# High-dimensional estimation
# via sum-of-squares proofs

Prasad Raghavendra*        Tselil Schramm†        David Steurer‡

September 12, 2019

## Abstract

Estimation is the computational task of recovering a *hidden parameter x* associated with a distribution $\mathcal{D}_x$, given a *measurement y* sampled from the distribution. High dimensional estimation problems arise naturally in statistics, machine learning, and complexity theory.

Many high dimensional estimation problems can be formulated as systems of polynomial equalities and inequalities, and thus give rise to natural probability distributions over polynomial systems. Sum-of-squares proofs provide a powerful framework to reason about polynomial systems, and further there exist efficient algorithms to search for low-degree 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 when the system is feasible. On the other hand, a general technique referred to as *pseudocalibration* has been developed towards showing lower bounds on the 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 existence of spectral algorithms. In this article we survey these developments.

# Contents

# 1 Introduction

In estimation problems, the goal is to recover a structured object from an observed input which partially obfuscates it. Formally, an estimation problem is specified by a family of distributions $\{\mathcal{D}_x\}$ over $\mathbb{R}^N$ parametrized by $x \in \mathbb{R}^n$. The input consists of a sample $y \in \mathbb{R}^N$ drawn from $\mathcal{D}_x$ for some $x \in \mathbb{R}^n$, and the goal is to recover the value of the parameter $x$. We refer to $x$ as the *hidden variable* or the *parameter*, and to the sample $y$ as 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 when recovery is information-theoretically possible, in many high-dimensional settings it is computationally intractable to recover $x$. For these reasons, we often seek 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 $\mathbb{E}_{y \sim \mathcal{D}_x}[\|\theta(y) - x\|^2]$.

In many cases, we can formulate such a minimization problem as a feasibility problem for a system of polynomial equations. By classical NP-completeness results, general polynomial systems in many variables are computationally intractable in the worst case. In our context, an estimation problem gives rise to a distribution over polynomial systems that encode it. We wish to study a typical system drawn from this distribution. If the underlying distributions are sufficiently well-behaved, polynomial systems yield an avenue to design algorithms for high-dimensional estimation problems.

In this survey, our tool for studying such polynomial systems will be sum-of-squares (SoS) proofs. Sum-of-squares proofs yield a complete proof system for reasoning about polynomial systems [Kri64, Ste74]. More importantly, SoS 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. Low-degree SoS proofs can be found efficiently, and the computational complexity of the algorithm grows exponentially with the degree of the polynomials involved in the proof.

The study of low-degree SoS proofs in the context of estimation problems suggests a rich family of questions. For natural estimation problems, if a polynomial system drawn from the corresponding distribution is feasible, can one harness sum-of-squares proofs towards solving the polynomial system? (surprisingly, the answer is often yes!) If a system from this distribution is typically infeasible, what is the smallest degree of a sum-of-squares refutation? Are there structural characterizations of the degree of SoS refutations in terms of the properties of the distribution? Is there a connection between the existence of low-degree SoS proofs and the spectra of random matrices associated with the distribution (yielding efficient spectral algorithms)? Over the past few years, significant strides have been made on all these fronts, exposing the contours of a rich theory that remains largely 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 $\mathbb{G}(n, \frac{1}{2})$. The goal is to recover the $k$-clique.

Formally, the structured family $\{\mathcal{D}_S\}$ is parametrized by subsets $S \subset \binom{[n]}{k}$. For a subset $S \in \binom{[n]}{k}$, the distribution $\mathcal{D}_S$ over measurements $G \in \{0,1\}^{\binom{n}{2}}$ is specified by the following sampling procedure:

- Sample a graph $G' = ([n], E(G'))$ from the Erdős-Rényi distribution $\mathbb{G}(n, \frac{1}{2})$ and set $G = ([n], E(G') \cup E(K_S))$ where $K_S$ denotes the clique on the vertices in $S$.

An application of the second moment method [GM75] shows that for all $k \gg 2 \log n$, the clique $S$ can be exactly recovered with high probability given the graph $G$. However, for any $k \ll \sqrt{n}$, there is no known polynomial time algorithm for the problem with the best algorithm being a brute force search running in time $n^{O(\log n)}$. Improving upon this runtime is an open problem dating back to Karp in 1976 [Kar76], but save for the spectral algorithm of Alon et al. for $k = \Omega(\sqrt{n})$ [AKS98a], the only progress has been in proving lower bounds against broad classes of algorithms (e.g. [Jer92, FK03, FGR+17, BHK+16]).

We will now see how to encode the problem as a polynomial system. For pairs $(S, G)$, let $y \in \{\pm 1\}^{\binom{n}{2}}$ denote the natural $\{\pm 1\}$-encoding of the graph $G$, namely, $y_{ij} = (1 - 2 \cdot \mathbf{1}[(i, j) \notin E(G)])$ 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*. We encode each constraint as a polynomial equality or inequality:

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

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

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

Note that when we are solving the estimation problem, 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 $x \in \{0,1\}^n$ which are supported on cliques of size at least $k$ in $G$.

**Refutation and distinguishing.** For every estimation problem that we will encounter in this survey, we can associate two related computational problems termed *refutation* and *distinguishing*. In estimation problems, we typically think of instances $y$ as having structure: we sample $y$ from a structured distribution $\mathcal{D}_x$, and we wish to recover the hidden variables $x$ that give structure to $\mathcal{D}_x$. But there may also be instances $y$ which do

not have structure. The goal of *refutation* is to certify that there is no hidden structure, when there is none.

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 the Erdős-Rényi random graph $\mathbb{G}(n, \frac{1}{2})$ (without a planted clique). With high probability, a graph $y \sim \mathbb{G}(n, \frac{1}{2})$ has no clique with significantly more than $2 \log n$ vertices. Therefore, for a fixed $k \gg 2 \log n$, given a graph $y \sim \mathbb{G}(n, \frac{1}{2})$, 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 $\pi$ on the hidden variables $x \in \mathbb{R}^n$, which in turn induces a distribution $\mathcal{D}_*$ on $\mathbb{R}^N$, obtained by first sampling $x \sim \pi$ and then sampling $y \sim \mathcal{D}_x$. The input consists of a sample $y$ which is with equal probability drawn from the structured distribution $\mathcal{D}_*$ or the null distribution $\mathcal{D}_\emptyset$. 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 $[n]$ of size $k$. In the distinguishing problem, the input is a graph drawn from either $\mathcal{D}_*$ or the null distribution $\mathbb{G}(n, \frac{1}{2})$, 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 $\{\mathcal{D}_v\}$ is parametrized by unit vectors $v \in \mathbb{R}^n$. A sample from $\mathcal{D}_v$ consists of a 4-tensor $\mathbf{T} = v^{\otimes 4} + \zeta$ 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 $v$.

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

$$x' = \operatorname{argmax}_{\|x\| \leqslant 1} \langle \mathbf{T}, x^{\otimes 4} \rangle$$

then one can show that $\|v - x'\|_2 \leqslant O(n^{1/2} \cdot \sigma)$ with high probability over $\zeta$. If $\mathbf{T} \sim \mathcal{D}_v$ then $\langle \mathbf{T}, v^{\otimes 4} \rangle = 1$. Furthermore, when $\sigma \ll n^{-1/2}$ it can be shown that $v \in \mathbb{R}^n$ is close to the unique maximizer of the function $\phi(x) = \langle \mathbf{T}, x^{\otimes 4} \rangle$. So the problem of recovering $v$ can be encoded as the following polynomial system:

$$x \text{ is in the unit sphere} \qquad \qquad \|x\|^2 \leqslant 1,$$

$$x \text{ has large value for } \mathbf{T} \qquad \qquad \sum_{i,j,k,\ell \in [n]^4} \mathbf{T}_{ijk\ell} x_i x_j x_k x_\ell \geqslant 1.$$

In the distinguishing and refutation versions of this problem, we will take the *null* distribution $\mathcal{D}_\emptyset$ to be the distribution over 4-tensors with independent Gaussian entries

sampled from $N(0, \sigma^2)$ (equivalent to the distribution of the noise $\zeta$ from $\mathcal{D}_v$). For a 4-tensor $\mathbf{T}$, the maximum of $\mathbf{T}(x) = \langle x^{\otimes 4}, \mathbf{T} \rangle$ over the unit ball is referred to as the *injective tensor norm* of the tensor $\mathbf{T}$, and is denoted by $\|\mathbf{T}\|_{\text{inj}}$. If $\mathbf{T} \sim \mathcal{D}_\emptyset$ then $\|\mathbf{T}\|_{\text{inj}} \leqslant O\left(n^{1/2} \cdot \sigma\right)$ with high probability over choice of $\mathbf{T}$ [ABAČ]. Thus when $\sigma \ll n^{-1/2}$, the refutation version of the TENSOR PCA problem reduces to certifying an upper bound on $\|\mathbf{T}\|_{\text{inj}}$. If we could compute $\|\mathbf{T}\|_{\text{inj}}$ exactly, then we can certify that $\mathbf{T} \sim \mathcal{D}_\emptyset$ for $\sigma$ as large as $\sigma = O(n^{-1/2})$.

The injective tensor norm is known to be computationally intractable in the worst case [Gur03, Gha10, BBH$^+$12]. Understanding the the function $\langle x^{\otimes k}, \mathbf{T} \rangle$ for random $\mathbf{T}$ is a deep topic in probability theory and statistical physics (e.g. [ABAČ]). As an estimation problem, TENSOR PCA was first considered by [MR14], and inspired multiple follow-up works concerned with spectral and SoS algorithms (e.g. [HSS15, HSSS16, RRS17, BGL17]).

**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 $X_\Omega$ for a subset $\Omega \subset [n] \times [n]$ with $|\Omega| = m$. The probability distribution $\mathcal{D}_X$ over measurements is obtained by picking the set $\Omega$ to be a uniformly random subset of $m$ entries.

To formulate a polynomial system for recovering a rank-$r$ matrix consistent with the measurement $X_\Omega$, we will use a $n \times r$ matrix of variables $B$, and write the following system of constraints on it:

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

Tensor completion is the analogous problem with $\mathbf{X}$ being a higher-order tensor namely, $\mathbf{X} = \sum_{i=1}^{r} a_i^{\otimes k}$ 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,

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

## 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 $\mathcal{A} = \{p_i(x) \geqslant 0\}_{i \in [m]}$ in variables $x_1, \ldots, x_n$. We will refer to these inequalities as the *axioms*. Starting with the axioms $\mathcal{A}$, a sum-of-squares proof of $q(x) \geqslant 0$ is given by an identity of the form,

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

where $\{s_j(x)\}_{j \in [m']}, \{a_i(x)\}_{i \in [m]}$ 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., $\max\{\deg(s_j^2), \deg(a_i^2 p_i)\}_{i,j}$.

Sum-of-squares proofs extend naturally to polynomial systems that involve a set of equalities $\{r_i(x) = 0\}$ along with a set of inequalities $\{p_i(x) \geqslant 0\}$. We can extend the definition syntactically by replacing 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 $\mathcal{A} \left|\frac{x}{d}\right. \{q(x) \geqslant 0\}$ 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 $\mathcal{A} \left|\frac{x}{d}\right. \{q(x) \geqslant 0\}$ 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 $\mathcal{A} \left|-\right. \{q(x) \geqslant 0\}$. Sum-of-squares proofs can also be used to certify the infeasibility, or *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,

$$-1 = \sum_{i \in [k]} s_i^2(x) + \sum_{i \in [m]} a_i^2(x) \cdot p_i(x) \tag{1.1}$$

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

The sum-of-squares proof system has been an object of study starting with the work of Hilbert and Minkowski more than a century ago (see [Rez00] for a survey). With no restriction on degree, Stengle's Positivestellensatz implies 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 an SoS proof of this fact.

The algorithmic implications of the sum-of-squares proof system were first realized in the works of Parrilo [Par00] and Lasserre [Las01], 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-$d$ SoS proof in time $n^{O(d)}$, if there exists one. This family of algorithms (called a hierarchy, as we have an algorithm for each even integer degree-$d$) are referred to as the sum-of-squares SDP hierarchy. We say that the SoS algorithm is *low-degree* if $d$ does not grow with $n$.

The SoS hierarchy has since emerged as a powerful tool for algorithm design. On the one hand, the first few levels of the SoS hierarchy systematically capture a vast majority of algorithms in combinatorial optimization and approximation algorithms developed over several decades. 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 [Tre12, BS14].

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, and clustering problems. 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 a low-degree SoS SDP can be used to actually retrieve $x$. We will discuss this broad principle and several 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 of infeasibility?* 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. [BHK$^+$16] 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 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 (spectral) 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 and more lightweight 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 $\{1, \ldots, n\}$. We sometimes use $\binom{[n]}{d}$ to denote the set of all subsets of $[n]$ of size $d$, and $[n]^{\leqslant d}$ 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 $x^{\leqslant d}$ 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 $\mathbb{R}[x]_{\leqslant d}$ 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 $\mu$ 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 $\mu$. For an event $\mathcal{E}$, we will use $\mathbf{1}[\mathcal{E}]$ as the indicator that $\mathcal{E}$ occurs. We use $\mathbb{G}(n, \frac{1}{2})$ to denote the Erdős-Rényi distribution with parameter $\frac{1}{2}$, or the distribution over graphs where each edge is included independently with probability $\frac{1}{2}$.

If $M$ is an $n \times m$ matrix, we use $\lambda_{\max}(M)$ to denote $M$'s largest eigenvalue. 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 $M$, $\|M\|_{\mathrm{F}} = \langle M, M \rangle^{1/2}$. 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 $A$. For an order-$k$ tensor $T \in (\mathbb{R}^n)^{\otimes k}$ and for a permutation of $[k]$ $i_1, \ldots, i_k$, we denote by $T_{\{i_1, \ldots, i_\ell\}\{i_{\ell+1}, \ldots, i_k\}}$ the $n^\ell \times n^{k-\ell}$ matrix reshaping given by ordering the modes of $T$ so that $i_1, \ldots, i_\ell$ index the rows and $i_{\ell+1}, \ldots, i_k$ index the columns.

**Pseudoexpectations.** For a polynomial system $\mathcal{P}$ in $n$ variables $x \in \mathbb{R}^n$ consisting of inequalities $\{p_i(x) \geqslant 0\}_{i \in [m]}$, we can write an SDP of size $n^{O(d)}$ which finds a degree-$d$ sum-of-squares refutation, if one exists (see [Rot13] for more discussion).

If there is no degree-$d$ refutation, the dual semidefinite program computes in time $n^{O(d)}$ a linear functional over degree-$d$ polynomials which we term a *pseudoexpectation*. Formally, a degree-$d$ pseudoexpectation $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$ 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. If 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 Eq. (1.1), we have $-1$. Applying $\tilde{\mathbb{E}}$ to the right-hand-side of Eq. (1.1), 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. □

The properties above imply that when $\mathcal{A} \vdash_d^x \{q(x) \geqslant 0\}$, then if $\tilde{\mathbb{E}}$ is a degree-$d$ pseudoexpectation operator for the polynomial system defined by $\mathcal{A}$, $\tilde{\mathbb{E}}[q(x)] \geqslant 0$ as well. This implies that $\tilde{\mathbb{E}}$ satisfies several useful inequalities; for example, the Cauchy-Schwarz inequality.

*Claim* 1.5. If $\tilde{\mathbb{E}}$ is a degree-$d$ pseudoexpectation and if $p, q$ are polynomials of degree at most $\frac{d}{2}$, then $\tilde{\mathbb{E}}[q(x) \cdot p(x)] \leqslant \frac{1}{2} \tilde{\mathbb{E}}[q(x)^2] + \frac{1}{2} \tilde{\mathbb{E}}[p(x)^2]$.

*Proof.* We have the following polynomial equality of degree at most $d$:

$$q(x)p(x) = \frac{1}{2} \cdot q(x)^2 + \frac{1}{2} \cdot q(x)^2 - \frac{1}{2} \left( q(x) - p(x) \right)^2.$$

Applying $\tilde{\mathbb{E}}$ to both sides, using that $\tilde{\mathbb{E}}[(q(x) - p(x))^2] \geqslant 0$, we have our conclusion. □

Other versions of the Cauchy-Schwarz inequality can be shown to hold for pseudoexpectations as well; see e.g. [BBH+12] for details.

# 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. This unifying perspective allows us to obtain algorithms with significantly better guarantees than what's known to be possible with other methods. We illustrate the power of this meta-theorem by applying it to matrix and tensor completion, tensor decomposition, and clustering.

