Decomposing Overcomplete 3rd Order Tensors using Sum-of-Squares Algorithms

Rong Ge[1] **and Tengyu Ma**[2]

1 Microsoft Research
1 Memorial Dr, Cambridge MA, USA
rongge@microsoft.com
2 Computer Science Department
Princeton University, USA
tengyu@cs.princeton.edu

Abstract

Tensor rank and low-rank tensor decompositions have many applications in learning and complexity theory. Most known algorithms use unfoldings of tensors and can only handle rank up to $n^{\lfloor p/2 \rfloor}$ for a p-th order tensor in $\mathbb{R}^{n^p}$. Previously no efficient algorithm can decompose 3rd order tensors when the rank is super-linear in the dimension. Using ideas from sum-of-squares hierarchy, we give the first quasi-polynomial time algorithm that can decompose a random 3rd order tensor decomposition when the rank is as large as $n^{3/2}/\operatorname{poly}\log n$.

We also give a polynomial time algorithm for certifying the injective norm of random low rank tensors. Our tensor decomposition algorithm exploits the relationship between injective norm and the tensor components. The proof relies on interesting tools for decoupling random variables to prove better matrix concentration bounds.

1998 ACM Subject Classification I.2.6 Learning

Keywords and phrases sum of squares, overcomplete tensor decomposition

Digital Object Identifier 10.4230/LIPIcs.APPROX-RANDOM.2015.829

1 Introduction

Tensors, as natural generalization of matrices, are often used to represent multi-linear relationships or data that involves higher order correlation. A p-th order tensor $T \in \mathbb{R}^{n^p}$ is a p-dimensional array indexed by $[n]^p$. A tensor T is rank-1 if it can be written as the outer-product of p vectors $T = a_1 \otimes \cdots \otimes a_p$, where $a_i \in \mathbb{R}^n$ (for $i = 1, \ldots, p$). Equivalently, $T_{i_1,\ldots,i_p} = \prod_{j=1}^{p} a_j(i_j)$ where $a_j(i_j)$ denotes the i_j-th entry of vector a_j.

Low rank tensors – similar to low rank matrices – are widely used in many applications. The rank of tensor T is defined as the minimum number m such that T can be written as the sum of m rank-1 tensors. This agrees with the definition of matrix rank. However, most of the corresponding tensor problems are much harder: for $p \geq 3$ computing the rank of the tensor (as well as many related problems) is NP-hard [22, 23]. Tensor rank is also not as well-behaved as matrix rank (see for example the survey [15]).

Unlike matrices, low rank tensor decompositions are often unique [24], which is important in many applications. In special cases (especially when rank m is less than dimension n) tensor decomposition can be efficiently computed. Such specialized tensor decompositions have been the key algorithmic ideas in many recent algorithms for learning latent variable models, including mixture of Gaussians, Independent Component Analysis, Hidden Markov Model and Latent Dirichlet Allocation (see [4]). In many cases tensor decomposition can be viewed

18th Int'l Workshop on Approximation Algorithms for Combinatorial Optimization Problems (APPROX'15) / 19th Int'l Workshop on Randomization and Computation (RANDOM'15).
Editors: Naveen Garg, Klaus Jansen, Anup Rao, and José D. P. Rolim; pp. 829–849
LIPICS Leibniz International Proceedings in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

as reinterpreting previous spectral learning results [14, 26, 2, 5]. This new interpretation has also inspired many new works (e.g. [3, 13, 19]).

A common limitation in early tensor decomposition algorithms is that they only work for the undercomplete case when rank m is at most the dimension n. Although there are some attempts to decompose tensors in the overcomplete case ($m > n$) [16, 13, 7, 18, 17], these works either require at least 4-th order tensors, or is polynomial time only when in mildly overcomplete case (when m is a constant factor larger than n). In many machine learning applications, the number of samples required to accurately estimate a 4-th order tensor is too large. In practice algorithms based on 3rd order tensor are much more preferable. Therefore we are interested in the key question: are there any efficient algorithms for overcomplete 3rd order tensor decomposition?

In the worst case setting, overcomplete 3rd order tensors are not well-understood. Kruskal [24] showed the tensor decomposition is unique when the rank $m \leq 1.5n - 1$ and the components are in general position, but there is no efficient algorithm known for finding this decomposition. Constructing an explicit 3rd order tensor with rank $\Omega(n^{1+\epsilon})$ will give nontrivial circuit complexity lowerbounds [29], while the best known rank bound for an explicit 3rd order matrix is only $3n - O(\log n)$ [1].

For many of the learning applications, it is natural to consider the average case problem where the components of the tensor are chosen according to a random distribution. In this case [6] give a polynomial time algorithm that can find the true components when $m = Cn$ for any constant $C > 0$ (however the runtime depends exponentially on C).

This paper also considers this average case setting and gives a quasi-polynomial algorithm for decomposing the tensor when m can be as large as $n^{3/2}$. The main idea of the algorithm is based on sum-of-squares (SoS) SDP hierarchy ([27, 25], see Section 2 and the recent survey [12]). The main difficulty in handling overcomplete 3rd order tensors is that there is no natural *unfolding* (i.e. mapping to a matrix) that can certify the rank of the tensor. We can unfold a 4-th order tensor T into a matrix M of size $n^2 \times n^2$ where $M_{(i_1,i_2),(i_3,i_4)} = T_{i_1,i_2,i_3,i_4}$. However, unfolding 3rd order tensor will result in a very unbalanced matrix of dimension $n \times n^2$ that cannot have rank more than n. Intuitively, the power of SoS-based algorithm is that it can provide higher-order "pseudo-moments" that will allow us to use nontrivial unfoldings.

In particular, the key component of the proof is a way of certifying *injective norm* (see Section 2) of random tensors, which is closely related to the problem of certifying the 2-to-4 norm of random matrices[8]. Recently, there has been an increasing number of applications of SoS hierarchy to learning problems. [9] give algorithms for finding the sparsest vectors in a subspace, which is closely related to many learning problems. [10] give a new algorithm for dictionary learning that can handle nearly linear sparsity, and also an algorithm for robust tensor decomposition.However their result requires a tensor of high order. [11] studies a related problem of tensor prediction, also using ideas of SoS hierarchies.

1.1 Our Results

In this paper we give a quasi-polynomial time algorithm for decomposing third-order tensors when the rank m is almost as large as $n^{3/2}$ and the components of the tensor is chosen randomly. More concretely, we define $\mathcal{D}_{m,n}$ to be a distribution of third order tensors of the following form:

$$T = \sum_{i=1}^{m} a_i^{\otimes 3},$$

where the vectors $a_i \in \mathbb{R}^n$ are uniformly random vectors in $\{\pm\frac{1}{\sqrt{n}}\}^n$ and $a_i^{\otimes 3}$ is short for $a_i \otimes a_i \otimes a_i$. Our goal is to recover these components a_i's. Since any permutation of a_i's is still a valid solution, we say two decompositions are ϵ-close if they are close after an arbitrary permutation:

▶ **Definition 1** (ϵ-close)**.** Two sets of vectors $\{a_i\}_{i\in[m]}$ and $\{\hat{a}_i\}_{i\in[m]}$ in $\mathbb{R}^n$ are ϵ-close if there exists a permutation $\pi : [m] \to [m]$ such that $\|\hat{a}_{\pi(i)} - a_i\| \le \epsilon$. Two decompositions of the tensor T are ϵ-close if their components are ϵ-close.

For tensors in distribution $\mathcal{D}_{m,n}$ our algorithm can recover the decomposition as long as $m \ll n^{3/2}$.

▶ **Theorem 2.** *Given a tensor $T = \sum_{i=1}^m a_i^{\otimes 3}$ sampled from distribution $\mathcal{D}_{m,n}$, when $m \ll n^{3/2}$ there is an algorithm that runs in time $n^{O(\log n)}$ and with high probability returns a decomposition $T \approx \sum_{i=1}^m \hat{a}_i^{\otimes 3}$ that is 0.1-close to the true decomposition.*

Our result easily generalizes to many other distributions for a_i (including a uniform random vector in unit sphere or a spherical Gaussian).

The algorithm does not output a very accurate solution (the accuracy can be improved to ϵ with an exponential dependency on $1/\epsilon$). However it is known that alternating minimization algorithms can refine the decomposition once we have a nice initial point[6]:

▶ **Theorem 3** ([6])**.** *Given a tensor T from distribution $\mathcal{D}_{m,n}$ ($m \ll n^{3/2}$), and an initial solution that is 0.1-close to the true decomposition, then for any $\epsilon > 0$ (that may depend on n) there is an algorithm that runs in time $\mathrm{poly}(n, \log 1/\epsilon)$ that with high probability finds a refined decomposition that is ϵ-close to the true decomposition.*

Combining the two results we have an algorithm that runs in time $n^{O(\log n)} \mathrm{poly}\log(1/\epsilon)$ that recovers a decomposition that is component-wise ϵ-close to the true decomposition.

