JMLR: Workshop and Conference Proceedings vol 35:1–37, 2014

Uniqueness of Tensor Decompositions with Applications to Polynomial Identifiability

Aditya Bhaskara BHASKARA@CS.PRINCETON.EDU
Google Research NYC

Moses Charikar MOSES@CS.PRINCETON.EDU
Princeton University

Aravindan Vijayaraghavan ARAVINDV@CS.PRINCETON.EDU
Carnegie Mellon University

Abstract

We give a robust version of the celebrated result of Kruskal on the uniqueness of tensor decompositions: given a tensor whose decomposition satisfies a robust form of Kruskal's rank condition, we prove that it is possible to approximately recover the decomposition if the tensor is known up to a sufficiently small (inverse polynomial) error.

Kruskal's theorem has found many applications in proving the *identifiability* of parameters for various latent variable models and mixture models such as Hidden Markov models, topic models etc. Our robust version immediately implies identifiability using only polynomially many samples in many of these settings – an essential first step towards efficient learning algorithms.

Our methods also apply to the "overcomplete" case, which has proved challenging in many applications. Given the importance of Kruskal's theorem in the tensor literature, we expect that our robust version will have several applications beyond the settings we explore in this work.

Keywords: Kruskal uniqueness theorem, tensor decomposition, latent variable models

1. Introduction

Statisticians have long studied the identifiability of probabilistic models, i.e., whether the parameters of a model can be learned from data (Teicher (1961, 1967); Tallis and Chesson (1982)). A central question in unsupervised learning is the efficient computation of such latent model parameters from observed data. The method of moments approach, pioneered by Pearson (1894), infers model parameters from empirical moments. This approach is very powerful and is used in many forms in applications. However in general, the method might require very high order moments and exponential sample complexity (Moitra and Valiant (2010); Belkin and Sinha (2010); Gravin et al. (2012)). Our focus in this work is to understand settings where moments of small order and polynomial sample complexity suffices.

Tensor decompositions have proved to be a valuable tool for reasoning about identifiability of probabilistic models in the algebraic statistics literature (Allman et al. (2009, 2011); Rhodes and Sullivant (2012)). Moments of data (which can be estimated empirically) are naturally represented by tensors (high dimensional analogs of matrices) and the low rank decomposition of such tensors can be used to deduce the parameters of the underlying model. A fundamental result of Kruskal (1977) on the uniqueness of tensor decompositions implies that the model parameters are uniquely identified by this procedure. In applications to learning theory, there are two issues that arise: first,

we cannot compute moments exactly (typically, using polynomial number of samples incurs an inverse polynomial error). Second, the modeling of the underlying problem (as a mixture model, say) is usually an approximate one, true up to some error. Both these issues result in small errors in our estimates for the entries of the tensor (i.e., the moments), as given by the model. This raises the questions: *Can we approximately recover the parameters given an approximation to the moment tensor? How small should the error be for such recovery to be possible?*

Our main technical contribution is establishing such a robust version of Kruskal's classic uniqueness theorem for tensor decompositions, tolerating inverse polynomial error. This directly implies *polynomial identifiability* (identifiability with polynomial samples) in a host of applications where Kruskal's theorem was used to prove identifiability assuming access to exact moment tensors (Allman et al. (2009)). To the best of our knowledge, no such robust version of Kruskal's theorem is known in the literature. Given the importance of this theorem in the tensor literature, we expect that this robust version will have applications beyond the settings we explore in this work.

We illustrate applications to learning several latent variable models. In particular, our results imply polynomial sample complexity bounds for learning multi-view mixture models, exchangeable (single) topic models, Hidden Markov Models, and mixtures of spherical Gaussians without separation assumptions.[1] These results also hold in the *overcomplete* setting, for which such bounds have often proved difficult.

1.1. Tensors, Kruskal's theorem, and our results

Tensors are mutlidimensional arrays – a generalization of vectors and matrices – which have been studied intensively as methods of extracting structure from data. The *low-rank decomposition* of a tensor often provides valuable insights into the structure of the data used to generate it. In sharp contrast to matrices, where a matrix of rank R (> 1) can be expressed in many ways as a sum of R rank-one matrices, higher order tensors typically have a *unique* decomposition up to much higher ranks (R roughly $c \cdot$ (dimension), as we will see). The classic result of Kruskal (1977) gives a sufficient condition for uniqueness, which has since found numerous applications.

Let us start with three dimensions. Suppose that a 3-tensor T has the following decomposition:[2]

$$T = [A\ B\ C] \equiv \sum_{r=1}^{R} A_r \otimes B_r \otimes C_r \tag{1}$$

Let the Kruskal rank or K-rank k_A of matrix A (formed by column vectors A_r) be the maximum value of k such that any k columns of A are linearly independent. k_B and k_C are similarly defined. Kruskal's result says that a sufficient condition for the uniqueness of the decomposition (1) is

$$k_A + k_B + k_C \geq 2R + 2 \tag{2}$$

Several alternate proofs of this fundamental result have been given (Jiang and Sidiropoulos (2004); Stegeman and Sidiropoulos (2007); Rhodes (2010); Landsberg (2012)). Sidiropoulos and Bro (2000) also extended this result to ℓ-order tensors.

We give robust versions of Kruskal's uniqueness theorem and its higher dimensional generalization. To this end, we need a natural robust analogue of Kruskal rank: we say that K-rank$_\tau(A) \geq k$

1. In this submission, we will focus on multi-view mixture models and Hidden Markov Models.
2. For $a, b, c \in \mathbb{R}^n$, $a \otimes b \otimes c$ is a *rank-one* tensor (dim $n \times n \times n$) whose j, k, l'th entry is $a_j b_k c_l$.

if every submatrix of A formed by k of its columns has minimum singular value at least $1/\tau$. (Note that this is related to, but much weaker than the Restricted Isometry Property (RIP)). A matrix is called bounded if its column vectors have bounded length, and we call two matrices (or tensors) close if the Frobenius norm of the difference is small. (See Section 2 for precise definitions.)

Our main result (for three dimensions, Theorem 5) can then be stated informally as follows:

Informal Theorem. *If any order 3 tensor T has a bounded rank R decomposition $[A\ B\ C]$, where the robust K-rank$_\tau$ k_A, k_B, k_C satisfy $k_A + k_B + k_C \geq 2R + 2$, then any decomposition $[A'\ B'\ C']$ that produces a tensor ε-close to T has A', B', C' being individually ε'-close to A, B and C respectively (up to permutation and re-scaling) for $\varepsilon < \varepsilon' \cdot poly(R, n, \tau)$.*

A similar result also holds for higher order tensors. For order ℓ tensors, a decomposition consists of $n \times R$ matrices $U^{(1)}, U^{(2)}, \ldots, U^{(\ell)}$, and if k_i denotes the robust K-rank$_\tau$ of $U^{(i)}$, then a sufficient condition for uniqueness (in the sense above) is

$$k_1 + k_2 + \cdots + k_\ell \geq 2R + (\ell - 1). \tag{3}$$

One way to interpret this result (as well as Kruskal's original theorem) is as saying that a *typical* tensor (in a probabilistic sense) of dimension $n^{\times \ell}$ and rank $R \leq \ell(n-1)/2$ has a unique decomposition. This is because the corresponding $n \times R$ matrices $U^{(i)}$ will have K-rank$_\tau = n$.

Discussion: Kruskal type results raise a natural question: can we show uniqueness of decomposition for super linear ranks? Note that a typical order ℓ tensor has rank $\Omega(n^{\ell-1})$, so there is indeed a large gap. It is known that Kruskal's rank conditions are best possible, however other assumptions could also imply uniqueness. Indeed for $\ell = 3$, algebraic geometry approaches (Chiantini and Ottaviani (2012)) show that *generic* (appropriately defined) tensors of rank up to $n^2/16$ have a unique decomposition. Obtaining a robust version of these results is a very interesting open problem. When $\ell \geq 5$, our claims can be strengthened if we do not assume worst case parameters. For instance, our uniqueness proof for higher dimensions (together with several ideas from random matrix theory) can be used to prove that tensors with a randomly perturbed decomposition[3] have unique decompositions for ranks up to $n^{\lfloor \frac{\ell-1}{2} \rfloor}$ (Bhaskara et al. (2014)).

Algorithms. Given the uniqueness results, it is natural to ask if there are efficient algorithms for recovering the decomposition under Kruskal's conditions. Such an algorithm can be used to recover the hidden parameters in various mixture models by estimating moments. This is a challenging open problem when the rank is $(1+\varepsilon)n$. For rank $\leq n$, and in particular when the decomposition matrices A, B, C have full column rank (any two of them being full rank also suffices), there are several algorithms which help compute the decomposition. The work of Harshman (1970) and that of Leurgans et al. (1993) give algorithms in this case. Anandkumar et al. (2012a) gave a power iteration type algorithm which is particularly efficient in practice.

Can our methods yield algorithms for the overcomplete case (rank $> n$)? In the recent work of Bhaskara et al. (2014), it was shown how some of our tools (such as the Khatri-Rao product) can be used for decomposition in the overcomplete case, provided we have access to higher order tensors. This is also in the spirit of the so-called FOOBI algorithm of De Lathauwer et al. (2007), which looks at fourth order moments, and the work of Goyal et al. (2014).

3. With high probability in a smoothed analysis setting.

For completeness, in appendix B, we give a simple (and general) SVD based algorithm (Theorem 24) to find low-rank tensor approximation, albeit in time exponential in the rank.[4] This can be viewed as a tensor analog of low-rank approximation, which is very well-studied for matrices.

Informal Theorem. *Given a tensor with a bounded, rank R decomposition up to an error ε, we can find a rank R approximation with error $O(\varepsilon)$ in time* $\exp(R^2 \log(n/\varepsilon))poly(n)$.

1.2. Applications to Latent Variable Models

Kruskal's uniqueness theorem has found a variety of applications in statistics and other fields. A robust version seems implicitly required in any application in which tensors are not given exactly. We will present some applications to learning, and in particular to the question of learning parameters in latent variable models.

Specifically, we will discuss multi-view models and Hidden Markov Models (HMMs), both which have been used extensively. Until very recently, the sample complexity of learning the parameters was not known in the overcomplete setting (rank $>$ dimension). The recent works of Bhaskara et al. (2014), Goyal et al. (2014) and Anderson et al. (2013) gave efficient algorithms (which imply sample complexity bounds) for these and related models in a smoothed analysis framework.

Our work implies polynomial sample complexity (even in the overcomplete case) whenever the parameters satisfy a certain Kruskal rank condition. This condition seems reasonable in practice (e.g., it also holds in a smoothed analysis framework). We will formally state the identifiability results for both models in Section 4. Below, we state the result for multi-view models, which illustrates the general 'template' of our results.

Multi-view models are mixture models with a discrete latent variable $h \in [R]$, such that $\Pr[h = r] = w_r$, for some *mixture weights* w_r (that form a probability distribution on $[R]$). We are given multiple observations or views $x^{(1)}, x^{(2)}, \ldots, x^{(\ell)}$ that are conditionally independent given the latent variable h, with $\mathbb{E}\left[x^{(j)} | h = r\right] = \mu_r^{(j)}$. Let $M^{(j)}$ be the $n \times R$ matrix whose columns are the means $\{\mu_r^{(j)}\}_{r \in [R]}$. The goal is to learn the matrices $\{M^{(j)}\}_{j \in [\ell]}$ and the weights $\{w_r\}_{r \in [R]}$.

Multi-view models are very expressive, and capture many well-studied models like Topic Models, Hidden Markov Models (HMMs), and random graph mixtures (Mossel and Roch (2006); Allman et al. (2009); Anandkumar et al. (2012c)). The techniques developed for this class have also been applied to phylogenetic tree models and certain tree mixtures (Chang (1996); Mossel and Roch (2006); Anandkumar et al. (2012b)).

Results. Suppose the dimension of the observations (n) is δR where δ is a small positive constant and R is the size of the range of the hidden variable, and hence the rank of the associated tensors. Then the result (formal version in Section 4) is the following:

Informal Theorem. *For a multi-view model with R topics or distributions, such that each of the parameter matrices $M^{(j)}$ has robust K-rank of at least δR for some constant δ, we can learn these parameters upto error ε with high probability using $poly_\delta(n, R)$ samples.*

The proofs follow from the fact that polynomially many samples suffice to estimate the ℓth moment tensor up to any inverse polynomial accuracy, followed by applying our robust uniqueness result the corresponding tensor (here $\ell = \lceil 2/\delta \rceil + 1$).

4. Our goal here is to show polynomial bounds on the sample complexity, thus we do not care about the run time.

1.3. Overview of Techniques

Robust Uniqueness of Tensor decompositions. Our proof broadly follows the outline of Kruskal's original proof (Kruskal (1977)): It proceeds by first establishing a *permutation lemma*, which gives sufficient conditions for concluding that the columns of two matrices are permutations of each other (up to scaling). Given two decompositions $[A\ B\ C]$ and $[A'\ B'\ C']$ for the same tensor, it is shown that A, A' satisfy the conditions of the lemma, and thus are permutations of each other (so also for B, C). Finally, it is shown that the three permutations for A, B and C (respectively) are identical.

The main challenge in adapting this proof is proving a robust version of the permutation lemma. The (robust) permutation lemma needs to establish that for every column of A', there is some column of A (close to being) parallel to it. Kruskal's proof uses downward induction to establish the following claim: for every set of $i \leq$ K-rank columns of A', there are at least i columns of A that are in the span of the chosen column vectors. The downward induction infers this by considering the intersection of columns that are close to $i+1$ dimensional spaces.

The natural analogue of this approach would be to consider columns of A which are "ε-close" to the span of i columns of A'. However, the inductive step involves considering combinations and intersections of the different spans that arise, and such arguments do not seem very tolerant to noise. In particular, we lose a factor of τn in each iteration, i.e., if the statement was true for $i+1$ with error ε_{i+1}, it will be true for i with error $\varepsilon_i = \tau n \cdot \varepsilon_{i+1}$. Since k steps of downward induction need to be unrolled, we recover a robust permutation lemma only when the error $< 1/(\tau n)^k$ to start with, which is exponentially small since k is typically $\Theta(n)$.

We overcome this issue by using a careful mix of combinatorial and linear algebra arguments: instead of keeping track of sets of vectors close to the span of i columns, we maintain an intersection of certain sets of vectors, and use the observation that they form a sunflower set system to obtain the desired bound on the size. This allows us to avoid losing any error in the recursion. We describe this in detail in Section 3.2. To carry forth this argument we crucially rely on the fact that A' is also "well-conditioned", which we need to establish initially (and is interesting in its own right).

Uniqueness for higher order tensors The idea here is to "combine the modes". Suppose we have a 4th order tensor $[ABCD] = \sum_{i=1}^{R} A_i \otimes B_i \otimes C_i \otimes D_i$ (and for simplicity suppose each matrix is $n \times R$). Now suppose we view $(C_i \otimes D_i)$ as an n^2 dimensional vector E_i, then what can we say about the robust K-rank of the $n^2 \times R$ dimensional matrix E (with columns E_i)? This notion is called the *Khatri-Rao Product*, and we can show that (robustly)

$$\text{K-rank}(E) \geq \text{K-rank}(C) + \text{K-rank}(D) - 1.$$

While this is tight in the worst case, it can be improved under stronger assumptions – and this increases the rank R for which we obtain uniqueness. As mentioned earlier, this forms the basis for the recent work of Bhaskara et al. (2014).

2. Notation and Preliminaries

We start with basic notation on tensors which we will use throughout the paper.

Tensors are higher dimensional arrays. An ℓth order, or ℓ-dimensional tensor is an element in $\mathbb{R}^{n_1 \times n_2 \times \cdots \times n_\ell}$, for positive integers n_i. The various "dimensions" $n_1, n_2, \ldots$ are referred to as the *modes* of the tensor.

