Rounding Sum-of-Squares Relaxations

Boaz Barak* Jonathan Kelner† David Steurer‡

June 10, 2014

Abstract

We present a general approach to rounding semidefinite programming relaxations obtained by the Sum-of-Squares method (Lasserre hierarchy). Our approach is based on using the connection between these relaxations and the Sum-of-Squares proof system to transform a *combining algorithm*—an algorithm that maps a distribution over solutions into a (possibly weaker) solution—into a *rounding algorithm* that maps a solution of the relaxation to a solution of the original problem.

Using this approach, we obtain algorithms that yield improved results for natural variants of three well-known problems:

1. We give a quasipolynomial-time algorithm that approximates $\max_{\|x\|_2=1} P(x)$ within an additive factor of $\varepsilon\|P\|_{spectral}$, where $\varepsilon > 0$ is a constant, P is a degree $d = O(1)$, n-variate polynomial with nonnegative coefficients, and $\|P\|_{spectral}$ is the spectral norm of a matrix corresponding to P's coefficients. Beyond being of interest in its own right, obtaining such an approximation for general polynomials (with possibly negative coefficients) is a long-standing open question in quantum information theory, and our techniques have already led to improved results in this area (Brandão and Harrow, STOC '13).
2. We give a polynomial-time algorithm that, given a subspace $V \subseteq \mathbb{R}^n$ of dimension d that (almost) contains the characteristic function of a set of size n/k, finds a vector $v \in V$ that satisfies $\mathbb{E}_i v_i^4 \geqslant \Omega(d^{-1/3}k(\mathbb{E}_i v_i^2)^2)$. This is a natural analytical relaxation of the problem of finding the sparsest element in a subspace, and is also motivated by a connection to the Small Set Expansion problem shown by Barak et al. (STOC 2012). In particular our results yield an improvement of the previous best known algorithms for small set expansion in a certain range of parameters.
3. We use this notion of L_4 vs. L_2 sparsity to obtain a polynomial-time algorithm with substantially improved guarantees for recovering a planted sparse vector v in a random d-dimensional subspace of $\mathbb{R}^n$. If v has μn nonzero coordinates, we can recover it with high probability whenever $\mu \leqslant O(\min(1, n/d^2))$. In particular, when $d \leqslant \sqrt{n}$, this recovers a planted vector with up to $\Omega(n)$ nonzero coordinates. When $d \leqslant n^{2/3}$, our algorithm improves upon existing methods based on comparing the L_1 and L_∞ norms, which intrinsically require $\mu \leqslant O\left(1/\sqrt{d}\right)$.

*Microsoft Research.

†Department of Mathematics, Massachusetts Institute of Technology.

‡Department of Computer Science, Cornell University.

Contents

1 Introduction

Convex programming is the algorithmic workhorse behind many applications in computer science and other fields. But its power is far from understood, especially in the case of *hierarchies* of linear programming (LP) and semidefinite programming (SDP) relaxations. These are systematic approaches to make a convex relaxation tighter by adding to it more constraints. Various such hierarchies have been proposed independently by researchers from several communities [Sho87, SA90, LS91, Nes00, Par00, Las01]. In general, these hierarchies are parameterized by a number ℓ called their *level*. For problems on n variables, the hierarchy of the ℓ^{th} level can be optimized in $n^{O(\ell)}$ time, where for the typical domains used in CS (such as $\{0,1\}^n$ or the n-dimensional unit sphere), n rounds correspond to the exact (or near exact) solution by brute force exponential-time enumeration.

There are several strong *lower bounds* (also known as *integrality gaps*) for these hierarchies, in particular showing that $\omega(1)$ levels (and often even $n^{\Omega(1)}$ or $\Omega(n)$ levels) of many such hierarchies can't improve by much on the known polynomial-time approximation guarantees for many NP hard problems, including SAT, Independent-Set, Max-Cut and more [Gri01b, Gri01a, ABLT06, dlVKM07, Sch08, Tul09, CMM09, BGMT12, BCV+12]. Unfortunately, there are many fewer *positive* results, and several of them only show that these hierarchies can match the performance of previously known (and often more efficient) algorithms, rather than using hierarchies to get genuinely new algorithmic results.[1] For example, Karlin, Mathieu and Nguyen [KMN11] showed that ℓ levels of the Sum of Squares hierarchy can approximate the Knapsack problem up to a factor of $1 + 1/\ell$, thus approaching the performance of the standard dynamic program. Guruswami and Sinop [GS11] and (independently) Barak, Raghavendra, and Steurer [BRS11] showed that some SDP hierarchies can match the performance of the [ABS10] algorithm for Small Set Expansion and Unique Games, and their techniques also gave improved results for some other problems (see also [RT12, AG11, AGS13]). Chlamtac and Singh [CS08] (building on [Chl07]) used hierarchies to obtain some new approximation guarantees for the independent set problem in 3-uniform hypergraphs. Bhaskara, Charikar, Chlamtac, Feige, and Vijayaraghavan [BCC+10] gave an LP-hierarchy based approximation algorithm for the k-densest subgraph problem, although they also showed a purely combinatorial algorithm with the same performance. The famous algorithm of Arora, Rao and Vazirani [ARV04] for Sparsest Cut can be viewed (in retrospect) as using a constant number of rounds of an SDP hierarchy to improve upon the performance of the basic LP for this problem. Perhaps the most impressive use of super-constant levels of a hierarchy to solve a new problem was the work of Brandão, Christandl and Yard [BCY11] who used an SDP hierarchy (first proposed by [DPS04]) to give a quasipolynomial time algorithm for a variant of the *quantum separability problem* of testing whether a given density matrix corresponds to a separable (i.e., non-entangled) quantum state or is ε-far from all such states (see Section 1.2).

One of the reasons for this paucity of positive results is that we have relatively few tools to *round* such convex hierarchies. A *rounding algorithm* maps a solution to the relaxation to a solution to the original program.[2] In the case of a hierarchy, the relaxation solution satisfies more constraints, but we do not always know how to take advantage of this when rounding. For example, [ARV04] used a very sophisticated analysis to get better rounding when the solution to a Sparsest Cut relaxation satisfies a constraint known as triangle inequalities, but we have no general tools to use the additional constraints that come from higher levels of the hierarchies, nor do we know if these can help in rounding or not. This lack of rounding techniques is particularly true for the *Sum of Squares* (SOS, also known as *Lasserre*) hierarchy [Par00, Las01].[3] This is the

[1] The book chapter [CT10] is a good source for several of the known upper and lower bounds, though it does not contain some of the more recent ones.

[2] While the name derives from the prototypical case of relaxing an integer program to a linear program by allowing the variables to take non-integer values, we use "rounding algorithm" for any mapping from relaxation solutions to actual solutions, even in cases where the actual solutions are themselves non-integer.

[3] While it is common in the TCS community to use *Lasserre* to describe the primal version of this SDP, and *Sum of Squares (SOS)*

strongest variant of the canonical semidefinite programming hierarchies, and has recently shown promise to achieve tasks beyond the reach of weaker hierarchies [BBH+12]. But there are essentially no general rounding tools that take full advantage of its power.[4]

In this work we propose a general approach to rounding SOS hierarchies, and instantiate this approach in two cases, giving new algorithms making progress on natural variants of two longstanding problems. Our approach is based on the intimate connection between the SOS hierarchy and the "Positivstellensatz"/"Sum of Squares" proof system. This connection was used in previous work for either negative results [Gri01b, Gri01a, Sch08], or positive results for specific instances [BBH+12, OZ13, KOTZ14], translating proofs of a bound on the actual value of these instances into proofs of bounds on the relaxation value. In contrast, we use this connection to give explicit rounding algorithms for general instances of certain computational problems.

1.1 The Sum of Squares hierarchy

Our work uses the *Sum of Squares (SOS)* semidefinite programming hierarchy and in particular its relationship with the Sum of Squares (or Positivstellensatz) proof system. We now briefly review both the hierarchy and proof system. See the introduction of [OZ13] and the monograph [Lau09] for a more in depth discussion of these concepts and their history. Underlying both the SDP and proof system is the natural approach to prove that a real polynomial P is nonnegative via showing that it equals a *sum of squares*: $P = \sum_{i=1}^{k} Q_i^2$ for some polynomials $Q_1, \ldots, Q_k$. The question of when a nonnegative polynomial has such a "certificate of non-negativity" was studied by Hilbert who realized this doesn't always hold and asked (as his 17*th* problem) whether a nonnegative polynomial is always a sum of squares of *rational* functions. This was proven to be the case by Artin, and also follows from the more general *Positivstellensatz* (or "Positive Locus Theorem") [Kri64, Ste74]. The Positivstellensatz/SOS proof system of Grigoriev and Vorobjov [GV01] is based on the Positivstellensatz as a way to refute the assertion that a certain set of polynomial equations

$$P_1(x_1, \ldots, x_n) = \ldots = P_k(x_1, \ldots, x_n) = 0 \tag{1.1}$$

can be satisfied by showing that there exists some polynomials $Q_1, \ldots, Q_k$ and a sum of squares polynomial S such that

$$\sum P_i Q_i = 1 + S \,. \tag{1.2}$$

([GV01] considered inequalities as well, although in our context one can always restrict to equalities without loss of generality.) One natural measure for the complexity of such proof is the *degree* of the polynomials $P_1 Q_1, \ldots, P_k Q_k$ and S.

The sum of squares semidefinite program was proposed independently by several authors [Sho87, Par00, Nes00, Las01] One way to describe it is as follows. If the set of equalities (1.1) is satisfiable then in particular there exists some random variable X over $\mathbb{R}^n$ such that

$$\mathbb{E}\, P_1(X_1, \ldots, X_n)^2 = \ldots = \mathbb{E}\, P_k(X_1, \ldots, X_n)^2 = 0 \,. \tag{1.3}$$

That is, X is some distribution over the non-empty set of solutions to (1.1).

For every degree ℓ, we can consider the linear operator $\mathcal{L} = \mathcal{L}_\ell$ that maps a polynomial P of degree at most ℓ into the number $\mathbb{E}\, P(X_1, \ldots, X_n)$. Note that by choosing the monomial basis, this operator can be described by a vector of length n^ℓ, or equivalently, by an $n^{\ell/2} \times n^{\ell/2}$ matrix. This operator satisfies the following conditions:

to describe the dual, in this paper we use the more descriptive SOS name for both programs. We note that in all the applications we consider, strong duality holds, and so these programs are equivalent.

[4] The closest general tool we are aware of is the repeated conditioning methods of [BRS11, GS11], though these can be implemented in weaker hierarchies too and so do not seem to use the full power of the SOS hierarchy. However, this technique does play a role in this work as well.

Normalization If P is the constant polynomial 1 then $\mathcal{L}P = 1$

Linearity $\mathcal{L}(P + Q) = \mathcal{L}P + \mathcal{L}Q$ for every P, Q of degree $\leqslant \ell$.

Positivity $\mathcal{L}P^2 \geqslant 0$ for every P of degree $\leqslant \ell/2$.

Following [BBH+12], we call a linear operator satisfying the above conditions *a level ℓ pseudoexpectation function*, or ℓ-p.e.f., and use the suggestive notation $\tilde{\mathbb{E}}\, P(X)$ to denote $\mathcal{L}P$. Correspondingly we will sometimes talk about a *level ℓ pseudodistribution* (or ℓ-p.d.) X, by which we mean that there is an associated level ℓ pseudoexpectation operator. (Note that if X is an actual random variable then it is in particular a level ℓ pseudodistribution for every ℓ.) Given the representation of $\mathcal{L}$ as an n^ℓ-dimensional vector it is possible to efficiently check that it satisfies the above conditions efficiently, and in particular the positivity condition corresponds to the fact that, when viewed as a matrix, $\mathcal{L}$ is positive semidefinite. Thus it is also possible to optimize over the set of operators satisfying these conditions in time $n^{O(\ell)}$, and this optimization procedure is known as the SOS SDP hierarchy. Clearly, as ℓ grows, the conditions become stricter. In Appendix A we collect some useful properties of these pseudoexpectations. In particular one can show (see Corollary A.3) that if $\tilde{\mathbb{E}}\, P^2(X) = 0$ then $\tilde{\mathbb{E}}\, P(X)Q(X) = 0$ for every polynomial Q (as long as Q, P have degrees at most $\ell/2$). Thus, if there is a refutation to (1.1) of the form (1.2) where all polynomials involved have degree at most ℓ then there would not exist a level 2ℓ pseudoexpectation operator satisfying (1.3). This connection goes both ways, establishing an equivalence between the degree of Positivstellensatz proofs and the level of the corresponding SOS relaxation.

Until recently, this relation was mostly used for *negative* results, translating proof complexity lower bounds into integrality gap results for the SOS hierarchy [BBH+12, OZ13, KOTZ14]. However, in 2012 Barak, Brandão, Harrow, Kelner, Steurer and Zhou [BBH+12] used this relation for *positive* results, showing that the SOS hierarchy can in fact solve some interesting instances of the Unique Games maximization problem that fool weaker hierarchies. Their idea was to use the analysis of the previous works that proved these integrality gaps for weaker hierarchies. Such proofs work by showing that (a) the weaker hierarchy outputs a large value on this particular instance but (b) the true value is actually small. [BBH+12]'s insight was that oftentimes the proof of (b) only uses arguments that can be captured by the SOS/Positivstellensatz proof system, and hence inadvertently shows that the SOS SDP value is actually small as well. Some follow up works [OZ13, KOTZ14] extended this to other instances, but all these results held for very specific instances which have been proven before to have small objective value.

In this work we use this relation to get some guarantees on the performance of the SOS SDP on *general* instances. We give a more detailed overview of our approach in Section 2, but the high level idea is as follows. For particular optimization problems, we design a "rounding algorithm" that on input the moment matrix of a distribution on *actual solutions* achieving a certain value v, outputs a solution with some value $\tilde{v}$ which is a function of v. We call such an algorithm a *combining algorithm*, since it "combines" a distribution over solutions into a single one. (Note that the solution output by the combining algorithm need not be in the support of the distribution, and generally, when $\tilde{v} \neq v$, it won't be.) We then "lift" the analysis of this combining algorithm into the SOS framework, by showing that all the arguments can be captured in this proof system. This in turns implies that the algorithm would still achieve the value $\tilde{v}$ even if it is only given a *pseudoexpectation* of the distribution of sufficiently high level ℓ, and hence in fact this combining algorithm is a rounding algorithm for the level ℓ SOS hierarchy. We apply this idea to obtain new results for two applications— optimizing polynomials with nonnegative coefficients over the unit sphere, and finding "analytically sparse" vectors inside a subspace.

Remark 1.1 (Relation to the Unique Games Conjecture.)**.** While the SOS hierarchy is relevant to many algorithmic applications, some recent work focused on its relation to Khot's Unique Games Conjecture (UGC) [Kho02]. On a high level, the UGC implies that the basic semidefinite program is an optimal efficient

algorithm for many problems, and hence in particular using additional constant or polylogarithmic levels of the SOS hierarchy will not help. More concretely, as discussed in Section 1.3 below, the UGC is closely related to the question of how hard it is to find sparse (or "analytically sparse") vectors in a given subspace. Our work shows how the SOS hierarchy can be useful in general, and in particular gives strong average-case results and nontrivial worst-case results for finding sparse vectors in subspaces. Therefore, it can be considered as giving some (far from conclusive) evidence that the UGC might be false.

1.2 Optimizing polynomials with nonnegative coefficients

Our first result yields an *additive* approximation to this optimization problem for polynomials with nonnegative coefficients, when the value is scaled by the spectral norm of an associated matrix. If P is an n-variate degree-t homogeneous polynomial with nonnegative coefficients, then it can be represented by a tensor $M \in \mathbb{R}^{n^t}$ such that $P(x) = M \cdot x^{\otimes t}$ for every $x \in \mathbb{R}^n$. It is convenient to state our result in terms of this tensor representation:

Theorem 1.2. *There is an algorithm A, based on $O(t^3 \log n/\varepsilon^2)$ levels of the SOS hierarchy, such that for every even*[5] *t and nonnegative $M \in \mathbb{R}^{n^t}$,*

$$\max_{\|x\|=1} M \cdot x^{\otimes t} \leqslant A(M) \leqslant \max_{\|x\|=1} M \cdot x^{\otimes t} + \varepsilon \|M\|_{spectral} ,$$

where $\cdot$ denotes the standard dot product, and $\|M\|_{spectral}$ denotes the spectral norm of M, when considered as an $n^{t/2} \times n^{t/2}$ matrix.

Note that the algorithm of Theorem 1.2 only uses a logarithmic number of levels, and thus it shows that this fairly natural polynomial optimization problem can be solved in quasipolynomial time, as opposed to the exponential time needed for optimizing over general polynomials of degree > 2. Indeed, previous work on the convergence of the Lasserre hierarchy for general polynomials [DW12] can be described in our language here as trying to isolate a solution in the support of the distribution, and this generally requires a linear number of levels. Obtaining the logarithmic bound here relies crucially on constructing a "combined" solution that is not necessarily in the support. The algorithm is also relatively simple, and so serves as a good demonstration of our general approach.

Relation to quantum information theory. An equivalent way to state this result is that we get an ε additive approximation in the case that $\|M\|_{spectral} \leqslant 1$, in which case the value $\max_{\|x\|=1} M \cdot x^{\otimes t}$ is in the interval $[0, 1]$. This phrasing is particularly natural in the context of quantum information theory. A general (potentially mixed) quantum state on 2ℓ-qubits is represented by an $n^2 \times n^2$ *density matrix* ρ for $n = 2^\ell$; ρ is a positive semidefinite matrix and has trace 1. If ρ is *separable*, which means that there is no entanglement between the first ℓ qubits and the second ℓ qubits, then $\rho = \mathbb{E}\, xx^* \otimes yy^*$ for some distribution over $x, y \in \mathbb{C}^n$, where v^* denotes the complex adjoint operation. If we further restrict the amplitudes of the states to be real, and enforce symmetry on the two halves, then this would mean that $\rho = \mathbb{E}\, x^{\otimes 4}$. (All our results should generalize to states without those restrictions to symmetry and real amplitudes, which we make just to simplify the statement of the problem and the algorithm.) A quantum *measurement operator* over this space is an $n^2 \times n^2$ matrix M of spectral norm $\leqslant 1$. The probability that the measurement accepts a state ρ is $\mathrm{Tr}(M\rho)$. Finding an algorithm that, given a measurement M, finds the separable state ρ that maximizes this probability is an important question in quantum information theory which amounts to finding a classical upper bound for the complexity class **QMA**(2) of Quantum Merlin Arthur proofs with two independent provers [HM13]. Note that if we consider symmetric real states then this is the same as finding $\mathrm{argmax}_{\|x\|=1} M \cdot x^{\otimes 4}$, and hence dropping the non-negativity constraint in our result would resolve this longstanding open problem. There is a

[5] The algorithm easily generalizes to polynomials of odd degree t and to non-homogenous polynomials, see Remark 3.5.

closely related dual form of this question, known as the *quantum separability problem*, where one is given a quantum state ρ and wants to find the test M that maximizes

$$\mathrm{Tr}(M\rho) - \max_{\rho' \text{ separable}} \mathrm{Tr}(M\rho') \tag{1.4}$$

or to simply distinguish between the case that this quantity is at least ε and the case that ρ is separable. The best result known in this area is the paper [BCY11] mentioned above, which solved the distinguishing variant of quantum separability problem in the case that measurements are restricted to so-called *Local Operations and one-way classical communication* (one-way LOCC) operators. However, they did not have a rounding algorithm, and in particular did not solve the problem of actually finding a separable state that maximizes the probability of acceptance of a given one-way LOCC measurement. The techniques of this work were used by Brandão and Harrow [BH13] to solve the latter problem, and also greatly simplify the proof of [BCY11]'s result, which originally involved relations between several measures of entanglement proved in several papers.[6] For completeness, in Appendix C we give a short proof of this result, specialized to the case of real vectors and polynomials of degree four (corresponding to quantum states of two systems, or two prover QMA proofs). We also show in Appendix B that in the case the measurement satisfies the stronger condition of having its ℓ_2 (i.e., Frobenius) norm be at most 1, there is a simpler and more efficient algorithm for estimating the maximum probability the measurement accepts a separable state, giving an ε additive approximation in $\mathrm{poly}(n)\exp(\mathrm{poly}(1/\varepsilon))$ time. In contrast, [BCY11]'s algorithm took quasipolynomial time even in this case.

Relation to small set expansion. Nonnegative tensors also arise naturally in some applications, and in particular in the setting of small set expansion for Cayley graphs over the cube, which was our original motivation to study them. In particular, one corollary of our result is:

Corollary 1.3 (Informally stated)**.** *There is an algorithm A, based on* $\mathrm{poly}(K(G))\log n$ *levels of the SOS hierarchy, that solves the* Small Set Expansion *problem on Cayley graphs G over* $\mathbb{F}_2^\ell$ *(where* $\ell = \log n$*) where* $K(G)$ *is a parameter bounding the spectral norm of an operator related to G's top eigenspace.*

We discuss the derivation and the meaning of this corollary in Section 6 but note that the condition of having small value $K(G)$ seems reasonable. Having $K(G) = O(1)$ implies that the graph is a small set expander, and in particular the known natural examples of Cayley graphs that are small set expanders, such as the noisy Boolean hypercube and the "short code" graph of [BGH+12] have $K(G) = O(1)$. Thus a priori one might have thought that a graph that is hard to distinguish from small set expanders would have a small value of $K(G)$.

1.3 Optimizing hypercontractive norms and finding analytically sparse vectors

Finding a sparse nonzero vector inside a d dimensional linear subspace $V \subseteq \mathbb{R}^n$ is a natural task arising in many applications in machine learning and optimization (e.g., see [DH13] and the references therein). Related problems are known under many names including the "sparse null space", "dictionary learning", "blind source separation", "min unsatisfy", and "certifying restricted isometry property" problems. (These problems all have the same general flavor but differ on various details such as worst-case vs. average case, affine vs. linear subspaces, finding a single vector vs. a basis, and more.) Problems of this type are often NP-hard, with some hardness of approximation results known, and conjectured average-case hardness (e.g., see [ABSS97, KZ12, GN10] and the references therein).

We consider a natural relaxation of this problem, which we call the *analytically sparse vector* problem (ASVP), which assumes the input subspace (almost) contains an actually sparse 0/1 vector, but allows the

[6] The paper [BH13] was based on a previous version of this work [BKS12] that contained only the results for nonnegative tensors.

algorithm to find a vector $v \in V$ that is only "analytically sparse" in the sense that $\|v\|_4/\|v\|_2$ is large. More formally, for $q > p$ and $\mu > 0$, we say that a vector v is μ L_q/L_p-*sparse* if $(\mathbb{E}_i v_i^q)^{1/q}/(E_i v_i^p)^{1/p} \geqslant \mu^{1/q-1/p}$. That is, a vector is μ L_q/L_p-sparse if it has the same q-norm vs p-norm ratio as a $0/1$ vector of measure at most μ.

This is a natural relaxation, and similar conditions have been considered in the past. For example, Spielman, Wang, and Wright [SWW12] used in their work on dictionary learning a subroutine that finds a vector v in a subspace that maximizes the ratio $\|v\|_\infty/\|v\|_1$ (which can be done efficiently via n linear programs). However, because any subspace of dimension d contains an $O(1/\sqrt{d})$ L_∞/L_1-sparse vector, this relaxation can only detect the existence of vectors that are supported on less than $O(n/\sqrt{d})$ coordinates. Some works have observed that the L_2/L_1 ratio is a much better proxy for sparsity [ZP01, DH13], but computing it is a non-convex optimization problem for which no efficient algorithm is known. Similarly, the L_4/L_2 ratio is a good proxy for sparsity for subspaces of small dimension (say $d = O(\sqrt{n})$) but it is non-convex, and it is not known how to efficiently optimize it.[7]

Nevertheless, because $\|v\|_4^4$ is a degree 4 polynomial, the problem of maximizing it for $v \in V$ of unit norm amounts to a polynomial maximization problem over the sphere, that has a natural SOS program. Indeed, [BBH+12] showed that this program does in fact yield a good approximation of this ratio for random subspaces. As we show in Section 5, we can use this to improve upon the results of [DH13] and find planted sparse vectors in random subspaces that are of not too large a dimension:

Theorem 1.4. *There is a constant $c > 0$ and an algorithm A, based on $O(1)$-rounds of the SOS program, such that for every vector $v_0 \in \mathbb{R}^n$ supported on at most $cn \min(1, n/d^2)$ coordinates, if $v_1, \ldots, v_d$ are chosen independently at random from the Gaussian distribution on $\mathbb{R}^n$, then given any basis for $V = \mathrm{span}\{v_0, \ldots, v_d\}$ as input, A outputs an ε-approximation of v_0 in $\mathrm{poly}(n, \log(1/\varepsilon))$ time.*

In particular, we note that this recovers a planted vector with up to $\Omega(n)$ nonzero coordinates when $d \leqslant \sqrt{n}$, and it can recover vectors with more than the $O(n/\sqrt{d})$ nonzero coordinates that are necessary for existing techniques whenever $d \ll n^{2/3}$.

Perhaps more significantly, we prove the following nontrivial *worst-case* bound for this problem:

Theorem 1.5. *There is a polynomial-time algorithm A, based on $O(1)$ levels of the SOS hierarchy, that on input a d-dimensional subspace $V \subseteq \mathbb{R}^n$ such that there is a $0/1$-vector $v \in V$ with at most μn nonzero coordinates, $A(V)$ outputs an $O(\mu d^{1/3})$ L_4/L_2-sparse vector in V.*

Moreover, this holds even if v is not completely inside V but only satisfies $\|\Pi_V v\|_2^2 \geqslant (1-\varepsilon)\|v\|_2^2$, for some absolute constant $\varepsilon > 0$, where Π_V is the projector to V.

