Types of Depth and Formula Size

TYPES OF DEPTH AND FORMULA SIZE

K. Kalorkoti
School of Informatics,
University of Edinburgh,
10 Crichton Street,
Edinburgh EH8 9LE, U.K.
kk@inf.ed.ac.uk

May 14, 2014

Abstract

We use a rank-based measure on rational expressions in indeterminates over a field and define notions of size and depth with associated subparts of formulae for expressions. Formulae are allowed to have as inputs expressions from a large set rather than just constants and indeterminates. A general lower bound is derived and this is used to deduce an exponential lower bound, subject to depth assumptions, on the formula size of the determinant with inputs restricted to the usual constants and indeterminates. The general bound is also used to show that a polynomial which is closely related to the determinant has exponential formula size if either (i) some types of operations do not occur in the formula or (ii) some assumptions on depth hold (the inputs allowed here are from a large set).

Keywords: Formula size, depth, determinant.
AMS Subject Classification: 03D15

1 Introduction

Algebraic complexity seeks to classify the computational cost of building objects by means of arithmetic operations ($+$, $-$, $\times$, and optionally $/$) starting with constants and indeterminates as inputs. In general once a subresult is obtained it can be used as many times as needed in subsequent calculations; this is the *circuit* model of computing. However there are good reasons to consider a restricted model in which subresults can be used only once, such a model corresponds to the notion of a *formula*. For example, the depth of a formula (defined below) corresponds to parallel computation time. Despite the apparent simplicity of these models, especially the latter, we still do not have strong lower bounds for such important objects as the determinant (see below for more details). In this paper we give a detailed classification of certain notions of size and depth for a formula and use them to show exponential lower bounds but under some assumptions on depth. Bounds are expressed, as usual, in terms of the number of input indeterminates.

Let k be a field and X a non-empty set of indeterminates over k. Normally one considers formulae over $k \cup X$, i.e., trees whose leaves are labeled by members of $k \cup X$ and whose non-leaf vertices are labeled by an operation from $\{+, -, \times\}$; divisions are sometimes also allowed but will not be in this paper. In fact for most measures multiplication by scalars is free and so subtraction can be left out as an operation, this makes little difference to the arguments of this paper. Leaving out division is not quite so straigthtforward. The arguments of Brent [3] and Strassen [10] can be used to show that for formulae over sufficiently large fields we can remove divisions at the cost of a polynomial increase in size; see the survey by Shpilka and Yehudayoff [9] for further details. However, to date, the largest unrestricted lower bound we have for the formula size of an explicit expression is only quadratic, see Kalorkoti [6], and it would surely be of interest to improve on this to some higher power.

In this paper the available inputs are extended from $k \cup X$ to a larger set for all results except for Theorem 2.1, see the final paragraph of this section for details. From now on we will use 'vertex' to mean a non-leaf of the tree. Edges are directed from the leaves towards the root. If (u, v) is a directed edge from u to v then we call u a *child* of v. Each vertex v has two children, a *left* child v_L and a *right* child v_R; thus (v_L, v) and (v_R, v) are edges. Paths always go from parent to child, i.e., consistently against the direction of edges. In a formula (whether standard or as extended later in this paper) each vertex v is labelled by an arithmetic operation and is also called a *computation vertex*. If v is labeled by $+$ or $-$ we call it *additive* otherwise *multiplicative*. Normally formulae are defined textually and it is observed that they correspond to trees, there is no harm in identifying them with their corresponding trees; using textual or graphical notation as is most convenient. At a vertex v the order of arguments for evaluation is the result of v_L followed by the result of v_R (this matters only when the operation is $-$ or $/$, when the latter operation is allowed).

We use $r(\Phi)$ to denote the result computed by a formula Φ; if the inputs are from $k \cup X$ then $r(\Phi)$ is a polynomial. (Note that on occasion we use r as an integer variable, however there can be no confusion since $r()$ is always used with an argument when denoting the result of a formula.) The *size* of Φ, denoted by $|\Phi|$ is the number of vertices, i.e., the number of operations. The *depth* of Φ, denoted by $d(\Phi)$, is the maximum number of computation vertices from the root of Φ to any leaf. Other notions of size and depth will be defined in §4. Note that in this paper we deal only with fan in 2 rather than the unbounded case, see [9] for further details on this.

In order to provide some context for the analysis presented in this paper, consider a polynomial $f = \sum_\nu a_\nu x^\nu$ where the $a_\nu \in k$ are constant coefficients and the x^ν are finitely many distinct power products, i.e., expressions of the form $x_1^{\nu_1} \cdots x_n^{\nu_n}$ where $\nu_i \in \mathbb{N}$, for $1 \leq i \leq n$ (the notation is discussed in the next section). We can build a formula Φ_f for f by using a balanced binary tree with multiplicative vertices to compute each power product, then multiplying each tree with the relevant constant to obtain the corresponding monomial and finally adding up all the monomials using a balanced tree. While this is a naive approach it has the interesting property that it is optimal with respect to the depth of non-scalar multiplications (this follows from a simple degree argument). On the other hand, with this approach, there must be at least as many multiplicative subtrees as there are monomials whereas a linear size formula such as $(z_1 + y_1) \cdots (z_n + y_n)$ has 2^n distinct power products just by allowing simple additive trees of depth 1 (i.e., the formuale $z_i + y_i$) to feed into the leaves of the single multiplicative tree.

The general analysis we provide applies to arbitrary formulae; identifying certain types of subformulae and associated notions of size and depth, motivated by the properties of the rank-based measure introduced in §3. There is a well known and widely used distinction between *scalar* and *non-scalar* multiplicative operations, the latter being multiplications by non-constants (and divisions by non-constants when / is allowed). The main contribution of this paper is to examine a finer subdivision of operations and perhaps point to ways in which further research can exploit them in order to relax the assumptions made for the lower bounds. In essence the method used is an improvement on an approach based on monomial counting; the latter is not powerful enough for the results given here. In terms of the previous paragraph, our specific lower bounds allow the depth of non-scalar multiplications to be a little more than the minimum necessary (up to an additive constant more) and layers of additive vertices to interleave with the multiplications, but with a restriction on the depth of each layer.

