Polynomial-Time Power-Sum Decomposition of Polynomials

Mitali Bafna
Department of Computer Science
Harvard University
Boston, USA
mitalibafna@gmail.com

Jun-Ting Hsieh
Computer Science Department
Carnegie Mellon University
Pittsburgh, USA
juntingh@cs.cmu.edu

Pravesh K. Kothari
Computer Science Department
Carnegie Mellon University
Pittsburgh, USA
praveshk@cs.cmu.edu

Jeff Xu
Computer Science Department
Carnegie Mellon University
Pittsburgh, USA
jeffxusichao@cmu.edu

***Abstract*—We give efficient algorithms for finding power-sum decomposition of an input polynomial $P(x) = \sum_{i \leqslant m} p_i(x)^d$ with component p_is. The case of linear p_is is equivalent to the well-studied tensor decomposition problem while the quadratic case occurs naturally in studying identifiability of non-spherical Gaussian mixtures from low-order moments.**

Unlike tensor decomposition, both the unique identifiability and algorithms for this problem are not well-understood. For the simplest setting of quadratic p_is and $d = 3$, prior work of [11] yields an algorithm only when $m \leqslant \widetilde{O}(\sqrt{n})$. On the other hand, the more general recent result of [13] builds an algebraic approach to handle any $m = n^{O(1)}$ components but only when d is large enough (while yielding no bounds for $d = 3$ or even $d = 100$) and only handles an inverse exponential noise.

Our results obtain a substantial quantitative improvement on both the prior works above even in the base case of $d = 3$ and quadratic p_is. Specifically, our algorithm succeeds in decomposing a sum of $m \sim \widetilde{O}(n)$ generic quadratic p_is for $d = 3$ and more generally the dth power-sum of $m \sim n^{2d/15}$ generic degree-K polynomials for any $K \geqslant 2$. Our algorithm relies only on basic numerical linear algebraic primitives, is *exact* (i.e., obtain arbitrarily tiny error up to numerical precision), and handles an inverse polynomial noise when the p_is have random Gaussian coefficients.

Our main tool is a new method for extracting the linear span of p_is by studying the linear subspace of low-order partial derivatives of the input P. For establishing polynomial stability of our algorithm in average-case, we prove inverse polynomial bounds on the *smallest* singular value of certain correlated random matrices with low-degree polynomial entries that arise in our analyses. Since previous techniques only yield significantly weaker bounds, we analyze the *smallest* singular value of matrices by studying the *largest* singular value of certain deviation matrices via *graph matrix decomposition* and the trace moment method.

I. Introduction

An n-variate polynomial $P(x)$ admits a power-sum decomposition if it can be written as $P(x) = \sum_{i \leqslant m} p_i(x)^d$ for some low-degree polynomials p_is. This work is about the algorithmic problem of computing such a decomposition when it exists and the related structural question of when such a decomposition, if it exists, is unique.

When p_is are *linear* forms $\langle v_i, x \rangle$ for $v_i \in \mathbb{R}^n$, the task of decomposing P is equivalent to decomposing the corresponding coefficient tensor $\sum_i v_i^{\otimes d}$ into rank 1 components. For $d = 2$, this corresponds to rank decomposition of matrices, which is unique only in degenerate settings. For $d = 3$, while the problem is already NP-hard [16], there is a long line of work on identifying natural sufficient conditions (e.g., Kruskal's condition [24]) that imply uniqueness of decomposition in all but degenerate settings. There are known efficient algorithms for decomposing tensors satisfying such non-degeneracy conditions and such algorithms form basic primitives in *tensor methods* [2], [4], [11], [14], [15], [20], [21], [23], [28]–[30], [33], [35]. An influential line of work has developed efficient learning algorithms for a long list of interesting statistical models (under appropriate assumptions) including Mixtures of Spherical Gaussians [11], [17], Independent Component Analysis [28], Hidden Markov Models [31], Latent Dirichlet Allocations [1], and Dictionary Learning [7] via reductions to tensor decomposition. Higher-degree power-sum decomposition is a natural generalization of the tensor decomposition problem and is equivalent to the well-studied problem of reconstructing certain classes of arithmetic circuits [13], [25], [26] with connections (see surveys [9], [40], [43]) to algebraic circuit lower bounds and derandomization.

a) Tensor Decomposition with Symmetries: Higher-degree power-sum decomposition is equivalent to a strict generalization of tensor decomposition where the components are symmetrized under a natural group action. For example, when $p_i(x) = x^\top A_i x$ are homogeneous quadratic polynomials for $n \times n$ matrices A_i, the coefficient tensor of P has the form $\mathbb{E}_{\sigma \sim \mathbb{S}_6} \sum_{i \leqslant m} \sigma(A_i^{\otimes 3})$ where $\mathbb{S}_6$ is the symmetric group on 6 elements and acts[1] by permuting the 6 indices involved in

[1]For example, for a symmetric matrix A, $\mathbb{E}_{\sigma \sim \mathbb{S}_6}[A^{\otimes 3}((i_1, i_2, i_3), (j_1, j_2, j_3))] = \mathbb{E}[A(e_1)A(e_2)A(e_3)]$ where the expectation is over the choice of a uniformly random perfect matching (e_1, e_2, e_3) of $\{i_1, i_2, i_3, j_1, j_2, j_3\}$.

any entry of $A_i^{\otimes 3}$. If not for the action of σ, the coefficient tensor would simply be a sum of tensor powers of vectorized A_is. The group action, however, has a drastic effect on the identifiability and algorithms for the problem. Specifically, the symmetrization causes the resulting tensor to have a large rank and thus any decomposition algorithm must strongly exploit the symmetries to succeed. In fact, in our full version, we exhibit a simple example of a sum of cubics of quadratics on 2 variables whose components are *not* uniquely identifiable even though the corresponding coefficient matrices of the quadratics are linearly independent. This is in contrast to the well-known result [15], [29] that 3rd order tensors with linearly independent components are uniquely identifiable and efficiently decomposable. This is similar to other orbit recovery problems that also reduce to tensor decomposition with symmetries such as multi-reference alignment and the cryo-EM [35], [38] problem where even establishing information-theoretic identifiability for generic parameters is significantly more challenging.

b) The Quadratic Case: Despite being the natural next step after linear p_is, power-sum decomposition of quadratic p_is is not well understood. In a seminal work, Ge, Huang and Kakade [11] (GHK from now) proved that the first 6 moments of a mixture of $m \sim \sqrt{n}$ non-spherical Gaussians with smoothed parameters *exactly* identify and (noise-resiliently) recover the m sets of means and covariances. Their analysis involves giving an algorithm (and uniqueness proof) for decomposing sums of cubics of *smoothed* quadratic *positive definite* polynomials but naturally generalizes to arbitrary smoothed quadratics. This is a striking result that exhibits a large gap between smoothed/generic parameters and arbitrary ones for mixtures of Gaussians as it is known that we need $\Omega(m)$ moments (an $n^{\Omega(m)}$-size object) to uniquely identify the parameters of arbitrary mixtures of m Gaussians [34]. Their approach uses a conceptually elegant "desymmetrize+tensor-decompose" strategy by first undoing the effect of the group action and then applying tensor decomposition. While their approach can potentially be extended to $m \geqslant \sqrt{n}$, it seems to encounter an inherent barrier at $m \geqslant n^{2/3}$ as we explain in Section II. Nevertheless, GHK conjectured that it should be possible to handle $m \approx n^{1-\delta}$ generic components for any $\delta > 0$ given $O(1)$-degree mixture moments which, in our context, corresponds to decomposing a sum of higher constant degree powers of quadratics.

c) The Garg-Kayal-Saha Algorithm: In a beautiful work of Garg, Kayal and Saha [13] (GKS from now), they suggest that there is an inherent barrier to extending the "desymmetrize+tensor-decompose" based approach of [11]. Instead, they work by exploiting an intriguing connection to algebraic circuit lower bounds and develop algorithms to recover any *polynomial* number of generic components from their power-sums of large enough degree. This algorithm however has two important deficiencies.

First, their strategy yields a decomposition algorithm for degree-d power-sums only when d is very large compared to the degree of the component p_is. In particular, they do not obtain any result for the simplest interesting setting of $d =$ 3rd (or even 100th) power of quadratics[2]. As a result, their techniques seem unsuitable to answer natural questions such as whether 6th moments of mixtures of non-spherical Gaussians (with generic parameters) can uniquely identify $m \geqslant n^{0.51}$ components of Gaussian mixture in n dimensions, or, whether a sum of $m \approx n$ cubics of generic quadratics can be uniquely decomposed. Second, their algorithm relies on algebraic methods for finding simultaneous vector-space decomposition. The resulting algorithm is not error-resilient and does not appear to handle even a small (e.g., $\exp(-n)$ in each entry) amount of noise in the input polynomial. In fact, GKS suggest finding a stable algorithm for power-sum decomposition as an open question.

d) This Work: In this paper, we give a conceptually simple algorithm that substantially improves the quantitative results in [13] for decomposing power-sums of low-degree polynomials. Somewhat surprisingly, our algorithm follows the "desymmetrize+decompose" approach similar to [11] while circumventing the barriers suggested by [13]. A key component is an efficient algorithm to extract the linear span of the coefficient tensors of (powers of) p_is from the subspace of "coordinate restrictions" of *partial derivatives* of $P = \sum_{i \leqslant m} p_i^d$ for $d \geqslant 3$. As a consequence of our algorithm, we obtain substantially improved guarantees even for the simplest non-trivial setting of sum of cubics of quadratics and handle $m \sim n$ components.

We give an error-tolerant implementation of our algorithm and prove that when each p_i has independent random Gaussian coefficients, the resulting algorithm tolerates an inverse polynomial amount of adversarial noise in the coefficients of the input polynomial. A key technical step in such an analysis requires establishing inverse polynomial lower bounds on the singular values of certain correlated random matrices whose entries are low-degree polynomials in the coefficients of p_is. Standard results (e.g., from [4]) for analyzing smallest singular values yield significantly weaker bounds in our setting. Instead, we rely on a new elementary but nimble method that lower bounds the *smallest* singular value of correlated random matrices by reducing the task to upper-bounding the much better understood *largest* eigenvalue of certain *deviation* matrices. Our analyses of the spectral norm of such matrices use the trace moment method combined with *graphical matrix decompositions* of random matrices that appear naturally in the analyses of sum-of-squares lower bound witnesses [6], [19] for average-case refutation problems. In particular, these sharper bounds are crucial in allowing us to handle $m \sim \widetilde{O}(n)$ components for decomposing sums of cubics of quadratics.

