# Smoothed Analysis of Tensor Decompositions

Aditya Bhaskara* Moses Charikar† Ankur Moitra‡ Aravindan Vijayaraghavan§

**Abstract**

Low rank decomposition of tensors is a powerful tool for learning generative models. The uniqueness of decomposition gives tensors a significant advantage over matrices. However, tensors pose significant algorithmic challenges and tensors analogs of much of the matrix algebra toolkit are unlikely to exist because of hardness results. Efficient decomposition in the overcomplete case (where rank exceeds dimension) is particularly challenging. We introduce a smoothed analysis model for studying these questions and develop an efficient algorithm for tensor decomposition in the highly overcomplete case (rank polynomial in the dimension). In this setting, we show that our algorithm is robust to inverse polynomial error – a crucial property for applications in learning since we are only allowed a polynomial number of samples. While algorithms are known for exact tensor decomposition in some overcomplete settings, our main contribution is in analyzing their stability in the framework of smoothed analysis.

Our main technical contribution is to show that tensor products of perturbed vectors are linearly independent in a robust sense (i.e. the associated matrix has singular values that are at least an inverse polynomial). This key result paves the way for applying tensor methods to learning problems in the smoothed setting. In particular, we use it to obtain results for learning multi-view models and mixtures of axis-aligned Gaussians where there are many more "components" than dimensions. The assumption here is that the model is not adversarially chosen, formalized by a perturbation of model parameters. We believe this an appealing way to analyze realistic instances of learning problems, since this framework allows us to overcome many of the usual limitations of using tensor methods.

*Google Research NYC. Email: bhaskara@cs.princeton.edu. Work done while the author was at EPFL, Switzerland.

†Princeton University. Email: moses@cs.princeton.edu. Supported by NSF awards CCF 0832797, AF 1218687 and CCF 1302518

‡Massachusetts Institute of Technology, Department of Mathematics and CSAIL. Email: moitra@mit.edu. Part of this work was done while the author was a postdoc at the Institute for Advanced Study and was supported in part by NSF grant No.DMS-0835373 and by an NSF Computing and Innovation Fellowship.

§Carnegie Mellon University. Email: aravindv@cs.cmu.edu. Supported by the Simons Postdoctoral Fellowship.

# 1 Introduction

## 1.1 Background

Tensor decompositions play a central role in modern statistics (see e.g. [27]). To illustrate their usefulness, suppose we are given a matrix $M = \sum_{i=1}^{R} a_i \otimes b_i$ When can we *uniquely* recover the factors $\{a_i\}_i$ and $\{b_i\}_i$ of this decomposition given access to $M$? In fact, this decomposition is almost never unique (unless we require that the factors $\{a_i\}_i$ and $\{b_i\}_i$ are orthonormal, or that $M$ has rank one). But given a tensor $T = \sum_{i=1}^{R} a_i \otimes b_i \otimes c_i$ there are general conditions under which $\{a_i\}_i$, $\{b_i\}_i$ and $\{c_i\}_i$ are uniquely determined (up to scaling) given $T$; perhaps the most famous such condition is due to Kruskal [24], which we review in the next section.

Tensor methods are commonly used to establish that the parameters of a generative model can be identified given third (or higher) order moments. In contrast, given just second-order moments (e.g. $M$) we can only hope to recover the factors up to a rotation. This is called the *rotation problem* and has been an important issue in statistics since the pioneering work of psychologist Charles Spearman (1904) [31]. Tensors offer a path around this obstacle precisely because their decompositions are often unique, and consequently have found applications in phylogenetic reconstruction [11], [29], hidden markov models [29], mixture models [20], topic modeling [5], community detection [3], etc.

However most tensor problems are hard: computing the rank [17], the best rank one approximation [18] and the spectral norm [18] are all $NP$-hard. Also many of the familiar properties of matrices do not generalize to tensors. For example, subtracting the best rank one approximation to a tensor can actually increase its rank [34] and there are rank three tensors that can be approximated arbitrarily well by a sequence of rank two tensors. One of the few algorithmic results for tensors is an algorithm for computing tensor decompositions in a restricted case. Let $A, B$ and $C$ be matrices whose columns are $\{a_i\}_i$, $\{b_i\}_i$ and $\{c_i\}_i$ respectively.

**Theorem 1.1.** *[25], [11] If rank$(A)$ = rank$(B)$ = $R$ and no pair of columns in $C$ are multiples of each other, then there is a polynomial time algorithm to compute the minimum rank tensor decomposition of $T$. Moreover the rank one terms in this decomposition are unique (among all decompositions with the same rank).*

If $T$ is an $n \times n \times n$ tensor, then $R$ can be at most $n$ in order for the conditions of the theorem to be met. This basic algorithm has been used to design efficient algorithms for phylogenetic reconstruction [11], [29], topic modeling [5], community detection [3] and learning hidden markov models and mixtures of spherical Gaussians [20]. However algorithms that make use of tensor decompositions have traditionally been limited to the full-rank case, and our goal is to develop *stable* algorithms that work for $R = \text{poly}(n)$. Recently Goyal et al [16] gave a robustness analysis for this decomposition, and we give an alternative proof in Appendix A.

In fact, this basic tensor decomposition can be bootstrapped to work even when $R$ is larger than $n$ (if we also increase the order of the tensor). The key parameter that dictates when one can efficiently find a tensor decomposition (or more generally, when it is unique) is the *Kruskal rank*:

**Definition 1.2.** The Kruskal rank (or Krank) of a matrix $A$ is the largest $k$ for which every set of $k$ columns are linearly independent. Also the *$\tau$-robust $k$-rank* is denoted by $\text{Krank}_\tau(A)$, and is the largest $k$ for which every $n \times k$ sub-matrix $A_{|S}$ of $A$ has $\sigma_k(A_{|S}) \geq 1/\tau$.

How can we push the above theorem beyond $R = n$? We can instead work with an order $\ell$ tensor. To be concrete set $\ell = 5$ and suppose $T$ is an $n \times n \times ... \times n$ tensor. We can "flatten" $T$ to get an

order three tensor

$$T = \sum_{i=1}^{R} \underbrace{A_i^{(1)} \otimes A_i^{(2)}}_{\text{factor}} \otimes \underbrace{A_i^{(3)} \otimes A_i^{(4)}}_{\text{factor}} \otimes \underbrace{A_i^{(5)}}_{\text{factor}}$$

Hence we get an order three tensor $\widehat{T}$ of size $n^2 \times n^2 \times n$. Alternatively we can define this "flattening" using the following operation:

**Definition 1.3.** The Khatri-Rao product of $U$ and $V$ which are size $m \times r$ and $n \times r$ respectively is an $mn \times r$ matrix $U \odot V$ whose $i^{th}$ column is $u_i \otimes v_i$.

Our new order three tensor $\widehat{T}$ can be written as:

$$\widehat{T} = \sum_{i=1}^{R} \left(A^{(1)} \odot A^{(2)}\right)_i \otimes \left(A^{(3)} \odot A^{(4)}\right)_i \otimes A^{(5)}$$

The factors are the columns of $A^{(1)} \odot A^{(2)}$, the columns of $A^{(3)} \odot A^{(4)}$ and the columns of $A^{(5)}$. The crucial point is that the Kruskal rank of the columns of $A^{(1)} \odot A^{(2)}$ is in fact at least the sum of the Kruskal rank of the columns of $A^{(1)}$ and $A^{(2)}$ (and similarly for $A^{(3)} \odot A^{(4)}$) [1], [9], but this is tight in the worst-case. Consequently this "flattening" operation allows us use the above algorithm unto $R = 2n$; since the rank ($R$) is larger than the largest dimension ($n$), this is called the *overcomplete* case.

Our main technical result is that in a natural *smoothed analysis* model, the Kruskal rank *robustly multiplies* and this allows us to give algorithms for computing a tensor decomposition even in the highly overcomplete case, for any $R = \text{poly}(n)$ (provided that the order of the tensor is large - but still a constant). Moreover our algorithms have immediate applications in learning mixtures of Gaussians and multiview mixture models.

## 1.2 Our Results

We introduce the following framework for studying tensor decomposition problems:

- An adversary chooses a tensor $T = \sum_{i=1}^{R} A_i^{(1)} \otimes A_i^{(2)} \otimes ... \otimes A_i^{(\ell)}$.
- Each vector $a_i^j$ is $\rho$-perturbed to yield $\tilde{a}_i^j$.[1]
- We are given $\widetilde{T} = \sum_{i=1}^{R} \widetilde{A}_i^{(1)} \otimes \widetilde{A}_i^{(2)} \otimes ... \otimes \widetilde{A}_i^{(\ell)}$ (possibly with noise.)

Our goal is to recover the factors $\{\widetilde{A}_i^{(1)}\}_i$, $\{\widetilde{A}_i^{(2)}\}_i$, ..., $\{\widetilde{A}_i^{(\ell)}\}_i$ (up to rescaling). This model is directly inspired by *smoothed analysis* which was introduced by Spielman and Teng [32], [33] as a framework in which to understand why certain algorithms perform well on realistic inputs.

In applications in learning, tensors are used to encode low-order moments of the distribution. In particular, each factor in the decomposition represents a "component". The intuition is that if these "components" are not chosen in a worst-case configuration, then we can obtain vastly improved learning algorithms in various settings. For example, as a direct consequence of our main result, we will give new algorithms for learning mixtures of spherical Gaussians again in the framework of smoothed analysis (without any additional separation conditions). There are no known polynomial

[1] An (independent) random gaussian with zero mean and variance $\rho^2/n$ in each coordinate is added to $a_i^j$ to obtain $\tilde{a}_i^j$. We note that we make the Gaussian assumption for convenience, but our analysis seems to apply to more general perturbations.

time algorithms to learn such mixures if the number of components ($k$) is larger than the dimension ($n$). But if their means are perturbed, we give a polynomial time algorithm for any $k = \text{poly}(n)$ by virtue of our tensor decomposition algorithm.

Our main technical result is the following:

**Theorem 1.4.** *Let $R \leq n^\ell/2$ for some constant $\ell \in \mathbb{N}$. Let $A^{(1)}, A^{(2)}, \ldots A^{(\ell)}$ be $n \times R$ matrices with columns of unit norm, and let $\widetilde{A}^{(1)}, \widetilde{A}^{(2)}, \ldots \widetilde{A}^{(\ell)} \in \mathbb{R}^{n \times m}$ be their respective $\rho$-perturbations. Then for $\tau = (n/\rho)^{3^\ell}$, the Khatri-Rao product satisfies*

$$\textit{Krank}_\tau \left( \widetilde{A}^{(1)} \odot \widetilde{A}^{(2)} \odot \ldots \odot \widetilde{A}^{(\ell)} \right) = R \quad \textit{w.p. at least} \quad 1 - \exp\left(-Cn^{1/3^\ell}\right) \tag{1}$$

In general the Kruskal rank *adds* [1, 9] but in the framework of smoothed analysis it *robustly multiplies*. What is crucial here is that we have a lower bound $\tau$ on how close these vectors are to linearly dependent. In almost all of the applications of tensor methods, we are not given $T$ exactly but rather with some noise. This error could arise, for example, because we are using a finite number of samples to estimate the moments of a distribution. It is the condition number of $\widetilde{A}^{(1)} \odot \widetilde{A}^{(1)} \odot \ldots \odot \widetilde{A}^{(\ell)}$ that will control whether various tensor decomposition algorithms work in the presence of noise.

Another crucial property our method achieves is exponentially small failure probability for any constant $\ell$, for our polynomial bound on $\tau$. In particular for $\ell = 2$, we show (in Theorem 3.1) for $\rho$-perturbations of two $n \times n^2/2$ matrices $U$ and $V$, the $\text{Krank}_\tau(\widetilde{U} \odot \widetilde{V}) = n^2/2$ for $\tau = \rho^2/n^{O(1)}$, with probability $1 - \exp(-\sqrt{n})$. We remark that it is fairly straightforward to obtain the above statement (for $\ell = 2$) for failure probability $\delta$, with $\tau = (n/\delta)^{O(1)}$ (see Remark 3.7 for more on the latter); however, this is not desirable since the running time has a polynomial dependence on the minimum singular value $1/\tau$ (and hence $\delta$).

We obtain the following main theorem from the above result and from analyzing the stability of the algorithm of Leurgans et al [25] (see Theorem 2.3):

**Theorem 1.5.** *Let $R \leq n^{\lfloor \frac{\ell-1}{2} \rfloor}/2$ for some constant $\ell \in \mathbb{N}$. Suppose we are given $\widetilde{T} + E$ where $\widetilde{T}$ and $E$ are order $\ell$-tensors and $\widetilde{T}$ has rank $R$ and is obtained from the above smoothed analysis model. Moreover suppose the entries of $E$ are at most $\varepsilon(\rho/n)^{3^\ell}$ where $\varepsilon < 1$. Then there is an algorithm to recover the rank one terms $\otimes_{i=1}^{\ell} \widetilde{a}_i^j$ up to an additive $\varepsilon$ error. The algorithm runs in time $n^{C3^\ell}$ and succeeds with probability at least $1 - \exp(-Cn^{1/3^\ell})$.*

As we discussed, tensor methods have had numerous applications in learning. However algorithms that make use of tensor decompositions have traditionally been limited to the full-rank case, and hence can only handle cases when the number of "components" is at most the dimension. However by using our main theorem above, we can get new algorithms for some of these problems that work even if there are many more "components" than dimensions.

### Multi-view Models (Section 4)

In this setting, each sample is composed of $\ell$ views $x^{(1)}, x^{(2)}, \ldots, x^{(\ell)}$ which are conditionally independent given which component $i \in [R]$ the sample is generated from. Hence such a model is specified by $R$ mixing weights $w_i$ and $R$ discrete distributions $\mu_i{}^{(1)}, \ldots, \mu_i{}^{(j)}, \ldots, \mu_i{}^{(\ell)}$, one for each view. Such models are very expressive and are used as a common abstraction for a number of inference problems. Anandkumar et al [2] gave algorithms in the full rank setting. However, in many practical settings like speech recognition and image classification, the dimension of the feature space is typically much smaller than the number of components. If we suppose that the

distributions that make up the multi-view model are $\rho$-perturbed (analogously to the tensor setting) then we can give the first known algorithms for the overcomplete setting. Suppose that the means $(\mu_i^{(j)})$ are $\rho$-perturbed to obtain $\{\widetilde{\mu}_i^{(j)}\}$. Then:

**Theorem 1.6.** *This is an algorithm to learn the parameters $w_i$ and $\{\widetilde{\mu}_i^{(j)}\}$ of an $\ell$-view multi-view model with $R \leq n^{\lfloor \frac{\ell-1}{2} \rfloor}/2$ components up to an accuracy $\varepsilon$. The running time and sample complexity are at most $\mathrm{poly}_\ell(n, 1/\varepsilon, 1/\rho)$ and succeeds with probability at least $1 - \exp(-Cn^{1/3^\ell})$ for some constant $C > 0$.*

**Mixtures of Axis-Aligned Gaussians (Section 5)**

Here we are given samples from a distribution $F = \sum_{i=1}^k w_i F_i(\mu_i, \Sigma_i)$ where $F_i(\mu_i, \Sigma_i)$ is a Gaussian with mean $\mu_i$ and covariance $\Sigma_i$ and each $\Sigma_i$ is diagonal. These mixtures are ubiquitous throughout machine learning. Feldman et al [14] gave an algorithm for PAC-learning mixtures of axis aligned Gaussians, however the running time is exponential in $k$, the number of components. Hsu and Kakade [20] gave a polynomial time algorithm for learning mixtures of spherical Gaussians provided that their means are full rank (hence $k \leq n$). Again, we turn to the framework of smoothed analysis and suppose that the means are $\rho$-perturbed. In this framework, we can give a polynomial time algorithm for learning mixtures of axis-aligned Gaussians for any $k = \mathrm{poly}(n)$. Suppose that the means of a mixture of axis-aligned Gaussians and suppose the means have been $\rho$-perturbed to obtain $\widetilde{\mu}_i$. Then

**Theorem 1.7.** *There is an algorithm to learn the parameters $w_i$, $\widetilde{\mu}_i$ and $\Sigma_i$ of a mixture of $k \leq n^{\lfloor \frac{\ell-1}{2} \rfloor}/(2\ell)$ axis-aligned Gaussians up to an accuracy $\varepsilon$. The running time and sample complexity are at most $\mathrm{poly}_\ell(n, 1/\varepsilon, 1/\rho)$ and succeeds with probability at least $1 - \exp(-Cn^{1/3^\ell})$ for some constant $C > 0$.*

We believe that our new algorithms for overcomplete tensor decomposition will have further applications in learning. Additionally this framework of studying distribution learning when the parameters of the distribution we would like to learn are not chosen adversarially, seems quite appealing.

**Remark 1.8.** Recall, our main technical result is that the Kruskal rank *robustly* multiplies. In fact, is is easy to see that for a generic set of vectors it multiplies [1]. This observation, in conjunction with the algorithm of Leurgans et al [25] yields an algorithm for tensor decomposition in the overcomplete case. Another approach to overcomplete tensor decomposition was given by [13] which works up to $r \leq n^{\lfloor \frac{\ell}{2} \rfloor}$. However these algorithms assume that we know $T$ *exactly*, and are not known to be stable when we are given $T$ with noise. The main issue is that these algorithms are based on solving a linear system which is full rank if the factors of $T$ are generic, but what controls whether or not these linear systems can handle noise is their condition number.

Alternatively, algorithms for overcomplete tensor decomposition that assume we know $T$ exactly would not have any applications in learning because we would need to take too many samples to have a good enough estimate of $T$ (i.e. the low-order moments of the distribution).

