

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

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, 6, 11, 13]. In this paper, we discuss the problem how the number of essential variables is affected by identification of variables (diagonalization). Salomaa [11] 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. *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 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, 7, 8, 9, 10, 12, 14]. In the case of Boolean functions, this quasi-order was studied in [4] where Theorem 2 was fundamental in deriving certain bounds on the essential arity of functions.

In this paper, we will generalize Theorem 2 to operations on arbitrary finite sets in Theorem 3. We will also strengthen Theorem 2 on Boolean functions in Theorem 6 by determining the Boolean functions f for which there exists a function g obtained from f by identification of variables such that $\text{ess } g = \text{ess } f - 1$.

2. VARIABLE IDENTIFICATION MINORS

Let A and B be arbitrary nonempty sets. A B -valued function of several variables on A is a mapping $f : A^n \rightarrow B$ for some positive integer n , called the *arity* of f .

Date: July 25, 2021.

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 $a_1, \dots, a_n, b \in A$ such that

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

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 $\sigma : \{1, \dots, n\} \rightarrow \{1, \dots, m\}$ such that

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

In the particular case that $n = m$ and σ is a permutation of $\{1, \dots, n\}$, we say that g is obtained from f by *permutation of variables*. For indices $i, j \in \{1, \dots, n\}$, $i \neq j$, if x_i and x_j are essential in f , then the function $f_{i \leftarrow j}$ obtained from f by the simple variable substitution

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

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

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

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

3. GENERALIZATION OF THEOREM 2

Theorem 3. *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, we will make use of the following theorem due to Salomaa [11, Theorem 1], which is a strengthening of Yablonski's [13] "fundamental lemma".

Theorem 4. *Let the function $f : M_1 \times \dots \times M_n \rightarrow N$ 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*

$$f(x_1, \dots, x_{j-1}, c, x_{j+1}, \dots, x_n)$$

depends essentially on all of its $n - 1$ variables.

We also need the following auxiliary lemma.

Lemma 5. *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, \dots, x_{k+1} is essential in $f_{i \leftarrow j}$.*

Proof. Since x_1 is essential in f , there are elements $a_1, \dots, a_n, b \in A$ such that

$$f(a_1, a_2, \dots, a_n) \neq f(b, a_2, \dots, a_n).$$

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 $f_{i \leftarrow j}$. If there are no such i and j with $i \neq 1$, then $i = 1 < j$ and we have that $b = a_l$ for some $1 < l \leq k+1$, $l \neq j$. For $m = 1, \dots, n$, let $c_m = a_m$ if $m \notin \{1, j, l\}$ and let $c_m = a_1$ if $m \in \{1, j, l\}$. Then $f(c_1, c_2, \dots, c_n)$ is distinct from at least one of $f(a_1, a_2, \dots, a_n)$ and $f(b, a_2, \dots, a_n)$. If $f(c_1, c_2, \dots, c_n) \neq f(a_1, a_2, \dots, a_n)$, then x_l is essential in $f_{1 \leftarrow j}$. If $f(c_1, c_2, \dots, c_n) \neq f(b, a_2, \dots, a_n)$, then x_l is essential in $f_{1 \leftarrow l}$. \square

Proof of Theorem 3. By Theorem 4, there exist $k+1$ constants $c_1, \dots, c_{k+1} \in A$ such that, after a suitable permutation of variables, the function

$$f(c_1, \dots, c_{k+1}, x_{k+2}, \dots, x_n)$$

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 5 there are indices $1 \leq l < m \leq k+1$ such that at least one of the variables x_1, \dots, x_{k+1} is essential in $f_{l \leftarrow m}$. 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, \dots, x_{k+1} is essential in $f_{1 \leftarrow 2}$, then we are done. Otherwise we have that for all $a_{k+2}, \dots, a_n \in A$,

$$f(c_1, c_1, c_3, c_4, \dots, c_{k+1}, a_{k+2}, \dots, a_n) = f(c_3, c_3, c_3, c_4, \dots, c_{k+1}, a_{k+2}, \dots, a_n).$$

Thus the variables x_{k+2}, \dots, x_n are essential in $f_{2 \leftarrow 3}$. If one of the variables x_1, \dots, x_{k+1} is essential in $f_{2 \leftarrow 3}$, then we are done. Otherwise we have that for all $a_{k+2}, \dots, a_n \in A$,

$$f(c_3, c_3, c_3, c_4, \dots, c_{k+1}, a_{k+2}, \dots, a_n) = f(c_3, c_4, c_4, c_4, \dots, c_{k+1}, a_{k+2}, \dots, a_n),$$

and so the variables x_{k+2}, \dots, x_n are essential in $f_{3 \leftarrow 4}$ and also at least one of x_1, \dots, x_{k+1} is essential in $f_{3 \leftarrow 4}$. \square

We would like to remark that our proof is considerably simpler than Salomaa's original proof of Theorem 2.

