Uniqueness of Tensor Decompositions with Applications to Polynomial Identifiability

Aditya Bhaskara[*]
Google Research NYC

Moses Charikar[†]
Princeton University

Aravindan Vijayaraghavan[‡]
Carnegie Mellon University

Abstract

We give a robust version of the celebrated result of Kruskal on the uniqueness of tensor decompositions: we prove that given a tensor whose decomposition satisfies a robust form of Kruskal's rank condition, 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.

Recently, algorithms based on tensor decompositions have been used to estimate the parameters of various hidden variable models efficiently in special cases as long as they satisfy certain "non-degeneracy" properties. Our methods give a way to go beyond this non-degeneracy barrier, and establish polynomial identifiablity of the parameters under much milder conditions. Given the importance of Kruskal's theorem in the tensor literature, we expect that this robust version will have several applications beyond the settings we explore in this work.

1   Introduction

Statisticians have long studied the identifiability of probabilistic models [Tei61, Tei67, TC82], i.e. whether the parameters of a model can be learned from data. A central question in unsupervised learning [Gha04] is the efficient computation of such latent model parameters from observed data. The method of moments approach, pioneered by Pearson [Pea94], infers model parameters from empirical moments. In general, very high order moments and exponential sample complexity may be needed for this approach to succeed [MV10, BS10, GLPR12]. Our focus in this work is to understand settings where 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 [AMR09, APRS11, RS12]. The moments are naturally represented by tensors (high dimensional analogs of matrices) and low rank decompositions of such tensors can be used to deduce the parameters of the underlying model. A fundamental result of Kruskal [Kru77] on uniqueness of tensor decompositions plays a crucial role in ensuring that the model parameters are correctly identified by this procedure. Unfortunately, this does not give any sample complexity bounds for learning within specified error bounds, since Kruskal's theorem only guarantees uniqueness for low rank decompositions of the *exact* moment tensors. A natural question here is: *Does uniqueness also hold for decompositions of empirical moment tensors which are approximations of the exact tensors? How small does the error in approximation of exact tensors need to be for uniqueness to hold?*

Our main technical contribution in this work is establishing such a robust version of Kruskal's classic uniqueness theorem for tensor decompositions for inverse polynomial error. This directly establishes polynomial identifiability in a host of applications where Kruskal's theorem was used [AMR09] to prove identifiability assuming access to exact moment tensors (because polynomially many samples from the distribution (typically) yield an approximation to these tensors up to $1/\text{poly}(n)$ error).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.

1.1   Tensors and their Decompositions

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 [Kru77] 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:

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

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 [tBS02, JS04, SS07, Rho10, Lan12]. Sidiropoulos and Bro [SB00] extended this result to ℓ-order tensors.

We give a robust version of of Kruskal's uniqueness theorem and its higher dimensional generalizations. To this end, we need a natural robust analogue of Kruskal rank: we say that K-rank$_\tau(A) \geq k$ if every submatrix of A formed by k of its columns has minimum singular value at least $1/\tau$. 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) can then be stated 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 $\varepsilon < \varepsilon' \cdot poly(R, n, \tau)$.*

(See Theorem 3.1 for a formal version.) 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

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

(See Theorem 3.14 for a formal version.)

How should we interpret the uniqueness theorems? The first observation is that a *typical* tensor of dimension $n^{\times \ell}$ and rank $\leq \ell n/2 - 1$ has a unique decomposition. This is because an $n \times (\ell n/2)$ matrix typically (for instance, a random one) has K-rank$_\tau$ equal to n. For $\ell \geq 3$, this corresponds to the so-called *over-complete* setting in applications to machine learning (to follow).

Second, we can view the Kruskal condition (Eq. 3) as a sufficient condition for proving a lower bound on tensor rank (because of the uniqueness, there cannot exist a smaller decomposition). Thus our robust version can be viewed as giving a lower bound on approximability using low-rank tensors (or a lower bound on the so-called *border rank*).

Can our uniqueness results go beyond linear rank ($\ell n/2$)? Typical order ℓ tensors have a rank $\Omega(n^{\ell-1})$, so this is a natural question. While Kruskal's conditions are known to be tight, it is possible that stronger assumptions help. Indeed for $\ell = 3$, algebraic geometry approaches [CO12] show that *typical* tensors of rank $n^2/16$ have a unique decomposition. Obtaining a robust version of these results is a very interesting open problem. However for $\ell \geq 5$, some of our tools, combined with many ideas from random matrix theory, can be used [BCMV13] to obtain robust uniqueness (as well as algorithms) for ranks up to $n^{\lfloor \frac{\ell-1}{2} \rfloor}$ in a *smoothed analysis* setting.

The other immediate question that arises is, can uniqueness proofs in the over-complete setting lead to new algorithmic tools? This is a challenging open problem (see Section 6) – we suspect that it is computationally hard to find decompositions of 3-tensors when the rank is $(1 + \varepsilon)n$. However, some of the tools (such as the Khatri-Rao product) can be used to obtain decompositions when we have access to higher order tensors. This is explored in recent work with Ankur Moitra [BCMV13].

In this work we show (Theorem 4.2) that a very simple SVD based algorithm can be used to find low-rank approximations, albeit in time exponential in the rank. This can be viewed as a tensor analog of low-rank approximation, which is very well-studied for matrices.

2

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 Related Work

Tensor decompositions have been applied in a variety of fields, from psychology and chemometrics, to statistics and machine learning. Consequently, there has been a lot of theoretical as well as heuristic work on algorithms for tensor decomposition. However all the algorithms with theoretical guarantees (we are aware of) require *non-degeneracy* assumptions, such as the matrices in the decomposition (i.e. A, B, C) having full rank. This approach thus intrinsically fails when the number of terms in the decomposition R is larger than the dimension n. This is often referred to as the *over-complete* case.

It is, for example, an explicitly stated barrier in the works of Anandkumar et al. [AGH+12], who give power iteration style algorithms for the case $R \leq n$. Even the earlier work of Leurgans et al. [LRA93] require the matrices to be of full rank.

The only works we are aware of in the over-complete case are those of de Lathauwer et al. [DLCC07] (the so-called FOOBI algorithm), and the subsequent works of Goyal et al. [GVX13] and [BCMV13]. De Lathauwer et al. [DLCC07] assumes that the tensor is generic (i.e. does not lie on any low dimensional manifold), which does not yield robustness. The last two works do give robust guarantees, however they work under assumptions that have a different flavor from our Kruskal rank conditions. For instance, [GVX13] assume that certain higher dimensional matrices obtained from the decomposition are full rank, while [BCMV13] analyse a smoothed analysis model for tensor decompositions.

1.3 Applications to Latent Variable Models

The robust version of Kruskal's theorem that we establish can be applied to get polynomial sample complexity bounds for learning several different latent variable models including multi-view mixture models, exchangeable (single) topic models, Hidden Markov Models, and mixtures of spherical Gaussians without separation assumptions.

We will formally state the identifiability and algorithmic results we obtain for each of these in Section 5. As an illustration, we describe the application to multi-view mixture models and Hidden Markov models:

Multi-view models are mixture models with a discrete latent variable $h \in [R]$, such that $\Pr[h = r] = w_r$. We are given multiple observations or views $x^{(1)}, x^{(2)}, \ldots, x^{(\ell)}$ that are conditionally independent given the latent variable h, with $\mathbb{E}[x^{(j)}|h = r] = \mu_r^{(j)}$. 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 mixing weights $\{w_r\}_{r \in [R]}$.

Multi-view models are very expressive, and capture many well-studied models like Topic Models [AHK12], Hidden Markov Models (HMMs) [MR06, AMR09, AHK12], random graph mixtures [AMR09], and the techniques developed for this class have also been applied to phylogenetic tree models [Cha96, MR06] and certain tree mixtures [AHHK12].

Hidden Markov Models (HMMs) are extensively used in speech recognition, image classification, bioinformatics etc[Edd96, GY08]. We follow the same setting as in [AMR09]: 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 represented by a 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 rth column M_r represents the probability distribution conditioned on the state $Z_t = r$ i.e.

$$\forall r \in [R], t \in [m], i \in [n], \quad \Pr[X_t = i | Z_t = r] = M_{ir}.$$

In many important applications of HMMs, n is much smaller than R. For instance in image classification, the commonly used SIFT features [Low99] are 128 dimensional, while the number of image classes is much larger, e.g. 256 classes in the Caltech-256 dataset [GHP07] and several thousands in the case of ImageNet [DDS$^+$09]. A similar situation arises in speech, where HMMs have been very successful. Further, in some other applications, even when the feature vectors lie in a large dimensional space ($n \gg R$), the set of relevant features or the effective feature space could be a space of much smaller dimension ($k < R$), that is unknown to us.

Results: Here is a representative result for multi-view models that applies when 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. In order to establish this, we apply our robust uniqueness result to the ℓth moment tensor for $\ell = \lceil 2/\delta \rceil + 1$.

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.*

Polynomial identifiability was not known previously for these models in the settings that we consider. For instance, the best known algorithms for learning multi-view mixture models in the worst case are from a recent work Rabani et al [RSS14], who give an algorithm that has complexity $R^{O(R^2)} + \text{poly}(n, R)$ and show that $\exp(\tilde{\Omega}(R))$ samples are necessary for learning multi-view models in one dimension ($n = 1$). Anandkumar et al. [AHK12] gave polynomial time algorithms in a restricted setting called the non-singular or non-degenerate setting. When each of these matrices $\left\{M^{(j)}\right\}_{j \in [\ell]}$ to have rank R in a robust sense i.e. $\sigma_R(M^{(j)}) \geq 1/\tau$ for all $j \in [\ell]$, their algorithm has running time $\text{poly}(R, n, \tau, 1/\varepsilon))$ for learning the parameters up to error ε. However, their algorithm fails even when $R = n+1$. (We now know [BCMV13] of algorithms for much larger R in a smoothed setting, but the results in the present paper apply whenever Kruskal's condition is satisfied).

1.4   Overview of Techniques

Robust Uniqueness of Tensor decompositions.   Our proof broadly follows the outline of Kruskal's original proof [Kru77]: It proceeds by establishing a *permutation lemma*, which gives sufficient conditions to conclude 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 first step in our argument is to prove that if A, B, C are "well-conditioned" (i.e., satisfy the K-rank conditions of the theorem), then any other "bounded" decomposition which is ε-close is also well-conditioned. This step is crucial to our argument, while an analogous step was not required in Kruskal's original proof.[1]   Besides, this statement is interesting in its own right: it implies, for

---

[1]Note that the uniqueness theorem, in hindsight, establishes that the other decomposition is also well-conditioned.

instance, that there cannot be a smaller rank (bounded) decomposition.