In recent work, Goyal et al [16] also made use of robust algorithms for overcomplete tensor decomposition, and their main application is underdetermined independent component analysis (ICA). The condition that they need to impose on the tensor holds generically (like ours, see e.g. Corollary 2.4) and can show in a smoothed analysis model that this condition holds with inverse polynomial failure probability. However here our focus was on showing a lower bound for the

condition number of $M^{\odot \ell}$ that does not depend (polynomially) on the failure probability. We focus on the failure probability being small (in particular, exponentially small), because in smoothed analysis, the perturbation is "one-shot" and if it does not result in an easy instance, you cannot ask for a new one!

### 1.3 Our Approach

Here we give some intuition for how we prove our main technical theorem, at least in the $\ell = 2$ case. Recall, we are given two matrices $U^{(1)}$ and $U^{(2)}$ whose $R$ columns are $\rho$-perturbed to obtain $\widetilde{U}^{(1)}$ and $\widetilde{U}^{(2)}$ respectively. Our goal is to prove that if $R \leq \frac{n^2}{2}$ then the matrix $\widetilde{U}^{(1)} \odot \widetilde{U}^{(2)}$ has smallest singular value that is at least $\text{poly}(1/n, \rho)$ with high probability. In fact, it will be easier to work with what we call the *leave-one-out distance* (see Definition 3.4) as a surrogate for the smallest singular value (see Lemma 3.5). Alternatively, if we let $x$ and $y$ be the first columns of $\widetilde{U}^{(1)}$ and $\widetilde{U}^{(2)}$ respectively, and we set

$$\mathcal{U} = \text{span}(\{\widetilde{U}_i^{(1)} \otimes \widetilde{U}_i^{(2)}, 2 \leq i \leq R\})$$

then we would like to prove that with high probability $x \otimes y$ has a non-negligible projection on the orthogonal complement of $\mathcal{U}$. This is the core of our approach. Set $\mathcal{V}$ to be the orthogonal complement of $\mathcal{U}$. In fact, we prove that for *any* dimension at least $\frac{n^2}{2}$ subspace $\mathcal{V}$, with high probability $x \otimes y$ has a non-negligible projection onto $\mathcal{V}$.

How can we reason about the projection of $x \otimes y$ onto an arbitrary (but large) dimensional subspace? If $\mathcal{V}$ were (say) the set of all low-rank matrices, then this would be straightforward. But what complicates this is that we are looking at the projection of a rank one matrix onto a large dimensional subspace of matrices, and these two spaces can be structured quite differently. A natural approach is to construct matrices $M_1, M_2, ..., M_p \in \mathcal{V}$ so that with high probability at least one quadratic form $x^T M_i y$ is non-negligible. Suppose the following condition were met (in which case we would be done): Suppose that there is a large set $S$ of indices so that each vector $x^T M_i$ has a large projection onto the orthogonal complement of $\text{span}(\{x^T M_i, i \in S\})$. In fact, if such a set $S$ exists with high probability then this would yield our main technical theorem in the $\ell = 2$ case. Our main step is in constructing a family of matrices $M_1, M_2, ...M_p$ that help us show that $S$ is large. We call this an $(\theta, \delta)$-orthogonal system (see Definition 3.13). The intuition behind this definition is that if we reveal a column in one of the $M_i$'s that has a significant orthogonal component to all of the columns that we have revealed so far, this is in effect a fresh source of randomness that can help us add another index to the set $S$. See Section 3 for a more complete description of our approach in the $\ell = 2$ case. The approach for $\ell > 2$ relies on the same basic strategy but requires a more delicate induction argument. See Section 3.4.

## 2 Prior Algorithms

Here we review the algorithm of Leurgans et al [25]. It has been discovered many times in different settings. It is sometimes referred to as "simultaneous diagonalization" or as Chang's lemma [11].

Suppose we are given a third-order tensor $T = \sum_{i=1}^{R} u_i \otimes v_i \otimes w_i$ which is $n \times m \times p$. Let $U, V$ and $W$ be matrices whose columns are $u_i$, $v_i$ and $w_i$ respectively. Suppose further that (1) $rank(U) = rank(V) = R$ and (2) k-rank$(W) \geq 2$. Then we can efficiently recover the factors of $T$.

We present the algorithm Decompose and its analysis assuming $n = m = R$. Any instance with $rank(U) = rank(V) = R$ can be reduced to this case as follows: find the span of the vectors $\{\widetilde{u}_{j,k}\}$, where $\widetilde{u}_{j,k}$ is the $n$ dimensional vector whose $i$th entry is $T_{ijk}$. This span must be precisely

the span of the columns of $U$.[2] Thus we can pick some orthonormal basis for this span, and write $T$ as an $R \times m \times p$ tensor. We can perform this operation again (along the second mode) to move to an $R \times R \times p$ tensor.

**Theorem 2.1.** *[25], [11] Given a tensor $T$ there exists an algorithm that runs in polynomial time and recovers the (unique) factors of $T$ provided that (1) $rank(U) = rank(V) = R$ and (2) $k\text{-}rank(W) \geq 2$.*

**Proof:** The algorithm is to pre-process as above (i.e., obtain $m = n = R$), and then run DECOMPOSE stated below. Let us thus analyze DECOMPOSE with $m, n$ being $R$.

We can write $T_a = UD_aV^T$ where $D_a = \text{diag}(a^Tw_1, a^Tw_2, ..., a^Tw_n)$ and similarly $T_b = UD_bV^T$ where $D_b = \text{diag}(b^Tw_1, b^Tw_2, ..., b^Tw_n)$. Moreover we can write $T_a(T_b)^{-1} = UD_aD_b^{-1}U^{-1}$ and $(T_b)^{-1}(T_a) = VD_b^{-1}D_aV^{-1}$. So we conclude $U$ and $V$ diagonalize $T_a(T_b)^{-1}$ and $(T_b)^{-1}T_a$ respectively. Note that almost surely the diagonals entries of $D_aD_b^{-1}$ are distinct (Claim A.4). Hence the eigendecompositions of $T_a(T_b)^{-1}$ and $(T_b)^{-1}(T_a)$ are unique, and we can pair up columns in $U$ and columns in $V$ based on their eigenvalues (we pair up $u$ and $v$ if their eigenvalues are equal). We can then solve a linear system to find the remaining factors (columns in $W$) and since this is a valid decomposition, we can conclude that these are also the true factors of $T$ appealing to Kruskal's uniqueness theorem [24]. ■

In fact, this algorithm is also stable, as Goyal et al [16] also recently showed. It is intuitive that if $U$ and $V$ are well-conditioned and each pair of columns in $W$ is well-conditioned then this algorithm can tolerate some inverse polynomial amount of noise. For completeness, we give a robustness analysis of DECOMPOSE in Appendix A.

**Condition 2.2.** *1. The condition numbers $\kappa(U), \kappa(V) \leq \kappa$,*

*2. The column vectors of $W$ are not close to parallel: for all $i \neq j$, $\|\frac{w_i}{\|w_i\|} - \frac{w_j}{\|w_j\|}\|_2 \geq \delta$ ,*

*3. The decompositions are bounded : for all $i$, $\|u_i\|_2, \|v_i\|_2, \|w_i\|_2 \leq C$.*

**Theorem 2.3.** *Suppose we are given tensor $T + E \in \mathbb{R}^{m \times n \times p}$ with the entries of $E$ being bounded by $\epsilon \cdot poly(1/\kappa, 1/n, 1/\delta)$ and moreover $T$ has a decomposition $T = \sum_{i=1}^{R} u_i \otimes v_i \otimes w_i$ that satisfies Condition 2.2. Then there exists an efficient algorithm that returns* each rank one term *in the decomposition of $T$ (up to renaming), within an additive error of $\epsilon$.*

As before, the algorithm is to preprocess so as to obtain $m = n = R$, and then run DECOMPOSE. The preprocessing step is slightly different because of the presence of error – instead of considering the span of the $\{\widetilde{u}_{j,k}\}$ as above, we need to look at the span of the top $R$ singular vectors of the matrix whose columns are $\widetilde{u}_{j,k}$. If $\|E\|_F$ is small enough (in terms of $\kappa, \delta, n$), the span of these top singular vectors suffices to obtain an approximation to the vectors $u_i$ (see Appendix A).

Note that the algorithm is limited by the condition that $rank(U) = rank(V) = R$ since this requires that $R \leq \min(m, n)$. But as we have seen before, by "flattening" a higher order tensor, we can handle overcomplete tensors. The following is an immediately corollary of Theorem 2.3:

[2] It is easy to see that the span is contained in the span of the columns of $U$. To see equality, we observe that if the span is $R - 1$ dimensional, then projecting each of the $u_i$s on to the span gives a *different* decomposition, and this contradicts Kruskal's uniqueness theorem, which holds in this case.

**Algorithm 1** DECOMPOSE, **Input:** $T \in \mathbb{R}^{R\times R\times R}$

1. Let $T_a = T(\cdot,\cdot,a), T_b = T(\cdot,\cdot,b)$ where $a, b$ are uniformly random unit vectors in $\Re^p$
2. Set $U$ to be the eigenvectors of $T_a(T_b)^{-1}$
3. Set $V$ to be the eigenvectors of $(T_b)^{-1}T_a$
4. Solve the linear system $T = \sum_{i=1}^{n} u_i \otimes v_i \otimes w_i$ for the vectors $w_i$
5. Output $U, V, W$

**Corollary 2.4.** *Suppose we are given an order-$\ell$ tensor $T + E \in \mathbb{R}^{n^{\times \ell}}$ with the entries of $E$ being bounded by $\epsilon \cdot poly_\ell(1/\kappa, 1/n, 1/\delta)$, and matrices $U^{(1)}, U^{(2)} \ldots U^{(\ell)} \in \mathbb{R}^{n\times r}$, whose columns give a rank-$r$ decomposition $T = \sum_{i=1}^{R} u_i^{(1)} \otimes u_i^{(2)} \otimes \cdots \otimes u_i^{(\ell)}$. If Condition 2.2 is satisfied by*

$$U = U^{(1)} \odot U^{(2)} \odot \ldots \odot U^{(\lfloor \frac{\ell-1}{2} \rfloor)}\ ,\ V = U^{(\lfloor \frac{\ell-1}{2} \rfloor+1)} \odot \ldots \odot U^{(2\lfloor \frac{\ell-1}{2} \rfloor)} \quad \text{and} \quad W = \begin{cases} U^{(\ell)} & \text{if } \ell \text{ is odd} \\ U^{(\ell-1)} \odot U^{(\ell)} & \text{otherwise} \end{cases}$$

*then there exists an efficient algorithm that computes each rank one term in this decomposition up to an additive error of $\epsilon$.*

Note that Corollary 2.4 does not require the decomposition to be symmetric. Further, any tri-partition of the $\ell$ modes that satisfies Condition 2.2 would have sufficed. To understand how large a rank we can handle, the key question is: *When does the Kruskal rank (or rank) of $\ell$-wise Khatri-Rao product become $R$?*

The following lemma is well-known (see [9] for a robust analogue) and is known to be tight in the worst case. This allows us to handle a rank of $R \approx \ell n/2$.

**Lemma 2.5.** $Krank(U \odot V) \geq \min\big(Krank(U) + Krank(V) - 1, R\big)$

But, for generic vectors set of vectors $U$ and $V$, a much stronger statement is true [1]: $\text{Krank}(U \odot V) \geq \min\big(\text{Krank}(U) \times \text{Krank}(V), R\big)$. Hence given a generic order $\ell$ tensor $T$ with $R \leq n^{\lfloor(\ell-1)/2\rfloor}$, "flattening" it to order three and appealing to Theorem 2.1 finds the factors uniquely. The algorithm of [13] follows a similar but more involved approach, and works for $R \leq n^{\lfloor(\ell)/2\rfloor}$.

However in learning applications we are not given $T$ exactly but rather an approximation to it. Our goal is to show that the Kruskal rank *robustly* multiplies typically, so that these types of tensor algorithms will not only work in the exact case, but are also necessarily stable when we are given $T$ with some noise. In the next section, we show that in the smoothed analysis model, the robust Kruskal rank multiplies on taking Khatri-Rao products. This then establishes our main result Theorem 1.5, assuming Theorem 3.3 which we prove in the next section.

**Proof of Theorem 1.5:** As in Corollary 2.4, let $U = \widetilde{U}^{(1)} \odot \ldots \odot \widetilde{U}^{(\lfloor \frac{\ell-1}{2} \rfloor)}$ , $V = \widetilde{U}^{(\lfloor \frac{\ell-1}{2} \rfloor+1)} \odot \ldots \odot \widetilde{U}^{(\ell-1)}$ and $W = \widetilde{U}^{(\ell)}$. Theorem 3.3 shows that with probability $1 - \exp\big(-n^{1/3^{O(\ell)}}\big)$ over the random $\rho$-perturbations, $\kappa_R(U), \kappa_R(V) \leq (n/\rho)^{3^\ell}$. Further, the columns $W$ are $\delta = \rho/n$ far from parallel with high probability. Hence, Corollary 2.4 implies Theorem 1.5. ■

## 3 The Khatri-Rao Product Robustly Multiplies

In the exact case, it is enough to show that the Kruskal rank almost surely multiplies and this yields algorithms for overcomplete tensor decomposition if we are given $T$ *exactly* (see Remark 1.8).

But if we want to prove that these algorithms are stable, we need to establish that even the robust Kruskal rank (possibly with a different threshold $\tau$) also multiplies. This ends up being a very natural question in *random matrix theory*, albeit the Khatri-Rao product of two perturbed vectors in $\mathbb{R}^n$ is far from a perturbed vector in $\mathbb{R}^{n^2}$.

Formally, suppose we have two matrices $U$ and $V$ with columns $u_1, u_2, \ldots, u_R$ and $v_1, v_2, \ldots, v_R$ in $\mathbb{R}^n$. Let $\widetilde{U}, \widetilde{V}$ be $\rho$-perturbations of $U, V$ i.e. for each $i \in [R]$, we perturb $u_i$ with an (independent) random gaussian perturbation of norm $\rho$ to obtain $\widetilde{u}_i$ (and similarly for $\widetilde{v}_i$). Then we show the following:

**Theorem 3.1.** *Suppose $U, V$ are $n \times R$ matrices and let $\widetilde{U}, \widetilde{V}$ be $\rho$-perturbations of $U, V$ respectively. Then for any constant $\delta \in (0,1)$, $R \leq \delta n^2$ and $\tau = n^{O(1)}/\rho^2$, the Khatri-Rao product satisfies $Krank_\tau(\widetilde{U} \odot \widetilde{V}) = R$ with probability at least $1 - \exp(-\sqrt{n})$.*

**Remark 3.2.** The natural generalization where the vectors $u_i$ and $v_i$ are in different dimensional spaces also holds. We omit the details here.

In general, a similar result holds for $\ell$-wise Khatri-Rao products which allows us to handle rank as large as $\delta n^{\lfloor \frac{\ell-1}{2} \rfloor}$ for $\ell = O(1)$. Note that this does not follow by repeatedly applying the above theorem (say applying the theorem to $U \odot V$ and then taking $\odot W$), because perturbing the entries of $(U \odot V)$ is not the same as $\widetilde{U} \odot \widetilde{V}$. In particular, we have only $\ell \cdot nR$ "truly" random bits, which are the perturbations of the columns of the base matrices. The overall structure of the proof is the same, but we need additional ideas followed by a delicate induction.

**Theorem 3.3.** *For any $\delta \in (0,1)$, let $R = \delta n^\ell$ for some constant $\ell \in \mathbb{N}$. Let $U^{(1)}, U^{(2)}, \ldots U^{(\ell)}$ be $n \times R$ matrices with unit column norm, and let $\widetilde{U}^{(1)}, \widetilde{U}^{(2)}, \ldots \widetilde{U}^{(\ell)} \in \mathbb{R}^{n \times m}$ be their respective $\rho$-perturbations. Then for $\tau = (n/\rho)^{3^\ell}$, the Khatri-Rao product satisfies*

$$Krank_\tau \left( \widetilde{U}^{(1)} \odot \widetilde{U}^{(1)} \odot \ldots \odot \widetilde{U}^{(\ell)} \right) = n^\ell/2 \quad w.p. \text{ at least } \; 1 - \exp\left(-\delta n^{1/3^\ell}\right) \tag{2}$$

Let $A$ denote the $n^\ell \times R$ matrix $\widetilde{U}^{(1)} \odot \widetilde{U}^{(2)} \odot \ldots \odot \widetilde{U}^{(\ell)}$ for convenience. The theorem states that the smallest singular value of $A$ is lower-bounded by $\tau$.

*How can we lower bound the smallest singular value of $A$?* We define a quantity which is can be used as a proxy for the least singular value and is simpler to analyze.

**Definition 3.4.** For any matrix $A$ with columns $A_1, A_2, \ldots A_R$, the leave-one-out distance is

$$\ell(A) = \min_i \; dist(A_i, span\{A_j\}_{j \neq i}).$$

The leave-one-out distance is a good proxy for the least singular value, if we are not particular about losing multiplicative factors that are polynomial in size of the matrix.

