JMLR: Workshop and Conference Proceedings vol 40:1–51, 2015

Tensor principal component analysis via sum-of-squares proofs

Sam Hopkins SAMHOP@CS.CORNELL.EDU
Jonathan Shi JSHI@CS.CORNELL.EDU
David Steurer DSTEURER@CS.CORNELL.EDU
Cornell University

Abstract

We study a statistical model for the *tensor principal component analysis problem* introduced by Montanari and Richard: Given a order-3 tensor $\mathbf{T}$ of the form $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$, where $\tau \geqslant 0$ is a signal-to-noise ratio, v_0 is a unit vector, and $\mathbf{A}$ is a random noise tensor, the goal is to recover the planted vector v_0. For the case that $\mathbf{A}$ has iid standard Gaussian entries, we give an efficient algorithm to recover v_0 whenever $\tau \geqslant \omega(n^{3/4} \log(n)^{1/4})$, and certify that the recovered vector is close to a maximum likelihood estimator, all with high probability over the random choice of $\mathbf{A}$. The previous best algorithms with provable guarantees required $\tau \geqslant \Omega(n)$.

In the regime $\tau \leqslant o(n)$, natural tensor-unfolding-based spectral relaxations for the underlying optimization problem break down. To go beyond this barrier, we use convex relaxations based on the sum-of-squares method. Our recovery algorithm proceeds by rounding a degree-4 sum-of-squares relaxations of the maximum-likelihood-estimation problem for the statistical model. To complement our algorithmic results, we show that degree-4 sum-of-squares relaxations break down for $\tau \leqslant O(n^{3/4}/\log(n)^{1/4})$, which demonstrates that improving our current guarantees (by more than logarithmic factors) would require new techniques or might even be intractable.

Finally, we show how to exploit additional problem structure in order to solve our sum-of-squares relaxations, up to some approximation, very efficiently. Our fastest algorithm runs in nearly-linear time using shifted (matrix) power iteration and has similar guarantees as above. The analysis of this algorithm also confirms a variant of a conjecture of Montanari and Richard about singular vectors of tensor unfoldings.

Keywords: Tensors, principal component analysis, parameter estimation, sum-of-squares method, semidefinite programming, spectral algorithms, shifted power iteration.

1. Introduction

Principal component analysis (PCA), the process of identifying a direction of largest possible variance from a matrix of pairwise correlations, is among the most basic tools for data analysis in a wide range of disciplines. In recent years, variants of PCA have been proposed that promise to give better statistical guarantees for many applications. These variants include restricting directions to the nonnegative orthant (nonnegative matrix factorization) or to directions that are sparse linear combinations of a fixed basis (SPARSE PCA). Often we have access to not only pairwise but also higher-order correlations. In this case, an analog of PCA is to find a direction with largest possible third moment or other higher-order moment (higher-order PCA or TENSOR PCA).

All of these variants of PCA share that the underlying optimization problem is NP-hard for general instances (often even if we allow approximation), whereas vanilla PCA boils down to an efficient eigenvector computation for the input matrix. However, these hardness result are not predictive in statistical settings where inputs are drawn from particular families of distributions, where efficent algorithm can often achieve much stronger guarantees than for general instances. Understanding the power and limitations of efficient algorithms for statistical models of NP-hard optimization problems is typically very challenging: it is not clear what kind of algorithms can exploit the additional structure afforded by statistical instances, but, at the same time, there are very few tools for reasoning about the computational complexity of statistical / average-case problems. (See (Berthet and Rigollet, 2013) and (Barak et al., 2013) for discussions about the computational complexity of statistical models for SPARSE PCA and random constraint satisfaction problems.)

We study a statistical model for the *tensor principal component analysis problem* introduced by Montanari and Richard (2014) through the lens of a meta-algorithm called the sum-of-squares method, based on semidefinite programming. This method can capture a wide range of algorithmic techniques including linear programming and spectral algorithms. We show that this method can exploit the structure of statistical TENSOR PCA instances in non-trivial ways and achieves guarantees that improve over the previous ones. On the other hand, we show that those guarantees are nearly tight if we restrict the complexity of the sum-of-squares meta-algorithm at a particular level. This result rules out better guarantees for a fairly wide range of potential algorithms. Finally, we develop techniques to turn algorithms based on the sum-of-squares meta-algorithm into algorithms that are truely efficient (and even easy to implement).

Montanari and Richard propose the following statistical model[1] for TENSOR PCA.

Problem 1 (Spiked Tensor Model for TENSOR PCA, Asymmetric) *Given an input tensor* $\mathbf{T} = \tau \cdot v^{\otimes 3} + \mathbf{A}$, *where* $v \in \mathbb{R}^n$ *is an arbitrary unit vector,* $\tau \geqslant 0$ *is the signal-to-noise ratio, and* $\mathbf{A}$ *is a random noise tensor with iid standard Gaussian entries, recover the signal* v *approximately.*

Montanari and Richard show that when $\tau \leqslant o(\sqrt{n})$ Problem 1 becomes information-theoretically unsovlable, while for $\tau \geqslant \omega(\sqrt{n})$ the maximum likelihood estimator (MLE) recovers v' with $\langle v, v' \rangle \geqslant 1 - o(1)$.

The maximum-likelihood-estimator (MLE) problem for Problem 1 is an instance of the following meta-problem for $k = 3$ and $f : x \mapsto \sum_{ijk} \mathbf{T}_{ijk} x_i x_j x_k$ (Montanari and Richard, 2014).

Problem 2 *Given a homogeneous, degree-*k *function* f *on* $\mathbb{R}^n$, *find a unit vector* $v \in \mathbb{R}^n$ *so as to maximize* $f(v)$ *approximately.*

For $k = 2$, this problem is just an eigenvector computation. Already for $k = 3$, it is NP-hard. Our algorithms proceed by relaxing Problem 2 to a convex problem. The latter can be solved either exactly or approximately (as will be the case of our faster algorithms). Under the Gaussian assumption on the noise in Problem 1, we show that for $\tau \geqslant \omega(n^{3/4} \log(n)^{1/4})$ the relaxation does not substantially change the global optimum.

Noise Symmetry. Montanari and Richard actually consider two variants of this model. The first we have already described. In the second, the noise is symmetrized, (to match the symmetry of potential signal tensors $v^{\otimes 3}$).

Problem 3 (Spiked Tensor Model for TENSOR PCA, Symmetric) *Given an input tensor* $\mathbf{T} = \tau \cdot v^{\otimes 3} + \mathbf{A}$, *where* $v \in \mathbb{R}^n$ *is an arbitrary unit vector,* $\tau \geqslant 0$ *is the signal-to-noise ratio, and* $\mathbf{A}$ *is a random* symmetric *noise tensor—that is,* $A_{ijk} = A_{\pi(i)\pi(j)\pi(k)}$ *for any permutation* π*—with otherwise iid standard Gaussian entries, recover the signal* v *approximately.*

[1] Montanari and Richard use a different normalization for the signal-to-noise ratio. Using their notation, $\beta \approx \tau / \sqrt{n}$.

It turns out that for our algorithms based on the sum-of-squares method, this kind of symmetrization is already built-in. Hence there is no difference between Problem 1 and Problem 3 for those algorithms. For our faster algorithms, such symmetrization is not built in. Nonetheless, we show that a variant of our nearly-linear-time algorithm for Problem 1 also solves Problem 3 with matching guarantees.

1.1. Results

Sum-of-squares relaxation. We consider the degree-4 sum-of-squares relaxation for the MLE problem. (See Section 1.2 for a brief discussion about sum-of-squares. All necessary definitions are in Section 2. See (Barak and Steurer, 2014) for more detailed discussion.) Note that the planted vector v has objective value $(1 - o(1))\tau$ for the MLE problem with high probability (assuming $\tau = \Omega(\sqrt{n})$ which will always be the case for us).

Theorem 1 *There exists a polynomial-time algorithm based on the degree-4 sum-of-squares relaxation for the MLE problem that given an instance of Problem 1 or Problem 3 with $\tau \geqslant n^{3/4}(\log n)^{1/4}/\varepsilon$ outputs a unit vector v' with $\langle v, v' \rangle \geqslant 1 - O(\varepsilon)$ with probability $1 - O(n^{-10})$ over the randomness in the input. Furthermore, the algorithm works by rounding any solution to the relaxation with objective value at least $(1 - o(1))\tau$. Finally, the algorithm also certifies that all unit vectors bounded away from v' have objective value significantly smaller than τ for the MLE problem Problem 2.*

We complement the above algorithmic result by the following lower bound.

Theorem 2 (Informal Version) *There is $\tau : \mathbb{N} \to \mathbb{R}$ with $\tau \leqslant O(n^{3/4}/\log(n)^{1/4})$ so that when $\mathbf{T}$ is an instance of Problem 1 with signal-to-noise ratio τ, with probability $1 - O(n^{-10})$, there exists a solution to the degree-4 sum-of-squares relaxation for the MLE problem with objective value at least τ that does not depend on the planted vector v. In particular, no algorithm can reliably recover from this solution a vector v' that is significantly correlated with v.*

Faster algorithms. We interpret a tensor-unfolding algorithm studied by Montanari and Richard as a spectral relaxation of the degree-4 sum-of-squares program for the MLE problem. This interpretation leads to an analysis that gives better guarantees in terms of signal-to-noise ratio τ and also informs a more efficient implementation based on shifted matrix power iteration.

Theorem 3 *There exists an algorithm with running time $\tilde{O}(n^3)$, which is linear in the size of the input, that given an instance of Problem 1 or Problem 3 with $\tau \geqslant n^{3/4}/\varepsilon$ outputs with probability $1 - O(n^{-10})$ a unit vector v' with $\langle v, v' \rangle \geqslant 1 - O(\varepsilon)$.*

We remark that unlike the previous polynomial-time algorithm this linear time algorithm does not come with a certification guarantee. In Section 4.1, we show that small adversarial perturbations can cause this algorithm to fail, whereas the previous algorithm is robust against such perturbations. We also devise an algorithm with the certification property and running time $\tilde{O}(n^4)$ (which is subquadratic in the size n^3 of the input).

Theorem 4 *There exists an algorithm with running time $\tilde{O}(n^4)$ (for inputs of size n^3) that given an instance of Problem 1 with $\tau \geqslant n^{3/4}(\log n)^{1/4}/\varepsilon$ for some ε, outputs with probability $1 - O(n^{-10})$ a unit vector v' with $\langle v, v' \rangle \geqslant 1 - O(\varepsilon)$ and certifies that all vectors bounded away from v' have MLE objective value significantly less than τ.*

Higher-order tensors. Our algorithmic results also extend in a straightforward way to tensors of order higher than 3. (See Section C for some details.) For simplicity we give some of these results only for the higher-order analogue of Problem 1; we conjecture however that all our results for Problem 3 generalize in similar fashion.

Theorem 5 *Let k be an odd integer, $v_0 \in \mathbb{R}^n$ a unit vector, $\tau \geqslant n^{k/4} \log(n)^{1/4}/\varepsilon$, and $\mathbf{A}$ an order-k tensor with independent unit Gaussian entries. Let $\mathbf{T}(x) = \tau \cdot \langle v_0, x\rangle^k + \mathbf{A}(x)$.*

1. *There is a polynomial-time algorithm, based on semidefinite programming, which on input $\mathbf{T}(x) = \tau \cdot \langle v_0, x\rangle^k + \mathbf{A}(x)$ returns a unit vector v with $\langle v_0, v\rangle \geqslant 1 - O(\varepsilon)$ with probability $1 - O(n^{-10})$ over random choice of $\mathbf{A}$.*
2. *There is a polynomial-time algorithm, based on semidefinite programming, which on input $\mathbf{T}(x) = \tau \cdot \langle v_0, x\rangle^k + \mathbf{A}(x)$ certifies that $\mathbf{T}(x) \leqslant \tau \cdot \langle v, x\rangle^k + O(n^{k/4} \log(n)^{1/4})$ for some unit v with probability $1 - O(n^{-10})$ over random choice of $\mathbf{A}$. This guarantees in particular that v is close to a maximum likelihood estimator for the problem of recovering the signal v_0 from the input $\tau \cdot v_0^{\otimes k} + \mathbf{A}$.*
3. *By solving the semidefinite relaxation approximately, both algorithms can be implemented in time $\tilde{O}(m^{1+1/k})$, where $m = n^k$ is the input size.*

For even k, the above all hold, except now we recover v with $\langle v_0, v\rangle^2 \geqslant 1 - O(\varepsilon)$, and the algorithms can be implemented in nearly linear time.

Remark 6 *When $\mathbf{A}$ is a* symmetric *noise tensor (the higher-order analogue of Problem 3), (1–2) above hold. We conjecture that (3) does as well.*

The last theorem, the higher-order generalization of Theorem 3, almost completely resolves a conjecture of Montanari and Richard regarding tensor unfolding algorithms for odd k. We are able to prove their conjectured signal-to-noise ratio τ for an algorithm that works mainly by using an unfolding of the input tensor, but our algorithm includes an extra random-rotation step to handle sparse signals. We conjecture but cannot prove that the necessity of this step is an artifact of the analysis.

Theorem 7 *Let k be an odd integer, $v_0 \in \mathbb{R}^n$ a unit vector, $\tau \geqslant n^{k/4}/\varepsilon$, and $\mathbf{A}$ an order-k tensor with independent unit Gaussian entries. There is a nearly-linear-time algorithm, based on tensor unfolding, which, with probability $1 - O(n^{-10})$ over random choice of $\mathbf{A}$, recovers a vector v with $\langle v, v_0\rangle^2 \geqslant 1 - O(\varepsilon)$. This continues to hold when $\mathbf{A}$ is replaced by a* symmetric *noise tensor (the higher-order analogue of Problem 3).*

1.2. Techniques

We arrive at our results via an analysis of Problem 2 for the function $f(x) = \sum_{ijk} \mathbf{T}_{ijk} x_i x_j x_k$, where $\mathbf{T}$ is an instance of Problem 1. The function f decomposes as $f = g + h$ for a signal $g(x) = \tau \cdot \langle v, x\rangle^3$ and noise $h(x) = \sum_{ijk} a_{ijk} x_i x_j x_j$ where $\{a_{ijk}\}$ are iid standard Gaussians. The signal g is maximized at $x = v$, where it takes the value τ. The noise part, h, is with high probability at most $\tilde{O}(\sqrt{n})$ over the unit sphere. We have insisted that τ be much greater than $\sqrt{n}$, so f has a unique global maximum, dominated by the signal g. The main problem is to find it.

To maximize g, we apply the *Sum of Squares meta-algorithm* (SoS). SoS provides a hierarchy of strong convex relaxations of Problem 2. Using convex duality, we can recast the optimization problem as one of efficiently certifying the upper bound on h which shows that optima of g are dominated by the signal. SoS efficiently finds boundedness certificates for h of the form

$$c - h(x) = s_1(x)^2 + \cdots + s_k(x)^2$$

where "=" denotes equality in the ring $\mathbb{R}[x]/(\|x\|^2 - 1)$ and where $s_1, \ldots, s_k$ have bounded degree, when such certificates exist. (The polynomials $\{s_i\}$ and $\{t_j\}$ certify that $h(x) \leqslant c$. Otherwise $c - h(x)$ would be negative, but this is impossible by the nonnegativity of squared polynomials.)

Our main technical contribution is an almost-complete characterization of certificates like these for such degree-3 random polynomials h when the polynomials $\{s_i\}$ have degree at most four. In particular, we show that with high probability in the random case a degree-4 certificate exists for $c = \tilde{O}(n^{3/4})$, and that with high probability, no significantly better degree-four certificate exists.

We apply this characterization in three ways to obtain three different algorithms. The first application shows the existence of a polynomial-time algorithm that maximizes f when $\tau \geqslant \tilde{\Omega}(n^{3/4})$ (and thus solves TPCA in the spiked tensor model for $\tau \geqslant \tilde{\Omega}(n^{3/4})$.) This first algorithm involves solving a large semidefinite program associated to the SoS relaxation. As a second application of this characterization, we avoid solving the semidefinite program. Instead, we give an algorithm running in time $\tilde{O}(n^4)$ which quickly constructs only a small portion of an almost-optimal SoS boundedness certificate; in the random case this turns out to be enough to find the signal v and certify the boundedness of g. (Note that this running time is only a factor of $n \operatorname{polylog} n$ greater than the input size n^3.)

Finally, we analyze a third algorithm for TPCA which simply computes the highest singular vector of a matrix unfolding of the input tensor. This algorithm was considered in depth by Montanari and Richard, who fully characterized its behavior in the case of even-order tensors (corresponding to $k = 4, 6, 8, \cdots$ in Problem 2). They conjectured that this algorithm successfully recovers the signal v at the signal-to-noise ratio τ of Theorem 4 for Problem 1 and Problem 3. Up to an extra random rotations step before the tensor unfolding in the case that the input comes from Problem 3 (and up to logarithmic factors in τ) we confirm their conjecture. We observe that their algorithm can be viewed as a method of rounding a non-optimal solution to the SoS relaxation to find the signal. We show, also, that for $k = 4$, the degree-4 SoS relaxation does no better than the simpler tensor unfolding algorithm as far as signal-to-noise ratio is concerned. However, for odd-order tensors this unfolding algorithm does not certify its own success in the way our other algorithms do.

1.3. Related Work

There is a vast literature on tensor analogues of linear algebra problems—too vast to attempt any survey here. Tensor methods for machine learning, in particular for learning latent variable models, have garnered recent attention, e.g., with works of Anandkumar et al. (Anandkumar et al., 2014b, 2013). These approaches generally involve decomposing a tensor which captures some aggregate statistics of input data into rank-one components. A recent series of papers analyzes the tensor power method, a direct analogue of the matrix power method, as a way to find rank-one components of random-case tensors (Anandkumar et al., 2014c,a).

Another recent line of work applies the Sum of Squares (a.k.a. Lasserre or Lasserre/Parrilo) hierarchy of convex relaxations to learning problems. See the survey of Barak and Steurer for references and discussion of these relaxations (Barak and Steurer, 2014). Barak, Kelner, and Steurer show how to use SoS to efficiently find sparse vectors planted in random linear subspaces, and the same authors give an algorithm for dictionary learning with strong provable statistical guarantees (Barak et al., 2014a,b). These algorithms, too, proceed by decomposition of an underlying random tensor; they exploit the strong (in many cases, the strongest-known) algorithmic guarantees offered by SoS for this problem in a variety of average-case settings.

Concurrently and independently of us, and also inspired by the recently-discovered applicability of tensor methods to machine learning, Barak and Moitra use SoS techniques formally related to ours to address the *tensor prediction* problem: given a low-rank tensor (perhaps measured with noise) only a subset of whose entries are revealed, predict the rest of the tensor entries (Barak and Moitra, 2015). They work with worst-case noise and study the number of revealed entries necessary for the SoS hierarchy to successfully predict the tensor. By constrast, in our setting, the entire tensor is revealed, and we study the signal-to-noise threshold necessary for SoS to recover its principal component under distributional assumptions on the noise that allow us to avoid worst-case hardness behavior.

Since Barak and Moitra work in a setting where few tensor entries are revealed, they are able to use algorithmic techniques and lower bounds from the study of sparse random constraint satisfaction problems (CSPs), in particular random 3XOR (Goerdt and Krivelevich, 2001; Friedman et al., 2005; Feige and Ofek, 2007; Feige et al., 2006). The tensors we study are much denser. In spite of the density (and even though our setting is real-valued), our algorithmic techniques are related to the same spectral refutations of random CSPs. However our lower bound techniques do not seem to be related to the proof-complexity techniques that go into sum-of-squares lower bound results for random CSPs.

The analysis of tractable tensor decomposition in the rank one plus noise model that we consider here (the *spiked tensor model*) was initiated by Montanari and Richard, whose work inspired the current paper (Montanari and Richard, 2014). They analyze a number of natural algorithms and find that tensor unfolding algorithms, which use the spectrum of a matrix unfolding of the input tensor, are most robust to noise. Here we consider more powerful convex relaxations, and in the process we tighten Montanari and Richard's analysis of tensor unfolding in the case of odd-order tensors.

Related to our lower bound, Montanari, Reichman, and Zeitouni prove strong impossibility results for the problem of detecting rank-one perturbations of Gaussian matrices and tensors using *any* eigenvalue of the matrix or unfolded tensor; they are able to characterize the precise threshold below which the entire spectrum of a perturbed noise matrix or unfolded tensor becomes indistinguishable from pure noise (Montanari et al., 2014). Our lower bounds are much coarser, applying only to the objective value of a relaxation (the analogue of just the top eigenvalue of an unfolding), but they apply to all degree-4 SoS-based algorithms, which are *a priori* a major generalization of spectral methods.

2. Preliminaries

2.1. Notation

We use $x = (x_1, \ldots, x_n)$ to denote a vector of indeterminates. The letters u, v, w are generally reserved for real vectors. The letters α, β are reserved for multi-indices; that is, for tuples $(i_1, \ldots, i_k)$ of indices. For $f, g : \mathbb{N} \to \mathbb{R}$ we write $f \lesssim g$ for $f = O(g)$ and $f \gtrsim g$ for $f = \Omega(g)$. We write $f = \tilde{O}(g)$ if $f(n) \leqslant g(n) \cdot \operatorname{polylog} n$, and $f = \tilde{\Omega}(g)$ if $f \geqslant g(n)/\operatorname{polylog} n$.

We employ the usual Loewner (a.k.a. positive semi-definite) ordering $\succeq$ on Hermitian matrices.

We will be heavily concerned with tensors and matrix flattenings thereof. In general, boldface capital letters $\mathbf{T}$ denote tensors and ordinary capital letters denote matrices A. We adopt the convention that unless otherwise noted for a tensor $\mathbf{T}$ the matrix T is the squarest-possible unfolding of $\mathbf{T}$. If $\mathbf{T}$ has even order k then T has dimensions $n^{k/2} \times n^{k/2}$. For odd k it has dimensions $n^{\lfloor k/2 \rfloor} \times n^{\lceil k/2 \rceil}$. All tensors, matrices, vectors, and scalars in this paper are real.

We use $\langle \cdot, \cdot \rangle$ to denote the usual entrywise inner product of vectors, matrices, and tensors. For a vector v, we use $\|v\|$ to denote its ℓ_2 norm. For a matrix A, we use $\|A\|$ to denote its operator norm (also known as the spectral or ℓ_2-to-ℓ_2 norm).

For a k-tensor $\mathbf{T}$, we write $\mathbf{T}(v)$ for $\langle v^{\otimes k}, \mathbf{T} \rangle$. Thus, $\mathbf{T}(x)$ is a homogeneous real polynomial of degree k.

We use $\mathcal{S}_k$ to denote the symmetric group on k elements. For a k-tensor $\mathbf{T}$ and $\pi \in \mathcal{S}_k$, we denote by $\mathbf{T}^\pi$ the k-tensor with indices permuted according to π, so that $\mathbf{T}^\pi_\alpha = \mathbf{T}_{\pi^{-1}(\alpha)}$. A tensor $\mathbf{T}$ is symmetric if for all $\pi \in \mathcal{S}_k$ it is the case that $\mathbf{T}^\pi = \mathbf{T}$. (Such tensors are sometimes called "supersymmetric.")

For clarity, most of our presentation focuses on 3-tensors. For an $n \times n$ 3-tensor $\mathbf{T}$, we use T_i to denote its $n \times n$ matrix slices along the first mode, i.e., $(T_i)_{j,k} = \mathbf{T}_{i,j,k}$.

We often say that an sequence $\{E_n\}_{n \in \mathbb{N}}$ of events occurs with high probability, which for us means that $\mathbb{P}(E_n \text{ fails}) = O(n^{-10})$. (Any other n^{-c} would do, with appropriate modifications of constants elsewhere.)

2.2. Polynomials and Matrices

Let $\mathbb{R}[x]_{\leqslant d}$ be the vector space of polynomials with real coefficients in variables $x = (x_1, \ldots, x_n)$, of degree at most d. We can represent a homogeneous even-degree polynomial $p \in \mathbb{R}[x]_d$ by an $n^{d/2} \times n^{d/2}$ matrix: a matrix M is a *matrix representation* for p if $p(x) = \langle x^{\otimes d/2}, M x^{\otimes d/2} \rangle$. If p has a matrix representation $M \succeq 0$, then $p = \sum_i p_i(x)^2$ for some polynomials p_i.

2.3. The Sum of Squares (SoS) Algorithm

Definition 8 *Let $\mathcal{L} : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$ be a linear functional on polynomials of degree at most d for some d even. Suppose that*

- $\mathcal{L}\, 1 = 1$.
- $\mathcal{L}\, p(x)^2 \geqslant 0$ *for all* $p \in \mathbb{R}[x]_{\leqslant d/2}$.

Then $\mathcal{L}$ is a degree-d pseudo-expectation. We often use the suggestive notation $\tilde{\mathbb{E}}$ for such a functional, and think of $\tilde{\mathbb{E}}\, p(x)$ as giving the expectation of the polynomial $p(x)$ under a pseudo-distribution *over $\{x\}$.*

For $p \in \mathbb{R}[x]_{\leqslant d}$ we say that the pseudo-distribution $\{x\}$ (or, equivalently, the functional $\tilde{\mathbb{E}}$) satisfies $\{p(x) = 0\}$ if $\tilde{\mathbb{E}}\, p(x)q(x) = 0$ for all $q(x)$ such that $p(x)q(x) \in \mathbb{R}[x]_{\leqslant d}$.

Pseudo-distributions were first introduced in (Barak et al., 2012) and are surveyed in (Barak and Steurer, 2014).

We employ the standard result that, up to negligible issues of numerical accuracy, if there exists a degree-d pseudo-distribution satisfying constraints $\{p_0(x) = 0, \ldots, p_m(x) = 0\}$, then it can be found in time $n^{O(d)}$ by solving a semidefinite program of size $n^{O(d)}$. (See (Barak and Steurer, 2014) for references.)

3. Certifying Bounds on Random Polynomials

Let $f \in \mathbb{R}[x]_d$ be a homogeneous degree-d polynomial. When d is even, f has square matrix representations of dimension $n^{d/2} \times n^{d/2}$. The maximal eigenvalue of a matrix representation M of f provides a natural certifiable upper bound on $\max_{\|v\|=1} f(v)$, as

$$f(v) = \langle v^{\otimes d/2}, M v^{\otimes d/2} \rangle \leqslant \max_{w \in \mathbb{R}^{n^{d/2}}} \frac{\langle w, Mw \rangle}{\langle w, w \rangle} = \|M\| \,.$$

When $f(x) = \mathbf{A}(x)$ for an even-order tensor $\mathbf{A}$ with independent random entries, the quality of this certificate is well characterized by random matrix theory. In the case where the entries of $\mathbf{A}$ are standard Gaussians, for instance, $\|M\| = \|A + A^T\| \leqslant \tilde{O}(n^{d/4})$ with high probability, thus certifying that $\max_{\|v\|=1} f(v) \leqslant \tilde{O}(n^{d/4})$.