▶ **Corollary 4.** *Given a tensor $T = \sum_{i=1}^m a_i^{\otimes 3}$ sampled from distribution $\mathcal{D}_{m,n}$, when $m \ll n^{3/2}$ for any $\epsilon > 0$ there is an algorithm that runs in time $n^{O((\log n))} \mathrm{poly}\log(1/\epsilon)$ and with high probability returns a decomposition $T \approx \sum_{i=1}^m \hat{a}_i^{\otimes 3}$ that is ϵ-close to the true decomposition.*

The main idea in proving Theorem 2 is the observation that when the tensor is generated randomly from $\mathcal{D}_{m,n}$, the true components are close to the maximizers of the multilinear form $T(x,x,x) = \sum_{i,j,k\in[n]} T_{i,j,k}x_ix_jx_k = \sum_{i=1}^m \langle a_i, x\rangle^3$. The maximum value of $T(x,x,x)$ on unit vectors $\|x\| = 1$ is known as the injective norm of the tensor. Computing or even approximating the injective norm is known to be hard [20, 21]. A key component of our approach is a sum-of-square algorithm (see Section 2 for preliminaries about sum-of-square algorithms) that certifies that the injective norm of a random tensor from $\mathcal{D}_{m,n}$ is small.

▶ **Theorem 5.** *For a tensor T in distribution $\mathcal{D}_{m,n}$, when $m \ll n^{3/2}$ with high probability the injective norm of T is bounded by $1 + o(1)$. Further, this can be certified in polynomial time.*

Our results (Theorem 2 and 5) still hold when we are given a tensor $\tilde{T}$ that is $1/\mathrm{poly}(n)$-close to T in the sense that the spectral norm of an unfolding of $\tilde{T} - T$ is $O(1/\mathrm{poly}\log(n))$. Theorem 3 (and hence Corollary 4) requires a tensor $\tilde{T}$ such that the unfolding of $\tilde{T} - T$ has spectral norm bounded by $\epsilon/\mathrm{poly}(n)$.

Organization

The rest of this paper is organized as follows: In Section 2 we introduce tensor notations and SoS hierarchies. Then we describe the main idea of the proof which relates tensor decomposition to the injective norm of tensor (Section 3). In Section 4 we give a polynomial time algorithm for certifying the injective norm of a random 3rd order tensor. Using this as a key tool in Section 5 we present the quasi-polynomial time algorithm that can decompose randomly generated tensors when $m \ll n^{3/2}$.

2 Preliminaries

2.1 Notations

In this paper we use $\|\cdot\|$ to denote the ℓ_2 norm of vectors and the spectral norm of matrices. That is, $\|v\| = \sqrt{\sum_i v_i^2}$ and $\|A\| = \sup_{\|u\|=1} \|Au\|$. Note that we will be using the sum-norm instead of expectation norm $\|v\|_{exp} = \sqrt{\mathbb{E}_i[v_i^2]}$ because the scaling of sum-norm is more natural for the tensor decomposition setting. We use $\langle u, v\rangle$ to denote the inner product of u and v. When A and B are two matrices, we use standard notation $A \preceq B$ to denote the fact that $B - A$ is a positive semidefinite. For a $m \times n$ matrix U and a $p \times q$ matrix V, we define the Kronecker product $U \otimes V$ as the $mp \times nq$ block matrix

$$U \otimes V = \begin{bmatrix} U_{1,1}V & \cdots & U_{1,n}V \\ \vdots & \ddots & \vdots \\ U_{m,1}V & \cdots & U_{m,n}V \end{bmatrix}$$

We use $\tilde{O}$ notations to hide dependencies on polylog factors in n and m. When we write $f \ll g$ we mean $f \le g/O(\operatorname{poly}\log n)$. Throughout the paper high probability means the probability is at least $1 - n^{-\omega(1)}$.

2.2 Tensors

Tensors are multi-dimensional arrays. In this paper for simplicity we only consider 3rd order symmetric tensors and their symmetric decompositions. For a third order symmetric tensor T, the value of $T_{i,j,k}$ only depends on the multi-set $\{i, j, k\}$, so $T_{i,j,k} = T_{j,i,k} = T_{k,i,j}$ (and more generally all the 6 permutations are equal). For a vector $v \in \mathbb{R}^n$, we use $v^{\otimes 3} \in \mathbb{R}^{n^3}$ to denote the symmetric third order tensor such that $v^{\otimes 3}_{i,j,k} = v_i v_j v_k$. Our goal is to decompose a tensor T as $T = \sum_{i=1}^m a_i^{\otimes 3}$.

There is a bijection between 3rd order symmetric tensors and homogeneous degree 3 polynomials. In particular, for a tensor T we define its corresponding polynomial $T(x, x, x) = \sum_{i,j,k=1}^n T_{i,j,k} x_i x_j x_k$. It is easy to verify that if $T = \sum_{i=1}^m a_i^{\otimes 3}$ then $T(x, x, x) = \sum_{i=1}^m \langle a_i, x\rangle^3$.

The injective norm $\|T\|_{inj}$ is defined to be the maximum value of the corresponding polynomial on the unit sphere, that is:

$$\|T\|_{inj} := \sup_{\|x\|=1} T(x, x, x).$$

It is not hard to prove when $m \ll n^{3/2}$, and the tensor T is chosen from the distribution $\mathcal{D}_{m,n}$, with high probability $1 - o(1) \le \|T\|_{inj} \le 1 + o(1)$, and in fact the value $T(x, x, x)$ is only close to 1 if x is close to one of the components a_i. We will give a (SoS) proof of this fact in Section 5

2.3 Sum-of-Square Algorithms and Proofs

Here we will only briefly introduce the notations and key concepts that are used in this paper, for more detailed discussions and references about SoS proofs we refer readers to [12] (especially Section 2).

Sum-of-squares proof system is a proof system for polynomial equalities and inequalities. Given a set of constraints $\{r_i(x) = 0\}$, and a degree bound d, we say there is a degree d SoS proof for $p(x) \ge q(x)$ if $p(x) - q(x)$ can be written as a sum of squares of polynomials modulo $r_i(x) = 0$, as defined formally below.

▶ **Definition 6** (SoS proof of degree d)**.** For a set of constraints $R = \{r_1(x) = 0, \dots, r_t(x) = 0\}$, and an integer d, we write

$$p(x) \succeq_{R,d} q(x)$$

if there exists polynomials $h_i(x)$ for $i = 0, 1, \dots, \ell$ and $g_j(x)$ for $j = 1, \dots, t$ such that $\deg(h_0^2(p(x) - q(x))) \le d$, $\deg(h_i) \le d/2$ (for $i > 0$) and $\deg(g_j r_j) \le d$ that satisfy

$$h_0(x)^2(p(x) - q(x)) = \sum_{i=1}^{\ell} h_i(x)^2 + \sum_{j=1}^{t} r_j(x) g_j(x),$$

We will drop the subscript d when it is clear form the context.

Note that the constraints set can be easily generalized to a set of inequalities by adding auxiliary variables. For example, constraint $r(x) \ge 0$ can be implemented as $r(x) = z^2$ where z is an auxiliary variable.

Many well-known inequalities can be proved using a low degree SoS proof, among them the most useful and important one is Cauchy-Schwarz inequality, which can be proved via degree-2 sum of squares. Another one is that $x^T Ax \preceq \|A\|\|x\|^2$. This is pretty useful when A is a random matrix where we can use random matrix theory to bound the spectral norm of A.

In order to turn an SoS arguments into an algorithm, we often consider the *pseudo-expectation*. Just as we have expectations for real distributions, we think of pseudo-expectation as expectations for pseudo-distributions that cannot be distinguished from true expectations using low degree polynomials. Pseudo-expectation can be viewed as a dual of SoS refutations.

▶ **Definition 7** (pseudo-expectation)**.** A degree d pseudo-expectation $\widetilde{\mathbb{E}}$ is a linear operator that maps degree d polynomials to reals. The operator satisfies $\widetilde{\mathbb{E}}[1] = 1$ and $\widetilde{\mathbb{E}}[p^2(x)] \ge 0$ for all polynomials $p(x)$ of degree at most $d/2$. We say a degree-d pseudo-expectation $\widetilde{\mathbb{E}}$ satisfies a set of equations $\{r_i(x) : i = 1 \dots, \ell\}$ if for any i and any $q(x)$ such that $\deg(r_i q) \le d$,

$$\widetilde{\mathbb{E}}\left[r_i(x)q(x)\right] = 0$$

By definition, if $p(x) \preceq_{R,d} q(x)$, and degree-d pseudo-expectation satisfies R, then we can take pseudo-expectation on both sides and obtain $\widetilde{\mathbb{E}}\left[p(x)\right] \le \widetilde{\mathbb{E}}\left[q(x)\right]$. We will use this property of pseudo-expectation many times in the proofs.

The relationship between pseudo-expectations and SoS refutations can be summarized in the following informal lemma:

▶ **Lemma 8** ([27, 25], c.f. [12], informal stated)**.** *For a set of constraints R, either there is an SoS refutation of degree d that refutes R, or there is a degree d pseudo-expectation that satisfies R. Such a refutation/pseudo-expectation can be found in* $\mathrm{poly}(tn^d)$ *time.*

3 Relating Tensor Decompositions and Injective Norm

In this section we introduce the main idea of our proof. Given a tensor $T = \sum_{i=1}^m a_i^{\otimes 3}$ from distribution $\mathcal{D}_{m,n}$, we first make some observations about its corresponding polynomial $T(x,x,x) = \sum_{i=1}^m \langle a_i, x\rangle^3$.

When $x = a_1$, we know $T(a_1,a_1,a_1) = 1 + \sum_{i=2}^m \langle a_i, a_1\rangle^3$. Here conditioned on a_1, the second term is a sum of independent random variables $(\langle a_i, a_1\rangle^3)$. By the distribution $\mathcal{D}_{m,n}$ we know these variables have mean 0 and absolute value around $1/n^{3/2}$. Standard concentration bounds show when $m \ll n^{3/2}$ with high probability $T(a_1,a_1,a_1) = 1 \pm o(1)$.

