THEORY OF COMPUTING, Volume 11 (946), 2015, pp. 357–393
www.theoryofcomputing.org

Non-Commutative Arithmetic Circuits with Division

Pavel Hrubeš[*] Avi Wigderson[†]

Received March 26, 2015; Revised November 8, 2015; Published December 20, 2015

Abstract: We initiate the study of the complexity of arithmetic circuits with division gates over non-commuting variables. Such circuits and formulae compute *non-commutative* rational functions, which, despite their name, can no longer be expressed as ratios of polynomials. We prove some lower and upper bounds, completeness and simulation results, as follows.

If X is an $n \times n$ matrix consisting of n^2 distinct mutually non-commuting variables, we show that:

(i). X^{-1} can be computed by a circuit of polynomial size.

(ii). Every formula computing some entry of X^{-1} must have size at least $2^{\Omega(n)}$.

We also show that matrix inverse is complete in the following sense:

(i). Assume that a non-commutative rational function f can be computed by a formula of size s. Then there exists an invertible $2s \times 2s$-matrix A whose entries are variables or field elements such that f is an entry of A^{-1}.

A conference version of this paper appeared in ITCS 2014 [25].

[*] The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Program (FP7/2007-2013) / ERC grant agreement no. 339691. The Institute of Mathematics is supported by RVO:67985840.

[†] This research was partially supported by NSF grant CCF-1412958.

ACM Classification: F.1.1, F.2.m

AMS Classification: 68Q17, 68W30

Key words and phrases: arithmetic circuits, non-commutative rational function, skew field

DOI: 10.4086/toc.2015.v011a946

(*ii*). If f is a non-commutative polynomial computed by a formula without inverse gates then A can be taken as an upper triangular matrix with field elements on the diagonal.

We show how divisions can be eliminated from non-commutative circuits and formulae which compute polynomials, and we address the non-commutative version of the "rational function identity testing" problem. As it happens, the complexity of both of these procedures depends on a single open problem in invariant theory.

1 Introduction

Arithmetic circuit complexity studies the computation of polynomials and rational functions using the basic operations addition, multiplication, and division. It is chiefly interested in *commutative* polynomials or rational functions, defined over a set of multiplicatively commuting variables (see the survey [46], or the book [7]). The dominant computational models are the *arithmetic circuit* and its weakening, the *arithmetic formula*. The main open problem is to present an explicit polynomial which cannot be computed by a circuit—or a formula—of polynomial size.

The complexity of computing polynomials (not allowing division) in *non-commuting* variables has also been considered, for example, in [39, 27]. This was motivated partly by an apparent lack of progress in proving lower bounds in the commutative setting, partly by an interest in computations in matrix algebras.[1] And indeed, we do have nontrivial lower bounds in this setting. Most notably, Nisan [39] has proved thirty years ago that any arithmetic *formula* computing the non-commutative determinant or permanent must have an exponential size, and also gave an exponential separation between the power of circuits and formulae in this model. Despite much effort, a similar lower bound for non-commutative arithmetic *circuits* has not been achieved—indeed, the best known lower bounds for non-commutative circuits are as weak as the commutative ones.

In this paper, we take the study of non-commutative computation one step further and consider the complexity of non-commutative circuits which contain division (equivalently, inverse) gates. Such a circuit computes a "non-commutative rational function"—a far more complicated object than its commutative counterpart. Traditionally, arithmetic circuit complexity focuses on the computation of polynomials, with rational functions receiving minor attention. This is mainly because any commutative rational function can be expressed as a pair of polynomials fg^{-1}. Even on the computational level, commutative rational functions do not substantially differ from polynomials—apart from the omnipresent threat of dividing by zero. In contrast, the structure of non-commutative rational functions is far more complex, giving rise to a host of new phenomena. It is not difficult to see that $x^{-1}+y^{-1}$ or $xy^{-1}x$ can no longer be expressed as fg^{-1} (or $g^{-1}f$), if x,y do not commute. More importantly, non-commutative rational functions may require *nested* inverses, as in $(u+xy^{-1}z)^{-1}$. Indeed, any number of inverse operations, and nested inverse operations, may be needed to represent a rational function. Moreover, there is no "canonical" representation of non-commutative rational functions. Despite these facts, or rather thanks to them, non-commutative rational functions possess quite a lot of structure. They form a skew

[1] Observe that Strassen's (and all subsequent) fast matrix multiplication algorithms *necessarily* work over non-commuting matrix entries.

field[2] which is equipped with invariants not present in the standard commutative field of fractions. Our main hope is that this additional structure may be useful for proving lower bounds, even for polynomials. We make initial steps in this direction.

Non-commutative rational functions arise naturally in a variety of settings, beyond the abstract mathematical fields of non-commutative algebra and geometry.[3] One area is linear system theory and control theory, where the order of actions clearly matters, and the dynamics is often given by a rational function and its iterates. The paper [32] surveys some of this work, and also demonstrates situations where results in the commutative case were proven by "lifting" them to the non-commutative setting. Another area is formal language theory, where regular expressions and formal series play an analogous role. Indeed, these two areas are tightly connected, and the book [4] surveys some of the connections between the algebraic and linguistic settings (and more).

Note that non-commutative rational functions can often be more natural than non-commutative polynomials. For example, the determinant as a non-commutative *polynomial* has no longer any relevance to solving linear equations or a geometrical interpretation. In [22], it was argued that the correct analogy of the commutative determinant is the *quasideterminant*, which is a set of non-commutative rational functions sharing and extending many of the useful and beautiful properties of the commutative determinant. This development has important consequences in a variety of mathematical areas. The inverse of a matrix is probably the best example of a notion that makes perfect sense in the non-commutative setting, as a set of rational functions (and indeed the quasi-determinant of a matrix may be naturally defined from the entries of its inverse). Matrix inverse further plays a key role in one definition, due to Cohn [8], of the universal skew field of non-commutative rational functions.

Matrix inverse turns out to be central from a computational perspective. In this paper we will focus on the complexity of computing the inverse X^{-1} of an $n \times n$ matrix X consisting of n^2 non-commuting variables. We show that X^{-1} can be computed by a polynomial size circuit, but on the other hand, every formula computing an entry of X^{-1} must have an exponential size. This provides a non-trivial example[4] of an exponential gap between circuit and formula size—a counterpart of the above mentioned result of Nisan. We also prove the following completeness result: if a rational function f can be computed by a formula of size s then f can be expressed as an entry of A^{-1}, where A is a $2s \times 2s$-matrix whose entries are variables or field elements. This is an analog of Valiant's [49] theorem on completeness of determinant in the commutative, division-free, setting.

To see the origins of the lower bounds, let us return to examples of rational expressions. We note that the expression $(x+xy^{-1}x)^{-1}$, which has *nested* inversions, can be simplified by Hua's identity to an equivalent expression without nesting: $x^{-1}-(x+y)^{-1}$. On the other hand, in the somewhat similar expression $(u+xy^{-1}z)^{-1}$, the nested inversion cannot be eliminated. This new phenomenon of nested inversion provides a new invariant not present in the commutative setting—the *height* of a rational function. The height is the minimum number of nested inversions in a formula computing this rational function. For a long time, it was not even clear that the height is unbounded, and it was a major result of C. Reutenauer [43] that it in fact is. Indeed, his result is much more precise and informative: any entry of

[2] A. k. a. "associative division algebra"—a field in which multiplication is not necessarily commutative.

[3] One of the best examples is the fact that the fundamental theorem of projective geometry follows rather simply from the fact that the following rational expression, $(x+xy^{-1}x)^{-1}+(x+y)^{-1}-x^{-1}$, is identically zero: this is called Hua's identity.

[4] The trivial example would be x^{2^n}.

the inverse of the generic $n \times n$ matrix X requires n nested inversions, namely has height n.

Our lower bound on formula size of matrix inverse is obtained by showing that a formula of size s can compute a function of height at most logarithmic in s. This is obtained via a general balancing procedure of formulae, which is a bit more involved than the usual one due to the non-commutativity and presence of inversion gates. Combined with Reutenauer's theorem, this implies that the inverse of an $n \times n$ matrix cannot be computed by a formula smaller than $2^{\Omega(n)}$. In circuit complexity, one keeps searching for properties that would imply that a function is hard to compute. For a polynomial f, there are not many such invariants at hand: for example, the degree or the number of variables, which both provide only very limited hardness results, and the more sophisticated rank of the partial derivative matrix used in Nisan's lower bound. In the context of non-commutative rational functions, we can now see that the inverse height is a new non-trivial invariant which can be successfully applied to obtain hardness results. Other non-trivial invariants are known in this setting, and it is quite possible that some of them can shed light on more classical problems of arithmetic circuit complexity.

We also prove a different characterization of the inverse height. We show that in a circuit, one never needs to use more inversion *gates* than the inversion height of the rational function computed, without significantly increasing the circuit size. Thus, e. g., the expression $x_1^{-1} + x_2^{-1} + \cdots + x_n^{-1}$ can be computed using only one inverse gate by an $O(n)$-sized circuit, and $n \times n$ matrix inverse can be computed by a polynomial size circuit with exactly n inverse gates.

We also consider the question of eliminating division from circuits or formulae whose output is a polynomial. Again, in the commutative setting this can be done with little overhead, as shown by Strassen [48]. His idea was to replace an inverse gate with an infinite power series expansion, and eventually truncate it according to the degree of the output polynomial. In the non-commutative setting, this approach faces a significant obstacle. In order to express f^{-1} as a power series, we need a point where f is non-zero, and so Strassen's argument hinges on the fact that a non-zero rational function does not vanish on some substitution from the underlying field (at least when the field is infinite). In contrast, assume that a non-commutative computation inverts the polynomial $xy - yx$. It is not identically zero, but vanishes on all inputs from any base field. A natural idea, which we indeed employ, is to evaluate the circuit on matrices instead of field elements. Extending relevant notions appropriately (namely, polynomials and power series with matrix coefficients), we can implement Strassen's idea and eliminate divisions, with the exception of one caveat—we don't know the size of matrices needed! As it turns out, it is a basic open problem, arising in non-commutative algebra as well as in commutative algebraic geometry, to determine a bound on the minimum size of matrices on which a nonzero rational expression does not vanish (resp., is invertible). Thus, our result is conditional: the simulation is polynomial in the size of the given circuit and in the size of the smallest dimension of matrices on which the given circuit can be correctly evaluated. Finally, we will see that this problem is also related to the question of deciding whether a rational expression computes the zero function. In the case of formulae, the "rational identity testing" problem can be decided by an efficient randomized algorithm, provided that the above matrix dimension is small.

Organization

In Section 2 we formally define our computational models, arithmetic circuits and formulae over non-commuting variables with division gates. We define rational functions, the skew field they live in and the notion of inverse height. Then we formally state our main results. In Section 3 we prove the circuit size upper bound on matrix inverse, and in Section 4 the formula size lower bound for it, via a general result about balancing formulae with division gates. In Section 5 we show that circuits require only as many inverse gates as the height, via an efficient simulation reducing the number of inverse gates to this bare minimum. In Section 6 we present several completeness results, most notably of matrix inverse for formulae. In Section 7 we define the identity testing problems for non-commutative polynomials and rational functions and discuss their complexities. In Section 8 we explain how to eliminate divisions when computing polynomials. In Section 9 we discuss some future directions and open problems. In Appendix A, we discuss a connection between the open problems and invariant theory.

2 Background and main results

Let $\mathbb{F}$ be a (commutative) field and $\bar{x} = x_1, \dots, x_n$ a set of variables. The ring of non-commutative polynomials in variables $\bar{x}$ will be denoted $\mathbb{F}\langle\bar{x}\rangle$, and $\mathbb{F}(\!\langle\bar{x}\rangle\!)$ denotes the free skew field of non-commutative rational functions. Two classical approaches to defining this field will be outlined below. For more detail see for example [9, 31]. The elements of $\mathbb{F}(\!\langle\bar{x}\rangle\!)$ are *non-commutative rational functions*, which we call simply *rational functions*.

Non-commutative arithmetic circuits with inverse gates

Non-commutative rational functions will be computed by means of *non-commutative arithmetic circuits with inverse gates*, which we call briefly *circuits*. This is a natural extension of both the notion of a commutative circuit with division gates, and the notion of a non-commutative circuit without divisions. We formally define the circuits, and then discuss how they lead to a definition of the free skew field.

A *circuit* Φ over a field $\mathbb{F}$ is a finite directed acyclic graph as follows. Nodes of in-degree zero are labeled by either a variable or a field element in $\mathbb{F}$. All the other nodes have in-degree one or two. The gates of in-degree one are labeled by $^{-1}$ and the gates of in-degree two by either $+$ or $\times$. The two edges going into a gate labeled by $\times$ are labeled by *left* and *right*, to determine the order of multiplication. The nodes are called input gates, inverse, sum and product gates. The nodes of out-degree zero are output gates. For nodes v, v_1, v_2, we write $v = v_1 \times v_2$ to indicate that v is a product gate with the two edges coming from v_1, v_2, and similarly for $v = v_1 + v_2$ or $v = v_1^{-1}$.

The *size* of a circuit Φ is the number of gates in Φ. Its *depth* is the length of the longest path in Φ. A *formula* is a circuit where every node has out-degree at most one.

For a node v in Φ, we denote Φ_v as the subcircuit of Φ rooted at v.

A node u in a circuit Φ in variables $x_1, \dots, x_n$ is intended to compute a non-commutative rational function $\widehat{u} \in \mathbb{F}(\!\langle x_1, \dots, x_n\rangle\!)$. However, the circuit may also contain division by zero, in which case we say that $\widehat{u}$ is *undefined*. The exact definition of $\widehat{u} \in \mathbb{F}(\!\langle x_1, \dots, x_n\rangle\!)$ is clear:

(i). If v is an input gate labeled by a (i. e., a is a variable or a field element), let $\widehat{v} := a$.

(*ii*). If $v = v_1 \times v_2$, resp. $v = v_1 + v_2$, let $\widehat{v} = \widehat{v}_1 \cdot \widehat{v}_2$, resp. $\widehat{v} = \widehat{v}_1 + \widehat{v}_2$, provided that both $\widehat{v}_1$ and $\widehat{v}_2$ are defined.

(*iii*). If $v = u^{-1}$, let $\widehat{v} := \widehat{u}^{-1}$, provided $\widehat{u}$ is defined and $\widehat{u} \neq 0$.

We say that a circuit Φ is a *correct circuit*, if $\widehat{u}$ is defined for every node in Φ. A correct circuit Φ *computes* a set of non-commutative rational functions $\widehat{\Phi} = \{\widehat{u}_1, \dots, \widehat{u}_m\}$, where $u_1, \dots, u_m$ are the output gates of Φ.

The free skew field—a computational definition

One classical definition of the field $\mathbb{F}\langle\!\langle \bar{x} \rangle\!\rangle$ is through the computation of its elements as above, with equivalence of elements defined through evaluating their circuits on matrix algebras as we outline now.

