Electronic Colloquium on Computational Complexity, Report No. 21 (2010)

Non-commutative circuits and the sum-of-squares problem

Pavel Hrubeš* Avi Wigderson* Amir Yehudayoff*

Abstract

We initiate a direction for proving lower bounds on the size of non-commutative arithmetic circuits. This direction is based on a connection between lower bounds on the size of *non-commutative* arithmetic circuits and a problem about *commutative* degree four polynomials, the classical sum-of-squares problem: find the smallest n such that there exists an identity

$$(x_1^2 + x_2^2 + \cdots + x_k^2) \cdot (y_1^2 + y_2^2 + \cdots + y_k^2) = f_1^2 + f_2^2 + \cdots + f_n^2, \tag{0.1}$$

where each $f_i = f_i(X, Y)$ is a bilinear form in $X = \{x_1, \ldots, x_k\}$ and $Y = \{y_1, \ldots, y_k\}$. Over the complex numbers, we show that a sufficiently strong *super-linear* lower bound on n in (0.1), namely, $n \geq k^{1+\epsilon}$ with $\varepsilon > 0$, implies an *exponential* lower bound on the size of arithmetic circuits computing the non-commutative permanent.

More generally, we consider such sum-of-squares identities for any biquadratic polynomial $h(X, Y)$, namely

$$h(X, Y) = f_1^2 + f_2^2 + \cdots + f_n^2. \tag{0.2}$$

Again, proving $n \geq k^{1+\epsilon}$ in (0.2) for *any* explicit h over the complex numbers gives an *exponential* lower bound for the non-commutative permanent. Our proofs relies on several new structure theorems for non-commutative circuits, as well as a non-commutative analog of Valiant's completeness of the permanent.

We proceed to prove such super-linear bounds in some restricted cases. We prove that $n \geq \Omega(k^{6/5})$ in (0.1), if $f_1, \ldots, f_n$ are required to have *integer* coefficients. Over the *real* numbers, we construct an explicit biquadratic polynomial h such that n in (0.2) must be at least $\Omega(k^2)$. Unfortunately, these results do not imply circuit lower bounds.

We also present other structural results about non-commutative arithmetic circuits. We show that any non-commutative circuit computing an *ordered* non-commutative polynomial can be efficiently transformed to a syntactically multilinear circuit computing that polynomial. The permanent, for example, is ordered. Hence, lower bounds on the size of syntactically multilinear circuits computing the permanent imply unrestricted non-commutative lower bounds. We also prove an exponential lower bound on the size of non-commutative syntactically multilinear circuit computing an explicit polynomial. This polynomial is, however, not ordered and an unrestricted circuit lower bound does not follow.

*School of Mathematics, Institute for Advanced Study, Princeton NJ 08540. Emails: `pahrubes@gmail.com`, `avi@ias.edu`, and `amir.yehudayoff@gmail.com`. Research supported by NSF Grant DMS-0835373.

ISSN 1433-8092

1 Introduction

1.1 Non-commutative computation

Arithmetic complexity theory studies computation of formal polynomials over some field or ring. Most of this theory is concerned with computation of commutative polynomials. The basic model of computation is that of *arithmetic circuit*. Despite decades of work, the best size lower bound for general circuits computing an explicit n-variate polynomial of degree d is $\Omega(n \log d)$, due to Baur and Strassen [30, 2]. Better lower bounds are known for a variety of more restricted computational models, such as monotone circuits, multilinear or bounded depth circuits (see, e.g., [6, 3]).

In this paper we deal with a different type of restriction. We investigate *non-commutative* polynomials and circuits; the case when the variables do not multiplicatively commute, i.e., $xy \neq yx$ if $x \neq y$, as in the case when the variables represent matrices over a field[1]. In a non-commutative circuit, a multiplication gate is given with an order in which its inputs are multiplied. Precise definitions appear in Section 2. A simple illustration of how absence of commutativity limits computation is the polynomial $x^2 - y^2$. If x, y commute, the polynomial can be computed as $(x - y)(x + y)$ using one multiplication. In the non-commutative case, two multiplications are required to compute it.

Surprisingly, while interest in non-commutative computations goes back at least to 1970 [33], no better lower bounds are known for general non-commutative circuits than in the commutative case. The seminal work in this area is [21], where Nisan proved exponential lower bounds on non-commutative *formula* size of determinant and permanent. He also gives an explicit polynomial that has linear size non-commutative circuits but requires non-commutative formulas of exponential size, thus separating non-commutative formulas and circuits.

One remarkable aspect of non-commutative computation is its connection with the celebrated approximation scheme for the (commutative) permanent [14]. The series of papers [7, 16, 1, 5] reduce the problem of approximating permanent to the problem of computing determinant of a matrix whose entries are elements of (non-commutative) Clifford algebras. However, already in the case of quaternions (the third Clifford algebra), determinant cannot be efficiently computed by means of arithmetic formulas. This was shown by Chien and Sinclair [4] who extend Nisan's techniques to this and other non-commutative algebras.

In this paper, we propose new directions towards proving lower bounds on non-commutative circuits. We present structure theorems for non-commutative circuits, which enable us to reduce circuit size lower bounds to apparently simpler problems. The foremost such problem is the so called sum-of-squares problem, a classical question on a border between algebra and topology. We also outline a connection with multilinear circuits, in which exciting progress was made in recent years. We then make modest steps towards the lower-bound goal, and present results some of which are of independent interest. Before we describe the results, we take a detour to briefly describe the sum-of-squares problem and its long history.

[1] As in this case, addition remains commutative, as well as multiplication by constants

1.2 The sum-of-squares problem

In this section all variables commute. Consider the polynomial

$$\mathsf{SOS}_k = (x_1^2 + x_2^2 + \cdots + x_k^2) \cdot (y_1^2 + y_2^2 + \cdots + y_k^2). \tag{1.1}$$

Given a field (or a ring) $\mathbb{F}$, define $\mathcal{S}_{\mathbb{F}}(k)$ as the smallest n such that there exists a polynomial identity

$$\mathsf{SOS}_k = z_1^2 + z_2^2 + \cdots + z_n^2\,, \tag{1.2}$$

where each $z_i = z_i(X, Y)$ is a bilinear form in variables $X = \{x_1, \ldots, x_k\}$ and $Y = \{y_1, \ldots, y_k\}$ over the field $\mathbb{F}$.

We refer to the problem of determining the value $\mathcal{S}_{\mathbb{F}}(k)$ as the *sum-of-squares* problem. Note that the problem is not interesting if $\mathbb{F}$ has characteristic two, for then $\mathcal{S}_{\mathbb{F}}(k) = 1$. Over other fields, the trivial bounds are

$$k \leq \mathcal{S}_{\mathbb{F}}(k) \leq k^2\,.$$

In Section 1.3, we describe the connection between the sum-of-squares problem and arithmetic complexity. At this point, let us discuss the mathematical significance of the sum-of-squares problem (much more can be found, e.g., in [29]). We focus on real sums of squares, for they are of the greatest historical importance[2]. Nontrivial identities exhibiting $\mathcal{S}_{\mathbb{R}}(k) = k$ initiated this story.

When $k = 1$, we have $x_1^2 y_1^2 = (x_1 y_1)^2$. When $k = 2$, we have

$$(x_1^2 + x_2^2) \cdot (y_1^2 + y_2^2) = (x_1 y_1 - x_2 y_2)^2 + (x_1 y_2 + x_2 y_1)^2.$$

Interpreting (x_1, x_2) and (y_1, y_2) as complex numbers α and β, this formula expresses the property

$$|\alpha|^2 |\beta|^2 = |\alpha\beta|^2 \tag{1.3}$$

of multiplication of complex numbers. The case $k = 1$ trivially expresses the same fact (1.3) for *real* α and β. In 1748, motivated by the number theoretic problem of expressing every integer as a sum of four squares, Euler proved an identity showing that $\mathcal{S}_{\mathbb{R}}(4) = 4$. When Hamilton discovered the *quaternion* algebra in 1843, this identity was quickly realized to express (1.3) for mutiplying quaternions. This was repeated in 1848 with the discovery of the *octonions* algebra, and the 8-square identity expressing (1.3) for octonions. Motivated by the study of division algebras, mathematicians tried to prove a 16-square identity in the following 50 years. Finally Hurwitz in 1898 proved that it is impossible, obtaining the first nontrivial lower bound:

Theorem 1.1. *[11] $\mathcal{S}_{\mathbb{R}}(k) > k$, except when $k \in \{1, 2, 4, 8\}$.*

The following interpretation of the sum-of-squares problem got topologists interested in this problem: if $z_1, \ldots, z_n$ satisfy (1.2), the map $z = (z_1, \ldots, z_n) : \mathbb{R}^k \times \mathbb{R}^k \to \mathbb{R}^n$ is a bilinear *normed* map. Namely, it satisfies $|z(\bar{x}, \bar{y})| = |\bar{x}||\bar{y}|$ for every $\bar{x}, \bar{y} \in \mathbb{R}^k$, where $|\cdot|$ is the Euclidean norm. This rigid structure allows for topological and algebraic geometry tools to yield the following, best known lower bound, which unfortunately gains only a factor of two over the trivial bound:

[2]The assumption that the z_i's in (1.2) are bilinear is satisfied automatically if the z_i's are real polynomials.

Theorem 1.2. *[13, 18]* $\mathcal{S}_{\mathbb{R}}(k) \geq (2 - o(1))k$.

As it happens, the trivial upper bound can be improved as well. There exists a normed bilinear map as above from $\mathbb{R}^k \times \mathbb{R}^{\rho(k)}$ to $\mathbb{R}^k$, with $\rho(k) = \Theta(\log k)$. This was shown by Radon and Hurwitz [24, 12], who computed the exact value of the optimal $\rho(k)$. Interestingly, such a map exists even if we require the polynomials z_i to have *integer*[3] coefficients, see [36, 19]. The existence of this integer bilinear normed map turns out to be related to Clifford algebras as well: it can be obtained using a matrix representation of a Clifford algebra with $\rho(k)$ generators. This can be seen to imply

Fact 1.3. $\mathcal{S}_{\mathbb{Z}}(k) \leq O(k^2/\log k)$.

This is the best known upper bound on $\mathcal{S}_{\mathbb{R}}$, or $\mathcal{S}_{\mathbb{F}}$ for any other field with char $\mathbb{F} \neq 2$. This motivated researchers to study integer sums of squares, and try to prove lower bounds on $\mathcal{S}_{\mathbb{Z}}$. Despite the effort [18, 34, 29], the asymptotic bounds on $\mathcal{S}_{\mathbb{Z}}$ remained as wide open as in the case of reals. One of the contributions of this paper is the first super-linear lower bound in the integer case. We show that $\mathcal{S}_{\mathbb{Z}}(k) \geq \Omega(k^{6/5})$.

To illustrate the subtlety of proving lower bounds on the sum-of-squares problem, let us mention that if we allow the z_i's to be *rational* functions rather than polynomials, the nature of the problem significantly changes. In 1965, Pfister [23] proved that if the z_i's are rational functions, SOS_k can be written as a sum of k squares whenever k is a power of two.

1.3 Non-commutative circuits and bilinear complexity

Conditional lower bounds on circuit complexity. The connection between the sum-of-squares problem and non-commutative lower bounds is that a sufficiently strong lower bound on $\mathcal{S}(k)$ implies an exponential lower bound for permanent. Here we present our main results, for a more detailed discussion, see Section 2.1. In the non-commutative setting, there are several options to define the permanent, we define it row-by-row, that is,

$$\mathrm{PERM}_n(X) = \sum_{\pi} x_{1,\pi(1)} x_{2,\pi(2)} \cdots x_{n,\pi(n)},$$

where π is a permutation of $[n] = \{1, \ldots, n\}$. The advertised connection can be summarized as follows[4].

Theorem 1.4. *Let $\mathbb{F}$ be an algebraically closed field. Assume that $\mathcal{S}_{\mathbb{F}}(k) \geq \Omega(k^{1+\varepsilon})$ for a constant $\varepsilon > 0$. Then PERM_n requires non-commutative circuits of size $2^{\Omega(n)}$.*

Theorem 1.4 is an instance of a general connection between non-commutative circuits and commutative degree four polynomials, which we now proceed to describe.

Let f be a *commutative* polynomial of degree four over a field $\mathbb{F}$. We say that f is *biquadratic* in variables $X = \{x_1, \ldots, x_k\}$ and $Y = \{y_1, \ldots, y_k\}$, if every monomial in f has the form $x_{i_1} x_{i_2} y_{j_1} y_{j_2}$. If f is biquadratic in variables X and Y, we define

[3] The coefficients of the z_i's can actually be taken to be in $\{-1, 0, 1\}$.

[4] If char $\mathbb{F} = 2$, the theorem holds trivially, since $\mathcal{S}_{\mathbb{F}}(k) = 1$.

sum-of-squares complexity: $\mathcal{S}_{\mathbb{F}}(f)$ is the smallest[5] n so that f can be written as

$$f = z_1^2 + \cdots + z_n^2,$$

bilinear complexity: $\mathcal{B}_{\mathbb{F}}(f)$ is the smallest n so that f can be written as

$$f = z_1 z_1' + \cdots + z_n z_n',$$

where each z_i and z_i' are bilinear forms in X, Y. We thus have $\mathcal{S}_{\mathbb{F}}(\mathsf{SOS}_k) = \mathcal{S}_{\mathbb{F}}(k)$, as defined in the previous section.

Let us first note that over certain fields, $\mathcal{S}_{\mathbb{F}}(f)$ and $\mathcal{B}_{\mathbb{F}}(f)$ are virtually the same:

Remark 1.5. *Clearly, $\mathcal{B}_{\mathbb{F}}(f) \leq \mathcal{S}_{\mathbb{F}}(f)$. If $\mathbb{F}$ is algebraically closed with char $\mathbb{F} \neq 2$, then $\mathcal{S}_{\mathbb{F}}(f) \leq 3\mathcal{B}_{\mathbb{F}}(f)$. This holds since $2zz' = (z+z')^2 + (\sqrt{-1}z)^2 + (\sqrt{-1}z')^2$.*

We now define the non-commutative version of SOS_k: the non-commutative *identity polynomial* is

$$\mathsf{ID}_k = \sum_{i,j \in [k]} x_i y_j x_i y_j . \tag{1.4}$$

We show that a lower bound on $\mathcal{B}_{\mathbb{F}}(\mathsf{SOS}_k)$ implies a lower bound on the size of non-commutative circuit computing ID_k.