**Lemma 3.5.** *For any matrix $A$ with columns $A_1, A_2, \ldots A_R$, we have $\frac{\ell(A)}{\sqrt{R}} \leq \sigma_{min}(A) \leq \ell(A)$.*

We will show that each of the vectors $A_i = \widetilde{u}_i^{(1)} \otimes \widetilde{u}_i^{(2)} \otimes \cdots \otimes \widetilde{u}_i^{(\ell)}$ has a reasonable projection (at least $n^{\ell/2}/\tau$) on the space orthogonal to the span of the rest of the vectors span $(\{A_j : j \in [R] - \{i\}\})$ with high probability. We do not have a good handle on the space spanned by the rest of the $R-1$ vectors, so we will prove a more general statement in Theorem 3.6: we will prove that a perturbed vector $\widetilde{x}^{(1)} \otimes \cdots \otimes \widetilde{x}^{(\ell)}$ has a reasonable projection onto *any* (fixed) subspace $\mathcal{V}$ w.h.p., as long as $\dim(\mathcal{V})$ is $\Omega(n^\ell)$. To say that a vector $w$ has a reasonable projection onto $\mathcal{V}$, we just need to exhibit a set of vectors in $\mathcal{V}$ such that one of them have a large inner product with $w$. This will imply our the required bound on the singular value of $A$ as follows:

1. Fix an $i \in [R]$ and apply Theorem 3.6 with $x^{(t)} = u_i^{(t)}$ for all $t \in [\ell]$, and $\mathcal{V}$ being the space orthogonal to rest of the vectors $A_j$.

2. Apply a union bound over all the $R$ choices for $i$.

We now state the main technical theorem about projections of perturbed product vectors onto arbitrary subspaces of large dimension.

**Theorem 3.6.** *For any constant $\delta \in (0,1)$, given any subspace $\mathcal{V}$ of dimension $\delta \cdot n^\ell$ in $\mathbb{R}^{n^{\times \ell}}$, there exists tensors $T_1, T_2, \ldots T_r$ in $\mathcal{V}$ of unit norm ($\|\cdot\|_F = 1$), such that for random $\rho$-perturbations $\widetilde{x}^{(1)}, \widetilde{x}^{(2)}, \ldots, \widetilde{x}^{(\ell)} \in \mathbb{R}^n$ of any vectors $x^{(1)}, x^{(2)}, \ldots, x^{(\ell)} \in \mathbb{R}^n$, we have*

$$\mathbf{Pr}\left[\exists j \in [r] \ s.t \ \|T_j\left(\widetilde{x}^{(1)}, \widetilde{x}^{(2)}, \ldots, \widetilde{x}^{(\ell)}\right)\| \geq \rho^\ell \left(\frac{1}{n}\right)^{3^\ell}\right] \geq 1 - \exp\left(-\delta n^{1/(2\ell)^\ell}\right) \tag{3}$$

**Remark 3.7.** Since the squared length of the projection is a degree $2\ell$ polynomial of the (Gaussian) variables $x_i$, we can apply standard anti-concentration results (Carbery-Wright, for instance) to conclude that the smallest singular value (in Theorem 3.6) is at least an inverse polynomial, with failure probability at most an inverse polynomial. This approach can only give a singular value lower bound of $\text{poly}_\ell(p/n)$ for a failure probability of $p$, which is not desirable since the running time depends on the smallest singular value.

**Remark 3.8.** For meaningful guarantees, we will think of $\delta$ as a small constant or $n^{-o(1)}$ (note the dependence of the error probability on $\delta$ in eq (3)). For instance, as we will see in section 3.4, we can not hope for exponential small failure probability when $\mathcal{V} \subseteq \mathbb{R}^{n^2}$ has dimension $n$.

The following restatement of Theorem 3.6 gives a sufficient condition about the singular values of a matrix $P$ of size $r \times n^\ell$, that gives a strong anti-concentration property for values attained by vectors obtained by the tensor product of perturbed vectors. This alternate view of Theorem 3.6 will be crucial in the inductive proof for higher $\ell$-wise products in section 3.4.

**Theorem 3.9** (Restatement of Theorem 3.6)**.** *Given any constant $\delta_\ell \in (0,1)$ and any matrix $T$ of size $r \times (n^\ell)$ such that $\sigma_{\delta n^\ell} \geq \eta$, then for random $\rho$-perturbations $\widetilde{x}^{(1)}, \widetilde{x}^{(2)}, \ldots, \widetilde{x}^{(\ell)} \in \mathbb{R}^n$ of any vectors $x^{(1)}, x^{(2)}, \ldots, x^{(\ell)} \in \mathbb{R}^n$, we have*

$$\mathbf{Pr}\left[\|M\left(\widetilde{x}^{(1)}, \widetilde{x}^{(2)}, \ldots, \widetilde{x}^{(\ell)}\right)\| \geq \eta\rho^\ell \left(\frac{1}{n}\right)^{3^{O(\ell)}}\right] \geq 1 - \exp\left(-\delta n^{1/3^\ell}\right) \tag{4}$$

**Remark 3.10.** Theorem 3.6 follows from the above theorem by choosing an orthonormal basis for $\mathcal{V}$ as the rows of $T$. The other direction follows by choosing $\mathcal{V}$ as the span of the top $\delta_\ell n^\ell$ right singular vectors of $T$.

**Remark 3.11.** Before proceeding, we remark that both forms of Theorem 3.6 could be of independent interest. For instance, it follows from the above (by a small trick involving partitioning the coordinates), that a vector $\widetilde{x}^{\otimes \ell}$ has a non-negligible projection into any $cn^\ell$ dimensional subspace of $\mathbb{R}^{n^\ell}$ with probability $1 - \exp(-f_\ell(n))$. For a vector $x \in \mathbb{R}^{n^\ell}$ whose entries are all independent Gaussians, such a claim follows easily, with probability roughly $1 - \exp(-n^\ell)$. The key difference for us is that $\widetilde{x}^{\otimes \ell}$ has essentially just $n$ bits of randomness, so many of the entries are highly correlated. So the theorem says that even such a correlated perturbation has enough mass in any large enough subspace, with high enough probability. A natural conjecture is that the probability bound can be improved to $1 - \exp(-\Omega(n))$, but it is beyond the reach of our methods.

### 3.1 Khatri-Rao Product of Two Matrices

We first show Theorem 3.9 for the case $\ell = 2$. This illustrates the main ideas underlying the general proof.

**Proposition 3.12.** *Let $0 < \delta < 1$ and $M$ be a $\delta n^2 \times n^2$ matrix with $\sigma_{\delta n^2}(M) \geq \tau$. Then for random $\rho$-perturbations $\widetilde{x}, \widetilde{y}$ of any two $x, y \in \mathbb{R}^n$, we have*

$$\mathbf{Pr}\left[\|M\left(\widetilde{x} \otimes \widetilde{y}\right)\| \geq \frac{\tau\rho}{n^{O(1)}}\right] \geq 1 - \exp\left(-\sqrt{\delta n}\right). \tag{5}$$

The high level outline is now the following. Let $\mathcal{U}$ denote the span of the top $\delta n^2$ singular vectors of $M$. We show that for $r = \Omega(\sqrt{n})$, there exist $n \times n$ matrices $M_1, M_2, \ldots, M_r$ whose columns satisfy certain orthogonal properties we define, and additionally $\mathrm{vec}(M_i) \in \mathcal{U}$ for all $i \in [r]$. We use the orthogonality properties to show that $(\widetilde{x} \otimes \widetilde{y})$ has an $\rho/\mathrm{poly}(n)$ dot-product with at least one of the $M_i$ with probability $\geq 1 - \exp(-r)$.

**The $\theta$-orthogonality property.** In order to motivate this, let us consider some matrix $M_i \in \mathbb{R}^{n \times n}$ and consider $M_i(x \otimes y)$. This is precisely $y^T M_i x$. Now suppose we have $r$ matrices $M_1, M_2, \ldots, M_r$, and we consider the sum $\sum_i (y^T M_i x)^2$. This is also equal to $\|Q(y)x\|^2$, where $Q(y)$ is an $r \times n$ matrix whose $(i,j)$th entry is $\langle y, (M_i)_j \rangle$ (here $(M_i)_j$ refers to the $j$th column in $M_i$).

Now consider some matrices $M_i$, and suppose we knew that $Q(\widetilde{y})$ has $\Omega(r)$ singular values of magnitude $\geq 1/n^2$. Then, an $\rho$-perturbed vector $\widetilde{x}$ has at least $\rho/n$ of its norm in the space spanned by the corresponding right singular vectors, with probability $\geq 1 - \exp(-r)$ (Fact 3.26). Thus we get

$$\mathbf{Pr}[\|Q(\widetilde{y})\widetilde{x}\| \geq \rho/n^3] \geq 1 - \exp(-r).$$

So the key is to prove that the matrix $Q(\widetilde{y})$ has a large number of "non-negligible" singular values with high probability (over the perturbation in $\widetilde{y}$). For this, let us examine the entries of $Q(\widetilde{y})$. For a moment suppose that $\widetilde{y}$ is a gaussian random vector $\sim \mathcal{N}(0, \rho^2 I)$ (instead of a perturbation). Then the $(i,j)$th entry of $Q(\widetilde{y})$ is precisely $\langle \widetilde{y}, (M_i)_j \rangle$, which is distributed like a one dimensional gaussian of variance $\rho^2 \|(M_i)_j\|^2$. If the entries for different $i, j$ were independent, standard results from random matrix theory would imply that $Q(\widetilde{y})$ has many non-negligible singular values.

However, this could be far from the truth. Consider, for instance, two vectors $(M_i)_j$ and $(M_{i'})_{j'}$ that are parallel. Then their dot products with $\widetilde{y}$ are highly correlated. However we note, that as long as $(M_{i'})_{j'}$ has a reasonable component orthogonal to $(M_i)_j$, the distribution of the $(i,j)$ and $(i',j')$th entries are "somewhat" independent. We will prove that we can roughly achieve such a situation. This motivates the following definition.

**Definition 3.13.** [Ordered $\theta$-orthogonality] A sequence of vectors $v_1, v_2, \ldots, v_n$ has the ordered $\theta$-orthogonality property if for all $1 \leq i \leq n$, $v_i$ has a component of length $\geq \theta$ orthogonal to $\mathrm{span}\{v_1, v_2, \ldots, v_{i-1}\}$.

Now we define a similar notion for a sequence of matrices $M_1, M_2, \ldots, M_r$, which says that a large enough subset of columns should have a certain $\theta$-orthogonality property. More formally,

**Definition 3.14** (Ordered $(\theta, \delta)$-orthogonal system)**.** A set of $n \times m$ matrices $M_1, M_2, \ldots, M_r$ form an *ordered $(\theta, \delta)$-orthogonal system* if there exists a permutation $\pi$ on $[m]$ such that the first $\delta m$ columns satisfy the followng property: for $i \leq \delta m$ and every $j \in [R]$, the $\pi(i)$th column of $M_j$ has a projection of length $\geq \theta$ orthogonal to the span of all the vectors given by the columns $\pi(1), \pi(2), \ldots, \pi(i-1), \pi(i)$ of all the matrices $M_1, M_2, \ldots M_r$ other than itself (i.e. the $\pi(i)$th column of $M_j$).

The following lemma shows the use of an ordered $(\theta, \delta)$ orthogonal system: a matrix $Q(\widetilde{y})$ constructed as above starting with these $M_i$ has many non-negligible singular values with high probability.

**Lemma 3.15** (Ordered $\theta$-orthogonality and perturbed combinations.)**.** *Let $M_1, M_2, \ldots, M_r$ be a set of $n \times m$ matrices of bounded norm ($\|\cdot\|_F \geq 1$) that are $(\theta, \delta)$ orthogonal for some parameters $\theta, \delta$, and suppose $r \leq \delta m$. Let $\widetilde{x}$ be an $\rho$-perturbation of $x \in \mathbb{R}^n$. Then the $r \times m$ matrix $Q(\widetilde{x})$ formed with the $j$th row of $(Q(\widetilde{x}))_j$ being $\widetilde{x}^T M_j$ satisfies*

$$\mathop{\mathbf{Pr}}_{x}\left[\sigma_{r/2}\left(Q(\widetilde{x})\right) \geq \frac{\rho\theta}{n^4}\right] \geq 1 - \exp\left(-r\right)$$

We defer the proof of this Lemma to section 3.3. Our focus will now be on constructing such a $(\theta, \delta)$ orthogonal system of matrices, given a subspace $\mathcal{V}$ of $\mathbb{R}^{n^2}$ of dimension $\Omega(n^2)$. The following lemma achieves this

**Lemma 3.16.** *Let $\mathcal{V}$ be a $\delta \cdot nm$ dimensional subspace $\mathbb{R}^{nm}$, and suppose $r, \theta, \delta'$ satisfy $\delta' \leq \delta/2$, $r \cdot \delta' m < \delta n/2$ and $\theta = 1/(nm^{3/2})$. Then there exist $r$ matrices $M_1, M_2, \ldots, M_r$ of dimension $n \times m$ with the following properties*

1. *$vec(M_i) \in \mathcal{V}$ for all $i \in [r]$.*
2. *$M_1, M_2, \ldots, M_r$ form an ordered $(\theta, \delta')$ orthogonal system.*

*In particular, when $m \leq \sqrt{n}$, they form an ordered $(\theta, \delta/2)$ orthogonal system.*

We remark that while $\delta$ is often a constant in our applications, $\delta'$ does not have to be. We will use this in the proof that follows, in which we use these above two lemmas regarding construction and use of an ordered $(\theta, \delta)$-orthogonal system to prove Proposition 3.12.

**Proof of Proposition 3.12** The proof follows by combining Lemma 3.16 and Lemma 3.15 in a fairly straightforward way. Let $\mathcal{U}$ be the span of the top $\delta n^2$ singular values of $M$. Thus $\mathcal{U}$ is a $\delta n^2$ dimensional subspace of $\mathbb{R}^{n^2}$. It has three steps:

1. We use Lemma 3.16 with $m = n, \delta' = \frac{\delta}{n^{1/2}}, \theta = \frac{1}{n^{5/2}}$ to obtain $r = \frac{n^{1/2}}{2}$ matrices $M_1, M_2, \ldots, M_r \in \mathbb{R}^{n \times n}$ having the $(\theta, \delta')$-orthogonality property.
2. Now, applying Lemma 3.15, we have that the matrix $Q(\widetilde{x})$, defined as before, (given by linear combinations along $\widetilde{x}$) , has $\sigma_{r/2}\left(Q(\widetilde{x})\right) \geq \frac{\rho\theta}{n^4}$ w.p $1 - \exp(-\sqrt{n})$.
3. Applying Fact 3.26 along with a simple averaging argument, we have that for one of the terms $M_i$, we have $|M_i(\widetilde{x} \otimes \widetilde{y})| \geq \rho\theta/n^6$ with probability $\geq 1 - \exp(-r/2)$ as required.

Please refer to Appendix B.2 for the complete details. □

The proof for higher order tensors will proceed along similar lines. However we require an additional pre-processing step and a careful inductive statement (Theorem 3.25), whose proof invokes Lemmas 3.16 and 3.15. The issues and details with higher order products are covered in Section 3.4. The following two sections are devoted to proving the two lemmas i.e. Lemma 3.16 and Lemma 3.15. These will be key to the general case ($\ell > 2$) as well.

### 3.2 Constructing the $(\theta, \delta)$-Orthogonal System (Proof of Lemma 3.16)

Recollect that $\mathcal{V}$ is a subspace of $\mathbb{R}^{n \cdot m}$ of dimension $\delta nm$ in Lemma 3.16. We will also treat a vector $M \in \mathcal{V}$ as a matrix of size $n \times m$, with its co-ordinates indexed by $[n] \times [m]$.

We want to construct many matrices $M_1, M_2, \ldots M_r \in \mathbb{R}^{n \times m}$ such that a reasonable fraction of the $m$ columns satisfy $\theta$-orthogonality property. Intuitively, such columns would have $\Omega(n)$ independent directions in $\mathbb{R}^n$, as choices for the $r$ matrices $M_1, M_2, \ldots, M_r$. Hence, we need to identify columns $i \in [m]$, such that the projection of $\mathcal{V}$ onto these $n$ co-ordinates (in column $i$) spans a large dimension, in a robust sense. This notion is formalized by defining the robust dimension of column projections, as follows.

**Definition 3.17** (Robust Dimension of projections)**.** For a subspace $\mathcal{V}$ of $\mathbb{R}^{n \cdot m}$, we define its robust dimension $\dim_i^\tau(\mathcal{V})$ to be

$$\dim_i^\tau(\mathcal{V}) = \max_d \text{ s.t. } \exists \text{ orthonormal } v_1, v_2, \ldots, v_d \in \mathbb{R}^n \text{ and } M_1, M_2, \ldots, M_d \in \mathcal{V}$$
$$\text{with } \forall t \in [d], \|M_t\| \leq \tau \text{ and } v_t = M_t(i).$$

This definition ensures that we do not take into account those spurious directions in $\mathbb{R}^n$ that are covered to an insignificant extent by projecting (unit) vectors in $\mathcal{V}$ to the $i$th column. Now, we would like to use the large dimension of $\mathcal{V}$ (dim=$\delta nm$) to conclude that there are many columns projections having large robust dimensions of around $\delta n$ .

**Lemma 3.18.** *In any subspace $\mathcal{V}$ in $\mathbb{R}^{p_1 \cdot p_2}$ of dimension $\dim(\mathcal{V})$ for any $\tau \geq \sqrt{p_2}$, we have*

$$\sum_{i \in [p_2]} \dim_i^\tau(\mathcal{V}) \geq \dim(\mathcal{V}) \tag{6}$$

**Remark 3.19.** This lemma will also be used in the first step of the proof of Theorem 3.6 to identify a *good block* of co-ordinates which span a large projection of a given subspace $\mathcal{V}$.

The above lemma is easy to prove if the dimension of the column projections used is the usual dimension of a vector space. However, with robust dimension, to carefully avoid spurious or insignificant directions, we identify the robust dimension with the number of large singular values of a certain matrix.

