UNIVERSITAT ROVIRA I VIRGILI

REPETITIONS IN PARTIAL WORDS

Robert George Mercas

ISBN: 978-84-693-7670-6
Dipòsit Legal: T-1749-2010

Robert Mercaş

Repetitions in Partial Words

Doctoral Thesis

Supervised by

Francine Blanchet-Sadri

Department of Romance Studies

Universitat Rovira i Virgili

Tarragona, Spain, 2010

Supervisor :

Professor Francine Blanchet-Sadri
Department of Computer Science
University of North Carolina
P.O. Box 26170
27402–6170 Greensboro, NC
USA

Tutor :

Dr. Maria Dolores Jiménez López
Grup de Recerca en Lingüística Matemàtica
Departament de Filologies Romániques
Universitat Rovira i Virgili
Av. Catalunya 35
43002 Tarragona
SPAIN

Abstract

The object of this thesis is repetitions in partial words, words that besides the regular letters, may have a number of unknown symbols, called “holes”, “wild cards” or “do not know” symbols. More specifically, we present and conclude an extension of the notion of repetition-freeness introduced by Axel Thue. We do a counting of the maximum distinct number of squares (2-repetitions) compatible with factors of a partial word. We also study some properties of unbordered and primitive partial words and give a characterization of the language of partial words having a critical factorization.

Contents

List of Figures

Acknowledgements

În primul rând aş vrea sa îi mulţumesc mamei mele pentru tot ceea ce a făcut pentru mine. În acelaşi timp aş vrea sa mulţumesc familiei care, mai uşor sau mai greu, m-a înţeles şi m-a susţinut. Aş vrea sa le mai mulţumesc surorii mele Delia, tatălui meu Aurel, bunicilor Maricica şi Sandu, matuşii Cristina şi Ciprianei.

I would like to thank professor Francine Blanchet-Sadri, for all her kindness and both scientific and financial help. I think that without her help I would have not managed finishing this thesis (maybe not even starting it). I would also want to thank professor Victor Mitrana for his help and advices. He put his confidence in me and I hope I have not let him down.

I would like to thank all my friends that have been there for me both for the good and the not so great moments of these last four years: Patricia Botez Voicu, Isabela Butnar, Adrian Dediu, Dragoş and Elena Deliu, Bogdan Drăgan, Cezara Drăgoi, Constantin Enea, Szilard Faszekas, Lourdes Fernandez Hernandez, Mihai and Raluca Ionescu, Anita Kerekes, Dan and Simona Libotean, Remco Loos, Călina and Florin Manea, Orsi and Zoltan Mecsei, Costel Moraru, Areti Panou, Ioana Scripa, Georgiana Stoica, Cristina and Cătălin Tîrnăucă, Danny Verboekeren. Tons of thanks as well for Lilica Voicu who helped me a lot during this time.

Moltes gràcies a la Gemma Bel Enguix i la Maria Dolores Jiménez López per l'ajuda donada. Ho agraeixo moltíssim. Quiero agradecer también al Prof. Carlos Martin Víde por aceptarme en 5th International PhD School in Formal Languages and Applications y por los valiosos consejos que me ha dado a lo largo del tiempo.

Special thanks to all professors of the 5th International PhD School in Formal Languages and Applications also to all my coauthors: Florin Manea, F. Blanchet-Sadri, Geoffrey Scott, Crystal D. Davis, Margaret Moorefield, Joel Dodge, Gemma Bel Enguix, Maria Dolores Jiménez López, Alexander Perekrestenko, Mihai Cucuringu, Kristen Wetzler, Emily Allen, Cameron Byrum, Elara Willet, Abraham Rashin, Christian Choffrut, Alfonso Ortega, Emilio del Rosal García, Diana Pérez, Manuel Alfonseca, Ilkyoo Choi, Jane Kim, William Severa, Sean Simmons, Eric Weissenstein.

I would also like to thank MEC (Ministerio de Educación del Gobierno de España, Mobility Grant), NSF (National Science Foundation of

United States, Grants No. DMS–0452020 and DMS–0754154), ESF (European Science Foundation, Grant AutoMathA: from Mathematics to Applications) and URV (Universitat Rovira i Virgili, Grant No. 1323 U07 E30 N-2008/InvAct/Bel, G./BJ01) for their financial help during these years.

Chapter 1

Introduction

The area of Combinatorics on Words took birth at the beginning of the last century, when Thue initiated a systematic study of words in a series of papers [Thu06, Thu10, Thu12, Thu14]. In these papers, several combinatorial problems that arose in the study of the sequences of symbols were considered, problems which were solved with the usual tools of discrete mathematics. While the second and the fourth paper deal with word problems for finitely presented semigroups (the so called *Thue system*), where he managed to solve the problem for special cases (in 1947 the general case is showed to be unsolvable independently in [Mar41] and [Pos47]), the first and third paper contained results regarding repetitions (consecutive occurrences of a factor) inside a word (see [Ber92, Thu06, Thu12], or "Section 1.6: Repetitions in words" from [AS03]).

Since then, several results on repetitions in words have been re-proven several times in various ways. According to Currie, [Cur93], "One reason for this sequence of rediscoveries is that non-repetitive sequences have been used to construct counterexamples in many areas of mathematics: ergodic theory, formal language theory, universal algebra and group theory, for example...."

The topic becomes of more interest in the early 80's due to its connection to avoidability of patterns [BEM79] and the theory of fixed points of iterated morphisms.

In the 90's Fraenkel and Simpson "restart" the study of squares, two consecutive repetitions of the same factor, in words. More precisely they are at first interested in infinite words avoiding large squares (squares of length 2×3 or higher), see [FS95] and later on in counting the total number of distinct squares a word might have [FS98]. A few years ago, Ilie gave a simpler proof of this result [Ili05], and then improved the result by a small margin [Ili07]. The subject also gained more interest in algorithmics, as Gusfield and Stoye gave in 2004 a linear time algorithm that marks end points of squares in the suffix tree of the word [GS04].

Periodicity and *borderedness* are two fundamental properties of words that play a role in several research areas including string searching algorithms [CP91, CR94, CR02, GS83], data compression [Sto88], theory of codes [BP85], sequence assembly [MS95] and superstrings [BJJ97] in computational biology, and serial data communication systems [BI80]. It is well known that these two word properties do not exist independently from each other.

Having as motivation some intriguing practical problems that appear as applications of the central topics in the field of Combinatorics on Words, such as gene comparison, Berstel and Boasson suggested the usage of partial words in this context, [BB99]. Partial words, a canonical extension of the classical words, are sequences that, besides regular letters, may have a number of unknown symbols, called “holes” or “do not know” symbols. More precisely, the holes, denoted by $\diamond$, can be taken as any of the letters of the alphabet the word is defined on. Molecular biology, in particular, has stimulated a considerable interest in the study of combinatorics on partial words; for example, the alignment of the DNA sequences is conceived as a construction of two compatible partial words [BB99, Leu05].

Until now, several combinatorial properties of the partial words have been investigated. Among these we mention periodicity, i.e., Fine and Wilf’s Theorem for weak and strong periods (see [SK01, BSH02, SG04, BSBS08, BSOR]), conjugacy and primitivity, i.e., Defect Theorem, Critical Factorization Theorem (see [BS04, BS05, BSD05, BSW07]), avoidability of sets, i.e., properties, decidability and complexity (see [BSBP07, BSBK$^+$09, BSJP09, BBSGR09, BBSG$^+$09]). Also, in [BS04], Blanchet-Sadri made a first step in investigating languages of partial words by introducing the concept of pcodes, sets of partial words fulfilling a code-like property. New approaches from this point of view were also given by Leupold in [Leu04], and Lischke in [Lis06]. Following all these, the study of repetitions in partial words comes somehow natural, considering the full words history.

This thesis contains three main parts structured as follows:

The first part of this work refers to the classical problem of repetitions. If we consider the English word `banana`, then we notice that the letter `b` is followed by a $\frac{5}{2}$-repetition of the word `an`, that is two full occurrences of `an` followed by the first letter `a`. It is easy to note that over a binary alphabet all words of length 4 or larger contain a repetition. Since in many applications the length of the words investigated can be arbitrarily long, it was natural to study infinite words. The concept of partial word is extended to that of infinite partial word. In this framework, we will present the problem of identifying and constructing k-free partial words, i.e., words that do not contain k consecutive factors which are pairwise compatible. The study is aimed in two directions: the interest is in both combinatorial and algorithmic aspects regarding the k-freeness of infinite partial words. We prove that in order to avoid squares (2-repetitions) and overlaps ($\frac{5}{2}$-repetitions) a ternary alphabet is necessary when we have an infinity of holes. For powers of 3 or larger, three or more repetitions of the same factor, we can construct such words over a binary alphabet. Also, we show that for alphabets of size eight, five and four, we can construct infinite full words in which, replacing arbitrary positions with holes keeps the partial word, square-, overlap- and, respectively, cube-free. Moreover we prove that these results are optimal. The result for cubes appeared as joint work with Florin Manea [MM07], the results for squares and overlaps represent joint work with Francine Blanchet-Sadri and Geoffrey Scott [BSMS09] and Francine Blanchet-Sadri, Abraham Rashin and Elara Willett [BSMRW09].

In the second part we study the maximum number of squares a (partial) word can have. The most common approach when counting the maximum number of distinct squares a word of a certain length can have, is that of counting the last occurrence of each of the squares. In the case of full words, Fraenkel and Simpson proved that, for a word w of length n, if we denote by $s_w(i)$ the number of squares that have their last occurrence at position i in the word w, then $s_w(i) < 3$. In the case of partial words, it is somehow natural to count the maximum number of distinct full words that are squares and are compatible with factors of our partial word. We prove that the bounds for these words are related to the length of the partial word, the alphabet size the words are defined on, and the number of holes they contain. Moreover, we prove that the number of distinct full squares compatible with factors of a partial word with one hole of length n is bounded by $\frac{7n}{2}$. These results represent a joint work with Francine Blanchet Sadri [BSM09] and also Geoffrey Scott [BSMS08].

The last part of the thesis refers to primitive and unbordered words. A word is called primitive if it cannot be expressed as a power of another word. A word is unbordered if it is primitive and, none of its proper prefixes equals one of its proper suffixes. It is well known that these two word properties do not exist independently from each other. A very important property of unbordered words is that none of the unbordered factors of a word overlap. At first we extend to partial words a result of Ehrenfeucht and Silberger [ES79] which states that if a word can be written as concatenation of non-empty prefixes of another word, then it can be written as a unique concatenation of non-empty unbordered prefixes of the second word. We also study properties of the longest unbordered prefix of a partial word and investigate the relationship between the minimal *weak* period of a partial word and the maximal length of its unbordered factors. Later on, we investigate the maximum number of holes a partial word can have and still fail to be bordered. Finally we discuss the so-called critical factorization theorem and look into some properties of it when the partial words are unbordered. Moreover, we show that the language generated by all partial words having a critical factorization is context sensitive according to the Chomsky hierarchy. This part represents joint work with Francine Blanchet Sadri together with Crystal D. Davis, Joel Dodge and Margaret Moorefield [BSDD^{+}09], and Emily Allen and Cameron Byrum [ABSBM09].

Chapter 2

Preliminaries

In this chapter we present the main definitions and results that are to be used throughout this paper.

In the following, we denote by $\mathbb{N}$ the set of natural numbers (note that $0 \in \mathbb{N}$). For $i, j \in \mathbb{N}$, $\{i, \ldots, j\}$ denotes the set $\{k \mid k \in \mathbb{N} \text{ and } i \leq k \leq j\}$.

2.1 Finite and infinite words

Let A be a non-empty finite set, called *alphabet*. An element a from A is usually called *symbol* or *letter*; if A has k elements it is called a *k-letter alphabet* and its cardinality is denoted by $||A||$.

A *finite word* w over the alphabet A is a finite sequence of letters from A; usually, a finite word is depicted as $w = a_0 \cdots a_{n-1}$, where $a_i \in A$ for $0 \leq i < n$. The sequence with no letters, or the *empty word*, is denoted by ε. Observe that a finite word $w = a_0 \cdots a_{n-1}$ can be defined as a mapping $w : \{0, \ldots, n-1\} \rightarrow A$, with $w(i) = a_i$, for $0 \leq i < n$.

For example the sequence of letters *banana* represents a word defined over the alphabet $A = \{a, b, n\}$.

Similarly, a *one-way infinite word* is depicted as: $w = a_0 a_1 a_2 \cdots$, and can be formally defined as a mapping from $\mathbb{N}$ to A, that associates to each position of the word the letter that is present at that position.

We denote by A^* the *set of finite words* over the alphabet A, by A^+ the *set of non-empty finite words* over A, and by A^ω the *set of one-way infinite words* over the same alphabet. It is not hard to see that A^* is the free monoid generated by A, under the operation of catenation of words (the catenation of two words u and v is defined as the string uv), while the unit element in this monoid is represented by the empty word ε. We stress out the fact that we can apply catenation also to pairs consisting of a finite word and a one-way infinite word, given that the left factor is finite.

The *length* of a finite word w over the alphabet A, denoted by $|w|$, is defined as the number of occurrences of the letters from A in that word. We denote by A^n the set of all *words of length* n over the alphabet A. A finite word v is said to be a *factor* of the (infinite) word u if $u = xvy$, where x is a finite word. Moreover, v is a *prefix* of u if $x = \varepsilon$ and v is finite, and v is a *suffix* of u if $y = \varepsilon$ and x is a finite word (note that v is infinite if u is infinite). We say that v is a proper factor of u if $v \neq \varepsilon$ and $v \neq u$. A *factorization* of a word w is a sequence of words $w_0, w_1, \ldots, w_i$ such that $w = w_0 w_1 \cdots w_i$.

Considering again the word *banana* of length 6, we have that *ba* is a prefix of it, *na* is a proper factor and a suffix, and ba, na, na is one of its factorizations.

For a word u over A, the powers of u are defined inductively by $u^0 = \varepsilon$ and, for any $n \geq 1$, $u^n = uu^{n-1}$. If u is non-empty, then v is a *root* of u if $u = v^n$ for some positive integer n. The shortest root of u, denoted by $\sqrt{u}$, is called the *primitive root* of u, and u is itself called *primitive* if $\sqrt{u} = u$. If $u = (\sqrt{u})^n$, then $\sqrt{u}$ is the unique primitive word v and n is the unique positive integer such that $u = v^n$. All positive powers of u have the same primitive root. As an example, the word *nana* has *na* as a primitive root and can be expressed as $nana = (na)^2$.

A word w is called *p-periodic*, if for all positions i, j with $0 \leq i, j < |w|$, $w(i) = w(j)$ whenever $i \equiv j \bmod p$. More precisely, all symbols that are a multiple of p apart are equal. The word *anana* has a period of 2

A non-empty word w is *unbordered* if $p(w) = |w|$. Otherwise, it is *bordered.* A non-empty word x is a *border* of a word w if $w = xv = ux$ for some non-empty words u and v. Unbordered words turn out to be primitive. Considering the previous examples, the word *banana* is unbordered, while the word *anana* has borders of length 1 and 3.

Given two alphabets A and B, a *morphism* is a mapping $\phi : A^* \to B^*$ that satisfies $\phi(uv) = \phi(u)\phi(v)$, for all $u, v \in A^*$. Since A^* is the free monoid generated by A, ϕ is completely defined by the values $\phi(a)$, for all $a \in A$, and $\phi(\varepsilon) = \varepsilon$. Given a morphism ϕ we can canonically define how this morphism works for infinite words. For example in the case of $w = a_0a_1a_2 \cdots \in A^\omega$, we have $\phi(w) = \phi(a_0)\phi(a_1)\phi(a_2)\cdots$. If $\phi : A^* \to A^*$ is a morphism such that there exists a letter $a \in A$ verifying $\phi(a) = aw$ with $w \in A^+$, then ϕ is said to be *prolongable* on $a \in A$. Because a is a prefix of $\phi(a)$, it follows that $\phi^i(a)$ is a prefix of $\phi^{i+1}(a)$. Consequently, the limit (called the infinite word defined by iterating the morphism ϕ) $w = \lim_{i\to\infty} \phi^i(a)$ exists. This infinite word is a fixed point of the morphism ϕ, i.e. $\phi(w) = w$.

For a more detailed presentation of these aspects, as well as for the proofs of the results cited here, we refer to [CK97, Lot97, KL06].

2.2 Finite and infinite partial words

A *partial word* of length n over the alphabet A is defined as a partial function $w : \{0, \ldots, n-1\} \to A$. For $i \in \{0, \ldots, n-1\}$, if $w(i)$ is defined we say that i belongs to the *domain* of w (denoted by $i \in D(w)$), otherwise we say that i belongs to *the set of holes* of w (denoted by $i \in H(w)$). A partial word having an empty set of holes is called *full word*.

Let $\diamond$ be a symbol that does not belong to A. For convenience, partial words are seen as full words over the extended alphabet $A_\diamond = A \cup \{\diamond\}$. If w is a partial word of length n over A, then w is the total function (or the full word) $w : \{0, \ldots, n-1\} \to A \cup \{\diamond\}$. For example, the partial word $w = a\diamond bb\diamond ab\diamond a$ will have $D(w) = \{0, 2, 3, 4, 5, 6, 8\}$ and $H(w) = \{1, 4, 7\}$.

Usually, a partial word w of length n is depicted as $w = a_0 \cdots a_{n-1}$, where $a_i = w(i)$. In this way, one can easily define the catenation of partial words, as the catenation of the corresponding full words over $A_\diamond$, and the length of partial words, as the length of the corresponding full words over $A_\diamond$. Quite naturally, we denote by $A_\diamond^*$ the set of *finite partial words* over the alphabet A, by $A_\diamond^+$ the set of *non-empty finite partial words* over A and by $A_\diamond^n$ the set of all partial words of length n over the same alphabet.

The partial words u and v are said to be *equal* if u and v have the same length, $D(u) = D(v)$ and $u(i) = v(i)$ for all $i \in D(u)$. If u and v are two partial words of equal length, then u is said to be *contained* in v, $u \subset v$, if all the elements of $D(u)$ are contained in $D(v)$ and $u(i) = v(i)$ for all $i \in D(u)$. We say that u is *properly contained* in v, $u \sqsubset v$, if $u \subset v$ and $u \neq v$. Note that for a full word u and a partial word v, with $|u| = |v|$, if $u \subset v$ then $u = v$.

The *least upper bound* of u and v is denoted by $u \vee v$. By this we mean $u \subset u \vee v$ and $v \subset u \vee v$ and $D(u \vee v) = D(u) \cup D(v)$. Note that if $u(i) \neq v(i)$ for some $i \in D(u) \cap D(v)$, then $u \vee v$ is not defined. For example, considering the words $u = a\diamond b\diamond\diamond c$ and $v = ab\diamond c\diamond c$ we have $u \vee v = abbc\diamond c$ with $a\diamond b\diamond\diamond c \subset abbc\diamond c$ and $ab\diamond c\diamond c \subset abbc\diamond c$.

Similarly to the case of full words, we say that the partial word v is a *factor* of the partial word u if there exist partial words x and y such that $u = xvy$. If $x = \varepsilon$ we say that v is a *prefix* of u, and if $y = \varepsilon$ we say that v is a *suffix* of u. Also, v is a *proper factor* if both $v \neq \varepsilon$ and $v \neq u$. If $w = a_0 \cdots a_{n-1}$, we denote by $w[i..j]$ the factor $a_i \cdots a_j$ of w, and by $w(i)$ the symbol a_i. We say that $w(i)$ is the symbol placed at the ith position in the partial word w. Just like the full word case, a *factorization* of a partial word w is a sequence of partial words $w_0, w_1, \ldots, w_i$ such that $w = w_0 w_1 \cdots w_i$, and is sometimes denoted as $(w_0, w_1, \ldots, w_{i-1})$.

The unique *maximal common prefix* of u and v will be denoted by $\text{pre}(u, v)$. For a subset X of $A_\diamond^*$, we denote by $P(X)$ the set of prefixes of elements in X and by $S(X)$ the set of suffixes of elements in X. If X is the singleton $\{w\}$, then $P(X)$ (respectively, $S(X)$) will be abbreviated by $P(w)$ (respectively, $S(w)$). The elements of the set $X = \{aba, ab\diamond\}$ have as maximal common prefix $\text{pre}(aba, ab\diamond) = ab$, $P(x) = \{\varepsilon, a, ab, aba, ab\diamond\}$ and $S(x) = \{\varepsilon, a, \diamond, ba, b\diamond, aba, ab\diamond\}$.

For a subset X of $A_\diamond^*$ and an integer $i \geq 0$, the set

$$\{w_0 w_1 \cdots w_{i-1} \mid w_0, \ldots, w_{i-1} \in X\}$$

is denoted by X^i. The submonoid of $A_\diamond^*$ generated by X will be denoted by X^* where $X^* = \bigcup_{i \geq 0} X^i$ and $X^0 = \{\varepsilon\}$. The subsemigroup of $A_\diamond^*$ generated by X is denoted by X^+ where $X^+ = \bigcup_{i > 0} X^i$. By definition, each partial word w in X^* admits at least one factorization $(w_0, w_1, \ldots, w_{i-1})$ whose elements are all in X. Such a factorization is called an X-*factorization.*

A *(strong) period* of a partial word w over A is a positive integer p such that $w(i) = w(j)$ whenever $i, j \in D(w)$ and $i \equiv j \bmod p$. In such a case, we call w (strongly) p-periodic. Similarly, a *weak period* of w is a positive integer p such that $w(i) = w(i+p)$ whenever $i, i+p \in D(w)$. In such a case, we call w weakly p-periodic. The partial word $abb\diamond bbcbb$ is weakly 3-periodic but it is not 3-periodic. The latter shows a difference between partial words and full words since every weakly p-periodic full word is p-periodic. Another difference worth noting is the fact that, even if the length of a partial word w is a multiple of a weak period of it, w is not necessarily a power of a shorter partial word. The minimum period of w is denoted by $p(w)$, and the minimum weak period by $p'(w)$.

For a partial word w, positive integer p and integer $0 \leq i < p$, define

$$w_{i,p} = w(i)w(i+p)w(i+2p)...w(i+jp)$$

where j is the largest non-negative integer such that $i + jp < |w|$. Then w is (strongly) p-periodic if and only if $w_{i,p}$ is (strongly) 1-periodic for all $0 \leq i < p$, and w is weakly p-periodic if and only if $w_{i,p}$ is weakly 1-periodic for all $0 \leq i < p$. Strongly 1-periodic partial words as well as the full factors, that is factors that are full words, of weakly 1-periodic partial words are over a singleton alphabet.

We say that two partial words u and v are *compatible*, denoted by $u \uparrow v$, if there exists a partial word w such that $u \subset w$ and $v \subset w$. We note that $u \uparrow v$ implies $v \uparrow u$.

Using the concatenation one gets the following straightforward rules:

Lemma 1. [BB99]

Let $u, v, w, x, y \in A_\diamond^*$. The following hold:

Multiplication: *If $u \uparrow v$ and $x \uparrow y$, then $ux \uparrow vy$.*

Simplification: *If $ux \uparrow vy$ and $|u| = |v|$, then $u \uparrow v$ and $x \uparrow y$.*

Weakening: *If $u \uparrow v$ and $w \subset u$, then $w \uparrow v$.*

The following result extends to partial words the *equidivisibility property* of words, or, *Lévi's lemma.*

Lemma 2. [BB99]

Let $u, v, x, y \in A_\diamond^*$ be such that $ux \uparrow vy$.

- *If $|u| \geq |v|$, then there exist $w, z \in A_\diamond^*$ such that $u = wz$, $v \uparrow w$, and $y \uparrow zx$.*

- *If $|u| \leq |v|$, then there exist $w, z \in A_\diamond^*$ such that $v = wz$, $u \uparrow w$, and $x \uparrow zy$.*

A partial word u is *primitive* if there exists no word v such that $u \subset v^n$ with $n \geq 2$. Note that the empty word is not primitive, and that if v is primitive and $v \subset u$, then u is primitive as well. If u is a non-empty partial word, then there exists a primitive word v and a positive integer n such that $u \subset v^n$. Uniqueness does not hold for partial words. For example, if $u = a\diamond$, then $u \subset a^2$ and $u \subset ab$ for distinct letters a, b.

For $u, v \in A_\diamond^*$, if there exists a primitive word x such that $uv \subset x^n$ for some positive integer n, then there exists a primitive word y such that $vu \subset y^n$. Consequently, if uv is primitive, then vu is primitive.

For partial words we have two distinct types of borders. If $w = x_1 v = u x_2$, where $x_1 \subset x$ and $x_2 \subset x$, we say that x is a *simple* border if $|x| \leq |v|$, and a *nonsimple (overlapping)* border, otherwise. A bordered partial word w is called *simply bordered* if a minimal border x exists such that $|w| \geq |2x|$. For example, the word $a\diamond b\diamond bb$ has the simple and minimal border abb and the nonsimple border $abbbb$, and thus it is simply bordered.

The notion of *one-way infinite partial word* extends the notion of partial word in a natural way. A one-way infinite partial word over the alphabet A is a partial function $w : \mathbb{N} \to A$. As in the case of finite partial words, for $i \in \mathbb{N}$, such that $w(i)$ is defined, we say that i belongs to the domain of w ($i \in D(w)$). Otherwise we say that i belongs to the set of holes of w ($i \in H(w)$). The infinite partial words that do not contain any hole are called infinite full words. One-way infinite partial words are seen as elements of $A_\diamond^\omega$ and an infinite partial word is usually depicted as $w = a_0 a_1 a_2 \cdots$, with $a_i \in A_\diamond$.

For a more detailed presentation of these aspects, as well as for the proofs of the results cited here, we refer to [BS08].

Chapter 3

Freeness for Partial Words

In [Thu06], Thue gives a construction of an infinite square-free word over a three-letter alphabet using a "uniform tag system" and an infinite square-free word over a four-letter alphabet using iterative morphisms (see [Ber92]). Later on, in [Thu12], the author introduces the so-called *Thue-Morse word*, and shows that all two-sided infinite overlap-free words are derived from this sequence. Moreover, he gives a complete description of circular overap-free words and gives an iterative morphism over three letters that constructs an infinite square-free word.

Because Thue's results were published in obscure Norwegian journals, they remained unknown for a long time and were independently rediscovered by S. E. Arshon in 1937 and by M. Morse and G. Hedlund from 1938 to 1944. For more information see [BEM79]. It is interesting to note, as the authors say, that while Thue saw analogies with the theory of Diophantine equations, the work of Morse and Hedlund was grounded in the investigation of flows on surfaces of negative curvature, and Arshon's work was done in order to answer a question which A.Y. Khinchin posed in 1933.

Since A^ω is an uncountable set, hence, there is no effective way to define the elements of this set, the focus is on infinite words that can be described through some precise method. The most frequently used method to define infinite words (as stated in the survey [KL06]) is that of *iterating a morphism* that is prolongable. In the following we present an example of an infinite word defined using this method.

Example 1. *(The Thue-Morse word)*

Let $\phi : \{a, b\}^* \rightarrow \{a, b\}^*$ be a morphism defined by $\phi(a) = ab$ and

$\phi(b) = ba$. We define $\tau_0 = a$ and $\tau_i = \phi^i(a)$. Remark that $\tau_{i+1} = \phi(\tau_i)$ and that $\tau_{i+1} = \tau_i\overline{\tau}_i$, where $\overline{x}$ is the word obtained from x by replacing each occurrence of a with b and each occurrence of b with a. We define the Thue-Morse word by:

$$\tau = \lim_{i\to\infty} \tau_i = \lim_{i\to\infty} \phi^i(a)$$

The Thue-Morse word τ is a fixed point for the morphism ϕ, i.e., $\phi(\tau) = \tau$.

We say that an (infinite) word w is *k-free* if there does not exist a word x such that x^k is a factor of w. Also, a (infinite) word is called *overlap-free* if it does not contain any two instances of the same factor overlapping. In other words it does not contain a factor of the form $ayaya$ with $a \in A$. For more details see [Lot97]. It is clear that any overlap-free word w is $k-$free, for $k \geq 3$, and, as well, any 2-free word is overlap-free. For simplicity, a 2-free word is said to be *square-free* , and a 3-free word is said to be *cube-free*.

A result that will be used throughout the chapter regarding the *Thue Morse infinite word* τ defined in Example 1, is the following:

Theorem 1. (Thue Theorem)[Ber92, Thu06, Thu12] *The Thue-Morse word τ is overlap-free.*

Remark 1. *As a consequence of Theorem 1, it follows that the Thue-Morse word τ is k-free for all $k \geq 3$.*

In the same papers A. Thue also gave a word that is square-free. The word was defined over a three-letter alphabet, using a morphism whose sum of the length of the images was 18. It has been shown by Carpi in [Car83] that this is actually the optimal bound for a morphism preserving square-freeness. Later on, however, Hall [Hal64] manages to give a simpler morphism that generates a fixed point word that is square-free. Let σ be the fixed point of the morphism $\psi : \{a, b, c\}^* \to \{a, b, c\}^*$ with $\psi(a) = abc$, $\psi(b) = ac$ and $\psi(c) = b$.

Theorem 2. [Lot97] *The word σ is square-free.*

A partial word $w \in A_\diamond^*$ is said to be *k-free* if for any non-empty factor $x_0 \cdots x_{k-1}$ of w, there does not exist a partial word u, such that $x_i \subset u$ for all $i \in \{0, \ldots, k-1\}$. Now looking at the definition of *overlap-freeness* we see that things are a bit different from the full word case. We notice that, for a word of the form $a_0w_0a_1w_1a_2$, we say that it is a *strong overlap* if $w_0 \uparrow w_1$ and a_0, a_1 and a_2 are pairwise compatible, and that it is a *weak overlap* if $w_0 \uparrow w_1$ and $a_0 \uparrow a_1$ and $a_1 \uparrow a_2$. Please note that the notions of strongly and weakly overlap-freeness come naturally from the definitions of strong and weak periods. Moreover, we see that a word is a strong overlap if it is compatible with a full word of the form $awawa$ and a weak overlap if both overlapping factors are both compatible with a word of the form awb, where a and b are letters and w is a word over the alphabet A.

Remark 2. *It is rather simple to note that any partial word w over A, with $|w| \geq 2$ and $H(w) \neq \emptyset$, cannot be square-free, since it contains at least one of the factors $a\diamond$ or $\diamond a$, where $a \in A_\diamond$. Also, if w is n-free, then w is m-free for $m \geq n$.*

Remark 3. *Inserting a hole is defined as replacing a letter with a hole in a fixed position of a word (the length of the word remains the same). When we introduce holes into arbitrary positions of a word, we impose the restriction that every two holes must have at least two non-hole letters between them.*

Without imposing this restriction, it would always be possible to obtain a repetition of order k of the form $\diamond^k$ or $\diamond^{k-i}a\diamond^{i-1}$, where a is a letter of the alphabet and k and i are integers with $i < k$.

Remark 4. *Since for all words of the form $a\diamond b$ we have $a\diamond \uparrow \diamond b$, all holes present after the first position determine a weak overlap. Hence, we will say that a word is a weak overlap, if it does not contain any overlap, except the trivial $a\diamond b$, for some letters a, b.*

For a more detailed presentation of these aspects, as well as for the proofs of the results cited here, we refer to [Ber92, BS08, CK97, KL06, Lot97].

3.1 Square-freeness

To generalize Thue's result, we wish to find a square-free partial word with infinitely many holes, and an infinite full word that remains square-free even after replacing an arbitrary selection of letters with holes. Unfortunately, every partial word containing at least one hole and having length at least two contains a square (as stated in Remark 2, either $a\diamond$ or $\diamond a$ cannot be avoided, where a denotes a letter from our alphabet). Furthermore, since it is obvious that if we replace with holes $2n$ consecutive letters in a full word, the corresponding factor of the resulting partial word will be a square, the restriction defined in Remark 3 must be applied.

Motivated by these observations, we call a word *trivial square* if it is of the form $a\diamond$, $\diamond a$ or $\diamond ab\diamond$, for any letters a, b. Any other square is called *non-trivial*, [BSMS09]. The study of squares of different lengths is not new. Following the results of Thue, see [Ber92, Thu06], Dekking studies in [Dek76] the properties of infinite binary sequences. One of his first results in the paper from 1976, states the following.

Theorem 3. [Dek76] *There exists an infinite binary sequence free from triple repetitions and free from repetitions of length* 4 *or greater.*

A second result talks about the number of words avoiding repetitions of length greater than 3.

Theorem 4. [Dek76] *Binary sequences that are free from triple repetitions, and contain no repetitions of length* 3 *or greater, are finite.*

The first result was improved after almost 30 years, in [RSW05] where a somehow simpler construction for an infinite binary word avoiding both cubes xxx and squares yy with $|y| \geq 4$ is given. In the same paper the authors also give a simpler construction for the implementation of an infinite binary word avoiding all squares except 0^2, 1^2, and $(01)^2$ a result originally from Fraenkel and Simpson [FS95]. This is done using a series of morphisms over two and four letters.

With these restrictions, the study of square-free partial words becomes much more subtle, as we will see in the following. First we find an infinite partial word containing infinitely many holes and avoiding all squares but the ones of the form $a\diamond$ and $\diamond a$, and later, we find an infinite full word that remains non-trivial square-free even after replacing an arbitrary selection of letters with holes. As a visual aid throughout this section, we will underline the first and $(n+1)^{th}$ symbol in a word that is a square and has length $2n$. These results represent joint work with Francine Blanchet-Sadri and Geoffrey Scott [BSMS09].

3.1.1 Square-free partial words

Let us first see if it is possible to have infinite words that do not contain squares.

Theorem 5. *There exist infinitely many infinite partial words with infinitely many holes over a three-letter alphabet that do not contain any squares other than squares of the form $\diamond a$ or $a\diamond$.*

Proof. Let σ be the fixed point defined in Theorem 2. We can define the word σ' by applying a morphism δ on the word σ that replaces a with $\psi^4(a)'$, b with $\psi^4(b)$, and c with $\psi^4(c)$ where

$$\psi^4(a) = abcacbabcbac\underline{a}bcacbacabcb$$

and

$$\psi^4(a)' = abcacbabcbac\underline{\diamond}bcacbacabcb$$

Here the a representing the $13th$ symbol of $\psi^4(a)$ is changed into a $\diamond$. Set $\sigma = a_0a_1\ldots$, and let $\sigma' = b_0b_1\ldots$ be the partial word $\delta(\sigma)$. We claim that σ' satisfies the desired property.

First, σ' contains no squares of length 4, other than $c\diamond$ and $\diamond b$. To see this, it is enough to check the word $bac\diamond bca$. Now, assume that σ' contains a non-trivial square. Then there exist integers $i \geq 0, k > 0$ such that $b_ib_{i+1}\ldots b_{i+k-1} \uparrow b_{i+k}b_{i+k+1}\ldots b_{i+2k-1}$. Since σ itself is square-free, the square in σ' must contain a hole. If $k < 7$, then the square factor is also a factor of $\psi^4(a)'$. It can be checked explicitly that $\psi^4(a)'$ is non-trivial square-free. Therefore, $k \geq 7$. We proceed by showing that if $b_{i+j} = \diamond$, then $b_{i+k+j} \in \{\diamond, a\}$ and that if $b_{i+k+j} = \diamond$, then $b_{i+j} \in \{\diamond, a\}$. This will show that every hole in σ' can be filled with the letter a while preserving the square factor in the word. However, the result of filling all holes in σ' with the letter a is the square-free word σ, so we will arrive at a contradiction. Since both implications are proved using the same logic, we will only show that if $b_{i+j} = \diamond$, then $b_{i+k+j} \in \{\diamond, a\}$. Let us consider the possibilities where the hole can appear. Suppose $b_{i+j} = \diamond$.

- If $0 \leq j < k-2$, then $b_{i+j}\ldots b_{i+j+2} = \diamond bc$. It is easy to check, by looking at the description of ψ, that the only factors of σ' compatible with $\diamond bc$ are $\diamond bc$ and abc. Since $b_{i+k+j}\cdots b_{i+k+j+2}$ must be compatible with $\diamond bc$, it follows that $b_{i+k+j} \in \{\diamond, a\}$.

- If $5 \leq j < k$, then $b_{i+j-5}\ldots b_{i+j} = bcbac\diamond$. It is easy to check that the only factors of σ' compatible with $bcbac\diamond$ are $bcbac\diamond$ and $bcbaca$. Therefore, $b_{i+k+j} \in \{\diamond, a\}$.

- If $k-2 \leq j < 5$, then $b_{i+j-1} \dots b_{i+j+1} = c\diamond b$. Since the only factors of σ' compatible with $c\diamond b$ are $c\diamond b$ and cab, it follows that $b_{i+k+j} \in \{\diamond, a\}$.

□

Corollary 1. *There exist infinitely many infinite partial words with an arbitrary number of holes over a three-letter alphabet that do not contain any squares other than squares of the form $\diamond a$ or $a\diamond$.*

Proof. If not all a's are replaced by $\psi^4(a)'$ (some could be replaced by $\psi^4(a)$ instead), then we get the result with an arbitrary number of holes. □

3.1.2 Generalization of square-freeness

We now turn our attention to words that remain non-trivial square-free after replacing an arbitrary collection of letters with holes.

We begin by stating an obvious remark that will be used several times throughout this section.

Remark 5. *Let $t_0 = a_0a_1a_2$ and $t_1 = b_0b_1b_2$ be full words. It is possible to insert holes into t_0 and t_1 such that the resulting partial words are compatible if and only if there exists i such that $a_i = b_i$ (or the letters in position i of t_0 and t_1 are equal). This is due to Remark 3 that states that every two holes must have at least two non-hole symbols between them.*

To insert holes in $t_0 = abb$ and $t_1 = acc$ in order to make them compatible (with the convention in Remark 3), one can create $t_0' = a\diamond b$ and $t_1' = ac\diamond$ respectively. However, this is impossible when $t_0 = abb$ and $t_1 = bcc$.

