NOVEL STRUCTURAL PROPERTIES AND AN IMPROVED BOUND FOR THE NUMBER OF DISTINCT SQUARES IN STRINGS

Adrien Thierry

NOVEL STRUCTURAL PROPERTIES AND IMPROVED BOUND FOR NUMBER OF DISTINCT SQUARES IN STRINGS

By **ADRIEN THIERRY**

A Thesis Submitted to the Department of Mathematics and Statistics and the School of Graduate Studies of McMaster University in Partial Fulfilment of the Requirement for the Degree of Doctor of Philosophy

Ph.D. Thesis McMaster University
Department of Mathematics and Statistics Hamilton, Ontario, Canada

TITLE:	Novel Structural Properties and Improved Bound for Number of Distinct Squares in Strings
AUTHOR:	Adrien Thierry M.Sc. Université Paris VI, Pierre et Marie Curie x Ecole Polytechnique B.Sc. Université Paris VI, Pierre et Marie Curie
SUPERVISORS:	Dr. Antoine Deza, Dr. Frantisek Franek
NUMBER OF PAGES:	72, X

Abstract

Combinatorics on words explore words – often called strings in the computer science community, or monoids in mathematics – and their structural properties. One of the most studied question deals with repetitions which are a form of redundancy. The thesis focuses on estimating the maximum number of distinct squares in a string of length n. Our approach is to study the combinatorial properties of these overlapping structures, nested systems, and obtain insights into the intricate patterns that squares create. Determining the maximum number of repetitions in a string is of interest in different fields such as biology and computer science. For example, the question arrises when one tries to bound the number of repetitions in a gene or in a computer file to be data compressed. Specific strings containing many repetitions are often of interest for additional combinatorial properties. After a brief review of earlier results and an introduction to the question of bounding the maximum number of distinct squares, we present the combinatorial insights and techniques used to obtain the main result of the thesis: a strengthening of the universal upper bound obtained by Fraenkel and Simpson in 1998.

Acknowledgements

I would like to express my gratitude to my supervisors, Dr. Antoine Deza and Dr. Frantisek Franek, for their invaluable guidance, generous support and continuous encouragement to my research studies and my life.

I appreciated the help and moral support from all my colleagues including the members of the department of Mathematics and Statistics and the members of ADVOL, the Advanced Optimization Laboratory.

Furthermore, I am grateful for the financial aid provided by McMaster University.

To my family and friends.

Table of Contents

List of Figures

Chapter 1

Introduction

"For the development of logical sciences it will be important, without consideration for possible applications, to find large domains for speculation about difficult problems. In this paper, we present some investigations in the theory of sequences of symbols, a theory that has some connections with number theory."

Axel Thue

"Uber die gegenseitige Lage gleicher Teile gewisser Zeichenreihen" [24] (1912), translated by Berstel [4] in "Axel Thue's papers on repetitions in words: a translation".

1.1 Preliminaries

Axel Thue can be regarded as one of the pioneers of combinatorics on words. In his 1906 paper "Über unendliche Zeichenreihen" [23] he constructed an infinite ternary word without squares – tandem repeats of two consecutive occurrences of a word – and an infinite binary word without cubes – tandem repeats of three consecutive occurrences of a word.

In the mid-twentieth century, combinatorics on words found their applications with the development of computer science and the rise of genome analysis. Computer files and DNA chromosomes can be viewed as strings constructed over finite alphabets; and the set of words with the operation of concatenation can be mathematically abstracted as a free monoid. Repetitions act as redundancy, and redundancy is an efficient way to secure the storage of important information. These applications motivate the efforts to understand the underpinning of repetitions.

A hundred years after Thue's pioneering work, squares in words are still studied. The thesis focuses on investigating the bound for the maximum number of distinct squares in a word when the types of the squares are counted rather than their occurrences. Our work aim at contributing to the understanding of the combinatorial structure of strings. We recall earlier results relevant to our strengthening of the bound for the maximum number of distinct squares, and highlight additional insights obtained as a consequence of our structural analysis.

1.2 Strings and repetitions in strings

A *word* – or equivalently a *string* – over a finite alphabet $\mathcal{A}$ is a sequence of *letters* drawn from the alphabet $\mathcal{A}$. While words can be finite or infinite, the thesis considers only finite words. The repetitions we investigate are so-called *tandem repetitions*, i.e. repeating occurrences of the same word when an occurrence starts right after the previous occurrence ends.

Significant research had been done concerning the maximum number of repetitions when the occurrences are being counted, we refer to [21] and references therein for more details.

The problem of counting types of squares rather than occurrences and the motivation for this approach was introduced by Fraenkel and Simpson in their seminal article "*How Many Squares Can a String Contain* ?" [11]. This problem became known as the problem of the *maximum number of distinct squares*. Fraenkel and Simpson provided a universal upper bound of $2n$ for a string of length n obtained via an insightful analysis of the combinatorial structure of squares starting at the same point. We focus on *primitively rooted squares*, i.e. squares of primitive strings. A string is *primitive* if it cannot be expressed as a multiple concatenation of some string. Fraenkel and Simpson [11] provided an upper bound for the number of distinct primitively rooted squares.

We use the term *square maximal* string for a string that attains the maximum number of distinct squares among all strings having the same length. For instance, the string 00000000000 has 5 distinct squares: 00, 0000, 000000, 00000000, and 0000000000; and thus is not square maximal as the string 01010010011 has 7 distinct squares: 00, 11, 0101, 1010, 100100, 010010, and 001001. In addition, 01010010011 is square maximal as no string of length 11 contains 8 or more distinct squares.

The combinatorics of strings can be counterintuitive. For example, it was commonly believed that, among all strings of a given length, the maximum number of distinct squares is achieved by a binary string. A counterexample to this intuition was exhibited by Deza, Franek, and Jiang [8] who showed that, for strings of length 33, the maximum number of distinct primitively rooted squares is at most 23 for binary strings, while there are ternary strings with 24 distinct primitively rooted squares.

1.3 A brief review of previous results

Fraenkel and Simpson [11] introduced in 1998 the problem of counting the number of distinct squares in a string of length n, and provided a universal upper bound of $2n$. Their proof is based on two insightful observations: (i) Crochemore-Rytter's Three Squares Lemma can be generalized, and (ii) counting the right-most occurrences yields combinatorial properties of squares that start at the same position. The original version of the Three Squares Lemma [6] appeared in 1995.

Lemma 1.1 (Three Squares Lemma – original formulation)**.** *Let* u_1^2, u_2^2, u_3^2

be three prefixes of a string x such that u_1, u_2, u_3 are primitive and $|u_3| < |u_2| < |u_1|$. Then $|u_2| + |u_3| < |u_1|$.

Fraenkel and Simpson observed that the proof actually only requires the primitiveness of u_3. Moreover, they noted that $\leq$ is required for $|u_2| + |u_3| \leq |u_1|$, and gave a counter-example for the sharp $<$. They provide a reformulation of the Three Squares Lemma.

Lemma 1.2 (Three Squares Lemma – reformulation)**.** *Let u_1^2, u_2^2, u_3^2 be three prefixes of a string x such that u_3 is primitive and $|u_3| < |u_2| < |u_1|$. Then $|u_2| + |u_3| \leq |u_1|$.*

The reformulation is of particular significance when considering exclusively the right-most occurrences of all squares, and yields the key insight that at most 2 squares can have their last occurrence starting at the same position.

Crochemore and Rytter [6] noted in 1995 that the efficiency of the string-matching algorithm proposed by Galil and Seiferas [13] relies on the periods of the prefixes of the pattern that one wants to find. Recall that a string-matching algorithm finds the occurrence of a pattern p in a string w. Studying these periods, Crochemore and Rytter noted that if three primitive squares start at the same position, the length of the longest one is at least the sum of the lengths of the two shorter ones, thus leading to Lemma 1.1.

Ilie [14] provided in 2005 a short proof of the main result in [11] without using the Three Squares Lemma 1.2. Instead, he used the fact that for three squares starting at the same position, u_3 is completely embedded in u_1 by

Periodicity Lemma 1.4 and Synchronization Principle 1.6.

In 2007, Ilie [15] considered a refinement of the counting of the rightmost occurrences of squares to obtain an asymptotic bound of $2n - \Theta(\log n)$. Illie commented: "*The improvement obtained is not very big. Essentially, we proved that the number of squares in w is bounded away from $2n$. Still, it is the only non-trivial improvement we know of. Moreover, Fraenkel and Simpson considered even improving the bound by a constant amount to be important.*"

Lam [18] proposed in 2011 a different approach based on bounding the number of double squares – i.e. two squares starting at the same position – improving the universal bound to $1.99n$. Lam's approach is discussed in Chapter 3.

Deza, Franek, and Jiang [7] considered in 2011 an approach inspired by a formulation used for polytopes and optimization. In addition to the length n, the number d of distinct letters of the string is considered as a parameter. They showed structural properties and computational applications, and presented their results in a so-called $(d, n-d)$ table. See Fig. 1.1 which succinctly presents the values of $\sigma_d(n)$ defined as the largest number of distinct squares in a string over all strings of length n with d distinct letters. This approach allows the determination of previously intractable instances, see [8].

		$n-d$																																	
		2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35
	2	2	2	3	3	4	5	6	7	7	8	9	10	11	12	12	13	13	14	15	16	17	18	19	20	20	21	22	23	23	23	24	25	26	27
	3	2	3	3	4	4	5	6	7	8	8	9	10	11	12	13	13	14	14	15	16	17	18	19	20	21	21	22	23	24	24	25	26	26	27
	4	2	3	4	4	5	5	6	7	8	9	9	10	11	12	13	14	14	15	15	16	17	18	19	20	21	22	22	23	24	25	25	26	27	
	5	2	3	4	5	5	6	6	7	8	9	10	10	11	12	13	14	15	15	16	16	17	18	19	20	21	22	23	23	24	25	26			
	6	2	3	4	5	6	6	7	7	8	9	10	11	11	12	13	14	15	16	16	17	17	18	19	20	21	22	23	24						
	7	2	3	4	5	6	7	7	8	8	9	10	11	12	12	13	14	15	16	17	17	18	18	19	20	21	22	23	24	25					
	8	2	3	4	5	6	7	8	8	9	9	10	11	12	13	13	14	15	16	17	18														
	9	2	3	4	5	6	7	8	9	9	10	10	11	12	13	14	14	15	16	17	18	19													
	10	2	3	4	5	6	7	8	9	10	10	11	11	12	13	14	15	15	16	17	18	19	20												
	11	2	3	4	5	6	7	8	9	10	11	11	12	12	13	14	15	16	16	17	18	19	20	21											
	12	2	3	4	5	6	7	8	9	10	11	12	12	13	13	14	15	16	17																
d	13	2	3	4	5	6	7	8	9	10	11	12	13	13	14	14	15	16	17	18															
	14	2	3	4	5	6	7	8	9	10	11	12	13	14	14	15	15	16	17	18	19														
	15	2	3	4	5	6	7	8	9	10	11	12	13	14	15	15	16	16	17	18	19	20													
	16	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	16	17	17	18	19														
	17	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	17	18	18															
	18	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	18	19															
	19	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	19	20														
	20	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	20														
	21	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21														
	22	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22													
	23	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23												
	24	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24											

Figure 1.1: $(d, n-d)$ table for $\sigma_d(n)$ values

In 2015, Deza, Franek and Thierry [9] strengthened the upper bound for the number of distinct square to $\lfloor 11n/6 \rfloor$. The title of the paper is a tribute to the pioneer work of Fraenkel and Simpson. The thesis is essentially documented in this publication. The combinatorial insights used to strengthen the upper bound include the notion of *inversion factors* which are discussed in Chapter 2. Inversion factors yield a generalization of the Three Squares Lemma by Bai, Deza, Franek, [1].

Lemma 1.3 (Three Squares Lemma – Generalization)**.** *Let v_2^2, v_3^2 be proper prefixes of v_1^2, then $|v_2|+|v_3| \leq |v_1|$ unless $v_3 = u_1^t, v_2 = u_1^p u_2$ and $v_1 = u_1^p u_2 u_1^q$ for some primitive string u_1, and its proper prefix u_2 and the integers $t > q$, $p \geq q \geq 1$.*

1.4 Definitions

An *alphabet* A is a finite set, and its elements are called *letters*. A sequence of length n of elements of A is a *word* w of length $n = |w|$, which can also be presented as an array $w[1\mathinner{..}n]$. We mostly use the term *string* instead of word. The multiplicative operation for strings is the same as for sequences: *concatenation*, i.e. listing one sequence after another; the notation xy indicates x concatenated with y; it is non-commutative. For positive integer n, the n^{th} *power* of a string w, denoted as w^n, is the concatenation of n copies of w. The power is *trivial* if $n = 1$, a *square* if $n = 2$, and a *cube* if $n = 3$. For a power w^n, w is referred to as a *root*. A string x is *primitive* if x cannot be expressed as a non-trivial power of another string. A *primitively rooted square* is a square whose root is primitive. For a string w, the shortest primitive x such that $x^n = w$ for some n is the *primitive root* of w. The empty string is denoted ε.

Given strings x_1, x_2, x_3, and their concatenation $x = x_1x_2x_3$, x_1 is a *prefix* of x, x_2 is a *factor* of x, and x_3 is a *suffix* of x. For a prefix x_1 of x, if $x_1 \neq x$, then x_1 is referred to as *proper prefix*, similarly for suffixes; if $x_1 \neq \varepsilon$, then it is referred to as a *non-trivial prefix*, and as a *trivial prefix* otherwise; similarly for suffixes. Note that the factor x_2 of $x = x_1x_2x_3$ starts at position $|x_1| + 1$ and ends at position $|x_1x_2|$.

A string $w = w[1\mathinner{..}n]$ has *period* $p \leq \frac{n}{2}$ if $w[i] = w[i{+}p]$ for $i = 1, \ldots, n{-}p$. Given strings x and y, not necessarily over the same alphabet, the *longest* *common prefix* of x and y is denoted as $\mathrm{lcp}(x, y)$, and is defined as $x[i] = y[i]$ for $= 1, \ldots, \mathrm{lcp}(x, y)$ while $x[\mathrm{lcp}(x, y) + 1] \neq y[\mathrm{lcp}(x, y)]$. Similarly, the *longest common suffix* of x and y is denoted as $\mathrm{lcs}(x, y)$ and $x[|x| - i] = y[|y| - i]$ for $i = 1, \ldots, \mathrm{lcs}(x, y)$ while $x[|x| - \mathrm{lcs}(x, y) - 1] \neq y[|y| - \mathrm{lcs}(x, y) - 1]$. Let $w = xy$, x be a proper prefix and y a proper suffix of w, the string $\widetilde{w} = yx$ is called a *conjugate*, or equivalently a *rotation*, of w.

Right and left cyclic shifts are defined recursively. For a substring $x[i\,..\,j]$ of $x[1\,..\,n]$, the substring $x[i+1\,..\,j+1]$ is a *right cyclic shift by one position* of $x[i\,..\,j]$ if $x[i] = x[j+1]$. The substring $x[i+k\,..\,j+k]$ is a *right cyclic shift by k positions* of $x[i\,..\,j]$ if $x[i+k-1\,..\,j+k-1]$ is a right cyclic shift of $x[i\,..\,j]$ by $k-1$ positions and $x[i+k\,..\,j+k]$ is a right cyclic shift by one position of $[x+k-1\,..\,j+k-1]$. Similarly, the substring $x[i-1\,..\,j-1]$ is a *left cyclic shift by one position* of $x[i\,..\,j]$ if $x[i-1] = x[j]$. The substring $x[i-k\,..\,j-k]$ is a *left cyclic shift by k positions* of $x[i\,..\,j]$ if $x[i-k+1\,..\,j-k+1]$ is a left cyclic shift of $x[i\,..\,j]$ by $k-1$ positions and $x[i+k\,..\,j+k]$ is a left cyclic shift by one position of $[x-k+1\,..\,j-k+1]$.

1.5 Basic tools

We recall Fine and Wilf's Periodicity Lemma 1.4, see for instance [21, 19], and Longest Common Border Lemma [9], and two corollaries of Periodicity Lemma 1.4: Common Factor Lemma 1.5 and Synchronization Principle 1.6. Periodicity Lemma 1.4 of Fine and Wilf [10] addressed in 1965 the following question: How long does a periodic sequence x of both periods m and n have to be to ensure that x also has a period $\gcd(m, n)$? While Fine and Wilf addressed both the discrete and the continuous versions of the problem, we only consider the discrete one. A proof can be found in [19].