**Proof:** Let $d = \dim(\mathcal{V})$. Let $B$ be a $(p_1 p_2) \times d$ matrix, with the $d$ columns comprising an orthonormal basis for $\mathcal{V}$. Clearly $\sigma_d(B) = 1$. Now, we split the matrix $B$ into $p_1$ blocks of size $p_1 \times d$ each. For $i \in [p_2]$, let $B_i \in \mathbb{R}^{p_1 \times d}$ be the projection of $B$ on the rows given by $[p_1] \times i$. Let $d_i = \max t$ such that $\sigma_t(B_i) \geq \frac{1}{\sqrt{p_2}}$.
We will first show that $\sum_i d_i \geq d$. Then we will show that $\dim_i^\tau(\mathcal{V}) \geq d_i$ to complete our proof.

Suppose for contradiction that $\sum_{i \in [p_2]} d_i < d$. Let $\mathcal{S}_i$ be the $(d - d_1)$-dimensional subspace of $\mathbb{R}^d$ spanned by the last $(d - d_1)$ right singular vectors of $B_i$. Hence,

$$\text{for unit vectors } \alpha \in \mathcal{S}_i \subseteq \mathbb{R}^d, \ \|B_i \alpha\| < \frac{1}{\sqrt{p_2}}.$$

Since, $d - \sum_{i \in [p_2]} d_i > 0$, there exists at least one unit vector $\alpha \in \bigcap_i \mathcal{S}_i^\perp$. Picking this unit vector $\alpha \in \mathbb{R}^d$, we can contradict $\sigma_d(B) \geq 1$

To establish the second part, consider the $d_i$ top left-singular vectors for matrix $B_i$ ($\in \mathbb{R}^{p_1}$) . These $d_i$ vectors can be expressed as small combinations ($\|\cdot\|_2 \leq \sqrt{p_2}$) of the columns of $B_i$ using

Lemma B.1. The corresponding $d_i$ small combinations of the columns of the whole matrix $B$, gives vectors in $\mathbb{R}^{p_1 p_2}$ which have length $\sqrt{p_2}$ as required (since column of $B$ are orthonormal). ■

We will construct the matrices $M_1, M_2, \ldots, M_r \in \mathbb{R}^{n \times m}$ in multiple stages. In each stage, we will focus on one column $i \in [m]$: we *fix* this column for all the matrices $M_1, M_2, \ldots, M_r$, so that this column satisfies the ordered $\theta$-orthogonal property w.r.t previously chosen columns, and then leave this column unchanged in the rest of the stages.

In each stage $t$ of this construction we will be looking at subspaces of $\mathcal{V}$ which are obtained by zero-ing out all the columns $J \subseteq [m]$ (i.e. all the co-ordinates $[n] \times J$), that we have fixed so far.

**Definition 3.20** (Subspace Projections)**.** For $J \subseteq [m]$, let $\mathcal{V}^{*J} \subseteq \mathbb{R}^{n \cdot (m - |J|)}$ represent the subspace obtained by projecting on to the co-ordinates $[n] \times ([m] - J)$, the subspace of $\mathcal{V}$ having zeros on all the co-ordinates $[n] \times J$.

$$\mathcal{V}^{*J} = \left\{ M' \in \mathbb{R}^{n \cdot (m-|J|)} : \exists M \in \mathcal{V} \text{ s.t. columns } M(i) = M'(i) \text{ for } i \in [m] - J, \text{ and 0 otherwise } . \right\}$$

The extension $\mathrm{Ext}_{*J}(M')$ for $M' \in \mathcal{V}^{*J}$ is the vector $M \in \mathcal{V}$ obtained by padding $M'$ with zeros in the coordinates $[n] \times J$ (columns given by $J$).

The following lemma shows that their dimension remains large as long as $|J|$ is not too large:

**Lemma 3.21.** *For any $J \subseteq [m]$ and any subspace $\mathcal{V}$ of $\mathbb{R}^{n \cdot m}$ of dimension $\delta \cdot nm$, the subspace having zeros in the co-ordinates $[n] \times J$ has $\dim\left(\mathcal{V}^{*J}\right) \geq n(\delta m - |J|)$.*

**Proof of Lemma 3.21:** Consider a constraint matrix $C$ of size $(1-\delta)nm \times nm$ which describes $\mathcal{V}$. $\mathcal{V}^{*J}$ is described by the constraint matrix of size $(1-\delta)nm \times n(m-|J|)$ obtained by removing the columns of $C$ corresponding to $[n] \times J$. Hence we get a subspace of dimension at least $n(m - |J|) - (1-\delta)nm$. ■

We now describe the construction more formally.

---

**The Iterative Construction of ordered $\theta$-orthogonal matrices.**

Initially set $J_0 = \emptyset$ and $M_j = 0$ for all $j \in [r]$, $\tau = \sqrt{m}$ and $s = \delta m/2$.
For $t = 1 \ldots s$,

1. Pick $i \in [m] - J_{t-1}$ such that $\dim_i^{\tau}\left(\mathcal{V}^{*J_{t-1}}\right) \geq \delta n/2$. If no such $i$ exists, report FAIL.
2. Choose $Z_1, Z_2, \ldots, Z_r \in \mathcal{V}^{*J_{t-1}}$ of length at most $\sqrt{mn}$ such that $i$th columns $Z_1(i), Z_2(i), \ldots, Z_r(i) \in \mathbb{R}^n$ are orthonormal, and also orthogonal to the columns $\{M_j(i')\}_{i' \in J_{t-1}, j \in [r]}$. If this is not possible, report FAIL.
3. Set for all $j \in [r]$, the new $M_j \leftarrow M_j + \mathrm{Ext}_{*J}(Z_j)$, where $\mathrm{Ext}_{*J}(Z_j)$ is the matrix padded with zeros in the columns corresponding to $J$. Set $J_t \leftarrow J_{t-1} \cup \{i\}$.

---

Let $J = J_s$ for convenience. We first show that the above process for constructing $M_1, M_2, \ldots, M_r$ completes successfully without reporting FAIL.

**Claim 3.22.** *For $r, s$ such that $s \leq \delta m/2$ and $r \cdot s \leq \delta n/3$, the above process does not FAIL.*

**Proof:** In each stage, we add one column index to $J$. Hence, $|J_t| \leq s$ at all times $t \in [s]$.

We first show that Step 1 of each iteration does not FAIL. From Lemma 3.21, we have $\dim\left(\mathcal{V}^{*J_t}\right) \geq \delta nm/2$. Let $\mathcal{W} = \mathcal{V}^{*J_t}$. Now, applying Lemma 3.18 to $\mathcal{W}$, we see that there exists $i \in [m] - J_t$ such that $\dim_i^\tau(\mathcal{W}) \geq \delta n/2$, as required. Hence, Step 1 does not fail.

$\dim_i^\tau(\mathcal{W}) \geq \delta n/2$ shows that there exist $Z'_1, Z'_2, \ldots Z'_{\delta n/2}$ with lengths at most $\sqrt{m}$ such that their $i$th columns $\{Z'_t(i)\}_{t \leq \delta n/2}$ are orthonormal. However, we additionally need to impose that the $i$th columns to also be orthogonal to the columns $\{M_j(i')\}_{j \in [r], i' \in J_{t-1}}$. Fortunately, the number of such orthogonality constraints is at most $r|J_{t-1}| \leq \delta n/3$. Hence, we can pick the $r < \delta n/6$ orthonormal $i$th columns $\{Z_j(i)\}_{j \in [r]}$ and their respective extensions $Z_j$, by taking linear combinations of $Z'_t$. Since the linear combinations result again in unit vectors in the $i$th column, the length of $Z_j \leq \sqrt{mn}$, as required. Hence, Step 2 does not FAIL as well. ■

**Completing the proof of Lemma 3.16.** We now show that since the process completes, then $M_1, M_2, \ldots, M_r$ have the required ordered $(\theta, \delta')$-orthogonal property for $\delta' = s/m$. We first check that $M_1, M_2, \ldots, M_r$ belong to $\mathcal{V}$. This is true because in each stage, $\mathrm{Ext}_{*J}(Z_j) \in \mathcal{V}$, and hence $M_j \in \mathcal{V}$ for $j \in [r]$. Further, since we run for $s$ stages, and each of the $Z_j$ are bounded in length by $\sqrt{mn}$, $\|M_j\|_F \leq s\sqrt{mn} \leq \sqrt{nm^3}$. Our final matrices $M_j$ will be scaled to $\|\cdot\|_F = 1$. The $s$ columns that satisfy the ordered $\theta$-orthogonality property are those of $J$, in the order they were chosen (we set this order to be $\pi$, and select an arbitrary order for the rest).

Suppose the column $i_t \in [m]$ was chosen at stage $t$. The key invariant of the process is that once a column $i_t$ is chosen at stage $t$, the $i_t^{th}$ column remains unchanged for each $M_j$ in all subsequent stages ($t+1$ onwards). By the construction, $Z_j(i_t) \in \mathbb{R}^n$ is orthogonal to $\{M_j(i)\}_{i \in J_{t-1}}$. Since $Z_j(i_t)$ has unit length and $M_j$ is of bounded length, we have the ordered $\theta$-orthogonal property as required, for $\theta = 1/\sqrt{nm^3}$. This concludes the proof. □

### 3.3 $(\theta, \delta)$-Orthogonality and $\rho$-Perturbed Combinations (Proof of Lemma 3.15)

Suppose $M_1, M_2, \ldots, M_r$ be a $(\theta, \delta)$-orthogonal set of matrices (dimensions $n \times m$). Without loss of generality, suppose that the permutation $\pi$ in the definition of orthogonality is the identity, and let $I$ be the first $\delta m$ columns.

Now let us consider an $\rho$-perturbed vector $\widetilde{x}$, and consider the matrix $Q(\widetilde{x})$ defined in the statement – it has dimensions $r \times m$, and the $(i,j)$th entry is $\langle \widetilde{x}, (M_i)_j \rangle$, which is distributed as a translated gaussian. Now for any column $i \in I$, the $i$th column in $Q(\widetilde{x})$ has every entry having an $(\rho \cdot \theta)$ 'component' independent of entries in the previous columns, and entries above. This implies that for a unit gaussian vector $g$, we have (by anti-concentration and $\theta$-orthogonality that

$$Pr[(g^T Q(\widetilde{x})_i)^2 < \theta^2/4n] < 1/2n. \tag{7}$$

Furthermore, the above inequality holds, even *conditioned* on the first $(i-1)$ columns of $Q(\widetilde{x})$.

**Lemma 3.23.** *Let $Q(\widetilde{x})$ be defined above, and fix some $i \in I$. Then for $g \sim \mathcal{N}(0,1)^n$, we have*

$$\mathbf{Pr}[(g^T Q(\widetilde{x})_i)^2 < \frac{\theta^2\rho^2}{4n^2} \mid Q(\widetilde{x})_1, \ldots, Q(\widetilde{x})_{(i-1)}] < \frac{1}{2n},$$

*for any given $Q(\widetilde{x})_1, Q(\widetilde{x})_2, \ldots, Q(\widetilde{x})_{(i-1)}$.*

**Proof:** Let $g = (g_1, g_2, \ldots, g_r)$. Then we have

$$\begin{aligned} g^T Q_i(\widetilde{x}) &= g_1(\widetilde{x}^T (M_1)_i) + g_2(\widetilde{x}^T (M_2)_i) + \cdots + g_r(\widetilde{x}^T (M_r)_i) \\ &= \langle \widetilde{x}, g_1(M_1)_i + g_2(M_2)_i + \ldots g_r(M_r)_i \rangle \end{aligned}$$

Let us denote the latter vector by $v_i$ for now, so we are interested in $\langle \widetilde{x}, v_i \rangle$. We show that $v_i$ has a non-negligible component orthogonal to the span of $v_1, v_2, \ldots v_{(i-1)}$. Let $\Pi$ be the matrix which projects orthogonal to the span of $(M_s)_{i'}$ for all $i' < i$. Thus any vector $\Pi u$ is also orthogonal to the span of $v_{i'}$ for $i' < i$.

Now by hypothesis, every vector $\Pi(M_s)_i$ has length $\geq \theta$. Thus the vector $\Pi\left(\sum_s g_s (M_s)_i\right) = \Pi v_i$ has length $\geq \theta/2$ with probability $\geq 1 - \exp(-r)$ (Lemma B.2).

Thus if we consider the distribution of $\langle \widetilde{x}, v_i \rangle = \langle x, v_i \rangle + \langle e, v_i \rangle$, it is a one-dimensional gaussian with mean $\langle x, v_i \rangle$ and variance $\rho^2$. From basic anti-concentration properties of a gaussian (that the mass in any $\rho \cdot (\text{variance})^{1/2}$ interval is at most $\rho$), the conclusion follows. ■

We can now do this for all $i \in I$, and conclude that the probability that Eq. (7) holds *for all* $i \in I$ is at most $1/(2n)^{|I|}$.

Now what does this imply about the singular values of $Q(\widetilde{x})$? Suppose it has $< r/2$ (which is $< |I|$) non-negligible singular values, then a gaussian random vector $g$, with probability *at least* $n^{-r}$, has a negligible component along *all* the corresponding singular vectors, and thus the length of $g^T Q(\widetilde{x})$ is negligible with at least this probability!

**Lemma 3.24.** *Let $M$ be a $t \times t$ matrix with spectral norm $\leq 1$. Suppose $M$ has at most $r$ singular values of magnitude $> \tau$. Then for $g \sim \mathcal{N}(0,1)^t$, we have*

$$\mathbf{Pr}[\|Mg\|_2^2 < 4t\tau^2 + \frac{t}{n^{2c}}] \geq \frac{1}{n^{cr}} - \frac{1}{2^t}.$$

**Proof:** Let $u_1, u_2, \ldots, u_r$ be the singular vectors corresponding to value $> \tau$. Consider the event that $g$ has a projection of length $< 1/n^c$ onto $u_1, u_2, \ldots, u_r$. This has probability $\geq \frac{1}{n^{cr}}$, by anti-concentration properties of the Gaussian (and because $\mathcal{N}(0,1)^t$ is rotationally invariant). For any such $g$, we have

$$\begin{aligned} \|Mg\|_2^2 &= \sum_{i=1}^{r} \langle g, u_i \rangle^2 + \tau^2 \|g\|^2 \\ &\leq \frac{r}{n^{2c}} + \tau^2 \|g\|_2^2. \end{aligned}$$

■

This contradicts the earlier anti-concentration bound, and so we conclude that the matrix has at least $r/2$ non-negligible singular values, as required.

### 3.4 Higher Order Products

We have a subspace $\mathcal{V} \in \mathbb{R}^{n^\ell}$ of dimension $\delta n^\ell$. The proof for higher order products proceeds by induction on the order $\ell$ of the product. Recall from Remark 3.8 that Proposition 3.12 and Theorem 3.3 do not get good guarantees for small values of $\delta$, like $1/n$. In fact, we can not hope to get such exponentially small failure probability in that case, since the all the $n$ degrees of freedom in $\mathcal{V}$ may be constrained to the first $n$ co-ordinates of $\mathbb{R}^{n^2}$ (all the independence is in just one mode). Here, it is easy to see that the best we can hope for is an inverse-polynomial failure probability. Hence, to get exponentially small failure probability, we will always need $\mathcal{V}$ to have a large dimension compared to the dimension of the host space in our inductive statements.

To carry out the induction, we will try to reduce this to a statement about $\ell - 1$ order products, by taking linear combinations (given by $\widetilde{x}^{(1)} \in \mathbb{R}^n$) along one of the modes. Loosely speaking,

Lemma 3.15 serves this function of "order reduction", however it needs a set of $r$ matrices in $\mathbb{R}^{n\times m}$ (flattened along all the other modes) which are ordered $(\theta,\delta)$ orthogonal.

Let us consider the case when $\ell = 3$, to illustrate some of the issues that arise. We can use Lemma 3.16 to come up with $r$ matrices in $\mathbb{R}^{n\times n^2}$ that are ordered $(\theta,\delta)$ orthogonal. These columns intuitively correspond to independent directions or degrees of freedom, that we can hope to get substantial projections on. However, since these are vectors in $\mathbb{R}^n$, the number of "flattened columns" can not be comparable to $n^2$ (in fact, $\delta m \ll n$) — hence, our induction hypothesis for $\ell = 2$ will give no guarantees, (due to Remark 3.8).

To handle this issue, we will first restrict our attention to a smaller block of co-ordinates of size $n_1 \times n_2 \times n_3$ (with $n_1 n_2 n_3 \ll n$) , that has reasonable size in all the three modes $(n_1, n_2, n_3 = n^{\Omega(1)})$. Additionally, we want $\mathcal{V}$'s projection onto this $n_1 \times n_2 \times n_3$ block spans a large subspace of (robust) dimension at least $\delta n_1 n_2 n_3$ (using Lemma 3.18).

Moreover, choosing the main inductive statement also needs to be done carefully. We need some property for choosing enough candidate "independent" directions $T_1, T_2, \ldots T_r \in \mathbb{R}^{n^\ell}$ (projected on the chosen block), such that our process of "order reduction" (by first finding $\theta$-orthogonal system and then combining along $\widetilde{x}^{(1)}$) maintains this property for order $\ell - 1$. This is where the alternate interpretation in Theorem 3.9 in terms of singular values helps: it suggests the exact property that we need! We ensure that the matrix formed by the flattened vectors $\mathrm{vec}(T_1), \mathrm{vec}(T_2), \ldots \mathrm{vec}(T_r)$ (projected onto the $n_1 \times n_2 \times n_3$ block) , as rows form a matrix with many large singular values.

We now state the main inductive claim. The claim assumes a block of co-ordinates of reasonable size in each mode that span many directions in $\mathcal{V}$, and then establishes the anti-concentration bound inductively.