Let R be a ring whose centre contains the field $\mathbb{F}$ (i. e., every element of $\mathbb{F}$ commutes with every element of R). Let Φ be a circuit in variables $x_1, \dots, x_n$ with a single output node. Then Φ can be viewed as computing a *partial* function $\widehat{\Phi}^R : R^n \to R$. That is, substitute $a_1, \dots, a_n$ for the variables $x_1, \dots, x_n$ and evaluate the circuit. $\widehat{\Phi}^R(a_1, \dots, a_n)$ is undefined if we come across an inverse gate whose input is not invertible in R. Note that

$$\widehat{\Phi}^R(x_1, \dots, x_n) = \widehat{\Phi}$$

if we interpret $x_1, \dots, x_n$ as elements of $R = \mathbb{F}\langle\!\langle x_1, \dots, x_n \rangle\!\rangle$.

Looking at rings of $k \times k$ matrices $M_{k \times k}(\mathbb{F})$, we can obtain the following characterization of circuits and non-commutative rational functions:

Theorem 2.1 (Amitsur, [2], cf. [31])**.**

(i). Φ is a correct circuit iff the domain of $\widehat{\Phi}^R$ is non-empty for some $R = M_{k \times k}(\mathbb{F})$.

(ii). For correct circuits Φ_1, Φ_2, $\widehat{\Phi}_1 = \widehat{\Phi}_2$ iff $\widehat{\Phi}_1^R$ and $\widehat{\Phi}_2^R$ agree on the intersection of their domains, for every $R = M_{k \times k}(\mathbb{F})$.

In fact, those conditions can be used to define the skew field $\mathbb{F}\langle\!\langle x_1, \dots, x_n \rangle\!\rangle$. It can be constructed as the set of all correct circuits modulo the equivalence class induced by (*ii*).

Matrix inverse (and the quasi-determinant)

Let $A \in \mathrm{Mat}_{n \times n}(R)$ be an $n \times n$ matrix whose entries are elements of a unital ring R. Then $A^{-1} \in \mathrm{Mat}_{n \times n}(R)$ is the $n \times n$ matrix such that

$$A \cdot A^{-1} = A^{-1} \cdot A = I_n\,,$$

where I_n is the identity matrix. The inverse A^{-1} does not always exists, but if it does, it is unique. We will be specifically interested in the inverse of the $n \times n$ *generic matrix* $X_n \in \mathrm{Mat}_{n \times n}(\mathbb{F}\langle\!\langle \bar{x} \rangle\!\rangle)$, which is the matrix $X_n = (x_{ij})_{i,j \in [n]}$ consisting of n^2 distinct variables.

Matrix inverse is a very close cousin of the *quasi-determinant*. In two influential papers, Gelfand and Retakh [22, 23] defined a non-commutative analog to the determinant, called *quasi-determinant*, which they argued to be the appropriate generalization of that fundamental polynomial. Its many beautiful properties and applications are surveyed in [21]. The quasi-determinant of a generic matrix is actually a *set* of n^2 rational functions, which can be simply defined from the entries of the matrix inverse. Indeed, the (i,j) quasi-determinant of X is simply the *inverse* of the (i,j) entry of X^{-1}. Thus, essentially everything we say about matrix inverse holds for the quasi-determinant as well.

That X_n^{-1} exists can be directly proved by induction on n, as in our construction in Section 3. However, one can also invoke an interesting theorem due to Cohn. Let R be a ring. A matrix $A \in \mathrm{Mat}_{n\times n}(R)$ is called *full in* R if it cannot be written as $A = B \cdot C$ with $B \in \mathrm{Mat}_{n\times k}(R)$, $C \in \mathrm{Mat}_{k\times n}(R)$ and $k < n$.

Theorem 2.2 (Cohn, [9]). *Let $A \in Mat_{n\times n}(\mathbb{F}\langle \bar{x}\rangle)$ be a matrix of non-commutative polynomials. Then A is invertible in the skew field $\mathbb{F}\langle\!\langle \bar{x}\rangle\!\rangle$ if and only if it is full in $\mathbb{F}\langle \bar{x}\rangle$. Moreover, if A is not full and its entries are polynomials of degree ≤ 1, then the entries of the factors B, C are without loss of generality degree ≤ 1 as well.*

This characterization of invertible matrices was then used by Cohn to give an alternative construction of the free field: we can identify an element of $\mathbb{F}\langle\!\langle \bar{x}\rangle\!\rangle$ with an element of A^{-1} for some full $A \in M_{n\times n}(\mathbb{F}\langle \bar{x}\rangle)$. This is another indication of the key role matrix inverse has in the study of non-commutative rational functions. Note that in the *commutative* polynomial ring, there exist matrices which are both full and singular (e. g., (7.2) in Section 7.1).

The height of a rational function

An important characteristic of a rational function is the number of nested inverse operations necessary to express it. For a circuit Φ, we define the *height* of Φ as the maximum number k such that there exists a path in Φ which contains k inverse gates. For example, the formula $xy^{-1} + zx^{-1}y^2$ has height 1 and $(1 + xy^{-1}x)^{-1}$ has height 2. For a rational function f, the height of f is the smallest height of some circuit computing f (in this definition, one may equivalently consider formulae). Naturally, the depth of a circuit computing f must be at least the height of f.

In the commutative setting, every rational function can be written as fg^{-1} for some polynomials f, g and so has height at most 1. In the non-commutative setting, there exist rational functions of an arbitrary height. This is in itself a remarkable and non-trivial fact. However, we will use a stronger statement:

Theorem 2.3 (Reutenauer, [43]). *The height of any entry of the generic inverse matrix X_n^{-1} is n.*

In Section 5, we will give a different characterization of the inverse height of f: in Corollary 5.3, we point out that a rational function of height k can be computed by a circuit which *altogether* uses only k inverses. Hence the height of f can also be defined as the smallest *number* of inverse gates needed to compute f by means of a circuit.

Main results

We shall prove the following two theorems about the complexity of matrix inverse (recall that X_n is a matrix of n^2 distinct variables):

Theorem 2.4. *X_n^{-1} can be computed by a circuit of size polynomial[5] in n.*

[5] In fact, we show that X_n^{-1} can be computed by a circuit of size $O(n^{\omega})$, where $2 \leq \omega < 3$ is the exponent of matrix multiplication.

Theorem 2.5. *Every formula computing some entry of X_n^{-1} has size $2^{\Omega(n)}$.*

Theorem 2.4 is an explicit construction. Theorem 2.5 is obtained by showing that a formula of size s can be balanced to obtain an equivalent formula of depth $O(\log s)$. This entails that if f can be computed by a formula of size s, then f has height at most logarithmic in s. This gives Theorem 2.5 by Theorem 2.3.

Theorems 2.4 and 2.5 can be strengthened as follows:

(i). X_n^{-1} can be computed by a polynomial size circuit which contains only n inverse gates (cf. Proposition 5.2).

(ii). Every formula computing some entry of X_n^{-1} has $2^{\Omega(n)}$ inverse gates (cf. Corollary 4.4).

In his seminal paper [49], Valiant has shown that an arithmetic formula can be expressed as the determinant of a linear size matrix whose entries are variables or field elements. This result considers commutative formulae without inverse gates. That commutativity is not essential was later shown in [26]. Here, we show that a similar relationship holds between non-commutative arithmetic formulae and the matrix inverse:

Theorem 2.6. *Assume that a rational function f can be computed by a formula Φ of size s. Then there exists $s' \leq 2s$ and an invertible $s' \times s'$-matrix A_Φ whose entries are variables or field elements such that f is an entry of A_Φ^{-1}.*

This is proved in Section 6. There, we also discuss some variants of the theorem. Namely, if f is computed without the use of inverse gates then A can be taken upper triangular, and we point out the connection with the non-commutative determinant.

We present several other results about the number of inverse gates in non-commutative circuits - how to minimize them when computing rational functions, and how to eliminate them when computing polynomials. More specifically:

- If f can be computed by a circuit of size s and height k, then f can be computed by a circuit of size $O(s(k+1))$ which contains k inverse gates.

In other words, if f can be computed by a circuit which contains at most k inverses on any *directed path*, it can be computed by a circuit with k inverse gates *in total*, with only a small increase in circuit size. (Proposition 5.2 in Section 5).

- Let $f \in \mathbb{F}\langle \bar{x} \rangle$ be a polynomial of degree d which is computable by a circuit with divisions of size s. Assume that there exist matrices $a_1, \dots, a_n \in R = \mathrm{Mat}_{m\times m}(\mathbb{F})$ such that $\widehat{\Phi}^R(a_1, \dots, a_n)$ is defined. Then f can be computed by a division-free circuit of size $O(sd^3m^3)$.

This is an analogy of the elimination of division gates from commutative circuits. However, we do not know how large can the parameter $m = m(s,d)$ be for the worst such circuit, and hence we do not know whether our construction is polynomial in s and d. (See Section 8).

A version of this parameter appears again in Section 7 in connection with the *rational identity testing problem*.

- For a correct formula Φ of size s, one can decide whether $\widehat{\Phi} = 0$ by a randomized algorithm which runs in time polynomial in $s \cdot w(s)$.

Here, $w(s)$ is defined as the smallest k so that every correct formula of size s can be correctly evaluated on some $p \times p$ matrices with $p \leq k$. The algorithm can proceed simply by evaluating Φ at random $p \times p$ matrices, for every $p \leq w(s)$. Via the completeness theorem above, we also consider a different version of the problem: given an $s \times s$ matrix A with linear functions as entries, decide whether A is invertible in the free skew field. Again, we obtain a probabilistic algorithm running in time $s\tilde{w}(s)$, where $\tilde{w}(s)$ is a related parameter. We note that in the recent paper [29], Ivanyos, Qiao and Subrahmanyam designed a *deterministic* algorithm for this problem with similar parameters.

An upper bound on $w(s)$ (and $\tilde{w}(s)$) can be obtained by solving the following basic problem in invariant theory. This is also the most important open problem our work suggests, and we conclude this section by stating it. In a slightly different formulation, it is presented in Section 9 as Problem 9.4 (that the formulations are equivalent follows from Proposition 7.3), and is also discussed in Section 7.1. $\mathbb{F}$ can be any field, but it is especially interesting for algebraically closed fields, and specifically for the complex numbers.

- Find an upper bound on the smallest $k = k(s)$, such that for all $Q_1, \ldots, Q_n \in \mathrm{Mat}_{s\times s}(\mathbb{F})$ if[6] $\sum_{i=1}^{n} Q_i \otimes a_i$ is invertible for some $m \in \mathbb{N}$ and $a_1, \ldots, a_n \in \mathrm{Mat}_{m\times m}(\mathbb{F})$ then $\sum_{i=1}^{n} Q_i \otimes a_i$ is invertible for some $a_1, a_2, \ldots, a_n \in \mathrm{Mat}_{p\times p}(\mathbb{F})$ with $p \leq k$.

In other words, we want to find k such that $\det(\sum_{i=1}^{n} Q_i \otimes a_i)$ does not identically vanish on $(\leq k) \times (\leq k)$ matrices. Note that the vanishing of the determinant is invariant to acting on the sequence Q_i with left and right multiplication by any two invertible matrices—this provides the connection to invariant theory. This connection is further discussed in Appendix A.

3 A polynomial-size circuit for matrix inverse

In this section, we show that X_n^{-1} can be computed by a polynomial size circuit, thus proving Theorem 2.4. The algorithm is implicit in Strassen's paper [47].

The construction of X^{-1}

Let $X = X_n = (x_{ij})_{i,j\in[n]}$ be a matrix consisting of n^2 distinct variables. We define the matrix X^{-1} recursively. If $n = 1$, let $X^{-1} := (x_{11}^{-1})$. If $n > 1$, divide X into blocks as

$$X = \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix}, \tag{3.1}$$

where a_1, a_4 are square matrices of dimensions $p \times p$ and $(n-p) \times (n-p)$, respectively, and a_2, a_3 are in general rectangular matrices of dimension $p \times (n-p)$ and $(n-p) \times p$, respectively. (Later, we will choose p as roughly $n/2$.) Assume we have already constructed the matrix a_1^{-1}. Let

$$z := a_4 - a_3 a_1^{-1} a_2,$$

[6] $\otimes$ is the Kronecker product.

and

$$X^{-1} := \begin{pmatrix} a_1^{-1}(I + a_2 z^{-1} a_3 a_1^{-1}) & -a_1^{-1} a_2 z^{-1} \\ -z^{-1} a_3 a_1^{-1} & z^{-1} \end{pmatrix} . \tag{3.2}$$

Here, we should argue that z^{-1} exists, which is however apparent from the fact that $z = a_4$ if we set $a_3 = 0$ and a_4 is invertible.

Correctness

We must show that X^{-1}, as constructed above, indeed satisfies $X \cdot X^{-1} = I$ and $X^{-1} \cdot X = I$. For $n = 1$, we have $x_{11} \cdot x_{11}^{-1} = x_{11}^{-1} \cdot x_{11} = 1$. Otherwise let $n > 1$ and let X be as in (3.1). Using some rearrangements and the definition of z, we obtain

$$\begin{aligned} X \cdot X^{-1} &= \begin{pmatrix} a_1 & a_2 \\ a_3 & a_4 \end{pmatrix} \begin{pmatrix} a_1^{-1}(I + a_2 z^{-1} a_3 a_1^{-1}) & -a_1^{-1} a_2 z^{-1} \\ -z^{-1} a_3 a_1^{-1} & z^{-1} \end{pmatrix} \\ &= \begin{pmatrix} I + a_2 z^{-1} a_3 a_1^{-1} - a_2 z^{-1} a_3 a_1^{-1} & -a_2 z^{-1} + a_2 z^{-1} \\ a_3 a_1^{-1} + (a_3 a_1^{-1} a_2 - a_4) z^{-1} a_3 a_1^{-1} & (a_4 - a_3 a_1^{-1} a_2) z^{-1} \end{pmatrix} \\ &= \begin{pmatrix} I & 0 \\ a_3 a_1^{-1} - z z^{-1} a_3 a_1^{-1} & z z^{-1} \end{pmatrix} = \begin{pmatrix} I & 0 \\ 0 & I \end{pmatrix} . \end{aligned}$$

The proof of $X^{-1} \cdot X = I$ is constructed in a similar fashion.

Complexity

Assume first that n is a power of two. Then X in (3.1) can be partitioned into four matrices of dimensions $n/2 \times n/2$. This shows that in order to compute the inverse of an $n \times n$ matrix, it is sufficient to compute the inverse of two $n/2 \times n/2$ matrices (a_1 and z), and add or multiply a constant number of $n/2 \times n/2$ matrices. Let $M(n)$ be the size of a smallest circuit which computes the product of two matrices and $C(n)$ the size of a smallest circuit computing X_n^{-1}. Then we obtain

$$C(n) \leq 2C(n/2) + c_1 M(n/2) + c_2 n^2 ,$$

for some constants c_1, c_2. If $M(n) = O(n^{\omega})$, with $2 \leq \omega$, this implies that $C(n) = O(n^{\omega})$.

If n is not a power of two, at each step partition X as evenly as possible, i. e., set $p := \lceil n/2 \rceil$. This gives that $C(n) \leq C(2^k)$, where k is the smallest integer such that $2^k \geq n$. Since $2^k \leq 2n$, this amounts to a loss of at most a constant factor.

