varepsilon_B^i \rangle| < \frac{10 \log d}{\sqrt{d}} \wedge |\langle w, \varepsilon_C^i \rangle| < \frac{10 \log d}{\sqrt{d}} \right\},$$

and S_L is the sum of *large* terms including all the other terms.

Bounding $|S_{L^c}|$: The sum S_{L^c} is just a weighted sum of η_i's, and the Bernstein's Inequality is exploited to bound it. Each term in the summation is bounded as

$$\left| \frac{1}{n} \langle u, \varepsilon_A^i \rangle \langle v, \varepsilon_B^i \rangle \langle w, \varepsilon_C^i \rangle \right| \leq O\left(\frac{\log^3 d}{nd^{3/2}} \right),$$

where the bound on the small terms is exploited. The variance term is also bounded as

$$O\left(\frac{\log^6 d}{nd^3} \right).$$

Applying Bernstein's inequality, with probability at least $1 - e^{-Cd\log n}$ (where C is a large enough constant), the sum of small terms $|S_{L^c}|$ is bounded by $\tilde{O}\left(\frac{1}{d\sqrt{n}}\right)$.

Bounding $|S_L|$: From RIP property (RIP), we know that noise matrices $E_A := [\varepsilon_A^1, \ldots, \varepsilon_A^n]$, $E_B := [\varepsilon_B^1, \ldots, \varepsilon_B^n]$ and $E_C := [\varepsilon_C^1, \ldots, \varepsilon_C^n]$ satisfy the weak RIP condition with high probability such that for any subset of $O\left(\frac{d}{\log^2 d}\right)$ number of columns, the spectral norm of matrices restricted to those columns is bounded by 2. Let L denote the set of large terms in the proposed partitioning, and $E_A[L]$, $E_B[L]$ and $E_C[L]$ be the matrices E_A, E_B and E_C restricted to the columns indexed by L. Applying Lemma 59, we have

$$|L| \leq \frac{3d}{25\log^2 d}.$$

Note that an additional factor 3 shows up here since the set of small terms is defined as the intersection of 3 sets comparing to what proved in Lemma 59. Therefore, RIP property of E_A, E_B and E_C implies that $E_A[L]$, $E_B[L]$ and $E_C[L]$ have spectral norm bounded by 2. Now applying triangle inequality, we have

$$|S_L| \leq \frac{1}{n}\sum_{i\in L} |\langle u, \varepsilon_A^i\rangle|\cdot|\langle v, \varepsilon_B^i\rangle|\cdot|\langle w, \varepsilon_C^i\rangle| \leq \frac{1}{n}\sum_{i\in L} |\langle u, \varepsilon_A^i\rangle|\cdot|\langle v, \varepsilon_B^i\rangle| \leq \frac{1}{n}\left\|E_A[L]^\top u\right\|\cdot\left\|E_B[L]^\top v\right\| \leq \frac{4}{n},$$

where the second step uses the fact that $|\langle w, \varepsilon_C^i\rangle| \leq 1$, the third step exploits Cauchy-Schwartz inequality, and the last step uses bounds $\|E_A[L]\| \leq 2$ and $\|E_B[L]\| \leq 2$. Notice the three matrices are already sampled before we do the ε-net argument, and therefore, we do not need to do union bound over all u, v, w for this event.

At this point, we have bounds on $|S_L|$ and $|S_{L^c}|$ for a fixed triple (u, v, w) in the ε-net. By applying union bound on all vectors in the ε-net, the bound holds for every triple (u, v, w) in the ε-net. The argument for other (u, v, w)'s which are not in the ε-net follows from their closest triples in the ε-net. □

Claim 6 (Bounding norm of (52*b*)) *With high probability over $\varepsilon_A^i, \varepsilon_B^i$'s and h_i's, we have*

$$\left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_A^i \otimes \varepsilon_B^i \otimes c_{h_i}\right\| \leq \tilde{O}\left(\frac{1}{n} + \sqrt{\frac{w_{\max}}{n\sqrt{d}}}\right).$$

Proof: The proof is similar to the previous claim. Let

$$T_2 = \frac{1}{n}\sum_{i=1}^n \eta_i \varepsilon_A^i \otimes \varepsilon_B^i \otimes c_{h_i},$$

where η_i's are independent random ± 1 variables with $\Pr[\eta_i = 1] = 1/2$. Similar to the previous claim, we first sample the vectors $\varepsilon_A^i, \varepsilon_B^i$ and h_i's, and therefore, the remaining random variables are just the η_i's. Assume the matrices E_A, E_B satisfy the RIP property, and the number of times $h_i = j$ for $j \in [k]$ is bounded by $[nw_{\min}/2, 2nw_{\max}]$. All the events happen with high probability when $n \geq \tilde{\Omega}(1/w_{\min})$ and $n \leq \mathrm{poly}(k)$.