Lemma 1.4 (Periodicity Lemma)**.** *Let w be a string having periods x and y. If $|w| \geq x + y - \gcd(x, y)$, then w has period $\gcd(x, y)$.*

Common Factor Lemma 1.5 provides a minimal length shared by two different repetitions that forces those two repetitions to have a common root. In particular, it implies that, if x and y are primitive, they are conjugate of each other.

Lemma 1.5 (Common Factor Lemma)**.** *Given strings x and y, integers $m \geq 2$ and $n \geq 2$, if x^m and y^n have a common factor of length $|x| + |y|$, then the primitive roots of x and y are conjugates of each other.*

Proof. Note that x^m has a period $|x|$ and y^n a period $|y|$, and Periodicity Lemma 1.4 applies. □

Synchronization Principle 1.6 prevents certain factors to occur at certain positions.

Lemma 1.6 (Synchronization Principle)**.** *If w is primitive, and w_1 is a proper prefix of $w = w_1w_2$, then $w_2w_1 \neq w$, or equivalently, w is not equal to any of its non-trivial rotations.*

Proof. Assume by contradiction that $w = w_1w_2 = w_2w_1$ for non empty strings w_1 and w_2. Without loss of generality, suppose that $|w_1| < |w_2|$. Since w_1 and w_2 are prefixes of w, w_1 is a prefix of w_2 and w_2w_1 is a factor of w_2^2. Since w_2 is a prefix of $w_1w_2, w_2 = {w_1}^n w_1'$ for a prefix w_1' of w_1 and an integer n, and $w_1w_2 = w_1^{n+1}w_1'$ is a factor of w_1^{n+2}. By Periodicity Lemma 1.4, w has a period $\gcd(|w_1|, |w_2|)$ and is not primitive, hence a contradiction. □

Longest Common Border Lemma 1.7 points out that the longest common prefix and the longest common suffix of two consecutive factors govern how many cyclical shifts of this factors can happen to the right for the first one, and to the left for the second one.

Lemma 1.7 (Longest Common Border Lemma)**.** *If x is primitive,* $\text{lcp}(x, \tilde{x}) + \text{lcs}(x, \tilde{x}) \leq |x| - 2$ *for all of its conjugates $\tilde{x}$.*

Proof. If $\text{lcp}(x, \tilde{x}) + \text{lcs}(x, \tilde{x}) \geq |x|$, then $x = \tilde{x}$, which contradicts the primitiveness of x by Synchronization Principle 1.6. Assume by contradiction that $\text{lcp}(x, \tilde{x}) + \text{lcs}(x, \tilde{x}) = |x| - 1$. Let $p = \text{lcp}(x, \tilde{x}), s = \text{lcs}(x, \tilde{x})$. Write $x = pas, \tilde{x} = pbs$ for the different letters a and b – by maximality of the lcp and the lcs. Thus, x has one more letter a and one letter b less than $\tilde{x}$, and the two strings cannot be conjugates of each other. It follows that $\text{lcp}(x, \tilde{x}) + \text{lcs}(x, \tilde{x}) \leq |x| - 2$. □

Longest Common Border Lemma 1.7 is illustrated in Figure. 1.2. One can check that $|\text{lcs}(cbab, abbaabab)| = 3$ and $|\text{lcp}(abbaabab, abc)| = 2$. Thus, u can be cyclically shifted 3 positions to the left, and 2 positions to the right.

$$w = cbab\underbrace{abbaabab}_{u}abc.$$

Figure 1.2: Cyclic shifts of u

Chapter 2

Double squares

Fraenkel and Simpson [11] conjectured that the maximal number of distinct squares in a string of length n is at most n. Their approach underlines the relevance of the structure formed by two squares starting at the same position as a tool to bound the number of distinct squares.

While double squares were not formally introduced in earlier papers, efforts concerning three squares and the structural or combinatorial approach to the problem can be found in Bland and Smyth [5], Kopylova and Smyth [17], Franek, Fuller, Simpson, and Smyth [12], and Simpson [20].

In several papers concerning the so-called New Periodicity Lemma, the relationships of three squares of which two start at the same position are investigated. Fan, Puglisi, Smyth, and Turpin [16] defined and studied the question. Similarly as for Three Squares Lemma 1.2, the methods used in [9] allow for a proof of the New Periodicity Lemma [3].

As mentioned in Section 1.3, Lam [18] considered a different approach. To the best of our knowledge, he was the first to consider the structure imposed by two sets of distinct double squares, each of them being their rightmost occurrence. We call such double squares *FS-double squares*, short for Fraenkel and Simpson double squares. Building on Fraenkel and Simpson's work, Lam's approach proposes to bound the number of distinct squares by bounding the number of double squares. For this purpose, Lam introduced the following structure: for a double square (u, U) there are $p \geq q \geq 1$ and a primitive x_1 and its non-trivial prefix x_2 so that $u = x_1^p x_2$ and $U = x_1^p x_2 x_1^q$. Calling $|x_1|$ the period of the double square and $|x_2|$ its co-period, Lam defined a taxonomy of relationships between two double squares (u, U) and (v, V) based on their periods, co-periods, and the starting or ending of v vis-à-vis u and U. He then consider an induction hypothesis based on the exhaustive lists of all types of relationships of two double squares. This induction yields a bound of $95n/48$ for the maximal number of distinct squares in a string of length n .

2.1 Structural approach to double squares

In Deza, Franek, and Thierry [9] a *double square* corresponds to two squares u and U starting at the same position and satisfying $|u| < |U| < 2|u| < 2|U|$. It imposes a unique structure called the *canonical factorization* consisting of a primitive u_1, a prefix u_2 of u_1, and $e_1 \geq e_2 \geq 1$ so that $u = u_1^{e_1} u_2$

and $U = u_1^{e_1}u_2u_1^{e_2}$, i.e. $UU = u_1^{e_1}u_2u_1^{e_1+e_2}u_2u_1^{e_2}$. We also showed that if u is primitive or if u is a right-most occurrence, then u_2 must be non-empty, i.e. a non-trivial proper prefix of u_1. Note that Bai, Franek and Smyth [2] proposed another proof of that factorization.

The factor $uu = {u_1}^{e_1}u_2{u_1}^{e_1}u_2$ is referred to as the *short square* of a double square $\mathcal{U}$, the factor $UU = u_1^{e_1}u_2u_1^{e_1+e_2}u_2u_1^{e_2}$ as the *long square.* We use the following notational convention: $\mathcal{U} := (u_1, u_2, e_1, e_2)$ indicates that (u_1, u_2, e_1, e_2) is the canonical factorization of $\mathcal{U}$. Moreover, the short square of $\mathcal{U}$ is uu and the long square of $\mathcal{U}$ is UU, while the short square of $\mathcal{V}$ is vv and the long one VV, for $\mathcal{W}$ it is ww and WW, and so forth.

Note that the uniqueness of the factorization applies only to double squares, and not to squares: for $u_1 = aab, u_2 = aa, u_1' = aaba$ and $u_2' = a$ we have $u_1u_2 = u_1'u_2'$.

The combinatorial structure of a double square contains another key structure. One can guess the existence of structural factors by looking at the canonical factorization of a double square $\mathcal{U}$: the two u_2's that appear at the positions $|u_1^{e_1}|$ and $|u_1^{e_1}u_2u_1^{e_1+e_2}|$ seem *unique* in the canonical factorization of $\mathcal{U}$. This property is discussed in Section 2.3.

2.2 Interruption of the period

The properties discussed in this section apply to double squares as defined in [2] even though they were originally discussed and proven for the slightly more restrictive notion of FS-double squares in [9]. A double square $\mathcal{U}$ has highly repetitive structure: ${u_1}^{e_1}u_2{u_1}^{e_2}{u_1}^{e_1}u_2{u_1}^{e_2}$.

$$UU = \underbrace{u_1^{e_1}u_2}_{w_1}\underbrace{u_1^{e_2}u_1^{e_1}u_2}_{w_2}\underbrace{u_1^{e_2}}_{w_3}.$$

Figure 2.1: 3 factors of $u_1^{e_1+e_2+1}$

The two occurrences of u_2's causing the *interrupt of the period* u_1 seem unique in UU, however they occur in UU multiply times as u_2 is a prefix of every occurrence of u_1. Nevertheless, since they interrupt the period, we can identify two unique factors *surrounding* the interrupts. We named these factors *inversion factors* since they have an *inverse* structure $\bar{u}_2u_2u_2\bar{u}_2$, i.e. the first half is an inverse of the second half. The two obvious occurrences of the inversion factor are indicated in Figure 2.2.

$$\mathcal{U} = u_1^{e_1}u_2u_1^{e_2}u_1^{e_1}u_2u_1^{e_2}.$$
$$\mathcal{U} = (u_2\bar{u}_2)^{e_1-1}u_2\underbrace{\bar{u}_2u_2u_2\bar{u}_2}_{\text{Inversion Factor.}}(u_2\bar{u}_2)^{e_2+e_1-2}u_2\underbrace{\bar{u}_2u_2u_2\bar{u}_2}_{\text{Inversion Factor.}}(u_2\bar{u}_2)^{e_2-1}.$$

Figure 2.2: Inversion factors.

The inversion factors play a key role in our approach, for if they are shown to be unique – and we will show that they are essentially unique, then any third square VV starting *close* to the beginning if $\mathcal{U}$ such that the first V contains the first inversion factor, then the second occurrence of V must contain the second occurrence of the inversion factor, and hence $|V| = |U|$ as the distance between the two inversion factors is exactly $|U|$. Simply put, any such third square VV must be either (i) small, and not containing the whole inversion factor, or (ii) of the size $|U|$, or (iii) very big and containing the first inversion factor and at least a part of the second inversion factor in the first occurrence of V. Therefore, this property significantly restricts the number of possible types and positions for VV, and by extension, dramatically restricts the number of possible types of double squares (v, V) with respect to $\mathcal{U}$.

2.3 Canonical inversion factors

After introducing the structure of double squares and the two occurrences of the inversion factor $\bar{u}_2u_2u_2\bar{u}_2$ illustrated in Figure 2.2, we precisely define this notion and describe its properties. We first present the original version of the inversion factor used in [9]. The *core of the interrupt*, a more compact version of the inversion factor introduced in Thierry [22], is presented in Section 2.4.

For a primitive string $x = x_2\bar{x}_2$ where x_2 is a proper prefix, and hence $\bar{x}_2$ is a proper non-trivial suffix, $x_2\bar{x}_2 \neq \bar{x}_2x_2$; which ensures a certain rarity of that factor. Since an occurrence of the inversion factor could be cyclically shifted, we need a structural definition that would consider such shift. For that purpose, we define the structure of the inversion factor, and investigate the places where it could occur. Finally, we prove that if such factor occurs within $\mathcal{U}$, it can only do so at, or close to, two specified positions:. $|{u_1}^{e_1-1}u_2|$

and $|{u_1}^{e_1}u_2{u_1}^{e_1+e_2-1}u_2|$. While Definition 2.1 and Lemma 2.3 were formulated in [9] for FS-double squares, we present them for double squares.

Definition 2.1 (Inversion Factor)**.** *Given a double square $\mathcal{U} := (u_1, u_2, e_1, e_2)$, a factor of UU of length $2|u_1|$ starting at position i is called an* inversion factor *of $\mathcal{U}$ if*

$$U^2[i+j] = U^2[i+j+|u_1|+|u_2|] \textit{ for } 0 \le j < |u_1|, \textit{ and}$$

$$U^2[i+j] = U^2[i+j+|u_1|] \textit{ for } |u_1| \le j \le |u_1|+|u_2|.$$

By Definition 2.1, the format of an inversion factor of $\mathcal{U}$ is $\bar{v}_2v_2v_2\bar{v}_2$ where $|v_2| = |u_2|, |\bar{v}_2| = |\bar{u}_2|$, and if $v_2 = u_2, \bar{v}_2 = \bar{u}_2$. We refer to such inversion factor as a *canonical* one, and denote as $N_1(\mathcal{U})$ and $N_2(\mathcal{U})$ the positions they start at $N_1(\mathcal{U}) = (e_1-1)|u_1|+|u_2|+1$ and $N_2(\mathcal{U}) = (2e_1+e_2-1)|u_1|+2|u_2|+1$.

The cyclic shifts of the inversion factor $\bar{u}_2u_2u_2\bar{u}_2$, have the same structural properties as the inversion factor, and, as such, are de facto inversion factors. We show that cyclic shifts are governed by the values of $\mathrm{lcp}(u_2\bar{u}_2, \bar{u}_2u_2)$ and $\mathrm{lcs}(u_2\bar{u}_2, \bar{u}_2u_2)$, and that we can find inversion factors starting at every position in $N_1(\mathcal{U}) - \mathrm{lcs}(u_1, \widetilde{u_1}), \ldots, N_1(\mathcal{U}) + \mathrm{lcp}(u_1, \widetilde{u_1})$ and every position in $N_2(\mathcal{U}) - \mathrm{lcs}(u_1, \widetilde{u_1}), \ldots, N_2(\mathcal{U}) + \mathrm{lcp}(u_1, \widetilde{u_1})$, provided that those positions exist as formalized in Definition 2.2 and illustrated in Figure 2.3 where the inversion factors are highlighted.

Definition 2.2. *Given a double square $\mathcal{U} = (u_1, u_2, e_1, e_2)$, the bounds for the ranges of the occurrences of the inversion factor are*

$$L_1(\mathcal{U}) = max\{1, N_1(\mathcal{U}) - \text{lcs}(u_1, \widetilde{u_1})\},$$

$$R_1(\mathcal{U}) = N_1(\mathcal{U}) + \text{lcp}(u_1, \widetilde{u_1}),$$

$$L_2(\mathcal{U}) = N_2(\mathcal{U}) - \text{lcs}(u_1, \widetilde{u_1}),\ \textit{and}$$

$$R_2(\mathcal{U}) = min\{|UU|, N_2(\mathcal{U}) + \text{lcp}(u_1, \widetilde{u_1})\}.$$

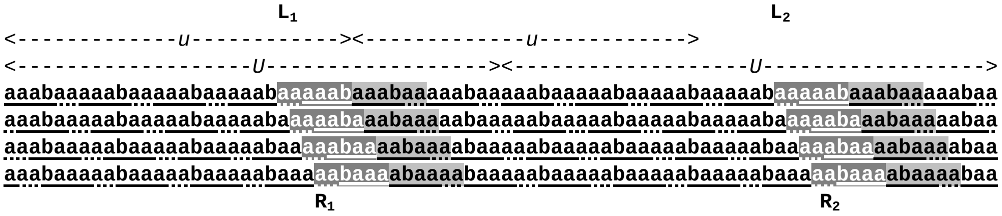

L₁
L₂
<-------------u------------><-------------u------------>
<-------------------U------------------><-------------------U------------------>
aaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaa
aaabaaaaabaaaaabaaaaabaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaabaaabaaaaabaa
aaabaaaaabaaaaabaaaaabaaaaabaaaabaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaabaaaabaa
aaabaaaaabaaaaabaaaaabaaaaabaaaabaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaabaaaabaa
R₁
R₂

Figure 2.3: The inversion factors and their right cyclical shifts starting from L_1 to R_1 and from L_2 to R_2.

When clear from the context, we omit the $\mathcal{U}$ designation from $N_1(\mathcal{U})$, $N_2(\mathcal{U})$, $L_1(\mathcal{U})$, $L_2(\mathcal{U})$, $R_1(\mathcal{U})$, and $R_2(\mathcal{U})$. Lemma 2.3 states that there are no further occurrences of inversion factors except the canonical ones and their cyclic shifts. The positions at which an inversion factor can occur are thus significantly restricted, and this is a key ingredient for strengthening Fraenkel and Simpson's bound.

Lemma 2.3 (Inversion Factor Lemma)**.** *An inversion factor of a double square* $\mathcal{U}$ *within the string* UU *starts at a position* i *if and only if* $i \in [L_1(\mathcal{U}), R_1(\mathcal{U})] \cup [L_2(\mathcal{U}), R_2(\mathcal{U})]$.

The rather tedious case-by-case proof of Lemma 2.3 presented in [9] applies even though the result here is stated for double squares rather than FS-double squares. Instead, we present a weaker version which follows directly from Synchronization Principle 1.6 and Common factor Lemma 1.5.