The condition that the vector is $0/1$ can be significantly relaxed, see Remark 4.12. Theorem 4.1 is also motivated by the Small Set Expansion problem. The current best known algorithms for Small Set Expansion and Unique Games [ABS10] reduce these problems into the task of finding a sparse vector in a subspace, and then find this vector using brute force enumeration. This enumeration is the main bottleneck in improving the algorithms' performance.[8] [BBH+12] showed that, at least for the Small Set Expansion

[7] It seems that what makes our relaxation different from the original problem is not so much the qualitative issue of considering analytically sparse vectors as opposed to actually sparse vectors, but the particular choice of the L_4/L_2 ratio, which on one hand seems easier (even if not truly easy) to optimize over than the L_2/L_1 ratio, but provides better guarantees than the L_∞/L_1 ratio. However, this choice does force us to restrict our attention to subspaces of low dimension, while in some applications such as certifying the restricted isometry property, the subspace in question is often the kernel of a "short and fat" matrix, and hence is almost full dimensional. Nonetheless, we believe it should be possible to extend our results to handle subspaces of higher dimension, perhaps at the some mild cost in the number of rounds.

[8] This is the only step that takes super-polynomial time in [ABS10]'s algorithm for Small Set Expansion. Their algorithm for Unique Games has an additional divide and conquer step that takes subexponential time, but, in our opinion, seems less inherently necessary. Thus we conjecture that if the sparse-vector finding step could be sped up then it would be possible to speed up the algorithm for both problems.

question, finding an L_4/L_2 *analytically sparse* vector would be good enough. Using their work we obtain the following corollary of Theorem 1.5:

Corollary 1.6 (Informally stated)**.** *There is an algorithm that given an n-vertex graph G that contains a set S of size $o(n/d^{1/3})$ with expansion at most ε, outputs a set S' of measure $\delta = o(1)$ with expansion bounded away from 1, i.e., $\Phi(S) \leqslant 1 - \Omega(1)$, where d is the dimension of the eigenspace of G's random walk matrix corresponding to eigenvalues larger than $1 - O(\varepsilon)$.*

The derivation and meaning of this result is discussed in Section 6. We note that this is the first result that gives an approximation of this type to the small set expansion in terms of the dimension of the top eigenspace, as opposed to an approximation that is polynomial in the number of vertices.

1.4 Related work

Our paper follows the work of [BBH+12], that used the language of pseudoexpectation to argue that the SOS hierarchy can solve specific interesting instances of Unique Games, and perhaps more importantly, how it is often possible to almost mechanically "lift" arguments about actual distributions to the more general setting of pseudodistribution. In this work we show how the same general approach be used to obtain positive results for general instances.

The fact that LP/SDP solutions can be viewed as expectations of distributions is well known, and several rounding algorithms can be considered as trying to "reverse engineer" a relaxation solution to get a good distribution over actual solutions.

Techniques such as randomized rounding, the hyperplane rounding of [GW95], and the rounding for TSP [GSS11, AKS12] can all be viewed in this way. One way to summarize the conceptual difference between our techniques and those approaches is that these previous algorithms often considered the relaxation solution as giving moments of an *actual* distribution on *"fake"* solutions. For example, in [GW95]'s Max Cut algorithm, where actual solutions are modeled as vectors in $\{\pm 1\}^n$, the SDP solution is treated as the moment matrix of a Gaussian distribution over real vectors that are not necessarily ± 1-valued. Similarly in the TSP setting one often considers the LP solution to yield moments of a distribution over spanning trees that are not necessarily TSP tours. In contrast, in our setting we view the solution as providing moments of a *"fake"* distribution on *actual* solutions.

Treating solutions explicitly as "fake distributions" is prevalent in the literature on *negative results* (i.e., integrality gaps) for LP/SDP hierarchies. For hierarchies weaker than SOS, the notion of "fake" is different, and means that there is a collection of local distributions, one for every small subset of the variables, that are consistent with one another but do not necessarily correspond to any global distribution. Fake distributions are also used in some positive results for hierarchies, such as [BRS11, GS11], but we make this more explicit, and, crucially, make much heavier use of the tools afforded by the Sum of Squares relaxation.

The notion of a "combining algorithm" is related to the notion of *polymorphisms* [BJK05] in the study of constraint satisfaction problems. A polymorphism is a way to combine a number of satisfying assignments of a CSP into a different satisfying assignments, and some relations between polymorphism, their generalization to approximation problems, rounding SDP's are known (e.g., see the talk [Rag10]). The main difference is polymorphisms operate on each bit of the assignment independently, while we consider here combining algorithms that can be very global.

In a follow up (yet unpublished) work, we used the techniques of this paper to obtain improved results for the *sparse dictionary learning* problem, recovering a set of vectors $x_1, \ldots, x_m \in \mathbb{R}^n$ from random samples of μ-sparse linear combinations of them for any $\mu = o(1)$, improving upon previous results that required $\mu \ll 1/\sqrt{n}$ [SWW12, AGM13, AAJ+13].

1.5 Organization of this paper

In Section 2 we give a high level overview of our general approach, as well as proof sketches for (special cases of) our main results. Section 3 contains the proof of Theorem 1.2— a quasipolynomial time algorithm to optimize polynomials with nonnegative coefficients over the sphere. Section 4 contains the proof of Theorem 1.5— a polynomial time algorithm for an $O(d^{1/3})$-approximation of the "analytical sparsest vector in a subspace" problem. In Section 5 we show how to use the notion of analytical sparsity to solve the question of finding a "planted" sparse vector in a random subspace. Section 6 contains the proofs of Corollaries 1.3 and 1.6 of our results to the small set expansion problem. Appendix A contains certain technical lemmas showing that pseudoexpectation operators obey certain inequalities that are true for actual expectations. Appendix C contains a short proof (written in classical notation, and specialized to the real symmetric setting) of [BCY11, BH13]'s result that the SOS hierarchy yields a good approximation to the acceptance probability of QMA(2) verifiers / measurement operators that have bounded one-way LOCC norm. Appendix B shows a simpler algorithm for the case that the verifier satisfies the stronger condition of a bounded $L2$ (Frobenius) norm. For the sake of completeness, Appendix D reproduces the proof from [BBH+12] of the relation between hypercontractive norms and small set expansion. Our papers raises many more questions than it answers, and some discussion of those appears in Section 7.

1.6 Notation

Norms and inner products. We will use linear subspaces of the form $V = R^{\mathcal{U}}$ where $\mathcal{U}$ is a finite set with an associated measure $\mu : \mathcal{U} \to [0,\infty]$. The p-norm of a vector $v \in V$ is defined as $\|v\|_p = (\sum_{\omega\in\mathcal{U}} \mu(\omega)|v_\omega|^p)^{1/p}$. Similarly, the inner product of $v, w \in V$ is defined as $\langle u, v\rangle = \sum_{\omega\in\mathcal{U}} \mu(\omega)u_\omega v_\omega$. We will only use two measures in this work: the *counting measure*, where $\mu(\omega) = 1$ for every $\omega \in \mathcal{U}$, and the *uniform measure*, where $\mu(\omega) = 1/|\mathcal{U}|$ for all $\omega \in \mathcal{U}$. (The norms corresponding to this measure are often known as the *expectation* norms.) We will use vector notation (i.e., letters such as u, v, and indexing of the form u_i) for elements of subspaces with the counting measure, and function notation (i.e., letters such as f, g and indexing of the form $f(x)$) for elements of subspaces with the uniform measure. The dot product notation $u \cdot v$ will be used exclusively for the inner product with the counting measure.

Pseudoexpectations. We use the notion of *pseudoexpectations* from [BBH+12]. A *level ℓ pseudoexpectation function* (ℓ-p.e.f.) $\tilde{\mathbb{E}}_{\mathcal{X}}$ is an operator mapping a polynomial P of degree at most ℓ into a number denoted by $\tilde{\mathbb{E}}_{x\sim\mathcal{X}} P(x)$ and satisfying the linearity, normalization, and positivity conditions as stated in Section 1.1. We sometimes refer to $\mathcal{X}$ as a *level ℓ pseudodistribution* (ℓ-p.d.) by which we mean that there exists an associated pseudoexpectation operator.[9] We will sometimes use the notation $\tilde{\mathbb{E}}\, P(\mathcal{X})$ when $\mathcal{X}$ is an actual random variable, in which case $\tilde{\mathbb{E}}\, P(\mathcal{X})$ simply equals $\mathbb{E}\, P(\mathcal{X})$. (We do so when we present arguments for actual distributions that we will later want to generalize to pseudodistributions.)

If P, Q are polynomials of degree at most $\ell/2$, and $\tilde{\mathbb{E}}_{\mathcal{X}}$ is an ℓ-p.e.f., we say that $\tilde{\mathbb{E}}_{\mathcal{X}}$ is *consistent* with the constraint $P(x) \equiv 0$ if it satisfies $\tilde{\mathbb{E}}_{x\sim\mathcal{X}} P(x)^2 = 0$. We say that it is consistent with the constraint $Q(x) \geqslant 0$, if it consistent with the constraint $Q(x) - S(x) \equiv 0$ for some polynomial S of degree $\leqslant \ell/2$ which is a *sum of squares*. (In the context of optimization, to enforce the inequality constraint $Q(x) \geqslant 0$, it is always possible to add an auxiliary variable y and then enforce the equality constraint $Q(x) - y^2 \equiv 0$.) Appendix A contains several useful facts about pseudoexpectations.

2 Overview of our techniques

Traditionally to design a mathematical-programming based approximation algorithm for some optimization problem O, one first decides what the relaxation is— i.e., whether it is a linear program, semidefinite program, or some other convex program, and what constraints to put in. Then, to demonstrate that the value of the

[9] In the paper [BBH+12] we used the name *level ℓ fictitious random variable* for $\mathcal{X}$, but we think the name pseudodistribution is better as it is more analogous to the name pseudoexpectation. The name "pseudo random variable" would of course be much too confusing.

program is not too far from the actual value, one designs a *rounding algorithm* that maps a solution of the convex program into a solution of the original problem of approximately the same value. Our approach is conceptually different— we design the rounding algorithm first, analyze it, and only then come up with the relaxation.

Initially, this does not seem to make much sense— how can you design an algorithm to round solutions of a relaxation when you don't know what the relaxation is? We do this by considering an idealized version of a rounding algorithm which we call a *combining algorithm*. Below we discuss this in more detail but roughly speaking, a combining algorithm maps a distribution over *actual solutions* of O into a single solution (that may or may not be part of the support of this distribution). This is a potentially much easier task than rounding relaxation solutions, and every rounding algorithm yields a combining algorithm. In the other direction, every combining algorithm yields a rounding algorithm for *some* convex programming relaxation, but in general that relaxation could be of exponential size. Nevertheless, we show that in several interesting cases, it is possible to transform a combining algorithm into a rounding algorithm for a not too large relaxation that we can efficiently optimize over, thus obtaining a feasible approximation algorithm. The main tool we use for that is the *Sum of Squares* proof system, which allows to lift certain arguments from the realm of combining algorithms to the realm of rounding algorithms.

We now explain more precisely the general approach, and then give an overview of how we use this approach for our two applications— finding "analytically sparse" vectors in subspaces, and optimizing polynomials with nonnegative coefficients over the sphere.

Consider a general optimization problem of minimizing some objective function in some set S, such as the n dimensional Boolean hypercube or the unit sphere. A *convex relaxation* for this problem consists of an embedding that maps elements in S into elements in some convex domain, and a suitable way to generalize the objective function to a convex function on this domain. For example, in linear programming relaxations we typically embed $\{0,1\}^n$ into the set $[0,1]^n$, while in semidefinite programming relaxations we might embed $\{0,1\}^n$ into the set of $n \times n$ positive semidefinite matrices using the map $x \mapsto X$ where $X_{i,j} = x_i x_j$. Given this embedding, we can use convex programming to find the element in the convex domain that maximizes the objective, and then use a *rounding algorithm* to map this element back into the domain S in a way that approximately preserves the objective value.

A *combining algorithm* C takes as input a *distribution* $\mathcal{X}$ over solutions in S and maps it into a single element $C(\mathcal{X})$ of S, such that the objective value of $C(\mathcal{X})$ is approximately close to the expected objective value of a random element in $\mathcal{X}$. Every rounding algorithm R yields a combining algorithm C. The reason is that if there is some embedding f mapping elements in S into some convex domain T, then for every distribution $\mathcal{X}$ over S, we can define $y_{\mathcal{X}}$ to be $\mathbb{E}_{x \in \mathcal{X}} f(x)$. By convexity, $y_{\mathcal{X}}$ will be in T and its objective value will be at most the average objective value of an element in $\mathcal{X}$. Thus if we define $C(\mathcal{X})$ to output $R(y_{\mathcal{X}})$ then C will be a combining algorithm with approximation guarantees at least as good as R's.

In the other direction, because the set of distributions over S is convex and can be optimized over by an $O(|S|)$-sized linear program, every combining algorithm can be viewed as a rounding algorithm for this program. However, $|S|$ is typically exponential in the bit description of the input, and hence this is not a very useful program. In general, we cannot improve upon this, because there is always a trivially lossless combining algorithm that "combines" a distribution $\mathcal{X}$ into a single solution x of the same expected value by simply sampling x from $\mathcal{X}$ at random. Thus for problems where getting an exact value is exponentially hard, this combining algorithm cannot be turned into a rounding algorithm for a subexponential-sized efficiently-optimizable convex program. However it turns out that at least in some cases, *nontrivial* combining algorithms can be turned into a rounding algorithm for an *efficient* convex program. A nontrivial combining algorithm C has the form $C(\mathcal{X}) = C'(M(\mathcal{X}))$ where C' is an efficient (say polynomial or quasipolynomial time) algorithm and $M(\mathcal{X})$ is a short (say polynomial or quasipolynomial size) *digest* of the distribution $\mathcal{X}$. In all the cases we consider, $M(\mathcal{X})$ will consist of all the moments up to some level ℓ of the random variable $\mathcal{X}$,

or some simple functions of it. That is, typically $M(X)$ is a vector in $\mathbb{R}^{m^\ell}$ such that for every $i_1, \ldots, i_\ell \in [m]$, $M_{i_1,\ldots,i_\ell} = \mathbb{E}_{x\sim X} x_{i_1} \cdots x_{i_\ell}$. We do not have a general theorem showing that any nontrivial combining algorithm can be transformed into a rounding algorithm for an efficient relaxation. However, we do have a fairly general "recipe" to use the *analysis* of nontrivial combining algorithms to transform them into rounding algorithms. The key insight is that many of the tools used in such analyses, such as the Cauchy–Schwarz and Hölder inequalities, and other properties of distributions, fall under the "Sum of Squares" proof framework, and hence can be shown to hold even when the algorithm is applied not to actual moments but to so-called "pseudoexpectations" that arise from the SOS semidefinite programming hierarchy.

We now turn to giving a high level overview of our results. For the sake of presentations, we focus on certain special cases of these two applications, and even for these cases omit many of the proof details and only provide rough sketches of the proofs. The full details can be found in Sections 5, 4 and 3.

2.1 Finding a planted sparse vector in a random low-dimensional subspace

We consider the following natural problem, which was also studied by Demanet and Hand [DH13]. Let $f_0 \in \mathbb{R}^{\mathcal{U}}$ be a sparse function over some universe $\mathcal{U}$ of size n. That is, f_0 is supported on at most μn coordinates for some $\mu = o(1)$. Let V be the subspace spanned by f_0 and d random (say Gaussian) functions $f_1, \ldots, f_d \in \mathbb{R}^{\mathcal{U}}$. Can we recover f_0 from any basis for V?

Demanet and Hand showed that if μ is very small, specifically $\mu \ll 1/\sqrt{d}$, then f_0 would be the most L_∞/L_1-sparse function in V, and hence (as mentioned above) can be recovered efficiently by running n linear programs. The SOS framework yields a natural and easy to describe algorithm for recovering f_0 as long as μ is a sufficiently small constant and the dimension d is at most $O(\sqrt{n})$. The algorithm uses the SOS program for finding the most L_4/L_2-sparse function in V, which, as mentioned above, is simply the polynomial optimization problem of maximizing $\|f\|_4^4$ over f in the intersection of V and the unit Euclidean sphere.

Since f_0 itself is in particular μ L_4/L_2-sparse , the optimum for the program is at least $1/\mu$. Thus a combining algorithm would get as input a distribution $\mathcal{D}$ over functions $f \in V$ satisfying $\|f\|_2 = 1$ and $\|f\|_4^4 \geqslant 1/\mu$, and need to output a vector closely correlated with f_0.[10] (We use here the *expectation* norms, namely $\|f\|_p^p = \mathbb{E}_\omega |f(\omega)|^p$.) For simplicity, assume that the f_i's are orthogonal to f_0 (they are nearly orthogonal, and so everything we say below will still hold up to a sufficiently good approximation, see Section 5). In this case, we can write every f in the support of $\mathcal{D}$ as $f = \langle f_0, f\rangle f_0 + f'$ where $f' \in V' = \mathrm{span}\{f_1, \ldots, f_d\}$. It is not hard to show using standard concentration of measure results (see e.g., [BBH+12, Theorem 7.1]) that if $d = O(\sqrt{n})$ then every $f' \in V'$ satisfies

$$\|f'\|_4 \leqslant C\|f'\|_2 \,, \tag{2.1}$$

for some constant C. Therefore using triangle inequality, and using the fact that $\|f'\|_2 \leqslant \|f\|_2 = 1$, it must hold that

$$\mu^{-1/4} \leqslant \|f\|_4 \leqslant \langle f, f_0\rangle \mu^{-1/4} + C \tag{2.2}$$

or

$$\langle f, f_0\rangle \geqslant 1 - C\mu^{1/4} = 1 - o(1) \tag{2.3}$$

for $\mu = o(1)$.

In particular this implies that if we apply a singular value decomposition (SVD) to the second moment matrix D of $\mathcal{D}$ (i.e., $D = \mathbb{E}_{f\in\mathcal{D}} f^{\otimes 2}$) then the top eigenvector will have $1 - o(1)$ correlation with f_0, and hence we can simply output it as our solution.

[10] Such a closely correlated vector can be corrected to output f_0 exactly, see Section 5.

To make this combining algorithm into a rounding algorithm we use the result of [BBH+12] that showed that (2.1) can actually be proven via a sum of squares argument. Namely they showed that there is a degree 4 sum of squares polynomial S such that

$$\|\Pi' f\|_4^4 + S(f) = C^4 \|f\|_2^4 . \tag{2.4}$$

(2.4) implies that even if $\mathcal{D}$ is merely a *pseudodistribution* then it must satisfy (2.1). (When the latter is raised to the fourth power to make it a polynomial inequality.) We can then essentially follow the argument, proving a version of (2.2) raised to the 4^{th} power by appealing to the fact that pseudodistributions satisfy Hölder's inequality, (Corollary A.11) and hence deriving that $\mathcal{D}$ will satisfy (2.3), with possibly slightly worse constants, even when it is only a pseudodistribution.

In Section 5, we make this precise and extend the argument to obtain nontrivial (but weaker) guarantees when $d \geqslant \sqrt{n}$. We then show how to use an additional correction step to recover the original function f_0 up to arbitrary accuracy, thus boosting our approximation of f_0 into an essentially exact one.

2.2 Finding "analytically sparse" vectors in general subspaces

We now outline the ideas behind the proof of Theorem 4.1— finding analytically sparse vectors in *general* (as opposed to random) subspaces. This is a much more challenging setting than random subspaces, and indeed our algorithm and its analysis is more complicated (though still only uses a constant number of SOS levels), and at the moment, the approximation guarantee we can prove is quantitatively weaker. This is the most technically involved result in this paper, and so the reader may want to skip ahead to Section 2.3 where we give an overview of the simpler result of optimizing over polynomials with nonnegative coefficients.

We consider the special case of Theorem 4.1 where we try to distinguish between a YES case where there is a 0/1 valued $o(d^{-1/3})$-sparse function that is completely contained in the input subspace, and a NO case where every function in the subspace has its four norm bounded by a constant times its two norm. That is, we suppose that we are given some subspace $V \subseteq \mathbb{R}^{\mathcal{U}}$ of dimension d and a distribution $\mathcal{D}$ over functions $f : \mathcal{U} \to \{0,1\}$ in V such that $\mathbb{P}_{\omega \in \mathcal{U}}[f(\omega) = 1] = \mu$ for every f in the support of $\mathcal{D}$, and $\mu = o(d^{-1/3})$. The goal of our combining algorithm is to output some function $g \in V$ such that $\|g\|_4^4 = \mathbb{E}_\omega g(\omega)^4 \gg (\mathbb{E}_\omega g(\omega)^2)^2 = \|g\|_2^4$. (Once again, we use the *expectation* inner product and norms, with uniform measure over $\mathcal{U}$.)

Since the f's correspond to sets of measure μ, we would expect the inner product $\langle f, f' \rangle$ of a typical pair f, f' (which equals the measure of the intersection of the corresponding sets) to be roughly μ^2. Indeed, one can show that if the average inner product $\langle f, f' \rangle$ is $\omega(\mu^2)$ then it's easy to find such a desired function g. Intuitively, this is because in this case the distribution $\mathcal{D}$ of sets does not have an equal chance to contain all the elements in $\mathcal{U}$, but rather there is some set I of $o(|\mathcal{U}|)$ coordinates which is favored by $\mathcal{D}$. Roughly speaking, that would mean that a random linear combination g of these functions would have most of its mass concentrated inside this small set I, and hence satisfy $\|g\|_4 \gg \|g\|_2$. But it turns out that letting g be a random gaussian function matching the first two moments of $\mathcal{D}$ is equivalent to taking such a random linear combination, and so our combining algorithm can obtain this g using moment information alone.

Our combining algorithm will also try all n *coordinate projection* functions. That is, let δ_ω be the function such that $\delta_\omega(\omega')$ equals $n = |\mathcal{U}|$ if $\omega = \omega'$ and equals 0 otherwise, (and hence under our expectation inner product $f(\omega) = \langle f, \delta_\omega \rangle$). The algorithm will try all functions of the form $\Pi \delta_u$, where Π is the projector to the subspace V. Fairly straightforward calculations show that the 2-norm squared of such a function is expected to be $(d/n)\|\delta_\omega\|_2^2 = d$, and it turns out in our setting we can assume that the norm is well concentrated around this expectation (or else we'd be able to find a good solution in some other way). Thus, if coordinate projection fails then it must hold that

$$O(d^2) = O(\mathop{\mathbb{E}}_\omega \|\Pi \delta_\omega\|_2^4) \geqslant \mathop{\mathbb{E}}_\omega \|\Pi \delta_\omega\|_4^4 = \mathop{\mathbb{E}}_{\omega,\omega'} \langle \Pi \delta_\omega, \delta_{\omega'} \rangle^4 . \tag{2.5}$$

It turns out that (2.5) implies some nontrivial constraints on the distribution $\mathcal{D}$. Specifically we know that

$$\mu = \mathop{\mathbb{E}}_{f\sim\mathcal{D}} \|f\|_4^4 = \mathop{\mathbb{E}}_{f\sim\mathcal{D},\omega\in\mathcal{U}} \langle f,\delta_\omega\rangle^4 .$$

But since $f = \Pi f$ and Π is symmetric, the RHS is equal to

$$\mathop{\mathbb{E}}_{f\sim\mathcal{D},\omega\in\mathcal{U}} \langle f,\Pi\delta_\omega\rangle^4 = \langle \mathop{\mathbb{E}}_{f\sim\mathcal{D}} f^{\otimes 4}, \mathop{\mathbb{E}}_{\omega\in\mathcal{U}} (\Pi\delta_\omega)^{\otimes 4}\rangle \leqslant \|\mathop{\mathbb{E}}_{f\sim\mathcal{D}} f^{\otimes 4}\|_2 \|\mathop{\mathbb{E}}_{\omega\in\mathcal{U}} (\Pi\delta_\omega)^{\otimes 4}\|_2 ,$$

where the last inequality uses Cauchy–Schwarz. If we square this inequality we get that

$$\mu^2 \leqslant \langle \mathop{\mathbb{E}}_{f\sim\mathcal{D}} f^{\otimes 4}, \mathop{\mathbb{E}}_{f\sim\mathcal{D}} f^{\otimes 4}\rangle \langle \mathop{\mathbb{E}}_{\omega\in\mathcal{U}]} (\Pi\delta_\omega)^{\otimes 4}, \mathop{\mathbb{E}}_{\omega\in\mathcal{U}} (\Pi\delta_\omega)^{\otimes 4}\rangle = \left(\mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}} \langle f,f'\rangle^4\right)\left(\mathop{\mathbb{E}}_{\omega,\omega'} \langle \Pi\delta_\omega,\Pi\delta_{\omega'}\rangle^4\right) .$$

But since Π is a projector satisfying $\Pi = \Pi^2$, we can use (2.5) and obtain that

$$\Omega(\mu^2/d^2) \leqslant \mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}} \langle f,f'\rangle^4 .$$

Since $d = o(\mu^{-3})$ this means that

$$\mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}} \langle f,f'\rangle^4 \gg \mu^8 . \tag{2.6}$$

Equation (2.6), contrasted with the fact that $\mathbb{E}_{f,f'\sim\mathcal{D}}\langle f,f'\rangle = O(\mu^2)$, means that the inner product of two random functions in $\mathcal{D}$ is somewhat "surprisingly unconcentrated", which seems to be a nontrivial piece of information about $\mathcal{D}$.[11] Indeed, because the f's are nonnegative functions, if we pick a random u and consider the distribution $\mathcal{D}_u$ where the probability of every function is reweighed proportionally to $f(u)$, then intuitively that should increase the probability of pairs with large inner products. Indeed, as we show in Lemma A.4, one can use Hölder's inequality to prove that there exist $\omega_1,\ldots,\omega_4$ such that under the distribution $\mathcal{D}'$ where every element f is reweighed proportionally to $f(\omega_1)\cdots f(\omega_4)$, it holds that

$$\mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}'} \langle f,f'\rangle \geqslant \left(\mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}} \langle f,f'\rangle^4\right)^{1/4} . \tag{2.7}$$

(2.7) and (2.6) together imply that $\mathcal{E}_{f,f'\sim\mathcal{D}'}\langle f,f'\rangle \gg \mu^2$, which, as mentioned above, means that we can find a function g satisfying $\|g\|_4 \gg \|g\|_2$ by taking a gaussian function matching the first two moments of $\mathcal{D}'$.

