Nonnegative rank measures and monotone algebraic branching programs

Hervé Fournier, Guillaume Malod, Maud Szusterman, Sébastien Tavenas

Nonnegative rank measures and monotone algebraic branching programs

Hervé Fournier[1], Guillaume Malod[1], Maud Szusterman[1], and Sébastien Tavenas[2]

[1]Université de Paris, IMJ-PRG, CNRS, F-75013 Paris, France
[2]Univ. Grenoble Alpes, Univ. Savoie Mont Blanc, CNRS, LAMA, 73000 Chambéry, France

July 31, 2019

Abstract

Inspired by Nisan's characterization of noncommutative complexity (Nisan 1991), we study different notions of nonnegative rank, associated complexity measures and their link with monotone computations. In particular we answer negatively an open question of Nisan asking whether nonnegative rank characterizes monotone noncommutative complexity for algebraic branching programs. We also prove a rather tight lower bound for the computation of elementary symmetric polynomials by algebraic branching programs in the monotone setting or, equivalently, in the homogeneous syntactically multilinear setting.

1 Introduction

Measures based on rank are one of the main tools to prove lower bounds in algebraic complexity theory. The complexity of first-order partial derivatives is the key ingredient for the best lower bound known for general circuits [2]. When looking at higher-order partial derivatives, one can consider their rank: the rank of partial derivatives, and some variants, have been intensively used to obtain lower bounds on restricted models [20, 21, 18]. Nisan [19] provided one of the earliest and cleanest examples of such a measure: when computing a polynomial over noncommuting variables by a so-called *algebraic branching program*, it gives an exact characterization of the complexity.[1] To state this result more precisely, let us give here the definition of algebraic branching programs used in [19].

Definition 1. *An* algebraic branching program *(ABP) is a layered directed acyclic graph with a source s and a sink t. The first layer contains only the source s, the last layer contains only the sink t. Edges can only appear between vertices of successive layers and carry a weight which is a linear form of the variables. The weight of a path from s to t is the product of the weights of its edges. The (homogeneous) polynomial computed by the ABP is the sum of the weights of the paths from s to t. The width of a layer is the number of vertices on that layer.*

[1]It was noticed in [8] (see also [17]) that this result actually follows from an older characterization for word series [11]. This characterization was also extended to tree series in [5], which can be applied to circuits.

This definition makes sense both in the usual case of commuting variables and in the case of noncommuting variables, over a free algebra, which we consider for the moment. For a noncommutative homogeneous polynomial P of degree d over variables in X, define matrices $M_i(P)$, for $0 \leqslant i \leqslant d$: the rows of $M_i(P)$ are indexed by all possible monomials u over X of degree i, the columns are indexed by all possible monomials v over X of degree $d-i$, and the coefficient (u, v) of $M_i(P)$ is the coefficient of the monomial uv in P. We call this matrix the i-th Nisan matrix of P. The characterization is then expressed by the following theorem.

Theorem 1 (Nisan [19], Fliess [11])**.** *The size of a smallest ABP computing a noncommutative polynomial P is the sum of the ranks of its Nisan matrices, i.e., $\sum_{i=0}^{d} \operatorname{rk} M_i(P)$. More precisely, the value $\operatorname{rk} M_i(P)$ is the width of the i-th layer in a smallest ABP computing P. It is also the smallest possible width of the i-th layer in any ABP computing P.*

Nisan also considers the case of monotone noncommutative computations. In this case Nisan does not obtain a characterization of monotone noncommutative complexity, but a sufficient tool for lower bounds, using the notion of nonnegative rank.

Definition 2. *An ABP over an ordered field is called* monotone *if all coefficients of linear forms on the edges are nonnegative.*

Definition 3. *The* nonnegative rank *of a nonnegative matrix M, $\operatorname{rk}^+ M$, is the smallest integer r such that M can be written as a sum of r rank-1 nonnegative matrices.*

Proposition 1 (Nisan [19])**.** *For a polynomial P with nonnegative coefficients, the value $\operatorname{rk}^+ M_i(P)$ is the smallest possible width of the i-th layer in a monotone ABP computing P. The size of a smallest monotone ABP computing P is therefore at least $\sum_{i=0}^{d} \operatorname{rk}^+ M_i(P)$.*

Nisan [19] leaves the tightness of the inequality in Proposition 1 as an open question: does nonnegative rank also provide a characterization of monotone noncommutative complexity? One of our main results is a negative answer to this question (Theorem 6).

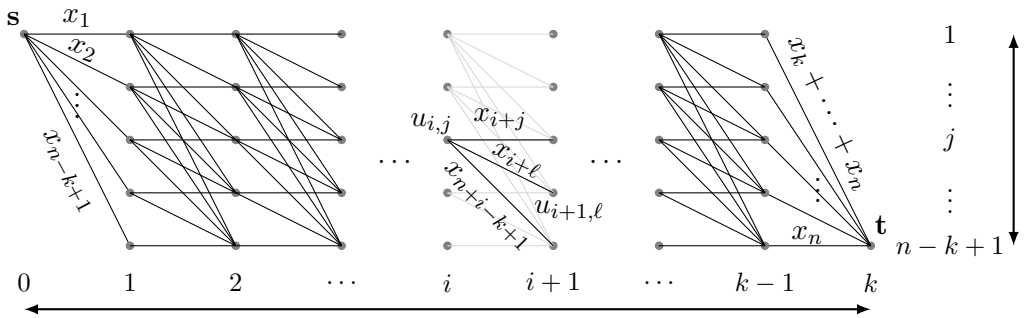
Before that, we consider in Section 2 a more general notion of monotone computation, which we call *weakly monotone*. Where monotonicity completely disallows cancellations, weak monotonicity allows them as long as any monomial appearing in the computation also appears in the end result. In other words, cancellations can be used to obtain the specific coefficients of a polynomial, but not to produce and then cancel out monomials outside the support of the polynomial. We strengthen Proposition 1 for weakly monotone noncommutative ABPs using a new rank measure. We then obtain a separation showing that weakly monotone noncommutative arithmetic formulas can be exponentially more powerful than monotone noncommutative ABPs. Thus weakly monotone lower bounds are stronger than monotone lower bounds.

In Section 3 we prove Theorem 6, answering Nisan's question, and more generally explore the link between nonnegative rank measures and the size of monotone noncommutative algebraic branching programs.

Finally, in Section 4 we focus on proving lower bounds for monotone *commutative* ABPs, building on ideas from the previous sections to develop new tools. Imposing monotonicity as a restriction on arithmetic computations to prove lower bounds has a long history [22, 15], which often involves hard polynomials and yields exponential lower bounds. We focus here on the elementary symmetric polynomials $e_{n,k}$.

While it is known that the elementary symmetric polynomials $e_{n,k}$ require monotone, or even homogeneous multilinear, formulas of size $k^{\Omega(k)}n$ [13], these can be efficiently computed: by arithmetic circuits of size $O(n \log k)$ [2]; by depth-3 arithmetic formulas of size $O(n^2)$, using interpolation; by monotone ABPs of size $O(k(n-k))$ by the following simple dynamic programming construction.

Apart from s and t, the ABP for $e_{n,k}$ has layers indexed by $1 \leqslant i \leqslant k-1$; the i-th layer contains $n-k+1$ vertices $u_{i,1}, \ldots, u_{i,n-k+1}$; vertex $u_{i,j}$ sends an edge of weight $x_{i+\ell}$ to vertex $u_{i+1,\ell}$ for all $2 \leqslant i \leqslant k-2$ and $j \leqslant \ell \leqslant n-k+1$; s sends an edge of weight x_ℓ to $u_{1,\ell}$ and $u_{k-1,\ell}$ sends an edge of weight $x_{k+\ell-1} + \cdots + x_n$ to t for $1 \leqslant \ell \leqslant n-k+1$. The monotone branching program obtained this way has size $(k-1)(n-k+1)+2$.



The existence of efficient computations imply that our lower bound techniques must be very precise. Surprisingly, there is also a very simple $\Omega(k(n-k+1))$ monotone lower bound in [16], but in a model where edge weights can only be a scalar or a scalar times a variable, not linear forms, which would only give a trivial lower bound in our setting. Our second main result is a similarly quadratic lower bound for our model, and for weakly monotone computations, at the cost of a significant increase in the complexity of the proof: we use a generalization of a combinatorial problem known as Galvin's problem.[2] Our lower bound can be equivalently stated as applying to homogeneous syntactically multilinear ABPs.

Let us add one remark on the definition of ABPs. This computation model is inherently homogeneous and we only consider nonzero homogeneous polynomials. We could also consider nonhomogeneous ABPs: these are directed acyclic graphs with a source and a sink, not necessarily layered, with arcs labelled with affine forms. In the noncommutative case, when computing a homogeneous polynomial, one can show that there is always a minimal-size ABP which is homogeneous and corresponds to Definition 1. We provide a proof sketch in Appendix A. This is also true in the commutative case for weakly monotone computations. Hence we shall consider only homogeneous branching programs.

Throughout the paper we use $\mathbb{R}$ in the case of an ordered field, but these results hold over any ordered field. When the field is not ordered we denote it by $\mathbb{K}$ and assume it is of characteristic 0.