Lemma 2.4 (Weak Inversion Factor Lemma)**.** *The factor* $\bar{u}_2 u_2 u_2 \bar{u}_2$ *in a double square* $\mathcal{U}$ *cannot be a subfactor of* ${u_1}^{e_1} u_2$*, nor a subfactor of* ${u_1}^{e_1+e_2} u_2$*, nor a subfactor of* ${u_1}^{e_2}$*.*

Proof. Note that ${u_1}^{e_1} u_2, {u_1}^{e_1+e_2} u_2$ and ${u_1}^{e_2}$ are all factors of ${u_1}^{e_1+e_2+1}$. If $\bar{u}_2 u_2 u_2 \bar{u}_2$ was a factor of ${u_1}^{e_1+e_2+1}$, there would exists a conjugate$\tilde{u_1}$ of u_1 for which $\tilde{u_1} = \bar{u_2} u_2 = u_2 \bar{u_2}$, contradicting lemma 1.6.

□

We use the following notation: for a square uu, $u_{[1]}$ indicate the first occurrence of u in uu, i.e. $\boldsymbol{u}u$, while $u_{[2]}$ indicate the second occurrence of u in uu, i.e. $u\boldsymbol{u}$. Recall that we are interested in counting distinct squares in a string, and that our strategy is to limit the number of FS-double squares. Given a double square $\mathcal{U}$, if another square vv were to start within the first occurrence of u and $v_{[1]}$ were to contain the first inversion factor, then either $v_{[1]}$ would contain the second inversion factor, forcing v to be long compared to u, or $v_{[2]}$ would have to contain it at the same position, forcing v to be of the same length as U. Hence the inversion factors can be seen as two notches that force some sort of alignment for double squares under certain conditions. Chapter 3 focuses on the different ways a double square can be articulated around those two notches.

2.4 Core of the interrupt

The core of the interrupt is a subfactor of the inversion factor introduced in Thierry [22] and identified as the element that makes an inversion factor essentially unique. Let $n_c(w)$ denote the *number of occurrences* of a letter c in a string w. Note that a string w and all of its conjugates have the same number of occurrences for all letters. Thus, if the numbers of occurrence for a given letter differ for two strings, they cannot be conjugates.

Definition 2.5 (Core of the Interrupt)**.** *Let $W = {x_1}^{e_1}x_2{x_1}^{e_2}$ for a primitive string x_1 and x_2 a proper prefix of $x_1 = x_2\bar{x}_2$. Let $\tilde{p}$ be the prefix of length $|\mathrm{lcp}(x_2\bar{x}_2, \bar{x}_2x_2)| + 1$ of $x_2\bar{x}_2$ and $\tilde{s}$ the suffix of length $|\mathrm{lcs}(x_2\bar{x}_2, \bar{x}_2x_2)| + 1$ of $\bar{x}_2x_2$. The factor $\tilde{s}\tilde{p}$ starting at position $|x_1^{e_1}| + |x_2| - |\mathrm{lcs}(x_2\bar{x}_2, \bar{x}_2x_2)| - 1$ is the* core of the interrupt *of W.*

Let $x_1 = aaabaaaaaabaaaa$, $x_2 = aaabaaaaaabaaa$, and $\bar{x}_2 = a$, then $x_1x_2x_1 = aaabaaaaaabaaaa\mathbf{aaabaaaaaabaaa}aaabaaaaaabaaaa$ is a prefix of $x_1x_2x_1^2$. Thus, $\mathrm{lcp}(x_2\bar{x}_2, \bar{x}_2x_2) = aaa$, $\tilde{p} = aaab, \mathrm{lcs}(x_2\bar{x}_2, \bar{x}_2x_2) = aaa$, and $\tilde{s} = baaa$. The core of the interrupt $\tilde{s}\tilde{p} = baaaaaab$ underbraced below in

$$x_1x_2x_1 = aaabaaaaaabaaaaaaabaaaaaa\,\underbrace{baaaaaab}_{\tilde{s}\tilde{p}}\,aaaaaabaaaa.$$

Figure 2.4: Example of the Core of the Interrupt

Theorem 2.6 shows the scarcity of the core of the interrupt.

Theorem 2.6. *Let x_1 be a primitive string, x_2 a proper prefix of x_1 and $W = {x_1}^{e_1}x_2{x_1}^{e_2}$ with $e_1 \geq 1, e_2 \geq 1, e_1 + e_2 \geq 3$. Let w_1 be the factor of* *length $|x_1|$ of W ending with the core of the interrupt of W, and let w_2 be the factor of length $|x_1|$ starting with the core of the interrupt of W. The strings w_1 and w_2 are not in the conjugacy class of x_1.*

Proof. Define $p = \operatorname{lcp}(x_2\bar{x}_2, \bar{x}_2x_2)$ and $s = \operatorname{lcs}(x_2\bar{x}_2, \bar{x}_2x_2)$ – note that p and s can be empty. By Longest Common Border Lemma 1.7, $|\operatorname{lcs}(x_2\bar{x}_2, \bar{x}_2x_2)| + |\operatorname{lcp}(x_2\bar{x}_2, \bar{x}_2x_2)| \leq |x_2\bar{x}_2| - 2$ when $x_2\bar{x}_2$ is primitive. Note that if $|\operatorname{lcs} x_2\bar{x}_2| + |\operatorname{lcp} x_2\bar{x}_2| = |x_1| - 2$, w_1 and w_2 are the same factor. Write $x_1 = pr_prr_ss$ and $\widetilde{x_1} = pr'_pr'r'_ss$ for the letters r_p, r'_p, r_s, r'_s, $r_p \neq r'_p, r_s \neq r'_s$ – by maximality of the longest common prefix and suffix - and the possibly empty or homographic strings r and r'. By construction, $w_1 = r'r'_ssprp$ and $w_2 = r'_ssprpr$. Note that $n_{r_p}(w_1) = n_{r_p}(\tilde{x_1}) + 1$ and that $n_{r'_p}(\tilde{x_1}) = n_{r'_p}(w_1) + 1$ and w_1 is not a conjugate of $\tilde{x_1}$, nor of x_1. Similarly for w_2, $n_{r'_s}(w_2) = n_{r'_s}(x_1) + 1$ and $n_{r_s}(x_1) = n_{r_s}(w_2) + 1$ and w_2 is not a conjugate of x_1. □

An illustration of the *tightness* of Theorem 2.6 consider $x_1 = aaabaaaaaabaaaa$, $x_2 = aaabaaaaaabaaa$ and $\bar{x}_2 = a$. We have $|x_1| = 15$, and

$$x_1x_2x_1 = aaabaaaaaabaaaaaaa\overbrace{baaaaaa\,\rlap{$\underbrace{\phantom{\mathbf{baaaaaab}}aaaaaab}_{w_2}$}\mathbf{baaaaaab}}^{w_1}aaaaaab\,aaaa$$

The core of the interrupt is presented in bold. The two factors w_1 and $w_2 = w_1 = baaaaaabaaaaaab$ (note that w_1 need not be equal to w_2), are not factors of x_1^2. Yet, the factor $aaaaaabaaaaaabaaaaaa$ of length $|x_1| + |\operatorname{lcs}(x_1, \widetilde{x_1})| + |\operatorname{lcp}(x_1, \widetilde{x_1})|$ and which contains the core of the interrupt is a factor of x_1^2. The same holds for the factors of length $|x_1| - 1$ that starts and ends with the core of the interrupt, *aaaaaabaaaaaab* and *baaaaaabaaaaaa*: they are both factors of x_1^2. Thus, Theorem 2.6 can be considered tight.

Chapter 3

FS-double square mates

The fact that the different types of configurations of two FS-double squares can be determined and their number is small is a key ingredient in our approach to bound the number of FS-double squares in a string. We call these configurations the *mate relation*, and show, using the inversion factors discussed in Chapter 2, that there are exactly five mate relations called α, β, γ, δ, and ε.

As in Chapter 2, $u_{[1]}$ denotes the first occurrence of u in $\boldsymbol{u}u$, while $u_{[2]}$ denotes the second occurrence of u in $u\boldsymbol{u}$. The double square $\mathcal{U}$ consists of the short square uu and the long square UU. The fact that (u_1, u_2, e_1, e_2) is the canonical factorization of the double square $\mathcal{U}$ is denoted by $\mathcal{U} := (u_1, u_2, e_1, e_2)$. $\mathcal{U}\prec \mathcal{V}$ indicates that the starting point of $\mathcal{U}$ precedes the starting point of $\mathcal{V}$, while $\mathcal{U} \mathrel{-\!\!\!\prec\!\!\!\prec} \mathcal{V}$ indicates that vv starts strictly before $R_1(\mathcal{U})$. For a double square $\mathcal{U}$, $u_1 = u_2\bar{u}_2$ and we adopt the notation $,\overline{u}_1 =$

$\bar{u}_2 u_2$.

Let us remark that Lam [18] put forth a relation of FS-double squares $\mathcal{U}$ and $\mathcal{V}$ that relied solely on the mutual configuration of uu and vv and the size of v. However, showing the exhaustiveness of Lam's taxonomy is challenging. In addition, for some of the relations hypothesized by Lam, no example of two FS-double squares satisfying such relations could be provided. In contrast, our mate relation taxonomy is shown to be exhaustive and an example is provided for each configuration.

3.1 α-mates

A double square $\mathcal{V}$ is a *right cyclic shift* of a double square $\mathcal{U}$ if vv is a right cyclic shift of uu and VV is the same right cyclic shift of UU. Clearly, a cyclic shift of a double square is again a double square. On the other hand, a cyclic shift of a FS-double square is a double square, but not necessarily a FS-double square, as the shifted short square, respectively long square, might not be the right-most occurrence. The first type of the mate relation between two FS-double squares formalizes this fact.

Definition 3.1. *Consider two FS-double squares $\mathcal{U} \prec \mathcal{V}$. Then $\mathcal{V}$ is an* α-mate *of $\mathcal{U}$ if it is a right cyclic shift of $\mathcal{U}$. The relation is denoted by $\mathcal{U} \underset{\alpha}{\rightarrow} \mathcal{V}$.*

The α-mate relation is illustrated in Figure 3.1 where the canonical inversion factors are highlighted, u_2, v_2 are underlined, and $\bar{u}_2, \bar{v}_2$ are under-dotted.

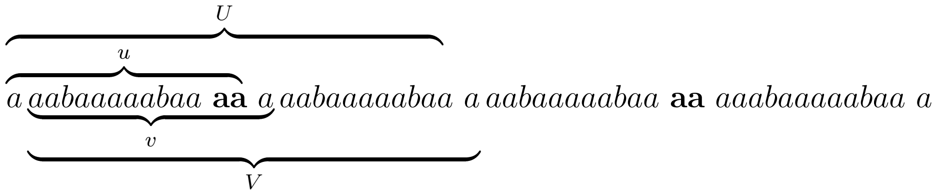

U
u
a aabaaaaabaa **aa** a aabaaaaabaa a aabaaaaabaa **aa** aaabaaaaabaa a
v
V

Figure 3.1: An example of an α-mate

The square UU can be cyclically shifted to the right by less than $|U|$ positions as otherwise UU would have another occurrence contradicting the definition of FS-double squares. Similarly, uu can potentially be cyclically shifted to the right by less than $|u|$ positions. Thus, $\mathcal{U}$ could be cyclically shifted to the right by less than $|u|$ positions to be still a FS-double square. The next observations shows that the shifts are even more limited.

Observation 3.2. *Let* $\mathcal{U} := (u_1, u_2, e_1, e_2)$ *be a FS-double square. The number of* α*-mates of* $\mathcal{U}$ *is bounded by* $|u_1| - 2$.

Proof. By definition, $UU = (u_2\bar{u}_2)^{e_1}u_2(u_2\bar{u}_2)^{e_2}(u_2\bar{u}_2)^{e_1}u_2(u_2\bar{u}_2)^{e_2}$; that is, $UU = (u_2\bar{u}_2)^{e_1}u_2(u_2\bar{u}_2)^{e_1}u_2(\bar{u}_2u_2)^{e_2}(u_2\bar{u}_2)^{e_2} = uu(\bar{u}_2u_2)^{e_2}(u_2\bar{u}_2)^{e_2}$. Thus, The square uu is followed by a factor $\bar{u}_2u_2$. Since $u_2\bar{u}_2$ is a prefix of uu, the number of right cyclic shifts of uu in $uu\bar{u}_2u_2$ is bounded by $\mathrm{lcp}(u_2\bar{u}_2, \bar{u}_2u_2)$. Thus, the number of right cyclic shifts of $\mathcal{U}$ in $UUu_2\bar{u}_2$ is bounded by $\mathrm{lcp}(u_2\bar{u}_2, \bar{u}_2u_2)$, and, by Longest Common Border Lemma 1.7, $\mathrm{lcp}(u_2\bar{u}_2, \bar{u}_2u_2) \leq |u_1| - 2$. □

As a corollary of Observation 3.2, an α-mate of $\mathcal{U}$ must start before $R_1(\mathcal{U})$, i.e. $\mathcal{U} \prec\!\!\prec \mathcal{V}$. Moreover, α-mate relation is transitive, i.e. $\mathcal{U} \underset{\alpha}{\to} \mathcal{V}$ and $\mathcal{V} \underset{\alpha}{\to} \mathcal{W}$ imply $\mathcal{U} \underset{\alpha}{\to} \mathcal{W}$.

3.2 β-mates

Another property that gives rise to a relation between FS-double squares is *right cyclic shift with shrinking of vv*. Consider $uu = {u_1}^{e_1}u_2{u_1}^{e_1}u_2 = u_1^t\mathbf{u_1^{e_1-t}u_2u_1^{e_1-t}u_2}(\bar{u}_2u_2)^t$ with $1 \leq t < e_1$. The square $(u_1^{e_1-t}u_2)(u_1^{e_1-t}u_2)$, denoted in bold, appears at the position $|u_1^t| + 1$. This may be coupled with a right cyclic shift of UU by $|{u_1}^t|$ positions. This property motivates Definition 3.3.

Definition 3.3. *Consider two FS-double squares $\mathcal{U} := (u_1, u_2, e_1, e_2)$ and $\mathcal{V}$ so that $\mathcal{U} \prec\!\!\prec \mathcal{V}$. If $v = \widetilde{u}_1^i\widetilde{u}_2$ with $1 \leq i < e_1$ where $\widetilde{u}_1$ is a cyclic shift of u_1 and $\widetilde{u}_2$ the prefix of length $|u_2|$ of $\widetilde{u}_1$, and VV a right cyclic shift of UU by ${u_1}^i$ positions, then $\mathcal{V}$ is said to be a β-*mate *of $\mathcal{U}$. The relation is denoted by $\mathcal{U} \underset{\beta}{\to} \mathcal{V}$.*

The β-mate relation is illustrated in Figure 3.2 where the canonical inversion factors are highlighted, u_2, v_2 are underlined, and $\bar{u}_2, \bar{v}_2$ are under-dotted.

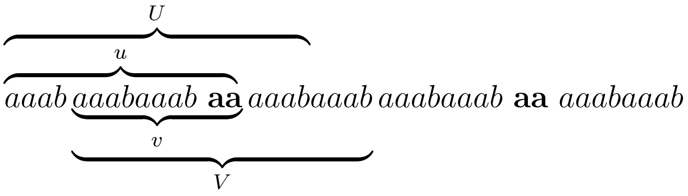


Figure 3.2: An example of a β-mate

As corollary of Definition 3.3,

> $\mathcal{U} \;{}_{\vec{\beta}}\; \mathcal{V} \;{}_{\vec{\alpha}}\; \mathcal{W}$, then $\mathcal{U} \;{}_{\vec{\beta}}\; \mathcal{W}$.
>
> $\mathcal{U} \;{}_{\vec{\beta}}\; \mathcal{V} \;{}_{\vec{\beta}}\; \mathcal{W}$, then $\mathcal{U} \;{}_{\vec{\beta}}\; \mathcal{W}$.
>
> $\mathcal{U} \;{}_{\vec{\alpha}}\; \mathcal{V} \;{}_{\vec{\beta}}\; \mathcal{W}$, then $\mathcal{U} \;{}_{\vec{\beta}}\; \mathcal{W}$.
>
> If $\mathcal{U} := (u_1, u_2, e_1, e_2)$, then β-mate $\mathcal{V} := (\widetilde{u}_1, \widetilde{u}_2, e_1{-}i, e_2{+}i)$ for some integer i, $1 \leq i \leq \lfloor \frac{e_1-e_2}{2} \rfloor$, where $\widetilde{u}_1$ is a cyclic shift of u_1 and $\widetilde{u}_2$ the prefix of length $|u_2|$ of $\widetilde{u}_1$.
>
> If a FS-double square $\mathcal{U} := (u_1, u_2, e_1, e_2)$ has a β-mate, then $e_1 \geq e_2{+}2 \geq 3$.