A similar story applies to f of odd degree with random coefficients, but with a catch: the certificates are not as good. For example, we expect a degree-3 random polynomial to be a smaller and simpler object than one of degree-4, and so we should be able to certify a tighter upper bound on $\max_{\|v\|=1} f(v)$. The matrix representations of f are now rectangular $n^2 \times n$ matrices whose top singular values are certifiable upper bounds on $\max_{\|v\|=1} f(v)$. But in random matrix theory, this maximum singular value depends (to a first approximation) only on the longer dimension n^2, which is the same here as in the degree-4 case. Again when $f(x) = \mathbf{A}(x)$, this time where $\mathbf{A}$ is an order-3 tensor of independent standard Gaussian entries, $\|M\| = \sqrt{\|AA^T\|} \geqslant \tilde{\Omega}(n)$, so that this method cannot certify better than $\max_{\|v\|=1} f(v) \leqslant \tilde{O}(n)$. Thus, the natural spectral certificates are unable to exploit the decrease in degree from 4 to 3 to improve the certified bounds.

To better exploit the benefits of square matrices, we bound the maxima of degree-3 homogeneous f by a degree-4 polynomial. In the case that f is multi-linear, we have the polynomial identity $f(x) = \frac{1}{3}\langle x, \nabla f(x) \rangle$. Using Cauchy-Schwarz, we then get $f(x) \leqslant \frac{1}{3}\|x\| \|\nabla f(x)\|$. This inequality suggests using the degree-4 polynomial $\|\nabla f(x)\|^2$ as a bound on f. Note that local optima of f on the sphere occur where $\nabla f(v) \propto v$, and so this bound is tight at local maxima. Given a random homogeneous f, we will associate a degree-4 polynomial related to $\|\nabla f\|^2$ and show that this polynomial yields the best possible degree-4 SoS-certifiable bound on $\max_{\|v\|=1} f(v)$.

Definition 9 *Let $f \in \mathbb{R}[x]_3$ be a homogeneous degree-3 polynomial with indeterminates $x = (x_1, \ldots, x_n)$. Suppose $A_1, \ldots, A_n$ are matrices such that $f = \sum_i x_i \langle x, A_i x \rangle$. We say that f is λ-bounded if there are matrices $A_1, \ldots, A_n$ as above and a matrix representation M of $\|x\|^4$ so that $\sum_i A_i \otimes A_i \preceq \lambda^2 \cdot M$.*

We observe that for f multi-linear in the coordinates x_i of x, up to a constant factor we may take the matrices A_i to be matrix representations of $\partial_i f$, so that $\sum_i A_i \otimes A_i$ is a matrix representation of the polynomial $\|\nabla f\|^2$. This choice of A_i may not, however, yield the optimal spectral bound λ^2.

The following theorem is the reason for our definition of λ-boundedness.

Theorem 10 *Let $f \in \mathbb{R}[x]_3$ be λ-bounded. Then $\max_{\|v\|=1} f(v) \leqslant \lambda$, and the degree-4 SoS algorithm certifies this. In particular, every degree-4 pseudo-distribution $\{x\}$ over $\mathbb{R}^n$ satisfies*

$$\tilde{\mathbb{E}} f \leqslant \lambda \cdot \left(\tilde{\mathbb{E}} \|x\|^4\right)^{3/4} \,.$$

Proof By Cauchy–Schwarz for pseudo-expectations, the pseudo-distribution satisfies $\left(\tilde{\mathbb{E}}\|x\|^2\right)^2 \leqslant \tilde{\mathbb{E}}\|x\|^4$ and $\left(\tilde{\mathbb{E}}\sum_i x_i\langle x, A_i x\rangle\right)^2 \leqslant \left(\tilde{\mathbb{E}}\sum_i x_i^2\right) \cdot \left(\sum_i \langle x, A_i x\rangle^2\right)$. Therefore,

$$\begin{aligned}
\tilde{\mathbb{E}} f &= \tilde{\mathbb{E}} \sum\nolimits_i x_i \cdot \langle x, A_i x\rangle \\
&\leqslant \left(\tilde{\mathbb{E}} \sum\nolimits_i x_i^2\right)^{1/2} \cdot \left(\tilde{\mathbb{E}} \sum\nolimits_i \langle x, A_i x\rangle^2\right)^{1/2} \\
&= \left(\tilde{\mathbb{E}}\|x\|^2\right)^{1/2} \cdot \left(\tilde{\mathbb{E}}\langle x^{\otimes 2}, \left(\sum\nolimits_i A_i \otimes A_i\right) x^{\otimes 2}\rangle\right)^{1/2} \\
&\leqslant \left(\tilde{\mathbb{E}}\|x\|^4\right)^{1/4} \cdot \left(\tilde{\mathbb{E}}\langle x^{\otimes 2}, \lambda^2 \cdot M x^{\otimes 2}\rangle\right)^{1/2} \\
&= \lambda \cdot \left(\tilde{\mathbb{E}}\|x\|^4\right)^{3/4} .
\end{aligned}$$

The last inequality also uses the premise $(\sum_i A_i \otimes A_i) \preceq \lambda^2 \cdot M$ for some matrix representation M of $\|x\|^4$, in the following way. Since $M' := \lambda^2 \cdot M - (\sum_i A_i \otimes A_i) \succeq 0$, the polynomial $\langle x^{\otimes 2}, M' x^{\otimes 2}\rangle$ is a sum of squared polynomials. Thus, $\tilde{\mathbb{E}}\langle x^{\otimes 2}, M' x^{\otimes 2}\rangle \geqslant 0$ and the desired inequality follows. ■

We now state the degree-3 case of a general λ-boundedness fact for homogeneous polynomials with random coefficients. The SoS-certifiable bound for a random degree-3 polynomial this provides is the backbone of our SoS algorithm for tensor PCA in the spiked tensor model.

Theorem 11 *Let $\mathbf{A}$ be a 3-tensor with independent entries from $\mathcal{N}(0,1)$. Then $\mathbf{A}(x)$ is λ-bounded with $\lambda = O(n^{3/4} \log(n)^{1/4})$, with high probability.*

The full statement and proof of this theorem, generalized to arbitrary-degree homogeneous polynomials, may be found as Theorem 56; we prove the statement above as a corollary in Section E. Here provide a proof sketch.

Proof [Proof sketch] We first note that the matrix slices A_i of $\mathbf{A}$ satisfy $\mathbf{A}(x) = \sum_i x_i\langle x, A_i x\rangle$. Using the matrix Bernstein inequality, we show that $\sum_i A_i \otimes A_i - \mathbb{E}\sum_i A_i \otimes A_i \preceq O(n^{3/2}(\log n)^{1/2}) \cdot \mathrm{Id}$ with high probability. At the same time, a straightforward computation shows that $\frac{1}{n}\mathbb{E}\sum_i A_i \otimes A_i$ is a matrix representation of $\|x\|^4$. Since Id is as well, we get that $\sum_i A_i \otimes A_i \preceq \lambda^2 \cdot M$, where M is some matrix representation of $\|x\|^4$ which combines Id and $\mathbb{E}\sum_i A_i \otimes A_i$, and $\lambda = O(n^{3/4}(\log n)^{1/4})$. ■

Corollary 12 *Let $\mathbf{A}$ be a 3-tensor with independent entries from $\mathcal{N}(0,1)$. Then, with high probability, the degree-4 SoS algorithm certifies that $\max_{\|v\|=1} \mathbf{A}(v) \leqslant O(n^{3/4}(\log n)^{1/4})$. Furthermore, also with high probability, every pseudo-distribution $\{x\}$ over $\mathbb{R}^n$ satisfies*

$$\tilde{\mathbb{E}}\,\mathbf{A}(x) \leqslant O(n^{3/4}(\log n)^{1/4})(\tilde{\mathbb{E}}\,\|x\|^4)^{3/4} .$$

Proof Immediate by combining Theorem 11 with Theorem 10. ■

4. Polynomial-Time Recovery via Sum of Squares

Here we give our first algorithm for tensor PCA: we analyze the quality of the natural SoS relaxation of tensor PCA using our previous discussion of boundedness certificates for random polynomials, and we show how to round this relaxation. We discuss also the robustness of the SoS-based algorithm to some amount of additional *worst-case* noise in the input. For now, to obtain a solution to the SoS relaxation we will solve a large semidefinite program. Thus, the algorithm discussed here is not yet enough to prove Theorem 4 and Corollary 4: the running time, while still polynomial, is somewhat greater than $\tilde{O}(n^4)$.

Tensor PCA with Semidefinite Programming

Input: $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$, where $v \in \mathbb{R}^n$ and $\mathbf{A}$ is some order-3 tensor.
Goal: Find $v \in \mathbb{R}^n$ with $|\langle v, v_0 \rangle| \geqslant 1 - o(1)$.

Algorithm 1 (Recovery) *Using semidefinite programming, find the degree-*4 *pseudo-distribution* $\{x\}$ *satisfying* $\{\|x\|^2 = 1\}$ *which maximizes* $\tilde{\mathbb{E}}\, \mathbf{T}(x)$. *Output* $\tilde{\mathbb{E}}\, x / \| \tilde{\mathbb{E}}\, x\|$.

Algorithm 2 (Certification) *Run Algorithm 1 to obtain v. Using semidefinite programming, find the degree-*4 *pseudo-distribution* $\{x\}$ *satisfying* $\{\|x\| = 1\}$ *which maximizes* $\tilde{\mathbb{E}}\, \mathbf{T}(x) - \tau \cdot \langle v, x\rangle^3$. *If* $\tilde{\mathbb{E}}\, \mathbf{T}(x) - \tau \cdot \langle v, x\rangle^3 \leqslant O(n^{3/4} \log(n)^{1/4})$, *output* CERTIFY. *Otherwise, output* FAIL.

The following theorem characterizes the success of Algorithm 1 and Algorithm 2

Theorem 13 (Formal version of Theorem 1) *Let* $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$, *where* $v_0 \in \mathbb{R}^n$ *and* $\mathbf{A}$ *has independent entries from* $\mathcal{N}(0,1)$. *Let* $\tau \gtrsim n^{3/4} \log(n)^{1/4}/\varepsilon$. *Then with high probability over random choice of* $\mathbf{A}$, *on input* $\mathbf{T}$ *or* $\mathbf{T}' := \tau \cdot v_0^{\otimes 3} + \frac{1}{|\mathcal{S}_3|} \sum_{\pi \in \mathcal{S}_3} \mathbf{A}^\pi$, *Algorithm 1 outputs a vector* v *with* $\langle v, v_0 \rangle \geqslant 1 - O(\varepsilon)$. *In other words, for this* τ, *Algorithm 1 solves both Problem 1 and Problem 3.*

For any unit $v_0 \in \mathbb{R}^n$ *and* $\mathbf{A}$, *if Algorithm 2 outputs* CERTIFY *then* $\mathbf{T}(x) \leqslant \tau \cdot \langle v, x \rangle^3 + O(n^{3/4} \log(n)^{1/4})$. *For* $\mathbf{A}$ *as described in either Problem 1 or Problem 3 and* $\tau \gtrsim n^{3/4} \log(n)^{1/4}/\varepsilon$, *Algorithm 2 outputs* CERTIFY *with high probability.*

The analysis has two parts. We show that

1. if there exists a sufficiently good upper bound on $\mathbf{A}(x)$ (or in the case of the symmetric noise input, on $\mathbf{A}^\pi(x)$ for every $\pi \in \mathcal{S}_3$) which is degree-4 SoS certifiable, then the vector recovered by the algorithm will be very close to v, and that

2. in the case of $\mathbf{A}$ with independent entries from $\mathcal{N}(0,1)$, such a bound exists with high probability.

Conveniently, Item 2 is precisely the content of Corollary 12. The following lemma expresses Item 1.

Lemma 14 *Suppose* $\mathbf{A}(x) \in \mathbb{R}[x]_3$ *is such that* $|\tilde{\mathbb{E}}\, \mathbf{A}(x)| \leqslant \varepsilon\tau \cdot (\tilde{\mathbb{E}}\, \|x\|^4)^{3/4}$ *for any degree-*4 *pseudo-distribution* $\{x\}$. *Then on input* $\tau \cdot v_0^{\otimes 3} + \mathbf{A}$, *Algorithm 1 outputs a unit vector* v *with* $\langle v, v_0 \rangle \geqslant 1 - O(\varepsilon)$.

Proof Algorithm 1 outputs $v = \tilde{\mathbb{E}}\, x / \| \tilde{\mathbb{E}}\, x\|$ for the pseudo-distribution that it finds, so we'd like to show $\langle v_0, \tilde{\mathbb{E}}\, x / \| \tilde{\mathbb{E}}\, x\| \rangle \geqslant 1 - O(\varepsilon)$. By pseudo-Cauchy-Schwarz (Lemma 47), $\| \tilde{\mathbb{E}}\, x\|^2 \leqslant \tilde{\mathbb{E}}\, \|x\|^2 = 1$, so it will suffice to prove just that $\langle v_0, \tilde{\mathbb{E}}\, x \rangle \geqslant 1 - O(\varepsilon)$.

If $\tilde{\mathbb{E}}\langle v_0, x\rangle^3 \geqslant 1 - O(\varepsilon)$, then by Lemma 50 (and linearity of pseudo-expectation) we would have

$$\langle v_0, \tilde{\mathbb{E}}\, x\rangle \;=\; \tilde{\mathbb{E}}\langle v_0, x\rangle \;\geqslant\; 1 - O(2\varepsilon) = 1 - O(\varepsilon)$$

So it suffices to show that $\tilde{\mathbb{E}}\langle v_0, x\rangle^3$ is close to 1.

Recall that Algorithm 1 finds a pseudo-distribution that maximizes $\tilde{\mathbb{E}}\,\mathbf{T}(x)$. We split $\tilde{\mathbb{E}}\,\mathbf{T}(x)$ into the signal $\tilde{\mathbb{E}}\langle v_0^{\otimes 3}, x^{\otimes 3}\rangle$ and noise $\tilde{\mathbb{E}}\,\mathbf{A}(x)$ components and use our hypothesized SoS upper bound on the noise.

$$\tilde{\mathbb{E}}\,\mathbf{T}(x) \;=\; \tau\cdot(\tilde{\mathbb{E}}\langle v_0^{\otimes 3}, x^{\otimes 3}\rangle) + \tilde{\mathbb{E}}\,\mathbf{A}(x) \;\leqslant\; \tau\cdot(\tilde{\mathbb{E}}\langle v_0^{\otimes 3}, x^{\otimes 3}\rangle) + \varepsilon\tau\,.$$

Rewriting $\langle v_0^{\otimes 3}, x^{\otimes 3}\rangle$ as $\langle v_0, x\rangle^3$, we obtain

$$\tilde{\mathbb{E}}\langle v_0, x\rangle^3 \;\geqslant\; \frac{1}{\tau}\cdot\tilde{\mathbb{E}}\,\mathbf{T}(x) - \varepsilon\,.$$

Finally, there exists a pseudo-distribution that achieves $\tilde{\mathbb{E}}\,\mathbf{T}(x) \geqslant \tau - \varepsilon\tau$. Indeed, the trivial distribution giving probability 1 to v_0 is such a pseudo-distribution:

$$\mathbf{T}(v_0) \;=\; \tau + \mathbf{A}(v_0) \;\geqslant\; \tau - \varepsilon\tau.$$

Putting it together,

$$\tilde{\mathbb{E}}\langle v_0, x\rangle^3 \;\geqslant\; \frac{1}{\tau}\cdot\tilde{\mathbb{E}}\,\mathbf{T}(x) - \varepsilon \;\geqslant\; \frac{(1-\varepsilon)\tau}{\tau} - \varepsilon = 1 - O(\varepsilon)\,.$$

■

Proof [Proof of Theorem 13] We first address Algorithm 1. Let $\tau, \mathbf{T}, \mathbf{T}'$ be as in the theorem statement. By Lemma 14, it will be enough to show that with high probability every degree-4 pseudo-distribution $\{x\}$ has $\tilde{\mathbb{E}}\,\mathbf{A}(x) \leqslant \varepsilon'\tau\cdot(\tilde{\mathbb{E}}\,\|x\|^4)^{3/4}$ and $\frac{1}{\mathcal{S}_3}\,\tilde{\mathbb{E}}\,\mathbf{A}^{\pi}(x) \leqslant \varepsilon'\tau\cdot(\tilde{\mathbb{E}}\,\|x\|^4)^{3/4}$ for some $\varepsilon' = \Theta(\varepsilon)$. By Corollary 12 and our assumptions on τ this happens for each permutation $\mathbf{A}^{\pi}$ individually with high probability, so a union bound over $\mathbf{A}^{\pi}$ for $\pi \in \mathcal{S}_3$ completes the proof.

Turning to Algorithm 2, the simple fact that SoS only certifies true upper bounds implies that the algorithm is never wrong when it outputs CERTIFY. It is not hard to see that whenever Algorithm 1 has succeeded in recovering v because $\tilde{\mathbb{E}}\,\mathbf{A}(x)$ is bounded, which as above happens with high probability, Algorithm 2 will output CERTIFY. ■

4.1. Semi-Random Tensor PCA

We discuss here a modified TPCA model, which will illustrate the qualitative differences between the new tensor PCA algorithms we propose in this paper and previously-known algorithms. The model is semi-random and semi-adversarial. Such models are often used in average-case complexity theory to distinguish between algorithms which work by solving robust maximum-likelihood-style problems and those which work by exploiting some more fragile property of a particular choice of input distribution.

Problem 4 (Tensor PCA in the Semi-Random Model) *Let $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$, where $v_0 \in \mathbb{R}^n$ and $\mathbf{A}$ has independent entries from $\mathcal{N}(0,1)$. Let $Q \in \mathbb{R}^{n \times n}$ with $\| \mathrm{Id} - Q \| \leqslant O(n^{-1/4})$, chosen adversarially depending on $\mathbf{T}$. Let $\mathbf{T}'$ be the 3-tensor whose $n^2 \times n$ matrix flattening is TQ. (That is, each row of T has been multiplied by a matrix which is close to identity.) On input $\mathbf{T}'$, recover v.*

Here we show that Algorithm 1 succeeds in recovering v in the semi-random model.

Theorem 15 *Let $\mathbf{T}'$ be the semi-random-model tensor PCA input, with $\tau \geqslant n^{3/4} \log(n)^{1/4}/\varepsilon$. With high probability over randomness in $\mathbf{T}'$, Algorithm 1 outputs a vector v with $\langle v, v_0 \rangle \geqslant 1 - O(\varepsilon)$.*

Proof By Lemma 14, it will suffice to show that $\mathbf{B} := (\mathbf{T}' - \tau \cdot v_0^{\otimes 3})$ has $\tilde{\mathbb{E}}\, \mathbf{B}(x) \leqslant \varepsilon' \tau \cdot (\tilde{\mathbb{E}}\, \|x\|^4)^{3/4}$ for any degree-4 pseudo-distribution $\{x\}$, for some $\varepsilon' = \Theta(\varepsilon)$. We rewrite $\mathbf{B}$ as

$$\mathbf{B} = (A + \tau \cdot v_0 (v_0 \otimes v_0)^T)(Q - \mathrm{Id}) + A$$

where $\mathbf{A}$ has independent entries from $\mathcal{N}(0,1)$. Let $\{x\}$ be a degree-4 pseudo-distribution. Let $f(x) = \langle x^{\otimes 2}, (A + \tau \cdot v_0(v_0 \otimes v_0)^T)(Q - \mathrm{Id})x \rangle$. By Corollary 12, $\tilde{\mathbb{E}}\, \mathbf{B}(x) = \tilde{\mathbb{E}}\, f(x) + O(n^{3/4} \log(n)^{1/4})(\tilde{\mathbb{E}}\, \|x\|^4)^{3/4}$ with high probability. By triangle inequality and sub-multiplicativity of the operator norm, we get that with high probability

$$\|(A + \tau \cdot v_0(v_0 \otimes v_0))(Q - \mathrm{Id})\| \leqslant (\|A\| + \tau)\|Q - \mathrm{Id}\,\| \leqslant O(n^{3/4}) \,,$$

where we have also used Lemma 55 to bound $\|A\| \leqslant O(n)$ with high probability and our assumptions on τ and $\|Q - \mathrm{Id}\,\|$. By an argument similar to that in the proof of Theorem 10 (which may be found in Lemma 51), this yields $\tilde{\mathbb{E}}\, f(x) \leqslant O(n^{3/4})(\tilde{\mathbb{E}}\, \|x\|^4)^{3/4}$ as desired. ■

5. Conclusion

5.1. Open Problems

One theme in this work has been efficiently certifying upper bounds on homogeneous polynomials with random coefficients. It is an interesting question to see whether one can (perhaps with the degree $d > 4$ SoS meta-algorithm) give an algorithm certifying a bound of $n^{3/4-\delta}$ over the unit sphere on a degree 3 polynomial with standard Gaussian coefficients. Such an algorithm would likely yield improved signal-to-noise guarantees for tensor PCA, and would be of interest in its own right.

Conversely, another problem is to extend our lower bound to handle degree $d > 4$ SoS. Together, these two problems suggest (as was independently suggested to us by Boaz Barak) the problem of characterizing the SoS degree required to certify a bound of $n^{3/4-\delta}$ as above.

Another problem is to simplify the linear time algorithm we give for tensor PCA under symmetric noise. Montanari and Richard's conjecture can be interpreted to say that the random rotations and decomposition into submatrices involved in our algorithm are unnecessary, and that in fact our linear time algorithm for recovery under asymmetric noise actually succeeds in the symmetric case.

5.2. Acknowledgements

We thank Moses Charikar for bringing to our attention the work of Montanari and Richard. We would like to thank Boaz Barak, Rong Ge, and Ankur Moitra for enlightening conversations. S. B. H.

acknowledges the support of an NSF Graduate Research Fellowship under award no. 1144153. D. S. acknowledges support from the Simons Foundation, the National Science Foundation, a Alread P. Sloan Fellowship, and a Microsoft Research Faculty Fellowship, A large portion of this work was completed while the authors were long-term visitors to the Simons Institute for the Theory of Computing (Berkeley) for the program on Algorithmic Spectral Graph Theory.

References

Anima Anandkumar, Rong Ge, Daniel Hsu, and Sham M. Kakade. A tensor spectral approach to learning mixed membership community models. *CoRR*, abs/1302.2684, 2013.

Anima Anandkumar, Rong Ge, and Majid Janzamin. Analyzing tensor power method dynamics: Applications to learning overcomplete latent variable models. *CoRR*, abs/1411.1488, 2014a. URL http://arxiv.org/abs/1411.1488.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15(1): 2773–2832, 2014b. URL http://dl.acm.org/citation.cfm?id=2697055.

Animashree Anandkumar, Rong Ge, and Majid Janzamin. Guaranteed non-orthogonal tensor decomposition via alternating rank-1 updates. *CoRR*, abs/1402.5180, 2014c.

Boaz Barak and Ankur Moitra. Tensor prediction, rademacher complexity and random 3-xor. *CoRR*, abs/1501.06521, 2015. URL http://arxiv.org/abs/1501.06521.

Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *CoRR*, abs/1404.5236, 2014.

Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In Howard J. Karloff and Toniann Pitassi, editors, *STOC*, pages 307–326. ACM, 2012. ISBN 978-1-4503-1245-5. URL http://dl.acm.org/citation.cfm?id=2213977.

Boaz Barak, Guy Kindler, and David Steurer. On the optimality of semidefinite relaxations for average-case and generalized constraint satisfaction. In *ITCS*, pages 197–214, 2013.

Boaz Barak, Jonathan A. Kelner, and David Steurer. Rounding sum-of-squares relaxations. In David B. Shmoys, editor, *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 31–40. ACM, 2014a. ISBN 978-1-4503-2710-7. doi: 10.1145/2591796.2591886. URL http://doi.acm.org/10.1145/2591796.2591886.

Boaz Barak, Jonathan A. Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. *CoRR*, abs/1407.1543, 2014b. URL http://arxiv.org/abs/1407.1543.

Quentin Berthet and Philippe Rigollet. Complexity theoretic lower bounds for sparse principal component detection. In *COLT 2013 - The 26th Annual Conference on Learning Theory, June 12-14, 2013, Princeton University, NJ, USA*, pages 1046–1066, 2013. URL http://jmlr.org/proceedings/papers/v30/Berthet13.html.

Andrew C Doherty and Stephanie Wehner. Convergence of sdp hierarchies for polynomial optimization on the hypersphere. *arXiv preprint arXiv:1210.5048*, 2012.

Uriel Feige and Eran Ofek. Easily refutable subformulas of large random 3cnf formulas. *Theory of Computing*, 3(1):25–43, 2007.

Uriel Feige, Jeong Han Kim, and Eran Ofek. Witnesses for non-satisfiability of dense random 3cnf formulas. In *FOCS*, pages 497–508, 2006.

Joel Friedman, Andreas Goerdt, and Michael Krivelevich. Recognizing more unsatisfiable random k-sat instances efficiently. *SIAM J. Comput.*, 35(2):408–430, 2005. doi: 10.1137/S009753970444096X. URL http://dx.doi.org/10.1137/S009753970444096X.

Andreas Goerdt and Michael Krivelevich. Efficient recognition of random unsatisfiable k-sat instances by spectral methods. In *STACS 2001, 18th Annual Symposium on Theoretical Aspects of Computer Science, Dresden, Germany, February 15-17, 2001, Proceedings*, pages 294–304, 2001. doi: 10.1007/3-540-44693-1_26. URL http://dx.doi.org/10.1007/3-540-44693-1_26.

Andrea Montanari and Emile Richard. A statistical model for tensor pca. In *Advances in Neural Information Processing Systems*, pages 2897–2905, 2014.

Andrea Montanari, Daniel Reichman, and Ofer Zeitouni. On the limitation of spectral methods: From the gaussian hidden clique problem to rank one perturbations of gaussian tensors. *arXiv preprint arXiv:1411.6149*, 2014.

Ryan O'Donnell and Yuan Zhou. Approximability and proof complexity. In Sanjeev Khanna, editor, *SODA*, pages 1537–1556. SIAM, 2013. ISBN 978-1-61197-251-1. doi: 10.1137/1.9781611973105. URL http://dx.doi.org/10.1137/1.9781611973105.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Soc., 2012.

Ryota Tomioka and Taiji Suzuki. Spectral norm of random tensors. *arXiv preprint arXiv:1407.1870*, 2014.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z. URL http://dx.doi.org/10.1007/s10208-011-9099-z.

Roman Vershynin. Introduction to the non-asymtotic analysis of random matrices. pages 210–268, 2011.

