Proc. Int. Cong. of Math. – 2018
Rio de Janeiro, Vol. 3 (3373–3408)

HIGH DIMENSIONAL ESTIMATION VIA SUM-OF-SQUARES PROOFS

Prasad Raghavendra, Tselil Schramm (צלי׳ל שרם) and David Steurer

Abstract

Estimation is the computational task of recovering a *hidden parameter* x associated with a distribution $\mathfrak{D}_x$, given a *measurement* y sampled from the distribution. High dimensional estimation problems can be formulated as system of polynomial equalities and inequalities, and thus give rise to natural probability distributions over polynomial systems.

Sum of squares proofs not only provide a powerful framework to reason about polynomial systems, but they are constructive in that there exist efficient algorithms to search for sum-of-squares proofs. The efficiency of these algorithms degrade exponentially in the degree of the sum-of-squares proofs.

Understanding and characterizing the power of sum-of-squares proofs for estimation problems has been a subject of intense study in recent years. On one hand, there is a growing body of work utilizing sum-of-squares proofs for recovering solutions to polynomial systems whenever the system is feasible. On the other hand, a broad technique referred to as *pseudocalibration* has been developed towards showing lower bounds on degree of sum-of-squares proofs. Finally, the existence of sum-of-squares refutations of a polynomial system has been shown to be intimately connected to the spectrum of associated low-degree matrix valued functions. This article will survey all of these developments in the context of estimation problems.

Contents

P. R. supported by NSF CCF-1408643 and NSF CCF-1718695.
MSC2010: 90C22.

1 Introduction

An estimation problem is specified by a family of distributions $\{\mathfrak{D}_x\}$ over $\mathbb{R}^N$ parametrized by $x \in \mathbb{R}^n$. The input consists of a sample $y \in \mathbb{R}^N$ drawn from $\mathfrak{D}_x$ for some $x \in \mathbb{R}^n$, and the goal is to recover the value of the parameter x. Here x is referred to as the *hidden variable* or the *parameter*, while the sample y is the *measurement* or the *instance*. Often, it is information theoretically impossible to recover hidden variables x in that their value is not completely determined by the measurements. Further, even if the hidden variable x is completely determined by the measurements, in many high-dimensional settings it is computationally intractable to recover x For these reasons, one often seeks to recover x approximately by minimizing the expected loss for an appropriate loss function. For example, if $\theta(y)$ denotes the estimate for x given the measurement y, a natural goal would be to minimize the expected mean-square loss given by $\mathbb{E}_{y\sim\mathfrak{D}_x}[\|\theta(y) - x\|^2]$.

Such a minimization problem can often be equivalently stated as the problem of finding a solution to a system of polynomial inequalities and equalities. By classical NP-completeness results, general polynomial systems in many variables are computationally intractable in the worst case. In the context of estimation problems, the estimation problem gives rise to a probability distribution over polynomial systems, and the goal is to reason about a typical system drawn from the distribution. If the underlying distributions are sufficiently well-behaved, polynomial systems yield an avenue to design algorithms for high-dimensional estimation problems.

The central tool that we will bring to bear on polynomial systems is that of sum-of-squares proofs. Sum-of-squares proofs yield a complete proof system to reason about polynomial systems Krivine [1964] and Stengle [1974]. More importantly, sum-of-squares proofs are constructive: the problem of finding a sum-of-squares proof can be formulated as a semidefinite program, and thus algorithms for convex optimization can be used to find a sum-of-squares proof when one exists. The computational complexity of the algorithm grows exponentially with the degree of the polynomials involved in the sum-of-squares proof. Thus, low-degree sum-of-squares proofs can be found efficiently.

Applying low-degree sum-of-squares proofs in the context of estimation problems lays open a rich family of questions. For natural distributions of polynomial systems, if a system drawn from the distribution is feasible, can one harness the sum-of-squares proofs towards actually solving the polynomial system? (surprisingly, the answer is yes!) If the system is typically infeasible, what is the smallest degree of a sum-of-squares refutation? Are there structural characterizations of the degree of sum-of-squares refutations in terms of the properties of the distribution? Is there a connection between the existence of

low-degree sum-of-squares proofs and the spectra of random matrices associated with the distribution? In the past few years, significant strides have been made on all these fronts, exposing the contours of a rich theory that lies hidden. This survey will be devoted to expounding some of the major developments in this context.

1.1 Estimation problems. We will start by describing a few estimation problems that will be recurring examples in our survey.

Example 1.1 (k-clique)**.** Fix a positive integer $k \leqslant n$. In the k-clique problem, a clique of size k is planted within a random graph drawn from the Erdős-Rényi distribution denoted $\mathbb{G}(n, 1/2)$. The goal is to recover the k clique. Formally, the structured family $\{\mathcal{G}\}$ is parametrized by subsets $S \subset \binom{[n]}{k}$. For a subset $S \in \binom{[n]}{k}$, the distribution $\mathcal{G}_S$ over $\{0,1\}^{\binom{n}{2}}$ is specified by the following sampling procedure:

- Sample a graph $G' = ([n], E(G'))$ from the Erdős-Rényi distribution $\mathbb{G}(n, 1/2)$ and set $G = ([n], E(G') \cup E(K_S))$ where K_S denotes the clique on the vertices in S. Let $y \in \{1,-1\}^{\binom{n}{2}}$ denote the natural $\{1,-1\}$-encoding of the graph G, namely, $y_{ij} = \frac{1}{2}(1 - 2\,\mathbf{1}[(i,j) \in E(G)])$ for all $i, j \in \binom{n}{2}$. Set $x := \mathbf{1}_S \in \{0,1\}^n$.

We will refer to the variables y_{ij} as *instance variables* as they specify the input to the problem. The variables x_i will be referred to as the *hidden variables*.

It is easy to see that for all $k \gg 2\log n$, the clique S can be exactly recovered with high probability given the graph G. However, there is no known polynomial time algorithm for the problem with the best algorithm being a brute force search running in time $n^{O(\log n)}$. We will now see how to encode the problem as a polynomial system by encoding the constraints one at a time, i.e.,

(1-1)

x_i are Boolean $\qquad \{x_i(1-x_i) = 0\}_{i\in[n]}$

(1-2)

if $(i,j) \notin E(G)$ then $\{i,j\}$ are not both in clique $\qquad \{(1-y_{ij})x_i x_j = 0\}_{\forall i,j \in \binom{[n]}{2}}$

(1-3)

at least k vertices in clique $\qquad \sum_{i\in[n]} x_i - k \geqslant 0$

Note that the instance variables y_{ij} are given, and the hidden variables $\{x_i\}$ are the unknowns in the polynomial system. It is easy to check that the only feasible solutions $x \in \mathbb{R}^n$ for this system of polynomial equations are Boolean vectors $x \in \{0,1\}^n$ wich are supported on cliques of size at least k in G.

For every estimation problem that we will encounter in this survey, one can associate two related computational problems termed refutation and distinguishing.

Estimation can be thought of as searching for a hidden structure within the input instance y. The goal of refutation is to certify that there is no hidden structure, when there is none. More precisely, a *null* distribution is a probability distribution over instances y for which there is no hidden structure x. For example, in the k-clique problem, the corresponding null distribution is just the Erdos-Renyi random graph $\mathbb{G}(n, 1/2)$ (without a planted clique in it). With high probability, a graph $y \sim \mathbb{G}(n, 1/2)$ has no clique with significantly more than $2 \log n$ vertices. Therefore, for a fixed $k \gg 2 \log n$, given a graph $y \sim \mathbb{G}(n, 1/2)$, the goal of a refutation algorithm is to certify that y has no clique of size k. Equivalently, the goal of a refutation algorithm is to certify the infeasibility of the associated polynomial system.

The most rudimentary computational task associated with estimation and refutation is that of distinguishing. The setup of the distinguishing problem is as follows. Fix a prior distribution π on the hidden variables $x \in \mathbb{R}^n$, which in turn induces a distribution $\mathfrak{D}_*$ on $\mathbb{R}^N$, obtained by first sampling $x \sim \pi$ and then sampling $y \sim \mathfrak{Y}_x$. The input consists of a sample y which is with equal probability drawn from the structured distribution $\mathfrak{D}_*$ or the null distribution $\mathfrak{D}_\varnothing$. The computational task is to identify which distribution the sample y is drawn from, with a probability of success $\frac{1}{2} + \delta$ for some constant $\delta > 0$. For example, the structured distribution for k-clique is obtained by setting the prior distribution of x to be uniform on subsets of size k. In the distinguishing problem, the input is a graph drawn from either $\mathfrak{D}_*$ or the null distribution $\mathbb{G}(n, 1/2)$ and the algorithm is required to identify the distribution. For every problem included in this survey, the distinguishing task is formally no harder than estimation or refutation, i.e., the existence of algorithms for estimation or refutation immediately implies a distinguishing algorithm.

Example 1.2. (tensor PCA) The family of structured distributions $\{\mu_x\}$ is parametrized by unit vectors $x \in \mathbb{R}^n$. A sample from μ_x consists of a symmetric 4-tensor $y = x^{\otimes 4} + \zeta$ where $\zeta \in \mathbb{R}^{n \times n \times n \times n}$ is a symmetric 4-tensor whose entries are i.i.d Gaussian random variables sampled from $N(0, \sigma^2)$. The goal is to recover a vector x' that is close as possible to x.

A canonical strategy to recover x given $y = x^{\otimes 4} + \zeta$ is to maximize the degree-4 polynomial associated with the symmetric 4 tensor y. Specifically, if we set

$$x' = \text{argmax}_{\|x'\| \leqslant 1} \langle y, x'^{\otimes 4} \rangle$$

then one can show that $\|x - x'\|_2 \leqslant O(n^{1/2} \cdot \sigma)$ with high probability over ζ. If $y \sim \mathfrak{Y}_x$ then $\langle y, x^{\otimes 4} \rangle = 1$. Furthermore, when $\sigma \ll n^{-1/2}$ it can be shown that $x \in \mathbb{R}^n$ is close to the unique maximizer of the function $\phi(z) = \langle y, z^{\otimes 4} \rangle$. So the problem of recovering

x can be encoded as following polynomial system:

$$\|x\|^2 \leqslant 1, \qquad \sum_{i,j,k,\ell \in [n]^4} y_{ijk\ell} x_i x_j x_k x_\ell \geqslant \tau. \tag{1-4}$$

where $\tau := 1$.

In the distinguishing and refutation versions of this problem, we will take the *null* distribution $\mathfrak{D}_{\varnothing}$ to be the distribution over 4-tensors with independent Gaussian entries sampled from $N(0, \sigma^2)$ (matching the distribution of the noise ζ from $\mathfrak{D}_*$). For a 4-tensor y, the maximum of $y(x) = \langle x^{\otimes 4}, y\rangle$ over the unit ball is referred to as the *injective tensor norm* of the tensor y, and is denoted by $\|y\|_{\text{inj}}$. If $y \sim \mathfrak{D}_{\varnothing}$ then $\|y\|_{\text{inj}} \leqslant O\left(n^{1/2} \cdot \sigma\right)$ with high probability over choice of y . Thus when $\sigma \ll n^{-1/2}$, the refutation version of the tensor PCA problem reduces to certifying an upper bound on $\|y\|_{\text{inj}}$. If we could compute $\|y\|_{\text{inj}}$ exactly, then we can certify that $y \sim \mathfrak{D}_{\varnothing}$ for σ as large as $\sigma = O(n^{-1/2})$. The injective tensor norm is known to be computationally intractable in the worst case Gurvits [2003], Gharibian [2010], and Barak, Brandão, Harrow, Kelner, Steurer, and Zhou [2012].

Example 1.3. (Matrix & Tensor Completion) In matrix completion, the hidden parameter is a rank-r matrix $X \in \mathbb{R}^{n \times n}$. For a parameter X, the measurement consists of a partial matrix revealing a subset of entries of X, namely X_Ω for a subset $\Omega \subset [n] \times [n]$ with $|\Omega| = m$. The probability distribution μ_X over measurements is obtained by picking the set Ω to be a uniformly random subset of m entries. To formulate a polynomial system for recovering a rank-r matrix consistent with the measurement X_Ω, we will use a $n \times r$ matrix of variables B, and write the following system of constraints on it:

$$(BB^T)_\Omega = X_\Omega \qquad (BB^T \text{ is consistent with measurement})$$

Tensor completion is the analogous problem with X being a higher-order tensor namely, $X = \sum_{i=1}^r a_i^{\otimes k}$ for some fixed $k \in \mathbb{N}$. The corresponding polynomial system is again over a $n \times r$ matrix of variables B with columns $b_1, \ldots, b_r$ and the following system of constraints,

$$\left(\sum_{i \in [r]} b_i^{\otimes k}\right)_\Omega = X_\Omega \qquad (\textstyle\sum_{i=1}^r b_i^{\otimes k} \text{ is consistent with measurement})$$

1.2 Sum-of-squares proofs. The sum-of-squares (SoS) proof system is a restricted class of proofs for reasoning about polynomial systems. Fix a set of polynomial inequalities $\mathfrak{A} = \{p_i(x) \geqslant 0\}_{i \in [m]}$ in variables $x_1, \ldots, x_n$. We will refer to these inequalities as the *axioms*. Starting with the axioms $\mathfrak{A}$, a sum-of-squares proof of $q(x) \geqslant 0$ is given by

an identity of the form,

$$\left(\sum_{i\in[m']} b_i^2(x)\right)\cdot q(x) = \sum_j s_j^2(x) + \sum_{i\in[m]} a_i^2(x)\cdot p_i(x)\,,$$

where $\{s_j(x)\}, \{a_i(x)\}_{i\in[m]}, \{b_i(x)\}_{i\in[m']}$ are real polynomials. It is clear that any identity of the above form manifestly certifies that the polynomial $q(x) \geqslant 0$, whenever each $p_i(x) \geqslant 0$ for real x. The degree of the sum-of-squares proof is the maximum degree of all the summands, i.e., $\max\{\deg(s_j^2), \deg(a_i^2 p_i)\}_{i,j}$.

The notion extends naturally to polynomial systems that involves a set of equations $\{r_i(x) = 0\}$ along with a set of inequalities $\{p_i(x) \geqslant 0\}$. A syntactic approach to extend the definition would be to replace each equality $r_i(x) = 0$ by a pair of inequalities $r_i(x) \geqslant 0$ and $-r_i(x) \geqslant 0$.

We will the use the notation $\mathcal{A} \vdash^x_d \{q(x) \geqslant 0\}$ to denote that the assertion that, there exists a degree d sum-of-squares proof of $q(x) \geqslant 0$ from the set of axioms $\mathcal{A}$. The superscript x in the notation $\mathcal{A} \vdash^x_d \{q(x) \geqslant 0\}$ indicates that the sum-of-squares proof is an identity of polynomials where x is the formal variable. We will drop the subscript or superscript when it is clear from the context, and just write $\mathcal{A} \vdash \{q(x) \geqslant 0\}$. Sum-of-squares proofs can also be used to certify the infeasibility, a.k.a., *refute* the polynomial system. In particular, a degree d sum-of-squares refutation of a polynomial system $\{p_i(x) \geqslant 0\}_{i\in[m]}$ is an identity of the form,

$$-1 = \sum_{i\in[k]} s_i^2(x) + \sum_{i\in[m]} a_i^2(x)\cdot p_i(x) \tag{1-5}$$

where $\max\{\deg(s_j^2), \deg(a_i^2 p_i)\}_{i,j}$ is at most d.

