

Lower bounds on the size of semidefinite programming relaxations

James R. Lee* Prasad Raghavendra† David Steurer‡

Abstract

We introduce a method for proving lower bounds on the efficacy of semidefinite programming (SDP) relaxations for combinatorial problems. In particular, we show that the cut, TSP, and stable set polytopes on n -vertex graphs are not the linear image of the feasible region of any SDP (i.e., any spectrahedron) of dimension less than 2^{n^δ} , for some constant $\delta > 0$. This result yields the first super-polynomial lower bounds on the semidefinite extension complexity of any explicit family of polytopes.

Our results follow from a general technique for proving lower bounds on the positive semidefinite rank of a matrix. To this end, we establish a close connection between arbitrary SDPs and those arising from the sum-of-squares SDP hierarchy. For approximating maximum constraint satisfaction problems, we prove that SDPs of polynomial-size are equivalent in power to those arising from degree- $O(1)$ sum-of-squares relaxations. This result implies, for instance, that no family of polynomial-size SDP relaxations can achieve better than a $7/8$ -approximation for MAX 3-SAT.

Keywords: semidefinite programming, sum-of-squares method, lower bounds on positive-semidefinite rank, approximation complexity, quantum learning, polynomial optimization.

*University of Washington.

†UC Berkeley.

‡Cornell University.

Contents

1	Introduction	3
1.1	Spectrahedral lifts of polytopes	4
1.2	Semidefinite relaxations and constraint satisfaction	5
1.3	Positive semidefinite rank and sum-of-squares degree	8
2	Proof overview and setup	11
2.1	Preliminaries	11
2.2	Factorizations, quantum learning, and pseudo-densities	12
3	PSD rank and sum-of-squares degree	17
3.1	Analysis of the separating functional	17
3.2	Degree reduction	21
3.3	Proof of the main theorem	23
4	Approximations for density operators	24
4.1	Approximation against a single test	24
4.2	Approximation against a family of tests	26
4.2.1	Junta approximation	30
5	The correlation polytope	31
5.1	Positive semidefinite rank	31
6	Optimality of low-degree sum-of-squares for max CSPs	34
6.1	The SDP approximation model	34
6.2	General SDPs vs. sum-of-squares	37
7	Nonnegative rank	38
7.1	Nonnegative rank vs. junta degree	38
7.2	The correlation polytope and lopsided disjointness	41
7.3	Unique games hardness for LPs	42
	References	43

1 Introduction

Convex characterizations and relaxations of combinatorial problems have been a consistent, powerful theme in the theory of algorithms since its inception. Linear and semidefinite programming relaxations have been particularly useful for the efficient computation of approximate solutions to NP-hard problems (see, for instance, the books [WS11, Vaz01]). In some sense, semidefinite programs (SDPs) can be seen as combining the rich expressiveness of linear programs with the global geometric power of spectral methods. For many fundamental combinatorial problems, this provides a genuinely new structural and computational perspective [GW95, KMS98, ARV09]. Indeed, for an array of optimization problems, the best-known approximation algorithms can only be achieved via SDP relaxations.

It has long been known that integrality gaps for linear programs (LPs) can often lead to gadgets for NP-hardness of approximation reductions (see, e.g., [LY93, CGH⁺05, HK03]). Furthermore, assuming the Unique Games Conjecture [Kho02], it is known that integrality gaps for SDPs can be translated directly into hardness of approximation results [KKMO04, Aus10, Rag08]. All of this suggests that the computational model underlying LPs and SDPs is remarkably powerful.

Thus it is a natural (albeit ambitious) goal to characterize the computational power of this model. If $P \neq NP$, we do not expect to find polynomial-size families of SDPs that yield arbitrarily good approximations to NP-hard problems. (See [Rot13, BDP13] for a discussion of how this follows formally from the assumption $NP \not\subseteq P/\text{poly}$.)

In the setting of linear programs (LPs), the search for a model and characterization began in a remarkable work of Yannakakis [Yan91]. He proved that the TSP and matching polytopes do not admit *symmetric* linear programming formulations of size $2^{o(n)}$, where n is the number of vertices in the underlying graph. In the process, he laid the structural framework (in terms of nonnegative factorizations) that would underlie all future work in the subject. It took over 20 years before Fiorini, Massar, Pokutta, Tiwary, and de Wolf [FMP⁺12] were able to remove the symmetry assumption and obtain a lower bound of $2^{\Omega(\sqrt{n})}$ on the size of any LP formulation. Soon afterward, Rothvoß [Rot14] gave a lower bound of $2^{\Omega(n)}$ on the size of any LP formulation for the matching polytope (and also TSP), completing Yannakakis' vision.

Despite the progress in understanding the power of LP formulations, it remained a mystery whether there were similar strong lower bounds in the setting of SDPs. An analogous positive semidefinite factorization framework was provided in [FMP⁺12, GPT11]. Following the LP methods of [CLRS13], the papers [LRST14, FSP13] proved exponential lower bounds on the size of *symmetric* SDP formulations for NP-hard constraint satisfaction problems (CSPs).

In the present work, we prove strong lower bounds on the size of general SDP formulations for the cut, TSP, and stable set polytopes. Moreover, we show that polynomial-size SDP relaxations cannot achieve arbitrarily good approximations for many NP-hard constraint satisfaction problems. For instance, no polynomial-size family of relaxations can achieve better than a 7/8-approximation for MAX 3-SAT. More generally, we show that the low-degree sum-of-squares SDP relaxations yield the best approximation among all polynomial-sized families of relaxations for max-CSPs.

This is achieved by relating arbitrary SDP formulations to those coming from the sum-of-squares SDP hierarchy¹ [Las01, Par00, Sho87], analogous to our previous work with Chan relating LP formulations to the Sherali–Adams hierarchy [CLRS13]. The SDP setting poses a number of significant challenges. At a very high level, our approach can be summarized as follows: Given an arbitrary SDP formulation of small size, we use methods from quantum entropy maximization and

¹This hierarchy is also frequently referred to as the Lasserre SDP hierarchy.

online convex optimization (often going by the name “matrix multiplicative weights update”) to *learn* an approximate low-degree sum-of-squares formulation on a subset of the input variables. In the next section, we present a formal overview of our results, and a discussion of the connections to quantum information theory, real algebraic geometry, and proof complexity.

Organization. The results of this work fall along two broad themes, lower bounds on spectrahedral lifts of specific polytopes and lower bounds on SDP relaxations for constraint satisfaction problems. Both sets of results are consequences of a general method for proving lower bounds on positive semidefinite rank. For the convenience of the reader, we have organized the two themes in two self-contained trajectories. Thus, lower bounds on spectrahedral lifts can be accessed via [Section 1.1](#), [Section 1.3](#), [Section 2](#), [Section 3](#) and [Section 5](#). The lower bounds for constraint satisfaction problems can be reached through [Section 1.2](#), [Section 1.3](#), [Section 2](#), [Section 3](#) and [Section 6](#).

We also present general results on approximating density operators against families of linear tests through quantum learning in [Section 4](#). Finally, in [Section 7](#), we exhibit applications of our techniques to non-negative rank; in particular, this is used to give a simple, self-contained proof of a lower bound on the nonnegative rank of the unique disjointness matrix.

1.1 Spectrahedral lifts of polytopes

Polytopes are an appealing and useful way to encode many combinatorial optimization problems. For example, the traveling salesman problem on n cities is equivalent to optimizing linear functions over the traveling salesman polytope, i.e., the convex hull of characteristic vectors $\mathbf{1}_C \in \{0, 1\}^{\binom{n}{2}} \subseteq \mathbb{R}^{\binom{n}{2}}$ of n -vertex Hamiltonian cycles C (viewed as edge sets). If a polytope admits polynomial-size LP or SDP formulations, then we can optimize linear functions over the polytope in polynomial time (exactly for LP formulations and up to arbitrary accuracy in the case of SDP formulations). Indeed, a large number of efficient, exact algorithms for combinatorial optimization problems can be explained by small LP or SDP formulations of the underlying polytope. (For approximation algorithms, the characterization in terms of compact formulations of polytopes is not as direct [[BFPS12](#)]. In [Section 1.2](#), we will give a direct characterization for approximation algorithms in terms of the original combinatorial problem.)

Positive semidefinite lifts. Fix a polytope $P \subseteq \mathbb{R}^n$ (e.g., the traveling salesman polytope described above). We are interested in the question of whether there exists a low-dimensional SDP that captures P . Let \mathcal{S}_+^k denote the cone of symmetric, $k \times k$ positive semidefinite matrices embedded naturally in $\mathbb{R}^{k \times k}$. If there exists an affine subspace $\mathcal{L} \subseteq \mathbb{R}^{k \times k}$ and a linear map $\pi : \mathbb{R}^{k \times k} \rightarrow \mathbb{R}^n$ such that

$$P = \pi(\mathcal{S}_+^k \cap \mathcal{L}),$$

one says that P admits a *positive-semidefinite (psd) lift of size k* . (This terminology is taken from [[FGP⁺14](#)].) We remark that the intersection of a PSD cone with an affine subspace is often referred to as a *spectrahedron*.

The point is that in order to optimize a linear function $\ell : \mathbb{R}^n \rightarrow \mathbb{R}$ over the polytope P , it is enough to optimize the linear function $\ell \circ \pi : \mathbb{R}^{k \times k} \rightarrow \mathbb{R}$ over the set $\mathcal{S}_+^k \cap \mathcal{L}$ instead,

$$\min_{x \in P} \ell(x) = \min_{y \in \mathcal{S}_+^k \cap \mathcal{L}} \ell \circ \pi(y).$$

Here, the optimization problem on the right is a semidefinite programming problem in k -by- k matrices. This idea also goes under the name of a *semidefinite extended formulation* [[FMP⁺12](#)].

The positive-semidefinite rank of explicit polytopes. We define the *positive-semidefinite (psd) rank* of a polytope P , denoted $\text{rk}_{\text{psd}}(P)$, to be the smallest number k such that there exists a psd lift of size k . (Our use of the word “rank” will make sense soon—see [Section 1.3](#).) Briët, Dadush, and Pokutta [BDP13] showed (via a counting argument) that there exist 0/1 polytopes in \mathbb{R}^n with exponential psd rank. In this work, we prove the first super-polynomial lower bounds on the psd rank of explicit polytopes: The correlation polytope $\text{CORR}_n \subseteq \mathbb{R}^{n^2}$ is given by

$$\text{CORR}_n = \text{conv}(\{xx^T : x \in \{0, 1\}^n\}).$$

In [Section 5.1](#), we show the following strong lower bound on its psd rank.

Theorem 1.1. *For every $n \geq 1$, we have*

$$\text{rk}_{\text{psd}}(\text{CORR}_n) \geq 2^{\Omega(n^{2/13})}.$$

The importance of the correlation polytope CORR_n lies in the fact that a number of interesting polytopes from combinatorial optimization contain a face that linearly projects to CORR_n . We first define a few different families of polytopes and then recall their relation to CORR_n .

For $n \geq 1$, let $K_n = ([n], \binom{[n]}{2})$ be the complete graph on n vertices. For a set $S \subseteq [n]$, we use $\partial S \subseteq \binom{[n]}{2}$ to denote the set of edges with one endpoint in S and the other in \bar{S} , and we use the notation $\mathbf{1}_{\partial S} \in \mathbb{R}^{\binom{[n]}{2}}$ to denote the characteristic vector of S . The *cut polytope on n vertices* is defined by

$$\text{CUT}_n = \text{conv}(\{\mathbf{1}_{\partial S} : S \subseteq [n]\}).$$

Similarly, if τ is a salesman tour of K_n (i.e., a Hamiltonian cycle), we use $\mathbf{1}_{E(\tau)} \in \mathbb{R}^{\binom{[n]}{2}}$ to denote the corresponding indicator of the edges contained in τ . In that case, the *TSP polytope* is given by

$$\text{TSP}_n = \text{conv}(\{\mathbf{1}_{E(\tau)} : \tau \text{ is a Hamiltonian cycle}\}).$$

Finally, consider an arbitrary n -vertex graph $G = ([n], E)$. We recall that a subset of vertices $S \subseteq [n]$ is an *independent set* (also called a stable set) if there are no edges between vertices in S . The *stable set polytope of G* is given by

$$\text{STAB}_n(G) = \text{conv}(\{\mathbf{1}_S \in \mathbb{R}^n : S \text{ is an independent set in } G\}).$$

By results of [DS90] and [FMP⁺12] (see [Proposition 5.2](#)), [Theorem 1.1](#) directly implies the following lower bounds on the psd rank of the cut, TSP, and stable set polytopes.

Corollary 1.2. *The following lower bounds hold for every $n \geq 1$,*

$$\begin{aligned} \text{rk}_{\text{psd}}(\text{CUT}_n) &\geq 2^{\Omega(n^{2/13})}, \\ \text{rk}_{\text{psd}}(\text{TSP}_n) &\geq 2^{\Omega(n^{1/13})}, \\ \max_{n\text{-vertex } G} \text{rk}_{\text{psd}}(\text{STAB}_n(G)) &\geq 2^{\Omega(n^{1/13})}. \end{aligned}$$

1.2 Semidefinite relaxations and constraint satisfaction

We now formalize a computational model of semidefinite relaxations for combinatorial optimization problems and prove strong lower bounds for it. Unlike the polytope setting in the previous section, this model also allows us to capture approximation algorithms directly.

Consider the following general optimization problem:² Given a low-degree function $f: \{0, 1\}^n \rightarrow \mathbb{R}$, represented by its coefficients as a multilinear polynomial,

$$\begin{aligned} & \text{maximize } f(x) \\ & \text{subject to } x \in \{0, 1\}^n. \end{aligned} \tag{1.1}$$

Many basic optimization problems are special cases of this general problem, corresponding to functions f of a particular form: For the problem of finding the maximum cut in a graph G with n vertices, the function f outputs on input $x \in \{0, 1\}^n$ the number of edges in G that cross the bipartition represented by x , i.e., $f(x)$ is the number of edges $\{i, j\} \in E(G)$ with $x_i \neq x_j$. Similarly, for MAX 3-SAT on a 3CNF formula φ with n variables, $f(x)$ is the number of clauses in φ satisfied by the assignment x . More generally, for any k -ary boolean constraint satisfaction problem, the function f counts the number of satisfied constraints. Note that in these examples, the functions have at most degree 2, degree 3, and degree k , respectively.

Algorithms with provable guarantees for these kinds of problems—either implicitly or explicitly—certify upper bounds on the optimal value of instances. (Indeed, for solving the decision version of these optimization problems, it is enough to provide such certificates.) It turns out that the best-known algorithms for these problems are captured by certificates of a particularly simple form, namely sums of squares of low-degree polynomials. The following upper bounds on problems of the form (1.1) are equivalent to the relaxations obtained by the sum-of-squares SDP hierarchy [Las01, Par00, Sho87]. For $f: \{0, 1\}^n \rightarrow \mathbb{R}$, we use $\deg(f)$ to denote the degree of the unique multilinear real polynomial agreeing with f on $\{0, 1\}^n$; see Section 2.1.

Definition 1.3. The *degree- d sum-of-squares upper bound* for a function $f: \{0, 1\}^n \rightarrow \mathbb{R}$, denoted $\overline{\text{sos}}_d(f)$, is the smallest number $c \in \mathbb{R}$ such that $c - f$ is a sum of squares of functions of degree at most $d/2$, i.e., there exists functions $g_1, \dots, g_t: \{0, 1\}^n \rightarrow \mathbb{R}$ for some $t \in \mathbb{N}$ with $\deg(g_1), \dots, \deg(g_t) \leq d/2$ such that the following identity between functions on the discrete cube holds:

$$c - f = g_1^2 + \dots + g_t^2.$$

Every function f satisfies $\overline{\text{sos}}_d(f) \geq \max(f)$ since sums of squares of real-valued functions are nonnegative pointwise. For $d \geq 1$, the problem of computing $\overline{\text{sos}}_d(f)$ for a given function $f: \{0, 1\}^n \rightarrow \mathbb{R}$ (of degree at most d) is a semidefinite program of size at most $1 + n^{d/2}$ (see, e.g., Theorem 3.8).³

The $\overline{\text{sos}}_d$ upper bound is equivalent to the degree- d sum-of-squares (also known as the level- $d/2$ Lasserre) SDP bound, and for small values of d , these upper bounds underlie the best-known approximation algorithms for several optimization problems. For example, the Goemans–Williamson algorithm for MAX CUT is based on the upper bound $\overline{\text{sos}}_2$. If we let $\alpha_{GW} \approx 0.878$ be the approximation ratio of this algorithm, then every graph G satisfies $\max(f_G) \geq \alpha_{GW} \cdot \overline{\text{sos}}_2(f_G)$ where the function f_G measures cuts in G , i.e., $f_G(x) := \sum_{ij \in E(G)} (x_i - x_j)^2$.

A natural generalization of low-degree sum-of-squares certificates is obtained by summing squares of functions in a low-dimensional subspace. We can formulate this generalization as a *non-uniform* model of computation that captures general semidefinite programming relaxations. First, we make the following definition for a subspace of functions.

²In this section, we restrict our discussion to optimization problems over the discrete cube. Some of our results also apply to other problems, e.g., the traveling salesman problem (albeit only for exact algorithms).

³Moreover, for every $d \in \mathbb{N}$, there exists an $n^{O(d)}$ -time algorithm based on the ellipsoid method that, given f, c , and $\varepsilon > 0$, distinguishes between the cases $\overline{\text{sos}}_d(f) \geq c$ and $\overline{\text{sos}}_d(f) \leq c - \varepsilon$ (assuming the binary encoding of f, c , and ε is bounded by $n^{O(d)}$).

Definition 1.4. For a subspace U of real-valued functions on $\{0, 1\}^n$, the *subspace- U sum-of-squares upper bound* for a function $f: \{0, 1\}^n \rightarrow \mathbb{R}$, denoted $\overline{\text{sos}}_U(f)$, is the smallest number $c \in \mathbb{R}$ such that $c - f$ is a sum of squares of functions from U , i.e., there exist $g_1, \dots, g_t \in U$ such that $c - f = g_1^2 + \dots + g_t^2$ is an identity of functions on $\{0, 1\}^n$.

Here, the subspace U can be thought of as “non-uniform advice” to an algorithm, where its dimension $\dim(U)$ is the size of advice. In fact, if we fix this advice U , the problem of computing $\overline{\text{sos}}_U(f)$ for a given function f has a semidefinite programming formulation of size $\dim(U)$.⁴ Moreover, it turns out that the generalization captures, in a certain precise sense, all possible semidefinite programming relaxations for (1.1). The dimension of the subspace corresponds to the size of the SDP. See Section 6.1 for a detailed discussion of the model.

In this work, we exhibit unconditional lower bounds in this powerful non-uniform model of computation. For example, we show that the MAX 3-SAT problem cannot be approximated to a factor better than $7/8$ using a polynomial-size family of SDP relaxations. Formally, we show the following lower bound for MAX 3-SAT.

Theorem 1.5. *For every $s > 7/8$, there exists a constant $\alpha > 0$ such that for every $n \in \mathbb{N}$ and every linear subspace U of functions $f: \{0, 1\}^n \rightarrow \mathbb{R}$ with*

$$\dim U \leq n^{\alpha \frac{\log n}{\log \log n}},$$

there exists a MAX 3-SAT instance \mathfrak{J} on n variables such that $\max(\mathfrak{J}) \leq s$ but $\overline{\text{sos}}_U(\mathfrak{J}) = 1$ (i.e., U fails to achieve a factor- s approximation for MAX 3-SAT).

Our main result is a characterization of an optimal semidefinite programming relaxation for the class of constraint satisfaction problems among all families of SDP relaxations of similar size. Roughly speaking, we show that the $O(1)$ -degree sum-of-squares relaxations are *optimal* among all polynomial-sized SDP relaxations for constraint satisfaction problems. Towards stating our main result, we define the class of constraint satisfaction problems. For the sake of clarity, we restrict ourselves to boolean constraint satisfaction problems although the results hold in greater generality.

For a finite collection \mathcal{P} of k -ary predicates $P: \{0, 1\}^k \rightarrow \{0, 1\}$, we let $\text{MAX-}\mathcal{P}$ denote the following optimization problem: An instance \mathfrak{J} consists of boolean variables X_1, \dots, X_n and a collection of \mathcal{P} -constraints $P_1(X) = 1, \dots, P_M(X) = 1$ over these variables. A \mathcal{P} -constraint is a predicate $P_0: \{0, 1\}^n \rightarrow \{0, 1\}$ such that $P_0(X) = P(X_{i_1}, \dots, X_{i_k})$ for some $P \in \mathcal{P}$ and *distinct* indices $i_1, \dots, i_k \in [n]$. The objective is to find an assignment $x \in \{0, 1\}^n$ that satisfies as many of the constraints as possible, that is, which maximizes

$$\mathfrak{J}(x) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{i=1}^M P_i(x).$$

We denote the optimal value of an assignment for \mathfrak{J} as $\text{opt}(\mathfrak{J}) = \max_{x \in \{0, 1\}^n} \mathfrak{J}(x)$. For example, MAX CUT corresponds to the case where \mathcal{P} consists of the binary inequality predicate. For MAX 3-SAT, \mathcal{P} contains all eight 3-literal disjunctions, e.g., $X_1 \vee \bar{X}_2 \vee \bar{X}_3$.

⁴Under mild conditions on the subspace U , there exists a boolean circuit of size $(\dim U)^{O(1)}$ that given a constant-degree function f , and number $c \in \mathbb{R}$ and $\varepsilon > 0$, distinguishes between the cases $\overline{\text{sos}}_U(f) \geq c$ and $\overline{\text{sos}}_U(f) \leq c - \varepsilon$ (assuming the bit encoding length of f, c , and ε is bounded by $(\dim U)^{O(1)}$). Note that since we will prove lower bounds against this model, the possibility that some subspaces might not correspond to small circuits does not weaken our results.

Next, we discuss how to compare the quality of upper bound certificates of the form $\overline{\text{sos}}_U$. Let Π be a boolean CSP and let Π_n be the restriction of Π to instances with n boolean variables. As discussed before, the problem Π_n could for example be MAX CUT on graphs with n vertices or MAX 3-SAT on formulas with n variables. We say that a subspace $U \subseteq \mathbb{R}^{\{0,1\}^n}$ achieves a (c, s) -approximation for Π_n if every instance $\mathfrak{J} \in \Pi_n$ satisfies

$$\max(\mathfrak{J}) \leq s \implies \overline{\text{sos}}_U(\mathfrak{J}) \leq c. \quad (1.2)$$

In other words, the upper bound $\overline{\text{sos}}_U$ allows us to distinguish⁵ between the cases $\max(\mathfrak{J}) \leq s$ and $\max(\mathfrak{J}) > c$ for all instances $\mathfrak{J} \in \Pi_n$.