2 A rank measure for weakly monotone computations

2.1 Weakly monotone computations

As defined before, the weight of a path is the product of the weights of its edges, i.e., a product of linear forms. Any of the monomials obtained when expanding completely this product, by choosing

[2] A different generalization of this combinatorial problem was recently used to prove almost quadratic lower bounds on the size of syntactically multilinear circuits [1].

one term in each linear form, is said to be *produced along the path.*

Definition 4. *An ABP is called* weakly monotone *if any monomial produced along a path from the source to the sink has a nonzero coefficient in the polynomial computed by this ABP.*

Note that this notion of monotonicity makes sense both in the commutative and noncommutative settings (Sections 2 and 3 deal with noncommutative computations, while we will use the commutative case in Section 4). We now define a new measure for weakly monotone computations. We will denote the support of a matrix M by $\operatorname{supp} M$: it is the subset of the coordinates of M which correspond to nonzero entries.

Definition 5. *The* weakly nonnegative rank *of a matrix M, denoted by $\operatorname{rk}^{\mathrm{w}} M$, is the smallest number r such that there exist $M_1, \ldots, M_r$ of rank 1 (with entries of any sign) such that $\operatorname{supp} M_i \subseteq \operatorname{supp} M$ for all i and $\sum_{i=1}^{r} M_i = M$.*

The usual nonnegative rank of a matrix already plays a role in several areas. For instance, the fact that the minimum number of facets of an extension of a polyhedron is equal to the nonnegative rank of its so-called slack matrix. In another direction, for a $0,1$-matrix M, $\log(\operatorname{rk} M)$ is a lower bound on the communication complexity of the associated problem. The *log-rank conjecture* stipulates that there is also a $\log^{O(1)}(\operatorname{rk} M)$ upper bound. This conjecture is known to be equivalent to the fact that for any $0,1$-matrix M, $\log(\operatorname{rk}^{+} M) = \log^{O(1)}(\operatorname{rk} M)$. The influence of communication complexity will be felt here too, as it can be seen from the use of the support of the matrix in the definition of weakly nonnegative rank. In fact, we will borrow a few more basic concepts from communication complexity.

Definition 6. *For a matrix M with rows indexed by a set I and columns indexed by a set J, a* combinatorial rectangle *is a subset of $I \times J$ of the form $A \times B$, with $A \subseteq I$ and $B \subseteq J$.*

A cover *of a matrix M is a set of combinatorial rectangles, included in the support of M and whose union is equal to the support of M. We define $\operatorname{cov} M$ as the smallest size of a cover of M.*

Proposition 2. *We have $\operatorname{cov} M \leqslant \operatorname{rk}^{\mathrm{w}} M$ and $\operatorname{rk} M \leqslant \operatorname{rk}^{\mathrm{w}} M$. For a nonnegative matrix M, $\operatorname{rk}^{\mathrm{w}} M \leqslant \operatorname{rk}^{+} M$.*

Let us remark that we can have $\operatorname{rk} M < \operatorname{rk}^{\mathrm{w}} M$: this is the case for the following matrix [6], for which $\operatorname{rk} R = 3$ and $\operatorname{cov} R = 4$:

$$R = \begin{pmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}.$$

The following proposition is the weak monotone version of Proposition 1.

Proposition 3. *For a noncommutative polynomial P, the smallest size of the i-th layer of a weakly monotone ABP computing P is equal to $\operatorname{rk}^{\mathrm{w}} M_i(P)$. Hence the weakly monotone ABP size of P is greater or equal to $\sum_i \operatorname{rk}^{\mathrm{w}} M_i(P)$.*

Proof. Let $\ell \in \{1, \ldots, d-1\}$ and let $M = M_\ell(P)$, $r = \operatorname{rk}^{\mathrm{w}} M$.

Consider a weakly monotone ABP computing P. Let s be the size of layer ℓ. Cutting the ABP at layer ℓ we get $P = \sum_{i=1}^{s} L_i R_i$. Let A_i be the matrix of the polynomial $L_i R_i$. The matrices $A_1, \ldots, A_s$ satisfy all the conditions to show that $\operatorname{rk}^{\mathrm{w}} M \leqslant s$.

Conversely, write $M_\ell(P) = A_1 + \ldots + A_r$ where A_i are rank 1 matrices with $\operatorname{supp} A_i \subseteq \operatorname{supp} M$. Each A_i can be interpreted as a product of two polynomials $L_i R_i$. It is easy to design a weakly monotone ABP with ℓ-layer of size r computing the polynomials $L_1, \ldots, L_r$ on the ℓ-th layer.

So we have proved that for any ℓ, the minimal size of the ℓ-th layer of a weakly monotone ABP computing P is equal to $\operatorname{rk}^{\mathrm{w}} M_\ell(P)$. The last inequality follows from the fact that the minimal size of a weakly monotone ABP computing P is greater or equal to the sum of the minimal size of the different layers. □

2.2 Separation of rank measures

We show now that we can have $\operatorname{rk}^{\mathrm{w}} M < \operatorname{rk}^{+} M$. In the following J is the matrix with all entries equal to 1 and $\|\cdot\|$ is the infinite norm.

Proposition 4. *Let M be a nonnegative matrix. For $\varepsilon > 0$ small enough, $N = M + \varepsilon J$ satisfies $\operatorname{rk}^{\mathrm{w}} N \leqslant \operatorname{rk} M + 1$ and $\operatorname{rk}^{+} N \geqslant \operatorname{cov} M$.*

Proof. We have $\operatorname{rk} N \leqslant \operatorname{rk} M + 1$ because J is of rank 1, and $\operatorname{rk} N = \operatorname{rk}^{\mathrm{w}} N$ since the support of N is full. Hence $\operatorname{rk}^{\mathrm{w}} N \leqslant \operatorname{rk} M + 1$.

It remains to show the lower bound on $r = \operatorname{rk}^{+} N$. Write $N = N_1 + \ldots + N_r$ with N_i nonnegative matrix of rank 1. Write $N_i = a_i b_i^T$ with a_i and b_i nonnegative satisfying $\|a_i\| = \|b_i\|$: this implies that $\|a_i\|, \|b_i\| \leqslant \sqrt{\|N\|} \leqslant 2\sqrt{\|M\|}$ for ε small enough. Let $\tilde{a}_i$ and $\tilde{b}_i$ be obtained from a_i and b_i by putting to 0 all entries smaller or equal to $\sqrt{\varepsilon}$. Let $\tilde{N}_i = \tilde{a}_i \tilde{b}_i^T$. The support of $\tilde{N}_i$ is a combinatorial rectangle. Moreover, $\operatorname{supp} \tilde{N}_i \subseteq \operatorname{supp} M$ since any nonzero entry of $\tilde{N}_i$ is greater than ε.

Let us show that $\operatorname{supp} M \subseteq \bigcup_i \operatorname{supp} \tilde{N}_i$. Let $\tilde{N} = \tilde{N}_1 + \ldots + \tilde{N}_r$. For any entry (x, y), we have $|N_i(x, y) - \tilde{N}_i(x, y)| \leqslant 2\sqrt{\varepsilon}\sqrt{\|M\|}$ so $\|N - \tilde{N}\| \leqslant 2r\sqrt{\varepsilon\|M\|}$. This shows that $\|M - \tilde{N}\| \leqslant \|M - N\| + \|N - \tilde{N}\| \leqslant \varepsilon + 2r\sqrt{\varepsilon\|M\|}$. Hence $\|M - \tilde{N}\|$ is smaller than the smallest nonzero entry of M for ε small enough. This proves that $\operatorname{supp} M$ is covered by $\bigcup_{i=1}^{r} \operatorname{supp} \tilde{N}_i$. □

We want to apply the previous proposition to a matrix with a large gap between rank and covering bound. Such examples are known: the $n \times n$ matrix defined by $M_{i,j} = (a_i - a_j)^2$ for distinct reals $a_1, \ldots, a_n$ has rank 3 but $\operatorname{cov} M = \Omega(\log n)$ [3]; the slack matrix of a generic polygon also exhibits such a gap [10] (note that this matrix is not explicit).

We will build on a third construction to obtain an exponential separation between weakly monotone formulas and monotone ABPs in the noncommutative setting. Let U_n be the matrix whose rows and columns are indexed by $\{0,1\}^n$ and which is defined by $U_n(u, v) = (\langle u, v\rangle - 1)^2$, where the scalar product is over $\mathbb{R}$.

Theorem 2 ([7], see also [9])**.** *It holds that $\operatorname{rk} U_n = O(n^2)$ and $\operatorname{cov} U_n = 2^{\Omega(n)}$.*

Using Proposition 4, this theorem gives a matrix with an exponential gap between weakly nonnegative rank and nonnegative rank.

Proposition 5. *For $\varepsilon > 0$ small enough, $\operatorname{rk}^{\mathrm{w}}(U_n + \varepsilon J) = O(n^2)$ and $\operatorname{rk}^{+}(U_n + \varepsilon J) = 2^{\Omega(n)}$.*

2.3 Separating noncommutative monotone and weakly monotone classes

Let us define a noncommutative polynomial over the variables $\{x_0, x_1\}$. For $u \in \{0,1\}^n$, let $x_u = x_{u_1} \ldots x_{u_n}$ and define $P = \sum_{u,v \in \{0,1\}^n} (\langle u, v\rangle - 1)^2 x_u x_v$. This polynomial was used in [14] to obtain the following separation.

Theorem 3 ([14])**.** *The noncommutative polynomial P defined above has formula size $O(n^3)$ but monotone circuit size $2^{\Omega(n)}$.*

As a consequence, we get a separation illustrating the difference between monotone and weakly monotone computations.