Appendix A. Linear Time Recovery via Further Relaxation

We now attack the problem of speeding up the algorithm from the preceding section. We would like to avoid solving a large semidefinite program to optimality: our goal is to instead use much faster linear-algebraic computations—in particular, we will recover the tensor PCA signal vector by performing a single singular vector computation on a relatively small matrix. This will complete the proofs of Theorem 4 and Theorem 3, yielding the desired running time.

Our SoS algorithm in the preceding section turned on the existence of the λ-boundedness certificate $\sum_i A_i \otimes A_i$, where A_i are the slices of a random tensor $\mathbf{A}$. Let $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$ be the spiked-tensor input to tensor PCA. We could look at the matrix $\sum_i T_i \otimes T_i$ as a candidate λ-boundedness certificate for $\mathbf{T}(x)$. The spectrum of this matrix must not admit the spectral bound that $\sum_i A_i \otimes A_i$ does, because $\mathbf{T}(x)$ is not globally bounded: it has a large global maximum near the signal v. This maximum plants a single large singular value in the spectrum of $\sum_i T_i \otimes T_i$. The associated singular vector is readily decoded to recover the signal.

Before stating and analyzing this fast linear-algebraic algorithm, we situate it more firmly in the SoS framework. In the following, we discuss *spectral SoS*, a convex relaxation of Problem 2 obtained by weakening the full-power SoS relaxation. We show that the spectrum of the aforementioned $\sum_i T_i \otimes T_i$ can be viewed as approximately solving the spectral SoS relaxation. This gives the fast, certifying algorithm of Theorem 4. We also interpret the tensor unfolding algorithm given by Montanari and Richard for TPCA in the spiked tensor model as giving a more subtle approximate solution to the spectral SoS relaxation. We prove a conjecture by those authors that the algorithm successfully recovers the TPCA signal at the same signal-to-noise ratio as our other algorithms, up to a small pre-processing step in the algorithm; this proves Theorem 3 (Montanari and Richard, 2014). This last algorithm, however, succeeds for somewhat different reasons than the others, and we will show that it consequently fails to certify its own success and that it is not robust to a certain kind of semi-adversarial choice of noise.

A.1. The Spectral SoS Relaxation

A.1.1. THE SOS ALGORITHM: MATRIX VIEW

To obtain spectral SoS, the convex relaxation of Problem 2 which we will be able to (approximately) solve quickly in the random case, we first need to return to the full-strength SoS relaxation and examine it from a more linear-algebraic standpoint.

We have seen in Section 2.2 that a homogeneous $p \in \mathbb{R}[x]_{2d}$ may be represented as an $n^d \times n^d$ matrix whose entries correspond to coefficients of p. A similar fact is true for non-homogeneous p. Let $\#\text{tuples}(d) = 1 + n + n^2 + \cdots + n^{d/2}$. Let $x^{\otimes \leqslant d/2} := (x^{\otimes 0}, x, x^{\otimes 2}, \ldots, x^{\otimes d/2})$. Then $p \in \mathbb{R}[x]_{\leqslant d}$ can be represented as an $\#\text{tuples}(d) \times \#\text{tuples}(d)$ matrix; we say a matrix M of these dimensions is a matrix representation of p if $\langle x^{\leqslant \otimes d/2}, M x^{\leqslant \otimes d/2} \rangle = p(x)$. For this section, we let $\mathcal{M}_p$ denote the set of all such matrix representation of p.

A degree-d pseudo-distribution $\{x\}$ can similarly be represented as an $\mathbb{R}^{\#\text{tuples}(d) \times \#\text{tuples}(d)}$ matrix. We say that M is a matrix representation for $\{x\}$ if $M[\alpha, \beta] = \tilde{\mathbb{E}}\, x^\alpha x^\beta$ wheneverα and β are multi-indices with $|\alpha|, |\beta| \leqslant d$.

Formulated this way, if $M_{\{x\}}$ is the matrix representation of $\{x\}$ and $M_p \in \mathcal{M}_p$ for some $p \in \mathbb{R}[x]_{\leqslant 2d}$, then $\tilde{\mathbb{E}}\, p(x) = \langle M_{\{x\}}, M_p \rangle$. In this sense, pseudo-distributions and polynomials, each represented as matrices, are dual under the trace inner product on matrices.

We are interested in optimization of polynomials over the sphere, and we have been looking at pseudo-distribution $\{x\}$ satisfying $\{\|x\|^2 - 1 = 0\}$. From this matrix point of view, the polynomial $\|x\|^2 - 1$ corresponds to a vector $w \in \mathbb{R}^{\#\text{tuples}(d)}$ (in particular, the vector w so that ww^T is a matrix representation of $(\|x\|^2 - 1)^2$), and a degree-4 pseudo-distribution $\{x\}$ satisfies $\{\|x\|^2 - 1 = 0\}$ if and only if $w \in \ker M_{\{x\}}$.

A polynomial may have many matrix representations, but a pseudo-distribution has just one: a matrix representation of a pseudo-distribution must obey strong symmetry conditions in order to assign the same pseudo-expectation to every representation of the same polynomial. We will have much more to say about constructing matrices satisfying these symmetry conditions when we state and prove our lower bounds, but here we will in fact profit from relaxing these symmetry constraints.

Let $p \in \mathbb{R}[x]_{\leqslant 2d}$. In the matrix view, the SoS relaxation of the problem $\max_{\|x\|^2=1} p(x)$ is the following convex program.

$$\max_{\substack{M : w \in \ker M \\ M \succeq 0 \\ \langle M, \mathcal{M}_1 \rangle = 1}} \min_{M_p \in \mathcal{M}_p} \langle M, M_p \rangle \,. \tag{A.1}$$

It may not be immediately obvious why this program optimizes only over M which are matrix representations of pseudo-distributions. If, however, some M does not obey the requisite symmetries, then $\min_{M_p \in \mathcal{M}_p} \langle M, M_p \rangle = -\infty$, since the asymmetry may be exploited by careful choice of $M_p \in \mathcal{M}_p$. Thus, at optimality this program yields M which is the matrix representation of a pseudo-distribution $\{x\}$ satisfying $\{\|x\|^2 - 1 = 0\}$.

A.1.2. RELAXING TO THE DEGREE-4 DUAL

We now formulate spectral SoS. In our analysis of full-power SoS for tensor PCA we have primarily considered pseudo-expectations of homogeneous degree-4 polynomials; our first step in further relaxing SoS is to project from $\mathbb{R}[x]_{\leqslant 4}$ to $\mathbb{R}[x]_4$. Thus, now our matrices M, M' will be in $\mathbb{R}^{n^2 \times n^2}$ rather than $\mathbb{R}^{\#\text{tuples}(2) \times \#\text{tuples}(2)}$. The projection of the constraint on the kernel in the non-homogeneous case implies $\operatorname{Tr} M = 1$ in the homogeneous case. The projected program is

$$\max_{\substack{\operatorname{Tr} M = 1 \\ M \succeq 0}} \min_{M_p \in \mathcal{M}_p} \langle M, M' \rangle \,.$$

We modify this a bit to make explicit that the relaxation is allowed to add and subtract arbitrary matrix representations of the zero polynomial; in particular $M_{\|x\|^4} - \mathrm{Id}$ for any $M_{\|x\|^4} \in \mathcal{M}_{\|x\|^4}$. This program is the same as the one which precedes it.

$$\max_{\substack{\operatorname{Tr} M = 1 \\ M \succeq 0}} \min_{\substack{M_p \in \mathcal{M}_p \\ M_{\|x\|^4} \in \mathcal{M}_{\|x\|^4} \\ c \in \mathbb{R}}} \langle M, M_p - c \cdot M_{\|x\|^4} \rangle + c \,. \tag{A.2}$$

By weak duality, we can interchange the min and the max in (A.2) to obtain the dual program:

$$\begin{aligned} \max_{\substack{\operatorname{Tr} M = 1 \\ M \succeq 0}} \min_{\substack{M_p \in \mathcal{M}_p \\ M_{\|x\|^4} \in \mathcal{M}_{\|x\|^4} \\ c \in \mathbb{R}}} \langle M, M_p - c \cdot M_{\|x\|^4} \rangle &\leqslant \min_{\substack{M_p \in \mathcal{M}_p \\ M_{\|x\|^4} \in \mathcal{M}_{\|x\|^4} \\ c \in \mathbb{R}}} \max_{\substack{\operatorname{Tr} M = 1 \\ M \succeq 0}} \langle M, M_p - c \cdot M_{\|x\|^4} \rangle + c & \text{(A.3)} \\ &= \min_{\substack{M_p \in \mathcal{M}_p \\ M_{\|x\|^4} \in \mathcal{M}_{\|x\|^4} \\ c \in \mathbb{R}}} \max_{\|v\|=1} \langle vv^T, M_p - c \cdot M_{\|x\|^4} \rangle + c & \text{(A.4)} \end{aligned}$$

We call this dual program the spectral SoS relaxation of $\max_{\|x\|=1} p(x)$. If $p = \sum_i \langle x, A_i x \rangle$ for $\mathbf{A}$ with independent entries from $\mathcal{N}(0,1)$, the spectral SoS relaxation achieves the same bound as our analysis of the full-strength SoS relaxation: for such p, the spectral SoS relaxation is at most $O(n^{3/2} \log(n)^{1/2})$ with high probability. The reason is exactly the same as in our analysis of the full-strength SoS relaxation: the matrix $\sum_i A_i \otimes A_i$, whose spectrum we used before to bound the full-strength SoS relaxation, is still a feasible dual solution.

A.2. Recovery via the $\sum_i T_i \otimes T_i$ Spectral SoS Solution

Let $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$ be the spiked-tensor input to tensor PCA. We know from our initial characterization of SoS proofs of boundedness for degree-3 polynomials that the polynomial $\mathbf{T}'(x) := (x \otimes x)^T (\sum_i T_i \otimes T_i)(x \otimes x)$ gives SoS-certifiable upper bounds on $\mathbf{T}(x)$ on the unit sphere. We consider the spectral SoS relaxation of $\max_{\|x\|=1} \mathbf{T}'(x)$,

$$\min_{\substack{M_{\mathbf{T}(x)} \in \mathcal{M}_{\mathbf{T}(x)} \\ M_{\|x\|^4} \in \mathcal{M}_{\|x\|^4} \\ c \in \mathbb{R}}} \|M_{\mathbf{T}(x)} - c \cdot M_{\|x\|^4}\| + c \,.$$

Our goal now is to guess a good $M' \in \mathcal{M}_{\mathbf{T}(x)}$. We will take as our dual-feasible solution the top singular vector of $\sum_i T_i \otimes T_i - \mathbb{E} \sum_i A_i \otimes A_i$. This is dual feasible with $c = n$, since routine calculation gives $\langle x^{\otimes 2}, (\mathbb{E} \sum_i A_i \otimes A_i) x^{\otimes 2} \rangle = \|x\|^4$. This top singular vector, which differentiates the spectrum of $\sum_i T_i \otimes T_i$ from that of $\sum_i A_i \otimes A_i$, is exactly the manifestation of the signal v_0 which differentiates $\mathbf{T}(x)$ from $\mathbf{A}(x)$. The following algorithm and analysis captures this.

Recovery and Certification with $\sum_i T_i \otimes T_i$

Input: $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$, where $v_0 \in \mathbb{R}^n$ and $\mathbf{A}$ is a 3-tensor.
Goal: Find $v \in \mathbb{R}^n$ with $|\langle v, v_0 \rangle| \geqslant 1 - o(1)$.

Algorithm 3 (Recovery) *Compute the top (left or right) singular vector v' of $M := \sum_i T_i \otimes T_i - \mathbb{E} \sum_i A_i \otimes A_i$. Reshape v' into an $n \times n$ matrix V'. Compute the top singular vector v of V'. Output $v/\|v\|$.*

Algorithm 4 (Certification) *Run Algorithm 3 to obtain v. Let $\mathbf{S} := \mathbf{T} - v^{\otimes 3}$. Compute the top singular value λ of*

$$\sum_i S_i \otimes S_i - \mathbb{E} \sum_i A_i \otimes A_i \,.$$

If $\lambda \leqslant O(n^{3/2} \log(n)^{1/2})$, output CERTIFY. *Otherwise, output* FAIL.

The following theorem describes the behavior of Algorithm 3 and Algorithm 4 and gives a proof of Theorem 4 and Corollary 4.

Theorem 16 (Formal version of Theorem 4) *Let $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$, where $v_0 \in \mathbb{R}^n$ and $\mathbf{A}$ has independent entries from $\mathcal{N}(0,1)$. In other words, we are given an instance of Problem 1. Let $\tau \geqslant n^{3/4} \log(n)^{1/4}/\varepsilon$. Then:*

— *With high probability, Algorithm 3 returns v with $\langle v, v_0 \rangle^2 \geqslant 1 - O(\varepsilon)$.*

— *If Algorithm 4 outputs* CERTIFY *then $\mathbf{T}(x) \leqslant \tau \cdot \langle v, x \rangle^3 + O(n^{3/4} \log(n)^{1/4})$ (regardless of the distribution of $\mathbf{A}$). If $\mathbf{A}$ is distributed as above, then Algorithm 4 outputs* CERTIFY *with high probability.*

— *Both Algorithm 3 and Algorithm 4 can be implemented in time $O(n^4 \log(1/\varepsilon))$.*

The argument that Algorithm 3 recovers a good vector in the spiked tensor model comes in three parts: we show that under appropriate regularity conditions on the noise $\mathbf{A}$ that $\sum_i T_i \otimes T_i - \mathbb{E}\, A_i \otimes A_i$ has a good singular vector, then that with high probability in the spiked tensor model those regularity conditions hold, and finally that the good singular vector can be used to recover the signal.

Lemma 17 *Let $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$ be an input tensor. Suppose $\| \sum_i A_i \otimes A_i - \mathbb{E} \sum_i A_i \otimes A_i \| \leqslant \varepsilon \tau^2$ and that $\| \sum_i v_0(i) A_i \| \leqslant \varepsilon \tau$. Then the top (left or right) singular vector v' of M has $\langle v', v_0 \otimes v_0 \rangle^2 \geqslant 1 - O(\varepsilon)$.*

Lemma 18 *Let $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$. Suppose $\mathbf{A}$ has independent entries from $\mathcal{N}(0,1)$. Then with high probability we have $\| \sum_i A_i \otimes A_i - \mathbb{E} \sum_i A_i \otimes A_i \| \leqslant O(n^{3/2} \log(n)^{1/2})$ and $\| \sum_i v_0(i) A_i \| \leqslant O(\sqrt{n})$.*

Lemma 19 *Let $v_0 \in \mathbb{R}^n$ and $v' \in \mathbb{R}^{n^2}$ be unit vectors so that $\langle v', v_0 \otimes v_0 \rangle \geqslant 1 - O(\varepsilon)$. Then the top right singular vector v of the $n \times n$ matrix folding V' of v' satisfies $\langle v, v_0 \rangle \geqslant 1 - O(\varepsilon)$.*

A similar fact to Lemma 19 appears in (Montanari and Richard, 2014).

The proofs of Lemma 17 and Lemma 19 follow here. The proof of Lemma 18 uses only standard concentration of measure arguments; we defer it to Section E.

Proof [Proof of Lemma 17] We expand M as follows.

$$\begin{aligned} M &= \sum_i \tau^2 \cdot (v_0^{\otimes 3})_i \otimes (v_0^{\otimes 3})_i + \tau \cdot ((v_0^{\otimes 3})_i \otimes A_i + A_i \otimes (v_0^{\otimes 3})_i) + A_i \otimes A_i - \mathbb{E}\, A_i \otimes A_i \\ &= \tau^2 \cdot (v_0 \otimes v_0)(v_0 \otimes v_0)^T + \tau \cdot v_0 v_0^T \otimes \sum_i v_0(i) A_i + \tau \cdot \sum_i v_0(i) A_i \otimes v_0 v_0^T + A_i \otimes A_i - \mathbb{E}\, A_i \otimes A_i \,. \end{aligned}$$

By assumption, the noise term is bounded in operator norm: we have $\| \sum_i A_i \otimes A_i - \mathbb{E} \sum_i A_i \otimes A_i \| \leqslant \varepsilon \tau^2$. Similarly, by assumption the cross-term has $\| \tau \cdot v_0 v_0^T \otimes \sum_i v_0(i) A_i \| \leqslant \varepsilon \tau^2$.

$$\tau \cdot \sum_i P_{u^\perp}((v_0^{\otimes 3})_i \otimes A_i + A_i \otimes (v_0^{\otimes 3})_i) P_{u^\perp} = \tau \cdot \sum_i v_0(i) P_{u^\perp}(v_0 v_0^T \otimes A_i + A_i \otimes v_0 v_0^T) P_{u^\perp} \,.$$

All in all, by triangle inequality,

$$\left\| \tau \cdot v_0 v_0^T \otimes \sum_i v_0(i) A_i + \tau \cdot \sum_i v_0(i) A_i \otimes v_0 v_0^T + A_i \otimes A_i - \mathbb{E}\, A_i \otimes A_i \right\| \leqslant O(\varepsilon \tau^2) \,.$$

Again by triangle inequality,

$$\|M\| \geqslant (v_0 \otimes v_0)^T M (v_0 \otimes v_0) = \tau^2 - O(\varepsilon \tau^2) \,.$$

Let u, w be the top left and right singular vectors of M. We have

$$u^T M w = \tau^2 \cdot \langle u, v_0 \otimes v_0 \rangle \langle w, v_0 \otimes v_0 \rangle + O(\varepsilon \tau^2) \geqslant \tau^2 - O(\varepsilon \tau^2) \,,$$

so rearranging gives the result. ■

Proof [Proof of Lemma 19] Let v_0, v', V', v, be as in the lemma statement. We know v is the maximizer of $\max_{\|w\|, \|w'\| = 1} w^T V' w'$. By assumption,

$$v_0^T V' v_0 = \langle v', v_0 \otimes v_0 \rangle \geqslant 1 - O(\varepsilon).$$

Thus, the top singular value of V' is at least $1 - O(\varepsilon)$, and since $\|v'\|$ is a unit vector, the Frobenius norm of V' is 1 and so all the rest of the singular values are $O(\varepsilon)$. Expressing v_0 in the right singular basis of V' and examining the norm of $V'v_0$ completes the proof. ■

Proof [Proof of Theorem 16] The first claim, that Algorithm 3 returns a good vector, follows from the previous three lemmas, Lemma 17, Lemma 18, Lemma 19. The next, for Algorithm 4, follows from noting that $\sum_i S_i \otimes S_i - \mathbb{E} \sum_i A_i \otimes A_i$ is a feasible solution to the spectral SoS dual. For the claimed runtime, since we are working with matrices of size n^4, it will be enough to show that the top singular vector of M and the top singular value of $\sum_i S_i \otimes S_i - \mathbb{E} \sum_i A_i \otimes A_i$ can be recovered with $O(\operatorname{poly}\log(n))$ matrix-vector multiplies.

In the first case, we start by observing that it is enough to find a vector w which has $\langle w, v' \rangle \geqslant 1 - \varepsilon$, where v' is a top singular vector of M. Let λ_1, λ_2 be the top two singular values of M. The analysis of the algorithm already showed that $\lambda_1/\lambda_2 \geqslant \Omega(1/\varepsilon)$. Standard analysis of the matrix power method now yields that $O(\log(1/\varepsilon))$ iterations will suffice.

We finally turn to the top singular value of $\sum_i S_i \otimes S_i - \mathbb{E} \sum_i A_i \otimes A_i$. Here the matrix may not have a spectral gap, but all we need to do is ensure that the top singular value is no more than $O(n^{3/2} \log(n)^{1/2})$. We may assume that some singular value is greater than $O(n^{3/2} \log(n)^{1/2})$. If all of them are, then a single matrix-vector multiply initialized with a random vector will discover this. Otherwise, there is a constant spectral gap, so a standard analysis of matrix power method says that within $O(\log n)$ iterations a singular value greater than $O(n^{3/2} \log(n)^{1/2})$ will be found, if it exists. ■

A.3. Nearly-Linear-Time Recovery via Tensor Unfolding and Spectral SoS

On input $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$, where as usual $v_0 \in \mathbb{R}^n$ and $\mathbf{A}$ has independent entries from $\mathcal{N}(0,1)$, Montanari and Richard's Tensor Unfolding algorithm computes the top singular vector u of the squarest-possible flattening of T into a matrix. It then extracts v with $\langle v, v_0 \rangle^2 \geqslant 1 - o(1)$ from u with a second singular vector computation.

Recovery with TT^T, a.k.a. Tensor Unfolding

Input: $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$, where $v_0 \in \mathbb{R}^n$ and $\mathbf{A}$ is a 3-tensor.
Goal: Find $v \in \mathbb{R}^n$ with $|\langle v, v_0 \rangle| \geqslant 1 - o(1)$.

Algorithm 5 (Recovery) *Compute the top eigenvector v of $M := T^T T$. Output v.*

We show that this algorithm successfully recovers a vector v with $\langle v, v_0 \rangle^2 \geqslant 1 - O(\varepsilon)$ when $\tau \geqslant n^{3/4}/\varepsilon$. Montanari and Richard conjectured this but were only able to show it when $\tau \geqslant n$. We also show how to implement the algorithm in time $\tilde{O}(n^3)$, that is to say, in time nearly-linear in the input size.

Despite its a priori simplicity, the analysis of Algorithm 5 is more subtle than for any of our other algorithms. This would not be true for even-order tensors, for which the square matrix unfolding tensor has one singular value asymptotically larger than all the rest, and indeed the corresponding singular vector is well-correlated with v_0. However, in the case of odd-order tensors the unfolding has no spectral gap. Instead, the signal v_0 has some second-order effect on the spectrum of the matrix unfolding, which is enough to recover it.

We first situate this algorithm in the SoS framework. In the previous section we examined the feasible solution $\sum_i T_i \otimes T_i - \mathbb{E} \sum_i A_i \otimes A_i$ to the spectral SoS relaxation of $\max_{\|x\|=1} \mathbf{T}(x)$. The tensor unfolding algorithm works by examining the top singular vector of the flattening T of $\mathbf{T}$, which is the top eigenvector of the $n \times n$ matrix $M = T^T T$, which in turn has the same spectrum as the $n^2 \times n^2$ matrix TT^T. The latter is also a feasible dual solution to the spectral SoS relaxation of $\max_{\|x\|=1} \mathbf{T}(x)$. However, the bound it provides on $\max_{\|x\|=1} \mathbf{T}(x)$ is much worse than that given by $\sum_i T_i \otimes T_i$. The latter, as we saw in the preceding section, gives the bound $O(n^{3/4} \log(n)^{1/4})$. The former, by contrast, gives only $O(n)$, which is the operator norm of a random $n^2 \times n$ matrix (see Lemma 55). This n versus $n^{3/4}$ is the same as the gap between Montanari and Richard's conjectured bound and what they were able to prove.

Theorem 20 *For an instance of Problem 1 with $\tau \geqslant n^{3/4}/\varepsilon$, with high probability Algorithm 5 recovers a vector v with $\langle v, v_0 \rangle^2 \geqslant 1 - O(\varepsilon)$. Furthermore, Algorithm 5 can be implemented in time $\tilde{O}(n^3)$.*

Lemma 21 *Let $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$ where $v_0 \in \mathbb{R}^n$ is a unit vector, so an instance of Problem 1. Suppose $\mathbf{A}$ satisfies $A^T A = C \cdot \mathrm{Id}_{n \times n} + E$ for some $C \geqslant 0$ and E with $\|E\| \leqslant \varepsilon \tau^2$ and that $\|A^T(v_0 \otimes v_0)\| \leqslant \varepsilon \tau$. Let u be the top left singular vector of the matrix T. Then $\langle v_0, u \rangle^2 \geqslant 1 - O(\varepsilon)$.*

Proof The vector u is the top eigenvector of the $n \times n$ matrix TT^T, which is also the top eigenvector of $M := TT^T - C \cdot \mathrm{Id}$. We expand:

$$\begin{aligned} u^T M u &= u^T \left[\tau^2 \cdot v_0 v_0^T + \tau \cdot v_0 (v_0 \otimes v_0)^T A + \tau \cdot A^T (v_0 \otimes v_0) v_0^T + E \right] u \\ &= \tau^2 \cdot \langle u, v_0 \rangle^2 + u^T \left[\tau \cdot v_0 (v_0 \otimes v_0)^T A + \tau \cdot A^T (v_0 \otimes v_0) v_0^T + E \right] u \\ &\leqslant \tau^2 \langle u, v_0 \rangle^2 + O(\varepsilon \tau^2)\,. \end{aligned}$$

Again by triangle inequality, $u^T M u \geqslant v_0^T M v = \tau^2 - O(\varepsilon \tau^2)$. So rearranging we get $\langle u, v_0 \rangle^2 \geqslant 1 - O(\varepsilon)$ as desired. ■

The following lemma is a consequence of standard matrix concentration inequalities; we defer its proof to Section E, Lemma 61.

Lemma 22 *Let $\mathbf{A}$ have independent entries from $\mathcal{N}(0, 1)$. Let $v_0 \in \mathbb{R}^n$ be a unit vector. With high probability, the matrix A satisfies $A^T A = n^2 \cdot \mathrm{Id} + E$ for some E with $\|E\| \leqslant O(n^{3/2})$ and $\|A^T(v_0 \otimes v_0)\| \leqslant O(\sqrt{n \log n})$.*

The final component of a proof of Theorem 20 is to show how it can be implemented in time $\tilde{O}(n^3)$. Since M factors as $T^T T$, a matrix-vector multiply by M can be implemented in time $O(n^3)$. Unfortunately, M does not have an adequate eigenvalue gap to make matrix power method efficient. As we know from Lemma 22, suppressing εs and constants, M has eigenvalues in the range $n^2 \pm n^{3/2}$. Thus, the eigenvalue gap of M is at most $g = O(1 + 1/\sqrt{n})$. For any number k of matrix-vector multiplies with $k \leqslant n^{1/2 - \delta}$, the eigenvalue gap will become at most $(1 + 1/\sqrt{n})^{n^{1/2-\delta}}$, which is subconstant. To get around this problem, we employ a standard trick to improve spectral gaps of matrices close to $C \cdot \mathrm{Id}$: remove $C \cdot \mathrm{Id}$.

Lemma 23 *Under the assumptions of Theorem 20, Algorithm 5 can be implemented in time* (n^3) *(which is linear in the input size,* n^3*).*

Proof Note that the top eigenvector of M is the same as that of $M - n^2 \cdot \mathrm{Id}$. The latter matrix, by the same analysis as in Lemma 21, is given by

$$M - n^2 \cdot \mathrm{Id} = \tau^2 \cdot v_0 v_0^T + M'$$

where $\|M'\| = O(\varepsilon\tau^2)$. Note also that a matrix-vector multiply by $M - n^2 \cdot \mathrm{Id}$ can still be done in time $O(n^3)$. Thus, $M - n^2 \cdot \mathrm{Id}$ has eigenvalue gap $\Omega(1/\varepsilon)$, which is enough so that the whole algorithm runs in time $\tilde{O}(n^3)$. ■

Proof [Proof of Theorem 20] Immediate from Lemma 21, Lemma 22, and Lemma 23. ■

A.4. Fast Recovery in the Semi-Random Model

There is a qualitative difference between the aggregate matrix statistics needed by our certifying algorithms (Algorithm 1, Algorithm 2, Algorithm 3, Algorithm 4) and those needed by rounding the tensor unfolding solution spectral SoS Algorithm 5. In a precise sense, the needs of the latter are greater. The former algorithms rely only on first-order behavior of the spectra of a tensor unfolding, while the latter relies on second-order spectral behavior. Since it uses second-order properties of the randomness, Algorithm 5 fails in the semi-random model.