We prove the following theorem, which shows that for every boolean CSP, the approximation guarantees obtained by the degree- d sum-of-squares upper bound (also known as the level- $d/2$ Lasserre SDP relaxation) are optimal among all semidefinite programming relaxations of size at most n^{cd} for some universal constant $c > 0$.

Theorem 1.6. *Let Π be boolean constraint satisfaction problem and let Π_n be the set of instances of Π on n variables. Suppose that for some $m, d \in \mathbb{N}$, the subspace of degree- d functions $f: \{0, 1\}^m \rightarrow \mathbb{R}$ fails to achieve a (c, s) -approximation for Π_m (in the sense of (1.2)). Then there exists a number $\alpha = \alpha(\Pi_m, c, s) > 0$ such that for all $n \in \mathbb{N}$, every subspace U of functions $f: \{0, 1\}^n \rightarrow \mathbb{R}$ with $\dim U \leq \alpha \cdot (n/\log n)^{d/4}$ fails to achieve a (c, s) -approximation for Π_n .*

The theorem has several immediate concrete consequences for specific boolean CSPs. First, we know that $O(1)$ -degree sos upper bounds do not achieve an approximation ratio better than $7/8$ for MAX 3-SAT [Gri01b, Sch08], therefore Theorem 1.6 implies that polynomial-size SDP relaxations for MAX 3-SAT cannot achieve an approximation ratio better than $7/8$. In fact, a quantitatively stronger version of the above theorem yields Theorem 1.5.

Another concrete consequence of this theorem is that if there exists a polynomial-size family of semidefinite programming relaxations for MAX CUT that achieves an approximation ratio better than α_{GW} , then also a $O(1)$ -degree sum-of-squares upper bound achieves such a ratio. This assertion is especially significant in light of the notorious Unique Games Conjecture one of whose implications is that it is NP-hard to approximate MAX CUT to a ratio strictly better than α_{GW} .

1.3 Positive semidefinite rank and sum-of-squares degree

In order to prove our results on spectrahedral lifts and semidefinite relaxations, the factorization perspective will be essential. In the LP setting, the characterization of polyhedral lifts and LP relaxations in terms of nonnegative factorizations is a significant contribution of Yannakakis [Yan91]. In the SDP setting, the analogous characterization is in terms of positive semidefinite factorizations [FMP⁺12, GPT11].

Definition 1.7 (PSD rank). Let $M \in \mathbb{R}_+^{p \times q}$ be a matrix with non-negative entries. We say that M admits a *rank- r psd factorization* if there exist positive semidefinite matrices $\{A_i : i \in [p]\}, \{B_j : j \in [q]\} \subseteq \mathcal{S}_+^r$ such that $M_{i,j} = \text{Tr}(A_i B_j)$ for all $i \in [p], j \in [q]$. We define $\text{rk}_{\text{psd}}(M)$ to be the smallest r such that M admits a rank- r psd factorization. We refer to this value as the *psd rank* of M .

⁵In order to distinguish between the cases $\max(\mathfrak{J}) \leq s$ and $\max(\mathfrak{J}) > c$ it is enough to check whether \mathfrak{J} satisfies $\overline{\text{sos}}_U(\mathfrak{J}) \leq c$. In the case $\max(\mathfrak{J}) \leq s$, we know that $\overline{\text{sos}}_U(\mathfrak{J}) \leq c$ by (1.2). On the other hand, in the case $\max(\mathfrak{J}) > c$, we know that $\overline{\text{sos}}_U(\mathfrak{J}) > c$ because $\overline{\text{sos}}_U(\mathfrak{J})$ is always an upper bound on $\max(\mathfrak{J})$.

Nonnegative factorizations correspond to the special case that the matrices $\{A_i\}$ and $\{B_j\}$ are restricted to be diagonal. A rank- r nonnegative factorization can equivalently be viewed as a sum of r rank-1 nonnegative factorizations (nonnegative rectangles). Indeed, this viewpoint is crucial for all lower bounds on nonnegative factorization. In contrast, rank- r psd factorizations do not seem to admit a good characterization in terms of rank-1 psd factorizations. This difference captures one of the main difficulties of proving psd rank lower bounds.

Main theorem. Consider a nonnegative function $f: \{0, 1\}^n \rightarrow \mathbb{R}_+$ on the n -dimensional discrete cube. We say that f has a *sum-of-squares (sos) certificate of degree d* if there exist functions $g_1, \dots, g_k: \{0, 1\}^n \rightarrow \mathbb{R}$ such that $\deg(g_1), \dots, \deg(g_k) \leq d/2$, and $f(x) = \sum_{i=1}^k g_i(x)^2$ for all $x \in \{0, 1\}^n$. (Here, the $\deg(g)$ denotes the degree of g as a multilinear polynomial. We refer to [Section 2.1](#) for the precise definition.) We then define the *sos degree of f* , denoted $\deg_{\text{sos}}(f)$, to be the minimal d such that f has a degree- d sos certificate.

This notion is closely related⁶ to (a special case of) the Positivstellensatz proof system of Grigoriev and Vorobjov [GV02]. We refer to the surveys [Lau09, BS14] and the introduction of [OZ12] for a review of such proof systems and their relationship to semidefinite programming.

With this notion in place, we can now present a representative theorem that embodies our approach. For a point $x \in \mathbb{R}^n$ and a subset $S \subseteq [n]$, we denote by $x_S \in \mathbb{R}^{|S|}$ the vector $x_S = (x_{i_1}, x_{i_2}, \dots, x_{i_{|S|}})$ where $S = \{i_1, i_2, \dots, i_{|S|}\}$ and $i_1 < i_2 < \dots < i_{|S|}$. For a function $f: \{0, 1\}^m \rightarrow \mathbb{R}_+$ and a number $n \geq m$, we define the following central object: The $\binom{n}{m} \times 2^m$ -dimensional real matrix M_n^f is given by

$$M_n^f(S, x) \stackrel{\text{def}}{=} f(x_S), \quad (1.3)$$

where $S \subseteq [n]$ runs over all subsets of size m and $x \in \{0, 1\}^m$.

Theorem 1.8 (Sum-of-squares degree vs. psd rank). *For every $m, d \geq 1$ and $f: \{0, 1\}^m \rightarrow \mathbb{R}_+$ with $\deg_{\text{sos}}(f) = d + 2$, there exists a constant $C > 0$ such that the following holds for all $n \geq 2m$,*

$$1 + n^{1+d/2} \geq \text{rk}_{\text{psd}}(M_n^f) \geq C \left(\frac{n}{\log n} \right)^{d/4}.$$

Remark 1.9. The reader might observe that the matrix in (1.3) looks very similar to the “pattern matrices” defined by Sherstov [She11]. This comparison is not unfounded; some high-level aspects of our proof are quite similar to his. Random restrictions are a powerful tool for analyzing functions over the discrete cube. We refer to [OD14, Ch. 4] for a discussion of their utility in the context of discrete Fourier analysis. They were also an important tool in the work [CLRS13] on lower bounds for LPs. Accordingly, one hopes that our methods may have additional applications in communication complexity. This would not be surprising, as there is a model of quantum communication that exactly captures psd rank (see [FMP⁺12]).

Connection to spectrahedral lifts of polytopes. The connection to psd lifts proceeds as follows. Let $\{x_1, x_2, \dots, x_v\} \subseteq P$ be such that $P = \text{conv}(V)$ is the convex hull of V , and also fix a representation

$$P = \{x \in \mathbb{R}^n : \langle a_i, x \rangle \leq b_i \ \forall i \in [m]\}.$$

The *slack matrix* S associated to P (and our chosen representation) is the matrix $S \in \mathbb{R}_+^{m \times v}$ given by $S_{i,j} = b_i - \langle a_i, x_j \rangle$. It is not difficult to see that $\text{rk}_{\text{psd}}(S)$ does not depend on the choice of representation. It turns out that the psd rank of S is precisely the minimum size of a psd lift of P .

⁶For the sake of simplicity, we have only defined this notion for functions on the discrete cube. In more general settings, one has to be a bit more careful; we refer to [GV02].

Proposition 1.10 ([FMP⁺12, GPT11]). *For every $n, k \geq 1$, every polytope $P \subseteq \mathbb{R}^n$ and every slack matrix S associated to P , it holds that $\text{rk}_{\text{psd}}(S) \leq k$ if and only if P admits a psd lift of size k .*

Thus our goal in this paper becomes one of proving lower bounds on psd rank. With this notation, we have a precise way to characterize the lack of previous progress: Before this work, there was no reasonable method available to prove lower bounds on the psd rank of explicit matrices. The characterization of [Proposition 1.10](#) explains our abuse of notation in [Theorem 1.1](#), writing $\text{rk}_{\text{psd}}(P)$ to denote the psd rank of any slack matrix associated to a polytope P .

[Theorem 1.8](#) is already enough to show that $\text{rk}_{\text{psd}}(\text{CORR}_n)$ must grow faster than any polynomial in n , as we will argue momentarily. In [Section 3.1](#), we present a more refined version (using a more robust version of sos degree) that will allow us to achieve a lower bound of the form $\text{rk}_{\text{psd}}(\text{CORR}_n) \geq 2^{\Omega(n^\delta)}$ for some $\delta > 0$.

Given [Theorem 1.8](#), in order to prove a lower bound on $\text{rk}_{\text{psd}}(\text{CORR}_n)$, we should find, for every $d \geq 1$, a number m and a function $f : \{0, 1\}^m \rightarrow \mathbb{R}_+$ such that $\deg_{\text{sos}}(f) \geq d$ and such that M_n^f is a submatrix of some slack matrix associated to CORR_n . To this end, it helps to observe the following (we recall the proof in [Section 5](#)).

Proposition 1.11. *If $f : \{0, 1\}^m \rightarrow \mathbb{R}_+$ is a nonnegative quadratic function over $\{0, 1\}^m$, then for any $n \geq m$, the matrix M_n^f is a submatrix of some slack matrix associated to CORR_n .*

Given the preceding proposition, the following result of Grigoriev on the Knapsack tautologies completes our quest for a lower bound.

Theorem 1.12 ([Gri01a]). *For every odd integer $m \geq 1$, the function $f : \{0, 1\}^m \rightarrow \mathbb{R}_+$ given by*

$$f(x) = \left(\frac{m}{2} - \sum_{i=1}^m x_i \right)^2 - \frac{1}{4} \quad (1.4)$$

has $\deg_{\text{sos}}(f) \geq m + 1$.

Note that since $m/2$ is not an integer, (1.4) is nonnegative for all $x \in \{0, 1\}^m$. It turns out that in order to prove stronger lower bounds for CORR_n , we will require a lower bound on the *approximate* sos degree of f . Thus the Knapsack tautologies (1.4) will be studied carefully in [Section 5.1](#). In [Section 2.2](#), we discuss the proof of [Theorem 1.8](#) in some detail. Then in [Section 3](#), we present a quantitatively stronger theorem and its proof.

Connection to semidefinite relaxations and constraint satisfaction. Fix now numbers $k, n \geq 1$ and a boolean CSP Π . Fix a pair of constants $0 \leq s \leq c \leq 1$. Suppose our goal is to show a lower bound on the size of SDP relaxations that yield a (c, s) -approximation on instances of size n . It turns out that this task reduces to proving a lower bound on the positive semidefinite rank of an explicit matrix M indexed by problem instances and problem solutions (points on the discrete cube in our case).

Proposition 1.13. *For any boolean CSP Π_n and any constants $0 \leq s < c \leq 1$, let U be a subspace of minimal dimension that achieves a (c, s) -approximation for Π_n . Denote the set of instances*

$$\Pi_n^{\leq s} = \{ \mathfrak{J} \mid \max(\mathfrak{J}) \leq s \}.$$

Let $M : \Pi_n^{\leq s} \times \{0, 1\}^n \rightarrow \mathbb{R}$ denote the matrix

$$M(\mathfrak{J}, x) = c - \mathfrak{J}(x).$$

Then, $\text{rk}_{\text{psd}}(M)^2 \geq \dim(U) \geq \text{rk}_{\text{psd}}(M)$.

Before describing the proof of this proposition, observe that together with our main theorem the proposition implies [Theorem 1.6](#) (optimality of degree- d sum-of-squares for approximating boolean CSPs): If \mathfrak{J}_0 is a $\text{MAX-}\mathcal{P}$ instance on m variables with $\max(\mathfrak{J}_0) \leq s$ and $\overline{\text{sos}}_d(\mathfrak{J}_0) > c$, then $f = c - \mathfrak{J}_0$ has sos degree larger than d . Our main theorem gives a lower bound on the psd rank of the matrix M_n^f . Since this matrix is a submatrix of the matrix in [Proposition 1.13](#), our psd rank lower bound implies a lower bound on the minimum dimension of a subspace achieving a (c, s) -approximation for $\text{MAX-}\mathcal{P}_n$.

Proof of [Proposition 1.13](#). Set $r = \dim(U)$. Fix a basis $q_1, \dots, q_r : \{0, 1\}^n \rightarrow \mathbb{R}$ for the subspace U . Define the function $Q : \{0, 1\}^n \rightarrow \mathcal{S}_r^+$ by setting $(Q(x))_{ij} := q_i(x)q_j(x)$ for all $i, j \in [r]$. Notice that for any $q \in U$, we can write $q = \sum_{i=1}^r \lambda_i q_i$ and thus $q(x)^2 = \text{Tr}(\Lambda Q(x))$ where $\Lambda \in \mathcal{S}_r^+$ is defined by $\Lambda_{ij} := \lambda_i \lambda_j$.

Since U achieves a (c, s) -approximation for Π_n , for every instance $\mathfrak{J} \in \Pi_n^{\leq s}$ we will have $\overline{\text{sos}}_U(\mathfrak{J}) \leq c$. By definition of $\overline{\text{sos}}_U(\mathfrak{J})$, this implies that $c - \mathfrak{J} = \sum_i g_i^2$ for $g_i \in U$. By expressing each g_i^2 as $g_i^2 = \text{Tr}(\Lambda_i Q(x))$ for some $\Lambda_i \in \mathcal{S}_r^+$ we get,

$$M(\mathfrak{J}, x) = c - \mathfrak{J}(x) = \sum_i \text{Tr}(\Lambda_i Q(x)) = \text{Tr}(\Lambda_{\mathfrak{J}} Q(x)).$$

This yields an explicit psd factorization of M certifying that $\text{rk}_{\text{psd}}(M) \leq \dim(U)$.

Conversely, by definition of $\text{rk}_{\text{psd}}(M)$ there exists positive semidefinite matrices $\{\Lambda_{\mathfrak{J}} : \mathfrak{J} \in \Pi_n^{\leq s}\}$, $\{Q(x) : x \in \{0, 1\}^n\} \subseteq \mathcal{S}_r^+$ such that $M(\mathfrak{J}, x) = \text{Tr}(\Lambda_{\mathfrak{J}} Q(x))$. Denote by $R(x) := Q(x)^{1/2}$ the positive semidefinite square root, and consider the subspace $\tilde{U} := \text{span}\{(R(x))_{ij}\} \subseteq \mathbb{R}^{\{0,1\}^n}$. Clearly, the dimension of \tilde{U} is at most $\text{rk}_{\text{psd}}(M)^2$. Further, for each instance $\mathfrak{J} \in \Pi_n^{\leq s}$, we can write

$$c - \mathfrak{J} = M_{\mathfrak{J}, x} = \text{Tr}(\Lambda_{\mathfrak{J}} Q(x)) = \text{Tr}(\Lambda_{\mathfrak{J}} R(x)^2) = \left\| \sqrt{\Lambda_{\mathfrak{J}}} R(x) \right\|_F^2.$$

Observe that $\left\| \sqrt{\Lambda_{\mathfrak{J}}} R(x) \right\|_F^2$ is a sum of squares of functions from the subspace \tilde{U} .⁷ Therefore we have $\overline{\text{sos}}_{\tilde{U}}(\mathfrak{J}) \leq c$, showing that $\overline{\text{sos}}_{\tilde{U}}$ yields a (c, s) -approximation to Π_n . Since U is the minimal subspace yielding a (c, s) -approximation, we have $\dim(U) \leq \dim(\tilde{U}) \leq \text{rk}_{\text{psd}}(M)^2$. \square

2 Proof overview and setup

2.1 Preliminaries

We write $[n] \stackrel{\text{def}}{=} \{1, 2, \dots, n\}$ for $n \in \mathbb{N}$. We will often use the notation \mathbb{E}_x to denote a uniform averaging operator where x assumes values over a finite set. For instance, if $x \in \{-1, 1\}^n$, then $\mathbb{E}_x f(x) = 2^{-n} \sum_{x \in \{-1, 1\}^n} f(x)$. The domain of the operator should always be readily apparent from context. We also use asymptotic notation: For two expressions A and B , we write $A \leq O(B)$ if there exists a universal constant C such that $A \leq C \cdot B$. We also sometimes write $A \lesssim B$ to denote $A \leq O(B)$. The notation $A \geq \Omega(B)$ similarly denotes $B \lesssim A$, and the notations $A = \Theta(B)$ and $A \asymp B$ are both used to denote the conjunction of $A \lesssim B$ and $B \lesssim A$. For a real number $x > 0$, we use $\log x$ to denote the natural logarithm of x .

Inner product spaces and norms. Let H denote a finite-dimensional vector space over \mathbb{R} equipped with an inner product $\langle \cdot, \cdot \rangle$ and the induced Euclidean norm $|\cdot|$. All vector spaces we consider here will be of this kind. We use $\mathcal{M}(H)$ to denote the set of self-adjoint linear operators on H , and

⁷Here, $\|\cdot\|_F$ denotes the Frobenius norm.

$\mathcal{D}(H) \subseteq \mathcal{M}(H)$ for the set of density operators on H , i.e., those positive semidefinite operators with trace one. We will use the standard Loewner ordering \geq on $\mathcal{M}(H)$.

If H comes equipped with a canonical (ordered) orthonormal basis (as will always be the case throughout), we represent linear operators on H by matrices with rows and columns indexed by the basis elements. In this case, $\mathcal{M}(H)$ consists of symmetric matrices and $\mathcal{D}(H)$ consists of symmetric, positive semidefinite matrices whose diagonal entries summing to one. If $A \in \mathcal{M}(H)$ is positive semidefinite, we use $A^{1/2}$ to denote the positive semidefinite square root of A .

Given a linear operator $A : H \rightarrow H$, we define the operator, trace, and Frobenius norms, respectively:

$$\begin{aligned}\|A\| &= \max_{x \neq 0} \frac{|Ax|}{|x|} \\ \|A\|_* &= \text{Tr}(\sqrt{A^T A}) \\ \|A\|_F &= \sqrt{\text{Tr}(A^T A)}.\end{aligned}$$

Recall $\text{Tr}(A^T B) \leq \|A\| \cdot \|B\|_*$ and the Cauchy-Schwarz inequality $\text{Tr}(A^T B) \leq \|A\|_F \|B\|_F$. For a matrix M , we use $\|M\|_\infty$ to denote the maximum absolute value of an entry in M .

Fourier analysis and degree over the discrete cube. We use $L^2(\{-1, 1\}^n)$ to denote the Hilbert space of real-valued functions $f : \{-1, 1\}^n \rightarrow \mathbb{R}$. This space is equipped with the natural inner product under the uniform measure: $\langle f, g \rangle = \mathbb{E}_x f(x)g(x)$. We recall the Fourier basis: For $S \subseteq [n]$, one has $\chi_S(x) = \prod_{i \in S} x_i$. The functions $\{\chi_S : S \subseteq [n]\}$ form an orthonormal basis for $L^2(\{-1, 1\}^n)$. We can decompose f in the Fourier basis as $f = \sum_{S \subseteq [n]} \hat{f}(S) \chi_S$.

We will use $\text{deg}(f)$ to denote the degree of f as a multi-linear polynomial on the discrete cube: $\text{deg}(f) \stackrel{\text{def}}{=} \max\{|S| : \hat{f}(S) \neq 0\}$. Note that by identifying $\{0, 1\}$ and $\{-1, 1\}$, we can define $\text{deg}(f)$ for functions $f : \{0, 1\}^n \rightarrow \mathbb{R}$ as well. (Since the change of domains is given by the linear map $x \mapsto 2x - 1$, the degree of polynomial representations do not change.)

If we are given a matrix-valued function $M : \{-1, 1\}^n \rightarrow \mathbb{R}^{p \times q}$, we can decompose M as $M = \sum_{S \subseteq [n]} \hat{M}_S \chi_S$ where $(\hat{M}_S)_{ij} = \widehat{M_{ij}}(S)$, and $\text{deg}(M) = \max\{\text{deg}(M_{ij}) : i \in [p], j \in [q]\}$. We refer to the book [O'D14] for additional background on boolean Fourier analysis.

Quantum information theory. The *von-Neumann entropy* of a density operator X is denoted $S(X) = -\text{Tr}(X \log X)$. For two density operators X and Y over the same vector space, the *quantum relative entropy* of X with respect to Y is the quantity $S(X \| Y) = \text{Tr}(X \cdot (\log X - \log Y))$. Here, the operator function \log is defined on positive operators as $\log X = -\sum_{k=0}^{\infty} \frac{1}{k} (\text{Id} - X)^k$. In general, for a function $g : I \rightarrow \mathbb{R}$ analytic on an open interval $I \subseteq \mathbb{R}$ and a symmetric operator $X \in \mathcal{M}(H)$, we define $g(X)$ via its Taylor series, with the understanding that the spectrum of X should lie in I . Finally, we will often use the notation $\mathcal{U} = \frac{\text{Id}}{\text{Tr}(\text{Id})}$ to denote the uniform density matrix (i.e., the maximally mixed state), where the dimension of the identity matrix Id is clear from context. We refer to [Wil13] for a detailed account of quantum information theory.

2.2 Factorizations, quantum learning, and pseudo-densities

First, we recall the setup of the main theorem in the paper. Fix $m \geq 1$, a function $f : \{0, 1\}^m \rightarrow \mathbb{R}_+$ and let $d + 2 = \text{deg}_{\text{sos}}(f)$. We define the matrix M_n^f as in (1.3). Our goal is to show a lower bound on the positive semidefinite rank of the matrix M_n^f .

Suppose we had a psd factorization

$$M_n^f(S, x) = \text{Tr}(P(S)Q(x)) \quad (2.1)$$

witnessing $\text{rk}_{\text{psd}}(M) \leq r$. First, we observe that lower bound on $\text{deg}_{\text{sos}}(f)$ already precludes certain *low degree* psd factorizations. More precisely, if $R(x) = Q(x)^{1/2}$ then $\text{deg}(R)$ is constrained to be at least $d/2$. For the sake of contradiction, let us suppose $\text{deg}(R) < d/2$. For any row $M_n^f(S, \cdot)$ of the matrix M_n^f we will have,

$$f(x_S) = \text{Tr}(P(S)R(x)^2) = \|\sqrt{P(S)}R(x)\|_F^2.$$

This contradicts $\text{deg}_{\text{sos}}(f) = d + 2$ since $\|\sqrt{P(S)}R(x)\|_F^2$ is a sum of squares of a polynomials of degree less than $d/2$.