The second and most technical step is to prove the robust permutation lemma. The (robust) Permutation lemma needs to establish that for every column of C', there is some column of C close to it. Kruskal's proof [Kru77] roughly uses downward induction to establish the following claim: for every set of $i \leq$ K-rank columns of C', there are at least i columns of C that are in the span of the chosen 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 C which are ε-close to the span of i columns of C'. 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 C' is also "well-conditioned".

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 general, it can often be improved – and this increases the R for which we obtain uniqueness. See Remark 3.13 in Section 3.4 for a discussion.

Algorithms for low-rank tensor approximation. A simple SVD based algorithm can be used to obtain a constant approximation to the best rank-R approximation problem for tensors, with running time $\exp(R^2)\text{poly}(n)$ (which is practical for small R). At a high level, our algorithm for finding a rank R approximation proceeds by finding a small ($O(R)$) dimensional space and then exhaustively searching inside it. This improves upon the naive search over an ε-net over $\mathbb{R}^n$, which takes time $\exp(Rn)\text{poly}(n)$.

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, but we will consider only real tensors. 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 2.1 (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

$$[U^{(1)} \ U^{(2)} \ \ldots \ U^{(\ell)}] := \sum_{r=1}^{R} U_r^{(1)} \otimes U_r^{(2)} \otimes \ldots \otimes U_r^{(\ell)}, \text{ where } A_r \text{ to denotes the } r\text{th column of } A.$$

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 2.2 (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 2.3 (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 are 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. We also note that this is somewhat in the spirit of (but much weaker than) the Restricted Isometry Property (RIP) [CT05] from the Compressed Sensing literature.

Another simple linear algebra definition we use is the following

Definition 2.4 (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.

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 $\leq \text{poly}(n)$. So also, our upper bounds on lengths $\rho_A, \rho_B, \ldots$ are all assumed to be between 1 and some $\text{poly}(n)$. This helps simplify the statements of our lemmas.

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 3.11, for instance, we use $\vartheta_{3.11}$. Unless specifically mentioned, they will be polynomials in the parameters mentioned above, so we do not mention them each time.

In our proofs we will require several simple (mostly elementary linear algebra) lemmas. The Section A is a medley of such lemmas. Most of the proofs are reasonably straightforward, and thus we place them in the Appendix.

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 [Kru77]. The key ingredient, which is a robust version of the so-called *permutation lemma* is presented in Section 3.2, since it seems interesting 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 3.1 (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*
$$\varepsilon = \varepsilon' / \big(R^6 \vartheta_{3.1}(\tau_A, \rho_A, \rho'_A, n_A) \vartheta_{3.1}(\tau_B, \rho_B, \rho'_B, n_B) \vartheta_{3.1}(\tau_C, \rho_C, \rho'_C, n_C) \big),$$
for some polynomial $\vartheta_{3.1}$ 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

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

The proof of Theorem 3.1 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

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

7

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 3.2 (Robust permutation lemma). *Suppose X, Y are ρ-bounded $n \times R$ matrices such that $K\text{-rank}_\tau(X)$ and $K\text{-rank}_\tau(Y)$ are $\geq k$, for some integer $k \geq 2$. Further, suppose that for $\varepsilon < 1/\vartheta_{3.2}$, 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 $\|X - Y\Pi\Lambda\|_F < \vartheta_{3.2} \cdot \varepsilon$. In fact, we can pick $\vartheta_{3.2} := (nR^2)\vartheta_{3.8}$.

Remark 3.3. 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, the proof of Lemma 3.7 actually shows that X and Y have columns that are permutations of each other up to scaling. However, as we will soon see in Lemma 3.4, 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 rank (or conditioned-ness) 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 3.2 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 3.4. *Suppose A, B, C, A', B', C' satisfy the conditions of Theorem 3.1, 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 $\varepsilon'' = \vartheta_{3.4} \cdot (\varepsilon + \varepsilon')$, where $\vartheta_{3.4} := 4R^3(\tau_A \tau_B \tau_C)^2 \rho_A \rho_B \rho_C (\rho'_A \rho'_B \rho'_C)^2$.

Remark 3.5. This lemma, together with its corollary Lemma 3.6 will imply the conditions of the permutation lemma. While the proof of the robust permutation lemma (Lemma 3.2) will directly apply this Lemma with $\varepsilon' = 0$, we will need the $\varepsilon' > 0$ case for establishing Lemma 3.6 that lets us conclude that $K\text{-rank}_{\tau\vartheta}(C') \geq K\text{-rank}_\tau(C)$ 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 give a rough intuition and defer the proof to the appendix B.1).

Intuition for Lemma 3.4. For convenience, let us define α to be the vector $x^T C$, and β the vector $x^T C'$. Let $t \leq R - k_C + 1$ be the number of entries of β of magnitude $> \varepsilon'$. Now from (6), we have

$$M := \sum_i \alpha_i A_i \otimes B_i = \sum_i \beta_i A'_i \otimes B'_i + Z, \text{ where error matrix } \|Z\|_F \leq \varepsilon \tag{7}$$

8

Now at most $t \le R - k_C + 1$ of the β_i are non-negligible. Thus the RHS above is essentially a rank $\le R - k_C + 1$ matrix. Now suppose $> R - k_C + 1$ of the α_i are non-negligible. Then it turns out to imply a contradiction to the Kruskal rank assumption for either A or B (because we cannot have many terms $A_i \otimes B_i$ with non-negligible coefficients summing to a very small rank matrix unless there was a small combination of the terms which itself summed up to zero). This is where the rank condition $k_A + k_B + k_C \ge 2R + 2$ is crucially used. However, carrying out this approach in the robust case (with errors) requires a delicate analysis that is handled in appendix B.1. $\qquad\square$

The next lemma uses the above to conclude that K-rank$_{\vartheta\tau}(C') \ge$ 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. Further, it says that any decomposition $[A' \ B' \ C']$ has *rank at least R*. Otherwise, we could add some zero-columns to each of A', B', C' and apply this lemma to conclude K-rank.$(A') \ge 2$, a contradiction if there exists a zero column.

Lemma 3.6. *Let A, B, C, A', B', C' be as in the setting of Theorem 3.1. Suppose $[A \ B \ C] =_\varepsilon [A' \ B' \ C']$, with $\varepsilon < 1/\vartheta_{3.6}$, where $\vartheta_{3.6} = R\tau_A\tau_B\tau_C\vartheta_{3.4} = 4R^4\tau_A^3\tau_B^3\tau_C^3\rho_A\rho_B\rho_C(\rho_A'\rho_B'\rho_C')^2$. Then A', B', C' have K-rank$_{\tau'}$ to be at least k_A, k_B, k_C respectively, where $\tau' := \vartheta_{3.6}$.*

Proof. We want to show that every $n_C \times k_C$ sub-matrix of C' has min. singular value $\sigma_{k_C} \ge \delta = 1/\tau_C'$. 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 $\|z^T C_S'\|_2 < \delta$. Then, it is easy to see (from Lemma 3.4) that

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

for $\varepsilon_1 = \vartheta_{3.4}(\varepsilon + \delta)$. 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$_{1/(R\varepsilon_1)}(C) \ge k_C$. $\qquad\square$

Let us check that the conditions of the robust permutation lemma hold with C', C taking the roles of X, Y in Lemma 3.2, and $k = k_C$, and $\tau = \vartheta_{3.6} \cdot \tau_C$. From Lemma 3.6, it follows that K-rank$_\tau(C)$ and K-rank$_\tau(C')$ are both $\ge k$, and setting $\varepsilon' = 0$ in Lemma 3.4, the other condition of Lemma 3.2 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 3.2). Recall that K-rank$_\tau(X)$ and K-rank$_\tau(Y)$ are $\ge k$, 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 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 3.2. For smaller sets S (and this definition is crucial to avoiding an accumulation of errors), we define: $T_S := \bigcap\limits_{|S'|=(k-1), S' \supset S} T_{S'}$.

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 3.8 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 $\varepsilon_1 \cdot \vartheta_{3.8}$-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

Lemma 3.7. *In the above notation, for any $S \subseteq [R]$ of size $k-1$, $|T_S|$ is precisely $k-1$.*

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 A.3 (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 A.2). Thus $|T_S| = k - 1$. $\square$

The next two lemmas are crucial to the analysis. The first is our main linear algebraic lemma.

Lemma 3.8. *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 $\vartheta_{3.8} \cdot \varepsilon$ close $\mathrm{span}(X_A)$, where $\vartheta_{3.8} := 4n\tau\rho$.*

We postpone the proof to the end of the section. A counting argument lies at the core of the claim. We present it in the language of sunflower set systems.

Definition 3.9 (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 3.10. *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 proof is by a counting argument which we also defer to the end of the section. With these lemmas in place, we can prove the main inductive claim.

Proof of Lemma 3.2. 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 3.7.

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 $\varepsilon_1 \cdot \vartheta_{3.8}$-close to $\text{span}(X_S)$. We claim that $W_i \cap W_j \subseteq T^*$ for any $i \neq j \in S$. 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 3.8, 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 3.10 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 $\varepsilon_1 \cdot \vartheta_{3.8}$-close to $\text{span}(X_S)$, and thus $|T^*| \leq |S|$ (by Lemma A.2), and thus we have $|T^*| = |S|$. Now we have equality in Lemma 3.10, 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 theorem follows by applying to singleton sets. $\qquad\square$

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

Proof of Lemma 3.8. 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:

$$y = u_1 + \sum_{j \neq 1} \alpha_{1j} x_j + z_1$$

$$y = u_2 + \sum_{j \neq 2} \alpha_{2j} x_j + z_2$$

$$\vdots$$

$$y = u_q + \sum_{j \neq q} \alpha_{rj} x_j + z_q,$$

where $u_i \in \text{span}(X_A)$ 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

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

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 A.1 (since $\text{K-rank}_\tau(X) \geq k$), 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

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

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

Proof of Lemma 3.10. 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 $i' \ne i$. Call the number of elements of the latter kind t_i. Then we must have

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

Now since all $T_i \subseteq [R]$, we have
$$\sum_i t_i + |T^*| \le R.$$

Combining the two, we obtain

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

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. $\qquad\square$

3.3   Wrapping up the proof

We are now ready to complete the robust Kruskal's theorem. From what we saw above, the main part that remains is to prove that the permutations in the various *dimensions* are equal.

