ON THE EFFECT OF VARIABLE IDENTIFICATION ON THE ESSENTIAL ARITY OF FUNCTIONS ON FINITE SETS

MIGUEL COUCEIRO AND ERKKO LEHTONEN

Abstract. We show that every function of several variables on a finite set of $k$ elements with $n > k$ essential variables has a variable identification minor with at least $n - k$ essential variables. This is a generalization of a theorem of Salomaa on the essential variables of Boolean functions. We also strengthen Salomaa’s theorem by characterizing all the Boolean functions $f$ having a variable identification minor that has just one essential variable less than $f$.

1. Introduction

Theory of essential variables of functions has been developed by several authors [2, 5, 6, 7, 14, 16]. In this paper, we discuss the problem how the number of essential variables is affected by identification of variables (diagonalization). Salomaa [14] proved the following two theorems: one deals with operations on arbitrary finite sets, while the other deals specifically with Boolean functions. We denote the number of essential variables of $f$ by $\text{ess } f$.

**Theorem 1.1.** Let $A$ be a finite set with $k$ elements. For every $n \leq k$, there exists an $n$-ary operation $f$ on $A$ such that $\text{ess } f = n$ and every identification of variables produces a constant function.

Thus, in general, essential variables can be preserved when variables are identified only in the case that $n > k$.

**Theorem 1.2.** For every Boolean function $f$ with $\text{ess } f \geq 2$, there is a function $g$ obtained from $f$ by identification of variables such that $\text{ess } g \geq \text{ess } f - 2$.

Identification of variables together with permutation of variables and cylindrification induces a quasi-order on operations whose relevance has been made apparent by several authors [3, 8, 9, 10, 12, 15, 18]. In the case of Boolean functions, this quasi-order was studied in [4] where Theorem 1.2 was fundamental in deriving certain bounds on the essential arity of functions.

In this paper, we will generalize Theorem 1.2 to operations on arbitrary finite sets in Theorem 3.1. We will also strengthen Theorem 1.2 on Boolean functions in Theorem 4.1 by determining the Boolean functions $f$ for which there exists a function $g$ obtained from $f$ by identification of variables such that $\text{ess } g = \text{ess } f - 1$.

Key words and phrases. Functions on finite sets; Boolean functions; essential variables; variable identification; arity gap; minors of functions.
2. Variable identification minors

Let $A$ and $B$ be arbitrary nonempty sets. A $B$-valued function of several variables on $A$ is a mapping $f : A^n \to B$ for some positive integer $n$, called the arity of $f$. $A$-valued functions on $A$ are called operations on $A$. Operations on $\{0, 1\}$ are called Boolean functions.

We say that the $i$-th variable is essential in $f$, or $f$ depends on $x_i$, if there are elements $a_1, \ldots, a_n, b \in A$ such that

$$f(a_1, \ldots, a_i, \ldots, a_n) \neq f(a_1, \ldots, a_{i-1}, b, a_{i+1}, \ldots, a_n).$$

The number of essential variables in $f$ is called the essential arity of $f$, and it is denoted by $\text{ess } f$. Thus the only functions with essential arity zero are the constant functions.

For an $n$-ary function $f$, we say that an $m$-ary function $g$ is obtained from $f$ by simple variable substitution if there is a mapping $\sigma : \{1, \ldots, n\} \to \{1, \ldots, m\}$ such that

$$g(x_1, \ldots, x_m) = f(x_{\sigma(1)}, \ldots, x_{\sigma(n)}).$$

In the particular case that $n = m$ and $\sigma$ is a permutation of $\{1, \ldots, n\}$, $i \neq j$, if $x_i$ and $x_j$ are essential in $f$, then the function $f_{i\leftrightarrow j}$ obtained from $f$ by the simple variable substitution

$$f_{i\leftrightarrow j}(x_1, \ldots, x_n) = f(x_1, \ldots, x_{i-1}, x_j, x_{i+1}, \ldots, x_n)$$

is called a variable identification minor of $f$, obtained by identifying $x_i$ with $x_j$. Note that $\text{ess } f_{i\leftrightarrow j} < \text{ess } f$, because $x_i$ is not essential in $f_{i\leftrightarrow j}$ even though it is essential in $f$.

We define a quasiorder on the set of all $B$-valued functions of several variables on $A$ as follows: $f \leq g$ if and only if $f$ is obtained from $g$ by simple variable substitution. If $f \leq g$ and $g \leq f$, we denote $f \equiv g$. If $f \leq g$ but $g \not\leq f$, we denote $f < g$. It can be easily observed that if $f \leq g$ then $\text{ess } f \leq \text{ess } g$, with equality if and only if $f \equiv g$.