Pseudo-densities and low degree psd factorizations. By appealing to convex duality, it is possible to construct a *certificate* that the matrix M_n^f does not admit *low degree* factorizations. The certificate consists of a linear functional that separates M_n^f from the convex hull of matrices that admit low degree psd factorizations. Formally, if we define the convex set C_d of non-negative matrices as,

$$C_d \stackrel{\text{def}}{=} \left\{ N : \binom{n}{m} \times \{0, 1\}^n \rightarrow \mathbb{R} \mid N(S, x) = \text{Tr}(P(S)R(x)^2), P(S) \geq 0, \text{deg}(R(x)) < d/2 \right\}$$

then we will construct a linear functional L on $\binom{n}{m} \times 2^n$ matrices such that

$$L(M_n^f) < 0, \text{ but } L(N) \geq 0 \text{ for all } N \in C_d. \quad (2.2)$$

The linear functional is precisely the one derived from what we refer to as a *pseudo-density*.

A *degree- d pseudo-density* is a mapping $D : \{0, 1\}^m \rightarrow \mathbb{R}$ such that $\mathbb{E}_x D(x) = 1$ and $\mathbb{E}_x D(x)g(x)^2 \geq 0$ for all functions $g : \{0, 1\}^m \rightarrow \mathbb{R}$ with $\text{deg}(g) \leq d/2$.⁸ Observe that for any probability distribution over $\{0, 1\}^n$, its density function relative to the uniform distribution on $\{0, 1\}^n$ satisfies the conditions of a degree- d pseudo-density for every $d \in \mathbb{N}$. One has the following characterization:

$$\text{deg}_{\text{sos}}(f) = \min \left\{ d \geq 0 : \mathbb{E}_x D(x)f(x) \geq 0 \text{ for every degree-}d \text{ pseudo-density } D \right\}. \quad (2.3)$$

In other words, the sos degree of a function is larger than d if and only if there exists a degree- d pseudo-density D such that $\mathbb{E}_x D(x)f(x) < 0$. To verify this, note that if $\text{deg}_{\text{sos}}(f) > d$, then f is not in the closed, convex cone generated by the squares of polynomials of degree at most $d/2$. Now the required pseudo-density D corresponds exactly to (the normal vector of) a hyperplane separating f from this cone.

Of course, if D is an actual density (with respect to the uniform measure on $\{0, 1\}^m$), then $\mathbb{E}_x D(x)f(x)$ is precisely the expectation of f under D . For a pseudo-density D , the corresponding linear functional $f \mapsto \mathbb{E}_x D(x)f(x)$ is referred to as a *pseudo-expectation* in previous papers (see, e.g., [BBH⁺12, CLRS13]), and the map D is called a *pseudo-distribution* in [BKS14]. Over finite domains, these notions are interchangeable. We use the language of densities here in anticipation of future applications to infinite domains and non-uniform background measures (in the context of nonnegative rank, this occurs already in [Section 7.2](#)).

⁸Note that a degree- d pseudo-density does not necessarily have degree d as a function on the discrete cube.

Now fix a degree- d pseudo-density D with $\mathbb{E}_x D(x)f(x) < 0$. We define the following linear functional on matrices $N: \binom{[n]}{m} \times \{0,1\}^n \rightarrow \mathbb{R}$:

$$L_D(N) \stackrel{\text{def}}{=} \mathbb{E}_{|S|=m} \mathbb{E}_x D(x_S) \cdot N(S, x). \quad (2.4)$$

Consider a matrix $N \in C_d$ which admits a low degree psd factorization given by $N(S, x) = \text{Tr}(P(S)R(x)^2)$. Then since D is a degree- d pseudo-density, we would have

$$L_D(N) = \mathbb{E}_{|S|=m} \mathbb{E}_x D(x_S) \text{Tr}(P(S)R(x)^2) = \mathbb{E}_{|S|=m} \mathbb{E}_x D(x_S) \|\sqrt{P(S)}R(x)\|_F^2 \geq 0.$$

However, since D is negatively correlated with f ,

$$L_D(M_n^f) = \mathbb{E}_{|S|=m} \mathbb{E}_x D(x_S) \cdot M_n^f(S, x) = \mathbb{E}_{|S|=m} \mathbb{E}_x D(x_S) \cdot f(x_S) < -\varepsilon. \quad (2.5)$$

for some $\varepsilon > 0$.

The core of our psd rank lower bound is to show that the linear functional L_D in fact separates the matrix M_n^f from *all* low rank psd factorizations, thereby certifying a lower bound on $\text{rk}_{\text{psd}}(M_n^f)$. Roughly speaking, the idea is to approximate an arbitrary psd factorization using low degree factorizations with respect to the linear functional L_D , and then appeal to the lower bound (2.2) for low degree factorizations.

Formally, for a number $r \geq 1$, consider the following set C_r of nonnegative matrices,

$$C_r \stackrel{\text{def}}{=} \left\{ N \in \mathbb{R}_+^{\binom{[n]}{m} \times \{0,1\}^n} : \text{rk}_{\text{psd}}(N) \leq r \cdot \|N\|_1, \|N\|_\infty \leq 1 \right\}.$$

Here, $\|N\|_1$ is the average of the entries of N and $\|N\|_\infty$ is the maximum entry of N . In the rest of the section, we will present an argument that unless r is very large, every matrix $N \in C_r$ satisfies $L_D(N) \geq -\varepsilon$. Since $L_D(M_n^f) < -\varepsilon$, this implies that the linear functional L_D separates M_n^f from the convex hull of C_r , thereby certifying a lower bound on $\text{rk}_{\text{psd}}(M_n^f)$.

Fix a matrix $N \in C_r$. It is instructive to have the situation $\|N\|_1, \|N\|_\infty = \Theta(1)$ in mind for the rest of this outline. By definition of C_r , the matrix N admits a psd factorization of rank $O(r)$. In light of the above discussion, our goal is to approximate the matrix N by a low degree factorization with respect to the functional L_D . A low degree approximation for N is constructed in two steps.

Well-behaved factorizations. The first step involves obtaining a nicer factorization of N . Toward this end, we define the quantity

$$\gamma_r(M) \stackrel{\text{def}}{=} \sup \left\{ \max_{i,j} \|A_i\| \cdot \|B_j\|_* : N_{ij} = \text{Tr}(A_i B_j), A_i, B_j \in \mathcal{S}_+^r \quad \forall i \in [p], j \in [q] \right\},$$

associated with a matrix $M \in \mathbb{R}_+^{p \times q}$. The following lemma is proved by Briët, Dadush, and Pokutta [BDP13] (see also the discussion in [FGP⁺14]).

Lemma 2.1 (Factorization rescaling). *For every nonnegative matrix M with $\text{rk}_{\text{psd}}(M) \leq r$, the following holds:*

$$\gamma_r(M) \leq r^2 \|M\|_\infty.$$

Applying the above lemma to the matrix N at hand, we get a psd factorization $N(S, x) = \text{Tr}(P(S)Q(x))$ wherein $\|P(S)\|$ and $\|Q(x)\|_*$ are bounded polynomially in r . This analytic control

on the factorization will be important for controlling error bounds, but also—in a more subtle way—for the next step.

Learning a low-degree quantum approximation. The next step of the argument exploits the following phenomenon concerning quantum learning. Fix a $k \geq 1$ and consider a matrix-valued function $Q : \{0, 1\}^n \rightarrow \mathcal{S}_+^k$ such that $\mathbb{E}_x \text{Tr}(Q(x)) = 1$. We will try to approximate Q by a simpler mapping with respect to a certain class of test functionals $\Lambda : \{0, 1\}^n \rightarrow \mathcal{S}_+^k$. If \tilde{Q} is the approximator, we would like that

$$\left| \mathbb{E}_x \text{Tr} \left(\Lambda(x)(Q(x) - \tilde{Q}(x)) \right) \right| \leq \varepsilon \quad (2.6)$$

for some parameter $\varepsilon > 0$. (In this case, \tilde{Q} and Q are indistinguishable to the test Λ up to accuracy ε .) One can set this up as a quantum learning problem in the following way. We define the density matrix $U_Q = \mathbb{E}_x(e_x e_x^T \otimes Q(x))$ and the PSD matrix $V_\Lambda = \sum_x(e_x e_x^T \otimes \Lambda(x))$.⁹ Note that $\mathbb{E}_x \text{Tr}(\Lambda(x)Q(x)) = \text{Tr}(V_\Lambda U_Q)$.

Now, if \mathcal{T} is a family of test functionals, then a canonical way of finding a “simple” approximation to U_Q that satisfies all the tests is via the following maximum-entropy (convex) optimization problem:

$$\max \{S(\tilde{U}) : \text{Tr}(\tilde{U}) = 1, \tilde{U} \geq 0, |\text{Tr}(V_\Lambda(U_Q - \tilde{U}))| \leq \varepsilon \forall \Lambda \in \mathcal{T}\}, \quad (2.7)$$

where we recall that $S(\cdot)$ denotes the quantum entropy functional. Moreover, one can attempt to solve this optimization by some form of projected sub-gradient descent. Interpretations of this algorithm go by many names, notably the “matrix multiplicative weights update method” and “mirror descent” with quantum entropy as the regularizer; see, e.g., [NY83, BT03, TRW05, AK07, WK12] and the recent survey [Bub14].

In our setting, we are not directly concerned with efficiency, but instead simplicity of the approximator. A key phenomenon is that when the class of tests \mathcal{T} is simple, the approximator inherits this simplicity. Moreover, one can tailor the nature of the approximator by choosing the sub-gradient steps wisely. In Section 4.2 (Theorem 4.5), we prove a generalization of the following approximation theorem. (Recall that $\mathcal{U} = \text{Id} / \text{Tr}(\text{Id})$ is the uniform density matrix.)

Theorem 2.2 (Approximation by a low-degree square). *Let $\kappa \geq 1$ and $\omega > 0$ be given. Define*

$$\mathcal{T}_{\kappa, \omega} = \left\{ \Lambda : \{0, 1\}^n \rightarrow \mathcal{S}_+^k : \text{deg}(\Lambda) \leq \kappa, \|\Lambda(x)\| \leq \omega \forall x \in \{0, 1\}^n \right\}.$$

For any $Q : \{0, 1\}^n \rightarrow \mathcal{S}_+^k$ with $\mathbb{E}_x \text{Tr}(Q(x)) = 1$, there is a matrix-valued function $R : \{0, 1\}^n \rightarrow \mathcal{S}_+^k$ with $\mathbb{E}_x \text{Tr}(R(x)^2) = 1$ satisfying

$$\frac{\text{deg}(R)}{\kappa} \lesssim (1 + S(U_Q \|\mathcal{U})) \frac{\omega}{\varepsilon}, \quad (2.8)$$

and for all tests $\Lambda \in \mathcal{T}_{\kappa, \omega}$,

$$\left| \mathbb{E}_x \text{Tr} \left(\Lambda(x)(Q(x) - R(x)^2) \right) \right| \leq \varepsilon.$$

In other words, the learning algorithm produces a hypothesis with error at most ε for all the tests in $\mathcal{T}_{\kappa, \omega}$; moreover, the hypothesis is the square of a polynomial whose degree is not much larger than that of the tests. The value ω corresponds to the ubiquitous “width” parameter and, as in most applications of the multiplicative weights method, bounding ω will be centrally important. The reader should also take note of the appearance of the relative entropy in the degree bound (2.8). It will turn out that low psd rank factorizations will give us functions $Q : \{0, 1\}^n \rightarrow \mathcal{S}_+^k$ with high

⁹In the quantum information literature, these are sometimes called QC states for “quantum/classical.”

entropy (and thus small relative entropy with respect to the uniform state); this is actually a direct consequence of the factorization rescaling in [Lemma 2.1](#).

Notice that the separating functional L_D induces a test of degree at most m . Therefore, if one takes for granted, as claimed above, that $S(U_Q \| \frac{\text{Id}}{\text{Tr}(\text{Id})})$ is small when Q comes from a low psd rank factorization, then [Theorem 2.2](#) suggests that we might think of $Q(x)$ as being a low-degree square.

Proof sketch for [Theorem 1.8](#). We have all the ingredients to sketch a proof of [Theorem 1.8](#). First, suppose that $\deg_{\text{sos}}(f) > d$ so that by [\(2.3\)](#), there exists a degree- d pseudo-density D with $\mathbb{E}_x f(x)D(x) < -\varepsilon\|f\|_\infty$ for some $\varepsilon > 0$. (We do not specify any quantitative bound on ε at the moment, but we write it this way to indicate how one can get improved bounds under stronger assumptions.)

Then from the definition of M_n^f , we have

$$L_D(M_n^f) < -\varepsilon\|M_n^f\|_\infty. \quad (2.9)$$

On the other hand, we will prove the following theorem in [Section 3.1](#).

Theorem 2.3. *For every $m, d \geq 1$, every $\varepsilon \in (0, 1]$, and every degree- d pseudo-density $D : \{0, 1\}^m \rightarrow \mathbb{R}$, there exists a number $\alpha > 0$ such that for every $n \geq 2m$ and every nonnegative matrix $N : \binom{[n]}{m} \times \{0, 1\}^n \rightarrow \mathbb{R}$ satisfying*

$$\|N\|_\infty \leq 1, \text{ and} \\ \frac{1}{\|N\|_1} \text{rk}_{\text{psd}}(N)^2 \leq \alpha(n/\log n)^{d/2},$$

we have $L_D(N) \geq -\varepsilon$.

Now if we consider the normalized matrix $N = M_n^f / \|M_n^f\|_\infty$, we see that it satisfies the first premise $\|N\|_\infty \leq 1$ but violates the conclusion of the theorem (because of [\(2.9\)](#)). Therefore we know that the second premise is violated, which gives the lower bound

$$\text{rk}_{\text{psd}}(N)^2 > \alpha(n/\log n)^{d/2} \cdot \|N\|_1 = \alpha(n/\log n)^{d/2} \mathbb{E}_x f(x).$$

Since this achieves our goal, we are left to explain why [Theorem 2.3](#) should be true, at least when we apply it with $N = M_n^f / \|M_n^f\|_\infty$. If we apply L_D to the right-hand side of [\(2.1\)](#)—our presumed factorization for M_n^f —we arrive at the expression

$$L_D(M_n^f) = \mathbb{E}_x \text{Tr} \left(\mathbb{E}_{|S|=m} D(x_S) P(S) Q(x) \right). \quad (2.10)$$

We can view this as a test on Q in the sense of [Theorem 2.2](#). Since $\deg(D) \leq m$ (because D is only a function of m variables), this is a low-degree test. [Theorem 2.2](#) then suggests that we can replace Q by a low-degree approximator R^2 , while losing only ε in the “accuracy” of the test.

Since the approximation property implies that $Q(x)$ and $R(x)^2$ should perform similarly under the test (up to the “accuracy” ε), we would conclude that $L_D(M_n^f) \geq -\varepsilon$, yielding the conclusion of [Theorem 2.3](#).

Random restriction and degree reduction. The one serious issue with the preceding argument is that our supposition is far too strong: One cannot expect to have $\deg(R) \leq d/2$. Indeed, the guarantee of [Theorem 2.2](#) tells us that the approximator $R(x)$ has degree at most $K \cdot \deg(D)$ for some (possibly large) number K (which itself depends on many parameters). To overcome this

problem, we use another crucial property of our functional (2.4): It is an expectation over small sets $S \subseteq [n]$. If we randomly choose such a subset with $|S| = m \ll n$ and randomly choose values $y_{\bar{S}}$ for the variables in \bar{S} , we expect that the resulting (partially evaluated) polynomial $R(x_S, x_{\bar{S}})|_{x_{\bar{S}}=y_{\bar{S}}}$ will satisfy $\deg(R(x_S, x_{\bar{S}})|_{x_{\bar{S}}=y_{\bar{S}}}) \ll \deg(R)$. (Strictly speaking, this will only be true in an approximate sense.)

It is precisely this degree reduction property of random restriction that saves the preceding sketch. In the next sections, we perform a more delicate quantitative analysis capable of achieving much stronger lower bounds. The norm $\|D\|_\infty$ of the pseudo-density will play a central role in this study. Thus in Section 5.1, we show that Grigoriev's proof of Theorem 1.12 can be carefully recast in the language of pseudo-densities such that the resulting pseudo-density has small norm.

3 PSD rank and sum-of-squares degree

We now move to proving the main technical theorems of the paper along the lines of the informal overview presented in Section 2.2.

3.1 Analysis of the separating functional

Recall that for a pseudo-density $D: \{0, 1\}^m \rightarrow \mathbb{R}$ and $n \geq 1$, we define a linear functional L_D on matrices $N: \binom{[n]}{m} \times \{0, 1\}^n \rightarrow \mathbb{R}$

$$L_D(N) \stackrel{\text{def}}{=} \mathbb{E}_x \mathbb{E}_S D(x_S) N(S, x),$$

where the expectation over S is a uniform average over all $S \subseteq [n]$ with $|S| = m$ (as will be the case throughout this section). We will use the notation $\|N\|_\infty = \max_{S, x} N(S, x)$ and $\|N\|_1 = \mathbb{E}_{S, x} N(S, x)$.

We prove the following quantitative version of Theorem 3.1. As discussed in Section 2.2, this theorem implies a lower bound on the $\text{rk}_{\text{psd}}(M_n^f)$ in terms of $\deg_{\text{SOS}}(f)$. This implication will be proved formally in Section 3.3.

Theorem 3.1 (Strengthening of Theorem 2.3). *For every $m, d \geq 1$, every $\varepsilon \in (0, 1]$, and every degree- d pseudo-density $D: \{0, 1\}^m \rightarrow \mathbb{R}$, there exists a number $\alpha > 0$ such that whenever $n \geq 2m$ and a nonnegative matrix $N: \binom{[n]}{m} \times \{0, 1\}^n \rightarrow \mathbb{R}$ satisfies*

$$\begin{aligned} \|N\|_\infty &\leq 1, \\ \frac{1}{\|N\|_1} \text{rk}_{\text{psd}}(N)^2 &\leq \alpha(n/\log n)^{d/2}, \end{aligned}$$

we have $L_D(N) \geq -\varepsilon$. Moreover, this holds for

$$\alpha = \left(\frac{C\varepsilon}{dm^2\|D\|_\infty} \right)^{d/2} \left(\frac{\varepsilon}{\|D\|_\infty} \right)^3,$$

where $C > 0$ is a universal constant.

The proof of this theorem consists of two parts. First, we observe that if D is a degree- d pseudo-density, then $L_D(N)$ is nonnegative for all matrices N that admit a factorization in terms of squares of low-degree polynomials, i.e., a factorization $N(S, x) = \text{Tr}(A_S^2 B_x^2)$ for symmetric matrices

$\{A_S\}$ and $\{B_x\}$ such that the function $x \mapsto B_x$ has degree at most $d/2$ over $\{0,1\}^n$.¹⁰ Indeed, consider such a factorization. Then,

$$L_D(N) = \mathbb{E}_x \mathbb{E}_S D(x_S) \text{Tr}(A_S^2 B_x^2) = \mathbb{E}_x \left(\mathbb{E}_S D(x_S) \|A_S B_x\|_F^2 \right) \geq 0,$$

where the inequality used the fact that D is a degree- d pseudo-density (hence $\mathbb{E}_x D(x)g(x)^2 \geq 0$ whenever $\deg(g) \leq d/2$).

As explained in [Section 2.2](#), this guarantee is not sufficient for us. The following theorem (proved in [Section 3.2](#)) allows us to analyze L_D even when the degree of the map $x \mapsto B_x$ is much larger than $d/2$. (For $m \leq n^{o(1)}$, it will be the case that the linear functional L_D is approximately nonnegative on N even when $x \mapsto B_x$ has degree up to $n^{o(1)}$).

Theorem 3.2 (Degree reduction). *Consider positive numbers $n \geq 1$ and $d, k, m \leq n$. Let $D: \{0,1\}^m \rightarrow \mathbb{R}$ be a degree- d pseudo-density. Let $N': \binom{[n]}{m} \times \{0,1\}^n \rightarrow \mathbb{R}$ be a matrix that admits a factorization $N'(S, x) = \text{Tr} A_S^2 B_x^2$ for symmetric matrices $\{A_S\}$ and $\{B_x\}$ such that the matrix-valued function $x \mapsto B_x$ has degree at most ℓ . Then,*

$$L_D(N') \gtrsim - \left(\frac{\ell m}{n - m} \right)^{d/4} \|D\|_\infty \cdot \left(\left(\max_S \|A_S^2\| \cdot \mathbb{E}_x \text{Tr}(B_x^2) \right) \cdot \left(\mathbb{E}_x \mathbb{E}_S N'(S, x) \right) \right)^{1/2}.$$

With this theorem in place, our goal is to approximate every matrix N with low psd rank by a matrix N' that satisfies the premise of [Theorem 3.2](#) for a reasonable value of ℓ . Here, our notion of approximation is fairly weak. We only require $L_D(N) \geq L_D(N') - \varepsilon$ for sufficiently small $\varepsilon > 0$. As a preliminary step, the following general theorem about psd factorizations allows us to assume that the factorization for N is appropriately scaled. Recall that $\mathcal{U} = \text{Id} / \text{Tr}(\text{Id})$ is the uniform density matrix.

Theorem 3.3 (psd factorization scaling). *For every nonnegative matrix $M \in \mathbb{R}^{p \times q}$ and every $\eta \in (0, 1]$, there exist psd matrices $\{P_i\}_{i \in [p]}$ and $\{Q_j\}_{j \in [q]}$ with the following properties:*

1. $M_{i,j} \leq \text{Tr}(P_i Q_j) \leq M_{i,j} + \eta \|M\|_\infty$,
2. $\frac{1}{p} \sum_{i=1}^p P_i = \text{Id}$,
3. $\|P_i\| \leq 2 \text{rk}_{\text{psd}}(M)^2 / \eta$ for all $i \in [p]$,
4. $Q_j \leq \|M\|_\infty (\eta + \text{rk}_{\text{psd}}(M)^2) \text{rk}_{\text{psd}}(M) \mathcal{U}$ for all $j \in [q]$.