On the other hand, suppose x is a random vector in the unit sphere, then $T(x,x,x) = \sum_{i=1}^m \langle a_i, x\rangle^3$ is again a sum of random variables. By concentration bounds we know for any particular x, when $m \ll n^{3/2}$ with high probability $T(x,x,x) = o(1)$. This can actually be generalized to all vectors x that do not have large correlation with a_i's using ϵ-net arguments.

▶ **Observation.** *For a random tensor $T \sim \mathcal{D}_{m,n}$, when $m = n^{3/2}$ with high probability $T(x,x,x) \le 1 + o(1)$ for $\|x\| = 1$. Further when $T(x,x,x)$ is close to 1 the vector x is close to one of the components a_i's.*

Later we will give a SoS proof for this observation. Based on this observation, if we want to find a component, then it suffices to find a vector x such that $T(x,x,x)$ is close to 1. Using the idea of pseudo-expectations, we can do this in two steps:

1. Find a pseudo-expectation $\tilde{\mathbb{E}}[x]$ that satisfies the constraint $\|x\|^2 - 1 = 0$ and maximizes $\tilde{\mathbb{E}}[T(x,x,x)]$.
2. "Sample" from this pseudo-distribution with psuedo-expectations $\tilde{\mathbb{E}}$ to get a vector x such that $T(x,x,x) \approx 1$, in particular x will be close to one of the components a_i's.

In Section 4 we will prove the first part of the observation. In particular we show even though we are maximizing over pseudo-expectation $\tilde{\mathbb{E}}[x]$ (instead of real distributions over x), we can still guarantee the maximum value $\tilde{\mathbb{E}}[T(x,x,x)]$ is at most $1 + 1/\log n$ with high probability.

In Section 5 we give algorithms for finding a component given a pseudo-expectation $\tilde{\mathbb{E}}$ with $\tilde{\mathbb{E}}[T(x,x,x)] \approx 1$. The main idea of our algorithm is similar to the robust tensor decomposition algorithm in [10]: first we show there must be a component a_i such that $\tilde{\mathbb{E}}[\langle a_i, x\rangle^d]$ is large for a large d, then we use ideas in [10] to find the component a_i.

4 Certifying Injective Norm

Algorithm 1 Certifying Injective Norm

Input: A random 3-tensor T

Output: If $\|T\|_{\text{inj}} > 1 + 1/\log n$, return NO. If $T \sim \mathcal{D}_{m,n}(m \ll n^{3/2})$, then w.h.p. return YES.

Solve the following optimization and obtain optimal value OPT

$$\begin{aligned} \text{Maximize} \quad & \tilde{\mathbb{E}}\left[T(x,x,x)\right] \\ \text{Subject to} \quad & \tilde{\mathbb{E}} \text{ is a degree-12 pseudo-expectation} && (1) \\ & \text{that satisfies } \{r(x) = \|x\|^2 - 1 = 0\} && (2) \end{aligned}$$

return YES if OPT $\le 1 + 1/\log n$ and NO otherwise.

In this section, we give Algorithm 1 based on SoS hierarchy that certifies the injective norm of random tensor. In particular, we will prove Theorem 5 which we restate in more details here.

▶ **Theorem 9.** *Algorithm 1 always returns NO when $\|T\|_{inj} > 1 + 1/\log n$. When $T \sim \mathcal{D}_{m,n}$ and $m \ll n^{3/2}$, Algorithm 1 returns YES with high probability over the randomness of T. Further, the same guarantee holds given an approximation $\tilde{T}$ where if $M \in \mathbb{R}^{n\times n^2}$ is an unfolding of $T - \tilde{T}$, $\|M\| \le 1/2\log n$.*

When $\|T\|_{\mathrm{inj}} > 1 + 1/\log n$, then by definition there must be a vector x^* that satisfies $\|x^*\| = 1$ and $T(x^*, x^*, x^*) > 1/\log n$. We can take $\widetilde{\mathbb{E}}$ to be the expectation of a distribution that is only supported on x^* (i.e. with probability 1 $x = x^*$). Clearly this pseudo-expectation is valid, and OPT will be at least larger than $1/\log n$. Hence the algorithm returns NO.

For random tensor T, we hope to show that with high probability, the tensor norm is less than $1 + 1/\log n$ can be proved via SoS.

▶ **Theorem 10.** *With high probability over the randomness of the tensor T, for $r(x) = \|x\|^2 - 1$,*

$$T(x,x,x) \preceq_{r,12} 1 + \widetilde{O}(m/n^{3/2}) \tag{3}$$

Note that taking pseudo-expectation $\widetilde{\mathbb{E}}$ on both hand sides of (3), for any degree-12 pseudo-expectation $\widetilde{\mathbb{E}}$ that is consistent with $r(x)$,

$$\widetilde{\mathbb{E}}\left[T(x,x,x)\right] \le 1 + \widetilde{O}(m/n^{3/2})$$

That is, when $m \ll n^{3/2}$, the objective value of the convex program in Algorithm 1 is less than $1 + 1/\log n$ with high probability for random tensor.

Now we need to prove Theorem 10. We first use Cauchy-Schwarz inequality to transform LHS of (3) to a degree-4 polynomial, which would then correspond to 4th order tensors and enable non-trivial unfoldings.

▶ **Claim 11.**

$$[T(x,x,x)]^2 \preceq_{r,12} \underbrace{\sum_{i=1}^m \langle a_i, x\rangle^4}_{\text{2-4 norm}} + \underbrace{\sum_{i\neq j} \langle a_i, a_j\rangle \langle a_i, x\rangle^2 \langle a_j, x\rangle^2}_{:=p(x)}. \tag{4}$$

Proof. This is a direct application of Cauchy-Schwarz inequality:

$$\left(T\cdot x^{\otimes 3}\right)^2 = \left\langle \sum_{i=1}^m \langle a_i, x\rangle^2 a_i, x\right\rangle^2 \preceq \left\|\sum_{i=1}^m \langle a_i, x\rangle^2 a_i\right\|^2 \|x\|^2 \preceq_r \left\|\sum_{i=1}^m \langle a_i, x\rangle^2 a_i\right\|^2$$

Expanding this quantity, and using the fact that $\|a_i\| = 1$, we get

$$\left\|\sum_{i=1}^m \langle a_i, x\rangle^2 a_i\right\|^2 = \sum_{i=1}^m \langle a_i, x\rangle^4 + \sum_{i\neq j} \langle a_i, a_j\rangle \langle a_i, x\rangle^2 \langle a_j, x\rangle^2. \tag{5}$$

◀

The first term is closely related to 2-to-4 norm of random matrices: let $A \in \mathbb{R}^{m\times n}$ be a matrix whose rows are equal to a_i's, then $\|A\|_{2\to 4} = \sup_{\|x\|=1} \|Ax\|_4$. Clearly, $\|A\|_{2\to 4}^4 = \sup_{\|x\|=1} \sum_{i=1}^m \langle a_i, x\rangle^4$ is the maximum value of the first term. This is considered in [8] where they gave a SoS proof that when $m \ll n^2$ the first term is bounded by $O(1)$. Here we are in the regime $m \ll n^{3/2}$ so we can improve the bound to $1 + o(1)$ (The proof is deferred to Appendix A.1):

▶ **Lemma 12.** *With high probability over the randomness of a_i's,*

$$\sum_{i=1}^m \langle a_i, x\rangle^4 \preceq_{r,12} 1 + \widetilde{O}(m/n^{3/2}) \tag{6}$$

The harder part of the proof is to deal with the second term $p(x)$ on the RHS of (4). The naive idea would be to let $y = x^{\otimes 2}$ and view $p(x)$ as a degree-2 polynomial of y,

$$q(y) = \sum_{i\neq j} \langle a_i, a_j\rangle \langle a_i \otimes a_i, y\rangle \langle a_j \otimes a_j, y\rangle = y^T N y. \tag{7}$$

Here N is an n^2 by n^2 random matrix that depends on a_i's. Suppose N has spectral norm less than $o(1)$, then we have $y^T N y \preceq \|N\|\|y\|^2$, and by replacing $y = x \otimes x$ we obtain $p(x) = q(x\otimes x) \preceq o(1)$. However, in our case the matrix N have spectral norm much larger than $o(1)$.

Our key insight is that we could have different ways to unfold $p(x)$ into a degree-2 polynomial. In particular, we use the following way of unfolding:

$$q'(y) = \sum_{i\neq j} \langle a_i, a_j\rangle \langle a_i \otimes a_j, y\rangle \langle a_i \otimes a_j, y\rangle = y^T M y \tag{8}$$

where M is the n^2 by n^2 matrix that encodes the coefficients of $q'(y)$,

$$M = \sum_{i\neq j} \langle a_i, a_j\rangle (a_i \otimes a_j)(a_i \otimes a_j)^T$$

It turns out that $q'(y)$ still have the property that $q'(x\otimes x) = p(x)$. The matrix M has much better spectral norm bound, which leads us to the bound for $p(x)$.