Proposition 1. *Let t be a full word over an alphabet A. If every factor of length n of t contains n distinct elements of A, then it is impossible to insert holes into t such that the resulting partial word contains a non-trivial square w_0w_1 with $w_0 \uparrow w_1$ and $|w_0| = |w_1| < n$.*

Proof. All positions i and $i+k$ have a different letter for $3 \leq k < n$ and thus the position i or $i+k$ must gain a hole. So there must be two holes at distance 1 or 2. □

Theorem 6. *There exists an infinite word over an eight-letter alphabet that remains non-trivial square-free after an arbitrary insertion of holes.*

Proof. Recall that σ is square-free according to Theorem 2.

We construct the desired word t by applying a uniform morphism δ on the word σ that replaces

- a with $defghijk$,
- b with $deghfkij$, and
- c with $dehfgjki$.

We claim that t satisfies our desired properties.

Assume that it is possible to change a selection of positions in t to holes such that the resulting partial word t' contains a non-trivial square. It is clear that t' has no factors that are non-trivial squares of length less or equal to 4. Therefore, we can restrict our attention to factors of the form w_0w_1 with $w_0 \uparrow w_1$ and $|w_0| = |w_1| \geq 3$. That is, if $t = a_0a_1a_2\ldots$ and $t' = b_0b_1b_2\ldots$, then there exist $i \geq 0$ and $k \geq 3$ such that

$$b_ib_{i+1}b_{i+2}\ldots b_{i+k-1} \uparrow b_{i+k}b_{i+k+1}b_{i+k+2}\ldots b_{i+2k-1}$$

There are two cases to be analyzed:

Case 1. $k \equiv 0 \bmod 8$

Setting $k = 8(m+1)$, note that $a_ia_{i+1}a_{i+2}\ldots a_{i+k-1}$ is of the form

$$w_{00}\delta(c_0)\delta(c_1)\ldots\delta(c_{m-1})w_{01}$$

and $a_{i+k}a_{i+k+1}a_{i+k+2}\ldots a_{i+2k-1}$ is of the form

$$w_{10}\delta(c_{m+1})\delta(c_{m+2})\ldots\delta(c_{2m})w_{11}$$

with $w_{01}w_{10} = \delta(c_m)$, $|w_{pr}| = |w_{qr}|$ with $w_{pr}, w_{qr} \in \{d,e,f,g,h,i,j,k\}^*$ and $c_l \in \{a,b,c\}$ for all $p,q,r \in \{0,1\}$ and $l \in \{0,1,\ldots,2m\}$.

Also note that if $c_p \neq c_{m+p+1}$ for any $0 \leq p < m$, then it is impossible to insert holes into $\delta(c_p)$ and $\delta(c_{m+p+1})$ such that the resulting partial words are compatible. Therefore, $c_p = c_{m+p+1}$ for all $0 \leq p < m$.

If $|w_{01}| \geq 5$, then by Remark 5, w_{11} must be a prefix of $\delta(c_m)$. Hence,

$$\underline{c_0}c_1\ldots c_{m-1}c_m\underline{c_{m+1}}c_{m+2}\ldots c_{2m}c_m$$

is a factor of σ. Since σ is square-free and $c_p = c_{m+p+1}$ for $0 \leq p < m$, this is a contradiction. If $|w_{01}| < 5$, then $|w_{00}| \geq 4$ and it follows that w_{00} and w_{10} are suffixes of $\delta(c_m)$. Then

$$\underline{c_m} c_0 c_1 \ldots c_{m-1} \underline{c_m} c_{m+1} c_{m+2} \ldots c_{2m}$$

is a factor of σ. Since σ is square-free, this is a contradiction.

Case 2. $k \not\equiv 0 \bmod 8$

Suppose that $a_{i+l} = a_{i+k+l} = d$ for some $0 \leq l < k-4$. Then the words

$$a_{i+l+1} \ldots a_{i+l+4} \text{ and } a_{i+k+l+1} \ldots a_{i+k+l+4}$$

can only be $efgh$, $eghf$, or $ehfg$. Since $k \not\equiv 0 \bmod 8$, it follows that $a_{i+l+1} \ldots a_{i+l+4}$ is different from $a_{i+k+l+1} \ldots a_{i+k+l+4}$. However, if we select any two different strings from $efgh$, $eghf$ and $ehfg$, it is easy to see that they cannot be made compatible through the introduction of holes. Therefore, it is clear that $b_{i+l+1} \ldots b_{i+l+4}$ is not compatible with $b_{i+k+l+1} \ldots b_{i+k+l+4}$. This contradicts with the assumption that

$$b_i b_{i+1} b_{i+2} \ldots b_{i+k-1} \uparrow b_{i+k} b_{i+k+1} b_{i+k+2} \ldots b_{i+2k-1}$$

Therefore, there is no l satisfying $0 \leq l < k-4$ such that $a_{i+l} = a_{i+k+l} = d$. In fact, this argument remains true if we replace the letter d with any letter in the set $\{d, e, f, g, h, i, j, k\}$. Thus, there exists no l satisfying $0 \leq l < k-4$ such that $a_{i+l} = a_{i+k+l}$. By Remark 5, it follows that $a_{i+l} = a_{i+k+l}$ for some $0 \leq l < 3$. If $k \geq 7$, this same l would satisfy $0 \leq l < k-4$. Therefore, $k < 7$.

We observe that every factor of length six of t contains no repeated letters. By Proposition 1, it follows that $k = 6$. Every factor of length 12 in t is contained in $\delta(c_1)\delta(c_2)\delta(c_3)$ for some $c_i \in \{a, b, c\}$. It is a tedious yet finite process to check that it is impossible to insert holes into any of the above factors to create a square. This can be done using a computer program.

Since all cases lead to contradiction we conclude that t satisfies the desired properties. □

Of course, it is natural to ask whether such a word can be constructed over a smaller alphabet. This question is intimately related to the study of full words of the form v_0awav_1, where $a \in A$ and $v_0, v_1, w \in A^*$.

Proposition 2. *Let* $t = v_0awav_1$ *be a full word over the alphabet* A*, where* $a \in A$ *and* $v_i, w \in A^*$*. If any of the following hold, then it is possible to insert holes into* t *so that the resulting partial word contains a non-trivial square:*

1. $|w| = 2$ *and* $|t| \geq 6$,
2. $|w| = 3$, $|t| \geq 8$ *and* $|v_i| \geq 1$,
3. $|w| = 4$ *and* $|v_i| \geq 2$,
4. $|w| = 5$, $|v_i| \geq 4$ *and* $|A| \leq 7$.

Proof. Let $b_i \in A$. For Statement 1, if t has factors of the form $ab_0b_1ab_2b_3$, $b_0ab_1b_2ab_3$, or $b_0b_1ab_2b_3a$, then by replacing b_0 and b_3 with holes into t we get partial words containing factors that are squares of the form $\underline{a}\diamond b_1\underline{a}b_2\diamond$, $\underline{\diamond}ab_1\underline{b_2}a\diamond$, or $\underline{\diamond}b_1a\underline{b_2}\diamond a$ respectively.

For Statement 2, if t has a factor $b_0ab_1b_2b_3ab_4b_5$ or $b_0b_1ab_2b_3b_4ab_5$, we can insert holes into t such that the resulting partial word has square factors $\underline{\diamond}ab_1\diamond\underline{b_3}a\diamond b_5$ or $\underline{b_0}\diamond ab_2\underline{\diamond}b_4a\diamond$ respectively.

For Statement 3, if t has a factor of the form $b_0b_1ab_2b_3b_4b_5ab_6b_7$, we can insert holes into t such that the resulting partial word has the square factor $\underline{\diamond}b_1a\diamond b_3\underline{b_4}\diamond ab_6\diamond$.

For Statement 4, it is obvious that $|t| \geq 15$. Hence, if t has a factor of the form

$$b_0b_1b_2b_3ab_4b_5b_6b_7b_8ab_9b_{10}b_{11}b_{12}$$

then we argue as follows. If $b_i = b_j$ for any $4 \leq i < j < 9$, then by the previous three statements we can insert holes into the factor such that the resulting partial word contains a non-trivial square (note that if $j = i+1$ or $j = i+2$, we could create the non-trivial squares $\underline{b_ib_i}$ or $\underline{b_i}\diamond\underline{b_i}b_k$, for some letter b_k). For the same reason, $b_i \neq a$ for $4 \leq i < 9$. Therefore, we assume that the letters b_i for $4 \leq i < 9$ are pairwise non-equal and distinct from a. Similarly, we can assume that $b_9 \neq b_i$ for $5 \leq i < 9$ and $b_9 \neq a$. If $b_9 = b_4$, then we can insert holes into the factor such that the resulting partial word contains the square $\underline{\diamond}b_3ab_4\diamond b_6\underline{b_7}\diamond ab_9b_{10}\diamond$. Thus, the letters b_i for $4 \leq i < 10$ are pairwise non-equal and distinct from a. Using the same logic, the letters

b_i for $3 \leq i < 9$ are pairwise non-equal and distinct from a. Since $\|A\| \leq 7$, we must have $b_3 = b_9$.

Next, b_{10} must be distinct from a and b_i for $6 \leq i < 10$, so either $b_{10} = b_5$ or $b_{10} = b_4$. If $b_{10} = b_5$, we can insert holes into the factor such that the resulting partial word contains the non-trivial square $\underline{\diamond}b_3a\diamond b_5b_6\underline{b_7}\diamond ab_9b_{10}\diamond$. Therefore, $b_{10} = b_4$. Using the same logic, we find that $b_2 = b_8$.

Finally, we can insert holes into our factor such that the obtained partial word contains the non-trivial square $\underline{\diamond}b_2b_3\diamond b_4b_5\diamond\underline{b_7}b_8\diamond b_9b_{10}\diamond b_{12}$. □

Corollary 2. *Let t be an infinite word over an alphabet A such that any partial word obtained by inserting holes in t is square-free. Then $\|A\| \geq 8$.*

Proof. Let t be an infinite word over the alphabet $A = \{a_0, a_1, \ldots, a_6\}$, where $a_i \neq a_j$ for all $0 \leq i < j \leq 6$. If t has a factor of the form v_0awav_1, where $a \in A$, $v_i, w \in A^*$, $2 \leq |w| \leq 5$ and $|v_i| \geq 4$, then according to the previous proposition it is possible to introduce holes into t to create square factors (note that if $|w| = 1$, then we can replace w with $\diamond$ to create the non-trivial square $\underline{a}\diamond\underline{a}b$). To avoid this, t must have a factor of the form

$$\underline{a_0}a_1a_2a_3a_4a_5a_6\underline{a_0}a_1a_2a_3a_4a_5a_6$$

up to an isomorphism between the letters. This implies that t contains squares that will certainly be preserved when holes are added. Therefore, at least eight letters are needed to create an infinite word satisfying our conditions. □

3.2 Overlap-freeness

As previously mentioned, Axel Thue was first to investigate avoidable regularities, especially words without overlapping factors and square-free words. His two papers [Thu06, Thu12] on this topic contain the definitions of the words τ, see example 1, and of a word similar to σ, see [AS99].

In order to prove the overlap-freeness, two lemmas were used.

Lemma 3. [Lot97] *Let $X = \{ab, ba\}$; if $x \in X^*$, then $axa \notin X^*$ and $bxb \notin X^*$.*

Lemma 4. [Lot97] *Let $w \in A^+$. If w has no overlapping factor, then $\phi(w)$ has no overlapping factor (ϕ is the morphism used in the Example 1).*

From Lemma 3 and Lemma 4, Theorem 1 regarding the overlap-freeness comes naturally.

Besides this result, in these papers Thue also gives a description of circular overlap-free words and mentions the problem of counting the number of overlap-free words over two letters.

Theorem 7. [Thu12] *Every circular overlap-free word over the two-letter alphabet $A = \{a, b\}$ is of the form $\phi^n(ab)$, $\phi^n(aab)$ or $\phi^n(abb)$ for some $n \geq 0$.*

In this section, consisting of results from [BSMS09], we will extend the concept of overlap-freeness to partial words. We use the standard definition of overlap-freeness given in the preliminaries, but we still adhere to the restriction described in Remark 3 when replacing an arbitrary selection of letters in a word with holes. First we prove the existence of an overlap-free partial word with infinitely many holes, and later on we find an infinite full word that remains overlap-free even after replacing an arbitrary selection of letters with holes. As a visual aid, we will underline the a_i's of the overlapping factor $\underline{a_0}w_0\underline{a_1}w_1\underline{a_2}$ to distinguish an overlap present in a sequence of letters. Moreover, let us note that our definition of weak overlap is actually a generalization of the overlap definitions used in [BSMS09] and [HHKS09], since here a factor is considered to be an overlap of length $2p+1$ if it has a weak period p, while in the previous papers, the factor had to have a strong period p.

3.2.1 Overlap-free partial words

Let us first look at a lower bound for the size of an alphabet necessary for constructing this type of words.

Lemma 5. *There exists an infinity of overlap-free infinite binary partial words containing one hole.*

Proof. Recall that the Thue-Morse word is overlap-free. We claim that the Thue-Morse word preceded by a hole, $\diamond\tau$, is also overlap-free. Let ϕ be the Thue-Morse morphism. Because τ is overlap-free, any overlap occurring in $\diamond\tau$ must contain the hole. It suffices, therefore, to show that $\diamond\phi^i(a)$ is overlap-free for any positive i. Note that

$$\phi^{i+3}(a) = \phi^i(a)\ \overline{\phi^i(a)}\ \overline{\phi^i(a)}\ \phi^i(a)\ \overline{\phi^i(a)}\ \phi^i(a)\ \phi^i(a)\ \overline{\phi^i(a)}$$

contains a copy of both $a\phi^i(a)$ and $b\phi^i(a)$ (due to the factors $\overline{\phi^i(a)}\phi^i(a)$ and $\phi^i(a)\phi^i(a)$). Since the Thue-Morse word is overlap-free, $\phi^{i+3}(a)$ is as well. Therefore, neither $a\phi^i(a)$ nor $b\phi^i(a)$ contain an overlap. This implies that $\diamond\phi^i(a)$ is overlap-free. Furthermore, adding a hole in front of an iteration's conjugate will not give us an overlap either. In order to obtain new words that are overlap-free, it is enough to add a hole in front of words obtained after taking out the prefix of length $|\phi^{i+3}(a)|$, from the Thue-Morse word. It is also clear that, $\diamond\overline{\tau}$ is overlap-free as well. □

Remark 6. *Over a binary alphabet all words of length greater than six with a hole in the third position contain an overlap.*

To see this, note that if the partial word has a factor of the form $a\diamond a$, $aa\diamond$ or $\diamond aa$, then it clearly contains an overlap. Therefore, we can assume that any overlap-free binary word with a hole in the third position has a prefix of the form $ab\diamond ab$. If this factor is followed by aa, then the word contains the overlap $\underline{a}b\diamond\underline{a}ba\underline{a}$. Similarly, if the factor is followed by ab, ba, or bb, it will contain $\underline{\diamond}a\underline{b}a\underline{b}$, $\underline{a}b\diamond\underline{a}bb\underline{a}$, or bbb respectively.

Proposition 3. *There is no infinite overlap-free binary partial word with more than one hole.*

Proof. To see this, note that by Remark 6, an infinite overlap-free binary partial word cannot contain a hole after the second position. However, it cannot contain holes in both the first and second positions, as an overlap of the form $\diamond\diamond a$ would clearly appear. Thus, only one hole is allowed. □

Please note that all previous results hold for both weak and strong overlaps, since a strong overlap implies a weak overlap. We conclude that in order to get a word that has an infinity of holes without failing to be overlap-free, we need at least a three-letter alphabet. We will now prove that the non-trivial square-free word given in Proposition 5 is also overlap-free.

Proposition 4. *There are infinitely many overlap-free infinite partial words with an arbitrary number of holes over a three-letter alphabet.*

Proof. In Theorem 5 we showed that the word σ' constructed there does not contain any squares other than squares of the form $\diamond b$ or $c\diamond$. Because σ is square-free and hence overlap-free, any overlap in σ' must contain a hole. So it remains only to show that σ' contains no overlaps of the form $a_0a_1a_2$ with

$a_0, a_1, a_2 \subset b$ or $a_0, a_1, a_2 \subset c$. Note that if $a_1 = \diamond$, this is a trivial overlap, hence the result will hold for both weak and strong overlaps. However, any such overlapping factor is so small that it would be contained in $\psi^4(a)'$. It is easy to check that $\psi^4(a)'$ does not contain any such overlapping factor. □

The concept of overlap-freeness is also investigated in [HHKS09] (only our notion of strong overlap is considered in the paper). Here a new concept of overlap is introduced:

Definition 1. *A partial word w is k-overlap-free if it is cube-free and, for any factor v of w, there is no overlap $xyxyx$ such that $v \subset xyxyx$ and $|x| = k$.*

Note that this means that a k-overlap-free partial word does not contain factors of the form $xyx'y'x''$ with x, x', x'', and respectively y, y' pairwise compatible non-empty partial words, and $|x| = k$. It is obvious that any k-overlap-free word is also k'-overlap-free for $k' \geq k$, and that a word is 1-overlap-free if and only if it is overlap-free. For example, the word $ab\diamond abaabba$ is 3-overlap-free but not 2-overlap-free since it contains the factor $\underline{ab}\diamond\underline{ab}a\underline{ab}$. Let us now recall the main result regarding this concept.

Theorem 8. [HHKS09] *There exist infinitely many 2-overlap-free binary partial words containing infinitely many holes.*

Since over a binary alphabet no overlap-free (1-overlap-free) words exist, this result is actually optimal.

3.2.2 Generalization of overlap-freeness

In the previous section, we gave infinite words that are (weakly) overlap-free even after carefully selected letters in the word were changed to holes. Now we will give overlap-free words that remain (weakly) overlap-free even after an arbitrary selection of their letters are changed to holes.

Proposition 5. *There is no infinite word over a four-letter alphabet that remains overlap-free after an arbitrary selection of its positions are changed to holes.*

Proof. Assume that such a word t exists over the four-letter alphabet A. Clearly, it contains no factors of the form bba or bab where $a, b \in A$, since holes could be introduced to form the overlap factors $bb\diamond$ and $b\diamond b$ respectively. If the word contains no factor of the form ba_0a_1b where $a_i \in A$ for all i, then every factor of t of length four contains no repeated letters. This would imply that t is of the form

$$\dots \underline{a_0}a_1a_2a_3\underline{a_0}a_1a_2a_3\underline{a_0}a_1a_2a_3 \dots$$

Therefore, we can assume that t has a factor of the form $a_0a_1a_2ba_3a_4ba_5a_6$, where $b \neq a_i$, for all $i > 0$. If $a_5 = a_3$, then $\underline{\diamond}ba_3\underline{a_4}ba_5\underline{\diamond}$ is an overlap (symmetrically, $a_4 \neq a_2$ to avoid the overlap $\underline{\diamond}a_2b\underline{\diamond}a_4b\underline{a_5}$). Therefore, $a_5 \neq a_3$. Similarly, we must have $a_5 \neq a_4$, $a_5 \neq a_6$, $a_4 \neq a_6$ and $a_4 \neq a_3$ to avoid the overlaps $a_4\diamond a_5$, $\diamond a_5a_6$, $\underline{\diamond}ba_3\underline{a_4}b\diamond\underline{a_6}$, and $\diamond a_3a_4$ respectively. Since b, a_4, a_5 and a_6 are pairwise non-equal, they must be four different letters. Since A is a four-letter alphabet and a_3 is distinct from b, a_4 and a_5, it follows that $a_3 = a_6$.

We use similar logic to determine that $a_1 = a_4$. We arrive at our desired contradiction by introducing holes to get the overlap $\underline{\diamond}a_1a_2\diamond\underline{a_3}a_4\diamond a_5\underline{a_6}$ □

This proposition gives us a lower bound of five for a minimum alphabet size necessary to construct a word that is strongly overlap-free after an arbitrary selection of its letters are changed to holes. Since strong overlap-freeness implies weak overlap-freeness, the result holds for this other case as well. The rest of this section contains results obtained together with Francine Blanchet-Sadri, Abraham Rashin and Elara Willett [BSMRW09].

First let us state a result regarding factors of a certain size of a prolongable morphism.

Lemma 6. *If in a prolongable morphism the set of factors of length n of the i-th iteration equals the set of factors of length n of the $i+1$-th iteration, it must be that it also equals the set of factors of length n of the fixed point of the morphism.*

Let us note that our definition of overlap is actually a generalization of the overlap definitions used in [BSMS09] and [HHKS09], since here a factor is considered to be an overlap of length $2p+1$ if it has a weak period p, while in the previous papers, the factor had to have a strong period p. Let us define a morphism $\gamma : \{a, b, c, d\}^* \to \{a, b, c, d\}^*$ with $\gamma(a) = ad$, $\gamma(b) = ac$, $\gamma(c) = cb$, and $\gamma(d) = ca$. Since a is a prefix of $\gamma(a)$, γ is prolongable. Thus, we define a fixed point of γ, $\Gamma = \lim_{i\to\infty} \gamma^i(a)$. Let us consider some properties of w.

Remark 7. *Both $\gamma^3(a) = adcacbad$ and $\gamma^4(a) = adcacbadcbacadca$ have only ac, ad, ba, ca, cb and dc as their length 2 factors. Thus, by Lemma 6, these are the only length 2 factors of Γ.*

Lemma 7. *The infinite full word Γ is square-free.*

Proof. It suffices to show that every $\gamma^n(a)$ is square-free. Clearly $\gamma^0(a) = \varepsilon$ is square-free. Now let $n \geq 0$ and $\gamma^n(a)$ be square-free. Suppose, for contradiction, that $\gamma^{n+1}(a)$ has a square factor of length $2p$ starting at position i. Since the letters b and d appear only at odd positions of $\gamma^{n+1}(a)$, hence, even distance apart, if p is odd, the factor would be of the form $\{a, c\}^*$. Since all binary words of length 5 contain squares, it must be that $p = 1$, which is a contradiction according to Remark 7.

Therefore p must be even, say $p = 2q$. If the factor starts at an even position, since $\gamma^{n+1}(a) = \gamma(\gamma^n(a))$ it follows that $\gamma^n(a)$ contains a square, contradiction with the initial assumption. Hence, the factor must start at an odd position. Since, $\gamma(f)$ ends in a different letter for all $f \in \{a, b, c, d\}$, it follows that we will have a factor that is a square starting with position $i - 1$, which is an even position. Following the previous reasoning we again reach a contradiction. □

Now let $\delta : \{a, b, c, d\}^* \to \{f, g, h, i, j\}^*$ be a morphism defined by $\delta(a) = fgifh$, $\delta(b) = fghij$, $\delta(c) = jigjh$, and $\delta(d) = jihgf$. We claim that $\delta(\Gamma)$ is overlap-free after an arbitrary (2-valid) insertion of holes.

Proposition 6. *There are no factors of $\delta(w)$ of length ≤ 21 that can be turned into weak overlaps by insertion of holes.*

Proof. It suffices to check that for all $p \leq 10$, there is no factor of $\delta(w)$ of length $2p + 1$ that contains a 2-valid weakly-p-periodic partial word. For every $p \leq 10$, its set of factors of length $2p + 1$ was computed, and each of these was checked for containment of 2-valid weakly-p-periodic words. We remind that, factors of the form $a\diamond b$ are considered to be trivial overlaps, and not weak overlaps. □

Let us recall that according to Remark 5, for two factors of length three to be compatible after hole insertion, it must be that they have at least two equal symbols.

Lemma 8. *In $\delta(\Gamma)$, any two length seven sequences starting with the same character will contain at least three consecutive mismatches if they are not identical.*

Proof. According to Remark 7 the only length two factors of Γ are ac, ad, ba, ca, cb and dc. We prove the lemma for sequences starting with letter f, the other cases being similar. If a sequence starts with f, then it must be either $fgifhji$, a prefix of both $\delta(ac)$ and $\delta(ad)$, $fghijfg$, prefix of $\delta(ba)$, $fjigjh$, suffix of $\delta(dc)$, or, $fhjigjh$ and $fhjihgf$, suffixes of $\delta(ac)$ and $\delta(ab)$. It is easy to check that each two of these blocks contain three consecutive mismatches once aligned. □

Proposition 7. *No factor of $\delta(w)$ of length $2p+1 > 21$ with p not divisible by 5 can be turned into a weak overlap.*

Proof. Let us assume towards a contradiction that there exists $a_0v_0a_1v_1a_2$, a factor that can be transformed into an overlap after insertion of holes. Since p is not divisible by 5, it follows that the images of δ in a_0v_0 and a_1v_1 will not be aligned. Let us look at the second position in a_0v_0. If this one aligns with the second position in a_1v_1, then applying Lemma 8 we get a contradiction. If the two positions do not match, following Remark 5 it must be that either the first or the third positions must match. Using the same technique we get a contradiction in both these cases. Therefore, no factor of $\delta(w)$ of length $2p+1 > 21$ with p not divisible by 5 can be turned into a weak overlap. □

Proposition 8. *No factor of $\delta(w)$ of length $2p+1 > 21$ with p divisible by 5 can be turned into a weak overlap.*

Proof. Let us assume towards a contradiction that a factor $a_0v_0a_1v_1a_2$ can be transformed into an overlap after insertion of holes. Since $|a_0v_0| = 5k$, for some $k > 2$, it follows that the images of δ will be aligned in a_0v_0 and a_1v_1. By looking at the blocks of δ we see that only the images of b and d do not contain 3 consecutive mismatches once aligned. Hence, we will consider the case when these two images are aligned, the other cases being straightforward by Remark 5.

We notice that the only character preceding d in Γ is a, and the only character preceding b is c, while the only character following d in Γ is c, and the only character following b is a. Let us assume that the block determined by $\delta(d)$ ends before the last position in a_iv_i with $i \in \{0, 1\}$. The character following this block is a j, while the one following the block $\delta(b)$ is an f. We notice that this letter together with the last two characters of the block gives us the sequences gfj and ijf, that will not match after a valid insertion of holes, by Remark 5.

If the block $\delta(d)$ starts at a position greater than 5, it follows that it is preceded by $\delta(a)$. Since $\delta(a)$ will align with a block $\delta(c)$ according to the previous observations, by Remark 5 we conclude that a matching is impossible. □

Theorem 9. *The infinite word $\delta(\Gamma)$ over a five-letter alphabet is weakly overlap-free after an arbitrary insertion of holes.*

Proof. This follows directly from Propositions 6, 7, and 8. □

3.3 Cube-freeness

In this section we analyze the concept of cube-freeness for partial words, while we still adhere to the restriction stated in Remark 3. We remind that in order for a partial word to be k-free it must be that the word does not contain a factor of the form $w_0w_1 \cdots w_{n-1}$ such that there exists a partial word u with $w_i \subset u$, for $0 \leq i < n$.

As mentioned before, the property of cube-freeness was for the first time analyzed by Axel Thue in [Thu06, Thu12]. The τ word present in Theorem 1 is overlap-free, hence, it is cube-free. As a visual aid throughout this section, we will underline the first, $(n + 1)$th and $(2n + 1)$th symbols in a word that is a cube and has length $3n$. The results for cubes appeared as joint work with Florin Manea [MM07].

3.3.1 Cube-free partial words

The main result we present here is that for $k \geq 3$ there exist k-free infinite partial words, containing an arbitrary number of holes, over binary alphabets. Moreover, we present an algorithm that, given a natural number n as input, constructs in $\mathcal{O}(n)$ time a cube-free partial word that contains exactly n holes.

Proposition 9. *There exist arbitrarily many cube-free infinite partial words, containing exactly one hole, over a binary alphabet.*

Proof. Assume that we replace an arbitrary position in τ with a hole and let τ' be the infinite partial word that we obtain in this manner. We will prove that for a non-empty factor $w_0 w_1 w_2$ of τ', and a partial word w such that $w_i \subset w$, for all $i \in \{0,1,2\}$, we have $|w_i| < 4$ and $|w_i| \neq 2$, for all $i \in \{0,1,2\}$.

Indeed, if none of the factors w_0, w_1, w_2 contains the hole inserted in τ, the result is an immediate consequence of Remark 1. Hence, we may assume that the hole is contained in one of the words w_0, w_1 or w_2.

Assume that there exist a non-empty factor $w_0 w_1 w_2$ of τ' and a partial word w, such that:

- one of the factors w_0, w_1 and w_2 contains a hole,
- $w_i \subset w$, for all $i \in \{0,1,2\}$,
- $|w_i| \geq 4$ or $|w_i| = 2$, for all $i \in \{0,1,2\}$.

Without loss of generality, we may assume that the hole was placed in w_0 (the other cases can be approached similarly). Also, let w_0' be the factor of τ in which a hole was inserted in order to obtain w_0; note that w_1, w_2 and $w_0' w_1 w_2$ are factors of τ, and we have $w_1 = w_2 = w$ and $w_0' \neq u$.

There are several cases to be analyzed:

Case 1: $|w_0| = 2k$, for $k \geq 1$, and the first symbol of w_0 is placed at an even position in τ'. Since $\tau = \phi(\tau)$, τ is cube-free and in τ' was inserted exactly one hole, it follows that $w_1 = w_2 = w = h(a_0 \cdots a_{k-1})$, where $a_j \in \{a,b\}$ for all $j \in \{0,\ldots,k-1\}$, and $w_0' = h(a_0 \cdots a_{l-1} a_l' a_{l+1} \cdots a_{k-1})$, where $a_l' \neq a_l$, for an integer l, with $l < k$. Moreover, one of the two letters of $\phi(a_l')$ was replaced with a hole to obtain w_0. If $a_l' = b$ it follows that $a_l = a$. But, since $\phi(a_l') = ba$ and $\phi(a_l) = ab$, we get that any partial word that can be obtained from $\phi(a_l')$ by replacing one of its letters with a hole cannot be contained in $\phi(a_l)$. The same argument holds in the case when $a_l' = a$ and $a_l = b$. Thus, we reach a contradiction.

Case 2: $|w_0| = 2k$, for $k \geq 1$, and the first symbol of w_0 is placed at an odd position in τ'. It follows that $w_1 = w_2 = w = b_0 h(a_0 \cdots a_{k-1}) b_1$ with $b_0, b_1, a_j \in \{a,b\}$, for all $j \in \{0,\ldots,k-1\}$. Note that $b_0 \neq b_1$, since

$b_1b_0 = \phi(f)$, for some $f \in \{a, b\}$. In this case, the word w_0' may have one of the following forms:

- $w_0' = b_0'\phi(a_0 \cdots a_{k-1})b_1$ with $b_0' \neq b_0$, or
- $w_0' = b_0\phi(a_0 \cdots a_{k-1})b_1'$ with $b_1' \neq b_1$, or
- $w_0' = b_0\phi(a_0 \cdots a_{l-1}a_l'a_{l+1} \cdots a_{k-1})b_1$, for some integer l with $a_l \neq a_l'$ and $0 \leq l < k$.

The last possibility leads to a contradiction similar to the one in Case 1. If $w_0' = b_0'h(a_0 \cdots a_{k-1})b_1$ with $b_0' \neq b_0$, since $b_0 \neq b_1$, it follows that τ contains the factor

$$h(a_0 \cdots a_{k-1})b_1w_1w_2 = \phi(a_0 \cdots a_{k-1})\underline{b_1}b_0\phi(a_0 \cdots a_{k-1})\underline{b_1}b_0\phi(a_0 \cdots a_{k-1})\underline{b_1}$$

a contradiction to the fact that τ is overlap-free.

Finally, if $w_0' = b_0\phi(a_0 \cdots a_{k-1})b_1'$, with $b_1' \neq b_1$, it follows $b_1' = b_0$, and hence $b_0b_0 = \phi(f)$, for some $f \in \{a, b\}$, again a contradiction.

Case 3: $|w_0| = 2k + 1$, for $k \geq 2$, and the first symbol of w_0 is placed at an even position in τ'. It follows that $w_1 = e\phi(a_0 \cdots a_{k-1})$, and $w_2 = \phi(b_0 \cdots b_{k-1})f$, where $e, f, a_j, b_j \in \{a, b\}$, for all $j \in \{0, \ldots, k-1\}$. We can easily observe that $a_j \neq b_j$, for $j \in \{0, \ldots, k-1\}$, and $a_{j-1} \neq b_j$, for $j \in \{1, \ldots, k-1\}$. Consequently, $a_j = a_{j+1}$ and $b_j = b_{j+1}$, for all $j \in \{0, \ldots, k-2\}$. Since τ is cube-free, it follows that $k = 2$. We may assume, without loss of generality, that $e = a$. Hence, $w_1 = w_2 = ababa$. But this is a contradiction to the fact that τ is overlap-free.

Case 4: $|w_0| = 2k + 1$, where $k \geq 2$ and the first symbol of w_0 is placed at an odd position in τ'. It follows that $w_1 = \phi(a_0 \cdots a_{k-1})e$, and $w_2 = f\phi(b_0 \cdots b_{k-1})$, where $e, f, a_j, b_j \in \{a, b\}$, for all $j \in \{0, \ldots, k-1\}$. As in the former case, we observe that $a_j \neq b_j$, for all $j \in \{0, \ldots, k-1\}$ and $a_{j+1} \neq b_j$, for all $j \in \{0, \ldots, k-2\}$. Consequently, $a_j = a_{j+1}$ and $b_j = b_{j+1}$, for all $j \in \{0, \ldots, k-2\}$. In particular, we obtain that $a_0 \neq b_{k-1}$. Thus, the first letter of $\phi(a_0)$ and last letter of $\phi(b_{k-1})$ coincide. Since e equals the last letter of $\phi(b_{k-1})$ and f equals the first letter of $\phi(a_0)$, it follows that $e = f$. This is not possible, since $ef = \phi(c)$, for some letter $c \in \{a, b\}$.

All the cases lead to a contradiction. Consequently, we have proved that for a non-empty factor $w_0w_1w_2$ of τ', and a partial word w, such that $w_i \subset w$, for all $i \in \{0,1,2\}$, we have $|w_0| < 4$ and $|w_0| \neq 2$.

Therefore, if we want to replace a letter of the infinite word τ with a hole, and obtain an infinite cube-free partial word τ', we should only verify that this replacement does not cause the apparition in τ' of a non-empty factor $w_0w_1w_2$ with $|w_0| = |w_1| = |w_2|$ and $|w_0| \in \{1,3\}$, for which there exists a partial word w such that $w_i \subset w$, for all $i \in \{0,1,2\}$.

We observe that there exist positions in τ where a substitution, respecting the restrictions described above, can be performed. For example, in the word

$$\phi^5(a) = abbabaabbaaba\underline{b}babaababbaabbabaab$$

which is a prefix of τ, the underlined letter can be replaced with a hole, and the partial word we obtain remains cube-free.

Also, we observe that $\phi^5(a)$ has an infinite number of occurrences as a factor of τ. For each such occurrence, we can construct a cube-free infinite partial word with exactly one hole, by replacing the 14th letter in $\phi^5(a)$ with $\diamond$. The infinite word we obtain will have the form

$$xabbabaabbaaba\diamond babaababbaabbabaaby$$

with $x \in \{a,b\}^*$ a prefix of τ, and $y \in \{a,b\}^\omega$ a suffix of τ.

In conclusion, we have proved that there exist infinitely many cube-free infinite partial words, containing exactly one hole. □

Since any cube-free infinite partial word is k-free, for $k \geq 3$ (as noted in Remark 2), we obtain, as a corollary of Proposition 9, the following result:

Corollary 3. *For $k \geq 3$, there exist infinitely many k-free infinite partial words, containing exactly one hole, over a binary alphabet.*

We also obtain, as another consequence, an already known result (see [Bra83, Lot97]):

Corollary 4. *For $k \geq 3$, there exist infinitely many k-free infinite full words, over a binary alphabet.*

Proof. Let τ' be one of the infinite k-free partial word constructed in the proof of Proposition 9. We replace the hole in τ' with an a letter; it is clear that the word obtained in this manner is a k-free infinite full word, for $k \geq 3$. This procedure can be applied to each of the infinite partial words constructed in the proof of Proposition 9 and obtain an infinite full word. Moreover, each two of these newly obtained infinite full words are different. □

Next, we extend the result stated in Proposition 9 in order to obtain cube-free partial words, with infinitely many holes. First, note the following.

Remark 8. *The word $\phi^k(a)$, where $k \geq 1$, has an infinite number of non-overlapping occurrences in τ, with its first letter placed at an even position. To begin with, $\phi^k(a)$ has one occurrence in τ, with the first letter placed at the position 0. Also, since $\phi^{i+1}(a) = \phi^i(a)\overline{\phi^i(a)}$, thus $\phi^{i+2}(a) = \phi^i(a)\overline{\phi^i(a)\phi^i(a)}\phi^i(a)$, and $|\phi^i(a)| = 2^i$, for all $i \geq 1$, it can be easily proved by induction that $\phi^k(a)$ occurs at least 2^l times in $\phi^{k+l+1}(a)$. Moreover, all of these occurrences have their first letter placed at an even position.*

Now let us look at partial words with infinitely many holes.

Theorem 10. *There exists a cube-free partial word, containing infinitely many holes, over a binary alphabet.*

Proof. From Remark 8 it follows that in the Thue-Morse word τ there exist an infinite number of non-overlapping occurrences of the word $\phi^5(a)$, each having its first letter placed at an even position. Furthermore, for each of these occurrences of $\phi^5(a)$, we replace its fourteenth letter (the underlined letter in the factor $abbabaabbaaba\underline{b}babaababbaabbabaab$) with a hole, in this manner resulting in an infinite partial word, with an infinite number of holes, τ'. It is clear that τ' can be obtained from τ applying on τ a morphism δ that takes a to $\phi^5(a) = abbabaabbaaba\diamond babaababbaabbabaab$ and b to $\phi^5(b)$. We claim that the partial word τ' is cube-free.

Note that if there exist a non-empty factor $w_0w_1w_2$ of τ' and a partial word w such that $w_i \subset w$, for $i \in \{0, 1, 2\}$, only a finite number of holes are contained in this factor; let n be this number. Consequently, to prove our claim it is sufficient to show that any word obtained by replacing n letters of τ with holes, at some of the aforementioned positions, is cube-free, for all integers $n \geq 0$.

