The power of sum-of-squares for detecting hidden structures

Samuel B. Hopkins* Pravesh K. Kothari † Aaron Potechin
Prasad Raghavendra Tselil Schramm‡ David Steurer§

September 10, 2019

Abstract

We study planted problems—finding hidden structures in random noisy inputs—through the lens of the sum-of-squares semidefinite programming hierarchy (SoS). This family of powerful semidefinite programs has recently yielded many new algorithms for planted problems, often achieving the best known polynomial-time guarantees in terms of accuracy of recovered solutions and robustness to noise. One theme in recent work is the design of spectral algorithms which match the guarantees of SoS algorithms for planted problems. Classical spectral algorithms are often unable to accomplish this: the twist in these new spectral algorithms is the use of spectral structure of matrices whose entries are low-degree polynomials of the input variables.

We prove that for a wide class of planted problems, including refuting random constraint satisfaction problems, tensor and sparse PCA, densest-k-subgraph, community detection in stochastic block models, planted clique, and others, eigenvalues of degree-d matrix polynomials are as powerful as SoS semidefinite programs of size roughly n^d. For such problems it is therefore always possible to match the guarantees of SoS without solving a large semidefinite program.

Using related ideas on SoS algorithms and low-degree matrix polynomials (and inspired by recent work on SoS and the planted clique problem [BHK+16]), we prove new nearly-tight SoS lower bounds for the tensor and sparse principal component analysis problems. Our lower bounds are the first to suggest that improving upon the signal-to-noise ratios handled by existing polynomial-time algorithms for these problems may require subexponential time.

*Cornell University, samhop@cs.cornell.edu Partially supported by an NSF GRFP under grant no. 1144153, by a Microsoft Research Graduate Fellowship, and by David Steurer's NSF CAREER award.

†Princeton University and IAS, kothari@cs.princeton.edu

‡UC Berkeley, tscrhamm@cs.berkeley.edu. Supported by an NSF Graduate Research Fellowship (1106400).

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

Contents

1 Introduction

Recent years have seen a surge of progress in algorithm design via the sum-of-squares (SoS) semidefinite programming hierarchy. Initiated by the work of [BBH+12], who showed that polynomial time algorithms in the hierarchy solve all known integrality gap instances for Unique Games and related problems, a steady stream of works have developed efficient algorithms for both worst-case [BKS14, BKS15, BKS17, BGG+16] and average-case problems [HSS15, GM15, BM16, RRS16, BGL16, MSS16a, PS17]. The insights from these works extend beyond individual algorithms to characterizations of broad classes of algorithmic techniques. In addition, for a large class of problems (including constraint satisfaction), the family of SoS semidefinite programs is now known to be as powerful as *any* semidefinite program (SDP) [LRS15].

In this paper we focus on recent progress in using Sum of Squares algorithms to solve average-case, and especially *planted* problems—problems that ask for the recovery of a planted *signal* perturbed by random *noise*. Key examples are finding solutions of random constraint satisfaction problems (CSPs) with planted assignments [RRS16] and finding planted optima of random polynomials over the n-dimensional unit sphere [RRS16, BGL16]. The latter formulation captures a wide range of unsupervised learning problems, and has led to many unsupervised learning algorithms with the best-known polynomial time guarantees [BKS15, BKS14, MSS16b, HSS15, PS17, BGG+16].

In many cases, classical algorithms for such planted problems are *spectral* algorithms—i.e., using the top eigenvector of a natural matrix associated with the problem input to recover a planted solution. The canonical algorithms for the *planted clique* [AKS98], *principal components analysis* (PCA) [Pea01], and *tensor decomposition* (which is intimately connected to optimizaton of polynomials on the unit sphere) [Har70] are all based on this general scheme. In all of these cases, the algorithm employs the top eigenvector of a matrix which is either given as input (the adjacency matrix, for planted clique), or is a simple function of the input (the empirical covariance, for PCA).

Recent works have shown that one can often improve upon these basic spectral methods using SoS, yielding better accuracy and robustness guarantees against noise in recovering planted solutions. Furthermore, for worst case problems—as opposed to the average-case planted problems we consider here—semidefinite programs are strictly more powerful than spectral algorithms.[1] *A priori* one might therefore expect that these new SoS guarantees for planted problems would not be achievable via spectral algorithms. But curiously enough, in numerous cases these stronger guarantees for planted problems can be achieved by spectral methods! The twist is that the entries of these matrices are low-degree polynomials in the input to the algorithm . The result is a new family of low-degree spectral algorithms with guarantees matching SoS but requriring only eigenvector computations instead of general semidefinite programming [HSSS16, RRS16, AOW15a].

This leads to the following question which is the main focus of this work.

Are SoS algorithms equivalent to low-degree spectral methods for planted problems?

We answer this question affirmatively for a wide class of distinguishing problems which includes refuting random CSPs, tensor and sparse PCA, densest-k-subgraph, community detection in stochastic block models, planted clique, and more. Our positive answer to this question implies that a light-weight algorithm—computing the top eigenvalue of a single matrix whose entries are

[1] For example, consider the contrast between the SDP algorithm for Max-Cut of Goemans and Williamson, [GW94], and the spectral algorithm of Trevisan [Tre09]; or the SDP-based algorithms for coloring worst-case 3-colorable graphs [KT17] relative to the best spectral methods [AK97] which only work for random inputs.

low-degree polynomials in the input—can recover the performance guarantees of an often bulky semidefinite programming relaxation.

To complement this picture, we prove two new SoS lower bounds for particular planted problems, both variants of component analysis: sparse principal component analysis and tensor principal component analysis (henceforth sparse PCA and tensor PCA, respectively) [ZHT06, RM14]. For both problems there are nontrivial low-degree spectral algorithms, which have better noise tolerance than naive spectral methods [HSSS16, DM14b, RRS16, BGL16]. Sparse PCA, which is used in machine learning and statistics to find important coordinates in high-dimensional data sets, has attracted much attention in recent years for being apparently computationally intractable to solve with a number of samples which is more than sufficient for brute-force algorithms [KNV+15, BR13b, MW15a]. Tensor PCA appears to exhibit similar behavior [HSS15]. That is, both problems exhibit *information-computation gaps*.

Our SoS lower bounds for both problems are the strongest yet formal evidence for information-computation gaps for these problems. We rule out the possibility of subexponential-time SoS algorithms which improve by polynomial factors on the signal-to-noise ratios tolerated by the known low degree spectral methods. In particular, in the case of sparse PCA, it appeared possible prior to this work that it might be possible in quasipolynomial time to recover a k-sparse unit vector v in p dimensions from $O(k \log p)$ samples from the distribution $\mathcal{N}(0, \mathrm{Id} + vv^\top)$. Our lower bounds suggest that this is extremely unlikely; in fact this task probably requires polynomial SoS degree and hence $\exp(n^{\Omega(1)})$ time for SoS algorithms. This demonstrates that (at least with regard to SoS algorithms) both problems are much harder than the *planted clique* problem, previously used as a basis for reductions in the setting of sparse PCA [BR13b].

Our lower bounds for sparse and tensor PCA are closely connected to the failure of low-degree spectral methods in high noise regimes of both problems. We prove them both by showing that with noise beyond what known low-degree spectral algorithms can tolerate, even low-degree *scalar* algorithms (the result of restricting low-degree spectral algorithms to 1×1 matrices) would require subexponential time to detect and recover planted signals. We then show that in the restricted settings of tensor and sparse PCA, ruling out these weakened low-degree spectral algorithms is enough to imply a strong SoS lower bound.

1.1 SoS and spectral algorithms for robust inference

We turn to our characterization of SoS algorithms for planted problems in terms of low-degree spectral algorithms. First, a word on planted problems. Many planted problems have several formulations: *search*, in which the goal is to recover a planted solution, *refutation*, in which the goal is to certify that no planted solution is present, and *distinguishing*, where the goal is to determine with good probability whether an instance contains a planted solution or not. Often an algorithm for one version can be parlayed into algorithms for the others, but distinguishing problems are often the easiest, and we focus on them here.

A distinguishing problem is specified by two distributions on instances: a *planted* distribution supported on instances with a hidden structure, and a *uniform* distribution, where samples w.h.p. contain no hidden structure. Given an instance drawn with equal probability from the planted or the uniform distribution, the goal is to determine with probability greater than $\frac{1}{2}$ whether or not the instance comes from the planted distribution. For example:

Planted clique *Uniform distribution:* $G(n, \frac{1}{2})$, the Erdős-Renyi distribution, which w.h.p. contains no clique of size $\omega(\log n)$. *Planted distribution:* The uniform distribution on graphs containing a n^ε-size clique, for some $\varepsilon > 0$. (The problem gets harder as ε gets smaller, since the distance between the distributions shrinks.)

Planted 3XOR *Uniform distribution:* a 3XOR instance on n variables and $m > n$ equations $x_i x_j x_k = a_{ijk}$, where all the triples (i, j, k) and the signs $a_{ijk} \in \{\pm 1\}$ are sampled uniformly and independently. No assignment to x will satisfy more than a 0.51-fraction of the equations, w.h.p. *Planted distribution:* The same, except the signs a_{ijk} are sampled to correlate with $b_i b_j b_k$ for a randomly chosen $b_i \in \{\pm 1\}$, so that the assignment $x = b$ satisfies a 0.9-fraction of the equations. (The problem gets easier as m/n gets larger, and the contradictions in the uniform case become more locally apparent.)

We now formally define a family of distinguishing problems, in order to give our main theorem. Let $\mathcal{I}$ be a set of instances corresponding to a product space (for concreteness one may think of $\mathcal{I}$ to be the set of graphs on n vertices, indexed by $\{0,1\}^{\binom{n}{2}}$, although the theorem applies more broadly). Let ν, our uniform distrbution, be a product distribution on $\mathcal{I}$.

With some decision problem $\mathcal{P}$ in mind (e.g. does G contain a clique of size $\geqslant n^\varepsilon$?), let $\mathcal{X}$ be a set of solutions to $\mathcal{P}$; again for concreteness one may think of $\mathcal{X}$ as being associated with cliques in a graph, so that $\mathcal{X} \subset \{0,1\}^n$ is the set of all indicator vectors on at least n^ε vertices.

For each solution $x \in \mathcal{X}$, let $\mu_{|_x}$ be the uniform distribution over instances $I \in \mathcal{I}$ that contain x. For example, in the context of planted clique, if x is a clique on vertices $1, \ldots, n^\varepsilon$, then $\mu_{|_x}$ would be the uniform distribution on graphs containing the clique $1, \ldots, n^\varepsilon$. We define the planted distribution μ to be the uniform mixture over μ_x, $\mu = U_{x \sim \mathcal{X}} \mu_{|_x}$.

The following is our main theorem on the equivalence of sum of squares algorithms for distinguishing problems and spectral algorithms employing low-degree matrix polynomials.

Theorem 1.1 (Informal)**.** *Let $N, n \in \mathcal{N}$, and let $\mathcal{A}, \mathcal{B}$ be sets of real numbers. Let $\mathcal{I}$ be a family of instances over $\mathcal{A}^N$, and let $\mathcal{P}$ be a decision problem over $\mathcal{I}$ with $\mathcal{X} = \mathcal{B}^n$ the set of possible solutions to $\mathcal{P}$ over $\mathcal{I}$. Let $\{g_j(x, I)\}$ be a system of $n^{O(d)}$ polynomials of degree at most d in the variables x and constant degree in the variables $\mathcal{I}$ that encodes $\mathcal{P}$, so that*

- *for $I \sim_\nu \mathcal{I}$, with high probability the system is unsatisfiable and admits a degree-d SoS refutation, and*
- *for $I \sim_\mu \mathcal{I}$, with high probability the system is satisfiable by some solution $x \in X$, and x remains feasible even if all but an $n^{-0.01}$-fraction of the coordinates of $\mathcal{I}$ are re-randomized according to ν.*

Then there exists a matrix whose entries are degree-$O(d)$ polynomials $Q : \mathcal{I} \to \mathbb{R}^{\binom{n}{\leqslant d} \times \binom{n}{\leqslant d}}$ such that

$$\mathop{\mathbb{E}}_{I \sim \nu} [\lambda^+_{max}(Q(I))] \leqslant 1, \quad \textit{while} \quad \mathop{\mathbb{E}}_{I \sim \mu} [\lambda^+_{max}(Q(I))] \geqslant n^{10d},$$

where $\lambda^+_{\max}$ denotes the maximum non-negative eigenvalue.

The condition that a solution x remain feasible if all but a fraction of the coordinates of $I \sim \mu_{|_x}$ are re-randomized should be interpreted as a noise-robustness condition. To see an example, in the context of planted clique, suppose we start with a planted distribution over graphs with a clique x of size $n^{\varepsilon+0.01}$. If a random subset of $n^{0.99}$ vertices are chosen, and all edges not entirely contained in that subset are re-randomized according to the $G(n, 1/2)$ distribution, then with high probability at least n^ε of the vertices in x remain in a clique, and so x remains feasible for the problem $\mathcal{P}$: G has a clique of size $\geqslant n^\varepsilon$?

1.2 SoS and information-computation gaps

Computational complexity of planted problems has become a rich area of study. The goal is to understand which planted problems admit efficient (polynomial time) algorithms, and to study the *information-computation gap* phenomenon: many problems have noisy regimes in which planted structures can be found by inefficient algorithms, but (conjecturally) not by polynomial time algorithms. One example is the *planted clique* problem, where the goal find a large clique in a sample from the uniform distribution over graphs containing a clique of size n^ε for a small constant $\varepsilon > 0$. While the problem is solvable for any $\varepsilon > 0$ by a brute-force algorithm requiring $n^{\Omega(\log n)}$ time, polynomial time algorithms are conjectured to require $\varepsilon \geqslant \frac{1}{2}$.

A common strategy to provide evidence for such a gap is to prove that powerful classes of efficient algorithms are unable to solve the planted problem in the (conjecturally) hard regime. SoS algorithms are particularly attractive targets for such lower bounds because of their broad applicability and strong guarantees.

In a recent work, Barak et al. [BHK+16] show an SoS lower bound for the planted clique problem, demonstrating that when $\varepsilon < \frac{1}{2}$, SoS algorithms require $n^{\Omega(\log n)}$ time to solve planted clique. Intriguingly, they show that in the case of planted clique that SoS algorithms requiring $\approx n^d$ time can distinguish planted from random graphs only when there is a *scalar-valued* degree $\approx d \cdot \log n$ polynomial $p(A) : \mathbb{R}^{n \times n} \to \mathbb{R}$ (here A is the adjacency matrix of a graph) with

$$\mathop{\mathbb{E}}_{G(n,1/2)} p(A) = 0, \qquad \mathop{\mathbb{E}}_{\text{planted}} p(A) \geqslant n^{\Omega(1)} \cdot \left(\mathop{\mathbb{V}}_{G(n,1/2)} p(A) \right)^{1/2} .$$

That is, such a polynomial p has much larger expectation in under the planted distribution than its standard deviation in uniform distribution. (The choice of $n^{\Omega(1)}$ is somewhat arbitrary, and could be replaced with $\Omega(1)$ or $n^{\Omega(d)}$ with small changes in the parameters.) By showing that as long as $\varepsilon < \frac{1}{2}$ any such polynomial p must have degree $\Omega(\log n)^2$, they rule out efficient SoS algorithms when $\varepsilon < \frac{1}{2}$. Interestingly, this matches the spectral distinguishing threshold—the spectral algorithm of [AKS98] is known to work when $\varepsilon \geqslant \frac{1}{2}$.

This stronger characterization of SoS for the planted clique problem, in terms of *scalar* distinguishing algorithms rather than *spectral* distinguishing algorihtms, may at first seem insignificant. To see why the scalar characterization is more powerful, we point out that if the degree-d moments of the planted and uniform distributions are known, determining the optimal scalar distinguishing polynomial is easy: given a planted distribution μ and a random distribution ν over instances $\mathcal{I}$, one just solves a linear algebra problem in the $n^{d \log n}$ coefficients of p to maximize the expectation over μ relative to ν:

$$\max_p \mathop{\mathbb{E}}_{\mathcal{I} \sim \mu} [p^2(\mathcal{I})] \quad s.t. \mathop{\mathbb{E}}_{\mathcal{I} \sim \nu} [p^2(\mathcal{I})] = 1 .$$

It is not difficult to show that the optimal solution to the above program has a simple form: it is the projection of the *relative density of ν with respect to μ* projected to the degree-$d \log n$ polynomials. So given a pair of distributions μ, ν, in $n^{O(d \log n)}$ time, it is possible to determine whether there exists a degree-$d \log n$ scalar distinguishing polynomial. Answering the same question about the existence of a spectral distinguisher is more complex, and to the best of our knowledge cannot be done efficiently.

Given this powerful theorem for the case of the planted clique problem, one may be tempted to conjecture that this stronger, *scalar* distinguisher characterization of the SoS algorithm applies more broadly than just to the planted clique problem, and perhaps as broadly as Theorem 1.1. If this conjecture is true, given a pair of distributions ν and μ with known moments, it would be possible in many cases to efficiently and mechanically determine whether polynomial-time SoS distinguishing algorithms exist!

Conjecture 1.2. *In the setting of Theorem 1.1, the conclusion may be replaced with the conclusion that there exists a scalar-valued polynomial $p : \mathcal{I} \to \mathbb{R}$ of degree $O(d \cdot \log n)$ so that*

$$\mathop{\mathbb{E}}_{uniform} p(I) = 0 \text{ and } \mathop{\mathbb{E}}_{planted} p(I) \geqslant n^{\Omega(1)} \left(\mathop{\mathbb{E}}_{uniform} p(I)^2 \right)^{1/2}$$

To illustrate the power of this conjecture, in the beginning of Section 6 we give a short and self-contained explanation of how this predicts, via simple linear algebra, our $n^{\Omega(1)}$-degree SoS lower bound for tensor PCA. As evidence for the conjecture, we verify this prediction by proving such a lower bound unconditionally.

We also note why Theorem 1.1 does not imply Conjecture 1.2. While, in the notation of that theorem, the entries of $Q(I)$ are low-degree polynomials in I, the function $M \mapsto \lambda^+_{\max}(M)$ is not (to the best of our knowledge) a low-degree polynomial in the entries of M (even approximately). (This stands in contrast to, say the operator norm or Frobenious norm of M, both of which are exactly or approximately low-degree polynomials in the entries of M.) This means that the final output of the spectral distinguishing algorithm offered by Theorem 1.1 is not a low-degree polynomial in the instance I.

1.3 Exponential lower bounds for sparse PCA and tensor PCA

Our other main results are strong exponential lower bound on the sum-of-squares method (specifically, against $2^{n^{\Omega(1)}}$ time or $n^{\Omega(1)}$ degree algorithms) for the tensor and sparse principal component analysis (PCA). We prove the lower bounds by extending the techniques pioneered in [BHK+16]. In the present work we describe the proofs informally, leaving full details to a forthcoming full version.

Tensor PCA. We start with the simpler case of tensor PCA, introduced by [RM14].

Problem 1.3 (Tensor PCA)**.** Given an order-k tensor in $(\mathbb{R}^n)^{\otimes k}$, determine whether it comes from:

- **Uniform Distribution**: each entry of the tensor sampled independently from $\mathcal{N}(0, 1)$.
- **Planted Distribution**: a spiked tensor, $\mathbf{T} = \lambda \cdot v^{\otimes k} + G$ where v is sampled uniformly from $\mathbb{S}^{n-1}$, and where G is a random tensor with each entry sampled independently from $\mathcal{N}(0, 1)$.

Here, we think of v as a signal hidden by Gaussian noise. The parameter λ is a signal-to-noise ratio. In particular, as λ grows, we expect the distinguishing problem above to get easier.

Tensor PCA is a natural generalization of the PCA problem in machine learning and statistics. Tensor methods in general are useful when data naturally has more than two modalities: for example, one might consider a recommender system which factors in not only people and movies but also time of day. Many natural tensor problems are NP hard in the worst-case. Though this is

not necessarily an obstacle to machine learning applications, it is important to have average-case models to in which to study algorithms for tensor problems. The spiked tensor setting we consider here is one such simple model.

Turning to algorithms: consider first the ordinary PCA problem in a spiked-matrix model. Given an $n \times n$ matrix M, the problem is to distinguish between the case where every entry of M is independently drawn from the standard Gaussian distribution $\mathcal{N}(0,1)$ and the case when M is drawn from a distribution as above with an added rank one shift $\lambda vv^\top$ in a uniformly random direction v. A natural and well-studied algorithm, which solves this problem to information-theoretic optimality is to threshold on the largest singular value/spectral norm of the input matrix. Equivalently, one thresholds on the maximizer of the degree two polynomial $\langle x, Mx\rangle$ in $x \in \mathbb{S}^{n-1}$.