A. Our results

Our main result gives a polynomial time algorithm (in the standard bit complexity model with exact rational arithmetic) for decomposing a sum of d-th powers of generic (e.g.,

[2]Indeed, while their bounds can likely be somewhat optimized, the smallest power of quadratics that their algorithm (as currently analyzed) succeeds in decomposing must be larger than 2^{335}.

smoothed) polynomials. We note that just as in standard tensor decomposition, sums of squares of low-degree polynomials are uniquely decomposable only in degenerate settings, so cubics of quadratics (i.e., $d = 3$) is the simplest non-trivial setting in this context.

Theorem I.1 (Decomposing Power-Sums of Smoothed Polynomials)**.** *There is an algorithm that takes input an n-variate degree-Kd (for d a multiple of 3) polynomial of the form $\widehat{P}(x) = \sum_{i \leqslant m} \widehat{A}_i(x)^d$ where $\widehat{A}_i = A_i + G_i$ for an arbitrary degree-K polynomial A_i and a degree-K polynomial G_i with independent $\mathcal{N}(0, \rho^2)$ coefficients, runs in time polynomial in the size of its input and $1/\rho$, and has the following guarantee: with probability at least 0.99 over the draw of G_is and internal randomness, it outputs the set $\{\widehat{A}_i \mid i \leqslant m\}$ up to permutation (and signs, if d is even) whenever*

- $m \leqslant \widetilde{O}(n)$ *for* $d = 3$, $K = 2$,
- $m \leqslant \widetilde{O}(n^2)$ *for* $d = 6$, $K = 2$,
- $m \leqslant \widetilde{O}(n^{2d/9})$ *for any* $d \geqslant 9$ *and* $K = 2$,
- $m \leqslant \widetilde{O}(n^{2Kd/3(5K-4)})$ *for all* $d \geqslant 9$ *and* $K \geqslant 2$.

The theorem above works more generally for any model of smoothing that independently perturbs the coefficients of each A_t with a distribution that allots a probability of at most $1/n^{O(d)}$ to any single point. In particular, a fine-enough discretization of any continuous smoothing suffices. As observed in [11], [13], identifying components of non-spherical mixtures of Gaussians from low-degree moments is equivalent[3] to decomposing the power-sum of quadratic polynomials. Thus, as an immediate corollary of the theorem above, we obtain:

Corollary I.2 (Moment Identifiability of Smoothed Mixtures of Gaussians)**.** *The parameters of a zero-mean mixture of Gaussians $\sum_{i \leqslant m} w_i \mathcal{N}(0, \Sigma_i)$, with arbitrary mixture weights w_i and smoothed[4] covariances Σ_i, are uniquely identifiable from the first $2d$ moments for any $m \leqslant \widetilde{O}(n^{2d/9})$. For $d = 3$ and 6, the bound improves to $m \leqslant \widetilde{O}(n)$ and $m \leqslant \widetilde{O}(n^2)$ respectively.*

a) Error-Resilience for Random Components: When $P(x) = \sum_i A_i(x)^d + E(x)$ where each A_i has independent, standard Gaussian coefficients, we prove that the our algorithm above in fact is error-resilient and tolerates an inverse polynomial error in every coefficient of the input $\widehat{P}$. Indeed, Theorem I.1 above is obtained essentially as a corollary (combined with simple algebraic tools) of this stronger analysis for random components.

[3]This follows from the fact that for $x \in \mathbb{R}^n$, the $2d$-th moment of $\mathcal{N}(0, \Sigma)$ in direction x equals $\mathbb{E}_{y \sim \mathcal{N}(0,\Sigma)}[\langle y, x\rangle^{2d}] = \frac{(2d)!}{2^d d!}\mathbb{E}[\langle y, x\rangle^2]^d = \frac{(2d)!}{2^d d!}(x^\top \Sigma x)^d$. Consequently, $\mathbb{E}_{y \sim \sum_i w_i \mathcal{N}(0,\Sigma_i)}[\langle x, y\rangle^{2d}] = \frac{(2d)!}{2^d d!}\sum_i w_i (x^\top \Sigma_i x)^d$.

[4]Any continuous smoothing suffices for this result. For e.g., for an arbitrary $\widehat{\Sigma}_i \succeq 0$, for $\rho = n^{-O(1)}$, add an independent and uniformly random entry from $[-\rho, \rho]$ to every off-diagonal entry of $\widehat{\Sigma}_i$ and a uniformly random entry from $[n\rho, 2n\rho]$ to every diagonal entry of $\widehat{\Sigma}_i$ to produce Σ_i. Note that the resulting matrix Σ_i is positive semidefinite.

Theorem I.3 (Power-sum Decomposition of Random Polynomials)**.** *There is a polynomial time algorithm that takes input an n-variate degree-Kd (for d a multiple of 3) polynomial of the form $\widehat{P}(x) = \sum_{i \leqslant m} A_i(x)^d + E(x)$ where A_i is a degree-K polynomial with independent $\mathcal{N}(0, 1)$ coefficients, and $E(x)$ is an arbitrary polynomial of degree Kd, and has the following guarantees: with probability at least 0.99 over the draw of A_is and internal randomness, it outputs the set $\{\widetilde{A}_i \mid i \leqslant m\}$ that contains an estimate of each A_i up to permutation (and signs, if d is even) with an error of at most $n^{O(1)} \|E\|_F^{1/d}$ whenever*

- $m \leqslant \widetilde{O}(n)$ *for* $d = 3$ *and* $K = 2$,
- $m \leqslant \widetilde{O}(n^2)$ *for* $d = 6$ *and* $K = 2$,
- $m \leqslant \widetilde{O}(n^{2d/9})$ *for any* $d \geqslant 9$ *and* $K = 2$,
- $m \leqslant \widetilde{O}(n^{2Kd/3(5K-4)})$ *for all* $d \geqslant 9$ *and* $K \geqslant 2$.

B. Discussion and comparison to prior works

Theorem I.3 shows that our algorithm tolerates an inverse polynomial amount of noise in each entry when the component A_is are random. Theorem I.1 is in fact an immediate corollary of our analysis for the random case combined with standard tools. Our result for generic (as opposed to random) p_is only handles an inverse exponential amount of noise. We believe that the same algorithm should handle inverse polynomial noise (i.e., is well-conditioned) in any reasonable smoothed analysis model. However, establishing such a result likely requires new techniques for analyzing condition numbers of matrices with dependent, low-degree polynomial entries in independent random variables.

For the simplest setting of sums of cubics of quadratics (i.e., $K = 2$ and $d = 3$), our theorem yields a polynomial time algorithm that succeeds whenever $m \leqslant \widetilde{O}(n)$. This improves on the algorithm implicit in [11] that succeeds[5] for $m \leqslant \widetilde{O}(\sqrt{n})$. As we discuss in Section II, natural extensions of their techniques to higher degree power-sums also appear to break down for $m \geqslant n$.

The work of [13] recently found a more sophisticated algorithm (that works in general on all large enough fields) that relies on simultaneous decomposition of vector spaces that escapes this barrier. In particular, they showed that for any K, $m = n^{O(1)}$, there is an algorithm that succeeds in decomposing a sum of m dth powers of generic degree-K polynomials for *large enough* d. Their algorithm however requires that d be very large as a function of K and $\log_n m$ and in particular, does not work for $d = 3$ (or even 100) for example. Their algorithm relies on exact algorithms for certain algebraic operations and does not appear to tolerate any more than an inverse exponential (in n) amount of noise in the input.

The corollary above immediately improves the moment identifiability of mixtures of smoothed centered Gaussians shown in both the works above. Extending our algorithm to the "asymmetric" case of sums of products of quadratics (instead of powers) will allow the above corollary to succeed

[5]Their algorithm succeeds more generally for smoothed A_is but in addition, needs access to $\sum_i A_i(x)^2$.

for Gaussians with arbitrary mean, but we do not pursue this goal in this paper. We also note that unlike [11], our theorem above does not immediately yield a polynomial time algorithm for learning mixtures of smoothed Gaussians from samples (similar to [13]). This is because samples from the mixture only give us access to the corresponding sum of powers of quadratics with inverse polynomial additive error in each entry while our current analysis for the case of smoothed components only handles an inverse exponential error.

a) Open Questions: Despite the progress in this work, we are far from understanding identifiability and algorithms for power-sum decomposition. Our result shows unique identifiability for sums of $\sim n$ cubics of quadratics. Could this be improved to n^2? Conversely, could we produce evidence of hardness of decomposing sums of $\omega(n)$ cubics of quadratics? Analogous questions arise for higher-degree polynomials and we mention one that eludes the current approach in both our work and [13]: is it possible to obtain efficient algorithms that succeed in decomposing sums of m d-th powers of degree-K polynomials where m grows as $n^{f(K)d}$ for some $f(K) \to \infty$ as $K \to \infty$?

In a different direction, a natural question is to generalize our result to obtain a polynomial time algorithm that decomposes power-sums of smoothed polynomials while tolerating an inverse polynomial entrywise error. Our current analysis obtains such a guarantee for power-sums of random polynomials but can only handle an inverse exponential error in the smoothed setting. We suspect that this goal requires new tools to analyze the smallest singular values of matrices whose entries are low-degree polynomials in independent Gaussians with *non-zero means*.

C. Brief overview of our techniques

Given (the special case of) sum of cubics of quadratics $P(x) = \sum_{i \leqslant m} (x^\top A_i x)^3$ for $n \times n$ symmetric matrices A_i with coefficient tensor $\sum_{i \leqslant m} \mathrm{Sym}_6(A_i^{\otimes 3})$, the main idea of the algorithm in [11] is a conceptually simple "desymmetrize + tensor-decompose" approach. Here, desymmetrization reverses the effect of the polynomial symmetry and yields $\sum_i A_i^{\otimes 3}$, and one can then apply standard tensor decomposition. While Sym_6 is a linear operator on 6th order tensors with an $\Omega(n^6)$-dimensional kernel, it turns out that it is invertible when restricted to tensors where the component A_is are restricted to a *known* generic subspace. The work of [11] shows how to estimate the span of A_is – i.e. this subspace – for $m \leqslant \widetilde{O}(\sqrt{n})$. But their techniques do not seem to extend to any $m \gg n^{2/3}$. Indeed, Garg, Kayal and Saha [13] comment that reduction to tensor decomposition of the sort above cannot yield algorithms that work for $m \gg n$. As a result, they build a considerably more sophisticated approach that relies on an algebraic algorithm for simultaneous decomposition of vector spaces.