The goal is to bound $\|T_2\|$. We construct an ε-net for vectors u and v with $\varepsilon = 1/n^2$. First, for any fixed pair (u, v), we bound $\|T_2(u, v, I)\|$ where $T_2(u, v, I)$ is a sum of independent zero mean

vectors. As introduced in Definition 58, consider partitioning on columns of E_A and E_B as

$$T_2(u, v, I) = \frac{1}{n}\sum_{i=1}^{n} \eta_i \langle u, \varepsilon_A^i\rangle\langle v, \varepsilon_B^i\rangle c_{h_i} = S_L + S_{L^c},$$

where S_{L^c} is the sum of *small* terms consisting of terms satisfying

$$\left\{|\langle u, \varepsilon_A^i\rangle| < \frac{10\log d}{\sqrt{d}} \ \wedge \ |\langle v, \varepsilon_B^i\rangle| < \frac{10\log d}{\sqrt{d}}\right\},$$

and S_L is the sum of *large* terms including all the other terms.

Bounding $\|S_L\|$: This is bounded in a similar way to the argument for bounding S_L in the previous claim. From RIP property (RIP), we know that noise matrices $E_A := [\varepsilon_A^1, \ldots, \varepsilon_A^n]$ and $E_B := [\varepsilon_B^1, \ldots, \varepsilon_B^n]$ satisfy the weak RIP condition with high probability. Let L be the set of large terms in the proposed partitioning, and $E_A[L]$, $E_B[L]$ be the matrices E_A, E_B restricted to the columns indexed by L. Applying Lemma 59, we have

$$|L| \le \frac{2d}{25\log^2 d}.$$

Therefore, RIP property of E_A and E_B implies that $E_A[L]$ and $E_B[L]$ have spectral norm bounded by 2. Applying triangle inequality, we have

$$\|S_L\| \le \frac{1}{n}\sum_{i\in L} |\langle u, \varepsilon_A^i\rangle| \cdot |\langle v, \varepsilon_B^i\rangle| \le \frac{1}{n}\left\|E_A[L]^\top u\right\| \cdot \left\|E_B[L]^\top v\right\| \le \frac{4}{n},$$

where Cauchy-Schwartz inequality is exploited in the second inequality, and the bounds $\|E_A[L]\| \le 2$ and $\|E_B[L]\| \le 2$ are used in the last inequality. Notice the two matrices are already sampled before we do the ε-net argument, and therefore, we do not need to do union bound over all u, v for this event.

Bounding $\|S_{L^c}\|$: Similar to how we bounded $|S_{L^c}|$ in the previous claim by applying Bernstein's inequality, it is tempting to apply vector Bernstein's inequality here. However, vector Bernstein's inequality does not utilize the fact that the matrix $C^\top$ has small $2 \to 3$ norm, and results in a suboptimal bound. Here, we try to exploit this additional property to to get a better bound.

Let L^c denote the set of small terms in the proposed partitioning on columns of E_A and E_B. Then, we have

$$\langle S_{L^c}, w\rangle = \frac{1}{n}\sum_{i\in L^c} \eta_i \langle u, \varepsilon_A^i\rangle\langle v, \varepsilon_B^i\rangle\langle w, c_{h_i}\rangle.$$

Now, we try to bound the above inner product $\langle S_{L^c}, w\rangle$ by considering an ε-net on w as well (Note that the ε-net on u and v are already considered). To do that we partition the inner products $\langle c_j, w\rangle$ into $t+1$ buckets ($t := \lceil \log_2 \sqrt{d}\rceil$) as defined in Definition 60 where

$$K_0 := \left\{j \in [k] : |\langle c_j, w\rangle| \le \frac{1}{\sqrt{d}}\right\},$$
$$K_l := \left\{j \in [k] : |\langle c_j, w\rangle| \in \left(\frac{2^{l-1}}{\sqrt{d}}, \frac{2^l}{\sqrt{d}}\right]\right\}, \quad l \in [t].$$

Let Q_l denote the sum of all terms that fall into bucket K_l as

$$Q_l := \frac{1}{n} \sum_{i \in L^c, h_i \in K_l} \eta_i \langle u, \varepsilon_A^i \rangle \langle v, \varepsilon_B^i \rangle \langle w, c_{h_i} \rangle. \tag{56}$$

Note that by construction of buckets, we have

$$\langle S_{L^c}, w \rangle = \sum_{l=0}^{t} Q_l.$$

There are only $O(\log d)$ terms in this summation, and therefore, it suffices to show each term Q_l is small.

For Q_0, it is a weighted sum of η_i's with weights bounded by $\tilde{O}(1/d^{3/2})$, so the situation is exactly the same as Claim 5.

For $Q_l, l \in [t]$, the argument is as follows. Let $p_l := 2^{l-1}$. Applying Lemma 61, we have

$$|K_l| \leq O\left(\frac{d^{3/2}}{p_l^3}\right).$$

As stated in the beginning of proof, each hidden state $h_i \in [k]$ appears in at most $O(2nw_{\max})$ samples w.h.p. Hence, the total number of terms in the summation form (56) for Q_l is w.h.p. bounded as

$$|\{i \in [n] : h_i \in K_l\}| \leq O\left(nw_{\max}\frac{d^{3/2}}{p_l^3}\right).$$