We prove this result by induction on n: for $n = 0$ and $n = 1$ it was already shown to be true in the proof of Proposition 9. We assume the statement holds for all $k < n$, and prove it for n.

Let τ' be a word obtained by replacing n letters of τ with holes, on n of the positions already defined. Assume, for the purpose of contradiction, that τ' contains a non-empty factor $w_0w_1w_2$ and there exists a partial word w such that $w_i \subset w$, where $i \in \{0, 1, 2\}$. Obviously, all holes must be contained in the factor $w_0w_1w_2$ since otherwise, we obtain, using the procedure described above, a non-cube-free infinite partial word with less than n holes, a contradiction to the induction hypothesis.

Note that in the infinite partial word τ' there are at least thirty-one letters between two distinct holes. Moreover, since $n \geq 2$, it follows that the factor $w_0w_1w_2$, whose length is divisible by three, has at least thirty-three symbols, and, consequently, $|w_i| \geq 11$, for all $i \in \{0, 1, 2\}$. Also, remark that any hole appearing in τ' replaces a b letter, and, consequently, the position that corresponds in w to that hole is occupied by an a letter (otherwise, the hole is not necessary, and, again, we obtain a contradiction to the induction hypothesis). Finally, note that all the holes are placed at an odd position in τ.

There are several cases to be analyzed.

Case 1. First let us assume that w_0 contains at least one hole. We have some factorizations $w_0 = w_{00}e_0w_{01}$, $w_1 = w_{10}e_1w_{11}$, and $w_2 = w_{20}e_2w_{21}$, with $e_0 = \diamond$ and $w = u_0au_1$, where $w_{ij} \subset u_j$, for $i \in \{0, 1, 2\}$ and $j \in \{0, 1\}$. Again, there are two cases to be discussed:

- $e_1 = a$, and,
- $e_1 = \diamond$.

Note that e_1 and e_2 cannot be simultaneously equal to $\diamond$, because, otherwise, none of the holes e_0, e_1 and e_2 is necessary, getting a contradiction with the induction hypothesis.

We will only describe how the first case leads to a contradiction, since the other one can be treated similarly. We remark that $|w_0| \geq 11$. Therefore, we have $|w_{00}| + |w_{01}| \geq 10$, so either $|w_{00}|$ or $|w_{01}|$ is at least 5. If $|w_{00}| \geq 5$, it follows that $baaba\diamond$ is a factor of w_0, and, consequently, $baabaa$ is a factor of u. Thus, $baabaa$ or a partial word contained in $baabaa$, with exactly one $\diamond$ replacing one of the a letters, is a factor of w_1. But, this leads to a contradiction. Indeed, in the case when no hole appears in this factor, since τ' was obtained by substituting some of the letters of $\tau = \phi(\tau)$ with holes, it follows that at least one of the two factors aa should be the image of a letter through the morphism ϕ, which is impossible. In the other case, when a hole replaces an a letter, considering the way we introduce holes into τ, it follows that $\diamond$ can replace only the last a in the sequence. This coincides with the letter denoted by e_1. we have a contradiction with the assumption that $e_1 \neq \diamond$.

If $|w_{00}| < 5$ and $|w_{01}| \geq 6$, it follows that $\diamond babaab$ is a factor of w_0. If $5 > |w_{00}| \geq 1$, it follows that $aababaab$ is a factor of w. Consequently $aababaab$ (or a partial word contained in $aababaab$, with exactly one $\diamond$ replacing an a letter) is a factor of w_1, again a contradiction, using the same reasoning as above.

If $|w_{00}| = 0$, it follows that $ababaababba$ is a factor of w. Hence, $ababaababba$ (or a partial word contained in $ababaababba$, with exactly one $\diamond$ replacing one of the a letters, other than the first) is a prefix of w_1. Note that no partial word contained in $ababaababba$, with a hole instead of an a other than the first one, can be obtained by the procedure that we use, since any hole should be followed by the factor $babaab$ or preceded by an a letter. Therefore, $ababaababba$ is a prefix of w_1. Also, note that the first symbol of w_1 is at an odd position in τ' (otherwise, the factor aa would have been the image of a letter through the morphism ϕ, a contradiction). In this case, since w_0 starts with $\diamond$, and a hole can be placed only at odd positions, we obtain that the length of the string w_{01} is odd. Since the first letter of w_1 is an a, it follows that the last letter of w_{01} is a b, as well as the last letter of w. This implies that the last letters of w_1 and w_2 are b letters. Since the last letter of w_2 is placed at an even position, it follows that the letter placed exactly after w_2 in τ' is an a. Consequently, $y_0 = w_{01}e_1$, $y_1 = w_{11}e_2$, $y_2 = w_{21}a$ and $y_0y_1y_2$ are factors of τ', and $u' = u_1a$ is a partial word, such that $y_i \subset u'$, for $i \in \{0, 1, 2\}$. Moreover, $y_0y_1y_2$ contains $n - 1$ holes, which is a contradiction to the induction hypothesis.

Case 2. w_2 contains at least one hole, and w_0 does not contain any hole, or w_1 contains at least one hole, and both w_0 and w_2 do not contain any holes.

We assume that w_0 does not contain any hole, and analyze the rest of the cases. If $w_1 = w_{10} \diamond w_{11}$, with $w_{10} \neq \varepsilon$, or $w_2 = w_{20} \diamond w_{21}$, with $w_{20} \neq \varepsilon$, we can apply similar arguments as in the previous case, and reach the same conclusion. If none of these cases occur, it follows that holes may replace only the letters placed at the first positions of w_1 and w_2. Therefore, $n \leq 2$. But, due to the induction hypothesis, we have $n \geq 2$, and, thus, we obtain $n = 2$. Hence, we have $w_1 = \diamond u$ and $w_2 = \diamond u$, for some word u, and no other $\diamond$ exists in τ'. Moreover, $w_0 = u = au$. It follows that, $au \diamond u \diamond u$ is a factor of τ', where u is a non-empty word that does not contain any hole. Thus, $ububu$ is a factor of τ, a contradiction to the fact that τ is overlap-free.

Since all the cases lead to a contradiction, we conclude that the assumption we made is false, and conclude our proof. □

Considering that in τ there exist infinitely many non-overlapping occurrences of $\phi^5(a)$ having their first letter placed at even positions, it follows that we can obtain an infinite number of cube-free infinite partial words with infinitely many holes. This can be done by choosing, randomly, infinitely many such occurrences of $\phi^5(a)$ and substitute, in each of them, the fourteenth letter with a hole, in the same way we have described in the proof of Theorem 10. It is clear that all the infinite partial words obtained in this manner are cube-free.

The following corollary is immediate:

Corollary 5. *For $k \geq 3$, there exist arbitrarily many k-free partial words, containing an infinite number of holes, over a binary alphabet.*

The proof of Theorem 10 provides an efficient solution to the following algorithmic problem: given the natural number n find a k-free partial word (for some $k \geq 3$) containing exactly n holes. In the following, we propose an algorithm that constructs a cube-free partial word with exactly n holes, offering, thus, a solution for this problem.

As stated in Remark 8, the word $\phi^{n+6}(a)$, with $n \geq 1$, contains at least 2^n non-overlapping occurrences of $\phi^5(a)$ having the first letter on an even position. Also, note that both the computational time and space needed to construct $\phi^n(a)$ are $\mathcal{O}(2^n)$. Thus, $\phi^{\lceil \log_2 n \rceil + 6}(a)$ has $\mathcal{O}(n)$ letters and can be constructed in $\mathcal{O}(n)$ time. Also, it contains n non-overlapping occurrences of $\phi^5(a)$, each having its first letter on an even position.

Algorithm 1 Construct-cube-free-word(n)

```
construct $\phi^{\lceil \log_2 n \rceil + 6}(a)$
identify $n$ non-overlapping occurrences of $\phi^5(a)$ in $\phi^{\lceil \log_2 n \rceil + 6}(a)$, having their first letters on even positions
for each of these occurrences do
    substitute its fourteenth letter with $\diamond$
end for
denote by $\tau'_{\lceil \log_2 n \rceil + 6}$ the word obtained after the $n$ substitutions were performed
return Construct-cube-free-word($n$):= $\tau'_{\lceil \log_2 n \rceil + 6}$ (the algorithm stops)
```

According to the proof of Theorem 10, Algorithm 1 constructs a cube-free partial word.

The running time of the above algorithm is clearly $\mathcal{O}(n)$. Indeed, we have already stated that the step where $\phi^{\lceil \log_2 n \rceil + 6}(a)$ is constructed can be performed in linear time. Furthermore, the identification of the occurrences of $\phi^5(a)$ as well as the step where the fourteenth letter of each of these strings is substituted with a hole can be completed in $\mathcal{O}(n)$ steps.

We note that it is impossible to solve this problem with an algorithm that requires less than n steps, since the string we construct must have at least n letters, the holes.

3.3.2 Generalization of cube-freeness

There have been studied applications of both partial and infinite words in the processing and analysis of DNA strings ([Har06, Leu05]), which are encoded over the four-letter alphabet $\{a, c, g, t\}$. Therefore, it seems interesting to us to analyze the existence and construction of k-free partial words that contain effectively four letters.

To begin with, we use a morphism δ defined on $\{a, b\}$ with values in $\{a, b, c, d\}^*$ defined as follows: $\delta(a) = abcd$ and $\delta(b) = badc$. Let $w = \delta(\tau)$ be the infinite word obtained by applying our morphism to the Thue-Morse word τ. We observe that if we delete the c and d letters from w, we obtain τ. Also, if we delete the a and b letters, we obtain the Thue-Morse word in which a is replaced by c and b by d, respectively. To keep the exposure simple, assume that the distance between two letters of w, placed at the positions n_0 and n_1 of w, respectively, is defined as $|n_0 - n_1|$. Note that the distance between two identical letters of w can be $4s, 4s+1$ or $4s+3$, for some integer $s > 0$.

It is not hard to see that w is cube-free. To prove this, assume, for the purpose of contradiction, that w contains a factor xxx, with $x \in \{a, b, c, d\}^+$. Also, assume that x has an a as its first letter. First, it follows that $|x| = 4k$, $|x| = 4k+1$ or $|x| = 4k+3$, with $k \geq 0$, since $|x|$ equals the distance between the first letter of the first x factor and the first letter of the second x factor, which are identical. If $|x| = 4k$ it follows that if we delete the c and d letters from xxx we obtain a non-empty factor yyy, for $y \in \{a, b\}^+$, contained in the Thue-Morse word τ. But, this would mean that τ is not cube-free, a contradiction to Theorem 1. If $|x|$ is odd it follows that the distance between the first letter of the first factor x and the first letter of the third factor x is $4p+2$, for some integer $p \geq 0$, a contradiction. If x has as first letter a b, a c or a d, similar arguments lead to a contradiction.

Also, we observe that any word that can be obtained from w by substituting one of its letters with a hole is still cube-free. Let w' be an infinite word obtained by replacing a letter of w with a hole. Also, assume that w' contains a non-empty factor $w_0w_1w_2$ and there exists a partial word u such that $w_i \subset u$ for all $i \in \{0, 1, 2\}$. First, note that $|w_0| > 1$. If $|w_0| = 4k$, with $k > 0$, and $\diamond$ replaces a c or a d letter, then we proceed as above and delete the c and the d letters, as well as the $\diamond$, and obtain that τ is not cube-free, a contradiction. The same strategy is applied for the case when $\diamond$ replaces an a or a b letter, but now the deleted symbols are a, b and $\diamond$. If $|w_0| = 4k+1$ or $|w_0| = 4k+3$, with $k \geq 0$, it follows, from the proof of the fact that w is cube-free, that one of the first symbols of w_0, w_1 or w_2 is a hole (otherwise the distance between two identical letters of w is $4p+2$, for some integer $p \geq 0$). But, this would mean that the symbols at the second positions of each of these factors coincide. Hence, the distance between the second letter of w_0 and the second letter of w_2, which are identical, is $4p+2$ with $p \geq 0$, a contradiction. Finally, with the same arguments, $|w_0| \neq 4k+2$. Consequently, the assumption that we made is false, and by replacing any letter by a hole, in w, we still obtain a cube-free word.

Remark that in the case of three-letter alphabets it is impossible to construct an infinite word w in which we can substitute randomly one of its letters with a hole and obtain a cube-free word in all the cases. Indeed, if such a word exists it follows that the number of letters between two identical letters of w is at least 2. But the words that verify this condition are of the form $\underline{a_0}a_1a_2\underline{a_0}a_1a_2\underline{a_0}a_1a_2\cdots$, where a_0, a_1 and a_2 are different letters of the alphabet. A word having this form is not cube-free, and, thus, the partial word obtained by replacing one of its letters with a hole is not cube-free, as well.

Theorem 11. *If w' is an infinite partial word obtained from $w = \delta(\tau)$ by replacing infinitely many of its letters with holes, such that each two consecutive holes are separated by at least two letters, then w' is cube-free.*

Proof. Assume, for the purpose of contradiction, that w' contains a non-empty factor $w_0w_1w_2$ and there exists a partial word u, such that $w_i \subset u$, for all $i \in \{0, 1, 2\}$.

It is not hard to see that $|w_0| \geq 3$. Note that there exists $k \leq |w_0|$ such that kth letters of w_0 and w_2 are both different from $\diamond$; this holds since otherwise, it is impossible to have at least two letters between every two consecutive holes according to Remark 5. This remark proves that the length of $|w_0|$ is even (otherwise, the distance between the two identical letters placed at the kth position of x_0 and x_2 is a number of the form $4m+2$, a contradiction). In a similar fashion we can show that there exists $l < |w_0|$ such that lth symbols of w_0 and w_1 are both different from $\diamond$. Combined with the fact that $|w_0|$ is even, this leads to the fact that $|w_0| = 4m$, for some integer $m \geq 0$.

Let w_0', w_1' and w_2' be the factors of w from which w_0, w_1 and, respectively, w_2 were obtained, by replacing some of their letters with holes. Since $|w_i'| = |w_i| = 4m$, for $i \in \{0, 1, 2\}$, it follows that these words have the following form: $w_i' = w_{i0}b_{i,0}\cdots b_{i,m-1}w_{i1}$, such that: $b_{i,j} = \delta(e_{i,j})$, with $e_{i,j} \in \{a, b\}$, $|w_{00}| = |w_{10}| = |w_{20}|$, $|w_{01}| = |w_{11}| = |w_{21}|$, $w_{01}w_{10} = \delta(e_0)$, and $w_{11}w_{20} = \delta(e_1)$, with $e_0, e_1 \in \{a, b\}$. We assume that, for $i \in \{0, 1, 2\}$, we have $w_i = u_{i0}c_{i,0}\cdots c_{i,m-1}u_{i1}$, where $c_{i,j}$ was obtained from $b_{i,j}$, u_{i0} from w_{i0}, and u_{i1} from w_{i1}, for all i and j, respectively, by replacing some of their letters with holes.

If there exist $j \in \{0, \ldots, m-2\}$ and $i, k \in \{0, 1, 2\}$ such that $i \neq k$ and $b_{i,j} \neq b_{k,j}$, it follows that $b_{i,j}$ and $b_{k,j}$ differ at every position, i.e., the lth letter of $b_{i,j}$ differs from the lth letter of $b_{k,j}$, for all integers $l \in \{0, 1, 2, 3\}$. Consequently, at least four letters in both these words must be substituted with holes in order to obtain $c_{i,j}$ and $c_{k,j}$, which are both included in the same partial word. But this is impossible, since two consecutive holes are separated by at least two letters, see Remark 5. Thus, we obtain that $b_{0,j} = b_{1,j} = b_{2,j}$, for all $0 \leq j \leq m-1$. This proves that $w'_i = w_{i0} z w_{i1}$, for $0 \leq i \leq 2$, and $z = \delta(z')$, for some factor z' of τ.

In the same manner, we can show that $w_{01}w_{10} = w_{11}w_{20}$, and, as a consequence $e_0 = e_1 = e$, for $e \in \{a, b\}$. Note that if $z \neq \varepsilon$, we deduce that w contains the factor $z\delta(e)z\delta(e)z$, and, since $z = \delta(z')$, it follows that τ is not overlap-free (having as a factor the word $z'ez'ez'$), a contradiction. Hence, we may assume, for the rest of the proof, that $z = \varepsilon$.

Moreover, we can obtain, similarly, that if $|w_{00}| \geq 3$, then $w_{00} = w_{10} = w_{20}$. But this would prove that w contains the factor $fw'_0w'_1w'_2$, where $f \in \{a, b, c, d\} \cup \{\varepsilon\}$ such that $fw_{00} = w_{01}w_{10} = w_{11}w_{20} = \delta(e)$. Consequently, w contains the factor $fw_{00}w_{01}w_{10}w_{11}w_{20}$, a contradiction to the fact that w is cube-free. Analogously, the case when $|w_{01}| \geq 3$ leads to a contradiction.

Thus, the only possibility left to be analyzed is when we have $|w_{i0}| = |w_{i1}| = 2$, for all $i \in \{0, 1, 2\}$. If $w_{00} = w_{10}$ or $w_{21} = w_{01}$, we obtain again, easily, a contradiction. Hence, we have $w_{00} \neq w_{10}$ (which implies that they differ at every position) and $w_{21} \neq w_{01}$ (also implying that they differ on every position). Since $w_{ij'}$ was obtained from w_{ij} by substituting some of their letters with holes, for all $i \in \{0, 1, 2\}$ and $j \in \{0, 1\}$, it follows that the strings $w'_{21}w'_{00}$ and $w'_{01}w'_{10}$ are both contained in the same word (which consists in the last two letters of u followed by the first two letters of u). Again, this means that at least four letters in the strings $w_{21}w_{00}$ and $w_{01}w_{10}$ were substituted with holes. According to Remark 5 this is impossible because each two consecutive holes are separated by at least two letters.

We have shown that all the cases lead to a contradiction, and, consequently, the assumption that we have made, namely that w' is not cube-free, is false. This concludes our proof. □

The following corollary results immediately.

Corollary 6. *If w' is an infinite partial word obtained from $w = \delta(\tau)$ by inserting an infinite number of holes, such that each two consecutive holes are separated by at least two letters, then w' is k-free, for every $k \geq 3$.*

This time, an algorithm that produces a k-free word with n holes, for $k \geq 3$, can be obtained more easily. We construct the prefix of length $3n-2$

Algorithm 2 Construct-cube-free-word-4-letters(n)

```
construct $\phi^{\lceil \log_2 \lceil \frac{3n}{4} \rceil \rceil}(a)$ (this word has $\lceil \frac{3n}{4} \rceil$ letters)
construct $u = \delta(\phi^{\lceil \log_2 \lceil \frac{3n}{4} \rceil \rceil}(a))$ (this word has at least $3n$ letters)
construct $v$ as the prefix of length $3n - 2$ of $u$
construct $v'$ from $v$ by replacing the letters at the positions $0, 3, \ldots, 3(n-1)$ with holes
return  Construct-cube-free-word-4-letters($n$):= $v'$
```

of $\delta(\tau)$ and replace n of its letters with holes, such that the number of letters between two consecutive holes is two. In this way Algorithm 2 will return a cube-free partial word (thus, k-free partial word) with exactly n holes.

The time complexity of this algorithm, as in the case of Algorithm 1, is clearly $\mathcal{O}(n)$, as well as its space complexity. Note, also, that the partial word produced by this algorithm has the minimal length that a cube-free partial word containing n holes can have. Indeed, if the length of a partial word w is less than $3n-2$ it follows that in this word one can find two holes that are separated by at most one letter, and, consequently, it has at least one factor of the form $\diamond\diamond a$, $\diamond a \diamond$ or $a\diamond\diamond$, for some letter a. In all these cases w is not cube-free.

3.4 Algorithms

In this chapter we propose two main algorithms. A first one that, given a finite partial word w and a natural number k, decides whether w is k-free or not (joint work with Florin Manea [MM07]), and a second one that, given a full word w, and two integers d and p, determines if partial words can be created by puncturing holes into w, such that the newly created partial words have period p and no two holes within distance d (joint work with Francine Blanchet-Sadri, Abraham Rashin and Elara Willett [BSMRW09]).

Now, let w be a partial word. If w is not k-free, the algorithm computes a non-empty factor $x_0 \cdots x_{k-1}$ of the input word w and a partial word u such that $x_i \subset u$, for all $i \in \{0, \ldots, k-1\}$. We analyze the soundness and the time complexity of this algorithm (on the random access machine model). In the following we assume that the input partial word w is over the alphabet A, and $*$ is a symbol not contained in A. Moreover, we denote by n the length of w.

First, we define the two-dimensional array $\uparrow [\,][\,]$, with n rows, $\left\lfloor \frac{n}{k} \right\rfloor$ columns, with elements from $A \cup \{\diamond\}$, as follows:

$$\uparrow [i][l] = \begin{cases} a & \text{if there exists an } a \text{ such that } w[i+hl] \subset a \text{ for all} \\ & h \in \{0,\ldots,k-1\}, \text{ and, for any symbol } b, \text{ such that} \\ & w[i+hl] \subset b \text{ for all } h \in \{0,\ldots,k-1\}, \text{ we have } a \subset b \\ * & \text{otherwise} \end{cases} \tag{3.1}$$

The usage of the symbol $\uparrow$ to denote this array is motivated by the fact that $\uparrow [i][l] \neq *$ if and only if every two symbols a and b in the set $\{w[i], \ldots, w[i+(k-1)l]\}$ are compatible (therefore, $a \uparrow b$), and both a and b are contained in $\uparrow [i][l]$.

The values stored in this array can be computed using the following relation:

$$\uparrow [i+l][l] = \begin{cases} \uparrow [i][l] & \text{if } w[i+lk] \subset \uparrow [i][l], \text{ and } w[i+lh] \neq \diamond, \\ & \text{for some } h \in \{1,\ldots,k-1\}; \\ w[i+lk] & \text{if } w[i+lh] = \diamond, \text{ for all } h \in \{1,\ldots,k-1\}; \\ * & \text{otherwise.} \end{cases} \tag{3.2}$$

An algorithm that effectively computes this array consists basically in the following two steps:

- for each possible value of l and for each i, such that $i \leq l$, we compute $\uparrow [i][l]$, using the definition 3.1 of the array $\uparrow [\,][\,]$.
- we use the relation 3.2, defined above, to compute recursively the elements $\uparrow [i+l][l], \ldots, \uparrow [i + l \left\lfloor \frac{(n-i-(k-1)l)}{l} \right\rfloor][l]$.

By a careful implementation of the above strategy, the time needed to compute all the elements of the array $\uparrow [\,][\,]$, on an input consisting of the partial word w, with $|w| = n$, and the natural number k, is $\mathcal{O}(\frac{n^2}{k})$.

Further we show how this array can be used to decide whether a given partial word is k-free or not.

In order to prove the soundness of Algorithm 3, we note the following immediate facts:

Algorithm 3 $Free(w, k)$: Testing k-freeness for a word w

```
$n := |w|$
for $l = 1$ to $\lfloor \frac{n}{k} \rfloor$ do
  $counter := 0,\ s := 0$
  for $i = 1$ to $n - l * k + 1$ do
    if $\uparrow [i][l] \neq *$ then
      $counter := counter + 1$
      if $s = 0$ then
        $s := i$
      end if
    else
      $counter := 0, s := 0$
    end if
    if $counter = l$ then
      print $w$ is not $k$-free. We have $x_j = w[s+jl..s+(j+1)l-1], j \in \{1, \dots, k\}$, and $u = \uparrow [s][l] \dots \uparrow [s+l-1][l]$
      return $Free(w, k) := False$ (the algorithm stops)
    else
      $counter := 0$
    end if
  end for
end for
return $Free(w, k) := True$ (the algorithm stops)
```

- For two partial words u and v, both of length l, we have $u \subset v$ if and only if $u(i) \subset v(i)$, for all $i \in \{1, \dots, l\}$.

- Consequently, for a non-empty factor $x = x_1 \cdots x_k$ of the partial word w there exists a partial word u, of length $l > 0$, such that $x_i \subset u$ if and only if the string $u' = \uparrow [r][l] \uparrow [r+1][l] \dots \uparrow [r+l-1][l]$ does not contain the symbol $*$, where r is the first position of the factor x in w.

- If the input word w is not k-free, and, by definition, there exists a factor $x_1 \cdots x_k$ of w and a non-empty partial word u, such that $x_i \subset u$, for all $i \in \{1, \dots, k\}$, then the length of a factor x_i is bounded by $\lfloor \frac{n}{k} \rfloor$.

The algorithm we propose identifies, if any, a non-empty factor $x_1 \cdots x_k$ of the input word w and a partial word u such that $x_i \subset u$, for all $i \in \{1, \dots, k\}$. According to the facts presented above, we note that w is not k-free if and only if the sequence $\uparrow [1][l], \dots, \uparrow [n - lk + 1][l]$ contains l consecutive positions that differ from $*$. Moreover, if this sequence contains

l such consecutive positions, starting from the position s, it follows that a possibility to choose the k factors $x_1, \ldots, x_k$ of w and the partial word u, proving that the input word is not k-free, is to set $u = \uparrow [s][l] \ldots \uparrow [s+l-1][l]$ and $x_i = w[s+(i-1)l..s+il-1]$, for $i \in \{1, \ldots, k\}$. Once such a possibility is discovered, the algorithm stops and concludes that w is not k-free; if no such possibility is identified for any $l \leq \lfloor \frac{n}{k} \rfloor$, the algorithm stops, and decides that w is k-free.

The overall time complexity of Algorithm 3 is clearly $\mathcal{O}(\frac{n^2}{k})$, where $n = |w|$, since the most time consuming operation is the computation of the array $\uparrow [\,][\,]$. The space needed by this algorithm is also $\mathcal{O}(\frac{n^2}{k})$.

Finally, we note that Algorithm 3 can be applied, as well, for full words. However, in this case, an algorithm working in time $\mathcal{O}(n \log n)$ can be developed using suffix arrays ([CR02, MM93]).

Now let us go to the second algorithm. We say two positions i, j in a partial word u are *d-proximal* if $0 < |j - i| \leq d$, where d denotes a positive integer. We say that u obeys the *hole constraint* d (or u is *d-valid*) if no two holes in u are d-proximal. When the value of d is clear from context, we may suppress reference to it, simply referring to the "hole constraint" or to "proximal" positions.

Let w be a length n full word defined over an alphabet A of size k. In this section, we present an $\mathcal{O}(nd)$ time algorithm, which finds, for given positive integers d and p both less than n, a *d-valid* p-periodic partial word contained in w, if any exists. In other words, it determines whether it is possible to insert holes into w with no two holes within distance d, such that the resulting partial word has strong period p. If this is possible, such a word is returned.

In order to work with words of length n more easily, we write them in rows of length p. For a partial word u and for an integer x, $0 \leq x < p$, we will call *column* x the sequence of positions (or letters at these positions) $x, x+p, \ldots, x+lp$, where l is the maximal integer such that $x + lp < n$. For example in Figure 3.1, if $w = caadabecabdaeecaad$, $p = 7$, and $d = 2$, then $u = caadabeca\diamond da\diamond ecaad$ is obtained using our algorithm. For an integer x, $0 \leq x < p$, let $S_x = \{w(i) \mid 0 \leq i < n, i \equiv x \bmod p\}$ be the set of distinct letters appearing in column x of w. We construct a new set of partial words

$$\begin{array}{cc|c|cc|c|c} c & a & a & d & a & b & e \\ c & a & b & d & a & e & e \\ c & a & a & d & & & \end{array} \qquad \begin{array}{cc|c|cc|c|c} c & a & a & d & a & b & e \\ c & a & \diamond & d & a & \diamond & e \\ c & a & a & d & & & \end{array}$$

Figure 3.1: The words w and u with columns 2 and 5 highlighted

$\Omega = \{\omega \mid \omega(i) \in S_i\}$, and call $u \subset w$ a partial word *induced* by the choice $\omega \in S_0 \times S_1 \times \cdots \times S_{p-1}$, if $u \subset \omega^l$, for some rational l. Now, u, induced by ω, is d-valid if and only if for any two proximal positions i and j ($0 \leq i, j < n$, $0 < |i - j| \leq d$), it is not the case that $u(i) = \diamond = u(j)$.

Remark 9. *The choice of letters $\omega \in \prod_{x=0}^{p-1} S_x$ induces a d-valid word if and only if for every two proximal positions i, j, $u(i) = \omega(i \bmod p)$ or $u(j) = \omega(j \bmod p)$.*

This suggests a geometric approach for determining which choices of letters do not cause a hole constraint violation. For $(a_0, b_0) \in A^2$, let the *cross centered at* (a_0, b_0) be the set $+(a_0, b_0) = \{(a, b) \in A^2 \mid a = a_0$ or $b = b_0\}$. Then, the choices of letters a for column x and b for column y that do not cause any hole constraint violations, are precisely those in the intersection of the crosses centered at $(w(i), w(j))$, for i, j proximal positions in columns x, y.

The subsets of A^2 formed by intersecting crosses, however, are of special forms. The following theorem describes these forms, and shows that they can be determined in l-linear time, where l is the number of crosses that need to be intersected (the number of distinct ordered pairs (i, j) where i, j are proximal positions in columns x, y, respectively).

Theorem 12. *Considered as a set of entries in a $k \times k$ matrix, any set T formed by intersecting crosses must be of one of the forms:*

1. *FULL: the universe A^2;*
2. *CROSS(a_0, b_0): a cross $+(a_0, b_0)$;*
3. *ROW(a_0): a row of the matrix $\{(a_0, b) \mid b \in A\}$;*
4. *COL(b_0): a column of the matrix $\{(a, b_0) \mid a \in A\}$;*

5. *TWO*$((a_1, b_1), (a_2, b_2))$*: a set of two points* (a_1, b_1) *and* (a_2, b_2) *in neither the same row nor column;*

6. *ONE*(a_0, b_0)*: a singleton set* $\{(a_0, b_0)\}$*;*

7. *NULL: the null set* $\emptyset$*.*

Proof. Let us denote by m the number of crosses that are intersected: $T = \bigcap_{s=1}^{m} +(a_s, b_s)$. If $m = 0$, then $T = A^2$ is FULL. Moreover, the form FULL is only possible for $m = 0$. If $m = 1$, then $T = +(a_1, b_1)$ is CROSS(a_1, b_1).

Now suppose that $m > 1$ and let $T' = \bigcap_{s=1}^{m-1} +(a_s, b_s)$. We consider what happens when we intersect $+(a_m, b_m)$ with T', for T' in each of the above forms.

Let $T' =$ CROSS(a_0, b_0). If $(a_m, b_m) = (a_0, b_0)$, then $T' = +(a_m, b_m)$, and we get that $T = T'$. If $a_m = a_0$ and $b_m \neq b_0$, then $T =$ ROW(a_0). If $b_m = b_0$ and $a_m \neq a_0$, then $T =$ COL(b_0). If $a_m \neq a_0$ and $b_m \neq b_0$, then $T =$ TWO$((a_0, b_m), (a_m, b_0))$. Therefore, intersecting $+(a_m, b_m)$ with a CROSS matrix results in a CROSS, ROW, COL, or TWO matrix, as depicted in Figure 3.2. a).

If $T' =$ ROW(a_0) and $a_m = a_0$, then $T = T' \subset +(a_m, b_m)$. Otherwise, $T =$ ONE(a_0, b_m). Furthermore, if $T' =$ COL(b_0) and $b_m = b_0$, then $T = T' \subset +(a_m, b_m)$. Otherwise, $T =$ ONE(a_m, b_0).

Now, let $T' =$ TWO$((a, b), (a', b'))$. If (a_m, b_m) is equal to (a, b') or to (a', b), then $T = T' \subset +(a_m, b_m)$. Now, if $a = a_m$ or $b = b_m$ then $(a, b) \in +(a_m, b_m)$, but $(a', b') \notin +(a_m, b_m)$, and so $T =$ ONE(a, b). Similarly, if $a' = a_m$ or $b' = b_m$ then $T =$ ONE(a', b'). Finally, if $a \neq a_m$, $b \neq b_m$, $a' \neq a_m$ and $b' \neq b_m$, then $(a, b), (a', b') \notin +(a_m, b_m)$, so $T =$ NULL. Therefore, intersecting $+(a_m, b_m)$ with a TWO matrix results in a TWO, ONE, or NULL matrix, as depicted in Figure 3.2. b).

If $T' =$ ONE(a_0, b_0) and $a_m = a_0$ or $b_m = b_0$, then $T' \subset +(a_m, b_m)$, so $T = T'$. Otherwise, $T =$ NULL. Finally, if $T' =$ NULL, then $T =$ NULL. □

Now, returning to the question of which $\omega \in \prod_{x=0}^{p-1} S_x$ induce d-valid partial words, for two columns $x, y < p$, we define the constraint matrix M^{xy}, to be a $k \times k$ matrix such that, for all $a, b \in A$, $M^{xy}(a, b)$ is $*$ if for every pair of proximal positions i, j in columns x, y, $(a, b) \in +(w(i), w(j))$, and 0 otherwise. Note that, trivially, the constraint matrix from x to y is

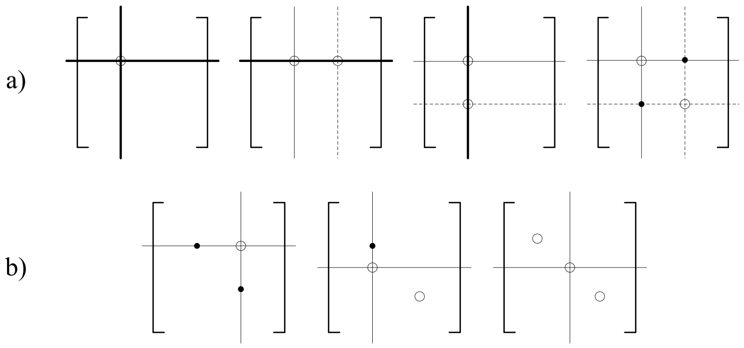

Figure 3.2: Intersection of different matrices

the transpose of the constraint matrix from y to x, and that $\omega \in \prod_{x=0}^{p-1} S_x$ induces a d-valid partial word if and only if for every $x, y \in \{0, \ldots, p-1\}$, $M^{xy}(\omega(x), \omega(y)) = *$.

The result of Theorem 12 is that the constraint matrices can be classified into a few simple types. Therefore, in practice, we store constraint matrices as objects that encode the form of the matrix (FULL, CROSS, TWO, etc.), and at most four characters to denote rows and columns (querying the position of stars in row a_0 of the object $<$ TWO$, (a, b), (a', b') >$ yields b if $a_0 = a$, b' if $a_0 = a'$ and NONE otherwise). These can be constructed and read in constant time.

Remark 10. *If columns x, y are proximal, that is x, y contain proximal positions, then $0 < |x - y| \leq d$ or $0 < p - |x - y| \leq d$.*

Fix some variables that will be shared by the algorithms: a table of constraint matrices, M; sets F_{ROW}, F_{ONE}, F_{TWO} and F_{CROSS}, where F_{FORM} contains (x, y) for which M^{xy} is of form FORM; a list of letters ω, where $\omega(x)$ is the letter chosen for column x. The following lemmas will be useful in proving the validity of our algorithms.

Lemma 9. *If $0 \leq x, y < p$ with $0 < |x - y| \leq d$, then M^{xy} is not FULL.*

Proof. The positions x and y in w are proximal since $0 < |x - y| \leq d$. Therefore at least one cross (namely, that centered at $(w(x), w(y))$ is used in the creation of the matrix M^{xy}, so it cannot be FULL. □

Furthermore, it follows from Theorem 12 that the types of constraints that one column can exert on another are limited.

Algorithm 4 Initializing the matrices

```
for $(x, y)$ columns within $d$ do
    $M^{xy} :=$ FULL
end for
for $i = 0$ to $n - y$ step $p$ do
    intersect $M^{xy}$ with cross centered at $(w(x + ip), w(y + ip))$
end for
```

Lemma 10. *If two columns x, y with $0 \leq x < y < p$, contain each at least two different letters, and M^{xy} is a CROSS matrix, then* $|x - y| \geq \max\{p - d, d + 1\}$.

Proof. Since M^{xy} is not a FULL matrix, by Remark 10, we have that $|x - y| \leq d$ or that $p - d \leq |x - y|$. Suppose that $|x - y| \leq d$, and let $y + sp$ be a position in column y, where $y \leq y + sp < n$. Thus, $x + sp$ is a position in column x, since $0 \leq x \leq x + sp < y + sp < n$. Furthermore, every position in column y is proximal to some position in column x. Since M^{xy} is a CROSS matrix, all ordered pairs $(w(i), w(j))$, for i, j proximal positions in columns x, y, must be equal. Therefore all letters in column y of w are equal, a contradiction. Therefore $|x - y| > d$ and $|x - y| \geq p - d$, so $|x - y| \geq \max\{p - d, d + 1\}$. □

There exist even more restrictions regarding CROSS matrices.

Lemma 11. *Let x_1, x_2, x_3 be distinct columns with at least two different letters each. If $M^{x_2x_3}$ and $M^{x_1x_3}$ are CROSS matrices, then $M^{x_1x_2}$ is neither a FULL nor a CROSS matrix.*

Henceforth, by *columns within d* we mean columns x, y such that $0 < |x - y| \leq d$ or $0 < p - |x - y| \leq d$. Any other pair of columns is necessarily related by a FULL constraint matrix and therefore can be ignored. Algorithm 4 computes all non-FULL constraint matrices of w in $\mathcal{O}(nd)$ time.