A natural generalization of this algorithm to the tensor PCA setting (restricting for simplicity $k = 3$ for this discussion) is the maximum of the degree-three polynomial $\langle T, x^{\otimes 3}\rangle$ over the unit sphere—equivalently, the (symmetric) injective tensor norm of T. This maximum can be shown to be much larger in case of the planted distribution so long as $\lambda \gg \sqrt{n}$. Indeed, this approach to distinguishing between planted and uniform distributions is information-theoretically optimal [PWB16, BMVX16]. Since recovering the spike v and optimizing the polynomial $\langle T, x^{\otimes 3}\rangle$ on the sphere are equivalent, tensor PCA can be thought of as an average-case version of the problem of optimizing a degree-3 polynomial on the unit sphere (this problem is NP hard in the worst case, even to approximate [HL09, BBH+12]).

Even in this average-case model, it is believed that there is a gap between which signal strengths λ allow recovery of v by brute-force methods and which permit polynomial time algorithms. This is quite distinct from the vanilla PCA setting, where eigenvector algorithms solve the spike-recovery problem to information-theoretic optimality. Nevertheless, the best-known algorithms for tensor PCA arise from computing convex relaxations of this degree-3 polynomial optimization problem. Specifically, the SoS method captures the state of the art algorithms for the problem; it is known to recover the vector v to $o(1)$ error in polynomial time whenever $\lambda \gg n^{3/4}$ [HSS15]. A major open question in this direction is to understand the complexity of the problem for $\lambda \leqslant n^{3/4-\varepsilon}$. Algorithms (again captured by SoS) are known which run in $2^{n^{O(\varepsilon)}}$ time [RRS16, BGG+16]. We show the following theorem which shows that the sub-exponential algorithm above is in fact nearly optimal for SoS algorithm.

Theorem 1.4. *For a tensor T, let*

$$SoS_d(T) = \max_{\tilde{\mathbb{E}}} \tilde{\mathbb{E}}[\langle T, x^{\otimes k}\rangle] \text{ such that } \tilde{\mathbb{E}} \text{ is a degree } d \text{ pseudoexpectation and satisfies } \{\|x\|^2 = 1\}^{2}$$

For every small enough constant $\varepsilon > 0$, if $T \in \mathbb{R}^{n\times n\times n}$ has iid Gaussian or $\{\pm 1\}$ entries, $\mathbb{E}_T\, SoS_d(T) \geqslant n^{k/4-\varepsilon}$, for every $d \leqslant n^{c\cdot\varepsilon}$ for some universal $c > 0$.

In particular for third order tensors (i.e $k = 3$), since degree $n^{\Omega(\varepsilon)}$ SoS is unable to certify that a random 3-tensor has maximum value much less than $n^{3/4-\varepsilon}$, this SoS relaxation cannot be used to distinguish the planted and random distributions above when $\lambda \ll n^{3/4-\varepsilon}$.[3]

[2] For definitions of pseudoexpectations and related matters, see the survey [BS14].

[3] In fact, our proof for this theorem will show somewhat more: that a large family of constraints—any valid constraint which is itself a low-degree polynomial of T—could be added to this convex relaxation and the lower bound would still obtain.

Sparse PCA. We turn to sparse PCA, which we formalize as the following planted distinguishing problem.

Problem 1.5 (Sparse PCA (λ, k))**.** Given an $n \times n$ symmetric real matrix A, determine whether A comes from:

- **Uniform Distribution**: each upper-triangular entry of the matrix A is sampled iid from $\mathcal{N}(0,1)$; other entries are filled in to preserve symmetry.
- **Planted Distribution**: a random k-sparse unit vector v with entries $\{\pm 1/\sqrt{k}, 0\}$ is sampled, and B is sampled from the uniform distribution above; then $A = B + \lambda \cdot vv^\intercal$.

We defer significant discussion to Section 6, noting just a few things before stating our main theorem on sparse PCA. First, the planted model above is sometimes called the *spiked Wigner* model—this refers to the independence of the entries of the matrix B. An alternative model for sparse PCA is the *spiked Wishart* model: A is replaced by $\sum_{i \leqslant m} x_i x_i^\intercal$, where each $x_i \sim \mathcal{N}(0, \mathrm{Id} + \beta vv^\intercal)$, for some number $m \in \mathbb{N}$ of samples and some signal-strength $\beta \in \mathbb{R}$. Though there are technical differences between the models, to the best of our knowledge all known algorithms with provable guarantees are equally applicable to either model; we expect that our SoS lower bounds also apply in the spiked Wishart model.

We generally think of k, λ as small powers of n; i.e. n^ρ for some $\rho \in (0,1)$; this allows us to generally ignore logarithmic factors in our arguments. As in the tensor PCA setting, a natural and information-theoretically optimal algorithm for sparse PCA is to maximize the quadratic form $\langle x, Ax \rangle$, this time over k-sparse unit vectors. For A from the uniform distribution standard techniques (ε-nets and union bounds) show that the maximum value achievable is $O(\sqrt{k} \log n)$ with high probability, while for A from the planted model of course $\langle v, Av \rangle \approx \lambda$. So, when $\lambda \gg \sqrt{k}$ one may distinguish the two models by this maximum value.

However, this maximization problem is NP hard for general quadratic forms A [CPR16]. So, efficient algorithms must use some other distinguisher which leverages the randomness in the instances. Essentially only two polynomial-time-computable distinguishers are known.[4] If $\lambda \gg \sqrt{n}$ then the maximum eigenvalue of A distinguishes the models. If $\lambda \gg k$ then the planted model can be distinguished by the presence of large diagonal entries of A. Notice both of these distinguishers fail for some choices of λ (that is, $\sqrt{k} \ll \lambda \ll \sqrt{n}, k$) for which brute-force methods (optimizing $\langle x, Ax \rangle$ over sparse x) could successfully distinguish planted from uniform A's. The theorem below should be interpreted as an impossibility result for SoS algorithms in the $\sqrt{k} \ll \lambda \ll \sqrt{n}, k$ regime. This is the strongest known impossibility result for sparse PCA among those ruling out classes of efficient algorithms (one reduction-based result is also know, which shows sparse PCA is at least as hard as the planted clique problem [BR13a]. It is also the first evidence that the problem may require subexponential (as opposed to merely quasi-polynomial) time.

Theorem 1.6. *If $A \in \mathbb{R}^{n \times n}$, let*

$$SoS_{d,k}(A) = \max_{\tilde{\mathbb{E}}} \tilde{\mathbb{E}} \langle x, Ax \rangle \text{ s.t. } \tilde{\mathbb{E}} \text{ is degree } d \text{ and satisfies } \left\{ x_i^3 = x_i, \|x\|^2 = k \right\} .$$

[4] If one studies the problem at much finer granularity than we do here, in particular studying λ up to low-order additive terms and how precisely it is possible to estimate the planted signal v, then the situation is more subtle [DM14a].

There are absolute constants $c, \varepsilon^* > 0$ so that for every $\rho \in (0,1)$ and $\varepsilon \in (0, \varepsilon^*)$, if $k = n^\rho$, then for $d \leqslant n^{c\cdot\varepsilon}$,

$$\mathop{\mathbb{E}}_{A\sim\{\pm 1\}^{\binom{n}{2}}} SoS_{d,k}(A) \geqslant \min(n^{1/2-\varepsilon}k, n^{\rho-\varepsilon}k)\,.$$

For more thorough discussion of the theorem, see Section 6.3.

1.4 Related work

On interplay of SoS relaxations and spectral methods. As we have already alluded to, many prior works explore the connection between SoS relaxations and spectral algorithms, beginning with the work of [BBH+12] and including the followup works [HSS15, AOW15b, BM16] (plus many more). Of particular interest are the papers [HSSS16, MS16b], which use the SoS algorithms to obtain *fast* spectral algorithms, in some cases running in time linear in the input size (smaller even than the number of variables in the associated SoS SDP).

In light of our Theorem 1.1, it is particularly interesting to note cases in which the known SoS lower bounds matching the known spectral algorithms—these problems include planted clique (upper bound: [AKS98], lower bound:[5] [BHK+16]), strong refutations for random CSPs (upper bound:[6] [AOW15b, RRS16], lower bounds: [Gri01b, Sch08, KMOW17]), and tensor principal components analysis (upper bound: [HSS15, RRS16, BGG+16], lower bound: this paper).

We also remark that our work applies to several previously-considered distinguishing and average-case problems within the sum-of-squares algorithmic framework: block models [MS16a] , densest-k-subgraph [BCC+10]; for each of these problems, we have by Theorem 1.1 an equivalence between efficient sum-of-squares algorithms and efficient spectral algorithms, and it remains to establish exactly what the tradeoff is between efficiency of the algorithm and the difficulty of distinguishing, or the strength of the noise.

To the best of knowledge, no previous work has attempted to characterize SoS relaxations for planted problems by simpler algorithms in the generality we do here. Some works have considered characterizing degree-2 SoS relaxations (i.e. basic semidefinie programs) in terms of simpler algorithms. One such example is recent work of Fan and Montanari [FM16] who showed that for some planted problems on sparse random graphs, a class of simple procedures called *local algorithms* performs as well as semidefinite programming relaxations.

On strong SoS lower bounds for planted problems. By now, there's a large body of work that establishes lower bounds on SoS SDP for various average case problems. Beginning with the work of Grigoriev [Gri01a], a long line work have established tight lower bounds for random constraint satisfaction problems [Sch08, BCK15, KMOW17] and planted clique [MPW15, DM15, HKP15, RS15, BHK+16]. The recent SoS lower bound for planted clique of [BHK+16] was particularly influential to this work, setting the stage for our main line of inquiry. We also draw attention to previous work on lower bounds for the tensor PCA and sparse PCA problems in the degree-4 SoS relaxation [HSS15, MW15b]—our paper improves on this and extends our understanding of lower bounds for tensor and sparse PCA to any degree.

[5] SDP lower bounds for the planted clique problem were known for smaller degrees of sum-of-squares relaxations and for other SDP relaxations before; see the references therein for details.

[6] There is a long line of work on algorithms for refuting random CSPs, and 3SAT in particular; the listed papers contain additional references.

Tensor principle component analysis was introduced by Montanari and Richard [RM14] who indentified information theoretic threshold for recovery of the planted component and analyzed the maximum likelihood estimator for the problem. The work of [HSS15] began the effort to analyze the sum of squares method for the problem and showed that it yields an efficient algorithm for recovering the planted component with strength $\tilde{\omega}(n^{3/4})$. They also established that this threshold is tight for the sum of squares relaxation of degree 4. Following this, Hopkins et al. [HSSS16] showed how to extract a linear time spectral algorithm from the above analysis. Tomioka and Suzuki derived tight information theoretic thresholds for detecting planted components by establishing tight bounds on the injective tensor norm of random tensors [TS14]. Finally, very recently, Raghavendra et. al. and Bhattiprolu et. al. independently showed sub-exponential time algorithms for tensor pca [RRS16, BGL16]. Their algorithms are spectral and are captured by the sum of squares method.

1.5 Organization

In Section 2 we set up and state our main theorem on SoS algorithms versus low-degree spectral algorithms. In Section 5 we show that the main theorem applies to numerous planted problems—we emphasize that checking each problem is very simple (and barely requires more than a careful definition of the planted and uniform distributions). In Section 3 and Section 4 we prove the main theorerm on SoS algorithms versus low-degree spectral algorithms.

In section 7 we get prepared to prove our lower bound for tensor PCA by proving a structural theorem on factorizations of low-degree matrix polynomials with well-behaved Fourier transforms. In section 8 we prove our lower bound for tensor PCA, using some tools proved in section 9.

Notation. For two matrices A, B, let $\langle A, B\rangle \stackrel{\text{def}}{=} \mathrm{Tr}(AB)$. Let $\|A\|_{Fr}$ denote the Frobenius norm, and $\|A\|$ its spectral norm. For matrix valued functions A, B over $\mathscr{I}$ and a distribution ν over $\mathscr{I} \sim \mathscr{I}$, we will denote $\langle A, B\rangle_\nu = \mathbb{E}_{\mathscr{I}\sim\nu}\langle A(\mathscr{I}), B(\mathscr{I})\rangle$ and by $\|A\|_{Fr,\nu} \stackrel{\text{def}}{=} (\mathbb{E}_{\mathscr{I}\sim\nu}\langle A(\mathscr{I}), A(\mathscr{I})\rangle)^{1/2}$.

For a vector of formal variables $x = (x_1, \ldots, x_n)$, we use $x^{\leqslant d}$ to denote the vector consisting of all monomials of degree at most d in these variables. Furthermore, let us denote $X^{\leqslant d} \stackrel{\text{def}}{=} (x^{\leqslant d})(x^{\leqslant d})^T$.

2 Distinguishing Problems and Robust Inference

In this section, we set up the formal framework within which we will prove our main result.

Uniform vs. Planted Distinguishing Problems

We begin by describing a class of *distinguishing* problems. For $\mathscr{A}$ a set of real numbers, we will use $\mathscr{I} = \mathscr{A}^N$ denote a space of instances indexed by N variables—for the sake of concreteness, it will be useful to think of $\mathscr{I}$ as $\{0,1\}^N$; for example, we could have $N = \binom{n}{2}$ and $\mathscr{I}$ as the set of all graphs on n vertices. However, the results that we will show here continue to hold in other contexts, where the space of all instances is $\mathbb{R}^N$ or $[q]^N$.

Definition 2.1 (Uniform Distinguishing Problem)**.** Suppose that $\mathscr{I}$ is the space of all instances, and suppose we have two distributions over $\mathscr{I}$, a product distribution ν (the "uniform" distribution), and an arbitrary distribution μ (the "planted" distribution).

In a *uniform distinguishing problem*, we are given an instance $\mathcal{I} \in \mathcal{I}$ which is sampled with probability $\frac{1}{2}$ from ν and with probability $\frac{1}{2}$ from μ, and the goal is to determine with probability greater than $\frac{1}{2} + \varepsilon$ which distribution $\mathcal{I}$ was sampled from, for any constant $\varepsilon > 0$.

Polynomial Systems

In the uniform distinguishing problems that we are interested in, the planted distribution μ will be a distribution over instances that obtain a large value for some optimization problem of interest (i.e. the max clique problem). We define polynomial systems in order to formally capture optimization problems.

Program 2.2 (Polynomial System)**.** Let $\mathcal{A}, \mathcal{B}$ be sets of real numbers, let $n, N \in \mathbb{N}$, and let $\mathcal{I} = \mathcal{A}^N$ be a space of instances and $\mathcal{X} \subseteq \mathcal{B}^n$ be a space of solutions. A *polynomial system* is a set of polynomial equalities

$$g_j(x, \mathcal{I}) = 0 \quad \forall j \in [m],$$

where $\{g_j\}_{j=1}^m$ are polynomials in the *program variables* $\{x_i\}_{i\in[n]}$, representing $x \in \mathcal{X}$, and in the *instance variables* $\{\mathcal{I}_j\}_{j\in[N]}$, representing $\mathcal{I} \in \mathcal{I}$. We define $\deg_{\text{prog}}(g_j)$ to be the degree of g_j in the program variables, and $\deg_{\text{inst}}(g_j)$ to be the degree of g_j in the instance variables.

Remark 2.3. For the sake of simplicity, the polynomial system Program 2.2 has no inequalities. Inequalities can be incorporated in to the program by converting each inequality in to an equality with an additional slack variable. Our main theorem still holds, but for some minor modifications of the proof, as outlined in Section 4.

A polynomial system allows us to capture problem-specific objective functions as well as problem-specific constraints. For concreteness, consider a quadtratic program which checks if a graph on n vertices contains a clique of size k. We can express this with the polynomial system over program variables $x \in \mathbb{R}^n$ and instance variables $\mathcal{I} \in \{0,1\}^{\binom{n}{2}}$, where $\mathcal{I}_{ij} = 1$ iff there is an edge from i to j, as follows:

$$\left\{ \sum\nolimits_{i\in[n]} x_i - k = 0 \right\} \cup \{x_i(x_i - 1) = 0\}_{i\in[n]} \cup \{(1 - \mathcal{I}_{ij})x_i x_j = 0\}_{i,j\in\binom{[n]}{2}}.$$

Planted Distributions

We will be concerned with planted distributions of a particular form; first, we fix a polynomial system of interest $\mathcal{S} = \{g_j(x, \mathcal{I})\}_{j\in[m]}$ and some set $\mathcal{X} \subseteq \mathcal{B}^n$ of feasible solutions for $\mathcal{S}$, so that the program variables x represent elements of $\mathcal{X}$. Again, for concreteness, if $\mathcal{I}$ is the set of graphs on n vertices, we can take $\mathcal{X} \subseteq \{0,1\}^n$ to be the set of indicators for subsets of at least n^ε vertices.

For each fixed $x \in \mathcal{X}$, let $\mu_{|x}$ denote the uniform distribution over $\mathcal{I} \in \mathcal{I}$ for which the polynomial system $\{g_j(x, \mathcal{I})\}_{j\in[m]}$ is feasible. The planted distribution μ is given by taking the uniform mixture over the $\mu_{|x}$, i.e., $\mu \sim U_{x\sim\mathcal{X}}[\mu_{|x}]$.

SoS Relaxations

If we have a polynomial system $\{g_j\}_{j\in[m]}$ where $\deg_{\text{prog}}(g_j) \leqslant 2d$ for every $j \in [m]$, then the degree-$2d$ sum-of-squares SDP relaxation for the polynomial system Program 2.2 can be written as,

Program 2.4 (SoS Relaxation for Polynomial System)**.** Let $\mathcal{S} = \{g_j(x, \mathcal{I})\}_{j\in[m]}$ be a polynomial system in instance variables $\mathcal{I} \in \mathcal{I}$ and program variables $x \in \mathcal{X}$. If $\deg_{\text{prog}}(g_j) \leqslant 2d$ for all $j \in [m]$, then an *SoS relaxation* for $\mathcal{S}$ is

$$\begin{aligned} &\langle G_j(\mathcal{I}), X\rangle = 0 \quad \forall j \in [m] \\ &X \succeq 0 \end{aligned}$$

where X is an $[n]^{\leqslant d} \times [n]^{\leqslant d}$ matrix containing the variables of the SDP and $G_j : \mathcal{I} \to \mathbb{R}^{[n]^{\leqslant d}\times[n]^{\leqslant d}}$ are matrices containing the coefficients of $g_j(x, \mathcal{I})$ in x, so that the constraint $\langle G_j(\mathcal{I}), X\rangle = 0$ encodes the constraint $g_j(x, \mathcal{I}) = 0$ in the SDP variables. Note that the entries of G_j are polynomials of degree at most $\deg_{\text{inst}}(g_j)$ in the instance variables.

Sub-instances

Suppose that $\mathcal{I} = \mathcal{A}^N$ is a family of instances; then given an instance $\mathcal{I} \in \mathcal{I}$ and a subset $S \subseteq [N]$, let $\mathcal{I}_S$ denote the sub-instance consisting of coordinates within S. Further, for a distribution Θ over subsets of $[N]$, let $\mathcal{I}_S \sim_\Theta \mathcal{I}$ denote a subinstance generated by sampling $S \sim \Theta$. Let $\mathcal{I}_\downarrow$ denote the set of all sub-instances of an instance $\mathcal{I}$, and let $\mathcal{I}_\downarrow$ denote the set of all sub-instances of all instances.

Robust Inference

Our result will pertain to polynomial systems that define planted distributions whose solutions to sub-instances generalize to feasible solutions over the entire instance. We call this property "robust inference."

Definition 2.5. Let $\mathcal{I} = \mathcal{A}^N$ be a family of instances, let Θ be a distribution over subsets of $[N]$, let $\mathcal{S}$ be a polynomial system as in Program 2.2, and let μ be a planted distribution over instances feasible for $\mathcal{S}$. Then the polynomial system $\mathcal{S}$ is said to satisfy the *robust inference property for probability distribution μ on $\mathcal{I}$ and subsampling distribution Θ*, if given a subsampling $\mathcal{I}_S$ of an instance $\mathcal{I}$ from μ, one can infer a setting of the program variables x^* that remains feasible to $\mathcal{S}$ for most settings of $\mathcal{I}_{\overline{S}}$.

Formally, there exists a map $x : \mathcal{I}_\downarrow \to \mathbb{R}^n$ such that

$$\mathop{\mathbb{P}}_{\mathcal{I}\sim\mu, S\sim\Theta, \tilde{\mathcal{I}}\sim\nu_{|\mathcal{I}_S}} [x(\mathcal{I}_S) \text{ is a feasible for } \mathcal{S} \text{ on } \mathcal{I}_S \circ \tilde{\mathcal{I}}] \geqslant 1 - \varepsilon(n, d)$$

for some negligible function $\varepsilon(n, d)$. To specify the error probability, we will say that polynomial system is $\varepsilon(n, d)$*-robustly inferable*.

Main Theorem

We are now ready to state our main theorem.