For the general setting, we choose $Y \subseteq X$ and set $Z = X - Y$. Apart from §2, our formulae will be extended in such a way as to compute elements of $k(Z)[[Y]]$, i.e., the ring of formal power series in Y with coefficients from $k(Z)$ (see Zarsiki and Samuel [12]). Note that we are working entirely with algebraic objects rather than functions. See [6] for a discussion of the relation between formula size in the algebraic setting as compared to the functional one. It must be stressed that the model does not compute all elements of $k(Z)[[Y]]$ but it does compute all elements of $k(Z)[Y]$ and hence all elements of $k[X]$, details are given in §4. For the general results, our formulae are allowed to have leaves labeled with members of $k(Z) \cup k[[Y]]$ rather than just $k \cup X$. Our aim is to deduce lower bounds so that this extension simply makes the bounds more powerful.

The use of formal power series is not necessary for the main results. However the proofs would

not be simplified by their removal. In the presence of formal power series, it is an easy matter to extend the results to include division but at the cost of some minor complications. This has been left out of the paper because the main application results would not go through with divisions allowed owing to the assumptions on depth. The use of formal power series makes available a greater range of possible transformations of formulae that could prove useful in deriving lower bounds by means of reductions.

2 Statement of the main result

The *non-scalar depth* of a formula Φ, denoted by $\mu(\Phi)$, is the maximum number of non-scalar multiplicative vertices on any path from the root to a leaf. The *non-scalar additive gap* of Φ, denoted by $\gamma(\Phi)$, is the maximum number of additive vertices between two consecutive non-scalar vertices over all paths from the root to a leaf or between the last non-scalar vertex of a path and a leaf (scalar multiplicative vertices are ignored). Note that if the root is additive or scalar then consecutive additive and scalar vertices starting at the root do not contribute to the non-scalar additive gap. It follows that $d(\Phi)$ cannot be bounded from above in terms of $\mu(\Phi)$ and $\gamma(\Phi)$. This fact is illustrated by the formula Φ_f for $f = \sum_\nu a_\nu x^\nu$ discussed in §1. Take f to be the determinant of an $n \times n$ matrix of distinct indeterminates, then $\mu(\Phi_f) = \lceil \lg n \rceil$, $\gamma(\Phi_f) = 0$ but $d(\Phi_f) \geq \lg(n!) \geq n \lg(n/e)$; we use lg for $\log_2$ throughout the paper.

THEOREM 2.1 *Let $M = (x_{ij})$ be an $n \times n$ matrix where the x_{ij} are distinct indeterminates over k, for $1 \leq i, j \leq n$. Let Φ be a formula for* $\det M$ *with leaves labeled by members of $k \cup X$. Suppose that $\mu(\Phi) \leq \lg c_1 n$ and $\gamma(\Phi) \leq c_2 \lg n + o(\lg n)$ where $c_1 c_2 < 2/3$. Then $|\Phi| \geq n^{\Omega(n)}$.*

The result applies to the permanent as well. The assumptions of the preceding result are quite strong but it is worth noting that, even so, the overall depth of a formula allowed by the assumptions can be as high as $c_2 \lg^2 n + o(\lg^2 n)$. Naturally $c_1 \geq 1$, since the degree of the result of Φ is at most $2^{\mu(\phi)}$, and thus $c_2 < 2/3$. Since a formula of size $n^{\Omega(n)}$ must have depth at least $\Omega(n \lg n)$, Theorem 2.1 shows that most of the depth of the formula consists of operations that do not contribute to the non-scalar depth or the additive gap.

See Bshouty, Cleve and Eberley [4] on size-depth trade offs. As noted in [4], the classical result of Brent [3] implies that if a formula has size S then it can be transformed into one of size $S^{O(1)}$ and depth $O(\lg S)$. Thus if the determinant has a polynomial size formula then it has one with depth $O(\lg n)$. There is therefore good reason to examine formulae with depth $O(\lg n)$; the results here allow us to go to depth proportional to $\lg^2 n$ but with strong restrictions.

The best known upper bound for the formula size of the determinant is $n^{O(\lg n)}$ obtained by Csanky [5] for fields of characteristic 0. The same bound was obtained for all fields by Borodin, von zur Gathen and Hopcroft [2]. It follows that formulae which satisfy the constraints of Theorem 2.1 cannot be optimal.

Raz [7] proves a lower bound of $n^{\Omega(\lg n)}$ for multilinear formulae for the determinant. The key underlying feature of multilinear formulae is that at each multiplication vertex the two subtrees have disjoint sets of indeterminates which is likely to be a significant restriction. Potentially, it puts quite strong constraints on the 'garbage collection' ability of algebraic computation, see Valiant [11]. The largest unrestricted lower bound is $\Omega(n^3)$ due to the author [6]. Shpilka and Wigderson [8] prove a lower bound of $\Omega(n^4/\lg n)$ for $\Sigma\Pi\Sigma$ circuits (these consists of a layer of additive vertices, then a layer of multiplicative ones and another additive layer). Finally, Agrawal and Vinay [1] show that if we allow unbounded fan in and are interested in proving exponential lower bounds for the circuit size of polynomials then we need only consider depth four.

3 Algebraic preliminaries

Set $Y = \{y_1 \ldots, y_m\}$. Let ν range over tuples from $\mathbb{N}^m$ (we include 0 in $\mathbb{N}$) and define $y^\nu = y_1^{\nu_1} \cdots y_m^{\nu_m}$ where $\nu = (\nu_1, \ldots, \nu_m)$. Define also $|\nu| = \sum_{i=1}^m \nu_i$ and addition of tuples to be component wise. An element f of $k(Z)[[Y]]$ has a unique expression $f = \sum_\nu f_\nu y^\nu$ where $f_\nu \in k(Z)$ for

all ν. We define $L_d(f,Y)$ to be the k-linear subspace of $k(Z)$ spanned by all f_ν with $|\nu| \leq d$ and set $D_d(f,Y) = \dim_k L_d(f,Y)$, as usual the dimension of the zero vector space is 0.

When f is a polynomial the measure is essentially the one based on matrix rank, e.g., see Shpilka and Wigderson [8], Raz [7] or Shpilka and Yehudayoff [9].