Sum-of-square proof system have been an object of study starting with the work of Hilbert and Minkoswki more than a century ago (see Reznick [2000] for a survey). With no restriction on degree, Stengle's Positivestellensatz imply that sum-of-squares proofs form a complete proof system, i.e., if the axioms $\mathcal{A}$ imply $q(x) \geqslant 0$, then there is a sum-of-squares proof of this fact.

The algorithmic implications of sum-of-squares proof system were realized starting with the work of Parrilo [2000] and Lasserre [2000/01], who independently arrived at families of algorithms for polynomial optimization using semidefinite programming (SDP). Specifically, these works observed that semidefinite programming can be used to find a degree-d sum-of-squares proof in time $n^{O(d)}$, if there exists one. This family of algorithms (called a hierarchy, as we have algorithms for each even integer degree d) are referred to as the low-degree sum-of-squares SDP hierarchy.

The SoS hierarchy has since emerged as one of the most powerful tools for algorithm design. On the one hand, a vast majority of algorithms in combinatorial optimization and

approximation algorithms developed over several decades can be systematically realized as being based on the first few levels of this hierarchy. Furthermore, the low-degree SoS SDP hierarchy holds the promise of yielding improved approximations to NP-hard combinatorial optimization problems, approximations that would beat the long-standing and universal barrier posed by the notorious unique games conjecture Trevisan [2012] and Barak and Steurer [2014].

More recently, the low-degree SoS SDP hierarchy has proved to be a very useful tool in designing algorithms for high-dimensional estimation problems, wherein the inputs are drawn from a natural probability distribution. For this survey, we organize the recent work on this topic into three lines of work.

- *When the polynomial system for an estimation problem is feasible, can sum-of-squares proofs be harnessed to retrieve the solution?* The answer is YES for many estimation problems including tensor decomposition, matrix and tensor completion. Furthermore, there is a simple and unifying principle that underlies all of these applications. Specifically, the underlying principle asserts that if there is a low-degree SoS proof that all solutions to the system are close to the hidden variable x, then low-degree SoS SDP can be used to actually retrieve x. We will discuss this broad principle and many of its implications in Section 2.

- *When the polynomial system is infeasible, what is the smallest degree at which it admits sum-of-squares proof?* The degree of the sum-of-squares refutation is critical for the run-time of the SoS SDP based algorithm. Recent work by Barak et al. Barak, Hopkins, Kelner, P. Kothari, Moitra, and A. Potechin [2016] introduces a technique referred to as "pseudocalibration" for proving lower bounds on the degree of SoS refutation, developed in the context of the work on k-clique. Section Section 3 is devoted to the heuristic technique of pseudocalibration, and the mystery surrounding its effectiveness.

- *Can the existence of degree-d of sum-of-square refutations be characterized in terms of properties of the underlying distribution?* In Section 4, we will discuss a result that shows a connection between the existence of low-degree sum-of-squares refutations and the spectra of certain low-degree matrices associated with the distribution. This connection implies that under fairly mild conditions, the SoS SDP based algorithms are no more powerful than a much simpler class of algorithms referred to as *spectral algorithms*. Roughly speaking, a spectral algorithm proceeds by constructing a matrix $M(x)$ out of the input instance x, and then using the eigenvalues of the matrix $M(x)$ to recover the desired outcome.

Notation. For a positive integer n, we use $[n]$ to denote the set $\{1, \dots, n\}$. We sometimes use $\binom{[n]}{d}$ to denote the set of all subsets of $[n]$ of size d, and $[n]^{\leq d}$ to denote the set of all multi-subsets of cardinality at most d.

If $x \in \mathbb{R}^n$ and $A \subset [n]$ is a multiset, then we will use the shorthand x^A to denote the monomial $x^A = \prod_{i \in A} x_i$. We will also use $x^{\leq d}$ to denote the $N \times 1$ vector containing all monomials in x of degree at most d (including the constant monomial 1), where $N = \sum_{i=0}^{d} n^i$. Let $\mathbb{R}[x]_{\leq d}$ denote the space of polynomials of degree at most d in variables x.

For a function $f(n)$, we will say $g(n) = O(f(n))$ if $\lim_{n\to\infty} \frac{g(n)}{f(n)} \leqslant C$ for some universal constant C. We say that $f(n) \ll g(n)$ if $\lim_{n\to\infty} \frac{f(n)}{g(n)} = 0$.

If μ is a distribution over the probability space $\mathcal{S}$, then we use the notation $x \sim \mu$ for $x \in \mathcal{S}$ sampled according to μ. For an event $\mathcal{E}$, we will use $\mathbf{1}[\mathcal{E}]$ as the indicator that $\mathcal{E}$ occurs. We use $\mathbb{G}(n, 1/2)$ to denote the Erdős-Rényi distribution with parameter $1/2$, or the distribution over graphs where each edge is included independently with probability $1/2$.

If M is an $n \times m$ matrix, we use $\lambda_{\max}(M)$ to denote M's largest eigenvector. When $n = m$, then $\mathrm{Tr}(M)$ denotes M's trace. If N is an $n \times m$ matrix as well, then we use $\langle M, N \rangle = \mathrm{Tr}(MN^\top)$ to denote the *matrix inner product*. We use $\|M\|_{\mathrm{F}}$ to denote the Frobenius norm of M, $\|M\|_{\mathrm{F}} = \langle M, M \rangle$. For a subset $S \subset [n]$, we will use $\mathbf{1}_S$ to denote the $\{0,1\}$ indicator vector of S in $\mathbb{R}^n$. We will also use $\mathbf{1}$ to denote the all-1's vector.

For two matrices A, B we use $A \otimes B$ to denote both the Kronecker product of A and B, and the order-4 tensor given by taking $A \otimes B$ and reshaping it with modes for the rows and columns of A and of B. We also use $A^{\otimes k}$ to denote the k-th Kronecker power of A, $A \otimes A \otimes \cdots \otimes A$.

Pseudoexpectations. If there is no degree-d refutation, the dual semidefinite program gives rise to a linear functional over degree d polynomials which we term a *pseudoexpectation*. Formally, a pseudoexpectation $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leq d} \to \mathbb{R}$ is a linear functional over polynomials of degree at most d with the properties that $\tilde{\mathbb{E}}[1] = 1$, $\tilde{\mathbb{E}}[p(x)a^2(x)] \geqslant 0$ for all $p \in \mathcal{P}$ and polynomials a such that $\deg(a^2 \cdot p) \leqslant d$, and $\tilde{\mathbb{E}}[q(x)^2] \geqslant 0$ whenever $\deg(q^2) \leqslant d$.

Claim 1.4. Suppose there exists a degree d pseudoexpectation $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$ for the polynomial system $\mathcal{P} = \{p_i(x) \geqslant 0\}_{i \in [m]}$, then $\mathcal{P}$ does not admit a degree d refutation.

Proof. Suppose $\mathcal{P}$ admits a degree d refutation. Applying the pseudoexpectation operator $\tilde{\mathbb{E}}$ to the left-hand-side of Equation (1-5), we have -1. Applying $\tilde{\mathbb{E}}$ to the right-hand-side of Equation (1-5), the first summand must be non-negative by definition of $\tilde{\mathbb{E}}$ since it is a sum of squares, and the second summand is non-negative, since we assumed that $\tilde{\mathbb{E}}$ satisfies the constraints of $\mathcal{P}$. This yields a contradiction. □

2 Algorithms for high-dimensional estimation

In this section, we prove a algorithmic meta-theorem for high-dimensional estimation that provides a unified perspective on the best known algorithms for a wide range of estimation problems. Through this unifying perspective we are also able to obtain algorithms with significantly than what's known to be possible with other methods.

2.1 Algorithmic meta-theorem for estimation. We consider the following general class of estimation problems, which will turn out to capture a plethora of interesting problems in a useful way: In this class, an estimation problem[1] is specified by a set $\mathcal{P} \subseteq \mathbb{R}^n \times \mathbb{R}^m$ of pairs (x, y), where x is called *parameter* and y is called *measurement*. Nature chooses a pair $(x^*, y^*) \in \mathcal{P}$, we are given the measurement y^* and our goal is to (approximately) recover the parameter x^*.

For example, we can encode compressed sensing with measurement matrix $A \in \mathbb{R}^{m \times n}$ and sparsity bound k by the following set of pairs,

$$\mathcal{P}_{A,k} = \{(x, y) \mid y = Ax,\ x \in \mathbb{R}^n \text{ is } k\text{-sparse}\} .$$

Similarly, we can encode matrix completion with observed entries $\Omega \subseteq [n] \times [n]$ and rank bound r by the set of pairs,

$$\mathcal{P}_{\Omega,r} = \{(X, X_\Omega) \mid X \in \mathbb{R}^{n \times n},\ \operatorname{rank} X \leqslant r\} .$$

For both examples, the measurement was a simple (linear) function of the parameter.

Identifiability. In general, an estimation problem $\mathcal{P} \subseteq \mathbb{R}^n \times \mathbb{R}^m$ may be ill-posed in the sense that, even ignoring computational efficiency, it may not be possible to (approximately) recover the parameter for a measurement y because we have $(x, y), (x', y) \in \mathcal{P}$ for two far-apart parameters x and x'.

For a pair $(x, y) \in \mathcal{P}$, we say that y *identifies* x *exactly* if $(x', y) \notin \mathcal{P}$ for all $x' \neq x$. Similarly, we say that y *identifies* x *up to error* $\varepsilon > 0$ if $\|x - x'\| \leqslant \varepsilon$ for all $(x', y) \in \mathcal{P}$. We say that x is identifiable (up to error ε) if every $(x, y) \in \mathcal{P}$ satisfies that y identifies x (up to error ε).

For example, for compressed sensing $\mathcal{P}_{A,k}$, it is not difficult to see that every k-sparse vector is identifiable if every subset of at most $2k$ columns of A is linearly independent. For tensor decomposition, it turns out, for example, that the observation

[1] In contrast to the discussion of estimation problems in Section 1, for every parameter, we have a set of possible measurements as opposed to a distribution over measurements. We can model distributions over measurements in this way by considering a set of "typical measurements". The viewpoint in terms of sets of possible measurements will correspond more closely to the kind of algorithms we consider.

$f(x) = \sum_{i=1}^{r} x_i^{\otimes 3}$ is enough to identify $x \in \mathbb{R}^{n \times r}$ (up to a permutation of its columns) if the columns $x_1, \ldots, x_r \in \mathbb{R}^n$ of x are linearly independent.

Identifiability proofs to efficient algorithms. By itself, identifiability typically only implies that there exists an inefficient algorithm to recover a vector x close to the parameter x^* from the observation y^*. But perhaps surprisingly, the notion of identifiability in a broader sense can also help us understand if there exists an efficient algorithm for this task. Concretely, if the *proof of identifiability* is captured by the sum-of-squares proof system at low degree, then there exists an efficient algorithm to (approximately) recover x from y.

In order to formalize this phenomenon, let the set $\mathcal{P} \subseteq \mathbb{R}^n \times \mathbb{R}^m$ be be described by polynomial equations

$$\mathcal{P} = \{(x, y) \mid \exists z.\ p(x, y, z) = 0\} ,$$

where $p = (p_1, \ldots, p_t)$ is a vector-valued polynomial and z are auxiliary variables.[2] (In other words, $\mathcal{P}$ is a projection of the variety given by the polynomials $p_1, \ldots, p_t$.) The following theorem shows that there is an efficient algorithm to (approximately) recover x^* given y^* if there exists a low-degree proof of the fact that the equation $p(x, y^*, z) = 0$ implies that x is (close to) x^*.

Theorem 2.1 (Meta-theorem for efficient estimation)**.** *Let p be a vector-valued polynomial and let the triples (x^*, y^*, z^*) satisfy $p(x^*, y^*, z^*) = 0$. Suppose $\mathcal{A} \vdash_{\ell}^{x,z} \{\|x^* - x\|^2 \leq \varepsilon\}$, where $\mathcal{A} = \{p(x, y^*, z) = 0\}$. Then, every level-$\ell$ pseudo-distribution D consistent with the constraints $\mathcal{A}$ satisfies*

$$\left\| x - \tilde{\mathbb{E}}_{D(x,z)} x \right\|^2 \leq \varepsilon .$$

Furthermore, for every $\ell \in \mathbb{N}$, there exists a polynomial-time algorithm (with running time $n^{O(\ell)}$)[3] that given a vector-valued polynomial p and a vector y outputs a vector $\hat{x}(y)$ with the following guarantee: if $\mathcal{A} \vdash_{\ell}^{x,z} \{\|x^* - x\|^2 \leq \varepsilon\}$ with a proof of bit-complexity at most n^{ℓ}, then $\|x^* - \hat{x}(y^*)\|^2 \leq \varepsilon + 2^{-n^{\ell}}$.

Despite not being explicitly stated, the above theorem is the basis for many recent advances in algorithms for estimation problems through the sum-of-squares method Barak, Kelner, and Steurer [2015, 2014], Hopkins, Shi, and Steurer [2015], Ma, Shi, and Steurer

[2] We allow auxiliary variables here because they might make it easier to describe the set $\mathcal{P}$. The algorithms we consider depend on the algebraic description of $\mathcal{P}$ we choose and different descriptions can lead to different algorithmic guarantees. In general, it is not clear what is the best possible description. However, typically, the more auxiliary variables the better.

[2016], Barak and Moitra [2016], A. Potechin and Steurer [2017], P. K. Kothari, Steinhardt, and Steurer [2018], and Hopkins and Li [2018].

2.2 Matrix and tensor completion. In matrix completion, we observe a few entries of a matrix and the goal is to fill in the missing entries. This problem is studied extensively both from practical and theoretical perspectives. One of its practical application is in recommender systems, which was the basis of the famous Netflix Prize competition. Here, we may observe a few movie ratings for each user and the goal is to infer a user's preferences for movies that the user hasn't rated yet.

In terms of provable guarantees, the best known polynomial time algorithm for matrix completion is based on a semidefinite programming relaxation. Let $X = \sum_{i=1}^{r} \sigma_i \cdot u_i v_i^\top \in \mathbb{R}^{n \times n}$ be a rank-r matrix such that its left and right singular vectors $u_1, \ldots, u_r, v_1, \ldots, v_r \in \mathbb{R}^n$ are μ-incoherent[4], i.e., they satisfy $\langle u_i, e_j \rangle^2 \leqslant \mu/n$ and $\langle v_i, e_j \rangle^2 \leqslant \mu/n$ for all $i \in [r]$ and $j \in [n]$. The algorithm observes the partial matrix X_Ω that contains a random cardinality m subset $\Omega \subseteq [n] \times [n]$ of the entries of X. If $m \geqslant \mu r n \cdot O(\log n)^2$, then with high probability over the choice of Ω the algorithm recovers X exactly Candès and Recht [2009], Gross [2011], Recht [2011], and Chen [2015]. This bound on m is best-possible in several ways. In particular, $m \geqslant \Omega(rn)$ appears to be necessary because an n-by-n rank-r matrix has $\Omega(r \cdot n)$ degrees of freedom (the entries of its singular vectors).