Theorem 1.6. *The size of a non-commutative circuit over $\mathbb{F}$ computing ID_k is at least $\Omega(\mathcal{B}_{\mathbb{F}}(\mathsf{SOS}_k))$.*

Theorem 1.6 is proved in Section 4. The lower bound given by the theorem is reminiscent of the tensor rank approach to lower bounds for commutative circuits, where a lower bound on tensor rank implies circuit lower bounds [31]. In the non-commutative case we can prove a much stronger implication. For every $\varepsilon > 0$, a $k^{1+\varepsilon}$ lower bound on $\mathcal{B}_{\mathbb{F}}(\mathsf{SOS}_k)$ gives an exponential lower bound for the permanent. Theorem 1.7, which is proved in Section 5, together with Remark 1.5 imply Theorem 1.4.

Theorem 1.7. *Assume that $\mathcal{B}_{\mathbb{F}}(\mathsf{SOS}_k) \geq \Omega(k^{1+\varepsilon})$ for some $\varepsilon > 0$. Then PERM_n requires non-commutative circuits of size $2^{\Omega(n)}$ over $\mathbb{F}$.*

The theorem is reminiscent of a result in Boolean complexity, where a sufficient *linear* lower bound on complexity of a bipartite graph implies an *exponential* circuit lower bound for a related function (see [15] for discussion.)

An important property that the non-commutative permanent shares with its commutative counterpart is its completeness for the class of explicit polynomials. This enables us to generalize Theorem 1.7 to the following theorem, which is proved in Section 5.1. Let $\{f_k\}$ be a family of commutative biquadratic polynomials such that the number of variables in f_k is polynomial in k. We call $\{f_k\}$ *explicit*, if there exists a polynomial-time algorithm which, given k and a degree-four monomial α as inputs[6], computes the coefficient of α in f_k. The polynomial SOS_k is clearly explicit.

[5]When no such n exists, $\mathcal{S}_{\mathbb{F}}(f)$ is infinite.

[6]We think of the input as given in a binary representation; the algorithm thus runs in time polynomial in $\log k$.

Theorem 1.8. *Let $\mathbb{F}$ be a field such that char $\mathbb{F} \neq 2$. Let $\{f_k\}$ be a family of explicit biquadratic polynomials. Assume that $\mathcal{B}_{\mathbb{F}}(f_k) \geq \Omega(k^{1+\epsilon})$ for some $\epsilon > 0$. Then* PERM_n *requires non-commutative circuits of size $2^{\Omega(n)}$ over $\mathbb{F}$.*

Lower bounds on sum-of-squares complexity in restricted cases. Remark 1.5 tells us that for some fields, $\mathcal{B}_{\mathbb{F}} = \Theta(\mathcal{S}_{\mathbb{F}})$, and hence to prove a circuit lower bound, it is sufficient to prove a lower bound on $\mathcal{S}_{\mathbb{F}}$. We prove lower bounds on $\mathcal{S}_{\mathbb{F}}(k)$ in some restricted cases. For more details, see Section 2.2.

Over $\mathbb{R}$, we find an explicit 'hard' polynomial (Theorem 1.9 is proved in Section 6).

Theorem 1.9. *There exists an explicit family $\{f_k\}$ of real biquadratic polynomials with coefficients in $\{0, 1, 2, 4\}$ such that $\mathcal{S}_{\mathbb{R}}(f_k) = \Theta(k^2)$.*

By Theorem 1.8, if the construction worked over the complex numbers $\mathbb{C}$ instead of $\mathbb{R}$, we would have an exponential lower bound on the size of non-commutative circuits for the permanent. Such a construction is not known.

In Section 7, we investigate sums of squares over integers. We prove the following:

Theorem 1.10. $\mathcal{S}_{\mathbb{Z}}(k) \geq \Omega(k^{6/5})$.

This result, too, does not imply a circuit lower bound. However, if we knew how to prove the same for $\mathbb{Z}[\sqrt{-1}]$ instead of $\mathbb{Z}$, we would get lower bounds for circuits over $\mathbb{Z}$. Such lower bounds are not known.

1.4 Ordered and multilinear circuits

An important restriction on computational power of circuits is multilinearity. This restriction has been extensively investigated in the commutative setting. A polynomial is multilinear, if every variable has individual degree at most one in it. Syntactically multilinear circuits are those in which every product gate multiplies gates with disjoint sets of variables. This model was first considered in [22], where lower bounds on constant depth multilinear circuits were proved (and later improved in [27]). In a breakthrough paper, Raz [25] proved super-polynomial lower bounds on multilinear formula size for the permanent and determinant. These techniques were extended by [28] to give a lower bound of about $n^{4/3}$ for the size multilinear *circuits*.

An interesting observation about non-commutative circuits is that if they compute a polynomial of a specific form, they are without loss of generality multilinear. Let us call a non-commutative polynomial f *ordered*, if the variables of f are divided into disjoint sets $X_1, \dots, X_d$ and every monomial in f has the form $x_1 \cdots x_d$ with $x_i \in X_i$. The non-commutative permanent, as defined above, is thus ordered. An *ordered circuit* is a natural model for computing ordered polynomials. Roughly, we require every gate to take variables from the sets X_i in the same interval $I \subset [d]$; see Section 8.1 for a precise definition. One property of ordered circuits is that they are automatically syntactically multilinear.

We show that any non-commutative circuit computing an ordered polynomial can be efficiently transformed to an ordered circuit, hence a multilinear one, computing the same polynomial. Such a reduction is not known in the commutative case, and gives hope that a progress on multilinear lower bounds for permanent or determinant will yield general non-commutative lower bounds. Theorem 1.11 is proved in Section 8.1.

Theorem 1.11. *Let f be an ordered polynomial of degree d. If f is computed by a non-commutative circuit of size s, it can be computed by an ordered circuit of size $O(d^3 s)$.*

Again, we fall short of utilizing this connection for general lower bounds. By a simple argument, we manage to prove an exponential lower bound on non-commutative multilinear circuits, as we state in the next theorem. However, the polynomial AP_k in question is not ordered, and we cannot invoke the previous result to obtain an unconditional lower bound (Theorem 1.12 is proved in Section 8.2).

Theorem 1.12. *Let*

$$\mathsf{AP}_k = \sum_{\sigma} x_{\sigma(1)} x_{\sigma(2)} \cdots x_{\sigma(k)},$$

where σ is a permutation of $[k]$. Then every non-commutative multilinear circuit computing AP_k is of size at least $2^{\Omega(k)}$.

1.5 A different perspective: lower bounds using rank

An extremely appealing way to obtain lower bounds is by using sub-additive measures, and matrix rank is perhaps the favorite measure across many computational models. It is abundant in communication complexity, and in circuit complexity it has also found its applications. Often, one cannot hope to find a unique matrix whose rank would capture the complexity of the investigated function. Instead, we can associate the function with a family of matrices, and the complexity of the function is related to the *minimum* rank of matrices in that family. Typically, the family consists of matrices which are in some sense “close”to some fixed matrix.

For arithmetic circuits, many of the known structure theorems [8, 21, 25, 9] invite a natural rank interpretation. This interpretation, however, has lead to lower bounds only for restricted circuits. We sketch below the rank problem which arises in the case of commutative circuits, and explain why it is considerably simpler in the case of non-commutative ones.

Let f be a commutative polynomial of degree d. Consider $N \times N$ matrices whose entries are elements of some field, and rows and columns are labelled by monomials of degree roughly $d/2$. Hence N is in general exponential in the degree of f. Associate with f a family $\mathcal{M}$ of all $N \times N$ matrices M with the following property: for every monomial α of degree d, the sum of all entries M_{β_1,β_2}, such that $\beta_1\beta_2 = \alpha$, is equal to the coefficient of α in f. In other words, we partition M into subsets T_α corresponding to the possible ways to write α as a product of two monomials, and we impose a condition on the sum of entries in every T_α. It can be shown that the circuit complexity of f can be lower bounded by the minimal rank of the matrices in $\mathcal{M}$.

Note that the sets T_α are of size exponential in d, the degree of f. The structure of the sets is not friendly either. Our first structure theorem for non-commutative circuits, which decomposes non-commutative polynomials to central polynomials, translates to a similar rank problem. However, the matrices $M \in \mathcal{M}$ will be partitioned into sets of size only d (instead of exponential in d). This is thanks to the fact that there are much fewer options to express a non-commutative monomial as a product of other monomials. Our second structure theorem, concerning block-central polynomials, gives a partition into sets of size at most two. The structure of these sets is quite simple too. However, not simple enough to allow us to prove a rank lower bound. In the rank formulation of circuit lower bounds, we can therefore see non-commutative circuits as a first step towards understanding commutative circuit lower bounds.

1.6 Structure of the paper

In Section 2 we outline our proofs of conditional lower bounds for non-commutative circuits and restricted lower bounds on the sum-of-squares complexity. In Section 3 we investigate the structure of non-commutative circuits. In Section 4 we present a connection between circuit complexity of degree-four polynomials and bilinear complexity. In Section 5 we show a reduction from circuit complexity of general polynomials to bilinear complexity of degree-four polynomials, in particular, we prove a conditional lower bound for permanent. In Section 6 we construct a polynomial whose sum-of-squares complexity over the reals in high. In Section 7 we prove a super-linear lower bound for the sum-of-squares complexity over the integers. Finally, in Section 8 we study a family of circuits we call *ordered*, which are in particular multilinear, and prove lower bound for the size of multilinear non-commutative circuits.

2 Overview of proofs

2.1 Conditional lower bounds on non-commutative circuit size

In this section we describe the path that leads from non-commutative circuit complexity to bilinear complexity.

Preliminaries. Let $\mathbb{F}$ be a field. A *non-commutative polynomial* is a formal sum of products of variables and field elements. We assume that the variables do not multiplicatively commute, that is, $xy \neq yx$ whenever $x \neq y$. However, the variables commute with elements of $\mathbb{F}$. The reader can imagine the variables as representing square matrices.

A *non-commutative arithmetic circuit* Φ is a directed acyclic graph as follows. Nodes (or gates) of in-degree zero are labelled by either a variable or a field element in $\mathbb{F}$. All the other nodes have in-degree two and they are labelled by either $+$ or $\times$. The two edges going into a gate v labelled by $\times$ are labelled by *left* and *right*. We denote by $v = v_1 \times v_2$ the fact that (v_1, v) is the left edge going into v, and (v_2, v) is the right edge going into v. (This is to determine the order of multiplication.) The *size* of a circuit Φ is the number of edges in Φ. The integer $\mathcal{C}(f)$ is the size of a smallest circuit computing f.

Note. *Unless stated otherwise, we refer to non-commutative polynomials as polynomials, and to non-commutative circuits as circuits.*

The proof is presented in three parts, which are an exploration of the structure of non-commutative circuits.

Part I: structure of circuits. The starting point of our trail is the structure of polynomials computed by non-commutative circuits, which we now explain. The methods we use are elementary, and are an adaptation of works like [8, 9] to the non-commutative world.

We start by defining the 'building blocks' of polynomials, which we call central polynomials. A homogeneous[7] polynomial f of degree d is called *central*, if there exist integers m and d_0, d_1, d_2 satisfying $d/3 \leq d_0 < 2d/3$ and $d_0 + d_1 + d_2 = d$ so that

$$f = \sum_{i \in [m]} h_i g \overline{h}_i, \tag{2.1}$$

where

(i). the polynomial g, which we call the *body*, is homogeneous of degree $\deg g = d_0$,

(ii). for every $i \in [m]$, the polynomials $h_i, \overline{h}_i$ are homogeneous of degrees $\deg h_i = d_1$ and $\deg \overline{h}_i = d_2$.

The *width* of a homogeneous polynomial f of degree d, denoted $w(f)$, is the smallest integer n so that f can be written as

$$f = f_1 + f_2 + \cdots f_n, \tag{2.2}$$

with each f_i a central polynomial. In Section 3.1 we show that the width of f is at most $O(d^3 \mathcal{C}(f))$, and so lower bounds on width imply lower bounds on circuit complexity. We prove this by induction on the circuit complexity of f.

Part II: degree-four. In the first part, we argued that a lower bound on width implies a lower bound on circuit complexity. In the case of degree-four, a central polynomial has a very simple structure: d_0 is always 2, and so the body must reside in one of three places: left (when $d_1 = 0$), center (when $d_1 = 1$), and right (when $d_1 = 2$). For a polynomial of degree four, we can thus write (2.2) with n at most order $\mathcal{C}(f)$, and each f_i of this special form.

This observation allows us to relate width and bilinear complexity, as the following proposition shows. For a more general statement, see Proposition 4.1, which also shows that the width and bilinear complexity are in fact equivalent.

Proposition 2.1. $w(\mathsf{ID}_k) \geq \mathcal{B}(\mathsf{SOS}_k)$.

Part I and Proposition 2.1 already imply Theorem 1.6, which states that a lower bound on bilinear complexity implies a lower bound on circuit complexity of ID_k.

[7]Recall that a polynomial f is *homogeneous*, if all monomials with a non-zero coefficient in f have the same degree, and that circuit Φ is homogeneous, if every gate in Φ computes a homogeneous polynomial.

Part III: general degree to degree-four. The argument presented in the second step can imply at most a quadratic lower bound on circuit size. To get exponential lower bounds, we need to consider polynomials of higher degrees. We think of the degree of a degree-$4r$ polynomial as divided into 4 groups, for which we try to mimic the special structure from part II: A *block-central* polynomial is a central polynomial so that $d_0 = 2r$ and $d_1 \in \{0, r, 2r\}$. The structure of block-central polynomials is similar to the structure of degree-four central polynomials in that the body is of fixed degree and it has three places it can reside in: left (when $d_1 = 0$), center (when $d_1 = r$), and right (when $d_1 = 2r$). In Section 5 we show that a degree-$4r$ polynomial f can be written as a sum of at most $O(r^3 2^r \mathcal{C}(f))$ block-central polynomials.

We thus reduced the analysis of degree-$4r$ polynomials to the analysis of degree-four polynomial. This reduction comes with a price, a loss of a factor of 2^r. We note that this loss is necessary. The proof is a rather technical case distinction. The idea behind it is a combinatorial property of intervals in the set $[4r]$, which allows us to transform a central polynomial to a sum of 2^r block-central polynomials.

Here is an example of this reduction in the case of the identity polynomial. The *lifted identity polynomial*, LID_r, is the polynomial in variables z_0, z_1 of degree $4r$ defined by

$$\mathsf{LID}_r = \sum_{e \in \{0,1\}^{2r}} z_e z_e \, ,$$

where for $e = (e_1, \dots, e_{2r}) \in \{0,1\}^{2r}$, we define $z_e = \prod_{i=1}^{2r} z_{e_i}$. The lifted identity polynomial is the high-degree counterpart of the identity polynomial, which allows us to prove that a super-linear lower bound implies an exponential one (the corollary is proved in Section 5):

Corollary 2.2. *If $\mathcal{B}(\mathsf{SOS}_k) \geq \Omega(k^{1+\epsilon})$ for some $\epsilon > 0$, then $\mathcal{C}(\mathsf{LID}_r) \geq 2^{\Omega(r)}$.*

To complete the picture, we show that LID_r is reducible to the permanent of dimension $4r$.