Our main idea comes as a surprise in the light of this discussion: we in fact give a conceptually simple "desymmetrize+tensor-decompose" based algorithm that substantially improves the bounds obtained in [13]. Our key idea is a "Span Finding algorithm" that recovers the linear span of A_is restricted to any $O(\sqrt{n})$ variables by computing the linear span of *restrictions* of partial derivatives of P and intersecting it with an appropriately constructed random subspace (see Section II for a more detailed overview).

Our algorithm is implemented using error-resilient numerical linear algebraic operations. In particular, to establish polynomial stability (Theorem I.3) for random A_i, we need to understand the *smallest* singular values (to obtain well-conditionedness) of certain correlated random matrices arising in our analyses. These random matrices are rather complicated with entries computed as low-degree polynomials (much smaller than the ambient dimension) of independent random variables. Standard techniques for analyzing such bounds (such as the "leave-one-out" method [39], [44], [45] employed in prior works on tensor decomposition [4], [33]) are inadequate for our purposes and yield weaker bounds (which, in particular, do not allow us to handle $m \sim \sqrt{n}$ for sum of cubics of quadratics, for example).

Instead, we rely on a new elementary method that establishes singular value lower bounds by studying *spectral norm* upper bounds of certain associated deviation matrices. We analyze and prove strong bounds on the spectral norm of such matrices using the graphical matrix decomposition technique that was introduced in [6], [19] and recently used and refined in several works [3], [12], [18], [22], [32], [37] on establishing sum-of-squares lower bounds for average-case problems and reducing the bounds to understanding certain combinatorial problems on graphs associated with the matrix. As far as we know, our work is the first use of this technique to prove singular value *lower bounds* and condition numbers in algorithms. We believe that the graphical matrix decomposition toolbox will find further applications in the analyses of numerical algorithms.

II. Technical Overview

In this section, we give a high-level overview of our algorithm and the key ideas that go into its design and analysis. Let's fix $P(x) = \sum_{t \leqslant m} A_t(x)^d + E(x)$ where $A_t(x)$ are homogeneous polynomials of degree K in n indeterminates $x_1, x_2, \ldots, x_n$. Throughout this paper, we will abuse notation slightly and use A_t to also denote the K-th order coefficient tensor of the associated polynomial. We will also use $\widetilde{O}$ to suppress $\mathrm{polylog}(n)$ factors. To begin with, we will focus on the case of *generic* A_ts – this simply means that A_ts do not satisfy any of some appropriate finite collection of polynomial equations. Eventually, as we explain in Section II-A, these equations will simply correspond to full-rankness of certain matrices that arise in our analyses. We will discuss a new method to prove strong polynomial condition number bounds for random A_ts in the following section. The results for smoothed/generic A_ts then follow via standard, simple tools.

Just like the special case of tensor decomposition (i.e., when A_t are linear forms), the decomposition is not uniquely identifiable from a sum of their quadratics (i.e., $d = 2$) except in degenerate cases (see Section B). Thus, the simplest non-trivial setting turns out to be $d = 3$.

In this section, we will focus on the simplest setting of $K = 2$ (and thus, A_t are simply $n \times n$ matrices) and $d = 3$. This, by itself, is an important special case and captures the question of identifiability of parameters from the 6th moments of a mixture of m n-dimensional Gaussians with zero-mean and smoothed covariance matrices, and our main results (Theorems I.1 and I.3) improve the current best identifiability results (Corollary I.2).

a) Structure of the Coefficient Tensor: Up to a constant scaling, the coefficient tensor of P equals $\sum_{t \leqslant m} \mathrm{Sym}(A_t^{\otimes 3})$. Here, $\mathrm{Sym} = \mathrm{Sym}_6$ acts on $A_t^{\otimes 3}$ by averaging over entries obtained by permuting the 6 elements involved.

b) Relationship to Tensor Decomposition: It is natural to compare our input to the related, *desymmetrized* tensor $\sum_t A_t^{\otimes 3}$, given which, we can immediately obtain the A_ts by applying standard tensor decomposition algorithms [15], [29] (see Fact III.7) whenever A_ts are linearly independent as vectors in $\binom{n+1}{2}$ dimensions. Our input, however, is not even close to a low-rank tensor because of the action of Sym_6 that generates essentially maximal rank terms even starting from a single generic A_t. Indeed, this effect is visible for just *bivariate* polynomials. In Appendix A of the full version of our work, we construct two different (and in fact, $\Omega(1)$-far in Frobenius norm) collections of robustly linearly independent bivariate quadratic polynomials such that the sums of their cubics have the same coefficient tensors. Thus, even though such A_ts can be uniquely and efficiently recovered from $\sum_t A_t^{\otimes 3}$ via standard tensor decomposition, it is information theoretically impossible to do so given $\sum_t \mathrm{Sym}(A_t^{\otimes 3})$.

c) The Ge-Huang-Kakade [11] Approach: The discussion above presents a conceptually simple way forward: if we could somehow compute the desymmetrized tensor (i.e., undo the effect of the group action) from the input, then we have reduced the problem to standard tensor decomposition. This is a bit tricky as the linear operation Sym_6 on 6th order tensors is a contraction that maps a $\binom{n+1}{2}^3 \sim n^6/8$-dimensional space into a $\binom{n+5}{6} \sim n^6/720$ dimensional subspace and is clearly not invertible (in fact, has a $\Omega(n^6)$-dimensional kernel) on arbitrary 6th order tensors. The main idea in GHK is to observe that Sym_6 can be invertible *when restricted* to 6th order tensors in some smaller subspace. In particular, let $B_1, B_2, \ldots, B_m$ be a basis for the span of the matrices A_t. Then, the desymmetrized coefficient tensor of P is a linear combination of $B_i \otimes B_j \otimes B_k$ – a subspace of m^3 dimension which is $\ll n^6/720$ if $m \ll n^2$. Proving such a claim requires analysis of the rank (and singular values, for polynomial error-stability) of the matrix representing Sym_6 on the linear span of A_ts and GHK managed to prove it for any $m \ll \sqrt{n}$.

To obtain the span of A_ts, GHK rely on access to $P_4 = \sum_{t \leqslant m} \mathrm{Sym}_4(A_t^{\otimes 2})$ in addition to the input tensor above. Plugging in e_a, e_b in the first two modes of this tensor yields an $n \times n$ matrix (i.e., a 2-D slice) of the form: $\sum_t A_t[a,b]A_t + \sum_t A_t[a] \otimes A_t[b]$ where $A_t[i]$ is the i-th column of A_t. As a, b vary, the first term generates the subspace of the span of A_ts. However, each such 2-D slice has an additive "error" that lies in the span of the rank 1 forms in the 2nd term above. The GHK idea is to zero out the rank 1 terms by projecting the 2-D slices to a subspace $\mathcal{S}^\perp$, where $\mathcal{S}$ contains the span of the rank 1 terms. To compute $\mathcal{S}$, they choose a subset $H \subseteq [n]$ and plug in $a, b, c \in H$ into three modes of P_4. The resulting 1-D slices are linear combinations of the columns $A_t[a]$ for $a \in H$ and $t \in [m]$. If $m|H| \ll n$, then all $A_t[a]$ are linearly independent generically, while if $|H|^3 \gg m|H|$, then there are enough slices to generate the span of $A_t[a]$ for all $a \in H$ and $t \in [m]$. This trade-off is optimized at $m \sim n^{2/3}$ and $|H| \sim n^{1/3}$. Given a good estimate of $\mathcal{S}$, we can now plug in $a, b \in H$ in two modes of P_4 and recover the span of A_t (restricted to columns in H) by projecting the resulting 2-D slices off $\mathcal{S}$. Repeating for disjoint choices of H completes the argument. In order to analyze the linear independence (and condition numbers) of the vectors arising in this analysis, GHK need to work with a somewhat smaller $m \sim \sqrt{n}$ in their argument.

d) Key Bottleneck in the GHK Approach: In our situation, we only have the sum of cubics P as input (but not P_4). But even given P_4, the crucial bottleneck is the need for recovering the span of a subset of columns of the A_ts. With more sophisticated analyses, given the above trade-offs, it's plausible that a sum of d-th powers of A_t allows handling m as large as $n^{1-O(1/d)}$, but there appears to be an inherent barrier at $m \sim n$. The GHK approach also seems to get unwieldy as it involves plugging in standard basis vectors in several modes of the tensor. This leads to more "spurious" terms that one must zero-out (instead of just the rank-1 terms for P_4).

Thus, even given higher powers, the GHK approach appears to have a natural break-point at $m \sim n$, and even handling $m \gg \sqrt{n}$ seems to require somewhat unwieldy analysis. In fact, in their recent work, Garg, Kayal and Saha [13] commented (see Page 17) *"However, we believe such an approach cannot be made to handle larger number of summands (say poly(n)) even in the quadratic case as the lower bounds for sums of powers of quadratics need substantially newer ideas than the linear case..."*.

e) The Garg-Kayal-Saha [13] Approach: In their beautiful recent work, GKS managed to find a different approach that escapes the above obstacles and showed an algorithm (that works on both finite fields and $\mathbb{Q}$) that for any K and $m = n^{O(1)}$, manages to decompose $P(x) = \sum_{t \leqslant m} A_t(x)^d$ for large enough d (and generic degree-K polynomials A_t). As discussed before, their approach requires d to be a large enough constant as a function of K and $\log_n m$ (though they remarked that the bounds could likely be improved, already for $K = 2$, they need $d \geqslant 2^{335}$ and $m \leqslant n^{d/1100}$). Their main idea, however, is relevant to our approach so we briefly describe it here.