Theorem 24 *Let* $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$*, where* $v_0 \in \mathbb{R}^n$ *is a unit vector and* $\mathbf{A}$ *has independent entries from* $\mathcal{N}(0,1)$*. There is* $\tau = \Omega(n^{7/8})$ *so that with high probability there is an adversarial choice of* Q *with* $\|Q - \mathrm{Id}\| \leqslant O(n^{-1/4})$ *so that the matrix* $(TQ)^T TQ = n^2 \cdot \mathrm{Id}$*. In particular, for such* τ*, Algorithm 5 cannot recover the signal* v_0*.*

Proof Let M be the $n \times n$ matrix $M := T^T T$. Let $Q = n \cdot M^{-1/2}$. It is clear that $(TQ)^T TQ = n^2\,\mathrm{Id}$. It suffices to show that $\|Q - \mathrm{Id}\| \leqslant n^{1/4}$ with high probability. We expand the matrix M as

$$M = \tau^2 \cdot v_0 v_0^T + \tau \cdot v_0 (v_0 \otimes v_0)^T A + \tau \cdot A^T (v_0 \otimes v_0) v_0^T + A^T A\,.$$

By Lemma 22, $A^T A = n^2 \cdot \mathrm{Id} + E$ for some E with $\|E\| \leqslant O(n^{3/2})$ and $\|A^T(v_0 \otimes v_0)\| \leqslant O(\sqrt{n \log n})$, both with high probability. Thus, the eigenvalues of M all lie in the range $n^2 \pm n^{1+3/4}$. The eigenvalues of Q in turn lie in the range

$$\frac{n}{(n^2 \pm O(n^{1+3/4}))^{1/2}} = \frac{1}{(1 \pm O(n^{-1/4}))^{1/2}} = \frac{1}{1 \pm O(n^{1/4})}\,.$$

Finally, the eigenvalues of $Q - \mathrm{Id}$ lie in the range $\frac{1}{1 \pm O(n^{1/4})} - 1 = \pm O(n^{-1/4})$, so we are done. ■

The argument that that Algorithm 3 and Algorithm 4 still succeed in the semi-random model is routine; for completeness we discuss here the necessary changes to the proof of Theorem 16. The non-probabilistic certification claims made in Theorem 16 are independent of the input model, so we show that Algorithm 3 still finds the signal with high probability and that Algorithm 4 still fails only with only a small probability.

Theorem 25 *In the semi-random model, $\varepsilon \geqslant n^{-1/4}$ and $\tau \geqslant n^{3/4}\log(n)^{1/4}/\varepsilon$, with high probability, Algorithm 3 returns v with $\langle v, v_0\rangle^2 \geqslant 1 - O(\varepsilon)$ and Algorithm 4 outputs* CERTIFY.

Proof We discuss the necessary modifications to the proof of Theorem 16. Since $\varepsilon \geqslant n^{-1/4}$, we have that $\|(Q - \mathrm{Id})v_0\| \leqslant O(\varepsilon)$. It suffices then to show that the probabilistic bounds in Lemma 18 hold with A replaced by AQ. Note that this means each A_i becomes A_iQ. By assumption, $\|Q \otimes Q - \mathrm{Id} \otimes \mathrm{Id}\,\| \leqslant O(\varepsilon)$, so the probabilistic bound on $\|\sum_i A_i \otimes A_i = \mathbb{E}\sum_i A_i \otimes A_i\|$ carries over to the semi-random setting. A similar argument holds for $\sum_i v_0(i)A_iQ$, which is enough to complete the proof. ■

A.5. Fast Recovery with Symmetric Noise

We suppose now that $\mathbf{A}$ is a symmetric Gaussian noise tensor; that is, that $\mathbf{A}$ is the average of $\mathbf{A}_0^\pi$ over all $\pi \in \mathcal{S}_k$, for some order-3 tensor $\mathbf{A}_0$ with iid standard Gaussian entries.

It was conjectured by Montanari and Richard (Montanari and Richard, 2014) that the tensor unfolding technique can recover the signal vector v_0 in the single-spike model $\mathbf{T} = \tau v_0^{\otimes 3} + \mathbf{A}$ with signal-to-noise ratio $\tau \geqslant \tilde{\Omega}(n^{3/4})$ under both asymmetric and symmetric noise.

Our previous techniques fail in this symmetric noise scenario due to lack of independence between the entries of the noise tensor. However, we sidestep that issue here by restricting our attention to an asymmetric block of the input tensor.

The resulting algorithm is not precisely identical to the tensor unfolding algorithm investigated by Montanari and Richard, but is based on tensor unfolding with only superficial modifications.

Fast Recovery under Symmetric Noise

Input: $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$, where $v_0 \in \mathbb{R}^n$ and $\mathbf{A}$ is a 3-tensor.
Goal: Find $v \in \mathbb{R}^n$ with $|\langle v, v_0\rangle| \geqslant 1 - o(1)$.

Algorithm 6 (Recovery) *Take X, Y, Z a random partition of $[n]$, and R a random rotation of $\mathbb{R}^n$. Let P_X, P_Y, and P_Z be the diagonal projectors onto the coordinates indicated by X, Y, and Z. Let* $\mathbf{U} := R^{\otimes 3}\mathbf{T}$*, so that we have the matrix unfolding $U := (R \otimes R)TR^T$ Using the matrix power method, compute the top singular vectors v_X, v_Y, and v_Z respectively of the matrices*

$$\begin{aligned} M_X &:= P_X U^T (P_Y \otimes P_Z) U P_X - n^2/9 \cdot \mathrm{Id} \\ M_Y &:= P_Y U^T (P_Z \otimes P_X) U P_Y - n^2/9 \cdot \mathrm{Id} \\ M_Z &:= P_Z U^T (P_X \otimes P_Y) U P_Z - n^2/9 \cdot \mathrm{Id}\ . \end{aligned}$$

Output the normalization of $R^{-1}(v_X + v_Y + v_Z)$.

Remark 26 (Implementation of Algorithm 6 in nearly-linear time.) *It is possible to implement each iteration of the matrix power method in Algorithm 6 in linear time. We focus on multiplying a vector by M_X in linear time; the other cases follow similarly.*

We can expand $M_X = P_X RT^T(R \otimes R)^T(P_Y \otimes P_Z)(R \otimes R)TR^T P_X - n^2/9 \cdot$ Id*. It is simple enough to multiply an n-dimensional vector by P_X, R, R^T, T, and* Id *in linear time. Furthermore multiplying* *an n^2-dimensional vector by T^T is also a simple linear time operation. The trickier part lies in multiplying an n^2-dimensional vector, say v, by the n^2-by-n^2 matrix $(R \otimes R)^T (P_Y \otimes P_Z)(R \otimes R)$.*

To accomplish this, we simply reflatten our tensors. Let V be the n-by-n matrix flattening of v. Then we compute the matrix $R^T P_Y R \cdot V \cdot R^T {P_Z}^T R$, and return its flattening back into an n^2-dimensional vector, and this will be equal to $(R \otimes R)^T (P_Y \otimes P_Z)(R \otimes R)\, v$.

Lemma 27 *Given a unit vector $u \in \mathbb{R}^n$, a random rotation R over $\mathbb{R}^n$, and a projection P to an m-dimensional subspace, with high probability*

$$\left| \|PRu\|^2 - m/n \right| \leqslant O(\sqrt{m/n^2} \log m)\,.$$

Proof Let γ be a random variable distributed as the norm of a vector in $\mathbb{R}^n$ with entries independently drawn from $\mathcal{N}(0, 1/n)$. Then because Gaussian vectors are rotationally invariant and Ru is a random unit vector, the coordinates of γRu are independent and Gaussian in any orthogonal basis.

So $\gamma^2 \|PRu\|^2$ is the sum of the squares of m independent variables drawn from $\mathcal{N}(0, 1/n)$. By a Bernstein inequality, $\left|\gamma^2 \|PRu\|^2 - m/n\right| \leqslant O(\sqrt{m/n^2} \log m)$ with high probability. Also by a Bernstein inequality, $\gamma^2 - 1 < O(\sqrt{1/n} \log n)$ with high probability. ■

Theorem 28 *For $\tau \geqslant n^{3/4}/\varepsilon$, with high probability, Algorithm 6 recovers a vector v with $\langle v, v_0 \rangle \geqslant 1 - O(\varepsilon)$ when $\mathbf{A}$ is a symmetric Gaussian noise tensor (as in Problem 3) and $\varepsilon \geqslant \log(n)/\sqrt{n}$.*

Furthermore the matrix power iteration steps in Algorithm 6 each converge within $\tilde{O}(-\log(\varepsilon))$ steps, so that the algorithm overall runs in almost linear time $\tilde{O}(n^3 \log(1/\varepsilon))$.

Proof Name the projections $U_X := (P_Y \otimes P_Z) U P_X$, $U_Y := (P_Z \otimes P_X) U P_Y$, and $U_Z := (P_X \otimes P_Y) U P_Z$.

First off, $\mathbf{U} = \tau (R v_0)^{\otimes 3} + \mathbf{A}'$ where $\mathbf{A}'$ is a symmetric Gaussian tensor (distributed identically to $\mathbf{A}$). This follows by noting that multiplication by $R^{\otimes 3}$ commutes with permutation of indices, so that $(R^{\otimes 3}\mathbf{A})^\pi = R^{\otimes 3}\mathbf{A}^\pi$, and from the rotational symmetry of Gaussian tensors.

Thus $U_X = \tau (P_Y \otimes P_Z)(R \otimes R)(v_0 \otimes v_0)(P_X R v_0)^T + (P_Y \otimes P_Z) A' P_X$, and

$$\begin{aligned} M_X + n^2/9 \cdot \mathrm{Id} &= U_X^T U_X \\ &= \tau^2 \|P_Y R v_0\|^2 \|P_Z R v_0\|^2 (P_X R v_0)(P_X R v_0)^T && \text{(A.5)} \\ &\quad + \tau (P_X R v_0)(v_0 \otimes v_0)^T (R \otimes R)^T (P_Y \otimes P_Z) A' P_X && \text{(A.6)} \\ &\quad + \tau P_X {A'}^T (P_Y \otimes P_Z)(R \otimes R)(v_0 \otimes v_0)(P_X R v_0)^T && \text{(A.7)} \\ &\quad + P_X {A'}^T (P_Y \otimes P_Z) A' P_X\,. && \text{(A.8)} \end{aligned}$$

Let S refer to Expression A.5. By Lemma 27, $\left| \|P R v_0\|^2 - \frac{1}{3} \right| < O(\sqrt{1/n} \log n)$ with high probability for $P \in \{P_X, P_Y, P_Z\}$. Hence $S = (\frac{1}{9} \pm O(\sqrt{1/n} \log n))\tau^2 (P_X R v_0)(P_X R v_0)^T$ and $\|S\| = (\frac{1}{27} \pm O(\sqrt{1/n} \log n))\tau^2$.

Let C refer to Expression A.6 so that Expression A.7 is C^T. Let also $A'' = (P_Y \otimes P_Z) A' P_X$. Note that, once the identically-zero rows and columns of A'' are removed, A'' is a matrix of iid standard Gaussian entries. Finally, let $v'' = P_Y R v_0 \otimes P_Z R v_0$. By some substitution and by noting that $\|P_X R\| \leqslant 1$, we have that $\|C\| \leqslant \tau \|v_0 {v''}^T A''\|$. Hence by Lemma 61, $\|C\| \leqslant O(\varepsilon \tau^2)$.

Let N refer to Expression A.8. Note that $N = {A''}^T A''$. Therefore by Lemma 22, $\|N - n^2/9 \cdot \mathrm{Id}\| \leqslant O(n^{3/2})$.

Thus $M_X = S + C + (N - n^2/9 \cdot \mathrm{Id})$, so that $\|M_X - S\| \leqslant O(\varepsilon\tau^2)$. Since S is rank-one and has $\|S\| \geqslant \Omega(\tau^2)$, we conclude that matrix power iteration converges in $\tilde{O}(-\log\varepsilon)$ steps.

The recovered eigenvector v_X satisfies $\langle v_X, M_X v_X\rangle \geqslant \Omega(\tau^2)$ and $\langle v_X, (M_X - S)v_X\rangle \leqslant O(\varepsilon\tau^2)$ and therefore $\langle v_X, S v_X\rangle = (\frac{1}{27} \pm O(\varepsilon + \sqrt{1/n}\log n))\tau^2$. Substituting in the expression for S, we conclude that $\langle P_X R v_0, v_X\rangle = (\frac{1}{\sqrt{3}} \pm O(\varepsilon + \sqrt{1/n}\log n))$.

The analyses for v_Y and v_Z follow in the same way. Hence

$$\begin{aligned}\langle v_X + v_Y + v_Z, Rv_0\rangle &= \langle v_X, P_X R v_0\rangle + \langle v_Y, P_Y R v_0\rangle + \langle v_Z, P_Z R v_0\rangle \\ &\geqslant \sqrt{3} - O(\varepsilon + \sqrt{1/n}\log n)\,.\end{aligned}$$

At the same time, since v_X, v_Y, and v_Z are each orthogonal to each other, $\|v_X + v_Y + v_Z\| = \sqrt{3}$. Hence with the output vector being $v := R^{-1}(v_X + v_Y + v_Z)/\|v_X + v_Y + v_Z\|$, we have

$$\langle v, v_0\rangle = \langle Rv, Rv_0\rangle = \tfrac{1}{\sqrt{3}}\langle v_X + v_Y + v_Z, Rv_0\rangle \geqslant 1 - O(\varepsilon + \sqrt{1/n}\log n)\,.$$

■

A.6. Numerical Simulations

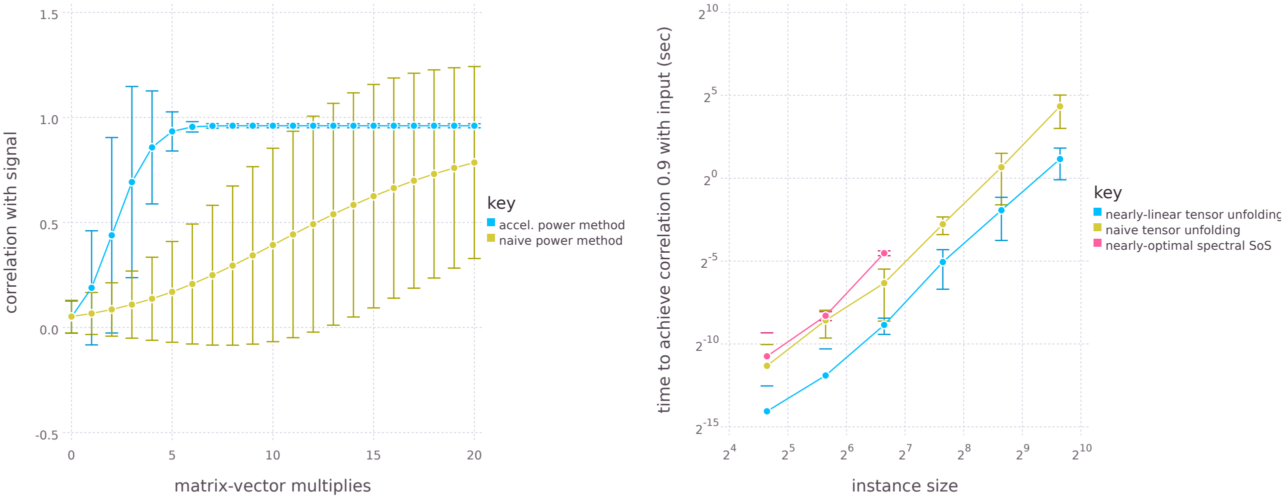

Figure 1: Numerical simulation of Algorithm 3 (“Nearly-optimal spectral SoS” implemented with matrix power method), and two implementations of Algorithm 5 (“Accelerated power method”/“Nearly-linear tensor unfolding” and “Naive power method”/“Naive tensor unfolding”. Simulations were run in Julia on a Dell Optiplex 7010 running Ubuntu 12.04 with two Intel Core i7 3770 processors at 3.40 ghz and 16GB of RAM. Plots created with Gadfly. Error bars denote 95% confidence intervals. Matrix-vector multiply experiments were conducted with $n = 200$. Reported matrix-vector multiply counts are the average of 50 independent trials. Reported times are in cpu-seconds and are the average of 10 independent trials. Note that both axes in the right-hand plot are log scaled.

We report now the results of some simple numerical simulations of the algorithms from this section. In particular, we show that the asymptotic running time differences among Algorithm 3, Algorithm 5 implemented naïvely, and the linear-time implementation of Algorithm 5 are apparent at reasonable values of n, e.g. $n = 200$.

Specifics of our experiments are given in Figure 1. We find pronounced differences between all three algorithms. The naïve implementation of Algorithm 5 is markedly slower than the linear implementation, as measured either by number of matrix-vector multiplies or processor time. Algorithm 3 suffers greatly from the need to construct an $n^2 \times n^2$ matrix; although we do not count the time to construct this matrix against its reported running time, the memory requirements are so punishing that we were unable to collect data beyond $n = 100$ for this algorithm.

Appendix B. Lower Bounds

We will now prove lower bounds on the performance of degree-4 SoS on random instances of the degree-4 and degree-3 homogeneous polynomial maximization problems. As an application, we show that our analysis of degree-4 for Tensor PCA is tight up to a small logarithmic factor in the signal-to-noise ratio.

Theorem 29 (Part one of formal version of Theorem 2) *There is $\tau = \Omega(n)$ and a function $\eta : \mathbf{A} \mapsto \{x\}$ mapping 4-tensors to degree-4 pseudo-distributions satisfying $\{\|x\|^2 = 1\}$ so that for every unit vector v_0, if $\mathbf{A}$ has unit Gaussian entries, then, with high probability over random choice of $\mathbf{A}$, the pseudo-expectation $\tilde{\mathbb{E}}_{x\sim\eta(\mathbf{A})}\, \tau \cdot \langle v_0, x\rangle^4 + \mathbf{A}(x)$ is maximal up to constant factors among $\tilde{\mathbb{E}}\, \tau \cdot \langle v_0, y\rangle^4 + \mathbf{A}(y)$ over all degree-4 pseudo-distributions $\{y\}$ satisfying $\{\|y\|^2 = 1\}$.*

Theorem 30 (Part two of formal version of Theorem 2) *There is $\tau = \Omega(n^{3/4}/(\log n)^{1/4})$ and a function $\eta : \mathbf{A} \mapsto \{x\}$ mapping 3-tensors to degree-4 pseudo-distributions satisfying $\{\|x\|^2 = 1\}$ so that for every unit vector v_0, if $\mathbf{A}$ has unit Gaussian entries, then, with high probability over random choice of $\mathbf{A}$, the pseudo-expectation $\tilde{\mathbb{E}}_{x\sim\eta(\mathbf{A})}\, \tau \cdot \langle v_0, x\rangle^3 + \mathbf{A}(x)$ is maximal up to logarithmic factors among $\tilde{\mathbb{E}}\, \tau \cdot \langle v_0, y\rangle^3 + \mathbf{A}(y)$ over all degree-4 pseudo-distributions $\{y\}$ satisfying $\{\|y\|^2 = 1\}$.*

The existence of the maps η depending only on the random part $\mathbf{A}$ of the tensor PCA input $v_0^{\otimes 3} + \mathbf{A}$ formalizes the claim from Theorem 2 that no algorithm can reliably recover v_0 from the pseudo-distribution $\eta(\mathbf{A})$.

The rest of this section is devoted to proving these theorems, which we eventually accomplish in Section B.2.

B.0.1. Discussion and Outline of Proof

Given a random 3-tensor $\mathbf{A}$, we will take the degree-3 pseudo-moments of our $\eta(\mathbf{A})$ to be $\varepsilon\mathbf{A}$, for some small ε, so that $\tilde{\mathbb{E}}_{x\sim\eta(\mathbf{A})}\, \mathbf{A}(x)$ is large. The main question is how to give degree-4 pseudo-moments to go with this. We will construct these from AA^T and its permutations as a 4-tensor under the action of $\mathcal{S}_4$.

We have already seen that a spectral *upper bound* on one of these permutations, $\sum_i A_i \otimes A_i$, provides a performance guarantee for degree-4 SoS optimization of degree-3 polynomials. It is not a coincidence that this SoS lower bound depends on the negative eigenvalues of the permutations of AA^T. Running the argument for the upper bound in reverse, a pseudo-distribution $\{x\}$ satisfying $\{\|x\|_2^2 = 1\}$ and with $\tilde{\mathbb{E}}\, \mathbf{A}(x)$ large must (by pseudo-Cauchy-Schwarz) also have $\tilde{\mathbb{E}}\langle x^{\otimes 2}, (\sum_i A_i \otimes A_i)\, x^{\otimes 2}\rangle$ large. The permutations of AA^T are all matrix representations of that same polynomial, $\langle x^{\otimes 2}, (\sum_i A_i \otimes A_i)\, x^{\otimes 2}\rangle$. Hence $\tilde{\mathbb{E}}\, \mathbf{A}(x)$ will be large only if the matrix representation of the pseudo-distribution $\{x\}$ is well correlated with the permutations of AA^T. Since this matrix representation will also need to be positive-semidefinite, control on the spectra of permutations of AA^T is therefore the key to our approach.

The general outline of the proof will be as follows:

1. Construct a pseudo-distribution that is well correlated with the permutations of AA^T and gives a large value to $\tilde{\mathbb{E}}\,\mathbf{A}(x)$, but which is not on the unit sphere.

2. Use a procedure modifying the first and second degree moments of the pseudo-distribution to force it onto a sphere, at the cost of violating the condition that $\tilde{\mathbb{E}}\,p(X)^2 \geqslant 0$ for all $p \in \mathbb{R}[x]_{\leqslant 2}$, then rescale so it lives on the unit sphere. Thus, we end up with an object that is no longer a valid pseudo-distribution but a more general linear functional $\mathcal{L}$ on polynomials.

3. Quantitatively bound the failure of $\mathcal{L}$ to be a pseudo-distribution, and repair it by statistically mixing the almost-pseudo-distribution with a small amount of the uniform distribution over the sphere. Show that $\tilde{\mathbb{E}}\,\mathbf{A}(x)$ is still large for this new pseudo-distribution over the unit sphere.

But before we can state a formal version of our theorem, we will need a few facts about polynomials, pseudo-distributions, matrices, vectors, and how they are related by symmetries under actions of permutation groups.

B.1. Polynomials, Vectors, Matrices, and Symmetries, Redux

Here we further develop the matrix view of SoS presented in Section A.1.1.

We will need to use general linear functionals $\mathcal{L} : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$ on polynomials as an intermediate step between matrices and pseudo-distributions. Like pseudo-distributions, each such linear-functional $\mathcal{L}$ has a unique matrix representation $M_{\mathcal{L}}$ satisfying certain maximal symmetry constraints. The matrix $M_{\mathcal{L}}$ is positive-semidefinite if and only if $\mathcal{L}\,p(x)^2 \geqslant 0$ for every p. If $\mathcal{L}$ satisfies this and $\mathcal{L}\,1 = 1$, then $\mathcal{L}$ is a pseudo-expectation, and $M_{\mathcal{L}}$ is the matrix representation of the corresponding pseudo-distribution.

B.1.1. Matrices for Linear Functionals and Maximal Symmetry

Let $\mathcal{L} : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$. $\mathcal{L}$ can be represented as an $n^{\#\text{tuples}(d)} \times n^{\#\text{tuples}(d)}$ matrix indexed by all d'-tuples over $[n]$ with $d' \leqslant d/2$. For tuples α, β, this matrix $M_{\mathcal{L}}$ is given by

$$M_{\mathcal{L}}[\alpha,\beta] \overset{\text{def}}{=} \mathcal{L}\,x^{\alpha}x^{\beta}\,.$$

For a linear functional $\mathcal{L} : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$, a polynomial $p(x) \in \mathbb{R}[x]_{\leqslant d}$, and a matrix representation M_p for p we thus have $\langle M_{\mathcal{L}}, M_p \rangle = \mathcal{L}\,p(x)$.

A polynomial in $\mathbb{R}[x]_{\leqslant d}$ may have many matrix representations, while for us, a linear functional $\mathcal{L}$ has just one: the matrix $M_{\mathcal{L}}$. This is because in our definition we have required that $M_{\mathcal{L}}$ obey the constraints

$$M_{\mathcal{L}}[\alpha,\beta] = M_{\mathcal{L}}[\alpha',\beta'] \qquad \text{when} \qquad x^{\alpha}x^{\beta} = x^{\alpha'}x^{\beta'}\,.$$

in order that they assign consistent values to each representation of the same polynomial. We call such matrices *maximally symmetric* (following Doherty and Wehner (Doherty and Wehner, 2012)).

We have particular interest in the maximally-symmetric version of the identity matrix. The degree-d symmetrized identity matrix $\mathrm{Id}^{\mathrm{sym}}$ is the unique maximally symmetric matrix so that

$$\langle x^{\otimes d/2}, \mathrm{Id}^{\mathrm{sym}}\, x^{\otimes d/2} \rangle = \|x\|_2^d. \tag{B.1}$$

The degree d will always be clear from context.

In addition to being a matrix representation of the polynomial $\|x\|_2^d$, the maximally symmetric matrix $\mathrm{Id}^{\mathrm{sym}}$ also serves a dual purpose as a linear functional. We will often be concerned with the expectation operator $\mathbb{E}^\mu$ for the uniform distribution over the n-sphere, and indeed for every polynomial $p(x)$ with matrix representation M_p,

$$\mathbb{E}^\mu p(x) = \frac{1}{n^2 + 2n} \langle \mathrm{Id}^{\mathrm{sym}}, M_p \rangle ,$$

and so $\mathrm{Id}^{\mathrm{sym}} /(n^2 + 2n)$ is the unique matrix representation of $\mathbb{E}^\mu$.