4. STRENGTHENING OF THEOREM 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* of f . 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 $\deg \mathbf{p}$ the degree of polynomial \mathbf{p} . If \mathbf{p} is the Zhegalkin polynomial of f , then we denote the Zhegalkin polynomial of $f_{i \leftarrow j}$ by $\mathbf{p}_{i \leftarrow j}$. Note that the only polynomials of degree 0 are the constant polynomials.

Theorem 6. *Let f be a Boolean function with at least 2 essential variables. Then the arity gap of f is 2 if and only if the Zhegalkin polynomial of f is of one of the following special forms:*

$$\bullet \quad x_{i_1} + x_{i_2} + \dots + x_{i_n} + c,$$

- $x_i x_j + x_i + c$,
- $x_i x_j + x_i x_k + x_j x_k + c$,
- $x_i x_j + x_i x_k + x_j x_k + x_i + x_j + c$,

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

We prove first an auxiliary lemma that takes care of the functions of essential arity at least 4 whose Zhegalkin polynomial has degree 2.

Lemma 7. *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 \mathbf{p} . We need to consider several cases and subcases.

Case 1. Assume first that \mathbf{p} is of the form

$$\mathbf{p} = x_i x_j + x_i x_k + x_j x_k + x_i \mathbf{a}_i + x_j \mathbf{a}_j + x_k \mathbf{a}_k + \mathbf{a},$$

where $\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k$ are polynomials of degree at most 1 and \mathbf{a} is a polynomial of degree at most 2 such that there are no occurrences of variables x_i, x_j, x_k in $\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k, \mathbf{a}$.

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

$$\mathbf{p}_{l \leftarrow i} = x_i x_j + x_i x_k + x_j x_k + x_i \mathbf{a}' + \mathbf{a}'',$$

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

Subcase 1.2. Assume that at least one of $\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k$ has degree 1, say $\deg \mathbf{a}_i = 1$. Then \mathbf{a}_i contains a variable x_l distinct from x_i, x_j, x_k , and so $\mathbf{a}_i = x_l + \mathbf{a}'_i$, where \mathbf{a}'_i has degree at most 1 and does not contain x_l . Consider

$$\mathbf{p}_{j \leftarrow k} = x_k(1 + \mathbf{a}_j + \mathbf{a}_k) + x_i \mathbf{a}_i + \mathbf{a}.$$

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

Subcase 1.2.1. Assume that variable x_k occurs in $\mathbf{p}_{j \leftarrow k}$ but there is a variable x_l that occurs in \mathbf{a}_j and \mathbf{a}_k but not in \mathbf{a}_i nor in \mathbf{a} such that x_l does not occur in $\mathbf{p}_{j \leftarrow k}$ (due to some cancelling terms in \mathbf{a}_j and \mathbf{a}_k). Write $\mathbf{a}_j = x_l + \mathbf{a}'_j$, $\mathbf{a}_k = x_l + \mathbf{a}'_k$, and consider

$$\begin{aligned} \mathbf{p}_{j \leftarrow l} &= x_i x_l + x_i x_k + x_l x_k + x_i \mathbf{a}_i + x_l + x_l \mathbf{a}'_j + x_k x_l + x_k \mathbf{a}'_k + \mathbf{a} \\ &= x_i x_l + x_i x_k + x_i \mathbf{a}_i + x_l + x_l \mathbf{a}'_j + x_k \mathbf{a}'_k + \mathbf{a}. \end{aligned}$$

Every essential variable of f except for x_j occurs in $\mathbf{p}_{j \leftarrow l}$, and hence $\text{gap } f = 1$.

Subcase 1.2.2. Assume that x_k does not occur in $\mathbf{p}_{j \leftarrow k}$. In this case $\mathbf{a}_j = \mathbf{a}_k + 1$. Consider

$$\mathbf{p}_{j \leftarrow i} = x_i(1 + \mathbf{a}_i + \mathbf{a}_j) + x_k \mathbf{a}_k + \mathbf{a}.$$

If any term of \mathbf{a}_j is cancelled by a term of \mathbf{a}_i , it still remains as a term of \mathbf{a}_k , and hence all variables occurring in $\mathbf{a}_i, \mathbf{a}_j, \mathbf{a}_k$ occur in $\mathbf{p}_{j \leftarrow i}$. If both x_i and x_k also occur in $\mathbf{p}_{j \leftarrow i}$, then all essential variables of f except for x_j occur in $\mathbf{p}_{j \leftarrow i}$, and so $\text{gap } f = 1$.