Proof. Let $r = \text{rk}_{\text{psd}}(M)$. By [Lemma 2.1](#), we have $\gamma := \gamma_r(M) \leq r^2 \|M\|_\infty$. Fix a factorization $M_{i,j} = \text{Tr}(A_i B_j)$ such that $\max_{i,j} \|A_i\| \cdot \|B_j\|_* = \|M\|_\infty \cdot \text{rk}_{\text{psd}}(M)^2$ and $A_i, B_j \in \mathbb{R}^{r \times r}$. By an appropriate normalization, we may assume $A_i, B_j \geq 0$ and $\|A_i\| \leq \gamma, \|B_j\|_* \leq 1$. To construct psd matrices $\{P_i\}$ and $\{Q_j\}$ with the desired properties, make the following definitions:

$$\begin{aligned} A &= \eta \|M\|_\infty \text{Id} + \frac{1}{p} \sum_{i=1}^p A_i \\ P_i &= A^{-1/2} (\eta \|M\|_\infty \text{Id} + A_i) A^{-1/2} \\ Q_j &= A^{1/2} B_j A^{1/2}. \end{aligned}$$

¹⁰For the convenience of the reader, we recall that the degree of the matrix-valued function $x \mapsto B_x$ is defined as the maximum degree of the functions $x \mapsto (B_x)_{ij}$ where i, j range over the indices of B_x .

Note that [Item 2](#) holds by construction. Also observe that

$$\text{Tr}(P_i Q_j) = M_{i,j} + \eta \|M\|_\infty \text{Tr}(A^{-1} A^{1/2} B_j A^{1/2}) = M_{i,j} + \eta \|M\|_\infty \text{Tr}(B_j),$$

verifying [Item 1](#). Finally, we have the inequalities for all $i \in [p], j \in [q]$,

$$\begin{aligned} \|P_i\| &\leq \frac{1}{\eta \|M\|_\infty} (\eta \|M\|_\infty + \|A_i\|) \leq 1 + \frac{\gamma}{\eta \|M\|_\infty}, \\ \|Q_j\|_* &\leq \|A\| \cdot \|B_j\|_* \leq \gamma + \eta \|M\|_\infty. \end{aligned}$$

The first inequality verifies [Item 3](#) since $r \geq 1$ and $\eta \leq 1$. The last inequality implies that

$$Q_j \leq (\gamma + \eta \|M\|_\infty) r \frac{\text{Id}}{\text{Tr}(\text{Id})} \leq r \|M\|_\infty (\eta + r^2) \frac{\text{Id}}{\text{Tr}(\text{Id})}$$

for all $j \in [q]$, verifying [Item 4](#). □

Consider a matrix of the form $N : \binom{[n]}{m} \times \{0, 1\}^n \rightarrow \mathbb{R}_+$ with $\|N\|_\infty \leq 1$ and let $\varepsilon > 0$ be given. Apply [Theorem 3.3](#) with a value $\eta \in (0, 1]$ to be chosen later to obtain a factorization

$$N(S, x) = \text{Tr}(P_S Q_x)$$

satisfying the conclusions of the theorem.

We can view the matrix-valued function $x \mapsto Q_x$ as a density matrix $Q = \frac{1}{\mathbb{E}_x(\text{Tr} Q_x)} \mathbb{E}_x(e_x e_x^T \otimes Q_x)$. (The first n bits in this density matrix are ‘‘classical’’ and their marginal distribution has density $x \mapsto \text{Tr} Q_x$. If we condition Q on an assignment $x \in \{0, 1\}^n$ to the first n bits, the resulting quantum state is $\frac{1}{\text{Tr} Q_x} Q_x$). Here, the normalization factor $\tau = \mathbb{E}_x \text{Tr} Q_x$ for the density matrix Q satisfies

$$\tau = \mathbb{E}_x \text{Tr} Q_x \stackrel{(\text{Thm 3.3(2)})}{=} \mathbb{E}_x \mathbb{E}_S \text{Tr} P_S Q_x \begin{cases} \stackrel{(\text{Thm 3.3(1)})}{\geq} \mathbb{E}_S \mathbb{E}_x N(S, x) = \|N\|_1, \\ \stackrel{(\text{Thm 3.3(1)})}{\leq} \mathbb{E}_S \mathbb{E}_x N(S, x) + \eta \leq 1 + \eta, \end{cases} \quad (3.1)$$

where the last inequality has used $\|N\|_\infty \leq 1$.

From [Theorem 3.3\(4\)](#), the density matrix Q satisfies

$$Q \leq \frac{(\eta + \text{rk}_{\text{psd}}(N)^2) \text{rk}_{\text{psd}}(N)}{\tau} \mathcal{U}.$$

Therefore,

$$S(Q \| \mathcal{U}) \leq \log(\eta \text{rk}_{\text{psd}}(N) / \tau) \leq \log(\text{rk}_{\text{psd}}(N) / \|N\|_1). \quad (3.2)$$

[Theorem 3.3\(1\)](#) allows us to lower bound $L_D(N)$ in terms of the matrix $(S, x) \mapsto \text{Tr}(P_S Q_x)$ and value $\|D\|_\infty$:

$$\begin{aligned} L_D(N) &= \mathbb{E}_x \mathbb{E}_S D(x_S) N(S, x) \\ &\geq \mathbb{E}_x \mathbb{E}_S D(x_S) \cdot \text{Tr}(P_S Q_x) - \eta \|D\|_\infty \quad (\text{by } \text{Theorem 3.3(1)}) \\ &= \tau \cdot \text{Tr}(FQ) - \eta \|D\|_\infty, \end{aligned} \quad (3.3)$$

where F is the symmetric matrix

$$F = \sum_{x \in \{0, 1\}^n} e_x e_x^T \otimes F_x \quad \text{with } F_x = \mathbb{E}_S D(x_S) P_S. \quad (3.4)$$

[Theorem 3.3\(2\)](#) allows us to upper bound the spectral norm of F by

$$\|F\| \leq \max_x \left\| \mathbb{E}_S D(x_S) P_S \right\| \leq \|D\|_\infty \cdot \left\| \mathbb{E}_S P_S \right\| \stackrel{(\text{Thm 3.3(2)})}{=} \|D\|_\infty. \quad (3.5)$$

The next theorem allows us to lower bound $\text{Tr}(FQ)$ by replacing Q with a simpler density matrix that is a low-degree polynomial in F . (See [Theorem 4.1](#), where a slightly more general version is proved.)

Theorem 3.4 (Density matrix approximation). *Let H be some finite-dimensional real inner-product space. Let $F \in \mathcal{M}(H)$ be a symmetric matrix and let $Q \in \mathcal{D}(H)$ be a density matrix. Then, for every $\varepsilon > 0$, there exists a degree- k univariate polynomial p with $k \lesssim (1 + S(Q \| \mathcal{U})) \cdot \|F\|/\varepsilon$ such that the density matrix $\tilde{Q} = \frac{1}{\text{Tr } p(F)^2} p(F)^2$ satisfies*

$$\text{Tr}(F\tilde{Q}) \leq \text{Tr}(FQ) + \varepsilon. \quad (3.6)$$

Apply [Theorem 3.4](#) to the density matrix Q and the symmetric matrix F defined above with the value ε (which is already fixed). Let p be the resulting degree- k polynomial, with k satisfying the bounds of the theorem.

Since the function $x \mapsto F_x$ has $\deg(F) \leq \deg(D) \leq m$ (since $D : \{0, 1\}^m \rightarrow \mathbb{R}$), the degree of the map $x \mapsto \tilde{Q}_x = \frac{1}{\mathbb{E}_x \text{Tr}(p(F_x)^2)} p(F_x)^2$ is at most $\deg(p) \cdot m = k \cdot m$. Applying [Theorem 3.2](#) to the matrix given by $N'(S, x) = \text{Tr}(P_S \cdot p(F_x)^2)$, we can give a lower bound:

$$\begin{aligned} \left(\mathbb{E}_x \text{Tr } p(F_x)^2 \right) \cdot \text{Tr}(F \cdot \tilde{Q}) &= \mathbb{E}_S \mathbb{E}_x D(x_S) \text{Tr}(P_S \cdot p(F_x)^2) \\ &\geq - \left(\frac{km^2}{n-m} \right)^{d/4} \cdot \|D\|_\infty \cdot \left(\left(\max_S \|P_S\| \cdot \mathbb{E}_x \text{Tr}(p(F_x)^2) \right) \cdot \left(\mathbb{E}_x \mathbb{E}_S N'(S, x) \right) \right)^{1/2}. \end{aligned}$$

Using the fact that $\mathbb{E}_x \mathbb{E}_S N'(S, x) = \mathbb{E}_S \mathbb{E}_x \text{Tr } P_S \cdot p(F_x)^2 = \mathbb{E}_x \text{Tr } p(F_x)^2$ from [Theorem 3.3\(2\)](#) and the fact that $\max_S \|P_S\| \leq 2\text{rk}_{\text{psd}}(N)^2/\eta$ from [Theorem 3.3\(3\)](#) yields

$$\text{Tr}(F \cdot \tilde{Q}) \geq - \left(\frac{km^2}{n-m} \right)^{d/4} \frac{\|D\|_\infty}{\sqrt{\eta}} \text{rk}_{\text{psd}}(N). \quad (3.7)$$

We have now assembled all components of the proof of [Theorem 3.1](#).

Proof of [Theorem 3.1](#). We lower bound the linear functional $L_D(N)$ by

$$\begin{aligned} L_D(N) &\stackrel{(3.3)}{\geq} \tau \cdot \text{Tr}(FQ) - \eta \|D\|_\infty \\ &\stackrel{(3.6)}{\geq} \tau \cdot \left(\text{Tr}(F \cdot \tilde{Q}) - \varepsilon \right) - \eta \|D\|_\infty \\ &\stackrel{(3.7)}{\geq} -c\tau \cdot \left(\frac{km^2}{n-m} \right)^{d/4} \cdot \frac{\|D\|_\infty}{\sqrt{\eta}} \text{rk}_{\text{psd}}(N) - \tau \cdot \varepsilon - \eta \|D\|_\infty, \end{aligned}$$

where $c > 0$ is a universal constant.

Now set $\eta := \min(\varepsilon/\|D\|_\infty, 1)$ and use [\(3.1\)](#) to bound $\tau \leq 1 + \eta \leq 2$. This yields

$$L_D(N) \geq -2c \left(\frac{km^2}{n-m} \right)^{d/4} \cdot \frac{\|D\|_\infty^{3/2}}{\sqrt{\varepsilon}} \text{rk}_{\text{psd}}(N) - 3\varepsilon \quad (3.8)$$

Now recall that our invocation of [Theorem 3.4](#) gives us a bound on $k = \deg(p)$:

$$k \lesssim (1 + S(Q \| \mathcal{U})) \cdot \|F\|/\varepsilon \stackrel{(3.5), (3.2)}{\lesssim} \log(\text{rk}_{\text{psd}}(N)/\|N\|_1) \frac{\|D\|_\infty}{\varepsilon}. \quad (3.9)$$

If $\text{rk}_{\text{psd}}(N)^2/\|N\|_1$ satisfies the upper bound in the theorem, then the degree bound (3.9) above gives

$$k \lesssim \frac{1}{\varepsilon} \cdot d \|D\|_\infty \cdot \log n.$$

Plugging this bound into (3.8) yields, for some constant $c' > 0$,

$$L_D(N) \geq - \left(\frac{c' d \|D\|_\infty m^2 \log n}{\varepsilon(n-m)} \right)^{d/4} \cdot \frac{\|D\|_\infty^{3/2}}{\sqrt{\varepsilon}} \text{rk}_{\text{psd}}(N) - 3\varepsilon$$

Since $\|N\|_1 \leq \|N\|_\infty \leq 1$, if $\text{rk}_{\text{psd}}(N)$ satisfies the upper bound in the theorem (for a sufficiently small constant α), this lower bound is $L_D(N) \geq -4\varepsilon$ as desired (up to scaling by a factor of 4). \square

3.2 Degree reduction

The next theorem is a restatement of [Theorem 3.2](#). One should simply note that for any symmetric matrix A , we have $\|A\|_F^2 = \text{Tr}(A^2)$.

Theorem 3.5 (Restatement of [Theorem 3.2](#)). *Let positive integers $n \geq 1$ and $m, d, \ell \leq n$ be given. Suppose $A: \binom{[n]}{m} \rightarrow \mathbb{R}^{p \times p}$ and $B: \{0, 1\}^n \rightarrow \mathbb{R}^{p \times p}$ are two functions taking symmetric matrices as values. Let $D: \{0, 1\}^m \rightarrow \mathbb{R}$ be a degree- d pseudo-density and suppose that $\deg(B) \leq \ell$. Then,*

$$\begin{aligned} & \mathbb{E}_{S,x} D(x_S) \|A(S)B(x)\|_F^2 \\ & \geq -2\|D\|_\infty \left(\frac{\ell m}{(n-m)} \right)^{d/4} \cdot \left(\max_S \|A(S)^2\| \right)^{1/2} \left(\mathbb{E}_{S,x} \|A(S)B(x)\|_F^2 \right)^{1/2} \cdot \left(\mathbb{E}_x \|B(x)\|_F^2 \right)^{1/2}, \end{aligned}$$

Proof. For the sake of this lemma, which uses Fourier analysis, we will think of B and D as functions on $\{-1, 1\}^n$. Since this is a linear transformation on the domain, it does not affect their degrees as multilinear polynomials.

For every $S \subseteq [n]$ with $|S| = m$, we decompose B into two parts $B = B_{S,\text{low}} + B_{S,\text{high}}$ such that $B_{S,\text{low}}$ is the part of B with degree at most $d/2$ in the variables S :

$$B_{S,\text{low}} = \sum_{\substack{\alpha \subseteq [n] \\ |\alpha \cap S| \leq d/2}} \hat{B}_\alpha \chi_\alpha.$$

(Recall [Section 2.1](#) for the Fourier-analytic definitions.)

The proof consists of two steps that are captured by the following two lemmas.

Lemma 3.6. *Let $\tau = \max_S \|A(S)^2\|$. Then,*

$$\mathbb{E}_{S,x} D(x_S) \|A(S)B(x)\|_F^2 \geq -2\sqrt{\tau} \|D\|_\infty \cdot \left(\mathbb{E}_{S,x} \|B_{S,\text{high}}(x)\|_F^2 \right)^{1/2} \cdot \left(\mathbb{E}_{S,x} \|A(S)B(x)\|_F^2 \right)^{1/2}$$

Proof. For ease of notation, we will treat $A = A(S)$ and $B = B(x)$ as matrix-valued random variables that are determined by choosing $x \in \{-1, 1\}^n$ and $S \subseteq [n]$ with $|S| = m$ uniformly and independently at random. In this notation, we are to lower bound the expectation $\mathbb{E} D(x_S) \|AB\|_F^2$ (over the joint distribution of x, S, A , and B).

Let $B_{\text{low}} = B_{S,\text{low}}(x)$ and $B_{\text{high}} = B_{S,\text{high}}(x)$ be matrix-valued random variables in the same probability space. By construction, the Fourier transforms of the functions $x \mapsto B_{S,\text{low}}(x)$ and

$x \mapsto B_{S,\text{high}}(x)$ have disjoint support for every subset S . Therefore, the expectation satisfies $\mathbb{E} B_{\text{low}} B_{\text{high}}^T = 0$. This fact allows us to control the expectations of $\|AB_{\text{low}}\|_F^2$ and $\|AB_{\text{high}}\|_F^2$,

$$\mathbb{E} \|AB_{\text{low}}\|_F^2 + \mathbb{E} \|AB_{\text{high}}\|_F^2 = \mathbb{E} \|AB\|_F^2$$

Here, we have used that the quadratic formula $\|AB\|_F^2 = \|AB_{\text{low}}\|_F^2 + \|AB_{\text{high}}\|_F^2 + 2\langle AB_{\text{low}}, AB_{\text{high}} \rangle$, where $\langle \cdot, \cdot \rangle$ is the inner product that induces $\|\cdot\|_F$, i.e., $\langle X, Y \rangle = \text{Tr}(X^T Y)$. Hence,

$$\mathbb{E} [\langle AB_{\text{low}}, AB_{\text{high}} \rangle \mid A] = \text{Tr}(A^2 \cdot \mathbb{E} B_{\text{low}} B_{\text{high}}^T) = 0.$$

Therefore,

$$\begin{aligned} & \left| \mathbb{E} [D(x_S) \|AB\|_F^2] - \mathbb{E} [D(x_S) \|AB_{\text{low}}\|_F^2] \right| \\ & \leq \|D\|_\infty \cdot \mathbb{E} \left[\left| \|AB\|_F + \|AB_{\text{low}}\|_F \right| \cdot \left| \|AB\|_F - \|AB_{\text{low}}\|_F \right| \right] \\ & \leq \|D\|_\infty \cdot \left(\mathbb{E} \left[\left| \|AB\|_F + \|AB_{\text{low}}\|_F \right|^2 \right] \cdot \mathbb{E} \left[\left| \|AB\|_F - \|AB_{\text{low}}\|_F \right|^2 \right] \right)^{1/2} \\ & \leq 2\|D\|_\infty \cdot (\mathbb{E} \|AB_{\text{high}}\|_F^2)^{1/2} \cdot (\mathbb{E} \|AB\|_F^2)^{1/2}. \end{aligned}$$

The first step used the identity $|x^2 - y^2| = |x + y| \cdot |x - y|$. In the second step, we applied Cauchy–Schwarz. The third step used the triangle inequality, $\left| \|AB\|_F - \|AB_{\text{low}}\|_F \right| \leq \|AB_{\text{high}}\|_F$.

Since $x \mapsto \|A(S)B_{S,\text{low}}(x)\|_F^2$ is a sum of squares of polynomials of degree at most $d/2$ in the variables S and D is a degree- d pseudo-density, the expectation $\mathbb{E} D(x_S) \|AB_{\text{low}}\|_F^2$ is non-negative. It follows that

$$\mathbb{E} [D(x_S) \|AB\|_F^2] \geq -2\|D\|_\infty \cdot (\mathbb{E} \|AB_{\text{high}}\|_F^2)^{1/2} \cdot (\mathbb{E} \|AB\|_F^2)^{1/2}.$$

We also have

$$\mathbb{E} \|AB_{\text{high}}\|_F^2 \leq \max_S \|A(S)\|^2 \cdot \mathbb{E} \|B_{\text{high}}\|_F^2 = \tau \mathbb{E} \|B_{\text{high}}\|_F^2.$$

This bound implies the desired lower bound

$$\mathbb{E} D(x_S) \|AB\|_F^2 \geq -2\sqrt{\tau} \|D\|_\infty \cdot (\mathbb{E} \|B_{\text{high}}\|_F^2)^{1/2} \cdot (\mathbb{E} \|AB\|_F^2). \quad \square$$

Lemma 3.7.

$$\mathbb{E}_{S,x} \|B_{S,\text{high}}(x)\|_F^2 \leq \frac{\ell^{d/2} m^{d/2}}{(n-m)^{d/2}} \cdot \mathbb{E}_{S,x} \|B(x)\|_F^2$$

Proof. By construction the Fourier transform of $B_{S,\text{high}}$ satisfies

$$B_{S,\text{high}} = \sum_{\substack{\alpha \subseteq [n] \\ |\alpha \cap S| > d/2}} \hat{B}(\alpha) \chi_\alpha.$$

Therefore,

$$\mathbb{E}_x \|B_{S,\text{high}}(x)\|_F^2 = \sum_{\substack{\alpha \subseteq [n] \\ |\alpha \cap S| > d/2}} \|\hat{B}(\alpha)\|_F^2.$$

The expectation satisfies

$$\mathbb{E}_S \mathbb{E}_x \|B_{S,\text{high}}(x)\|_F^2 = \sum_{\alpha \subseteq [n]} \|\hat{B}(\alpha)\|_F^2 \cdot \mathbb{P} \{ |\alpha \cap S| > d/2 \}.$$

Since B has degree at most ℓ , we can upper bound the probability of the event $\{|\alpha \cap S| > d/2\}$,

$$\mathbb{P}\{|\alpha \cap S| > d/2\} \leq \binom{\ell}{d/2} \binom{n}{m-d/2} / \binom{n}{m} \leq \frac{\ell^{d/2} m^{d/2}}{(n-m)^{d/2}}.$$

Together with $\sum_{\alpha} \|\hat{B}_{\alpha}\|_F^2 = \mathbb{E}_x \|B(x)\|_F^2$, the desired bound on the expected norm of $B_{S,\text{high}}$ follows:

$$\mathbb{E}_S \mathbb{E}_x \|B_{S,\text{high}}(x)\|_F^2 \leq \frac{\ell^{d/2} m^{d/2}}{(n-m)^{d/2}} \cdot \mathbb{E}_x \|B(x)\|_F^2. \quad \square$$

We combine the previous two lemmas to lower bound the correlation between the pseudo-density $D(x_S)$ and the norms $\|A(S)B(x)\|_F^2$,

$$\begin{aligned} \mathbb{E}_{S,x} D(x_S) \|A(S)B(x)\|_F^2 &\geq -2\sqrt{\tau} \|D\|_{\infty} \left(\mathbb{E}_{S,x} \|A(S)B(x)\|_F^2 \right)^{1/2} \left(\mathbb{E}_{S,x} \|B_{S,\text{high}}(x)\|_F^2 \right)^{1/2} \\ &\quad \text{(using Lemma 3.6)} \\ &\geq -2\sqrt{\tau} \|D\|_{\infty} \frac{\ell^{d/4} m^{d/4}}{(n-m)^{d/4}} \left(\mathbb{E}_{S,x} \|A(S)B(x)\|_F^2 \right)^{1/2} \left(\mathbb{E}_x \|B(x)\|_F^2 \right)^{1/2} \\ &\quad \text{(using Lemma 3.7)}. \quad \square \end{aligned}$$

3.3 Proof of the main theorem

For a function $f: \{0,1\}^m \rightarrow [0,1]$ and an integer $n \geq m$, let $M_n^f: \binom{[n]}{m} \times \{0,1\}^n \rightarrow [0,1]$ be the matrix,

$$M_n^f(S, x) \stackrel{\text{def}}{=} f(x_S).$$

Theorem 3.8. *For any $m, d \geq 1$, the following holds. Let $f: \{0,1\}^m \rightarrow [0,1]$ be a nonnegative function with $d+2 = \deg_{\text{sos}}(f)$. Then for $n \geq 2m$,*

$$1 + n^{1+d/2} \geq \text{rk}_{\text{psd}}(M_n^f) \geq C_f \left(\frac{n}{\log n} \right)^{\frac{d}{4}}, \quad (3.10)$$

where $C_f > 0$ is a constant depending only on f .