Proof of Theorem 3.1. 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 $k_A, k_B, k_C \ge 2$ (since $k_A, k_B, k_C \le R$, and $k_A + k_B + k_C \ge 2R + 2$). Thus from the Lemmas 3.6 and 3.4 (setting $\varepsilon' = 0$), we obtain that C, C' satisfy the hypothesis of the Robust permutation lemma (Lemma 3.2) with C', C set to X, Y respectively, and the parameters

$$\text{“}\tau\text{”} := \vartheta_{3.6} \ ; \quad \text{“}\varepsilon\text{”} := \vartheta_{3.4}\varepsilon.$$

Hence, we apply Lemma 3.2 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 $\varepsilon_2 = \vartheta_{3.2}\vartheta_{3.4} \cdot \varepsilon$,

$$\left\|A' - A\Pi_A\Lambda_A\right\|_F < \varepsilon_2, \ \ \left\|B' - B\Pi_B\Lambda_B\right\|_F < \varepsilon_2 \ \text{ and } \ \left\|C' - C\Pi_C\Lambda_C\right\|_F < \varepsilon_2 \tag{8}$$

We now 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 just use a robust analog of Kruskal's argument in the exact case. However, error analysis is a little tedious and we refer the interested reader to the appendix B.3 for details.

$\qquad\square$

3.4 Uniqueness Theorem for Higher Order Tensors

We show the uniqueness theorem for higher order tensors by a reduction to third order tensors as in [SB00]. 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 3.11 (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 ith column equal to $M_i = A_i \otimes B_i$ (viewed as a vector) is the Khatri-Rao product.

The following Lemma 3.12 (proof in the appendix A.1) 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 3.12 (K-rank of the Khatri-Rao product). *For two matrices A, B with R columns with robust K-rank $k_A = K\text{-}rank_{\tau_1}(A)$ and $k_B = K\text{-}rank_{\tau_2}(B)$, the K-rank of the Khatri-Rao product $M = A \odot B$ is super-additive:*

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

Remark 3.13. While the above lemma is tight in the worst case (see appendix A.1), we could expect that *typically*, the robust K-rank of the $A \odot B$ grows multiplicatively (for $R = \Omega(k_A k_B)$), since the vectors are now in $n_A n_B$ dimensions. In fact, we show that this can be formalized in our subsequent work with Ankur Moitra [BCMV13], by considering a smoothed setting. This at once leads to uniqueness results under weaker K-rank conditions for tensors of order > 3 in the smoothed analysis setting.

Theorem 3.14 (Uniqueness of Decompositions for Higher Orders). *Suppose we are given an order ℓ tensor (with $\ell \leq R$), $T = [U^{(1)} U^{(2)} \ldots U^{(\ell)}]$, where $\forall j \in [\ell]$ the n_j-by-R matrix $U^{(j)}$ is ρ_j-bounded, with $K\text{-}rank_{\tau_j}(U^{(j)}) = k_j \geq 2$ satisfying*

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

Then for every $0 < \varepsilon' < 1$, there exists $\varepsilon = \left(\vartheta_{3.14}^{(\ell)}\left(\frac{\varepsilon'}{R} \right) \right) \cdot \left(\prod_{j\in[\ell]} \vartheta_{3.14}(\tau_j, \rho_j, \rho_j', n_j) \right)^{-1}$ such that, for any other $(\rho_1', \rho_2', \ldots, \rho_\ell')$-bounded decomposition $[V^{(1)} V^{(2)} \ldots V^{(\ell)}]$ 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

$$\left\| \prod_{j\in[\ell]} \Lambda^{(j)} - I \right\|_F \leq \varepsilon' \quad and \; \forall j \in [\ell], \; \left\| V^{(j)} - U^{(j)}\Pi\Lambda^{(j)} \right\|_F \leq \varepsilon' \tag{9}$$

Setting $\vartheta_{3.14}^{(\ell)}(x) = x^{2^\ell}$ and $\vartheta_{3.14}(\tau_j, \rho_j, \rho_j', n_j) = (\tau_j\rho_j\rho_j'n_j)^{O(1)}$ 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 $[U^{(1)} \ U^{(2)} \ \ldots \ U^{(\ell)}]$ and $[V^{(1)} \ V^{(2)} \ \ldots \ V^{(\ell)}]$ are close, we conclude that $[U^{(1)} \ U^{(2)} \ \ldots \ (U^{(\ell-1)} \odot U^{(\ell)})]$ and $[V^{(1)} \ V^{(2)} \ \ldots \ (V^{(\ell-1)} \odot V^{(\ell)})]$ are close, and use the inductive hypothesis, which holds because of Lemma 3.12 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 A.4 – 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 of Theorem 3.14. We will prove by induction on ℓ. The base case of $\ell = 3$ is established by Theorem 3.1. 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 3.14 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

$$\left\| \sum_{r \in [R]} U_r^{(1)} \otimes U_r^{(2)} \otimes \cdots \otimes U_r^{(\ell-2)} \otimes \tilde{U}_r - \sum_{r \in [R]} V_r^{(1)} \otimes V_r^{(2)} \otimes \cdots \otimes V_r^{(\ell-2)} \otimes \tilde{V}_r \right\|_F < \varepsilon_\ell \quad (10)$$

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 3.12, 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 A.4. To ensure these, we will set

$$\varepsilon_\ell^{-1} = \vartheta_{3.14}^\ell \left(\frac{R}{\varepsilon'_\ell}\right) \cdot \left(\prod_{j \in [\ell-2]} \vartheta_{3.14}(\tau_j, \rho_j, \rho'_j, n_j)\right) \vartheta_{3.14}(\tilde{\tau}, \tilde{\rho}, \tilde{\rho}', \tilde{n}),$$

where $\vartheta_{3.14}^\ell = x^{O(2^\ell)}$. 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.

14

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 (9) 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 A.4, 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 A.4. 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 A.4 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 A.4 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 (9) is satisfied. Finally, Lemma A.4 shows that

$$\forall j \in \{\ell-1, \ell\} \quad \left\| V_r^{(j)} - U_{\pi(r)}^{(j)} \Lambda^{(j)}(r) \right\|_2 < \sqrt{\delta_r}, \quad \forall r \in [R]$$

$$\left\| V^{(j)} - U^{(j)} \Pi \Lambda^{(j)} \right\|_F < R^{1/4} \sqrt{\varepsilon'_{\ell-1}} \quad (\text{ by Cauchy-Schwartz inequality}).$$

$$< \varepsilon'_\ell$$

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 5).

Corollary 3.15 (Unique Symmetric Decompositions). *For every $0 < \eta < 1$, $\tau, \rho, \rho' > 0$ and $\ell, R \in \mathbb{N}$, $\exists \varepsilon_\ell = \vartheta_{3.15}^{(\ell)}(\frac{1}{\eta}, R, n, \tau, \rho, \rho')$ 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 $K\text{-rank}_\tau(U) = k \geq \frac{2R-1}{\ell} + 1$, and for any other ρ' bounded, symmetric, rank-R decomposition of T which is ε-close, i.e.,

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

15

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

$$\|V - U\Pi\|_F \leq \eta \tag{11}$$

The mild intricacy here is that applying Theorem 3.14 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 A.2.

4 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 [Lan12]. 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 4.1 (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

$$T = \sum_{i=1}^{R} a_i \otimes b_i \otimes c_i + \mathcal{N}, \tag{12}$$

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 3.1, 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 4.2. *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 4.3. *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. $\qquad \square$

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 4.4. *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

$$T - \mathcal{N} - \sum_i \widetilde{a}_i \otimes \widetilde{b}_i \otimes \widetilde{c}_i = \left(\sum_i a_i \otimes b_i \otimes c_i - \widetilde{a}_i \otimes b_i \otimes c_i \right)$$
$$+ \left(\sum_i \widetilde{a}_i \otimes b_i \otimes c_i - \widetilde{a}_i \otimes \widetilde{b}_i \otimes c_i \right)$$
$$+ \left(\sum_i \widetilde{a}_i \otimes \widetilde{b}_i \otimes c_i - \widetilde{a}_i \otimes \widetilde{b}_i \otimes \widetilde{c}_i \right).$$

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:

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

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

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

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 4.5. *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 $\text{poly}(n) \cdot \exp(O(\ell R^2) \log(\ell R\rho/\varepsilon))$.*

4.1 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 4.4 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 [BPR96] (which are based on the decidability of the *existential theory of reals*). The best algorithms here end up giving a running time which is $\exp(\mathrm{poly}(R))$ 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.

5 Polynomial Identifiability of Latent Variable and Mixture Models

We now show how our robust uniqueness theorems for tensor decompositions can be used for learning latent variable models, with polynomial sample complexity bounds.

Definition 5.1 (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 $N \leq \mathrm{poly}(m, 1/\eta)$ samples and finds with probability $1 - o(1)$ estimates of the hidden variables Υ' such that $\|\Upsilon' - \Upsilon\|_\infty < \eta$.

Consider a simple mixture-model, where each sample is generated from mixture of R distributions $\{\mathcal{D}_r\}_{r \in [R]}$, with mixing probabilities $\{w_r\}_{r \in [R]}$. Here the latent variable h corresponds to the choice of distribution and it can have $[R]$ possibilities. First the distribution $h = r$ is picked with probability w_r, and then the data is sampled according to $\mathcal{D}_r$, which has parameter $\mu_r \in \mathbb{R}^n$ (say its mean). Let $M_{n \times R}$ represent the matrix of these R parameters. The goal is to learn these hidden parameters (M and weights $\{w_r\}$) after observing many samples. This setting captures many latent variable models including topic models, HMMs, gaussian mixtures etc.

While practitioners typically use Expectation-Maximization (EM) methods to learn the parameters, a good alternative in the case of mixture models is *the method of moments* approach (

starting from the work by Pearson [Pea94] for univariate gaussians), which tries to identify the parameters by estimating higher order moments. However, one drawback is that the number of moments required is typically as large as the number of mixtures R (or parameters), resulting in a sample complexity that is exponential in R [MV10, BS10, FOS05, FSO06, RSS14].