While tensors have classically been defined over complex numbers for certain applications, we will consider only real tensors here. We now define the *rank* of a tensor. Firstly, a rank-1 tensor as a product $a^{(1)} \otimes a^{(2)} \otimes \ldots \otimes a^{(\ell)}$, where $a^{(i)}$ is an n_i dimensional vector.

Definition 1 (Tensor rank, Rank R decomposition) *The* rank *of a tensor $T \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_\ell}$ is defined to be the smallest R for which there exist R rank-1 tensors $T^{(i)}$ whose sum is T.*

A rank-R decomposition of T is given by a set of matrices $U^{(1)}, U^{(2)}, \ldots, U^{(\ell)}$ with $U^{(i)}$ of dimension $n_i \times R$, such that we can write $T = [U^{(1)}\ U^{(2)}\ \ldots\ U^{(\ell)}]$, which is defined by $[U^{(1)}\ U^{(2)}\ \ldots\ U^{(\ell)}] := \sum_{r=1}^{R} U_r^{(1)} \otimes U_r^{(2)} \otimes \ldots \otimes U_r^{(\ell)}$*, where A_r to denotes the rth column of A.*

Third order tensors (or 3-tensors) play a central role in understanding properties of tensors in general (as in many other areas of mathematics, the jump in complexity occurs most dramatically when we go from two to three dimensions, in this case from matrices to 3-tensors). For 3-tensors, we often write the decomposition as $[A\ B\ C]$, where A, B, C have dimensions n_A, n_B, n_C respectively.

Definition 2 (ε-close, ρ-bounded) *Two tensors T_1 and T_2 are said to be ε-close if the Frobenius norm of the difference is small, i.e., $\|T_1 - T_2\|_F \leq \varepsilon$. We will sometimes write this as $T_1 =_\varepsilon T_2$.*

An $n \times R$ matrix A is said to be ρ-bounded if each of the columns has length at most ρ, for some parameter ρ. A tensor $[U^{(1)}\ U^{(2)}\ \ldots\ U^{(\ell)}]$ is called $(\rho_1, \rho_2, \ldots, \rho_\ell)$-bounded if the matrix $U^{(i)}$ is ρ_i bounded for all i.

Unless mentioned specifically, the errors in the paper will be ℓ_2 (or Frobenius norm, which is the square root of the sum of squares of entries in a matrix/tensor), since they add up conveniently.

We next define the notion of Kruskal rank, and its robust counterpart.

Definition 3 (Kruskal rank, K-rank$_\tau(.)$) *Let A be an $n \times R$ matrix. The K-rank (or Kruskal rank) of A is the largest k for which* every *set of k columns of A is linearly independent.*

*Let τ be a parameter. The τ-*robust *k-rank is denoted by K-rank$_\tau(A)$, and is the largest k for which every $n \times k$ sub-matrix $A_{|S}$ of A has $\sigma_k(A_{|S}) \geq 1/\tau$.*

Note that we only have a lower bound on the (kth) smallest singular value of A, and not for example the condition number $\sigma_{\max}/\sigma_k$. This is because we will usually deal with matrices that are also ρ-bounded, so such a bound will automatically hold, but our definition makes the notation a little cleaner. Another simple linear algebra definition we use is the following

Definition 4 (ε-close to a space) *Let V be a subspace of $\mathbb{R}^n$, and let Π be the projection matrix onto V. Let $u \in \mathbb{R}^n$. We say that u is ε-close to V if $\|u - \Pi u\| \leq \varepsilon$.*

Other notation. For $z \in \mathbb{R}^d$, diag(z) is the $d \times d$ diagonal matrix with the entries of z occupying the diagonal. For a vector $z \in \mathbb{R}^d$, $nz(z)$ denotes the number of non-zero entries in z. Further, $nz_\varepsilon(z)$ denotes the number of entries of magnitude $\geq \varepsilon$. As is standard, we denote by $\sigma_i(A)$ the ith largest singular value of a matrix A. Also, we abuse the notation of $\otimes$ at times, with $u \otimes v$ sometimes referring to a matrix of dimension $dim(u) \times dim(v)$, and sometimes a $dim(u) \cdot dim(v)$ vector. This will always be clear from context. Finally, poly$_\delta(n)$ refers to any polynomial in the parameter n, with $\delta > 0$ being treated as a constant.

Normalization. To avoid complications due to scaling, we will assume that our tensors are scaled such that all the $\tau_A, \tau_B, \ldots,$ are ≥ 1 and $\leq$ poly(n). So also, our upper bounds on lengths $\rho_A, \rho_B, \ldots$ are all assumed to be between 1 and some poly(n).

Error polynomials. We will, in many places, encounter statements such as "if $Q_1 \leq \varepsilon$, then $Q_2 \leq (3n^2\gamma)\cdot\varepsilon$", with polynomials ϑ (in this case $3n^2\gamma$) involving the variables $n, R, k_A, k_B, k_C, \tau, \rho, \ldots$. In order to keep track of these, we use the notation $\vartheta_1, \vartheta_2, \ldots$. Sometimes, to refer to a polynomial introduced in Lemma 5, for instance, we use ϑ_5. Unless specifically mentioned, they will be polynomials in the parameters mentioned above, so we do not mention them each time.

3. Uniqueness of Tensor Decompositions

First we consider third order tensors and prove our robust uniqueness theorem for 3-tensors (Sections 3.1 and 3.3). Our proof broadly follows along the lines of Kruskal's original proof of the uniqueness theorem Kruskal (1977). The key ingredient, which is a robust version of the so-called *permutation lemma* is presented in Section 3.2, since it seems interesting in its own right. Finally we will see how to reduce the case of higher order tensors, to that of third order tensors (Section 3.4).

3.1. Uniqueness Theorem for Third Order Tensors

Theorem 5 (Unique Decompositions) *Suppose a rank-R tensor $T = [A\ B\ C]$ is (ρ_A, ρ_B, ρ_C)-bounded, with K-rank$_{\tau_A}(A) = k_A$, K-rank$_{\tau_B}(B) = k_B$, K-rank$_{\tau_C}(C) = k_C$ satisfying $k_A + k_B + k_C \geq 2R + 2$. Then for every $0 < \varepsilon' < 1$, there exists*

$$\varepsilon = \varepsilon' / \big(R^6 \vartheta_5(\tau_A, \rho_A, \rho'_A, n_A)\vartheta_5(\tau_B, \rho_B, \rho'_B, n_B)\vartheta_5(\tau_C, \rho_C, \rho'_C, n_C)\big),$$

for some polynomial ϑ_5 such that for any other $(\rho'_A, \rho'_B, \rho'_C)$-bounded decomposition $[A'\ B'\ C']$ of rank R that is ε-close to $[A\ B\ C]$, there exists an $(R \times R)$ permutation matrix Π and diagonal matrices $\Lambda_A, \Lambda_B, \Lambda_C$ such that

$$\|\Lambda_A\Lambda_B\Lambda_C - I\|_F \leq \varepsilon' \text{ and } \quad \big\|A' - A\Pi\Lambda_A\big\|_F \leq \varepsilon' \quad (\text{similarly for } B \text{ and } C) \tag{4}$$

Eq. (4) says that A, B, C are scaled permutations of A', B', C' respectively, and that the scalings in each term multiply to one (approximately). The proof broadly has two parts. First, we prove that if $[A, B, C] \approx [A', B', C']$, then A is essentially a permutation of A', B of B', and C of C'. Second, we prove that the permutations in the (three) different "modes" (or dimensions) are indeed equal. Let us begin by describing a lemma which is key to the first step.

The Permutation Lemma This is the core of Kruskal's argument for the uniqueness of tensor decompositions. Given two matrices X and Y, how does one conclude that the columns are permutations of each other? Kruskal gives a very clever sufficient condition, involving looking at *test vectors* w, and considering the number of non-zero entries of $w^T X$ and $w^T Y$. The intuition is that if X and Y are indeed permutations, these numbers are precisely equal for all w.

More precisely, suppose X, Y are $n \times R$ matrices of rank k. Let $nz(x)$ denote the number of non-zero entries in a vector x. The lemma then states that if for all w, we have

$$nz(w^T X) \leq R - k + 1 \implies nz(w^T Y) \leq nz(w^T X),$$

then the matrices X and Y have columns which are permutations of each other up to a scaling. That is, there exists an $R \times R$ permutation matrix Π, and a diagonal matrix Λ s.t. $Y = X\Pi\Lambda$. We prove a robust version of this lemma, stated as follows (recall the definition of $nz_\varepsilon(.)$, Section 2)

Lemma 6 (Robust permutation lemma) *Suppose X, Y are ρ-bounded $n \times R$ matrices such that K-rank$_\tau(X)$ and K-rank$_\tau(Y)$ are $\geq k$, for some integer $k \geq 2$. Further, suppose that for $\varepsilon < 1/\vartheta_6$, the matrices satisfy:*

$$\forall w \text{ s.t. } \quad nz(w^T X) \leq R - k + 1, \text{ we have } nz_\varepsilon(w^T Y) \leq nz(w^T X), \tag{5}$$

then there exists an $R \times R$ permutation matrix Π, and a diagonal matrix Λ s.t. X and Y satisfy $\|X - Y\Pi\Lambda\|_F < \vartheta_6 \cdot \varepsilon$. In fact, we can pick $\vartheta_6 := (nR^2)\vartheta_{12}$.

Remark 7 *To see why this condition involving $nz_\varepsilon(\cdot)$ helps, let us imagine that $nz_\varepsilon(w^T Y) \leq nz(w^T X)$* for all *$w$. Then considering a random w easily shows that X and Y must have columns that are permutations of each other up to scaling. However, as we will soon see in Lemma 8, we only have this condition for those w with $nz(w^T X) \geq R - k + 1$.*

A key component in the proofs that follow is to view the three-dimensional tensor $[A\ B\ C]$ as a bunch of *matrix slices*, and argue about the ranks of weighted combinations of these slices. One observation, which follows from the Cauchy-Schwarz inequality, is the following: if $[A\ B\ C] =_\varepsilon [A'\ B'\ C']$, then by taking a combination of "matrix" slices along the third mode (with weights given by $x \in \mathbb{R}^{n_C}$,

$$\forall x \in \mathbb{R}^{n_C}, \quad \left\| A \operatorname{diag}(x^T C)\ B^T - A' \operatorname{diag}(x^T C')\ (B')^T \right\|_F^2 \leq \varepsilon^2 \|x\|_2^2 . \tag{6}$$

We now state the key technical lemma which allows us to verify that the hypotheses of Lemma 6 hold. It says for any $k_C - 1$ vectors of C' there are at least as many columns of C which are close to the span of the chosen columns from C'.

Lemma 8 *Suppose A, B, C, A', B', C' satisfy the conditions of Theorem 5, and suppose $[A\ B\ C] =_\varepsilon [A'\ B'\ C']$. Then for any unit vector x, we have*

$$\forall \varepsilon', \quad nz_{\varepsilon'}(x^T C') \leq R - k_C + 1 \implies nz_{\varepsilon''}(x^T C) \leq nz_{\varepsilon'}(x^T C')$$

for $\varepsilon'' = \vartheta_8 \cdot (\varepsilon + \varepsilon')$, where $\vartheta_8 := 4R^3(\tau_A\tau_B\tau_C)^2\rho_A\rho_B\rho_C(\rho'_A\rho'_B\rho'_C)^2$.

Remark 9 *This lemma, together with its corollary Lemma 10 will imply the conditions of the permutation lemma. While the proof of the robust permutation lemma (Lemma 6) will directly apply this Lemma with $\varepsilon' = 0$, we will need the $\varepsilon' > 0$ case for establishing Lemma 10 that lets us conclude that K-rank$_{\tau\vartheta}(C') \geq$ K-rank$_\tau(C)$ for some error polynomial ϑ. This is essential in our proof of the robust permutation lemma, and it also has other implications, as we will see.*

The proof of the lemma is quite involved and tricky – we defer the proof to the appendix D.1). The next lemma uses the above to conclude that K-rank$_{\vartheta\tau}(C') \geq$ K-rank$_\tau(C)$, for some polynomial ϑ. i.e. if T has a well-conditioned decomposition which satisfies the Kruskal conditions, then any other bounded decomposition that approximates T sufficiently well should also be well-conditioned.

Lemma 10 *Let A, B, C, A', B', C' be as in the setting of Theorem 5. Suppose $[A\ B\ C] =_\varepsilon [A'\ B'\ C']$, with $\varepsilon < 1/\vartheta_{10}$, where $\vartheta_{10} = R\tau_A\tau_B\tau_C\vartheta_8 = 4R^4\tau_A^3\tau_B^3\tau_C^3\rho_A\rho_B\rho_C(\rho'_A\rho'_B\rho'_C)^2$. Then A', B', C' have K-rank$_{\tau'}$ to be at least k_A, k_B, k_C respectively, where $\tau' := \vartheta_{10}$.*

Proof We want to show that every $n_C \times k_C$ sub-matrix of C' has min. singular value $\sigma_{k_C} \geq \delta = 1/\tau'_C$. For contradiction, let C'_S be a $n_C \times k_C$ sub-matrix of C' and unit vector $z \in \mathbb{R}^{n_C}$ such that $\left\|z^T C'_S\right\|_2 < \delta$. Then, it is easy to see (from Lemma 8) that

$$\sum_{i \in S} \left\langle z, C'_i \right\rangle^2 < \delta^2 \implies nz_\delta(z^T C') \leq n_C - k_C \implies nz_{\varepsilon_1}(z^T C) \leq n_C - k_C$$

for $\varepsilon_1 = \vartheta_8(\varepsilon + \delta)$. Now, picking the sub-matrix of C given by the these k_C co-ordinates of $z^T C$ that are small, we can contradict K-rank$_{1/(R\varepsilon_1)}(C) \geq k_C$. ■

Let us check that the conditions of the robust permutation lemma hold with C', C taking the roles of X, Y in Lemma 6, and $k = k_C$, and $\tau = \vartheta_{10} \cdot \tau_C$. From Lemma 10, it follows that K-rank$_\tau(C)$ and K-rank$_\tau(C')$ are both $\geq k$, and setting $\varepsilon' = 0$ in Lemma 8, the other condition of Lemma 6 holds. Now, we proceed to prove the robust permutation lemma.

3.2. A Robust Permutation Lemma

Let us now prove the robust version of the permutation lemma (Lemma 6). Recall that K-rank$_\tau(X)$ and K-rank$_\tau(Y)$ are $\geq k$, and that the matrices X, Y are $n \times R$.

Kruskal's proof of the permutation lemma proceeds by induction. Roughly, he considers the span of some set of i columns of X (for $i < k$), and proves that there exist at least i columns of Y which lie in this span. The hypothesis of his lemma implies this for $i = k - 1$, and the proof proceeds by downward induction. Note that $i = 1$ implies for every column of X, there is at least one column of Y in its span. Since no two columns of X are *parallel*, and the number of columns is equal in X, Y, there must be precisely one column, and this completes the proof.

The natural way to mimic this proof, as mentioned in the introduction, accumulates errors in each inductive step. Thus the trick is to define the sets of columns differently. We start by introducing some notation. If V is a matrix and S a subset of the columns, denote by $\mathrm{span}(V_S)$ the span of the columns of V indexed by S. Now for $S \subseteq [R]$ of size $(k-1)$, we define T_S to be the set of indices corresponding to columns of Y which are ε_1-close to $\mathrm{span}(X_S)$, where $\varepsilon_1 := (nR)\varepsilon$, and ε is as defined in the statement of Lemma 6. For smaller sets S (and this definition is crucial to avoiding an accumulation of errors), we define: $T_S := \bigcap_{|S'|=(k-1), S' \supset S} T_{S'}$.