Lemma 2.3. *There exists a matrix M of dimension $4r \times 4r$ whose nonzero entries are variables z_0, z_1 so that the permanent of M is LID_r.*

To prove the lemma, the matrix M is constructed explicitly, see Section 5. The conditional lower bound on the permanent, Theorem 1.7, follows from Corollary 2.2 and Lemma 2.3.

An important property that non-commutative permanent shares with its commutative counterpart is completeness for the class of explicit polynomials. This enables us to argue that a super-linear lower bound on the bilinear complexity of an explicit degree-four polynomial implies an exponential lower bound on permanent. In the commutative setting, this a consequence of the VNP completeness of permanent, as given in [32]. In the non-commutative setting, one can prove a similar result [10], see Section 5.1 for more details.

2.2 Restricted lower bounds on sum-of-squares complexity

We now discuss the lower bounds for restricted sum-of-squares problems we prove: an explicit lower bound over $\mathbb{R}$ and a lower bound for SOS_k over integers. For more details and formal definitions, see Sections 6 and 7.

We phrase the problem of lower bounding $\mathcal{S}_{\mathbb{R}}(g)$ in terms of matrices of real vectors. Let $V = \{\mathbf{v}_{i,j} : i, j \in [k]\}$ be a $k \times k$ matrix whose entries are vectors in $\mathbb{R}^n$. We call V a *vector matrix*, and n is called the *height* of V. The matrix V defines a biquadratic polynomial $f(V)$ in $X = \{x_1, \ldots, x_k\}$ and $Y = \{y_1, \ldots, y_k\}$ by

$$f(V) = \sum_{i_1 \leq i_2, j_1 \leq j_2} a_{i_1,i_2,j_1,j_2} x_{i_1} x_{i_2} y_{j_1} y_{j_2},$$

where a_{i_1,i_2,j_1,j_2} is equal to $\mathbf{v}_{i_1,j_1} \cdot \mathbf{v}_{i_2,j_2} + \mathbf{v}_{i_1,j_2} \cdot \mathbf{v}_{i_2,j_1}$, up to a small correction factor which is not important at this point. We can think of the coefficients as given by the permanent of the 2×2 sub-matrix[8] of V define by i_1, i_2 and j_1, j_2.

The following lemma, whose version is proved in Section 6, gives the connection between sum-of-squares complexity and vector matrices.

Lemma 2.4. *Let g be a biquadratic polynomial. Then $\mathcal{S}_{\mathbb{R}}(g) \leq n$ is equivalent to the existence a vector matrix V of height n so that $g = f(V)$.*

As long as it is finite, the height of a vector matrix for any polynomial does not exceed k^2, and a counting argument shows that this holds for “almost” all polynomials. The problem is to construct explicit polynomials that require large height. Even a super-linear lower bound seems nontrivial, since the permanent condition does not talk about inner products of pairs of vectors, but rather about the sum of inner products of two such pairs. In Sections 6 we manage to construct an explicit polynomial which requires near-maximal height $\Omega(k^2)$. In our proof, the coefficients impose (through the 2×2 permanent conditions) either equality or orthognality constraints on the vectors in the matrix, and eventually the existence of many pairwise orthogonal ones. In a crucial way, we employ the fact that over $\mathbb{R}$, if two unit vectors have inner product one, they must be equal. This property[9] fails over $\mathbb{C}$, but it is still possible that even over $\mathbb{C}$ our construction has similar height (of course, if this turns out to be even $k^{1+\epsilon}$, we get an exponential lower bound for non-commutative circuits).

The construction, however, does not shed light on the classical sum-of-squares problem which is concerned specifically with the polynomial SOS_k. In the case of SOS_k, the conditions on the matrix V from Lemma 2.4 are especially nice and simple: (1) all vectors in V are unit vectors, (2) in each row and column the vectors are pairwise orthogonal, and (3) every 2×2 permanent (of inner products) must be zero.

As mentioned in the introduction, the best upper bounds for the sum-of-squares problem have *integer* coefficients, and so a lot of effort was invested into proving lower bounds in the integer case. Despite that, previously known lower bounds do not even reach $2k$. In Section 7 we prove the first super-linear lower bound, $\mathcal{S}_{\mathbb{Z}}(k) = \Omega(k^{6/5})$. Over integers, we take advantage of the fact that the unit vectors in V must have entries in $\{-1, 0, 1\}$ and there is exactly one nonzero entry in each vector. The nonzero coordinate can be thus thought of as a “color” in $[n]$, which is signed by plus or minus. This gives rise to the earlier studied notion of *intercalate matrices* (see, [34] and the book [29]). The integer sum-of-squares problem can thus be phrased in terms of minimizing the number of colors in a signed intercalate matrix, which can be approached as an elementary combinatorial problem.

[8] In some cases, e.g., when $i_1 = i_2$, this matrix can become 1×2, 2×1 or even 1×1, but we still think of it as a 2×2 matrix. This is also where the correction factor comes from.

[9] Here, the inner product of two complex vectors a, b is $\sum_i a_i b_i$, rather than $\sum_i a_i \bar{b}_i$, with $\bar{b}$ the complex conjugate of b.

Our strategy for proving the integer lower bound has three parts. The first step uses a simple counting argument to show that there must exist a sub-matrix in which one color appears in every row and every column. In the second step we show that the permanent conditions give rise to a “forbidden configuration” in such sub-matrices. In the last step we conclude that any matrix without this forbidden configuration must have many colors.

3 Non-commutative circuits

In this section we study the structure of non-commutative circuits. We use the following notation. For a node v in a circuit Φ, we denote by Φ_v the sub-circuit of Φ rooted at v. Every node v computes a polynomial $\widehat{\Phi}_v$ in the obvious way. A *monomial* α is a product of variables, and $\mathsf{COEF}_\alpha(f)$ is the coefficient of α in the polynomial f. Denote by $\deg f$ the degree of f, and if v is a node in a circuit Φ, denote by $\deg v$ the degree of $\widehat{\Phi}_v$.

3.1 Structure of non-commutative circuits

In this section we describe the structure of the polynomials computed by non-commutative circuits. The methods we use are elementary, and are an adaptation of works like [8, 9] to the non-commutative world.

We start by defining the ‘building blocks’ of polynomials, which we call central polynomials. Recall that a polynomial f is *homogeneous*, if all monomials with a non-zero coefficient in f have the same degree, and that circuit Φ is homogeneous, if every gate in Φ computes a homogeneous polynomial. A homogeneous polynomial f of degree d is called *central*, if there exist integers m and d_0, d_1, d_2 satisfying

$$d/3 \leq d_0 < 2d/3 \ \text{ and } \ d_0 + d_1 + d_2 = d$$

so that

$$f = \sum_{i \in [m]} h_i g \overline{h}_i, \tag{3.1}$$

where

(i). the polynomial g is homogeneous of degree $\deg g = d_0$,

(ii). for every $i \in [m]$, the polynomials $h_i, \overline{h}_i$ are homogeneous of degrees $\deg h_i = d_1$ and $\deg \overline{h}_i = d_2$.

Remark 3.1. *In the definition of central polynomial, no assumption on the size of m is made. Hence we can without loss of generality assume that $h_i = c_i\alpha_i$ and $\overline{h}_i = \beta_i$, where α_i is a monomial of degree d_1, β_i is a monomial of degree d_2, and c_i is a field element.*

The *width* of a homogeneous polynomial f of degree d, denoted $w(f)$, is the smallest integer n so that f can be written as

$$f = f_1 + f_2 + \cdots + f_n,$$

where $f_1, \ldots, f_n$ are central polynomials of degree d. The following proposition shows that the width of a polynomial is a lower bound for its circuit complexity. We will later relate width and bilinear complexity.

Proposition 3.2. *Let f be a homogeneous polynomial of degree $d \geq 2$. Then*

$$\mathcal{C}(f) \geq \Omega(d^{-3} w(f)).$$

Proof. We start by observing that the standard homogenization of commutative circuits [31, 3] works for non-commutative circuits as well.

Lemma 3.3. *Let g be a homogeneous polynomial of degree d. Then there exists a homogeneous circuit of size $O(d^2\mathcal{C}(f))$ computing g.*

Assume that we have a homogeneous circuit Φ of size s computing f. We will show that $w(f) \leq ds$. By Lemma 3.3, this implies that $w(f) \leq O(d^3\mathcal{C}(f))$, which completes the proof. Without loss of generality, we can also assume that no gate v in Φ computes the zero polynomial (gates that compute the zero polynomial can be removed, decreasing the circuit size).

For a multiset of pairs of polynomials $\mathcal{H} = \{\langle h_i, \overline{h}_i\rangle : i \in [m]\}$, define

$$g \times \mathcal{H} = \sum_{i\in[m]} h_i g \overline{h}_i \,.$$

Let $\mathcal{G} = \{g_1, \dots, g_t\}$ be the set of homogeneous polynomials g of degree $d/3 \leq \deg g < 2d/3$ so that there exists a gate in Φ computing g. We show that for every gate v in Φ so that $\deg v \geq d/3$ there exist multisets of pairs of homogeneous polynomials $\mathcal{H}_1(v), \dots, \mathcal{H}_t(v)$ satisfying

$$\widehat{\Phi}_v = \sum_{i\in[t]} g_i \times \mathcal{H}_i(v). \tag{3.2}$$

We prove (3.2) by induction on the depth of Φ_v. If $\deg(v) < 2d/3$ then $\widehat{\Phi}_v = g_i \in \mathcal{G}$ for some $i \in [t]$. Thus (3.2) is true, setting $\mathcal{H}_i(v) = \{\langle 1, 1\rangle\}$ and $\mathcal{H}_j(v) = \{\langle 0, 0\rangle\}$ for $j \neq i$ in $[t]$. Otherwise, we have $\deg v \geq 2d/3$. When $v = v_1 + v_2$, we do the following. Since Φ is homogeneous, v_1, v_2 and v have the same degree which is at least $2d/3$. Induction thus implies: for every $e \in \{1, 2\}$,

$$\widehat{\Phi}_{v_e} \quad = \quad \sum_{i\in[t]} g_i \times \mathcal{H}_i(v_e).$$

This gives

$$\widehat{\Phi}_v = \widehat{\Phi}_{v_1} + \widehat{\Phi}_{v_2} = \sum_{i\in[t]} g_i \times \big(\mathcal{H}_i(v_1) \cup \mathcal{H}_i(v_2)\big).$$

When $v = v_1 \times v_2$, we have $\deg v = \deg v_1 + \deg v_2$. Since $\deg v \geq 2d/3$, either (a) $\deg v_1 \geq d/3$ or (b) $\deg v_2 \geq d/3$. In the case (a), by induction,

$$\widehat{\Phi}_{v_1} = \sum_{i\in[t]} g_i \times \mathcal{H}_i(v_1)\,.$$

Defining $\mathcal{H}_i(v) = \{\langle h, \overline{h}\widehat{\Phi}_{v_2}\rangle \ : \ \langle h, \overline{h}\rangle \in \mathcal{H}_i(v_1)\}$, we obtain

$$\widehat{\Phi}_v = \widehat{\Phi}_{v_1}\widehat{\Phi}_{v_2} = \Big(\sum_{i\in[t]} g_i \times \mathcal{H}_i(v_1)\Big)\widehat{\Phi}_{v_2} = \sum_{i\in[t]} g_i \times \mathcal{H}_i(v).$$

Since $\widehat{\Phi}_{v_2}$ is a homogeneous polynomial, $\mathcal{H}_i(v)$ consists of pairs of homogeneous polynomials. In case (b), define $\mathcal{H}_i(v) = \{\langle \widehat{\Phi}_{v_1} h, \overline{h} \rangle \ : \ \langle h, \overline{h} \rangle \in \mathcal{H}_i(v_2)\}$.

Applying (3.2) to the output gate of Φ, we obtain

$$f = \sum_{i \in [t]} g_i \times \mathcal{H}_i \,,$$

where $\mathcal{H}_i$ are multisets of pairs of homogeneous polynomials. For every $i \in [t]$ and every $r \leq d - \deg g_i$, define $\mathcal{H}_i^r = \{\langle h, \overline{h} \rangle \in \mathcal{H}_i \ : \ \deg(h) = r, \ \deg \overline{h} = d - \deg g_i - r\}$. Then $g_i \times \mathcal{H}_i^r$ is a central polynomial. Moreover, since f is homogeneous of degree d, we obtain

$$f = \sum_{i \in [t]} \sum_{r=0}^{d - \deg g_i} g_i \times \mathcal{H}_i^r \,.$$

Since $t \leq s$, the proof is complete. QED

3.2 Degree four polynomials

Before we describe the specific structure of degree four polynomials, let us give general definitions. For a monomial α and a variable x, we say that x *occurs at position* i *in* α, if $\alpha = \alpha_1 x \alpha_2$ and $\deg \alpha_1 = i - 1$. Let $X_1, \dots, X_r$ be (not necessarily disjoint) sets of variables. For a polynomial f, let $f[X_1, \dots, X_r]$ be the homogeneous polynomial of degree r so that for every monomial α,

$$\mathsf{COEF}_\alpha(f[X_1, \dots, X_r]) = \begin{cases} \mathsf{COEF}_\alpha(f) & \text{if } \alpha = x_1 x_2 \cdots x_r \text{ with } x_i \in X_i \text{ for every } i \in [r], \\ 0 & \text{otherwise.} \end{cases}$$

In other words, $f[X_1, \dots X_r]$ is the part of f consisting of monomials degree r with the property that if a variable x occurs at a position i then $x \in X_i$.

Claim 3.4. *Let f be a central polynomial so that $f = f[X_1, X_2, X_3, X_4]$. Then, either*

$$f = g[X_1, X_2] h[X_3, X_4] \quad \text{or} \quad f = \sum_{i \in [m]} h_i[X_1] g[X_2, X_3] \overline{h}_i[X_4] \,,$$

where $g, h, h_i, \overline{h}_i$ are some polynomials.

Proof. As f is central of degree four, $\deg g = d_0 = 2$, and $d_1 \in \{0, 1, 2\}$. If $d_1 = 1$, then $d_2 = 1$ as well, and

$$f = f[X_1, X_2, X_3, X_4] = \sum_{i \in [m]} \big(h_i g \overline{h}_i\big)[X_1, X_2, X_3, X_4] = \sum_{i \in [m]} h_i[X_1] g[X_2, X_3] \overline{h}_i[X_4].$$

If $d_1 = 0$, then $d_2 = 2$, and

$$f = f[X_1, X_2, X_3, X_4] = \sum_{i \in [m]} \big(g \overline{h}_i\big)[X_1, X_2, X_3, X_4] = g[X_1, X_2] \Big(\sum_{i \in [m]} h_i[X_1] \overline{h}_i[X_4] \Big).$$

A similar argument holds when $d_1 = 2$. QED

Claim 3.4 implies the following lemma.

