Discrete Mathematics and Theoretical Computer Science DMTCS vol. **19**:1, 2017, #16

On universal partial words[*]

Herman Z. Q. Chen[1] Sergey Kitaev[2] Torsten Mütze[3] Brian Y. Sun[4†]

[1] *School of Science, Tianjin Chengjian University, P.R. China*
[2] *Department of Computer and Information Sciences, University of Strathclyde, Glasgow, UK*
[3] *Institut für Mathematik, TU Berlin, Germany*
[4] *College of Mathematics and System Science, Xinjiang University, Urumqi, Xinjiang 830046, P. R. China*

received 10th Nov. 2016, revised 8th May 2017, accepted 16th May 2017.

A *universal word* for a finite alphabet A and some integer $n \geq 1$ is a word over A such that every word in A^n appears exactly once as a subword (cyclically or linearly). It is well-known and easy to prove that universal words exist for any A and n. In this work we initiate the systematic study of universal *partial* words. These are words that in addition to the letters from A may contain an arbitrary number of occurrences of a special 'joker' symbol $\diamond \notin A$, which can be substituted by any symbol from A. For example, $u = 0\diamond 011100$ is a linear partial word for the binary alphabet $A = \{0, 1\}$ and for $n = 3$ (e.g., the first three letters of u yield the subwords 000 and 010). We present results on the existence and non-existence of linear and cyclic universal partial words in different situations (depending on the number of $\diamond$s and their positions), including various explicit constructions. We also provide numerous examples of universal partial words that we found with the help of a computer.

Keywords: universal word, partial word, De Bruijn graph, Eulerian cycle, Hamiltonian cycle

1 Introduction

De Bruijn sequences are a centuries-old and well-studied topic in combinatorics, and over the years they found widespread use in real-world applications, e.g., in the areas of molecular biology [10, 23], computer security [25], computer vision [24], robotics [28] and psychology experiments [27]. More recently, they have also been studied in a more general context by constructing *universal cycles* for other fundamental combinatorial structures such as permutations or subsets of a fixed ground set (see e.g. [4, 20, 8, 19, 29]).

In the context of words over a finite alphabet A, we say that a word u is *universal for* A^n if u contains every word of length $n \geq 1$ over A exactly once as a subword. We distinguish *cyclic universal words* and *linear universal words*. In the cyclic setting, we view u as a cyclic word and consider all subwords of length n cyclically across the boundaries of u. In the linear setting, on the other hand, we view u as a linear word and only consider subwords that start and end within the index range of letters of u. From

[*]The last author was supported by the Scientific Research Program of the Higher Education Institution of Xinjiang Uygur Autonomous Region (No. XJEDU2016S032) and the Scientific Research Initiative Foundation of Xinjiang University for Graduated Ph.D Students (No. BS160206).

[†]Corresponding author. Email: brianys1984@126.com.

ISSN 1365–8050

this definition it follows that the length of a cyclic or linear universal word must be $|A|^n$ or $|A|^n + n - 1$, respectively. For example, for the binary alphabet $A = \{0, 1\}$ and for $n = 3$, $u = 0001011100$ is a linear universal word for A^3. Observe that a cyclic universal word for A^n can be easily transformed into a linear universal word for A^n, so existence results in the cyclic setting imply existence results for the linear setting. Note also that reversing a universal word, or permuting the letters of the alphabet yields a universal word again. The following classical result is the starting point for our work (see [11, 31, 22]).

Theorem 1 *For any finite alphabet A and any $n \geq 1$, there exists a cyclic universal word for A^n.*

The standard proof of Theorem 1 is really beautiful and concise, using the De Bruijn graph, its line graph and Eulerian cycles (see [8] and Section 2 below).

1.1 Universal partial words

In this paper we consider the universality of *partial words*, which are words that in addition to letters from A may contain any number of occurrences of an additional special symbol $\diamond \notin A$. The idea is that every occurrence of $\diamond$ can be substituted by any symbol from A, so we can think of $\diamond$ as a 'joker' or 'wildcard' symbol. Formally, we define $A_\diamond := A \cup \{\diamond\}$ and we say that a word $v = v_1 v_2 \cdots v_n \in A^n$ appears as a *factor* in a word $u = u_1 u_2 \cdots u_m \in A_\diamond^m$ if there is an integer i such that $u_{i+j} = \diamond$ or $u_{i+j} = v_j$ for all $j = 1, 2, \ldots, n$. In the cyclic setting we consider the indices of u in this definition modulo m. For example, in the linear setting and for the ternary alphabet $A = \{0, 1, 2\}$, the word $v = 021$ occurs twice as a factor in $u = 120\diamond120021$ because of the subwords $0\diamond1$ and 021 of u, whereas v does not appear as a factor in $u' = 12\diamond11\diamond$.

Partial words were introduced in [1], and they too have real-world applications (see [6] and references therein). In combinatorics, partial words appear in the context of primitive words [5], of (un)avoidability of sets of partial words [7, 2], and also in the study of the number of squares [17] and overlap-freeness [18] in (infinite) partial words. The concept of partial words has been extended to pattern-avoiding permutations in [9].

The notion of universality given above extends straightforwardly to partial words, and we refer to a universal partial word as an *upword* for short. Again we distinguish cyclic upwords and linear upwords. The simplest example for a linear upword for A^n is $\diamond^n := \diamond\diamond\cdots\diamond$, the word consisting of n many $\diamond$s, which we call *trivial*. Let us consider a few more interesting examples of linear upwords over the binary alphabet $A = \{0, 1\}$. We have that $\diamond\diamond0111$ is a linear upword for A^3, whereas $\diamond\diamond01110$ is *not* a linear upword for A^3, because replacing the first two letters $\diamond\diamond$ by 11 yields the same factor 110 as the last three letters. Similarly, $0\diamond1$ is *not* a linear upword for A^2 because the word $10 \in A^2$ does not appear as a factor (and the word $01 \in A^2$ appears twice as a factor).

1.2 Our results

In this work we initiate the systematic study of universal partial words. It turns out that these words are rather shy animals, unlike their ordinary counterparts (universal words without 'joker' symbols). That is, in stark contrast to Theorem 1, there are no general existence results on upwords, but also many non-existence results. The borderline between these two cases seems rather complicated, which makes the subject even more interesting (this is true also for non-binary alphabets, as the constructions of the follow-up paper [16] indicate). In addition to the size of the alphabet A and the length n of the factors, we also consider the number of $\diamond$s and their positions in an upword as problem parameters.

n	k	
1	1	$\diamond$
2	1	$\diamond$011 (Thm. 9, Thm. 17)
	2	— (Thm. 6)
3	1	$\diamond$00111010 (Thm. 9)
	2	0$\diamond$011100 (Thm. 10)
	3	— (Thm. 6)
	4	— (Thm. 7)
4	1	$\diamond$00011110100101100 (Thm. 9)
	2	0$\diamond$010011011110000 (Thm. 10)
	3	01$\diamond$0111100001010 (Thm. 10)
	4	— (Thm. 6)
	5	— (Thm. 7)
	6	01100$\diamond$011110100
	7	— (Thm. 7)
	8	0011110$\diamond$0010110
5	1	$\diamond$000011111011100110001011010100100 0 (Thm. 9)
	2	0$\diamond$010110000011010011101111110010001 (Thm. 10)
	3	01$\diamond$011000001000111001010111110100 (Thm. 10)
	4	011$\diamond$011111000001010010001101011 0 (Thm. 10)
	5	— (Thm. 6)
	6	00101$\diamond$0010011101111100000110101
	7	010011$\diamond$0100000101011011111100011
	8	0100110$\diamond$0100000111001011111 0110
	9	01110010$\diamond$0111110110100110000010
	10	010011011$\diamond$0100011111000000101011
	11	0101000001$\diamond$0101111100111011 0001
	12	01010000011$\diamond$0101101111100010011
	13	001001101011$\diamond$001010000011111011
	14	0011101111100$\diamond$00110100010101100
	15	01010000010011$\diamond$0101101111100011
	16	001000001101011$\diamond$001010011111011

Tab. 1: Examples of linear upwords for A^n, $A = \{0, 1\}$, with a single $\diamond$ at position k from the beginning or end for $n = 1, 2, 3, 4, 5$ and all possible values of k (upwords where the $\diamond$ is closer to the end of the word than to the beginning can be obtained by reversal). A dash indicates that no such upword exists.