Corollary 7. *The forms (as per Theorem 12) of all the non-FULL constraint matrices for w can be determined in $\mathcal{O}(nd)$ time via Algorithm 4.*

Note that given two proximal columns x and y, and a letter a chosen for column x, there are either zero, one, or $\|S_y\|$ choices of a letter for column y that do not conflict with the choice of letter a for column x. This observation suggests an algorithm for labeling multiple columns. Let us now construct

Algorithm 5 Fill(x, a)

```
 1: initialize $Q$ to be an empty queue accepting columns
 2: choose letter $a$ for column $x$
 3: add $x$ to $Q$
 4: while dequeue $y$ from $Q$ do
 5:    let $b = \omega(y)$
 6:    for $z$ a neighbor of $y$ do
 7:       let row be the $b$ row of the matrix $M^{yz}$
 8:       remove edges between $y$ and $z$
 9:       if row has all $*$'s then
10:          next (go to line 4)
11:       else if row has exactly one $*$, say at position $c$ then
12:          if letter $c$ has already been chosen for column $z$ then
13:             next (go to line 4)
14:          else if column $z$ is unlabeled then
15:             choose letter $c$ for column $z$
16:             add $z$ to $Q$
17:             next (go to line 4)
18:          end if
19:       end if
20:       undo all recent labellings and edge erasures
21:       return false
22:    end for
23: end while
24: return true
```

a directed graph G that has vertex set $\{0, \ldots, p-1\}$ and edge set consisting of edges (x, y) labeled by M^{xy} when columns x, y are within d.

Theorem 13. *For a column x and a letter $a \in S_x$, Algorithm 5 correctly chooses letters for some additional columns such that, after the completion of this algorithm no undetermined column is constrained by an already determined column. Additionally, if the constraint matrices have already been computed, the running-time of Algorithm 5 is $\mathcal{O}(m)$, where m is the number of edges that are traversed.*

Proof. The problem of finding a choice of letters for the columns is equivalent to finding a labeling of the vertices of G, such that every vertex x is labeled with a letter $\omega(x)$ that occurs in column x of w, and for any two columns x and y within d, the $(\omega(x), \omega(y))$-entry of M^{xy} is a $*$. If such a labeling exists, then it induces a p-periodic d-valid partial word contained in w, by replacing every non-$\omega(x)$ letter in each column x with a hole.

The algorithm starts by assuming a labeling of vertex x by the letter a, and then performs a breadth-first search on the graph G, starting at x. This is implemented using a queue. Suppose that a vertex y has been marked by letter b and that we are now traversing an edge from y to z. Then the constraint matrix M^{yz} either uniquely determines the label c on the sink vertex z, or it imposes no constraint at all, or there are no choices, in which case the algorithm immediately fails. In the former case, either the unique label is applied ($\omega(z)$ is set to c) and vertex z is added to the queue for later traversal, or if $\omega(z)$ has already been set to a different value, the algorithm fails because there cannot be any labeling of G with $\omega(x) = a$ and $\omega(z)$ with its original value. In the case when no constraint is imposed (the b row is filled with $*$'s), this matrix is ignored, since for any value of $\omega(z)$ the matrix will not cause a contradiction. In all cases, the edge (y, z) and its opposite (z, y) are marked as having been traversed, so that they will not be visited again. In conclusion, an undetermined column is marked exactly when it is constrained by an already determined column, thus, ensuring that at the end of the algorithm no determined column will constrain an undetermined column. This algorithm visits m edges, no more than once each. On each edge, it performs a constant time operation. Thus, Algorithm 5 runs in $\mathcal{O}(m)$ time.

Please note that undoing all recent labellings and edge erasures, while keeping the algorithm's runtime within $\mathcal{O}(m)$, is solved in constant time by implementing data structures that could be "marked" in a particular state, and reset to this state later on. These data structures are used for the sets of neighbors of a vertex, the sets F_{FORM} (of edges of each type), and the set of labeled vertices. While, all the F_{FORM}'s and labellings can be reset in constant time, the vertex neighbor sets can be reset in $\mathcal{O}(l)$ time, where l is the number of vertices visited during this run of the algorithm. Since the number of vertices visited is less than the number of edges visited, $l < m$, the overall algorithm runs in $\mathcal{O}(m)$ time. □

The next lemma will help us prove that we never need to run Algorithm 5 ("Fill(x, a)") on a vertex more than twice.

Lemma 12. *Suppose that x and y are vertices of G such that $M^{xy} = TWO$ $((a, b), (a', b'))$, Fill(x, a) returns true, and $\omega \in \prod_{z=0}^{p-1} S_z$ induces a d-valid* *partial word u with $\omega(x) = a'$. Then, there exists a choice ω' of letters for the columns, that induces a d-valid partial word with $\omega'(x) = a$.*

Proof. Let T be the set of vertices of G that are labeled by $\text{Fill}(x, a)$, and Q be the labeling of T. For every vertex x of G, let $\omega'(x) = Q(x)$ if $x \in T$ and $\omega'(x) = \omega(x)$ otherwise. Since the labeling Q of T was generated by $\text{Fill}(x, a)$, we know that no letter choice for a vertex outside T is constrained by any of the letter choices specified in Q. Furthermore, since ω induced a d-valid partial word, we know that no constraint matrix is violated by two letter choices in ω. Therefore the letter choices in ω' do not violate any constraint matrices, so ω' induces a d-valid partial word. Also, clearly $\omega'(x) = a$, so we have our result. □

Algorithm 6 traverses all edges corresponding to non-FULL matrices and finds a consistent labeling of the vertices of G if any exists.

Theorem 14. *Algorithm 6 returns a d-valid p-periodic partial word contained in w, unless no such word exists. The running-time of the algorithm is $\mathcal{O}(nd)$.*

Proof. If there is a NULL matrix between two columns, then no consistent labeling of the vertices exists, so the algorithm fails. If any column in w has all letters equal, then that letter must be assigned for the column, and $\text{Fill}(x, w(x))$ ran. There can only be one consistent labeling of all vertices if it succeeds (note that the determination of whether a column has only one character can be performed in $\mathcal{O}(\frac{n}{p})$ time, and thus, it can be performed for all columns in $\mathcal{O}(n)$ time). Similarly, if there is a ROW or ONE matrix M^{xy} with a $*$ in row a, then a must be chosen for column x. We run $\text{Fill}(x, a)$, and it must succeed for there to be a consistent labeling of the vertices of G.

If $M^{xy} = \text{TWO}((a, b), (a', b'))$ then we know that any consistent labeling of the vertices of G must have column x labeled with either a or a'. But by Lemma 12, if some consistent labeling of G exists and $\text{Fill}(x, a)$ returns true, then there exists a consistent labeling of G that agrees on all choices of letters made by $\text{Fill}(x, a)$. Therefore in this case we can simply continue. Otherwise we try $\text{Fill}(x, a')$. If this fails, then we return false.

Algorithm 6 Traversing the entire graph

```
initialize matrices
for $(x, y)$ columns within $d$ do
    if $M^{xy}$ = NULL then
        return false
    end if
    add $(x, y)$ to $F_{\text{FORM}}$
end for
for column $x$ do
    if $\|S_x\| = 1$ then
        Fill$(x, w(x))$
    end if
end for
while exists $(x, y)$ with $M^{xy}$ of form ROW$(a)$, in $F_{\text{ROW}}$ do
    if not Fill$(x, a)$ then return false
end while
while exists $(x, y)$ with $M^{xy}$ of form ONE$(a, b)$, in $F_{\text{ONE}}$ do
    if not Fill$(x, a)$ then return false
end while
while exists $(x, y)$ with $M^{xy}$ of form TWO$((a, b), (a', b'))$, in $F_{\text{TWO}}$ do
    if not Fill$(x, a)$ and not Fill$(x, a')$ then return false
end while
for column $x$ do
    if column $x$ is unlabeled then
        choose $w(x)$ for column $x$
    end if
end for
for $i$ from 0 to $n - 1$ do
    let $u(i)$ be $w(i)$ if $w(i) = \omega(i \bmod p)$ and $\diamond$ otherwise
end for
return $u$
```

At this point in the algorithm, any unlabeled vertices x, y are related by either a FULL or CROSS matrix, since all other types of matrices have already been taken into account. Consider a graph T' with the so-far unlabeled vertices of G as the vertex set, and an edge between x and y if and only if M^{xy} is a CROSS matrix. We can satisfy all remaining constraints (the CROSS matrices) by considering every connected component of T' separately. But, by Lemma 11, this graph has no connected components of size greater than two (since only crosses are left, connecting more than two of them falls in Lemma 11).

We claim that we can label any remaining vertex x with $w(x)$ (the first letter appearing in column x) without introducing any new contradictions. This is clearly true for any isolated vertex in T', since these are unconstrained. Now consider x, y vertices in T' related by CROSS(a, b). Every proximal pair of positions i, j in columns x, y must have $w(i) = a$ and $w(j) = b$. But between any two columns that have proximal pairs, at least one of them has its first (top) position proximal to some position in the other column. Therefore $w(x) = a$ or $w(y) = b$ (or both). Therefore these choices satisfy the constraint matrix. If the algorithm reached this step, then there exists a p-periodic d-valid partial word contained in w, namely the one induced by ω.

Each matrix is visited at most twice (this worst case scenario is achieved precisely if the edge is examined twice in the loop starting on line 19). There are at most $2pd$ matrices in question, and analyzing a row of a matrix takes constant time. Thus, the running-time is $\mathcal{O}(pd)$ plus the running-time of checking which columns are uniform, and of constructing the constraint matrices ($\mathcal{O}(nd)$ by Corollary 7). Therefore, the total running-time of Algorithm 6 is $\mathcal{O}(nd)$. □

Let us now look at an example of how this algorithm works.

Example 2. *Let us take the full word*

$$w = acbbabcaaababbaaacbbabcaa$$

and see if it is possible to introduce holes that are not 2-proximal, such that the obtained word is 8-periodic. First we will arrange the word in rows of length eight:

$$\begin{array}{cccccccc}
a & c & b & b & a & b & c & a \\
a & a & b & a & b & b & a & a \\
a & c & b & b & a & b & c & a \\
a & & & & & & &
\end{array}$$

Now, let us look at the types of constraint matrices created by the intersections of columns. We see that after using Algorithm 4 the matrices M^{01} , M^{06}, M^{23}, M^{24}, M^{56} are of type COL; M^{12}, M^{17}, M^{35}, M^{45}, M^{67} are of type ROW; M^{02}, M^{07}, M^{57} are of type CROSS; M^{13}, M^{34}, M^{46} are of type TWO; and all the rest of the matrices are FULL ones.

Let us now run Algorithm 5 for column 0 and letter a. This determines a choice of a or c for both columns 1 and 6. Choosing c for column 1 gives us as in columns 3 and 7 and b in column 2. Running now the algorithm for letter a and column 3 we get a for column 4 and b for column 5. Since the a of column 4 determines an a in column 6, and for column 6 we have both a and c as possible solutions, the algorithm ends correctly. A 2-valid 8-periodic partial word contained in w would be

$$acb\diamond ab\diamond aa\diamond ba\diamond baaacb\diamond ab\diamond aa$$

Moreover, arranging the word in rows of length eight we now have:

$$\begin{array}{cccccccc}
a & c & b & \diamond & a & b & \diamond & a \\
a & \diamond & b & a & \diamond & b & a & a \\
a & c & b & \diamond & a & b & \diamond & a \\
a & & & & & & &
\end{array}$$

3.5 Conclusion

In this section we have presented results regarding the extension of repetition on full words, to partial words.

The concept of square (overlap, respectively, cube)-free morphisms is well defined for full words. If Berstel in 1979 and Crochemore in 1982 give characterizations of the square-free morphisms, Bean, Ehrenfeucht and McNulty investigate the k-free morphisms, in [BEM79]. Would be somehow interesting the study of such morphism from the partial words point of view.

Even more, in [BEM79] the authors also introduce the concept of so-called avoidable patterns. The study of avoidable patterns on partial words has recently been initiated in a couple of papers [BSMSW10, BSSW09], and lot of progress has been made already. In [BSMSW10] the authors extend the problem of avoiding binary patterns to partial words. Hence, they show that the classification of unavoidable binary patterns, started by Scmidt [Sch86, Sch89], continued by Roth [Rot92] and completed by Cassaigne [Cas93], is very similar to the one of partial words with infinitely many holes. The only problems arise for the case of overlaps, which, unlike the full word case, are not avoidable over binary alphabets.

Dejean in [Dej72] improves some of the original inequalities of Thue and introduces the notion of threshold repetitiveness. This concept requires that the length of a word y separating two occurrences of x is bounded from below by the length of x times some factor. Actually in this paper it is proved that for an alphabet of size 3 this threshold is $\frac{7}{4}$ and it is conjectured that for 4, the value is $\frac{7}{5}$. In fact, this result and several other bounds were proved in 1984 by Pansiot and by Moulin-Ollagnier in 1992 up to an alphabet of size 11 and for bigger size alphabets it was conjectured that this threshold is $\frac{n}{n-1}$. These other bounds were proved by Currie and Mohammad-Noori in 2004 for $12 \leq n \leq 14$ [MNC07]; Currie and Rampersad [CR09c] and Rao [Rao09] for $15 \leq n \leq 26$; Currie and Rampersad for $27 \leq n \leq 32$ [CR09b, CR09a]; and Carpi for $n \geq 33$ in [Car07]. For partial words, this topic is much simpler. Since all letters followed by a hole create a repetition of degree 2, it must be that the threshold is at least 2. For binary words, using the results from [BSMS09, HHKS09] one can see that overlaps cannot be avoided, while 2-overlaps are always avoidable. Hence, the threshold in this case is $\frac{5}{2}$. For alphabets larger than two, the result is actually proved to be 2, [BSMS09], three letters being enough for the construction of square-free partial words with infinitely many holes.

Another research direction that has been investigated lately has to do with the avoidance of large squares. A result due to Entringer, Jackson and Schatz [EJS74], says that for infinite words avoiding patterns of the form xx, the bound $|x| \geq 3$ is optimal, holds for infinite partial words containing infinitely many holes. Recently, in [BSCM09], it has been proved that the result stands for partial words with infinitely many holes as well. Moreover, a well known result of Fraenkel and Simpson [FS95], that states that for full words over a binary alphabet one can construct an infinite word containing less than four distinct squares, has also been extended to partial words. In this case it has been proved that the result holds for partial words with at most two holes, and the bound is optimal. In 1976 [Dek76], Dekking shows that there exists an infinite cube-free binary word that avoids all squares xx with $|x| \geq 4$, and that the bound of four is best possible. In the same paper of Blanchet-Sadri, Choi and Mercaş [BSCM09], it has been shown that the result holds for partial words with at most two holes, and proved that for five or more holes the length of such words is less than 18. Moreover, infinite cube-free binary partial words containing more than five holes and less than eleven squares exist, and the bound is the best possible.

From the algorithmic point of view, there are several problems that seem interesting to us, and we have not been able yet to solve efficiently nor to prove that they are intractable. First, given a k-free full word, what is the maximum number of letters that can be replaced with holes in this word, such that the partial word we obtain is also k-free? Second, given a partial word over an alphabet A, find, if any, a possibility to replace each hole with a letter from A such that the word obtained in this fashion is k-free; is there a method to compute the number of all these possibilities efficiently? Some recent work has been done in this direction [DMT09], and the results look promising.

Chapter 4

Counting Distinct Squares in Partial Words

Computing repetitions such as squares in sequences or strings of symbols from a finite alphabet is profoundly connected to numerous fields such as biology, computer science, and mathematics [Smy03]. The literature has generally considered problems in which a period u of a repetition is invariant. It has been required that occurrences of u match each other exactly. In some applications however, such as DNA sequence analysis, it becomes interesting to relax this condition and to recognize u' as an occurrence of u if u' is *compatible* with u.

Counting the number of squares in a word can be done in various ways. The counting of all squares in a full word gives quadratic results as referred to the length of the word, result that can be obtained by looking at one-letter words. A first approach, according to [Ili07], was to count the number of primitively rooted squares, squares of the form xx, where x is not an integer power of another word. In [Cro81], Crochemore proved that the number of such occurrences is $O(n \log n)$, where n is the length of the word, and the upper bound is reached by the Fibonacci words.

A well known result of Fraenkel and Simpson [FS98] states that the number of distinct squares in a word of length n is bounded by $2n$ since at each position there are at most two distinct squares whose last occurrence start. In [Ili07], Ilie improves this bound to $2n - \Theta(\log n)$. Based on numerical evidence, it has been conjectured that this number is actually less than n.

In this section, we investigate the problem of counting distinct squares in partial words, or sequences over a finite alphabet that may contain some "do not know" symbols or "holes." At first, after making some remarks about the maximum number of distinct full squares compatible with factors of a partial word, we give some lower bounds for that number. These bounds are related to the length of the word, the alphabet size this word is defined on, and the number of holes it contains. In the following part of this section, we show that for partial words with one hole, there may be more than two squares that have their last occurrence starting at the same position. We prove that if such is the case, then the hole is in the shortest square. There, we also construct for $k \geq 2$, a partial word with one hole over a k-letter alphabet that has more than k squares whose last occurrence start at position 0. Actually in [HHK09] it has been proven that the maximum number of squares starting on one position is at most $2k$, where k is the size of the alphabet. All these results are from [BSMS08], a joint work with Francine Blanchet-Sadri and Geoffrey Scott. In the last part of this chapter, representing a work done together with Francine Blanchet-Sadri [BSM09], we prove that, if it is the case that there are more than three squares that have their last occurrence starting at the same position, then the length of the shortest square is at most half the length of the third shortest square. As a result, we show that the number of distinct full squares compatible with factors of a partial word with one hole of length n is bounded by $\frac{7n}{2}$.

4.1 A first counting of distinct squares

In a full word, every factor of length $2n$ contains at most one square factor ww with $|w| = n$. In a square partial word w_0w_1 where $w_0 \uparrow w_1$, we call the word $v = w_0 \vee w_1$ the *general form* of the square. For example, the general form of the square $ab\diamond\diamond c\diamond a\diamond d\diamond\diamond\diamond$ is $abd\diamond c\diamond$. We observe that in partial words, a square w_0w_1 may be compatible with more than one distinct full square of length $2|w_0|$. For example, the word $aa\diamond aa\diamond$ over the alphabet $\{a, b, c\}$ is compatible with three distinct full squares of length 6: $(aaa)^2$, $(aab)^2$ and $(aac)^2$. It is easy to see that if $aa\diamond aa\diamond$ is a word over an alphabet of size k, then it is compatible with exactly k squares of length 6. Whenever we talk about a full square compatible with a general form, we refer to a square that has the first half compatible with the general form. In general, if $w = a_0a_1 \ldots a_{2m-1}$ is a partial word over a k-letter alphabet A that is a square, then w is compatible with exactly $k^{\|H(v)\|}$ squared full words of length m, where $v = a_0a_1 \dots a_{m-1} \vee a_ma_{m+1} \dots a_{2m-1}$.

At this point, we see that the study of distinct squares in partial words is quite different from the study of distinct squares in full words. In the case of full words, there exists an upper bound for the number of distinct squares in a word of length n, no matter what the alphabet size is. The same statement is certainly untrue for partial words. For example, the number of distinct non-empty full squares compatible with $\diamond\diamond$ is equal to k, where k is the alphabet size.

Let w be a partial word over a k-letter alphabet A. We will denote by $f_k(w)$ the number of distinct non-empty full squares over A compatible with factors of w, and by $g_{h,k}(n)$ the maximum of the $f_k(w)$'s where w ranges over all partial words over A with h holes of length n. Note that the number of all distinct full square non-empty words compatible with factors of $\diamond^n$, where n is a positive integer, over A, is equal to the number of all distinct full non-empty words of length $i \leq \lfloor \frac{n}{2} \rfloor$ over A. Using this remark,

$$g_{n,k}(n) = \sum_{i=1}^{\lfloor \frac{n}{2} \rfloor} k^i = \frac{k\ (k^{\lfloor \frac{n}{2} \rfloor} - 1)}{k-1} \tag{4.1}$$

Note that if n is odd, then $g_{n-1,k}(n-1) = g_{n,k}(n)$ and $g_{n-1,k}(n) = g_{n,k}(n)$. The first equality follows directly from (4.1). For the second equality, note that the number of distinct non-empty full squares compatible with factors of $\diamond^{n-1}a$ over the k-letter alphabet A where $a \in A$ is at least $g_{n-1,k}(n-1) = g_{n,k}(n)$ (those compatible with factors of $\diamond^{n-1}$). Thus, $g_{n-1,k}(n) \geq g_{n,k}(n)$. Since the function $g_{h,k}(n)$ is clearly monotonically increasing with respect to h, k, and n, it follows that $g_{n-1,k}(n) \leq g_{n,k}(n)$. Thus, $g_{n-1,k}(n) = g_{n,k}(n)$.

As we have seen earlier with the word $\diamond\diamond$, the number of distinct non-empty full squares compatible with factors of a partial word may be unbounded if we allow the alphabet size to grow arbitrarily large. However, we can often write this number as a function of the alphabet size. The following proposition shows that this number is indeed a polynomial in the alphabet size.

Proposition 10. *Let w be a partial word of length n over a k-letter alphabet, and let S_1 be the set of general forms of all factors of w that are squares. Let S_m be the set of all partial words v that can be written as $v = u_0 \vee u_1 \vee \dots \vee$*

u_{m-1}, where $u_i \in S_1$ for all $0 \leq i < m$ and $u_i \neq u_j$ for all $i < j < m$. Then the number of full distinct squares compatible with factors of w is given by

$$\sum_{m=1}^{\lfloor \frac{n}{2} \rfloor} ((-1)^{m-1} \sum_{s \in S_m} k^{\|H(s)\|}) \tag{4.2}$$

Proof. For a set X of partial words, denote by $\hat{X}$ the set of all full words compatible with elements of X. The number of full distinct square words compatible with factors of w is given by $\|\hat{S_1}\|$. By the principle of inclusion-exclusion,

$$\hat{S_1} = \sum_{m=1}^{\lfloor \frac{n}{2} \rfloor} ((-1)^{m-1} \sum_{s \in S_m} \|\hat{\{s\}}\|)$$

Since $\|\hat{\{s\}}\| = k^{\|H(s)\|}$, the proof is complete. □

To generalize the study of counting distinct squares in words to partial words, we are interested in the limiting behavior of $g_{h,k}(n)$ as k increases. However, as we have seen with the word $w = \diamond\diamond$, the value $\lim_{k \to \infty} f_k(w)$ may be infinity. Following Proposition 10, if we treat k as an unknown variable, the number of distinct non-empty full squares compatible with factors in any partial word is a polynomial with respect to k. If we consider all such polynomials corresponding to words of length n containing h holes, the maximal such polynomial would describe this limiting behavior. Given a finite length n, there exist only finitely many partial words of length n up to an isomorphism between letters. Therefore, a lower bound for $g_{h,k}(n)$ can be given using the leading term of this well defined maximal polynomial, $m_{h,k}(n)$.

The next results give bounds on the leading term in $m_{h,k}(n)$. We begin by defining a *free hole* of a square. Let w be a partial word over an alphabet A that contains a factor v that is a square. A hole in v is called a *free hole* of v if the square v is preserved even after we replace the hole with any letter of A. For example, consider the partial word $w = ab\diamond a\diamond\underline{\diamond}$ over the alphabet $\{a, b, c\}$. The underlined hole is a free hole of the squares $ab\diamond a\diamond\diamond$ and $\diamond\diamond$, but not of $\diamond a\diamond\diamond$. It is easy to see that the number of free holes of a square factor is exactly twice the number of holes in the general form of that square. Two free holes in positions i and j in a square v are aligned if $i = j + \frac{|v|}{2}$ or $j = i + \frac{|v|}{2}$ and $v(i) = v(j) = \diamond$.

Note that the degree of $m_{h,k}(n)$ is $\lfloor \frac{h}{2} \rfloor$. To see this, let w be a word of length n with h holes over a k-letter alphabet. Clearly, any factor of w that is a square has at most $\lfloor \frac{h}{2} \rfloor$ holes in its general form. Thus, by (4.2) there can be no term of $m_{h,k}(n)$ with k raised to a power higher than $\lfloor \frac{h}{2} \rfloor$. Also note that the word $w = \diamond^h a^{n-h}$ achieves this bound. The following technical lemma will assist us in proving results about the coefficients of $m_{h,k}(n)$.

Lemma 13. *Let l be a positive integer, let w be a partial word of length n, and let $0 \leq p_1 \leq p_2 < n$. Then there are at most $\lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1$ factors $v = w(i)w(i+1)\ldots w(i+2l-1)$ of length $2l$ in w such that $i \leq p_1$ and $i + l > p_2$.*

Proof. Assume that there exist $\lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 2$ such factors of length $2l$ in w. Since all of these factors have the same length, no two of them may start at the same position. Therefore, $p_1 \geq \lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1$. In particular, one of these factors must start at a position no later than $p_1 - (\lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1)$. This gives us that $l > ((p_2 - p_1) + \lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1)$ from the condition that $i + l > p_2$. For any factor $v = w(i)w(i+1)\ldots w(i+2l-1)$ of length $2l$ in w, we know that the length of w must exceed $2l + i$. Since there exist $\lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 2$ such factors, at least one must start at a position i satisfying $i \geq \lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1$. Therefore, we obtain the contradiction

$$n \geq 2(p_2 - p_1 + \lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 2) + \lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1$$

$$n \geq 3\lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 2(p_2 - p_1 + 1) + 3$$

$$n \geq n - 2(p_2 - p_1 + 1) - 2 + 2(p_2 - p_1 + 1) + 3$$

□

Intuitively, the above lemma states that for any $l > 0$, there can be at most $\lfloor \frac{n-2(p_2-p_1+1)}{3} \rfloor + 1$ factors of length $2l$ that use the letters $w(p_1)w(p_1 + 1)\ldots w(p_2)$ in their first half. We will use this lemma to find upper bounds for the leading term of $m_{h,k}(n)$.

Theorem 15. *The leading term in $m_{2h,k}(n)$ is $(\lfloor \frac{n-2h}{3} \rfloor + 1)k^h$.*

Proof. The degree of $m_{2h,k}(n)$ being h, it only remains to show that the coefficient of k^h in $m_{2h,k}(n)$ is equal to $\lfloor \frac{n-2h}{3} \rfloor + 1$. We will give a lower bound of this coefficient by constructing a word with the given leading term. Consider any word w of length n containing $2h$ holes and the factor

$$a^{\lfloor \frac{n-2h}{3} \rfloor} \diamond^h a^{\lfloor \frac{n-2h}{3} \rfloor} \diamond^h a^{\lfloor \frac{n-2h}{3} \rfloor}$$

The following is an exhaustive list of general forms of factors of w that are squares containing $2h$ free holes:

$$\begin{array}{ccccc}
aaa & \cdots & aa\diamond\diamond & \cdots & \diamond\diamond \\
aaa & \cdots & a\diamond\diamond\diamond & \cdots & \diamond a \\
 & & \vdots & & \\
a\diamond\diamond & \cdots & \diamond\diamond aa & \cdots & aa \\
\diamond\diamond\diamond & \cdots & \diamond aaa & \cdots & aa
\end{array}$$

These $\lfloor \frac{n-2h}{3} \rfloor + 1$ partial words are pairwise compatible, but for any words v_1, v_2 in the above list, $\|H(v_1 \vee v_2)\| < h$. Therefore, by (4.2) we see that the coefficient of k^h in $m_{2h,k}(n)$ will be at least $\lfloor \frac{n-2h}{3} \rfloor + 1$.

Note that the coefficient of k^h corresponding to a word w is equal to the number of distinct factors in w, that are squares with $2h$ free holes. Let

$$w = w_0 \diamond_0 w_1 \diamond_1 w_2 \diamond_2 \ldots \diamond_{2h-1} w_{2h}$$

where $w_i \in A^*$ for all $0 \leq i \leq 2h$ and $\diamond_i = \diamond$ for all $0 \leq i < 2h$. Note that all factors of w with $2h$ free holes that are squares must have the same length (because in a square the free hole $\diamond_0$ is aligned with $\diamond_h$, the length of all such square factors will be twice the distance between $\diamond_0$ and $\diamond_h$). We observe that all factors of w that are squares containing $2h$ free holes must contain the first h holes of w in their first half. Therefore, every such factor contains $\diamond_0 w_1 \diamond_1 \ldots \diamond_{h-1}$ in its first half. The length of $\diamond_0 w_1 \diamond_1 \ldots \diamond_{h-1}$ is at least h, so by Lemma 13, there exist at most $\lfloor \frac{n-2h}{3} \rfloor + 1$ such factors. □

Proposition 11. *The leading term in* $m_{2h+1,k}(n)$ *is at least* $(2\lfloor \frac{n-2h}{3} \rfloor + 1)k^h$.

Proof. The degree of $m_{2h+1,k}(n)$ being h, it only remains to show that the coefficient of k^h in $m_{2h+1,k}(n)$ is at least $2\lfloor \frac{n-2h}{3} \rfloor + 1$. Consider any word w of length n containing $2h+1$ holes and the factor

$$a^{\lfloor \frac{n-2h}{3} \rfloor} \diamond^h a^{\lfloor \frac{n-2h}{3} \rfloor - 1} \diamond^{h+1} a^{\lfloor \frac{n-2h}{3} \rfloor}$$

The following is an exhaustive list of general forms of factors of w that are squares containing $2h$ free holes:

$$\begin{array}{ll}
a^{\lfloor\frac{n-2h}{3}\rfloor-1}a\diamond^{h-1}\diamond & a^{\lfloor\frac{n-2h}{3}\rfloor-2}a\diamond^{h-1}\diamond \\
a^{\lfloor\frac{n-2h}{3}\rfloor-1}\diamond\diamond^{h-1}a & a^{\lfloor\frac{n-2h}{3}\rfloor-2}\diamond\diamond^{h-1}a \\
\vdots & \vdots \\
a\diamond^{h-1}\diamond a^{\lfloor\frac{n-2h}{3}\rfloor-1} & \diamond^{h-1}\diamond a a^{\lfloor\frac{n-2h}{3}\rfloor-2} \\
\diamond^{h-1}\diamond a^{\lfloor\frac{n-2h}{3}\rfloor-1}a &
\end{array}$$

There are $\lfloor\frac{n-2h}{3}\rfloor+1$ words in the left column and $\lfloor\frac{n-2h}{3}\rfloor$ words in the right column. It is easy to check that if we select two compatible words v_1, v_2 from the above list of $(2\lfloor\frac{n-2h}{3}\rfloor+1)$ partial words, $\|H(v_1 \vee v_2)\| < h$. Using (4.2) we get that the coefficient of k^h in $m_{2h+1,k}(n)$ will be at least $2\lfloor\frac{n-2h}{3}\rfloor+1$. □

Proposition 12. *The leading term in* $m_{2h+1,k}(n)$ *is at most* $(2\lfloor\frac{n-2h}{3}\rfloor+3)k^h$ *for* $h > 1$.

Proof. Let w be a word of length n containing $2h+1$ holes for some $h > 1$. Then w is of the form $w_0\diamond_0 w_1\diamond_1 w_2\diamond_2 \ldots \diamond_{2h} w_{2h+1}$ where $\diamond_i = \diamond$ for all i. We need to count the number of distinct factors of w that are squares containing $2h$ free holes. Let S denote the set of all such factors in w. Note that for every $s \in S$, there exists a hole in w that is not a free hole of s. Let S_j denote the set of all $s \in S$ having the property that $\diamond_j$ is not a free hole of s. Clearly, we have the partition $S = \cup_{0\leq j\leq 2h} S_j$.

First, assume that there exists $j \notin \{0, h, 2h\}$ such that $S_j \neq \emptyset$. Then $w_j\diamond_j w_{j+1} \uparrow w_k$ for some $j \neq k$. If there exists an i distinct from j such that $S_i \neq \emptyset$, then in one of the squares of S_i, the hole $\diamond_j$ is aligned with $\diamond_{k-1}$ or $\diamond_k$. In these cases, we get that $|w_{j+1}| \geq |w_k|$ or $|w_j| \geq |w_k|$ respectively. Both cases contradict with $w_j\diamond_j w_{j+1} \uparrow w_k$. Thus, $S_i = \emptyset$ for all $i \neq j$. Hence, we can replace $w_j\diamond w_{j+1}$ in w with w_k and preserve all squares. The resulting word has only $2h$ holes. From Theorem 15,

$$\|S\| \leq \lfloor\frac{n-2h}{3}\rfloor + 1$$

Next, let us consider the case where $S_j = \emptyset$ for every $j \notin \{0, h, 2h\}$. Note that all squares in S_0 have length equal to the distance between $\diamond_1$ and $\diamond_{h+1}$ in w, since these two holes are aligned in each square of S_0. Using the same argument, all squares in S_{2h} have length equal to the distance between $\diamond_1$ and $\diamond_{h+1}$ in w. Therefore, the length of squares in S_0 is equal to the length of the squares in S_{2h}. Note that all squares in S_0 and S_{2h} contain the factor $\diamond_1 w_2 \diamond_2 \ldots \diamond_{h-1}$ in their first half. The length of this common factor is at least $h-1$. By Lemma 13, $\|S_0 \cup S_{2h}\| \leq \lfloor \frac{n-2(h-1)}{3} \rfloor + 1 = \lfloor \frac{n-2h+5}{3} \rfloor$. Since all squares in S_h have the same length and contain the factor $\diamond_0 w_1 \diamond_1 \ldots \diamond_{h-1}$, it follows from Lemma 13 that $\|S_h\| \leq \lfloor \frac{n-2h}{3} \rfloor + 1$. Therefore,

$$\|S\| \leq \lfloor \frac{n-2h}{3} \rfloor + 1 + \lfloor \frac{n-2h+5}{3} \rfloor \leq 2\lfloor \frac{n-2h}{3} \rfloor + 3$$

The upper bound for $\|S\|$ reached in the second case is always greater than or equal to the upper bound reached in the first case. Therefore,

$$\|S\| \leq 2\lfloor \frac{n-2h}{3} \rfloor + 3$$

□

Although the following bound is trivial, it yields an interesting construction.

Proposition 13. *The leading term in $m_{3,k}(n)$ is at most $\frac{3n}{4}k$.*

Proof. Let $w = w_0 \diamond w_1 \diamond w_2 \diamond w_3$ be a partial word of length n with three holes. We wish to count the number of possible factors of w that are squares containing two free holes. Let S_1 be all such factors wherein the first hole of w is *not* free. Define S_2 and S_3 similarly. We wish to find the size of $S = \cup_{1 \leq i \leq 3} S_i$. The types of factors in S_1, S_2, and S_3 are illustrated below (the first half of each factor is written above the second half to show the alignment of the holes):

S_1	$(w_0 \diamond w_1)''$	$\diamond$	w_2'
	w_2''	$\diamond$	w_3'
S_2	w_0''	$\diamond$	$(w_1 \diamond w_2)'$
	$(w_1 \diamond w_2)''$	$\diamond$	w_3'
S_3	w_0''	$\diamond$	w_1'
	w_1''	$\diamond$	$(w_2 \diamond w_3)'$

where v' and v'' denote a prefix and suffix of a word v respectively. Because all factors in S_1 have the second and third holes of w aligned, all factors in

S_1 have the same length. Therefore, each factor in S_1 ends at a different position of $\diamond w_3$. Also, the first element of the second half of each factor in S_1 occurs at a different position of $w_2\diamond$. Therefore, $\|S_1\| \leq |w_3| + 1$ and $\|S_1\| \leq |w_2| + 1$. We can use similar reasoning to arrive at the following relations:

$$\begin{array}{lll} \|S_1\| \leq |w_2| + 1 & \|S_2\| \leq |w_0| + 1 & \|S_3\| \leq |w_0| + 1 \\ \|S_1\| \leq |w_3| + 1 & \|S_2\| \leq |w_3| + 1 & \|S_3\| \leq |w_1| + 1 \end{array}$$

Because $\|S\| = \|S_1\| + \|S_2\| + \|S_3\|$ and $n = |w_0| + |w_1| + |w_2| + |w_3| + 3$, we determine that

$$\|S\| \leq |w_2| + 1 + |w_3| + 1 + |w_1| + 1 = n - |w_0|$$

$$\|S\| \leq |w_2| + 1 + |w_3| + 1 + |w_0| + 1 = n - |w_1|$$

$$\|S\| \leq |w_3| + 1 + |w_0| + 1 + |w_1| + 1 = n - |w_2|$$

$$\|S\| \leq |w_2| + 1 + |w_0| + 1 + |w_1| + 1 = n - |w_3|$$

Therefore,

$$\|S\| \leq n - \max\{|w_0|, |w_1|, |w_2|, |w_3|\} \leq n - \lceil \frac{n-3}{4} \rceil \leq \frac{3n}{4}$$

□

As we show next, we can improve the bound for the case when there are only two holes present in the word.

Proposition 14. *If* $n \equiv 2 \bmod 6$*, then*

$$m_{2,k}(n) - (\frac{n+1}{3})k \geq \frac{n-2}{2}$$

Proof. Using Theorem 15 and the fact that $n \equiv 2 \bmod 6$, the leading term in $m_{2,k}(n)$ is $(\frac{n+1}{3})k$. Therefore, $m_{2,k}(n) - (\frac{n+1}{3})k$ is the constant term of the polynomial $m_{2,k}(n)$. It suffices to construct a partial word w with two holes over a k-letter alphabet A with $|w| = n \equiv 2 \bmod 6$ such that w contains $(\frac{n+1}{3})k + \frac{n-2}{2}$ distinct squares. Consider the word

$$w = (ab)^l \diamond (ab)^l \diamond (ab)^l$$

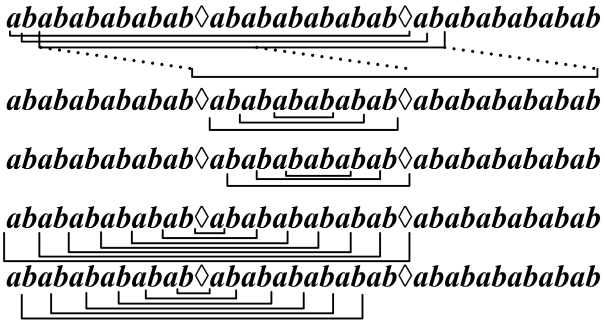
abababababab◊abababababab◊abababababab
abababababab◊abababababab◊abababababab
abababababab◊abababababab◊abababababab
abababababab◊abababababab◊abababababab
abababababab◊abababababab◊abababababab

Figure 4.1: Squares in $(ab)^6\diamond(ab)^6\diamond(ab)^6$

of length n over A, such that a, b are distinct letters of A with $l = \frac{n-2}{6}$. The following is an exhaustive list of general forms of factors of w that are squares:

$$\begin{array}{llll}
(ab)^l\diamond, & b(ab)^{l-1}\diamond a, & \ldots, & \diamond(ab)^l \\
ab, & (ab)^2, & \ldots, & (ab)^{\lfloor \frac{l}{2} \rfloor} \\
ba, & (ba)^2, & \ldots, & (ba)^{\lceil \frac{l}{2} \rceil} \\
(ab)^0 a, & (ab)^1 a, & \ldots, & (ab)^{l-1} a \\
(ba)^0 b, & (ba)^1 b, & \ldots, & (ba)^{l-1} b
\end{array}$$

Figure 4.1 illustrates these squares for $n = 38$. These general forms are pairwise incompatible. Thus, there are a total of

$$(2l+1)k + \lfloor \frac{l}{2} \rfloor + \lceil \frac{l}{2} \rceil + l + l = (\frac{n-2}{3} + 1)k + 3l = (\frac{n+1}{3})k + \frac{n-2}{2}$$

distinct full words that are squares compatible with factors of w. □

4.2 Partial words with one hole

At each position in a full word there are at most two distinct squares whose last occurrence starts, and thus the number of distinct squares in a word of length n is bounded by $2n$ as stated in the following theorem.