Lemma 3.5. *If $f = f[X_1, X_2, X_3, X_4]$, then $w(f)$ is the smallest n so that f can be written as $f = f_1 + \cdots + f_n$, where for every $t \in [n]$, either*

(a) $f_t = g_t[X_1, X_2]h_t[X_3, X_4]$, *or*

(b) $f_t = \sum_{i \in [m]} h_{t,i}[X_1]g_t[X_2, X_3]\overline{h}_{t,i}[X_4]$,

where $g_t, h_t, h_{t,i}, \overline{h}_{t,i}$ are some polynomials.

4 Degree four and bilinear complexity

In this section we related the width of degree four polynomials to their highway number. We consider polynomials of a certain structure. Let f be a polynomial in variables $X = \{x_1, \ldots, x_k\}$ and $Y = \{y_1, \ldots, y_k\}$ so that $f = f[X, Y, X, Y]$, i.e.,

$$f = \sum_{i_1, j_1, i_2, j_2 \in [k]} a_{i_1,j_1,i_2,j_2} x_{i_1} y_{j_1} x_{i_2} y_{j_2}. \tag{4.1}$$

For a non-commutative polynomial g, we define $g^{(c)}$ to be the polynomial g understood as a commutative polynomial. For example, if $g = xy + yx$, then $g^{(c)} = 2xy$. We say that f is (X, Y)-*symmetric*, if for every $i_1, j_1, i_2, j_2 \in [k]$,

$$a_{i_1,j_1,i_2,j_2} = a_{i_2,j_1,i_1,j_2} = a_{i_1,j_2,i_2,j_1} = a_{i_2,j_2,i_1,j_1} .$$

In particular, if f is of the form (4.1), the polynomial $f^{(c)}$ is biquadratic. In the following proposition, we relate the width of a polynomial f and $\mathcal{B}(f^{(c)})$.

Proposition 4.1. *Let f be a homogeneous polynomial of degree four of the form* (4.1). *Then*

(i). $\mathcal{B}(f^{(c)}) \leq w(f)$, *and*

(ii). If char $\mathbb{F} \neq 2$ and f is (X, Y)-symmetric, then $w(f) \leq 4\mathcal{B}(f^{(c)})$.

Proof. We start by proving (i). Using Lemma 3.5, we can write $f = f_1 + \cdots + f_n$, where for every $t \in [n]$, either

(a) $f_t = g_t[X, Y]h_t[X, Y]$, or

(b) $f_t = \sum_{i \in [m]} h_{t,i}[X]g_t[Y, X]\overline{h}_{t,i}[Y]$.

The commutative polynomial $f_t^{(c)}$ is a product of two bilinear forms in X and Y: in case (a), of $g_t[X,Y]^{(c)}$ and $h_t[X,Y]^{(c)}$, and in case (b), of $g_t[Y,X]^{(c)}$ and $\sum_{i\in[m]} h_{t,i}[X]^{(c)}\overline{h}_{t,i}[Y]^{(c)}$. Altogether $f^{(c)} = f_1^{(c)} + \cdots + f_n^{(c)}$, where each $f_t^{(c)}$ is a product of two bilinear forms, and hence $\mathcal{B}(f^{(c)}) \leq n$.

We now prove (ii). Assume that

$$f^{(c)} = z_1 z_1' + \cdots + z_n z_n', \tag{4.2}$$

where z_t and z_t', $t \in [n]$, are bilinear in X and Y. Write

$$z_t = \sum_{j\in[k]} x_j g_{t,j} \qquad \text{and} \qquad z_t' = \sum_{j\in[k]} x_j h_{t,j},$$

where $g_{t,j}$ and $h_{t,j}$ are homogeneous degree one polynomials in the variables Y. Let f_t be the non-commutative polynomial

$$\begin{aligned} f_t = \sum_m (x_j g_{t,m}) \sum_j (x_j h_{t,j}) + \sum_m (x_j h_{t,m}) \sum_j (x_j g_{t,j}) + \\ + \sum_m (x_m \sum_j (g_{t,j} x_j) h_{t,m}) + \sum_m (x_m \sum_j (h_{t,j} x_j) g_{t,m}) \end{aligned}$$

with summations ranging over $[k]$. We can see that f_t is a sum of four central polynomials. It is therefore sufficient to show that

$$f = \frac{1}{4}(f_1 + \cdots + f_n). \tag{4.3}$$

First, note that $f_t^{(c)} = 4 z_t z_t'$ and hence

$$f^{(c)} = \frac{1}{4}(f_1^{(c)} + \cdots + f_n^{(c)}) \tag{4.4}$$

Second, note that if g is (X,Y)-symmetric and $\alpha = x_{i_1} y_{j_1} x_{i_2} y_{j_2}$ is a non-commutative monomial, then

$$\mathsf{COEF}_{\alpha^{(c)}}(g^{(c)}) = N(i_1, j_1, i_2, j_2)\mathsf{COEF}_{\alpha}(g),$$

where

$$N(i_1, j_1, i_2, j_2) = \begin{cases} 1 & \text{if } \ i_1 = i_2 \text{ and } j_1 = j_2, \\ 2 & \text{if } \ i_1 = i_2 \text{ and } j_1 \neq j_2, \\ 2 & \text{if } \ i_1 \neq i_2 \text{ and } j_1 = j_2, \\ 4 & \text{if } \ i_1 \neq i_2 \text{ and } j_1 \neq j_2. \end{cases}$$

Fix a monomial $\alpha = x_{i_1} y_{j_1} x_{i_2} y_{j_2}$ and consider the coefficient of α in the two sides of (4.3). Since f is (X,Y)-symmetric, we have

$$\mathsf{COEF}_{\alpha^{(c)}}(f^{(c)}) = N(i_1, j_1, i_2, j_2)\mathsf{COEF}_{\alpha}(f)\,.$$

Hence (4.4) tells us that

$$\mathsf{COEF}_\alpha(f) = \frac{\mathsf{COEF}_{\alpha^{(c)}}\left(f_1^{(c)} + \cdots + f_n^{(c)}\right)}{4N(i_1, j_1, i_2, j_2)}.$$

Since f_t is (X,Y)-symmetric, we have $\mathsf{COEF}_{\alpha^{(c)}} f_t^{(c)} = N(i_1, j_1, i_2, j_2)\mathsf{COEF}_\alpha(f_t)$. Hence

$$\mathsf{COEF}_\alpha(f) = \frac{1}{4}(\mathsf{COEF}_\alpha(f_1) + \cdots + \mathsf{COEF}_\alpha(f_t))$$

and equation (4.3) follows. QED

Proof of Theorem 1.6. Recall the definition of the identity polynomial,

$$\mathsf{ID}_k = \sum_{i,j\in[k]} x_i y_j x_i y_j .$$

The commutative polynomial $\mathsf{ID}_k^{(c)}$ is the polynomial SOS_k

$$\mathsf{SOS}_k = \sum_{i\in[k]} x_i^2 \sum_{j\in[k]} y_j^2 .$$

The theorem follows from Proposition 3.2 and 4.1. QED

Let us note that it is not necessary to separate variables in ID_k into two disjoint sets X and Y. In the non-commutative setting, this is just a cosmetic detail. This is a consequence of a more general phenomenon discussed in Section 8.1.

Remark 4.2. $w(\mathsf{ID}_k) = w(\sum_{i,j\in[k]} x_i x_j x_i x_j)$.

Proof. Denote $g = \sum_{i,j\in[k]} x_i x_j x_i x_j$. Clearly $w(g) \le w(\mathsf{ID}_k)$ and we must prove the opposite inequality. Let $X := \{x_i : i \in [k]\}$. Let us write ID_k as $\sum_{i,j\in[k]} x_{i,0} x_{j,1} x_{i,0} x_{j,1}$ in variables $X_b = \{x_{i,b} : i \in [k]\}$, $b \in \{0,1\}$. If f is a homogeneous polynomial of degree r in X and $e = \langle e_1, \ldots e_r\rangle \in \{0,1\}^n$, let f^e be the polynomial s.t. $f^e = f^e[X_{e_1}, \ldots X_{e_r}]$ and

$$\mathsf{COEF}_{x_{i_1,e_1}\ldots x_{i_r,e_r}}(f^e) = \mathsf{COEF}_{x_{i_1}\ldots x_{i_r}}(f) .$$

Hence $\mathsf{ID}_k = g^{0101}$. Moreover, if $g = f_1 + \ldots f_n$ then $g^{0101} = f_1^{0101} + \ldots f_n^{0101}$. It is thus sufficient to prove that if f_j is central, then $f_j^{(0101)}$ is also central. This follows from the following: $(gh)^{(0101)} = g^{(01)}h^{(01)}$, if g, h are homogeneous polynomials of degree two, and $(hg\overline{h})^{(0101)} = h^{(0)}g^{(10)}\overline{h}^{\,(1)}$, if $h, g, \overline{h}$ are homogeneous polynomials of degrees one, two and one. QED

5 Higher degrees

In this section, we show that a sufficiently strong lower bound on the width of a degree four polynomial implies an exponential lower bound on the width, and hence circuit size, of a related high degree polynomial.

Let f be a homogeneous polynomial of degree $4r$. We assume that f contains only two variables z_0 and z_1. We define $f^{(\lambda)}$ to be the polynomial obtained by replacing degree r monomials in f by new variables. Formally, for every monomial α of degree r in variables z_0, z_1, introduce a new variable x_α. The polynomial $f^{(\lambda)}$ is defined as the homogenous degree four polynomial in the 2^r variables $X = \{x_\alpha : \deg \alpha = r\}$ satisfying

$$\mathsf{COEF}_{x_{\alpha_1} x_{\alpha_2} x_{\alpha_3} x_{\alpha_4}}(f^{(\lambda)}) = \mathsf{COEF}_{\alpha_1 \alpha_2 \alpha_3 \alpha_4}(f)\,. \tag{5.1}$$

Remark 5.1. *Let g be a homogeneous degree four polynomial in k variables. If $k \leq 2^r$, then there exists a polynomial f of degree $4r$ in variables z_0, z_1 such that $g = f^{(\lambda)}$ (up to a renaming of variables).*

Proof. For $e = (e_1, \ldots, e_r) \in \{0,1\}^r$, let z_e be the monomial $\prod_{j=1}^r z_{e_j}$. If $k \leq 2^r$ and $i \in [k]$, let $(i) \in \{0,1\}^r$ be the binary representation of i. If

$$g = \sum_{i_1 j_1 i_2 j_2 \in [k]} a_{i_1 j_1 i_2 j_2} x_{i_1} x_{j_1} x_{i_2} x_{j_2}\,,$$

let

$$f = \sum_{i_1 j_1 i_2 j_2 \in [k]} a_{i_1 j_1 i_2 j_2} z_{(i_1)} z_{(j_1)} z_{(i_2)} z_{(j_2)}\,.$$

QED

We now relate $w(f)$ and $w(f^{(\lambda)})$. To do so, we need a modified version of Proposition 3.2. Let f be a homogeneous polynomial of degree $4r$. We say that f is *block-central*, if either

I. $f = gh$, where g, h are homogeneous polynomials with $\deg g = \deg h = 2r$, or

II. $f = \sum_{i \in [m]} h_i g \overline{h}_i$, where $g, h_i, \overline{h}_i$ are homogeneous polynomials of degrees $\deg g = 2r$ and $\deg h_i = \deg \overline{h}_i = r$ for every $i \in [m]$.

Every block-central polynomial is also central. The following lemma shows that every central polynomial can be written as a sum of 2^r block-central polynomials. The lemma thus enables us to consider a simpler problem, i.e., lower bounding the width with respect to block-central polynomials. However, this simplification comes with a price, namely, a loss of a factor of 2^r.

Lemma 5.2. *Let f be a central polynomial of degree $4r$ in two variables z_0, z_1. Then there exist $n \leq 2^r$ and block-central polynomials $f_1, \ldots, f_n$ so that $f = f_1 + \cdots + f_n$.*

Proof. Let $M(k)$ be the set of monomials in variables z_0, z_1 of degree exactly k. The size of $M(k)$ is 2^k. As f is central, by Remark 3.1, we can write f as

$$f = \sum_{\alpha \in M(d_1), \omega \in M(d_2)} c(\alpha, \omega) \alpha G \omega \,, \tag{5.2}$$

where $c(\alpha, \omega)$ is a field element, G is a homogeneous polynomial of degree d_0 with $4r/3 \le d_0 < 8r/3$, and $d_0 + d_1 + d_2 = 4r$.

Our goal is to write f is a sum of block-central polynomials, namely, we wish to write f as a sum of polynomials of either type I or type II. We use the parameters d_0, d_1, d_2 to determine the type of these polynomials, according to the following case distinction.

Assume first that $d_0 + 2d_1 \le 3r$. We express f as a sum of type I polynomials. There are two sub-cases to consider.

1. $d_0 + d_1 \le 2r$**:** Every monomial $\omega \in M(d_2)$ can be written as $\omega_1\omega_2$, where $\omega_1 \in M(t)$, $\omega_2 \in M(d_2 - t)$ and $t = 2r - (d_0 + d_1)$. Then (5.2) can be written as

$$f = \sum_{\alpha \in M(d_1), \omega_1 \in M(t)} f_{\alpha,\omega_1} \,,$$

where

$$f_{\alpha,\omega_1} = (\alpha G \omega_1)\Big(\sum_{\omega_2 \in M(t-d_2)} c(\alpha, \omega_1\omega_2)\omega_2 \Big).$$

As $d_2 - t = 2r$, each f_{α,ω_1} is of type I. There are at most $|M(d_1)||M(t)| = 2^{2r-d_0}$ such f_{α,ω_1}. Since $d_0 \ge 4r/3$, there are at most $2^{2r/3}$ of them.

2. $d_0 + d_1 > 2r$**:** We can write $G = \sum_{\gamma \in M(t)} G_\gamma \gamma$, where $t = d_0 + d_1 - 2r$, and G_γ are some polynomials of degree $d_0 - t$. Then

$$f = \sum_{\alpha \in M(d_1), \gamma \in M(t)} f_{\alpha,\gamma} \,,$$

where

$$f_{\alpha,\gamma} = (\alpha G_\gamma)\Big(\sum_{\omega \in M(d_2)} c(\alpha, \omega)\gamma\omega \Big).$$

Each $f_{\alpha,\gamma}$ is of type I, and the number of such $f_{\alpha,\gamma}$ is $2^{d_0+2d_1-2r} \le 2^r$, as $d_0 + 2d_1 \le 3r$.