Moreover, it is easy to see that the constructed circuit has height n.

4 Matrix inverse has exponential formula size

In this section, we prove Theorem 2.5. For this purpose, we show that a formula of size s can be balanced to obtain an equivalent formula of depth $O(\log s)$. Both the statement and its proof are analogous to the commutative version given by Brent in [6]. On the other hand, the fact that this statement does hold for non-commutative rational functions is in itself slightly surprising.

Proposition 4.1. *Assume that a non-commutative rational function f can be computed by a formula of size s. Then f can be computed by a formula of depth $O(\log s)$.*

This immediately implies:

Corollary 4.2. *If f can be computed by a formula of size s then f has height $O(\log s)$.*

This, together with Theorem 2.3, implies Theorem 2.5.

Let us first prove Proposition 4.1:

Proof of Proposition 4.1. The proof is almost identical to Brent's commutative version. Hence we just outline the argument and point out the peculiarities arising in the non-commutative setting.

The idea is to simultaneously prove the following two statements by induction on s. Let Φ be a correct formula of size s then, for sufficiently large s and suitable constants $c_1, c_2 > 0$, the following hold:

(i). $\widehat{\Phi}$ can be computed by a formula of depth $c_1 \log s$,

(ii). if z is a variable occurring at most once in Φ then

$$\widehat{\Phi} = (Az+B)(Cz+D)^{-1},$$

where A,B,C,D are non-commutative rational functions which do not depend on z and each is computable by a formula of depth $\le c_2 \log s$. Moreover, $C\widehat{\Psi}+D \neq 0$ for any Ψ such that $\Phi(z/\Psi)$ is correct.

Here $\Phi(z/\Psi)$ means that Ψ is substituted for z in Φ. Furthermore, for a node v in Φ, Φ_v will stand for the subformula of Φ with the output v and $\Phi_{v:=z}$ is the formula obtained by replacing Φ_v in Φ by the variable z.

On the inductive step, (i) is obtained roughly as follows. Find a node v in Φ such that both Φ_v and $\Phi_{v:=z}$ are small (of size at most $2s/3$). Apply part (i) of the inductive assumption to bound the depth of $\widehat{\Phi}_v$. Apply (ii) to write $\widehat{\Phi}_{v:=z} = (Az+B)(Cz+D)^{-1}$, with A,B,C,D having small depth, which altogether gives a small depth formula for $\widehat{\Phi} = (A\widehat{\Phi}_v+B)(C\widehat{\Phi}_v+D)^{-1}$. To prove (ii), find an appropriate node v on the path between z and the output of Φ. (An "appropriate v" is a node v such that $\Phi_{v:=z_1}$ is small and Φ_{u_1} is small, where either $v = u_1^{-1}$, $v = u_1 + u_2$, $v = u_1 \times u_2$, or $v = u_2 \times u_1$, where u_2 does not depend on z.) Use the inductive assumptions to write

$$\begin{aligned} \widehat{\Phi}_{v:=z_1} &= (A_1 z_1 + B_1)(C_1 z_1 + D_1)^{-1}, \\ \widehat{\Phi}_v &= (A_2 z + B_2)(C_2 z + D_2)^{-1}, \end{aligned}$$

and compose these expressions to obtain (ii).

The main point that needs to be checked is that the representation

$$f = (Az+B)(Cz+D)^{-1} \tag{4.1}$$

is well-behaved in the non-commutative setting. If A,B,C,D do not depend on z, we will say that f in (4.1) has *z-normal form*. It is not immediately clear that if Φ is as in (ii), then $\widehat{\Phi}$ has a z-normal form (even if we require nothing about the complexity of A,B,C,D).

To see this, assume that z appears at most once in Φ. Let d be the length of the shortest path from z to the root of Φ and proceed by induction on d. (Hence, $d = 0$ if $\Phi = z$ and we set $d := -1$ if z does not appear in Φ.) If $d \leq 1$, we must convert the following

$$A\,,\; z\,,\; z^{-1}\,,\; z+A\,,\; A\cdot z\,,\; z\cdot A\,,$$

with A not depending on z, to a z-normal form. This is clear, perhaps with the exception of $z\cdot A$. If $A \neq 0$, this can be written as $zA = z(A^{-1})^{-1} = (1z+0)(0z+A^{-1})^{-1}$; if $A = 0$ then $z\cdot A = (0z+0)(0z+1)^{-1}$. If $d > 1$, it is enough to show that normal forms can be composed in the following sense. If

$$f = (A_1 z_1 + B_1)(C_1 z_1 + D_1)^{-1}\,,$$
$$z_1 = (A_2 z + B_2)(C_2 z + D_2)^{-1}\,,$$

then

$$f = (Az+B)(Cz+D)^{-1}\,,$$

where $A = A_1A_2 + B_1C_2$, $B = A_1B_2 + B_1D_2$, $C = C_1A_2 + D_1C_2$ and $D = C_1B_2 + D_1D_2$. To see this, let $h_1 := A_2z + B_2$ and $h_2 := C_2z + D_2$ so that $z_1 = h_1h_2^{-1}$. Then

$$\begin{aligned} f &= (A_1h_1h_2^{-1} + B_1)(C_1h_1h_2^{-1} + D_1)^{-1} \\ &= (A_1h_1 + B_1h_2)h_2^{-1}((C_1h_1 + D_1h_2)h_2^{-1})^{-1} \\ &= (A_1h_1 + B_1h_2)h_2^{-1}h_2(C_1h_1 + D_1h_2)^{-1} \\ &= (A_1h_1 + B_1h_2)(C_1h_1 + D_1h_2)^{-1}\,, \end{aligned}$$

and substitute for h_1, h_2.

This shows that $\widehat{\Phi}$ indeed has a z-normal form. Furthermore, the outlined compositionality of z-normal forms is quite sufficient to reconstruct Brent's proof. □

Let us note that the proof of Proposition 4.1 does not rely on the fact that the formula computes an element of a *free* skew field, and the proposition can be generalized as follows (recall the definition of $\widehat{\Phi}^R$ from Section 2):

Proposition 4.3. *Let $\mathbb{F}$ be a field and R a skew field whose centre contains $\mathbb{F}$. Assume that $\Psi(x_1,\dots,x_n)$ is a formula of size s and $\widehat{\Psi}^R(a_1,\dots,a_n)$ is defined for given $a_1,\dots,a_n \in R$. Then there exists a formula $\Phi(x_1,\dots,x_n)$ of depth $O(\log s)$ such that*

$$\widehat{\Phi}^R(a_1,\dots,a_n) = \widehat{\Psi}^R(a_1,\dots,a_n)\,.$$

This, together with Lemma 4.5 below, implies that Corollary 4.2 can be rephrased in terms of the number of inverse gates only:

Corollary 4.4. *Assume that f can be computed by a formula with k inverse gates. Then f has height $O(\log k)$.*

Consequently, any formula computing an entry of X_n^{-1} must have an exponential number of inverse gates. Corollary 4.4 follows from the following lemma, where we apply Proposition 4.3 to the formula $\Psi(z_1,\dots,z_m)$.

Lemma 4.5. *Assume that a rational function $f \in R = \mathbb{F} \langle\!\!\!\!\langle \bar{x} \rangle\!\!\!\!\rangle$ can be computed by a formula with k inverse gates. Then there exists a formula $\Psi(z_1, \dots, z_m)$ such that*

(i). *Ψ has size $O(k)$,*

(ii). *$f = \widehat{\Psi}^R(h_1, \dots, h_m)$ for some polynomials $h_1, \dots, h_m \in \mathbb{F}\langle \bar{x} \rangle$.*

Proof. Let us call a gate u in a formula Φ *trivial* if Φ_u does not contain any inverse gate. Let us call u *maximal* if it is trivial and for every trivial $v \neq u$, u is not a gate in Φ_v.

Assume that f can be computed by a formula with k inverse gates. It is sufficient to show that f can be computed by a formula Φ with $m = O(k)$ maximal gates. For let $u_1, \dots, u_m$ be the maximal gates in Φ. Introduce new variables $z_1, \dots, z_m$, and let Ψ be the formula obtained by replacing every u_i by z_i in Φ. Clearly, $f = \widehat{\Psi}^R(h_1, \dots, h_m)$, where h_i is the polynomial computed by Φ_{u_i}. Moreover, Ψ is a formula with m leaves. If we assume that Ψ does not contain redundant inverse gates $(u^{-1})^{-1}$, then Ψ has size $O(m)$.

To construct a formula with $O(k)$ maximal gates computing f, assume that $k > 0$. First, show that any formula with k inverses can be transformed to an equivalent formula Φ with k inverses such that every maximal gate u occurs in Φ in one of the following contexts:

(i). u^{-1},

(ii). $u \times v^{-1} \times u'$ or $u' \times v^{-1} \times u$, where u' is itself maximal,

(iii). $v_1 \times v_2 + u$, where $v_1 \times v_2$ is non-trivial.

This is easily proved by induction on k. If $k = 1$, we are supposed to compute f by a formula of the form $u_1 \times v^{-1} \times u_2 + u_3$, where v, u_1, u_2, u_3 do not contain inverse gates.

Finally, let us argue that Φ contains $O(k)$ maximal gates. For every inverse gate, there are at most 3 maximal gates corresponding to the conditions (i) and (ii). This also means that the number of non-trivial product gates $v_1 \times v_2$ in Ψ is $O(k)$ and so there are $O(k)$ maximal gates corresponding to the condition (iii). □

5 Height versus the number of inverse gates

Recall that the height of a circuit is the maximum number of inverse gates along *some directed path* in the circuit. Here we show that a circuit of height k can be transformed to an equivalent circuit which *altogether* uses only k inverse gates. This means that the height of f can be equivalently defined as the smallest number of inverse gates needed to compute f by a circuit.

This is based on the observation that $x_1^{-1}, \dots, x_n^{-1}$ can be simultaneously computed using one inverse gate only:

Lemma 5.1. *The rational functions $x_1^{-1}, \dots, x_n^{-1}$ can be simultaneously computed by a circuit of size $O(n)$ which contains only one inverse gate.*

Proof. Let $z := x_1 x_2 \cdots x_n$. As $z^{-1} = x_n^{-1} x_{n-1}^{-1} \cdots x_1^{-1}$, we have for every $j \in \{1, \dots, n\}$

$$x_j^{-1} = (x_{j+1} x_{j+2} \cdots x_n) z^{-1} (x_1 x_2 \cdots x_{j-1}) .$$

Hence $x_1^{-1},\dots,x_n^{-1}$ can all be computed using just the inverse z^{-1}. To see that this gives a linear size circuit, it is sufficient to construct a linear size circuit simultaneously computing the polynomials $f_i = x_i x_{i+1}\cdots x_n$ and $g_i = x_1 x_2 \cdots x_i$, for $i \in \{1,\dots,n\}$. This is clear because $f_i = x_i f_{i+1}$ and $g_{i+1} = g_i x_{i+1}$. □

Proposition 5.2. *Assume that a rational function f can be computed by a circuit of size s and height k. Then f can be computed by a circuit of size $O(s(k+1))$ which contains k inverse gates.*

Proof. Assume that Φ is a circuit of size s, inverse height k, which computes f. We construct the new circuit by induction on k. If $k=0$, the statement is trivial, and so assume that $k>0$. Let $u_1^{-1},\dots,u_n^{-1}$ be the inverse gates in Φ such that Φ_{u_i} does not contain any inverse gate. By the previous lemma, the rational functions computed by $u_1^{-1},\dots,u_n^{-1}$ can be computed by a circuit Ψ of size $c\cdot s$ which contains only one inverse gate. Let Φ' be the circuit obtained from Φ by replacing the gate u_i^{-1} by a fresh variable z_i, for every $i\in\{1,\dots,n\}$. The circuit Φ' has inverse height $k-1$ and size $\le s$, and so we can construct an equivalent circuit of size csk with only $k-1$ division gates. Feeding the outputs of Ψ into the circuit Φ', we obtain a circuit computing f which has k inverse gates and size $csk + cs = cs(k+1)$. □

Corollary 5.3. *The following are equivalent:*

(i). f has inverse height $\le k$,

(ii). f can be computed by a circuit with $\le k$ inverse gates.

As follows from Corollary 4.4, the proposition or its corollary do not hold for formulae. Moreover, every formula computing $x_1^{-1}+\cdots+x_n^{-1}$ must contain a linear number of inverse gates (cf. Corollary 6.2 and Proposition 7.7.7 of [8]).

6 Formula completeness of matrix inverse

Here we prove Theorem 2.6. Our construction of matrices from formulae is similar to Malcolmson's approach for the construction of the skew field [35]. After this proof, we proceed in the following subsections to discuss the special case of formulae without inverse gates, in which the computation produces a non-commutative polynomial, and compare with the commutative case.

Proof of Theorem 2.6. The matrix A is constructed by induction on s. We retain the property that f is the entry in the upper-right corner of A^{-1}. This entry will be denoted $\mathcal{R}A^{-1}$.

Let A,B be invertible matrices of dimension $p\times p$ and $q\times q$. For $k\in\{p,q\}$, let u_k, resp. v_k, be the $1\times k$ vector whose first, resp. last, component is 1 and the others are zero. Furthermore, let a^t be the first column of A. The key observation are the following equalities (the unspecified entries are zero):

$$\mathcal{R}A^{-1}\cdot\mathcal{R}B^{-1} = \mathcal{R}\begin{pmatrix} A & -v_p^t\cdot u_q \\ & B \end{pmatrix}^{-1}, \tag{6.1}$$

$$\mathcal{R}A^{-1}+\mathcal{R}B^{-1} = \mathcal{R}\begin{pmatrix} A & a^t\cdot u_q & -v_p^t \\ & B & v_q^t \\ & & 1 \end{pmatrix}^{-1}, \tag{6.2}$$

$$(\mathcal{R}A^{-1})^{-1} = \mathcal{R}\begin{pmatrix} v_p^t & A \\ 0 & -u_p \end{pmatrix}^{-1} \qquad \text{if } \mathcal{R}A^{-1}\neq 0. \tag{6.3}$$

To prove (6.1) and (6.2), note that

$$\begin{pmatrix} a_1 & a_2 \\ 0 & a_4 \end{pmatrix}^{-1} = \begin{pmatrix} a_1^{-1} & -a_1^{-1}a_2a_4^{-1} \\ 0 & a_4^{-1} \end{pmatrix},$$

whenever the right hand side makes sense. This follows from (3.2), noting that (3.2) holds whenever X is a matrix such that the right hand side makes sense. This gives

$$\begin{aligned}\mathcal{R}\begin{pmatrix} A & -v_p^tu_q \\ & B \end{pmatrix}^{-1} &= \mathcal{R}\begin{pmatrix} A^{-1} & A^{-1}v_p^tu_qB^{-1} \\ & B^{-1} \end{pmatrix} \\ &= \mathcal{R}(A^{-1}v_p^tu_qB^{-1}) = (\mathcal{R}A^{-1})(\mathcal{R}B^{-1}).\end{aligned}$$

Similarly, one can verify that the entry in the upper-right corner of the matrix in (6.2) is the first entry of the $p\times 1$-vector

$$w \;=\; A^{-1}v_p^t + A^{-1}a^tu_qB^{-1}v_q^t \;=\; A^{-1}v_p^t + u_p^tu_qB^{-1}v_q^t \;=\; A^{-1}v_p^t + (\mathcal{R}B^{-1})u_p^t\,,$$

where we used that $A^{-1}a^t = u_p^t$. The first entry of w is therefore $\mathcal{R}A^{-1}+\mathcal{R}B^{-1}$.