## 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,

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

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

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

For both examples, the measurement was a simple (linear) function of the parameter. This is not always the case; consider for example the following clustering problem. There are two distinct centers $\mu_1, \mu_2 \in \mathbb{R}^n$, and we observe $m$ samples $y_1, \ldots y_m \in \mathbb{R}^n$ such that each sample is closer to either $\mu_1$ or $\mu_2$. Then we can encode the problem of finding $\mu_1$ and $\mu_2$ as follows,

$$\mathcal{P}_{\mu,m} = \left\{ ((\mu_1, \mu_2), Y) \mid \mu_1 \neq \mu_2 \in \mathbb{R}^n, \ Y \in \mathbb{R}^{n \times m}, \ \forall i \in [m], \ \big| \|y_i - \mu_1\| - \|y_i - \mu_2\| \big| > 0 \right\} .$$

**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

---

[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.

say that $x$ is identifiable (up to error $\varepsilon$) if every $(x, y) \in \mathcal{P}$ satisfies that $y$ identifies $x$ (up to error $\varepsilon$).

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

**From 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 $y^*$ (e.g. by brute-force search over the set of all $x$). 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

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

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 $\mathcal{A} \left|\frac{x,z}{d}\right. \{\|x^* - x\|^2 \leqslant \varepsilon\}$, where $\mathcal{A} = \{p(x, y^*, z) = 0\}$. Then, every degree-d pseudo-distribution D consistent with the constraints $\mathcal{A}$ satisfies*

$$\left\| x^* - \underset{D(x,z)}{\tilde{\mathbb{E}}} \; x \right\|^2 \leqslant \varepsilon \,.$$

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

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 [BKS15, BKS14, HSS15, MSS16, BM16, PS17, KSS18, HL18].

---

[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 which description is best. However, typically, the more auxiliary variables the better.

[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 $n$.

*Proof.* Let $D$ be a degree-$d$ pseudo-distribution $D$ with $D \mathrel{\vphantom{|}\smash{\left|\frac{\phantom{xx}}{\phantom{x}}\right.}^{x,z}} \mathcal{A}$. Since degree-$d$ sum-of-squares proofs are sound for degree-$d$ pseudo-distributions, we have $D \mathrel{\vphantom{|}\smash{\left|\frac{\phantom{xx}}{d}\right.}^{x,z}} \{\|x^* - x\|^2 \leqslant \varepsilon\}$. In particular, $\tilde{\mathbb{E}}_{D(x,z)}\|x^* - x\|^2 \leqslant \varepsilon$. By Cauchy–Schwarz for pseudo-distributions (Claim 1.5), every vector $u \in \mathbb{R}^n$ satisfies

$$\langle u, \underset{D(x,z)}{\tilde{\mathbb{E}}}\, x^* - x\rangle = \underset{D(x,z)}{\tilde{\mathbb{E}}}\, \langle u, x^* - x\rangle \leqslant \left(\tilde{\mathbb{E}}\|u\|^2\right)^{1/2} \cdot \left(\tilde{\mathbb{E}}\|x^* - x\|^2\right)^{1/2} \leqslant \|u\| \cdot \varepsilon^{1/2}.$$

By choosing $u = \tilde{\mathbb{E}}_{D(x,z)}\, x^* - x$, we obtain the desired conclusion about $\tilde{\mathbb{E}}_{D(x,z)}\, x$.

Given a measurement $y^*$, the algorithm computes a degree-$d$ pseudo-distribution $D(x, z)$ that satisfies $\mathcal{A}$ up to error $2^{-n^{100d}}$ and outputs $\hat{x}(y^*) = \tilde{\mathbb{E}}_{D(x,z)}\, x$. We are guaranteed that such a pseudo-distribution exists, e.g. the distribution that places all its probability mass on the vector $x^*$. If the proof $\mathcal{A} \mathrel{\vphantom{|}\smash{\left|\frac{x}{d}\right.}} \{\|x^* - x\|^2 \leqslant \varepsilon\}$ has bit-complexity $n^d$, it follows that $D(x)$ satisfies $\{\|x^* - x\|^2 \leqslant \varepsilon\}$ up to error $2^{-n^d}$. In particular, $\tilde{\mathbb{E}}_{D(x)}\|x^* - x\|^2 \leqslant \varepsilon + 2^{-n^d}$. By the same argument as before, it follows that $\|x^* - \hat{x}(y^*)\|^2 \leqslant \varepsilon + 2^{-n^d}$. $\quad\square$

## 2.2 Matrix and tensor completion

In matrix completion, we observe a few entries of a low-rank 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 applications 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 $X = \sum_{i=1}^r \sigma_i \cdot u_i v_i^{\mathsf{T}} \in \mathbb{R}^{n \times n}$ 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 $\mu$-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 $X_\Omega$ 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 $\Omega$ the algorithm recovers $X$ exactly [CR09, Gro11, Rec11, Che15]. This bound on $m$ is nearly optimal in that $m \geqslant \Omega(rn)$ appears to be necessary because an $n$-by-$n$ rank-$r$ matrix has $\Omega(rn)$ 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 [CR09, Gro11, Rec11, Che15]. 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).

---

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

**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 $X = \sum_{i=1}^{r} a_i a_i^\mathsf{T}$ for $\mu$-incoherent orthonormal vectors $a_1, \ldots, a_r \in \mathbb{R}^n$. The following theorem shows that, with high probability over the choice of $\Omega$, the matrix $X$ is identified by the partial matrix $X_\Omega$. 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 $X_\Omega$.

**Theorem 2.2** (implicit in [CR09, Gro11, Rec11, Che15]). *Let $X = \sum_{i=1}^{r} a_i a_i^\mathsf{T} \in \mathbb{R}^{n \times n}$ be an r-dimensional projector and $a_1, \ldots, a_r \in \mathbb{R}^n$ 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 the n-by-r matrix variable $B$,*

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

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

$$\mathcal{A} \left|\frac{B}{4}\right. \left\{ \left\| BB^\mathsf{T} - X \right\|_\mathrm{F} = 0 \right\} .$$

*Proof.* The analyses of the aforementioned algorithm for matrix completion [CR09, Gro11, Rec11, Che15] show the following: let $\overline{\Omega}$ be the complement of $\Omega$ in $[n] \times [n]$. Then if $X$ satisfies our incoherence assumptions, with high probability over the choice of $\Omega$, there exists[5] a symmetric matrix $M$ with $M_{\overline{\Omega}} = 0$ 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 $M_{\overline{\Omega}} = 0$ and $\mathcal{A}$ contains the equation $(BB^\mathsf{T})_\Omega = X_\Omega$, we have $\mathcal{A} \left|\frac{B}{}\right. \langle M, BB^\mathsf{T} \rangle = \langle M, X \rangle \geqslant r$. At the same time, we have

$$\mathcal{A} \left|\frac{}{}\right. \langle M, BB^\mathsf{T} \rangle \leqslant \langle X, BB^\mathsf{T} \rangle + 0.9 \langle \mathrm{Id}_n - X, BB^\mathsf{T} \rangle = 0.1 \langle X, BB^\mathsf{T} \rangle + 0.9 r ,$$

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

$$\mathcal{A} \left|\frac{}{}\right. \langle X, BB^\mathsf{T} \rangle \geqslant r .$$

Together with the facts $\|X\|_\mathrm{F}^2 = r$ and $\mathcal{A} \left|\frac{}{}\right. \left\| BB^\mathsf{T} \right\|_\mathrm{F}^2 = r$, we obtain $\mathcal{A} \left|\frac{}{}\right. \left\| X - BB^\mathsf{T} \right\|_\mathrm{F}^2 = 0$ as desired. $\square$

---

[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 [Gro11, Rec11, Che15].

**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 [PS17] (of tensors with orthogonal low-rank decompositions) and approximate recovery [BM16] (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** ([PS17]). *Let $a_1, \ldots, a_r \in \mathbb{R}^n$ be orthonormal vectors with incoherence $\mu = \max_{i,j} n \cdot \langle a_i, e_j \rangle^2$ and let $\mathbf{X} = \sum_{i=1}^{r} a_i^{\otimes 3}$ 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 the n-by-r matrix variable $B$ with columns $b_1, \ldots, b_r$,*

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

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

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

*Proof.* Let $A \in \mathbb{R}^{n \times r}$ be the matrix with columns $a_1, \ldots, a_r$. Analogous to the proof for matrix completion, the heart of the proof is the existence of a 3-tensor $\mathbf{T}$ that satisfies the following properties: $\mathbf{T}_{\overline{\Omega}} = 0$, $\langle \mathbf{T}, a_i^{\otimes 3} \rangle = 1$, and

$$\{\|x\|^2 = 1\} \Big|_{6}^{x} \quad \langle \mathbf{T}, x^{\otimes 3} \rangle \leqslant 1 - \tfrac{1}{100} \left( 1 - \sum_{i=1}^{r} \langle a_i, x \rangle^2 \right) - \tfrac{1}{100} \left( \sum_{i \neq j} \langle a_i, x \rangle^2 \langle a_j, x \rangle^2 \right). \quad (2.1)$$

These properties imply that $a_1, \ldots, a_r$ are the unique global maximizers of the cubic polynomial $\langle \mathbf{T}, x^{\otimes 3} \rangle$ over the unit sphere. (We remark that for matrix completion, the spectral properties of the matrix $M$ imply that the unique global optimizers of the quadratic polynomial $\langle M, x^{\otimes 2} \rangle$ are the unit vectors in the span of $a_1, \ldots, a_r$.)

The proof that this tensor $\mathbf{T}$ exists follows the same approach as the proof of existence of the matrix $M$ for matrix completion in Theorem 2.2 and proceeds by an iterative construction [Rec11, Gro11]. The main difference is due to the fact that for $M$ we only

need to ensure spectral properties, whereas for $\mathbf{T}$ we need to ensure the existence of (higher-degree) sum-of-squares proofs Eq. (2.1). We refer to previous literature for details of the proof that such $\mathbf{T}$ exists with high probability over the choice of $\Omega$ [PS17].

Similar to the proof for matrix completion, we have by the properties of $\mathbf{T}$ that $\langle \mathbf{T}, \mathbf{X} \rangle = r$ and $\mathcal{A} \vdash \{\langle \mathbf{T}, \sum_{i=1}^{r} b_i^{\otimes 3} \rangle = \langle \mathbf{T}, \mathbf{X} \rangle = r\}$. By Eq. (2.1) and linearity,

$$\mathcal{A} \vdash \langle \mathbf{T}, \textstyle\sum_{i=1}^{r} b_i^{\otimes 3} \rangle \leqslant r - \tfrac{1}{100} \textstyle\sum_{i=1}^{r}(1 - \textstyle\sum_{j=1}^{r}\langle a_j, b_i \rangle^2) - \tfrac{1}{100} \textstyle\sum_{i=1}^{r} \textstyle\sum_{j\neq k}\langle a_j, b_i \rangle^2 \langle a_k, b_i \rangle^2 \,.$$

Because $\mathcal{A}$ includes the equations $\|b_1\|^2 = \cdots = \|b_r\|^2 = 1$ and because the final term is a sum of squares, we conclude that $\mathcal{A} \vdash \sum_{j=1}^{r}\langle a_j, b_i \rangle^2 = 1$ for all $i \in [r]$ and $\mathcal{A} \vdash \langle a_j, b_i \rangle^2 \langle a_k, b_i \rangle^2 = 0$ for all $i, j, k \in [r]$ with $j \neq k$. We also have the following claim:

*Claim* 2.4. When $\{a_i\}_{i\in[r]}$ are orthogonal and $\mathcal{A} \vdash \{\sum_{j\in[r]}\langle a_j, b_i \rangle^2 = 1\}_{i\in[r]}$ and $\mathcal{A} \vdash \{\langle a_{j_1}, b_i \rangle^2 \langle a_{j_2}, b_i \rangle^2 = 0\}_{j_1 \neq j_2 \in [r]}$, then $\mathcal{A} \vdash \|b_i^{\otimes 3} - \sum_{j=1}^{r}\langle a_j, b_i \rangle^3 a_j^{\otimes 3}\|^2 = 0$.

We give the (easy) proof of Claim 2.4 in Appendix A. Thus, from the orthonormality of the $a_i$,

$$\mathcal{A} \vdash r = \left\langle \mathbf{T}, \sum_{i=1}^{r} b_i^{\otimes 3} \right\rangle = \sum_{i,j}\langle a_j, b_i \rangle^3 \langle \mathbf{T}, a_j^{\otimes 3} \rangle = \sum_{i,j}\langle a_j, b_i \rangle^3 = \left\langle \mathbf{X}, \sum_{i=1}^{r} b_i^{\otimes 3} \right\rangle .$$

Together with the facts $\|\mathbf{X}\|_{\mathrm{F}}^2 = r$ and $\mathcal{A} \vdash \left\| \sum_{j=1}^{r} b_i^{\otimes 3} \right\|_{\mathrm{F}}^2 = r$, we obtain $\mathcal{A} \vdash \left\| \mathbf{X} - \sum_{j=1}^{r} b_i^{\otimes 3} \right\|_{\mathrm{F}}^2 = 0$ as desired. $\qquad\square$

## 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 $a_1, \dots, a_r \in \mathbb{R}^n$. It turns out that under mild conditions on the components $a_1, \dots, a_r$, the noise, and the tensor order $k$, this estimation task is possible information theoretically. For example, generic components $a_1, \dots, a_r \in \mathbb{R}^n$ with $r \leqslant \Omega(n^2)$ are identified by their 3-tensor $\sum_{i=1}^{r} a_i^{\otimes 3}$ [CO12] (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 [LCC07, MR05, AFH$^+$12, HK13, BCMV14, BKS15, MSS16, AGMR17]. 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 2.5** (Tensor decomposition, single component recovery, 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 $\max_{i \in [r]} \frac{1}{\|a_i\| \cdot \|u\|} |\langle a_i, u \rangle| \geq 0.9$.

Algorithms for Problem 2.5 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 Jennrich [Har70, LRA93] can solve Problem 2.5 for up to $r \leq n$ generic components if the tensor order is at least 3. (Concretely, the algorithm works for 3-tensors with linearly independent components.) Essentially the same algorithm works up to $\Omega(n^2)$ generic[7] components if the tensor order is at least 5. A more sophisticated algorithm [LCC07] solves Problem 2.5 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 is $n^{1+\Omega(1)}$, 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.

**Theorem 2.6** ([MSS16] building on [BKS15, GM15, HSSS16]). *There exists a polynomial-time algorithm to solve Problem 2.5 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 2.5 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,

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

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.

---

[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.

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 2.6.

**Theorem 2.7** (Robust Jennrich's algorithm [MSS16, SS17]). *There exists $\varepsilon > 0$ and a randomized polynomial-time algorithm that given a 3-tensor $\mathbf{T} \in (\mathbb{R}^n)^{\otimes 3}$ 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 $\|\mathrm{Id}_r - A^\top A\| \leqslant \varepsilon$, where $A \in \mathbb{R}^{n \times r}$ is the matrix with columns $a_1, \dots, a_r$. Suppose $\left\| T - \sum_i a_i^{\otimes 3} \right\|_F^2 \leqslant \varepsilon \cdot r$ and that $\max\{\|T\|_{\{1,3\}\{2\}}, \|T\|_{\{1\}\{2,3\}}\} \leqslant 10$. Then, with at least inverse polynomial probability, $\max_{i \in [r]}\langle a_i, u\rangle \geqslant 0.9$.*

We will apply Theorem 2.7 to the noisy copy of the 6-tensor $\mathbf{X}$ returned by the SoS algorithm, viewing it as a 3-tensor in the lifted/squared components $a_1 \otimes a_1, \dots, a_r \otimes a_r \in \mathbb{R}^{n^2}$; these lifted components are in $n^2$ dimensions, and may be linearly independent and close to orthogonal for $r \gg n$.[9] To ensure that our approximation to $\mathbf{X}$ meets the conditions of the theorem, we can add constraints to the SoS SDP to bound the spectral norm of rectangular reshapings of $\mathbf{X}$; see [MSS16] for details.

*Proof sketch.* We apply the following version of Jennrich's algorithm to $\mathbf{T}$: Choose a Gaussian vector $g \sim \mathcal{N}(0, \mathrm{Id})$ and compute the $d \times d$ matrix $T(g)$ given by the random contraction,

$$T(g) = \sum_{j \in [n]} g_j \cdot T_i,$$

where $T_i$ is the $n \times n$ matrix resulting from the restriction of $\mathbf{T}$ to coordinate $i$ in the third mode. Then, output the top eigenvector of $T(g)$.

To analyze this algorithm, write $\mathbf{T}$ as a sum of signal and noise terms $\mathbf{T} = S + E$, where $S = \sum_{i=1}^r a_i^{\otimes 3}$ and $\|E\|_F \leqslant \varepsilon \cdot r$. Notice that when $E = 0$,

$$T(g) = \sum_{i \in [r]} \langle g, a_i \rangle \cdot a_i a_i^\top,$$

and with probability 1 the values $\langle a_i, g \rangle$ are distinct. So when $\|\mathrm{Id}_r - A^\top A\| = 0$, the eigenvectors of $T(g)$ are exactly the $a_i$. To establish the theorem, it remains to show that when $\|E\|_F^2 \leqslant \varepsilon r$ and when the orthogonality defect is at most $\varepsilon$, the top eigenvector is still close to some $a_i$ with reasonable probability. Though the full proof is not complicated, we defer it to Appendix A. $\square$

## 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.

---

[9]To ensure that the lifted/squared components are close to orthogonal, we must stipulate conditions for $\mathbf{T}$.

