Polynomial-time tensor decompositions with sum-of-squares

Tengyu Ma* Jonathan Shi† David Steurer‡

October 21, 2016

Abstract

We give new algorithms based on the sum-of-squares method for tensor decomposition. Our results improve the best known running times from quasi-polynomial to polynomial for several problems, including decomposing random overcomplete 3-tensors and learning overcomplete dictionaries with constant relative sparsity. We also give the first robust analysis for decomposing overcomplete 4-tensors in the smoothed analysis model.

A key ingredient of our analysis is to establish small spectral gaps in moment matrices derived from solutions to sum-of-squares relaxations. To enable this analysis we augment sum-of-squares relaxations with spectral analogs of maximum entropy constraints.

*Princeton University. tengyu@cs.princeton.edu. Supported by Simons Award in Theoretical Computer Science, IBM PhD Fellowship, Dodds Fellowship and Siebel Scholarship.

†Cornell University, jshi@cs.cornell.edu. Supported by David Steurer's NSF CAREER award.

‡Cornell University, dsteurer@cs.cornell.edu. Supported by a Microsoft Research Fellowship, a Alfred P. Sloan Fellowship, an NSF CAREER award, and the Simons Collaboration for Algorithms and Geometry.

Contents

1 Introduction

Tensors are arrays of (real) numbers with multiple indices—generalizing matrices (two indices) and vectors (one index) in a natural way. They arise in many different contexts, e.g., moments of multivariate distributions, higher-order derivatives of multivariable functions, and coefficients of multivariate polynomials. An important ongoing research effort aims to extend algorithmic techniques for vectors and matrices to more general tensors. A key challenge is that many tractable matrix computations (like rank and spectral norm) become NP-hard in the tensor setting (even for just three indices) [Hås90, HL13]. However, recent work gives evidence that it is possible to avoid this computational intractability and develop provably efficient algorithms, especially for low-rank tensor decompositions, by making suitable assumptions about the input and allowing for approximations [AGJ15, AGJ14, GM15, HSS15, HSSS16]. These algorithms lead to the best known provable guarantees for a wide range of unsupervised learning problems [AGH+14, BCMV14, GVX14, AGHK14], including learning mixtures of Gaussians [GHK15], Latent Dirichlet topic modeling [AFH+15], and dictionary learning [BKS15]. Low-rank tensor decompositions are useful for these learning problems because they are often unique up to permuting the factors—in contrast, low-rank matrix factorizations are unique only up to unitary transformation. In fact, as far as we are aware, in all natural situations where finding low-rank tensor decompositions is tractable, the decompositions are also unique.

We consider the following (symmetric) version of the tensor decomposition problem: Let $a_1, \ldots, a_n \in \mathbb{R}^d$ be d-dimensional unit vectors. We are given (approximate) access to the first k moments $\mathcal{M}_1, \ldots, \mathcal{M}_k$ of the uniform distribution over $a_1, \ldots, a_n$, that is,

$$\mathcal{M}_t = \frac{1}{n} \sum_{i=1}^{n} a_i^{\otimes t} \quad \text{for } t \in \{1, \ldots, k\}\,. \tag{1.1}$$

The goal is to approximately recover the vectors $a_1, \ldots, a_n$. What conditions on the vectors $a_1, \ldots, a_n$ and the number of moments k allow us to efficiently and robustly solve this problem?

A classical algorithm based on (simultaneous) matrix diagonalization [Har70, LRA93, attributed to Jennrich] shows that whenever the vectors $a_1, \ldots, a_n$ are linearly independent, $k = 3$ moments suffice to recover the vectors in polynomial time. (This algorithm is also robust against polynomially small errors in the input moment tensors [AGH+15, GVX14, BCMV14].) Therefore an important remaining algorithmic challenge for tensor decomposition is the *overcomplete* case, when the number of vectors (significantly) exceeds their dimension. Several recent works studied this case with different assumptions on the vectors and the number of moments. In this work, we give a unified algorithmic framework for overcomplete tensor decomposition that achieves—and in many cases surpasses—the previous best guarantees for polynomial-time algorithms.

In particular, some decompositions that previously required quasi-polynomial time to find are reduced to polynomial time in our framework, including the case of general tensors with order logarithmically large in its overcompleteness n/d [BKS15] and random order-3 tensors with rank $n \leqslant d^{3/2}/\log^{O(1)}(d)$ [GM15]. Iterative methods may also achieve fast local convergence guarantees for incoherent order-3 tensors with rank $o(d^{3/2})$, which become global convergence guarantees under no more than constant overcompleteness [AGH+14]. In the smoothed analysis model, where

each vector of the desired decomposition is assumed to have been randomly perturbed by an inverse polynomial amount, polynomial-time decomposition was achieved for order-5 tensors of rank up to $d^2/2$ [BCMV14]. Our framework extends this result to order-4 tensors, for which the corresponding analysis was previously unknown for any superconstant overcompleteness.

The starting point of our work is a new analysis of the aforementioned matrix diagonalization algorithm that works for the case when $a_1, \ldots, a_n$ are linearly independent. A key ingredient of our analysis is a powerful and by now standard concentration bound for Gaussian matrix series [Oli10, Tro12]. An important feature of our analysis is that it is captured by the sum-of-squares (SoS) proof system in a robust way. This fact allows us to use Jennrich's algorithm as a rounding procedure for sum-of-squares relaxations of tensor decomposition, which is the key idea behind improving previous quasi-polynomial time algorithms based on these relaxations [BKS15, GM15].

The main advantage that sum-of-squares relaxations afford for tensor decomposition is that they allow us to efficiently hallucinate faithful *higher-degree moments* for a distribution given only its lower-degree moments. We can now run classical tensor decomposition algorithms like Jennrich's on these hallucinated higher-degree moments (akin to *rounding*). The goal is to show that those algorithms work as well as they would on the true higher moments. What is challenging about it is that the analysis of Jennrich's algorithm relies on small spectral gaps that are difficult to reason about in the sum-of-squares setting. (Previous sum-of-squares based methods for tensor decomposition also followed this outline but used simpler, more robust rounding algorithms which required quasi-polynomial time.)

To this end, we view solutions to sum-of-squares relaxations as *pseudo-distributions*, which generalize classical probability distributions in a way that takes computational efficiency into account.[1] More concretely, pseudo-distributions are indistinguishable from actual distributions with respect to tests captured by a restricted system of proofs, called *sum-of-squares proofs*.

An interesting feature of how we use pseudo-distributions is that our relaxations search for pseudo-distributions of large *entropy* (via an appropriate surrogate). This objective is surprising, because when we consider convex relaxations of NP-hard search problems, the intended solutions typically correspond to atomic distributions which have entropy 0. Here, high entropy in the pseudo-distribution allows us to ensure that rounding results in a useful solution. This appears to be related to the way in which many randomized rounding procedures use maximum-entropy distributions [Gha14], but differs in that the aforementioned rounding procedures focus on the entropy of the rounding process rather than the entropy (surrogate) of the solution to the convex relaxation. A measure of "entropy" has also been directly ascribed to pseudo-distributions previously [LRS15], and the principle of maximum entropy has been applied to pseudo-distributions as well [BHK+16], but these have previously occurred separately, and our application is the first to encode a surrogate notion of entropy directly into the sum-of-squares proof system.

Our work also takes inspiration from a recent work that uses sum-of-squares techniques to design fast spectral algorithms for a range of problems including tensor decomposition [HSSS16]. Their algorithm also proceeds by constructing surrogates for higher moments and applying a

[1] In particular, the set of constant-degree moments of n-variate pseudo-distributions admits an $n^{O(1)}$-time separation oracle based on computing eigenvectors.

classical tensor decomposition algorithm on these surrogates. The difference is that the surrogates in [HSSS16] are explicitly constructed as low-degree polynomial of the input tensor, whereas our surrogates are computed by sum-of-squares relaxations. The explicit surrogates of [HSSS16] allow for a direct (but involved) analysis through concentration bounds for matrix polynomials. In our case, a direct analysis is not possible because we have very little control over the surrogates computed by sum-of-squares relaxations. Therefore, the challenge for us is to understand to what extent classical tensor decomposition algorithms are compatible with the sum-of-squares proof system. Our analysis ends up being less technically involved compared to [HSSS16] (using the language of pseudo-distributions and sum-of-squares proofs).

1.1 Results for tensor decomposition

Let $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}^d$ be a set of unit vectors. We study the task of approximately recovering this set of vectors given (noisy) access to its first k moments (1.1). We organize this overview of our results based on different kinds of assumptions imposed on the set $\{a_1, \ldots, a_n\}$ and the order of tensor/moments that we have access to. All of our algorithms are randomized and may fail with some small probability over their internal randomness, say probability at most 0.01. (Standard arguments allow us to amplify this probability at the cost of a small increase in running time.)

Orthogonal vectors. This scenario often captures the case of general linearly independent vectors because knowledge of the second moments of $a_1, \ldots, a_n$ allows us to orthonormalize the vectors (this process is sometimes called "whitening"). Many efficient algorithms are known in this case. Our contribution here is in improving the error tolerance. For a symmetric 3-tensor $E \in (\mathbb{R}^d)^{\otimes 3}$, we use $\|E\|_{\{1\},\{2,3\}}$ to denote the spectral norm of E as a d-by-d^2 matrix (using the first mode of E to index rows and the last two modes of E to index the columns). This norm is at most $\sqrt{d}$ times the injective norm $\|E\|_{\{1\},\{2\},\{3\}}$ (the maximum of $\langle E, x \otimes y \otimes z\rangle$ over all unit vectors $x, y, z \in \mathbb{R}^d$). The previous best error tolerance for this problem required the error tensor $E = T - \sum_{i=1}^n a_i^{\otimes 3}$ to have injective norm $\|E\|_{\{1\},\{2\},\{3\}} \ll 1/d$. Our algorithm requires only $\|E\|_{\{1\},\{2,3\}} \ll 1$, which is satisfied in particular when $\|E\|_{\{1\},\{2\},\{3\}} \ll 1/\sqrt{d}$.

Theorem 1.1. *There exists a polynomial-time algorithm that given a symmetric 3-tensor $T \in (\mathbb{R}^d)^{\otimes 3}$ outputs a set of vectors $\{a'_1, \ldots, a'_{n'}\} \subseteq \mathbb{R}^d$ such that for every orthonormal set $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}^d$, the Hausdorff distance[2] between the two sets is at most*

$$\mathrm{dist}_H\left(\{a_1, \ldots, a_n\}, \{a'_1, \ldots, a'_{n'}\}\right)^2 \leqslant O(1) \cdot \left\| T - \sum_{i=1}^n a_i^{\otimes 3} \right\|_{\{1\},\{2,3\}} . \tag{1.2}$$

Under the additional assumption $\|T - \sum_{i=1}^n a_i^{\otimes 3}\|_{\{1\},\{2,3\}} \leqslant 1/\log d$, the running time of the algorithm can be improved to $O(d^{1+\omega}) \leqslant d^{3.33}$ using fast matrix multiplication, where ω is the number such that two $n \times n$ matrices can be multiplied together in time n^ω (See Theorem 10.2).

It is also possible to replace the spectral norm $\|\cdot\|_{\{1\},\{2,3\}}$ in the above theorem statement by constant-degree sum-of-squares relaxations of the injective norm of 3-tensors. (See Remark 5.3 for

[2]The Hausdorff distance $\mathrm{dist}_H(X, Y)$ between two finite sets X and Y measures the length of the largest gap between the two sets. Formally, $\mathrm{dist}_H(X, Y)$ is the maximum of $\max_{x\in X} \min_{y\in Y} \|x - y\|$ and $\max_{y\in Y} \min_{x\in X} \|x - y\|$.

details.) If the error E has Gaussian distribution $\mathcal{N}(0, \sigma^2 \cdot \mathrm{Id}_d^{\otimes 3})$, then this norm is w.h.p. bounded by $\sigma \cdot d^{3/4}(\log d)^{O(1)}$ [HSS15], whereas the norm $\|\cdot\|_{\{1\},\{2,3\}}$ has magnitude $\Omega(\sigma \cdot d)$. We prove Theorem 1.1 in Section 5.2.

Random vectors. We consider the case that $a_1, \ldots, a_n$ are chosen independently at random from the unit sphere of $\mathbb{R}^d$. For $n \leqslant d$, this case is roughly equivalent to the case of orthonormal vectors. Thus, we are interested in the "overcomplete" case $n \gg d$, when the rank is larger than the dimension. Previous work found the decomposition in quasi-polynomial time when $n \leqslant d^{3/2}/\log^{O(1)} d$ [GM15], or in time subquadratic in the input size when $n \leqslant d^{4/3}/\log^{O(1)} d$ [HSSS16]. Our polynomial-time algorithm therefore is an improvement when n is between $d^{4/3}$ and $d^{3/2}$ (up to logarithmic factors).

Theorem 1.2. *There exists a polynomial-time algorithm A such that with probability $1 - d^{-\omega(1)}$ over the choice of random unit vectors $a_1, \ldots, a_n \in \mathbb{R}^d$, every symmetric 3-tensor $T \in (\mathbb{R}^d)^{\otimes 3}$ satisfies*

$$\mathrm{dist}_H\left(A(T), \{a_1, \ldots, a_n\}\right)^2 \leqslant O\left(\left(\frac{n}{d^{1.5}}\right)^{\Omega(1)} + \left\|T - \sum_{i=1}^n a_i^{\otimes 3}\right\|_{\{1\},\{2,3\}}\right). \tag{1.3}$$

Again it is possible to replace the spectral norm $\|\cdot\|_{\{1\},\{2,3\}}$ in the above theorem statement by constant-degree sum-of-squares relaxations of the injective norm of 3-tensors, which as mentioned before give better bounds for Gaussian error tensors. We prove Theorem 1.2 in Section 7.

Smoothed vectors. Next, we consider a more general setup where the vectors $a_1, \ldots, a_n \in \mathbb{R}^d$ are smoothed, i.e., randomly perturbed. This scenario is significantly more general than random vectors. Again we are interested in the overcomplete case $n \gg d$. The previous best work [BCMV14] showed that the fifth moment of smoothed vectors $a_1, \ldots, a_n$ with $n \leqslant d^2/2$ is enough to approximately recover the vectors even in the presence of a polynomial amount of error. For fourth moments of smoothed vectors, no such result was known even for lower overcompleteness, say $n = d^{1.01}$.

We give an interpretation of the 4-tensor decomposition algorithm FOOBI[3] [LCC07] as a special case of a sum-of-squares based decomposition algorithm. We show that the sum-of-squares based algorithm works in the smoothed setting even in the presence of a polynomial amount of error. We define a condition number $\kappa(\cdot)$ for sets of vectors $a_1, \ldots, a_n \in \mathbb{R}^d$ (a polynomial in the condition number of two matrices, one with columns $\{a_i^{\otimes 2} \mid i \in [n]\}$ and one with columns $\{a_i \otimes (a_i \otimes a_j - a_j \otimes a_i) \otimes a_j \mid i \neq j \in [n]\}$). First, we show that the algorithm can tolerate error $\ll 1/\kappa$ which could be independent of the dimension. Concretely, our algorithm will output a set of vectors $\hat{a}_1, \ldots, \hat{a}_n$ which will be close to $\{a_1, \ldots, a_n\}$ up to permutations and sign flip with a relative error that scales linearly in the relative error of the input and the condition number κ. Second, we show that for smoothed vectors this condition number is at least inverse polynomial with probability exponentially close to 1.

[3]The FOOBI algorithm is known to work for overcomplete 4-tensors when there is no error in the input. Researchers [BCMV14] asked if this algorithm tolerates a polynomial amount of error. Our work answers this question affirmatively for a variant of FOOBI (based on sum-of-squares).

Theorem 1.3. *There exists a polynomial-time algorithm such that for every symmetric 4-tensor $T \in (\mathbb{R}^d)^{\otimes 4}$ and every set $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}^d$ of vectors not necessarily unit length, there exists a permutation $\pi : [n] \to [n]$ so that the output $\{a'_1, \ldots, a'_n\}$ of the algorithm on input T satisfies*

$$\max_{i \in [n]} \frac{\left\| a_i - a'_{\pi(i)} \right\|}{\|a_i\|} \leqslant O(1) \cdot \frac{\left\| T - \sum_{i=1}^n a_i^{\otimes 4} \right\|_{\{1,2\},\{3,4\}}}{\sigma_n \left(\sum_{i=1}^n (a_i^{\otimes 2})(a_i^{\otimes 2})^\mathsf{T} \right)} \cdot \kappa(a_1, \ldots, a_n)\,, \tag{1.4}$$

where $\sigma_n(A)$ refers to the nth singular value of the matrix A, here the smallest non-zero singular value.

We say that a distribution over vectors $a_1, \ldots, a_n \in \mathbb{R}^d$ is γ-smoothed if $a_i = a_i^0 + \gamma \cdot g_i$, where $a_1^0, \ldots, a_n^0$ are fixed vectors and $g_1, \ldots, g_n$ are independent Gaussian vectors from $\mathcal{N}(0, \frac{1}{d} \mathrm{Id}_d)$.

Theorem 1.4. *Let $\varepsilon > 0$ and $n, d \in \mathbb{N}$ with $n \leqslant d^2/10$. Then, for any γ-smoothed distribution over vectors $a_1, \ldots, a_n$ in $\mathbb{R}^d$,*

$$\mathbb{P}\left\{ \kappa(a_1, \ldots, a_n) \leqslant \mathrm{poly}(d, \gamma) \right\} \geqslant 1 - \exp(-d^{\Omega(1)})\,.$$

The above theorems together imply a polynomial-time algorithm for approximately decomposing overcomplete smoothed 4-tensors even if the input error is polynomially large. The error probability of the algorithm is exponentially small over the choice of the smoothing. It is an interesting open problem to extend this result to overcomplete smoothed 3-tensors, even for lower overcompleteness $n = d^{1.01}$. Theorem 1.3 and Theorem 1.4 are proved in Section 8.

Separated unit vectors. In the scenario, when inner products among the vectors $a_1, \ldots, a_n \in \mathbb{R}^d$ are bounded by $\rho < 1$ in absolute value, the previous best decomposition algorithm shows that moments of order $(\log n)/\log \rho$ suffice [SW15]. Our algorithm requires moments of higher order (by a factor logarithmic in the desired accuracy) but in return tolerates up to constant spectral error. This increased error tolerance also allows us to apply this result for dictionary learning with up to constant sparsity (see Section 1.2).

Theorem 1.5. *There exists an algorithm A with polynomial running time (in the size of its input) such that for all $\eta, \rho \in (0, 1)$ and $\sigma \geqslant 1$, for every set of unit vectors $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}^d$ with $\|\sum_{i=1}^n a_i a_i^\mathsf{T}\| \leqslant \sigma$ and $\max_{i \neq j} |\langle a_i, a_j \rangle| \leqslant \rho$, when the algorithm is given a symmetric k-tensor $T \in (\mathbb{R}^d)^{\otimes k}$ with $k \geqslant O\left(\frac{1+\log \sigma}{\log \rho}\right) \cdot \log(1/\eta)$, then its output $A(T)$ is a set of vectors $\{a'_1, \ldots, a'_{n'}\} \subseteq \mathbb{R}^d$ such that*

$$\mathrm{dist}_H\left(\{a_1'^{\otimes 2}, \ldots, a_n'^{\otimes 2}\}, \{a_1^{\otimes 2}, \ldots, a_n^{\otimes 2}\} \right)^2 \leqslant O\left(\eta + \left\| T - \sum_{i=1}^n a_i^{\otimes k} \right\|_{\{1,\ldots,\lfloor k/2 \rfloor\},\{\lfloor k/2 \rfloor + 1,\ldots,k\}} \right)\,. \tag{1.5}$$

We also show that a simple spectral algorithm with running time close to d^k (the size of the input) achieves similar guarantees (see Remark 10.3). However, the error tolerance of this algorithm is in terms of an *unbalanced* spectral norm: $\|T - \sum_{i=1}^n a_i^{\otimes k}\|_{\{1,\ldots,k/3\},\{k/3+1,\ldots,k\}}$ (the spectral norm of the tensor viewed as a $d^{k/3}$-by-$d^{2k/3}$ matrix). This norm is always larger than the balanced spectral norm in the theorem statement. In particular, for dictionary learning applications, this norm is larger than 1, which renders the guarantee of the simpler spectral algorithm vacuous in this case. We prove Theorem 1.5 in Section 5.3.

General unit vectors. In this scenario, the number of moments that our algorithm requires is constant as long as $\sum_i a_i a_i^\mathsf{T}$ has constant spectral norm and the desired accuracy is constant.

Theorem 1.6. *There exists an algorithm A (see Algorithm 4) with polynomial running time (in the size of its input) such that for all $\varepsilon \in (0,1), \sigma \geqslant 1$, for every set of unit vectors $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}^d$ with $\|\sum_{i=1}^n a_i a_i^\mathsf{T}\| \leqslant \sigma$ and every symmetric $2k$-tensor $T \in (\mathbb{R}^d)^{\otimes 2k}$ with $k \geqslant (1/\varepsilon)^{O(1)} \cdot \log(\sigma)$ and $\left\|T - \sum_i a_i^{\otimes 2k}\right\|_{\{1,\ldots,k\},\{k+1,\ldots,2k\}} \leqslant 1/3$, we have*

$$\operatorname{dist}_H \left(A(T), \{a_1^{\otimes 2}, \ldots, a_n^{\otimes 2}\}\right)^2 \leqslant O(\varepsilon) .$$

The previous best algorithm for this problem required tensors of order $(\log \sigma)/\varepsilon$ and had running time $d^{O((\log \sigma)/\varepsilon^{O(1)} + \log n)}$ [BKS15, Theorem 4.3]. We require the same order of the tensor and the runtime is improved to be polynomial in the size of the inputs (that is, $d^{\operatorname{poly}((\log \sigma)/\varepsilon)}$).

We also remark that a bit surprisingly we can handle 1/3 error in spectral norm, and this is possible partly due to the choice of working with high order tensors. As a sanity check, we note that information-theoretically the components are identifiable: under the assumptions, the only vectors u that satisfy $\langle T, u^{\otimes 2k}\rangle \geqslant 1/3$ are those vectors close to one of the a_i's. We also note that the rounding algorithm of the sum-of-squares relaxation of this simple inefficient test requires a bit new idea beyond what we used previously. Here the difficulty is to make the runtime $d^{\operatorname{poly}((\log \sigma)/\varepsilon)}$ instead of $d^{\operatorname{poly}(\sigma/\varepsilon)}$. See Section 9 for details.

Spectral algorithms without sum-of-squares. Finally, using a similar rounding technique directly on a orthogonal tensor (without using sum-of-squares and the pseudo-moment), we also obtain a fast and robust algorithm for orthogonal tensor decomposition. See Section 10 for details.

1.2 Applications of tensor decomposition

Tensor decomposition has a wide range of applications. We focus here on learning sparse dictionaries, which is an example of the more general phenomenon of using tensor decomposition to learn latent variable models. Here, we obtain the first polynomial-time algorithms that work in the overcomplete regime up to constant sparsity.

Dictionary learning is an important problem in multiple areas, ranging from computational neuroscience [OF97, OF96a, OF96b], machine learning [EP07, MRBL07], to computer vision and image processing [EA06, MLB+08, YWHM08]. The general goal is to find a good basis for given data. More formally, in the dictionary learning problem, also known as sparse coding, we are given samples of a random vector $y \in \mathbb{R}^n$, of the form $y = Ax$ where A is some unknown matrix in $\mathbb{R}^{n \times m}$, called *dictionary*, and x is sampled from an unknown distribution over sparse vectors. The goal is to approximately recover the dictionary A.

We consider the same class of distributions over sparse vectors $\{x\}$ as [BKS15], which as discussed in [BKS15] admits a wide-range of non-product distributions over sparse vectors. (The case of product distributions reduces to the significantly easier problem of independent component analysis.) We say that $\{x\}$ is (k, τ)-*nice* if $\mathbb{E}\, x_i^k = 1$ for every $i \in [m]$, $\mathbb{E}\, x_i^{k/2} x_j^{k/2} \leqslant \tau$ for all $i \neq j \in [m]$,

and $\mathbb{E}\, x^\alpha = 0$ for every non-square degree-k monomial x^α. Here, τ is a measure of the relative sparsity of the vectors $\{x\}$.

We give an algorithm that for nice distributions solves the dictionary learning problem in polynomial time when the desired accuracy is constant, the overcompleteness of the dictionary is constant (measured by the spectral norm $\|A\|$), and the sparsity parameter τ is a sufficiently small constant (depending only on the desired accuracy and $\|A\|$). The previous best algorithm [BKS15] requires quasi-polynomial time in this setup (but works in polynomial-time for polynomial sparsity $\tau \leqslant n^{-\Omega(1)}$).

Theorem 1.7. *There exists an algorithm $\mathcal{R}$ parameterized by $\sigma \geqslant 1, \eta \in (0,1)$, such that for every dictionary $A \in \mathbb{R}^{n\times m}$ with $\|A\| \leqslant \sigma$ and every (k,τ)-nice distribution $\{x\}$ over $\mathbb{R}^m$ with $k \geqslant k(\eta,\sigma) = O((\log\sigma)/\eta)$ and $\tau \leqslant \tau(k) = k^{-O(k)}$, the algorithm given $n^{O(k)}$ samples from $\{y = Ax\}$ outputs in time $n^{O(k)}$ vectors $a'_1, \dots, a'_m$ that are $O(\eta)^{1/2}$-close to the columns of A.*

Since previous work [BKS15] provides a black box reduction from dictionary learning to tensor decomposition, the theorem above follows from Theorem 1.6. Our Theorem 1.5 implies a dictionary learning algorithm with better parameters for the case that the columns of A are separated.

1.3 Polynomial optimization with few global optima

Underlying our algorithms for the tensor decomposition is an algorithm for solving general systems of polynomial constraints with the property that the total number of different solutions is small and that there exists a short certificate for that fact in form of a sum-of-squares proof.

Let $\mathcal{A}$ be a system of polynomial constraints over real variables $x = (x_1, \dots, x_d)$ and let $P\colon \mathbb{R}^d \to \mathbb{R}^{d^\ell}$ be a polynomial map of degree at most ℓ—for example, $P(x) = x^{\otimes\ell}$. We say that solutions $a_1, \dots, a_n \in \mathbb{R}^d$ to $\mathcal{A}$ are *unique* under the map P if the vectors $P(a_1), \dots, P(a_n)$ are orthonormal up to error 0.01 (in spectral norm) and every solution a to $\mathcal{A}$ satisfies $P(a) \approx P(a_i)$ for some $i \in [n]$. We encode this property algebraically by requiring that the constraints in $\mathcal{A}$ imply the constraint $\sum_{i=1}^n \langle P(a_i), P(x)\rangle^4 \geqslant 0.99 \cdot \|P(x)\|^4$. We say that the solutions $a_1, \dots, a_n$ are *ℓ-certifiably unique* if in addition this implication has a degree-ℓ sum-of-squares proof.

The following theorem shows that if polynomial constraints have certifiably unique solutions (under a given map P), then we can find them efficiently (under the map P).

Theorem 1.8 (Informal statement of Theorem 5.2)**.** *Given a system of polynomial constraints $\mathcal{A}$ and a polynomial map P such that there exists ℓ-certifiably unique solutions $a_1, \dots, a_n$ for $\mathcal{A}$, we can find in time $d^{O(\ell)}$ vectors 0.1-close to $P(a_1), \dots, P(a_n)$ in Hausdorff distance.*

2 Techniques

Here is the basic idea behind using sum-of-squares for tensor decomposition: Let $a_1, \dots, a_n \in \mathbb{R}^d$ be unit vectors and suppose we have access to their first three moments $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ as in (1.1). Since the task of recovering $a_1, \dots, a_n$ is easier the more moments we know, we would make a lot of progress if we could compute higher moments of $a_1, \dots, a_n$, say the fourth moment $\mathcal{M}_4$. A

natural approach toward that goal is to compute a probability distribution D over the sphere of $\mathbb{R}^d$ such that D matches the moments of $a_1, \ldots, a_k$ that we know, i.e., $\mathbb{E}_{D(u)}\, u = \mathcal{M}_1$, $\mathbb{E}_{D(u)}\, u^{\otimes 2} = \mathcal{M}_2$, $\mathbb{E}_{D(u)}\, u^{\otimes 3} = \mathcal{M}_3$, and then use the fourth moment $\mathbb{E}_D\, u^{\otimes 4}$ as an estimate for $\mathcal{M}_4$.

There are two issues with this approach: (1) computing such a distribution D is intractable and (2) even if we could compute such a distribution it is not clear if its fourth moment will be close to the fourth moments $\mathcal{M}_4$ we are interested in.

We address issue (1) by relaxing D to be a pseudo-distribution (solution to sum-of-squares relaxations). Then, we can match the given moments efficiently.

Issue (2) is related to the uniqueness of the tensor decomposition, which relies on properties of the vectors $a_1, \ldots, a_n$. Here, the general strategy is to first prove that this uniqueness holds for actual distributions and then transfer the uniqueness proof to the sum-of-squares proof system, which would imply that uniqueness also holds for pseudo-distributions.

In subsection 2.1 below, we demonstrate our key rounding idea on the (nearly) orthogonal tensor decomposition problem. Then in subsection 2.2 we discuss the high level insight for the robust 4th-order tensor decomposition algorithm and in subsection 2.3 the techniques for random 3rd-order tensor decomposition.