Now the sum Q_l in (56) is a weighted sum of η_i's and the Bernstein's inequality is exploited to bound it. Each term in the summation is bounded as

$$\tilde{O}\left(\frac{p_l}{nd^{3/2}}\right),$$

where the bound on the small terms and the bound on terms in bucket K_l are exploited. The variance term is also bounded as

$$O\left(\frac{w_{\max}}{np_l d^{3/2}}\right).$$

Applying Bernstein's inequality, with probability at least $1 - e^{-Cd\log n}$ for large enough constant C, we have

$$Q_l \leq \tilde{O}\left(\frac{p_l}{\sqrt{d}n} + \sqrt{\frac{w_{\max}}{np_l\sqrt{d}}}\right) \leq \tilde{O}\left(\frac{1}{n} + \sqrt{\frac{w_{\max}}{n\sqrt{d}}}\right).$$

At this point, we have bounds on $\|S_L\|$ and $\|S_{L^c}\|$ for a fixed pair of vectors (u, v) in the ε-net. By applying union bound on all vectors in the ε-net, the bound holds for every pair (u, v) in the ε-net. The argument for other (u, v)'s which are not in the ε-net follows from their closest pairs in the ε-net. □

Now we are ready to bound the last term (52c).

Claim 7 (Bounding norm of (52c)) *With high probability over ε_A^i's and h_i's, we have*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \varepsilon_A^i \otimes b_{h_i} \otimes c_{h_i} \right\| \leq \tilde{O}\left(\frac{1}{n} + \sqrt{\frac{w_{\max}}{n}}\right).$$

Proof: Again, rewrite the tensor as

$$T_3 = \frac{1}{n}\sum_{i=1}^{n} \eta_i \varepsilon_A^i \otimes b_{h_i} \otimes c_{h_i}, \tag{57}$$

where η_i's are independent random ± 1 variables with $\Pr[\eta_i = 1] = 1/2$. First sample ε_A^i and h_i's, and therefore, the remaining random variables are just the η_i's. In addition, assume $E_A := [\varepsilon_A^1, \varepsilon_A^2, \ldots, \varepsilon_A^n]$ satisfies the RIP property (RIP) and each $h_i \in [k]$ appears between $nw_{\min}/2$ and $2nw_{\max}$ times where both events happen with high probability.

The goal is to bound norm of T_3 in (57) which is defined as

$$\|T_3\| := \sup_{\|u\|=\|v\|=\|w\|=1} |T_3(u,v,w)| = \sup_{\|u\|=\|v\|=\|w\|=1} \left| \frac{1}{n}\sum_{i=1}^{n} \langle u, \varepsilon_A^i\rangle \langle v, b_{h_i}\rangle \langle w, c_{h_i}\rangle \right|. \tag{58}$$

In order to bound the above, we provide an ε-net argument similar to what we did for bounding S_{L^c} in the previous claim with the difference that here we apply bucketing to all three matrices E_A, B and C. First, for any fixed triple (u,v,w), we partition the inner products in (58) into buckets as defined in Definition 60. Let K_l^a, K_l^b and K_l^c denote the bucketing of matrices E_A, B and C, respectively.

In addition, we merge the buckets $K_0^a, K_1^a, \ldots, K_{4\log\log d}^a$ into K_0^a. This means K_0^a now contains all i's with inner product

$$|\langle \varepsilon_A^i, u\rangle| \le \frac{16\log d}{\sqrt{d}},$$

and K_l^a's for $1 \le l \le 4\log\log d$ are empty. Let

$$J_{l_1,l_2,l_3} := \left\{ i \in [n] : i \in K_{l_1}^a \wedge h_i \in K_{l_2}^b \wedge h_i \in K_{l_3}^c \right\},$$

and Q_{l_1,l_2,l_3} be the sum of terms in summation (58) on this set, i.e.,

$$Q_{l_1,l_2,l_3} := \frac{1}{n}\sum_{i\in J_{l_1,l_2,l_3}} \langle u, \varepsilon_A^i\rangle \langle v, b_{h_i}\rangle \langle w, c_{h_i}\rangle. \tag{59}$$

Note that by construction of buckets, the summation in (58) is expanded as

$$\frac{1}{n}\sum_{i=1}^{n} \langle u, \varepsilon_A^i\rangle \langle v, b_{h_i}\rangle \langle w, c_{h_i}\rangle = \sum_{l_1,l_2,l_3=0}^{t} Q_{l_1,l_2,l_3}.$$

There are only $O(t^3) = O(\log^3 d)$ terms in this summation, and therefore, it suffices to show each term Q_{l_1,l_2,l_3} is small.

For $Q_{0,0,0}$, it is a weighted sum of η_i's with weights bounded by $\tilde{O}(1/d^{3/2})$, and therefore, it follows from the same arguments as Claim 5.

For Q_{l_1,l_2,l_3} with $\max\{l_1,l_2,l_3\} > 0$, let $p_l := 2^{\max\{l_1,l_2,l_3\}-1}$. By Lemma 61 and Lemma 62, the total number of terms in the summation form (59) for Q_{l_1,l_2,l_3} is w.h.p. bounded as

$$|J_{l_1,l_2,l_3}| \le O\left(nw_{\max} \frac{d^{3/2}}{p_l^3} \right),$$

and there exists an ε-net of size

$$\exp\left(O\left(\frac{d^{3/2}}{p_l^3}\log n\right)\right)$$

with $\varepsilon < 1/n^2$. For every u, v, w in the ε-net, this term $n \cdot Q_{l_1,l_2,l_3}$ is a weighted sum of η_i's, and the Bernstein's inequality is exploited to bound it. Each term in the summation is bounded as $\frac{8p_l^3}{d^{3/2}}$, where the bound on the terms in buckets are exploited. The variance term is also bounded as

$$O\left(nw_{\max}\frac{p_l^3}{d^{3/2}}\right).$$

Applying Bernstein's inequality, with probability at least $1 - \exp\left(-C\frac{d^{3/2}}{p_l^3}\log n\right)$ for large enough constant C, we have

$$nQ_{l_1,l_2,l_3} \leq \tilde{O}\left(1 + \sqrt{nw_{\max}}\right).$$

Taking the union bound over all triples in ε-net, this bound holds for all such triples. For u, v, w which are not in the ε-net, the bound follows from the closest point in the ε-net.