We restrict our attention to the quadratic case ($K = 2$) from here on. The GKS approach relies on the linear span of *partial derivatives* of the input polynomial P. In fact, taking rth partial derivatives of P is essentially the same (though, more principled and easier to analyze) as "plugging in" all possible standard basis vectors in r modes of the input coefficient tensor as in GHK. GKS observed that for $r < d$, the

$n_r = \binom{n+r-1}{r}$ many r-th partial derivatives of P are all of the form $\sum_{t \leqslant m} A_t(x)^{d-r} Q_t(x)$ for some degree-r polynomials Q_t. This linear subspace is *strictly contained within* the space of all polynomial multiples of $A_t(x)^{d-r}$ – the containment is strict because the latter space is of dimension $\sim mn_r \gg n_r$ for generic A_ts. However, if we were to project each of the n_r partial derivatives down to be a function of some small enough $\ell = o(n)$ variables y, then, the dimension counting above is no longer an obstruction to the span being all multiples of the projected $A_t(x)^{d-r}$. Indeed, for generic A_t, the subspace $\mathcal{U}$ of the *projected* partial derivatives does in fact equal the subspace $\mathcal{V}$ of all multiples of $B_t(y)^{d-r}$, where $B_t = M^\top A_t M$ (the projection of A_t) and $B_t(y) = A_t(My) = y^\top M^\top A_t M y$ for an $n \times \ell$ projection matrix M.

f) Key Bottleneck in the GKS Approach: If we take $r = d-1$, then, it appears that the partial derivatives give us access to the subspace of span of *multiples of* B_ts (of degree $2d-r = d+1$ for $K = 2$). If we could extract the span of quadratics B_t from this subspace, we could implement desymmetrization and tensor decomposition to obtain at least the B_ts (i.e., the projected A_ts).

Unfortunately, this hope did not materialize for GKS who managed only to recover the span of $B_t(y)^{d-r}$ for $r < 2d/3$. This is because their analysis of a certain "multi-GCD" requires that the subspaces $B_t(y)^{d-r} y_T$ for $|T| = r$ for each $t \in [m]$ only have trivial (i.e., 0) pairwise intersection. This condition is impossible if $r \geqslant 2(d-r)$ or $r \geqslant 2d/3$; for example, if $d = 3$ and $r = 2$, then the degree-4 polynomial $B_t(y) B_{t'}(y)$ is clearly in the subspaces corresponding to both t and t', which is a non-trivial intersection! Thus, the GKS analysis is restricted to work with $r < 2d/3$ and in particular, only manages to recover the span of $B_t(y)^{d-r}$ (for $d - r > d/3$). This route rules out the desymmetrization + tensor decomposition approach.

As a result, GKS used a more complicated sequence of operations that involves taking projections of the partial derivatives and algorithms for simultaneous decomposition of vector spaces into irreducibles which they analyze by studying the associated "adjoint algebra". The two-step projection step requires that d be very large as a function of $\log_n m$ (and degree K of the A_ts).

g) Summary: The "desymmetrize + tensor decompose" approach of GHK is elegant and simple but suffers from an inherent bottleneck for going beyond $m \sim n$ (or even $n^{2/3}$) for sums of cubics (or higher powers) of quadratics and gets unwieldy as d gets large. The GKS approach manages to handle any $m = n^{O(1)}$ components but only for very large d and relies on a somewhat complicated algebraic algorithm. While GKS do not do this, finding a polynomially conditioned variant of their algorithm will likely require significant effort.

A. Our approach and outline of our algorithm

Somewhat surprisingly, we manage to find an algorithm that achieves the best of both worlds. Our algorithm relies on the conceptually simple approach of desymmetrizing the input tensor (as in GHK) while at the same time managing to not only hit $m \sim n$ when $K = 2$ and $d = 3$ but also get a substantially improved trade-off compared to GKS for all m, d, K. Further, we find a polynomially stable implementation of our algorithm when A_ts are random by establishing condition number upper bounds on the structured random matrices that arise in our analysis.

In the following, we explain the main components in our algorithm and analysis: insights that rescue the simple "desymmetrize + tensor decompose" approach, the resulting algorithm, and a new method to prove strong condition number upper bounds on structured random matrices. We will focus on the case when the A_ts have independent $\mathcal{N}(0,1)$ entries in the following section. For this setting, we obtain an algorithm with polynomial error-stability guarantees. Our result for generic (or smoothed) A_t is a simple corollary of this result using standard tools.

a) Recovering the Span of B_ts: Recall that the GKS observation shows that given a polynomial $P(x) = \sum_{t \leqslant m} A_t(x)^d$ for quadratic A_ts, the subspace $\mathcal{U}$ spanned by r-th partial derivatives of P, when *projected* to a sufficiently small dimension $\ell = o(n)$ variables y, equals the span $\mathcal{V} = \text{span}(B_t(y)^{d-r} y_T \mid t \in [m], T \in [\ell]^r)$ for $B_t(y) = A_t(My)$ where M is an $n \times \ell$ projection matrix.

GKS then perform a multi-GCD step that recovers the span of $B_t(y)^{d-r}$ from $\mathcal{V}$ and their analysis requires the subspaces $\{B_t(y)^{d-r} y_T \mid T \in [\ell]^r\}$ for each $t \in [m]$ to have only trivial pairwise intersection (i.e. $= \{0\}$) . Our key idea is to observe that this assumption is not crucial! We can extract the span of powers of B_t as long as these subspaces do not have a large intersection. As discussed before, when $r \geqslant 2d/3$, their analysis fails because of some obvious intersections between the above subspaces. We substantially improve their analysis by observing that for random polynomials these obvious intersections between the subspaces *are the only ones* possible!

More precisely, let's restrict to $d = 3$ and consider the subspace of projected (we in fact show that simply *restricting* the variables suffices) partial derivatives of order $r = 2$ of P. Then, the subspace of restricted 2nd order partial derivatives of P contains homogeneous polynomials of degree 4. For random A_ts, we fully characterize the set of quadratic polynomials $\{q_t \mid t \leqslant m\}$ that satisfy the polynomial equality $\sum_{t \leqslant m} B_t(y) q_t(y) = 0$. Observe that for any $s \neq t \in [m]$, $q_s = B_t$ and $q_t = -B_s$ is clearly in the solution space. Such solutions span a subspace of dimension $\binom{m}{2}$, and we prove that these solutions are in fact the *only* solutions whenever $m \leqslant \widetilde{O}(n)$.

This understanding immediately allows us to use a simple subroutine to recover the span of $\{B_t(y) \mid t \leqslant m\}$. Specifically, we take a random homogeneous quadratic polynomial $p(y)$ and let $\mathcal{V}_p$ be the subspace of quartic multiples of p, that is, $\mathcal{V}_p = \text{span}(p(y) y_S \mid |S| = 2)$. Then, any non-zero $f(y) = p(y) q_0(y) \in \mathcal{V} \cap \mathcal{V}_p$ must be a solution to $\sum_{t \leqslant m} B_t(y) q_t(y) = p(y) q_0(y)$. The above characterization of the solution subspace allows us to conclude that whenever q_0 is non-zero, it lies in the span of $B_t(y)$. Thus, we have confirmed that $\mathcal{V} \cap \mathcal{V}_p = \text{span}(p(y) B_t(y) \mid t \leqslant m)$, and dividing this

subspace by p immediately yields $\mathrm{span}(B_t(y) \mid t \leqslant m)$!

Thus, to summarize, our algorithm for finding the span of B_ts is simple:

1) Restrict all 2nd order partial derivatives of P to some ℓ variables ($\ell = O(\sqrt{n})$ suffices),
2) Find intersection of this subspace with $\mathcal{V}_p$ for a random homogeneous quadratic polynomial p and divide the resulting subspace by p.

The analog of this result for $d = 3D$ powers of quadratics relies on a similar lemma that characterizes the solution space of $\sum_{t \leqslant m} B_t(y)^D q_t(y) = 0$. For sums of powers of degree $K > 2$ polynomials however, the characterization gets a little more involved as unlike in the case of quadratic B_ts, q_t will have a larger degree than B_t^D, which makes the solution space larger.

b) Noise Resilient Implementation: For obtaining a noise-resilient version of the above method, we first need a noise-robust version of the GKS observation that the subspace $\mathcal{U}$ of restricted partial derivatives equals the subspace $\mathcal{V}$ spanned by multiples of $B_t(y)$, and also a robust way of obtaining a basis for $\mathcal{V}$. This amounts to understanding the smallest nonzero singular value of various matrices. Finally, we robustly compute the intersection of two subspaces given a basis for each by looking at the largest singular values of the sum of the corresponding projection matrices, allowing us to obtain a subspace close to the span of B_t.

c) Desymmetrization: The above discussions show how we can estimate the span of $B_t(y)$ for a restriction of the quadratic A_t to some $\ell = O(\sqrt{n})$ variables. Given this subspace, we apply desymmetrization *directly to the restricted polynomial* $P(My)$. To analyze this step, we need to understand the invertibility (and condition numbers) of the matrix representing the Sym_6 linear transform on the subspace of the linear span of B_t.

d) Aggregating Restrictions: For a given restriction (via an $n \times \ell$ matrix M), the above steps give us access to the tensor $\sum_{t \leqslant m} B_t^{\otimes 3}$ where $B_t = M^\top A_t M$ is the $\ell \times \ell$ matrix of the restricted A_t. We would like to piece together such restrictions to obtain $\sum_{t \leqslant m} A_t^{\otimes 3}$. We show how to do this by working with a simple n^6-size *pseudorandom* set of restriction matrices M such that the average over the corresponding restricted 3rd order tensor gives us the unrestricted 3rd order tensor up to a known scaling. Our construction is a simple modification of the standard construction of 6-wise independent hash families.

e) Tensor decomposition and taking sD-th roots: Given an estimate of $\sum_{t \leqslant m} A_t^{\otimes 3}$, we can apply the standard polynomially-stable tensor decomposition algorithms (Fact III.7) to recover the A_ts. When we work with higher ($d = 3D$) powers of quadratics (or degree-K polynomials, more generally), this step only gives us $\mathrm{Sym}_{KD}(A_t^{\otimes D})$. The task of recovering A_t given $\mathrm{Sym}_{KD}(A_t^{\otimes D})$ is a certain simple "deconvolution" problem. We give a noise-robust algorithm for this task that relies on a simple semidefinite program analyzed in Lemma III.9.