B.1.2. The Monomial-Indexed (i.e. Symmetric) Subspace

We will also require vector representations of polynomials. We note that $\mathbb{R}[x]_{\leqslant d/2}$ has a canonical embedding into $\mathbb{R}^{\#\mathrm{tuples}(d)}$ as the subspace given by the following family of constraints, expressed in the basis of d'-tuples for $d' \leqslant d/2$:

$$\mathbb{R}[x]_{\leqslant d/2} \simeq \{ p \in \mathbb{R}^{\#\mathrm{tuples}(d)} \text{ such that } p_\alpha = p_{\alpha'} \text{ if } \alpha' \text{ is a permutation of } \alpha \} .$$

We let Π be the projector to this subspace. For any maximally-symmetric M we have $\Pi M \Pi = M$, but the reverse implication is not true (for readers familiar with quantum information: any M which has $M = \Pi M \Pi$ is *Bose-symmetric*, but may not be *PPT-symmetric*; maximally symmetric matrices are both. See (Doherty and Wehner, 2012) for further discussion.)

If we restrict attention to the embedding this induces of $\mathbb{R}[x]_{d/2}$ (i.e. the homogeneous degree-$d/2$ polynomials) into $\mathbb{R}^{n^{d/2}}$, the resulting subspace is sometimes called the *symmetric subspace* and in other works is denoted by $\vee^{d/2}\mathbb{R}^n$. We sometimes abuse notation and let Π be the projector from $\mathbb{R}^{n^{d/2}}$ to the canonical embedding of $\mathbb{R}[x]_{d/2}$.

B.1.3. Maximally-Symmetric Matrices from Tensors

The group $\mathcal{S}_d$ acts on the set of d-tensors (canonically flattened to matrices $\mathbb{R}^{n^{\lfloor d/2 \rfloor} \times n^{\lceil d/2 \rceil}}$) by permutation of indices. To any such flattened $M \in \mathbb{R}^{n^{\lfloor d/2 \rfloor} \times n^{\lceil d/2 \rceil}}$, we associate a family of maximally-symmetric matrices $\mathrm{Sym}\, M$ given by

$$\mathrm{Sym}\, M \stackrel{\mathrm{def}}{=} \left\{ t \sum_{\pi \in \mathcal{S}_d} \pi \cdot M \text{ for all } t \geqslant 0 \right\} .$$

That is, $\mathrm{Sym}\, M$ represents all scaled averages of M over different possible flattenings of its corresponding d-tensor. The following conditions on a matrix M are thus equivalent: (1) $M \in \mathrm{Sym}\, M$, (2) M is maximally symmetric, (3) a tensor that flattens to M is invariant under the index-permutation action of $\mathcal{S}_d$, and (4) M may be considered as a linear functional on the space of homogeneous polynomials $\mathbb{R}[x]_d$. When we construct maximally-symmetric matrices from un-symmetric ones, the choice of t is somewhat subtle and will be important in not being too wasteful in intermediate steps of our construction.

There is a more complex group action characterizing maximally-symmetric matrices in $\mathbb{R}^{\#\mathrm{tuples}(d) \times \#\mathrm{tuples}(d)}$, which projects to the action of $\mathcal{S}_{d'}$ under the projection of $\mathbb{R}^{\#\mathrm{tuples}(d) \times \#\mathrm{tuples}(d)}$ to $\mathbb{R}^{n^{d'/2} \times n^{d'/2}}$. We will never have to work explicitly with this full symmetry group; instead we will be able to construct linear functionals on $\mathbb{R}[x]_{\leqslant d}$ (i.e. maximally symmetric matrices in $\mathbb{R}^{\#\mathrm{tuples}(d) \times \#\mathrm{tuples}(d)}$) by symmetrizing each degree (i.e. each $d' \leqslant d$) more or less separately.

B.2. Formal Statement of the Lower Bound

We will warm up with the degree-4 lower bound, which is conceptually somewhat simpler.

Theorem 31 (Degree-4 Lower Bound, General Version) *Let* $\mathbf{A}$ *be a* 4*-tensor and let* $\lambda > 0$ *be a function of n. Suppose the following conditions hold:*

- $\mathbf{A}$ is significantly correlated with $\sum_{\pi \in \mathcal{S}_4} \mathbf{A}^\pi$.
  $\langle \mathbf{A}, \sum_{\pi \in \mathcal{S}_4} \mathbf{A}^\pi \rangle \geqslant \Omega(n^4)$.
- Permutations have lower-bounded spectrum.
  For every $\pi \in \mathcal{S}_4$*, the Hermitian* $n^2 \times n^2$ *unfolding* $\frac{1}{2}(A^\pi + (A^\pi)^T)$ *of* $\mathbf{A}^\pi$ *has no eigenvalues smaller than* $-\lambda^2$.
- Using $\mathbf{A}$ as 4-th pseudo-moments does not imply that $\|x\|^4$ is too large.
  For every $\pi \in \mathcal{S}_4$*, we have* $\langle \mathrm{Id}^{\mathrm{sym}}, A^\pi \rangle \leqslant O(\lambda^2 n^{3/2})$
- Using $\mathbf{A}$ for 4-th pseudo-moments does not imply first and second degree moments are too large.
  Let $\mathcal{L} : \mathbb{R}[x]_4 \to \mathbb{R}$ *be the linear functional given by the matrix representation* $M_{\mathcal{L}} := \frac{1}{\lambda^2 n^2} \sum_{\pi \in \mathcal{S}_4} A^\pi$*. Let*

  $$\delta_2 \stackrel{\mathrm{def}}{=} \max_{i \neq j} \left| \mathcal{L} \, \|x\|_2^2 x_i x_j \right|$$
  $$\delta_2' \stackrel{\mathrm{def}}{=} \max_{i} \left| \mathcal{L} \, \|x\|_2^2 x_i^2 \right|$$

  Then $n^{3/2}\delta_2' + n^2 \delta_2 \leqslant O(1)$.

*Then there is a degree-*4 *pseudo-distribution* $\{x\}$ *satisfying* $\{\|x\|_2^2 = 1\}$ *so that* $\tilde{\mathbb{E}}\, \mathbf{A}(x) \geqslant \Omega(n^2/\lambda^2) + \Theta(\mathbb{E}^\mu\, \mathbf{A}(x))$.

The degree-3 version of our lower bound requires bounds on the spectra of the flattenings not just of the 3-tensor $\mathbf{A}$ itself but also of the flattenings of an associated 4-tensor, which represents the polynomial $\langle x^{\otimes 2}, (\sum_i A_i \otimes A_i) x^{\otimes 2} \rangle$.

Theorem 32 (Degree-3 Lower Bound, General Version) *Let* $\mathbf{A}$ *be a* 3*-tensor and let* $\lambda > 0$ *be a function of n. Suppose the following conditions hold:*

- $\mathbf{A}$ is significantly correlated with $\sum_{\pi \in \mathcal{S}_3} \mathbf{A}^\pi$.
  $\langle \mathbf{A}, \sum_{\pi \in \mathcal{S}_3} \mathbf{A}^\pi \rangle \geqslant \Omega(n^3)$.
- Permutations have lower-bounded spectrum.
  For every $\pi \in \mathcal{S}_3$*, we have*

  $$-2\lambda^2 \cdot \Pi \, \mathrm{Id} \, \Pi \preceq \frac{1}{2}\Pi(\sigma \cdot A^\pi (A^\pi)^T + \sigma^2 \cdot A^\pi (A^\pi)^T)\Pi + \frac{1}{2}\Pi(\sigma \cdot A^\pi (A^\pi)^T + \sigma^2 \cdot A^\pi (A^\pi)^T)^T \Pi \,.$$
- Using AA^T for 4-th moments does not imply $\|x\|^4$ is too large.
  For every $\pi \in \mathcal{S}_3$*, we have* $\langle \mathrm{Id}^{\mathrm{sym}}, A^\pi (A^\pi)^T \rangle \leqslant O(\lambda^2 n^2)$

— Using A and AA^T for 3rd and 4th moments do not imply first and second degree moments are too large.

Let $\pi \in \mathcal{S}_3$. Let $\mathcal{L} : \mathbb{R}[x]_4 \to \mathbb{R}$ be the linear functional given by the matrix representation $M_{\mathcal{L}} := \frac{1}{\lambda^2 n^2} \sum_{\pi' \in \mathcal{S}_4} \pi' \cdot AA^T$. Let

$$\delta_1 \overset{\text{def}}{=} \max_i \left| \frac{1}{\lambda n^{3/2}} \langle \mathrm{Id}_{n \times n}\,, A_i^\pi \rangle \right|$$

$$\delta_2 \overset{\text{def}}{=} \max_{i \neq j} \left| \mathcal{L}\, \|x\|_2^2 x_i x_j \right|$$

$$\delta_2' \overset{\text{def}}{=} \max_i \left| \mathcal{L}\, \|x\|_2^2 x_i^2 - \tfrac{1}{n}\, \mathcal{L}\, \|x\|_2^4 \right|$$

Then $n\delta_1 + n^{3/2}\delta_2' + n^2\delta_2 \leqslant O(1)$.

Then there is a degree-4 pseudo-distribution $\{x\}$ satisfying $\{\|x\|_2^2 = 1\}$ so that

$$\tilde{\mathbb{E}}\, \mathbf{A}(x) \geqslant \Omega\left(\frac{n^{3/2}}{\lambda}\right) + \Theta(\mathbb{E}^\mu\, \mathbf{A}(x))\,.$$

B.2.1. PROOF OF THEOREM 30

We prove the degree-3 corollary; the degree-4 case is almost identical using Theorem 31 and Lemma 63 in place of their degree-3 counterparts.

Proof Let $\mathbf{A}$ be a 3-tensor. If $\mathbf{A}$ satisfies the conditions of Theorem 32 with $\lambda = O(n^{3/4} \log(n)^{1/4})$, we let $\eta(\mathbf{A})$ be the pseudo-distribution described there, with

$$\tilde{\mathbb{E}}_{x \sim \eta(\mathbf{A})}\, \mathbf{A}(x) \geqslant \Omega\left(\frac{n^{3/2}}{\lambda}\right) + \Theta(\mathbb{E}^\mu\, \mathbf{A}(x))$$

If $\mathbf{A}$ does not satisfy the regularity conditions, we let $\eta(\mathbf{A})$ be the uniform distribution on the unit sphere. If $\mathbf{A}$ has unit Gaussian entries, then Lemma 62 says that the regularity conditions are satisfied with this choice of λ with high probability. The operator norm of $\mathbf{A}$ is at most $O(\sqrt{n})$, so $\mathbb{E}^\mu\, \mathbf{A}(x) = O(\sqrt{n})$ (all with high probability) (Tomioka and Suzuki, 2014). We have chosen λ and τ so that when the conditions of Theorem 32 and the bound on $\mathbb{E}^\mu\, \mathbf{A}(x)$, obtain,

$$\tilde{\mathbb{E}}_{x \sim \eta(\mathbf{A})}\, \tau \cdot \langle v_0, x \rangle^3 + \mathbf{A}(x) \geqslant \Omega\left(\frac{n^{3/4}}{\log(n)^{1/4}}\right)\,.$$

On the other hand, our arguments on degree-4 SoS certificates for random polynomials say with high probability every degree-4 pseudo-distribution $\{y\}$ satisfying $\{\|y\|^2 = 1\}$ has $\tilde{\mathbb{E}}\, \tau \cdot \langle v, y \rangle^3 + \mathbf{A}(y) \leqslant O(n^{3/4} \log(n)^{1/4})$. Thus, $\{x\}$ is nearly optimal and we are done. ■

B.3. In-depth Preliminaries for Pseudo-Expectation Symmetries

This section gives the preliminaries we will need to construct maximally-symmetric matrices (a.k.a. functionals $\mathcal{L} : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$) in what follows. For a non-maximally-symmetric $M \in \mathbb{R}^{n^2 \times n^2}$ under the action of $\mathcal{S}_4$ by permutation of indices, the subgroup $C_3 < \mathcal{S}_4$ represents all the significant permutations whose spectra may differ from one another in a nontrivial way. The lemmas that follow will make this more precise. For concreteness, we take $C_3 = \langle \sigma \rangle$ with $\sigma = (234)$, but any other choice of 3-cycle would lead to a merely syntactic change in the proof.

Lemma 33 *Let $\mathcal{D}_8 < \mathcal{S}_4$ be given by $\mathcal{D}_8 = \langle(12), (34), (13)(24)\rangle$. Let $C_3 = \{(), \sigma, \sigma^2\} = \langle\sigma\rangle$, where $()$ denotes the identity in $\mathcal{S}_4$. Then $\{gh : g \in \mathcal{D}_8, h \in C_3\} = \mathcal{S}_4$.*

Proof The proof is routine; we provide it here for completeness. Note that C_3 is a subgroup of order 3 in the alternating group $\mathcal{A}_4$. This alternating group can be decomposed as $\mathcal{A}_4 = \mathcal{K}_4 \cdot C_3$, where $\mathcal{K}_4 = \langle(12)(34), (13)(24)\rangle$ is a normal subgroup of $\mathcal{A}_4$. We can also decompose $\mathcal{S}_4 = C_2 \cdot \mathcal{A}_4$ where $C_2 = \langle(12)\rangle$ and $\mathcal{A}_4$ is a normal subgroup of $\mathcal{S}_4$. Finally, $\mathcal{D}_8 = C_2 \cdot \mathcal{K}_4$ so by associativity, $\mathcal{S}_4 = C_2 \cdot \mathcal{A}_4 = C_2 \cdot \mathcal{K}_4 \cdot C_3 = \mathcal{D}_8 \cdot C_3$. ■

This lemma has two useful corollaries:

Corollary 34 *For any subset $S \subseteq \mathcal{S}_4$, we have $\{ghs : g \in \mathcal{D}_8, h \in C_3, s \in S\} = \mathcal{S}_4$.*

Corollary 35 *Let $M \in \mathbb{R}^{n^2 \times n^2}$. Let the matrix M' be given by*

$$M' \overset{\text{def}}{=} \frac{1}{2}\Pi\left(M + \sigma \cdot M + \sigma^2 \cdot M\right)\Pi + \frac{1}{2}\Pi\left(M + \sigma \cdot M + \sigma^2 \cdot M\right)^T \Pi\,.$$

Then $M' \in \operatorname{Sym} M$.

Proof Observe first that $M + \sigma \cdot M + \sigma^2 \cdot M = \sum_{\pi \in C_3} \pi \cdot M$. For arbitrary $N \in \mathbb{R}^{n^2 \times n^2}$, we show that $\frac{1}{2}\Pi N \Pi + \frac{1}{2}\Pi N^T \Pi = \frac{1}{8}\sum_{\pi \in \mathcal{D}_8} \pi \cdot N$. First, conjugation by Π corresponds to averaging M over the group $\langle(12), (34)\rangle$ generated by interchange of indices in row and column indexing pairs, individually. At the same time, $N + N^T$ is the average of M over the matrix transposition permutation group $\langle(13)(24)\rangle$. All together,

$$M' = \frac{1}{8}\sum_{g \in \mathcal{D}_8}\sum_{h \in C_3}(gh) \cdot M = \frac{1}{8}\sum_{\pi \in \mathcal{S}_4} \pi \cdot M$$

and so $M' \in \operatorname{Sym} M$. ■

We make an useful observation about the nontrivial permutations of M, in the special case that $M = AA^T$ for some 3-tensor $\mathbf{A}$.

Lemma 36 *Let $\mathbf{A}$ be a 3-tensor and let $A \in \mathbb{R}^{n^2 \times n}$ be its flattening, where the first and third modes lie on the longer axis and the third mode lies on the shorter axis. Let A_i be the $n \times n$ matrix slices of $\mathbf{A}$ along the first mode, so that*

$$A = \begin{pmatrix} A_1 \\ A_2 \\ \vdots \\ A_n \end{pmatrix}.$$

Let $P : \mathbb{R}^{n^2} \to \mathbb{R}^{n^2}$ be the orthogonal linear operator so that $[Px](i, j) = x(j, i)$. Then

$$\sigma \cdot AA^T = \left(\sum_i A_i \otimes A_i\right)P \qquad \text{and} \qquad \sigma^2 \cdot AA^T = \sum_i A_i \otimes A_i^T\,.$$

Proof We observe that $AA^T[(j_1, j_2), (j_3, j_4)] = \sum_i A_{ij_1j_2}A_{ij_3j_4}$ and that $(\sum_i A_i \otimes A_i)[(j_1, j_2), (j_3, j_4)] = \sum_i A_{ij_1j_3}A_{ij_2j_4}$. Multiplication by P on the right has the effect of switching the order of the second indexing pair, so $[(\sum_i A_i \otimes A_i)P][(j_1, j_2), (j_3, j_4)] = \sum_i A_{ij_1j_4}A_{ij_2j_3}$. From this it is easy to see that $\sigma \cdot AA^T = (234) \cdot AA^T = (\sum_i A_i \otimes A_i)P$.

Similarly, we have that

$$(\sigma^2 AA^T)[(j_1, j_2), (j_3, j_4)] = ((243) \cdot AA^T)[(j_1, j_2), (j_3, j_4)] = \sum_k A_{ij_1j_3} A_{ij_4j_2} ,$$

from which we see that $\sigma^2 \cdot AA^T = \sum_i A_i \otimes A_i^T$. ■

Permutations of the Identity Matrix. The nontrivial permutations of $\mathrm{Id}_{n^2 \times n^2}$ are:

$$\begin{aligned} \mathrm{Id}\,[(j,k),(j',k')] &= \delta(j,k)\delta(j',k') \\ \sigma \cdot \mathrm{Id}\,[(j,k),(j',k')] &= \delta(j,j')\delta(k,k') \\ \sigma^2 \cdot \mathrm{Id}\,[(j,k),(j',k')] &= \delta(j,k')\delta(j',k)\,. \end{aligned}$$

Since $(\mathrm{Id} + \sigma \cdot \mathrm{Id} + \sigma^2 \cdot \mathrm{Id})$ is invariant under the action of $\mathcal{D}_8$, we have $(\mathrm{Id} + \sigma \cdot \mathrm{Id} + \sigma^2 \cdot \mathrm{Id}) \in \mathrm{Sym}\, M$; up to scaling this matrix is the same as $\mathrm{Id}^{\mathrm{sym}}$ defined in (B.1). We record the following observations:

- Id, $\sigma \cdot \mathrm{Id}$, and $\sigma^2 \cdot \mathrm{Id}$ are all symmetric matrices.
- Up to scaling, $\mathrm{Id} + \sigma^2\, \mathrm{Id}$ projects to identity on the canonical embedding of $\mathbb{R}[x]_2$.
- The matrix $\sigma \cdot \mathrm{Id}$ is rank-1, positive-semidefinite, and has $\Pi(\sigma \cdot \mathrm{Id})\Pi = \sigma \cdot \mathrm{Id}$.
- The scaling $[1/(n^2 + 2n)](\mathrm{Id} + \sigma \cdot \mathrm{Id} + \sigma^2 \cdot \mathrm{Id})$ is equal to a linear functional $\mathbb{E}^{\mu} : \mathbb{R}[x]_4 \to \mathbb{R}$ giving the expectation under the uniform distribution over the unit sphere S^{n-1}.

B.4. Construction of Initial Pseudo-Distributions

We begin by discussing how to create an initial guess at a pseudo-distribution whose third moments are highly correlated with the polynomial $\mathbf{A}(x)$. This initial guess will be a valid pseudo-distribution, but will fail to be on the unit sphere, and so will require some repairing later on. For now, the method of creating this initial pseudo-distribution involves using a combination of symmetrization techniques to ensure that the matrices we construct are well defined as linear functionals over polynomials, and spectral techniques to establish positive-semidefiniteness of these matrices.

B.4.1. EXTENDING PSEUDO-DISTRIBUTIONS TO DEGREE FOUR

In this section we discuss a construction that takes a linear functional $\mathcal{L} : \mathbb{R}[x]_{\leqslant 3} \to \mathbb{R}$ over degree-3 polynomials and yields a degree-4 pseudo-distribution $\{x\}$. We begin by reminding the reader of the Schur complement criterion for positive-semidefiniteness of block matrices.

Theorem 37 *Let M be the following block matrix.*

$$M \overset{\mathrm{def}}{=} \begin{pmatrix} B & C^T \\ C & D \end{pmatrix}$$

where $B \succeq 0$ and is full rank. Then $M \succeq 0$ if and only if $D \succeq CB^{-1}C^T$.

Suppose we are given a linear functional $\mathcal{L} : \mathbb{R}[x]_{\leqslant 3} \to \mathbb{R}$ with $\mathcal{L}\,1 = 1$. Let $\mathcal{L}|_1$ be $\mathcal{L}$ restricted to $\mathbb{R}[x]_1$ and similarly for $\mathcal{L}|_2$ and $\mathcal{L}|_3$. We define the following matrices:

— $M_{\mathcal{L}|_1} \in \mathbb{R}^{n\times 1}$ is the matrix representation of $\mathcal{L}|_1$.

— $M_{\mathcal{L}|_2} \in \mathbb{R}^{n\times n}$ is the matrix representation of $\mathcal{L}|_2$.

— $M_{\mathcal{L}|_3} \in \mathbb{R}^{n^2\times n}$ is the matrix representation of $\mathcal{L}|_3$.

— $V_{\mathcal{L}|_2} \in \mathbb{R}^{n^2\times 1}$ is the vector flattening of $M_{\mathcal{L}|_2}$.

Consider the block matrix $M \in \mathbb{R}^{\#\text{tuples}(2)\times\#\text{tuples}(2)}$ given by

$$M \stackrel{\text{def}}{=} \begin{pmatrix} 1 & M_{\mathcal{L}|_1}^T & V_{\mathcal{L}|_2}^T \\ M_{\mathcal{L}|_1} & M_{\mathcal{L}|_2} & M_{\mathcal{L}|_3}^T \\ V_{\mathcal{L}|_2} & M_{\mathcal{L}|_3} & D \end{pmatrix},$$

with $D \in \mathbb{R}^{n^2\times n^2}$ yet to be chosen. By taking

$$B = \begin{pmatrix} 1 & M_{\mathcal{L}|_1}^T \\ M_{\mathcal{L}|_1} & M_{\mathcal{L}|_2} \end{pmatrix} \qquad C = \begin{pmatrix} V_{\mathcal{L}|_2} & M_{\mathcal{L}|_3} \end{pmatrix},$$

we see by the Schur complement criterion that M is positive-semidefinite so long as $D \succeq CB^{-1}C^T$. However, not any choice of D will yield M maximally symmetric, which is necessary for M to define a pseudo-expectation operator $\tilde{\mathbb{E}}$.

We would ideally take D to be the spectrally-least maximally-symmetric matrix so that $D \succeq CB^{-1}C^T$. But this object might not be well defined, so we instead take the following substitute.

Definition 38 *Let $\mathcal{L}, B, C$ as be as above. The* symmetric Schur complement $D \in \operatorname{Sym} CB^{-1}C^T$ *is $t\sum_{\pi\in\mathcal{S}_4}\pi\cdot(CB^{-1}C^T)$ for the least t so that $t\sum_{\pi\in\mathcal{S}_4}\pi\cdot(CB^{-1}C^T) \succeq CB^{-1}C^T$. We denote by $\tilde{\mathbb{E}}^{\mathcal{L}}$ the linear functional $\tilde{\mathbb{E}}^{\mathcal{L}} : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$ whose matrix representation is M with this choice of D, and note that $\tilde{\mathbb{E}}^{\mathcal{L}}$ is a valid degree-4 pseudo-expectation.*

Example 1 (Recovery of Degree-4 Uniform Moments from Symmetric Schur Complement) *Let $\mathcal{L} : \mathbb{R}[x]_{\leqslant 3} \to \mathbb{R}$ be given by $\mathcal{L}\,p(x) := \mathbb{E}^\mu\, p(x)$. We show that $\tilde{\mathbb{E}}^{\mathcal{L}} = \mathbb{E}^\mu$. In this case it is straightforward to compute that $CB^{-1}C^T = \sigma\cdot\mathrm{Id}/n^2$. Our task is to pick $t \geqslant 0$ minimal so that $\frac{t}{n^2}\Pi(\mathrm{Id}+\sigma\cdot\mathrm{Id}+\sigma^2\cdot\mathrm{Id})\Pi \succeq \frac{1}{n^2}\Pi(\sigma\cdot\mathrm{Id})\Pi$.*

We know that $\Pi(\sigma\cdot\mathrm{Id})\Pi = \sigma\cdot\mathrm{Id}$. Furthermore, $\Pi\,\mathrm{Id}\,\Pi = \Pi(\sigma^2\cdot\mathrm{Id})\Pi$, and both are the identity on the canonically-embedded subspace $\mathbb{R}[x]_2$ in $\mathbb{R}^{\#\text{tuples}(4)}$. We have previously observed that $\sigma\cdot\mathrm{Id}$ is rank-one and positive-semidefinite, so let $w \in \mathbb{R}^{\#\text{tuples}(4)}$ be such that $ww^T = \sigma\cdot\mathrm{Id}$.

We compute $w^T(\mathrm{Id}+\sigma\cdot\mathrm{Id}+\sigma^2\cdot\mathrm{Id})w = 2\|w\|_2^2 + \|w\|_2^4 = 2n+n^2$ and $w^T(\sigma\cdot\mathrm{Id})w = \|w\|_2^4 = n^2$. Thus $t = n^2/(n^2+2n)$ is the minimizer. By a previous observation, this yields $\mathbb{E}^\mu$.

To prove our lower bound, we will generalize the above example to the case that we start with an operator $\mathcal{L} : \mathbb{R}[x]_{\leqslant 3} \to \mathbb{R}$ which does not match $\mathbb{E}^\mu$ on degree-3 polynomials.

B.4.2. SYMMETRIES AT DEGREE THREE

We intend on using the symmetric Schur complement to construct a pseudo-distribution from some $\mathcal{L} : \mathbb{R}[x]_{\leqslant 3} \to \mathbb{R}$ for which $\mathcal{L}\,\mathbf{A}(x)$ is large. A good such $\mathcal{L}$ will have $\mathcal{L}\,x_i x_j x_k$ correlated with $\sum_{\pi \in \mathcal{S}_3} \mathbf{A}^{\pi}_{ijk}$ for all (or many) indices i, j, k. That is, it should be correlated with the coefficient of the monomial $x_i x_j x_k$ in $\mathbf{A}(x)$. However, if we do this directly by setting $\mathcal{L}\,x_i x_j x_k = \sum_{\pi} \mathbf{A}^{\pi}_{ijk}$, it becomes technically inconvenient to control the spectrum of the resulting symmetric Schur complement. To this avoid, we discuss how to utilize a decomposition of $M_{\mathcal{L}|_3}$ into nicer matrices if such a decomposition exists.

