Sample Complexity Analysis for Learning Overcomplete Latent Variable Models through Tensor Methods

Anima Anandkumar* Rong Ge† Majid Janzamin‡

January 22, 2016

Abstract

We provide guarantees for learning latent variable models emphasizing on the overcomplete regime, where the dimensionality of the latent space can exceed the observed dimensionality. In particular, we consider multiview mixtures, spherical Gaussian mixtures, ICA, and sparse coding models. We provide tight concentration bounds for empirical moments through novel covering arguments. We analyze parameter recovery through a simple tensor power update algorithm. In the semi-supervised setting, we exploit the label or prior information to get a rough estimate of the model parameters, and then refine it using the tensor method on unlabeled samples. We establish that learning is possible 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. Our concentration bound analysis also leads to minimax sample complexity for semi-supervised learning of spherical Gaussian mixtures. In the unsupervised setting, we use a simple initialization algorithm based on SVD of the tensor slices, and provide guarantees under the stricter condition that $k \leq \beta d$ (where constant β can be larger than 1), where the tensor method recovers the components under a polynomial running time (and exponential in β). Our analysis establishes that a wide range of overcomplete latent variable models can be learned efficiently with low computational and sample complexity through tensor decomposition methods.

Keywords: Unsupervised and semi-supervised learning, latent variable models, overcomplete representation, tensor decomposition, sample complexity analysis.

1 Introduction

It is imperative to incorporate latent variables in any modeling framework. Latent variables can capture the effect of hidden causes which are not directly observed. Learning these hidden factors is central to many applications, e.g., identifying the latent diseases through observed symptoms, identifying the latent communities through observed social ties, and so on. Moreover, latent variable models (LVMs) can provide an efficient representation of the observed data, and learning these representations can lead to improved performance on various tasks such as classification. The recent performance gains in domains such as speech and computer vision can be largely attributed

*University of California, Irvine. Email: a.anandkumar@uci.edu

†Microsoft Research, New England. Email: rongge@microsoft.com

‡University of California, Irvine. Email: mjanzami@uci.edu

to efficient representation learning (Bengio et al., 2012). Moreover, it has been shown that learning overcomplete representations is crucial to achieving these impressive gains (Coates et al., 2011b).

In an overcomplete representation, the dimensionality of the latent space exceeds the observed dimensionality. Overcomplete representations are known to be more robust to noise, and can provide greater flexibility in modeling (Lewicki and Sejnowski, 2000). Although overcomplete representations have led to huge performance gains in practice, theoretical guarantees for learning are mostly lacking. In many domains, we face the challenging task of unsupervised or semi-supervised learning, since it is expensive to obtain labeled samples and we typically have access to a large number of unlabeled samples, e.g. (Coates et al., 2011b; Le et al., 2011). Therefore, it is imperative to develop novel guaranteed methods for efficient unsupervised/semi-supervised learning of overcomplete models.

In this paper, we bridge the gap between theory and practice, and establish that a wide range of overcomplete LVMs can be learned efficiently through simple spectral learning techniques. We perform spectral decomposition of the higher order moment tensors (estimated using unlabeled samples) to obtain the model parameters. A recent line of work has shown that tensor decompositions can be employed for unsupervised learning of a wide range of LVMs, e.g., independent components (De Lathauwer et al., 2007), topic models, Gaussian mixtures, hidden Markov models (Anandkumar et al., 2014a), network community models (Anandkumar et al., 2013b), and so on. It involves decomposition of a multivariate moment tensor, and is guaranteed to provide a consistent estimate of the model parameters. The sample and computational requirements are only a low order polynomial in the latent dimensionality for the tensor method (Anandkumar et al., 2014a; Song et al., 2013). However, a major drawback behind these works is that they mostly consider the undercomplete setting, where the latent dimensionality cannot exceed the observed dimensionality.

In this work, we establish guarantees for tensor decomposition in learning overcomplete LVMs, such as multiview mixtures, independent component analysis, 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, which limits the redundancy among the components. We establish that this constraint not only makes learning well-posed but also enables efficient learning through tensor methods. 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, Independent Component Analysis (ICA), Gaussian mixtures and sparse coding models. For the learning algorithm, we exploit the tensor decomposition algorithm in (Anandkumar et al., 2014b), which performs alternating asymmetric power updates on the input tensor modes (or performs symmetric power updates if the input tensor is symmetric). Under the semi-supervised setting, we establish that highly overcomplete models can be learned efficiently through tensor decomposition methods. 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 require stricter conditions on the extent of overcompleteness for guaranteed learning. In addition, we provide tight concentration bounds on the empirical tensors through novel covering arguments, which imply efficient sample complexity bounds for learning using the tensor method.

We now summarize the results for learning multiview mixtures model with incoherent components[1]. Let k be the number of hidden components, and d be the observed dimensionality. In the semi-supervised setting, we prove guaranteed learning when $k = o(d^{p/2})$, where p is the order of observed moment employed for tensor decomposition. We prove that in the "low" noise regime (where the norm of noise is of the same order as that of the component means), having an extremely small number of labeled samples for each label is sufficient (scaling as $\operatorname{polylog}(d, k)$ independent of the final precision). This is far less than 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. (2011a). Furthermore, we show that the sample complexity bounds for unlabeled samples is $\tilde{\Omega}(k)$. Note that this is the *minimax* bound up to polylog factors.

We also provide *unsupervised* learning guarantees when no label is available. Here, the initialization is obtained by performing a rank-1 SVD on the random slices of the moment tensor. This imposes additional conditions on rank and sample complexity. We prove that when $k \leq \beta d$ (for arbitrary constant β which can be larger than 1), the model parameters can be learned using a polynomial number of initializations (which depends on β and scales as k^{β^2}) and sample complexity scales as $\tilde{\Omega}(kd)$, which is efficient.

We also provide semi-supervised and unsupervised learning guarantees for ICA model. By semi-supervised setting in ICA, we mean some prior information is available which provides good initializations (with a constant ℓ_2 error on the columns) for the tensor decomposition algorithm. In the semi-supervised setting, we show that when the number of components scales as $k = \Theta(d^2)/\operatorname{polylog}(d)$, 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 $k = \Theta(d)$, the ICA model can be learned with $n \geq \tilde{\Omega}(k^3)$ in time $k^{\Omega(k^2/d^2)}$.

We also provide learning results for the sparse coding model, when the coefficients are independently drawn from a Bernoulli-Gaussian distribution. Note that this corresponds to a sparse ICA model since the hidden coefficients are independent. Let s be the expected sparsity level of the hidden variables. In the semi-supervised setting (where prior information gives good initialization), we require $\tilde{\Omega}(\max\{sk, s^2k^2/d^3\})$ number of unlabeled samples for learning as long as $k = o(d^2)$. Note that in the special case when s is a constant, the sample complexity is akin to learning multiview models, where $s = 1$; and when $s = \Theta(k)$, it is akin to learning the "dense" ICA model, where $s = k$. Thus, the sparse coding model bridges the range of models between multiview mixtures and ICA. Furthermore, we also extend the learning results to dependent sparsity setting, but with worse performance guarantees.

Although we prove strong theoretical guarantees for learning overcomplete models, there are two main caveats for our approach. We recover the model parameters with an *approximation* error, which decays with the dimension d. Concretely, for the p^{th} order tensor, the approximation error is $\tilde{O}\big(\sqrt{k/d^{p-1}}\big)$, which decays since $k = o(d^{p/2})$. This is because the actual mixture components are not the stationary points of the tensor algorithm updates (even in the noiseless setting) since

[1] We use the term incoherence to say that the deterministic condition in the appendix of Anandkumar et al. (2014b) is satisfied which basically imposes soft-orthogonality constraints on the components. It is also shown that this condition is satisfied whp when the components are uniformly i.i.d. drawn from unit sphere.

the components are not strictly orthogonal. This bias can be presumably removed by performing joint updates (e.g alternating least squares) where the objective is to fit the learnt vectors to the input tensor and we leave this for future study. Second, the setting is not suited for topic models, where there is a non-negativity constraint on the topic-word matrix. Here, incoherence can only be enforced through sparsity, and since our method does not exploit sparsity, we believe that other formulations may be better suited for learning in this setting.

Overview of techniques: We establish tight concentration bounds for empirical tensors when the samples are drawn from multiview linear mixtures, Gaussian mixtures, ICA or sparse coding models. The concentration bound involves bounding the spectral norm of the error tensor, and this relies on the construction of *ε-nets* to cover all vectors (on the sphere). A naive ε-net argument is however too loose since it results in a large number of vectors without a "fine-grained" distinction between them. A more refined notion is to employ an *entropy-concentration* trade-off, as proposed in Rudelson and Vershynin (2009), where the vectors in the ε-net are classified into sparse and dense vectors, and to analyze them separately. The sparse vectors can result in large correlations, but the number of such vectors is small, while the dense vectors have small correlations, although their number is larger. In our setting, however, this classification is still not enough, and we need a more refined analysis. We group the data samples into "buckets" based on their correlation with a given vector, and bound each "bucket" separately. We impose additional conditions on the factor and noise matrices to bound the size of the buckets.

For the multiview linear mixtures, we impose a restricted isometry property (RIP) on the noise matrices and a bounded $2 \rightarrow 3$ norm condition on the factor matrices (which is weaker than RIP). For Gaussian mixtures, the RIP property on noise is satisfied, and we only require a condition of bounded $2 \rightarrow 3$ norm on the matrix of component mean vectors. These constraints allow us to bound the size of the "buckets", where each bucket corresponds to noise or factor vectors with a certain level of correlation with a fixed vector. Intuitively, the number of samples having a high correlation with a fixed vector (i.e. size of a "bucket") cannot be too large due to RIP/bounded 2-to-3 norm constraints. We apply Bernstein's bound on each of these buckets separately and combine them to obtain the final bound. Our construction has only a logarithmic number of buckets (since we vary correlation levels geometrically), and therefore the overall concentration bound only has additional logarithmic factors when we combine the results.

For the ICA model, the conditions and analysis are somewhat different. This is because all the hidden sources "mix" together in each sample, in contrast to the mixture model, where each sample is generated from only one component. Establishing concentration bounds involves two steps, viz., first having a bound on the fourth order empirical moment of the hidden sources, assuming they are sub-Gaussian and kurtotic,[2] and then converting the bound to the observed space. This involves a spectral norm bound on the linear map between the hidden sources and the observations.

We then consider the sparse coding model, where the hidden variables are assumed to be sparsely activated. In the special case, when the hidden variables are independent, this corresponds to a sparse ICA model. We derive the concentration bound for Bernoulli-Gaussian variables, assuming that the dictionary has the RIP property (e.g., Gaussian matrix). In this case, we establish that the concentration bound depends only on the sparsity level, and not on the total number of

[2] Note that while the kurtosis (4th order cumulant) of a Gaussian random variable is zero, the kurtotis of sub-Gaussian random variables is in general nonzero. In addition, note that this analysis can be also extended to sub-exponential random variables.

dictionary elements. Here, we partition the vectors into "buckets" based on their correlation with the dictionary elements and the RIP property allows us to bound the size of buckets, as before in the case of multiview mixtures. In addition, we exploit the sparsity of elements to obtain a tighter bound for the sparse coding setting.

Thus, we obtain tight concentration bounds for empirical tensors for multiview and Gaussian mixtures, ICA and sparse coding models. The conditions on noise (RIP) and factor matrices (bounded 2-to-3 norm) are fairly benign and natural to impose. Our novel bucketing arguments could be applicable in other settings involving matrix and tensor concentration bounds.

We then employ the concentration bounds in conjunction with the alternating rank-1 updates algorithm to obtain learning guarantees for the above models. In our recent work (Anandkumar et al., 2014b), we establish local and global convergence guarantees for this algorithm when the components are incoherent. We combine these guarantees with the concentration bounds to establish that a wide range of latent variable models can be learned with low computational and sample complexities.

1.2 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 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 $k = \operatorname{poly}(d)$, 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.

1.3 Notations and tensor preliminaries

Define $[n] := \{1, 2, \ldots, n\}$. Let $\|u\|_p$ denote the ℓ_p norm of vector u, and the induced $q \to p$ norm of matrix A 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 $f(d) \le O(g(d))$ and $g(d) \ge \Omega(f(d))$ for additional clarity. We also use the asymptotic notation $f(d) = \tilde{O}(g(d))$ if and only if $f(d) \le \alpha g(d)$ for all $d \ge d_0$, 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 $f(d) \ge \alpha g(d)$ for all $d \ge d_0$, for some $d_0 > 0$ and $\alpha = \text{polylog}(d)$.

