Electronic Colloquium on Computational Complexity, Revision 2 of Report No. 59 (2014)

Sum-of-Squares Proofs and the Quest toward Optimal Algorithms

Boaz Barak and David Steurer

Abstract. In order to obtain the best-known guarantees, algorithms are traditionally tailored to the particular problem we want to solve. Two recent developments, the *Unique Games Conjecture (UGC)* and the *Sum-of-Squares (SOS) method*, surprisingly suggest that this tailoring is not necessary and that a single efficient algorithm could achieve best possible guarantees for a wide range of different problems.

The *Unique Games Conjecture (UGC)* is a tantalizing conjecture in computational complexity, which, if true, will shed light on the complexity of a great many problems. In particular this conjecture predicts that a *single concrete algorithm* provides optimal guarantees among all efficient algorithms for a large class of computational problems.

The *Sum-of-Squares (SOS) method* is a general approach for solving systems of polynomial constraints. This approach is studied in several scientific disciplines, including real algebraic geometry, proof complexity, control theory, and mathematical programming, and has found applications in fields as diverse as quantum information theory, formal verification, game theory and many others.

We survey some connections that were recently uncovered between the Unique Games Conjecture and the Sum-of-Squares method. In particular, we discuss new tools to rigorously bound the running time of the SOS method for obtaining approximate solutions to hard optimization problems, and how these tools give the potential for the sum-of-squares method to provide new guarantees for many problems of interest, and possibly to even refute the UGC.

Mathematics Subject Classification (2010). Primary 68Q25; Secondary 90C22.

Keywords. Sum of squares, semidefinite programming, unique games conjecture, small-set expansion

1. Introduction

A central mission of theoretical computer science is to understand which computational problems can be solved efficiently, which ones cannot, and what it is about a problem that makes it easy or hard. To illustrate these kind of questions, let us consider the following parameters of an undirected d-regular graph[1] $G = (V, E)$:

[1] An undirected d-regular graph $G = (V, E)$ consists of a set of *vertices* V, which we sometimes identify with the set $[n] = \{1, \ldots, n\}$ for some integer n, and a set of *edges* E, which are 2-element subsets of V, such that every vertex is part of exactly d edges. The assumption that G is regular is not important and made chiefly for notational simplicity. For vertex sets $S, T \subseteq V$, we let

ISSN 1433-8092

- The *smallest connected component of* G is the size of the smallest non-empty set $S \subseteq V$ such that $E(S, V \setminus S) = \emptyset$.
- The *independent-set number of* G is the size of the largest set $S \subseteq V$ such that $E(S, S) = \emptyset$.
- The *(edge) expansion*[2] *of* G, denoted ϕ_G, is the minimum *expansion* $\phi_G(S)$ of a vertex set $S \subseteq V$ with size $1 \leq |S| \leq |V|/2$, where

$$\phi_G(S) = \frac{|E(S, V \setminus S)|}{d|S|}\,.$$

  The expansion $\phi_G(S)$ measures the probability that a step of the random walk on G leaves S conditioned on starting in S.

All these parameters capture different notions of well-connectedness of the graph G. Computing these can be very useful in many of the settings in which we use graphs to model data, whether it is communication links between servers, social connections between people, genes that are co-expressed together, or transitions between states of a system.

The computational complexity of the first two parameters is fairly well understood. The smallest connected component is easy to compute in time linear in the number $n = |V|$ of vertices by using, for example, breadth-first search from every vertex in the graph. The independent-set number is **NP**-hard to compute, which means that, assuming the widely believed conjecture that $\mathbf{P} \neq \mathbf{NP}$, it cannot be computed in time polynomial in n. In fact, under stronger (but still widely believed) quantitative versions of the $\mathbf{P} \neq \mathbf{NP}$ conjecture, for every k it is infeasible to decide whether or not the maximum independent set is larger than k in time $n^{o(k)}$ [DF95, CHKX06] and hence we cannot significantly beat the trivial $O(n^k)$-time algorithm for this problem. Similarly, while we can approximate the independent-set number trivially within a factor of n, assuming such conjectures, there is no polynomial-time algorithm to approximate it within a factor of $n^{1-\varepsilon(n)}$ where $\varepsilon(n)$ is some function tending to zero as n grows [Hås96, Kho01].

So, connectivity is an easy problem and independent set a hard one, but what about expansion? Here the situation is more complicated. We know that we can't efficiently compute ϕ_G exactly, and we can't even get an arbitrarily good approximation [AMS11], but we actually do have efficient algorithms with non-trivial approximation guarantees for ϕ_G. Discrete versions of *Cheeger's inequality* [Che70, Dod84, AM85, Alo86] yield such an estimate, namely

$$\tfrac{d-\lambda_2}{2d} \leq \phi_G \leq 2\sqrt{\tfrac{d-\lambda_2}{2d}}\,, \tag{1}$$

$E(S, T)$ denote the set of edges $\{s, t\} \in E$ with $s \in S$ and $t \in T$.

[2] The expansion of a graph is closely related to other quantities, known as *isoperimetric constant*, *conductance* or *sparsest cut*. These quantities are not identical but are the same up to scaling and a multiplicative factor of at most 2. Hence, they are computationally equivalent for our purposes. We also remark that expansion is often not normalized by the degree. However for our purposes this normalization is useful.

where $\lambda_2(G)$ denotes the (efficiently computable) second largest eigenvalue of the G's adjacency matrix.[3] In particular, we can use (1) to efficiently distinguish between graphs with ϕ_G close to 0 and graphs with ϕ_G bounded away from 0. But can we do better? For example, could we efficiently compute a quantity c_G such that $c_G \leq \phi_G \leq O(c_G^{0.51})$? We simply don't know.[4]

This is not an isolated example, but a pattern that keeps repeating. Over the years, computer scientists have developed sophisticated tools to come up with algorithms on one hand, and hardness proofs showing the limits of efficient algorithms on the other hand. But those two rarely match up. Moreover, the cases where we do have tight hardness results are typically in settings, such as the independent set problem, where there is no way to significantly beat the trivial algorithm. In contrast, for problems such as computing expansion, where we already know of an algorithm giving non-trivial guarantees, we typically have no proof that this algorithm is *optimal*. In other words, the following is a common theme:

> *If you already know an algorithm with non-trivial approximation guarantees for a problem, it's very hard to rule out that cleverer algorithms couldn't get even better guarantees.*

In 2002, Subhash Khot formulated a conjecture, known as the *Unique Games Conjecture (UGC)* [Kho02]. A large body of follow up works has shown that this conjecture (whose description is deferred to Section 1.1 below) implies many hardness results that overcome the above challenge and match the best-known algorithms even in cases when they achieve non-trivial guarantees. In fact, beyond just resolving particular questions, this line of works obtained far-reaching complementary *meta algorithmic* and *meta hardness* results. By this we mean results that give an efficient *meta algorithm* $\mathcal{A}$ (i.e., an algorithm that can be applied to a family of problems, and not just a single one) that is *optimal* within a broad domain $\mathcal{C}$, in the sense that (assuming the UGC) there is no polynomial-time algorithm that performs better than $\mathcal{A}$ on any problem in $\mathcal{C}$. It is this aspect of the Unique Games Conjecture result that we find most exciting, and that shows promise of going beyond the current state where the individual algorithmic and hardness results form "isolated islands of knowledge surrounded by a sea of ignorance"[5] into a more unified theory of complexity.

The meta-algorithm that the UGC predicts to be optimal is based on *semidefinite programming* and it uses this technique in a very particular and quite restricted way. (In many settings, this meta-algorithm can be implemented in near-linear time [Ste10].) We will refer to this algorithm as the *UGC meta-algorithm*. It can be viewed as a common generalization of several well known algorithms, including those that underlie Cheeger's Inequality, Grothendieck's Inequality [Gro53],

[3] The *adjacency matrix* of a graph G is the $|V| \times |V|$ matrix A with 0/1 entries such that $A_{u,v} = 1$ iff $\{u, v\} \in E$.

[4] As we will mention later, there are algorithms to approximate ϕ_G up to factors depending on the number n of vertices, which give better guarantees than (1) for graphs where ϕ_G is sufficiently small as a function of n.

[5] Paraphrasing John Wheeler.

the Goemans–Williamson MAX CUT algorithm [GW95], and the Lovász ϑ function [Lov79]. As we've seen for the example of Cheeger's Inequality, in many of those settings this meta-algorithm gives *non-trivial approximation guarantees* which are the best known, but there are no hardness results ruling out the existence of better algorithms. The works on the UGC has shown that this conjecture (and related ones) imply that this meta-algorithm is *optimal* for a vast number of problems, including all those examples above. For example, a beautiful result of Raghavendra [Rag08] showed that for every constraint-satisfaction problem (a large class of problems that includes many problems of interest such as MAX k-SAT, k-COLORING, and MAX-CUT), the UGC meta-algorithm gives the best estimate on the maximum possible fraction of constraints one can satisfy. Similarly, the UGC (or closely related variants) imply there are no efficient algorithms that give a better estimate for the sparsest cut of a graph than the one implied by Cheeger's Inequality [RST12] and no better efficient estimate for the maximum correlation of a matrix with ± 1-valued vectors than the one given by Grothendieck's Inequality.[6] To summarize:

> *If true, the Unique Games Conjecture tells us not only* which *problems in a large class are easy and which are hard, but also* why *this is the case. There is a* single unifying reason*, captured by a concrete meta-algorithm, that explains all the easy problem in this class. Moreover, in many cases where this meta-algorithm already gives non-trivial guarantees, the UGC implies that no further efficient improvements are possible.*

All this means that the Unique Games Conjecture is certainly a very attractive proposition, but the big question still remains unanswered—is this conjecture actually true? While some initial results supported the UGC, more recent works, although still falling short of disproving the conjecture, have called it into question. In this survey we discuss the most promising current approach to refute the UGC, which is based on the *Sum of Squares (SOS) method* [Sho87, Nes00, Par00, Las01]. The SOS method could potentially refute the Unique Games Conjecture by beating the guarantees of the UGC meta-algorithm on problems on which the conjecture implies the latter's optimality. This of course is interesting beyond the UGC, as it means we would be able to improve the known guarantees for many problems of interest. Alas, analyzing the guarantees of the SOS method is a very challenging problem, and we still have relatively few tools to do so. However, as we will see, we already know that at least in some contexts, the SOS method can yield better results than what was known before. The SOS method is itself a meta algorithm, so even if it turns out to refute the UGC, this does not mean we need to give up on the notion of explaining the complexity of wide swaths of problems via a single algorithm; we may just need to consider a different algorithm. To summarize, regardless of whether it refutes the UGC or not, understanding the power of the SOS

[6] See [RS09b] for the precise statement of Grothendieck's Inequality and this result. Curiously, the UGC implies that Grothendieck's Inequality yields the best efficient approximation factor for the correlation of a matrix with ± 1-valued vectors even though we don't actually know the numerical value of this factor (known as Grothendieck's constant).

method is an exciting research direction that could advance us further towards the goal of a unified understanding of computational complexity.

1.1. The UGC and SSEH conjectures. Instead of the Unique Games Conjecture, in this survey we focus on a related conjecture known as the *Small-Set Expansion Hypothesis (SSEH)* [RS10]. The SSEH implies the UGC [RS10], and while there is no known implication in the other direction, there are several results suggesting that these two conjectures are probably equivalent [RS10, RST10, RS09a, ABS10, BBH$^+$12]. At any rate, most (though not all) of what we say in this survey applies equally well to both conjectures, but the SSEH is, at least in our minds, a somewhat more natural and simpler-to-state conjecture.

Recall that for a d-regular graph $G = (V, E)$ and a vertex set $S \subseteq V$, we defined its expansion as $\phi_G(S) = |E(S, V \setminus S)|/(d|S|)$. By Cheeger's inequality (1), the second largest eigenvalue yields a non-trivial approximation for the minimum expansion $\phi_G = \min_{1 \leq |S| \leq |V|/2} \phi_G(S)$, but it turns out that eigenvalues and similar methods do not work well for the problem of approximating the minimum expansion of smaller sets. The Small-Set Expansion Hypothesis conjectures that this problem is inherently difficult.