Moreover, if there exists an $\varepsilon \in (0,1]$, and a degree- d pseudo-density $D: \{0,1\}^m \rightarrow \mathbb{R}$ with $\mathbb{E}_x D(x)f(x) < -\varepsilon$, then for every $n \geq 2m$, we have

$$\text{rk}_{\text{psd}}(M_n^f) \geq \left(\frac{c\varepsilon n}{dm^2 \|D\|_{\infty} \log n} \right)^{d/4} \left(\frac{\varepsilon}{\|D\|_{\infty}} \right)^{3/2} \sqrt{\mathbb{E}_x f(x)}, \quad (3.11)$$

where $c > 0$ is a universal constant.

Proof. Let $d+2 = \deg_{\text{sos}}(f)$ and consider a degree- d pseudo-density with $\mathbb{E} Df < -\varepsilon$ for some $\varepsilon > 0$. Recall the linear functional L_D defined in Section 3.1. One observes that $L_D(M_n^f) < -\varepsilon$.

By (the contrapositive of) Theorem 3.1, it follows that $\text{rk}_{\text{psd}}(M_n^f)^2 \geq \alpha(n/\log n)^{d/2} \cdot \|M_n^f\|_1$, where α is a constant depending only on the parameters ε, m, d , and the pseudo-density D . Note that $\|M_n^f\|_1 = \mathbb{E} f$. This immediately implies (3.10). Likewise, (3.11) follows directly from Theorem 3.1.

Let us now prove that $\text{rk}_{\text{psd}}(M_n^f) \leq 1 + n^{1+d/2}$ by exhibiting an explicit factorization of M_n^f . Let $\mathcal{F} \stackrel{\text{def}}{=} \{A \subseteq [n] : |A| \leq 1 + d/2\}$ and set $r = |\mathcal{F}|$. For $x \in \{0,1\}^n$, we use the notation $x^A := \prod_{i \in A} x_i$. Suppose $f = \sum_{j=1}^t g_j^2$ for some $\{g_j: \{0,1\}^m \rightarrow \mathbb{R}\}$ such that $\deg(g_j) \leq 1 + d/2$ for $j \in [t]$.

For each function $j \in [t]$ and subset $S \subseteq [n]$ with $|S| = m$, define the function $g_{S,j} : \{0, 1\}^n \rightarrow \mathbb{R}$ by $g_{S,j}(x) = g_j(x_S)$. We associate the coefficient vector $\hat{g}_{S,j} : \mathcal{F} \rightarrow \mathbb{R}$ associated to $g_{S,j}$ by letting $\hat{g}_{S,j}(A)$ be the coefficient of the monomial $\prod_{i \in A} x_i$ in $g_{S,j}$. Finally, for every $|S| = m$ and $x \in \{0, 1\}^m$, we define $r \times r$ PSD matrices indexed by \mathcal{F} as follows: $(Q_x)_{A,B} := x^A x^B$ and $(P_S)_{A,B} := \sum_{j=1}^t \hat{g}_{S,j}(A) \hat{g}_{S,j}(B)$. It is easy to check that

$$\text{Tr}(P_S Q_x) = \sum_{j=1}^t \left(\sum_{A \in \mathcal{F}} x^A \hat{g}_{S,j}(A) \right)^2 = \sum_{j=1}^t g_{S,j}(x)^2 = f(x_S) = M_n^f(S, x),$$

which yields an explicit psd factorization of M_n^f with matrices $\{P_S\}, \{Q_x\}$ of dimension $r = \sum_{i \leq 1+d/2} \binom{n}{i} \leq 1 + n^{1+d/2}$. \square

4 Approximations for density operators

We turn now to a central theme of our approach: High-entropy states can be approximated by “simple” states if the approximation is only with respect to “simple” tests. In our setting, “simple” will mean low-degree. In [Section 4.1](#), we present a basic version of this principle with respect to a single test functional. This suffices for essentially all our applications to psd rank lower bounds.

We believe that the maximum-entropy approximation framework is a powerful one, so [Section 4.2](#) is devoted to a more general exploration of the principle. In particular, we state and prove approximation theorems for density operators with respect to families of tests. In the rest of this section, we fix a finite-dimensional real inner product space H .

4.1 Approximation against a single test

The following theorem shows that a linear functional over density matrices with high entropy is approximately minimized at a density matrix that is the square of a low-degree polynomial in the linear functional. We recall that $U = \frac{\text{Id}}{\text{Tr}(\text{Id})}$ is the uniform density matrix.

Theorem 4.1 (Density matrix approximation). *Let $F \in \mathcal{M}(H)$ be a symmetric matrix and let $Q \in \mathcal{D}(H)$ be a density matrix. Then, for every $\varepsilon \in (0, \frac{1}{2})$, there exists a degree- k univariate polynomial p with $k \leq O(\|F\|/\varepsilon) \cdot S(Q \| U) + O\left(\frac{\log 1/\varepsilon}{\log \log 1/\varepsilon}\right)$ such that*

$$\text{Tr} \left(F \cdot \frac{1}{\text{Tr}(p(F)^2)} p(F)^2 \right) \leq \text{Tr}(FQ) + \varepsilon.$$

Moreover, the polynomial p depends only on ε , the operator norm $\|F\|$, and the relative entropy $S(Q \| U)$.

The proof consists of two steps. First, we will show that the theorem holds with $\frac{1}{\text{Tr}(p(F)^2)} p(F)^2$ replaced by $e^{-\lambda F} / \text{Tr}(e^{-\lambda F})$ for $\lambda \leq (1/\varepsilon) \cdot S(Q \| U)$. Then, we will approximate the matrix exponential by the square of a low-degree polynomial.

Lemma 4.2. *For every symmetric matrix F and every density matrix Q ,*

$$\text{Tr} \left(F \cdot \frac{1}{\text{Tr} e^{-\lambda F}} e^{-\lambda F} \right) \leq \text{Tr}(FQ) + \varepsilon,$$

as long as $\lambda \geq 1/\varepsilon \cdot S(Q \| U)$.

Proof. By the duality formula for quantum entropy (see, e.g., [Car10, Thm. 2.13]), the function $f : X \mapsto \lambda \operatorname{Tr}(FX) + S(X \parallel \mathcal{U})$ over the set of density matrices is minimized at $X^* = e^{-\lambda F} / \operatorname{Tr}(e^{-\lambda F})$. Therefore, using the fact $S(X^* \parallel \mathcal{U}) \geq 0$, we get

$$\lambda \operatorname{Tr}(FX^*) \leq f(X^*) \leq f(Q) = \lambda \operatorname{Tr}(FQ) + S(Q \parallel \mathcal{U}),$$

which implies that $\operatorname{Tr}(FX^*) \leq \operatorname{Tr}(FQ) + S(Q \parallel \mathcal{U})/\lambda \leq \operatorname{Tr}(FQ) + \varepsilon$, as desired. \square

Next we observe that one can pass from univariate approximations of e^x to approximations of e^F in the trace norm.

Lemma 4.3. *Let $\delta \in (0, 1]$ and $\tau > 0$ be given. Suppose there exists a univariate polynomial $p(x)$ such that for every $x \in [-\tau/2, \tau/2]$,*

$$|e^x - p(x)| \leq \delta e^x. \quad (4.1)$$

Then for every $F \in \mathcal{M}(H)$ with $\|F\| \leq \tau$, we have

$$\left\| \frac{e^F}{\operatorname{Tr}(e^F)} - \frac{p(F/2)^2}{\operatorname{Tr}(p(F/2)^2)} \right\|_* \leq 6\delta. \quad (4.2)$$

Proof. Under the assumptions, for every $x \in [-\tau, \tau]$, one has

$$\begin{aligned} |e^x - p(x/2)^2| &= |e^{x/2} - p(x/2)| \cdot |e^{x/2} + p(x/2)| \\ &\leq e^{x/2}(2 + \delta) |e^{x/2} - p(x/2)| \\ &\leq \delta e^x (2 + \delta) \\ &\leq 3\delta e^x, \end{aligned} \quad (4.3)$$

where the last line follows from $\delta \leq 1$. Note the elementary equality: For all $x, y, x', y' > 0$,

$$\frac{x}{y} - \frac{x'}{y'} = \frac{x - x'}{y} + \frac{y - y'}{yy'} x'. \quad (4.4)$$

Let $\lambda_1, \lambda_2, \dots, \lambda_n \in [-\tau, \tau]$ denote the eigenvalues of F . We conclude that

$$\begin{aligned} \sum_{i=1}^n \left| \frac{e^{\lambda_i}}{\sum_{i=1}^n e^{\lambda_i}} - \frac{p(\lambda_i/2)^2}{\sum_{i=1}^n p(\lambda_i/2)^2} \right| &\stackrel{(4.4)}{\leq} \sum_{i=1}^n \frac{|e^{\lambda_i} - p(\lambda_i/2)^2|}{\sum_{i=1}^n e^{\lambda_i}} + \frac{p(\lambda_i/2)^2 |\sum_{i=1}^n e^{\lambda_i} - p(\lambda_i/2)^2|}{(\sum_{i=1}^n e^{\lambda_i})(\sum_{i=1}^n p(\lambda_i/2)^2)} \\ &\stackrel{(4.3)}{\leq} 3\delta + \sum_{i=1}^n \frac{p(\lambda_i/2)^2 (3\delta \sum_{i=1}^n e^{\lambda_i})}{(\sum_{i=1}^n e^{\lambda_i})(\sum_{i=1}^n p(\lambda_i/2)^2)} \\ &\leq 6\delta. \end{aligned}$$

Since e^F and $p(F)$ are simultaneously diagonalizable, the preceding inequality is precisely our goal (4.2). \square

The following corollary of Lemma 4.3 follows by checking that the Taylor expansion of e^x satisfies the approximation guarantee (4.1).

Corollary 4.4. *For every $\varepsilon \in (0, \frac{1}{2})$ and every symmetric matrix $F \in \mathcal{M}(H)$, there is a number $k \leq 3e \left(\|F\|_\infty + \frac{\log(1/\varepsilon)}{\log \log(1/\varepsilon)} \right)$ and a univariate degree- k polynomial p_k with non-negative coefficients such that*

$$\left\| \frac{e^F}{\operatorname{Tr}(e^F)} - \frac{p_k(F/2)^2}{\operatorname{Tr}(p_k(F/2)^2)} \right\|_* \leq \varepsilon. \quad (4.5)$$

Proof. Let $p_k(x) = \sum_{t=0}^k \frac{x^t}{k!}$. By Taylor's theorem, we have

$$|e^x - p_k(x)| \leq e^x \frac{x^{k+1}}{(k+1)!}.$$

Define $\tau = \|F\|_\infty$ and choose $k = \left\lceil 3e \left(\tau + \frac{\log(1/\varepsilon)}{\log \log(1/\varepsilon)} \right) \right\rceil$ so that for $x \in [-\tau/2, \tau/2]$, we have $\frac{x^{k+1}}{(k+1)!} \leq \varepsilon/6$. Finally, apply [Lemma 4.3](#). \square

The proof of the main theorem in this section follows by combining [Lemma 4.2](#) and [Corollary 4.4](#).

Proof of Theorem 4.1. Fix $\varepsilon \in (0, \frac{1}{2})$ and $F \in \mathcal{M}(H)$, $q \in \mathcal{D}(H)$. Choose $\lambda = (2/\varepsilon) \cdot S(Q \| \mathcal{U})$ and $F' = -\lambda F$. Let p_k be the polynomial from [Corollary 4.4](#) for $k = 3e \left(\|F'\| + \frac{\log(1/\varepsilon')}{\log \log(1/\varepsilon')} \right)$ and $\varepsilon' = \varepsilon/(2\|F\|)$. Note that $k \leq O(\|F\|/\varepsilon) \cdot S(Q \| \mathcal{U}) + O\left(\frac{\log 1/\varepsilon}{\log \log 1/\varepsilon}\right)$. Moreover,

$$\begin{aligned} \text{Tr} \left(F \cdot \frac{1}{\text{Tr}(p_k(F'/2)^2)} p_k(F'/2)^2 \right) &\leq \text{Tr} \left(F \cdot \frac{1}{\text{Tr}(e^{F'})} e^{F'} \right) + \varepsilon' \cdot \|F\| && \text{(by Corollary 4.4)} \\ &\leq \text{Tr}(FQ) + \frac{\varepsilon}{2} + \varepsilon' \cdot \|F\| && \text{(by Lemma 4.2)} \end{aligned}$$

Since $\varepsilon/2 + \varepsilon' \cdot \|F\| \leq \varepsilon$, the polynomial $p(x) = p_k(-\lambda x/2)$ satisfies the desired bound

$$\text{Tr} \left(F \cdot \frac{1}{\text{Tr}(p(F)^2)} p(F)^2 \right) \leq \text{Tr}(FQ) + \varepsilon.$$

\square

4.2 Approximation against a family of tests

Let $\mathcal{T} \subseteq \mathcal{M}(H)$ denote a compact set of matrices, and set $\Delta(\mathcal{T}) := \sup_{A \in \mathcal{T}} \|A\|$. For $A \in \mathcal{M}(H)$, we define the associated dual gauge

$$[A]_{\mathcal{T}} \stackrel{\text{def}}{=} \sup_{B \in \mathcal{T}} \text{Tr}(BA).$$

One should think of \mathcal{T} as a set of test functionals; for $A, A' \in \mathcal{M}(H)$, the value $[A - A']_{\mathcal{T}}$ measures the extent to which A and A' are distinguishable using tests from \mathcal{T} . It is important to note that if \mathcal{T} is not centrally symmetric, then $[\cdot]_{\mathcal{T}}$ might also fail to be symmetric. For future reference, we observe that fact that for any $A \in \mathcal{M}(H)$,

$$|[A]_{\mathcal{T}}| \leq \Delta(\mathcal{T}) \|A\|_* \tag{4.6}$$

$$[A + A']_{\mathcal{T}} \leq [A]_{\mathcal{T}} + [A']_{\mathcal{T}}. \tag{4.7}$$

Our main approximation theorem asserts that, with respect to tests from a convex set \mathcal{T} , a high-entropy density operator can be well-approximated by the square of a low-degree polynomial in some element of \mathcal{T} .

Theorem 4.5 (Approximation by a low-degree square). *For every $\varepsilon \in (0, \frac{1}{2})$, the following holds. Let $\mathcal{T} \subseteq \mathcal{M}(H)$ be compact and convex, and let $Q \in \mathcal{D}(H)$ be a density matrix. Then there exists a number*

$$k \lesssim (1 + S(Q \| \mathcal{U})) \frac{\Delta(\mathcal{T})}{\varepsilon},$$

a univariate degree- k polynomial p , and an element $F \in \mathcal{T}$ such that $\text{Tr}(p(F)^2) = 1$ and

$$[Q - p(F)^2]_{\mathcal{T}} \leq \varepsilon.$$

Just as for [Theorem 4.1](#), this is proved in two steps: First we find an initial approximator of a simple form, and then we construct from that a low-degree approximator. In the next argument, it is helpful to have the following fact: If $X(t)$ is continuously differentiable matrix-valued function, then for any $\beta \in \mathbb{R}$, we have the Duhamel formula:

$$\frac{d}{dt} e^{\beta X(t)} = \int_0^\beta e^{\alpha X(t)} \frac{dX(t)}{dt} e^{(\beta-\alpha)X(t)} d\alpha. \quad (4.8)$$

This can be verified immediately by showing that both sides satisfy the differential equation

$$\frac{\partial F}{\partial \beta} = e^{\beta X} \frac{dX}{dt} + X(t)F(\beta, t)$$

with $F(0, t) = 0$ for all t . (This argument is taken from [\[Wil67\]](#).)

We will only require (4.8) for $\beta = 1$. For example, (4.8) and cyclicity of the trace yields

$$\text{Tr} \left(\frac{d}{dt} e^{X(t)} \right) = \text{Tr} \left(e^{X(t)} \frac{dX(t)}{dt} \right). \quad (4.9)$$

Denote $X'(t) = \frac{dX(t)}{dt}$. If we know that $X(t)$ is symmetric, and its eigenvalues are $\{\lambda_i\}$, then by diagonalizing in the basis of $X(t)$, we can also derive

$$\begin{aligned} \text{Tr} \left(X'(t) \frac{d}{dt} e^{X(t)} \right) &= \int_0^1 \text{Tr} \left(X'(t) e^{\alpha X(t)} X'(t) e^{(1-\alpha)X(t)} \right) d\alpha \\ &= \sum_{i,j} (X'(t))_{ij}^2 e^{\lambda_i} \int_0^1 e^{\alpha(\lambda_j - \lambda_i)} d\alpha \\ &= \sum_{i,j} (X'(t))_{ij}^2 \frac{e^{\lambda_i} - e^{\lambda_j}}{\lambda_i - \lambda_j} \quad \text{using } \int_0^1 e^{\alpha x} d\alpha = \frac{e^x - 1}{x} \\ &\leq \sum_{i,j} (X'(t))_{ij}^2 e^{\max(\lambda_i, \lambda_j)}, \end{aligned}$$

where in the final line we have used the fact that if $a \geq b$, then

$$\frac{e^a - e^b}{a - b} = \frac{e^a(1 - e^{b-a})}{a - b} \leq e^a,$$

since $e^{b-a} \geq 1 + (b-a)$. Thus we have

$$\begin{aligned} \text{Tr} \left(X'(t) \frac{d}{dt} e^{X(t)} \right) &\leq 2 \sum_i e^{\lambda_i} \sum_j (X'(t))_{ij}^2 \\ &\leq 2 \text{Tr}(e^{X(t)}) \max_i \sum_j (X'(t))_{ij}^2 \\ &= 2 \text{Tr}(e^{X(t)}) \|X'(t)\|_{2 \rightarrow \infty}^2 \\ &\leq 2 \text{Tr}(e^{X(t)}) \|X'(t)\|^2. \end{aligned} \quad (4.10)$$

Together these imply

$$\text{Tr} \left(X'(t) \frac{d}{dt} \frac{e^{X(t)}}{\text{Tr}(e^{X(t)})} \right) = \text{Tr} \left(\frac{X'(t)}{\text{Tr}(e^{X(t)})} \frac{de^{X(t)}}{dt} \right) - \frac{\text{Tr}(X'(t)e^{X(t)})}{\text{Tr}(e^{X(t)})^2} \text{Tr} \left(\frac{de^{X(t)}}{dt} \right)$$

$$\begin{aligned}
&= \operatorname{Tr} \left(\frac{X'(t)}{\operatorname{Tr}(e^{X(t)})} \frac{de^{X(t)}}{dt} \right) - \frac{\operatorname{Tr}(X'(t)e^{X(t)})}{\operatorname{Tr}(e^{X(t)})} \cdot \frac{\operatorname{Tr}(e^{X(t)}X'(t))}{\operatorname{Tr}(e^{X(t)})} \\
&= \frac{1}{\operatorname{Tr}(e^{X(t)})} \operatorname{Tr} \left(X'(t) \frac{de^{X(t)}}{dt} \right) - \left(\frac{\operatorname{Tr}(X'(t)e^{X(t)})}{\operatorname{Tr}(e^{X(t)})} \right)^2 \\
&\stackrel{(4.10)}{\leq} 2\|X'(t)\|^2. \tag{4.11}
\end{aligned}$$

We will use this for the following lemma.

Lemma 4.6 (Sparse approximation by mirror descent). *For every $\varepsilon > 0$, the following holds. Let $C \subseteq \mathcal{M}(H)$ be a compact set, and let $Q, Q_0 \in \mathcal{D}(H)$ be density matrices. If one defines $h = \lceil \frac{8}{\varepsilon^2} S(Q \| Q_0) \Delta(\mathcal{T})^2 \rceil$ then there exist $A_1, A_2, \dots, A_h \in \mathcal{T}$ such that*

$$\tilde{Q} \stackrel{\text{def}}{=} \frac{\exp \left(\log Q_0 - \frac{\varepsilon}{4\Delta(\mathcal{T})^2} \sum_{i=1}^h A_i \right)}{\operatorname{Tr} \left(\exp \left(\log Q_0 - \frac{\varepsilon}{4\Delta(\mathcal{T})^2} \sum_{i=1}^h A_i \right) \right)} \in \mathcal{D}(H) \tag{4.12}$$

satisfies

$$[Q - \tilde{Q}]_{\mathcal{T}} \leq \varepsilon. \tag{4.13}$$

Proof. Consider for $t \geq 0$, the density matrix

$$Q_t = \frac{\exp \left(\log Q_0 - \int_0^t \Lambda_s ds \right)}{\operatorname{Tr} \left(\exp \left(\log Q_0 - \int_0^t \Lambda_s ds \right) \right)},$$

where $s \mapsto \Lambda_s \in \mathcal{T}$ is any measurable function.

First, one calculates

$$\frac{d}{dt} \log Q_t = -\Lambda_t - \operatorname{Id} \frac{d}{dt} \log \operatorname{Tr} \left(\exp \left(\log Q_0 - \int_0^t \Lambda_s ds \right) \right) \tag{4.14}$$

Now, we have

$$\frac{d}{dt} \log \operatorname{Tr} \left(\exp \left(\log Q_0 - \int_0^t \Lambda_s ds \right) \right) = \frac{\frac{d}{dt} \operatorname{Tr} \left(\exp \left(\log Q_0 - \int_0^t \Lambda_s ds \right) \right)}{\operatorname{Tr} \left(\exp \left(\log Q_0 - \int_0^t \Lambda_s ds \right) \right)} \stackrel{(4.9)}{=} -\operatorname{Tr}(\Lambda_t Q_t),$$

and thus

$$\frac{d}{dt} S(Q \| Q_t) = \operatorname{Tr} \left(Q \frac{d}{dt} \log Q_t \right) = -\operatorname{Tr}(\Lambda_t Q) + \operatorname{Tr}(Q) \operatorname{Tr}(\Lambda_t Q_t) = -\operatorname{Tr}(\Lambda_t(Q - Q_t)), \tag{4.15}$$

where in the final line we have used $\operatorname{Tr}(Q) = 1$.

Let $T = \frac{2}{\varepsilon} S(Q \| Q_0)$. Suppose the map $t \mapsto \Lambda_t \in \mathcal{T}$ is such that

$$\operatorname{Tr}(\Lambda_t(Q - Q_t)) > \frac{\varepsilon}{2} \quad \forall t \in [0, T]. \tag{4.16}$$

Then from (4.15) and (4.16), we arrive at

$$S(Q \| Q_T) < S(Q \| Q_0) - \frac{\varepsilon}{2} T = 0,$$

which contradicts the fact that $S(Q \| Q_T) \geq 0$.

Finally, we define the elements $A_1, \dots, A_h \in \mathcal{T}$ and corresponding approximators $\tilde{Q}_0, \tilde{Q}_1, \dots, \tilde{Q}_h \in \mathcal{D}(H)$ inductively. Define, for $i = 0, 1, 2, \dots, h$, the times $t_i = i \frac{\varepsilon}{4\Delta(\mathcal{T})^2}$. We will choose the map $t \mapsto \Lambda_t$ and put $\tilde{Q}_i = Q_{t_i}$.