2.1 Rounding pseudo-distributions by matrix diagonalization

Our main departure from previous tensor decomposition algorithms based on sum-of-squares [BKS15, GM15] lies in *rounding*: the procedure to extract an actual solution from a pseudo-distribution over solutions. The previous algorithms rounded a pseudo-distribution D by directly using the first moments (or the mean) $\mathbb{E}_{D(u)}\, u$, which requires D to concentrate strongly around the desired solution. Our approach here instead uses Jennrich's (simultaneous) matrix diagonalization [Har70, LRA93], to extract the desired solution as a *singular vector* of a matrix of the form $\mathbb{E}_{D(u)}\langle g, u\rangle u u^\intercal$, for a random vector g.[4] This permits us to impose much weaker conditions on D.

For the rest of this subsection, we assume that we have an actual distribution D that is supported on vectors close to some orthonormal basis $a_1, \ldots, a_d$ of $\mathbb{R}^d$, and we will design a rounding algorithm that extracts the vectors a_i from the low-degree moments of D. This is a much simpler task than rounding from a pseudo-distribution, though it captures most of the essential difficulties. Since pseudo-distributions behave similarly to actual distributions on the low-degree moments, the techniques involved in rounding from actual distributions will turn out to be easily generalizable to the case of pseudo-distributions.

Let D be a distribution over the unit sphere in $\mathbb{R}^d$. Suppose that this distribution is supported on vectors close to some orthonormal basis $a_1, \ldots, a_d$ of $\mathbb{R}^d$, in the sense that the distribution satisfies the constraint

$$\left\{ \sum_{i=1}^{d} \langle a_i, u\rangle^3 \geqslant 1 - \varepsilon \right\}_{D(u)} . \tag{2.1}$$

[4] In previous treatments of simultaneous diagonalization, multiple matrices would be used for noise tolerance—increasing the confidence in the solution when more than one matrix agrees on a particular singular vector. This is unnecessary in our setting, since as we'll see, the SoS framework itself suffices to certify the correctness of a solution.

(This constraint implies $\{\max_{i\in[d]}\langle a_i, u\rangle \geqslant 1-\varepsilon\}_{D(u)}$ because $\sum_{i=1}^{d}\langle a_i, u\rangle^3 \leqslant \max_{i\in[d]}\langle a_i, u\rangle$ by orthonormality.) The analysis of [BKS15] shows that reweighing the distribution D by a function of the form $u \mapsto \langle g, u\rangle^{2k}$ for $g \sim \mathcal{N}(0, \mathrm{Id}_d)$ and some $k \leqslant O(\log d)$ creates, with significant probability, a distribution D' such that for one of the basis vectors a_i, almost all of the probability mass of D' is on vectors close to a_i, in the sense that

$$\max_{i\in[d]} \mathop{\mathbb{E}}_{D'(u)} \langle a_i, u\rangle \geqslant 1 - O(\varepsilon)\,, \text{where } D'(u) \propto \langle g, u\rangle^{2k} D(u)\,.$$

In this case, we can extract a vector close to one of the vectors a_i by computing the mean $\mathbb{E}_{D'(u)}\, u$ of the reweighted distribution. This rounding procedure takes quasi-polynomial time because it requires access to logarithmic-degree moments of the original pseudo-distribution D.

To avoid this quasi-polynomial running time, our strategy is to instead modify the original distribution D in order to create a small bias in one of the directions a_i such that a modified moment matrix of D has a one-dimensional eigenspace close to a_i. (This kind of modification is much less drastic than the kind of modification in previous works. Indeed, reweighing a distribution such that it concentrates around a particular vector seems to require logarithmic degree.)

Concretely, we will study the spectrum of matrices of the following form, for $g \sim \mathcal{N}(0, \mathrm{Id}_d)$:

$$M_g = \mathop{\mathbb{E}}_{D(u)} \langle g, u\rangle \cdot uu^{\mathsf{T}}.$$

Our goal is to show that with good probability, M_g has a one-dimensional eigenspace close to one of the vectors a_i.

However, this is not actually true for a naïve distribution: although we have encoded the basis vectors a_i into the distribution D by means of constraint (2.1), we cannot yet conclude that the eigenspaces of M_g have anything to do with them. We can understand this as the error allowed in (2.1) being highly under-constrained. For example, the distribution could be a uniform mixture of vectors of the form $a_i + \varepsilon w$ for some fixed vector w, which causes w to become by far the most significant contribution to the spectrum of M_g. More generally, an arbitrary spectrally small error could still completely displace all of the eigenspaces of M_g.

An interpretation of this situation is that we have permitted D itself to contain a large amount of information that we do not actually possess. Constraint (2.1) is consistent with a wide range of possible solutions, yet in the pathological example above, the distribution does not at all reflect this uncertainty, instead settling arbitrarily on some particular biased solution: it is this bias that disrupts the usefulness of the rounding procedure.

A similar situation has previously arisen in strategies for rounding convex relaxations—specifically, when the variables of the relaxations were interpreted as the marginals of some probability distribution over solutions, then actual solutions were constructed by sampling from that distribution. In that context, a workaround was to sample those solutions from the maximum-entropy distributions consistent with those marginals [Gha14], to ensure that the distribution faithfully reflected the ignorance inherent in the relaxation solution rather than incorporating arbitrary information. Our situation differs in that it is the solution to the convex relaxation itself which is misbehaving, rather than some aspect of the rounding process, but the same approach carries over here as well.

Therefore, suppose that D satisfies the maximum-entropy constraint $\|\mathbb{E}_{D(u)} uu^\mathsf{T}\| \leqslant 1/n$. This essentially enforces D to be a uniform distribution over vectors close to $a_1, \ldots, a_n$. For the sake of demonstration, we assume that D is a uniform distribution over $a_1, \ldots, a_n$. Moreover, since our algorithm is invariant under linear transformations, we may assume that the components $a_1, \ldots, a_n$ are the standard basis vectors $e_1, \ldots, e_n \in \mathbb{R}^d$. We first decompose M_g along the coordinate g_1,

$$M_g = g_1 \cdot M_{e_1} + M_{g'}, \quad \text{where } g' = g - g_1 \cdot e_1 .$$

Note that under our simplified assumption for D, by simple algebraic manipulation we have $M_{e_1} = \mathbb{E}_{D(u)} u_1 uu^\mathsf{T} = e_1 e_1^\mathsf{T}$. Moreover, by definition, g_1 and g' are independent. It turns out that the entropy constraint implies $\mathbb{E}_{g'} \|M_{g'}\| \lesssim \sqrt{\log d} \cdot 1/n$ (using concentration bounds for Gaussian matrix series [Oli10]). Therefore, if we condition on the event $g_1 > \eta^{-1}\sqrt{\log d}$, we have that $M_g = g_1 e_1 e_1^\mathsf{T} + M_{g'}$ consists of two parts: a rank-1 single part $g_1 e_1 e_1^\mathsf{T}$ with with eigenvalue larger than $\eta^{-1}\sqrt{\log d}$, and a noise part which has spectral norm at most $\lesssim \sqrt{\log d}$. Hence, by the eigenvector perturbation theorem we have that the top eigenvector is $O(\eta^{1/2})$-close to e_1 as desired.

Taking $\eta = 0.1$, we see with $1/\operatorname{poly}(d)$ probability the event $g_1 > \eta^{-1}\sqrt{\log d}$ will happen, and therefore by repeating this procedure $\operatorname{poly}(d)$ times, we obtain a vector that is $O(\eta^{1/2})$-close to e_1. We can find other vectors similarly by repeating the process (in a slightly more delicate way), and the accuracy can also be boosted (see Sections 4 and 5 for details).

2.2 Overcomplete fourth-order tensor

In this section, we give a high-level description of a robust sum-of-squares version of the tensor decomposition algorithm FOOBI [LCC07]. For simplicity of the demonstration, we first work with the noiseless case where we are given a tensor $T \in (\mathbb{R}^d)^{\otimes 4}$ of the form

$$T = \sum_{i=1}^{n} a_i^{\otimes 4} . \tag{2.2}$$

We will first review the key step of FOOBI algorithm and then show how to convert it into a sum-of-squares algorithm that will naturally be robust to noise.

To begin with, we observe that by viewing T as a $d^2 \times d^2$ matrix of rank n, we can easily find the span of the $a_i^{\otimes 2}$'s by low-rank matrix factorization. However, since the low rank matrix factorization is only unique up to unitary transformation, we are not able to recover the $a_i^{\otimes 2}$'s from the subspace that they live in. The key observation of [LCC07] is that the $a_i^{\otimes 2}$'s are actually the only "rank-1" vectors in the span, under a mild algebraic independence condition. Here, a d^2-dimensional vector is called "rank-1" if it is a tensor product of two vectors of dimension d.

Lemma 2.1 ([LCC07]). *Suppose the following set of vectors is linearly independent,*

$$\left\{ a_i^{\otimes 2} \otimes a_j^{\otimes 2} - (a_i \otimes a_j)^{\otimes 2} \;\middle|\; i \neq j \right\} . \tag{2.3}$$

Then every vector $x^{\otimes 2}$ in the linear span of $a_1^{\otimes 2}, \ldots, a_n^{\otimes 2}$ is a multiple of one of the vectors $a_i^{\otimes 2}$.

This observation leads to the algorithm FOOBI, which essentially looks for rank-1 vectors in the span of $a_i^{\otimes 2}$'s. The main drawback is that it uses simultaneous diagonalization as a sub-procedure, which is unlikely to tolerate noise better than inverse polynomial in d, and in fact no noise tolerance guarantee has been explicitly shown for it before.

Our approach starts with rephrasing the original proof of Lemma 2.1 into the following SoS proof (which only uses polynomial inequalities that can be proved by SoS).

Proof of Lemma 2.1. Let $\alpha_1, \dots, \alpha_n$ be multipliers such that $x^{\otimes 2} = \sum_{i=1}^n \alpha_i \cdot a_i^{\otimes 2}$.[5] Then, these multipliers satisfy the following quadratic equations:

$$x^{\otimes 4} = \sum\nolimits_{i,j} \alpha_i \alpha_j \cdot a_i^{\otimes 2} \otimes a_j^{\otimes 2}\,,$$
$$x^{\otimes 4} = \sum\nolimits_{i,j} \alpha_i \alpha_j \cdot (a_i \otimes a_j)^{\otimes 2}\,.$$

Together, the two equations imply that

$$0 = \sum\nolimits_{i \neq j} \alpha_i \alpha_j \cdot \left(a_i^{\otimes 2} \otimes a_j^{\otimes 2} - (a_i \otimes a_j)^{\otimes 2} \right)\,.$$

By assumption, the vectors $a_i^{\otimes 2} \otimes a_j^{\otimes 2} - (a_i \otimes a_j)^{\otimes 2}$ are linearly independent for $i \neq j$. Therefore, from the equation above, we conclude $\sum_{i \neq j} \alpha_i^2 \alpha_j^2 = 0$, meaning that at most one of α_i can be non-zero. Furthermore this argument is a SoS proof, since for any matrix $A \in \mathbb{R}^{D \times D}$ with linearly independent columns and any vector polynomial $v \in \mathbb{R}[x]^D$, the inequality $\|v\|^2 \leqslant \frac{1}{\sigma_{\min}(A)^2} \|Av\|^2$ can be proved by SoS (here $\sigma_{\min}(A)$ denotes the least singular value of matrix A). So choosing A to be the matrix with columns $a_i^{\otimes 2} \otimes a_j^{\otimes 2} - (a_i \otimes a_j)^{\otimes 2}$ for $i \neq j$ and v to be the vector with entries $\alpha_i \alpha_j$, we find by SoS proof that $\|\alpha\|_4^4 - \|\alpha\|_2^4 = 0$. □

When there is noise present, we cannot find the true subspace of the $a_i^{\otimes 2}$'s and instead we only have an approximation, denoted by V, of that subspace. We will modify the proof above by starting with a polynomial inequality

$$\| \operatorname{Id}_V x^{\otimes 2} \|^2 \geqslant (1 - \delta) \|x^{\otimes 2}\|^2\,, \tag{2.4}$$

which constrains $x^{\otimes 2}$ to be close to the estimated subspace V (where δ is a small number that depends on error and condition number). Then an extension of the proof of Lemma 2.1 will show that equation (2.4) implies (via a SoS proof) that for some small enough δ,

$$\sum_{i \neq j} \alpha_i^2 \alpha_j^2 \leqslant o(1)\,. \tag{2.5}$$

Note that $\alpha = K x^{\otimes 2}$ is a linear transformation of $x^{\otimes 2}$, and furthermore K is the pseudo-inverse of the matrix with columns $a_i^{\otimes 2}$. Moreover, if we assume for a moment that α has 2-norm 1 (which is not true in general), then the equation above further implies that

$$\sum_{i=1}^n \langle K_i, x^{\otimes 2} \rangle^4 = \|\alpha\|_4^4 \geqslant 1 - o(1)\,, \tag{2.6}$$

[5] technically, $\alpha_1, \dots, \alpha_n$ are polynomials in x so that $x^{\otimes 2} = \sum_{i=1}^n \alpha_i \cdot a_i^{\otimes 2}$ holds

where $K_i \in \mathbb{R}^{d^2}$ is the i-th row of K. This effectively gives us access to the 4-tensor $\sum_i K_i^{\otimes 4}$ (which has ambient dimension d^2 when flattened into a matrix), since equation (2.6) is anyway the constraint that would have been used by the SoS algorithm if given the tensor $\sum_i K_i^{\otimes 4}$ as input. Note that because the K_i are not necessarily (close to) orthogonal, we cannot apply the SoS orthogonal tensor decomposition algorithm directly. However, since we are working with a 4-tensor whose matrix flattening has higher dimension d^2, we can whiten K_i effectively in the SoS framework and then use the orthogonal SoS tensor decomposition algorithm to find the K_i's, which will in turn yield the a_i's.

Many details were omitted in the heuristic argument above (for example, we assumed α to have norm 1). The full argument follows in Section 8.

2.3 Random overcomplete third-order tensor

In the random overcomplete setting, the input tensor is of the form

$$T = \sum_{i=1}^{n} a_i^{\otimes 3} + E\,,$$

where each a_i is drawn uniformly at random from the Euclidean unit sphere, we have $d < n \leqslant d^{1.5}/(\log d)^{O(1)}$, and E is some noise tensor such that $\|E\|_{\{1\},\{2,3\}} < \varepsilon$ or alternatively such that a constant-degree sum-of-squares relaxation of the injective norm of E is at most ε.

Our original rounding approach depends on the target vectors a_i being orthonormal or nearly so. But when $n \gg d$ in this overcomplete setting, orthonormality fails badly: the vectors a_i are not even linearly independent.

We circumvent this problem by embedding the vectors a_i in a larger ambient space—specifically by taking the tensor powers $a'_1 = a_1^{\otimes 2}, \ldots, a'_n = a_n^{\otimes 2}$. Now the vectors $a'_1, \ldots, a'_n$ are linearly independent (with probability 1) and actually close to orthonormal with high probability. Therefore, if we had access to the order-6 tensor $\sum (a'_i)^{\otimes 3} = \sum a_i^{\otimes 6}$, then we could (almost) apply our rounding method to recover the vectors a'_i.

The key here will be to use the sum-of-squares method to generate a pseudo-distribution over the unit sphere having T as its third-order moments tensor, and then to extract from it the set of order-6 pseudo-moments estimating the moment tensor $\sum_i a_i^{\otimes 6}$. This pseudo-distribution would obey the constraint $\{(u \otimes u \otimes u)^{\mathsf{T}} T \geqslant 1 - \varepsilon\}$, which implies the constraint $\{\sum_i \langle a_i, u\rangle^3 \geqslant 1 - \varepsilon\}$, saying, informally, that our pseudo-distribution is close to the actual uniform distribution over $\{a_i\}$. Substituting $v = u^{\otimes 2}$, we obtain an implied pseudo-distribution in v which therefore ought to be close to the uniform distribution over $\{a'_i\}$, and we should therefore be able to round the order-3 pseudo-moments of v to recover $\{a'_i\}$.

Only two preconditions need to be checked: first that $\sum_i (a'_i)(a'_i)^{\mathsf{T}}$ is not too large in spectral norm, and second that our pseudo-distribution in v satisfies the constraint $\{\sum_i \langle a'_i, v\rangle^3 \geqslant 1 - O(\varepsilon)\}$. The first precondition is true (except for a spurious eigenspace which can harmlessly be projected away) and is essentially equivalent to a line of matrix concentration arguments previously made in [HSSS16]. The second precondition follows from a line of constant-degree sum-of-squares proofs, notably extending arguments made in [GM15] stating that the constraints $\{\sum_i \langle a_i, u\rangle^3 \geqslant 1 - \varepsilon, \|u\|^2 = 1\}$

imply with constant-degree sum-of-squares proofs that $\{\sum_i \langle a_i, u\rangle^k \geqslant 1 - O(\varepsilon) - \tilde{O}(n/d^{3/2})\}$ for some higher powers k. The rigorous verification of these conditions is detailed in Section 7.

3 Preliminaries

Unless explicitly stated otherwise, $O(\cdot)$-notation hides absolute multiplicative constants. Concretely, every occurrence of $O(x)$ is a placeholder for some function $f(x)$ that satisfies $\forall x \in \mathbb{R}. |f(x)| \leqslant C|x|$ for some absolute constant $C > 0$. Similarly, $\Omega(x)$ is a placeholder for a function $g(x)$ that satisfies $\forall x \in \mathbb{R}. |g(x)| \geqslant |x|/C$ for some absolute constant $C > 0$.

For a matrix A, let A^+ denote the Moore-Penrose pseudo-inverse of A. For a symmetric positive semidefinite matrix B, let $B^{1/2}$ denote the square root of B, i.e. the unique symmetric positive-semidefinite matrix L such that $L^2 = B$.

The Kronecker product of two matrices A and B is denoted by $A \otimes B$. A useful identity is that $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$ whenever the matrix multiplications are defined. The norm $\|\cdot\|$ denotes the Euclidean norm for vectors and the spectral norm for matrices.

Let $T \in (\mathbb{R}^d)^{\otimes k}$ be a *k-tensor* over $\mathbb{R}^d$ such that $T = \sum_{i_1,\dots,i_k} T_{i_1\cdots i_k} e_{i_1} \otimes \cdots \otimes e_{i_k}$, where $e_1, \dots, e_d$ is the standard basis of $\mathbb{R}^d$. We say T is *symmetric* if the entries $(T_{i_1,\dots,i_k})$ are invariant under permuting the indices. The k index positions of T are called *modes*. The injective norm $\|T\|_{\text{inj}}$ is the maximum value of $\langle T, x_1 \otimes \cdots \otimes x_k\rangle$ over all vectors $x_1, \dots, x_k \in \mathbb{R}^d$ with $\|x_1\| = \cdots = \|x_k\| = 1$. A useful class of multilinear operations on tensors has the form $T \mapsto (A_1 \otimes \cdots \otimes A_k)T$, where $A_1, \dots, A_k$ are matrices with d columns. (This notation is the same as the Kronecker product notation for matrices, that is, $(A_1 \otimes \cdots \otimes A_k)T = \sum_{i_1,\dots,i_k} T_{i_1\cdots i_k}(A_1 e_{i_1}) \otimes \cdots \otimes (A_k e_{i_k})$.) If some of the matrices A_i are row vectors, and the others are the identity matrix, then the corresponding operation is called tensor contraction. For example, for a third-order tensor $T \in (\mathbb{R}^d)^{\otimes 3}$ and a vector $g \in \mathbb{R}^d$, we call $(\mathrm{Id} \otimes \mathrm{Id} \otimes g^\mathsf{T})T$ the contraction of the third mode of T with g. (Some authors use the notation $T(\mathrm{Id}, \mathrm{Id}, g)$ to denote this operation.)

For a bipartition A, B of the index set $[k]$ of T, we let $\|T\|_{A,B}$ denote the spectral norm of the matrix unfolding $T_{A,B}$ of T with rows indexed by the indices in A and columns indexed by indices in B. Concretely,

$$\|T\|_{A,B} = \max_{\substack{x \in (\mathbb{R}^d)^{\otimes |A|},\ y \in (\mathbb{R}^d)^{\otimes |B|} \\ \|x\| \leqslant 1,\ \|y\| \leqslant 1}} \sum_{i_1,\dots,i_k} T_{i_1\cdots i_k} \cdot x_{i_A} y_{i_B} \,,$$

Here, $i_A = i_{a_1} \cdots i_{a_{|A|}}$ and $i_B = i_{b_1} \cdots i_{b_{|B|}}$ are multi-indices, where $A = \{a_1, \dots, a_{|A|}\}$ and $B = \{b_1, \dots, b_{|B|}\}$. For $k = 2$, $\|T\|_{\{1\},\{2\}}$ is the spectral norm of T viewed as a d-by-d matrix. For $k = 3$, $\|T\|_{\{1,2\},\{3\}}$ is the spectral norm of T viewed as a d^2-by-d matrix with rows indexed by the first two modes of T and columns indexed by the last index of T. For symmetric 3-tensors, all norms $\|T\|_{\{1,2\},\{3\}}$, $\|T\|_{\{1,3\},\{2\}}$, and $\|T\|_{\{2,3\},\{1\}}$ are the same.

3.1 Pseudo-distributions

Pseudo-distributions generalize probability distributions in a way that allows us to optimize efficiently over moments of pseudo-distributions. We represent a discrete probability distribution

D over $\mathbb{R}^n$ by its probability mass function $D\colon \mathbb{R}^n \to \mathbb{R}$ such that $D(x)$ is the probability of x under the distribution for every $x \in \mathbb{R}^n$. This function is nonnegative point-wise and satisfies $\sum_{x\in\mathrm{supp}(D)} D(x) = 1$. For pseudo-distributions we relax the nonnegative requirement and only require that the function passes a set of simple nonnegativity tests.

A *degree-d pseudo-distribution over* $\mathbb{R}^n$ is a finitely[6] supported function $D\colon \mathbb{R}^n \to \mathbb{R}$ such that $\sum_{x\in\mathrm{supp}(D)} D(x) = 1$ and $\sum_{x\in\mathrm{supp}(D)} D(x)f(x)^2 \geqslant 0$ for every function $f\colon \mathbb{R}^n \to \mathbb{R}$ of degree at most $d/2$. We define the *pseudo-expectation* of a (possibly vector-valued or matrix-valued) function f with respect to D as

$$\tilde{\mathbb{E}}_D f \stackrel{\text{def}}{=} \sum_{x\in\mathrm{supp}(D)} D(x)f(x)\,.$$

In order to emphasize which variable is bound by the pseudo-expectation, we write $\tilde{\mathbb{E}}_{D(x)} f(x)$. (This notation is useful if $f(x)$ is a more complicated expression involving several variables.)

Note that a degree-∞ pseudo-distribution D satisfies $D(x) \geqslant 0$ for all $x \in \mathbb{R}^n$. Therefore, D is an actual probability distribution (with finite support). The pseudo-expectation $\tilde{\mathbb{E}}_D f = \mathbb{E}_D f$ of a function f is its expected value under the distribution D.

Our algorithms will not work with pseudo-distributions (as finitely-supported functions on $\mathbb{R}^n$) directly. Instead the algorithms will work with moment tensors $\tilde{\mathbb{E}}_{D(x)}(1, x_1, \ldots, x_n)^{\otimes d}$ of pseudo-distributions and the associated linear functional $p \mapsto \tilde{\mathbb{E}}_{D(x)} p(x)$ on polynomials p of degree at most d.

Unlike actual probability distribution, pseudo-distributions admit general, efficient optimization algorithms. In particular, the set of low-degree moments of pseudo-distributions has an efficient separation oracle.

Theorem 3.1 ([Sho87, Par00, Las01])**.** *For $n, d \in \mathbb{N}$, the following set admits an $n^{O(d)}$-time weak separation oracle (in the sense of [GLS81]),*

$$\left\{ \tilde{\mathop{\mathbb{E}}_{D(x)}} (1, x_1, \ldots, x_n)^{\otimes d} \;\middle|\; \text{degree-}d\text{ pseudo-distribution } D \text{ over } \mathbb{R}^n \right\}.$$

This theorem, together with the equivalence of separation and optimization [GLS81] allows us to solve a wide range of optimization and feasibility problems over pseudo-distributions efficiently.

The following definition captures what kind of linear constraints are induced on a pseudo-distribution over $\mathbb{R}^n$ by a system of polynomial constraints over $\mathbb{R}^n$.

Definition 3.2. Let D be a degree-d pseudo-distribution over $\mathbb{R}^n$. For a system of polynomial constraints $\mathcal{A} = \{f_1 \geqslant 0, \ldots, f_m \geqslant 0\}$ with $\deg(f_i) \leqslant \ell$ for every i, we say that D *satisfies the polynomial constraints $\mathcal{A}$ at degree ℓ*, denoted $D \models_\ell \mathcal{A}$, if $\tilde{\mathbb{E}}_D \left(\prod_{i\in S} f_i\right) h \geqslant 0$ for every $S \subseteq [m]$ and every sum-of-squares polynomial h on $\mathbb{R}^n$ with $|S|\ell + \deg h \leqslant d$.

This is a relaxation (to pseudo-distributions) of the statement that the probability mass of a true distribution contains only solutions to $\mathcal{A}$. Indeed, if an actual distribution D is supported on the solutions to $\mathcal{A}$, then D satisfies $D \models_\ell \mathcal{A}$ regardless of the value of ℓ.

[6] We restrict these functions to be finitely supported in order to avoid integrals and measurability issues. It turns out to be without loss of generality in our context.

We say that D satisfies $\mathcal{A}$ (without further specifying the degree) if $D \models_\ell \mathcal{A}$ for $\ell = \max_{\{f \geqslant 0\} \subseteq \mathcal{A}} \deg f$. We say that a system $\mathcal{A}$ of polynomial constraints in variables x is *explicitly bounded* if it contains a constraint of the form $\{\|x\|^2 \leqslant M\}$. The following theorem follows from Theorem 3.1 and [GLS81]. We give a proof in Appendix B for completeness.

Theorem 3.3. *There exists a $(n + |\mathcal{A}|)^{O(d)}$-time algorithm that, given any explicitly bounded and satisfiable system $\mathcal{A}$ of polynomial constraints in n variables, outputs (up to arbitrary accuracy) a degree-d pseudo-distribution that satisfies $\mathcal{A}$.*

3.2 Sum of squares proofs

Let $x = (x_1, \ldots, x_n)$ be a tuple of indeterminates. Let $\mathbb{R}[x]$ be the set of polynomials in these indeterminates with real coefficients. A polynomial p is a *sum-of-squares* if there are polynomials $q_1, \ldots, q_r$ such that $p = q_1^2 + \cdots + q_r^2$. Let $f_1, \ldots, f_r$ and g be multivariate polynomials in $\mathbb{R}[x]$. A *sum-of-squares proof* that the constraints $\{f_1 \geqslant 0, \ldots, f_r \geqslant 0\}$ imply the constraint $\{g \geqslant 0\}$ consists of sum-of-squares polynomials $(p_S)_{S \subseteq [n]}$ in $\mathbb{R}[x]$ such that

$$g = \sum_{S \subseteq [n]} p_S \cdot \prod_{i \in S} f_i \,.$$

We say that this proof has *degree* ℓ if every set $S \subseteq [n]$ satisfies $\deg(p_S \cdot \prod_{i \in S} f_i) \leqslant \ell$ (in particular, this would imply $p_S = 0$ for every set S such that $\deg \prod_{i \in S} f_i > \ell$). If there exists a degree-ℓ sum-of-squares proof that $\{f_1 \geqslant 0, \ldots, f_r \geqslant 0\}$ implies $\{g \geqslant 0\}$, we write

$$\{f_1 \geqslant 0, \ldots, f_r \geqslant 0\} \vdash_\ell \{g \geqslant 0\} \,.$$

In order to emphasize the indeterminates for the proofs, we sometimes write $\{f_1(x) \geqslant 0, \ldots, f_r(x) \geqslant 0\} \vdash_{x,\ell} \{g(x) \geqslant 0\}$.

Sum-of-squares proofs obey the following inference rules, for all polynomials $f, g \colon \mathbb{R}^n \to \mathbb{R}$ and $F \colon \mathbb{R}^n \to \mathbb{R}^m, G \colon \mathbb{R}^n \to \mathbb{R}^k, H \colon \mathbb{R}^p \to \mathbb{R}^n$,

$$\frac{\mathcal{A} \vdash_\ell \{f \geqslant 0, g \geqslant 0\}}{\mathcal{A} \vdash_\ell \{f + g \geqslant 0\}}, \quad \frac{\mathcal{A} \vdash_\ell \{f \geqslant 0\},\ \mathcal{A} \vdash_{\ell'} \{g \geqslant 0\}}{\mathcal{A} \vdash_{\ell+\ell'} \{f \cdot g \geqslant 0\}}, \qquad \text{(addition and multiplication)}$$

$$\frac{\mathcal{A} \vdash_\ell \mathcal{B},\ \mathcal{B} \vdash_{\ell'} C}{\mathcal{A} \vdash_{\ell \cdot \ell'} C}, \qquad \text{(transitivity)}$$

$$\frac{\{F \geqslant 0\} \vdash_\ell \{G \geqslant 0\}}{\{F(H)) \geqslant 0\} \vdash_{\ell \cdot \deg(H)} \{G(H) \geqslant 0\}}. \qquad \text{(substitution)}$$

Sum-of-squares proofs are sound and complete for polynomial constraints over pseudo-distributions, in the sense that sum-of-squares proofs allow us to reason about what kind of polynomial constraints are satisfied by a pseudo-distribution. We defer the proofs of the following lemmas to Appendix B.