□

N.2. ICA

For ICA, the tensor we are considering is given in Equation (17).

Lemma 63 *Suppose entries of h are independent subgaussian variables with $\mathbb{E}[h_i] = 1$. Given n samples $x^i = Ah^i$ where $\|A\| \leq O(\sqrt{k/d})$, let $W = \frac{1}{n}\sum_{i=1}^n x^i(x^i)^\top$, $\widehat{T}_{i_1,i_2,i_3,i_4} = W_{i_1,i_2}W_{i_3,i_4} + W_{i_1,i_3}W_{i_2,i_4} + W_{i_1,i_4}W_{i_2,i_3}$ and $\widehat{M}_4 = \frac{1}{n}\sum_{i=1}^n (x^i)^{\otimes 4} - \widehat{T}$. Let M_4 be defined as Equation (17), then when $n \geq d$ with high probability $\|\widehat{M}_4 - M_4\| \leq \tilde{O}\left(\frac{k^2}{n} + \sqrt{\frac{k^4}{d^3 n}}\right)$.*

The proof directly follows from Claims 9 and 10, which bound the perturbation of the two terms separately.

Before bounding the 4-th order term we first give the following claim which bounds a sum of subgaussian variables raised to the 4-th power.

Claim 8 *Suppose x_i's are independent q-subgaussian variables, then for any $d > 10$, $|\sum_{i=1}^n x_i^4 - \mathbb{E}[\sum_{i=1}^n x_i^4]| \leq \tilde{O}(q^4 d^2/n + \sqrt{q^8 d/n})$ with probability $\exp(-d)$.*

Proof: We shall prove $\Pr[|\sum_{i=1}^n x_i^4 - \text{med}(\sum_{i=1}^n x_i^4)|] \leq \tilde{O}(q^4 d^2/n + \sqrt{q^8 d/n}) \leq \exp(-\omega(d\log n))$, where med is the median of the distribution. This (stronger) concentration implies $\mathbb{E}[|\sum_{i=1}^n x_i^4 - \text{med}(\sum_{i=1}^n x_i^4)|] \leq \tilde{O}(q^4/n)$ (by simple integration), therefore we get the desired bound in the claim.

For the deviation from the median, we use the standard symmetrization argument: it is good enough to take two independent samples $x^1, x^2, ..., x^n$ and $y^1, y^2, ..., y^n$ with the same distribution, and bound $|\frac{1}{n}\sum_{i\in[n]} x_i^4 - y_i^4|$. In order to bound the sum, we rewrite it in the following form

$$Q = \frac{1}{n}\sum_{i\in[n]} \eta_i |x_i^4 - y_i^4|.$$

Now we divide the sum into multiple buckets according to the magnitude of $x_i^4 - y_i^4$. Let $t := \lceil \log_2 d^2 + 10 \log_2 \log_2 n \rceil$ (where C' is a large enough constant), then the buckets are defined as

$$\begin{aligned} K_0 &:= \left\{ i \in [n] : |x_i^4 - y_i^4| \le q^4 \right\}, \\ K_l &:= \left\{ i \in [n] : |x_i^4 - y_i^4| \in \left(2^{l-1} q^4, 2^l q^4 \right] \right\}, \quad l \in [t]. \\ K_{t+1} &:= \left\{ i \in [n] : |x_i^4 - y_i^4| > 2^t q^4 \right\} \end{aligned}$$

Let Q_l denote the sum of all terms that fall into bucket K_l, use $\mathbf{1}_{K_l}$ as the indicator variable for K_l, then Q_l can be written as

$$Q_l := \frac{1}{n} \sum_{i \in [n]} \mathbf{1}_{K_l} |x_i^4 - y_i^4| \eta_i. \tag{60}$$

Note that by construction of buckets, the original summation is equal to $\sum_{l=0}^{t+1} Q_l$. There are only $O(\log d)$ terms in this summation. Therefore it suffices to show each term Q_l is small.

Since the variables x_i's and y_i's are independent 1-subgaussian random variables, we have

$$\Pr\left[|x_i^4 - y_i^4| \ge \lambda q^4 \right] \le 2 \Pr\left[|x_i^4| \ge (\lambda q/2)^{1/4} \right] \le 4 \exp\left(-\frac{\sqrt{\lambda}}{8} \right), \tag{61}$$

Where the last inequality uses q-subgausian property.