In a recent exciting line of work [MR06, AHK12, HK12, AFH+12, AGH+12], it is shown that poly(R, n) samples suffice for identifiability in a special case called the *non-singular* or *non-degenerate* case i.e. when the matrix M has full rank (rank $= R$)^2 for many of these models. Their algorithms for this case proceed by reducing the problem of finding the latent variables (the means and weights) to the problem of decomposing *Symmetric Orthogonal Tensors* of order 3, which are known to be solvable in poly(n, R) time using power-iteration type methods [KR01, ZG01, AGH+12]. However, their approach crucially relies on these non-degeneracy conditions, and are not robust: even in the case when these R-means reside in a $(R-1)$-dimensional space, these algorithms fail, and the best known sample complexity upper bounds for these problems are $\exp(R)\text{poly}(n)$.

In many settings like speech recognition and image classification, the dimension of the feature space n is typically much smaller than R, the number of topics or clusters. For instance, the (effective) feature space corresponds to just the low-frequency components in the fourier spectrum for speech, or the local neighborhood of a pixel in images (SIFT features [Low99]). These are typically much smaller than the different kinds of objects or patterns (topics) that are possible. Further, in other settings, the set of relevant features (the effective feature space) could be a space of much smaller dimension $(k < R)$ that is unknown to us even when the feature vectors are actually represented in a large dimensional space $(n \gg R)$.

In this section, we show that we can use our Robust Uniqueness results for Tensor Decompositions (Theorem 3.1 and Theorem 3.14) to go past the non-degeneracy barrier and prove that poly(R, n) samples suffice even under the milder condition that no $k = \delta R$ of these parameters (means) lie in a $(k-1)$ dimensional space (for some constant $\delta > 0$). These results apply to learning latent variable models like Hidden Markov models, Multi-view mixture models, Mixtues of Gaussians, Topic models etc. One interesting aspect of our approach is that, unlike previous works, we get a smooth tradeoff : we get polynomial identifiability under successively milder conditions by using higher order tensors $(\ell \approx 2/\delta)$. This reinforces the intuition that higher moments capture more information at the cost of efficiency.

In the rest of this section, we will first describe Multi-view models and show how the robust uniqueness theorems for tensor decompositions immediately imply polynomial identifiability in this model. We will then see two popular latent variable models which fit into the multi-view mixture model: the exchangeable (single) Topic Model and Hidden Markov models. We note that the results of this section (for $\ell = 3$ views) also apply to other latent variable models like *Latent Dirichlet Allocation (LDA)* and *Independent Component Analysis (ICA)* that were studied in [AGH+12]. We omit the details in this version of the paper.

5.1   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 Mod-

---

^{2}For polynomial identifiability, $\sigma_R \geq 1/\text{poly}(n)$.

els [AHK12], Hidden Markov Models (HMMs) [MR06, AMR09, AHK12], random graph mixtures [AMR09]. Allman et al [AMR09] refer to these models by *finite mixtures of finite measure products*. We first introduce some notation, along the lines of [AMR09, AHK12].

Definition 5.2 (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.

The following key lemma shows how to obtain a higher order tensor (to apply our results from previous sections) in terms of the hidden parameters that we need to recover. It follows easily because of conditional independence.

Lemma 5.3 ([AMR09, AHK12]). *In the notation established above for multi-view models, $\forall \ell \in \mathbb{N}$ the ℓth moment tensor*

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

In our usual representation of tensor decompositions,

$$\mathbb{E}\left[x^{(1)} \otimes \ldots x^{(j)} \otimes \ldots x^{(\ell)}\right] = \left[M^{(1)} \; M^{(2)} \; \ldots \; M^{(\ell)}\right].$$

In the worst case, the best known algorithms for this problem are from a recent work Rabani et al [RSS14], who give an algorithm that has complexity $R^{O(R^2)} + \text{poly}(n, R)$ (both running time and sample complexity). In fact they also show a sample complexity lower-bound of $\tilde{\Omega}(\exp(R))$ for learning multiview models in one dimension ($k = n = 1$). When $k \equiv \text{K-rank}_\tau(M) \geq R$ for each of these matrices (the non-degenerate or non-singular setting), Anandkumar et al. [AHK12] give a polynomial time algorithm to learn the hidden variables using only $\text{poly}(R, \tau, n)$ samples (hence polynomial identifiability). We now show how our tensor decomposition results immediately imply polynomial identifiability even when $k = \delta R$ for any constant $\delta > 0$.

Theorem 5.4 (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 5.2), with the parameters satisfying:*

(a) For each mixture $r \in [R]$, the mixture weight $w_r > \gamma$.

(b) For each $j \in [\ell]$, $K\text{-}rank_\tau(M^{(j)}) \geq k \geq \frac{2R}{\ell} + 1$.

then there is a algorithm that given any $\eta > 0$ uses $N = \vartheta_{5.4}{}^{(\ell)}\left(\frac{1}{\eta}, R, n, \tau, 1/\gamma\right)$ 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

$$\forall j \in [\ell], \quad \left\|M^{(j)} - \tilde{M}^{(j)}\right\|_F \leq \eta \quad and \quad \forall r \in [R], \ |w_r - \tilde{w}_r| < \eta \tag{13}$$

The function $\vartheta_{5.4}{}^{(\ell)}(\cdot, \ldots, \cdot) = \text{poly}(Rn/(\gamma\eta))^{2^\ell} \text{poly}(n, \tau, 1/\gamma)^\ell$ 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 $\text{poly}(1/w_r)$ 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 $K\text{-}rank_\tau(M^{(j)})$ have bounds k_j which are potentially different, and satisfy the same condition as in Theorem 3.14.*

Proof Sketch. We will consider the ℓth moment tensor for $\ell = \lceil 2/\delta \rceil + 1$. The proof is simple, and proceeds in three steps. First, we use enough samples to obtain an estimate $\tilde{T}$ of the ℓth moment tensor T, upto inverse polynomial error. Then we find a good rank-R approximation to $\tilde{T}$ (it exists because T has rank R). We then use the Robust Uniqueness theorem for tensor decompositions to claim that the terms of this decomposition are in fact close to the hidden parameters. The error analysis is a little tedious, but straightforward, so we defer it to the appendix C. $\qquad\square$

We will now see two popular latent variable models which fit into the multi-view mixture model: the exchangeable (single) Topic Model and Hidden Markov models. While the topic model fits directly into the multi-view framework, in the case of Hidden Markov Models we need to come up these independent views cleverly to apply our results. We note that the results of this section (for $\ell = 3$ views) also apply to other latent variable models like *Latent Dirichlet Allocation (LDA)* and *Independent Component Analysis (ICA)* that were studied in [AGH$^+$12]. Anandkumar et al. [AFH$^+$12, AGH$^+$12] show how we can obtain third order tensors by looking at "third" moments and applying suitable transformations. Applying our robust uniqueness theorem (Theorem 3.1) to these 3-tensors identify the parameters. We omit the details in this version of the paper.

5.2 Exchangeable (single) Topic Model

The simplest latent variable model that fits the multi-view setting is the Exchangeable Single Topic model as given in [AHK12]. This is a simple bag-of-words model for documents, in which the words in a document are assumed to be exchangeable. This model can be viewed as first picking the topic $r \in [R]$ of the document, with probability w_r. Given a topic $r \in [R]$, each word in the document is sampled independently at random according to the probability distribution $\mu_r \in \mathbb{R}^n$ (n is the dictionary size). In other words, the topic $r \in R$ is a latent variable such that the ℓ words in a document are conditionally i.i.d given r.
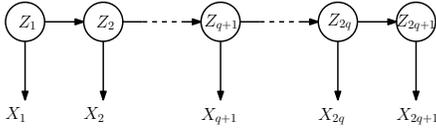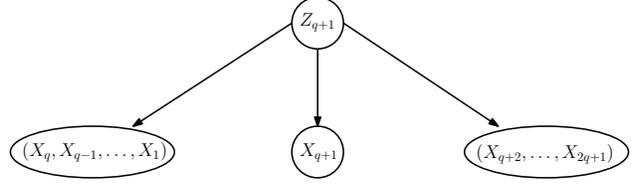
Figure 1: An HMM with $2q + 1$ time steps.

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

The views in this case correspond to the words in a document. This is a special case of the multi-view model since the distribution of each of the views $j \in [\ell]$ is identical. As in [AHK12, AGH$^+$12], we will represent the ℓ words in a document by indicator vectors $x^{(1)}, x^{(2)}, \ldots, x^{(\ell)} \in \{0,1\}^n$ ($c_{max} = 1$ here). Hence, the $(i_1, i_2, \ldots, i_\ell)$ entry of the tensor $\mathbb{E}\left[x^{(1)} \otimes x^{(2)} \otimes \ldots x^{(\ell)}\right]$ corresponds to the probability that the first words is i_1, the second word is i_2, ... and the ℓth word is i_ℓ. The following is a simple corollary of Theorem 5.4.

Corollary 5.5 (Polynomial Identifiability of Topic Model)**.** *The following statement holds for any constant $\delta > 0$. Suppose we are given documents generated by the topic model described above, where the topic probabilities of the R topics are $\{w_r\}_{r\in[R]}$, and the probability distribution of words in a topic r are given by $\mu_r \in \mathbb{R}^n$ (represented as a n-by-R matrix M). If $\forall r \in [R]$ $w_r > \gamma$, and if $K\text{-rank}_\tau(M) \geq k \geq 2R/\ell + 1$,*
then there is a algorithm that given any $\eta > 0$ uses $N = \vartheta_{5.4}{}^{(\ell)}\left(\frac{1}{\eta}, R, n, \tau, 1/\gamma, 1\right)$ samples, and finds with high probability M' and $\{w_r'\}_{r\in[R]}$ such that

$$\left\|M - M'\right\|_F \leq \eta \quad and \quad \forall r \in [R], \; \left|w_r - w_r'\right| < \eta \tag{14}$$

5.3 Hidden Markov Models

The next latent variable model that we consider are (discrete) Hidden Markov Model which is extensively used in speech recognition, image classification, bioinformatics etc. We follow the same setting as in [AMR09]: there is a hidden state sequence $Z_1, Z_2, \ldots, Z_m$ taking values in $[R]$, that forms a stationary Markov chain $Z_1 \rightarrow Z_2 \rightarrow \cdots \rightarrow Z_m$ 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[3] $\{1, 2, \ldots, n\}$ 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 rth 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.