For a $B$-valued function $f$ of several variables on $A$, we denote the maximum essential arity of a variable identification minor of $f$ by

$$\text{ess}^\prec f = \max_{g < f} \text{ess } g,$$

and we define the arity gap of $f$ by $\text{gap } f = \text{ess } f - \text{ess}^\prec f$.

3. Generalization of Theorem 1.2

Theorem 3.1. Let $A$ be a finite set of $k \geq 2$ elements, and let $B$ be a set with at least two elements. Every $B$-valued function of several variables on $A$ with $n > k$ essential variables has a variable identification minor with at least $n - k$ essential variables.

In the proof of Theorem 3.1, we will make use of the following theorem due to Salomaa (Theorem 1 in [14]), which is a strengthening of Yablonski’s [16] “fundamental lemma”.

**Theorem 3.2.** Let the function \( f : M_1 \times \cdots \times M_n \rightarrow N \) depend essentially on all of its \( n \) variables, \( n \geq 2 \). Then there is an index \( j \) and an element \( c \in M_j \) such that the function \( f(\ldots, x_{j-1}, c, x_{j+1}, \ldots, x_n) \) depends essentially on all of its \( n - 1 \) variables.

We also need the following auxiliary lemma.

**Lemma 3.3.** Let \( f \) be an \( n \)-ary function with \( \text{ess} f = n > k \). Then there are indices \( 1 \leq i < j \leq k + 1 \) such that at least one of the variables \( x_1, \ldots, x_{k+1} \) is essential in \( f_{i \rightarrow j} \).

**Proof.** Since \( x_1 \) is essential in \( f \), there are elements \( a_1, \ldots, a_n, b \in A \) such that
\[
\text{ ess } f(a_1, a_2, \ldots, a_n) \neq f(b, a_2, \ldots, a_n).
\]
Thus there are indices \( 1 \leq i < j \leq k + 1 \) such that \( a_i = a_j \). If \( i \neq 1 \), then it is clear that \( x_1 \) is essential in \( f_{i \rightarrow j} \). If there are no such \( i \) and \( j \) with \( i \neq 1 \), then \( i = 1 < j \) and we have that \( b = a_1 \) for some \( 1 < l \leq k + 1, l \neq j \). For \( m = 1, \ldots, n \), let \( c_m = a_m \) if \( m \notin \{1, j, l\} \) and let \( c_m = a_j \) if \( m \in \{1, j, l\} \). Then \( f(c_1, c_2, \ldots, c_n) \) is distinct from at least one of \( f(a_1, a_2, \ldots, a_n) \) and \( f(b, a_2, \ldots, a_n) \). If \( f(c_1, c_2, \ldots, c_n) \neq f(a_1, a_2, \ldots, a_n) \), then \( x_1 \) is essential in \( f_{1 \rightarrow j} \). If \( f(c_1, c_2, \ldots, c_n) \neq f(b, a_2, \ldots, a_n) \), then \( x_l \) is essential in \( f_{1 \rightarrow l} \).

**Proof of Theorem 3.1.** By Theorem 3.2, there exist \( k + 1 \) constants \( c_1, \ldots, c_k+1 \in A \) such that, after a suitable permutation of variables, the function
\[
f(c_1, \ldots, c_{k+1}, x_{k+2}, \ldots, x_n)
\]
depends on all of its \( n - k - 1 \) variables. There are indices \( 1 \leq i < j \leq k + 1 \) such that \( c_i = c_j \), and by Lemma 3.3 there are indices \( 1 \leq l < m \leq k + 1 \) such that at least one of the variables \( x_1, \ldots, x_{k+1} \) is essential in \( f_{i \rightarrow m} \). With a suitable permutation of variables, we may assume that \( i = 1, j = 2, 1 \leq l \leq 3, m = l + 1 \).

If one of the variables \( x_1, \ldots, x_{k+1} \) is essential in \( f_{1 \rightarrow 2} \), then we are done. Otherwise, we have that for all \( a_{k+2}, \ldots, a_n \in A \),

\[
f(c_1, c_2, \ldots, c_{k+1}, a_{k+2}, \ldots, a_n) = f(c_3, c_3, c_4, \ldots, c_{k+1}, a_{k+2}, \ldots, a_n).
\]
Thus the variables \( x_{k+2}, \ldots, x_n \) are essential in \( f_{2 \rightarrow 3} \). If one of the variables \( x_1, \ldots, x_{k+1} \) is essential in \( f_{2 \rightarrow 3} \), then we are done. Otherwise, we have that for all \( a_{k+2}, \ldots, a_n \in A \),

\[
f(c_3, c_3, c_4, \ldots, c_{k+1}, a_{k+2}, \ldots, a_n) = f(c_3, c_4, c_4, \ldots, c_{k+1}, a_{k+2}, \ldots, a_n),
\]
and so the variables \( x_{k+2}, \ldots, x_n \) are essential in \( f_{3 \rightarrow 4} \) and also at least one of \( x_1, \ldots, x_{k+1} \) is essential in \( f_{3 \rightarrow 4} \). This completes the proof of Theorem 3.1. \( \square \)