Lemma 39 *Let $\mathcal{L} : \mathbb{R}[x]_{\leqslant 3} \to \mathbb{R}$, and suppose that $M_{\mathcal{L}|_3} = \frac{1}{k}(M^1_{\mathcal{L}|_3} + \cdots + M^k_{\mathcal{L}|_3})$ for some $M^1_{\mathcal{L}|_3}, \ldots, M^k_{\mathcal{L}|_3} \in \mathbb{R}^{n^2 \times n}$. Let $D_1, \ldots, D_k$ be the respective symmetric Schur complements of the family of matrices*

$$\left\{ \begin{pmatrix} 1 & M^T_{\mathcal{L}|_1} & V^T_{\mathcal{L}|_2} \\ M_{\mathcal{L}|_1} & M_{\mathcal{L}|_2} & (M^i_{\mathcal{L}|_3})^T \\ V_{\mathcal{L}|_2} & M^i_{\mathcal{L}|_3} & \bullet \end{pmatrix} \right\}_{1 \leqslant i \leqslant k} .$$

Then the matrix

$$M \overset{\text{def}}{=} \frac{1}{k} \sum_{i=1}^{k} \begin{pmatrix} 1 & M^T_{\mathcal{L}|_1} & V^T_{\mathcal{L}|_2} \\ M_{\mathcal{L}|_1} & M_{\mathcal{L}|_2} & (M^i_{\mathcal{L}|_3})^T \\ V_{\mathcal{L}|_2} & M^i_{\mathcal{L}|_3} & D_i \end{pmatrix}$$

is positive-semidefinite and maximally symmetric. Therefore it defines a valid pseudo-expectation $\tilde{\mathbb{E}}^{\mathcal{L}}$. (This is a slight abuse of notation, since the pseudo-expectation defined here in general differs from the one in Definition 38.)

Proof Each matrix in the sum defining M is positive-semidefinite, so $M \geq 0$. Each D_i is maximally symmetric and therefore so is $\sum_{i=1}^{k} D_i$. We know that $M_{\mathcal{L}|_3} = \sum_{i=1}^{k} M^i_{\mathcal{L}|_3}$ is maximally-symmetric, so it follows that M is the matrix representation of a valid pseudo-expectation. ■

B.5. Getting to the Unit Sphere

Our next tool takes a pseudo-distribution $\tilde{\mathbb{E}}$ that is slightly off the unit sphere, and corrects it to give a linear functional $\mathcal{L} : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$ that lies on the unit sphere.

We will also characterize how badly the resulting linear functional deviates from the nonnegativity condition ($\mathcal{L}\,p(x)^2 \geqslant 0$ for $p \in \mathbb{R}[x]_{\leqslant 2}$) required to be a pseudo-distribution

Definition 40 *Let $\mathcal{L} : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$. We define*

$$\lambda_{\min} \mathcal{L} \overset{\text{def}}{=} \min_{p \in \mathbb{R}[x]_{\leqslant d/2}} \frac{\mathcal{L}\,p(x)^2}{\mathbb{E}^{\mu}\,p(x)^2}$$

where $\mathbb{E}^{\mu}\,p(x)^2$ is the expectation of $p(x)^2$ when x is distributed according to the uniform distribution on the unit sphere.

Since $\mathbb{E}^{\mu}\,p(x)^2 \geqslant 0$ for all p, we have $\mathcal{L}\,p(x)^2 \geqslant 0$ for all p if and only if $\lambda_{\min} \mathcal{L} \geqslant 0$. Thus $\mathcal{L}$ on the unit sphere is a pseudo-distribution if and only if $\mathcal{L}\,1 = 1$ and $\lambda_{\min} \mathcal{L} \geqslant 0$.

Lemma 41 *Let $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$ be a valid pseudodistribution. Suppose that:*

1. $c := \tilde{\mathbb{E}} \|x\|_2^4 \geqslant 1$.
2. *$\tilde{\mathbb{E}}$ is close to lying on the sphere, in the sense that there are $\delta_1, \delta_2, \delta_2' \geqslant 0$ so that:*
   (a) $|\frac{1}{c} \tilde{\mathbb{E}} \|x\|_2^2 x_i - \mathcal{L}' \, x_i| \leqslant \delta_1$ *for all i.*
   (b) $|\frac{1}{c} \tilde{\mathbb{E}} \|x\|_2^2 x_i x_j - \mathcal{L}' \, x_i x_j| \leqslant \delta_2$ *for all $i \neq j$.*
   (c) $|\frac{1}{c} \tilde{\mathbb{E}} \|x\|_2^2 x_i^2 - \mathcal{L}' \, x_i^2| \leqslant \delta_2'$ *for all i.*

Let $\mathcal{L} : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$ be as follows on homogeneous p:

$$\mathcal{L}\, p(x) \overset{\text{def}}{=} \begin{cases} \tilde{\mathbb{E}}\, 1 & \textit{if} \deg p = 0 \\ \frac{1}{c} \tilde{\mathbb{E}}\, p(x) & \textit{if} \deg p = 3, 4 \\ \frac{1}{c} \tilde{\mathbb{E}}\, p(x) \|x\|_2^2 & \textit{if} \deg p = 1, 2\,. \end{cases}$$

Then $\mathcal{L}$ satisfies $\mathcal{L}\, p(x)(\|x\|_2^2 - 1) = 0$ for all $p(x) \in \mathbb{R}[x]_{\leqslant 2}$ and has $\lambda_{\min} \mathcal{L} \geqslant -\frac{c-1}{c} - O(n)\delta_1 - O(n^{3/2})\delta_2' - O(n^2)\delta_2$.

Proof It is easy to check that $\mathcal{L}\, p(x)(\|x\|_2^2 - 1) = 0$ for all $p \in \mathbb{R}[x]_{\leqslant 2}$ by expanding the definition of $\mathcal{L}$.

Let the linear functional $\mathcal{L}' : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$ be defined over homogeneous polynomials p as

$$\mathcal{L}'\, p(x) \overset{\text{def}}{=} \begin{cases} c & \text{if} \deg p = 0 \\ \tilde{\mathbb{E}}\, p(x) & \text{if} \deg p = 3, 4 \\ \tilde{\mathbb{E}}\, p(x) \|x\|_2^2 & \text{if} \deg p = 1, 2\,. \end{cases}$$

Note that $\mathcal{L}'\, p(x) = c\, \mathcal{L}\, p(x)$ for all $p \in \mathbb{R}[x]_{\leqslant 4}$. Thus $\lambda_{\min} \mathcal{L} \geqslant \lambda_{\min} \mathcal{L}' / c$, and the kernel of $\mathcal{L}'$ is identical to the kernel of $\mathcal{L}$.

In particular, since $(\|x\|_2^2 - 1)$ is in the kernel of $\mathcal{L}'$, either $\lambda_{\min} \mathcal{L}' = 0$ or

$$\lambda_{\min} \mathcal{L}' = \min_{p \in \mathbb{R}[x]_{\leqslant 2}, p \perp (\|x\|_2^2 - 1)} \frac{\mathcal{L}'\, p(x)^2}{\mathbb{E}^{\mu}\, p(x)^2}.$$

Here $p \perp (\|x\|^2 - 1)$ means that the polynomials p and $\|x\|^2 - 1$ are perpendictular in the coefficient basis. That is, if $p(x) = p_0 + \sum_i p_i x_i + \sum_{ij} p_{ij} x_i x_j$, this means $\sum_{ii} p_{ii} = p_0$. The equality holds because any linear functional on polynomials $\mathcal{K}$ with $(\|x\|^2 - 1)$ in its kernel satisfies $\mathcal{K}(p(x) + \alpha(\|x\|^2 - 1))^2 = \mathcal{K} p(x)^2$ for every α. The functionals $\mathcal{L}'$ and $\mathbb{E}^{\mu}$ in particular both satisfy this.

Let $\Delta := \mathcal{L}' - \tilde{\mathbb{E}}$, and note that Δ is nonzero only when evaluated on the degree-1 or -2 parts of polynomials. It will be sufficient to bound Δ, since assuming $\lambda_{\min} \mathcal{L}' \neq 0$,

$$\begin{aligned} \lambda_{\min} \mathcal{L}' &= \min_{p \in \mathbb{R}[x]_{\leqslant 2}, p \perp (\|x\|_2^2 - 1)} \frac{\Delta p(x)^2 + \tilde{\mathbb{E}}\, p(x)^2}{\mathbb{E}^{\mu}\, p(x)^2} \\ &\geqslant \min_{p \in \mathbb{R}[x]_{\leqslant 2}, p \perp (\|x\|_2^2 - 1)} \frac{\Delta p(x)^2}{\mathbb{E}^{\mu}\, p(x)^2}\,. \end{aligned}$$

Let $p \in \mathbb{R}[x]_{\leqslant 2}$. We expand p in the monomial basis: $p(x) = p_0 + \sum_i p_i x_i + \sum_{i,j} p_{ij} x_i x_j$. Then

$$p(x)^2 = p_0^2 + 2p_0 \sum_i p_i x_i + 2p_0 \sum_{ij} p_{ij} x_i x_j + \left(\sum_i p_i x_i\right)^2 + 2\left(\sum_i p_i x_i\right)\left(\sum_{ij} p_{ij} x_i x_j\right) + \left(\sum_{ij} p_{ij} x_i x_j\right)^2 .$$

An easy calculation gives

$$\mathbb{E}^\mu\, p(x)^2 = p_0^2 + \frac{2p_0}{n}\sum_i p_{ii} + \frac{1}{n}\sum_i p_i^2 + \frac{1}{n^2+2n}\left(\left(\sum_i p_{ii}\right)^2 + \sum_{ij} p_{ij}^2 + \sum_i p_{ii}^2\right).$$

The condition $p \perp (\|x\|_2^2 - 1)$ yields $p_0 = \sum_i p_{ii}$. Substituting into the above, we obtain the sum of squares

$$\mathbb{E}^\mu\, p(x)^2 = p_0^2 + \frac{2p_0^2}{n} + \frac{1}{n}\sum_i p_i^2 + \frac{1}{n^2+2n}\left(p_0^2 + \sum_{ij} p_{ij}^2 + \sum_i p_{ii}^2\right).$$

Without loss of generality we assume $\mathbb{E}^\mu\, p(x)^2 = 1$, so now it is enough just to bound $\Delta p(x)^2$. We have assumed that $|\Delta x_i| \leqslant c\delta_1$ and $|\Delta x_i x_j| \leqslant c\delta_2$ for $i \neq j$ and $|\Delta x_i^2| \leqslant c\delta_2'$. We also know $\Delta 1 = c - 1$ and $\Delta p(x) = 0$ when p is a homogeneous degree-3 or -4 polynomial. So we expand

$$\Delta p(x)^2 = p_0^2(c-1) + 2p_0 \sum_i p_i \Delta x_i + 2p_0 \sum_{ij} p_{ij} \Delta x_i x_j + \sum_{i,j} p_i p_j \Delta x_i x_j$$

and note that this is maximized in absolute value when all the signs line up:

$$|\Delta p(x)^2| \leqslant p_0^2(c-1) + 2c\delta_1 |p_0| \sum_i |p_i| + 2|p_0|\left(c\delta_2 \sum_{i\neq j} |p_{ij}| + c\delta_2' \sum_i |p_{ii}|\right) + c\delta_2 \left(\sum_i |p_i|\right)^2 + c\delta_2' \sum_i p_i^2 .$$

We start with the second term. If $p_0^2 = \alpha$ for $\alpha \in [0,1]$, then $\sum_i p_i^2 \leqslant n(1-\alpha)$ by our assumption that $\mathbb{E}^\mu\, p(x)^2 = 1$. This means that

$$2c\delta_1 |p_0| \sum_i |p_i| \leqslant 2c\delta_1 \sqrt{\alpha n \sum_i p_i^2} \leqslant 2c\delta_1 n \sqrt{\alpha(1-\alpha)} \leqslant O(n) c\delta_1 ,$$

where we have used Cauchy-Schwarz and the fact $\max_{0\leqslant\alpha\leqslant 1} \alpha(1-\alpha) = (1/2)^2$. The other terms are all similar:

$$\begin{aligned} p_0^2(c-1) &\leqslant c - 1 \\ 2|p_0| c\delta_2 \sum_{i\neq j} |p_{ij}| \leqslant 2c\delta_2 \sqrt{\alpha n^2 \sum_{ij} p_{ij}^2} \leqslant 2c\delta_2 O(n^2)\sqrt{\alpha(1-\alpha)} &\leqslant O(n^2) c\delta_2 \\ 2|p_0| c\delta_2' \sum_i |p_{ii}| \leqslant 2c\delta_2' \sqrt{\alpha n \sum_i p_{ii}^2} &\leqslant O(n^{3/2}) c\delta_2' \\ c\delta_2 \left(\sum_i |p_i|\right)^2 \leqslant c\delta_2 n \sum_i p_i^2 &\leqslant O(n^2) c\delta_2 \end{aligned}$$

$$c\delta_2' \sum_i p_i^2 \leqslant O(n) c\delta_2' \,,$$

where in each case we have used Cauchy-Schwarz and our assumption $\mathbb{E}^\mu \, p(x)^2 = 1$.

Putting it all together, we get

$$\lambda_{\min} \Delta \geqslant -(c-1) - O(n)c\delta_1 - O(n^{3/2})c\delta_2' - O(n^2)c\delta_2 \,.$$

■

B.6. Repairing Almost-Pseudo-Distributions

Our last tool takes a linear functional $\mathcal{L} : \mathbb{R}[x]_{\leqslant d}$ that is "almost" a pseudo-distribution over the unit sphere, in the precise sense that all conditions for being a pseudo-distribution over the sphere are satisfied except that $\lambda_{\min} \mathcal{L} = -\varepsilon$. The tool transforms it into a bona fide pseudo-distribution at a slight cost to its evaluations at various polynomials.

Lemma 42 *Let $\mathcal{L} : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$ and suppose that*

— $\mathcal{L}\,1 = 1$

— $\mathcal{L}\, p(x)(\|x\|^2 - 1) = 0$ *for all* $p \in \mathbb{R}[x]_{\leqslant d-2}$.

— $\lambda_{\min} \mathcal{L} = -\varepsilon$.

Then the operator $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$ *given by*

$$\tilde{\mathbb{E}} \, p(x) \overset{\text{def}}{=} \frac{1}{1+\varepsilon}(\mathcal{L}\, p(x) + \varepsilon \, \mathbb{E}^\mu \, p(x))$$

is a valid pseudo-expectation satisfying $\{\|x\|^2 = 1\}$.

Proof It will suffice to check that $\lambda_{\min} \tilde{\mathbb{E}} \geqslant 0$ and that $\tilde{\mathbb{E}}$ has $\tilde{\mathbb{E}}(\|x\|_2^2 - 1)^2 = 0$ and $\tilde{\mathbb{E}}\,1 = 1$. For the first, let $p \in \mathbb{R}[x]_{\geqslant 2}$. We have

$$\frac{\tilde{\mathbb{E}} \, p(x)^2}{\mathbb{E}^\mu \, p(x)^2} = \left(\frac{1}{1+\varepsilon}\right)\left(\frac{\mathbb{E}^0 \, p(x)^2 + \varepsilon \, \mathbb{E}^\mu \, p(x)^2}{\mathbb{E}^\mu \, p(x)^2}\right) \geqslant \left(\frac{1}{1+\varepsilon}\right)(-\varepsilon + \varepsilon) \geqslant 0 \,.$$

Hence, $\lambda_{\min} \tilde{\mathbb{E}} \geqslant 0$.

It is straightforward to check the conditions that $\tilde{\mathbb{E}}\,1 = 1$ and that $\tilde{\mathbb{E}}$ satisfies $\{\|x\|^2 - 1 = 0\}$, since $\tilde{\mathbb{E}}$ is a convex combination of linear functionals that already satisfy these linear constraints. ■

B.7. Putting Everything Together

We are ready to prove Theorem 31 and Theorem 32. The proof of Theorem 31 is somewhat simpler and contains many of the ideas of the proof of Theorem 32, so we start there.

B.7.1. THE DEGREE-4 LOWER BOUND

Proof [Proof of Theorem 31] We begin by constructing a degree-4 pseudo-expectation $\tilde{\mathbb{E}}^0 : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$ whose degree-4 moments are biased towards $\mathbf{A}(x)$ but which does not yet satisfy $\{\|x\|_2^2 - 1 = 0\}$.

Let $\mathcal{L} : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$ be the functional whose matrix representation when restricted to $\mathcal{L}|_4 : \mathbb{R}[x]_4 \to \mathbb{R}$ is given by $M_{\mathcal{L}|_4} = \frac{1}{|\mathcal{S}_4| n^2} \sum_{\pi \in \mathcal{S}_4} A^\pi$, and which is 0 on polynomials of degree at most 3.

Let $\tilde{\mathbb{E}}^0 := \mathbb{E}^\mu + \varepsilon \mathcal{L}$, where ε is a parameter to be chosen soon so that $\tilde{\mathbb{E}}^0\, p(x)^2 \geqslant 0$ for all $p \in \mathbb{R}[x]_{\leqslant 2}$. Let $p \in \mathbb{R}[x]_{\leqslant 2}$. We expand p in the monomial basis as $p(x) = p_0 + \sum_i p_i x_i + \sum_{ij} p_{ij} x_i x_j$. Then

$$\mathbb{E}^\mu\, p(x)^2 \geqslant \frac{1}{n^2} \sum_{ij} p_{ij}^2\,.$$

By our assumption on negative eigenvalues of A^π for all $\pi \in \mathcal{S}_4$, we know that $\mathcal{L}\, p(x)^2 \geqslant \frac{-\lambda^2}{n^2} \sum_{ij} p_{ij}^2$. So if we choose $\varepsilon \leqslant 1/\lambda^2$, the operator $\tilde{\mathbb{E}}^0 = \tilde{\mathbb{E}}^\mu + \mathcal{L}/\lambda^2$ will be a valid pseudo-expectation. Moreover $\tilde{\mathbb{E}}^0$ is well correlated with A, since it was obtained by maximizing the amount of $\mathcal{L}$, which is simply the (maximally-symmetric) dual of A. However the calculation of $\tilde{\mathbb{E}}^0\, \|x\|_2^4$ shows that this pseudo-expectation is not on the unit sphere, though it is close. Let c refer to

$$c := \tilde{\mathbb{E}}^0\, \|x\|_2^4 = \mathbb{E}^\mu\, \|x\|_2^4 + \frac{1}{\lambda^2} \mathcal{L}\, \|x\|_2^4 = 1 + \frac{1}{|\mathcal{S}_4| n^2 \lambda^2} \sum_{\pi \in \mathcal{S}_4} \langle \mathrm{Id}^{\mathrm{sym}}, A^\pi \rangle = 1 + O(n^{-1/2})\,.$$

We would like to use Lemma 41 together with $\tilde{\mathbb{E}}^0$ to obtain some $\mathcal{L}^1 : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$ with $\|x\|_2^2 - 1$ in its kernel and bounded $\lambda_{\min} \mathcal{L}^1$ while still maintaining a high correlation with A. For this we need ξ_1, ξ_2, ξ_2' so that

- $\left| \frac{1}{c} \tilde{\mathbb{E}}^0\, \|x\|_2^2 x_i - \tilde{\mathbb{E}}^0\, x_i \right| \leqslant \xi_1$ for all i.
- $\left| \frac{1}{c} \tilde{\mathbb{E}}^0\, \|x\|_2^2 x_i x_j - \tilde{\mathbb{E}}^0\, x_i x_j \right| \leqslant \xi_2$ for all $i \neq j$.
- $\left| \frac{1}{c} \tilde{\mathbb{E}}^0\, \|x\|_2^2 x_i^2 - \tilde{\mathbb{E}}^0\, x_i^2 \right| \leqslant \xi_2'$ for all i.

Since $\tilde{\mathbb{E}}^0\, p(x) = 0$ for all homogeneous odd-degree p, we may take $\xi_1 = 0$. For ξ_2, we have that when $i \neq j$,

$$\left| \tfrac{1}{c} \tilde{\mathbb{E}}^0\, \|x\|_2^2 x_i x_j - \tilde{\mathbb{E}}^0\, x_i x_j \right| = \left| \frac{1}{c\lambda^2} \mathcal{L}\, \|x\|_2^2 x_i x_j \right| \leqslant \delta_2\,,$$

where we recall δ_2 and δ_2' defined in the theorem statement. Finally, for ξ_2', we have

$$\left| \tfrac{1}{c} \tilde{\mathbb{E}}^0\, \|x\|_2^2 x_i^2 - \tilde{\mathbb{E}}^0\, x_i^2 \right| \leqslant \left| \frac{1}{c\lambda^2} \mathcal{L}\, \|x\|_2^2 x_i^2 \right| + \left| \tfrac{1}{c} \mathbb{E}^\mu\, \|x\|_2^2 x_i^2 - \mathbb{E}^\mu\, x_i^2 \right| \leqslant \delta_2' + \tfrac{c-1}{cn}\,.$$

Thus, Lemma 41 yields $\mathcal{L}^1 : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$ with $\|x\|_2^2 - 1$ in its kernel in the sense that $\mathcal{L}^1\, p(x)(\|x\|_2^2 - 1) = 0$ for all $p \in \mathbb{R}[x]_{\leqslant 2}$. If we take $\xi_2 = \delta_2$ and $\xi_2' = \delta_2' + \frac{c-1}{cn}$, then $\lambda_{\min} \mathcal{L}^1 \geqslant -\frac{c-1}{c} - n^2 \delta_2 - n^{3/2}(\delta_2' + \frac{c-1}{cn}) = -O(1)$. Furthermore, $\mathcal{L}^1\, \mathbf{A}(x) = \frac{1}{c\lambda^2} \mathcal{L}\, \mathbf{A}(x) = \Theta(\frac{1}{\lambda^2} \mathcal{L}\, \mathbf{A}(x))$.

So by Lemma 42, there is a degree-4 pseudo-expectation $\tilde{\mathbb{E}}$ satisfying $\{\|x\|_2^2 = 1\}$ so that

$$\tilde{\mathbb{E}}\, \mathbf{A}(x) = \Theta\left(\frac{1}{\lambda^2} \mathcal{L}\, \mathbf{A}(x) \right) + \Theta(\mathbb{E}^\mu\, \mathbf{A}(x))$$

$$
\begin{aligned}
&= \Theta\left(\frac{1}{|\mathcal{S}_4|n^2\lambda^2}\langle A, \sum_{\pi\in\mathcal{S}_4} A^\pi\rangle\right) + \Theta(\mathbb{E}^\mu \mathbf{A}(x)) \\
&\geqslant \Omega\left(\frac{n^2}{\lambda^2}\right) + \Theta(\mathbb{E}^\mu \mathbf{A}(x)) .
\end{aligned}
$$

■

B.7.2. THE DEGREE-3 LOWER BOUND

Now we turn to the proof of Theorem 32.

Proof [Proof of Theorem 32] Let $\mathbf{A}$ be a 3-tensor. Let $\varepsilon \geqslant 0$ be a parameter to be chosen later. We begin with the following linear functional $\mathcal{L} : \mathbb{R}[x]_{\leqslant 3} \to \mathbb{R}$. For any monomial x^α (where α is a multi-index of degree at most 3),

$$
\mathcal{L}\, x^\alpha \stackrel{\text{def}}{=} \begin{cases} \mathbb{E}^\mu\, x^\alpha & \text{if } \deg x^\alpha \leqslant 2 \\ \frac{\varepsilon}{n^{3/2}} \sum_{\pi\in\mathcal{S}_3} \mathbf{A}^\pi_\alpha & \text{if } \deg x^\alpha = 3 \end{cases} .
$$

The functional $\mathcal{L}$ contains our current best guess at the degree 1 and 2 moments of a pseudo-distribution whose degree-3 moments are ε-correlated with $\mathbf{A}(x)$.

The next step is to use symmetric Schur complement to extend $\mathcal{L}$ to a degree-4 pseudo-expectation. Note that $M_{\mathcal{L}|_3}$ decomposes as

$$
M_{\mathcal{L}|_3} = \sum_{\pi\in\mathcal{S}_3} \Pi A^\pi
$$

where, as a reminder, A^π is the $n^2{\times}n$ flattening of $\mathbf{A}^\pi$ and Π is the projector to the canonical embedding of $\mathbb{R}[x]_2$ into $\mathbb{R}^{n^2}$. So, using Lemma 39, we want to find the symmetric Schur complements of the following family of matrices (with notation matching the statement of Lemma 39):

$$
\left\{ \begin{pmatrix} 1 & M_{\mathcal{L}|_1}^T & V_{\mathcal{L}|_2}^T \\ M_{\mathcal{L}|_1} & M_{\mathcal{L}|_2} & \frac{\varepsilon}{n^{3/2}}(\Pi A^\pi)^T \\ V_{\mathcal{L}|_2} & \frac{\varepsilon}{n^{3/2}}\Pi A^\pi & \bullet \end{pmatrix} \right\}_{\pi\in\mathcal{S}_3} .
$$

Since we have the same assumptions on A^π for all $\pi \in \mathcal{S}_3$, without loss of generality we analyze just the case that π is the identity permutation, in which case $A^\pi = A$.

Since $\mathcal{L}$ matches the degree-one and degree-two moments of the uniform distribution on the unit sphere, we have $M_{\mathcal{L}|_1} = \mathbf{0}$, the n-dimensional zero vector, and $M_{\mathcal{L}|_2} = \frac{1}{n}\,\mathrm{Id}_{n\times n}$. Let $w \in \mathbb{R}^{n^2}$ be the n^2-dimensional vector flattening of $\mathrm{Id}_{n\times n}$. We observe that $ww^T = \sigma \cdot \mathrm{Id}$ is one of the permutations of $\mathrm{Id}_{n^2\times n^2}$. Taking B and C as follows,

$$
B = \begin{pmatrix} 1 & \mathbf{0} \\ \mathbf{0} & \frac{1}{n}\,\mathrm{Id}_{n\times n} \end{pmatrix} \qquad C = \begin{pmatrix} w & \frac{\varepsilon}{n^{3/2}} A \end{pmatrix} ,
$$

we compute that

$$
CB^{-1}C^T = \frac{1}{n^2}(\sigma \cdot \mathrm{Id}) + \frac{\varepsilon^2}{n^2}\Pi A A^T \Pi .
$$

Symmetrizing the Id portion and the AA^T portion of this matrix separately, we see that the symmetric Schur complement that we are looking for is the spectrally-least $M \in \mathrm{Sym}\left(\frac{1}{n^2}(\sigma \cdot \mathrm{Id}) + \frac{\varepsilon^2}{n^2}AA^T\right)$ so that

$$\begin{aligned} M &= \frac{t}{n^2}\left[3\,\mathrm{Id}^{\mathrm{sym}} + \frac{\varepsilon^2}{2}\left(\Pi(AA^T + \sigma \cdot AA^T + \sigma^2 \cdot AA^T)\Pi + \Pi(AA^T + \sigma \cdot AA^T + \sigma^2 \cdot AA^T)^T\Pi\right)\right] \\ &\succeq \frac{1}{n^2}(\sigma \cdot \mathrm{Id}) + \frac{\varepsilon^2}{n^2}\Pi AA^T \Pi\,. \end{aligned}$$

Here we have used Corollary 35 and Corollary 34 to express a general element of $\mathrm{Sym}(\frac{\varepsilon^2}{n^2}\Pi AA^T\Pi)$ in terms of $\Pi, AA^T, \sigma \cdot AA^T$, and $\sigma^2 \cdot AA^T$.