We begin by setting $\tilde{Q}_0 = \mathcal{U}$ and $\Lambda_0 = 0$. Now if $[Q - \tilde{Q}_i]_{\mathcal{T}} \leq \varepsilon$, then we are done. Otherwise, let $A_{i+1} \in \mathcal{T}$ be such that

$$\text{Tr}(A_{i+1}Q - A_{i+1}\tilde{Q}_i) > \varepsilon, \quad (4.17)$$

and define $\Lambda_t = A_{i+1}$ for $t \in (t_i, t_{i+1}]$.

Finally, observe that for $t \in (t_i, t_{i+1})$, we have

$$\frac{d}{dt} \text{Tr}(A_{i+1}Q - A_{i+1}Q_t) = \frac{d}{dt} \text{Tr}(\Lambda_t Q_t) \stackrel{(4.11)}{\geq} -2\|\Lambda_t\|^2.$$

where we have used the fact that $\Lambda_t = -\frac{d}{dt} \log \left(e^{\log Q_0 - \int_0^t \Lambda_s ds} \right)$. We conclude that

$$\begin{aligned} \text{Tr}(A_{i+1}Q - A_{i+1}Q_{t_{i+1}}) &\geq \text{Tr}(A_{i+1}Q - A_{i+1}Q_{t_i}) - 2\|\Lambda_t\|^2(t_{i+1} - t_i) \\ &\geq \text{Tr}(A_{i+1}Q - A_{i+1}Q_{t_i}) - \frac{\varepsilon}{2} \\ &> \frac{\varepsilon}{2}, \end{aligned}$$

using (4.17). Thus we either find an approximator \tilde{Q}_i for some $i = 0, 1, \dots, h$ satisfying (4.13) or (4.16) holds. But we have already seen that the latter possibility cannot happen. Observe that the approximators \tilde{Q}_i are all of the desired form (4.12). \square

Proof of Theorem 4.5. First, we apply Lemma 4.6 with $Q_0 = \mathcal{U}$ to obtain an approximation \tilde{Q} of the form

$$\tilde{Q} = \frac{e^{\lambda F}}{\text{Tr}(e^{\lambda F})}$$

with $|\lambda| \leq 1 + \frac{1}{\delta} S(Q \| U)$ and which satisfies

$$[Q - \tilde{Q}]_{\mathcal{T}} \leq \varepsilon/2. \quad (4.18)$$

Note here that since \mathcal{T} is assumed to be convex, we have $F \in \mathcal{T}$ (see the form of (4.12)).

Then we apply Corollary 4.4 to λF to obtain a degree- k polynomial p_k such that

$$\left\| \tilde{Q} - \frac{p_k(\lambda F/2)^2}{\text{Tr}(p_k(\lambda F/2)^2)} \right\|_* \leq \frac{\varepsilon}{2\Delta(\mathcal{T})}, \quad (4.19)$$

where

$$k \lesssim |\lambda|\Delta(\mathcal{T}) + \frac{\log(\Delta(\mathcal{T})/\varepsilon)}{\log \log(\Delta(\mathcal{T})/\varepsilon)} \lesssim (1 + S(Q \| \mathcal{U})) \frac{\Delta(\mathcal{T})}{\varepsilon}.$$

Thus we conclude that

$$\begin{aligned} \left[Q - \frac{p_k(\lambda F/2)^2}{\text{Tr}(p_k(\lambda F/2)^2)} \right]_{\mathcal{T}} &\stackrel{(4.7)}{\leq} [Q - \tilde{Q}]_{\mathcal{T}} + \left[\tilde{Q} - \frac{p_k(\lambda F/2)^2}{\text{Tr}(p_k(\lambda F/2)^2)} \right]_{\mathcal{T}} \\ &\stackrel{(4.6)}{\leq} [Q - \tilde{Q}]_{\mathcal{T}} + \Delta(\mathcal{T}) \left\| \frac{p_k(\lambda F/2)^2}{\text{Tr}(p_k(\lambda F/2)^2)} - \tilde{Q} \right\|_* \\ &\stackrel{(4.18) \wedge (4.19)}{\leq} \varepsilon. \quad \square \end{aligned}$$

4.2.1 Junta approximation

We record here the following application to “classical” functions by restricting [Lemma 4.6](#) to the diagonal case. If X is a finite set, and \mathcal{T} is a collection of real-valued functions on X , we extend the notation $\Delta(\mathcal{T}) = \sup_{g \in \mathcal{T}} \|g\|_\infty$. If μ is a measure on X , and $f : X \rightarrow \mathbb{R}_+$ satisfies $\mathbb{E}_\mu f = 1$, we abuse notation by writing

$$D(f \parallel \mu) = \mathbb{E}_\mu [f \log f].$$

for the relative entropy between $f\mu$ and μ . We will also allow ourselves to conflate μ with the corresponding density by writing $\mu(x)$ for $\mu(\{x\})$ and $x \in X$. One should note that an analog of [Lemma 4.6](#) for the special case of probability distributions (instead of density matrices) can be proved exactly along the same lines, but without the use of matrix inequalities.

Corollary 4.7 (Sparse approximation of functions by mirror descent). *For every $\varepsilon > 0$, the following holds. Let X be a finite set equipped with a probability measure μ . Let $\mathcal{T} \subseteq L^2(X, \mu)$ be a compact set of functions, and let $f : X \rightarrow \mathbb{R}_+$ be such that $\mathbb{E}_\mu f = 1$. If one defines $h = \lceil \frac{8}{\varepsilon^2} D(f \parallel \mu) \Delta(\mathcal{T})^2 \rceil$ then there exist functions $g_1, g_2, \dots, g_h \in \mathcal{T}$ such that*

$$\tilde{f} \stackrel{\text{def}}{=} \frac{\exp\left(\frac{\varepsilon}{4\Delta(\mathcal{T})^2} \sum_{i=1}^h g_i\right)}{\sum_{x \in X} \exp\left(\frac{\varepsilon}{4\Delta(\mathcal{T})^2} \sum_{i=1}^h g_i(x)\right) \mu(x)} \quad (4.20)$$

so that $\mathbb{E}_\mu \tilde{f} = 1$, and for every $g \in \mathcal{T}$,

$$\mathbb{E}_{x \sim \mu} g(x)(f(x) - \tilde{f}(x)) \leq \varepsilon. \quad (4.21)$$

Proof. H the Euclidean space $\mathbb{R}^{\{0,1\}^n}$, and let $\{e_x : x \in \{0,1\}^n\}$ be an orthonormal basis of H . We will represent f by the diagonal matrix $M(f) \in \mathcal{D}(H)$ defined by

$$M(f) = \sum_{x \in \{0,1\}^n} f(x) e_x e_x^T \mu(x).$$

We also lift each test g to a matrix $M(g) = \sum_{x \in \{0,1\}^n} g(x) e_x e_x^T$ and the set $M(\mathcal{T})$ now denotes a class of matrix tests.

Furthermore, we write

$$Q_0 = \sum_{x \in \{0,1\}^n} e_x e_x^T \mu(x)$$

so that $S(M(f) \parallel Q_0) = D(f \parallel \mu)$.

Applying [Lemma 4.6](#) yields an approximation \tilde{Q} to $M(f)$ of the form

$$\tilde{Q} = Q_0 \cdot M,$$

where M is a diagonal matrix. Furthermore, by construction, the function $\tilde{f} : \{0,1\}^n \rightarrow \mathbb{R}$ given by $\tilde{f}(x) = \langle e_x, M e_x \rangle$ has the form [\(4.20\)](#). Finally, the approximation guarantee $[M(f) - \tilde{Q}]_{M(\mathcal{T})} \leq \varepsilon$ is precisely [\(4.21\)](#). \square

We now apply the preceding corollary to prove an approximation-by-juntas theorem. An essentially equivalent result for Boolean domains is proved in [\[CLRS13\]](#), but it is instructive to see that it falls easily out of the learning framework. Fix $n \geq 1$ and a finite set X . We recall that for a subset $S \subseteq \{1, \dots, n\}$, a function $f : X^n \rightarrow \mathbb{R}$ is called an S -junta if f only depends (at most) on the coordinates in S . In other words, for all $x, x' \in X^n$, if $x|_S = x'|_S$ then $f(x) = f(x')$. We say that f is a k -junta if it is an S -junta for a set with $|S| = k$.

Theorem 4.8 (Junta approximation). *Let X be an arbitrary finite set, and let μ denote a probability measure on X^n . Consider a non-negative function $f : X^n \rightarrow \mathbb{R}_+$ with $\mathbb{E}_\mu f = 1$, and let \mathcal{T} be a collection of k -juntas. Then for every $\varepsilon > 0$, there exists a non-negative k' -junta $\tilde{f} : X^n \rightarrow \mathbb{R}_+$ with $\mathbb{E}_\mu \tilde{f} = 1$, where*

$$k' \lesssim \frac{k}{\varepsilon^2} D(f \parallel \mu) \Delta(\mathcal{T})^2,$$

and such that for every $g \in \mathcal{T}$,

$$\mathbb{E}_{x \sim \mu} g(x) (f(x) - \tilde{f}(x)) \leq \varepsilon. \quad (4.22)$$

Proof. Applying Corollary 4.7 yields an approximation \tilde{f} . One simply notes that from (4.20), \tilde{f} is an hk -junta where $h \lesssim \frac{1}{\varepsilon^2} D(f \parallel \mu) \Delta(\mathcal{T})^2$ \square

5 The correlation polytope

Recall the correlation polytope $\text{CORR}_n \subseteq \mathbb{R}^{n^2}$ given by

$$\text{CORR}_n = \text{conv}(\{xx^T : x \in \{0, 1\}^n\}).$$

This polytope is also known as the *Boolean quadric polytope* [Pad89] for the following reason.

Proposition 5.1 (Restatement of Proposition 1.11). *If $f : \{0, 1\}^m \rightarrow \mathbb{R}_+$ is a nonnegative quadratic function over $\{0, 1\}^m$, then for any $n \geq m$, M_n^f is a submatrix of some slack matrix associated to CORR_n .*

Proof. Let $\langle A, B \rangle = \text{Tr}(A^T B)$ denote the Frobenius inner product on \mathbb{R}^{n^2} . Suppose that $f(x) = \sum_{i \leq j} a_{ij} x_i x_j + a_0 \geq 0$ for all $x \in \{0, 1\}^n$. We claim that this gives a valid linear inequality for CORR_n as follows: For all $x \in \{0, 1\}^n$,

$$f(x) = \langle A, xx^T \rangle + a_0 \geq 0,$$

where A is the matrix $A = (a_{ij})$. Since this inequality holds at the vertices, it holds for all of CORR_n . \square

We now recall the relationship between the correlation, cut, TSP, and stable set polytopes. The first fact is from [DS90], while the second two are taken from [FMP⁺12].

Proposition 5.2. *For every $n \geq 1$, the following hold:*

1. CORR_n is linearly isomorphic to CUT_{n+1} .
2. There exists a number $a_n \leq O(n^2)$ such that some face of TSP_{a_n} linearly projects to CORR_n .
3. There exists a graph H_n on $b_n \leq O(n^2)$ vertices such that some face of $\text{STAB}_{b_n}(H_n)$ linearly projects to CORR_n .

5.1 Positive semidefinite rank

We will now prove a lower bound on the psd rank of CORR_n . Our first goal is to construct a suitable family of pseudo-densities. We will employ Grigoriev's work [Gri01a] on degree lower bounds for Positivstellensatz calculus refutations. The primary difficulty will be in expressing Grigoriev's lower bound using a pseudo-density of small norm.

Theorem 5.3. Fix an odd integer $m \geq 3$. There exists a degree- m pseudo-density $D : \{0, 1\}^m \rightarrow \mathbb{R}$ such that

$$\mathbb{E}_x D(x) \left(\sum_{i=1}^m x_i - \frac{m}{2} \right)^2 = 0,$$

and

$$\|D\|_\infty \leq m^{3/2}.$$

Proof. Grigoriev constructs a linear functional \mathcal{G} on the space of m -variate real polynomials modulo the ideal \mathcal{I} generated by $\{X_i^2 - X_i : i = 1 \in [m]\}$:

$$\mathcal{G} : \mathbb{R}[X_1, \dots, X_m]/\mathcal{I} \rightarrow \mathbb{R}.$$

His functional satisfies

$$\mathcal{G}(p(X)^2) \geq 0 \quad \forall p \in \mathbb{R}[X_1, \dots, X_m]/\mathcal{I}, \quad \deg(p) \leq m/2, \quad (5.1)$$

and

$$\mathcal{G}\left(\left(\sum_{i=1}^m X_i - \frac{m}{2}\right)^2\right) = 0. \quad (5.2)$$

The functional is uniquely defined by the values

$$\mathcal{G}(X^S) \stackrel{\text{def}}{=} \frac{\binom{m/2}{|S|}}{\binom{m}{|S|}},$$

for each multilinear monomial $X^S = \prod_{i \in S} X_i$ with $S \subseteq [m]$. Observe that $m/2$ is not an integer and the (generalized) binomial coefficient $\binom{m/2}{k}$ is defined using the formal expression

$$\binom{r}{k} = \frac{r \cdot (r-1) \cdots (r-k+1)}{k \cdot (k-1) \cdots 1}.$$

It is easy to check that \mathcal{G} satisfies (5.2):

$$\begin{aligned} \mathcal{G}\left(\left(\sum_{i=1}^m X_i - \frac{m}{2}\right)^2\right) &= \sum_{i=1}^m \mathcal{G}(X_i^2) + 2 \sum_{i \neq j} \mathcal{G}(X_i X_j) - m \sum_{i=1}^m \mathcal{G}(X_i) + \frac{m^2}{4} \\ &= \frac{m}{2} + m(m-1) \frac{m-2}{4(m-1)} - m \frac{m}{2} + \frac{m^2}{4} = 0. \end{aligned}$$

Grigoriev shows that \mathcal{G} satisfies (5.1) [Gri01a, Lem. 1.4].

We will construct a pseudo-density $D : \{0, 1\}^m \rightarrow \mathbb{R}$ such that $\mathbb{E}_x D(x)p(x) = \mathcal{G}(p(X_1, \dots, X_m))$ for every multilinear polynomial p . Observe that \mathcal{G} is invariant under permutation of variables $\{X_1, \dots, X_m\}$. For $w = 0, 1, \dots, m$, let c_w denote the unique degree m polynomial such that,

$$c_w(t) = \begin{cases} 1 & \text{if } t = w \\ 0 & \text{if } t \in \{0, 1, \dots, m\} \setminus \{w\} \end{cases}$$

We claim that for any univariate real polynomial p with $\deg(p) \leq m$,

$$\sum_{w=0}^m p(w) \cdot c_w(t) = p(t). \quad (5.3)$$

Both sides of the claimed identity are univariate polynomials in t of degree at most m and agree with each other on the $m + 1$ points given by $t \in \{0, 1, \dots, m\}$. Hence, the two polynomials are identically equal.

For each $x \in \{0, 1\}^m$, let $|x|$ denote its hamming weight, and define

$$D(x) \stackrel{\text{def}}{=} 2^m \cdot \frac{c_{|x|}(m/2)}{\binom{m}{|x|}}.$$

We claim that D satisfies $\mathbb{E}_x D(x)p(x) = \mathcal{G}(p(X))$ for every polynomial multilinear real polynomial p . To see this, consider any monomial $x^S = \prod_{i \in S} x_i$ with $S \subseteq [m]$. Put $\ell = |S|$. Then we have:

$$\begin{aligned} \mathbb{E}_x D(x)x^S &= \mathbb{E}_x D(x) \cdot \frac{1}{\binom{m}{\ell}} \left(\sum_{T \subseteq [m], |T|=\ell} x^T \right) \quad (\text{symmetry of } D) \\ &= \mathbb{E}_x D(x) \cdot \frac{1}{\binom{m}{\ell}} \binom{|x|}{\ell} \\ &= \sum_{w=0}^m \frac{\binom{m}{w}}{2^m} \mathbb{E}_x \left[D(x) \cdot \frac{1}{\binom{m}{\ell}} \binom{|x|}{\ell} \mid |x| = w \right] \\ &= \sum_{w=0}^m c_w(m/2) \frac{\binom{w}{\ell}}{\binom{m}{\ell}} \\ &= \frac{\binom{m/2}{\ell}}{\binom{m}{\ell}} \\ &= \mathcal{G}(X^S) \quad (\text{using (5.3) with } p(t) = \binom{t}{\ell}). \end{aligned}$$

Finally, in order to bound $\|D\|_\infty$, observe that the polynomials $c_w(t)$ are given by the interpolation formula

$$c_w(t) = \frac{\prod_{a=0, a \neq w}^m (t - a)}{\prod_{a=0, a \neq w}^m (w - a)}.$$

For an $x \in \{0, 1\}^m$ with $|x| = w$ we have,

$$\begin{aligned} |D(x)| &= \left| 2^m \cdot \frac{c_w(m/2)}{\binom{m}{w}} \right| = \left| \frac{\prod_{a=0, a \neq w}^m (m - 2a)}{\prod_{a=0, a \neq w}^m (w - a)} \cdot \frac{1}{\binom{m}{w}} \right| \\ &= \left| \frac{\prod_{a=0, a \neq w}^m (m - 2a)}{(w - 1)!(m - w)!} \cdot \frac{1}{\binom{m}{w}} \right| \\ &= m \cdot \frac{1}{2^{m-1}} \cdot \frac{(m-1)}{\binom{(m-1)/2}} \cdot \frac{w}{|m - 2w|} \\ &\leq \frac{m}{\sqrt{\frac{3}{2}(m-1) + 1}} \left(\frac{m+1}{2} \right) \\ &\leq m^{3/2}, \end{aligned}$$

where in the last step we have used Stirling's approximation for the inequality $\binom{m-1}{(m-1)/2} \leq 2^{m-1} / \sqrt{\frac{3}{2}(m-1) + 1}$, valid for $m \geq 3$. \square

Theorem 5.4. *There is a constant $\alpha > 0$ such that for every $n \geq 1$,*

$$\text{rk}_{\text{psd}}(\text{CORR}_n) \geq 2^{\alpha n^{2/13}}.$$

Proof. For odd integers $m \geq 1$, define $f : \{0, 1\}^m \rightarrow \mathbb{R}_+$ by

$$f(x) = \frac{1}{m^2} \left(\left(\sum_{i=1}^m x_i - \frac{m}{2} \right)^2 - \frac{1}{4} \right), \quad (5.4)$$

and let $M_n^f : \binom{[n]}{m} \times \{0, 1\}^n \rightarrow \mathbb{R}_+$ be given by $M_n^f(S, x) = f(x_S)$ as in (1.3).

By [Theorem 5.3](#), there exists a degree- m pseudo-density $D : \{0, 1\}^m \rightarrow \mathbb{R}$ such that $\mathbb{E}_x D(x) f(x) = -\frac{1}{4m^2}$ and $\|D\|_\infty \leq m^{3/2}$. Fix $\varepsilon = 1/(4m^2)$ and $d = m$ and apply [Theorem 3.8](#) to conclude that there is a constant $\alpha' > 0$ such that for $n \geq 2m$, we have

$$\text{rk}_{\text{psd}}(M_n^f) \geq \left(\frac{\alpha' n}{m^{13/2} \log n} \right)^{m/4} \cdot m^{-21/4} \cdot m^{-1/2}$$

Choosing $n \geq \frac{2}{\alpha'} m^{13/2} \log n$, an easy calculation shows that

$$\text{rk}_{\text{psd}}(M_n^f) \geq 2^{\Omega(n^{2/13})}.$$

By [Proposition 5.1](#), we have $\text{rk}_{\text{psd}}(\text{CORR}_n) \geq \text{rk}_{\text{psd}}(M_n^f)$, completing the proof. \square

6 Optimality of low-degree sum-of-squares for max CSPs

Constraint satisfaction problems form a broad class of discrete optimization problems that include, for example, MAX CUT and MAX 3-SAT. For simplicity of presentation, we will focus on constraint satisfaction problems with a boolean alphabet, though similar ideas extend to larger domains (see an analogous generalization in [Section 7](#)). We begin our presentation with a formal definition of semidefinite programming relaxations for max-CSPs.

6.1 The SDP approximation model

In order to write an SDP relaxation for a max-CSP, one needs to *linearize* the objective function. For $n \in \mathbb{N}$, let $\text{MAX-}\Pi_n$ be the set of MAX- Π instances on n variables. An *SDP-relaxation of size r* for $\text{MAX-}\Pi_n$ consists of the following.

Linearization: Let r be a natural number. For every $\mathfrak{J} \in \text{MAX-}\Pi_n$, we associate a vector $\tilde{\mathfrak{J}} \in \mathbb{R}^{r \times r}$ and for every assignment $x \in \{0, 1\}^n$, we associate a point $\tilde{x} \in \mathbb{R}^{r \times r}$, such that $\mathfrak{J}(x) = \langle \tilde{\mathfrak{J}}, \tilde{x} \rangle$ for all $\mathfrak{J} \in \text{MAX-}\Pi_n$ and all $x \in \{0, 1\}^n$.

Feasible region: The feasible region is a closed, convex (possibly unbounded) spectrahedron $\mathcal{S} \subseteq \mathbb{R}^{r \times r}$ described as the intersection of the cone of $r \times r$ PSD matrices with an affine linear subspace:

$$\mathcal{S} = \{y \in \mathbb{R}^{r \times r} \mid Ay = b, y \in \mathcal{S}_r^+\},$$

such that $\tilde{x} \in \mathcal{S}$ for all assignments $x \in \{0, 1\}^n$. Note that the spectrahedron \mathcal{S} is independent of the instance \mathfrak{J} of $\text{MAX-}\Pi_n$.

Given an instance $\mathfrak{J} \in \text{MAX-}\Pi_n$, the SDP relaxation \mathcal{S} has value

$$\mathcal{S}(\mathfrak{J}) \stackrel{\text{def}}{=} \max_{y \in \mathcal{S}} \langle \tilde{\mathfrak{J}}, y \rangle.$$

Since $\tilde{x} \in \mathcal{S}$ for all assignments $x \in \{0, 1\}^n$ and $\langle \tilde{\mathfrak{J}}, \tilde{x} \rangle = \mathfrak{J}(x)$, we have $\mathcal{S}(\mathfrak{J}) \geq \text{opt}(\mathfrak{J})$ for all instances $\mathfrak{J} \in \text{MAX-}\Pi_n$.