Tensor preliminaries

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

The different dimensions of the tensor are referred to as *modes*. For instance, for a matrix, the first mode refers to columns and the second mode refers to rows. In addition, *fibers* are higher order analogues of matrix rows and columns. A fiber is obtained by fixing all but one of the indices of the tensor (and is arranged as a column vector). For instance, for a matrix, its mode-1 fiber is any matrix column while a mode-2 fiber is any row. For a third order tensor $T \in \mathbb{R}^{d\times d\times d}$, the mode-1 fiber is given by $T(:, j, l)$, mode-2 by $T(i, :, l)$ and mode-3 by $T(i, j, :)$. Similarly, *slices* are obtained by fixing all but two of the indices of the tensor. For example, for the third order tensor T, the slices along 3rd mode are given by $T(:, :, l)$.

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

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

In particular, for vectors $u, v, w \in \mathbb{R}^d$, we have [3]

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

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

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

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

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

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

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

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

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

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

For vector $v \in \mathbb{R}^d$, we define

$$v^{\otimes p} := v \otimes v \otimes \cdots \otimes v \in \bigotimes^p \mathbb{R}^d$$

as its p-th tensor power.

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

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

2 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 CP tensor decomposition form when appropriate modifications are performed. This is done for multiview linear mixtures model, 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).

Therefore, an efficient tensor decomposition method leads to efficient learning procedure for a wide range of latent variable models. In Section 4.1, we provide the tensor decomposition algorithm introduced in Anandkumar et al. (2014b), and exploit it for learning latent variable models providing sample complexity results in the subsequent sections. Note that the sample complexity guarantees are argued through tensor concentration bounds proposed in Section 3.

2.1 Multiview linear mixtures model

Consider a multiview linear mixtures model as in Figure 1 with k components and $p \geq 3$ views. Throughout the paper, we assume $p = 3$ for simplicity, while the results can be also extended to

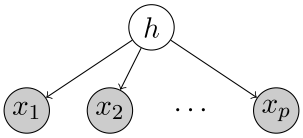

Figure 1: Multi-view mixtures model.

higher-order. 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

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

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.

For this model, the third order observed moment has the form (See Anandkumar et al. 2014a)

$$\mathbb{E}[x_1 \otimes x_2 \otimes x_3] = \sum_{j \in [k]} w_j a_j \otimes b_j \otimes c_j. \tag{5}$$

The decomposition in (5) is referred to as the CP decomposition (Carroll and Chang, 1970), and k denotes the CP tensor rank. 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 (5).

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

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

where $\varepsilon_A, \varepsilon_B, \varepsilon_C \in \mathbb{R}^d$ 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, let all the vectors $a_h, b_h, c_h, h \in [k]$, have unit ℓ_2 norm. Furthermore, since w_j's are the mixture probabilities, for simplicity we consider $w_j = \Theta(1/k), j \in [k]$. We call this model $\mathcal{S}$.

When $\zeta^2 = \Theta(1/d)$, the norm of the noise is roughly the same as the norm of the components. We call this the *low noise regime.* When $\zeta^2 = \Theta(1)$, the norm of noise in *every dimension* is roughly the same as the norm of the components. We call this the *high noise regime.*

2.2 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 1 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 1 (Hsu and Kakade 2012)**.** *If*

$$M_3 := \mathbb{E}[x \otimes x \otimes x] - \zeta^2 \sum_{i \in [d]} \left(\mathbb{E}[x] \otimes e_i \otimes e_i + e_i \otimes \mathbb{E}[x] \otimes e_i + e_i \otimes e_i \otimes \mathbb{E}[x] \right), \tag{6}$$

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 (6) is

$$\widehat{M}_3 := \widehat{\mathcal{M}}_3 - \zeta^2 \sum_{i \in [d]} \left(\widehat{\mathcal{M}}_1 \otimes e_i \otimes e_i + e_i \otimes \widehat{\mathcal{M}}_1 \otimes e_i + e_i \otimes e_i \otimes \widehat{\mathcal{M}}_1 \right). \tag{7}$$

Remark 1 (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 $\widehat{\mathcal{M}}_2 - \widehat{\mathcal{M}}_1 \widehat{\mathcal{M}}_1^\top$.

2.3 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

$$x = Ah + z.$$

Figure 2 depicts a graphical representation of the ICA model where the coordinates of h are independent.

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

Lemma 2 (Comon and Jutten 2010)**.** *Define*

$$M_4 := \mathbb{E}[x \otimes x \otimes x \otimes x] - T, \tag{8}$$

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

$$T_{i_1,i_2,i_3,i_4} := \mathbb{E}[x_{i_1} x_{i_2}]\mathbb{E}[x_{i_3} x_{i_4}] + \mathbb{E}[x_{i_1} x_{i_3}]\mathbb{E}[x_{i_2} x_{i_4}] + \mathbb{E}[x_{i_1} x_{i_4}]\mathbb{E}[x_{i_2} x_{i_3}], \quad i_1, i_2, i_3, i_4 \in [d]. \tag{9}$$

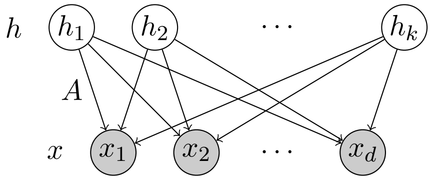

Figure 2: Graphical representation of ICA model $x = Ah$, where the coordinates of h are independent.

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

$$M_4 = \sum_{j \in [k]} \kappa_j a_j \otimes a_j \otimes a_j \otimes a_j. \tag{10}$$

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.

Sparse ICA

We also consider the sparse ICA model, which is the ICA with the additional constraint that the hidden vector h is sparse.

This is related to the dictionary learning or sparse coding model $x = Ah$ where the observations $x \in \mathbb{R}^d$ are sparse combination of dictionary atoms $a_j \in \mathbb{R}^d, j \in [k]$ through sparse vector $h \in \mathbb{R}^k$. If in addition, the coordinates of h are random and independent, the dictionary learning model is the same as the sparse ICA model. Others have studied the general sparse coding problem which are briefly mentioned in the related works section.

3 Tensor Concentration Bounds

In this section, we provide tensor concentration results for the proposed latent variable models. For each LVM, consider the higher-order observed moment (tensor) described in Section 2. The tensor concentration result bounds the spectral norm of error between the true moment tensor and its empirical estimate given n samples.

3.1 Multiview linear mixtures model

For the multiview linear mixtures model, we provide the tensor concentration result for the 3rd order observed moment in (5).

Consider the multiview linear mixtures model described in Section 2.1 denoted as model $\mathcal{S}$. Let $x_1^i, x_2^i, x_3^i, i \in [n]$, denote n samples of views x_1, x_2, x_3, respectively. Since the main focus is on recovering the components, we bound the spectral norm of difference between the empirical tensor estimate

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

and

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

where the expectation is conditioned on the choice of hidden states for n samples, and taken over the randomness of noise. Here, $h_i \in [k]$ denotes the hidden state for sample $i \in [n]$. Notice that tensor $\tilde{T}$ has the same form as true tensor T in (5) where

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

Here $\tilde{w}_j, j \in [k]$ are the empirical frequencies of different hidden states $h \in [k]$. It is easy to see that if $n \geq \Omega\left(\frac{\log k}{w_{\min}}\right)$, then all the empirical frequencies $\tilde{w}_j$ are within $[w_j/2, 2w_j]$. Therefore, tensor decomposition of $\tilde{T}$ has the same eigenvectors and similar eigenvalues as the true expectation (over both the noise and the hidden variables), and hence, it suffices to bound $\|\hat{T} - \tilde{T}\|$ provided as follows.

Theorem 1 (Tensor concentration bound for multiview linear mixtures model)**.** *Consider n samples $\{(x_1^i, x_2^i, x_3^i), i \in [n]\}$ from the multiview linear mixtures 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 defined in* (12) *satisfy the RIP condition in (RIP) (see Remark 3 for details on RIP condition). For $\hat{T}$ and $\tilde{T}$ as above, if $n = \operatorname{poly}(d)$, we have with high probability (over the choice of hidden state h and the noise)*

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

See the proof in Appendix C.1. The main ideas are described later in this section.

The above bound holds for any level of noise, but in each specific regime of noise, one of the terms is dominant and the bound is simplified. We now provide the bound for the high noise $\zeta^2 = \Theta(1)$ and low noise $\zeta^2 = \Theta\left(1/d\right)$ regimes which were introduced in Section 2.1. In the high noise regime $\zeta^2 = \Theta(1)$, the term $\zeta^3\sqrt{\frac{d}{n}}$ in Theorem 1 is dominant, and in the low noise regime $\zeta^2 = \Theta\left(1/d\right)$, the term $\zeta\sqrt{w_{\max}\frac{d}{n}}$ in Theorem 1 is dominant. This concentration bound is later used in Section 5 to provide sample complexity guarantees for learning multiview linear mixtures model.

Remark 2 (Application of Theorem 1 to whitening-based approaches). In the undercomplete setting, a guaranteed approach for tensor decomposition is to first orthogonalize the tensor through the *whitening* step, and then perform the orthogonal tensor eigen-decomposition through the power method (Anandkumar et al., 2014a). The whitening step leads to dependency to the condition number in the sample complexity result. Applying the proposed tensor concentration bound in Theorem 1 to this approach, we get similar dependency to the condition number, but better dependency in the dimension d. This improvement comes at the cost of additional bounded $2 \to 3$ norm condition on the factor matrices.

Concretely, following the analysis in Anandkumar et al. (2014a); Song et al. (2013), we have the error in recovery (up to permutation) as

$$\|\hat{a}_i - a_i\| \leq \frac{32\sqrt{2}\epsilon_{\text{triples}}}{\sigma_{\min}^3 w_{\min}^{1.5}} + \frac{512\epsilon_{\text{pairs}}^3}{\sigma_{\min}^3 w_{\min}^{1.5}}, \tag{11}$$

where $\epsilon_{\text{triples}} := \|\hat{T} - \tilde{T}\|$ is the error in estimating the third order moment, ϵ_{pairs} is the error in estimating the second order moments and $\sigma_{\min}$ is the k^{th} singular value of the factor matrices. While the ϵ_{pairs} can be obtained by matrix Bernstein's bounds as before (e.g. see Anandkumar et al. (2012)), we have an improved bound for $\epsilon_{\text{triples}}$ from Theorem 1, compared to previous results. Note that the first term corresponding to $\epsilon_{\text{triples}}$ is the dominant one and we improve its scaling.

Remark 3 (RIP property). Given n samples for the model $\mathcal{S}$ proposed in Section 2.1, define noise matrix

$$E_A := [\varepsilon_A^1, \varepsilon_A^2, \dots, \varepsilon_A^n] \in \mathbb{R}^{d \times n}, \tag{12}$$

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 when $n = \text{poly}(d)$, 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.

Proof ideas: The basic idea for proving the concentration result in Theorem 1 is an ε-net argument. We construct an ε-net and then show that with high probability the norm of error tensor is bounded for every vector in the ε-net.

In some cases even a usual ε-net of size $e^{O(d)}$ is good enough. But, in many other cases the usual ε-net construction does not provide a useful result since the failure probability is not small enough, and the union bound argument over all vectors in the ε-net fails (or incurs additional polynomial factors in the sample complexity result). In particular, for a vector with high correlation with the data, we get a worse concentration bound. But, the key observation is that there can not be too many vectors that have high correlation with the data. Therefore, for each fixed vector in the ε-net, we partition the terms in the error into two sets; one set corresponds to the small terms (where the vector is not highly correlated with the data) and the other set corresponds to the large terms. For the small terms, the usual ε-net argument still works. For the large terms, we show that the number of such terms is limited. This is done either by RIP property of the noise matrices or by the bounded $2 \to 3$ norm of factor matrices $A^\top$, $B^\top$ and $C^\top$. See the proofs of Claims 1-3 for more details. This partitioning argument is inspired by the entropy-concentration trade-off proposed in (Rudelson and Vershynin, 2009); however, here we have a finer partitioning into several sets, while in (Rudelson and Vershynin, 2009) the partitioning is done into only two sets.

Spherical Gaussian mixtures: Similar tensor concentration bound as above holds for the spherical Gaussian mixtures model with exploiting symmetrization trick as follows. In the spherical Gaussian mixtures model, the modified higher order moment (tensor) in (6) is symmetric, and hence noise matrices E_A, E_B and E_C are all the same. This can cause a problem because some square terms in the error tensor are not zero mean and we need to show their concentration around the mean. The well-known *symmetrization technique* can be exploited here where we draw two independent set of samples, and show the difference between the two is with high probability small. This technique is widely applied to show concentration around the median, and in all our cases the median is very close to the mean.

3.2 ICA and sparse ICA

For the ICA model, we provide the tensor concentration result for the modified 4th order observed moment (tensor) in (8) in both dense and sparse cases.