Lemma 3.4 (Soundness)**.** *If $D \models_\ell \mathcal{A}$ for a pseudo-distribution D and there exists a sum-of-squares proof $\mathcal{A} \vdash_{\ell'} \mathcal{B}$, then $D \models_{\ell \cdot \ell'} \mathcal{B}$.*

Lemma 3.5 (Completeness)**.** *Suppose $d \geqslant \ell' \geqslant \ell$, and $\mathcal{A}$ is a collection of polynomial constraints with degree at most ℓ, and $\mathcal{A} \vdash \{x_1^2 + \cdots + x_n^2 \leqslant B\}$ for some finite B. Let $\{g \geqslant 0\}$ be a polynomial constraint with degree ℓ'. If every degree-d pseudo-distribution D that satisfies $D \models_\ell \mathcal{A}$ also satisfies $D \models_{\ell'} \{g \geqslant 0\}$, then for every $\varepsilon > 0$, there is a sum-of-squares proof $\mathcal{A} \vdash_d \{g \geqslant -\varepsilon\}$.* [7]

3.3 Matrix constraints and sum-of-squares proofs

In sections 4 and 9, we still state positive-semidefiniteness constraints on matrices, which will be implied by sum-of-squares proofs. We define notation to express what it means for a matrix constraint to be implied by sum-of-squares. While the duality between proof systems and convex relaxations also holds in the matrix case [Cim12], and it is possible to give a full treatment of matrix constraints in sum-of-squares, here we give an abridged and simplified treatment sufficient for our purposes.

Definition 3.6. Let $\mathcal{A}$ be a set of polynomial constraints in indeterminant x, and M is a symmetric $p \times p$ matrix with entries in $\mathbb{R}[x]$. Then we write $\mathcal{A} \vdash_\ell \{M \succeq 0\}$ if there exists a set of polynomials $q_1(x), \ldots q_m(x)$ and a set of vectors $v_1(x), \ldots, v_m(x)$ of p-dimension with entries in $\mathbb{R}[x]$ such that $\mathcal{A} \vdash_{\ell_i} \{q_i \geqslant 0\}$ where $\ell_i + 2\deg(v_i) \leqslant \ell$ for every i, and

$$M = \sum_{i=1}^{m} q_i(x) v_i(x) v_i(x)^\mathsf{T} . \tag{3.1}$$

The proof that sum-of-squares is sound for these matrix constraints is very similar to the analogous proof of Lemma 3.4 (see Appendix B).

Lemma 3.7. *Let D be a pseudo-distribution of degree d and $d \geqslant \ell\ell'$. Suppose $D \models_\ell \mathcal{A}$, and $\mathcal{A} \vdash_{\ell'} M \succeq 0$. Then $\tilde{\mathbb{E}}[M] \succeq 0$.*

We now give some basic properties of these matrix sum-of-squares proofs.

Lemma 3.8. *Suppose A, B' are symmetric matrix polynomials such that $\vdash \{A \succeq 0,\ B \succeq 0\}$. Then $\vdash \{A \otimes B \succeq 0\}$.*

Proof. Express $A = \sum_{i=1}^n q_i(x) u_i(x) u_i(x)^\mathsf{T}$ and $B = \sum_{i=1}^m r_i(x) v_i(x) v_i(x)^\mathsf{T}$. Then

$$A \otimes B = \sum_{i=1}^{n} \sum_{j=1}^{m} q_i(x) r_j(x) \big[u_i(x) \otimes v_j(x)\big]\big[u_i(x) \otimes v_j(x)\big]^\mathsf{T} . \qquad \square$$

Lemma 3.9. *Suppose A, B, A', B' are symmetric matrix polynomials such that $\vdash \{A \succeq 0, B' \succeq 0, A \succeq A', B \succeq B'\}$. Then $\vdash \{A \otimes B \succeq A' \otimes B', B \otimes A \succeq B' \otimes A'\}$.*

Proof. By Lemma 3.8, we have $\vdash A \otimes (B - B') \succeq 0$ and $\vdash (A - A') \otimes B' \succeq 0$. Adding the two equations we complete the proof. We may also take the tensor powers in the other order. $\square$

[7] The completeness claim stated here does not match the strength of the corresponding soundness claim. This reflects an impreciseness in how we count the degrees of intermediate sum-of-squares proofs (in particular our degree accounting is not tight under proof composition), and does not reflect than the power of the proofs themselves.

Lemma 3.10. *Let $u = [u_1, \dots, u_d]$ be an indeterminate. Then $\vdash \{uu^\mathsf{T} \preceq \|u\|^2 \cdot \mathrm{Id}_d\}$.*

Proof. The conclusion follows from the following explicit decomposition

$$\vdash \|u\|^2 \,\mathrm{Id} - uu^\mathsf{T} = \sum_{1 \leqslant i < j \leqslant d} (u_i e_j - u_j e_i)(u_i e_j - u_j e_i)^\mathsf{T} \succeq 0 \qquad \square$$

4 Rounding pseudo-distributions

4.1 Rounding by matrix diagonalization

The following theorem analyzes a form of Jennrich's algorithm for tensor decomposition through matrix diagonalization, when applied to the moments of a pseudo-distribution. We show that if the pseudo-distribution $D(u)$ has good correlation with some vector $a^{\otimes k}$, then with good chance a simple random contraction of the $(k+2)$-th moments of the pseudo-distribution will return a matrix with top eigenvector close to a.

Theorem 4.1 below is the key ingredient toward a polynomial-time algorithm. It states that in order for Jennrich's approach to successfully extract a solution in polynomial time, the correlation of the desired solution with the $(k+2)$-th moments of the pseudo-distribution only needs to be large compared to the spectral norm of the covariance matrix of the pseudo-distribution. This covariance matrix can be made as small as $O(1/n)$ in spectral norm in many situations, including—as a toy example—when D is a uniform distribution over n orthogonal unit vectors. Therefore in this sense the condition (4.1) below is a fairly weak requirement, which is key to the polynomial-time algorithm in Section 5.1.

Theorem 4.1. *Let $k \in \mathbb{N}$ be even and $\varepsilon \in (0,1)$. Let D be a degree-$O(k)$ pseudo-distribution over $\mathbb{R}^d$ that satisfies $\{\|u\|^2 \leqslant 1\}_{D(u)}$, let $a \in \mathbb{R}^d$ be a unit vector. Suppose that*

$$\tilde{\mathbb{E}}_{D(u)} \langle a, u \rangle^{k+2} \geqslant \Omega\left(\frac{1}{\varepsilon\sqrt{k}}\right) \cdot \left\|\tilde{\mathbb{E}}_{D(u)} uu^\mathsf{T}\right\| . \tag{4.1}$$

Then, with probability at least $1/d^{O(k)}$ over the choice of $g \sim \mathcal{N}(0, \mathrm{Id}_d^{\otimes k})$, the top eigenvector $u^\star$ of the following matrix M_g satisfies $\langle a, u^\star \rangle^2 \geqslant 1 - O(\varepsilon)$,

$$M_g := \tilde{\mathbb{E}}_{D(u)} \langle g, u^{\otimes k} \rangle \cdot uu^\mathsf{T} . \tag{4.2}$$

As before, we decompose M_g into two parts, with $M_{a^{\otimes k}}$ and $M_{g'}$ defined in analogy with M_g.

$$M_g = \langle g, a^{\otimes k} \rangle \cdot M_{a^{\otimes k}} + M_{g'} \quad \text{where } g' = g - \langle g, a^{\otimes k} \rangle \cdot a^{\otimes k} . \tag{4.3}$$

Our proof of Theorem 4.1 consists of two propositions: one about the good part $M_{a^{\otimes k}}$ and one about the noise part $M_{g'}$. The first proposition shows that $M_{a^{\otimes k}}$ is close to a multiple of aa^T in spectral norm (which means that the top eigenvector of $M_{a^{\otimes k}}$ is close to a).

Proposition 4.2. *In the setting of Theorem 4.1, for $t = \tilde{\mathbb{E}}_{D(u)} \langle a, u \rangle^{k+2}$,*

$$\left\| M_{a^{\otimes k}} - t \cdot aa^\mathsf{T} \right\| \leqslant O(\varepsilon) \cdot t . \tag{4.4}$$

The second proposition shows that $M_{g'}$ has small spectral norm in expectation.

Proposition 4.3. *In the setting of Theorem 4.1, let $g' = g - \langle g, a^{\otimes k}\rangle \cdot a^{\otimes k}$. Then, for $t = \tilde{\mathbb{E}}_{D(u)}\langle a, u\rangle^{k+2}$,*

$$\mathop{\mathbb{E}}_{g'} \left\| M_{g'} \right\| \leqslant O(\varepsilon^2 k \log d)^{1/2} \cdot t\,.$$

Before proving the above propositions, we demonstrate how they allow us to prove Theorem 4.1.

Proof of Theorem 4.1. We are to show that with probability $1/d^{O(k)}$ over the choice of the Gaussian vector g, there exists $s \in \mathbb{R}$ such that $\left\| s \cdot M_g - aa^\top \right\| \leqslant O(\varepsilon)$. By Davis-Kahan Theorem (see Theorem A.4), this implies the conclusion of Theorem 4.1. Let $t = \tilde{\mathbb{E}}_{D(u)}\langle a, u\rangle^{k+2}$. For a parameter $\tau = \Omega(k \log d)^{1/2}$, we bound the spectral norm conditioned on the event $\langle g, a^{\otimes k}\rangle \geqslant \tau$,

$$\begin{aligned}
&\mathop{\mathbb{E}}_{g} \left[\left\| \tfrac{1}{\langle g, a^{\otimes k}\rangle \cdot t} M_g - aa^\top \right\| \;\middle|\; \langle g, a^{\otimes k}\rangle \geqslant \tau \right] \\
&\leqslant \left\| \tfrac{1}{t} M_{a^{\otimes k}} - aa^\top \right\| + \mathop{\mathbb{E}}_{g} \left[\tfrac{1}{\langle g, a^{\otimes k}\rangle \cdot t} \left\| M_{g'} \right\| \;\middle|\; \langle g, a^{\otimes k}\rangle \geqslant \tau \right] && \text{(by (4.3))} \\
&\leqslant \left\| \tfrac{1}{t} M_{a^{\otimes k}} - aa^\top \right\| + \tfrac{1}{\tau \cdot t} \cdot \mathop{\mathbb{E}}_{g'} \left\| M_{g'} \right\| && \text{(by independence of } \langle g, a^{\otimes k}\rangle \text{ and } g') \\
&\leqslant O(\varepsilon) + \tfrac{1}{\tau} \cdot O(\varepsilon^2 k \log d)^{1/2} && \text{(by Proposition 4.2 and 4.3)} \\
&\leqslant O(\varepsilon)\,. && (4.5)
\end{aligned}$$

By Markov's inequality, it follows that conditioned on $\langle g, a^{\otimes k}\rangle \geqslant \tau$, the event $\left\| \frac{1}{\langle g, a^{\otimes k}\rangle \cdot t} M_g - aa^\top \right\| \leqslant O(\varepsilon)$ has probability at least $\Omega(1)$. The theorem follows because the event $\langle g, a^{\otimes k}\rangle \geqslant \tau$ has probability at least $d^{-O(k)}$. □

Proof of Proposition 4.2. We are to bound the spectral norm of $M_{a^{\otimes k}} - t \cdot aa^\top$ for $t = \tilde{\mathbb{E}}_{D(u)}\langle a, u\rangle^{k+2}$. Let $\alpha = \|\tilde{\mathbb{E}}_{D(u)}\, uu^\top\|$. Let $\mathrm{Id}_1 = aa^\top$ be the projector onto the subspace spanned by a and let $\mathrm{Id}_{-1} = \mathrm{Id} - \mathrm{Id}_1$ be the projector on the orthogonal complement. By our choice of t, we have $\mathrm{Id}_1 M_{a^{\otimes k}} \mathrm{Id}_1 = t \cdot aa^\top$.

Since $\mathrm{Id}_{-1} \mathrm{Id}_1 = 0$, we can upper bound the spectral norm of $M_{a^{\otimes k}} - t \cdot a^{\otimes k} a^{\otimes k^\top}$,

$$\begin{aligned}
\|M_{a^{\otimes k}} - t \cdot \mathrm{Id}_1\| &\leqslant \|\mathrm{Id}_1 (M_{a^{\otimes k}} - t \cdot \mathrm{Id}_1)\, \mathrm{Id}_1\| + \|\mathrm{Id}_{-1} M_{a^{\otimes k}} \mathrm{Id}_{-1}\| + 2\,\|\mathrm{Id}_1 M_{a^{\otimes k}} \mathrm{Id}_{-1}\| \\
&\leqslant \|\mathrm{Id}_{-1} M_{a^{\otimes k}} \mathrm{Id}_{-1}\| + 2\,\|\mathrm{Id}_1 M_{a^{\otimes k}} \mathrm{Id}_{-1}\| && \text{(because } \mathrm{Id}_1 M_{a^{\otimes k}} \mathrm{Id}_1 = t \cdot \mathrm{Id}_1) \\
&\leqslant \|\mathrm{Id}_{-1} M_{a^{\otimes k}} \mathrm{Id}_{-1}\| + 2\,\|\mathrm{Id}_1 M_{a^{\otimes k}} \mathrm{Id}_1\|^{1/2} \cdot \|\mathrm{Id}_{-1} M_{a^{\otimes k}} \mathrm{Id}_{-1}\|^{1/2} && \text{(because } M_{a^{\otimes k}} \succeq 0) \\
&\leqslant \|\mathrm{Id}_{-1} M_{a^{\otimes k}} \mathrm{Id}_{-1}\| + 2\sqrt{\alpha} \cdot \|\mathrm{Id}_{-1} M_{a^{\otimes k}} \mathrm{Id}_{-1}\|^{1/2}\,. && (4.6)
\end{aligned}$$

It remains to bound the spectral norm of $\mathrm{Id}_{-1} M_{a^{\otimes k}} \mathrm{Id}_{-1}$,

$$\begin{aligned}
\|\mathrm{Id}_{-1} M_{a^{\otimes k}} \mathrm{Id}_{-1}\| &= \left\| \mathop{\tilde{\mathbb{E}}}_{D(u)} \langle a, u\rangle^k \cdot \mathrm{Id}_{-1}\, uu^\top\, \mathrm{Id}_{-1} \right\| \\
&\leqslant \mathop{\tilde{\mathbb{E}}}_{D(u)} \langle a, u\rangle^k \cdot (1 - \langle a, u\rangle^2) && \text{(because } \vdash \mathrm{Id}_{-1}\, uu^\top\, \mathrm{Id}_{-1} \preceq (\|u\|^2 - \langle a_1, u\rangle^2)\,\mathrm{Id}) \\
&\leqslant \tfrac{2}{k-2} \mathop{\tilde{\mathbb{E}}}_{D(u)} \langle a, u\rangle^2 && \text{(using } \vdash_{k+2} x^{k-2} \cdot (1 - x^2) \leqslant \tfrac{2}{k-2}\text{; see below)}
\end{aligned}$$

$$\leqslant \frac{2}{k-2} \cdot \alpha \tag{4.7}$$

Basic calculus shows that the inequality $x^{k-2} \cdot (1-x^2) \leqslant \frac{2}{k-2}$ holds for all $x \in \mathbb{R}$. Since it is a true univariate polynomial inequality in x, it has a sum-of-squares proof with degree no larger than the degree of the involved polynomials, which is $k+2$ in our case.

Combining (4.6) and (4.7), yields as desired that

$$\left\| M_{a^{\otimes k}} - t \cdot aa^\mathsf{T} \right\| \leqslant O\left(\tfrac{1}{\sqrt{k}}\right) \cdot \alpha \leqslant O(\varepsilon) \cdot t \,,$$

where the second step uses the condition of Theorem 4.1 on $t = \tilde{\mathbb{E}}_{D(u)} \langle a, u \rangle$ and $\alpha = \|\tilde{\mathbb{E}}_{D(u)} uu^\mathsf{T}\|$. □

Proof of Proposition 4.3. The matrix $M_{g'} = \tilde{\mathbb{E}}_{D(u)} \langle g', u^{\otimes k} \rangle \cdot uu^\mathsf{T}$, whose spectral norm we are to bound, is a random contraction of the third-order tensor $T = \tilde{\mathbb{E}}_{D(u)} u \otimes u \otimes (u^{\otimes k})$. Corollary 6.6 gives the following bound on the expected norm of a random contraction in terms of spectral norms of two matrix unfoldings of T—which turn out to be the same in our case due to the symmetry of T.

$$\mathop{\mathbb{E}}_{g'} \|M_{g'}\| \leqslant O(\log d)^{1/2} \cdot \max\{\|T\|_{\{1\},\{23\}}, \|T\|_{\{2\},\{13\}}\} = O(\log d)^{1/2} \cdot \left\| \tilde{\mathbb{E}}_{D(u)} u^{\otimes k} u^\mathsf{T} \right\| . \tag{4.8}$$

Theorem 6.1 shows that for any pseudo-distribution D that satisfies $\{\|u\|^2 \leqslant 1\}$,

$$\left\| \tilde{\mathbb{E}}_{D(u)} u^{\otimes k} u^\mathsf{T} \right\| \leqslant \left\| \tilde{\mathbb{E}}_{D(u)} uu^\mathsf{T} \right\| . \tag{4.9}$$

The statement of the lemma follows by combining the previous bounds (4.8) and (4.9),

$$\mathop{\mathbb{E}}_{g'} \|M_{g'}\| \leqslant O(\log d)^{1/2} \cdot \left\| \tilde{\mathbb{E}}_{D(u)} u^{\otimes k} u^\mathsf{T} \right\| \leqslant O(\log d)^{1/2} \cdot \left\| \tilde{\mathbb{E}}_{D(u)} uu^\mathsf{T} \right\| \leqslant O(\varepsilon^2 k \log d)^{1/2} \cdot t \,,$$

using condition (4.1) of Theorem 4.1 which yields $t = \tilde{\mathbb{E}}_{D(u)} \langle a, u \rangle^{k+2} \geqslant \Omega\left((\varepsilon\sqrt{k})^{-1}\right) \|\tilde{\mathbb{E}}_D uu^\mathsf{T}\|$. □

4.2 Improving accuracy of a found solution

We need one more technical ingredient before analyzing our main algorithm. Previously, the run-time of the sum-of-squares algorithm in [BKS15] (on which our algorithm is based) depended exponentially on the accuracy parameter $1/\varepsilon$, and we give here a simple boosting technique that allows us to remove this dependency and achieve polynomially small error.

Here we have a set of nearly isotropic vectors $a_1, \ldots, a_n$. We give a sum-of-squares proof that if $\sum_{i=1}^n \langle a_i, u \rangle^4$ is only ε off from its maximum possible value, and if u has constant correlation with some a_i, then u must in fact be $(1 - O(\varepsilon))$-correlated with a_i. Intuitively, the former constraint forces D to roughly be a mixture distribution over vectors that are close to $a_1, \ldots, a_n$, and the latter one forces it to actually only be close to a_i. We then briefly show how this proof implies an algorithm to boost the accuracy when we already know a vector b that is 0.01-close to a solution, by solving for a pseudo-distribution with the added constraint $\{\langle b, u \rangle^2 \geqslant 0.9\}$.

Theorem 4.4. *Let $\varepsilon > 0$ be smaller than some constant. Let $a_1, \dots, a_n \in \mathbb{R}^d$ be unit vectors such that $\|\sum_{i=1}^n a_i a_i^\top\| \leqslant 1 + \varepsilon$. Define the following systems of constraints, for each $j \in [n]$ or unit vector $b \in \mathbb{R}^d$:*

$$\mathcal{A}_j := \left\{ \|u\|^2 \leqslant 1, \sum_{i=1}^n \langle a_i, u\rangle^4 \geqslant 1 - \varepsilon,\ \langle a_j, u\rangle^2 \geqslant \tfrac{1}{2} \right\}_{D(u)}$$
$$\mathcal{B}_b := \left\{ \|u\|^2 \leqslant 1, \sum_{i=1}^n \langle a_i, u\rangle^4 \geqslant 1 - \varepsilon,\ \langle b, u\rangle^2 \geqslant 0.9 \right\}_{D(u)} .$$

Then $\mathcal{A}_j \vdash_4 \{\langle a_j, u\rangle^2 \geqslant 1 - 10\varepsilon\}$ for all $j \in [n]$, and also $\mathcal{B}_b \vdash \mathcal{A}_j$ and $\langle a_i, b\rangle^2 \geqslant 0.8$ for some $j \in [n]$.

Proof. We have the following sum-of-squares proof:

$$\begin{aligned}
\mathcal{A} \vdash_{u,4} \quad 1 - \varepsilon &\leqslant \sum_{i=1}^n \langle a_i, u\rangle^4 \\
&\leqslant \langle a_j, u\rangle^4 + \left(\sum_{i \neq j} \langle a_i, u\rangle^2\right)^2 && \text{(by adding only square terms)} \\
&\leqslant \langle a_j, u\rangle^4 + \left(1 + \varepsilon - \langle a_j, u\rangle^2\right)^2 && \text{(using } \vdash_u \textstyle\sum_{i=1}^n \langle a_i, u\rangle^2 \leqslant (1+\varepsilon)\|u\|^2\text{)} \\
&\leqslant \langle a_j, u\rangle^2 + \left(\tfrac{1}{2} + \varepsilon\right)\left(1 + \varepsilon - \langle a_j, u\rangle^2\right) && \text{(since } \vdash 1/2 \leqslant \langle a_j, u\rangle^2 \leqslant 1\text{)} \\
&\leqslant \left(\tfrac{1}{2} - \varepsilon\right)\langle a_j, u\rangle^2 + \tfrac{1}{2} + 2\varepsilon\,, && (4.10)
\end{aligned}$$

which means that $\mathcal{A} \vdash_{u,4} \langle a_j, u\rangle^2 \geqslant 1 - 10\varepsilon$ for $\varepsilon > 0$ small enough.

To show that $\mathcal{B}_b \vdash \mathcal{A}_i$ for some i, it is enough to show that if $\mathcal{B}_b$ is consistent (i.e. there exists a pseudo-distribution satisfying $\mathcal{B}_b$), then there exists $i \in [n]$ such that $\langle a_i, b\rangle^2 \geqslant 0.8$, because it implies $\{\langle a_i, u\rangle^2 \geqslant 1/2\}$ by triangle inequality.

For the sake of contradiction, assume that $\langle a_i, b\rangle^2 < 0.8$ for all $i \in [n]$. Then, by triangle inequality (see Lemma A.2), $\mathcal{B}_b \vdash \{\forall i \in [n]. \langle a_i, u\rangle^2 \leqslant 0.99\}$ which when combined with $\left\|a_i a_i^\top\right\|^2 \leqslant 1 + \varepsilon$ using substitution, contradicts the assumption that $\mathcal{B}_b \vdash \{\sum_{i=1}^n \langle a_i, u\rangle^4 \geqslant 1 - \varepsilon\}$ for small enough $\varepsilon > 0$. □

Corollary 4.5. *Let D be a degree-ℓ pseudo-distribution over $\mathbb{R}^d$ such that $D \models_{\ell/4} \mathcal{B}_b$, with $\mathcal{B}_b$ as defined in Theorem 4.4. Then, there exists $i \in [n]$ such that $\left\|\tilde{\mathbb{E}}_{D(u)} u^{\otimes 2} - a_i^{\otimes 2}\right\|^2 \leqslant O(\varepsilon)$ and $\langle a_i, b\rangle^2 \geqslant 0.8$.*

Proof. By Theorem 4.4, $\mathcal{B}_b \vdash_4 \{\langle a_i, u\rangle^2 \geqslant 1 - 10\varepsilon\}$ for some i. It follows by Lemma 3.4 that

$$\left\|\tilde{\mathbb{E}}_{D(u)} u^{\otimes 2} - a_i^{\otimes 2}\right\|^2 \leqslant 2 - 2\left\langle \tilde{\mathbb{E}}_{D(u)} u^{\otimes 2}, a_i^{\otimes 2}\right\rangle = 2 - 2\tilde{\mathbb{E}}_{D(u)} \langle a_i, u\rangle^2 \leqslant 20\varepsilon\,. \qquad \square$$

5 Decomposition with sum-of-squares

In this section, we give a generic sum-of-squares algorithm (Algorithm 1 and Theorem 5.2) that will be used for various different settings in the following subsections (Section 5.2 for orthogonal tensors, Section 5.3 for tensors with separated components), and in the section 7 for random 3-tensor and Section 8 for robust FOOBI.

5.1 General algorithm for tensor decomposition

In this section, we provide a general sum-of-squares tensor decomposition that serve as the main building block for sections later. We will need the following lemma, which appears in [BKS15, Proof of Lemma 6.1].

Lemma 5.1. *Let $\varepsilon \in (0,1)$ and $\{a_1, \ldots, a_n\}$ be a set of unit vectors in $\mathbb{R}^d$ with $\|\sum_{i=1}^n a_i a_i^\mathsf{T}\| \leqslant 1+\varepsilon$. Then, for all even integers $k \in \mathbb{N}$, there exists a sum-of-squares proof that*

$$\left\{ \|u\|^2 \leqslant 1, \sum_{i=1}^n \langle a_i, u\rangle^4 \geqslant 1-\varepsilon \right\} \vdash_{u,k+2} \left\{ \sum_{i=1}^n \langle a_i, u\rangle^{k+2} \geqslant 1 - O(k\varepsilon) \right\}. \tag{5.1}$$

Our main algorithm below finds the solutions to a system of polynomial constraints $\mathcal{A}$, when given a "hint" in the form of a polynomial transformation of formal variables $P(\cdot)$. Roughly P should be an "orthogonalizing" map so that if $a_1, \ldots, a_n$ are the desired solutions to the constraints $\mathcal{A}$, then $P(a_1), \ldots, P(a_n)$ are nearly an orthonormal basis, or more precisely $\|\sum_{i=0}^n P(a_i)P(a_i)^\mathsf{T}\| \leqslant 1+\varepsilon$ while $\|P(a_i)\|^2 \geqslant 1-\varepsilon$ for all i. We then only require that a sum-of-squares proof exists certifying that the solutions to $\mathcal{A}$ after being mapped by P are actually close to $P(a_1), \ldots P(a_n)$; more precisely, that $\mathcal{A} \vdash_\ell \{\sum_{i=1}^n \langle P(a_i), P(u)\rangle^4 \geqslant 1-\varepsilon\}_u$ for some ℓ. The existence of this sum-of-squares certificate then allows us to recover the solutions $P(a_i)$ up to $O(\varepsilon)$ accuracy by solving for pseudo-distributions and then rounding them.

We later show how Algorithm 1 can be applied to a variety of tensor rank decomposition problems by the design of an appropriate orthogonalizing transform P. For example, in Section 7 $P(\cdot)$ orthogonalizes an overcomplete tensor by lifting the variables to a higher-dimensional space, and $P(\cdot)$ serves as a whitening transformation on a far-from-orthogonal tensor in Section 8.

The main technical difficulty in this analysis was in making the run-time polynomial (as opposed to quasi-polynomial in [BKS15]) for the nearly-orthogonal case where P is the identity transform.

Theorem 5.2. *For every $\ell \in \mathbb{N}$, there exists an $n^{O(\ell)}$-time algorithm (see Algorithm 1) with the following property: Let $\varepsilon > 0$ be smaller than some constant. Let $d, d' \in \mathbb{N}$ be numbers. Let $P\colon \mathbb{R}^d \to \mathbb{R}^{d'}$ be a polynomial with $\deg P \leqslant \ell$. Let $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}^d$ be a set of vectors such that $b_1 = P(a_1), \ldots, b_n = P(a_n) \in \mathbb{R}^{d'}$ all have norm at least $1-\varepsilon$ and $\|\sum_{i=1}^n b_i b_i^\mathsf{T}\| \leqslant 1+\varepsilon$. Let $\mathcal{A}$ be a system of polynomial inequalities in variables $u = (u_1, \ldots, u_d)$ such that the vectors $a_1, \ldots, a_n$ satisfy $\mathcal{A}$ and*

$$\mathcal{A} \vdash_{u,\ell} \left\{ \sum_{i=1}^n \langle b_i, P(u)\rangle^4 \geqslant (1-\varepsilon)\,\|P(u)\|^4 \right\}. \tag{5.2}$$

Then, the algorithm on input $\mathcal{A}$ and P outputs a set of unit vectors $\{b'_1, \ldots, b'_n\} \subseteq \mathbb{R}^{d'}$ such that

$$\operatorname{dist}_H\left(\{b_1^{\otimes 2}, \ldots, b_n^{\otimes 2}\}, \{(b'_1)^{\otimes 2}, \ldots, (b'_n)^{\otimes 2}\} \right) \leqslant O(\varepsilon)^{1/2}.$$

Proof of Theorem 5.2. We analyze Algorithm 1. By Corollary 4.5, if there exists a pseudo-distribution $D'(u)$ that satisfies constraints (5.5), then the top eigenvector of $\tilde{\mathbb{E}}_{D'(u)} P(u)P(u)^\mathsf{T}$ is $O(\varepsilon)^{1/2}$-close to one of the vectors $b_1, \ldots, b_n$. The fact that we add in step 4, the constraint $\{\langle P(u), b'_i\rangle \leqslant 0.1\}$

Algorithm 1 General tensor decomposition algorithm

Parameters: numbers $\varepsilon > 0$, $n, \ell \in \mathbb{N}$.
Given: system $\mathcal{A}$ of polynomial inequalities over $\mathbb{R}^d$ and polynomial $P\colon \mathbb{R}^d \to \mathbb{R}^{d'}$.
Find: vectors $b'_1, \dots, b'_n \in \mathbb{R}^{d'}$.
Algorithm:

- For i from 1 to n, do the following:

  1. Compute a degree-$(k+2)\ell$ pseudo-distribution $D(u)$ over $\mathbb{R}^d$, with $k = O(1)$, that satisfies the constraints

  $$\begin{aligned} &\mathcal{A} \cup \{1 + \varepsilon \geqslant \|P(u)\|^2 \geqslant 1 - \varepsilon\} \\ &\left\| \tilde{\mathbb{E}}_{D(u)} P(u)P(u)^\mathsf{T} \right\| \leqslant \frac{1+\varepsilon}{n-i+1} \,. \end{aligned} \tag{5.3}$$

  2. Choose standard Gaussian vectors $g^{(1)}, \dots, g^{(T)} \sim \mathcal{N}(0, \mathrm{Id}_{(d')^k})$ and $T = d^{O(1)}$ and compute the top eigenvectors of the following matrices for all $t \in [T]$:

  $$\tilde{\mathbb{E}}_{D(u)} \langle g^{(t)}, P(u)^{\otimes k} \rangle \cdot P(u)P(u)^\mathsf{T} \in \mathbb{R}^{d' \times d'} \,. \tag{5.4}$$

  3. Check if for one of the normalized top eigenvectors $b^\star$ computed in the previous step, there exists a degree-4ℓ pseudo-distribution $D'(u)$ that satisfies the constraints

  $$\mathcal{A} \cup \left\{1 + \varepsilon \geqslant \|P(u)\|^2 \geqslant 1 - \varepsilon, \langle b^\star, P(u) \rangle^2 \geqslant 0.99\right\}. \tag{5.5}$$

  4. Set b'_i to be the top eigenvector of the matrix $\tilde{\mathbb{E}}_{D'(u)} P(u)P(u)^\mathsf{T}$ and add to $\mathcal{A}$ the constraint $\{\langle P(u), b'_i \rangle^2 \leqslant 0.01\}$.

also implies by Corollary 4.5 that in some iteration i, we can never find a vector b'_i that is close to one vector b'_j from a previous iteration $j < i$. Therefore, it remains to show that in each of the n iterations with high probability we can find a pseudo-distribution $D'(u)$ that satisfies (5.5).