Conjecture 1.1 (Small-Set Expansion Hypothesis [RS10]). *For every $\varepsilon > 0$ there exists $\delta > 0$ such that given any graph $G = (V, E)$, it is* **NP***-hard to distinguish between the case* ***(i)*** *that there exists a subset $S \subseteq V$ with $|S| = \delta|V|$ such that $\phi_G(S) \leq \varepsilon$ and the case* ***(ii)*** *that $\phi_G(S) \geq 1 - \varepsilon$ for every S with $|S| \leq \delta|V|$.*

As mentioned above, the SSEH implies that (1) yields an optimal approximation for ϕ_G. More formally, assuming the SSEH, there is some absolute constant $c > 0$ such that for every $\phi \geq 0$, it is **NP**-hard to distinguish between the case that a given graph G satisfies $\phi_G \leq \phi$ and the case that $\phi_G \geq c\sqrt{\phi}$ [RST12]. Given that the SSEH conjectures the difficulty of approximating expansion, the reader might not be so impressed that it also implies the optimality of Cheeger's Inequality. However, we should note that the SSEH merely conjectures that the problem becomes harder as δ becomes smaller, without postulating any quantitative relation between δ and ε, and so it is actually surprising (and requires a highly non-trivial proof) that it implies such quantitatively tight bounds. Even more surprising is that (through its connection with the UGC) the SSEH implies tight hardness result for a host of other problems, including every constraint satisfaction problem, Grothendieck's problem, and many others, which a priori seem to have nothing to do with graph expansion.

Remark 1.2. While we will stick to the SSEH in this survey, for completeness we present here the definition of the Unique Games Conjecture. We will not use this definition in the proceeding and so the reader can feel free to skip this remark. The UGC can be thought of as a more structured variant of the SSEH where we restrict to graphs and sets that satisfy some particular properties. Because we restrict both the graphs and the sets, a priori it is not clear which of these conjectures should be stronger. However it turns out that the SSEH implies the UGC [RS10]. It is an open problem whether the two conjectures are equivalent,

though the authors personally suspect that this is the case. We say that an n-vertex graph $G = (V, E)$ is *δ-structured* if there is a partition of V into δn sets $V_1, \ldots, V_{\delta n}$ each of size $1/\delta$, such that for every $i \neq j$, either $E(V_i, V_j) = \emptyset$ or $E(V_i, V_j)$ is a *matching* (namely for every $u \in V_i$ there is exactly one $v \in V_j$ such that $\{u, v\} \in E$). We say a set $S \subseteq V$ is *δ-structured* if $|S \cap V_i| = 1$ for all i (and so in particular, $|S| = \delta n$). The Unique Games Conjecture states that for every $\varepsilon > 0$ there exists a $\delta > 0$ such that it is **NP** hard, given a δ-structured G, to distinguish between the case **(i)** that there exists a δ-structured S such that $\phi_G(S) \leq \varepsilon$ and the case **(ii)** that every δ-structured S satisfies $\phi_G(S) \geq 1 - \varepsilon$. The conjecture can also be described in the form of so-called "two prover one round games" (hence its name); see Khot's surveys [Kho10a, Kho10b].

1.2. Organization of this survey and further reading. In the rest of this survey we describe the Sum of squares algorithm, some of its applications, and its relation to the Unique Games and Small-Set Expansion Conjectures. We start by defining the Sum of Squares algorithm, and how it relates to classical questions such as Hilbert 17^{th} problem. We will demonstrate how the SOS algorithm is used, and its connection to the UGC/SSEH, by presenting Cheeger's Inequality (1) as an instance of this algorithm. The SSEH implies that the SOS algorithm cannot yield better estimates to ϕ_G than those obtained by (1). While we do not know yet whether this is true or false, we present two different applications where the SOS does beat prior works— finding a planted sparse vector in a random subspace, and *sparse coding*— learning a set of vectors A given samples of random sparse linear combinations of vectors in A. We then discuss some of the evidence for the UGC/SSEH, how this evidence is challenged by the SOS algorithm and the relation between the UGC/SSEH and the problem of (approximately) finding sparse vectors in arbitrary (not necessarily random) subspaces. Much of our discussion is based on the papers [ABS10, BGH$^+$12, BBH$^+$12, BKS14b, BKS14a]. See also [Bar12, Bar14b, Bar14a] for informal overviews of some of these issues.

For the reader interested in learning more about the Unique Games Conjecture, there are three excellent surveys on this topic. Khot's CCC survey [Kho10b] gives a fairly comprehensive overview of the state of knowledge on the UGC circa 2010, while his ICM survey [Kho10a] focuses on some of the techniques and connections that arose in the works around the UGC. Trevisan [Tre12] gives a wonderfully accessible introduction to the UGC, using the Max-Cut problem as a running example to explain in detail the UGC's connection to semidefinite programming. As a sign of how rapidly research in this area is progressing, this survey is almost entirely disjoint from [Kho10a, Kho10b, Tre12]. While the former surveys mostly described the implications of the UGC for obtaining very strong hardness and "meta hardness" results, the current manuscript is focused on the question of whether the UGC is actually true, and more generally understanding the power of the SOS algorithm to go beyond the basic LP and SDP relaxations.

Our description of the SOS algorithm barely scratches the surface of this fascinating topic, which has a great many applications that have nothing to do with the UGC or even approximation algorithms at large. The volume [BPT13] and

the monograph [Lau09] are good sources for some of these topics. The SOS algorithm was developed in slightly different forms by several researchers, including Shor [Sho87], Nesterov [Nes00], Parrilo [Par00], and Lasserre [Las01]. It can be viewed as a strengthening of other "meta-algorithms" proposed by [SA90, LS91] (also known as linear and semi-definite programming hierarchies).[7] Our description of the SOS meta algorithm follows Parrilo's, while the description of the dual algorithm follows Lasserre, although we use the pseudoexpectation notation introduced in [BBH$^+$12] instead of Lasserre's notion of "moment matrices". The Positivstellensatz/SOS proof system was first studied by Grigoriev and Vorobjov [GV01] and Grigoriev [Gri01] proved some degree lower bounds for it, that were later rediscovered and expanded upon by [Sch08, Tul09]. All these are motivated by the works in real geometry related to Hilbert's 17$^{\text{th}}$ problem; see Reznick's survey [Rez00] for more on this research area. One difference between our focus here and much of the other literature on the SOS algorithm is that we are content with proving that the algorithm supplies an *approximation* to the true quantity, rather than exact convergence, but on the other hand are much more stringent about using only very low degree (preferably constant or polylogarithmic in the number of variables).

2. Sums of Squares Proofs and Algorithms

One of the most common ways of proving that a quantity is non-negative is by expressing it as a *Sum of Squares* (SOS). For example, we can prove the Arithmetic-Mean Geometric-Mean inequality $ab \leq a^2/2 + b^2/2$ by the identity $a^2 + b^2 - 2ab = (a-b)^2$. Thus a natural question, raised in the late 19$^{\text{th}}$ century, was whether *any* non-negative (possibly multivariate) polynomial can be written as a sum of squares of polynomials. This was answered negatively by Hilbert in 1888, who went on to ask as his 17$^{\text{th}}$ problem whether any such polynomial can be written as a sum of squares of *rational* functions. A positive answer was given by Artin [Art27], and considerably strengthened by Krivine and Stengle. In particular, the following theorem is a corollary of their results, which captures much of the general case.

Theorem 2.1 (Corollary of the Positivstellensatz [Kri64, Ste74]). *Let $P_1, \ldots, P_m \in \mathbb{R}[x] = \mathbb{R}[x_1, \ldots, x_n]$ be multivariate polynomials. Then, the system of polynomials equations $\mathcal{E} = \{P_1 = 0, \ldots, P_m = 0\}$ has no solution over $\mathbb{R}^n$ if and only if, there exists polynomials $Q_1, \ldots, Q_m \in \mathbb{R}[x]$ such that $S \in \mathbb{R}[x]$ is a* sum of squares *of polynomials and*

$$-1 = S + \sum Q_i \cdot P_i \,. \tag{2}$$

We say that the polynomials $S, Q_1, \ldots, Q_m$ in the conclusion of the theorem form an *SOS proof* refuting the system of polynomial equations[8] $\mathcal{E}$. Clearly the

[7] See [Lau03] for a comparison.

[8] In this survey we restrict attention to polynomial *equalities* as opposed to *inequalities*, which turns out to be without loss of generality for our purposes. If we have a system of polynomial inequalities $\{P_1 \geq 0, \ldots, P_m \geq 0\}$ for $P_i \in \mathbb{R}[x]$, the Positivstellensatz certificates of infeasibility take the form $-1 = \sum_{\alpha \subseteq [n]} Q_\alpha P_\alpha$, where each $Q_\alpha \in \mathbb{R}[X]$ is a sum of squares and $P_\alpha = \prod_{i \in \alpha} P_i$.

existence of such polynomials implies that $\mathcal{E}$ is unsatisfiable—the interesting part of Theorem 2.1 is the other direction. We say that a SOS refutation $S_1, Q_1, ..., Q_m$ has *degree* ℓ if the maximum degree of the polynomials $Q_i P_i$ involved in the proof is at most ℓ [GV01]. By writing down the coefficients of these polynomials, we see that a degree-ℓ SOS proof can be written using $mn^{O(\ell)}$ numbers.[9]

In the following lemma, we will prove a special case of Theorem 2.1, where the solution set of $\mathcal{E}$ is a subset of the hypercube $\{\pm 1\}^n$. Here, the degree of SOS refutations is bounded by $2n$. (This bound is not meaningful computationally because the size of degree-$\Omega(n)$ refutations is comparable to the number of points in $\{\pm 1\}^n$.)

Lemma 2.2. *Let $\mathcal{E} = \{P_0 = 0, x_1^2 - 1 = 0, \ldots, x_n^2 - 1 = 0\}$ for some $P_0 \in \mathbb{R}[x]$. Then, either the system $\mathcal{E}$ is satisfiable or it has a degree-$2n$ SOS refutation.*

Proof. Suppose the system is not satisfiable, which means that $P_0(x) \neq 0$ for all $x \in \{\pm 1\}^n$. Since $\{\pm 1\}^n$ is a finite set, we may assume $P_0^2 \geq 1$ over $\{\pm 1\}^n$. Now interpolate the real-valued function $\sqrt{P_0^2 - 1}$ on $\{\pm 1\}^n$ as a multilinear (and hence degree at most n) polynomial in $R \in \mathbb{R}[x]$. Then, $P_0^2 - 1 - R^2$ is a polynomial of degree at most $2n$ that vanishes over $\{\pm 1\}^n$. (Since we can replace x_i^2 by 1 in any monomial, we can assume without loss of generality that P_0 is multilinear and hence has degree at most n.) This means that we can write $P_0^2 - 1 - R^2$ in the form $\sum_{i=1}^n Q_i \cdot (x_i^2 - 1)$ for polynomials Q_i with $Q_i \leq \deg 2n - 2$. (This fact can be verified either directly or by using that $x_1^2 - 1, \ldots, x_n^2 - 1$ is a Gröbner basis for $\{\pm 1\}^n$.) Putting things together, we see that $-1 = R^2 + (-P_0) \cdot P_0 + \sum_{i=1}^n Q_i \cdot (x_i^2 - 1)$, which is a SOS refutation for $\mathcal{E}$ of the form in Theorem 2.1. □

2.1. From proofs to algorithms.

The Sum of Squares algorithm is based on the following theorem, which was discovered in different forms by several researchers:

Theorem 2.3 (SOS Theorem [Sho87, Nes00, Par00, Las01], informally stated). *If there is a degree-ℓ SOS proof refuting $\mathcal{E} = \{P_1 = 0, \ldots, P_m = 0\}$, then such a proof can be found in $mn^{O(\ell)}$ time.*

Proof sketch. We can view a degree-ℓ SOS refutation $-1 = S + \sum_i Q_i P_i$ for $\mathcal{E}$ as a system of linear equations in $mn^{O(\ell)}$ variables corresponding to the coefficients of the unknown polynomials $S, Q_1, \ldots, Q_m$. We only need to incorporate the non-linear constraint that S is a sum of squares. But it is not hard to see that a degree-ℓ polynomial S is a sum of squares if and only if there exists a positive-semidefinite matrix M such that $S = \sum_{\alpha,\alpha'} M_{\alpha,\alpha'} x^\alpha x^{\alpha'}$, where α and α' range over all monomials x^α and $x^{\alpha'}$ of degree at most $\ell/2$. Thus, the task of finding a

However, we can transform inequalities $\{P_i \geq 0\}$ to equivalent equalities $\{P_i' = P_i - y_i^2 = 0\}$, where $y_1, \ldots, y_m$ are fresh variables. This transformation makes it only easier to find certificates, because $\sum_{\alpha \subseteq [n]} Q_\alpha P_\alpha = S' + \sum_i Q_i' P_i'$ for $S' = \sum_{\alpha \subseteq [n]} Q_\alpha y_\alpha^2$, where $y_\alpha = \prod_{i \in \alpha} y_i$. It also follows that the transformation can only reduce the degree of SOS refutations.