Theorem 2 (Tensor concentration bound for ICA)**.** *Consider n samples $x^i = Ah^i, i \in [n]$ from the ICA model with mixing matrix $A \in \mathbb{R}^{d\times k}$. Suppose $\|A\| \leq O(1+\sqrt{k/d})$ and the entries of $h \in \mathbb{R}^k$ are independent subgaussian variables with $\mathbb{E}[h_j^2] = 1$ and constant nonzero 4th order cumulant. For the 4th order cumulant M_4 in (8) and its empirical estimate $\widehat{M_4}$, if $n \geq d$, we have with high probability*

$$\|\widehat{M}_4 - M_4\| \leq \tilde{O}\left(\frac{m^2}{n} + \sqrt{\frac{m^4}{d^3 n}}\right), \quad m := \max(d,k).$$

See the proof in Appendix C.2. We have an improved bound for the sparse ICA setting as follows.

Theorem 3 (Tensor concentration bound for sparse overcomplete ICA)**.** *In the ICA model $x = Ah$, suppose $h_j = s_j g_j$ where s_j's are i.i.d. Bernoulli random variables with $\Pr[s_j = 1] = s/k$, and g_j's are independent 1-subgaussian random variables. Consider n independent samples $x^i = Ah^i, i \in [n]$, where each h^i is distributed as h. Suppose A satisfies (RIP) property (see Remark 3 for details on RIP condition). For the 4th order cumulant M_4 in (8) and its empirical estimate $\widehat{M_4}$, if $n, k \geq d$, we have with high probability*

$$\|\widehat{M}_4 - M_4\| \leq \tilde{O}\left(\frac{s^2}{n} + \sqrt{\frac{s^4}{d^3 n}}\right).$$

See the proof in Appendix C.3.

Dependence on k: It may seem counter-intuitive that the bound in Theorem 3 does not depend on k. The dependency on k is actually in the expectation where the expected tensor $\mathbb{E}\big[x^{\otimes 4}\big]$ in M_4 is close to $\frac{s}{k}\sum_{j\in[k]} a_j^{\otimes 4}$. We typically require the deviation to be less than the expected value.

Proof ideas: The proof ideas are similar to the multiview mixtures model where we provide ε-net arguments and partition the terms to small and large ones. In addition, for the ICA model, we exploit the subgaussian property of h_j's to provide concentration bound for the summation of subgaussian random variables raised to the 4th power (see Claim 4). This implies the concentration bound for the 4th order term $\mathbb{E}[x^{\otimes 4}]$ in M_4 (see Claim 5). For the 2nd order term T in M_4, the bound is argued using Matrix Bernstein's inequality (see Claim 6). For the sparse ICA model, the RIP property of A is exploited to bound the size of intersection between the support of (partitioned) vectors in the ε-net and the support of sparse vectors h^i (see Claim 7).

4 Learning Algorithm

In this section, we first introduce the tensor decomposition algorithm. Then, we provide some basic definitions and assumptions incorporated throughout the learning results. We conclude the section stating the organization of learning guarantees which are proposed in subsequent sections.

4.1 Tensor decomposition algorithm

We exploit the tensor decomposition algorithm in (Anandkumar et al., 2014b) to learn the parameters of the latent variable models. This is given in Algorithm 1. The main step in (14) basically performs alternating *asymmetric power updates*[4] on the different tensor modes. 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. For vectors $v, w \in \mathbb{R}^d$, recall the definition of multilinear form $T(I, v, w) \in \mathbb{R}^d$ in (2) where $T(I, v, w)$ is a multilinear combination of the tensor mode-1 fibers.

Intuition about the performance of tensor power update under non-orthogonal components is provided in (Anandkumar et al., 2014b), which is reminded here. For a rank-k tensor T as in (4), suppose we start at the correct vectors $\widehat{a} = a_j$ and $\widehat{b} = b_j$, for some $j \in [k]$. Then the numerator of tensor power update in (14) is expanded as

$$T\left(\widehat{a}, \widehat{b}, I\right) = T\left(a_j, b_j, I\right) = w_j c_j + \sum_{i \neq j} w_i \langle a_j, a_i \rangle \langle b_j, b_i \rangle c_i. \tag{13}$$

We observe that under orthogonal components the second term is zero, and thus the true vectors a_j, b_j and c_j are stationary points for the power update procedure. However under incoherent (soft-orthogonal) components, the stationary points of the power update procedure are approximate estimates of the true components with small error.

The purpose of clustering step is to identify which initializations are successful in recovering the true components under unsupervised setting. For more detailed discussion on the algorithm, see Anandkumar et al. (2014b).

Notice that in this paper, the input tensor T is the higher order moment of the LVMs described in Section 2. More details are stated in the learning results provided in next sections.

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 mutiview mixtures model desribed in Section 2.1 where the goal is to decompose the empirical moment tensor $\hat{T}$ of the form

$$\widehat{T} := \frac{1}{n} \sum_{l \in [n]} x_1^{(l)} \otimes x_2^{(l)} \otimes x_3^{(l)}, \tag{16}$$

where $x_r^{(l)}$ is the l^{th} sample from view $r \in [3]$. Applying the power update (14) in Algorithm 1 to $\hat{T}$, we have

$$\tilde{c} := \widehat{T}(\widehat{a}, \widehat{b}, I) = \frac{1}{n} X_3 \big(X_1^\top \widehat{a} * X_2^\top \widehat{b} \big), \tag{17}$$

where $*$ corresponds to the *Hadamard* product. Here, $X_r := \begin{bmatrix} x_r^{(1)} & x_r^{(2)} & \cdots & x_r^{(n)} \end{bmatrix} \in \mathbb{R}^{d \times n}$. Thus, the update can be computed efficiently using simple matrix and vector operations. It is easy to see

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

```
Algorithm 1 Tensor decomposition via alternating power updates (Anandkumar et al., 2014b)
Input: Tensor $T \in \mathbb{R}^{d \times d \times d}$, number of initializations $L$, number of iterations $N$.
  for $\tau = 1$ to $L$ do
    Initialize unit vectors $\widehat{a}_\tau^{(0)} \in \mathbb{R}^d$, $\widehat{b}_\tau^{(0)} \in \mathbb{R}^d$, and $\widehat{c}_\tau^{(0)} \in \mathbb{R}^d$ as
      • Semi-supervised setting: label information is exploited. See equation (19).
      • Unsupervised setting: SVD-based technique in Procedure 3 when $k \leq \beta d$ (for arbitrary constant $\beta$).
    for $t = 0$ to $N - 1$ do
      Asymmetric power updates (see (2) for the definition of the multilinear form):
        $\widehat{a}_\tau^{(t+1)} = \frac{T\left(I, \widehat{b}_\tau^{(t)}, \widehat{c}_\tau^{(t)}\right)}{\left\|T\left(I, \widehat{b}_\tau^{(t)}, \widehat{c}_\tau^{(t)}\right)\right\|}, \quad \widehat{b}_\tau^{(t+1)} = \frac{T\left(\widehat{a}_\tau^{(t)}, I, \widehat{c}_\tau^{(t)}\right)}{\left\|T\left(\widehat{a}_\tau^{(t)}, I, \widehat{c}_\tau^{(t)}\right)\right\|}, \quad \widehat{c}_\tau^{(t+1)} = \frac{T\left(\widehat{a}_\tau^{(t)}, \widehat{b}_\tau^{(t)}, I\right)}{\left\|T\left(\widehat{a}_\tau^{(t)}, \widehat{b}_\tau^{(t)}, I\right)\right\|}. \quad (14)$
    end for
    weight estimation:
        $\widehat{w}_\tau = T\left(\widehat{a}_\tau^{(N)}, \widehat{b}_\tau^{(N)}, \widehat{c}_\tau^{(N)}\right). \quad (15)$
  end for
  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.
  return the center member of these $k$ clusters as estimates $(\widehat{w}_j, \widehat{a}_j, \widehat{b}_j, \widehat{c}_j), j \in [k]$.
```

that the above update in (17) is easily parallelizable, and especially, the different initializations can be parallelized, making the algorithm scalable for large problems.

Basic definitions and assumptions

The error bounds in the subsequent results are provided in terms of distance between the estimated and the true vectors.

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

$$\operatorname{dist}(u, v) := \sup_{z \perp u} \frac{\langle z, v \rangle}{\|z\| \cdot \|v\|} = \sup_{z \perp v} \frac{\langle z, u \rangle}{\|z\| \cdot \|u\|}. \tag{18}$$

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

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

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

Procedure 2 Clustering process (Anandkumar et al., 2014b)

```
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 $\epsilon$.
  for $i = 1$ to $k$ do
    Among the remaining 4-tuples, choose $\widehat{a}, \widehat{b}, \widehat{c}$ which correspond to the largest $|T(\widehat{a}, \widehat{b}, \widehat{c})|$.
    Do $N$ more iterations of alternating updates in (14) starting from $\widehat{a}, \widehat{b}, \widehat{c}$.
    Let the output of iterations denoted by $(\widehat{a}, \widehat{b}, \widehat{c})$ be the center of cluster $i$.
    Remove all the tuples with $\max\{|\langle \widehat{a}_\tau, \widehat{a}\rangle|, |\langle \widehat{b}_\tau, \widehat{b}\rangle|, |\langle \widehat{c}_\tau, \widehat{c}\rangle|\} > \epsilon/2$.
  end for
  return the $k$ cluster centers.
```

Procedure 3 SVD-based initialization when $k = O(d)$ (Anandkumar et al., 2014b)

```
Input: Tensor $T \in \mathbb{R}^{d\times d\times d}$.
  Draw a random standard Gaussian vector $\theta \sim \mathcal{N}(0, I_d)$.
  Compute $u_1$ and $v_1$ as the top left and right singular vectors of $T(I, I, \theta) \in \mathbb{R}^{d\times d}$.
  $\widehat{a}^{(0)} \leftarrow u_1$, $\widehat{b}^{(0)} \leftarrow v_1$.
  Initialize $\widehat{c}^{(0)}$ by update formula in (14).
  return $\left(\widehat{a}^{(0)}, \widehat{b}^{(0)}, \widehat{c}^{(0)}\right)$.
```

Here, we review some of the assumptions and settings assumed throughout the learning results provided in next sections. Consider tensor decomposition form in (4). 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 linear mixtures model. For simplicity and 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. Also, 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 6 for more details).

In this paper, we focus on learning in the challenging overcomplete regime where the number of components/mixtures is larger than observed dimension. Precisely, we assume $k \geq \Omega(d)$. Note that the results can be easily adapted to the highly undercomplete regime when $k \leq o(d)$.

Learning results organization

In Section 2, we described how learning different latent variable models can be formulated as a tensor decomposition problem by performing appropriate modifications on the observed moments. For those LVMs, the tensor concentration bounds are provided in Section 3. We then proposed the tensor decomposition algorithm in Section 4.1 which is robust to noise. Employing all these techniques and results, we finally provide learning results for different latent variable models including multiview linear mixtures, ICA and sparse ICA in the subsequent sections. We consider two settings, viz., semi-supervised setting, where a small amount of label information is available, and 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 only works when $k \leq \beta d$ for any constant β. See the following two sections for learning guarantees.

5 Learning Multiview Linear Mixtures Model

In this section, we provide the semi-supervised and unsupervised learning results for the multiview linear mixtures model described in Section 2.1.

5.1 Semi-supervised learning

In the semi-supervised setting, label information is exploited to build good initialization vectors for tensor decomposition Algorithm 1 as follows. For the multiview linear mixtures model in Figure 1, let

$$x_{1,j}^{(l)}, x_{2,j}^{(l)}, x_{3,j}^{(l)} \in \mathbb{R}^d, \quad j \in [k], l \in [m_j],$$

denote $m = \sum_{j \in [k]} m_j$ samples of vectors corresponding to different labels, where the samples with subscript j have label j. Then, for any $j \in [k]$, we have the empirical estimate of mixture components as

$$\widehat{a}_j := \frac{1}{m_j} \sum_{l \in [m_j]} x_{1,j}^{(l)}, \quad \widehat{b}_j := \frac{1}{m_j} \sum_{l \in [m_j]} x_{2,j}^{(l)}, \quad \widehat{c}_j := \frac{1}{m_j} \sum_{l \in [m_j]} x_{3,j}^{(l)}. \tag{19}$$

Given n unlabeled samples, let

$$\epsilon_R := \begin{cases} \tilde{O}\left(k\sqrt{d}/\sqrt{n}\right) + \tilde{O}\left(\sqrt{k}/d\right), & \zeta^2 = \Theta(1), \\ \tilde{O}\left(\sqrt{k/n}\right) + \tilde{O}\left(\sqrt{k}/d\right), & \zeta^2 = \Theta\left(\frac{1}{d}\right), \end{cases} \tag{20}$$

denote the recovery error. We first provide the settings of Algorithm 1 which include input tensor T, number of iterations N and the initialization setting.