▶ **Lemma 13.** *When $m \ll n^{3/2}$, the matrix $M = \sum_{i\neq j} \langle a_i, a_j\rangle (a_i \otimes a_j)(a_i \otimes a_j)^T$ has spectral norm at most $\widetilde{O}(m/n^{3/2})$ and as a direct consequence,*

$$p(x) \preceq_{r,4} \widetilde{O}(m/n^{3/2})$$

First we give an informal and suboptimal bound for intuition. Let B be the $n^2 \times m^2$ matrix whose (i,j)-column ($i, j \in [m]$) is $a_i \otimes a_j$ (viewed as an n^2 dimensional vector). Then M can be written as $M = B\,\mathrm{diag}(\langle a_i, a_j\rangle)_{i\neq j} B^T$. Note that B can also be written as $A \otimes A$ where $\otimes$ is the Kronecker product of two matrices, so we have $\|B\| = \|A\|^2 \lesssim m/n$. Then we can bound the norm of M by $\|M\| \le \|B\| \|\,\mathrm{diag}(b)\| \|B\| \le (m/n) \cdot \max_{i,j} |\langle a_i, a_j\rangle| \cdot (m/n) \lesssim m^2/n^{5/2}$, where we used the incoherence of a_i's, that is, $|\langle a_i, a_j\rangle| \lesssim 1/\sqrt{n}$. This will only be $o(1)$ when $m \lesssim n^{1.25}$.

Intuitively, this proof is not tight because we ignored potential cancellation caused by the randomness of $\langle a_i, a_j\rangle$. Note that $\langle a_i, a_j\rangle$ have expectation 0, but we treated them all as positive $1/\sqrt{n}$. If we assume that $\langle a_i, a_j\rangle$'s are independent $\pm 1/\sqrt{n}$, then $M =$

$\sum_{i \neq j} \langle a_i, a_j \rangle (a_i \otimes a_j)(a_i \otimes a_j)^T$ would be a sum of PSD matrices with random weights and we can apply more standard matrix concentration bounds to make sure cancellations happen.

However, $\langle a_i, a_j \rangle$ are of course not independent and our key idea is to decouple the randomness of $\langle a_i, a_j \rangle$.

Proof. (Sketch) We first replace the vectors a_i's with $\sigma_i a_i$ where σ_i is a random ± 1 variable. This is OK because the distribution of a_i and $\sigma_i a_i$ are the same. Now we first sample the a_i's, conditioned on the samples $M = \sum_{i \neq j} \sigma_i \sigma_j \langle a_i, a_j \rangle (a_i \otimes a_j)(a_i \otimes a_j)^T$ (where only σ_i's are still random). Now since the vectors a_i's are all fixed, the correlation between different terms only depends on scalar variables $\sigma_i \sigma_j$, and we never use the term σ_i^2 (because $i \neq j$).

By a result of [28], in this case we can decouple the product $\sigma_i \sigma_j$. In particular, in order to prove concentration properties for M, it suffices to prove concentration for a different matrix $\sum_{i \neq j} \sigma_i \tau_j \langle a_i, a_j \rangle (a_i \otimes a_j)(a_i \otimes a_j)^T$. Here $\tau \in \{\pm 1\}^m$ is an independent copy of σ_i's. In this way we have decoupled the randomness in σ_i and τ_i, and the rest of the Lemma can follow from careful matrix concentration analysis. ◀

We give the full proof of Lemma 13 in Appendix A.2.

Proof Sketch of Main Theorem

Theorem 10 follows directly from Lemma 12 and Lemma 13. Using Lemma 8, we get the main Theorem 9 in the noiseless case. When there is noise, since we have bounds on spectral norm of an unfolding of $\tilde{T} - T$, it implies (by Lemma 33) $[\tilde{T} - T](x, x, x) \preceq_{r,12} 1/2 \log n$.it is easy to verify that $\tilde{T}(x, x, x) = T(x, x, x) + [\tilde{T} - T](x, x, x) \preceq_{r,12} 1 + 1/\log n$, so Theorem 9 still holds. We give more details in Appendix A.3.

5 Quasi-polynomial Time Algorithm for Tensor Decomposition

In this section we give a quasi-polynomial time algorithm for decomposing random 3rd order tensors in distribution $\mathcal{D}_{m,n}$. In particular, we prove Theorem 2 which we restate with more details below:

▶ **Theorem 14.** *Let T be a tensor chosen from $\mathcal{D}_{m,n}$, when $m \ll n^{3/2}$ with high probability over the randomness of T Algorithm 2 returns $\{\hat{a}_i\}$ that is 0.1-close to $\{a_i\}$ in time $n^{O(\log n)}$. Further, the same guarantee holds given an approximation $\tilde{T}$ where if $M \in \mathbb{R}^{n \times n^2}$ is an unfolding of $T - \tilde{T}$, $\|M\| \leq 1/10 \log n$.*

A key component of our algorithm is a way of sampling pseudo-distributions given in [10]:

▶ **Theorem 15** (Theorem 5.1 in [10]). *For every $k \geq 0$, there exists a randomized algorithm with running time $n^{O(k)}$ and success probability $2^{-k/\operatorname{poly}(\epsilon)}$ for the following problem: Given a degree-k pseudo distribution $\{u\}$ over $\mathbb{R}^n$ that satisfies the polynomial constraint $\|u\|^2 = 1$ and the condition $\widetilde{\mathbb{E}}[\langle c, u \rangle^k] \geq e^{-\epsilon k}$ for some unit vector $c \in \mathbb{R}^n$, output a unit vector $c' \in \mathbb{R}^n$ with $\langle c, c' \rangle \geq 1 - O(\epsilon)$.*

The basic idea of Algorithm 2 is as follows. At each iteration, the algorithm tries to find a new vector $\hat{a}_i$. As we discussed in Section 3, in order to find a vector close to a_i it finds a vector x with large $T(x, x, x)$ value. Moreover, It enforces that the new vector is different from all previous found vectors by the set of polynomial equations $\{\langle s, x \rangle^2 \leq 1/8 : s \in S\}$. Intuitively, if we haven't found all of the vectors a_i's any of the remaining a_i's will satisfy the set of constraints $\{\langle s, x \rangle^2 \leq 1/8 : s \in S\}$ and $T(x, x, x) \geq 1 - 1/\log n$. Therefore each time we can find a valid pseudo-expectation $\widetilde{\mathbb{E}}$.

What we need to prove is for any pseudo-expectation $\widetilde{\mathbb{E}}$ we found, it always satisfies $\widetilde{\mathbb{E}}[\langle a_i, x\rangle^k] \geq e^{-\epsilon k}$ for some $k = O((\log n)/\epsilon)$ for some small enough constant ϵ. Then by Theorem 15we can obtain a new vector that is $O(\epsilon)$-close to one of the a_i's. We formalize this in the following lemma:

Algorithm 2 Overcomplete Random 3-Tensor Decomposition

Input: Random 3-tensor $T = \sum_{i=1}^m a_i^{\otimes 3} \sim \mathcal{D}_{m,n}$.
Output: $\hat{a}_1, \ldots, \hat{a}_m \in \mathbb{R}^n$ s.t. $\{\hat{a}_i\}$ is 0.1-close to $\{a_i\}$

1: $S \leftarrow \emptyset$
2: **repeat**
3: Using semidefinite programming to find a degree $k = O(\log n)$ pseudo-expectation $\widetilde{\mathbb{E}}$ that satisfies the constraints $\{T(x,x,x) \geq 1 - 1/\log n, \|x\|^2 = 1\}$ and $\{\langle s, x\rangle^2 \leq 1/8 : s \in S\}$.
4: Run the algorithm in Theorem 5.1 of [10] (for $n^{O(k)}$ times) with input $\widetilde{\mathbb{E}}$ and obtain vector c such that $T(c,c,c) \geq 0.99$.
5: add vector c to S.
6: **until** $|S| = m$
7: **return** $\{\hat{a}_i\} = S$.

▶ **Lemma 16.** *When T is chosen from $\mathcal{D}_{m,n}$ where $m \ll n^{3/2}$, with high probability over the randomness of T, the pseudo-expectation found in Step 3 of Algorithm 2 satisfies the following: there exists an a_i such that $\tilde{E}[\langle a_i, x\rangle^k] \geq e^{-\epsilon k}$ for sufficiently small constant ϵ (where the pseudo-expectation has degree $4k$ and $k = O((\log n)/\epsilon)$). In particular, applying Theorem 15, repeat the algorithm for $n^{O(k)}$ time will give a vector c such that $\langle c, a_i\rangle \geq 1 - O(\epsilon)$.*

The main intuition is to use Cauchy-Schwarz and Hölder inequalities (like what we used in Claim 11) to raise the power in the sum $\sum_{i=1}^m \langle a_i, x\rangle^d$ (we start with $d = 3$ and hope to get to $d = k$). When the degree is high enough we can afford to do an averaging argument and lose a factor of m to go from the sum to a individual vector, because $e^{-\epsilon k} = \mathrm{poly}(m)$. The detailed proof is given in Appendix B.1.

Now we are ready to prove Theorem 14.

Proof. (sketch) We prove Theorem 14 by induction. Suppose s already contains a set of vectors $\hat{a}_i$'s, where for each $\hat{a}_i$ there is a corresponding a_j that satisfies $\|\hat{a}_i - a_j\| \leq 0.1$. We would like to show with high probability in the next iteration, the algorithm finds a new component that is different from all the previously found a_i's.

In order to do that, we need to show the following:

1. The SDP in Step 3 of Algorithm 2 is feasible and gives a valid pseudo-expectation.
2. For any valid pseudo-expectation, with high probability we get an unit vector c that satisfies $T(c,c,c) \geq 0.99$, and c is far from all the previously found a_i's.
3. For any unit vector c such that $T(c,c,c) \geq 0.99$, there must be a component a_i such that $\|a_i - c\| \leq 0.1$.

In these three steps, Step 1 follows because we can take $\widetilde{\mathbb{E}}$ to be the expectation of a true distribution: $x = a_i$ with probability 1 for some unfound a_i. Step 2 is basically Lemma 16, when we choose ϵ to be a small enough constant, it is easy to prove that all the vectors that satisfy $\langle c, a_i\rangle \geq 1 - O(\epsilon)$ must satisfy $T(c,c,c) \geq 0.99$. Step 3 is the second part of our observation in Section 3, which we prove in the appendix. ◀

The details in this proof can be found in Appendix B.2.