Theorem 16 (4)**.** *Any full word of length n has at most $2n$ distinct squares.*

A short proof of Theorem 16 is given in [5]. It follows from the unique decomposition of words into primitive ones, and synchronization (a word w

is primitive if and only if in ww there exist exactly two factors equal to w, namely the prefix and the suffix).

We now consider the one-hole case which behaves very differently from the zero-hole case. We will also count each square at the position where its last occurrence starts. If the last occurrence of a square in a partial word starts at position i, then it is a *square at position i*. In the case of partial words with one hole, there may be more than two squares that have their last occurrence starting at the same position. Such is the case with $a\diamond aababaab$ that has three squares at position 0: $a\diamond aa$, $a\diamond aaba$ and $a\diamond aababaab$. We will prove that if there are more than two squares at some position, then the hole is in the shortest square. We will also construct for $k \geq 2$, a partial word with one hole over a k-letter alphabet that has more than k squares at position 0. But first, we recall some results that will be useful for our purposes.

Lemma 14. [BB99] *Let $x, y \in A_\diamond^*$ be such that xy has at most one hole. If $xy \uparrow yx$, then there exist $z \in A^*$ and integers m, n such that $x \subset z^m$ and $y \subset z^n$.*

Lemma 15. [Ili07] *Let $w \in A^*$. If $w = z_1z_2z_3 = z_2z_3z_4 = z_3z_4z_5$ for some $z_i \in A^* \setminus \{\varepsilon\}$, then there exist $x \in A^*$ primitive and integers p, q and r, $1 \leq p \leq r < q$, such that $x = x'x''$ for some $x' \in A^*$ and $x'' \in A^* \setminus \{\varepsilon\}$, and $z_1 = x^p$, $z_2 = x^{q-r}$, $z_3 = x^{r-p}x'$, $z_4 = x''x^{p-1}x'$, and $z_5 = x''x^{q-r-1}x'$.*

Theorem 17. *If a partial word with one hole has at least three distinct squares at the same position, then the hole is in the shortest square.*

Proof. Let uu', vv' and ww' be the three shortest squares whose last occurrence start at the same position, and assume that $|w| < |v| < |u|$. It is impossible for these three squares to be all full (otherwise the subword u^2, a full word, would have three squares starting at its position 0).

For a contradiction, let us assume that ww' is full (here $w = w'$). If $w^2 \leq u$, then the prefix of length $|w^2|$ of u' is a later occurrence of a square compatible with w^2. And so we must have $v < u < w^2$. If the hole is in u' but not in v', then $v = v'$, and by replacing the hole with the corresponding letter in u, we obtain the full word u^2 that has three distinct squares at position 0, a contradiction. If the hole is in v', then set $w^2 = uz_3$, $u = vz_2$ and $v = wz_1$. We get $w = z_1z_2z_3$, $v = z_1z_2z_3z_1$ and $u = z_1z_2z_3z_1z_2$. Let

w_2 and w_3 be the prefixes of length $|w|$ of v' and u' respectively. Since z_2z_3 is a prefix of both v and v', let z_4 be such that $w, w_2 \subset z_2z_3z_4$. Note that $|z_4| = |z_1|$. Two cases occur.

Case 1. The hole is in the suffix of length $|v| - |w|$ of v'.

In this case, let z_5 be such that $w = z_3z_4z_5$. Note that $|z_5| = |z_2|$. Here $w = z_1z_2z_3 = z_2z_3z_4 = z_3z_4z_5$ and by Lemma 15, there exist $x \in A^*$ primitive and integers p, q and r, $1 \leq p \leq r < q$, such that $x = x'x''$ for some $x' \in A^*$, $x'' \in A^* \setminus \{\varepsilon\}$, and $z_1 = x^p$, $z_2 = x^{q-r}$, $z_3 = x^{r-p}x'$, $z_4 = x''x^{p-1}x'$, and $z_5 = x''x^{q-r-1}x'$. We have $w = z_1z_2z_3 = x^qx'$, $v = wz_1 = x^qx'x^p$ and $u = vz_2 = x^qx'x^px^{q-r}$. If $x' = \varepsilon$, then a later occurrence of a square compatible with w^2 exists, and so we assume that $x' \neq \varepsilon$. Since the hole is in the suffix of length $|v| - |w|$ of v', the hole is in the suffix of length $|x^p|$ of v'. We can write $v' = x^qx'x^sx_1x_2x^{p-s-1}$ where $0 \leq s < p$, $|x_1| = |x'|$ and $|x_2| = |x''|$, and where the hole is in x_1 or x_2. Since $u \uparrow u'$, we have $z_1z_2z_3z_1z_2 \uparrow z_3z_4x^sx_1x_2x^{p-s-1}\cdots$, or $x^qx'x^px^{q-r} \uparrow x^rx'x^sx_1x_2x^{p-s-1}\cdots$. The fact that $r < q$ implies that $x^{q-r}x'x^px^{q-r} \uparrow x'x^sx_1x_2x^{p-s-1}\cdots$. If $s > 0$, then $x'x''x' = x'x'x''$ and $x''x' = x'x''$, and the latter being an equation of commutativity implies that a word y exists such that $x' = y^m$ and $x'' = y^n$ for some integers m, n. In this case, there is obviously a later occurrence of a square compatible with w^2. If $s = 0$, then $x^{q-r}x'x^px^{q-r} \uparrow x'x_1x_2x^{p-1}\cdots$. Since $q > r$, by looking at the prefixes of length $|xx'|$ we get $x'x''x' \uparrow x'x_1x_2$ and deduce $x''x' \uparrow x_1x_2$.

If the hole is in x_1, then $x_2 = x''$ and $x''x' \uparrow x_1x''$. By weakening, we get $x''x_1 \uparrow x_1x''$, an equation of commutativity that satisfies the conditions of Lemma 14 since $x''x_1$ has only one hole. Similarly as above, a word y exists such that $x_1 \subset y^m$ and $x'' = y^n$ for some integers m, n. Set $x_1 = y^ty'y^{m-t-1}$ where $0 \leq t < m$ and y' is the factor that contains the hole. Since $x_1 \subset x'$, we deduce that $x' = y^ty''y^{m-t-1}$ for some y''. The compatibility $x''x' \uparrow x_1x''$ implies $y^ny^ty''y^{m-t-1} \uparrow y^ty'y^{m-t-1}y^n$ and by simplification $y^ny'' \uparrow y'y^n$. Since $x'' \neq \varepsilon$, we have $n > 0$ and obtain $y'' = y$. We get $x' = y^m$, and there is obviously a later occurrence of a square compatible with w^2. We argue similarly in the case where the hole is in x_2.

Case 2. The hole is not in the suffix of length $|v| - |w|$ of v'.

In this case, set $w = z_2z_3z_4$ and $w_2 = z_2z_3z_4'$ and the hole is in z_4'.

Also, set $w = z_3 z_4'' z_5$ and $w_3 = z_3 z_4' z_5$ where both $z_4' \subset z_4$ and $z_4' \subset z_4''$, and $|z_5| = |z_2|$. We treat the case where $z_4'' \neq z_4$ and leave the case where $z_4'' = z_4$ to the reader.

If $z_4'' = z_4$, then $w = z_1 z_2 z_3 = z_2 z_3 z_4 = z_3 z_4 z_5$ and by Lemma 15, there exist $x \in A^*$ primitive and integers p, q and r, $1 \leq p \leq r < q$, such that $x = x'x''$ for some $x' \in A^*$, $x'' \in A^* \setminus \{\varepsilon\}$, and $z_1 = x^p$, $z_2 = x^{q-r}$, $z_3 = x^{r-p}x'$, $z_4 = x''x^{p-1}x'$, and $z_5 = x''x^{q-r-1}x'$. We have $w = z_1 z_2 z_3 = x^q x'$, $v = wz_1 = x^q x' x^p$ and $u = vz_2 = x^q x' x^p x^{q-r}$. Since the hole is in z_4', we can write $v' = x^{q-p}x'(x''x')^s x_2 x_1 (x''x')^{p-s-1} x^p$ where $0 \leq s < p$, $|x_1| = |x'|$, $|x_2| = |x''|$, and where the hole is in x_1 or x_2. Since $u \uparrow u'$, we have

$$z_1 z_2 z_3 z_1 z_2 \uparrow z_3 z_4' z_1 \cdots$$

or

$$x^q x' x^p x^{q-r} \uparrow x^{r-p} x' (x''x')^s x_2 x_1 (x''x')^{p-s-1} x^p \cdots$$

or

$$x^q x' x^p x^{q-r} \uparrow x^{r-p+s} x' x_2 x_1 (x''x')^{p-s-1} x^p \cdots$$

The fact that $r - p + s < q$ implies that

$$x^{q-r+p-s} x' x^p x^{q-r} \uparrow x' x_2 x_1 (x''x')^{p-s-1} x^p \cdots$$

Since $q - r + p - s > p - s$, we get by simplification $x'x'' = x''x'$, and the latter being an equation of commutativity implies that a word y exists such that $x' = y^m$ and $x'' = y^n$ for some integers m, n. In this case, there is obviously a later occurrence of a square compatible with w^2.

If $z_4'' \neq z_4$, then put $z_1 = x^p$ where x is primitive and p is a positive integer. Since $z_1 z_2 z_3 = z_2 z_3 z_4$ and the equation $z_1(z_1 z_2 z_3) = (z_1 z_2 z_3) z_4$ is one of conjugacy, we can write $z_4 = x'' x^{p-1} x'$, where $x = x'x''$ with x'' non-empty, and $z_1 z_2 z_3 = x^q x'$ for some $q \geq p$. Since $z_1 z_2 z_3 = x^q x'$ and $z_1 = x^p$, we have $z_2 z_3 = x^{q-p} x'$. Say $z_2 = x^t y'$ where $t \geq 0$, and y' is a prefix of x with $y' \neq x$. Set $x = y'y''$ with y'' non-empty. If $y' = \varepsilon$, we have $z_2 = x^t$ and $z_3 = x^{q-p-t} x'$ and in this case $z_4'' = z_4$, a contradiction. This can be seen by using the equality $z_2 z_3 z_4 = z_3 z_4'' z_5$. And so $y' \neq \varepsilon$. Since z_4' has the length of z_1, write $z_4' = (x''x')^s x_2 x_1 (x''x')^{p-s-1}$ where $0 \leq s < p$, $|x_1| = |x'|$, $|x_2| = |x''|$, and where the hole is in x_1 or x_2. There are three cases to consider: (2.1) $t < q - p - 1$; (2.2) $t = q - p - 1$; and (2.3) $t = q - p$.

For (2.1), $z_2 = x^t y'$ and $z_3 = y'' x^{q-p-t-1} x'$. Since $z_1 z_2 z_3 = z_3 z_4'' z_5$, we have $x^q x' = y'' x^{q-p-t-1} x' \cdots = y'' x \cdots$. Since $q > p+t+1 > 0$, the prefixes of length $|x|$ are $y'y''$ and $y''y'$ respectively. The equality $y'y'' = y''y'$ holds, and so by commutativity, y' and y'' are positive powers of a common word, leading to x not being primitive, a contradiction.

For 2.2), $z_2 = x^t y'$ and $z_3 = y'' x'$. Since $z_1 z_2 z_3 = z_3 z_4'' z_5$, we have $x^q x' = y'' x' \cdots$. We consider the case where $|x'| \geq |y'|$ and then the case where $|x'| < |y'|$. If $|x'| \geq |y'|$ or y' is a prefix of x', then since $q = p + t + 1 > 0$, the prefixes of length $|x|$ are $y'y''$ and $y''y'$ respectively and again, the equality $y'y'' = y''y'$ holds, and as above leads to a contradiction. If $|x'| < |y'|$ or x' is a prefix of y', then since $z_1 z_2 z_3 \uparrow z_3 z_4' z_5$, we have $x^q x' \uparrow y'' x' (x'' x')^s x_2 x_1 (x'' x')^{p-s-1} \cdots$.

If $s > 0$, then the fact that the prefixes of length $|x|$ are compatible implies that $y'y'' = y''y'$. If $s = 0$ and the hole is in x_1, then $x_2 = x''$ and $y'' x' x'' = y'' x = y'' y' y''$ is a prefix of $z_3 z_4' z_5$ in which case $y'y'' = y''y'$ as above. If $s = 0$ and the hole is in x_2, then $x_1 = x'$ and set $y' = x'y$ for some $y \neq \varepsilon$. Here, $x'' = yy''$, and put $x_2 = y_1 y_2$ where $y_1 \subset y$ and $y_2 \subset y''$. We get $x^q x' \uparrow y'' x' x_2 x_1 (x'' x')^{p-1} \cdots = y'' x' y_1 y_2 x' (x'' x')^{p-1} \cdots$.

If the hole is in y_2, then $y_1 = y$ and $y'' x' y_1 = y'' x' y = y'' y'$ is a prefix of $z_3 z_4' z_5$ and the result again follows since $y'y'' = y''y'$. If the hole is in y_1, then $y'y'' \uparrow y'' x' y_1$ or $x' y y'' \uparrow y'' x' y_1$, and by weakening $(x' y_1) y'' \uparrow y'' (x' y_1)$. The latter being an equation of commutativity, by Lemma 14, we get that $x' y_1 \subset z^m$ and $y'' = z^n$ for some word z and positive integers m, n. Set $x' y_1 = z^k z' z^{m-k-1}$ where $0 \leq k < m$ and z' is the factor that contains the hole. Since $x' y_1 \subset x' y$, we deduce that $x' y = z^k z'' z^{m-k-1}$ for some z''. The compatibility $x' y y'' \uparrow y'' x' y_1$ implies $z^k z'' z^{m-k-1} z^n \uparrow z^n z^k z' z^{m-k-1}$. By simplification we obtain $z'' z^n \uparrow z^n z'$, and since $n > 0$ we get $z'' = z$, and thus $y' = x' y = z^m$. The result follows since $x = y'y'' = z^{m+n}$ with $m + n > 1$.

For (2.3), $z_2 = x^t y'$ and $z_3 = y \neq \varepsilon$. Here $x' = y'y$, and so $x = x'x'' = y' y x''$ and $y'' = yx''$. Here

$$x^q x' = z_1 z_2 z_3 \uparrow z_3 z_4' z_5 = y (x'' x')^s x_2 x_1 (x'' x')^{p-s-1} \cdots .$$

If $s > 0$, then $x^q x' = x' x'' y' \cdots = y' y'' y' \cdots = y' y x'' y' \cdots$ and $y (x'' x')^s \cdots = y x'' x' \cdots = y x'' y' y \cdots$, and so $y' y x'' = y x'' y'$ or $y'y'' = y''y'$, again leading to a contradiction.

If $s = 0$ and the hole is in x_1, then $x_2 = x''$ and set $x_1 = y_2 y_1$ where $y_2 \subset y'$ and $y_1 \subset y$. We get $y'y'' \cdots = x^q x' \uparrow yx_2x_1 \cdots = yx''y_2y_1 \cdots = y''y_2y_1 \cdots$. If the hole is in y_1, then $y_2 = y'$ and $y''y'$ is a prefix of $z_3z_4'z_5$ and the result again follows since $y'y''$ is a prefix of $x^q x'$. If the hole is in y_2, then $y'y'' \uparrow y''y_2$, and by weakening $y_2y'' \uparrow y''y_2$. By Lemma 14, a word z exists such that $y_2 \subset z^m$ and $y'' = z^n$ for some positive integers m, n. Set $y_2 = z^k z' z^{m-k-1}$ where $0 \leq k < m$ and z' is the factor that contains the hole. Since $y_2 \subset y'$, we deduce that $y' = z^k z'' z^{m-k-1}$ for some z''. The compatibility $y'y'' \uparrow y''y_2$ implies $z^k z'' z^{m-k-1} z^n \uparrow z^n z^k z' z^{m-k-1}$. Since $n > 0$, by simplification we obtain $z'' = z$. We get $y' = z^m$ and x is not primitive.

If $s = 0$ and the hole is in x_2, then $x_1 = x'$ and $y'yx''y'y \cdots = x'x''x' \cdots = x^q x' \uparrow yx_2x_1 \cdots = yx_2x' \cdots = yx_2y'y \cdots$. We deduce $y'yx'' \uparrow yx_2y'$, and by weakening $y'yx_2 \uparrow yx_2y'$. By Lemma 14, we get that $yx_2 \subset z^m$ and $y' = z^n$ for some word z and positive integers m, n. Set $yx_2 = z^k z' z^{m-k-1}$ where $0 \leq k < m$ and z' is the factor that contains the hole. Since $yx_2 \subset yx''$, we deduce that $yx'' = z^k z'' z^{m-k-1}$ for some z''. The compatibility $y'yx'' \uparrow yx_2y'$ implies $z^n z^k z'' z^{m-k-1} \uparrow z^k z' z^{m-k-1} z^n$. Since $n > 0$, by simplification we obtain $z'' = z$. We get $y'' = yx'' = z^m$ and the result follows since $x = y'y'' = z^{m+n}$ with $m + n > 1$. □

Proposition 15. *For $k \geq 2$, there exists a partial word with one hole over a k-letter alphabet that has more than k squares at position 0.*

Proof. Let $\Sigma = \{a_1, a_2, \ldots\}$ be an infinite ordered set. We build a sequence of partial words with one hole, $(DS_i)_{i \geq 2}$, where DS_i contains $i + 1$ squares with their last occurrence starting at position 0. In order to do this, we build an intermediary sequence of partial words with one hole $(DS_i')_{i \geq 2}$ and denote by $DS_i'(a)$, the word DS_i' in which the hole has been replaced by the letter a. Let $DS_2 = a_1 \diamond a_1 a_1 a_2 a_1 a_2 a_1 a_1 a_2$, and for $i \geq 3$,

$$\begin{array}{lcl} DS_{i-1}' & = & DS_{i-1} a_{i-1} \\ DS_i & = & DS_{i-1}' DS_{i-1}'(a_i) \end{array}$$

In other words, DS_i consists of the concatenation of DS_{i-1} with the last letter of the smallest alphabet used for creating DS_{i-1}, concatenated again with the same factor in which the hole has been replaced by a letter not present in the word so far. For example,

$$\begin{aligned} DS_2' &= a_1\diamond a_1a_1a_2a_1a_2a_1a_1a_2a_2 \\ DS_3 &= a_1\diamond a_1a_1a_2a_1a_2a_1a_1a_2a_2a_1a_3a_1a_1a_2a_1a_2a_1a_1a_2a_2 \end{aligned}$$

the latter having three squares other than itself at position 0: $a_1\diamond a_1a_1$, $a_1\diamond a_1a_1a_2a_1$ and $a_1\diamond a_1a_1a_2a_1a_2a_1a_1a_2$. For $k \geq 2$, DS_k, a partial word with one hole over a k-letter alphabet, has $k+1$ squares. This is due to the fact that all previous squares cannot reappear later in the word because of the newly introduced letter. □

The concept is furthermore investigated in [HHK09]. Here, the authors extend the bound on the number of full squares that have their last compatible occurrences starting at a certain position in a partial words with one hole. Using the pigeonhole principle, they show that the bound goes from 2, in the case of full words, to $2k$, where k is the size of the alphabet the word is defined on.

Proposition 16. [HHK09] *Let w be a partial word defined over an alphabet of size k, such that w contains only one hole. Then the number of full squares that have their last compatible occurrences in w starting at a position i is $2k$.*

This result is actually an improvement of Proposition 15, an example being straightforward:

Example 3. [HHK09] *Construct a partial word w such that the number of full squares that have their last compatible occurrences in w starting at a position 0 is $2k$. Let $A = \{a_1, \dots, a_k\}$ and for $w = uv$ define $u^{-1}w = v$ the left quotient of w by u (if u is not a prefix of w, then $u^{-1}w$ is undefined), and similarly define the right quotient wu^{-1}. Now, we define $w = w_{2k}$ recursively. Let $w_0 = \diamond a_k a_{k-1} \cdots a_1$ and, for $j \in \{1, 2, \dots, k\}$, set*

$$w_{2j-1} = w_{2j} - 2w_{2j-2}(aj)$$

$$w_2j = w_{2j-1}(\diamond^{-1}w_{2j-1})a_j^{-1}$$

For $w = w_{2k}$, we have $2k$ full squares that have their last compatible occurrences in w starting at a position 0.

4.3 Bound on the number of squares

In this section, we prove that, in a partial word with one hole, if three squares have their last occurrence starting at the same position, then the length of the shortest square is at most half the length of the third shortest square. As a result, we show that the number of distinct full squares compatible with factors of a partial word with one hole of length n is bounded by $\frac{7n}{2}$. But first, we recall some lemmas that will be useful for our purposes.

Fundamental results on periodicity of full words include a theorem of Fine and Wilf which considers the simultaneous occurrence of different periods in a word. The following lemma extends this result to partial words with one hole.

Lemma 16. [BB99] *Let $w \in A_\diamond^*$ be weakly p-periodic and weakly q-periodic. If $H(w)$ is a singleton and $|w| \geq p+q$, then w is (strongly) $\gcd(p,q)$-periodic.*

The following lemmas on commutativity and conjugacy will be useful for our purposes.

Lemma 17. [BB99] *Let $x, y \in A^+$ and let $z \in A_\diamond^*$ be such that $H(z)$ is a singleton. If $z \subset xy$ and $z \subset yx$, then $xy = yx$.*

Lemma 18. [BSBL06] *Let $x, y, z \in A_\diamond^*$ be such that $|x| = |y| > 0$. Then $xz \uparrow zy$ if and only if xzy is weakly $|x|$-periodic.*

Lemma 19. [BSL02] *Let $x, y \in A_\diamond^+$ and $z \in A^*$. If $xz \uparrow zy$, then there exist $v, w \in A^*$ and an integer $n \geq 0$ such that $x \subset vw$, $y \subset wv$, and $z = (vw)^n v$. Consequently, if $xz \uparrow zy$, then xzy is (strongly) $|x|$-periodic.*

Theorem 18. *Let ww', vv' and uu' be three squares at the same position, with $|w| < |v| < |u|$. If $H(uu')$ is a singleton, then $|ww'| \leq |u|$.*

Proof. Since $|w| < |v| < |u|$, let us denote $v = wz_1$ and $u = vz_2$, for some partial words z_1, z_2 over the alphabet A. By contradiction, let us assume that $|ww'| > |u|$, and denote $ww' = uz_3$, where $z_3 \in A_\diamond^*$. According to Theorem 17, the hole is in ww'. We have $w' = z_1z_2z_3$, $w = z_1'z_2'z_3'$, $v = z_1'z_2'z_3'z_1$ and $u = z_1'z_2'z_3'z_1z_2$, where $z_i' \uparrow z_i$ for all $i \in \{1,2,3\}$. Since $v \uparrow v'$, we get that there exists $z_4 \in A^*$ such that $z_2z_3z_4$ is a prefix of v' and $|z_4| = |z_1|$, and by looking at the prefixes of length $|w|$ of u and u', we get that there exists $z_5 \in A^*$, with $|z_5| = |z_2|$, such that $z_1'z_2'z_3' \uparrow z_3z_4z_5$ (see

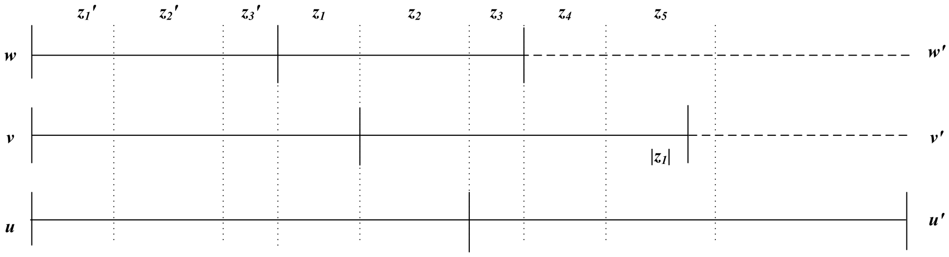

Figure 4.2: The case when $|ww'| > |u|$

Figure 4.2). There are six cases to consider: Case 1 (respectively, Case 2, Case 3, Case 4, Case 5, Case 6) where the hole is in z_3 (respectively, z_2, z_1, z_3', z_2', z_1').

Case 1. The hole is in z_3.

We have $w' = z_1z_2z_3$, $w = z_1z_2z_3'$, $v = z_1z_2z_3'z_1$ and $u = z_1z_2z_3'z_1z_2$ where $z_3 \subset z_3'$. Since $z_1z_2z_3' \uparrow z_2z_3z_4$, we get $z_1z_2z_3 \uparrow z_2z_3z_4$ by weakening. By Lemma 18, we get

$$z_1z_2z_3z_4 \text{ is weakly } |z_1|\text{-periodic} \tag{4.3}$$

Now, since w is full, the prefixes of length $|w|$ of v and v', and respectively of u and u' are compatible, and $z_2z_3z_4 \subset w$ and $z_3z_4z_5 \subset w$, we get $z_2z_3z_4 \uparrow z_3z_4z_5$. Using Lemma 18 again, we get

$$z_2z_3z_4z_5 \text{ is weakly } |z_2|\text{-periodic} \tag{4.4}$$

Finally, applying the weakening rule for the prefixes of length $|w|$ of u and u', we get $z_1z_2z_3 \uparrow z_3z_4z_5$. After using Lemma 18, we get

$$z_1z_2z_3z_4z_5 \text{ is weakly } |z_1z_2|\text{-periodic} \tag{4.5}$$

From (4.3) and (4.5) we get that $z_1z_2z_3z_4$ is weakly $|z_1|$- and weakly $|z_1z_2|$-periodic. Applying Lemma 16, we get that $z_1z_2z_3z_4$ is $\gcd(|z_1|, |z_1z_2|)$-periodic. Hence there exists a word $x \in A^*$ of length $\gcd(|z_1|, |z_1z_2|)$, such that $z_1 = x^m$ and $z_1z_2 = x^{m+n}$ for some integers $m, n > 0$.

From (4.4) and (4.5) we get that $z_2z_3z_4z_5$ is weakly $|z_2|$- and weakly $|z_1z_2|$-periodic. Applying Lemma 16, we get that $z_2z_3z_4z_5$ is $\gcd(|z_2|, |z_1z_2|)$-

periodic. Since $\gcd(|z_1|, |z_1z_2|) = \gcd(|z_2|, |z_1z_2|)$, we get that $z_2z_3z_4z_5$ is $|x|$-periodic. Because $|z_1| \geq |x|$ and $|z_2| \geq |x|$ we get that $z_1z_2z_3z_4z_5$ is $|x|$-periodic.

Because z_1 and z_5 share a prefix of length $\min(|x^m|, |x^n|)$ with $m, n > 0$, z_5 is $|x|$-periodic and $|z_5| = |x^n|$, we get that $z_5 = x^n = z_2$. Since $z_3z_4z_5$ is $|x|$-periodic, $|z_5| \geq |x|$ and $|z_4| = |x^m|$, we get that $z_4 = x^m = z_1$.

Since $z_1z_2z_3z_4$ is $|x|$-periodic and $z_1z_2 = x^{m+n}$, it results that $z_3 \subset (x'x'')^px'$ and $z_4 = (x''x')^m$ where $x = x'x''$ and $p \geq 0$ is an integer. But $z_4 = z_1 = x^m$. Hence, $x'x'' = x''x'$ and there exists a full word y, such that $x' = y^q$ and $x'' = y^r$ for some integers $q, r \geq 0$.

Since $v' \uparrow v$, we have that $z_2z_3z_1z_1 \uparrow z_1z_2z_3'z_1$. By cancellation, we get $z_2z_3z_1 \uparrow z_1z_2z_3'$. Replacing z_1 by x^m and z_2 by x^n, we get $x^nz_3x^m \uparrow x^mx^nz_3'$, and consequently $z_3x^m \uparrow x^mz_3'$ by cancellation. By Lemma 19, there exist full words y', y'' such that $z_3 \subset y'y''$, $z_3' = y''y'$, and $x^m = (y'y'')^ry'$ for some integer $r \geq 0$. By Lemma 17, since $z_3 \subset y'y''$ and $z_3 \subset z_3' = y''y'$, we get $y'y'' = y''y'$. The latter implies that there exists a full word z such that y' and y'' are powers of z. We obtain $x^m = z^{m'}$ for some integer m', and x and z are hence powers of a common word z'. We conclude that z_1, z_2, z_3, z_3', z_4 and z_5 are contained in powers of z', implying that there is a later occurrence of a square compatible with w^2.

Case 2. The hole is in z_2.

Hence, $w' = z_1z_2z_3$, $w = z_1z_2'z_3$, where $z_2' \uparrow z_2$, $v = z_1z_2'z_3z_1$ and $u = z_1z_2'z_3z_1z_2$.

Since $z_1z_2'z_3 \uparrow z_2z_3z_4$, we get $z_1z_2z_3 \uparrow z_2z_3z_4$ by applying weakening. Using Lemma 18, we get

$$z_1z_2z_3z_4 \text{ is weakly } |z_1|\text{-periodic} \tag{4.6}$$

Now, since $w' \subset w$, by looking at the prefixes of length $|w|$, $z_2z_3z_4 \subset w$ of v' and $z_3z_4z_5 \subset w$ of u', we get $z_2z_3z_4 \uparrow z_3z_4z_5$. Using Lemma 19, we get

$$z_2z_3z_4z_5 \text{ is } |z_2|\text{-periodic} \tag{4.7}$$

Finally, for the prefixes of length $|w|$ of u and u', we have $z_1z_2'z_3 = z_3z_4z_5$.

Using Lemma 19, it results that

$$z_1z_2'z_3z_4z_5 \text{ is } |z_1z_2|\text{-periodic} \tag{4.8}$$

From (4.6) and (4.8) we get that $z_1z_2z_3z_4$ is weakly $|z_1|$- and $|z_1z_2|$-periodic. Applying Lemma 16, we have $z_1z_2z_3z_4$ $\gcd(|z_1|, |z_1z_2|)$-periodic. Hence, there exists a word $x \in A^*$ of length $\gcd(|z_1|, |z_1z_2|)$, such that $z_1 = x^m$ and $z_1z_2 \subset x^{m+n}$, for some integers $m, n > 0$.

From (4.7) and (4.8) we get that $z_2z_3z_4z_5$ is $|z_2|$- and $|z_1z_2|$-periodic. Applying Lemma 16, we get that $z_2z_3z_4z_5$ is $\gcd(|z_2|, |z_1z_2|)$-periodic. It follows that $z_1z_2z_3z_4z_5$ is $|x|$-periodic.

Because z_1 and z_5 share a prefix of length $\min(|x^m|, |x^n|)$, and $|z_5| = |x^n|$, we get that $z_5 = x^n$. Since $z_3z_4z_5$ is $|x|$-periodic, $|z_5| \geq |x|$ and $|z_4| = |x^m|$, we get that $z_4 = x^m = z_1$.

Since $z_1z_2z_3$ is $|x|$-periodic, it results that $z_3 = (x'x'')^px'$ with $x = x'x''$ and some integer $p \geq 0$. By looking at the prefixes of length $|w|$ of u and u', we notice that $z_1z_2'z_3 = z_3z_1z_5$. This implies that $z_1z_2'z_3z_1z_5$ is $|z_1z_2'|$-periodic, and so $z_3z_1z_5 = (x'x'')^px'(x'x'')^m(x'x'')^n$ is $|z_1z_2'|$-periodic. Hence, by looking at the suffix of length $|x|$ and the corresponding factor starting $|z_1z_2'|$ positions before, we get that $x'x'' = x''x'$. Results that there exist integers q, r with $q, r \geq 0$ and a word y such that $x' = y^q$ and $x'' = y^r$. But since $z_1z_2'z_3 = z_3z_1z_5$ we get that $y^{m(q+r)}z_2'y^{p(q+r)+q} = y^{(q+r)(m+n+p)+q}$. Hence, $z_2' = y^{n(q+r)}$.

We get, again, another factor compatible with $y^{2(q+r)(m+n+p)+2q} \supset ww'$, starting with position $|y|$, which is a contradiction.

Case 3. The hole is in z_1.

In this case we have that $w' = z_1z_2z_3$, $w = z_1'z_2z_3$, with $z_1' \uparrow z_1$, $v = z_1'z_2z_3z_1$ and $u = z_1'z_2z_3z_1z_2$. Also, $v' = z_2z_3z_4z_1''$, where $z_1 \uparrow z_1''$.

Since $z_1'z_2z_3 = z_2z_3z_4$, we get by Lemma 19 that

$$z_1'z_2z_3z_4 \text{ is } |z_1|\text{-periodic} \tag{4.9}$$

Now, looking at the prefixes of length $|w|$, $z_1'z_2z_3$ of u and $z_3z_4z_5$ of u', we get $z_1'z_2z_3 = z_3z_4z_5$. Hence

$$z_1'z_2z_3z_4z_5 \text{ is } |z_1z_2|\text{-periodic} \tag{4.10}$$

Finally, for the prefixes of length $|w|$ of v' and u', we have $z_2z_3z_4 = z'_1z_2z_3 = z_3z_4z_5$. This implies that

$$z_2z_3z_4z_5 \text{ is } |z_2|\text{-periodic} \tag{4.11}$$

From (4.9) and (4.10) we get that $z'_1z_2z_3z_4$ is $\gcd(|z_1|, |z_1z_2|)$-periodic. Hence there exists a word $x \in A^*$ such that $z_1 \subset z'_1 = x^m$ and $z_1z_2 \subset z'_1z_2 = x^{m+n}$ for some integers $m, n > 0$ and $|x| = \gcd(|z_1|, |z_1z_2|)$. In addition, $z_2 = x^n$, $z_3 = (x'x'')^px'$, $z_4 = (x''x')^m$ where $p \geq 0$ is an integer and $x = x'x''$.

From (4.10) and (4.11) we get that $z_2z_3z_4z_5$ is $\gcd(|z_2|, |z_1z_2|)$-periodic. Since $\gcd(|z_1|, |z_1z_2|) = \gcd(|z_2|, |z_1z_2|)$, we get that $z_2z_3z_4z_5$ is $|x|$-periodic. And so $z_5 = (x''x')^n$.

Using the simplification rule for $v \uparrow v'$, we get that there exists z''_1 such that $z_1 \uparrow z''_1$ and z''_1 shares a prefix of length $\min(|x^m|, |x^n|)$ with z_5. Hence, we get that either $x'x'' = x''x'$, $x_1x'' \uparrow x''x'$ or $x'x_2 \uparrow x''x'$, where the hole is in x_1 or x_2, and $x_1 \subset x'$ and $x_2 \subset x''$. Following the previous cases, the first option leads us to a later occurrence of a factor compatible with the full square w^2.

Let us now consider the case where $x_1x'' \uparrow x''x'$ (the other case is similar). By Lemma 19, there exist full words y', y'' such that $x_1 \subset y'y''$, $x' = y''y'$, and $x'' = (y'y'')^ry'$ for some integer $r \geq 0$. By Lemma 17, since $x_1 \subset y'y''$ and $x_1 \subset x' = y''y'$, we get $y'y'' = y''y'$. The latter implies that there exists a full word z such that y' and y'' are powers of z. We obtain $x^m = z^{m'}$ for some integer m', and x and z are hence powers of a common word z'. We conclude that z_1, z'_1, z_2, z_3, z_4 and z_5 are contained in powers of z', implying that there is a later occurrence of a square compatible with w^2.

Case 4. The hole is in z'_3.

Looking at the prefixes of length $|w|$ of v and v', we have $z_1z_2z'_3 \uparrow z_2z_3z_4$. Applying weakening and Lemma 18, we get that $z_1z_2z'_3z_4$ is weakly $|z_1|$-periodic. Also, by looking at the prefixes of length $|w|$ of u and u' we get that $z_1z_2z'_3 \uparrow z_3z_4z_5$. We apply weakening and Lemma 18 again, and get that $z_1z_2z'_3z_4z_5$ is weakly $|z_1z_2|$-periodic. Using Lemma 16, it follows that $z_1z_2z'_3z_4$ is $\gcd(|z_1|, |z_1z_2|)$-periodic. Hence, there exists x such that $z_1 = x^m$ and $z_1z_2 = x^{m+n}$, for some positive integers m, n, with $|x| = \gcd(|z_1|, |z_1z_2|)$.

If we denote $x = x'x''$, it results that $z'_3 \subset (x'x'')^p x'$, for some integer $p \geq 0$, and $z_4 = (x''x')^m$.