We would like to remark that our proof is considerably simpler than Salomaa’s original proof of Theorem 1.2.
4. Strengthening of Theorem 1.2

It is well-known that every Boolean function is represented by a unique multilinear polynomial over the two-element field. Such a representation is called the Zhegalkin polynomial (or the Reed–Muller polynomial) of $f$ [11, 13, 17]. It is clear that a variable is essential in $f$ if and only if it occurs in the Zhegalkin polynomial of $f$. We denote by $\deg p$ the degree of polynomial $p$. If $p$ is the Zhegalkin polynomial of $f$, then we denote the Zhegalkin polynomial of $f_{i\leftarrow j}$ by $p_{i\leftarrow j}$. Note that the only polynomials of degree zero are the constant polynomials.

**Theorem 4.1.** Let $f$ be a Boolean function with at least two essential variables. Then the arity gap of $f$ is two if and only if the Zhegalkin polynomial of $f$ is of one of the following special forms:

- $x_{i_1} + x_{i_2} + \cdots + x_{i_a} + c$,
- $x_i x_j + x_i + c$,
- $x_i x_j + x_i x_k + x_j x_k + c$,
- $x_i x_j + x_i x_k + x_j x_k + x_i + x_j + c$,

where $c \in \{0, 1\}$. Otherwise the arity gap of $f$ is one.

We prove first an auxiliary lemma that takes care of the functions of essential arity at least four whose Zhegalkin polynomial has degree two.

**Lemma 4.2.** If $f$ is a Boolean function with at least four essential variables and the Zhegalkin polynomial of $f$ has degree two, then the arity gap of $f$ is one.

**Proof.** Denote the Zhegalkin polynomial of $f$ by $p$. We need to consider several cases and subcases.

**Case 1.** Assume first that $p$ is of the form

$$p = x_i x_j + x_i x_k + x_j x_k + x_i a_i + x_j a_j + x_k a_k + a,$$

where $a_i, a_j, a_k$ are polynomials of degree at most 1 and $a$ is a polynomial of degree at most 2 such that there are no occurrences of variables $x_i, x_j, x_k$ in $a_i, a_j, a_k, a$.

**Subcase 1.1.** Assume that $\deg a_i = \deg a_j = \deg a_k = 0$. Then $a$ contains a variable $x_l$ distinct from $x_i, x_j, x_k$, and we can write $a = x_l a' + a''$, where $a'$ and $a''$ do not contain $x_l$. Then $f_{i\leftarrow l}$ is represented by the polynomial

$$p_{i\leftarrow l} = x_i x_j + x_i x_k + x_j x_k + x_i a' + a'',$

where all essential variables of $f$ except for $x_l$ occur, because no terms cancel, and hence gap $f = 1$.

**Subcase 1.2.** Assume that at least one of $a_i, a_j, a_k$ has degree 1, say $\deg a_i = 1$. Then $a_i$ contains a variable $x_j$ distinct from $x_i, x_j, x_k$, and so $a_i = x_i a'_i + a''$, where $a'_i$ has degree at most 1 and does not contain $x_i$. Consider

$$p_{j\leftarrow i} = x_k (1 + a_j + a_k) + x_i a_i + a.$$

If all essential variables of $f$ except for $x_j$ occur in $p_{j\leftarrow i}$, then gap $f = 1$ and we are done. Otherwise we need to analyze three different subcases.

**Subcase 1.2.1.** Assume that variable $x_k$ occurs in $p_{j\leftarrow i}$ but there is a variable $x_l$ that occurs in $a_i$ and $a_k$ but not in $a_j$, nor in $a$ such that $x_l$ does not occur in $p_{j\leftarrow i}$ (due to some cancelling terms in $a_j$ and $a_k$). Write $a_j = x_l + a'_j, a_k = x_l + a'_k$, and consider

$$p_{j\leftarrow l} = x_i x_l + x_i x_k + x_l x_k + x_i a_i + x_l + x_l a'_j + x_k x_l + x_k a'_k + a$$

$$= x_i x_l + x_i x_k + x_i a_i + x_l + x_l a'_j + x_k a'_k + a.$$
Every essential variable of \( f \) except for \( x_j \) occurs in \( p_{j+1} \), and hence \( \text{gap} \, f = 1 \).

**Subcase 1.2.2.** Assume that \( x_k \) does not occur in \( p_{j+1} \). In this case \( a_j = a_k + 1 \).

Consider
\[
(14) \quad p_{j+1} = x_i(1 + a_i + a_j) + x_k a_k + a.
\]

If any term of \( a_j \) is cancelled by a term of \( a_i \), it still remains as a term of \( a_k \), and hence all variables occurring in \( a_i, a_j, a_k \) occur in \( p_{j+1} \). If both \( x_i \) and \( x_k \) also occur in \( p_{j+1} \), then all essential variables of \( f \) except for \( x_j \) occur in \( p_{j+1} \), and so \( \text{gap} \, f = 1 \).