6 Conclusion

In this paper we give the first algorithm that can decompose an overcomplete 3rd order tensor when the rank m is almost $n^{3/2}$ that matches the $n^{p/2}$ bounds for even order tensors. Our argument is based on a special unfolding of the tensor and a decoupling argument for matrix concentration. We feel such techniques can be useful in other settings.

Tensor decompositions are widely applied in machine learning for learning latent variable models. Although the SoS based algorithm have poor dependency on the accuracy ϵ, in the case of tensor decomposition we can actually use SoS as an initialization algorithm. We hope such ideas can help solving more problems in machine learning.

Acknowledgements. We would like to thank Anima Anandkumar, Boaz Barak, Johnathan Kelner, David Steurer, Venkatesan Guruswami for helpful discussions at various stages of this work

References

1 Boris Alexeev, Michael A Forbes, and Jacob Tsimerman. Tensor rank: Some lower and upper bounds. In *Computational Complexity (CCC), 2011 IEEE 26th Annual Conference on*, pages 283–291. IEEE, 2011.

2 A. Anandkumar, D. P. Foster, D. Hsu, S. M. Kakade, and Y. K. Liu. Two SVDs Suffice: Spectral Decompositions for Probabilistic Topic Modeling and Latent Dirichlet Allocation. *to appear in the special issue of Algorithmica on New Theoretical Challenges in Machine Learning*, July 2013.

3 A. Anandkumar, R. Ge, D. Hsu, and S. M. Kakade. A Tensor Spectral Approach to Learning Mixed Membership Community Models. In *Conference on Learning Theory (COLT)*, June 2013.

4 A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor Methods for Learning Latent Variable Models. *J. of Machine Learning Research*, 15:2773–2832, 2014.

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

6 Anima Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed Non-Orthogonal Tensor Decomposition via Alternating Rank-1 Updates. *arXiv preprint arXiv:1402.5180*, Feb. 2014.

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

8 Boaz Barak, Fernando G.S.L. Brandao, Aram W. Harrow, Jonathan Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *Proceedings of the Forty-fourth Annual ACM Symposium on Theory of Computing*, STOC'12, pages 307–326, New York, NY, USA, 2012. ACM.

9 Boaz Barak, Jonathan A. Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *STOC*, pages 31–40, 2014.

10 Boaz Barak, Jonathan A. Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC'15, 2015.

11 Boaz Barak and Ankur Moitra. Tensor prediction, rademacher complexity and random 3-XOR. `http://arxiv.org/abs/1501.06521`, 2015.

12 Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. In *Proceedings of International Congress of Mathematicians (ICM)*, 2014. To appear.
13 Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pages 594–603. ACM, 2014.
14 Joseph T. Chang. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. *Mathematical Biosciences*, 137:51–73, 1996.
15 Pierre Comon. Tensor: a partial survey. *Signal Processing Magazine*, page 11, 2014.
16 Lieven De Lathauwer, Joséphine Castaing, and Jean-François Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *Signal Processing, IEEE Transactions on*, 55(6):2965–2973, 2007.
17 Ignat Domanov and Lieven De Lathauwer. Canonical polyadic decomposition of third-order tensors: relaxed uniqueness conditions and algebraic algorithm. *arXiv preprint arXiv:1501.07251*, 2015.
18 Ignat Domanov and Lieven De Lathauwer. Canonical polyadic decomposition of third-order tensors: reduction to generalized eigenvalue decomposition. *SIAM Journal on Matrix Analysis and Applications*, 35(2):636–660, 2014.
19 Rong Ge, Qingqing Huang, and Sham M. Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the Forty-seventh Annual ACM Symposium on Theory of Computing*, STOC'15, 2015.
20 Leonid Gurvits. Classical deterministic complexity of edmonds' problem and quantum entanglement. In *Proceedings of the Thirty-fifth Annual ACM Symposium on Theory of Computing*, STOC'03, pages 10–19, New York, NY, USA, 2003. ACM.
21 Aram W Harrow and Ashley Montanaro. Testing product states, quantum merlin-arthur games and tensor optimization. *Journal of the ACM (JACM)*, 60(1):3, 2013.
22 Johan Håstad. Tensor rank is np-complete. *Journal of Algorithms*, 11(4):644–654, 1990.
23 Christopher J. Hillar and Lek-Heng Lim. Most tensor problems are NP hard. *arXiv preprint arXiv:0911.1393*, 2009.
24 J.B. Kruskal. Three-way arrays: Rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977.
25 Jean B Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.
26 Elchanan Mossel and Sébastian Roch. Learning nonsingular phylogenies and hidden Markov models. *Annals of Applied Probability*, 16(2):583–614, 2006.
27 Pablo A Parrilo. *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization.* PhD thesis, California Institute of Technology, 2000.
28 Victor H. de la Pena and S. J. Montgomery-Smith. Decoupling inequalities for the tail probabilities of multivariate u-statistics. *The Annals of Probability*, 23(2):pp. 806–816, 1995.
29 Volker Strassen. Vermeidung von divisionen. *Journal für die reine und angewandte Mathematik*, 264:184–202, 1973.
30 Joel A Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012.

A Omitted Proofs in Section 4

A.1 Proof of Lemma 12

We first restate the lemma here.

▶ **Lemma 17.** *With high probability over the randomness of a_i's,*

$$\sum_{i=1}^{m} \langle a_i, x\rangle^4 \preceq_{r,12} 1 + \widetilde{O}(m/n^{3/2}) \tag{9}$$

Recall [8] showed that when $m \ll n^2$,

$$\sum_{i=1}^{m} \langle a_i, x\rangle^4 \leq O(1) \tag{10}$$

Here in order to improve this bound, we consider the square of the LHS of (6) and apply Cauchy-Schwarz (similar to Claim 11),

$$\begin{aligned}\left(\sum_{i=1}^{m} \langle a_i, x\rangle^4\right)^2 &= \left\langle \sum_{i=1}^{m} \langle a_i, x\rangle^3 a_i, x \right\rangle^2 \\ &\preceq \left\|\sum_{i=1}^{m} \langle a_i, x\rangle^3 a_i\right\|^2 \|x\|^2 \qquad \text{by Cauchy-Schwarz} \\ &\preceq_r \left\|\sum_{i=1}^{m} \langle a_i, x\rangle^3 a_i\right\|^2 = \sum_{i=1}^{m} \langle a_i, x\rangle^6 + \sum_{i\neq j} \langle a_i, a_j\rangle \langle a_i, x\rangle^3 \langle a_j, x\rangle^3 \end{aligned} \tag{11}$$

We will bound the first term of (11) by $1 + o(1)$. We simply let $y = x^{\otimes 3}$ and let B be the matrix whose ith row is $a_i^{\otimes 3}$. Then $f(y) = \|By\|^2$ has the property that $f(x^{\otimes 3}) = \sum_{i=1}^{m} \langle a_i, x\rangle^6$. Therefore it suffices to prove that $f(y) \preceq (1 + o(1)\|y\|^2$ or equivalently $\|B\| \leq 1 + o(1)$.

Consider the matrix BB^T. It is a n by n matrix with diagonal entries 1 and off diagonal entries of the form $\langle a_i^{\otimes 3}, a_j^{\otimes 3}\rangle = \langle a_i, a_j\rangle^3$. By the incoherence of a_i's, we have $\langle a_i, a_j\rangle^3 \lesssim 1/n^{3/2}$. Then by Gershgorin disk theorem, we have $\|BB^T\| \leq 1 + \widetilde{O}(m/n^{3/2}) = 1 + \delta$. It follows that $\|B\| \leq 1 + \widetilde{O}(m/n^{3/2})$. Therefore,

$$\sum_{i=1}^{m} \langle a_i, x\rangle^6 = \|Bx^{\otimes 3}\|^2 \preceq (1 + \widetilde{O}(m/n^{3/2}))\|x^{\otimes 3}\| \leq_r 1 + \widetilde{O}(m/n^{3/2}) \tag{12}$$

For the second term of (11), we apply Cauchy-Schwarz again:

$$\begin{aligned}\left(\sum_{i\neq j} \langle a_i, a_j\rangle \langle a_i, x\rangle^3 \langle a_j, x\rangle^3\right)^2 &\preceq \left(\sum_{i\neq j} \langle a_i, a_j\rangle^2 \langle a_i, x\rangle^2 \langle a_j, x\rangle^2\right)\left(\sum_{i\neq j} \langle a_i, x\rangle^4 \langle a_j, x\rangle^4\right) \\ &\preceq \left(\frac{1}{n} \cdot \sum_i \langle a_i, x\rangle^2 \sum_j \langle a_j, x\rangle^2\right)\left(\sum_i \langle a_i, x\rangle^4 \sum_j \langle a_j, x\rangle^4\right)\end{aligned} \tag{13}$$

Note that the matrix $A = [a_1|\dots|a_m]$ has spectral norm bound $\|A\| \lesssim \sqrt{m/n}$, and therefore

$$\sum_i \langle a_i, x\rangle^2 = \|A^T x\|^2 \preceq \|A\|^2\|x\|^2 \preceq_r \|A\|^2$$

Then using Equation 10, and the equation above, we have

$$\text{RHS of (13)} \;\; \preceq_r \frac{1}{n}\cdot\frac{m}{n}\cdot\frac{m}{n}\cdot O(1)\cdot O(1) \le \widetilde{O}(m^2/n^3) \tag{14}$$

Then by 13 and 14 and Lemma 34, we have that

$$\sum_{i\neq j}\langle a_i, a_j\rangle\langle a_i, x\rangle^3\langle a_j, x\rangle^3 \preceq_r \widetilde{O}(m^2/n^3) \tag{15}$$

Hence, combining equation (15), (12) and (11) we have that

$$\begin{aligned}\left(\sum_{i=1}^m \langle a_i, x\rangle^4\right)^2 &\preceq_r \sum_{i=1}^m \langle a_i, x\rangle^6 + \sum_{i\neq j}\langle a_i, a_j\rangle\langle a_i, x\rangle^3\langle a_j, x\rangle^3 \\ &\preceq_r 1 + \widetilde{O}(m/n^{3/2}) + \widetilde{O}(m/n^{3/2}) = 1 + \widetilde{O}(m/n^{3/2})\end{aligned} \tag{16}$$

Using Lemma 34 again, we complete the proof of Lemma 6.