We first focus on linear upwords. For linear upwords containing a *single* $\diamond$, we have the following results: For non-binary alphabets A (i.e., $|A| \geq 3$) and $n \geq 2$, there is *no* linear upword for A^n with a single $\diamond$ at all (Theorem 5 below). For the binary alphabet $A = \{0,1\}$, the situation is more interesting (see Table 1): Denoting by k the position of the $\diamond$, we have that for $n \geq 2$, there is *no* linear upword for A^n with $k = n$ (Theorem 6), and there are *no* linear upwords in the following three cases: $n = 3$ and $k = 4$, and $n = 4$ and $k \in \{5,7\}$ (Theorem 7). We conjecture that these are the only non-existence cases for a binary alphabet (Conjecture 8). To support this conjecture, we performed a computer-assisted search and indeed found linear upwords for all values of $2 \leq n \leq 13$ and all possible values of k other than the ones excluded by the beforementioned results. Some of these examples are listed in Table 1, and the remaining ones are available on the third author's website [30]. We also prove the special cases $k = 1$ and $k \in \{2,3,\ldots,n-1\}$ of our conjecture (Theorems 9 and 10, respectively).

For linear upwords containing *two* $\diamond$s we have the following results: First of all, Table 2 shows examples of linear upwords with two $\diamond$s for the binary alphabet $A = \{0,1\}$ for $n = 2,3,4,5$. We establish a sufficient condition for non-existence of binary linear upwords with two $\diamond$s (Theorem 11), which in particular shows that a $(1-o(1))$-fraction of all ways of placing two $\diamond$s among the $N = \Theta(2^n)$ positions does not yield a valid upword. Moreover, we conclude that there are only two binary linear upwords where the two $\diamond$s are adjacent (Corollary 12), namely $\diamond\diamond$ for $n = 2$ and $\diamond\diamond 0111$ for $n = 3$ (see Table 2). We also construct an infinite family of binary linear upwords with two $\diamond$s (Theorem 13). Let us now discuss cyclic upwords. Note that the trivial solution $\diamond^n$ is a cyclic upword only for $n = 1$. For the cyclic setting we have the following rather general non-existence result: If $\gcd(|A|,n) = 1$, then there is no cyclic upword for A^n (Corollary 16). In particular, for a binary alphabet $|A| = 2$ and odd n, there is no cyclic upword for A^n. In fact, we know only of a single cyclic upword for the binary alphabet $A = \{0,1\}$ and any $n \geq 2$, namely $\diamond 001 \diamond 110$ for $n = 4$ (up to cyclic shifts, reversal and letter permutations).

1.3 Outline of this paper

This paper is organized as follows. In Section 2 we introduce some notation and collect basic observations that are used throughout the rest of the paper. In Sections 3 and 4 we prove our results on linear upwords containing a single or two $\diamond$s, respectively. Section 5 contains the proofs on cyclic upwords. Finally, Section 6 discusses possible directions for further research, including some extensions of our results to non-binary alphabets.

2 Preliminaries

For the rest of this paper, we assume w.l.o.g. that the alphabet is $A = \{0,1,\ldots,\alpha-1\}$, so $\alpha \geq 2$ denotes the size of the alphabet. We often consider the special case $\alpha = 2$ of the binary alphabet, and then for $x \in \{0,1\}$ we write $\overline{x}$ for the complement of x. Moreover, for any word u, we let $|u|$ denote its length. As we mentioned before, reversing a universal word and/or permuting the letters of the alphabet again yields a universal word. We can thus assume w.l.o.g. that in an upword u the letters of A appear in increasing order from left to right, i.e., the first occurence of symbol i is before the first occurence of symbol j whenever $i < j$. Moreover, if u can be factored as $u = xyz$, where x and z do not contain any $\diamond$s, then we can assume that $|x| \leq |z|$.

One standard approach to prove the existence of universal words is to define a suitable graph and to search for a Hamiltonian path/cycle in this graph (another more algebraic approach uses irreducible polynomials). Specifically, the *De Bruijn graph* G_A^n has as vertices all elements from A^n (all words of

n	
2	$\diamond\diamond$ (Cor. 12)
3	$\diamond\diamond$0111 (Cor. 12, Thm. 17)
	$\diamond$001011$\diamond$
4	$\diamond$00011$\diamond$1001011 (Thm. 13)
	$\diamond$0001011$\diamond$10011
	001$\diamond$110$\diamond$001
5	$\diamond$0100$\diamond$10101100000111011110010
	$\diamond$0000111$\diamond$100010010101100110111 (Thm. 13)
	$\diamond$00001001$\diamond$10001101011111011001
	$\diamond$0000100111$\diamond$100011001010110111
	$\diamond$00001010111$\diamond$10001101100100111
	0$\diamond$0011$\diamond$01000101011011111100000
	0$\diamond$010110$\diamond$00011101111100100110
	0$\diamond$0101110$\diamond$00011011001001111110
	0$\diamond$010111110$\diamond$00011011001001110
	0$\diamond$0101101110$\diamond$0001100100111110
	00$\diamond$0011$\diamond$00101011011111010000
	01$\diamond$01100101110$\diamond$0100000111110
	01$\diamond$0110010111110$\diamond$01000001110
	01$\diamond$010000010101100011111 0$\diamond$110
	001$\diamond$0101$\diamond$001110111110000010
	011$\diamond$011010010$\diamond$0111110000010
	011$\diamond$0110101001000$\diamond$011111000
	011$\diamond$01111100011010100000 10$\diamond$10
	011$\diamond$011010000011111000100101$\diamond$1
	01001$\diamond$1110$\diamond$010000011011001

Tab. 2: Examples of linear upwords for A^n, $A = \{0, 1\}$, with two $\diamond$s for $n = 2, 3, 4, 5$.

length n over A), and a directed edge from a vertex u to a vertex v whenever the last $n-1$ letters of u are the same as the first $n-1$ letters of v. We call such an edge (u,v) an *x-edge*, if the last letter of v equals x. Figure 1 (a) and (b) shows the graph G_A^n, $A=\{0,1\}$, for $n=2$ and $n=3$, respectively. Clearly, a linear universal word for A^n corresponds to a Hamiltonian path in G_A^n, and a cyclic universal word to a Hamiltonian cycle in this graph. Observe furthermore that G_A^n is the line graph of G_A^{n-1}. Recall that the *line graph* $L(G)$ of a directed graph G is the directed graph that has a vertex for every edge of G, and a directed edge from e to e' if in G the end vertex of e equals the starting vertex of e'. Therefore, the problem of finding a Hamiltonian path/cycle in G_A^n is equivalent to finding an Eulerian path/cycle in G_A^{n-1}. The existence of an Eulerian path/cycle follows from the fact that the De Bruijn graph is connected and that each vertex has in- and out-degree α (this is one of Euler's famous theorems [12], see also [3, Theorem 1.6.3]). This proves Theorem 1. In fact, this existence proof can be easily turned into an algorithm to actually find (many) universal words (using Hierholzer's algorithm [21] or Fleury's algorithm [14]).

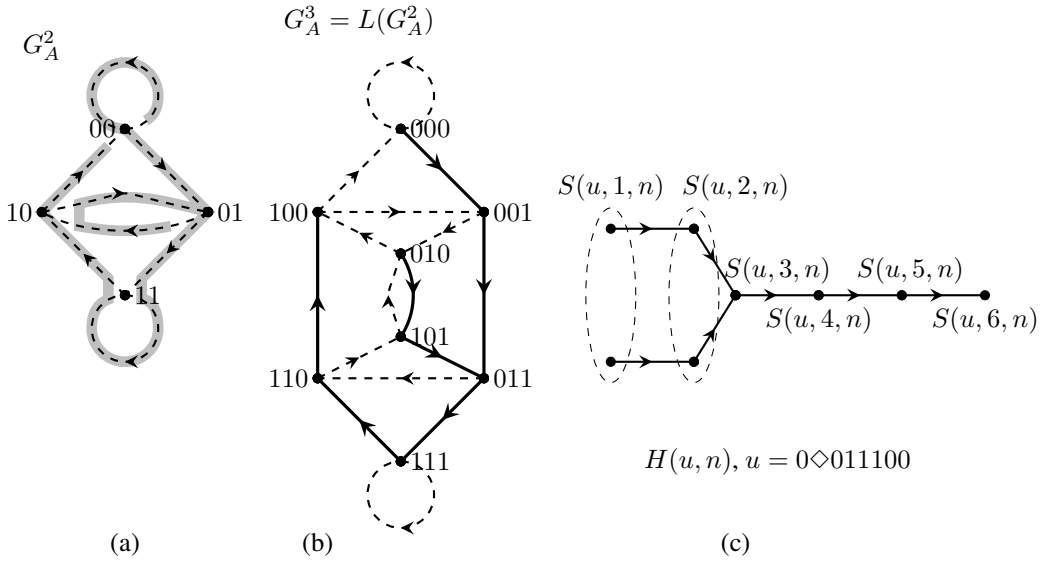

Fig. 1: The De Bruijn graphs G_A^2 (a) and $G_A^3 = L(G_A^2)$ (b) for $A=\{0,1\}$ with a spanning subgraph $H(u,n)$ of G_A^3 for the linear upword $u = 0\diamond 011100$ for A^3 ($H(u,n)$ is shown by solid edges). Part (c) if the figure shows a schematic drawing of the graph $H(u,n)$. $H(u,n)$ is the line graph of the highlighted sequences of edges in G_A^2.