Corollary 5.6 (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 :*

---

[3]in general, we can also allow x_t to be certain continuous distributions like multivariate gaussians

(a) The stationary distribution $\{w_r\}_{r\in[R]}$ has $\forall r \in [R]$ $w_r > \gamma_1$,

(b) The observation matrix M has $K\text{-rank}_\tau(M) \geq k \geq \delta R$,

(c) 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 $N = \vartheta_{5.4}^{(\frac{1}{\delta}+1)}\left(\frac{1}{\eta}, R, n, \tau, \frac{1}{\gamma_1\gamma_2}\right)$ 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

$$\left\|M - M'\right\|_F \leq \eta, \quad \left\|P - P'\right\|_F \leq \eta \quad and \; \forall r \in [R], \; |w_r - \tilde{w}_r| < \eta \tag{15}$$

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.

Proof sketch. Here, the idea is to cleverly come up with three independent of the HMM, so that it fits into our framework. The proof follows along the lines of Allman et al [AMR09], so we only sketch the proof here. We now show to cast this HMM into a multi-view model (Def. 5.2) using a nice trick of [AMR09]. We can then apply Theorem 5.4 and prove identifiability (Corollary 5.6). 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 5.2. 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 3.12). Showing that these are indeed the transition matrices is fairly straightforward, and we refer to Allman et al. [AMR09] for the details.

$$A = ((\ldots(M\tilde{P}) \odot M)\tilde{P}) \odot M)\ldots\tilde{P}) \odot M)\tilde{P} \tag{16}$$
$$B = M \tag{17}$$
$$C = ((\ldots(MP) \odot M)P) \odot M)\ldots P) \odot M)P \tag{18}$$

(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 $K\text{-rank}_\tau(M) \geq k$, and thus $K\text{-rank}_{\tau_2\tau}(MP) \geq k$ (because P is a stochastic matrix with all eigenvalues $\geq \tau_2$). Now by the property of the Khatri-Rao product (Lemma 3.12), we have $K\text{-rank}_{(\tau\tau_2)\tau}(C^{(2)}) \geq \min\{R, 2k\}$. We can continue this argument, to eventually conclude that $K\text{-rank}_{\tau'}(C^{(q)}) = \min\{R, qk\} = R$ for $\tau' = \tau^q\gamma_2^{q^2}(qk)^{q/2}$.

Precisely the same argument lets us conclude that $K\text{-rank}_{\tau'}(A) \geq R$, for the $\tau' = \tau^q\gamma_2^{q^2}(qk)^{q/2}$. Now since $K\text{-rank}_\tau(B) \geq 2$, we have that the conditions of Theorem 3.1 hold. Now using the

arguments of Theorem 5.4 (here, we use Theorem 3.1 instead of Theorem 3.14), 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 $\delta = \text{poly}(1/\eta, \dots)$. 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 [AMR09] by a trick which is similar in spirit to Lemma A.4. 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. $\quad\square$

Remark: *Allman et al. [AMR09] show identifiability under weaker conditions than Corollary 5.6 when they have infinite samples. This is because they prove their results for* generic *values of the parameters M, P (this formally means their results hold for all M, P except a set of measure zero, but they do not give an explicit characterization). Our bounds are weaker, but hold whenever the $K\text{-rank}_\tau(M) \geq \delta n$ condition holds. Further, the main advantage is that our result is* robust *to noise: the case when we only have finite samples.*

5.4 Mixtures of Spherical Gaussians

Suppose we have a mixture of R spherical gaussians in $\mathbb{R}^n$, with mixing weights $w_1, w_2, \dots w_R$, means $\mu_1, \mu_2, \dots, \mu_r$, and the common variance σ^2. Let us denote this mixture distribution by $\mathcal{D}$, and the $n \times R$ matrix of means by M.

We define the μ-tensor of ℓth order to be

$$\text{Mom}_\ell := \sum_i w_i \mu_i^{\otimes \ell}.$$

The empirical mean $\mu := \text{Mom}_1$, and can be estimated by drawing samples $x \sim D$, and computing $\mathbb{E}[x]$. Similarly, we will show how to compute Mom_ℓ for larger ℓ by computing higher order moment tensors, assuming we know the value of σ. We can then use the robust Kruskal's theorem (Theorem 3.14) and the sampling lemma (Lemma F.2) to conclude the following theorem.

Theorem 5.7. *Suppose we have a mixture of gaussians given by $\mathcal{D}$, with hidden parameters $\{w_r\}_{r \in [R]}$ and M (in particular, we assume we know σ)[4]. Suppose also that $\forall r \in [R]$ $w_r > \gamma$, and $K\text{-rank}_\tau(M) = k$ for some $k \geq \delta R$.*

Then there is a algorithm that given any $\eta > 0$ and σ, uses $N = \vartheta_{5.7}^{(1/\delta)}\left(\frac{1}{\eta}, R, n, \tau, 1/\gamma\right)$ samples drawn from $\mathcal{D}$, and finds with high probability M' and $\{w_r'\}_{r \in [R]}$ such that

$$\left\|M - M'\right\|_F \leq \eta \quad \text{and} \quad \forall r \in [R], \ \left|w_r - w_r'\right| < \eta \tag{19}$$

---

[4]As will be clear, it suffices to know it up to an inverse polynomial error, so from an algorithmic viewpoint, we can "try all possible" values.

Proof. This will follow the same outline as Theorem 5.4. So, we sketch the proof here. The theorem works for error polynomial $\vartheta_{5.7}$ being essentially as the same error polynomial in $\vartheta_{5.4}$. However, we first need to gain access to an order ℓ-tensor, where each rank-1 term corresponds to a mean μ_r. Hence, we show how to obtain this order-ℓ tensor of means, by subtracting out terms involving σ, by our estimates of moments upto ℓ.

Pick $\ell = \lceil \frac{2}{\delta} \rceil + 2$. We will use order ℓ tensors given by the ℓth moment. We will first show how to obtain Mom_ℓ, from which we learn the parameters. The computation of Mom_ℓ will be done inductively. Note that Mom_1 is simply $\mathbb{E}[x]$. Now observe that

$$
\mathbb{E}\left[x^{\otimes 2}\right] = \mathbb{E}[x \otimes x] = \mathbb{E}\left[\sum_i w_i(\mu_i + \varepsilon_i) \otimes (\mu_i + \varepsilon_i)\right]
$$

$$
= \mathbb{E}\left[\sum_i w_i \mu_i \otimes \mu_i\right] + \mathbb{E}\left[\sum_i w_i \varepsilon_i \otimes \varepsilon_i\right]
$$

$$
= \mathrm{Mom}_2 + \sigma^2 I.
$$

We compute $\mathbb{E}\left[x^{\otimes 2}\right]$ by sampling, and since we know σ, we can find Mom_2 up to any polynomially small error. In general, we have

$$
\mathbb{E}\left[x^{\otimes \ell}\right] = \sum_i w_i \mathbb{E}\left[(\mu_i + \varepsilon_i)^{\otimes \ell}\right] \tag{20}
$$

$$
= \sum_i w_i \sum_{x_j \in \{\mu_i, \varepsilon_i\}} \mathbb{E}[x_1 \otimes x_2 \otimes \ldots \otimes x_\ell]. \tag{21}
$$

The last summation has 2^ℓ terms. One of them is $\mu_i^{\otimes \ell}$, which produces Mom_ℓ on the RHS. The other terms have the form $x_1 \otimes x_2 \otimes \ldots \otimes x_\ell$, where some of the x_i are μ_i and the rest ε_i, and there is at least one ε_i.

If a term has r terms being μ_i and $\ell - r$ being ε_i, the tensor obtained is essentially a *permutation* of $\widehat{\mu}(r, \ell) := \mu_i^{\otimes r} \otimes \varepsilon_i^{\otimes (\ell - r)}$. By permutation, we mean that the $(j_1, \ldots, j_\ell)$th entry of the tensor would correspond to the $(j_{\pi(1)}, \ldots, j_{\pi(\ell)})$th entry of $\widehat{\mu}(r, \ell)$, for some permutation π. Thus we focus on showing how to evaluate the tensor $\widehat{\mu}(r, \ell)$ for different r, ℓ.

Note that if $\ell - r$ is odd, we have that $\mathbb{E}[\widehat{\mu}(r, \ell)] = 0$. This is because the odd moments of a Gaussian with mean zero, are all zero (since it is symmetric). If we have $\ell - r$ being even, we can describe the tensor $\mathbb{E}\left[\varepsilon_i^{\ell - r}\right]$ explicitly as follows. Consider an index $(j_1, \ldots, j_{\ell - r})$, and bucket the j into groups of equal coordinates. For example for index $(1, 2, 3, 2)$, the buckets are $\{(1), (22), (3)\}$. Now suppose the bucket sizes are $b_1, \ldots, b_t$ (they add up to $\ell - r$). Then the $(j_1, \ldots, j_{\ell - r})$th entry of $\varepsilon_i^{\otimes (\ell - r)}$ is precisely the product $m_{b_1} m_{b_2} \ldots m_{b_t}$, where m_s is the sth moment of the univariate Gaussian $\mathcal{N}(0, \sigma^2)$.

The above describes the entries of $\mathbb{E}\left[\varepsilon_i^{\ell - r}\right]$. Now $\mathbb{E}[\widehat{\mu}(r, \ell)]$ is precisely $\mathrm{Mom}_r \otimes \mathbb{E}\left[\varepsilon_i^{\ell - r}\right]$ (since the μ_i is fixed). Thus, since we have inductively computed Mom_r for $r < \ell$, this gives a procedure to compute each entry of $\mathbb{E}[\widehat{\mu}(r, \ell)]$. Thus each of the 2^ℓ terms in the RHS of (20) except Mom_ℓ can be calculated using this process. The LHS can be estimated to any inverse polynomial small error by sampling (Lemma F.2). Thus we can estimate Mom_ℓ up to a similar error.

Hence, we can use the algorithm from Section 4 and apply Corollary 3.15 to obtain vectors

26

$\{\tilde{u}_r\}_{r \in [R]}$ such that

$$\forall r \in [R] \left\| u_r - w^{1/\ell} \mu_r \right\| < \eta.$$

Similarly, applying the same process with Mom_ℓ (the Kruskal conditions also hold for $\ell - 1$) we get η-close approximations to $w_i^{1/(\ell-1)} \mu_r$. Now, we appeal to Lemma D.1 to obtain $\{w_r, \mu_r\}_{r \in [R]}$. $\square$

Remark: Note that the previous proof worked even when the gaussians are not spherical: they just need to have the same known covariance matrix Σ.