Definition 7. *A formula is called* weakly monotone *if any monomial produced by the computation (before any possible cancellations) has a nonzero coefficient in the computed polynomial. More formally, a formula is weakly monotone if any monomial produced by a parse tree is present in the computed polynomial.*

Theorem 4. *For $\varepsilon > 0$ small enough, the noncommutative polynomial $P + \varepsilon(x_0 + x_1)^{2n}$ has weakly monotone formula size $O(n^3)$ but requires monotone ABP size $2^{\Omega(n)}$.*

Proof. Let $Q = P + \varepsilon(x_0 + x_1)^{2n}$ for some $\varepsilon > 0$ small enough. The polynomial P has formula size $O(n^3)$ by the upper bound from Theorem 3. The polynomial $\varepsilon(x_0 + x_1)^{2n}$ has formula size $O(n)$. Since the support of Q is full, the formula obtained for Q by summing these two formulas is weakly monotone.

The middle Nisan matrix of Q is $M_n(Q) = U_n + \varepsilon J$ so $\operatorname{rk}^+ M_n(Q) = 2^{\Omega(n)}$ by Proposition 5. It follows from Proposition 1 that Q has monotone ABP size $2^{\Omega(n)}$. □

3 Monotone noncommutative complexity vs monotone rank measures

This section is devoted to the comparison between nonnegative rank measures and the size of monotone noncommutative algebraic branching programs, in particular Nisan's question on the tightness of the lower bound for monotone noncommutative ABPs. Let us start by a particular case where the inequality is tight.

We work over a field $\mathbb{K}$ of characteristic zero. We say a vector v is a weakly monotone linear combination of $u_1, \dots, u_p$ if there exist scalars λ_i for $1 \leqslant i \leqslant p$ such that no cancellations occur:

$$v = \sum_{i \in [1,p]} \lambda_i u_i \quad \text{with} \quad \operatorname{supp}(v) = \bigcup_{\substack{i \in [1,p] \\ \lambda_i \neq 0}} \operatorname{supp}(u_i).$$

3.1 In the case of ranks at most 2

In the case where each Nisan matrix is of rank at most 2, we prove that an algebraic branching program of minimal size can be chosen to be monotone (or weakly monotone). Since $\operatorname{rk} M \leqslant 2$ implies $\operatorname{rk} M = \operatorname{rk}^{\mathrm{w}} M = \operatorname{rk}^+ M$, it means that the measures $\sum_i \operatorname{rk}^+ M_i(P)$ and $\sum_i \operatorname{rk}^{\mathrm{w}} M_i(P)$ do characterize the monotone and weakly monotone ABP size in this case.

Lemma 1. *If a homogeneous noncommutative polynomial P of degree d with nonnegative coefficients satisfies $\operatorname{rk} M_i(P) = 2$ for all $0 < i < d$, then there exists a monotone ABP of width 2 computing P. Hence the minimal size of a monotone ABP computing P is equal to $\sum_{i=0}^{d} \operatorname{rk}^+ M_i(P)$.*

Proof. For any $i \in \{1, \dots, d\}$, the matrix $M_i(P)$ is a matrix of size $k \times \ell$, of nonnegative rank 2 and so also of rank 2.

Claim 1. *There exist U_i and V_i two columns of $M_i(P)$ which nonnegatively generate all columns of $M_i(P)$.*

Proof. Up to removing the all-zeros columns, we can assume all columns of $M_i(P)$ are nonzero. The ℓ column vectors $W_1, \dots, W_\ell$ of $M_i(P)$ lie in a 2-dimensional vector space $\mathcal{P}$ of $\mathbb{R}^k$. Let us consider the vectors $\frac{W_1}{\|W_1\|_1}, \dots, \frac{W_\ell}{\|W_\ell\|_1}$. They are still in $\mathcal{P}$, but, as they are nonnegative, they are also in the affine hyperplane $\mathcal{H} \stackrel{\text{def}}{=} \{X \in \mathbb{R}^k \mid \sum X_i = 1\}$. The intersection of $\mathcal{P}$ and $\mathcal{H}$ is an affine line, so we can find U_i and V_i among the $(W_j)_{1 \leqslant j \leqslant \ell}$ such that for all j, $\frac{W_j}{\|W_j\|}$ is in the closed affine segment $\left[\frac{U_i}{\|U_i\|}, \frac{V_i}{\|V_i\|}\right]$. As U_i and V_i are nonnegative, they nonnegatively generate all columns of $M_i(P)$. □

For $i \in \{1, \dots, d-1\}$, let $Q_1^{(i)}$ and $Q_2^{(i)}$ be the polynomials corresponding to the columns U_i and V_i. Let X be the set of variables of P, then

$$P(X) = \sum_{\substack{m \deg d-i-1 \\ \text{monomial}}} P_m^{(i+1)}(X) \cdot m = \sum_{\substack{m \deg d-i-1 \\ \text{monomial}}} \sum_{v \in X} P_{vm}^{(i)}(X) \cdot vm.$$

So, each column of $M_{i+1}(P)$ can be obtained by $P_m^{(i+1)}(X) = \sum_{v \in X} P_{vm}^{(i)}(X) \cdot v$, and so by a monotone combination $\sum_{v \in X} (\alpha_v Q_1^{(i)} + \beta_v Q_2^{(i)}) \cdot v = Q_1^{(i)} \left(\sum_{v \in X} \alpha_v v\right) + Q_2^{(i)} \left(\sum_{v \in X} \alpha_v v\right)$. This designs a monotone ABP of width 2 computing $Q_i^{(1)}$ and $Q_i^{(2)}$ on layer i for all i. □

In order to prove the analog of Lemma 1 in the weakly monotone case, we will use several times the following very simple observation.

Proposition 6. *Let $\mathcal{E}_i$ be the i-th coordinate hyperplane of $\mathbb{K}^n$. and $\mathcal{E} = \bigcap_{i \in I} \mathcal{E}_i$ for $I \subseteq [n]$. Let $v \in \mathcal{E}$, and $u_1, \dots, u_p \in \mathbb{K}^n$. Assume $v = \sum_{j=1}^p \lambda_j u_j$ is a weakly monotone linear combination. Then for all j such that $\lambda_j \neq 0$, we have $u_j \in \mathcal{E}$.*

In particular, if $v \neq 0$ is a weakly monotone linear combination $v = \sum_j \lambda_j u_j$, then there exists j_0 such that $\lambda_{j_0} \neq 0$ and $u_{j_0} \in \mathcal{E} \setminus \{0\}$.

Proof. For the first point, if $\lambda_j \neq 0$, then, by definition, $\operatorname{supp}(u_j) \subseteq \operatorname{supp}(v)$, which implies that $u_j \in \mathcal{E}$. The second point follows directly by noticing that if $v \neq 0$, then at least one of the λ_j has to be nonzero. □

Lemma 2. *If P is a homogeneous noncommutative polynomial of degree d such that $\operatorname{rk} M_i(P) = 2$ for all $0 < i < d$, there exists a weakly monotone ABP of width 2 computing P. Hence the minimal size of a weakly monotone ABP computing P is equal to $\sum_{i=0}^d \operatorname{rk}^{\mathrm{w}} M_i(P)$.*

Proof. For any $i \in \{1, \dots, d\}$, the matrix $M_i(P)$ is a matrix of size $k \times \ell$, of nonnegative rank 2 and so also of rank 2.

Claim 2. *There exist U_i and V_i two columns of $M_i(P)$ which nonnegatively generate all columns of $M_i(P)$.*

Proof. Up to removing the all-zeros columns, we can assume all columns of $M_i(P)$ are nonzero. The ℓ column vectors $W_1, \dots, W_\ell$ of $M_i(P)$ lie in a 2-dimensional vector space $\mathcal{P}$ of $\mathbb{R}^k$. Let us consider the vectors $\frac{W_1}{\|W_1\|_1}, \dots, \frac{W_\ell}{\|W_\ell\|_1}$. They are still in $\mathcal{P}$, but, as they are nonnegative, they are

also in the affine hyperplane $\mathcal{H} \stackrel{\text{def}}{=} \{X \in \mathbb{R}^k \mid \sum X_i = 1\}$. The intersection of $\mathcal{P}$ and $\mathcal{H}$ is an affine line, so we can find U_i and V_i among the $(W_j)_{1 \leqslant j \leqslant \ell}$ such that for all j, $\frac{W_j}{\|W_j\|}$ is in the closed affine segment $\left[\frac{U_i}{\|U_i\|}, \frac{V_i}{\|V_i\|}\right]$. As U_i and V_i are nonnegative, they nonnegatively generate all columns of $M_i(P)$. □

For $i \in \{1, \dots, d-1\}$, let $Q_1^{(i)}$ and $Q_2^{(i)}$ be the polynomials corresponding to the columns U_i and V_i. Let X be the set of variables of P, then

$$P(X) = \sum_{\substack{m \deg d-i-1 \\ \text{monomial}}} P_m^{(i+1)}(X) \cdot m = \sum_{\substack{m \deg d-i-1 \\ \text{monomial}}} \sum_{v \in X} P_{vm}^{(i)}(X) \cdot vm.$$