Once again, this combining algorithm can be turned into an algorithm that uses $O(1)$ levels of the SOS hierarchy. The main technical obstacle (which is still not very hard) is to prove another appropriate generalization of Hölder's inequality for pseudoexpectations (see Lemma A.4). Generalizing to the setting that in the YES case the function is only approximately in the vector space is a bit more cumbersome. We need to consider apart from f the function $\overline{f}$ that is obtained by first projecting f to the subspace and then "truncating" it by rounding each coordinate where f is too small to zero. Because this truncation operation is not a low degree polynomial, we include the variables corresponding to $\overline{f}$ as part of the relaxation, and so our pseudoexpectation operator also contains the moments of these functions as well.

[11] Interestingly, this part of the argument does not require μ to be $o(d^{-1/3})$, and some analogous "non-concentration" property of $\mathcal{D}$ can be shown to hold for a hard to round $\mathcal{D}$ for any $\mu = o(1)$. However, we currently know how to take advantage of this property to obtain a combining algorithm only in the case that $\mu \ll d^{-1/3}$.

2.3 Optimizing polynomials with nonnegative coefficients

We now consider the task of maximizing a polynomial with nonnegative coefficients over the sphere, namely proving Theorem 3.1. We consider the special case of Theorem 3.1 where the polynomial is of degree 4. That is, we are given a parameter $\varepsilon > 0$ and an $n^2 \times n^2$ nonnegative matrix M with spectral norm at most 1 and want to find an ε additive approximation to the maximum of

$$\sum_{i,j,k,l} M_{i,j,k,l} x_i x_j x_k x_l \,, \tag{2.8}$$

over all $x \in R^n$ with $\|x\| = 1$, where in this section we let $\|x\|$ be the standard (counting) Euclidean norm $\|x\| = \sqrt{\sum_i x_i^2}$.

One can get some intuition for this problem by considering the case where M is 0/1 valued and x is $0/k^{-1/2}$ valued for some k. In this case one can think of M as a 4-uniform hypergraph on n vertices and x as a subset $S \subseteq [n]$ that maximizes the number of edges inside S divided by $|S|^2$, and so this problem is related to some type of a densest subgraph problem on a hypergraph.[12]

Let's assume that we are given a distribution $\mathcal{X}$ over unit vectors that achieve some value v in (2.8). This is a non convex problem, and so generally the average of these vectors would not be a good solution. However, it turns out that the vector x^* defined such that $x_i^* = \sqrt{\mathbb{E}_{x\sim\mathcal{X}} x_i^2}$ can sometimes be a good solution for this problem. Specifically, we will show that if it fails to give a solution of value at least $v - \varepsilon$, then we can find a new distribution $\mathcal{X}'$ obtained by reweighing elements $\mathcal{X}$ that is in some sense "simpler" than $\mathcal{X}$. More precisely, we will define some nonnegative potential function Ψ such that $\Psi(\mathcal{X}) \leqslant \log n$ for all $\mathcal{X}$ and $\Psi(\mathcal{X}') \leqslant \Psi(\mathcal{X}) - \Omega(\varepsilon^2)$ under the above conditions. This will show that we will need to use this reweighing step at most logarithmically many times.

Indeed, suppose that

$$\sum_{i,j,k,l} M_{i,j,k,l} x_i^* x_j^* x_k^* x_l^* = (x^{*\otimes 2})^T M x^{*\otimes 2} \leqslant v - \varepsilon \,. \tag{2.9}$$

We claim that in contrast

$$y^T M y \geqslant v \,, \tag{2.10}$$

where y is the n^2-dimensional vector defined by $y_{i,j} = \sqrt{\mathbb{E}_{x\sim\mathcal{X}} x_i^2 x_j^2}$. Indeed, (2.10) follows from the non-negativity of M and the Cauchy–Schwarz inequality since

$$v = \sum_{i,j,k,l} M_{i,j,k,l} \mathop{\mathbb{E}}_{x\in\mathcal{X}} x_i x_j x_k x_l \leqslant \sum_{i,j,k,l} M_{i,j,k,l} \sqrt{\mathop{\mathbb{E}}_{x\sim\mathcal{X}} x_i^2 x_j^2} \sqrt{\mathop{\mathbb{E}}_{x\sim\mathcal{X}} x_k^2 x_l^2} = y^T M y$$

Note that since $\mathcal{X}$ is a distribution over unit vectors, both x^* and y are unit vectors, and hence (2.9) and (2.10) together with the fact that M has bounded spectral norm imply that

$$\varepsilon \leqslant y^T M y - (x^{*\otimes 2})^T M x^{*\otimes 2} = (y - x^{*\otimes 2})^T M (y + x^{*\otimes 2}) \leqslant \|y - x^{*\otimes 2}\| \cdot \|y + x^{*\otimes 2}\| \leqslant 2\|y - x^{*\otimes 2}\| \,. \tag{2.11}$$

However, it turns out that $\|y - x^{*\otimes 2}\|$ equals $\sqrt{2}$ times the *Hellinger distance* of the two distributions D, D^* over $[n] \times [n]$ defined as follows: $\mathbb{P}[D = (i,j)] = \mathbb{E}\, x_i^2 x_j^2$ while $\mathbb{P}[D^* = (i,j)] = (\mathbb{E}\, x_i^2)(\mathbb{E}\, x_j^2)$ (see Section 3). At this point we can use standard information theoretic inequalities to derive from (2.11) that

[12] The condition of maximizing $|E(S)|/|S|^2$ is related to the *log density* condition used by [BCC+10] in their work on the densest subgraph problem, since, assuming that the set $[n]$ of all vertices is not the best solution, the set S satisfies that $\log_{|S|} |E(S)| > \log_n |E|$. However, we do not know how to use their algorithm to solve this problem. Beyond the fact that we consider the hypergraph setting, their algorithm manages to find a set of nontrivial density under the assumption that there is a "log dense" subset, but it is not guaranteed to find the "log dense" subset itself.

there is $\Omega(\varepsilon^2)$ *mutual information* between the two parts of D. Another way to say this is that the entropy of the second part of D drops on average by $\Omega(\varepsilon^2)$ if we condition on the value of the first part. To say the same thing mathematically, if we define $D(\mathcal{X})$ to be the distribution $(\mathbb{E}_{x\sim\mathcal{X}} x_1^2, \ldots, \mathbb{E}_{x\sim\mathcal{X}} x_n^2)$ over $[n]$ and $D(\mathcal{X}|i)$ to be the distribution $\frac{1}{\mathbb{E}_{x\sim\mathcal{X}} x_i^2}(\mathbb{E}_{x\sim\mathcal{X}} x_i^2 x_1^2, \ldots, \mathbb{E}_{x\sim\mathcal{X}} x_i^2 x_n^2)$ then

$$\mathop{\mathbb{E}}_{i\sim D(x)} H(\mathcal{X}|i) \leqslant H(\mathcal{X}) - \Omega(\varepsilon^2)\,.$$

But one can verify that $D(\mathcal{X}|i) = D(\mathcal{X}_i)$ where $\mathcal{X}_i$ is the distribution over x's such that $\mathbb{P}[\mathcal{X}_i = x] = x_i^2\,\mathbb{P}[\mathcal{X} = x]/\,\mathbb{E}_{\mathcal{X}}\, x_i^2$, which means that if we define $\Psi(\mathcal{X}) = H(D(\mathcal{X}))$ then we get that

$$\mathop{\mathbb{E}}_{i\sim D(x)} \Psi(\mathcal{X}_i) \leqslant \Psi(\mathcal{X}) - \Omega(\varepsilon^2)$$

and hence Ψ is exactly the potential function we were looking for.

To summarize our combining algorithm will do the following for $t = O(\log n/\varepsilon^2)$ steps: given the first moments of the distribution $\mathcal{X}$, define the vector x^* as above and test if it yields an objective value of at least $v - \varepsilon$. Otherwise, pick i with probability $\mathbb{E}_{x\sim\mathcal{X}} x_i^2$ and move to the distribution $\mathcal{X}_i$. Note that given d level moments for $\mathcal{X}$, we can compute the $d-1$ level moments of $\mathcal{X}_i$, and hence the whole algorithm can be carried out with only access to level $O(\log n/\varepsilon^2)$ moments of $\mathcal{X}$. We then see that the only properties of the moments used in this proof are linearity, the fact that $\sum x_i^2$ can always be replaced with 1 in any expression, and the Cauchy–Schwarz inequality used for obtaining (2.10). It turns out that all these properties hold even if we are not given access to the moments of a true distribution $\mathcal{X}$ but are only given access to a level d *pseudoexpectation* operator $\tilde{\mathbb{E}}$ for d equalling some constant times $\log n/\varepsilon^2$. Such pseudoexpectations operators can be optimized over in d levels of the SOS hierarchy, and hence this combining algorithm is in fact a rounding algorithm.

3 Approximation for nonnegative tensor maximization

In this section we prove Theorem 1.2, giving an approximation algorithm for the maximum over the sphere of a polynomial with nonnegative coefficients. We will work in the space R^n endowed with the *counting* measure for norms and inner products. We will define the *spectral norm* of a degree-$2t$ homogeneous polynomial M in $x = x(x_1, \ldots, x_n)$, denoted by $\|M\|_{\text{spectral}}$, to be the minimum of the spectral norm of Q taken over all quadratic forms Q over $(\mathbb{R}^n)^{\otimes t}$ such that $Q(x^{\otimes t}) = M(x)$ for every x. Note that we can compute the spectral norm of a homogeneous polynomial in polynomial time using semidefinite programming. Thus we can restate our main theorem of this section as:

Theorem 3.1 (Theorem 1.2, restated)**.** *Let M be a degree-$2t$ homogeneous polynomial in $x = (x_1, \ldots, x_n)$ with nonnegative coefficients. Then, there is an algorithm, based on $O(t^3 \log n/\varepsilon^2)$ levels of the SOS hierarchy, that finds a unit vector $x^* \in \mathbb{R}^n$ such that*

$$M(x^*) \geqslant \max_{x\in\mathbb{R}^n, \|x\|=1} M(x) - \varepsilon\|M\|_{\text{spectral}}\,.$$

To prove Theorem 3.1 we first come up with a *combining algorithm*, namely an algorithm that takes (the moment matrix of) a distribution $\mathcal{X}$ over unit vectors $x \in R^n$ such that $M(x) \geqslant v$ and find a unit vector x^* such that $M(x^*) \geqslant v - \varepsilon$. We then show that the algorithm will succeed even if $\mathcal{X}$ is merely a level $O(t \log n/\varepsilon^2)$ *pseudo distribution*; that is, the moment matrix is a pseudoexpectation operator. The combining algorithm is very simple:

Combining algorithm for polynomials with nonnegative coefficients:

Input: distribution $\mathcal{X}$ over unit $x \in \mathbb{R}^n$ such that $M(x) = \nu$.

Operation: Do the following for $t^2 \log n/\varepsilon^2$ steps:

Direct rounding: For $i \in [n]$, let $x^*_i = \sqrt{\mathbb{E}_{x\sim\mathcal{X}} x_i^2}$. If $M(x^*) \geqslant \nu - 4\varepsilon$ then output x^* and quit.

Conditioning: Try to find $i_1, \ldots, i_{t-1} \in [n]$ such that the distribution $\mathcal{X}_{i_1,\ldots,i_{t-1}}$ satisfies $\Psi(\mathcal{X}_{i_1,\ldots,i_{t-1}}) \leqslant \Psi(\mathcal{X}) - \varepsilon^2/t^2$, and set $\mathcal{X} = \mathcal{X}_{i_1,\ldots,i_{t-1}}$, where:

- $\mathcal{X}_{i_1,\ldots,i_{t-1}}$ is defined by letting $\mathbb{P}[\mathcal{X}_{i_1,\ldots,i_{t-1}} = x]$ be proportional to $\mathbb{P}[\mathcal{X} = x] \cdot \prod_{j=1}^{t-1} x_{i_j}^2$ for every $x \in \mathbb{R}^n$.
- $\Psi(\mathcal{X})$ is defined to be $H(A(\mathcal{X}))$ where $H(\cdot)$ is the Shannon entropy function and $A(\mathcal{X})$ is the distribution over $[n]$ obtained by letting $\mathbb{P}[A(\mathcal{X}) = i] = \mathbb{E}_{x\sim\mathcal{X}} x_i^2$ for every $i \in [n]$.

Clearly $\Psi(\mathcal{X})$ is always in $[0, \log n]$, and hence if we can show that we always succeed in at least one of the steps, then eventually the algorithm will output a good x^*. We now show that if the direct rounding step fails, then the conditioning step must succeed. We do the proof under the assumption that $\mathcal{X}$ is an actual distribution. Almost of all of this analysis holds verbatim when $\mathcal{X}$ is a pseudodistribution of level at least $2t^2 \log n/\varepsilon^2$, and we note the one step where the extension requires using a nontrivial (though easy to prove) property of pseudoexpectations, namely that they satisfy the Cauchy–Schwarz inequality.

Some information theory facts. We recall some standard relations between various entropy and distance measures. Let X and Y be two jointly distributed random variables. We denote the joint distribution of X and Y by $\{XY\}$, and their marginal distributions by $\{X\}$ and $\{Y\}$. We let $\{X\}\{Y\}$ denote the product of the distributions $\{X\}$ and $\{Y\}$ (corresponding to sampling X and Y independently from their marginal distribution). Recall that the *Shannon entropy* of X, denoted by $H(X)$, is defined to be $\sum_{x\in\mathrm{Support}(X)} \mathbb{P}[X = x] \log(-\mathbb{P}[X = x])$. The *mutual information* of X and Y is defined as $I(X, Y) \stackrel{\mathrm{def}}{=} H(X) - H(X \mid Y)$, where $H(X \mid Y)$ is *conditional entropy* of X with respect to Y, defined as $\mathbb{E}_{y\sim\{Y\}} H(X \mid Y = y)$. The *Hellinger distance* between two distributions p and q is defined by $d_H(p, q) \stackrel{\mathrm{def}}{=} \left(1 - \sum_i \sqrt{p_i q_i}\right)^{1/2}$. (In particular, $d_H(p, q)$ equals $1/\sqrt{2}$ times the Euclidean distance of the unit vectors $\sqrt{p}$ and $\sqrt{q}$.) The following inequality (whose proof follows by combining standard relations between the Hellinger distance, Kullback–Leibler divergence, and mutual information) would be useful for us

Lemma 3.2. *For any two jointly-distributed random variables X and Y,*

$$2d_H\Big(\{XY\}, \{X\}\{Y\}\Big)^2 \leqslant I(X, Y)$$

3.1 Direct Rounding

Given $\mathcal{X}$, we define the following correlated random variables $A_1, \ldots, A_t$ over $[n]$: the probability that $(A_1, \ldots, A_t) = (i_1, \ldots, i_t)$ is equal to $\mathbb{E}_{x\sim\mathcal{X}} x_{i_1}^2 \cdots x_{i_t}^2$. Note that for every i, the random variable A_i is distributed according to $A(\mathcal{X})$. (Note that even if $\mathcal{X}$ is only a pseudodistribution, $A_1, \ldots, A_t$ are actual random variables.) The following lemma gives a sufficient condition for our direct rounding step to succeed:

Lemma 3.3. *Let $M, \mathcal{X}$ be as above. If $d_H(\{A_1 \cdots A_t\}, \{A_1\} \cdots \{A_t\}) \leqslant \varepsilon$, then the unit vector x^* with $x^*_i = (\mathbb{E}_{x\sim\mathcal{X}} x_i^2)^{1/2}$ satisfies $M(x^*) \geqslant \nu - 4\varepsilon\|M\|_{\mathrm{spectral}}$. Moreover, this holds even if $\mathcal{X}$ is a level $\ell \geqslant 2t$ pseudodistribution.*

Proof. Let Q be a quadratic form with $Q(x^{\otimes t}) = M(x)$. Let $y \in (\mathbb{R}^n)^{\otimes t}$ be the vector $y_{i_1 \cdots i_t} = (\tilde{\mathbb{E}}_{x \sim_c X} x_{i_1}^2 \cdots x_{i_t}^2)^{1/2}$. Then,

$$\tilde{\mathbb{E}}\, M(x^*) = \langle \hat{M}, \tilde{\mathbb{E}}\, x^{*\otimes 2t} \rangle \leqslant \langle \hat{M}, y \otimes y \rangle = Q(y) \tag{3.1}$$

Here, the vector $\hat{M} \in (\mathbb{R}^n)^{\otimes 2t}$ contains the coefficients of M. In particular, $\hat{M} \geqslant 0$ entry-wise. The inequality in (3.1) uses Cauchy–Schwarz; namely that $\tilde{\mathbb{E}}\, x^\alpha x^\beta \leqslant (\tilde{\mathbb{E}}(x^\alpha)^2 \cdot \tilde{\mathbb{E}}(x^\beta)^2)^{1/2} = y_\alpha y_\beta$. The final equality in (3.1) uses that y is symmetric.

Next, we bound the difference between $Q(y)$ and $M(x^*)$

$$Q(y) - M(x^*) = Q(y) - Q(x^{*\otimes t}) = \langle y + x^{*\otimes t}, Q(y - x^{*\otimes t}) \rangle \leqslant \|Q\| \cdot \|y + x^{*\otimes t}\| \cdot \|y - x^{*\otimes t}\| \,. \tag{3.2}$$

(Here, $\langle \cdot, Q\, \cdot \rangle$ denotes the symmetric bilinear form corresponding to Q.)

Since both $x^{*\otimes t}$ and y are unit vectors, $\|y + x^{*\otimes t}\| \leqslant 2$. By construction, the vector y corresponds to the distribution $\{A_1 \cdots A_t\}$ and $x^{*\otimes t}$ corresponds to the distribution $\{A_1\} \cdots \{A_t\}$. In particular, $d_H(\{A_1 \cdots A_t\}, \{A_1\} \cdots \{A_t\}) = \frac{1}{\sqrt{2}} \|y - x^{*\otimes t}\|$. Together with the bounds (3.1) and (3.2),

$$M(x^*) \geqslant \tilde{\mathbb{E}}\, M(x) - 4\|Q\| \cdot d_H(\{A_1 \cdots A_t\}, \{A_1\} \cdots \{A_t\}) \,. \qquad \square$$

To verify this carries over when $\mathcal{X}$ is a pseudodistribution, we just need to use the fact that Cauchy–Schwarz holds for pseudoexpectations (Lemma A.2).

3.2 Making Progress

The following lemma shows that if the sufficient condition above is violated, then on expectation we can always make progress. (Because $A_1, \ldots, A_t$ are actual random variables, it automatically holds regardless of whether $\mathcal{X}$ is an actual distribution or a pseudodistribution.)

Lemma 3.4. *If $d_H(\{A_1 \cdots A_t\}, \{A_1\} \cdots \{A_t\}) \geqslant \varepsilon$, then $H(A_t \mid A_1 \cdots A_{t-1}) \leqslant H(A) - 2\varepsilon^2/t^2$*

Proof. The bound follows by combining a hybrid argument with Lemma 3.2.

Let $A'_1, \ldots, A'_t$ be independent copies of $A_1, \ldots, A_t$ so that

$$\{A_1 \cdots A_t \cdots A'_1 \cdots A'_t\} = \{A_1 \cdots A_t\}\{A_1\} \cdots \{A_t\} \,.$$

We consider the sequence of distributions $D_0, \ldots, D_t$ with

$$D_i = \{A_1 \cdot A_i \cdots A'_{i+1} \cdots A'_t\} \,.$$

By assumption, $d_H(D_0, D_t) \geqslant \varepsilon$. Therefore, there exists an index i such that $d_H(D_{i-1}, D_i) \geqslant \varepsilon/t$. Let $X = A_1 \cdots A_{i-1}$ and $Y = A_i A'_{i+1} \cdots A'_t$. Then, $D_i = \{XY\}$ and $D_{i-1} = \{X\}\{Y\}$. By Lemma 3.2,

$$H(Y) - H(Y \mid X) = I(X, Y) \geqslant 2 d_H(\{XY\}, \{X\}\{Y\})^2 \geqslant 2\varepsilon^2/t^2 \,.$$

Since $A'_{i+1}, \ldots, A'_t$ are independent of $A_1, \ldots, A_i$,

$$H(Y) - H(Y \mid X) = H(A_i) - H(A_i \mid A_1 \cdots A_{i-1}) \,.$$

By symmetry and the monotonicity of entropy under conditioning, we conclude

$$H(A_t \mid A_1 \cdots A_{t-1}) \leqslant H(A) - 2\varepsilon^2/t^2 \,. \qquad \square$$

Lemma 3.4 implies that if our direct rounding fails then the expectation of $H(A_1)$ conditioned on $A_2, \ldots, A_t$ is at most $H(A) - 2\varepsilon^2/t^2$, but in particular this means there exist $i_1, \ldots, i_{t-1}$ so that $H(A_t | A_1 = i_1, \ldots, A_{t-1} = i_{t-1}) \leqslant H(A) - 2\varepsilon^2/t^2$. The probability of i under this distribution $A_t | A_1 = i_1, \ldots, A_{t-1} = i_{t-1}$ is proportional to $\mathbb{E}_{x \sim \mathcal{X}} x_i^2 \cdot \prod_{j=1}^{t-1} x_{i_j}^2$, which means that it exactly equals the distribution $A(\mathcal{X}_{i_1,\ldots,i_{t-1}})$. Thus we see that $\Psi(\mathcal{X}_{i_1,\ldots,i_{t-1}}) \leqslant \Psi(\mathcal{X}) - 2\varepsilon^2/t^2$. This concludes the proof of Theorem 3.1. □

Remark 3.5 (Handling odd degrees and non homogenous polynomials)**.** If the polynomial P is not homogenous but only has monomials of even degree, we can homogenize it by multiplying every monomial with an appropriate power of $(\sum x_i^2)$ which is identically equal to 1 on the sphere. To handle odd degree monomials we can introduce a new variable x_0 and set a constraint that it must be identically equal to $1/\sqrt{t}$. (Note that if the pseudoexpectation operator is consistent with this constraint then our rounding algorithm will in fact output a vector that satisfies it.) This way we can represent every odd degree monomial $\alpha_S \prod_{i \in S} x_i$ with the even degree monomial $t\alpha_S x_0 \prod_{i \in S} x_i$. The maximum of P on the unit sphere is equal to the maximum of the new polynomial P' on the sphere of radius $\sqrt{1 + 1/t}$, which, because P' is homogenous, equals $(1 + 1/t)^{t/2}$ times the maximum of P' of the sphere. We simple define the spectral norm of P as the spectral norm of P'.

4 Finding an "analytically sparse" vector in a subspace

In this section we prove Theorem 1.5. We let $\mathcal{U}$ be a universe of size n and $L_2(\mathcal{U})$ be the vector space of real-valued functions $f : \mathcal{U} \to \mathbb{R}$. The measure on the set $\mathcal{U}$ is the uniform probability distribution and hence we will use the inner product $\langle f, g \rangle = \mathbb{E}_\omega f(\omega) g(\omega)$ and norm $\|f\|_p = (\mathbb{E}_\omega f(\omega)^p)^{1/p}$ for $f, g : \mathcal{U} \to \mathbb{R}$ and $p \geqslant 1$.

Theorem 4.1 (Theorem 1.5, restated)**.** *There is a constant $\varepsilon > 0$ and a polynomial-time algorithm A, based on $O(1)$ levels of the SOS hierarchy, that on input a projector operator Π such that there exists a μ-sparse Boolean function f satisfying $\|\Pi f\|_2^2 \geqslant (1 - \varepsilon)\|f\|_2^2$, outputs a function $g \in$ Image(Π) such that*

$$\|g\|_4^4 \geqslant \Omega\left(\frac{\|g\|_2^4}{\mu(\operatorname{rank}\Pi)^{1/3}}\right).$$

We will prove Theorem 4.1 by first showing a combining algorithm and then transforming it into a rounding algorithm. Note that the description of the combining algorithm is independent of the actual relaxation used, since it assumes a true distribution on the solutions, and so we first describe the algorithm before specifying the relaxation. In our actual relaxation we will use some auxiliary variables that will make the analysis of the algorithm simpler.

Combining algorithm for finding an analytically sparse vector:

Input: Distribution $\mathcal{D}$ over Boolean (i.e., 0/1 valued) functions $f \in L_2(\mathcal{U})$ that satisfy:

- $\mu(f) = \mathbb{P}[f(\omega) = 1] = 1/\lambda$.
- $\|\Pi f\|_2^2 \geqslant (1 - \varepsilon)\|f\|_2^2$.

Goal: Output g such that

$$\|g\|_4^4 \geqslant \gamma \|g\|_2^4 \text{ where } \gamma = \Omega(1/\mu(\operatorname{rank}\Pi)^{1/3}) \tag{4.1}$$

Operation: Do the following:

Coordinate projection rounding: For $\omega \in \mathcal{U}$, let $\delta_\omega : \mathcal{U} \to \mathbb{R}$ be the function that satisfies $\langle f, \delta_\omega \rangle = f(\omega)$ for all $f \in L_2(\mathcal{U})$. Go over all vectors of the form $g_\omega = \Pi \delta_\omega$ for $\omega \in \mathcal{U}$ and if there is one that satisfies (4.1) then output it. Note that the output of this procedure is independent of the distribution $\mathcal{D}$.

Random function rounding: Choose a random gaussian vector $t \in L_2(\mathcal{U})$ and output $g = \Pi t$ if it satisfies (4.1). (Note that this is also independent of the distribution $\mathcal{D}$.)

Conditioning: Go over all choices for $\omega_1, \ldots, \omega_4 \in \mathcal{U}$ and modify the distribution $\mathcal{D}$ to the distribution $\mathcal{D}_{\omega_1,\ldots,\omega_4}$ defined such that $\mathbb{P}_{\mathcal{D}_{\omega_1,\ldots,\omega_4}}[f]$ is proportional to $\mathbb{P}_{\mathcal{D}}[f] \prod_{j=1}^4 f(\omega_j)^2$ for every f.

Gaussian rounding: For every one of these choices, let t be a random Gaussian that matches the first two moments of the distribution $\mathcal{D}$, and output $g = \Pi t$ if it satisfies (4.1).