3.3 γ-mates

We have considered the case where $\mathcal{V}$ is a right cyclic shift of $\mathcal{U}$, i.e. vv and VV are shifts of respectively uu and UU, and the case where VV is a right cyclic shift of UU. Let us consider the case where vv is a right cyclic shift of UU which *blows up* the size of VV.

Definition 3.4. *Consider FS-double squares $\mathcal{U} \;-\!\!\!\ll \mathcal{V}$. If $|v| = |U|$, then $\mathcal{V}$ is said to be a γ-mate of $\mathcal{U}$. The relation is denoted by by $\mathcal{U} \;{}_{\vec{\gamma}}\; \mathcal{V}$.*

The γ-mate relation is illustrated in Figure 3.3 where thee canonical inversion factors are highlighted, u_2, v_2 are underlined, and $\bar{u}_2, \bar{v}_2$ are under-dotted.

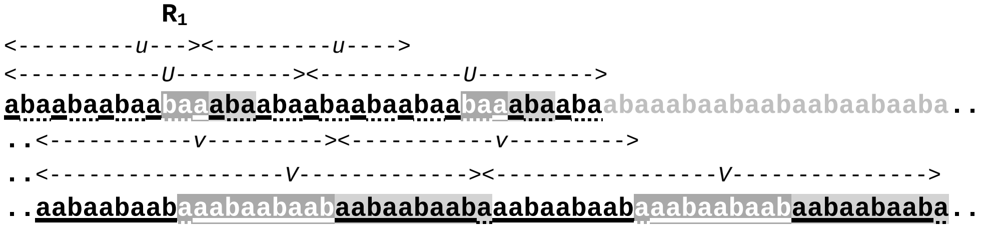

R₁
<---------u---><---------u---->
<-----------U---------><-----------U--------->
abaabaabaabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaaba..
..<-----------v---------><-----------v--------->
..<-----------------V-------------><----------------V-------------->
..aabaabaabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaaba..

Figure 3.3: An example of a γ-mate

The structure of α-mates and β-mates is straightforward and we can determine their canonical factorizations. On the other hand, γ-mates are more challenging. Nevertheless, we can handle γ-mates as follows. Let a FS-double square $\mathcal{V}$ be a γ-mate of a FS-double square $\mathcal{V} := (u_1, u_2, e_1, e_2)$. Then $v = s_2{u_1}^{e_1-t-1}u_2{u_1}^{e_2+t}s_1$ or $v = {u_1}^{e_1-t}u_2{u_1}^{e_2+t}$ for some $e_1 - t \geq 1$ and some s_1, s_2 so that $s_1s_2 = u_1$. We define a *type* of $\mathcal{V}$ as:

$$type(\mathcal{V}) := \begin{cases} (e_1 - t, e_2 + t) & \text{if } v = {u_1}^{e_1-t}u_2{u_1}^{e_2+t}, \\ (e_1 - t, e_2 + t) & \text{if } s_2{u_1}^{e_1-t-1}u_2{u_1}^{e_2+t}s_1 \text{and} |s_1| \leq |u_1| - lcs(u_1, \overline{u}_1), \\ (e_1 - t - 1, e_2 + t + 1) & \text{otherwise.} \end{cases}$$

While characterizing V^2 is challenging, Lemma 3.5 states some of its properties.

Lemma 3.5. *Let a FS-double square $\mathcal{V} := (v_1, v_2, e'_1, e'_2)$ be a γ-mate of a FS-double square $\mathcal{U} := (u_1, u_2, e_1, e_2)$, and let the type of $\mathcal{V}$ be (p, q) where $p, q \geq 2$. Then $e'_1 = e'_2$ and $|v_2| \leq min(p, q)|u_1|$. Moreover, either $|v_2| < |u_1|$ or there is a factor $({u_1}^q u_2)({u_1}^q u_2)$ in V^2.*

Proof. **Case 1.** Assume $v^2 = [{u_1}^p u_2 {u_1}^q][{u_1}^p u_2 {u_1}^q]$.

Case 1.a. Assume $p \geq q$. By Synchronization Principle Lemma 1.6, the leftmost possible beginning of $V_{[2]}$ can be at $|{u_1}^p u_2 {u_1}^{p+q} u_2| + 1$ and so ${u_1}^p u_2$ is a prefix of ${v_1}^{e'_2}$ and v_2 is a factor of ${u_1}^q$. First we prove that $|v_1| > (p-1)|u_1|$. Assume that $|v_1| \leq (p-1)|u_1|$. Then ${u_1}^p$ contains a factor of size $|v_1| + |u_1|$, and the same factor is also contained in ${v_1}^{e'_2}$ as ${u_1}^p u_2$ is a prefix of ${v_1}^{e'_2}$. If $e'_2 \geq 2$, then by the Common Factor Lemma 1.5, $u_1 = v_1$ and so ${u_1}^p u_2$ is a prefix of ${u_1}^{e'_2}$ and thus ${u_1}^p u_2 u_1$ is a prefix of ${u_1}^{e'_2+1}$, which contradicts Synchronization Principle Lemma 1.6. Therefore $e'_2 = 1$ and so $|v_1| \geq p|u_1| + |u_2| > (p-1)|u_1|$, hence a contradiction. Thus, $|v_1| > (p-1)|u_1| \geq q|u_1|$ and since v_2 is a factor in ${u_1}^q$, $e'_1 = e'_2$. If $V_{[2]}$ begins even more to the right, this makes v_2 smaller and ${v_1}^{e'_2}$ bigger, thus the same argument can be applied.

Case 1.b. Assume $p < q$. By Synchronization Principle Lemma 1.6, the leftmost possible beginning of $V_{[2]}$ can be at $|{u_1}^p u_2 {u_1}^{p+q} u_2 {u_1}^{q-p}| + 1$ and so ${u_1}^p u_2 {u_1}^{q-p}$ is a prefix of ${v_1}^{e'_2}$ and v_2 is a factor of ${u_1}^p$. Let $r = max(p, q-p)$. First we prove that $|v_1| > (r-1)|u_1|$. Assume that $|v_1| \leq (r-1)|u_1|$. Then either ${u_1}^p$ or ${u_1}^{q-p}$ contains a factor of size $|v_1| + |u_1|$ and the same factor is also contained in ${v_1}^{e'_2}$ as ${u_1}^p u_2 {u_1}^{q-p}$ is a prefix of ${v_1}^{e'_2}$. If $e'_2 \geq 2$, then by the Common Factor Lemma 1.5, $u_1 = v_1$ and so ${u_1}^p u_2 {u_1}^{q-p}$ is a prefix of ${u_1}^{e'_2}$, which contradicts Synchronization Principle Lemma 1.6. Therefore

$e'_2 = 1$ and so $|v_1| \geq q|u_1| + |u_2| > (r-1)|u_1|$, hence a contradiction. Thus, $|v_1| > (r-1)|u_1| \geq p|u_1|$ and since v_2 is a factor in ${u_1}^p$, $e'_1 = e'_2$. If $V_{[2]}$ begins even more to the right, this makes v_2 smaller and ${v_1}^{e'_2}$ bigger, thus the same argument can be applied.

Case 2. Assume that $v^2 = [s_2{u_1}^{p-1}u_2{u_1}^q s_1][s_2{u_1}^{p-1}u_2{u_1}^q s_2]$ and $|s_1| \leq |u_1| - lcs(u_1,\overline{u}_1)$. Thus, $|s_2| > lcs(u_1,\overline{u}_1)$.

Case 2.a. Assume $p \geq q$. By Synchronization Principle Lemma 1.6, the leftmost possible beginning of $V_{[2]}$ can be at $|s_2{u_1}^{p-1}u_2{u_1}^{p+q}u_2s_1| + 1$. If it started to the left of this point, by Synchronization Principle Lemma 1.6, s_2 would have to be a suffix of u_1u_2 and so s_2 would be a common suffix of u_1 and $\overline{u}_1$, and so $|s_2| \leq lcs(u_1,\overline{u}_1)$, hence a contradiction. Therefore the same arguments as in the case $v^2 = [{u_1}^p u_2 {u_1}^q][{u_1}^p u_2 {u_1}^q]$ can be applied.

Case 2.b. Assume $p < q$. By Synchronization Principle Lemma 1.6 and since $|s_2| > lcs(u_1,\overline{u}_1)$, the leftmost possible beginning of $V_{[2]}$ can be at $|s_2{u_1}^p u_2{u_1}^{p+q}u_2{u_1}^{q-p}s_1| + 1$. Again, if it started to the left of this point, by Synchronization Principle Lemma 1.6, s_2 would have to be a suffix of u_1u_2 and so s_2 would be a common suffix of u_1 and $\overline{u}_1$, and so $|s_2| \leq lcs(u_1,\overline{u}_1)$, hence a contradiction. Therefore, the same arguments as in the case $v^2 = [{u_1}^p u_2 {u_1}^q][{u_1}^p u_2 {u_1}^q]$ can be applied.

If $|v_2| \geq |u_1|$, then a prefix of $V_{[2]}$ must align with the last u_1 of ${u_1}^p u_2 {u_1}^{q+p} u_2 {u_1}^q$ and so ${u_1}^p u_2 {u_1}^{q+p} u_2 {u_1}^q$ is extended for sure by another u_2, i.e. V^2 contains a factor ${u_1}^q u_2 {u_1}^q u_2$. □

3.4 δ-mates

We are able to carry on the induction hypothesis in Chapter 4 provided two neighbouring FS-double squares either have a sufficiently large difference – called *gap* – between their respective starts, or a sufficiently large difference – called *tail* — between the respective ends of their short squares. The three types we have investigated so far, α-, β-, and γ-mates have small gaps and tails, and are some kind of *shifts*: complete shifts of the short and the long square or α-mates, shifts of the long square, while *shrinking* the short square for β-mates, shifts of the long square to become the short square for γ-mates. In contrast, the δ-mates relation ensures a sufficiently large tail.

Definition 3.6. *Consider FS-double squares $\mathcal{U} \lessdot\!\!\lessdot \mathcal{V}$. If $|v| > |U|$, then $\mathcal{V}$ is a δ-mate of $\mathcal{U}$. The relation is denoted by $\mathcal{U} \underset{\delta}{\rightarrow} \mathcal{V}$.*

The δ-mate relation is illustrated in Figure 3.4 where thee canonical inversion factors are highlighted, u_2, v_2 are underlined, and $\bar{u}_2, \bar{v}_2$ are under-dotted.

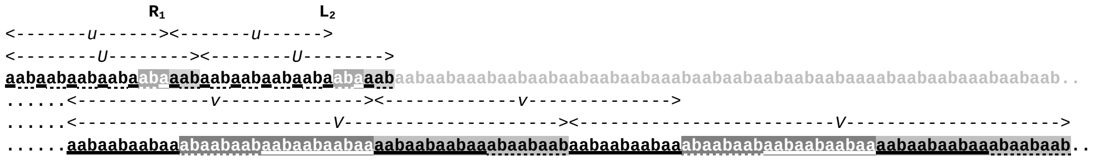

R₁
L₂
<-------u------><-------u------>
<--------U--------><--------U-------->
aabaabaabaabaabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaabaaabaabaabaabaabaaaabaabaabaaabaabaab..
......<-------------v--------------><-------------v-------------->
......<-------------------------V--------------------><------------------------V--------------------->
......aabaabaabaaabaabaabaabaabaabaaaabaabaabaaabaabaabaabaabaaabaabaabaabaabaabaaaabaabaabaaabaabaab..

Figure 3.4: An example of a δ-mate

Observation 3.7. *Let $\mathcal{V}$ be a δ-mate of $\mathcal{U}$, then $v_{[1]}$ ends after $L_2(\mathcal{U})$. Thus, $v_{[1]}$ must contain the whole first inversion factor.*

Proof. Note that, by definition, $v_{[1]}$ contains the first inversion factor as $|v| > |U|$. Recall that the second inversion factor is exactly at a distance $|U|$. Suppose, by contradiction, that $v_{[1]}$ ends before $L_2(\mathcal{U})$. Set $x = |v| - |U| \geq 1$ and y the position at which the first inversion factor starts relative to $v_{[1]}$. Then $v_{[1]}$ contains the inversion factor at position y and $v_{[2]}$ at position $y - x < y$, hence a contradiction. □

3.5 ε-mates

Finally, we consider double squares starting after the beginning of the inversion factor of $\mathcal{U}$, i.e. after $R_1(\mathcal{U})$.

Definition 3.8. *Consider FS-double squares $\mathcal{U}{:=}\ (u_1, u_2, e_1, e_2) \prec \mathcal{V}$. Let $s(\mathcal{V})$ denote the starting position of $\mathcal{V}$, and $e(u_{[1]})$ denote the end of $u_{[1]}$. If $R_1(\mathcal{U}) \leq s(\mathcal{V})$, then $\mathcal{V}$ is said to be an* ε-mate *of $\mathcal{U}$, denoted by $\mathcal{U} \underset{\varepsilon}{\rightarrow} \mathcal{V}$. If moreover $e(u_{[1]}) < s(\mathcal{V})$, then $\mathcal{V}$ is said to be a* super-ε-mate *of $\mathcal{U}$.*

Figure 3.5 illustrates an ε-mate and Figure 3.6 illustrates a super-ε-mate.

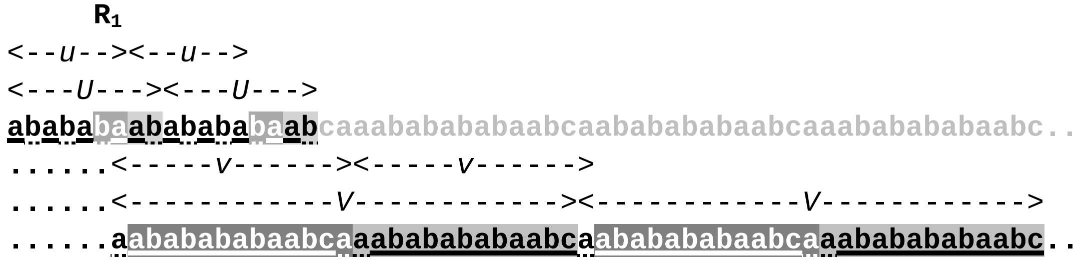

R₁
<--u--><--u-->
<---U---><---U--->
abababaababababaabcaaabababababaabcaabababababaabcaaabababababaabc..
......<-----v------><-----v------>
......<------------V------------><------------V------------>
......aabababababaabcaaabababababaabcaabababababaabcaaabababababaabc..

Figure 3.5: An example of an ε-mate

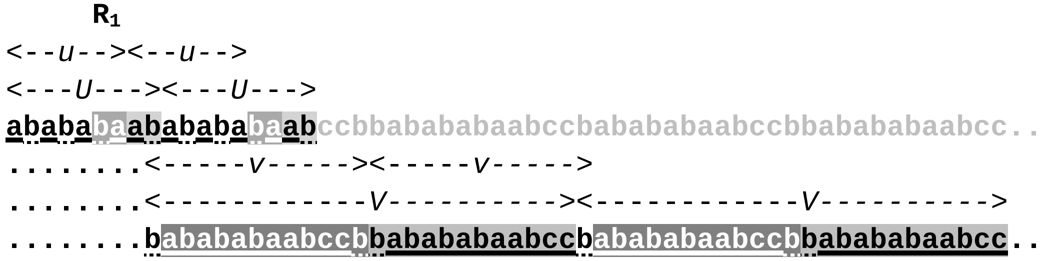

R₁
<--u--><--u-->
<---U---><---U--->
abababaababababaabccbbababababaabccbababababaabccbbababababaabcc..
........<-----v-----><-----v----->
........<------------V----------><------------V---------->
........bababababaabccbbababababaabccbababababaabccbbababababaabcc..

Figure 3.6: An example of a super-ε-mate

While the gap $G(\mathcal{U},\mathcal{V})$ and tail $T(\mathcal{U},\mathcal{V})$ are analyzed in detail in Chapter 4, we prove first some of auxiliary properties of ε-mates with respect to the gap and tail. The *gap* between $\mathcal{U}$ and $\mathcal{V}$ indicates how far v starts after u, the tail how far v ends after u. If $ut = gv$, $G(\mathcal{U},\mathcal{V}) = g, T(\mathcal{U},\mathcal{V}) = t$. Unlike the gap, the tail may not exist. Lemma 3.9 is used in Chapter 4 to carry on the induction

Lemma 3.9. *Let* $\mathcal{U} \underset{\gamma}{\rightarrow} \mathcal{V} \underset{\varepsilon}{\rightarrow} \mathcal{W}$*,* $\mathcal{V}$ *be a* γ*-mate of* $\mathcal{U}$ *of type* (e_1-t, e_2+t)*,* $2 \le e_1 - t$ *and* $2 \le e_2 + t$*, and* $\mathcal{W}$ *be an* ε*-mate but not a super-*ε*-mate of* $\mathcal{V}$*,* *then* $|G(\mathcal{U},\mathcal{W})| \geq t|u_1|$ *and* $|T(\mathcal{U},\mathcal{W})| \geq (e_1+e_2)|u_1|$.