Theorem 2.6. *Suppose that $\mathcal{S}$ is a polynomial system as defined in Program 2.2, of degree at most $2d$ in the program variables and degree at most k in the instance variables. Let $B > d \cdot k \in \mathbb{N}$ such that*

1. *The polynpomial system* $\mathcal{S}$ *is* $\frac{1}{n^{8B}}$*-robustly inferable with respect to the planted distribution* μ *and the sub-sampling distribution* Θ*.*

2. *For* $\mathcal{I} \sim \nu$*, the polynomial system* $\mathcal{S}$ *admits a degree-*d *SoS refutation with numbers bounded by* n^B *with probability at least* $1 - \frac{1}{n^{8B}}$*.*

Let $D \in \mathbb{N}$ *be such that for any subset* $\alpha \subseteq [N]$ *with* $|\alpha| \geqslant D - 2dk$*,*

$$\mathop{\mathbb{P}}_{S \sim \Theta}[\alpha \subseteq S] \leqslant \frac{1}{n^{8B}}$$

There exists a degree $2D$ *matrix polynomial* $Q : \mathcal{I} \to \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}}$ *such that,*

$$\frac{\mathbb{E}_{\mathcal{I} \sim \mu}[\lambda^+_{max}(Q(\mathcal{I}))]}{\mathbb{E}_{\mathcal{I} \sim \nu}[\lambda^+_{max}(Q(\mathcal{I}))]} \geqslant n^{B/2}$$

Remark 2.7. Our argument implies a stronger result that can be stated in terms of the eigenspaces of the subsampling operator. Specifically, suppose we define

$$\mathcal{S}_\varepsilon \stackrel{\text{def}}{=} \left\{ \alpha \mid \mathop{\mathbb{P}}_{S \sim \Theta}\{\alpha \subseteq S\} \leqslant \varepsilon \right\}$$

Then, the distinguishing polynomial exhibited by Theorem 2.6 satisfies $Q \in$ span{ monomials $\mathcal{I}_\alpha | \alpha \in \mathcal{S}_\varepsilon$}. This refinement can yield tighter bounds in cases where all monomials of a certain degree are not equivalent to each other. For example, in the Planted Clique problem, each monomial consists of a subgraph and the right measure of the degree of a sub-graph is the number of vertices in it, as opposed to the number of edges in it.

In Section 5, we will make the routine verifications that the conditions of this theorem hold for a variety of distinguishing problems: planted clique (Lemma 5.2), refuting random CSPs (Lemma 5.4, stochastic block models (Lemma 5.6), densest-k-subgraph (Lemma 5.8), tensor PCA (Lemma 5.10), and sparse PCA (Lemma 5.12). Now we will proceed to prove the theorem.

3 Moment-Matching Pseudodistributions

We assume the setup from Section 2: we have a family of instances $\mathcal{I} = \mathcal{A}^N$, a polynomial system $\mathcal{S} = \{g_j(x, \mathcal{I})\}_{j \in [m]}$ with a family of solutions $\mathcal{X} = \mathcal{B}^n$, a "uniform" distribution ν which is a product distribution over $\mathcal{I}$, and a "planted" distribution μ over $\mathcal{I}$ defied by the polynomial system $\mathcal{S}$ as described in Section 2.

The contrapositive of Theorem 2.6 is that if $\mathcal{S}$ is robustly inferable with respect to μ and a distribution over sub-instances Θ, and if there is no spectral algorithm for distinguishing μ and ν, then with high probability there is no degree-d SoS refutation for the polynomial system $\mathcal{S}$ (as defined in Program 2.4). To prove the theorem, we will use duality to argue that if no spectral algorithm exists, then there must exist an object which is in some sense close to a feasible solution to the SoS SDP relaxation.

Since each $\mathcal{I}$ in the support of μ is feasible for $\mathcal{S}$ by definition, a natural starting point is the SoS SDP solution for instances $\mathcal{I} \sim_\mu \mathcal{I}$. With this in mind, we let $\Lambda : \mathcal{I} \to (\mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}})_+$ be an arbitrary function from the support of μ over $\mathcal{I}$ to PSD matrices. In other words, we take

$$\Lambda(\mathcal{I}) = \hat{\mu}(\mathcal{I}) \cdot M(\mathcal{I})$$

where $\hat{\mu}$ is the relative density of μ with respect to ν, so that $\hat{\mu}(\mathcal{I}) = \mu(\mathcal{I})/\nu(\mathcal{I})$, and M is some matrix valued function such that $M(\mathcal{I}) \succeq 0$ and $\|M(\mathcal{I})\| \leqslant B$ for all $\mathcal{I} \in \mathcal{J}$. Our goal is to find a PSD matrix-valued function P that matches the low-degree moments of Λ in the variables $\mathcal{I}$, while being supported over most of $\mathcal{J}$ (rather than just over the support of μ).

The function $P : \mathcal{J} \to (\mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}})_+$ is given by the following exponentially large convex program over matrix-valued functions,

Program 3.1 (Pseudodistribution Program)**.**

$$\begin{aligned} \min \quad & \|P\|_{Fr,\nu}^2 && (3.1)\\ s.t. \quad & \langle Q, P\rangle_\nu = \langle Q, \Lambda'\rangle_\nu \quad \forall Q : \mathcal{J} \to \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}},\ \deg_{\text{inst}}(Q) \leqslant D && (3.2)\\ & P \succeq 0 \\ & \Lambda' = \Lambda + \eta \cdot \mathrm{Id}, \quad 2^{-2^{2^n}} > \eta > 0 && (3.3) \end{aligned}$$

The constraint (3.2) fixes $\mathbb{E}\,\mathrm{Tr}(P)$, and so the objective function (3.1) can be viewied as minimizing $\mathbb{E}\,\mathrm{Tr}(P^2)$, a proxy for the collision probability of the distribution, which is a measure of entropy.

Remark 3.2. We have perturbed Λ in (3.3) so that we can easily show that strong duality holds in the proof of Claim 3.4. For the remainder of the paper we ignore this perturbation, as we can accumulate the resulting error terms and set η to be small enough so that they can be neglected.

The dual of the above program will allow us to relate the existence of an SoS refutation to the existence of a spectral algorithm.

Program 3.3 (Low-Degree Distinguisher)**.**

$$\begin{aligned} \max \quad & \langle \Lambda, Q\rangle_\nu \\ s.t. \quad & Q : \mathcal{J} \to \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}},\ \deg_{\text{inst}}(Q) \leqslant D \\ & \|Q_+\|_{Fr,\nu}^2 \leqslant 1, \end{aligned}$$

where Q_+ is the projection of Q to the PSD cone.

Claim 3.4. Program 3.3 is a manipulation of the dual of Program 3.1, so that if Program 3.1 has optimum $c > 1$, Program 3.3 as optimum at least $\Omega(\sqrt{c})$.

Before we present the proof of the claim, we summarize its central consequence in the following theorem: if Program 3.1 has a large objective value (and therefore does not provide a feasible SoS solution), then there is a spectral algorithm.

Theorem 3.5. *Fix a function $M : \mathcal{J} \to \mathbb{R}_+^{[n]^{\leqslant d} \times [n]^{\leqslant d}}$ be such that $\mathrm{Id} \succeq M \succeq 0$. Let $\lambda_{\max}^+(\cdot)$ be the function that gives the largest non-negative eigenvalue of a matrix. Suppose $\Lambda = \mu \cdot M$ then the optimum of Program 3.1 is equal to $\mathrm{opt} > 1$ only if there exists a low-degree matrix polynomial Q such that,*

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\mu} [\lambda_{max}^+(Q(\mathcal{I}))] \geqslant \Omega(\sqrt{\mathrm{opt}}/n^d)$$

while,

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} [\lambda_{max}^+(Q(\mathcal{I}))] \leqslant 1\,.$$

Proof. By Claim 3.4, if the value of Program 3.1 is opt > 1, then there is a polynomial Q achieves a value of $\Omega(\sqrt{\text{opt}})$ for the dual. It follows that

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\mu}[\lambda^{+}_{max}(Q(\mathcal{I}))] \geqslant \frac{1}{n^d}\mathop{\mathbb{E}}_{\mathcal{I}\sim\mu}[\langle \mathrm{Id}, Q(\mathcal{I}))\rangle] \geqslant \frac{1}{n^d}\langle \Lambda, Q\rangle_\nu = \Omega(\sqrt{\text{opt}}/n^d),$$

while

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}[\lambda^{+}_{max}(Q(\mathcal{I}))] \leqslant \sqrt{\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}[\lambda^{+}_{\max}(Q(\mathcal{I}))^2]} \leqslant \sqrt{\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\|Q_+(\mathcal{I})\|^2_{Fr}} \leqslant 1.$$

□

It is interesting to note that the specific structure of the PSD matrix valued function M plays no role in the above argument—since M serves as a proxy for monomials in the solution as represented by the program variables $x^{\otimes d}$, it follows that the choice of how to represent the planted solution is not critical. Although seemingly counterintuitive, this is natural because the property of being distinguishable by low-degre distinguishers or by SoS SDP relaxations is a property of ν and μ.

We wrap up the section by presenting a proof of the Claim 3.4.

Proof of Claim 3.4. We take the Lagrangian dual of Program 3.1. Our dual variables will be some combination of low-degree matrix polynomials, Q, and a PSD matrix A:

$$\mathcal{L}(P, Q, A) = \|P\|^2_{Fr,\nu} - \langle Q, P - \Lambda'\rangle_\nu - \langle A, P\rangle_\nu \qquad s.t. \quad A \succeq 0.$$

It is easy to verify that if P is not PSD, then A can be chosen so that the value of $\mathcal{L}$ is ∞. Similarly if there exists a low-degree polynomial upon which P and Λ differ in expectation, Q can be chosen as a multiple of that polynomial so that the value of $\mathcal{L}$ is ∞.

Now, we argue that Slater's conditions are met for Program 3.1, as $P = \Lambda'$ is strictly feasible. Thus strong duality holds, and therefore

$$\min_P \max_{A\succeq 0, Q} \mathcal{L}(P, Q, A) \leqslant \max_{A\succeq 0, Q} \min_P \mathcal{L}(P, Q, A).$$

Taking the partial derivative of $\mathcal{L}(P, Q, A)$ with respect to P, we have

$$\frac{\partial}{\partial P}\mathcal{L}(P, Q, A) = 2 \cdot P - Q - A.$$

where the first derivative is in the space of functions from $\mathcal{I} \to \mathbb{R}^{[n]^{\leqslant d}\times[n]^{\leqslant d}}$. By the convexity of $\mathcal{L}$ as a function of P, it follows that if we set $\frac{\partial}{\partial P}\mathcal{L} = 0$, we will have the minimizer. Substituting, it follows that

$$\begin{aligned}\min_P \max_{A\succeq 0, Q} \mathcal{L}(P, Q, A) &\leqslant \max_{A\succeq 0, Q} \frac{1}{4}\|A + Q\|^2_{Fr,\nu} - \frac{1}{2}\langle Q, A + Q - \Lambda'\rangle_\nu - \frac{1}{2}\langle A, A + Q\rangle_\nu \\ &= \max_{A\succeq 0, Q} \langle Q, \Lambda'\rangle_\nu - \frac{1}{4}\|A + Q\|^2_{Fr,\nu} \qquad (3.4)\end{aligned}$$

Now it is clear that the maximizing choice of A is to set $A = -Q_-$, the negation of the negative-semi-definite projection of Q. Thus (3.4) simplifies to

$$\min_P \max_{A\succeq 0, Q} \mathcal{L}(P, Q, A) \leqslant \max_Q \langle Q, \Lambda'\rangle_\nu - \frac{1}{4}\|Q_+\|^2_{Fr,\nu}$$

$$\leqslant \max_{Q} \langle Q, \Lambda \rangle_\nu + \eta \operatorname{Tr}_\nu(Q_+) - \frac{1}{4}\|Q_+\|^2_{Fr,\nu}, \tag{3.5}$$

where we have used the shorthand $\operatorname{Tr}_\nu(Q_+) \overset{\text{def}}{=} \mathbb{E}_{\mathcal{I}\sim\nu} \operatorname{Tr}(Q(\mathcal{I})_+)$. Now suppose that the low-degree matrix polynomial Q^* achieves a right-hand-side value of

$$\langle Q^*, \Lambda \rangle_\nu + \eta \cdot \operatorname{Tr}_\nu(Q^*_+) - \frac{1}{4}\|Q^*_+\|^2_{Fr,\nu} \geqslant c.$$

Consider $Q' = Q^*/\|Q^*_+\|_{Fr,\nu}$. Clearly $\|Q'_+\|_{Fr,\nu} = 1$. Now, multiplying the above inequality through by the scalar $1/\|Q^*_+\|_{Fr,\nu}$, we have that

$$\begin{aligned}\langle Q', \Lambda \rangle_\nu &\geqslant \frac{c}{\|Q^*_+\|_{Fr,\nu}} - \eta \cdot \frac{\operatorname{Tr}_\nu(Q^*_+)}{\|Q^*_+\|_{Fr,\nu}} + \frac{1}{4}\|Q^*_+\|_{Fr,\nu} \\ &\geqslant \frac{c}{\|Q^*_+\|_{Fr,\nu}} - \eta \cdot n^d + \frac{1}{4}\|Q^*_+\|_{Fr,\nu}.\end{aligned}$$

Therefore $\langle Q', \Lambda \rangle_\nu$ is at least $\Omega(c^{1/2})$, as if $\|Q^*_+\|_{Fr,\nu} \geqslant \sqrt{c}$ then the third term gives the lower bound, and otherwise the first term gives the lower bound.

Thus by substituting Q', the square root of the maximum of (3.5) within an additive ηn^d lower-bounds the maximum of the program

$$\begin{aligned} \max \quad & \langle Q, \Lambda \rangle_\nu \\ s.t. \quad & Q : \mathcal{I} \to \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}}, \quad \deg_{\text{inst}}(Q) \leqslant D \\ & \|Q_+\|^2_{Fr,\nu} \leqslant 1. \end{aligned}$$

This concludes the proof. □

4 Proof of Theorem 2.6

We will prove Theorem 2.6 by contradiction. Let us assume that there exists no degree-$2D$ matrix polynomial that distinguishes ν from μ. First, the lack of distinguishers implies the following fact about scalar polynomials.

Lemma 4.1. *Under the assumption that there are no degree-$2D$ distinguishers, for every degree-D scalar polynomial Q,*

$$\|Q\|^2_{Fr,\mu} \leqslant n^B \|Q\|^2_{Fr,\nu}$$

Proof. Suppose not, then the degree-$2D$ 1×1 matrix polynomial $\operatorname{Tr}(Q(\mathcal{I})^2)$ will be a distinguisher between μ and ν. □

Constructing Λ. First, we will use the robust inference property of μ to construct a pseudo-distribution Λ. Recall again that we have defined $\hat{\mu}$ to be the relative density of μ with respect to ν, so that $\hat{\mu}(\mathcal{I}) = \mu(\mathcal{I})/\nu(\mathcal{I})$. For each subset $S \subseteq [N]$, define a PSD matrix-valued function $\Lambda_S : \mathcal{I} \to (\mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}})_+$ as,

$$\Lambda_S(\mathcal{I}) = \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}}[\hat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})] \cdot x(\mathcal{I}_S)^{\leqslant d}(x(\mathcal{I}_S)^{\leqslant d})^T$$

where we use $\mathcal{I}_S$ to denote the restriction of $\mathcal{I}$ to $S \subset [N]$, and $\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}$ to denote the instance given by completing the sub-instance $\mathcal{I}_S$ with the setting $\mathcal{I}'_{\overline{S}}$. Notice that Λ_S is a function depending only on $\mathcal{I}_S$—this fact will be important to us. Define $\Lambda \stackrel{\text{def}}{=} \mathbb{E}_{S\sim\Theta} \Lambda_S$. Observe that Λ is a PSD matrix-valued function that satisfies

$$\langle \Lambda_{\emptyset,\emptyset}, 1\rangle_\nu = \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} \mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}\sim\nu} [\hat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})] = \mathop{\mathbb{E}}_{S} \mathop{\mathbb{E}}_{\mathcal{I}_{\overline{S}}} \mathop{\mathbb{E}}_{\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}\sim\nu} [\hat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})] = 1 \tag{4.1}$$

Since $\Lambda(\mathcal{I})$ is an average over $\Lambda_S(\mathcal{I})$, each of which is a feasible solution with high probability, $\Lambda(\mathcal{I})$ is close to a feasible solution to the SDP relaxation for $\mathcal{I}$. The following Lemma formalizes this intuition.

Define $\mathcal{G} \stackrel{\text{def}}{=} \text{span}\{\chi_S \cdot G_j \mid j \in [m], S \subseteq [N]\}$, and use $\Pi_{\mathcal{G}}$ to denote the orthogonal projection into $\mathcal{G}$.

Lemma 4.2. *Suppose Program 2.2 satisfies the ε-robust inference property with respect to planted distribution μ and subsampling distribution Θ and if $\|x(\mathcal{I}_S)^{\leqslant d}\|_2^2 \leqslant K$ for all $\mathcal{I}_S$ then for every $G \in \mathcal{G}$, we have*

$$\langle \Lambda, G\rangle_\nu \leqslant \sqrt{\varepsilon} \cdot K \cdot \left(\mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\tilde{\mathcal{I}}_{\overline{S}}\sim\nu} \mathop{\mathbb{E}}_{\mathcal{I}\sim\mu} \|G(\mathcal{I}_S \circ \mathcal{I}_{\overline{S}})\|_2^2 \right)^{1/2}$$

Proof. We begin by expanding the left-hand side by substituting the definition of Λ. We have

$$\begin{aligned}
\langle \Lambda, G\rangle_\nu &= \mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} \langle \Lambda_S(\mathcal{I}_S), G(\mathcal{I})\rangle \\
&= \mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}\sim\nu} \hat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}) \cdot \langle x(\mathcal{I}_S)^{\leqslant d} (x(\mathcal{I}_S)^{\leqslant d})^T, G(\mathcal{I})\rangle
\end{aligned}$$

And because the inner product is zero if $x(\mathcal{I}_S)$ is a feasible solution,

$$\begin{aligned}
&\leqslant \mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}\sim\nu} \hat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}) \cdot \mathbb{I}[x(\mathcal{I}_S) \text{ is infeasible for } \mathcal{S}(\mathcal{I})] \cdot \left\|x(\mathcal{I}_S)^{\leqslant d}\right\|_2^2 \cdot \|G(\mathcal{I})\|_{Fr} \\
&\leqslant \mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}\sim\nu} \hat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}) \cdot \mathbb{I}[x(\mathcal{I}_S) \text{ is infeasible for } \mathcal{S}(\mathcal{I})] \cdot K \cdot \|G(\mathcal{I})\|_{Fr}
\end{aligned}$$

And now letting $\tilde{\mathcal{I}}_{\overline{S}}$ denote the completion of $\mathcal{I}_S$ to $\mathcal{I}$, so that $\mathcal{I}_S \circ \tilde{\mathcal{I}}_{\overline{S}} = \mathcal{I}$, we note that the above is like sampling $\mathcal{I}'_{\overline{S}}, \tilde{\mathcal{I}}_{\overline{S}}$ independently from ν and then reweighting by $\hat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})$, or equivalently taking the expectation over $\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}} = \mathcal{I}' \sim \mu$ and $\tilde{\mathcal{I}}_{\overline{S}} \sim \nu$:

$$= \mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\mathcal{I}'\sim\mu} \mathop{\mathbb{E}}_{\tilde{\mathcal{I}}_{\overline{S}}\sim\nu} \cdot \mathbb{I}[x(\mathcal{I}_S) \text{ is infeasible for } \mathcal{S}(\mathcal{I}_S \circ \tilde{\mathcal{I}}_{\overline{S}})] \cdot K \cdot \|G(\mathcal{I}_S \circ \tilde{\mathcal{I}}_S)\|_{Fr}$$

and by Cauchy-Schwarz,

$$\leqslant K \cdot \left(\mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\mathcal{I}'\sim\mu} \mathop{\mathbb{E}}_{\tilde{\mathcal{I}}_{\overline{S}}\sim\nu} \cdot \mathbb{I}[x(\mathcal{I}_S) \text{ is infeasible for } \mathcal{S}(\mathcal{I}_S \circ \tilde{\mathcal{I}}_{\overline{S}})] \right)^{1/2} \cdot \left(\mathop{\mathbb{E}}_{S\sim\Theta} \mathop{\mathbb{E}}_{\mathcal{I}'\sim\mu} \mathop{\mathbb{E}}_{\tilde{\mathcal{I}}_{\overline{S}}\sim\nu} \|G(\mathcal{I}_S \circ \tilde{\mathcal{I}}_S)\|_{Fr}^2 \right)^{1/2}$$

The lemma follows by observing that the first term in the product above is exactly the non-robustness of inference probability ε. □

Corollary 4.3. *If $G \in \mathcal{G}$ is a degree-D polynomial in $\mathcal{I}$, then under the assumption that there are no degree-$2D$ distinguishers for ν, μ,*

$$\langle \Lambda, G \rangle_\nu \leqslant \sqrt{\varepsilon} \cdot K \cdot n^B \cdot \|G\|_{Fr,\nu}$$

Proof. For each fixing of $\tilde{\mathcal{I}}_S$, $\|G(\mathcal{I}_S \circ \tilde{\mathcal{I}}_S)\|_2^2$ is a degree-$2D$-scalar polynomial in $\mathcal{I}$. Therefore by Lemma 4.1 we have that,

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\mu} \|G(\mathcal{I}_S \circ \tilde{\mathcal{I}}_S)\|_{Fr}^2 \leqslant n^B \cdot \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} \|G(\mathcal{I}_S \circ \tilde{\mathcal{I}}_S)\|_{Fr}^2 .$$

Substituting back in the bound in Lemma 4.2 the corollary follows. □

Now, since there are no degree-D matrix distinguishers Q, for each S in the support of Θ we can apply reasoning similar to Theorem 3.5 to conclude that there is a high-entropy PSD matrix-valued function P_S that matches the degree-D moments of Λ_S.