If \( x_k \) does not occur in \( p_{j+1} \), then \( a_k = 0 \) and so \( a_j = 1 \). Then
\[
(15) \quad p_{l+1} = x_i x_j + x_i x_k + x_j x_k + x_i + x_i a_i' + x_j a_j + a,
\]

and every essential variable of \( f \) except for \( x_i \) occurs in \( p_{l+1} \). Thus \( \text{gap} \, f = 1 \).

If \( x_i \) does not occur in \( p_{j+1} \), then \( a_j = a_i + 1 \), and hence \( a_i = a_k \). Consider then
\[
(16) \quad p_{l+1} = x_i(1 + a_i + a_k) + x_j a_j + a = x_k + x_j a_j + a.
\]

Again all essential variables of \( f \) except for \( x_i \) occur in \( p_{l+1} \), and so \( \text{gap} \, f = 1 \).

**Subcase 1.2.3.** Assume that both \( x_i \) and \( x_k \) occur in \( p_{j+1} \) but there is a variable \( x_l \) occurring in \( a_i \) and in \( a_j \), but neither in \( a_k \) nor in \( a \) such that \( x_l \) does not occur in \( p_{j+1} \) (due to some cancelling terms in \( a_i \) and \( a_j \)). Write \( a_i = x_l + a_i', \ a_j = x_l + a_j', \) and consider
\[
(17) \quad p_{j+1} = x_i x_l + x_i x_k + x_l x_k + x_i x_l + x_j a_j' + x_i a_i' + x_j a_j' + x_k a_k + a
\]

Every essential variable of \( f \) except for \( x_j \) occurs in \( p_{j+1} \), and so \( \text{gap} \, f = 1 \).

**Case 2.** Assume then that \( p \) is of the form
\[
(18) \quad p = x_i x_j + x_i x_k a_i k + x_i a_i + x_j a_j + x_k a_k + a,
\]

where \( a_i k \) is a polynomial of degree \( 0; \ a_i, a_j, a_k \) are polynomials of degree at most 1; and \( a \) is a polynomial of degree at most 2 such that variables \( x_i, x_j, x_k \) do not occur in \( a_i k, a_i, a_j, a_k, a \). Note that \( a_i k \) and \( a_k \) cannot both be 0, for otherwise \( x_k \) would not occur in \( p \). Consider
\[
(19) \quad p_{j+1} = x_i(1 + a_i + a_j) + x_i x_k a_i k + x_k a_k + a.
\]

By the above observation that \( a_i k \) and \( a_k \) are not both 0, \( x_k \) occurs in \( p_{j+1} \). If all essential variables of \( f \) except for \( x_j \) occur in \( p_{j+1} \), then \( \text{gap} \, f = 1 \) and we are done. Otherwise we distinguish between two cases.

**Subcase 2.1.** Assume that \( x_i \) does not occur in \( p_{j+1} \). In this case \( a_j = a_i + 1 \), \( a_k = 0 \), and \( a_k \neq 0 \). Consider
\[
(20) \quad p_{l+1} = x_j x_k + x_k a_i k + x_k a_i + x_j a_j + x_k a_k + a
\]

Both \( x_j \) and \( x_k \) occur in \( p_{l+1} \), because the term \( x_j x_k \) cannot be cancelled. If any term of \( a_i \) is cancelled by a term of \( a_k \), it still remains in \( x_j a_i \). Thus, all essential variables of \( f \) except for \( x_i \) occur in \( p_{l+1} \), and hence \( \text{gap} \, f = 1 \).

**Subcase 2.2.** Assume that \( x_i \) occurs in \( p_{j+1} \) but there is a variable \( x_l \) occurring in \( a_i \) and \( a_j \), but not in \( a_k, a_k \), nor in \( a \) such that \( x_l \) does not occur in \( p_{j+1} \) (due to some cancelling terms in \( a_i \) and \( a_j \)). Consider
\[
(21) \quad p_{k+1} = x_i x_j + x_i x_l a_i k + x_i a_i + x_j a_j + x_l a_k + a.
\]
If $a_k = 1$, then the terms $x_i x_j$ in $x_i a_i$ and in $x_i x_j a_k$ cancel each other. These are the only terms that may be cancelled out. Nevertheless, $x_i$ occurs also in $a_j$, and so all essential variables of $f$ except for $x_k$ occur in $p_{k-1}$. Therefore gap $f = 1$ also in this case.