We now discuss how this standard approach of proving the existence of universal words can be extended to universal partial words. Specifically, we collect a few simple but powerful observations that will be used in our proofs later on.

For any vertex v of G_A^n, we let $\Gamma^+(v)$ and $\Gamma^-(v)$ denote the sets of out-neighbours and in-neighbours of v, respectively (both are sets of vertices of G_A^n). As we mentioned before, we clearly have $|\Gamma^+(v)| = |\Gamma^-(v)| = \alpha$.

Observation 2 *For any vertex $v = v_1v_2\cdots v_n$ of G_A^n and its set of out-neighbours $\Gamma^+(v)$, there are $\alpha - 1$ vertices different from v with the same set of out-neighbours $\Gamma^+(v)$, given by $xv_2v_3\cdots v_n$, where $x \in A \setminus \{v_1\}$. For any vertex $v = v_1v_2\cdots v_n$ of G_A^n and its set of in-neighbours $\Gamma^-(v)$, there are $\alpha - 1$ vertices different from v with the same set of in-neighbours $\Gamma^-(v)$, given by $v_1v_2\cdots v_{n-1}x$, where $x \in A \setminus \{v_n\}$.*

For any linear upword u for A^n, we define a spanning subgraph $H(u,n)$ of the De Bruijn graph G_A^n as follows, see Figure 1 (c): For any $i = 1, 2, \ldots, N-n+1$, we let $S(u,i,n)$ denote the set of all words that are obtained from the subword of u of length n starting at position i by replacing any occurences of $\diamond$ by a letter from the alphabet A. Clearly, if there are d many $\diamond$s in this subword, then there are α^d different possibilities for substitution, so we have $|S(u,i,n)| = \alpha^d$. Note that the sets $S(u,i,n)$ form a partition of the vertex set of G_A^n (and $H(u,n)$). The directed edges of $H(u,n)$ are given by all the edges of G_A^n induced between every pair of consecutive sets $S(u,i,n)$ and $S(u,i+1,n)$ for $i = 1, 2, \ldots, N-n$. For example, for the linear upword $u = 0\diamond 011100$ over the binary alphabet $A = \{0,1\}$ for $n = 3$ we have $S(u,1,n) = \{000, 010\}$, $S(u,2,n) = \{001, 101\}$, $S(u,3,n) = \{011\}$, $S(u,4,n) = \{111\}$, $S(u,5,n) = \{110\}$ and $S(u,6,n) = \{100\}$, and the spanning subgraph $H(u,n)$ of G_A^3 is shown in Figure 1 (c). To give another example with the same A and n, for the linear upword $u = \diamond\diamond 0111$ we have $S(u,1,n) = \{000, 010, 100, 110\}$, $S(u,2,n) = \{001, 101\}$, $S(u,3,n) = \{011\}$, $S(u,4,n) = \{111\}$, and then $H(u,n)$ is a binary tree of depth 2 with an additional edge emanating from the root.

The following observation follows straightforwardly from these definitions.

Observation 3 *Let $u = u_1u_2\cdots u_N$ be a linear upword for A^n. A vertex in $S(u,i,n)$, $i = 1, 2, \ldots, N-n$, has out-degree 1 in $H(u,n)$ if $u_{i+n} \in A$, and out-degree α if $u_{i+n} = \diamond$. A vertex in $S(u,i,n)$, $i = 2, 3, \ldots, N-n+1$, has in-degree 1 in $H(u,n)$ if $u_{i-1} \in A$, and in-degree α if $u_{i-1} = \diamond$. The vertices in $S(u,1,n)$ have in-degree 0, and the vertices in $S(u,N-n+1,n)$ have out-degree 0.*

By this last observation, the graph $H(u,n)$ is determined only by the positions of the $\diamond$s in u. Intuitively, the $\diamond$s lead to branching in the graph $H(u,n)$ due to the different possibilities of substituting symbols from A. In particular, if u has no $\diamond$s, then $H(u,n)$ is just a spanning path of G_A^n (i.e., a Hamiltonian path, so we are back in the setting of Theorem 1). So when searching for a linear universal partial word u with a particular number of $\diamond$s at certain positions, we essentially search for a copy of the spanning subgraph $H(u,n)$ in G_A^n. We will exploit this idea both in our existence and non-existence proofs. For the constructions it is particularly useful (and for our computer-searches it is computationally much more efficient) to not search for a copy of $H(u,n)$ in G_A^n directly, but to rather search for the corresponding sequences of edges in G_A^{n-1}, which can be seen as generalizations of Eulerian paths that were used before in the proof of Theorem 1 (see Figure 1 (a)). For example, to search for a linear upword u with a single $\diamond$ at position $k \in \{1, 2, \ldots, n-1\}$, we can prescribe the first $k-1$ letters and the n letters after the $\diamond$ (with a particular choice of symbols from A, or by iterating over all possible choices), and search for an Eulerian path in the subgraph of G_A^{n-1} that remains when deleting from it all edges that correspond to the prescribed prefix of u (see the proofs of Theorems 9 and 10 below). This idea can be generalized straightforwardly to search for upwords with other $\diamond$ patterns (see for example the proof of Theorem 13 below).

The next lemma will be used repeatedly in our proofs (both for existence and non-existence of upwords). The proof uses the previous two graph-theoretical observations to derive dependencies between letters of an upword.

Lemma 4 *Let $u = u_1u_2\cdots u_N$ be a linear upword for A^n, $A = \{0,1,\ldots,\alpha-1\}$, $n \geq 2$, such that $u_k = \diamond$ and $u_{k+n} \neq \diamond$ (we require $k+n \leq N$). Then for all $i = 1,2,\ldots,n-1$ we have that if $u_i \neq \diamond$, then $u_{k+i} = u_i$. Moreover, we have that if $u_n \neq \diamond$, then $\alpha = 2$ and $u_{k+n} = \overline{u_n}$.*

Proof: By Observation 3, each vertex in the set $S(u,k+1,n)$ has in-degree α in $H(u,n)$, and each vertex in $S(u,k,n)$ has out-degree 1. By Observation 2, for each $v = v_1v_2\cdots v_n \in S(u,k+1,n)$ there are $\alpha-1$ other vertices (different from the ones in $S(u,k+1,n)$) in G_A^n with the same set $\Gamma^-(v)$ of α many in-neighbors, namely $v_x := v_1\cdots v_{n-1}x$, where $x \in A \setminus \{v_n\}$ (see Figure 2). As the in-degree of every vertex of G_A^n is exactly α, and in $H(u,n)$ all vertices except the ones in $S(u,1,n)$ already have in-degree at least 1, it follows that each of the vertices v_x must be equal to one of the vertices in $S(u,1,n)$. It follows that if $u_i \neq \diamond$ then $u_{k+i} \neq \diamond$ and $u_i = v_i = u_{k+i}$ for all $i = 1,2,\ldots,n-1$. Moreover, if $u_n \neq \diamond$ and $\alpha \geq 3$, then there are at least two vertices v_x, $x \in A \setminus \{v_n\}$, ending with different symbols x, each of which must be equal to one of the vertices in $S(u,1,n)$, which is impossible because all words in this set end with the same symbol u_n. It follows that if $u_n \neq \diamond$ then we must have $\alpha = 2$ and $u_n = x \neq v_n = u_{k+n}$, so $u_{k+n} = \overline{u_n}$. □

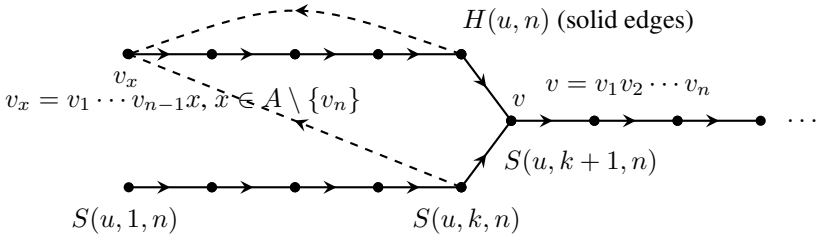

Fig. 2: Illustration of the proof of Lemma 4.