Consider a particular iteration $i_0 \in [n]$ of Algorithm 1. We may assume that the vectors $b'_1, \dots, b'_{i_0-1}$ are close to $b_1, \dots, b_{i_0-1}$. First we claim that there exists a pseudo-distribution satisfying conditions (5.3) in step 1, including the additional constraints added to $\mathcal{A}$ in previous iterations. Indeed, the uniform distribution over vectors $a_i, \dots, a_n$ satisfies all of those conditions. By assumption, we have a sum-of-squares proof $\mathcal{A} \vdash_{u,\ell} \{\sum_{i=1}^n \langle b_i, P(u)\rangle^4 \geqslant 1-\varepsilon\}$. Lemma 5.1 then implies $\mathcal{A} \vdash_{u,(k+2)\ell} \{\sum_{i=1}^n \langle b_i, P(u)\rangle^k \geqslant 1 - O(k\varepsilon)\}$ for an absolute constant parameter k to be determined later. Since $\mathcal{A}$ includes the added constraints $\{\langle b_1, P(u)\rangle^2 \leqslant 0.1, \dots, \langle b_{i_0-1}, P(u)\rangle^2 \leqslant 0.1\}$, it follows by $\left\|\sum_{i=1}^n b_i b_i^\mathsf{T}\right\|^2 \leqslant 1+O(\varepsilon)$ and substitution that $\mathcal{A} \vdash \{\sum_{i=1}^{i_0-1}\langle b_i, P(u)\rangle^k \leqslant (0.1)^{k-2}\cdot(1+O(\varepsilon))\}$, here choosing k so that $(0.1)^{k-2}\cdot(1+O(\varepsilon)) \leqslant 0.001$. Therefore, $\mathcal{A} \vdash_{(k+2)\ell} \{\sum_{i=i_0}^n \langle b_i, P(u)\rangle^k \geqslant 0.99\}$ and so $\tilde{\mathbb{E}}_{D(u)} \sum_{i=i_0}^n \langle b_i, P(u)\rangle^k \geqslant 0.99$ for any degree-$(k+2)\ell$ pseudo-distribution D that satisfies constraints (5.3). In particular, by averaging, there exists an index $i^\star \in \{i_0, \dots, n\}$ such that

$$\tilde{\mathbb{E}}_{D(u)} \langle b_{i^\star}, P(u)\rangle^k \geqslant \frac{0.99}{n - i_0 + 1} \geqslant 0.9 \cdot \left\| \tilde{\mathbb{E}}_{D(u)} P(u)P(u)^\mathsf{T} \right\| .$$

By Theorem 4.1, for each of the matrices (5.4) in step 2, its top eigenvector is 0.001-close to $b_{i^\star}$ with probability at least $d^{-O(1)}$. Therefore, we find at least one of those vectors with probability no smaller than $1 - d^{\Omega(1)}$. In this case, a pseudo-distribution $D'(u)$ as required in step 3 exists, as an atomic distribution supported only on $b_{i^\star}$ is an example that satisfies the conditions. □

5.2 Tensors with orthogonal components

We apply Theorem 5.2 to orthogonal tensors with noise.

Theorem (Restatement of Theorem 1.1)**.** *There exists a polynomial-time algorithm that given a symmetric 3-tensor $T \in (\mathbb{R}^d)^{\otimes 3}$ outputs a set of vectors $\{a'_1, \dots, a'_{n'}\} \subseteq \mathbb{R}^d$ such that for every orthonormal set $\{a_1, \dots, a_n\} \subseteq \mathbb{R}^d$, the Hausdorff distance between the two sets is at most*

$$\operatorname{dist}_H\left(\{a_1, \dots, a_n\}, \{a'_1, \dots, a'_{n'}\}\right)^2 \leqslant O(1) \cdot \left\| T - \sum\nolimits_{i=1}^n a_i^{\otimes 3} \right\|_{\{1\},\{2,3\}} . \tag{5.6}$$

Proof. We feed Algorithm 1 with the inputs $P(u) = u$ and $\mathcal{A} = \left\{\langle T, u^{\otimes 3}\rangle \geqslant 1-\varepsilon\right\}$ where $\varepsilon = \|E\|_{\{1\},\{2,3\}}$ and $E = T - \sum_i a_i^{\otimes 3}$. We have

$$\begin{aligned} \mathcal{A} \vdash_4 \sum_{i=1}^n \langle a_i, u\rangle^3 &= \langle T, u^{\otimes 3}\rangle - \langle E, u^{\otimes 3}\rangle \\ &= \langle T, u^{\otimes 3}\rangle - \varepsilon \\ &\geqslant 1 - 2\varepsilon . \end{aligned}$$

Here at the second line we used that

$$\vdash \langle E, u^{\otimes 3}\rangle \leqslant \|E\|_{\{1\},\{2,3\}} \leqslant \varepsilon . \tag{5.7}$$

We verify that $\mathcal{A}$ satisfies the requirement (5.2),

$$\begin{aligned}
\mathcal{A} \vdash \sum_{i=1}^{n} \langle a_i, u\rangle^4 &\geqslant \left(\sum_{i=1}^{n} \langle a_i, u\rangle^4\right)\left(\sum_{i=1}^{n} \langle a_i, u\rangle^2\right) && \text{(using orthonormality)} \\
&\geqslant \left(\sum_{i=1}^{n} \langle a_i, u\rangle^3\right)^2 && \text{(Cauchy-Schwarz: Lemma A.1)} \\
&\geqslant 1 - 4\varepsilon\,.
\end{aligned}$$

Therefore calling Algorithm 1, we can recover $\hat{a}_i$ which is, up to sign flip, close to a_i with error $O(\varepsilon^{1/2})$. We determine the sign by finding the $\tau \in \{-1, +1\}$ such that $\langle T, \tau \hat{a}_i^{\otimes 3}\rangle \geqslant 1 - \varepsilon$ and set the output a_i' to $\tau \hat{a}_i$. □

Remark 5.3. Note that in the proof of Theorem 1.1, the conclusion of equation (5.7) is the only thing we used about the error term E. Therefore, define the following SoS relaxation of the injective norm:

$$\|E\|_{\mathrm{SoS}} = \inf_{c \in \mathbb{R}} \left[\{\|u\|^2 \leqslant 1\} \vdash \{\langle E, u^{\otimes 3}\rangle \leqslant c\} \right] .$$

Then we can replace the right hand side of equation (5.6) by $O(1) \cdot \left\| T - \sum_{i=1}^{n} a_i^{\otimes 3} \right\|_{\mathrm{SoS}}$.

5.3 Tensors with separated components

The following lemma shows that for separated vectors the sum of higher-order outer products has spectral norm that decrease exponentially with the tensor order.

Lemma 5.4. *Let $a_1, \ldots, a_n$ be unit vectors in $\mathbb{R}^d$. Then, for every $k \in \mathbb{N}$,*

$$\left\| \sum_{i=1}^{n} \left(a_i a_i^{\mathsf{T}}\right)^{\otimes (k+1)} \right\| \leqslant 1 + \left(\max_{i \neq j} |\langle a_i, a_j\rangle|\right)^k \cdot \left\| \sum_{i=1}^{n} a_i a_i^{\mathsf{T}} \right\| .$$

Proof. Let $A = \sum_i \left(a_i a_i^{\mathsf{T}}\right)^{\otimes k+1}$. For a unit vector $x \in (\mathbb{R}^d)^{\otimes k+1}$ we'll bound the quadratic form $x^{\mathsf{T}} A x$.

First, without loss of generality we can assume that x is in the subspace V spanned by $\{a_i^{\otimes k+1}\}_i$. This is because if x had a component y orthogonal to V, then $y^{\mathsf{T}} a_i^{\otimes k+1} = 0$ for all $i \in [n]$ by definition, so that $Ay = 0$ and y can make no nonzero contribution to the quadratic form above.

Also let $W = (A^{1/2})^+$ so that W is a whitening transform and WAW is a projector onto V. Then suppose $x = \sum_i c_i W a_i^{\otimes k+1}$, so that $\sum_i c_i^2 = \|x\|^2 = 1$. Then

$$\begin{aligned}
x^{\mathsf{T}} A x &= \sum_{ij} c_i c_j (a_i^{\otimes k+1})^{\mathsf{T}} W A W a_j^{\otimes k+1} \\
&= \sum_i c_i^2 + \sum_{i \neq j} c_i c_j \langle a_i, a_j\rangle^{k+1} \\
&\leqslant \sum_i c_i^2 + \left(\max_{i \neq j} |\langle a_i, a_j\rangle|\right)^k \sum_{i \neq j} c_i c_j \langle a_i, a_j\rangle
\end{aligned}$$

$$\leqslant 1 + \left(\max_{i\neq j} |\langle a_i, a_j\rangle|\right)^k \left\|\sum_{i=1}^n a_i a_i^\mathsf{T}\right\|,$$

where in the last step we let $A' = \sum_i a_i a_i^\mathsf{T}$ and $W' = (A'^{1/2})^+$, and apply the inequality $\sum_{i,j} c_i c_j \langle a_i, a_j\rangle = \sum_{ij} c_i c_j a_i^\mathsf{T} W' A' W' a_j = x'^\mathsf{T} A' x' \leqslant \|A'\|$, where $x' = \sum_i c_i W' a_i$ is a unit vector. □

Lemma 5.5. *Let $a \in \mathbb{R}^d$ and $b \in (\mathbb{R}^d)^{\otimes k}$ be unit vectors such that $\langle a^{\otimes k}, b\rangle^2 \geqslant 1 - \varepsilon$. Let B be the reshaping of the vector b into a d-by-d^{k-1} matrix. Then the top left singular vector $a' \in \mathbb{R}^d$ of B satisfies $\langle a', a\rangle^2 \geqslant 1 - O(\varepsilon)$.*

Proof. Let c be the top right singular vector of B. Then, $\langle a' \otimes c, b\rangle \geqslant \langle a^{\otimes k}, b\rangle \geqslant 1 - \varepsilon$. Therefore, $\|a' \otimes c - b\| \leqslant O(\varepsilon)^{1/2}$. By triangle inequality, $\|a' \otimes c - a \otimes a^{\otimes k-1}\| \leqslant O(\varepsilon)^{1/2}$, which means that as desired $|\langle a, a'\rangle| \geqslant \langle a' \otimes c, a \otimes a^{\otimes(k-1)}\rangle \geqslant 1 - O(\varepsilon)$. □

Theorem (Restatement of Theorem 1.5)**.** *There exists an algorithm A with polynomial running time (in the size of its input) such that for all $\eta, \rho \in (0,1)$ and $\sigma \geqslant 1$, for every set of unit vectors $\{a_1, \dots, a_n\} \subseteq \mathbb{R}^d$ with $\|\sum_{i=1}^n a_i a_i^\mathsf{T}\| \leqslant \sigma$ and $\max_{i\neq j} |\langle a_i, a_j\rangle| \leqslant \rho$, when the algorithm is given a symmetric k-tensor $T \in (\mathbb{R}^d)^{\otimes k}$ with $k \geqslant O\left(\frac{1+\log\sigma}{\log\rho}\right) \cdot \log(1/\eta)$, then its output $A(T)$ is a set of vectors $\{a'_1, \dots, a'_{n'}\} \subseteq \mathbb{R}^d$ such that*

$$\operatorname{dist}_H\left(\{a_1'^{\otimes 2}, \dots, a_n'^{\otimes 2}\}, \{a_1^{\otimes 2}, \dots, a_n^{\otimes 2}\}\right)^2 \leqslant O\left(\eta + \left\|T - \sum\nolimits_{i=1}^n a_i^{\otimes k}\right\|_{\{1,\dots,\lfloor k/2\rfloor\},\{\lfloor k/2\rfloor+1,\dots,k\}}\right). \quad (5.8)$$

Proof. We use Algorithm 1 from Theorem 5.2. Let $E = T - \sum_i a_i^{\otimes k}$. We may assume that $\|E\|_{\{1,\dots,\lfloor k/2\rfloor\},\{\lfloor k/2\rfloor+1,\dots,k\}} \leqslant \eta$, since otherwise the theorem follows from the case when $\eta = \|E\|_{\{1,\dots,\lfloor k/2\rfloor\},\{\lfloor k/2\rfloor+1,\dots,k\}}$. Let P be the polynomial map $P(x) = x^{\otimes\lceil k/4\rceil}$ and let $\mathcal{A}$ be the system of polynomial inequalities

$$\mathcal{A} = \{\langle T, u^{\otimes k}\rangle \geqslant 1 - \eta, \|u\|^2 = 1\}. \quad (5.9)$$

in variables $u = (u_1, \dots, u_d)$. Since $\|E\|_{\{1,\dots,\lfloor k/2\rfloor\},\{\lfloor k/2\rfloor+1,\dots,k\}} \leqslant \eta$, all of the vectors $a_1, \dots, a_n$ satisfy $\mathcal{A}$. Let $b_1, \dots, b_n$ be the unit vectors $b_i = P(a_i)$. By Lemma 5.4 and the condition on k, these vectors satisfy $\|\sum_i b_i b_i^\mathsf{T}\| \leqslant 1 + \rho^{\lceil k/4\rceil}\sigma \leqslant 1 + \eta$. Then, we have the following sum-of-squares proof

$$\mathcal{A} \vdash_{u,k} \sum_{i=1}^n \langle b_i, P(u)\rangle^4 = \sum_{i=1}^n \langle a_i, u\rangle^{4\lceil k/4\rceil} \geqslant \sum_{i=1}^n \langle a_i, u\rangle^k = \langle T, u^{\otimes k}\rangle - \langle E, u^{\otimes k}\rangle \quad (5.10)$$

$$\geqslant 1 - \eta - \|T\|_{\{1,\dots,\lfloor k/2\rfloor\},\{\lfloor k/2\rfloor+1,\dots,k\}} \geqslant 1 - 2\eta. \quad (5.11)$$

It follows that $\mathcal{A}$ and P satisfy the conditions of Theorem 5.2. Thus, Algorithm 1 on input $\mathcal{A}$ and P recovers vectors $b'_1, \dots, b'_n$ with Hausdorff distance at most $O(\sqrt{\eta})$ from $b_1, \dots, b_n$. By Lemma 5.5, the top left singular vectors of the d-by-$d^{\lceil k/4\rceil-1}$ matrix reshapenings of $b'_1, \dots, b'_n$ are $O(\sqrt{\eta})$-close to the vectors $a_1, \dots, a_n$ up to sign. (If k is odd, then we may determine the signs of the a_i by checking if $\langle T, a_i'^{\otimes k}\rangle \geqslant 1 - O(\eta)$ or $\langle T, a_i'^{\otimes k}\rangle \leqslant -1 + O(\eta)$ for each output vector a'_i.) □

6 Spectral norms and tensor operations

In this section, we provide several bounds regarding the spectral norms of moments of the lifted vectors, and the spectral norm of random contraction of a tensor, which are crucial in our analysis in previous sections. We suggest readers who are more interested in applications of the algorithms jump to Section 7 and 8.

6.1 Spectral norms and pseudo-distributions

Theorem 6.1. *Let D be a degree-$4(p+q)$ pseudo-distribution over $\mathbb{R}^d$ that satisfies $\{\|u\|^2 \leqslant 1\}_{D(u)}$. Then, for all $p, q \in \mathbb{N}$,*

$$\left\| \tilde{\mathbb{E}}_{D(u)} u^{\otimes p} \left(u^{\otimes q}\right)^{\mathsf{T}} \right\| \leqslant \left\| \tilde{\mathbb{E}}_{D(u)} u u^{\mathsf{T}} \right\| . \tag{6.1}$$

The theorem follows by combining Lemma 6.2 and Lemma 6.4 proved below. Lemma 6.2 reduces Theorem 6.1 to the case when $p = q$.

Lemma 6.2. *Let D be a degree-$4(p+q)$ pseudo-distribution over $\mathbb{R}^d$ that satisfies $\{\|u\|^2 \leqslant 1\}_{D(u)}$. Then, for all $p, q \in \mathbb{N}$,*

$$\left\| \tilde{\mathbb{E}}_{D(u)} u^{\otimes p} \left(u^{\otimes q}\right)^{\mathsf{T}} \right\|^2 \leqslant \left\| \tilde{\mathbb{E}}_{D(u)} \left(u^{\otimes p}\right) \left(u^{\otimes p}\right)^{\mathsf{T}} \right\| \cdot \left\| \tilde{\mathbb{E}}_{D(u)} \left(u^{\otimes q}\right) \left(u^{\otimes q}\right)^{\mathsf{T}} \right\| . \tag{6.2}$$

Proof. For all unit vectors $x \in (\mathbb{R}^d)^{\otimes p}$ and $y \in (\mathbb{R}^d)^{\otimes q}$

$$\begin{aligned}
\langle x, \left(\tilde{\mathbb{E}}_{D(u)} u^{\otimes p} \left(u^{\otimes q}\right)^{\mathsf{T}} \right) y \rangle &= \tilde{\mathbb{E}}_{D(u)} \langle x, u^{\otimes p} \rangle \langle u^{\otimes q}, y \rangle \\
&\leqslant \left(\tilde{\mathbb{E}}_{D(u)} \langle x, u^{\otimes p} \rangle^2 \right)^{1/2} \cdot \left(\tilde{\mathbb{E}}_{D(u)} \langle u^{\otimes q}, y \rangle^2 \right)^{1/2} \\
&\qquad \text{(Cauchy–Schwarz for pseudo-expectations)} \\
&\leqslant \left\| \tilde{\mathbb{E}}_{D(u)} \left(u^{\otimes p}\right) \left(u^{\otimes p}\right)^{\mathsf{T}} \right\|^{1/2} \cdot \left\| \tilde{\mathbb{E}}_{D(u)} \left(u^{\otimes q}\right) \left(u^{\otimes q}\right)^{\mathsf{T}} \right\|^{1/2} .
\end{aligned} \tag{6.3}$$

The lemma follows from this bound by choosing x and y as the top left and right singular vectors of the matrix $\tilde{\mathbb{E}}_{D(u)} u^{\otimes p} (u^{\otimes q})^{\mathsf{T}}$. □

Towards proving Theorem 6.1 for the case of $p = q$, we first establish the following lemma which says that tensoring with vector with norm less 1 won't increase the spectral norm.

Lemma 6.3. *Let $g(u, v)$ be a polynomial in indeterminates u, v. Let D be a degree-4 pseudo-distribution over $\mathbb{R}^d$ that satisfies $\{\|u\|^2 \leqslant 1, g(u, v) \geqslant 0\}_{D(u,v)}$. Then, for all $p \in \mathbb{N}$,*

$$\left\| \tilde{\mathbb{E}}_{D(u,v)} g(u, v) (u \otimes v) (u \otimes v)^{\mathsf{T}} \right\| \leqslant \left\| \tilde{\mathbb{E}}_{D(v)} g(u, v) v v^{\mathsf{T}} \right\| . \tag{6.4}$$

Proof. We have the sum-of-squares proof that

$$\begin{aligned}
&\vdash g(u,v)\left(\mathrm{Id}\otimes vv^\top - (u\otimes v)(u\otimes v)^\mathsf{T}\right)\\
&= g(u,v)\left(\mathrm{Id} - uu^\top\right)\otimes vv^\mathsf{T}\\
&= g(u,v)\left(1-\|u\|^2\right)\mathrm{Id}\otimes vv^\top + g(u,v)(\|u\|^2\,\mathrm{Id} - uu^\top)\otimes vv^\top\\
&\geq 0 \qquad\qquad \text{(by } 1-\|u\|^2 \geqslant 0 \text{ and } \vdash \|u\|^2\,\mathrm{Id} - uu^\top \succeq 0 \text{ (Lemma 3.10))}
\end{aligned}$$

Therefore, we obtain that

$$\tilde{\mathbb{E}}_{D(u,v)}[g(u,v)\,\mathrm{Id}\otimes vv^\top] - \tilde{\mathbb{E}}_{D(u,v)}\left[g(u,v)\,(u\otimes v)(u\otimes v)^\mathsf{T}\right] \succeq 0$$

The desired inequality follows,

$$\left\|\tilde{\mathbb{E}}_{D(u,v)} g(u,v)\,(u\otimes v)(u\otimes v)^\mathsf{T}\right\| \leqslant \left\|\tilde{\mathbb{E}}_{D(u,v)} \mathrm{Id}\otimes g(u,v)vv^\top\right\| = \left\|\tilde{\mathbb{E}}_{D(v)} g(u,v)vv^\mathsf{T}\right\|$$

□

The following statement follows straightforward from the Lemma 6.3 by induction on p.

Lemma 6.4. *Let D be a degree-$4p$ pseudo-distribution over $\mathbb{R}^d$ that satisfies $\{\|u\|^2 \leqslant 1\}_{D(u)}$. Then, for all $p \in \mathbb{N}$,*

$$\left\|\tilde{\mathbb{E}}_{D(u)}\left(uu^\mathsf{T}\right)^{\otimes p}\right\| \leqslant \left\|\tilde{\mathbb{E}}_{D(u)} uu^\mathsf{T}\right\|^p. \tag{6.5}$$

6.2 Spectral norm of random contraction

The following theorem shows that a random contraction of a 3-tensor has spectral norm at most $O(\sqrt{\log d})$ factor larger than the spectral norm of its matrix unfoldings.

Theorem 6.5. *Let $T \in \mathbb{R}^p\otimes\mathbb{R}^q\otimes\mathbb{R}^r$ be an order-3 tensor. Let $g \in \mathcal{N}(0,\mathrm{Id}_r)$. Then for any $t \geqslant 0$,*

$$\mathbb{P}_g\left\{\left\|\left(\mathrm{Id}\otimes\mathrm{Id}\otimes g^\mathsf{T}\right)T\right\|_{\{1\},\{2\}} \geqslant t\cdot\max\left\{\|T\|_{\{1\},\{2,3\}}, \|T\|_{\{2\},\{1,3\}}\right\}\right\} \leqslant 2(p+q)\cdot e^{-t^2/2}, \tag{6.6}$$

and consequently,[8]

$$\mathbb{E}_g\left[\left\|\left(\mathrm{Id}\otimes\mathrm{Id}\otimes g^\mathsf{T}\right)T\right\|_{\{1\},\{2\}}\right] \leqslant O(\log(p+q))^{1/2}\cdot\max\left\{\|T\|_{\{1\},\{2,3\}}, \|T\|_{\{2\},\{1,3\}}\right\}. \tag{6.7}$$

Proof. Let T_i denote the ith third-mode slice of T so that $T_i = \left(\mathrm{Id}\otimes\mathrm{Id}\otimes e_i^\mathsf{T}\right)T$ reshaped as a p-by-q matrix. Note that when regarded as a p-by-q matrix, the contraction $\left(\mathrm{Id}\otimes\mathrm{Id}\otimes g^\mathsf{T}\right)T$ is a Gaussian matrix series with coefficients $T_1,\dots,T_r$, so that

$$\left\|\left(\mathrm{Id}\otimes\mathrm{Id}\otimes g^\mathsf{T}\right)T\right\|_{\{1\},\{2\}} = \left\|\sum\nolimits_{i=1}^r g_iT_i\right\|,$$

[8] For large enough p and q, the constant hidden in the big-Oh notation below is at most 2

where $g_1, \dots, g_r$ are independent standard Gaussians with $g_i = \langle g, e_i \rangle$. Therefore, by concentration of Gaussian matrix series [Oli10, Theorem 1] (also see [Tro12, Corollary 4.2]), we have

$$\mathbb{P}\left\{\|\left(\mathrm{Id} \otimes \mathrm{Id} \otimes g^\top\right) T\| \geqslant t\sigma\right\} \leqslant 2(p+q)e^{-t^2/2}\,,$$

where $\sigma = \max\left\{\left\|\sum_i T_i T_i^\top\right\|, \left\|\sum_i T_i^\top T_i\right\|\right\}^{1/2}$.

For U and V sets of indices, let $T_{U,V}$ denote the matrix unfolding of T with rows indexed by U and columns indexed by V, so that $\|T\|_{U,V} = \|T_{U,V}\|$. We claim that $\sum_i T_i T_i^\top = (T_{\{1\},\{2,3\}})^\top (T_{\{1\},\{2,3\}})$ and $\sum_i T_i^\top T_i = (T_{\{2\},\{1,3\}})^\top (T_{\{2\},\{1,3\}})$, which completes the proof. These identities are forced by the observations that both of these objects are matrix quantities that are quadratic in T, with the first object being a sum over the 2nd and 3rd indices of the two copies of T, and the second object being a sum over the 1st and 3rd indices. □

The following corollary of Theorem 6.5 handles a larger class of random contractions.

Corollary 6.6. *Let $T \in \mathbb{R}^p \otimes \mathbb{R}^q \otimes \mathbb{R}^r$ be an order-3 tensor. Let $g \sim \mathcal{N}(0, \Sigma)$ with covariance matrix Σ satisfying $0 \preceq \Sigma \preceq \mathrm{Id}_r$. Then for any $t \geqslant 0$,*

$$\mathop{\mathbb{P}}_{g}\left\{\left\|\left(\mathrm{Id} \otimes \mathrm{Id} \otimes g^\top\right) T\right\|_{\{1\},\{2\}} \geqslant t \cdot \max\left\{\|T\|_{\{1\},\{2,3\}}, \|T\|_{\{2\},\{1,3\}}\right\}\right\} \leqslant 4(p+q) \cdot e^{-t^2/2}\,. \tag{6.8}$$

Proof. We reduce to the case $\Sigma = \mathrm{Id}_p$ and apply Theorem 6.5. Concretely, let $g' = g + h$ and $g'' = g - h$ where h is a random variable with distribution $\mathcal{N}(0, \mathrm{Id}_p - \Sigma)$ that is independent of g. By this construction, g' and g'' both have marginal distribution $\mathcal{N}(0, \mathrm{Id}_p)$, and $g = \frac{1}{2}(g' + g'')$. Therefore we can invoke Theorem 6.5 for random variables g' and g''. Letting $\sigma = \max\{\|T\|_{\{1\},\{2,3\}}, \|T\|_{\{2\},\{1,3\}}\}$, using the union bound and the triangle inequality, we have that

$$\begin{aligned}
&\mathop{\mathbb{P}}_{g}\left\{\left\|\left(\mathrm{Id} \otimes \mathrm{Id} \otimes g^\top\right) T\right\|_{\{1\},\{2\}} \geqslant t\sigma\right\} \\
&= \mathop{\mathbb{P}}_{g,h}\left\{\left\|\left(\mathrm{Id} \otimes \mathrm{Id} \otimes (g' + g'')^\top\right) T\right\|_{\{1\},\{2\}} \geqslant 2t\sigma\right\} \\
&\leqslant \mathop{\mathbb{P}}_{g,h}\left\{\left\|\left(\mathrm{Id} \otimes \mathrm{Id} \otimes (g')^\top\right) T\right\|_{\{1\},\{2\}} + \left\|\left(\mathrm{Id} \otimes \mathrm{Id} \otimes (g'')^\top\right) T\right\|_{\{1\},\{2\}} \geqslant 2t\sigma\right\} \\
&\leqslant \mathop{\mathbb{P}}_{g,h}\left\{\left\|\left(\mathrm{Id} \otimes \mathrm{Id} \otimes (g')^\top\right) T\right\|_{\{1\},\{2\}} \geqslant t\sigma\right\} + \mathop{\mathbb{P}}_{g,h}\left\{\left\|\left(\mathrm{Id} \otimes \mathrm{Id} \otimes (g'')^\top\right) T\right\|_{\{1\},\{2\}} \geqslant t\sigma\right\} \\
&\leqslant 4(p+q) \cdot e^{-t^2/2}\,,
\end{aligned}$$

where the second line uses the triangle inequality, the third line uses the union bound, and the fourth line uses Theorem 6.5 applied to g' and g''. □

Corollary 6.6 and Theorem 6.1 together imply the following theorem..