[9] It can be shown that the decomposition of S into sums of squares will not require more than n^ℓ terms; also in all the settings we consider, there are no issues of accuracy in representing real numbers, and so a degree ℓ-proof can be written down using $mn^{O(\ell)}$ bits.

degree-ℓ SOS refutation reduces to the task of solving linear systems of equations with the additional constraint that matrix formed by some of the variables is positive-semidefinite. *Semidefinite programming* solves precisely this task and is computationally efficient.[10] □

Remark 2.4 (*What does "efficient" mean?*). In the applications we are interested in, the number of variables n corresponds to our "input size". The equation systems $\mathcal{E}$ we consider can always be solved via a "brute force" algorithm running in $\exp(O(n))$ time, and so degree-ℓ SOS proofs become interesting when ℓ is much smaller than n. Ideally we would want $\ell = O(1)$, though $\ell = \text{polylog}(n)$ or even, say, $\ell = \sqrt{n}$, is still interesting.

Theorem 2.3 yields the following *meta algorithm* that can be applied on any problem of the form

$$\min_{x \in \mathbb{R}^n \,:\, P_1(x)=\cdots=P_m(x)=0} P_0(x) \tag{3}$$

where $P_0, P_1, \ldots, P_m \in \mathbb{R}[x]$ are polynomials. The algorithm is parameterized by a number ℓ called its *degree* and operates as follows:

The degree-ℓ Sum-of-Squares Algorithm

Input: Polynomials $P_0, \ldots, P_m \in \mathbb{R}[x]$
Goal: Estimate $\min P_0(x)$ over all $x \in \mathbb{R}^n$ such that $P_1(x) = \ldots = P_m(x) = 0$
Operation: Output the smallest value $\varphi^{(\ell)}$ such that there does *not* exist a degree-ℓ SOS proof refuting the system,

$$\{P_0 = \varphi^{(\ell)}, P_1 = 0, \ldots, P_m(x) = 0\}\,.^{11}$$

We call $\varphi^{(\ell)}$ the *degree-ℓ SOS estimate* for (3), and by Theorem 2.3 it can be computed in $n^{O(\ell)}$ time. For the actual minimum value φ of (3), the corresponding system of equations $\{P_0 = \varphi, P_1 = 0, \ldots, P_m = 0\}$ is satisfiable, and hence in particular cannot be refuted by an SOS proof. Thus, $\varphi^{(\ell)} \leq \varphi$ for any ℓ. Since higher degree proofs are more powerful (in the sense that they can refute more equations), it holds that

$$\varphi^{(2)} \leq \varphi^{(4)} \leq \varphi^{(6)} \leq \cdots \leq \min_{x \in \mathbb{R}^n \,:\, P_1(x)=\cdots=P_m(x)=0} P_0(x)\,.$$

(We can assume degrees of SOS proofs to be even.) As we've seen in Lemma 2.2, for the typical domains we are interested in Computer Science, such as when the

[10] In this survey we ignore issues of numerical accuracy which turn out to be easily handled in our setting.

[11] As in other cases, we are ignoring here issues of numerical accuracy. Also, we note that when actually executing this algorithm, we will not need to check all the (uncountably many) values $\varphi^{(\ell)} \in \mathbb{R}$, but it suffices to enumerate over a sufficiently fine discretization of the interval $[-M, +M]$ for some number M depending on the polynomials $P_0, \ldots, P_m$. This step can be carried out in polynomial time in all the settings we consider.

set of solutions of $\{P_1 = 0, \ldots, P_m = 0\}$ is equal to $\{\pm 1\}^n$, this sequence is finite in the sense that $\varphi^{(2n)} = \min_{x \in \{\pm 1\}^n} P_0(x)$.

The SOS algorithm uses semidefinite programming in a much more general way than many previous algorithms such as [Lov79, GW95]. In fact, the UGC meta-algorithm is the same as the base case (i.e., $\ell = 2$) of the SOS algorithm.

Recall that the UGC and SSEH imply that in many settings, one cannot improve on the approximation guarantees of the UGC meta-algorithm without using $\exp(n^{\Omega(1)})$ time. Thus in particular, if those conjectures are true then in those settings, using the SOS meta algorithm with degree, say, $\ell = 10$ (or even $\ell = \text{polylog}(n)$ or $\ell = n^{o(1)}$) will not yield significantly better guarantees than $\ell = 2$.

Remark 2.5 (Comparison with local-search based algorithms). Another approach to optimize over non-linear problems such as (3) is to use local-search algorithms such as gradient descent that make local improvement steps, e.g., in the direction of the gradient, until a local optimum is reached. One difference between such local search algorithms and the SOS algorithm is that the latter sometimes succeeds in optimizing highly non-convex problems that have exponential number of local optima. As an illustration, consider the polynomial $P(x) = n^4 \sum_{i=1}^n (x_i^2 - x_i)^2 + (\sum_{i=1}^n x_i)^2$. Its unique global minimum is the point $x = 0$, but it is not hard to see that it has an exponential number of local minima (for every $x \in \{0,1\}^n$, $P(x) < P(y)$ for every y with $\|y - x\| \in [1/n, 2/n]$, and so there must be a local minima in the ball of radius $1/n$ around x). Hence, gradient descent or other such algorithms are extremely likely to get stuck in one of these suboptimal local minima. However, since P is in fact a sum of squares with constant term 0, the degree-4 SOS algorithm will output P's correct global minimum value.

2.2. Pseudodistributions and pseudoexpectations.

Suppose we want to show that the level-ℓ SOS meta-algorithm achieves a good approximation of the minimum value of P_0 over the set $\mathcal{Z} = \{x \in \mathbb{R}^n \mid P_1(x) = \cdots = P_m(x) = 0\}$ for a particular kind of polynomials $P_0, P_1, \ldots, P_m \in \mathbb{R}[x]$. Since the estimate $\varphi^{(\ell)}$ always lower bounds this quantity, we are to show that

$$\min_{\mathcal{Z}} P_0 \leq f(\varphi^{(\ell)}) \tag{4}$$

for some particular function f (satisfying $f(\varphi) \geq \varphi$) which captures our approximation guarantee. (E.g., a factor c approximation corresponds to the function $f(\varphi) = c\varphi$.)

If we expand out the definition of $\varphi^{(\ell)}$, we see that to prove Equation (4) we need to show that for every φ if there does not exists a degree-ℓ proof that $P_0(x) \neq \varphi$ for all $x \in \mathcal{Z}$, then there exists an $x \in \mathcal{Z}$ such that $P_0(x) \leq f(\varphi)$. So, to prove a result of this form, we need to find ways to use the *non-existence* of a proof. Here, *duality* is useful.

> *Pseudodistributions are the dual object to SOS refutations, and hence the* non-existence *of a refutation implies the* existence *of a pseudodistribution.*

We now elaborate on this, and explain both the definition and intuition behind pseudodistributions. In Section 3 we will give a concrete example, by showing how one can prove that degree-2 SOS proofs capture Cheeger's Inequality using such an argument. Results such as the analysis of the Goemans-Williamson Max Cut algorithm [GW95], and the proof of Grothendieck's Inequality [Gro53] can be derived using similar methods.

Definition 2.6. Let $\mathbb{R}[x]_\ell$ denote the set of polynomials in $\mathbb{R}[x]$ of degree at most ℓ. A *degree-ℓ pseudoexpectation operator for* $\mathbb{R}[x]$ is a linear operator $\mathcal{L}$ that maps polynomials in $\mathbb{R}[x]_\ell$ into $\mathbb{R}$ and satisfies that $\mathcal{L}(1) = 1$ and $\mathcal{L}(P^2) \geq 0$ for every polynomial P of degree at most $\ell/2$.

The term pseudoexpectation stems from the fact that for every distribution $\mathcal{D}$ over $\mathbb{R}^n$, we can obtain such an operator by choosing $\mathcal{L}(P) = \mathbb{E}_{\mathcal{D}} P$ for all $P \in \mathbb{R}[x]$. Moreover, the properties $\mathcal{L}(1) = 1$ and $\mathcal{L}(P^2) \geq 0$ turn out to capture to a surprising extent the properties of distributions and their expectations that we tend to use in proofs. Therefore, we will use a notation and terminology for such pseudoexpectation operators that parallels the notation we use for distributions. In fact, all of our notation can be understood by making the thought experiment that there exists a distribution as above and expressing all quantities in terms of low-degree moments of that distribution (so that they also make sense if we only have a pseudoexpectation operator that doesn't necessarily correspond to a distribution).

In the following, we present the formal definition of our notation. We denote pseudoexpectation operators as $\tilde{\mathbb{E}}_{\mathcal{D}}$, where $\mathcal{D}$ acts as index to distinguish different operators. If $\tilde{\mathbb{E}}_{\mathcal{D}}$ is a degree-ℓ pseudoexpectation operator for $\mathbb{R}[x]$, we say that $\mathcal{D}$ is a *degree-ℓ pseudodistribution* for the indeterminates x. In order to emphasize or change indeterminates, we use the notation $\tilde{\mathbb{E}}_{y\sim\mathcal{D}} P(y)$. In case we have only one pseudodistribution $\mathcal{D}$ for indeterminates x, we denote it by $\{x\}$. In that case, we also often drop the subscript for the pseudoexpectation and write $\tilde{\mathbb{E}}\, P$ for $\tilde{\mathbb{E}}_{\{x\}} P$.

We say that a degree-ℓ pseudodistribution $\{x\}$ satisfies a system of polynomial equations $\{P_1 = 0, \ldots, P_m = 0\}$ if $\tilde{\mathbb{E}}\, Q \cdot P_i = 0$ for all $i \in [m]$ and all polynomials $Q \in \mathbb{R}[x]$ with $\deg Q \cdot P_i \leq \ell$. We also say that $\{x\}$ satisfies the constraint $\{P(x) \geq 0\}$ if there exists some sum-of-squares polynomial $S \in \mathbb{R}[x]$ such that $\{x\}$ satisfies the polynomial equation $\{P = S\}$. It is not hard to see that if $\{x\}$ was an actual distribution, then these definitions imply that all points in the support of the distribution satisfy the constraints. We write $P \succcurlyeq 0$ to denote that P is a sum of squares of polynomials, and similarly we write $P \succcurlyeq Q$ to denote $P - Q \succcurlyeq 0$.