So, each column of $M_{i+1}(P)$ can be obtained by $P_m^{(i+1)}(X) = \sum_{v \in X} P_{vm}^{(i)}(X) \cdot v$, and so by a monotone combination $\sum_{v \in X} (\alpha_v Q_1^{(i)} + \beta_v Q_2^{(i)}) \cdot v = Q_1^{(i)} \left(\sum_{v \in X} \alpha_v v\right) + Q_2^{(i)} \left(\sum_{v \in X} \alpha_v v\right)$. This designs a monotone ABP of width 2 computing $Q_i^{(1)}$ and $Q_i^{(2)}$ on layer i for all i. □

Then we can easily conclude:

Theorem 5. *Let P be a noncommutative polynomial, homogeneous of degree $d > 0$, such that $\operatorname{rk} M_i(P) \leqslant 2$ for all i. Then the minimal size of a weakly monotone ABP computing P is equal to $\sum_{i=0}^{d} \operatorname{rk}^{\mathrm{w}} M_i(P)$. Moreover, if P is nonnegative, the minimal size of a monotone ABP computing P is equal to $\sum_{i=0}^{d} \operatorname{rk}^{+} M_i(P)$.*

Proof. Assume P is nonnegative homogeneous of degree $d > 0$. We prove the second point by induction on d. If $d = 1$ the polynomial P is linear with nonnegative coefficients, $P \neq 0$, and thus $\operatorname{rk}^{+} M_0(P) + \operatorname{rk}^{+} M_1(P) = 2$, which is the size of a minimal monotone ABP. Assume now that $d > 1$. If $\operatorname{rk} M_i(P) = 2$ for all $0 < i < d$, then the minimal size of a monotone ABP computing P is equal to $\sum_{i=0}^{d} \operatorname{rk}^{+} M_i(P)$ by Lemma 1. Otherwise, there exists $0 < i < d$ such that $\operatorname{rk}^{+} M_i(P) = 1$. It means that $P = QR$ with Q and R homogeneous of degree i and $d - i$. By induction the minimal size of a monotone ABP computing Q is equal to $\sum_{i=0}^{d} \operatorname{rk}^{+} M_i(Q)$ and similarly for R. The conclusion follows easily for P.

The proof of the first point is analogous, using Lemma 2. □

3.2 Separation of monotone rank measure and ABP size

We now prove a separation between the sum-of-ranks measure and the minimal noncommutative ABP size, both in the monotone and in the weakly monotone cases.

If $X = X_1 \uplus \ldots \uplus X_d$ is a partition of the set of variables, a noncommutative polynomial f is called *ordered* over the family $X_1, \dots, X_d$ if it is homogeneous of degree d and if each monomial m from f is of the form $v_1 v_2 \cdots v_d$, where $v_i \in X_i$ for each i.

Lemma 3. *There exists a noncommutative ordered degree 3 polynomial H with nonnegative coefficients in $\mathbb{R}$ over the set of variables (X, Y, Z) with $|X| = 4, |Y| = 2, |Z| = 4$, such that $\operatorname{rk}^{+} M_i(H) = \operatorname{rk}^{\mathrm{w}} M_i(H) = \operatorname{rk} M_i(H) = 3$ for $i \in \{1, 2\}$, so that $\sum_{i=0}^{3} \operatorname{rk}^{+} M_i(H) = \sum_{i=0}^{3} \operatorname{rk}^{\mathrm{w}} M_i(H) = 8$, but the minimal size of a monotone ABP and of a weakly monotone ABP is 9.*

Proof. Define the vectors

$$A = \begin{pmatrix}1\\0\\0\\1\end{pmatrix}, B = \begin{pmatrix}0\\1\\1\\0\end{pmatrix}, C = \begin{pmatrix}1\\0\\1\\0\end{pmatrix}, D = \begin{pmatrix}0\\1\\0\\1\end{pmatrix}$$

(they correspond to the columns of the matrix R of Section 2). Then, $\mathrm{rk}(A,B,C,D) = 3$ and $\mathrm{rk}^{\mathrm{w}}(A,B,C,D) = 4$, since $\mathrm{cov}(A,B,C,D) = 4$. Define the matrices $M \in \mathbb{R}_{\geqslant 0}^{4\times 8}$ and $N \in \mathbb{R}_{\geqslant 0}^{8\times 4}$:

$$M \stackrel{\text{def}}{=} \begin{pmatrix} A & B & \frac{A+C}{2} & \frac{B+C}{2} & C & C & \frac{C+D}{2} & \frac{C+D}{2}\end{pmatrix}$$

and

$$N \stackrel{\text{def}}{=} \begin{pmatrix} A & B & \frac{A+C}{2} & \frac{B+C}{2} \\ C & C & \frac{C+D}{2} & \frac{C+D}{2}\end{pmatrix}.$$

As $\frac{C+D}{2} = \frac{A+B}{2}$, the columns of M are monotone linear combinations of A, B and C. Moreover, the columns of N are monotone linear combinations of

$$\begin{pmatrix}A\\C\end{pmatrix}, \begin{pmatrix}B\\C\end{pmatrix} \text{ and } \begin{pmatrix}C\\D\end{pmatrix}.$$

Hence, $\mathrm{rk}^{+} M = \mathrm{rk}^{+} N = 3$. This shows that $\mathrm{rk}^{\mathrm{w}} M = \mathrm{rk}^{\mathrm{w}} N = 3$.

Let $X = \{x_1, x_2, x_3, x_4\}$, $Y = \{y_1, y_2\}$ and $Z = \{z_1, z_2, z_3, z_4\}$. We consider the ordered polynomial $H \in \mathbb{R}_{\geqslant 0}[X, Y, Z]$:

$$\begin{aligned} H \stackrel{\text{def}}{=}\ & x_1y_1z_1 + x_4y_1z_1 + x_2y_1z_2 + x_3y_1z_2 + x_1y_1z_3 + \frac{1}{2}x_3y_1z_3 + \frac{1}{2}x_4y_1z_3 + \frac{1}{2}x_1y_1z_4 \\ & + \frac{1}{2}x_2y_1z_4 + x_3y_1z_4 + x_1y_2z_1 + x_3y_2z_1 + x_1y_2z_2 + x_3y_2z_2 + \frac{1}{2}x_1y_2z_3 + \frac{1}{2}x_2y_2z_3 \\ & + \frac{1}{2}x_3y_2z_3 + \frac{1}{2}x_4y_2z_3 + \frac{1}{2}x_1y_2z_4 + \frac{1}{2}x_2y_2z_4 + \frac{1}{2}x_3y_2z_4 + \frac{1}{2}x_4y_2z_4. \end{aligned}$$

One can verify than the middle Nisan matrices of H are $M_1(H) = M$ and $M_2(H) = N$.

Assume that there exists a weakly monotone noncommutative homogeneous ABP of size $8 = \sum \mathrm{rk}^{\mathrm{w}} M_i(H)$ computing H. It means that the ABP has exactly $\mathrm{rk}^{\mathrm{w}} M_i$ nodes at layer i for $0 \leqslant i \leqslant 3$. In particular, such an ABP has three nodes at layer 1, each one computing a polynomial $P_1^{(1)}(X)$, $P_2^{(1)}(X)$ and $P_3^{(1)}(X)$ and has also three nodes at layer 2 which compute the polynomials $P_1^{(2)}(X,Y)$, $P_2^{(2)}(X,Y)$ and $P_3^{(2)}(X,Y)$. The goal is to show that these triplets of polynomials are precisely defined and there is no way to link them together in a weakly monotone ABP. By definition of the Nisan matrix, we can see columns of M as polynomials in $\mathbb{R}[X]$ and columns of N as polynomials in $\mathbb{R}[X,Y]$.

Claim 3. *The polynomials $P_1^{(1)}$, $P_2^{(1)}$ and $P_3^{(1)}$ weakly monotonically generate the columns of M and the polynomials $P_1^{(2)}$, $P_2^{(2)}$ and $P_3^{(2)}$ weakly monotonically generate the columns of N.*

Proof. Let us show the result at layer 1, the case of layer 2 is symmetrical. Consider a column C of the first Nisan matrix: say it corresponds to the coefficient of y_jz_k in H. If we instantiate the variables y_j and z_k to 1 and the other variables from $Y \cup Z$ to 0 in the ABP, we get C

as a linear combination of columns representing $P_1^{(1)}(X)$, $P_2^{(1)}(X)$ and $P_3^{(1)}(X)$. More precisely, $C = \sum_{s=1}^{3} \lambda_s P_s^{(1)}$ where $\lambda_s \neq 0$ if and only if we can read the monomial $y_j z_k$ between the node corresponding to $P_s^{(1)}(X)$ and the output of the ABP.

It remains to show that this linear combination $C = \sum_{s=1}^{3} \lambda_s P_s^{(1)}$ is weakly monotone. Assume this is not, it means for some i the coefficient of x_i is 0 in C but there exists s such that $\lambda_s \neq 0$ and the coefficient of x_i in $P_s^{(1)}$ is different to 0. It means that the coefficient of $x_i y_j z_k$ is 0 in H but there is a path in the ABP with nonzero coefficient for this monomial (otherwise the scalar in front of $P_s^{(1)}(X)$ would be 0). It contradicts the fact that the ABP is weakly monotone. □

Claim 4. *If three vectors U, V and W weakly monotonically generate the family (A, B, C) then (up-to permuting the names of U, V and W), $U \in \mathbb{R}A$,$V \in \mathbb{R}B$ and $W \in \mathbb{R}C$.*

Proof. As $\mathrm{rk}(A, B, C) = 3$, we can consider the vector-space $\mathcal{F}$ generated by $\{A, B, C\}$, namely $\mathcal{F} = \{T \in \mathbb{R}^4 \mid t_1 + t_2 - t_3 - t_4 = 0\}$. So the vectors U, V and W must form a basis of $\mathcal{F}$ and so, have to lie in $\mathcal{F}$. For $1 \leqslant i \leqslant 4$, let $\mathcal{E}_i$ be the i-th coordinate hyperplane of $\mathbb{R}^4$.