Proof. The position of v^2 is ${u_1}^t s_1 \left[s_2 {u_1}^{e_1-t-1} u_2 {u_1}^{e_2+t} s_1\right] \left[s_2 {u_1}^{e_1-t-1} u_2 {u_1}^{e_2+t} s_1\right]$. Since $\mathcal{V}$ is a γ-mate of $\mathcal{U}$, by Lemma 3.5 $e'_1 = e'_2$ and so $\mathcal{V}$ cannot have a β-mate, see Lemma 3.12. Thus $w_{[1]}$ must end past the end of $v_{[1]}$ and thus by Synchronization Principle Lemma 1.6, $|w| \geq |v|$. Therefore, $G \geq t|u_1|$ and $T \geq (e_1+e_2)|u_1|$. □

Lemma 3.10. *Let $\mathcal{V}$ be a super-ε-mate of $\mathcal{U}$, then either*

(*a*) $|G(\mathcal{U},\mathcal{V})| \geq (2e_1+e_2-3)|u_1|+2|u_2|$ *and* $|T(\mathcal{U},\mathcal{V})| \geq (e_1+e_2-2)|u_1|+|u_2|$, *or*

(*b*) $|G(\mathcal{U},\mathcal{V})| \geq e_1|u_1|+|u_2|$ *and* $|T(\mathcal{U},\mathcal{V})| \geq (e_1+e_2-1)|u_1|+|u_2|$.

Proof. We use $e(x)$ to indicate the end point of x and $s(x)$ to indicate the starting point of x. If v^2 were a factor of ${u_1}^{e_1+e_2-1}u_2$, then there would be a farther copy of v^2 in ${u_1}^{e_1+e_2}u_2$ – just starting $|u_1|$ positions to the right, which is a contradiction as v^2 must be a rightmost occurrence. Hence $e(v^2) > |u_1{}_1^e u_2 {u_1}^{e_1+e_2-1}u_2|$. Let us assume that $v_{[1]}$ is a factor in ${u_1}^{e_1}u_2{u_1}^{(e_1+e_2-1)}u_2$. Then ${u_1}^{(e_1+e_2)}u_2$ and v^2 both contain a common factor of size $|v|+|u_1|$, and thus by Common Factor Lemma 1.5, $v = {v_1}^k$ for some conjugate v_1 of u_1 and some $k \geq 1$. If $k=1$, then $s(v_{[1]}) \geq |u_1{}_1^e u_2 {u_1}^{(e_1+e_2-3)}u_2|$ and so $G \geq |u_1{}_1^e u_2 {u_1}^{(e_1+e_2-3)}u_2|$. Moreover $s(v_{[2]}) = s(v_{[1]}) + |u_1|$ and so $T \geq |{u_1}^{(e_1+e_2-2)}u_2|$, i.e. (a) holds.

Let us assume that $k \geq 2$ and consider two sub-cases

(i) $v_{[1]}$ starts in $\overline{u}_2$ and ends in $\overline{u}_2$ Then, there are $s_1s_2 = \overline{u}_2$ so that $v = (s_2u_2s_1)^k$ and v^2s_2 is a suffix of ${u_1}^{e_1}u_2{u_1}^{(e_1+e_2)}$.

(i_1) Let $|s_2| \leq lcs(u_1, \overline{u}_1)$. Then, we can assume without loss of generality that $v = {u_1}^k$ as otherwise we can cyclically shift the whole structure $|s_2|$ positions to the left. Hence $V = {u_1}^{2k-1}t_1$ for some non-trivial proper prefix t_1 of u_1. Let $t_1t_2 = u_1$. Then, the prefix ${u_1}^3$ of $V_{[2]}$ must align by Synchronization Principle Lemma 1.6 with $t_2u_1u_1$ and hence $t_2u_2 = u_1$. Therefore $|t_2| = |\overline{u}_2|$ and since t_2 is a suffix of $u_1 = u_2\overline{u}_2$, in fact $t_2 = \overline{u}_2$, Hence $u_1 = \overline{u}_2u_2$, a contradiction.

(i_2) Let $|s_2| > lcs(u_1, \overline{u}_1)$. Hence $V = (s_2u_2s_1)^{2k-1}t_1$ where t_1 is a non-trivial proper prefix of $s_2u_2s_1$. Let $t_1t_2 = s_2u_2s_1$. Then, the prefix $(s_2u_2s_1)^3$ of $V_{[2]}$ must align by Synchronization Principle Lemma 1.6 with $t_2u_2u_2s_1s_2u_2s_1s_2u_2$ and so either $t_2u_2 = s_2$ or $t_2u_2 = s_su_2s_1s_2$. In either case, s_2 is a suffix of t_2u_2 and since s_2 is a suffix if $\overline{u}_2$, s_2 is both a suffix of u_1 and of $\overline{u}_1$. Hence $|s_2| \leq lcs(u_1, \overline{u}_1)$, a contradiction.

(ii) $v_{[1]}$ starts in u_2 and ends in u_2. Then, there are $s_1s_2 = u_2$ so that $v = (s_2\overline{u}_2s_1)^k$ and v^2s_2 is a suffix of $u_1{}_1^e u_2{u_1}^{(e_1+e_2)}u_2$.

(ii_1) Let $|s_2| \leq lcs(u_1, \overline{u}_1)$. Then, without loss of generality, we can assume $v = (\overline{u}_2u_2)^k$ and v^2 is a suffix of $u_1{}_1^e u_2{u_1}^{(e_1+e_2)}u_2$ as otherwise we could cyclically shift the whole structure $|s_2|$ positions to the left. Then, a suffix $(\overline{u}_2u_2)(\overline{u}_2u_2)(\overline{u}_2u_2)(\overline{u}_2u_2)$ of v^2 must align with
$(\overline{u}_2u_2)(\overline{u}_2u_2)(\overline{u}_2u_2)(u_2\overline{u}_2)(u_2\overline{u}_2)$ giving $\overline{u}_2u_2 = u_2\overline{u}_2$, a contradiction.

(ii_2) Let $|s_2| > lcs(u_1, \overline{u}_1)$. Then, $v = (s_2\overline{u}_2 s_1)^k$ and so $V = (s_2\overline{u}_2 s_1)^{2k-1}t_1$ and $t_1 t_2 = s_2\overline{u}_2 s_1$. Then, a prefix $(s_2\overline{u}_2 s_1)^3$ of $V_{[2]}$ must align by Synchronization Principle Lemma 1.6 with $t_2 s_1 s_2 \overline{u}_2 s_1 s_2 \overline{u}_2$ and so $t_2 = s_2\overline{u}_2$. Since $t_1 t_2 = s_2\overline{u}_2 s_1$, then $t_1 t_2 s_2 = s_2\overline{u}_2 s_1 s_2 = s_2\overline{u}_2 u_2$, i.e. $t_1 t_2 s_2 = s_2\overline{u}_1$ and so s_2 is both a suffix of $\overline{u}_1$ and a suffix of u_2 and hence of u_1, and so $|s_2| \leq lcs(u_1, \overline{u}_1)$, a contradiction.

Considering the end of v^2 in the next $\overline{u}_2$ yields a contradiction using the same argumentation as for (i), and considering the end of v^2 in the next u_2 yields a contradiction using the same argumentation as for (ii). Thus, the only remaining case is when $v_{[1]}$ is not a factor in ${u_1}^{e_1} u_2 {u_1}^{(e_1+e_2-1)} u_2$, i.e. $e(v_{[1]}) > {u_1}^{e_1} u_2 {u_1}^{(e_1+e_2-1)} u_2$ and so $G \geq |u_{1}{}_1^{e} u_2|$ and $T \geq |{u_1}^{(e_1+e_2-1)} u_2|$, i.e. case ($b$) holds. □

3.6 Exhaustivity of the mate relation

In Sections 3.1, 3.2, 3.3, 3.4 and 3.5 we discussed five relations, namely α, β, γ, δ, and ε-mates. We now show that there is no other relation. For this purpose, we use the inversion factors to limit the position and the size of a third square with respect to a given FS-double square $\mathcal{U}$. Using the result for vv and VV of $\mathcal{V}$ we establish that only the five mates types we have defined are possible.

Lemma 3.11. *Let w be a string starting with a FS-double square $\mathcal{U} := (u_1, u_2, e_1, e_2)$ and v^2 be a rightmost occurrence in w starting at a position*

$i_v < R_1(\mathcal{U})$, *then*

(*a*) *either* $|v| < |u|$ *and* $v = \widetilde{u}_1^{\,j}\widetilde{u}_2$ *for some* $1 \leq j < e_1$ *where* $\widetilde{u}_1$ *is a cyclic shift of* u_1 *and* $\widetilde{u}_2$ *the prefix of length* $|u_2|$ *of* $\widetilde{u}_1$*;*

(*b*) *or* $|v| = |u|$ *and* $v = \widetilde{u}_1^{\,e_1}\widetilde{u}_2$ *where* $\widetilde{u}_1$ *is a cyclic shift of* u_1 *and* $\widetilde{u}_2$ *the prefix of length* $|u_2|$ *of* $\widetilde{u}_1$*;*

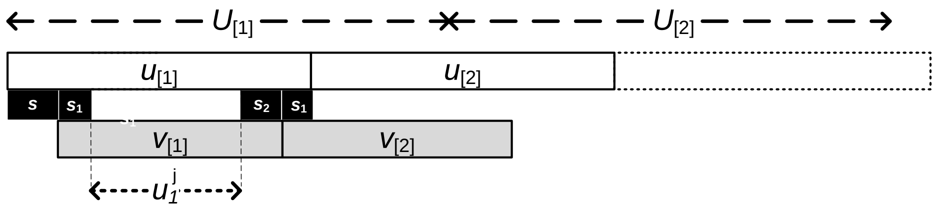
(*c*) *or* $|v| = |U|$*;*

(*d*) *or* $|v| > |U|$ *and either* $s_1\overline{u}_2u_2{u_1}^{(e_1+e_2-1)}u_2$ *is a prefix of* v *for some suffix* s_1 *of* u_2*, or* $s_1{u_1}^{i}u_2{u_1}^{(e_1+e_2-1)}u_2$ *is a prefix of* v *for some suffix* s_1 *of* u_1 *and some* $i \geq 1$*.*

Proof. Without loss of generality we can assume that *lcp*$(u_1,\overline{u}_1)= 0$ and thus $R_1(\mathcal{U}) = N_1(\mathcal{U})$. If *lcp*$(u_1,\overline{u}_1) \neq 0$, instead of showing the argument for ${u_1}^{e_1}u_2{u_1}^{(e_1+e_2)}u_2{u_1}^{e_2}$ we can use $s_2{w_1}^{e_1}w_2{w_1}^{(e_1+e_2)}w_2{w_1}^{(e_2-1)}s_1$ where w_1, respectively w_2, is a right cyclic shift of w_1, respectively w_2 .by *lcp*$(u_1,\overline{u}_1)$ positions. That is, $s_1s_2 = w_1$, $|s_1| =$ *lcp*$(u_1,\overline{u}_1)$, and thus *lcp*$(w_1, \overline{w}_1) = 0$. The proof considers all possible endings of $v_{[1]}$. Recall that $e(x)$ denotes the ending position of x in w.

(1) **Case** $e(v_{[1]}) \leq e(u_{[1]})$**.** Note that $e(vv) > e(U_{[1]}) = e({u_1}^{e_1}u_2{u_1}^{e_2})$ as otherwise there would be a farther copy of vv in $U_{[2]}$. Since $v_{[1]}$ does not contain any inversion factors, neither $v_{[2]}$ can, and thus cannot contain the inversion factor at N_1. Therefore, $v_{[1]}$ must end in the suffix $\overline{u}_2u_2$ of $u_{[1]}$. Let s be the offset of $v_{[1]}$ in $u_{[1]}$, and s_1 be the overlap between $u_{[1]}$ and $v_{[2]}$, i.e. $svs_1 = u = {u_1}^{e_1}u_2$, see the diagram bellow for an

illustration.

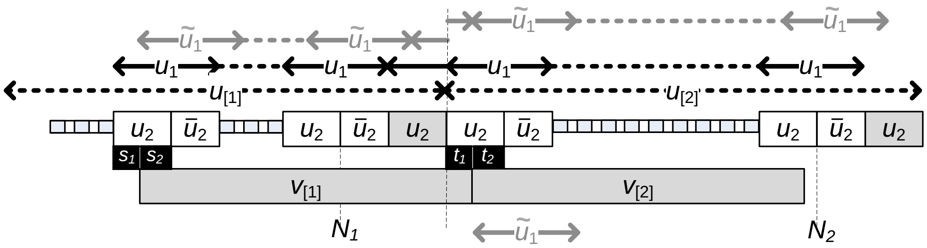


Then s_1 is both a prefix of v and a suffix of u. Since s_1 is the overlap of u_1 and v_1, $|s_1| < |u_1|$ and s_1 is a suffix of $\overline{u}_2u_2$. Thus, $v = t_1{u_1}^it_2$ for some suffix t_1 of u_1, some prefix t_2 of u_1, and some $i \geq 0$. Since $U_{[1]} = {u_1}^{e_1}u_2{u_1}^{e_2} = u{u_1}^{e_2}$ is a non-trivial proper prefix of sv^2, and so $svs_1{u_1}^{e_2}$ is a non-trivial proper prefix of sv^2, implying that $s_1{u_1}^{e_2}$ is a non-trivial proper prefix of v and, therefore, $v = s_1u_i^js_2$ for some prefix s_2 of u_1 and some $j \geq 1$. Thus, $v = t_1{u_1}^it_2 = s_1{u_1}^js_2$. Since t_1 is a suffix of u_1 and t_2 a prefix of u_1, by Synchronization Principle Lemma 1.6, $t_1 = s_1$ and $t_2 = s_2$. Therefore, s_1 is a suffix of u_1. Since s_2s_1 is a suffix of u, then $s_2s_1 = {u_1}^iu_2$ for some $i \geq 0$. Since $|s_2| + |s_1| < 2|u_1|$, either $i = 0$ or $i = 1$, which proves that either $s_2s_1 = u_1u_2$ or $s_2s_1 = u_2$. In the former case, $|v| = (j+1)|u_1| + |u_2|$ and so $v = \widetilde{u}_1^{(j+1)}\widetilde{u}_2$, while in the latter case $v = {\widetilde{u}_1}^j\widetilde{u}_2$, where in both cases $\widetilde{u}_1$, respectively $\widetilde{u}_2$, is a left cyclic rotation of u_1, respectively u_2, by $|s_1|$ positions. The left cyclic shift is possible as s_1 is both a suffix of u_1 and a suffix of $\overline{u}_1 = \overline{u}_2u_2$. Therefore, $v = {\widetilde{u}_1}^j\widetilde{u}_2$ and $1 \leq j \leq e_1$ and so when $j < e_1$, case (a) holds, and when $j = e_1$, case (b) holds.

(2) **Case** $e(u_{[1]}) < e(v_{[1]}) \leq e(u_{[1]}u_1)$**.** We consider four sub-cases based on

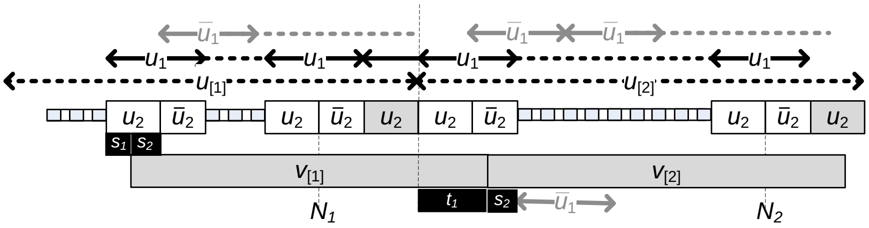
where $v_{[1]}$ starts and ends.

(2.1) $v_{[1]}$ **starts in a** u_2 **and ends in the first** u_2 **of** $u_{[2]}$**.** Let s_1 be the offset of $v_{[1]}$ in the u_2 it starts in, s_2 be the overlap of $v_{[1]}$ and the u_2 it starts in, let t_1 be the overlap of $v_{[1]}$ with the u_2 it ends in, and t_2 be the overlap of $v_{[2]}$ with the u_2 where $v_{[1]}$ ends. Let $\widetilde{u}_1 = s_2\overline{u}_2 s_1$; as a conjugate of u_1, it is primitive.