The duality between SOS proofs and pseudoexpectations is expressed in the following theorem. We say that a system $\mathcal{E}$ of polynomial equations is *explicitly bounded* if there exists a linear combination of the constraints in $\mathcal{E}$ that has the form $\{\sum_i x_i^2 + S = M\}$ for $M \in \mathbb{R}$ and $S \in \mathbb{R}[x]$ a sum-of-squares polynomial. (Note that in this case, every solution $x \in \mathbb{R}^n$ of the system $\mathcal{E}$ satisfies $\sum_i x_i^2 \leq M$.)

Theorem 2.7. *Let $\mathcal{E} = \{P_1 = 0, \ldots, P_m = 0\}$ be a set of polynomial equations with $P_i \in \mathbb{R}[x]$. Assume that $\mathcal{E}$ is explicitly bounded in the sense above. Then, exactly*

one of the following two statements holds: (a) there exists a degree-ℓ SOS proof refuting $\mathcal{E}$, or (b) there exists a degree-ℓ pseudodistribution $\{x\}$ that satisfies $\mathcal{E}$.

Proof. First, suppose there exists a degree-ℓ refutation of the system $\mathcal{E}$, i.e., there exists polynomials $Q_1, \ldots, Q_m \in \mathbb{R}[x]$ and a sum-of-squares polynomial $R \in \mathbb{R}[x]$ so that $-1 = R + \sum_i Q_i P_i$ and $\deg Q_i P_i \leq \ell$. Let $\{x\}$ be any pseudodistribution. We are to show that $\{x\}$ does not satisfy $\mathcal{E}$. Indeed, $\tilde{\mathbb{E}} \sum_i Q_i P_i = -\tilde{\mathbb{E}}\, 1 - \tilde{\mathbb{E}}\, R \leq -1$, which means that $\tilde{\mathbb{E}}\, Q_i P_i \neq 0$ for at least one $i \in [m]$. Therefore, $\{x\}$ does not satisfy $\mathcal{E}$.

Next, suppose there does not exist a degree-ℓ refutation of the system $\mathcal{E}$. We are to show that there exists a pseudodistribution that satisfies $\mathcal{E}$. Let $\mathcal{C}$ be the cone of all polynomials of the form $R + \sum_i Q_i P_i$ for sum-of-squares R and polynomials Q_i with $\deg Q_i P_i \leq \ell$. Since $\mathcal{E}$ does not have a degree-ℓ refutation, the constant polynomial -1 is not contained in $\mathcal{C}$. We claim that from our assumption that the system $\mathcal{E}$ is explicitly bounded it follows that -1 also cannot lie on the boundary of $\mathcal{C}$. Assuming this claim, the hyperplane separation theorem implies that there exists a linear form L such that $L(-1) < 0$ but $L(P) \geq 0$ for all $P \in \mathcal{C}$. By rescaling, we may assume that $L(1) = 1$. Now this linear form satisfies all conditions of a pseudoexpectation operator for the system $\mathcal{E}$.

Proof of claim. We will show that if -1 lies on the boundary of $\mathcal{C}$, then also $-1 \in \mathcal{C}$. If -1 is on the boundary of $\mathcal{C}$, then there exists a polynomial $P \in \mathbb{R}[X]_\ell$ such that $-1 + \varepsilon P \in \mathcal{C}$ for all $\varepsilon > 0$ (using the convexity of $\mathcal{C}$). Since $\mathcal{E}$ is explicitly bounded, for every polynomial $P \in \mathbb{R}[X]_\ell$, the cone $\mathcal{C}$ contains a polynomial of form $N - P - R$ for a sum-of-square R and a number N. (Here, the polynomial $N - P - R \in \mathcal{C}$ is a certificate that $P \leq N$ over the solution set of $\mathcal{E}$. Such a certificate is easy to obtain when $\mathcal{E}$ is explicitly bounded. We are omitting the details.) At this point, we see that -1 is a nonnegative combination of the polynomials $-1 + \varepsilon P$, $N - P - R$, and R for $\varepsilon < 1/N$. Since these polynomials are contained in $\mathcal{C}$, their nonnegative combination -1 is also contained in the cone $\mathcal{C}$. □

Recipe for using pseudoexpectations algorithmically. In many applications we will use the following dual form of the SOS algorithm:

The degree-ℓ Sum-of-Squares Algorithm (dual form)

Input: Polynomials $P_0, \ldots, P_m \in \mathbb{R}[x]$
Goal: Estimate $\min P_0(x)$ over all x with $P_1(x) = \ldots = P_m(x) = 0$
Operation: Output the smallest value $\varphi^{(\ell)}$ such that there is a degree-ℓ pseudodistribution $\{x\}$ satisfying the system,

$$\{P_0 = \varphi^{(\ell)}, P_1 = 0, \ldots, P_m(x) = 0\}\,.$$

Theorem 2.7 shows that in the cases we are interested in, both variants of the SOS algorithm will output the same answer. Regardless, a similar proof to

that of Theorem 2.3 shows that the dual form of the SOS algorithm can also be computed in time $n^{O(\ell)}$. Thus, when using the SOS meta-algorithm, instead of trying to argue from the non-existence of a proof, we will use the existence of a pseudodistribution. Specifically, to show that the algorithm provides an $f(\cdot)$ approximation in the sense of (4), what we need to show is that given a degree-ℓ pseudodistribution $\{x\}$ satisfying the system $\{P = \varphi, P_1 = 0, \ldots, P_m = 0\}$, we can find some particular x^* that satisfies $P(x^*) \leq f(\varphi)$. Our approach to doing so (based on the authors' paper with Kelner [BKS14b]) can be summarized as follows:

> *Solve the problem pretending that* $\{x\}$ *is an* actual distribution *over solutions, and if all the steps you used have low-degree SOS proofs, the solution still works even when* $\{x\}$ *is a low-degree* pseudodistribution.

It may seem that coming up with an algorithm for the actual distribution case is trivial, as any element in the support of the distribution would be a good solution. However note that even in the case of a real distribution, the algorithm does not get sampling access to the distribution, but only access to its low-degree moments. Depending on the reader's temperament, the above description of the algorithm, which "pretends" pseudodistributions are real ones, may sound tautological or just wrong. Hopefully it will be clearer after the next two sections, where we use this approach to show how the SOS algorithm can match the guarantee of Cheeger's Inequality for computing the expansion, to find planted sparse vectors in random subspaces, and to approximately recover sparsely used dictionaries.

3. Approximating expansion via sums of squares

Recall that the *expansion*, ϕ_G, of a d-regular graph $G = (V, E)$ is the minimum of $\phi_G(S) = |E(S, V \setminus S)||/(d|S|)$ over all sets S of size at most $|V|/2$. Letting $x = \mathbb{1}_S$ be the characteristic vector[12] of the set S the expression $|E(S, V \setminus S)|$ can be written as $\sum_{\{i,j\} \in E} (x_i - x_j)^2$ which is a quadratic polynomial in x. Therefore, for every k, computing the value $\phi_G(k) = \min_{|S|=k} |E(S, V \setminus S)|/(dk)$ can be phrased as the question of minimizing a polynomial P_0 over the set of x's satisfying the equations $\{x_i^2 - x_i = 0\}_{i=1}^n$ and $\{\sum_{i=1}^n x_i = k\}$. Let $\phi_G^{(\ell(k))}$ be the degree-ℓ SOS estimate for $\phi_G(k)$. We call $\phi_G^{(\ell)} = \min_{k \leq n/2} \phi_G(k)$ the degree-ℓ SOS estimate for ϕ_G. Note that $\phi_G^{(\ell)}$ can be computed in $n^{O(\ell)}$ time. For the case $\ell = 2$, the following theorem describes the approximation guarantee of the estimate $\phi_G^{(\ell)}$.

Theorem 3.1. *There exists an absolute constant c such that for every graph G*

$$\phi_G \leq c\sqrt{\phi_G^{(2)}} \tag{5}$$

Before we prove Theorem 3.1, let us discuss its significance. Theorem 3.1 is essentially a restatement of Cheeger's Inequality in the SOS language—the degree 2-SOS algorithm is the UGC meta algorithm which is essentially the same as the

[12] The i-th coordinate of vector $\mathbb{1}_S$ is equal 1 if $i \in S$ and equal 0 otherwise.

algorithm based on the second-largest eigenvalue.[13] There are examples showing that (5) is tight, and so we cannot get better approximation using degree 2 proofs. But can we get a better estimate using degree 4 proofs? Or degree $\log n$ proofs? We don't know the answer, but if the Small-Set Expansion Hypothesis is true, then beating the estimate (5) is **NP**-hard, which means (under standard assumptions) that to do so we will need to use proofs of degree at least $n^{\Omega(1)}$.

This phenomenon repeats itself in other problems as well. For example, for both the Grothendieck Inequality and the Max Cut problems, the SSEH (via the UGC) predicts that beating the estimate obtained by degree-2 proofs will require degree $\ell = n^{\Omega(1)}$. As in the case of expansion, we have not been able to confirm or refute these predictions. However, we will see some examples where using higher degree proofs *does* help, some of them suspiciously close in nature to the expansion problem.

One such example comes from the beautiful work of Arora, Rao and Vazirani [ARV09] who showed that

$$\phi_G \leq O(\sqrt{\log n}) \cdot \phi_G^{(6)} \ ,$$

which is better than the guarantee of Theorem 3.1 for $\phi_G \ll 1/\log n$. However, this is not known to contradict the SSEH or UGC, which apply to the case when ϕ_G is a small constant.

As we will see in Section 5, for the small set expansion problem of approximating $\phi_G(S)$ for small sets S, we can beat the degree 2 bounds with degree $\ell = n^\tau$ proofs where τ is a parameter tending to zero with the parameter ε of the SSEH [ABS10]. This yields a sub-exponential algorithm for the small-set expansion problem (which can be extended to the Unique Games problem as well) that "barely misses" refuting the SSEH and UGC. We will also see that degree $O(1)$ proofs have surprising power in other settings that are closely related to the SSEH/UGC, but again at the moment still fall short of refuting those conjectures.

3.1. Proof of Theorem 3.1. This proof is largely a reformulation of the standard proof of a discrete variant of Cheeger's Inequality, phrased in the SOS language of pseudodistributions, and hence is included here mainly to help clarify these notions, and to introduce a tool— sampling from a distribution matching first two moments of a pseudodistribution— that will be useful for us later on. By the dual formulation, to prove Theorem 3.1 we need to show that given a pseudodistribution $\{x\}$ over characteristic vectors of size-k sets S of size $k \leq n/2$ with $|E(S, V \setminus S)| = \varphi dk$, we can find a particular set S^* of size at most $n/2$ such that $E(S^*, V \setminus S^*) \leq O(\sqrt{\varphi})d|S^*|$. For simplicity, we consider the case $k = n/2$ (the other cases can be proven in a very similar way). The distribution $\{x\}$ satisfies the constraints $\{\sum x_i = n/2\}$, $\{x_i^2 = x_i\}$ for all i, and $\{\sum_{\{i,j\}\in E}(x_i - x_j)^2 = \varphi d \sum_i x_i\}$. The algorithm to find S^* is quite simple:

[13] The second-largest eigenvalue is directly related to the minimum value of φ such that there exists a degree-2 pseudodistribution satisfying the more relaxed system $\{\sum_{\{ij\}\in E}(x_i - x_j)^2 = \varphi \cdot dn/2, \sum_i x_i = n/2, \sum_i x_i^2 = n/2\}$.

1. Choose $(y_1, \ldots, y_n)$ from a random Gaussian distribution with the same quadratic moments as $\{x\}$ so that $\mathbb{E}\, y_i = \tilde{\mathbb{E}}\, x_i$ and $\mathbb{E}\, y_i y_j = \tilde{\mathbb{E}}\, x_i x_j$ for all $i, j \in [n]$. (See details below.)