Low-degree sum-of-squares relaxations. We will now describe the low-degree sum-of-squares relaxation as it applies to a max-CSP. Let Π be a max-CSP with arity k . Given an instance \mathfrak{J} of $\text{MAX-}\Pi_n$, we recall that we think of it as a function $\mathfrak{J} : \{0, 1\}^n \rightarrow \mathbb{R}$ given by $\mathfrak{J}(x) = \frac{1}{m} \sum_{i=1}^m P_i(x)$ where $\{P_i\}_{i \in [m]}$ are the constraints in \mathfrak{J} . Define the cone $C_d^{\text{sos}} \subseteq \mathbb{R}^{\{0,1\}^n}$ as the cone generated by squares of polynomials of degree at most $d/2$, i.e.,

$$C_d^{\text{sos}} = \text{Cone}(\{g^2 \mid g : \{0, 1\}^n \rightarrow \mathbb{R}, \deg(g) \leq d/2\}).$$

The *degree- d sos relaxation* for \mathfrak{J} is given by

$$\overline{\text{sos}}_d(\mathfrak{J}) \stackrel{\text{def}}{=} \min \{c \mid c - \mathfrak{J} \in C_d^{\text{sos}}\} \quad (6.1)$$

We will now write the dual formulation of the above semidefinite program to expose the underlying spectrahedron and linearization. The dual of (6.1) can be written as,

$$\begin{aligned} \overline{\text{sos}}_d(\mathfrak{J}) &= \max_{D: \{0,1\}^n \rightarrow \mathbb{R}} \langle D, \mathfrak{J} \rangle \\ &\text{subject to } \langle D, 1 \rangle = 1, \\ &\langle D, h \rangle \geq 0 \quad \forall h \in C_d^{\text{sos}}. \end{aligned} \quad (6.2)$$

The function $D : \{0, 1\}^n \rightarrow \mathbb{R}$ is referred to as a pseudo-density over $\{0, 1\}^n$, since it satisfies that for every degree $d/2$ function g , $\mathbb{E}_x D(x)g^2(x) \geq 0$.

Notice that all the constraints on the pseudodensity $D : \{0, 1\}^n \rightarrow \mathbb{R}$ correspond to inner products with functions of degree at most d . Hence, without loss of generality, we may assume $\deg(D) \leq d$. Alternately, the convex program (6.2) can be written succinctly in terms of the low-degree part of D . We will now carry this out explicitly and thereby identify the feasible region associated with the degree- d sos relaxation.

To this end, set $\mathcal{F} := \{A : A \subseteq [n], |A| \leq d/2\}$ and let $r = |\mathcal{F}| \leq \sum_{i=0}^{d/2} \binom{n}{i}$. Recall that $\mathcal{S}_r^+ \subseteq \mathbb{R}^{r \times r}$ is the cone of $r \times r$ PSD matrices. We will index the matrices in \mathcal{S}_r^+ using elements of \mathcal{F} in the natural way. Define a matrix $Y : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}$ as follows,

$$Y(A, B) = \left\langle D, \prod_{i \in A \cup B} x_i \right\rangle.$$

By definition of Y , it is clear that $Y(A, B) = Y(B, A) = Y(A \cup B, \emptyset)$ for all $A, B \in \mathcal{F}$. Moreover, we have $Y(\emptyset, \emptyset) = \langle D, 1 \rangle = 1$. Furthermore, the matrix Y is PSD since, for all $\hat{g} : \mathcal{F} \rightarrow \mathbb{R}$, we have

$$\begin{aligned} \langle \hat{g}, Y \hat{g} \rangle &= \sum_{A, B \in \mathcal{F}} \hat{g}_A \hat{g}_B Y(A, B) \\ &= \left\langle D, \sum_{A, B \in \mathcal{F}} \hat{g}_A \hat{g}_B \prod_{i \in A \cup B} x_i \right\rangle \end{aligned}$$

$$\begin{aligned}
&= \left\langle D, \left(\sum_{A \in \mathcal{F}} \hat{g}_A \prod_{i \in A} x_i \right)^2 \right\rangle && \text{using } x_i^2 = x_i \quad \forall i \in [n], x \in \{0, 1\}^n \\
&\geq 0
\end{aligned}$$

where the final inequality used the fact that $\langle D, g^2 \rangle \geq 0$ for all functions g with $\deg(g) \leq d/2$.

From the above discussion, it is clear that the *feasible region* of the degree d -sos relaxation (6.2) corresponds to the spectrahedron,

$$\mathcal{S} \stackrel{\text{def}}{=} \{Y \in \mathbb{R}^{r \times r} \mid Y \in \mathcal{S}_r^+, Y(\emptyset, \emptyset) = 1 \text{ and } Y_{A,B} = Y_{B,A} = Y_{A \cup B, \emptyset} \quad \forall A, B \in \mathcal{F}\}$$

Now we describe the *linearization* associated with the degree- d sos relaxation. For every assignment $x \in \{0, 1\}^n$, associate the matrix $\tilde{x} : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}$ given by

$$\tilde{x}(A, B) \stackrel{\text{def}}{=} \prod_{i \in A \cup B} x_i. \quad (6.3)$$

By definition, we have $\tilde{x}(A, B) = \tilde{x}(B, A) = \tilde{x}(A \cup B, \emptyset)$ and $\tilde{x}(\emptyset, \emptyset) = 1$. Moreover, the matrix \tilde{x} is positive semidefinite since it can be written as $\tilde{x} = XX^T$ wherein $X : \mathcal{F} \rightarrow \mathbb{R}$ is given by $X(A) = \prod_{i \in A} x_i$. Therefore, for each assignment x , we have $\tilde{x} \in \mathcal{S}$.

Finally, given an instance $\mathfrak{J} \in \text{MAX-}\Pi_n$ its linearization $\tilde{\mathfrak{J}}$ is written as follows. Fix $d \geq 2\lceil k/2 \rceil$, and for every subset $S \subseteq [d]$ with $|S| \leq k$, define a disjoint union $S = A_S \cup B_S$ where A_S contains (up to) the $\lceil k/2 \rceil$ smallest elements of S , and B_S contains the rest (or is empty).

Each constraint P_0 in \mathfrak{J} is of the form $P_0(X) = P(X_{i_1}, \dots, X_{i_k})$ for a predicate $P : \{0, 1\}^n \rightarrow \{0, 1\}$ in Π . Therefore the function $\mathfrak{J} : \{0, 1\}^n \rightarrow \mathbb{R}$ given by $\mathfrak{J}(x) = \frac{1}{m} \sum_{i=1}^m P_i(x)$ can be expressed as a degree- k multilinear polynomial in x , i.e.,

$$\mathfrak{J}(x) = \sum_{A \subseteq [n], |A| \leq k} \hat{\mathfrak{J}}_A \prod_{i \in A} x_i.$$

The linearization $\tilde{\mathfrak{J}} : \mathcal{F} \times \mathcal{F} \rightarrow \mathbb{R}$ is given by,

$$\tilde{\mathfrak{J}}(A, B) \stackrel{\text{def}}{=} \begin{cases} \hat{\mathfrak{J}}_S & \text{if } A = A_S, B = B_S \\ 0 & \text{otherwise} \end{cases} \quad (6.4)$$

From (6.3) and (6.4), for every instance $\mathfrak{J} \in \text{MAX-}\Pi_n$ and every assignment $x \in \{0, 1\}^n$ we have

$$\langle \tilde{\mathfrak{J}}, \tilde{x} \rangle = \sum_{A, |A| \leq d/2} \hat{\mathfrak{J}}_A \prod_{i \in A} x_i = \mathfrak{J}(x).$$

Now the degree- d sos relaxation corresponding to an instance $\mathfrak{J} \in \text{MAX-}\Pi_n$ in (6.1) and (6.2) can be equivalently formulated as

$$\overline{\text{sos}}_d(\mathfrak{J}) \stackrel{\text{def}}{=} \max_{y \in \mathcal{S}} \langle \tilde{\mathfrak{J}}, y \rangle.$$

(c, s) -approximations. For $0 \leq s \leq c \leq 1$, a sequence of SDP relaxations $\{\mathcal{S}_n\}_{n=1}^\infty$ for $\text{MAX-}\Pi$ is said to achieve a (c, s) -approximation to $\text{MAX-}\Pi$ if for each $n \in \mathbb{N}$ and every instance \mathfrak{J} of $\text{MAX-}\Pi_n$ with $\text{opt}(\mathfrak{J}) \leq s$, we have $\mathcal{S}_n(\mathfrak{J}) \leq c$. In order to study (c, s) -approximations for $\text{MAX-}\Pi$, we recall (from Proposition 1.13) the set of matrices $\{M_{c,s}^{n,\Pi}\}_{n=1}^\infty$ associated with it, defined as:

$$M_{c,s}^{n,\Pi}(\mathfrak{J}, x) = c - \mathfrak{J}(x),$$

where the first index of $M_{c,s}^{n,\Pi}$ ranges over all instances on n variables satisfying $\text{opt}(\mathfrak{J}) \leq s$. A simple consequence of Proposition 1.10 is the following.

Proposition 6.1. *There exists a sequence of SDP relaxations \mathcal{S}_n of size r_n achieving a (c, s) -approximation to $\text{MAX-}\Pi_n$ if and only if $\text{rk}_{\text{psd}}(M_{c,s}^{n,\Pi}) \leq r_n$.*

6.2 General SDPs vs. sum-of-squares

Our main theorem is that general SDP relaxations for max-CSPs are no more powerful than low-degree sum-of-squares relaxations in the polynomial-size regime.

Theorem 6.2. *Fix a positive number $d \in \mathbb{N}$, and a k -ary CSP $\text{MAX-}\Pi$ with $d \geq 2\lceil k/2 \rceil$. Suppose that the degree- d sos relaxation cannot achieve a (c, s) -approximation for $\text{MAX-}\Pi$. Then no sequence of SDP relaxations of size at most $o\left(\left(\frac{n}{\log n}\right)^{d/4}\right)$ can achieve a (c, s) -approximation for $\text{MAX-}\Pi$.*

Proof. Given that the degree- d sos relaxation cannot achieve a (c, s) -approximation, there exists an instance \mathfrak{J} of $\text{MAX-}\Pi_m$ for some m such that $\text{opt}(\mathfrak{J}) \leq s$ but $\text{deg}_{\text{sos}}(c - \mathfrak{J}) > d$.

By [Proposition 6.1](#), it is sufficient to lower bound the psd rank of the matrix $M_{c,s}^{n,\Pi}$. Fix $f = c - \mathfrak{J}$ and define the matrix $M_n^f : \binom{[m]}{d} \times \{0, 1\}^n \rightarrow [0, 1]$ as $M_n^f(S, x) \stackrel{\text{def}}{=} f(x_S)$. Since M_n^f is a submatrix of $M_{c,s}^{n,\Pi}$, we have $\text{rk}_{\text{psd}}(M_{c,s}^{n,\Pi}) \geq \text{rk}_{\text{psd}}(M_n^f)$. By [Theorem 1.8](#), for some constant $C \geq 1$, we have

$$\text{rk}_{\text{psd}}(M_n^f) \geq \left(\frac{n}{C \log n}\right)^{d/4}.$$

This implies that no sequence of SDP relaxations of size at most $o\left(\left(\frac{n}{\log n}\right)^{d/4}\right)$ can achieve a (c, s) -approximation for $\text{MAX-}\Pi$. \square

For a stronger quantitative bound, we require the following simple fact.

Fact 6.3. *For every positive even integer d , every degree- d pseudo-density $D : \{0, 1\}^m \rightarrow \mathbb{R}$ and every subset $\alpha \subseteq [m]$, $|\alpha| \leq d$, we have*

$$|\mathbb{E}_x D(x) \chi_\alpha(x)| \leq 1.$$

Proof. Write $\chi_\alpha = \chi_A \chi_B$ for some A, B with $|A|, |B| \leq d/2$ and observe that

$$\mathbb{E}_x D(x) \chi_\alpha(x) = \mathbb{E}_x D(x) \left(1 - \frac{(\chi_A - \chi_B)^2}{2}\right) \leq \mathbb{E}_x D(x) \cdot 1 = 1,$$

where we used the fact that $\mathbb{E}_x D(x) p(x)^2 \geq 0$ whenever $\text{deg}(p) \leq d/2$. Using $\chi_\alpha = \frac{1}{2}(\chi_A + \chi_B)^2 - 1$, we get the other direction of the inequality. \square

Theorem 6.4. *Fix a k -ary CSP $\text{MAX-}\Pi$ and a monotone increasing function $d : \mathbb{N} \rightarrow \mathbb{N}$ such that the following three conditions are true: $d(1) \geq 2\lceil k/2 \rceil$, and $d(n) \leq n$ for all $n \geq 1$, and $\lim_{n \rightarrow \infty} d(n) = \infty$. Fix $\varepsilon > 0$ and $0 < s < c \leq 1$. There is a constant $K > 0$ such that the following holds.*

Suppose that for every $n \geq 1$, the degree- $d(n)$ sos relaxation cannot achieve a $(c + \varepsilon, s)$ -approximation for $\text{MAX-}\Pi_n$. Then for all $n \geq 1$, no SDP relaxation of size at most $Kn^{d(n)^2/8}$ can achieve a (c, s) -approximation for $\text{MAX-}\Pi_N$ for every $N > n^{4d(n)}$.

Proof. Without loss of generality, we may assume that $d(n) \geq 24$ is always an even integer. Given that the degree- $d(n)$ sos relaxation cannot achieve a $(c + \varepsilon, s)$ -approximation for $\text{MAX-}\Pi_n$, there exists an instance \mathfrak{J} of $\text{MAX-}\Pi_n$ such that $\text{opt}(\mathfrak{J}) \leq s$, along with a degree- $d(n)$ pseudo-density $D(x)$ such that $\mathbb{E}_x D(x)(c - \mathfrak{J}(x)) < -\varepsilon$. The pseudo-density $D(x)$ can be written as

$$D(x) = \sum_{\alpha \subseteq [n], |\alpha| \leq d(n)} \mathbb{E}_x [D(x) \chi_\alpha(x)] \cdot \chi_\alpha(x),$$

where for each α , $|\mathbb{E}_x[D(x)\chi_\alpha(x)]| \leq 1$ by [Fact 6.3](#). Hence,

$$\|D\|_\infty \leq \sum_{i=0}^{d(n)} \binom{n}{i} \leq 1 + n^{d(n)}.$$

Fix $f = c - \mathfrak{J}$ and define the matrix $M_N^f : \binom{N}{n} \times \{0,1\}^n \rightarrow [0,1]$ as $M_N^f(S, x) \stackrel{\text{def}}{=} f(x_S)$. By [Theorem 3.8 \(3.11\)](#), whenever $N > n^{4d(n)}$, we have $\text{rk}_{\text{psd}}(M_n^f) \geq n^{d(n)^2/8}$. As $M_{c,s}^{N,\Pi}$ contains M_N^f as a submatrix, the same lower bound applies to $M_{c,s}^{N,\Pi}$. This implies that no SDP relaxation of size at most $n^{d(n)^2/4}$ can achieve a (c, s) -approximation for $\text{MAX-}\Pi_N$ when $N > n^{4d(n)}$. \square

Using known lower bounds for low degree sum-of-squares relaxations for max-CSPs [[Gri01b](#), [Sch08](#), [Tul09](#)], [Theorem 6.4](#) implies lower bounds against general SDP relaxations for a range of specific max-CSPs. For instance, the lower bounds of Grigoriev [[Gri01b](#)] and Schoenebeck [[Sch08](#)] imply a lower bound for MAX 3-SAT (see [Theorem 1.5](#)).

Theorem 6.5 ([[Gri01b](#), [Sch08](#)]). *For every $\varepsilon > 0$, there exists a constant c_ε such that the following holds. For every $n \geq 1$, there is a $\text{MAX-}\Pi_n$ instance \mathfrak{J}_n such that $\text{opt}(\mathfrak{J}_n) \leq 7/8 + \varepsilon$, but $\overline{\text{sos}}_{c_\varepsilon n}(\mathfrak{J}) = 1$.*

Observe that one can obtain the bound of [Theorem 1.5](#) using the preceding result as follows. In [Theorem 6.4](#), choose $n \asymp \log N$ and $d(n) \asymp \frac{\log N}{\log \log N}$ so that $n^{4d(n)} \asymp N$. In that case, the lower bound obtained is of the order $N^{d(n)/32} \asymp N^{\Omega(\log N / \log \log N)}$.

7 Nonnegative rank

[Theorem 3.8](#) exhibits a connection between psd rank and sos degree. There is a similar connection between nonnegative rank and junta-degree. The results of [Section 7.1](#) generalize those of [[CLRS13](#)], while the method of proof is closely related. As opposed to [[CLRS13](#)], we use the learning approach of [Section 4](#) to approximate by juntas. In [Section 7.2](#), we demonstrate an application to the correlation polytope.

7.1 Nonnegative rank vs. junta degree

We recall that the *nonnegative rank* of a matrix $M \in \mathbb{R}_+^{p \times q}$ is the smallest integer $r \geq 1$ such that there exist $v_1, \dots, v_p, u_1, \dots, u_q \in \mathbb{R}_+^r$ satisfying $M_{ij} = \langle u_i, v_j \rangle$ for all $i \in [p], j \in [q]$. We denote the minimal value r by $\text{rk}_+(M)$.

Junta degree and pseudo-densities. Fix a finite set X . For a nonnegative function $f : X^n \rightarrow \mathbb{R}_+$, we say that f has a *nonnegative junta certificate of degree d* if there exist nonnegative d -juntas $g_1, g_2, \dots, g_k : X^n \rightarrow \mathbb{R}_+$ such that $f = \sum_{i=1}^k g_i$ (as functions on the discrete cube). The *junta degree of f* , denoted $\text{deg}_J(f)$, is the minimal d such that f has a nonnegative junta certificate of degree d .

Consider an arbitrary measure μ on X^n . A function $D : X^n \rightarrow \mathbb{R}$ is called a *d -local pseudo-density (with respect to the measure μ)* if $\mathbb{E}_\mu D = 1$ and furthermore $\mathbb{E}_{x \sim \mu} D(x)g(x) \geq 0$ for all nonnegative d -juntas g . If a measure μ is unspecified, we always refer to the uniform measure by default. The following characterization is immediate from the fact that the set of functions satisfying $\text{deg}_J(f) \leq d$ is a closed convex cone.

Lemma 7.1. *For every $f : X^n \rightarrow \mathbb{R}_+$ and $d \geq 0$, we have $\text{deg}_J(f) > d$ if and only if there exists a d -local pseudo-density such that $\mathbb{E}_x D(x)f(x) < 0$.*

We also define a more quantitative notion: *Approximate* junta degree with respect to an arbitrary measure. Given $\varepsilon > 0$ and a measure μ on X^n , we define

$$\deg_J^\varepsilon(f; \mu) \stackrel{\text{def}}{=} 1 + \max \left\{ d : \exists \text{ a } d\text{-local pseudo-density } D \text{ wrt } \mu \text{ and } \mathbb{E}_{x \sim \mu} D(x)f(x) < -\varepsilon \|D\|_\infty \mathbb{E}_\mu f \right\},$$

where we take the maximum to be equal to -1 if no such pseudo-density exists. (See [Section 7.2](#) for an example where a biased measure μ is used to analyze the nonnegative rank of the lopsided disjointness matrix.) We can now state our main theorem on nonnegative rank.

For any measure μ on X , we use μ^n to denote the corresponding product measure on X^n . In the following theorem, we write $\|f\|_1 := \mathbb{E}_{\mu^m} f$.

Theorem 7.2. *For any finite set X , any measure μ on X , and any $\varepsilon > 0$, the following holds. For any $f : X^m \rightarrow \mathbb{R}_+$ and all $n \geq 2m$,*

$$1 + n^{d+1} \geq \text{rk}_+(M_n^f) \geq \left(\frac{c\varepsilon^2 n}{m^2(d \log n + \log(\|f\|_\infty / \|f\|_1))} \right)^d, \quad (7.1)$$

where $d + 1 = \deg_J^\varepsilon(f; \mu^m)$ and $c > 0$ is a universal constant.

Proof. The left-hand-side inequality of (7.1) follows from the fact that the cone of nonnegative d -juntas is spanned by $\sum_{i=0}^{d+1} \binom{n}{i} \leq 1 + n^{d+1}$ nonnegative d -juntas. We move on to right-hand inequality.

We will use $\langle \cdot, \cdot \rangle$ for the inner product on $L^2(X^n; \mu^n)$, i.e. $\langle g, h \rangle = \mathbb{E}_{\mu^n} [gh]$. Consider a rank- r nonnegative factorization

$$M_n^f(S, x) = \sum_{i=1}^r \lambda_i(S) q_i(x). \quad (7.2)$$

By rescaling, we may assume that $\mathbb{E}_{\mu^n} q_i = 1$ for each $i \in [r]$. Observe that, by taking expectation on both sides with respect to μ^n , for any fixed S we have

$$\sum_{i=1}^r \lambda_i(S) = \mathbb{E}_{x \sim \mu^n} M_n^f(S, x) = \mathbb{E}_{\mu^m} f = \|f\|_1. \quad (7.3)$$

Let $\Lambda_\tau = \{i : \|q_i\|_\infty \leq \tau\}$. Note that for $i \notin \Lambda_\tau$, we must have

$$\lambda_i(S) \leq \frac{\|f\|_\infty}{\tau} \quad \forall |S| = m. \quad (7.4)$$

Let $D : X^m \rightarrow \mathbb{R}$ be a d -local pseudo-density witnessing $\deg_J^\varepsilon(f; \mu^m) > d$. For $S \subseteq [n]$ with $|S| = m$, we define a function $D_S : X^n \rightarrow \mathbb{R}$ by $D_S(x) = D(x_S)$. Note that each D_S is clearly an m -junta.

For some $\delta > 0$ to be chosen later, for each q_i with $i \in \Lambda_\tau$, we apply [Theorem 4.8](#) to obtain a density \tilde{q}_i that is a k' -junta for $k' = O(\|D\|_\infty^2 \frac{m}{\delta^2} \log \tau)$, and such that for every $S \subseteq [n]$ with $|S| = m$, we have

$$\langle D_S, q_i \rangle \geq \langle D_S, \tilde{q}_i \rangle - \delta. \quad (7.5)$$

Let J_i denote the set of coordinates on which q_i depends, so that $|J_i| \leq k'$.