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

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

*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 from the constraint $B^{\mathsf{T}} B = \mathrm{Id}$, $\mathcal{A} \left|\frac{B}{}\right. \left\| \sum_{i=1}^r b_i^{\otimes 6} \right\|_{\mathrm{F}}^2 = \left\| \sum_{i=1}^r b_i^{\otimes 3} \right\|_{\mathrm{F}}^2 = r$. Thus, by the equality $\sum_i a_i^{\otimes 3} = \sum_j b_j^{\otimes 3}$, we have $\mathcal{A} \left|-\right. \sum_{i,j} \langle a_i, b_j \rangle^3 = r$. It suffices to show $\mathcal{A} \left|-\right. \sum_{i,j} \langle a_i, b_j \rangle^6 \geqslant r$.

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

$$\mathcal{A} \left|-\right. \quad r = \sum_{i,j} \langle a_i, b_j \rangle^3 \leqslant \tfrac{1}{2} \sum_{i,j} \langle a_i, b_j \rangle^2 + \tfrac{1}{2} \sum_{i,j} \langle a_i, b_j \rangle^4 \leqslant \tfrac{1}{2} r + \tfrac{1}{2} \sum_{i,j} \langle a_i, b_j \rangle^4.$$

We conclude that $\mathcal{A} \left|-\right. \sum_{i,j} \langle a_i, b_j \rangle^4 = r$. 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 $\mathcal{A} \left|-\right. \sum_{i,j} \langle a_i, b_j \rangle^6 = r$ as desired. □

**Incoherent components.** The following lemma shows that a 6-tensor is identifiable from a 3-tensor with the same components if the components satisfy a set of simple deterministic conditions . Furthermore, this fact has a low-degree sum-of-squares proof. These conditions allow for overcomplete tensors with components $a_1, \ldots, a_r \in \mathbb{R}^n$ such that $r \geqslant n^{1+\Omega(1)}$. In fact, together with the techniques in Section 2.3, the following lemma gives a polynomial-time algorithm to solve Problem 2.5 for tensor order 3 and up to $\tilde{\Omega}(n^{1.25})$ components that are drawn uniformly at random from the unit sphere.

For $\sigma \geqslant 1$ and $\rho > 0$, we say that unit vectors $a_1, \ldots, a_r \in \mathbb{R}^n$ are $(\sigma, \rho)$-incoherent if $\sum_{i=1}^r a_i a_i^{\mathsf{T}} \preceq \sigma \cdot \mathrm{Id}$ and $|\langle a_i, a_j \rangle| \leqslant \rho$ for all $i \neq j$. Random unit vectors satisfy this property for $\sigma \leqslant \tilde{O}(r/n)$ and $\rho \leqslant \tilde{O}(1/\sqrt{n})$.

Let $B$ be an n-by-r matrix of variables and let $b_1, \ldots, b_r$ be the columns of $B$. Consider the following system of polynomial constraints

$$\mathcal{B}_\varepsilon = \left\{ \|b_i\|^2 = 1 \; \forall i \in [r], \; \left\| \sum_{i=1}^r b_i^{\otimes 3} \right\|_{\mathrm{F}}^2 \geqslant (1-\varepsilon) \cdot r, \; \left\| \sum_{i=1}^r b_i^{\otimes 6} \right\|_{\mathrm{F}}^2 \leqslant (1+\varepsilon) \cdot r \right\}. \quad (2.2)$$

We observe that $(\sigma, \rho)$-incoherent unit vectors satisfy $\mathcal{B}_\varepsilon$ for $\varepsilon = \rho\sigma$. In particular, if $a_1, \ldots, a_r$ are $(\sigma, \rho)$-incoherent unit vectors, then $\left\| \sum_{i=1}^r a_i^{\otimes 3} \right\|_F^2 = r + \sum_{i \neq j} \langle a_i, a_j \rangle^3 \geqslant (1 - \rho\sigma) \cdot r$. For a similar reason, $\left\| \sum_{i=1}^r a_i^{\otimes 6} \right\|_F^2 \leqslant (1 + \rho^4\sigma) \cdot r \leqslant (1 + \rho\sigma) \cdot r$.

**Lemma 2.9.** *Let $a_1, \ldots, a_r \in \mathbb{R}^n$ be $(\sigma, \rho)$-incoherent unit vectors. Let $B$ be an $n$-by-$r$ matrix of variables, $b_1, \ldots, b_r$ the columns of $B$, and $\mathcal{A}$ the following system of polynomial constraints,*

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

*Then,*

$$\mathcal{A} \left|\frac{B}{12}\right. \left\{ \left\| \sum_{i=1}^r a_i^{\otimes 6} - \sum_{i=1}^r b_i^{\otimes 6} \right\|_F^2 \leqslant O(\rho\sigma^2) \cdot \left\| \sum_{i=1}^r a_i^{\otimes 6} + \sum_{i=1}^r b_i^{\otimes 6} \right\|_F^2 \right\}.$$

The proof follows the same strategy as our proof of Lemma 2.8. We aim to lower bound first $\sum_{i,j} \langle a_i, b_j \rangle^4$ and then $\sum_{i,j} \langle a_i, b_j \rangle^6$.

*Proof.* Since $\left\| \sum_{i=1}^r a_i^{\otimes 3} \right\|_F^2 \geqslant (1 - \rho\sigma) \cdot r$ and $\mathcal{A} \vdash \left\| \sum_{i=1}^r b_i^{\otimes 3} \right\|_F^2 \geqslant (1 - \rho\sigma) \cdot r$, it holds that $\mathcal{A} \vdash \sum_{i,j} \langle a_i, b_j \rangle^3 \geqslant (1 - \rho\sigma) \cdot r$. At the same time, since $\left\| \sum_{i=1}^r a_i^{\otimes 6} \right\|_F^2 \leqslant (1 + \rho\sigma) \cdot r$ and $\mathcal{A} \vdash \left\| \sum_{i=1}^r b_i^{\otimes 6} \right\|_F^2 \leqslant (1 + \rho\sigma) \cdot r$, it suffices to show $\mathcal{A} \vdash \sum_{i,j} \langle a_i, b_j \rangle^6 \geqslant (1 - 10\rho\sigma^2) \cdot r$. Indeed,

$$\mathcal{A} \vdash \quad (1 - \rho\sigma) \cdot r \leqslant \sum_{i,j} \langle a_i, b_j \rangle^3$$
$$= \sum_j \left\langle b_j, \sum_i \langle a_i, b_j \rangle^2 a_i \right\rangle$$
$$\leqslant \sum_j \tfrac{1}{2} \|b_j\|^2 + \tfrac{1}{2} \left\| \sum_i \langle a_i, b_j \rangle^2 a_i \right\|^2$$
$$= \tfrac{1}{2} r + \tfrac{1}{2} \sum_{i,j} \langle a_i, b_j \rangle^4 + \tfrac{1}{2} \sum_j \sum_{i \neq i'} \langle a_i, a_{i'} \rangle \langle a_i, b_j \rangle^2 \langle a_{i'}, b_j \rangle^2$$
$$\leqslant \tfrac{1}{2} r + \tfrac{1}{2} \sum_{i,j} \langle a_i, b_j \rangle^4 + \tfrac{1}{2} \rho\sigma^2 r, \ .$$

where to obtain the final line we have used that $|\langle a_i, a_{i'} \rangle| \leqslant \rho$ and $\sum_{i=1}^r \langle a_i, b_j \rangle^2 = b_j^\top \left( \sum_{i=1}^r a_i a_i^\top \right) b_j \leqslant \sigma$ by the assumption that $\sum_i a_i a_i^\top \preceq \sigma \mathrm{Id}$. It follows that $\mathcal{A} \vdash \sum_{i,j} \langle a_i, b_j \rangle^4 \geqslant (1 - \rho\sigma^2 - 2\rho\sigma) \cdot r$. By applying the above reasoning to $\sum_{i,j} \langle a_i, b_j \rangle^4$ instead of $\sum_{i,j} \langle a_i, b_j \rangle^3$, we obtain $\mathcal{A} \vdash \sum_{i,j} \langle a_i, b_j \rangle^6 \geqslant (1 - 3\rho\sigma^2 - 4\rho\sigma) \cdot r \geqslant (1 - 10\rho\sigma^2) \cdot r$ as desired. Concretely,

$$\mathcal{A} \vdash \quad (1 - \rho\sigma^2 - 2\rho\sigma) \cdot r \leqslant \sum_{i,j} \langle a_i, b_j \rangle^4$$
$$= \sum_j \left\langle b_j, \sum_i \langle a_i, b_j \rangle^3 a_i \right\rangle$$
$$\leqslant \tfrac{1}{2} r + \tfrac{1}{2} \sum_{i,j} \langle a_i, b_j \rangle^6 + \tfrac{1}{2} \sum_j \sum_{i \neq i'} \langle a_i, a_{i'} \rangle \langle a_i, b_j \rangle^3 \langle a_{i'}, b_j \rangle^3$$
$$\leqslant \tfrac{1}{2} r + \tfrac{1}{2} \sum_{i,j} \langle a_i, b_j \rangle^6 + \tfrac{1}{2} \rho \sum_j \sum_{i \neq i'} \langle a_i, b_j \rangle^2 \langle a_{i'}, b_j \rangle^2$$
$$\leqslant \tfrac{1}{2}(1 + \rho\sigma^2) \cdot r + \tfrac{1}{2} \sum_{i,j} \langle a_i, b_j \rangle^6 . \quad \square$$

---

[9] A formal reason for these bounds is that the assignment $b_i = a_i$ satisfies the constraints $\mathcal{A}$.

**Random components.** Let $a_1, \ldots, a_r \in \mathbb{R}^n$ be uniformly random unit vectors with $r \leqslant n^{O(1)}$. Let $B$ be an $n$-by-$r$ matrix of variables and let $b_1, \ldots, b_r$ be the columns of $B$. With high probability, the vectors $a_1, \ldots, a_r$ satisfy $\mathcal{B}_\varepsilon$ for $\varepsilon \leqslant \tilde{O}(r/n^{1.5})$, as defined in Eq. (2.2). 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, $\left\| \sum_{i=1}^r b_i^{\otimes 3} \right\|_F^2 = r + \sum_{i \neq j} \langle b_i, b_j \rangle^3 \geqslant (1 + \tilde{O}(r/n^{1.5})) \cdot r$ and $\left\| \sum_{i=1}^r b_i^{\otimes 6} \right\|_F^2 \leqslant (1 + \tilde{O}(r/n^3)) \cdot r$.

**Lemma 2.10** (implicit in [GM15]). *Let $\varepsilon > 0$ and $a_1, \ldots, a_r \in \mathbb{R}^n$ be random unit vectors with $r \leqslant \varepsilon \cdot \tilde{\Omega}(n^{1.5})$. Let $B$ be an $n$-by-$r$ matrix of variables, $b_1, \ldots, b_r$ the columns of $B$, and $\mathcal{A}$ the following system of polynomial constraints,*

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

*Then,*

$$\mathcal{A} \left|\frac{B}{12}\right. \left\{ \left\| \sum_{i=1}^r a_i^{\otimes 6} - \sum_{i=1}^r b_i^{\otimes 6} \right\|_F^2 \leqslant O(\varepsilon) \cdot \left\| \sum_{i=1}^r a_i^{\otimes 6} + \sum_{i=1}^r b_i^{\otimes 6} \right\|_F^2 \right\} .$$

*Proof.* With high probability over the choice of $a_1, \ldots, a_n$, we have $\left\| \sum_{i=1}^r a_i^{\otimes 3} \right\|_F \geqslant (1 - \varepsilon) \cdot r$ and $\left\| \sum_{i=1}^r a_i^{\otimes 6} \right\|_F \geqslant (1 + \varepsilon) \cdot r$. Therefore, it holds $\mathcal{A} \left|- \sum_{i,j} \langle a_i, b_j \rangle^3 \geqslant 1 - \varepsilon \right.$ and it suffices to show $\mathcal{A} \left|- \sum_{i,j} \langle a_i, b_j \rangle^6 \geqslant 1 - 10\varepsilon \right.$.

The work [GM15] shows that, with high probability over the choice of $a_1, \ldots, a_n$,

$$\{\|x\|^2 = 1\} \left|- \right. \begin{cases} \sum_{i \neq j} \langle a_i, a_j \rangle \langle a_i, x \rangle^2 \langle a_j, x \rangle^2 \leqslant \varepsilon \\ \\ \sum_{i \neq j} \langle a_i, a_j \rangle \langle a_i, x \rangle^3 \langle a_j, x \rangle^3 \leqslant \varepsilon \end{cases} .$$

Under these conditions, the same reasoning as in the proof of Lemma 2.9 allows us to conclude $\mathcal{A} \left|- \sum_{i,j} \langle a_i, b_j \rangle^4 \geqslant (1 - 3\varepsilon) \cdot r \right.$ and $\mathcal{A} \left|- \sum_{i,j} \langle a_i, b_j \rangle^6 \geqslant (1 - 7\varepsilon) \cdot r \right.$. $\square$

## 2.5 Clustering

We consider the following clustering problem: given a set of points $y_1, \ldots, y_n \in \mathbb{R}^d$, 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 $S_1, \ldots, S_k$ 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 SoS allows 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 [Das99, AK01, VW04, AM05, KMV10, MV10, BS10]. 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 2.11** ([HL18, KSS18, DKS18]). *There exists an algorithm that given $k \in \mathbb{N}$ with $k \leqslant n$ and vectors $\hat{y}_1, \ldots, \hat{y}_n \in \mathbb{R}^d$ 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 $\hat{y}_1, \ldots, \hat{y}_n$ be a sample from the uniform mixture of k spherical Gaussians $\mathcal{N}(\mu_1, \mathrm{Id}), \ldots, \mathcal{N}(\mu_k, \mathrm{Id})$ with mean separation $\min_{i \neq j} \|\mu_i - \mu_j\| \geqslant O(\sqrt{\log k})$ 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 $\hat{y}_i$ and $\hat{y}_j$ 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 $O(k^{1/4})$ [VW04].

**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 the parameter and a "typical sample" $Y = y_1, \ldots, y_n$ of the mixture as the 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,

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

where $\mu_\kappa = \mathbb{E}_{i \in S_\kappa} y_i$ is the mean of cluster $S_\kappa \subseteq [n]$, and where $(1, v)$ is the vector of dimension $\dim(v) + 1$ with a 1 in the first coordinate (we extend $y_i$ and $g$ in this way so that the bound includes all moments of order at most $\ell$).

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 2.11 follows from Theorem 2.1 using the fact that under the conditions of Theorem 2.11, the following sum-of-squares proof exists with high probability for $\ell \leqslant (\log k)^{O(1)}$,

$$\mathcal{A}(\widehat{Y}) \left|\frac{X,z}{\ell} \left\{ \|X - X^*\|_{\mathrm{F}}^2 \leqslant 0.1 \cdot \|X^*\|_{\mathrm{F}}^2 \right\} \right. ,$$

where $X^*$ is the ground-truth clustering matrix (corresponding to the Gaussian components), and $\widehat{Y} = \hat{y}_1, \ldots, \hat{y}_n$ is the input samples observed.

# 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 [FS02], $k$-CLIQUE [FK00], and SPARSEST CUT [KV15], 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 [Gri01b, Gri01a] (the first two of these bounds were later independently rediscovered by Schoenebeck [Sch08]). Few other lower bounds against SoS were known. Most of the subsequent works (e.g. [Tul09, BCV+12]) built on the $k$-SAT lower bounds via reductions; 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 [MPW15, DM15, HKP+16] introduced higher-degree sum-of-squares lower bounds for $k$-CLIQUE, culminating in the work of Barak et al. [BHK+16]. Barak et al. go beyond proving lower bounds for the $k$-CLIQUE problem specifically, introducing a beautiful and general framework, called *pseudocalibration*, for proving SoS lower bounds. Though their work settles the degree of SoS refutations for $k$-CLIQUE in $\mathbb{G}(n, \frac{1}{2})$, it brings up intriguing new questions. 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 $\mathcal{P} = \{p_i(x, y) \geqslant 0\}_{i \in [m]}$ 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 $\mathcal{D}_\emptyset$ denotes the null distribution, then $\mathcal{P}_y$ is infeasible w.h.p. when $y \sim \mathcal{D}_\emptyset$, and we are interested in the minimum degree of sum-of-squares refutation.