Because we will make use of this fact later, we will note when certain properties hold not just for expectations of actual probability distributions but for *pseudoexpectations* as well. The extension to pseudo-expectations is typically not deep, but can be cumbersome, and so the reader might want to initially restrict attention to the case of a combining algorithm, where we only deal with actual expectations. We show the consequences for each of the steps failing, and then combine them together to get a contradiction.

4.1 Random function rounding

We start by analyzing the random function rounding step. Let $e_1, \ldots, e_n$ be an orthonormal basis for the space of functions $L_2(\mathcal{U})$. Let t be a standard Gaussian function in $L_2(\mathcal{U})$, i.e., $t = \xi_1 e_1 + \ldots + \xi_n e_n$ for independent standard normal variable $\xi_1, \ldots, \xi_n$ (each with mean 0 and variance 1). The following lemmas combined show what are the consequences if $\|\Pi t\|_4$ is not much bigger than $\|\Pi t\|_2$.

Lemma 4.2. *For any $f, g \colon \mathcal{U} \to \mathbb{R}$,*

$$\mathop{\mathbb{E}}_t \langle f, t\rangle\langle g, t\rangle = \langle f, g\rangle \,.$$

Proof. Write f and g in the basis $\{e_1, \ldots, e_n\}$: i.e., $f = \sum_i a_i e_i$ and $g = \sum_j b_j e_j$. Then, because this is an orthonormal basis, $\langle f, g\rangle$ is equal to $\sum_i a_i b_i$ and $\langle f, t\rangle\langle g, t\rangle = \sum_{i,j} a_i b_j \xi_i \xi_j$, which has expectation $\sum_i a_i b_i$. Hence, the left-hand side is the same as the right-hand side. □

Lemma 4.3. *The 4^{th} moment of $\|\Pi t\|_4$ satisfies*

$$\mathop{\mathbb{E}}_t \|\Pi t\|_4^4 \geqslant \mathop{\mathbb{E}}_\omega \|\Pi \delta_\omega\|_2^4 \,.$$

Proof. By the previous lemma, the Gaussian variable $\Pi t(\omega) = \langle \Pi\delta_\omega, t\rangle$ has variance $\|\Pi\delta_\omega\|_2^2$. Therefore,

$$\begin{aligned}\mathop{\mathbb{E}}_t \|\Pi t\|_4^4 &= \mathop{\mathbb{E}}_t \mathop{\mathbb{E}}_\omega \Pi t(\omega)_4^4 = \mathop{\mathbb{E}}_\omega \mathop{\mathbb{E}}_t \langle \delta_\omega, \Pi t\rangle^4 \\ &= 3 \mathop{\mathbb{E}}_\omega \left(\mathop{\mathbb{E}}_t \langle \Pi\delta_\omega, t\rangle^2\right)^4 = 3 \mathop{\mathbb{E}}_\omega \|\Pi\delta_\omega\|_2^4 \,,\end{aligned}$$

since $3 = \mathbb{E}_{X\sim N(0,1)} X^4$. □

Lemma 4.4. *The 4^{th} moment of $\|\Pi t\|_2$ satisfies*

$$\mathop{\mathbb{E}}_t \|\Pi t\|_2^4 \leqslant 10 \cdot (\operatorname{rank}\Pi)^2 \,.$$

Proof. The random variable $\|\Pi t\|_2^2$ has a χ^2-distribution with $k = \operatorname{rank}\Pi$ degrees of freedom. The mean of this distribution is k and the variance is $2k$. It follows that $\mathbb{E}_t \|\Pi t\|_2^4 \leqslant 10(\operatorname{rank}\Pi)^2$. □

4.2 Coordinate projection rounding

We now turn to showing the implications of the failure of projection rounding. We start by noting the following technical lemma, that holds for both the expectation and counting inner products:

Lemma 4.5. *Let x and y be two independent, vector-valued random variables. Then,*

$$\mathbb{E}\langle x, y\rangle^4 \leqslant \left(\mathbb{E}\langle x, x'\rangle^4\right)^{1/2} \cdot \left(\mathbb{E}\langle y, y'\rangle^4\right)^{1/2} .$$

Moreover, this holds even if x, y come from a level $\ell \geqslant 8$ pseudodistribution.

Proof. By Cauchy–Schwarz,

$$\tilde{\mathbb{E}}_{x,y}\langle x, y\rangle^4 = \langle \tilde{\mathbb{E}}_x x^{\otimes 4}, \tilde{\mathbb{E}}_y y^{\otimes 4}\rangle \leqslant \|\tilde{\mathbb{E}}_x x^{\otimes 4}\|_2 \cdot \|\tilde{\mathbb{E}}_y y^{\otimes 4}\|_2 = \left(\tilde{\mathbb{E}}_{x,x'}\langle x, x'\rangle^4\right)^{1/2} \cdot \left(\tilde{\mathbb{E}}_{y,y'}\langle y, y'\rangle^4\right)^{1/2} .$$

We now consider the case of pseudodistributions. In this case the pseudoexpectation over two independent x and x' is obtained using Lemma A.5. Let X and Y be the n^4-dimensional vectors $\tilde{\mathbb{E}}\, x^{\otimes 4}$ and $\tilde{\mathbb{E}}\, y^{\otimes 4}$ respectively.

We can use the standard Cauchy–Schwarz to argue that $X \cdot Y \leqslant \|X\|_2 \cdot \|Y\|_2$, and so what is left is to argue that $\|X\|_2^2 = \tilde{\mathbb{E}}_{x,x'}\langle x, x'\rangle^4$, and similarly for Y. This holds by linearity for the same reason this is true for actual expectations, but for the sake of completeness, we do this calculation. We use the counting inner product for convenience. Because the lemma's statement is scale free, this will imply it also for the expectation norm.

$$\tilde{\mathop{\mathbb{E}}_{x,x'}}\langle x, x'\rangle^4 = \tilde{\mathop{\mathbb{E}}_{x,x'}} \sum_{i,j,k,l} x_i x_j x_k x_l x'_i x'_j x'_k x'_l = \sum_{i,j,k,l} (\tilde{\mathop{\mathbb{E}}_{x}}\, x_i x_j x_k x_l)(\tilde{\mathop{\mathbb{E}}_{x'}}\, x'_i x'_j x'_k x'_l) ,$$

where the last equality holds by independence. But this is simply equal to

$$\sum_{i,j,k,l} (\tilde{\mathop{\mathbb{E}}_{x}}\, x_i x_j x_k x_l)^2 = \|X\|_2^2$$

□

The following lemma shows a nontrivial consequence for $\|\Pi\delta_\omega\|_4^4$ being small:

Lemma 4.6 (Coordinate projection rounding)**.** *For any distribution $\mathcal{D}$ over $L_2(\mathcal{U})$,*

$$\mathop{\mathbb{E}}_{f\sim\mathcal{D}} \|\Pi f\|_4^4 \leqslant \left(\mathbb{E}_{f,f'\sim\mathcal{D}}\langle f, \Pi f'\rangle^4\right)^{1/2} \cdot \left(\mathbb{E}_\omega \|\Pi\delta_\omega\|_4^4\right)^{1/2} .$$

Moreover, this holds even if $\mathcal{D}$ is a level $\ell \geqslant 8$ pseudodistribution. (Note that ω is simply the uniform distribution over $\mathcal{U}$, and hence the last term of the right hand side always denotes an actual expectation.)

Proof. By the previous lemma,

$$\begin{aligned}\tilde{\mathop{\mathbb{E}}_{f\sim\mathcal{D}}} \|\Pi f\|_4^4 = \tilde{\mathop{\mathbb{E}}_{f\sim\mathcal{D}}} \mathop{\mathbb{E}}_{\omega} \langle \delta_\omega, \Pi f\rangle^4 &\leqslant \left(\tilde{\mathbb{E}}_{f,f'\sim\mathcal{D}}\langle f, \Pi f'\rangle^4\right)^{1/2} \cdot \left(\mathbb{E}_{\omega,\omega'}\langle \delta_\omega, \Pi\delta_{\omega'}\rangle^4\right)^{1/2} \\ &= \left(\tilde{\mathbb{E}}_{f,f'\sim\mathcal{D}}\langle f, \Pi f'\rangle^4\right)^{1/2} \cdot \left(\mathbb{E}_\omega \|\Pi\delta_\omega\|_4^4\right)^{1/2} .\end{aligned}$$

□

4.3 Gaussian Rounding

In this subsection we analyze the gaussian rounding step. Let t be a random function with the Gaussian distribution that matches the first two moments of a distribution $\mathcal{D}$ over $L_2(\mathcal{U})$.

Lemma 4.7. *The 4^{th} moment of $\|\Pi t\|_4$ satisfies*

$$\mathop{\mathbb{E}}_{t}\|\Pi t\|_4^4 = 3 \mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}} \left\langle (\Pi f)^2, (\Pi f')^2 \right\rangle .$$

Moreover, this holds even if $\mathcal{D}$ is a level $\ell \geqslant 100$ pseudodistribution. (Note that even in this case t is still an actual distribution.)

Proof.

$$\mathop{\mathbb{E}}_{t}\|\Pi t\|_4^4 \;=\; \mathop{\mathbb{E}}_{t}\mathop{\mathbb{E}}_{\omega}\Pi t(\omega)^4 \;=\; 3\mathop{\mathbb{E}}_{\omega}\left(\mathbb{E}_t\,\Pi t(\omega)^2\right)^2 \;=\; 3\mathop{\tilde{\mathbb{E}}}_{\omega}\left(\mathbb{E}_{f\sim\mathcal{D}}\,\Pi f(\omega)^2\right)^2 \;=\; 3\mathop{\tilde{\mathbb{E}}}_{f,f'\sim\mathcal{D}}\left\langle (\Pi f)^2, (\Pi f')^2\right\rangle . \quad \square$$

Fact 4.8. *If $\{A, B, C, D\}$ have Gaussian distribution, then*

$$\mathbb{E}\,ABCD = \mathbb{E}\,AB\cdot\mathbb{E}\,CD \;+\; \mathbb{E}\,AC\cdot\mathbb{E}\,BD \;+\; \mathbb{E}\,BC\cdot\mathbb{E}\,AD .$$

Lemma 4.9. *The fourth moment of $\|\Pi t\|_2$ satisfies*

$$\mathop{\mathbb{E}}_{t}\|\Pi t\|_2^4 \leqslant 3\left(\mathop{\mathbb{E}}_{f\sim\mathcal{D}}\|\Pi f\|_2^2\right)^2 .$$

Moreover, this holds even if $\mathcal{D}$ is a level $\ell \geqslant 4$ pseudodistribution.

Proof. By the previous fact,

$$\begin{aligned}\mathop{\mathbb{E}}_{t}\|\Pi t\|_2^4 &= \mathop{\mathbb{E}}_{\omega,\omega'}\mathop{\mathbb{E}}_{t}\Pi t(\omega)^2\cdot\Pi t(\omega')^2 \\ &= \mathop{\mathbb{E}}_{\omega,\omega'}\mathop{\tilde{\mathbb{E}}}_{f}\Pi f(\omega)^2\cdot\mathop{\tilde{\mathbb{E}}}_{f}\Pi f(\omega')^2 + 2\left(\mathop{\tilde{\mathbb{E}}}_{f}\Pi f(\omega)\Pi f(\omega')\right)^2 \leqslant 3\left(\mathop{\tilde{\mathbb{E}}}_{f}\|\Pi f\|_2^2\right)^2 . \quad \square\end{aligned}$$

4.4 Conditioning

We now show the sense in which conditioning can make progress. Let $\mathcal{D}$ be a distribution over $L_2(\mathcal{U})$. For $\omega \in \mathcal{U}$, let $\mathcal{D}_\omega$ be the distribution $\mathcal{D}$ reweighed by $f(\omega)^2$ for $f \sim \mathcal{D}$. That is, $\mathbb{P}_{\mathcal{D}_\omega}\{f\} \propto f(\omega)^2\cdot\mathbb{P}_{\mathcal{D}}\{f\}$, or in other words, for every function $P(\cdot)$, $\mathbb{E}_{f\sim\mathcal{D}_\omega}P(f) = (\mathbb{E}_{f\sim\mathcal{D}}f(\omega)^2P(f))/(\mathbb{E}_{f\sim\mathcal{D}}f(\omega)^2)$. Similarly, we write $\mathcal{D}_{\omega_1,\dots,\omega_r}$ for the distribution $\mathcal{D}$ reweighed by $f(\omega_1)^2\cdots f(\omega_r)^2$.

Lemma 4.10 (Conditioning)**.** *For every even $r \in \mathbb{N}$, there are points $\omega_1,\dots,\omega_r \in \mathcal{U}$ such that the reweighed distribution $\mathcal{D}' = \mathcal{D}_{\omega_1,\dots,\omega_r}$ satisfies*

$$\mathop{\mathbb{E}}_{f,g\sim\mathcal{D}'}\left\langle f^2, g^2\right\rangle \geqslant \left(\mathbb{E}_{f,g\sim\mathcal{D}}\left\langle f^2, g^2\right\rangle^r\right)^{1/r}$$

Moreover, this holds even if $\mathcal{D}$ is a level $\ell \geqslant 10r$ pseudodistribution.

Proof. We have that

$$\max_{\omega_1,\ldots,\omega_r} \tilde{\mathbb{E}}_{f,g\sim\mathcal{D}_{\omega_1,\ldots,\omega_r}} \left\langle f^2, g^2 \right\rangle = \max_{\omega_1,\ldots,\omega_r} \left(\frac{\tilde{\mathbb{E}}\, f(\omega_1)^2 \cdots f(\omega_r)^2 \cdot g(\omega_1)^2 \cdots g(\omega_r)^2 \left\langle f^2, g^2 \right\rangle}{\left(\tilde{\mathbb{E}}_f\, f(\omega_1)^2 \cdots f(\omega_r)^2\right)\left(\tilde{\mathbb{E}}_g\, g(\omega_1)^2 \cdots g(\omega_r)^2\right)} \right)$$

but using $\mathbb{E}(X)/\,\mathbb{E}(Y) \leqslant \max(X/Y)$ and $\mathbb{E}_{\omega_1,\ldots,\omega_r} f(\omega_1)^2 \cdots f(\omega_r)^2 g(\omega_1)^2 \cdots g(\omega_r)^2 = \left\langle g^2, f^2 \right\rangle^r$, the RHS is lower bounded by

$$\frac{\mathbb{E}_{\omega_1,\ldots,\omega_r} \tilde{\mathbb{E}}\, f(\omega_1)^2 \cdots f(\omega_r)^2 \cdot g(\omega_1)^2 \cdots g(\omega_r)^2 \left\langle f^2, g^2 \right\rangle}{\mathbb{E}_{\omega_1,\ldots,\omega_r} \left(\tilde{\mathbb{E}}_f\, f(\omega_1)^2 \cdots f(\omega_r)^2\right)\left(\tilde{\mathbb{E}}_g\, g(\omega_1)^2 \cdots g(\omega_r)^2\right)} = \frac{\mathbb{E}_{f,g\sim\mathcal{D}} \left\langle f^2, g^2 \right\rangle^{r+1}}{\tilde{\mathbb{E}}_{f,g\sim\mathcal{D}} \left\langle f^2, g^2 \right\rangle^r}$$

Now, if $\mathcal{D}$ was an actual expectation, then we could use Hölder's inequality to lower bound the numerator of the RHS by $\left(\mathbb{E}_{f,g\sim\mathcal{D}} \left\langle f^2, g^2 \right\rangle^r\right)^{(r+1)/r}$ which would lower bound the RHS by $\left(\mathbb{E}_{f,g\sim\mathcal{D}} \left\langle f^2, g^2 \right\rangle^r\right)^{1/r}$. For pseudoexpectations this follows by appealing to Lemma A.4. □

4.5 Truncating functions

The following observation would be useful for us for analyzing the case that the distribution is over functions that are not completely inside the subspace. Note that if the function f is inside the subspace, we can just take $\overline{f} = f$ in Lemma 4.11, and so the reader may want to skip this section in a first reading and just pretend that $\overline{f} = f$ below.

Lemma 4.11. *Let $\varepsilon < 1/400$, Π be a projector on $\mathbb{R}^{\mathcal{U}}$ and suppose that $f : \mathcal{U} \to \{0,1\}$ satisfies that $\mathbb{P}[f(\omega) = 1] = \mu$ and $\|\Pi f\|_2^2 \geqslant (1-\varepsilon)\mu$. Then there exists a function $\overline{f} : \mathcal{U} \to \mathbb{R}$ such that:*

1. $\|\Pi\overline{f}\|_4^4 \geqslant \Omega(\mu)$.
2. *For every $\omega \in \mathcal{U}$, $\Pi f(\omega)^2 \geqslant \Omega(|\overline{f}(\omega)|)$.*

Proof. Fix $\tau > 0$ to be some sufficiently small constant (e.g., $\tau = 1/2$ will do). Let $f' = \Pi f$. We define $\overline{f} = f' \cdot 1_{|f'|\geqslant\tau}$ (i.e., $\overline{f}(\omega) = f'(\omega)$ if $|f'(\omega)| \geqslant \tau$ and $\overline{f}(\omega) = 0$ otherwise) and define $\underline{f} = f' \cdot 1_{|f'|<\tau}$. Clearly $f'(\omega)^2 \geqslant \tau|\overline{f}(\omega)|$ for every $\omega \in \mathcal{U}$.

Since $\underline{f}(x) \neq 0$ if and only if $f'(x) \in (0,\tau)$, clearly $|\underline{f}(x)| \leqslant |f(x) - f'(x)|$ and hence $\|\underline{f}\|_2^2 \leqslant \varepsilon\mu$. Using $f' = \overline{f} + \underline{f}$, we see that $\Pi\overline{f} = f + (f' - f) - \underline{f} + (\Pi\overline{f} - \overline{f})$. Now since f' is in the subspace, $\|\Pi\overline{f} - \overline{f}\|_2 \leqslant \|f' - \overline{f}\|_2 = \|\underline{f}\|_2$ and hence for $g = (f' - f) - \underline{f} + (\Pi\overline{f} - \overline{f})$, $\|g\|_2 \leqslant 3\sqrt{\varepsilon\mu}$. Therefore the probability that $g(\omega) \geqslant 10\sqrt{\varepsilon}$ is at most $\mu/2$. This means that with probability at least $\mu/2$ it holds that $f(\omega) = 1$ and $g(\omega) \leqslant 10\sqrt{\varepsilon}$, in which case $\overline{f}(\omega) \geqslant 1 - 10\sqrt{\varepsilon} \geqslant 1/2$. In particular, we get that $\mathbb{E}\,\overline{f}(\omega)^4 \geqslant \Omega(\mu)$. □

Remark 4.12 (Non-Boolean functions)**.** The proof of Lemma 4.11 establishes much more than its statement. In particular note that we did not make use of the fact that f is nonnegative, and a function f into $\{0, \pm 1\}$ with $\mathbb{P}[f(\omega) \neq 0] = \mu$ would work just the same. We also did not need the nonzero values to have magnitude exactly one, since the proof would easily extend to the case where they are in $[1/c, c]$ for some constant c. One can also allow some nonzero values of the function to be outside that range, as long as their total contribution to the 2-norm squared is much smaller than μ.

4.6 Putting things together

We now show how the above analysis yields a combining algorithm, and we then discuss the changes needed to extend this argument to pseudodistributions, and hence obtain a rounding algorithm.

Let $\mathcal{D}$ be a distribution over Boolean functions $f\colon \mathcal{U} \to \{0,1\}$ with $\|f\|_2^2 = \mu$ and $\|\Pi f\|_2^2 \geqslant 0.99\|f\|_2^2$. The goal is to compute a function $t\colon \mathcal{U} \to \mathbb{R}$ with $\|\Pi t\|_4 \gg \|t\|_2^2$, given the low-degree moments of $\mathcal{D}$.

Suppose that random-function rounding and coordinate-projection rounding fail to produce a function t with $\|\Pi t\|_4^4 \geqslant \gamma\|t\|_2^4$. Then, $\mathbb{E}_\omega\|\Pi\delta_\omega\|_2^4 \leqslant O(\gamma)\cdot(\operatorname{rank}\Pi)^2$ (from failure of random-function rounding and Lemmas 4.3 and 4.4). By the failure of coordinate-projection rounding (and using Lemma 4.6 applied to the distribution over $\overline{f}$) we get that

$$\left(\mathop{\mathbb{E}}_{f\sim\mathcal{D}}\|\Pi\overline{f}\|_4^4\right)^2 \leqslant O(\gamma)\cdot \mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}}\left\langle\overline{f},\overline{f}'\right\rangle^4\cdot\mathop{\mathbb{E}}_{\omega}\|\Pi\delta_\omega\|_2^4.$$

Combining the two bounds, we get

$$\mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}}\left\langle\overline{f},\overline{f}'\right\rangle^4 \geqslant \Omega(1/(\gamma\operatorname{rank}\Pi)^2)\left(\mathop{\mathbb{E}}_{f\sim\mathcal{D}}\|\Pi\overline{f}\|_4^4\right)^2$$

Since (by Lemma 4.11), $(\Pi f)(\omega)^2 \geqslant \Omega(|\overline{f}(\omega)|)$ for every $\omega\in\mathcal{U}$ and f in the support of $\mathcal{D}$, we have $\langle(\Pi f)^2,(\Pi f')^2\rangle \geqslant \Omega(\langle\overline{f},\overline{f}'\rangle)$ for all f, f' in the support. Thus,

$$\mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}}\left\langle(\Pi f)^2,(\Pi f')^2\right\rangle^4 \geqslant \Omega(1/(\gamma\operatorname{rank}\Pi)^2)\left(\mathop{\mathbb{E}}_{f\sim\mathcal{D}}\|\Pi\overline{f}\|_4^4\right)^2$$

By the reweighing lemma, there exists $\omega_1,\dots,\omega_4\in\mathcal{U}$ such that the reweighted distribution $\mathcal{D}' = \mathcal{D}_{\omega_1,\dots,\omega_4}$ satisfies

$$\mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}'}\left\langle(\Pi f)^2,(\Pi f')^2\right\rangle \geqslant \left(\mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}}\left\langle(\Pi f)^2,(\Pi f')^2\right\rangle^4\right)^{1/4} \geqslant \Omega(1/(\gamma\operatorname{rank}\Pi))^{1/2}\left(\mathop{\mathbb{E}}_{f\sim\mathcal{D}}\|\Pi\overline{f}]\|_4^4\right)^{1/2}$$

The failure of Gaussian rounding (applied to $\mathcal{D}'$) implies

$$\mathop{\mathbb{E}}_{f,f'\sim\mathcal{D}'}\left\langle(\Pi f)^2,(\Pi f')^2\right\rangle \leqslant O(\gamma)\left(\mathop{\mathbb{E}}_{f\sim\mathcal{D}'}\|\Pi f\|_2^2\right)^2.$$

Combining these two bounds, we get

$$\mathop{\mathbb{E}}_{f\sim\mathcal{D}}\|\Pi\overline{f}\|_4^4 \leqslant O(\gamma^3\operatorname{rank}\Pi)\cdot\left(\mathop{\mathbb{E}}_{f\sim\mathcal{D}'}\|\Pi f\|_2^2\right)^4$$

By the properties of $\mathcal{D}$ and Lemma 4.11, the left-hand side is $\Omega(\mu)$ and the right-hand side is $O(\gamma^3\operatorname{rank}\Pi\mu^4)$. Therefore, we get

$$\gamma \geqslant \Omega\left(\frac{1}{(\operatorname{rank}\Pi)^{1/3}\mu}\right)$$

Extending to pseudodistributions. We now consider the case that $\mathcal{D}$ is a pseudodistribution of some large constant level ℓ. (We have not tried to optimize it at all, though $\ell = 100$ should follow easily from the proofs above.) Most of the statements above just go through as is, given that the analysis of all individual steps does extend (as noted) for pseudoexpectations. One issue is that the truncation operation used to obtain $\overline{f}$ is not a low degree polynomial. While it may be possible to approximate it with such a polynomial, we sidestep

the issue by simply adding $\overline{f}$ as additional auxiliary variables to our program, and enforcing the conclusions of Lemma 4.11 as constraints that the pseudoexpectation operator must be consistent with. This is another example of how we design our relaxation to fit the rounding/combining algorithm, rather than the other way around. With this step, we can replace statements such as "(*) holds for all functions in the support of $\mathcal{D}$" (where (*) is some equality or inequality constraint in the variables $f, \overline{f}$) with the statement "$\mathcal{D}$ is consistent with (*)" and thus complete the proof. □

5 Finding planted sparse vectors

As an application of our work, we show how we can find sparse (or analytically sparse) vectors inside a sufficiently generic subspace. In particular, this improves upon a recent result of Demanet and Hand [DH13] who used the L_∞/L_1 optimization procedure of Spielman et al. [SWW12] to show one can recover a μ-sparse vector planted in a random d-dimension subspace $V' \subseteq \mathbb{R}^n$ when $\mu \ll 1/\sqrt{d}$. Our result, combined with the bound on the SDP value of the $2 \to 4$ norm of a random subspace from [BBH+12], implies that if $d = O(\sqrt{n})$ then we can in fact recover such a vector as long as $\mu \ll 1$.

Problem: PlantedRecovery$(\mu, d, |\mathcal{U}|, \varepsilon)$

Input: An arbitrary basis for a linear subspace $V = \operatorname{span}(V' \cup \{f_0\})$, where:

- $V' \subseteq \mathbb{R}^{\mathcal{U}}$ is a **random d-dimensional subspace**, chosen as the span of d vectors drawn independently from the standard Gaussian distribution on $\mathbb{R}^{\mathcal{U}}$, and
- f_0 is an arbitrary **μ-sparse vector**, i.e., $S = \operatorname{supp}(f_0)$ has $|S| \leqslant \mu|\mathcal{U}|$.

Goal: Find a vector $f \in V$ with $\langle f, f_0\rangle^2 \geqslant (1-\varepsilon)\|f\|_2 \|f_0\|_2$.