3 Linear upwords with a single $\diamond$

3.1 Non-existence results

Our first result completely excludes the existence of linear upwords with a single $\diamond$ for non-binary alphabets.

Theorem 5 *For $A = \{0,1,\ldots,\alpha-1\}$, $\alpha \geq 3$, and any $n \geq 2$, there is no linear upword for A^n with a single $\diamond$.*

Proof: Suppose that such an upword $u = u_1u_2\cdots u_{k-1}\diamond u_{k+1}\cdots u_N$ exists. We claim that the $\diamond$ in u is preceded or followed by at least n symbols from A. If not, then u would have at most αn different factors, which is strictly less than α^n for $\alpha \geq 3$ and $n \geq 2$. So we assume w.l.o.g. that the $\diamond$ in u is followed by at least n symbols from A, i.e., $k+n \leq N$. By Lemma 4 we have $u_i = \diamond$ or $u_{k+i} = u_i$ for all $i = 1,2,\ldots,n-1$ and $u_n = \diamond$, which implies $k = n$ and therefore $u_{n+i} = u_i$ for all $i = 1,\ldots,n-1$. But this means that the word $v := u_{n+1}\cdots u_{2n} \in A^n$ appears twice as a factor in u starting at positions

1 and $n+1$ (in other words, the vertex $v \in S(u, n+1, n)$ is identical to a vertex from $S(u, 1, n)$ in $H(u, n)$), a contradiction. □

Our next result excludes several cases with a single $\diamond$ for a binary alphabet.

Theorem 6 *For $A = \{0, 1\}$ and any $n \geq 2$, there is no linear upword for A^n with a single $\diamond$ at position n from the beginning or end.*

Proof: We first consider the case $n = 2$. Suppose that there is an upword $u = u_1 \diamond u_3$ for A^n. Assuming w.l.o.g. that $u_1 = 0$, we must have $u_3 = 1$, otherwise the word 00 would appear twice as a factor. But then the word 10 does not appear as a factor in $u = 0 \diamond 1$, while 01 appears twice, a contradiction.

For the rest of the proof we assume that $n \geq 3$. Suppose there was an upword $u = u_1 u_2 \cdots u_{n-1} \diamond u_{n+1} \cdots u_N$ with $N = 2^n - 1$. Note that $N - n \geq n$, or equivalently $2^n \geq 2n+1$, holds by our assumption $n \geq 3$, so the $\diamond$ in u is followed by at least n more symbols from A. Applying Lemma 4 yields that $u_{n+i} = u_i$ for all $i = 1, \ldots, n-1$, which means that the word $v := u_{n+1} \cdots u_{2n} \in A^n$ appears twice as a factor in u starting at positions 1 and $n+1$, a contradiction. □

In contrast to Theorem 5, for a binary alphabet we can only exclude the following three more (small) cases in addition to the cases excluded by Theorem 6 (all the exceptions are marked in Table 1).

Theorem 7 *For $A = \{0, 1\}$, there is no linear upword for A^n with a single $\diamond$ at position k from the beginning or end in the following three cases: $n = 3$ and $k = 4$, and $n = 4$ and $k \in \{5, 7\}$.*

Proof: Suppose that there is an upword $u = u_1 u_2 u_3 \diamond u_5 u_6 u_7$ for the case $n = 3$. Applying Lemma 4 twice to u and its reverse we obtain that $u_5 u_6 u_7 = u_1 u_2 \overline{u_3}$ and $u_1 u_2 u_3 = \overline{u_5} u_6 u_7$, a contradiction.

To prove the second case suppose that there is an upword of the form $u = u_1 u_2 u_3 u_4 \diamond u_6 \cdots u_{15}$ for $n = 4$. Applying Lemma 4 twice to u and its reverse we obtain that u has the form $u = u_1 u_2 u_3 u_4 \diamond u_1 u_2 u_3 \overline{u_4} u_{10} u_{11} \overline{u_1} u_2 u_3 u_4$. We assume w.l.o.g. that $u_1 = 0$. The word $z := 0000$ must appear somewhere as a factor in u, and since $u_{12} = \overline{u_1} = 1$, the only possible starting positions for z in u are $1, 2, \ldots, 8$. However, the starting positions $1, 2, 5, 6, 7$ can be excluded immediately, as they would cause z to appear twice as a factor in u. On the other hand, if z starts at positions 3, 4 or 8, then the neighboring letters must both be 1, causing 0101, 1010 or 1101, respectively, to appear twice as a factor in u, a contradiction.

The proof of the third case proceeds very similarly to the second case, and allows us to conclude that such an upword u must have the form $u = u_1 u_2 u_3 u_4 u_5 u_6 \diamond u_1 u_2 u_3 \overline{u_4 u_3} u_4 u_5 u_6$. We assume w.l.o.g. that $u_3 = 0$. The word $z := 0000$ must appear somewhere as a factor in u, and since $u_{12} = \overline{u_3} = 1$ the only possible starting positions for z in u are $1, 2, \ldots, 8$. The starting positions $1, 3, 4, 6, 8$ can be excluded immediately, as they would cause z to appear twice as a factor in u. On the other hand, if z starts at positions 2, 5 or 7, then the neighboring letters must both be 1, causing 0011, 0101 or 0000, respectively, to appear twice as a factor in u, a contradiction.

□

3.2 Existence results

We conjecture that for a binary alphabet and a single $\diamond$, the non-existence cases discussed in the previous section are the only ones.

Conjecture 8 *For $A = \{0,1\}$ and any $n \geq 1$, there is a linear upword for A^n containing a single $\diamond$ at position k in every case not covered by Theorem 6 or Theorem 7.*

Recall the numerical evidence for the conjecture discussed in the introduction. In the remainder of this section we prove some cases of this general conjecture.

Theorem 9 *For $A = \{0,1\}$ and any $n \geq 2$, there is a linear upword for A^n with a single $\diamond$ at the first position that begins with $\diamond 0^{n-1}1$.*

Note that by Lemma 4, *every* linear upword for A^n with a single $\diamond$ of the form $u = \diamond u_2 u_3 \cdots u_N$ satisfies the conditions $u_2 = u_3 = \cdots = u_n = \overline{u_{n+1}}$, i.e., w.l.o.g. it begins with $\diamond 0^{n-1}1$ (up to letter permutations).

Proof: Consider the word $v = v_1 v_2 \cdots v_{n+1} := \diamond 0^{n-1}1$ and the corresponding three edges $(0^{n-1}, 0^{n-1})$, $(10^{n-2}, 0^{n-1})$ and $(0^{n-1}, 0^{n-2}1)$ in the De Bruijn graph G_A^{n-1}. Denote the graph obtained from G_A^{n-1} by removing these three edges and the isolated vertex 0^{n-1} by G'. Clearly, the edges of G' form a connected graph, and every vertex in G' has in- and out-degree exactly two, except the vertex $y := 0^{n-2}1$ which has one more out-edge than in-edges and the vertex $z := 10^{n-2}$ which has one more in-edge than out-edges. Therefore, G' has an Eulerian path starting at y and ending at z, and this Eulerian path yields the desired upword that begins with v. □

For any binary word $w \in A^k$, $A = \{0,1\}$, and any $n \geq 1$, we write $c(w,n) = c_1 c_2 \cdots c_n$ for the word given by $c_i = w_i$ for $i = 1,2,\ldots,k$, $c_i = c_{i-k}$ for all $i = k+1, k+2, \ldots, n-1$ and $c_n = \overline{c_{n-k}}$. Informally speaking, $c(w,n)$ is obtained by concatenating infinitely many copies of w, truncating the resulting word at length n and complementing the last symbol. For example, we have $c(011,7) = 0110111$ and $c(011,8) = 01101100$. Using this terminology, the starting segment of the linear upword from Theorem 9 can be written as $\diamond c(0,n)$. The next result is a considerable extension of the previous theorem.