Theorem 6.7. *Let $k \in \mathbb{N}$ and D be a degree-$(4k+10)$ pseudo-distribution over $\mathbb{R}^d$ that satisfies $\{\|u\|^2 \leqslant 1\}_{D(u)}$. Let $g \sim \mathcal{N}(0, \Sigma)$ be a Gaussian vector with covariance $\Sigma \preceq \mathrm{Id}_d^{\otimes k}$. Then,*

$$\mathop{\mathbb{E}}_{g \sim \mathcal{N}(0, \mathrm{Id}_d)}\left\|\tilde{\mathbb{E}}_{D(u)} \langle g, u^{\otimes k}\rangle \cdot uu^\top\right\| \lesssim \sqrt{k \log d}\left\|\tilde{\mathbb{E}}_{D(u)} uu^\top\right\|\,. \tag{6.9}$$

We can apply Corollary 6.6 repeatedly to obtain a bound for random contraction over a larger number of modes.

Theorem 6.8. *Let $T \in \mathbb{R}^{p\times q\times r_1\times\cdots\times r_s}$ be an order-$(s+2)$ tensor, and $g_1 \sim \mathcal{N}(0,\Sigma_1),\ldots,g_s \sim \mathcal{N}(0,\Sigma_s)$ be independent Gaussian random variables with covariance $\Sigma_i \preceq \mathrm{Id}_{r_i}$ for each $i \in [r]$. Let $\bar{r} = \max_{i\in[s]}\{r_i+2\}$. Then for any $t \geqslant 0$,*

$$\mathop{\mathbb{P}}_{g}\left\{\left\|\left(\mathrm{Id}\otimes\mathrm{Id}\otimes {g_1}^{\mathsf{T}}\otimes\ldots\otimes {g_s}^{\mathsf{T}}\right)T\right\|_{\{1\},\{2\}} \geqslant t^s\cdot\max_{S\subset[s]:1\in S,2\notin S}\left\{\|T\|_{S,S^c}\right\}\right\} \leqslant 4(p+q)\,\bar{r}^{s-1}\,e^{-t^2/2}\,. \quad (6.10)$$

Proof. We prove by induction on s. The base case is exactly Corollary 6.6. For $s \geqslant 2$, suppose we have proved the $(s-1)$-case.

Let $T' = \left(\mathrm{Id}\otimes\mathrm{Id}\otimes\mathrm{Id}\otimes {g_2}^{\mathsf{T}}\cdots\otimes {g_s}^{\mathsf{T}}\right)T$ be an order-3 tensor. Then we have that

$$\left(\mathrm{Id}\otimes\mathrm{Id}\otimes {g_1}^{\mathsf{T}}\otimes\ldots\otimes {g_s}^{\mathsf{T}}\right)T = \left(\mathrm{Id}\otimes\mathrm{Id}\otimes {g_1}^{\mathsf{T}}\right)T'\,.$$

Then using Corollary 6.6 on T' and g_1, and then taking the expectation over $g_2,\ldots,g_s$, we have

$$\mathop{\mathbb{P}}_{g_1,\ldots,g_s}\left\{\left\|\left(\mathrm{Id}\otimes\mathrm{Id}\otimes {g_1}^{\mathsf{T}}\otimes\ldots\otimes {g_s}^{\mathsf{T}}\right)T\right\|_{\{1\},\{2\}} \geqslant t\cdot\max\left\{\|T'\|_{\{1\},\{2,3\}},\|T'\|_{\{2\},\{1,3\}}\right\}\right\} \leqslant 4(p+q)\,e^{-t^2/2}\,. \quad (6.11)$$

We view T' as an order-$(s+1)$ tensor by merging the 2nd and 3rd modes, that is, $\left(\mathrm{Id}\otimes(\mathrm{Id}\otimes\mathrm{Id})\otimes {g_2}^{\mathsf{T}}\cdots\otimes {g_s}^{\mathsf{T}}\right)T$, and then apply the inductive hypothesis. We obtain

$$\mathop{\mathbb{P}}_{g_2,\ldots,g_s}\left\{\|T'\|_{\{1\},\{2,3\}} \geqslant t^{s-1}\cdot\max_{S\subset[s]:1\in S,\{2,3\}\cap S=\emptyset}\left\{\|T\|_{S,S^c}\right\}\right\} \leqslant 4(p+qr_1)\,\bar{r}^{s-2}\cdot e^{-t^2/2}\,. \quad (6.12)$$

Similarly, we have that

$$\mathop{\mathbb{P}}_{g_2,\ldots,g_s}\left\{\|T'\|_{\{1,3\},\{2\}} \geqslant t^{s-1}\cdot\max_{S\subset[s]:\{1,3\}\in S,2\notin S=\emptyset}\left\{\|T\|_{S,S^c}\right\}\right\} \leqslant 4(pr_1+q)\,\bar{r}^{s-2}\cdot e^{-t^2/2}\,. \quad (6.13)$$

Using equations (6.11), (6.12), (6.13), and applying union bound we obtain

$$\mathop{\mathbb{P}}_{g}\left\{\left\|\left(\mathrm{Id}\otimes\mathrm{Id}\otimes {g_1}^{\mathsf{T}}\otimes\cdots\otimes {g_s}^{\mathsf{T}}\right)T\right\|_{\{1\},\{2\}} \geqslant t^s\cdot\max_{S\subset[s]:1\in S,2\notin S}\left\{\|T\|_{S,S^c}\right\}\right\} \leqslant 4(p+q)\,\bar{r}^{s-1}\,e^{-t^2/2}\,,$$

and complete the inductive proof.

□

7 Decomposition of random overcomplete 3-tensors

In this section, we assume that we are given a random 3rd order overcomplete symmetric tensor T of the following form

$$T = \sum\nolimits_{i=1}^{n} a_i^{\otimes 3} + E\,, \quad (7.1)$$

where $n \leqslant d^{1.5}/(\log d)^{O(1)}$, the vectors a_i are drawn independently at random from the Euclidean unit sphere, and the error tensor E satisfies $\|E\|_{\{1\},\{2,3\}} \leqslant \varepsilon$.

Let $\mathrm{Id}_{\mathrm{sym}}$ be the projection to the symmetric subspace of $(\mathbb{R}^d)^{\otimes 2}$ (the span of all $x^{\otimes 2}$ for $x \in \mathbb{R}^d$), and let $\Phi = \frac{1}{\sqrt{d}} \sum_{i=1}^d e_i^{\otimes 2} \in (\mathbb{R}^d)^{\otimes 2}$. Let $\mathrm{Id}_{\mathrm{sym}'}$ be the projection to the subspace orthogonal to Φ:

$$\mathrm{Id}_{\mathrm{sym}'} = \mathrm{Id}_{\mathrm{sym}} - \Phi\Phi^{\mathsf{T}} . \tag{7.2}$$

Algorithm 2 Polynomial-time algorithm for random overcomplete 3-tensor decomposition

Input: Number $\varepsilon > 0$ and $n \in \mathbb{N}$ and symmetric tensor $T \in (\mathbb{R}^d)^{\otimes 3}$ of the form (7.1).
Find: $\hat{a}_1, \ldots, \hat{a}_n \in \mathbb{R}^d$.
Algorithm:

1. Call Algorithm 1 with

$$\mathcal{A} = \left\{ \langle T, u^{\otimes 3} \rangle \geqslant 1 - \varepsilon, \|u\|^2 = 1 \right\} , \tag{7.3}$$

$$P(u) = \mathrm{Id}_{\mathrm{sym}'}\, u^{\otimes 2} , \tag{7.4}$$

   where $\mathrm{Id}_{\mathrm{sym}'}$ is defined in (7.2). Suppose the outputs of Algorithm 1 are $\hat{b}_1, \ldots, \hat{b}_n$.

2. Let $\hat{a}_i$ be τ_i the top eigenvector of the matrix reshaping of $\hat{b}_i$, where $\tau_i \in \{1, -1\}$ is chosen so that $T\hat{a}_1 > 0$.

Theorem (Restatement of Theorem 1.2)**.** *With probability $1 - d^{-\omega(1)}$ over the choice of random unit vectors $a_1, \ldots, a_n \in \mathbb{R}^d$, when given a symmetric 3-tensor $T \in (\mathbb{R}^d)^{\otimes 3}$ as input, the output $\hat{a}_1, \ldots, \hat{a}_n \in \mathbb{R}^d$ of Algorithm 2 satisfies*

$$\mathrm{dist}_H \left(\{\hat{a}_1, \ldots, \hat{a}_n\}, \{a_1, \ldots, a_n\} \right)^2 \leqslant O\left(\left(\frac{n}{d^{1.5}} \right)^{\Omega(1)} + \left\| T - \sum\nolimits_{i=1}^n a_i^{\otimes 3} \right\|_{\{1\},\{2,3\}} \right) . \tag{7.5}$$

Theorem 1.2 follows immediately from Theorem 5.2 and the following proposition:

Proposition 7.1. *With probability $1 - d^{-\omega(1)}$ over the choice of random unit vectors $a_1, \ldots, a_n$, the parameters $P(\cdot)$ and $\mathcal{A}$ defined in Algorithm 2 satisfy the requirements of Theorem 5.2. In particular, let $c_i = P(a_i) = \mathrm{Id}_{\mathrm{sym}'}\, a_i^{\otimes 2}$. Then*

$$\left\| \sum_{i=1}^n c_i c_i^{\mathsf{T}} \right\| \leqslant 1 + \delta , \tag{7.6}$$

where $\delta = \tilde{O}(\sqrt{n}/d + n/d^{1.5})$, and

$$\mathcal{A} \vdash \sum_{i=1}^n \langle c_i, P(u) \rangle^4 \geqslant \left(1 - O(\varepsilon + 1/d)\right) \|P(u)\|^4 . \tag{7.7}$$

We first show that by a simple extension of [GM15, Theorem 4.2 and Lemma 8], $\mathcal{A}$ implies that the sum of the terms $\langle a_i, u \rangle^8$ is large. Note that $c_i \approx a_i^{\otimes 2}$ and $P(u) \approx u^{\otimes 2}$, and therefore this is already fairly close to our target inequality (7.7).

Lemma 7.2 (Simple extension of [GM15, Theorem 4.2 and Lemma 8]). *With probability* $1 - d^{-\omega(1)}$ *over the choice of random unit vectors* a_i*,*

$$\mathcal{A} \vdash \left\{ \sum_{i=1}^{n} \langle a_i, u \rangle^3 \geqslant 1 - \varepsilon, \|u\|^2 = 1 \right\} \vdash \left\{ \sum_{i=1}^{n} \langle a_i, u \rangle^8 \geqslant 1 - O(\varepsilon) - \delta \right\} . \tag{7.8}$$

where $\delta = \widetilde{O}(n/d^{3/2})$.

Proof. Using the proof of [GM15, Theorem 4.2] (specifically Lemma 3 and Claim 1), and the proof of Lemma 5 (specifically equation (11) and equation (15)) we have[9]

$$\mathcal{A} \vdash \left\{ \sum_{i=1}^{n} \langle u, a_i \rangle^4 \geqslant 1 - O(\varepsilon) - \delta, \text{ and } \sum_{i=1}^{n} \langle u, a_i \rangle^6 \geqslant 1 - O(\varepsilon) - \delta \right\} . \tag{7.9}$$

where $\delta = O(n \log^{O(1)} d/d^{3/2})$. Then we extend the proof using the same idea to higher powers:

$$\begin{aligned} \vdash \left(\sum_{i=1}^{n} \langle u, a_i \rangle^6 \right)^2 &= \left\langle \left[\sum_{i=1}^{n} \langle u, a_i \rangle^5 a_i \right], u \right\rangle^2 \leqslant \left\| \sum_{i=1}^{n} \langle u, a_i \rangle^5 a_i \right\|^2 \\ &= \sum_{i=1}^{n} \langle u, a_i \rangle^{10} + \sum_{i \neq j} \langle u, a_i \rangle^5 \langle u, a_j \rangle^5 \langle a_i, a_j \rangle \\ &\leqslant \sum_{i=1}^{n} \langle u, a_i \rangle^{10} + \left(\sum_{i=1}^{n} \langle u, a_i \rangle^4 \right) \left(\sum_{i=1}^{n} \langle u, a_i \rangle^4 \right) \max_{i \neq j} |\langle a_i, a_j \rangle| , \end{aligned} \tag{7.10}$$

where the first line uses the Cauchy-Schwarz inequality (Lemma A.1) and the last line uses the fact that $D(u)$ satisfies the constraint $-1 \leqslant \langle u, a_i \rangle \leqslant 1$. By [GM15, Lemma 2], we have that

$$\mathcal{A} \vdash \sum_{i=1}^{n} \langle u, a_i \rangle^4 \leqslant 1 + \delta . \tag{7.11}$$

Combining the equation above, equation (7.10), equation (7.9), and the fact that with high probability $\langle a_i, a_j \rangle \leqslant \widetilde{O}(1/\sqrt{d})$, we obtain

$$\mathcal{A} \vdash \sum_{i=1}^{n} \langle u, a_i \rangle^{10} \geqslant 1 - O(\varepsilon) - \delta . \tag{7.12}$$

Therefore, using the fact that $\langle u, a_i \rangle^2 \leqslant 1$, we complete the proof. □

Lemma 7.3 (Rephrasing of [HSSS16, Lemma 5.9]). *Let* $a_1, \ldots, a_n \in \mathbb{R}^d$ *be independent random vectors drawn uniformly from the Euclidean unit sphere with* $1 \leqslant n \leqslant d^{1.5}/\log^{O(1)} d$*, and let* C *be the matrix with columns* $c_i = \mathrm{Id}_{\mathrm{sym}'} a_i^{\otimes 2}$*. Then*

$$\| C^{\mathsf{T}} C - \mathrm{Id}_n \| \leqslant \delta , \tag{7.13}$$

where $\delta = \tilde{\mathrm{O}}(\sqrt{n}/d + n/d^{1.5})$.

[9] Technically [GM15] only proved the case when the vectors a_i are uniform over $\{\pm 1/\sqrt{d}\}^d$, though the proofs work for the uniform distribution over the unit sphere as well.

Though [HSSS16, Lemma 5.9] assumes $n \geqslant d$, its proof can also handle $n \leqslant d$ if the error bound is relaxed to $\tilde{O}(\sqrt{n}/d)$. See specifically the end of the first paragraph of its proof. Also while [HSSS16, Lemma 5.9] assumes Gaussian random vectors, its proof reduces to the case on the unit sphere. Therefore we omit the proof of Lemma 7.3.

Finally we prove Proposition 7.1.

Proof of Proposition 7.1. Equation (7.6) follows from Lemma 7.3. To prove equation (7.3), we essentially just replace $a_i^{\otimes 2}$ in equation (7.8) by c_i and bound the approximation error. We have

$$\begin{aligned}
\mathcal{A} \vdash \sum_{i=1}^n \langle \mathrm{Id}_{\mathrm{sym}'} u^{\otimes 2}, c_i\rangle^4 &= \sum_{i=1}^n \langle u^{\otimes 2}, \mathrm{Id}_{\mathrm{sym}'} a_i^{\otimes 2}\rangle^4 = \sum_{i=1}^n \left(\langle u^{\otimes 2}, a_i^{\otimes 2}\rangle - \langle \Phi, a_i^{\otimes 2}\rangle\right)^4 \\
&= \sum_{i=1}^n \left(\langle u^{\otimes 2}, a_i^{\otimes 2}\rangle - 1/d\right)^4 \\
&\geqslant (1-1/d)\sum_{i=1}^n \langle u, a_i\rangle^8 - O(1/d) \\
&\geqslant 1 - O(\varepsilon) - O(1/d) - \tilde{O}(n/d^{3/2}). \qquad (7.14)
\end{aligned}$$

where the second last step uses $\vdash (x-y^2)^4 \geqslant (1-y^2)x^4 - O(y^2)$, and the last step uses equation (7.8). □

8 Robust decomposition of overcomplete 4-tensors

In this section we provide a sum-of-squares version of the FOOBI algorithm [LCC07]. FOOBI yields the rank decomposition of a 4th order tensor $T = \sum_{i=1}^n a_i^{\otimes 4}$ under the mild condition that the set $\{a_i^{\otimes 2} \otimes a_j^{\otimes 2} - (a_i \otimes a_j)^{\otimes 2}\}_{i \neq j}$ is linearly independent. FOOBI has not been formally shown to be robust to noise, though it's believed to tolerate spectral noise with magnitude up to some inverse polynomial of dimension. In contrast, the noise tolerance of our sum-of-squares version depends only on the condition number of certain matrices, and not directly on the dimension.

In Section 8.3 we additionally show, under a smoothed analysis model where each component a_i of the input tensor is randomly perturbed, that the relevant condition numbers are never smaller than some inverse polynomial of the dimension, with high probability over the random perturbations.

Throughout this section we will work with an input tensor T of the form

$$T = \sum_{i=1}^n a_i^{\otimes 4} + E$$

where $n \leqslant d^2$ and E is a symmetric noise tensor with bounded spectral norm $\|E\|_{\{1,2\},\{3,4\}}$.

For a matrix M, we use $\sigma_{\max}(M), \sigma_{\min}(M)$ to denote its largest and smallest singular values respectively, and $\sigma_k(M)$ to denote its kth largest singular value.

Let $A \in \mathbb{R}^{d^2 \times n}$ be the matrix with columns $a_i^{\otimes 2}$ for $i = 1, \ldots, n$. The guarantees of our algorithm will depend on the following 4th order condition number of A:

Definition 8.1. For a full rank matrix $A \in \mathbb{R}^{d^2 \times n}$ with columns $a_i^{\otimes 2}$ for $i = 1, \dots, n$, let $\kappa(A)$ defined as

$$\kappa(A) = \sigma_{\max}^{1.5}(Q)/\sigma_n^{1.5}(Q) + \sigma_{\max}^{2.5}(Q)/(\sigma_{\min}^2(B)\sigma_n^{0.5}(Q)) \,. \tag{8.1}$$

where $Q = AA^\mathsf{T}$ and $B \in \mathbb{R}^{d^4 \times n(n-1)}$ is the matrix with columns $b_{i,j} = a_i^{\otimes 2} \otimes a_j^{\otimes 2} - (a_i \otimes a_j)^{\otimes 2}$ for every $i \neq j$.

Theorem 8.2 (Restatement of Theorem 1.3)**.** *Let $\delta > 0$. Let $T \in (\mathbb{R}^d)^{\otimes 4}$ be a symmetric 4-tensor and $\{a_1, \dots, a_n\} \subseteq \mathbb{R}^d$ be a set of vectors. Define $E = T - \sum_{i=1}^n a_i^{\otimes 4}$ and define A as the matrix with columns $a_i^{\otimes 2}$. If $\|E\|_{\{1,2\},\{3,4\}} \leqslant \delta\, \sigma_n(AA^\mathsf{T})$ and Algorithm 3 outputs $\{\hat{a}_1, \dots, \hat{a}_n\}$ on input T, then there exists a permutation $\pi : [n] \to [n]$ so that for every $i \in [n]$,*

$$\left\| a_i - \hat{a}_{\pi(i)} \right\| \leqslant O\big(\delta\, \kappa(A)\big)\, \|a_i\| \,. \tag{8.2}$$

Algorithm 3 Sum-of-Squares FOOBI for robust overcomplete 4-tensor decomposition

Input: Number $\delta > 0$ and symmetric tensor $T \in (\mathbb{R}^d)^{\otimes 4}$.
Find: $\hat{a}_1, \dots, \hat{a}_n \in \mathbb{R}^d$.
Algorithm:

1. Compute the best rank-n approximation[10] $\widetilde{Q}$ of the $d^2 \times d^2$ matrix reshaping of T. Let $\widetilde{S}$ be the column span of $\widetilde{Q}$.

2. Run Algorithm 1 with inputs $P(\cdot)$ and $\mathcal{A}$ set to

$$P(x) = (\widetilde{Q}^+)^{1/2} x^{\otimes 2} \,, \tag{8.3}$$

$$\mathcal{A} = \left\{ \| \mathrm{Id}_{\widetilde{S}}\, x^{\otimes 2} \|^2 \geqslant (1 - 3\delta)\, \|x\|^4 \right\}_x \,. \tag{8.4}$$

   Suppose the algorithm outputs $\hat{c}_1, \dots, \hat{c}_n$.

3. Output $\hat{a}_1, \dots, \hat{a}_n$ such that for each $i \in [n]$, the matrix $\hat{a}_i \hat{a}_i^\mathsf{T}$ is the best rank-1 approximation of the matrix reshaping of $\widetilde{Q}^{1/2} \hat{c}_i$.

Let $\widetilde{Q}$ be the best rank-n approximation of the $d^2 \times d^2$ matrix reshaping of T, and let $\widetilde{S}$ be the column space of $\widetilde{Q}$. These two objects serve as our initial best-guess approximations of $Q = AA^\mathsf{T}$ and the subspace S spanned by $\{a_1^{\otimes 2}, \dots, a_n^{\otimes 2}\}$ (also the column space of Q), which we do not have access to. We define $B \in \mathbb{R}^{d^4 \times n(n-1)}$ as the matrix with columns $b_{i,j} = a_i^{\otimes 2} \otimes a_j^{\otimes 2} - (a_i \otimes a_j)^{\otimes 2}$ for $i \neq j$.

One of the core techniques in the analysis will be to use (the following rephrased version of) Davis and Kahan's "sin θ" Theorem, which bounds the principle angle between the column spaces of two matrices that are spectrally close to each other.

Theorem 8.3 (Direct consequence of Davis-Kahan Theorem [DK70])**.** *Suppose symmetric PSD matrices $Q \in \mathbb{R}^{D \times D}$ and $\widetilde{Q} \in \mathbb{R}^{D \times D}$ of rank $n \leqslant D$ satisfy $\|Q - \widetilde{Q}\| \leqslant \delta\, \sigma_n(Q)$. Let S and $\widetilde{S}$ be the column spaces of Q and $\widetilde{Q}$ respectively, and assume $\delta \leqslant \frac{1}{2}$. Then we have*

$$\sin(S, \widetilde{S}) \stackrel{\text{def}}{=} \| \mathrm{Id}_S - \mathrm{Id}_{\widetilde{S}} \mathrm{Id}_S \| = \| \mathrm{Id}_{\widetilde{S}} - \mathrm{Id}_S \mathrm{Id}_{\widetilde{S}} \| \leqslant \delta/(1 - \delta) \,. \tag{8.5}$$

Consequently,

$$\|\operatorname{Id}_S - \operatorname{Id}_{\widetilde{S}}\| \leqslant O(\delta)\,. \tag{8.6}$$

Theorem 8.2 follows from the analysis of Algorithm 1 as given in Theorem 5.2, as long as we can verify its two conditions, which we restate in the following two propositions. While Proposition 8.4 follows quickly from Theorem 8.3, we prove Proposition 8.5 over the next three subsections.

Proposition 8.4. *Let $P(x)$ and $\mathcal{A}$ be as defined in Algorithm 3. Then each vector $a_1, \dots, a_n$ satisfies $\mathcal{A}$.*

Proof. By Theorem 8.3, $\|\operatorname{Id}_{\widetilde{S}} a_i^{\otimes 2}\| \geqslant \|\operatorname{Id}_S a_i^{\otimes 2}\| - \|(\operatorname{Id}_S - \operatorname{Id}_{\widetilde{S}})\, a_i^{\otimes 2}\| \geqslant (1 - 2\delta)\,\|a_i\|^4$. □

Proposition 8.5. *Let $P(x)$ and $\mathcal{A}$ be as defined in Algorithm 3. Then*

$$\mathcal{A} \vdash_8 \sum_{i=1}^n \langle P(a_i), P(x)\rangle^4 \geqslant (1-\tau)\,\|P(x)\|^4\,,$$

where $\tau \leqslant O(\delta\,\sigma_{\max}^2(Q)/\sigma_{\min}^2(B)) + O(\delta\,\sigma_{\max}(Q)/\sigma_n(Q))$.

Proof of Theorem 8.2. By Theorem 5.2 along with Proposition 8.4 and Proposition 8.5, step 2 in Algorithm 3 must yield vectors $\hat{c}_1, \dots, \hat{c}_n$ that are respectively $O(\tau)$-close to $P(a_1), \dots, P(a_n)$, where $\tau \leqslant O(\delta\,\sigma_{\max}^2(Q)/\sigma_{\min}^2(B)) + O(\delta\,\sigma_{\max}(Q)/\sigma_n(Q))$. Then

$$\begin{aligned}
\|a_i^{\otimes 2} - \widetilde{Q}^{1/2}\hat{c}_i\| &\leqslant \|a_i^{\otimes 2} - \widetilde{Q}^{1/2}P(a_i)\| + \|\widetilde{Q}^{1/2}(P(a_i) - \hat{c}_i)\| \\
&\leqslant \|a_i^{\otimes 2} - \operatorname{Id}_{\widetilde{S}}\, a_i^{\otimes 2}\| + \sigma_{\max}^{1/2}(Q)\cdot O(\tau) \\
&\leqslant \|(\operatorname{Id}_S - \operatorname{Id}_{\widetilde{S}})a_i^{\otimes 2}\| + \sigma_{\max}^{1/2}(Q)\cdot O(\tau) \\
&\leqslant O(\delta\,\sigma_{\max}^{1/2}(Q)) + \sigma_{\max}^{1/2}(Q)\cdot O(\tau) \\
&\leqslant \sigma_{\max}^{1/2}(Q)\cdot O(\tau) \\
&\leqslant \sigma_{\max}^{1/2}(Q)/\sigma_n^{1/2}(Q)\cdot O(\tau)\cdot\|a_i^{\otimes 2}\| \\
&= O(\delta\,\kappa(A))\,\|a_i^{\otimes 2}\|\,.
\end{aligned}$$

Therefore taking the best rank-1 approximation of the matrix reshaping of $\widetilde{Q}^{1/2}\hat{c}_i$ gives an $O(\delta\,\kappa(A))$-approximation of a_i. □

8.1 Noiseless case

We first prove Proposition 8.5 in the noiseless case, when $T = \sum a_i^{\otimes 4}$ and $\widetilde{Q} = Q$ and $\widetilde{S} = S$. In this scenario, we find that the left-hand side of the conclusion of Proposition 8.5 becomes

$$\begin{aligned}
\sum_{i=1}^n \langle P(a_i), P(x)\rangle^4 &= \sum_{i=1}^n \left\langle (Q^+)^{1/2} a_i^{\otimes 2},\, (Q^+)^{1/2} x^{\otimes 2}\right\rangle^4 \\
&= \sum_{i=1}^n \left[(a_i^{\otimes 2})^{\mathsf{T}} Q^+ x^{\otimes 2}\right]^4 \\
&= \left\|A^{\mathsf{T}} Q^+ x^{\otimes 2}\right\|_4^4\,.
\end{aligned}$$

The term $\|P(x)\|_2^4$ on the right becomes $\|(Q^+)^{1/2}x^{\otimes 2}\|_2^4$. Thus Proposition 8.5 becomes

Proposition 8.6 (Noiseless Proposition 8.5)**.** *Let*

$$\mathcal{A}' = \left\{ \| \operatorname{Id}_S x^{\otimes 2} \|_2^2 \geqslant (1 - c\,\delta) \, \|x\|_2^4 \right\}_x , \tag{8.7}$$

for some constant $c \geqslant 0$. Then

$$\mathcal{A}' \vdash_8 \left\| A^\mathsf{T} Q^+ x^{\otimes 2} \right\|_4^4 \geqslant (1 - \tau) \left\| (Q^+)^{1/2} x^{\otimes 2} \right\|_2^4 ,$$

where $\tau \leqslant O(\delta \, \sigma_{\max}^2(Q)/\sigma_{\min}^2(B))$, where c is treated as a constant in the big-O notation.

Proof. We write $x^{\otimes 2}$ as a linear combination of the vectors $a_i^{\otimes 2}$ plus some term orthogonal to S.

$$\begin{aligned} \vdash \; x^{\otimes 2} &= \operatorname{Id}_S x^{\otimes 2} + \operatorname{Id}_{S^\perp} x^{\otimes 2} \\ &= \left[\sum_{i=1}^n \alpha_i \, a_i^{\otimes 2} \right] + \operatorname{Id}_{S^\perp} x^{\otimes 2} , \end{aligned}$$

where $\alpha = A^+ x^{\otimes 2}$ is a n-dimensional vector with polynomial entries. Since $Q = AA^\mathsf{T}$, it follows that $A^\mathsf{T} Q^+ x^{\otimes 2} = \alpha$ and $\|(Q^+)^{1/2} x^{\otimes 2}\|_2^4 = \|\alpha\|_2^4$, so that it will suffice to show $\|\alpha\|_4^4 \geqslant (1 - \tau) \, \|\alpha\|_2^4$.

We consider $x^{\otimes 4}$:

$$\begin{aligned} \vdash \; x^{\otimes 4} = x^{\otimes 2} \otimes x^{\otimes 2} &= \left[\sum_{i=1}^n \alpha_i \, a_i^{\otimes 2} \right] \otimes \left[\sum_{i=1}^n \alpha_i \, a_i^{\otimes 2} \right] + \zeta \\ &= \left[\sum_{i,j \in [n]} \alpha_i \alpha_j \, a_i^{\otimes 2} \otimes a_j^{\otimes 2} \right] + \zeta , \end{aligned} \tag{8.8}$$

with the error term $\zeta = (\operatorname{Id}_S x^{\otimes 2} + \operatorname{Id}_{S^\perp} x^{\otimes 2})^{\otimes 2} - (\operatorname{Id}_S x^{\otimes 2})^{\otimes 2}$, so that $\mathcal{A}' \vdash \|\zeta\|_2^2 \leqslant O(\delta) \, \|x\|_2^8$, since $\mathcal{A}' \vdash \| \operatorname{Id}_{S^\perp} x^{\otimes 2} \|_2^2 \leqslant O(\delta) \, \|x\|_2^4$ from the definition of $\mathcal{A}'$.