A.2 Proof of Lemma 13

We first restate the lemma:

▶ **Lemma 18.** *When $m \ll n^{3/2}$, the matrix $M = \sum_{i\neq j}\langle a_i, a_j\rangle(a_i\otimes a_j)(a_i\otimes a_j)^T$ has spectral norm at most $\widetilde{O}(m/n^{3/2})$ and as a direct consequence,*

$$p(x) \preceq_{r,4} \widetilde{O}(m/n^{3/2})$$

Proof. As suggested in the proof sketch, we first use a simple symmetrization which allows us to focus on the randomness of signs of $\langle a_i, a_j\rangle$. For simplicity of notation, let $Q_{ij} := \langle a_i, a_j\rangle(a_i\otimes a_j)(a_i\otimes a_j)^T$. Let $\sigma \in \{\pm 1\}^m$ be uniform random ± 1 vector and define M' as

$$M' = \sum_{i\neq j}\sigma_i\sigma_j Q_{ij}.$$

We claim that M' has the same distribution as M, since a_i has the same distribution as $\sigma_i a_i$. Then from now on we condition on the event that a_i's have incoherence property and low spectral norm, that is, $\langle a_i, a_j\rangle \lesssim 1/\sqrt{n}$, $\|A\| = \|[a_1|a_2\dots|a_m]\| \lesssim \sqrt{m/n}$, and we will only focus on the randomness of σ. Ideally we want to write M' as a sum of independent random matrices so that we can apply matrix Bernstein inequality. However, now the random coefficients are $\sigma_i\sigma_j$, and they are not independent with each other.

A key observation here is that the sum is only over the indices (i,j) with $i\neq j$, therefore we can use Theorem 1 of [28] (restated as Theorem 29 in the end) to decouple the correlation first.

Theorem 29 basically says that to study the concentration of a sum of the form $\sum_{i\neq j} f_{ij}(X_i, X_j)$, it is up to constant factor similar to the concentration of the sum $\sum_{i\neq j} f_{ij}(X_i, Y_j)$ where Y_i is an independent copy of X_i. Applying the theorem to our situation, we have that there exists absolute constant C such that

$$\Pr[\|M'\| \ge t] \le C\Pr[M'' \ge t/C] \tag{17}$$

where

$$M'' := \sum_{i \neq j} \sigma_i \tau_j Q_{ij},$$

and σ, τ are independently uniform over $\{-1, +1\}^m$.

Now it suffices to bound the norm of M''. We proceed by rewriting M'' as

$$M'' = \sum_i \sigma_i \sum_{j \neq i} \tau_j Q_{ij} := \sum_i \sigma_i T_i,$$

where

$$T_i := \sum_{j \neq i} \tau_j Q_{ij} \tag{18}$$

We study the properties of T_i first.

▶ **Claim 19.** *With high probability over the randomness of a_i's, for all i, $T_i \preceq \tilde{O}(\sqrt{m}/n)(a_i a_i^T) \otimes I$.*

Proof. Recall that $Q_{ij} = \langle a_i, a_j \rangle (a_i \otimes a_j)(a_i \otimes a_j^T)$. In the definition 18 of T_i, the index i is fixed and we take sum over j. Therefore it will be convenient to write Q_{ij} as $Q_{ij} = \langle a_i, a_j \rangle (a_i a_i^T) \otimes (a_j a_j)^T$ where $\otimes$ is the Kronecker product between matrices. Then T_i can be written as

$$T_i = (a_i a_i^T) \otimes \left(\sum_j \tau_j \langle a_i, a_j \rangle a_j a_j^T \right).$$

We apply the Matrix Bernstein inequality (Theorem 30) on the right factor. Matrix Bernstein bound requires spectral norm bound for individual matrices, and a variance bound.

For the spectral norm of individual matrices, we check that $\|\tau_j \langle a_i, a_j \rangle a_j a_j^T\| \lesssim 1/\sqrt{n}$ (by incoherence). For variance we know

$$\| \mathbb{E}[\sum_j \tau_j^2 (\langle a_i, a_j \rangle a_j a_j^T)^2]\| = \|A \operatorname{diag}(\langle a_i, a_j \rangle^2)_{j \neq i} A^T\| \lesssim m/n^2,$$

where we used the spectral norm of A and the fact that $\langle a_i, a_j \rangle^2 \lesssim 1/n$.

Therefore by Matrix Bernstein's inequality (Theorem 30) we have that whp, over the randomness of τ,

$$\| \sum_j \tau_j \langle a_i, a_j \rangle a_j a_j^T \| \leq \widetilde{O}(\sqrt{m}/n).$$

Using the fact that for two matrices P and Q, if $P \preceq Q$ and R is PSD, then $R \otimes P \preceq R \otimes Q$ (see Claim 20), it follows that

$$T_i \preceq (a_i a_i^T) \otimes (\widetilde{O}(\sqrt{m}/n) \cdot I).$$

Finally we use union bound and conclude with high probability this is true for any i. ◀

Now we can apply matrix Bernstein for the sum $M'' = \sum_{i=1}^m \sigma_i T_i$. The individual spectral norm is bounded by $\tilde{O}(\sqrt{m}/n)$ by the Claim 19. The variance is

$$\| \sum_{i=1}^m T_i^2 \| \leq \tilde{O}(m/n^2) \| \sum_{i=1}^m ((a_i a_i^T) \otimes I)^2 \| = \tilde{O}(m/n^2) \|(AA^T) \otimes I\| = \tilde{O}(m^2/n^3).$$

Using matrix Bernstein inequality, we know with high probability $\|M''\| \leq \tilde{O}(m/n^{3/2})$.

Using (17), we get that whp, $\|M'\| \leq \widetilde{O}(m/n^{3/2})$. Since M' and M has the same distribution, we conclude that whp, $\|M\| \leq \widetilde{O}(m/n^{3/2})$. ◀

We complete the proof by providing the following claim about Kronecker products.

▶ **Claim 20.** *If $P \preceq Q$ and R is psd, then $R \otimes P \preceq R \otimes Q$.*

Proof. It suffices to prove this when $R = uu^T$ (as we can always decompose R as sum of rank one components). In that case, for any $y \in \mathbb{R}^{n^2}$, we can write $y = u \otimes v + z$ where z is orthogonal to $u \otimes e_i$ for all $i \in [n]$. Now $(R \otimes P)z = 0$, therefore

$$y^T(R \otimes P)y = (u \otimes v)^T(R \otimes P)(u \otimes v) = (u^T Ru)(v^T Pv) \leq (u^T Ru)(v^T Qv) = y^T(R \otimes Q)y.$$

Therefore $R \otimes P \preceq R \otimes Q$. ◀

A.3 Main Theorem for Certifying Injective Norm

Now we are ready to prove Theorem 9.

▶ **Theorem 21.** *Algorithm 1 always returns NO when $\|T\|_{inj} > 1+1/\log n$. When $T \sim \mathcal{D}_{m,n}$ and $m \ll n^{3/2}$, Algorithm 1 returns YES with high probability over the randomness of T. Further, the same guarantee holds given an approximation $\tilde{T}$ where if $M \in \mathbb{R}^{n \times n^2}$ is an unfolding of $T - \tilde{T}$, $\|M\| \leq 1/2 \log n$.*

Proof. We first prove whenever $\|T\|_{\text{inj}} > 1 + 1/\log n$, the algorithm returns NO. This is because a large injective norm implies there exists an unit vector x^* with $T(x^*, x^*, x^*) = 1$. We can construct a pseudo-expectation $\widetilde{\mathbb{E}}$ as $\widetilde{\mathbb{E}}[p(x)] = p(x^*)$. Clearly this is a valid pseudo-expectation (it is even the expectation of a true distribution: $x = x^*$ with probability 1). Also, we know $\widetilde{\mathbb{E}}[T(x, x, x)] = T(x^*, x^*, x^*) > 1+1/\log n$, so in particular $OPT > 1+1/\log n$ and the algorithm must return NO.

Next we show the algorithm returns YES with high probability when T is chosen from $\mathcal{D}$. This follows directly from Theorem 10, which in turn follows from Lemmas 12 and 13. In particular, we know there is a degree-12 SoS proof that shows $T(x, x, x) \leq 1 + \tilde{O}(m/n^{3/2}) \leq 1 + 1/2 \log n$, so by Lemma 8 this must also hold for any pseudo-expectation.