If x_k does not occur in $\mathbf{p}_{j \leftarrow i}$, then $\mathbf{a}_k = 0$ and so $\mathbf{a}_j = 1$. Then

$$\mathbf{p}_{l \leftarrow i} = x_i x_j + x_i x_k + x_j x_k + x_i + x_i \mathbf{a}'_i + x_j + \mathbf{a},$$

and every essential variable of f except for x_l occurs in $\mathbf{p}_{l \leftarrow i}$. Thus $\text{gap } f = 1$.

If x_i does not occur in $\mathbf{p}_{j \leftarrow i}$, then $\mathbf{a}_j = \mathbf{a}_i + 1$, and hence $\mathbf{a}_i = \mathbf{a}_k$. Consider then

$$\mathbf{p}_{i \leftarrow k} = x_k(1 + \mathbf{a}_i + \mathbf{a}_k) + x_j \mathbf{a}_j + \mathbf{a} = x_k + x_j \mathbf{a}_j + \mathbf{a}.$$

Again all essential variables of f except for x_i occur in $\mathbf{p}_{i \leftarrow k}$, and so $\text{gap } f = 1$.

Subcase 1.2.3. Assume that both x_i and x_k occur in $\mathbf{p}_{j \leftarrow k}$ but there is a variable x_l occurring in \mathbf{a}_i and in \mathbf{a}_j but not in \mathbf{a}_k nor in \mathbf{a} such that x_l does not occur in $\mathbf{p}_{j \leftarrow k}$ (due to some cancelling terms in \mathbf{a}_i and \mathbf{a}_j). Write $\mathbf{a}_i = x_l + \mathbf{a}'_i$, $\mathbf{a}_j = x_l + \mathbf{a}'_j$, and consider

$$\begin{aligned} \mathbf{p}_{j \leftarrow l} &= x_i x_l + x_i x_k + x_l x_k + x_i x_l + x_i \mathbf{a}'_i + x_l + x_l \mathbf{a}'_j + x_k \mathbf{a}_k + \mathbf{a} \\ &= x_i x_k + x_l x_k + x_i \mathbf{a}'_i + x_l + x_l \mathbf{a}'_j + x_k \mathbf{a}_k + \mathbf{a}. \end{aligned}$$

Every essential variable of f except for x_j occurs in $\mathbf{p}_{j \leftarrow l}$, and so $\text{gap } f = 1$.

Case 2. Assume then that \mathbf{p} is of the form

$$\mathbf{p} = x_i x_j + x_i x_k \mathbf{a}_{ik} + x_i \mathbf{a}_i + x_j \mathbf{a}_j + x_k \mathbf{a}_k + \mathbf{a},$$

where \mathbf{a}_{ik} is a polynomial of degree 0; \mathbf{a}_i , \mathbf{a}_j , \mathbf{a}_k are polynomials of degree at most 1; and \mathbf{a} is a polynomial of degree at most 2 such that variables x_i , x_j , x_k do not occur in \mathbf{a}_{ik} , \mathbf{a}_i , \mathbf{a}_j , \mathbf{a}_k , \mathbf{a} . Note that \mathbf{a}_{ik} and \mathbf{a}_k cannot both be 0, for otherwise x_k would not occur in \mathbf{p} . Consider

$$\mathbf{p}_{j \leftarrow i} = x_i(1 + \mathbf{a}_i + \mathbf{a}_j) + x_i x_k \mathbf{a}_{ik} + x_k \mathbf{a}_k + \mathbf{a}.$$

By the above observation that \mathbf{a}_{ik} and \mathbf{a}_k are not both 0, x_k occurs in $\mathbf{p}_{j \leftarrow i}$. If all essential variables of f except for x_j occur in $\mathbf{p}_{j \leftarrow i}$, 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 $\mathbf{p}_{j \leftarrow i}$. In this case $\mathbf{a}_j = \mathbf{a}_i + 1$, $\mathbf{a}_{ik} = 0$, and $\mathbf{a}_k \neq 0$. Consider

$$\begin{aligned} \mathbf{p}_{i \leftarrow k} &= x_j x_k + x_k \mathbf{a}_{ik} + x_k \mathbf{a}_i + x_j \mathbf{a}_j + x_k \mathbf{a}_k + \mathbf{a} \\ &= x_j x_k + x_k(\mathbf{a}_i + \mathbf{a}_k) + x_j + x_j \mathbf{a}_i + \mathbf{a}. \end{aligned}$$

Both x_j and x_k occur in $\mathbf{p}_{i \leftarrow k}$, because the term $x_j x_k$ cannot be cancelled. If any term of \mathbf{a}_i is cancelled by a term of \mathbf{a}_k , it still remains in $x_j \mathbf{a}_i$. Thus, all essential variables of f except for x_i occur in $\mathbf{p}_{i \leftarrow k}$, and hence $\text{gap } f = 1$.