For $Q_l (0 \le l \le 2 \log \log n)$, we apply Bernstein's inequality directly. Each term in the summation is bounded as $\tilde{O}(q^4)$, and the variance term is bounded as $\tilde{O}(q^8/n)$. By Bernstein's inequality with probability at least $1 - \exp(-\omega d \log n)$, we have

$$Q_0 \le \tilde{O}\left(\frac{q^4 d}{n} + \sqrt{\frac{q^8 d}{n}} \right).$$

For $Q_l, 2 \log \log n < l \le t$, we bound the number of terms in bucket K_l. By the bound in Equation (61), we know the probability that there are more than $\tilde{O}(d 2^{-l/2})$ items in K_l is bounded by $\exp(-\omega d \log n)$. Each term in Q_l is bounded by $2^l q^4/n$, therefore the sum Q_l is bounded by

$$n Q_l \le \tilde{O}(d 2^{-l/2}) 2^l q^4 \le \tilde{O}(q^4 d 2^{l/2}) \le \tilde{O}(q^4 d^2).$$

Here the last inequality uses the fact that $l \le t$, which implies $2^{l/2} = \tilde{O}(d)$.

For the last term Q_{t+1}, again by Equation (61, we know with probability $1 - \exp(-\omega d \log n)$, there is only one term in the sum and that particular term is smaller than $\tilde{O}(q^4 d^2/n)$.

Now by union bound, with probability $1 - \exp(-\omega d \log n)$ all the terms are bounded by $\tilde{O}(q^4 d^2/n + \sqrt{q^8 d/n})$, which implies the whole sum is bounded by $\tilde{O}(q^4 d^2/n + \sqrt{q^8 d/n})$. □

Now we are ready bound the 4-th order term.

Claim 9 *Suppose entries of h are independent subgaussian variables with $\mathbb{E}[h_i] = 1$, given n samples $x^i = Ah^i$ where $\|A\| \le O(\sqrt{k/d})$, with high probability*

$$\left\| \frac{1}{n} \sum_{i \in [n]} (x^i)^{\otimes 4} - \mathbb{E}[x^{\otimes 4}] \right\| \le \tilde{O}\left(\frac{k^2}{n} + \sqrt{\frac{k^4}{d^3 n}} \right).$$

Proof: The goal is to bound the deviation of

$$\frac{1}{n}\sum_{i\in[n]}(x^i)^{\otimes 4} = \frac{1}{n}\sum_{i\in[n]}(Ah^i)^{\otimes 4}$$

from its mean. We bound this by an ε-net argument.

Construct an ε-net for the unit ball in $\mathbb{R}^d$ with $\varepsilon = 1/n^2$, the size of the ε-net is $\exp(O(d\log n))$. For any fixed u in the ε-net, let $v = A^\top u$, we know $\langle u, x^i\rangle = \langle u, Ah^i\rangle = \langle v, h^i\rangle$. Therefore, for any u in the ε-net, we would like to bound

$$Q = \frac{1}{n}\sum_{i\in[n]}(\langle v, h^i\rangle^4 - \mathbb{E}[\langle v, h^i\rangle^4]).$$

Notice that $\|v\|$ is bounded by $\|A\|\|u\| = O(\sqrt{k/d})$, and h^i's have subgaussian entries. Therefore $\langle v, h^i\rangle$ is a $O(\sqrt{k/d})$-subgaussian random variable. By Claim 8, we know $|Q| \le \tilde{O}(k^2/n + \sqrt{k^3/d^3n})$ with probability $\exp(-Cd\log n)$ for large enough constant C. Taking union bound over every u in the ε-net, we know the bound is true for any vector u in the net.

The argument for other vectors u's which are not in the ε-net follows from their closest vector in the ε-net. □

For the extra term T in Equation (17), it can be decomposed into the sum of three terms, each of which is an outerproduct of two matrices. Therefore it is good enough to apply a matrix concentration bound.

Claim 10 *Suppose entries of h are independent subgaussian variables with $\mathbb{E}[h_i] = 1$. Given n samples $x^i = Ah^i$ where $\|A\| \le O(\sqrt{k/d})$, let $W = \frac{1}{n}\sum_{i=1}^n x^i(x^i)^\top$, $\widehat{T}_{i_1,i_2,i_3,i_4} = W_{i_1,i_2}W_{i_3,i_4} + W_{i_1,i_3}W_{i_2,i_4} + W_{i_1,i_4}W_{i_2,i_3}$. Let T be defined as in Equation 17), then with high probability when $n \ge d$ $\|\widehat{T} - T\| \le \tilde{O}\left(\sqrt{\frac{k^4}{d^3n}}\right)$.*

Proof: For simplicity we consider one of the terms $\widehat{T}_1[i_1, i_2, i_3, i_4] = W_{i_1,i_2}W_{i_3,i_4} := W \otimes W$, all the terms follow from symmetry.