B. Overview of singular value lower bounds

For establishing polynomial stability of our algorithm for random A_ts and proving Theorem I.3, we need to understand the condition number and in particular, the smallest singular value of certain random matrices that arise in our analyses. Analyzing the smallest singular value of random matrices turns out to be more challenging than the much better understood largest singular value. For matrices with independent and *identically* distributed random subgaussian entries, a sharp bound was only achieved in the breakthrough work of [39] via a sophisticated analysis via the "leave-one-out" distance method. The matrices that arise in our analyses are significantly more involved. The entries are not independent but are instead computed as low-degree polynomials of independent random variables that are of polynomially smaller number than the dimension of the matrix. As a result, the entries exhibit large correlations, and the leave-one-out method appears hard to implement for such matrices.

Instead, we adopt a different, more elementary but nimble method that obtains estimates of the smallest singular values via upper bounds on the *largest* singular values of certain *deviation* matrices. To see this method on a simple toy example, consider an $n \times m$ matrix (for $m \ll n$) G of independent $\mathcal{N}(0,1)$ entries. Then, we can write $G^\top G = n(1 \pm O(\frac{1}{\sqrt{n}})) \cdot \mathbb{I} + \mathsf{offdiag}(G^\top G)$ where $\mathsf{offdiag}(G^\top G)$ zeros out the diagonal entries of $G^\top G$. To establish a lower bound on the mth singular value of G, it is thus enough to observe that $\left\|\mathsf{offdiag}(G^\top G)\right\|_{\mathrm{op}} \leqslant \widetilde{O}(\sqrt{mn})$ with high probability.

This argument works as long as $m \leqslant n/\operatorname{polylog}(n)$ and gives a sharp (up to the leading constant) estimate on the smallest singular value. Note that in this argument, we effectively "charge" the spectral norm of the off-diagonal "deviation" matrix to the smallest entry of the diagonal part. Such a strategy works so long as all columns of G are of roughly similar length.

It turns out that despite its simplicity, this technique is surprisingly resilient for our purposes and unlike methods from prior works, it easily applies to the involved matrices that arise in our analysis, yielding bounds that are essentially sharp so long as we can keep the dimensions of the matrix somewhat "lopsided" (i.e. $m \ll n$ in the example above). This turns out to not be a handicap in our setting.

In our analysis, the problem now reduces to bounding the spectral norm of certain correlated, low-degree polynomial-entry random matrices arising from the off-diagonal part of the matrices we analyze. While this can be quite complicated, we rely on the recent advances in understanding the spectral norm of such matrices [3], [6], [22] in the context of proving Sum-of-Squares lower bounds for average-case optimization problems. This technique relies on decomposing random matrices into a linear combination of certain structured random matrices called *graph matrices*. We rely on the tools from prior works that reduce the task of analyzing the spectral norm of such matrices to analyzing combinatorial properties of the underlying "graph".

This technique gets us started but hits a snag as it turns out that some of the deviation matrices simply *do not have small spectral norms*. We handle such terms by proving that the large spectral norm can be "blamed" on having large *positive* eigenvalues that cannot affect the bounds on the smallest singular value. Formally, we provide a charging argument, reminiscent of the positivity analyses in the construction of sum-of-squares lower bounds [6], [12], [18], [22], to handle such terms and establish the required bounds on the spectral norm.

While somewhat technical, the proofs of singular value lower bounds for all the matrices in our analyses follow the same blueprint. We give a more detailed exposition of these tools (by means of an example) in Section 6 of the full version before applying them to the matrices relevant to us.

III. Decomposing Power-Sums of Quadratics

In this section, we describe our efficient algorithm to decompose powers of low-degree polynomials. To keep the exposition simpler, we will analyze the algorithm for the case of quadratic p_is in this section and postpone the analysis for higher-degree p_is to the next section.

Specifically, we will prove that there is a polynomially stable and exact algorithm for decomposing power-sums of *random* quadratics. The same algorithm's recovery guarantees hold more generally for power-sums of *smoothed* quadratic polynomials though our current analysis only derives an inverse exponential error tolerance. Our algorithms work in the standard bit complexity model for exact rational arithmetic.

Theorem III.1. *There is an algorithm that takes input parameters $n, m, D \in \mathbb{N}$, an accuracy parameter $\tau > 0$, and the coefficient tensor $\widehat{P}$ of a degree-$6D$ polynomial $\widehat{P}$ in n variables with total bit complexity $\mathrm{size}(\widehat{P})$, runs in time $(\mathrm{size}(\widehat{P})n)^{O(D)}\,\mathrm{polylog}(1/\tau)$, and outputs a sequence of symmetric matrices $\widetilde{A}_1, \widetilde{A}_2, \ldots, \widetilde{A}_m \in \mathbb{R}^{n\times n}$ with the following guarantee.*

Suppose $\widehat{P}(x) = \sum_{t=1}^{m} A_t(x)^{3D} + E(x)$ where each A_t is an $n \times n$ symmetric matrix of independent $\mathcal{N}(0,1)$ entries, $\|E\|_F \leqslant n^{-O(D)}$ and $m \leqslant (\frac{n}{\mathrm{polylog}(n)})^D$ if $D \leqslant 2$ and $m \leqslant (\frac{n}{\mathrm{polylog}(n)})^{2D/3}$ if $D > 2$. Then, with probability at least 0.99 over the draw of A_is and internal randomness of the algorithm, for odd D,

$$\min_{\pi\in\mathbb{S}_m}\max_{t\in[m]} \left\|\widetilde{A}_t - A_{\pi(t)}\right\|_F \leqslant \mathrm{poly}(n)\left(\|E\|_F^{1/D} + \tau^{1/D}\right),$$

and for even D,

$$\min_{\pi\in\mathbb{S}_m}\max_{t\in[m]}\min_{\sigma\in\{\pm1\}} \left\|\widetilde{A}_t - \sigma A_{\pi(t)}\right\|_F \leqslant \mathrm{poly}(n)\left(\|E\|_F^{1/3D} + \tau^{1/3D}\right).$$

Observe that for odd D, we are able to recover A_ts up to permutation while for even D, we recover A_t up to permutation and signings. Such a guarantee is also the best possible given $P(x)$.

a) The Algorithm: : Our proof of Theorem III.1 uses the following algorithm (that works as stated for decomposing powers of degree-K A_ts more generally, but we will analyze for quadratic A_ts in this section).

Algorithm III-.1 (Decomposing Power Sums)**.**

Input: *Coefficient Tensor of a n-variate degree-$3KD$ polynomial $\widehat{P}(x) = P(x) + E(x)$ where $P(x) = \sum_{t\in[m]} A_t(x)^{3D}$ for degree-K polynomials A_t.*

Output: *Estimates $\widetilde{A}_1, \widetilde{A}_2, \ldots, \widetilde{A}_m$ of the coefficient tensors of $A_1, A_2, \ldots, A_m$.*

Operation:

1) ***Construct Pseudorandom Restrictions:*** *Construct the collection $\mathcal{S}$ of $\leqslant 3KD\ell$-size subsets of $[n]$, $|\mathcal{S}| = n^{O(D)}$ using the algorithm from Lemma III.6.*
2) ***Desymmetrize Pseudorandom Restrictions of Coefficient Tensor:*** *For each $S \in \mathcal{S}$:*
   a) ***Find Subspace of Restricted Partials:*** *Compute the linear span $\widetilde{\mathcal{V}}_D$ of coefficient vectors of M_S-restrictions of $2D$-th order partial derivatives of $\widehat{P}$.*
   b) ***Span-finding:*** *Find the span of restricted $A_t(x)^D$'s.*
   c) ***Desymmetrize:*** *Compute the desymmetrized restricted coefficient tensor.*
3) ***Aggregate Restricted Tensors:*** *Use restricted desymmetrized tensors from all restrictions in the pseudorandom set to construct the desymmetrized tensor.*
4) ***Decompose Tensor:*** *Apply tensor decomposition to the desymmetrized tensor.*
5) ***Take D-th Root of a Single Polynomial:*** *using Lemma III.9.*

b) Algorithm Overview: In this section, we henceforth restrict our attention to quadratic A_t's. Like in the case of cubics of quadratics discussed in Section II our algorithm first desymmetrizes the input coefficient tensor and then applies tensor decomposition to recover estimates of the individual components. Specifically (when $E = 0$), given the coefficient tensor of P has the form $\mathrm{Sym}_{6D}(\sum_{t\leqslant m}\mathrm{Sym}_{2D}(A_t^{\otimes D})^{\otimes 3})$, our goal is to "undo" the effect of the outer application of Sym and this is accomplished in the first three steps that are direct analogs of the ones discussed in the special case analyzed in Section II. After performing the desymmetrization step, for higher powers of quadratics we only recover estimates of $\mathrm{Sym}(A^{\otimes D})$ at the end of this procedure. The final (and extra, compared to the cubic case) step in the algorithm takes D-th root of noisy estimates of single polynomials, i.e. obtains an estimate of $A(x)$ from an estimate of $A(x)^D$.

Specifically, in Step 2a, we compute the $\sim n^{2D}$ different $2D$-th order partial derivatives $\partial_I\widehat{P}(x)$ of $\widehat{P}$ as $I = \{i_1, \ldots, i_{2D}\} \in [n]^{2D}$ ranges over all multisets of size $2D$. We then restrict each of these degree-$4D$ polynomials to some

fixed set of $\ell = o(n)$ variables which, in order to distinguish from the original set of indeterminates x, we will call y. The effect of this restriction is to transform A_t into $B_t = M^\top A_t M$ for a $n \times \ell$ restriction matrix M defined below.

Definition III.2 (Restriction matrix)**.** Given a set $S \subseteq [n]$ with $|S| = \ell$, we denote $M_S \in \mathbb{R}^{n\times\ell}$ to be the matrix whose columns consist of standard unit vectors e_j for $j \in S$. We write $P \circ M_S$ for the polynomial (in indeterminates y) defined by $P \circ M_S(y) = P(M_S y)$.

For each M_S, we let $\mathcal{R}_S$ be the linear operator that takes an $n\times n$ matrix $A \in \mathbb{R}^{n\times n}$, into $\mathcal{R}_S(A) = (M_S M_S^\top)A(M_S M_S^\top)$ – i.e., zeros out the (i,j) entry of A if i or j is not in S.