In this section, we will show how the above algorithm is captured by sum-of-squares and, in particular, Theorem 2.1. We remark that this fact follows directly by inspecting the analysis of the original algorithm Candès and Recht [2009], Gross [2011], Recht [2011], and Chen [2015]. The advantage of sum-of-squares here is two-fold: First, it provides a unified perspective on algorithms for matrix completion and other estimation problems. Second, the sum-of-squares approach for matrix completion extends in a natural way to tensor completion (in a way that the original approach for matrix completion does not).

Identifiability proof for matrix completion. For the sake of clarity, we consider a simplified setup where the matrix X is assumed to be a rank-r projector so that $X = \sum_{i=1}^{r} a_i a_i^\top$ for μ-incoherent orthonormal vectors $a_1, \ldots, a_r \in \mathbb{R}^n$. The following theorem shows that, with high probability over the choice of Ω, the matrix X is identified by the partial matrix X_Ω. Furthermore, the proof of this fact is captured by sum-of-squares. Together with Theorem 2.1, the following theorem implies that there exists a polynomial-time algorithm to recover X from X_Ω.

[3] In order to be able to state running times in a simple way, we assume that the total bit-complexity of (x, y, z) and the vector-valued polynomial p (in the monomial basis) is bounded by a fixed polynomial in n.

[4] Random unit vectors satisfy this notion of μ-incoherence for $\mu \leqslant O(\log n)$. In this sense, incoherent vectors behave similar to random vectors.

Theorem 2.2 (implicit in Candès and Recht [2009], Gross [2011], Recht [2011], and Chen [2015])**.** *Let $X = \sum_{i=1}^{r} a_i a_i^\top \in \mathbb{R}^{n\times n}$ be an r-dimensional projector and $a_1, \dots, a_r \in \mathbb{R}^n$ orthonormal with incoherence $\mu = \max_{i,j} n \cdot \langle a_i, e_j\rangle^2$. Let $\Omega \subseteq [n]\times[n]$ be a random symmetric subset of size $|\Omega| = m$. Consider the system of polynomial equations in n-by-r matrix variable B,*

$$\mathfrak{A} = \left\{ \left(BB^\top\right)_\Omega = X_\Omega,\ B^\top B = \mathrm{Id}_r \right\} .$$

Suppose $m \geqslant \mu r n \cdot O(\log n)^2$. Then, with high probability over the choice of Ω,

$$\mathfrak{A} \vdash^{\frac{B}{4}} \left\{ \left\| BB^\top = X \right\|_{\mathrm{F}} = 0 \right\} .$$

Proof. The analyses of the aforementioned algorithm for matrix completion Candès and Recht [2009], Gross [2011], Recht [2011], and Chen [2015] show the following: with high probability over the choice of Ω, there exists[5] a symmetric matrix M with $M_{\bar{\Omega}} = 0$ and $0.9(\mathrm{Id}_n - X) \preceq M - X \preceq 0.9(\mathrm{Id}_n - X)$. As we will see, this matrix also implies that the above proof of identifiability exists.

Since $0 \preceq X$ and $X - 0.9(\mathrm{Id}_n - X) \preceq M$, we have

$$\langle M, X\rangle \geqslant \langle X, X\rangle - 0.9\langle \mathrm{Id}_n - X, X\rangle = \langle X, X\rangle = r .$$

Since $M_{\bar{\Omega}} = 0$ and $\mathfrak{A}$ contains the equation $(BB^\top)_\Omega = X_\Omega$, we have $\mathfrak{A} \vdash^{B} \langle M, BB^\top\rangle = \langle M, X\rangle \geqslant r$. At the same time, we have

$$\mathfrak{A} \vdash \langle M, BB^\top\rangle \leqslant \langle X, BB^\top\rangle + 0.9\langle \mathrm{Id}_n - X, BB^\top\rangle = 0.1\langle X, BB^\top\rangle + 0.9r ,$$

where the first step uses $M \preceq X + 0.9(\mathrm{Id} - X)$ and the second step uses $\mathfrak{A} \vdash \langle \mathrm{Id}_n, BB^\top\rangle = r$ because $\langle \mathrm{Id}_n, BB^\top\rangle = \mathrm{Tr}\, B^\top B$ and $\mathfrak{A}$ contains the equation $B^\top B = \mathrm{Id}_r$. Combining the lower and upper bound on $\langle M, BB^\top\rangle$, we obtain

$$\mathfrak{A} \vdash \langle X, BB^\top\rangle \geqslant r .$$

Together with the facts $\|X\|_{\mathrm{F}}^2 = r$ and $\mathfrak{A} \vdash \left\|BB^\top\right\|_{\mathrm{F}}^2 = r$, we obtain $\mathfrak{A} \vdash \left\|X - BB^\top\right\|_{\mathrm{F}}^2 = 0$ as desired. □

[5] Current proofs of the existence of this matrix proceed by an ingenious iterative construction of this matrix (alternatingly projecting to two affine subspaces). The analysis of this iterative construction is based on matrix concentration bounds. We refer to prior literature for details of this proof Gross [2011], Recht [2011], and Chen [2015].

Identifiability proof for tensor completion. Tensor completion is the analog of matrix completion for tensors. We observe a few of the entries of an unknown low-rank tensor and the goal is to fill in the missing entries. In terms of provable guarantees, the best known polynomial-time algorithms are based on sum-of-squares, both for exact recovery A. Potechin and Steurer [2017] (of tensors with orthogonal low-rank decompositions) and approximate recovery Barak and Moitra [2016] (of tensors with general low-rank decompositions).

Unlike for matrix completion, there appears to be a big gap between the number of observed entries required by efficient and inefficient algorithms. For 3-tensors, all known efficient algorithms require $r \cdot \tilde{O}(n^{1.5})$ observed entries (ignoring the dependence on incoherence) whereas information-theoretically $r \cdot O(n)$ observed entries are enough. The gap for higher-order tensors becomes even larger. It is an interesting open question to close this gap or give formal evidence that the gap is inherent.

As for matrix completion, we consider the simplified setup that the unknown tensor has the form $X = \sum_{i=1}^{r} a_i^{\otimes 3}$ for incoherent, orthonormal vectors $a_1, \ldots, a_r \in \mathbb{R}^n$. The following theorem shows that with high probability, X is identifiable from $rn^{1.5} \cdot (\mu \log n)^{O(1)}$ random entries of X and this fact has a low-degree sum-of-squares proof.

Theorem 2.3 (A. Potechin and Steurer [2017]). *Let $a_1, \ldots, a_r \in \mathbb{R}^n$ orthonormal vectors with incoherence $\mu = \max_{i,j} n \cdot \langle a_i, e_j \rangle^2$ and let $X = \sum_{i=1}^{r} a_i^{\otimes 3}$ be their 3-tensor. Let $\Omega \subseteq [n]^3$ be a random symmetric subset of size $|\Omega| = m$. Consider the system of polynomial equations in n-by-r matrix variable B with columns $b_1, \ldots, b_r$,*

$$\mathcal{A} = \left\{ \left(\sum_{i=1}^{r} b_i^{\otimes 3} \right)_{\Omega} = X_{\Omega},\ B^{\mathsf{T}} B = \mathrm{Id}_r \right\}$$

Suppose $m \geq rn^{1.5} \cdot (\mu \log n)^{O(1)}$. Then, with high probability over the choice of Ω,

$$\mathcal{A} \vdash_{O(1)}^{B} \left\{ \left\| \sum_{i=1}^{r} b_i^{\otimes 3} - X \right\|_{\mathrm{F}}^2 = 0 \right\}$$

2.3 Overcomplete tensor decomposition. Tensor decomposition refers to the following general class of estimation problems: Given (a noisy version of) a k-tensor of the form $\sum_{i=1}^{r} a_i^{\otimes k}$, the goal is to (approximately) recover one, most, or all of the component vectors $a_1, \ldots, a_r \in \mathbb{R}^n$. It turns out that under mild conditions on the components $a_1, \ldots, a_r$, the noise, and the tensor order k, this estimation task is possible information theoretically. For example, generic components $a_1, \ldots, a_r \in \mathbb{R}^n$ with $r \leqslant \Omega(n^2)$ are

identified by their 3-tensor $\sum_{i=1}^{r} a_i^{\otimes 3}$ Chiantini and Ottaviani [2012] (up to a permutation of the components). Our concern will be what conditions on the components, the noise, and the tensor order allow us to efficiently recover the components.

Besides being significant in its own right, tensor decomposition is a surprisingly versatile and useful primitive to solve other estimation problems. Concrete examples of problems that can be reduced to tensor decomposition are latent Dirichlet allocation models, mixtures of Gaussians, independent component analysis, noisy-or Bayes nets, and phylogenetic tree reconstruction Lathauwer, Castaing, and Cardoso [2007], Mossel and Roch [2005], Anandkumar, Foster, Hsu, S. Kakade, and Liu [2012], Hsu and S. M. Kakade [2013], Bhaskara, Charikar, Moitra, and Vijayaraghavan [2014], Barak, Kelner, and Steurer [2015], Ma, Shi, and Steurer [2016], and Arora, Ge, Ma, and Risteski [2016]. Through these reductions, better algorithms for tensor decomposition can lead to better algorithms for a large number of other estimation problems.

Toward better understanding the capabilities of efficient algorithms for tensor decomposition, we focus in this section on the following more concrete version of the problem.

Problem 2.4 (Tensor decomposition, one component, constant error)**.** Given an order-k tensor $\sum_{i=1}^{r} a_i^{\otimes k}$ with component vectors $a_1, \ldots, a_r \in \mathbb{R}^n$, find a vector $u \in \mathbb{R}^n$ that is close[6] to one of the component vectors in the sense that $\max_{i \in [r]} \frac{1}{\|a_i\| \cdot \|u\|} |\langle a_i, u \rangle| \geqslant 0.9$.

Algorithms for Problem 2.4 can often be used to solve a-priori more difficult versions of the tensor decomposition that ask to recover most or all of the components or that require the error to be arbitrarily small.

A classical spectral algorithm attributed to Harshman [1970] and Leurgans, Ross, and Abel [1993] can solve Problem 2.4 for up to $r \leqslant n$ generic components if the tensor order is at least 3. (Concretely, the algorithm works for 3-tensors with linearly independent components.) Essentially the same algorithms works up to $\Omega(n^2)$ generic[7] components if the tensor order is at least 5. A more sophisticated algorithm Lathauwer, Castaing, and Cardoso [2007] solves Problem 2.4 for up to $\Omega(n^2)$ generic[8] components if the tensor order is at least 4. However, these algorithms and their analyses break down if the tensor order is only 3 and the number of components issue $n^{1+\Omega(1)}$, even if the components are random vectors.

In this and the subsequent section, we will discuss a polynomial-time algorithm based on sum-of-squares that goes beyond these limitations of previous approaches.

[6] This notion of closeness ignores the sign of the components. If the tensor order is odd, the sign can often be recovered as part of some postprocessing. If the tensor order is even, the sign of the components is not identified.

[7] Here, the vectors $a_1^{\otimes 2}, \ldots, a_r^{\otimes 2}$ are assumed to be linearly independent.

[8] Concretely, the vectors $\{a_i^{\otimes 2} \otimes a_j^{\otimes 2} \mid i \neq j\} \cup \{(a_i \otimes a_j)^{\otimes 2} \mid i \neq j\}$ are assumed to be linearly independent.

Theorem 2.5 (Ma, Shi, and Steurer [2016] building on Barak, Kelner, and Steurer [2015], Ge and Ma [2015], and Hopkins, Schramm, Shi, and Steurer [2016])**.** *There exists a polynomial-time algorithm to solve Problem 2.4 for tensor order 3 and $\tilde{\Omega}(n^{1.5})$ components drawn uniformly at random from the unit sphere.*

The strategy for this algorithm consists of two steps:

1. use sum-of-squares in order to lift the given order-3 tensor to a noisy version of the order-6 tensor with the same components,

2. apply Jennrich's classical algorithm to decompose this order-6 tensor.

While Problem 2.4 falls outside of the scope of Theorem 2.1 (Meta-theorem for efficient estimation) because the components are only identified up to permutation, the problem of lifting a 3-tensor to a 6-tensor with the same components is captured by Theorem 2.1. Concretely, we can formalize this lifting problem as the following set of parameter–measurement pairs,

$$\mathcal{P}_{3,6;r} = \left\{ (x, y) \;\middle|\; x = \sum_{i=1}^{r} a_i^{\otimes 6},\ y = \sum_{i=1}^{r} a_i^{\otimes 3},\ a_1, \dots, a_r \in \mathbb{R}^n \right\} \subseteq \mathbb{R}^{n^6} \times \mathbb{R}^{n^3}.$$

In Section 2.4, we give the kind of sum-of-squares proofs that Theorem 2.1 requires in order to obtain an efficient algorithm to solve the above estimation problem of lifting 3-tensors to 6-tensors with the same components.

The following theorem gives an analysis of Jennrich's algorithm that we can use to implement the second step of the above strategy for Theorem 2.5.

Theorem 2.6 (Ma, Shi, and Steurer [2016] and Schramm and Steurer [2017])**.** *There exists $\varepsilon > 0$ and a randomized polynomial-time algorithm that given a 3-tensor $T \in (\mathbb{R}^n)^{\otimes 3}$ outputs a unit vector $u \in \mathbb{R}^n$ with the following guarantees: Let $a_1, \dots, a_r \in \mathbb{R}^n$ be unit vectors with orthogonality defect $\|\mathrm{Id}_r - A^{\mathsf{T}}A\| \leqslant \varepsilon$, where $A \in \mathbb{R}^{n \times r}$ is the matrix with columns $a_1, \dots, a_r$. Suppose $\left\| T - \sum_i a_i^{\otimes 3} \right\|_{\mathrm{F}}^2 \leqslant \varepsilon \cdot r$ and $\max\{\|T\|_{\{1,3\}\{2\}}, \|T\|_{\{1\}\{2,3\}}\} \leqslant 10$. Then, with at least inverse polynomial probability, $\max_{i \in [r]} \langle a_i, u \rangle \geqslant 0.9$.*

2.4 Tensor decomposition: lifting to higher order. In this section, we give low-degree sum-of-squares proofs of identifiability for the different version of the estimation problem of lifting 3-tensors to 6-tensors with the same components. These sum-of-squares proofs are a key ingredient of the algorithms for overcomplete tensor decomposition discussed in Section 2.3.

We first consider the problem of lifting 3-tensors with orthonormal components. By itself, this lifting theorem cannot be used for overcomplete tensor decomposition. However it turns out that this special case best illustrates the basic strategy for lifting tensors to higher-order tensors with the same components.