Theorem 10 *For $A = \{0,1\}$, any $n \geq 3$ and any $k \in \{2,3,\ldots,n-1\}$, there is a linear upword for A^n with a single $\diamond$ at the k-th position that begins with $01^{k-2}\diamond c(01^{k-1}, n)$.*

The idea of the proof of Theorem 10 is a straightforward generalization of the approach we used to prove Theorem 9 before, and boils down to showing that the De Bruijn graph G_A^{n-1} without the edges that are given by the prescribed upword prefix still has an Eulerian path.

Proof: The words $0\diamond c(01,3)100 = 0\diamond 011100$, $0\diamond c(01,4)11011110000 = 0\diamond 010011011110000$ and $01\diamond c(011,4)100001010 = 01\diamond 0111100001010$ from Table 1 show that the statement is true for $n = 3$ and $n = 4$. For the rest of the proof we assume that $n \geq 5$. Consider the word $w = w_1 w_2 \cdots w_{k+n} := 01^{k-2}\diamond c(01^{k-1}, n)$. For $i = 1,2,\ldots,k$ we let v_i^0 and v_i^1 denote the two words from $S(w,i,n-1)$ obtained by substituting $\diamond$ in w by 0 or 1, respectively. Moreover, let $v_{k+1} = w_{k+1} \cdots w_{k+n-1}$ be the unique word from $S(w,k+1,n-1)$ and $v_{k+2} = w_{k+2} \cdots w_{k+n}$ the unique word from $S(w,k+2,n-1)$, and define $V^0 := \{v_i^0 \mid i = 1,2,\ldots,k\}$, $V^1 := \{v_i^1 \mid i = 1,2,\ldots,k\}$, $V^2 := \{v_{k+1}, v_{k+2}\}$ and $V' := V^0 \cup V^1 \cup V^2$. We proceed to show that $|V'| = 2k+1$, i.e., only two of the words just defined coincide, namely $v_1^1 = v_{k+1}$ (v_1^1 is given by the first $n-1$ letters of $w = 01^{k-1}c(01^{k-1}, n)$, and v_{k+1} is given by the first $n-1$ letters of $c(01^{k-1}, n)$, which are equal). In other words, the corresponding set of vertices in G_A^{n-1} has size $2k+1$ (see Figure 3). If $k = 2$, then this can be verified directly by considering the number of leading and trailing 0s and 1s of the vertices in V^0, V^1 and V^2. We now

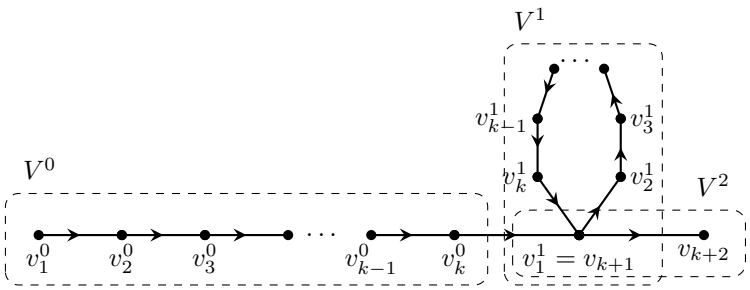

Fig. 3: Subgraph of G_A^{n-1} constructed in the proof of Theorem 10.

assume that $k \geq 3$. Every word from V^0, except possibly v_1^0, contains the factor 00 exactly once and is uniquely identified by the position of this factor, proving that $|V^0| = k$. The words in V^1 are all uniquely identified by the number of leading 1s, which equals 0 for v_1^1 and $k - i + 1$ for $i = 2, 3, \ldots, k$, implying that $|V^1| = k$. We now show that V^0 and V^1 are disjoint. To prove this we use again that all the words in V^0, except possibly v_1^0, contain the factor 00, and that moreover no word from V^1 contains this factor. However v_1^0 does not contain the factor 00 only in the case $k = n - 1$, and then v_1^0 starts and ends with 0, unlike any of the words from V^1 in this case, proving that V^0 and V^1 are disjoint. It remains to show that $v_{k+2} \notin V^0 \cup V^1$. If $k = n-1$, then $v_{k+2} = 1^{n-1}$ and all other words from V^0 and V^1 contain at least one 0, so $v_{k+2} \notin V^0 \cup V^1$. If $k \leq n - 2$, then the word $v_{k+2} = w_{k+2} \cdots w_{k+n}$ satisfies $w_{k+n} = \overline{w_n}$, i.e., its last letter and the one k positions to the left of it are complementary (recall the definition of $c(01^{k-1}, n)$), a property that does not hold for any of the words in V^1, implying that $v_{k+2} \notin V^1$. Moreover, in this case all words from V^0 contain the factor 00 exactly once and are uniquely identified by the position of this factor, and v_{k+2} might contain the factor 00 only at the last two positions, so the only potential conflict could arise in the case $k = n-2$ when $v_1^0 = 01^{n-4}00$ ends with 00. However, in this case $v_{k+2} = 1^{n-3}00$ is still different from v_1^0. We conclude that $v_{k+2} \notin V^0 \cup V^1$ in all cases. Combining these observations shows that $|V'| = |V^0| + |V^1| + |V^2| - 1 = 2k + 1$, as claimed.

Consider the set of $2k + 1$ edges $E' := \{(v_i^0, v_{i+1}^0) \mid i = 1, 2, \ldots, k-1\} \cup \{(v_i^1, v_{i+1}^1) \mid i = 1, 2, \ldots, k-1\} \cup \{(v_k^0, v_{k+1}), (v_k^1, v_{k+1}), (v_{k+1}, v_{k+2})\}$ in the De Bruijn graph G_A^{n-1} (see Figure 3). They span a subgraph on V' that has the following pairs of out-degrees and in-degrees: $(1, 0)$ for the vertex v_1^0, $(0, 1)$ for the vertex v_{k+2}, $(1, 1)$ for the vertices v_i^0 and v_i^1, $i = 2, 3, \ldots, k$, $(2, 2)$ for the vertex $v_1^1 = v_{k+1}$.

We denote the graph obtained from G_A^{n-1} by removing the edges in E' and the isolated vertex $v_1^1 = v_{k+1}$ by G'. Clearly, every vertex in G' has the same in- and out-degree (1 or 2), except the vertex v_{k+2} which has one more out-edge than in-edges, and the vertex v_1^0 which has one more in-edge than out-edges. To complete the proof of the theorem we show that G' contains an Eulerian path (which must start at v_{k+2} and end at v_1^0), and to do this, it suffices (by the before-mentioned degree conditions) to show that G' is connected.

We first consider the case $k \leq n - 2$: From any vertex $v \in G'$, we follow 0-edges until we either reach the vertex 0^{n-1} or a vertex from V' for which the next 0-edge is from E' (this could happen right at the beginning if $v \in V'$). In this case we follow 1-edges until we reach the vertex 1^{n-1}, and from there we follow 0-edges until we reach 0^{n-1}. (We only ever follow edges in forward direction.) We claim that in

this process we never use an edge from E', which shows that G' is connected. To see this suppose we encounter a vertex v' from V' for which the next 0-edge is from E'. This means that v' has $k-1$ trailing 1s (here we use that $k \leq n-2$), so following a 1-edge leads to a vertex that has k trailing 1s, and in the next step to a vertex that has $k+1$ trailing 1s. Note that the vertices in $V' \setminus \{v_{k+2}\}$ have at most $k-1$ trailing 1s, and v_{k+2} has at most k trailing 1s, so we avoid any edges from E' on our way to 1^{n-1}. Moreover, on the way from 1^{n-1} to 0^{n-1} via $1^{n-1-i}0^i$, $i = 1, 2, \ldots, n-1$, we do not use any edges from E' either, because any vertex from $V' \setminus \{v_{k+2}\}$ that starts with a 1 has at least two transitions from 1s to 0s, or vice versa, when reading it from left to right (using again $k \leq n-2$), and $0^{n-1} \notin V'$.

Now consider the case $k = n-1$: From any vertex $v \in G'$, we follow 0-edges until we either reach the vertex 0^{n-1} or the only vertex $v_1^0 = 01^{n-3}0$ from $V' \setminus \{v_{k+1}\}$ for which the next 0-edge is from E'. In this case we follow a single 1-edge to $1^{n-3}01 = v_3^1$, and from there we follow 0-edges until we reach 0^{n-1}. Similarly to before, we need to argue that we never use an edge from E' in this process. On the way from $1^{n-3}01 = v_3^1$ to 0^{n-1} we never use any edges from E', because any vertices on this path except the first one $1^{n-3}01$ and the last two 10^{n-2} and 0^{n-1} contain the factor 010, so all these vertices are different from V' (for $n \geq 5$ and $k = n-1$ no word from V' contains 010 as a factor), implying that all edges except possibly the last one are safe. However, since $0^{n-1} \notin V'$, the last edge is safe, too.