Let $T_1 = \mathbb{E}[xx^\top] \otimes \mathbb{E}[xx^\top] = \mathbb{E}[W] \otimes \mathbb{E}[W]$. For $\widehat{T}_1$, we know $\widehat{T}_1 - T_1 = (W - \mathbb{E}[W]) \otimes \mathbb{E}[W] + \mathbb{E}[W] \otimes (W - \mathbb{E}[W]) + (W - \mathbb{E}[W]) \otimes (W - \mathbb{E}[W])$. By property of the outerproduct we know $\|A \otimes B\| \le \|A\|\|B\|$ for all matrices A, B, therefore

$$\|\widehat{T}_1 - T_1\| \le 2\|W - \mathbb{E}[W]\|\|\mathbb{E}[W]\| + \|W - \mathbb{E}[W]\|^2. \tag{62}$$

We bound $\|W - \mathbb{E}[W]\|$ by Matrix Bernstein's inequality. For technical reasons we first construct $W' = \frac{1}{n}\sum_{i=1}^n x^i(x^i)^\top \mathbf{1}_{\|x^i\|\le O(\sqrt{k}logn)}$ where $\mathbf{1}_{\|x^i\|\le O(\sqrt{k}\log n)}$ is an indicator variable. Since $x = Ah$ and entries of h are subgaussian, these variables are 1 with probability $1 - n^{-\log n}$, therefore W and W' are equal with high probability at it suffices to apply Matrix Bernstein's bound on W'.

For W', each term has norm bounded by $\tilde{O}(k)$, and the variance term $\mathbb{E}[W'(W')^\top]$ is equal to

$$\frac{1}{n}\mathbb{E}[\|x^i\|^2 x^i(x^i)^\top \mathbf{1}_{\|x^i\|\le O(\sqrt{k}\log n)}] \preceq \frac{1}{n}\tilde{O}(k)\mathbb{E}[x^i(x^i)^\top] = \frac{1}{n}\tilde{O}(k)AA^\top.$$

Also, we know $\|A\| \le O(\sqrt{k/d})$, therefore the variance is bounded by $\tilde{O}(k^2/dn)$. Matrix Bernstein's inequality implies $\|W' - \mathbb{E}[W']\| \le \tilde{O}(k/n + k/\sqrt{dn})$. Since W is equal to W' with

high probability and $\|\mathbb{E}[W] - \mathbb{E}[W']\|$ is negligible, we also know $\|W - \mathbb{E}[W]\| \leq \tilde{O}(k/\sqrt{dn})$ (when $n \geq d$).

On the other hand, $\mathbb{E}[W] = AA^\top$ which has spectral norm k/d. By Equation (62) we know $\|\widehat{T}_1 - T_1\| \leq \tilde{O}\left(\sqrt{\frac{k^4}{d^3 n}}\right)$. □

N.3. Sparse ICA

In this part we prove concentration bounds for sparse coding in the sparse ICA setting (where h_i's are independent and sparse). The proof can be generalized to the case when h_i's are negatively correlated or more generally when concentration bounds hold for h_i's.

Lemma 64 *Suppose $h_i = s_i w_i$ where s_i's are i.i.d. $0/1$ variables with probability s/k of being 1, w_i's are independent 1-subgaussian random variables. Given n independent samples of the form $x^i = Ah^i$ (each h^i is distributed as h), if A satisfy RIP-property (RIP) then $\|\frac{1}{n}\sum_{i=1}^n (x^i)^{\otimes 4} - \mathbb{E}[(x^i)^{\otimes 4}]\| \leq \tilde{O}(s^2/n + \sqrt{s^4/d^3 n})$.*

Proof: The proof uses ideas from both Claim 7 and 8 . Without loss of generality we assume $s/k < 1/2$ (otherwise h_i's are 2-subgaussian) so Claim 9 implies the desired bound).

Here, we first partition the entries of vector $v = A^\top u \in \mathbb{R}^k$ into different vectors v_l according to the magnitude of entries (this is very similar to Claim 7). In particular, we partition entries (inner products) $v_j = \langle u, a_j \rangle, j \in [k]$, into $t+1$ buckets ($t := \lceil \log_2 \sqrt{d} \rceil$) where (similar to Definition 60)

$$K_0 := \left\{ j \in [k] : |\langle u, a_j \rangle| \leq \frac{1}{\sqrt{d}} \right\},$$
$$K_l := \left\{ j \in [k] : |\langle u, a_j \rangle| \in \left(\frac{2^{l-1}}{\sqrt{d}}, \frac{2^l}{\sqrt{d}} \right] \right\}, \quad l \in [t].$$

In addition, we merge the buckets $K_0, K_1, \ldots, K_{\frac{1}{2}\log\log d}$ into K_0. This means K_0 now contains all j's with inner product

$$|\langle u, a_j \rangle| \leq \frac{\sqrt{\log d}}{\sqrt{d}},$$

and K_l's for $1 \leq l \leq \frac{1}{2}\log\log d$ are empty. Now, let v_l denote the restriction of vector v to entries indexed by K_l, i.e.,

$$v_l(j) := \begin{cases} v(j), & j \in K_l, \\ 0, & j \notin K_l. \end{cases}$$

Let $p_l := 2^{l-1}$. By RIP property of matrix A, and exploiting Lemma 62, the number of nonzero entries in v_l is bounded as

$$\|v_l\|_0 = |K_l| \leq O\left(\frac{d}{p_l^2}\right), \quad l > \frac{1}{2}\log\log d.$$

Now, we follow the ideas of Claim 8, and apply the symmetrization trick to see that it is good enough to bound the $\|\frac{1}{n}\sum_{i=1}^n (x^i)^{\otimes 4} - (y^i)^{\otimes 4}\|$ where y^i is an independent copy of x^i (the difference between mean and median here is negligible because our distributions have first and second

moments polynomial in parameters, and strong exponential concentration). For any vector u (and the corresponding v), we would like to bound the sum

$$Q = \frac{1}{n}\sum_{i=1}^{n} \eta_i |\langle v, x^i\rangle^4 - \langle v, y^i\rangle^4|.$$