To prove (6.3), apply (3.2), with $a_1 = A$ and $a_4 = 0$, to show that the entry in *lower*-right corner of

$$\begin{pmatrix} A & v_p^t \\ -u_p & 0 \end{pmatrix}^{-1}$$

is $(u_pA^{-1}v_p^t)^{-1} = (\mathcal{R}A^{-1})^{-1}$. We have

$$\begin{aligned}\begin{pmatrix} v_p^t & A \\ 0 & -u_p \end{pmatrix}^{-1} &= \left(\begin{pmatrix} A & v_p^t \\ -u_p & 0 \end{pmatrix}\begin{pmatrix} & I \\ 1 & \end{pmatrix}\right)^{-1} \\ &= \begin{pmatrix} & 1 \\ I & \end{pmatrix}\begin{pmatrix} A & v_p^t \\ -u_p & 0 \end{pmatrix}^{-1},\end{aligned}$$

and so $\mathcal{R}\begin{pmatrix} v_p^t & A \\ 0 & -u_p \end{pmatrix} = (\mathcal{R}A^{-1})^{-1}$.

Equipped with (6.1)-(6.3), the statement of the theorem is directly proved by induction. If $s = 1$, f is either a variable or a field element and we have

$$f = \mathcal{R}\begin{pmatrix} 1 & f \\ 0 & -1 \end{pmatrix}^{-1}.$$

If $s > 1$, consider the output node of the size-s formula computing f and apply (6.1)-(6.3) accordingly. □

The matrix A_Φ can be written as $Q_0 + x_1Q_1 + \cdots + x_nQ_n$, where Q_0 is a matrix in $\mathbb{F}$ and $Q_1, \dots Q_n$ are $0,1$-matrices. In general, if A is a matrix of the form

$$Q_0 + x_1Q_1 + \cdots + x_nQ_n \quad \text{with} \quad Q_0, \dots, Q_n \in \mathrm{Mat}_{s' \times s'}(\mathbb{F}),$$

R is a ring with $R \supseteq \mathbb{F}$, and $a_1, \dots, a_n \in R$, we define

$$A(a_1, \dots, a_n) := Q_0 + a_1Q_1 + \cdots + a_nQ_n;$$

it is a matrix in $\mathrm{Mat}_{s' \times s'}(R)$. Theorem 2.6 can be generalized in the following manner:

Proposition 6.1. *Let $\mathbb{F}$ be a field and R a ring whose centre contains $\mathbb{F}$. Assume that $\Phi(x_1, \dots, x_n)$ is a formula of size s and $a_1, \dots, a_n \in R$ are such that $\widehat{\Phi}^R(a_1, \dots, a_n)$ is defined. Then $A_\Phi(a_1, \dots, a_n)$ is invertible in $M_{s' \times s'}(R)$, and $\widehat{\Phi}^R(a_1, \dots, a_n)$ is an entry in A_Φ^{-1}.*

The proof is almost identical to that of Theorem 2.6. The only difference is that we do not assume that R is a skew field, and we must check that the constructed matrix A is indeed invertible. This follows from the assumption that $\widehat{\Phi}^R(a_1, \dots, a_n)$ is defined.

By means of Lemma 4.5, Proposition 6.1 implies:

Corollary 6.2. *Assume that a rational function f can be computed by a formula with k inverses. Then there exists an $O(k) \times O(k)$ invertible matrix A whose entries are polynomials such that f is an entry of A^{-1}.*

The matrix inverse representation of rational functions will be directly used in the next section, on the *rational identity testing* problem. Before that we take a detour to discuss the special case of formulae without inverse gates.

6.1 Triangular matrices

As remarked in the "Main results" section, commutatively or not, the determinant is complete for formulae without inverse gates. That is, a polynomial f can be written as $f = \det(A)$, where A is a matrix of variables or field elements whose dimension is linear in the formula size of f. One difference between the commutative and non-commutative settings is the following: the commutative determinant can be computed by a formula of size $n^{O(\log n)}$, but non-commutatively, it requires a formula of exponential size (see [39]). However, let us remark here that this gulf is by no means necessary. Inspecting the construction presented in [26], one can see that it uses only matrices A of a specific form, and one can show that the determinant of such matrices can be computed by a non-commutative formula of quasipolynomial size.

Let us first see what happens with our construction in Theorem 2.6, if we assume that f is a non-commutative polynomial computed by a formula without inverse gates. The constructed matrix A is then upper triangular with field elements on the diagonal. We summarize the properties of such matrices in the following proposition:

Proposition 6.3.

(i). Let A be an $n \times n$-upper triangular matrix whose entries are variables or field elements such that A has non-zero field elements on the diagonal. Then the entries of A^{-1} are polynomials. Moreover, every entry can be computed by a formula without inverse gates of size $n^{O(\log n)}$.

(ii). Assume that a polynomial f can be computed by a formula without inverse gates of size s. Then there exists a $2s \times 2s$-upper triangular matrix A whose entries are variables or field elements, A has 1 or -1 on the diagonal, and f is the entry in the upper-right corner of A^{-1}.

Proof. (*i*) Without loss of generality assume that A has the constant 1 on the diagonal, and write $A = I - J$, where J is an upper triangular matrix with zeros on the diagonal. This implies that $J^n = 0$ and therefore

$$A^{-1} = (I-J)^{-1} = I + J + J^2 + \cdots + J^{n-1}\,.$$

In general, if B is $n \times n$ matrix then every entry of B^k can be computed by a formula of size $n^{O(\log k)}$. This means that A^{-1} has formula size $n^{O(\log n)}$.

(*ii*) Inspect the construction in the proof of Theorem 2.6. The equation (6.3) is never applied since f is computed without inverse gates. Equations (6.1) and (6.2) convert upper-triangular matrices A, B to an upper triangular matrix. □

One may wonder whether Proposition 6.3 can be improved to show that A^{-1} can be computed by a polynomial size formula, if A is upper triangular. However, this problem is—up to a polynomial—equivalent to the problem of computing the product $A_1 A_2 \cdots A_k$ of $n \times n$ matrices $A_1, \ldots, A_k$. We have seen one half of this statement in the proof of Proposition 6.3: in order to compute A^{-1}, it is sufficient to compute $J, J^2, \ldots, J^{n-1}$. Conversely, suppose that we want to compute the product of $n \times n$ matrices $A_1 A_2 \cdots A_k$. Let A be the $((k+1)n) \times ((k+1)n)$–matrix

$$A := \begin{pmatrix} I & -A_1 & & & \\ & I & -A_2 & & \\ & & \ddots & & \\ & & & I & -A_k \\ & & & & I \end{pmatrix}. \tag{6.4}$$

One can check that the $n \times n$-block in the upper right corner of A^{-1} equals $A_1 A_2 \cdots A_k$. This means that the inverse of an upper triangular matrix has a polynomial size formula iff the iterated matrix product can be computed by a polynomial size formula, which is believed to be unlikely.

This observation can be used to strengthen Proposition 6.3 to apply to algebraic branching programs instead of formulae. An *algebraic branching program*, ABP, is a directed acyclic graph without multiple edges and with exactly one source and one sink such that every edge is labeled by a (not necessarily homogeneous) linear function $\sum_i a_i x_i + b$. An ABP computes a polynomial which is the sum, over all paths from the source to the sink, of products of the linear functions on that path. We are considering non-commutative computation, and the order of multiplication is taken from the source to the sink. The size of an ABP is the number of vertices.

Algebraic branching programs can simulate formulae (without inverse gates), and are believed to be more powerful. This is because the product $A_1 A_2 \cdots A_k$ of $n \times n$ matrices can be computed by an ABP of size $O(kn^2)$, but the best known formula has size $n^{O(\log k)}$. In fact, algebraic branching programs can be characterized in terms of matrix product:

Lemma 6.4. *Assume that a polynomial f in n variables can be computed by an ABP of size s. Then there exist $k \times k$ matrices $A_1, \ldots, A_{2s}$ with $k = O(ns^2)$ such that their entries are variables or field elements and f is the entry in the upper-right corner of the product $A_1 A_2 \cdots A_{2s}$.*

Proof. Let f have n variables and let C be an ABP of size s computing f. First, show that f can be computed by an ABP B with the following properties:

(i). Every edge is labeled by a variable or a field element.

(ii). B has $2s+1$ levels: the vertices can be partitioned into disjoint sets $B_0, B_1, \ldots B_{2s}$ with every edge going from B_i to B_{i+1} , and with B_0 and B_{2s} containing only the source and the sink, respectively.

(iii). For every $i \in \{1, \ldots, 2s-1\}$, B_i has size $k = O(ns^2)$.

(i) is obtained by taking every edge in C labeled by $a_1 x_1 + \cdots + a_n x_n + b$ and replacing it by n new vertices and $2n+1$ edges, labeled by $a_1, x_1, \ldots, a_n, x_n$ and b respectively. Since C has at most $\binom{s}{2}$ edges, the new ABP has $k = O(ns^2)$ vertices. Moreover, since C had depth at most s, the new ABP has depth at most $2s$. The conditions (ii) and (iii) will be guaranteed by adding a copy of every vertex to every level $B_1, \ldots, B_{2s-1}$, with appropriate labels of edges.

For $i \in 1, \ldots, 2s-1$, let v_i be the vector of the k polynomials computed by the nodes in B_i and let $v_0 := (1, 0, \ldots, 0)$ and $v_{2s} := (0, \ldots, 0, f)$. The condition ($i$) guarantees that we can find, for every $i \in \{1, \ldots, 2s\}$, a $k \times k$ matrix A_i of variables or field elements such that

$$v_i = v_{i-1} A_i \,.$$

Hence $v_{2s} = v_0 A_1 A_2 \cdots A_{2s}$, which implies that f is the entry in the upper-right corner of $A_1 A_2 \cdots A_{2s}$. □

Proposition 6.5.

(i). Let A be an $n \times n$-upper triangular matrix whose entries are variables or field elements such that A has non-zero field elements on the diagonal. Then every entry of A^{-1} can be computed by a polynomial size ABP.

(ii). Assume that a polynomial f in n variables can be computed by an ABP of size s. Then there exists a $k \times k$-upper triangular matrix A with $k = O(ns^3)$ whose entries are variables or field elements, A has 1 on the diagonal, and f is the entry in the upper-right corner of A^{-1}.

Proof. (*i*) is as in Proposition 6.3, where we note that J^k can be computed by a polynomial-size ABP.

(*ii*). Let $A_1, \dots, A_{2s}$ be the matrices from the previous lemma. Let

$$A := \begin{pmatrix} I & A_1 & & & & \\ & I & A_2 & & & \\ & & \ddots & & & \\ & & & & I & A_{2s} \\ & & & & & I \end{pmatrix}.$$

be as in (6.4). Then the upper-right block in A^{-1} is $A_1 \cdots A_{2s}$ and hence f is the entry in the upper-right corner of A^{-1}. □

6.2 The determinant of nearly triangular matrices

We now discuss the connection between matrix inverse of triangular matrices and the determinant of *nearly* triangular matrices. If $X_n = (x_{i,j})_{i,j\in[n]}$, let

$$\begin{aligned} \det(X) &= \sum_{\sigma\in S_n} sgn(\sigma) x_{1,\sigma(1)} \cdots x_{n,\sigma(n)}\,, \\ \operatorname{perm}(X) &= \sum_{\sigma\in S_n} x_{1,\sigma(1)} \cdots x_{n,\sigma(n)}\,. \end{aligned}$$

An $n \times n$-matrix A will be called *nearly (upper) triangular*, if for every $i, j \in [n]$, $A_{i,j} \in \{1, -1\}$, if $j = i - 1$, and $A_{i,j} = 0$, if $j < i - 1$. That is, A is upper triangular, except for a string of 1 and -1 below the main diagonal.

As an application of Propositions 6.3 and 6.5, we obtain:

Proposition 6.6.

(i). Let A be a nearly triangular matrix consisting of variables or field elements. Then $\det(A)$ *can be computed by a non-commutative formula of size $n^{O(\log n)}$ without inverse gates, and also by a polynomial size ABP.*

(ii). Let f be a polynomial in n variables which is computed by a) a formula of size s without inverse gates, or b) an ABP of size s. Then $f = \det(A)$, where A is a $k \times k$-nearly triangular matrix whose entries are variables or field elements where a) $k = 2s$, or b) $k = O(ns^3)$.

Proof. To prove (*i*), extend A to an upper triangular matrix

$$B = \begin{pmatrix} u^t & A \\ 0 & v \end{pmatrix} \quad \text{with} \quad u = (1, 0, \dots, 0)\,, v = (0, \dots, 0, 1)\,.$$

Let g be the entry in the upper right corner of B^{-1}. By Propositions 6.3 and 6.5, g can be computed by a quasipolynomial size formula and a polynomial-size ABP. Commutatively, we would be done, since $\det(B)$ is either 1 or -1 and A is the minor of $B_{n+1,1}$. Hence g is equal—up to a sign—to $\det(A)$.

Non-commutatively, one must check that the inverse of an upper triangular matrix can indeed be expressed in terms of determinants of minors. This we leave as an exercise. (Note that in the definition of the determinant, variables of X are multiplied row by row.)

Similarly, if A is the matrix from Proposition 6.3 or 6.5, such that f is the entry in the upper right corner of A^{-1}, we can argue that f is—up to a sign—equal to the determinant of the minor of $A_{2s,1}$, and this minor is a nearly upper triangular matrix. (The sign can be accounted for by adding an appropriate row and column.) □

A well-known, albeit not well-used, approach to lower-bounds on commutative formula size, is to try to bound the smallest s such that $f = \det(A)$, where A is an $s \times s$ matrix of variables or field elements. The previous proposition shows that we can without loss of generality assume that A is nearly upper triangular. This restricted problem may perhaps be easier to solve. Also, one could hope to prove a lower bound even for the determinant itself: to ask what is the smallest s such that $\det(X) = \det(A)$, where A is a nearly triangular matrix of variables or field elements.

However, the following shows that the modified problem is different only in the non-commutative setting:

Corollary 6.7.

(i). Assume that $\det(X_n) = \det(A)$ *or* $\mathrm{perm}(X_n) = \det(A)$, *where A is a nearly triangular matrix with entries variables or field elements, of dimension $s \times s$. Then $s \geq 2^{\Omega(n)}$.*

(ii). Assuming commutativity of variables, there exists a polynomial-size nearly triangular matrix A of variables or field elements such that $\det(X_n) = \det(A)$.