If $d_0 + 2d_2 \le 3r$, the argument is analogous. Hence we are in the situation $d_0 + 2d_1 > 3r$ and $d_0 + 2d_2 > 3r$. In this case, we express f as a sum of centralpolynomials of type II. There are four sub-cases to consider.

1. $d_1 \ge r$ **and** $d_2 \ge r$**:** For $\alpha \in M(d_1)$, write $\alpha = \alpha_1\alpha_2$ with $\alpha_1 \in M(r)$ and $\alpha_2 \in M(d_1 - r)$. For $\omega \in M(d_2)$, write $\omega = \omega_1\omega_2$ with $\omega_1 \in M(d_2 - r)$ and $\omega_2 \in M(r)$. Then

$$f = \sum_{\alpha_2 \in M(d_1 - r), \omega_1 \in M(d_2 - r)} f_{\alpha_2,\omega_1} \,,$$

where

$$f_{\alpha_2,\omega_1} = \sum_{\alpha_1,\omega_2 \in M(r)} c(\alpha_1\alpha_2, \omega_1\omega_2)\alpha_1(\alpha_2 G\omega_1)\omega_2 \,.$$

Each f_{α_2,ω_1} is of type II. There are $2^{2r-d_0} \leq 2^{2r/3}$ such f_{α_2,ω_1}, since $d_0 \geq 4r/3$.

2. $d_1 < r$ **and** $d_2 \geq r$**:** Write $G = \sum_{\gamma \in M(r-d_1)} \gamma G_\gamma$, where G_γ is a homogeneous polynomial of degree $d_0 - (r - d_1)$. Write $\omega = \omega_1\omega_2$ with $\omega_1 \in M(d_2 - r)$ and $\omega_2 \in M(r)$. Then

$$f = \sum_{\gamma \in M(r-d_1), \omega_1 \in M(d_2-r)} f_{\gamma,\omega_1} \,,$$

where

$$f_{\gamma,\omega_1} = \sum_{\alpha \in M(d_1), \omega_2 \in M(r)} c(\alpha, \omega_1\omega_2)\alpha\gamma(G_\gamma\omega_1)\omega_2 \,.$$

Each f_{γ,ω_1} is of type II, and there are $2^{d_2-d_1} < 2^r$ such f_{γ,ω_1}, since $d_0 + 2d_1 > 3r$.

3. $d_1 \geq r$ **and** $d_2 < r$**:** This is the previous case with d_2 and d_1 interchanged.

4. $d_1 < r$ **and** $d_2 < r$**:** Write $G = \sum_{\gamma_1 \in M(r-d_1), \gamma_2 \in M(r-d_2)} \gamma_1 G_{\gamma_1,\gamma_2}\gamma_2$, where G_{γ_1,γ_2} is a homogeneous polynomial of degree $2r$. Then

$$f = \sum_{\gamma_1 \in M(r-d_1), \gamma_2 \in M(r-d_2)} f_{\gamma_1,\gamma_2},$$

where

$$f_{\gamma_1,\gamma_2} = \sum_{\alpha \in M(d_1), \omega \in M(d_2)} c(\alpha,\omega)\alpha\gamma_1 G_{\gamma_1,\gamma_2}\gamma_2\omega.$$

Each f_{γ_1,γ_2} is of type II, and there are $2^{d_0-2r} \leq 2^{2r/3}$ such f_{γ_1,γ_2}, since $d_0 \leq 8r/3$.

QED

We can now relate the width of f and $f^{(\lambda)}$.

Proposition 5.3. *Let f be a homogeneous polynomial of degree $4r$ in the variables z_0, z_1. Then $w(f) \geq 2^{-r}w(f^{(\lambda)})$.*

Proof. Assume $w(f) = n$. Lemma 5.2 implies $f = f_1 + \cdots + f_{n'}$, where $n' \leq 2^r n$ and f_j are block-central polynomials. Equation (5.1) implies

$$f^{(\lambda)} = f_1^{(\lambda)} + \cdots + f_{n'}^{(\lambda)}.$$

It is thus sufficient to show that every $f_t^{(\lambda)}$ is a central polynomial, for then $w(f^{(\lambda)}) \leq n' \leq 2^r n$.

In order to do so, let us extend the definition of $(.)^{(\lambda)}$ as follows. If g is a polynomial of degree ℓr in the variables z_0, z_1, let $g^{(\lambda)}$ be the homogeneous polynomial of degree ℓ in X so that

$$\mathsf{COEF}_{x_{\alpha_1}\cdots x_{\alpha_k}}(g^{(\lambda)}) = \mathsf{COEF}_{\alpha_1\cdots\alpha_k}(g).$$

If g, h are homogeneous polynomials whose degree is divisible by r, we obtain $(gh)^{(\lambda)} = g^{(\lambda)}h^{(\lambda)}$. Hence if $f_t = g_t h_t$ a block-centralpolynomial of type I, then $f_t^{(\lambda)} = g_t^{(\lambda)} h_t^{(\lambda)}$ is a central polynomial of type (a) according to Lemma 3.5 with $X = X_1 = X_2 = X_3 = X_4$. If $f_t = \sum_i h_{t,i} g_t \overline{h}_{t,i}$ is a block-centralpolynomial of type II, $f_t^{(\lambda)} = \sum_i h_{t,i}^{(\lambda)} g_t^{(\lambda)} \overline{h}_{t,i}{}^{(\lambda)}$, and hence $f_t^{(\lambda)}$ is a centralpolynomial of type (b) according to Lemma 3.5. QED

By Remark 5.1, we can start with a degree four polynomial in $k \leq 2^r$ variables and "lift" it to a polynomial f of degree $4r$ such that $f^{(\lambda)} = g$. We can then deduce that a sufficiently strong lower bound on the bilinear complexity of g implies an exponential lower bound for the circuit complexity of f. We apply this to the specific case of the identity polynomial. The *lifted identity polynomial*, LID_r, is the polynomial in variables z_0, z_1 of degree $4r$ defined by

$$\mathsf{LID}_r = \sum_{e \in \{0,1\}^{2r}} z_e z_e \,,$$

where for $e = (e_1, \ldots, e_s) \in \{0,1\}^s$, we define $z_e = \prod_{i=1}^{s} z_{e_i}$.

Corollary 5.4 (Corollary 2.2 restated)**.** *If $\mathcal{B}(\mathsf{SOS}_k) \geq \Omega(k^{1+\epsilon})$ for some $\epsilon > 0$, then $\mathcal{C}(\mathsf{LID}_r) \geq 2^{\Omega(r)}$.*

Proof. The definition of LID_r can be equivalently written as

$$\mathsf{LID}_r = \sum_{e_1, e_2 \in \{0,1\}^r} z_{e_1} z_{e_2} z_{e_1} z_{e_2} \,.$$

By definition, $\mathsf{LID}_r^{(\lambda)} = \sum_{i,j \in [k]} x_i x_j x_i x_j$ with $k = 2^r$. Hence, by Remark 4.2, $w(\mathsf{LID}_r^{(\lambda)}) = w(\mathsf{ID}_k)$. By Proposition 5.3, $w(\mathsf{LID}_r) \geq 2^{-r} w(\mathsf{LID}_r^{(\lambda)})$. Hence $w(\mathsf{LID}_r) \geq 2^{-r} w(\mathsf{ID}_k)$. By Proposition 4.1, $w(\mathsf{ID}_k) \geq \mathcal{B}(\mathsf{ID}_k)$. If $\mathcal{B}(\mathsf{ID}_k) \geq c k^{1+\epsilon}$ for some constants $c, \epsilon > 0$, we have $w(\mathsf{LID}_r) \geq c 2^{-r} 2^{r(1+\epsilon)} = c 2^{\epsilon r}$. By Proposition 3.2, $\mathcal{C}(\mathsf{LID}_r) \geq \Omega(r^{-3} 2^{\epsilon r}) = 2^{\Omega(r)}$. QED

One motivation for studying the lifted identity polynomial is that we believe it is hard for non-commutative circuits. However, note that an apparently similar polynomial has small circuit size. For $e = (e_1, \ldots, e_s) \in \{0,1\}^s$, let $e^\star = (e_s, \ldots, e_1)$. The polynomial

$$\sum_{e \in \{0,1\}^{2r}} z_e z_{e^\star} \,,$$

has a non-commutative circuit of linear size. This result can be found in [21], where it is also shown that the non-commutative formula complexity of this polynomial is exponential in r.

We now show that LID_r is reducible to the permanent of dimension $4r$.

Lemma 5.5 (Lemma 2.3 restated)**.** *There exists a matrix M of dimension $4r \times 4r$ whose nonzero entries are variables z_0, z_1 so that the permanent of M is LID_r.*

Proof. For $j \in \{0,1\}$, let D_j be the $2r \times 2r$ matrix with z_j on the diagonal and zero everywhere else. The matrix M is defined as

$$M = \begin{bmatrix} D_0 & D_1 \\ D_1 & D_0 \end{bmatrix}.$$

The permanent of M taken row by row is

$$\mathsf{PERM}(M) = \sum_{\sigma} M_{1,\sigma(1)} M_{2,\sigma(2)} \cdots M_{4r,\sigma(4r)},$$

where σ is a permutation of $[4r]$. The permutations that give nonzero value in $\mathsf{PERM}(M)$ satisfy: for every $i \in [2r]$, if $\sigma(i) = i$ then $\sigma(2r+i) = 2r+i$, and if $\sigma(i) = 2r+i$ then $\sigma(2r+i) = i$. By definition of M, this means that for every such σ and $i \in [2r]$, $M_{i,\sigma(i)} = M_{i+2r,\sigma(i+2r)}$. Moreover, given the values of such a σ on $[2r]$, it can be uniquely extended to all of $[4r]$. QED

Theorem 1.7 follows from Corollary 2.2 and Lemma 2.3.

5.1 Explicit polynomials and completeness of non-commutative permanent

The (conditional) exponential lower bound on the circuit size of permanent can be significantly generalized. An important property that non-commutative permanent shares with its commutative counterpart is completeness for the class of explicit polynomials. This enables us to argue that a super-linear lower bound on width of an explicit degree four polynomial implies an exponential lower bound on permanent.

Let $\{f_k\}$ be an infinite family of non-commutative polynomials over $\mathbb{F}$ so that every f_k has at most $p(k)$ variables and degree at most $p(k)$, where $p : \mathbb{N} \to \mathbb{N}$ is a polynomial. We call $\{f_k\}$ *explicit*, if there exists a polynomial time algorithm which, given k and a monomial α is input, computes $\mathsf{COEF}_\alpha(f_k)$. Hence PERM_k and other families of polynomials are explicit in this sense. In the commutative setting, the following theorem is a consequence of the VNP completeness of permanent, as given in [32]. In the non-commutative setting, one can prove a similar result [10].

Theorem 5.6. *Assume that $\{f_k\}$ is an explicit family of non-commutative polynomials such that $\mathcal{C}(f_k) \geq 2^{\Omega(k)}$. Then $\mathcal{C}(\mathrm{PERM}_k) \geq 2^{\Omega(k)}$.*

Proof of Theorem 1.8 For a commutative biquadratic polynomial in k variables

$$f = \sum_{i_1,j_1,i_2,j_2 \in [k]} a_{i_1,j_1,i_2,j_2} x_{i_1} y_{j_1} x_{i_2} y_{j_2} \, ,$$

define f' as the non-commutative polynomial

$$f' = \sum_{i_1,j_1,i_2,j_2 \in [k]} a_{i_1,j_1,i_2,j_2} x_{i_1} y_{j_1} x_{i_2} y_{j_2} \, .$$

This is to guarantee that $f' = f'[X,Y,X,Y]$ and $(f')^{(c)} = f$ is as required in Proposition 4.1. Let r be the smallest integer so that $2^r \geq 2k$. Let $f^\star$ be the polynomial given by Remark 5.1 so that $(f^\star)^{(\lambda)} = f'$. If f is explicit, $f^\star$ is explicit.

Let $\{f_k\}$ be as in the assumption. As in the proof of Corollary 2.2, we conclude that $f_k^\star$ require exponential size non-commutative circuits. By Theorem 5.6, this implies an exponential lower bound for permanent.

QED

6 Real sum-of-squares

In this section, we prove Theorem 1.9. We construct a real biquadratic polynomial f in the variables $X = \{x_1, \dots, x_k\}$ and $Y = \{y_1, \dots, y_k\}$ over $\mathbb{R}$, so that f can be written as $f = \sum_{i\in[n]} z_i^2$ with z_i bilinear in X, Y, but every such n is at least $k^2/4$. The construction of f is in polynomial time with respect to the length of the binary representation of k.