Proof of Theorem 4.1. Denote the Zhegalkin polynomial of $f$ by $p$. It is straightforward to verify that if $p$ has one of the special forms listed in the statement of the theorem, then $f$ does not have a variable identification minor of essential arity $\text{ess } f - 1$ but it has one of essential arity $\text{ess } f - 2$. For the converse implication, we will prove by induction on $\text{ess } f$ that if $p$ is not of any of the special forms, then there is a variable identification minor of essential arity $g$ such that $\text{ess } g = \text{ess } f - 1$, i.e., $f$ has arity gap $1$.

If $\text{ess } f = 2$ and $p$ is not of any of the special forms, then $p = x_i x_j + c$ or $p = x_i x_j + x_i + x_j + c$ where $c \in \{0, 1\}$, and in both cases $p_{j-1} = x_i + c$. In this case gap $f = 1$.

If $\text{ess } f = 3$, then $p$ has one of the following forms:

- $x_i x_j x_k + x_i x_j + x_i x_k + x_j x_k + a_i x_i + a_j x_j + a_k x_k + c$,
- $x_i x_j x_k + x_i x_j + a_i x_i + a_j x_j + a_k x_k + c$,
- $x_i x_j x_k + a_i x_i + a_j x_j + a_k x_k + c$,
- $x_i x_j + x_i x_k + x_j x_k + x_k + c$,
- $x_i x_j + x_i x_k + a_i x_i + a_j x_j + a_k x_k + c$,
- $x_i x_k + a_i x_i + a_j x_j + a_k x_k + c$,
- $x_i x_k + a_i x_i + a_j x_j + a_k x_k + c$,

where $a_i, a_j, a_k, c \in \{0, 1\}$. It is easy to verify that in each case $p_{j-1}$ contains the term $x_i x_k$, and hence both $x_i$ and $x_k$ are essential in $f_{j-1}$, and so gap $f = 1$.

For the sake of induction, assume then that the claim holds for $2 \leq \text{ess } f < n$, $n \geq 4$. Consider the case that $\text{ess } f = n$. Since the case where $\text{deg } p = 1$ is ruled out by the assumption that $p$ does not have any of the special forms and the case where $\text{deg } p = 2$ is settled by Lemma 4.2, we can assume that $\text{deg } p \geq 3$. Choose a variable $x_m$ from a term of the highest possible degree in $p$, and write

$$p = x_m q + r,$$

where the polynomials $q$ and $r$ do not contain $x_m$. We clearly have that $\text{deg } q = \text{deg } p - 1$, and $q$ and $r$ represent functions with less than $n$ essential variables. Of course, every essential variable of $f$ except for $x_m$ occurs in $q$ or $r$. We have three different cases to consider, depending on the comparability under inclusion of the sets of variables occurring in $q$ and $r$.

Case 1. Assume that there is a variable $x_i$ that occurs in $q$ but does not occur in $r$, and there is a variable $x_j$ that occurs in $r$ but does not occur in $q$. Write

$$q = x_i q' + q'', \quad r = x_j t' + t'' ,$$

where $q'$, $q''$, $t'$, $t''$ do not contain $x_i$, $x_j$. Then

$$p = x_m x_i q' + x_m q'' + x_j t' + t'' ,$$

and we have that

$$p_{j-1} = x_m x_i q' + x_m q'' + x_i t' + t'' ,$$

where no terms can cancel. Hence all essential variables of $f$ except for $x_j$ are essential in $f_{j-1}$, and so gap $f = 1$. 

□
Assume that every variable occurring in \( r \) occurs in \( q \). In this case \( q \) represents a function \( q \) of essential arity \( \text{ess } f - 1 \), containing all essential variables of \( f \) except for \( x_m \). We also have that \( \deg q = \deg p - 1 \geq 2 \).

**Subcase 2.1.** If \( \text{ess } f \geq 5 \), then \( \text{ess } q \geq 4 \), and we can apply the inductive hypothesis, which tells us that there are variables \( x_i \) and \( x_j \) such that \( \text{ess } q_{i\sim j} = \text{ess } q - 1 \). Hence \( f_{i\sim j} \) is represented by the polynomial \( p_{i\sim j} = x_m q_{i\sim j} + t_{i\sim j} \), and all essential variables of \( f \) except for \( x_i \) occur in \( p_{i\sim j} \), since no terms can cancel between \( x_m q_{i\sim j} \) and \( t_{i\sim j} \). Thus gap \( f = 1 \).