The techniques we used to prove bounds on $\sum_{i=1}^n \eta_i w_i$ (either Bernstein's inequality, or bounding the number of terms and use triangle inequality) all works if we just know an *upperbound* of w_i. Therefore we can safely replace Q by Q':

$$Q' = \frac{1}{n}\sum_{i=1}^{n} \eta_i (t+1)^3 \sum_{l=0}^{t} |\langle v_l, x^i\rangle^4 + \langle v_l, y^i\rangle^4|.$$

Here the corresponding coefficient $(t+1)^3\sum_{l=0}^t |\langle v_l, x^i\rangle^4 + \langle v_l, y^i\rangle^4|$ is larger than $|\langle v, x^i\rangle^4 + \langle v, y^i\rangle^4|$ because $(\sum_{i=1}^n a_i)^4 \le (n\sum_{i=1}^n a_i^2)^2 \le n^3(\sum_{i=1}^n a_i^4)$ (the two steps are Cauchy-Schwartz).

We then break Q' into the sum of $t+1$ terms $Q'_0, Q'_1, ..., Q'_l$, where

$$Q'_l = (t+1)^3 \frac{1}{n}\sum_{i=1}^{n} \eta_i |\langle v_l, x^i\rangle^4 + \langle v_l, y^i\rangle^4|.$$

All these terms can be bounded in the same way as Claim 8. Especially, $Q'_0 \le \tilde{O}(s^2 + \sqrt{s^4/d^3 n})$ directly from Claim 8. For the other terms, we need the tail behavior of $\langle v_l, x^i\rangle^4$. The tail behavior of this variable comes from two phenomena: the first is how large is the intersection of the supports of v_l and x^i, the second is given the intersection the tail behavior of the sum of subgaussian variables $\sum_j v_l[j]w^i[j]$ (recall that $x^i[j] = s^i[j]w^i[j]$ where $s^i[j]$ specifies support). The first part (the intersection of support) can be bounded by Chernoff bound $\Pr[\sum_j s_j \ge (1+\delta)\mu] \le (e^\delta/(1+\delta)^{(1+\delta)})^\mu$; the second part would just follow from subgaussian bounds. Suppose we are interested in a bucket Q'_l where v_l has entries in $(\theta/2, \theta]$, we discuss the tail behavior in cases where $1/\theta^2 \ge s$ and $1/\theta^2 \le s$.

In the first case ($1/\theta^2 \ge s$) most of $\langle v_l, x^i\rangle^4$ are of size s^2/k^2 which is very small. For any $q \in [\sqrt{s/k\theta^2}\,\mathrm{poly}\log n, s]$, the probability that $\langle v_l, x^i\rangle^4 \in (q^4\theta^4/2, q^4\theta^4]$ is $\exp(-\tilde{\Omega}(q))$. In this range with probability $\exp(-\tilde{O}(1/\theta^2))$, the sum of all terms is bounded by $\frac{1}{n}\tilde{O}(q^4\theta^4 \cdot (1/\theta^2 q)) = \tilde{O}(q^3\theta^2/n) \le \tilde{O}(s^2/n)$ (the last inequality uses the fact that $\theta^2 \le 1/s$). For $q \in (s, \sqrt{s/\theta^2}\log^2 n]$, the probability that $\langle v_l, x^i\rangle^4 \in (q^4\theta^4/2, q^4\theta^4]$ is $\exp(-\tilde{\Omega}(q^2/s))$. In this range $\exp(-\tilde{O}(1/\theta^2))$, the sum of all terms is bounded by $\frac{1}{n}\tilde{O}(q^4\theta^4 \cdot (1/\theta^2)/(q^2/s)) = \tilde{O}(q^2\theta^2 s/n) \le \tilde{O}(s^2/n)$ (where the last inequality uses the fact that $q^2 = \tilde{O}(s/\theta^2)$). When $q > \sqrt{s/\theta^2}\log^2 n$ with high probability there are no terms in this range. Therefore, in the first case, by union bound Q'_l is always bounded by $\tilde{O}(s^2/n) + o(s^4/d^3 n)$.

In the second case ($1/\theta^2 \le s$) again most of $\langle v_l, x^i\rangle^4$ are of size s^2/k^2 which is very small. The only difference in this case is the two ranges: instead of separated at s they are separated at $1/\theta^2$ because there are at most $\tilde{O}(1/\theta^2)$ entries in v_l. For any $q \in [\sqrt{s/k\theta^2}\,\mathrm{poly}\log n, 1/\theta^2]$, the probability that $\langle v_l, x^i\rangle^4 \in (q^4\theta^4/2, q^4\theta^4]$ is $\exp(-\tilde{\Omega}(q))$. In this range with probability $\exp(-\tilde{O}(1/\theta^2))$, the sum of all terms is bounded by $\frac{1}{n}\tilde{O}(q^4\theta^4 \cdot (1/\theta^2 q)) = \tilde{O}(q^3\theta^2/n) \le \tilde{O}(s^2/n)$ (the final inequality uses the fact that $1/\theta^2 \le s$). For $q \in (1/\theta^2, \sqrt{s/\theta^2}\log^2 n]$, the probability that $\langle v_l, x^i\rangle^4 \in (q^4\theta^4/2, q^4\theta^4]$ is $\exp(-\tilde{\Omega}(q^2\theta^2))$. In this range $\exp(-\tilde{O}(1/\theta^2))$, the sum of all