Settings of Algorithm 1 in Theorem 4:

- 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 (5) as the input to Algorithm 1.
- Number of iterations: $N = \Theta\left(\log\left(1/\epsilon_R\right)\right)$.
- Initialization: Exploit the empirical estimates in (19) as initialization vectors.

Conditions for Theorem 4:

- Rank condition: $\Omega(d) \leq k \leq o(d^{3/2})$.
- The columns of factor matrices are uniformly i.i.d. drawn from unit d-dimensional sphere $\mathcal{S}^{d-1}$ (see Remark 6 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 [5]

$$m_j \geq \tilde{\Omega}\left(\zeta^2 d\right), \; j \in [k]. \tag{21}$$

[5] 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 $\zeta^2 d \geq \omega(1)$.

- Given n unlabeled 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 3 for details on RIP condition). The number of samples n satisfies

$$n \geq \begin{cases} \tilde{\Omega}\left(k^2 d\right), & \zeta^2 = \Theta(1), \\ \tilde{\Omega}\left(k\right), & \zeta^2 = \Theta\left(\frac{1}{d}\right), \end{cases} \tag{22}$$

where ζ^2 is the variance of each entry of observation vectors.

Theorem 4 (Semi-supervised learning of multiview linear mixtures model)**.** *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.*

$$\operatorname{dist}\left(\hat{a}_j, a_j\right) \leq \epsilon_R, \quad j \in [k],$$

where $\operatorname{dist}(\cdot,\cdot)$ *function and ϵ_R are defined in* (18) *and* (20)*, respectively. Similar error bounds hold for other factor matrices B and C. In addition, the weight estimates $\widehat{w}_j, j \in [k]$ satisfy w.h.p.*

$$|\widehat{w}_j - w_j| \leq \epsilon_R / k, \quad j \in [k].$$

See Appendix B for the proof.

Approximation error in recovery: The recovery error ϵ_R involves two terms. One arises due to empirical estimation of 3rd order moment (given by $\tilde{O}\big(k\sqrt{d}/\sqrt{n}\big)$ or $\tilde{O}\big(\sqrt{k/n}\big)$) and is inevitable. The other term is due to non-orthogonality of columns of factor matrices (given by $\tilde{O}\big(\sqrt{k}/d\big)$) which is an approximation error in recovery of the tensor components. Note that the latter goes to zero for large enough d since we have $k \leq o(d^{3/2})$.

Remark 4 (Minimax sample complexity). Note that the number of labeled samples required is much smaller than the number of unlabeled samples, i.e., $\sum_{j \in [k]} m_j \ll n$. Thus, we provide efficient learning guarantees for overcomplete multiview Gaussian mixtures in the semi-supervised setting under a small number of labeled samples. Furthermore, in the low noise regime $\zeta^2 = \Theta\left(\frac{1}{d}\right)$, the sample complexity bounds for unlabeled samples is $\tilde{\Omega}(k)$, which is the *minimax* bound up to polylog factors.

Remark 5 (Different noise regime). For brevity, both semi-supervised and unsupervised learning results for multiview linear mixtures model in this section are provided in low noise $\zeta^2 = \Theta(1/d)$ and high noise $\zeta^2 = \Theta(1)$ regimes. But, notice that the result for general regime of noise (all different magnitudes of ζ) can be provided according to the general tensor concentration bound proposed in Theorem 1.

Remark 6 (Random assumption on factor matrices). 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 is a reasonable assumption for continuous models including the multiview linear 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. Moreover, the random assumption is provided for simplicity, while the original conditions for the guarantees of Algorithm 1 are deterministic; see Anandkumar et al. (2014b). They also show that random matrices satisfy these deterministic assumptions with high probability.

Remark 7 (Bounded 2 → 3 norm assumption). Notice that the bounded 2 → 3 norm assumption in tensor concentration bound in Theorem 1 is a weaker condition than assuming incoherence property for learning result in Theorem 4 which is needed for the algorithm guarantees. Furthermore, it is discussed in Anandkumar et al. (2014b) that under the assumptions $k \leq o(d^{3/2})$ and uniform draws of columns of A, B and C from unit sphere, the bound on 2 → 3 norm is satisfied.

Remark 8 (Spherical Gaussian mixtures). Similar learning results as in Theorem 4 hold for the spherical Gaussian mixtures. It is discussed in Section 2.2 how learning this model can be reduced to the tensor decomposition problem. Here, the 3rd order empirical (modified) moment $\widehat{M}_3$ in (7) is considered as the input of Algorithm 1 with symmetric updates. Thus, we show minimax unlabeled sample complexity for semi-supervised learning of overcomplete spherical Gaussian mixtures.

5.2 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 proposed in Procedure 3. 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 5:

- 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 (5) as the input to Algorithm 1.
- Number of iterations: $N = \Theta\left(\log\left(1/\epsilon_R\right)\right)$.
- The initialization in each run of Algorithm 1 is performed by SVD-based technique proposed in Procedure 3, with the number of initializations as

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

Conditions for Theorem 5:

- Rank condition: $k = \Theta(d)$.
- The columns of factor matrices are uniformly i.i.d. drawn from unit d-dimensional sphere $\mathcal{S}^{d-1}$.
- The number of samples n satisfies

$$n \geq \begin{cases} \tilde{\Omega}\left(k^4\right), & \zeta^2 = \Theta(1), \\ \tilde{\Omega}\left(k^2\right), & \zeta^2 = \Theta\left(\frac{1}{d}\right). \end{cases}$$

Theorem 5 (Unsupervised learning of multiview linear mixtures model)**.** *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 (up to permutation) satisfying w.h.p.*

$$\operatorname{dist}\left(\hat{a}_j, a_j\right) \leq \epsilon_R, \quad j \in [k],$$

where $\operatorname{dist}(\cdot,\cdot)$ *function and ϵ_R are defined in* (18) *and* (20)*, respectively. Similar error bounds hold for other factor matrices B and C. In addition, the weight estimates $\widehat{w}_j, j \in [k]$ satisfy w.h.p.*

$$|\widehat{w}_j - w_j| \leq \epsilon_R/k, \quad j \in [k].$$

See Appendix B for the proof.

Remark 9 (Comparison with "whitening + moment-based" techniques in the undercomplete setting when $k \approx d$)*.* Here, we discuss how our approach makes a huge improvement on sample complexity for learning multiview linear mixtures model and spherical Gaussian mixtures with the additional *incoherence* property we assume.

Multiview linear mixtures model: We compare 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 analysis shows lower sample complexity under incoherent factors.

Spherical Gaussian mixtures: As mentioned in Remark 8, 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 also 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 $n \geq \tilde{\Omega}(k^3)$. But, our tight tensor concentration analysis leads to the better sample complexity of $n \geq \tilde{\Omega}(k^2)$. Note that this comparison is in the low noise regime $\zeta^2 = \Theta\left(\frac{1}{d}\right)$.

6 Learning Independent Component Analysis (ICA)

In this section, we propose the semi-supervised and unsupervised learning results for the ICA model described in Section 2.3. Unlike multi-view models, the standard ICA model does not have noise. This is because in ICA every sample is a mixture of many components (compared to multi-view models), and therefore, the noise is already "built in" the model.

6.1 Semi-supervised learning

By semi-supervised setting in ICA, we mean some prior information is available which provides good initializations for the components.

Given n samples of observations $x^i = Ah^i, i \in [n]$, let

$$\tilde{\epsilon}_R := \tilde{O}\left(k^2 / \min\left\{n, \sqrt{d^3 n}\right\}\right) + \tilde{O}\left(\frac{\sqrt{k}}{d^{3/2}}\right) \tag{23}$$

denote the recovery error.

Settings of Algorithm 1 in Theorem 6:

- Given n samples $x^i = Ah^i, i \in [n]$, consider the empirical estimate of 4th order (modified) moment M_4 (see (8)) as the input to Algorithm 1 with symmetric 4th order updates. See Anandkumar et al. (2014b) for higher order extension of the algorithm.
- Number of iterations: $N = \Theta\left(\log\left(1/\tilde{\epsilon}_R\right)\right)$.
- Initialization: it is assumed that for any $j \in [k]$, an approximation of a_j denoted by $\widehat{a}_j^{(0)}$ is given satisfying
$$\min_{z \in \{-1,1\}} \|z\widehat{a}_j^{(0)} - a_j\| \leq \frac{w_{\max}}{w_{\min}}.$$

Note that the initialization up to sign recovery is only required.

Conditions for Theorem 6:

- Rank condition: $\Omega(d) \leq k \leq o(d^2)$.
- The columns of factor matrices are uniformly i.i.d. drawn from unit d-dimensional sphere $\mathcal{S}^{d-1}$.
- The entries of h are independent subgaussian variables with $\mathbb{E}[h_j^2] = 1$ and constant nonzero 4th order cumulant.
- The number of samples n satisfies

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

Theorem 6 (Semi-supervised learning of ICA). *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 mixing matrix A satisfying w.h.p.*

$$\operatorname{dist}\left(\hat{a}_j, a_j\right) \leq \tilde{\epsilon}_R, \quad j \in [k],$$

where $\operatorname{dist}(\cdot,\cdot)$ *function and* $\tilde{\epsilon}_R$ *are defined in* (18) *and* (23)*, respectively. In addition, the weight estimates* $\widehat{w}_j, j \in [k]$ *satisfy w.h.p.*

$$|\widehat{w}_j - w_j| \leq \tilde{\epsilon}_R, \quad j \in [k].$$

See Appendix B for the proof.

Approximation error for recovery: Notice that the approximation error recovery for the ICA model is $\tilde{O}\big(\sqrt{k}/d^{3/2}\big)$, while for the multiview mixtures model the approximation is $\tilde{O}\big(\sqrt{k}/d\big)$. The difference is because of different tensor orders for the two models.

Weight recovery comparison with multiview mixtures model: Comparing the weight recovery error for ICA model in Theorems 6 and 7 with the multiview linear mixtures model in Theorems 4 and 5, we observe that the factor $1/k$ does not exist in the ICA model. This is because of different assumptions on the weights in the tensor form. For the multiview mixtures model, it is assumed $w_j = \Theta(1/k), j \in [k]$. But, in the ICA model, the weights are the 4th order cumulants κ_j (see (10)) which are assumed to be constant.

Remark 10 (Efficient sample complexity). 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. In the unsupervised setting, previous results require large polynomial sample complexity, e.g., Goyal et al. (2013) need $\operatorname{poly}(k^4)$ number of unlabeled samples for learning ICA, where the poly factor is not explicitly characterized.

6.2 Unsupervised learning

In the unsupervised setting, the initialization is performed by doing rank-1 SVD on random slices of the moment tensor proposed in Procedure 3. 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 7:

- Given n samples $x^i = Ah^i, i \in [n]$, consider the empirical estimate of 4th order (modified) moment M_4 (see (8)) as the input to Algorithm 1 with symmetric 4th order updates. See Anandkumar et al. (2014b) for higher order extension of the algorithm.
- Number of iterations: $N = \Theta\left(\log\left(1/\tilde{\epsilon}_R\right)\right)$.
- The initialization is performed by 4-th order generalization [6] of SVD-based technique in Procedure 3, with the number of initializations as
$$L \geq k^{\Omega(k^2/d^2)}.$$

Conditions for Theorem 7:

- Rank condition: $k = \Theta(d)$.
- The columns of factor matrices are uniformly i.i.d. drawn from unit d-dimensional sphere $\mathcal{S}^{d-1}$.
- The entries of h are independent subgaussian variables with $\mathbb{E}[h_j^2] = 1$ and constant nonzero 4th order cumulant.
- The number of samples n satisfies
$$n \geq \tilde{\Omega}\left(k^3\right).$$

Theorem 7 (Unsupervised learning of ICA)**.** *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 mixing matrix A (up to permutations) satisfying w.h.p.*

$$\operatorname{dist}\left(\hat{a}_j, a_j\right) \leq \tilde{\epsilon}_R, \quad j \in [k],$$

where $\operatorname{dist}(\cdot,\cdot)$ *function and* $\tilde{\epsilon}_R$ *are defined in* (18) *and* (23)*, respectively. In addition, the weight estimates $\widehat{w}_j, j \in [k]$ satisfy w.h.p.*

$$|\widehat{w}_j - w_j| \leq \tilde{\epsilon}_R, \quad j \in [k].$$

See Appendix B for the proof.