Proof. *(i)* By [39], both $\det(X_n)$ and $\mathrm{perm}(X_n)$ require ABP of size $2^{\Omega(n)}$, but $\det(A)$ can be computed by a polynomial-size ABP.

(ii) Commutatively, $\det(X_n)$ can be computed by a branching program of a polynomial size, and use part *(ii)* of Proposition 6.6. □

7 The rational identity testing problem

We will now address the following basic question: how can we decide whether two rational expressions define the same rational function? This is equivalent to testing whether a single rational expression defines the zero function. This problem can take several forms, and we will focus on deciding whether a *formula* computes the zero function, and refer to this question as the *rational identity testing* problem. As we shall see, the complexity of this problem will depend on a natural problem in (commutative) invariant theory of a simple linear-algebraic flavor, which will appear again in the next section on the elimination of divisions.

In the commutative setting, the rational identity testing problem can be reduced to the well-known *polynomial identity testing* problem, which can be solved quite efficiently by a polynomial time randomized algorithm, by means of the Schwarz-Zippel Lemma. The reduction is possible due to the fact that a commutative rational function can be written as a ratio of two polynomials.

Given the complex structure of rational expressions, it is not even clear that the rational identity testing problem is decidable. This was shown by Cohn and Reutenauer in [10]. The algorithm eventually requires deciding whether a set of (commutative) polynomial equations has a solution, which puts this problem in PSPACE. We will outline a probabilistic algorithm whose efficiency depends on an extra parameter w arising from the above mentioned invariant-theory problem, and assuming the bound is polynomial in s will yield a BPP algorithm for the problem.

The parameter is defined as follows:

- $w(s)$ is the smallest k so that for every correct formula $\Phi(x_1,\dots,x_n)$ of size s there exists $p \le k$ and $a_1,\dots,a_n \in R := \mathrm{Mat}_{p\times p}(\bar{\mathbb{F}})$ such that $\widehat{\Phi}^R(a_1,\dots,a_n)$ is defined.[7]

We will sketch a randomized algorithm for rational identity testing which runs in time polynomial in s and $w(s)$. We will also consider a different version of the parameter. Recall that a linear matrix is of the form

$$A = Q_0 + x_1Q_1 + \dots + x_nQ_n \quad \text{with} \quad Q_0,\dots,Q_n \in \mathrm{Mat}_{s\times s}(\mathbb{F})\,. \tag{7.1}$$

The second parameter is defined as:

- $\tilde{w}(s)$ is the smallest k so that for every $s \times s$ matrix A of the form (7.1) with $Q_0 = 0$, if A is invertible in $\mathbb{F}\langle\!\!\langle x_1,\dots,x_n\rangle\!\!\rangle$ then there exists $p \le k$ and $a_1,\dots,a_n \in \mathrm{Mat}_{p\times p}(\bar{\mathbb{F}})$ such that $Q_1 \otimes a_1 + \cdots Q_n \otimes a_n$ is invertible in $\mathrm{Mat}_{sp\times sp}(\bar{\mathbb{F}})$.

That both $w(s)$ and $\tilde{w}(s)$ are finite essentially follows from Theorem 2.1 part (i). We will prove this in Section 7.1 where the two parameters are further discussed. Note that the absence of the constant term Q_0 in the definition of $\tilde{w}$ is just cosmetic and, without loss of generality, one can assume $n \le s^2$ (cf. Claims 7.5 and 7.8).

Let us first observe that the rational identity problem is essentially equivalent to the following problem: decide whether a formula Φ is a correct formula. For, in order to see whether Φ is correct, we must only check that for every inverse gate u^{-1} in Φ, Φ_u doesn't compute the zero function. Conversely, a correct formula Φ computes the zero function if and only if Φ^{-1} is not a correct formula. Using the construction in Theorem 2.6, we can give the following criterion for the correctness of a formula:

Proposition 7.1. *Let $\mathbb{F}$ be a field and R a ring whose centre contains $\mathbb{F}$. For a formula Φ and $a_1,\dots,a_n \in R$, the following are equivalent:*

(i). $\widehat{\Phi}^R(a_1,\dots,a_n)$ is defined.

(ii). For every gate u in Φ, $A_{\Phi_u}(a_1,\dots,a_n)$ is invertible in Mat(R).

Proof. (i)→(ii) follows from Proposition 6.1. To prove the converse, assume that $\widehat{\Phi}^R(a_1,\dots,a_n)$ is not defined. Then there exists a gate u^{-1} in Φ such that $\widehat{\Phi}_u(a_1,\dots,a_n)$ is defined but $b := \widehat{\Phi}_u(a_1,\dots,a_n)$ is not invertible in R. Let $A := A_{\Phi_u}(a_1,\dots,a_n)$. From Proposition 6.1, we know that A is invertible and $b = \mathcal{R}A^{-1}$, where we invoke the notation from the proof Theorem 2.6. It is sufficient to show that $B := A_{\Phi_{u^{-1}}}(a_1,\dots,a_n)$ is not invertible.

[7] $\bar{\mathbb{F}}$ is the algebraic closure of $\mathbb{F}$.

From the construction of $A_{\Phi_{u^{-1}}}$, we have

$$B = \begin{pmatrix} v_p^t & A \\ 0 & -u_p \end{pmatrix},$$

and $b = u_p A^{-1} v_p^t$. Assume that A is a $k \times k$ matrix. Then there exist invertible $(k+1)\times(k+1)$ matrices C, D such that

$$B = C \begin{pmatrix} b & 0 \\ 0 & I_k \end{pmatrix} D.$$

This can be seen from

$$\begin{pmatrix} v_p^t & A \\ 0 & -u_p \end{pmatrix} = \begin{pmatrix} 0 & I_k \\ 1 & -u_p A^{-1} \end{pmatrix} \cdot \begin{pmatrix} u_p A^{-1} v_p^t & 0 \\ 0 & I_k \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ v_p^t & A \end{pmatrix}.$$

Hence, if B is invertible then so is

$$\begin{pmatrix} b & 0 \\ 0 & I_k \end{pmatrix}$$

and b must be invertible. Therefore, B is not invertible. □

Specializing to the case $R = \mathbb{F}\langle\!\langle x_1, \ldots, x_n \rangle\!\rangle$, we obtain:

Corollary 7.2. *Φ is a correct formula iff for every gate u in Φ, A_{Φ_u} is invertible.*

Hence, the problem of deciding whether Φ is a correct formula can be reduced to the problem of deciding whether a matrix A, whose entries are degree one polynomials, is invertible. The algorithm in [10] in fact solves this latter problem. The essence of the algorithm is Theorem 2.2. By the second part of the theorem, A of dimension $m \times m$ is invertible iff it cannot be written as $B \cdot C$ where B, C consist of degree-one polynomials and have dimensions $m \times k$ and $k \times m$, respectively, with $k < m$. Taking the coefficients of the polynomials in B, C as unknowns, the problem is expressible in terms of solvability of a set of (commutative) polynomial equations over $\mathbb{F}$.

However, we will use a different invertibility test:

Proposition 7.3. *Let A be as in (7.1). Then A is invertible in $Mat_{s\times s}(\mathbb{F}\langle\!\langle \bar{x} \rangle\!\rangle)$ iff there exists $k \in \mathbb{N}$ and $a_1, \ldots, a_n \in Mat_{k\times k}(\mathbb{F})$ such that*

$$Q_0 \otimes I_k + Q_1 \otimes a_1 + \cdots + Q_n \otimes a_n$$

is invertible in $Mat_{sk\times sk}(\mathbb{F})$.

Proof. If A is invertible, there is $A^{-1} \in \mathrm{Mat}_{s\times s}(\mathbb{F}\langle\!\langle \bar{x} \rangle\!\rangle)$ with $A \cdot A^{-1} = A^{-1} \cdot A = I_s$. The entries of A^{-1} are rational functions. By Theorem 2.1 part (*i*), there exist $k \in \mathbb{N}$ and $a_1, \ldots, a_n \in R := \mathrm{Mat}_{k\times k}(\bar{\mathbb{F}})$ such that every $A^{-1}_{i,j}$ is defined on $a_1, \ldots, a_n$. (To obtain $a_1, \ldots, a_n$ which work for every i, j, it is enough to consider a circuit in which every $A^{-1}_{i,j}$ is computed by some gate.) Evaluating A^{-1} at this point gives a matrix $B \in \mathrm{Mat}_{s\times s}(R)$ such that

$$A(a_1, \ldots, a_n) \cdot B = B \cdot A(a_1, \ldots, a_n) = I_s,$$

and so $A(a_1,\dots,a_n)$ is invertible in $\mathrm{Mat}_{s\times s}(R)$. It is an $s\times s$-matrix with $k\times k$-matrices as entries, and it can be identified with the $sk\times sk$-matrix

$$A(a_1,\dots,a_n)' = Q_0\otimes I_k + \sum_{i=1}^{n} Q_i\otimes a_i\,.$$

Clearly, $A(a_1,\dots,a_n)$ is invertible iff $A(a_1,\dots,a_n)'$ is invertible. Hence $A(a_1,\dots,a_n)'$ is invertible.

If A is not invertible then it is not full by Theorem 2.2. Hence $A(a_1,\dots,a_n)$ is not full in R for every $R=\mathrm{Mat}_{k\times k}(\mathbb{F})$ and every $a_1,\dots,a_n$. Hence $A(a_1,\dots,a_n)$ and $A(a_1,\dots,a_n)'$ are never invertible. □

The quantity $\tilde{w}$ is utilized as follows. Let A be an $s\times s$ matrix as in (7.1), and k a parameter. Introduce nk^2 distinct *commutative* variables $y^i_{p,q}, i\in\{1,\dots,n\}, p,q\in\{1,\dots,k\}$. For every variable x_i, consider the $k\times k$ matrix

$$x_i^\star := (y^i_{p,q})_{p,q\in\{1,\dots,k\}}\,,$$

and let

$$A^{(k)} := Q_0\otimes I_k + Q_1\otimes x_1^\star + \cdots + Q_n\otimes x_n^\star\,.$$

It is an $sk\times sk$ matrix whose entries are commutative (linear) polynomials in the auxiliary variables Y.

Then we obtain the following invertibility test:

Proposition 7.4. *Let A be as (7.1). Then A is invertible iff there exists $k\le\tilde{w}(s)$ such that* $\det(A^{(k)})\neq 0$ *(as a polynomial in $\mathbb{F}[Y]$).*

Proof. We first note that the assumption $Q_0=0$ in the definition of $\tilde{w}$ is inessential:

Claim 7.5. *A is invertible iff there exists $k\le\tilde{w}(s)$ and $a_1,\dots a_n\in Mat_{k\times k}(\bar{\mathbb{F}})$ such that*

$$Q_0\otimes I_k + \sum_{i=1}^{n} Q_i\otimes a_i$$

is invertible.

Proof of Claim 7.5. If $Q_0\otimes I_k+\sum_{i=1}^n Q_i\otimes a_i$ is invertible then A is invertible by Proposition 7.3. Conversely, if A is invertible then so is $x_0Q_0+\sum_{i=1}^n x_iQ_i$, where x_0 is a fresh variable. Hence there exist $k\le\tilde{w}(s)$ and $a_0,\dots,a_n\in\mathrm{Mat}_{k\times k}(\bar{\mathbb{F}})$ such that $Q_0\otimes a_0+\sum_{i=1}^n Q_i\otimes a_i$ is invertible. Since $\bar{\mathbb{F}}$ is infinite, we can assume that a_0 is invertible and hence $Q_0\otimes I_k+\sum_{i=1}^n Q_i\otimes(a_ia_0^{-1})$ is invertible. □

This shows that A is invertible (over the free skew field) iff there exists $k\le\tilde{w}(s)$ such that $A^{(k)}$ is invertible (over the field of fractions $\mathbb{F}(Y)$). The matrix $A^{(k)}$ has entries from the commutative ring $\mathbb{F}[Y]\subseteq\mathbb{F}(Y)$ and so it is invertible iff $\det(A^{(k)})\neq 0$. □

Setting $m:=s\tilde{w}(s)$, each $A^{(k)}$ has dimension at most $m\times m$, and each entry is a linear polynomial in the commuting variables Y. Hence $\det(A^{(k)})$ is a polynomial of degree at most m, and it can be computed by a commutative arithmetic circuit of size polynomial in m. This allows to test invertibility of A by a randomized algorithm running in time polynomial in s and $\tilde{w}(s)$. We note that in [29], a *deterministic* algorithm was designed which runs in time polynomial in s and a parameter similar to $\tilde{w}$.

By means of Proposition 7.1, this also gives an algorithm for rational identity testing whose complexity depends on $\tilde{w}$. However, such an algorithm can also be obtained using the parameter w instead. Proposition 7.1 gives:

Proposition 7.6. *For a formula Φ of size s, the following are equivalent:*

(i). Φ is a correct formula.

(ii). There exists $k \leq w(s)$ such that for every gate u in Φ, $\det(A^{(k)}_{\Phi_u}) \neq 0$.

By Theorem 2.6, each of the matrices $A^{(k)}_{\Phi_u}$ has size at most $2sw(s) \times 2sw(s)$ and each entry is again a linear polynomial in the commuting variables Y. This allows to reduce the non-commutative rational identity testing problem to s instances of the commutative polynomial identity testing problem.

Non-commutative *polynomial* identity testing

Let us add some comments about the special case of rational identity testing: decide whether a formula or a circuit *without* inverse gates computes the zero polynomial. In [41], Raz and Shpilka show that the problem can be decided by a polynomial time deterministic algorithm for a formula or even an arithmetic branching program. For a non-commutative circuit, no such deterministic algorithm is known. In [5], Bogdanov and Wee show it can be decided by a polynomial time randomized algorithm.

The main point is given by the celebrated Amitsur-Levitzki theorem ([3], discussed in a greater detail in Section 9): if a polynomial f vanishes on all $k \times k$ matrices over an infinite field then f must have degree at least $2k$. Hence, assume that we are given a division free circuit Φ computing a polynomial $f(x_1, \dots, x_n)$ of degree $< 2k$. To check whether $f(x_1, \dots, x_n) = 0$, it is now sufficient to check whether $f(a_1, \dots, a_n) = 0$ for all $k \times k$ matrices $a_1, \dots, a_n$, over an infinite field. As above, we can interpret $f(a_1, \dots, a_n)$ as a matrix of k^2 polynomials in nk^2 commuting variables, each of degree $< 2k$, and hence reduce the non-commutative identity testing problem to k^2 instances of the commutative identity testing. This yields a polynomial time randomized algorithm, whenever we focus on a class of circuits computing polynomials of a polynomial degree.

7.1 The two parameters

We now give more details about the parameters w and $\tilde{w}$. We start by showing that they are indeed well-defined, and that $\tilde{w}$ can be used to upper-bound w.