6   Discussion and Open Problems

The most natural open problem arising from our work is that of computing approximate small rank decompositions efficiently. While the problem is NP hard in general, we suspect that *well conditioned* assumptions regarding robust Kruskal ranks being sufficiently large, as in the uniqueness theorem (Theorem 3.1) for decompositions of 3-tensors for instance, could help. In particular,

Question 6.1. *Suppose T is a 3-tensor, that is promised to have a rank R decomposition $[A\ B\ C]$, with $k_A = K\text{-}rank_\tau(A)$ (similarly k_B and k_C) satisfying $k_A + k_B + k_C \geq 2R + 2$. Can we find the decomposition A, B, C (up to a specified error ε) in time polynomial in n, R and $1/\varepsilon$?*

In the special case that the decomposition $[A\ B\ C]$ is known to be orthogonal (i.e., the columns of A, B, C are mutually orthogonal), which in particular implies $n \geq R$, then iterative methods like power iteration [AGH+12], and "alternating least squares" (ALS) [CLdA09] [5] converge in polynomial time.

A result in the spirit of finding alternative sufficient conditions for uniqueness was by Chiantini and Ottaviani [CO12], who use ideas from algebraic geometry (in particular a notion called *weak defectivity*), to prove that *generic* $n \times n \times n$ tensors of rank $k \leq n^2/16$ have a unique decomposition (here the word 'generic' is meant to mean all except a measure zero set of rank k tensors, which they characterize in terms of weak defectivity). Note that this is much stronger than the bound obtained by Kruskal's theorem, which is roughly $3n/2$. It is also roughly the best one can hope for, since every 3-tensor has rank at most n^2 (and a random tensor has rank $\geq n^2/2$). It would be very interesting to prove robust versions of their results, as it would imply identifiability for a much larger range of parameters in the models we consider.

A third question is that of certifying that a given decomposition is unique. Kruskal's rank condition, while elegant, is not known to be verifiable in polynomial time (unless we are certifying that the K-rank is essentially the rank). Given an $n \times R$ matrix, certifying that every k columns are linearly independent is known to be NP-hard [Kha95, TP12]. Even the average case version i.e. when the matrix is random with independent gaussian entries, has received much attention as it is related to certifying the Restricted Isometry Property (RIP), which plays a key role in compressed sensing [CT05, KZ11]. It is thus an fascinating open question to find uniqueness (and robust uniqueness) theorems which natural parameters that can be computed efficiently.

From the perspective of learning latent variable models, it would be very interesting to obtain efficient learning algorithms with polynomial running times for the settings considered in Section 5. Recall that we give algorithms which need only polynomial samples (in the dimension n, and

---

[5]This is the method of choice in practice for computing tensor decompositions.

number of mixtures R), when the parameters satisfy the robust Kruskal conditions. In subsequent work with Ankur Moitra [BCMV13], we address this partially and give efficient algorithms through tensor decompositions in a smoothed analysis setting. Finally, we believe that our approach can be extended to learning the parameters of general mixtures of gaussians [MV10, BS10], and more generally to a broader class of problems on learning probabilistic problems.

7   Acknowledgements

References

[AFH$^+$12]   Anima Anandkumar, Dean Foster, Daniel Hsu, Sham Kakade, and Yi-Kai Liu. A spectral algorithm for latent dirichlet allocation. In *Advances in Neural Information Processing Systems 25*, pages 926–934, 2012.

[AGH$^+$12]   Anima Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *arXiv preprint arXiv:1210.7559*, 2012.

[AGHK13]   Anima Anandkumar, Rong Ge, Daniel Hsu, and Sham M Kakade. A tensor spectral approach to learning mixed membership community models. *arXiv preprint arXiv:1302.2684*, 2013.

[AGM12]   Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *Foundations of Computer Science (FOCS), 2012 IEEE 53rd Annual Symposium on*, pages 1–10. IEEE, 2012.

[AHHK12]   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, 2012.

[AHK12]   Animashree Anandkumar, Daniel Hsu, and Sham M Kakade. A method of moments for mixture models and hidden markov models. *arXiv preprint arXiv:1203.0683*, 2012.

[AK01]   Sanjeev Arora and Ravi Kannan. Learning mixtures of arbitrary gaussians. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 247–257. ACM, 2001.

[AM05]   Dimitris Achlioptas and Frank McSherry. On spectral learning of mixtures of distributions. In *Learning Theory*, pages 458–469. Springer, 2005.

[AMR09]   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.

[APRS11]   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.

[AS12]   Pranjal Awasthi and Or Sheffet. Improved spectral-norm bounds for clustering. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 37–49. Springer, 2012.

[AY09]   Evrim Acar and Bülent Yener. Unsupervised multiway data analysis: A literature survey. *Knowledge and Data Engineering, IEEE Transactions on*, 21(1):6–20, 2009.

[BCMV13]   Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. *Manuscript*, 2013.

[BIWX11]   Arnab Bhattacharyya, Piotr Indyk, David P. Woodruff, and Ning Xie. The complexity of linear dependence problems in vector spaces. In Chazelle [Cha11], pages 496–508.

[BPR96]   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.

[Bro97]   Rasmus Bro. Parafac. tutorial and applications. *Chemometrics and intelligent laboratory systems*, 38(2):149–171, 1997.

[BS10]   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.

[BV09]   S Charles Brubaker and Santosh S Vempala. Random tensors and planted cliques. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques*, pages 406–419. Springer, 2009.

[Cat44]   Raymond B Cattell. 'parallel proportional profiles and other principles for determining the choice of factors by rotation. *Psychometrika*, 9(4):267–283, 1944.

[CC70]   J Douglas Carroll and Jih-Jie Chang. Analysis of individual differences in multidimensional scaling via an n-way generalization of eckart-young decomposition. *Psychometrika*, 35(3):283–319, 1970.

[Cha96]   Joseph T Chang. Full reconstruction of markov models on evolutionary trees: identifiability and consistency. *Mathematical biosciences*, 137(1):51–73, 1996.

[Cha11]   Bernard Chazelle, editor. *Innovations in Computer Science - ICS 2010, Tsinghua University, Beijing, China, January 7-9, 2011. Proceedings*. Tsinghua University Press, 2011.

[CLdA09]   P. Comon, X. Luciani, and A. L. F. de Almeida. Tensor decompositions, alternating least squares and other tales. *Journal of Chemometrics*, 23(7-8):393–405, 2009.

[CO12]   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.

[CT05]   E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theor.*, 51(12):4203–4215, December 2005.

[Das99]   Sanjoy Dasgupta. Learning mixtures of gaussians. In *Foundations of Computer Science, 1999. 40th Annual Symposium on*, pages 634–644. IEEE, 1999.

[DDF⁺90]   Scott Deerwester, Susan T. Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[DDS⁺09]   Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[DLCC07]   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.

[DLR77]   Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 1–38, 1977.

[dlVKKV05] W Fernandez de la Vega, Marek Karpinski, Ravi Kannan, and Santosh Vempala. Tensor decomposition and approximation schemes for constraint satisfaction problems. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 747–754. ACM, 2005.

[DM07]   Petros Drineas and Michael W Mahoney. A randomized algorithm for a tensor-based generalization of the singular value decomposition. *Linear algebra and its applications*, 420(2):553–571, 2007.

[DS07]   Sanjoy Dasgupta and Leonard Schulman. A probabilistic analysis of em for mixtures of separated, spherical gaussians. *The Journal of Machine Learning Research*, 8:203–226, 2007.

[DSL08]   Vin De Silva and Lek-Heng Lim. Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3):1084–1127, 2008.

[Edd96]   Sean R Eddy. Hidden markov models. *Current opinion in structural biology*, 6(3):361–365, 1996.

[EY36]   Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[FOS05]   Jon Feldman, Ryan ODonnell, and Rocco A. Servedio. Learning mixtures of product distributions over discrete domains. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '05, pages 501–510, Washington, DC, USA, 2005. IEEE Computer Society.

[FSO06]   Jon Feldman, Rocco A. Servedio, and Ryan O'Donnell. Pac learning axis-aligned mixtures of gaussians with no separation assumption. In *Proceedings of the 19th annual conference on Learning Theory*, COLT'06, pages 20–34, Berlin, Heidelberg, 2006. Springer-Verlag.

[Gha04]   Zoubin Ghahramani. Unsupervised learning. In *Advanced Lectures on Machine Learning*, pages 72–112. Springer, 2004.

[GHP07]   Gregory Griffin, Alex Holub, and Pietro Perona. Caltech-256 object category dataset. 2007.

[GLPR12]   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.

[GVL12]   Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHUP, 2012.

[GVX13]   Navin Goyal, Santosh Vempala, and Ying Xiao. Fourier pca. *CoRR*, abs/1306.5825, 2013.

[GY08]   Mark Gales and Steve Young. The application of hidden markov models in speech recognition. *Foundations and Trends in Signal Processing*, 1(3):195–304, 2008.

[Har70]   Richard A Harshman. Foundations of the parafac procedure: models and conditions for an explanatory multimodal factor analysis. 1970.

[Hås90]   Johan Håstad. Tensor rank is np-complete. *Journal of Algorithms*, 11(4):644–654, 1990.

[Hit27]   Frank Lauren Hitchcock. The expression of a tensor or a polyadic as a sum of products. *J. Math. Phys.*, 6:164–189, 1927.

[HK12]   Daniel Hsu and Sham M Kakade. Learning gaussian mixture models: Moment methods and spectral decompositions. *arXiv preprint arXiv:1206.5766*, 2012.

[HK13]   Daniel Hsu and Sham M Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Proceedings of the 4th conference on Innovations in Theoretical Computer Science*, pages 11–20. ACM, 2013.

[HKL12]   Daniel Hsu, Sham Kakade, and Percy Liang. Identifiability and unmixing of latent parse trees. In *Advances in Neural Information Processing Systems 25*, pages 1520–1528, 2012.

[HKZ12]   Daniel Hsu, Sham M Kakade, and Tong Zhang. A spectral algorithm for learning hidden markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

[HL13]   Christopher Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *arXiv preprint arXiv:0911.1393v4*, 2013.