LEMMA 3.1 *For all $f, g \in k(Z)[[Y]]$ and all $d \geq 0$*

1. $D_d(f \pm g, Y) \leq D_d(f,Y) + D_d(g,Y)$.
2. $D_d(fg,Y) \leq D_d(f,Y)D_d(g,Y)$.

PROOF. If $fg = 0$ the claims are trivial so assume this is not the case. Let S_f and S_g be bases for $L_d(f,Y)$ and $L_d(g,Y)$ respectively. If $f \pm g = 0$ the first inequality is trivial, otherwise $\sum_\nu f_\nu y^\nu \pm \sum_\nu g_\nu y^\nu = \sum_\nu (f_\nu \pm g_\nu) y^\nu$ so that $S_f \cup S_g$ is a spanning set for $L_d(f \pm g, Y)$.

For the second inequality we have

$$\Big(\sum_\mu f_\mu y^\mu\Big)\Big(\sum_\nu g_\nu y^\nu\Big) = \sum_{i=0}^{\infty} \sum_{|\mu|+|\nu|=i} f_\mu g_\nu y^{\mu+\nu}$$

so that $S_f S_g = \{f'g' \mid f' \in S_f, g' \in S_g\}$ is a spanning set for $L_d(fg,Y)$ whose cardinality is no larger than $D_d(f,Y)D_d(g,Y)$. □

Recall that by definition $k(Z)k[[Y]] = \{fg \mid f \in k(Z), g \in k[[Y]]\}$.

LEMMA 3.2 *Let $f, g \in k(Z)[[Y]]$. Then, for all $d \geq 0$,*

1. $D_d(fg,Y) \leq 1$, *whenever* $f, g \in k(Z)k[[Y]]$.
2. $D_d(fg,Y) \leq D_d(f,Y)$, *whenever* $g \in k(Z)k[[Y]]$.
3. $D_d(fg,Y) \leq D_d(f,Y)$, *whenever* $g \in k(Z) \cup k[[Y]]$.

PROOF. We may assume that $fg \neq 0$. The first inequality follows from the second: $D_d(fg,Y) = D_d(1 \cdot fg, Y) \leq D_d(1,Y) = 1$. The second inequality follows from the third since $g = h_1 h_2$ with $h_1 \in k(Z)$ and $h_2 \in k[[Y]]$ so that $D_d(fg,Y) = D_d(fh_1h_2,Y) \leq D_d(fh_1,Y) \leq D_d(f,Y)$. For the third inequality, if $g \in k(Z)$ and $\{g_1, \ldots, g_r\}$ is a spanning set for $L_d(f,Y)$ then $\{g_1 g, \ldots, g_r g\}$ is a spanning set for $L_d(fg,Y)$. On the other hand if $g \in k[[Y]]$ then each coefficient of fg of degree d is a k-linear combination of the coefficients of f of degree at most d, i.e., it is in the space $L_d(f,Y)$. □

4 Formulae

Recall from §1 that our formulae are allowed to have leaves labeled with members of $k(Z) \cup k[[Y]]$ rather than just $k \cup X$ (where $Y \subseteq X$ and $Z = X - Y$). In all other respects there is no difference from the standard definition.

Any formula Φ can be converted to a formula Ψ with the following properties:

1. $r(\Psi) = r(\Phi)$ and $|\Psi| \leq |\Phi|$.
2. The result at each computation vertex is not a member of $k(Z) \cup k[[Y]]$.
3. If the result at a multiplicative vertex is fg with $f \in k(Z)$ and $g \in k[[Y]]$ but $fg \notin k(Z) \cup k[[Y]]$ then the subformula rooted at v is just $f \times g$.

These properties simplify some definitions and proofs. The formula Ψ is obtained from Φ by replacing each maximal size subtree whose result f is in $k(Z) \cup k[[Y]]$ with a leaf labeled by f. Likewise for multiplicative vertices whose result is in $k(Z)k[[Y]] - (k(Z) \cup k[[Y]])$. Clearly $|\Psi| \leq |\Phi|$ and the depth of Ψ is no larger than that of Φ. Indeed the number of vertices in Ψ of any given

type is no more than the number in Φ. The same applies to the notions of depth defined below. From now on we will assume that formulae have the stated properties.

A simple induction argument shows that the set of expressions computed by our model is $\{f_1g_1 + \cdots + f_rg_r \mid r \geq 1, f_i \in k(Z), g_i \in k[[Y]], \text{ for } 1 \leq i \leq r\}$, i.e., the smallest subring of $k(Z)[[Y]]$ that contains $k(Z) \cup k[[Y]]$ (this is a standard algebraic fact). Note that if $f = f_1g_1 + \cdots + f_rg_r$ is an element of this ring then $D_d(f, Y) \leq r$, for all $d \geq 0$, by the first part of Lemmas 3.1 and 3.2. Thus $D_d(f, Y)$ is bounded from above independently of d; naturally this is not true for arbitrary members of $k(Z)[[Y]]$, e.g., consider $\sum_{i=0}^{\infty} z^i y^i$. The bounds we provide in the rest of the paper do not depend on the degree and so we will use $D(f, Y)$ rather than $D_d(f, Y)$. Also, if Φ is a formula, we use $D(\Phi, Y)$ as short hand for $D(r(\Phi), Y)$.

An *extended formula* is the same one defined at the start of this section but we allow the leaves to be labeled by elements of $k(Z)[[Y]]$ rather than $k(Z) \cup k[[Y]]$ (or just $k \cup X$ for the usual definition). This is a convenient device that allows us to collapse subformulae and replace them by leaves labeled with the expressions computed by the collapsed subformulae. Clearly the resulting formula computes the same expression. For the rest of this section we will use the term 'formula' to mean an extended formula unless otherwise stated. The notion of an extended formula is used later on in Lemmas 4.3 and 5.1 however it is necessary to have the preceding definition in place so that preliminary results can be applied to the two lemmas.