Lemma 4.4. *If there are no degree-D matrix distinguishers Q for μ, ν, then for each $S \sim \Theta$, there exists a solution P_S to Program 3.1 (with the variable $\Lambda := \Lambda_S$) and*

$$\|P_S\|_{Fr,\nu} \leqslant n^{(B+d)/4} \leqslant n^{B/2} \tag{4.2}$$

This does not follow directly from Theorem 3.5, because a priori a distinguisher for some specific S may only apply to a small fraction of the support of μ. However, we can show that Program 3.1 has large value for Λ_S only if there is a distinguisher for μ, ν.

Proof. By Claim 3.4, it suffices for us to argue that there is no degree-D matrix polynomial Q which has large inner product with Λ_S relative to its Frobenius norm. So, suppose by way of contradiction that Q is a degree-D matrix that distinguishes Λ_S, so that $\langle Q, \Lambda_S \rangle_\nu \geqslant n^{B+d}$ but $\|Q\|_{Fr,\nu} \leqslant 1$.

It follows by definition of Λ_S that

$$\begin{aligned} n^{B+d} \leqslant \langle Q, \Lambda_S \rangle_\nu &= \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu} \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}\sim\nu} \hat{\mu}(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}) \cdot \langle Q(\mathcal{I}), x(\mathcal{I}_S)^{\leqslant d} (x(\mathcal{I}_S)^{\leqslant d})^\top \rangle \\ &= \mathop{\mathbb{E}}_{\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}}\sim\mu} \left\langle \mathop{\mathbb{E}}_{\mathcal{I}_{\overline{S}}\sim\nu} Q(\mathcal{I}_S \circ \mathcal{I}_{\overline{S}}), x(\mathcal{I}_S)^{\leqslant d} (x(\mathcal{I}_S)^{\leqslant d})^\top \right\rangle \\ &\leqslant \mathop{\mathbb{E}}_{\mu} \left[\lambda^+_{\max} \left(\mathop{\mathbb{E}}_{\mathcal{I}_{\overline{S}}\sim\nu} Q(\mathcal{I}_S \circ \mathcal{I}_{\overline{S}}) \right) \right] \cdot \left\| x(\mathcal{I}_S)^{\leqslant d} \right\|_2^2 . \end{aligned}$$

So, we will show that $Q_S(\mathcal{I}) = \mathbb{E}_{\mathcal{I}'_{\overline{S}}\sim\nu} Q(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})$ is a degree-D distinguisher for μ. The degree of Q_S is at most D, since averaging over settings of the variables cannot increase the degree. Applying our assumption that $\|x(\mathcal{I}_S)^{\leqslant d}\|_2^2 \leqslant K \leqslant n^d$, we already have $\mathbb{E}_\mu \lambda^+_{\max}(Q_S) > n^B$. It remains to show that $\mathbb{E}_\nu \lambda^+_{\max}(Q_S)$ is bounded. For this, we use the following fact about the trace.

Fact 4.5 (See e.g. Theorem 2.10 in [CC09])**.** *For a function $f : \mathbb{R} \to \mathbb{R}$ and a symmetric matrix A with eigendecomposition $\sum \lambda \cdot vv^\top$, define $f(A) = \sum f(\lambda) \cdot vv^\top$. If $f : \mathbb{R} \to \mathbb{R}$ is continuous and convex, then the map $A \to \mathrm{Tr}(f(A))$ is convex for symmetric A.*

The function $f(t) = (\max\{0, t\})^2$ is continuous and convex over $\mathbb{R}$, so the fact above implies that the map $A \to \|A_+\|_{Fr}^2$ is convex for symmetric A. We can take Q_S to be symmetric without loss of generality, as in the argument above we only consider the inner product of Q_S with symmetric matrices. Now we have that

$$\|(Q_S(I))_+\|_{Fr}^2 = \left\|\left(\mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}}\left[Q(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})\right]\right)_+\right\|_{Fr}^2 \leqslant \mathop{\mathbb{E}}_{\mathcal{I}'_{\overline{S}}}\left\|\left(Q(\mathcal{I}_S \circ \mathcal{I}'_{\overline{S}})\right)_+\right\|_{Fr}^2,$$

where the inequality is the definition of convexity. Taking the expectation over $\mathcal{I} \sim \nu$ gives us that $\|(Q_S)_+\|_{Fr,\nu}^2 \leqslant \|Q_+\|_{Fr,\nu}^2 \leqslant 1$, which gives us our contradiciton. □

Now, analogous to Λ, set $P \stackrel{\text{def}}{=} \mathbb{E}_{S\sim\Theta} P_S$.

Random Restriction. We will exploit the crucial property that Λ and P are averages over functions that depend on subsets of variables. This has the same effect as a random restriction, in that $\langle P, R\rangle_\nu$ essentially depends on the low-degree part of R. Formally, we will show the following lemma.

Lemma 4.6. *(Random Restriction) Fix $D, \ell \in \mathbb{N}$. For matrix-valued functions $R : \mathcal{I} \to \mathbb{R}^{\ell\times\ell}$ and a family of functions $\{P_S : \mathcal{I}_S \to \mathbb{R}^{\ell\times\ell}\}_{S\subseteq[N]}$, and a distribution Θ over subsets of $[N]$,*

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), R(\mathcal{I})\rangle \geqslant \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), R_S^{<D}(\mathcal{I}_S)\rangle - \rho(D,\Theta)^{1/2}\cdot\left(\mathop{\mathbb{E}}_{S\sim\Theta}\|P_S\|_{Fr,\nu}^2\right)^{\frac{1}{2}}\|R\|_{Fr,\nu}$$

where

$$\rho(D,\Theta) = \max_{\alpha,|\alpha|\geqslant D}\mathop{\mathbb{P}}_{S\sim\Theta}[\alpha \subseteq S].$$

Proof. We first re-express the left-hand side as

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), R(\mathcal{I})\rangle = \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), R_S(\mathcal{I}_S)\rangle$$

where $R_S(\mathcal{I}_S) \stackrel{\text{def}}{=} \mathbb{E}_{\mathcal{I}_{\overline{S}}}[R(\mathcal{I})]$ obtained by averaging out all coordinates outside S. Splitting the function R_S into its low-degree and high-degree parts, $R_S = R_S^{\leqslant D} + R_S^{>D}$, then applying a Cauchy-Schwartz inequality we get

$$\mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), R_S(\mathcal{I}_S)\rangle \geqslant \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), R_S^{<D}(\mathcal{I}_S)\rangle - \left(\mathop{\mathbb{E}}_{S\sim\Theta}\|P_S\|_{Fr,\nu}^2\right)^{1/2}\cdot\left(\mathop{\mathbb{E}}_{S\sim\Theta}\|R_S^{\geqslant D}\|_{Fr,\nu}^2\right)^{1/2}.$$

Expressing $R^{\geqslant D}(\mathcal{I})$ in the Fourier basis, we have that over a random choice of $S \sim \Theta$,

$$\mathop{\mathbb{E}}_{S\sim\Theta}\|R_S^{\geqslant D}\|_{Fr,\nu}^2 = \sum_{\alpha,|\alpha|\geqslant D}\mathop{\mathbb{P}}_{S\sim\Theta}[\alpha\subseteq S]\cdot\hat{R}_\alpha^2 \leqslant \rho(D,\Theta)\cdot\|R\|_{Fr}^2$$

Substituting into the above inequality, the conclusion follows. □

Equality Constraints. Since Λ is close to satisfying all the equality constraints $\mathcal{G}$ of the SDP, the function P approximately satisfies the low-degree part of $\mathcal{G}$. Specifically, we can prove the following.

Lemma 4.7. *Let $k \geqslant \deg_{\text{inst}}(G_j)$ for all $G_j \in \mathcal{S}$. With P defined as above and under the conditions of Theorem 2.6 for any function $G \in \mathcal{G}$,*

$$\left|\langle P, G^{\leqslant D}\rangle_v\right| \leqslant \frac{2}{n^{2B}}\|G\|_{Fr,v}$$

Proof. Recall that $\mathcal{G} = \text{span}\{\chi_S \cdot G_j \mid j \in [m], S \subseteq [N]\}$ and let $\Pi_{\mathcal{G}}$ be the orthogonal projection into $\mathcal{G}$. Now, since $G \in \mathcal{G}$,

$$G^{\leqslant D} = (\Pi_{\mathcal{G}} G)^{\leqslant D} = (\Pi_{\mathcal{G}} G^{\leqslant D-2k})^{\leqslant D} + (\Pi_{\mathcal{G}} G^{>D-2k})^{\leqslant D}. \tag{4.3}$$

Now we make the following claim regarding the effect of projection on to the ideal $\mathcal{G}$, on the degree of a polynomial.

Claim 4.8. For every polynomial Q, $\deg(\Pi_{\mathcal{G}} Q) \leqslant \deg(Q) + 2k$. Furthermore for all α, $\Pi_{\mathcal{G}} Q^{>\alpha}$ has no monomials of degree $\leqslant \alpha - k$

Proof. To establish the first part of the claim it suffices to show that $\Pi_{\mathcal{G}} Q \in \text{span}\{\chi_S \cdot G_j \mid |S| \leqslant \deg(Q) + k\}$, since $\deg(G_j) \leqslant k$ for all $j \in [m]$. To see this, observe that $\Pi_{\mathcal{G}} Q \in \text{span}\{\chi_S \cdot G_j \mid |S| \leqslant \deg(Q) + k\}$ and is orthogonal to every $\chi_S \cdot G_j$ with $|S| > \deg(Q) + k$:

$$\langle \Pi_{\mathcal{G}} Q, \chi_S \cdot G_j\rangle_v = \langle Q, \Pi_{\mathcal{G}} \chi_S \cdot G_j\rangle_v = \langle Q, \chi_S \cdot G_j\rangle_v = \langle Q G_j, \chi_S\rangle_v = 0,$$

where the final equality is because $\deg(\chi_S) > \deg(G_j) + \deg(Q)$. On the other hand, for every subset S with $\deg(\chi_S) \leqslant \alpha - k$,

$$\langle \Pi_{\mathcal{G}} Q^{>\alpha}, \chi_S \cdot G_j\rangle = \langle Q^{>\alpha}, \Pi_{\mathcal{G}} \chi_S \cdot G_j\rangle = \langle Q^{>\alpha}, \chi_S \cdot G_j\rangle = 0, \text{ since } \alpha > \deg(G_j) + \deg(\chi_S)$$

This implies that $\Pi_{\mathcal{G}} Q^{>\alpha} \in \text{span}\{\chi_S \cdot G_j \mid |S| > \alpha - k\}$ which implies that $\Pi_{\mathcal{G}} Q^{>\alpha}$ has no monomials of degree $\leqslant \alpha - k$. □

Incorporating the above claim into (4.3), we have that

$$G^{\leqslant D} = \Pi_{\mathcal{G}} G^{\leqslant D-2k} + (\Pi_{\mathcal{G}} G^{\geqslant D-2k})^{[D-3k,D]},$$

where the superscript $[D-3k, D]$ denotes the degree range. Now,

$$\langle P, G^{\leqslant D}\rangle_v = \langle P, \Pi_{\mathcal{G}} G^{\leqslant D-2k}\rangle_v + \langle P, (\Pi_{\mathcal{G}} G^{\geqslant D-2k})^{[D-3k,D]}\rangle_v$$

And since $\Pi_{\mathcal{G}} G^{\leqslant D-2k}$ is of degree at most D we can replace P by Λ,

$$= \langle \Lambda, \Pi_{\mathcal{G}} G^{\leqslant D-2k}\rangle_v + \langle P, (\Pi_{\mathcal{G}} G^{\geqslant D-2k})^{[D-3k,D]}\rangle_v$$

Now bounding the first term using Corollary 4.3 with a n^B bound on K,

$$\leqslant \left(\frac{1}{n^{8B}}\right)^{1/2} \cdot n^B \cdot (n^B \cdot \|\Pi_{\mathcal{G}} G^{\leqslant D-2k}_{\emptyset,\emptyset}\|_{Fr,v}) + \langle P, (\Pi_{\mathcal{G}} G^{\geqslant D-2k})^{[D-3k,D]}\rangle$$

And for the latter term we use Lemma 4.6,

$$\leqslant \frac{1}{n^{2B}}\|\Pi_{\mathcal{G}} G^{\leqslant D-2k}_{\emptyset,\emptyset}\|_{Fr,\nu} + \frac{1}{n^{4B}}\left(\mathop{\mathbb{E}}_{S}\|P_S\|^2_{Fr,\nu}\right)^{1/2}\|G\|_{Fr,\nu},$$

where we have used the fact that $(\Pi_{\mathcal{G}} G^{\geqslant D-2k})^{[D-3k,D]}$ is high degree. By property of orthogonal projections, $\|\Pi_{\mathcal{G}} G^{\geqslant D-2k}\|_{Fr,\nu} \leqslant \|G^{\geqslant D-2k}\|_{Fr,\nu} \leqslant \|G\|_{Fr,\nu}$. Along with the bound on $\|P_S\|_{Fr,\nu}$ from (4.2), this implies the claim of the lemma. □

Finally, we have all the ingredients to complete the proof of Theorem 2.6.

Proof of Theorem 2.6. Suppose we sample an instance $\mathcal{I} \sim \nu$, and suppose by way of contradiction this implies that with high probability the SoS SDP relaxation is infeasible. In particular, this implies that there is a degree-d sum-of-squares refutation of the form,

$$-1 = a^{\mathcal{I}}(x) + \sum_{j\in[m]} g^{\mathcal{I}}_j(x)\cdot q^{\mathcal{I}}_j(x),$$

where $a^{\mathcal{I}}$ is a sum-of-squares of polynomials of degree at most $2d$ in x, and $\deg(q^{\mathcal{I}}_j) + \deg(g^{\mathcal{I}}_j) \leqslant 2d$. Let $A^{\mathcal{I}} \in \mathbb{R}^{[n]^{\leqslant d}\times[n]^{\leqslant d}}$ be the matrix of coefficients for $a^{\mathcal{I}}(c)$ on input $\mathcal{I}$, and let $G^{\mathcal{I}}$ be defined similarly for $\sum_{j\in[m]} g_j(x)\cdot q_j(x)$. We can rewrite the sum-of-squares refutation as a matrix equality,

$$-1 = \langle X^{\leqslant d}, A^{\mathcal{I}}\rangle + \langle X^{\leqslant d}, G^{\mathcal{I}}\rangle,$$

where $G^{\mathcal{I}} \in \mathcal{G}$, the span of the equality constraints of the SDP.

Define $s : \mathcal{I} \to \{0,1\}$ as

$$s(\mathcal{I}) \stackrel{\text{def}}{=} \mathbb{I}[\exists \text{ a degree-}2d \text{ sos-refutation for } \mathcal{S}(\mathcal{I})]$$

By assumption, $\mathbb{E}_{\mathcal{I}\sim\nu}[s(\mathcal{I})] = 1 - \frac{1}{n^{8B}}$. Define matrix valued functions $A, G : \mathcal{I} \to \mathbb{R}^{[n]^{\leqslant d}\times[n]^{\leqslant d}}$ by setting,

$$A(\mathcal{I}) \stackrel{\text{def}}{=} s(\mathcal{I})\cdot A^{\mathcal{I}}$$
$$G(\mathcal{I}) \stackrel{\text{def}}{=} s(\mathcal{I})\cdot G^{\mathcal{I}}$$

With this notation, we can rewrite the sos-refutation identity as a polynomial identity in X and $\mathcal{I}$,

$$-s(\mathcal{I}) = \langle X^{\leqslant d}, A(\mathcal{I})\rangle + \langle X^{\leqslant d}, G(\mathcal{I})\rangle\,.$$

Let $\mathbf{e}_{\emptyset,\emptyset}$ denote the $[n]^{\leqslant d}\times[n]^{\leqslant d}$ matrix with the entry corresponding to $(\emptyset,\emptyset)$ equal to 1, while the remaining entries are zero. We can rewrite the above equality as,

$$-\langle X^{\leqslant d}, s(\mathcal{I})\cdot\mathbf{e}_{\emptyset,\emptyset}\rangle = \langle X^{\leqslant d}, A(\mathcal{I})\rangle + \langle X^{\leqslant d}, G(\mathcal{I})\rangle\,.$$

for all $\mathcal{I}$ and formal variables X.

Now, let $P = \mathbb{E}_{S\sim\Theta} P_S$ where each P_S is obtained by from the Program 3.1 with Λ_S. Substituting $X^{\leqslant d}$ with $P(\mathcal{I})$ and taking an expectation over $\mathcal{I}$,

$$\langle P, s(\mathcal{I})\cdot\mathbf{e}_{\emptyset,\emptyset}\rangle_\nu = \langle P, A\rangle_\nu + \langle P, G\rangle_\nu \tag{4.4}$$

$$\geqslant \langle P, G\rangle_\nu \tag{4.5}$$

where the inequality follows because $A, P \succeq 0$. We will show that the above equation is a contradiction by proving that LHS is less than -0.9, while the right hand side is at least -0.5. First, the right hand side of (4.4) can be bounded by Lemma 4.7

$$\begin{aligned}
\langle P, G\rangle_\nu &= \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), G(\mathcal{I})\rangle \\
&\geqslant \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), G^{\leqslant D}(\mathcal{I})\rangle - \frac{1}{n^{4B}}\cdot\left(\mathop{\mathbb{E}}_{S}\|P_S\|_{Fr,\nu}^2\right)^{1/2}\cdot\|G\|_{Fr,\nu} \qquad \text{(random restriction Lemma 4.6)} \\
&\geqslant -\frac{2}{n^{2B}}\cdot\|G\|_{Fr,\nu} - \frac{1}{n^{4B}}\left(\mathop{\mathbb{E}}_{S}\|P_S\|_{Fr,\nu}^2\right)^{1/2}\|G\|_{Fr,\nu} \qquad \text{(using Lemma 4.7)} \\
&\geqslant -\frac{1}{2}
\end{aligned}$$

where the last step used the bounds on $\|P_S\|_{Fr,\nu}$ from (4.2) and on $\|G\|_{Fr,\nu}$ from the n^B bound assumed on the SoS proofs in Theorem 2.6.

Now the negation of the left hand side of (4.4) is

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P(\mathcal{I}), s(\mathcal{I})\cdot\mathbf{e}_{\emptyset,\emptyset}\rangle \geqslant \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}[P_{\emptyset,\emptyset}(\mathcal{I})\cdot 1] - \mathbb{E}[(s-1)^2]^{1/2}\cdot\|P\|_{Fr,\nu}$$

The latter term can be simplified by noticing that the expectation of the square of a 0,1 indicator is equal to the expectation of the indicator, which is in this case $\frac{1}{n^{8B}}$ by assumption. Also, since 1 is a constant, $P_{\emptyset,\emptyset}$ and $\Lambda_{\emptyset,\emptyset}$ are equivalent:

$$\begin{aligned}
&= \mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}[\Lambda_{\emptyset,\emptyset}(\mathcal{I})\cdot 1] - \frac{1}{n^{4B}}\cdot\|P\|_{Fr,\nu} \\
&= 1 - \frac{1}{n^{4B}}\cdot\|P\|_{Fr,\nu} \qquad \text{(using (4.1))} \\
&= 1 - \frac{1}{n^{3B}} \qquad \text{(using (4.2))}
\end{aligned}$$

We have the desired contradiction in (4.4). □

4.1 Handling Inequalities

Suppose the polynomial system Program 2.2 includes inequalities of the form $h(\mathcal{I}, x) \geqslant 0$, then a natural approach would be to introduce a slack variable z and set $h(\mathcal{I}, x) - z^2 = 0$. Now, we can view the vector (x, z) consisting of the original variables along with the slack variables as the hidden planted solution. The proof of Theorem 2.6 can be carried out as described earlier in this section, with this setup. However, in many cases of interest, the inclusion of slack variables invalidates the robust inference property. This is because, although a feasible solution x can be recovered from a subinstance $\mathcal{I}_S$, the value of the corresponding slack variables could potentially depend on $\mathcal{I}_{\overline{S}}$. For instance, in a random CSP, the value of the objective function on the assignment x generated from $\mathcal{I}_S$ depends on all the constraints outside of S too.