2. Output the set $S^* = \{i \mid y_i \geq 1/2\}$ (which corresponds to the 0/1 vector closest to y).

We remark that the set produces by the algorithm might have cardinality larger than $n/2$, in which case we will take the complement of S^*.

Sampling from a distribution matching two moments. We will first give a constructive proof the well-known fact that for every distribution over $\mathbb{R}^n$, there exists an n-dimensional Gaussian distribution with the same quadratic moments. Given the moments of a distribution $\{x\}$ over $\mathbb{R}^n$, we can sample a Gaussian distribution $\{y\}$ matching the first two moments of $\{x\}$ as follows. First, we can assume $\mathbb{E}\, x_i = 0$ for all i by shifting variables if necessary. Next, let $v^1, \ldots, v^n$ and $\lambda_1, \ldots, \lambda_n$ be the eigenvectors and eigenvalues of the matrix $M_{i,j} = \mathbb{E}\, x_i x_j$. (Note that M is positive semidefinite and so $\lambda_1, \ldots, \lambda_n \geq 0$.) Choose i.i.d random standard Gaussian variables $w_1, \ldots, w_n$ and define $y = \sum_k \sqrt{\lambda_k} w_k v^k$. Since $\mathbb{E}\, w_k w_{k'}$ equals 1 if $k = k'$ and equals 0 otherwise,

$$\mathbb{E}\, y_i y_j = \sum_k \lambda_k (v^k)_i (v^k)_j = M_{i,j}\ .$$

One can verify that if $\{x\}$ is a degree-2 pseudodistribution then the second moment matrix M of the shifted version of x (such that $\tilde{\mathbb{E}}\, x_i = 0$ for all i) is positive-semidefinite, and hence the above can be carried for pseudodistributions of degree at least 2 as well. Concretely, if we let $\bar{x} = \tilde{\mathbb{E}}\, x$ be the mean of the pseudodistribution, then $M = \tilde{\mathbb{E}}(x - \bar{x})(x - \bar{x})^\top$. This matrix is positive semidefinite because every test vector $z \in \mathbb{R}^n$ satisfies $z^\top M z = \tilde{\mathbb{E}}\big(z^\top (x - \bar{x})\big)^2 \geq 0$.

Analyzing the algorithm. The analysis is based on the following two claims: (i) the set S^* satisfies $n/3 \leq |S^*| \leq 2n/3$ with constant probability and (ii) in expectation $|E(S^*, V \setminus S^*)| \leq O(\sqrt{\varphi} dn)$.

We will focus on two extreme cases that capture the heart of the arguments for the claims. In the first case, all variables y_i have very small variance so that $\mathbb{E}\, y_i^2 \approx (\mathbb{E}\, y_i)^2$. In this case, because our constraints imply that $\mathbb{E}\, y_i^2 = \mathbb{E}\, y_i$, every variable satisfies either $\mathbb{E}\, y_i^2 \approx 0$ or $\mathbb{E}\, y_i^2 \approx 1$, which means that the distribution of the set S^* produced by the algorithm is concentrated around a particular set, and it is easy to verify that this set satisfies the two claims. In the second, more interesting case, all variables y_i have large variance, which means $\mathbb{E}\, y_i^2 = 1/2$ in our setting.

In this case, each event $\{y_i \geq 1/2\}$ has probability $1/2$ and therefore $\mathbb{E}|S^*| = n/2$. Using that the quadratic moments of $\{y\}$ satisfy $\mathbb{E} \sum_i y_i = n/2$ and $\mathbb{E}(\sum_i y_i)^2 = (n/2)^2$, one can show that these events cannot be completely correlated, which

allows us to control the probability of the event $n/3 \leq |S^*| \leq 2n/3$ and establishes (i). For the second claim, it turns out that by convexity considerations it suffices to analyze the case that all edges contribute equally to the term $\frac{1}{|E|}\sum_{\{i,j\}\in E}\tilde{\mathbb{E}}(x_i - x_j)^2 = \varphi$, so that $\tilde{\mathbb{E}}(x_i - x_j)^2 = \varphi$ for all $\{i,j\} \in E$. So we see that $\{y_i, y_j\}$ is a 2-dimensional Gaussian distribution with mean $(\frac{1}{2}, \frac{1}{2})$ and covariance $\frac{1}{4}\left(\begin{smallmatrix} 1 & 1-2\varphi \\ 1-2\varphi & 1 \end{smallmatrix}\right)$ Thus, in order to bound the expected value of $|E(S^*, V \setminus S^*)|$, we need to bound the probability of the event "$y_i \geq 1/2$ and $y_j < 1/2$" for this particular Gaussian distribution, which amounts to a not-too-difficult calculation that indeed yields an upper bound of $O(\sqrt{\varphi})$ on this probability. □

4. Machine learning with Sum of Squares

In this section, we illustrate the computational power of the sum-of-squares method with applications to two basic problems in unsupervised learning. In these problems, we are given samples of an unknown distribution from a fixed, parametrized family of distributions and the goal is to recover the unknown parameters from these samples. Despite the average-case nature of these problems, most of the analysis in these applications will be for deterministic problems about polynomials that are interesting in their own right.

The first problem is SPARSE VECTOR RECOVERY. Here, we are given a random basis of a d-dimensional linear subspace $U \subseteq \mathbb{R}^n$ of the form

$$U = \mathrm{Span}\{x^{(0)}, x^{(1)}, \ldots, x^{(d)}\},$$

where $x^{(0)}$ is a sparse vector and $x^{(1)}, \ldots, x^{(d)}$ are independent standard Gaussian vectors. The goal is to reconstruct the vector $x^{(0)}$. This is a natural problem in its own right, and is also a useful subroutine in various settings; see [DH13]. Demanet and Hand [DH13] gave an algorithm (based on [SWW12]) that recovers $x^{(0)}$ by searching for the vector x in U that maximizes $\|x\|_\infty/\|x\|_1$ (which can be done efficiently by n linear programs). It is not hard to show that $x^{(0)}$ has to have less than $|n|/\sqrt{d}$ coordinates for it to be maximize this ratio,[14] and hence this was a limitation of prior techniques. In contrast, as long as d is not too large (namely, $d = O(\sqrt{n})$), the SOS method can recover $x^{(0)}$ as long as it has less than εn coordinates for some constant $\varepsilon > 0$ [BKS14b].

The second problem we consider is SPARSE DICTIONARY LEARNING, also known as SPARSE CODING. Here, we are given independent samples $y^{(1)}, \ldots, y^{(R)} \in \mathbb{R}^n$ from an unknown distribution of the form $\{y = Ax\}$, where $A \in \mathbb{R}^{n\times m}$ is a matrix and x is a random m-dimensional vector from a distribution over sparse vectors. This problem, initiated by the work Olshausen and Field [OF96] in computational neuroscience, has found a variety of uses in machine learning, computer vision, and image processing (see, e.g. [AAJ+13] and the references therein). The appeal of this problem is that intuitively data should be sparse in the "right" representation

[14] See Lemma 5.2 below for a related statement.

(where every coordinate corresponds to a meaningful feature), and finding this representation can be a useful first step for further processing, just as representing sound or image data in the Fourier or Wavelet bases is often a very useful primitive. While there are many heuristics use to solve this problem, prior works giving rigorous recovery guarantees such as [SWW12, AAJ+13, AGM13] all required the vector x to be *very* sparse, namely less than $\sqrt{n}$ nonzero entries.[15] In contrast, the SOS method can be used to approximately recover the dictionary matrix A as long as x has $o(n)$ nonzero (or more generally, significant) entries [BKS14a].

4.1. Sparse vector recovery. We say a vector x is μ-sparse if the 0/1 indicator $\mathbb{1}_{\operatorname{supp} x}$ of the support of x has norm-squared $\mu = \|\mathbb{1}_{\operatorname{supp} x}\|_2^2$. The ratio $\mu/\|\mathbb{1}\|_2^2$ is the fraction of non-zero coordinates in x.

Theorem 4.1. *There exists a polynomial-time approximation algorithm for* SPARSE VECTOR RECOVERY *with the following guarantees: Suppose the input of the algorithm is an arbitrary basis of a $d+1$-dimensional linear subspace $U \subseteq \mathbb{R}^n$ of the form $U = \operatorname{Span}\{x^{(0)}, x^{(1)} \dots, x^{(d)}\}$ such that $x^{(0)}$ is a μ-sparse unit vector with $\mu \leq \varepsilon \cdot \|\mathbb{1}\|_2^2$ and $x^{(1)}, \dots, x^{(d)}$ are standard Gaussian vectors orthogonal to $x^{(0)}$ with $d \ll \sqrt{n}$. Then, with probability close to 1, the algorithm outputs a unit vector x that has correlation $\langle x, x^{(0)} \rangle^2 \geq 1 - O(\varepsilon)$ with $x^{(0)}$.*

Our algorithm will follow the general recipe we described in Section 2.2:

> *Find a system of polynomial equations $\mathcal{E}$ that captures the intended solution $x^{(0)}$, then pretend you are given a distribution $\{u\}$ over solutions of $\mathcal{E}$ and show how you could recover a single solution u^* from the low order moments of $\{u\}$.*

Specifically, we come up with a system $\mathcal{E}$ so that desired vector $x^{(0)}$ satisfies all equations, and it is essentially the only solution to $\mathcal{E}$. Then, using the SOS algorithm, we compute a degree-4 pseudodistribution $\{u\}$ that satisfies $\mathcal{E}$. Finally, as in Section 3.1, we sample a vector u^* from a Gaussian distribution that has the same quadratic moments as the pseudodistribution $\{u\}$.

How to encode this problem as a system of polynomial equations? By Cauchy–Schwarz, any μ-sparse vector x satisfies $\|x\|_2^2 \leq \|x\|_{2p}^2 \cdot \|\mathbb{1}_{\operatorname{supp} x}\|_q = \|x\|_{2p}^2 \cdot \mu^{1-1/p}$ for all $p, q \geq 1$ with $1/p + 1/q = 1$. In particular, for $p = 2$, such vectors satisfy $\|x\|_4^4 \geq \|x\|_2^4/\mu$. This fact motivates our encoding of SPARSE VECTOR RECOVERY as a system of polynomial equations. If the input specifies subspace $U \subseteq \mathbb{R}^n$, then we compute the projector P into the subspace U and choose the following polynomial equations: $\|u\|_2^2 = 1$ and $\|Pu\|_4^4 = 1/\mu_0$, where $\mu_0 = \|x^{(0)}\|_2^4/\|x^{(0)}\|_4^4$. (We assume here the algorithm is given $\mu_0 \leq \mu$ as input, as we can always guess a sufficiently close approximation to it.)

[15] If the distribution x consists of m independent random variables then better guarantees can be achieved using *Independent Component Analysis (ICA)* [Com94]. See [GVX14] for the current state of art in this setting. However we are interested here in the more general case.

Why does the sum-of-squares method work? The analysis of algorithm has two ingredients. The first ingredient is a structural property about projectors of random subspaces.

Lemma 4.2. *Let $U' \subseteq \mathbb{R}^n$ be a random d-dimensional subspace with $d \ll \sqrt{n}$ and let P' be the projector into U'. Then, with high probability, the following sum-of-squares relation over $\mathbb{R}[u]$ holds for $\mu' \geq \Omega(1) \cdot \|\mathbb{1}\|_2^2$,*

$$\|P'u\|_4^4 \preccurlyeq \|u\|_2^4/\mu' \,.$$

Proof outline. We can write $P' = B^\top B$ where B is a $d \times n$ matrix whose rows are an orthogonal basis for the subspace U'. Therefore, $P'u = B^\top x$ where $x = Bu$, and so to prove Lemma 4.2 it suffices to show that under these conditions, $\|B^\top x\|_4^4 \preccurlyeq O(\|x\|_2^4/\|\mathbb{1}\|_2^4)$. The matrix $B^\top$ will be very close to having random independent Gaussian entries, and hence, up to scaling, $\|B^\top x\|_4^4$ will be (up to scaling), close to $Q(x) = \frac{1}{n}\sum\langle w_i, x\rangle^4$ where $w_1, \ldots, w_d \in \mathbb{R}^d$ are chosen independently at random from the standard Gaussian distribution. The expectation of $\langle w, x\rangle^4$ is equal $3\sum_{i,j} x_i^2 x_j^2 = 3\|x\|_2^4$. Therefore, to prove the lemma, we need to show that for $n \gg d^2$, the polynomial $Q(x)$ is with high probability close to its expectation, in the sense that the $d^2 \times d^2$ matrix corresponding to Q's coefficients is close to its expectation in the spectral norm. This follows from standard matrix concentration inequalities, see [BBH+12, Theorem 7.1[16]]). □

The following lemma is the second ingredient of the analysis of the algorithm.