Notice that $\mathbb{R}A = \mathcal{F} \cap \mathcal{E}_2 \cap \mathcal{E}_3$. By Proposition 6, since A is a weakly monotone linear combination of U, V, W, (at least) one of the vectors $\{U, V, W\}$ must belong to $\mathcal{E}_2 \cap \mathcal{E}_3$. Since this vector lies also in $\mathcal{F}$, it is in $\mathbb{R}A$.

In the same way, since $\mathbb{R}B = \mathcal{F} \cap \mathcal{E}_1 \cap \mathcal{E}_4$ and $\mathbb{R}C = \mathcal{F} \cap \mathcal{E}_2 \cap \mathcal{E}_4$, one vector of $\mathcal{B}$ must belong to $\mathbb{R}B$ and one must belong to $\mathbb{R}C$.

Since $\mathbb{R}A, \mathbb{R}B, \mathbb{R}C$ are 3 distinct one-dimensional linear subspaces, each one of these spaces has to contain one of the vectors U, V, W. □

Claim 5. *If three vectors Q, R and S weakly monotonically generate the columns of N then, up to permuting the names of Q, R and S, $Q \in \mathbb{R}\binom{A}{C}, R \in \mathbb{R}\binom{B}{C}, S \in \mathbb{R}\binom{C}{D}$.*

Proof. Let us define $\mathcal{B} = \{Q, R, S\}$. We can easily see that the columns of N lie in the vector space given by the intersection of the three hyperplanes

$$\mathcal{F}_1 = \{T \in \mathbb{R}^8 \mid t_1 + t_2 = t_3 + t_4\},$$
$$\mathcal{F}_2 = \{T \in \mathbb{R}^8 \mid t_5 + t_6 = t_7 + t_8\},$$
$$\mathcal{F}_3 = \left\{ T \in \mathbb{R}^8 \mid \sum_{i=1}^{4} t_i = \sum_{j=5}^{8} t_j \right\}.$$

As $\mathrm{rk}(Q, R, S) = \mathrm{rk}\, N$, the vectors Q, R and S are in $\mathcal{F}_1 \cap \mathcal{F}_2 \cap \mathcal{F}_3$.

For $1 \leqslant i \leqslant 8$, we denote by $\mathcal{E}_i$ the i-th coordinate hyperplane of $\mathbb{R}^8$. Looking at the two first columns of N we can notice that

$$\mathbb{R}\binom{A}{C} = \mathcal{F}_1 \cap \mathcal{F}_2 \cap \mathcal{F}_3 \cap \mathcal{E}_2 \cap \mathcal{E}_3 \cap \mathcal{E}_6 \cap \mathcal{E}_8$$

and

$$\mathbb{R}\binom{B}{C} = \mathcal{F}_1 \cap \mathcal{F}_2 \cap \mathcal{F}_3 \cap \mathcal{E}_1 \cap \mathcal{E}_4 \cap \mathcal{E}_6 \cap \mathcal{E}_8$$

are two distinct one-dimensional spaces of $\mathbb{R}^8$. By Proposition 6, there is at least one vector of $\mathcal{B}$ (let us suppose this is Q) such that $Q = q\binom{A}{C}$ and at least one other vector of $\mathcal{B}$ (assume this is R) such that $R = r\binom{B}{C}$.

Finally we need to identify the last vector S. For that, we decompose $S = \binom{S_1}{S_2}$ where S_1 is the projection of S on its first four coordinates and S_2 the projection on the last four. Now, we know that $\mathcal{B}$ weakly monotonically generates the last two columns of N. So there exist $a_1, a_2, a_3, b_1, b_2, b_3 \in \mathbb{R}$ such that:

$$a_1 q \begin{pmatrix} A \\ C \end{pmatrix} + a_2 r \begin{pmatrix} B \\ C \end{pmatrix} + a_3 \begin{pmatrix} S_1 \\ S_2 \end{pmatrix} = \begin{pmatrix} \frac{A+C}{2} \\ \frac{C+D}{2} \end{pmatrix}$$

and

$$b_1 q \begin{pmatrix} A \\ C \end{pmatrix} + b_2 r \begin{pmatrix} B \\ C \end{pmatrix} + b_3 \begin{pmatrix} S_1 \\ S_2 \end{pmatrix} = \begin{pmatrix} \frac{B+C}{2} \\ \frac{C+D}{2} \end{pmatrix}.$$

By Proposition 6, as $\begin{pmatrix} \frac{A+C}{2} \\ \frac{C+D}{2} \end{pmatrix} \in \mathcal{E}_2$ and $\begin{pmatrix} \frac{B+C}{2} \\ \frac{C+D}{2} \end{pmatrix} \in \mathcal{E}_4$, we know that $a_2 = b_1 = 0$. Moreover, as A and $(A + C)$ are not colinear, it means that S_1 belongs to the plane $\mathrm{vect}(A, A + C)$ (by the way, this space is inside $\mathcal{E}_2$). Similarly, $S_1 \in \mathrm{vect}(B, B + C)$. As $B \notin \mathcal{E}_2$, these two planes are distinct, so the intersection is of dimension at most 1. Moreover, $\mathrm{vect}(C)$ is in the intersection, and so, $S_1 \in \mathrm{vect}(C)$. There exists $s \neq 0$ such that $S_1 = sC$. As $a_1 qA + a_3 sC = \frac{A+C}{2}$, it implies that $a_1 q = a_3 s = \frac{1}{2}$. Then, we have $\frac{C}{2} + \frac{S_2}{2s} = \frac{C+D}{2}$, i.e., $S_2 = sD$. □

Consequently, by Claim 3 and Claim 5, one node at layer 2 computes the polynomial whose matrix is $s\binom{C}{D}$ (with $s \neq 0$). By instantiating y_1 to 0 and y_2 to $1/s$, this node computes exactly the polynomial corresponding to D as a weakly monotone linear combination of the nodes at layer 1. By Claim 4, the nodes at layer 1 are polynomials associated to A, B and C (up to scalar multiplication). This would imply that $\mathrm{rk}^{\mathrm{w}}(A, B, C, D) = 3$, which is false. Hence, there does not exist a weakly monotone ABP of size 8.

To complete the proof, we show there is a monotone ABP of size 9 computing H. There are two natural monotone ABPs of size 9, let us describe one of them. One can compute the four polynomials associated to A, B, C and D at the first layer. It gives the following monotone ABP of size 9:

$$H = \frac{1}{2}\begin{pmatrix} x_1 + x_4 & x_2 + x_3 & x_1 + x_3 & x_2 + x_4 \end{pmatrix} \begin{pmatrix} y_1 & 0 & 0 \\ 0 & y_1 & 0 \\ y_2 & y_2 & y_1 \\ 0 & 0 & y_2 \end{pmatrix} \begin{pmatrix} 2z_1 + z_3 \\ 2z_2 + z_4 \\ z_3 + z_4 \end{pmatrix}.$$

□

Theorem 6. *There exists a noncommutative homogeneous degree 3 polynomial P over 4 variables such that $\mathrm{rk}^{+} M_i(P) = \mathrm{rk}^{\mathrm{w}} M_i(P) = \mathrm{rk}\, M_i(P) = 3$ for $i \in \{1, 2\}$, so that $\sum_{i=0}^{3} \mathrm{rk}^{+} M_i(P) = \sum_{i=0}^{3} \mathrm{rk}^{\mathrm{w}} M_i(P) = 8$, but the minimal size of a weakly monotone or monotone ABP computing P is 9.*

Proof. Consider the noncommutative polynomial $P = H(x_1, x_2, x_3, x_4, x_1, x_2, x_1, x_2, x_3, x_4)$. As H is ordered, and as the previous substitution follows this order, it is injective over the set of monomials which appear in H, that is to say, if m_1 and m_2 are two monomials from H which give

the same monomial in P, then $m_1 = m_2$. It directly implies that the substitution establishes a bijection between the set of monomials which appear in H and the ones which appear in P. We will say that this substitution is faithful.

Any ABP A which computes the polynomial H can be transformed into an ABP B which computes P with layers of same size by a direct substitution of the variables. Moreover, if A is monotone, then B is immediately monotone. Then, if A is weakly monotone, the faithfulness property implies that B is also weakly monotone.