**Theorem 3.25** (Main Inductive Claim)**.** *Let $T_1, T_2, \ldots, T_r \in \mathbb{R}^{n^{\times \ell}}$ be $r$ tensors with bounded norm ($\|\cdot\|_F \le 1$) and $I_1, I_2, \ldots I_\ell \subseteq [n]$ be sets of indices of sizes $n_1, n_2, \ldots n_\ell$. Let $T$ be the $r \times n^\ell$ matrix obtained with rows $vec(T_1), vec(T_2), \ldots, vec(T_r)$. Suppose*

- *$\forall j \in [r], P_j$ is $T_j$ restricted to the block $I_1 \times \cdots \times I_\ell$, and matrix $P \in \mathbb{R}^{r\times(n_1\cdot n_2 \ldots n_\ell)}$ has $j$th row as $vec(P_j)$,*
- *$r \ge \delta_\ell n_1 n_2 \ldots n_\ell$ and $\forall t \in [\ell-1], n_t \ge (n_{t+1} n_{t+2} \ldots n_\ell)^2$,*
- *$\sigma_r(P) \ge \eta$.*

*Then for random $\rho$-perturbations $\widetilde{x}^{(1)}, \widetilde{x}^{(2)} \ldots \widetilde{x}^{(\ell)}$ of any $x^{(1)}, x^{(2)} \ldots x^{(\ell)} \in \mathbb{R}^n$, we have*

$$\Pr_{\widetilde{x}^{(1)},\ldots\widetilde{x}^{(\ell)}}\left[\|T\left(\widetilde{x}^{(1)} \otimes \cdots \otimes \widetilde{x}^{(\ell)}\right)\| \ge \rho^\ell \left(\frac{\eta}{n_1}\right)^{3^\ell}\right] \ge 1 - \exp\left(-\delta_\ell n_\ell\right)$$

Before we give a proof of the main inductive claim, we first present a standard fact that relates the singular value of matrices and some anti-concentration properties of randomly perturbed vectors. This will also establish the base case of our main inductive claim.

**Fact 3.26.** *Let $M$ be a matrix of size $m \times n$ with $\sigma_r(M) \ge \eta$. Then for any unit vector $u \in \mathbb{R}^n$ and an random $\rho$-perturbation $\widetilde{x}$ of it, we have*

$$\|M\widetilde{x}\|_2 \ge \eta\rho/n^2 \quad w.p \quad 1 - n^{-\Omega(r)}$$

**Proof of Theorem 3.25:** The proof proceeds by induction. The base case ($\ell = 1$) is handled by Fact 3.26. Let us assume the theorem for $(\ell-1)$-wise products. The inductive proof will have two main steps:

1. Suppose we flatten the tensors $\{P_j\}_{j\in[r]}$ along all but the first mode, and imagine them as matrices of size $n_1 \times (n_2 n_3 \dots n_\ell)$. We can use Lemma 3.16 to construct ordered $(\theta, \delta')$ orthogonal system w.r.t vectors in $\mathbb{R}^{n_1}$ (columns correspond to $[m] = [n_2 \dots n_\ell]$).

2. When we take combinations along $\widetilde{x}^{(1)}$ as $T\left(\widetilde{x}^{(1)}, \cdot, \cdot, \dots, \cdot\right)$, these tensors will now satisfy the condition required for $(\ell-1)$-order products in the inductive hypothesis, because of Lemma 3.15.

Unrolling this induction allows us to take combinations along $\widetilde{x}^{(1)}, \widetilde{x}^{(2)}, \dots$ as required, until we are left with the base case. For notational convenience, let $y = \widetilde{x}^{(1)}$, $\delta_\ell = \delta$, $r_\ell = r$ and $N = n_1 n_2 \dots n_\ell$.

To carry out the first step, we think of $\{P_j\}_{j\in[r]}$ as matrices of size $n_1 \times (n_2 n_3 \dots n_\ell)$. We then apply Lemma 3.16 with $n = n_1$, $m = \frac{N}{n_1} = n_2 n_3 \dots n_\ell \le \sqrt{n_1}$ ; hence there exists $r'_\ell = n_2 \dots n_\ell$ matrices $\{Q_q\}_{q\in[r'_\ell]}$ with $\|\cdot\|_F \le 1$ which are ordered $(\theta, \delta'_\ell)$-orthogonal for $\delta'_\ell = \delta_\ell/3$. Further, since $Q_q$ are in the row-span of $P$, there exists matrix of coefficients $\alpha = (\alpha(q,j))_{q\in[r'_\ell], j\in[r_\ell]}$ such that

$$\forall q \in [r'_\ell],\ Q_q = \sum_{j\in[r_\ell]} \alpha(q,j) P_j \tag{8}$$

$$\|\alpha(q)\|_2^2 = \sum_{j\in[r_\ell]} \alpha(q,j)^2 \le 1/\eta \quad \text{(since } \sigma_r(P) \ge \eta \text{ and } \|Q_q\|_F \le 1) \tag{9}$$

Further, $Q_q$ is the projection of $\sum_{j\in[r_\ell]} \alpha_{q,j} T_j$ onto co-ordinates $I_1 \times I_2 \dots \times I_\ell$. Suppose we define a new set of matrices$\{W_q\}_{q\in[r'_\ell]}$ in $\mathbb{R}^{n\times(\frac{N}{n_1})}$ by flattening the following into a matrix with $n$ rows:

$$W_q = \left(\sum_{j\in[r]} \alpha_{q,j} T_j\right)_{[n]\times(I_2\times\dots\times I_\ell)}.$$

In other words, $Q_q$ is obtained by projecting $W_q$ on to the $n_1$ rows given by $I_1$. Note that $\{W_q\}_{q\in[r'_\ell]}$ is also ordered $(\theta'_\ell, \delta'_\ell)$ orthogonal for $\theta'_\ell = \theta\eta$.

To carry out the second part, we apply Lemma 3.15 with $\{W_q\}$ and infer that the $r'_\ell \times (N/n_1)$ matrix $W(y)$ with $q$th row being $y^T W_q$ has $\sigma_{r_{\ell-1}}(W(y)) \ge \eta'_\ell = \theta^2\rho^2/n_1^4$ with probability $1 - \exp(-\Omega(r'_\ell))$, where $r_{\ell-1} = r'_\ell/2$.
We will like to apply the inductive hypothesis for $(\ell-1)$ with $P$ being $W(y)$; however $W(y)$ *does not* have full (robust) row rank. Hence we will consider the top $r_{\ell-1}$ right singular vectors of $W(y)$ to construct an $r_{\ell-1}$ tensors of order $\ell$, whose projections to the block $I_2 \times \dots \times I_\ell$, lead to a well-conditioned $r_{\ell-1} \times (n_2 n_3 \dots n_\ell)$ matrix for which our inductive hypothesis holds.

Let the top $r_{\ell-1}$ right singular vectors of $W(y)$ be $Z_1, Z_2, \dots Z_{r_{\ell-1}}$. Hence, from Lemma B.1, we have a coefficient $\beta$ of size $r_{\ell-1} \times r_\ell$ such that

$$\forall j' \in [r_{\ell-1}] \quad Z_{j'} = \sum_{q\in[r'_\ell]} \beta_{j',q} W_q(y) \text{ and } \|\beta(j')\|_2 \le 1/\eta'_\ell.$$

Now let us try to represent these new vectors in terms of the original row-vectors of $P$, to construct the required tensor of order $(\ell-1)$ . Consider the $r_{\ell-1} \times r_\ell$ matrix $\Lambda = \beta\alpha$. Clearly,

$$\text{rownorm}(\Lambda) \le \text{rownorm}(\beta) \cdot \|\alpha\|_F \le \sqrt{r'_\ell} \cdot \text{rownorm}(\beta) \cdot \text{rownorm}(\alpha) \le \frac{r'_\ell}{\eta_\ell \eta'_\ell}.$$

Define $\forall j' \in [r_{\ell-1}]$, an order $\ell$ tensor $T'_{j'} = \sum_{j\in[r]} \lambda_{j',j} T_j$; from the previous equation, $\|T'_{j'}\|_F \leq r'_\ell/(\eta_\ell \eta'_\ell)$ . We need to get a normalized order $(\ell-1)$ tensor: so, we consider $\widehat{T}_{j'} = T'_{j'}/\|T'_{j'}(y)\|_F$, and $\widehat{T}$ be the $r_{\ell-1} \times (n^\ell)$ matrix with $j'$th row being $\widehat{T}_{j'}$. Hence,

$$\sigma_{r_{\ell-1}}\left(\widehat{T}(y,\cdot,\cdot,\ldots,\cdot)\right) \geq \frac{\eta_\ell^3}{r'_\ell n_1^3}.$$

We also have $r_{\ell-1} \geq \frac{1}{2}\cdot n_2 n_3 \ldots n_\ell$. By the inductive hypothesis

$$\|\widehat{T}\left(y, \widetilde{x}^{(2)}, \ldots, \widetilde{x}^{(\ell)}\right)\| \geq \eta' \equiv \rho^{\ell-1}\left(\frac{\eta_\ell^3}{n_1^4 n_2}\right)^{3^{\ell-1}} \quad \text{w.p } 1-\exp\left(-\Omega(n_\ell)\right) \tag{10}$$

Hence, for one of the $j' \in [r_{\ell-1}]$, $\left|\widehat{T}_{j'}\left(\widetilde{x}^{(1)}, \widetilde{x}^{(2)}, \ldots \widetilde{x}^{(\ell)}\right)\right| \geq \eta'/\sqrt{r_{\ell-1}}$. Finally, since $\widehat{T}_{j'}$ is given by a small combination of the $\{T_j\}_{j\in[r]}$, we have from Cauchy-Schwartz

$$\|T\left(\widetilde{x}^{(1)}, \widetilde{x}^{(1)}, \ldots, \widetilde{x}^{(1)}\right)\| \geq \eta' \cdot \left(\frac{\eta^3}{\sqrt{r_\ell^2 n_1^4}}\right).$$

■

The main required theorem now follows by just showing the exists of the $n_1 \times n_2 \times \cdots \times n_\ell$ block that satisfies the theorem conditions. This follows from Lemma 3.18.

**Proof of Theorem 3.6:** First we set $n_1, n_2, n_\ell$ by the recurrence $\forall t \in [\ell]$, $n_t = 2(n_{t+1}\cdot n_{t+2}\ldots n_\ell)^2$ and $n_1 = O(n)$. It is easy to see that this is possible for $n_\ell = n^{1/3^\ell}$. Now, we partition the set of co-ordinates $[n]^\ell$ into blocks of size $n_1 \times n_2 \times \ldots n_\ell$. Let $p_1 = n_1 \cdot n_2 \ldots n_\ell$ and $p_2 = n^\ell/p_1$. Applying Lemma 3.18 we see that there exists indices $I_1, I_2, \ldots I_\ell$ of sizes $n_1, n_2, \ldots, n_\ell$ respectively such that projection $\mathcal{W} = \mathcal{V}_{|I_1\times I_2 \times \cdots \times I_\ell}$ on this block of co-ordinates has dimension $\dim_I^\tau(\mathcal{W}) \geq n_1 n_2 \ldots n_\ell/4$. Let $r = n_1 n_2 \ldots n_\ell$. Now we construct $P'$ with the rows of $P'$ being an orthonormal basis for $\mathcal{W}$, and let $T'$ be the corresponding vectors in $\mathcal{V}$. Note that $\forall j \in [r]$, $\|T'_j\| \leq n^\ell$. Let $P$ be the re-scaling of the matrix so that for the $j$th row$(j \in [r])$, $P_j = P'_j/\|T'_j\|$ and $T_j = T'_j/\|T'_j\|$. Hence $\sigma_r(P) \geq 1/n^\ell$. Applying Theorem 3.25 with this choice of $P, T$, we get the required result. ■

# 4 Learning Multi-view Mixture Models

We now see how Theorem 1.5 immediately gives efficient learning algorithms for broad class of discrete mixture models called multi-view models in the *over-complete* setting. In a multi-view mixture model, for each sample we are given a few different observations or views $x^{(1)}, x^{(2)}, \ldots, x^{(\ell)}$ that are conditionally independent given which component $i \in [R]$ the sample is from. Typically, the $R$ components in the mixture are discrete distributions. Multi-view models are very expressive, and capture many well-studied models like Topic Models [2], Hidden Markov Models (HMMs) [29, 1, 2], and random graph mixtures [1]. They are also sometimes referred to as *finite mixtures of finite measure products*[1] or *mixture-learning with multiple snapshots* [30].

In this section, we will assume that each of the components in the mixture is a discrete distribution with support of size $n$. We first introduce some notation, along the lines of [2].

**Parameters and the model:** Let the $\ell$-view mixture model be parameterized by a set of $\ell$ vectors in $\mathbb{R}^n$ for each mixture component, $\left\{\mu_i{}^{(1)}, \mu_i{}^{(2)}, \ldots, \mu_i{}^{(\ell)}\right\}_{i\in[R]}$, and mixing weights $\{w_i\}_{i\in[R]}$ , that add up to 1. Each of these parameter vectors are normalized : in this work, we will assume that $\|\mu_i{}^{(j)}\|_1 = 1$ for all $i \in [R], j \in [\ell]$. Finally, for notational convenience we think of the parameters are represented by $n \times R$ matrices (one per view) $M^{(1)}, M^{(2)}, \ldots, M^{(\ell)}$, with $M^{(j)}$ formed by concatenating the vectors $\mu_i{}^{(j)}$ $(1 \le i \le R)$.

Samples from the multi-view model with $\ell$ views are generated as follows:

1. The mixture component $i$ $(i \in [R])$ is first picked with probability $w_i$
2. The views $x^{(1)}, \ldots, x^{(j)}, \ldots, x^{(\ell)}$ are indicator vectors in $n$-dimensions, that are drawn according to the distribution $\mu_i{}^{(1)}, \ldots, \mu_i{}^{(j)}, \ldots, \mu_i{}^{(\ell)}$.

The state-of-the-art algorithms for learning multi-view mixture models have guarantees that mirror those for mixtures of gaussians. In the worst case, the best known algorithms for this problem are from a recent work Rabani et al [30], who give an algorithm that has complexity $R^{O(R^2)} + \mathrm{poly}(n, R)$. In fact they also show a sample complexity lower-bound of $\exp(\tilde{\Omega}(R))$ for learning multi-view models in one dimension ($n = 1$). Polynomial time algorithms were given by Anandkumar et al. [2] in a restricted setting called the non-singular or non-degenerate setting. When each of these matrices $\left\{M^{(j)}\right\}_{j\in[\ell]}$ to have rank $R$ in a robust sense i.e. $\sigma_R(M^{(j)}) \ge 1/\tau$ for all $j \in [\ell]$, their algorithm runs in just $\mathrm{poly}(R, n, \tau, 1/\varepsilon))$ time to learn the parameters up to error $\varepsilon$. However, their algorithm fails even when $R = n + 1$.

However, in many practical settings like speech recognition and image classification, the dimension of the feature space is typically much smaller than the number of components or clusters i.e. $n \ll R$. To the best of our knowledge, there was no efficient algorithm for learning multi-view mixture models in such *over-complete settings*. We now show how Theorem 1.5 gives a polynomial time algorithm to learn multi-view mixture models in a smoothed sense, even in the over-complete setting $R \gg n$.

**Theorem 4.1.** *Let $(w_i, \mu_i{}^{(1)}, \ldots, \mu_i{}^{(\ell)})$ be a mixture of $R = O(n^{\ell/2-1})$ multi-view models with $\ell$ views, and suppose the means $\left(\mu_i{}^{(j)}\right)_{i\in[R],j\in[\ell]}$ are perturbed independently by gaussian noise of magnitude $\rho$. Then there is a polynomial time algorithm to learn the weights $w_i$, the perturbed parameter vectors $\left\{\tilde{\mu}_i^{(j)}\right\}_{j\in[\ell],i\in[R]}$ up to an accuracy $\varepsilon$ when given samples from this distribution. The running time and sample complexity is $\mathrm{poly}_\ell(n, 1/\rho, 1/\varepsilon)$.*

The conditional independence property is very useful in obtaining a higher order tensor, in terms of the hidden parameter vectors that we need to recover. This allows us to use our results on tensor decompositions from previous sections.

**Lemma 4.2** ([1])**.** *In the notation established above for multi-view models, $\forall \ell \in \mathbb{N}$ the $\ell^{th}$ moment tensor*

$$Mom_\ell = \mathbf{E}\left[x^{(1)} \otimes \ldots x^{(j)} \otimes \ldots x^{(\ell)}\right] = \sum_{r\in[R]} w_r \mu_r^{(1)} \otimes \mu_r^{(2)} \cdots \otimes \mu_r^{(j)} \otimes \cdots \otimes \mu_r^{(\ell)}. \tag{11}$$

Our algorithm to learn multi-view models consists of three steps:

1. Obtain a good empirical estimate $\widehat{T}$ of the order $\ell$ tensor $\mathrm{Mom}_\ell$ from $N = \mathrm{poly}_\ell(n, R, 1/\rho, 1/\varepsilon)$ samples (given by Lemma C.3)

$$\widehat{T} = \frac{1}{N}\sum_{t=1} x_t{}^{(1)} \otimes x_t{}^{(2)} \otimes \cdots \otimes x_t{}^{(\ell)}.$$

2. Apply Theorem 1.5 to $\widehat{T}$ and recover the parameters $\widehat{\mu}_i^{(j)}$ upto scaling.

3. Normalize the parameter vectors $\widehat{\mu}_i^{(j)}$ to having $\ell_1$ norm of 1, and hence figure out the weights $\widehat{w}_i$ for $i \in [R]$.