Since the hole is in z'_3 then, either $z'_3 = (x'x'')^{p_1} x'_1 (x''x')^{p_2}$, where $p_1 + p_2 = p$ and x'_1 has a hole, or $z'_3 = (x'x'')^{p_1} x'x'_2 (x'x'')^{p_2} x'$, where $p_1+p_2+1 = p$ and x'_2 has a hole. Because $z'_3 \subset z_3$, it implies that $z_3 = (x'x'')^{p_1} x_1 (x''x')^{p_2}$ where $x'_1 \subset x_1$, or $z_3 = (x'x'')^{p_1} x'x_2 (x'x'')^{p_2} x'$ where $x'_2 \subset x_2$. But also, $z_1 z_2 z'_3 \uparrow z_2 z_3 z_4$. Hence, we get that $x^m z'_3 \uparrow z_3 z_4$. This is equivalent to one of the following cases: $x^m (x'x'')^{p_1} x'_1 (x''x')^{p_2} \uparrow (x'x'')^{p_1} x_1 (x''x')^{p_2} (x''x')^m$ when we get $x_1 = x'$, or $x^m (x'x'')^{p_1} x'x'_2 (x'x'')^{p_2} x' \uparrow (x'x'')^{p_1} x'x_2 (x'x'')^{p_2} x' (x''x')^m$ when we get $x_2 = x''$. In either case, $z_3 = (x'x'')^p x'$.

Since $z_1 z_2 z'_3 \uparrow z_3 z_4 z_5$, there is the possibility that $z_5 = (x''x')^n$ if $n \leq p_2$. We leave this case to the reader and assume that $n > p_2$. We get that $z_5 = (x''x')^{n_1} x'' x_1 (x''x')^{n_2}$ with $x'_1 \subset x_1$, or $z_5 = (x''x')^{n_1} x_2 x' (x''x')^{n_2}$ with $x'_2 \subset x_2$ (in either case $n_1 + n_2 + 1 = n$). Since $v \uparrow v'$, it follows that z_5 and z_1 share a prefix of length $|x|$, and so z_5 has $x'x''$ as a prefix. There are three cases to consider: (4.1) $x'x'' = x''x'$; (4.2) $x'x'' = x''x_1$; and (4.3) $x'x'' = x_2 x'$. For (4.1), there exists a full word y such that x' and x'' are powers of y. It follows that z_1, z_2, z_3, z'_3 and z_4 are contained in powers of y, implying that there is a later occurrence of a partial word that is compatible with the full square $(w')^2$. For (4.2) and (4.3), $n_1 = 0$ and we can denote z_5 as $x'x''(x''x')^{n-1}$. Furthermore, since $z_1 z_2 z'_3 \uparrow z_3 z_4 z_5$ we get that either $x^{m+n+p_1} x'_1 (x''x')^{p_2} \uparrow x^{m+p} x'x'x''(x''x')^{n-1}$ or $x^{m+n+p_1} x'x'_2 (x'x'')^{p_2} x' \uparrow x^{m+p} x'x'x''(x''x')^{n-1}$. We prove the first case (the other is similar).

If $p > n + p_1$, then $p_2 > n$ a contradiction. If $p = n + p_1$, then $x'_1 x'' x' \uparrow x'x'x''$ and $x'x'' = x''x'$, the same contradiction as before follows. If $p < n + p_1$, we get $x^{n-p_2} x'_1 \uparrow x'x'x''(x''x')^{n-1-p_2}$. If $p_2 < n - 1$, then again $x'x'' = x''x'$. If $p_2 = n - 1$, then $x'x'' \uparrow x''x'_1$. By Lemma 19, there exist full words y', y'' such that $x' = y'y''$, $x'_1 \subset y''y'$, and $x'' = (y'y'')^r y'$ for some integer $r \geq 0$. By Lemma 17, since $x'_1 \subset y''y'$ and $x'_1 \subset x' = y'y''$, we get $y'y'' = y''y'$. The latter implies that there exists a full word z such that y' and y'' are powers of z. We obtain x' and x'' are powers of z. We conclude that z_1, z_2, z_3, z'_3, z_4 and z_5 are contained in powers of z, implying that there is a later occurrence of a square compatible with $(w')^2$.

Case 5. The hole is in z'_2.

Looking at the prefixes of length $|w|$ of v and v', we have $z_1 z'_2 z_3 \uparrow z_2 z_3 z_4$.

Applying weakening and Lemma 18, we get that $z_1z'_2z_3z_4$ is weakly $|z_1|$-periodic. Also, by looking at the prefixes of length $|w|$ of u and u' we get that $z_1z'_2z_3 \uparrow z_3z_4z_5$. By applying Lemma 19, we get that $z_1z'_2z_3z_4z_5$ is $|z_1z_2|$-periodic. Using Lemma 16, it follows that $z_1z'_2z_3z_4$ is $\gcd(|z_1|, |z_1z_2|)$-periodic. Hence, there exists x, such that $z_1 = x^m$ and $z_1z'_2 \subset x^{m+n}$, for some positive integers m, n and $|x| = \gcd(|z_1|, |z_1z_2|)$. Hence we have $z_3 = (x'x'')^p x'$ and $z_4 = (x''x')^m$, where $x = x'x''$ and $p \geq 0$.

Since the hole is in z'_2 then, either $z'_2 = (x'x'')^{n_1}x'_1x''(x'x'')^{n_2}$ where x'_1 has a hole, or $z'_2 = (x'x'')^{n_1}x'x'_2(x'x'')^{n_2}$ where x'_2 has a hole (in either case $n_1+n_1+1 = n$). Because $z'_2 \subset z_2$, it implies that $z_2 = (x'x'')^{n_1}x_1x''(x'x'')^{n_2}$ where $x'_1 \subset x_1$, or $z_2 = (x'x'')^{n_1}x'x_2(x'x'')^{n_2}$ where $x'_2 \subset x_2$. But also, $z_1z'_2z_3 \uparrow z_2z_3z_4$. This is equivalent to one of the following cases:

$$x^m(x'x'')^{n_1}x'_1x''(x'x'')^{n_2}(x'x'')^p x' \uparrow (x'x'')^{n_1}x_1x''(x'x'')^{n_2}(x'x'')^p x'(x''x')^m$$

when we get $x_1 = x'$, or

$$x^m(x'x'')^{n_1}x'x'_2(x'x'')^{n_2}(x'x'')^p x' \uparrow (x'x'')^{n_1}x'x'_2(x'x'')^{n_2}(x'x'')^p x'(x''x')^m$$

when we get $x_2 = x''$. In either case, $z_2 = (x'x'')^n$.

We treat the second case (the other is similar). Since $z_1z'_2z_3 \uparrow z_3z_4z_5$, there is the possibility that $z_5 = (x''x')^n$ if $n \leq n_2 + p$. We leave this case to the reader and assume that we have $n > n_2 + p$. We get that $z_5 = (x''x')^{n_1-p}x_2x'(x''x')^{n_2+p}$ with $x'_2 \subset x_2$. Since $v \uparrow v'$, it follows that z_5 and z_1 share a prefix of length $|x|$, and so z_5 has $x'x''$ as a prefix. There are two cases to consider: (5.1) $x'x'' = x''x'$; and (5.2) $x'x'' = x_2x'$. For (5.1), there exists a full word y such that x' and x'' are powers of y. It follows that z_1, z_2, z'_2, z_3 and z_4 are contained in powers of y, implying that there is a later occurrence of a partial word that is compatible with the full square $(w')^2$. For (5.2), $n_1 = p$ and we can denote z_5 as $x'x''(x''x')^{n-1}$. Furthermore, since $z_1z'_2z_3 \uparrow z_3z_4z_5$ we get that $x^{n_1-p}x'_1(x''x')^{n_2+p+1} \uparrow x'x'x''(x''x')^{n-1}$.

Since $n > n_2 + p$, we have $n_1 \geq p$. If $n_1 = p$ or $n_1 \geq p+2$, then $x'x'' = x''x'$. If $n_1 = p+1$, then $x'x'' \uparrow x''x'_1$. By Lemma 19, there exist full words y', y'' such that $x' = y'y''$, $x'_1 \subset y''y'$, and $x'' = (y'y'')^r y'$ for some integer $r \geq 0$. By Lemma 17, since $x'_1 \subset y''y'$ and $x'_1 \subset x' = y'y''$, we get $y'y'' = y''y'$. The latter implies that there exists a full word z such that y' and y'' are powers of z. We obtain x' and x'' are powers of z. We conclude that z_1, z_2, z_2', z_3, z_4 and z_5 are contained in powers of z, implying that there is a later occurrence of a square compatible with $(w')^2$.

Case 6. The hole is in z_1'.

Since $v \uparrow v'$, we have that $z_1'z_2z_3 \uparrow z_2z_3z_4$. Applying Lemma 19, we get that $z_1'z_2z_3z_4$ is $|z_1|$-periodic. Since $u \uparrow u'$, we get that $z_1'z_2z_3 \uparrow z_3z_4z_5$, and $z_1'z_2z_3z_4z_5$ is $|z_1z_2|$-periodic. Hence, $z_1'z_2z_3z_4$ is $|x| = \gcd(|z_1|, |z_1z_2|)$-periodic, where $z_1' \subset x^m$ and $z_2 = x^n$, for some word x and integers $m, n > 0$. This implies that there exist $x', x'' \in A^*$, such that $x = x'x''$, and $z_3 = (x'x'')^p x'$, for some integer $p \geq 0$, $z_4 = (x''x')^m$ and $z_5 = (x''x')^n$ (because $z_2z_3z_4z_5$ is $|z_1z_2|$-periodic and $|z_2z_3z_4| > |z_1z_2|$).

Since $v \uparrow v'$, if $|z_2| \geq |z_1|$, then z_5 and z_1 share a common prefix of length $|x^m|$. It follows that $z_1 = (x''x')^m$. But, since $z_1' \subset z_1$, it results that $z_1' \subset (x''x')^m$, and recall that $z_1' \subset (x'x'')^m$. If $m > 1$, then we get $x'x'' = x''x'$ and so there exists y, such that $x' = y^q$ and $x'' = y^r$ for some non-negative integers q, r, giving us a contradiction with the assumption that there is no later occurrence of a factor compatible with $(w')^2$. If $m = 1$, then we also get $x'x'' = x''x'$ by Lemma 17. Hence, we may assume that $|z_2| < |z_1|$ and z_1 has as a prefix $(x''x')^n$. Let $z_6 \in A^*$, where $|z_6| = |z_2|$, such that $z_1'z_2z_3z_1 \uparrow z_3z_4z_1z_6$ (the prefixes of length $|v|$ of u and u' are compatible). By using simplification we get that $z_2x'z_1 \uparrow x'z_1z_6$ and $z_2x'z_1z_6$ is $|z_2|$-periodic by Lemma 19. Since $|z_1| = |x^m| > |z_2|$, it follows that $z_1 = (x''x')^m$. Since $z_1' \subset (x'x'')^m$, we get a contradiction as before.

Since Cases 1–6 lead to contradiction we conclude that $z_3 = \varepsilon$. □

Let us now assume that the hole is at a position i in a word of length n. The upper bound for the maximum number of factors, compatible with distinct squares, would be achieved if all these factors, starting before the hole, would contain the hole (this way more than two squares can start at the same position). Note that in the case when a square containing a hole has its last occurrence at a certain position, no other full word that is a square can have its last occurrence starting at the same position (otherwise a later occurrence of the same full word, or a word compatible with it, would appear later in the word). Let us look at the start position j of a square containing the hole and denote it as j (obviously, there are at most i such squares). Let us denote the length of such square by n_j.

Hence, if at position j we have a square of length n_j, then according to Theorem 18, up to position $2n_j + j$ we will have counted at most three distinct squares. Using an induction we notice that up to position $2^m n_j + j$ we will have counted at most $2m + 1$ distinct squares. Since the length of the word is n, we have that the maximum value for m, for squares starting at position j, is bounded by $\log(\frac{n-j}{n_j})$.

Note that the maximum is achieved for the case when n_j is minimum. Hence we can replace in our formula n_j by $i - j$, which is the smallest length a square starting at position j and containing the hole may have.

Theorem 19. *The number of distinct full squares compatible with factors in any partial word with one hole of length n is at most $\frac{7n}{2}$.*

Proof. Using the previous remarks it is easy to see that the number of squares at position j and containing the hole is $2\log(\frac{n-j}{i-j}) + 1$. Hence, we get that the total number of distinct squares that we can obtain is

$$\textstyle\sum_{j=0}^{i-1}(2\log(\frac{n-j}{i-j}) + 1) = i + \frac{2}{\ln 2}\sum_{j=0}^{i-1}\ln(\frac{n-j}{i-j})$$

The sum from the previous formula is equal to $\sum_{x=n-i+1}^{n}\ln(x) - \sum_{y=1}^{i}\ln(y)$ and implicitly, less or equal to $\int_{n-i+1}^{n+1}\ln(x)\mathrm{d}x - \int_1^i \ln(y)\mathrm{d}y$. After integrating we get

$$(n+1)\ln(n+1) - (n-i+1)\ln(n-i+1) - i\ln(i)$$

Since the maximum is obtained for $i = \frac{n+1}{2}$, the function is hence less than $(n+1)\ln 2$. Using Theorem 16 for the rest of the word, we get that the number of distinct squares, compatible with factors of the word, is bounded by

$$2n - \tfrac{n+1}{2} + \tfrac{2}{\ln 2}(n+1)\ln 2 = \tfrac{7n}{2} + \tfrac{3}{2}$$

Since the last position in the word contains no squares, we get that the maximum number of factors compatible with distinct full squares is smaller than $\frac{7n}{2}$. □

This bound can be slightly improved by using Ilie's $2n - \Theta(\log n)$ [Ili07].

Motivated by the same conclusion (the results have been investigated independently at the same time), in [HHK09], the authors show that, by restricting the alphabet to the binary case, the bound is improved to $3n$, where n is the length of the word.

Theorem 20. [HHK09] *For any binary partial word w, of length n, containing only one hole, the maximum number of distinct squares is at most $3n$.*

The proof of this result makes use of Theorem 16, which states that for binary words we cannot have more than 4 last occurrences of factors compatible with squares at each position.

4.4 Conclusion

Although the computations done so far show that the actual bound for the one-hole partial words give us at most n distinct squares in any word of length n, the results obtained here using the approach of Fraenkel and Simpson make the bound directly dependable on the size of the alphabet. From our point of view, finding a dependency between the maximum number of squares starting at one position and the length of the word might be a solution. Solving this problem, at least partially, could also give a new perspective to the study of maximum distinct squares within a full word.

Note as well that for arbitrarily large alphabets of size k, we get an upper bound for all words containing h holes and having length n

$$g_{h,k}(n) \leq m_{h,k}(n) + k^{\lfloor \frac{h}{2} \rfloor}$$

This is due to the fact that the leading term is always maximal in $m_{h,k}$, hence adding one to its coefficient we get an upper bound.

In order to improve the bound stated in Theorem 19, we need to somehow limit to less than 3.5 the average number of squares that have their last occurrence starting at the positions of the partial word. This requirement draws attention to positions i where three or more squares have their last occurrences. Is it true that at positions "neighboring" to i, no squares can have their last occurrences? In fact, if at position i we have at least three factors compatible with full squares, this does not imply that at position $i + 1$ we will have less. Indeed, consider the example

$$ab\diamond abcabbeabcabdabcabbeabcabbabcabbeabcabdabcabbeabcab$$

where at position 0 we have factors compatible with the squares

$$(abc)^2,\ (abdabcabbeabc)^2 \text{ and } (abbabcabbeabcabdabcabbeabc)^2$$

at position 1 factors compatible with the squares

$$(bca)^2,\ (beabcab)^2,\ (bdabcabbeabca)^2 \text{ and } (bbabcabbeabcabdabcabbeabca)^2$$

and at position 2 factors compatible with

$$a^2,\ (cab)^2,\ (dabcabbeabcab)^2 \text{ and } (babcabbeabcabdabcabbeabcab)^2$$

In [Ili07], Ilie gave a relation between the lengths of squares at positions neighboring a position where two squares have their last occurrences. More precisely, he showed that if $v^2 < u^2$ are two squares at position i and w^2 is a square at position $i+1$, then either $|w| \in \{|v|, |u|\}$ or $|w| \geq 2|v|$ (see Lemma 2 in [Ili07]). Referring to the above example, we observe that such is not the case with partial words with one hole.

Chapter 5

Unbordered Partial Words

A word is called *bordered* if one of its proper prefixes is one of its suffixes. The length of the longest such prefix (also called longest border) is the length of the word minus the length of its shortest period. The word is called *unbordered* otherwise. In other words, it is *unbordered* if it has no proper period. For example, *abaabb* is unbordered while *abaab* is bordered. Unbordered words turn out to be primitive, that is, they cannot be written as a power of another word. Moreover, unborderedness has the following important property: different occurrences of an unbordered factor never overlap in the same word. A related property is that no primitive word u can be an inside factor of uu. Fast algorithms for testing primitivity of words can be based on this property [CR94].

The study of unbordered partial words was initiated in [BS05]. Later on, Blanchet-Sadri and Wetzler extended the well known critical factorization theorem to partial words and their result states that the minimal weak period of a *non-special* partial word can be locally determined in at least one position [BSW07]. The first two sections of this chapter, together with the last one, represent joint work with Francine Blanchet-Sadri, Crystal D. Davis, Joel Dodge and Margaret Moorefield, see [BSDD$^+$09], while the third and the fourth Sections contain work done together with Emily Allen, Francine Blanchet-Sadri and Cameron Byrum, see [ABSBM09].

5.1 Concatenations of prefixes

First let us state an equivalent definition for the notion of a weak period:

Lemma 20. *For an integer p, the partial word $u \in A_\diamond^*$ is weakly p-periodic if and only if the containments $u \subset xv$ and $u \subset wx$ hold for some partial words x, v, w satisfying $|v| = |w| = p$.*

Proof. Write u as $v_1 v_2 \cdots v_k r$ where $|v_1| = |v_2| = \cdots = |v_k| = p$ and $0 \leq |r| < p$, and v_k as st where $|s| = |r|$. Set $x_1 = v_1 \cdots v_{k-1} s$ and $x_2 = v_2 \cdots v_k r$.

If the containments $u \subset xv$ and $u \subset wx$ hold for some partial words x, v, w satisfying $|v| = |w| = p$, then both $v_1 \cdots v_{k-1} s \subset x$ and $v_2 \cdots v_k r \subset x$ hold, and so $v_1 \cdots v_{k-1} s \uparrow v_2 \cdots v_k r$. By Simplification, $v_1 \uparrow v_2, \ldots, v_{k-1} \uparrow v_k$ and $s \uparrow r$. Now, let $i, i+p \in D(u)$. Then $i = lp + j$ for some $0 \leq l < k$ and $0 \leq j < p$. If $l < k-1$, then we get $u(i) = v_{l+1}(j) = v_{l+2}(j) = u(i+p)$ since $v_{l+1} \uparrow v_{l+2}$ and $j \in D(v_{l+1}) \cap D(v_{l+2})$, and if $l = k-1$, then $u(i) = v_k(j) = s(j) = r(j) = u(i+p)$ since $s \uparrow r$ and $j \in D(s) \cap D(r)$. In either case, u is weakly p-periodic. Conversely, if p is a weak period of u, then $v_i \uparrow v_{i+1}$ for all $1 \leq i < k$ and $s \uparrow r$. Thus $x_1 \uparrow x_2$, and there exists x such that $x_1 \subset x$ and $x_2 \subset x$. Setting $v = tr$ and $w = v_1$, we get $u = x_1 v \subset xv$ and $u = w x_2 \subset wx$ with $|v| = |w| = p$. □

For $u, v \in A_\diamond^*$, we write $u \ll v$ if there exists a sequence $v_0, \ldots, v_{n-1}$ of prefixes of v such that $u = v_0 \cdots v_{n-1}$. Obviously, $\varepsilon \ll u$ and $u \ll u$. Also, if $u \ll v$ and $v \ll w$, then $u \ll w$.

Theorem 21. [ES79] *Let $u \in A^+, v \in A^*$ be such that $u \ll v$. Then there exists a unique sequence $v_0, \ldots, v_{n-1}$ of non-empty unbordered prefixes of v such that $u = v_0 \cdots v_{n-1}$.*

Our main result in this section is to extend Theorem 21 to partial words (see Theorem 22). In order to do this, we introduce two types of bordered partial words: the *well bordered* and the *badly bordered* partial words.

Definition 2. *Let $u \in A_\diamond^+$ be bordered. Let x be a minimal border of u, and set $u = x_1 v = w x_2$ where $x_1 \subset x$ and $x_2 \subset x$. We call u* well bordered *if x_1 is unbordered. Otherwise, we call u* badly bordered.

Note that if a non-empty partial word u is well bordered then x_2 can be either bordered or unbordered, and the same is true if u is badly bordered.

For convenience, we will at times refer to a minimal border of a well bordered partial word as a *good border* and of a badly bordered partial word as a *bad border.*

As a result of x being a bad border, we have the following Lemma.

Lemma 21. *Let $u \in A_\diamond^+$ be badly bordered. Let x be a minimal border of u, and set $u = x_1 v = w x_2$ where $x_1 \subset x$ and $x_2 \subset x$. Then there exists i such that $i \in H(x_1)$ and $i \in D(x_2)$.*

Proof. Since x_1 is bordered, $x_1 = r_1 s_1 = s_2 r_2$ for non-empty partial words r_1, r_2, s_1, s_2 where $s_1 \subset s$ and $s_2 \subset s$ for some s. If no i exists such that $i \in H(x_1)$ and $i \in D(x_2)$, then x_2 must also be bordered. So $x_2 = r_1' s_1' = s_2' r_2'$ where $r_1' \subset r_1$, $r_2' \subset r_2$, $s_1' \subset s$ and $s_2' \subset s$, thus $s_2 \uparrow s_1'$. This means that there exists a border of u of length shorter that $|x|$ which contradicts the fact that x is a minimal border of u. □

Our goal is to extend Theorem 21 to partial words or to construct, given any partial words u and v satisfying $u \ll v$, a unique sequence of non-empty unbordered prefixes of v, $v_0, \ldots, v_{n-1}$, such that $u \uparrow v_0 \cdots v_{n-1}$. We will see that if during the construction of the sequence a badly bordered prefix is encountered, then the desired sequence may not exist. We first prove two propositions.

Proposition 17. *If $v \in A_\diamond^*$, then there do not exist two distinct compatible sequences of non-empty unbordered prefixes of v.*

Proof. Suppose that $v_0 \cdots v_{n-1} \uparrow v_0' \cdots v_{m-1}'$ where each v_i and each v_i' is a non-empty unbordered prefix of v. If there exists $i \geq 0$ such that $|v_0| = |v_0'|, \ldots, |v_{i-1}| = |v_{i-1}'|$ and $|v_i| < |v_i'|$, then $v_0 = v_0', \ldots, v_{i-1} = v_{i-1}'$ and v_i is a prefix of v_i'. By simplification, $v_i \cdots v_j x \uparrow v_i'$ where $i \leq j < n-1$ and x is a non-empty prefix of v_{j+1}. The fact that x, v_i' are prefixes of v satisfying $|v_i'| > |x|$ implies that x is a prefix of v_i'. In addition, x is compatible with the suffix of length $|x|$ of v_i', and consequently v_i' is bordered. Similarly, there exists no $i \geq 0$ such that $|v_0| = |v_0'|, \ldots, |v_{i-1}| = |v_{i-1}'|$ and $|v_i| > |v_i'|$. Clearly, $n = m$ and uniqueness follows. □

Proposition 18. *Let $u \in A_\diamond^+$ be bordered. Let x be a minimal border of u, and set $u = x_1 v = w x_2$ where $x_1 \subset x$ and $x_2 \subset x$. Then the following hold:*

1. *The partial word x is unbordered.*
2. *If u is well bordered, then $u = x_1 u' x_2 \subset x u' x$ for some u'.*

Proof. For Statement 1, assume that r is a border of x, that is, $x \subset rs$ and $x \subset s'r$ for some non-empty partial words r, s, s'. Since $u \subset xv$ and $x \subset rs$, we have $u \subset rsv$, and similarly, since $u \subset wx$ and $x \subset s'r$, we have $u \subset ws'r$. Then r is a border of u. Since x is a minimal border of u, we have $|x| \leq |r|$ contradicting the fact that $|r| < |x|$. This proves (1).

For Statement 2, if $|v| < |x|$, then $u = wtv$ for some t. Here $x_1 = wt = t'w'$ for some t', w' satisfying $|t| = |t'|$ and $|v| = |w| = |w'|$. Since $x_1 \uparrow x_2$, we have $t'w' \uparrow tv$ and by simplification, $t' \uparrow t$. The latter implies the existence of a partial word t'' such that $t' \subset t''$ and $t \subset t''$. So $x_1 = t'w' \subset t''w'$ and $x_1 = wt \subset wt''$. Then t'' is a border of x_1 and x_1 is bordered. According to the definition of u being well bordered, x_1 is an unbordered partial word and this leads to a contradiction. Hence, we have $|v| \geq |x|$ and, for some u', we have $v = u'x_2$ and $w = x_1u'$, and $u = wx_2 = x_1u'x_2 \subset xu'x$. This proves (2). □

Note that Proposition 18 implies that if $u \in A^+$ is bordered, then u is well bordered. In this case, $u = xu'x$ where x is the minimal border of u.

Lemma 22. *If $u, v \in A_\diamond^+$ are such that $u = v_0 \cdots v_{n-1}$ where $v_0, \ldots, v_{n-1}$ is a sequence of non-empty unbordered prefixes of v, then there exists a unique sequence $v'_0, \ldots, v'_{m-1}$ of non-empty unbordered prefixes of v such that $u \uparrow v'_0 \cdots v'_{m-1}$ (the desired sequence is just $v_0, \ldots, v_{n-1}$).*

Proof. The statement follows immediately from Proposition 17. □

The badly bordered partial words are now split into the *specially bordered* and the *non-specially bordered* partial words according to the following definition.

Definition 3. *Let $u \in A_\diamond^+$ be a partial word that is badly bordered. Let x be a minimal border of u, and set $u = x_1v = wx_2$ where $x_1 \subset x$ and $x_2 \subset x$. If there exists a proper factor x' of u such that $x_1 \not\uparrow x'$ and $x' \uparrow x_2$, then we call u* specially bordered. *Otherwise, we call u* non-specially bordered.

Lemma 23. *Let $v \in A_\diamond^+$ be badly bordered. Let y be a minimal border of v, and set $v = y_1w' = wy_2$ where $y_1 \subset y$ and $y_2 \subset y$ (and thus $y_1 \uparrow y_2$). If there exists a sequence $v_0, \ldots, v_{m-1}$ of non-empty unbordered prefixes of v such that $v \uparrow v_0 \cdots v_{m-1}$, then $|y_1| < |v_{m-1}|$ and v is specially bordered.*

Proof. By Definition 2, y_1 is bordered. If $|y_1| = |v_{m-1}|$, then both y_1 and v_{m-1} are prefixes of v, and thus $y_1 = v_{m-1}$. We get that y_1 is unbordered, a contradiction. If $|y_1| > |v_{m-1}|$, then set $y_2 = z_1v'$ where $|v'| = |v_{m-1}|$. Since both y_1 and v_{m-1} are prefixes of v, we get that v_{m-1} is a prefix of y_1. So $y_1 = v_{m-1}z_2$ for some z_2, and $v = v_{m-1}z_2w' = wz_1v'$ with $v_{m-1} \uparrow v'$. Thus v has a border of length $|v_{m-1}| < |y_1| = |y|$ contradicting the fact that y is a minimal border. And so $|y_1| < |v_{m-1}|$.

Since $v \uparrow v_0 \cdots v_{m-1}$, we have $|v_{m-1}| \leq |v|$. Since v_{m-1} is a prefix of v, and $v = y_1w'$ and $|v_{m-1}| > |y_1|$ there exists z_1 such that $y_1z_1 = v_{m-1}$. Since $v = wy_2$ and v_{m-1} is compatible with a suffix of v, we have $v_{m-1} \uparrow z_2y_2$ for some z_2. Thus, we get that $v_{m-1} = y_1z_1 \uparrow z_2y_2$. Since $v_{m-1} \uparrow z_2y_2$, set $v_{m-1} = z_3y_3$ where $z_3 \uparrow z_2$ and $y_3 \uparrow y_2$. So $v_{m-1} = z_3y_3 = y_1z_1$. If $y_3 \uparrow y_1$, then v_{m-1} is bordered, a contradiction with the fact that v_{m-1} is unbordered. Thus $y_3 \not\uparrow y_1$, and since v_{m-1} is a prefix of v, we have that v is specially bordered. □

The following example illustrates Lemma 23.

Example 4. *Consider the partial word*

$$v = aa\diamond aabbaaaaa\diamond b$$

Here, v is specially bordered (indeed, it has the factor abb such that $aa\diamond \not\uparrow abb$ and $a\diamond b \uparrow abb$) and is compatible with a sequence of some of its unbordered prefixes. Indeed, the compatibility

$$aa\diamond aabbaaaaa\diamond b \uparrow (aa\diamond aabb)(aa\diamond aabb)$$

holds. The shortest border of v is aab which has length shorter than $aa\diamond aabb$.

Lemma 24. *Let $v \in A_\diamond^+$ be a well bordered word. Then there exists a longest sequence $v_0, v_1, \ldots, v_{m-1}$ of non-empty prefixes of v such that $v \uparrow v_0v_1 \cdots v_{m-1}$, v_j is unbordered for every $1 \leq j < m$, and v_0 is unbordered or badly bordered. Moreover, if v_0 is badly bordered, then no sequence of non-empty unbordered prefixes of v exists that is compatible with v.*

Proof. Let y_0 be a minimal border of $w_0 = v$, and set $w_0 = x_0w_1' = w_1x_0'$ where $x_0 \subset y_0$ and $x_0' \subset y_0$ (and thus $x_0 \uparrow x_0'$). By Definition 2, x_0 is unbordered, and

$$v = w_1x_0' \uparrow w_1x_0 \tag{5.1}$$

where both w_1 and x_0 are prefixes of w_0 (and hence of v). If w_1 is unbordered, then v is compatible with a sequence of its non-empty unbordered prefixes.

If w_1 is badly bordered, then no sequence $v'_0, \ldots, v'_{m'-1}$ of non-empty unbordered prefixes of v exists that is compatible with w_1 unless w_1 is specially bordered and $|y_1| < |v'_{m'-1}|$ by Lemma 23 (here y_1 is a minimal border of w_1). If this is the case, then w_1 may be compatible with such a sequence of non-empty unbordered prefixes of v, and if so replace w_1 on the right hand side of the compatibility in (1) by $v'_0 \cdots v'_{m'-1}$. If this is not the case, then no sequence of non-empty unbordered prefixes of v exists that is compatible with v.

If w_1 is well bordered, then repeat the process. Let y_1 be a minimal border of w_1, and set $w_1 = x_1 w'_2 = w_2 x'_1$ where $x_1 \subset y_1$ and $x'_1 \subset y_1$ (and thus $x_1 \uparrow x'_1$). By Definition 2, x_1 is unbordered, and

$$v = w_2 x'_1 x'_0 \uparrow w_2 x_1 x_0 \tag{5.2}$$

where both w_2 and x_1 are prefixes of w_1 (and hence of v, since w_1 is a prefix of v) and x_0 is a prefix of v.

Let $w_0, w_1, \ldots, w_{j-1}$ be the longest sequence of non-empty well bordered prefixes defined in this manner. For all $0 \leq k < j$, let y_k be a minimal border of w_k, and set $w_k = x_k w'_{k+1} = w_{k+1} x'_k$ where $x_k \subset y_k$ and $x'_k \subset y_k$ (and thus $x_k \uparrow x'_k$). Again by Definition 2, $x_0, \ldots, x_{j-1}$ are unbordered. We have $w_{j-1} = w_j x'_{j-1} \uparrow w_j x_{j-1}$ and thus by induction,

$$v = w_j x'_{j-1} \cdots x'_0 \uparrow w_j x_{j-1} \cdots x_0 \tag{5.3}$$

where $w_j, x_{j-1}, \ldots, x_0$ are prefixes of w_0 (and hence of v). Now, if w_j is unbordered, then v is compatible with a sequence of some of its non-empty unbordered prefixes. If w_j is badly bordered, then proceed as in the case above when w_1 is badly bordered.

We can thus equate v with sequences of shorter and shorter factors that are some of its prefixes or compatible with some of its prefixes and the existence of the required sequence $v_0, \ldots, v_{m-1}$ is established. □

Theorem 22. *Let $u, v \in A_\diamond^+$ be such that $u \ll v$, and let $v_0, \ldots, v_{m-1}$ be a longest sequence of non-empty prefixes of v satisfying $u \uparrow v_0 \cdots v_{m-1}$. Then, either all v_i's are unbordered, or u is not compatible with the concatenation of any sequence of unbordered prefixes of v. In the latter case, some of the v_i's are badly bordered while the others are unbordered.*

Proof. If $v_0, \dots, v_{m-1}$ are unbordered, then by Lemma 22 we get the unique sequence of non-empty unbordered prefixes of v whose concatenation is compatible with u. If any of the prefixes are well or badly bordered, then proceed as in Lemma 23 or Lemma 24. □

Example 5. *Consider the partial words*

$$u = aaaa\diamond babbaaaaa\diamond baa \text{ and } v = aa\diamond babbaaaaa\diamond b$$

We have a factorization of u in terms of non-empty prefixes of v. Here, the compatibility

$$u \uparrow (a)(a)(aa\diamond babbaaaaa\diamond b)(a)(a)$$

consists of unbordered and badly bordered prefixes of v and is a longest such sequence ($aa\diamond babbaaaaa\diamond b$ is specially bordered and is not compatible with any sequence of non-empty unbordered prefixes of v). We can check that no sequence of non-empty unbordered prefixes of v is compatible with u.

5.2 More results on concatenations of prefixes

In this section, we give more results on concatenations of prefixes. In particular, we study properties of the longest unbordered prefix of a partial word. We also investigate the relationship between the minimal weak period of a partial word and the maximal length of its unbordered factors. Our main results in this section (Theorems 23 and 24) extend a result of Ehrenfeucht and Silberger [ES79] which states that if $u = xv$ is a non-empty unbordered word where x is the longest unbordered proper prefix of u, then v is unbordered.

If $u \in A_\diamond^+$, then $\text{unb}(u)$ denotes the longest unbordered prefix of u. A result of Ehrenfeucht and Silberger shows that if $u, v \in A^*$ are such that $u = \text{unb}(u)v$, then $v \ll \text{unb}(u)$ [ES79]. This does not extend to partial words as $u = (ab)(\diamond b) = \text{unb}(u)v$ provides a counterexample. However, the following lemma does hold.

Lemma 25. *Let $u \in A_\diamond^+, v \in A_\diamond^*$ be such that $u = unb(u)v$. Then $u \ll unb(u)$ if and only if $v \ll unb(u)$.*

Proof. If $v \ll \text{unb}(u)$, then obviously $u \ll \text{unb}(u)$. For the other direction, since $u \ll \text{unb}(u)$, we can write $u = u_0 u_1 \cdots u_{n-1}$ where each u_i is a non-empty prefix of $\text{unb}(u)$. We can suppose that $v \neq \varepsilon$. Then $\text{unb}(u) = u_0 \cdots u_k u'$ for some $k < n-1$ and some prefix u' of u_{k+1}. Since $\text{unb}(u)$ is unbordered, we have that $u' = \varepsilon$, that $k = 0$, and hence that $\text{unb}(u) = u_0$. It follows that $v = u_1 \cdots u_{n-1}$ and $v \ll \text{unb}(u)$. □

We get the following corollary.

Corollary 8. *Let $u \in A_\diamond^*, v \in A_\diamond^+$. Then the following hold:*

1. *If $u \ll unb(v)$, then $u \ll v$.*
2. *If $w \in A_\diamond^*$ is such that $v = unb(v)w$ and $w \ll unb(v)$, then $u \ll v$ if and only if $u \ll unb(v)$.*

Proof. Statement 1 holds trivially. For Statement 2, by Lemma 25, $w \ll \text{unb}(v)$ if and only if $v \ll \text{unb}(v)$. Now, if $u \ll v$, then since $v \ll \text{unb}(v)$, by transitivity we get $u \ll \text{unb}(v)$. □

Statement 2 of Corollary 8 is not true in general. Indeed, $u = ababac\diamond aab$ and $v = abac\diamond aba$ provide a counterexample. To see this, $v = (abac)(\diamond aba) = \text{unb}(v)w$ and we have $u \ll v$ since $u = (ab)(abac\diamond a)(ab)$ where ab and $abac\diamond a$ are prefixes of v. However $u \not\ll \text{unb}(v)$ (here $w \not\ll \text{unb}(v)$). However, for $u, v \in A^*$, $u \ll v$ if and only if $u \ll \text{unb}(v)$ [ES79].

For $u, v \in A_\diamond^*$, when both $u \ll v$ and $v \ll u$ we write $u \approx v$. The relation $\approx$ is an equivalence relation. A result on words states that for $u, v \in A^*$, $u \approx v$ if and only if $\text{unb}(u) = \text{unb}(v)$ [ES79]. Remember that $P(u)$ represents the set of prefixes of u, while $S(u)$ its set of suffixes. For partial words, the following holds.

Proposition 19. *For $u, v \in A_\diamond^*$, if $u \approx v$, then $unb(u) = unb(v)$.*

Proof. Suppose that $u \approx v$. Set $v = \text{unb}(v)w$ for some partial word w. Since $u \ll v$, we can write $u = v_0 \cdots v_{n-1}$ where each v_i is a non-empty prefix of v. Since $v \ll u$, there exists a sequence of non-empty prefixes of u, say $u_0, \ldots, u_{m-1}$, such that $v = u_0 u_1 \cdots u_{m-1}$. Since $\text{unb}(v)$ is a prefix of v, we have $\text{unb}(v) = u_0 \ldots u_k u'$ where u' is a prefix of u_{k+1} and $k < m-1$. Since $\text{unb}(v)$ is unbordered, we have $u' = \varepsilon$, $k = 0$, and $\text{unb}(v) = u_0$. Therefore, $\text{unb}(v)$ is an unbordered prefix of u. Hence, it is a prefix of $\text{unb}(u)$. Similarly, $\text{unb}(u)$ is a prefix of $\text{unb}(v)$. □

The converse of Proposition 19 does not necessarily hold for partial words as is seen by considering $u = aba\diamond$ and $v = ab\diamond b$. We have $\text{unb}(u) = ab = \text{unb}(v)$ but $u \not\approx v$.