In the other direction, if in a weakly monotone noncommutative ABP A computing P we replace the variables x_1 and x_2 in the second layer by y_1 and y_2 and the variables x_1, x_2, x_3 and x_4 in the third layer by z_1, z_2, z_3 and z_4, then we get a new ABP B which computes the polynomial $H(x_1, x_2, x_3, x_4, y_1, y_2, z_1, z_2, z_3, z_4)$. The fact that this transformation preserves the monotonicity is still immediate. The faithfulness property implies it also preserves the weak monotonicity. So, the theorem follows from Lemma 3. □

Corollary 1 (Gap increasing with the degree and the number of variables)**.** *Let P be the polynomial defined in Theorem 6. Let $m, n \geqslant 1$. Let $X_1, \ldots, X_n$ be n sets of distinct variables, with each set of size 4. Let $Q(X_1, \ldots, X_n) = \sum_{j=1}^{n} P^m(X_j)$. This is a polynomial of degree $3m$ in $4n$ variables such that $\sum_{i=0}^{3m} \mathrm{rk}^+ M_i(Q) = \sum_{i=0}^{3m} \mathrm{rk}^{\mathrm{w}} M_i(Q) = 7mn - n + 2$ but the minimal size of a monotone or weakly monotone ABP for it is equal to $8mn - n + 2$.*

Proof. Let us first consider the case $n = 1$. From Theorem 6, one can easily checked that $\mathrm{rk}\, M_i(P^m(X_1)) = 1$ for i multiple of 3 and $\mathrm{rk}\, M_i(P^m(X_1)) = \mathrm{rk}^+ M_i(P^m(X_1)) = 3$ otherwise, and that a minimal (weakly) monotone ABP computing $P^m(X_1)$ has $8m + 1$ nodes.

Consider a weakly monotone ABP for Q. Assume there is an internal node α and two distinct indices k and k' such that α depends on at least one variable of X_k and one variable of $X_{k'}$. Consequently, one path of the ABP produces a monomial which contains both a variable in X_k and a variable not in X_k. Since $Q = \sum_{j=1}^{n} P^m(X_j)$, a given monomial in Q can only contain variables coming from a single X_k. The above statement thus contradicts the fact that the ABP is weakly monotone. Hence, we can partition the internal nodes of the ABP into n parts, each one related to one variable set X_j. As mentioned earlier, a minimal weakly monotone ABP for $P^m(X_j)$ has $8m - 1$ internal nodes. The minimal size of a weakly monotone ABP is therefore $8mn - n + 2$. The same is true of a monotone ABP computing Q.

Let us compute the sum of ranks for Q. If $0 < i < 3m$, the i-th Nisan matrix of Q is block-diagonal with n blocks, where the j-th block corresponds to the i-th Nisan matrix of $P^m(X_j)$. As the nonnegative rank of a block-diagonal matrix is equal to the the sum of the nonnegative ranks of its blocks, $\mathrm{rk}\, M_i(Q) = \mathrm{rk}^+ M_i(Q) = n$ for $i \in \{3, 6, \ldots, 3m-3\}$ and $\mathrm{rk}\, M_i(Q) = \mathrm{rk}^+ M_i(Q) = 3n$ for $i \equiv 1, 2 \mod 3$. Summing over the different layers we get that the sum-of-ranks measure for Q, both for usual rank and nonnegative rank, and thus for weakly nonnegative rank, is equal to $7mn - n + 2$. □

An upper bound on the size of a monotone ABP computing a homogeneous degree d polynomial P is obtained by summing, for each $\ell \in \{0, \ldots, d\}$ the minimal number of rows extracted from $M_\ell(P)$ whose cone contains all other columns of $M_\ell(P)$. The example above shows that this is not a characterization of monotone size: for the polynomial H built in Lemma 3, it is needed to extract 4 rows in both $M_1(H)$ and $M_2(H)$. The same remark applies in the weakly monotone setting (about the minimum number of extracted rows which weakly monotonically generate all the rows).

4 Lower bounds for monotone commutative ABPs

4.1 Lower bound tools for monotone and weakly monotone ABPs

Consider a homogeneous degree d commutative polynomial P. For $\ell \in \{0, \dots, d\}$, we define the set $\mathcal{M}_\ell(P)$ of matrices, whose rows are indexed by commutative degree-ℓ monomials and whose columns indexed by degree-$(d-\ell)$ commutative monomials. A matrix M belongs to $\mathcal{M}_\ell(P)$ if:

(a) For any degree d commutative monomial m such that m does not appear in P and any (m_1, m_2) satisfying $m = m_1 m_2$, m_1 of degree ℓ and m_2 of degree $d-\ell$, we have $M_{m_1,m_2} = 0$;

(b) For any other degree d commutative monomial m, $\sum_{m_1 m_2 = m} M_{m_1,m_2}$ is equal to the coefficient of m in P.

For a matrix M whose rows and columns are indexed by noncommutative monomials, we define M^{com} the matrix obtained by summing rows and columns indexed by the same *commutative* monomial.

Proposition 7. *A homogeneous degree-d noncommutative polynomial Q computes commutatively P without cancelling monomials if and only if $M_\ell(Q)^{com} \in \mathcal{M}_\ell(P)$ for all $\ell \in \{0, \dots, d\}$.*

Proof. The polynomial Q computes commutatively P if and only if, for each ℓ, the matrix $M := M_\ell(Q)^{com}$ satisfies the following: for any degree d commutative monomial m, $\sum_{m_1 m_2 = m} M_{m_1,m_2}$ is equal to the coefficient of m in P.

The polynomial Q does not cancel monomials if and only if, for all monomial m not appearing in P and for all decomposition $m = m_1 m_2$, there is no noncommutative monomial $m' = m'_1 m'_2$ in Q such that m'_i computes commutatively m_i for $i \in \{1, 2\}$.

Together, these two statements prove the proposition. □

For a homogeneous degree d polynomial P and $\ell \in \{0, \dots, d\}$ consider the support matrix $S_\ell(P)$ indexed by degree-ℓ commutative monomials on the rows, degree-$(d-\ell)$ commutative monomials on the column, such that $S_\ell(P)_{m_1,m_2} = 1$ if the coefficient of $m_1 m_2$ in P is nonzero and $S_\ell(P)_{m_1,m_2} = 0$ otherwise.

Definition 8. *For M, S two matrices of the same size we define $\mathrm{rk}^{\mathrm{w}}(M, S)$ to be the smallest r such that there exist rank 1 matrices $M_1, \dots, M_r$ such that $\mathrm{supp}(M_i) \subseteq \mathrm{supp}(S)$ and $M = \sum_{i=1}^r M_i$.*

Notice that $\mathrm{rk}^{\mathrm{w}} M$, defined in Section 2, is nothing but $\mathrm{rk}^{\mathrm{w}}(M, M)$.

Theorem 7. *The size of a monotone ABP computing a homogeneous commutative polynomial P of degree d is at least $\sum_{\ell=0}^d \min\{\mathrm{rk}^+ M \mid M \in \mathcal{M}_\ell(P),\ M \geqslant 0\}$. If the ABP is weakly monotone the bound becomes $\sum_{\ell=0}^d \min\{\mathrm{rk}^{\mathrm{w}}(M, S_\ell(P)) \mid M \in \mathcal{M}_\ell(P)\}$.*

Proof. Let $\ell \in \{1, \dots, d-1\}$. Consider an ABP computing P with minimal number of nodes at level ℓ: say it is w. Cutting this ABP at layer ℓ gives a decomposition $P = \sum_{i=1}^w Q_i R_i$. For $i \in \{1, \dots, w\}$ let M_i be the matrix of $Q_i R_i$. All matrices M_i are of rank 1 and we have $\sum_{i=1}^w M_i \in \mathcal{M}_\ell(P)$. If the ABP is monotone, the matrices M_i are nonnegative and we get $\min\{\mathrm{rk}^+ M \mid M \in \mathcal{M}_\ell(P),\ M \geqslant 0\} \leqslant w$. If the ABP is weakly monotone, we have $\mathrm{supp}(M_i) \subseteq \mathrm{supp}(S_\ell(P))$. Hence $\min\{\mathrm{rk}^{\mathrm{w}}(M, S_\ell(P)) \mid M \in \mathcal{M}_\ell(P)\} \leqslant w$. □

For two same-sized matrices M, S, let $\mathrm{cov}(M, S)$ be the smallest number of combinatorial rectangles included in the support of S and whose union covers the support of M.

Proposition 8. $\mathrm{cov}(M, S) \leqslant \mathrm{rk}^{\mathrm{w}}(M, S)$.

Proof. Let $r = \mathrm{rk}^{\mathrm{w}}(M, S)$ and write $M = \sum_{i=1}^{r} M_i$ with M_i of rank 1, $\mathrm{supp}(M_i) \subseteq \mathrm{supp}(S)$. We have $\mathrm{supp}(M) \subseteq \bigcup_{i=1}^{r} \mathrm{supp}(M_i)$: this shows that $\mathrm{cov}(M, S) \leqslant r$. □

Corollary 2. *Any weakly monotone ABP computing P has size greater or equal to*

$$\sum_{\ell=0}^{d} \min\{\mathrm{cov}(M, S_\ell(P)) \mid M \in \mathcal{M}_\ell(P)\}.$$

4.2 Application to the elementary symmetric polynomials

For n positive integer we write $[n] = \{1, \ldots, n\}$. For $0 \leqslant k \leqslant n$, let $e_{n,k}$ be the elementary symmetric polynomial of degree k over the variables $x_1, \ldots, x_n$: $e_{n,k} = \sum_{I \in \binom{[n]}{k}} \prod_{i \in I} x_i$. Notice that $S_j(e_{n,k})$ is exactly the disjointness matrix $D_{n,j,k-j}$ with rows indexed by elements of $\binom{[n]}{j}$ and columns indexed by elements of $\binom{[n]}{k-j}$, and whose entry in row A and column B is 1 if $A \cap B = \emptyset$ and 0 otherwise.

To get lower bounds for $e_{n,k}$ using Corollary 2 we need to show that, for enough values of j and for any $M \in \mathcal{M}_j(e_{n,k})$, $\mathrm{cov}(M, D_{n,j,k-j})$ is large.