The goal here should be thought of as recovering f_0 to arbitrarily high precision ("exactly"), and thus the running time of an algorithm should be logarithmic in $1/\varepsilon$. We note that f_0 is not required to be random, and it may be chosen adversarially based on the choice of V'. We will prove the following theorem, which is this section's main result:

Theorem 5.1. *(Theorem 1.4, restated) For some absolute constant $K > 0$, there is an algorithm that solves* PlantedRecovery$(\mu, d, |\mathcal{U}|, \varepsilon)$ *with high probability in time* $\operatorname{poly}(|\mathcal{U}|, \log(1/\varepsilon))$ *for any $\mu < K\mu_0(d)$, where*

$$\mu_0(d) = \begin{cases} 1 & \text{if } d \leqslant \sqrt{\mathcal{U}}, \text{ and} \\ n/d^2 & \text{if } d \geqslant \sqrt{\mathcal{U}}. \end{cases}$$

Our algorithm will work in two stages. It will first solve a constant-degree sum-of-squares relaxation to find a somewhat noisy approximate solution. It will then solve an auxiliary linear program that converts any sufficiently good approximate solution into an exact one.

The first stage is based on the following theorem (proven in Section 5.1), which shows that we can approximately recover a vector when it is planted in a subspace consisting of vectors with substantially smaller L_4/L_2 ratio, provided that we can certify this property of the subspace using a low-degree sum-of-squares proof. To avoid unnecessary notation, we will use a degree 4 certificate in the statement and proof of the theorem; the proof goes through in greater generality, but this suffices for our application.

Theorem 5.2. *Let $V = \operatorname{span}(V' \cup \{f_0\})$, where $f_0 \in \mathbb{R}^{\mathcal{U}}$ is a vector with $\|f_0\|_4/\|f_0\|_2 \geqslant C$, and $V' \subseteq \mathbb{R}^{\mathcal{U}}$ is a linear subspace with*

$$\max_{0 \neq f \in V'} \frac{\|f\|_4}{\|f\|_2} \leqslant c. \tag{5.1}$$

Furthermore, assume that (5.1) *has a degree 4 sum-of-squares proof, i.e., that*

$$\|\Pi_{V'} f\|_4^4 = c^4 \|\Pi_{V'} f\|_2^4 - S, \tag{5.2}$$

where $\Pi_{V'}$ *is the orthogonal projection onto* V'*, and* S *is a degree 4 sum of squares.*

There is a polynomial-time algorithm based on a constant-degree sum-of-squares relaxation that returns a vector $f \in V$ *with* $\langle f, f_0 \rangle \geqslant \left(1 - (c/C)^{\Omega(1)}\right) \|f_0\|_2 \|f\|_2$.

If V' is a random subspace of dimension d, [BBH+12, Theorem 7.1] showed that (5.1) has a degree 4 sum-of-squares proof with high probability for $c = O(1)$ when $d \leqslant \sqrt{|\mathcal{U}|}$, and for $c = O\left(d^{1/2}/|\mathcal{U}|^{1/4}\right)$ when $d \geqslant \sqrt{|\mathcal{U}|}$.[13] We can concisely write these two cases together in our present notation as $c = O\left(\mu_0(d)^{-1/4}\right)$.

Since f_0 is μ-sparse, we know that $\|f_0\|_4 \geqslant \mu^{-1/4} \|f_0\|_2$, so we can take $C = \mu^{-1/4}$. We can thus solve a constant-degree sum-of-squares program to obtain a vector f with $\langle f, f_0 \rangle = (1 - O(1)) \|f_0\|_2 \|f\|_2$ whenever $c \ll O(\mu^{-1/4})$, i.e., when

$$\mu \ll O\left(\frac{1}{c^4}\right) = O\left(\mu_0(d)\right). \tag{5.3}$$

For the second stage, we will consider the following linear program, which can be thought of as searching for a sparse vector in V with a large inner product with f:

$$\arg\min_{y \in V} \|y\|_1 \text{ such that } \langle y, f \rangle = 1. \tag{5.4}$$

In Section 5.2, we will prove the following theorem, which provides conditions under which the linear program will exactly recover f_0 from any f that is reasonably correlated to it:

Theorem 5.3. *Let* $V = \operatorname{span}(V' \cup \{f_0\})$*, and suppose that the following conditions hold:*

- $\operatorname{supp}(f_0) = S$, $|S| = \mu n$ [f_0 is a μ-sparse vector]
- $\|V'\|_{2:1} \leqslant \alpha$ *where* $\|V'\|_{2:1} = \max \|f'\|_2 / \|f'\|_1$ *for all* $0 \neq f' \in V'$) [V' doesn't contain any $1/\alpha^2$ L_2/L_1-sparse vectors]
- $\langle f_0, f \rangle \geqslant (1 - \varepsilon) \|f_0\|_2 \|f\|_2$ [f is correlated with f_0]
- $\langle f', f \rangle \leqslant \eta \|f'\|_2 \|f\|_2$ *for all* $f' \in V'$ [f is not very correlated with anything in V'].

If

$$\frac{\eta}{1 - \varepsilon} < \frac{1}{\alpha \sqrt{\mu}} - 2,$$

then $f_0 / \langle f_0, f \rangle$ *is the unique optimal solution to* (5.4).

Remark 5.4. Because we believe the result might be useful elsewhere, we state the theorem in much more generality than needed for our application. In particular in our application we only need the trivial bound $\eta \leqslant 1$. Also, a bound on $\|V'\|_{2:1}$ can also be derived using the relations between the 4 norm and the 2 norm on vectors in V'.

To prove Theorem 5.1, we take f to be the vector with $\langle f', f \rangle \geqslant \left(1 - (c/C)^{\Omega(1)}\right) \|f_0\|_2 \|f\|_2$ given by Theorem 5.2 and solve the linear program from Equation (5.4). The theorem is vacuous for $d > \sqrt{K} n$, so we may assume that d is less than any fixed constant times n. In this case, the following classic result on almost-spherical sections of the ℓ_1 ball ([Kas77, FLM77], as stated in [DH13]) guarantees that $\|V'\|_{2:1} \leqslant O(1)$:

[13] Their proof actually directly shows that the polynomial P in the RHS of (5.2) has $\|P\|_{\text{spectral}} \leqslant c^4$, which corresponds to a degree 4 SOS proof via Lemma A.12.

Lemma 5.5. *Fix $\delta \in (0,1)$, let $d \leqslant \delta n$, and let $W \subseteq \mathbb{R}^{\mathcal{U}}$ be a random d-dimensional subspace given by the span of d independent standard Gaussians. There exists a constant $C_\delta > 0$ and absolute constants $\gamma_1, \gamma_2 > 0$ such that*

$$C_\delta \left\|f'\right\|_2^2 \leqslant \left\|f'\right\|_1^2 \leqslant \left\|f'\right\|_2^2$$

for all $f' \in W$ with probability $1 - \gamma_1 e^{-\gamma_2 n}$.

By Cauchy-Schwarz, we have $\langle f', f\rangle \leqslant \|f'\|_2 \|f\|_2$, so Theorem 5.3 implies that we recover f_0 exactly as long as[14]

$$\frac{1}{1 - O(c/C)} < \sqrt{\frac{C_\delta}{\mu}} - 2. \tag{5.5}$$

C_δ is a constant for any fixed δ, so, by taking K sufficiently small in the statement of the theorem, we may assume that $\mu < C_\delta/16$, and thus that the right-hand side of (5.5) is at least 2. In this case, we can recover f_0 as long as $c \ll O(C)$. Combining this with Equation (5.3) and choosing K appropriately thus completes the proof of Theorem 5.1. □

It thus suffices to prove Theorems 5.2 and 5.3, which we will do in sections 5.1 and 5.2, respectively. We note that Theorems 5.2 and 5.3 hold for any V' that meets certain norm requirements, and they do not require V' to be a uniformly random subspace. As such, the results of this section hold in a broader context. (For example, they immediately generalize to other distributions of subspaces that meet the norm bounds.) We hope that the technical results of this section will find other uses, so we have stated them in a somewhat general way to facilitate their application in other settings.

5.1 Recovering f_0 approximately (Proof of Theorem 5.2)

In this section, we prove Theorem 5.2, which allows us to recover a vector that is reasonably well-correlated with f_0. The basic idea is that f_0 has a much larger L_4/L_2 ratio than anything in V', so maximizing the L_4/L_2 ratio should give a vector near f_0.

The key ingredient of the theorem is the following lemma about (pseudo-)distributions supported on L_4/L_2-sparse functions in V'. Note that this lemma does not need the space to be random, but only that it can be certified to have no L_4/L_2 sparse vectors by the SOS SDP.

Lemma 5.6. *Let $V' \subseteq \mathbb{R}^{\mathcal{U}}$ be a linear subspace such that*

$$\max_{0 \neq f \in V'} \frac{\|f\|_4}{\|f\|_2} \leqslant c \tag{5.6}$$

Let f_0 be a unit function in $V'^{\perp}$ with $\|f_0\|_4 = C > 100c$, and let $\mathcal{X}$ be a distribution over $\mathbb{R}^{\mathcal{U}}$ over unit functions $f \in \mathrm{Span} V' \cup \{f_0\}$ satisfying $\|f\|_4 \geqslant C$. Then

$$\mathbb{E}\langle x, f_0\rangle^2 \geqslant 1 - O(c/C)\,.$$

Moreover this holds even if $\mathcal{X}$ is a pseudodistribution of level $\ell \geqslant 8$, as long as (5.6) has a degree 4 sum-of-squares proof.

We can obtain a pseudodistribution $\mathcal{X}$ meeting the requirements in Lemma 5.6 by solving a degree 8 sum-of-squares program that maximizes $\|f\|_4^4$ over $f \in V$ with $\|f\|_2^2 = 1$. If we sample a random Gaussian consistent with the first two moments of $\mathcal{X}$, then we will obtain a vector g whose expected 2-norm squared is 1 and whose expected inner product with f_0 is $(1 - o(1))\|f_0\|$, so Lemma 5.6 therefore implies Theorem 5.2.

[14] We note that the last two bounds were somewhat weak: Lemma 5.5 holds for subspaces of linear dimension, but we only applied it to a subspace with $d \leqslant \sqrt{|\mathcal{U}|}$; and the application of Cauchy-Schwarz could have been tightened using a better analysis. However, these were sufficient to prove Theorem 5.1.

Proof of Lemma 5.6. Write every vector f in the support of $\mathcal{X}$ in the form $f = \alpha f_0 + f'$ where $f' \in V'$ and $\alpha = \langle f, f_0 \rangle$. We know that

$$C = \|f\|_4 \leqslant \alpha \|f_0\|_4 + \|f'\|_4 \leqslant \alpha C + c\|f'\|_2 \leqslant \alpha C + c, \tag{5.7}$$

so

$$\alpha \geqslant 1 - c/C\,.$$

This concludes the proof for actual expectations. To argue about pseudoexpectations, we need to use only constraints involving polynomials, and therefore we use

$$C^4 = \|f\|_4^4 = \tilde{\mathbb{E}}_f \mathop{\mathbb{E}}_\omega (\alpha f_0(\omega) + f'(\omega))^4$$

which equals

$$\tilde{\mathbb{E}}_f\, \alpha^4 \|f_0\|_4^4 + 4\,\tilde{\mathbb{E}}\, \alpha^3 \langle f_0^3, f' \rangle + 6\,\tilde{\mathbb{E}}\, \alpha^2 \langle f_0^2, f'^2 \rangle + 4\,\tilde{\mathbb{E}}\, \alpha \langle f_0, f'^3 \rangle + \tilde{\mathbb{E}} \|f'\|_4^4\,.$$

The existence of a degree 4 sum-of-squares proof of (5.6) implies that $\mathcal{X}$ must be consistent with the constraint $\|f'\|_4^4 \leqslant c^4$. We can thus use Cauchy–Schwarz and Hölder's inequality (Lemma A.10 and Corollary A.11), to bound all of the terms except the first one by a constant times $|\alpha|^3 C^3 c$, and so we get

$$C^4 \leqslant \tilde{\mathbb{E}}\, \alpha^4 C^4 + 15|\alpha|^3 C^3 c.$$

Using the fact that the expectation is consistent with the constraint $|\alpha| \leqslant 1$, we obtain

$$\tilde{\mathbb{E}}\, \alpha^4 \geqslant 1 - 15c/C\,.$$

Since we satisfy $|\alpha| \leqslant 1$, we know that $\tilde{\mathbb{E}}\, \alpha^6 \leqslant 1$, so we can apply Cauchy-Schwarz to show that

$$\tilde{\mathbb{E}}\, \alpha^4 \leqslant \sqrt{\tilde{\mathbb{E}}\, \alpha^2}\, \sqrt{\tilde{\mathbb{E}}\, \alpha^6} \leqslant \sqrt{\tilde{\mathbb{E}}\, \alpha^2}\,,$$

which allows us to conclude that

$$\tilde{\mathbb{E}}\, \alpha^2 \geqslant 1 - 30c/C. \qquad \square$$

5.2 Recovering f_0 exactly (Proof of Theorem 5.3)

In this section, we prove Theorem 5.3, which allows us to use a vector near f_0 to recover f_0 exactly (up to the precision used when solving the linear program). Intuitively, this is relies on the same tendency towards sparsity of vectors with minimal 1-norm that underlies the earlier works that are based on L_∞/L_1-sparsity. Minimizing the L_∞/L_1-sparsity amounts to solving the linear program in (5.4) with y equal to each of the unit basis vectors, and then taking the best of the $|\mathcal{U}|$ solutions. When f_0 is sparse enough, it will have at least one fairly large coefficient, and f_0 will then be sufficiently correlated with the corresponding unit basis vector for the linear program to find it. This breaks down when $\mu = \Omega(1/\sqrt{d})$, at which point any one basis vector is expected to be more correlated with some vector in V' than it is with f_0. Here, instead of using the unit basis vectors, we use a vector y that shares many coordinates with f_0, which then lets us handle a much broader range of μ.

Proof of Theorem 5.3. To analyze the optimum of (5.4), we decompose $y \in V$ as $y = tf_0 + f'$ for $t \in \mathbb{R}$ and $f' \in V'$. We will show that

$$\frac{\|f_0\|_1}{\langle f_0, f \rangle} \leqslant \frac{\|y\|_1}{\langle y, f \rangle} = \frac{\|tf_0 + f'\|_1}{t\langle f_0, f \rangle + \langle f', f \rangle} \tag{5.8}$$

for all $y \in V$, with equality only if $f' = 0$, which immediately implies Theorem 5.3.

Let f'_S and $f'_{\overline{S}}$ be the vectors obtained from f' by zeroing out the coordinates outside S and $\overline{S}$, respectively, so that $f' = f'_S + f'_{\overline{S}}$. Since f_0 is zero outside of S, we have

$$\left\|tf_0 + f'\right\|_1 = \left\|tf_0 + f'_S\right\|_1 + \left\|f'_{\overline{S}}\right\|_1 \geqslant t\,\|f_0\|_1 - \left\|f'_S\right\|_1 + \left\|f'_{\overline{S}}\right\|_1 . \tag{5.9}$$

Equation (5.9) and the inequality

$$\frac{A+B}{C+D} \geqslant \min\left\{\frac{A}{B}, \frac{C}{D}\right\}$$

give

$$\frac{\|tf_0 + f'\|_1}{t\langle f_0, f\rangle + \langle f', f\rangle} \geqslant \frac{t\,\|f_0\|_1 - \left\|f'_S\right\|_1 + \left\|f'_{\overline{S}}\right\|_1}{t\langle f_0, f\rangle + \langle f', f\rangle} \geqslant \min\left\{\frac{\|f_0\|_1}{\langle f_0, f\rangle}, \frac{\left\|f'_{\overline{S}}\right\|_1 - \left\|f'_S\right\|_1}{\langle f', f\rangle}\right\}, \tag{5.10}$$

where the second inequality in (5.10) is strict unless the two terms inside the min are equal. To prove the inequality asserted in (5.8), and thus Theorem 5.3, it therefore suffices to show that

$$\frac{\|f_0\|_1}{\langle f_0, f\rangle} < \frac{\left\|f'_{\overline{S}}\right\|_1 - \left\|f'_S\right\|_1}{\langle f', f\rangle} \tag{5.11}$$

for all $0 \neq f' \in V'$.

We can bound the left-hand side of (5.11) using the assumptions that $\langle f_0, f\rangle \geqslant (1-\varepsilon)\,\|f_0\|_2\,\|f\|_2$ and that f_0 is μ-sparse:

$$\frac{\|f_0\|_1}{\langle f_0, f\rangle} \leqslant \frac{\|f_0\|_1}{(1-\varepsilon)\,\|f_0\|_2\,\|f\|_2} \leqslant \frac{\sqrt{\mu}\,\|f_0\|_2}{(1-\varepsilon)\,\|f_0\|_2\,\|f\|_2} = \frac{\sqrt{\mu}}{(1-\varepsilon)\,\|f\|_2}.$$

To bound the numerator of the right-hand side of (5.11), we need to show that f' cannot have too large a fraction of its 1-norm concentrated in the coordinates in S. We first note that, if this occurred, it would lead to a large contribution to the 2-norm:

$$\left\|f'_S\right\|_1 = \mu \mathop{\mathbb{E}}_{i\in S} |f'(i)| \leqslant \mu\sqrt{\mathop{\mathbb{E}}_{i\in S} f'(i)^2} \leqslant \mu\sqrt{\frac{1}{\mu}\mathop{\mathbb{E}}_{i} f'(i)^2} = \sqrt{\mu}\|f'\|_2.$$

Combining this with our assumption that $\|V'\|_{2:1} \leqslant \alpha$, gives

$$\left\|f'_{\overline{S}}\right\|_1 - \left\|f'_S\right\|_1 = \left\|f'\right\|_1 - 2\left\|f'_S\right\|_1 \geqslant \alpha^{-1}\left\|f'\right\|_2 - 2\sqrt{\mu}\left\|f'\right\|_2 = \left(\alpha^{-1} - 2\sqrt{\mu}\right)\left\|f'\right\|_2,$$

and thus

$$\frac{\left\|f'_{\overline{S}}\right\|_1 - \left\|f'_S\right\|_1}{\langle f', f\rangle} \geqslant \frac{\left(\alpha^{-1} - 2\sqrt{\mu}\right)\|f'\|_2}{\eta\,\|f'\|_2\,\|f\|_2} = \frac{\left(\alpha^{-1} - 2\sqrt{\mu}\right)}{\eta\,\|f\|_2}.$$

If $\frac{\eta}{1-\varepsilon} < (\alpha\sqrt{\mu})^{-1} - 2$, this implies that

$$\frac{\left\|f'_{\overline{S}}\right\|_1 - \left\|f'_S\right\|_1}{\langle f', f\rangle} > \frac{\left(\alpha^{-1} - 2\sqrt{\mu}\right)}{(1-\varepsilon)\left((\alpha\sqrt{\mu})^{-1} - 2\right)\|f\|_2} = \frac{\sqrt{\mu}}{(1-\varepsilon)\,\|f\|_2} \geqslant \frac{\|f_0\|_1}{\langle f_0, f\rangle},$$

from which our desired result follows. □

6 Results for Small Set Expansion

As stated in Corollaries 1.3 and 1.6, our results imply two consequences for the Small Set Expansion problem of [RS10]. This is the problem of deciding, given an input graph G and parameters δ, ε, whether there is a measure-δ subset S of G's vertices where all but an ε fraction of S's edges stay inside it, or that G is a *small set expander* in the sense that every sufficiently small set has almost all its edges leaving it. Beyond being a natural problem in its own right, Small Set Expansion is also closely related to the Unique Games problem whose conjectured hardness is known as Khot's "Unique Games Conjecture" [Kho02]. [RS10] gave a reduction from Small Set Expansion to Unique Games. While a reduction in the other direction is not known, all currently known algorithmic and integrality gap results apply to both problems equally well (e.g., [ABS10, RST10, BGH+12, BBH+12]), and thus they are likely to be computationally equivalent.

We give an algorithm to solve Small Set Expansion in quasipolynomial time on an interesting family of Cayley graphs, and a new polynomial-time approximation algorithm for this problem on general graphs, with the approximation guarantee depending on the dimension of the input graph's top eigenspace.

6.1 Small-set expansion of Cayley graphs

We consider the problem of solving the small set expansion problem on Cayley graphs over $\mathbb{F}_2^\ell$. One reason to consider such graphs is that, until recently, the hardest looking instances for this problem were graphs of this type (i.e., the noisy hypercube [KV05] and the "short code" graph [BGH+12]). [BBH+12] showed that these instances can in fact be solved via constant rounds of the SOS hierarchy, but we still do not have any other good candidate hard instances, and so it is natural to ask whether Cayley graphs can provide such candidates. Also, since the SOS algorithm does not make use of the algebraic structure of Cayley graphs, it is plausible that if this algorithm can efficiently solve the Small-Set Expansion problem on Cayley graphs, then it can in fact solve it on all graphs.

Let G be a Cayley graph on $\mathbb{F}_2^\ell$ with $n = 2^\ell$ vertices. Let $V_{\geqslant\lambda}$ be the linear subspace spanned by the eigenfunctions of G with eigenvalue at least λ. (We identify G here with its random-walk matrix.) Let P_λ be the degree-4 polynomial $P_\lambda(f) = \|\Pi_{\geqslant\lambda} f\|_4^4$, where $\Pi_{\geqslant\lambda}$ is the projector into $V_{\geqslant\lambda}$. We define $K_\lambda(G) = \|P_\lambda\|_{\text{spectral}}$. In this section, we describe approximation algorithms with running times that depend on $K_\lambda(G)$. The algorithms run in quasipolynomial time if $K_\lambda(G)$ is polylogarithmic. We will show interesting families of graphs with $K_\lambda(G) = O(1)$. (See Theorem 6.3.)

The following theorem shows that low-degree sum-of-squares relaxations can detect L_4/L_2-sparse functions in the subspaces $V_{\geqslant\lambda}$ (in the case when $K_\lambda(G)$ is not too large). This result follows from Theorem 3.1 and the fact that the polynomial P_λ has nonnegative coefficients in an appropriate basis.

Theorem 6.1. *Sum-of-squares relaxations of degree $\varepsilon^{-O(1)} K_\lambda(G)^{O(1)} \log n$ provide an additive ε-approximation to the maximum of $\|f\|_4/\|f\|_2$ over all non-zero functions $f \in V_{\geqslant\lambda}$.*

Proof. The problem of maximizing $\|f\|_4/\|f\|_2$ over the subspace $V_{\geqslant\lambda}$ is equivalent to maximizing the polynomial P_λ over functions with norm 1. (Also notice that $\|f\|_4 \geqslant \|f\|_2$ for every function f.) In order to apply Theorem 3.1, we need to verify that P_λ has nonnegative coefficients in an appropriate basis. Since G is a Cayley graph over $\mathbb{F}_2^\ell$, we can take the characters $\{\chi_\alpha\}_{\alpha\in\mathbb{F}_2^\ell}$ as an eigenbasis. (Here, $\chi_\alpha(x) = (-1)^{\sum_i \alpha_i x_i}$.) If we represent $f = \sum_\alpha \hat{f}_\alpha \chi_\alpha$ in this eigenbasis and let $S_{\geqslant\lambda} = \{\alpha \mid \lambda_\alpha \geqslant \lambda\}$ be the indices of the eigenfunctions with eigenvalue at least λ, then

$$P_\lambda(f) = \|\Pi_{\geqslant\lambda} f\|_4^4 = \mathbb{E}\Big(\sum_{\alpha\in S_{\geqslant\lambda}} \hat{f}_\alpha \chi_\alpha\Big)^4 = \sum_{\alpha,\beta,\alpha',\beta'\in S_{\geqslant\lambda}} \hat{f}_\alpha \hat{f}_\beta \hat{f}_{\alpha'} \hat{f}_{\beta'}\, \mathbb{E}\,\chi_\alpha\chi_\beta\chi_{\alpha'}\chi_{\beta'} = \sum_{\substack{\alpha,\beta,\alpha',\beta'\in S_{\geqslant\lambda} \\ \alpha+\beta=\alpha'+\beta'}} \hat{f}_\alpha \hat{f}_\beta \hat{f}_{\alpha'} \hat{f}_{\beta'}\,.$$

It follows that P_λ has nonnegative coefficients in the monomial basis corresponding to the eigenfunctions of G. By Theorem 3.1, sum-of-squares relaxations of degree $\varepsilon^{-O(1)}\|P_\lambda\|_{\text{spectral}}^{O(1)} \log n$ provides an additive approximation to the maximum of P_λ over functions with $\|f\|^2 = \sum_\alpha \hat{f}_\alpha^2 = 1$. □

Using the characterization of small-set expansion in terms of L_4/L_2-sparse functions [BBH+12], Theorem 6.1 implies the following approximation algorithm for small-set expansion on Cayley graphs. This theorem implies Corollary 1.3.

Theorem 6.2. *For some absolute constant $C \geqslant 1$ and all $\mu, \varepsilon > 0$ small enough, sum-of-squares relaxations of degree $K_\lambda(G)^{O(1)} \log n$ can distinguish between the following two cases with $\lambda = 1 - C\varepsilon$.*

Yes: *The Cayley graph G contains a vertex set of measure at most μ and expansion at most ε.*

No: *All vertex sets of measure at most $C\sqrt{\mu}$ in G have expansion at least $1 - 1/C$.*

Proof. We will show that the maximum of $\|f\|_4/\|f\|_2$ over $f \in V_{\geqslant\lambda}$ distinguishes the two cases (by a constant margin). Therefore, Theorem 6.1 implies that we can distinguish between the cases using sum-of-squares relaxations.

Yes-case: Let f be the indicator function of a set with measure at most μ and expansion at most ε. Then, $\|\Pi_{\geqslant\lambda} f\|^2 \geqslant 0.99\|f\|^2$. It follows that $\|\Pi_{\geqslant\lambda} f\|_4^4 \geqslant 0.9\|f\|_4^4$. (See Lemma 4.11.) Therefore, $\|\Pi_{\geqslant\lambda} f\|_4^4/\|\Pi_{\geqslant\lambda} f\|_2^4 \geqslant \Omega(1)\|f\|_4^4/\|f\|_2^4 = \Omega(1) \cdot 1/\mu$.