We now take the inner product of both sides of (7.2) with the function $\mathbb{E}_{|S|=m} D_S$. On the left-hand side, using our assumption on the pseudo-density D ,

$$\mathbb{E}_S \mathbb{E}_{x \sim \mu^n} D_S(x) M_n^f(S, x) < -\varepsilon \|D\|_\infty \|f\|_1. \quad (7.6)$$

We break the right-hand side of (7.2) into two parts. First, using (7.4),

$$\sum_{i \notin \Lambda_\tau} \mathbb{E}_S \mathbb{E}_{x \sim \mu^n} D_S(x) \lambda_i(S) q_i(x) \geq -\frac{\|f\|_\infty}{\tau} \|D\|_\infty r. \quad (7.7)$$

For the second part, we use (7.5) so that for every $i \in \Lambda_\tau$ and $|S| = m$,

$$\begin{aligned} \mathbb{E}_{x \sim \mu^n} D_S(x) q_i(x) &\geq -\delta + \mathbb{E}_{x \sim \mu^n} D_S(x) \tilde{q}_i(x) \\ &= -\delta + \mathbb{E}_{\substack{y \in X^S \\ y \sim \mu^m}} D_S(y) \mathbb{E}_{x \sim \mu^n} [\tilde{q}_i(x) \mid x_S = y] \\ &\geq -\delta - \|D\|_\infty \mathbf{1}_{\{|S \cap J_i| > d\}}, \end{aligned}$$

where in the final line we have used the facts that the function $y \mapsto \mathbb{E}_{x \sim \mu^n} [\tilde{q}_i(x) \mid x_S = y]$ is a nonnegative $(S \cap J_i)$ -junta, and that $D_S : X^m \rightarrow \mathbb{R}$ is a d -local pseudo-density.

This implies that for $i \in \Lambda_\tau$,

$$\begin{aligned} \mathbb{E}_S \lambda_i(S) \mathbb{E}_{x \sim \mu^n} D_S(x) q_i(x) &\geq -\delta \mathbb{E}_S \lambda_i(S) - \|D\|_\infty \|\lambda_i\|_\infty \mathbb{P}(|S \cap J_i| > d) \\ &\geq -\delta \mathbb{E}_S \lambda_i(S) - \|D\|_\infty \|\lambda_i\|_\infty \frac{\binom{|J_i|}{d} \binom{n}{m-d/2}}{\binom{n}{m}} \\ &\geq -\delta \mathbb{E}_S \lambda_i(S) - \|D\|_\infty \|\lambda_i\|_\infty \frac{|J_i|^d m^d}{(n-m)^d} \\ &\geq -\delta \mathbb{E}_S \lambda_i(S) - \|D\|_\infty \|f\|_1 \frac{(k')^d (2m)^d}{n^d}, \end{aligned}$$

where in the final line we have used $\|\lambda_i\|_\infty \leq \|f\|_1$ from (7.3) and also $|J_i| \leq k'$ and $n \geq 2m$.

Combining this with (7.7) and (7.6), we conclude that

$$\begin{aligned} -\varepsilon \|D\|_\infty \|f\|_1 &> \sum_{i=1}^r \mathbb{E}_S \lambda_i(S) \mathbb{E}_{x \sim \mu^n} D_S(x) q_i(x) \\ &\geq -\frac{\|f\|_\infty}{\tau} \|D\|_\infty r - |\Lambda_\tau| \cdot \|D\|_\infty \|f\|_1 \frac{(k')^d (2m)^d}{n^d} - \delta \sum_{i \in \Lambda_\tau} \mathbb{E}_S \lambda_i(S) \\ &\geq -r \|D\|_\infty \left(\frac{\|f\|_\infty}{\tau} + \|f\|_1 \frac{(k')^d (2m)^d}{n^d} \right) - \delta \|f\|_1. \end{aligned}$$

Let us now set $\tau = 3r \frac{\|f\|_\infty}{\|f\|_1}$ and $\delta = \varepsilon \|D\|_\infty / 3$, yielding

$$r \geq \frac{1}{3} \left(\frac{n}{2k'm} \right)^d \geq \left(c \frac{\varepsilon^2 n}{m^2 \log \tau} \right)^d \quad (7.8)$$

for some universal constant $c > 0$.

Now, if $r \geq n^d$, then we are done. Otherwise, (7.8) yields

$$r \geq \left(\frac{c \varepsilon^2 n}{m^2 (d \log n + \log(\|f\|_\infty / \|f\|_1))} \right)^d,$$

completing the proof. \square

7.2 The correlation polytope and lopsided disjointness

We now illustrate a particularly simple application of our method to nonnegative rank.

Lemma 7.3. *There is a constant $\varepsilon_0 > 0$ such that for all $m \geq 3$, the following holds. Define $f : \{0, 1\}^m \rightarrow \mathbb{R}_+$ by*

$$f(x) = \left(1 - \sum_{i=1}^m x_i\right)^2, \quad (7.9)$$

and let μ be the measure on $\{0, 1\}$ satisfying $\mu(0) = 1 - 2/m$ and $\mu(1) = 2/m$. Then,

$$\deg^{\varepsilon_0}(f; \mu^m) \geq \frac{m}{2} + 1.$$

Plugging this result into [Theorem 7.2](#) yields the following.

Theorem 7.4. *There is a constant $c > 0$ such that for every $m \geq 3$ and $n \geq 2m$, we have*

$$\text{rk}_+(M_n^f) \geq \left(\frac{cn}{m^3 \log n}\right)^{m/2}.$$

In particular, by setting $m = m(n)$ appropriately, [Proposition 5.1](#) implies that $\text{rk}_+(\text{CORR}_n) \geq 2^{\Omega(n^{1/3})}$. One should note that this is somewhat weaker than the lower bound $\text{rk}_+(\text{CORR}_n) \geq 2^{\Omega(n)}$ proved in [\[FMP⁺12\]](#).

Proof of Lemma 7.3. For $x \in \{0, 1\}^m$, let $|x|$ denote the hamming weight of x . Define the pseudo-density $D : \{0, 1\}^m \rightarrow \mathbb{R}$ with respect to μ^m by

$$D(x) \stackrel{\text{def}}{=} \begin{cases} -\frac{1}{\mu^m(\mathbf{0})} & |x| = 0 \\ \frac{2}{m\mu^m(x)} & |x| = 1 \\ 0 & |x| > 1 \end{cases}$$

We now verify that D is a d -local pseudo-density (with respect to μ^m) for $d = \frac{m}{2} + 1$. Observe first that

$$\mathbb{E}_{x \sim \mu^m} D(x) = -1 + m \cdot \frac{2}{m} = 1.$$

Let $\beta = \frac{2}{m} \left(1 - \frac{2}{m}\right)^{m-1}$ denote $\mu^m(1, 0, \dots, 0)$. Consider a subset $S \subseteq [m]$ and some fixed string $b \in \{0, 1\}^S$. Let $\mathbf{1}_b : \{0, 1\}^m \rightarrow \{0, 1\}$ denote the indicator of whether $x_S = b$. If $b = \mathbf{0}$, then

$$\mathbb{E}_{x \sim \mu^m} D(x) \mathbf{1}_b(x) = \beta (m - |S|) - \left(1 - \frac{2}{m}\right)^m = \left(1 - \frac{2}{m}\right)^{m-1} \left(1 - 2 \frac{|S| - 1}{m}\right)$$

The latter quantity is nonnegative as long as $|S| \leq \frac{m}{2} + 1$.

If $|b| > 1$, then $\mathbb{E}_{x \sim \mu^m} D(x) \mathbf{1}_b(x) = 0$, and if $|b| = 1$, then $\mathbb{E}_{x \sim \mu^m} D(x) \mathbf{1}_b(x) \geq 0$ since $D(x) \geq 0$ on the support of $\mathbf{1}_b$. But any nonnegative S -junta is a nonnegative combination of the functions $\mathbf{1}_b$ as b ranges over $\{0, 1\}^S$. We conclude that as long as $d \leq \frac{m}{2} + 1$, D is a d -local pseudo-density.

Moreover we have

$$\mathbb{E}_{x \sim \mu^m} D(x) f(x) = \mathbb{E}_{x \sim \mu^m} D(x) \left(1 - 2 \sum_{i=1}^m x_i + \sum_{i=1}^m x_i^2 + 2 \sum_{i \neq j} x_i x_j\right) = -1.$$

Also observe that since $m \geq 3$,

$$\|D\|_\infty = |D(\mathbf{0})| = \left(1 - \frac{2}{m}\right)^{-m} \leq 27.$$

Lastly, it is easy to see that $\mathbb{E}_{\mu^m} f \geq \Omega(1)$. These facts together imply that for some universal constant $\varepsilon_0 > 0$, we have $\deg_J^{\varepsilon_0}(f; \mu^m) \geq m/2 + 1$, as desired. \square

An interesting feature of the pseudodensity D is that it is supported only on $x \in \{0, 1\}^m$ with $|x| \leq 1$. Therefore, the lower bound on the approximate junta degree established in [Lemma 7.3](#) applies to any function $f : \{0, 1\}^m \rightarrow \mathbb{R}_+$ that satisfies

$$f(x) = \begin{cases} 1 & \text{for } |x| = 0 \\ 0 & \text{for } |x| = 1. \end{cases}$$

Moreover, the lower bound on $\text{rk}_+(M_n^f)$ also applies in this general setting. To restate this generalization of [Theorem 7.4](#), let us interpret an element of $\{0, 1\}^n$ as a subset of $\{1, 2, \dots, n\}$.

Corollary 7.5 (Lopsided unique disjointness). *There is a fixed constant $c > 0$ such that for every $m \geq 3$ and $n \geq 2m$, given a matrix $M : \binom{[n]}{m} \times 2^{[n]} \rightarrow \mathbb{R}_+$ satisfying*

$$M(S, T) = \begin{cases} 1 & \text{if } |S \cap T| = 1 \\ 0 & \text{if } |S \cap T| = 0, \end{cases}$$

we have $\text{rk}_+(M) \geq \left(\frac{cn}{m^3 \log n}\right)^{m/2}$.

In other words, the lower bound of [Theorem 7.4](#) applies to all matrices that have a subset of entries corresponding to the *unique disjointness* problem.

7.3 Unique games hardness for LPs

As an illustrative application of the relation between nonnegative rank and junta-degree ([Theorem 7.2](#)), we present an LP hardness result for the Unique Games problem.¹¹

Fix an integer $q \geq 1$. An instance \mathfrak{J} of unique games UG^q consists of variables X_1, \dots, X_n taking values in $[q]$ and a collection of predicates P_1, \dots, P_M over these variables. Each constraint P_i is over a pair of distinct variables $\{X_{a_i}, X_{b_i}\}$ and is specified by a bijection $\pi_i : [q] \rightarrow [q]$ as follows:

$$P_i(X_{a_i}, X_{b_i}) \stackrel{\text{def}}{=} \mathbf{1}[\pi_i(X_{a_i}) = \pi_i(X_{b_i})].$$

The goal is to find an assignment that maximizes, over $x \in [q]^n$, the number of satisfied constraints:

$$\mathfrak{J}(x) \stackrel{\text{def}}{=} \frac{1}{M} \sum_{i=1}^M P_i(x)$$

Recall that $\text{opt}(\mathfrak{J}) = \max_{x \in [q]^n} \mathfrak{J}(x)$. Let UG_n^q denote the family of Unique Games instances on n variables. The authors of [\[CMM09\]](#) exhibit a strong integrality gap for Sherali-Adams linear programming relaxations of UG^q

¹¹We thank Ola Svensson for the suggestion to make this explicit.

Theorem 7.6 ([CMM09]). Fix a number $t \geq 1$ and let $q = 2^t$. Then for every $\delta > 0$, there exist $\gamma, \varepsilon > 0$, an $m \geq 1$, and an instance $\mathfrak{J} \in \text{UG}_m^q$ such that

$$\text{opt}(\mathfrak{J}) \leq \frac{1}{q} + \delta,$$

but $\text{deg}_j^\varepsilon(1 - \delta - \mathfrak{J}) \geq m^\gamma$.

In fact, the authors of [CMM09] construct a lower bound for the d -round Sherali-Adams LP relaxation (where $d \asymp m^\gamma$). But there is an equivalence between such lower bounds and the existence of a d -local pseudo-density; we refer to [CLRS13] for a discussion. Applying Theorem 7.2 (with $X = [q]$ and μ as the uniform measure on $[q]$), we obtain the following corollary.

Let $M_{c,s}^{n, \text{UG}^q}$ denote the matrix with entries

$$M_{c,s}^{n, \text{UG}^q}(\mathfrak{J}, x) = c - \mathfrak{J}(x),$$

where \mathfrak{J} runs over all UG_n^q instances with $\text{opt}(\mathfrak{J}) \leq s$, and all values $x \in [q]^n$.

Corollary 7.7. For every $t \geq 1$, $\delta > 0$, and $d \geq 1$, there exists a constant $c > 0$ such that for all $n \geq 1$,

$$\text{rk}_+ \left(M_{1-\delta, 1/q+\delta}^{n, \text{UG}^q} \right) \geq cn^d,$$

where $q = 2^t$.

In the language of [CLRS13] (see also Section 6 for related definitions in the SDP setting), this shows that polynomial-size families of LP relaxations cannot achieve a $(1 - \delta, \frac{1}{q} + \delta)$ -approximation for the Unique Games problem.

Acknowledgments

This work was supported, in large part, by NSF grant CCF-1407779. A significant fraction of the project was completed during a long-term visit of the authors to the Simons Institute for the Theory of Computing (Berkeley) for the program on Algorithmic Spectral Graph Theory. The authors would also like to thank Paul Beame, Siu-On Chan, Daniel Dadush, Troy Lee, Sebastian Pokutta, Pablo Parrilo, Mohit Singh, Ola Svensson, Thomas Rothvoß, and Rekha Thomas for valuable discussions and comments.

References

- [AK07] Sanjeev Arora and Satyen Kale. A combinatorial, primal-dual approach to semidefinite programs. In David S. Johnson and Uriel Feige, editors, *STOC*, pages 227–236. ACM, 2007. 15
- [ARV09] Sanjeev Arora, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. *J. ACM*, 56(2):Art. 5, 37, 2009. 3
- [Aus10] Per Austrin. Towards sharp inapproximability for any 2-CSP. *SIAM J. Comput.*, 39(6):2430–2463, 2010. 3
- [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. In *STOC*, pages 307–326, 2012. 13

- [BDP13] Jop Briët, Daniel Dadush, and Sebastian Pokutta. On the existence of 0/1 polytopes with high semidefinite extension complexity. In *Algorithms - ESA 2013 - 21st Annual European Symposium, Sophia Antipolis, France, September 2-4, 2013. Proceedings*, pages 217–228, 2013. [3](#), [5](#), [14](#)
- [BFPS12] Gábor Braun, Samuel Fiorini, Sebastian Pokutta, and David Steurer. Approximation limits of linear programs (beyond hierarchies). In *FOCS*, pages 480–489, 2012. [4](#)
- [BKS14] Boaz Barak, Jonathan A. Kelner, and David Steurer. Rounding sum-of-squares relaxations. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 31–40, 2014. [13](#)
- [BS14] Boaz Barak and David Steurer. Sum-of-squares proofs and the quest toward optimal algorithms. *CoRR*, abs/1404.5236, 2014. [9](#)
- [BT03] Amir Beck and Marc Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett.*, 31(3):167–175, 2003. [15](#)
- [Bub14] S. Bubeck. Theory of convex optimization for machine learning. [arXiv:1405.4980](#), 2014. [15](#)
- [Car10] Eric Carlen. Trace inequalities and quantum entropy: an introductory course. In *Entropy and the quantum*, volume 529 of *Contemp. Math.*, pages 73–140. Amer. Math. Soc., Providence, RI, 2010. [25](#)
- [CGH⁺05] Julia Chuzhoy, Sudipto Guha, Eran Halperin, Sanjeev Khanna, Guy Kortsarz, Robert Krauthgamer, and Joseph Naor. Asymmetric k -center is $\log^* n$ -hard to approximate. *J. ACM*, 52(4):538–551, 2005. [3](#)
- [CLRS13] Siu On Chan, James R. Lee, Prasad Raghavendra, and David Steurer. Approximate constraint satisfaction requires large LP relaxations. In *54th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pages 350–359, 2013. [3](#), [9](#), [13](#), [30](#), [38](#), [43](#)
- [CMM09] M. Charikar, K. Makarychev, and Y. Makarychev. Integrality gaps for Sherali-Adams relaxations. In *Proc. STOC*, pages 283–292. ACM, 2009. [42](#), [43](#)
- [DS90] Caterina De Simone. The cut polytope and the Boolean quadric polytope. *Discrete Math.*, 79(1):71–75, 1989/90. [5](#), [31](#)
- [FGP⁺14] Hamza Fawzi, João Gouveia, Pablo A. Parrilo, Richard Z. Robinson, and Rekha R. Thomas. Positive semidefinite rank. *Arxiv*, arXiv:1407.4095, 2014. [4](#), [14](#)
- [FMP⁺12] Samuel Fiorini, Serge Massar, Sebastian Pokutta, Hans Raj Tiwary, and Ronald de Wolf. Linear vs. semidefinite extended formulations: exponential separation and strong lower bounds. In *STOC*, pages 95–106, 2012. [3](#), [4](#), [5](#), [8](#), [9](#), [10](#), [31](#), [41](#)
- [FSP13] H. Fawzi, J. Saunderson, and P. A. Parrilo. Equivariant semidefinite lifts and sum-of-squares hierarchies. [arXiv:1312.6662](#), 2013. [3](#)
- [GPT11] J. Gouveia, P. A. Parrilo, and R. Thomas. Lifts of convex sets and cone factorizations. [arXiv:1111.3164](#), 2011. [3](#), [8](#), [10](#)

- [Gri01a] Dima Grigoriev. Complexity of positivstellensatz proofs for the knapsack. *Computational Complexity*, 10(2):139–154, 2001. [10](#), [31](#), [32](#)
- [Gri01b] Dima Grigoriev. Linear lower bound on degrees of Positivstellensatz calculus proofs for the parity. *Theoret. Comput. Sci.*, 259(1-2):613–622, 2001. [8](#), [38](#)
- [GV02] Dima Grigoriev and Nicolai Vorobjov. Complexity of Null- and Positivstellensatz proofs. *Ann. Pure Appl. Logic*, 113(1-3):153–160, 2002. First St. Petersburg Conference on Days of Logic and Computability (1999). [9](#)
- [GW95] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. Assoc. Comput. Mach.*, 42(6):1115–1145, 1995. [3](#)
- [HK03] Eran Halperin and Robert Krauthgamer. Polylogarithmic inapproximability. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing*, pages 585–594 (electronic). ACM, New York, 2003. [3](#)
- [Kho02] Subhash Khot. On the power of unique 2-prover 1-round games. In *STOC*, pages 767–775, 2002. [3](#)
- [KKMO04] Subhash Khot, Guy Kindler, Elchanan Mossel, and Ryan O’Donnell. Optimal inapproximability results for max-cut and other 2-variable csps? In *FOCS*, pages 146–154, 2004. [3](#)
- [KMS98] David Karger, Rajeev Motwani, and Madhu Sudan. Approximate graph coloring by semidefinite programming. *J. ACM*, 45(2):246–265, 1998. [3](#)
- [Las01] Jean B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM J. Optim.*, 11(3):796–817, 2000/01. [3](#), [6](#)
- [Lau09] Monique Laurent. Sums of squares, moment matrices and optimization over polynomials. In *Emerging applications of algebraic geometry*, volume 149 of *IMA Vol. Math. Appl.*, pages 157–270. Springer, New York, 2009. [9](#)
- [LRST14] James R. Lee, Prasad Raghavendra, David Steurer, and Ning Tan. On the power of symmetric LP and SDP relaxations. In *IEEE 29th Conference on Computational Complexity, CCC 2014, Vancouver, BC, Canada, June 11-13, 2014*, pages 13–21, 2014. [3](#)
- [LY93] Carsten Lund and Mihalis Yannakakis. On the hardness of approximating minimization problems. In *STOC*, pages 286–293, 1993. [3](#)
- [NY83] A. S. Nemirovsky and D. B. Yudin. *Problem complexity and method efficiency in optimization*. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1983. Translated from the Russian and with a preface by E. R. Dawson, Wiley-Interscience Series in Discrete Mathematics. [15](#)
- [O’D14] Ryan O’Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. [9](#), [12](#)
- [OZ12] Ryan O’Donnell and Yuan Zhou. Approximability and proof complexity. *CoRR*, abs/1211.1958, 2012. [9](#)

- [Pad89] Manfred Padberg. The Boolean quadric polytope: some characteristics, facets and relatives. *Math. Programming*, 45(1, (Ser. B)):139–172, 1989. [31](#)
- [Par00] Pablo Parrilo. *Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization*. PhD thesis, California Institute of Technology, 2000. [3](#), [6](#)
- [Rag08] Prasad Raghavendra. Optimal algorithms and inapproximability results for every CSP? [extended abstract]. In *STOC'08*, pages 245–254. ACM, New York, 2008. [3](#)
- [Rot13] Thomas Rothvoß. Some 0/1 polytopes need exponential size extended formulations. *Math. Program.*, 142(1-2, Ser. A):255–268, 2013. [3](#)
- [Rot14] Thomas Rothvoß. The matching polytope has exponential extension complexity. In *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pages 263–272, 2014. [3](#)
- [Sch08] G. Schoenebeck. Linear level Lasserre lower bounds for certain k-CSPs. In *Proc. FOCS*, pages 593–602. IEEE, 2008. [8](#), [38](#)
- [She11] Alexander A. Sherstov. The pattern matrix method. *SIAM J. Comput.*, 40(6):1969–2000, 2011. [9](#)
- [Sho87] N. Z. Shor. An approach to obtaining global extremums in polynomial mathematical programming problems. *Cybernetics*, 23(5):695–700, 1987. [3](#), [6](#)
- [TRW05] Koji Tsuda, Gunnar Rätsch, and Manfred K. Warmuth. Matrix exponentiated gradient updates for on-line learning and Bregman projection. *J. Mach. Learn. Res.*, 6:995–1018, 2005. [15](#)
- [Tul09] Madhur Tulsiani. CSP gaps and reductions in the lasserre hierarchy. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pages 303–312, 2009. [38](#)
- [Vaz01] Vijay V. Vazirani. *Approximation algorithms*. Springer-Verlag, Berlin, 2001. [3](#)
- [Wil67] R. M. Wilcox. Exponential operators and parameter differentiation in quantum physics. *J. Mathematical Phys.*, 8:962–982, 1967. [27](#)
- [Wil13] Mark M. Wilde. *Quantum information theory*. Cambridge University Press, Cambridge, 2013. [12](#)
- [WK12] Manfred K. Warmuth and Dima Kuzmin. Online variance minimization. *Mach. Learn.*, 87(1):1–32, 2012. [15](#)
- [WS11] D. P. Williamson and D. B. Shmoys. *The design of approximation algorithms*. Cambridge University Press, Cambridge, 2011. [3](#)
- [Yan91] Mihalis Yannakakis. Expressing combinatorial optimization problems by linear programs. *J. Comput. System Sci.*, 43(3):441–466, 1991. [3](#), [8](#)