Subcase 2.2. Assume that x_i occurs in $\mathbf{p}_{j \leftarrow i}$ but there is a variable x_l occurring in \mathbf{a}_i and \mathbf{a}_j but not in \mathbf{a}_{ik} , \mathbf{a}_k , nor in \mathbf{a} such that x_l does not occur in $\mathbf{p}_{j \leftarrow i}$ (due to some cancelling terms in \mathbf{a}_i and \mathbf{a}_j). Consider

$$\mathbf{p}_{k \leftarrow l} = x_i x_j + x_i x_l \mathbf{a}_{ik} + x_i \mathbf{a}_i + x_j \mathbf{a}_j + x_l \mathbf{a}_k + \mathbf{a}.$$

If $\mathbf{a}_{ik} = 1$, then the terms $x_i x_l$ in $x_i \mathbf{a}_i$ and in $x_i x_l \mathbf{a}_{ik}$ cancel each other. These are the only terms that may be cancelled out. Nevertheless, x_l occurs also in \mathbf{a}_j , and so all essential variables of f except for x_k occur in $\mathbf{p}_{k \leftarrow l}$. Therefore $\text{gap } f = 1$ also in this case. \square

Proof of Theorem 6. Denote the Zhegalkin polynomial of f by \mathbf{p} . It is straightforward to verify that if \mathbf{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 \mathbf{p} is not of any of the special forms, then there is a variable identification minor g of f such that $\text{ess } g = \text{ess } f - 1$, i.e., f has arity gap 1.

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

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

$$\begin{aligned} & x_i x_j x_k + x_i x_j + x_i x_k + x_j x_k + a_i x_i + a_j x_j + a_k x_k + c, \\ & x_i x_j x_k + x_i x_k + x_j x_k + a_i x_i + a_j x_j + a_k x_k + c, \\ & x_i x_j x_k + x_i x_j + a_i x_i + a_j x_j + a_k x_k + c, \\ & x_i x_j + x_i x_k + x_j x_k + x_k + c, \\ & x_i x_j + x_i x_k + x_j x_k + x_i + x_j + x_k + c, \\ & x_i x_j + x_i x_k + a_i x_i + a_j x_j + a_k x_k + c, \\ & x_i x_k + a_i x_i + a_j x_j + a_k x_k + c, \end{aligned}$$

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

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 $\deg \mathbf{p} = 1$ is ruled out by the assumption that \mathbf{p} does not have any of the special forms and the case where $\deg \mathbf{p} = 2$ is settled by Lemma 7, we can assume that $\deg \mathbf{p} \geq 3$. Choose a variable x_m from a term of the highest possible degree in \mathbf{p} , and write

$$\mathbf{p} = x_m \mathbf{q} + \mathbf{r},$$

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

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

$$\mathbf{q} = x_i \mathbf{q}' + \mathbf{q}'', \quad \mathbf{r} = x_j \mathbf{r}' + \mathbf{r}'',$$

where $\mathbf{q}', \mathbf{q}'', \mathbf{r}', \mathbf{r}''$ do not contain x_i, x_j . Then

$$\mathbf{p} = x_m x_i \mathbf{q}' + x_m \mathbf{q}'' + x_j \mathbf{r}' + \mathbf{r}'',$$

and we have that

$$\mathbf{p}_{j \leftarrow i} = x_m x_i \mathbf{q}' + x_m \mathbf{q}'' + x_i \mathbf{r}' + \mathbf{r}'',$$

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

Case 2. Assume that every variable occurring in \mathbf{r} occurs in \mathbf{q} . In this case \mathbf{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 $\deg \mathbf{q} = \deg \mathbf{p} - 1 \geq 2$.

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 $\text{ess } q_{i \leftarrow j} = \text{ess } q - 1$. Hence $f_{i \leftarrow j}$ is represented by the polynomial $\mathbf{p}_{i \leftarrow j} = x_m \mathbf{q}_{i \leftarrow j} + \mathbf{r}_{i \leftarrow j}$, and all essential variables of f except for x_i occur in $\mathbf{p}_{i \leftarrow j}$, since no terms can cancel between $x_m \mathbf{q}_{i \leftarrow j}$ and $\mathbf{r}_{i \leftarrow j}$. Thus $\text{gap } f = 1$.

Subcase 2.2. If $\text{ess } f = 4$, then $\text{ess } q = 3$, and we can apply the inductive hypothesis as above unless $\mathbf{q} = x_i x_j + x_i x_k + x_j x_k + c$ or $\mathbf{q} = x_i x_j + x_i x_k + x_j x_k +$

$x_i + x_j + c$. If this is the case, consider first the case where \mathbf{q} contains a variable $x_l \in \{x_i, x_j, x_k\}$ that does not occur in \mathbf{r} . Consider then

$$\mathbf{p}_{m \leftarrow l} = x_l \mathbf{q} + \mathbf{r}.$$

Then $x_l \mathbf{q}$ contains the term $x_i x_j x_k$, which cannot be cancelled. Namely, all other terms of $x_l \mathbf{q}$ have degree at most 2, and since there are at most two variables occurring in \mathbf{r} , the terms of \mathbf{r} also have degree at most 2. Thus, all variables of f except for x_m occur in $\mathbf{p}_{m \leftarrow l}$, and so the arity gap of f is 1.