Lemma 4.3. *Let $U' \subseteq \mathbb{R}^n$ be a linear subspace and let P' be the projector into U'. Let $x^{(0)} \in \mathbb{R}^n$ be a μ-sparse unit vector orthogonal to U' and let $U = \mathrm{Span}\{x^{(0)}\} \oplus U'$ and P the projector on U. Let $\{u\}$ be a degree-4 pseudodistribution that satisfies the constraints $\{\|u\|_2^2 = 1\}$ and $\{\|Pu\|_4^4 = 1/\mu_0\}$, where $\mu_0 = \|x^{(0)}\|_2^4/\|x^{(0)}\|_4^4 \leq \mu$. Suppose $\|P'u\|_4^4 \preccurlyeq \|u\|_2^4/\mu'$ is a sum-of-squares relation in $\mathbb{R}[u]$. Then, $\{u\}$ satisfies*

$$\tilde{\mathbb{E}}\|P'u\|_2^2 \leq 4\big(\tfrac{\mu}{\mu'}\big)^{1/4} \,.$$

Note that the conclusion of Lemma 4.3 implies that a vector u^* sampled from a Gaussian distribution with the same quadratic moments as the computed pseudodistribution also satisfies $\mathbb{E}_{u^*}\|P'u^*\|_2^2 \leq 4(\mu/\mu')^{1/4}$ and $\mathbb{E}\|u^*\|_2^2 = 1$. By Markov inequality, $\|u^* - x^{(0)}\|_2^2 \leq 16(\mu/\mu')^{1/4}$ holds with probability at least $3/4$. Since u^* is Gaussian, it satisfies $\|u^*\|_2^2 \geq 1/4$ with probability at least $1/2$. If both events occur, which happens with probability at least $1/4$, then $\langle u^*, x^{(0)}\rangle^2 \geq (1 - O(\mu/\mu'))\|u^*\|_2^2$, thus establishing Theorem 4.1.

Proof of Lemma 4.3 There are many ways in which pseudodistributions behave like actual distributions, as far as low degree polynomials are concerned. To prove Lemma 4.3, we need to establish the following two such results:

[16] The reference is for the arxiv version `arXiv:1205.4484v2` of the paper.

Lemma 4.4 (Hölder's inequality for pseudoexpectation norms). *Suppose a and b are nonnegative integers that sum to a power of 2. Then, every degree-$(a+b)$ pseudodistribution $\{u, v\}$ satisfies*

$$\tilde{\mathbb{E}}\,\mathbb{E}_i\, u_i^a v_i^b \le \left(\tilde{\mathbb{E}}\,\mathbb{E}_i\, u_i^{a+b}\right)^{a/(a+b)} \cdot \left(\tilde{\mathbb{E}}\,\mathbb{E}_i\, v_i^{a+b}\right)^{b/(a+b)} .$$

Proof sketch. The proof of the general case follows from the case $a = b = 2$ by an inductive argument. The proof for the case $a = b = 1$ follows from the fact that the polynomial $\alpha\,\mathbb{E}_i\, u_i^2 + \beta\,\mathbb{E}_i\, v_i^2 - \sqrt{\alpha\beta}\,\mathbb{E}_i\, u_i v_i \in \mathbb{R}[u, v]$ is a sum of squares for all $\alpha, \beta \ge 0$ and choosing $\alpha = 1/\,\tilde{\mathbb{E}}\,\mathbb{E}_i\, u_i^2$ and $\beta = 1/\,\tilde{\mathbb{E}}\,\mathbb{E}_i\, v_i^2$ □

Lemma 4.5 (Triangle inequality for pseudodistribution ℓ_4 norm). *Let $\{u, v\}$ be a degree-4 pseudodistribution. Then,*

$$\left(\tilde{\mathbb{E}}\|u+v\|_4^4\right)^{1/4} \le \left(\tilde{\mathbb{E}}\|u\|_4^4\right)^{1/4} + \left(\tilde{\mathbb{E}}\|v\|_4^4\right)^{1/4} .$$

Proof. The inequality is invariant with respect to the measure used for the inner norm $\|\cdot\|_4$. For simplicity, suppose $\|x\|_4^4 = \mathbb{E}\, x_i^4$. Then, $\|u+v\|_4^4 = \mathbb{E}_i\, u_i^4 + 4\,\mathbb{E}_i\, u_i^3 v_i + 6\,\mathbb{E}_i\, u_i v_i^3 + \mathbb{E}_i\, v_i^4$. Let $A = \tilde{\mathbb{E}}\,\mathbb{E}_i\, u_i^4$ and $B = \tilde{\mathbb{E}}\,\mathbb{E}_i\, v_i^4$. Then, Lemma 4.5 allows us to bound the pseudoexpectations of the terms $\mathbb{E}_i\, u_i^a v_i^b$, so that as desired

$$\tilde{\mathbb{E}}\|u+v\|_4^4 \le A + 4A^{3/4}B^{1/4} + 6A^{1/2}B^{1/2} + 4A^{1/3}B^{3/4} + B = (A^{1/4} + B^{1/4})^4 . \ \square$$

We can now prove Lemma 4.1. Let $\alpha_0 = \langle u, x^{(0)}\rangle \in \mathbb{R}[u]$. By construction, the polynomial identity $\|Pu\|_4^4 = \|\alpha_0 x^{(0)} + P'u\|_4^4$ holds over $\mathbb{R}[u]$. By the triangle inequality for pseudodistribution ℓ_4 norm, for $A = \tilde{\mathbb{E}}\,\alpha_0^4\|x^{(0)}\|_4^4$ and $B = \tilde{\mathbb{E}}\|P'u\|_4^4$

$$\left(\tfrac{1}{\mu_0}\right)^{1/4} = \left(\tilde{\mathbb{E}}\|Pu\|_4^4\right)^{1/4} \le A^{1/4} + B^{1/4}$$

By the premises of the lemma, $A = \tilde{\mathbb{E}}\,\alpha_0^4/\mu_0$ and $B \le 1/\mu'$. Together with the previous bound, it follows that $(\tilde{\mathbb{E}}\,\alpha_0^4)^{1/4} \ge 1 - (\mu_0/\mu')^{1/4}$. Since $\alpha_0^2 \preccurlyeq \|u\|_2^2$ and $\{u\}$ satisfies $\|u\|_2^2 = 1$, we have $\tilde{\mathbb{E}}\,\alpha_0^2 \ge \tilde{\mathbb{E}}\,\alpha_0^4 \ge 1 - 4(\mu_0/\mu')^{1/4}$. Finally, using $\|u - x^{(0)}\|_2^2 = \|u\|_2^2 - \alpha_0^2$, we derive the desired bound $\tilde{\mathbb{E}}\|u - x^{(0)}\|_2^2 = 1 - \tilde{\mathbb{E}}\,\alpha_0^2 \le 4(\mu_0/\mu')^{1/4}$ thus establishing Lemma 4.5 and Theorem 4.1. □

4.2. Sparse dictionary learning.

A *κ-overcomplete dictionary* is a matrix $A \in \mathbb{R}^{n\times m}$ with $\kappa = m/n \ge 1$ and isotropic unit vectors as columns (so that $\|A^\top u\|_2^2 = \kappa\|u\|_2^2$). We say a distribution $\{x\}$ over $\mathbb{R}^m$ is (d, τ)-nice if it satisfies $\mathbb{E}_i\, x_i^d = 1$ and $\mathbb{E}_i\, x_i^{d/2} x_j^{d/2} \le \tau$ for all $i \ne j \in [m]$, and it satisfies that non-square monomial degree-d moments vanish so that $\mathbb{E}\, x^\alpha = 0$ for all non-square degree-d monomials x^α, where $x^\alpha = \prod x_i^{\alpha_i}$ for $\alpha \in \mathbb{Z}^n$. For $d = O(1)$ and $\tau = o(1)$, a nice distribution satisfies that $\mathbb{E}\,\frac{1}{m}\sum_i x_i^4 \gg \left(\frac{1}{m}\sum_i x_i^2\right)^2$ which means that it is approximately sparse in the sense that the square of the entries of x has large variance (which means that few of the entries have very big magnitude compared to the rest).

Theorem 4.6. *For every $\varepsilon > 0$ and $\kappa \geq 1$, there exists d and τ and a quasipolynomial-time algorithm algorithm for* SPARSE DICTIONARY LEARNING *with the following guarantees: Suppose the input consists of $n^{O(1)}$ independent samples[17] from a distribution $\{y = Ax\}$ over $\mathbb{R}^n$, where $A \in \mathbb{R}^{n\times m}$ is a κ-overcomplete dictionary and the distribution $\{x\}$ over $\mathbb{R}^m$ is (d,τ)-nice. Then, with high probability, the algorithm outputs a set of vectors with Hausdorff distance[18] at most ε from the set of columns of A.*

Encoding as a system of polynomial equations. Let $y^{(1)}, \dots, y^{(R)}$ be independent samples from the distribution $\{y = Ax\}$. Then, we consider the polynomial $P = \frac{1}{R}\sum_i \langle y^{(i)}, u\rangle^d \in \mathbb{R}[u]_d$. Using the properties of nice distributions, a direct computation shows that with high probability P satisfies the relation

$$\|A^\top u\|_d^d - \tau\|u\|_2^d \preccurlyeq P \preccurlyeq \|A^\top u\|_d^d + \tau\|u\|_2^d\,.$$

(Here, we are omitting some constant factors, depending on d, that are not important for the following discussion.) It follows that $P(a^{(i)}) = 1 \pm \tau$ for every column $a^{(i)}$ of A. It's also not hard to show that every unit vector a^* with $P(a^*) \approx 1$ is close to one of the columns of A. (Indeed, every unit vector satisfies $P(a^*) \leq \max_i \langle a^{(i)}, a^*\rangle^{d-2}\kappa + \tau$. Therefore, $P(a^*) \approx 1$ implies that $\langle a^{(i)}, a^*\rangle^2 \geq \kappa^{-\Omega(1/d)}$, which is close to 1 for $d \gg \log\kappa$.) What we will show is that pseudodistributions of degree $O(\log n)$ allow us to find all such vectors.

Why does the sum-of-squares method work? In the following, $\varepsilon > 0$ and $\kappa \geq 1$ are arbitrary constants that determine constants $d = d(\varepsilon,\kappa) \geq 1$ and $\tau = \tau(\varepsilon,\kappa) > 0$ (as in the theorem).