Since $x^{\otimes 4}$ is invariant with respect to permutation of its tensor modes, we can also write it as

$$\vdash \; x^{\otimes 4} = \left[\sum_{i=1}^n \alpha_i \alpha_j \, (a_i \otimes a_j)^{\otimes 2} \right] + \zeta' , \tag{8.9}$$

where similarly $\mathcal{A}' \vdash \|\zeta'\|_2^2 \leqslant O(\delta) \, \|x\|_2^8$.

Therefore, taking the difference of constraints (8.8) and (8.9) and recalling the definition of B being the matrix with columns $b_{i,j} = a_i^{\otimes 2} \otimes a_j^{\otimes 2} - (a_i \otimes a_j)^{\otimes 2}$, we obtain

$$\mathcal{A}' \; \vdash \; \left\| \sum_{i \neq j} \alpha_i \alpha_j \, b_{i,j} \right\|_2^2 = \|\zeta' - \zeta\|_2^2 \leqslant O(\delta) \, \|x\|_2^8 ,$$

so that therefore since $\|Bv\|_2^2 \geqslant \sigma_{\min}^2(B) \, \|v\|_2^2$ for all vectors v,

$$\mathcal{A}' \; \vdash \; \sum_{i \neq j} \alpha_i^2 \alpha_j^2 \cdot \sigma_{\min}^2(B) \leqslant \left\| \sum_{i \neq j} \alpha_i \alpha_j \, b_{ij} \right\|_2^2 \leqslant O(\delta) \, \|x\|_2^8$$

$$\leqslant O(\delta)\left[\sigma_{\max}(Q)\, x^{\otimes 2} Q^+ x^{\otimes 2}\right]^2 \leqslant O(\delta)\,\sigma_{\max}^2(Q)\,\|\alpha\|_2^4\,.$$

Hence, substituting in the above inequality,

$$\mathcal{A}' \;\vdash\; \|\alpha\|_4^4 = \|\alpha\|_2^4 - \sum_{i\neq j}\alpha_i^2\alpha_j^2 \geqslant (1 - O(\delta\,\sigma_{\max}^2(Q)/\sigma_{\min}^2(B)))\,\|\alpha\|_2^4\,. \qquad \square$$

8.2 Noisy case

At this point we've proved a version of Proposition 8.5 in the special case where there is no noise. In order to handle noise we need to show two things: first that the noisy set of polynomial constraints $\mathcal{A}$ used in Algorithm 3 implies the noiseless version $\mathcal{A}'$ from Proposition 8.6, and second that the desired conclusion of $\mathcal{A}$ in Proposition 8.5 follows from its noiseless counterpart in Proposition 8.6.

The first step follows immediately from Theorem 8.3:

Lemma 8.7. *Let $\mathcal{A}$ be defined as in Algorithm 3 and $\mathcal{A}'$ be defined as in Proposition 8.6. Then $\mathcal{A} \vdash \mathcal{A}'$.*

Proof. By Theorem 8.3, $\|\operatorname{Id}_S x^{\otimes 2}\|_2^2 = \|\operatorname{Id}_{\widetilde{S}} x^{\otimes 2}\|_2^2 - (x^{\otimes 2})^\mathsf{T}(\operatorname{Id}_{\widetilde{S}} - \operatorname{Id}_S)\, x^{\otimes 2} \geqslant (1 - O(\delta))\,\|x\|_2^4$. $\square$

For the second step, we have by Proposition 8.6 and Lemma 8.7 a statement of the form

$$\mathcal{A} \;\vdash_8\; \left\|A^\mathsf{T} Q^+ x^{\otimes 2}\right\|_4^4 \geqslant (1-\tau')\,\left\|(Q^+)^{1/2} x^{\otimes 2}\right\|_2^4\,,$$

but to prove Proposition 8.5 we need a statement of the form (after expanding out the function P)

$$\mathcal{A} \;\vdash_\ell\; \left\|A^\mathsf{T} \widetilde{Q}^+ x^{\otimes 2}\right\|_4^4 \geqslant (1-\tau)\,\left\|(\widetilde{Q}^+)^{1/2} x^{\otimes 2}\right\|_2^4\,.$$

Thus what remains is to show that not too much is lost when we approximate Q^+ with $\widetilde{Q}^+$.

Lemma 8.8. *Suppose symmetric PSD matrices $Q \in \mathbb{R}^{D\times D}$ and $\widetilde{Q} \in \mathbb{R}^{D\times D}$ both of rank $n \leqslant D$ satisfy $\|Q - \widetilde{Q}\| \leqslant \delta\,\sigma_n(Q)$. Then $\|Q(Q^+ - \widetilde{Q}^+)\| \leqslant O(\delta)$ and similarly $\left\|Q^{1/2}\big((Q^+)^{1/2} - (\widetilde{Q}^+)^{1/2}\big)\right\| \leqslant O(\delta)$.*

Proof. By Theorem 8.3, $\|QQ^+ - \widetilde{Q}\widetilde{Q}^+\| \leqslant \|\operatorname{Id}_S - \operatorname{Id}_{\widetilde{S}}\| \leqslant O(\delta)$, where S and $\widetilde{S}$ are the column spaces of Q and $\widetilde{Q}$ respectively. Then by adding and subtracting a term of $Q\widetilde{Q}^+$,

$$\left\|Q(Q^+ - \widetilde{Q}^+) + (Q - \widetilde{Q})\,\widetilde{Q}^+\right\| \leqslant O(\delta)\,.$$

By triangle inequality,

$$\left\|Q(Q^+ - \widetilde{Q}^+)\right\| \;\leqslant\; O(\delta) + \left\|(Q - \widetilde{Q})\,\widetilde{Q}^+\right\| \;\leqslant\; O(\delta) + \|Q - \widetilde{Q}\| \cdot \|\widetilde{Q}^+\| \;\leqslant\; O(\delta)\,.$$

The analogous result for $\left\|Q^{1/2}\big((Q^+)^{1/2} - (\widetilde{Q}^+)^{1/2}\big)\right\|$ is obtained by substituting $(Q^+)^{1/2}$ for Q^+ and $(\widetilde{Q}^+)^{1/2}$ for $\widetilde{Q}^+$ in the above argument. $\square$

We also show that not too much is lost when approximating a vector with high ℓ_4/ℓ_2 ratio.

Lemma 8.9. *For $\gamma, \beta, \tau \leqslant 1/2$, let $\mathcal{B}$ be the set of polynomial inequalities*

$$\mathcal{B} = \left\{-\beta\|v\|_2^2 \leqslant \|u\|_2^2 - \|v\|_2^2 \leqslant \beta\|v\|_2^2,\ \ \|u-v\|_2^2 \leqslant \gamma\|v\|_2^2\right\} \cup \left\{\|v\|_4^4 \geqslant (1-\tau)\|v\|_2^4\right\}.$$

Then we have

$$\mathcal{B} \vdash_4 \left\{\|u\|_4^4 \geqslant (1 - \tau - O(\sqrt{\gamma} + \beta))\,\|u\|_2^4\right\}.$$

Proof of Lemma 8.9. We have that $\left\{\|u-v\|_2^2 \leqslant \gamma\|v\|_2^2\right\} \vdash (u_i - v_i)^2 \leqslant \gamma\|v\|^2$. Moreover, since $\mathcal{B} \vdash \|u\|_2^2 \leqslant (1+\gamma)\|v\|_2^2$ we have that $\mathcal{B} \vdash (u_i + v_i)^2 \leqslant \|u+v\|_2^2 \leqslant O(\|v\|_2^2)$. Therefore it follows that

$$\begin{aligned}\mathcal{B} \ \vdash\ v_i^2 - u_i^2 \ &= \ (v_i - u_i)(u_i + v_i) \ \leqslant\ \sqrt{\gamma}/2 \cdot (u_i + v_i)^2 + 1/\sqrt{\gamma} \cdot (u_i - v_i)^2 \\ &\leqslant\ O(\sqrt{\gamma}\,\|v\|_2^2),\end{aligned}$$

where we used the AM-GM inequality. It follows that for every i,

$$\mathcal{B} \ \vdash\ \sum_{j\neq i} u_j^2 - v_j^2 \ \leqslant\ \|u\|^2 - \|v\|^2 + v_i^2 - u_i^2 \ \leqslant\ O((\sqrt{\gamma} + \beta)\,\|v\|_2^2).$$

Therefore by two rounds of adding-and-subtracting,

$$\begin{aligned}\mathcal{B} \ \vdash\ \sum_{i\neq j} u_i^2 u_j^2 &= \sum_i u_i^2 \left(\sum_{j\neq i} u_j^2 - \sum_{j\neq i} v_j^2\right) + \sum_i v_i^2 \left(\sum_{j\neq i} u_j^2 - \sum_{j\neq i} v_j^2\right) + \sum_{i\neq j} v_i^2 v_j^2 \\ &\leqslant \sum_i u_i^2 \cdot O((\sqrt{\gamma}+\beta)\,\|v\|_2^2) + \sum_i v_i^2 \cdot O((\sqrt{\gamma}+\beta)\,\|v\|_2^2) + \sum_{i\neq j} v_i^2 v_j^2 \\ &= O(\sqrt{\gamma}+\beta)\,(\|u\|_2^2\|v\|_2^2 + \|v\|_2^4) + \|v\|_4^4 - \|v\|_2^4 \\ &\leqslant (\tau + O(\sqrt{\gamma}+\beta))\,\|v\|_2^4 \\ &\leqslant (\tau + O(\sqrt{\gamma}+\beta))\,\|u\|_2^4.\end{aligned}$$

Here in the second line we used the axiom that $\|u\|_2^2 - \|v\|_2^2 \leqslant \beta\|v\|_2^2$, the second-to-last line uses the axiom $\|v\|_4^4 \geqslant (1-\tau)\|v\|_2^4$, and the last one uses $\|u\|_2^2 - \|v\|_2^2 \geqslant -\beta\|v\|_2^2$ so that $\|v\|_2^2 \leqslant (1-\beta)^{-1}\|u\|_2^2$. Rearranging the final inequality above we obtain the desired result. □

Proof of Proposition 8.5. We know by Proposition 8.6 and Lemma 8.7

$$\mathcal{A} \ \vdash_8\ \left\|A^\top Q^+ x^{\otimes 2}\right\|_4^4 \geqslant (1-\tau')\,\left\|(Q^+)^{1/2} x^{\otimes 2}\right\|_2^4, \tag{8.10}$$

where $\tau' \leqslant O(\delta\,\sigma_{\max}^2(Q)/\sigma_{\min}^2(B))$.

By Lemma 8.8,

$$\begin{aligned}\vdash\ \left\|A^\top Q^+ x^{\otimes 2} - A^\top \widetilde{Q}^+ x^{\otimes 2}\right\|_2^2 &= \left\|A^\top (Q^+ - \widetilde{Q}^+)\, x^{\otimes 2}\right\|_2^2 \\ &= \left\|A^+ Q (Q^+ - \widetilde{Q}^+)\, x^{\otimes 2}\right\|_2^2 \\ &\leqslant \|A^+\|_2^2 \cdot O(\delta^2) \cdot \|x\|_2^4 \\ &\leqslant O(\delta^2)\,\sigma_n^{-1}(Q)\,\|x\|_2^4.\end{aligned}$$

Also, using Lemma 8.8 after adding and subtracting a term of $Q^+Q\widetilde{Q}^+$,

$$\begin{aligned}\vdash \ \left\|A^\mathsf{T}Q^+x^{\otimes 2}\right\|_2^2 - \left\|A^\mathsf{T}\widetilde{Q}^+x^{\otimes 2}\right\|_2^2 &= (x^{\otimes 2})^\mathsf{T}Q^+AA^\mathsf{T}Q^+x^{\otimes 2} - (x^{\otimes 2})^\mathsf{T}\widetilde{Q}^+AA^\mathsf{T}\widetilde{Q}^+x^{\otimes 2}\\ &= (x^{\otimes 2})^\mathsf{T}\left[Q^+QQ^+ - \widetilde{Q}^+Q\widetilde{Q}^+\right]x^{\otimes 2}\\ &= (x^{\otimes 2})^\mathsf{T}\left[Q^+Q(Q^+ - \widetilde{Q}^+) + (Q^+ - \widetilde{Q}^+)Q\widetilde{Q}^+\right]x^{\otimes 2}\\ &\leqslant (x^{\otimes 2})^\mathsf{T}\left[O(\delta)\,Q^+ + O(\delta)\,\widetilde{Q}^+\right]x^{\otimes 2}\\ &\leqslant O(\delta)\,\sigma_n^{-1}(Q)\,\|x\|_2^4\,,\end{aligned}$$

and similarly in the other direction. Furthermore,

$$\vdash \ \left\|A^\mathsf{T}Q^+x^{\otimes 2}\right\|_2^2 = \left\|A^+x^{\otimes 2}\right\|_2^2 \geqslant \sigma_{\max}^{-1}(Q)\,\|x\|_2^4\,.$$

Combining the above three inequalities with Lemma 8.9 and (8.10),

$$\mathcal{A} \ \vdash_8 \ \left\|A^\mathsf{T}\widetilde{Q}^+x^{\otimes 2}\right\|_4^4 \geqslant (1-\tau)\left\|A^\mathsf{T}\widetilde{Q}^+x^{\otimes 2}\right\|_2^4\,,$$

where $\tau \leqslant O\big(\delta\,\sigma_{\max}^2(Q)/\sigma_{\min}^2(B) + \delta\,\sigma_{\max}(Q)/\sigma_n(Q)\big)$.

Finally, using the fact that $A^+Q^{1/2}$ is a whitened matrix and therefore has orthonormal rows and then using triangle inequality with Lemma 8.8,

$$\begin{aligned}\vdash \ \left\|A^\mathsf{T}\widetilde{Q}^+x^{\otimes 2}\right\|_2^2 &= \left\|A^+Q^{1/2}\cdot Q^{1/2}(\widetilde{Q}^+)^{1/2}\cdot(\widetilde{Q}^+)^{1/2}x^{\otimes 2}\right\|_2^2\\ &= \left\|Q^{1/2}(\widetilde{Q}^+)^{1/2}\cdot(\widetilde{Q}^+)^{1/2}x^{\otimes 2}\right\|_2^2\\ &\geqslant \left\|Q^{1/2}(Q^+)^{1/2}\cdot(\widetilde{Q}^+)^{1/2}x^{\otimes 2}\right\|_2^2 - \left\|Q^{1/2}[(Q^+)^{1/2} - (\widetilde{Q}^+)^{1/2}]\cdot(\widetilde{Q}^+)^{1/2}x^{\otimes 2}\right\|_2^2\\ &\geqslant \left\|\mathrm{Id}_S\cdot(\widetilde{Q}^+)^{1/2}x^{\otimes 2}\right\|_2^2 - O(\delta)\cdot\left\|(\widetilde{Q}^+)^{1/2}x^{\otimes 2}\right\|_2^2\\ &\geqslant (1-O(\delta))\,\left\|(\widetilde{Q}^+)^{1/2}x^{\otimes 2}\right\|_2^2\,.\end{aligned}$$

Combining the above two inequalities we obtain the theorem. □

8.3 Condition number under smooth analysis

In this section we prove that the condition number $\kappa(A)$ is at least inverse polynomial under the smooth analysis framework [ST04]. We work with the same ρ-perturbation model as introduced by [BCMV14]: Each $\tilde{a}_i$ is generated by adding a Gaussian random variable with covariance matrix $\frac{\rho}{d}\,\mathrm{Id}_d$ to a_i. We are given a symmetric 4th order tensor $\sum_{i=1}^n \tilde{a}_i^{\otimes 4}$ (with noise). Let $\widetilde{A}$ be the corresponding matrix with columns $\tilde{a}_i^{\otimes 2}$. We will give an upper bound on $\kappa(\widetilde{A})$. Suppose the vectors a_i have bounded norm; then $\sigma_{\max}(Q)$ is bounded, and therefore an upper bound on $\kappa(\widetilde{A})$ follows from establishing lower bounds on $\sigma_{\min}(\widetilde{B})$ and $\sigma_{\min}(\widetilde{A}\widetilde{A}^\mathsf{T})$. The lower bound on the latter follows from [BCMV14] and therefore we focus on the former.

Theorem 8.10. *Let $n \leqslant \frac{d^2}{10}$ and $\tilde{a}_1, \ldots, \tilde{a}_n$ be independent ρ-perturbations of $a_1, \ldots, a_n$. Let $\widetilde{B} \in \mathbb{R}^{d^4 \times n(n-1)}$ be the matrix with columns $\tilde{b}_{ij} = \tilde{a}_i^{\otimes 2} \otimes \tilde{a}_j^{\otimes 2} - (\tilde{a}_i \otimes \tilde{a}_j)^{\otimes 2}$ for $i \neq j$. Then with probability $1 - \exp(-d^{\Omega(1)})$, we have $\sigma_{\min}(\widetilde{B}) \geqslant \mathrm{poly}(1/d, \rho)$.*

We will bound the smallest singular value using the leave-one-out distance defined by [RV09].

Lemma 8.11. *[RV09] For matrix $A \in \mathbb{R}^{d\times n}$ with columns $A_i, i \in [n]$, let S_{-i} be the span of the columns without A_i, and $d(A) = \min_{i\in[n]} \|(\mathrm{Id} - \mathrm{Id}_{S_{-i}}) A_i\|$. Then $\sigma_{\min}(A) \geqslant \frac{1}{\sqrt{n}} d(A)$.*

To bound $d(\widetilde{B})$ from below, we use [BCMV14, Theorem 3.9] as our main tool.

Theorem 8.12. *[BCMV14, Theorem 3.9] Let $\delta \in (0,1)$ be a constant and W be an operator from $\mathbb{R}^{n^\ell}$ to $\mathbb{R}^m$ such that $\sigma_{\delta n^\ell}(W) \geqslant \eta$. Then for any $a_1, \ldots, a_\ell \in \mathbb{R}^d$ and their ρ-perturbations $\tilde{a}_1, \ldots, \tilde{a}_\ell$,*

$$\mathbb{P}\left[\left\|W(\tilde{a}_1 \otimes \cdots \otimes \tilde{a}_\ell)\right\| \geqslant \eta\rho^\ell d^{-O(3^\ell)}\right] \leqslant 1 - \exp(-\delta d^{1/3^\ell}) \tag{8.11}$$

Towards bounding the least singular value of $\widetilde{B}$ using Theorem 8.12, we need to address two issues that don't exist in [BCMV14]. The first one is that Theorem 8.12 requires $\tilde{a}_1, \ldots, \tilde{a}_\ell$ to be independent perturbations of $a_1, \ldots, a_\ell$. However, we need to deal with $\tilde{a}_i \otimes \tilde{a}_i \otimes \tilde{a}_j \otimes \tilde{a}_j$ which is a correlated perturbation of $a_i \otimes a_i \otimes a_j \otimes a_j$. We will use (a simpler version of) the decoupling technique of [dlPG99] and focus on a sub-matrix of $\widetilde{B}$ where the noise is un-correlated.

The second difficulty is that the columns of $\widetilde{B}$ are also correlated since each $\tilde{a}_i$ is used in n columns. Therefore when the leave-one-out distance of $\widetilde{B}$ is under consideration, the column $\widetilde{B}_{ij}$ and the subspace of the rest of the columns have correlated randomness, which prevents us from using Theorem 8.12 directly. We will address this issue by projecting $\widetilde{B}_{ij}$ into a smaller subspace which is un-correlated with $\widetilde{B}_{ij}$ and then apply Theorem 8.12.

Proof of Theorem 8.10. We partition $[d]$ into 4 disjoint subsets L_1, L_2, L_3, L_4 of size $d/4$. Let $\widetilde{B}'$ be the set of rows of $\widetilde{B}$ indexed by $L_1 \times L_2 \times L_3 \times L_4$. That is, the columns of $\widetilde{B}'$ are $\tilde{a}_{i,L_1} \otimes (\tilde{a}_{i,L_2} \otimes \tilde{a}_{j,L_3} - \tilde{a}_{j,L_2} \otimes \tilde{a}_{i,L_3}) \otimes \tilde{a}_{j,L_4}$, for $i \neq j$, where $\tilde{a}_{i,L}$ denotes the restriction of vector a_i to the subset L.

We fix a column $\widetilde{B}'_{ij}$ with $i \neq j$. Let $V = \mathrm{span}\{\widetilde{B}'_{k\ell} : (k,\ell) \neq (i,j)\}$. Clearly V is correlated with $\widetilde{B}'_{ij}$. We define the following subspace that contains V,

$$\hat{V} = \mathrm{span}\Big\{ \begin{aligned} &\tilde{a}_{j,L_1} \otimes x \otimes y \otimes \tilde{a}_{i,L_4}\,, \\ &\tilde{a}_{k,L_1} \otimes \tilde{a}_{k,L_2} \otimes x \otimes y\,, \\ &\tilde{a}_{k,L_1} \otimes x \otimes \tilde{a}_{k,L_3} \otimes y\,, \\ &x \otimes y \otimes \tilde{a}_{k,L_3} \otimes \tilde{a}_{k,L_4}\,, \\ &x \otimes \tilde{a}_{k,L_2} \otimes y \otimes \tilde{a}_{k,L_4} \end{aligned} \Big| \; x, y \otimes \mathbb{R}^{d/4}, k \notin \{i,j\} \Big\} \tag{8.12}$$

Therefore by definition $\hat{V} \supset V$, and thus $\hat{V}^\perp \subset V^\perp$ where $V^\perp$ denotes the subspace orthogonal to V. Observe that by the definition of $\hat{V}$, we have $\tilde{a}_{i,L_1} \otimes \tilde{a}_{i,L_2} \otimes \tilde{a}_{j,L_3} \otimes \tilde{a}_{j,L_4}$ is independent from $\hat{V}$. Moreover, $\hat{V}$ has dimension at most $d^2 + d^2 + 4nd^2 < d^4/2$. Then by Theorem 8.12 we obtain that with probability at least $1 - \exp(-d^{\Omega(1)})$,

$$\left\|\mathrm{Id}_{\hat{V}^\perp} \tilde{a}_{i,L_1} \otimes \tilde{a}_{i,L_2} \otimes \tilde{a}_{j,L_3} \otimes \tilde{a}_{j,L_4}\right\| \geqslant \mathrm{poly}(1/d, \rho)\,.$$

Consequently,

$$\left\|\mathrm{Id}_{V^\perp} \widetilde{B}'_{ij}\right\| \geqslant \left\|\mathrm{Id}_{\hat{V}^\perp} \widetilde{B}'_{ij}\right\| = \left\|\mathrm{Id}_{\hat{V}^\perp} \tilde{a}_{i,L_1} \otimes \tilde{a}_{i,L_2} \otimes \tilde{a}_{j,L_3} \otimes \tilde{a}_{j,L_4}\right\| \geqslant \mathrm{poly}(1/d, \rho)\,,$$

where the first inequality follows from $V \subset \hat{V}$ and second one follows from the fact that $\tilde{a}_{i,L_1} \otimes \tilde{a}_{j,L_2} \otimes \tilde{a}_{i,L_3} \otimes \tilde{a}_{j,L_4}$ is orthogonal to the subspace $\hat{V}^\perp$.

Then taking union bound over all $i \neq j$, we obtain that $d(\widetilde{B}') \geqslant \mathrm{poly}(1/d, \rho)$ occurs with probability $1 - \exp(-d^{\Omega(1)})$. Therefore $\sigma_{\min}(\widetilde{B}') \geqslant \mathrm{poly}(1/d, \rho)$ which in turn implies that $\sigma_{\min}(\widetilde{B}) \geqslant \sigma_{\min}(\widetilde{B}') \geqslant \mathrm{poly}(1/d, \rho)$.

□

9 Tensor decomposition with general components

In this section we prove Theorem 1.6 (tensor decomposition with general components). The key ingredient is a scheme for rounding pseudo-distributions (see Theorem 9.1 below) that improves over our previous scheme (Theorem 4.1): The improved rounding scheme only requires moments of degree logarithmic in the overcompleteness parameter σ.

9.1 Improved rounding of pseudo-distributions

Theorem 9.1. *Let $s, \sigma \geqslant 1$ and $\varepsilon \in (0,1)$. Let D be a degree-s pseudo-distribution over $\mathbb{R}^d$ that satisfies the constraint $\{\|u\|^2 \leqslant 1\}$, and let a be a unit vector in $\mathbb{R}^d$. Suppose that $s \geqslant O(1/\varepsilon) \cdot \log(\sigma/\varepsilon)$ and,*

$$\tilde{\mathbb{E}}_{D(u)} \langle a, u\rangle^{2s+2} \geqslant \Omega(1/\sigma) \cdot \left\|\tilde{\mathbb{E}}\left[uu^\mathsf{T}\right]\right\| \geqslant d^{-O(1)}\,. \tag{9.1}$$

Then, with probability at least $1/d^{O(s^3)}$ over the choice of independent random variables $g_1, \ldots, g_s \sim \mathcal{N}(0, \mathrm{Id}_{d^2})$, the top eigenvector $u^\star$ of the following matrix satisfies that $\langle a, u^\star\rangle^2 \geqslant 1 - O(\varepsilon)$,

$$M_g = \tilde{\mathbb{E}}_{D(u)} \langle g_1, u^{\otimes 2}\rangle \cdots \langle g_s, u^{\otimes 2}\rangle \cdot uu^\mathsf{T} \tag{9.2}$$

We start by defining some notations for convenience. Let

$$p_g(u) = \langle g_1, u^{\otimes 2}\rangle \cdots \langle g_s, u^{\otimes 2}\rangle\,.$$

Moreover, Let

$$\alpha_j = \langle g_j, a^{\otimes 2}\rangle, \qquad g_j' = g_j - \alpha_j a^{\otimes 2}, \quad \text{and} \quad \beta_j = \langle g_j', u\rangle\,.$$

Therefore we have that $\langle g_j, u^{\otimes 2}\rangle = \langle \alpha_j a^{\otimes 2}, u^{\otimes 2}\rangle + \langle g_j', u^{\otimes 2}\rangle = \alpha_j \langle a, u\rangle^2 + \beta_j$, and it follows that $p_g(u) = \prod_{1 \leqslant j \leqslant s} (\alpha_j \langle a, u\rangle^2 + \beta_j)$.

Theorem 9.1 follows from the following proposition and a variant of Wedin's Theorem (see Lemma A.5).

Proposition 9.2. *In the setting of Theorem 9.1, let $\mathrm{Id}_{-1} = \mathrm{Id} - aa^\mathsf{T}$. Then, with at least $(\Omega(1/n) - 1/d^{O(1)}) \cdot 1/d^{O(s^3)}$ probability over randomness of g, we have*

$$\max\left\{\left\|\tilde{\mathbb{E}}\left[p_g(u)\,\mathrm{Id}_{-1}\, uu^\mathsf{T}\, \mathrm{Id}_{-1}\right]\right\|, \left\|\tilde{\mathbb{E}}\left[p_g(u)\,\mathrm{Id}_{-1}\, uu^\mathsf{T}\, \mathrm{Id}_1\right]\right\|\right\} \leqslant \varepsilon\, \tilde{\mathbb{E}}\left[p_g(u)\langle u, a\rangle^2\right]\,. \tag{9.3}$$

Proposition above follows from the following two propositions, one of which lowerbounds the RHS of (9.3) and the other upperbounds the LHS of (9.3).

Proposition 9.3. *In the setting of Theorem 9.1, let $\tau = s\sqrt{s \log d}$. Conditioned on the event that $\alpha_1, \ldots, \alpha_s \geqslant \tau$, we have with at least $\Omega(1/n)$ probability over the choice of g'*

$$\tilde{\mathbb{E}}\left[p_g(u)\langle u, a\rangle^2\right] \geqslant 0.9\alpha_1 \ldots \alpha_s \, \tilde{\mathbb{E}}[\langle a, u\rangle^{2s+2}]$$

Proposition 9.4. *In the setting of Theorem 9.1, let $\tau = s\sqrt{s \log d}$ and $\mathrm{Id}_{-1} = \mathrm{Id} - aa^\top$. Conditioned on the event that $\alpha_1, \ldots, \alpha_s \geqslant \tau$, we have with at least $1 - d^{-\Omega(1)}$ probability over the choice of g',*

$$\max\left\{\left\|\tilde{\mathbb{E}}\left[p_g(u)\,\mathrm{Id}_{-1}\, uu^\top\, \mathrm{Id}_{-1}\right]\right\|, \left\|\tilde{\mathbb{E}}\left[p_g(u)\,\mathrm{Id}_{-1}\, uu^\top\, \mathrm{Id}_1\right]\right\|\right\} \leqslant O(\varepsilon\alpha_1 \ldots \alpha_s) \cdot \tilde{\mathbb{E}}\left[\langle a, u\rangle^{2s+2}\right]. \quad (9.4)$$

We first prove Proposition 9.3. We need the following three lemmas.