If v is an unbordered word and w is a proper prefix of v for which $u \ll w$, then uv and wv are unbordered [ES79]. For partial words, we can prove the following.

Lemma 26. *Let $u \in A_\diamond^*$ be unbordered. Then the following hold:*

1. *If $v \in P(u)$ and $v \neq u$, then vu is unbordered.*

2. *If $v \in S(u)$ and $v \neq u$, then uv is unbordered.*

Proof. Let us prove Statement 1 (the proof of Statement 2 is similar). Set $u = vx$ for some x. If $vu = vvx$ is bordered, then there exist non-empty partial words r, s, s' such that $vvx \subset rs$ and $vvx \subset s'r$. If $|r| \leq |v|$, then $u = vx$ is bordered by r. And if $|r| > |v|$, then $r = v'y$ where $|v'| = |v|$ and this implies that $u = vx$ is bordered by y. In either case, we get a contradiction with the assumption that u is unbordered. □

Lemma 27. *If $v \in A_\diamond^*$ is unbordered and $u \ll v$ and $u \neq v$, then uv is unbordered.*

Proof. Since $u \ll v$, we can write $u = v_0 v_1 \cdots v_{n-1}$ where each v_i is a prefix of v. Therefore, any prefix of u is a concatenation of prefixes of v. Assume that uv is bordered by y. If $|y| > |u|$, then set $y = u'y'$ with $u \subset u'$. We get y' a border of v contradicting the fact that v is unbordered. If $|y| \leq |u|$, then we have the following two cases:

Case 1. y contains a prefix of v_0

Here y contains a prefix of v and also a suffix of v and therefore, y is a border of the unbordered word v.

Case 2. $v_0 \cdots v_k v' \subset y$ where v' is a prefix of v_{k+1}

If $v' = \varepsilon$, then $v_0 \cdots v_k \subset y$ where v_k is a prefix of v. This results in a suffix of y containing both a prefix and a suffix of v. Similarly, if $v' \neq \varepsilon$, then factor y as $y = y_1 y_2$ where $v' \subset y_2$. Because v' is a prefix of v, we can write $v = v'z \subset y_2 z$. But because $|y_2| < |v|$ and we have assumed that uv is bordered by $y = y_1 y_2$, we must have that $v = z'v''$ with $v'' \subset y_2$. Therefore

y_2 is a border for v. In either case, we get a contradiction with the fact that v is unbordered. □

A result of Ehrenfeucht and Silberger [ES79] states that if $u = \text{punb}(u)v$ is a non-empty unbordered word where $\text{punb}(u)$ the longest proper unbordered prefix of u, then v is unbordered. The partial word $u = ab\diamond ac$ where $\text{punb}(u) = ab$ and $v = \diamond ac$ and the partial word $u = abaca\diamond c$ where $\text{punb}(u) = abac$ and $v = a\diamond c$ provide counterexamples for partial words. However, when v is full, the following theorem does hold.

Theorem 23. *Let $u \in A_\diamond^+$ be unbordered. Then the following hold:*

1. *Let x be the longest proper unbordered prefix of u and let v be such that $u = xv$. If $v \in A^*$, then v is unbordered.*
2. *Let y be the longest proper unbordered suffix of u and let w be such that $u = wy$. If $w \in A^*$, then w is unbordered.*

Proof. We prove Statement 1 (Statement 2 can be proved similarly). Assume that v is bordered. Since v is full, there exist non-empty words z, v' such that $v = zv'z$ where z is the minimal border of v. Then $u = \text{punb}(u)zv'z$, so that $\text{punb}(u)z$ is a proper prefix of u such that $|\text{punb}(u)z| > |\text{punb}(u)|$. It follows that $\text{punb}(u)z$ is bordered, and there exist non-empty partial words r, r_1, r_2, s_1, s_2 such that $\text{punb}(u)z = r_1s_1 = s_2r_2$, $r_1 \subset r$ and $r_2 \subset r$ (here r is a minimal border). Let us consider the following two cases:

Case 1. $|r| > |z|$

In this case, $r_2 = x'z$ where x' is a non-empty suffix of $\text{punb}(u)$. Since $r_1 \uparrow r_2$, there exist partial words x'', z' such that $r_1 = x''z'$ where $x'' \uparrow x'$ and $z' \uparrow z$. But then, $x''z's_1 = r_1s_1 = \text{punb}(u)z = s_2r_2 = s_2x'z$. It follows that x'' is a prefix of $\text{punb}(u)$ and x' is a suffix of $\text{punb}(u)$ that are compatible. As a result, $\text{punb}(u)$ is bordered.

Case 2. $|r| \leq |z|$

In this case, r_2 is a suffix of z and set $z = sr_2$ for some s. We get $u = \text{punb}(u)zv'z = r_1s_1v'sr_2 \subset rs_1v'sr$, whence r is a border of the unbordered partial word u. □

A closer look at the proof of Theorem 23 allows us to show the following.

Theorem 24. *Let $u \in A_{\diamond}^{+}$. Then the following hold:*

1. *Let x be the longest proper unbordered prefix of u and let v be such that $u = xv$. If v is bordered, then set $v = z_1v_1 = v_2z_2$ where $z_1 \subset z, z_2 \subset z$ and where z is a minimal border of v. Then xz_1 has a minimal border r such that $|r| \leq |z|$. Moreover, if v is well bordered, then $|x| \geq |r|$.*

2. *Let y be the longest proper unbordered suffix of u and let w be such that $u = wy$. If w is bordered, then set $w = z_1v_1 = v_2z_2$ where $z_1 \subset z, z_2 \subset z$ and where z is a minimal border of w. Then z_2y has a minimal border r such that $|r| \leq |z|$. Moreover, if w is well bordered, then $|y| \geq |r|$.*

Proof. We prove Statement 1 (Statement 2 can be proved similarly). Then $u = \text{punb}(u)z_1v_1$, so that $\text{punb}(u)z_1$ is a proper prefix of u longer than $\text{punb}(u)$. It follows that $\text{punb}(u)z_1$ is bordered, and there exist non-empty partial words r, r_1, r_2, s_1, s_2 such that $\text{punb}(u)z_1 = r_1s_1 = s_2r_2$, $r_1 \subset r$ and $r_2 \subset r$ with r a minimal border. If $|r| > |z|$, then $r_2 = x'z_1$ where x' is a non-empty suffix of $\text{punb}(u)$. Since $r_1 \uparrow r_2$, there exist partial words x'', z' such that $r_1 = x''z'$ where $x'' \uparrow x'$ and $z' \uparrow z_1$. But then, $x''z's_1 = r_1s_1 = \text{punb}(u)z_1 = s_2r_2 = s_2x'z_1$. It follows that x'' is a prefix of $\text{punb}(u)$ and x' is a suffix of $\text{punb}(u)$ that are compatible. As a result, $\text{punb}(u)$ is bordered, which contradicts that $\text{punb}(u)$ is the longest unbordered proper prefix of u. And so $|r| \leq |z|$ and r_2 is a suffix of z_1. Set $z_1 = sr_2$ for some suffix s of s_2 ($s_2 = \text{punb}(u)s$). If we further assume that v is well bordered, then we claim that $|\text{punb}(u)| \geq |r|$. To see this, if $|\text{punb}(u)| < |r|$, then set $r_1 = \text{punb}(u)t$ and $z_1 = ts_1$ for some t. Since $r_1 \uparrow r_2$, there exist x', t' such that $r_2 = x't'$ and $\text{punb}(u) \uparrow x'$ and $t \uparrow t'$. Since r_2 is a suffix of z_1, we have that t' is a suffix of z_1. Consequently, t is a prefix of z_1 and t' is a suffix of z_1 that are compatible. So z_1 is bordered and we get a contradiction with v's well borderedness, establishing our claim. □

The maximum length of the unbordered factors of a partial word u is denoted by $\mu(u)$. Recall that $p(u)$ denotes the minimal period of a (full) word u. Ehrenfeucht and Silberger studied the relationship between $p(u)$ and $\mu(u)$ in [ES79]. Clearly, $\mu(u) \leq p(u)$. Here, we investigate the relationship between the minimal weak period of a partial word u, $p'(u)$, and $\mu(u)$.

Proposition 20. *For all $u \in A_{\diamond}^{*}$, $\mu(u) \leq p'(u) \leq p(u)$.*

Proof. Let w be a factor of u such that $|w| > p'(u)$. Factor w as $w = xw_1 = w_2y$ where $|w_1| = |w_2| = p'(u)$. We have $x(i) = w(i)$ and $y(i) = w(i+p'(u))$. This means that whenever $x(i) \neq y(i)$, $i \in H(x)$ or $i \in H(y)$. Therefore $x \uparrow y$ and w is bordered. So we must have that $\mu(u) \leq p'(u)$. □

For any partial word u, Proposition 20 gives an upper bound for the maximum length of the unbordered factors of u: $\mu(u) \leq p'(u)$. This relationship cannot be replaced by $\mu(u) < p'(u)$ as is seen by considering $u = aba\diamond$ with $\mu(u) = p'(u) = 2$.

For any $v, w \in A_\diamond^*$, if there exists a partial word u such that $u \ll w$ and $u \subset v$, then we say that v contains a concatenation of prefixes of w. Otherwise, we say that v contains no concatenation of prefixes of w. Similarly, if $u \in P(w)$ and $u \subset v$, then we say that v contains a prefix of w.

The following result extends to partial words a result on words which states that if u, v are words such that $u = \text{unb}(u)v\text{unb}(u)$ and $\text{unb}(u)$ is not a factor of v, then $v\text{unb}(u)$ is unbordered (Corollary 2.5 in [Duv82]).

Proposition 21. *Let $u, v \in A_\diamond^*$ be such that $u = hvh$ where h abbreviates $\text{unb}(u)$. If h is not compatible with any factor of v, then vh is unbordered if one of the following holds:*

1. *v is full,*

2. *v contains a prefix of h or a concatenation of prefixes of h.*

Proof. For Statement 1, suppose that v is full and there exist non-empty x, w_1, w_2 such that $vh \subset xw_1$ and $vh \subset w_2x$. We must have that $|x| \leq |v|$ or else h, which is unbordered, would be bordered by a factor of x. If $|h| < |x|$, then there exists $x' \in S(x)$ such that $h \subset x'$ and because $|x| \leq |v|$, there exists v' a factor of v with $v' \subset x'$ and this says that $v' \uparrow h$, contradicting our assumption. Now, if $|h| \geq |x|$, then set $v = rv'$ and $h = h's$ where $|r| = |s| = |x|$. In this case, $r \subset x$ and $s \subset x$, and there exist non-empty $r \in P(v)$ and $s \in S(h)$ such that $r \uparrow s$. But r is full and so $r \uparrow s$ implies that $s \subset r$. But then, by Lemma 26, we have that hs is unbordered, and so hr is an unbordered prefix of u with length greater than $|h|$. This contradicts the assumption that $h = \text{unb}(u)$, hence vh must be unbordered.

For Statement 2, first assume that v contains a prefix of h. Let $v' \in P(h)$ be such that $v' \subset v$. By Lemma 26, since h is unbordered, we have that $v'h$ is unbordered. Now, assume that v contains a concatenation of prefixes of h.

Let v' be such that $v' \ll h$ and $v' \subset v$. By Lemma 27, since h is unbordered and $v' \ll h$, we have that $v'h$ is unbordered. In either case, since $v' \subset v$, vh is unbordered as well. □

5.3 Simply bordered partial words

Once the number of holes in a partial word of a fixed length reaches a certain bound, the word will have a simple border. In this section, we give a closed formula for that bound and show that it is constant over all alphabets of size at least two.

It follows from Proposition 18 that if u is a full bordered word, then $x_1 = x$ is unbordered. In this case, x is the minimal border of u and $u = xu'x$. Thus, a bordered full word is always simply bordered and has a unique minimal border. Since borders for partial words are defined using containment, it is possible to have numerous borders having the same length. Thus, a partial word does not necessarily have a unique minimal border.

In [BS07], an open problem related to borderedness in the context of partial words was suggested by the fact that every partial word of length five that has more than two holes is simply bordered. The partial word $aa\diamond\diamond b$ shows that this bound on the number of holes for length five is tight. For length six, every partial word with more than two holes is simply bordered as well. What is the maximum number of holes $m_k(n)$ a partial word of length n over an alphabet of size k can have and still fail to be simply bordered? Some values for small n follow: $m_2(5) = 2$, $m_2(6) = 2$, $m_2(7) = 3$, $m_2(8) = 4$, $m_2(9) = 5$, $m_2(10) = 5$, $m_2(11) = 6$, $m_2(12) = 7$, $m_2(13) = 8$, $m_2(14) = 8$, and $m_2(15) = 9$. The following theorem gives an answer to this problem.

Theorem 25. *For $k \geq 2$ and $l \geq 1$, the following equalities hold: $m_k(1) = 1$, $m_k(2l) = 2l - \left(\left\lfloor \sqrt{l} \right\rfloor + \left\lceil \frac{l}{\lfloor \sqrt{l} \rfloor} \right\rceil\right)$ and $m_k(2l+1) = m_k(2l) + 1$.*

Proof. Let A be a k-letter alphabet where $k \geq 2$, and let a, b be two distinct letters of A. We prove the lower bound by constructing a partial word $w(n)$ of length n over A with $m_k(n)$ holes, that is not simply bordered. Take $w(1) = \diamond$, and for $l \geq 1$,

$$\begin{aligned} w(2l) &= (a\diamond^{\lfloor \sqrt{l} \rfloor - 1})^{\frac{l}{\lfloor \sqrt{l} \rfloor}} \diamond^{l - \lfloor \sqrt{l} \rfloor} b^{\lfloor \sqrt{l} \rfloor} \\ w(2l+1) &= (a\diamond^{\lfloor \sqrt{l} \rfloor - 1})^{\frac{l}{\lfloor \sqrt{l} \rfloor}} \diamond^{l + 1 - \lfloor \sqrt{l} \rfloor} b^{\lfloor \sqrt{l} \rfloor} \end{aligned}$$

where a fractional power of the form $(a_0 \cdots a_{i-1})^{\frac{mi+j}{i}}$ with $0 \leq j < i$ is equal to $(a_0 \cdots a_{i-1})^m a_0 \cdots a_{j-1}$. It is easy to check that for this construction we never have a prefix of $w(n)$ of length at most $\left\lfloor \frac{n}{2} \right\rfloor = l$ compatible with a suffix. For $n \geq 4$, this is due to the fact that no factor of length $\lfloor \sqrt{l} \rfloor + 1$ of the prefix is compatible with the suffix $\diamond b^{\lfloor \sqrt{l} \rfloor}$, since each such factor has an a among its last $\lfloor \sqrt{l} \rfloor$ positions.

Now, we prove the upper bound. Let us observe that the simply bordered words of odd length are not influenced by the middle character. Hence, this character can always be replaced by a hole so that the number of holes is maximal. Because of that we can only look at the even length case. Let us consider a partial word $w = a_0 \cdots a_{2l-1}$ of length $n = 2l \geq 4$ that is not simply bordered. Obviously both a_0 and a_{2l-1} are distinct letters of the alphabet A in order to avoid a trivial one-letter border. Note that any two factors $a_0 \cdots a_{i-1}$ and $a_{2l-i} \cdots a_{2l-1}$ differ in at least one position for any $0 < i \leq l$. In order to avoid having the second half of w formed only of letters, we need in the first half at least two occurrences of letters. Let us suppose that a_i is the second occurrence of a letter in the first half of w (the first occurrence is a_0, that is, a_0 and a_i are letters and between them there are only holes). This implies that $a_{2l-i} \cdots a_{2l-1} \in A^*$. In other words, the suffix of length l of the word ends with a word of length i over A, since otherwise we again would get compatibility for a shorter factor. Now if we look at the prefix of length $2i$, we observe that we need a second incompatibility relation with the suffix of the same length. This implies that there exists another occurrence of a letter either in the prefix at a position j, with $j \leq 2i$, or in the suffix at position j, with $j > n - 2i$. Continuing the reasoning, and looking at the problem for the following occurrences of letters in each half, we will finally get an expression of the form $i + \frac{l}{i}$ for which we have to find the minimum value, for $0 < i \leq l$. Calculating the first derivative of $i + \frac{l}{i}$ and equating to zero, we get that $i = \sqrt{l}$. Hence the minimum number of letters (it is the number of holes that we wish to be maximized) is $\left\lfloor \sqrt{l} \right\rfloor + \left\lceil \frac{l}{\lfloor \sqrt{l} \rfloor} \right\rceil$, i.e., the number of consecutive occurrences of letters from the end of the word plus the number of occurrences of letters in the first half of the word. Furthermore, we observe that the upper bound coincides with the lower bound and the obtained computer values. □

Note that Theorem 25 implies that the equality $m_k(n) = m_2(n)$ holds for all $k \geq 2, n \geq 1$.

5.4 Bordered partial words

The previously defined concept of the maximum number of holes a "non-simply bordered" partial word may have can be extended to an "unbordered" partial word. Let $\hat{m}_k(n)$ be the maximum number of holes a partial word of length n over a k-letter alphabet can have and still fail to be bordered. For all integers $k \geq 2$ and $n \geq 1$, the inequality

$$\hat{m}_k(n) \leq m_2(n) \tag{5.4}$$

holds. To see this, consider a partial word u of length n over a k-letter alphabet with more than $m_k(n)$ holes. The word u necessarily has a simple border, so u is bordered and $\hat{m}_k(n)$ cannot be greater than $m_k(n)$. The inequality then follows by Theorem 25 which implies that $m_k(n) = m_2(n)$.

Next, let us refine the upper bound (5.4).

Proposition 22. *For all integers $k \geq 2$ and $n \geq 2$, we have the upper bound*

$$\hat{m}_k(n) \leq \left\lfloor n - \sqrt{\frac{2k}{k-1}(n-1)} \right\rfloor$$

Proof. Consider a partial word u of length n over a k-letter alphabet. Say $u = x_1 v = w x_2$, for some partial words x_1, x_2, v and w with x_1, x_2 of length i. For u not to have a border of length i, there must exist a pair of corresponding positions from x_1, x_2 whose letters are non-compatible. Since there exist $n-1$ possible border lengths for u, there must exist at least $n-1$ such pairs of non-compatible letters for u to be unbordered.

For a given number of letters $n-h$, the maximum number of non-compatible pairs will occur when each symbol of the alphabet appears equally, which would be $\frac{n-h}{k}$ times. Thus, the maximum number of non-compatible pairs is bounded above by

$$\left(\frac{n-h}{k}\right)^2 (k-1+k-2+\cdots+1) = \left(\frac{n-h}{k}\right)^2 \left(\frac{k(k-1)}{2}\right)$$

If there are strictly less than $n-1$ non-compatible pairs of letters in u, then u is necessarily bordered. So when $n-1 > \frac{(n-h)^2(k-1)}{2k}$ holds, u will be bordered. Thus, a word with $h > n - \sqrt{\frac{2k}{k-1}(n-1)}$ holes is necessarily bordered. So we have $\hat{m}_k(n) \leq \lfloor n - \sqrt{\frac{2k}{k-1}(n-1)} \rfloor$, for all integers $k \geq 2$ and $n \geq 2$. □

5.4.1 A formula for two-letter alphabets

First, we consider the 2-letter alphabet $\{a, b\}$. For $n \geq 1$, the upper bound

$$\hat{m}_2(n) \leq \lfloor n - 2\sqrt{n-1} \rfloor \tag{5.5}$$

follows from Proposition 22 by letting $k = 2$ (note that the case when $n = 1$ is trivial since $\diamond$ is an unbordered word of length one with one hole). We will show that this upper bound is also a lower bound.

Proposition 23. *For all integers $i, j, k \geq 0$ where $k \leq i$, the partial word given by $(a\diamond^i)^j a\diamond^k ab^{i+1}$ is an unbordered word of length $(i+1)(j+1)+k+2$.*

Proof. Assume that $i, j \geq 1$ (the other cases are similar). Consider a prefix of length l with $1 \leq l < (i+1)$. This gives us the prefix $a\diamond^{l-1}$ and the corresponding suffix b^l. Thus, there is no border of this length. Next, consider a prefix of length l with $(i+1) \leq l < (i+1)j+k+2$. Since an a will appear within at least the last $i+1$ letters of the prefix and the corresponding position in the suffix will be b, there cannot be a border of this length. Now, consider a prefix of length l with $(i+1)j+k+2 \leq l \leq (i+1)(j+1)+k+1$. The prefix ends with $ab^{i'}$, where $i' \leq i$. Since the suffix ends with b^{i+1}, the last a in the prefix does not agree with the corresponding b in the suffix. □

Proposition 24. *For all integers $n \geq 5$, we have the lower bound*

$$\hat{m}_2(n) \geq \lfloor n - 2\sqrt{n-1} \rfloor$$

Proof. First, assume there exists an integer $l \geq 2$ such that $n = l^2 + 1$. We construct the binary word $(a\diamond^{l-1})^{l-1}ab^l$ of length l^2+1 which is unbordered for all integers $l \geq 2$ by Proposition 23. This word has $(l-1)^2$ holes. Making the substitution $n = l^2 + 1$ we have

$$\lfloor n - 2\sqrt{n-1} \rfloor = \lfloor l^2 + 1 - 2\sqrt{l^2} \rfloor = l^2 + 1 - 2l = (l-1)^2$$

Thus, there exists an unbordered word of length n with $\lfloor n - 2\sqrt{n-1} \rfloor$ holes.

Now, assume n cannot be written in the form l^2+1 for any integer l. Let

$$i = \left\lfloor \frac{-1+\sqrt{1+4(n-2)}}{2} \right\rfloor + 1, j = \lceil \sqrt{n} \rceil - 3, k = n-(i+1)(j+1)-2$$

Let $u = (a\diamond^i)^j a \diamond^k ab^{i+1}$ whose length is given by $(i+1)(j+1)+k+2$ which is equivalent to $(i+1)(j+1)+n-(i+1)(j+1)-2+2 = n$. The number of holes in u is given by $ij+k = ij+n-(i+1)(j+1)-2 = n-i-j-3 = n - \left\lfloor \frac{-1+\sqrt{1+4(n-2)}}{2} \right\rfloor - \lceil \sqrt{n} \rceil - 1$. In order to show that u has $\lfloor n-2\sqrt{n-1} \rfloor$ holes, it suffices to show $\left\lfloor \frac{-1+\sqrt{1+4(n-2)}}{2} \right\rfloor + \lceil \sqrt{n} \rceil + 1 = \lceil 2\sqrt{n-1} \rceil$. First we note that for any integer $n \geq 5$, there exists a unique integer $m \geq 2$ such that $(m-1)^2 < n \leq m^2$. The next four bounds will be useful:

First, for $(m-1)(m-2)+2 \leq n < m(m-1)+2$, we have that $0 \leq 4(m-1)(m-2) \leq 4(n-2) < 4m(m-1)$. Thus, after adding 1 to, taking the square root of, subtracting 1 from, and dividing by 2 each part of the inequality yields $m-2 \leq \frac{-1+\sqrt{1+4(n-2)}}{2} < m-1$, and we get $\left\lfloor \frac{-1+\sqrt{1+4(n-2)}}{2} \right\rfloor = m-2$.

Second, for $(m-1)^2 < n \leq m^2$, we have that $(m-1) < \sqrt{n} \leq m$ and so $\lceil \sqrt{n} \rceil = m$.

Third, for $(m-1)^2+1 < n \leq m(m-1)+1$, we have $0 \leq (m-1)^2+1 < n \leq m(m-1)+1.25$. Thus, after subtracting 1 from, taking the square root of, and multiplying by 2 each part of the inequality, this yields $2m-2 < \lceil 2\sqrt{n-1} \rceil \leq 2m-1$. Hence, we have $\lceil 2\sqrt{n-1} \rceil = 2m-1$.

Fourth, for $m(m-1)+2 \leq n \leq m^2+1$, we have $m(m-1)+1.25 < n \leq m^2+1$. Thus, after subtracting 1 from, taking the square root of, and multiplying by 2 each part of the inequality, this yields $2m-1 < 2\sqrt{n-1} \leq 2m$. Thus, we have $\lceil 2\sqrt{n-1} \rceil = 2m$.

Now, if $(m-1)^2+1 < n \leq m(m-1)+1$, then $\left\lfloor \frac{-1+\sqrt{1+4(n-2)}}{2} \right\rfloor = m-2$, $\lceil \sqrt{n} \rceil = m$, and $\lceil 2\sqrt{n-1} \rceil = 2m-1$. If $m(m-1)+2 \leq n \leq m^2$, then

$\left\lfloor \frac{-1+\sqrt{1+4(n-2)}}{2} \right\rfloor = m-1$, $\lceil \sqrt{n} \rceil = m$, and $\lceil 2\sqrt{n-1} \rceil = 2m$.

Finally we claim that u is unbordered. By Proposition 23, it suffices to show that $k \leq i$. This is equivalent to demonstrating

$$n \leq 2\lceil \sqrt{n} \rceil - \left\lfloor \frac{-1+\sqrt{1+4(n-2)}}{2} \right\rfloor + \lceil \sqrt{n} \rceil \left\lfloor \frac{-1+\sqrt{1+4(n-2)}}{2} \right\rfloor - 1$$

Again, let m be the unique integer such that $(m-1)^2 < n \leq m^2$. If $(m-1)^2+1 < n \leq m(m-1)+1$, then $\left\lfloor \frac{-1+\sqrt{1+4(n-2)}}{2} \right\rfloor = m-2$ and $\lceil \sqrt{n} \rceil = m$. So we have $n \leq m(m-1)+1 = 2m-(m-2)+m(m-2)-1$. If $m(m-1)+1 < n \leq m^2$, then $\left\lfloor \frac{-1+\sqrt{1+4(n-2)}}{2} \right\rfloor = m-1$ and $\lceil \sqrt{n} \rceil = m$. We get

$$n \leq m^2 = 2m-(m-1)+m(m-1)-1$$

□

Theorem 26. *For any integer $n \geq 1$, $\hat{m}_2(n) = \lfloor n - 2\sqrt{n-1} \rfloor$.*

Proof. For $n = 1$, the result is trivial as mentioned earlier. For $n = 2$, note that a word with at least one hole necessarily has a border of length one. An unbordered word of length two with no hole is ab, and $\hat{m}_2(2) = 0 = \lfloor 2 - 2\sqrt{1} \rfloor$. For $n = 3$, $\hat{m}_2(3) = 0 = \lfloor 3 - 2\sqrt{2} \rfloor$, and an example of an unbordered word of length three with no hole is abb. As in the case of words of length two, a word that has one hole will be bordered. For $n = 4$, we can argue similarly. Thus, we have as example $abbb$, and $\hat{m}_2(4) = 0 = \lfloor 4 - 2\sqrt{3} \rfloor$. For $n \geq 5$, the result follows from (5.5) and Proposition 24. □

5.4.2 A lower bound for three-letter alphabets

Now, we consider the 3-letter alphabet $\{a, b, c\}$. For $n \geq 2$, the upper bound

$$\hat{m}_3(n) \leq \left\lfloor n - \sqrt{3(n-1)} \right\rfloor \tag{5.6}$$

follows from Proposition 22 by letting $k = 3$. We will give a lower bound for $\hat{m}_3(n)$.

Proposition 25. *For all integers $i, j, k \geq 0$ with $k \leq i$, the partial word given by $(a\diamond^i)^j a \diamond^k c^i b$ is an unbordered word of length $(i+1)(j+1)+k+1$.*

Proof. Assume that $i, j, k > 0$ (the other cases are similar). Consider a possible border length l with $1 \leq l \leq i+1$. This yields a prefix that begins with a and a suffix which begins in b or c, so there is no border of length l. If $i+2 \leq l \leq j(i+1)+1$, we have the letter a within the last $i+1$ positions of the prefix which will correspond with c or b in the last $i+1$ positions of the suffix, so there is no border of this length. If $j(i+1)+2 \leq l \leq j(i+1)+1+k$, we have a appearing within the last $k+1$ positions of the prefix, and since $k \leq i$, the last $k+1$ positions of the suffix are c's and b's. Finally, if $j(i+1)+k+2 \leq l \leq (j+1)(i+1)+k$, we have a prefix that ends in c and a suffix which ends in b. □

Proposition 26. *For all integers $i, j \geq 2$ and $k \geq 0$, the partial word given by $(a\diamond^i)^j(b\diamond^{i+1})^k c^i b$ is an unbordered word of length $(i+1)(j+k+1)+k$.*

Proof. Assume that $i \geq 2, j \geq 2, k \geq 1$ (the case where $k = 0$ is similar). Consider a possible border length l with $1 \leq l \leq i+1$. Our prefix will begin with a which is not equal to the corresponding b or c in the suffix. For $i+2 \leq l \leq j(i+1)$, we have a within the last $i+1$ positions of the prefix, and the last $i+1$ positions of the suffix are $c^i b$. So there is no border of length l. For $j(i+1)+1 \leq l \leq j(i+1)+k(i+2)$, we have one of the following three cases:

If our prefix ends in b, then we have $l = j(i+1)+m(i+2)+1$ for some integer m with $0 \leq m < k$. In this case, the a at position $l-1-m(i+2)-2(i+1)$ of the prefix will correspond with b at this position of the suffix. So there is no border of length l. If our prefix ends with $b\diamond^{i'}$ such that $1 \leq i' \leq i$, then our prefix contains the letter b within the last $i+1$ positions, but not at the last position. However, the suffix will have c's in all of these positions. If our prefix ends with $b\diamond^{i+1}$ so that we have $l = j(i+1)+m(i+2)$ where $1 \leq m \leq k$, then the a at position $l-m(i+2)-(i+1)$ of the prefix will correspond with b at this position of the suffix. So there is no border of length l.

Finally, consider the case where $j(i+1)+k(i+2)+1 \leq l \leq j(i+1)+k(i+2)+i$. We will have a prefix which ends with c and a suffix which ends with b, so we have no border for this length. □

Proposition 27. *For any integer $n > 9$, we have the lower bound $\hat{m}_3(n) \geq n - \lceil 2\sqrt{n+3} \rceil + 2$.*

Proof. Let $l = \lceil \sqrt{n} \rceil$, so that $l \geq 4$ and $(l-1)^2 < n \leq l^2$. To show that $\hat{m}_3(n) \geq n - \lceil 2\sqrt{n+3} \rceil + 2$ is equivalent to showing that

$$\hat{m}_3(n) \geq \begin{cases} n - 2\lceil \sqrt{n} \rceil + 1 & \text{if } l^2 - 2 \leq n \leq l^2 \\ n - 2\lceil \sqrt{n} \rceil + 2 & \text{if } l(l-1) - 2 \leq n \leq l^2 - 3 \\ n - 2\lceil \sqrt{n} \rceil + 3 & \text{if } (l-1)^2 + 1 \leq n \leq l(l-1) - 3 \end{cases}$$

We consider the following five cases, and in each case demonstrate that there exists an unbordered word of length n with the required number of holes. Note that in each of the cases we have $l = \lceil \sqrt{n} \rceil$.

First, if $(l-1)^2 + 1 \leq n \leq l(l-1) - 3$, then let $u = (a\diamond^i)^j (b\diamond^{i+1})^k c^i b$ where $i = l-2$, $j = (l-1)l - (n+1)$, and $k = n - (l-1)^2$. The length of u is $(i+1)(j+k+1)+k = (l-1)((l-1)l-(n+1)+n-(l-1)^2+1)+n-(l-1)^2 = n$. The number of holes in u is $ij + (i+1)k = (l-2)(l(l-1) - (n+1)) + (l-1)(n-(l-1)^2) = n - 2l + 3$. By Proposition 26, to show u is unbordered it suffices to show $i, j \geq 2$. This case only holds for $l \geq 4$, so $i \geq 2$. Since $n \leq l(l-1) - 3$, we have $2 \leq l(l-1) - n - 1 = j$. Thus, there exists an unbordered word of length n with $n - 2\lceil \sqrt{n} \rceil + 3$ holes.

Second, if $l(l-1) - 2 \leq n \leq l(l-1)$, then let $u = (a\diamond^i)^j a \diamond^k c^i b$ where $i = l-1$, $j = l-3$, and $k = n - (l-1)^2$. The length of u is $(i+1)(j+1) + k + 1 = l(l-2) + n - l^2 + 2l - 1 + 1 = n$, and the number of holes in u is $ij + k = (l-1)(l-3) + n - l^2 + 2l - 1 = n - 2l + 2$. By Proposition 25, u is unbordered if $k \leq i$. Since $n \leq l(l-1)$, we have $n - l^2 + 2l - 1 = k \leq l - 1 = i$. Thus, there exists an unbordered word of length n with $n - 2\lceil \sqrt{n} \rceil + 2$ holes.

Third, if $l(l-1) + 1 \leq n \leq l^2 - 4$, then let $u = (a\diamond^i)^j (b\diamond^{i+1})^k c^i b$ where $i = l-1$, $j = l^2 - 2 - n$, and $k = n - l(l-1)$. The length of u is $(i+1)(j+k+1)+k = l(l^2-2-n+n-l^2+l+1)+n-l(l-1) = n$, and the number of holes in u is $ij + (i+1)k = (l-1)(l^2-2-n) + l(n - l(l-1)) = n - 2l + 2$. By Proposition 26, u is unbordered if $i, j \geq 2$. Since $n \geq 7$, it must be that $l \geq 3$, and so $i \geq 2$. Since $n \leq l^2 - 4$, we have $2 \leq l^2 - 2 - n = j$. Thus, there exists an unbordered word of length n with $n - 2\lceil \sqrt{n} \rceil + 2$ holes.

Fourth, if $n = l^2 - 3$, then let $u = (a\diamond^i)^2(b\diamond^{i+1})^k c^i b$ where $i = l - 2$, $j = 2$, and $k = l - 3$. The length of u is $(i + 1)(j + k + 1) + k = (l - 1)(2 + l - 3 + 1) + l - 3 = l^2 - 3 = n$, and the number of holes in u is $ij+(i+1)k = (l-2)(2)+(l-1)(l-3) = l^2-2l-1 = l^2-3-2l+2 = n-2l+2$. By Proposition 26, u is unbordered since $j = 2$ and $i \geq 2$, because $n \geq 7$ and $l \geq 4$. Thus, there exists an unbordered word of length n with $n-2\lceil\sqrt{n}\rceil+2$ holes.

Fifth, if $l^2 - 2 \leq n \leq l^2$, then let $u = (a\diamond^i)^j a\diamond^k c^i b$ where $i = l - 1$, $j = l - 2$, and $k = n - l(l - 1) - 1$. The length of u is $(i + 1)(j + 1) + k + 1 = l(l - 1) + n - l(l - 1) - 1 + 1 = n$. The number of holes in u is $ij + k = (l - 1)(l - 2) + n - l(l - 1) - 1 = n - 2l + 1$. By Proposition 25, u is unbordered if $k \leq i$. Since $n \leq l^2$, we have $n-l^2+l-1 = k \leq l-1 = i$. Thus, there exists an unbordered word of length n with $n - 2\lceil\sqrt{n}\rceil + 1$ holes. □

Note that our upper bound and lower bound for $\hat{m}_3(n)$ are equal for $n \leq 27$. We believe that our lower bound is tight and have the following conjecture.

Conjecture 1. *The equality* $\hat{m}_3(n) = n - \lceil 2\sqrt{n+3}\rceil + 2$ *holds for all* $n \geq 6$.

5.4.3 A lower bound for four-letter alphabets

Finally, we consider the 4-letter alphabet $\{a, b, c, d\}$. By letting $k = 4$ in Proposition 22, we have the upper bound

$$\hat{m}_4(n) \leq \left\lfloor n - \sqrt{\frac{8}{3}(n-1)} \right\rfloor \tag{5.7}$$

for $n \geq 2$. We will give a lower bound for $\hat{m}_4(n)$.

Proposition 28. *The partial word* $a\diamond^i(b\diamond^{i+1})^j c^i d$ *is an unbordered word of length* $(i + 2)(j + 1) + i$, *for all* $i, j \geq 0$ *and distinct letters* a, b, c, d.

Proof. We assume that $i, j \geq 1$ (the other cases are similar). Consider a border length l. If $1 \leq l \leq i + 1$, then we have a prefix which begins with a and a suffix which begins with c or d. If $i + 2 \leq l < (j + 1)(i + 2)$, then the prefix ends in either $b\diamond^{i'}$ or $b\diamond^{i+1}$, where $0 \leq i' \leq i$. If the prefix ends with $b\diamond^{i'}$, then we have the letter b appearing within the last $i + 1$ positions of the prefix which will correspond with either c or d in the suffix. If the prefix ends with $b\diamond^{i+1}$, then our prefix will begin with a, and our suffix will begin with b. If $(j+1)(i+2) \leq l \leq 2i+1+j(i+2)$, then our prefix ends with the letter c while our suffix ends with the letter d. In each case, there is no border of length l. □

Proposition 29. *For integers $n \geq 7$, we have the lower bound*

$$\hat{m}_4(n) \geq \begin{cases} l(l-2) & \textit{if } n = l^2 - 2 \textit{ for some integer } l \\ l^2 - l - 1 & \textit{if } n = l^2 + l - 2 \textit{ for some integer } l \\ \hat{m}_3(n) & \textit{otherwise} \end{cases}$$

Proof. First, suppose that $n = l^2 - 2$ for some integer l. Let $i = j = l - 2$. The word $a\diamond^i(b\diamond^{i+1})^j c^i d$ is unbordered by Proposition 28. The length of this word is $2i + 2 + j(i+2) = 2(l-2) + 2 + (l-2)l = l^2 - 2 = n$. The number of holes in the word is $i + j(i+1) = l - 2 + (l-2)(l-1) = l(l-2)$.