The main inductive claim will be that for every $S \subseteq [R]$ of size $\leq (k-1)$, we have $|T_S| = |S|$. Suppose we have this claim for a singleton, say $S = \{i\}$. Now if y is a column of Y which is in $\mathrm{span}(X_{S'})$ for all $(k-1)$ element subsets S' (of $[R]$) which contain i, by Lemma 12 which we will prove (applied with $A = \{i\}$ and B being any set of size $(k-1)$ not containing i), we will obtain that y is $\varepsilon_1 \cdot \vartheta_{12}$-close to $\mathrm{span}(X_{\{i\}})$, completing the proof of the permutation lemma. Thus it remains to show the inductive claim. The base case is the following, proved in Appendix C.1

Lemma 11 *In the above notation, for any $S \subseteq [R]$ of size $k-1$, $|T_S|$ is precisely $k-1$.*

The next two lemmas are crucial to the analysis. The first is our main linear algebraic lemma, and the second is a counting argument which lies at the heart of the proof. It is stated in the language of sunflower set systems.

Lemma 12 *Let X be a matrix as above, with K-rank$_\tau(X) \geq k$. Let $A, B \subseteq [R]$, with $|B| = q$ and $A \cap B = \emptyset$. For $1 \leq i \leq q$, define T_i to be the union of A with all elements of B except the ith one (when indexed in some way). Suppose further that $|A| + |B| \leq k$. Then if $y \in \mathbb{R}^n$ is ε-close to $\mathrm{span}(X_{T_i})$ for each i, it is in fact $\vartheta_{12} \cdot \varepsilon$ close $\mathrm{span}(X_A)$, where $\vartheta_{12} := 4n\tau\rho$.*

Definition 13 (Sunflower set system) *A set system $\mathcal{F}$ is said to be a "sunflower on $[R]$ with core T^*" if $\mathcal{F} \subseteq 2^{[R]}$, and for any $F_1, F_2 \in \mathcal{F}$, we have $F_1 \cap F_2 \subseteq T^*$.*

Lemma 14 *Let $\{T_1, T_2, \ldots, T_q\}$, $q \geq 2$, be a sunflower on $[R]$ with core T^*, and suppose $|T_1| + |T_2| + \cdots + |T_q| \geq R + (q-1)\theta$, for some θ. Then we have $|T^*| \geq \theta$, and furthermore, equality occurs iff $T^* \subseteq T_i$ for all $1 \leq i \leq q$.*

The proofs of the lemmas are deferred to Appendix C.1. With these lemmas in place, we can prove the main inductive claim.

Proof [Proof of Lemma 6.] We need to prove the following inductive claim:

Claim. For every $S \subseteq [R]$ of size $\leq (k-1)$, we have $|T_S| = |S|$.

We show this by downward induction on $|S|$, for which the base case $|S| = k-1$ is proved in Lemma 11. Now consider some S of size $|S| \leq k-2$. W.l.o.g., we may suppose it is $\{R - |S| + 1, \ldots, R\}$. Let W_i denote $T_{S \cup \{i\}}$, for $1 \leq i \leq R - |S|$, and let us write $q = R - |S|$. By the inductive hypothesis, $|W_i| \geq |S| + 1$ for all i.

Let us define T^* to be the set of indices of the columns of Y which are $\varepsilon_1 \cdot \vartheta_{12}$-close to $\mathrm{span}(X_S)$. We claim that $W_i \cap W_j \subseteq T^*$ for any $i \neq j \notin S$. This can be seen as follows: first note that $W_i \cap W_j$ is contained in the intersection of $T_{S'}$, where the intersection is over $S' \supset S$ such that $|S'| = k-1$, and S' contains either i or j. Now consider any $k - |S|$ element set B which contains both i, j (note $|S| \leq k-2$). The intersection above includes sets which contain S along with all of B except the rth element (indexed arbitrarily), for each r. Thus by Lemma 12, we have that $W_i \cap W_j \subseteq T^*$.

Thus the sets $\{W_1, \ldots, W_q\}$ form a sunflower family with core T^*. Further, we can check that the condition of Lemma 14 holds with $\theta = |S|$: since $|W_j| \geq |S| + 1$ by the inductive hypothesis, it suffices to verify that

$$R + (q-1)|S| \leq q(|S| + 1), \text{ which is true since } R = q + |S|.$$

Thus we must have $|T^*| \geq |S|$.

But now, note that T^* is defined as the columns of Y which are $\varepsilon_1 \cdot \vartheta_{12}$-close to $\mathrm{span}(X_S)$, and thus $|T^*| \leq |S|$ (by Lemma 29), and thus we have $|T^*| = |S|$. Now we have equality in Lemma 14, and so the 'furthermore' part of the lemma implies that $T^* \subseteq W_i$ for all i.

Thus we have $T_S = \bigcap_i W_i = T^*$ (the first equality follows from the definition of T_S), thus completing the proof of the claim, by induction.

Once we have the claim, the lemma follows by applying to singleton sets. ■

3.3. Wrapping up the proof

We are now ready to complete the robust Kruskal's theorem. From what we have seen (and the permutation lemma), we have that A is a permutation of A', B of B' and C of C' (with scaling).

So we only need to prove that these three permutations are in fact identical, and that the scalings multiply to the identity (up to small error). Here we can imitate Kruskal's argument, carefully making each step robust. We do this in appendix D.3.

3.4. Uniqueness Theorem for Higher Order Tensors

We show the uniqueness theorem for higher order tensors by a reduction to third order tensors. This is along the lines of Sidiropoulos and Bro (2000). This reduction will proceed inductively. We will convert an order ℓ tensor to a order $(\ell - 1)$ tensor by "combining" two of the modes of the tensor. This is done by using the Khatri-Rao product:

Definition 15 (Khatri-Rao product) *Given two matrices A (size $n_1 \times R$) and B (size $n_2 \times R$), the $(n_1 n_2) \times R$ matrix $M = A \odot B$ constructed with the i^{th} column equal to $M_i = A_i \otimes B_i$ (viewed as a vector) is called the Khatri-Rao product.*

The following lemma (proof in the appendix C.3) relates the robust K-rank of $A \odot B$ with the robust K-rank of A and B. This is the main tool in the proof of uniqueness in the general case, as mentioned in the outline.

Lemma 16 (K-rank of the Khatri-Rao product) *For two matrices A, B with R columns with robust K-rank $k_A =$ K-rank$_{\tau_1}(A)$ and $k_B =$ K-rank$_{\tau_2}(B)$, the K-rank of the Khatri-Rao product $M = A \odot B$ is super-additive:*

$$\textit{K-rank}_{(\tau_1\tau_2\sqrt{k_A+k_B})}(M) \geq \min\{k_1 + k_2 - 1, R\}.$$

Given this lemma, the proof of the uniqueness theorem goes as follows: suppose we have a decomposition $[A_1, A_2, \ldots, A_\ell]$ which satisfies the Kruskal rank condition

$$\text{K-rank}_\tau(A_1) + \text{K-rank}_\tau(A_2) + \cdots + \text{K-rank}_\tau(A_\ell) \geq 2R + \ell - 1,$$

then we can first show the uniqueness of the tensor with decomposition $[A_1 \odot A_2, A_3, \ldots, A_\ell]$ by induction (the inductive condition will hold because of Lemma 16), and then observe that $A_1 \odot A_2$ uniquely identifies A_1 and A_2. The last observation is straightforward, but a robust version introduces additional loss in the parameters.

The full details (and the formal statements) are presented in Section A.

4. Polynomial Identifiability of Latent Variable and Mixture Models

We now show how our robust uniqueness theorem can be used for learning latent variable models with polynomial sample complexity.

Definition 17 (Polynomial Identifiability) *An instance of a hidden variable model of size m with hidden variables set Υ is said to be polynomial identifiable if there is an algorithm that given any $\eta > 0$, uses only $N \leq$ poly$(m, 1/\eta)$ samples and finds with probability $1 - o(1)$ estimates of the hidden variables Υ' such that $\|\Upsilon' - \Upsilon\|_\infty < \eta$.*

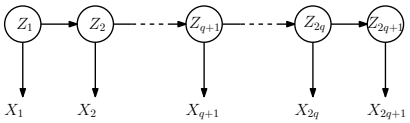

Figure 1: An HMM with $2q + 1$ time steps.

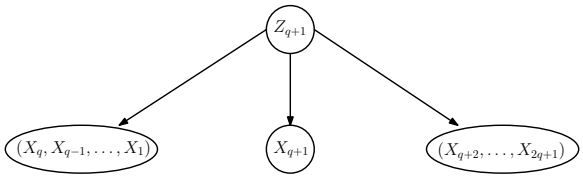

Figure 2: Embedding the HMM into the Multi-view model

We illustrate two results on polynomial identifiability implied by the robust Kruskal theorem. The first will be on Multi-view mixture models, and the second on Hidden Markov Models (HMMs).

We start by recalling the multi-view model defined in Section 1.2 (see F for a formal set up). To apply Kruskal's theorem, we need to estimate (accurately enough) a tensor that encodes the parameters of the model. This is done by the following lemma (proof is easy by conditional independence).

Lemma 18 (Allman et al. (2009); Anandkumar et al. (2012c)) *In the notation established above for multi-view models, $\forall \ell \in \mathbb{N}$ the ℓ^{th} moment tensor*

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

Since we can estimate the LHS up to any inverse polynomial accuracy using samples, using the robust Kruskal theorem, we immediately obtain the informal theorem on multi-view models (Section 1.2). We state and prove it formally in Appendix F, Theorem 37. We now move to the application of our results to Hidden Markov Models.

4.1. Hidden Markov Models

Hidden Markov Models are extensively used in speech recognition, image classification, bioinformatics etc. We follow the same setting as in Allman et al. (2009): there is a hidden state sequence $Z_1, Z_2, \ldots, Z_m$ taking values in $[R]$, that forms a stationary Markov chain $Z_1 \to Z_2 \to \cdots \to Z_m$ with transition matrix P and initial distribution $w = \{w_r\}_{r \in [R]}$ (assumed to be the stationary distribution). The observation X_t is from the set of discrete events[5] $\{1, 2, \ldots, n\}$ and it is represented by an indicator vector in $x^{(t)} \in \mathbb{R}^n$. Given the state Z_t at time t, X_t (and hence $x^{(t)}$) is conditionally independent of all other observations and states. The matrix M (of size $n \times R$) represents the probability distribution for the observations: the r^{th} column M_r represents the probability distribution conditioned on the state $Z_t = r$ i.e.

$$\forall r \in [R], \forall j \in [n], \quad \Pr\left[X_j = i | Z_j = r\right] = M_{ir}.$$

The HMM model described above is shown in Fig. 1.

Our result here states that we can recover the parameters of an HMM using polynomial many samples (see Corollary 38 for a formal statement).

5. In general, we can also allow x_t to be certain continuous distributions like multivariate gaussians.

Acknowledgments

We thank Ravi Kannan for valuable discussions about the algorithmic results in this work, and Daniel Hsu for helpful pointers to the literature. The third author would also like to thank Siddharth Gopal for some useful pointers about HMM models in speech and image recognition.

References

Elizabeth S Allman, Catherine Matias, and John A Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

Elizabeth S Allman, Sonia Petrovic, John A Rhodes, and Seth Sullivant. Identifiability of two-tree mixtures for group-based models. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(3):710–722, 2011.

Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *arXiv preprint arXiv:1210.7559*, 2012a.

Anima Anandkumar, Daniel Hsu, Furong Huang, and Sham Kakade. Learning mixtures of tree graphical models. In *Advances in Neural Information Processing Systems 25*, pages 1061–1069, 2012b.

Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden markov models. *arXiv preprint arXiv:1203.0683*, 2012c.

Joseph Anderson, Mikhail Belkin, Navin Goyal, Luis Rademacher, and James R. Voss. The more, the merrier: the blessing of dimensionality for learning large gaussian mixtures. *CoRR*, abs/1311.2891, 2013.

Saugata Basu, Richard Pollack, and Marie-Françoise Roy. On the combinatorial and algebraic complexity of quantifier elimination. *J. ACM*, 43(6):1002–1045, November 1996. ISSN 0004-5411. doi: 10.1145/235809.235813. URL http://doi.acm.org/10.1145/235809.235813.

Mikhail Belkin and Kaushik Sinha. Polynomial learning of distribution families. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 103–112. IEEE, 2010.

Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the 46th annual ACM Symposium on Theory of Computing (STOC)*. ACM, 2014.

Joseph T Chang. Full reconstruction of markov models on evolutionary trees: identifiability and consistency. *Mathematical biosciences*, 137(1):51–73, 1996.

L. Chiantini and G. Ottaviani. On generic identifiability of 3-tensors of small rank. *SIAM Journal on Matrix Analysis and Applications*, 33(3):1018–1037, 2012. doi: 10.1137/110829180.

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.

Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier pca. In *Proceedings of the 46th annual ACM Symposium on Theory of Computing (STOC)*. ACM, 2014.

Nick Gravin, Jean Lasserre, Dmitrii V Pasechnik, and Sinai Robins. The inverse moment problem for convex polytopes. *Discrete & Computational Geometry*, 48(3):596–621, 2012.

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

Tao Jiang and Nicholas D Sidiropoulos. Kruskal's permutation lemma and the identification of candecomp/parafac and bilinear models with constant modulus constraints. *Signal Processing, IEEE Transactions on*, 52(9):2625–2636, 2004.

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

J.M. Landsberg. *Tensors:: Geometry and Applications*. Graduate Studies in Mathematics Series. American Mathematical Society, 2012. ISBN 9780821869079. URL http://books.google.com.au/books?id=JTjv3DTvxZIC.

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.

Ankur Moitra and Gregory Valiant. Settling the polynomial learnability of mixtures of gaussians. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 93–102. IEEE, 2010.

Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden markov models. *The Annals of Applied Probability*, pages 583–614, 2006.

Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

John A Rhodes. A concise proof of kruskal's theorem on tensor decomposition. *Linear Algebra and Its Applications*, 432(7):1818–1824, 2010.

John A Rhodes and Seth Sullivant. Identifiability of large phylogenetic mixture models. *Bulletin of mathematical biology*, 74(1):212–231, 2012.

Nicholas D Sidiropoulos and Rasmus Bro. On the uniqueness of multilinear decomposition of n-way arrays. *Journal of chemometrics*, 14(3):229–239, 2000.

Alwin Stegeman and Nicholas D Sidiropoulos. On kruskal's uniqueness condition for the candecomp/parafac decomposition. *Linear Algebra and its applications*, 420(2):540–552, 2007.

GM Tallis and P Chesson. Identifiability of mixtures. *J. Austral. Math. Soc. Ser. A*, 32(3):339–348, 1982.

Henry Teicher. Identifiability of mixtures. *The annals of Mathematical statistics*, 32(1):244–248, 1961.

Henry Teicher. Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38(4):1300–1302, 1967.

Appendix A. Uniqueness Theorem for Higher Order Tensors

We show the uniqueness theorem for higher order tensors by a reduction to third order tensors as in Sidiropoulos and Bro (2000). This reduction will proceed inductively, i.e., the robust uniqueness of order ℓ tensors is deduced from that of order $(\ell-1)$ tensors. We will convert an order ℓ tensor to a order $(\ell-1)$ tensor by combining two of the components together (say last two) as a $n_{\ell-1}n_\ell$ dimensional vector ($U^{(\ell-1)} \otimes U^{(\ell)}$ say). This is precisely captured by the Khatri-Rao product of two matrices:

Definition 19 (Khatri-Rao product) *Given two matrices A (size $n_1 \times R$) and B (size $n_2 \times R$), the $(n_1 n_2) \times R$ matrix $M = A \odot B$ constructed with the i^{th} column equal to $M_i = A_i \otimes B_i$ (viewed as a vector) is the Khatri-Rao product.*

The following Lemma 20 (proof in the appendix C.3) relates the robust K-rank of $A \odot B$ with the robust K-rank of A and B. This turns out to be crucial to the proof of uniqueness in the general case, which we present right after.