6.3 Sparse ICA

For the sparse ICA model introduced in Section 2.3, suppose the entries of h are i.i.d. Bernoulli-subgaussian random entries, where the probability of each Bernoulli variable being 1 is s/k, and therefore, s is the expected number of nonzero entries in h. More precisely, suppose $h_j = s_j g_j$ where s_j's are i.i.d. Bernoulli random variables with $\Pr[s_j = 1] = s/k$, and g_j's are independent 1-subgaussian random variables.

In the following theorem, we provide both semi-supervised and unsupervised learning of sparse ICA model. Note that the error recovery is changed as

$$\tilde{\epsilon}_R := \tilde{O}\left(sk/\min\left\{n, \sqrt{d^3 n}\right\}\right) + \tilde{O}\left(\frac{\sqrt{k}}{d^{3/2}}\right).$$

[6] 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.

Theorem 8 (Semi-supervised and unsupervised learning of sparse ICA)**.** *Similar semi-supervised and unsupervised learning guarantees as in Theorems 6 and 7 hold for the sparse ICA model with the following sample complexity requirements. 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 addition, here we assume that mixing matrix A satisfies the RIP property in (RIP) (see Remark 3 for details on RIP condition).

See Appendix B for the proof.

Remark 11 (Comparison with multiview mixtures and ICA)*.* In terms of sparsity of latent vector h, the sparse ICA spans between multiview Gaussian mixtures (where h has one nonzero entry in vector representation), 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.

Comparison with previous approaches: The dictionary learning problem is also studied in Arora et al. (2013); Agarwal et al. (2013); Barak et al. (2014). 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) use the sum of squares framework and can handle the sparsity level up to (small enough) constant times k, but with the expense of computational complexity which scales as $k^{O(\log k)}$, and the size of the dictionary $k = O(d)$. In addition, when the sparsity level is smaller as $k^{1-\delta}$ for some $0 < \delta < 1$, their algorithm runs in polynomial time $k^{O(1/\delta)}$. They can also go to higher level of overcompleteness with the expense of reducing sparsity level. They do not need the assumptions that the dictionary is incoherent or that the coefficients are independent. They only have approximate recovery and note that exact recovery is impossible (from an identifiability standpoint) unless further assumptions are imposed. In contrast, we have a polynomial time method for incoherent dictionaries and independent coefficients which can handle arbitrary sparsity level, and provides approximate recovery. Moreover, we can handle larger dictionary sizes k at the expense of more computation.

Below, we show how we can extend our analysis to dependent sparsity setting, but with worse performance guarantees.

Extension to dependent sparsity

In this section, we consider the noiseless sparse coding model $x = Ah$, but with no independence assumption on the latent entries h_i's. The analysis can be extended to noisy case.

We assume the following moment conditions on h in the dependent sparsity model. Note that these assumptions are comparable with the moment assumptions in Barak et al. (2014).

$$\begin{aligned} \mathbb{E}\left[h_i^4\right] = \mathbb{E}\left[h_i^2\right] &= \beta s/k, \\ \mathbb{E}\left[h_i^2 h_j^2\right] &\leq \tau, \quad i \neq j, \\ \mathbb{E}\left[h_i^3 h_j\right] &= 0, \quad i \neq j, \end{aligned}$$

with parameters s and τ, where s is the expected number of nonzero entries in h, and β is a universal constant. The first condition represents the normalization factor which depends on the sparsity level. The second condition limits the sparsity level and the amount of correlation between different entries of vector h. To provide more intuition about these parameters, assume that the entries of h are distributed as Bernoulli-Gaussian random variables with each entry being nonzero with probability s/k. Then, we have $\tau = \rho p + (1-\rho)p^2$, where ρ is the correlation coefficient between h_i^2 and h_j^2 for $i \neq j$.

Theorem 9 (Noiseless sparse coding with dependent sparsity)**.** *Consider the described dictionary learning model $x = Ah$ where the moments of random vector h satisfy the conditions stated before the theorem. Let the noiseless 4th order observed moment $\mathbb{E}\left[x^{\otimes 4}\right]$ be the input to Algorithm 1 with symmetric 4th order updates. Let the initialization in each run of Algorithm 1 is performed by 4th order generalization of the SVD-based technique proposed in Procedure 3. Let $\tilde{\epsilon}_R := \tilde{O}(\tau k/s) + \tilde{O}\left(\sqrt{k}/d^{3/2}\right)$, and suppose*

$$k = \Theta(d), \quad N = \Theta\left(\log\left(1/\tilde{\epsilon}_R\right)\right), \quad L \geq k^{\Omega\left(k^2/d^2\right)}.$$

In addition, assume that the columns of dictionary A are uniformly i.i.d. drawn from unit d-dimensional sphere $\mathcal{S}^{d-1}$. If

$$\tau \leq \tilde{O}\left(\frac{s/k}{d}\right),$$

then whp

$$\operatorname{dist}\left(\hat{a}_j, a_j\right) \leq \tilde{\epsilon}_R, \quad j \in [k].$$

See Appendix B for the proof.

Comparing with the dictionary learning result by Barak et al. (2014), their algorithm is based on sum-of-squares techniques, and do not require any incoherence assumptions on the dictionary atoms. They can also handle higher levels of sparsity and correlation. On the other hand, they have a quasi-polynomial algorithm in the regime of high sparsity (small enough constant times k), while our algorithm is very simple and efficient.

The above analysis is in the noiseless regime, and the generalization to noisy case can be investigated as a future work which involves the sample complexity analysis in the dependent sparsity case.

7 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[7] $w_j = \Pr[h = j] = \frac{1}{k}, j \in [k]$. 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 $n = 1000$.

[7] 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.

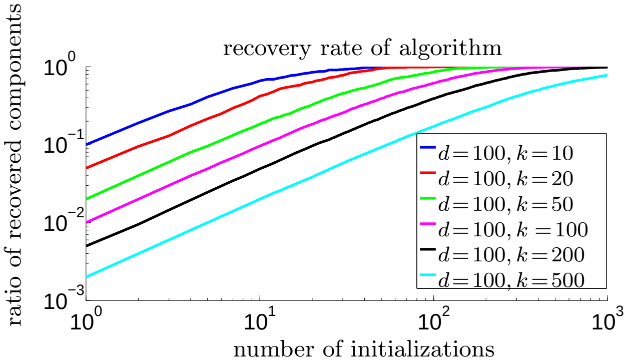

Figure 3: Ratio of recovered components vs. the number of initializations. The figure is an average over 10 random runs.

Notice that the empirical tensor $\widehat{T}$ in (16) is not explicitly computed, and the tensor power updates in the algorithm are computed through the multilinear form stated in (17). This leads to efficient computational complexity. See Section 4.1 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 4, we let

$$t_{\mathrm{S}} := t_1 (\log d)^2 \sqrt{\frac{k}{n}} + t_2 (\log d)^2 \frac{\sqrt{k}}{d}, \tag{24}$$

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 (14). 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

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

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

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

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 (20) which claims that the square recovery error is bounded as $\tilde{O}(k)$ when d and n are fixed as here.

Acknowledgements

A. Anandkumar is supported in part by Microsoft Faculty Fellowship, NSF Career award CCF-1254106, NSF award CCF-1219234, ONR award N00014-14-1-0665, ARO YIP award W911NF-13-1-0084, and AFOSR YIP award FA9550-15-1-0221. M. Janzamin is supported by NSF Award CCF-1219234, ARO Award W911NF-12-1-0404 and ARO YIP Award W911NF-13-1-0084.

Appendix

More Matrix and Tensor Notations

The outer product operator $\otimes$ defined earlier for vectors can be also generalized to higher order tensors. For instance, given matrices $A, B \in \mathbb{R}^{d \times d}$, the 4th order tensor $T \in \mathbb{R}^{d \times d \times d \times d}$ is defined as

$$T := A \otimes B \Leftrightarrow T_{i_1,i_2,i_3,i_4} = A_{i_1,i_2} B_{i_3,i_4}.$$

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,

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

A Recap of Guarantees for Algorithm 1

In this section, we recap the local and global convergence guarantees of Algorithm 1 provided in Anandkumar et al. (2014b). These results are required for proving unsupervised and semi-supervised learning results provided in this paper.

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

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

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

A.1 Local convergence guarantee

The local convergence result is provided in the following theorem which bounds the estimation error after N 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.

Settings of Algorithm 1 in Theorem 10:

- Number of iterations: $N = \Theta\left(\log\left(\frac{1}{\gamma \hat{\epsilon}_R}\right)\right)$, where $\gamma := \frac{w_{\max}}{w_{\min}}$.

Conditions for Theorem 10:

- 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 $\mathcal{S}^{d-1}$.
- Rank condition: $k = o\left(d^{1.5}\right)$.
- Perturbation tensor Ψ satisfies the bound

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

- 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

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

  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 (14).

Theorem 10 (Local convergence guarantee of Algorithm 1 (Anandkumar et al., 2014b))**.** *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 w.h.p. after N iterations as*

$$\max\left\{\operatorname{dist}\left(\widehat{a}^{(N)}, a_j\right), \operatorname{dist}\left(\widehat{b}^{(N)}, b_j\right), \operatorname{dist}\left(\widehat{c}^{(N)}, c_j\right)\right\} \leq O(\hat{\epsilon}_R), \tag{27}$$

where $\hat{\epsilon}_R$ is defined in (25)*. Furthermore, the weight estimate $\widehat{w} = \widehat{T}\left(\widehat{a}^{(N)}, \widehat{b}^{(N)}, \widehat{c}^{(N)}\right)$ in* (15) *satisfies w.h.p.*

$$|\widehat{w} - w_j| \leq O(w_{\min}\hat{\epsilon}_R).$$

Note that the recovery error $\hat{\epsilon}_R$ 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. The above local convergence result can be also interpreted as an approximate local identifiability result for tensor decomposition under incoherent factors.

A.2 Global convergence guarantee when $k = O(d)$

Theorem 10 provides local convergence guarantee given good initialization vectors for different components. The global convergence guarantee is presented in the following theorem where the SVD-based initialization method in Procedure 3 is exploited to provide good initialization vectors when $k = O(d)$.

Settings of Algorithm 1 in Theorem 11:

- Number of iterations: $N = \Theta\left(\log\left(\frac{1}{\gamma\hat{\epsilon}_R}\right)\right)$, 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 11:

- Rank-k decomposition and perturbation conditions as[8]

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

  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: $k = O(d)$.

[8] Note that the perturbation condition is stricter than the corresponding condition in the local convergence guarantee (Theorem 10).

Theorem 11 (Global convergence guarantee of Algorithm 1 when $k = O(d)$, (Anandkumar et al., 2014b))**.** *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.,*

$$\max\left\{\operatorname{dist}\left(\widehat{a}_j, a_j\right), \operatorname{dist}\left(\widehat{b}_j, b_j\right), \operatorname{dist}\left(\widehat{c}_j, c_j\right)\right\} \le O(\hat{\epsilon}_R),$$
$$|\widehat{w}_j - w_j| \le O(w_{\min}\hat{\epsilon}_R),$$

where $\hat{\epsilon}_R$ is defined in (25)*.*

The number of initialization trials L is polynomial when γ is a constant, and $k = O(d)$.

B Proof of Learning Theorems

The semi-supervised and unsupervised learning results for each latent variable model are proved by combining the corresponding tensor concentration bound proposed in Section 3 and the convergence guarantees of the tensor decomposition algorithm recapped in Appendix A.

Proof of Theorem 4: The result is proved by applying the tensor concentration bound in Theorem 1 to the local convergence result of Algorithm 1 recapped in Theorem 10. Note that in the high noise regime $\zeta^2 = \Theta(1)$, the term $\zeta^3\sqrt{\frac{d}{n}}$ in Theorem 1 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 Theorem 1 is dominant.

Note that the sub-Gaussian property of conditional observed distributions is used to provide the labeled sample complexity. Since the distribution of observed variables given hidden state is sub-Gaussian with covariance matrix $\zeta^2 I$ as in model $\mathcal{S}$ described in Section 2.1, we have the following concentration bound where with probability at least $1-\delta$, the empirical estimate $\widehat{a}_j^{(0)}$ satisfies

$$\left\|\widehat{a}_j^{(0)} - a_j\right\| \le C_1\sqrt{\frac{\zeta^2 d\log(1/\delta)}{m_j}}, \quad j \in [k],$$

for some constant $C_1 > 0$. □

Proof of Theorem 5: The result is proved by applying the tensor concentration bound in Theorem 1 to the global convergence result of Algorithm 1 recapped in Theorem 11. The dominant error bounds in Theorem 1 are the same as what stated in the proof of Theorem 4. □

Proof of Theorem 6: The result is proved by applying the tensor concentration bound in Theorem 2 to the local convergence result of Algorithm 1 in the 4th order case. See Anandkumar et al. (2014b) for the generalization of convergence result to higher order cases. □

Proof of Theorem 7: The result is proved by applying the tensor concentration bound in Theorem 2 to the global convergence result of Algorithm 1 recapped in Theorem 11. 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 $\psi := \|\Psi\| \le \frac{w_{\min}\sqrt{\log k}}{\alpha_0^2 d}$. □

Proof of Theorem 8: The learning results for the sparse ICA are proved similar to the ICA case, with the difference that the sparse ICA concentration bound in Theorem 3 is exploited here.
□

Proof of Theorem 9: Given linear model $x = Ah$, the 4th order observed moment is expanded as

$$\mathbb{E}\left[x^{\otimes 4}\right] = \mathbb{E}\left[h^{\otimes 4}\right]\left(A^\top, A^\top, A^\top, A^\top\right), \tag{28}$$

where the multilinear notation defined in (1) is exploited.