**Pseudodensities.** By Theorem 1.4, to rule out degree-$d$ sum-of-squares refutations for $\mathcal{P}_y$, it is sufficient to construct the dual 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 related objects called *pseudodensities* rather than *pseudoexpectation functionals*. Towards defining pseudodensities, we first pick a natural background measure $\sigma$ for $x \in \mathbb{R}^n$, and we use $\mathbb{E}_x$ to denote the expectation over the background measure $\sigma$. The choice of background measure itself is not too important, but for the example we will consider, it will be convenient to pick $\sigma$ to be uniform distribution over $\{0, 1\}^n$.

**Definition 3.1.** A function $\bar{\mu} : \{0, 1\}^n \rightarrow \mathbb{R}$ is a pseudodensity for a polynomial system

$\mathcal{P} = \{p_i(x) \geqslant 0\}_{i \in [m]}$ if $\tilde{\mathbb{E}}_{\bar{\mu}} : \mathbb{R}[x]_{\leqslant d} \to \mathbb{R}$ defined as follows:

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

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

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

(scaling) $\qquad\qquad\qquad \mathbb{E}_x \bar{\mu}_y(x) = 1$ $\qquad\qquad\qquad\qquad\qquad\qquad$ (3.1)

(PSDness) $\qquad\qquad\qquad \mathbb{E}_x q(x)^2 \bar{\mu}_y(x) \geqslant 0 \qquad \forall q \in \mathbb{R}[x]_{\leqslant d/2}$ $\qquad\qquad$ (3.2)

($\mathcal{P}$ constraints) $\quad \mathbb{E}_x p(x)a^2(x) \cdot \bar{\mu}_y(x) \geqslant 0 \qquad \forall p \in \mathcal{P}, a \in \mathbb{R}[x], \deg(a^2 \cdot p) \leqslant d.$ $\quad$ (3.3)

## 3.1 Pseudocalibration

*Pseudocalibration* is a heuristic for constructing pseudodensities for non-feasible systems in such settings. It was first introduced in [BHK+16] 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. [Gri01b, Sch08]).

At a high level, pseudocalibration leverages the existence of a structured distribution of estimation problems to construct pseudodensities. For each $x \in \{0, 1\}^n$, let $\mathcal{D}_x$ be a distribution over $\{\pm 1\}^N$ such that $(x, y)$ are a feasible pair for $\mathcal{P}$.[10] Let $\mathcal{J}_*$ denote the joint *structured* distribution over feasible pairs $y^* \in \{\pm 1\}^N$ and $x^*$ sampled from $\sigma$, i.e. $\mathbb{P}_{\mathcal{J}_*}\{(x, y)\} = \sigma(x) \cdot \mathbb{P}_{\mathcal{D}_x}\{y\}$. Let us define a joint null distribution $\mathcal{J}_{\emptyset}$ on pairs $(x, y)$ to be

$$\mathcal{J}_{\emptyset} \stackrel{\text{def}}{=} \sigma \times \mathcal{D}_{\emptyset}.$$

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

$$\mu_*(x, y) = \frac{\mathbb{P}_{\mathcal{J}_*}(x, y)}{\mathbb{P}_{\mathcal{J}_{\emptyset}}(x, y)} = \frac{\mathbb{P}_{\mathcal{D}_*}\{y\}}{\mathbb{P}_{\mathcal{D}_{\emptyset}}\{y\}} \cdot \frac{\mathbb{P}_{\mathcal{J}_*}\{x|y\}}{\sigma(x)}$$

At first glance, a candidate construction of a pseudodensity $\bar{\mu}_y$ for $y \sim \mathcal{D}_{\emptyset}$ would be the partially-evaluated relative joint density $\mu_*$ namely

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

---

[10]Again, the choice $y \in \{\pm 1\}^N$ is not fundamental, and we make it for for simplicity of presentation. Also, for calculations it will sometimes be convenient to define $\mathcal{D}_x$ so that $(x, y)$ is feasible only with high probability over $y \sim \mathcal{D}_x$; however this does not greatly impact the arguments, and we neglect this detail in our exposition.

This construction $\bar{\mu}_y$ already satisfies two of the three conditions for being a pseudo-density (Eq. (3.2) and Eq. (3.3)). This is because for any polynomial $p(x, y)$,

$$\mathbb{E}_x p(x)\bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathcal{D}_*}\{y\}}{\mathbb{P}_{\mathcal{D}_0}\{y\}} \cdot \mathbb{E}_x p(x)\frac{\mathbb{P}_{\mathcal{J}_*}\{x|y\}}{\sigma(x)} = \frac{\mathbb{P}_{\mathcal{D}_*}\{y\}}{\mathbb{P}_{\mathcal{D}_0}\{y\}} \cdot \mathbb{E}_{x \sim \mathcal{J}_*(\cdot|y)} p(x) .$$

From the above equality, Eq. (3.2) follows directly because

$$\mathbb{E}_x q(x)^2 \bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathcal{D}_*}\{y\}}{\mathbb{P}_{\mathcal{D}_0}\{y\}} \cdot \mathbb{E}_{x \sim \mathcal{J}_*(\cdot|y)} q^2(x) \geqslant 0 .$$

Similarly, Eq. (3.3) is again an immediate consequence of the fact that $\mathcal{J}_*$ is supported on feasible pairs for $\mathcal{P}$,

$$\mathbb{E}_x p(x)a^2(x)\bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathcal{D}_*}\{y\}}{\mathbb{P}_{\mathcal{D}_0}\{y\}} \cdot \mathbb{E}_{x \sim \mathcal{J}_*(\cdot|y)} p(x)a^2(x) \geqslant 0 .$$

However, the scaling constraint Eq. (3.1) is far from satisfied because,

$$\mathbb{E}_x \bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathcal{D}_*}\{y\}}{\mathbb{P}_{\mathcal{D}_0}\{y\}} \cdot \mathbb{E}_{x \sim \mathcal{J}(\cdot|y)} 1 = \frac{\mathbb{P}_{\mathcal{D}_*}\{y\}}{\mathbb{P}_{\mathcal{D}_0}\{y\}}$$

is a quantity that may be really large for $y \in \mathrm{supp}(\mathcal{D}_*)$ and $0$ otherwise (recall that $\mathcal{D}_*$ has low entropy compared to $\mathcal{D}_0$). As a saving grace, the constraint Eq. (3.1) is satisfied in expectation over $y$, i.e.,

$$\mathbb{E}_{y \sim \mathcal{D}_0} \mathbb{E}_x \bar{\mu}_y(x) = \mathbb{E}_{y \sim \mathcal{D}_0} \mathbb{E}_x \mu^*(x, y) = \mathbb{E}_{(x,y) \sim \mathcal{J}_0} \mu^*(x, y) = 1 ,$$

since $\mu^*$ is a density.

The relative joint density $\mu_*(x, y)$ faces an inherent limitation in that it is only nonzero on $\mathrm{supp}(\mathcal{D}_*)$, which accounts for a negligible fraction of $y \sim \mathcal{D}_0$. 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 as $\mu_*$, but has a much higher entropy: that is, its mass is not concentrated on a small fraction of instances.

A natural way to achieve this is to simply project the joint density $\mu_*$ in to the space of low-degree polynomials. Formally, let $L_2(\mathcal{J}_0)$ denote the vector space of functions over $\mathbb{R}^N \times \mathbb{R}^n$ equipped with the inner product $\langle f, g \rangle_{\mathcal{J}_0} = \mathbb{E}_{(x,y) \sim \mathcal{J}_0} f(x, y)g(x, y)$. For $d, D \in \mathbb{N}$, let $V_{d,D} \subseteq L_2(\mathcal{J}_0)$ denote the following vector space

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

If $\Pi_{d,D}$ denotes the projection on to $V_{d,D}$, then the pseudo-calibration recipe suggests the use of the following candidate pseudodensity:

**Definition 3.2.** For $D \in \mathbb{N}$, the *D-pseudocalibrated function* $\bar{\mu}(x, y)$ is defined as

$$\bar{\mu}_y(x) = \Pi_{d,D} \circ \mu_*(x, y) \tag{3.4}$$

where $d$ is the target degree for the pseudodistribution.

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

$$
\begin{aligned}
\mathbb{E}\,\bar{\mu}_y(x)p(x, y) &= \underset{(x,y)\sim\mathcal{J}_{\emptyset}}{\mathbb{E}}\,(\Pi_{d,D} \circ \mu_*(x, y))p(x, y) \\
&= \underset{(x,y)\sim\mathcal{J}_{\emptyset}}{\mathbb{E}}\,\mu_*(x, y)p(x, y) \qquad \text{(because } p \in V_{d,D}) \\
&= \underset{(x,y)\sim\mathcal{J}_*}{\mathbb{E}}\,p(x, y) \geqslant 0\,.
\end{aligned}
$$

We additionally require that the constraints of this form are satisfied for each $y \sim \mathcal{D}_{\emptyset}$, and not just in expectation. Often, this follows using fairly straightforward arguments. In fact, for equality constraints constraints of the form $\{p(x, y) = 0\}$, one can show that the pseudocalibrated construction satisfies these constraints with high probability under very mild conditions on the joint distribution $\mathcal{J}_*$. Specifically, the following theorem holds.

**Theorem 3.3.** *Suppose $\{p(x, y) = 0\} \in \mathcal{P}$ is always satisfied for $(x, y) \sim \mathcal{J}_*$ and let $B :=$ $\max_{(x,y)\in\mathcal{J}_{\emptyset}} |p(x, y)|$ and let $D_y := \deg_y(p)$ and $d_x := \deg_x(p)$. If $d \geqslant d_x$ and $\bar{\mu}_y$ is the $D$-pseudocalibrated function defined in Eq. (3.4) then*

$$
\underset{y\sim\mathcal{D}_{\emptyset}}{\mathbb{P}}\,[|\underset{x}{\mathbb{E}}\,p(x, y)\bar{\mu}_y(x)| \geqslant \varepsilon] \leqslant \frac{B^2}{\varepsilon^2} \cdot \left\|\Pi_{d,D+2D_y} \circ \mu_* - \Pi_{d,D-1} \circ \mu_*\right\|^2_{2,\mathcal{J}_{\emptyset}}
$$

*where $\Pi_{d,D}$ for $d, D \in \mathbb{N}$ denotes the projection on to $V_{d,D}$, the 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 $\mu_*$ decays with increasing degree in $y$, then for $D$ chosen large enough, the $D$-pseudocalibrated function $\bar{\mu}_y$ satisfies the same equality constraints as those satisfied by $\mu_*$, 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 3.3 to Appendix B.

**Verifying non-negativity of squares.** The chief obstacle in establishing $\bar{\mu}(\cdot, y)$ as a valid pseudodensity is in proving that it satisfies the constraint $\mathbb{E}_x\,p(x, y)^2\bar{\mu}(x, y) \geqslant 0$, for every polynomial $p$ of degree at most $\frac{d}{2}$ in $x$. As we will see in Claim 3.4, this condition is equivalent to establishing the positive-semidefiniteness (PSDness) of the matrix

$$
M_d(y) \stackrel{\text{def}}{=} \underset{x}{\mathbb{E}}\left[\left(x^{\leqslant d/2}\right)\left(x^{\leqslant d/2}\right)^{\top} \cdot \bar{\mu}(x, y)\right], \tag{3.5}
$$

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

*Claim 3.4.* $\mathbb{E}_x\,q(x, y)^2\bar{\mu}(x, y) \geqslant 0$ for all polynomials $q(x, y)$ of degree at most $d/2$ in $x$ if and only if the matrix $M_d(y)$ is positive semidefinite.

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

$$\mathbb{E}_x q(x, y)^2 \bar{\mu}(x, y) = \mathbb{E}_x[\hat{q}(y)^\top (x^{\leqslant d/2})(x^{\leqslant d/2})^\top \hat{q}(y) \cdot \mu(x, y)] = \hat{q}(y)^\top M_d(y)\hat{q}(y) \geqslant 0,$$

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

To prove the contrapositive, we note that if $M_d(y)$ is not positive-semidefinite, then there is some negative eigenvector $v(y)$ so that $v(y)^\top M_d(y)v(y) < 0$. Taking $q(x, y) = \langle v(y), x^{\leqslant d/2} \rangle$, we have our conclusion. $\qquad\square$

Each entry of $M_d(y)$ is a degree-$D$ polynomial in $y \sim \mathcal{D}_\emptyset$. 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 $M_d(y)$. For this reason proving PSDness for $M_d(y)$ is a delicate process. Though the known lower bounds for planted clique, random SAT refutation, and other problems all use the same construction for $\bar{\mu}$, the current proofs of PSDness are very tailored to the specific choice of $\mathcal{D}_\emptyset$, and in some cases they are quite technical. We will expand further in Section 3.3.

**Pseudocalibration: a partial answer, and many questions**

While Theorem 3.3 establishes some desirable properties for the pseudocalibrated function $\bar{\mu}$, we are left with many unanswered questions. Ideally, we would be able to identify simple, general sufficient conditions on the structured distribution $\mathcal{D}_*$ and on $d$ the degree in $x$ and $D$ the degree in $y$, for which $\bar{\mu}$ yields a valid pseudodensity. The following conjecture stipulates one such choice of conditions:

**Conjecture 3.5.** *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 $\bar{\mu}(\cdot, y)$ is a valid degree-$d$ pseudodistribution which satisfies $\mathcal{P}$ with high probability over $y \sim \mathcal{D}_\emptyset$ if and only if there is no polynomial $q(y)$ of degree at most $D$ in $y$ such that $\mathbb{E}_{y \sim \mathcal{D}_\emptyset}[q(y)] = 0$ and*

$$n^{\omega(d)} \cdot \sqrt{\mathop{\mathbb{E}}_{y \sim \mathcal{D}_\emptyset}[q(y)^2]} < \mathop{\mathbb{E}}_{y \sim \mathcal{D}_*}[q(y)].$$

The upper and lower bounds on $D$ stated in Conjecture 3.5 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 [BHK+16], TENSOR PCA [HKP+17], and random $k$-SAT and $k$-XOR [Gri01b, Sch08]. 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 $\mu'(x, y)$ such that $\mu'(\cdot, y)$ is a valid pseudodensity satisfying $\mathcal{P}$ with high probability over $y \sim \mathcal{D}_\emptyset$?

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. [BHK+16]). It is a necessary condition that $\mu'$ and $\mu_*$ agree at least on the moments of $y$ which span the constraints of $\mathcal{P}$ (otherwise $\mu'$ cannot satisfy $\mathcal{P}$ in expectation). However, there are alternative ways to ensure this, while also choosing $\mu'$ to have higher entropy than $\mu_*$.

In [HKP+17], the authors give a different construction, in which rather than setting $\mu' = \Pi_{d,D}\mu_*$, they choose the function $\mu'$ which minimizes the Frobenius norm under the constraint that $\mathbb{E}_x\, x^{\leqslant d/2}(x^{\leqslant d/2})^\top \mu'(x, y)$ is positive semidefinite for every $y \in \mathrm{supp}(\mathcal{D}_\emptyset)$, and that $\Pi_{d,D}\mu'(x, y) = \Pi_{d,D}\mu_*(x, y)$. Though in [HKP+17] 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, which we discuss further in Section 4.

## 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 pseudocalibration recipe laid out in Eq. (3.4).

**The null and structured distributions.**   Recall that $\mathcal{D}_\emptyset$ is the uniform distribution over the hypercube $\{\pm 1\}^{\binom{[n]}{2}}$, corresponding to $\mathbb{G}(n, \frac{1}{2})$. For $\mathcal{J}_*$ we use the joint distribution over tuples of instance and hidden variables $(y^*, x^*)$ described in Example 1.1, with a small twist designed to ease calculations: Rather than sampling $x^*$ from $\pi$ 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 describing 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 $E^{\leqslant D}$ be the set of subsets of edges with cardinality at most $D$. Following the pseudocalibration recipe from Eq. (3.4), we project into the span of low-degree polynomials in $y$ using the monomial basis: for each $\alpha \in \mathcal{A}$ we will compute the Fourier coefficient

$$\mathbb{E}_{y \sim \mathcal{D}_\emptyset}\left[y^\alpha \cdot \tilde{\mathbb{E}}[x^A](y)\right] = \sum_{y \in \{\pm 1\}^E} y^\alpha \cdot \mathbb{P}_{\mathcal{D}_\emptyset}\{y\} \cdot \mathbb{E}_{x \sim \sigma}\, x^A \cdot \bar{\mu}_y = \mathbb{E}_{(x,y) \sim \mathcal{D}_*}[y^\alpha x^A].$$

The right-hand side can be simplified further. For $(x, y) \sim \mathcal{J}_*$, 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 $\alpha$ and $A$, which we denote $v(\alpha) \cup A$, are contained in the clique:

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

Now expressing $\tilde{\mathbb{E}}[x^A](y)$ via its Fourier decomposition, we have

$$\tilde{\mathbb{E}}(y)[x^A] = \sum_{\alpha\in E^{\leqslant D}} \left(\tfrac{2k}{n}\right)^{|v(\alpha)\cup A|} \cdot y^\alpha. \tag{3.6}$$

It is an exercise to verify that scaling (Eq. (3.1)) holds up to errors of $o(1)$ for the pseudodistribution given by these moments when $k^2 \ll n$ and $D \ll \log n$, since for such $n, k, D$, the projection $\|(\Pi_{0,D} - \Pi_{0,0})\mu_*\| \leqslant o(1)$.[11] In a similar way one can also verify that the $\mathcal{P}$ constraints (Eq. (3.3)) are satisfied (since the conditions of Theorem 3.3 are met for the natural polynomial system for clique) and that the condition of Conjecture 3.5 holds. In the following subsection, we will discuss at a high level [BHK+16]'s proof that the positive semidefiniteness constraint (Eq. (3.2)) holds.