**Subcase 2.2.** If \( \text{ess } f = 4 \), then \( \text{ess } q = 3 \), and we can apply the inductive hypothesis as above unless \( q = x_i x_j + x_i x_k + x_j x_k + c \) or \( q = x_i x_j + x_i x_k + x_j x_k + x_i + x_j + c \). If this is the case, consider first the case where \( q \) contains a variable \( x_l \in \{ x_i, x_j, x_k \} \) that does not occur in \( \tau \). Consider then

\[
\begin{align*}
\text{Case 2. Assume that every variable occurring in } q \text{ occurs in } \tau \text{ but there is a variable } x_l \text{ that occurs in } \tau \text{ but does not occur in } q. \text{ If } \deg \tau = 1, \text{ then } \tau = x_l + \tau' \text{ where } \tau' \text{ does not contain } x_l. \text{ Then } p_{m\sim l} = x_l q + x_l + \tau', \text{ where the only term that may cancel out is } x_l, \text{ and this happens if } q \text{ has a constant term 1. Nevertheless, } x_l \text{ occurs in } t_{m\sim l} \text{ because } \deg q \geq 2. \text{ Of course, all other essential variables of } f \text{ except for } x_m \text{ also occur in } p_{m\sim l}, \text{ so gap } f = 1. \text{ We may thus assume that } \deg \tau \geq 2.
\end{align*}
\]

**Subcase 3.1.** Assume first that \( \text{ess } f = 4 \) (in which case \( \tau \) contains three variables and \( q \) contains at most two variables) and \( \tau = x_i x_j + x_i x_k + x_k x_j + c \) or \( \tau = x_i x_j + x_i x_k + x_j x_k + x_i + x_j + c \). Since we assume that \( \deg p \geq 3 \), we have that \( \deg q \geq 2 \) and hence \( q \) contains at least two variables. Thus exactly two variables occur in \( q \) and so also \( \deg q = 2 \). Then \( q = x_\alpha x_\beta + b_1 x_\alpha + b_2 x_\beta + d \) where \( \alpha, \beta \in \{ i, j, k \} \) and \( b_1, b_2, d \in \{ 0, 1 \} \). Let \( \gamma \in \{ i, j, k \} \setminus \{ \alpha, \beta \} \). Then \( p_{m\sim \gamma} \) contains the term \( x_i x_j x_k \), and hence all essential variables of \( f \) except for \( x_m \) occur in \( p_{m\sim \gamma} \), and so gap \( f = 1 \).

**Subcase 3.2.** Assume then that \( \text{ess } f > 4 \) or \( \text{ess } f = 4 \) but \( \tau \) does not have any of the special forms. In this case we can apply the inductive hypothesis on the function \( \tau \) represented by \( \tau \). Let \( x_i \) and \( x_j \) be such that \( \text{ess } \tau_{i\sim j} = \text{ess } \tau - 1 \). If \( q_{i\sim j} \neq 0 \), then \( x_m \) and all other essential variables of \( f \) except for \( x_i \) occur in \( p_{j\sim i} \), and we are done—the arity gap of \( f \) is 1. We may thus assume that \( q_{j\sim i} = 0 \). Write \( q \) and \( \tau \) in the form

\[
\begin{align*}
q &= x_i x_j a_1 + x_i a_2 + x_j a_3 + a_4, \\
\tau &= x_i x_j b_1 + x_j b_2 + x_i b_3 + b_4.
\end{align*}
\]
where the polynomials $a_1$, $a_2$, $a_3$, $a_4$, $b_1$, $b_2$, $b_3$, $b_4$ do not contain $x_i$, $x_j$. Define the polynomials $q_1$, \ldots, $q_7$ as follows (cf. the proof of Theorem 4 in Salomaa [14]):

$q_1$ consists of the terms common to $a_1$, $a_2$, and $a_3$.
$q_i$, $i = 2, 3$, consists of those terms common to $a_1$ and $a_i$ which are not in $q_1$.
$q_4$ consists of those terms common to $a_2$ and $a_3$ which are not in $q_1$.
$q_{i+j}$, $i = 1, 2, 3$, consists of the remaining terms in $a_i$.

Define the polynomials and $r_1$, \ldots, $r_7$ similarly in terms of the $b_i$'s. Note that for any $i \neq j$, $q_i$ and $q_j$ do not have any terms in common, and similarly $r_i$ and $r_j$ do not have any terms in common. Hence,

\[
q = x_i x_j (q_1 + q_2 + q_3 + q_5) + \\
x_i (q_1 + q_2 + q_4 + q_6) + \\
x_j (q_1 + q_3 + q_4 + q_7) + a_4,
\]

(29)

\[
r = x_i x_j (r_1 + r_2 + r_3 + r_5) + \\
x_i (r_1 + r_2 + r_4 + r_6) + \\
x_j (r_1 + r_3 + r_4 + r_7) + b_4.
\]

(30)

Identification of $x_i$ with $x_j$ yields

\[
q_{j \leftarrow i} = x_i (q_1 + q_5 + q_6 + q_7) + a_4,
\]

(31)

\[
r_{j \leftarrow i} = x_i (r_1 + r_5 + r_6 + r_7) + b_4.
\]

(32)

Since we are assuming that $q_{j \leftarrow i} = 0$, we have that $q_1 = q_5 = q_6 = q_7 = a_4 = 0$. On the other hand, $q \neq 0$, so $q_2$, $q_3$, $q_4$ are not all zero. Thus

\[
q = x_i x_j (q_2 + q_3) + x_i (q_2 + q_4) + x_j (q_3 + q_4).
\]

(33)

All essential variables of $f$ except for $x_j$ are contained in $r_{j \leftarrow i}$.