Expanding $\mathbb{E}\left[h^{\otimes 4}\right]$, and treating $\sum_{i\in[k]} \mathbb{E}[h_i^4]\ e_i^{\otimes 4}$ as the main signal, the remaining term is

$$R := \sum_{i\neq j} \mathbb{E}[h_i^2 h_j^2]\ e_i^{\otimes 2} \otimes e_j^{\otimes 2},$$

where we also exploited the assumption that the expectation of terms involving odd powers of h_i are zero. Then, from (28), the spectral norm of perturbation tensor is bounded as

$$\|\Psi\| := \left\|R\left(A^\top, A^\top, A^\top, A^\top\right)\right\| = \left\|\sum_{i\neq j} \mathbb{E}[h_i^2 h_j^2]\ a_i^{\otimes 2} \otimes a_j^{\otimes 2}\right\| \leq \tau \|A\|^4,$$

where we used $\mathbb{E}[h_i^2 h_j^2] \leq \tau$ in the last inequality. Imposing condition $\|\Psi\| \leq \tilde{O}\left(w_{\min}/d\right)$, and then applying Theorem 11, the result is proved. Note that $w_{\min} := \min_{i\in[k]} \mathbb{E}[h_i^4] = \beta s/k$. □

C Proof of Tensor Concentration Bounds

In this section, we provide the proof of tensor concentration bounds for different latent variable models including multiview linear mixtures model, ICA and sparse ICA. 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 concentration 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.

C.1 Multiview linear mixtures model

In this section, we prove the tensor concentration result for the multiview linear mixtures model provided in Theorem 1.

Proof of Theorem 1: Expanding the difference $\hat{T} - \tilde{T}$, we have

$$\begin{aligned}
\hat{T} - \tilde{T} = {} & \frac{1}{n}\zeta^3 d^{1.5} \sum_{i\in[n]} \varepsilon_A^i \otimes \varepsilon_B^i \otimes \varepsilon_C^i && (29a)\\
& + \frac{1}{n}\zeta^2 d \sum_{i\in[n]} \left(a_{h_i} \otimes \varepsilon_B^i \otimes \varepsilon_C^i + \varepsilon_A^i \otimes b_{h_i} \otimes \varepsilon_C^i + \varepsilon_A^i \otimes \varepsilon_B^i \otimes c_{h_i}\right) && (29b)\\
& + \frac{1}{n}\zeta\sqrt{d} \sum_{i\in[n]} \left(a_{h_i} \otimes b_{h_i} \otimes \varepsilon_C^i + a_{h_i} \otimes \varepsilon_B^i \otimes c_{h_i} + \varepsilon_A^i \otimes b_{h_i} \otimes c_{h_i}\right). && (29c)
\end{aligned}$$

There are three types of terms in the above difference which are bounded separately in Claims 1-3 in Section C.1.2. Combining the results of claims, the theorem follows directly.

□

C.1.1 Basic definitions and lemmata

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

Definition 2 (Small and large terms)**.** *Consider matrices $E_A := [\varepsilon_A^1, \varepsilon_A^2, \dots, \varepsilon_A^n] \in \mathbb{R}^{d\times n}$, 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 3. *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 2, 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

$$|L| \leq \frac{d}{25\log^2 d}. \tag{30}$$

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 (30) holds. □

The above partitioning into small and large sets is good when all we care about is the inner-products between a fixed vector and the noise vectors. However, when we are also interested in the inner-products between a fixed vector and columns of A, B, C, it is often not tight enough, and in order to get a tight bound, we propose the following finer partitioning.

Definition 3 (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 unit 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 $z^l \in \mathbb{R}^k, l \in \{0,1,2,\ldots,t\}$, 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 4. *Consider matrix $C := [c_1, c_2, \ldots, c_k] \in \mathbb{R}^{d\times k}$ 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 3. 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: For the first part, we know the number of non-zero entries in z^l is $|K_l|$. For any unit vector w, we have

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

which implies the desired bound on $|K_l|$.

Let $q_l := O\big(\frac{d^{3/2}}{p_l^3}\big)$ be the maximum number of nonzero entries in z^l. 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 5. *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 3. 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 3. The ε-net is constructed in the same way as in the previous lemma. □

C.1.2 Proof of claims

In this section, we separately bound different error terms (29a)-(29c). Among all the terms, the terms like (29c) 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 (29c) can be adapted to bound all the other terms. Here for clarity we start from the simplest term (29a), and point out new ideas in the proofs of (29b) and (29c).

Claim 1 (Bounding norm of (29a))**.** *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

$$T_1 = \frac{1}{n} \sum_{i=1}^{n} \eta_i \varepsilon_A^i \otimes \varepsilon_B^i \otimes \varepsilon_C^i, \tag{31}$$

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 (31) which is defined as

$$\|T_1\| := \sup_{\|u\|=\|v\|=\|w\|=1} |T_1(u,v,w)| = \sup_{\|u\|=\|v\|=\|w\|=1} \left| \frac{1}{n} \sum_{i=1}^{n} \eta_i \langle u, \varepsilon_A^i \rangle \langle v, \varepsilon_B^i \rangle \langle w, \varepsilon_C^i \rangle \right|. \tag{32}$$

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 2, 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

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

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 $E_A := [\varepsilon_A^1, \dots, \varepsilon_A^n]$, $E_B := [\varepsilon_B^1, \dots, \varepsilon_B^n]$ and $E_C := [\varepsilon_C^1, \dots, \varepsilon_C^n]$ 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 3, 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 3. 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 2 (Bounding norm of (29*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 2, 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 3, we have

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

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

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

where Cauchy-Schwartz inequality is exploited in the second inequality, and the bounds $\|E_A[L]\| \leq 2$ and $\|E_B[L]\| \leq 2$ 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-nets 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 3 where

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

Let Q_l denote the sum of all terms that fall into bucket K_l as

$$Q_l := \frac{1}{n} \sum_{i \in L^c, h_i \in K_l} \eta_i \langle u, \varepsilon_A^i \rangle \langle v, \varepsilon_B^i \rangle \langle w, c_{h_i} \rangle. \tag{33}$$

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 1.

For $Q_l, l \in [t]$, the argument is as follows. Let $p_l := 2^{l-1}$. Applying Lemma 4, 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 (33) 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 (33) 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 (notice below that $p_l \leq O(\sqrt{d})$)

$$Q_l \leq \tilde{O}\left(\frac{p_l}{\sqrt{d}n} + \sqrt{\frac{w_{\max}}{np_l\sqrt{d}}} \right) \leq \tilde{O}\left(\frac{1}{n} + \sqrt{\frac{w_{\max}}{n\sqrt{d}}} \right).$$

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 (29c).

Claim 3 (Bounding norm of (29c))**.** *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

$$T_3 = \frac{1}{n} \sum_{i=1}^{n} \eta_i \varepsilon_A^i \otimes b_{h_i} \otimes c_{h_i}, \tag{34}$$

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 $E_A := [\varepsilon_A^1, \varepsilon_A^2, \dots, \varepsilon_A^n]$ 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 (34) which is defined as

$$\|T_3\| := \sup_{\|u\|=\|v\|=\|w\|=1} |T_3(u,v,w)| = \sup_{\|u\|=\|v\|=\|w\|=1} \left| \frac{1}{n} \sum_{i=1}^{n} \langle u, \varepsilon_A^i \rangle \langle v, b_{h_i} \rangle \langle w, c_{h_i} \rangle \right|. \tag{35}$$

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 (35) into buckets as defined in Definition 3. 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 $K_0^a, K_1^a, \dots, K_{4 \log\log d}^a$ into K_0^a. This means K_0^a now contains all i's with inner product

$$|\langle \varepsilon_A^i, u \rangle| \leq \frac{16 \log d}{\sqrt{d}},$$

and K_l^a's for $1 \leq l \leq 4 \log\log d$ 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 (35) on this set, i.e.,

$$Q_{l_1,l_2,l_3} := \frac{1}{n} \sum_{i \in J_{l_1,l_2,l_3}} \langle u, \varepsilon_A^i \rangle \langle v, b_{h_i} \rangle \langle w, c_{h_i} \rangle. \tag{36}$$

Note that by construction of buckets, the summation in (35) 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 1.

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 4 and Lemma 5, the total number of terms in the summation form (36) for Q_{l_1,l_2,l_3} is w.h.p. bounded as

$$|J_{l_1,l_2,l_3}| \leq O\left(nw_{\max}\frac{d^{3/2}}{p_l^3}\right),$$

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.

□

C.2 ICA

In this section, we prove the tensor concentration result for the ICA model provided in Theorem 2.

Recall the 4th order modified moment tensor in equation (8) as

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

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 $\widehat{M_4}$ be the empirical estimate of M_4 given n samples.

Proof of Theorem 2: Let $W := \frac{1}{n}\sum_{i=1}^n x^i (x^i)^\top$, and therefore, the empirical estimate of T is given by

$$\widehat{T}_{i_1,i_2,i_3,i_4} = W_{i_1,i_2}W_{i_3,i_4} + W_{i_1,i_3}W_{i_2,i_4} + W_{i_1,i_4}W_{i_2,i_3}. \tag{37}$$

Then, the empirical estimate of M_4 is given by

$$\widehat{M_4} = \frac{1}{n}\sum_{i=1}^n (x^i)^{\otimes 4} - \widehat{T}.$$

The proof directly follows from Claims 5 and 6, which bound the perturbation of the two terms separately. Claim 5 bounds the 4th order term perturbation $\mathbb{E}[x^{\otimes 4}] - \frac{1}{n}\sum_{i=1}^n (x^i)^{\otimes 4}$, and Claim 6 bounds the 2nd order term perturbation $T - \widehat{T}$. □

C.2.1 Proof of claims

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 4. *Suppose $h_i, i \in [n]$, are independent q-subgaussian random variables. Then, for any $d \geq 1$, with probability at least $1 - e^{-\omega(d \log n)}$ we have*

$$\left| \frac{1}{n} \sum_{i=1}^n \left(h_i^4 - \mathbb{E}\left[h_i^4\right] \right) \right| \leq \tilde{O}\left(\frac{q^4 d^2}{n} + \sqrt{\frac{q^8 d}{n}} \right).$$

(Notice that here d is intended to be the dimension in later applications. However, for this claim we can choose d to be an arbitrary real number that is at least 1.)
Proof: We prove

$$\Pr\left[\frac{1}{n} \left| \sum_{i=1}^n h_i^4 - \mathrm{med}\Big(\sum_{i=1}^n h_i^4 \Big) \right| \leq \tilde{O}\left(\frac{q^4 d^2}{n} + \sqrt{\frac{q^8 d}{n}} \right) \right] \geq 1 - e^{-\omega(d \log n)}, \tag{38}$$

where $\mathrm{med}(\cdot)$ is the median of the distribution. By doing simple integration (for d from 1 to ∞), this concentration bound implies

$$\left| \mathbb{E}\left[\sum_{i=1}^n h_i^4 \right] - \mathrm{med}\Big(\sum_{i=1}^n h_i^4 \Big) \right| \leq \tilde{O}\left(\frac{q^4}{\sqrt{n}} \right).$$

Therefore, when $d \geq 1$ the difference between mean and median is negligible, and we get the desired bound in the claim.

In order to prove the deviation bound from the median in (38), we use the standard symmetrization argument: it is enough to take two independent sample sets $\{h_1, h_2, \dots, h_n\}$ and $\{\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_n\}$ with the same distribution, and bound $\left| \frac{1}{n} \sum_{i \in [n]} \big(h_i^4 - \tilde{h}_i^4 \big) \right|$. In order to bound the sum, we rewrite it in the form

$$Q = \frac{1}{n} \sum_{i \in [n]} \eta_i |h_i^4 - \tilde{h}_i^4|,$$

where η_i's are independent random ± 1 variables with $\Pr[\eta_i = 1] = 1/2$.