## 3.3 Positive-semidefiniteness of matrix polynomials

To prove that the pseudocalibrated function $\bar{\mu}$ is a valid pseudodistribution, it remains to show that $\bar{\mu}$ satisfies the PSDness constraint Eq. (3.2). From Claim 3.4, we have that this is equivalent to proving that the matrix $M_d(y)$ defined in Eq. (3.5) is positive-semidefinite with high probability over $y \sim \mathcal{D}_\emptyset$. Here we discuss, at a very high level, the proof of this fact for the planted clique problem from [BHK+16].

Let $S, T \subset [n]^{\leqslant d/2}$ be multisets that index the rows and columns of $M_d$. In Eq. (3.6), we have shown that the entries of $M_d$ have the form

$$[M_d(y)]_{S,T} = \tilde{\mathbb{E}}[x^{S\cup T}] = \sum_{\alpha\in E^{\leqslant D}} \left(\frac{2k}{n}\right)^{|v(\alpha)\cup S\cup T|} \cdot y^\alpha,$$

where $E^{\leqslant D}$ is the set of all subsets of edge variables with cardinality at most $D$. Each entry of $M_d$ is a degree-$D$ polynomial in the random variable $y$, and because $M_d$ does not correspond in a natural way to a sum of independent random matrices, we cannot apply black-box matrix concentration results to $M_d$.

Since $\alpha \in E^{\leqslant D}$ corresponds to a subset of edge variables, it is natural to associate with each $\alpha, S, T$ a colored subgraph or *shape* $\sigma$ on $|v(\alpha) \cup S \cup T|$ vertices. The shape $\sigma$ is a graph $H_\sigma$ with vertex set isomorphic to $v(\alpha) \cup S \cup T$, and edge set isomorphic to $\alpha$. Further, vertices isomorphic to $S$ are assigned colors $L = \ell_1, \ldots, \ell_{|S|}$, and vertices isomorphic to $T$ are assigned colors $R = r_1, \ldots, r_{|T|}$ (note that a single vertex may receive more than one color). For example, if $S = \{a, b, k\}$, $T = \{i, j, k\}$, and $\alpha = \{(a, i), (b, u), (u, j)\}$, we would have the corresponding shape

---

[11]This is equivalent to checking that the variance of Eq. (3.6) is $o(1)$ for $A = \emptyset$.
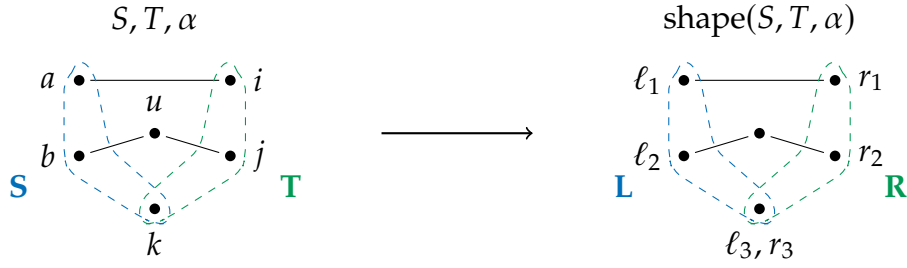
$$S, T, \alpha \qquad\qquad \text{shape}(S, T, \alpha)$$

For each such shape $\sigma$, we define the $n^{\leqslant d} \times n^{\leqslant d}$ $\sigma$-matrix polynomial $M_\sigma(y)$, so that for $S, T \in [n]^{\leqslant d}$,

$$[M_\sigma(y)]_{S,T} = \sum_{\alpha \in E^{\leqslant D}} \mathbf{1}[\text{shape}(S, T, \alpha) = \sigma] \cdot y^\alpha. \tag{3.7}$$

Or alternatively for $S, T, L \subset [n]$, let $H_\sigma(S, T, U)$ be the labeled copy of $H_\sigma$ in which $R$ is labeled with $S$, $L$ with $T$, and the remaining vertices with $U$, and every vertex of $H_\sigma$ receives a unique, single label. Then Eq. (3.7) is equivalent to summing over products of edges for all valid labelings of the uncolored vertices of $H_\sigma$:

$$[M_\sigma(y)]_{S,T} = \sum_{U \in \binom{[n] \setminus (S \cup T)}{|V(H) \setminus (L \cup R)|}} \mathbf{1}[H_\sigma(S, T, U) \text{ valid}] \cdot y^{E(H_\sigma(S,T,U))}.$$

The matrices $\{M_\sigma\}$ form a natural basis for expressing $M_d(y)$:

$$M_d(y) = \sum_\sigma \left(\frac{2k}{n}\right)^{|v(\sigma)|} \cdot M_\sigma(y).$$

In [BHK$^+$16], the authors characterize the spectrum of $M_\sigma(y)$. Incredibly, the spectral properties of the $\sigma$-matrix polynomials determined by the connectivity of $H_\sigma$.

**Theorem 3.6** ([BHK$^+$16, MP16]). *Suppose that $H_\sigma$ has $t = O(\log n)$ vertices, and that $H_\sigma$ has exactly $p$ vertex-disjoint paths from $L \setminus R$ to $R \setminus L$, and that $|R \cap L| = c$. Then with high probability over $y \sim \{\pm 1\}^N$,*

$$\left\|M_\sigma(y)\right\| \leqslant 2^{O(t)} (\log n)^{O(t+p-c)} \cdot n^{\frac{t-p-c}{2}}.$$

In order to characterize the spectrum of $M_d(y)$, it does not suffice to understand the spectrum of each $M_\sigma$ individually; one must account for the interactions of the spectra of the $M_\sigma$. This is challenging because different shapes exhibit very different spectral characteristics. For example, for $\sigma_1$ the shape given by two horizontal parallel paths of length 1, the matrix $M_{\sigma_1}$ has spectral norm of magnitude $\tilde{O}(n)$.



$$\sigma_1 \qquad\qquad\qquad \sigma_2$$

On the other hand, for $\sigma_2$ the shape given by two vertical parallel paths of length 1, the matrix $M_{\sigma_2}$ has spectral norm $\Omega(n^2)$.[12]

To prove that $M_d(y)$ is indeed PSD with high probability, [BHK$^+$16] employ a delicate iterative charging scheme, partitioning the $\sigma$ into groups according to the number of disjoint paths from $L \setminus R$ to $R \setminus L$ and charging the spectral norm of each such group to the positive-semidefinite shapes. We will not give further details, and refer the interested reader to [BHK$^+$16].

**Symmetric matrix polynomials.** The matrix $M_d(y)$ is not completely devoid of structure: it is what we call a *symmetric matrix polynomial*, since the matrix remains fixed under permutations of the elements of $[n]$. Specifically, if we let $\pi \in \mathcal{S}_n$, and if $\pi(M)$ and $\pi(y)$ are the natural action induced by permutations of the vertices of $G$ on the rows/columns of $M$ and edges in $y$, then

$$\pi(M_\sigma(\pi(y))) = M_\sigma(y).$$

For a familiar example of a symmetric matrix polynomial, consider the adjacency matrix of $G$ with entries expressed as degree-1 polynomials of $y$.

One compelling question is whether the symmetry of $M_d$ can be useful in characterizing its spectrum.

**Question 3.7.** Suppose that $A(y)$ is a symmetric matrix polynomial. What are sufficient conditions on $A$ such that $A(y) \succeq 0$ with high probability over $y \sim \mathcal{D}_\emptyset$?

For some classes of symmetric matrices, such as association scheme matrices, the above question is fully answered (see e.g. [GS]). There is hope that if this question is answered in greater generality, it will lead to a "book proof" of the pseudocalibration method.

# 4 Connection to spectral algorithms

Sum-of-squares SDPs yield a systematic framework that captures and generalizes a loosely defined class of algorithms often referred to as *spectral algorithms*. We say that an algorithm is a "spectral algorithm" if on input $y$ the algorithm constructs a matrix $M(y)$ that can be easily computed from $y$, whose eigenvalues or eigenvectors manifestly yield a solution to the problem at hand.[13] We will give a more concrete definition for the notion of a spectral algorithm a little later in this section.

Although spectral algorithms are typically subsumed by sum-of-squares SDPs, they tend to be simpler to implement and more efficient. Furthermore, in many cases such

---

[12]Another way to see that this is true without Theorem 3.6 is that if $A$ is the signed adjacency matrix of the random graph, then $\|A\|_F^2 = n(n-1)$ and with high probability $\|A\| = n$, and (excluding entries corresponding to $S, T$ with nontrivial intersection) $M_{\sigma_1} = A \otimes A$, while $M_{\sigma_2} = v(A)v(A)^\top$ where $v(A)$ is the $n^2 \times 1$ reshaping of $A$.

[13]In other contexts, "spectral algorithms" may sometimes describe algorithms that also modify $M(y)$ as the algorithm proceeds; for simplicity and because our main result will be an *equivalence* between SoS and spectral algorithms, we consider only this narrower class.

as $k$-CLIQUE [AKS98b] and tensor decomposition [Har70], 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 (corresponding to an SoS SDP of degree 2) yields a 0.878 approximation for MAX CUT [GW95], and has no known analogues among spectral algorithms. Contrary to this, in many random settings, the best known SoS algorithms yield guarantees that are no better than the corresponding spectral algorithms. Recent work explains this phenomenon by showing an equivalence between spectral algorithms and their sum-of-squares counterparts for a broad family of problems [HKP+17].

## 4.1  Spectral algorithms and sum-of-squares proofs

Before formally stating the equivalence of SoS and spectral algorithms from [HKP+17], we first demonstrate that SoS often captures spectral algorithms. Let us begin by considering a classic example of a simple spectral algorithm for the $k$-CLIQUE problem.

**Spectral algorithms for $k$-CLIQUE.**   In a graph $G = (V, E)$ with adjacency matrix $A_G$, if a subset $S \subset V$ of $k$ vertices forms a clique then,

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

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 left-hand side by

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

The eigenvalue of this matrix thus certifies 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 $\mathbb{G}(n, \tfrac{1}{2})$, the matrix $A_G - \tfrac{1}{2}J$ is a random symmetric matrix whose off-diagonal entries are i.i.d uniform over $\{\pm\tfrac{1}{2}\}$. By a classical result in random matrix theory (Bai-Yin's law, see e.g. [Tao12]), we will have that $\lambda_{\max}\left(A_G - \tfrac{1}{2}J\right) = O(\sqrt{n})$ 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 $\mathbb{G}(n, \tfrac{1}{2})$ by computing the largest eigenvalue of the associated matrix valued function $P(G) = A_G - \tfrac{1}{2}J$.

This algorithm also gives a degree-2 sum-of-squares proof; if $x \in \{0, 1\}^n$ are the indicator variables for vertex membership in the clique polynomial system $\mathcal{A}$ (as in Example 1.1), then we have that the clique size $k = \sum_i x_i$, and

$$\mathcal{A} \vdash \left(\textstyle\sum_i x_i\right)^2 = \sum_{i,j} x_i x_j - 2x_i x_j \cdot \mathbf{1}[(i, j) \notin E(G)]$$

$$= x^\top (2A_G + 2\mathrm{Id} - J)x$$
$$= x^\top \left(2 \cdot \lambda_{\max}(A_G + \mathrm{Id} - \tfrac{1}{2}J) \cdot \mathrm{Id}\right) x + \textstyle\sum_j s_j(x)^2$$
$$= \|x\|^2 \cdot (2 \cdot \lambda_{\max}(A_G - \tfrac{1}{2}J) + 2) + \textstyle\sum_j s_j(x)^2$$
$$= (\textstyle\sum_i x_i) \cdot (2 \cdot \lambda_{\max}(A_G - \tfrac{1}{2}J) + 2) + \textstyle\sum_j s_j(x)^2,$$

where in the first line we have used that $\{x_i x_j = 0\}_{(i,j)\notin E(G)} \in \mathcal{A}$, to derive the third line we have used that for a symmetric matrix $M$, $\lambda_{\max}(M) \cdot \mathrm{Id} - M$ is positive-semidefinite and therefore its eigendecomposition gives that $x^\top (\lambda_{\max}(M) \cdot \mathrm{Id} - M) x$ is a sum-of-squares. To derive the last line we have used that $\{x_i^2 = x_i\}_{i\in[n]} \in \mathcal{A}$.

**Spectral algorithms for injective tensor norm.** Now we will see another example of a more complex spectral algorithm which is captured by sum-of-squares. Recall that the injective tensor norm (see Example 1.2) of a symmetric 4-tensor $\mathbf{T} \in \mathbb{R}^{[n]\times[n]\times[n]\times[n]}$ is given by $\max_{\|x\|\leqslant 1}\langle x^{\otimes 4}, \mathbf{T}\rangle$. The injective tensor norm $\|\mathbf{T}\|_{\mathrm{inj}}$ is computationally intractable in the worst case [HL13]. We will now describe a sequence of spectral algorithms that certify tighter bounds for the injective tensor norm of a tensor $\mathbf{T}$ drawn from the null distribution (with entries drawn i.i.d from $\mathcal{N}(0,1)$).

Let $T = T_{\{1,2\},\{3,4\}}$ denote the $n^2 \times n^2$ matrix obtained by reshaping the tensor $\mathbf{T}$. Then,

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

Thus $\lambda_{\max}(T)$ is a spectral upper bound on $\|\mathbf{T}\|_{\mathrm{inj}}$. Since each entry of $\mathbf{T}$ is drawn independently from $\mathcal{N}(0,1)$, we have from the Bai-Yin law that $\lambda_{\max}(T) \leqslant O(n)$ with high probability [Tao12]. We also recall that the injective norm of a random $\mathcal{N}(0,1)$ tensor $T$ is at most $O(\sqrt{n})$ with high probability [ABAČ, MR14] . Taking these facts together, $\lambda_{\max}(T)$ certifies an upper bound that is $O(\sqrt{n})$-factor approximation to $\|\mathbf{T}\|_{\mathrm{inj}}$.

We will now describe a sequence of improved approximations to the injective tensor norm via spectral methods, which also yield an analysis of the SoS SDP for TENSOR PCA. Fix a positive integer $k \in \mathbb{N}$. The polynomial $\mathbf{T}(x) = \langle x^{\otimes 4}, \mathbf{T}\rangle$ can be written as,

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

The tensored vector $x^{\otimes 2k}$ is symmetric, and is invariant under permutations of its modes. Let $\Sigma_{2k}$ 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 $\Pi$ to the modes of $A$. By averaging over all permutations $\Pi, \Pi' \in \Sigma_{2k}$, we can write

$$\mathbf{T}(x) = \left(\mathop{\mathbb{E}}_{\Pi,\Pi'\in\Sigma_{2k}} \langle \Pi \circ x^{\otimes 2k}, T^{\otimes k}(\Pi' \circ x^{\otimes 2k})\rangle\right)^{1/k}$$

$$= \left(\left\langle x^{\otimes 2k}, \left(\mathop{\mathbb{E}}_{\Pi,\Pi'\in\Sigma_{2k}} \Pi \circ T^{\otimes k} \circ \Pi'\right) x^{\otimes 2k}\right\rangle\right)^{1/k}$$

$$\leqslant \lambda_{\max}\left(\mathop{\mathbb{E}}_{\Pi,\Pi'\in\Sigma_{2k}} \Pi \circ T^{\otimes k} \circ \Pi'\right)^{1/k} \cdot \|x\|_2^4. \tag{4.1}$$

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

$$P_k(\mathbf{T}) \overset{\text{def}}{=} \underset{\Pi, \Pi' \in \Sigma_{2k}}{\mathbb{E}} \Pi \circ T^{\otimes k} \circ \Pi'$$

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

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

$$P_2(\mathbf{T})_{ijk\ell, abcd} = \frac{1}{(4!)^2} \cdot \left( T_{ijab} \cdot T_{k\ell cd} + T_{ijac} \cdot T_{k\ell bd} + T_{ijad} \cdot T_{k\ell bc} + \cdots \right),$$

where we sum over the $(4!)^2$ pairs of permutations of $i, j, k, \ell$ and $a, b, c, d$. Thus a typical entry of $P_k(\mathbf{T})$ with no repeated indices is an average of a super-exponentially large (in $k$) number of i.i.d. random variables. We will call this number $N_k$ for convenience.