Orthonormal components. The following lemma shows that for orthonormal components, the 3-tensor identifies the 6-tensor with the same set of components and that this fact has a low-degree sum-of-squares proof.

Lemma 2.7. *Let $a_1, \dots, a_r \in \mathbb{R}^n$ be orthonormal. Let $\mathfrak{A} = \{\sum_{i=1}^r a_i^{\otimes 3} = \sum_{i=1}^r b_i^{\otimes 3}, B^\top \cdot B = \mathrm{Id}\}$, where B is an n-by-r matrix of variables and $b_1, \dots, b_r$ are the columns of B. Then,*

$$\mathfrak{A} \vdash_{12} \left\{ \left\| \sum_{i=1}^r a_i^{\otimes 6} - \sum_{i=1}^r b_i^{\otimes 6} \right\|_{\mathrm{F}}^2 = 0 \right\}.$$

Proof. By orthonormality, $\left\|\sum_{i=1}^r a_i^{\otimes 6}\right\|_{\mathrm{F}}^2 = \left\|\sum_{i=1}^r a_i^{\otimes 3}\right\|_{\mathrm{F}}^2 = r$ and $\mathfrak{A} \vdash \left\|\sum_{i=1}^r b_i^{\otimes 6}\right\|_{\mathrm{F}}^2 = \left\|\sum_{i=1}^r b_i^{\otimes 3}\right\|_{\mathrm{F}}^2 = r$. Thus, $\mathfrak{A} \vdash \sum_{i,j} \langle a_i, b_j\rangle^3 = r$ it suffices to show $\mathfrak{A} \vdash \sum_{i,j} \langle a_i, b_j\rangle^6 \geqslant r$.

Using $\sum_{i=1}^r a_i a_i^\top \preceq \mathrm{Id}$, a sum-of-squares version of Cauchy–Schwarz, and the fact that $\mathfrak{A}$ contains the constraints $\|b_1\|^2 = \cdots = \|b_r\|^2 = 1$,

$$\mathfrak{A} \vdash \quad r = \sum_{i,j} \langle a_i, b_j\rangle^3 \leqslant \tfrac{1}{2} \sum_{i,j} \langle a_i, b_j\rangle^2 + \tfrac{1}{2} \sum_{i,j} \langle a_i, b_j\rangle^4 \leqslant \tfrac{1}{2} r + \tfrac{1}{2} \sum_{i,j} \langle a_i, b_j\rangle^4.$$

We conclude that $\mathfrak{A} \vdash \sum_{i,j} \langle a_i, b_j\rangle^4 = r$. Applying the same reasoning to $\sum_{i,j} \langle a_i, b_j\rangle^4$ instead of $\sum_{i,j} \langle a_i, b_j\rangle^3$ yields $\mathfrak{A} \vdash \sum_{i,j} \langle a_i, b_j\rangle^6 = r$ as desired. □

Random components. Let $a_1, \dots, a_r \in \mathbb{R}^n$ be uniformly random unit vectors with $r \leqslant n^{O(1)}$. Let B be an n-by-r matrix of variables and let $b_1, \dots, b_r$ be the columns of B. Consider the following system of polynomial constraints

(2-1)

$$\mathfrak{B}_\varepsilon = \left\{ \|b_1\|^2 = \cdots = \|b_r\|^2 = 1,\ \left\|\sum_{i=1}^r b_i^{\otimes 3}\right\|_{\mathrm{F}}^2 \geqslant (1-\varepsilon)\cdot r,\ \left\|\sum_{i=1}^r b_i^{\otimes 6}\right\|_{\mathrm{F}}^2 \leqslant (1+\varepsilon)\cdot r \right\}.$$

With high probability, the vectors $a_1, \dots, a_r$ satisfy $\mathfrak{B}_\varepsilon$ for $\varepsilon \leqslant \tilde{O}(r/n^{1.5})$. Concretely, with high probability, every pair $(i, j) \in [r]^2$ with $i \neq j$ satisfies $\langle a_i, a_j\rangle^2 \leqslant \tilde{O}(1/n)$. Thus, $\left\|\sum_{i=1}^r b_i^{\otimes 3}\right\|_{\mathrm{F}}^2 = r + \sum_{i\neq j} \langle b_i, b_j\rangle^3 \geqslant (1 + \tilde{O}(r/n^{1.5}))\cdot r$ and $\left\|\sum_{i=1}^r b_i^{\otimes 6}\right\|_{\mathrm{F}}^2 \leqslant (1 + \tilde{O}(r/n^3))\cdot r$.

Lemma 2.8 (implicit in Ge and Ma [2015]). *Let $\varepsilon > 0$ and $a_1, \dots, a_r \in \mathbb{R}^n$ be random unit vectors with $r \leqslant \varepsilon \cdot \tilde{\Omega}(n^{1.5})$. Let B be an n-by-r matrix of variables, $b_1, \dots, b_r$ the columns of B, and $\mathfrak{A}$ the following system of polynomial constraints,*

$$\mathfrak{A} = \mathfrak{B}_\varepsilon \bigcup \left\{ \sum_{i=1}^r a_i^{\otimes 3} = \sum_{i=1}^r b_i^{\otimes 3} \right\} .$$

Then,

$$\mathfrak{A} \vdash_{\frac{B}{12}} \left\{ \left\| \sum_{i=1}^r a_i^{\otimes 6} - \sum_{i=1}^r b_i^{\otimes 6} \right\|_{\mathrm{F}}^2 \leqslant O(\varepsilon) \cdot \left\| \sum_{i=1}^r a_i^{\otimes 6} \right\|_{\mathrm{F}}^2 \right\} .$$

2.5 Clustering. We consider the following clustering problem: given a set of points $y_1, \dots, y_n \in \mathbb{R}^d$, the goal is to output a k-clustering matrix $X \in \{0,1\}^{n \times n}$ of the points such that the points in each cluster are close to each other as possible. Here, we say that a matrix $X \in \{0,1\}^{n \times n}$ is a k-clustering if there is a partition $S_1, \dots, S_k$ of $[n]$ such that $X_{ij} = 1$ if and only if there exists $\ell \in [k]$ with $i, j \in S_\ell$.

In this section, we will discuss how sum-of-squares allow us to efficiently find clusterings with provable guarantees that are significantly stronger than for previous approaches. For concreteness, we consider in the following theorem the extensively studied special case that the points are drawn from a mixture of spherical Gaussians such that the means are sufficiently separated Dasgupta [1999], Arora and Kannan [2001], Vempala and Wang [2004], Achlioptas and McSherry [2005], Kalai, Moitra, and Valiant [2010], Moitra and Valiant [2010], and Belkin and Sinha [2010]. Another key advantage of the approach we discuss is that it continues to work even if the points are not drawn from a mixture of Gaussians and the clusters only satisfy mild bounds on their empirical moment tensors.

Theorem 2.9 (Hopkins and Li [2018], P. K. Kothari, Steinhardt, and Steurer [2018], and Diakonikolas, Kane, and Stewart [2018]). *There exists an algorithm that given $k \in \mathbb{N}$ with $k \leqslant n$ and vectors $y_1, \dots, y_n \in \mathbb{R}^d$ outputs a k-clustering matrix $X \in \{0,1\}^{n \times n}$ in quasi-polynomial time $n + (dk)^{(\log k)^{O(1)}}$ with the following guarantees: Let $y_1, \dots, y_n$ be a sample from the uniform mixture of k spherical Gaussians $N(\mu_1, \mathrm{Id}), \dots, N(\mu_k, \mathrm{Id})$ with mean searation $\min_{i \neq j} \|\mu_i - \mu_j\| \geqslant O(\sqrt{\log k})$ and $n \geqslant (dk)^{(\log k)^{O(1)}}$. Let $X^* \in \{0,1\}^{n \times n}$ be the k-clustering matrix corresponding to the Gaussian components (so that $X^*_{ij} = 1$ if y_i and y_j were drawn from the same Gaussian component and $X^*_{ij} = 0$ otherwise). Then with high probability,*

$$\|X - X^*\|_{\mathrm{F}}^2 \leqslant 0.1 \cdot \|X^*\|_{\mathrm{F}}^2 .$$

We remark that the same techniques also give a sequence of polynomial-time algorithms that approach the logarithmic separation of the algorithm above. Concretely, for every $\varepsilon > 0$, there exists an algorithm that works if the mean separation is at least $O_\varepsilon(k^\varepsilon)$.

These algorithms for clustering points drawn from mixtures of separated spherical Gaussians constitute a significant improvement over previous algorithms that require separation at least $O(k^{1/4})$ Vempala and Wang [2004].

Sum-of-squares approach to learning mixtures of spherical Gaussians. In order to apply Theorem 2.1, we view the clustering matrix X corresponding to the Gaussian components as parameter and a "typical sample" $y_1, \dots, y_n$ of the mixture as measurement. Here, typical means that the empirical moments in each cluster are close to the moments of a spherical Gaussian distribution. Concretely, we consider the following set of parameter–measurement pairs,

$$\mathcal{P}_{k,\varepsilon,\ell} = \left\{ (X, Y) \;\middle|\; \begin{array}{l} X \text{ is } k\text{-clustering matrix with clusters } S_1, \dots, S_k \subseteq [n] \\ \forall \kappa \in [k].\ \left\| \mathbb{E}_{i \in S_\kappa} (1, x_i - \mu_\kappa)^{\otimes \ell} - \mathbb{E}_{x \sim N(0,\mathrm{Id})} (1, x)^{\otimes \ell} \right\|_{\mathrm{F}} \leqslant \varepsilon \end{array} \right\} \subseteq \{0,1\}^{k \times n} \times \mathbb{R}^{d \times n},$$

where $\mu_\kappa = \mathbb{E}_{i \in S_\kappa} x_i$ is the mean of cluster $S_\kappa \subseteq [n]$.

It is straightforward to express $\mathcal{P}_{k,\varepsilon,\ell}$ in terms of a system of polynomial constraints $\mathcal{A} = \{p(X, Y, z) = 0\}$, so that $\mathcal{P}_{k,\varepsilon,\ell} = \{(X, Y) \mid \exists z.\ p(X, Y, z) = 0\}$. Theorem 2.9 follows from Theorem 2.1 using the fact that under the conditions of Theorem 2.9, the following sum-of-squares proof exists with high probability for $\ell \leqslant (\log k)^{O(1)}$,

$$\mathcal{A} \vdash_{\ell}^{X,z} \left\{ \|X - X^*\|_{\mathrm{F}}^2 \leqslant 0.1 \cdot \|X^*\|_{\mathrm{F}}^2 \right\},$$

where X^* is the ground-truth clustering matrix (corresponding to the Gaussian components).

3 Lower bounds

In this section, we will be concerned with showing lower bounds on the minimum degree of sum-of-squares refutations for polynomial systems, especially those arising out of estimation problems.

The turn of the millennium saw several works that rule out degree-2 sum-of-squares refutations for a variety of problems, such as max cut Feige and Schechtman [2002], k-clique Feige and Krauthgamer [2000], and sparsest cut Khot and Vishnoi [2015], among others. These works, rather than explicitly taking place in the context of sum-of-squares proofs, were motivated by the desire to show tightness for specific SDP relaxations.

Around the same time, Grigoriev proved *linear* lower bounds on the degree of sum-of-squares refutations for k-XOR, k-SAT, and knapsack Grigoriev [2001b,a] (these bounds

were later independently rediscovered by Schoenebeck [2008]). Few other lower bounds against SoS were known. Most of the subsequent works (e.g. Tulsiani [2009] and Bhaskara, Charikar, Vijayaraghavan, Guruswami, and Zhou [2012]) built on the k-SAT lower bounds via reduction; in essence, techniques for proving lower bounds against higher-degree sum-of-squares refutations were ad hoc and few.

In recent years, a series of papers Meka, A. Potechin, and Wigderson [2015], Deshpande and Montanari [2015], and Hopkins, P. Kothari, A. H. Potechin, Raghavendra, and Schramm [2016] introduced higher-degree sum-of-squares lower bounds for k-clique, culminating in the work of Barak et al. Barak, Hopkins, Kelner, P. Kothari, Moitra, and A. Potechin [2016]. Barak et al. go beyond proving lower bounds for the k-clique problem specifically, introducing a beautiful and general framework for proving SoS lower bounds. Though their work settles the k-clique refutation problem in $\mathbb{G}(n, 1/2)$, it leaves more questions than answers. In particular, it gives rise to a compelling conjecture, which if proven, would settle the degree needed to refute a broad class of estimation problems, including densest k-subgraph, community detection problems, graph coloring, and more. We devote this section to describing the technique of pseudocalibration.

Let us begin by recalling some notation. Let $\mathcal{P} = \{p_i(x, y) \geq 0\}_{i \in [m]}$ be a polynomial system associated with an estimation problem. The polynomial system is over hidden variables $x \in \mathbb{R}^n$, with coefficients that are functions of the measurement/instance variables $y \in \mathbb{R}^N$. We will use $\mathcal{P}_y$ to denote the polynomial system for a fixed y. Let $\mathcal{P}$ have degree at most d_x in x and degree at most d_y in y. If $\mathfrak{D}_\varnothing$ denotes the null distribution, then $\mathcal{P}_y$ is infeasible w.h.p. when $y \sim \mathfrak{D}_\varnothing$, and we are interested in the minimum degree of sum-of-squares refutation.

Pseudodensities. By Claim 1.4, to rule out degree-d sum-of-squares refutations for $\mathcal{P}_y$, it is sufficient to construct the dual object namely the pseudoexpectation functional $\tilde{\mathbb{E}}_y$ with the properties outlined in Section 1.2. However, it turns out to be conceptually cleaner to think about constructing a related object called *pseudodensities* rather than *pseudoexpectation functionals*. Towards defining pseudodensities, we first pick a natural background measure σ for $x \in \mathbb{R}^n$. Therefore, $\mathbb{E}_x$ will denote the expectation over the background measure σ. The choice of background measure itself is not too important, but for the example we will consider, it will be convenient to pick σ to be uniform distribution over $\{0, 1\}^n$.

Definition 3.1. A function $\bar{\mu} : \{0, 1\}^n \to \mathbb{R}$ is a pseudodensity for a polynomial system $\mathcal{P} = \{p_i(x) \geq 0\}_{i \in [m]}$ if $\tilde{\mathbb{E}}_{\bar{\mu}} : \mathbb{R}[x]_{\leq d} \to \mathbb{R}$ defined as follows:

$$\tilde{\mathop{\mathbb{E}}_{\bar{\mu}}}[p(x)] \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{x} \bar{\mu}(x) p(x)$$

is a valid pseudoexpectation operator, namely, it satisfies the constraints outlined in Section 1.2.