These arguments show that G' is connected, so it has an Eulerian path, and this Eulerian path yields the desired upword that begins with w. This completes the proof. □

4 Linear upwords with two $\diamond$s

In this section we focus on binary alphabets. Many of the non-existence conditions provided in this section can be generalized straightforwardly to non-binary alphabets, as we briefly discuss in Section 6 below.

4.1 Non-existence results

Theorem 11 *For $A = \{0, 1\}$ and any $n \geq 5$, there is no linear upword for A^n with two $\diamond$s of the form $u = x\diamond y\diamond z$ if $|x|, |y|, |z| \geq n$ or $|x| = n-1$ or $|z| = n-1$ or $|y| \leq n-2$.*

As Table 2 shows, there are examples of linear upwords with two $\diamond$s whenever the conditions in Theorem 11 are violated. Put differently, for every upword $u = x\diamond y\diamond z$ in the table for $n \geq 5$ we have that one of the numbers $|x|, |y|, |z|$ is at most $n-1$, $|x| \neq n-1$, $|z| \neq n-1$ and $|y| \geq n-1$. Note that already by the first condition $|x|, |y|, |z| \geq n$, a $(1-o(1))$-fraction of all choices of placing two $\diamond$s among $N = \Theta(2^n)$ positions are excluded as possible candidates for upwords.

Proof: We first assume that $|x|, |y|, |z| \geq n$, i.e., $y_n, z_n \in A$. Applying Lemma 4 yields $z_i = y_i = x_i \in A$ for $i = 1, 2, \ldots, n-1$ and $z_n = y_n = \overline{x_n}$, so the word $y_1y_2 \cdots y_n = z_1z_2 \cdots z_n$ appears twice as a factor in u, a contradiction.

We now assume that $|x| = n-1$ (the case $|z| = n-1$ follows by symmetry). Note that the number of factors of u is at most $2(|y|+1) + 4(|z|+1)$: This is because every subword ending at the first $\diamond$ or at a letter from y contains at most one $\diamond$, giving rise to two factors, and every subword ending at the second $\diamond$ or at a letter from z contains at most two $\diamond$s, giving rise to four factors. This number is at most $2n + 4n = 6n$ for $|y|, |z| \leq n-1$, which is strictly less than 2^n for $n \geq 5$. Therefore, we must have $|y| \geq n$ or $|z| \geq n$ in this case. We assume w.l.o.g. that $|y| \geq n$, i.e., $y_n \in A$. Applying Lemma 4 yields $y_i = x_i \in A$ for $i = 1, 2, \ldots, n-1$, implying that the word $y_1y_2 \cdots y_n$ appears twice as a factor in u, a contradiction.

We now assume that $|y| \leq n-2$. In this case we must have $|x| \geq n$ or $|z| \geq n$, because if $|x|, |z| \leq n-1$ then the number of factors of u is at most $2(|y|+1)+4(|z|+1) \leq 2(n-1)+4n \leq 6n$, which is strictly less than 2^n for $n \geq 5$. We assume w.l.o.g. that $|z| \geq n$. Let $k := |y|+1 \leq n-1$ and consider the subword $y' := y\diamond z_1 z_2 \cdots z_{n-k}$ of u, which is well-defined since $|z| \geq n$ (k is the position of the $\diamond$ in y'). Since $k \leq n-1$ we have $y'_n = z_{n-k} \in A$. Applying Lemma 4 yields that $|x| = |y|$. Moreover, if $k = 1$ ($|x| = |y| = 0$) then the same lemma yields $y'_2 = y'_3 = \cdots = y'_{n-1} = \overline{y'_n}$, i.e., $z_1 = z_2 = \cdots = z_{n-2} = \overline{z_{n-1}}$ and $z_{n-1} = z_{n-3}$, a contradiction. On the other hand, if $k \geq 2$, then $z_{i+k\ell} = y_i = x_i$ for all $i = 1, 2, \ldots, k-1$ and $\ell = 0, 1, \ldots$ with $i + k\ell \leq n-1$, i.e., the factors obtained from the subword y' in u appear twice, starting at position 1 and position $k+1$, a contradiction. □

Corollary 12 *For $A = \{0,1\}$ and any $n \geq 2$, $\diamond\diamond$ for $n = 2$ and $\diamond\diamond 0111$ for $n = 3$ are the only linear upwords for A^n containing two $\diamond$s that are adjacent (up to reversal and letter permutations).*

Proof: The non-existence of linear upwords with two adjacent $\diamond$s for $n \geq 5$ follows from Theorem 11, because for such an upword $u = x\diamond\diamond z$ the subword y between the two $\diamond$s is empty, so $|y| = 0 \leq n-2$. For $n = 4$ and $|y| = 0$ the estimate in the third part in the proof of Theorem 11 can be strengthened to show that if $|x|, |z| \leq n-1$, then the number of factors of u is strictly less than $4n \leq 2^n$ unless $u = u_1u_2u_3\diamond\diamond u_6u_7u_8$, which means we can continue the argument as before, leading to a contradiction. The exceptional case $u = u_1u_2u_3\diamond\diamond u_6u_7u_8$ can be excluded as follows: Applying Lemma 4 shows that $u_2 = u_6$ and $u_3 = u_7$, and then it becomes clear that the factor 0000, at whatever position within u it is placed, would appear twice. For $n = 3$ the only possible linear upwords with two adjacent $\diamond$s by Lemma 4 are $u = \diamond\diamond u_3\overline{u_3u_3}u_6$, which leads to $\diamond\diamond 0111$ (w.l.o.g. $u_3 = 0$, and for 111 to be covered we must have $u_6 = 1$), and $u = u_1\diamond\diamond u_4$ is impossible because $u_1 0 u_4$ appears twice as a factor (starting at positions 1 and 2). For $n = 2$ the only possible linear upword with two $\diamond$s is $\diamond\diamond$. □

4.2 Existence results

Our next result provides an infinite number of binary linear upwords with two $\diamond$s (see Table 2).

Theorem 13 *For $A = \{0,1\}$ and any $n \geq 4$, there is a linear upword for A^n with two $\diamond$s that begins with $\diamond 0^{n-1}1^{n-2}\diamond 10^{n-2}1$.*

Proof: Consider the word $w = w_1w_2\cdots w_{3n-1} := \diamond 0^{n-1}1^{n-2}\diamond 10^{n-2}1$. It is easy to check that w yields $3n+1$ different factors $x_1x_2\cdots x_n \in A^n$, and each of these factors gives rise to an edge $(x_1x_2\cdots x_{n-1}, x_2x_3\cdots x_n)$ in the De Bruijn graph G_A^{n-1}. The set E' of these edges and their end vertices V' form a connected subgraph that has in- and out-degree 1 for all vertices in V' except for $v'_0 := 0^{n-1}$, $v'_1 := 1^{n-1}$, $v'_2 := 10^{n-2}$ and $v'_3 := 1^{n-2}0$ which have in- and out-degree 2, and $y := 0^{n-2}1$ and $z := 01^{n-2}$ which have in-degree 2 and out-degree 1, or in-degree 1 and out-degree 2, respectively. We denote the graph obtained from G_A^{n-1} by removing the edges in E' and the vertices v'_0, v'_1, v'_2 and v'_3 by G'. Clearly, every vertex in G' has the same in- and out-degree (1 or 2), except the vertex y which has only one outgoing edge, and the vertex z which has only one incoming edge. To complete the proof of the theorem we show that G' contains an Eulerian path (which must start at y and end at z), and to do this, it suffices (by the before-mentioned degree conditions) to show that G' is connected.

If $n = 4$, then G' consists only of the edges $(y, 010), (010, 101), (101, z)$ (a connected graph), so for the rest of the proof we assume that $n \geq 5$. Consider a vertex v in G' other than z.

If v ends with 0, consider the (maximum) number k of trailing 0s. Note that $k \leq n-3$, as the vertices v'_2 and v'_0 that correspond to the cases $k \in \{n-2, n-1\}$ are not in G'. From v we follow 1-edges and 0-edges alternatingly, starting with a 1-edge, until we either reach the vertex $s := 1^{n-3}01$ or the vertex $t := 010101\cdots \in A^{n-1}$ (this could happen right at the beginning if $v = t$). From s or t we follow 1-edges until the vertex z.