When $k \ll \sqrt{n}$, a typical entry of $P_k(\mathbf{T})$ contains no repeated indices, and this implies that the variance of a typical entry of $P_k(\mathbf{T})$ is equal to $\frac{1}{N_k}$. For the moment, let us assume that the spectrum of $P_k(\mathbf{T})$ 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, $\lambda_{\max}(P_k(\mathbf{T})) \leqslant O(n^k \cdot \frac{1}{N_k^{1/2}})$ with high probability, certifying that $\|\mathbf{T}\|_{\text{inj}} \leqslant \frac{n}{N_k^{1/2k}}$. On accounting for the symmetries of $\mathbf{T}$, it is not difficult to see that $N_k = k! \left( \frac{1}{2^k} \frac{2k!}{k!} \right)^2 \gg (k!)^3$. Consequently, as per this heuristic argument, $\lambda_{\max}(P_k(\mathbf{T}))$ would certify an upper bound of $\|\mathbf{T}\|_{\text{inj}} \leqslant O(\frac{n}{k^{3/2}})$.

Unfortunately, the entries of $P_k(\mathbf{T})$ are not independent random variables and not all entries of $P_k(T)$ are typical as described above. Although the heuristic bound on $\lambda_{\max}(P_k(\mathbf{T}))$ is not quite accurate, a careful analysis via the trace method shows that the upper bound $\lambda_{\max}(P_k(\mathbf{T}))^{1/k}$ decreases polynomially in $k$ [BGL17, RRS17].

**Theorem 4.1.** *[BGL17] For $4 \leqslant k \leqslant n$, if $\mathbf{T}$ is a symmetric 4-tensor with i.i.d. entries from a subgaussian measure then*

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

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

Thus the matrix polynomial $P_k(\mathbf{T})$ 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.

**Spectral algorithms from Sum-of-Squares analyses.** In fact, this spectral algorithm was discovered in the context of analyzing SoS refutation algorithms, and the upper bound certificate produced by the above spectral algorithm can again be cast as a degree $4k$ sum-of-squares proof. In particular, if $\lambda_{\max}(P_k(\mathbf{T})) \leqslant \tau$ for some tensor $\mathbf{T}$ and $\tau \in \mathbb{R}$ then,

$$\tau - \mathbf{T}(x)^k = \tau \|x\|_2^{4k} - \langle x^{\otimes 2k}, P_k(\mathbf{T}) x^{\otimes 2k} \rangle + \tau(1 - \|x\|_2^{4k})$$
$$= \langle x^{\otimes 2k}, (\tau \cdot \text{Id} - P_k(\mathbf{T})) x^{\otimes 2k} \rangle + \tau(1 - \|x\|_2^{4k})$$

$$= \langle x^{\otimes 2k}, (\tau \cdot \mathrm{Id} - P_k(\mathbf{T}))x^{\otimes 2k} \rangle + (1 - \|x\|_2^2)\left( \tau \cdot \sum_{i=0}^{2k-1} \|x\|_2^{2i} \right)$$

$$= \sum_j s_j^2(x) + (1 - \|x\|_2^2)\left( \tau \cdot \sum_{i=0}^{2k-1} \|x\|_2^{2i} \right),$$

The final step in the calculation again uses the fact that if a matrix $M \succeq 0$, then the polynomial $\langle x^{\otimes 2k}, Mx^{\otimes 2k} \rangle$ is a sum-of-squares $\sum_j s_j^2(x)$. Therefore, the degree-$4k$ sum-of-squares obtains an approximation guarantee that is no worse than the somewhat ad hoc 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 ad hoc algorithms. It also exemplifies the trend of SoS inspiring new spectral algorithms which take advantage of polynomial identities and problem symmetry to improve upon simpler spectral algorithms (see also [HSS15, AOW15, BGL17, RRS17]). There is a line of work in which the spectral certificates used by SoS for refutation are modified and compressed to give efficient, lightweight spectral algorithms that run in subquadratic or near-linear time; we refer the interested reader to [HSSS16, MS16, SS17].

**Refuting Random CSPs.** The basic scheme used to upper bound the injective tensor norm (see Eq. (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 [Fei02, BB02, DLS14, BKS13, CLP02]. 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 homogeneous degree-4 linear system of $m$ equations over $n$ variables $\{X_1, \dots, X_n\}$ in $\mathbb{F}_2$. A random 4-XOR instance is one where each equation is sampled uniformly at random (avoiding repetition). For $m \gg n$, with high probability over the choice of the constraints, no assignment satisfies more than a $\frac{1}{2} + o(1)$ fraction of constraints.

To formulate a polynomial system, we will use the natural $\{\pm 1\}$-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 $\{\mathbf{T}_{ijk\ell}\}_{i,j,k,\ell \in \binom{[n]}{4}}$, with $\mathbf{T}_{ijk\ell} = \pm 1$ if we have the equation $x_i x_j x_k x_\ell = \pm 1$, and $\mathbf{T}_{ijk} = 0$ otherwise. To certify that no assignment satisfies more than $\varepsilon m$ constraints, we will need to refute the following polynomial system.

$$\left\{x_i^2 - 1\right\}_{i \in [n]} \qquad \text{and} \qquad \left\{\langle \mathbf{T}, x^{\otimes 4} \rangle \geqslant \varepsilon \cdot m \right\} \tag{4.2}$$

A refutation for this system with $\varepsilon < 1 - \eta$ for $\eta$ independent of $m, n$ is called a *strong refutation*. 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 $\mathbf{T}$ 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 Eq. (4.1) can still be utilized to obtain the following strong refutation algorithm.

**Theorem 4.3.** *[RRS17] For all $\delta \in [0, 1)$, the degree $n^\delta$ sum-of-squares SDP can strongly 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 [RRS17]. Moreover, these bounds on the degree of sum-of-squares refutations tightly match corresponding lower bounds for CSPs shown in [KMOW17, BCK15].

## 4.2 Equivalence of spectral algorithms and sum-of-squares refutations

The algorithms described above will serve as blueprints for a class of spectral algorithms that will characterize the power of SoS SDPs.

**Defining spectral algorithms.** Here, we will consider spectral algorithms for distinguishing problems. Recall that in a distinguishing problem, the input consists of a sample $y$ drawn from one of two distributions, say a structured distribution $\mathcal{D}_*$ or a null distribution $\mathcal{D}_\emptyset$, and the algorithm's goal is to identify the distribution the sample is drawn from. We think of samples from the structured distribution $\mathcal{D}_*$ as having an underlying hidden structure, while samples from the null distribution $\mathcal{D}_\emptyset$ typically do not.

A *spectral* algorithm $\mathcal{A}$ for the distinguishing problem proceeds as follows. Given an instance $y$, the algorithm $\mathcal{A}$ computes a matrix $P(y)$ whose entries are given by low-degree polynomials in $y$, such that $\lambda_{\max}(P(y))$ indicates whether $y \sim \mathcal{D}_*$ or $y \sim \mathcal{D}_\emptyset$.

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

$$\mathop{\mathbb{E}}_{y \sim \mathcal{D}_*} \lambda_{\max}^+(P(y)) \gg \mathop{\mathbb{E}}_{y \sim \mathcal{D}_\emptyset} \lambda_{\max}^+(P(y))$$

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 SDP in a wide variety of settings.

Spectral algorithms as defined in Theorem 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. On the other hand, 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 spectral algorithms are no weaker than sum-of-squares SDPs for a wide variety of estimation problems.

**Robust inference.** Many estimation problems share the "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 $\mathcal{D}_*$ for the $k$-CLIQUE problem. A graph $G \sim \mathcal{D}_*$ consists of a $k$-clique embedded in to an Erdős-Rényi 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 the *robust inference* property, where a random subsample $G'$ can reveal non-trivial structure in the instance.

Though the subsample does not determine the planted clique in $G$, the information revealed is substantial. For example, as long as $\delta \cdot k \gg 2\log n$, observing $G'$ allows us to distinguish whether $G$ is sampled from the structured distribution $\mathcal{D}_*$ or the null distribution $\mathcal{D}_\emptyset$. 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 $x \in \mathbb{R}^n$. We define the *subsampling distribution* $\Upsilon$ to be 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 $\mathcal{I}$ denote the collection of all instances, and $\mathcal{I}_\downarrow$ denote the collection of all sub-instances.

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

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

The 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 [HKP+17] for a detailed discussion). The existence of the robust inference property has a dramatic implication for the power of low-degree sum-of-squares SDPs: they are no more powerful than spectral algorithms. This assertion is formalized in the following theorem.

**Theorem 4.6** ([HKP+17]). *Suppose $\mathcal{P} = \{p_i(x, y) \geqslant 0\}_{i \in [m]}$ is a polynomial system with degree $d_x$ and $d_y$ over $x$ and $y$ respectively. Fix $B \geqslant d_x \cdot d_y \in \mathbb{N}$. If the degree-d sum-of-squares SDP relaxation can be used to distinguish between the structured distribution $\mathcal{D}_*$ and the null distribution $\mathcal{D}_\emptyset$, namely,*

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

- *For $y \sim \mathcal{D}_\emptyset$, 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 $Q : \mathcal{I} \to \mathbb{R}^{[n]^{\leqslant d} \times [n]^{\leqslant d}}$ such that,*

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

*where $D \in \mathbb{N}$ is 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 the typical parameter settings of random CSPs, community detection and densest subgraph we have $D = O(d \log n)$ (see [HKP+17] 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.[14] From a theoretical standpoint, it reduces the task of showing lower bounds against the complicated sum-of-squares SDP to that of understanding the spectrum of low-degree matrix polynomials over the two distributions.

**Future directions.** 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 suggested 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 $(\alpha_k \pm o(1)) \cdot n$ with high probability. For every $\beta \in (0, 1)$, the existence of a size $\beta \cdot n$-independent set can

---

[14]We comment however that the matrix polynomial $Q(y)$ is non-constructive and non-uniform, and arises as the dual object of a exponentially-sized convex program. For this reason, the theorem does not automatically give efficient spectral algorithms matching the guarantees of SoS.

be formulated as the following polynomial system.

$$
\mathcal{P}_\beta(G) : \left\{ \{x_i(1 - x_i) = 0\}_{i \in [n]}, \qquad \{x_i x_j = 0\}_{(i,j) \in E(G)}, \qquad \sum_{i \in [n]} x_i \geqslant \beta \cdot n . \right\}
$$

For each degree-$d \in \mathbb{N}$ define the degree-$d$ SoS SDP refutation threshold to be

$$
\alpha_k^{(d)} \stackrel{\text{def}}{=} \text{ smallest } \beta \text{ such that } \mathop{\mathbb{P}}_{G \sim \mathbb{G}([n], k/n)} \left[ \mathcal{P}_\beta(G) \big|\frac{x}{d} \perp \right] = 1 - o_n(1)
$$

It is natural to ask if the degree-$d$ sum-of-squares SDP refutation threshold steadily improves with $k$.

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

A natural structured distribution $\mathcal{D}_\beta$ for the problem is the following: For each subset $S \in \binom{[n]}{\beta \cdot n}$, define $\mu_S$ as $\mathbb{G}(n, k/n)$ conditioned on $S$ being an independent set. For $D \in \mathbb{N}$ let $\gamma_k^{(D)} \in (0, 1)$ be the largest value of $\beta$ for which distribution of eigenvalues of every degree-$D$ matrix polynomial in the structured distribution $\mathcal{D}_\beta$ and null distribution $\mathcal{D}_\emptyset$ converge to each other in distribution. In other words, $\gamma_k^{(D)}$ is the precise threshold of independent set size $\beta$ below which the spectrum of degree-$D$ matrix polynomials fails to distinguish the structured and null distributions. It is natural to conjecture that if the empirical distribution of eigenvalues looks 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 there a universal constant $C > 0$ such that $\alpha_k^{(d)} \geqslant \gamma_k^{(C \cdot d)}$?

This question strengthens Theorem 4.6 in that we ask for the degree $d$ of the SoS refutation to differ from the degree of the spectral algorithm by only a *constant* factor; in Theorem 4.6, the degrees differ by a factor that depends on the robustness of the polynomial system, which may grow with $n$. On the other hand, this question differs from Theorem 4.6 because if we ask for the spectra of matrices from the structured and null distributions to converge in distribution, we are working with a different class of spectral algorithms. Depending on the notion of convergence, we may not be able to reason about the value of the maximum positive eigenvalue, or other non-smooth tests. Question 4.8 and its variants are an intriguing direction for future research.

# 5 Concluding remarks

We have now seen how the sum-of-squares algorithm may be used as a tool for solving estimation problems, via low-degree SoS proofs of identifiability of parameters from measurements. This proofs-to-algorithms perspective has unified and simplified previous algorithmic results, as well as lead to stronger novel ones.

On the other hand, we have surveyed recent progress towards characterizing the limitations of SoS algorithms for estimation problems in the regime where the estimation problem is information-theoretically solvable (but perhaps not computationally tractable). The *pseudocalibration* heuristic of [BHK+16] suggests exposing the limitations of SoS by comparing *structured* and *null* distributions over measurements; when SoS fails to distinguish these distributions, the SoS algorithm fails to solve the estimation problem. But many questions remain: which properties of the structured distribution dictate whether low-degree SoS proofs of identifiability (and therefore algorithms) exist? How does the degree of SoS proofs scale with the amount of information in the measurement?

Remarkably, all current evidence is consistent with the conjecture that low-degree sum-of-squares proofs are only as powerful as low-degree polynomial tests for a broad family of estimation problems (Conjecture 3.5). Affirming this conjecture will establish a beautiful theory of the power of semidefinite programs, and bring new insight to the study of information-computation gaps. Refuting this conjecture may lead to exciting algorithmic discoveries, and a fine-grained understanding of the difficulty of estimation problems.

# References

[ABAČ]     Antonio Auffinger, Gérard Ben Arous, and Jiří Čeraný, *Random matrices and complexity of spin glasses*, Communications on Pure and Applied Mathematics **66**, no. 2, 165–201. 4, 30

[AFH+12]   Anima Anandkumar, Dean P. Foster, Daniel J. Hsu, Sham Kakade, and Yi-Kai Liu, *A spectral algorithm for latent dirichlet allocation*, NIPS, 2012, pp. 926–934. 13

[AGMR17]   Sanjeev Arora, Rong Ge, Tengyu Ma, and Andrej Risteski, *Provable learning of noisy-or networks*, Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017 (Hamed Hatami, Pierre McKenzie, and Valerie King, eds.), ACM, 2017, pp. 1057–1066. 13

[AK01]     Sanjeev Arora and Ravi Kannan, *Learning mixtures of arbitrary gaussians*, STOC, ACM, 2001, pp. 247–257. 18

[AKS98a]   Noga Alon, Michael Krivelevich, and Benny Sudakov, *Finding a large hidden clique in a random graph*, Random Struct. Algorithms **13** (1998), no. 3-4, 457–466. 2

[AKS98b]   Noga Alon, Michael Krivelevich, and Benny Sudakov, *Finding a large hidden clique in a random graph*, Proceedings of the Eighth International Conference "Random Structures and Algorithms" (Poznan, 1997), vol. 13, 1998, pp. 457–466. MR 1662795 29

[AM05]     Dimitris Achlioptas and Frank McSherry, *On spectral learning of mixtures of distributions*, COLT, Lecture Notes in Computer Science, vol. 3559, Springer, 2005, pp. 458–469. 18

[AOW15]    Sarah R. Allen, Ryan O'Donnell, and David Witmer, *How to refute a random CSP*, FOCS, IEEE Computer Society, 2015, pp. 689–708. 32

[BB02]     Eli Ben-Sasson and Yonatan Bilu, *A gap in average proof complexity*, Electronic Colloquium on Computational Complexity (ECCC) (2002), no. 003. 32

[BBH⁺12]   Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou, *Hypercontractivity, sum-of-squares proofs, and their applications*, STOC, ACM, 2012, pp. 307–326. 4, 7

[BCK15]    Boaz Barak, Siu On Chan, and Pravesh K. Kothari, *Sum of squares lower bounds from pairwise independence [extended abstract]*, STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing, ACM, New York, 2015, pp. 97–106. MR 3388187 33

[BCMV14]   Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan, *Smoothed analysis of tensor decompositions*, STOC, ACM, 2014, pp. 594–603. 13

[BCV⁺12]   Aditya Bhaskara, Moses Charikar, Aravindan Vijayaraghavan, Venkatesan Guruswami, and Yuan Zhou, *Polynomial integrality gaps for strong SDP relaxations of densest k-subgraph*, SODA, SIAM, 2012, pp. 388–405. 20

[BGL17]    Vijay Bhattiprolu, Venkatesan Guruswami, and Euiwoong Lee, *Sum-of-squares certificates for maxima of random tensors on the sphere*, APPROX-RANDOM, LIPIcs, vol. 81, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2017, pp. 31:1–31:20. 4, 31, 32