**Proof of Theorem 4.1:** The proof follows from a direct application of Theorem 1.5. Hence, we just sketch the details. We first obtain a good empirical estimate of $\mathrm{Mom}_\ell$ that is given in equation (11) using Lemma C.3. Applying Theorem 1.5 to $\widehat{T}$, we recover each rank-1 term in the decomposition $w_i {\mu_i}^{(1)} \otimes {\mu_i}^{(2)} \otimes \cdots \otimes {\mu_i}^{(\ell)}$ up to error $\varepsilon$ in frobenius norm ($\|\cdot\|_F$). However, we know that each of the parameter vectors are of unit $\ell_1$ norm. Hence, by scaling all the parameter vectors to unit $\ell_1$ norm, we obtain all the parameters up to the required accuracy. ■

# 5 Learning Mixtures of Axis-Aligned Gaussians

Let $F$ be a mixture of $k = \mathrm{poly}(n)$ axis-aligned Gaussians in $n$ dimensions, and suppose further that the means of the components are perturbed by Gaussian noise of magnitude $\rho$. We restrict to Gaussian noise not because our results change, but for notational convenience.

**Parameters:** The mixture is described by a set of $k$ mixing weights $w_i$, means $\mu_i$ and covariance matrices $\Sigma_i$. Since the mixture is axis-aligned, each covariance $\Sigma_i$ is diagonal and we will denote the $j^{th}$ diagonal of $\Sigma_i$ as $\sigma_{ij}^2$. Our main result in this section is the following:

**Theorem 5.1.** *Let $(w_i, \mu_i, \Sigma_i)$ be a mixture of $k = n^{\lfloor \frac{\ell-1}{2} \rfloor}/(2\ell)$ axis-aligned Gaussians and suppose $\{\widetilde{\mu}_i\}_{i\in[k]}$ are the $\rho$-perturbations of $\{\mu_i\}_{i\in[k]}$ (that have polynomially bounded length). Then there is a polynomial time algorithm to learn the parameters $(w_i, \widetilde{\mu}_i, \Sigma_i)_{i\in[k]}$ up to an accuracy $\varepsilon$ when given samples from this mixture. The running time and sample complexity is $\mathrm{poly}_\ell(\frac{n}{\rho\varepsilon})$.*

Next we outline the main steps in our learning algorithm:

1. We first pick an appropriate $\ell$, and estimate $\mathcal{M}_\ell := \sum_i w_i \widetilde{\mu}_i^{\otimes \ell}$.[3]

2. We run our decomposition algorithm for *overcomplete* tensors on $\mathcal{M}_\ell$ to recover $\widetilde{\mu}_i, w_i$.

3. We then set up a system of linear equations and solve for $\sigma_{ij}^2$.

We defer a precise description of the second and third steps to the next subsections (in particular, we need to describe how we obtain $\mathcal{M}_\ell$ from the moments of $F$ and we need to describe the linear system that we will use to solve for $\sigma_{ij}^2$).

## 5.1 Step 2: Recovering the Means and Mixing Weights

Our first goal in this subsection is to construct the tensor $\mathcal{M}_\ell$ defined above from random samples. In fact, if we are given many samples we can estimate a related tensor (and our error will be an inverse polynomial in the number of samples we take). Unlike the multi-view mixture model, we do not have $\ell$ independent views in this case. Let us consider the tensor $\mathbf{E}[x^{\otimes \ell}]$:

$$\mathbf{E}[x^{\otimes \ell}] = \sum_i w_i (\widetilde{\mu}_i + \eta_i)^{\otimes \ell}.$$

[3] We do not estimate the entire tensor, but only a relevant "block", as we will see.

Here we have used $\eta_i$ to denote a Gaussian random variable whose mean is zero and whose covariance is $\Sigma_i$. Now the first term in the expansion is the one we are interested in, so it would be nice if we could "zero out" the other terms. Our observation here is that if we restrict to $\ell$ distinct indices $(j_1, j_2, \ldots, j_\ell)$, then this coordinate will only have contribution from the means. To see this, note that the term of interest is

$$\sum_i \left[w_i \prod_{t=1}^{\ell} (\widetilde{\mu}_i(j_t) + \eta_i(j_t))\right]$$

Since the Gaussians are axis aligned, the $\eta_i(j_t)$ terms are independent for different $t$, and each is a random variable of zero expectation. Thus the term in the summation is precisely $\sum_i w_i \prod_{t=1}^{\ell} \widetilde{\mu}_i(j_t)$.

Our idea to estimate the means is now the following: we partition the indices $[n]$ into $\ell$ roughly equal parts $S_1, S_2, \ldots, S_\ell$, and estimate a tensor of dimension $|S_1| \times |S_2| \times \cdots \times |S_\ell|$.

**Definition 5.2** (Co-ordinate partitions)**.** Let $S_1, S_2, \ldots, S_\ell$ be a partition of $[n]$ into $\ell$ pieces of equal size (roughly). Let $\widetilde{\mu}_i^{(t)}$ denotes the vector $\widetilde{\mu}_i$ restricted to the coordinates $S_t$, and for a sample $x$, let $x^{(t)}$ denote its restriction to the coordinates $S_t$.

Now, we can estimate the order $\ell$ tensor $\mathbf{E}[x^{(1)} \otimes x^{(2)} \cdots \otimes x^{(\ell)}]$ to any inverse polynomial accuracy using polynomial samples (see Lemma C.3 or [20] for details), where

$$\mathbf{E}[x^{(1)} \otimes x^{(2)} \cdots \otimes x^{(\ell)}] = \sum_i w_i \big(\widetilde{\mu}_i^{(1)} \otimes \widetilde{\mu}_i^{(2)} \otimes \cdots \otimes \widetilde{\mu}_i^{(\ell)}\big).$$

Now applying the main tensor decomposition theorem (Theorem 1.5) to this order $\ell$ tensor, we obtain a set of vectors $\nu_i^{(1)}, \nu_i^{(2)}, \ldots, \nu_i^{(t)}$ such that

$$\nu_i^{(t)} = c_{it}\widetilde{\mu}_i^{(t)}, \text{ and for all } t,\ c_{i1}c_{i2}\cdots c_{i\ell} = 1/w_i.$$

Now we show how to recover the means $\widetilde{\mu}_i$ and weights $w_i$.

**Claim 5.3.** *The algorithm recovers the perturbed means $\{\widetilde{\mu}_i\}_{i\in[R]}$ and weights $w_i$ up to any accuracy $\varepsilon$ in time $\text{poly}_\ell(n, 1/\varepsilon)$*

So far, we have portions of the mean vectors, each scaled differently (upto some $\varepsilon/\text{poly}_\ell(n)$ accuracy. We need to estimate the scalars $c_{i1}, c_{i2}, \ldots, c_{i\ell}$ up to a scaling (we need another trick to then find $w_i$). To do this, the idea is to take a different partition of the indices $S_1', S_2', \ldots, S_\ell'$, and 'match' the coordinates to find the $\widetilde{\mu}_i$. In general, this is tricky since some portions of the vector may be zero, but this is another place where the perturbation in $\widetilde{\mu}_i$ turns out to be very useful (alternately, we can also apply a random basis change, and a more careful analysis to doing this 'match').

**Claim 5.4.** *Let $\mu$ be any $d$ dimensional vector. Then a coordinate-wise $\sigma$-perturbation of $\mu$ has length $\geq d\sigma^2/10$ w.p. $\geq 1 - \exp(-d)$.*

The proof is by a basic anti-concentration along with the observation that coordinates are independently perturbed and hence the failure probability multiplies.

Let us now define the partition $S_t'$. Suppose we divide $S_1$ and $S_2$ into two roughly equal parts each, and call the parts $A_1, B_1$ and $A_2, B_2$ (respectively). Now consider a partition with $S_1' = A_1 \cup A_2$ and $S_2' = B_1 \cup B_2$, and $S_t' = S_t$ for $t > 2$. Consider the solution $\nu_i'$ we obtain using the decomposition algorithm, and look at the vectors $\nu_1, \nu_2, \nu_1', \nu_2'$. For the sake of exposition, suppose we did not have any error in computing the decomposition. We can scale $\nu_1'$ such that

the sub-vector corresponding to $A_1$ is precisely equal to that in $\nu_1$. Now look at the remaining sub-vector of $\nu_1$, and suppose it is $\gamma$ times the "$A_2$ portion" of $\nu_2$. Then we must have $\gamma = c_2/c_1$.

To see this formally, let us fix some $i$ and write $v_{11}$ and $v_{12}$ to denote the sub-vectors of $\widetilde{\mu}_i^{(1)}$ restricted to coordinates in $A_1$ and $B_1$ respectively. Write $v_{21}$ and $v_{22}$ to represent sub-vectors of $\widetilde{\mu}_i^{(2)}$ restricted to $A_2$ and $B_2$ respectively. Then $\nu_1$ is $c_1 v_{11} \oplus c_1 v_{12}$ (where $\oplus$ denotes concatenation). So also $\nu_2$ is $c_2 v_{21} \oplus c_2 v_{22}$. Now we scaled $\nu_1'$ such that the $A_1$ portion agrees with $\nu_1$, thus we made $\nu_1'$ equal to $c_1 v_{11} \oplus c_1 v_{21}$. Thus by the way $\gamma$ is defined, we have $c_1\gamma = c_2$, which is what we claimed.

We can now compute the entire vector $\widetilde{\mu}_i$ up to scaling, since we know $c_1/c_2$, $c_1/c_3$, and so on. Thus it remains to find the mixture weights $w_i$. Note that these are all non-negative. Now from the decomposition, note that for each $i$, we can find the quantity

$$C_\ell := w_i \|\widetilde{\mu}_i\|^\ell.$$

The trick now is to note that by repeating the entire process above with $\ell$ replaced by $\ell + 1$, the conditions of the decomposition theorem still hold, and hence we compute

$$C_{\ell+1} := w_i \|\widetilde{\mu}_i\|^{\ell+1}.$$

Thus taking the ratio $C_{\ell+1}/C_\ell$ we obtain $\|\widetilde{\mu}_i\|$. This can be done for each $i$, and thus using $C_\ell$, we obtain $w_i$. This completes the analysis assuming we can obtain $\widetilde{\mu}_i^{(t)}$ without any error. Please see lemma C.4 for details on how to recover the weights $w_i$ in the presence of errors. This establishes the above claim about recovering the means and weights.

### 5.2 Step 3: Recovering the Variances

Now that we know the values of $w_i$ and all the means $\widetilde{\mu}_i$, we show how to recover the variances. This can be done in many ways, and we will outline one which ends up solving a linear system of equations. Recall that for each Gaussian, the covariance matrix is diagonal (denoted $\Sigma_i$, with $j$th entry equal to $\sigma_{ij}^2$).

Let us show how to recover $\sigma_{i1}^2$ for $1 \le i \le R$. The same procedure can be applied to the other dimensions to recover $\sigma_{ij}^2$ for all $j$. Let us divide the set of indices $\{2, 3, \ldots, n\}$ into $\ell$ (nearly equal) sets $S_1, S_2, \ldots, S_\ell$. Now consider the expression

$$\mathcal{N}_1 = \mathbf{E}[x(1)^2 (x_{|S_1} \otimes x_{|S_2} \otimes \cdots \otimes x_{|S_\ell})].$$

This can be evaluated as before. Write $\widetilde{\mu}_i^{(t)}$ to denote the portion of $\widetilde{\mu}_i$ restricted to $S_t$, and similarly $\eta_i^{(t)}$ to denote the portion of the noise vector $\eta_i$. This gives

$$\mathcal{N}_1 = \sum_i w_i (\widetilde{\mu}_i(1)^2 + \sigma_{i1}^2)(\widetilde{\mu}_i^{(1)} \otimes \widetilde{\mu}_i^{(2)} \otimes \cdots \otimes \widetilde{\mu}_i^{(\ell)}).$$

Now recall that we *know* the vectors $\widetilde{\mu}_i$ and hence each of the tensors $\widetilde{\mu}_i^{(1)} \otimes \widetilde{\mu}_i^{(2)} \otimes \cdots \otimes \widetilde{\mu}_i^{(\ell)}$. Further, since our $\widetilde{\mu}_i$ are the perturbed means, our theorem (Theorem 3.3) about the condition number of Khatri-Rao products implies that the matrix (call it $\mathcal{M}$) whose columns are the flattened $\prod_t \widetilde{\mu}_i^{(t)}$ for different $i$, is well conditioned, i.e., has $\sigma_R(\cdot) \ge 1/\mathrm{poly}_\ell(n/\rho)$. This implies that a system of linear equations $\mathcal{M}z = z'$ can be solved to recover $z$ up to a $1/\mathrm{poly}_\ell(n/\rho)$ accuracy (assuming we know $z'$ up to a similar accuracy).

Now using this with $z'$ being the flattened $\mathcal{N}_1$ allows us to recover the values of $w_i(\widetilde{\mu}_i(1) + \sigma_{i1}^2)$ for $1 \le i \le R$. From this, since we know the values of $w_i$ and $\widetilde{\mu}_i(1)$ for each $i$, we can recover the values $\sigma_{i1}^2$ for all $i$. As mentioned before, we can repeat this process for other dimensions and recover $\sigma_{ij}^2$ for all $i, j$.

# 6 Acknowledgements

We thank Ryan O'Donnell for suggesting that we extend our techniques for learning mixtures of spherical Gaussians to the more general problem of learning axis-aligned Gaussians.

## References


[1] E. Allman, C. Matias and J. Rhodes. Identifiability of Parameters in Latent Structure Models with many Observed Variables. *Annals of Statistics*, pages 3099–3132, 2009. 1.1, 1.2, 1.8, 2, 4, 4.2

[2] A. Anandkumar, D. Hsu and S. Kakade. A method of moments for mixture models and hidden Markov models. In *COLT* 2012. 1.2, 4, 4

[3] A. Anandkumar, R. Ge, D. Hsu and S. Kakade. A Tensor Spectral Approach to Learning Mixed Membership Community Models. In *COLT* 2013. 1.1, 1.1

[4] A. Anandkumar, R. Ge, D. Hsu, S. Kakade and M. Telgarsky. Tensor Decompositions for Learning Latent Variable Models. *arxiv:1210.7559*, 2012.

[5] A. Anandkumar, D. Foster, D. Hsu, S. Kakade, Y. Liu. A Spectral Algorithm for Latent Dirichlet Allocation. In *NIPS*, pages 926–934, 2012. 1.1, 1.1

[6] S. Arora, R. Ge, A. Moitra and S. Sachdeva. Provable ICA with Unknown Gaussian Noise, and Implications for Gaussian Mixtures and Autoencoders. In *NIPS*, pages 2384–2392, 2012.

[7] M. Belkin, L. Rademacher and J. Voss. Bling Signal Separation in the Presence of Gaussian Noise. In *COLT* 2013.

[8] M. Belkin and K. Sinha. Polynomial Learning of Distribution Families. In *FOCS*, pages 103–112, 2010.

[9] A. Bhaskara, M. Charikar and A. Vijayaraghavan. Uniqueness of Tensor Decompositions with Applications to Polynomial Identifiability. *arxiv:1304.8087*, 2013. 1.1, 1.2, 2, A

[10] J. Chang. Full Reconstruction of Markov Models on Evolutionary Trees: Identifiability and Consistency. *Mathematical Biosciences*, pages 51–73, 1996.

[11] P. Comon. Independent Component Analysis: A New Concept? *Signal Processing*, pages 287–314, 1994. 1.1, 1.1, 2, 2.1

[12] S. Dasgupta. Learning Mixtures of Gaussians. In *FOCS*, pages 634–644, 1999.

[13] L. De Lathauwer, J Castaing and J. Cardoso. Fourth-order Cumulant-based Blind Identification of Underdetermined Mixtures. *IEEE Trans. on Signal Processing*, 55(6):2965–2973, 2007. 1.8, 2

[14] J. Feldman, R. A. Servedio, and R. O'Donnell. PAC Learning Axis-aligned Mixtures of Gaussians with No Separation Assumption. In *COLT*, pages 20–34, 2006. 1.2

[15] A. Frieze, M. Jerrum, R. Kannan. Learning Linear Transformations. In *FOCS*, pages 359–368, 1996.

[16] N. Goyal, S. Vempala and Y. Xiao. Fourier PCA. *arxiv:1306.5825*, 2013. 1.1, 1.2, 2, A

[17] J. Håstad. Tensor Rank is $NP$-Complete. *Journal of Algorithms*, pages 644–654, 1990. 1.1

[18] C. Hillar and L-H. Lim. Most Tensor Problems are $NP$-Hard. *arxiv:0911.1393v4*, 2013. 1.1

[19] R. Horn and C. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[20] D. Hsu and S. Kakade. Learning Mixtures of Spherical Gaussians: Moment Methods and Spectral Decompositions. In *ITCS*, pages 11–20, 2013. 1.1, 1.1, 1.2, 5.1

[21] A. Hyvärinen, J. Karhunen and E. Oja. *Independent Component Analysis*. Wiley Interscience, 2001.

[22] A. T. Kalai, A. Moitra, and G. Valiant. Efficiently Learning Mixtures of Two Gaussians. In *STOC*, pages 553-562, 2010.

[23] A. T. Kalai, A. Samorodnitsky and S-H Teng. Learning and Smoothed Analysis. In *FOCS*, pages 395–404, 2009.

[24] J. Kruskal. Three-way Arrays: Rank and Uniqueness of Trilinear Decompositions. *Linear Algebra and Applications*, 18:95–138, 1977. 1.1, 2

[25] S. Leurgans, R. Ross and R. Abel. A Decomposition for Three-way Arrays. *SIAM Journal on Matrix Analysis and Applications*, 14(4):1064–1083, 1993. 1.1, 1.2, 1.8, 2, 2.1

[26] B. Lindsay. *Mixture Models: Theory, Geometry and Applications*. Institute for Mathematical Statistics, 1995.