Lemma 20 (K-rank of the Khatri-Rao product) *For two matrices A, B with R columns with robust K-rank $k_A = $ K-rank$_{\tau_1}(A)$ and $k_B = $ K-rank$_{\tau_2}(B)$, the K-rank of the Khatri-Rao product $M = A \odot B$ is super-additive:*

$$\text{K-rank}_{(\tau_1\tau_2\sqrt{k_A+k_B})}(M) \geq \min\{k_1 + k_2 - 1, R\}.$$

Theorem 21 (Uniqueness of Decompositions for Higher Orders) *Suppose we are given an order ℓ tensor (with $\ell \leq R$), $T = [U^{(1)}\ U^{(2)}\ \dots\ U^{(\ell)}]$, where $\forall j \in [\ell]$ the n_j-by-R matrix $U^{(j)}$ is ρ_j-bounded, with K-rank$_{\tau_j}(U^{(j)}) = k_j \geq 2$ satisfying*

$$\sum_{j=1}^{\ell} k_j \geq 2R + \ell - 1.$$

Then for every $0 < \varepsilon' < 1$, there exists $\varepsilon = \left(\vartheta_{21}^{(\ell)}\left(\frac{\varepsilon'}{R}\right)\right) \cdot \left(\prod_{j\in[\ell]} \vartheta_{21}(\tau_j, \rho_j, \rho'_j, n_j)\right)^{-1}$ such that, for any other $(\rho'_1, \rho'_2, \dots, \rho'_\ell)$-bounded decomposition $[V^{(1)}\ V^{(2)}\ \dots\ V^{(\ell)}]$ which is ε-close to T, there exists an $R \times R$ permutation matrix Π and diagonal matrices $\{\Lambda^{(j)}\}_{j\in[\ell]}$ such that

$$\left\| \prod_{j\in[\ell]} \Lambda^{(j)} - I \right\|_F \leq \varepsilon' \quad \text{and } \forall j \in [\ell],\ \left\| V^{(j)} - U^{(j)}\Pi\Lambda^{(j)} \right\|_F \leq \varepsilon' \tag{7}$$

Setting $\vartheta_{21}^{(\ell)}(x) = x^{2^\ell}$ and $\vartheta_{21}(\tau_j, \rho_j, \rho'_j, n_j) = (\tau_j \rho_j \rho'_j n_j)^{O(1)}$ suffice for the theorem.

Proof Outline. The proof proceeds by induction on ℓ. The base case is $\ell = 3$, and for higher ℓ, the idea is to reduce to the case of $\ell - 1$ by taking the Khatri-Rao product of the vectors in two of the dimensions. That is, if $[U^{(1)}\ U^{(2)}\ \dots\ U^{(\ell)}]$ and $[V^{(1)}\ V^{(2)}\ \dots\ V^{(\ell)}]$ are close, we conclude that $[U^{(1)}\ U^{(2)}\ \dots\ (U^{(\ell-1)} \odot U^{(\ell)})]$ and $[V^{(1)}\ V^{(2)}\ \dots\ (V^{(\ell-1)} \odot V^{(\ell)})]$ are close, and use the inductive hypothesis, which holds because of Lemma 20 we mentioned above. We then need an additional step to conclude that if $A \odot B$ and $C \odot D$ are close, then so are A, C and B, D up to some loss (Lemma 31 – this is where we have a *square root* loss, which is why we have a bad dependence on the ε' in the statement). We now formalize this outline.

Proof [Proof of Theorem 21] We will prove by induction on ℓ. The base case of $\ell = 3$ is established by Theorem 5. Thus consider some $\ell \geq 4$, and suppose the theorem is true for $\ell - 1$. Furthermore, suppose the parameters ε and ε' in the statement of Theorem 21 for $(\ell - 1)$ be $\varepsilon_{\ell-1}$ and $\varepsilon'_{\ell-1}$. We will use these to define ε_ℓ and ε'_ℓ which correspond to parameters in the statement for ℓ.

Now consider $U^{(i)}$ and $V^{(i)}$ as in the statement of the theorem. Let us assume without loss of generality that $k_1 \geq k_2 \geq \cdots \geq k_\ell$. Also let $K = \sum_{j\in[\ell]} k_j$. We will now combine the last two components $(\ell - 1)$ and ℓ by the Khatri-Rao product.

$$\tilde{U} = U^{(\ell-1)} \odot U^{(\ell)} \text{ and } \tilde{V} = V^{(\ell-1)} \odot V^{(\ell)}.$$

Since we know that the two representations are close in Frobenius norm, we have

$$\left\| \sum_{r\in[R]} U_r^{(1)} \otimes U_r^{(2)} \otimes \cdots \otimes U_r^{(\ell-2)} \otimes \tilde{U}_r - \sum_{r\in[R]} V_r^{(1)} \otimes V_r^{(2)} \otimes \cdots \otimes V_r^{(\ell-2)} \otimes \tilde{V}_r \right\|_F < \varepsilon_\ell \quad (8)$$

Let us first check that the conditions for $(\ell - 1)$-order tensors hold for $\tilde{\tau} = (\tau_{\ell-1}\tau_\ell\sqrt{K}) \leq (\tau_{\ell-1}\tau_\ell\sqrt{3R})$. From Lemma 20, K-rank$_{\tilde{\tau}}(\tilde{U}) \geq \min\{k_\ell + k_{\ell-1} - 1, R\}$.

Suppose first that $k_\ell + k_{\ell-1} \leq R + 1$, then

$$\sum_{j\in[\ell-1]} k'_j \geq \sum_{j\in[\ell-2]} k_j + k_{\ell-1} + k_\ell - 1 \geq 2R + (\ell - 1) - 1.$$

Otherwise, if $k_\ell + k_{\ell-1} > R + 1$, then $k_{\ell-3} + k_{\ell-2} \geq R + 2$ (due to our ordering, and $\ell \geq 4$). Hence

$$\sum_{j\in[\ell-1]} k'_j \geq (\ell - 4) + (R + 2) + (R + 1) \geq 2R + \ell - 1$$

We now apply the inductive hypothesis on this $(\ell - 1)$th order tensor. Note that $\tilde{\rho} \leq (\rho_{\ell-1}\rho_\ell)$, $\tilde{\rho'} \leq (\rho'_{\ell-1}\rho'_\ell)$, $\tilde{\tau} \leq \left(2\tau_{\ell-1}\tau_\ell\sqrt{R}\right)$ and $\tilde{n} = n_{\ell-1}n_\ell$.

We will in fact apply it with $\varepsilon'_{\ell-1} < \min\{(R \cdot \tau_{\ell-1}\tau_\ell \cdot \rho'_{\ell-1}\rho'_\ell)^{-2},\ (\varepsilon'_\ell)^2/R\}$, so that we can later use Lemma 31. To ensure these, we will set

$$\varepsilon_\ell^{-1} = \vartheta_{21}^\ell \left(\frac{R}{\varepsilon'_\ell}\right) \cdot \left(\prod_{j\in[\ell-2]} \vartheta_{21}(\tau_j, \rho_j, \rho'_j, n_j)\right) \vartheta_{21}(\tilde{\tau}, \tilde{\rho}, \tilde{\rho'}, \tilde{n}),$$

where $\vartheta_{21}^\ell = x^{O(2^\ell)}$. From the values of $\tilde{\tau}, \tilde{\rho}, \tilde{n}$ above, this can easily be seen to be of the form in the statement of the theorem.

The inductive hypothesis implies that there is a permutation matrix Π and scalar matrices $\{\Lambda^{(1)}, \Lambda^{(2)}, \ldots, \Lambda^{(\ell-2)}, \Lambda'\}$, such that $\left\|\Lambda^{(1)}\Lambda^{(2)}\ldots\Lambda^{(\ell-2)}\Lambda' - I\right\| < \varepsilon'_{\ell-1}$ and

$$\forall j \in [\ell-2] \quad \left\|V^{(j)} - U^{(j)}\Pi\Lambda^{(j)}\right\|_F < \varepsilon'_{\ell-1}$$
$$\left\|\tilde{V} - \tilde{U}\Pi\Lambda'\right\|_F < \varepsilon'_{\ell-1}$$

Since $\varepsilon'_{\ell-1} < \varepsilon'_\ell$, equation (7) is satisfied for $j \in [\ell-2]$. We thus need to show that $\left\|V^{(j)} - U^{(j)}\Pi\Lambda^{(j)}\right\|_F < \varepsilon'_\ell$ for $j = \ell-1$ and ℓ. To do this, we appeal to Lemma 31, to say that if the Frobenius norm of the difference of two tensor products $u \otimes v$ and $u' \otimes v'$ is small, then the component vectors are nearly parallel.

Let us first set the parameters for applying Lemma 31. Each column vector is of length at most $L_{\max} \leq \tilde{\rho'} \leq (\rho'_{\ell-1}\rho'_\ell)$ and length at least $L_{\min} \geq 1/\tilde{\tau} \geq (2\tau_{\ell-1}\tau_\ell\sqrt{R})$. Hence, because of our choice of $\varepsilon'_{\ell-1} \ll \left(4\sqrt{R}(\tau_{\ell-1}\tau_\ell)(\rho'_{\ell-1}\rho'_\ell)\right)^{-1}$ earlier, the conditions of Lemma 31 are satisfied with $\delta \leq \varepsilon'_\ell$. Let $\delta_r = \left\|\tilde{V}_r - \tilde{U}_{\pi(r)}\Lambda'(r)\right\|_2$.

Now applying Lemma 31 with $\delta = \delta_r$, to column r, we see that there are scalars $\alpha_r(\ell-1)$ and $\alpha_r(\ell)$ such that

$$|1 - \alpha_r(\ell-1)\alpha_r(\ell)| < \frac{\varepsilon'_{\ell-1}}{L^2_{\min}} \leq \varepsilon'_\ell.$$

By setting for all $r \in [R]$, $\Lambda^{(\ell-1)}(r) = \alpha(\ell-1)_r$ and $\Lambda^{(\ell)}(r) = \alpha(2)\Lambda'(r)$, we see that the first part of (7) is satisfied. Finally, Lemma 31 shows that

$$\forall j \in \{\ell-1, \ell\} \quad \left\|V^{(j)}_r - U^{(j)}_{\pi(r)}\Lambda^{(j)}(r)\right\|_2 < \sqrt{\delta_r}\,, \quad \forall r \in [R]$$
$$\left\|V^{(j)} - U^{(j)}\Pi\Lambda^{(j)}\right\|_F < R^{1/4}\sqrt{\varepsilon'_{\ell-1}} \quad \text{(by Cauchy-Schwartz inequality).}$$
$$< \varepsilon'_\ell$$

This completes the proof of the theorem. ∎

We show a similar result for symmetric tensors, which shows robust uniqueness upto permutations (and no scaling) which will be useful in applications to mixture models (Section 4).

Corollary 22 (Unique Symmetric Decompositions) *For every $0 < \eta < 1$, $\tau, \rho, \rho' > 0$ and $\ell, R \in \mathbb{N}$, $\exists \varepsilon_\ell = \vartheta^{(\ell)}_{22}(\frac{1}{\eta}, R, n, \tau, \rho, \rho')$ such that, for any ℓ-order symmetric tensor (with $\ell \leq R$)*

$$T = \sum_{r\in[R]} \bigotimes_{j=1}^{\ell} U_r$$

where the matrix U is ρ-bounded with K-rank$_\tau(U) = k \geq \frac{2R-1}{\ell} + 1$, and for any other ρ' bounded, symmetric, rank-R decomposition of T which is ε-close, i.e.,

$$\left\|\sum_{r\in[R]} \bigotimes_{j=1}^{\ell} V_r - \sum_{r\in[R]} \bigotimes_{j=1}^{\ell} U_r\right\|_F \leq \varepsilon$$

there exists an $R \times R$ permutation matrix Π such that

$$\|V - U\Pi\|_F \leq \eta \tag{9}$$

The mild intricacy here is that applying Theorem 21 gives a bunch of scalar matrices whose product is close to the identity, while we want each of the matrices to be so. This turns out to be easy to argue – see Section C.4.

Appendix B. Computing Tensor Decompositions

For matrices, the theory of low rank approximation is well understood, and they are captured using singular values. In contrast, the tensor analog of the problem is in general ill-posed: for instance, there exist rank-3 tensors with arbitrarily good rank 2 approximations Landsberg (2012). For instance if u, v are orthogonal vectors, we have

$$u \otimes v \otimes v + v \otimes u \otimes v + v \otimes v \otimes u = \frac{1}{\varepsilon}\left[(v + \varepsilon u) \otimes (v + \varepsilon u) \otimes (v + \varepsilon u) - v \otimes v \otimes v\right] + \mathcal{N},$$

where $\|\mathcal{N}\|_F \leq O(\varepsilon)$, while it is known that the LHS has rank 3. However note that the rank-2 representation with error ε uses vectors of length $1/\varepsilon$, and such *cancellations*, in a sense are responsible for the ill-posedness.

Hence in order to make the problem well-posed, we will impose a boundedness assumption.

Definition 23 (ρ-bounded Low-rank Approximation) *Suppose we are given a parameter R and an $m \times n \times p$ tensor T which can be written as*

$$T = \sum_{i=1}^{R} a_i \otimes b_i \otimes c_i + \mathcal{N}, \tag{10}$$

where $a_i \in \mathbb{R}^m, b_i \in \mathbb{R}^n, c_i \in \mathbb{R}^p$ satisfy $\max\{\|a_i\|_2, \|b_i\|_2, \|c_i\|_2\} \leq \rho$, and $\mathcal{N}$ is a noise *tensor which satisfies $\|\mathcal{N}\|_F \leq \varepsilon$, for some small enough ε. The ρ-bounded low-rank decomposition problem asks to recover a* good *low rank approximation, i.e.,*

$$T = \sum_{i=1}^{R} a'_i \otimes b'_i \otimes c'_i + \mathcal{N}',$$

such that a'_i, b'_i, c'_i are vectors with norm at most ρ, and $\|\mathcal{N}'\|_F \leq O(1) \cdot \varepsilon$.

We note that if the decomposition into $[A\ B\ C]$ above satisfies the conditions of Theorem 5, then solving the ρ-bounded low-rank approximation problem would allow us to recover A, B, C up to a small error. The algorithmic result we prove is the following.

Theorem 24 *The ρ-bounded low-rank approximation problem can be solved in time poly$(n) \cdot \exp(R^2 \log(R\rho/\varepsilon))$.*

In fact, the $O(1)$ term in the error bound $\mathcal{N}' \leq O(1) \cdot \varepsilon$ will just be 5. Our algorithm is extremely simple conceptually: we identify three R-dimensional spaces by computing appropriate SVDs, and prove that for the purpose of obtaining an approximation with $O(\varepsilon)$ error, it suffices to look for

a_i, b_i, c_i in these spaces. We then find the approximate decomposition by a brute force search using an epsilon-net. Note that the algorithm has a polynomial running time for constant R, which is typically when the low rank approximation problem is interesting.

Proof In what follows, let M_A denote the $m \times np$ matrix whose columns are the so-called j,kth *modes* of the tensor T, i.e., the m dimensional vector of T_{ijk} values obtained by fixing j,k and varying i. Similarly, we define M_B $(n \times mp)$ and M_C $(p \times mn)$. Also, we denote by A the $m \times R$ matrix with columns being a_i. Similarly define B $(n \times R), C$ $(p \times R)$.

The outline of the proof is as follows: we first observe that the matrices M_A, M_B, M_C are all approximately rank R. We then let V_A, V_B and V_C be the span of the top R singular vectors of M_A, M_B and M_C respectively, and show that it suffices to search for a_i, b_i, and c_i in these spans. We note that we do not (and in fact cannot, as simple examples show) obtain the *true* span of the a_i, b_i and c_i's in general. Our proof carefully gets around this point. We then construct an ε-net for V_A, V_B, V_C, and try out all possible R-tuples. This gives the roughly $\exp(R^2)$ running time claimed in the Theorem.

We now make formal claims following the outline above.