A multiplicative vertex is *non-scalar* if the neither of the results of its left and right children is in k. Let v be a non-scalar multiplicative vertex of a formula Φ whose left and right children have results f, g respectively. We say that v is

- *separated*: if $f \in k(Z)k[[Y]]$ or $g \in k(Z)k[[Y]]$.
- *essential*: if $f, g \notin k(Z)k[[Y]]$;

Naturally these definitions are relative to Y and Z which we have fixed throughout. The reason for this distinction is that if v is a separated vertex with, e.g., $g \in k(Z)k[[Y]]$ then $D(fg, Y) \leq D(f, Y)$, by Lemma 3.2. Note that, by definition, a separated vertex is non-scalar, thus a multiplicative vertex that is not essential is either separated or scalar. Note also that in a non-extended formula every essential vertex is the root of a subtree that has at least four leaves.

Let $v_1, \ldots, v_n$ be a path in a formula where v_1 is the root and both children of v_n are leaves. We identify three consecutive sections of the path, some of which might be empty:

1. *Upper essential-free section* $v_1, \ldots, v_r$: no essential vertices.
2. *Essential section* $v_{r+1}, \ldots, v_s$: where v_{r+1} and v_s are essential, other vertices are arbitrary.
3. *Lower essential-free section* $v_{s+1}, \ldots, v_n$: no essential vertices.

If a path has no essential vertices then it consists of only an upper section and is called an *essential free path*. Let $u_1 \ldots, u_r$ be an essential-free path. We divide it into two subparts (again either part can be empty):

1. *Separated section* $u_1, \ldots, u_s$: where u_s is separated (hence non-scalar);
2. *Additive section* $u_{s+1}, \ldots, u_r$: all vertices are additive or scalar.

Figure 1 is helpful in understanding the preceding definitions as well as those in the next paragraph. The various measures are understood to be the maximum over all paths.

For a formula Φ the *essential depth*, denoted by $e(\Phi)$, is the maximum number of essential vertices on any path from the root to a leaf. The *essential additive gap*, denoted by $g(\Phi)$, is the maximum number of additive vertices between any two consecutive essential vertices in any path from the root to a leaf. Note that if $e(\Phi) \leq 1$ then $g(\Phi) = 0$. The *lower separated additive depth*, denoted by $a_L(\Phi)$, is the maximum number of additive vertices in the separated section of the lower essential-free part of any path from the root to a leaf. Naturally we could define an upper counterpart but it will not play any role in our analysis.

Upper essential free section.	v_1 $\vdots$ v_r	
Essential section.	v_{r+1} $\vdots$ v_s	Essential depth, $e(\Phi)$. Essential additive gap, $g(\Phi)$.
Lower essential free section.	v_{s+1} $\vdots$ v_t	Lower separated additive depth, $a_L(\Phi)$.
	v_{t+1} $\vdots$ v_n	

Figure 1: Sections of a path where v_{r+1} and v_s are essential, v_t is separated.

Before proceeding with the next few definitions it is worthwhile commenting that the main general result, the second part of Theorem 5.1, can be established by using only the notions of depth introduced above. In the lemmas of this section we would replace all occurrences of formula size of various types by the obvious upper bound implied by the appropriate depth. However the proofs are more natural in the form given and of course the various subformulae could in principle be much smaller than the upper bound based on depth; the price is the need to introduce some extra definitions.

A formula is *essential-free* if it has no essential vertices, thus every path from the root to a leaf consists only of an essential free section. Note that such a formula is not required to have any computation vertices at all; thus every leaf is an essential-free formula no matter how it is labelled. For such a formula Φ, we will use $|\Phi|_a$ to denote the total number of additive vertices in the separated sections of all paths from the root to a leaf counting each vertex only once (this is not the same as the total number of additive vertices in Φ). Given an arbitrary formula Φ the last vertex of every non-empty essential section has two essential-free formulae attached to it. Let $\Phi_1, \ldots, \Phi_r$ be all the essential-free formulae thus obtained. We define $|\Phi|_L = \max_{1 \leq i \leq r} |\Phi_i|_a$. If Φ has no essential vertices then $|\Phi|_L = 0$.

A formula Φ is *essential* if the root is essential, and so $|\Phi| \geq 1$. The *essential size* of such a Φ, denoted by $|\Phi|_e$, is the number of essential vertices of Φ. For an arbitrary formula Ψ its essential size is $|\Psi|_e = \max |\Phi|_e$ where Φ ranges over all essential subformulae of Ψ. If Φ has no essential vertices then $|\Phi|_e = 0$. In order to avoid misunderstanding, we note here that a formula which is not essential is not necessarily an essential-free formula since it could have an essential vertex other than the root.

Let Φ be a formula and let $\Phi_1, \ldots, \Phi_r$ be all the maximal essential subformulae of Φ. Replacing each Φ_i with a leaf labeled by $r(\Phi_i)$ yields an essential-free formula Ψ (which could consist of just a leaf). Define $|\Phi|_U$ to be $|\Psi|_a$. Note that if Φ is essential-free then $|\Phi|_U = |\Phi|_a$.

Since most of the preceding definitions are not standard, a summary of them is given in Figure 2 as an aid to the reader.

We now define a function α on formulae as follows.

1. $\alpha(\Phi) = 1$, if Φ is essential-free.
2. $\alpha(\Phi_1 \pm \Phi_2) = 1 + \max(\alpha(\Phi_1), \alpha(\Phi_2))$, where $\Phi_1 \pm \Phi_2$ has an essential vertex.
3. $\alpha(\Phi_1 \times \Phi_2) = \max(\alpha(\Phi_1), \alpha(\Phi_2))$, if the operation is not essential.
4. $\alpha(\Phi_1 \times \Phi_2) = \alpha(\Phi_1) + \alpha(\Phi_2)$, if the operation is essential.

Extended formula: a formula with leaves labeled by elements of $k(Z)[[Y]]$.

Non-scalar vertex: a multiplicative vertex for which neither of the results of its left and right children is in k.

Separated vertex: a non-scalar vertex s.t. the result of at least one child is in $k(Z)k[[Y]]$.

Essential vertex: a non-scalar vertex s.t. the result of neither child is in $k(Z)k[[Y]]$.

Essential depth, $e(\Phi)$: maximum number of essential vertices on any path from the root to a leaf.

Essential additive gap, $g(\Phi)$: maximum number of additive vertices between any two consecutive essential vertices in any path from the root to a leaf.

Lower separated additive depth, $a_L(\Phi)$: maximum number of additive vertices in the separated section of the lower essential-free part of any path from the root to a leaf.

Essential-free formula: a formula with no essential vertices.