Proposition 7.7. *Both w and $\tilde{w}$ are well-defined non-decreasing functions and $w(s) \leq \tilde{w}(s^2 + s)$.*

Proof. We start by proving that $\tilde{w}(s)$ is well-defined (finite) for every s. To see this, we first note the following:

Claim 7.8. *Assume that $\sum_{i=1}^{n} x_i Q_i$ is invertible, where $Q_1, \dots, Q_n \in Mat_{s\times s}(\mathbb{F})$. Then there exists $M \subseteq \{1, \dots, n\}$ with $|M| \leq s^2$, such that $\sum_{i \in M} x_i Q_i$ is invertible.*

Proof of Claim 7.8. Assume that Q_n lies in the linear span of $Q_1,\ldots,Q_{n-1}$, i. e., $Q_n = \sum_{i=1}^{n-1} c_i Q_i$ for some constants $c_1,\ldots c_{n-1}$ from $\mathbb{F}$. Proposition 7.3 gives that $\sum_{i=1}^{n} Q_i \otimes a_i$ is invertible in $\mathrm{Mat}_{ks\times ks}(\mathbb{F})$ for some $a_1,\ldots,a_n \in \mathrm{Mat}_{k\times k}(\mathbb{F})$. We have

$$\sum_{i=1}^{n} Q_i \otimes a_i = \sum_{i=1}^{n-1} Q_i \otimes a_i + \left(\sum_{i=1}^{n-1} c_i Q_i\right) \otimes a_n = \sum_{i=1}^{n-1} Q_i \otimes (a_i + c_i a_n)\,.$$

Hence, by the same proposition, $\sum_{i=1}^{n-1} x_i Q_i$ is invertible in $\mathbb{F}\langle\!(x_1,\ldots,x_{n-1})\!\rangle$. To conclude, it is enough to pick M so that $\{Q_i : i \in M\}$ is a basis of the linear span of $Q_1,\ldots,Q_n$. □

The claim shows that in the definition of $\tilde{w}(s)$ we can assume that $n = s^2$, which we now proceed to do. Let T be the set of non-invertible linear matrices of the form $A = \sum_{i=1}^{s^2} x_i Q_i$ with $Q_1,\ldots,Q_{s^2} \in \mathrm{Mat}_{s\times s}(\mathbb{F})$. Proposition 7.3 asserts that $A \in T$ iff

$$\det\left(\sum_{i=1}^{s^2} Q_i \otimes a_i\right) = 0$$

for every $k \in \mathbb{N}$ and $a_1,\ldots,a_{s^2} \in \mathrm{Mat}_{k\times k}(\mathbb{F})$. We can view A as the s^2-tuple of the $s\times s$ matrices $Q_1,\ldots,Q_{s^2}$ and T as a subset of $\mathbb{F}^{s^4}$. Then we see that T is an algebraic set defined by the equations

$$\det\left(\sum_{i=1}^{s^2} Q_i \otimes a_i\right) = 0\,,$$

for all possible a_i's. By Hilbert's basis theorem, a finite subset of the equations is enough to define T. Switching to the complement of T, this means that there exists $k \in \mathbb{N}$ such that A is invertible iff there exists $p \le k$ and $a_1,\ldots,a_{s^2} \in \mathrm{Mat}_{p\times p}(\mathbb{F})$ such that $\sum_{i=1}^{s^2} Q_i \otimes a_i$ is invertible—and $\tilde{w}(s)$ is the smallest such k.

We now prove the inequality $w(s) \le \tilde{w}(s^2+s)$. Let Φ be a correct formula of size s in variables $x_1,\ldots,x_n$. For a gate v in Φ, let s_v be the size of Φ_v and let A_{Φ_v} be of dimension $k_v \times k_v$. Let $k := \sum_v k_v$. It is easy to see that $\sum_v s_v \le (s^2+s)/2$. Since $k_v \le 2s_v$ (Theorem 2.6), we have that $k \le s^2+s$. Let A be a $k\times k$-matrix which is block-diagonal with the blocks being the matrices A_{Φ_v}, for all gates v. Then A is invertible and of the form $Q_0 + \sum_{i=1}^{n} x_i Q_i$ with $Q_0,\ldots,Q_n \in \mathrm{Mat}_{k\times k}(\mathbb{F})$. By Claim 7.5, there exist $p \le \tilde{w}(k)$ and $a_1,\ldots,a_n \in R := \mathrm{Mat}_{p\times p}(\overline{\mathbb{F}})$ so that $Q_0 \otimes I_k + \sum_{i=1}^{n} Q_i \otimes a_i$ is invertible. That is, $A(a_1,\ldots,a_n)$ is invertible in $\mathrm{Mat}_{k\times k}(R)$—and hence all the blocks A_{Φ_v} are. This shows that $\Phi^R(a_1,\ldots,a_n)$ is defined by Proposition 7.1 and so $w(s) \le \tilde{w}(k) = \tilde{w}(s^2+s)$.

This entails that w is well-defined. $\tilde{w}$ is non-decreasing because an invertible matrix A can be enlarged to a block-diagonal $(s+1)\times(s+1)$ invertible matrix A', the blocks being A and x_1. w is non-decreasing because a correct formula Φ of size s can be modified to a correct formula of size $s+1$ which contains all the inverse gates of Φ. (Hint: Φ^{-1} works whenever $\widehat{\Phi} \neq 0$.) □

In [25], $\tilde{w}$ was originally defined with the additional constraint $n \le s$ on A in (7.1). Here, we have removed this assumption: Claim 7.8 shows that we can always assume $n \le s^2$ and hence it is unnecessary to have an explicit constraint on n.

Amitsur-Levitzki theorem (cf. Section 9) shows that both $w(s)$ and $\tilde{w}(s)$ grow with s. To see that $w(s)$ exceeds 1 for some s, it is enough to take the formula $(xy-yx)^{-1}$. Similarly, the following matrix (borrowed from [19]) shows that $\tilde{w}(3) \geq 2$:

$$\begin{pmatrix} 0 & x & y \\ -x & 0 & z \\ -y & -z & 0 \end{pmatrix}. \tag{7.2}$$

Proposition 7.6 shows that in order to obtain an efficient rational identity testing algorithm, it is enough to show that $w(s)$ is small. The other parameter $\tilde{w}$ is designed to solve the more general problem of deciding whether a linear matrix is invertible. We do not have a reason to believe that the inequality in Proposition 7.7 is even close to being tight. Hence, estimating w in terms of $\tilde{w}$ can turn out to be far too generous. However, the main appeal of $\tilde{w}$ is the simplicity of its definition. By means of Proposition 7.3, $\tilde{w}$ can be introduced without any reference to computations or even the skew field: in the definition of $\tilde{w}$, the assumption "A is invertible over $\mathbb{F}\langle\!\langle \bar{x} \rangle\!\rangle$" can be replaced by the assumption "$\sum_{i=1}^{n} Q_i \otimes a_i$ is invertible for some $m \in \mathbb{N}$ and $a_1, \ldots, a_n \in \mathrm{Mat}_{m\times m}(\mathbb{F})$." We also note that the suspected gap between w and $\tilde{w}$ disappears if in the definition of $\tilde{w}$, we consider only matrices A which come from a representation of some formula Φ as in Theorem 2.6—i. e., $A = A_\Phi$ for some Φ.

8 How to eliminate divisions

A classical result of Strassen [48], see also [7] Chapter 7.1, asserts that division gates are not very helpful when computing commutative polynomials. If a commutative polynomial f of degree k can be computed by a circuit with divisions of size s then it can be computed by a circuit without divisions of size $O(sk^2)$. (The original argument assumes that the underlying field is infinite. It was noted in [28] that a similar statement holds over any field.) However, when dealing with non-commutative computations, the situation seems to be much more complicated. To outline the main issue, assume that we have computed a polynomial f using only one inverse g^{-1} for a polynomial g. Commutatively, if $g \neq 0$ and $\mathbb{F}$ is infinite, there must exist $a \in \mathbb{F}^n$ such that $g(a) \neq 0$. We can then rewrite g^{-1} as a power series around the point a. Supposing f has degree k, it is enough to truncate the series up to terms of degree k, obtaining a computation of f without divisions. Non-commutatively, no such substitution from $\mathbb{F}$ may exist. For example, $g = xy-yx$ is a non-zero polynomial which vanishes on every substitution from $\mathbb{F}$. An obvious remedy is to allow substitutions which are $m \times m$ matrices over $\mathbb{F}$, and then to work with power series in this matrix algebra. In the example $xy-yx$, it is easy to find 2×2 matrices a, b such that $ab-ba$ is invertible, and the power series argument yields a satisfactory answer. However, as g gets more complicated, we must substitute $m \times m$ matrices with a larger m. Computations with $m \times m$ matrices have cost roughly m^3 and in order to eliminate divisions efficiently, we want m to have polynomial size. The question now becomes: what is the dimension of matrices we need to make g invertible? We do not know how to answer this question and state it as Problem 9.3 in Section 9. As opposed to the running example, we cannot in general assume that g is a polynomial: it may itself contain inverses, and nested inverses. This makes the bound on m quite challenging; see Section 9 for further discussion. In Proposition 8.3, we will prove only a conditional result: if a polynomial f is computed by a circuit with divisions Φ such that

$\widehat{\Phi}$ is defined for some matrices of small size, then f can be computed by a small circuit without division gates.

We now proceed towards proving Proposition 8.3. This requires introducing some definitions as well as extending definitions from Section 2. Let R be an arbitrary (unital) ring and let $\bar{x} = \{x_1, \dots, x_n\}$ a set of variables. $R\langle\bar{x}\rangle$ will denote the set of polynomials in variables $\bar{x}$ with coefficients from R. The ring R is not in general commutative, and the variables do not multiplicatively commute, but we will assume that variables commute with elements of R. More exactly, a polynomial in $R\langle\bar{x}\rangle$ is a sum

$$\sum_{\alpha} c_\alpha \alpha \,,$$

where α ranges over products of variables from $\bar{x}$ and $c_\alpha \in R$, with only finitely many c_α's being non-zero. Addition and multiplication are given by

$$\begin{aligned} \sum_{\alpha} c_\alpha \alpha + \sum_{\alpha} c'_\alpha \alpha &= \sum_{\alpha} (c_\alpha + c'_\alpha)\alpha \,, \\ \sum_{\alpha} c_\alpha \alpha \cdot \sum_{\alpha} c'_\alpha \alpha &= \sum_{\alpha,\beta} (c_\alpha c'_\beta)(\alpha\beta) \,. \end{aligned}$$

We will extend $R\langle\bar{x}\rangle$ to the ring of power-series[8] $R\{\bar{x}\}$. A power series $f \in R\{x\}$ is an infinite sum

$$f^{(0)} + f^{(1)} + f^{(2)} + \cdots \,,$$

where every $f^{(k)} \in R\langle\bar{x}\rangle$ is a homogeneous polynomial of degree k. Addition and multiplication are defined in the obvious way:

$$(f+g)^{(k)} = f^{(k)} + g^{(k)} \,, \qquad (f \cdot g)^{(k)} = \sum_{i+j=k} f^{(i)} \cdot g^{(j)} \,.$$

If $f^{(0)} = a \in R$ is invertible in R then f is invertible in $R\{\bar{x}\}$. Its inverse is given by

$$f^{-1} \;=\; (a-(a-f))^{-1} \;=\; a^{-1}(1-(1-fa^{-1}))^{-1} \;=\; a^{-1}\sum_{i=0}^{\infty}(1-fa^{-1})^i \,.$$

If we denote the partial sum $f^{(0)} + f^{(1)} + \cdots + f^{(k)}$ by $f^{(\le k)}$ this can be written as

$$(f^{-1})^{(\le k)} = a^{-1}\left(\sum_{i=0}^{k}(1-f^{(\le k)}a^{-1})^i\right)^{(\le k)} \,. \tag{8.1}$$

We also need to extend the definition of a circuit so that it can compute power series over R: this is achieved by taking circuits as introduced in Section 2, but allowing them to use elements of R in the computation (in the same way elements of the fields were). Such a circuit will be called an *R-circuit*. All other notions are generalized in an obvious manner. Mainly, an R-circuit computes an element of $R\{\bar{x}\}$: evaluate the circuit gate-by-gate over $R\{\bar{x}\}$. This either produces an element of $R\{\bar{x}\}$, or the evaluation fails due to attempted inversion of a non-invertible elements in $R\{\bar{x}\}$.

[8] This ring is sometimes denoted $R\langle\langle\bar{x}\rangle\rangle$.

Lemma 8.1. *Assume that $f \in R\{\bar{x}\}$ can be computed by an R-circuit of size s and depth d. Then for every $k \in \mathbb{N}$, $f^{(\leq k)} \in R\langle\bar{x}\rangle$ can be computed by an R-circuit without division gates of size $O(sk^3)$ and depth $O(d\log^2 k)$.*

Proof. We divide the proof into two steps.

Step 1. Let Φ be the circuit computing f, and let us fix k. We will construct a division-free circuit Φ' of size $O(sk)$ and depth $O(d\log k)$ computing a polynomial $g \in R\langle\bar{x}\rangle$ such that $f^{(\leq k)} = g^{(\leq k)}$. Let $\lambda(z)$ be the univariate polynomial $\sum_{i=0}^{k}(1-z)^i$. Clearly, it can be computed by a division free circuit $\Lambda(z)$ of size $O(k)$ and depth $O(\log k)$ (actually, even of size $O(\log k)$). For any inverse gate u^{-1} in Φ, u computes an invertible element $\widehat{u} \in R\{\bar{x}\}$ and so $a := \widehat{u}^{(0)} \in R$ is invertible. Then Φ' is obtained by simultaneously replacing every inverse gate u^{-1} by the circuit $a^{-1}\Lambda(ua^{-1})$ and appropriately rewiring the inputs and outputs.

Since $a^{-1} \in R$, Φ' is an R-circuit without divisions, and it computes a polynomial $g \in R\langle\bar{x}\rangle$. That g satisfies $g^{(\leq k)} = f^{(\leq k)}$ is easily proved by induction on the size s. Note that for every f_1, f_2,

$$(f_1 \circ f_2)^{(\leq k)} = (f_1^{(\leq k)} \circ f_2^{(\leq k)})^{(\leq k)},$$

where $\circ \in \{+,\cdot\}$, and (8.1) gives $(f_1^{-1})^{(\leq k)} = a^{-1}(\lambda(f_1^{(\leq k)}a^{-1}))^{(\leq k)}$, where $a = f_1^{(0)}$.

Step 2. Given a division-free Φ' of size s' and depth d' computing a polynomial g, $g^{(\leq k)}$ can be computed by a division-free circuit of size $O(s'k^2)$ and depth $O(d'\log k)$. This is a standard homogenization argument.