[BHK⁺16]   Boaz Barak, Samuel B. Hopkins, Jonathan A. Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin, *A nearly tight sum-of-squares lower bound for the planted clique problem*, FOCS, IEEE Computer Society, 2016, pp. 428–437. 2, 6, 20, 21, 24, 25, 26, 27, 28, 37

[BKS13]    Boaz Barak, Guy Kindler, and David Steurer, *On the optimality of semidefinite relaxations for average-case and generalized constraint satisfaction*, ITCS, ACM, 2013, pp. 197–214. 32

[BKS14]    Boaz Barak, Jonathan A. Kelner, and David Steurer, *Rounding sum-of-squares relaxations*, STOC, ACM, 2014, pp. 31–40. 9

[BKS15]    _____, *Dictionary learning and tensor decomposition via the sum-of-squares method*, STOC, ACM, 2015, pp. 143–151. 9, 13, 14

[BM16]     Boaz Barak and Ankur Moitra, *Noisy tensor completion via the sum-of-squares hierarchy*, COLT, JMLR Workshop and Conference Proceedings, vol. 49, JMLR.org, 2016, pp. 417–445. 9, 12

[BS10]     Mikhail Belkin and Kaushik Sinha, *Toward learning gaussian mixtures with arbitrary separation*, COLT, Omnipress, 2010, pp. 407–419. 18

[BS14]     Boaz Barak and David Steurer, *Sum-of-squares proofs and the quest toward optimal algorithms*, Electronic Colloquium on Computational Complexity (ECCC) **21** (2014), 59. 5

[Che15]    Yudong Chen, *Incoherence-optimal matrix completion*, IEEE Trans. Information Theory **61** (2015), no. 5, 2909–2923. 10, 11

[CLP02]    Andrea Crisanti, Luca Leuzzi, and Giorgio Parisi, *The 3-SAT problem with large number of clauses in the ∞-replica symmetry breaking scheme*, Journal of Physics A: Mathematical and General **35** (2002), no. 3, 481. 32

[CO12]     Luca Chiantini and Giorgio Ottaviani, *On generic identifiability of 3-tensors of small rank*, SIAM J. Matrix Analysis Applications **33** (2012), no. 3, 1018–1037. 13

[CR09]     Emmanuel J. Candès and Benjamin Recht, *Exact matrix completion via convex optimization*, Foundations of Computational Mathematics **9** (2009), no. 6, 717–772. 10, 11

[Das99]    Sanjoy Dasgupta, *Learning mixtures of gaussians*, FOCS, IEEE Computer Society, 1999, pp. 634–644. 18

[DKS18]    Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart, *List-decodable robust mean estimation and learning mixtures of spherical gaussians mixture models, robustness, and sum of squares proofs*, STOC, ACM, 2018, p. (to appear). 19

[DLS14]    Amit Daniely, Nati Linial, and Shai Shalev-Shwartz, *From average case complexity to improper learning complexity*, STOC, ACM, 2014, pp. 441–448. 32

[DM15]     Yash Deshpande and Andrea Montanari, *Improved sum-of-squares lower bounds for hidden clique and hidden submatrix problems*, COLT, JMLR Workshop and Conference Proceedings, vol. 40, JMLR.org, 2015, pp. 523–562. 20

[Fei02]    Uriel Feige, *Relations between average case complexity and approximation complexity*, Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing, ACM, New York, 2002, pp. 534–543. MR 2121179 32

[FGR⁺17]   Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Srinivas Vempala, and Ying Xiao, *Statistical algorithms and a lower bound for detecting planted cliques*, J. ACM **64** (2017), no. 2, 8:1–8:37. 2

[FK00]    Uriel Feige and Robert Krauthgamer, *Finding and certifying a large hidden clique in a semirandom graph*, Random Structures Algorithms **16** (2000), no. 2, 195–208. MR 1742351 20

[FK03]    _____ , *The probable value of the Lovász-Schrijver relaxations for maximum independent set*, SIAM J. Comput. **32** (2003), no. 2, 345–370. MR 1969394 2

[FS02]    Uriel Feige and Gideon Schechtman, *On the optimality of the random hyperplane rounding technique for MAX CUT*, Random Structures Algorithms **20** (2002), no. 3, 403–440, Probabilistic methods in combinatorial optimization. MR 1900615 20

[Gha10]   Sevag Gharibian, *Strong np-hardness of the quantum separability problem*, Quantum Information & Computation **10** (2010), no. 3, 343–360. 4

[GM75]    G. R. Grimmett and C. J. H. McDiarmid, *On colouring random graphs*, Mathematical Proceedings of the Cambridge Philosophical Society **77** (1975), no. 2, 313–324. 2

[GM15]    Rong Ge and Tengyu Ma, *Decomposing overcomplete 3rd order tensors using sum-of-squares algorithms*, APPROX-RANDOM, LIPIcs, vol. 40, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2015, pp. 829–849. 14, 18

[Gri01a]   Dima Grigoriev, *Complexity of positivstellensatz proofs for the knapsack*, Computational Complexity **10** (2001), no. 2, 139–154. 20

[Gri01b]   _____ , *Linear lower bound on degrees of positivstellensatz calculus proofs for the parity*, Theor. Comput. Sci. **259** (2001), no. 1-2, 613–622. 20, 21, 24

[Gro11]   David Gross, *Recovering low-rank matrices from few coefficients in any basis*, IEEE Trans. Information Theory **57** (2011), no. 3, 1548–1566. 10, 11, 12

[GS]      Chris Godsil and Sung Y Song, *Association schemes*, Handbook of Combinatorial Designs,, 325–330. 28

[Gur03]   Leonid Gurvits, *Classical deterministic complexity of edmonds' problem and quantum entanglement*, Proceedings of the thirty-fifth annual ACM symposium on Theory of computing, ACM, 2003, pp. 10–19. 4

[GW95]    Michel X. Goemans and David P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. Assoc. Comput. Mach. **42** (1995), no. 6, 1115–1145. MR 1412228 29

[Har70]   Richard A Harshman, *Foundations of the parafac procedure: Models and conditions for an" explanatory" multi-modal factor analysis*. 14, 29

[HK13]    Daniel J. Hsu and Sham M. Kakade, *Learning mixtures of spherical gaussians: moment methods and spectral decompositions*, ITCS, ACM, 2013, pp. 11–20. 13

[HKP⁺16]   Samuel B. Hopkins, Pravesh Kothari, Aaron Henry Potechin, Prasad Raghavendra, and Tselil Schramm, *On the integrality gap of degree-4 sum of squares for planted clique*, SODA, SIAM, 2016, pp. 1079–1095. 20

[HKP⁺17]   Samuel B. Hopkins, Pravesh K. Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer, *The power of sum-of-squares for detecting hidden structures*, FOCS, IEEE Computer Society, 2017, pp. 720–731. 24, 25, 29, 34, 35

[HL13]   Christopher J. Hillar and Lek-Heng Lim, *Most tensor problems are NP-hard*, J. ACM **60** (2013), no. 6, Art. 45, 39. MR 3144915 30

[HL18]   Sam B. Hopkins and Jerry Li, *Mixture models, robustness, and sum of squares proofs*, STOC, ACM, 2018, p. (to appear). 9, 19

[HSS15]   Samuel B. Hopkins, Jonathan Shi, and David Steurer, *Tensor principal component analysis via sum-of-square proofs*, COLT, JMLR Workshop and Conference Proceedings, vol. 40, JMLR.org, 2015, pp. 956–1006. 4, 9, 32

[HSSS16]   Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer, *Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors*, STOC, ACM, 2016, pp. 178–191. 4, 14, 32

[Jer92]   Mark Jerrum, *Large cliques elude the metropolis process*, Random Struct. Algorithms **3** (1992), no. 4, 347–360. 2

[Kar76]   Richard M. Karp, *The probabilistic analysis of some combinatorial search algorithms*, 1–19. MR 0445898 2

[KMOW17]   Pravesh K. Kothari, Ryuhei Mori, Ryan O'Donnell, and David Witmer, *Sum of squares lower bounds for refuting any CSP*, STOC, ACM, 2017, pp. 132–145. 33

[KMV10]   Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant, *Efficiently learning mixtures of two gaussians*, STOC, ACM, 2010, pp. 553–562. 18

[Kri64]   Jean-Louis Krivine, *Anneaux préordonnés*, Journal d'analyse mathématique **12** (1964), no. 1, 307–326. 1

[KSS18]   Pravesh K. Kothari, Jacob Steinhardt, and David Steurer, *Robust moment estimation and improved clustering via sum-of-squares*, STOC, ACM, 2018, p. (to appear). 9, 19

[KV15]   Subhash A. Khot and Nisheeth K. Vishnoi, *The unique games conjecture, integrability gap for cut problems and embeddability of negative-type metrics into $\ell_1$*, J. ACM **62** (2015), no. 1, Art. 8, 39. MR 3323774 20

[Las01]   Jean B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM J. Optim. **11** (2000/01), no. 3, 796–817. MR 1814045 5

[LCC07]   Lieven De Lathauwer, Joséphine Castaing, and Jean-François Cardoso, *Fourth-order cumulant-based blind identification of underdetermined mixtures*, IEEE Trans. Signal Processing **55** (2007), no. 6-2, 2965–2973. 13, 14

[LRA93]   S. E. Leurgans, R. T. Ross, and R. B. Abel, *A decomposition for three-way arrays*, SIAM J. Matrix Anal. Appl. **14** (1993), no. 4, 1064–1083. MR 1238921 14

[MP16]    Dhruv Medarametla and Aaron Potechin, *Bounds on the norms of uniform low degree graph matrices*, Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques, APPROX/RANDOM 2016, September 7-9, 2016, Paris, France (Klaus Jansen, Claire Mathieu, José D. P. Rolim, and Chris Umans, eds.), LIPIcs, vol. 60, Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 2016, pp. 40:1–40:26. 27

[MPW15]   Raghu Meka, Aaron Potechin, and Avi Wigderson, *Sum-of-squares lower bounds for planted clique*, STOC, ACM, 2015, pp. 87–96. 20

[MR05]    Elchanan Mossel and Sébastien Roch, *Learning nonsingular phylogenies and hidden markov models*, STOC, ACM, 2005, pp. 366–375. 13

[MR14]    Andrea Montanari and Emile Richard, *A statistical model for tensor pca*, Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2 (Cambridge, MA, USA), NIPS'14, MIT Press, 2014, pp. 2897–2905. 4, 30

[MS16]    Andrea Montanari and Nike Sun, *Spectral algorithms for tensor completion*. 32

[MSS16]   Tengyu Ma, Jonathan Shi, and David Steurer, *Polynomial-time tensor decompositions with sum-of-squares*, FOCS, IEEE Computer Society, 2016, pp. 438–446. 9, 13, 14, 15, 45

[MV10]    Ankur Moitra and Gregory Valiant, *Settling the polynomial learnability of mixtures of gaussians*, FOCS, IEEE Computer Society, 2010, pp. 93–102. 18

[Par00]   Pablo A Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, Ph.D. thesis, California Institute of Technology, 2000. 5

[PS17]    Aaron Potechin and David Steurer, *Exact tensor completion with sum-of-squares*, COLT, Proceedings of Machine Learning Research, vol. 65, PMLR, 2017, pp. 1619–1673. 9, 12, 13

[Rec11]   Benjamin Recht, *A simpler approach to matrix completion*, Journal of Machine Learning Research **12** (2011), 3413–3430. 10, 11, 12

[Rez00]   Bruce Reznick, *Some concrete aspects of Hilbert's 17th Problem*, Real algebraic geometry and ordered structures (Baton Rouge, LA, 1996), Contemp. Math., vol. 253, Amer. Math. Soc., Providence, RI, 2000, pp. 251–272. MR 1747589 5

[Rot13]    Thomas Rothvoß, *The lasserre hierarchy in approximation algorithms*. 7

[RRS17]    Prasad Raghavendra, Satish Rao, and Tselil Schramm, *Strongly refuting random csps below the spectral threshold*, STOC, ACM, 2017, pp. 121–131. 4, 31, 32, 33

[Sch08]    Grant Schoenebeck, *Linear level lasserre lower bounds for certain k-csps*, FOCS, IEEE Computer Society, 2008, pp. 593–602. 20, 21, 24

[SS17]     Tselil Schramm and David Steurer, *Fast and robust tensor decomposition with applications to dictionary learning*, COLT, Proceedings of Machine Learning Research, vol. 65, PMLR, 2017, pp. 1760–1793. 15, 32

[Ste74]    Gilbert Stengle, *A nullstellensatz and a positivstellensatz in semialgebraic geometry*, Mathematische Annalen **207** (1974), no. 2, 87–97. 1

[Tao12]    Terence Tao, *Topics in random matrix theory*, vol. 132, American Mathematical Soc., 2012. 29, 30

[Tre12]    Luca Trevisan, *On Khot's unique games conjecture*, Bull. Amer. Math. Soc. (N.S.) **49** (2012), no. 1, 91–111. MR 2869009 5

[Tul09]    Madhur Tulsiani, *CSP gaps and reductions in the lasserre hierarchy*, STOC, ACM, 2009, pp. 303–312. 20

[VW04]     Santosh Vempala and Grant Wang, *A spectral algorithm for learning mixture models*, J. Comput. Syst. Sci. **68** (2004), no. 4, 841–860. 18, 19

## A    Continued proofs of identifiability

Here, we fill in some details from the proofs in Section 2.

*Claim* (Restatement of Claim 2.4). When $\{a_i\}_{i \in [r]}$ are orthogonal and

1. $\mathcal{A} \models \left\{ \sum_{j \in [r]} \langle a_j, b_i \rangle^2 = 1 \right\}_{i \in [r]}$, and

2. $\mathcal{A} \models \{ \langle a_{j_1}, b_i \rangle^2 \langle a_{j_2}, b_i \rangle^2 = 0 \}_{j_1 \neq j_2 \in [r]}$,

then

$$\mathcal{A} \models \| b_i^{\otimes 3} - \sum_{j=1}^{r} \langle a_j, b_i \rangle^3 a_j^{\otimes 3} \|^2 = 0.$$

*Proof.* It follows from the orthogonality of the $a_i$ and from the first condition that $\mathcal{A} \models \| b_i - \sum_{j=1}^{r} \langle a_j, b_i \rangle a_j \|^2 = 0$. To prove the claim, we now use that in turn, $\mathcal{A} \models \| b_i^{\otimes 3} - $

$(\sum_{j=1}^{r}\langle a_j, b_i\rangle a_j)^{\otimes 3}\|^2 = 0$ and verify

$$\mathcal{A}\vdash\quad \left\|\left(\sum_{j=1}^{r}\langle a_j, b_i\rangle a_j\right)^{\otimes 3} - \sum_{j=1}^{r}\langle a_j, b_i\rangle^3 \cdot a_j^{\otimes 3}\right\|^2$$

$$= \left\|\sum_{\substack{j_1, j_2, j_3 \in [r] \\ \text{not all equal}}} \langle a_{j_1}, b_i\rangle\langle a_{j_2}, b_i\rangle\langle a_{j_3}, b_i\rangle \cdot a_{j_1} \otimes a_{j_2} \otimes a_{j_3}\right\|^2$$

$$\leqslant C(r) \sum_{\substack{j_1, j_2, j_3 \in [r] \\ \text{not all equal}}} \langle a_{j_1}, b_i\rangle^2\langle a_{j_2}, b_i\rangle^2\langle a_{j_3}, b_i\rangle^2$$

$$= 0 \,.$$

Where we have used that $\vdash^{x} \{(x_1 + \cdots + x_r)^2 \leqslant C(r) \cdot (x_1^2 + \cdots + x_r^2)\}$ for some function $C(r)$, and the second condition of the claim. We conclude that

$$\mathcal{A}\vdash \left\|\sum_{i=1}^{r} b_i^{\otimes 3} - \sum_{i,j}\langle a_j, b_i\rangle^3 a_j^{\otimes 3}\right\|_{\mathrm{F}}^2 = 0 \,.$$

as desired. $\qquad\qquad\square$

We now give a full proof of the robust version of Jennrich's algorithm.