The proof we described is to be modified as follows.

- As earlier, construct Λ_S using only the robust inference property of original variables x, and the corresponding matrix functions P_S.
- Convert each inequality of the form $h_i(\mathcal{I}, x) \geqslant 0$, in to an equality by setting $h_i(\mathcal{I}, x) = z_i^2$.
- Now we define a pseudo-distribution $\tilde{\Lambda}_S(\mathcal{I}_S)$ over original variables x and slack variables z as follows. It is convenient to describe the pseudo-distribution in terms of the corresponding pseudo-expectation operator. Specifically, if $x(\mathcal{I}_S)$ is a feasible solution for Program 2.2 then define

$$\tilde{E}[z_\sigma x_\alpha] \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } \sigma_i \text{ odd for some } i \\ \prod_{i\in\sigma}(h_i(\mathcal{I}, x(\mathcal{I}_S)))^{\sigma_i/2} \cdot x(\mathcal{I}_S)_\alpha & \text{otherwise} \end{cases}$$

  Intuitively, the pseudo-distribution picks the sign for each z_i uniformly at random, independent of all other variables. Therefore, all moments involving an odd power of z_i are zero. On the other hand, the moments of even powers of z_i are picked so that the equalities $h_i(\mathcal{I}, x) = z_i$ are satisfied.

  It is easy to check that $\tilde{\Lambda}$ is psd matrix valued, satisfies (4.1) and all the equalities.
- While Λ_S in the original proof was a function of $\mathcal{I}_S$, $\tilde{\Lambda}_S$ is not. However, the key observation is that, $\tilde{\Lambda}_S$ is degree at most $k \cdot d$ in the variables outside of S. Each function $h_i(\mathcal{I}, x(\mathcal{I}_S))$ is degree at most k in $\mathcal{I}_{\overline{S}}$, and the entries of $\tilde{\Lambda}_S(\mathcal{I}_S)$ are a product of at most d of these polynomials.
- The main ingredient of the proof that is different from the case of equalities is the random restriction lemma which we outline below. The error in the random restriction is multiplied by $D^{dk/2} \leqslant n^{B/2}$; however this does not substantially change our results, since Theorem 2.6 requires $\rho(D, \Theta) < n^{-8B}$, which leaves us enough slack to absorb this factor (and in every application $\rho(D, \Theta) = p^{O(D)}$for some $p < 1$ sufficiently small that we meet the requirement that $D^{dk}\rho(D - dk, \Theta)$ is monotone non-increasing in D).

Lemma 4.9. *[Random Restriction for Inequalities] Fix $D, \ell \in \mathbb{N}$. Consider a matrix-valued function $R : \mathcal{I} \to \mathbb{R}^{\ell\times\ell}$ and a family of functions $\{P_S : \mathcal{I} \to \mathbb{R}^{\ell\times\ell}\}_{S\subseteq[N]}$ such that each P_S has degree at most dk in $\mathcal{I}_{\overline{S}}$. If Θ is a distribution over subsets of $[N]$ with*

$$\rho(D, \Theta) = \max_{\alpha, |\alpha|\geqslant D} \mathbb{P}_{S\sim\Theta}[\alpha \subseteq S],$$

and the additional requirement that $D^{dk} \cdot \rho(D - dk, \Theta)$ is monotone non-increasing in D, then

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), R(\mathcal{I})\rangle \geqslant \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), \tilde{R}_S^{<D}(\mathcal{I}_S)\rangle - D^{dk/2} \cdot \rho(D - dk, \Theta)^{1/2} \cdot \left(\mathop{\mathbb{E}}_{S\sim\Theta}\|P_S\|_{2,\nu}^2\right)^{\frac{1}{2}} \|R\|_{Fr,\nu}$$

Proof.

$$\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\mathop{\mathbb{E}}_{S\sim\Theta}\langle P_S(\mathcal{I}_S), R(\mathcal{I})\rangle = \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), \tilde{R}_S(\mathcal{I})\rangle$$

where $\tilde{R}_S(\mathcal{I})$ is now obtained by averaging out the values for all monomials whose degree in $\overline{S}$ is $> dk$. Writing $\tilde{R}_S = \tilde{R}_S^{\leqslant D} + \tilde{R}_S^{>D}$ and applying a Cauchy-Schwartz inequality we get,

$$\mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), \tilde{R}_S(\mathcal{I})\rangle \geqslant \mathop{\mathbb{E}}_{S\sim\Theta}\mathop{\mathbb{E}}_{\mathcal{I}\sim\nu}\langle P_S(\mathcal{I}_S), \tilde{R}_S^{<D}(\mathcal{I})\rangle - \left(\mathop{\mathbb{E}}_{S\sim\Theta}\|P_S\|_{Fr,\nu}^2\right)^{1/2} \cdot \left(\mathop{\mathbb{E}}_{S\sim\Theta}\|\tilde{R}_S^{\geqslant D}\|_{Fr,\nu}\right)^{1/2}$$

Over a random choice of S,

$$\mathop{\mathbb{E}}_{S\sim\Theta}\|\tilde{R}_S^{\geqslant D}\|_{Fr,\nu}^2 = \sum_{\alpha,|\alpha|\geqslant D} \mathop{\mathbb{P}}_{S\sim\Theta}[|\alpha\cap\overline{S}|\leqslant dk]\cdot\hat{R}_\alpha^2 \leqslant D^{dk}\cdot\rho(D-dk,\Theta)\cdot\|R\|_{Fr}^2,$$

where we have used that $D^{dk}\rho(D-dk,\Theta)$ is a monotone non-increasing function of D. Substituting this in the earlier inequality the Lemma follows. □

5 Applications to Classical Distinguishing Problems

In this section, we verify that the conditions of Theorem 2.6 hold for a variety of canonical distinguishing problems. We'll rely upon the (simple) proofs in Appendix A, which show that the ideal term of the SoS proof is well-conditioned.

Problem 5.1 (Planted clique with clique of size n^δ)**.** Given a graph $G = (V, E)$ on n vertices, determine whether it comes from:

- **Uniform Distribution**: the uniform distribution over graphs on n vertices ($G(n, \frac{1}{2})$).
- **Planted Distribution**: the uniform distribution over n-vertex graphs with a clique of size at least n^δ

The usual polynomial program for *planted clique* in variables $x_1, \ldots, x_n$ is:

$$\begin{aligned}\text{obj} &\leqslant \sum_i x_i\\ x_i^2 &= x_i \quad \forall i\in[n]\\ x_i x_j &= 0 \quad \forall (i,j)\in E\end{aligned}$$

Lemma 5.2. *Theorem 2.6 applies to the above planted clique program, so long as* obj $\leqslant n^{\delta-\varepsilon}$ *for any* $\varepsilon \geqslant \frac{c\cdot d}{D-6d}$ *for a fixed constant* c.

Proof. For planted clique, for our notion of "instance degree", rather than the multiplicity of instance variables, the "degree" of $\mathcal{I}_\alpha$ will be the number of distinct vertices incident on the edges in α. The proof of Theorem 2.6 proceeds identically with this notion of degree, but we will be able to achieve better bounds on D relative to d.

In this case, the instance degree of the SoS relaxation is $k = 2$. We have from Corollary A.3 that the degree-d SoS refutation is well-conditioned, with numbers bounded by $n^{c_1\cdot d}$ for some constant $c_1/2$. Define $B = c_1 d \geqslant dk$.

Our subsampling distribution Θ is the distribution given by including every vertex with probability ρ, producing an induced subgraph of $\approx \rho n$ vertices. For any set of edges α of instance degree at most $D - 6d$,

$$\mathop{\mathbb{P}}_{S\sim\Theta}[\alpha\subseteq S]\leqslant\rho^{D-6d},$$

since the instance degree corresponds to the number of vertices incident on α.

This subsampling operation satisfies the subsample inference condition for the clique constraints with probability 1, since a clique in any subgraph of G is also a clique in G. Also, if there is a clique of size n^δ in G, then by a Chernoff bound

$$\mathop{\mathbb{P}}_{S\sim\Theta}[\exists \text{ clique of size } \geqslant(1-\beta)\rho n^\delta\in S]\geqslant 1-\exp(-\frac{\beta^2\rho n^\delta}{2}).$$

Choosing $\beta = \sqrt{\frac{10B \log n}{\rho n^{\delta}}}$, this gives us that Θ gives n^{-10B}-robust inference for the planted clique problem, so long as obj $\leqslant \rho n/2$. Choosing $\rho = n^{-\varepsilon}$ for ε so that

$$\rho^{D-6d} \leqslant n^{-8B} \implies \varepsilon \geqslant \frac{c_2 d}{D-6d},$$

for some constant c_2, all of the conditions required by Theorem 2.6 now hold. □

Problem 5.3 (Random CSP Refutation at clause density α)**.** Given an instance of a Boolean k-CSP with predicate $P : \{\pm 1\}^k \to \{\pm 1\}$ on n variables with clause set C, determine whether it comes from:

- **Uniform Distribution**: $m \approx \alpha n$ constraints are generated as follows. Each k-tuple of variables $S \in [n]^k$ is independently with probability $p = \alpha n^{-k+1}$ given the constraint $P(x_S \circ z_S) = b_S$ (where $\circ$ is the entry-wise multiplication operation) for a uniformly random $z_S \in \{\pm 1\}^k$ and $b_S \in \{\pm 1\}$.
- **Planted Distribution**: a planted solution $y \in \{\pm 1\}^n$ is chosen, and then $m \approx \alpha n$ constraints are generated as follows. Each k-tuple of variables $S \in [n]^k$ is independently with probability $p = \alpha n^{-k+1}$ given the constraint $P(x_S \circ z_S) = b_S$ for a uniformly random $z_S \in \{\pm 1\}^k$, but $b_S = P(y_S \circ z_S)$ with probability $1 - \delta$ and b_S is uniformly random otherwise.

The usual polynomial program for *random CSP refutation* in variables $x_1, \ldots, x_n$ is:

$$\text{obj} \leqslant \sum_{S \in [n]^k} \mathbb{I}[\exists \text{ constraint on } S] \cdot \left(\frac{1 + P(x_S \circ z_S) \cdot b_S}{2} \right)$$
$$x_i^2 = 1 \quad \forall i \in [n]$$

Lemma 5.4. *If $\alpha \geqslant 1$, then Theorem 2.6 applies to the above* random k-CSP refutation *problem, so long as* obj $\leqslant (1 - \delta - \varepsilon)m$ *for any $\varepsilon \geqslant \frac{c \cdot d \log n}{D-3d}$, where c is a fixed constant.*

Proof. In this case, the instance degree of the SoS relaxation $k = 1$. We have from Corollary A.3 that the degree-d SoS refutation is well-conditioned, with numbers bounded by $n^{c_1 d}$ for some constant c_1. Define $B = c_1 d$.

Our subsampling distribution Θ is the distribution given by including each constraint independently with probability ρ, producing an induced CSP instance on n variables with approximately ρm constraints. Since each constraint survives the subsampling with probability ρ, for any $\alpha \in \binom{C}{D-3d}$,

$$\mathbb{P}_{S \sim \Theta}[\alpha \subseteq S] \leqslant \rho^{D-3d}.$$

The subsample inference property clearly holds for the boolean constraints $\{x_i^2 = 1\}_{i \in [n]}$, as a Boolean assignment to the variables is valid regardless of the number of constraints. Before subsampling there are at least $(1 - \delta)m$ satisfied constraints, and so letting O_S be the number of constraints satisfied in sub-instance S, we have by a Chernoff bound

$$\mathbb{P}_{S \sim \Theta}[O_S \geqslant (1 - \beta) \cdot \rho(1 - \delta)m] \geqslant 1 - \exp\left(-\frac{\beta^2 \rho (1 - \delta) m}{2} \right).$$

Choosing $\beta = \sqrt{\frac{10B \log n}{\rho(1-\delta)m}} = o(1)$ (with overwhelming probability since we have $\alpha \geqslant 1 \implies \mathbb{E}[m] \geqslant n$), we have that Θ gives us n^{-10B}-robust inference for the random CSP refutation problem, so long as $\text{obj} \leqslant (1 - o(1))\rho(1-\delta)m$. Choosing $\rho = (1-\varepsilon)$ so that

$$\rho^{D-3d} \leqslant n^{-8B} \implies \varepsilon \geqslant \frac{c_2 d \log n}{D - 3d},$$

for some constant c_2. The conclusion follows (after making appropriate adjustments to the constant). □

Problem 5.5 (Community detection with average degree d (stochastic block model))**.** Given a graph $G = (V, E)$ on n vertices, determine whether it comes from:

- **Uniform Distribution**: $G(n, b/n)$, the distribution over graphs in which each edge is included independently with probability b/n.
- **Planted Distribution**: the stochastic block model—there is a partition of the vertices into two equally-sized sets, Y and Z, and the edge (u, v) is present with probability a/n if $u, v \in Y$ or $u, v \in Z$, and with probability $(b-a)/n$ otherwise.

Letting $x_1, \ldots, x_n$ be variables corresponding to the membership of each vertex's membership, and let A be the adjacency of the graph. The canonical polynomial optimization problem is

$$\begin{aligned} \text{obj} &\leqslant x^\top A x \\ x_i^2 &= 1 \qquad \forall i \in [n] \\ \sum_i x_i &= 0. \end{aligned}$$

Lemma 5.6. *Theorem 2.6 applies to the* community detection *problem so long as* $\text{obj} \leqslant (1-\varepsilon)\frac{(2a-b)}{4}n$*, for* $\varepsilon > \frac{c \cdot d \log n}{D-3d}$ *where* c *is a fixed constant.*

Proof. The degree of the SoS relaxation in the instance is $k = 1$. Since we have only hypercube and balancedness constraints, we have from Corollary A.3 that the SoS ideal matrix is well-conditioned, with no number in the SoS refutation larger than $n^{c_1 d}$ for some constant c_1. Let $B = c_1 d$.

Consider the solution x which assigns $x_i = 1$ to $i \in Y$ and $x_i = -1$ to $i \in Z$. Our subsampling operation is to remove every edge independently with probability $1 - \rho$. The resulting distribution Θ and the corresponding restriction of x clearly satisfies the Booleanity and balancedness constraints with probability 1. Since each edge is included independently with probability ρ, for any $\alpha \in \binom{E}{D-3d}$,

$$\mathop{\mathbb{P}}_{S \sim \Theta}[\alpha \subseteq S] \leqslant \rho^{D-3d}.$$

In the sub-instance, the expected value (over the choice of planted instance and over the choice of sub-instance) of the restricted solution x is

$$\frac{\rho a}{n} \cdot \left(\binom{|Y|}{2} + \binom{|Z|}{2} \right) - \rho \frac{b-a}{n} \cdot |Y| \cdot |Z| = \frac{(2a-b)\rho n}{4} - \rho a,$$

and by a Chernoff bound, the value in the sub instance is within a $(1-\beta)$-factor with probability $1 - n^{-10B}$ for $\beta = \sqrt{\frac{10B \log n}{n}}$. On resampling the edges outside the sub-instance from the uniform

distribution, this value can only decrease by at most $(1-\rho)(1+\beta)nb/2$ w.h.p over the choice of the outside edges.

If we set $\rho = (1-\varepsilon(2a-b)/10b)$, then $\rho^{D-3d} \leqslant n^{-8B}$ for $\varepsilon \geqslant \frac{c_2(2a-b)\log n}{D-3d}$. for some constant c_2, while the objective value is at least $(1-\varepsilon)\frac{(2a-b)n}{4}$. The conclusion follows (after making appropriate adjustments to the constant). □

Problem 5.7 (Densest-k-subgraph)**.** Given a graph $G = (V, E)$ on n vertices, determine whether it comes from:

- **Uniform Distribution**: $G(n, p)$.
- **Planted Distribution**: A graph from $G(n, p)$ with an instance of $G(k, q)$ planted on a random subset of k vertices, $p < q$.

Letting A be the adjacency matrix, the usual polynomial program for *densest-k-subgraph* in variables $x_1, \ldots, x_n$ is:

$$\begin{aligned} \text{obj} &\leqslant x^\top A x \\ x_i^2 &= x_i \quad \forall i \in [n] \\ \sum_i x_i &= k \end{aligned}$$

Lemma 5.8. *When $k^2(p+q) \gg d \log n$, Theorem 2.6 applies to the* densest-k-subgraph *problem with* $\text{obj} \leqslant (1-\varepsilon)(p+q)\binom{k}{2}$ *for any $\varepsilon > \frac{c \cdot d \log n}{D-3d}$ for a fixed constant c.*

Proof. The degree of the SoS relaxation in the instance is $k = 1$. We have from Corollary A.3 that the SoS proof has no values larger than $n^{c_1 d}$ for a constant c_1; fix $B = c_1 d$.

Our subsampling operation is to include each edge independently with probability ρ, and take the subgraph induced by the included edges. Clearly, the Booleanity and sparsity constraints are preserved by this subsampling distribution Θ. Since each edge is included independently with probability ρ, for any $\alpha \in \binom{E}{D-3d}$,

$$\mathop{\mathbb{P}}_{S \sim \Theta}[\alpha \subseteq S] \leqslant \rho^{D-3d}.$$

Now, the expected objective value (over the instance and the sub-sampling) is at least $\rho(p+q)\binom{k}{2}$, and applying a Chernoff bound, we hace that the probability the sub-sampled instance has value less than $(1-\beta)\rho(p+q)\binom{k}{2}$ is at most n^{-10B} if we choose $\beta = \sqrt{\frac{10B\log n}{\rho(p+q)\binom{k}{2}}}$ (which is valid since we assumed that $d \log n \ll (p+q)k^2$). Further, a dense subgraph on a subset of the edges is still dense when more edges are added back, so we have the n^{-10B}-robust inference property.

Thus, choosing $\rho = (1-\varepsilon)$ and setting

$$\rho^{D-3d} \leqslant n^{-8B} \implies \varepsilon \geqslant \frac{c_2 d \log n}{D-3d},$$

for some constant c_2, which concludes the proof (after making appropriate adjustments to the constant). □

Problem 5.9 (Tensor PCA)**.** Given an order-k tensor in $(\mathbb{R}^n)^{\otimes k}$, determine whether it comes from:

- **Uniform Distribution**: each entry of the tensor sampled independently from $\mathcal{N}(0, 1)$.

- **Planted Distribution**: a spiked tensor, $\mathbf{T} = \lambda \cdot v^{\otimes k} + G$ where v is sampled uniformly from $\{\pm\frac{1}{\sqrt{n}}\}^n$, and where G is a random tensor with each entry sampled independently from $\mathcal{N}(0,1)$.

Given the tensor $\mathbf{T}$, the canonical program for the tensor PCA problem in variables $x_1, \ldots, x_n$ is:

$$\begin{aligned} \text{obj} &\leqslant \langle x^{\otimes k}, \mathbf{T} \rangle \\ \|x\|_2^2 &= 1 \end{aligned}$$

Lemma 5.10. *For $\lambda n^{-\varepsilon} \gg \log n$, Theorem 2.6 applies to the* tensor PCA *problem with* $\text{obj} \leqslant \lambda n^{-\varepsilon}$ *for any $\varepsilon \geqslant \frac{c \cdot d}{D-3d}$ for a fixed constant c.*

Proof. The degree of the SoS relaxation in the instance is $k = 1$. Since the entries of the noise component of the tensor are standard normal variables, with exponentially good probability over the input tensor $\mathbf{T}$ we will have no entry of magnitude greater than n^d. This, together with Corollary A.3, gives us that except with exponentially small probability the SoS proof will have no values exceeding $n^{c_1 d}$ for a fixed constant c_1.

Our subsampling operation is to set to zero every entry of $\mathbf{T}$ independently with probability $1 - \rho$, obtaining a sub-instance $\mathbf{T}'$ on the nonzero entries. Also, for any $\alpha \in \binom{[n]^k}{D-3d}$,

$$\mathop{\mathbb{P}}_{S \sim \Theta}[\alpha \in S] \leqslant \rho^{D-3d}.$$

This subsampling operation clearly preserves the planted solution unit sphere constraint. Additionally, let $\mathcal{R}$ be the operator that restricts a tensor to the nonzero entries. We have that $\langle \mathcal{R}(\lambda \cdot v^{\otimes k}), v^{\otimes k} \rangle$ has expectation $\lambda \cdot \rho$, since every entry of $v^{\otimes k}$ has magnitude $n^{-k/2}$. Applying a Chernoff bound, we have that this quantity will be at least $(1-\beta)\lambda\rho$ with probability at least n^{-10B} if we choose $\beta = \sqrt{\frac{10B \log n}{\lambda \rho}}$.