Lemma 9.5. *Let $\varepsilon \in (0, 1/3)$ and $0 \leqslant \delta \leqslant \varepsilon$. Suppose $0 \leqslant \kappa \leqslant \delta\alpha_j$ for every $j \in [s]$, then there exists a SoS proof*

$$\vdash_x x^2 \prod_{j\in[s]} (\alpha_j x^2 + \kappa) \leqslant \alpha_1 \ldots \alpha_s \left((1-\varepsilon)^s x^2 + (1 + O(\delta))^s x^{2s+2}\right)$$

Proof. Since this is a univariate polynomial inequality, it suffice to show that it's true for every x, which will imply that there is also a SoS proof. For $x \in \mathbb{R}$ such that $x^2 \leqslant 1 - 2\varepsilon$, we have that

$$\begin{aligned} x^2 \prod_{j\in[s]} (\alpha_j x^2 + \kappa) &\leqslant x^2 \prod_{j\in[s]} (1-\varepsilon)\alpha_j && \text{(by } \varepsilon \geqslant \delta \text{ and } \delta\alpha_j \geqslant \kappa\text{)} \\ &\leqslant (1-\varepsilon)^s \alpha_1 \ldots \alpha_s x^2 \end{aligned}$$

For $x \in \mathbb{R}$ such that $x^2 \geqslant 1 - 2\varepsilon \geqslant 1/3$, we have that

$$\begin{aligned} x^2 \prod_{j\in[s]} (\alpha_j x^2 + \kappa) &\leqslant x^2 \prod_{j\in[s]} \alpha_j (1 + O(\delta)) x^2 && \text{(by } x^2 \geqslant 1/3 \text{ and } \delta\alpha_j \geqslant \kappa\text{)} \\ &\leqslant \alpha_1 \ldots \alpha_s (1 + O(\delta))^s x^{2s+2} \end{aligned}$$

Hence we obtain a proof for the nonnegativity of the target polynomial. It is known that every nonnegative univariate polynomial admits a sum-of-squares proof. Therefore the inequality above has a sum-of-squares proof. □

Lemma 9.6. *In the setting of Theorem 9.1, let $\tau = s\sqrt{s \log d}$, $\kappa = O(\sqrt{s \log d})$. Conditioned on the event that $\alpha_1, \ldots, \alpha_j \geqslant \tau$, we have*

$$\tilde{\mathbb{E}}\left[\langle a, u\rangle^2 \prod_{j\in[s]} (\alpha_j \langle a, u\rangle^2 + \kappa)\right] \leqslant \alpha_1 \ldots \alpha_s \cdot O(\tilde{\mathbb{E}}\left[\langle a, u\rangle^{2s+2}\right]).$$

Proof. By Lemma 9.5, we have,

$$\begin{aligned} \tilde{\mathbb{E}}\left[\langle a, u\rangle^2 \prod_{j\in[s]} (\alpha_j \langle a, u\rangle^2 + \kappa)\right] &\leqslant \alpha_1 \ldots \alpha_s \left((1-\varepsilon)^s \, \tilde{\mathbb{E}}[\langle a, u\rangle^2] + (1 + O(1/s))^s \, \tilde{\mathbb{E}}\left[\langle a, u\rangle^{2s+2}\right]\right) \\ &\leqslant \alpha_1 \ldots \alpha_s \cdot O(\tilde{\mathbb{E}}\left[\langle a, u\rangle^{2s+2}\right]) \\ &\qquad \text{(by } (1-\varepsilon)^s \leqslant \varepsilon/\sigma \text{ and } \tilde{\mathbb{E}}\left[\langle u, a\rangle^{2s+2}\right] \geqslant \tfrac{1}{\sigma}\left\|\tilde{\mathbb{E}}[uu^\top]\right\|\text{)} \end{aligned}$$

□

Proof of Proposition 9.3. We have

$$\begin{aligned}
\mathop{\mathbb{E}}_{g'}\left[\tilde{\mathbb{E}}\left[p_g(u)\langle u,a\rangle^2\right] \mid \alpha_1,\dots,\alpha_s\right] &= \mathop{\mathbb{E}}_{g'}\left[\tilde{\mathbb{E}}\left[\prod_{1\leqslant j\leqslant s}(\alpha_j\langle a,u\rangle^2+\beta_j)\langle u,a\rangle^2\right] \mid \alpha_1,\dots,\alpha_s\right] \\
&= \tilde{\mathbb{E}}\left[\prod_{j\in[s]}(\alpha_j\langle a,u\rangle^2+\mathbb{E}[\beta_j])\cdot\langle u,a\rangle^2\right] \\
&\qquad \text{(by linearity of pseudo-expectation and independence of } g'_1,\dots,g'_s) \\
&= \alpha_1\dots\alpha_s\,\tilde{\mathbb{E}}[\langle a,u\rangle^{2s+2}] \qquad (9.5)
\end{aligned}$$

Moreover, we bound the variance,

$$\begin{aligned}
\mathop{\mathbb{E}}_{g'}\left[\left(\tilde{\mathbb{E}}\left[p_g(u)\langle u,a\rangle^2\right]\right)^2 \mid \alpha_1,\dots,\alpha_s\right] &\leqslant \mathop{\mathbb{E}}_{g'}\left[\tilde{\mathbb{E}}\left[p_g(u)^2\langle u,a\rangle^4\right] \mid \alpha_1,\dots,\alpha_s\right] \\
&= \tilde{\mathbb{E}}\left[\prod_{j\in[s]}\mathop{\mathbb{E}}_{g_j}\left[\alpha_j\langle a,u\rangle^2+\beta_j\right]^2\cdot\langle u,a\rangle^4\right] \\
&= \tilde{\mathbb{E}}\left[\prod_{j\in[s]}(\alpha_j^2\langle a,u\rangle^4+1)\cdot\langle u,a\rangle^4\right] \\
&\leqslant \tilde{\mathbb{E}}\left[\prod_{j\in[s]}(\alpha_j^2\langle a,u\rangle^2+1)\cdot\langle u,a\rangle^2\right] \qquad \text{(by } \langle u,a\rangle^2\leqslant 1) \\
&= \alpha_1\dots\alpha_s\,\tilde{\mathbb{E}}[\langle a,u\rangle^{2s+2}] \qquad \text{(By Lemma 9.6)}
\end{aligned}$$

Therefore, by Paley-Zygmund inequality, we have that with probability

$$\mathop{\mathbb{P}}_{g'}\left[\tilde{\mathbb{E}}\left[p_g(u)\langle u,a\rangle^2\right] \geqslant 0.9\alpha_1\dots\alpha_s\,\tilde{\mathbb{E}}[\langle a,u\rangle^{2s+2}] \mid \alpha_1,\dots,\alpha_s\right] \geqslant \frac{1}{100}\alpha_1\dots\alpha_s\,\tilde{\mathbb{E}}[\langle a,u\rangle^{2s+2}] \geqslant \Omega(1/n)\,,$$

which completes the proof. □

Towards proving Proposition 9.4, we start with the following Lemma.

Lemma 9.7. *Let $\varepsilon>0$, $k\in\mathbb{N}$, $a\in\mathbb{R}^d$ with $\|a\|=1$ and $\mathcal{A}=\{\|u\|^2\leqslant 1\}$. Then, there exists a matrix sum-of-squares proof,*

$$\mathcal{A}\vdash_{u,1/\varepsilon}\left(\langle a,u\rangle^{2k}-(1-\varepsilon)^k\right)\cdot(\mathrm{Id}_{-1}\,u)(\mathrm{Id}_{-1}\,u)^{\mathsf{T}}\preceq O(\varepsilon)\cdot\langle a,u\rangle^{2k+2}\cdot\mathrm{Id}\ .$$

Proof. Let $r=1/\varepsilon$. We may assume r is an integer and that $\varepsilon>0$ is small enough such that $(1-\varepsilon)^r\geqslant 1/3$. Then, the univariate polynomial inequality $x^{2k}-(1-\varepsilon)^k\leqslant 3x^{2k}\cdot x^{2r}$ holds for all $x\in\mathbb{R}$. (For $x^2<1-\varepsilon$, the left-hand side is negative. For $x^2\geqslant 1-\varepsilon$, the right-hand side is at least x^{2k} because $3x^{2r}\geqslant x^{2r}/(1-\varepsilon)^r\geqslant 1$.) It follows that there exists a sum-of-squares proof

$$\vdash_x x^{2k}-(1-\varepsilon)^k\leqslant 3x^{2k}\cdot x^{2r}. \qquad (9.6)$$

Similarly, there exists a sum-of-squares proof (see the texts below equation (4.7) as well)

$$\vdash_x x^{2r}(1-x^2) \leqslant O(1/r)\cdot x^2 = O(\varepsilon)\cdot x^2. \tag{9.7}$$

Therefore,

$$\begin{aligned}
\mathcal{A} \vdash_{u,1/\varepsilon} & \left(\langle a,u\rangle^{2k} - (1-\varepsilon)^k\right)\cdot (\mathrm{Id}_{-1}\, u)(\mathrm{Id}_{-1}\, u)^\top \\
&\preceq 3\langle a,u\rangle^{2k+2r}\cdot (\mathrm{Id}_{-1}\, u)(\mathrm{Id}_{-1}\, u)^\top && \text{(by (9.6))} \\
&\preceq 3\langle a,u\rangle^{2k+2r}\cdot (1-\langle a,u\rangle^2)\,\mathrm{Id} && \text{(because } \vdash vv^\top \preceq \|v\|^2\,\mathrm{Id} \text{ by Lemma 3.10)} \\
&\preceq O(\varepsilon)\cdot \langle a,u\rangle^{2k+2}\cdot \mathrm{Id}\ . && \text{(by (9.7))}
\end{aligned}$$

□

Proof of Proposition 9.4. We only bound $\left\|\tilde{\mathbb{E}}\left[p_g(u)\,\mathrm{Id}_{-1}\, uu^\top\,\mathrm{Id}_{-1}\right]\right\|$. The other term can be controlled similarly and the detailed proof are left to the readers. Let $\alpha_S = \prod_{j\in S}\alpha_j$ and $\beta_S = \prod_{j\in S}\beta_j$. By the fact that $\langle g_j, u^{\otimes 2}\rangle = \alpha_j\langle a,u\rangle^2 + \beta_j$, we have,

$$p_g(u)\,\mathrm{Id}_{-1}\, uu^\top\,\mathrm{Id}_{-1} = \sum_{S\subset[s],L=S^c} \underbrace{\alpha_S\beta_L\langle a,u\rangle^{2|S|}\,\mathrm{Id}_{-1}\, uu^\top\,\mathrm{Id}_{-1}}_{W_S(u)}, \tag{9.8}$$

where each summand is denoted by $W_S(u)$. Observe that W_S can be written as

$$W_S(u) = \left(\mathrm{Id}\otimes\mathrm{Id}\otimes \underbrace{g_{j_1}'^\top\otimes\cdots\otimes g_{j_r}'^\top}_{\{j_1,\ldots,j_r\}=T}\right)\cdot\left(\langle a,u\rangle^{2|S|}(\mathrm{Id}_{-1}\, u)\otimes(\mathrm{Id}_{-1}\, u)\otimes u^{\otimes|L|}\right)$$

Then by Theorem 6.8, with probability at least $1-2^{-s}d^{-\Omega(1)}$ over the choice of $g_1',\ldots,g_s'$ we have ,

$$\begin{aligned}
\left\|\tilde{\mathbb{E}}\left[W_S\right]\right\| &\leqslant \alpha_S O(s\log d)^{|L|/2}\cdot \max_{J\in[|L|+2]:1\in J,2\notin J}\left\|\tilde{\mathbb{E}}\left[\langle a,u\rangle^{2|S|}(\mathrm{Id}_{-1}\, u)\otimes(\mathrm{Id}_{-1}\, u)\otimes u^{\otimes|L|}\right]\right\|_{J,J^c} \\
&\leqslant \alpha_S O(s\log d)^{|L|/2}\cdot\left\|\tilde{\mathbb{E}}\left[\langle a,u\rangle^{2|S|}(\mathrm{Id}_{-1}\, u)\otimes(\mathrm{Id}_{-1}\, u)\right]\right\| \qquad \text{(Lemma 6.3 and } \|u\|^2\leqslant 1.)
\end{aligned}$$

By Lemma 9.7 we have that

$$\{\|u\|^2\leqslant 1\}\vdash_{u,1/\varepsilon}\left(\langle a,u\rangle^{2|S|} - (1-\varepsilon)^{|S|}\right)\cdot(\mathrm{Id}_{-1}\, u)(\mathrm{Id}_{-1}\, u)^\top \preceq O(\varepsilon)\cdot\langle a,u\rangle^{2|S|+2}\cdot\mathrm{Id}\ .$$

Therefore taking pseudo-expectation, we obtain that

$$\tilde{\mathbb{E}}\left[\langle a,u\rangle^{2|S|}(\mathrm{Id}_{-1}\, u)\otimes(\mathrm{Id}_{-1}\, u)\right] \preceq \tilde{\mathbb{E}}\left[(1-\varepsilon)^{|S|}(\mathrm{Id}_{-1}\, u)\otimes(\mathrm{Id}_{-1}\, u)\right] + O(\varepsilon)\,\tilde{\mathbb{E}}\left[\langle a,u\rangle^{2|S|+2}\,\mathrm{Id}\right] \tag{9.9}$$

Then using the fact that

$$\left\|\tilde{\mathbb{E}}\left[(1-\varepsilon)^{|S|}(\mathrm{Id}_{-1}\, u)\otimes(\mathrm{Id}_{-1}\, u)\right]\right\| = \left\|\mathrm{Id}_{-1}\,\tilde{\mathbb{E}}\left[(1-\varepsilon)^{|S|}uu^\top\right]\mathrm{Id}_{-1}\right\| \leqslant (1-\varepsilon)^{|S|}\left\|\tilde{\mathbb{E}}\left[uu^\top\right]\right\|,$$

and equation (9.9), we have

$$\left\|\tilde{\mathbb{E}}\left[W_S\right]\right\| \leqslant \alpha_S O(s\log d)^{|L|/2}\cdot\left((1-\varepsilon)^{|S|}\left\|\tilde{\mathbb{E}}\left[uu^\top\right]\right\| + O(\varepsilon)\,\tilde{\mathbb{E}}\left[\langle a,u\rangle^{2|S|+2}\,\mathrm{Id}\right]\right) \tag{9.10}$$

Taking union bound over all subset S, with probability at least $1 - d^{-\Omega(1)}$, we have equation (9.10) holds for every $S \subset [s]$. Taking the sum of equation (9.10) over S, we conclude that

$$\left\|\tilde{\mathbb{E}}\left[p_g(u)\,\mathrm{Id}_{-1}\,uu^\top\,\mathrm{Id}_{-1}\right]\right\| \leqslant \sum_S \left\|\tilde{\mathbb{E}}\left[W_S\right]\right\| \qquad \text{(by equation (9.8))}$$

$$\leqslant \left\|\tilde{\mathbb{E}}\left[uu^\top\right]\right\| \cdot \prod_{j\in[s]}((1-\varepsilon)\alpha_j + O(\sqrt{s\log d})) + O(\varepsilon)\,\tilde{\mathbb{E}}\left[\langle a,u\rangle^2 \prod_{j\in[s]}(\alpha_j\langle a,u\rangle^2 + O(\sqrt{s\log d}))\right] \quad (9.11)$$

When $\alpha_j \geqslant \tau$, where $\tau = s\sqrt{s\log d}$ and $s = \frac{c}{\varepsilon}\log(\sigma/\varepsilon)$ where c is a sufficiently large absolute constant, using the fact that $\left\|\tilde{\mathbb{E}}\left[uu^\top\right]\right\| \leqslant \sigma/n$, we have

$$\left\|\tilde{\mathbb{E}}\left[uu^\top\right]\right\| \prod_{j\in[s]}((1-\varepsilon)\alpha_j + O(\sqrt{s\log d})) \leqslant (1-\varepsilon/2)^s\alpha_1\ldots\alpha_s \leqslant \frac{\varepsilon}{n}\alpha_1\ldots\alpha_s\,.$$

Regarding the second term on the RHS of (9.11), by Lemma 9.6, we have that when $\alpha_j \geqslant \tau$,

$$\tilde{\mathbb{E}}\left[\langle a,u\rangle^2 \prod_{j\in[s]}(\alpha_j\langle a,u\rangle^2 + O(\sqrt{s\log d}))\right] \leqslant \alpha_1\ldots\alpha_s \cdot O(\tilde{\mathbb{E}}\left[\langle a,u\rangle^{2s+2}\right])$$

Therefore, plugging in the two bounds above into equation (9.11), we obtain that

$$\left\|\tilde{\mathbb{E}}\left[p_g(u)\,\mathrm{Id}_{-1}\,uu^\top\,\mathrm{Id}_{-1}\right]\right\| \leqslant O(\varepsilon\alpha_1\ldots\alpha_s)\cdot\tilde{\mathbb{E}}\left[\langle a,u\rangle^{2s+2}\right].$$

□

9.2 Finding all components

In this section we prove Theorem 1.6 (restated below) using iteratively the rounding scheme that is developed in the subsection before.

Theorem (Restatement of Theorem 1.6)**.** *There exists an algorithm A (see Algorithm 4) with polynomial running time (in the size of its input) such that for all $\varepsilon \in (0,1), \sigma \geqslant 1$, for every set of unit vectors $\{a_1,\ldots,a_n\} \subseteq \mathbb{R}^d$ with $\|\sum_{i=1}^n a_i a_i^\top\| \leqslant \sigma$ and every symmetric $2k$-tensor $T \in (\mathbb{R}^d)^{\otimes 2k}$ with $k \geqslant (1/\varepsilon)^{O(1)}\cdot\log(\sigma)$ and $\left\|T - \sum_i a_i^{\otimes 2k}\right\|_{\{1,\ldots,k\},\{k+1,\ldots,2k\}} \leqslant 1/3$, we have*

$$\mathrm{dist}_H\left(A(T), \{a_1^{\otimes 2},\ldots,a_n^{\otimes 2}\}\right)^2 \leqslant O(\varepsilon)\,.$$

Proof of Theorem 1.6. We analyze Algorithm 4. Let $\eta = c_0\varepsilon^{1/2}$ where c_0 is a large enough absolute constant. Let $\mathcal{A}'$ be the constraint that $\sum_i a_i, u^{2s+2} \geqslant 1/3$. Then we have $\mathcal{A} \vdash \mathcal{A}'$. We first observe that as long as a vector a satisfies $\mathcal{A}$, then a has to be $O(\varepsilon^{1/2})$-close to one of the a_i's up to sign flip. This is because

$$1/3 \leqslant \langle T, u^{\otimes 2s+2}\rangle - 1/3 \leqslant \langle u, a_i\rangle^{2s+2} \leqslant \max_i\langle a_i,u\rangle^{2s}\left(\sum_i\langle a_i,u\rangle^2\right) \leqslant \sigma\max_i\langle a_i,u\rangle^{2s}\,.$$

Algorithm 4 Tensor decomposition with general components

Parameters: numbers $\varepsilon > 0$, $n \in \mathbb{N}$.

Given: $2k$-th order tensor T

Find: Set of vectors $S = \{\hat{a}_1, \ldots, \hat{a}_{n'}\} \subset \mathbb{R}^d$ with $n' \leqslant n$.

- Let $s = k - 1$, $\ell = O(s)$, and $\eta = O(\varepsilon)^{1/2}$.

$$\mathcal{A} = \left\{\|u\|^2 = 1\right\} \cup \left\{\langle T, u^{2s+2}\rangle \geqslant 2/3\right\} . \tag{9.12}$$

- For i from 1 to n, do the following:

  1. Compute a ℓ-degree pseudo-distribution $D(u)$ over $\mathbb{R}^d$ that satisfies the constraints

$$\mathcal{A} \text{ and } \left\|\tilde{\mathbb{E}}_{D(u)}\, uu^{\mathsf{T}}\right\| \leqslant \frac{\sigma}{n-i+1} . \tag{9.13}$$

  2. Repeat $T = d^{O(s^3)}$ rounds of the following:
     - Choose standard Gaussian vectors $g_1, \ldots, g_s \sim \mathcal{N}(0, \mathrm{Id}_{d^2})$ and compute the top eigenvectors $a^\star$ of the following matrix,

$$\tilde{\mathbb{E}}_{D(u)} \langle g_s, u^{\otimes 2}\rangle \ldots \langle g_s, u^{\otimes 2}\rangle \cdot uu^{\mathsf{T}} \in \mathbb{R}^{d\times d} . \tag{9.14}$$

     - Check if $a^\star$ satisfies $\mathcal{A}$. If yes, let $\hat{a}_i = \hat{a}^\star$ and $S \leftarrow S \cup \{\hat{a}_i\}$, add to $\mathcal{A}$ the constraint $\{\langle u, \hat{a}_i\rangle^2 \leqslant 1 - 5\eta\}$, and break the (inner) loop.
  3. If no new $\hat{a}_i$ is found in the previous step, stop the algorithm.

That is, we can always check whether $a^\star$ is what we wanted as in the second bullet of step 2. Therefore, it remains to show that as long as there exists a_j that is $\eta^{1/2}$-far away (up to sign flip) to the set S, we will find a new vector after Step 2 in the next iteration.

We assume that after iteration i_0, the set $W = \{a_j : \forall i \in [i_0], \langle a_j, \hat{a}_i\rangle^2 \leqslant 1-\eta\}$ is not empty. We will show that after iteration i_0+1, we will find a new vector in W up to $O(\varepsilon)^{1/2}$ error. We claim first that in the i_0+1 iteration there exists a pseudo-distribution $D(u)$ that satisfies (9.13). Indeed, this is because the actual uniform distribution over the finite set W satisfies constraint (9.13). Here we used the fact that for every $j \in [n]$ we have $\langle T, a_j^{\otimes 2k}\rangle \geqslant \langle \sum_i a_i^{\otimes 2k}, a_j^{\otimes 2k}\rangle - 1/3 \geqslant \langle a_j, a_j\rangle^k - 1/3 = 2/3$.

Since constraints (9.13) enforce that for every $i \leqslant i_0$ pseudo-distribution $D(u)$ satisfies that $\langle u, \hat{a}_i\rangle^2 \leqslant 1-\eta$, and moreover, we have $\|a_i - \tau_i \hat{a}_i\|^2 \leqslant O(\varepsilon)$ for some $\tau_i \in \{-1,+1\}$, by Lemma A.2, we conclude that $D(u)$ also satisfies the constraint that $\langle u, a_i\rangle^2 \leqslant 1-\eta/2$ (here we use the fact that $\eta = c_0\varepsilon$ with large enough constant c_0). These implies that $D(u)$ satisfies that $\sum_{i=1}^{i_0}\langle u, a_i\rangle^{2s+2} \leqslant (1-\eta)^{2s}\sum_{i=1}^{i_0}\langle u, a_i\rangle^2$. Therefore, we have $\tilde{\mathbb{E}}\left[\sum_{i=1}^{i_0}\langle u, a_i\rangle^{2s+2}\right] \leqslant (1-\eta)^{2s}\,\tilde{\mathbb{E}}\left[\langle u, a_i\rangle^2\right] \leqslant \sigma^{-1}\cdot\sigma/(n-i_0+1) \leqslant 1/3$. Thus by constraint (9.13) we have $\tilde{\mathbb{E}}\left[\sum_{i>i_0}\langle u, a_i\rangle^{2s+2}\right] \geqslant 1/3$. Therefore, there exists $i^\star > i_0$ such that $\tilde{\mathbb{E}}\left[\langle u, a_{i^\star}\rangle^{2s+2}\right] \geqslant \frac{1}{3(n-i_0+1)}$. Then by Theorem 9.1 we obtain that with $1/d^{O(s^3)}$ probability, in each step of the inner loop we can find $\hat{a}_i$ that is $O(\varepsilon^{1/2})$-close to $a_{i^\star}$, and therefore at the end of the inner loop with high probability we found a new vector $\hat{a}_{i_0+1}$ which is close $O(\varepsilon^{1/2})$-close to $a_{i^\star}$.

□

10 Fast orthogonal tensor decomposition without sum-of-squares

In this section, we give an algorithm (see Theorem 10.2) with quasi-linear running time (in the size of the input) that finds a component of an orthogonal 3-tensor in the presence of spectral norm error at most $1/\log d$. The previous best known algorithm for orthogonal 3-tensor is by [AGH+14, Theorem 5.1] which takes similar runtime and tolerates $1/d$ error in injective norm. It is known that for any symmetric tensor E the spectral norm can be bounded by injective norm with multiplicative factor $\sqrt{d}$, that is, $\|E\|_{\{1\}\{2,3\}} \leqslant \sqrt{d}\cdot\|E\|_{\{1\}\{2\}\{3\}}$. Therefore, our robustness guarantee is at least $\sqrt{d}$ factor better than tensor power method.

The key step of Algorithm is the following simple Theorem that finds a single component. It is in fact an analog of Theorem 4.1 without sum-of-squares. Here we analyze the success probability much more carefully for achieving quasi-linear time.

Theorem 10.1. *Let $a_1, \dots, a_n \in \mathbb{R}^d$ be orthonormal vectors. Let $T \in (\mathbb{R}^d)^{\otimes 3}$ be a symmetric 3-tensor such that $\|T - \sum_i a_i^{\otimes 3}\|_{\{1\},\{2,3\}} \leqslant \tau$. Let g be a standard d-dimensional Gaussian vector. Let $\delta \in [0,1]$. Then, with probability $1/(d^{1+\delta}(\log d)^{O(1)})$ over the choice of g, the top eigenvector of the following matrix is $O(\tau/\delta)$-close to a_1,*

$$M_g := (\mathrm{Id}\otimes\mathrm{Id}\otimes g^{\mathsf{T}})T\,.$$

At the same time, the ratio between the top eigenvalue and the second largest eigenvalue in absolute value is at least $1+\delta/3-O(\tau)$.

Proof. Let $E = T - \sum_i a_i^{\otimes 3}$. Then,

$$M_g = (\mathrm{Id} \otimes \mathrm{Id} \otimes g^\mathsf{T})E + \sum_{i=1}^{n} \langle g, a_i \rangle \cdot a_i^{\otimes 2}\,, \tag{10.1}$$

Since E is symmetric and $\|E\|_{\{1\},\{2,3\}} \leqslant \tau$, Theorem 6.5 implies that with probability at least $1 - 1/d^2$ over the choice of g,

$$\left\|(\mathrm{Id} \otimes \mathrm{Id} \otimes g^\mathsf{T})E\right\|_{\{1\},\{2\}} \leqslant 2(\log d)^{1/2}\tau\,. \tag{10.2}$$

Let $t = \sqrt{2 \log d}$. By the fact that $\langle g, a_1 \rangle, \ldots, \langle g, a_n \rangle$ are independent standard Gaussian variables and standard estimates on their cumulative density function, the following event happens with probability at least $\frac{1}{d^{(1+\delta)} \cdot (\log d)^{O(1)}}$

$$\langle g, a_1 \rangle \geqslant (1 + \delta/3) \cdot t \quad \text{and} \quad \max_{i \in \{2,\ldots,n\}} |\langle g, a_i \rangle| \leqslant t \tag{10.3}$$

Conditioned on the events in (10.2) and (10.3), we have the following bound on the spectral norms of M_g and $M_g - \delta/3 \cdot t \cdot a_1^{\otimes 2}$, which implies that the top eigenvector of M_g is $O(\tau/\delta)$-close to a_1 (by [HSSS16, Lemma A.1]),

$$\begin{aligned}
&\left\|M_g\right\|_{\{1\},\{2\}} - \left\|M_g - \frac{1}{3}\delta t \cdot a_1^{\otimes 2}\right\|_{\{1\},\{2\}} && (10.4)\\
&\geqslant \left\|\sum_{i=1}^{n} \langle g, a_i \rangle \cdot a_i a_i^\mathsf{T}\right\| - \left\|(\langle g, a_1 \rangle - \frac{1}{3}\delta t) \cdot a_1 a_1^\mathsf{T} + \sum_{i=2}^{n} \langle g, a_i \rangle \cdot a_i a_i^\mathsf{T}\right\| - 2(\log n)^{1/2}\tau \\
&\qquad \text{(conditioned on event in (10.2))} \\
&\geqslant \frac{1}{3}\delta t - 2(\log d)^{1/2}\tau && \text{(conditioned on event in (10.3))} \\
&\geqslant (1 - O(\tau/\delta)) \cdot \frac{1}{3}\delta t\,. && (10.5)
\end{aligned}$$

The probability that the events of (10.2) and (10.3) happen simultaneously is at least

$$\frac{1}{d^{1+\delta}(\log d)^{O(1)}} - \frac{1}{d^2} \geqslant \frac{1}{d^{1+\delta}(\log d)^{O(1)}}\,.$$

This bound implies the first part of the theorem. To see the eigengap bound, we first observe that the largest eigenvalue of M_g is at least $\langle g, a_1 \rangle - \left\|(\mathrm{Id} \otimes \mathrm{Id} \otimes g^\mathsf{T})E\right\|_{\{1\},\{2\}} \geqslant (1 + \delta/3)t - 2(\log d)^{1/2}\tau$. On the other hand, by eigenvalue interlacing, the second larges eigenvalue of M_g is bounded by the top eigenvalue of $M_g - \langle g, a_1 \rangle a_1 a_1^\mathsf{T} = (\mathrm{Id} \otimes \mathrm{Id} \otimes g^\mathsf{T})E + \sum_{i=2}^{n} \langle g, a_i \rangle \cdot a_i^{\otimes 2}$, which in turn is bounded above by $2(\log d)^{1/2}\tau + t$. Therefore the eigenvalue gap statement follows by recalling $t = \sqrt{2 \log d}$. □

We remark that we can amplify the success probability of the algorithm by running it repeatedly with independent randomness.