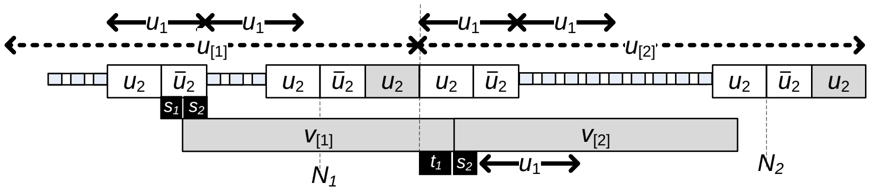
By Synchronization Principle Lemma 1.6, $t_1 = s_1$ and $t_2 = s_2$, and so $s_1 v_{[2]}$ is a non-trivial proper prefix of ${u_1}^{(e_1+e_2)}u_2$. Thus, the suffix $u_2 s_1$ of v must align with $u_1 u_2 = (u_2\overline{u}_2)u_2$ of ${u_1}^{(e_1+e_2)}u_2$, and so s_1 is a prefix of $u_2\overline{u}_2$. Thus, s_1 is a prefix of both, $u_2\overline{u}_2$ and $\overline{u}_2 u_2$. Therefore, $|s_1| \leq lcp(u_1,\overline{u}_1) = 0$, and so s_1 is empty. Thus, $v = u_1^j u_2$ for $1 \leq j \leq e_1$ and either (a) or (b) holds. Note that Synchronization Principle Lemma 1.6 applies even if $e_1 = 1$, since then $v_{[1]}$ must start in the first u_2 of $u_{[1]}$.

(2.2) $v_{[1]}$ **starts in a** u_2 **and ends in the first** $\overline{u}_2$ **of** $u_{[2]}$**.** Let s_1 and s_2 be as in case (2.1), and t_1 be the overlap of $v_{[1]}$ and $u_{[2]}$.



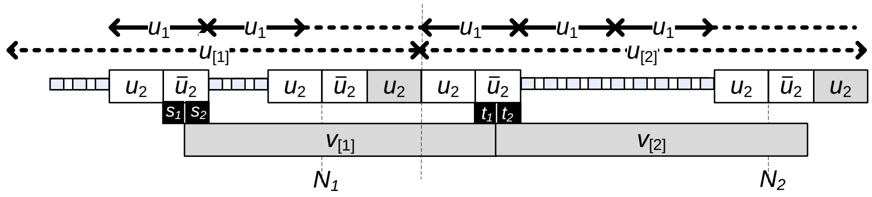
The factor ${u_1}^{(e_1+e_2)}u_2$ has $u_2\overline{u}_1\overline{u}_1$ as a prefix as $e_1 + e_2 \geq 2$. The factor v has $s_2\overline{u}_1$ as a prefix. Thus ${u_1}^{(e_1+e_2)}u_2$ has also $t_1s_2\overline{u}_1$ as a prefix. Since $|t_1s_2| < |u_2|+|u_1|$, this contradicts Synchronization Principle Lemma 1.6 as $\overline{u}_1$ is primitive being a conjugate of u_1.

(2.3) $v_{[1]}$ **starts in a** $\overline{u}_2$ **and ends in the first** u_2 **of** $u_{[2]}$**.** Let s_1 be the offset of $v_{[1]}$ in $\overline{u}_2$ it starts in, s_2 be the overlap of $v_{[1]}$ and the $\overline{u}_2$ it starts in, and t_1 be the overlap of $v_{[1]}$ with $u_{[2]}$.



The factor v has s_2u_1 as a prefix, and so ${u_1}^{(e_1+e_2)}$ has as a prefix u_1u_1 and $t_1s_2u_1$. Since $|t_1s_2| < |u_1|$, this contradicts Synchronization Principle Lemma 1.6.

(2.4) $v_{[1]}$ **starts in a** $\overline{u}_2$ **and ends in the first** $\overline{u}_2$ **of** $u_{[2]}$**.** Let s_1 and s_2 be as in case (2,3), t_1 be the overlap of $v_{[1]}$ and the $\overline{u}_2$ it ends in, and t_2 be the overlap of $v_{[2]}$ with the $\overline{u}_2$ in which $v_{[1]}$ ends.



By Synchronization Principle Lemma 1.6, $t_1 = s_1$ and $t_2 = s_2$. Since $u_2 s_1 v_{[2]}$ is a prefix of ${u_1}^{(e_1+e_2)} u_2$, the suffix $u_2 u_2 s_1$ of $v_{[2]}$ must align with $u_1 u_2$ in ${u_1}^{(e_1+e_2)} u_2$. Thus, $u_2 u_2 s_1$ is a prefix of $u_2 \overline{u}_2 u_2$, hence $u_2 s_1$ is a prefix of $\overline{u}_2 u_2$. Thus, $u_2 \overline{u}_2 = u_2 s_1 s_2$ is a prefix of $\overline{u}_2 u_2 s_2$, giving $u_2 \overline{u}_2 = \overline{u}_2 u_2$; s a contradiction as $u_2 \overline{u}_2$ is primitive.

(3) **Case** $e(u_{[1]} u_1) < e(v_{[1]}) < R_2$**.** Then, $v_{[1]}$ contains the inversion factor at R_1. Thus, $v_{[2]}$ must contain the inversion factor at R_2 and it must be placed in $v_{[2]}$ in the same position mate to the beginning of $v_{[2]}$ as in $v_{[1]}$, and therefore $|v| = R_2 - R_1 = |U|$. Thus, case (c) holds.

(4) **Case** $R_2 \leq e(v_{[1]})$**.** Since $e(v_{[1]}) \geq R_2 \geq N_2 = {u_1}^{e_1} u_2 {u_1}^{(e_1+e_2-1)} u_2$, either $s_1 \overline{u}_2 u_2 {u_1}^{(e_1+e_2-1)} u_2$ for some suffix s_1 of u_2 is a prefix of v, or $s_1 {u_1}^i u_2 {u_1}^{(e_1+e_2-1)} u_2$ for some suffix s_1 of u_1 and some $i \geq 1$ is a prefix of v, i.e. case (d) holds.

. □

Lemma 3.11 is a key tool in the analysis of the mate relation. The cases (a) and (b) relate to α-mates or β-mates,(c) to γ-mates, and (d) to δ-mates. If the starting position i_v of vv is not strictly smaller than R_1, it relates to

ε-mates. To extend the analysis to double squares, we apply Lemma 3.11 to both vv and VV.

Lemma 3.12. *Let x be a string starting with a FS-double square $\mathcal{U} := (u_1, u_2, e_1, e_2)$ and let $\mathcal{V}$ be a FS-double square so that $\mathcal{U} \prec \mathcal{V}$, then either*

(α) *$\mathcal{V}$ is an α-mate of $\mathcal{U}$, or*

(β) *$\mathcal{V}$ is a β-mate of $\mathcal{U}$ and $e_1 > e_2 + 1$, or*

(γ) *$\mathcal{V}$ is a γ-mate of $\mathcal{U}$, or*

(δ) *$\mathcal{V}$ is a δ-mate of $\mathcal{U}$, or*

(ε) *$\mathcal{V}$ is an ε-mate of $\mathcal{U}$ and the end of $v_{[1]} >$ the end of $u_{[1]}$.*

Proof. Let i_v be the starting position of $\mathcal{V}$ and recall that $e(w)$ denote the end point of a factor w. If $i_v \geq R(\mathcal{U})$, $\mathcal{V}$ is an ε-mate of $\mathcal{U}$ by definition of ε-mate. Thus, we can assume that $i_v < R(\mathcal{U})$. Applying Lemma 3.11 to v^2 gives the following possibilities:

(*i*) $v = \widetilde{u}_1^{\,i}\widetilde{u}_2$ for $1 \leq i < e_1$ where $\widetilde{u}_1$ is a cyclic shift of u_1 and $\widetilde{u}_2$ the prefix of length $|u_2|$ of $\widetilde{u}_1$, by Lemma 3.11 (a).

(*ii*) $v = \widetilde{u}_1^{\,e_1}\widetilde{u}_2$ where $\widetilde{u}_1$ is a cyclic shift of u_1 and $\widetilde{u}_2$ is the prefix of length $|u_2|$ of $\widetilde{u}_1$, by Lemma 3.11 (b).

(*iii*) $|v| = |U|$, by Lemma 3.11 (c).

(*iv*) $e(v_{[1]}) - e(u_{[1]}) \geq (e_1 + e_2 - 1)|u_1| + |u_2|$, by Lemma 3.11 ($d$).

Applying Lemma 3.11 to V^2 gives the following possibilities:

(I) $V = {\widehat{u}_1}^j\widehat{u}_2$ for $1 \leq j \leq e_1$ where $\widehat{u}_1$ is a cyclic shift of u_1 and $\widehat{u}_2$ the prefix of length $|u_2|$ of $\widetilde{u}_1$, by Lemma 3.11 (a, b).

(II) $|V| = |U|$, by Lemma 3.11 (c).

(III) Either $s_1\overline{u}_2u_2{u_1}^{(e_1+e_2-1)}u_2$ for some suffix s_1 of u_2 is a prefix of V, or $s_1{u_1}^iu_2{u_1}^{(e_1+e_2-1)}u_2$ for some suffix s_1 of u_1 and some $j \geq 1$ is a prefix of V, by Lemma 3.11 (d).

Considering all possible combinations yields:

Combining (i) and (I) is impossible. Since v is a prefix of V, $\widetilde{u}_1 = \widehat{u}_1$ and $\widetilde{u}_2 = \widehat{u}_2$. Since $j > i$ as $|V| > |v|$, this contradicts the properties of the canonical factorization.

Combining (i) and (II) is possible and yields case (β). Since v is a prefix of V, $\widetilde{u}_1 = \widehat{u}_1$ and $\widetilde{u}_2 = \widehat{u}_2$ and so $\mathcal{V}$ must be a β-mate of $\mathcal{U}$. Since $|V| = |U| = (e_1+e_2)|u_1|+|u_2|$, $V = {\widetilde{u}_1}^i\widetilde{u}_2\widetilde{u}_1^{(e_2+e_1-i)}$. Since $i \geq e_2+e_1-i$ as otherwise there would be a farther copy of v^2, $2i \geq e_1 + e_2$. Since $1 \leq i < e_1$, $i = e_1 - k$ for some $1 \leq k < e_1$. Thus, $2(e_1 - k) \geq e_1 + e_2$, so $2e_1 - 2k \geq= e_1 + e_2$, and thus $e_1 \geq e_2 + 2$.

Combining (i) and (III) is impossible. Since v^2 is a prefix of V^2, ${\widetilde{u}_1}^i\widetilde{u}_2{\widetilde{u}_1}^i\widetilde{u}_2$ is a prefix of V^2. At the same time either $s_1{u_1}^ju_2{u_1}^{(e_1+e_2-1)}u_2$ is a prefix of V or $s_1\overline{u_2}u_2{u_1}^{(e_1+e_2-1)}u_2$ is a prefix of V. Due to the Synchronization Principle Lemma 1.6, in both cases, ${\widetilde{u}_1}^i\widetilde{u}_2\widetilde{u}_1^{(e_1+e_2-1)}\widetilde{u}_2$

is a prefix of V and so $\widetilde{u}_1^{\,i}\widetilde{u}_2\widetilde{u}_1^{\,i}\widetilde{u}_2$ is a prefix of V. Thus, v^2 is a factor in $V_{[1]}$ and, consequently, it has a farther copy in $V_{[2]}$, a contradiction.

Combining (ii) and (I) is impossible since $j \leq e_1$ implies that $|V| \leq |v|$, hence a contradiction.

Combining (ii) and (II) is possible and yields that $\mathcal{V}$ is an α-mate of $\mathcal{U}$, hence case (α).

Combining (ii) and (III) is impossible for the same reasons as for the combination (i) and (III).

Combining (iii) and (I) or (II) is impossible due to the size of v being bigger than the size of V.

Combining (iii) and (III) is possible and yields case (γ).

Combining (iv) and (I) or (II) is impossible due to the size of v being bigger than the size of V.

Combining (iv) and (III) yields case (δ).

□

Proposition 3.13. *For any two FS-double squares $\mathcal{U} \prec \mathcal{V}$, $\mathcal{V}$is an α-, β-, γ-, δ-, or ε-mate of $\mathcal{U}$.*

Proof. If $\mathcal{U} \prec\!\!\prec \mathcal{V}$, $\mathcal{V}$ is an α-, β-, γ or δ-mate by Lemma 3.12. If $R_1 \leq s(v)$ where $s(v)$ is the starting position of $\mathcal{V}$, $\mathcal{V}$ is an ε-mate of $\mathcal{U}$ by definition. □

Chapter 4

At most $\lfloor 11n/6 \rfloor$ distinct squares

After introducing structural properties of FS-double squares in Chapters 2 and 3, we present our main result: A string of length n has at most $\lfloor 11n/6 \rfloor$ distinct squares. This bound is obtained by showing that the number of FS-double squares is at most $\lfloor 5n/6 \rfloor$. The proof of the $\lfloor 5n/6 \rfloor$ upper bound is carried by induction, starting from the right end of the string to the beginning while counting the number of FS-double squares. The induction hypothesis is $\Delta(x) \leq \frac{5}{6}|x| - \frac{1}{3}|u|$ where $\Delta(x)$ indicates the number of FS-double squares in a string x, and u is the root of the small square of the leftmost FS-double square $\mathcal{U}$. Gap-Tail Lemma 4.2 shows how to make an induction step.

Definition 4.1. *Let $\mathcal{U}$ and $\mathcal{V}$ be FS-double squares such that $\mathcal{U} \prec \mathcal{V}$. Let g be a string such that gv is a prefix of u or u is a prefix of gv, then, $G(\mathcal{U},\mathcal{V}) = g$ and is called the* gap *between $\mathcal{U}$ and $\mathcal{V}$. If there is a string t such that $gv = ut$, then $T(\mathcal{U},\mathcal{V}) = t$ and is called the* tail.

Note that while the gap always exists,t the tail may not if v is completely contained in u. Figure 4.1 illustrates the gap and tail between $\mathcal{U}$ and $\mathcal{V}$.

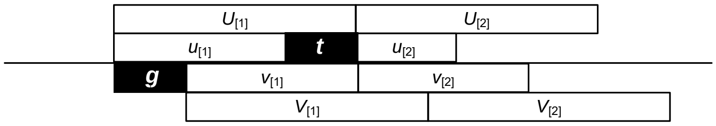


Figure 4.1: Gap g and tail t between $\mathcal{U}$ and $\mathcal{V}$

Lemma 4.2 (Gap-Tail Lemma)**.** *Let x be a string starting with a FS-double square $\mathcal{U}$, $\mathcal{V}$ a FS-double square so that $\mathcal{U} \prec \mathcal{V}$, x' the suffix of x starting with $\mathcal{V}$, d the number of FS-double squares strictly preceding $\mathcal{V}$ (including $\mathcal{U}$), then $\Delta(x') \leq \frac{5}{6}|x'| - \frac{1}{3}|v|$ implies $\Delta(x) \leq \frac{5}{6}|x| - \frac{1}{3}|u| + d - \frac{1}{2}|G(\mathcal{U},\mathcal{V})| - \frac{1}{3}|T(\mathcal{U},\mathcal{V})|$ where $T(\mathcal{U},\mathcal{V})$ is the tail between $\mathcal{U}$ and $\mathcal{V}$.*

Proof. Since $ut = gv$, $|u| + |t| = |g| + |v|$, and $|v| = |u| + |t| - |g|$. Thus $\Delta(x) = d + \Delta(x') \leq d + \frac{5}{6}|x'| - \frac{1}{3}|v|$. It follows that $\Delta(x) \leq \frac{5}{6}(|x'| + |g|) - \frac{1}{3}|u| + d - \frac{5}{6}|g| + \frac{1}{3}|g| - \frac{1}{3}|t| = \frac{5}{6}|x| - \frac{1}{3}|u| + d - \frac{1}{2}|G(\mathcal{U},\mathcal{V})| - \frac{1}{3}|T(\mathcal{U},\mathcal{V})|$. □

In other words, we can carry on the induction from x' to x as long as the gap and the tail is large enough to offset for d – i.e. $d \leq \frac{1}{2}|G| + \frac{1}{3}|T|$. Lemma 4.2 indicates where the challenge is: Ff $\mathcal{U} \underset{\alpha}{\rightarrow} \mathcal{V}$, the gap and the tail may be both as short as 1; if $\mathcal{U} \underset{\beta}{\rightarrow} \mathcal{V}$, then the gap can be very small and the tail is empty; if $\mathcal{U} \underset{\beta}{\rightarrow} \mathcal{W}$, and $\mathcal{U} \underset{\gamma}{\rightarrow} \mathcal{V}$, and $\mathcal{U} \prec \mathcal{W} \prec \mathcal{V}$, then the gap and the tail may be too short. To address these possibilities, we introduce the notion of a family for the leading FS-double square.