Consider then the case that \mathbf{q} and \mathbf{r} contain the same variables, i.e., x_i, x_j, x_k . If $\deg \mathbf{r} \leq 2$, then it is easily seen that $\mathbf{p}_{m \leftarrow i}$ contains the term $x_i x_j x_k$, and all essential variables of f except for x_m are essential in $f_{m \leftarrow i}$. Otherwise, we can apply the inductive hypothesis on the function r represented by \mathbf{r} and we obtain variables x_α and x_β such that $\text{ess } r_{\alpha \leftarrow \beta} = \text{ess } r - 1$. It can be easily verified that no identification of variables brings \mathbf{q} into the zero polynomial, so x_m and two other variables will occur in $\mathbf{p}_{\alpha \leftarrow \beta} = x_m \mathbf{q}_{\alpha \leftarrow \beta} + \mathbf{r}_{\alpha \leftarrow \beta}$. We have that $\text{gap } f = 1$ also in this case.

Case 3. Assume that every variable occurring in \mathbf{q} occurs in \mathbf{r} but there is a variable x_l that occurs in \mathbf{r} but does not occur in \mathbf{q} . If $\deg \mathbf{r} = 1$, then $\mathbf{r} = x_l + \mathbf{r}'$ where \mathbf{r}' does not contain x_l . Then $\mathbf{p}_{m \leftarrow l} = x_l \mathbf{q} + x_l + \mathbf{r}'$, where the only term that may cancel out is x_l , and this happens if \mathbf{q} has a constant term 1. Nevertheless, x_l occurs in $\mathbf{r}_{m \leftarrow l}$ because $\deg \mathbf{q} \geq 2$. Of course, all other essential variables of f except for x_m also occur in $\mathbf{p}_{m \leftarrow l}$, so $\text{gap } f = 1$. We may thus assume that $\deg \mathbf{r} \geq 2$.

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

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

$$\begin{aligned} \mathbf{q} &= x_i x_j \mathbf{a}_1 + x_i \mathbf{a}_2 + x_j \mathbf{a}_3 + \mathbf{a}_4, \\ \mathbf{r} &= x_i x_j \mathbf{b}_1 + x_i \mathbf{b}_2 + x_j \mathbf{b}_3 + \mathbf{b}_4, \end{aligned}$$

where the polynomials $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3, \mathbf{a}_4, \mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3, \mathbf{b}_4$ do not contain x_i, x_j . Define the polynomials $\mathbf{q}_1, \dots, \mathbf{q}_7$ as follows (cf. the proof of Theorem 4 in Salomaa [11]):

\mathbf{q}_1 consists of the terms common to $\mathbf{a}_1, \mathbf{a}_2$, and \mathbf{a}_3 .

$\mathbf{q}_i, i = 2, 3$, consists of those terms common to \mathbf{a}_1 and \mathbf{a}_i which are not in \mathbf{q}_1 .

\mathbf{q}_4 consists of those terms common to \mathbf{a}_2 and \mathbf{a}_3 which are not in \mathbf{q}_1 .

$\mathbf{q}_{4+i}, i = 1, 2, 3$, consists of the remaining terms in \mathbf{a}_i .

Define the polynomials and $\mathbf{r}_1, \dots, \mathbf{r}_7$ similarly in terms of the \mathbf{b}_i 's. Note that for any $i \neq j$, \mathbf{q}_i and \mathbf{q}_j do not have any terms in common, and similarly \mathbf{r}_i and \mathbf{r}_j

do not have any terms in common. Hence,

$$\begin{aligned}\mathbf{q} &= x_i x_j (\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_3 + \mathbf{q}_5) + x_i (\mathbf{q}_1 + \mathbf{q}_2 + \mathbf{q}_4 + \mathbf{q}_6) + x_j (\mathbf{q}_1 + \mathbf{q}_3 + \mathbf{q}_4 + \mathbf{q}_7) + \mathbf{a}_4, \\ \mathbf{r} &= x_i x_j (\mathbf{r}_1 + \mathbf{r}_2 + \mathbf{r}_3 + \mathbf{r}_5) + x_i (\mathbf{r}_1 + \mathbf{r}_2 + \mathbf{r}_4 + \mathbf{r}_6) + x_j (\mathbf{r}_1 + \mathbf{r}_3 + \mathbf{r}_4 + \mathbf{r}_7) + \mathbf{b}_4.\end{aligned}$$