**Subcase 3.2.1.** Assume that there is a variable $x_l$ occurring in $b_4$ that does not occur in $r_1$, $r_5$, $r_6$, $r_7$. Consider

\[
p_{m \leftarrow i} = x_l q + r = x_l q + x_i x_j b_1 + x_i b_2 + x_j b_3 + b_4.
\]

(34)

Cancelling may only happen between a term of $x_l q$ and a term of $r$. No term of $b_4$ can be cancelled, because every term of $x_l q$ contains $x_i$ or $x_j$ but the terms of $b_4$ do not contain either. The variables that do not occur in $b_4$ occur in some terms of $b_1$, $b_2$, $b_3$ that do not contain $x_l$. Thus, all essential variables of $f$ except for $x_m$ occur in $p_{m \leftarrow i}$, and so in this case $f$ has arity gap 1.

**Subcase 3.2.2.** Assume that all variables of $r$ except for $x_i$, $x_j$ occur already in $r_1 + r_5 + r_6 + r_7$. Consider

\[
p_{m \leftarrow i} = x_i x_j (q_2 + q_4 + r_1 + r_2 + r_3 + r_5) + \\
x_i (q_2 + q_4 + r_1 + r_2 + r_4 + r_6) + \\
x_j (r_1 + r_3 + r_4 + r_7) + b_4.
\]

(35)

**Subcase 3.2.2.1.** Assume first that $x_i$ does not occur in $p_{m \leftarrow i}$ in (35). Then

\[
q_2 + q_4 + r_1 + r_2 + r_3 + r_5 = 0,
\]

(36)

\[
q_2 + q_4 + r_1 + r_2 + r_4 + r_6 = 0,
\]

(37)

and since the $r_i$’s do not have terms in common, we have that

\[
r_1 + r_2 = q_2 + q_4, \quad r_3 = r_4 = r_5 = r_6 = 0.
\]

(38)
Then all variables of \( r \) except for \( x_i, x_j \) occur already in \( r_1 + r_7 \). Consider

\[
p_{m \rightarrow j} = x_i x_j (q_3 + q_4 + t_1 + t_2 + t_3 + t_5) + \\
x_i (t_1 + t_2 + t_4 + t_6) + \\
x_j (q_3 + q_4 + t_1 + t_3 + t_4 + t_7) + b_4
\]

\[
= x_i x_j (q_2 + q_3) + \\
x_i (t_1 + t_2) + \\
x_j (q_2 + q_3 + t_2 + t_7) + b_4.
\]

(39)

All variables of \( r_1 \) are there on the fifth line of (39). If a term of \( r_7 \) is cancelled by a term of \( q_2 + q_3 \) on the sixth line, it still remains on the fourth line, so all variables of \( r_7 \) are also there. We still need to verify that the variables \( x_i \) and \( x_j \) are not cancelled out from (39). If \( q_2 + q_3 \neq 0 \) then we are done. Assume then that \( q_2 + q_3 = 0 \), in which case \( q_4 \neq 0 \). Since

\[
r_1 + r_2 + r_4 + r_6 = r_1 + r_2 = q_2 + q_4 = q_4 \neq 0,
\]

we have \( x_i \) in (39). Since

\[
q_3 + q_4 + t_1 + t_3 + t_4 + t_7 = q_4 + t_1 + t_7
\]

and \( r_1 + t_7 \) contains all variables of \( r \) except for \( x_i, x_j \), but \( q_4 \) does not, \( q_4 + t_1 + t_7 \neq 0 \), so we also have \( x_j \) in (39). Thus, the arity gap of \( f \) equals 1 in this case.