Remark 6.1. *In the case of $\mathbb{R}$, the condition that z_i are bilinear is satisfied automatically, provided z_i is a polynomial.*

6.1 Real sums-of-squares and vector matrices

We phrase the problem of lower bounding $\mathcal{S}_{\mathbb{R}}(f)$ in terms of matrices of real vectors. Let $V = \{\mathbf{v}_{i,j} : i \in [r], j \in [s]\}$ be a matrix whose entries are vectors in $\mathbb{R}^n$. We call V a *vector matrix*, and n is called the *height* of V. Let $U = \{\mathbf{u}_{i,j} : i \in [r], j \in [s]\}$ be a vector matrix of arbitrary height. We say that U and V are *equivalent,* if for every $i_1, i_2 \in [r]$, $j_1, j_2 \in [s]$,

$$\mathbf{v}_{i_1,j_1} \cdot \mathbf{v}_{i_2,j_2} = \mathbf{u}_{i_1,j_1} \cdot \mathbf{u}_{i_2,j_2},$$

where for two vectors $\mathbf{w}_1, \mathbf{w}_2$ in $\mathbb{R}^m$, $\mathbf{w}_1 \cdot \mathbf{w}_2$ is the standard inner product in $\mathbb{R}^m$. We say that U and V are *similar*, if for every $i_1, i_2 \in [r]$ and $j_1, j_2 \in [s]$

$$\mathbf{v}_{i_1,j_1} \cdot \mathbf{v}_{i_2,j_2} + \mathbf{v}_{i_1,j_2} \cdot \mathbf{v}_{i_2,j_1} = \mathbf{u}_{i_1,j_1} \cdot \mathbf{u}_{i_2,j_2} + \mathbf{u}_{i_1,j_2} \cdot \mathbf{u}_{i_2,j_1}\,.$$

It is more convenient to consider the four different cases of this equality:

$$\mathbf{v}_{i,j} \cdot \mathbf{v}_{i,j} = \mathbf{u}_{i,j} \cdot \mathbf{u}_{i,j}\,. \tag{6.1}$$

$$\mathbf{v}_{i,j_1} \cdot \mathbf{v}_{i,j_2} = \mathbf{u}_{i,j_1} \cdot \mathbf{u}_{i,j_2}\,,\ \text{if } j_1 \neq j_2\,. \tag{6.2}$$

$$\mathbf{v}_{i_1,j} \cdot \mathbf{v}_{i_2,j} = \mathbf{u}_{i_1,j} \cdot \mathbf{u}_{i_1,j}\,,\ \text{if } i_1 \neq i_2\,. \tag{6.3}$$

$$\mathbf{v}_{i_1,j_1} \cdot \mathbf{v}_{i_2,j_2} + \mathbf{v}_{i_1,j_2} \cdot \mathbf{v}_{i_2,j_1} = \mathbf{u}_{i_1,j_1} \cdot \mathbf{u}_{i_2,j_2} + \mathbf{u}_{i_1,j_2} \cdot \mathbf{u}_{i_2,j_1}\,,\ \text{if } i_1 \neq i_2,\ j_1 \neq j_2\,. \tag{6.4}$$

A $k \times k$ vector matrix V defines a polynomial $f(V)$ in the variables X, Y by

$$f(V) = \sum_{i_1 \leq i_2, j_1 \leq j_2} a_{i_1,i_2,j_1,j_2} x_{i_1} x_{i_2} y_{j_1} y_{j_2},$$

with

$$a_{i_1,i_2,j_1,j_2} = \begin{cases} \mathbf{v}_{i,j}\cdot\mathbf{v}_{i,j} & \text{if } i_1 = i_2 = i,\ j_1 = i_2 = j, \\ 2\mathbf{v}_{i,j_1}\cdot\mathbf{v}_{i,j_2} & \text{if } i_1 = i_2 = i,\ j_1 < j_2, \\ 2\mathbf{v}_{i_1,j}\cdot\mathbf{v}_{i_2,j} & \text{if } i_1 < i_2,\ j_1 = j_2 = j, \\ 2(\mathbf{v}_{i_1,j_1}\cdot\mathbf{v}_{i_2,j_2} + \mathbf{v}_{i_1,j_2}\cdot\mathbf{v}_{i_2,j_1}) & \text{if } i_1 < i_2,\ j_1 < j_2. \end{cases} \tag{6.5}$$

Note that if U and V are similar then $f(U) = f(V)$.

Lemma 6.2. *If V be a $k \times k$ a vector matrix. Then the following are equivalent:*

(i). There exist bilinear forms $z_1, \ldots, z_n$ so that $f(V) = \sum_{i\in[n]} z_i^2$.

(ii). There exists a vector matrix U of height n so that U and V are similar.

Proof. Assume that

$$f(V) = \sum_{i\in[n]} z_i^2, \tag{6.6}$$

where each z_i is bilinear. For $\ell \in [n]$ and $i, j \in [k]$, let $u_{i,j}[\ell]$ be the coefficient of $x_i y_j$ in z_ℓ, and let $\mathbf{u}_{i,j} = (u_{i,j}[1], \ldots, u_{i,j}[n])$. Let $U = \{\mathbf{u}_{i,j} : i, j \in [k]\}$ be the $k \times k$ vector matrix of height n. Equation (6.6) can be written as

$$f(V) = \Big(\sum_{i,j\in[k]} \mathbf{u}_{i,j} x_i y_j\Big)\cdot\Big(\sum_{i,j\in[k]} \mathbf{u}_{i,j} x_i y_j\Big). \tag{6.7}$$

The right hand side of (6.7) can be written as

$$\sum_{i,j} \big((\mathbf{u}_{i,j}\cdot\mathbf{u}_{i,j}) x_i^2 y_j^2\big) + 2\sum_{i,j_1<j_2} \big((\mathbf{u}_{i,j_1}\cdot\mathbf{u}_{i,j_2}) x_i^2 y_{j_1} y_{j_2}\big) + 2\sum_{i_1<i_2,j} \big((\mathbf{u}_{i_1,j}\cdot\mathbf{u}_{i_2,j}) x_{i_1} x_{i_2} y_j^2\big) +$$
$$+2\sum_{i_1<i_2,j_1<j_2} \big((\mathbf{u}_{i_1,j_1}\cdot\mathbf{u}_{i_2,j_2} + \mathbf{u}_{i_1,j_2}\cdot\mathbf{u}_{i_2,j_1}) x_{i_1} x_{i_2} y_{j_1} y_{j_2}\big).$$

Comparing the coefficients on the left and right hand side of (6.7), we obtain

$$a_{i_1,i_2,j_1,j_2} = \begin{cases} \mathbf{u}_{i,j}\cdot\mathbf{u}_{i,j} & \text{if } i_1 = i_2 = i,\ j_1 = i_2 = j, \\ 2\mathbf{u}_{i,j_1}\cdot\mathbf{u}_{i,j_2} & \text{if } i_1 = i_2 = i,\ j_1 < j_2, \\ 2\mathbf{u}_{i_1,j}\cdot\mathbf{u}_{i_2,j} & \text{if } i_1 < i_2,\ j_1 = j_2 = j, \\ 2(\mathbf{u}_{i_1,j_1}\cdot\mathbf{u}_{i_2,j_2} + \mathbf{u}_{i_1,j_2}\cdot\mathbf{u}_{i_2,j_1}) & \text{if } i_1 < i_2,\ j_1 < j_2. \end{cases} \tag{6.8}$$

By (6.5), this means that U and V are similar. Conversely, if U is a vector matrix similar to V, we obtain (6.6) by means of (6.5), (6.7) and (6.8). QED

In particular, the lemma shows that $f(V)$ *can always* be written as a sum of real bilinear squares. Moreover, the proof of the lemma entails the converse: if f can be written as sum of bilinear squares, then $f = f(V)$ for some vector matrix V.

6.2 A hard vector matrix

In this section we construct a hard polynomial by describing its vector matrix M. Let $\mathbf{e}_i(j)$, $i, j \in [\ell]$, be ℓ^2 orthonormal vectors in $\mathbb{R}^{\ell^2}$. The matrix M will be $k \times k$ matrix with $k = 2(\ell - 1)$ whose entries are vectors $\mathbf{e}_i(j)$.

For $t \in [\ell]$, let $L(t)$ be the following $2 \times 2(\ell - 1)$ matrix

$$L(t) = \begin{bmatrix} \mathbf{e}_1(t) & \mathbf{e}_1(t) & \mathbf{e}_2(t) & \mathbf{e}_2(t) & \mathbf{e}_3(t) & \mathbf{e}_3(t) & \cdots & \mathbf{e}_{\ell-2}(t) & \mathbf{e}_{\ell-1}(t) & \mathbf{e}_{\ell-1}(t) \\ \mathbf{e}_1(t) & \mathbf{e}_2(t) & \mathbf{e}_2(t) & \mathbf{e}_3(t) & \mathbf{e}_3(t) & \mathbf{e}_4(t) & \cdots & \mathbf{e}_{\ell-1}(t) & \mathbf{e}_{\ell-1}(t) & \mathbf{e}_\ell(t) \end{bmatrix}.$$

Let M be the matrix

$$M = \begin{bmatrix} L(1) & L(1) \\ L(1) & L(2) \\ L(2) & L(2) \\ \cdots & \cdots \\ L(\ell-2) & L(\ell-1) \\ L(\ell-1) & L(\ell-1) \\ L(\ell-1) & L(\ell) \end{bmatrix}.$$

The following theorem shows that $f(M)$ is hard, in the sense of sum-of-squares complexity.

Theorem 6.3. $\mathcal{S}_{\mathbb{R}}(f(M)) = \ell^2$.

The proof is based on the following lemma.

Lemma 6.4. *If U and M are similar, then U and M are equivalent.*

Let us first show that Lemma 6.4 implies Theorem 6.3. The matrix M consists of ℓ^2 orthonormal vectors. Hence any matrix U equivalent to M has height at least ℓ^2. The theorem follows from Lemmas 6.4 and 6.2.

We now proceed to prove Lemma 6.4. Let us state two more definitions. A vector matrix V is called *normal*, if for every $\mathbf{v}$ in V, we have $\mathbf{v}\cdot\mathbf{v} = 1$. Two vector matrices V_1, V_2 of the same height are called *orthogonal*, if for every $\mathbf{v}_1$ in V_1 and $\mathbf{v}_2$ in V_2, we have $\mathbf{v}_1\cdot\mathbf{v}_2 = 0$.

The proof of Lemma 6.4 starts with the following three simple claims. In the claims, we denote elements of V, U by $\mathbf{v}, \mathbf{u}$, and elements of V_p, U_p by $\mathbf{v}^p, \mathbf{u}^p$, where p is an integer. In a crucial way, we employ the following property of real vectors: if $\mathbf{v}\cdot\mathbf{v} = \mathbf{u}\cdot\mathbf{u} = 1$ and $\mathbf{v}\cdot\mathbf{u} = 1$, then $\mathbf{v} = \mathbf{u}$.

Claim 6.5. *If U and V are similar and V is normal, then U is normal.*

Proof. This follows from condition (6.1). QED

Claim 6.6. *Let*

$$V = \begin{bmatrix} V_1 \\ V_1 \end{bmatrix} \quad \textit{and} \quad U = \begin{bmatrix} U_1 \\ U_2 \end{bmatrix}.$$

If V is normal and U and V are similar, then $U_1 = U_2$. The same holds for $V = [V_1\, V_1]$ and $U = [U_1 U_2]$.

Proof. Let V_1 be $r \times c$ vector matrix. For every $i \in [r]$ and $j \in [c]$, $\mathbf{v}_{i+r,j} = \mathbf{v}_{i,j}$ and so $\mathbf{v}_{i+r,j} \cdot \mathbf{v}_{i,j} = 1$. Since V is normal and U and V are similar, by (6.1), $\mathbf{u}_{i,j} \cdot \mathbf{u}_{i,j} = \mathbf{u}_{i+r,j} \cdot \mathbf{u}_{i+r,j} = 1$. By (6.3), $\mathbf{u}_{i+r,j} \cdot \mathbf{u}_{i,j} = 1$, and so $\mathbf{u}_{i+r,j} = \mathbf{u}_{i,j}$. QED

Claim 6.7. *Let*

$$V = \begin{bmatrix} V_1 & V_2 \\ V_3 & V_4 \end{bmatrix} \quad \textit{and} \quad U = \begin{bmatrix} U_1 & U_2 \\ U_3 & U_4 \end{bmatrix}$$

with V a normal matrix. Assume that V_1 and V_4 are orthogonal and either

(i). V_2, V_3 are orthogonal and U_2, U_3 are orthogonal, or

(ii). $V_1 = V_2 = V_3$, $U_1 = U_2 = U_3$ and U_1, V_1 are equivalent.

Then U_1 and U_4 are orthogonal.

Proof. Let V_1 and V_4 be of sizes $r_1 \times c_1$ and $r_2 \times c_2$. From condition (6.4), we have that for every $i_1 \in [r_1], j_1 \in [c_1], i_1 \in [r_2]$ and $j_2 \in [c_2]$,

$$\mathbf{v}_{i_1,j_1} \cdot \mathbf{v}_{r_1+i_2,c_1+j_2} + \mathbf{v}_{i_1,c_1+j_2} \cdot \mathbf{v}_{r_1+i_2,j_1} = \mathbf{u}_{i_1,j_1} \cdot \mathbf{u}_{r_1+i_2,c_1+j_2} + \mathbf{u}_{i_1,c_1+j_2} \cdot \mathbf{u}_{r_1+i_2,j_1} ,$$

which gives

$$\mathbf{v}^1_{i_1,j_1} \cdot \mathbf{v}^4_{i_2,j_2} + \mathbf{v}^2_{i_1,j_2} \cdot \mathbf{v}^3_{i_2,j_1} = \mathbf{u}^1_{i_1,j_1} \cdot \mathbf{u}^4_{i_2,j_2} + \mathbf{u}^2_{i_1,j_2} \cdot \mathbf{u}^3_{i_2,j_1} .$$

The property that is common to both cases in the assumption of the claim is that $\mathbf{v}^2_{i_1,j_2} \cdot \mathbf{v}^3_{i_2,j_1} = \mathbf{u}^2_{i_1,j_2} \cdot \mathbf{u}^3_{i_2,j_1}$. Therefore, $\mathbf{u}^1_{i_1,j_1} \cdot \mathbf{u}^4_{i_2,j_2} = \mathbf{v}^1_{i_1,j_1} \cdot \mathbf{v}^4_{i_2,j_2} = 0$. QED

The three claims imply the following, using the special structure of the matrix L.

Claim 6.8. *If a vector matrix U is similar to $L(t)$, then U is equivalent to $L(t)$.*

Proof. Since $L(t)$ is normal, so is U by Claim 6.5. By Claim 6.6, U is of the form

$$U = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_2 & \mathbf{u}_3 & \mathbf{u}_3 & \cdots & \mathbf{u}_{\ell-2} & \mathbf{u}_{\ell-1} & \mathbf{u}_{\ell-1} \\ \mathbf{u}_1 & \mathbf{u}_2 & \mathbf{u}_2 & \mathbf{u}_3 & \mathbf{u}_3 & \mathbf{u}_4 & \cdots & \mathbf{u}_{\ell-1} & \mathbf{u}_{\ell-1} & \mathbf{u}_\ell \end{bmatrix} .$$

It is thus sufficient to prove that for every $i < j$ in $[\ell]$, the two vectors $\mathbf{u}_i$ and $\mathbf{u}_j$ are orthogonal. This follows by induction on j. If $j = i + 1$, apply case (ii) of Claim 6.7 to the matrices

$$\begin{bmatrix} \mathbf{e}_i(t) & \mathbf{e}_i(t) \\ \mathbf{e}_i(t) & \mathbf{e}_j(t) \end{bmatrix} \text{ and } \begin{bmatrix} \mathbf{u}_i & \mathbf{u}_i \\ \mathbf{u}_i & \mathbf{u}_j \end{bmatrix} .$$

Otherwise, $j \geq i + 2$ and U contains a submatrix

$$\begin{bmatrix} \mathbf{u}_i & \mathbf{u}_{j-1} \\ \mathbf{u}_i & \mathbf{u}_j \end{bmatrix} ,$$

which is similar to the matrix

$$\begin{bmatrix} \mathbf{e}_i(t) & \mathbf{e}_{j-1}(t) \\ \mathbf{e}_i(t) & \mathbf{e}_j(t) \end{bmatrix}.$$

By assumption $\mathbf{u}_i$ and $\mathbf{u}_{j-1}$ are orthogonal. Now case (i) of Claim 6.7 implies that $\mathbf{u}_i$ and $\mathbf{u}_j$ are orthogonal. QED

Proof of Lemma 6.4. If U is similar to M, by Claim 6.8, U has the form

$$U = \begin{bmatrix} U(1) & U(1) \\ U(1) & U(2) \\ U(2) & U(2) \\ \cdots & \cdots \\ U(\ell-2) & U(\ell-1) \\ U(\ell-1) & U(\ell-1) \\ U(\ell-1) & U(\ell) \end{bmatrix},$$

where $U(t)$ is equivalent to $L(t)$. It is now sufficient to prove that $U(i)$ and $U(j)$ are orthogonal whenever $i < j$. This follows by a similar argument as the one in Claim 6.8. QED

7 Integer sums-of-squares

In this section, we prove Theorem 1.10. More exactly, we prove that in any identity of the form

$$(x_1^2 + \cdots + x_k^2) \cdot (y_1^2 + \cdots + y_k^2) = z_1^2 + \cdots + z_n^2\,, \tag{7.1}$$

where z_i are bilinear forms with integer coefficients, n must be at least $\Omega(k^{6/5})$.