No-case: Let $\mu' = C/\sqrt{\mu}$. By [BBH+12, Theorem 2.4], graphs with this kind of small-set expansion satisfy $\|f\|_4^4/\|f\|_2^4 \leqslant O(1)/(\mu')^2 \ll 1/\mu$ for all functions $f \in V_{\geqslant\lambda}(G)$. □

The following theorem shows that there are interesting Cayley graphs that satisfy $K_\lambda(G) = O(1)$ for $\lambda = \Omega(1)$. We consider constructions based on the long code and the short code [BGH+12]. These constructions are parameterized by the size of the graph and its eigenvalue gap. In the context of the Unique Games Conjecture and the Small-Set Expansion Hypothesis, the most relevant case is that the eigenvalue gap is a constant. (The eigenvalue gap corresponds to the gap to perfect completeness.)

Theorem 6.3. *Long-code and short-code based graphs with constant eigenvalue gap satisfy $K_\lambda(G) = O(1)$ for all $\lambda = \Omega(1)$.*

Proof. By [BBH+12, Lemma 5.1], there exists a constant C such that $P_\lambda(f) = C\|f\|_2^4 - S(f)$ where $S(\cdot)$ is a sum of squares (the same constant C works for both graph constructions). Therefore, by Lemma A.12, $\|P_\lambda\|_{\text{spectral}} \leqslant C$. □

6.2 Approximating small-set expansion using ASVP

The approximation algorithm for the analytical sparse vector problem (Theorem 4.1) implies the following approximation algorithm for small-set expansion. An algorithm for the same problem with the factor $(\dim V_{\geqslant\lambda})^{1/3}$ replaced by a constant would refute the Small-Set Expansion Hypothesis [RS10, RST12].[15]

Theorem 6.4. *For some absolute constant $C \geqslant 1$ and all $\mu, \varepsilon > 0$ small enough, sum-of-squares relaxations with constant degree can solve the promise problem on regular graphs G:*

Yes: *The graph contains a vertex set with measure at most $\mu/(\dim V_{\geqslant\lambda})^{1/3}$ and expansion at most ε, where $\lambda = 1 - C \cdot \varepsilon$.*

No: *All vertex sets of measure at most $C\sqrt{\mu}$ have expansion at least $1 - 1/C$.*

[15] It's plausible that, under standard complexity assumptions such as **NP** $\not\subseteq$ **SUBEXP**, even a smaller improvement to a $(\dim V_{\geqslant\lambda})^{o(1)}$ factor instead of $(\dim V_{\geqslant\lambda})^{1/3}$ would refute this hypothesis, though we have no proof of such an implication.

Proof. Suppose G satisfies the **Yes** property. Let f be the indicator functions of a set with measure at most $\mu' = \mu/(\dim V_{\geqslant\lambda})^{1/3}$ and expansion at most ε. Then, $\|\Pi_{\geqslant\lambda} f\|_2^2 \geqslant (1 - 1/C')\|f\|_2^2$, where we can make C' as large as we like by making C larger. By Theorem 4.1, constant-degree sum-of-squares relaxations allow us to find an L_4/L_2-sparse function $g \in V_{\geqslant\lambda}$, so that $\|g\|_4^4 \geqslant \Omega(1/\mu)\|g\|_2^4$. By [BBH+12, Theorem 2.4] (see also Appendix D), such a function certifies that we are not in the **No** case. □

7 Discussion and open questions

A general open question is to find other applications of our approach for rounding sum-of-squares relaxations. Natural candidates would be problems where it seems that they do not display a "dichotomy" behavior, where beating some simple algorithm is likely to be exponentially hard, but rather suggest more of a smooth tradeoff between time and performance. As far as we are aware, all known "robust"[16] lower-bound results for the sum-of-squares method are *non-constructive*, i.e., they show that hard instances for the sos method exist but do not give an efficient way of constructing them. More concretely, the results use the *probabilistic method* and show that with high probability, random instances are hard for sum-of-squares relaxations [Gri01b, Sch08]. Therefore, SOS seems promising for problems where random instances do not seem to be the most difficult, e.g., problems related to the Unique Games Conjecture. A concrete problem of that type to look at is Sparsest Cut. In particular, can we obtain even a small improvement[17] to [ARV04]'s algorithm using more SOS levels? In fact, we believe that even finding a natural reinterpretation of [ARV04] result in our framework would be interesting. That said, our result for finding a planted sparse vector shows that SOS can be useful for average-case problems as well, and in particular we believe SOS might be a strong tool for solving unsupervised learning problems, especially for nonlinear models.

A relaxation-based approximation algorithm can be thought of as having three components: the relaxation, the rounding algorithm, and its analysis. In our approach there is almost no creativity in choosing the relaxation, which is simply taken to be a sufficiently high level of the SOS hierarchy. (Though there may be some flexibility in how we represent solutions.) Can we similarly show a "universal" rounding algorithm, thus pushing all the creative choices into the *analysis*? A related question is whether one can formulate a theorem giving a translation from combining algorithms into rounding algorithms under sufficiently general conditions, so that results like ours would follow as special cases, and as mentioned in Section 1.4, have already made some progress in this direction.

The notion of "analytically sparse" vectors seems potentially useful for more applications. It would be interesting to explore the different choices for L_q/L_p sparsity, and what tradeoffs they yield in terms of computation time versus usefulness as a proxy for actual sparsity. In particular, for the planted sparse vector question, it is natural to conjecture that there is an analytical relaxation that we can optimize over in $n^{O(\ell)}$ time, and can detect sparse vectors in random subspaces of dimension $n^{1-1/\ell}$.

In the context of the Small-Set-Expansion Hypothesis / Unique Games Conjecture, the most important question is whether our results of Section 4 can be further improved. We do not know of any candidate hard instances for this problem (in the relevant range of parameters) and so conjecture that our algorithm (or at least our analysis of it) is not optimal and can be improved further.

Related to the question of finding hard instances, our work suggests a different type of negative results for convex relaxations. While integrality gaps are instances that are hard for a particular relaxation, regardless of the rounding algorithm, one can consider the notion of "combining gaps". These will be instances where

[16] For KNAPSACK-like problems, there exist explicit lower bounds [Gri01a], but here low-degree sum-of-squares proofs provide very good approximation (in this sense, the lower bound is not robust).

[17] An approach to obtain constant-factor approximations for SPARSEST CUT in subexponential time is outlined in the dissertation [Ste10a, Chapter 9]. However, this approach also works with weaker hierarchies. An approach tailored to sum-of-squares would be interesting.

there is a distribution of good solutions, but a particular combining algorithm C fails to find one. Hence, viewing C as a rounding algorithm, such a result shows that C will fail regardless of the relaxation used. (Karloff's work [Kar99] on hard instances for the [GW95] hyperplane cut rounding algorithms can be viewed as such an example.) Studying such gaps can shed more light on our approach and computational difficulty in general. In particular, it might be interesting to consider this question for random satisfiable instances of SAT or other constraint satisfaction problems.

Acknowledgements.. We thank Aram Harrow, Alex Wein, and an anonymous STOC referee for pointing out many typos and errors in a previous version of this paper.

References

[AAJ+13] Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon, *Learning sparsely used overcomplete dictionaries via alternating minimization*, arXiv preprint 1310.7991 (2013), http://arxiv.org/abs/1310.7991. 7

[ABLT06] Sanjeev Arora, Béla Bollobás, László Lovász, and Iannis Tourlakis, *Proving integrality gaps without knowing the linear program*, Theory of Computing **2** (2006), no. 1, 19–51, Preliminary version in FOCS '02. 1

[ABS10] Sanjeev Arora, Boaz Barak, and David Steurer, *Subexponential algorithms for unique games and related problems*, FOCS, 2010, pp. 563–572. 1, 6, 28

[ABSS97] Sanjeev Arora, László Babai, Jacques Stern, and Z. Sweedyk, *The hardness of approximate optima in lattices, codes, and systems of linear equations*, J. Comput. Syst. Sci. **54** (1997), no. 2, 317–331. 5

[AG11] Sanjeev Arora and Rong Ge, *New tools for graph coloring*, APPROX-RANDOM, 2011, pp. 1–12. 1

[AGM13] Sanjeev Arora, Rong Ge, and Ankur Moitra, *New algorithms for learning incoherent and overcomplete dictionaries*, arXiv preprint 1308.6723 (2013), http://arxiv.org/abs/1308.6273. 7

[AGS13] Sanjeev Arora, Rong Ge, and Ali Kemal Sinop, *Towards a better approximation for sparsest cut?*, FOCS, 2013. 1

[AKS12] Hyung-Chan An, Robert Kleinberg, and David B. Shmoys, *Improving christofides' algorithm for the s-t path tsp*, STOC, 2012, pp. 875–886. 7

[ARV04] Sanjeev Arora, Satish Rao, and Umesh V. Vazirani, *Expander flows, geometric embeddings and graph partitioning*, STOC, 2004, pp. 222–231. 1, 30

[BBH+12] Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou, *Hypercontractivity, sum-of-squares proofs, and their applications*, STOC, 2012, pp. 307–326. 2, 3, 6, 7, 8, 10, 11, 23, 24, 28, 29, 30, 35, 36, 37, 38, 42, 45

[BCC+10] Aditya Bhaskara, Moses Charikar, Eden Chlamtac, Uriel Feige, and Aravindan Vijayaraghavan, *Detecting high log-densities: an $O(n^{1/4})$ approximation for densest k-subgraph*, STOC, 2010, pp. 201–210. 1, 13

[BCV+12] Aditya Bhaskara, Moses Charikar, Aravindan Vijayaraghavan, Venkatesan Guruswami, and Yuan Zhou, *Polynomial integrality gaps for strong sdp relaxations of densest k-subgraph*, SODA, 2012, pp. 388–405. 1

[BCY11] Fernando G. S. L. Brandão, Matthias Christandl, and Jon Yard, *A quasipolynomial-time algorithm for the quantum separability problem*, STOC, 2011, pp. 343–352. 1, 5, 8, 39

[BGH+12] Boaz Barak, Parikshit Gopalan, Johan Håstad, Raghu Meka, Prasad Raghavendra, and David Steurer, *Making the long code shorter*, FOCS, 2012, pp. 370–379. 5, 28, 29, 45

[BGMT12] Siavosh Benabbas, Konstantinos Georgiou, Avner Magen, and Madhur Tulsiani, *Sdp gaps from pairwise independence*, Theory of Computing **8** (2012), no. 1, 269–289. 1

[BH13] Fernando G. S. L. Brandão and Aram Wettroth Harrow, *Quantum de finetti theorems under local measurements with applications*, STOC, 2013, pp. 861–870. 5, 8, 39

[BJK05] Andrei Bulatov, Peter Jeavons, and Andrei Krokhin, *Classifying the complexity of constraints using finite algebras*, SIAM Journal on Computing **34** (2005), no. 3, 720–742, Preliminary version in ICALP '00. 7

[BKS12] Boaz Barak, Jonathan Kelner, and David Steurer, *Iterative rounding for sum-of-squares relaxations*, Preliminary version of the current work, unpublished., 2012. 5

[BRS11] Boaz Barak, Prasad Raghavendra, and David Steurer, *Rounding semidefinite programming hierarchies via global correlation*, FOCS, 2011, To appear. arXiv:1104.4680v1. 1, 2, 7, 40

[Chl07] Eden Chlamtac, *Approximation algorithms using hierarchies of semidefinite programming relaxations*, FOCS, 2007, pp. 691–701. 1

[CMM09] Moses Charikar, Konstantin Makarychev, and Yury Makarychev, *Integrality gaps for sherali-adams relaxations*, STOC, 2009, pp. 283–292. 1

[CS08] Eden Chlamtac and Gyanit Singh, *Improved approximation guarantees through higher levels of sdp hierarchies*, APPROX-RANDOM, 2008, pp. 49–62. 1

[CT10] Eden Chlamtac and Madhur Tulsiani, *Convex relaxations and integrality gaps*, 2010, Chapter in Handbook on Semidefinite, Cone and Polynomial Optimization. 1

[DH13] L. Demanet and P. Hand, *Recovering the Sparsest Element in a Subspace*, October 2013, Arxiv preprint 1310.1654. 5, 6, 10, 23, 24

[DKL11] Etienne De Klerk and Monique Laurent, *On the lasserre hierarchy of semidefinite programming relaxations of convex polynomial optimization problems*, SIAM Journal on Optimization **21** (2011), no. 3, 824–832. 35

[dlVKM07] Wenceslas Fernandez de la Vega and Claire Kenyon-Mathieu, *Linear programming relaxations of maxcut*, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 2007, pp. 53–61. 1

[DPS04] Andrew C Doherty, Pablo A Parrilo, and Federico M Spedalieri, *Complete family of separability criteria*, Physical Review A **69** (2004), no. 2, 022308. 1

[DW12] Andrew C Doherty and Stephanie Wehner, *Convergence of sdp hierarchies for polynomial optimization on the hypersphere*, arXiv preprint arXiv:1210.5048 (2012). 4

[FLM77] Tadeusz Figiel, Joram Lindenstrauss, and Vitali D Milman, *The dimension of almost spherical sections of convex bodies*, Acta Mathematica **139** (1977), no. 1, 53–94. 24

[GN10] Lee-Ad Gottlieb and Tyler Neylon, *Matrix sparsification and the sparse null space problem*, APPROX-RANDOM, 2010, pp. 205–218. 5

[Gri01a] Dima Grigoriev, *Complexity of positivstellensatz proofs for the knapsack*, computational complexity **10** (2001), no. 2, 139–154. 1, 2, 30

[Gri01b] Dima Grigoriev, *Linear lower bound on degrees of positivstellensatz calculus proofs for the parity*, Theor. Comput. Sci. **259** (2001), no. 1-2, 613–622. 1, 2, 30

[GS11] Venkatesan Guruswami and Ali Kemal Sinop, *Lasserre hierarchy, higher eigenvalues, and approximation schemes for graph partitioning and quadratic integer programming with psd objectives*, FOCS, 2011, pp. 482–491. 1, 2, 7

[GSS11] Shayan Oveis Gharan, Amin Saberi, and Mohit Singh, *A randomized rounding approach to the traveling salesman problem*, FOCS, 2011, pp. 550–559. 7

[GV01] Dima Grigoriev and Nicolai Vorobjov, *Complexity of null-and positivstellensatz proofs*, Annals of Pure and Applied Logic **113** (2001), no. 1, 153–160. 2

[GW95] Michel X. Goemans and David P. Williamson, *Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming*, J. ACM **42** (1995), no. 6, 1115–1145, Preliminary version in STOC '94. 7, 31

[HM13] Aram Wettroth Harrow and Ashley Montanaro, *Testing product states, quantum merlin-arthur games and tensor optimization*, J. ACM **60** (2013), no. 1, 3, Preliminary version in FOCS '10. 4

[Kar99] Howard J. Karloff, *How good is the goemans-williamson max cut algorithm?*, SIAM J. Comput. **29** (1999), no. 1, 336–350, Preliminary version in STOC '96. 31

[Kas77] Boris Sergeevich Kashin, *Diameters of some finite-dimensional sets and classes of smooth functions*, Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya **41** (1977), no. 2, 334–351, [Math. USSR-Izv. 11, 317333 (1978)]. 24

[Kho02] Subhash Khot, *On the power of unique 2-prover 1-round games*, IEEE Conference on Computational Complexity, 2002, p. 25. 3, 28

[KMN11] Anna R. Karlin, Claire Mathieu, and C. Thach Nguyen, *Integrality gaps of linear and semi-definite programming relaxations for knapsack*, IPCO, 2011, pp. 301–314. 1

[KOTZ14] M. Kauers, R. O'Donnell, L.-Y. Tan, and Y. Zhou, *Hypercontractive inequalities via sos, and the frankl-rödl graph*, SODA, 2014. 2, 3

[Kri64] Jean-Louis Krivine, *Anneaux préordonnés*, Journal d'analyse mathématique **12** (1964), no. 1, 307–326. 2

[KV05] Subhash Khot and Nisheeth K. Vishnoi, *The unique games conjecture, integrality gap for cut problems and embeddability of negative type metrics into ℓ_1*, FOCS, 2005, pp. 53–62. 28, 45

[KZ12] Pascal Koiran and Anastasios Zouzias, *Hidden cliques and the certification of the restricted isometry property*, CoRR **abs/1211.0665** (2012). 5

[Las01] Jean B. Lasserre, *Global optimization with polynomials and the problem of moments*, SIAM Journal on Optimization **11** (2001), no. 3, 796–817. 1, 2, 35

[Lau09] Monique Laurent, *Sums of squares, moment matrices and optimization over polynomials*, Emerging applications of algebraic geometry, Springer, 2009, Updated version available on the author's homepage at http://homepages.cwi.nl/~monique/files/moment-ima-update-new.pdf, pp. 157–270. 2

[LS91] László Lovász and Alexander Schrijver, *Cones of matrices and set-functions and 0-1 optimization*, SIAM Journal on Optimization **1** (1991), no. 2, 166–190. 1

[Nes00] Y. Nesterov, *Squared functional systems and optimization problems*, High performance optimization **13** (2000), 405–440. 1, 2

[O'D07] Ryan O'Donnell, *Analysis of boolean functions*, Lecture Notes. Available online at , 2007. 45

[OZ13] Ryan O'Donnell and Yuan Zhou, *Approximability and proof complexity*, SODA, 2013, pp. 1537–1556. 2, 3

[Par00] Pablo A Parrilo, *Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization*, Ph.D. thesis, California Institute of Technology, 2000. 1, 2

[Rag10] Prasad Raghavendra, *Complexity of constraint satisfaction problems: Exact and approximate*, 2010, Talk at the Institute for Advanced Study, video available on http://video.ias.edu/csdm/complexityconstraint. 7

[RS10] Prasad Raghavendra and David Steurer, *Graph expansion and the unique games conjecture*, STOC, 2010, pp. 755–764. 28, 29, 42

[RST10] Prasad Raghavendra, David Steurer, and Prasad Tetali, *Approximations for the isoperimetric and spectral profile of graphs and related parameters*, STOC, 2010, pp. 631–640. 28

[RST12] Prasad Raghavendra, David Steurer, and Madhur Tulsiani, *Reductions between expansion problems*, IEEE Conference on Computational Complexity, 2012, pp. 64–73. 29, 42

[RT12] Prasad Raghavendra and Ning Tan, *Approximating csps with global cardinality constraints using sdp hierarchies*, SODA, 2012, pp. 373–387. 1

[SA90] Hanif D Sherali and Warren P Adams, *A hierarchy of relaxations between the continuous and convex hull representations for zero-one programming problems*, SIAM Journal on Discrete Mathematics **3** (1990), no. 3, 411–430. 1

[Sch08] Grant Schoenebeck, *Linear level Lasserre lower bounds for certain k-CSPs*, FOCS, 2008, pp. 593–602. 1, 2, 30

[Sho87] NZ Shor, *An approach to obtaining global extremums in polynomial mathematical programming problems*, Cybernetics and Systems Analysis **23** (1987), no. 5, 695–700. 1, 2

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

[Ste10a] David Steurer, *On the complexity of unique games and graph expansion*, Tech. Report TR-887-10, Princeton University, 2010, Available at ftp://ftp.cs.princeton.edu/techreports/2010/887.pdf. 30

[Ste10b] ———, *Subexponential algorithms for d-to-1 two-prover games and for certifying almost perfect expansion*, Manuscript, available from the author's website., 2010. 43

[SWW12] Daniel A. Spielman, Huan Wang, and John Wright, *Exact recovery of sparsely-used dictionaries*, Journal of Machine Learning Research - Proceedings Track **23** (2012), 37.1–37.18. 6, 7, 23

[Tul09] Madhur Tulsiani, *Csp gaps and reductions in the lasserre hierarchy*, STOC, 2009, pp. 303–312. 1

[ZP01] Michael Zibulevsky and Barak A Pearlmutter, *Blind source separation by sparse decomposition in a signal dictionary*, Neural computation **13** (2001), no. 4, 863–882. 6

A Pseudoexpectation toolkit

We recall here the definition of pseudoexpectation from [BBH+12] and prove some of its useful properties. Some of these were already proven in [BBH+12] but others are new.

Definition A.1. Let $\tilde{\mathbb{E}}$ be a functional that maps polynomial P over $\mathbb{R}^n$ of degree at most r into a real number which we denote by $\tilde{\mathbb{E}}_x P(x)$ or $\tilde{\mathbb{E}} P$ for short. We say that $\tilde{\mathbb{E}}$ is a *level-r pseudo-expectation functional* (r-p.e.f. for short) if it satisfies:

Linearity For every polynomials P, Q of degree at most r and $\alpha, \beta \in \mathbb{R}$, $\tilde{\mathbb{E}}(\alpha P + \beta Q) = \alpha \tilde{\mathbb{E}} P + \beta \tilde{\mathbb{E}} Q$.

Positivity For every polynomial P of degree at most $r/2$, $\tilde{\mathbb{E}} P^2 \geqslant 0$.

Normalization $\tilde{\mathbb{E}} 1 = 1$ where on the RHS, 1 denotes the degree-0 polynomial that is the constant 1.

The functional $\tilde{\mathbb{E}}$ can be represented by a table of size $n^{O(r)}$ containing the pseudo-expectations of every monomial of degree at most r (or some other linear basis for polynomials of degree at most r). For a linear functional $\tilde{\mathbb{E}}$, the map $P \mapsto \tilde{\mathbb{E}} P^2$ is a quadratic form. Hence, $\tilde{\mathbb{E}}$ satisfies the positivity condition if and only if the corresponding quadratic form is positive semidefinite. It follows that the convex set of level-r pseudo-expectation functionals over $\mathbb{R}^n$ admits an $n^{O(r)}$-time separation oracle, and hence the r-round SoS relaxation can be solved up to accuracy ε in time $(mn \cdot \log(1/\varepsilon))^{O(r)}$.

For every random variable X over $\mathbb{R}^n$, the functional $\tilde{\mathbb{E}} P := \mathbb{E} P(X)$ is a level-r pseudo-expectation functional for every r. As $r \to \infty$, this hierarchy of pseudo-expectations will converge to the expectations of a true random variable [Las01], in general the convergence is not guaranteed to happen in a finite number of steps [DKL11], although for most problems of interest in TCS, n levels would suffice for either exact convergence or sufficiently close approximation.

We now record various useful ways in which pseudoexpectations behave close to actual expectations.

For two polynomials P and Q, we write $P \preceq Q$ if $Q = P + \sum_{i=1}^m R_i^2$ for some polynomials $R_1, \ldots, R_m$.

If P and Q have degree at most r, then $P \preceq Q$ implies that $\tilde{\mathbb{E}} P \leqslant \tilde{\mathbb{E}} Q$ every r-p.e.f. $\tilde{\mathbb{E}}$. This follows using linearity and positivity, as well as the (not too hard to verify) observation that if $Q - P = \sum_i R_i^2$ then it must hold that $\deg(R_i) \leqslant \max\{\deg(P), \deg(Q)\}/2$ for every i.

One of the most useful properties of pseudo-expectation is that it satisfies the Cauchy–Schwarz inequality:

Lemma A.2 (Pseudo Cauchy–Schwarz,[BBH+12])**.** *Let P and Q be two polynomials of degree at most r. Then, $\tilde{\mathbb{E}} PQ \leqslant \sqrt{\tilde{\mathbb{E}} P^2} \cdot \sqrt{\tilde{\mathbb{E}} Q^2}$ for any degree-$2r$ pseudo-expectation functional $\tilde{\mathbb{E}}$.*

Proof. We first consider the case $\tilde{\mathbb{E}} P^2, \tilde{\mathbb{E}} Q^2 > 0$. Then, by linearity of $\tilde{\mathbb{E}}$, we may assume that $\tilde{\mathbb{E}} P^2 = \tilde{\mathbb{E}} Q^2 = 1$. Since $2PQ \preceq P^2 + Q^2$ (by expanding the square $(P - Q)^2$), it follows that $\tilde{\mathbb{E}} PQ \leqslant \frac{1}{2} \tilde{\mathbb{E}} P^2 + \frac{1}{2} \tilde{\mathbb{E}} Q^2 = 1$ as desired. It remains to consider the case $\tilde{\mathbb{E}} P^2 = 0$. In this case, $2\alpha PQ \preceq P^2 + \alpha^2 Q^2$ implies that $\tilde{\mathbb{E}} PQ \leqslant \alpha \cdot \frac{1}{2} \tilde{\mathbb{E}} Q^2$ for all $\alpha > 0$. Thus $\tilde{\mathbb{E}} PQ = 0$, as desired. □

In particular this implies the following corollary

Corollary A.3 ([BBH+12]). *If P is a polynomial of degree $\leqslant r$, and $\tilde{\mathbb{E}}_x$ is a $2r$-p.e.f. such that $\tilde{\mathbb{E}}\, P(x)^2 = 0$, then $\tilde{\mathbb{E}}\, P(x)Q(x) = 0$ for every Q of degree $\leqslant r$.*

Proof. By Lemma A.2,

$$\tilde{\mathbb{E}}\, PQ \leqslant \sqrt{\tilde{\mathbb{E}}\, P^2}\sqrt{\tilde{\mathbb{E}}\, Q^2} = 0$$

□

In this paper we also need the following variant of Hölder's inequality:

Lemma A.4 (Pseudoexpectation Hölder). *Let $d, c, k \in \mathbb{N}$, $\mathcal{D}$ be a level $\ell \geqslant 10dck$ pseudodistribution over $\mathbb{R}^n$, and P a sum of squares n-variate polynomial of degree d, then*

$$\tilde{\mathbb{E}}_{X\sim\mathcal{D}} P(X)^{r'} \geqslant \left(\tilde{\mathbb{E}}_{X\sim\mathcal{D}} P(X)^{r}\right)^{r'/r}$$

where $r = ck$ and $r' = (c+1)k$.