[JS04]   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.

[KB09]   Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. *SIAM review*, 51(3):455–500, 2009.

[Kha95]   Leonid Khachiyan. On the complexity of approximating extremal determinants in matrices. *J. Complex.*, 11(1):138–153, March 1995.

[Kie00]   Henk AL Kiers. Towards a standardized notation and terminology in multiway analysis. *Journal of chemometrics*, 14(3):105–122, 2000.

[KK10]   Amit Kumar and Ravindran Kannan. Clustering with spectral norm and the k-means algorithm. In *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*, pages 299–308. IEEE, 2010.

[KM11]   Tamara G Kolda and Jackson R Mayo. Shifted power method for computing tensor eigenpairs. *SIAM Journal on Matrix Analysis and Applications*, 32(4):1095–1124, 2011.

[KMV10]   Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant. Efficiently learning mixtures of two gaussians. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 553–562. ACM, 2010.

[Knu]   Donald Knuth. The art of computer programming vol. 2, (1997).

[KR01]   Eleftherios Kofidis and Phillip A. Regalia. On the best rank-1 approximation of higher-order supersymmetric tensors. *SIAM J. Matrix Anal. Appl.*, 23(3):863–884, March 2001.

[Kru]   JB Kruskal. Statement of some current results about three-way arrays, 1983. *Unpublished manuscript, AT&T Bell Labs, Murray Hill, NC. Pdf available from http://three-mode. leidenuniv. nl.*

[Kru77]   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.

[KZ11]   Pascal Koiran and Anastasios Zouzias. On the certification of the restricted isometry property. *CoRR*, abs/1103.4984, 2011.

[Lan12]   J.M. Landsberg. *Tensors:: Geometry and Applications*. Graduate Studies in Mathematics Series. American Mathematical Society, 2012.

[Low99]   David G Lowe. Object recognition from local scale-invariant features. In *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, volume 2, pages 1150–1157. Ieee, 1999.

[LRA93]   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.

[MR06]   Elchanan Mossel and Sébastien Roch. Learning nonsingular phylogenies and hidden markov models. *The Annals of Applied Probability*, pages 583–614, 2006.

[MV10]   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.

[Paa00]   Pentti Paatero. Construction and analysis of degenerate parafac models. *Journal of Chemometrics*, 14(3):285–299, 2000.

[Pea94]   Karl Pearson. Contributions to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society of London. A*, 185:71–110, 1894.

[Rho10]   John A Rhodes. A concise proof of kruskal's theorem on tensor decomposition. *Linear Algebra and Its Applications*, 432(7):1818–1824, 2010.

[RS12]   John A Rhodes and Seth Sullivant. Identifiability of large phylogenetic mixture models. *Bulletin of mathematical biology*, 74(1):212–231, 2012.

[RSS14]   Yuval Rabani, Leonard Schulman, and Chaitanya Swamy. Learning mixtures of arbitrary distributions over large discrete domains. In *ITCS*, 2014.

[SB00]   Nicholas D Sidiropoulos and Rasmus Bro. On the uniqueness of multilinear decomposition of n-way arrays. *Journal of chemometrics*, 14(3):229–239, 2000.

[SC10]   Alwin Stegeman and Pierre Comon. Subtracting a best rank-1 approximation may increase tensor rank. *Linear Algebra and its Applications*, 433(7):1276–1300, 2010.

[SS07]   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.

[tBS02]   Jos MF ten Berge and Nikolaos D Sidiropoulos. On uniqueness in candecomp/parafac. *Psychometrika*, 67(3):399–409, 2002.

[TC82]   GM Tallis and P Chesson. Identifiability of mixtures. *J. Austral. Math. Soc. Ser. A*, 32(3):339–348, 1982.

[Tei61]   Henry Teicher. Identifiability of mixtures. *The annals of Mathematical statistics*, 32(1):244–248, 1961.

[Tei67]   Henry Teicher. Identifiability of mixtures of product measures. *The Annals of Mathematical Statistics*, 38(4):1300–1302, 1967.

[TP12]   A. M. Tillmann and M. E. Pfetsch. The Computational Complexity of the Restricted Isometry Property, the Nullspace Property, and Related Concepts in Compressed Sensing. *ArXiv e-prints*, May 2012.

[VW04]   Santosh Vempala and Grant Wang. A spectral algorithm for learning mixture models. *Journal of Computer and System Sciences*, 68(4):841–860, 2004.

[WGPY97] PC Woodland, MJF Gales, D Pye, and SJ Young. The development of the 1996 htk broadcast news transcription system. In *DARPA speech recognition workshop*, pages 73–78, 1997.

[YEG$^+$02] Steve Young, Gunnar Evermann, Mark Gales, Thomas Hain, Dan Kershaw, Xunying Liu, Gareth Moore, Julian Odell, Dave Ollason, Dan Povey, et al. The htk book. *Cambridge University Engineering Department*, 3, 2002.

[ZG01] Tong Zhang and Gene H. Golub. Rank-one approximation to high order tensors. *SIAM J. Matrix Anal. Appl.*, 23(2):534–550, February 2001.

A   A Medley of Auxiliary Lemmas

We now list some of the (primarily linear algebra) lemmas we used in our proofs. They range in difficulty from trivial to 'straightforward', but we include them for completeness.

Lemma A.1. *Suppose X is a matrix in $\mathbb{R}^{n \times k}$ with $\sigma_k \geq 1/\tau$. Then if $\left\| \sum_i \alpha_i X_i \right\|_2 < \varepsilon$, 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 A.2. *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:*

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

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 $S = \{i : \|\Pi_{\mathcal{S}}^{\perp} A_i\| \leq \varepsilon\}$ 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 $1/\tau$.

$\square$

Lemma A.3. *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

$$\Pr\left[|\langle u_i, r \rangle| > \frac{\varepsilon}{10t}\right] \quad \text{for all } i. \tag{22}$$

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. (22) holds. This implies the lemma (in fact we obtain $\sqrt{d}$ in the denominator). $\square$

Lemma A.4. *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*

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

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

$$\left\| u \otimes v - u' \otimes v' \right\|_F^2 < \delta^2$$

$$\left\| (1 - \alpha_1 \alpha_2) u' \otimes v' + \beta_1 \alpha_2 \tilde{u}_\perp \otimes v' + \beta_2 \alpha_1 u' \otimes \tilde{v}_\perp + \beta_1 \beta_2 \tilde{u}_\perp \otimes \tilde{v}_\perp \right\|_F^2 < \delta^2$$

$$L_{\min}^4 (1 - \alpha_1 \alpha_2)^2 + \beta_1^2 \alpha_2^2 L_{\min}^2 + \beta_2^2 \alpha_1^2 L_{\min}^2 + \beta_1^2 \beta_2^2 < \delta^2 \tag{23}$$

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

$$L_{\min}^2 \leq \|v\|^2 = \alpha_2^2 \left\| v' \right\|^2 + \beta_2^2$$

$$L_{\min}^2 - \delta \leq \alpha_2^2 L_{\max}^2$$

$$\text{Hence,} \quad \alpha_2 \geq \frac{L_{\min}}{2 L_{\max}}$$

Now, using (23), we see that $\beta_1 < \sqrt{\delta}$. $\qquad \square$

Lemma A.5. *For $\lambda \geq 0$, 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*

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

then we have

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

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

Finally, $\|v - u\|_2 \leq \|v - \lambda u\|_2 + |\lambda - 1| \|u\|_2 \leq \varepsilon$ (since $\lambda \geq 0$). Hence, the lemma follows. $\quad \square$

A.1 Khatri-Rao product

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 of Lemma 3.12. Let $\tau = \tau_1 \tau_2 \sqrt{k_A + k_B}$. Suppose for contradiction M has K-rank$_\tau(M) < k = k_A + k_B - 1 \leq R$ (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 $x = \Pi_{\mathcal{S}}^{\perp} A_k / \|\Pi_{\mathcal{S}}^{\perp} A_k\|$ (it exists because $\text{K-rank}_\tau(M) < R$).
Pre-multiplying the expression in (A.1) 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 A.2), and there are only $k - k_A + 1 \leq k_B$ terms in the expression. Again, by Lemma A.2 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. $\qquad\square$

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 $\text{K-rank}_{10}(A) = n$, but for any τ, we have $\text{K-rank}_\tau(A \odot A) = 2n - 1$, 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).

A.2   Symmetric Decompositions

Proof of Corollary 3.15. Applying Theorem 3.14 with $\varepsilon' < \eta(2\rho\tau\sqrt{R})^{-1}$, to obtain a permutation matrix Π and scalar matrices Λ_j such that

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

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 (A.2) 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. $\quad\square$

B   Complete proofs for the Robust Uniqueness Theorem for 3-tensors.

B.1   Proof of Lemma 3.4

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

$$M := \sum_i \alpha_i A_i \otimes B_i = \sum_i \beta_i A_i' \otimes B_i' + Z, \tag{24}$$

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

$$\sigma_{R-k_C+2}(M) \leq \sigma_{t+1}(M) < \varepsilon + (R\rho_A'\rho_B')\varepsilon' \tag{25}$$

We will now show that if $x^T C$ has too many co-ordinates which are larger than ε'' then we will contradict (25). 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 (25). 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 well-known fact about singular values of products, which is proved by considering the variational characterization of singular values:

Fact B.1. *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*

$$\sigma_\ell(PQ) \geq \sigma_{\ell+m-i}(P)\sigma_i(Q) \tag{26}$$

Now, let us view M as PQ, where $P = A$, and $Q = \text{diag}(\alpha)B^T$. 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 (25) by setting $\ell = R - k_C + 2$ and $i = k_B$ since

$$\frac{\delta}{\tau_A\tau_B} = \frac{\varepsilon''}{\vartheta\tau_A\tau_B} > (R\rho_A'\rho_B'\varepsilon' + \varepsilon) \quad \text{by our choice of } \vartheta_{3.4}.$$

(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 second 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 $Q_J = \text{diag}(\alpha)_J B_J^T$. 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 $\sigma_{k_B}(\text{diag}(\alpha)_J) \geq \delta$, and applying Fact B.1 with $P = \text{diag}(\alpha)_J, Q = B_J^T$ 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. (24)) 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. (25).