Lemma 4.7. *Let $P \in \mathbb{R}[u]$ be a degree-d polynomial with $\pm(P - \|A^\top u\|_d^d) \preccurlyeq \tau\|u\|_2^d$ for some κ-overcomplete dictionary A. Let $\mathcal{D}$ be a degree-$O(\log n)$ pseudodistribution that satisfies the constraints $\{\|u\|_2^2 = 1\}$ and $\{P(u) = 1 - \tau\}$. Let $W \in \mathbb{R}[u]$ be a product of $O(\log n)$ random linear forms[19]. Then, with probability at least $n^{-O(1)}$ over the choice of W, there exists a column $a^{(i)}$ of A such that*

$$\tfrac{1}{\tilde{\mathbb{E}}_{\mathcal{D}} W^2}\,\tilde{\mathbb{E}}_{\mathcal{D}}\, W^2 \cdot \big(\|u\|^2 - \langle a^{(i)}, u\rangle^2\big) \leq \varepsilon\,.$$

If $\tilde{\mathbb{E}}_{\mathcal{D}}$ is a pseudoexpectation operator, then $\tilde{\mathbb{E}}_{\mathcal{D}'} \colon P \mapsto \tilde{\mathbb{E}}\, W^2 P / \tilde{\mathbb{E}}\, W^2$ is also a pseudoexpectation operator (as it satisfies linearity, normalization, and nonnegativity). (This transformation corresponds to reweighing the pseudodistribution

[17] Here, we also make the mild assumption that the degree-$2d$ moments of x are bounded by $n^{O(1)}$.

[18] The *Hausdorff distance* between two sets of vectors upper bounds the maximum distance of a point in one of the sets to its closest point in the other set. Due to the innate symmetry of the sparse dictionary problem (replacing a column $a^{(i)}$ of A by $-a^{(i)}$ might not affect the input distribution), we measure the Hausdorff distance after symmetrizing the sets, i.e., replacing the set S by $S \cup -S$.

[19] Here, a random linear form means a polynomial $\langle u, v\rangle \in \mathbb{R}[u]$ where v is a random unit vector in $\mathbb{R}^n$.

$\mathcal{D}$ by the polynomial W^2.) Hence, the conclusion of the lemma gives us a new pseudodistribution $\mathcal{D}'$ such that $\tilde{\mathbb{E}}_{\mathcal{D}'}\|u\|_2^2 - \langle a^{(i)}, u\rangle^2 \leq \varepsilon$. Therefore, if we sample a Gaussian vector a^* with the same quadratic moments as $\mathcal{D}'$, it satisfies $\|a^*\|_2^2 - \langle a^{(i)}, a^*\rangle^2 \leq 4\varepsilon$ with probability at least $3/4$. At the same time, it satisfies $\|a^*\|^2 \geq 1/4$ with probability at least $1/2$. Taking these bounds together, a^* satisfies $\langle a^{(i)}, a^*\rangle^2 \geq (1 - 16\varepsilon)\|a^*\|^2$ with probability at least $1/4$.

Lemma 4.7 allows us to reconstruct one of the columns of A. Using similar ideas, we can iterate this argument and recover one-by-one all columns of A. We omit the proof of Lemma 4.7, but the idea behind it is to first give an SOS proof version of our argument above that maximizers of P must be close to one of the $a^{(i)}$'s. We then note that if a distribution $\mathcal{D}$ is supported (up to noise) on at most m different vectors, then we can essentially isolate one of these vectors by re-weighing $\mathcal{D}$ using the product of the squares of $O(\log m)$ random linear forms. It turns out, this latter argument has a low degree SOS proof as well, which means that in our case that given $\mathcal{D}$ satisfying the constraint $\{P(u) = 1 - \tau\}$, we can isolate one of the $a^{(i)}$'s even when $\mathcal{D}$ is not an actual distribution but merely a pseudodistribution.

5. Hypercontractive norms and small-set expansion

So far we have discussed the Small-Set Expansion Hypothesis and the Sum of Squares algorithm. We now discuss how these two notions are related. One connection, mentioned before, is that the SSEH predicts that in many settings the guarantees of the degree-2 SOS algorithm are best possible, and so in particular it means that going from degree 2 to say degree 100 should not give any substantial improvement in terms of guarantees. Another, perhaps more meaningful connection is that there is a candidate approach for refuting the SSEH using the SOS algorithm. At the heart of this approach is the following observation:

> *The small-set expansion problem is a special case of the problem of finding "sparse" vectors in a linear subspace.*

This may seem strange, as a priori, the following two problem seem completely unrelated: **(i)** Given a graph $G = (V, E)$, find a "small" subset $S \subseteq V$ with low expansion $\phi_G(S)$, and **(ii)** Given a subspace $W \subseteq \mathbb{R}^n$, find a "sparse" vector in W. The former is a combinatorial problem on graphs, and the latter a geometric problem on subspaces. However, for the right notions of "small" and "sparse", these turn out to be essentially the same problem. Intuitively, the reason is the following: the expansion of a set S is proportional to the quantity $x^\top Lx$ where x is the characteristic vector of S (i.e. x_i equals 1 if $i \in S$ and equals 0 otherwise), and L is the *Laplacian matrix* of G (defined as $L = I - d^{-1}A$ where I is the identity, d is the degree, and A is G's adjacency matrix). Let $v_1, \ldots, v_n$ be the eigenvectors of L and $\lambda_1, \ldots, \lambda_n$ the corresponding eigenvalues. Then $x^\top Lx = \sum_{i=1}^n \lambda_i \langle v_i, x\rangle^2$.

Therefore if $x^\top Lx$ is smaller than $\varphi\|x\|^2$ and c is a large enough constant, then most of the mass of x is contained in the subspace $W = \mathrm{Span}\{v_i : \lambda_i \leq c\varphi\}$. Since S is small, x is sparse, and so we see that there is a sparse vector that is "almost" contained in W. Moreover, by projecting x into W we can also find a "sparse" vector that is actually contained in W, if we allow a slightly softer notion of "sparseness", that instead of stipulating that most coordinates are zero, only requires that the distribution of coordinates is very "spiky" in the sense that most of its mass is dominated by the few "heavy hitters".

Concretely, for $p > 1$ and $\delta \in (0,1)$, we say that a vector $x \in \mathbb{R}^n$ is (δ, p)-*sparse* if $\mathbb{E}_i\, x_i^{2p} \geq \delta^{1-p}(\mathbb{E}_i\, x_i^2)^p$. Note that a characteristic vector of a set of measure δ is (δ, p)-sparse for any p. The relation between small-set-expansion and finding sparse vectors in a subspace is captured by the following theorem:

Theorem 5.1 (Hypercontractivity and small-set expansion [BBH$^+$12], informal statement). *Let $G = (V, E)$ be a d-regular graph with Laplacian L. Then for every $p \geq 2$ and $\varphi \in (0,1)$,*

1. (Non-expanding small sets imply sparse vectors.) *If there exists $S \subseteq V$ with $|S| = o(|V|)$ and $\phi_G(S) \leq \varphi$ then there exists an $(o(1), p)$-sparse vector $x \in W_{\leq \varphi + o(1)}$ where for every λ, $W_{\leq \lambda}$ denotes the span of the eigenvectors of L with eigenvalue smaller than λ.*

2. (Sparse vectors imply non-expanding small sets.) *If there exists a $(o(1), p)$-sparse vector $x \in W_{\leq \varphi}$, then there exists $S \subseteq V$ with $|S| = o(|V|)$ and $\phi_G(S) \leq \rho$ for some constant $\rho < 1$ depending on φ.*

The first direction of Theorem 5.1 follows from the above reasoning, and was known before the work of [BBH$^+$12]. The second direction is harder, and we omit the proof here. The theorem reduces the question of determining whether there for small sets S, the minimum of $\phi_G(S)$ is close to one or close to zero, into the question of bounding the maximum of $\mathbb{E}_i\, x_i^{2p}$ over all unit vectors in some subspace. The latter question is a polynomial optimization problem of the type the SOS algorithm is designed for! Thus, we see that we could potentially resolve the SSEH if we could answer the following question:

> *What is the degree of SOS proofs needed to certify that the $2p$-norm is bounded for all (Euclidean norm) unit vectors in some subspace W?*

We still don't know the answer to this question in full generality, but we do have some interesting special cases. Lemma 4.2 of Section 4.1 implies that if W is a random subspace of dimension $\ll \sqrt{n}$ then we can certify that $\mathbb{E}_i\, x_i^4 \leq O(\mathbb{E}_i\, x_i^2)^2$ for all $x \in W$ via a degree-4 SOS proof. This is optimal, as the 4-norm simply won't be bounded for dimensions larger than $\sqrt{n}$:

Lemma 5.2. *Let $W \subseteq \mathbb{R}^n$ have dimension d and $p \geq 2$, then there exists a unit vector $x \in W$ such that*

$$\mathbb{E}_i\, x_i^{2p} \geq \tfrac{d^p}{n}(\mathbb{E}_i\, x_i^2)^p$$

Hence in particular any subspace of dimension $d \gg n^{1/p}$ contains a $(o(1), p)$-sparse vector.

Proof of Lemma 5.2. Let P be the matrix corresponding to the projection operator to the subspace W. Note that P has d eigenvalues equalling 1 and the rest equal 0, and hence $\mathrm{Tr}(P) = d$ and the Frobenius norm squared of P, defined as $\sum P_{i,j}^2$, also equals d. Let $x^i = Pe^i$ where e^i is the i^{th} standard basis vector. Then $\sum x_i^i$ is the trace of P which equals d and hence using Cauchy-Schwarz

$$\sum (x_i^i)^2 \geq \frac{1}{n}\left(\sum x_i\right)^2 = \frac{\mathrm{Tr}(P)^2}{n} = \frac{d^2}{n} \,.$$

On the other hand,

$$\sum_i \sum_j (x^i)_j^2 = \sum_{i,j} (Pe^i)_j^2 = \sum P_{i,j}^2 = d \,.$$

Therefore, by the inequality $(\sum a_i)/(\sum b_i) \leq \max a_i/b_i$, there exists an i such that if we let $x = x^i$ then $x_i^2 \geq \frac{d}{n}\sum_j x_j^2 = d\,\mathbb{E}\, x_j^2$. Hence, just the contribution of the i^{th} coordinate to the expectation achieves $\mathbb{E}_j\, x_j^{2p} \geq \frac{d^p}{n}\left(\mathbb{E}_j\, x_j^2\right)^p$. □

Lemma 5.2 implies the following corollary:

Corollary 5.3. *Let $p, n \in \mathbb{N}$, and W be subspace of $\mathbb{R}^n$. If $\mathbb{E}_i\, x_i^{2p} \leq O(\mathbb{E}_i\, x_i^2)^p)$, then there is an $O(n^{1/p})$-degree SOS proof for this fact. (The constants in the $O(\cdot)$ notation can depend on p but not on n.)*

Proof sketch. By Lemma 5.2, the condition implies that $d = \dim W \leq O(n^{1/p})$, and it is known that approximately bounding a degree-$O(1)$ polynomial on the d-dimensional sphere requires an SOS proof of at most $O(d)$ degree (e.g., see [DW12] and the references therein). □

Combining Corollary 5.3 with Theorem 5.1 implies that for every ε, δ there exists some τ (tending to zero with ε), such that if we want to distinguish between the case that an n-vertex graph G satisfies $\phi_G(S) \leq \varepsilon$ for every $|S| \leq \delta n$, and the case that there exists some S of size at most δn with $\phi_G(S) \geq 1 - \varepsilon$, then we can do so using a degree n^τ SOS proofs, and hence in $\exp(O(n^\tau))$ time. This is much better than the trivial $\binom{n}{\delta n}$ time algorithm that enumerates all possible sets. Similar ideas can be used to achieve an algorithm with a similar running time for the problem underlying the Unique Games Conjecture [ABS10]. If these algorithms could be improved so the exponent τ tends to zero with n for a fixed ε, this would essentially refute the SSEH and UGC.