Proof. We'll do the proof by induction on r. The base case is $r = c$ in which case this is simply the pseudoexpectation Cauchy Schwarz that $\tilde{\mathbb{E}}\, P(X)^{2c} \geqslant (\tilde{\mathbb{E}}\, P(X)^c)^2$. Define $\mathcal{D}'$ to be the pseudodistribution obtained by reweighing $\mathcal{D}$ according to $P(X)^{r-c}$. Using $\tilde{\mathbb{E}}_{\mathcal{D}'} P(X)^{2c} \geqslant \left(\tilde{\mathbb{E}}_{\mathcal{D}'} P(X)^c\right)^2$ we can write

$$\frac{\tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r+x}}{\tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r-c}} = \frac{\tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r-c}P(X)^{2c}}{\tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r-c}} \geqslant \left(\frac{\tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r-c}P(X)^{c}}{\tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r-c}}\right)^2$$

moving things around we get that

$$\tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r+c} \geqslant \left(\tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r}\right)^2 / \tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r-c}$$

which using our induction hypothesis on r vs $r-c$, we can lower bound by

$$\left(\tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r}\right)^2 / \left(\tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r}\right)^{(r-c)/r} = \left(\tilde{\mathbb{E}}_{\mathcal{D}} P(X)^{r}\right)^{(r+c)/r}$$

□

We sometime would need to extend a pseudoexpectation of one random variable to a pseudoexpectation of two independent copies of it. The following lemma would be useful there

Lemma A.5. *Suppose that X and Y are two pseudodistributions of level ℓ. Then we can define a level ℓ pseudoexpectation operator on X, Y such that for every two polynomials P Q of degree at most $\ell/2$, $\tilde{\mathbb{E}}\, P(X)Q(Y) = (\tilde{\mathbb{E}}\, P(X))(\tilde{\mathbb{E}}\, Q(Y))$.*

Proof. We define the pseudoexpectation operator in the obvious way—for every set of ℓ indices $i_1, \ldots, i_k, j_{k+1}, \ldots, j_\ell$ we let $\tilde{\mathbb{E}}\, X_{i_1}\cdots X_{i_k}\cdot Y_{j_{k+1}}\cdots Y_{j_d} = (\tilde{\mathbb{E}}\, X_{i_1}\cdots X_{i_k})\cdot(\tilde{\mathbb{E}}\, Y_{j_{k+1}}\cdots Y_{j_d})$ and extend it linearly to all monomials. Clearly $\tilde{\mathbb{E}}\, 1 = 1$ and so the only thing left to do is to prove that for every polynomial P of degree $\leqslant \ell/2$ in the X, Y variables $\tilde{\mathbb{E}}\, P(X,Y)^2 \geqslant 0$.

Write $P(X,Y) = \sum M_i(X)N_i(Y)$ where M_i, N_i are monomials, then $P(X,Y)^2 = \sum_{i,j} M_i(X)M_j(X)N_i(Y)N_j(Y)$ and so under our definition

$$\tilde{\mathbb{E}}\, P(X,Y)^2 = \sum_{i,j}(\tilde{\mathbb{E}}\, M_i(X)M_j(X))(\tilde{\mathbb{E}}\, N_i(Y)N_j(Y)) = \langle A, B\rangle$$

where A and B are the matrices defined by $A_{i,j} = \tilde{\mathbb{E}}\, M_i(X)M_j(X)$ and $B_{i,j} = \tilde{\mathbb{E}}\, N_i(Y)N_j(Y)$. But the pseudoexpectation conditions on X, Y implies that both these matrices are p.s.d and so their dot product is nonnegative. □

We would like to understand how polynomials behave on linear subspaces of $\mathbb{R}^n$. A map $P\colon \mathbb{R}^n \to \mathbb{R}$ is *polynomial* over a linear subspace $V \subseteq \mathbb{R}^n$ if P restricted to V agrees with a polynomial in the coefficients for some basis of V. Concretely, if $g_1, \dots, g_m$ is an (orthonormal) basis of V, then P is *polynomial* over V if $P(f)$ agrees with a polynomial in $\langle f, g_1\rangle, \dots, \langle f, g_m\rangle$. We say that $P \preceq Q$ holds over a subspace V if $P - Q$, as a polynomial over V, is a sum of squares.

Lemma A.6 ([BBH+12])**.** *Let P and Q be two polynomials over $\mathbb{R}^n$ of degree at most r, and let $B\colon \mathbb{R}^n \to \mathbb{R}^k$ be a linear operator. Suppose that $P \preceq Q$ holds over the kernel of B. Then, $\tilde{\mathbb{E}}\, P \leqslant \tilde{\mathbb{E}}\, Q$ holds for any r-p.e.f. $\tilde{\mathbb{E}}$ over $\mathbb{R}^n$ that satisfies $\tilde{\mathbb{E}}_f \|Bf\|^2 = 0$.*

Proof. Since $P \preceq Q$ over the kernel of B, we can write $Q(f) = P(f) + \sum_{i=1}^m R_i^2(f) + \sum_{j=1}^k (Bf)_j S_j(f)$ for polynomials $R_1, \dots, R_m$ and $S_1, \dots, S_k$ over $\mathbb{R}^n$. By positivity, $\tilde{\mathbb{E}}_f R_i^2(f) \geqslant 0$ for all $i \in [m]$. We claim that $\tilde{\mathbb{E}}_f (Bf)_j S_j(f) = 0$ for all $j \in [k]$ (which would finish the proof). This claim follows from the fact that $\tilde{\mathbb{E}}_f (Bf)_j^2 = 0$ for all $j \in [k]$ and Lemma A.2. □

Lemma A.7 ([BBH+12])**.** *The relation $P^2 \preceq P$ holds if and only if $0 \preceq P \preceq 1$. Furthermore, if $P^2 \preceq P$ and $0 \preceq Q \preceq P$, then $Q^2 \preceq Q$.*

Proof. If $P \succeq 0$, then $P \preceq 1$ implies $P^2 \preceq P$. (Multiplying both sides with a sum of squares preserves the order.) On the other hand, suppose $P^2 \preceq P$. Since $P^2 \succeq 0$, we also have $P \succeq 0$. Since $1 - P = P - P^2 + (1-P)^2$, the relation $P^2 \preceq P$ also implies $P \preceq 1$.

For the second part of the lemma, suppose $P^2 \preceq P$ and $0 \preceq Q \preceq P$. Using the first part of the lemma, we have $P \preceq 1$. It follows that $0 \preceq Q \preceq 1$, which in turn implies $Q^2 \preceq Q$ (using the other direction of the first part of the lemma). □

Fact A.8. *If f is a d-f.r.v. over $\mathbb{R}^{\mathcal{U}}$ and $\{P_v\}_{v\in\mathcal{U}}$ are polynomials of degree at most k, then g with $g(v) = P_v(f)$ is a level-(d/k) pseudodistribution over $\mathbb{R}^{\mathcal{U}}$. (For a polynomial Q of degree at most d/k, the pseudo-expectation is defined as $\tilde{\mathbb{E}}_g\, Q(\{g(v)\}_{v\in\mathcal{U}}) := \tilde{\mathbb{E}}_f\, Q(\{P_v(f)\}_{v\in\mathcal{U}})$.)*

Lemma A.9 ([BBH+12])**.** *For $f, g \in L_2(\mathcal{U})$,*

$$\langle f, g\rangle \preceq \tfrac{1}{2}\|f\|^2 + \tfrac{1}{2}\|g\|^2 .$$

Proof. The right-hand side minus the LHS equals the square polynomial $\frac{1}{2}\langle f - g, f - g\rangle$ □

Here is another form of the Cauchy–Schwarz inequality.

Lemma A.10 (Function Cauchy–Schwarz inequality,[BBH+12])**.** *If (f, g) is a level-2 p.d. over $\mathbb{R}^{\mathcal{U}} \times \mathbb{R}^{\mathcal{U}}$, then*

$$\tilde{\mathbb{E}}_{f,g} \langle f, g\rangle \leqslant \sqrt{\tilde{\mathbb{E}}_f \|f\|^2} \cdot \sqrt{\tilde{\mathbb{E}}_g \|g\|^2} .$$

Proof. Let $\bar{f} = f/\sqrt{\tilde{\mathbb{E}}_f \|f\|^2}$ and $\bar{g} = g/\sqrt{\tilde{\mathbb{E}}_g \|g\|^2}$. Note $\tilde{\mathbb{E}}_{\bar{f}} \|\bar{f}\|^2 = \tilde{\mathbb{E}}_{\bar{g}} \|\bar{g}\|^2 = 1$. Since by Lemma A.9, $\langle \bar{f}, \bar{g}\rangle \preceq 1/2\|\bar{f}\|^2 + 1/2\|\bar{g}\|^2$, we can conclude the desired inequality,

$$\tilde{\mathbb{E}}_{f,g} \langle f, g\rangle = \sqrt{\tilde{\mathbb{E}}_f \|f\|^2} \cdot \sqrt{\tilde{\mathbb{E}}_g \|g\|^2}\, \tilde{\mathbb{E}}_{\bar{f},\bar{g}} \langle \bar{f}, \bar{g}\rangle \leqslant \sqrt{\tilde{\mathbb{E}}_f \|f\|^2} \cdot \sqrt{\tilde{\mathbb{E}}_g \|g\|^2} \cdot \underbrace{\left(\tfrac{1}{2}\, \tilde{\mathbb{E}}_{\bar{f}} \|\bar{f}\|^2 + \tfrac{1}{2}\, \mathbb{E}_{\bar{g}} \|\bar{g}\|^2\right)}_{=1} . \qquad \square$$

And it implies another form of Hölder's inequality

Corollary A.11 (Function Hölder's inequality,[BBH+12]). *If (f,g) is a level 4 p.d. over $\mathbb{R}^{\mathcal{U}} \times \mathbb{R}^{\mathcal{U}}$, then*

$$\tilde{\mathbb{E}}_{f,g} \mathop{\mathbb{E}}_{\omega\in\mathcal{U}} f(\omega)g(\omega)^3 \leqslant \left(\tilde{\mathbb{E}}_{f}\|f\|_4^4\right)^{1/4}\left(\tilde{\mathbb{E}}_{g}\|g\|_4^4\right)^{3/4}.$$

Proof. Using Lemma A.2 twice, we have

$$\tilde{\mathbb{E}}_{f,g} \mathop{\mathbb{E}}_{\omega\in\mathcal{U}} f(\omega)g(\omega)^3 \leqslant \left(\tilde{\mathbb{E}}_{f,g} \mathop{\mathbb{E}}_{\omega\in\mathcal{U}} f(\omega)^2 g(\omega)^2\right)^{1/2}\left(\tilde{\mathbb{E}}_{g}\|g\|_4^4\right)^{1/2} \leqslant \left(\tilde{\mathbb{E}}_{f}\|f\|_4^4\right)^{1/4}\left(\tilde{\mathbb{E}}_{g}\|g\|_4^4\right)^{3/4}.$$

□

A.1 Spectral norm and SOS proofs

Here we note the following alternative characterization of the spectral norm of a polynomial:

Lemma A.12. *Let P be a degree-4 homogenous polynomial , then $\|P\|_{\text{spectral}} \leqslant c$ if and only if there is a sum of squares degree 4 polynomial S such that $P(x) = c\|x\|_2^4 - S(x)$.*

Proof. Suppose that $\|P\|_{\text{spectral}} \leqslant C$. Then there is an $n^2 \times n^2$ matrix M such that $M \cdot x^{\otimes 4} = P(x)$ for all x $M = cI - S$ where I is the $n^2 \times n^2$ identity and S is a positive semidefinite matrix. That is, $S = \sum \lambda_i Q_i^{\otimes 2}$ for some $\lambda_i \geqslant 0$ and $Q_i \in \mathbb{R}^{n^2}$. Now, if we consider S as a degree 4 polynomial $S(x) = S \cdot x^{\otimes 4}$ then it equals $\sum \lambda_i (Q_i \cdot x^{\otimes 2})^2$ and hence it is a sum of squares, and it satisfies

$$P(x) = cI \cdot x^{\otimes 4} - S(x) = c\sum_{i,j} x_i^2 x_j^2 - S(x) = c\|x\|_2^4 - S(x).$$

On the other hand, suppose that $P(x) = c\|x\|_2^4 - \sum R_i(x)^2$ where the R_i's are quadratic polynomials. We can let $r_i \in \mathbb{R}^{n^2}$ be such that $r_i \cdot x^{\otimes 2} = R_i(x)$, and then let M be the quadratic operator on R^{n^2} such that $M(y) = c\|y\|_2^2 - \sum (r_i \cdot y)^2$ for every $y \in \mathbb{R}^{n^2}$. One can easily verify that the spectral norm of M is at most c and $M \cdot x^{\otimes 4} = P(x)$ for every $x \in \mathbb{R}^n$. □

B Low-Rank Tensor Optimization

For a vector $x \in \mathbb{R}^n$, let $\|x\|$ denote the Euclidean norm of x. For a polynomial $P \in \mathbb{R}[X_1, \ldots, X_n]$, we define its **norm** as $\|P\| \stackrel{\text{def}}{=} \max\{|P(x)| \mid \|x\| = 1\}$.

Consider an n-variate degree-4 polynomial P of the form $P(x) = \sum_{i=1}^r Q_i(x)^2$ for quadratic polynomials $Q_1, \ldots, Q_r$.

Theorem B.1. *There exists an algorithm that, given P and ε, computes $\|P\|$ up to multiplicative error ε in time* $\text{poly}(n) \cdot \exp(O(r \log(1/\varepsilon)))$.

Proof. For $\lambda \in \mathbb{R}^n$, consider the polynomial $Q_\lambda(x) = \sum_{i=1}^r \lambda_i Q_i(x)$.

First, we claim that $\max_{\|\lambda\|=1}\|Q_\lambda\| = \|P\|^{1/2}$. On the one hand, $\|Q_\lambda\| \leqslant \|\lambda\| \cdot \|P\|^{1/2}$ by Cauchy–Schwarz. On the other hand, if $\lambda = \frac{1}{P(x^*)^{1/2}}(Q_1(x), \ldots, Q_r(x))$ for some vector $x^* \in \mathbb{R}^n$, then $Q_\lambda(x^*) = P(x^*)^{1/2}$. Therefore, if we choose x^* as a unit vector that maximizes P, then $\|Q_\lambda\| \geqslant P(x^*)^{1/2} = \|P\|^{1/2}$.

Next, we claim that we can compute $\max_{\|\lambda\|=1}\|Q_\lambda\|$ up to error ε in time $\text{poly}(n) \cdot \exp(O(r \log(1/\varepsilon)))$. Since Q_λ is quadratic, we can compute $\|Q_\lambda\|$ in polynomial time. (The norm of Q_λ is equal to the largest singular value of the coefficient matrix of Q_λ.) The idea is to compute $\|Q_\lambda\|$ for all vectors $\lambda \in N_\varepsilon$, where N_ε is an ε-net of the unit ball in $\mathbb{R}^r$. (There exist such nets of size $(1/\varepsilon)^{O(r)}$.) Let λ^* be the vector that

achieves the maximum, x^* be the corresponding input, and u^* be the vector $(Q_1(x^*), \ldots, Q_r(x^*))$. Thus $\max_{\|\lambda\|=1} \|Q_\lambda\|^2 = \langle \lambda^*, u^* \rangle^2$ and $\|u^*\| = \|P\|$. Therefore, for every λ

$$\|Q_\lambda\|^2 \geqslant \|Q_\lambda(x^*)\|^2 = \langle \lambda, u^* \rangle^2 .$$

But if $\|\lambda - \lambda^*\| \leqslant \varepsilon$ then

$$|\langle \lambda^*, u^* \rangle - \langle \lambda, u^* \rangle| \leqslant \|\lambda - \lambda^*\| \|u^*\| = \|\lambda - \lambda^*\| \|P\| .$$

Thus if $\|\lambda - \lambda'\| \leqslant \varepsilon$ then we get a $1 - O(\varepsilon)$ multiplicative approximation to $\|P\|$. □

Corollary B.2. *If M is a symmetric $n^2 \times n^2$ PSD matrix with Frobenius norm at most 1 then we can compute an ε additive approximation to*

$$\max_{\|x\|=1} \langle M, x^{\otimes 4} \rangle$$

in $\mathrm{poly}(n) \exp(\mathrm{poly}(1/\varepsilon))$ *time.*

Proof. Write M in its eigenbasis as $M = \sum \lambda_i Q_i^{\otimes 2}$ for $n \times n$ matrices $\{Q_i\}$ with Frobenius norm at most 1, and let $M' = \sum_{\lambda_i \geqslant \varepsilon} \lambda_i Q_i^{\otimes 2}$. Since $\sum \lambda_i^2 = 1$ we know that the rank of M' is at most $1/\varepsilon^2$, and therefore we can compute a $1 \pm \varepsilon$ multiplicative approximation to the maximum of $\langle M', x^{\otimes 4} \rangle$ over unit x (which in particular implies an ε additive approximation since this value is bounded by 1). But this implies an ε-additive approximation for this maximum over M since these quantities can differ by at most ε. □

C LOCC Polynomial Optimization

Let $P \in \mathbb{R}[X]_4$ be a degree-4 homogeneous polynomial of the form $P(X) = A_1(X) \cdot B_1(X) + \cdots + A_m(X) \cdot B_m(X)$ for quadratic polynomials $A_1, \ldots, A_m \in \mathbb{R}[X]_2$ with $0 \preceq A_i \preceq \|X\|^2$ and quadratic polynomials $B_1, \ldots, B_m \in \mathbb{R}[X]_2$ with $B_i \succeq 0$ and $\sum_i B_i \preceq \|X\|^2$. Note that this corresponds to the tensor corresponding to P having a one-way local operations and classical communication (LOCC) norm bounded by 1 [BCY11]. Without loss of generality, we may assume $\sum_i B_i = \|X\|^2$. (We can choose $A_m = 0$ and choose B_m appropriately without changing P.) Our goal is to compute the norm of P, defined as $\|P\| = \max_{\|x\|=1} |P(x)|$. In the quantum setting, this corresponds to finding the maximum probability of acceptance by a separable state for the measurement operator P.

In this section, we will show that sum-of-squares relaxations provide good approximation for the norm of polynomials of the form above. (For the case that the variables of $A_1, \ldots, A_m$ are disjoint from the variables for $B_1, \ldots, B_m$, the theorem is due to Brandão, Christandl, and Yard [BCY11]. Up to the Gaussian rounding step, the proof here is essentially the same as the proof by Brandão and Harrow [BH13].)

Theorem C.1. *Sum-of-squares relaxations with degree d achieve the following approximations for the norm of degree-4 polynomials P of the form above:*

- *the value of the relaxation is at most $3\|P\| + \varepsilon$ for degree $d \geqslant O(1/\varepsilon^2) \cdot \log n$.*
- *in the case that the variables in $A_1, \ldots, A_m$ are disjoint from the variables in $B_1, \ldots, B_m$, the value of the relaxation is at most $\|P\| + \varepsilon$ for degree $d \geqslant O(1/\varepsilon^2) \log n$.*

Direct Rounding. As direct rounding for a distribution $\{X\}$, we choose a Gaussian variable with the same first two moments as $\{X\}$. To analyze this rounding procedure, the following lemma is useful.

The lemma considers an arbitrary distribution over unit vectors in $\mathbb{R}^n$ (intended to maximize P). We express the second moment $\rho = \mathbb{E} XX^\top$ of this distribution as convex combination $\rho = \sum_i \beta_i \rho_i$ for $\beta_i = \mathbb{E}_X B_i(X)$ and $\rho_i = \mathbb{E}_X XX^T B_i(X)/\beta_i$. By the assumptions on the distribution $\{X\}$, the matrices ρ and $\rho_1, \ldots, \rho_m$

are positive semidefinite and have trace 1 (*density matrices*). The quantum entropy $H(\rho) = -\operatorname{Tr}\rho\log\rho$ is concave so that $H(\rho) \geqslant \sum_i \beta_i H(\rho_i)$. The assumption of the lemma is that the inequality is approximately tight. Roughly speaking, this condition means that the ρ_i matrices are close ρ. For the distribution $\{X\}$, this condition means that reweighing by the polynomials B_i does not affect second moments of the distribution. We say that the distribution has *low global correlation* with respect to the polynomials $B_1, \dots, B_m$. (This notion is related to but distinct from the notion of global correlation in [BRS11]). The lemma asserts that if the distribution $\{X\}$ has low global correlation with respect to the polynomials $B_1, \dots, B_m$, then sampling X independently for the A-part and B-part of the polynomial P gives roughly the same value as sampling X in a correlated way. (The next lemma explains why our direct rounding achieves at least the quantity corresponding to sampling X independently for the two parts.)

Lemma C.2. *Let $\{X\}$ be a distribution over $\mathbb{R}^n$ that satisfies the constraint $\|X\|^2 = 1$. Suppose $\sum_i \beta_i H(\rho_i) \geqslant H(\sum_i \beta_i \rho_i) - \varepsilon^2$ for $\rho_i = \mathbb{E}_X XX^\top B_i(X)/\beta_i$ and $\beta_i = \mathbb{E}_X B_i(X)$. Then,*

$$\sum_i \mathop{\mathbb{E}}_X A_i(X) \cdot \mathop{\mathbb{E}}_X B_i(X) \geqslant \sum_i \mathop{\mathbb{E}}_X A_i(X) B_i(X) \quad - \varepsilon\,.$$

Moreover, the statement holds if $\{X\}$ is a degree-4 pseudo-distribution.

Proof. Consider the block-diagonal density matrix $\rho = \sum_i \beta_i \rho_i \otimes e_i e_i^\top$, and the block-diagonal measurement matrix $A = \sum_i A_i \otimes e_i e_i^\top$. (In this construction, we identify the quadratic polynomial A with its representation as a symmetric square matrix.) Furthermore, consider the partial traces $\rho_A = \operatorname{Tr}_B \rho = \sum_i \beta_i \rho_i$ and $\rho_B = \operatorname{Tr}_A \rho = \sum_i \beta_i e_i e_i^T$. We can express the two sides of the conclusion of the lemma as follows,

$$\sum_i \mathop{\mathbb{E}}_X A_i(X) \cdot B_i(X) = \operatorname{Tr} A\rho\,,$$
$$\sum_i \mathop{\mathbb{E}}_X A_i(X) \cdot \mathop{\mathbb{E}}_X B_i(X) = \operatorname{Tr} A(\rho_A \otimes \rho_B)\,.$$

Since A has spectral norm at most 1, we can bound the difference by$|\operatorname{Tr} A(\rho - \rho_A \otimes \rho_B)| \leqslant \|\rho - \rho_A \otimes \rho_B\|_*$. (Here, $\|\cdot\|_*$ is the trace norm—the dual of the spectral norm.) By Pinsker's inequality, $\|\rho - \rho_A \otimes \rho_B\|^2 \leqslant H(\rho_A) + H(\rho_B) - H(\rho)$. By the chain rule, $H(\rho) = H(\rho_B) + \sum_i \beta_i H(\rho_i)$. The assumption of the lemma allows us to bound the trace norm by $\|\rho - \rho_A \otimes \rho_B\|^2 \leqslant H(\sum_i \beta_i \rho_i) - \sum_i \beta_i H(\rho_i) \leqslant \varepsilon^2$. At this point, the conclusion of the lemma follows from the bound $|\operatorname{Tr} A(\rho - \rho_A \otimes \rho_B)| \leqslant \|\rho - \rho_A \otimes \rho_B\|_* \leqslant \varepsilon$. □

The following lemma shows that Gaussian rounding achieves a value at least as large as the value achieved by sampling $\{X\}$ independently for the A-part and B-part of the polynomial P.

Lemma C.3. *Let $\{X\}$ be a distribution over $\mathbb{R}^n$ that satisfies the constraint $\|X\|^2 = 1$. Suppose $\{X'\}$ is a Gaussian distribution with the same first two moments as $\{X\}$. Then, $\{X'\}$ satisfies $\mathbb{E}_{X'} \|X'\|^2 = 1$, $\mathbb{E}_{X'} \|X'\|^4 = 3$, and*

$$\mathop{\mathbb{E}}_{X'} \sum_i A_i(X') \cdot B_i(X') \geqslant \sum_i \mathop{\mathbb{E}}_X A_i(X) \cdot \mathop{\mathbb{E}}_X B_i(X)\,.$$

Moreover, the statement holds if $\{X\}$ is a degree-4 pseudo-distribution.

Proof. Using the assumption $A_i, B_i \succeq 0$, the lemma follows from the fact that Gaussian variables P, Q satisfy $\mathbb{E}\, P^2 Q^2 \geqslant \mathbb{E}\, P^2\, \mathbb{E}\, Q^2$. □

The previous two lemmas together yield the following corollary.

Corollary C.4. *Let $\{X\}$ be a distribution over $\mathbb{R}^n$ that satisfies the constraint $\|X\|^2 = 1$. Suppose $\{X'\}$ and $\varepsilon > 0$ are as in the previous two lemmas, that is, $\{X'\}$ is a Gaussian distribution with the same first two moments as $\{X\}$ and $\sum_i \beta_i H(\rho_i) \geqslant H(\sum_i \beta_i \rho_i) - \varepsilon^2$ for $\rho_i = \mathbb{E}_X XX^\top B_i(X)/\beta_i$ and $\beta_i = \mathbb{E}_X B_i(X)$. Then,*

$$\mathop{\mathbb{E}}_{X'} P(X') \geqslant \mathop{\mathbb{E}}_{X} P(X) - \varepsilon \text{ and } \mathop{\mathbb{E}}_{X'} \|X'\|^4 = 3\,.$$

Moreover, the statement holds if $\{X\}$ is a degree-4 pseudo-distribution.

Making progress. The following lemma shows that there exists a low-degree polynomial so that reweighing by the polynomial results in a distribution that has low global correlation with respect to the polynomials $B_1, \ldots, B_m$.

Lemma C.5. *Let $\{X\}$ be a distribution over $\mathbb{R}^n$ that satisfies the constraint $\|X\|^2 = 1$. Then, there exists a polynomial $B \in \mathbb{R}[X]_{2d}$ of the form $B = B_{i(1)} \cdots B_{i(d)}$ with $d = O(1/\varepsilon^2) \log n$ such that $\sum_i \beta_i H(\rho_i) \geqslant H(\sum_i \beta_i \rho_i) - \varepsilon^2$ for $\rho_i = \mathbb{E}_X XX^\top B(X) B_i(X)/\beta_i$ and $\beta_i = \mathbb{E}_X B(X) B_i(X)$. Moreover, the statement holds if $\{X\}$ is a degree-$d+4$ pseudo-distribution.*

Proof. By contraposition, suppose that $\sum_i \beta_i H(\rho_i) < H(\sum_i \beta_i \rho_i) - \eta$ holds for all polynomials B of the form $B = B_{i(1)} \cdots B_{i(d')}$ with $d' \leqslant d = 10/\varepsilon^2 \cdot \log n$. Then, we can greedily construct a sequence of polynomial $B_{i^*(1)}, \ldots, B_{i^*(d)}$ such that in each step the entropy decreases by at least η. In particular, $H(\rho^*) \leqslant H(\rho) - \eta \cdot d$ for $\rho^* \propto \mathbb{E}_X XX^\top B_{i^*(1)} \cdots B_{i^*(d)}(X)$ and $\rho = \mathbb{E}_X XX^\top$. Since $H(\rho) \leqslant \log n$ and $H(\rho^*) \geqslant 0$, we have $\eta \geqslant 1/d \cdot \log n = \varepsilon^2/10$. As desired it follows that there exists a polynomial B of the desired form such that $\sum_i \beta_i H(\rho_i) \geqslant H(\sum_i \beta_i \rho_i) - \varepsilon^2$. □

Putting things together. The following lemma combines the conclusion about direct-rounding and making-progress.