Any spectrally small M satisfying the above suffices for us. Taking $t = 1$, canceling some terms, and making the substitution $3\,\mathrm{Id}^{\mathrm{sym}} - \sigma \cdot \mathrm{Id} = 2\Pi\,\mathrm{Id}\,\Pi$, we see that it is enough to have

$$-2\,\Pi\,\mathrm{Id}\,\Pi \preceq \frac{\varepsilon^2}{2}\Pi(\sigma \cdot AA^T + \sigma^2 \cdot AA^T)\Pi + \frac{\varepsilon^2}{2}\Pi(\sigma \cdot AA^T + \sigma^2 \cdot AA^T)^T\Pi\,,$$

which by the premises of the theorem holds for $\varepsilon = 1/\lambda$. Pushing through our symmetrized Schur complement rule with our decomposition of $M_{\mathcal{L}|_3}$ (Lemma 39), this ε yields a valid degree-4 pseudo-expectation $\tilde{\mathbb{E}}^0 : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$. From our choice of parameters, we see that $\tilde{\mathbb{E}}^0|_4$, the degree-4 part of $\tilde{\mathbb{E}}^0$, is given by $\tilde{\mathbb{E}}^0|_4 = \frac{n^2+2n}{n^2}\mathbb{E}^\mu + \mathcal{L}$, where $\mathcal{L} : \mathbb{R}[x]_4 \to \mathbb{R}$ is as defined in the theorem statement. Furthermore, $\tilde{\mathbb{E}}^0\, p(x) = \mathbb{E}^\mu\, p(x)$ for p with $\deg p \leqslant 2$.

We would like to know how big $\tilde{\mathbb{E}}^0 \|x\|_2^4$ is. We have

$$c := \tilde{\mathbb{E}}^0 \|x\|_2^4 = \left(1 + \frac{1}{n}\right)\mathbb{E}^\mu \|x\|_2^4 + \mathcal{L}\|x\|_2^4 = 1 + \frac{1}{n} + \mathcal{L}\|x\|_2^4\,.$$

We have assumed that $\langle \mathrm{Id}^{\mathrm{sym}}, AA^T\rangle \leqslant O(\lambda^2 n^2)$. Since $\mathrm{Id}^{\mathrm{sym}}$ is maximally symmetric, we have $\langle \mathrm{Id}^{\mathrm{sym}}, \sum_{\pi \in \mathcal{S}_4} \pi \cdot AA^T\rangle = \langle \mathrm{Id}^{\mathrm{sym}}, |\mathcal{S}_4|AA^T\rangle$ and so

$$\mathcal{L}\|x\|_2^4 = \frac{1}{\lambda^2 n^2}\langle \mathrm{Id}^{\mathrm{sym}}, M_{\mathcal{L}|_4}\rangle = \frac{1}{n^2\lambda^2}\Theta(\langle \mathrm{Id}^{\mathrm{sym}}, \sum_{\pi \in \mathcal{S}_4} \pi \cdot AA^T\rangle) \leqslant O(1)\,.$$

Finally, our assumptions on $\langle A, \sum_{\pi \in \mathcal{S}_3} A^\pi\rangle$ yield

$$\tilde{\mathbb{E}}^0\, \mathbf{A}(x) = \frac{\varepsilon}{n^{3/2}}\langle A, \sum_{\pi \in \mathcal{S}_3} A^\pi\rangle \geqslant \Omega\left(\frac{n^{3/2}}{\lambda}\right)\,.$$

We have established the following lemma.

Lemma 43 *Under the assumptions of Theorem 32 there is a degree-*4 *pseudo-expectation operator* $\tilde{\mathbb{E}}^0$ *so that*

— $c := \tilde{\mathbb{E}}^0 \|x\|_2^4 = 1 + O(1)$.

— $\tilde{\mathbb{E}}^0\, \mathbf{A}(x) \geqslant \Omega(n^{3/2}/\lambda)$.

— $\tilde{\mathbb{E}}^0\, p(x) = \mathbb{E}^\mu\, p(x)$ *for all* $p \in \mathbb{R}[x]_{\leqslant 2}$.

— $\tilde{\mathbb{E}}^0|_4 = (1+\frac{1}{n})\,\mathbb{E}^\mu|_4 + \mathcal{L}$.

Now we would like feed $\tilde{\mathbb{E}}^0$ into Lemma 41 to get a linear functional $\mathcal{L}^1 : \mathbb{R}[x]_{\leqslant 4} \to \mathbb{R}$ with $\|x\|_2^2 - 1$ in its kernel (equivalently, which satisfies $\{\|x\|_2^4 - 1 = 0\}$), but in order to do that we need to find ξ_1, ξ_2, ξ_2' so that

— $\left|\frac{1}{c}\,\tilde{\mathbb{E}}^0\,\|x\|_2^2 x_i - \tilde{\mathbb{E}}^0\,x_i\right| \leqslant \xi_1$ for all i.

— $\left|\frac{1}{c}\,\tilde{\mathbb{E}}^0\,\|x\|_2^2 x_i x_j - \tilde{\mathbb{E}}^0\,x_i x_j\right| \leqslant \xi_2$ for all $i \neq j$.

— $\left|\frac{1}{c}\,\tilde{\mathbb{E}}^0\,\|x\|_2^2 x_i^2 - \tilde{\mathbb{E}}^0\,x_i^2\right| \leqslant \xi_2'$ for all i.

For ξ_1, we note that for every i, $\tilde{\mathbb{E}}^0\,x_i = 0$ since $\tilde{\mathbb{E}}^0$ matches the uniform distribution on degree one and two polynomials. Thus, $\left|\frac{1}{c}\,\tilde{\mathbb{E}}^0\,\|x\|_2^2 x_i - \tilde{\mathbb{E}}^0\,x_i\right| = \left|\frac{1}{c}\,\tilde{\mathbb{E}}^0\,\|x\|_2^2 x_i\right|$.

We know that $M_{\tilde{\mathbb{E}}^0|_3}$, the matrix representation of the degree-3 part of $\tilde{\mathbb{E}}^0$, is $\frac{1}{|\mathcal{S}_3|n^{3/2}\lambda}A$. Expanding $\tilde{\mathbb{E}}^0\,\|x\|_2^2 x_i$ with matrix representations, we get

$$\left|\tfrac{1}{c}\,\tilde{\mathbb{E}}^0\,\|x\|_2^2 x_i\right| = \frac{1}{|\mathcal{S}_3| c n^{3/2}\lambda}\left|\langle \mathrm{Id}_{n\times n}\,, \sum_{\pi\in\mathcal{S}_3} A_i\rangle\right| \leqslant \delta_1\,,$$

where δ_1 is as defined in the theorem statement.

Now for ξ_2 and ξ_2'. Let $\mathcal{L}$ be the operator in the theorem statement. By the definition of $\tilde{\mathbb{E}}^0$, we get

$$\tilde{\mathbb{E}}^0|_4 \leqslant \left[\left(1+\frac{1}{n}\right)\mathbb{E}^\mu|_4 + \mathcal{L}\right].$$

In particular, for $i \neq j$,

$$\left|\tfrac{1}{c}\,\tilde{\mathbb{E}}^0\,\|x\|_2^2 x_i x_j - \tilde{\mathbb{E}}^0\,x_i x_j\right| = \left|\tfrac{1}{c}\,\mathcal{L}\,\|x\|_2^2 x_i x_j\right| \leqslant \delta_2\,.$$

For $i = j$,

$$\begin{aligned}
\left|\tfrac{1}{c}\,\tilde{\mathbb{E}}^0\,\|x\|_2^2 x_i^2 - \tilde{\mathbb{E}}^\mu\,x_i^2\right| &= \tfrac{1}{c}\left|\mathcal{L}\,\|x\|_2^2 x_i^2 + \left(1+\frac{1}{n}\right)\mathbb{E}^\mu\,\|x\|_2^2 x_i^2 - c\,\mathbb{E}^\mu\,x_i^2\right| \\
&= \tfrac{1}{c}\left|\mathcal{L}\,\|x\|_2^2 x_i^2 - \tfrac{1}{n}\,\mathcal{L}\,\|x\|_2^4 + \tfrac{1}{n}\,\mathcal{L}\,\|x\|_2^4 + \frac{1}{n}\left(1+\frac{1}{n}\right)\mathbb{E}^\mu\,\|x\|_2^4 - c\,\mathbb{E}^\mu\,x_i^2\right| \\
&= \tfrac{1}{c}\left|\mathcal{L}\,\|x\|_2^2 x_i^2 - \tfrac{1}{n}\,\mathcal{L}\,\|x\|_2^4 + \tfrac{1}{n}\,\tilde{\mathbb{E}}^0\,\|x\|_2^4 - c\,\mathbb{E}^\mu\,x_i^2\right| \\
&= \tfrac{1}{c}\left|\mathcal{L}\,\|x\|_2^2 x_i^2 - \tfrac{1}{n}\,\mathcal{L}\,\|x\|_2^4\right| \\
&\leqslant \delta_2'\,.
\end{aligned}$$

Thus, we can take $\xi_1 = \delta_1$, $\xi_2 = \delta_2$, $\xi_2' = \delta_2'$, and $c = \tilde{\mathbb{E}}^0\,\|x\|_2^4 = 1 + O(1)$, and apply Lemma 41 to conclude that

$$\lambda_{\min}\,\mathcal{L}^1 \geqslant -\tfrac{c-1}{c} - O(n)\xi_1 - O(n^{3/2})\xi_2' - O(n^2)\xi_2 = -O(1)\,.$$

The functional $\mathcal{L}^1$ loses a constant factor in the value assigned to $\mathbf{A}(x)$ as compared to $\tilde{\mathbb{E}}^0$:

$$\mathcal{L}^1\, \mathbf{A}(x) = \frac{\tilde{\mathbb{E}}^0\, \mathbf{A}(x)}{c} \geqslant \Omega\left(\frac{n^{3/2}}{\lambda}\right).$$

Now using Lemma 42, we can correct the negative eigenvalue of $\mathcal{L}^1$ to get a pseudo-expectation

$$\tilde{\mathbb{E}} \overset{\text{def}}{=} \Theta(1)\, \mathcal{L}^1 + \Theta(1)\, \mathbb{E}^{\mu}\ .$$

By Lemma 42, the pseudo-expectation $\tilde{\mathbb{E}}$ satisfies $\{\|x\|_2^2 = 1\}$. Finally, to complete the proof, we have:

$$\tilde{\mathbb{E}}\, \mathbf{A}(x) = \Omega\left(\frac{n^{3/2}}{\lambda}\right) + \Theta(1)\, \mathbb{E}^{\mu}\, \mathbf{A}(x)\,.$$

■

Appendix C. Higher-Order Tensors

We have heretofore restricted ourselves to the case $k = 3$ in our algorithms for the sake of readability. In this section we state versions of our main results for general k and indicate how the proofs from the 3-tensor case may be generalized to handle arbitrary k. Our policy is to continue to treat k as constant with respect to n, hiding multiplicative losses in k in our asymptotic notation.

The case of general odd k may be reduced to $k = 3$ by a standard trick, which we describe here for completeness. Given $\mathbf{A}$ an order-k tensor, consider the polynomial $\mathbf{A}(x)$ and make the variable substitution $y_\beta = x^\beta$ for each multi-index β with $|\beta| = (k+1)/2$. This yields a degree-3 polynomial $\mathbf{A}'(x, y)$ to which the analysis in Section 3 and Section 4 applies almost unchanged, now using pseudo-distributions $\{x, y\}$ satisfying $\{\|x\|^2 = 1, \|y\|^2 = 1\}$. In the analysis of tensor PCA, this change of variables should be conducted after the input is split into signal and noise parts, in order to preserve the analysis of the second half of the rounding argument (to get from $\tilde{\mathbb{E}}\langle v_0, x\rangle^k$ to $\tilde{\mathbb{E}}\langle v_0, x\rangle$), which then requires only syntactic modifications to Lemma 50. The only other non-syntactic difference is the need to generalize the λ-boundedness results for random polynomials to handle tensors whose dimensions are not all equal; this is already done in Theorem 56.

For even k, the degree-k SoS approach does not improve on the tensor unfolding algorithms of Montanari and Richard (Montanari and Richard, 2014). Indeed, by performing a similar variable substitution, $y_\beta = x^\beta$ for all $|\beta| = k/2$, the SoS algorithm reduces exactly to the eigenvalue/eigenvector computation from tensor unfolding. If we perform instead the substitution $y_\beta = x^\beta$ for $|\beta| = k/2 - 1$, it becomes possible to extract v_0 directly from the degree-2 pseudo-moments of an (approximately) optimal degree-4 pseudo-distribution, rather than performing an extra step to recover v_0 from v well-correlated with $v_0^{\otimes k/2}$. Either approach recovers v_0 only up to sign, since the input is unchanged under the transformation $v_0 \mapsto -v_0$.

We now state analogues of all our results for general k. Except for the above noted differences from the $k = 3$ case, the proofs are all easy transformations of the proofs of their degree-3 counterparts.

Theorem 44 *Let k be an odd integer, $v_0 \in \mathbb{R}^n$ a unit vector, $\tau \gtrsim n^{k/4} \log(n)^{1/4}/\varepsilon$, and $\mathbf{A}$ an order-k tensor with independent unit Gaussian entries.*

1. *There is an algorithm, based on semidefinite programming, which on input* $\mathbf{T}(x) = \tau \cdot \langle v_0, x\rangle^k + \mathbf{A}(x)$ *returns a unit vector* v *with* $\langle v_0, v\rangle \geqslant 1 - \varepsilon$ *with high probability over random choice of* $\mathbf{A}$.
2. *There is an algorithm, based on semidefinite programming, which on input* $\mathbf{T}(x) = \tau \cdot \langle v_0, x\rangle^k + \mathbf{A}(x)$ *certifies that* $\mathbf{T}(x) \leqslant \tau \cdot \langle v, x\rangle^k + O(n^{k/4} \log(n)^{1/4})$ *for some unit* v *with high probability over random choice of* $\mathbf{A}$. *This guarantees in particular that* v *is close to a maximum likelihood estimator for the problem of recovering the signal* v_0 *from the input* $\tau \cdot v_0^{\otimes k} + \mathbf{A}$.
3. *By solving the semidefinite relaxation approximately, both algorithms can be implemented in time* $\tilde{O}(m^{1+1/k})$, *where* $m = n^k$ *is the input size.*

For even k, *the above all hold, except now we recover* v *with* $\langle v_0, v\rangle^2 \geqslant 1 - \varepsilon$, *and the algorithms can be implemented in nearly-linear time.*

The next theorem partially resolves a conjecture of Montanari and Richard regarding tensor unfolding algorithms for odd k. We are able to prove their conjectured signal-to-noise ratio τ, but under an asymmetric noise model. They conjecture that the following holds when $\mathbf{A}$ is symmetric with unit Gaussian entries.

Theorem 45 *Let* k *be an odd integer,* $v_0 \in \mathbb{R}^n$ *a unit vector,* $\tau \gtrsim n^{k/4}/\varepsilon$, *and* $\mathbf{A}$ *an order-*k *tensor with independent unit Gaussian entries. There is a nearly-linear-time algorithm, based on tensor unfolding, which, with high probability over random choice of* $\mathbf{A}$, *recovers a vector* v *with* $\langle v, v_0\rangle^2 \geqslant 1 - \varepsilon$.

Acknowledgments

We thank Moses Charikar for bringing to our attention the work of Montanari and Richard. We would like to thank Boaz Barak, Rong Ge, and Ankur Moitra for enlightening conversations. The first-named author acknowledges the support of an NSF Graduate Research Fellowship under award no. 1144153. A large portion of this work was completed while the authors were long-term visitors to the Simons Institute for the Theory of Computing (Berkeley) for the program on Algorithmic Spectral Graph Theory.

Appendix D. Pseudo-Distribution Facts

Lemma 46 (Quadric Sampling) *Let* $\{x\}$ *be a pseudo-distribution over* $\mathbb{R}^n$ *of degree* $d \geqslant 2$. *Then there is an actual distribution* $\{y\}$ *over* $\mathbb{R}^n$ *so that for any polynomial* p *of degree at most* 2, $\mathbb{E}[p(y)] = \tilde{\mathbb{E}}[p(x)]$. *Furthermore,* $\{y\}$ *can be sampled from in time* poly n.

Lemma 47 (Pseudo-Cauchy-Schwarz, Function Version, (Barak et al., 2012)) *Let* x, y *be vector-valued polynomials. Then*

$$\langle x, y\rangle \preceq \frac{1}{2}(\|x\|^2 + \|y\|^2).$$

See (Barak et al., 2014a) for the cleanest proof.

Lemma 48 (Pseudo-Cauchy-Schwarz, Powered Function Version) *Let* x, y *be vector-valued polynomials and* $d > 0$ *an integer. Then*

$$\langle x, y\rangle^d \preceq \frac{1}{2}(\|x\|^{2d} + \|y\|^{2d}).$$

Proof Note that $\langle x, y\rangle^d = \langle x^{\otimes d}, y^{\otimes d}\rangle$ and apply Lemma 47. ■

Yet another version of pseudo-Cauchy-Schwarz will be useful:

Lemma 49 (Pseudo-Cauchy-Schwarz, Multiplicative Function Version, (Barak et al., 2012)) *Let $\{x, y\}$ be a degree d pseudo-distribution over a pair of vectors, $d \geqslant 2$. Then*

$$\tilde{\mathbb{E}}[\langle x, y\rangle] \leqslant \sqrt{\tilde{\mathbb{E}}[\|x\|^2]}\sqrt{\tilde{\mathbb{E}}[\|y\|^2]}.$$

Again, see (Barak et al., 2014a) for the cleanest proof.

We will need the following inequality relating $\tilde{\mathbb{E}}\langle x, v_0\rangle^3$ and $\tilde{\mathbb{E}}\langle x, v_0\rangle$ when $\tilde{\mathbb{E}}\langle x, v_0\rangle^3$ is large.

Lemma 50 *Let $\{x\}$ be a degree-4 pseudo-distribution satisfying $\{\|x\|^2 = 1\}$, and let $v_0 \in \mathbb{R}^n$ be a unit vector. Suppose that $\tilde{\mathbb{E}}\langle x, v_0\rangle^3 \geqslant 1 - \varepsilon$ for some $\varepsilon \geqslant 0$. Then $\tilde{\mathbb{E}}\langle x, v_0\rangle \geqslant 1 - 2\varepsilon$.*

Proof Let $p(u)$ be the univariate polynomial $p(u) = 1 - 2u^3 + u$. It is easy to check that $p(u) \geqslant 0$ for $u \in [-1, 1]$. It follows from classical results about univariate polynomials that $p(u)$ then can be written as

$$p(u) = s_0(u) + s_1(u)(1 + u) + s_2(u)(1 - u)$$

for some SoS polynomials s_0, s_1, s_2 of degrees at most 2. (See (O'Donnell and Zhou, 2013), fact 3.2 for a precise statement and attributions.)

Now we consider

$$\tilde{\mathbb{E}}\, p(\langle x, v_0\rangle) \geqslant \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)(1 + \langle x, v_0\rangle)] + \tilde{\mathbb{E}}[s_2(\langle x, v_0\rangle)(1 - \langle x, v_0\rangle)]\,.$$

We have by Lemma 47 that $\langle x, v_0\rangle \preceq \frac{1}{2}(\|x\|^2 + 1)$ and also that $\langle x, v_0\rangle \succeq -\frac{1}{2}(\|x\|^2 + 1)$. Multiplying the latter SoS relation by the SoS polynomial $s_1(\langle x, v_0\rangle)$ and the former by $s_2(\langle x, v_0\rangle)$, we get that

$$\begin{aligned}\tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)(1 + \langle x, v_0\rangle)] &= \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)] + \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)\langle x, v_0\rangle] \\ &\geqslant \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)] - \frac{1}{2}\tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)(\|x\|^2 + 1)] \\ &\geqslant \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)] - \tilde{\mathbb{E}}[s_1(\langle x, v_0\rangle)] \\ &\geqslant 0\,,\end{aligned}$$

where in the second-to-last step we have used the assumption that $\{x\}$ satisfies $\{\|x\|^2 = 1\}$. A similar analysis yields

$$\tilde{\mathbb{E}}[s_2(\langle x, v_0\rangle)(1 - \langle x, v_0\rangle)] \geqslant 0\,.$$

All together, this means that $\tilde{\mathbb{E}}\, p(\langle x, v_0\rangle) \geqslant 0$. Expanding, we get $\tilde{\mathbb{E}}[1 - 2\langle x, v_0\rangle^3 + \langle x, v_0\rangle] \geqslant 0$. Rearranging yields

$$\tilde{\mathbb{E}}\langle x, v_0\rangle \geqslant 2\,\tilde{\mathbb{E}}\langle x, v_0\rangle^3 - 1 \geqslant 2(1 - \varepsilon) - 1 \geqslant 1 - 2\varepsilon\,.$$

■

We will need a bound on the pseudo-expectation of a degree-3 polynomial in terms of the operator norm of its coefficient matrix.

Lemma 51 *Let $\{x\}$ be a degree-4 pseudo-distribution. Let $M \in \mathbb{R}^{n^2 \times n}$. Then $\tilde{\mathbb{E}}\langle x^{\otimes 2}, Mx\rangle \leqslant \|M\|(\tilde{\mathbb{E}}\,\|x\|^4)^{3/4}$.*

Proof We begin by expanding in the monomial basis and using pseudo-Cauchy-Schwarz:

$$\begin{aligned}
\tilde{\mathbb{E}}\langle x^{\otimes 2}, Mx\rangle &= \tilde{\mathbb{E}} \sum_{ijk} M_{(j,k),i} x_i x_j x_k \\
&= \tilde{\mathbb{E}} \sum_{i} x_i \sum_{jk} M_{(j,k),i} x_j x_k \\
&\leqslant (\tilde{\mathbb{E}}\,\|x\|^2)^{1/2} \left[\tilde{\mathbb{E}} \sum_i \left(\sum_{jk} M_{(j,k),i} x_i x_j \right)^2 \right]^{1/2} \\
&\leqslant (\tilde{\mathbb{E}}\,\|x\|^4)^{1/4} \left[\tilde{\mathbb{E}} \sum_i \left(\sum_{jk} M_{(j,k),i} x_i x_j \right)^2 \right]^{1/2}
\end{aligned}$$

We observe that MM^T is a matrix representation of $\sum_i \left(\sum_{jk} M_{(j,k),i} x_i x_j\right)^2$. We know $MM^T \preceq \|M\|^2\,\mathrm{Id}$, so

$$\tilde{\mathbb{E}} \sum_i \left(\sum_{jk} M_{(j,k),i} x_i x_j \right)^2 \leqslant \|M\|^2\,\tilde{\mathbb{E}}\,\|x\|^4 .$$

Putting it together, we get $\tilde{\mathbb{E}}\langle x^{\otimes 2}, Mx\rangle \leqslant \|M\|(\tilde{\mathbb{E}}\,\|x\|^4)^{3/4}$ as desired. ■

Appendix E. Concentration bounds

E.1. Elementary Random Matrix Review

We will be extensively concerned with various real random matrices. A great deal is known about natural classes of such matrices; see the excellent book of Tao (Tao, 2012) and the notes by Vershynin and Tropp (Vershynin, 2011; Tropp, 2012).

Our presentation here follows Vershynin's (Vershynin, 2011). Let X be a real random variable. The subgaussian norm $\|X\|_{\psi_2}$ of X is $\sup_{p \geqslant 1} p^{-1/2}(\mathbb{E}|X|^p)^{1/p}$. Let $\{a\}$ be a distribution on $\mathbb{R}^n$. The subgaussian norm $\|a\|_{\psi_2}$ of $\{a\}$ is the maximal subgaussian norm of the one-dimensional marginals: $\|a\|_{\psi_2} = \sup_{\|u\|=1} \|\langle a, u\rangle\|_{\psi_2}$. A family of random variables $\{X_n\}_{n \in \mathbb{N}}$ is subgaussian if $\|X_n\|_{\psi_2} = O(1)$. The reader may easily check that an n-dimensional vector of independent standard Gaussians or independent ± 1 variables is subgaussian.

It will be convenient to use the following standard result on the concentration of empirical covariance matrices. This statement is borrowed from (Vershynin, 2011), Corollary 5.50.

Lemma 52 *Consider a sub-gaussian distribution $\{a\}$ in $\mathbb{R}^m$ with covariance matrix Σ, and let $\delta \in (0,1), t \geqslant 1$. If $a_1, \ldots, a_N \sim \{a\}$ with $N \geqslant C(t/\delta)^2 m$ then $\|\frac{1}{N}\sum a_i a_i^T - \Sigma\| \leqslant \delta$ with probability at least $1 - 2\exp(-t^2 m)$. Here $C = C(K)$ depends only on the sub-gaussian norm $K = \|a\|_{\psi_2}$ of a random vector taken from this distribution.*

We will also need the matrix Bernstein inequality. This statement is borrowed from Theorem 1.6.2 of Tropp (Tropp, 2012).

Theorem 53 (Matrix Bernstein) *Let $S_1, \ldots, S_m$ be independent square random matrices with dimension n. Assume that each matrix has bounded deviation from its mean: $\|S_i - \mathbb{E}\, S_i\| \leqslant R$ for all i. Form the sum $Z = \sum_i S_i$ and introduce a variance parameter*

$$\sigma^2 = \max\{\|\mathbb{E}(Z - \mathbb{E}\, Z)(Z - \mathbb{E}\, Z)^T\|, \|\mathbb{E}(Z - \mathbb{E}\, Z)^T(Z - \mathbb{E}\, Z)\|\}\,.$$

Then

$$\mathbb{P}\{\|Z - \mathbb{E}\, Z\| \geqslant t\} \leqslant 2n \exp\left(\frac{t^2/2}{\sigma^2 + Rt/3}\right) \qquad \textit{for all } t \geqslant 0\,.$$

We will need bounds on the operator norm of random square rectangular matrices, both of which are special cases of Theorem 5.39 in (Vershynin, 2011).

Lemma 54 *Let A be an $n \times n$ matrix with independent entries from $\mathcal{N}(0,1)$. Then with probability $1 - n^{-\omega(1)}$, the operator norm $\|A\|$ satisfies $\|A\| \leqslant O(\sqrt{n})$.*

Lemma 55 *Let A be an $n^2 \times n$ matrix with independent entries from $\mathcal{N}(0,1)$. Then with probability $1 - n^{-\omega(1)}$, the operator norm $\|A\|$ satisfies $\|A\| \leqslant O(n)$.*

E.2. Concentration for $\sum_i A_i \otimes A_i$ and Related Ensembles

Our first concentration theorem provides control over the nontrivial permutations of the matrix AA^T under the action of $\mathcal{S}_4$ for a tensor **A** with independent entries.