Subcase 3.2.2.2. Assume then that \( x_i \) occurs in \( p_{m \rightarrow i} \) in (35). Nothing cancels out on the third line of (35), and therefore the variables of \( r_1 \) and \( r_7 \) occur in \( p_{m \rightarrow i} \). Terms of \( r_5 \) may be cancelled out by terms of \( q_2 + q_4 \) on the first line of (35) but such terms will remain on the second line. Thus the variables of \( r_5 \) occur in \( p_{m \rightarrow i} \). A similar argument shows that the variables of \( r_6 \) also occur in \( p_{m \rightarrow i} \). In order for \( f \) to have arity gap 1, we still need to verify that \( x_j \) occurs in \( p_{m \rightarrow i} \). If \( q_2 + q_4 + t_1 + t_2 + t_3 + t_5 \neq 0 \), then we are done. We may thus assume that

\[
q_2 + q_4 + t_1 + t_2 + t_3 + t_5 = 0.
\]

By the assumption that \( x_i \) occurs in \( p_{m \rightarrow i} \), the second line of (35) does not vanish, i.e.,

\[
0 \neq q_2 + q_4 + t_1 + t_2 + t_4 + t_6 = t_3 + t_4 + t_5 + t_6.
\]

If the third line of (35) does not vanish either, i.e., \( t_1 + t_2 + t_4 + t_7 \neq 0 \), then we have both \( x_i \) and \( x_j \) and we are done. We may thus assume that \( t_1 + t_2 + t_4 + t_7 = 0 \), i.e., \( t_1 = t_3 = t_4 = t_7 = 0 \). Then all variables of \( r \) except for \( x_i, x_j \) occur already in \( r_5 + r_6 \). Equation (42) implies that \( t_2 + t_5 = q_2 + q_4 \). Consider

\[
p_{m \rightarrow j} = x_i x_j (q_3 + q_4 + t_1 + t_2 + t_3 + t_5) + \\
x_i (t_1 + t_2 + t_4 + t_6) + \\
x_j (q_3 + q_4 + t_1 + t_3 + t_4 + t_7) + b_4
\]

\[
= x_i x_j (q_2 + q_3) + \\
x_i (q_2 + q_4 + t_5 + t_6) + \\
x_j (q_3 + q_4) + b_4.
\]

(44)

Assume first that \( q_2 + q_3 = 0 \), in which case \( q_4 \neq 0 \). If a term of \( r_5 + r_6 \) is cancelled by a term of \( q_4 \) on the fifth line of (44), it will still remain on the sixth line. Therefore we have in \( p_{m \rightarrow j} \) all variables of \( r \) except for \( x_i \) and \( x_j \). Since \( t_5 + t_6 \) contains all variables of \( r \) except for \( x_i, x_j \) but \( q_2 + q_4 = q_4 \) does not, the fifth line of (44) does...
not vanish, and so we have $x_i$. We also have $x_j$ because $q_3 + q_4 = q_4 \neq 0$ on the sixth line. In this case $f$ has arity gap 1.

Assume then that $q_2 + q_3 \neq 0$. Then the fourth line of (44) does not vanish and both $x_i$ and $x_j$ occur in $p_{m+j}$. If any term of $t_5 + t_6$ is cancelled by a term of $q_2$ on the fifth line of (44), it still remains on the fourth line, and if it is cancelled by a term of $q_1$, it remains on the sixth line. Thus all variables of $t$ occur in $p_{m+j}$, and $f$ has arity gap 1 again. This completes the proof of Theorem 4.1. □

5. Concluding remarks

We do not know whether the upper bound on arity gap given by Theorem 3.1 is sharp. For base sets $A$ with $k \geq 3$ elements, we do not know whether there exists an operation $f$ on $A$ with $\text{ess} f \geq k + 1$ and $\text{gap} f \geq 3$. We know that for all $k \geq 2$, there are operations on a $k$-element set $A$ with arity gap 2. Consider for instance the quasi-linear functions of Burle [1]. A function $f$ is quasi-linear if it has the form

\begin{equation}
(45) \quad f = g(h_1(x_1) \oplus h_2(x_2) \oplus \cdots \oplus h_n(x_n)),
\end{equation}

where $h_1, \ldots, h_n : A \to \{0, 1\}$, $g : \{0, 1\} \to A$ are arbitrary mappings and $\oplus$ denotes addition modulo 2. It is easy to verify that if those $h_i$’s that are nonconstant coincide (and $g$ is not a constant map), then $f$ has arity gap 2.

In general, if there is an operation $f$ on a $k$-element set $A$ with with gap $f = m$, then there are operations of arity gap $m$ on all sets $B$ of at least $k$ elements. Namely, it is easy to see that any operation $g$ on $B$ of the form

\begin{equation}
(46) \quad g = \phi(f(\gamma(x_1), \gamma(x_2), \ldots, \gamma(x_n))),
\end{equation}

where $\gamma : B \to A$ is surjective and $\phi : A \to B$ is injective, satisfies $\text{ess} g = \text{ess} f$ and $\text{gap} g = \text{gap} f$.

References


(M. Couceiro) Department of Mathematics, Statistics and Philosophy, University of Tampere, FI-33014 Tampereen yliopisto, Finland

E-mail address: miguel.couceiro@uta.fi

(E. Lehtonen) Institute of Mathematics, Tampere University of Technology, P.O. Box 553, FI-33101 Tampere, Finland

E-mail address: erkko.lehtonen@tut.fi