Now we partition the terms in the summation for Q into multiple buckets according to the magnitude of $\left| h_i^4 - \tilde{h}_i^4 \right|$. Let $t := \lceil \log_2 d^2 + C' \log_2 \log_2 n \rceil$ (where C' is a large enough constant). Then the buckets are defined as

$$\begin{aligned} K_0 &:= \left\{ i \in [n] : |h_i^4 - \tilde{h}_i^4| \leq q^4 \right\}, \\ K_l &:= \left\{ i \in [n] : |h_i^4 - \tilde{h}_i^4| \in \left(2^{l-1} q^4, 2^l q^4 \right] \right\}, \quad l \in [t], \\ K_{t+1} &:= \left\{ i \in [n] : |h_i^4 - \tilde{h}_i^4| > 2^t q^4 \right\}. \end{aligned}$$

Let Q_l denote the sum of all terms that fall into bucket K_l as

$$Q_l := \frac{1}{n} \sum_{i\in[n], i\in K_l} |h_i^4 - \tilde{h}_i^4|\eta_i. \tag{39}$$

Note that by construction of buckets, the original summation $Q = \sum_{l=0}^{t+1} Q_l$. There are only $O(\log d)$ terms in this summation, and therefore, it suffices to show each term Q_l is small.

Note that since h_i's and $\tilde{h}_i$'s are q-subgaussian random variables, we have

$$\begin{aligned}\Pr\left[|h_i^4 - \tilde{h}_i^4| \geq \lambda q^4\right] &\leq \Pr\left[h_i^4 \geq \lambda q^4/2\right] + \Pr\left[\tilde{h}_i^4 \geq \lambda q^4/2\right] \\ &= 2\Pr\left[|h_i| \geq (\lambda q^4/2)^{1/4}\right] \\ &\leq 4\exp\left(-\frac{\sqrt{\lambda}}{2\sqrt{2}}\right), \end{aligned} \tag{40}$$

where the last inequality uses q-subgaussian property.

For $Q_l, 0 \leq l \leq 2\log\log n$, we apply Bernstein's inequality directly. Each term in the summation for Q_l is bounded as $\tilde{O}(q^4/n)$, and the variance term is also bounded as $\tilde{O}(q^8/n)$. By applying Bernstein's inequality, with probability at least $1 - e^{-\omega(d\log n)}$, we have

$$Q_l \leq \tilde{O}\left(\frac{q^4 d}{n} + \sqrt{\frac{q^8 d}{n}}\right), \quad 0 \leq l \leq 2\log\log n.$$

For $Q_l, 2\log\log n < l \leq t$, we first bound the number of terms in bucket K_l. From (40), we have

$$\Pr\left[|K_l| \geq \tilde{\Omega}\left(d2^{-l/2}\right)\right] \leq e^{-\omega(d\log n)}.$$

Each term in the summation Q_l is bounded by $2^l q^4/n$, and therefore, by applying triangle inequality we have with probability at least $1 - e^{-\omega(d\log n)}$

$$Q_l \leq \tilde{O}\left(d2^{-l/2}\right)\frac{2^l q^4}{n} \leq \tilde{O}\left(\frac{q^4 d 2^{l/2}}{n}\right) \leq \tilde{O}\left(\frac{q^4 d^2}{n}\right), \quad 2\log\log n < l \leq t.$$

Here the last inequality uses the fact that $l \leq t$, which implies $2^{l/2} = \tilde{O}(d)$.

For the last term Q_{t+1}, again from (40), we have with probability at least $1 - e^{-\omega(d\log n)}$, 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 at least $1 - e^{-\omega(d\log n)}$ all the terms are bounded by $\tilde{O}\big(q^4 d^2/n + \sqrt{q^8 d/n}\big)$, which implies the summation Q is also bounded by

$$\tilde{O}\left(\frac{q^4 d^2}{n} + \sqrt{\frac{q^8 d}{n}}\right).$$

□

Now we are ready to bound the 4-th order term perturbation $\mathbb{E}[x^{\otimes 4}] - \frac{1}{n}\sum_{i=1}^{n}(x^i)^{\otimes 4}$.

Claim 5. *Suppose $\|A\| \leq O(\sqrt{k/d})$ and the entries of $h \in \mathbb{R}^k$ are independent subgaussian variables with $\mathbb{E}[h_j^2] = 1$. Given n samples $x^i = Ah^i, i \in [n]$, we have with high probability*

$$\left\| \frac{1}{n} \sum_{i \in [n]} (x^i)^{\otimes 4} - \mathbb{E}[x^{\otimes 4}] \right\| \leq \tilde{O}\left(\frac{k^2}{n} + \sqrt{\frac{k^4}{d^3 n}} \right).$$

Proof: The desired spectral norm in the lemma is defined as

$$\sup_{\|u\|=1} \left| \frac{1}{n} \sum_{i \in [n]} \langle u, x^i \rangle^4 - \mathbb{E}[\langle u, x \rangle^4] \right|.$$

In order to bound it, we provide an ε-net argument. Construct an ε-net for vectors u in the unit ball $\mathcal{S}^{d-1}$ with $\varepsilon = 1/n^2$. By standard construction, size of the ε-net is $e^{O(d \log n)}$. For any fixed u in the ε-net, let $v := A^\top u$. Since $x^i = Ah^i, i \in [n]$, we have $\langle u, x^i \rangle = \langle v, h^i \rangle$. Therefore, for any fixed u (and the corresponding v) in the ε-net, we would like to bound

$$Q := \frac{1}{n} \sum_{i \in [n]} \left(\langle v, h^i \rangle^4 - \mathbb{E}[\langle v, h^i \rangle^4] \right).$$

Since h^i's have independent subgaussian entries, we know that $\langle v, h^i \rangle$ is $\|v\|$-subgaussian. On the other hand, we have

$$\|v\| \leq \|A\| \|u\| = O(\sqrt{k/d}),$$

and therefore, $\langle v, h^i \rangle$ is a $O(\sqrt{k/d})$-subgaussian random variable. By Claim 4, with probability at least $1 - e^{-Cd \log n}$ (for large enough constant C) we have

$$|Q| \leq \tilde{O}\left(\frac{k^2}{n} + \sqrt{\frac{k^4}{d^3 n}} \right).$$

By applying union bound on all vectors in the ε-net, the bound holds for every vector u in the ε-net. The argument for other u's which are not in the ε-net follows from their closest vectors in the ε-net.

□

The 2nd order term T in (9) is sum of three terms, each of which is an outer-product of two matrices. Hence, it is good enough to apply a matrix concentration for bounding this term.

Claim 6. *Suppose $\|A\| \leq O(\sqrt{k/d})$ and the entries of $h \in \mathbb{R}^k$ are independent subgaussian variables with $\mathbb{E}[h_j^2] = 1$. Given n samples $x^i = Ah^i, i \in [n]$, for T in (9) and the empirical estimate $\widehat{T}$ in (37), if $n \geq d$, we have with high probability*

$$\|\widehat{T} - T\| \leq \tilde{O}\left(\sqrt{\frac{k^4}{d^3 n}} \right).$$

Proof: Recall $W := \frac{1}{n} \sum_{i=1}^n x^i (x^i)^\top$. We prove the result for the first term

$$\widehat{T}_1[i_1, i_2, i_3, i_4] = W_{i_1, i_2} W_{i_3, i_4},$$

or equivalently $\widehat{T}_1 = W \otimes W$. The analysis for the other two terms follow similarly from symmetry.

Let $T_1 = \mathbb{E}[xx^\top] \otimes \mathbb{E}[xx^\top] = \mathbb{E}[W] \otimes \mathbb{E}[W]$. We have

$$\widehat{T}_1 - T_1 = (W - \mathbb{E}[W]) \otimes \mathbb{E}[W] + \mathbb{E}[W] \otimes (W - \mathbb{E}[W]) + (W - \mathbb{E}[W]) \otimes (W - \mathbb{E}[W]).$$

For any matrices A and B, we have $\|A \otimes B\| \leq \|A\|\|B\|$. Thus,

$$\|\widehat{T}_1 - T_1\| \leq 2\|W - \mathbb{E}[W]\| \cdot \|\mathbb{E}[W]\| + \|W - \mathbb{E}[W]\|^2. \tag{41}$$

We bound $\|W - \mathbb{E}[W]\|$ by Matrix Bernstein's inequality. For applying Matrix Bernstein's inequality, we need a bound on the norm of each term in the summation form of W, i.e., bound on $\|x^i(x^i)^\top\|$ which holds almost surely. Therefore, we apply the Bernstein's inequality on the bounded version of W as

$$W' := \frac{1}{n}\sum_{i=1}^{n} x^i(x^i)^\top \mathbf{1}_{\|x^i\| \leq O(\sqrt{k}\log n)},$$

where $\mathbf{1}_{\|x^i\| \leq O(\sqrt{k}\log n)}$ is an indicator variable. Since $x = Ah$ and entries of h are subgaussian, the indicator 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 the summation W', the norm of each term is bounded by $\tilde{O}(k/n)$, and for the variance term, we have

$$\mathbb{E}\left[W'(W')^\top\right] = \frac{1}{n}\mathbb{E}\left[\|x^i\|^2 x^i(x^i)^\top \mathbf{1}_{\|x^i\| \leq O(\sqrt{k}\log n)}\right] \preceq \frac{1}{n}\tilde{O}(k)\mathbb{E}\left[x^i(x^i)^\top\right] = \frac{1}{n}\tilde{O}(k)AA^\top.$$

Since $\|A\| \leq O(\sqrt{k/d})$, it is concluded that the variance is bounded by $\tilde{O}(k^2/dn)$. Therefore, Matrix Bernstein's inequality implies that with probability at least $1 - d/n$,

$$\|W' - \mathbb{E}[W']\| \leq \tilde{O}\left(\frac{k}{n} + \frac{k}{\sqrt{dn}}\right).$$

Since W is equal to W' with high probability and $\|\mathbb{E}[W] - \mathbb{E}[W']\|$ is negligible, we also have $\|W - \mathbb{E}[W]\| \leq \tilde{O}(k/\sqrt{dn})$ (when $n \geq d$).

On the other hand, $\mathbb{E}[W] = AA^\top$, and therefore, $\|\mathbb{E}[W]\| \leq k/d$. From (41), we have

$$\|\widehat{T}_1 - T_1\| \leq \tilde{O}\left(\sqrt{\frac{k^4}{d^3 n}}\right).$$

□

C.3 Sparse ICA

In this section, we prove the tensor concentration result for the sparse ICA model provided in Theorem 3. This is the sparse coding problem 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.

The proof of Theorem 3 is similar to the proof of Theorem 2, where the 4th order term perturbation $\mathbb{E}[x^{\otimes 4}] - \frac{1}{n}\sum_{i=1}^{n}(x^i)^{\otimes 4}$, and the 2nd order term perturbation $T - \widehat{T}$ are separately bounded in the following two claims. First, we bound the perturbation of the 4th order term in the following claim. Note that this is the sparse version of Claim 5.

Claim 7. *Consider the sparse ICA model described in Theorem 3. Given n independent samples $x^i = Ah^i, i \in [n]$, we have with high probability*

$$\left\| \frac{1}{n} \sum_{i=1}^{n} (x^i)^{\otimes 4} - \mathbb{E}[x^{\otimes 4}] \right\| \le \tilde{O}\left(\frac{s^2}{n} + \sqrt{\frac{s^4}{d^3 n}} \right).$$

Proof: The proof uses ideas from both Claims 2 and 4 . Without loss of generality, we assume $s/k < 1/2$. Otherwise, h_j's are 2-subgaussian, and therefore the dense case argument in Claim 5 implies the desired bound.