**Theorem** (Restatement of Theorem 2.7). *There exists $\varepsilon > 0$ and a randomized polynomial-time algorithm that given a 3-tensor $\mathbf{T} \in (\mathbb{R}^n)^{\otimes 3}$ outputs a unit vector $u \in \mathbb{R}^n$ with the following guarantees: Let $a_1, \ldots, a_r \in \mathbb{R}^n$ 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 $a_1, \ldots, a_r$. Suppose $\|T - \sum_i a_i^{\otimes 3}\|_{\mathrm{F}}^2 \leqslant \varepsilon \cdot r$ and that $\max\{\|T\|_{\{1,3\}\{2\}}, \|T\|_{\{1\}\{2,3\}}\} \leqslant 10$. Then, with at least inverse polynomial probability, $\max_{i \in [r]}\langle a_i, u\rangle \geqslant 0.9$.*

*Proof.* For a vector $v \in \mathbb{R}^n$, define the linear operator $\mathcal{M}_v = \mathrm{Id} \otimes \mathrm{Id} \otimes v^\mathsf{T}$ from $(\mathbb{R}^n)^{\otimes 3} \to (\mathbb{R}^n)^{\otimes 2}$. We apply the following version of Jennrich's algorithm to $\mathbf{T}$: Choose a Gaussian vector $g \sim \mathcal{N}(0, \mathrm{Id})$ and apply $\mathcal{M}_g$ to the $d^3 \times 1$ reshaping of $\mathbf{T}$. Reshape the resulting vector $\mathcal{M}_g\mathbf{T}$ to an $n \times n$ matrix:

$$(\mathcal{M}_g\mathbf{T})_{\{1\}\{2\}} = \sum_{j \in [n]} g_j \cdot T_i,$$

where $T_i$ is the $n \times n$ matrix resulting from the restriction of $\mathbf{T}$ to coordinate $i$ in the third mode. Then, output the top eigenvector of $(\mathcal{M}_g\mathbf{T})_{\{1\}\{2\}}$.

We let $\mathbf{T} = S + E$ where $S = \sum_i a_i^{\otimes 3}$. First, we claim that $\|\mathcal{M}_{a_i}E\|_{\mathrm{F}}^2 \leqslant 4\varepsilon$ for at least half of the indices $i \in [r]$. This claim follows by averaging from the following bound

$$\sum_{i=1}^{r} \left\| \mathcal{M}_{a_i} E \right\|_F^2 = \left\langle E, \left( \sum_{i=1}^{r} \mathcal{M}_{a_i}{}^\mathsf{T} \mathcal{M}_{a_i} \right) E \right\rangle$$

$$\leqslant \|E\|_F^2 \cdot \lambda_{\max}\left( \sum_{i=1}^{r} \mathcal{M}_{a_i}{}^\mathsf{T} \mathcal{M}_{a_i} \right) = \|E\|_F^2 \cdot \lambda_{\max}(AA^\mathsf{T}) \leqslant (1+\varepsilon)\varepsilon r \leqslant 2\varepsilon r \,. \quad \text{(A.1)}$$

Second, we claim $\|(\mathcal{M}_{a_i} S)_{\{1\}\{2\}} - a_i a_i{}^\mathsf{T}\| \leqslant 2\varepsilon$ for all $i \in [r]$. Indeed, by our assumption that the orthogonality defect is bounded,

$$(\mathcal{M}_{a_i} S)_{\{1\}\{2\}} - a_i a_i{}^\mathsf{T} = \sum_{j \neq i} \langle a_i, a_j \rangle \cdot a_j a_j{}^\mathsf{T} \preceq \varepsilon \cdot AA^\mathsf{T} \preceq (1+\varepsilon)\varepsilon \cdot \mathrm{Id}$$

and, by the same reasoning, $(\mathcal{M}_{a_i} S)_{\{1\}\{2\}} - a_i a_i{}^\mathsf{T} \succeq -(1+\varepsilon)\varepsilon \cdot \mathrm{Id}$. Taken together, these bounds imply $\|(\mathcal{M}_{a_i} \mathbf{T})_{\{1\}\{2\}} - a_i a_i{}^\mathsf{T}\| \leqslant 6\varepsilon$ for at least half of the indices $i \in [r]$.

To finish the analysis, we consider an index $i \in [r]$ that satisfies $\|(\mathcal{M}_{a_i} \mathbf{T})_{\{1\}\{2\}} - a_i a_i{}^\mathsf{T}\| \leqslant 6\varepsilon$. Decomposing $g = \sum_i \langle a_i, g \rangle \cdot a_i + g'$, we write $\mathcal{M}_g \mathbf{T}$ as the sum $\mathcal{M}_g \mathbf{T} = \langle g, a_i \rangle \mathcal{M}_{a_i} \mathbf{T} + \mathcal{M}_{g'} \mathbf{T}$. By a matrix Chernoff bound, the assumption $\max\{\|\mathbf{T}\|_{\{1,3\}\{2\}}, \|\mathbf{T}\|_{\{1\}\{2,3\}}\} \leqslant 10$ implies that with high probability $\|\mathcal{M}_{g'} \mathbf{T}\|_{\{1\}\{2\}} \leqslant O(\sqrt{\log n})$. (See [MSS16] for details.) Since $\langle g, a_i \rangle$ is independent of $g'$, the event $\langle g, a_i \rangle \geqslant 1/\varepsilon \cdot \|\mathcal{M}_{g'} \mathbf{T}\|_{\{1\}\{2\}}$ has inverse polynomial probability in $n$ (but exponentially small probability in $1/\varepsilon$). It is straightforward to verify that in this event $\frac{1}{\langle g, a_i \rangle} \cdot (\mathcal{M}_g \mathbf{T})_{\{1\}\{2\}}$ is at most $7\varepsilon$ far from $a_i a_i{}^\mathsf{T}$ in spectral norm. For small enough $\varepsilon$, these events imply that the algorithm outputs a unit vector $u$ that satisfies the conclusion of the theorem. $\qquad\square$

# B  Approximate preservation of equalities under low-degree projection

We now prove Theorem 3.3.

**Theorem** (Restatement of Theorem 3.3). *Suppose $\{p(x, y) = 0\} \in \mathcal{P}$ is always satisfied for $(x, y) \sim \mathcal{J}_*$ and let $B := \max_{(x,y) \in \mathcal{J}_\emptyset} |p(x, y)|$ and let $D_y := \deg_y(p)$ and $d_x := \deg_x(p)$. If $d \geqslant d_x$ and $\bar{\mu}_y$ is the $D$-pseudocalibrated function defined in Eq. (3.4) then*

$$\mathop{\mathbb{P}}_{y \sim \mathcal{D}_\emptyset}\left[ \left| \mathop{\mathbb{E}}_x p(x, y) \bar{\mu}_y(x) \right| \geqslant \varepsilon \right] \leqslant \frac{B^2}{\varepsilon^2} \cdot \left\| \Pi_{d, D+2D_y} \circ \mu_* - \Pi_{d, D-1} \circ \mu_* \right\|_{2, \mathcal{J}_\emptyset}^2$$

*where $\Pi_{d,D}$ for $d, D \in \mathbb{N}$ denotes the projection on to $V_{d,D}$, the span of polynomials of degree at most $D$ in $y$ and degree $d$ in $x$.*

We begin with a couple of observations. Notice that $L_2(\mathcal{J}_\emptyset) = L_2(\mathcal{D}_\emptyset) \otimes L_2(\sigma)$ since the distribution $\mathcal{J}_\emptyset(x, y) = \mathcal{D}_\emptyset(y) \cdot \sigma(x)$. By Gram-Schmidt orthogonalization, the vector space $L_2(\mathcal{D}_\emptyset)$ can be written as a direct sum of

$$L_2(\mathcal{D}_\emptyset) = \bigoplus_{i=0}^{\infty} \mathcal{Y}_i$$

such that the following properties hold:

1. If $f \in \mathcal{Y}_i$ and $g \in \mathcal{Y}_j$ with $i < j$, then $\langle f, g \rangle_{\mathcal{D}_\emptyset} = 0$.

2. For each $i \in \mathbb{N}$ and $f \in \mathcal{Y}_i$, $\deg_y(f) \leq i$.

Similarly, one can decompose $L_2(\sigma) = \bigoplus_{i=0}^{\infty} \mathcal{X}_i$ with similar properties. Abusing notation, we will use $\Pi_i^{\mathcal{Y}}$ to also denote the projection operator on to the space $\mathcal{Y}_i$. Let $\Pi_{\leq D}^{\mathcal{Y}}$ denote the projection operator on to the space $\bigoplus_{i=0}^{D} \mathcal{Y}_i$, and let $\Pi_{[a,b]}^{\mathcal{Y}}$ denote the projection on to $\bigoplus_{i=a}^{b} \mathcal{Y}_i$. Let $\Pi_{\leq d}^{\mathcal{X}}$, $\mathcal{X}_{[a,b]}$ be similar operators for $L_2(\sigma)$. Both $\Pi_{\leq D}^{\mathcal{Y}}$ and $\Pi_{\leq d}^{\mathcal{X}}$ operators have a natural action on tensor product space $L_2(\mathcal{J}_\emptyset)$. In fact, the pseudocalibrated function can be defined as $\bar{\mu} = \Pi_{\leq D}^{\mathcal{Y}} \circ \Pi_{\leq d}^{\mathcal{X}} \circ \mu_*$.

We will require the following lemma, which relates the product of projections of polynomials to the projection of their product:

**Lemma B.1.** *Suppose $f, g \in L_2(\mathcal{J}_\emptyset)$ are polynomials and that $\deg_y(g) = D_y$. Then the following relationship between the the projection of the product and the product of projections holds:*

$$\Pi_{\leq D+D_y}^{\mathcal{Y}} \circ (fg) = (\Pi_{\leq D}^{\mathcal{Y}} \circ f)g + \Pi_{\leq D+D_y}^{\mathcal{Y}} \circ \left( (\Pi_{[D,D+2D_y]}^{\mathcal{Y}} \circ f)g \right). \tag{B.1}$$

*Proof.* Using the decomposition of the projector $\Pi^{\mathcal{Y}} = \Pi_{<D}^{\mathcal{Y}} + \Pi_{[D,D+2D_y]}^{\mathcal{Y}} + \Pi_{>D+2D_y}^{\mathcal{Y}}$ and that $f \in L_2(\mathcal{J}_\emptyset)$, we can express the left-hand-side term of Eq. (B.1) as

$$\Pi_{\leq D+D_y}^{\mathcal{Y}} \circ (fg) = \Pi_{\leq D+D_y}^{\mathcal{Y}} \circ \left( \left( \Pi_{<D}^{\mathcal{Y}} \circ f + \Pi_{[D,D+2D_y]}^{\mathcal{Y}} \circ f + \Pi_{>D+2D_y}^{\mathcal{Y}} \circ f \right) g \right),$$

$$= \Pi_{\leq D+D_y}^{\mathcal{Y}} \circ \left( (\Pi_{\leq D}^{\mathcal{Y}} \circ f)g \right) + \Pi_{\leq D+D_y}^{\mathcal{Y}} \circ \left( (\Pi_{[D,D+2D_y]}^{\mathcal{Y}} \circ f)g \right) + (\Pi_{>D+2D_y}^{\mathcal{Y}} \circ f)g.$$

In the first term, $\deg((\Pi_{\leq D}^{\mathcal{Y}} \circ f)g) \leq D + D_y$, so we may drop the leading projector to obtain the first right-hand-side of Eq. (B.1). For the third term, note that

$$\Pi_{\leq D+D_y}^{\mathcal{Y}} \circ \left( (\Pi_{>D+2D_y}^{\mathcal{Y}} \circ f)g \right) = 0,$$

for otherwise, we would have a polynomial $h \in \mathcal{Y}_{\leq D+D_y}$ such that

$$\left\langle h, \left( \Pi_{>D+2D_y}^{\mathcal{Y}} \circ f \right) g \right\rangle_{\mathcal{D}_\emptyset} = \left\langle hg, \Pi_{>D+2D_y}^{\mathcal{Y}} \circ f \right\rangle \neq 0,$$

a contradiction since $\deg_y(hg) \leq D + 2D_y$ while $\Pi_{>D+2D_y}^{\mathcal{Y}} \circ f \in \mathcal{Y}_{>D+2D_y}$. Putting these facts together, Eq. (B.1) follows immediately. $\square$

*Proof of Theorem 3.3.* Set $f = \Pi_{\leq d}^{\mathcal{X}} \circ \mu_*$ and $g = p$ in the statement of Theorem B.1. Rearranging, we have

$$(\Pi_{\leq D}^{\mathcal{Y}} \circ \Pi_{\leq d}^{\mathcal{X}} \circ \mu_*)p = \Pi_{\leq D+D_y}^{\mathcal{Y}} \circ \left( (\Pi_{\leq d}^{\mathcal{X}} \circ \mu_*)p \right) - \Pi_{\leq D+D_y}^{\mathcal{Y}} \circ \left( (\Pi_{[D,D+2D_y]}^{\mathcal{Y}} \circ \Pi_{\leq d}^{\mathcal{X}} \circ \mu_*)p \right),$$

and applying the $\mathbb{E}_x$ operator on both sides, we get

$$\mathbb{E}_x \left[ (\Pi_{\leq D}^{\mathcal{Y}} \circ \Pi_{\leq d}^{\mathcal{X}} \circ \mu_*)p \right] = \Pi_{\leq D+D_y}^{\mathcal{Y}} \circ \left( \mathbb{E}_x \left[ (\Pi_{\leq d}^{\mathcal{X}} \circ \mu_*)p \right] \right) - \Pi_{\leq D+D_y}^{\mathcal{Y}} \circ \left( \mathbb{E}_x \left[ (\Pi_{[D,D+2D_y]}^{\mathcal{Y}} \circ \Pi_{\leq d}^{\mathcal{X}} \circ \mu_*)p \right] \right).$$

By definition of the pseudocalibrated function $\bar{\mu}_y$, the left hand side is equal to $\mathbb{E}_x\, p(x,y)\bar{\mu}_y(x)$. Since $\deg_x(p) \leqslant d$, we have $\mathbb{E}_x(\Pi^{\mathcal{X}}_{\leqslant d} \circ \mu_*)p = \mathbb{E}_x\, \mu_* p$. Further $p(x,y) = 0$ for each $(x,y) \in \mathrm{supp}(\mu_*)$ implies that that $\mathbb{E}_x\, \mu_*(x,y)p(x,y) = 0$ for all $y$. Thus the first term in the right-hand side, given by $\Pi^{\mathcal{Y}}_{\leqslant D+D_y} \circ \left(\mathbb{E}_x\left[(\Pi^{\mathcal{X}}_{\leqslant d} \circ \mu_*)p\right]\right)$, is 0.

Therefore we have the following inequality for each $y$,

$$\mathbb{E}_x\, p(x,y)\bar{\mu}_y(x) \geqslant -\Pi^{\mathcal{Y}}_{\leqslant D+D_y} \circ \left(\mathbb{E}_x\left[(\Pi^{\mathcal{Y}}_{[D,D+2D_y]} \circ \Pi^{\mathcal{X}}_{\leqslant d} \circ \mu_*)p\right]\right).$$

Now, we apply Chebyshev's inequality to the right-hand side of the above,

$$\mathop{\mathbb{P}}_{y\sim\mathcal{D}_0}[|\mathbb{E}_x\, p(x,y)\bar{\mu}_y(x)| \geqslant \varepsilon] \leqslant \frac{1}{\varepsilon^2} \mathop{\mathbb{E}}_{y\sim\mathcal{D}_0} \left(\Pi^{\mathcal{Y}}_{\leqslant D+D_y} \circ \left(\mathbb{E}_x(\Pi^{\mathcal{Y}}_{[D,D+2D_y]} \circ \Pi^{\mathcal{X}}_{\leqslant d} \circ \mu_*)p)\right)\right)^2,$$

$$= \frac{1}{\varepsilon^2} \left\|\Pi^{\mathcal{Y}}_{\leqslant D+D_y} \circ (\mathbb{E}_x(\Pi^{\mathcal{Y}}_{[D,D+2D_y]} \circ (\Pi^{\mathcal{X}}_{\leqslant d} \circ \mu_*))p)\right\|^2_{2,\mathcal{J}_0},$$

And since norms decrease under projection,

$$\leqslant \frac{1}{\varepsilon^2} \left\|\mathbb{E}_x(\Pi^{\mathcal{Y}}_{[D,D+2D_y]} \circ \Pi^{\mathcal{X}}_{\leqslant d} \circ \mu_*)p\right\|^2_{2,\mathcal{J}_0},$$

Now, since we have assumed that $\max_{(x,y)\in\mathcal{J}_0} |p(x,y)| \leqslant B$,

$$\leqslant \frac{B^2}{\varepsilon^2} \mathop{\mathbb{E}}_{y\sim\mathcal{D}_0} \left(\mathbb{E}_x\left|(\Pi^{\mathcal{Y}}_{[D,D+2D_y]} \circ \Pi^{\mathcal{X}}_{\leqslant d} \circ \mu_*)\right|\right)^2,$$

$$\leqslant \frac{B^2}{\varepsilon^2} \mathop{\mathbb{E}}_{y\sim\mathcal{D}_0} \mathbb{E}_x \left|(\Pi^{\mathcal{Y}}_{[D,D+2D_y]} \circ \Pi^{\mathcal{X}}_{\leqslant d} \circ \mu_*)\right|^2,$$

$$= \frac{B^2}{\varepsilon^2} \left\|\Pi^{\mathcal{Y}}_{[D,D+2D_y]} \circ \Pi^{\mathcal{X}}_{\leqslant d} \circ \mu_*\right\|^2_{2,\mathcal{J}_0}.$$

This concludes the proof. □