For any restriction matrix M, let $B_t = M^\top A_t M$. Let $\mathcal{V}_D$ be the span of polynomials of the form $B_t(y)^D y_T$:

$$\mathcal{V}_D \coloneqq \operatorname{span}\left(B_t(y)^D y_T \mid t \in [m], T \in [\ell]^{2D}\right) .$$

Then, any $2D$th order partial derivative of P, when restricted via M, is in $\mathcal{V}_D$. We prove that for small enough m, ℓ, the linear span of the restricted partials of P is in fact *equal* to the linear span of the polynomials $B_t(y)^D y_T$ (we prove an error-tolerant version in full version).

Lemma III.3 (Analysis of the Subspace of Restricted Partials of $\widehat{P}$)**.** *Fix $D \in \mathbb{N}$. Let $m, \ell, n \in \mathbb{N}$ be parameters such that $m \leqslant (\frac{\ell}{\operatorname{polylog}(\ell)})^{2D}$ if $D \leqslant 2$, and $m \leqslant (\frac{\ell}{\operatorname{polylog}(\ell)})^{D}$ if $D > 2$, and that $m\ell^{2D} \leqslant (\frac{n}{\operatorname{polylog}(n)})^{2D}$. Given $\widehat{P} = \sum_{t\in[m]} A_t(x)^{3D} + E(x)$ where each A_t is a degree-2 homogeneous polynomial with i.i.d. $\mathcal{N}(0,1)$ entries, and a restriction matrix $M \in \mathbb{R}^{n\times\ell}$, we have that with probability $1 - n^{-\Omega(D)}$ over the choice of A_t's, Algorithm Partial-Derivative outputs a subspace $\widetilde{\mathcal{V}}_D$ of $\mathbb{R}^{\ell_{4D}}$ that satisfies:*

$$\left\|\widetilde{\mathcal{V}}_D - \mathcal{V}_D\right\|_F \leqslant O\left(\frac{\|E\|_F}{(n\ell)^D}\right) ,$$

with $B_t(y) = A_t \circ M(y)$.

Consider $\mathcal{W}_D = \operatorname{span}(B_t(y)^D \mid t \in [m])$, for the next step, we show we can extract a subspace $\widetilde{\mathcal{W}}_D \approx \mathcal{W}_D$ given a basis for $\widetilde{\mathcal{V}}_D$, by proving for a random degree-$2D$ polynomial $p(y)$ the intersection (computed in Step 2b) of $\mathcal{V}_D$ with the linear span of polynomials of the form $p(y)y_T$ (for $|T| = 2D$) equals that of $B_t(y)^D$ with high probability over p and B_t's:

Lemma III.4 (Extracting Span of $B_t(y)^D$)**.** *Let D, m, ℓ be the same parameters as Lemma III.3. Given degree-2 homogeneous polynomials B_t for $t \in [m]$ in ℓ variables with coefficients drawn i.i.d from $\mathcal{N}(0,1)$, with probability $1 - \ell^{-\Omega(D)}$, the span-finding algorithm outputs $\widetilde{\mathcal{W}}_D$ that satisfies:*

$$\left\|\widetilde{\mathcal{W}}_D - \mathcal{W}_D\right\|_F \leqslant O\left(m\ell^{4D}\|\mathcal{V}_D - \widetilde{\mathcal{V}}_D\|_F\right).$$

Finally, we show that on the subspace of linear span of $B_t(y)^D$, the outer Sym_{6D} operation is invertible in an error-tolerant way via the least squares algorithm. This gives us a desymmetrized, M-restricted 3rd order tensor.

Lemma III.5 (Desymmetrization of Restricted $\widehat{P}$ via Least--Squares)**.** *Let $D, m, \ell \in \mathbb{N}$ such that $m \leqslant (\frac{\ell}{\operatorname{polylog}(\ell)})^{2D}$. For each $t \in [m]$, let B_t be a degree-2 homogeneous polynomial in ℓ variables with i.i.d. $\mathcal{N}(0,1)$ entries. Suppose $\widetilde{\mathcal{W}}_D$ is a subspace of $\mathbb{R}^{\ell_{2D}}$ such that $\left\|\widetilde{\mathcal{W}}_D - \mathcal{W}_D\right\|_F \leqslant 1/(m^{3.5}\ell^{O(D)})$, then with probability $1 - n^{\Omega(D)}$ over the choice of B_t's, Algorithm Desym outputs a tensor $\widetilde{T}$ such that:*

$$\left\|\widetilde{T} - \sum_{t\in[m]} \left(\operatorname{Sym}(B_t^{\otimes D})\right)^{\otimes 3}\right\|_F \leqslant \operatorname{poly}(m)\left(\ell^{O(D)}\left\|\widetilde{\mathcal{W}}_D - \mathcal{W}_D\right\|_F + \|E\|_F\right) .$$

We show how to aggregate the desymmetrized estimates above for $n^{O(D)}$ pseudorandom restriction matrices to obtain the estimate of the unrestricted tensor we need.

Lemma III.6 (Aggregating Pseudorandom Restrictions)**.** *Let $D, n, \ell, m \in \mathbb{N}$ such that $6D \leqslant \ell \leqslant n$. There is an $n^{O(D)}$-time computable collection $\mathcal{S}$ of subsets of $[n]$ such that each $S \in \mathcal{S}$ satisfies $\ell \leqslant |S| \leqslant 6D\ell$ and that*

$$\mathbb{E}_{S\sim\mathcal{S}} \sum_{t=1}^{m} \left(\operatorname{Sym}\left(\mathcal{R}_S(A_t)^{\otimes D}\right)\right)^{\otimes 3} = C \circ \sum_{t=1}^{m} \left(\operatorname{Sym}(A_t^{\otimes D})\right)^{\otimes 3}$$

where $C \in (\mathbb{R}^n)^{\otimes 6D}$ is a fixed tensor whose entries depend only on the entry locations, and each entry of C has value within $((\ell/2n)^{6D}, 1)$.

Given such a partially desymmetrized tensor, an application of off-the-shelf algorithms for 3rd order tensor decomposition allows us obtain $\operatorname{Sym}(A_t^{\otimes D})$ for $t \leqslant m$ in Step 4. We will specifically use:

Fact III.7 (Stable Tensor Decomposition, symmetric case of Theorem 2.3 in [4])**.** *There exists an algorithm that takes input a $n \times n \times n$ tensor $\widetilde{T}$ and an accuracy parameter $\tau > 0$, runs in time $(\operatorname{size}(T)n)^{O(1)} \operatorname{polylog}(1/\tau)$ and outputs a sequence of vectors $\widetilde{v}_1, \widetilde{v}_2, \ldots, \widetilde{v}_r$ with the following guarantee. If $\widetilde{T} = \sum_i v_i^{\otimes 3} + E$ for an arbitrary $n\times n\times n$ tensor E and the matrix with v_is as rows has a condition number (ratio of largest to r-th smallest singular value) at most $\kappa < \infty$. Then,*

$$\min_{\pi\in\mathbb{S}_r} \max_{i\leqslant r} \left\|\widetilde{v}_i - v_{\pi(i)}\right\|_2 \leqslant \operatorname{poly}(\kappa, n)\|E\|_F + \tau .$$

To apply this fact, we will need the following bound on the condition number κ of the matrix with $\operatorname{Sym}(A_t^{\otimes D})$ as columns:

Lemma III.8 (Condition number)**.** *Under the same assumptions as Lemma III.3, let A_D be the $n_{2D} \times m$ matrix whose columns are the coefficient vectors of $A_t(x)^D$ for $t \in [m]$. Then, with probability $1 - n^{-\Omega(D)}$, the condition number $\kappa(A_D) \leqslant O(1)$.*

Recall that for any natural number k, we write $n_k = \binom{n+k-1}{k}$ for the number of distinct degree k monomials in n variables.

Finally, in Step 5, we extract A_t from $\operatorname{Sym}(A_t^{\otimes D})$ (i.e.

desymmetrize a single noisy power). Note that in this step, we do not need randomness/genericity of the A_t.

Lemma III.9 (Stable Computation of D-th Roots)**.** *Let $D, n \in \mathbb{N}$ and $\delta \geqslant 0$. Let $P \in \mathbb{R}^{n\times n}$ be an unknown symmetric matrix. Suppose $\widetilde{P_D}(x)$ is a homogeneous degree-D polynomial in n variables such that its coefficient tensor satisfies $\left\|\widetilde{P_D} - \mathrm{Sym}(P^{\otimes D})\right\|_F \leqslant \delta$. There is an algorithm that runs in $n^{O(D)}$ time and outputs $\widetilde{Q} \in \mathbb{R}^{n\times n}$ such that if D is odd, then*

$$\left\|\widetilde{Q} - P\right\|_F \leqslant O(\sqrt{n}\delta^{1/D}),$$

and if D is even, then

$$\min_{\sigma\in\{\pm 1\}} \left\|\widetilde{Q} - \sigma P\right\|_F \leqslant O(n\delta^{1/3D}) \cdot \|P\|_{\max}.$$

c) Putting things together: We will prove each of the above lemmas and provide details of each step in the following subsections. Here, we use them to finish the proof of Theorem III.1.

Proof of Theorem III.1: For $D \leqslant 2$, we set $\ell = \sqrt{n}$ and $m \leqslant (\frac{n}{\mathrm{polylog}(n)})^D$ such that $m \leqslant (\frac{\ell}{\mathrm{polylog}(\ell)})^{2D}$. For $D > 2$, we set $\ell = n^{2/3}$ and $m \leqslant (\frac{n}{\mathrm{polylog}(n)})^{2D/3}$ such that $m \leqslant (\frac{\ell}{\mathrm{polylog}(\ell)})^D$. In both cases, we have $m\ell^{2D} \leqslant (\frac{n}{\mathrm{polylog}(n)})^{2D}$.

We consider the collection $\mathcal{S}$ of subsets of $[n]$ from Lemma III.6 with parameter ℓ such that $|\mathcal{S}| = n^{O(D)}$ and $\ell \leqslant |S| \leqslant 6D\ell$ for all $S \in \mathcal{S}$. Thus for $m \leqslant (\frac{n}{\mathrm{polylog}(n)})^D$, the parameters $m, n, |S|$ satisfy $m \leqslant (\frac{|S|}{\mathrm{polylog}(n)})^{2D}$ and $m|S|^{2D} \leqslant (\frac{n}{\mathrm{polylog}(n)})^{2D}$.