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. (24) as

$$M' = \sum_{i \in S_1} \alpha_i (A_i \otimes B_i) + \sum_{j \in S_2 \setminus S_1} \alpha_j (A_j \otimes B_j) = \sum_{i=1}^{t} \beta_i (A'_i \otimes B'_i) + Err, \tag{27}$$

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 (27) 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 A.1 since:

$$\|(Err)y\|_2 < \vartheta_1(\delta + \varepsilon' + \varepsilon) = R\rho_A \rho_B \rho'_A \rho'_B(\delta + \varepsilon' + \varepsilon)$$

$$= R\rho_A \rho_B \rho'_A \rho'_B \left(\frac{\varepsilon''}{\vartheta} + \varepsilon' + \varepsilon\right)$$

$$< \frac{1}{\tau_A} \cdot \frac{\varepsilon''}{R\tau_B}$$

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.

B.2   Proof of Lemma 3.6

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 3.4, we have

$$nz_{\varepsilon_1}(zC) \leq n - k, \text{ where } \varepsilon_1 = \vartheta_{3.4}(\varepsilon + \delta).$$

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.

B.3   Wrapping up the Proof of Theorem 3.1

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 3.6 and 3.4 (setting $\varepsilon' = 0$), we obtain that C, C' satisfy the hypothesis of the Robust permutation lemma (Lemma 3.2) with C', C set to X, Y respectively, and the parameters

$$\text{``}\tau\text{''} := \vartheta_{3.6} \; ; \quad \text{``}\varepsilon\text{''} := \vartheta_{3.4}\varepsilon.$$

Hence, we apply Lemma 3.2 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 $\varepsilon_2 = \vartheta_{3.2}\vartheta_{3.4} \cdot \varepsilon$,

$$\left\| A' - A\Pi_A\Lambda_A \right\|_F < \varepsilon_2, \;\; \left\| B' - B\Pi_B\Lambda_B \right\|_F < \varepsilon_2 \;\; \text{and} \;\; \left\| C' - C\Pi_C\Lambda_C \right\|_F < \varepsilon_2 \qquad (28)$$

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 (28) for A we have

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

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

$$\left\| Z - Z' \right\| \leq \varepsilon_3 = \varepsilon + 2\varepsilon_2 R \rho'_C (\rho'_A + \rho'_B) \quad \text{where}$$

$$Z = \sum_{r\in[R]} \langle v, A_r \rangle \langle w, B_r \rangle C_r \quad \text{and} \quad Z' = \sum_{r\in[R]} \Lambda_A(r)\Lambda_B(r) \langle v, A_{\pi_A(r)} \rangle \langle w, B_{\pi_B(r)} \rangle C'_r$$

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 $\Pi_{\mathcal{V}}^{\perp} A_{r^*}$ and w as unit vector along $\Pi_{\mathcal{W}}^{\perp} B_{r^*}$. 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 A.2). Further, $|V| = k_A - 1$ implies that at most $R - k_A + 1 \leq k_C - 1$ terms of Z is non-zero.

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

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 A.2. 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 (28) 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

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

41

We will show each β_r is negligible. Since $R + 2 \le k_A + k_B$, let $S, W \subseteq [R] - \{r\}$ be disjoint sets of indices not containing r, such that $|S| = k_A - 1$ and $|W| \le k_B - 1$. 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$_{\tau_A}(A) \ge k_A$ and K-rank$_{\tau_B}(B) \ge k_B$, we have that $\left\| \Pi_{\mathcal{S}}^{\perp} A_r \right\| \ge 1/\tau_A$ (similarly for B_r). Hence, from Lemma A.2

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

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

Let us now set the ε for the above to hold (note that $\vartheta_{3.2}$ involves a τ term which depends on $\vartheta_{3.6}$)

$$\varepsilon := \frac{\varepsilon'}{6(R \tau_A \tau_B \tau_C) \rho_A' \rho_B' \rho_C' \cdot \vartheta_{3.4} \vartheta_{3.2}},$$

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

C   Multi-view models

Proof of Theorem 5.4. Set $\eta' = \frac{\eta \gamma}{16 \ell n}$. We know from Lemma F.1 that the ℓth moment tensor can be estimated to accuracy

$\varepsilon_1 = \left(\ell \cdot \vartheta_{3.14}^{(\ell)}(R/\eta') \cdot \vartheta_{3.14}(\tau/\gamma, c_{max} \sqrt{n}, c_{max} \sqrt{n}, n) \right)^{-1}$ in $\left\| \cdot \right\|_F$ 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 4 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

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

This is possible since the algorithm in Section 4 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 3.14 to ℓ^{th} moment tensor T to claim that these $\tilde{U}^{(j)}$ are close to $M^{(j)}$ upto permutations. When we apply Theorem 3.14, 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 3.14 with our choice of ε_1, and assuming that the permutation is identity without loss of generality, we get

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

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

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

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

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).

$$\forall j \in [\ell], r \in [R], \quad \tilde{v}_r^{(j)} = \text{sgn}\left(\Lambda^{(j)}(r)\right) \cdot \tilde{u}_r^{(j)} \tag{29}$$

This ensures that

$$\forall j \in [\ell-1], r \in [R] \ \left\|\tilde{v}_r^{(j)} - \left|\Lambda^{(j)}(r)\right| \mu_r^{(j)}\right\| < \eta' \leq \frac{\eta\gamma}{16n\ell} \quad \text{and} \tag{30}$$

$$\forall r \in [R] \ \left\|\tilde{v}_r^{(\ell)} - \left|\Lambda^{(\ell)}(r)\right| w_r \mu_r^{(\ell)}\right\| < \eta' \leq \frac{\eta\gamma}{16\ell n} \tag{31}$$

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 A.5 we get that the required estimates $\tilde{v}_r^{(j)}$ (i.e. $\tilde{\mu}_r^{(j)}$) satisfy:

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

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 (C) and (C) we get that

$$\forall r \in [R] \quad \left|\Lambda^{(\ell)}(r) - 1\right| \leq \frac{\eta\gamma}{8\sqrt{n}}$$

$$\text{Hence from (31),} \quad \left\|\tilde{v}_r^{(\ell)} - w_r \mu_r^{(\ell)}\right\| \leq \frac{\eta\gamma}{4\sqrt{n}}$$

$$\left\|\tilde{w}_r \tilde{\mu}_r^{(\ell)} - w_r \mu_r^{(\ell)}\right\| \leq \frac{\eta\gamma}{4\sqrt{n}}$$

$$w_r \left\|\frac{\tilde{w}_r}{w_r} \tilde{\mu}_r^{(\ell)} - \mu_r^{(\ell)}\right\| \leq \frac{\eta\gamma}{4\sqrt{n}}$$

Using the fact that $w_r \geq \gamma$ and using Lemma A.5, 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. $\qquad\square$

D   Mixtures of Gaussians

The following lemma (used in the proof of Theorem 5.7) allows us to recover the weights after obtaining estimates to $w_r^{1/\ell}\mu_r$ and $w_r^{1/(\ell-1)}\mu_r$ through decompositions for the $\ell-1$ and the ℓ moment tensors.

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

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

Then,

$$\left| \left(\frac{|\langle u, v \rangle|}{\|u\|} \right)^{\ell(\ell-1)} - w \right| < \delta' \tag{32}$$

Proof. From (D.1) and triangle inequality, we see that

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

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

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

Now, substituting the values for α_1, α_2, we see that

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

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

$\square$

E Properties of Tensors

E.1 A necessary condition for Uniqueness

Consider a 3-tensor T of rank R represented by $[A\ B\ C]$ where these three matrices are of size $n \times R$.

$$T = \sum_{r \in [R]} A_r \otimes B_r \otimes C_r.$$

We now show a necessary condition in terms of the n^2 dimensional vectors $A_r \otimes B_r$ from the decomposition.

Claim E.1 (A necessary condition for uniqueness). *Suppose for a subset $S \subset [R]$, there exist $\{\alpha_r\}$ with $\|\alpha\| = 1$.*

$$\sum_{r \in S} \alpha_r A_r \otimes B_r = 0$$

then there exists multiple rank-R decompositions for T

Proof. Consider any fixed non-zero vector u (it can be also chosen to be not close to any of the other vectors in S). This is because $\sum_{r \in S} A_r \otimes B_r \otimes u = \sum_{r \in S} \alpha_r (A_r \otimes B_r) \otimes u = 0$.
Hence, $T = \sum_{r \in S} A_r \otimes B_r \otimes (C_r + \alpha_r u) + \sum_{r' \in [R] \backslash S} A_{r'} \otimes B_{r'} \otimes C_{r'}$. $\qquad \square$

The above example showed that one necessary condition is that the $A \odot B$ should be full rank R (and well-conditioned). These examples are ruled out when the Kruskal ranks of A and B are such that $k_A + k_B \geq R$ by Lemma 3.12.

F   Sampling Error Estimates for Higher Moment Tensors

In this section, we show error estimates for ℓ-order tensors obtained by looking at the ℓth 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 F.1 (Error estimates for Multiview mixture model). *For every $\ell \in \mathbb{N}$, suppose we have a multi-view model, with parameters $\{w_r\}_{r \in [R]}$ and $\{M^{(j)}\}_{j \in [\ell]}$, 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

$$\left\| \mathbb{E}\left[x^{(1)} \otimes x^{(2)} \otimes \ldots x^{(\ell)} \right] - \frac{1}{N}\left(\sum_{t \in [N]} x(t)^{(1)} \otimes x(t)^{(2)} \otimes x(t)^{(\ell)} \right) \right\|_\infty < \varepsilon \qquad (33)$$

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 (33) to not occur $\exp\left(-\frac{(\varepsilon n^{-\ell/2})^2 N^2}{2 N c_{max}^\ell} \right) = \exp\left(-\varepsilon^2 N / \left(2(c_{max} n)^\ell \right) \right)$. 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. $\qquad \square$

Lemma F.2 (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 $N = \Omega(poly(\frac{1}{\varepsilon})), \sigma^2, n, R)$ such that if $x^{(1)}, x^{(2)}, \ldots, x^{(N)} \in R^n$ were the N samples, then

$$\forall \{i_1, i_2, \ldots, i_\ell\} \in [n]^\ell, \left| \mathbb{E}\left[\prod_{j\in[\ell]} x_{i_j} \right] - \frac{1}{N}\left(\sum_{t\in[N]} \prod_{j\in[\ell]} x_{i_j}^{(t)} \right) \right| < \varepsilon. \tag{34}$$

In other words,

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

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

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

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