Altogether we have obtained a circuit of size $O(sk^3)$ and depth $O(d\log^2 k)$ computing $f^{(\leq k)}$. □

Next, we consider the ring of $m \times m$ matrices

$$R := \mathrm{Mat}_{m\times m}(\mathbb{F}).$$

We want to interpret a polynomial $f \in R\langle\bar{x}\rangle$ as an $m \times m$ matrix $f^\star$ whose entries are polynomials in $\mathbb{F}\langle\bar{x}\rangle$. Let $f \in R\langle\bar{x}\rangle$ be written as $f = \sum_\alpha c_\alpha \alpha$, where $c_\alpha \in R$ and α ranges over products of the variables $\bar{x}$. Then $f^\star \in \mathrm{Mat}_{m\times m}(\mathbb{F}\langle\bar{x}\rangle)$ is the matrix with

$$f^\star_{i,j} = \sum_\alpha (c_\alpha)_{i,j}\alpha\,,\ i,j \in \{1,\dots,m\}.$$

Lemma 8.2. *Assume that $f \in R\langle\bar{x}\rangle$ can be computed by an R-circuit without divisions of size s and depth d. Then $f^\star$ can be computed by a circuit over $\mathbb{F}$ without divisions of size $O(sm^3)$ and depth $O(d\log m)$.*

Proof. Note that $(f_1+f_2)^\star = f_1^\star + f_2^\star$ and $(f_1 \cdot f_2)^\star = f_1^\star \cdot f_2^\star$, where on the right hand side we see a sum, resp. product, of $m \times m$ matrices. This means that a sum and a product gate in an R-circuit can be simulated by a sum, resp. product, of $m \times m$ matrices over $\mathbb{F}$. This gives an increase in size of factor at most $O(m^3)$ and depth by a factor of $O(\log m)$. □

Proposition 8.3. *Let $f \in \mathbb{F}\langle\bar{x}\rangle$ be a polynomial of degree k which is computable by a circuit Φ of size s and depth d. Assume that there exist matrices $a_1,\dots,a_n \in R = Mat_{m\times m}(\mathbb{F})$ such that $\widehat{\Phi}^R(a_1,\dots,a_n)$ is defined. Then f can be computed by a circuit without divisions of size $O(sk^3m^3)$ and depth $O(d\log^2 k\log m)$.*

Proof. $\mathbb{F}$ can be embedded into R via the map $a \in \mathbb{F} \to aI_m \in R$. Similarly, each variable x_i is mapped to x_iI_m. So we can view f as en element of $R\langle\bar{x}\rangle$ and Φ as an R-circuit. Consider $g := f(x_1+a_1,\dots,x_n+a_n)$ and the R-circuit $\Phi' := \Phi(x_1+a_1,\dots,x_n+a_n)$. The assumption that $\widehat{\Phi}^R(a_1,\dots,a_n)$ is defined guarantees that Φ' is a correct R-circuit computing $g \in R\langle\bar{x}\rangle \subseteq R\{\bar{x}\}$. By Lemma 8.1, $g^{(\le k)}$ can be computed by a division-free R-circuit Φ'' of size $O(sk^3)$ and depth $O(d\log^2 k)$. Since g has degree k, we have $g^{(\le k)} = g$ and Ψ'' computes g. Moreover, $f = g(x_1-a_1,\dots,x_n-a_n)$ and so $\Phi''(x_1-a_1,\dots,x_n-a_n)$ computes f. Lemma 8.2 gives that $f^\star$ can be computed by a division-free circuit over $\mathbb{F}$ of size $O(sk^3m^3)$ and depth $O(d\log^2 k\log m)$. But $f^\star$ is simply the diagonal matrix with f on the diagonal and the statement of the proposition follows. □

Invoking Proposition 4.1 gives:

Corollary 8.4. *Let $f \in \mathbb{F}\langle\bar{x}\rangle$ be a polynomial of degree k which is computable by a formula Φ of size s. Assume that there exist matrices $a_1,\dots,a_n \in Mat_{m\times m}(\mathbb{F})$ such that $\widehat{\Phi}(a_1,\dots,a_n)$ is defined. Then f can be computed by a formula without divisions of size $s^{O(\log^2 k\log m)}$.*

Recalling the definition of $w(s)$ from Section 7, the corollary implies that

- if $\mathbb{F}$ is algebraically closed then f can be computed by a formula without divisions of size $s^{O(\log^2 k\log w(s))}$.

The bounds presented in Proposition 8.3, and Corollary 8.4, are not intended to be optimal. A more careful calculation would show

- a size upper bound $O(sk^2\log k\cdot m^\omega)$, where $\omega < 3$ is the (hitherto unknown) exponent of matrix multiplication;
- a depth upper bound $O(d\log k\log m)$ assuming that $\mathbb{F}$ is infinite (owing to savings in Step 2 in the proof of Lemma 8.1).

9 Open problems

Problem 9.1. Give an explicit polynomial $f \in \mathbb{F}\langle\bar{x}\rangle$ which cannot be computed by a polynomial size formula *with divisions*.

Nisan's result [39] gives the solution for formulae without division gates. An obvious approach to Problem 9.1 is to show that division gates can be eliminated without increasing the formula size too much. This leads to the following question:

Problem 9.2. Assume that a polynomial $f \in \mathbb{F}\langle\bar{x}\rangle$ of degree k can be computed by a circuit Φ with divisions of size s. Give a non-trivial upper bound on the size of a circuit without divisions computing f. Similarly for some other complexity measure of f, such as formula size.

A conditional answer to Problem 9.2 was given in Proposition 8.3. There, we constructed a circuit without divisions computing f under the assumption that there exist matrices of small dimension for which the original circuit Φ is defined. (That is, matrices in $R = \mathrm{Mat}_{m\times m}(\mathbb{F})$ such that, when we evaluate Φ in R, we never come across an inverse gate computing a non-invertible matrix.) Hence:

Problem 9.3. Assume that $\Phi(x_1,\dots,x_n)$ is a correct circuit of size s. What is the smallest m so that there exist $a_1,\dots,a_n \in R = \mathrm{Mat}_{m\times m}(\mathbb{F})$ for which $\Phi^R(a_1,\dots,a_n)$ is defined? Similarly, for some other complexity measure, such as formula size.

As was explained in Section 7, the question is also relevant in the rational identity testing problem: to decide whether a formula computes the zero rational function. There, we also mentioned a related version of the question:

Problem 9.4. Find an upper bound on the smallest $k = k(s)$, such that for all $Q_1,\dots,Q_n \in \mathrm{Mat}_{s\times s}(\mathbb{F})$ if $\sum_{i=1}^n x_iQ_i$ is invertible in $\mathbb{F}\langle\!\langle x_1,\dots,x_n\rangle\!\rangle$ then $\sum_{i=1}^n Q_i \otimes a_i$ is invertible for some $a_1,a_2,\dots,a_n \in \mathrm{Mat}_{p\times p}(\mathbb{F})$ with $p \le k$.

Problems 9.3 and 9.4 are interesting from a purely algebraic perspective. The connection between them is explained in Section 7.1. Here, let us give a few comments about Problem 9.3. First, such an m always exists, and can be bounded by a function of s. This follows from Theorem 2.1 (cf. Proposition 7.1). Second, let us recall the celebrated Amitsur-Levitzki theorem [3]: for every p there exists a non-zero polynomial $f_p \in \mathbb{F}\langle x_1,\dots,x_{2p}\rangle$ of degree $2p$ such that f_p vanishes on all $p\times p$ matrices. Conversely, every non-zero polynomial vanishing on all $p\times p$ matrices over an infinite field must have degree at least $2p$. The converse can be strengthened to show that if $0 \ne f \in \mathbb{F}\langle x_1,\dots,x_n\rangle$ has degree $< 2p$, there exist $p\times p$ matrices $a_1,\dots,a_n$ such that the matrix $f(a_1,\dots,a_n)$ is invertible—indeed most tuples $(a_1,\dots,a_n)$ will satisfy this property. This follows from another theorem of Amitsur (see [44] Theorem 3.2.6 and Exercise 2.4.2, as well as [34], Proposition 2.4). To apply this to Problem 9.3, suppose that the circuit Φ in Problem 9.3 contains a gate computing f_p^{-1}, where f_p is the Amitsur-Levitzki polynomial of degree $2p$. Then m must be at least $p+1$, which shows that m grows with s. On the other hand, assume that Φ contains only one inverse gate computing g^{-1}, for some polynomial g of degree k. Then m can be taken $\le k/2+1$. A similar bound can be obtained for any Φ of inverse height one. However, we do not know how to compose this argument, and what happens for circuits of general height—even the case of circuits of height two is far from clear. As far as we know, the strongest upper-bound known in Problem 9.4 is $k(s) = 2^{O(s\log s)}$. This was obtained in a recent paper [29] by Ivanyos, Qiao and Subrahmanyam, improving an earlier result of Derksen [12]. These results involve tools from invariant theory, and we explore this general connection in the Appendix below.

A Appendix: The Invariant Theory angle

In this section, we give the necessary background from Invariant Theory to explain how our main open problem (Problem 9.4) arises there. We also mention a few other computational complexity questions which have recently been cast in this language. We will present only, in high level, the fragment of the theory which is relevant to us. Often, the stated results are known in greater generality. One can find much more in the books [11, 13, 33]. We stress that in this section all variables *commute*!

Fix a field $\mathbb{F}$ (while problems are interesting in every field, results mostly work for infinite fields only, and sometimes just for characteristic zero or algebraically closed ones). Let G be a group, and V a representation of G, namely a vector space on which G acts; for every $g,h \in G$ and $v \in V$ we have $gv \in V$ and $g(hv) = (gh)v$. When G acts on V, it also acts on $\mathbb{F}[V]$, the polynomial functions on V, also called the *coordinate ring* of V. We will denote gp the action of a group element g on a polynomial $p \in \mathbb{F}[X]$. In our setting V will have finite dimension (say m), and so $\mathbb{F}[V]$ is simply $\mathbb{F}[x_1, x_2, \ldots, x_m] = \mathbb{F}[X]$, the polynomial ring over $\mathbb{F}$ in m variables.

A polynomial $p(X) \in \mathbb{F}[X]$ is *invariant* if it is unchanged by this action, namely for every $g \in G$ we have $gp = p$. All invariant polynomials clearly form a subring of $\mathbb{F}[X]$, denoted $\mathbb{F}[X]^G$, called the ring of invariants of this action. Understanding the invariants of group actions is the main subject of Invariant Theory. In our setting, all these rings will be *finitely generated*, namely there will be a finite set of polynomials $\{q_1, q_2, \ldots q_t\}$ in $\mathbb{F}[X]^G$ so that for every polynomial $p \in \mathbb{F}[X]^G$ there is a t-variate polynomial r over $\mathbb{F}$ so that $p = r(q_1, q_2, \ldots, q_t)$. The following example must be familiar to the reader.

Example A.1. Let $G = S_m$, the symmetric group on m letters, acting on the set of m formal variables X (and hence the vector space they generate) by simply permuting them. Then the set of invariant polynomials are simply all *symmetric polynomials*. As is well known, they are generated in the sense above by the m *elementary* symmetric polynomials (namely $q_1 = \sum_i x_i, q_2 = \sum_{i<j} x_i x_j, \ldots, q_m = \prod_i x_i$). Another generating set of the same size is provided by the *sums-of-powers* (namely $q'_1 = \sum_i x_i, q'_2 = \sum_i x_i^2, \ldots, q'_m = \sum_i x_i^m$).

This example demonstrates a nearly perfect understanding of the ring of invariants. The first basic requirement is a finite set of generators of the ring of invariants. Establishing this for a group action is often called the “First Fundamental Theorem,” or FFT. The second requirement (naturally called the “Second Fundamental Theorem,” or SFT, when established) is describing all algebraic relations between the given generating invariants. In the case of symmetric polynomials above, they are algebraically independent. Hence, we know FFT and SFT in this example.

Further requirements have to do with the explicitness and constructivity of the given invariants, their number, as a function of natural parameters like the dimension of the space V, size (when finite) or “dimension” of the group G. Finally, a more modern request is that the given invariants would be easy to compute. For the action of the symmetric group above, we are in an “optimal” situation. There are exactly m generating invariants (the dimension of V), explicitly given and very easy to compute. Some of these explicitness and computational notions are formally defined and discussed in [17], section 1.2.

This set of “computational” properties is clearly directly related to the efficiency of solving perhaps the most basic *orbit* problem of this setting: given two points $u, v \in V$, are they in the same orbit under the group action? If they are, clearly their evaluation at every invariant is identical (and the converse can be achieved with a somewhat more general notion of “separating invariants”). Many basic problems in many mathematical disciplines can be viewed in this way (e. g., Is a given knot unknotted? Can one turn a polygon into another via (straight) cutting and pasting?). More recently, basic problems of computational complexity were cast in these terms. Valiant [49] showed that to separate the arithmetic classes VP and VNP it suffices to show that the permanent polynomial is not a *linear projection* of a determinant of a not much larger matrix. While projection is not a group operation, the Geometric Complexity Theory (GCT) project of Mulmuley and Sohoni (see, e. g., [37, 36] for surveys) describes it in similar terms, namely the intersection of the orbit *closure* of varieties defined respectively by permanent and determinant. In this last motivation the groups acting are linear groups.

Most work and knowledge in Invariant Theory concerns linear groups.[9] The first seminal results

[9] Here, we consider only actions on vector spaces, the real interest of algebraic geometry is their actions on general affine

came from Hilbert [24], who proved[10] the first and second fundamental theorems for the natural actions of the general and special linear groups, $GL(V)$ and $SL(V)$ on a vector space V. Again, a very familiar very special case, in which knowledge is complete, is the following.

Example A.2. Consider the action of the group $SL_n(\mathbb{F})$ on the vector space of $n \times n$ matrices in $\mathrm{Mat}_{n\times n}(\mathbb{F})$, simply by matrix multiplication. Namely, $A \in SL_n$ acts on M by AM. The entries of such a generic matrix M may be viewed as $m = n^2$ variables $X = \{x_{ij}\}$. In this case all polynomial invariants are generated by the determinant $det(X)$.

We shall be interested in invariants[11] of actions of SL_n on d-tuples of $n \times n$ matrices. So now the number of variables is $m = dn^2$. The most well understood is the action of a single copy of SL_n by simultaneous conjugation of the d matrices, and the one we care about the action of two copies, $SL_n \times SL_n$, by simultaneous multiplication on the left and right. We define both next and discuss what is known, but first point out that these are very special cases of the general setting of *quivers* and their representations [20] and many of the results generalize to this setting.

Now a typical element of our vector space is a d-tuple of $n \times n$ matrices $(M_1, M_2, \dots, M_d)$ and the underlying variables X are now the $m = dn^2$ entries. Consider the action of a matrix $A \in SL_n$ on this tuple by simultaneous conjugation, namely transforming it to the tuple $(A^{-1}M_1A, A^{-1}M_2A, \dots, A^{-1}M_dA)$. Which polynomials in X are invariant under this action? The first and second fundamental theorem were proved by Procesi, Formanek, Razmyslov, and Donkin [40, 18, 42, 16]. More precisely, the invariants are generated by traces of products of length at most n^2 of the given matrices, namely by the set

$$\left\{\mathrm{Tr}(M_{i_1}M_{i_2}\cdots M_{i_t}) : t \le n^2\right\}.$$

These polynomials are explicit, have small degree and are easily computable. The one possible shortcoming, the exponential size of this set (a serious impediment to, e. g., solving the orbit problem above), was recently improved to quasi-polynomial by Forbes and Shpilka [17], who “derandomized” a probabilistic construction of Mulmuley [38]. Furthermore, [17] connects the problem of finding few invariants to solving the Polynomial Identity Testing problem (in the commutative setting).