It remains to address the noise introduced by $G_{\mathbf{T}'}$ and resampling all the entries outside of the subinstance $\mathbf{T}'$. Each of these entries is a standard normal entry. The quantity $\langle (\text{Id} - \mathcal{R})(N), v^{\otimes k} \rangle$ is a sum over at most n^k i.i.d. Gaussian entries each with standard deviation $n^{-k/2}$ (since that is the magnitude of $(v^{\otimes k})_\alpha$. The entire quantity is thus a Gaussian random variable with mean 0 and variance 1, and therefore with probability at least n^{-10B} this quantity will not exceed $\sqrt{10B \log n}$. So long as $\sqrt{10B \log n} \ll \lambda\rho$, the signal term will dominate, and the solution will have value at least $\lambda\rho/2$.

Now, we set $\rho = n^{-\varepsilon}$ so that

$$\rho^{D-3d} \leqslant n^{-8B} \implies \varepsilon \geqslant \frac{2c_1 d}{D-3d},$$

which concludes the proof (after making appropriate adjustments to the constant c_1). □

Problem 5.11 (Sparse PCA)**.** Given an $n \times m$ matrix M in $\mathbb{R}^n$, determine whether it comes from:

- **Uniform Distribution**: each entry of the matrix sampled independently from $\mathcal{N}(0,1)$.
- **Planted Distribution**: a random vector with k non-zero entries $v \in \{0, \pm 1/\sqrt{k}\}^n$ is chosen, and then the ith column of the matrix is sampled independently by taking $s_i v + \gamma_i$ for a uniformly random sign $s_i \in \{\pm 1\}$ and a standard gaussian vector $\gamma_i \sim \mathcal{N}(0, \text{Id})$.

The canonical program for the sparse PCA problem in variables $x_1, \ldots, x_n$ is:

$$\begin{aligned} \text{obj} &\leqslant \|M^\top x\|_2^2 \\ x_i^2 &= x_i \quad \forall i \in [n] \\ \|x\|_2^2 &= k \end{aligned}$$

Lemma 5.12. *For $kn^{-\varepsilon/2} \gg \log n$, Theorem 2.6 applies to the* sparse PCA *problem with* obj $\leqslant k^{2-\varepsilon} m$ *for any $\varepsilon > \frac{c \cdot d}{D-6d}$ for a fixed constant c.*

Proof. The degree of the SoS relaxation in the instance is 2. Since the entries of the noise are standard normal variables, with exponentially good probability over the input matrix M we will have no entry of magnitude greater than n^d. This, together with Corollary A.3, gives us that except with exponentially small probability the SoS proof will have no values exceeding $n^{c_1 d}$ for a fixed constant c_1.

Our subsampling operation is to set to zero every entry of M independently with probability $1-\rho$, obtaining a sub-instance M on the nonzero entries. Also, for any $\alpha \in \binom{M}{D-6d}$,

$$\mathop{\mathbb{P}}_{S \sim \Theta}[\alpha \in S] \leqslant \rho^{D-6d}.$$

This subsampling operation clearly preserves the constraints on the solution variables.

We take our subinstance solution $y = \sqrt{k}v$, which is feasible. Let $\mathcal{R}$ be the subsampling operator that zeros out a set of columns. On subsampling, and then resampling the zeroed out columns from the uniform distribution, we can write the resulting $\tilde{M}$ as

$$\tilde{M}^\top = \mathcal{R}(sv^T) + G^\top$$

where G^T is a random Gaussian matrix. Therefore, the objective value obtained by the solution $y = \sqrt{k}v$ is

$$\tilde{M}^\top y = \sqrt{k} \cdot \mathcal{R}(sv^\top)v + \sqrt{k} \cdot G^\top v$$

The first term is a vector u_{signal} with m entries, each of which is a sum of k Bernoulli random variables, all of the same sign, with probability ρ of being nonzero. The second term is a vector u_{noise} with m entries, each of them an independent Gaussian variable with variance bounded by k. We have that

$$\mathop{\mathbb{E}}_{\Theta}[\|u_{signal}\|_2^2] = (\rho k)^2 m,$$

and by Chernoff bounds we have that this concentrates within a $(1-\beta)$ factor with probability $1 - n^{-10B}$ if we take $\beta = \sqrt{\frac{10B \log n}{(\rho k)^2 m}}$.

The expectation of $\langle u_{signal}, u_{noise} \rangle$ is zero, and applying similar concentration arguments we have that with probability $1 - n^{10B}$, $|\langle u_{signal}, u_{noise} \rangle| \leqslant (1+\beta)\rho k$. Taking the union bound over these events and applying Cauchy-Schwarz, we have that

$$\|\mathcal{R}(M)y\|_2^2 \geqslant (\rho k)^2 m - 2km = \rho^2 k^2 m - 2km.$$

so long as $\rho k \gg 1$, the first term dominates.

Now, we set $\rho = n^{-\varepsilon}$ for $\varepsilon < 1$ so that

$$\rho^{D-6d} \leqslant n^{-8B} \implies \varepsilon \geqslant \frac{c_2 d}{D-6d},$$

for some constant c_2, which concludes the proof. □

Remark 5.13. For tensor PCA and sparse PCA, the underlying distributions were Gaussian. Applying Theorem 2.6 in these contexts yields the existence of distinguishers that are *low-degree* in a non-standard sense. Specifically, the degree of a monomial will be the number of distinct variables in it, irrespective of the powers to which they are raised.

6 Exponential lower bounds for PCA problems

In this section we give an overview of the proofs of our SoS lower bounds for the tensor and sparse PCA problems. We begin by showing how Conjecture 1.2 predicts such a lower bound in the tensor PCA setting. Following this we state the key lemmas to prove the exponential lower bounds; since these lemmas can be proved largely by techniques present in the work of Barak et al. on planted clique [BHK+16], we leave the details to a forthcoming full version of the present paper.

6.1 Predicting sos lower bounds from low-degree distinguishers for Tensor PCA

In this section we demonstrate how to predict using Conjecture 1.2 that when $\lambda \ll n^{3/4-\varepsilon}$ for $\varepsilon > 0$, SoS algorithms cannot solve Tensor PCA. This prediction is borne out in Theorem 1.4.

Theorem 6.1. *Let μ be the distribution on $\mathbb{R}^{n\otimes n\otimes n}$ which places a standard Gaussian in each entry. Let ν be the density of the Tensor PCA planted distribution with respect to μ, where we take the planted vector v to have each entry uniformly chosen from $\{\pm\frac{1}{\sqrt{n}}\}$.*[7] *If $\lambda \leqslant n^{3/4-\varepsilon}$, there is no degree $n^{o(1)}$ polynomial p with*

$$\mathop{\mathbb{E}}_{\mu} p(A) = 0, \quad \mathop{\mathbb{E}}_{planted} p(A) \geqslant n^{\Omega(1)} \cdot \left(\mathop{\mathbb{V}}_{\mu} p(A)\right)^{1/2}.$$

We sketch the proof of this theorem. The theorem follows from two claims.

Claim 6.2.

$$\max_{\substack{\deg p \leqslant n^{o(1)} \\ , \mathbb{E}_\mu p(T)=0}} \frac{\mathbb{E}_\nu p(T)}{\left(\mathbb{E}_\mu p(T)^2\right)^{1/2}} = (\mathop{\mathbb{E}}_{\mu}(\nu^{\leqslant d}(T)-1)^2)^{1/2} \tag{6.1}$$

where $\nu^{\leqslant d}$ is the orthogonal projection (with respect to μ) of the density ν to the degree-d polynomials. Note that the last quantity is just the 2 norm, or the variance, of the truncation to low-degree polynomials of the density ν of the planted distribution.

Claim 6.3. $(\mathbb{E}_\mu(v^{\leqslant d}(T)-1)^2)^{1/2} \ll 1$ when $\lambda \leqslant n^{3/4-\varepsilon}$ for $\varepsilon \geqslant \Omega(1)$ and $d = n^{o(1)}$.

The theorem follows immediately. We sketch proofs of the claims in order.

[7] This does not substantially modify the problem but it will make calculations in this proof sketch more convenient.

Sketch of proof for Claim 6.2. By definition of ν, the maximization is equivalent to maximizing $\mathbb{E}_\mu \nu(T) \cdot p(T)$ among all p of degree $d = n^{o(1)}$ and with $\mathbb{E}_\mu p(T)^2 = 1$ and $\mathbb{E}_\mu p(T) = 0$. Standard Fourier analysis shows that this maximum is achieved by the orthogonal projection of $\nu - 1$ into the span of degree d polynomials.

To make this more precise, recall that the Hermite polynomials provide an orthonormal basis for real-valued polynomials under the multivariate Gaussian distribution. (For an introduction to Hermite polynomials, see the book [O'D14].) The tensor $T \sim \mu$ is an n^3-dimensional multivariate Gaussian. For a (multi)-set $W \subseteq [n]^3$, let H_W be the W-th Hermite polynomial, so that $\mathbb{E}_\mu H_W(T)H_{W'}(T) = \mathbb{I}_{W=W'}$.

Then the best p (ignoring normalization momentarily) will be the function

$$p(A) = \nu^{\leqslant d}(A) - 1 = \sum_{1 \leqslant |W| \leqslant d} (\mathop{\mathbb{E}}_{T \sim \mu} \nu(T) H_W(T)) \cdot H_W(A)$$

Here $\mathbb{E}_\mu \nu(T)H_W(T) = \widehat{\nu}(W)$ is the W-th Fourier coefficient of ν. What value for (6.1) is achieved by this p? Again by standard Fourier analysis, in the numerator we have,

$$\mathop{\mathbb{E}}_{\nu} p(T) = \mathop{\mathbb{E}}_{\nu}(\nu^{\leqslant D}(T) - 1) = \mathop{\mathbb{E}}_{\mu} \nu(T) \cdot (\nu^{\leqslant D}(T) - 1) = \mathop{\mathbb{E}}_{\mu}(\nu^{\leqslant d}(T) - 1)^2$$

Comparing this to the denominator, the maximum value of (6.1) is $(\mathbb{E}_\mu(v^{\leqslant d}(T) - 1)^2)^{1/2}$. This is nothing more than the 2-norm of the projection of $\nu - 1$ to degree-d polynomials! □

The following fact, used to prove Claim 6.3, is an elementary computation with Hermite polynomials.

Fact 6.4. *Let $W \subseteq [n]^3$. Then $\widehat{\nu}(W) = \lambda^{|W|} n^{-3|W|/2}$ if W, thought of as a 3-uniform hypergraph, has all even degrees, and is 0 otherwise.*

To see that this calculation is straightforward, note that $\widehat{\nu(W)} = \mathbb{E}_\mu \nu(T)H_W(T) = \mathbb{E}_\nu H_W(T)$, so it is enough to understand the expectations of the Hermite polynomials under the planted distribution.

Sketch of proof for Claim 6.3. Working in the Hermite basis (as described above), we get $\mathbb{E}_\mu(v^{\leqslant d}(T) - 1)^2 = \sum_{1 \leqslant |W| \leqslant d} \widehat{\nu}(W)^2$. For the sake of exposition, we will restrict attention in the sum to W in which no element appears with multiplicity larger than 1 (other terms can be treated similarly).

What is the contribution to $\sum_{1 \leqslant |W| \leqslant d} \widehat{\nu}(W)^2$ of terms W with $|W| = t$? By the fact above, to contribute a nonzero term to the sum, W,considered as a 3-uniform hypergraph must have even degrees. So, if it has t hyperedges, it contains at most $3t/2$ nodes. There are $n^{3t/2}$ choices for these nodes, and having chosen them, at most $t^{O(t)}$ 3-uniform hypergraphs on those nodes. Hence,

$$\sum_{1 \leqslant |W| \leqslant d} \widehat{\nu}(W)^2 \leqslant \sum_{t=1}^{d} n^{3t/2} t^{O(t)} \lambda^{2t} n^{-3t} .$$

So long as $\lambda^2 \leqslant n^{3/2-\varepsilon}$ for some $\varepsilon = \Omega(1)$ and $t \leqslant d \leqslant n^{O(\varepsilon)}$, this is $o(1)$. □

6.2 Main theorem and proof overview for Tensor PCA

In this section we give an overview of the proof of Theorem 1.4. The techniques involved in proving the main lemmas are technical refinements of techniques used in the work of Barak et al. on SoS lower bounds for planted clique [BHK+16]; we therefore leave full proofs to a forthcoming full version of this paper.

To state and prove our main theorem on tensor PCA it is useful to define a Boolean version of the problem. For technical convenience we actually prove an SoS lower bound for this problem; then standard techniques (see Section C) allow us to prove the main theorem for Gaussian tensors.

Problem 6.5 (k-Tensor PCA, signal-strength λ, boolean version)**.** Distinguish the following two distributions on $\Omega_k \stackrel{\text{def}}{=} \{\pm 1\}^{\binom{n}{k}}$.

- *the uniform distribution:* $A \sim \Omega$ chosen uniformly at random.
- *the planted distribution:* Choose $v \sim \{\pm 1\}^n$ and let $B = v^{\otimes k}$. Sample A by rerandomizing every coordinate of B with probability $1 - \lambda n^{-k/2}$.

We show that the natural SoS relaxation of this problem suffers from a large integrality gap, when λ is slightly less than $n^{k/4}$, even when the degree of the SoS relaxation is $n^{\Omega(1)}$. (When $\lambda \gg n^{k/4-\varepsilon}$, algorithms with running time $2^{n^{O(\varepsilon)}}$ are known for $k = O(1)$ [RM14, HSS15, HSSS16, BGL16, RRS16].)

Theorem 6.6. *Let $k = O(1)$. For $A \in \Omega_k$, let*

$$SoS_d(A) \stackrel{\text{def}}{=} \max_{\tilde{\mathbb{E}}} \tilde{\mathbb{E}}\langle x^{\otimes k}, A\rangle \text{ s.t. } \tilde{\mathbb{E}} \text{ is a degree-}d \text{ pseudoexpectation satisfying } \{\|x\|^2 = 1\}.$$

There is a constant c so that for every small enough $\varepsilon > 0$, if $d \leqslant n^{c\cdot\varepsilon}$, then for large enough n,

$$\mathop{\mathbb{P}}_{A\sim\Omega}\{SoS_d(A) \geqslant n^{k/4-\varepsilon}\} \geqslant 1 - o(1)$$

and

$$\mathop{\mathbb{E}}_{A\sim\Omega} SoS_d(A) \geqslant n^{k/4-\varepsilon}.$$

Moreover, the latter also holds for A with iid entries from $\mathcal{N}(0,1)$.[8]

To prove the theorem we will exhibit for a typical sample A from the uniform distribution a degree $n^{\Omega(\varepsilon)}$ pseudodistribution $\tilde{\mathbb{E}}$ which satisfies $\{\|x\|^2 = 1\}$ and has $\tilde{\mathbb{E}}\langle x^{\otimes k}, A\rangle \geqslant n^{k/4-\varepsilon}$. The following lemma ensures that the pseudo-distribution we exhibit will be PSD.

Lemma 6.7. *Let $d \in \mathbb{N}$ and let $N_d = \sum_{s\leqslant d} n(n-1)\cdots(n-(s-1))$ be the number of $\leqslant d$-tuples with unique entries from $[n]$. There is a constant ε^* independent of n such that for any $\varepsilon < \varepsilon^*$ also independent of n, the following is true. Let $\lambda = n^{k/4-\varepsilon}$. Let $\mu(A)$ be the density of the following distribution (with respect to the uniform distribution on $\Omega = \{\pm 1\}^{\binom{n}{k}}$).*

The Planted Distribution: *Choose $v \sim \{\pm 1\}^n$ uniformly. Let $B = v^{\otimes k}$. Sample A by*

- *replacing every coordinate of B with a random draw from $\{\pm 1\}$ independently with probability $1 - \lambda n^{-k/2}$,*

[8] For technical reasons we do not prove a tail bound type statement for Gaussian A, but we conjecture that this is also true.

- *then choosing a subset $S \subseteq [n]$ by including every coordinate with probability $n^{-\varepsilon}$,*
- *then replacing every entry of B with some index outside S independently with a uniform draw from $\{\pm 1\}$.*

Let $\Lambda : \Omega \to \mathbb{R}^{N_d \times N_d}$ be the following function

$$\Lambda(A) = \mu(A) \cdot \mathop{\mathbb{E}}_{v|A} v^{\otimes \leqslant 2d}$$

Here we abuse notation and denote by $x^{\leqslant \otimes 2d}$ the matrix indexed by tuples of length $\leqslant d$ with unique entries *from $[n]$. For $D \in \mathbb{N}$, let $\Lambda^{\leqslant D}$ be the projection of Λ into the degree-D real-valued polynomials on $\{\pm 1\}^{\binom{n}{k}}$. There is a universal constant C so that if $Cd/\varepsilon < D < n^{\varepsilon/C}$, then for large enough n*

$$\mathop{\mathbb{P}}_{A \sim \Omega} \{\Lambda^{\leqslant D}(A) \succeq 0\} \geqslant 1 - o(1)\,.$$

For a tensor A, the moment matrix of the pseudodistribution we exhibit will be $\Lambda^{\leqslant D}(A)$. We will need it to satisfy the constraint $\{\|x\|^2 = 1\}$. This follows from the following general lemma. (The lemma is much more general than what we state here, and uses only the vector space structures of space of real matrices and matrix-valued functions.)

Lemma 6.8. *Let $k \in \mathbb{N}$. Let V be a linear subspace of $\mathbb{R}^{N \times M}$. Let $\Omega = \{\pm 1\}^{\binom{n}{k}}$. Let $\Lambda : \Omega \to V$. Let $\Lambda^{\leqslant D}$ be the entrywise orthogonal projection of Λ to polynomials of degree at most D. Then for every $A \in \Omega$, the matrix $\Lambda^{\leqslant D}(A) \in V$.*

Proof. The function Λ is an element of the vector space $\mathbb{R}^{N \times M} \otimes \mathbb{R}^{\Omega}$. The projection $\Pi_V : \mathbb{R}^{N \times M} \to V$ and the projection $\Pi_{\leqslant D}$ from $\mathbb{R}^{\Omega}$ to the degree-D polynomials commute as projections on $\mathbb{R}^{N \times M} \otimes \mathbb{R}^{\Omega}$, since they act on separate tensor coordinates. It follows that $\Lambda^{\leqslant D} \in V \otimes (\mathbb{R}^{\Omega})^{\leqslant D}$ takes values in V. □

Last, we will require a couple of scalar functions of $\Lambda^{\leqslant D}$ to be well concentrated.

Lemma 6.9. *Let $\Lambda, d, \varepsilon, D$ be as in Lemma 6.7. The function $\Lambda^{\leqslant D}$ satisfies*

- $\mathbb{P}_{A \sim \Omega}\{\Lambda^{\leqslant D}_{\emptyset,\emptyset}(A) = 1 \pm o(1)\} \geqslant 1 - o(1)$ *(Here $\Lambda_{\emptyset,\emptyset} = 1$ is the upper-left-most entry of Λ.)*
- $\mathbb{P}_{A \sim \Omega}\{\langle \Lambda^{\leqslant D}(A), A\rangle = (1 \pm o(1)) \cdot n^{3k/4-\varepsilon}\} \geqslant 1 - o(1)$ *(Here we are abusing notation to write $\langle \Lambda^{\leqslant D}(A), A\rangle$ for the inner product of the part of $\Lambda^{\leqslant D}$ indexed by monomials of degree k and A.)*

The Boolean case of Theorem 6.6 follows from combining the lemmas. The Gaussian case can be proved in a black-box fashion from the Boolean case following the argument in Section C.

The proofs of all the lemmas in this section follow analogous lemmas in the work of Barak et al. on planted clique [BHK+16]; we defer them to the full version of the present work.

6.3 Main theorem and proof overview for sparse PCA

In this section we prove the following main theorem. Formally, the theorem shows that with high probability for a random $n \times n$ matrix A, even high-degree SoS relaxations are unable to certify that no sparse vector v has large quadratic form $\langle v, Av\rangle$.

Theorem 6.10 (Restatement of Theorem 1.6)**.** *If $A \in \mathbb{R}^{n \times n}$, let*

$$SoS_{d,k}(A) = \max_{\tilde{\mathbb{E}}} \tilde{\mathbb{E}}\langle x, Ax\rangle \text{ s.t. } \tilde{\mathbb{E}} \text{ is degree } d \text{ and satisfies } \left\{x_i^3 = x_i, \|x\|^2 = k\right\} .$$

There are absolute constants $c, \varepsilon^* > 0$ so that for every $\rho \in (0,1)$ and $\varepsilon \in (0, \varepsilon^*)$, if $k = n^\rho$, then for $d \leqslant n^{c \cdot \varepsilon}$,

$$\mathbb{P}_{A \sim \{\pm 1\}^{\binom{n}{2}}} \{SoS_{d,k}(A) \geqslant \min(n^{1/2-\varepsilon}k, n^{\rho-\varepsilon}k)\} \geqslant 1 - o(1)$$

and

$$\mathbb{E}_{A \sim \{\pm 1\}^{\binom{n}{2}}} SoS_{d,k}(A) \geqslant \min(n^{1/2-\varepsilon}k, n^{\rho-\varepsilon}k) .$$

Furthermore, the latter is true also if A is symmetric with iid entries from $\mathcal{N}(0,1)$.[9]

We turn to some discussion of the theorem statement. First of all, though it is technically convenient for A in the theorem statement above to be a ± 1 matrix, the entries may be replaced by standard Gaussians (see Section C).