Identification of x_i with x_j yields

$$\begin{aligned}\mathbf{q}_{j \leftarrow i} &= x_i (\mathbf{q}_1 + \mathbf{q}_5 + \mathbf{q}_6 + \mathbf{q}_7) + \mathbf{a}_4, \\ \mathbf{r}_{j \leftarrow i} &= x_i (\mathbf{r}_1 + \mathbf{r}_5 + \mathbf{r}_6 + \mathbf{r}_7) + \mathbf{b}_4.\end{aligned}$$

Since we are assuming that $\mathbf{q}_{j \leftarrow i} = 0$, we have that $\mathbf{q}_1 = \mathbf{q}_5 = \mathbf{q}_6 = \mathbf{q}_7 = \mathbf{a}_4 = 0$. On the other hand, $\mathbf{q} \neq 0$, so $\mathbf{q}_2, \mathbf{q}_3, \mathbf{q}_4$ are not all zero. Thus

$$\mathbf{q} = x_i x_j (\mathbf{q}_2 + \mathbf{q}_3) + x_i (\mathbf{q}_2 + \mathbf{q}_4) + x_j (\mathbf{q}_3 + \mathbf{q}_4).$$

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

Subcase 3.2.1. Assume that there is a variable x_t occurring in \mathbf{b}_4 that does not occur in $\mathbf{r}_1, \mathbf{r}_5, \mathbf{r}_6, \mathbf{r}_7$. Consider

$$\mathbf{p}_{m \leftarrow t} = x_t \mathbf{q} + \mathbf{r} = x_t \mathbf{q} + x_i x_j \mathbf{b}_1 + x_i \mathbf{b}_2 + x_j \mathbf{b}_3 + \mathbf{b}_4.$$

Cancelling may only happen between a term of $x_t \mathbf{q}$ and a term of \mathbf{r} . No term of \mathbf{b}_4 can be cancelled, because every term of $x_t \mathbf{q}$ contains x_i or x_j but the terms of \mathbf{b}_4 do not contain either. The variables that do not occur in \mathbf{b}_4 occur in some terms of $\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3$ that do not contain x_t . Thus, all essential variables of f except for x_m occur in $\mathbf{p}_{m \leftarrow t}$, and so in this case f has arity gap 1.

Subcase 3.2.2. Assume that all variables of \mathbf{r} except for x_i, x_j occur already in $\mathbf{r}_1 + \mathbf{r}_5 + \mathbf{r}_6 + \mathbf{r}_7$. Consider

$$\begin{aligned}\mathbf{p}_{m \leftarrow i} &= x_i x_j (\mathbf{q}_2 + \mathbf{q}_4 + \mathbf{r}_1 + \mathbf{r}_2 + \mathbf{r}_3 + \mathbf{r}_5) + \\ (1) \quad & x_i (\mathbf{q}_2 + \mathbf{q}_4 + \mathbf{r}_1 + \mathbf{r}_2 + \mathbf{r}_4 + \mathbf{r}_6) + \\ & x_j (\mathbf{r}_1 + \mathbf{r}_3 + \mathbf{r}_4 + \mathbf{r}_7) + \mathbf{b}_4.\end{aligned}$$

Subcase 3.2.2.1. Assume first that x_i does not occur in $\mathbf{p}_{m \leftarrow i}$ in (1). Then

$$\begin{aligned}\mathbf{q}_2 + \mathbf{q}_4 + \mathbf{r}_1 + \mathbf{r}_2 + \mathbf{r}_3 + \mathbf{r}_5 &= 0, \\ \mathbf{q}_2 + \mathbf{q}_4 + \mathbf{r}_1 + \mathbf{r}_2 + \mathbf{r}_4 + \mathbf{r}_6 &= 0,\end{aligned}$$

and since the \mathbf{r}_i 's do not have terms in common, we have that

$$\mathbf{r}_1 + \mathbf{r}_2 = \mathbf{q}_2 + \mathbf{q}_4, \quad \mathbf{r}_3 = \mathbf{r}_4 = \mathbf{r}_5 = \mathbf{r}_6 = 0.$$

Then all variables of \mathbf{r} except for x_i, x_j occur already in $\mathbf{r}_1 + \mathbf{r}_7$. Consider

$$\begin{aligned}\mathbf{p}_{m \leftarrow j} &= x_i x_j (\mathbf{q}_3 + \mathbf{q}_4 + \mathbf{r}_1 + \mathbf{r}_2 + \mathbf{r}_3 + \mathbf{r}_5) + \\ & x_i (\mathbf{r}_1 + \mathbf{r}_2 + \mathbf{r}_4 + \mathbf{r}_6) + \\ (2) \quad & x_j (\mathbf{q}_3 + \mathbf{q}_4 + \mathbf{r}_1 + \mathbf{r}_3 + \mathbf{r}_4 + \mathbf{r}_7) + \mathbf{b}_4 \\ &= x_i x_j (\mathbf{q}_2 + \mathbf{q}_3) + \\ & x_i (\mathbf{r}_1 + \mathbf{r}_2) + \\ & x_j (\mathbf{q}_2 + \mathbf{q}_3 + \mathbf{r}_2 + \mathbf{r}_7) + \mathbf{b}_4.\end{aligned}$$