terms is bounded by $\frac{1}{n}\tilde{O}(q^4\theta^4 \cdot (1/\theta^2)/(q^2\theta^2)) = \tilde{O}(q^2/n) \leq \tilde{O}(s^2/n)$ (where the last inequality uses the fact that $q^2 = \tilde{O}(s/\theta^2) = \tilde{O}(s^2)$). When $q > \sqrt{s/\theta^2}\log^2 n$ with high probability there are no terms in this range. Therefore, in the first case, by union bound Q'_l is always bounded by $\tilde{O}(s^2/n) + o(s^4/d^3n)$.

Combining all the terms, we know the sum is bounded by $\tilde{O}(s^2/n + \sqrt{s^4/d^3n})$ as desired. □

Remark 65 *It may seem counter-intuitive that the bound in Lemma 64 does not depend on k. The dependency on k is actually in the expectation: the expected tensor will be close to $\frac{s}{k}\sum_{i=1}^{k} a_i^{\otimes 4}$. Therefore we need the error to be much smaller than $\frac{s}{k}$ even in the semi-supervised setting. The number of samples required is roughly $O(sk)$ for small s which agrees with our intuition in the mixture model. The number of samples when s is roughly k also matches the ICA bound in Lemma 63*

When the hidden variables are really independent, we are in the sparse ICA model instead of sparse coding. In this case we can use the same formula as Equation (17) to get a low rank tensor. In the next claim we bound the perturbation of the T term.

Claim 11 *Suppose $h_i = s_i w_i$ where s_i's are i.i.d. $0/1$ variables with probability s/k of being 1, w_i's are independent 1-subgaussian random variables. Given n samples $x^i = Ah^i$ where $\|A\| \leq O(\sqrt{k/d})$, let $W = \frac{1}{n}\sum_{i=1}^{n} x^i(x^i)^\top$,$\widehat{T}_{i_1,i_2,i_3,i_4} = W_{i_1,i_2}W_{i_3,i_4} + W_{i_1,i_3}W_{i_2,i_4} + W_{i_1,i_4}W_{i_2,i_3}$. Let T be defined as in Equation 17), then with high probability when $n \geq d$ $\|\widehat{T} - T\| \leq \tilde{O}\left(\sqrt{\frac{s^4}{d^3n}}\right)$.*

Proof: The proof is very similar to Claim 10.

Using the same idea, we consider one of the terms $\widehat{T}_1[i_1, i_2, i_3, i_4] = W_{i_1,i_2}W_{i_3,i_4} := W \otimes W$, all the terms follow from symmetry. Similar to Equation 62, we have the following fact

$$\|\widehat{T}_1 - T_1\| \leq 2\|W - \mathbb{E}[W]\|\|\mathbb{E}[W]\| + \|W - \mathbb{E}[W]\|^2.$$

We bound $\|W - \mathbb{E}[W]\|$ by Matrix Bernstein's inequality. For technical reasons we first construct $W' = \frac{1}{n}\sum_{i=1}^{n} x^i(x^i)^\top \mathbf{1}_{\|x^i\| \leq O(\sqrt{s}\log n)}$ where $\mathbf{1}_{\|x^i\| \leq O(\sqrt{s}\log n)}$ is an indicator variable. Since $x = Ah$ and entries of h are subgaussian, these variables are 1 with probability $1 - n^{-\log n}$, therefore W and W' are equal with high probability at it suffices to apply Matrix Bernstein's bound on W'.

For W', each term has norm bounded by $\tilde{O}(s)$, and the variance term $\mathbb{E}[W'(W')^\top]$ is equal to

$$\frac{1}{n}\mathbb{E}[\|x^i\|^2 x^i(x^i)^\top \mathbf{1}_{\|x^i\| \leq O(\sqrt{s}\log n)}] \preceq \frac{1}{n}\tilde{O}(s)\mathbb{E}[x^i(x^i)^\top] = \frac{1}{n}\tilde{O}(s^2/k)AA^\top.$$

Also, we know $\|A\| \leq O(\sqrt{k/d})$, therefore the variance is bounded by $\tilde{O}(s^2/dn)$. Matrix Bernstein's inequality implies $\|W' - \mathbb{E}[W']\| \leq \tilde{O}(s/n + s/\sqrt{dn})$. Since W is equal to W' with high probability and $\|\mathbb{E}[W] - \mathbb{E}[W']\|$ is negligible, we also know $\|W - \mathbb{E}[W]\| \leq \tilde{O}(s/\sqrt{dn})$ (when $n \geq d$).

On the other hand, $\mathbb{E}[W] = \frac{s}{k}AA^\top$ which has spectral norm s/d. By Equation (62) we know $\|\widehat{T}_1 - T_1\| \leq \tilde{O}\left(\sqrt{\frac{s^4}{d^3n}}\right)$. □