Consider a set $S \in \mathcal{S}$ and the corresponding restriction matrix M_S, and let $B_t = M_S^\top A_t M_S \in \mathbb{R}^{|S|\times|S|}$. By Lemma III.3, III.4 and III.5 (assuming $\|E\|_F \leqslant n^{-\Omega(D)}$), after Steps 2a, 2b and 2c, we obtain tensor $\widetilde{T}_S \in \mathbb{R}^{\ell^{6D}}$ such that

$$\left\|\widetilde{T}_S - \sum_{t\in[m]} \left(\mathrm{Sym}(B_t^{\otimes D})\right)^{\otimes 3}\right\|_F \leqslant n^{O(D)} \cdot \|E\|_F,$$

with probability $1 - n^{\Omega(D)}$ over the randomness of the input. By union bound over $\mathcal{S}$, we get the same guarantees for all $S \in \mathcal{S}$ with probability $1 - \frac{1}{\mathrm{poly}(n)}$.

Next, observe that $\sum_{t\in[m]} \left(\mathrm{Sym}(B_t^{\otimes D})\right)^{\otimes 3} \in (\mathbb{R}^\ell)^{\otimes 6D}$ is simply a sub-tensor obtained by removing zero entries from the tensor $\sum_{t=1}^m \left(\mathrm{Sym}(\mathcal{R}_S(A_t)^{\otimes D}\right)^{\otimes 3} \in (\mathbb{R}^n)^{\otimes 6D}$ according to $S \subseteq [n]$. Therefore, for each $S \in \mathcal{S}$ we have an estimate of $\sum_{t=1}^m \left(\mathrm{Sym}(\mathcal{R}_S(A_t)^{\otimes D})\right)^{\otimes 3}$, then if we average over all $S \in \mathcal{S}$, by Lemma III.6, we get a tensor $\widetilde{R}'_D \in \mathbb{R}^{n^{6D}}$ such that

$$\left\|\widetilde{R}'_D - C \circ \sum_{t\in[m]} \left(\mathrm{Sym}(A_t^{\otimes D})\right)^{\otimes 3}\right\|_F \leqslant n^{O(D)} \cdot \|E\|_F$$

where the error bound is by triangle inequality, and C is a known tensor with entries within $((\ell/2n)^{6D}, 1)$. Thus, by normalizing $\widetilde{R}'_D$ according to C, we get a tensor $\widetilde{R}_D$ such that

$$\left\|\widetilde{R}_D - \sum_{t\in[m]} \left(\mathrm{Sym}(A_t^{\otimes D})\right)^{\otimes 3}\right\|_F \leqslant n^{O(D)} \cdot \|E\|_F.$$

Next, by the tensor decomposition algorithm (Fact III.7) and the condition number upper bound from Lemma III.8, Step 4 runs in $n^{O(D)}\,\mathrm{polylog}(\tau)$ time and outputs tensors $\widetilde{A_1^D}, \ldots, \widetilde{A_m^D}$ such that

$$\min_{\pi\in\mathbb{S}_m} \max_{t\in[m]} \left\|\widetilde{A_t^D} - \mathrm{Sym}\left(A_{\pi(t)}^{\otimes D}\right)\right\|_F \leqslant n^{O(D)}\|E\|_F + \tau.$$

Finally, by Lemma III.9 we can extract $\widetilde{A}_t \in \mathbb{R}^{n\times n}$ from $\widetilde{A_t^D}$. For odd D, using the fact that $x^{1/D}$ is a concave function when $D \geqslant 1$, we get that

$$\begin{aligned}\min_{\pi\in\mathbb{S}_m} \max_{t\in[m]} \left\|\widetilde{A}_t - A_{\pi(t)}\right\|_F &\leqslant O(\sqrt{n})\left(n^{O(D)}\|E\|_F + \tau\right)^{1/D} \\ &\leqslant \mathrm{poly}(n)\left(\|E\|_F^{1/D} + \tau^{1/D}\right).\end{aligned}$$

For even D, since $\|A_t\|_{\max} \leqslant \mathrm{polylog}\, n$ with high probability by standard concentration results, we get that

$$\begin{aligned}&\min_{\pi\in\mathbb{S}_m} \max_{t\in[m]} \min_{\sigma\in\{\pm1\}} \left\|\widetilde{A}_t - \sigma A_{\pi(t)}\right\|_F \\ &\leqslant O(n)\left(n^{O(D)}\|E\|_F + \tau\right)^{1/3D} \|A_t\|_{\max} \\ &\leqslant \mathrm{poly}(n)\left(\|E\|_F^{1/3D} + \tau^{1/3D}\right).\end{aligned}$$

This completes the proof. ∎

APPENDIX

A. Non-identifiability of sum of cubics of linearly independent quadratics

Lemma A.1. *Let $a = \sqrt{6}$ and consider the following distinct sets of bivariate quadratic polynomials in variables x, y:*

$$\mathcal{S}_1 = \{x^2 + axy,\ x^2 + y^2,\ y^2 + axy\},$$
$$\mathcal{S}_2 = \{x^2,\ x^2 + axy + y^2,\ y^2\}.$$

Then, the polynomials in each set have linearly independent coefficient matrices but the sum of cubics of polynomials in either sets is equal.

Proof: It is easy to verify that in both sets, the coefficient matrices of the polynomials are linearly independent. The sum of cubics of $\mathcal{S}_1$ is

$$\begin{aligned}&\sum_{p\in\mathcal{S}_1} p(x,y)^3 = 2x^6 + 3ax^5y \\ &+ 3(a^2+1)x^4y^2 + 2a^3x^3y^3 + 3(a^2+1)x^2y^4 + 3axy^5 + 2y^6\end{aligned}$$

whereas

$$\begin{aligned}&\sum_{p\in\mathcal{S}_2} p(x,y)^3 = 2x^6 + 3ax^5y + 3(a^2+1)x^4y^2 \\ &+ (6a + a^3)x^3y^3 + 3(a^2+1)x^2y^4 + 3axy^5 + 2y^6\end{aligned}$$

Thus by setting $a = \sqrt{6}$, we have $2a^3 = 6a + a^3$, meaning $\sum_{p\in\mathcal{S}_1} p(x,y)^3 = \sum_{p\in\mathcal{S}_2} p(x,y)^3$. ∎

B. Non-identifiability of sum of squares of quadratics

We observe that sum-of-squares of even two random homogeneous quadratics cannot be uniquely decomposed.

Lemma A.2 (Non-Indentifiability of Generic Sum of Quadratics of Quadratics)**.** *Let A_1, A_2 be $n \times n$ matrices of independent $\mathcal{N}(0,1)$ entries up to symmetry. Then, with probability 1 over the draw of A_1, A_2, there exist symmetric A_1', A_2' such that $\|A_i' - A_j\|_2 \geqslant 1/n^{O(1)}$ for every i, j such that $(x^\top A_1 x)^2 + (x^\top A_2 x)^2 = (x^\top A_1' x)^2 + (x^\top A_2' x)^2$ for every x.*

Proof: Let V_1, V_2 be the vectorization of upper-triangular entries of A_1, A_2 respectively. Since the coefficient tensor of $(x^\top A_1 x)^2 + (x^\top A_2 x)^2$ is a linear transformation (scaling of Sym operation) applied to $A_1^{\otimes 2} + A_2^{\otimes 2}$, it is enough to find V_1', V_2' distinct form V_1, V_2 such that $V_1^{\otimes 2} + V_2^{\otimes 2} = V_1'^{\otimes 2} + V_2'^{\otimes 2}$. The (since V_1, V_2 are random Gaussian, the rank decomposition is unique w.p. 1) orthogonal decomposition of the matrix $V_1 V_1^\top + V_2 V_2^\top$ uses orthogonal vectors V_1', V_2' that are different from V_1, V_2 (in fact must have a distance of at least $1/n \|V_1\|_2$). Taking A_i' to be the matrix whose upper triangular entries are given by V_i' for $i = 1, 2$ completes the proof. ∎

In this section, we derive Theorem I.1 as a corollary of our proof of Theorem I.3 combined with some elementary algebraic considerations.

We will rely on the following lemma that shows that whenever a matrix with low-degree polynomial entries in some variable A has full column rank for some real assignment to variables A, it must in fact have non-trivially lower bounded singular value for any A' after a small random perturbation in each entry. Specifically, we prove:

Lemma A.3. *Let $\mathcal{G}$ be a product distribution on N dimensional vectors such that the marginal of any coordinate of G is distributed so that no single point has probability $\geqslant 2^{-N^{O(1)}}$ (for e.g., uniform distribution on $N^{O(1)}$ bit rational numbers in any constant length interval suffices). Let $M(A)$ be a $R \times S$ matrix such that each entry of $M(A)$ is a degree-d polynomial in the N-dimensional vector A with each entry upper bounded by $2^{N^{O(1)}}$. Suppose there is a point $A' \in \mathbb{R}^N$ such that $M(A')$ has full rank R. Then, for any vector $B \in \mathbb{R}^N$ with rational entries of bit complexity at most $N^{O(1)}$,*

$$\Pr_{G\sim\mathcal{G}}[M(B+G) \text{ has } R\text{-th singular value} \geqslant 2^{-(SN)^{O(1)}}] \geqslant 1 - 2^{-N^{O(1)}}.$$

Our proof relies on the following variant of the classical Schwartz-Zippel lemma and a simple observation about eigenvalues of matrices with polynomial bit complexity entries.

Fact A.4 (Corollary of Generalized DeMillo–Lipton–Zippel Lemma, Theorem 4.6 in [5])**.** *Let $p(x_1, x_2, \ldots, x_n)$ be an n-variate degree-d polynomial over any field $\mathbb{F}$. Suppose p is not identically equal to 0. Let $S_1, S_2, \ldots, S_n$ be finite subsets of $\mathbb{F}$ of size $s \geqslant dn^2$. Then, if $x_i \sim S_i$ is chosen uniformly at random and independently for every i, then,*

$$\mathbf{Pr}[p(x) = 0] \leqslant dn/s\,.$$

Lemma A.5 (Gapped Eigenvalues from Polynomial Bit Complexity)**.** *Let A be a $n \times r$ matrix of N-bit rational entries. Suppose A has rank r. Then, the r-th smallest singular value of A is at least $2^{-O(Nn^3)}$.*

Proof: Let $B = A^\top A$ and let B' be the matrix of integers obtained by clearing the denominators of the rational numbers appearing in the entries of B. The bit complexity by B' is then larger than that of B by at most an additive Nn^2 and is thus at most $4Nn^2$. Further, by the Gershgorin circle theorem, the largest eigenvalue of B' is at most $n2^{4Nn^2}$.