Next, suppose that $n = l^2 + l - 2$ for some integer l. Now let $i = l - 2, j = l - 1$. We have the word $a\diamond^i(b\diamond^{i+1})^j c^i d$, which is unbordered by Proposition 28. The length of this word is $2i + 2 + j(i+2) = 2(l-2) + 2 + (l-1)l = l^2 + l - 2 = n$. The number of holes in this word is $i + j(i+1) = l - 2 + (l-1)(l-1) = l^2 - l - 1$.

For all other n, consider an unbordered word with $\hat{m}_3(n)$ holes, which is still unbordered over an alphabet of size 4. □

Note that our lower bound can be improved when $n = 24, 35, 48, 63, 80$ and 99. For instance, $\hat{m}_4(24) = 16 > \hat{m}_3(24) = 15$.

$\hat{m}_4(24) = 16$	$a\diamond^1 b\diamond^2 c\diamond^2 c\diamond^3 b\diamond^8 add$
$\hat{m}_4(35) = 25$	$a\diamond^2 a\diamond^2 b\diamond^3 c\diamond^3 c\diamond^4 b\diamond^{11} addd$ $a\diamond^2 d\diamond^2 d\diamond^3 b\diamond^3 b\diamond^4 a\diamond^{11} cccd$
$\hat{m}_4(48) = 36$	$a\diamond^1 b\diamond^2 b\diamond^2 c\diamond^3 c\diamond^3 c\diamond^3 a\diamond^4 a\diamond^{18} bddd$ $a\diamond^1 d\diamond^2 a\diamond^2 a\diamond^3 b\diamond^3 b\diamond^3 b\diamond^4 c\diamond^{18} dcdd$ $a\diamond^1 d\diamond^2 d\diamond^2 b\diamond^3 b\diamond^3 b\diamond^3 a\diamond^4 a\diamond^{18} cccd$
$\hat{m}_4(63) \geq 49$	$a\diamond^2 a\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^4 c\diamond^4 a\diamond^5 c\diamond^{22} dcddd$ $a\diamond^2 a\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 c\diamond^4 c\diamond^4 a\diamond^5 c\diamond^{22} bdddd$ $a\diamond^2 a\diamond^3 d\diamond^3 a\diamond^3 d\diamond^3 d\diamond^4 b\diamond^4 b\diamond^5 a\diamond^{22} ccccd$
$\hat{m}_4(80) \geq 64$	$a\diamond^1 b\diamond^2 b\diamond^2 c\diamond^3 c\diamond^3 c\diamond^3 c\diamond^3 c\diamond^3 c\diamond^3 c\diamond^3 a\diamond^4 a\diamond^{34} bddd$ $a\diamond^1 d\diamond^2 a\diamond^2 a\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^4 c\diamond^{34} dcdd$ $a\diamond^1 d\diamond^2 d\diamond^2 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 a\diamond^4 a\diamond^{34} cccd$
$\hat{m}_4(99) \geq 81$	$a\diamond^2 a\diamond^2 b\diamond^3 c\diamond^3 c\diamond^3 c\diamond^3 c\diamond^3 c\diamond^3 c\diamond^3 c\diamond^3 c\diamond^3 c\diamond^3 c\diamond^4 b\diamond^{43} addd$ $a\diamond^2 d\diamond^2 d\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^3 b\diamond^4 a\diamond^{43} cccd$

This leads us to the following conjecture.

Conjecture 2. *The equality* $\hat{m}_4(l^2 - 1) = (l - 1)^2$ *holds for all* $l > 2$.

5.5 Critical factorizations

In this section, we first discuss so-called critical factorizations of a partial word w, then study some of their properties when w is unbordered (Proposition 30, and Corollaries 9 and 10), and finally investigate the position in the Chomsky hierarchy of the set of all partial words having a critical factorization (Theorems 27 and 28).

If w is a *non-special* partial word of length at least two, then there exists a factorization (u, v) of w with $u, v \neq \varepsilon$ such that the minimal local period of w at position $|u| - 1$ (as defined below) equals the minimal weak period of w [BSD05, BSW07]. Such a factorization (u, v) of w is called *critical* and the position $|u| - 1$ is called a *critical point* of w.

Definition 4. [BSD05] *Let* $w \in A_\diamond^+$. *A positive integer* p *is called* a local period of w at position i *if there exist* $u, v, x, y \in A_\diamond^+$ *such that* $w = uv$, $|u| = i + 1$, $|x| = p$, $x \uparrow y$, *and such that one of the following conditions holds for some partial words* r, s:

1. $u = rx$ *and* $v = ys$ *(internal square),*

2. $x = ru$ and $v = ys$ (left-external square if $r \neq \varepsilon$),

3. $u = rx$ and $y = vs$ (right-external square if $s \neq \varepsilon$),

4. $x = ru$ and $y = vs$ (left- and right-external square if $r, s \neq \varepsilon$).

The minimal local period of w at position i is denoted by $p(w, i)$. Clearly, $1 \leq p(w, i) \leq p'(w) \leq |w|$.

There exist unbordered partial words that have no critical factorizations, like $w = a\diamond bc$ [Wet].

We now investigate some of the properties of an unbordered partial word of length at least two and how they relate to its critical factorizations (if any).

Definition 5. *Let $u, v \in A_\diamond^+$. We say that u and v overlap if there exist partial words r, s satisfying one of the following conditions:*

1. $r \uparrow s$ with $u = ru'$ and $v = v's$,

2. $r \uparrow s$ with $u = u'r$ and $v = sv'$,

3. $u = ru's$ with $u' \uparrow v$,

4. $v = rv's$ with $v' \uparrow u$.

Otherwise we say that u and v do not overlap.

Proposition 30. *Let $u, v \in A_\diamond^+$. If $w = uv$ is unbordered, then $|u| - 1$ is a critical point of w if and only if u and v do not overlap.*

Proof. Let us first consider the first implication and let us suppose u and v overlap. If we have Type 1 overlap, then $w = ru'v's$ and $r \uparrow s$ for some partial words r, s, u', v'. This contradicts the fact that w is unbordered. If we have Type 2 overlap, then $w = u'rsv'$ and there is an internal square at position $|u| - 1$ of length $k = |r| = |s|$, so $p(w, |u| - 1) \leq k$. But because w is unbordered, $p'(w) = |w|$. Of course we have that $k < |w|$ (otherwise we have Type 1 overlap), so this contradicts that $|u| - 1$ is a critical point of w. If we have Type 3 overlap, then $w = ru'sv$ and there is a right-external square of length $|u's|$ at position $|u| - 1$. Because $v \neq \varepsilon$, $|u's| < |w| = p'(w)$ and we have that $|u| - 1$ cannot be a critical point of w, a contradiction. The case for Type 4 overlap is very similar to Type 3.

For the other direction we have that u and v do not overlap and let us suppose that $|u|-1$ is not a critical point of w.

Since $|u|-1$ is not a critical point, there exist x and y defined as in Definition 4, with the length of x strictly smaller than the minimal weak period of w. Let us now look at all the four conditions of the definition. If we have an internal square, then according to Definition 5 we have a Type 2 overlap of u and v, which is a contradiction with our assumption. For a left-external, respectively right-external, square we get that either u is compatible with a factor of v, or v is compatible with a factor of u. Both cases contradict with the fact that u and v do not overlap, giving us a Type 4, respectively Type 3, overlap.

In the case we have a left- and right-external square we get that $x = ru$ and $y = vs$, where $x \uparrow y$ and $r, s \neq \varepsilon$. If $|r| < |v|$, then there exists v' with $|v'| > 0$, such that $v = rv'$. Hence, since $ru \uparrow rv's$ we get a Type 2 overlap, $u \uparrow v's$, which is a contradiction with our initial assumption. If $|r| \geq |v|$, then there exists r' such that $r = vr'$. This implies that $|w| = |uv| \leq |vr'u| = |ru| = |x| < p'(w) \leq |w|$, which is a contradiction. □

Corollary 9. *Let $u, v \in A_\diamond^+$. If $w = uv$ is unbordered and $|u|-1$ is a critical point of w, then $w' = vu$ is unbordered as well.*

Proof. This is immediately implied by Proposition 30 and the fact that if $w' = vu$ is bordered, then u and v must overlap. □

Corollary 10. *Let $u, v \in A_\diamond^+$. If $w = uv$ is unbordered and $|u|-1$ is a critical point of w, then $|v|-1$ is a critical point of $w' = vu$.*

Proof. By Proposition 30, u and v do not overlap. By Corollary 9, w' is unbordered. Then by Proposition 30, the point $|v|-1$ is critical for w'. □

We end this section by considering the language

$$CrFa = \{w \mid w \text{ is a partial word over } A \text{ that has a critical factorization}\}$$

where A denotes an arbitrary non-unary fixed finite alphabet (we will assume that a and b are two distinct letters of A). What is the position of $CrFa$ in the Chomsky hierarchy? Due to [Zha], we know that $CrFa$ it is not regular. We prove that $CrFa$ is a context sensitive language that is not context-free.

Let us first recall a version of the pumping lemma that is due to Bader and Moura [BM82], and is a generalization of the well known Ogden's Lemma.

Lemma 28. [BM82] *For any context-free language L, there exists $n \in \mathbb{N}$, the set of non-negative integers, such that for all $z \in L$, if d positions in z are "distinguished" and e positions are "excluded," with $d > n^{(e+1)}$, then there exist u, v, w, x, y such that $z = uvwxy$ and*

1. *vx contains at least one distinguished position and no excluded positions,*
2. *if r is the number of distinguished positions and s is the number of excluded positions in vwx, then $r \leq n^{(s+1)}$,*
3. *for all $i \in \mathbb{N}$, $uv^iwx^iy \in L$.*

The above lemma says that for any context-free language L, there exists a natural number n, such that in any word $z \in L$, by marking any d positions as "distinguished" and e positions as "excluded" with $d > n^{(e+1)}$, we can decompose z in five contiguous factors that satisfy the three statements. It is easy to observe that the only restrictions imposed by d and e are on the three inner factors v, w and x.

Theorem 27. *The language $CrFa$ is not context-free.*

Proof. Let us assume that the language $CrFa$ is context-free. This implies that the previously defined pumping lemma holds. Let us take the word

$$z = ba^{3n^3}ba^{n^3}\diamond^{n^3}a^{n^3}ba^{3n^3}b$$

where n is the natural number from the lemma, and mark all symbols except the first and the last one as distinguished and these two as excluded. It is easy to check that $p'(z) = 3n^3 + 1$, z has a critical factorization (b, $a^{3n^3}ba^{n^3}\diamond^{n^3}a^{n^3}ba^{3n^3}b$) and the number of distinguished positions is greater than $n^{(2+1)}$. From Lemma 28(1) we get that the first and the last occurrences of b will never be part of either v or x.

Let us first consider the case when $u = \varepsilon$. This implies, by Lemma 28(1), that $v = \varepsilon$. Hence, w contains exactly one excluded position, implying $x = a^k$, where $0 < k \leq n^2$ by Lemma 28(2). In this case, for $i = 0$, we obtain the word

$$ba^{3n^3-k}ba^{n^3}\diamond^{n^3}a^{n^3}ba^{3n^3}b$$

which is not in $CrFa$, contradicting Lemma 28(3). To see that this word does not have a critical factorization, note that it has minimal weak period greater than $3n^3+1$. However, the minimal local periods at the positions defined by the factorization $(b, a^{3n^3-k}, b, a^{n^3}\diamond^{n^3}a^{n^3}, b, a^{3n^3}, b)$ are $3n^3-k+1, 3n^3-k+1, n^3+1, n^3+1, 3n^3+1$ and $3n^3+1$ respectively, while the minimal local period at any other position is 1. Similarly we easily prove that it is impossible to have $y=\varepsilon$.

From now on, let us consider the cases where both u and y are nonempty. Then each of u and y contains an excluded position and so vwx will all be distinguished. And therefore the length of vwx is at most n by Lemma 28(2).

When vwx matches a^* and vwx is part of the 1st group of a's, then $vwx = a^k$ for some $0 < k \leq n$, and $v = a^{k_1}$ and $x = a^{k_2}$ with $k_1 > 0$ or $k_2 > 0$. In this case take $i = 0$. The 1st group of a's is then reduced to $3n^3 - k_1 - k_2$, giving us the word

$$ba^{3n^3-k_1-k_2}ba^{n^3}\diamond^{n^3}a^{n^3}ba^{3n^3}b$$

that does not have a critical factorization (again, the minimal weak period is greater than $3n^3+1$ while the minimal local periods are smaller than or equal to $3n^3+1$). A similar argument works for the 2nd, 3rd and 4th groups of a's. We are left with the cases when vwx matches a^*ba^*, or $a^*\diamond^*$ or $\diamond^*a^*$.

If x matches a^*ba^*, then v is a string of a's of length at most $n-1$ with the a's either from the 1st group or the 3rd group. In both cases, taking $i=0$, we get a contradiction with the fact that the words

$$ba^{4n^3-k_1}\diamond^{n^3}a^{n^3}ba^{3n^3}b$$

and

$$ba^{3n^3}ba^{n^3}\diamond^{n^3}a^{4n^3-k_2}b$$

are in $CrFa$ for some $0 \leq k_1, k_2 < n$. To see that the first word does not have a critical factorization, note that it has minimal weak period greater than $4n^3+1$. However, the minimal local periods at the positions defined by the factorization $(b, a^{4n^3-k_1}\diamond^{n^3}a^{n^3}, b, a^{3n^3}, b)$ are $4n^3-k_1+1, n^3+1, 3n^3+1$ and $3n^3+1$ respectively, while the minimal local period at any other position is 1. The case where v matches a^*ba^* is solved analogously to the previous one and hence we will omit its proof. By taking $i = 2$, a contradiction is reached in the cases where $v = a^{k_1}$ and $x = a^{k_2}$ for some k_1, k_2 with the a's in v from the 1st group of a's, and the ones in x from the 2nd group of a's (respectively, with the a's in v from the 3rd group of a's, and the ones in x from the 4th group of a's).

If $v = a^{k_1}\diamond^{k_2}$ or $x = \diamond^{k_1}a^{k_2}$ for some k_1, k_2, then we get that $x = \diamond^{k_3}$, respectively $v = \diamond^{k_3}$, with $0 < k_1 + k_2 + k_3 \leq n$. In both cases, taking $i = 2$, we obtain a word that does not have a critical factorization. When $v = \diamond^{k_1}a^{k_2}$ or $x = a^{k_1}\diamond^{k_2}$, we proceed similarly. The case $vx = a^k$ where $0 < k \leq n$, with the a's from the 2nd or the 3rd group, is solved similarly.

Since all cases lead to contradictions we conclude that our assumption is false, hence the language $CrFa$ is not context-free. □

Theorem 28. *The language $CrFa$ is context sensitive.*

Proof. To prove this we will give an LBA (linear bounded automaton) that recognizes all partial words having a critical factorization. We recall that the factorization (u, v) of input partial word w is critical if the minimal local period of w at position $|u| - 1$ is equal to the minimal weak period of w, $p'(w)$.

Our LBA will have an input tape of size $3|w|$ and five auxiliary tapes of size at most $|w| + 1$, that we are going to describe next. We will denote the word on the input tape as inp.

The input tape will contain, starting from position $|w|$, the input word while all other positions will be filled in with $\diamond$'s. Position $|w|$ (respectively, $2|w| - 1$) on the input tape can be easily recognized by using an auxiliary symbol \$ (respectively, #).

The first auxiliary tape, let us call it P, will have size $|w|$ and will be used for the identification of the minimal weak period of our input word w. This can be easily done by using an unary numbering system that adds 1's until the minimal weak period is discovered. Since the minimal weak period of a word is greater than or equal to one, we start with a 1 symbol on the tape.

The second tape, Z, will be used for remembering the current position in the word. Hence, for position $i < |w|$, the head will be positioned on the input tape on the $(|w| + i)$th cell, and Tape Z will contain i ones. The tape is initialized with one 1 and has size $|w| + 1$.

The following tape, X, will have size $p'(w)$ and will be used for checking the size of the current minimal local period.

The last two tapes, called Y_1 and Y_2, will have sizes $p'(w)$. They will be used to save the words of length at most $p'(w)$, positioned to the left and right of the current position. More exactly these tapes will contain x and y from the definition of critical factorization.

We now describe how the LBA works, using the notation $|T|$ for denoting the number of symbols present on Tape T:

1. Starting at position $|w|$ on the input tape, the head marks the current position and then moves to the right $|P|$ positions and checks if the symbols are compatible. This step is repeated until the condition is violated. If this happens, then a 1 is added to Tape P and all symbols are unmarked. If the end of the word is reached, then the head moves left to the position $|w|$ and repeats the step for the first unmarked symbol. The step is repeated until all symbols are marked or $|P| = |w|$. This will give us the minimal weak period of the word.

2. Increment the value of X.

3. Starting at position i, where i represents the sum between $|w|$ and the number of 1's on Tape Z, the LBA copies the suffix of length $|X|$ (recall that the number of symbols present on Tape X, or $|X|$, is bounded by $p'(w)$) of the word inp$[0..i)$ on Tape Y_1 and the prefix of length $|X|$ of the word inp$[i..3|w|)$ on Tape Y_2.

4. Next the LBA checks if the word on Tape Y_1 is compatible with the word on Tape Y_2. This can easily be done just by comparing one symbol at a time while going in parallel on the two tapes. If the words are compatible and the sum of 1's in X is equal to $p'(w)$, then the automaton stops and outputs the position where a critical factorization is present (the LBA will accept the word). If the words are compatible and the sum of 1's in X is not equal to $p'(w)$, then the automaton fills the X tape with 1's and goes to the next step.

5. If X is full, then the tape is brought to the initial configuration and the LBA adds a 1 on Z. If Z is full, then the automaton stops and concludes that a critical factorization does not exist, hence, the LBA will reject the word. Otherwise, the LBA goes to Step 2.

It is easy to check that the algorithm will always stop. Since the construction of a linear bounded automaton that recognizes all partial words over $\{a, b\}$ having a critical factorization was possible, we conclude that $CrFa$ is a context sensitive language. □

5.6 Conclusion

In conclusion, note that the following conjecture is somehow natural, since increasing the length of a partial word by one is possible through the addition of at most one hole.

Conjecture 3. *The inequalities $\hat{m}_k(n) \leq \hat{m}_k(n+1) \leq \hat{m}_k(n) + 1$ hold for all $k \geq 2, n \geq 1$.*

This result would imply that, over the same alphabet a word of length $n+1$ can have at most one hole more than a word of length n. This would actually help us lower the upper bound in the case of the alphabets of 3 or more letters.

Now let us give some remarks about conjugacy on partial words. We call a word u a conjugate of v, and we write $u \sim v$, if $u = xy$ while $v = yx$ for some x and y. Equivalently, u and v are conjugate if and only if there exists a word z such that $uz = zv$ [LS62]. Clearly, $\sim$ is an equivalence relation. For two words u and v, $(\sqrt{u})^m \sim (\sqrt{v})^n$ if and only if both $m = n$ and $\sqrt{u} \sim \sqrt{v}$. Thus, every conjugate of a non-primitive non-empty word is bordered. A main result of Ehrenfeucht and Silberger states that if u is a primitive word such that $a \in \alpha(u)$, then there exists an unbordered conjugate av of u [ES79]. In other words, if u is such that $u = \sqrt{u}$ and $a \in \alpha(u)$, then there exist x, y such that $u = xay$ and $v = ayx$ is unbordered. For instance, if $u = aba$, then $x = ab$ and $y = \varepsilon$ work for the letter $a \in \alpha(u)$.

Now, partial words u and v are *conjugate* if there exist partial words x and y such that $u \subset xy$ and $v \subset yx$. Again, we denote u is a conjugate of v by $u \sim v$. Here, the relation $\sim$ is not an equivalence relation: it is both reflexive and symmetric, but not transitive [BSL02]. Note that the conjugates $a\diamond b$, $\diamond ba$ and $ba\diamond$ of $u = a\diamond b$ are bordered. However, we can easily extend Ehrenfeucht and Silberger's result by showing that if u is a primitive partial word and a is a letter of the alphabet A that appears in the spelling of u, then there exists an unbordered (full) conjugate av of u.

Index

Bibliography

[ABSBM09] Emily Allen, F. Blanchet-Sadri, Cameron Byrum, and Robert Mercaş. How many holes can an unbordered partial word contain? In Adrian Horia Dediu, Armand Mihai Ionescu, and Carlos Martín-Vide, editors, *Language and Automata Theory and Applications 2009*, volume 5457 of *Lecture Notes in Computer Science*, pages 176–187, Berlin, Germany, 2009. Springer-Verlag.

[AS99] Jean-Paul Allouche and Jeffrey Shallit. The ubiquitous Prouhet-Thue-Morse sequence. In C. Ding, T. Helleseth, and H. Niederreiter, editors, *Sequences and their Applications: Proceedings of SETA '98*, Discrete Mathematics and Theoretical Computer Science, pages 1–16. Springer-Verlag, 1999.

[AS03] Jean-Paul Allouche and Jeffrey Shallit. *Automatic Sequences: Theory, Applications, Generalizations*. Cambridge University Press, 2003.

[BB99] Jean Berstel and Luc Boasson. Partial words and a theorem of Fine and Wilf. *Theoretical Computer Science*, 218:135–141, 1999.

[BBSG$^+$09] Brandon Blakeley, F. Blanchet-Sadri, Josh Gunter, Sean Simmons, and Eric Weissenstein. Classifying all avoidable sets of partial words of size two. preprint, 2009.

[BBSGR09] Brandon Blakeley, Francine Blanchet-Sadri, Josh Gunter, and Narad Rampersad. On the complexity of deciding avoidability of sets of partial words. In *DLT '09: Proceedings of the 13th*

International Conference on Developments in Language Theory, pages 113–124, Berlin, Heidelberg, 2009. Springer-Verlag.

[BEM79] Dwight R. Bean, Andrzej Ehrenfeucht, and George McNulty. Avoidable patterns in strings of symbols. *Pacific Journal of Mathematics*, 85:261–294, 1979.

[Ber92] Jean Berstel. Axel Thue's work on repetitions in words. In Pierre Leroux and Christophe Reutenauer, editors, *Invited Lecture at the 4th Conference on Formal Power Series and Algebraic Combinatorics*, pages 65–80, 1992.

[BI80] P. Bylanski and D.G.W. Ingram. *Digital Transmission Systems*. Inspec/IEE, 1980.

[BJJ97] Dany Breslauer, Tao Jiang, and Zhigen Jiang. Rotations of periodic strings and short superstrings. *Journal of Algorithms*, 24(2):340–353, 1997.

[BM82] Christopher Bader and Arnaldo Moura. A generalization of Ogden's lemma. *Journal of ACM*, 29(2):404–407, 1982.

[BP85] Jean Berstel and Dominique Perrin. *Theory of Codes*. Academic Press, 1985.

[Bra83] F.-J. Brandenburg. Uniformly growing k-th power-free homomorphisms. *Theoretical Computer Science*, 23:69–82, 1983.

[BS04] F. Blanchet-Sadri. Codes, orderings, and partial words. *Theoretical Computer Science*, 329(1–3):177–202, 2004.

[BS05] F. Blanchet-Sadri. Primitive partial words. *Discrete Applied Mathematics*, 148(3):195–213, 2005.

[BS07] F. Blanchet-Sadri. Open problems on partial words. In Gemma Bel Bel-Enguix, Maria Dolores Jiménez-López, and Carlos Martín-Vide, editors, *New Developments in Formal Languages and Applications*, pages 11–58. Springer-Verlag, Berlin, 2007.

[BS08] F. Blanchet-Sadri. *Algorithmic Combinatorics on Partial Words*. Chapman & Hall/CRC Press, 2008.

[BSBK+09] F. Blanchet-Sadri, Naomi C. Brownstein, Andy Kalcic, Justin Palumbo, and Tracy Weyand. Unavoidable sets of partial words. *Theory of Computing Systems*, 45:381–406, 2009.

[BSBL06] F. Blanchet-Sadri, D. Dakota Blair, and Rebeca V. Lewis. Equations on partial words. In Markus Lohrey, editor, *Mathematical Foundations of Computer Science 2006*, volume 4162 of *Lecture Notes in Computer Science*, pages 167–178, Berlin/ Heidelberg, Germany, 2006. Springer-Verlag.

[BSBP07] F. Blanchet-Sadri, Naomi C. Brownstein, and Justin Palumbo. Two element unavoidable sets of partial words. In Kai Salomaa and Markus Holzer, editors, *Developments in Language Theory*, volume 4588 of *Lecture Notes in Computer Science*, pages 96–107, Berlin, Germany, 2007. Springer Berlin / Heidelberg.

[BSBS08] F. Blanchet-Sadri, Deepak Bal, and Gautam Sisodia. Graph connectivity, partial words, and a theorem of Fine and Wilf. *Information and Computation*, 206(5):676–693, 2008.

[BSCM09] F. Blanchet-Sadri, Ilkyoo Choi, and Robert Mercaş. Avoiding large squares in partial words. preprint, 2009.

[BSD05] F. Blanchet-Sadri and S. Duncan. Partial words and the critical factorization theorem. *Journal of Combinatorial Theory Series A*, 109(2):221–245, 2005.

[BSDD+09] F. Blanchet-Sadri, Crystal D. Davis, Joel Dodge, Robert Mercaş, and Margaret Moorefield. Unbordered partial words. *Discrete Applied Mathematics*, 157(5-6):890 – 900, 2009.

[BSH02] F. Blanchet-Sadri and Robert A. Hegstrom. Partial words and a theorem of Fine and Wilf revisited. *Theoretical Computer Science*, 270(1-2):401–419, 2002.

[BSJP09] F. Blanchet-Sadri, R. Jungers, and J. Palumbo. Testing avoidability on sets of partial words is hard. *Theoretical Computer Science*, 410:968–972, 2009.

[BSL02] F. Blanchet-Sadri and D. K. Luhmann. Conjugacy on partial words. *Theoretical Computer Science*, 289(1):297–312, 2002.

[BSM09] F. Blanchet-Sadri and Robert Mercaş. A note on the number of squares in a partial word with one hole. *RAIRO Theoretical Informatics and Applications*, 43:767–774, 2009.

[BSMRW09] F. Blanchet-Sadri, Robert Mercaş, Abraham Rashin, and Elara Willett. An answer to a conjecture on overlaps in partial words using periodicity algorithms. In Adrian Horia Dediu, Armand Mihai Ionescu, and Carlos Martín-Vide, editors, *Language and Automata Theory and Applications 2009*, volume 5457 of *Lecture Notes in Computer Science*, pages 188–199, Berlin, Germany, 2009. Springer-Verlag.

[BSMS08] F. Blanchet-Sadri, Robert Mercaş, and Geoffrey Scott. Counting distinct squares in partial words. In Erzsébet Csuhaj-Varju and Zoltan Esik, editors, *12th International Conference on Automata and Formal Languages*, pages 122–133, Balatonfüred, Hungary, 2008.

[BSMS09] F. Blanchet-Sadri, Robert Mercaş, and Geoffrey Scott. A generalization of Thue freeness for partial words. *Theoretical Computer Science*, 410(8-10):793–800, 2009.

[BSMSW10] F. Blanchet-Sadri, Robert Mercaş, Sean Simmons, and Eric Weissenstein. Avoidable binary patterns in partial words. In Adrian Horia Dediu, Henning Fernau, and Carlos Martín-Vide, editors, *Language and Automata Theory and Applications 2010*, volume 6031 of *Lecture Notes in Computer Science*, pages –, Berlin, Germany, 2010. Springer-Verlag.

[BSOR] F. Blanchet-Sadri, Taktin Oey, and Timothy Rankin. Computing weak periods of partial words. *International Journal of Foundations of Computer Science*, to appear.

[BSSW09] F. Blanchet-Sadri, Sean Simmons, and Eric Weissenstein. Avoidable patterns on partial words. preprint, 2009.

[BSW07] F. Blanchet-Sadri and Nathan D. Wetzler. Partial words and the critical factorization theorem revisited. *Theoretical Computer Science*, 385(1-3):179–192, 2007.

[Car83] Arturo Carpi. On the size of a squarefree morphism on a three letter alphabet. *Information Processing Letters*, 16(5):231–236, 1983.

[Car07] Arturo Carpi. On Dejean's conjecture over large alphabets. *Theoretical Computer Science*, 385(1-3):137–151, 2007.

[Cas93] Julien Cassaigne. Unavoidable binary patterns. *Acta Informatica*, 30:385–395, 1993.

[CK97] Christian Choffrut and Juhani Karhumäki. Combinatorics of words. In Grzegorz Rozenberg and Arto Salomaa, editors, *Handbook of Formal Languages*, volume 1, pages 329–438. Springer-Verlag, 1997.

[CP91] Maxime Crochemore and Dominique Perrin. Two-way string-matching. *Journal of the ACM*, 38(3):650–674, 1991.

[CR94] Maxime Crochemore and Wojciech Rytter. *Text algorithms.* Oxford University Press, Inc., New York, NY, USA, 1994.

[CR02] Maxime Crochemore and Wojciech Rytter. *Jewels of Stringology: Text algorithms.* World Scientific, Singapore, 2002.

[CR09a] James Currie and Narad Rampersad. Dejean's conjecture holds for n$\geq$27, 2009. to appear.

[CR09b] James Currie and Narad Rampersad. Dejean's conjecture holds for n$\geq$30. *Theoretical Computer Science*, 410(30-32):2885–2888, 2009.

[CR09c] James Currie and Narad Rampersad. A proof of Dejean's conjecture. preprint, 2009.

[Cro81] Maxime Crochemore. An optimal algorithm for computing the repetitions in a word. *Information Processing Letters*, 12(5):244–250, 1981.

[Cur93] James D. Currie. Open problems in pattern avoidance. *American Mathematical Monthly*, 100(1):790–793, 1993.

[Dej72] Françoise Dejean. Sur un théorème de Thue. *Journal of Combinatorial Theory*, 13(1):90–99, 1972.

[Dek76] F.M. Dekking. On repetitions of blocks in binary sequences. *Journal of Combinatorial Theory*, 20(3):262–299, 1976.

[DMT09] Adrian Diaconu, Florin Manea, and Catalin Tiseanu. Combinatorial queries and updates on partial words. In Miroslaw Kutylowski, Maciej Gebala, and Witold Charatonik, editors, *17th International Symposium on Fundamentals of Computation Theory*, volume 5699 of *Lecture Notes in Computer Science*, pages 96–108, Berlin/ Heidelberg, Germany, 2009. Springer.

[Duv82] Jean-Pierre Duval. Relationship between the period of a finite word and the length of its unbordered segments. *Discrete Mathematics*, 40(1):31–44, 1982.

[EJS74] Roger C. Entringer, Douglas E. Jackson, and J.A. Schatz. On nonrepetitive sequences. *Journal of Combinatorial Theory*, 16(2):159–164, 1974.

[ES79] Andrzej Ehrenfeucht and D.M. Silberger. Periodicity and unbordered segments of words. *Discrete Mathematics*, 26(2):101–109, 1979.

[FS95] Aviezri S. Fraenkel and R. Jamie Simpson. How many squares must a binary sequence contain? *Electronic Journal of Combinatorics*, 2:R2, 1995.

[FS98] Aviezri S. Fraenkel and R. Jamie Simpson. How many squares can a string contain? *Journal of Combinatorial Theory*, 82(1):112–120, 1998.

[GS83] Zvi Galil and Joel Seiferas. Time-space optimal string matching. *Journal of computer and system sciences*, 26:280–294, 1983.

[GS04] Dan Gusfield and Jens Stoye. Linear time algorithms for finding and representing all the tandem repeats in a string. *Journal of Computer and System Sciences*, 69(4):525–546, 2004.

[Hal64] M. Hall, Jr. Generators and relations in groups – the Burnside problem. In T. L. Saaty, editor, *Lectures on Modern Mathematics*, volume 2, pages 42–92. Wiley, 1964.

[Har06] Tero Harju. Combinatorics on words. In Zoltan Esik, Carlos Martín-Vide, and Victor Mitrana, editors, *Recent Advances in Formal Languages and Applications*, volume 25, pages 381–392, Berlin, Germany, 2006. Springer-Verlag.

[HHK09] Vesa Halava, Tero Harju, and Tomi Kärki. On the number of squares in partial words. Technical Report 896, Turku Center for Computer Science, Turku, Finland, 2009.

[HHKS09] Vesa Halava, Tero Harju, Tomi Kärki, and Patrice Séébold. Overlap-freeness in infinite partial words. *Theoretical Computer Science*, 410(8-10):943–948, 2009.

[Ili05] Lucian Ilie. A simple proof that a word of length n has at most $2n$ distinct squares. *Journal of Combinatorial Theory*, 112(1):163–164, 2005.

[Ili07] Lucian Ilie. A note on the number of squares in a word. *Theoretical Computer Science*, 380(3):373–376, 2007.

[KL06] Juhani Karhumäki and Arto Lepistö. Combinatorics on infinite words. In Zoltan Esik, Carlos Martín-Vide, and Victor Mitrana, editors, *Recent Advances in Formal Languages and Applications*, volume 25, pages 393–410, Berlin, Germany, 2006. Springer-Verlag.

[Leu04] Peter Leupold. Languages of partial words - how to obtain them and what properties they have. *Formal Grammars*, 7:179–192, 2004.

[Leu05] Peter Leupold. Partial words for DNA coding. In Grzegorz Rozenberg, Peng Yin, Erik Winfree, John H. Reif, Byoung-Tak Zhang, Max H. Garzon, Matteo Cavaliere, Mario J. Prez-Jimnez, Lila Kari, and Sudheer Sahu, editors, *10th International Workshop on DNA Computing*, volume 3384 of *Lecture*

Notes in Computer Science, pages 224–234, Berlin, Germany, 2005. Springer-Verlag.

[Lis06] Gerhard Lischke. Restoration of punctured languages and similarity of languages. *Mathematical Logic Quarterly*, 52(1):20–28, 2006.

[Lot97] M. Lothaire. *Combinatorics on Words*. Cambridge University Press, 1997.

[LS62] R. C. Lyndon and Marcel-Paul Schützenberger. The equation $a^m = b^n c^p$ in a free group. *Michigan Mathematical Journal*, 9(4):289–298, 1962.

[Mar41] A.A. Markov. Impossibility of certain algorithms in the theory of associative systems. *Doklady Akademii Nauk SSSR*, 55:587–590, 1941.

[MM93] Udi Manber and Gene Myers. Suffix arrays: a new method for on-line string searches. *SIAM J. Comput.*, 22(5):935–948, 1993.

[MM07] Florin Manea and Robert Mercaş. Freeness of partial words. *Theoretical Computer Science*, 389(1-2):265–277, 2007.

[MNC07] M. Mohammad-Noori and James D. Currie. Dejean's conjecture and sturmian words. *European Journal of Combinatorics*, 28(3):876–890, 2007.

[MS95] D. Margaritis and S. S. Skiena. Reconstructing strings from substrings in rounds. In *FOCS '95: Proceedings of the 36th Annual Symposium on Foundations of Computer Science*, pages 613–620, Washington, DC, USA, 1995. IEEE Computer Society.

[Pos47] E.L. Post. Recursive unsolvability of a problem of Thue. *Journal of Symbolic Logic*, 11:1–11, 1947.

[Rao09] Michaël Rao. Last cases of Dejean's conjecture. In *WORDS 2009, the 7th International Conference on Words*, 2009. to appear.

[Rot92] Peter Roth. Every binary pattern of length six is avoidable on the two-letter alphabet. *Acta Informatica*, 29(1):95–107, 1992.

[RSW05] Narad Rampersad, Jeffrey Shallit, and Ming-Wei Wang. Avoiding large squares in infinite binary words. *Theoretical Computer Science*, 339(1):19–34, 2005.

[Sch86] Ursula Schmidt. *Motifs inévitables dans les mots.* Rapport LITP 86–63, Paris VI, 1986.

[Sch89] Ursula Schmidt. Avoidable patterns on two letters. *Theoretical Computer Science*, 63(1):1–17, 1989.

[SG04] Arseny M. Shur and Yulia V. Gamzova. Partial words and the interaction property of periods. *Izvestiya RAN*, 68(2):191–214, 2004.

[SK01] Arseny M. Shur and Yulia V. Konovalova. On the periods of partial words. In *MFCS '01: Proceedings of the 26th International Symposium on Mathematical Foundations of Computer Science*, volume 2136, pages 657–665, London, UK, 2001. Springer-Verlag.

[Smy03] William Fennell Smyth. *Computing Patterns in Strings.* Pearson Addison-Wesley, 2003.

[Sto88] James A. Storer. *Data compression: methods and theory.* Computer Science Press, Inc., New York, NY, USA, 1988.

[Thu06] Axel Thue. Über unendliche Zeichenreihen. *Norske Vid. Selsk. Skr. I, Mat. Nat. Kl. Christiana*, 7:1–22, 1906. (Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, Norway (1977), pp. 139–158).

[Thu10] Axel Thue. Die lösung eines Spezialfalles eines generellen logischen Problems. *Norske Videnskabers Selskabs Skrifter, I Mathematisch-Naturwissenschaftliche Klasse Christiana*, 8, 1910. (Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, Norway (1977), pp. 273–310).

[Thu12] Axel Thue. Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen. *Norske Videnskabers Selskabs Skrifter, I Mathematisch-Naturwissenschaftliche Klasse Christiana*, 1:1–67, 1912. (Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, Norway (1977), pp. 413–478).

[Thu14] Axel Thue. Probleme über Veränderungen von Zeichenreihen nach gegebenen Regeln. *Norske Videnskabers Selskabs Skrifter, I Mathematisch-Naturwissenschaftliche Klasse Christiana*, 7, 1914. (Reprinted in *Selected Mathematical Papers of Axel Thue*, T. Nagell, editor, Universitetsforlaget, Oslo, Norway (1977), pp. 493–524).

[Wet] Nathan D. Wetzler. Unbordered partial words and the critical factorization theorem. Personal communication.

[Zha] Jeffery Zhang. The language CF is not regular. Personal communication.