All variables of \mathbf{r}_1 are there on the fifth line of (2). If a term of \mathbf{r}_7 is cancelled by a term of $\mathbf{q}_2 + \mathbf{q}_3$ on the sixth line, it still remains on the fourth line, so all variables of \mathbf{r}_7 are also there. We still need to verify that the variables x_i and x_j

are not cancelled out from (2). 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

$$\mathfrak{r}_1 + \mathfrak{r}_2 + \mathfrak{r}_4 + \mathfrak{r}_6 = \mathfrak{r}_1 + \mathfrak{r}_2 = q_2 + q_4 = q_4 \neq 0,$$

we have x_i in (2). Since

$$q_3 + q_4 + \mathfrak{r}_1 + \mathfrak{r}_3 + \mathfrak{r}_4 + \mathfrak{r}_7 = q_4 + \mathfrak{r}_1 + \mathfrak{r}_7$$

and $\mathfrak{r}_1 + \mathfrak{r}_7$ contain all variables of \mathfrak{r} except for x_i, x_j , but q_4 does not, $q_4 + \mathfrak{r}_1 + \mathfrak{r}_7 \neq 0$, so we also have x_j in (2). Thus, the arity gap of f equals 1 in this case.

Subcase 3.2.2.2. Assume then that x_i occurs in $\mathfrak{p}_{m \leftarrow i}$ in (1). Nothing cancels out on the third line of (1), and therefore the variables of \mathfrak{r}_1 and \mathfrak{r}_7 occur in $\mathfrak{p}_{m \leftarrow i}$. Terms of \mathfrak{r}_5 may be cancelled out by terms of $q_2 + q_4$ on the first line of (1) but such terms will remain on the second line. Thus the variables of \mathfrak{r}_5 occur in $\mathfrak{p}_{m \leftarrow i}$. A similar argument shows that the variables of \mathfrak{r}_6 also occur in $\mathfrak{p}_{m \leftarrow i}$. In order for f to have arity gap 1, we still need to verify that x_j occurs in $\mathfrak{p}_{m \leftarrow i}$. If $q_2 + q_4 + \mathfrak{r}_1 + \mathfrak{r}_2 + \mathfrak{r}_3 + \mathfrak{r}_5 \neq 0$, then we are done. We may thus assume that

$$(3) \quad q_2 + q_4 + \mathfrak{r}_1 + \mathfrak{r}_2 + \mathfrak{r}_3 + \mathfrak{r}_5 = 0.$$

By the assumption that x_i occurs in $\mathfrak{p}_{m \leftarrow i}$, the second line of (1) does not vanish, i.e.,

$$0 \neq q_2 + q_4 + \mathfrak{r}_1 + \mathfrak{r}_2 + \mathfrak{r}_4 + \mathfrak{r}_6 = \mathfrak{r}_3 + \mathfrak{r}_4 + \mathfrak{r}_5 + \mathfrak{r}_6.$$

If the third line of (1) does not vanish either, i.e., $\mathfrak{r}_1 + \mathfrak{r}_3 + \mathfrak{r}_4 + \mathfrak{r}_7 \neq 0$, then we have both x_i and x_j and we are done. We may thus assume that $\mathfrak{r}_1 + \mathfrak{r}_3 + \mathfrak{r}_4 + \mathfrak{r}_7 = 0$, i.e., $\mathfrak{r}_1 = \mathfrak{r}_3 = \mathfrak{r}_4 = \mathfrak{r}_7 = 0$. Then all variables of \mathfrak{r} except for x_i, x_j occur already in $\mathfrak{r}_5 + \mathfrak{r}_6$. Equation (3) implies that $\mathfrak{r}_2 + \mathfrak{r}_5 = q_2 + q_4$. Consider

$$\begin{aligned} \mathfrak{p}_{m \leftarrow j} &= x_i x_j (q_3 + q_4 + \mathfrak{r}_1 + \mathfrak{r}_2 + \mathfrak{r}_3 + \mathfrak{r}_5) + \\ &\quad x_i (\mathfrak{r}_1 + \mathfrak{r}_2 + \mathfrak{r}_4 + \mathfrak{r}_6) + \\ &\quad x_j (q_3 + q_4 + \mathfrak{r}_1 + \mathfrak{r}_3 + \mathfrak{r}_4 + \mathfrak{r}_7) + \mathfrak{b}_4 \\ (4) \quad &= x_i x_j (q_2 + q_3) + \\ &\quad x_i (q_2 + q_4 + \mathfrak{r}_5 + \mathfrak{r}_6) + \\ &\quad x_j (q_3 + q_4) + \mathfrak{b}_4. \end{aligned}$$