To show that $\mathcal{P}_y$ is does not admit a degree d SoS refutation for most $y \sim \mathfrak{D}_\varnothing$, it suffices for us to show that with high probability over $y \sim \mathfrak{D}_\varnothing$, we can construct a pseudodensity $\bar{\mu}_y : \{0,1\}^n \to \mathbb{R}$. More precisely, with high probability over the choice of $y \sim \mathfrak{D}_\varnothing$, the following must hold:

(3-1)
(scaling) $$\mathop{\mathbb{E}}_x \bar{\mu}_y(x)dx = 1$$

(3-2)
(positive semidefiniteness) $$\mathop{\mathbb{E}}_x q(x)^2 \bar{\mu}_y(x)dx \geq 0 \qquad \forall q \in \mathbb{R}[x]_{\leq d}$$

(3-3)
(constraints from $\mathcal{P}$) $$\mathop{\mathbb{E}}_x p(x)a^2(x) \cdot \mu_y(x)dx \geq 0 \qquad \forall p \in \mathcal{P}, a \in \mathbb{R}[x], \deg(a^2 \cdot p) \leq$$

3.1 Pseudocalibration. *Pseudocalibration* is a heuristic for constructing pseudodensities for non-feasible systems in such settings. It was first introduced in Barak, Hopkins, Kelner, P. Kothari, Moitra, and A. Potechin [2016] for the k-clique problem, but the heuristic is quite general and can be seen to yield lower bounds for other problems as well (e.g. Grigoriev [2001b] and Schoenebeck [2008]).

At a high level, pseudocalibration leverages the existence of the structured/structured distribution of estimation problem, to construct pseudodistributions. Let $\mathfrak{J}_*$ denote the joint *structured* distribution over $y^* \in \{\pm 1\}^N$ and x^* is sampled from σ, i.e., $\mathbb{P}_{\mathfrak{J}_*}\{(x,y)\} = \sigma(x) \cdot \mathbb{P}_{\mathfrak{J}_x}\{y\}$.

Let us define a joint null distribution $\mathfrak{J}_\varnothing$ on pairs (x,y) to be

$$\mathfrak{J}_\varnothing \stackrel{\text{def}}{=} \sigma \times \mathfrak{D}_\varnothing \,.$$

As we describe pseudocalibration, $\mathfrak{J}_\varnothing$ will serve as the background measure for us. Let $\mu_* : \{\pm 1\}^N \times \{0,1\}^n \to \mathbb{R}^+$ denote the density of the joint structured distribution $\mathfrak{J}_*$ with respect to the background measure $\mathfrak{J}_\varnothing$, namely

$$\mu^*(x,y) = \frac{\mathbb{P}_{\mathfrak{J}_*}(x,y)}{\mathbb{P}_{\mathfrak{J}_\varnothing}(x,y)} = \frac{\mathbb{P}_{\mathfrak{D}_*}\{y\}}{\mathbb{P}_{\mathfrak{D}_\varnothing}\{y\}} \cdot \frac{\mathbb{P}_{\mathfrak{J}_*}\{x|y\}}{\sigma(x)}$$

At first glance, a candidate construction of pseudodensities $\bar{\mu}_y$ would be the partially-evaluated relative joint density μ_* namely

$$\bar{\mu}_y = \mu_*(y,\cdot)\,.$$

This construction already satisfies two of the three constraints namely Equation (3-2) and Equation (3-3). Note that for any polynomial $p(x, y)$,

$$\mathop{\mathbb{E}}_{x} p(x)\bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathfrak{D}_*}\{y\}}{\mathbb{P}_{\mathfrak{D}_\varnothing}\{y\}} \cdot \mathop{\mathbb{E}}_{x} p(x) \frac{\mathbb{P}_{\mathfrak{J}_*}\{x|y\}}{\sigma(x)} = \frac{\mathbb{P}_{\mathfrak{D}_*}\{y\}}{\mathbb{P}_{\mathfrak{D}_\varnothing}\{y\}} \cdot \mathop{\mathbb{E}}_{x\sim\mathfrak{J}_{|y}} p(x)\,.$$

From the above equality, Equation (3-2) follows directly because

$$\mathop{\mathbb{E}}_{x} q(x)^2\bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathfrak{D}_*}\{y\}}{\mathbb{P}_{\mathfrak{D}_\varnothing}\{y\}} \cdot \mathop{\mathbb{E}}_{x\sim\mathfrak{J}_{|y}} q^2(x) \geq 0\,.$$

Similarly, Equation (3-3) is again an immediate consequence of the fact that $\mathfrak{J}$ is supported on feasible pairs for $\mathcal{P}$,

$$\mathop{\mathbb{E}}_{x} p(x)a^2(x)\bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathfrak{D}_*}\{y\}}{\mathbb{P}_{\mathfrak{D}_\varnothing}\{y\}} \cdot \mathop{\mathbb{E}}_{x\sim\mathfrak{J}_{|y}} p(x)a^2(x) \geq 0\,.$$

However, the scaling constraint Equation (3-1) is far from satisfied because,

$$\mathop{\mathbb{E}}_{x} \bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathfrak{D}_*}\{y\}}{\mathbb{P}_{\mathfrak{D}_\varnothing}\{y\}} \mathop{\mathbb{E}}_{x\sim\mathfrak{J}_{|y}} 1 = \frac{\mathbb{P}_{\mathfrak{D}_*}\{y\}}{\mathbb{P}_{\mathfrak{D}_\varnothing}\{y\}}$$

is a quantity that is really large for $y \in \mathrm{supp}(\mathfrak{D}_*)$ and 0 otherwise. As a saving grace, the constraint Equation (3-1) is satsified in expectation over y, i.e.,

$$\mathop{\mathbb{E}}_{y\sim\mathfrak{D}_\varnothing} \mathop{\mathbb{E}}_{x} \bar{\mu}_y(x) = \mathop{\mathbb{E}}_{y\sim\mathfrak{D}_\varnothing} \mathop{\mathbb{E}}_{x} \mu^*(x, y) = \mathop{\mathbb{E}}_{(x,y)\sim\mathfrak{J}_\varnothing} \mu^*(x, y) = 1\,,$$

since μ^* is a density.

The relative joint density $\mu_*(y, x)$ faces an inherent limitation in that it is only nonzero on $\mathrm{supp}(\mathfrak{D}_*)$, which accounts for a negligible fraction of $y \sim \mathfrak{D}_\varnothing$. Intuitively, the constraints of $\mathcal{P}$ are low-degree polynomials in x and y. Therefore, our goal is to construct a $\bar{\mu}_y$ that has the same low-degree structure of μ_* but has a much higher entropy a.k.a., its mass is not too all concentrated on a small fraction of instances.

The most natural candidate to achieve this is to just project the joint density μ_* in to the space of low-degree polynomials. Formally, let $L_2(\mathfrak{J}_\varnothing)$ denote the vector space of functions over $\mathbb{R}^N \times \mathbb{R}^n$ equipped with the inner product $\langle f, g\rangle_{\mathfrak{D}_\varnothing} = \mathbb{E}_{y\sim\mathfrak{J}_\varnothing} f(x, y)g(x, y)$. For $d_x, D_y \in \mathbb{N}$, let $V_{d_x,D_y} \subseteq L_2(\mathfrak{J}_\varnothing)$ denote the following vector space

$$V_{d_x,D_y} = \mathrm{span}\{q(x, y) \in \mathbb{R}[x, y] \,|\, \deg_x(q) \leqslant d_x, \deg_y(q) \leqslant D_y\}$$

If Π_{d_x,D_y} denote the projection on to V_{d_x,D_y}, then the pseudo-calibration recipe suggests the use of the following pseudodistribution:

(3-4) (Pseudo-calibration) $\bar{\mu}_y(x) = \Pi_{d_x,D_y} \circ \mu_*(x, y)$

where d_x is the target degree for the pseudodistribution and $D_y \in \mathbb{N}$ is to be chosen sufficiently large given d_x.

Consider a constraint $\{p(x,y) \geq 0\} \in \mathcal{P}$ in the polynomial system. As long as $\deg_x(p) \leq d$ and $\deg_y(p) \leq D_y$, the pseudodensity $\bar{\mu}_y$ satisfies the constraint in expectation over y. This is immediate from the following calculation,

$$\begin{aligned}\mathbb{E}\,\bar{\mu}_y(x)p(x,y) &= \mathop{\mathbb{E}}_{(x,y)\sim\mathfrak{J}_\varnothing}(\Pi_{d_x,D_y}\circ\mu_*(x,y))p(x,y)\\ &= \mathop{\mathbb{E}}_{(x,y)\sim\mathfrak{J}_\varnothing}\mu_*(x,y)p(x,y)\\ &= \mathop{\mathbb{E}}_{(x,y)\sim\mathfrak{J}_*}p(x,y)\geqslant 0\,.\end{aligned}$$

We require that the constraints are satisfied for each $y \sim \mathfrak{D}_\varnothing$, rather than in expectation. Under mild conditions on the joint distribution $\mathfrak{J}_*$, the pseudocalibrated construction satisfies the constraints approximately with high probability over $y \in \mathfrak{D}_\varnothing$. Specifically, the following theorem holds.

Theorem 3.2. *Suppose $\{p(x,y) \geq 0\} \in \mathcal{P}$ is always satisfied for $(x,y) \sim \mathfrak{J}_*$ and let $B := \max_{(x,y)\in\mathfrak{J}_\varnothing} p(x,y)$ and let $d_y := \deg_y(p)$. If $\bar{\mu}_y$ is the pseudocalibrated pseudodensity as defined in Equation (3-4) then*

$$\mathop{\mathbb{P}}_{y\in\mathfrak{D}_\varnothing}[\mathop{\mathbb{E}}_x p(x,y)\bar{\mu}_y(x) \lessapprox -\varepsilon] \leq \frac{B^2}{\varepsilon^2}\|\Pi_{d,D_y+2d_y}\circ\mu_* - \Pi_{d,D_y}\circ\mu_*)\|^2_{2,\mathfrak{J}_\varnothing}$$

where $\Pi_{d,D}$ is the projection on to span of polynomials of degree at most D in y and degree d in x.

The theorem suggests that if the projection of the structured density μ_* decays with increasing degree then the pseudocalibrated density $\bar{\mu}_y$ satisfies the same constraints as those satisfied by μ_*, with high probability. This decay in the Fourier spectrum of the structured density is a common feature in all known applications of pseudocalibration. We defer the proof of the Theorem 3.2 to the full version.

Verifying non-negativity of squares. In light of Theorem 3.2, the chief obstacle in establishing $\bar{\mu}(y,\cdot)$ as a valid pseudodensity is in proving that it satisfies the constraint $\mathbb{E}_{\mathfrak{J}_\varnothing} p(y,x)^2\bar{\mu}(y,x)dx \geq 0$, for every polynomial p of degree at most $\frac{d}{2}$ in x. As we will see in Claim 3.3, this condition is equivalent to establishing the positive-semidefiniteness (PSDness) of the matrix

$$M_d(y) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_x\left(x^{\leqslant d/2}\right)\left(x^{\leqslant d/2}\right)^\top\cdot\bar{\mu}(y,x)dx,$$

where $x^{\leq d/2}$ is the $O(n^{d/2}) \times 1$ vector whose entries contain all monomials in x of degree at most $d/2$.

Claim 3.3. $\mathbb{E}_{\mathfrak{D}_\varnothing} q(y,x)^2 \bar{\mu}(y,x) \geq 0$ for all polynomials $q(y,x)$ of degree at most $d/2$ in x if and only if the matrix $M(y) \stackrel{\text{def}}{=} \tilde{\mathbb{E}}[(x^{\leq d/2})(x^{\leq d/2})^\top](y)$ is positive semidefinite.

Proof. The first direction is given by expressing $q(y,x)$ with its vector of coefficients of monomials of x, $\hat{q}(y)$, so that $\langle \hat{q}(y), x^{\leq d/2} \rangle = q(y,x)$. Then

$$\mathop{\mathbb{E}}_{\mathfrak{D}_\varnothing} q(y,x)^2 \bar{\mu}(y,x) = \mathop{\mathbb{E}}_{\mathfrak{D}_\varnothing} [\hat{q}(y)^\top (x^{\leq 2})(x^{\leq 2})^\top \hat{q}(y)] = \hat{q}(y)^\top M(y) \hat{q}(y) \geq 0,$$

by the positive-semidefiniteness of $M(y)$.

We now prove the contrapositive: if $M(y)$ is not positive-semidefinite, then there is some negative eigenvector $v(y)$ so that $v(y)^\top M(y) v(y) < 0$. Taking $q(y,x) = \langle v(y), x^{\leq d/2} \rangle$, we have our conclusion. □

Each entry of $M_d(y)$ is a degree-D polynomial in $y \sim \mathfrak{D}_\varnothing$. Since the entries of $M_d(y)$ are not independent, and because $M_d(y)$ cannot be decomposed easily into a sum of independent random matrices, standard black-box matrix concentration arguments such as matrix Chernoff bounds and Wigner-type laws do not go far towards characterizing the spectrum of $M_d(y)$. This ends up being a delicate and involved process, and the current proofs are very tailored to the specific choice of $\mathfrak{D}_\varnothing$, and in some cases they are quite technical.

3.1.1 Pseudocalibration: a partial answer, and many questions. While Theorem 3.2 establishes some desirable properties for μ, we are left with many unanswered questions. Ideally, we would be able to identify simple, general sufficient conditions on the structured distribution $\mathfrak{D}_*$ and on d the degree in x and D the degree in y, for which the answer to the above questions is affirmative. The following conjecture stipulates one such choice of conditions:

Conjecture 3.4. *Suppose that $\mathcal{P}$ contains no polynomial of degree more than k in y. Let $D = O(kd \log n)$ and $D = \Omega(kd)$. Then the D-pseudocalibrated function $\bar{\mu}(y,\cdot)$ is with high probability a valid degree-d pseudodistribution which satisfies $\mathcal{P}$ if and only if there is no polynomial $q(y)$ of degree D in y such that*

$$\mathop{\mathbb{E}}_{y \sim \mathfrak{D}_\varnothing} [q(y)] < n^{O(d)} \cdot \mathop{\mathbb{E}}_{y \sim \mathfrak{D}_*} [q(y)].$$

The upper and lower bounds on D stated in Conjecture 3.4 may not be precise; what is important is that D not be too much larger than $O(kd)$. In support of this conjecture, we list several refutation problems for which the conjecture has been proven: k-clique Barak,

Hopkins, Kelner, P. Kothari, Moitra, and A. Potechin [2016], tensor PCA Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer [2017], and random k-SAT and k-XOR Grigoriev [2001b] and Schoenebeck [2008]. However, in each of these cases, the proofs have been somewhat ad hoc, and do not generalize well to other problems of interest, such as densest-k-subgraph, community detection, and graph coloring.

Resolving this conjecture, which will likely involve discovering the "book" proof of the above results, is an open problem which we find especially compelling.

Variations. The incompleteness of our understanding of the pseudocalibration technique begs the question, is there a different choice of function $\mu'(y, x)$ such that $\mu'(y, \cdot)$ is a valid pseudodensity satisfying $\mathcal{P}$ with high probability over $y \sim \mathfrak{D}_\varnothing$? Indeed, already among the known constructions there is some variation in the implementation of the low-degree projection: the truncation threshold is not always a sharp degree D, and is sometimes done in a gradual fashion to ease the proofs (see e.g. Barak, Hopkins, Kelner, P. Kothari, Moitra, and A. Potechin [2016]). It is a necessary condition that μ' and μ_* agree at least on the moments of y which span the constraints of $\mathcal{P}$. However, there are alternative ways to ensure this, while also choosing μ' to have higher entropy than μ_*.

In Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer [2017], the authors give a different construction, in which rather than projecting μ_* to the span of low-degree polynomials in y, they choose the function μ' which minimizes energy under the constraint that $\int x^{\otimes d} \mu'(y, x) dx$ is positive semidefinite for every $y \in \operatorname{supp}(\mathfrak{D}_\varnothing)$, and that $\mathbb{E}_y \mu'(y, x) p(y, x) = \mathbb{E}_y \mu_*(y, x) p(y, x)$ for every $p(y, x)$ of degree at most D in y. Though in Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer [ibid.] this did not lead to unconditional lower bounds, it was used to obtain a characterization of sum-of-squares algorithms in terms of spectral algorithms.

3.2 Example: k-clique . In the remainder of this section, we will work out the pseudocalibration construction for the k-clique problem (see Example 1.1 for a definition). We'll follow the outline of the pseudocalibration recipe laid out in Equation (3-4), filling in the blanks as we go along.

The null and structured distributions. Recall that $\mathfrak{D}_\varnothing$ is the uniform distribution over the hypercube $\{\pm 1\}^{\binom{[n]}{2}}$, corresponding to $\mathbb{G}(n, 1/2)$. For $\mathfrak{D}_*$ we use the joint distribution over tuples of instance and solution variables (y^*, x^*) described in Example 1.1, with a small twist designed to ease calculations: Rather than sampling x^* from π the uniform distribution over the indicators $\mathbf{1}_S \in \{0, 1\}^n$ for $|S| = k$, we sample x^* by choosing every coordinate to be 1 with probability $\frac{2k}{n}$, and 0 otherwise.

Pseudomoments. Instead of directly constructing the pseudodensity $\bar{\mu}$, it will be more convenient for us to work with the *pseudomoments*. So for each monomial x^A where the multiset $A \subset [n]$ has cardinality at most d, we will directly define the function $\tilde{\mathbb{E}}_{\bar{\mu}(y)}[x^A] : \{\pm 1\}^{\binom{[n]}{2}} \to \mathbb{R}$. For convenience, and to emphasize the dependence on y, we will equivalently write $\tilde{\mathbb{E}}[x^A](y)$.

Let $\mathcal{A} \subseteq E^{\leq D}$ be a set of subsets of edges of cardinality at most D (we will specify D later). Following the pseudocalibration recipe from Equation (3-4), for each $\alpha \in \mathcal{A}$ we will set

$$\mathop{\mathbb{E}}_{y\sim\mathfrak{D}_\varnothing}\left[y^\alpha \cdot \tilde{\mathbb{E}}[x^A](y)\right] = \sum_{y\in\{\pm1\}^E} \int x^A \cdot \mu_*(y,x)dx = \mathop{\mathbb{E}}_{(y,x)\sim\mathfrak{D}_*}[y^\alpha x^A].$$

The right-hand side can be simplified further. For $(y,x) \sim \mathfrak{D}_*$, if any vertices of A are not chosen to be in the clique, then x^A is zero. Similarly, if any edge $e \in \alpha$ has an endpoint not in the clique, then $y^{\{e\}}$ is independent of $y^{A\setminus\{e\}}$ and of expectation 0. Thus, the expression is equal to the probability that all vertices of α and A, which we denote $v(\alpha) \cup A$, are contained in the clique:

$$\mathop{\mathbb{E}}_{(y,x)\sim\mathfrak{D}_*}[x^A y^\alpha] = \mathop{\mathbb{P}}_{x\sim\mathfrak{D}_*}[x_i = 1,\ \forall i \in v(\alpha)\cup A] = \left(\frac{2k}{n}\right)^{|v(\alpha)\cup A|}.$$

For convenience, we will let $\lambda \stackrel{\text{def}}{=} (\frac{2k}{n})$. Now expressing $\tilde{\mathbb{E}}[x^A](y)$ via its Fourier decomposition, we have

$$\tilde{\mathbb{E}}(y)[x^A] = \sum_{\alpha\in\mathcal{A}} \left(\frac{2k}{n}\right)^{|v(\alpha)\cup A|} \cdot y^\alpha. \tag{3-5}$$

4 Connection to Spectral Algorithms

Sum-of-squares SDPs yield a systematic framework that capture and generalize a loosely defined class of algorithms often referred to as *spectral algorithms*. The term "spectral algorithm" refers to an algorithm that on an input x associates a matrix $M(x)$ that can be easily computed from x and whose eigenvalues and eigenvectors manifestly yield a solution to the problem at hand. We will give a more concrete definition for the notion of a spectral algorithms a little later in this section.

Although spectral algorithms are typically subsumed by the sum-of-squares SDPs, the spectral versions tend to be simpler to implement and more efficient. Furthermore, in many cases such as the k-clique Alon, Krivelevich, and Sudakov [1998] and tensor decompositionHarshman [1970], the first algorithms discovered for the problem were spectral. From

a theoretical standpoint, spectral algorithms are much simpler to study and could serve as stepping stones to understanding the limits of sum-of-squares SDPs.

In the worst case, sum-of-squares SDPs often yield strictly better guarantees than corresponding spectral algorithms. For instance, the Goemans-Williamson SDP yields a 0.878 approximation for max cut Goemans and Williamson [1995], has no known analogues among spectral algorithms. Contrary to this, in many random settings, the best known sum-of-squares SDP algorithms yield guarantees that are no better than the corresponding spectral algorithms. Recent work explains this phenomena by showing an equivalence between spectral algorithms and their sum-of-squares SDP counterparts for a broad family of problems Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer [2017]. To formally state this equivalence, we will need a few definitions. Let us begin by considering a classic example of a spectral algorithm for the k-clique problem. In a graph $G = (V, E)$, if a subset $S \subset V$ of k vertices forms a clique then,

$$\left\langle 1_S, \left(A_G - \frac{J}{2}\right) 1_S \right\rangle = \frac{k(k-2)}{2}.$$

where $J \in \mathbb{R}^{n \times n}$ denotes the $n \times n$ matrix consisting of all ones. On the other hand, we can upper bound the right hand side by

$$\left\langle 1_S, \left(A_G - \frac{J}{2}\right) 1_S \right\rangle \leqslant \|1_S\|_2^2 \|A_G - \frac{J}{2}\|_{\text{op}} = k \cdot \lambda_{\max}\left(A_G - \frac{J}{2}\right).$$

thereby certifying an upper bound on the size of the clique k, namely,

$$k \leqslant 2\lambda_{\max}\left(A_G - \frac{J}{2}\right) + 2.$$

In particular, for a graph G drawn from the null distribution namely, Erdos-Renyi distribution $G(n, \frac{1}{2})$, the matrix $A_G - \frac{J}{2}$ is a random matrix whose entries are i.i.d uniformly over $\{\pm\frac{1}{2}\}$. By Matrix Chernoff inequality Tropp [2015] , we will have that $\lambda_{\max}(A_G - J/2) = O(\sqrt{n})$ with high probability. Thus one can certify an upper bound of $O(\sqrt{n})$ on the size of the clique in a random graph drawn from $G(n, \frac{1}{2})$ by computing the largest eigenvalue of the associated matrix valued function $P(G) = A_G - \frac{1}{2}J$.

Injective tensor norm. Recall that the injective tensor norm (see Example 1.2) of a symmetric 4-tensor $T \in \mathbb{R}^{[n]\times[n]\times[n]\times[n]}$ is given by $\max_{\|x\|\leqslant 1}\langle x^{\otimes 4}, T\rangle$. The injective tensor norm $\|T\|_{\text{inj}}$ is computationally intractable in the worst case Hillar and Lim [2013]. We will now describe a sequence of spectral algorithms that certify tighter bounds for the injective tensor norm of a tensor T drawn from the null distribution, namely a tensor T whose entries are i.i.d Gaussian random variables from $N(0, 1)$.

Let $T_{2,2}$ denotes the $n^2 \times n^2$ matrix obtained by flattening the tensor T then,

$$\|T\|_{\text{inj}} = \text{argmax}_{\|x\|_2 \leq 1} \langle T, x^{\otimes 4} \rangle = \text{argmax}_{\|x\|_2 \leq 1} \langle x^{\otimes 2}, T_{2,2} x^{\otimes 2} \rangle \leqslant \lambda_{\max}(T_{2,2})$$

Thus $\lambda_{\max}(T_{2,2})$ is a spectral upper bound on $\|T\|_{\text{inj}}$. Since each entry of T is drawn independently from $N(0,1)$, we have $\lambda_{\max}(T_{2,2}) \leqslant O(n)$ with high probability Tropp [2015]. Note that the injective norm of a random $N(0,1)$ tensor T is at most $O(\sqrt{n})$ with high probability Tomioka and Suzuki [2014] and Montanari and Richard [2014] . In other words, $\lambda_{\max}(T_{2,2})$ certifies an upper bound that is $n^{1/2}$-factor approximation to $\|T\|_{\text{inj}}$. We will now describe a sequence of improved approximations to the injective tensor norm via spectral methods. Fix a positive integer $k \in \mathbb{N}$. The polynomial $T(x) = \langle x^{\otimes 4}, T \rangle$ can be written as,

$$T(x) = \langle x^{\otimes 2}, T_{2,2} x^{\otimes 2} \rangle = \langle x^{\otimes 2k}, T_{2,2}^{\otimes k} x^{\otimes 2k} \rangle^{1/k}.$$

The tensor $x^{\otimes 2k}$ is symmetric, and is invariant under permutations of its modes. Let Σ_{2k} denote the set of all permutations of $\{1, \ldots, 2k\}$. For a permutation $\Pi \in \Sigma_{2k}$ and a $2k$-tensor $A \in \mathbb{R}^{[n]^{2k}}$, let $\Pi \circ A$ denote the $2k$-tensor obtained by applying the permutation Π to the modes of A. By averaging over all permutations $\Pi, \Pi' \in \Sigma_{2k}$, we can write

$$\begin{aligned} T(x) &= \left(\mathop{\mathbb{E}}_{\Pi, \Pi' \in \Sigma_{2k}} \langle \Pi \circ x^{\otimes 2k}, T_{2,2}^{\otimes k} (\Pi' \circ x^{\otimes 2k}) \rangle \right)^{1/2k} \\ &= \left(\langle x^{\otimes 2k}, \left(\mathop{\mathbb{E}}_{\Pi, \Pi' \in \Sigma_{2k}} \Pi \circ T_{2,2}^{\otimes k} \circ \Pi' \right) x^{\otimes 2k} \rangle \right)^{1/2k} \\ &\leqslant \lambda_{\max} \left(\mathop{\mathbb{E}}_{\Pi, \Pi' \in \Sigma_{2k}} \Pi \circ T_{2,2}^{\otimes k} \circ \Pi' \right)^{1/2k} \cdot \|x\|_2^2 . \end{aligned} \tag{4-1}$$

Therefore for every $k \in \mathbb{N}$, if we denote

$$P_k(T) \stackrel{\text{def}}{=} \mathop{\mathbb{E}}_{\Pi, \Pi' \in \Sigma_{2k}} \Pi \circ T_{2,2}^{\otimes k} \circ \Pi'$$

then $\|T\|_{\text{inj}} \leqslant \lambda_{\max}(P_k(T))^{1/k}$.

The entries of $P_k(T)$ are degree k polynomials in the entries of T. For example, a generic entry of $P_2(T)$ looks like,

$$P_2(T)_{ijk\ell, i'j'k'\ell'} = \frac{1}{(4!)^2} \cdot \big(T_{iji'j'} \cdot T_{k\ell k'\ell'} + T_{iji'k'} \cdot T_{k\ell j'\ell'} + T_{iji'\ell'} \cdot T_{k\ell j'k'} + \cdots 4!^2 \text{ terms} \cdots$$

Thus a typical entry of $P_k(T)$ with no repeated indices is an average of a super-exponentially large number, say N_k, of i.i.d. random variables. This implies that the variance of a typical entry of $P_k(T)$ is equal to $\frac{1}{N_k}$. For the moment, let us assume that

the spectrum of $P_k(T)$ has a distribution that is similar to that of a random matrix with i.i.d. Gaussian entries with variance $\frac{1}{N_k}$. Then, $\lambda_{\max}(P_k(T)) \leqslant O(n^k \cdot \frac{1}{N_k^{1/2}})$ with high probability, certifying that $\|T\|_{\text{inj}} \leqslant \frac{n}{N_k^{1/2k}}$. On accounting for the symmetries of T, it is easy to see that $N_k = k!\left(\frac{1}{2^k}\frac{2k!}{k!}\right)^2 \gg (k!)^2$. Consequently, as per this heuristic argument, $\lambda_{\max}(P_k(T))$ would certify an upper bound of $\|T\|_{\text{inj}} \leqslant O(\frac{n}{k^{3/4}})$.

Unfortunately, the entries of $P_k(T)$ are not independent random variables and not all entries of $P_k(T)$ are typical as described above. Although the heuristic bound on $\lambda_{\max}(P_k(T))$ is not quite accurate, a careful analysis via the trace method shows that the upper bound $\lambda_{\max}(P_k(T))^{1/k}$ decreases polynomially in k Bhattiprolu, Guruswami, and Lee [2017] and Raghavendra, Rao, and Schramm [2017].

Theorem 4.1. *Bhattiprolu, Guruswami, and Lee [2017] For $4 \leqslant k \leqslant n^{2/3}$ if T is a symmetric 4-tensor with i.i.d. entries from a subgaussian measure then*

$$\lambda_{\max}(P_k(T))^{1/k} \leqslant \tilde{O}\left(\frac{n}{k^{1/2}}\right)$$

then with probability $1 - o(1)$. Here $\tilde{O}$ notation hides factors polylogarithmic in n.

Thus the matrix polynomial $P_k(T)$ yields a $n^{O(k)}$-time algorithm to certify an upper bound of $\tilde{O}(n/k^{1/2})$ on the injective tensor norm of random 4-tensors with Gaussian entries.

Note that the upper bound certificate produced by the above spectral algorithm can be cast as a degree $4k$ sum-of-squares proof. In particular, if $\lambda_{\max}(P_k(T)) \leqslant B$ for some tensor T and $B \in \mathbb{R}$ then,

$$\begin{aligned}
B - T(x)^k &= B\|x\|_2^{4k} - \langle x^{\otimes 2k}, P_k(T)x^{\otimes 2k}\rangle + B(1 - \|x\|_2^{4k}) \\
&= \langle x^{\otimes 2k}, (B \cdot \mathrm{Id} - P_k(T))x^{\otimes 2k}\rangle + B(1 - \|x\|_2^{4k}) \\
&= \langle x^{\otimes 2k}, (B \cdot \mathrm{Id} - P_k(T))x^{\otimes 2k}\rangle + (1 - \|x\|_2^2)\left(B \cdot \sum_{i=0}^{2k-1} \|x\|_2^{2i}\right) \\
&= \sum_j s_j^2(x) + (1 - \|x\|_2^2)\left(B \cdot \sum_{i=0}^{2k-1} \|x\|_2^{2i}\right)
\end{aligned}$$

The final step in the calculation uses the fact that if a matrix $M \succeq 0$ is positive semidefinite, then the polynomial $\langle x^{\otimes 2k}, Mx^{\otimes 2k}\rangle$ is a sum-of-squares. Therefore, the degree $4k$ sum-of-squares SDP to obtain the same approximation guarantee at least as good as the somewhat adhoc spectral algorithm described above. This is a recurrent theme where the sum-of-squares SDP yields a unified and systematic algorithm that subsumes a vast majority of more adhoc approaches to algorithm design.