Claim 25 *Let V_A be the span of the top R singular vectors of M_A, and let Π_A be the projection matrix onto V_A (i.e., $\Pi_A v$ is the projection of $v \in \mathbb{R}^n$ onto V_A). Then we have*

$$\|M_A - \Pi_A M_A\|_F \leq \varepsilon$$

Proof Because the top R singular vectors give the best possible rank-R approximation of a matrix for every R, for any R-dimensional subspace S, if Π_S is the projection matrix onto S, we have

$$\|M_A - \Pi_A M_A\|_F \leq \|M_A - \Pi_S M_A\|_F$$

Picking S to be the span of the vectors $\{a_1, \ldots, a_R\}$, we obtain

$$\|M_A - \Pi_S M_A\|_F \leq \|\mathcal{N}\|_F \leq \varepsilon.$$

The first inequality above is because the j,kth mode of the tensor $\sum_i a_i \otimes b_i \otimes c_i$ is a vector in the span of $\{a_1, \ldots, a_R\}$, in particular, it is equal to $\sum_i b_i(j)c_i(k)a_i$, where $b_i(j)$ denotes the jth coordinate of b_i.

This completes the proof. ■

Next, we will show that looking for a_i, b_i, c_i in the spaces V_A, V_B, V_C is sufficient. The natural choices are $\Pi_A a_i, \Pi_B b_i, \Pi_C c_i$, and we show that this choice in fact gives a good approximation. For convenience let $\widetilde{a}_i := \Pi_A a_i$, and $a_i^{\perp} := a_i - \widetilde{a}_i$.

Claim 26 *For $T, V_A, \widetilde{a}_i, \ldots$ as defined above, we have*

$$\left\| T - \mathcal{N} - \sum_i \widetilde{a}_i \otimes \widetilde{b}_i \otimes \widetilde{c}_i \right\|_F \leq 3\varepsilon.$$

Proof The proof is by a *hybrid argument*. We write

$$\begin{aligned} T - \mathcal{N} - \sum_i \widetilde{a}_i \otimes \widetilde{b}_i \otimes \widetilde{c}_i = &\big(\sum_i a_i \otimes b_i \otimes c_i - \widetilde{a}_i \otimes b_i \otimes c_i\big) \\ &+ \big(\sum_i \widetilde{a}_i \otimes b_i \otimes c_i - \widetilde{a}_i \otimes \widetilde{b}_i \otimes c_i\big) \\ &+ \big(\sum_i \widetilde{a}_i \otimes \widetilde{b}_i \otimes c_i - \widetilde{a}_i \otimes \widetilde{b}_i \otimes \widetilde{c}_i\big). \end{aligned}$$

We now bound each of the terms in the parentheses, and then appeal to triangle inequality (for the Frobenius norm). Now, the first term is easy:

$$\left\| \sum_i a_i \otimes b_i \otimes c_i - \widetilde{a}_i \otimes b_i \otimes c_i \right\|_F = \|M_A - \Pi_A M_A\|_F \leq \varepsilon.$$

One way to bound the second term is as follows. Note that:

$$\sum_i a_i \otimes b_i \otimes c_i - a_i \otimes \widetilde{b}_i \otimes c_i = \Big(\sum_i \widetilde{a}_i \otimes b_i \otimes c_i - \widetilde{a}_i \otimes \widetilde{b}_i \otimes c_i\Big) + \Big(\sum_i a_i^\perp \otimes b_i \otimes c_i - a_i^\perp \otimes \widetilde{b}_i \otimes c_i\Big).$$

Now let us denote the two terms in the parenthesis on the RHS by G, H – these are tensors which we view as mnp dimensional vectors. We have $\|G + H\|_2 \leq \varepsilon$, because the Frobenius norm of the LHS is precisely $\|M_B - \Pi_B M_B\|_F \leq \varepsilon$. Furthermore, $\langle G, H \rangle = 0$, because $\langle \widetilde{a}_i, a_j^\perp \rangle = 0$ for any i, j (one vector lies in the span V_A and the other orthogonal to it). Thus we have $\|G\|_2 \leq \varepsilon$ (since in this case $\|G + H\|_2^2 = \|G\|_2^2 + \|H\|_2^2$).

A very similar proof lets us conclude that the Frobenius norm of the third term is also $\leq \varepsilon$. This completes the proof of the claim, by our earlier observation. ■

The claim above shows that there exist vectors $\widetilde{a}_i, \widetilde{b}_i, \widetilde{c}_i$ of length at most ρ in V_A, V_B, V_C resp., which give a rank-R approximation with error at most 4ε. Now, we form an $\varepsilon/(R\rho^2)$-net over the ball of radius ρ in each of the spaces V_A, V_B, V_C. Since these spaces have dimension R, the nets have size

$$\Big(\frac{O(R\rho^2)}{\varepsilon}\Big)^R \leq \exp(O(R)\log(R\rho/\varepsilon)).$$

Thus let us try all possible candidates for $\widetilde{a}_i, \widetilde{b}_i, \widetilde{c}_i$ from these nets. Suppose we have $\widehat{a}_i, \widehat{b}_i, \widehat{c}_i$ being vectors which are $\varepsilon/(6R\rho^2)$-close to $\widetilde{a}_i, \widetilde{b}_i, \widetilde{c}_i$ respectively, it is easy to see that

$$\left\| \sum_i \widetilde{a}_i \otimes \widetilde{b}_i \otimes \widetilde{c}_i - \widehat{a}_i \otimes \widehat{b}_i \otimes \widehat{c}_i \right\|_F \leq \sum_i \left\| \widetilde{a}_i \otimes \widetilde{b}_i \otimes \widetilde{c}_i - \widehat{a}_i \otimes \widehat{b}_i \otimes \widehat{c}_i \right\|_F$$

Now by a hybrid argument exactly as above, and using the fact that all the vectors involved are $\leq \rho$ in length, we obtain that the LHS above is at most ε.

Thus the algorithm finds vectors such that the error is at most 5ε. The running time depends on the time taken to try all possible candidates for $3R$ vectors, and evaluating the tensor for each. Thus it is $\text{poly}(m, n, p) \cdot \exp(O(R^2)\log(R\rho/\varepsilon))$. ■

This argument generalizes in an obvious way to order ℓ tensors, and gives the following. We omit the proof.

Theorem 27 *There is an algorithm, that when given an order ℓ tensor of size n with a rank R approximation of error ε (in $\|\cdot\|_F$), finds a rank-R approximation of error $O(\ell\varepsilon)$ in time poly$(n) \cdot \exp(O(\ell R^2)\log(\ell R\rho/\varepsilon))$.*

B.1. Removing the ρ-boundedness Assumption

The assumption that there exists a low-rank decomposition which is ρ-bounded seems appropriate in many settings, but it is natural to ask if it can be removed. Note that the Claim 26 still holds, i.e., the spaces V_A, V_B, V_C as defined earlier still contain vectors which give a good approximation. However, we cannot use the same searching algorithm, since we do not have a bound on the lengths of the vectors we should search for.

Another way to look at the above is as follows: let us consider some orthonormal basis for each of V_A, V_B, V_C (call them $\{v_A^j\}_{j=1}^R$, etc), and write our tensor in this basis, plus some noise. Formally, we write $\mathbf{e}_i$ of the standard basis as a combination of the vectors v_A^j, plus noise, then write out $\mathbf{e}_i \otimes \mathbf{e}_j \otimes \mathbf{e}_k$ as a combination of $v_A^{j_1} \otimes v_B^{j_2} \otimes v_C^{j_3}$, plus noise. This transforms the original low-rank approximation problem to one for an $R \times R \times R$ tensor, and we need to find a decomposition with $\widetilde{a}_i, \widetilde{b}_i, \widetilde{c}_i$ being R dimensional vectors. However our main problem persists – we do not know a bound on the vectors in the decomposition.

We thank Ravi Kannan for suggesting the following to get around this issue: the key is to simply view this as a system of polynomial inequalities! Let us have $3R^2$ variables, R each for the entries $\widetilde{a}_i, \widetilde{b}_i, \widetilde{c}_i$. Now the fact that the squared Frobenius error with the original tensor is small ($\leq 25\varepsilon^2$) can be written down as a constraint the variables need to satisfy. We can then solve for our variables using algorithmic results on solving general polynomial systems over the reals Basu et al. (1996) (which are based on the decidability of the *existential theory of reals*). The best algorithms here end up giving a running time which is $\exp(\text{poly}(R))$ for our problem without the ρ-boundedness assumption. We will not go into the details here.

However the algorithms to solve polynomial systems of equalities are extremely involved, as opposed to the simple search process under the ρ-bounded assumption.

Appendix C. Auxiliary Lemmas and Complete proofs

We start by listing some of the (primarily linear algebra) lemmas we used in our proofs, which are proven in the full version of the paper.

Lemma 28 *Suppose X is a matrix in $\mathbb{R}^{n\times k}$ with $\sigma_k \geq 1/\tau$. Then if $\|\sum_i \alpha_i X_i\|_2 < \varepsilon$, for some α_i, we have $\|\alpha\| = \sqrt{\sum_i \alpha_i^2} \leq \tau\varepsilon$.*

Proof From the singular value condition, we have for any $y \in \mathbb{R}^k$,

$$\|Xy\|_2^2 \geq \sigma_k^2 \|y\|^2,$$

from which the lemma follows by setting y to be the vector of α_i. ■

Lemma 29 *Let $A \in \mathbb{R}^{n\times R}$ have K-rank$_\tau = k$ and be ρ-bounded. Then,*

1. *If $\mathcal{S} = \mathrm{span}(S)$, where S is a set of at most $k-1$ column vectors of A, then each unit vector in $\mathcal{S}$ has a small representation in terms of the columns denoted by S:*

$$v = \sum_{i \in S} z_i A_i \implies \frac{1}{(\rho^2+1)k} \leq (\sum_i z_i^2)/\left\|v\right\|^2 \leq \max\{\tau^2, 1\}$$

2. *If $\mathcal{S} = \mathrm{span}(S)$ where S is any subset of $k-1$ column vectors S of A, the other columns are far from the span $\mathcal{S}$:*

$$\forall j \in [R] \setminus S, \quad \left\|\Pi_{\mathcal{S}}^{\perp} A_j\right\| \geq \frac{1}{\tau}$$

3. *If $\mathcal{S}$ is any ℓ-dimensional space with $\ell < k$, then at most ℓ column vectors of A are ε-close to it for $\varepsilon = 1/(\tau\sqrt{\ell})$:*

$$\left|\{i : \left\|\Pi_{\mathcal{S}}^{\perp} A_i\right\| \leq \frac{1}{\tau\sqrt{\ell}}\}\right| \leq \ell$$

Proof We now present the simple proofs of the three parts of the lemma.

1. The first part simply follows because from change of basis. Let M be the $n \times n$ matrix, where the first $|S|$ columns of M correspond to S and the rest of the $n - |S|$ columns being unit vectors orthogonal to $\mathcal{S}$. Since $A_{|S}$ is well-conditioned, then $\lambda_{\max}(M) \leq (\rho+1)\sqrt{n}$ and $\lambda_{\min}(M) \geq 1/\max\tau, 1$. The change of basis matrix is exactly M^{-1}: hence $z = (M)^{-1}v$. Thus, $\lambda_{\min}(M^{-1}) \leq \|z\| \leq \lambda_{\max}(M^{-1}) = 1/\lambda_{\min}(M) \leq \max\{1, \tau\}$.

2. Let $S = \{1, \ldots, k-1\}$ and $j = k$ without loss of generality. Let $v = \sum_{i \in S} z_i A_i$ be a vector ε-close to A_k. Let M' be the $n \times k$ matrix restricted to first k columns: i.e. $M' = A|_{S \cup \{j\}}$. Hence, the vector $z = (z_1, \ldots, z_{k-1}, -1)$ has square length $1 + \sum_i z_i^2$, and $\|M'z\| = \varepsilon$. Thus,

$$\varepsilon \geq \lambda_{\min}(M')\sqrt{1 + \sum_i z_i^2} \geq 1/\tau$$

3. Let $\varepsilon = 1/(\tau\sqrt{k})$. For contradiction, assume that $S = \{i : \left\|\Pi_{\mathcal{S}}^{\perp} A_i\right\| \leq \varepsilon\}$ is of size $\ell + 1$. Let $v_i = \Pi_{\mathcal{S}} A_i \in \mathcal{S}$. Since $\{v_i\}_{i \in S}$ are $\ell + 1$ vectors in a ℓ dimension space,

$$\exists \{\alpha_i\}_{i \in S} \text{ with } \sum_i \alpha_i^2 = 1, \quad \text{s.t} \sum_i \alpha_i v_i = 0$$

Hence, $\left\|\sum_{i \in S} \alpha_i A_i\right\| \leq \left\|\sum_{i \in S} \alpha_i \Pi_{\mathcal{S}}^{\perp} A_i\right\| \leq (\sum_{i \in S} |\alpha_i|)\varepsilon \leq \sqrt{|S|}\varepsilon$ (where the last inequality follows from Cauchy-Schwarz inequality). But these set of α_i contradict the fact that the minimum singular value of any n-by-k submatrix of A is at least $1/\tau$.

■

C.1. Missing proofs – Permutation lemma

We first prove the base case, Lemma 11.

Proof Let V be the $(n-k+1)$ dimensional space orthogonal to the span of X_S, and let t be the number of columns of Y which have a projection $> \varepsilon_1$ onto V. From Lemma 30 (applied to the projections to V), there is a unit vector $w \in V$ (a random vector suffices) with dot-product of magnitude $> \varepsilon_1/(Rn) = \varepsilon$ with each of the t columns. From the hypothesis, since $w \in V$ ($\implies nz(w^T X) \leq R-k+1$), we have $t \leq R-k+1$. Thus at least $(k-1)$ of the columns are ε_1-close to $\mathrm{span}(X_S)$. Now since K-rank$_\tau(Y) \geq k$, it follows that k columns of Y cannot be ε_1-close to the $(k-1)$-dimensional space $\mathrm{span}(X_S)$ (Lemma 29). Thus $|T_S| = k-1$. ■

Lemma 30 *Let $u_1, \ldots, u_t \in \mathbb{R}^d$ (for some t, d) satisfy $\|u_i\|_2 \geq \varepsilon > 0$ for all i. Then there exists a unit vector $w \in \mathbb{R}^d$ s.t. $|\langle u_i, w\rangle| > \frac{\varepsilon}{20dt}$ for all $i \in [t]$.*

Proof The proof is by a somewhat standard probabilistic argument.

Let $r \sim \mathbb{R}^d$ be a random vector drawn from a uniform spherical Gaussian with a unit variance in each direction. It is well-known that for any $y \in \mathbb{R}^d$, the inner product $\langle y, r\rangle$ is distributed as a univariate Gaussian with mean zero, and variance $\|y\|_2^2$. Thus for each y, from standard anti-concentration properties of the Gaussian, we have

$$\Pr\left[|\langle u_i, r\rangle| \leq \frac{\|u_i\|}{10t}\right] \leq \frac{1}{2t}.$$

Thus by a union bound, with probability at least $1/2$, we have

$$\Pr\left[|\langle u_i, r\rangle| > \frac{\varepsilon}{10t}\right] \qquad \text{for all } i. \tag{11}$$

Next, since $\mathbb{E}\left[\|r\|_2^2\right] = d$, $\Pr[\|r\|_2^2 > 4d] < 1/4$, and thus there exists a vector r s.t. $\|r\|_2^2 \leq 4d$, and Eq. (11) holds. This implies the lemma (in fact we obtain $\sqrt{d}$ in the denominator). ■

C.1.1. PROOFS OF LEMMAS 12 AND 14

Let us now prove the main linear algebraic lemma and the sunflower lemma.

Proof [Proof of Lemma 12] W.l.o.g., let us suppose $B = \{1, \ldots, q\}$. Also, let x_j denote the jth column of X. From the hypothesis, we can write:

$$\begin{aligned} y &= u_1 + \sum_{j \neq 1} \alpha_{1j} x_j + z_1 \\ y &= u_2 + \sum_{j \neq 2} \alpha_{2j} x_j + z_2 \\ &\vdots \\ y &= u_q + \sum_{j \neq q} \alpha_{rj} x_j + z_q, \end{aligned}$$

where $u_i \in \mathrm{span}(X_A)$ and z_i are the *error* vectors, which by hypothesis satisfy $\|z_i\|_2 < \varepsilon$. We will use the fact that $|A| + |B| \leq k$ to conclude that *each* α_{ij} is tiny. This then implies the desired conclusion.