If v ends with 1, then we do the following: If $v \neq s$ we follow a single 0-edge, and then proceed as before until the vertex z. If $v = s$ we directly follow 1-edges until z. (Note that we only ever follow edges in forward direction.)

We claim that in this process we never use an edge from E', which shows that G' is connected. To see this we first consider the case that we start at a vertex v with $k \leq n-3$ trailing 0s. If $k \geq 2$, then the vertex reached from v via a 1-edge is not in V', because no vertex in V' has a segment of $k \leq n-3$ consecutive 0s surrounded by 1s. Also, none of the next vertices before reaching t is from V', because all contain the factor 0010, unlike any word in V'. If $k = 1$, then the vertex reached from v by following a 1-edge is either $s \in V'$ (then we stop) or not in V', as no other vertex from V' ends with 101. If it is not in V', then the next vertex reached via a 0-edge could be in V', but all the subsequent vertices until (and including) t are not, since they all contain the factor 0101, unlike any word in V'. This shows that none of the edges traversed from v to s or t is from E'. Moreover, none of the vertices traversed between s and z or between t and z is from V', because they all contain the factor 0101 or 1011, unlike any word in V', so we indeed reach z without using any edges from E'.

Now consider the case that we start at a vertex v that ends with 1. The only interesting case is $v \neq s$. There are only two 0-edges in E' starting at a vertex that ends with 1, namely the edges starting at v'_1 and s. However, v is different from v'_1 because v'_1 is not part of G', and v is different from s by assumption. We conclude that the 1-edge we follow is not from E'.

These arguments show that G' is connected, so it has an Eulerian path, and this Eulerian path yields the desired upword that begins with w. This completes the proof. □

5 Cyclic upwords

Throughout this section, all indices are considered modulo the size of the corresponding word. All the notions introduced in Section 2 can be extended straightforwardly to cyclic upwords, where factors are taken cyclically across the word boundaries. In particular, when defining the graph $H(u,n)$ for some cyclic upword $= u_1u_2\cdots u_N$ we consider the subsets of words $S(u,i,n)$ cyclically for all $i = 1,2,\ldots,N$. Then the first two statements of Observation 3 hold for all vertices $S(u,i,n)$, $i = 1,2,\ldots,N$. The next lemma is the analogue of Lemma 4 for cyclic upwords.

Lemma 14 *Let $u = u_1u_2\cdots u_N$ be a cyclic upword for A^n, where $A = \{0,1,\ldots,\alpha-1\}$ and $n \geq 2$. If $u_k = \diamond$ then $u_{k+n} = \diamond$.*

Proof: Suppose that $u_k = \diamond$ and $u_{k+n} \neq \diamond$. By Observation 3, each vertex in the set $S(u,k+1,n)$ has in-degree α in $H(u,n)$, and each vertex in $S(u,k,n)$ has out-degree 1. By Observation 2, for each $v = v_1v_2\cdots v_n \in S(u,k+1,n)$ there are $\alpha-1$ other vertices (different from the ones in $S(u,k+1,n)$) in G^n_A with the same set $\Gamma^-(v)$ of α many in-neighbors, namely $v_x := v_1\cdots v_{n-1}x$, where $x \in A\setminus\{v_n\}$. As the in-degree of every vertex of G^n_A is exactly α, and in $H(u,n)$ all vertices already have in-degree at

least 1, it follows that the vertices v_x can not be part of $H(u,n)$, a contradiction to the fact that $H(u,n)$ is a spanning subgraph of G_A^n. □

Lemma 15 immediately yields the following corollary, which captures various rather severe conditions that a cyclic upword must satisfy, relating its length N, the size α of the alphabet, and the value of the parameter n.

Corollary 15 *Let $u = u_1u_2\cdots u_N$ be a cyclic upword for A^n, where $A = \{0,1,\ldots,\alpha-1\}$ and $n \geq 2$, with at least one $\diamond$. Then we have $N = \alpha^{n-d}$ for some d, $1 \leq d \leq n-1$, such that n divides dN.*

Proof: By Lemma 14, for any $\diamond$ in u, the other two symbols in distance n from it must be $\diamond$s as well. Thus, the indices $1,2,\ldots,N$ are partitioned into $\gcd(n,N)$ many residue classes modulo n, and all symbols at positions from the same residue class are either all $\diamond$s or all letters from A. Let d denote the number of $\diamond$s among any n consecutive symbols of u, then we have $1 \leq d \leq n-1$ (there is at least one $\diamond$, but not all letters can be $\diamond$s), and any starting position in u gives rise to α^d different factors, implying that $N = \alpha^{n-d}$. Furthermore, the d many $\diamond$s within any n consecutive letters of u are partitioned into $n/\gcd(n,N)$ many blocks with the same $\diamond$ pattern, so $n/\gcd(n,N)$ must divide d, and this condition is equivalent to n dividing $d\gcd(n,N)$ and to n dividing dN. □

As an immediate corollary of our last result, we can exclude the existence of cyclic upwords for many combinations of α and n.

Corollary 16 *Let $A = \{0,1,\ldots,\alpha-1\}$ and $n \geq 2$. If $\gcd(\alpha,n) = 1$, then there is no cyclic upword for A^n. In particular, for $\alpha = 2$ and odd n, there is no cyclic upword for A^n.*

Proof: Suppose that such an upword $u = u_1u_2\cdots u_N$ exists. Then by Corollary 15 we have $N = \alpha^{n-d}$ for some d, $1 \leq d \leq n-1$, such that n divides dN. However, as $\gcd(\alpha,n) = 1$, n does not divide $N = \alpha^{n-d}$, so n must divide d, which is impossible, yielding a contradiction. □

By Corollaries 15 and 16, for a binary alphabet ($\alpha = 2$), the only remaining potential parameter values for cyclic upwords are $n = 2$ and $d = 1$, $n = 4$ and $d \in \{1,2\}$, $n = 6$ and $d = 3$, $n = 8$ and $d \in \{1,2,\ldots,6\}$, $n = 10$ and $d = 5$, $n = 12$ and $d \in \{3,6,9\}$, etc. The case $n = 2$ and $d = 1$ can be easily exluded: w.l.o.g. such a word has the form $\diamond 0$, leading to the factor 00 appearing twice (and 11 does not appear as a factor at all). However, for $n = 4$ and $d = 1$ we have the cyclic upword $\diamond 001\diamond 110$, which we already mentioned in the introduction. This is the only cyclic upword for a binary alphabet that we know of. Cyclic upwords for any even alphabet size $\alpha \geq 4$ and $n = 4$ have been constructed in the follow-up paper [16].

6 Outlook

In this paper we initiated the systematic study of universal partial words, and we hope that our results and the numerous examples of upwords provided in the tables (see also the extensive data available on the website [30]) generate substantial interest for other researchers to continue this exploration, possibly in one of the directions suggested below.

Concerning the binary alphabet $A = \{0,1\}$, it would be interesting to achieve complete classification of linear upwords containing a single $\diamond$, as suggested by Conjecture 8. For two $\diamond$s such a task seems somewhat more challenging (recall Table 2, Theorem 11 and see the data from [30]). Some examples

n	
3	$\diamond\diamond\diamond$
4	$\diamond\diamond\diamond$01111 (Thm. 17)
	$\diamond\diamond$001$\diamond$11010
	$\diamond$001$\diamond$110$\diamond$00
	0$\diamond$001$\diamond$110$\diamond$0
5	$\diamond$0010$\diamond$0111$\diamond$10011011000001
	$\diamond$0000111$\diamond$10001001101100101$\diamond$1
	$\diamond$00001110$\diamond$10001010011010111$\diamond$
	$\diamond$0000100111$\diamond$10001101100101$\diamond$1
	$\diamond$0000101110$\diamond$10001101010011111$\diamond$
	$\diamond$00001111101$\diamond$10001011001$\diamond$01
	$\diamond$00001010111 0$\diamond$10001101001111$\diamond$
	$\diamond$0000101001110$\diamond$10001101011111$\diamond$
	$\diamond$0000110110011 1$\diamond$1000100101$\diamond$1
	$\diamond$0000110101001110$\diamond$1000101111$\diamond$
	$\diamond$0000110110010011 1$\diamond$1000101$\diamond$1
	$\diamond$0000100101011111100$\diamond$1101$\diamond$00
	0$\diamond$1100$\diamond$0011111011010001 01$\diamond$1

Tab. 3: Examples of linear upwords for A^n, $A = \{0, 1\}$, with three $\diamond$s for $n = 3, 4, 5$.

of binary linear upwords with three $\diamond$s are listed in Table 3, and deriving some general existence and non-existence results for this setting would certainly be of interest.