4.1 Families

Double squares are in the same family if they are related to each other. Those relationships rely on the mate relations discussed in Chapter 3.

Definition 4.3. *Let x be a string starting with a double square $\mathcal{U}$. If $\mathcal{U}$ has a β-mate, then the $\mathcal{U}$-family consists of $\mathcal{U}$ and all of its $\alpha-$, $\beta-$ and γ-mates. If $\mathcal{U}$ has no β-mates, then the $\mathcal{U}$-family consists only of $\mathcal{U}$ and its α-mates.*

Instead of $\mathcal{U}$-family we may use the expression family of $\mathcal{U}$. From the definition, there are exactly three types of $\mathcal{U}$-families:

- *α-family* consisting exclusively of α-mates of $\mathcal{U}$, in which case either there is no other FS-double square in x, or the first FS-double square after the family must be a γ-mate, or a δ-mate, or an ε-mate of $\mathcal{U}$.

- *$(\alpha+\beta)$-family* consisting of all α-mates and β-mates of $\mathcal{U}$, in which case either there is no other FS-double square in x, or the first FS-double square after the family must be a δ-mate or an ε-mate of $\mathcal{U}$.

- *$(\alpha+\beta+\gamma)$-family* consisting of all α-mates, β-mates, and γ-mates of $\mathcal{U}$, in which case either there is no other FS-double square in x, or the first FS-double square after the family must be a δ-mate or an ε-mate of $\mathcal{U}$.

See Figures 4.2, 4.1.1, and 4.1.2 for examples of these families. A key challenge is to estimate the sizes of these families. To that end, we provide an estimate for the number of cyclic shifts.

Lemma 4.4. *Let x be a string starting with a FS-double square $\mathcal{U} = (u_1, u_2, e_1, e_2)$.*

(a) If $e_1 > e_2$, then the number of cyclic shifts of $\mathcal{U}$ is at most $|u_1| - 2$.

(b) If $e_1 = e_2$, then the number of cyclic shifts of $\mathcal{U}$ is at most $|u_2| - 1$.

Proof. Since $UU = (u_2\bar{u}_2)^{e_1}u_2(u_2\bar{u}_2)^{e_1+e_2}u_2(u_2\bar{u}_2)^{e_2}$, the number of cyclic shifts of $\mathcal{U}$ at most $lcp(u_2\bar{u}_2) + lcp(\bar{u}_2u_2) \leq |u_1| - 2$ by Lemma 1.7. For (b), assume we can cyclically shift more than $|u_2|$ positions. Then $x = U^2u_2z$ for some, possibly empty, z. Thus, $x = {u_1}^e u_2 {u_1}^e \boldsymbol{{u_1}^e u_2 {u_1}^e u_2} z$, i.e. there is a farther occurrence of u^2 shown in bold, a contradiction. □

4.1.1 α-families

Figure 4.2 illustrates an α-family where u_2 is underlined and $\bar{u}_2$ is underdotted.

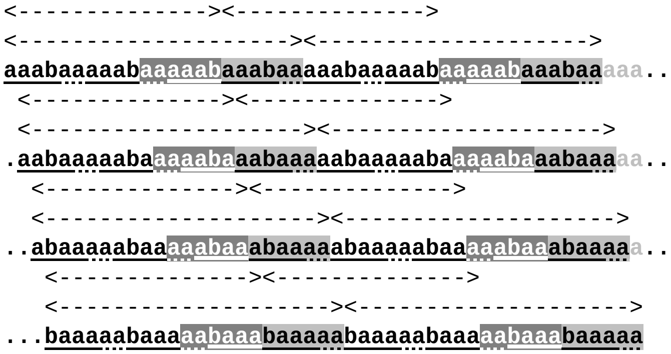

aaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaabaaaaa..
.aabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaabaaaaa..
..abaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaabaaaaa..
...baaaaabaaaaabaaabaaaaabaaaaabaaaaabaaabaaaaa

Figure 4.2: An example of an α-family of $\mathcal{U} = (u_1, u_2, e_1, e_2)$

Lemma 4.4 gives a bound on the size of an α-family of $\mathcal{U} = (u_1, u_2, e_1, e_2)$, namely $|u_1| - 2$ if $e_1 > e_2$ or $|u_2|$ if $e_1 = e_2$. We first consider the case when there are no other FS-double squares besides the α-family.

Observation 4.5. *Let x be a string starting with an α-family of a FS-double square $\mathcal{U} = (u_1, u_2, e_1, e_2)$ with no additional FS-double squares in x, then $\Delta(x) \leq \frac{5}{6}|x| - \frac{1}{3}|u|$.*

Proof. The size of the $\mathcal{U}$-family satisfies $f < |u_1|$, $|u| = e_1|u_1| + |u_2|$, and $|x| \geq |U^2| + f = 2(e_1+e_2)|u_1| + 2|u_2| + f$. Thus, $\frac{5}{6}|x| - \frac{1}{3}|u| \geq \frac{5}{6}(2e_1+e_2)|u_1| + \frac{5}{6}2|u_2| - \frac{2}{3}p|u_1| - \frac{1}{3}|u_2| = \frac{6e_1+5e_2}{6}|u_1| + \frac{8}{6}|u_2| > \frac{11}{6}|u_1| > f = \Delta(x)$. □

The situation is more complicated when α-family of $\mathcal{U}$ is followed by yet another FS-double square $\mathcal{V}$ which, as mentioned previously, can only be a γ-mate, a δ-mate, or an ε-mate.

Observation 4.6. *Let x be a string starting with an α-family of a FS-double square $\mathcal{U} = (u_1, u_2, e_1, e_2)$, $\mathcal{V}$ the first FS-double square that is not a member of the $\mathcal{U}$ family, x' he suffix of x starting with $\mathcal{V}$, then $\Delta(x') \leq \frac{5}{6}|x'| - \frac{1}{3}|v|$ implies $\Delta(x) \leq \frac{5}{6}|x| - \frac{1}{3}|u|$.*

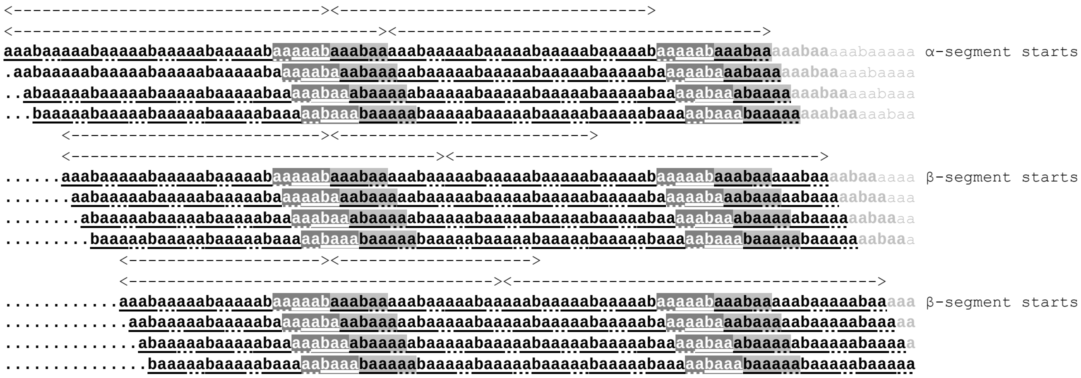
Proof. The size of the $\mathcal{U}$-family satisfies $f < |u_1|$. Let $\mathcal{W}$ be the last member of the α-family of $\mathcal{U}$. Note that $\mathcal{W}$ may equal to $\mathcal{U}$ when the $\mathcal{U}$-family consists only of $\mathcal{U}$. Since $\mathcal{V}$ is neither an α-mate nor a β-mate of $\mathcal{W}$ – as otherwise it would be a member of the family by transitivity of the relations. Then, either it is a γ-mate or a δ-mate, or an ε-mate of $\mathcal{W}$. If it is a γ-mate or a δ-mate, then $|v| \geq |W|$ and the size of the tail between $\mathcal{W}$ and $\mathcal{V}$ is at least $e_2|u_1|$. Since $e_2 \geq 1$, the size of the tail is at least $|u_1|$. Therefore, the size of the gap

G between $\mathcal{U}$ and $\mathcal{V}$ is at least f, the size of the tail T between $\mathcal{U}$ and $\mathcal{V}$ is at least $f+|u_1| \geq 2f$. Thus, $\frac{1}{2}|G|+\frac{1}{3}|T| \geq \frac{1}{2}f+\frac{1}{3}2f = \frac{7}{6}f > f$. If $\mathcal{V}$ is an ε-mate of $\mathcal{W}$, then the gap between $\mathcal{W}$ and $\mathcal{V}$ is at least u_1. Hence, the gap between $\mathcal{U}$ and $\mathcal{V}$ is at least $f + |u_1| \geq 2f$. . Therefore, $\frac{1}{2}|G| + \frac{1}{3}|T| \geq \frac{1}{2}2f = f$. By Lemma 4.2, $\Delta(x) \leq \frac{5}{6}|x| - \frac{1}{3}|u|$. □

4.1.2 $(\alpha+\beta)$-families

Figure 4.3 illustrates an $(\alpha+\beta)$-family. We list a few observations:

- The first so-called *α-segment* consists of $\mathcal{U}$ and its α-mates. The size of this segment is at most $lcp(u_1,\overline{u}_1) \leq |u_1| - 2$ by Lemma 4.4. Note that $e_1 > e_2$ as otherwise there would not be any β-mates of $\mathcal{U}$. All the FS-double squares in this segments have the first exponent equal to e_1 and the second exponent equal to e_2. Thus, we say that the type of the segment is (e_1, e_2).



<--------------------------------><-------------------------------->
<--------------------------------------><-------------------------------------->
aaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaa α-segment starts
.aabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaaaabaaaa
..abaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaabaaaaabaaa
...baaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaaaaaabaaaaabaa
<--------------------------><-------------------------->
<--------------------------------------><-------------------------------------->
......aaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaabaaaaaa β-segment starts
.......aabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaa
........abaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaaa
.........baaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaaaaabaaa
<----------------------><---------------------->
<--------------------------------------><-------------------------------------->
............aaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaa β-segment starts
.............aabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaa
..............abaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaa
...............baaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaaaabaaabaaaaabaaaaabaaaaa

Figure 4.3: An example of an $(\alpha+\beta)$-family

- There must be a β-mate of $\mathcal{U}$ and possibly its α-mates. All the FS-double squares in the segment have the first exponent equal to $e_1 - i_1$ and the second exponent equal to $e_2 + i_1$ for some $1 \leq i_1 \leq \lfloor \frac{e_1 - e_2}{2} \rfloor$, thus we say that the type of the segment is $(e_1 - i_1, e_2 + i_1)$. This so-calledβ-*segment* has size $\leq$ $lcp(u_1,\overline{u}_1)$ + $lcs(u_1,\overline{u}_1) \leq$ $|u_1| - 2$ if $e_1 - i_1 > e_2 + i_1$, or $\leq |u_2| - 1 \leq |u_1| - 2$ if $e_1 - i_1 = e_2 + i_1$.
- There may be another β-segment of type $(e_1 - i_2, e_2 + i_2)$ for some $1 \leq i_1 < i_2 \leq \lfloor \frac{e_1 - e_2}{2} \rfloor$, etc.
- Either there is no other FS-double square in x, or the first FS-double square after the last member of the last β-segment must be either a δ-mate or an ε-mate of $\mathcal{U}$, since if it were a γ-mate, the $\mathcal{U}$-family would be an $(\alpha+\beta+\gamma)$-family discussed in Section 4.1.3.

As for the α-family, we consider two cases: first when there are no other FS-double squares in x except the $\mathcal{U}$-family, and second when there are. For technical reasons related to the proof of Lemma 4.8, we present Observation 4.7 in a more involved setting – not only assuming no other FS-double squares except the ones in the family, we allow possibly other FS-double squares but with a highly specific property – in essence the ε-mates that are not super-ε-mates; note that it does cover the case there is no other FS-double square outside of the family.

Observation 4.7. *Let x be a string starting with an $(\alpha+\beta)$-family of a FS-double square $\mathcal{U}$ $= (u_1, u_2, e_1, e_2)$ and $\mathcal{V}$ the last member of the $\mathcal{U}$-family.*

Assume that for any other FS-double square $\mathcal{W}$, $R_1(\mathcal{U}) \leq s(\mathcal{W}) \leq e(u_{[1]})$, where $s(\mathcal{W})$ is the starting point of $\mathcal{W}$ and $e(u_{[1]})$ is the end point of $u_{[1]}$, then $\Delta(x) \leq \frac{5}{6}|x| - \frac{1}{3}|u|$.

Proof. Let the type of $\mathcal{V}$ be $(e_1 - t, e_2 + t)$, thus $2t \leq e_1 - e_2$. Since every FS-double square $\mathcal{W}$ after $\mathcal{V}$ starts after R_1 but before $e(u_{[1]})$,, the total number of FS-double squares in x is the number of FS-double squares in the $\mathcal{U}$-family plus at most $|u_1|$ additional FS-double squares. Thus, $f \leq (t+2)|u_1|$. Since $|x| \geq |U^2| + f = 2(e_1+e_2)|u_1| + 2|u_2| + f$, we get $\frac{5}{6}|x| - \frac{1}{3}|u| \geq \frac{5}{6}2(e_1+e_2)|u_1| + \frac{5}{6}2|u_2| - \frac{1}{3}e_1|u_1| - \frac{1}{3}|u_2| = \frac{4e_1+5e_2}{3}|u_1| + \frac{4}{3}|u_2| > \frac{4e_1-4e_2}{3}|u_1| + \frac{9e_2}{3}|u_1| > \frac{8t}{3}|u_1| + 2|u_1| > t|u_1| + 2|u_1| \geq f = \Delta(x)$. □

Observation 4.8 deals with the situation when the (α+β)-family is followed by another FS-double square. We need a slightly more involved context than for the α-families and, instead of just the first FS-double square not in the family, we consider all of the FS-double squares not in the family.

Observation 4.8. *Let x be a string starting with an (α+β)-family of a FS-double square $\mathcal{U} = (u_1, u_2, e_1, e_2)$ and assume there are some FS-double squares in x not members of the $\mathcal{U}$ family. Let for any FS-double square $\mathcal{V}$ that is not a member of the $\mathcal{U}$-family $\Delta(x') \leq \frac{5}{6}|x'| - \frac{1}{3}|v|$ where x' is a suffix of x starting at the same position as $\mathcal{V}$, then $\Delta(x) \leq \frac{5}{6}|x| - \frac{1}{3}|u|$.*

Proof. Let the last β-segment be of type $(e_1 - t, e_2 + t)$. Thus, $e_1 - t \geq e_2 + t$, $2t \leq e_1 - e_2$, and the size of the $\mathcal{U}$ family is at most $(t+1)|u_1|$. By Lemma 3.12, $\mathcal{V}$ is either a δ-mate, a γ-mate, or an ε-mate of $\mathcal{U}$. Since $\mathcal{U}$

family is an $(\alpha+\beta)$-family, $\mathcal{V}$ cannot be γ-mate of $\mathcal{U}$ as otherwise it would be an $(\alpha+\beta+\gamma)$-family. Thus, the size of the $\mathcal{U}$ family is $f \leq (t+1)|u_1|$.

Case when $\mathcal{V}$ is a δ-mate of $\mathcal{U}$. Then $T(\mathcal{U},\mathcal{V}) \geq f$, $T(\mathcal{U},\mathcal{V}) \geq (e_1+e_2-1)|u_1|+|u_2|$ and so $\frac{1}{2}|G|+\frac{1}{3}|T| > \frac{1}{2}f+\frac{e_1+e_2-1}{3}|u_1| > \frac{1}{2}f+\frac{e_1-e_2}{3}|u_1|+\frac{2e_2-1}{3}|u_1| \geq \frac{1}{2}f+\frac{2t}{3}|u_1|+\frac{1}{3}|u_1| > \frac{1}{2}f+\frac{2t+1}{3}|u_1| > \frac{1}{2}f+\frac{t+1}{2}|u_1| \geq \frac{1}{2}f+\frac{1}{2}f = f$. Thus, by Lemma 4.2, $\Delta(x) \leq \frac{1}{2}|x| - \frac{1}{3}|u|$.