Proposition 9. *For n, j, k fixed, assume $\mathrm{cov}(M, D_{n,j,k-j}) \leqslant m$ for some $M \in \mathcal{M}_j(e_{n,k})$. Then there exists $A_1, \ldots, A_m \subseteq [n]$ with the following property:*

$$\text{For all } B \in \binom{[n]}{k}, \text{ there is } i \in \{1, \ldots, m\} \text{ such that } |A_i \cap B| = j. \tag{1}$$

Proof. Let $M \in \mathcal{M}_j(e_{n,k})$. Assume $U_1 \times V_1, \ldots, U_m \times V_m$ is a set of combinatorial rectangles from $\binom{[n]}{j} \times \binom{[n]}{k-j}$ included in the support of $D_{n,j,k-j}$ and covering $\mathrm{supp}\, M$. Notice that such a combinatorial rectangle $U \times V$ is included in the support of $D_{n,j,k-j}$ if and only if $\left(\bigcup_{u \in U} u\right) \cap \left(\bigcup_{v \in V} v\right) = \emptyset$. For $i \in \{1, \ldots, m\}$, let $A_i = \bigcup_{u \in U_i} u$. From the previous remark the set of combinatorial rectangles $R_1, \ldots, R_m$ defined by $R_i = \binom{A_i}{j} \times \binom{[n] \setminus A_i}{k-j}$ is included in the support of $D_{n,j,k-j}$ and covers $\mathrm{supp}\, M$.

Let us show that the family $\{A_1, \ldots, A_m\}$ satisfies Equation (1). Let $B \in \binom{[n]}{k}$. The monomial $\prod_{i \in B} x_i$ appears in $e_{n,k}$ so one non-zero entry of M is of the form (I, J) with $I \in \binom{[n]}{j}$, $J \in \binom{[n]}{k-j}$ and $I \cup J = B$. Therefore $(I, J) \in R_i$ for some $i \in \{1, \ldots, m\}$, i.e. $|A_i \cap B| = |I| = j$. □

We will now relate our lower bound endeavor to a combinatorial question known as Galvin's problem: for n a multiple of 4, prove a lower bound on the size m of a family $\{A_1, \ldots, A_m\} \subseteq \binom{[n]}{n/2}$ such that for any $B \in \binom{[n]}{n/2}$, there exists i such that $|A_i \cap B| = n/4$. Proving a lower bound on a family $\{A_1, \ldots, A_m\}$ satisfying Equation (1) for the parameters $k = n/2$ and $j = n/4$ is a generalization of Galvin's problem because the sets A_i can be of arbitrary size, instead of $n/2$ in the original problem.

We first give a lower bound for the middle elementary symmetric polynomial. The argument is similar to the solution of Galvin's original problem presented in [12, Theorem 11.1], which we reproduce here for completeness. It is based on the following result, restricted here to the case of codes over an alphabet with 2 elements (we denote by Δ the symmetric difference between two sets):

Theorem 8 ([12], Theorem 1.10)**.** *Suppose $0 < \delta < \frac{1}{2}$ is given. Then there exists $\varepsilon > 0$ such that for any d even satisfying $\delta n < d < (1-\delta)n$, any family of distinct subsets $C_1, \dots, C_m \subseteq [n]$ such that, for all $i \neq j$, $|C_i \Delta C_j| \neq d$, has size $m \leqslant (2-\varepsilon)^n$.*

Lemma 4. *There exists $\alpha > 0$ such that for $n \in 4\mathbb{N} \setminus \{0\}$, $k = n/2$ and j odd, any family $\{A_1, \dots, A_m\}$ satisfying Equation (1) has size $m \geqslant \alpha n$.*

Proof. Assume there exists $A_1, \dots, A_m \subseteq [n]$ such that $\mathcal{F} = \{A_1, \dots, A_m\}$ satisfies Equation (1). Let V be the subspace of $\mathbb{F}_2^n$ spanned by the characteristic vectors of the elements of $\mathcal{F}$. By assumption, for all $B \in \binom{[n]}{n/2}$, there exists $F \in \mathcal{F}$ such that $|B \cap F| = j$; this means that $\langle \chi(B), \chi(F) \rangle = 1 \neq 0$ because j is odd. Hence $V^\perp$ contains no vector of weight $n/2$. Because $V^\perp$ is a vector space, it implies that for any $C, D \subseteq [n]$ such that $\chi(C), \chi(D) \in V^\perp$, $|C \Delta D| \neq n/2$.

By Theorem 8, $|V^\perp| \leqslant (2-\varepsilon)^n$ for some constant $\varepsilon > 0$. This means that $\dim V^\perp \leqslant (1-\alpha)n$ for some $\alpha > 0$. It follows that $m = |\mathcal{F}| \geqslant \dim V \geqslant \alpha n$. □

Lemma 5. *For $n \in 4\mathbb{N}$, every weakly monotone ABP computing $e_{n,n/2}$ has size $\Omega(n^2)$.*

Proof. There exists $\alpha > 0$ such that for $n \in 4\mathbb{N}$, $k = n/2$ and j odd, any family $\{A_1, \dots, A_m\}$ satisfying Equation (1) has size $m \geqslant \alpha n$ by Lemma 4. It follows from Proposition 9 that for all $M \in \mathcal{M}_j(e_{n,n/2})$, $\mathrm{cov}(M, D_{n,j,n/2-j}) \geqslant \alpha n$. The lower bound is obtained by Corollary 2. □

From the simple observation $e_{n,k}(x_1, \dots, x_m, 0, \dots, 0) = e_{m,k}(x_1, \dots, x_m)$, Lemma 5 yields quadratic lower bounds on the size of weakly monotone ABPs computing $e_{n,k}$ for $\delta n \leqslant k \leqslant n/2$ for a fixed $\delta > 0$. However we need to be more careful to get a quadratic lower bound for e.g. $e_{n,2n/3}$. Indeed the simple reduction

$$e_{n,k}(x_1, \dots, x_n) = \prod_{i=1}^{n} x_i \cdot e_{n,n-k}\left(\frac{1}{x_1}, \dots, \frac{1}{x_n}\right)$$

uses divisions, which are not allowed in our model and would cost too much to remove.

In an ABP, the formal degree $\mathrm{fdeg}_t(\alpha)$ of a node α with respect to a variable t is defined as the maximum degree in t of the polynomial computed along a path from the source to α, which is also the maximal degree in t of a monomial produced along a path from the source to α. By definition, the formal degree of the source is 0. Let us denote by $\hat{\alpha}$ the polynomial computed at the node α. Remark that $\mathrm{fdeg}_t(\alpha) \geqslant \deg_t(\hat{\alpha})$. The formal degree in t of an ABP is the formal degree in t of its output.

Let us show now that we can always extract the part of maximal formal degree without changing the size of an ABP. We denote by $[t^k]f$ the coefficient of the homogeneous component of f of degree k in t.

Lemma 6. *Let A be an ABP of size s and of formal degree k in the variable t computing a polynomial f. Then there exists A' an ABP of size at most s such that A' computes $[t^k]f$.*

Moreover, if A is weakly monotone, then it is also the case for A'.

Proof. We show by induction that we can construct A' such that for every node α in A, there is α' in A' which computes the coefficient of the homogeneous part of $\hat{\alpha}$ of degree $\mathrm{fdeg}(\alpha)$. That is, α' computes $[t^{\mathrm{fdeg}(\alpha)}]\hat{\alpha}$.

The result is immediate for an ABP with only one node since it computes 0. Let us consider a node $\alpha = \ell_1\beta_1 + \ldots + \ell_m\beta_m$ where $\beta_1, \ldots, \beta_m$ are predecessors of the node α in A and where $\ell_1, \ldots, \ell_m$ are linear forms. Let us denote by I and J the subsets of indices:

$$I = \{i \in [m] \mid \mathrm{fdeg}(\beta_i) + 1 = \mathrm{fdeg}(\alpha) \text{ and } [t]\ell_i \neq 0\}$$

and

$$J = \{i \in [m] \mid \mathrm{fdeg}(\beta_i) = \mathrm{fdeg}(\alpha) \text{ and } [t]\ell_i = 0\}.$$

Then,

$$\begin{aligned}[t^{\mathrm{fdeg}(\alpha)}]\hat{\alpha} &= \sum_{i\in I}[t^{\mathrm{fdeg}(\alpha)}]\ell_i\hat{\beta}_i + \sum_{i\in J}[t^{\mathrm{fdeg}(\alpha)}]\ell_i\hat{\beta}_i \\ &= \sum_{i\in I}([t]\ell_i)\,[t^{\mathrm{fdeg}(\beta)}]\hat{\beta}_i + \sum_{i\in J}\ell_i[t^{\mathrm{fdeg}(\beta)}]\hat{\beta}_i.\end{aligned}$$

So we just have to define

$$\alpha' \stackrel{\mathrm{def}}{=} \sum_{i\in I}([t]\ell_i)\beta'_i + \sum_{i\in J}\ell_i\beta'_i.$$

The second point is a consequence of the fact that if some monomial m cancels in A', the monomial $t^k m$ cancels in A. □

Lemma 7. *If there is a weakly monotone ABP of size s computing the polynomial $e_{n,p}$, then for all $q \leqslant p$, there is a weakly monotone ABP of size at most s which computes the polynomial $e_{n-q,p-q}$.*

Proof. Let us replace the variables $x_{n-q+1}, \ldots, x_n$ in the weakly monotone ABP A computing $e_{n,k}$ by the variable t. As the monomial $x_1x_2\cdots x_{p-q}x_{n-q+1}\cdots x_n$ appears in $e_{n,p}$ and gives a monomial of degree q with respect to t, the formal degree in t of our new ABP is exactly q. As we work over a field of characteristic 0, all monomials of the form $t^\ell \prod_{i\in I_\ell} x_i$ (for some $0 \leqslant \ell \leqslant q$) with $I_\ell \subseteq [n-q]$ of size $p-\ell$ appear in $e_{n,p}(x_1, \ldots, x_{n-q}, t, \ldots, t)$. It implies that the ABP we get after substitution is still weakly monotone.