Lemma C.6. *Let $\{X\}$ be a distribution over $\mathbb{R}^n$ that satisfies the constraints $\|X\|^2 = 1$ and $P(X) \geqslant c$. Then, there exists a polynomial $B \in \mathbb{R}[X]_{2d}$ of the form $B = B_{i(1)} \cdots B_{i(d)}$ with $d = O(1/\varepsilon^2) \log n$ such that the Gaussian distribution X' that matches the first two moments of $\{X\}$ reweighted by $B(X)$ satisfies*

$$\mathop{\mathbb{E}}_{X'} P(X) \geqslant c - \varepsilon \text{ and } \mathop{\mathbb{E}}_{X'} \|X'\|^4 = 3\,.$$

(Concretely, $\{X'\}$ is the Gaussian distribution that satisfies $\mathbb{E}_{X'} Q(X') = \mathbb{E}_X Q(X)B(X)/\mathbb{E}_X B(X)$ for quadratic polynomial Q). Moreover, the statement holds for degree-$2d+4$ pseudo-distributions.

Proof. Take the polynomial B as in Lemma C.5. Reweigh the distribution $\{X\}$ by the polynomial B. Apply Corollary C.4 to the resulting distribution. □

At this time, we have all ingredients for the proof of Theorem C.1.

Proof of Theorem C.1. Let $\{X\}$ be a degree-$d+4$ pseudo-distribution over $\mathbb{R}^n$ that satisfies the constraints $\|X\|^2 = 1$ and $P(X) \geqslant c$ for $d = O(1/\varepsilon^2) \log n$. By the previous lemma, there exists a distribution $\{X'\}$ over $\mathbb{R}^n$ such that $\mathbb{E}_{X'} P(X')/\mathbb{E}_{X'} \|X'\|^4 \geqslant c/3 - \varepsilon$. It follows that there exists a vector $x \in \mathbb{R}^n$ with $P(x)/\|x\|^4 \geqslant c/3 - \varepsilon$. (We can also find such a vector efficiently because we can sample from the distribution $\{X'\}$ efficiently and the random variables $P(X')$ and $\|X'\|$ are well-behaved.) By homogeneity, we get $\|P\| \geqslant c/3 - \varepsilon$.

In the case that the variables Y in $A_1, \ldots, A_m$ are disjoint from the variables Z in $B_1, \ldots, B_m$, we can modify the direct-rounding distribution $\{X'\} = \{(Y', Z')\}$ slightly and sample the variables Y' for the A_i polynomials independently from the variables Z' for the B_i polynomials. By Lemma C.2, we still have $\mathbb{E}_{X'} P(X') \geqslant c - \varepsilon$. We can assume that $\mathbb{E}\|Y'\|^2 = \mathbb{E}\|Z'\|^2 = 1/2$ (by adding the corresponding constraint to the sos relaxation). Therefore, $\mathbb{E}\|Y'\|^2\|X'\|^2 = 1/4$. It follows that there exists a vector $x = (y, z)$ in $\mathbb{R}^n$ with $P(y, z)/(\|y\|^2 \cdot \|z\|^2) \geqslant 4(c - \varepsilon)$. By homogeneity, we can assume that $\|y\|^2 = 1/2$ and $\|z\|^2 = 1/2$. In this case, $\|x\|^2 = 1$ and $P(x)gc - \varepsilon$ as desired. □

D The 2-to-q norm and small-set expansion

This appendix reproduces from [BBH+12] the proof that a graph is a *small-set expander* if and only if the projector to the subspace of its adjacency matrix's top eigenvalues has a bounded $2 \to q$ norm for even $q \geqslant 4$. We also note that while [BBH+12] stated their result for the decision question, it does yield an efficient algorithm to transform a vector in the top eigenspace with large 4 norm into a small set that does not expand.

Notation. For a regular graph $G = (V, E)$ and a subset $S \subseteq V$, we define the *measure* of S to be $\mu(S) = |S|/|V|$ and we define $G(S)$ to be the distribution obtained by picking a random $x \in S$ and then outputting a random neighbor y of x. We define the *expansion* of S, to be $\Phi_G(S) = \mathbb{P}_{y \in G(S)}[y \notin S]$, where y is a random neighbor of x. For $\delta \in (0,1)$, we define $\Phi_G(\delta) = \min_{S \subseteq V : \mu(S) \leqslant \delta} \Phi_G(S)$. We often drop the subscript G from Φ_G when it is clear from context. We identify G with its normalized adjacency (i.e., random walk) matrix. For every $\lambda \in [-1, 1]$, we denote by $V_{\geqslant \lambda}(G)$ the subspace spanned by the eigenvectors of G with eigenvalue at least λ. The projector into this subspace is denoted $P_{\geqslant \lambda}(G)$. For a distribution D, we let $\mathsf{cp}(D)$ denote the collision probability of D (the probability that two independent samples from D are identical).

Our main theorem of this section is the following:

Theorem D.1. *For every regular graph G, $\lambda > 0$ and even q,*

1. (Norm bound implies expansion) *For all $\delta > 0, \varepsilon > 0$, $\|P_{\geqslant \lambda}(G)\|_{2 \to q} \leqslant \varepsilon / \delta^{(q-2)/2q}$ implies that $\Phi_G(\delta) \geqslant 1 - \lambda - \varepsilon^2$.*

2. (Expansion implies norm bound) *There is a constant c such that for all $\delta > 0$, $\Phi_G(\delta) > 1 - \lambda 2^{-cq}$ implies $\|P_{\geqslant \lambda}(G)\|_{2 \to q} \leqslant 2/\sqrt{\delta}$. Moreover there is an efficient algorithm such that given a function $f \in V_{\geqslant \lambda}(G)$ such that $\|f\|_q > 2\|f\|_2/\sqrt{\delta}$ finds a set S of measure less than δ such that $\Phi_G(S) \leqslant 1 - \lambda 2^{-cq}$.*

Corollary D.2. *If there is a polynomial-time computable relaxation $\mathcal{R}$ yielding good approximation for the $2 \to q$, then the* Small-Set Expansion Hypothesis *of [RS10] is false.*

Proof. Using [RST12], to refute the small-set expansion hypothesis it is enough to come up with an efficient algorithm that given an input graph G and sufficiently small $\delta > 0$, can distinguish between the *Yes* case: $\Phi_G(\delta) < 0.1$ and the *No* case $\Phi_G(\delta') > 1 - 2^{-c \log(1/\delta')}$ for any $\delta' \geqslant \delta$ and some constant c. In particular for all $\eta > 0$ and constant d, if δ is small enough then in the *No* case $\Phi_G(\delta^{0.4}) > 1 - \eta$. Using Theorem D.1, in the *Yes* case we know $\|V_{1/2}(G)\|_{2 \to 4} \geqslant 1/(10\delta^{1/4})$, while in the *No* case, if we choose η to be smaller then $\eta(1/2)$ in the Theorem, then we know that $\|V_{1/2}(G)\|_{2 \to 4} \leqslant 2/\sqrt{\delta^{0.2}}$. Clearly, if we have a good approximation for the $2 \to 4$ norm then, for sufficiently small δ we can distinguish between these two cases. □

The first (easier) part of Theorem D.1 is proven in Section D.1. The second part will follow from the following lemma:

Lemma D.3. *Set $e = e(\lambda, q) := 2^{cq}/\lambda$, with a constant $c \leqslant 100$. Then for every $\lambda > 0$ and $1 \geqslant \delta \geqslant 0$, if G is a graph that satisfies*

$$\mathsf{cp}(G(S)) \leqslant 1/(e|S|) \tag{D.1}$$

for all S with $\mu(S) \leqslant \delta$, then $\|f\|_q \leqslant 2\|f\|_2/\sqrt{\delta}$ for all $f \in V_{\geqslant \lambda}(G)$. Moreover, there is an efficient algorithm that given a function $f \in V_{\geqslant \lambda}(G)$ such that $\|f\|_q > 2\|f\|_2/\sqrt{\delta}$ finds a set S that violates (D.1).

Proving the second part of Theorem D.1 from Lemma D.3. We use the variant of the local Cheeger bound obtained in [Ste10b, Theorem 2.1], stating that if $\Phi_G(\delta) \geqslant 1-\eta$ then for every $f \in L_2(V)$ satisfying $\|f\|_1^2 \leqslant \delta \|f\|_2^2$, $\|Gf\|_2^2 \leqslant c\sqrt{\eta}\|f\|_2^2$. The proof follows by noting that for every set S, if f is the characteristic function of S then $\|f\|_1 = \|f\|_2^2 = \mu(S)$, and $\mathsf{cp}(G(S)) = \|Gf\|_2^2/(\mu(S)|S|)$. Because this local Cheeger bound is algorithmic (and transforms a function with large L_2/L_1 ratio into a set by simply using a threshold cut), this part is algorithmic as well. □

Proof of Lemma D.3. Fix $\lambda > 0$. We assume that the graph satisfies the condition of the Lemma with $e = 2^{cq}/\lambda$, for a constant c that we'll set later. Let $G = (V, E)$ be such a graph, and f be function in $V_{\geqslant\lambda}(G)$ with $\|f\|_2 = 1$ that maximizes $\|f\|_q$. We write $f = \sum_{i=1}^m \alpha_i \chi_i$ where $\chi_1, \ldots, \chi_m$ denote the eigenfunctions of G with values $\lambda_1, \ldots, \lambda_m$ that are at least λ. Assume towards a contradiction that $\|f\|_q > 2/\sqrt{\delta}$. We'll prove that $g = \sum_{i=1}^m (\alpha_i/\lambda_i)\chi_i$ satisfies $\|g\|_q \geqslant 10\|f\|_q/\lambda$. This is a contradiction since (using $\lambda_i \in [\lambda, 1]$) $\|g\|_2 \leqslant \|f\|_2/\lambda$, and we assumed f is a function in $V_{\geqslant\lambda}(G)$ with a maximal ratio of $\|f\|_q/\|f\|_2$. (To prove the "moreover" part, where we don't assume f is the maximal function, we repeat this process with g until we get stuck.)

Let $U \subseteq V$ be the set of vertices such that $|f(x)| \geqslant 1/\sqrt{\delta}$ for all $x \in U$. Using Markov and the fact that $\mathbb{E}_{x\in V}[f(x)^2] = 1$, we know that $\mu(U) = |U|/|V| \leqslant \delta$, meaning that under our assumptions any subset $S \subseteq U$ satisfies $\mathsf{cp}(G(S)) \leqslant 1/(e|S|)$. On the other hand, because $\|f\|_q^q \geqslant 2^q/\delta^{q/2}$, we know that U contributes at least half of the term $\|f\|_q^q = \mathbb{E}_{x\in V} f(x)^q$. That is, if we define α to be $\mu(U)\,\mathbb{E}_{x\in U} f(x)^q$ then $\alpha \geqslant \|f\|_q^q/2$. We'll prove the lemma by showing that $\|g\|_q^q \geqslant 10\alpha/\lambda$.

Let c be a sufficiently large constant ($c = 100$ will do). We define U_i to be the set $\{x \in U : f(x) \in [c^i/\sqrt{\delta}, c^{i+1}/\sqrt{\delta})\}$, and let I be the maximal i such that U_i is non-empty. Thus, the sets $U_0, \ldots, U_I$ form a partition of U (where some of these sets may be empty). We let α_i be the contribution of U_i to α. That is, $\alpha_i = \mu_i\,\mathbb{E}_{x\in U_i} f(x)^q$, where $\mu_i = \mu(U_i)$. Note that $\alpha = \alpha_0 + \cdots + \alpha_I$. We'll show that there are some indices $i_1, \ldots, i_J$ such that:

(i) $\alpha_{i_1} + \cdots + \alpha_{i_J} \geqslant \alpha/(2c^{10})$.

(ii) For all $j \in [J]$, there is a nonnegative function $g_j : V \to \mathbb{R}$ such that $\mathbb{E}_{x\in V} g_j(x)^q \geqslant e\alpha_{i_j}/(10c^2)^{q/2}$.

(iii) For every $x \in V$, $g_1(x) + \cdots + g_J(x) \leqslant |g(x)|$.

Showing these will complete the proof, since it is easy to see that for two nonnegative functions and even q, g', g'', $\mathbb{E}(g'(x) + g''(x))^q \geqslant \mathbb{E}\, g'(x)^q + \mathbb{E}\, g''(x)^q$, and hence **(ii)** and **(iii)** imply that

$$\|g\|_4^4 = \mathbb{E}\, g(x)^4 \geqslant (e/(10c^2)^{q/2}) \sum_j \alpha_{i_j} \,. \tag{D.2}$$

Using **(i)** we conclude that for $e \geqslant (10c)^q/\lambda$, the right-hand side of (D.2) will be larger than $10\alpha/\lambda$.

We find the indices $i_1, \ldots, i_J$ iteratively. We let $\mathcal{I}$ be initially the set $\{0..I\}$ of all indices. For $j = 1, 2, \ldots$ we do the following as long as $\mathcal{I}$ is not empty:

1. Let i_j be the largest index in $\mathcal{I}$.
2. Remove from $\mathcal{I}$ every index i such that $\alpha_i \leqslant c^{10}\alpha_{i_j}/2^{i-i_j}$.

We let J denote the step when we stop. Note that our indices $i_1, \ldots, i_J$ are sorted in descending order. For every step j, the total of the α_i's for all indices we removed is less than $c^{10}\alpha_{i_j}$ and hence we satisfy **(i)**. The crux of our argument will be to show **(ii)** and **(iii)**. They will follow from the following claim:

Claim D.4. *Let $S \subseteq V$ and $\beta > 0$ be such that $|S| \leqslant \delta$ and $|f(x)| \geqslant \beta$ for all $x \in S$. Then there is a set T of size at least $e|S|$ such that $\mathbb{E}_{x\in T} g(x)^2 \geqslant \beta^2/4$.*

The claim will follow from the following lemma:

Lemma D.5. *Let D be a distribution with $\mathsf{cp}(D) \leqslant 1/N$ and g be some function. Then there is a set T of size N such that $\mathbb{E}_{x\in T} g(x)^2 \geqslant (\mathbb{E} g(D))^2/4$.*

Proof. Identify the support of D with the set $[M]$ for some M, we let p_i denote the probability that D outputs i, and sort the p_i's such that $p_1 \geqslant p_2 \cdots p_M$. We let β' denote $\mathbb{E} g(D)$; that is, $\beta' = \sum_{i=1}^{M} p_i g(i)$. We separate to two cases. If $\sum_{i>N} p_i g(i) \geqslant \beta'/2$, we define the distribution D' as follows: we set $\mathbb{P}[D' = i]$ to be p_i for $i > N$, and we let all $i \leqslant N$ be equiprobable (that is be output with probability $(\sum_{i=1}^{N} p_i)/N$). Clearly, $\mathbb{E}|g(D')| \geqslant \sum_{i>N} p_i g(i) \geqslant \beta'/2$, but on the other hand, since the maximum probability of any element in D' is at most $1/N$, it can be expressed as a convex combination of flat distributions over sets of size N, implying that one of these sets T satisfies $\mathbb{E}_{x\in T}|g(x)| \geqslant \beta'/2$, and hence $\mathbb{E}_{x\in T} g(x)^2 \geqslant \beta'^2/4$.

The other case is that $\sum_{i=1}^{N} p_i g(i) \geqslant \beta'/2$. In this case we use Cauchy–Schwarz and argue that

$$\beta'^2/4 \leqslant \left(\sum_{i=1}^{N} p_i^2\right)\left(\sum_{i=1}^{N} g(i)^2\right) . \tag{D.3}$$

But using our bound on the collision probability, the right-hand side of (D.3) is upper bounded by $\frac{1}{N}\sum_{i=1}^{N} g(i)^2 = \mathbb{E}_{x\in[N]} g(x)^2$. □

Proof of Claim D.4 from Lemma D.5. By construction $f = Gg$, and hence we know that for every x, $f(x) = \mathbb{E}_{y\sim x} g(y)$. This means that if we let D be the distribution $G(S)$ then

$$\mathbb{E}|g(D)| = \mathop{\mathbb{E}}_{x\in S}\mathop{\mathbb{E}}_{y\sim x}|g(y)| \geqslant \mathop{\mathbb{E}}_{x\in S}|\mathop{\mathbb{E}}_{y\sim x} g(y)| = \mathop{\mathbb{E}}_{x\in S}|f(x)| \geqslant \beta .$$

By the expansion property of G, $\mathsf{cp}(D) \leqslant 1/(e|S|)$ and thus by Lemma D.5 there is a set T of size $e|S|$ satisfying $\mathbb{E}_{x\in T} g(x)^2 \geqslant \beta^2/4$. □

We will construct the functions $g_1, \ldots, g_J$ by applying iteratively Claim D.4. We do the following for $j = 1, \ldots, J$:

1. Let T_j be the set of size $e|U_{i_j}|$ that is obtained by applying Claim D.4 to the function f and the set U_{i_j}. Note that $\mathbb{E}_{x\in T_j} g(x)^2 \geqslant \beta_{i_j}^2/4$, where we let $\beta_i = c^i/\sqrt{\delta}$ (and hence for every $x \in U_i$, $\beta_i \leqslant |f(x)| \leqslant c\beta_i$).
2. Let g'_j be the function on input x that outputs $\gamma \cdot |g(x)|$ if $x \in T_j$ and 0 otherwise, where $\gamma \leqslant 1$ is a scaling factor that ensures that $\mathbb{E}_{x\in T_j} g'(x)^2$ equals exactly $\beta_{i_j}^2/4$.
3. We define $g_j(x) = \max\{0, g'_j(x) - \sum_{k<j} g_k(x)\}$.

Note that the second step ensures that $g'_j(x) \leqslant |g(x)|$, while the third step ensures that $g_1(x) + \cdots + g_j(x) \leqslant g'_j(x)$ for all j, and in particular $g_1(x) + \cdots + g_J(x) \leqslant |g(x)|$. Hence the only thing left to prove is the following:

Claim D.6. $\mathbb{E}_{x\in V} g_j(x)^q \geqslant e\alpha_{i_j}/(10c)^{q/2}$

Proof. Recall that for every i, $\alpha_i = \mu_i \mathbb{E}_{x\in U_i} f(x)^q$, and hence (using $f(x) \in [\beta_i, c\beta_i)$ for $x \in U_i$):

$$\mu_i \beta_i^q \leqslant \alpha_i \leqslant \mu_i c^q \beta_i^q . \tag{D.4}$$

Now fix $T = T_j$. Since $\mathbb{E}_{x\in V} g_j(x)^q$ is at least (in fact equal) $\mu(T)\,\mathbb{E}_{x\in T} g_j(x)^q$ and $\mu(T) = e\mu(U_{i_j})$, we can use (D.4) and $\mathbb{E}_{x\in T} g_j(x)^q \geqslant (E_{x\in T} g_j(x)^2)^{q/2}$, to reduce proving the claim to showing the following:

$$\mathop{\mathbb{E}}_{x\in T} g_j(x)^2 \geqslant (c\beta_{i_j})^2/(10c^2) = \beta_{i_j}^2/10 . \tag{D.5}$$

We know that $\mathbb{E}_{x\in T} g'_j(x)^2 = \beta_{i_j}^2/4$. We claim that (D.5) will follow by showing that for every $k < j$,

$$\mathop{\mathbb{E}}_{x\in T} g'_k(x)^2 \leqslant 100^{-i'} \cdot \beta_{i_j}^2/4 \,, \tag{D.6}$$

where $i' = i_k - i_j$. (Note that $i' > 0$ since in our construction the indices $i_1, \ldots, i_J$ are sorted in descending order.)

Indeed, (D.6) means that if we let momentarily $\|g_j\|$ denote $\sqrt{\mathbb{E}_{x\in T} g_j(x)^2}$ then

$$\|g_j\| \geqslant \|g'_j\| - \|\textstyle\sum_{k<j} g_k\| \geqslant \|g'_j\| - \sum_{k<j}\|g_k\| \geqslant \|g'_j\|(1 - \sum_{i'=1}^{\infty} 10^{-i'}) \geqslant 0.8\|g'_j\| \,. \tag{D.7}$$

The first inequality holds because we can write g_j as $g'_j - h_j$, where $h_j = \min\{g'_j, \sum_{k<j} g_k\}$. Then, on the one hand, $\|g_j\| \geqslant \|g'_j\| - \|h_j\|$, and on the other hand, $\|h_j\| \leqslant \|\sum_{k<j} g_k\|$ since $g'_j \geqslant 0$. The second inequality holds because $\|g_k\| \leqslant \|g'_k\|$. By squaring (D.7) and plugging in the value of $\|g'_j\|^2$ we get (D.5).

Proof of (D.6). By our construction, it must hold that

$$c^{10}\alpha_{i_k}/2^{i'} \leqslant \alpha_{i_j} \,, \tag{D.8}$$

since otherwise the index i_j would have been removed from the $\mathcal{I}$ at the k^{th} step. Since $\beta_{i_k} = \beta_{i_j} c^{i'}$, we can plug (D.4) in (D.8) to get

$$\mu_{i_k} c^{10+4i'}/2^{i'} \leqslant c^4 \mu_{i_j}$$

or

$$\mu_{i_k} \leqslant \mu_{i_j}(2/c)^{4i'} c^{-6} \,.$$

Since $|T_i| = e|U_i|$ for all i, it follows that $|T_k|/|T| \leqslant (2/c)^{4i'} c^{-6}$. On the other hand, we know that $\mathbb{E}_{x\in T_k} g'_k(x)^2 = \beta_{i_k}^2/4 = c^{2i'}\beta_{i_j}^2/4$. Thus,

$$\mathop{\mathbb{E}}_{x\in T} g'_k(x)^2 \leqslant 2^{4i'} c^{2i'-4i'-6}\beta_{i_j}^2/4 \leqslant (2^4/c^2)^{i'}\beta_{i_j}^2/4 \,,$$

and now we just choose c sufficiently large so that $c^2/2^4 > 100$. □

□

D.1 Norm bound implies small-set expansion

In this section, we show that an upper bound on $2 \to q$ norm of the projector to the top eigenspace of a graph implies that the graph is a small-set expander. This proof appeared elsewhere implicitly [KV05, O'D07] or explicitly [BGH+12, BBH+12] and is presented here only for completeness. Fix a graph G (identified with its normalized adjacency matrix), and $\lambda \in (0, 1)$, letting $V_{\geqslant\lambda}$ denote the subspace spanned by eigenfunctions with eigenvalue at least λ.

If p, q satisfy $1/p + 1/q = 1$ then $\|x\|_p = \max_{y:\|y\|_q\leqslant 1} |\langle x, y\rangle|$. Indeed, $|\langle x, y\rangle| \leqslant \|x\|_p\|y\|_q$ by Hölder's inequality, and by choosing $y_i = \mathrm{sign}(x_i)|x_i|^{p-1}$ and normalizing one can see this equality is tight. In particular, for every $x \in L(\mathcal{U})$, $\|x\|_q = \max_{y:\|y\|_{q/(q-1)}\leqslant 1} |\langle x, y\rangle|$ and $\|y\|_{q/(q-1)} = \max_{\|x\|_q\leqslant 1} |\langle x, y\rangle|$. As a consequence

$$\|A\|_{2\to q} = \max_{\|x\|_2\leqslant 1} \|Ax\|_q = \max_{\|x\|_2\leqslant 1, \|y\|_{q/(q-1)}\leqslant 1} |\langle Ax, y\rangle| = \max_{\|y\|_{q/(q-1)}\leqslant 1} |\langle A^T y, x\rangle| = \|A^T\|_{q/(q-1)\to 2}$$

Note that if A is a projection operator, $A = A^T$. Thus, part 1 of Theorem D.1 follows from the following lemma:

Lemma D.7. *Let $G = (V, E)$ be regular graph and $\lambda \in (0, 1)$. Then, for every $S \subseteq V$,*

$$\Phi(S) \geqslant 1 - \lambda - \|V_\lambda\|^2_{q/(q-1)\to 2}\mu(S)^{(q-2)/q}$$

Proof. Let f be the characteristic function of S, and write $f = f' + f''$ where $f' \in V_\lambda$ and $f'' = f - f'$ is the projection to the eigenvectors with value less than λ. Let $\mu = \mu(S)$. We know that

$$\Phi(S) = 1 - \langle f, Gf\rangle/\|f\|_2^2 = 1 - \langle f, Gf\rangle/\mu\,, \tag{D.9}$$

And $\|f\|_{q/(q-1)} = \left(\mathbb{E}\, f(x)^{q/(q-1)}\right)^{(q-1)/q} = \mu^{(q-1)/q}$, meaning that $\|f'\| \leqslant \|V_\lambda\|_{q/(q-1)\to 2}\mu^{(q-1)/q}$. We now write

$$\begin{aligned}\langle f, Gf\rangle = \langle f', Gf'\rangle + \langle f'', Gf''\rangle \leqslant \|f'\|_2^2 + \lambda\|f''\|_2^2 &\leqslant \|\mathcal{V}\|^2_{q/(q-1)\to 2}\|f\|^2_{q/(q-1)} + \lambda\mu \\ &\leqslant \|\mathcal{V}\|^2_{2\to q}\mu^{2(q-1)/q} + \lambda\mu\,. \end{aligned}\tag{D.10}$$

Plugging this into (D.9) yields the result. □