Refuting Random CSPs. The basic scheme used to upper bound the injective tensor norm (see Equation (4-1)) can be harnessed towards refuting random constraint satisfaction problems (CSPs). Fix a positive integer $k \in \mathbb{N}$. In general, a random k-CSP instance consists of a set of variables V over a finite domain, and a set of randomly sampled constraints each of which is on a subset of at most k variables. The problem of refuting random CSPs has been extensively studied for its numerous connections and applications Feige [2002], Ben-Sasson and Bilu [2002], Daniely, Linial, and Shalev-Shwartz [2014], Barak, Kindler, and Steurer [2013], and Crisanti, Leuzzi, and Parisi [2002]. For the sake of concreteness, let us consider the example of random 4-xor.

Example 4.2 (4-xor)**.** In the 4-xor problem, the input consists of a linear system over $\mathbb{F}_2$-valued variables $\{X_1, \ldots, X_n\}$ such that each equation has precisely 4 variables in it. A random 4-xor instance is one where each equation is sampled uniformly at random (avoiding repetition). Let m denote the number of equations, and n the number of variables. For $m \gg n$, with high probability over the choice of the constraints, every assignment satisfies at most $\frac{1}{2} + o(1)$ fraction of constraints. The goal of refutation algorithm is to certify that there no assignment that satisfies $\frac{1}{2} + o(1)$ fraction of constraints. To formulate a polynomial system, we will use the natural ± 1-encoding of $\mathbb{F}_2$, i.e., $x_i = 1 \iff X_i = 0$ and $x_i = -1 \iff X_i = 1$. An equation of the form $X_i + X_j + X_k + X_\ell = 0/1$ translates in to $x_i x_j x_k x_\ell = \pm 1$. We can specify the instance using a symmetric 4-tensor $\{T_{ijk\ell}\}_{i,j,k,\ell \in \binom{[n]}{4}}$, with $T_{ijk\ell} = \pm 1$ if we have the equation $x_i x_j x_k x_\ell = \pm 1$, and $T_{ijk} = 0$ otherwise. To certify that no assignment satisfies more than εm constraints, we will need to refute the following polynomial system.

$$\{x_i^2 - 1\}_{i \in [n]} \qquad \text{and} \qquad \{\langle T, x^{\otimes 4} \rangle \geq \varepsilon \cdot m\} \tag{4-2}$$

This system is analogous to the injective tensor norm, except the maximization is over the boolean hypercube $x \in \{\pm 1\}^n$, as opposed to the unit ball. Unlike the case of random Gaussian tensors, the tensor T of interest in 4-xor is a sparse tensor with about $n^{1+o(1)}$ non-zero entries. While this poses a few technical challenges, the basic schema from Equation (4-1) can still be utilized to obtain the following refutation algorithm.

Theorem 4.3. *Raghavendra, Rao, and Schramm [2017] For all $\delta \in [0, 1)$, the degree n^δ sum-of-squares SDP can refute random* 4-xor *instances with $m > \tilde{\Omega}(n^{2-\delta})$ with high probability.*

The refutation algorithm for XOR can be used as a building block to obtain sum-of-squares refutations for all random k-CSPs Raghavendra, Rao, and Schramm [ibid.]. Moreover, these bounds on the degree of sum-of-squares refutations tightly match corresponding lower bounds for CSPs shown in P. K. Kothari, Mori, O'Donnell, and Witmer [2017] and Barak, Chan, and P. K. Kothari [2015].

Defining spectral algorithms. The above-described algorithms will serve as blue-prints for the class of spectral algorithms that we will formally define now. The problem setup that is most appropriate for our purposes is that of distinguishing problem. Recall that in a distinguishing problem, the input consists of a x sample drawn from one of two distributions say $\mathfrak{D}_*$ or $\mathfrak{D}_\varnothing$ and the algorithm's goal is to identify the distribution the sample is drawn from. Furthermore, one of the distributions $\mathfrak{D}_*$ is referred to as the structured distribution is guaranteed to have an underlying hidden structure that is planted within, while samples from the null distribution $\mathfrak{D}_\varnothing$ typically do not.

A *spectral* algorithm $\mathfrak{A}$ to distinguish between samples from a structured distribution $\mathfrak{D}_*$ and a null distribution $\mathfrak{D}_\varnothing$ proceeds as follows. Given an instance x, the algorithm $\mathfrak{A}$ computes a matrix $P(x)$ whose entries are given by low-degree polynomials in x, such that $\lambda_{\max}(P(x)) > 0$ indicates whether $x \sim \mathfrak{D}_*$ or $x \sim \mathfrak{D}_\varnothing$.

Definition 4.4. (Spectral Algorithm) A spectral algorithm $\mathfrak{A}$ consists of a matrix valued polynomial $P : \mathcal{P} \to \mathbb{R}^{N \times N}$. The algorithm $\mathfrak{A}$ is said to distinguish between samples from structured distribution $\mathfrak{D}_*$ and a null distribution $\mathfrak{D}_\varnothing$ if,

$$\mathop{\mathbb{E}}_{y \sim \mathfrak{D}_*} \lambda^+_{\max}(P(y)) \gg \mathop{\mathbb{E}}_{y \sim \mathfrak{D}_\varnothing} \lambda^+_{\max}(P(y))$$

where $\lambda^+_{\max}(M) \stackrel{\text{def}}{=} \max(\lambda_{\max}(M), 0)$ for a matrix M.

In general, a spectral algorithm could conceivably use the entire spectrum of the matrix $P(y)$ instead of the largest eigenvalue, and perform some additional computations on the spectrum. However, a broad range of spectral algorithms can be cast into this framework and as we will describe in this section, this restricted class of spectral algorithms already subsumes the sum-of-squares SDPs in a wide variety of settings.

Spectral algorithms as defined in Definition 4.4 are a simple and highly structured class of algorithms, in contrast to algorithms for solving a sum-of-squares SDP. The feasible region for a sum-of-squares SDP is the intersection of the positive semidefinite cone with polynomially many constraints, some of which are equality constraints. Finding a feasible solution to the SDP involves an iterated sequence of eigenvalue computations. Furthermore, the feasible solution returned by the SDP solver is by no-means guaranteed to be a low-degree function of the input instance. Instead a spectral algorithm involves exactly one eigenvalue computation of a matrix whose entries are low-degree polynomials in the instance. In spite of their apparent simplicity, we will now argue that they are no weaker than sum-of-squares SDPs for a wide variety of estimation problems.

Robust Inference. Many estimation problems share a useful property that we will refer to as "robust inference" property. Specifically, the structured distributions underlying

these estimation problems are such that, a randomly chosen subsampling of the instance is sufficient to recover a non-negligible fraction of the planted structure. For example, consider the structured distribution $\mathfrak{D}_*$ for the k-clique problem. A graph $G \sim \mathfrak{D}_*$ consists of a k-clique embedded in to a Erdos-Renyi random graph. Suppose we subsample an induced subgraph G' of G, by randomly sampling a subset $S \subset V$ of vertices of size $|S| = \delta|V|$. With high probability, G' contains $\Omega(\delta \cdot k)$ of the planted clique in G. Therefore, the maximum clique in G' yields a clique of size $\Omega(\delta \cdot k)$ in the original graph G. This is an example of *robust inference* property, where a random subsample G' can reveal non-trivial structure in the instance. While the subsample does not determine the planted clique in G, the information revealed is substantial. For example, as long as $\delta \cdot k \gg 2 \log n$, G' is sufficient to distinguish whether G is sampled from the structured distribution $\mathfrak{D}_*$ or the null distribution $\mathfrak{D}_\varnothing$. Moreover, the maximum clique in G' can be thought of as a feasible solution to a relaxed polynomial system where the clique size sought after is $\delta \cdot k$, instead of k.

Let $\mathcal{P}$ denote a polynomial system defined on instance variables $y \in \mathbb{R}^N$ and in solution variables $x \in \mathbb{R}^n$ Let Υ denote the *subsampling distribution* namely, a probability distribution over subsets of instance variables $[N]$. Given an instance $y \in \mathbb{R}^N$, a subsample z can be sampled by first picking $S \sim \Upsilon$ and setting $z = y_S$. Let $\mathbb{I}$ denote the collection of all instances, and $\mathbb{I}_\downarrow$ denote the collection of all sub-instances.

Definition 4.5. A polynomial system $\mathcal{P}$ is ε-robustly inferable with respect to a subsampling distribution Υ and a structured distribution $\mathfrak{D}_*$, if there exists a map $\zeta : \mathbb{I}_\downarrow \to \mathbb{R}^n$ such that,

$$\mathop{\mathbb{P}}_{\substack{y \sim \mathfrak{D}_* \\ S \sim \Upsilon}} [\zeta(y_S) \text{ is feasible for } \mathcal{P}] \geq 1 - \varepsilon$$

Robust inference property arises in a broad range of estimation problems including stochastic block models, densest k-subgraph, tensor PCA, sparse PCA and random CSPs (see Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer [2017] for a detailed discussion). The existence of robust inference property has a stark implication on the power of low-degree sum-of-squares SDPs, namely they are no more powerful than spectral algorithms. This assertion is formalized in the following theorem.

Theorem 4.6. *Suppose $\mathcal{P} = \{p_i(x, y) \geq 0\}_{i \in [m]}$ is a polynomial system with degree d_x and d_y over x and y respectively. Fix $B \geq d_x \cdot d_y \in \mathbb{N}$. If the degree d sum-of-squares SDP relaxation can be used to distinguish between the structured distribution $\mathfrak{D}_*$ and the null distribution $\mathfrak{D}_\varnothing$, namely,*

- *For $y \sim \mathfrak{D}_*$, the polynomial system $\mathcal{P}$ is not only satisfiable, but is $1/n^{8B}$-robustly inferable with respect to a sub-sampling distribution Υ.*

- *For $y \sim \mathfrak{D}_\emptyset$, the polynomial system $\mathcal{P}$ is not only infeasible but admits a degree d sum-of-squares refutation with numbers bounded by n^B with probability at least $1 - 1/n^{8B}$.*

Then, there exists a degree $2D$ matrix polynomial $Q : \mathbb{R} \to \mathbb{R}^{[n]^{\leq d} \times [n]^{\leq d}}$ such that,

$$\frac{\mathbb{E}_{y \sim \mathfrak{D}_*}[\lambda^+_{\max}(Q(y))]}{\mathbb{E}_{y \sim \mathfrak{D}_\emptyset}[\lambda^+_{\max}(Q(y))]} \geqslant n^{B/2}$$

where $D \in \mathbb{N}$ be smallest integer such that for every subset $\alpha \subset [N]$ with $|\alpha| \geqslant D - 2d_x d_y$, $\mathbb{P}_{S \sim \Upsilon}[\alpha \subseteq S] \leqslant \frac{1}{n^{8B}}$.

The degree D of the spectral distinguisher depends on the sub-sampling distribution. Intuitively, the more robustly inferable (a.k.a inferable from smaller subsamples) the problem is, the smaller the degree of the distinguisher D. For the k-clique problem with a clique size of $n^{1/2-\varepsilon}$, we have $D = O(d/\varepsilon)$. For random CSPs, community detection and densest subgraph we have $D = O(d \log n)$ (see Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer [2017] for details).

From a practical standpoint, the above theorem shows that sum-of-squares SDPs can often be replaced by their more efficient spectral counterparts. From a theoretical standpoint, it reduces the task of showing lower bounds against the complicated algorithm namely the sum-of-squares SDP to that of understanding the spectrum of low-degree matrix polynomials over the two distributions.

Future work. The connection in Theorem 4.6 could potentially be tightened, leading to a fine-grained understanding of the power of sum-of-squares SDPs. We will use a concrete example to expound on the questions laid open by Theorem 4.6, but the discussion is applicable more broadly too.

Consider the problem of certifying an upper bound on the size of maximum independent sets in sparse random graphs. Formally, let G be a sparse random graph drawn from $\mathbb{G}(n, k/n)$ by sampling each edge independently with probability k/n. There exists a constant $\alpha_k \in (0,1)$ such that the size of the largest independent set in G is $(\alpha \pm o(1)) \cdot n$ with high probability. For every $\beta \in (0,1)$, the existence of a size $\beta \cdot n$-independent set can be formulated as the following polynomial system.

$$\mathcal{P}_\beta(G) : \left\{ \{x_i^2 - x_i = 0\}_{i \in [n]}, \qquad \{x_i x_j = 0\}_{(i,j) \in E(G)}, \qquad \sum_{i \in [n]} x_i \geqslant \beta \cdot n \,. \right\}$$

For each degree $d \in \mathbb{N}$ define

$$\alpha_d^{(k)} \stackrel{\text{def}}{=} \text{ smallest } \beta \text{ such that } \lim_{n \to \infty} \mathbb{P}_{G \sim \mathbb{G}([n], k/n)}[\mathcal{P}_\beta \vdash_d \perp] = 1$$

It is natural to ask if the approximation obtained by the degree d sum-of-squares SDP steadily improves with k.

Question 4.7. Is $\{\alpha_d^{(k)}\}_{d\in\mathbb{N}}$ a strictly decreasing sequence?

We can associate the following structured distribution $\mathfrak{Y}_\beta$ with the problem. For each subset $S \in \binom{[n]}{\beta\cdot n}$, define μ_S as $\mathbb{G}(n, k/n)$ conditioned on S being an independent set. For $D \in \mathbb{N}$ define, Let $\gamma_D^{(k)} \in (0,1)$ be the largest value of β for which distribution of eigenvalues of low-degree matrix polynomials in the structured distribution $\mathfrak{Y}_\beta$ and null distribution $\mathfrak{D}_\varnothing$ converge to each other in distribution. In other words, $\gamma_D^{(k)}$ is the precise threshold of independent set size β below which the structured and the null distributions have same empirical distribution of eigenvalues. It is natural to conjecture that if the empirical distribution of eigenvalues look alike then the sum-of-squares SDP cannot distinguish between the two. Roughly speaking, the conjecture formalizes the notion that sum-of-squares SDPs are no more powerful than spectral algorithms.

Question 4.8. Is $\alpha_d^{(k)} \geq \gamma_{O(d)}^{(k)}$?

References

Dimitris Achlioptas and Frank McSherry (2005). "On Spectral Learning of Mixtures of Distributions". In: *COLT*. Vol. 3559. Lecture Notes in Computer Science. Springer, pp. 458–469 (cit. on p. 3389).