Since B' is a symmetric matrix with integer entries and has full rank, the determinant of B', $\det(B')$ is a non-zero integer and thus at least 1 in magnitude. Since $\det(B')$ is the product of all r eigenvalues of B' each of which is at most $n2^{4Nn^2}$, the smallest eigenvalue must be at least $n^{-n}2^{-4Nn^3} \leqslant 2^{-5Nn^3}$ and large enough n. This completes the proof. ∎

Proof of Lemma A.3: For any fixed $B \in \mathbb{R}^N$, consider the determinant $\det(Q)$ of the $R \times R$ matrix $Q = M(A + B)M(A+B)^\top$. This is a polynomial of degree $2Rd$ in A. For $A^* = A' - B$, from the hypothesis, $M(A^* + B)M(A^* + B)^\top$ has full rank R. Thus, $\det(Q)$ is not identically equal to 0 as a polynomial of A.

Let $G \in \mathbb{R}^N$ be sampled from $\mathcal{G}$. For each entry i of G, let S_i be the support of the distribution that G_i is drawn from. Then, we know that this support is of size at least $2^{N^{O(1)}}$. Thus, by the generalized De-Millo-Lipton-Zippel lemma (Fact A.4), the probability that $M(B+G)M(B+G)^\top$ is singular is at most $2^{-N^{O(1)}}$.

Further, the entries of $M(B + G)M(B + G)^\top$ have bit complexity at most $N^{O(1)}$. Thus, by Lemma A.5, whenever the matrix $M(B+G)M(B+G)^\top$ is non-singular, it's smallest eigenvalue is lower bounded by $2^{-(SN)^{O(1)}}$. ∎

We can now finish the proof of Theorem I.1 and we defer this to the full version of our paper.

ACKNOWLEDGMENT

P.K. thanks Ankit Garg for preliminary discussions on power-sum decomposition. M.B. thanks Akash Sengupta for several helpful discussions and for suggesting that a version of Lemma 4.11 of the full version should be true.

REFERENCES

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

[2] Anima Anandkumar, Rong Ge, Daniel J. Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models (A survey for ALT). In *ALT*, volume 9355 of *Lecture Notes in Computer Science*, pages 19–38. Springer, 2015.

[3] Kwangjun Ahn, Dhruv Medarametla, and Aaron Potechin. Graph matrices: Norm bounds and applications, 2016.

[4] A. Bhaskara, M. Charikar, A. Moitra, and A. Vijayaraghavan. Smoothed analysis of tensor decompositions. In *Symposium on Theory of Computing, STOC 2014*, pages 594–603, 2014.

[5] Anurag Bishnoi, Pete L. Clark, Aditya Potukuchi, and John R. Schmitt. On zeros of a polynomial in a finite grid. *Combin. Probab. Comput.*, 27(3):310–333, 2018.

[6] Boaz Barak, Samuel Hopkins, Jonathan Kelner, Pravesh K Kothari, Ankur Moitra, and Aaron Potechin. A nearly tight sum-of-squares lower bound for the planted clique problem. *SIAM Journal on Computing*, 48(2):687–735, 2019.

[7] Boaz Barak, Jonathan A Kelner, and David Steurer. Dictionary learning and tensor decomposition via the sum-of-squares method. In *Proceedings of the forty-seventh annual ACM Symposium on Theory of Computing*, pages 143–151, 2015.

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

[9] Xi Chen, Neeraj Kayal, and Avi Wigderson. Partial derivatives in arithmetic complexity and beyond. *Found. Trends Theor. Comput. Sci.*, 6(1-2):1–138, 2011.

[10] N. Fleming, P. Kothari, and T. Pitassi. Semialgebraic proofs and efficient algorithm design. *Foundations and Trends® in Theoretical Computer Science*, 14(1-2):1–221, 2019.

[11] R. Ge, Q. Huang, and S. M. Kakade. Learning mixtures of gaussians in high dimensions. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015*, pages 761–770, 2015.

[12] M. Ghosh, F. Jeronimo, C. Jones, A. Potechin, and G. Rajendran. Sum-of-squares lower bounds for sherrington-kirkpatrick via planted affine planes. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 954–965, Los Alamitos, CA, USA, nov 2020. IEEE Computer Society.

[13] Ankit Garg, Neeraj Kayal, and Chandan Saha. Learning sums of powers of low-degree polynomials in the non-degenerate case. In *2020 IEEE 61st Annual Symposium on Foundations of Computer Science (FOCS)*, pages 889–899. IEEE, 2020.

[14] Rong Ge and Tengyu Ma. Decomposing Overcomplete 3rd Order Tensors using Sum-of-Squares Algorithms. In *Approximation, Randomization, and Combinatorial Optimization. Algorithms and Techniques (APPROX/RANDOM 2015)*, 2015.

[15] Richard A Harshman. Foundations of the parafac procedure: Models and conditions for an" explanatory" multi-modal factor analysis. 1970.

[16] Johan Håstad. Tensor rank is NP-complete. *Journal of Algorithms*, 11(4):644–654, 1990.

[17] D. Hsu and S. M. Kakade. Learning mixtures of spherical gaussians: moment methods and spectral decompositions. In *Innovations in Theoretical Computer Science, ITCS '13*, pages 11–20, 2013.

[18] Jun-Ting Hsieh and Pravesh K Kothari. Algorithmic thresholds for refuting random polynomial systems. In *Proceedings of the 2022 Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, pages 1154–1203. SIAM, 2022.

[19] Samuel B Hopkins, Pravesh K Kothari, and Aaron Potechin. SoS and planted clique: Tight analysis of MPW moments at all degrees and an optimal lower bound at degree four. *arXiv preprint arXiv:1507.05230*, 2015.

[20] Samuel B. Hopkins, Jonathan Shi, and David Steurer. Tensor principal component analysis via sum-of-square proofs. In *COLT*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 956–1006. JMLR.org, 2015.

[21] Samuel B Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer. Speeding up sum-of-squares for tensor decomposition and planted sparse vectors. *arXiv preprint arXiv:1512.02337*, 2015.

[22] Chris Jones, Aaron Potechin, Goutham Rajendran, Madhur Tulsiani, and Jeff Xu. Sum-of-squares lower bounds for sparse independent set. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 406–416. IEEE, 2022.

[23] Bohdan Kivva and Aaron Potechin. Exact nuclear norm, completion and decomposition for random overcomplete tensors via degree-4 sos. *arXiv preprint arXiv:2011.09416*, 2020.

[24] Joseph B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra Appl.*, 18(2):95–138, 1977.

[25] Neeraj Kayal and Chandan Saha. Reconstruction of non-degenerate homogeneous depth three circuits. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, page 413–424, New York, NY, USA, 2019. Association for Computing Machinery.

[26] Neeraj Kayal, Chandan Saha, and Ramprasad Saptharishi. A super-polynomial lower bound for regular arithmetic formulas. In *Proceedings of the Forty-Sixth Annual ACM Symposium on Theory of Computing*, STOC '14, page 146–153, New York, NY, USA, 2014. Association for Computing Machinery.

[27] Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[28] Lieven De Lathauwer, Josphine Castaing, and Jean-Franois Cardoso. Fourth-order cumulant-based blind identification of underdetermined mixtures. *IEEE Transactions on Signal Processing*, 55(6):2965–2973, 2007.

[29] S. E. Leurgans, R. T. Ross, and R. B. Abel. A decomposition for three-way arrays. *SIAM J. Matrix Anal. Appl.*, 14(4):1064–1083, 1993.

[30] Peter McCullagh. *Tensor methods in statistics*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1987.

[31] E. Mossel and S. Roch. Learning nonsingular phylogenies and hidden Markov models. In *Proc. 37th Annual ACM Symposium on Theory of Computing (STOC)*, pages 366–375, 2005.

[32] Sidhanth Mohanty, Prasad Raghavendra, and Jeff Xu. Lifting sum-of-squares lower bounds: degree-2 to degree-4. In *Proccedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020*, pages 840–853. ACM, 2020.

[33] Tengyu Ma, Jonathan Shi, and David Steurer. Polynomial-time tensor decompositions with sum-of-squares. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 438–446, 2016.

[34] A. Moitra and G. Valiant. Settling the polynomial learnability of mixtures of Gaussians. In *FOCS*, pages 93–102, 2010.

[35] Ankur Moitra and Alexander S Wein. Spectral methods from tensor networks. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, pages 926–937, 2019.

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

[37] Aaron Potechin and Goutham Rajendran. Machinery for proving sum-of-squares lower bounds on certification problems. *CoRR*, abs/2011.04253, 2020.

[38] Amelia Perry, Jonathan Weed, Afonso S. Bandeira, Philippe Rigollet, and Amit Singer. The sample complexity of multi-reference alignment. *CoRR*, abs/1707.00943, 2017.

[39] M. Rudelson and R. Vershynin. The Littlewood-Offord Problem and invertibility of random matrices. *Advances in Mathematics*, 218(2):600–633, 2008.

[40] Ramprasad Saptharishi. A survey of lower bounds in arithmetic complexity. *Github survey*, 2015.

[41] Gilbert Stewart and Jiguang Sun. *Matrix Perturbation Theory*. Academic Press, 1990.

[42] Warren Schudy and Maxim Sviridenko. Concentration and moment inequalities for polynomials of independent random variables. In *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms*, pages 437–446. SIAM, 2012.

[43] Amir Shpilka and Amir Yehudayoff. Arithmetic circuits: A survey of recent results and open questions. *Foundations and Trends® in Theoretical Computer Science*, 5(3–4):207–388, 2010.

[44] Terence Tao and Van H Vu. Inverse Littlewood-Offord theorems and the condition number of random discrete matrices. *Annals of Mathematics*, pages 595–632, 2009.

[45] Terence Tao and Van Vu. Random matrices: The distribution of the smallest singular values. *Geometric And Functional Analysis*, 20(1):260–297, 2010.