$|\Phi|_a$, for an essential free formula: the total number of additive vertices in the separated sections of all paths from the root to a leaf counting each vertex only once (this is not the same as the total number of additive vertices in Φ).

$|\Phi|_L$, for an arbitrary formula: the last vertex of every non-empty essential section of Φ has two essential-free formulae attached to it. Let $\Phi_1, \ldots, \Phi_r$ be all the essential-free formulae thus obtained. Define $|\Phi|_L = \max_{1 \le i \le r} |\Phi_i|_a$; if Φ has no essential vertices then $|\Phi|_L = 0$.

Essential formula: a formula whose root is an essential vertex.

Essential size, $|\Phi|_e$: if Φ is essential then the number of essential vertices of Φ. Otherwise $\max |\Psi|_e$ where Ψ ranges over all essential subformulae of Φ.

$|\Phi|_U$, for an arbitrary formula: let $\Phi_1, \ldots, \Phi_r$ be all the maximal essential subformulae of Φ. Replace each Φ_i with a leaf labeled by $r(\Phi_i)$ to yield an essential-free formula Ψ (which could consist of just a leaf). $|\Phi|_U$ is defined to be $|\Psi|_a$; if Φ is essential-free then $|\Phi|_U = |\Phi|_a$.

Figure 2: Summary of definitions for a formula Φ (see also Figure 1).

LEMMA 4.1 *Let Φ be a formula that has an essential vertex.*

1. *Suppose that $\Phi = \Phi_1 \pm \Phi_2$ and one of Φ_1, Φ_2 is essential-free. Let Ψ be the subformula from Φ_1, Φ_2 that has an essential vertex. Then $\alpha(\Phi) = 1 + \alpha(\Psi)$.*
2. *Suppose that $\Phi = \Phi_1 \times \Phi_2$ and the operation is not essential. Let Ψ be the subformula from Φ_1, Φ_2 whose result is not in $k(Z)k[[Y]]$. Then $\alpha(\Phi) = \alpha(\Psi)$.*

PROOF. The first part is immediate from the definition of α and the fact that its value is at least 1. For the second let Ψ' be the subformula from Φ_1, Φ_2 whose result is in $k(Z)k[[Y]]$. It follows from the assumptions at the start of this section that Ψ' is essential-free and so $\alpha(\Psi') = 1$. □

As a consequence of this lemma, when evaluating α on a formula we either obtain 1 straight away because the formula is non-essential or follow some path to an essential operation (since at additive vertices we add one to the maximum value of α on the left and right subformuale).

The key part of the next lemma is the bound for essential formulae. However we need to introduce an auxiliary definition, $g^*(\Phi)$, for the induction proof since a subformula of an essential formula need not itself be essential even if it has an essential vertex (i.e., its root need not be an essential vertex). However, if Φ does have an essential vertex at all then $\Phi \times \Phi$ is an essential formula.

LEMMA 4.2 *Let Φ be a formula and define $g^*(\Phi) = g(\Phi \times \Phi)$. Then $\alpha(\Phi) \leq g^*(\Phi)(|\Phi|_e - 1)/2 + |\Phi|_e + 1$. If Φ is essential then $\alpha(\Phi) \leq g(\Phi)(|\Phi|_e - 1)/2 + |\Phi|_e + 1$*

PROOF. We use induction on $d(\Phi)$, the depth of Φ, to prove the first part. If Φ is essential-free the result is trivial since $\alpha(\Phi) = 1$ and $g^*(\Phi) = 0$; this also covers the base case $d(\Phi) = 0$. Assume now that $d(\Phi) > 0$ and Φ has an essential vertex. There are three cases to consider.

Case 1: $\Phi = \Phi_1 \pm \Phi_2$. Then $\alpha(\Phi) = 1 + \max(\alpha(\Phi_1), \alpha(\Phi_2))$. As observed above, the value of α is obtained by following some path P to an essential vertex. Suppose there are c additive vertices on the path P to an essential vertex that is the root of a subformula $\Psi_1 \times \Psi_2$. Note that $r(\Phi) \notin k(Z)k[[Y]]$ since Φ has an essential vertex and so $\Phi \times \Phi$ is essential, hence $g^*(\Phi) \geq c$. Now

$$\begin{aligned}
\alpha(\Phi) &= c + \alpha(\Psi_1 \times \Psi_2) \\
&\leq c + g^*(\Psi_1)(|\Psi_1|_e - 1)/2 + |\Psi_1|_e + 1 + g^*(\Psi_2)(|\Psi_2|_e - 1)/2 + |\Psi_2|_e + 1 \\
&\leq c + g^*(\Phi)(|\Psi_1|_e + |\Psi_2|_e - 2)/2 + |\Psi_1|_e + |\Psi_2|_e + 2 \\
&\leq g^*(\Phi)(|\Phi|_e - 1)/2 + |\Phi|_e + 1 - g^*(\Phi) + c \\
&\leq g^*(\Phi)(|\Phi|_e - 1)/2 + |\Phi|_e + 1.
\end{aligned}$$

The inequality $g^*(\Psi_1) \leq g^*(\Phi)$ is justified the fact that $\Psi_1 \times \Psi_2$ is essential so that $g^*(\Psi_1) \leq g(\Psi_1 \times \Psi_2) \leq g^*(\Phi)$. Similarly for $g^*(\Psi_2) \leq g^*(\Phi)$.