Theorem 56 *Let $c \in \{1,2\}$ and $d \geqslant 1$ an integer. Let $A_1, \ldots, A_{n^c}$ be iid random matrices in $\{\pm 1\}^{n^d \times n^d}$ or with independent entries from $\mathcal{N}(0,1)$. Then, with probability $1 - O(n^{-100})$,*

$$\left\| \sum_{i \in [n^c]} A_i \otimes A_i - \mathbb{E}\, A_i \otimes A_i \right\| \precsim \sqrt{d} n^{(2d+c)/2} \cdot (\log n)^{1/2}\,.$$

and

$$\left\| \sum_{i \in [n^c]} A_i \otimes A_i^T - \mathbb{E}\, A_i \otimes A_i^T \right\| \precsim \sqrt{d} n^{(2d+c)/2} \cdot (\log n)^{1/2}\,.$$

We can prove Theorem 11 as a corollary of the above.

Proof [Proof of Theorem 11] Let A have iid Gaussian entries. We claim that $\mathbb{E}\, A \otimes A$ is a matrix representation of $\|x\|^4$. To see this, we compute

$$\begin{aligned} \langle x^{\otimes 2}, \mathbb{E}(A \otimes A) x^{\otimes 2}\rangle &= \mathbb{E}\langle x, Ax\rangle^2 \\ &= \sum_{i,j,k,l} \mathbb{E}\, A_{ij} A_{kl} x_i x_j x_k x_l \\ &= \sum_{ij} x_i^2 x_j^2 \\ &= \|x\|^4\,. \end{aligned}$$

Now by Theorem 56, we know that for A_i the slices of the tensor $\mathbf{A}$ from the statement of Theorem 11,

$$\sum_i A_i \otimes A_i \preceq n \mathbb{E} A \otimes A + \lambda^2 \cdot \mathrm{Id}$$

for $\lambda = O(n^{3/4} \log(n)^{1/4})$. Since $n = O(\lambda)$ and both Id and $\mathbb{E} A \otimes A$ are matrix representations of $\|x\|^4$, we are done. ■

Now we prove Theorem 56. We will prove only the statement about $\sum_i A_i \otimes A_i$, as the case of $\sum_i A_i \otimes A_i^T$ is similar.

Let $A_1, \ldots, A_{n^c}$ be as in Theorem 56. We first need to get a handle on their norms individually, for which we need the following lemma.

Lemma 57 *Let A be a random matrix in $\{\pm 1\}^{n^d \times n^d}$ or with independent entries from $\mathcal{N}(0,1)$. For all $t \geqslant 1$, the probability of the event $\{\|A\| > tn^{d/2}\}$ is at most $2^{-t^2 n^d/K}$ for some absolute constant K.*

Proof The subgaussian norm of the rows of A is constant and they are identically and isotropically distributed. Hence Theorem 5.39 of (Vershynin, 2011) applies to give the result. ■

Since the norms of the matrices $A_1, \ldots, A_{n^c}$ are concentrated around $n^{d/2}$ (by Lemma 57), it will be enough to prove Theorem 56 after truncating the matrices $A_1, \ldots, A_{n^c}$. For $t \geqslant 1$, define iid random matrices $A'_1, \ldots, A'_{n^c}$ such that

$$A'_i \stackrel{\text{def}}{=} \begin{cases} A_i & \text{if } \|A_i\| \leqslant tn^{d/2}, \\ 0 & \text{otherwise} \end{cases}$$

for some t to be chosen later. Lemma 57 allows us to show that the random matrices $A_i \otimes A_i$ and $A'_i \otimes A'_i$ have almost the same expectation. For the remainder of this section, let K be the absolute constant from Lemma 57.

Lemma 58 *For every $i \in [n^c]$ and all $t \geqslant 1$, the expectations of $A_i \otimes A_i$ and $A'_i \otimes A'_i$ satisfy*

$$\left\| \mathbb{E}[A_i \otimes A_i] - \mathbb{E}[A'_i \otimes A'_i] \right\| \leqslant O(1) \cdot 2^{-tn^d/K} \,.$$

Proof Using Jensen's inequality and that $A_i = A'_i$ unless $\|A_i\| > tn^{d/2}$, we have

$$\begin{aligned}
\| \mathbb{E} A_i \otimes A_i - A'_i \otimes A'_i \| &\leqslant \mathbb{E} \|A_i \otimes A_i - A'_i \otimes A'_i\| \quad \text{Jensen's inequality} \\
&= \int_{tn^{d/2}}^{\infty} \mathbb{P}(\|A_i\| \geqslant \sqrt{s})\, ds \quad \text{since } A_i = A'_i \text{ unless } \|A_i\| \geqslant tn^{d/2} \\
&\leqslant \int_{tn^{d/2}}^{\infty} 2^{-s/K}\, ds \quad \text{by Lemma 57} \\
&\leqslant \sum_{i=0}^{\infty} 2^{-tn^{d/2}/K} \cdot 2^{-i/K} \quad \text{discretizing the intergral} \\
&= O(2^{-tn^{d/2}/K}) \quad \text{as desired.}
\end{aligned}$$

■

Lemma 59 *Let $B'_1, \dots, B'_{n^c}$ be i.i.d. matrices such that $B'_i = A'_i \otimes A'_i - \mathbb{E}[A'_i \otimes A'_i]$. Then for every $C \geqslant 1$ with $C \leqslant 3t^2 n^{c/2}$,*

$$\mathbb{P}\left\{\left\|\sum_{i\in[n^c]} B'_i\right\| > C \cdot n^{(2d+c)/2}\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-C^2}{6t^4}\right).$$

Proof For $R = 2t^2 n^d$, the random matrices $B'_1, \dots, B'_{n^c}$ satisfy $\{\|B'_i\| \leqslant R\}$ with probability 1. Therefore, by the Bernstein bound for non-symmetric matrices (Tropp, 2012, Theorem 1.6),

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{n^c} B'_i\right\| \geqslant s\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-s^2/2}{\sigma^2 + Rs/3}\right),$$

where $\sigma^2 = \max\{\|\sum_i \mathbb{E}\, B'_i(B'_i)^\top\|, \|\sum_i \mathbb{E}(B'_i)^\top B'_i\|\} \leqslant n^c \cdot R^2$. For $s = C \cdot n^{(2d+c)/2}$, the probability is bounded by

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{n} B'_i\right\| \geqslant s\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-C^2 \cdot n^{(2d+c)}/2}{4t^4 \cdot n^{2d+c} + 2t^2 C \cdot n^{(4d+c)/2}/3}\right).$$

Since our parameters satisfy $t^2 C \cdot n^{(4d+c)/2}/3 \leqslant t^4 n^{(2d+c)}$, this probability is bounded by

$$\mathbb{P}\left\{\left\|\sum_{i=1}^{n} B'_i\right\| \geqslant s\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-C^2}{6t^4}\right).$$

■

At this point, we have all components of the proof of Theorem 56.

Proof [Proof of Theorem 56 for $\sum_i A_i \otimes A_i$ (other case is similar)] By Lemma 59,

$$\mathbb{P}\left\{\left\|\sum_i A'_i \otimes A'_i - \sum_i \mathbb{E}[A'_i \otimes A'_i]\right\| > C \cdot n^{(2d+c)/2}\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-C^2}{Kt^4}\right).$$

At the same time, by Lemma 57 and a union bound,

$$\mathbb{P}\left\{A_1 = A'_1, \dots, A_n = A'_{n^c}\right\} \geqslant 1 - n^c \cdot 2^{-t^2 n^d/K}.$$

By Lemma 58 and triangle inequality,

$$\left\|\sum_i \mathbb{E}[A_i \otimes A_i] - \sum_i \mathbb{E}[A'_i \otimes A'_i]\right\| \leqslant n^c \cdot 2^{-tn^d/K}.$$

Together, these bounds imply

$$\mathbb{P}\left\{\left\|\sum_i A_i \otimes A_i - \sum_i \mathbb{E}[A_i \otimes A_i]\right\| > C \cdot n^{(2d+c)/2} + n^c \cdot 2^{-tn^d/K}\right\} \leqslant 2n^{2d} \cdot \exp\left(\frac{-C^2}{Kt^4}\right) + n^c \cdot 2^{-t^2 n^d/K}.$$

We choose $t = 1$ and $C = 100\sqrt{2Kd \log n}$ and assume that n is large enough so that $C \cdot n^{(2d+c)/2} \geqslant n^c \cdot 2^{-tn^d/K}$ and $2n^{2d} \cdot \exp\left(\frac{-C^2}{Kt^4}\right) \geqslant n^c \cdot 2^{-t^2 n^d/K}$. Then the probability satisfies

$$\mathbb{P}\left\{\left\|\sum_i A_i \otimes A_i - \sum_i \mathbb{E}[A_i \otimes A_i]\right\| > 20n^{(2d+c)/2}\sqrt{2Kd \log n}\right\} \leqslant 4n^{-100}.$$

■

E.3. Concentration for Spectral SoS Analyses

Lemma 60 (Restatement of Lemma 18) *Let* $\mathbf{T} = \tau \cdot v_0^{\otimes 3} + \mathbf{A}$. *Suppose* $\mathbf{A}$ *has independent entries from* $\mathcal{N}(0,1)$. *Then with probability* $1 - O(n^{-100})$ *we have* $\| \sum_i A_i \otimes A_i - \mathbb{E} \sum_i A_i \otimes A_i \| \leqslant O(n^{3/2} \log(n)^{1/2})$ *and* $\| \sum_i v_0(i) A_i \| \leqslant O(\sqrt{n})$.

Proof The first claim is immediate from Theorem 56. For the second, we note that since v_0 is a unit vector, the matrix $\sum_i v_0(i) A_i$ has independent entries from $\mathcal{N}(0,1)$. Thus, by Lemma 54, $\| \sum_i v_0(i) A_i \| \leqslant O(\sqrt{n})$ with probability $1 - O(n^{-100})$, as desired. ■

Lemma 61 (Restatement of Lemma 22 for General Odd k**)** *Let* $\mathbf{A}$ *be a* k*-tensor with* k *an odd integer, with independent entries from* $\mathcal{N}(0,1)$. *Let* $v_0 \in \mathbb{R}^n$ *be a unit vector, and let* V *be the* $n^{(k+1)/2} \times n^{(k-1)/2}$ *unfolding of* $v_0^{\otimes k}$. *Let* A *be the* $n^{(k+1)/2} \times n^{(k-1)/2}$ *unfolding of* $\mathbf{A}$. *Then with probability* $1 - O(n^{-100})$, *the matrix* A *satisfies* $A^T A = n^{(k+1)/2} I + E$ *for some* E *with* $\|E\| \leqslant O(n^{k/2} \log(n))$ *and* $\|A^T V\| \leqslant O(n^{(k-1)/4} \log(n)^{1/2})$.

Proof With $\delta = O(1/\sqrt{n})$ and $t = 1$, our parameters will satisfy $n^{(k+1)/2} \geqslant (t/\delta)^2 n^{(k-1)/2}$. Hence, by Lemma 52,

$$\|E\| = \|A^T A^T - n^{(k+1)/2} I\| = \left\| \sum_{|\alpha| = (k+1)/2} a_\alpha a_\alpha^T - n^{(k+1)/2} \cdot \mathrm{Id} \right\| \leqslant n^{(k+1)/2} \cdot O\left(\frac{1}{\sqrt{n}} \right) = O(n^{k/2})$$

with probability at least $1 - 2\exp(-n^{(k+1)/2}) \geqslant 1 - O(n^{-100})$.

It remains to bound $\|A^T V\|$. Note that $V = uw^T$ for fixed unit vectors $u \in \mathbb{R}^{(k-1)/2}$ and $w \in \mathbb{R}^{(k+1)/2}$. So $\|A^T V\| \leqslant \|A^T u\|$. But $A^T u$ is distributed according to $\mathcal{N}(0,1)^n$ and so $\|A^T u\| \leqslant O(\sqrt{n \log n})$ with probability $1 - n^{-100}$ by standard arguments. ■

E.4. Concentration for Lower Bounds

The next theorems collects the concentration results necessary to apply our lower bounds Theorem 31 and Theorem 32 to random polynomials.

Lemma 62 *Let* $\mathbf{A}$ *be a random* 3*-tensor with unit Gaussian entries. For a real parameter* λ, *let* $\mathcal{L} : \mathbb{R}[x]_4 \to \mathbb{R}$ *be the linear operator whose matrix representation* $M_{\mathcal{L}}$ *is given by* $M_{\mathcal{L}} := \frac{1}{n^2 \lambda^2} \sum_{\pi \in \mathcal{S}_4} \pi \cdot AA^T$. *There is* $\lambda = O(n^{3/2}/\log(n)^{1/2})$ *so that with probability* $1 - O(n^{-50})$ *the following events all occur for every* $\pi \in \mathcal{S}_3$.

$$\begin{aligned} &-2\lambda^2 \cdot \Pi \,\mathrm{Id}\, \Pi \\ &\qquad \preceq \frac{1}{2} \Pi \left[\sigma \cdot A^\pi (A^\pi)^T + \sigma^2 \cdot A^\pi (A^\pi)^T + (\sigma \cdot A^\pi (A^\pi)^T)^T + (\sigma^2 \cdot A^\pi (A^\pi)^T)^T \right] \Pi \end{aligned} \tag{E.1}$$

$$\langle \mathbf{A}, \sum_{\pi \in \mathcal{S}_3} \mathbf{A}^\pi \rangle = \Omega(n^3) \tag{E.2}$$

$$\langle \mathrm{Id}^{\mathrm{sym}}, A^\pi (A^\pi)^T \rangle = O(n^3) \tag{E.3}$$

$$n\left(\max_i \left|\frac{1}{\lambda n^{3/2}}\langle \mathrm{Id}_{n\times n}, A_i^\pi\rangle\right|\right) = O(1) \tag{E.4}$$

$$n^2\left(\max_{i\neq j} \left|\mathcal{L}\,\|x\|^2 x_i x_j\right|\right) = O(1) \tag{E.5}$$

$$n^{3/2}\left(\max_i \left|\mathcal{L}\,\|x\|^2 x_i^2 - \tfrac{1}{n}\,\mathcal{L}\,\|x\|^4\right|\right) = O(1/n) \tag{E.6}$$

Proof For (E.1), from Theorem 56, Lemma 36, the observation that multiplication by an orthogonal operator cannot increase the operator norm, a union bound over all π, and the triangle inequality, it follows that:

$$\left\|\sigma\cdot A^\pi(A^\pi)^T - \mathbb{E}[\sigma\cdot A^\pi(A^\pi)^T] + \sigma^2\cdot A^\pi(A^\pi)^T - \mathbb{E}[\sigma^2\cdot A^\pi(A^\pi)^T]\right\| \leqslant 2\lambda^2.$$

with probability $1-n^{-100}$. By the definition of the operator norm and another application of triangle inequality, this implies

$$\begin{aligned}-4\lambda^2\,\mathrm{Id} \preceq\ & \sigma\cdot A^\pi(A^\pi)^T + \sigma^2\cdot A^\pi(A^\pi)^T + (\sigma\cdot A^\pi(A^\pi)^T)^T + (\sigma^2\cdot A^\pi(A^\pi)^T)^T \\ & - \mathbb{E}[\sigma\cdot A^\pi(A^\pi)^T] - \mathbb{E}[\sigma^2\cdot A^\pi(A^\pi)^T] - \mathbb{E}[(\sigma\cdot A^\pi(A^\pi)^T)^T] - \mathbb{E}[(\sigma^2\cdot A^\pi(A^\pi)^T)^T]\,.\end{aligned}$$

We note that $\mathbb{E}[\sigma\cdot A^\pi(A^\pi)^T] = \sigma\cdot\mathrm{Id}$ and $\mathbb{E}[\sigma^2\cdot A^\pi(A^\pi)^T] = \sigma^2\cdot\mathrm{Id}$, and the same for their transposes, and that $\Pi(\sigma\cdot\mathrm{Id}+\sigma^2\cdot\mathrm{Id})\Pi \succeq 0$. So, dividing by 2 and projecting onto the Π subspace:

$$\begin{aligned}&-2\lambda^2\cdot\Pi\,\mathrm{Id}\,\Pi \\ &\quad\preceq \frac{1}{2}\Pi\left(\sigma\cdot A^\pi(A^\pi)^T + \sigma^2\cdot A^\pi(A^\pi)^T + (\sigma\cdot A^\pi(A^\pi)^T)^T + (\sigma^2\cdot A^\pi(A^\pi)^T)^T\right)\Pi\,.\end{aligned}$$

We turn to (E.2). By a Chernoff bound, $\langle \mathbf{A},\mathbf{A}\rangle = \Omega(n^3)$ with probability $1-n^{-100}$. Let $\pi\in\mathcal{S}_3$ be a nontrivial permutation. To each multi-index α with $|\alpha|=3$ we associate its orbit O_α under $\langle\pi\rangle$. If α has three distinct indices, then $|O_\alpha|>1$ and $\sum_{\beta\in O_\alpha} A_\beta A_\beta^\pi$ is a random variable X_α with the following properties:

- $|X_\alpha| < O(\log n)$ with probability $1-n^{-\omega(1)}$.
- X_α and $-X_\alpha$ are identically distributed.

Next, we observe that we can decompose

$$\langle \mathbf{A},\mathbf{A}^\pi\rangle = \sum_{|\alpha|=3}\mathbf{A}_\alpha\mathbf{A}_\alpha^\pi = R + \sum_{O_\alpha} X_\alpha\,,$$

where R is the sum over multi-indices α with repeated indices, and therefore has $|R| = \tilde{O}(n^2)$ with probability $1-n^{-100}$. By a standard Chernoff bound, $|\sum_{O_\alpha} X_\alpha| = O(n^2)$ with probability $1-O(n^{-100})$. By a union bound over all π, we get that with probability $1-O(n^{-100})$,

$$\langle \mathbf{A}, \sum_{\pi\in\mathcal{S}_3}\mathbf{A}^\pi\rangle = n^3 - O(n^2) = \Omega(n^3)\,,$$

establishing (E.2).

Next up is (E.3). Because A^π are identically distributed for all $\pi \in \mathcal{S}_3$ we assume without loss of generality that $A^\pi = A$. The matrix $\mathrm{Id}^{\mathrm{sym}}$ has $O(n^2)$ nonzero entries. Any individual entry of AA^T is with probability $1 - n^{-\omega(1)}$ at most $O(n)$. So $\langle \mathrm{Id}^{\mathrm{sym}}, AA^T \rangle = O(n^3)$ with probability $1 - O(n^{-100})$.

Next, (E.4). As before, we assume without loss of generality that π is the trivial permutation. For fixed $1 \leqslant i \leqslant n$, we have $\langle \mathrm{Id}_{n\times n}, A_i \rangle = \sum_j A_{ijj}$, which is a sum of n independent unit Gaussians, so $|\langle \mathrm{Id}_{n\times n}, A_i \rangle| \leqslant O(\sqrt{n}\log n)$ with probability $1 - n^{-\omega(1)}$. By a union bound over i this also holds for $\max_i |\langle \mathrm{Id}_{n\times n}, A_i\rangle|$. Thus with probability $1 - O(n^{-100})$,

$$n\left(\max_i \left|\frac{1}{n^{3/2}\lambda}\langle \mathrm{Id}_{n\times n}, A_i\rangle\right|\right) \leqslant \frac{\tilde{O}(1)}{\lambda}\,.$$

Last up are (E.5) and (E.6). Since we will do a union bound later, we fix $i, j \leqslant n$. Let $w \in \mathbb{R}^{n^2}$ be the matrix flattening of $\mathrm{Id}_{n\times n}$. We expand $\mathcal{L}\,\|x\|^2 x_i x_j$ as

$$\begin{aligned}\mathcal{L}\,\|x\|^2 x_i x_j = \frac{1}{n^2 O(\lambda^2)}&(w^T \Pi(AA^T + \sigma\cdot AA^T + \sigma^2 \cdot AA^T)\Pi(e_i \otimes e_j)\\ &+ w^T\Pi(AA^T + \sigma\cdot AA^T + \sigma^2\cdot AA^T)^T \Pi(e_i\otimes e_j))\,.\end{aligned}$$

We have $\Pi w = w$ and we let $e_{ij} := \Pi(e_i \otimes e_j) = \frac{1}{2}(e_i\otimes e_j + e_j \otimes e_i)$. So using Lemma 36,

$$\begin{aligned}n^2 O(\lambda^2)\,\mathcal{L}\,\|x\|^2 x_i x_j &= w^T(AA^T + \sigma\cdot AA^T + \sigma^2\cdot AA^T)e_{ij}\\ &\quad + w^T(AA^T + \sigma\cdot AA^T + \sigma^2\cdot AA^T)^T e_{ij}\\ &= w^T\Big(AA^T e_{ij}\\ &\quad + \frac{1}{2}\sum_k A_k e_j \otimes A_k e_i + \frac{1}{2}\sum_k A_k e_i \otimes A_k e_j\\ &\quad + \frac{1}{2}\sum_k A_k^T e_j \otimes A_k^T e_i + \frac{1}{2}\sum_k A_k^T e_i \otimes A_k^T e_j\\ &\quad + \frac{1}{2}\sum_k A_k e_i \otimes A_k^T e_j + \frac{1}{2}\sum_k A_k e_j \otimes A_k^T e_i\\ &\quad + \frac{1}{2}\sum_k A_k^T e_i \otimes A_k e_j + \frac{1}{2}\sum_k A_k^T e_j \otimes A_k e_i\Big)\,.\end{aligned}$$

For $i \neq j$, each term $w^T(A_k e_j \otimes A_k e_i)$ (or similar, with various transposes) is the sum of n independent products of pairs of independent unit Gaussians, so by a Chernoff bound followed by a union bound, with probability $1 - n^{-\omega(1)}$ all of them are $O(\sqrt{n}\log n)$. There are $O(n)$ such terms, for an upper bound of $O(n^{3/2}(\log n))$ on the contribution from the tensored parts.

At the same time, $w^T A$ is a sum $\sum_k a_{kk}$ of n rows of A and Ae_{ij} is the average of two rows of A; since $i \neq j$ these rows are independent from $w^T A$. Writing this out, $w^T AA^T e_{ij} = \frac{1}{2}\sum_k \langle a_{kk}, a_{ij} + a_{ji}\rangle$. Again by a standard Chernoff and union bound argument this is in absolute value at most $O(n^{3/2}(\log n))$ with probability $1 - n^{-\omega(1)}$. In sum, when $i \neq j$, with probability at least $1 - n^{-\omega(1)}$, we get $|\mathcal{L}\,\|x\|^2 x_i x_j| = O(1/n^2 \log n)$. After a union bound, the maximum over all i, j is $O(1/n^2)$. This concludes (E.5).

In the $i = j$ case, since $\sum_k \langle w, A_k e_i \otimes A_k e_i\rangle = \sum_{j,k}\langle e_j, A_k e_i\rangle^2$ is a sum of n^2 independent square Gaussians, by a Bernstein inequality, $|\sum_k \langle w, A_k e_i \otimes A_k e_i\rangle - n^2| \leqslant O(n\log^{1/2} n)$ with probability

$1 - n^{-\omega(1)}$. The same holds for the other tensored terms, and for $w^T AA^T e_{ii}$, so when $i = j$ we get that $|O(\lambda^2)\,\mathcal{L}\,\|x\|^2 x_i^2 - 5| \leqslant O((\log^{1/2} n)/n)$ with probability $1 - n^{-\omega(1)}$. Summing over all i, we find that $|O(\lambda^2)\,\mathcal{L}\,\|x\|^4 - 5n| \leqslant O(\log^{1/2} n)$, so that $O(\lambda^2)|\,\mathcal{L}\,\|x\|^2 x_i^2 - \frac{1}{n}\,\mathcal{L}\,\|x\|^4| \leqslant O((\log^{1/2} n)/n)$ with probability $1 - n^{-\omega(1)}$. A union bound over i completes the argument. ■

Lemma 63 *Let* $\mathbf{A}$ *be a random* 4*-tensor with unit Gaussian entries. There is* $\lambda^2 = O(n)$ *so that when* $\mathcal{L} : \mathbb{R}[x]_4 \to \mathbb{R}$ *is the linear operator whose matrix representation* $M_{\mathcal{L}}$ *is given by* $M_{\mathcal{L}} := \frac{1}{n^2\lambda^2}\sum_{\pi \in \mathcal{S}_4} A^\pi$, *with probability* $1 - O(n^{-50})$ *the following events all occur for every* $\pi \in \mathcal{S}_4$.

$$-\lambda^2 \preceq \frac{1}{2}(A^\pi + (A^\pi)^T) \tag{E.7}$$

$$\langle \mathbf{A}, \sum_{\pi \in \mathcal{S}_4} \mathbf{A}^\pi \rangle = \Omega(n^4) \tag{E.8}$$

$$\langle \mathrm{Id}^{\mathrm{sym}}, A^\pi \rangle = O(\lambda^2 \sqrt{n}) \tag{E.9}$$

$$n^2 \max_{i \neq j} |\,\mathcal{L}\,\|x\|^2 x_i x_j| = O(1) \tag{E.10}$$

$$n^{3/2} \max_{i} |\,\mathcal{L}\,\|x\|^2 x_i^2| = O(1)\,. \tag{E.11}$$

Proof

For (E.7), we note that $\frac{1}{2}A^\pi + (A^\pi)^T$ is an $n^2 \times n^2$ matrix with unit Gaussian entries. Thus, by Lemma 54, we have $\frac{1}{2}\|A^\pi + (A^\pi)^T\| \leqslant O(n) = O(\lambda)$. For (E.8) only syntactic changes are needed from the proof of (E.3). For (E.9), we observe that $\langle \mathrm{Id}^{\mathrm{sym}}, A^\pi \rangle$ is a sum of $O(n^2)$ independent Gaussians, so is $O(n \log n) \leqslant O(\lambda^2 \sqrt{n})$ with probability $1 - O(n^{-100})$. We turn finally to (E.10) and (E.11). Unlike in the degree 3 case, there is nothing special here about the diagonal so we will able to bound these cases together. Fix $i, j \leqslant n$. We expand $\mathcal{L}\,\|x\|^2 x_i x_j$ as $\frac{1}{n^2\lambda^2}\sum_{\pi \in \mathcal{S}_4} w^T A^\pi(e_i \otimes e_j)$. The vector $A^\pi(e_i \otimes e_j)$ is a vector of unit Gaussians, so $w^T A^\pi(e_i \otimes e_j) = O(\sqrt{n}\log n)$ with probability $1 - n^{-\omega(1)}$. Thus, also with probability $1 - n^{-\omega(1)}$, we get $n^2 \max_{i,j} |\,\mathcal{L}\,\|x\|^2 x_i x_j| = O(1)$, which proves both (E.10) and (E.11). ■