The next step would be to consider the situation of more than three $\diamond$s present in a linear upword. The following easy-to-verify example in this direction was communicated to us by Rachel Kirsch [16].

Theorem 17 *For $A = \{0, 1\}$ and any $n \geq 2$, $\diamond^{n-1}01^n$ is a linear upword for A^n with $n - 1$ many $\diamond$s.*

Complementing Theorem 17, we can prove the following non-existence result in this direction, but it should be possible to obtain more general results.

Theorem 18 *For $A = \{0, 1\}$, any $n \geq 4$ and any $2 \leq d \leq n - 2$, there is no linear upword for A^n that begins with $\diamond^d x_{d+1}x_{d+2} \dots x_{n+2}$ with $x_i \in A$ for all $i = d + 1, \dots, n + 2$.*

The proof of Theorem 18 is easy by applying Lemma 4 to the first and second $\diamond$. We leave the details to the reader.

It would also be interesting to find examples of binary cyclic upwords other than $\diamond 001 \diamond 110$ for $n = 4$ mentioned before.

Finally, a natural direction would be to search for (linear or cyclic) upwords for *non-binary* alphabets, but we anticipate that no non-trivial upwords exist in most cases (recall Theorem 5). As evidence for this we have the following general non-existence result in this setting.

Theorem 19 *For $A = \{0, 1, \dots, \alpha - 1\}$, $\alpha \geq 3$, and any $d \geq 2$, for large enough n there is no linear or cyclic upword for A^n with exactly d many $\diamond$s.*

Theorem 19 shows in particular that for a fixed alphabet size α and a fixed number $d \geq 2$ of diamonds, there are only finitely many possible candidates for upwords with d diamonds (which in principle could all be checked by exhaustive search). The proof idea is that for fixed d and large enough n, such an upword must contain a $\diamond$ and a symbol from A in distance n, and then applying Lemma 4 or Lemma 14 yields a contradiction (recall the proof of Theorem 5). We omit the details here. On the positive side, upwords for even alphabet sizes $\alpha \geq 4$ and $n = 4$ have been constructed in [16] (and these upwords are even cyclic).

A question that we have not touched in this paper is the algorithmic problem of efficiently generating upwords. As a preliminary observation in this direction we remark here that some of the linear upwords constructed in Theorem 9 and 10 can also be obtained by straightforward modifications of the FKM de Bruijn sequences constructed in [15, 13], for which efficient generation algorithms are known [26].

Acknowledgements

The authors thank Martin Gerlach for his assistance in our computer searches, Rachel Kirsch and her collaborators [16], as well as Artem Pyatkin for providing particular examples of small upwords. The second author is grateful to Sergey Avgustinovich for helpful discussions on universal partial words, and to Bill Chen and Arthur Yang for their hospitality during the author's visit of the Center for Combinatorics at Nankai University in November 2015. This work was supported by the 973 Project, the PCSIRT Project of the Ministry of Education and the National Science Foundation of China. We also thank the anonymous referees of this paper for several valuable suggestions and references that helped improving the presentation.

References

[1] J. Berstel and L. Boasson. Partial words and a theorem of Fine and Wilf. *Theoret. Comput. Sci.*, 218(1):135–141, 1999. WORDS (Rouen, 1997).

[2] B. Blakeley, F. Blanchet-Sadri, J. Gunter, and N. Rampersad. On the complexity of deciding avoidability of sets of partial words. *Theoret. Comput. Sci.*, 411(49):4263–4271, 2010.

[3] J. Bang-Jensen and G. Z. Gutin. *Digraphs: Theory, Algorithms and Applications*. Springer, 2nd edition, 2008.

[4] A. Burstein and S. Kitaev. On unavoidable sets of word patterns. *SIAM J. Discrete Math.*, 19(2):371–381, 2005.

[5] F. Blanchet-Sadri. Primitive partial words. *Discrete Applied Mathematics*, 148(3):195–213, 2005.

[6] F. Blanchet-Sadri. Open problems on partial words. In G. Bel-Enguix, M. D. Jiménez-López, and C. Martín-Vide, editors, *New Developments in Formal Languages and Applications*, pages 11–58. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[7] F. Blanchet-Sadri, N. C. Brownstein, A. Kalcic, J. Palumbo, and T. Weyand. Unavoidable sets of partial words. *Theory Comput. Syst.*, 45(2):381–406, 2009.

[8] F. Chung, P. Diaconis, and R. Graham. Universal cycles for combinatorial structures. *Discrete Math.*, 110(1-3):43–59, 1992.

[9] A. Claesson, V. Jelínek, E. Jelínková, and S. Kitaev. Pattern avoidance in partial permutations. *Electron. J. Combin.*, 18(1):Paper 25, 41, 2011.

[10] P. E. C. Compeau, P. A. Pevzner, and G. Tesler. How to apply De Bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987–991, 2011.

[11] N. G. de Bruijn. A Combinatorial Problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, 49:758–764, 1946.

[12] L. Euler. Solutio problematis ad geometriam situs pertinentis. *Comment. Academiae Sci. I. Petropolitanae*, 8:128–140, 1736.

[13] H. Fredricksen and I. J. Kessler. An algorithm for generating necklaces of beads in two colors. *Discrete Math.*, 61(2-3):181–188, 1986.

[14] M. Fleury. Deux problemes de geometrie de situation. *Journal de mathematiques elementaires*, pages 257–261, 1883.

[15] H. Fredricksen and J. Maiorana. Necklaces of beads in k colors and k-ary de Bruijn sequences. *Discrete Math.*, 23(3):207–210, 1978.

[16] B. Goeckner, C. Groothuis, C. Hettle, B. Kell, P. Kirkpatrick, R. Kirsch, and R. Solava. Universal partial words over non-binary alphabets. *arXiv:1611.03928*, Nov 2016.

[17] V. Halava, T. Harju, and T. Kärki. Square-free partial words. *Inform. Process. Lett.*, 108(5):290–292, 2008.

[18] V. Halava, T. Harju, T. Kärki, and P. Séébold. Overlap-freeness in infinite partial words. *Theoret. Comput. Sci.*, 410(8-10):943–948, 2009.

[19] A. E. Holroyd, F. Ruskey, and A. Williams. Shorthand universal cycles for permutations. *Algorithmica*, 64(2):215–245, 2012.

[20] G. H. Hurlbert. *Universal cycles: On beyond de Bruijn*. ProQuest LLC, Ann Arbor, MI, 1990. Thesis (Ph.D.)–Rutgers The State University of New Jersey - New Brunswick.

[21] C. Hierholzer and C. Wiener. Ueber die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren. *Math. Ann.*, 6(1):30–32, 1873.

[22] A. Lempel. m-ary closed sequences. *J. Combinatorial Theory Ser. A*, 10:253–258, 1971.

[23] P. Leupold. Partial words for DNA coding. In *DNA computing*, volume 3384 of *Lecture Notes in Comput. Sci.*, pages 224–234. Springer, Berlin, 2005.

[24] J. Pagès, J. Salvi, C. Collewet, and J. Forest. Optimised De Bruijn patterns for one-shot shape acquisition. *Image and Vision Computing*, 23(8):707 – 720, 2005.

[25] A. Ralston. De Bruijn sequences—a model example of the interaction of discrete mathematics and computer science. *Math. Mag.*, 55(3):131–143, 1982.

[26] F. Ruskey, C. Savage, and T. M. Y. Wang. Generating necklaces. *J. Algorithms*, 13(3):414–430, 1992.

[27] H. Sohn, D. L. Bricker, J. R. Simon, and Y. Hsieh. Optimal sequences of trials for balancing practice and repetition effects. *Behavior Research Methods, Instruments, & Computers*, 29(4):574–581, 1997.

[28] E. R. Scheinerman. Determining planar location via complement-free De Brujin sequences using discrete optical sensors. *IEEE Transactions on Robotics and Automation*, 17(6):883–889, Dec 2001.

[29] B. Stevens and A. Williams. The coolest way to generate binary strings. *Theory Comput. Syst.*, 54(4):551–577, 2014.

[30] currently `http://www.math.tu-berlin.de/~muetze`.

[31] M. Yoeli. Binary ring sequences. *Amer. Math. Monthly*, 69:852–855, 1962.