Case 2: $\Phi = \Phi_1 \times \Phi_2$ where the operation is not essential. We may assume w.l.o.g. that Φ_2 is essential-free. Thus

$$\alpha(\Phi) = \alpha(\Phi_1) \leq g^*(\Phi_1)(|\Phi_1|_e - 1)/2 + |\Phi_1|_e + 1 \leq g^*(\Phi)(|\Phi|_e - 1)/2 + |\Phi|_e + 1.$$

Case 3: $\Phi = \Phi_1 \times \Phi_2$ and the operation is essential. Then

$$\begin{aligned}
\alpha(\Phi) &= \alpha(\Phi_1) + \alpha(\Phi_2) \\
&\leq g^*(\Phi_1)(|\Phi_1|_e - 1)/2 + |\Phi_1|_e + 1 + g^*(\Phi_2)(|\Phi_2|_e - 1)/2 + |\Phi_2|_e + 1 \\
&\leq g^*(\Phi)(|\Phi_1|_e + |\Phi_2|_e - 2)/2 + |\Phi|_e + 1 \\
&\leq g^*(\Phi)(|\Phi|_e - 3)/2 + |\Phi|_e + 1 \\
&\leq g^*(\Phi)(|\Phi|_e - 1)/2 + |\Phi|_e + 1.
\end{aligned}$$

Finally, if Φ is an essential formula then $g^*(\Phi) = g(\Phi)$. □

LEMMA 4.3 *Let Φ be an essential-free formula whose leaves are labeled by $f_1, \ldots, f_r \in k(Z)[[Y]] - (k(Z) \cup k[[Y]])$ and $h_1, \ldots h_s \in k(Z) \cup k[[Y]]$. Let $\delta = \max_{1 \leq i \leq r} D(f_i, Y)$. If $s = 0$ then $D(\Phi, Y) \leq r\delta(|\Phi|_a + 1)$ else $D(\Phi, Y) \leq (r\delta + 2)(|\Phi|_a + 1)$*

PROOF. We proceed by induction on $|\Phi|$. If Φ has no separated vertices then it consists only of additive and scalar vertices, hence $r(\Phi) = \sum_{i=1}^{r} a_i f_i + \sum_{j=1}^{s} b_j h_j$ where $a_i, b_j \in k$, for $1 \leq i \leq r$ and $1 \leq j \leq s$. If $s = 0$ then $D(f, Y) \leq \sum_{i=1}^{r} D(f_i, Y) \leq r\delta$, by the first part of Lemma 3.1. Otherwise $\sum_{j=1}^{s} b_j h_j = g_1 + g_2$ where $g_1 \in k(Z)$ and $g_2 \in k[[Y]]$ and so $D(\sum_{j=1}^{s} b_j h_j, Y) \leq D(g_1, Y) + D(g_2, Y) \leq 2$ by the first part of Lemmas 3.1 and 3.2. Thus $D(\Phi, Y) \leq r\delta + 2$. For the rest of the proof we define γ to be $r\delta$ if $s = 0$ otherwise $r\delta + 2$. Assume now that Φ has at least one separated vertex. We have two cases.

Case 1: $\Phi = \Phi_1 \times \Phi_2$ and the operation must be non essential. If $r(\Phi_2) \in k(Z)k[[Y]]$ then $D(\Phi, Y) \leq D(\Phi_1, Y) \leq \gamma(|\Phi_1|_a + 1) \leq \gamma(|\Phi|_a + 1)$. Similarly if $r(\Phi_1) \in k(Z)k[[Y]]$.

Case 2: $\Phi = \Phi_1 \pm \Phi_2$ so that $D(\Phi, Y) \leq D(\Phi_1, Y) + D(\Phi_2, Y) \leq \gamma(|\Phi_1|_a + 1) + \gamma(|\Phi_2|_a + 1) = \gamma(|\Phi|_a + 1)$. Here we have used the fact that Φ has a separated vertex so that $|\Phi|_a = |\Phi_1|_a + |\Phi_2|_a + 1$. □

LEMMA 4.4 *Let Φ be an essential formula with leaves labeled by members of $k(Z) \cup k[[Y]]$. Then $D(\Phi, Y) \leq 2^{\alpha(\Phi)}(|\Phi|_L + 1)^{|\Phi|_e + 1}$.*

PROOF. We use an auxiliary definition which is necessary because a subformula of an essential formula is not necessarily essential. Let Ψ be any formula and let $\Psi_1, \ldots, \Psi_r$ be all of its maximal essential-free subformulae. We define $|\Psi|_l = \max_{1 \le i \le r} |\Psi_i|_a$. Note that for an essential formula Φ we have $|\Phi|_L = |\Phi|_l$ and so it suffices to establish the result for $|\Phi|_l$ in place of $|\Phi|_L$ where Φ is any formula. We proceed by induction on the maximum number of vertices on a path from the root to an essential vertex. If this number is 0 then Φ is essential-free and, by Lemma 4.3, $D(\Phi, Y) \le 2(|\Phi|_a + 1) = 2^{\alpha(\Phi)}(|\Phi|_l + 1)^{|\Phi|_e+1}$. There are three cases to consider for the induction step.

Case 1: $\Phi = \Phi_1 \pm \Phi_2$. We have

$$\begin{aligned} D(\Phi, Y) &\le D(\Phi_1, Y) + D(\Phi_2, Y) \\ &\le 2^{\alpha(\Phi_1)}(|\Phi_1|_l + 1)^{|\Phi_1|_e+1} + 2^{\alpha(\Phi_2)}(|\Phi_2|_l + 1)^{|\Phi_2|_e+1} \\ &\le 2 \cdot 2^{\max(\alpha(\Phi_1), \alpha(\Phi_2))}(|\Phi|_l + 1)^{|\Phi|_e+1} \\ &= 2^{\alpha(\Phi)}(|\Phi|_l + 1)^{|\Phi|_e+1}. \end{aligned}$$

Case 2: $\Phi = \Phi_1 \times \Phi_2$ where this operation is not essential. If $r(\Phi_2) \in k(Z)k[[Y]]$ then $D(\Phi, Y) \le D(\Phi_1, Y) \le 2^{\alpha(\Phi_1)}(|\Phi_1|_l + 1)^{|\Phi_1|_e+1} \le 2^{\alpha(\Phi)}(|\Phi|_l + 1)^{|\Phi|_e+1}$. Similarly if $r(\Phi_1) \in k(Z)k[[Y]]$.