Remark 6.11 (Relation to the spiked-Wigner model of sparse principal component analysis). To get some intuition for the theorem statement, it is useful to return to a familiar planted problem: the spiked-Wigner model of sparse principal component analysis. Let W be a symmetric matrix with iid entries from $\mathcal{N}(0,1)$, and let v be a random k-sparse unit vector with entries $\{\pm 1/\sqrt{k}, 0\}$. Let $B = W + \lambda vv^\intercal$. The problem is to distinguish between a single sample from B and a sample from W. There are two main algorithms for this problem, both captured by the SoS hierarchy. The first, applicable when $\lambda \gg \sqrt{n}$, is vanilla PCA: the top eigenvalue of B will be larger than the top eigenvalue of W. The second, applicable when $\lambda \gg k$, is diagonal thresholding: the diagonal entries of B which corresponds to nonzero coordinates will be noticeably large. The theorem statement above (transferred to the Gaussian setting, though this has little effect) shows that once λ is well outside these parameter regimes, i.e. when $\lambda < n^{1/2-\varepsilon}, k^{1-\varepsilon}$ for arbitrarily small $\varepsilon > 0$, even degree $n^{\Omega(\varepsilon)}$ SoS programs do not distinguish between B and W.

Remark 6.12 (Interpretation as an integrality gap). A second interpretation of the theorem statement, independent of any planted problem, is as a strong integrality gap for random instances for the problem of maximizing a quadratic form over k-sparse vectors. Consider the actual maximum of $\langle x, Ax\rangle$ for random ($\{\pm 1\}$ or Gaussian) A over k-sparse unit vectors x. There are roughly $2^{k \log n}$ points in a $\frac{1}{2}$-net for such vectors, meaning that by standard arguments,

$$\max_{\|x\|=1, x \text{ is } k\text{-sparse}} \langle x, Ax\rangle \leqslant O(\sqrt{k} \log n) .$$

With the parameters of the theorem, this means that the integrality gap of the degree $n^{\Omega(\varepsilon)}$ SoS relaxation is at least $\min(n^{\rho/2-\varepsilon}, n^{1/2-\rho/2-\varepsilon})$ when $k = n^\rho$.

Remark 6.13 (Relation to spiked-Wishart model). Theorem 1.6 most closely concerns the spiked-Wigner model of sparse PCA; this refers to independence of the entries of the matrix A. Often, sparse PCA is instead studied in the (perhaps more realistic) *spiked-Wishart model*, where the input

[9] For technical reasons we do not prove a tail bound type statement for Gaussian A, but we conjecture that this is also true.

is m samples $x_1, \dots, x_m$ from an n-dimensional Gaussian vector $\mathcal{N}(0, \mathrm{Id} + \lambda \cdot vv^\top)$, where v is a unit-norm k-sparse vector. Here the question is: as a function of the sparsity k, the ambient dimension n, and the signal strength λ, how many samples m are needed to recover the vector v? The natural approach to recovering v in this setting is to solve a convex relaxation of the problem of maximizing he quadratic form of the empirical covariance $M = \sum_{i \leqslant m} x_i x_i^\intercal$ over k-sparse unit vectors (the maximization problem itself is NP-hard even to approximate [CPR16]).

Theoretically, one may apply our proof technique for Theorem 1.6 directly to the spiked-Wishart model, but this carries the expense of substantial technical complication. We may however make intelligent guesses about the behavior of SoS relaxations for the spiked-Wishart model on the basis of Theorem 1.6 alone. As in the spiked Wigner model, there are essentially two known algorithms to recover a planted sparse vector v in the spiked Wishart model: vanilla PCA and diagonal thresholding [DM14b]. We conjecture that, as in the spiked Wigner model, the SoS hierarchy requires $n^{\Omega(1)}$ degree to improve the number of samples required by these algorithms by any polynomial factor. Concretely, considering the case $\lambda = 1$ for simplicity, we conjecture that there are constants c, ε^* such that for every $\varepsilon \in (0, \varepsilon^*)$ if $m \leqslant \min(k^{2-\varepsilon}, n^{1-\varepsilon})$ and $x_1, \dots, x_m \sim \mathcal{N}(0, \mathrm{Id})$ are iid, then with high probability for every $\rho \in (0, 1)$ if $k = n^\rho$,

$$SoS_{d,k}\left(\sum_{i \leqslant m} x_i x_i^\intercal\right) \geqslant \min(n^{1-\varepsilon} k, k^{2-\varepsilon})$$

for all $d \leqslant n^{c \cdot \varepsilon}$.

Lemmas for Theorem 1.6. Our proof of Theorem 1.6 is very similar to the analogous proof for Tensor PCA, Theorem 6.6. We state the analogues of Lemma 6.7 and Lemma 6.9. Lemma 6.8 can be used unchanged in the sparse PCA setting.

The main lemma, analogous to Lemma 6.7 is as follows.

Lemma 6.14. *Let $d \in \mathbb{N}$ and let $N_d = \sum_{s \leqslant d} n(n-1)\cdots(n-(s-1))$ be the number of $\leqslant d$-tuples with unique entries from $[n]$. Let $\mu(A)$ be the density of the following distribution on $n \times n$ matrices A with respect to the uniform distribution on $\{\pm 1\}^{\binom{n}{2}}$.*

Planted distribution: *Let $k = k(n) \in \mathbb{N}$ and $\lambda = \lambda(n) \in \mathbb{R}$, and $\gamma > 0$, and assume $\lambda \leqslant k$. Sample a uniformly random k-sparse vector $v \in \mathbb{R}^n$ with entries $\pm 1, 0$. Form the matrix $B = vv^\top$. For each nonzero entry of B independently, replace it with a uniform draw from $\{\pm 1\}$ with probability $1 - \lambda/k$ (maintaining the symmetry $B = B^\top$). For each zero entry of B, replace it with a uniform draw from $\{\pm 1\}$ (maintaining the same symmetry). Finally, choose every $i \in [n]$ with probability $n^{-\gamma}$ independently; for those indices that were not chosen, replace every entry in the corresponding row and column of B with random ± 1 entries.[10] Output the resulting matrix A. (We remark that this matrix is a Boolean version of the more standard spiked-Wigner model $B + \lambda vv^\top$ where B has iid standard normal entries and v is a random k-sparse unit vector with entries from $\{\pm 1/\sqrt{k}, 0\}$.)*

Let $\Lambda : \{\pm 1\}^{\binom{n}{2}} \to \mathbb{R}^{N_d \times N_d}$ be the following function

$$\Lambda(A) = \mu(A) \cdot \mathop{\mathbb{E}}_{v|A} v^{\otimes \leqslant 2d}$$

[10] This additional $n^{-\gamma}$ noising step is a technical convenience which has the effect of somewhat decreasing the number of nonzero entries of v and decreasing the signal-strength λ.

where the expectation is with respect to the planted distribution above. For $D = D(n) \in \mathbb{N}$, let $\Lambda^{\leqslant D}$ be the entrywise projection of Λ into the Boolean functions of degree at most D.

There are constants $C, \varepsilon^* > 0$ such that for every $\gamma > 0$ and $\rho \in (0,1)$ and every $\varepsilon \in (0, \varepsilon^*)$ (all independent of n), if $k = n^\rho$ and $\lambda \leqslant \min\{n^{\rho-\varepsilon}, n^{1/2-\varepsilon}\}$, and if $Cd/\varepsilon < D < n^{\varepsilon/C}$, then for large enough n

$$\mathbb{P}_{A \sim \{\pm 1\}^{\binom{n}{2}}} \{\Lambda^{\leqslant D}(A) \succeq 0\} \geqslant 1 - o(1)\,.$$

Remark 6.15. We make a few remarks about the necessity of some of the assumptions above. A useful intuition is that the function $\Lambda^{\leqslant D}(A)$ is (with high probability) positive-valued when the parameters $\rho, \varepsilon, \gamma$ of the planted distribution are such that there is no degree-D polynomial $f : \{\pm 1\}^{\binom{n}{2}} \to \mathbb{R}$ whose values distinguish a typical sample from the planted distribution from a null model: a random symmetric matrix with iid entries.

At this point it is useful to consider a more familiar planted model, which the lemma above mimics. Let W be a $n \times n$ symmetric matrix with iid entries from $\mathcal{N}(0,1)$. Let $v \in \mathbb{R}^n$ be a k-sparse unit vector, with entries in $\{\pm 1/\sqrt{k}, 0\}$. Let $A = W + \lambda v v^\intercal$. Notice that if $\lambda \gg k$, then diagonal thresholding on the matrix W identifies the nonzero coordinates of v. (This is the analogue of the covariance-thresholding algorithm in the spiked-Wishart version of sparse PCA.) On the other hand, if $\lambda \gg \sqrt{n}$ then (since typically $\|W\| \approx \sqrt{n}$), ordinary PCA identifies v. The lemma captures computational hardness for the problem of distinguishing a single sample from A from a sample from the null model W both diagonal thresholding and ordinary PCA fail.

Next we state the analogue of Lemma 6.9.

Lemma 6.16. *Let $\Lambda, d, k, \lambda, \gamma, D$ be as in Lemma 6.14. The function $\Lambda^{\leqslant D}$ satisfies*

- $\mathbb{P}_{A \sim \{\pm 1\}^{\binom{n}{k}}} \{\Lambda^{\leqslant D}_{\emptyset,\emptyset}(A) = 1 \pm o(1)\} \geqslant 1 - o(1)$.
- $\mathbb{P}_{A \sim \{\pm 1\}^{\binom{n}{k}}} \{\langle \Lambda^{\leqslant D}(A), A \rangle = (1 \pm o(1)) \cdot \lambda n^{\Theta(-\gamma)}\} \geqslant 1 - o(1)$.

References

[AK97] Noga Alon and Nabil Kahale, *A spectral technique for coloring random 3-colorable graphs*, SIAM J. Comput. **26** (1997), no. 6, 1733–1748. 1

[AKS98] Noga Alon, Michael Krivelevich, and Benny Sudakov, *Finding a large hidden clique in a random graph*, Random Struct. Algorithms **13** (1998), no. 3-4, 457–466. 1, 4, 8

[AOW15a] Sarah R. Allen, Ryan O'Donnell, and David Witmer, *How to refute a random CSP*, 2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015, IEEE Computer Soc., Los Alamitos, CA, 2015, pp. 689–708. MR 3473335 1

[AOW15b] Sarah R. Allen, Ryan O'Donnell, and David Witmer, *How to refute a random CSP*, FOCS, IEEE Computer Society, 2015, pp. 689–708. 8

[BBH+12] 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*, STOC, ACM, 2012, pp. 307–326. 1, 6, 8

[BCC+10] Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan, *Detecting high log-densities—an $O(n^{1/4})$ approximation for densest k-subgraph*, STOC'10—Proceedings of the 2010 ACM International Symposium on Theory of Computing, ACM, New York, 2010, pp. 201–210. MR 2743268 8

[BCK15] Boaz Barak, Siu On Chan, and Pravesh K. Kothari, *Sum of squares lower bounds from pairwise independence [extended abstract]*, STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing, ACM, New York, 2015, pp. 97–106. MR 3388187 8

[BGG+16] Vijay V. S. P. Bhattiprolu, Mrinal Kanti Ghosh, Venkatesan Guruswami, Euiwoong Lee, and Madhur Tulsiani, *Multiplicative approximations for polynomial optimization over the unit sphere*, Electronic Colloquium on Computational Complexity (ECCC) **23** (2016), 185. 1, 6, 8

[BGL16] Vijay V. S. P. Bhattiprolu, Venkatesan Guruswami, and Euiwoong Lee, *Certifying random polynomials over the unit sphere via sum of squares hierarchy*, CoRR **abs/1605.00903** (2016). 1, 2, 9, 31

[BHK+16] Boaz Barak, Samuel B. Hopkins, Jonathan A. Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin, *A nearly tight sum-of-squares lower bound for the planted clique problem*, FOCS, IEEE Computer Society, 2016, pp. 428–437. 1, 4, 5, 8, 29, 31, 32

[BKS14] Boaz Barak, Jonathan A. Kelner, and David Steurer, *Rounding sum-of-squares relaxations*, STOC, ACM, 2014, pp. 31–40. 1

[BKS15] _____, *Dictionary learning and tensor decomposition via the sum-of-squares method*, STOC, ACM, 2015, pp. 143–151. 1

[BKS17] Boaz Barak, Pravesh Kothari, and David Steurer, *Quantum entanglement, sum of squares, and the log rank conjecture*, CoRR **abs/1701.06321** (2017). 1

[BM16] Boaz Barak and Ankur Moitra, *Noisy tensor completion via the sum-of-squares hierarchy*, COLT, JMLR Workshop and Conference Proceedings, vol. 49, JMLR.org, 2016, pp. 417–445. 1, 8

[BMVX16] Jess Banks, Cristopher Moore, Roman Vershynin, and Jiaming Xu, *Information-theoretic bounds and phase transitions in clustering, sparse pca, and submatrix localization*, CoRR **abs/1607.05222** (2016). 6

[BR13a] Quentin Berthet and Philippe Rigollet, *Complexity theoretic lower bounds for sparse principal component detection*, COLT, JMLR Workshop and Conference Proceedings, vol. 30, JMLR.org, 2013, pp. 1046–1066. 7

[BR13b] Quentin Berthet and Philippe Rigollet, *Computational lower bounds for sparse pca*, COLT (2013). 2

[BS14] Boaz Barak and David Steurer, *Sum-of-squares proofs and the quest toward optimal algorithms*, CoRR **abs/1404.5236** (2014). 6

[CC09] Eric Carlen and ERIC CARLEN, *Trace inequalities and quantum entropy: An introductory course*, 2009. 17

[CPR16] Siu On Chan, Dimitris Papailliopoulos, and Aviad Rubinstein, *On the approximability of sparse PCA*, COLT, JMLR Workshop and Conference Proceedings, vol. 49, JMLR.org, 2016, pp. 623–646. 7, 34

[DM14a] Yash Deshpande and Andrea Montanari, *Information-theoretically optimal sparse PCA*, ISIT, IEEE, 2014, pp. 2197–2201. 7

[DM14b] _____, *Sparse PCA via covariance thresholding*, NIPS, 2014, pp. 334–342. 2, 34

[DM15] _____, *Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems*, COLT, JMLR Workshop and Conference Proceedings, vol. 40, JMLR.org, 2015, pp. 523–562. 8

[DX13] Feng Dai and Yuan Xi, *Spherical harmonics*, arXiv preprint arXiv:1304.2585 (2013). 43

[Fil16] Yuval Filmus, *An orthogonal basis for functions over a slice of the boolean hypercube*, Electr. J. Comb. **23** (2016), no. 1, P1.23. 43, 44

[FM16] Zhou Fan and Andrea Montanari, *How well do local algorithms solve semidefinite programs?*, CoRR **abs/1610.05350** (2016). 8

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

[Gri01a] Dima Grigoriev, *Complexity of positivstellensatz proofs for the knapsack*, Computational Complexity **10** (2001), no. 2, 139–154. 8

[Gri01b] _____, *Linear lower bound on degrees of positivstellensatz calculus proofs for the parity*, Theor. Comput. Sci. **259** (2001), no. 1-2, 613–622. 8

[GW94] Michel X. Goemans and David P. Williamson, *.879-approximation algorithms for MAX CUT and MAX 2sat*, STOC, ACM, 1994, pp. 422–431. 1

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

[HKP15] Samuel B. Hopkins, Pravesh K. Kothari, and Aaron Potechin, *Sos and planted clique: Tight analysis of MPW moments at all degrees and an optimal lower bound at degree four*, CoRR **abs/1507.05230** (2015). 8

[HL09] Christopher J. Hillar and Lek-Heng Lim, *Most tensor problems are NP hard*, CoRR **abs/0911.1393** (2009). 6

[HSS15] Samuel B. Hopkins, Jonathan Shi, and David Steurer, *Tensor principal component analysis via sum-of-square proofs*, COLT, JMLR Workshop and Conference Proceedings, vol. 40, JMLR.org, 2015, pp. 956–1006. 1, 2, 6, 8, 9, 31

[HSSS16] Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer, *Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors*, STOC, ACM, 2016, pp. 178–191. 1, 2, 8, 9, 31

[KMOW17] Pravesh K. Kothari, Ryuhei Mori, Ryan O'Donnell, and David Witmer, *Sum of squares lower bounds for refuting any CSP*, CoRR **abs/1701.04521** (2017). 8

[KNV+15] Robert Krauthgamer, Boaz Nadler, Dan Vilenchik, et al., *Do semidefinite relaxations solve sparse pca up to the information limit?*, The Annals of Statistics **43** (2015), no. 3, 1300–1322. 2

[KT17] Ken-ichi Kawarabayashi and Mikkel Thorup, *Coloring 3-colorable graphs with less than* $n^{1/5}$ *colors*, J. ACM **64** (2017), no. 1, 4:1–4:23. 1

[LRS15] James R. Lee, Prasad Raghavendra, and David Steurer, *Lower bounds on the size of semidefinite programming relaxations*, STOC, ACM, 2015, pp. 567–576. 1

[MPW15] Raghu Meka, Aaron Potechin, and Avi Wigderson, *Sum-of-squares lower bounds for planted clique [extended abstract]*, STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing, ACM, New York, 2015, pp. 87–96. MR 3388186 8

[MS16a] Andrea Montanari and Subhabrata Sen, *Semidefinite programs on sparse random graphs and their application to community detection*, STOC, ACM, 2016, pp. 814–827. 8

[MS16b] Andrea Montanari and Nike Sun, *Spectral algorithms for tensor completion*, CoRR **abs/1612.07866** (2016). 8

[MSS16a] Tengyu Ma, Jonathan Shi, and David Steurer, *Polynomial-time tensor decompositions with sum-of-squares*, CoRR **abs/1610.01980** (2016). 1

[MSS16b] ———, *Polynomial-time tensor decompositions with sum-of-squares*, FOCS, IEEE Computer Society, 2016, pp. 438–446. 1

[MW15a] Tengyu Ma and Avi Wigderson, *Sum-of-squares lower bounds for sparse PCA*, NIPS, 2015, pp. 1612–1620. 2

[MW15b] ———, *Sum-of-squares lower bounds for sparse PCA*, CoRR **abs/1507.06370** (2015). 8

[O'D14] Ryan O'Donnell, *Analysis of boolean functions*, Cambridge University Press, 2014. 30

[Pea01] Karl Pearson, *On lines and planes of closes fit to systems of points in space*, Philosophical Magazine **2** (1901), 559–572. 1

[PS17] Aaron Potechin and David Steurer, *Exact tensor completion with sum-of-squares*, CoRR **abs/1702.06237** (2017). 1

[PWB16] Amelia Perry, Alexander S. Wein, and Afonso S. Bandeira, *Statistical limits of spiked tensor models*, CoRR **abs/1612.07728** (2016). 6

[RM14] Emile Richard and Andrea Montanari, *A statistical model for tensor PCA*, NIPS, 2014, pp. 2897–2905. 2, 5, 9, 31

[RRS16] Prasad Raghavendra, Satish Rao, and Tselil Schramm, *Strongly refuting random csps below the spectral threshold*, CoRR **abs/1605.00058** (2016). 1, 2, 6, 8, 9, 31

[RS15] Prasad Raghavendra and Tselil Schramm, *Tight lower bounds for planted clique in the degree-4 SOS program*, CoRR **abs/1507.05136** (2015). 8

[RW17] Prasad Raghavendra and Benjamin Weitz, *On the bit complexity of sum-of-squares proofs*, CoRR **abs/1702.05139** (2017). 39, 41

[Sch08] Grant Schoenebeck, *Linear level lasserre lower bounds for certain k-csps*, FOCS, IEEE Computer Society, 2008, pp. 593–602. 8

[Tre09] Luca Trevisan, *Max cut and the smallest eigenvalue*, STOC, ACM, 2009, pp. 263–272. 1

[TS14] Ryota Tomioka and Taiji Suzuki, *Spectral norm of random tensors*, arXiv preprint arXiv:1407.1870 (2014). 9

[Wei17] Benjamin Weitz, *Polynomial proof systems, effective derivations, and their applications in the sum-of-squares hierarchy*, Ph.D. thesis, UC Berkeley, 2017. 41, 42

[ZHT06] Hui Zou, Trevor Hastie, and Robert Tibshirani, *Sparse principal component analysis*, Journal of Computational and Graphical Statistics **15** (2006), no. 2, 265–286. 2

A Bounding the sum-of-squares proof ideal term

We give conditions under which sum-of-squares proofs are well-conditioned, using techniques similar to those that appear in [RW17] for bounding the bit complexity of SoS proofs. We begin with some definitions.