Noga Alon, Michael Krivelevich, and Benny Sudakov (1998). "Finding a large hidden clique in a random graph". In: *Proceedings of the Eighth International Conference "Random Structures and Algorithms" (Poznan, 1997)*. Vol. 13. 3-4, pp. 457–466. MR: 1662795 (cit. on p. 3397).

Anima Anandkumar, Dean P. Foster, Daniel J. Hsu, Sham Kakade, and Yi-Kai Liu (2012). "A Spectral Algorithm for Latent Dirichlet Allocation". In: *NIPS*, pp. 926–934 (cit. on p. 3386).

Sanjeev Arora, Rong Ge, Tengyu Ma, and Andrej Risteski (2016). "Provable learning of Noisy-or Networks". *CoRR* abs/1612.08795 (cit. on p. 3386).

Sanjeev Arora and Ravi Kannan (2001). "Learning mixtures of arbitrary gaussians". In: *STOC*. ACM, pp. 247–257 (cit. on p. 3389).

Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou (2012). "Hypercontractivity, sum-of-squares proofs, and their applications". In: *STOC*. ACM, pp. 307–326 (cit. on p. 3377).

Boaz Barak, Siu On Chan, and Pravesh K. Kothari (2015). "Sum of squares lower bounds from pairwise independence [extended abstract]". In: *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing*. ACM, New York, pp. 97–106. MR: 3388187 (cit. on p. 3401).

Boaz Barak, Samuel B. Hopkins, Jonathan A. Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin (2016). "A Nearly Tight Sum-of-Squares Lower Bound for the Planted Clique Problem". In: *FOCS*. IEEE Computer Society, pp. 428–437 (cit. on pp. 3379, 3391, 3392, 3395, 3396).

Boaz Barak, Jonathan A. Kelner, and David Steurer (2014). "Rounding sum-of-squares relaxations". In: *STOC*. ACM, pp. 31–40 (cit. on p. 3382).

– (2015). "Dictionary Learning and Tensor Decomposition via the Sum-of-Squares Method". In: *STOC*. ACM, pp. 143–151 (cit. on pp. 3382, 3386, 3387).

Boaz Barak, Guy Kindler, and David Steurer (2013). "On the optimality of semidefinite relaxations for average-case and generalized constraint satisfaction". In: *ITCS*. ACM, pp. 197–214 (cit. on p. 3401).

Boaz Barak and Ankur Moitra (2016). "Noisy Tensor Completion via the Sum-of-Squares Hierarchy". In: *COLT*. Vol. 49. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 417–445 (cit. on pp. 3382, 3383, 3385).

Boaz Barak and David Steurer (2014). "Sum-of-squares proofs and the quest toward optimal algorithms". *Electronic Colloquium on Computational Complexity (ECCC)* 21, p. 59 (cit. on p. 3379).

Mikhail Belkin and Kaushik Sinha (2010). "Toward Learning Gaussian Mixtures with Arbitrary Separation". In: *COLT*. Omnipress, pp. 407–419 (cit. on p. 3389).

Eli Ben-Sasson and Yonatan Bilu (2002). "A Gap in Average Proof Complexity". *Electronic Colloquium on Computational Complexity (ECCC)* 003 (cit. on p. 3401).

Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan (2014). "Smoothed analysis of tensor decompositions". In: *STOC*. ACM, pp. 594–603 (cit. on p. 3386).

Aditya Bhaskara, Moses Charikar, Aravindan Vijayaraghavan, Venkatesan Guruswami, and Yuan Zhou (2012). "Polynomial integrality gaps for strong SDP relaxations of Densest k-subgraph". In: *SODA*. SIAM, pp. 388–405 (cit. on p. 3391).

Vijay Bhattiprolu, Venkatesan Guruswami, and Euiwoong Lee (2017). "Sum-of-Squares Certificates for Maxima of Random Tensors on the Sphere". In: *APPROX-RANDOM*. Vol. 81. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 31:1–31:20 (cit. on p. 3400).

Emmanuel J. Candès and Benjamin Recht (2009). "Exact Matrix Completion via Convex Optimization". *Foundations of Computational Mathematics* 9.6, pp. 717–772 (cit. on pp. 3383, 3384).

Yudong Chen (2015). "Incoherence-Optimal Matrix Completion". *IEEE Trans. Information Theory* 61.5, pp. 2909–2923 (cit. on pp. 3383, 3384).

Luca Chiantini and Giorgio Ottaviani (2012). "On Generic Identifiability of 3-Tensors of Small Rank". *SIAM J. Matrix Analysis Applications* 33.3, pp. 1018–1037 (cit. on p. 3386).

Andrea Crisanti, Luca Leuzzi, and Giorgio Parisi (2002). "The 3-SAT problem with large number of clauses in the∞-replica symmetry breaking scheme". *Journal of Physics A: Mathematical and General* 35.3, p. 481 (cit. on p. 3401).

Amit Daniely, Nati Linial, and Shai Shalev-Shwartz (2014). "From average case complexity to improper learning complexity". In: *STOC*. ACM, pp. 441–448 (cit. on p. 3401).

Sanjoy Dasgupta (1999). "Learning Mixtures of Gaussians". In: *FOCS*. IEEE Computer Society, pp. 634–644 (cit. on p. 3389).

Yash Deshpande and Andrea Montanari (2015). "Improved Sum-of-Squares Lower Bounds for Hidden Clique and Hidden Submatrix Problems". In: *COLT*. Vol. 40. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 523–562 (cit. on p. 3391).

Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart (2018). "List-Decodable Robust Mean Estimation and Learning Mixtures of Spherical Gaussians Mixture Models, Robustness, and Sum of Squares Proofs". In: *STOC*. ACM, (to appear) (cit. on p. 3389).

Uriel Feige (2002). "Relations between average case complexity and approximation complexity". In: *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*. ACM, New York, pp. 534–543. MR: 2121179 (cit. on p. 3401).

Uriel Feige and Robert Krauthgamer (2000). "Finding and certifying a large hidden clique in a semirandom graph". *Random Structures Algorithms* 16.2, pp. 195–208. MR: 1742351 (cit. on p. 3390).

Uriel Feige and Gideon Schechtman (2002). "On the optimality of the random hyperplane rounding technique for MAX CUT". *Random Structures Algorithms* 20.3. Probabilistic methods in combinatorial optimization, pp. 403–440. MR: 1900615 (cit. on p. 3390).

Rong Ge and Tengyu Ma (2015). "Decomposing Overcomplete 3rd Order Tensors using Sum-of-Squares Algorithms". In: *APPROX-RANDOM*. Vol. 40. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, pp. 829–849 (cit. on pp. 3387, 3389).

Sevag Gharibian (2010). "Strong NP-hardness of the quantum separability problem". *Quantum Information & Computation* 10.3, pp. 343–360 (cit. on p. 3377).

Michel X. Goemans and David P. Williamson (1995). "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming". *J. Assoc. Comput. Mach.* 42.6, pp. 1115–1145. MR: 1412228 (cit. on p. 3398).

Dima Grigoriev (2001a). "Complexity of Positivstellensatz proofs for the knapsack". *Computational Complexity* 10.2, pp. 139–154 (cit. on p. 3390).

– (2001b). "Linear lower bound on degrees of Positivstellensatz calculus proofs for the parity". *Theor. Comput. Sci.* 259.1-2, pp. 613–622 (cit. on pp. 3390, 3392, 3396).

David Gross (2011). "Recovering Low-Rank Matrices From Few Coefficients in Any Basis". *IEEE Trans. Information Theory* 57.3, pp. 1548–1566 (cit. on pp. 3383, 3384).

Leonid Gurvits (2003). "Classical deterministic complexity of Edmonds' Problem and quantum entanglement". In: *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*. ACM, pp. 10–19 (cit. on p. 3377).

Richard A Harshman (1970). "Foundations of the PARAFAC procedure: Models and conditions for an" explanatory" multi-modal factor analysis" (cit. on pp. 3386, 3397).

Christopher J. Hillar and Lek-Heng Lim (2013). "Most tensor problems are NP-hard". *J. ACM* 60.6, Art. 45, 39. MR: 3144915 (cit. on p. 3398).

Sam B. Hopkins and Jerry Li (2018). "Mixture Models, Robustness, and Sum of Squares Proofs". In: *STOC*. ACM, (to appear) (cit. on pp. 3382, 3383, 3389).

Samuel B. Hopkins, Pravesh K. Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer (2017). "The Power of Sum-of-Squares for Detecting Hidden Structures". In: *FOCS*. IEEE Computer Society, pp. 720–731 (cit. on pp. 3396, 3398, 3403, 3404).

Samuel B. Hopkins, Pravesh Kothari, Aaron Henry Potechin, Prasad Raghavendra, and Tselil Schramm (2016). "On the Integrality Gap of Degree-4 Sum of Squares for Planted Clique". In: *SODA*. SIAM, pp. 1079–1095 (cit. on p. 3391).

Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer (2016). "Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors". In: *STOC*. ACM, pp. 178–191 (cit. on p. 3387).

Samuel B. Hopkins, Jonathan Shi, and David Steurer (2015). "Tensor principal component analysis via sum-of-square proofs". In: *COLT*. Vol. 40. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 956–1006 (cit. on p. 3382).

Daniel J. Hsu and Sham M. Kakade (2013). "Learning mixtures of spherical gaussians: moment methods and spectral decompositions". In: *ITCS*. ACM, pp. 11–20 (cit. on p. 3386).

Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant (2010). "Efficiently learning mixtures of two Gaussians". In: *STOC*. ACM, pp. 553–562 (cit. on p. 3389).

Subhash A. Khot and Nisheeth K. Vishnoi (2015). "The unique games conjecture, integrability gap for cut problems and embeddability of negative-type metrics into ℓ_1". *J. ACM* 62.1, Art. 8, 39. MR: 3323774 (cit. on p. 3390).

Pravesh K. Kothari, Ryuhei Mori, Ryan O'Donnell, and David Witmer (2017). "Sum of squares lower bounds for refuting any CSP". In: *STOC*. ACM, pp. 132–145 (cit. on p. 3401).

Pravesh K. Kothari, Jacob Steinhardt, and David Steurer (2018). "Robust moment estimation and improved clustering via sum-of-squares". In: *STOC*. ACM, (to appear) (cit. on pp. 3382, 3383, 3389).

Jean-Louis Krivine (1964). "Anneaux préordonnés". *Journal d'analyse mathématique* 12.1, pp. 307–326 (cit. on p. 3374).

Jean B. Lasserre (2000/01). "Global optimization with polynomials and the problem of moments". *SIAM J. Optim.* 11.3, pp. 796–817. MR: 1814045 (cit. on p. 3378).

Lieven De Lathauwer, Joséphine Castaing, and Jean-François Cardoso (2007). "Fourth-Order Cumulant-Based Blind Identification of Underdetermined Mixtures". *IEEE Trans. Signal Processing* 55.6-2, pp. 2965–2973 (cit. on p. 3386).

S. E. Leurgans, R. T. Ross, and R. B. Abel (1993). "A decomposition for three-way arrays". *SIAM J. Matrix Anal. Appl.* 14.4, pp. 1064–1083. MR: 1238921 (cit. on p. 3386).

Tengyu Ma, Jonathan Shi, and David Steurer (2016). "Polynomial-Time Tensor Decompositions with Sum-of-Squares". In: *FOCS*. IEEE Computer Society, pp. 438–446 (cit. on pp. 3382, 3386, 3387).

Raghu Meka, Aaron Potechin, and Avi Wigderson (2015). "Sum-of-squares Lower Bounds for Planted Clique". In: *STOC*. ACM, pp. 87–96 (cit. on p. 3391).

Ankur Moitra and Gregory Valiant (2010). "Settling the Polynomial Learnability of Mixtures of Gaussians". In: *FOCS*. IEEE Computer Society, pp. 93–102 (cit. on p. 3389).

Andrea Montanari and Emile Richard (2014). "A statistical model for tensor PCA". *CoRR* abs/1411.1076 (cit. on p. 3399).

Elchanan Mossel and Sébastien Roch (2005). "Learning nonsingular phylogenies and hidden Markov models". In: *STOC*. ACM, pp. 366–375 (cit. on p. 3386).

Pablo A Parrilo (2000). "Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization". PhD thesis. California Institute of Technology (cit. on p. 3378).

Aaron Potechin and David Steurer (2017). "Exact tensor completion with sum-of-squares". In: *COLT*. Vol. 65. Proceedings of Machine Learning Research. PMLR, pp. 1619–1673 (cit. on pp. 3382, 3383, 3385).

Prasad Raghavendra, Satish Rao, and Tselil Schramm (2017). "Strongly refuting random CSPs below the spectral threshold". In: *STOC*. ACM, pp. 121–131 (cit. on pp. 3400, 3401).

Benjamin Recht (2011). "A Simpler Approach to Matrix Completion". *Journal of Machine Learning Research* 12, pp. 3413–3430 (cit. on pp. 3383, 3384).

Bruce Reznick (2000). "Some concrete aspects of Hilbert's 17th Problem". In: *Real algebraic geometry and ordered structures (Baton Rouge, LA, 1996)*. Vol. 253. Contemp. Math. Amer. Math. Soc., Providence, RI, pp. 251–272. MR: 1747589 (cit. on p. 3378).

Grant Schoenebeck (2008). "Linear Level Lasserre Lower Bounds for Certain k-CSPs". In: *FOCS*. IEEE Computer Society, pp. 593–602 (cit. on pp. 3391, 3392, 3396).

Tselil Schramm and David Steurer (2017). "Fast and robust tensor decomposition with applications to dictionary learning". In: *COLT*. Vol. 65. Proceedings of Machine Learning Research. PMLR, pp. 1760–1793 (cit. on p. 3387).

Gilbert Stengle (1974). "A Nullstellensatz and a Positivstellensatz in semialgebraic geometry". *Mathematische Annalen* 207.2, pp. 87–97 (cit. on p. 3374).

Ryota Tomioka and Taiji Suzuki (2014). "Spectral norm of random tensors". *arXiv preprint arXiv:1407.1870* (cit. on p. 3399).

Luca Trevisan (2012). "On Khot's unique games conjecture". *Bull. Amer. Math. Soc. (N.S.)* 49.1, pp. 91–111. MR: 2869009 (cit. on p. 3379).

Joel A. Tropp (2015). "An Introduction to Matrix Concentration Inequalities". *Foundations and Trends in Machine Learning* 8.1-2, pp. 1–230 (cit. on pp. 3398, 3399).

Madhur Tulsiani (2009). "CSP gaps and reductions in the lasserre hierarchy". In: *STOC*. ACM, pp. 303–312 (cit. on p. 3391).

Santosh Vempala and Grant Wang (2004). "A spectral algorithm for learning mixture models". *J. Comput. Syst. Sci.* 68.4, pp. 841–860 (cit. on pp. 3389, 3390).

Received 2018-02-27.

Prasad Raghavendra
U. C. Berkeley
nrprasad@gmail.com

Tselil Schramm (צליל שרם)
MIT and Harvard
tschramm@cs.berkeley.edu

David Steurer
ETH Zürich
dsteurer@gmail.com