[27] P. McCullagh. *Tensor Methods in Statistics*. Chapman and Hall/CRC, 1987. 1.1

[28] A. Moitra and G. Valiant. Setting the Polynomial Learnability of Mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

[29] E. Mossel and S. Roch. Learning Nonsingular Phylogenies and Hidden Markov Models. In *STOC*, pages 366–375, 2005. 1.1, 1.1, 4

[30] Y. Rabani, L. Schulman and C. Swamy. Learning mixtures of arbitrary distributions over large discrete domains. . In *ITCS* 2014. 4, 4

[31] C. Spearman. General Intelligence. *American Journal of Psychology*, pages 201–293, 1904. 1.1

[32] D.A. Spielman, S.H. Teng. Smoothed Analysis of Algorithms: Why the Simplex Algorithm usually takes Polynomial Time. *Journal of the ACM*, pages 385–463, 2004. 1.2

[33] D.A. Spielman, S.H. Teng. Smoothed Analysis: An Attempt to Explain the Behavior of Algorithms in Practice. *Communications of the ACM*, pages 76-84, 2009. 1.2

[34] A. Stegeman and P. Comon. Subtracting a Best Rank-1 Approximation may Increase Tensor Rank. *Linear Algebra and Its Applications*, pages 1276–1300, 2010. 1.1

[35] H. Teicher. Identifiability of Mixtures. *Annals of Mathematical Statistics*, pages 244–248, 1961.

[36] S. Vempala, Y. Xiao. Structure from Local Optima: Learning Subspace Juntas via Higher Order PCA. *Arxiv:*abs/1108.3329, 2011.

[37] P. Wedin. Perturbation Bounds in Connection with Singular Value Decompositions. *BIT*, 12:99–111, 1972.

# A Stability of the recovery algorithm

In this section we prove Theorem 2.3, which shows that the algorithm from section 2 is actually robust to errors, under Condition 2.2. This consists of two parts: first proving that the preprocessing step indeed allows us to recover $u_i$ (approximately), and second, that DECOMPOSE is robust to noise.

### Stability of the preprocessing

Suppose we are given $T + E$, where $T = \sum_i u_i \otimes v_i \otimes w_i$, and $E$ is a tensor each of whose entries is $< \epsilon \cdot \text{poly}(1/\kappa, 1/n, 1/\delta)$. Let $\widetilde{u}_{j,k}$ be vectors in $\Re^m$ defined as before, and let $\widetilde{U}$ be the $m \times np$ matrix whose columns are $\widetilde{u}_{j,k}$ (for different $j, k$). Let $u_i'$ be the projection of $u_i$ onto the span of the top $R$ singular vectors of $\widetilde{U}$. By Claim 4.4 in [9], we have $\|T - \sum_i u_i' \otimes v_i \otimes w_i\|_F < 2\|E\|_F$, and thus from the robust version of Kruskal's uniqueness theorem [9], we must have that $u_i'$ and $u_i$ are $\epsilon \cdot \text{poly}(1/\kappa, 1/n, 1/\delta)$ close. Repeating the above along the second mode allows us to move to an $R \times R \times p$ tensor.

### Stability of DECOMPOSE

Next, we establish that DECOMPOSE is stable (in what follows, we have $m = n = R$). Intuitively, DECOMPOSE is stable provided that the matrices $U$ and $V$ are well-conditioned and the eigenvalues of the matrices that we need to diagonalize are separated.

The main step in DECOMPOSE is an eigendecomposition, so first we will establish perturbation bounds. The standard perturbation bounds are known as sin $\theta$ theorems following Davis-Kahan and Wedin. However these bounds hold most generally for the singular value decomposition of an arbitrary (not necessarily symmetric) matrix. We require perturbation bounds for eigen-decompositions of general matrices. There are known bounds due to Eisenstat and Ipsen, however the notion of separation required there is difficult to work with and for our purposes it is easier to prove a direct bound in our setting.

Suppose $M = UDU^{-1}$ and $\widehat{M} = M(I + E) + F$ and $M$ and $\widehat{M}$ are $n \times n$ matrices. In order to relate the eigendecompositions of $M$ and $\widehat{M}$ respectively, we will first need to establish that the eigenvalues of $M$ are all distinct. We thank Santosh Vempala for pointing out an error in an earlier version. We incorrectly used the Bauer-Fike Theorem to show that $\widehat{M}$ is diagonalizable, but this theorem only shows that each eigenvalue of $\widehat{M}$ is close to some eigenvalue of $M$, but does not show that there is a one-to-one mapping. Fortunately there is a fix for this that works under the same conditions (but again see [16] for an earlier, alternative proof that uses a "homotopy argument").

**Definition A.1.** Let $\text{sep}(D) = \min_{i \neq j} |D_{i,i} - D_{j,j}|$.

Our first goal is to prove that $\widehat{M}$ is diagonalizable, and we will do this by establishing that its eigenvalues are distinct if the error matrices $E$ and $F$ are not too large. Consider

$$U^{-1}(M(I + E) + F)U = D + R$$

where $R = U^{-1}(ME + F)U$. We can bound each entry in $R$ by $\kappa(U)(\|ME\|_2 + \|F\|_2)$. Hence if $E$ and $F$ are not too large, the eigenvalues of $D + R$ are close to the eigenvalues of $D$ using Gershgorin's disk theorem, and the eigenvalues of $D + R$ are the same as the eigenvalues of $\widehat{M}$ since these matrices are similar. So we conclude:

**Lemma A.2.** *If $\kappa(U)(\|ME\|_2 + \|F\|_2) < sep(D)/(2n)$ then the eigenvalues of $\widehat{M}$ are distinct and it is diagonalizable.*

Next we prove that the eigenvectors of $\widehat{M}$ are also close to those of $M$ (this step will rely on $\widehat{M}$ being diagonalizable). This technique is standard in numerical analysis, but it will be more convenient for us to work with relative perturbations (i.e. $\widehat{M} = M(I+E)+F$) so we include the proof of such a bound for completeness

Consider a right eigenvector $\widehat{u}_i$ of $\widehat{M}$ with eigenvalue $\widehat{\lambda}_i$. We will assume that the conditions of the above corollary are met, so that there is a unique eigenvector $u_i$ of $M$ with eigenvalue $\lambda_i$ which it is paired with. Then since the eigenvectors $\{u_i\}_i$ of $M$ are full rank, we can write $\widehat{u}_i = \sum_j c_j u_j$. Then

$$\begin{aligned} \widehat{M}\widehat{u}_i &= \widehat{\lambda}_i \widehat{u}_i \\ \sum_j c_j \lambda_j u_j + (ME+F)\widehat{u}_i &= \widehat{\lambda}_i \widehat{u}_i \\ \sum_j c_j(\lambda_j - \widehat{\lambda}_i)u_j &= -(ME+F)\widehat{u}_i \end{aligned}$$

Now we can left multiply by the $j^{th}$ row of $U^{-1}$; call this vector $w_j^T$. Since $U^{-1}U = I$, we have that $w_j^T u_i = \mathbf{1}_{i=j}$. Hence

$$c_j(\lambda_j - \widehat{\lambda}_i) = -w_j^T(ME+F)\widehat{u}_i$$

So we conclude:

$$\|\widehat{u}_i - u_i\|_2^2 = 2dist(\widehat{u}_i, span(u_i))^2 \leq 2\sum_{j\neq i}\Big(\frac{(w_j^T(ME+F)\widehat{u}_i)}{|\lambda_j - \widehat{\lambda}_i|}\Big)^2 \leq 8\sum_{j\neq i}\frac{\|U^{-1}(ME+F)\widehat{u}_i\|_2^2}{sep(D)^2}$$

where we have used the condition that $\kappa(U)(\|ME\|_2 + \|F\|_2) < sep(D)/2$ to lower bound the denominator. Furthermore: $\|U^{-1}ME\widehat{u}_i\|_2 = \|DU^{-1}E\widehat{u}_i\|_2 \leq \frac{\sigma_{max}(E)\lambda_{max}(D)}{\sigma_{min}(U)}$ since $\widehat{u}_i$ is a unit vector.

**Theorem A.3.** *If $\kappa(U)(\|ME\|_2 + \|F\|_2) < sep(D)/2$, then*

$$\|\widehat{u}_i - u_i\|_2 \leq 3\frac{\sigma_{max}(E)\lambda_{max}(D) + \sigma_{max}(F)}{\sigma_{min}(U)sep(D)}$$

Now we are ready to analyze the stability of DECOMPOSE: Let $T = \sum_{i=1}^n u_i \otimes v_i \otimes w_i$ be an $n \times n \times p$ tensor that satisfies Condition 2.2. In our settings of interest we are not given $T$ exactly but rather a good approximation to it, and here let us model this noise as an additive error $E$ that is itself an $n \times n \times p$ tensor.

**Claim A.4.** *With high probability, $sep(D_a D_b^{-1}), sep(D_b D_a^{-1}) \geq \frac{\delta}{\sqrt{p}}$.*

**Proof:** Fix some $i, j$. The $(i,i)$th entry of $D_a D_b^{-1}$ is precisely $\frac{\langle w_i, a\rangle}{\langle w_i, b\rangle}$. Note that $\mathbf{Pr}[|\langle w_i, b\rangle| > n\|w_i\|]$ is $\exp(-n)$, thus the denominators are all at least $1/(Cn)$ in magnitude with probability $1-\exp(-n)$.

Now given $b$ for which this happens, we have $\frac{\langle w_i,a\rangle}{\langle w_i,b\rangle} - \frac{\langle w_j,a\rangle}{\langle w_j,b\rangle} = c_i\langle w_i, a\rangle - c_j\langle w_j, a\rangle$ where $c_i, c_j$ have magnitude $> 1/(Cn)$. Because $w_i$ has at least a $\delta$ component orthogonal to $w_j$, anti-concentration of Gaussians implies that the difference above is at least $\delta/C^2n^6$ with probability at least $1 - 1/n^4$. Thus we can take a union bound over all pairs. ■

We will make crucial use of the following matrix identity:

$$(A+Z)^{-1} = A^{-1} - A^{-1}Z(I+A^{-1}Z)^{-1}A^{-1}$$

Let $N_a = T_a + E_a$ and $N_b = T_b + E_b$. Then using the above identity we have:

$$N_a(N_b)^{-1} = T_a(T_b)^{-1}(I+F) + G$$

where $F = -E_b(I + (T_b)^{-1}E_b)^{-1}(T_b)^{-1}$ and $G = E_a(T_b)^{-1}$

**Claim A.5.** $\sigma_{max}(F) \leq \frac{\sigma_{max}(E_b)}{\sigma_{min}(T_b)-\sigma_{max}(E_b)}$ *and* $\sigma_{max}(G) \leq \frac{\sigma_{max}(E_a)}{\sigma_{min}(T_b)}$

**Proof:** Using Weyl's Inequality we have

$$\sigma_{max}(F) \leq \frac{\sigma_{max}(E_b)}{1-\frac{\sigma_{max}(E_b)}{\sigma_{min}(T_b)}} \times \frac{1}{\sigma_{min}(T_b)} = \frac{\sigma_{max}(E_b)}{\sigma_{min}(T_b)-\sigma_{max}(E_b)}$$

as desired. The second bound is obvious. ■

We can now use Theorem A.3 to bound the error in recovering the factors $U$ and $V$ by setting e.g. $M = T_a(T_b)^{-1}$. Additionally, the following claim establishes that the linear system used to solve for $W$ is well-conditioned and hence we can also bound the error in recovering $W$.

**Claim A.6.** $\kappa(U \odot V) \leq \frac{\min(\sigma_{max}(U),\sigma_{max}(V))}{\max(\sigma_{min}(U),\sigma_{min}(V))} \leq \min(\kappa(U),\kappa(V))$

These bounds establish what we qualitatively asserted: DECOMPOSE is stable provided that the matrices $U$ and $V$ are well-conditioned and the eigenvalues of the matrices that we need to diagonalize are separated.

# B K-rank of the Khatri-Rao product

## B.1 Leave-One-Out Distance

Recall: we defined the leave-one-out distance in Section 3. Here we establish that is indeed equivalent to the smallest singular value, up to polynomial factors. In our main proof, this quantity will be much easer to work with since it allows us to translate questions about a set of vectors being well-conditioned to reasoning about projection of each vector onto the orthogonal complement of the others.

**Proof of Lemma 3.5:** Using the variational characterization for singular values: $\sigma_{min}(A) = \min_{u,\|u\|_2=1} \|Au\|_2$. Then let $i = \text{argmax}|u_i|$. Clearly $|u_i| \geq 1/\sqrt{m}$ since $\|u\|_2 = 1$. Then $\|A_i + \sum_{j\neq i} A_j \frac{u_j}{u_i}\|_2 = \frac{\sigma_{min}(A)}{u_i}$. Hence

$$\ell(A) \leq dist(A_i, span\{A_j\}_{j\neq i}) \leq \frac{\sigma_{min}(A)}{u_i} \leq \sigma_{min}(A)\sqrt{m}$$

Conversely, let $i = \text{argmin}_i dist(A_i, span\{A_j\}_{j\neq i})$. Then there are coefficients (with $u_i = 1$) such that

$$\|A_i u_i + \sum_{j\neq i} A_j u_j\|_2 = \ell(A).$$

Clearly $\|u\|_2 \geq 1$ since $u_i = 1$. And we conclude that

$$\ell(A) = \|A_i u_i + \sum_{j\neq i} A_j u_j\|_2 \geq \frac{\|A_i u_i + \sum_{j\neq i} A_j u_j\|_2}{\|u\|_2} \geq \sigma_{min}(A).$$

■

## B.2 Proof of Proposition 3.12

We now give the complete details of the proof of Proposition 3.12, that shows how the Kruskal rank multiplies in the smoothed setting for two-wise products. The proof follows by just combining Lemma 3.16 and Lemma 3.15.

Let $\mathcal{U}$ be the span of the top $\delta n^2$ singular values of $M$. Thus $\mathcal{U}$ is a $\delta n^2$ dimensional subspace of $\mathbb{R}^{n^2}$. Using Lemma 3.16 with:

$$r = \frac{n^{1/2}}{2}, \quad m = n, \quad \delta' = \frac{\delta}{n^{1/2}},$$

we obtain $n \times n$ matrices $M_1, M_2, \ldots, M_r$ having the $(\theta, \delta')$-orthogonality property. Note that in this setting, $\delta' m = \frac{n^{1/2}}{2}$.

Thus by applying Lemma 3.15, we have that the matrix $Q(\widetilde{x})$, defined as before, satisfies

$$\mathop{\mathbf{Pr}}_{x}\left[\sigma_{r/2}\left(Q(\widetilde{x})\right) \geq \frac{\rho\theta}{n^4}\right] \geq 1 - \exp(-r). \tag{12}$$

Now let us consider

$$\sum_s (\widetilde{y}^T M_s \widetilde{x})^2 = \|\widetilde{y}^T Q(\widetilde{x})\|^2.$$

Since $Q(\widetilde{x})$ has many non-negligible singular values (Eq.(12)), we have (by Fact 3.26 for details) that an $\rho$-perturbed vector has a non-negligible norm when multiplied by $Q$. More precisely, $\mathbf{Pr}[\|\widetilde{y}^T Q(\widetilde{x})\| \geq \rho\theta/n^4] \geq 1 - \exp(-r/2)$. Thus for one of the terms $M_s$, we have $|M_s(\widetilde{x} \otimes \widetilde{y})| \geq \rho\theta/n^5$ with probability $\geq 1 - \exp(-r/2)$.

Now this *almost* completes the proof, but recall that our aim is to argue about $M(\widetilde{x} \otimes \widetilde{y})$, where $M$ is the given matrix. $\mathrm{vec}(M_s)$ is a vector in the span of the top $\delta n^2$ (right) singular vectors of $M$, and $\sigma_{\delta n^2} \geq \tau$, thus we can write $M_s$ as a combination of the rows of $M$, with each weight in the combination being $\leq n/\tau$ (Lemma B.1). This implies that for at least one row $M^{(j)}$ of the matrix $M$, we must have

$$\|M^{(j)}(\widetilde{x} \otimes \widetilde{y}\| \geq \frac{\theta\rho\tau}{n^6} = \frac{\rho\tau}{n^{O(1)}}.$$

(Otherwise we have a contradiction). This completes the proof.

□

Before we give the complete proofs of the two main lemmas regarding ordered $(\theta, \delta)$ orthogonal systems (Lemma 3.16 and Lemma 3.15), we start with a simple lemma about top singular vectors of matrices, which is very useful to obtain linear combinations of small length.

**Lemma B.1** (Expressing top singular vectors as small combinations of columns)**.** *Suppose we have a $m \times n$ matrix $M$ with $\sigma_t(M) \geq \eta$, and let $v_1, v_2, \ldots v_t \in \mathbb{R}^m$ be the top $t$ left-singular vectors of $M$. Then these top $t$ singular vector can be expressed using* small *linear combinations of the columns $\{M(i)\}_{i\in[n]}$ i.e.*

$$\forall k \in [t],\ \exists \{\alpha_{k,i}\}_{i\in[n]} \text{ such that } v_k = \sum_{i\in[n]} \alpha_{k,i} M(i) \text{ and} \sum_i \alpha_{k,i}^2 \leq 1/\eta^2$$

**Proof:** Let $\ell$ correspond to the number of non-zero singular values of $M$. Using the SVD, there exists matrices $V \in \mathbb{R}^{m\times\ell}, U \in \mathbb{R}^{n\times\ell}$ with orthonormal columns (both unitary matrices), and a diagonal matrix $\Sigma \in \mathbb{R}^{\ell\times\ell}$ such that $M = V\Sigma U^T$. Since the $n \times \ell$ matrix $V = M(U\Sigma^{-1})$, the $t$ columns of $V$ corresponding to the top $t$ singular values ($\sigma_t(M) \geq \eta$) correspond to linear combinations which are small i.e. $\forall k \in [t],\ \|\alpha_k\| \leq 1/\eta$. ■

### B.3 Constructing the $(\theta, \delta)$-Orthogonal System (Proof of Lemma 3.16)

Let $\mathcal{V}$ be a subspace of $\mathbb{R}^{n\cdot m}$, with its co-ordinates indexed by $[n] \times [m]$. Further,remember that the vectors in $\mathbb{R}^{n\cdot m}$ are also treated as matrices of size $n \times m$.