Assume first that $q_2 + q_3 = 0$, in which case $q_4 \neq 0$. If a term of $\mathfrak{r}_5 + \mathfrak{r}_6$ is cancelled by a term of q_4 on the second line of (4), it will still remain on the third line. Therefore we have in $\mathfrak{p}_{m \leftarrow j}$ all variables of \mathfrak{r} except for x_i and x_j . Since $\mathfrak{r}_5 + \mathfrak{r}_6$ contains all variables of \mathfrak{r} except for x_i, x_j but $q_2 + q_4 = q_4$ does not, the second line of (4) 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 third line. In this case f has arity gap 1.

Assume then that $q_2 + q_3 \neq 0$. Then the first line of (4) does not vanish and both x_i and x_j occur in $\mathfrak{p}_{m \leftarrow j}$. If any term of $\mathfrak{r}_5 + \mathfrak{r}_6$ is cancelled by a term of q_2 on the second line of (4), it still remains on the first line, and if it is cancelled by a term of q_4 , it remains on the third line. Thus all variables of \mathfrak{r} occur in $\mathfrak{p}_{m \leftarrow j}$, and f has arity gap 1 again. This completes the proof of Theorem 6. \square

5. CONCLUDING REMARKS

We do not know whether the upper bound on arity gap given by Theorem 3 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

$$f = g(h_1(x_1) \oplus h_2(x_2) \oplus \cdots \oplus h_n(x_n)),$$

where $h_1, \dots, h_n : A \rightarrow \{0, 1\}$, $g : \{0, 1\} \rightarrow A$ 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

$$g = \phi(f(\gamma(x_1), \gamma(x_2), \dots, \gamma(x_n))),$$

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

REFERENCES

- [1] G. A. BURLE, The classes of k -valued logics containing all one-variable functions, *Diskretnyi Analiz* **10** (1967) 3–7 (in Russian).
- [2] K. N. ČIMEV, *Separable Sets of Arguments of Functions*, Studies 180/1986, Computer and Automation Institute, Hungarian Academy of Sciences, Budapest, 1986.
- [3] M. COUCEIRO, On the lattice of equational classes of Boolean functions and its closed intervals, Technical report A367, University of Tampere, 2006.
- [4] M. COUCEIRO, M. POUZET, On a quasi-ordering on Boolean functions, arXiv:math.CO/0601218, 2006.
- [5] R. O. DAVIES, Two theorems on essential variables, *J. London Math. Soc.* **41** (1966) 333–335.
- [6] A. EHRENFEUCHT, J. KAHN, R. MADDUX, J. MYCIELSKI, On the dependence of functions on their variables, *J. Combin. Theory Ser. A* **33** (1982) 106–108.
- [7] O. EKIN, S. FOLDES, P. L. HAMMER, L. HELLERSTEIN, Equational characterizations of Boolean function classes, *Discrete Math.* **211** (2000) 27–51.
- [8] A. FEIGELSON, L. HELLERSTEIN, The forbidden projections of unate functions, *Discrete Appl. Math.* **77** (1997) 221–236.
- [9] E. LEHTONEN, Descending chains and antichains of the unary, linear, and monotone subfunction relations, *Order* **23** (2006) 129–142.
- [10] N. PIPPENGER, Galois theory for minors of finite functions, *Discrete Math.* **254** (2002) 405–419.
- [11] A. SALOMAA, On essential variables of functions, especially in the algebra of logic, *Ann. Acad. Sci. Fenn. Ser. A I. Math.* **339** (1963) 3–11.
- [12] C. WANG, Boolean minors, *Discrete Math.* **141** (1991) 237–258.
- [13] S. V. YABLONSKI, Functional constructions in a k -valued logic, *Tr. Mat. Inst. Steklova* **51** (1958) 5–142 (in Russian).
- [14] I. E. ZVEROVICH, Characterizations of closed classes of Boolean functions in terms of forbidden subfunctions and Post classes, *Discrete Appl. Math.* **149** (2005) 200–218.

DEPARTMENT OF MATHEMATICS, STATISTICS AND PHILOSOPHY, UNIVERSITY OF TAMPERE, FI-33014 TAMPEREEN YLIOPISTO, FINLAND

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

INSTITUTE OF MATHEMATICS, TAMPERE UNIVERSITY OF TECHNOLOGY, P.O. Box 553, FI-33101 TAMPERE, FINLAND

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