7.1 Sum-of-squares and intercalate matrices

Following Yiu [34], we phrase $\mathcal{S}_{\mathbb{Z}}(k)$ in a more combinatorial language (though we deviate from Yiu's notation). We call a $k \times k$ matrix $M = (M_{i,j})_{i,j\in[k]}$ with non-zero integer entries an *intercalate matrix*, if

1) $|M_{i,j_1}| \neq |M_{i,j_2}|$, whenever $j_1 \neq j_2$,

2) $|M_{i_1,j}| \neq |M_{i_2,j}|$, whenever $i_1 \neq i_2$,

3) if $i_1 \neq i_2$, $j_1 \neq j_2$ and $M_{i_1,j_1} = \pm M_{i_2,j_2}$, then $M_{i_1,j_2} = \mp M_{i_2,j_1}$.

We call $C = C(M) = \{|M_{ij}| \;:\; i,j \in [k]\}$ the *set of colors* in M. We say that M *has n colors*, if $|C| = n$.

Condition 1) says that no color appears twice in the same row of M, condition 2) says that no color appears twice in the same column of M. Condition 3) then requires that for every 2×2 submatrix

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

of M, either $|a|, |b|, |c|, |d|$ are all different, or the submatrix is of the form

$$\begin{pmatrix} \epsilon_1 a & \epsilon_2 b \\ \epsilon_3 b & \epsilon_4 a \end{pmatrix},$$

where $|a| \neq |b|$ and $\epsilon_i \in \{+1, -1\}$ satisfy $\epsilon_1\epsilon_2\epsilon_3\epsilon_4 = -1$. The following are examples of 2×2 intercalate matrices:

$$\begin{pmatrix} 1 & 2 \\ 3 & -4 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 2 & -1 \end{pmatrix}, \text{ and } \begin{pmatrix} -1 & -2 \\ 2 & -1 \end{pmatrix}.$$

The following matrices are *not* intercalate:

$$\begin{pmatrix} 1 & 2 \\ 3 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}, \text{ and } \begin{pmatrix} -1 & 2 \\ 2 & -1 \end{pmatrix}.$$

The following proposition relates intercalate matrices and integer sum-of-squares formulas.

Proposition 7.1. *The following are equivalent:*

(i). There exists an identity (0.1) *where $z_1, \ldots z_n$ are bilinear forms with integer coefficients.*

(ii). There exists an intercalate $k \times k$ matrix with n colors.

Proof. The proof is analogous to the proof of Lemma 6.2. Note that if V is a $k \times k$ vector matrix consisting of k^2 orthonormal real vectors then, by definition of $f(V)$, $(x_1^2 + \cdots + x_k^2)(y_1^2 + \cdots + y_k^2) = f(V)$. As in Lemma 6.2, we can show that (i) is equivalent to the following: there exists a $k \times k$ vector matrix U consisting of vectors $\mathbf{u}_{i,j} \in \mathbb{Z}^n$, $i, j \in [k]$, with the properties ($\mathbf{u}\cdot\mathbf{v}$ denotes the usual inner product in $\mathbb{R}^n$)

i) $\mathbf{u}_{i,j}\cdot\mathbf{u}_{i,j} = 1$, for every i, j,

ii) $\mathbf{u}_{i,j_1}\cdot\mathbf{u}_{i,j_2} = 0$, whenever $j_1 \neq j_2$,

iii) $\mathbf{u}_{i_1,j}\cdot\mathbf{u}_{i_1,j} = 0$, whenever $i_1 \neq i_2$,

iv) $\mathbf{u}_{i_1,j_1}\cdot\mathbf{u}_{i_2,j_2} + \mathbf{u}_{i_1,j_2}\cdot\mathbf{u}_{i_2,j_1} = 0$, for every $i_1 \neq i_2, j_1 \neq j_2$.

However, since we are dealing with vectors in $\mathbb{Z}^n$, condition i) implies a stronger property

v) $\mathbf{u}_{ij} \in \{0, 1, -1\}^n$ and $\mathbf{u}_{ij}$ has exactly one non-zero entry.

Here is how a matrix U with properties i) through v) corresponds to an intercalate matrix. Given an intercalate matrix M with colors $\{a_1, \ldots, a_n\}$, define V as follows: for every $\ell \in [n]$ and $i, j \in [k]$, define $u_{i,j}[\ell] = \mathrm{sgn}(M_{i,j})$, if $M_{i,j} = a_\ell$, and $u_{i,j}[\ell] = 0$ otherwise (where $u_{i,j})[\ell]$ denotes the ℓth coordinate of $\mathbf{u}_{i,j}$). Conversely, given such a matrix U, define an intercalate matrix with colors $\{1, \ldots, n\}$ as $M_{i,j} = u_{i,j}[\ell] \cdot \ell$, where ℓ is the unique coordinate such that $u_{i,j}[\ell] \neq 0$. It is straightforward to verify that the required properties of V resp. M are satisfied. QED

7.2 The number of colors in intercalate matrices

We say that two integer matrices M and M' *are equivalent*, if M' can be obtained from M by

1) permuting rows and columns,
2) multiplying rows and columns by minus one, and
3) renaming colors, that is, if $\theta : \mathbb{N} \to \mathbb{N}$ is a one-to-one map, we have $M'_{i,j} = \mathrm{sgn}(M_{i,j}) \cdot \theta(|M_{i,j}|)$, for every $i, j \in [k]$.

Here are two elementary properties of intercalate matrices.

Fact 7.2. *A submatrix of an intercalate matrix is an intercalate matrix.*

Fact 7.3. *If M and M' are equivalent, then M is intercalate if and only if M' is intercalate.*

We say that a $k \times k$ matrix M is *full*, if for every $i \in [k]$, we have $M_{i,i} = 1$.

The following lemma is the main step in the proof of our main theorem.

Lemma 7.4. *Let M be a $k \times k$ full intercalate matrix. Then M has at least $\Omega(k^{3/2})$ colors.*

Lemma 7.4 implies the following theorem, which gives Theorem 1.10 by Proposition 7.1 .

Theorem 7.5. *Any $k \times k$ intercalate matrix has at least $\Omega(k^{6/5})$ colours.*

Proof. Let M be a $k \times k$ intercalate matrix with n colors. We show that M contains a $s \times s$ submatrix $M^{(0)}$ which is equivalent to a full intercalate matrix, with $s \geq k^2/n$. For a color a, let $M_a = \{(i,j) \in [k] \times [k] \; : \; |M_{i,j}| = a\}$. The sets M_a form a partition of $[k] \times [k]$ to n pairwise disjoint sets, and hence there exists some a so that $s := |M_a| \geq k^2/n$. Let $M^{(0)}$ be the submatrix of M obtained by deleting rows and columns that do not contain a. Since the color a never occurs twice in the same row or column in $M^{(0)}$, $M^{(0)}$ is $s \times s$ matrix, and we can permute rows and columns of $M^{(0)}$ to obtain a matrix $M^{(1)}$ in which the diagonal entries satisfy $|M^{(1)}_{i,i}| = a$. We can thus multiply some of the rows of $M^{(1)}$ by minus one to obtain a matrix $M^{(2)}$ in which the diagonal entries have $M^{(2)}_{i,i} = a$. Finally, we can rename the colors of $M^{(2)}$ to obtain a matrix $M^{(3)}$ with $M^{(3)}_{i,i} = 1$ for every $i \in [k]$. Altogether, $M^{(3)}$ is a full intercalate matrix equivalent to $M^{(0)}$.

$M^{(0)}$ contains at most n colors. Hence Lemma 7.4 tells us that $n \geq \Omega(s^{3/2})$. Since $s \geq k^2/n$, we have $n \geq \Omega(k^3/n^{3/2})$, which implies $n \geq \Omega(k^{6/5})$. QED

7.3 Number of colors in full intercalate matrices

The definition of intercalatness immediately implies the following:

Fact 7.6. *If M is a full intercalate matrix, then $M_{i,j} = -M_{j,i}$ for every $i \neq j$.*

We now describe a few combinatorial properties of full intercalate matrices.

Lemma 7.7. *Assume that M is 6×6 intercalate matrix of the form*[10]

$$\begin{pmatrix} 1 & 2 & 3 & & & \\ & 1 & 4 & & & \\ & & 1 & & & \\ & & & 1 & 2 & 3 \\ & & & & 1 & b \\ & & & & & 1 \end{pmatrix}.$$

Then $b = -4$.

Proof. Let $M_{1,4} = c$. By Fact 7.6, M has the form

$$\begin{pmatrix} 1 & 2 & 3 & c & & \\ -2 & 1 & 4 & & & \\ -3 & -4 & 1 & & & \\ -c & & & 1 & 2 & 3 \\ & & & & 1 & b \\ & & & & & 1 \end{pmatrix}.$$

Property 3) in the definition of intercalate matrices implies that $M_{2,5} = M_{3,6} = M_{4,1} = -c$, as $M_{2,1} = -M_{4,5}$ and $M_{3,1} = -M_{4,6}$. Using Fact 7.6, we thus conclude that M has the form

$$\begin{pmatrix} 1 & 2 & 3 & c & & \\ -2 & 1 & 4 & & -c & \\ -3 & -4 & 1 & & & -c \\ -c & & & 1 & 2 & 3 \\ & c & & & 1 & b \\ & & c & & & 1 \end{pmatrix}.$$

Here we have $M_{5,2} = -M_{3,6}$ and hence $M_{5,6} = M_{3,2}$. In other words, $b = -4$. QED

Let M be a $k \times k$ matrix. A triple (i, j_1, j_2) such that $1 \leq i < j_1 < j_2 \leq k$ is called a *position* in M. Let (a, b) be an ordered pair of natural numbers. We say that (a, b) *occurs* in position (i, j_1, j_2) in M, if $|M_{i,j_1}| = a$ and $|M_{i,j_2}| = b$.

[10] The empty entries are some unspecified integers.

Proposition 7.8. *Let M be a full intercalate matrix. Then every pair (a, b) occurs in at most two different positions in M.*

Proof. Assume that (a, b) occurs at three distinct positions $(i(p), j_1(p), j_2(p))$, $p \in \{0, 1, 2\}$, in M. By renaming colors, we can assume without loss of generality that $(a, b) = (2, 3)$. We show that M contains 9×9 submatrix M' equivalent to a matrix of the form

$$\begin{pmatrix} A_1 & & \\ & A_2 & \\ & & A_3 \end{pmatrix},$$

where

$$A_i = \begin{pmatrix} 1 & 2 & 3 \\ & 1 & c_i \\ & & 1 \end{pmatrix}.$$

This will imply a contradiction: Lemma 7.7 implies that $c_2 = -c_1$, $c_3 = -c_1$ and $c_3 = -c_2$, and hence $c_1 = -c_1$, which is impossible, as $c_1 \neq 0$.

We first show that the nine indices $I = \{i(p), j_1(p), j_2(p) \ : \ p \in \{0, 1, 2\}\}$ are all distinct. There are a few cases to consider.

(i). The definition of position guarantees that

$$|\{i(p), j_1(p), j_2(p)\}| = 3$$

for every $p \in \{0, 1, 2\}$.

(ii). Since no color can appear twice in the same row,

$$|\{i(0), i(1), i(2)\}| = |\{j_1(0), j_1(1), j_1(2)\}| = |\{j_2(0), j_2(1), j_2(2)\}| = 3.$$

(iii). Since $|M_{i(p),j_1(p)}| = |M_{i(q),j_1(q)}| = 2$, M being intercalate implies

$$|M_{i(p),j_1(q)}| = |M_{i(q),j_1(p)}|.$$

Assume, for the sake of contradiction, that $j_2(p) = j_1(q)$ for some $p \neq q$. Thus, $|M_{i(p),j_1(q)}| = |M_{i(p),j_2(p)}| = 3$, and so $|M_{i(q),j_1(p)}| = 3$. But $j_1(p) \neq j_2(q)$, as $j_1(p) < j_2(p) = j_1(q) < j_2(q)$. This contradicts property (1 in the definition of intercalate matrices, since $|M_{i(q),j_1(p)}| = |M_{i(q),j_2(q)}|$.

(iv). Assume, for the sake of contradiction, that $i(q) = j_e(p)$ for some $p \neq q$ and $e = 1, 2$. Since M is full, $M_{i(q),j_e(p)} = 1$. As above, we conclude that $|M_{i(p),j_e(q)}| = 1$. But $i(p) \neq j_e(q)$, since $i(p) < j_e(p) = i(q) < j_e(q)$. Thus the color 1 appear twice in the row $i(p)$, which is a contradiction.

Let M' be the 9×9 submatrix of M defined by the set of rows and columns I. Permuting rows and columns of M', we obtain a matrix of the form

$$\begin{pmatrix} B_1 & & \\ & B_2 & \\ & & B_3 \end{pmatrix},$$

where

$$B_i = \begin{pmatrix} 1 & \epsilon_i 2 & \delta_i 3 \\ & 1 & \\ & & 1 \end{pmatrix}$$

and $\epsilon_i, \delta_i \in \{1, -1\}$. Multiplying rows and columns by minus one where appropriate, we conclude that M' is of the desired form. QED

We are now ready for the proof of the lemma.