By Lemma 6, there is a weakly monotone ABP of size at most s which computes

$$[t^q]e_{n,p}(x_1, \ldots, x_{n-q}, t, \ldots, t) = e_{n-q,p-q}.$$

□

Theorem 9. *Every weakly monotone ABP, or equivalently every homogeneous syntactically multilinear ABP, computing $e_{n,k}$ has size $\Omega(\min\{k^2, (n-k)^2\})$.*

Proof. Let us first prove the lower bound when n and k are even.

If $k \leqslant n/2$, then as mentioned previously, any weakly monotone ABP of size s implies a weakly monotone ABP of size at most s for $e_{2k,k}$ by putting some variables to 0. So in this case $s = \Omega(k^2)$ by Lemma 5.

Otherwise, we have $k > n/2$. It means that $k \geqslant 2k - n > 0$. Then a weakly monotone ABP of size s for $e_{n,k}$ gives a weakly monotone ABP of size at most s for $e_{2n-2k,n-k}$ by Lemma 7, choosing the parameters $p = k$ and $q = 2k - n$. The lower bound $s = \Omega((n-k)^2)$ follows from Lemma 5.

The lower bound is obtained for n odd by noticing that $e_{2\lfloor n/2 \rfloor,k}$ can be reduced to $e_{n,k}$ by putting one variable to 0. Moreover, $e_{n,k}$ reduces to $e_{n-1,k-1}$ by Lemma 7. So the lower bound holds for n and k of any parity.

This lower bound also holds in the homogeneous syntactically multilinear model: indeed, any such ABP computing $e_{n,k}$ is weakly monotone because $e_{n,k}$ has all degree k monomials in its support. □

References

[1] Noga Alon, Mrinal Kumar, and Ben Lee Volk. Unbalancing sets and an almost quadratic lower bound for syntactically multilinear arithmetic circuits. In *33rd Computational Complexity Conference, CCC 2018, June 22-24, 2018, San Diego, CA, USA*, pages 11:1–11:16, 2018. URL: https://doi.org/10.4230/LIPIcs.CCC.2018.11, doi:10.4230/LIPIcs.CCC.2018.11.

[2] Walter Baur and Volker Strassen. The complexity of partial derivatives. *Theoretical Computer Science*, 22(3):317 – 330, 1983. URL: http://www.sciencedirect.com/science/article/pii/030439758390110X, doi:https://doi.org/10.1016/0304-3975(83)90110-X.

[3] LeRoy B. Beasley and Thomas J. Laffey. Real rank versus nonnegative rank. *Linear Algebra Appl.*, 431(12):2330–2335, 2009. URL: https://doi.org/10.1016/j.laa.2009.02.034, doi:10.1016/j.laa.2009.02.034.

[4] Jean Berstel, Jr. and Christophe Reutenauer. *Rational Series and Their Languages*. Springer-Verlag, Berlin, Heidelberg, 1988.

[5] Symeon Bozapalidis and Olympia Louscou-Bozapalidou. The rank of a formal tree power series. *Theoretical Computer Science*, 27(1):211 – 215, 1983. URL: http://www.sciencedirect.com/science/article/pii/0304397583901007, doi:https://doi.org/10.1016/0304-3975(83)90100-7.

[6] Joel E Cohen and Uriel G Rothblum. Nonnegative ranks, decompositions, and factorizations of nonnegative matrices. *Linear Algebra and its Applications*, 190:149–168, 1993.

[7] Ronald de Wolf. Nondeterministic quantum query and communication complexities. *SIAM J. Comput.*, 32(3):681–699, 2003. URL: https://doi.org/10.1137/S0097539702407345, doi:10.1137/S0097539702407345.

[8] Nathanaël Fijalkow, Guillaume Lagarde, Pierre Ohlmann, and Olivier Serre. Lower bounds for arithmetic circuits via the hankel matrix. *Electronic Colloquium on Computational Complexity (ECCC)*, 25:180, 2018. URL: https://eccc.weizmann.ac.il/report/2018/180.

[9] Samuel Fiorini, Serge Massar, Sebastian Pokutta, Hans Raj Tiwary, and Ronald de Wolf. Exponential lower bounds for polytopes in combinatorial optimization. *J. ACM*, 62(2):17:1–17:23, 2015. URL: http://doi.acm.org/10.1145/2716307, doi:10.1145/2716307.

[10] Samuel Fiorini, Thomas Rothvoß, and Hans Raj Tiwary. Extended formulations for polygons. *Discrete & Computational Geometry*, 48(3):658–668, 2012. URL: https://doi.org/10.1007/s00454-012-9421-9, doi:10.1007/s00454-012-9421-9.

[11] Michel Fliess. Matrices de Hankel. *J. Math. Pures Appl. (9)*, 53:197–222, 1974.

[12] Peter Frankl and Vojtěch Rödl. Forbidden intersections. *Trans. Amer. Math. Soc.*, 300(1):259–286, 1987. URL: https://doi.org/10.2307/2000598, doi:10.2307/2000598.

[13] Pavel Hrubes and Amir Yehudayoff. Homogeneous formulas and symmetric polynomials. *Computational Complexity*, 20(3):559–578, 2011. URL: https://doi.org/10.1007/s00037-011-0007-3, doi:10.1007/s00037-011-0007-3.

[14] Pavel Hrubes and Amir Yehudayoff. Formulas are exponentially stronger than monotone circuits in non-commutative setting. In *Proceedings of the 28th Conference on Computational Complexity, CCC 2013, K.lo Alto, California, USA, 5-7 June, 2013*, pages 10–14, 2013. URL: https://doi.org/10.1109/CCC.2013.11, doi:10.1109/CCC.2013.11.

[15] Mark Jerrum and Marc Snir. Some exact complexity results for straight-line computations over semirings. *J. ACM*, 29(3):874–897, July 1982. URL: http://doi.acm.org/10.1145/322326.322341, doi:10.1145/322326.322341.

[16] Stasys Jukna and Georg Schnitger. On the optimality of Bellman-Ford-Moore shortest path algorithm. *Theor. Comput. Sci.*, 628:101–109, 2016. URL: https://doi.org/10.1016/j.tcs.2016.03.014, doi:10.1016/j.tcs.2016.03.014.

[17] Adam Klivans and Amir Shpilka. Learning restricted models of arithmetic circuits. *Theory of Computing*, 2(10):185–206, 2006. URL: http://www.theoryofcomputing.org/articles/v002a010, doi:10.4086/toc.2006.v002a010.

[18] Mrinal Kumar and Shubhangi Saraf. On the power of homogeneous depth 4 arithmetic circuits. *SIAM J. Comput.*, 46(1):336–387, 2017. URL: https://doi.org/10.1137/140999335, doi:10.1137/140999335.

[19] Noam Nisan. Lower bounds for non-commutative computation (extended abstract). In *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing, May 5-8, 1991, New Orleans, Louisiana, USA*, pages 410–418, 1991. URL: https://doi.org/10.1145/103418.103462, doi:10.1145/103418.103462.

[20] Noam Nisan and Avi Wigderson. Lower bounds on arithmetic circuits via partial derivatives. *Comput. Complexity*, 6(3):217–234, 1996/97. URL: https://doi.org/10.1007/BF01294256, doi:10.1007/BF01294256.

[21] Ran Raz. Multi-linear formulas for permanent and determinant are of super-polynomial size. *J. ACM*, 56(2):8:1–8:17, 2009. URL: https://doi.org/10.1145/1502793.1502797, doi:10.1145/1502793.1502797.

[22] C.P. Schnorr. A lower bound on the number of additions in monotone computations. *Theoretical Computer Science*, 2(3):305 – 315, 1976. URL: http://www.sciencedirect.com/science/article/pii/0304397576900839, doi:https://doi.org/10.1016/0304-3975(76)90083-9.

A From nonhomogeneous to homogeneous ABPs

The nonhomogeneous model described in the introduction can be shown to be equivalent (without changing the size) to one where the vertices are not layered, but still with linear forms on the edges except one possible additional scalar edge from the source to the sink.

In the noncommutative case, this second model can be seen as a special case of *linear representations* of word series as in [11], see also [4]. The basic result there is that the minimal size of a linear representation is the rank of the so-called Hankel matrix. We can take from the proof of this theorem the construction of a minimal-size linear representation from the Hankel matrix. In the case of a noncommutative homogeneous polynomial, this linear representation will be exactly an ABP in the sense of Definition 1.

In the weakly monotone case, when computing a homogeneous polynomial, each node has to compute a homogeneous polynomial. So this second model can be layered to conform to Definition 1. Note that there cannot be a scalar edge from source to sink in this case.