Case 3: $\Phi = \Phi_1 \times \Phi_2$ and this operation is essential. We have

$$\begin{aligned} D(\Phi, Y) &\le D(\Phi_1, Y) D(\Phi_2, Y) \\ &\le 2^{\alpha(\Phi_1)}(|\Phi_1|_l + 1)^{|\Phi_1|_e+1} \cdot 2^{\alpha(\Phi_2)}(|\Phi_2|_l + 1)^{|\Phi_2|_e+1} \\ &\le 2^{\alpha(\Phi_1)+\alpha(\Phi_2)}(|\Phi|_l + 1)^{|\Phi_1|_e+|\Phi_2|_e+2} \\ &\le 2^{\alpha(\Phi)}(|\Phi|_l + 1)^{|\Phi|_e+1}. \end{aligned}$$

This completes the proof. □

We end this section with a remark. One reason for allowing non-extended formulae to have leaves labeled by members of $k(Z) \cup k[[Y]]$ without charge is because $D(f, Y) \le 1$ for all $f \in k(Z) \cup k[[Y]]$. On this basis we could allow leaves to be labeled by members of $k(Z)k[[Y]]$. However we would lose the other main important property, exploited in Lemma 4.3, that if Φ has only additive vertices and has leaves labeled by members of $k(Z) \cup k[[Y]]$ then $D(\Phi, Y) \le 2$. If we allow leaves to be labeled by members of $k(Z)k[[Y]]$ then the bound in Lemma 4.3 must be replaced by $(r\delta + s)(|\Phi|_a + 1)$ and consequently the factor $2^{\alpha(\Phi)}$ in Lemma 4.4 must be replaced with one that is too large for subsequent bounds.

5 Lower bounds

From now on, unless otherwise stated, 'formula' means a tree with leaves labeled by elements of $k(Z) \cup k[[Y]]$, i.e., we exclude extended formulae. For a formula Φ let $\alpha^*(\Phi)$ denote the maximum value of α over all maximal essential subformulae of Φ.

LEMMA 5.1 *Let Φ be a formula that computes $f \in k(Z)[[Y]]$. Then*

1. *$D(f, Y) \le 2(|\Phi|_U + 1)$, if Φ has no essential operations;*
2. *$D(f, Y) \le |\Phi| 2^{\alpha^*(\Phi)}(|\Phi|_L + 1)^{|\Phi|_e+1}(|\Phi|_U + 1)$, otherwise.*

PROOF. If Φ has no essential operations then it is an essential-free formula with inputs from $k(Z) \cup k[[Y]]$ and the result follows from Lemma 4.3.

Suppose now that Φ has an essential operation. Let $\Phi_1, \ldots, \Phi_r$ be all the maximal essential subformulae of Φ and set $f_i = r(\Phi_i)$, for $1 \le i \le r$. By Lemma 4.4 we have

$$D(f_i, Y) \le 2^{\alpha(\Phi_i)}(|\Phi_i|_a + 1)^{|\Phi_i|_e+1} \le 2^{\alpha^*(\Phi)}(|\Phi|_L + 1)^{|\Phi|_e+1}.$$

Now replace each Φ_i with a leaf labeled by f_i. This yields an essential-free extended formula Ψ whose result is that of Φ. As pointed out in §4, p.5, every essential vertex in Φ is the root of a subtree with at least four leaves; it follows that $|\Phi| \geq 4r - 1 \geq r + 2$ as $r \geq 1$. By Lemma 4.3,

$$\begin{aligned} D(f, Y) &\leq (r2^{\alpha^*(\Phi)}(|\Phi|_L + 1)^{|\Phi|_e+1} + 2)(|\Psi|_a + 1) \\ &\leq (r + 2)2^{\alpha^*(\Phi)}(|\Phi|_L + 1)^{|\Phi|_e+1}(|\Psi|_a + 1) \\ &\leq |\Phi|\, 2^{\alpha^*(\Phi)}(|\Phi|_L + 1)^{|\Phi|_e+1}(|\Phi|_U + 1). \end{aligned}$$

This completes the proof. □

THEOREM 5.1 *Let Φ be a formula that computes $f \in k(Z)[[Y]]$. Then*

1. *$D(f, Y) \leq 2(|\Phi| + 1)$, if Φ has no essential operations;*
2. *$D(f, Y) \leq |\Phi|^2 2^{\alpha^*(\Phi)}(|\Phi|_L + 1)^{|\Phi|_e+1} \leq |\Phi|^2 2^{(g(\Phi)/2+a_L(\Phi)+1)2^{e(\Phi)}-g(\Phi)}$, otherwise.*

PROOF. The result follows from Lemma 5.1. For the first part we simply note that $|\Phi|_U \leq |\Phi|$ and for the second part $|\Phi|_U \leq |\Phi| - 1$. The rest follows from Lemma 4.2 and by noting that $|\Phi|_L \leq 2^{a_L(\Phi)} - 1$, $|\Phi|_e \leq 2^{e(\Phi)} - 1$ □

THEOREM 5.2 *Let $Y = \{y_{ij} \mid 1 \leq i, j \leq n\}$ and $Z = \{z_{ij} \mid 1 \leq i, j \leq n\}$ be disjoint sets of n^2 distinct indeterminates over k. Let M be the $n \times n$ matrix with (i, j)th entry $z_{ij}y_{ij}$, for $1 \leq i, j \leq n$. Suppose that Φ is a formula for $\det M$ and that Φ either has no essential operations or that $e(\Phi) \leq \lg c_1 n$, $g(\Phi) \leq c_2 \lg n + o(\lg n)$ and $a_L(\Phi) \leq c_3 \lg n + o(\lg n)$ where $(c_2/2 + c_3)c_1 < 1$. Then $|\Phi| \geq n^{\Omega(n)}$.*

PROOF. Since

$$\det M = \sum_{\pi \in S_n} (-1)^{\mathrm{sgn}(\pi)} z_{1\pi(1)} \cdots z_{n\pi(n)} y_{1\pi(1)} \cdots y_{n\pi(n)}$$

it follows that $D(\det M, Y) = n! \geq (n/e)^n = 2^{n \lg n - n \lg e}$. We may assume that the formula Φ for $\det M$ satisfies the assumptions at the start of §4. If Φ has no essential operations then it follows from the first part of Theorem 5.1 that $|\Phi| \geq 2^{n \lg n - n \lg e - 1} - 1$ irrespective of depth.

Suppose now that $|\Phi|$ has an essential operation. Let $\epsilon = (c_2/2 + c_3)c_1$. By the second part of Theorem 5.1, $|\Phi| \geq 2^{((1-\epsilon)n \lg n - o(n \lg n))/2}$. □

Note that in the preceding result, $\det(z_{ij})$ and $\det(y_{ij})$ are given to us for free as are any other members of $k(Z) \cup k[[Y]]$. Thus the non-scalar depth of a formula for $\det M$ can be less than $\lg \deg(\det M) = 2 \lg n$.