Proof of Lemma 7.4. There are at least $k^3/8$ different positions in M. From n colors, one can build at most n^2 ordered pairs. Proposition 7.8 implies that any such pair appears in at most two positions in M. Thus, $2n^2 \geq k^3/8$ and so $n \geq \Omega(k^{3/2})$. QED

8 Multilinear and ordered circuits

8.1 Ordered circuits

An interesting property of non-commutative polynomials and circuits is that we can treat occurrences of x at different positions as distinct variables. For example, we could have defined the identity polynomial as a polynomial in $4k$ variables

$$\mathsf{ID}_k{}' = \sum_{i,j \in [k]} x_{1,i} x_{2,j} x_{3,i} x_{4,j} \ ,$$

or as the polynomial in only k variables

$$\mathsf{ID}_k{}'' = \sum_{i,j \in [k]} x_i x_j x_i x_j \ .$$

These modification are not important in the non-commutative setting; the circuit complexity of ID_k, $\mathsf{ID}_k{}'$ and $\mathsf{ID}_k{}''$ differ by at most a constant factor. We discuss this phenomenon in this section.

A homogeneous polynomial f of degree r is called *ordered*, if there exist *disjoint* sets of variables $X_1, \dots, X_r$ so that $f = f[X_1, \dots, X_r]$ with the definition of $f[X_1, \dots, X_r]$ from Section 3.2. In other words, f is ordered if every variable that occurs at position i in some monomial in f is in X_i.

An *interval* I is a set of the form $I = [j_1, j_2] = \{i : \ j_1 \leq i \leq j_2\}$. A polynomial g *is of type* $[j_1, j_2]$, if $g = g[X_{j_1}, \dots, X_{j_2}]$. It is a homogeneous polynomial of degree $j_2 - j_1 + 1$. A constant polynomial is of type $I = \emptyset$.

We now define ordered circuits, which are a natural model for computing ordered polynomials. In an ordered circuit Φ, every gate v is associated with an interval $I_v = I_v(\Phi) \subseteq [r]$. A circuit Φ is called *ordered*, if it satisfies the following properties:

(i). Every gate v in Φ computes a polynomial of type I_v.

(ii). If $v = v_1 + v_2$, then $I_v = I_{v_1} = I_{v_2}$.

(iii). If $v = v_1 \times v_2$ with $I_v = [i, j]$, then there exists $i - 1 \leq \ell \leq j$ so that $I_{v_1} = [i, \ell]$ and $I_{v_2} = [\ell + 1, j]$.

We can also define an ordered version for a general polynomial. Let f be a homogeneous polynomial of degree r in the variables $X = \{x_1, \dots, x_k\}$. We define the *ordered version* of f, denoted $f^{(ord)}$, as follows. For every $j \in [r]$ and $i \in [k]$, introduce a new variable $x_{j,i}$, and let $X_j = \{x_{j,1}, \dots, x_{j,k}\}$. For a monomial $\alpha = x_{i_1} \dots x_{i_r}$, let $\alpha^{(ord)} := x_{1,i_1} \dots x_{r,i_r}$. The polynomial $f^{(ord)}$ is the ordered polynomial in the variables $X_1, \dots, X_r$ defined by

$$\mathsf{COEF}_{\alpha^{(ord)}}(f^{(ord)}) = \mathsf{COEF}_{\alpha}(f)\,.$$

Given $f^{(ord)}$ we can easily recover f by substituting $x_{j,i} = x_i$ for every $j \in [r]$ and $i \in [k]$. When f is already ordered, then $f^{(ord)}$ and f are the same polynomials, up to renaming of variables.

The following theorem shows that non-commutative circuits computing f can be efficiently simulated by ordered circuits computing $f^{(ord)}$. In particular, if f is already ordered, then a general circuit computing f can be efficiently simulated by an ordered circuit. Every ordered circuit is syntactically multilinear, as defined in Section 8.2 below. This implies that non-commutative circuits for ordered polynomials can be efficiently simulated by syntactically multilinear circuits. We do know such a result in the commutative world: the best known transformation of a commutative circuit to a syntactically multilinear circuit increases the size by a factor of 2^r (instead of r^3 here).

The theorem is a stronger version of Theorem 1.11 which was stated in the introduction.

Theorem 8.1. *Let Φ be a circuit of size s computing a homogeneous polynomial f of degree r. Then there is an ordered circuit Ψ of size $O(r^3 s)$ that computes $f^{(ord)}$.*

Proof. Before we prove the theorem we introduce some notation. If g is a polynomial (not necessarily homogeneous) and $I = [j_1, j_2] \subseteq [r]$ is a nonempty interval, define $g^{(I)}$ as the polynomial of type I defined by

$$\mathsf{COEF}_{\alpha^{(I)}}(g^{(I)}) = \mathsf{COEF}_{\alpha}(g),$$

where $\alpha^{(I)} = \prod_{j=j_1}^{j_2} x_{j,i_j}$ and $\alpha = \prod_{j=j_1}^{j_2} x_{i_j}$, and if $I = \emptyset$, $g^{(I)}$ is the constant term in g. We thus have that $f^{(ord)} = f^{(I)}$ with $I = [1, r]$.

We prove the theorem by describing how to construct Ψ. We duplicate each gate v in Φ into $O(r^2)$ gates in Ψ, which we denote (v, I) with $I \subseteq [r]$ an interval. Every (v, I) will compute the polynomial $\widehat{\Phi}_v^{(I)}$. If v is an input gate labelled by a field element, set $(v, \emptyset) = \widehat{\Phi}_v$ and $(v, I) = 0$ for every nonempty I. If v is an input gate labelled by a variable x_i, set $(v, [j, j]) = x_{j,i}$ and $(v, I) = 0$ when I is not a singleton. If $v = v_1 + v_2$, set $(v, I) = (v_1, I) + (v_2, I)$ for all I. If $v = v_1 \times v_2$ and $I = [i, j]$, set

$$(v, I) = \sum_{i-1 \leq \ell \leq j} (v_1, [i, \ell]) \times (v_2, [\ell + 1, j]).$$

Associate with the gate (v, I) in Ψ the interval I. Thus, Ψ is ordered. By induction, every gate (v, I) computes $\widehat{\Phi}_v^{(I)}$, and hence Ψ computes $f^{(ord)}$. For every gate v in Φ, there are at most $O(r^3)$ edges in Ψ, and so the size of Ψ is as claimed. $\square$

8.2 Multilinear circuits

In this section we prove an exponential lower bound for the size of non-commutative syntactically multilinear circuits (a circuit Φ is *syntactically multilinear*, if for every product gate $v = v_1 \times v_2$ in Φ, the two circuits Φ_{v_1} and Φ_{v_2} do not share variables). Note that an ordered circuit is automatically syntactically multilinear. By means of Proposition 8.1, a lower bound on syntactically multilinear circuits computing an ordered polynomial would imply an unconditional lower bound. However, our lower bound involves a polynomial which is *not* ordered.

We now define the multilinear version of central polynomials. Let f be a multilinear polynomial of degree d. We say that f is *ml-central*, if f is central as in (3.1), and for every $i \in [m]$, the polynomial $h_i g \overline{h}_i$ is multilinear; in particular, the polynomials $h_i, g, \overline{h}_i$ have distinct variables.

The following lemma describes the structure of multilinear circuits.

Lemma 8.2. *Let f be a homogeneous multilinear polynomial of degree $d \geq 2$. Assume that there is a syntactically multilinear circuit Φ of size s computing f. Then there exist $n \leq O(d^3 s)$ and ml-central polynomials $f_1, \ldots, f_n$ such that $f = f_1 + \cdots + f_n$.*

Proof. The proof is almost identical to Proposition 3.2. QED

Our lower bound is based on counting monomials. The following lemma is the basic observation for the lower bound.

Lemma 8.3. *Let f be a ml-central polynomial of degree k in k variables. Then f has at most $2^{-\Omega(k)}k!$ monomials with nonzero coefficients.*

Proof. Write f as $f = \sum_{i \in [m]} h_i g \overline{h}_i$ with every $h_i g \overline{h}_i$ multilinear. Let X be the set of variables in f and X_0 the set of variables in g. Every monomial with a nonzero coefficient in f has the form $\alpha_1 \gamma \alpha_2$, where (1) γ is a multilinear monomial of degree d_0 in variables X_0, and (2) α_1, α_2 are multilinear monomials in the variables $X \setminus X_0$ of degrees d_1, d_2, and α_1,α_2 have distinct variables. Since $d_0 + d_1 + d_2 = k$, we have $|X_0| = d_0$. There are thus $d_0!$ βs in (1), and at most $(d_1 + d_2)!$ pairs α_1, α_2 in (2). Hence f contains at most

$$d_0!(d_1 + d_2)! = d_0!(k - d_0)! = \frac{k!}{\binom{k}{d_0}}$$

monomials with non-zero coefficients. Since $k/3 \leq d_0 < 2k/3$, this is at most $2^{-\Omega(k)}k!$. QED

Define the all-permutations polynomial, AP_k, as a polynomial in variables $x_1, \ldots, x_k$

$$\mathsf{AP}_k = \sum_{\sigma} x_{\sigma(1)} x_{\sigma(2)} \cdots x_{\sigma(k)} \, ,$$

where σ is a permutation of $[k]$. Note that $\mathsf{AP}_k^{(ord)}$ is a polynomial in k^2 variables,

$$\mathsf{AP}_k^{(ord)} = \sum_{\sigma} x_{1,\sigma(1)} x_{2,\sigma(2)} \cdots x_{k,\sigma(k)}.$$

In other words, $\mathsf{AP}_k^{(ord)} = \mathrm{PERM}_k$.

Proof of Theorem 1.12. Assume that AP_k is computed by such a circuit of size s. By Lemma 8.2, AP_k can be written as a sum of $O(k^3 s)$ ml-centralpolynomials. By Lemma 8.3, AP_k can thus have at most $O(2^{-\Omega(k)} k! k^3 s)$ monomials with nonzero coefficients. However, AP_k has $k!$ monomials. QED

References

[1] A. Barvinok. A simple polynomial time algorithm to approximate the permanent within a simply exponential factor. Random Structures and Algorithms 14(1), pages 29–61, 1999.

[2] W. Baur and V. Strassen. The complexity of partial derivatives. Theoretical computer science (22), pages 317–330, 1983.

[3] P. Burgisser. Completeness and reduction in algebraic complexity theory. Springer-Verlag Berlin Heidelberg 2000.

[4] S. Chien and A. Sinclair. Algebras with polynomial identities and computing the determinant. SIAM Journal on Computing 37, pages 252–266, 2007.

[5] S. Chien, L. Rasmussen and A. Sinclair. STOC 02', pages 222-231, 2002.

[6] J. von zur Gathen. Algebraic complexity theory. Ann. Rev. Comp. Sci. (3), pages 317–347, 1988.

[7] C. Godsil and I. Gutman. On the matching polynomial of a graph. Algebraic Methods in Graph Theory, pages 241–249, 1981.

[8] L. Hyafil. On the parallel evaluation of multivariate polynomials. SIAM J. Comput. 8(2), pages 120–123, 1979.

[9] P. Hrubes and A. Yehudayoff. Homogeneous formulas and symmetric polynomials. arXiv:0907.2621

[10] P. Hrubes, A. Wigderson and A. Yehudayoff. Completeness and separations with less relations. In preparation.

[11] A. Hurwitz. Über die Komposition der quadratischen Formen von beliebigvielen Variabeln. *Nach. Ges. der Wiss. Göttingen*, pages 309–316, 1898.

[12] A. Hurwitz. Über die Komposition der quadratischen Formen. *Math. Ann.*, 88, pages 1–25, 1923.

[13] I. M. James. On the immersion problem for real projective spaces. *Bull. Am. Math. Soc.*, 69, pages 231–238, 1967.

[14] M. Jerrum, A. Sinclair and E. Vigoda. A polynomial-time approximation algorithm for the permanent of a matrix with nonnegative entries. J. ACM 51(4), pages 671–697, 2004.

[15] S. Jukna Boolean function complexity: advances and frontiers. Book in preparation.

[16] N. Karmarkar, R. Karp, R. Lipton, L. Lovasz and M. Luby. A Monte-Carlo algorithm for estimating the permanent. SIAM Journal on Computing 22(2), pages 284–293, 1993.

[17] T. Kirkman. On pluquatemions, and horaoid products of sums of n squares. *Philos. Mag. (ser. 3)*, 33, pages 447–459; 494-509, 1848.

[18] K. Y. Lam. Some new results on composition of quadratic forms. *Inventiones Mathematicae.*, 1985.

[19] T. Y. Lam and T. Smith. On Yuzvinsky's monomial pairings. *Quart. J. Math. Oxford. (2)*, 44, pages 215–237, 1993.

[20] K. Mulmuley. On P vs. NP, Geometric Complexity Theory, and the Riemann Hypothesis. Technical Report, Computer Science department, The University of Chicago, 2009.

[21] N. Nisan. Lower bounds for non-commutative computation. STOC 91', pages 410–418, 1991.

[22] N. Nisan and A. Wigderson. Lower bounds on arithmetic circuits via partial derivatives. Computational Complexity, vol. 6, pages 217–234, 1996.

[23] A. Pfister. Zur Darstellung definiter Funktionen als Summe von Quadraten. *Inventiones Mathematicae.*, 1967.

[24] J. Radon. Lineare scharen orthogonalen matrizen. *Abh. Math. Sem. Univ. Hamburg 1*, pages 2–14, 1922.

[25] R. Raz. Multi-linear formulas for permanent and determinant are of super-polynomial size. Journal of the Association for Computing Machinery 56 (2), 2009.

[26] R. Raz. Elusive functions and lower bounds for arithmetic circuits. To appear in Theory of Computing.

[27] R. Raz and A. Yehudayoff. Lower bounds and separation for constant depth multilinear circuits. Proceedings of Computational Complexity, pages 128-139, 2008.

[28] R. Raz, A. Shpilka and A. Yehudayoff. A lower bound for the size of syntactically multilinear arithmetic circuits. SIAM Journal on Computing 38 (4), pages 1624-1647, 2008.

[29] D. B. Shapiro. Composition of quadratic forms. W. de Gruyter Verlag, 2000.

[30] Die berechnungskomplexitat von elementarsymmetrischen funktionen und von interpolationskoeffizienten. Numerische Mathematik (20), pages 238–251, 1973.

[31] V. Strassen. Vermeidung von Divisionen. J. Reine Angew. Math. 264, pages 182–202, 1973.

[32] L. G. Valiant. Completeness classes in algebra. STOC '79, pages 249–261.

[33] S. Winograd. On the number of multiplications needed to compute certain functions. Comm. on Pure and Appl. Math. (23), pages 165–179, 1970.

[34] P. Yiu. Sums of squares formulae with integer coefficients. *Canad. Math. Bull.* , 30, pages 318-324, 1987.

[35] P. Yiu. On the product of two sums of 16 squares as a sum of squares of integral bilinear forms. *Quart. J. Math. Oxford. (2)* , 41, pages 463–500, 1990.

[36] S. Yuzvinsky. A series of monomial pairings. *Linear and multilinear algebra*, 15, pages 19–119, 1984.

ECCC ISSN 1433-8092

http://eccc.hpi-web.de