By equating the first and ith equations ($i \geq 2$), we obtain

$$u_1 + \sum_{j \neq 1} \alpha_{1j} x_j + z_1 = u_i + \sum_{j \neq i} \alpha_{ij} x_j + z_i.$$

Thus we have a combination of the vectors x_i being equal to $z_i - z_1$, which by hypothesis is small: $\|z_i - z_1\|_2 \leq 2\varepsilon$. Now the key is to observe that the coefficient of x_i is precisely α_{1i}, because it is zero in the ith equation. Thus by Lemma 28 (since K-rank$_\tau(X) \geq k$), we have that $|\alpha_{1i}| \leq 2\tau\varepsilon$.

Since we have this for all i, we can use the first equation to conclude that

$$\|y - u_1\|_2 \leq \sum_{j \neq 1} |\alpha_{1j}| \, \|x_j\|_2 + \|z_1\|_2 \leq 2q\tau\rho\varepsilon + \varepsilon < 4n\tau\rho\varepsilon$$

The last inequality is because $q < n$, and this completes the proof. ■

Proof [Proof of Lemma 14] The proof is by a counting argument. By the sunflower structure, each T_i has some intersection with T^* (possibly empty), and some elements which do not belong to $T_{i'}$ for any $i' \neq i$. Call the number of elements of the latter kind t_i. Then we must have

$$R + (q-1)\theta \leq \sum_i |T_i| = \sum_i (t_i + |T_i \cap T^*|) \leq \sum_i t_i + q|T^*|.$$

Now since all $T_i \subseteq [R]$, we have

$$\sum_i t_i + |T^*| \leq R.$$

Combining the two, we obtain

$$R + (q-1)\theta \leq R + (q-1)|T^*| \implies |T^*| \geq \theta,$$

as desired. For equality to occur, we must have equality in each of the places above, in particular, we must have $|T_i \cap T^*| = |T^*|$ for all i, which implies $T^* \subseteq T_i$ for all i. ■

C.2. Lemmas about the Khatri-Rao product

Lemma 31 *Suppose* $\|u \otimes v - u' \otimes v'\|_F < \delta$, *and* $L_{\min} \leq \|u\|, \|v\|, \|u'\|, \|v'\| \leq L_{\max}$, *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*

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

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

$$\begin{aligned} \left\|u \otimes v - u' \otimes v'\right\|_F^2 &< \delta^2 \\ \left\|(1-\alpha_1\alpha_2)u' \otimes v' + \beta_1\alpha_2\tilde{u}_\perp \otimes v' + \beta_2\alpha_1 u' \otimes \tilde{v}_\perp + \beta_1\beta_2\tilde{u}_\perp \otimes \tilde{v}_\perp\right\|_F^2 &< \delta^2 \\ L_{\min}^4(1-\alpha_1\alpha_2)^2 + \beta_1^2\alpha_2^2 L_{\min}^2 + \beta_2^2\alpha_1^2 L_{\min}^2 + \beta_1^2\beta_2^2 &< \delta^2 \end{aligned} \tag{12}$$

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

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

Now, using (12), we see that $\beta_1 < \sqrt{\delta}$. ■

Lemma 32 *For $\lambda \ge 0$, a vector $v \in \mathbb{R}^n$ with $\|v\|_1 \in [1-\varepsilon/4, 1+\varepsilon/4]$, a probability vector $u \in \mathbb{R}^n$ ($\|u\|_1 = \sum_i u_i = 1$), if*

$$\|v - \lambda u\|_2 \le \frac{\varepsilon}{4\sqrt{n}}$$

then we have

$$1 - \varepsilon/2 \le \lambda \le 1 + \varepsilon/2 \quad \textit{and} \quad \|v-u\|_2 \le \varepsilon$$

Proof First we have $\|v - \lambda u\|_1 \le \varepsilon/4$ by Cauchy-Schwartz. Hence, by triangle inequality, $|\lambda|\,\|u\|_1 \le 1 + \varepsilon/2$.
Since $\|u\|_1 = 1$, we get $\lambda \le 1 + \varepsilon/2$. Similarly $\lambda \ge 1 - \varepsilon/2$.

Finally, $\|v-u\|_2 \le \|v - \lambda u\|_2 + |\lambda - 1|\,\|u\|_2 \le \varepsilon$ (since $\lambda \ge 0$). Hence, the lemma follows. ■

C.3. Khatri-Rao product adds up

We now prove the lemma that shows that the K-rank of the Khatri-Rao product is at least additive in the worst case.

Proof [Proof of Lemma 20] Let $\tau = \tau_1\tau_2\sqrt{k_A + k_B}$. Suppose for contradiction M has K-rank$_\tau(M) < k = k_A + k_B - 1 \le R$ (otherwise we are done).
Without loss of generality let the sub-matrix M' of size $(n_1 n_2) \times k$, formed by the first k columns of M have $\lambda_k(M) < 1/\tau$. Note that for a vector $z \in R^{nR}$, $\|z\|_2 = \|Z\|_F$ where Z is the natural $n \times R$ matrix representing z. Hence

$$\exists \{\alpha_i\}_{i\in[k]} \text{ with } \sum_{i\in[k]} \alpha_i^2 = 1 \quad \text{s.t.} \quad \left\|\sum_{i\in[k]} \alpha_i A_i \otimes B_i\right\|_F < \varepsilon.$$

Clearly $\exists i^* \in [k]$ s.t $|\alpha_i| \geq 1/\sqrt{k}$: let $i^* = k$ without loss of generality. Let $\mathcal{S} = \text{span}(\{A_1, A_3, \ldots A_{k_A-1}\})$, and pick $x = \Pi_{\mathcal{S}}^{\perp} A_k / \left\|\Pi_{\mathcal{S}}^{\perp} A_k\right\|$ (it exists because K-rank$_\tau(M) < R$).
Pre-multiplying the expression in (C.3) by x, we get

$$\left\| \sum_{i=k_A}^{k} \beta_i B_i \right\| < \varepsilon \text{ where } \beta_i = \alpha_i \langle x, A_i \rangle$$

But $|\beta_k| \geq 1/(\sqrt{k}\tau_1)$ (by Lemma 29), and there are only $k - k_A + 1 \leq k_B$ terms in the expression. Again, by Lemma 29 applied to these (at most) k_B columns of B, we get that $1/\varepsilon < \tau_1\tau_2\sqrt{k}$, which establishes the lemma. ■

Remark. Note that the bound of the lemma is tight in general. For instance, if A is an $n \times 2n$ matrix s.t. the first n columns correspond to one orthonormal basis, and the next n columns to another (and the two bases are random, say). Then K-rank$_{10}(A) = n$, but for any τ, we have K-rank$_\tau(A \odot A) = 2n - 1$, since the first n terms and the next n terms of $A \odot A$ add up to the same vector (as a matrix, it is the identity).

C.4. Symmetric Decompositions

Proof [Proof of Corollary 22] Applying Theorem 21 with $\varepsilon' < \eta(2\rho\tau\sqrt{R})^{-1}$, to obtain a permutation matrix Π and scalar matrices Λ_j such that

$$\forall j \in [\ell] \, \|V - U\Pi\Lambda_j\|_F < \varepsilon'$$
$$\text{By triangle inequality,} \quad \forall j, j' \in [\ell], \ \left\|U\Pi(\Lambda_j - \Lambda_{j'})\right\|_F < 2\varepsilon'$$

Since Π is a permutation matrix and U has columns of length at least $1/\tau$, we get that

$$\forall r \in [R], j \in [\ell], j' \in [\ell], \quad \left|\Lambda_j(r) - \Lambda_{j'}(r)\right| < \varepsilon'\tau$$

However, we also know that

$$\left\| \prod_{j \in \ell}, \ \Lambda_j - I \right\| \leq \varepsilon'$$
$$\forall r \in [R], \quad (1 - \varepsilon') \leq \prod_{j \in [\ell]} \Lambda_j(i) \leq 1 + \varepsilon'$$

Hence, substituting (C.4) in the last inequality, it is easy to see that $\forall i \in [n], |\lambda_j(i) - 1| < 2\varepsilon'\tau$. But since each column of A is ρ-bounded, this shows that $\|A' - A\Pi\|_F < 2\varepsilon'\tau\rho\sqrt{R} \leq \eta$, as required. ■

Appendix D. Complete proofs for the Robust Uniqueness Theorem for 3-tensors.

D.1. Proof of Lemma 8

W.l.o.g., we may assume that $k_A \geq k_B$ (the proof for $k_A < k_B$ will follow along the same lines). For convenience, let us define α to be the vector $x^T C$, and β the vector $x^T C'$. Let t be the number

of entries of β of magnitude $> \varepsilon'$. The assumption of the lemma implies that $t \leq R - k_C + 1$. Now from (6), we have

$$M := \sum_i \alpha_i A_i \otimes B_i = \sum_i \beta_i A'_i \otimes B'_i + Z, \tag{13}$$

where Z is an error matrix satisfying $\|Z\|_F \leq \varepsilon$. Now, since the RHS has at most t terms with $|\beta_i| > \varepsilon'$, we have that σ_{t+1} of the LHS is at most $R\rho'_A\rho'_B\varepsilon' + \varepsilon$. Using the value of t, we obtain

$$\sigma_{R-k_C+2}(M) \leq \sigma_{t+1}(M) < \varepsilon + (R\rho'_A\rho'_B)\varepsilon' \tag{14}$$

We will now show that if $x^T C$ has too many co-ordinates which are larger than ε'' then we will contradict (14). One tricky case we need to handle is the following: while each of these non-negligible co-ordinates of $x^T C$ will give rise to a large rank-1 term, they can be canceled out by combinations of the rank-1 terms corresponding to entries of $x^T C$ which are slightly smaller than ε''. Hence, we will also set a smaller threshold δ and first handle the case when there are many co-ordinates in $x^T C$ which are larger than δ. δ is chosen so that the terms with $(x^T C)_i < \delta$ can not cancel out any of the large terms ($(x^T C)_i \geq \varepsilon''$).

Define $S_1 = \{i : |(x^T C)_i| > \varepsilon''\}$ and $S_2 = \{i : |(x^T C)_i| > \delta\}$, where $\delta = \varepsilon''/\vartheta$ for some error polynomial $\vartheta = 2R^2\rho_A\rho_B\rho_C\rho'_A\rho'_B\rho'_C\tau_A\tau_B\tau_C$ (which is always > 1). Thus we have $S_1 \subseteq S_2$. We consider two cases.

Case 1: $|S_2| \geq k_B$**.**

In this case we will give a lower bound on $\sigma_{R-k_C+2}(M)$, which gives a contradiction to (14). The intuition is roughly that A, B have k_A, k_B *large* singular values, and thus the product should have enough large ones as well. To formalize this, we use the following standard fact about singular values of products, which is proved by considering the variational characterization of singular values:

Fact 33 *Let P, Q be matrices of dimensions $p \times m$ and $m \times q$ respectively. Then for all ℓ, i such that $\ell \leq \min\{p, q\}$, we have*

$$\sigma_\ell(PQ) \geq \sigma_{\ell+m-i}(P)\sigma_i(Q) \tag{15}$$

Now, let us view M as PQ, where $P = A$, and $Q = \mathrm{diag}(\alpha)B^T$. We will show that $\sigma_{k_B}(Q) \geq \delta/\tau_B$, and that $\sigma_{2R+2-k_B-k_C}(A) \geq 1/\tau_A$. These will then imply a contradiction to (14) by setting $\ell = R - k_C + 2$ and $i = k_B$ since

$$\frac{\delta}{\tau_A\tau_B} = \frac{\varepsilon''}{\vartheta\tau_A\tau_B} > (R\rho'_A\rho'_B\varepsilon' + \varepsilon) \text{ by our choice of } \vartheta_8.$$

(It is easy to check that $\ell \leq \min\{k_A, k_B\} \leq \min\{n_A, n_B\}$, and thus we can use the fact above.)

Thus we only need to show the two inequalities above. The latter is easy, because by the hypothesis we have $2R + 2 - k_B - k_C \leq k_A$, and we know that $\sigma_{k_A}(A) \geq 1/\tau_A$, by the definition of K-rank$_{\tau_A}(A)$. Thus it remains to prove the first inequality. To see this, let $J \subset S_2$ of size k_B. Let B_J^T and Q_J be the submatrices of B^T and Q restricted to rows of J. Thus we have $Q_J = \mathrm{diag}(\alpha)_J B_J^T$. Because of the Kruskal condition, every k_B sized sub matrix of B is well-conditioned, and thus $\sigma_{k_B}(B_J) = \sigma_{k_B}(B_J^T) \geq 1/\tau_B$.

Further, since $|\alpha_j| > \delta \ \forall j \in J$, multiplication by the diagonal cannot lower the singular values by much, and we get $\sigma_{k_B}(Q_J) \geq \delta/\tau_B$. This can also be seen formally by noting that $\sigma_{k_B}(\mathrm{diag}(\alpha)_J) \geq \delta$, and applying Fact 33 with $P = \mathrm{diag}(\alpha)_J, Q = B_J^T$ and $\ell = m = i = k_B$.

Finally, since Q is essentially Q_J along with additional rows, we have $\sigma_{\tau_B}(Q) \geq \sigma_{\tau_B}(Q_J) \geq \delta/\tau_B$. From the argument earlier, we obtain a contradiction in this case.

Case 2: $|S_2| < k_B$**.**

Roughly, by defining S_1, S_2, we have divided the coefficients α_i into large ($\geq \varepsilon''$), small, and tiny ($< \delta$). In this case, we have that the number of large and small terms together (in M, see Eq. (13)) is at most k_B. For contradiction, we can assume the number of large ones is $\geq t+1$, since we are done otherwise. The aim is to now prove that this implies a lower bound on $\sigma_{t+1}(M)$, which gives a contradiction to Eq. (14).

Now let us define $M' = \sum_{i \in S_2} \alpha_i (A_i \otimes B_i)$. Thus M and M' are equal up to *tiny* terms. Further, let Π be the matrix which projects a vector onto the span of $\{B'_i : |\beta_i| \geq \varepsilon'\}$, i.e., the span of the columns of B' which correspond to $|\beta_i| \geq \varepsilon'$. Because there are at most t such β_i, this is a space of dimension $\leq t$. Thus we can rewrite Eq. (13) as

$$M' = \sum_{i \in S_1} \alpha_i (A_i \otimes B_i) + \sum_{j \in S_2 \setminus S_1} \alpha_j (A_j \otimes B_j) = \sum_{i=1}^{t} \beta_i (A'_i \otimes B'_i) + Err, \tag{16}$$

where we assumed w.l.o.g. that $|\beta_i| \geq \varepsilon'$ for $i \in [t]$, and Err is an error matrix of Frobenius norm at most $\varepsilon + R(\rho_A \rho_B \delta + \rho'_A \rho'_B \varepsilon') \leq \varepsilon + (R\rho_A \rho_B \rho'_A \rho'_B)(\delta + \varepsilon')$.

Now because $|S_1| \geq t+1$, and K-rank$_{\tau_B}(B) \geq k_B \geq t+1$, there must be one vector among the B_i, $i \in S_1$, which has a reasonably large projection orthogonal to the span above, i.e., which satisfies

$$\|B_i - \Pi B_i\|_2 \geq 1/(\tau_B \sqrt{R}).$$

Let us pick a unit vector y along $B_i - \Pi B_i$. Consider the equality (16) and multiply by y on both sides. We obtain

$$\sum_{i \in S_2} \alpha_i \langle B_i, y \rangle A_i = (Err)y.$$

Thus we have a combination of the A_i's, with at least one coefficient being $> \varepsilon''/(R\tau_B)$, having a magnitude at most $\|(Err)y\|_2 < \vartheta_1(\delta + \varepsilon' + \varepsilon)$, where ϑ_1 was specified above.
Now $k_A \geq k_B \geq |S_2|$. So, we obtain a contradiction by Lemma 28 since:

$$\begin{aligned} \|(Err)y\|_2 < \vartheta_1(\delta + \varepsilon' + \varepsilon) &= R\rho_A \rho_B \rho'_A \rho'_B (\delta + \varepsilon' + \varepsilon) \\ &= R\rho_A \rho_B \rho'_A \rho'_B \left(\frac{\varepsilon''}{\vartheta} + \varepsilon' + \varepsilon\right) \\ &< \frac{1}{\tau_A} \cdot \frac{\varepsilon''}{R\tau_B} \end{aligned}$$

The last inequality follows because $\vartheta = 2R^2 \rho_A \rho_B \rho_C \rho'_A \rho'_B \rho'_C \tau_A \tau_B \tau_C$.
This completes the proof in this case, hence concluding the proof of the lemma.