Finally, we get to the action we care about. Here a pair of matrices $(A,B) \in SL_n \times SL_n$ acts on $(M_1, M_2, \dots, M_d)$ to give $(AM_1B, AM_2B, \dots, AM_dB)$. Note that whether the symbolic matrix of linear forms $z_1M_1 + z_2M_2 + \cdots + z_dM_d$ is *full*[12] or not is unchanged by this action, which is the first relationship to the body of our paper. But in this case we only know an infinite set of generating invariants. They were determined (for arbitrary quivers) by [14, 45, 15] (and also for this specific left-right action in [1]). The invariants can be described in several ways. The papers [15, 14] describe them in terms of determinants of block matrices. Let $(X_1, X_2, \dots, X_d)$ be the generic matrices of our variables X. For the left-right action the invariants are determinants of sums of the matrices $X_i \otimes T_i$ for some arbitrary d-tuple of $k \times k$ matrices $(T_1, T_2, \dots, T_d)$. Namely, this is the set

$$\left\{\det\left(\sum_{i=1}^{d} X_i \otimes T_i\right) : k \in \mathbb{N},\, T_i \in \mathrm{Mat}_{k\times k}(\mathbb{F})\right\}.$$

varieties.

[10] Among many other foundational results—this paper and its 1890 non-constructive predecessor contain in particular the Nullstellensatz, the finite basis theorem and other cornerstones of commutative algebra and algebraic geometry.

[11] Often called semi-invariants, to distinguish them from invariants of GL_n.

[12] As in the definition before Theorem 2.2.

These are precisely the polynomials we care about in the body of this paper, in both the non-commutative PIT problem, as well as in eliminating divisions from non-commutative circuits. By Hilbert's basis theorem, we know that a finite subset of this set generates all invariants, and hence in particular we can take them to be all invariants of degree below some finite bound (which puts an upper bound on the dimension k of the auxiliary matrices T_i). Attempts of giving an upper bound on the degree are surveyed by Derksen in [12], who obtains an upper bound of $2^{O(n^2)}$ on the degree. Recently, the bound has been improved to $2^{O(n\log n)}$ by Ivanyos, Qiao and Subrahmanyam in [29], see also [30]. Note that a polynomial upper bound here implies one for Problem 9.4, and vice versa.

Acknowledgement. We thank Susan Durst, Peter Malcolmson and Aidan Schofield for useful references, Amir Yehudayoff and Klim Efremenko for comments on an earlier version of the paper.

References

[1] BHARAT ADSUL, SURESH NAYAK, AND K. V. SUBRAHMANYAM: A geometric approach to the Kronecker problem II: Invariants of matrices for simultaneous left-right actions, 2010. Available at http://www.cmi.ac.in/~kv/ANS10.pdf. 388

[2] SHIMSHON A. AMITSUR: Rational identities and applications to algebra and geometry. *J. Algebra*, 3(3):304–359, 1966. [doi:10.1016/0021-8693(66)90004-4] 362

[3] SHIMSHON A. AMITSUR AND JACOB LEVITZKI: Minimal identities for algebras. *Proc. Amer. Math. Soc.*, 1(4):449–463, 1950. [doi:10.1090/S0002-9939-1950-0036751-9] 380, 386

[4] JEAN BERSTEL AND CHRISTOPHE REUTENAUER: *Rational series and their languages*. Springer, 1988. 359

[5] ANDREJ BOGDANOV AND HOETECK WEE: More on noncommutative polynomial identity testing. In *IEEE Conference on Computational Complexity (CCC)*, pp. 92–99. IEEE Comp. Soc. Press, 2005. [doi:10.1109/CCC.2005.13] 380

[6] RICHARD P. BRENT: The parallel evaluation of general arithmetic expressions. *J. ACM*, 21(2):201–206, 1974. [doi:10.1145/321812.321815] 366

[7] PETER BÜRGISSER, MICHAEL CLAUSEN, AND MOHAMMAD AMIN SHOKROLLAHI: *Algebraic complexity theory*. Volume 315 of *Grundlehren der mathematischen Wissenschaften*. Springer, 1997. [doi:10.1007/978-3-662-03338-8] 358, 382

[8] PAUL MORITZ COHN: *Free rings and their relations*. Academic Press, 1985. 359, 370

[9] PAUL MORITZ COHN: *Skew Fields: Theory of General Division Rings*. Volume 57 of *Encyclopedia of Mathematics*. Cambridge Univ. Press, 1995. 361, 363

[10] PAUL MORITZ COHN AND CHRISTOPHE REUTENAUER: On the construction of the free field. *Int. J. Algebra Comput.*, 9(3-4):307–324, 1999. [doi:10.1142/S0218196799000205] 377, 378

[11] DAVID COX, JOHN LITTLE, AND DONAL O'SHEA: *Ideals, Varieties, and Algorithms*. Undergraduate Texts in Mathematics. Springer, 2007. [doi:10.1007/978-0-387-35651-8] 386

[12] HARM DERKSEN: Polynomial bounds for rings of invariants. *Proc. Amer. Math. Soc.*, 129(4):955–963, 2001. [doi:10.1090/S0002-9939-00-05698-7] 386, 389

[13] HARM DERKSEN AND GREGOR KEMPER: *Computational Invariant Theory*. Volume 130 of *Encyclopaedia of Math. Sci.* Springer, 2002. [doi:10.1007/978-3-662-04958-7] 386

[14] HARM DERKSEN AND JERZY WEYMAN: Semi-invariants of quivers and saturation for Littlewood-Richardson coefficients. *J. Amer. Math. Soc.*, 13(3):467–479, 2000. [doi:10.1090/S0894-0347-00-00331-3] 388

[15] MÁTYÁS DOMOKOS AND ALEXANDR N. ZUBKOV: Semi-invariants of quivers as determinants. *Transform. Groups*, 6(1):9–24, 2001. [doi:10.1007/BF01236060] 388

[16] STEPHEN DONKIN: Invariants of several matrices. *Inventiones Mathematicae*, 110(1):389–401, 1992. [doi:10.1007/BF01231338] 388

[17] MICHAEL A. FORBES AND AMIR SHPILKA: Explicit Noether normalization for simultaneous conjugation via polynomial identity testing. In *Proc. 17th Internat. Workshop on Randomization and Computation (RANDOM'13)*, pp. 527–542, 2013. [doi:10.1007/978-3-642-40328-6_37, arXiv:1303.0084] 387, 388

[18] EDWARD FORMANEK: Invariants and the ring of generic matrices. *J. Algebra*, 89(1):178–223, 1984. [doi:10.1016/0021-8693(84)90240-0] 388

[19] MARC FORTIN AND CHRISTOPHE REUTENAUER: Commutative/non-commuative rank of linear matrices and subspaces of matrices of low rank. *Séminaire Lotharingien de Combinatoire*, 52, 2004. Available at EUDML. 382

[20] PETER GABRIEL: Unzerlegbare Darstellungen I. *Manuscripta Mathematica*, 6(1):71–103, 1972. [doi:10.1007/BF01298413] 388

[21] ISRAEL MOISEEVICH GELFAND, SERGEI GELFAND, VLADIMIR S. RETAKH, AND ROBERT LEE WILSON: Quasideterminants. *Adv. Math.*, 193(1):56–141, 2005. [doi:10.1016/j.aim.2004.03.018, arXiv:math/0208146] 363

[22] ISRAEL MOISEEVICH GELFAND AND VLADIMIR S. RETAKH: Determinants of matrices over noncommutative rings. *Funct. Anal. Appl.*, 25(2):91–102, 1991. [doi:10.1007/BF01079588] 359, 362

[23] ISRAEL MOISEEVICH GELFAND AND VLADIMIR S. RETAKH: A theory of noncommutative determinants and characteristic functions of graphs. *Funct. Anal. Appl.*, 26(4), 1992. [doi:10.1007/BF01075044] 362

[24] DAVID HILBERT: Über die vollen Invariantensysteme. *Math. Ann.*, 42(3):313–373, 1893. [doi:10.1007/BF01444162] 388

[25] PAVEL HRUBEŠ AND AVI WIGDERSON: Non-commutative circuits with division. In *Proc. 5th Conf. Innovations in Theoret. Comp. Sci. (ITCS'14)*, pp. 49–66, 2014. [doi:10.1145/2554797.2554805] 357, 381

[26] PAVEL HRUBEŠ, AVI WIGDERSON, AND AMIR YEHUDAYOFF: Relationless completeness and separations. In *IEEE Conference on Computational Complexity (CCC)*, pp. 280–290, 2010. Preliminary version in ECCC. [doi:10.1109/CCC.2010.34] 364, 372

[27] PAVEL HRUBEŠ, AVI WIGDERSON, AND AMIR YEHUDAYOFF: Non-commutative circuits and the sum of squares problem. *J. Amer. Math. Soc.*, 24(3):871–898, 2011. Preliminary versions in ECCC and STOC'10. [doi:10.1090/S0894-0347-2011-00694-2] 358

[28] PAVEL HRUBEŠ AND AMIR YEHUDAYOFF: Arithmetic complexity in ring extensions. *Theory of Computing*, 7(8):119–129, 2011. [doi:10.4086/toc.2011.v007a008] 382

[29] GÁBOR IVANYOS, YOUMING QIAO, AND K. V. SUBRAHMANYAM: Non-commutative Edmonds' problem and matrix semi-invariants. *CoRR*, 2015. [arXiv:1508.00690] 365, 379, 386, 389

[30] GÁBOR IVANYOS, YOUMING QIAO, AND K. V. SUBRAHMANYAM: On generating the ring of matrix semi-invariants. *CoRR*, 2015. [arXiv:1508.01554] 389

[31] DMITRY KALIUZHNYI-VERBOVETSKYI AND VICTOR VINNIKOV: Noncommutative rational functions, their difference-differential calculus and realizations. *Multidim. Systems and Sig. Proc.*, 23(1):49–77, 2012. [doi:10.1007/s11045-010-0122-3, arXiv:1003.0695] 361, 362

[32] DMITRY S. KALIUZHNYI-VERBOVETSKYI AND VICTOR VINNIKOV: Singularities of noncommutative rational functions and minimal factorizations. *Lin. Alg. Appl.*, 430(4):869–889, 2009. [doi:10.1016/j.laa.2008.08.027] 359

[33] HANSPETER KRAFT AND CLAUDIO PROCESI: *Classical Invariant Theory, a Primer*. 1996. Available at `http://jones.math.unibas.ch/~kraft/Papers/KP-Primer.pdf`. 386

[34] TSIU-KWEN LEE AND YIQIANG ZHOU: Right ideals generated by an idempotent of finite rank. *Lin. Alg. and Appl.*, 431(11):2118–2126, 2009. [doi:10.1016/j.laa.2009.07.005] 386

[35] PETER MALCOLMSON: A prime matrix ideal yields a skew field. *J. of the London Math. Soc.*, 18(2):221–233, 1978. [doi:10.1112/jlms/s2-18.2.221] 370

[36] KETAN D. MULMULEY: On P vs. NP and geometric complexity theory. *J. ACM*, 58(2):5:1–5:26, 2011. [doi:10.1145/1944345.1944346] 387

[37] KETAN D. MULMULEY: The GCT program toward the P vs. NP problem. *Com. of the ACM*, 55(6):98–107, 2012. [doi:10.1145/2184319.2184341] 387

[38] KETAN D. MULMULEY: Geometric complexity theory V: Equivalence between blackbox derandomization of polynomial identity testing and derandomization of Noether's normalization lemma. *CoRR*, p. 60pp, 2012. Extended Abstract in FOCS'12. [arXiv:1209.5993] 388

[39] NOAM NISAN: Lower bounds for non-commutative computation (extended abstract). In *Proc. 23rd STOC*, pp. 410–418. ACM Press, 1991. [doi:10.1145/103418.103462] 358, 372, 376, 385

[40] CLAUDIO PROCESI: The invariant theory of $n \times n$ matrices. *Adv. Math.*, 19(3):306–381, 1976. [doi:10.1016/0001-8708(76)90027-X] 388

[41] RAN RAZ AND AMIR SHPILKA: Deterministic polynomial identity testing in non-commutative models. *Comput. Complexity*, 14(1):1–19, 2005. Preliminary version in CCC'04. [doi:10.1007/s00037-005-0188-8] 380

[42] YURI PTIRIMOVICH RAZMYSLOV: Trace identities of full matrix algebras over a field of characteristic zero. *Izv. Akad. Nauk SSSR Ser. Mat.*, 8(4):727–760, 1974. [doi:10.1070/IM1974v008n04ABEH002126] 388

[43] CHRISTOPHE REUTENAUER: Inversion height in free fields. *Selecta Mathematica*, 2(1):93–109, 1996. [doi:10.1007/BF01587940] 359, 363

[44] LOUIS HALLE ROWEN: *Polynomial identities in ring theory*. Volume 84. Academic Press, 1980. 386

[45] AIDAN SCHOFIELD AND MICHEL VAN DEN BERGH: Semi-invariants of quivers for arbitrary dimension vectors. *Indag. Math*, 12(1):125–138, 2001. [doi:10.1016/S0019-3577(01)80010-0, arXiv:math/9907174] 388

[46] AMIR SHPILKA AND AMIR YEHUDAYOFF: Arithmetic circuits: A survey of recent results and open questions. *Found. and Trends in Theor. Comput. Sci.*, 5(3-4):207–388, 2010. [doi:10.1561/0400000039] 358

[47] VOLKER STRASSEN: Gaussian elimination is not optimal. *Numer. Math.*, 13(4):354–356, 1969. [doi:10.1007/BF02165411] 365

[48] VOLKER STRASSEN: Vermeidung von Divisionen. *J. of Reine Angew. Math.*, 1973(264):182–202, 1973. [doi:10.1515/crll.1973.264.184] 360, 382

[49] LESLIE G. VALIANT: Completeness classes in algebra. In *Proc. 11th STOC*, pp. 249–261. ACM Press, 1979. [doi:10.1145/800135.804419] 359, 364, 387

AUTHORS

Pavel Hrubeš
Institute of Mathematics of CAS, Prague
hrubes@math.cas.cz
`http://www.math.cas.cz/homepage/main_page.php?id_membre=189`

Avi Wigderson
Institute for Advanced Study, Princeton
avi@ias.edu
http://www.math.ias.edu/avi/home

ABOUT THE AUTHORS

PAVEL HRUBEŠ graduated from Charles University in Prague in 2008 with supervisor Pavel Pudlák. After a period of joyful postdocs in U.S. and Canada, he became a mature worker in the factory of academia. He contributes mainly to arithmetic circuit and proof complexity. His hobbies include mountaineering and brooding.

AVI WIGDERSON was born in Haifa, Israel in 1956, and received his Ph. D. in 1983 at Princeton University under Dick Lipton. He enjoys and is fascinated with studying the power and limits of efficient computation, and the remarkable impact of this field on understanding our world. Avi's other major source of fascination and joy are his three kids, Eyal, Einat, and Yuval, and his granddaughter Tamar.