We now give the complete proof of lemma 3.18 that shows that the average robust dimension of column projections is large if the dimension of $\mathcal{V}$ is large .

**Proof of Lemma 3.18:** Let $d = \dim(\mathcal{V})$. Let $B$ be a $p_1p_2 \times d$ matrix composed of a orthonormal basis (of $d$ vectors) for $\mathcal{V}$ i.e. the $j^{th}$ column of $B$ is the $j^{th}$ basis vector ($j \in [d]$) of $\mathcal{V}$. Clearly $\sigma_d(B) = 1$.
For $i \in [p_2]$, let $B_i$ be the $p_1 \times d$ matrix obtained by projecting the columns of $B$ on just the rows given by $[p_1] \times i$. Hence, $B$ is obtained by just concatenating the columns as $B^T = \left[{B_1}^T \| {B_2}^T \| \ldots \| {B_p}^T\right]$. Finally, let $d_i = \max t$ such that $\sigma_t(B_i) \geq \frac{1}{\sqrt{p_2}}$.

We will first show that $\sum_i d_i \geq d$. Then we will show that $\dim_i^\tau(\mathcal{V}) \geq d_i$ to complete our proof. Suppose for contradiction that $\sum_{i\in[p_2]} d_i < d$. Let $\mathcal{S}_i$ be the $(d - d_1)$-dimensional subspace of $\mathbb{R}^d$ spanned by the last $(d - d_1)$ right singular vectors of $B_i$. Hence,

$$\text{for unit vectors } \alpha \in \mathcal{S}_i \subseteq \mathbb{R}^d,\ \|B_i\alpha\| < \frac{1}{\sqrt{p_2}}.$$

Since, $d - \sum_{i\in[p_2]} d_i > 0$, there exists at least one unit vector $\alpha \in \bigcap_i \mathcal{S}_i^\perp$. Picking this unit vector $\alpha \in \mathbb{R}^d$, we have $\|B\alpha\|_2^2 = \sum_{i\in[p_2]} \|B_i\alpha\|_2^2 < p_2 \cdot (\frac{1}{\sqrt{p_2}})^2 < 1$. This contradicts $\sigma_d(B) \geq 1$

To establish the second part, consider some $B_i$ ($i \in [p_2]$). We pick $d_i$ orthonormal vectors $\in \mathbb{R}^{p_1}$ corresponding to the top $d_i$ left-singular vectors of $B_i$. By using Lemma B.1, we know that each of these $j \in [d_i]$ vectors can be expressed as a small combination $\vec{\alpha_j}$ of the columns of $B_i$ s.t. $\|\vec{\alpha_j}\| \leq \sqrt{p_2}$. Further, if we associate with each of these $j \in [d_i]$ vectors, the vector $w_j \in \mathbb{R}^{(p_1p_2)}$ given by the same combination $\vec{\alpha_j}$ of the columns of $B$, we see that $\|w_j\| \leq \sqrt{p_2}$ since the columns of the matrix $B$ are orthonormal. ■

### B.4 Implications of Ordered $(\theta, \delta)$-Orthogonality: Details of Proof of Lemma 3.15

Here we show some auxiliary lemmas that are used in the Proof of Lemma B.4.

**Claim B.2.** *Suppose $v_1, v_2, \ldots, v_m$ are a set of vectors in $\Re^n$ of length $\leq 1$, having the $\theta$-orthogonal property. Then we have*

*(a) For $g \sim \mathcal{N}(0,1)^n$, we have $\sum_i \langle v_i, g\rangle^2 \geq \theta^2/2$ with probability $\geq 1 - \exp(-\Omega(m))$,*

*(b) For $g \sim \mathcal{N}(0,1)^m$, we have $\|\sum_i g_i v_i\|^2 \geq \theta^2/2$ with probability $\geq 1 - \exp(-\Omega(m))$.*

*Furthermore, part (a) holds even if $g$ is drawn from $u + g'$, for any fixed vector $u$ and $g' \sim \mathcal{N}(0,1)^n$.*

**Proof:** First note that we must have $m \leq n$, because otherwise $\{v_1, v_2, \ldots, v_m\}$ cannot have the $\theta$-orthogonal property for $\theta > 0$. For any $j \in [m]$, we claim that

$$\mathbf{Pr}[(\langle v_j, g\rangle^2 < \theta^2/2) \mid v_1, v_2, \ldots, v_{j-1}] < 1/2. \tag{13}$$

To see this, write $v_j = v'_j + v_j^\perp$, where $v_j^\perp$ is orthogonal to the span of $\{v_1, v_2, \ldots, v_{j-1}\}$. Since $j \in I$, we have $\|v_j^\perp\| \geq \theta$. Now given the vectors $v_1, v_2, \ldots, v_{j-1}$, the value $\langle v'_j, g\rangle$ is fixed, but $\langle v_j^\perp, g\rangle$ is distributed as a Gaussian with variance $\theta^2$ (since $g$ is a Gaussian of unit variance in each direction).

Thus from a standard anti-concentration property for the one-dimensional Gaussian, $\langle v_j, g\rangle$ cannot have a mass $> 1/2$ in any $\theta^2$ length interval, in particular, it cannot lie in $[-\theta^2/2, \theta^2/2]$ with probability $> 1/2$. This proves Eq. (13). Now since this is true for any conditioning $v_1, v_2, \ldots, v_{j-1}$ and for all $j$, it follows (see Lemma B.3 for a formal justification) that

$$\mathbf{Pr}[\langle v_j, g\rangle^2 < \theta^2/2 \text{ for all } j] < \frac{1}{2^m} < \exp(-m/2).$$

This completes the proof of the claim, part (a). Note that even if we had $g$ replaced by $u+g$ throughout, the anti-concentration property still holds (we have a shifted one-dimensional Gaussian), thus the proof goes through verbatim.

Let us now prove part (b). First note that if we denote by $M$ the $n \times m$ matrix whose columns are the $v_i$, then part (a) deals with the distribution of $g^T M M^T g$, where $g \sim \mathcal{N}(0,1)^n$. Part (b) deals with the distribution of $g^T M^T M g$, where $g \sim \mathcal{N}(0,1)^m$. But since the eigenvalues of $MM^T$ and $M^T M$ are precisely the same, due to the rotational invariance of Gaussians, these two quantities are distributed exactly the same way. This completes the proof. ■

**Lemma B.3.** *Suppose we have random variables $X_1, X_2, \ldots, X_r$ and an event $f(\cdot)$ which is defined to occur if its argument lies in a certain interval (e.g. $f(X)$ occurs iff $0 < X < 1$). Further, suppose we have $\mathbf{Pr}[f(X_1)] \leq p$, and $\mathbf{Pr}[f(X_i)|X_1, X_2, \ldots, X_{i-1}] \leq p$ for all $X_1, X_2, \ldots, X_{i-1}$. Then*

$$Pr[f(X_1) \wedge f(X_2) \wedge \cdots \wedge f(X_r)] \leq p^r.$$

# C Applications to Mixture Models

## C.1 Sampling Error Estimates for Multi-view Models

In this section, we show error estimates for $\ell$-order tensors obtained by looking at the $\ell^{th}$ moment of the multi-view model.

**Lemma C.1** (Error estimates for Multiview mixture model)**.** *For every $\ell \in \mathbb{N}$, suppose we have a multi-view model, with parameters $\{w_r\}_{r\in[R]}$ and $\{M^{(j)}\}_{j\in[\ell]}$, the $n$ dimensional sample vectors $x^{(j)}$ have $\|x^{(j)}\|_\infty \leq 1$. Then, for every $\varepsilon > 0$, there exists $N = O(\varepsilon^{-2}\sqrt{\ell \log n})$ such that if $N$ samples $\{x(1)^{(j)}\}_{j\in[\ell]}, \{x(2)^{(j)}\}_{j\in[\ell]}, \ldots, \{x(N)^{(j)}\}_{j\in[\ell]}$ are generated, then with high probability*

$$\|\mathbf{E}\, x^{(1)} \otimes x^{(2)} \otimes \ldots x^{(\ell)} - \frac{1}{N}\left(\sum_{t\in[N]} x(t)^{(1)} \otimes x(t)^{(2)} \otimes x(t)^{(\ell)}\right)\|_\infty < \varepsilon \tag{14}$$

**Proof:** We first bound the $\|\cdot\|_\infty$ norm of the difference of tensors i.e. we show that

$$\forall \{i_1, i_2, \ldots, i_\ell\} \in [n]^\ell, \left|\mathbf{E}\prod_{j\in[\ell]} x_{i_j}^{(j)} - \frac{1}{N}\left(\sum_{t\in[N]}\prod_{j\in[\ell]} x(t)_{i_j}^{(j)}\right)\right| < \varepsilon/n^{\ell/2}.$$

Consider a fixed entry $(i_1, i_2, \ldots, i_\ell)$ of the tensor.

Each sample $t \in [N]$ corresponds to an independent random variable with a bound of 1. Hence, we have a sum of $N$ bounded random variables. By Bernstein bounds, probability for (14) to not occur $\exp\left(-\frac{\left(\varepsilon n^{-\ell/2}\right)^2 N^2}{2N}\right) = \exp\left(-\varepsilon^2 N/\left(2n^\ell\right)\right)$. We have $n^\ell$ events to union bound over. Hence $N = O(\varepsilon^{-2} n^\ell \sqrt{\ell \log n})$ suffices. Note that similar bounds hold when the $x^{(j)} \in \mathbb{R}^n$ are generated from a multivariate gaussian. ■

## C.2 Error Analysis for Multi-view Models

**Lemma C.2.** *Suppose $\|u \otimes v - u' \otimes v'\|_F < \delta$, and $L_{\min} \le \|u\|, \|v\|, \|u'\|, \|v'\| \le L_{\max}$, with $\delta < \frac{\min\{L_{\min}^2, 1\}}{(2\max\{L_{\max}, 1\})}$. If $u = \alpha_1 u' + \beta_1 \tilde{u}_\perp$ and $v = \alpha_2 v' + \beta_2 \tilde{v}_\perp$, where $\tilde{u}_\perp$ and $\tilde{v}_\perp$ are unit vectors orthogonal to $u', v'$ respectively, then we have*

$$|1 - \alpha_1\alpha_2| < \delta/L_{\min}^2 \quad \textit{and} \quad \beta_1 < \sqrt{\delta},\ \beta_2 < \sqrt{\delta}.$$

**Proof:** We are given that $u = \alpha_1 u' + \beta_1 \tilde{u}_\perp$ and $v = \alpha_2 v' + \beta_2 \tilde{v}_\perp$. Now, since the tensored vectors are close

$$\begin{aligned} \|u \otimes v - u' \otimes v'\|_F^2 &< \delta^2 \\ \|(1-\alpha_1\alpha_2)u' \otimes v' + \beta_1\alpha_2 \tilde{u}_\perp \otimes v' + \beta_2\alpha_1 u' \otimes \tilde{v}_\perp + \beta_1\beta_2 \tilde{u}_\perp \otimes \tilde{v}_\perp\|_F^2 &< \delta^2 \\ L_{\min}^4(1-\alpha_1\alpha_2)^2 + \beta_1^2\alpha_2^2 L_{\min}^2 + \beta_2^2\alpha_1^2 L_{\min}^2 + \beta_1^2\beta_2^2 &< \delta^2 \end{aligned} \tag{15}$$

This implies that $|1 - \alpha_1\alpha_2| < \delta/L_{\min}^2$ as required.

Now, let us assume $\beta_1 > \sqrt{\delta}$. This at once implies that $\beta_2 < \sqrt{\delta}$. Also

$$\begin{aligned} L_{\min}^2 \le \|v\|^2 &= \alpha_2^2\|v'\|^2 + \beta_2^2 \\ L_{\min}^2 - \delta &\le \alpha_2^2 L_{\max}^2 \\ \text{Hence,} \quad \alpha_2 &\ge \frac{L_{\min}}{2L_{\max}} \end{aligned}$$

Now, using (15), we see that $\beta_1 < \sqrt{\delta}$. ■

## C.3 Sampling Error Estimates for Gaussians

**Lemma C.3** (Error estimates for Gaussians)**.** *Suppose $x$ is generated from a mixture of $R$-gaussians with means $\{\mu_r\}_{r\in[R]}$ and covariance $\Sigma_i$ that is diagonal , with the means satisfying $\|\mu_r\| \le B$. Let $\sigma = \max_i \sigma_{\max}(\Sigma_i)$*
*For every $\varepsilon > 0, \ell \in \mathbb{N}$, there exists $N = \Omega(\text{poly}(\frac{1}{\varepsilon})), \sigma^2, n, R)$ such that if $x^{(1)}, x^{(2)}, \ldots, x^{(N)} \in R^n$ were the $N$ samples, then*

$$\forall\{i_1, i_2, \ldots, i_\ell\} \in [n]^\ell, \left|\mathbf{E}\prod_{j\in[\ell]} x_{i_j} - \frac{1}{N}\left(\sum_{t\in[N]}\prod_{j\in[\ell]} x_{i_j}^{(t)}\right)\right| < \varepsilon. \tag{16}$$

*In other words,*

$$\|\mathbf{E}\, x^{\otimes\ell} - \frac{1}{N}\Big(\sum_{t\in[N]} (x^{(t)})^{\otimes\ell}\Big)\|_\infty < \varepsilon$$

**Proof:** Fix an element $(i_1, i_2, \ldots, i_\ell)$ of the $\ell$-order tensor. Each point $t \in [N]$ corresponds to an i.i.d random variable $Z^t = x_{i_1}^{(t)} x_{i_2}^{(t)} \ldots x_\ell^{(t)}$. We are interested in the deviation of the sum $S = \frac{1}{N}\sum_{t\in[N]} Z^t$. Each of the i.i.d rvs has value $Z = x_{i_1} x_{i_2} \ldots x_\ell$. Since the gaussians are axis-aligned and each mean is bounded by $B$, $|Z| < (B + t\sigma)^\ell$ with probability $O\left(\exp(-t^2/2)\right)$. Hence, by using standard sub-gaussian tail inequalities, we get

$$\mathbf{Pr}\,|S - \mathbf{E}\,z| > \varepsilon < \exp\left(-\frac{\varepsilon^2 N}{(M + \sigma\ell\log n)^\ell}\right)$$

Hence, to union bound over all $n^\ell$ events $N = O\left(\varepsilon^{-2}(\ell \log nM)^\ell\right)$ suffices. ■

## C.4 Recovering Weights in Gaussian Mixtures

We now show how we can approximate upto a small error the weight $w_i$ of a gaussian components in a mixture of gaussians, when we have good approximations to $w_i\mu_i^{\otimes\ell}$ and $w_i\mu_i^{\otimes(\ell-1)}$.

**Lemma C.4** (Recovering Weights)**.** *For every $\delta' > 0, w > 0, L_{\min} > 0, \ell \in \mathbb{N}$, $\exists\delta = \Omega\left(\frac{\delta_1 w^{1/(\ell-1)}}{\ell^2 L_{\min}}\right)$ such that, if $\mu \in \mathbb{R}^n$ be a vector with length $\|\mu\| \geq L_{\min}$, and suppose*

$$\|v - w^{1/\ell}\mu\| < \delta \quad \text{and } \|u - w^{1/(\ell-1)}\mu\| < \delta.$$

*Then,*

$$\left|\left(\frac{|\langle u, v\rangle|}{\|u\|}\right)^{\ell(\ell-1)} - w\right| < \delta' \tag{17}$$

**Proof:** From (C.4) and triangle inequality, we see that

$$\|w^{-1/\ell}v - w^{-1/(\ell-1)}u\| \leq \delta(w^{-1/(\ell)} + w^{-1/(\ell-1)}) = \delta_1.$$

Let $\alpha_1 = w^{-1/(\ell-1)}$ and $\alpha_2 = w^{-1/\ell}$. Suppose $v = \beta u + \varepsilon\tilde{u}_\perp$ where $\tilde{u}_\perp$ is a unit vector perpendicular to $u$. Hence $\beta = \langle v, u\rangle/\|u\|$.

$$\begin{aligned}
\|\alpha_1 v - \alpha_2 u\|^2 = \|(\beta\alpha_1 - \alpha_2)u + \alpha_1\varepsilon\tilde{u}_\perp\| < \delta_1^2 \\
(\beta\alpha_1 - \alpha_2)^2\|u\|^2 + \alpha_1^2\varepsilon^2 \leq \delta_1^2 \\
\left|\beta - \frac{\alpha_2}{\alpha_1}\right| < \frac{\delta_1}{L_{\min}}
\end{aligned}$$

Now, substituting the values for $\alpha_1, \alpha_2$, we see that

$$\left|\beta - w^{\frac{1}{(\ell-1)} - \frac{1}{\ell}}\right| < \frac{\delta_1}{L_{\min}}.$$

$$\begin{aligned}
\left|\beta - w^{1/(\ell(\ell-1))}\right| < \frac{\delta}{w^{1/(\ell-1)}L_{\min}} \\
\left|\beta^{\ell(\ell-1)} - w\right| \leq \delta' \quad \text{when } \delta \ll \frac{\delta' w^{1/(\ell-1)}}{\ell^2 L_{\min}}
\end{aligned}$$

■