When we are only given tensor $\tilde{T}$ such that the unfolding of $\tilde{T} - T$ has spectral norm $1/2 \log n$. Let M be the unfolding of $\tilde{T} - T$, and $y = x \otimes x$, then by Lemma 33 we know $(x^T My)^2 \preceq \|x\|^2\|M\|^2\|y\|^2$, which implies (by Lemma 33) $[\tilde{T} - T](x, x, x) = x^T My \preceq_{r,12} \|M\| \leq 1/2 \log n$. Combining the two terms we know

$$\tilde{T}(x, x, x) = T(x, x, x) + [\tilde{T} - T](x, x, x) \preceq_{r,12} 1 + 1/\log n.$$

◀

B Omitted Proof in Section 5

B.1 Proof of Lemma 16

We first restate the lemma here:

▶ **Lemma 22.** *When T is chosen from $\mathcal{D}_{m,n}$ where $m \ll n^{3/2}$, with high probability over the randomness of T, the pseudo-expectation found in Step 3 of Algorithm 2 satisfies the following: there exists an a_i such that $\tilde{E}[\langle a_i, x\rangle^k] \geq e^{-\epsilon k}$ for sufficiently small constant ϵ (where the pseudo-expectation has degree $4k$ and $k = O((\log n)/\epsilon)$). In particular, applying Theorem 15, repeat the algorithm for $n^{O(k)}$ time will give a vector c such that $\langle c, a_i\rangle \geq 1 - O(\epsilon)$.*

First we will show that for a valid pseudo-expectation, the sum of $\langle a_i, x\rangle^4$ and $\langle a_i, x\rangle^6$ are also bounded. This actually follows directly from the proof of Lemma 12 and 13.

▶ **Lemma 23.** *With high probability over the randomness of T, we have that for any degree-12 pseudo expectation $\widetilde{\mathbb{E}}$ that satisfies the constraints $\{\|x\|^2 = 1, T(x,x,x) \geq 1-\tau\}$, it also satisfies*

$$1+\epsilon \geq \widetilde{\mathbb{E}}\left[\sum_{i=1}^{m}\langle a_i, x\rangle^4\right] \geq 1-\epsilon \tag{19}$$

$$1+\epsilon \geq \widetilde{\mathbb{E}}\left[\sum_{i=1}^{m}\langle a_i, x\rangle^6\right] \geq 1-\epsilon \tag{20}$$

for $\epsilon = \widetilde{O}(m/n^{3/2}) + O(\tau)$.

Proof. We essentially just take pseudo-expectation on the SoS proofs for Lemma 12 and 13. The upper bounds follows directly by taking pseudo-expectation on equation (9) and (12). Fo the lower bounds, by taking pseudo-expectation over the SoS equation in Lemma 13, we have that $\widetilde{\mathbb{E}}\,[p(x)] \leq \widetilde{O}(m/n^{3/2})$. Taking pseudo-expectation over Claim 11, using the assumption that $\widetilde{\mathbb{E}}$ satisfies $T(x,x,x) \geq 1-\tau$, we have that

$$1-\tau \leq \widetilde{\mathbb{E}}\left[[T(x,x,x)]^2\right] \leq \widetilde{\mathbb{E}}\left[\langle a_i, x\rangle^4\right] + \widetilde{\mathbb{E}}\,[p(x)] \leq \widetilde{\mathbb{E}}\left[\langle a_i, x\rangle^4\right] + \widetilde{O}(m/n^{3/2}) \tag{21}$$

which implies

$$\widetilde{\mathbb{E}}\left[\langle a_i, x\rangle^4\right] \geq 1-\tau-\widetilde{O}(m/n^{3/2}). \tag{22}$$

For proving the lower bounds in (20), we first pseudo-expectation on equation 15, we have that

$$\widetilde{\mathbb{E}}\left[\sum_{i\neq j}\langle a_i, a_j\rangle\langle a_i, x\rangle^3\langle a_j, x\rangle^3\right] \leq \widetilde{O}(m^2/n^3)$$

Then taking pseudo-expectation over equation (16), we obtain that

$$\widetilde{\mathbb{E}}\left[\left(\sum_{i=1}^{m}\langle a_i, x\rangle^4\right)^2\right] \leq \widetilde{\mathbb{E}}\left[\sum_{i=1}^{m}\langle a_i, x\rangle^6\right] + \widetilde{\mathbb{E}}\left[\sum_{i\neq j}\langle a_i, a_j\rangle\langle a_i, x\rangle^3\langle a_j, x\rangle^3\right]$$

Note that by equation (22) and Cauchy-Schwarz, we have

$$\widetilde{\mathbb{E}}\left[\left(\sum_{i=1}^{m}\langle a_i, x\rangle^4\right)^2\right] \geq \left(\widetilde{\mathbb{E}}\left[\sum_{i=1}^{m}\langle a_i, x\rangle^4\right]\right)^2 \geq 1-O(\tau)-\widetilde{O}(m/n^{3/2})$$

Combining the two equations above, we obtain that

$$\widetilde{\mathbb{E}}\left[\sum_{i=1}^{m}\langle a_i, x\rangle^6\right] \geq 1-O(\tau)-\widetilde{O}(m/n^{3/2})$$

◀

Next we are going to prove that $\widetilde{\mathbb{E}}$ also satisfies the condition of Theorem 5.2 of [10].

▶ **Lemma 24.** *For $k = O((\log n)/\epsilon)$ with constant $\epsilon < 1$, If $\widetilde{\mathbb{E}}$ is a degree-k pseudo-expectation that satisfies equation* (20) *and* (19)*, then there must exists $i \in [m]$ such that $\widetilde{\mathbb{E}}[\langle a_i, x\rangle^k] \geq e^{-(2\epsilon+\delta)k}$ with $\delta = \widetilde{O}(m/n^{3/2})$.*

Proof. By equation (2.5) of [10], we the following SoS version of Holder inequality. For any integer t, d and $k = t(d-2)$,

$$\|v\|_d^{dt} \preceq_k \|v\|_k^k \cdot \|v\|^{2t}$$

Let $v_i = \langle a_i, x\rangle^2$, we have

$$\left(\sum_{i=1}^m \langle a_i, x\rangle^{2d}\right)^t \preceq_k \sum_{i=1}^m \langle a_i, x\rangle^{2k} \cdot \left(\sum_{i=1}^m \langle a_i, x\rangle^4\right)^t \tag{23}$$

By Lemma 12, we have that with high probability over randomness of a_i's, $\sum_{i=1}^m \langle a_i, x\rangle^4 \preceq 1 + \widetilde{O}(m/n^{3/2})$, and it follows that

$$\left(\sum_{i=1}^m \langle a_i, x\rangle^4\right)^t \leq (1 + \widetilde{O}(m/n^{3/2}))^t \tag{24}$$

By picking $d = 3$, we have $t = k$. Taking $t = O(\log m/\epsilon)$ and combining equation (23) and (24), we have that

$$\left(\sum_{i=1}^m \langle a_i, x\rangle^6\right)^k \preceq_k \sum_{i=1}^m \langle a_i, x\rangle^{2k} \cdot \left(\sum_{i=1}^m \langle a_i, x\rangle^4\right)^k \preceq_k (1 + \widetilde{O}(m/n^{3/2}))^k \sum_{i=1}^m \langle a_i, x\rangle^{2k}$$

Applying pseudo-expectation on both hands, we obtain,

$$\widetilde{\mathbb{E}}\left[\left(\sum_{i=1}^m \langle a_i, x\rangle^6\right)^k\right] \leq (1 + \widetilde{O}(m/n^{3/2}))^k \cdot \widetilde{\mathbb{E}}\left[\sum_{i=1}^m \langle a_i, x\rangle^{2k}\right]$$

Note that by Cauchy-Schwarz and equation (20), we have

$$(1-\epsilon)^k \leq \widetilde{\mathbb{E}}\left[\sum_{i=1}^m \langle a_i, x\rangle^6\right]^k \leq \widetilde{\mathbb{E}}\left[\left(\sum_{i=1}^m \langle a_i, x\rangle^6\right)^k\right]$$

Combining the two equations above, we obtain that for $\delta = \widetilde{O}(m/n^{3/2})$,

$$\widetilde{\mathbb{E}}\left[\sum_{i=1}^m \langle a_i, x\rangle^{2k}\right] \geq (1-\delta)^k (1-\epsilon)^k \tag{25}$$

Therefore by averaging argument, there exists i such that

$$\widetilde{\mathbb{E}}[\langle a_i, x\rangle^{2k}] \geq (1-\delta)^k/m = e^{-\delta k - \log m - \epsilon k}$$

when $k \geq (\log m)/\epsilon$, we have that $\widetilde{\mathbb{E}}[\langle a_i, x\rangle^{2k}] \geq e^{-(2\epsilon+\delta)k}$ ◀

Lemma 16 follows directly from the two lemmas above.

B.2 Proof of Theorem 14

In this section we prove the main theorem in Section 5.