D.2. Proof of Lemma 10

By symmetry, let us just show this for matrix C' (dimensions $n \times R$), and let $k = k_C$ for convenience. We need to show that every n-by-k submatrix of C' has minimum singular value $\geq \delta = 1/\tau'_C$.

For contradiction let C'_S be the submatrix corresponding to the columns in S ($|S| = k$), such that $\sigma_k(C'_S) < \delta$. Let us consider a left singular vector z which corresponds to $\sigma_k(C'_S)$, and suppose z is normalized to be unit length. Then we have

$$\sum_{i \in S} \langle z, C'_i \rangle^2 < \delta^2$$

Thus $|\langle z, C'_i \rangle| < \delta$ for all $i \in S$, so we have $nz_\delta(z^T C') \leq n - k$. Now from Lemma 8, we have

$$nz_{\varepsilon_1}(zC) \leq n - k, \text{ where } \varepsilon_1 = \vartheta_8(\varepsilon + \delta).$$

Let J denote the set of indices in $z^T C$ which are $< \varepsilon_1$ in magnitude (by the above, we have $|J| \geq k$). Thus we have $\|zC_J\|_2 < R\varepsilon_1$, which leads to a contradiction if we have K-rank$_{1/(R\varepsilon_1)}(C) \geq k$.

Since this is true for our choice of parameters, the claim follows.

D.3. Wrapping up the Proof of Theorem 5

Suppose we are given an $\varepsilon' < 1$ as in the statement of the theorem. For a moment, suppose ε is small enough, and A, B, C, A', B', C' satisfying the conditions of the theorem produce tensors which are ε-close.

From the hypothesis, note that $k_A, k_B, k_C \geq 2$ (since $k_A, k_B, k_C \leq R$, and $k_A + k_B + k_C \geq 2R + 2$). Thus from the Lemmas 10 and 8 (setting $\varepsilon' = 0$), we obtain that C, C' satisfy the hypothesis of the Robust permutation lemma (Lemma 6) with C', C set to X, Y respectively, and the parameters

$$\text{“}\tau\text{”} := \vartheta_{10}\,; \quad \text{“}\varepsilon\text{”} := \vartheta_8 \varepsilon.$$

Hence, we apply Lemma 6 to A, B and C, and get that there exists permutation matrices Π_A, Π_B and Π_C and scalar matrix $\Lambda_A, \Lambda_B, \Lambda_C$ such that for $\varepsilon_2 = \vartheta_6 \vartheta_8 \cdot \varepsilon$,

$$\left\|A' - A\Pi_A\Lambda_A\right\|_F < \varepsilon_2, \ \left\|B' - B\Pi_B\Lambda_B\right\|_F < \varepsilon_2 \text{ and } \left\|C' - C\Pi_C\Lambda_C\right\|_F < \varepsilon_2 \tag{17}$$

We follow the outline given in the proof sketch. **To show $\Pi_A = \Pi_B = \Pi_C$:**

Let us assume for contradiction that $\Pi_A \neq \Pi_B$. We will use an index where the permutations disagree to obtain a contradiction to the assumptions on the K-rank .

For notational convenience, let $\pi_A : [R] \to [R]$ correspond to the permutation given by Π_A, with $\pi_A(r)$ being the column that A'_r maps to. Permutation $\pi_B : [R] \to [R]$ similarly corresponds to Π_B. Using (17) for A we have

$$\begin{aligned}\left\|\sum_{r\in[R]} \left(A'_r - \Lambda_A(r) A_{\pi_A(r)}\right) \otimes B'_r \otimes C'_r\right\|_F &\leq \sum_{r\in[R]} \left\|\left(A'_r - \Lambda_A(r) A_{\pi_A(r)}\right) \otimes B'_r \otimes C'_r\right\|_F \\ &\leq \varepsilon_2 \sqrt{R} \rho'_B \rho'_C \text{ using Cauchy-Schwarz}\end{aligned}$$

By a similar argument, and using triangle inequality (along with $\varepsilon_2 \leq 1 \leq \rho'_B$) we get

$$\left\|\sum_{r\in[R]} A'_r \otimes B'_r \otimes C'_r - \sum_{r\in[R]} \Lambda_A(r)\Lambda_B \cdot A_{\pi_A(r)} \otimes B_{\pi_B(r)} \otimes C'_r\right\|_F \leq 2\varepsilon_2 \sqrt{R}(\rho'_B \rho'_C + \rho'_A \rho'_C)$$

Let us take linear combinations given by unit vectors v and w, of the given tensor $T = [A\ B\ C]$ along the first and second dimensions. By combining the above inequality along with the fact that the two decompositions are ε-close i.e. $\left\|\sum_{r\in[R]} A_r \otimes B_r \otimes C_r - A'_r \otimes B'_r \otimes C'_r\right\|_F \leq \varepsilon$, we have

$$\left\|Z - Z'\right\| \leq \varepsilon_3 = \varepsilon + 2\varepsilon_2 R\rho'_C(\rho'_A + \rho'_B) \text{ where}$$
$$Z = \sum_{r\in[R]} \langle v, A_r\rangle \langle w, B_r\rangle C_r \quad \text{and } Z' = \sum_{r\in[R]} \Lambda_A(r)\Lambda_B(r) \left\langle v, A_{\pi_A(r)}\right\rangle \left\langle w, B_{\pi_B(r)}\right\rangle C'_r$$

Note that the ε term above is negligible compared to the second term involving ε_2.
We know that $\pi_A \neq \pi_B$, so there exist $s \neq t \in [R]$ such that $r^* = \pi_A(s) = \pi_B(t)$. We will now use this r^* to pick v and w carefully so that the vector Z' is negligible while Z is large. We partition $[R]$ into V, W with $|V| = k_A - 1$ and $|W| \leq k_B - 1$, so that $\pi_A(t) \in V$ and $\pi_B(s) \in W$ and for each $r \in [R] - \{s, t\}$, either $\pi_A(r) \in V$ or $\pi_B(r) \in W$. Such a partitioning is possible since $R \leq k_A + k_B - 2$.

Let $\mathcal{V} = \text{span}(V)$ and $\mathcal{W} = \text{span}(W)$. We know that $r^* = \pi_A(s) \notin S$ and $r^* = \pi_A(t) \notin T$. Hence, pick v as unit vector along $\Pi^{\perp}_{\mathcal{V}} A_{r^*}$ and w as unit vector along $\Pi^{\perp}_{\mathcal{W}} B_{r^*}$. By this choice, we ensure that $Z' = 0$ (since $v \perp \mathcal{V}$ and $w \perp \mathcal{W}$).

However, K-rank$_{\tau_A}(A) \geq k_A$ and K-rank$_{\tau_B}(B) \geq k_B$, so $\langle v, A_{r^*}\rangle \langle w, B_{r^*}\rangle \geq 1/\tau_A\tau_B$ (by Lemma 29). Further, $|V| = k_A - 1$ implies that at most $R - k_A + 1 \leq k_C - 1$ terms of Z is non-zero.

$$\left\|\sum_{r\in[R]\setminus V} \beta_r C_r\right\| \leq \varepsilon_3 \text{ where } \beta_r = \langle v, A_r\rangle \langle w, B_r\rangle$$

Further, $|\beta_{r^*}| \geq (\tau_A\tau_B)^{-1}$, and since K-rank$_{\tau_C}(C) = k_C \geq R - |V| + 1$, we have a contradiction if $\varepsilon_3 < (\tau_A\tau_B\tau_C)^{-1}$ due to Lemma 29. This will be true for our choice of parameters. Hence $\Pi_A = \Pi_B$, and similarly $\Pi_A = \Pi_C$. Let us denote $\Pi = \Pi_A = \Pi_B = \Pi_C$. In the remainder, we assume Π is the identity, since this is without loss of generality.

To show $\Lambda_A\Lambda_B\Lambda_C =_{\varepsilon'} I_R$:

Let us denote $\beta_i = \lambda_A(i)\lambda_B(i)\lambda_C(i)$. From (17) and triangle inequality, we have as before

$$\left\|\sum_{r\in[R]} A'_r \otimes B'_r \otimes C'_r - \sum_{r\in[R]} \Lambda_A(r)\Lambda_B(r)\Lambda_C(r) \cdot A_{\pi_A(r)} \otimes B_{\pi_B(r)} \otimes C_{\pi_C(r)}\right\|_F \leq 5\varepsilon_2\sqrt{R}\rho'_A\rho'_B\rho'_C$$

Combining this with the fact that the decompositions are ε-close we get

$$\left\|\sum_{r\in[R]} (1 - \beta_r) A_r \otimes B_r \otimes C_r\right\| < \varepsilon_4 = \varepsilon + 5\sqrt{R}\rho'_A\rho'_B\rho'_C\varepsilon_2 \leq 6\sqrt{R}\rho'_A\rho'_B\rho'_C\varepsilon_2.$$

By taking linear combinations given by unit vectors x, y along the first two dimensions (i.e. xA and yB) we have

$$\left\|\sum_{r\in[R]} (1 - \beta_r)(xA_r)(yB_r)C_r\right\| < \varepsilon_4.$$

We will show each β_r is negligible. Since $R+2 \leq k_A+k_B$, let $S, W \subseteq [R]-\{r\}$ be disjoint sets of indices not containing r, such that $|S| = k_A-1$ and $|W| \leq k_B-1$. Let $\mathcal{S} = \text{span}(\{A_j : j \in S\})$ and $\mathcal{W} = \text{span}(\{B_j : j \in W\})$. Let x and y be unit vectors along $\Pi_{\mathcal{S}}^{\perp} A_r$ and $\Pi_{\mathcal{W}}^{\perp} B_r$ respectively.

Since K-rank$_{\tau_A}(A) \geq k_A$ and K-rank$_{\tau_B}(B) \geq k_B$, we have that $\left\|\Pi_{\mathcal{S}}^{\perp} A_r\right\| \geq 1/\tau_A$ (similarly for B_r). Hence, from Lemma 29

$$(1-\beta_r)(\frac{1}{\tau_A \tau_B}) \|C_r\| < \varepsilon_4 \implies 1-\beta_r < \varepsilon_4 \tau_A \tau_B \tau_C.$$

Thus, $\|\Lambda_A \Lambda_B \Lambda_C - I\| \leq \varepsilon_4 \tau_A \tau_B \tau_C \leq \varepsilon'$ (our choice of ε will ensure this). This implies the theorem.

Let us now set the ε for the above to hold (note that ϑ_6 involves a τ term which depends on ϑ_{10})

$$\varepsilon := \frac{\varepsilon'}{6(R\tau_A\tau_B\tau_C)\rho'_A\rho'_B\rho'_C \cdot \vartheta_8\vartheta_6},$$

which can easily be seen to be of the form in the statement of the theorem. This completes the proof.

Appendix E. Sampling Error Estimates for Higher Moment Tensors

In this section, we show error estimates for ℓ-order tensors obtained by looking at the ℓ^{th} moment of various hidden variable models. In most of these models, the sample is generated from mixture of R distributions $\{\mathcal{D}_r\}_{r\in[R]}$, with mixing probabilities $\{w_r\}_{r\in[R]}$. First the distribution $\mathcal{D}_r$ is picked with probability w_r, and then the data is sampled according to $\mathcal{D}_i$, which is characteristic to the application.

Lemma 34 (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 $\{M^{(j)}\}_{j\in[\ell]}$, such that every entry of $x^{(j)} \in \mathbb{R}^n$ is bounded by c_{max} (or if it is multivariate gaussian). Then, for every $\varepsilon > 0$, there exists $N = O(c_{max}^{\ell}\varepsilon^{-2}\sqrt{\ell \log n})$ 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

$$\left\| \mathbb{E}\left[x^{(1)} \otimes x^{(2)} \otimes \ldots x^{(\ell)}\right] - \frac{1}{N}\left(\sum_{t\in[N]} x(t)^{(1)} \otimes x(t)^{(2)} \otimes x(t)^{(\ell)}\right)\right\|_{\infty} < \varepsilon \tag{18}$$

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

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

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 c_{max}^{ℓ}. Hence, we have a sum of N bounded random variables. By Bernstein bounds, probability for (18) to not occur $\exp\left(-\frac{\left(\varepsilon n^{-\ell/2}\right)^2 N^2}{2Nc_{max}^{\ell}}\right) = \exp\left(-\varepsilon^2 N/\left(2(c_{max}n)^{\ell}\right)\right)$. We have n^{ℓ} events to union

bound over. Hence $N = O(\varepsilon^{-2}(c_{max}n)^{\ell}\sqrt{\ell \log n})$ suffices. Note that similar bounds hold when the $x^{(j)} \in \mathbb{R}^n$ are generated from a multivariate gaussian. ■

Lemma 35 (Error estimates for Gaussians) *Suppose x is generated from a mixture of R-gaussians with means $\{\mu_r\}_{r\in[R]}$ and covariance $\sigma^2 I$, with the means satisfying $\|\mu_r\| \leq c_{max}$.*
For every $\varepsilon > 0, \ell \in \mathbb{N}$, there exists $N = poly\left(\frac{1}{\varepsilon}, \sigma^2, n, R\right)$ such that if $x^{(1)}, x^{(2)}, \ldots, x^{(N)} \in R^n$ were the N samples, then

$$\forall\{i_1, i_2, \ldots, i_\ell\} \in [n]^\ell, \left| \mathbb{E}\left[\prod_{j\in[\ell]} x_{i_j}\right] - \frac{1}{N}\left(\sum_{t\in[N]}\prod_{j\in[\ell]} x_{i_j}^{(t)}\right)\right| < \varepsilon. \tag{19}$$

In other words,

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

Proof Fix an element $(i_1, i_2, \ldots, i_\ell)$ of the ℓ-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 spherical (axis-aligned suffices) and each mean is bounded by c_{max}, $|Z| < (c_{max} + t\sigma)^\ell$ with probability $O\left(\exp(-t^2/2)\right)$. Hence, by using standard sub-gaussian tail inequalities, we get

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

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

Appendix F. Applications to Polynomial Identifiability

F.1. Multi-view Mixture Model

Multi-view models are mixture models with a latent variable h, where we are given multiple observations or views $x^{(1)}, x^{(2)}, \ldots, x^{(\ell)}$ that are conditionally independent given the latent variable h. Multi-view models are very expressive, and capture many well-studied models like Topic Models Anandkumar et al. (2012c), Hidden Markov Models (HMMs) Mossel and Roch (2006); Allman et al. (2009); Anandkumar et al. (2012c), random graph mixtures Allman et al. (2009). Allman et al Allman et al. (2009) refer to these models by *finite mixtures of finite measure products*. We first introduce some notation, along the lines of Allman et al. (2009); Anandkumar et al. (2012c).