Definition A.1. Let $\mathcal{P}$ be a polynomial optimization problem and let $\mathcal{D}$ be the uniform distribution over the set of feasible solutions S for $\mathcal{P}$. Define the degree-$2d$ moment matrix of $\mathcal{D}$ to be $X_{\mathcal{D}} = \mathbb{E}_{s\sim\mathcal{D}}[\hat{s}^{\otimes 2d}]$, where $\hat{s} = [1\ s]^\top$.

- We say that $\mathcal{P}$ *is k-complete on up to degree* $2d$ if every zero eigenvector of $X_{\mathcal{D}}$ has a degree-k derivation from the ideal constraints of $\mathcal{P}$.

Theorem A.2. *Let $\mathcal{P}$ be a polynomial optimization problem over variables $x \in \mathbb{R}^n$ of degree at most $2d$, with objective function $f(x)$ and ideal constraints $\{g_j(x) = 0\}_{j\in[m]}$. Suppose also that $\mathcal{P}$ is $2d$-complete up to degree $2d$. Let G be the matrix of ideal constraints in the degree-$2d$ SoS proof for $\mathcal{P}$. Then if*

- *the SDP optimum value is bounded by $n^{O(d)}$*
- *the coefficients of the objective function are bounded by $n^{O(d)}$,*
- *there is a set of feasible solutions $\mathcal{S} \subseteq \mathbb{R}^n$ with the property that for each $\alpha \subseteq [n]^d$, $|\alpha| \leqslant d$ for which χ_α is not identically zero over the solution space, there exists some $s \in \mathcal{S}$ such that the square monomial $\chi_\alpha(s)^2 \geqslant n^{-O(d)}$,*

it follows that the SoS certificate for the problem is well-conditioned, with no value larger than $n^{O(d)}$.

To prove this, we essentially reproduce the proof of the main theorem of [RW17], up to the very end of the proof at which point we slightly deviate to draw a different conclusion.

Proof. Following our previous convention, the degree-$2d$ sum-of-squares proof for $\mathcal{P}$ is of the form

$$\text{sdpOpt} - f(x) = a(x) + g(x),$$

where the $g(x)$ is a polynomial in the span of the ideal constraints, and A is a sum of squares of polynomials. Alternatively, we have the matrix characterization,

$$\text{sdpOpt} - \langle F, \hat{x}^{\otimes 2d} \rangle = \langle A, \hat{x}^{\otimes 2d} \rangle + \langle G, \hat{x}^{\otimes 2d} \rangle,$$

where $\hat{x} = [1\ x]^\top$, F, A, and G are matrix polynomials corresponding to f, a, and g respectively, and with $A \succeq 0$.

Now let $s \in \mathcal{S}$ be a feasible solution. Then we have that

$$\text{sdpOpt} - \langle F, s^{\otimes 2d} \rangle = \langle A, s^{\otimes 2d} \rangle + \langle G, s^{\otimes 2d} \rangle = \langle A, s^{\otimes 2d} \rangle,$$

where the second equality follows because each $s \in S$ is feasible. By assumption the left-hand-side is bounded by $n^{O(d)}$.

We will now argue that the diagonal entries of A cannot be too large. Our first step is to argue that A cannot have nonzero diagonal entries unless there is a solution element in the solution Let $X_{\mathcal{D}} = \mathbb{E}[x^{\otimes 2d}]$ be the $2d$-moment matrix of the uniform distribution of feasible solutions to $\mathcal{P}$. Define Π to be the orthogonal projection into the zero eigenspace of $X_{\mathcal{D}}$. By linearity and orthonormality, we have that

$$\begin{aligned}\langle X_{\mathcal{D}}, A \rangle &= \left\langle X_{\mathcal{D}}, (\Pi + \Pi^\perp) A (\Pi + \Pi^\perp) \right\rangle \\ &= \left\langle X_{\mathcal{D}}, \Pi^\perp A \Pi^\perp \right\rangle + \left\langle X_{\mathcal{D}}, \Pi A \Pi^\perp \right\rangle + \left\langle X_{\mathcal{D}}, \Pi^\perp A \Pi \right\rangle + \langle X_{\mathcal{D}}, \Pi A \Pi \rangle .\end{aligned}$$

By assumption $\mathcal{P}$ is $2d$-complete on $\mathcal{D}$ up to degree $2d$, and therefore Π is derivable in degree $2d$ from the ideal constraints $\{g_j\}_{j \in [m]}$. Therefore, the latter three terms may be absorbed into G, or more formally, we can set $A' = \Pi^\perp A \Pi^\perp$, $G' = G + (\Pi + \Pi^\perp) A (\Pi + \Pi^\perp) - \Pi^\perp A \Pi^\perp$, and re-write the original proof

$$\text{sdpOpt} - \langle F, \hat{x}^{\otimes 2d} \rangle = \langle A', \hat{x}^{\otimes 2d} \rangle + \langle G', \hat{x}^{\otimes 2d} \rangle. \tag{A.1}$$

The left-hand-side remains unchanged, so we still have that it is bounded by $n^{O(d)}$ for any feasible solution $s \in \mathcal{S}$. Furthermore, the nonzero eigenspaces of $X_{\mathcal{D}}$ and A' are identical, and so A' cannot be nonzero on any diagonal entry which is orthogonal to the space of feasible solutions.

Now, we argue that every diagonal entry of A' is at most $n^{O(d)}$. To see this, for each diagonal term χ_α^2, we choose the solution $s \in \mathcal{S}$ for which $\chi_\alpha(s)^2 \geqslant n^{-O(d)}$. We then have by the PSDness of A' that

$$A'_{\alpha,\alpha} \cdot \chi_\alpha(s)^2 \leqslant \langle s^{\otimes 2d}, A' \rangle \leqslant n^{O(d)},$$

which then implies that $A'_{\alpha,\alpha} \leqslant n^{O(d)}$. It follows that $\text{Tr}(A') \leqslant n^{O(d)}$, and again since A' is PSD,

$$\|A'\|_F \leqslant \sqrt{\text{Tr}(A')} \leqslant n^{O(d)}. \tag{A.2}$$

Putting things together, we have from our original matrix identity (A.1) that

$$\begin{aligned}\|G'\|_F &= \|\operatorname{sdpOpt} - A' - F\|_F \\ &\leqslant \|\operatorname{sdpOpt}\|_F + \|A'\|_F + \|F\|_F \quad \text{(triangle inequality)} \\ &\leqslant \|\operatorname{sdpOpt}\|_F + n^{O(d)} + \|F\|_F \quad \text{(from (A.2))}.\end{aligned}$$

Therefore by our assumptions that $\|\operatorname{sdpOpt}\|, \|F\|_F = n^{O(d)}$, the conclusion follows. □

We now argue that the conditions of this theorem are met by several general families of problems.

Corollary A.3. *The following problems have degree-$2d$ SoS proofs with all coefficients bounded by $n^{O(d)}$:*

1. *The hypercube: Any polynomial optimization problem with the only constraints being $\{x_i^2 = x_i\}_{i\in[n]}$ or $\{x_i^2 = 1\}_{i\in[n]}$ and objective value at most $n^{O(d)}$ over the set of integer feasible solutions. (Including* MAX k-CSP*).*

2. *The hypercube with balancedness constraints: Any polynomial optimization problem with the only constraints being $\{x_i^2 - 1\}_{i\in[n]} \cup \{\sum_i x_i = 0\}$. (Including* COMMUNITY DETECTION*).*

3. *The unit sphere: Any polynomial optimization problem with the only constraints being $\{\sum_{i\in[n]} x_i^2 = 1\}$ and objective value at most $n^{O(d)}$ over the set of integer feasible solutions. (Including* TENSOR PCA*).*

4. *The sparse hypercube: As long as $2d \leqslant k$, any polynomial optimization problem with the only constraints being $\{x_i^2 = x_i\}_{i\in[n]} \cup \{\sum_{i\in[n]} x_i = k\}$, or $\{x_i^3 = x_i\}_{i\in[n]} \cup \{\sum_{i\in[n]} x_i^2 = k\}$, and objective value at most $n^{O(d)}$ over the set of integer feasible solutions. (Including* DENSEST k-SUBGRAPH *and the Boolean version of* SPARSE PCA*).*

5. *The* MAX CLIQUE *problem.*

We prove this corollary below. For each of the above problems, it is clear that the objective value is bounded and the objective function has no large coefficients. To prove this corollary, we need to verify the completeness of the constraint sets, and then demonstrate a set of feasible solutions so that each square term receives non-negligible mass from some solution.

A large family of completeness conditions were already verified by [RW17] and others (see the references therein):

Proposition A.4 (Completeness of canonical polynomial optimization problems (from Corollary 3.5 of [RW17])). *The following pairs of polynomial optimization problems $\mathcal{P}$ and distributions over solutions $\mathcal{D}$ are complete:*

1. *If the feasible set is $x \in \mathbb{R}^n$ with $\{x_i^2 = 1\}_{i\in[n]}$ or $\{x_i^2 = x_i\}_{i\in[n]}$, $\mathcal{P}$ is d-complete up to degree d (e.g. if $\mathcal{P}$ is a CSP). This is still true of the constraints $\{x_i^2 = 1\}_{i\in[n]} \cup \{\sum_i x_i = 0\}$ (e.g. if $\mathcal{P}$ is a community detection problem).*

2. *If the feasible set is $x \in \mathbb{R}^n$ with $\sum_{i\in[n]} x_i^2 = \alpha$, then $\mathcal{P}$ is d-complete on $\mathcal{D}$ up to degree d (e.g. if $\mathcal{P}$ is the tensor PCA problem).*

3. *If $\mathcal{P}$ is the* MAX CLIQUE *problem with feasible set $x \in \mathbb{R}^n$ with $\{x_i^2 = x_i\}_{i\in[n]} \cup \{x_i x_j = 0\}_{(i,j)\in E}$, then $\mathcal{P}$ is d-complete on $\mathcal{D}$ up to degree d.*

A couple of additional examples can be found in the upcoming thesis of Benjamin Weitz [Wei17]:

Proposition A.5 (Completeness of additional polynomial optimization problems) [Wei17]**.** *The following pairs of polynomial optimization problems $\mathcal{P}$ and distributions over solutions $\mathcal{D}$ are complete:*

1. *If $\mathcal{P}$ is the* densest k-subgraph *relaxation, with feasible set $x \in \mathbb{R}^n$ with $\{x_i^2 = x_i\}_{i\in[n]} \cup \{\sum_{i\in[n]} x_i = k\}$, $\mathcal{P}$ is d-complete on $\mathcal{D}$ up to degree $d \leqslant k$.*

2. *If $\mathcal{P}$ is the* sparse PCA *relaxation with sparsity k, with feasible set $x \in \mathbb{R}^n$ with $\{x_i^3 = x_i\}_{i\in[n]} \cup \{\sum_{i\in[n]} x_i^2 = k\}$, $\mathcal{P}$ is d-complete up to degree $d \leqslant k/2$.*

Proof of Corollary A.3. We verify the conditions of Theorem A.2 separately for each case.

1. The hypercube: the completeness conditions are satisfied by Proposition A.4. We choose the set of feasible solutions to contain a single point, $s = \vec{1}$, for which $\chi_\alpha^2(s) = 1$ always.

2. The hypercube with balancedness constraints: the completeness conditions are satisfied by Proposition A.4. We choose the set of feasible solutions to contain a single point, s, some perfectly balanced vector, for which $\chi_\alpha^2(s) = 1$ always.

3. The unit sphere: the completeness conditions are satisfied by Proposition A.4. We choose the set of feasible solutions to contain a single point, $s = \frac{1}{\sqrt{n}} \cdot \vec{1}$, for which $\chi_\alpha^2(s) \geqslant n^{-d}$ as long as $|\alpha| \leqslant d$, which meets the conditions of Theorem A.2.

4. The sparse hypercube: the completeness conditions are satisfied by Proposition A.5. Here, we choose the set of solutions $\mathcal{S} = \{x \in \{0,1\}^n \mid \sum_i x_i = k\}$. as long as $k > d$, for any $|\alpha| \leqslant d$ we have that $\chi_S(x)^2 = 1$ when s is 1 on α.

5. The max clique problem: the completeness conditions are satisfied by Proposition A.4. We choose the solution set $\mathcal{S}$ to be the set of $0,1$ indicators for cliques in the graph. Any α that corresponds to a non-clique in the graph has χ_α identically zero in the solution space. Otherwise, $\chi_\alpha(s)^2 = 1$ when $s \in \mathcal{S}$ is the indicator vector for the clique on α.

This concludes the proof. □

B Lower bounds on the nonzero eigenvalues of some moment matrices

In this appendix, we prove lower bounds on the magnitude of nonzero eigenvalues of covariance matrices for certain distributions over solutions. Many of these bounds are well-known, but we re-state and re-prove them here for completeness. We first define the property we want:

Definition B.1. Let $\mathcal{P}$ be a polynomial optimization problem and let $\mathcal{D}$ be the uniform distribution over the set of feasible solutions S for $\mathcal{P}$. Define the degree-$2d$ moment matrix of $\mathcal{D}$ to be $X_{\mathcal{D}} = \mathbb{E}_{x\sim\mathcal{D}}[\hat{x}^{\otimes 2d}]$, where $\hat{x} = [1\ x]^\top$.

- We say that $\mathcal{D}$ *is δ-spectrally rich up to degree* $2d$ if every nonzero eigenvalue of $X_{\mathcal{D}}$ is at least δ.

Proposition B.2 (Spectral richness of polynomial optimization problems)**.** *The following distributions over solutions $\mathcal{D}$ are polynomially spectrally rich:*

1. *If $\mathcal{D}$ is the uniform distribution over $\{\pm 1\}^n$, then $\mathcal{D}$ is polynomially spectrally rich up to degree $d \leqslant n$.*

2. *If $\mathcal{D}$ is the uniform distribution over $\alpha \cdot \mathcal{S}_{n-1}$, then $\mathcal{D}$ is polynomially spectrally rich up to degree $d \leqslant n$.*

3. *If $\mathcal{D}$ is the uniform distribution over $x \in \{1, 0\}^n$ with $\|x\|_0 = k$, then if $2d \leqslant k$, $\mathcal{D}$ is polynomially spectrally rich up to degree d.*

4. *If $\mathcal{D}$ is the uniform distribution over $x \in \{\pm 1, 0\}^n$ with $\|x\|_0 = k$, then if $2d \leqslant k$, $\mathcal{D}$ is polynomially spectrally rich up to degree d.*

Proof. In the proof of each statement, denote the $2d$th moment matrix of $\mathcal{D}$ by $X_{\mathcal{D}} \overset{\text{def}}{=} \mathbb{E}_{x\sim\mathcal{D}}[x^{\otimes 2d}]$. Because $X_{\mathcal{D}}$ is a sum of rank-1 outer-products, an eigenvector of $X_{\mathcal{D}}$ has eigenvalue 0 if and only if it is orthogonal to every solution in the support of $\mathcal{D}$, and therefore the zero eigenvectors correspond exactly to the degree at most d constraints that can be derived from the ideal constraints.

Now, let $p_1(x), \ldots, p_r(x)$ be a basis for polynomials of degree at most $2d$ in x which is orthonormal with respect to $\mathcal{D}$, so that

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}}[p_i(x)p_j(x)] = \begin{cases} 1 & i = j \\ 0 & \text{otherwise} \end{cases}$$

If $\hat{p}_i$ is the representation of p_i in the monomial basis, we have that

$$(\hat{p}_i)^\top X_{\mathcal{D}} \hat{p}_j = \mathop{\mathbb{E}}_{x\sim\mathcal{D}}[p_i(x)p_j(x)].$$

Therefore, the matrix $R = \sum_i e_i(\hat{p}_i)^\top$ diagonalizes $X_{\mathcal{D}}$,

$$RX_{\mathcal{D}}R^\top = \mathrm{Id}\,.$$

It follows that the minimum non-zero eigenvalue of $X_{\mathcal{D}}$ is equal to the smallest eigenvalue of $(RR^\top)^{-1}$, which is in turn equal to $\frac{1}{\sigma_{\max}(R)^2}$ where $\sigma_{\max}(R)$ is the largest singular value of R. Therefore, for each of these cases it suffices to bound the singular values of the change-of-basis matrix between the monomial basis and an orthogonal basis over $\mathcal{D}$. We now proceed to handle each case separately.

1. $\mathcal{D}$ uniform over hypercube: In this case, the monomial basis *is* an orthogonal basis, so R is the identity on the space orthogonal to the ideal constraints, and $\sigma_{\max}(R) = 1$, which completes the proof.

2. $\mathcal{D}$ uniform over sphere: Here, the canonical orthonormal basis the spherical harmonic polynomials. Examining an explicit characterization of the spherical harmonic polynomials (given for example in [DX13], Theorem 5.1), we have that when expressing p_i in the monomial basis, no coefficient of a monomial (and thus no entry of $\hat{p}_i$) exceeds $n^{O(d)}$, and since there are at most n^d polynomials each with $\sum_{i=0}^{d}\binom{n}{d} \leqslant n^d$ coefficients, employing the triangle inequality we have that $\sigma_{\max}(R) \leqslant n^{O(d)}$, which completes the proof.

3. $\mathcal{D}$ uniform over $\{x \in \{0, 1\}^k \mid \|x\|_0 = k\}$: In this case, the canonical orthonormal basis is the correctly normalized Young's basis (see e.g. [Fil16] Theorems 3.1,3.2 and 5.1), and agan we have that when expressing an orthonormal basis polynomial p_i in the monomial basis, no coefficient exceeds $n^{O(d)}$. As in the above case, this implies that $\sigma_{\max}(R) \leqslant n^{O(d)}$ and completes the proof.

4. $\mathcal{D}$ uniform over $\{x \in \{\pm 1, 0\}^k \mid \|x\|_0 = k\}$: Again the canonical orthonormal basis is Young's basis with a correct normalization. We again apply [Fil16] Theorems 3.1,3.2, but this time we calculate the normalization by hand: we have that in expressing each p_i, no element of the monomial basis has coefficient larger than $n^{O(d)}$ multiplied by the quantity

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}} \left[\prod_{i=1}^{d} (x_{2i-1} - x_{2i})^2 \right] = O(1).$$

This gives the desired conclusion.

□

C From Boolean to Gaussian lower bounds

In this section we show how to prove our SoS lower bounds for Gaussian PCA problems using the lower bounds for Boolean problems in a black-box fashion. The techniques are standard and more broadly applicable than the exposition here but we prove only what we need.

The following proposition captures what is needed for tensor PCA; the argument for sparse PCA is entirely analogous so we leave it to the reader.

Proposition C.1. *Let $k \in \mathbb{N}$ and let $A \sim \{\pm 1\}^{\binom{n}{k}}$ be a symmetric random Boolean tensor. Suppose that for every $A \in \{\pm 1\}^{\binom{n}{k}}$ there is a degree-d pseudodistribution $\tilde{\mathbb{E}}$ satisfying $\{\|x\|^2 = 1\}$ such that*

$$\mathop{\mathbb{E}}_{A} \tilde{\mathbb{E}} \langle x^{\otimes k}, A \rangle = C\,.$$

Let $T \sim \mathcal{N}(0,1)^{\binom{n}{k}}$ be a Gaussian random tensor. Then

$$\mathop{\mathbb{E}}_{T} \max_{\tilde{\mathbb{E}}} \tilde{\mathbb{E}} \langle x^{\otimes k}, T \rangle \geqslant \Omega(C)$$

where the maximization is over pseudodistributions of degree d which satisfy $\{\|x\|^2 = 1\}$.

Proof. For a tensor $T \in (\mathbb{R}^n)^{\otimes k}$, let $A(T)$ have entries $A(T)_\alpha = \operatorname{sign}(T_\alpha)$. Now consider

$$\mathop{\mathbb{E}}_{T} \tilde{\mathbb{E}}_{A(T)} \langle x^{\otimes k}, T \rangle = \sum_{\alpha} \mathop{\mathbb{E}}_{T} \tilde{\mathbb{E}}_{A(T)} x^\alpha T_\alpha$$

where α ranges over multi-indices of size k over $[n]$. We rearrange each term above to

$$\mathop{\mathbb{E}}_{A(T)} (\tilde{\mathbb{E}}_{A(T)} x^\alpha) \cdot \mathop{\mathbb{E}}_{T_\alpha \,|\, A(T)} T_\alpha = \mathop{\mathbb{E}}_{A(T)} (\tilde{\mathbb{E}}_{A(T)} x^\alpha) \cdot A(T)_\alpha \cdot \mathbb{E}\,|g|$$

where $g \sim \mathcal{N}(0,1)$. Since $\mathbb{E}\,|g|$ is a constant independent of n, all of this is

$$\Omega(1) \cdot \sum_{\alpha} \mathop{\mathbb{E}}_{A} \tilde{\mathbb{E}}_A x^\alpha \cdot A_\alpha = C\,.$$

□