Before proving Theorem 2.1 we need a technical lemma.

LEMMA 5.2 *Let $X = \{x_1, \ldots, x_n\}$, $Y = \{y_1, \ldots, y_n\}$, $Z = \{z_1, \ldots, z_n\}$ and $\{\xi_1, \ldots, \xi_n\}$ be disjoint sets of distinct indeterminates over k and $f \in k[X]$ with $f \notin k$. Set $k_\xi = k(\xi_1, \ldots, \xi_n)$. Then there do not exist $g \in k_\xi[[Y]]$ and $h \in k_\xi(Z)$ such that*

$$f(\xi_1 + y_1z_1, \ldots, \xi_n + y_nz_n) = g(y_1, \ldots, y_n)h(z_1, \ldots, z_n).$$

PROOF. Suppose there are g and h such that the equation holds so that $g(y_1, \ldots, y_n) = f(\xi_1 + y_1z_1, \ldots, \xi_n + y_nz_n)/h(z_1, \ldots, z_n)$ in $k_\xi(Z)[[Y]]$. Since $f(\xi_1 + y_1z_1, \ldots, \xi_n + y_nz_n)$ is a polynomial in Y and $h(z_1, \ldots, z_n)$ is free of Y it follows that $g(y_1, \ldots, y_n)$ is also a polynomial in Y. Setting $y_i \mapsto 0$, for $1 \leq i \leq n$ we obtain $f(\xi_1, \ldots, \xi_n) = g(0, \ldots, 0)h(z_1, \ldots, z_n)$. Note that $g(0, \ldots, 0) \neq 0$ since $f(x_1, \ldots, x_n) \neq 0$ and $\xi_1, \ldots, \xi_n$ are indeterminates over k so that $f(\xi_1, \ldots, \xi_n) \neq 0$. Thus

$$f(\xi_1 + y_1z_1, \ldots, \xi_n + y_nz_n) = g(y_1, \ldots, y_n)f(\xi_1, \ldots, \xi_n)/g(0, \ldots, 0).$$

Hence $\deg_Z f(\xi_1 + y_1z_1, \ldots, \xi_n + y_nz_n) = 0$. However $\deg_Z f(\xi_1 + y_1z_1, \ldots, \xi_n + y_nz_n) = \deg_X f(x_1, \ldots, x_n) > 0$ which is a contradiction. □

Proof of Theorem 2.1. Let $Y = \{y_{ij} \mid 1 \leq i, j \leq n\}$, $Z = \{z_{ij} \mid 1 \leq i, j \leq n\}$ and $\{\xi_{ij} \mid 1 \leq i, j \leq n\}$ be disjoint sets each of n^2 distinct inderminates over k. We may assume that the formula Φ for $\det M$ satisfies the assumptions at the start of §4. Replace each leaf of Φ labeled with x_{ij} by the formula $\xi_{ij} + z_{ij} \times y_{ij}$. Let Ψ be the formula thus obtained, which is viewed as a formula with scalars from the field $k(\{\xi_{ij} \mid 1 \leq i, j \leq n\})$. Clearly $|\Psi| \leq 3|\Phi| + 2$ and $r(\Psi) = \det(\xi_{ij} + z_{ij}y_{ij})$. The highest degree terms of $\det(\xi_{ij} + z_{ij}y_{ij})$ are $\det(z_{ij}y_{ij})$ so that $D(\Psi, Y) \geq n!$. By Lemma 5.2, all of the non-scalar vertices of Φ become essential in Ψ and Ψ satisfies the assumptions at the start of §4. The only active vertices are those of the subtrees $\xi_{ij} + z_{ij} \times y_{ij}$. It follows that $e(\Psi) = \mu(\Phi)$, $g(\Psi) \leq \gamma(\Phi)$ and $a_L(\Psi) \leq \gamma(\Phi) + 1$. The result now follows as in Theorem 5.2. □

Acknowledgment. The author is grateful to an anonymous referee for very helpful comments.

References

[1] M. Agrawal and V. Vinay, Arithmetic Circuits: A Chasm at Depth Four, *49th Annual IEEE Symposium on Foundations of Computer Science*, (2008), pp. 67–72

[2] A. Borodin, J. von zur Gathen and J. Hopcroft, Fast parallel matrix and GCD computations, *Inform. and Control*, Vol. 52, No. 3 (1982), pp. 241–256.

[3] R.P. Brent, The parallel evaluation of general arithmetic expressions, *JACM*, Vol. 21, No. 2 (1974), pp. 201–206.

[4] N.H. Bshouty, R. Cleve and W. Eberley, Size-Depth Tradeoffs for Algebraic Formulas, *SIAM J. Comput.*, Vol. 24, No. 4 (1995), pp. 682–705.

[5] L. Csanky, Fast parallel inversion algorithms, *SIAM J. Comput.*, Vol. 5, No. 4 (1976), pp. 618–623.

[6] K. Kalorkoti, A lower bound for the formula size of rational functions, *SIAM J. Comput.*, Vol. 14, No. 3 (1985), pp. 678–687.

[7] R. Raz, Multi-linear formulas for permanent and determinant are of super-polynomial size, *Proceedings of the 36th Annual STOC*, pp. 633–641, 2004.

[8] A. Shpilka and A. Wigderson, Depth-3 arithmetic circuits over fields of characteristic zero, *Computational Complexity*, Vol. 10, No. 1 (2001), pp. 1–27.

[9] A. Shpilka and A. Yehudayoff, Arithmetic Circuits: a survey of recent results and open questions, *Foundations and Trends in Theoretical Computer Science*, Vol. 5, No. 3/4 (2009) pp. 207–388.

[10] V. Strassen , Vermeidung von divisionen, *Journal für die Reine un Angewandte Mathematik*, 264 (1973), pp. 182–202.

[11] L.G. Valiant Why is Boolean Complexity Theory Difficult?, in 'Boolean Function Complexity', edited by M.S. Paterson, *LMS Lecture Note Series*, 169 (1992), pp. 84–90.

[12] O. Zariski and P. Samuel, *Commutative Algebra*, Vols I, II, (Van Nostrand, Princeton, NJ, 1958).