Theorem 10.2. *There exists a randomized algorithm with running time $d^3 \cdot (\log d)^{O(1)}$ that given a symmetric 3-tensor $T \in (\mathbb{R}^d)^{\otimes 3}$ such that $\|T - \sum_{i=1}^{n} a_i^{\otimes 3}\|_{\{1\},\{2,3\}} \leqslant 1/\log d$ for some set of orthonormal vector $\{a_1, \ldots, a_n\} \subseteq \mathbb{R}^d$ outputs with probability $\Omega(1)$ a vector unit v such that*

$$\min_{i \in [n]} \|v - a_i\|^2 \leqslant \frac{1}{2^d} + \left\|T - \sum_{i=1}^{n} a_i^{\otimes 3}\right\|_{\{1\},\{2\},\{3\}}\,.$$

Furthermore, there exists a randomized algorithm with running time $d^{1+\omega} \cdot (\log d)^{O(1)} \leqslant O(d^{3.33})$ *that given* T *as before, with probability at least* $\Omega(1)$ *outputs a set of vectors* $\{a'_1, \ldots, a'_n\}$ *with Hausdorff distance at most* $\frac{1}{2^d} + \left\| T - \sum_{i=1}^n a_i^{\otimes 3} \right\|_{\{1\},\{2\},\{3\}}$ *from* $\{a_1, \ldots, a_n\}$. *Here* ω *is the matrix multiplication exponent.*

Proof. We may assume that d is larger than some constant. We run $d^{1+\omega} \cdot O(\log d)^{O(1)}$ iterations of the following procedure which can be carried out in $\tilde{O}(d^3)$ time. We will discuss how to speed the algorithm at the end.

1. Choose a standard Gaussian vector g and compute $M_g = (\mathrm{Id} \otimes \mathrm{Id} \otimes g^\mathsf{T})T$.
2. Run $O(\log d)^2$ iterations of the matrix power method on M_g viewed as a d-by-d matrix (with random initialization) and set u to be the top eigenvector calculated in this way.
3. Check that $|\langle u^{\otimes 3}, T\rangle| \geqslant 0.9$.
4. Run $O(\log \log d)$ iterations of the tensor power method on T starting from u. Output the final iterate v of the method.

The analysis of the tensor power method [AGH+14, Lemma 5.1] shows that whenever the check in step 3 succeeds then the final output v satisfies the desired accuracy guarantee of the theorem. It remains to show that the check in step 3 succeeds with probability at least $1/(\log d)^{O(1)}$ over the randomness of the algorithm. (We obtain success probability $\Omega(1)$ by repeating the algorithm $(\log d)^{O(1)}$ times.) We apply Theorem 10.1 for $\delta = O(1/\log d)$ and $\tau = 1/\log d$ such that for every $i \in [n]$, the distance guarantee for the top eigenvector of M_g is at most 0.001 and the ratio between first and second eigenvalue is at least $1 + 1/\log d$. By symmetry, for every index $i \in [n]$, the probability that the top eigenvector of M_g is 0.001-close to a_i is at least $1/d(\log d)^{O(1)}$. Since the vectors $a_1, \ldots, a_n$ are orthonormal these events are disjoint. Therefore, with probability at least $1/(\log d)^{O(1)}$ over the choice of g, the top eigenvector of M_g is 0.001-close to one of the vectors $a_1, \ldots, a_n$. Since the multiplicative gap between the top eigenvalue and the remaining eigenvalues of M_g is at least $1 + 1/\log d$ (by Theorem 10.1 for our choice of δ and τ), it follows that with constant probability over the choice of the random initialization of the matrix power method, the second step of the algorithm recovers a vector that is 0.001-close to the top eigenvector of M_g. In this case, the resulting vector u satisfies the check $|\langle u^{\otimes 3}, T\rangle| \geqslant 0.9$.

In order to find all components in time $d^{1+\omega} \cdot O(\log d)^{O(1)}$ we run $d \cdot (\log d)^{O(1)}$ independent evaluations of the above algorithm. Note that each run involves multiplication of a $d^2 \times d$ matrix with a d dimensional vector and therefore in total we are to multiply a $d^2 \times d$ matrix with $d \times d$ matrix. Therefore, using fast matrix multiplication, we can "parallelize" all of the required linear algebra operations and speedup the running time from $d^4(\log d)^{O(1)}$ to the desired $O(d^{1+\omega}) \cdot (\log d)^{O(1)}$. □

Remark 10.3 (Extension to other settings). The same rounding idea in Theorem 10.1 can be extended to the setting when the components $a_1, \ldots, a_n$ are close to isotropic in the sense that $\left\| \sum_i a_i a_i^\mathsf{T} - \mathrm{Id}_d \right\| \leqslant \sigma$. The success probability will decrease to roughly $1/d^{1+\mathrm{poly}(\sigma)}$, and therefore when σ is at most a constant, the overall runtime will remain polynomial in d.

Suppose $a_1, \ldots, a_n$ are separate vectors as in the setting of Theorem 1.5, we can apply the idea in paragraph above to the 3-tensor $\sum_i b_i^{\otimes 3}$ where $b_i = a_i^{\otimes k/3}$ and $k \geqslant O\left(\frac{1+\log \sigma}{\log \rho}\right) \cdot \log(1/\eta)$ is a multiple

of 3. By Lemma 5.4 and the condition on k, we have that b_i are in nearly isotropic position with $\left\|\sum_i b_i b_i^\top\right\| \leqslant 1 + \eta$. Hence, using idea above we have a spectral algorithm without sum-of-squares for this setting. As noted before (below Theorem 1.5), the error tolerance of this algorithm is in terms of an *unbalanced* spectral norm: $\|T - \sum_{i=1}^n a_i^{\otimes 2k}\|_{\{1,\ldots,2k/3\},\{2k/3+1,\ldots,2k\}}$, which limits its application, for example, to dictionary learning.

References

[AFH+15] Anima Anandkumar, Dean P. Foster, Daniel J. Hsu, Sham M. Kakade, and Yi-Kai Liu, *A spectral algorithm for latent dirichlet allocation*, Algorithmica **72** (2015), no. 1, 193–214. 1

[AGH+14] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, Sham M. Kakade, and Matus Telgarsky, *Tensor decompositions for learning latent variable models*, Journal of Machine Learning Research **15** (2014), no. 1, 2773–2832. 1, 46, 48

[AGH+15] Anima Anandkumar, Rong Ge, Daniel J. Hsu, Sham M. Kakade, and Matus Telgarsky, *Tensor decompositions for learning latent variable models (A survey for ALT)*, ALT, Lecture Notes in Computer Science, vol. 9355, Springer, 2015, pp. 19–38. 1

[AGHK14] Animashree Anandkumar, Rong Ge, Daniel J. Hsu, and Sham M. Kakade, *A tensor approach to learning mixed membership community models*, Journal of Machine Learning Research **15** (2014), no. 1, 2239–2312. 1

[AGJ14] Anima Anandkumar, Rong Ge, and Majid Janzamin, *Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models*, CoRR **abs/1411.1488** (2014). 1

[AGJ15] Animashree Anandkumar, Rong Ge, and Majid Janzamin, *Learning overcomplete latent variable models through tensor methods*, COLT, JMLR Workshop and Conference Proceedings, vol. 40, JMLR.org, 2015, pp. 36–112. 1

[BCMV14] Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan, *Smoothed analysis of tensor decompositions*, STOC, ACM, 2014, pp. 594–603. 1, 2, 4, 38, 39

[BHK+16] Boaz Barak, Samuel B. Hopkins, Jonathan Kelner, Pravesh K. Kothari, Ankur Moitra, and Aaron Potechin, *A nearly tight sum-of-squares lower bound for the planted clique problem*, FOCS, IEEE Computer Society, 2016. 2

[BKS15] Boaz Barak, Jonathan A. Kelner, and David Steurer, *Dictionary learning and tensor decomposition via the sum-of-squares method*, STOC, ACM, 2015, pp. 143–151. 1, 2, 6, 7, 8, 9, 19, 21

[BPT13] Grigoriy Blekherman, Pablo A. Parrilo, and Rekha R. Thomas (eds.), *Semidefinite optimization and convex algebraic geometry.*, Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 2013 (English). 54

[Cim12] J. Cimprič, *Real algebraic geometry for matrices over commutative rings*, Journal of Algebra **359** (2012), 89 – 103. 16

[DK70] Chandler Davis and W. M. Kahan, *The rotation of eigenvectors by a perturbation. iii*, SIAM Journal on Numerical Analysis **7** (1970), no. 1, 1–46. 33, 53

[dlPG99] Víctor H de la Peña and Evarist Giné, *Decoupling. from dependence to independence. randomly stopped processes. u-statistics and processes. martingales and beyond, probability and its applications*, 1999. 39

[EA06] Michael Elad and Michal Aharon, *Image denoising via sparse and redundant representations over learned dictionaries*, Image Processing, IEEE Transactions on **15** (2006), no. 12, 3736–3745. 6

[EP07] Andreas Argyriou Theodoros Evgeniou and Massimiliano Pontil, *Multi-task feature learning*, Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference, vol. 19, MIT Press, 2007, pp. 41–48. 6

[Gha14] Shayan Oveis Gharan, *New rounding techniques for the design and analysis of approximation algorithms*, Ph.D. thesis, STANFORD UNIVERSITY, 2014. 2, 9

[GHK15] Rong Ge, Qingqing Huang, and Sham M. Kakade, *Learning mixtures of gaussians in high dimensions*, STOC, ACM, 2015, pp. 761–770. 1

[GLS81] Martin Grötschel, László Lovász, and Alexander Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica **1** (1981), no. 2, 169–197. 14, 15, 54

[GM15] Rong Ge and Tengyu Ma, *Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms*, APPROX-RANDOM, LIPIcs, vol. 40, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2015, pp. 829–849. 1, 2, 4, 8, 12, 30, 31

[GVX14] Navin Goyal, Santosh Vempala, and Ying Xiao, *Fourier PCA and robust tensor decomposition*, STOC, ACM, 2014, pp. 584–593. 1

[Har70] Richard A Harshman, *Foundations of the parafac procedure: Models and conditions for an" explanatory" multi-modal factor analysis*. 1, 8

[Hås90] Johan Håstad, *Tensor rank is np-complete*, J. Algorithms **11** (1990), no. 4, 644–654. 1

[HL13] Christopher J. Hillar and Lek-Heng Lim, *Most tensor problems are np-hard*, J. ACM **60** (2013), no. 6, 45. 1

[HSS15] Samuel B. Hopkins, Jonathan Shi, and David Steurer, *Tensor principal component analysis via sum-of-square proofs*, COLT, JMLR Workshop and Conference Proceedings, vol. 40, JMLR.org, 2015, pp. 956–1006. 1, 4

[HSSS16] Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer, *Fast spectral algorithms from sum-of-squares proofs: Tensor decomposition and planted sparse vectors*, Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing (New York, NY, USA), STOC 2016, ACM, 2016, pp. 178–191. 1, 2, 3, 4, 12, 31, 32, 47

[Las01] Jean B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim. **11** (2000/01), no. 3, 796–817. MR 1814045 (2002b:90054) 14

[LCC07] Lieven De Lathauwer, Joséphine Castaing, and Jean-François Cardoso, *Fourth-order cumulant-based blind identification of underdetermined mixtures*, IEEE Trans. Signal Processing **55** (2007), no. 6-2, 2965–2973. 4, 10, 32

[LRA93] S.E. Leurgans, R.T. Ross, and R.B. Abel, *A decomposition for three-way arrays.*, SIAM J. Matrix Anal. Appl. **14** (1993), no. 4, 1064–1083 (English). 1, 8

[LRS15] James R. Lee, Prasad Raghavendra, and David Steurer, *Lower bounds on the size of semidefinite programming relaxations*, Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing (New York, NY, USA), STOC '15, ACM, 2015, pp. 567–576. 2

[MLB+08] Julien Mairal, Marius Leordeanu, Francis Bach, Martial Hebert, and Jean Ponce, *Discriminative sparse image models for class-specific edge detection and image interpretation*, Computer Vision–ECCV 2008, Springer, 2008, pp. 43–56. 6

[MRBL07] Y Marc'Aurelio Ranzato, Lan Boureau, and Yann LeCun, *Sparse feature learning for deep belief networks*, Advances in neural information processing systems **20** (2007), 1185–1192. 6

[OF96a] Bruno A Olshausen and David J Field, *Emergence of simple-cell receptive field properties by learning a sparse code for natural images*, Nature **381** (1996), no. 6583, 607–609. 6

[OF96b] _____, *Natural image statistics and efficient coding**, Network: computation in neural systems **7** (1996), no. 2, 333–339. 6

[OF97] Bruno A. Olshausen and David J. Field, *Sparse coding with an overcomplete basis set: A strategy employed by v1?*, Vision Research **37** (1997), no. 23, 3311 – 3325. 6

[Oli10] Roberto I. Oliveira, *Sums of random Hermitian matrices and an inequality by Rudelson.*, Electron. Commun. Probab. **15** (2010), 203–212 (English). 2, 10, 28

[Par00] Pablo A Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, Ph.D. thesis, Citeseer, 2000. 14

[RV09] Mark Rudelson and Roman Vershynin, *Smallest singular value of a random rectangular matrix*, Communications on Pure and Applied Mathematics **62** (2009), no. 12, 1707–1739. 39

[Sho87] N. Z. Shor, *An approach to obtaining global extrema in polynomial problems of mathematical programming*, Kibernetika (Kiev) (1987), no. 5, 102–106, 136. MR 931698 (89d:90202) 14

[ST04] Daniel A. Spielman and Shang-Hua Teng, *Smoothed analysis of algorithms: Why the simplex algorithm usually takes polynomial time*, J. ACM **51** (2004), no. 3, 385–463. 38

[SW15] Tselil Schramm and Benjamin Weitz, *Low-rank matrix completion with adversarial missing entries*, CoRR **abs/1506.03137** (2015). 5

[Tro12] Joel A. Tropp, *User-friendly tail bounds for sums of random matrices*, Foundations of Computational Mathematics **12** (2012), no. 4, 389–434. 2, 28

[YWHM08] Jianchao Yang, John Wright, Thomas Huang, and Yi Ma, *Image super-resolution as sparse representation of raw image patches*, Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, IEEE, 2008, pp. 1–8. 6

[YWS15] Yi Yu, Tengyao Wang, and Richard J Samworth, *A useful variant of the davis–kahan theorem for statisticians*, Biometrika **102** (2015), no. 2, 315–323. 53

A Toolbox

Lemma A.1 (sums-of-squares proof for Cauchy-Schwarz inequality)**.** *Let $x_1, \dots, x_n$ and $y_1, \dots, y_n$ be polynomials in some indeterminates. Then*

$$\vdash \left\{ \left(\sum_{i=1}^{n} x_i y_i \right)^2 \leqslant \left(\sum_{i=1}^{n} x_i^2 \right) \left(\sum_{i=1}^{n} y_i^2 \right) \right\} .$$

Proof. The difference between the RHS and the LHS is a sum of squares.

$$\left(\sum_{i=1}^{n} x_i^2 \right) \left(\sum_{i=1}^{n} y_i^2 \right) - \left(\sum_{i=1}^{n} x_i y_i \right)^2 = \sum_{i,j} (x_i y_j - x_j y_i)^2 . \qquad \square$$

Lemma A.2. *Let u be indeterminate and a be a unit vector. Let $\mathcal{A} = \{\|u\|^2 = 1, \langle u, a\rangle^2 \leqslant \tau\}$. Then for any unit vector b such that $\|a - b\| \leqslant 2\delta$, we have that*

$$\mathcal{A} \vdash \{\langle u, b\rangle^2 \leqslant \left(\sqrt{\tau} + \sqrt{\delta}\right)^2\} .$$

Proof. First of all, by Lemma A.3, $\mathcal{A} \vdash \langle u, a\rangle \leqslant \sqrt{\tau}$ and $\mathcal{A} \vdash \langle u, a\rangle \geqslant -\sqrt{\tau}$. To bound $\langle u, b\rangle$, we decompose u and b into their components parallel to and perpendicular to a.

Let $u' = u - \langle u, a\rangle a$, so that u' is a vector polynomial in u that satisfies $u = \langle u, a\rangle a + u'$ and $\langle u', a\rangle = 0$. Then $\|u\|^2 = \|u'\|^2 + \langle u, a\rangle^2 \|a\|^2$ and therefore $\mathcal{A} \vdash \|u'\|^2 \leqslant 1$.

Similarly, let $b' = b - \langle b, a\rangle a$, so that $b = \langle b, a\rangle a + b'$ and $\langle b', a\rangle = 0$. Since $\|b - a\|^2 \leqslant 2\delta$, we have $\langle b, a\rangle \geqslant 1 - \delta$, meaning that $\|b'\|^2 = \|b\|^2 - \langle b, a\rangle \|a\|^2 \leqslant 1 - (1 - \delta) = \delta$.

Then we are ready to bound $\langle u, b\rangle$:

$$\begin{aligned}\langle u, b\rangle &= \left\langle \langle u, a\rangle a + u', \langle b, a\rangle a + b' \right\rangle \\ &= \langle u, a\rangle\langle b, a\rangle + \langle u', b'\rangle\,.\end{aligned}$$

Since $\mathcal{A} \vdash \langle u, a\rangle\langle b, a\rangle \leqslant \sqrt{\tau}$ and also $\langle u', b\rangle^2 \leqslant \|u'\|^2\|b'\|^2 \leqslant \delta\|u'\|^2$ which in turn implies (by $\mathcal{A} \vdash \|u'\|^2 \leqslant 1$ and Lemma A.3) that $\mathcal{A} \vdash \langle u', b'\rangle \leqslant \sqrt{\delta}$, we conclude that $\mathcal{A} \vdash \langle u, b\rangle \leqslant \sqrt{\tau} + \sqrt{\delta}$.

Similarly, $\mathcal{A} \vdash \langle u, b\rangle \geqslant -\sqrt{\tau} - \sqrt{\delta}$. Hence

$$\langle u, b\rangle^2 - \left(\sqrt{\tau} + \sqrt{\delta}\right)^2 = \left(\langle u, b\rangle - \left(\sqrt{\tau} + \sqrt{\delta}\right)\right)\left(\langle u, b\rangle + \left(\sqrt{\tau} + \sqrt{\delta}\right)\right) \leqslant 0\,,$$

as desired. □

Lemma A.3. *For a positive real number a, and x be an indeterminate, then we have that*

$$\left\{x^2 \leqslant a^2\right\} \vdash \{x \leqslant a, x \geqslant -a\} \tag{A.1}$$

Proof. The first statement simply follows from the following two polynomial identities,

$$a - x = \tfrac{1}{2a}\left(a^2 - x^2 + (a - x)^2\right)\,,$$

and similarly,

$$x + a = \tfrac{1}{2a}\left(a^2 - x^2 + (a + x)^2\right)\,.$$

□

Theorem A.4 (Consequence of Davis-Kahan Theorem [DK70]. c.f [YWS15])**.** *Let $\Sigma, \hat{\Sigma}$ be symmetric matrices in $\mathbb{R}^{d\times d}$. Let $v_1, \tilde{v}_1$ be their top eigenvector respectively and let $\lambda_1 \geqslant \lambda_2 \ldots$ and $\hat{\lambda}_1 \geqslant \hat{\lambda}_2 \ldots$ be their eigenvalues, respectively. Then,*

$$\|v_1 - \hat{v}_1\| \leqslant \frac{\sqrt{2}\|\Sigma - \hat{\Sigma}\|}{|\lambda_1 - \hat{\lambda}_2|}\,.$$

Lemma A.5 (Consequence of Theorem A.4)**.** *Let a be unit vector and $\mathrm{Id}_1 = aa^\top$, $\mathrm{Id}_{-1} = \mathrm{Id} - aa^\top$. Suppose symmetric matrix M satisfies that*

$$\max\left\{\|\mathrm{Id}_{-1}\, M\, \mathrm{Id}_{-1}\|\,, \|\mathrm{Id}_1\, M\, \mathrm{Id}_{-1}\|\right\} \leqslant \varepsilon a^\top M a$$

Then, a is $3\sqrt{2}\varepsilon$-close to the top eigenvector of M in Euclidean distance.

Proof. Let $t = a^\top M a$ and $\widehat{M} = taa^\top = \mathrm{Id}_1\, M\, \mathrm{Id}_1$. Then we have that $M = (\mathrm{Id}_1 + \mathrm{Id}_{-1})M(\mathrm{Id}_1 + \mathrm{Id}_{-1}) = \widehat{M} + \mathrm{Id}_1\, M\, \mathrm{Id}_{-1} + \mathrm{Id}_{-1}\, M\, \mathrm{Id}_1 + \mathrm{Id}_{-1}\, M\, \mathrm{Id}_{-1}$. Therefore by the assumption we have $\|\widehat{M} - M\| \leqslant 3\varepsilon t$. Therefore using Theorem A.4 with $\Sigma = M$ and $\hat{\Sigma} = \widehat{M}$ we obtain that the top eigenvector of M is $3\sqrt{2}\varepsilon$-close to M in Euclidean distance. □

B Missing proofs in Section 3

Proof of Theorem 3.3. Suppose $\mathcal{A} = \{f_1 \geqslant 0, \dots, f_m \geqslant 0\}$ with $f_1, \dots, f_m \in \mathbb{R}[x]$ and $\deg(f) \leqslant d$. Let $m = |\mathcal{A}|$ and let $\ell = \max_{i\in[m]} \deg(f_i)$. We use an optimization algorithm to find such pseudo-distribution D. Here the variables are all the moments $\tilde{\mathbb{E}}_{D(x)} [\prod_{i\in S} x_i]$ for all S with $|S| \leqslant d$. The constraints are linear constraints over these variables

$$\left\{ \tilde{\mathop{\mathbb{E}}_{D}} \left(\prod_{i\in S} f_i \right) h^2 \geqslant 0 \;\middle|\; S \subseteq [m],\, h \in \mathbb{R}[x],\, |S|\ell + \deg\left(h^2\right) \leqslant d \right\} .$$

We can separate over these constraints in time $(n+m)^{O(d)}$. Indeed, for every fixed choice of S, the set of constraints of the form $\tilde{\mathbb{E}}_D \left(\prod_{i\in S} f_i\right) h^2 \geqslant 0$ may be written as a single matrix constraint $\tilde{\mathbb{E}}_D \left(\prod_{i\in S} f_i\right) \left[(1,x)^{\otimes d-|S|\ell}\right] \left[(1,x)^{\otimes d-|S|\ell}\right]^{\mathsf{T}} \succeq 0$, with the equivalence established by mapping h to a vector of coefficients. Therefore, by Theorem 3.1 and the equivalence of optimization and separation [GLS81], we can find moments $\tilde{\mathbb{E}}_{D(x)}(1,x)^{\otimes d}$ of a degree-d pseudo-distribution in time $(n+m)^{O(d)}$. A standard multivariate polynomial interpolation argument allows us to recover the underlying pseudo-distribution D [BPT13]. □

Proof of Lemma 3.4. Suppose D is a degree-d pseudo-distribution. Let $\mathcal{A} = \{f_1 \geqslant 0, \dots, f_n \geqslant 0\}$ and let $\mathcal{B} = \{g_1 \geqslant 0, \dots, g_m \geqslant 0\}$. Moreover, $\mathcal{A} \vdash_{\ell'} \mathcal{B}$ means that for every constraint $\{g_j \geqslant 0\}$ in $\mathcal{B}$, there are sums-of-squares polynomials $p_{j,S}$ for each $S \subseteq [n]$ such that $g_j = \sum_{S\subseteq[n]} p_{j,S} \prod_{i\in S} f_i$ where each summand $p_{j,S} \prod_{i\in S} f_i$ has degree at most ℓ'.

Consider some set $T \subset [m]$ and some sum-of-squares polynomial h' such that $|T|\ell\ell' + \deg h' \leqslant d$. We would like to show that

$$\tilde{\mathop{\mathbb{E}}_{D}} \left(\prod_{j\in T} g_j \right) h' \geqslant 0 . \tag{B.1}$$

$\mathcal{A} \vdash_{\ell'} \mathcal{B}$ means that for every constraint $\{g_j \geqslant 0\}$ in $\mathcal{B}$, there are sums-of-squares polynomials $p_{j,S}$ for each $S \subseteq [n]$ such that $g_j = \sum_{S\subseteq[n]} p_{j,S} \prod_{i\in S} f_i$ where each summand $p_{j,S} \prod_{i\in S} f_i$ has degree at most ℓ'. Substituting g_j in equation (B.1), it suffices to show that

$$\tilde{\mathop{\mathbb{E}}_{D}} \left(\prod_{j\in T} \left(\sum_{S\subseteq[n]} p_{j,S} \prod_{i\in S} f_i \right) \right) h' \geqslant 0 . \tag{B.2}$$

We expand the outer product over T and see that the polynomial inside the pseudo-expectation is in fact a sum of many polynomials of the form $q_1 \dots q_{|T|} \left(\prod_{i\in W} f_i\right) h'$, where each of the q_i is equal to $p_{j,S}$ for some $j \in T$ and some $S \subset [n]$, and where $W \subset S$ is a multi-set, with $\deg\left(q_1 \dots q_{|T|} \left(\prod_{i\in W} f_i\right)\right) \leqslant |T|\ell$. Moreover, we note that since each q_i is a sum of squares, $q_1 \dots q_{|T|} \left(\prod_{i\in W} f_i\right)$ can be written as $q \prod_{i\in W'} f_i$ where q is a sum of squares and W' is the set of elements that appear in W an odd number of times. We calculate

$$\deg(q) = \deg(q_1 \dots q_{|T|}) + \deg\left(\prod_{i\in W\setminus W'} f_i \right)$$

$$\leqslant (|T|\ell - |W|) + (|W| - |W'|)\ell'$$
$$\leqslant |T|\ell\ell' - |W'|\ell' ,$$

where we used $\deg\left(q_1 \dots q_{|T|}\left(\prod_{i\in W} f_i\right)\right) \leqslant |T|\ell$ in combination with $\deg\left(\prod_{i\in W} f_i\right) \geqslant |W|$, along with the fact that therefore $|W| \leqslant |T|\ell$. Therefore since qh' is a sum of squares and $|W'|\ell' + \deg(qh') \leqslant d$, by the definition of $D \models_\ell \mathcal{A}$, we have $\tilde{\mathbb{E}}\left[q_1 \dots q_{|T|}\left(\prod_{i\in W} f_i\right) h'\right] = \tilde{\mathbb{E}}\left[qh'\left(\prod_{i\in W'} f_i\right)\right] \geqslant 0$. Then by linearity of pseudo-expectation we prove equation (B.2), which completes the proof. □

Proof of Lemma 3.5. We prove the contrapositive. Let $\mathcal{A} = \{f_1 \geqslant 0, \dots, f_m \geqslant 0\}$. Assume that $\mathcal{A} \not\vdash_d \{g \geqslant -\varepsilon\}$ for some $\varepsilon > 0$.

A polynomial h satisfies $\mathcal{A} \vdash_d \{h \geqslant 0\}$ precisely when $h = \sum_{S\subset[m]} p_S \prod_{i\in S} f_i$ for some sum-of-squares polynomials p_S where the degree of each summand is at most d. We observe that $\mathcal{H} = \{h \mid \mathcal{A} \vdash_d \{h \geqslant 0\}\}$ is a convex cone. Let $\bar{\mathcal{H}}$ be its closure. We argue that $g \notin \bar{\mathcal{H}}$.

Indeed, if there exists a sequence of polynomial g_k that converges to g (in coefficients), then there exists a sufficiently large K such that for $k \in K$, $\{\|x\|^2 \leqslant B\} \vdash g_k(x) - g(x) \leqslant \varepsilon/2$. Therefore $\mathcal{A} \vdash g + \varepsilon = g_k + (g - g_k + \varepsilon) \geqslant 0$. This contradicts our assumption.

Then by the hyperplane separation theorem, there exists a linear functional L over the space of all degree-d polynomials such that $L[g] < 0$ and $L[h] > 0$ for all $h \in \mathcal{H}$. Since $1 \in \mathcal{H}$, we have $L(1) \geqslant 0$. We can scale L properly so that $L(1) = 1$ and therefore L defines a pseudo-distribution D. In particular, D is a pseudo-distribution such that $D \models_\ell \mathcal{A}$ because $(\prod_{i\in S} f_i)h \in \mathcal{H}$ holds whenever $|S|\ell + \deg(h) \leqslant d$ and thus $\tilde{\mathbb{E}}_D\left[(\prod_{i\in S} f_i)h\right] \geqslant 0$. However, we also have $D \not\models_{\ell'} \mathcal{B}$ since $L(g) < 0$. □

Proof of Lemma 3.7. Let $\mathcal{A} = \{f_1 \geqslant 0, \dots, f_k \geqslant 0\}$. For any vector $z \in \mathbb{R}^p$, we prove $\langle z, \tilde{\mathbb{E}}[M] z\rangle = \tilde{\mathbb{E}}\left[z^\intercal M z\right] \geqslant 0$. Indeed, $\mathcal{A} \vdash_\ell \mathcal{M}$ implies the existence of q_i, v_i's that satisfy equation (3.1), where q_i can be written as $q_i = \sum_S p_{i,S} \prod_{j\in S} f_j$. Therefore, $\tilde{\mathbb{E}}\left[z^\intercal M z\right] = \tilde{\mathbb{E}}\left[\sum_i \sum_S \langle z, v_i(x)\rangle^2 p_{i,S} \prod_{j\in S} f_j\right]$. For fixed i, S, we have that $\deg(\langle z, v_i(x)\rangle^2) = 2\deg(v_i)$, and $|S| \leqslant \ell' - 2\deg(v_i)$. Therefore, we have $|S|\ell + 2\deg(v_i) \leqslant \ell\ell' \leqslant d$, and by $D \models_\ell \mathcal{A}$ we obtain that $\tilde{\mathbb{E}}\left[\sum_i \sum_S \langle z, v_i(x)\rangle^2 p_{i,S} \prod_{j\in S} f_j\right] \geqslant 0$, which completes the proof. □