Let η_i's be independent random ± 1 variables with $\Pr[\eta_i = 1] = 1/2$. We equivalently bound

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \eta_i \left((x^i)^{\otimes 4} - \mathbb{E}\left[(x^i)^{\otimes 4} \right] \right) \right\| := \sup_{\|u\|=1} \left| \frac{1}{n} \sum_{i \in [n]} \eta_i \left(\langle u, x^i \rangle^4 - \mathbb{E}[\langle u, x^i \rangle^4] \right) \right|.$$

In order to bound it, we provide an ε-net argument. Construct an ε-net for vectors u in the unit ball $\mathcal{S}^{d-1}$ with $\varepsilon = 1/n^2$. By standard construction, size of the ε-net is $e^{O(d \log n)}$. For any fixed u in the ε-net, let $v := A^\top u$. Since $x^i = Ah^i, i \in [n]$, we have $\langle u, x^i \rangle = \langle v, h^i \rangle$. Therefore, for any fixed u (and the corresponding v) in the ε-net, we would like to bound

$$\left| \frac{1}{n} \sum_{i \in [n]} \eta_i \left(\langle v, h^i \rangle^4 - \mathbb{E}[\langle v, h^i \rangle^4] \right) \right|. \tag{42}$$

Now, we follow the ideas of Claim 4, and apply the standard symmetrization trick: it is enough to take two independent sample sets $\{h^1, h^2, \dots, h^n\}$ and $\{\tilde{h}^1, \tilde{h}^2, \dots, \tilde{h}^n\}$ with the same distribution, and bound $\left| \frac{1}{n} \sum_{i \in [n]} \eta_i \left(\langle v, h^i \rangle^4 - \langle v, \tilde{h}^i \rangle^4 \right) \right|$ instead of (42). Note that the difference between mean and median here is negligible because our distributions have first and second moments polynomial in parameters, and strong exponential concentration. Therefore, for any vector u (and the corresponding v), we would like to bound the sum

$$\frac{1}{n} \sum_{i \in [n]} \eta_i \left| \langle v, h^i \rangle^4 - \langle v, \tilde{h}^i \rangle^4 \right|.$$

The techniques we use to prove bounds on sums of random variables $\sum_{i=1}^{n} \eta_i z_i$ (either Bernstein's inequality, or bounding the number of terms and then using triangle inequality) all works if we just know an *upper bound* of z_i. Therefore, we can equivalently bound

$$Q = \frac{1}{n} \sum_{i \in [n]} \eta_i \left(\langle v, h^i \rangle^4 + \langle v, \tilde{h}^i \rangle^4 \right),$$

where the subtraction is replaced with addition.

Now, we 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 2). 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 3)

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

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 5, the number of nonzero entries in v_l is bounded as

$$\|v_l\|_0 = |K_l| \leq O\left(\frac{d}{p_l^2}\right), \quad l > \frac{1}{2}\log\log d.$$

Exploiting the above partitioning, the term $\langle v, h^i\rangle^4$ in summation Q can be upper bounded as

$$\langle v, h^i\rangle^4 = \Big(\sum_{l=0}^{t}\langle v_l, h^i\rangle\Big)^4 \leq \Big((t+1)\sum_{l=0}^{t}\langle v_l, h^i\rangle^2\Big)^2 \leq (t+1)^3\sum_{l=0}^{t}\langle v_l, h^i\rangle^4,$$

where the equality is concluded from the fact that nonzero values of v_l's are derived from partitioning of values of v, and Cauchy-Schwartz inequality is exploited in the last two steps. Applying this upper bound on Q, we would like to bound

$$Q' := \frac{1}{n}\sum_{i\in[n]}\eta_i(t+1)^3\sum_{l=0}^{t}\left(\langle v_l, h^i\rangle^4 + \langle v_l, \tilde{h}^i\rangle^4\right).$$

In order to bound Q', we break it into sum of $t+1$ terms as $Q' = \sum_{l=0}^{t} Q'_l$ where

$$Q'_l := \frac{1}{n}(t+1)^3\sum_{i\in[n]}\eta_i\left(\langle v_l, h^i\rangle^4 + \langle v_l, \tilde{h}^i\rangle^4\right).$$

All terms Q'_l can be bounded in the same way as Claim 4. Especially, directly from Claim 4, we have

$$Q'_0 \leq \tilde{O}\left(\frac{s^2}{n} + \sqrt{\frac{s^4}{d^3 n}}\right).$$

For the other terms $Q'_l, l > \frac{1}{2}\log\log d$, we need to analyze the tail behavior of $\langle v_l, h^i\rangle^4$. The tail behavior of this variable is affected by two phenomena: 1) the size of intersection of the supports of v_l and h^i, and 2) given the intersection, the tail behavior of

$$\langle v_l, h^i\rangle = \sum_{j\in[k]:s^i[j]=1} v_l[j]g^i[j], \tag{43}$$

which is a sum of subgaussian random variables. Recall that $h^i[j] = s^i[j]g^i[j]$ where $s^i \in \mathbb{R}^k$ with i.i.d. Bernoulli random entries specifies the support of h^i.

The first part (the intersection of supports) can be bounded by Chernoff bound as

$$\Pr\Big[\sum_{j\in[k]} s^i[j] \geq (1+\delta)s\Big] \leq \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^s.$$

The second part follows from subgaussian concentrations bounds. Let $\theta_l := \frac{2^l}{\sqrt{d}}$. For bucket K_l, and subsequently Q'_l where v_l has entries in the interval $(\theta_l/2, \theta_l]$, we discuss the tail behavior in two cases where $1/\theta_l^2 \geq s$ and $1/\theta_l^2 \leq s$.

Case 1 ($1/\theta_l^2 \geq s$): In this case, most of $\langle v_l, h^i\rangle^4$ are of size s^2/k^2 which is very small. For any $q \in \big[\sqrt{s/(k\theta_l^2)}\,\mathrm{polylog}(n), s\big]$, since the summation in (43) is $\sqrt{s}\theta_l$-subgaussian, with probability at least $1 - e^{-\tilde{\Omega}(q)}$, we have

$$\langle v_l, h^i\rangle^4 \in \left(q^4\theta_l^4/2, q^4\theta_l^4\right].$$

Therefore in this range, with probability at least $1 - e^{-\tilde{\Omega}(1/\theta_l^2)}$, the summation Q'_l is bounded by

$$\frac{1}{n}\tilde{O}\left(\frac{q^4\theta_l^4}{\theta_l^2 q}\right) = \tilde{O}\left(\frac{q^3\theta_l^2}{n}\right) \leq \tilde{O}\left(\frac{s^2}{n}\right),$$

where the last inequality uses the fact that $\theta_l^2 \leq 1/s$.

For any $q \in \big(s, \sqrt{s/\theta_l^2}\log^2 n\big]$, since the summation in (43) is $\sqrt{s}\theta_l$-subgaussian, with probability at least $1 - e^{-\tilde{\Omega}(q^2/s)}$, we have

$$\langle v_l, h^i\rangle^4 \in \left(q^4\theta_l^4/2, q^4\theta_l^4\right].$$

Therefore in this range, with probability at least $1 - e^{-\tilde{\Omega}(1/\theta_l^2)}$, the summation Q'_l is bounded by

$$\frac{1}{n}\tilde{O}\left(q^4\theta_l^4\frac{1}{\theta_l^2 q^2/s}\right) = \tilde{O}\left(\frac{q^2\theta_l^2 s}{n}\right) \leq \tilde{O}\left(\frac{s^2}{n}\right),$$

where the last inequality uses the fact that $q^2 = \tilde{O}(s/\theta_l^2)$.

When $q > \sqrt{s/\theta_l^2}\log^2 n$, there are no term $\langle v_l, h^i\rangle^4$ in this range with high probability. Therefore, in the first case, by doing union bound Q'_l is always bounded by

$$\tilde{O}\left(\frac{s^2}{n}\right) + o\left(\frac{s^4}{d^3 n}\right).$$

Case 2 ($1/\theta_l^2 \leq s$): In this case, again most of $\langle v_l, h^i\rangle^4$ are of size s^2/k^2 which is very small. The only difference with case 1 is the two ranges where instead of being separated at s, they are separated at $1/\theta_l^2$ because there are at most $\tilde{O}(1/\theta_l^2)$ nonzero entries in v_l as shown earlier.

For any $q \in \big[\sqrt{s/(k\theta_l^2)}\,\mathrm{polylog}(n), 1/\theta_l^2\big]$, since the summation in (43) is $\sqrt{s}\theta_l$-subgaussian, with probability at least $1 - e^{-\tilde{\Omega}(q)}$, we have

$$\langle v_l, h^i\rangle^4 \in \left(q^4\theta_l^4/2, q^4\theta_l^4\right].$$

Therefore in this range, with probability at least $1 - e^{-\tilde{\Omega}(1/\theta_l^2)}$, the summation Q'_l is bounded by

$$\frac{1}{n}\tilde{O}\left(\frac{q^4\theta_l^4}{\theta_l^2 q}\right) = \tilde{O}\left(\frac{q^3\theta_l^2}{n}\right) \leq \tilde{O}\left(\frac{s^2}{n}\right),$$

where the last inequality uses the fact that $\theta_l^2 \leq 1/s$.

For any $q \in \left(1/\theta_l^2, \sqrt{s/\theta_l^2}\log^2 n\right]$, since the summation in (43) is $\sqrt{s}\theta_l$-subgaussian, with probability at least $1 - e^{-\tilde{\Omega}(q^2\theta_l^2)}$, we have

$$\langle v_l, h^i\rangle^4 \in \left(q^4\theta_l^4/2, q^4\theta_l^4\right].$$

Therefore in this range, with probability at least $1 - e^{-\tilde{\Omega}(1/\theta_l^2)}$, the summation Q_l' is bounded by

$$\frac{1}{n}\tilde{O}\left(q^4\theta_l^4\frac{1}{\theta_l^2 q^2\theta_l^2}\right) = \tilde{O}\left(\frac{q^2}{n}\right) \leq \tilde{O}\left(\frac{s^2}{n}\right),$$

where the last inequality uses the fact that $q^2 = \tilde{O}(s/\theta_l^2) \leq \tilde{O}(s^2)$.

When $q > \sqrt{s/\theta_l^2}\log^2 n$, there are no term $\langle v_l, h^i\rangle^4$ in this range with high probability. Therefore, in the second case, by doing union bound Q_l' is always bounded by

$$\tilde{O}\left(\frac{s^2}{n}\right) + o\left(\frac{s^4}{d^3 n}\right).$$

Combining the bounds on all terms finishes the proof. □

In the next claim we bound the perturbation of the 2nd order term T. Note that this is the sparse version of Claim 6.

Claim 8. *Consider the same sparse setting as in Theorem 3. Given n samples $x^i = Ah^i, i \in [n]$, where $\|A\| \leq O(\sqrt{k/d})$, for T in* (9) *and the empirical estimate $\widehat{T}$ in* (37)*, if $n \geq d$, we have with high probability*

$$\|\widehat{T} - T\| \leq \tilde{O}\left(\sqrt{\frac{s^4}{d^3 n}}\right).$$

Proof: The proof is very similar to Claim 6. Recall $W := \frac{1}{n}\sum_{i=1}^n x^i(x^i)^\top$. We prove the result for the first term

$$\widehat{T}_1[i_1, i_2, i_3, i_4] = W_{i_1,i_2}W_{i_3,i_4},$$

or equivalently $\widehat{T}_1 = W \otimes W$. The analysis for the other two terms follow similarly from symmetry. As in (41), we have

$$\|\widehat{T}_1 - T_1\| \leq 2\|W - \mathbb{E}[W]\| \cdot \|\mathbb{E}[W]\| + \|W - \mathbb{E}[W]\|^2.$$

We bound $\|W - \mathbb{E}[W]\|$ by Matrix Bernstein's inequality. As in Claim 6, 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, the indicator 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 the summation W', the norm of each term is bounded by $\tilde{O}(s/n)$, and for the variance term, we have

$$\mathbb{E}\left[W'(W')^\top\right] = \frac{1}{n}\mathbb{E}\left[\|x^i\|^2 x^i (x^i)^\top \mathbf{1}_{\|x^i\| \leq O(\sqrt{s}\log n)}\right] \preceq \frac{1}{n}\tilde{O}(s)\mathbb{E}\left[x^i(x^i)^\top\right] = \frac{1}{n}\tilde{O}(s^2/k)AA^\top.$$

Since $\|A\| \leq O(\sqrt{k/d})$, it is concluded that the variance is bounded by $\tilde{O}(s^2/dn)$. Therefore, Matrix Bernstein's inequality implies that with probability at least $1 - d/n$,

$$\|W' - \mathbb{E}[W']\| \leq \tilde{O}\left(\frac{s}{n} + \frac{s}{\sqrt{dn}}\right).$$

Since W is equal to W' with high probability and $\|\mathbb{E}[W] - \mathbb{E}[W']\|$ is negligible, we also have $\|W - \mathbb{E}[W]\| \leq \tilde{O}(s/\sqrt{dn})$ (when $n \geq d$).

On the other hand, $\mathbb{E}[W] = \frac{s}{k}AA^\top$, and therefore, $\|\mathbb{E}[W]\| \leq s/d$. From (41), we have

$$\|\widehat{T}_1 - T_1\| \leq \tilde{O}\left(\sqrt{\frac{s^4}{d^3 n}}\right).$$

□

References

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

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.

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

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, 2011a.

Adam Coates, Andrew Y Ng, and Honglak Lee. An analysis of single-layer networks in unsupervised feature learning. In *International Conference on Artificial Intelligence and Statistics*, pages 215–223, 2011b.

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

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

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.

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

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.