▶ **Theorem 25.** *Let T be a tensor chosen from $\mathcal{D}_{m,n}$, when $m \ll n^{3/2}$ with high probability over the randomness of T Algorithm 2 returns $\{\hat{a}_i\}$ that is 0.1-close to $\{a_i\}$ in time $n^{O(\log n)}$. Further, the same guarantee holds given an approximation $\tilde{T}$ where if $M \in \mathbb{R}^{n\times n^2}$ is an unfolding of $T - \tilde{T}$, $\|M\| \le 1/10\log n$.*

As suggested in the proof sketch, we prove this theorem by induction. The induction hypothesis is that all vectors $s_i \in S$ are 0.1-close (in ℓ_2 norm) to distinct components a_i's. We break the proof into three claims:

▶ **Claim 26.** *With high probability over the tensor T, suppose all the previously found s_i's are 0.1-close (in ℓ_2 norm) to some components a_j's, then there exists a pseudo-expectation that satisfies Step 3 in Algorithm 2.*

Proof. We first prove that with high probability $T(a_i, a_i, a_i) \ge 1 - 1/\log n$ for all i. This is easy because $T(a_i, a_i, a_i) = 1 + \sum_{j\neq i}\langle a_i, a_j\rangle^3$. Conditioned on a_i, the values $\langle a_i, a_j\rangle$ are sub-Gaussian random variables with mean 0 and variance $1/n$, so by standard concentration bounds we know with high probability $\sum_{j\neq i}\langle a_i, a_j\rangle^3 \ge -1/\log n$. We can then take the union bound and conclude $T(a_i, a_i, a_i) \ge 1 - 1/\log n$ for all i.

Now for simplicity of notation, assume that $S = \{s_1, \dots, s_t\}$ for some $t < m$, where s_i is 0.1-close to a_i. We can construct a pseudo-expectation $\widetilde{\mathbb{E}}[p(x)] = p(a_{t+1})$. Clearly this is a valid pseudo-expectation that satisfies $\|x\|^2 = 1$. For the inequality constraints we also know $\langle a_{t+1}, s_i\rangle^2 \le 2(\langle a_{t+1}, a_i\rangle^2 + \langle a_{t+1}, a_i - s_i\rangle^2) < 1/8$ (where the whole proof only uses Cauchy-Schwarz and $(A + B)^2 \le 2(A^2 + B^2)$, so the proof is SoS). Therefore the system in Step 3 must have a feasible solution. ◀

▶ **Claim 27.** *For any valid pseudo-expectation in Step 3, with high probability we get an unit vector c that satisfies $T(c, c, c) \ge 0.99$, and c is far from all the previously found a_i's.*

Proof. By Lemma 16 we know there must be a vector a_i such that $\widetilde{\mathbb{E}}[\langle a_i, x\rangle^k] \ge e^{-\epsilon k}$ for sufficiently small constant ϵ. We show that this vector a_i cannot be among the previously found ones. By Lemma 32 we know that for even number k,

$$(\langle s_i, x\rangle + \langle s_i - a_i, x\rangle)^k \le 2^{k-1}(\langle s_i - a_i, x\rangle^k + \langle s_i, x\rangle^k)$$

Taking pseudo-expectations over both sides, we have that

$$\widetilde{\mathbb{E}}[\langle a_i, x\rangle^k] \preceq_{2k} 2^{k-1}(\widetilde{\mathbb{E}}[\langle s_i, x\rangle^k] + k\widetilde{\mathbb{E}}[\langle s_i - a_i, x\rangle^k]) \preceq_{\|x|^2=1,2k} e^{-\epsilon k}$$

where we've used the constraint $\langle s_i, x\rangle^2 \le 1/8$ and induction hypothesis $\|s_i - a_i\| \le 0.1$.

Now applying Theorem 15 we get a vector c that is has inner-product $1 - O(\epsilon)$ with a_i. Therefore $T(c, c, c) = T(a_i, a_i, a_i) + T(c - a_i, a_i, a_i) + T(c, c - a_i, a_i) + T(c, c, a_i) \ge 1 - 1/\log n - 3\|T\|_{\mathrm{inj}}\|c - a_i\| \ge 0.99$. Here $T(x, y, z) = \sum_{i_1,i_2,i_3} T_{i_1,i_2,i_3} x_{i_1} y_{i_2} z_{i_3}$ is the multilinear form for the tensor, and note that this step of the proof does not need to be SoS because we already have the vector c from Theorem 15. ◀

▶ **Claim 28.** *For any unit vector c such that $T(c, c, c) \ge 0.99$, there must be a component a_i such that $\|a_i - c\| \le 0.1$.*

Proof. We define the following trivial pseudo-expectation $\widetilde{\mathbb{E}}^c$ defined by c: $\widetilde{\mathbb{E}}^c[p(x)] = p(c)$. Then we know that $\widetilde{\mathbb{E}}^c$ does satisfy equation $T(x,x,x) \geq 0.99$, and the degree of $\widetilde{\mathbb{E}}^c$ can be any finite number. Therefore, by Lemma 24, we have that $\widetilde{\mathbb{E}}^c\left[\langle a_i, x\rangle^k\right] \geq e^{-(2\epsilon+\delta)k}$ for $k = O(\log n)$. Therefore using the definition of $\widetilde{\mathbb{E}}^c$, we have that $\widetilde{\mathbb{E}}^c\left[\langle a_i, x\rangle^k\right] = \langle a_i, c\rangle^k \geq e^{-(2\epsilon+\delta)k}$. Taking $\epsilon = 0.001$ and then we have that $\langle a_i, c\rangle \geq 0.999 - \delta$ and it follows that $\|a_i - c\| \leq 0.99$. ◀

These three claims finishes the induction in the noiseless case. For the noisy case, we can handle it the same ways as Theorem 9: note that $[\tilde{T} - T](x,x,x) \preceq_{\|x\|^2=1,12} 1/2\log n$ and this additional term does not change any part of the proof.

Finally, the runtime of Line 3 in Algorithm 2 is $n^{O(k)}$, and the run-time of line 4 is also $n^{O(k)}$. Therefore the total runtime is $n^{O(k)}$.

C Matrix Concentrations

In this section we introduce theorems used to prove matrix concentrations. First we need the following lemma for decoupling the randomness in the sum.

▶ **Theorem 29** (Special case of Theorem 1 of [28]). *Let $X_1, \ldots, X_n$, $Y_1, \ldots, Y_n$ are independent random variables on a measurable space over S, where X_i and Y_i has the same distribution for $i = 1, \ldots, n$. Let $f_{ij}(\cdot,\cdot)$ be a family of functions taking $S \times S$ to a Banach space $(B, \|\cdot\|)$. Then there exists absolute constant C, such that for all $n \geq 2$, $t > 0$,*

$$\Pr\left[\left\|\sum_{i\neq j} f_{ij}(X_i, X_j)\right\| \geq t\right] \leq C \Pr\left[\left\|\sum_{i\neq j} f_{ij}(X_i, Y_j)\right\| \geq t/C\right]$$

We also need the Matrix Bernstein's Inequality:

▶ **Theorem 30** (Matrix Bernstein, [30]). *Consider a finite sequence $\{X_k\}$ of independent, random symmetric matrices with dimension d. Assume that each random matrix satisfies*

$$\mathbb{E}[X_k] = 0 \text{ and } \|X_k\| \leq R \text{ almost surely.}$$

Then, for all $t \geq 0$,

$$\Pr[\|\sum_k X_k\| \geq t] \leq d \cdot \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right) \text{ where } \sigma^2 := \|\sum_k \mathbb{E}[X_k^2]\|.$$

D Sum-of-Square Proofs

In this section we state some lemmas that can be proved by low-degree SoS proofs. Most of these lemmas can be found in [12] and [9] but we still give the proofs here for completeness.

▶ **Lemma 31.** *[SoS proof for Cauchy-Schwarz] Cauchy-Schwarz inequality can be proved by degree-2 sum of squares proofs,*

$$\left(\sum_{i=1}^n a_i^2\right)\left(\sum_{i=1}^n b_i^2\right) - \left(\sum_i a_i b_i\right)^2 = \sum_{i,j}(a_i b_j - a_j b_i)^2$$

▶ **Lemma 32.** *For any vector x, y, we have that for even number k,*

$$\|x + y\|^k \preceq_k 2^{k-1}(\|x\|^k + \|y\|^k)$$

Proof. Note that it suffices to prove it for one dimensional vector x, y. We prove by induction. For $k = 2$, it just follows Cauchy-Schwarz. Suppose it is true for $k-2$ case, we have

$$(x+y)^k = (x+y)^{k-2}(x+y)^2 \preceq 2^{k-3}(x^{k-2}+y^{k-2}) \cdot 2(x^2+y^2)$$

Note that

$$2(x^k+y^k) - (x^{k-2}+y^{k-2})(x^2+y^2) = (x^2-y^2)^2(x^{k-4}+x^{k-6}y^2+\cdots+y^{k-4}) \succeq 0$$

Combing the two equations above we obtain the desired result. ◀

▶ **Lemma 33.** *Suppose M is $m \times n$ matrix with spectral norm $\|M\|$, then*

$$(x^T My)^2 \preceq_4 \|x\|^2\|y\|^2\|M\|^2$$

Proof. Assume $m \le n$ without loss of generality, and suppose M has singular decomposition $M = U\Sigma V^T$ where $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_m)$. Let $z = x^T U$ and $w = V^T y$. Then

$$(x^T My)^2 = \left(\sum_{i=1}^m \sigma_i z_i w_i\right)^2 \preceq_4 \left(\sum_{i=1}^m \sigma_i^2 z_i^2\right)\left(\sum_{i=1}^m w_i^2\right) \le \|M\|^2\|z\|^2\|w\|^2 = \|x\|^2\|y\|^2\|M\|^2$$

◀

▶ **Lemma 34.** *For a nonnegative real number a and a set of polynomial R and positive integer k, if a polynomial $p(x)$ satisfy $p(x) \preceq_{R,k} a^2$, then $p(x) \preceq_{R,k'} a$ for $k' = \max\{k, 2\deg(p)\}$.*

Proof. By a simple manipulation of algebra, we have that

$$p(x) - a \preceq_{R,k} \frac{1}{2a}(p(x)-a)^2 \preceq_{R,k'} 0.$$

◀