Definition 36 (Multi-view mixture models)

- *The latent variable h is a discrete random variable having domain $[R]$, so that $\Pr[h = r] = w_r, \forall r \in [R]$.*

- *The views* $\{x^{(j)}\}_{j\in[\ell]}$ *are random vectors* $\in \mathbb{R}^n$ *(with* ℓ_1 *norm at most* 1*), that are conditionally independent given* h*, with means* $\mu^{(j)} \in \mathbb{R}^n$ *i.e.*

$$\mathbb{E}\left[x^{(j)}|h=r\right] = \mu_r^{(j)} \text{ and } \mathbb{E}\left[x^{(i)}\otimes x^{(j)}|h=r\right] = \mu_r^{(i)}\otimes\mu_r^{(j)} \text{ for } i\neq j$$

- *Denote by* $M^{(j)}$*, the* $n \times R$ *matrix with the means* $\{\mu_r^{(j)}\}_{r\in[R]}$ *(normalized with* ℓ_1 *norm at most 1) comprising its columns i.e.*

$$M^{(j)} = [\mu_1^{(j)}|\ldots|\mu_r^{(j)}|\ldots|\mu_R^{(j)}].$$

The parameters of the model to be learned are the matrices $\{M^{(j)}\}_{j\in[\ell]}$ and the mixing weights $\{w_r\}_{r\in[R]}$. In many settings, the n-dimensional vectors $x^{(j)}$ are actually indicator vectors: this is commonly used to encode the case when the observation is one of n discrete events.

Our main theorem is formally the following.

Theorem 37 (Polynomial Identifiability of Multi-view mixture model) *The following statement holds for any constant integer* ℓ*. Suppose we are given samples from a multi-view mixture model (see Def 36), with the parameters satisfying:*

1. *For each mixture* $r \in [R]$*, the mixture weight* $w_r > \gamma$*.*
2. *For each* $j \in [\ell]$*,* k*-rank*$_\tau(M^{(j)}) \geq k \geq \frac{2R}{\ell} + 1$*.*

then there is a algorithm that given any $\eta > 0$ *uses* $N = \vartheta_{37}{}^{(\ell)}\left(\frac{1}{\eta}, R, n, \tau, 1/\gamma\right)$ *samples, and finds with high probability* $\{\tilde{M}^{(j)}\}_{j\in[\ell]}$ *and* $\{\tilde{w}_r\}_{r\in[R]}$ *(upto renaming of the mixtures* $\{1,2,\ldots,R\}$*) such that*

$$\forall j \in [\ell], \quad \left\|M^{(j)} - \tilde{M}^{(j)}\right\|_F \leq \eta \quad \text{and} \quad \forall r \in [R],\ |w_r - \tilde{w}_r| < \eta \tag{20}$$

The function $\vartheta_{37}{}^{(\ell)}(\cdot,\ldots,\cdot) = \text{poly}(Rn/(\gamma\eta))^{2^\ell}\text{poly}(n,\tau,1/\gamma)^\ell$ is a polynomial for constant ℓ and satisfies the theorem.

Remarks:

1. *Note that the condition* (a) *in the theorem about the mixing weights* $w_r > \gamma$ *is required to recover* all *the parameters, since we need poly*$(1/w_r)$ *samples before we see a sample from mixture* r*. However, by setting* $\gamma \ll \varepsilon'$*, the above algorithm can still be used to recover the mixtures components of weight larger than* ε'*.*
2. *The theorem also holds when for different* j*, the* k*-rank*$_\tau(M^{(j)})$ *have bounds* k_j *which are potentially different, and satisfy the same condition as in Theorem 21.*

Proof [Proof of Theorem 37] Set $\eta' = \frac{\eta\gamma}{16\ell n}$. We know from Lemma 34 that the ℓ^{th} moment tensor can be estimated to accuracy $\varepsilon_1 = \left(\ell \cdot \vartheta_{21}{}^{(\ell)}(R/\eta') \cdot \vartheta_{21}(\tau/\gamma, c_{max}\sqrt{n}, c_{max}\sqrt{n}, n)\right)^{-1}$ in $\|\cdot\|_F$ norm using $N = O(\varepsilon_1^{-2}R(c_{max})^\ell\sqrt{\ell\log n})$ samples. This estimated tensor $\tilde{T}$ has a rank-R decomposition upto error ε_1.

Next, we will apply our algorithm for getting approximate low-rank tensor decompositions from Section B on $\tilde{T}$. Since each $\mu_r^{(j)}$ is a probability distribution, we can obtain vectors $\{\tilde{u}_r^{(j)}\}_{j\in[\ell],r\in[R]}$ (let us call the corresponding $n \times R$ matrices $\tilde{U}^{(j)}$) such that

$$\forall j \in [\ell-1], r \in [R] \ \left\|\tilde{u}_r^{(j)}\right\|_1 \in [1-\delta, 1+\delta] \quad \text{where } \delta = \varepsilon_1\sqrt{R} < \frac{\eta}{2\ell}.$$

This is possible since the algorithm in Section B searches for the vectors $\tilde{u}_r^{(j)}$, by just enumerating over ε-nets on an R-dimensional space. An alternate way to see this is to obtain any decomposition and scale all but the last column in the matrices $\tilde{U}^{(j)}$ so that they have ℓ_1 norm of 1 (upto error δ). Note that this step of finding an ε-close rank-R decomposition can also just comprise of brute force enumeration, if we are only concerned with polynomial identifiability. Hence, we have obtained a rank-R decomposition which is $O(\ell\varepsilon_1)$ far in $\|\cdot\|_F$.

Now, we apply Theorem 21 to ℓ^{th} moment tensor T to claim that these $\tilde{U}^{(j)}$ are close to $M^{(j)}$ upto permutations. When we apply Theorem 21, we absorb the co-efficients w_r into $M^{(\ell)}$. In other words

$$U^{(j)} = M^{(j)} \ \text{ for all } j \in [\ell-1], \quad \text{and} \quad U^{(\ell)} = M^{(\ell)}\text{diag}(w).$$

We know that K-rank$_\tau(M^{(j)}) = k_j$, and K-rank$_{\tau/\gamma}(U^{(\ell)}) = k_\ell$. We now apply Theorem 21 with our choice of ε_1, and assuming that the permutation is identity without loss of generality, we get

$$\forall r \in [R] \ \left\|\tilde{u}_r^{(j)} - \Lambda^{(j)}(r)\mu_r^{(j)}\right\| < \eta' \le \frac{\eta\gamma}{16n\ell} \quad \forall j \in [\ell-1]$$
$$\text{and} \ \left\|\tilde{u}_r^{(\ell)} - \Lambda^{(\ell)}(r)w_r\mu_r^{(\ell)}\right\| < \eta' \le \frac{\eta\gamma}{16\ell n}$$

for some scalar matrices Λ_j (on R-dims) such that

$$\left\|\prod \Lambda^{(j)} - I_R\right\| \le \frac{\eta}{16\ell n}$$

Note that the entries in the diagonal matrices Λ_j (the scalings) may be negative. We first transform the vectors so that each of the entries in Λ_j are non-negative (this is possible since the product of Λ_j is close to the identity matrix, which only has non-negative entries).

$$\forall j \in [\ell], r \in [R], \quad \tilde{v}_r^{(j)} = \text{sgn}\left(\Lambda^{(j)}(r)\right) \cdot \tilde{u}_r^{(j)} \tag{21}$$

This ensures that

$$\forall j \in [\ell-1], r \in [R] \ \left\|\tilde{v}_r^{(j)} - \left|\Lambda^{(j)}(r)\right|\mu_r^{(j)}\right\| < \eta' \le \frac{\eta\gamma}{16n\ell} \quad \text{and} \tag{22}$$
$$\forall r \in [R] \ \left\|\tilde{v}_r^{(\ell)} - \left|\Lambda^{(\ell)}(r)\right|w_r\mu_r^{(\ell)}\right\| < \eta' \le \frac{\eta\gamma}{16\ell n} \tag{23}$$

Moreover, the $\mu_r^{(j)}$ correspond to probability vectors which have $\|\mu^{(j)}\|_1 = 1$, we have ensured that $\left\|\tilde{v}_r^{(j)}\right\|_1 \in [1-\delta, 1+\delta]$. Applying Lemma 32 we get that the required estimates $\tilde{v}_r^{(j)}$ (i.e. $\tilde{\mu}_r^{(j)}$) satisfy:

$$\forall j \in [\ell-1], r \in [R], \ \left\|\tilde{v}_r^{(j)} - \mu_r^{(j)}\right\| \le \frac{\eta\gamma}{4\ell\sqrt{n}} \quad \text{and} \quad \left|\Lambda^{(j)}(r)\right| \in \left[1 - \frac{\eta\gamma}{8\ell\sqrt{n}}, 1 - \frac{\eta\gamma}{8\ell\sqrt{n}}\right]$$

Now, set $\tilde{\mu}_r^{(\ell)} = \frac{\tilde{v}_r^{(\ell)}}{\left\|\tilde{v}_r^{(\ell)}\right\|_1}$, and $\tilde{w}_r = \left\|\tilde{v}_r^{(\ell)}\right\|_1$, for all $r \in [R]$. Now, from equations (F.1) and (F.1) we get that

$$\begin{aligned} \forall r \in [R] \qquad & \left|\Lambda^{(\ell)}(r) - 1\right| \leq \frac{\eta\gamma}{8\sqrt{n}} \\ \text{Hence from (23),} \quad & \left\|\tilde{v}_r^{(\ell)} - w_r \mu_r^{(\ell)}\right\| \leq \frac{\eta\gamma}{4\sqrt{n}} \\ & \left\|\tilde{w}_r \tilde{\mu}_r^{(\ell)} - w_r \mu_r^{(\ell)}\right\| \leq \frac{\eta\gamma}{4\sqrt{n}} \\ & w_r \left\|\frac{\tilde{w}_r}{w_r} \tilde{\mu}_r^{(\ell)} - \mu_r^{(\ell)}\right\| \leq \frac{\eta\gamma}{4\sqrt{n}} \end{aligned}$$

Using the fact that $w_r \geq \gamma$ and using Lemma 32, we see that $\tilde{w}_r$ and $\tilde{\mu}_r^{(\ell)}$ are also η-close estimates to w_r and $\mu_r^{(\ell)}$ respectively, for all r. ■

F.2. Hidden Markov Models

Our result on HMMs can be formally stated as follows.

Corollary 38 (Polynomial Identifiability of Hidden Markov models) *The following statement holds for any constant $\delta > 0$. Suppose we are given a Hidden Markov model as described above, with parameters satisfying:*

1. *The stationary distribution $\{w_r\}_{r\in[R]}$ has $\forall r \in [R]$ $w_r > \gamma_1$,*
2. *The observation matrix M has K-rank$_\tau(M) \geq k \geq \delta R$,*
3. *The transition matrix P has minimum singular value $\sigma_R(P) \geq \gamma_2$,*

then there is a algorithm that given any $\eta > 0$ uses $N = \vartheta_{37}{}^{(\frac{1}{\delta}+1)}\left(\frac{1}{\eta}, R, n, \tau, \frac{1}{\gamma_1\gamma_2}\right)$ samples of $m = 2\lceil\frac{1}{\delta}\rceil + 3$ consecutive observations (of the Markov Chain), and finds with high probability, P', M' and $\{\tilde{w}_r\}_{r\in[R]}$ such that

$$\left\|M - M'\right\|_F \leq \eta, \quad \left\|P - P'\right\|_F \leq \eta \text{ and } \forall r \in [R], \ |w_r - \tilde{w}_r| < \eta \tag{24}$$

Further, this algorithm runs in time $n^{O_\delta(R^2 \log(\frac{1}{\eta\gamma_1}))}\left(n \cdot \frac{\tau}{\gamma_1\gamma_2}\right)^{O_\delta(1)}$ time.

Remark: *Allman et al. (2009) show identifiability under weaker conditions than ours. This is because they prove their results for* generic *values of the parameters (this formally means they hold for all M, P except a set of measure zero, but they do not give a characterization). Our bounds are weaker, but hold whenever the K-rank$_\tau(M) \geq \delta n$ condition holds. Further, since we rely on robust uniqueness, our result holds when we only have a finite number of samples.*

Proof [Proof sketch] Here, the idea is to cleverly come up with three independent views of the HMM, so that it fits into the multi-view framework. The proof follows along the lines of Allman

et al. (2009), so we only sketch it here. We now show to cast this HMM into a multi-view model (Def. 36) using a nice trick of Allman et al. (2009). We can then apply Theorem 37 and prove identifiability (Corollary 38). We will choose $m = 2q + 1$ where $q = \lceil \frac{1}{\delta} \rceil + 1$, and then use the hidden state Z_{q+1} as the latent variable h of the Multi-view model. We will use three different views ($\ell = 3$) as shown in Fig. 2: the first view A comprises the tuple of observations $(X_q, X_{q-1}, \ldots, X_1)$ (ordered this way for convenience), the second view B is the observation X_{q+1}, while the third view C comprises the tuple $V_3 = (X_{q+2}, X_{q+3}, \ldots, X_{2q+1})$. This fits into the Multi-view model since the three views are conditionally independent given the latent variable $h = Z_{q+1}$.

Abusing notation a little, let A, B, C be matrices of dimensions $n^q \times R, n \times R, n^q \times R$ respectively. They denote the conditional probability distributions as in Definition 36. For convenience, let $\tilde{P} = \text{diag}(w)P^T\text{diag}(w)^{-1}$, which is the "reverse transition" matrix of the Markov chain given by P. We can now write the matrices A, B, C in terms of M and the transition matrices. The matrix product $X \odot Y$ refers to the Khatri-Rao product (Lemma 20). Showing that these are indeed the transition matrices is fairly straightforward, and we refer to Allman et al. Allman et al. (2009) for the details.

$$A = ((\ldots(M\tilde{P}) \odot M)\tilde{P}) \odot M) \ldots \tilde{P}) \odot M)\tilde{P} \tag{25}$$

$$B = M \tag{26}$$

$$C = ((\ldots(MP) \odot M)P) \odot M) \ldots P) \odot M)P \tag{27}$$

(There are precisely q occurrences of M, P (or $\tilde{P}$) in the first and third equalities). Now we can use the properties of the Khatri-Rao product. For convenience, define $C^{(1)} = MP$, and $C^{(j)} = (C^{(j-1)} \odot M)P$ for $j \geq 2$, so that we have $C = C^{(q)}$. By hypothesis, we have K-rank$_\tau(M) \geq k$, and thus K-rank$_{\tau_2\tau}(MP) \geq k$ (because P is a stochastic matrix with all eigenvalues $\geq \tau_2$). Now by the property of the Khatri-Rao product (Lemma 20), we have K-rank$_{(\tau\tau_2)\tau}(C^{(2)}) \geq \min\{R, 2k\}$. We can continue this argument, to eventually conclude that K-rank$_{\tau'}(C^{(q)}) = \min\{R, qk\} = R$ for $\tau' = \tau^q\gamma_2^{q^2}(qk)^{q/2}$.

Precisely the same argument lets us conclude that K-rank$_{\tau'}(A) \geq R$, for the $\tau' = \tau^q\gamma_2^{q^2}(qk)^{q/2}$. Now since K-rank$_\tau(B) \geq 2$, we have that the conditions of Theorem 5 hold. Now using the arguments of Theorem 37 (here, we use Theorem 5 instead of Theorem 21), we get matrices A', B', C' and weights w' such that

$$\left\|A' - A\right\|_F < \delta \quad \text{and similarly for } B, C$$
$$\left\|w' - w\right\| < \delta$$

for some $\delta = \text{poly}(1/\eta, \ldots)$. Note that $M = B$. We now need to argue that we can obtain a good estimate P' for P, from A', B', C'. This is done in Allman et al. (2009) by a trick which is similar in spirit to Lemma 31. It uses the property that the matrix C above is full rank (in fact well conditioned, as we saw above), and the fact that the columns of M are all probability distributions.

Let $D = C^{(q-1)}$, as defined above. Hence, $C = (D \odot M)P$. Now note that all the columns of M represent probability distributions, so they add up to 1. Thus given $D \odot M$, we can combine (simply add) appropriate rows together to get D. Thus by performing this procedure (adding rows) on C, we obtain DP. Now, if we had performed the entire procedure by replacing q with $(q-1)$ (we should ensure that $(q-1)k \geq R$ for the Kruskal rank condition to hold), we would obtain the

matrix D. Now knowing D and DP, we can recover the matrix P, since D is well-conditioned. ■