Thus, the question is whether Corollary 5.3 is the best we could do. As we've seen, Lemma 4.2 shows that for random subspaces we can do much better, namely certify the bound with a constant degree proof. Two other results are known of that flavor. Barak, Kelner and Steurer [BKS14b] showed that if a d-dimensional subspace W does not contain a $(\delta, 2)$-sparse vector, then there is an $O(1)$-degree SOS proof that it does not contain (or even almost contains) a vector with $O(\frac{\delta n}{d^{1/3}})$

nonzero coordinates. If the dependence on d could be eliminated (even at a significant cost to the degree), then this would also refute the SSEH. Barak, Brandão, Harrow, Kelner, Steurer and Zhou [BBH+12] gave an $O(1)$-degree SOS proof for the so-called "Bonami-Beckner-Gross $(2,4)$ hypercontractivity theorem" (see [O'D14, Chap. 9]). This is the statement that for every constant k, the subspace $W_k \subseteq \mathbb{R}^{2^t}$ containing the evaluations of all degree $\leq k$ polynomials on the points $\{\pm 1\}^t$ does not contain an $(o(1),2)$-sparse vector, and specifically satisfies for all $x \in W_k$,

$$\mathbb{E}\, x_i^4 \leq 9^k (\mathbb{E}\, x_i^2)^2 \,. \tag{6}$$

On its own this might not seem so impressive, as this is just one particular subspace. However, this particular subspace underlies much of the evidence that has been offered so far in support of both the UGC and SSEH conjectures. The main evidence for the UGC/SSEH consists of several papers such as [KV05, KS09, RS09a, BGH+12] that verified the predictions of these conjectures by proving that various natural algorithms indeed fail to solve some of the computational problems that are hard if the conjectures are true. These results all have the form of coming up with a "hard instance" G on which some algorithm $\mathcal{A}$ fails, and so to prove such a result one needs to do two things: **(i)** compute (or bound) the true value of the parameter on G, and **(ii)** show that the value that $\mathcal{A}$ outputs on G is (sufficiently) different than this true value. It turns out that all of these papers, the proof of **(i)** can be formulated as low degree SOS proof, and in fact the heart of these proofs is the bound (6). Therefore, the results of [BBH+12] showed that all these "hard instances" can in fact be solved by the SOS algorithm using a constant degree. This means that at the moment, we don't even have any example of an instance for the problems underlying the SSEH and UGC that can be reasonably *conjectured* (let alone proved) hard for the constant degree SOS algorithm. This does not mean that such instances do not exist, but is suggestive that we have not yet seen the last algorithmic word on this question.

Acknowledgments. We thank Amir Ali Ahmadi for providing us with references on the diverse applications of the SOS method.

References

[AAJ+13] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning Sparsely Used Overcomplete Dictionaries via Alternating Minimization. *arXiv preprint 1310.7991*, 2013.

[Alo86] N. Alon. Decomposition of the complete$r$-graph into complete$r$-partite$r$-graphs. *Graphs and Combinatorics*, 2(1):95–100, 1986.

[AM85] N. Alon and V. D. Milman. λ_1, Isoperimetric inequalities for graphs, and superconcentrators. *J. Comb. Theory, Ser. B*, 38(1):73–88, 1985.

[AMS11] C. Ambühl, M. Mastrolilli, and O. Svensson. Inapproximability Results for Maximum Edge Biclique, Minimum Linear Arrangement, and Sparsest Cut. *SIAM J. Comput.*, 40(2):567–596, 2011.

[ABS10] S. Arora, B. Barak, and D. Steurer. Subexponential Algorithms for Unique Games and Related Problems. In *FOCS*, pages 563–572, 2010.

[AGM13] S. Arora, R. Ge, and A. Moitra. New Algorithms for Learning Incoherent and Overcomplete Dictionaries. *arXiv preprint 1308.6723*, 2013.

[ARV09] S. Arora, S. Rao, and U. V. Vazirani. Expander flows, geometric embeddings and graph partitioning. *J. ACM*, 56(2), 2009.

[Art27] E. Artin. Über die zerlegung definiter funktionen in quadrate. In *Abhandlungen aus dem mathematischen Seminar der Universität Hamburg*, volume 5, pages 100–115. Springer, 1927.

[Bar12] B. Barak. Truth vs. Proof in Computational Complexity. *Bulletin of the European Association for Theoretical Computer Science*, (108), October 2012.

[Bar14a] B. Barak. Fun and Games with Sums of Squares, Feb. 2014. Windows on Theory blog, `http://windowsontheory.org`

[Bar14b] B. Barak. Structure vs. Combinatorics in Computational Complexity. *Bulletin of the EATCS*, (112):115–126, February 2014.

[BBH+12] B. Barak, F. G. S. L. Brandão, A. W. Harrow, J. A. Kelner, D. Steurer, and Y. Zhou. Hypercontractivity, sum-of-squares proofs, and their applications. In *STOC*, pages 307–326, 2012.

[BGH+12] B. Barak, P. Gopalan, J. Håstad, R. Meka, P. Raghavendra, and D. Steurer. Making the Long Code Shorter. In *FOCS*, pages 370–379, 2012.

[BKS14a] B. Barak, J. Kelner, and D. Steurer. Dictionary Learning via the Sum-of-Squares Method. Unpublished manuscript, 2014.

[BKS14b] B. Barak, J. Kelner, and D. Steurer. Rounding Sum of Squares Relaxations. In *STOC*, 2014.

[BPT13] G. Blekherman, P. A. Parrilo, and R. R. Thomas. *Semidefinite optimization and convex algebraic geometry*, volume 13. Siam, 2013.

[Che70] J. Cheeger. A lower bound for the smallest eigenvalue of the Laplacian. *Problems in analysis*, 625:195–199, 1970.

[CHKX06] J. Chen, X. Huang, I. A. Kanj, and G. Xia. Strong computational lower bounds via parameterized complexity. *Journal of Computer and System Sciences*, 72(8):1346–1367, 2006.

[Com94] P. Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

[DH13] L. Demanet and P. Hand. Recovering the Sparsest Element in a Subspace, Oct. 2013. Arxiv preprint 1310.1654.

[Dod84] J. Dodziuk. Difference equations, isoperimetric inequality and transience of certain random walks. *Transactions of the American Mathematical Society*, 284(2):787–794, 1984.

[DW12] A. C. Doherty and S. Wehner. Convergence of SDP hierarchies for polynomial optimization on the hypersphere. *arXiv preprint arXiv:1210.5048*, 2012.

[DF95] R. G. Downey and M. R. Fellows. Fixed-parameter tractability and completeness II: On completeness for W[1]. *Theoretical Computer Science*, 141(1):109–131, 1995.

[GW95] M. X. Goemans and D. P. Williamson. Improved Approximation Algorithms for Maximum Cut and Satisfiability Problems Using Semidefinite Programming. *J. ACM*, 42(6):1115–1145, 1995.

[GVX14] N. Goyal, S. Vempala, and Y. Xiao. Fourier PCA. In *STOC*, 2014. Also available as arXiv report 1306.5825.

[Gri01] D. Grigoriev. Linear lower bound on degrees of Positivstellensatz calculus proofs for the parity. *Theor. Comput. Sci.*, 259(1-2):613–622, 2001.

[GV01] D. Grigoriev and N. Vorobjov. Complexity of Null-and Positivstellensatz proofs. *Annals of Pure and Applied Logic*, 113(1):153–160, 2001.

[Gro53] A. Grothendieck. Résumé de la théorie métrique des produits tensoriels topologiques. *Bol. Soc. Mat. Sao Paulo*, 8(1-79):88, 1953.

[Hås96] J. Håstad. Clique is Hard to Approximate Within $n^{1-\epsilon}$. In *FOCS*, pages 627–636, 1996.

[Kho01] S. Khot. Improved Inaproximability Results for MaxClique, Chromatic Number and Approximate Graph Coloring. In *FOCS*, pages 600–609, 2001.

[Kho02] S. Khot. On the Power of Unique 2-Prover 1-Round Games. In *IEEE Conference on Computational Complexity*, page 25, 2002.

[Kho10a] S. Khot. Inapproximability of np-complete problems, discrete fourier analysis, and geometry. In *International Congress of Mathematics*, volume 5, 2010.

[Kho10b] S. Khot. On the Unique Games Conjecture (Invited Survey). In *2012 IEEE 27th Conference on Computational Complexity*, pages 99–121. IEEE, 2010.

[KS09] S. Khot and R. Saket. SDP Integrality Gaps with Local ℓ_1-Embeddability. In *FOCS*, pages 565–574, 2009.

[KV05] S. Khot and N. K. Vishnoi. The Unique Games Conjecture, Integrality Gap for Cut Problems and Embeddability of Negative Type Metrics into ℓ_1. In *FOCS*, pages 53–62, 2005.

[Kri64] J.-L. Krivine. Anneaux préordonnés. *Journal d'analyse mathématique*, 12(1):307–326, 1964.

[Las01] J. B. Lasserre. Global Optimization with Polynomials and the Problem of Moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[Lau03] M. Laurent. A Comparison of the Sherali-Adams, Lovász-Schrijver, and Lasserre Relaxations for 0-1 Programming. *Math. Oper. Res.*, 28(3):470–496, 2003.

[Lau09] M. Laurent. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*, pages 157–270. Springer, 2009.

[Lov79] L. Lovász. On the Shannon capacity of a graph. *Information Theory, IEEE Transactions on*, 25(1):1–7, 1979.

[LS91] L. Lovász and A. Schrijver. Cones of matrices and set-functions and 0-1 optimization. *SIAM Journal on Optimization*, 1(2):166–190, 1991.

[Nes00] Y. Nesterov. Squared functional systems and optimization problems. *High performance optimization*, 13:405–440, 2000.

[O'D14] R. O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. To be published in May 2014.

[OF96] B. A. Olshausen and D. J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, 1996.

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

[Rag08] P. Raghavendra. Optimal algorithms and inapproximability results for every CSP? In *STOC*, pages 245–254, 2008.

[RS09a] P. Raghavendra and D. Steurer. Integrality Gaps for Strong SDP Relaxations of UNIQUE GAMES. In *FOCS*, pages 575–585, 2009.

[RS09b] P. Raghavendra and D. Steurer. Towards computing the Grothendieck constant. In *SODA*, pages 525–534, 2009.

[RS10] P. Raghavendra and D. Steurer. Graph expansion and the unique games conjecture. In *STOC*, pages 755–764, 2010.

[RST10] P. Raghavendra, D. Steurer, and P. Tetali. Approximations for the isoperimetric and spectral profile of graphs and related parameters. In *STOC*, pages 631–640, 2010.

[RST12] P. Raghavendra, D. Steurer, and M. Tulsiani. Reductions between Expansion Problems. In *IEEE Conference on Computational Complexity*, pages 64–73, 2012.

[Rez00] B. Reznick. Some concrete aspects of Hilbert's 17th problem. *Contemporary Mathematics*, 253:251–272, 2000.

[Sch08] G. Schoenebeck. Linear Level Lasserre Lower Bounds for Certain k-CSPs. In *FOCS*, pages 593–602, 2008.

[SA90] H. D. Sherali and W. P. Adams. A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems. *SIAM Journal on Discrete Mathematics*, 3(3):411–430, 1990.

[Sho87] N. Shor. An approach to obtaining global extremums in polynomial mathematical programming problems. *Cybernetics and Systems Analysis*, 23(5):695–700, 1987.

[SWW12] D. A. Spielman, H. Wang, and J. Wright. Exact Recovery of Sparsely-Used Dictionaries. *Journal of Machine Learning Research - Proceedings Track*, 23:37.1–37.18, 2012.

[Ste74] G. Stengle. A Nullstellensatz and a Positivstellensatz in semialgebraic geometry. *Mathematische Annalen*, 207(2):87–97, 1974.

[Ste10] D. Steurer. Fast SDP Algorithms for Constraint Satisfaction Problems. In *SODA*, pages 684–697, 2010.

[Tre12] L. Trevisan. On Khot's Unique Games Conjecture. *Bulletin (New Series) of the American Mathematical Society*, 49(1), 2012.

[Tul09] M. Tulsiani. CSP gaps and reductions in the lasserre hierarchy. In *STOC*, pages 303–312, 2009.

Boaz Barak

E-mail: info@boazbarak.org

David Steurer

E-mail: dsteurer@cs.cornell.edu

ECCC ISSN 1433-8092

http://eccc.hpi-web.de