Case when $\mathcal{V}$ is an ε-mate of $\mathcal{U}$. If there were no super-ε-mates of $\mathcal{U}$, then by Lemma 4.7, $\Delta(x) \leq \frac{5}{6}|x| - \frac{1}{3}|u|$. Thus, assume there is a super-ε-mate, and let $\mathcal{V}$ be the first super-ε-mate of $\mathcal{U}$. Between the first ε-mate of $\mathcal{U}$ and $\mathcal{V}$ there are at most $|u_1|$ FS-double squares, $\Delta(x) \leq \Delta(x') + (t+2)|u_1|$. By the assumption of this lemma, $\Delta(x') \leq \frac{1}{2}|x'| - \frac{1}{3}|v|$. By Lemma 3.10, there are two subcases.

Subcase (a). $G(\mathcal{U},\mathcal{V}) \geq (2e_1+e_2-3)|u_1|+2|u_2|$ and $T(\mathcal{U},\mathcal{V}) \geq (e_1+e_2-3)|u_1|+|u_2|$. Since $e_2 \geq 1$ and $t \geq 2$, then $\frac{1}{2}|G|+\frac{1}{3}|T| > \frac{2e_1+e_2-3}{2}|u_1| + \frac{e_1+e_2-2}{3}|u_1| = \frac{8e_1+5e_2-13}{6}|u_1| = \frac{8e_1-8e_2}{6}|u_1| + \frac{13e_2-13}{6}|u_1| > \frac{16t}{6}|u_1| = t|u_1| + \frac{10t}{6}|u_1| \geq t|u_1| + \frac{20}{6}|u_1| \geq t|u_1|+2|u_1|$ as $t \geq 2$.

Subcase (b). $G(\mathcal{U},\mathcal{V}) \geq e_1|u_1|+|u_2|$ and $T(\mathcal{U},\mathcal{V}) \geq (e_1+e_2-1)|u_1| + |u_2|$. Then $\frac{1}{2}|G|+\frac{1}{3}|T| > \frac{e_1}{2}|u_1| + \frac{e_1+e_2-1}{3}|u_1| = \frac{5e_1+2e_2-2}{6}|u_1| = \frac{5e_1-5e_2}{6}|u_1| + \frac{7e_2-2}{6}|u_1| \geq \frac{10t}{6}|u_1| + \frac{5}{6}|u_1| = t|u_1| + \frac{4t}{6}|u_1| + \frac{5}{6}|u_1| \geq t|u_1| + \frac{8}{6}|u_1| + \frac{5}{6}|u_1| = t|u_1| + \frac{13}{6}|u_1| > t|u_1| + 2|u_1|$ as $t \geq 2$. □

4.1.3 $(\alpha+\beta+\gamma)$-families

We consider the last case where the $\mathcal{U}$ family is composed of α, β and γ-mates. Figure 4.4 illustrates an $(\alpha+\beta+\gamma)$-family. Estimating the size of an $(\alpha+\beta+\gamma)$-family is challenging. Since there must be some β-mates of $\mathcal{U}$, $e_1 \geq e_2 + 2$. The family consists of segments. We list a few observations.

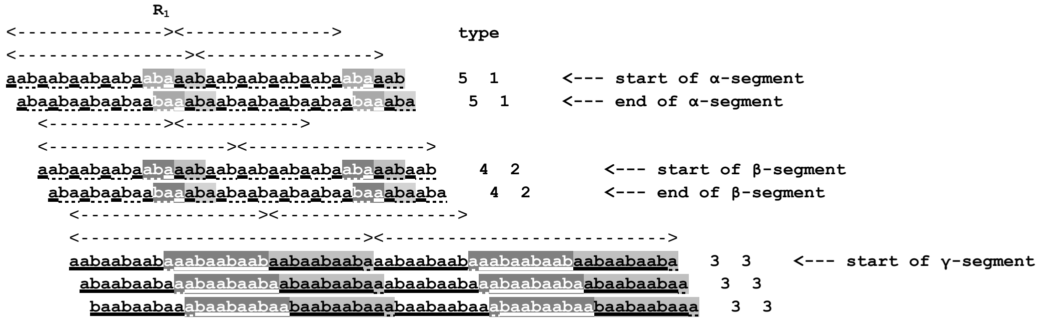

R₁
type
aabaabaabaabaabaaabaabaabaabaabaaabaaab 5 1 <--- start of α-segment
abaabaabaabaabaaabaabaabaabaabaaabaaaba 5 1 <--- end of α-segment
aabaabaabaabaabaaabaabaabaabaabaaabaaabaab 4 2 <--- start of β-segment
abaabaabaabaabaaabaabaabaabaabaaabaaabaaba 4 2 <--- end of β-segment
aabaabaabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaaba 3 3 <--- start of γ-segment
abaabaabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaabaa 3 3
baabaabaaabaabaabaabaabaabaaabaabaabaaabaabaabaabaabaabaaa 3 3

Figure 4.4: An example of an $(\alpha+\beta+\gamma)$-family

- The first so-called α-segment consists of $\mathcal{U}$ and possibly its right cyclic shifts, i.e. its α-mates. The size of the segment is at most $lcp(u_1,\overline{u}_1) \leq |u_1| - 2$ by Lemma 4.4. All the FS-double squares in this segments have the first exponent equal to e_1 and the second exponent equal to e_2, thus we say that the type of the segment is (e_1, e_2).

- There must be a β-mate of $\mathcal{U}$ and possibly its right cyclic shifts. All the FS-double squares in the segment have the first exponent equal to $e_1 - i_1$ and the second exponent equal to $e_2 + i_1$ for some $1 \leq i_1 \leq \frac{e_1 - e_2}{2}$, thus we say that the type of the segment is $(e_1 - i_1, e_2 + i_1)$. This so-called

β-segment has size $\leq lcp(u_1,\overline{u}_1)+lcs(u_1,\overline{u}_1) \leq |u_1|-2$ if $e_1-i_1 > e_2+i_1$ if $e_1-i_1 > e_2+i_1$ by Lemma 4.4, or $\leq |u_2|-1 \leq |u_1|-2$ if $e_1-i_1 = e_2+i_1$. Hence the β-segment has size $\leq |u_1|-2$.

- There may be another β-segment of type (e_1-i_2, e_2+i_2) for some $1 \leq i_1 < i_2 \leq \lfloor\frac{e_1-e_2}{2}\rfloor$, etc. There may be t such β-segments where $2t \leq e_1-e_2$. Let the last β-segment has type (p,q); then $p \geq q$.

- There must be $\mathcal{G}$, a γ-mate of $\mathcal{U}$. Consider all the γ-mates of $\mathcal{U}$ of which $\mathcal{G}$ is the first one. They form what we call a γ-segment. Since all the FS-double squares in the γ-segment have the short square of the same length $|U^2|$ and since they have equal exponents by Lemma 3.5, by Lemma 3.12 they are all α-mates of $\mathcal{G}$. Thus, the γ-segment consists of a γ-mate of $\mathcal{U}$ and its right cyclic shifts. The shorter square of $\mathcal{G}$ is $[s_1{u_1}^i u_2 {u_1}^{(e_1+e_2-i-1)} s_2][s_1{u_1}^i u_2 {u_1}^{(e_1+e_2-i-1)} s_2]$ for some $1 \leq i \leq p$ and some s_1 and s_2 such that $s_2s_1 = u_1$. In order to estimate the size of the γ-segment, we have to estimate how many right cyclic shifts $\mathcal{G}$ can have. It is possible to have a double square in a string that is not a FS-double square as there might be farther occurrences of the short or the long square of the double square. Thus, we overestimate the sizes of families, as we count the double squares and up to $|u_1|$ cyclic shifts for each α-segment or β-segment. Note that every segment can have at most $lcs(u_1,\overline{u}_1)+lcp(u_1,\overline{u}_1) \leq |u_1|-2$ members. Thus, we can set every segment to have a *hole*. So if there is a farther FS-double square that can be assigned to the hole, we will say that it *complements* the segment and thus does not need to be counted as its count was already part of the overestimation. If there is a farther FS-double square $\mathcal{V}$ containing a farther copy of ${u_1}^r u_2 {u_1}^r u_2$ and thus implying that though there is a double square of type (r, r'), it is not a FS-double square, we will say that $\mathcal{V}$ *replaces* the double square structure of type (r, r').

Observation 4.9. *The size of an* $(\alpha+\beta+\gamma)$*-family satisfies* $f \leq \frac{2}{3}(e_1+1)|u_1|$ *if* $e_1 \geq 4$*,* $f \leq \frac{2e_1}{3}|u_1|$ *otherwise.*

Proof. There are two cases:

Case when $\mathcal{G}$, the first member of the γ-segment, is of type $(e_1 - t, e_2 + t)$ **and** $e_1 - t > 2(e_2 + t)$, hence $e_1 \geq 4$ as $e_2 \geq 1, t \geq 1$. Since $e_1 - t > 2(e_2+t)$, $3t < e_1 - 2e_2$, $3t \leq e_1 - 2e_2 - 1$, and thus $6t \leq 2e_1 - 4e_2 - 2$. By Lemmas 3.5 and 4.4, $\mathcal{G}$ has $\leq (e_2+t) - 1$ cyclic shifts. Thus, we start with $\mathcal{U}$ of type (e_1, e_2) and end with the last member of the γ-segment that is of type $(e_1 - t - (e_2 + t - 1)), (e_2 + t + (e_2 + t - 1))$, thus there are at most $(2e_2 + 2t - 1) - e_2 + 1 = e_2 + 2t$ members in the $(\alpha+\beta+\gamma)$-family. Then $3f = 3e_2 + 6t \leq 3e_2 + 2e_1 - 4e_2 - 2 = 2e_1 - e_2 - 2 \leq 2e_1 - 3 < 2e_1 + 2 = 2(e_1+1)$ as $q \geq 1$. Thus, $f < \frac{2}{3}(e_1+1)|u_1|$.

Case when $\mathcal{G}$, the first member of the γ-segment, is of type $(e_1 - t, e_2 + t)$ **and** $e_1 - t \leq 2(e_2 + t)$. There are two subcases.

Subcase (a). $e_1 - t \le e_2 + t$. By Lemma 3.5, G^2 of $\mathcal{G}$ contains a further copy of ${u_1}^{e_2+t}u_2{u_1}^{e_2+t}u_2$ and so $\mathcal{G}$ either *replaces* a possible member of the α-segment or a β-segment, or it *complements* the α-segment or a β-segment. Thus, $f \le \frac{1}{2}(e_1 - e_2)|u_1| < \frac{2e_1}{3}|u_1|$.

Subcase (b). $e_1 - t > e_2 + t$, hence $e_1 \ge 4$ as $e_2 \ge 1, t \ge 1$. Either g_2 of $\mathcal{G}$ is small, i.e. $|g_2| < |u_1|$ and then $\mathcal{G}$ has less than $|u_1|$ shifts, and so $f \le \frac{1}{2}(e_1 - e_2)|u_1| + |u_1| \le \frac{2}{3}(e_1 + 1)|u_1|$, or $|g_2| \ge |u_1|$. Thus, assume that $|g_2| \ge |u_1|$. We can further assume by Lemma 3.5 that the last member of the γ-segment is of type $(e_2 + t, e_1 - t)$, since if it were shifted any further, it would start "replacing" or "completing" the members of the α-segment or the β-segments, so we do not need to count them. Since $e_1 - t \le 2(e_2 + t)$, then $e_1 - 2e_2 \ge 3t$. Thus, $3f = 3(e_1 - t - e_2 - 1)|u_1| = (3e_1 - 3t - 3e_2 + 3)|u_1| \le (3e_1 - 3e_2 + 3 + 2e_2 - e_1)|u_1| = (2e_1 - e_2 + 3)|u_1| \le (2e_1 + 2)|u_1| = 2(e_1 + 1)|u_1|$. Therefore, $f \le \frac{2}{3}(e_1 + 1)|u_1|$. □

As for α-families and $(\alpha+\beta)$-families, we must consider two cases. First in Observation 4.10 when the $(\alpha+\beta+\gamma)$-family is not followed by any FS-double square, and then in Observation 4.11 when it does.

Observation 4.10. *Let x be a string starting with an $(\alpha+\beta+\gamma)$-family of a FS-double square $\mathcal{U}$ and assume there is no other FS-double square, then $\Delta(x) \le \frac{5}{6}|x| - \frac{1}{3}|u|$.*

Proof. The size of the family satisfies $f \le \frac{2}{3}(e_1 + 1)|u_1|$. Thus, $|x| \ge f + |U^2| = f + 2(e_1 + e_2)|u_1| + 2|u_2|$, and so $\frac{5}{6}|x| - \frac{1}{3}|u| \ge \frac{5}{6}f + \frac{5}{6}(2(e_1 + e_2)|u_1| +$

$\frac{5}{6}2|u_2| - \frac{1}{3}e_1|u_1| - \frac{1}{3}|u_2| = \frac{5}{6}f + \frac{8}{6}p|u_1| + \frac{10}{6}e_2|u_1| + \frac{3}{6}|u_2| > \frac{5}{6}f + \frac{30}{18}p|u_1| > \frac{5}{6}f + \frac{2}{18}p|u_1| + \frac{28}{18}p|u_1| \geq \frac{5}{6}f + \frac{2}{18}(p+1)|u_1| \geq \frac{5}{6}f + \frac{1}{6}f = f = \Delta(x)$. □

Observation 4.11. *Let x be a string starting with an $(\alpha+\beta+\gamma)$-family of a FS-double square $\mathcal{U}$, $\mathcal{V}$ the first FS-double square not in the $\mathcal{U}$ family, and x' the suffix of x starting at the same position as $\mathcal{V}$, if $\delta(x') \leq \frac{5}{6}|x'| - \frac{1}{3}|v|$, then $\delta(x) \leq \frac{5}{6}|x| - \frac{1}{3}|u|$.*

Proof. Recall that $s(w)$ and $e(w)$ indicates respectively the starting and end point of the factor w. $\mathcal{V}$ can be either a δ-mate or ε-mate of $\mathcal{U}$. Let $\mathcal{G} = (g_1, g_2, f_1, f_2)$ be the last member of the γ-segment and its type be $(f_1, f_2) = (e_1 - t, e_2 + t)$. Then $g^2 = {u_1}^t s_1[s_2 {u_1}^{(e_1-t-1)} u_2 u_1{}_2^e s_1][s_2 {u_1}^{(e_1-t-1)} u_2 u_1{}_2^e s_1]$. If $e(v_{[1]} \leq e(g_{[1]})$, then $\mathcal{U}$ would be a β-mate of $\mathcal{G}$ by Lemma 3.12 – which is impossible as $f_1 = f_2$ by Lemma 3.5 Thus $e(v_{[1]}) > e(g_{[1]})$.

Case when $\mathcal{V}$ is a δ-mate. Then $T(\mathcal{U},\mathcal{V}) \geq (e_1 + e_2 - 1)|u_1|$. In fact, $v_{[1]}$ contains an inversion factor from $\big[L_1(\mathcal{U}), R_1(\mathcal{U})\big]$. If $s(v_{[2]}) \leq R_2(\mathcal{U})$, then $v_{[2]}$ would contain an inversion factor from $\big[L_2(\mathcal{U}), R_2(\mathcal{U})\big]$ at the same position, giving $|v| = |U|$, a contradiction. Hence $s(v_{[2]}) > R_2(\mathcal{U})$ and by Synchronization Principle Lemma 1.6, $T(\mathcal{U},\mathcal{V}) \geq (e_1 + e_2)|u_1|$. Since $G(\mathcal{U},\mathcal{V}) \geq f$, we have $\frac{1}{2}|G| + \frac{1}{3}|T| \geq \frac{1}{2}f + \frac{1}{3}(e_1 + e_2)|u_1| \geq \frac{1}{2}f + \frac{1}{3}(e_1 + 1)|u_1| \geq \frac{1}{2}f + \frac{1}{2}f = f$ as $e_2 \geq 1$ and $\frac{1}{2}f \leq \frac{1}{3}(e_1 + 1)|u_1|$.

Case when $\mathcal{V}$ is an ε-mate of $\mathcal{U}$, but not a super-ε-mate. Then $s(v_{[1]}) \leq$

$e(u_{[1]})$ and $e(v_{[1]}) > e(g_{[1]})$. By Synchronization Principle Lemma 1.6, $T(\mathcal{U},\mathcal{V}) \geq (e_1+e_2)|u_1|$ and so $\frac{1}{2}|G|+\frac{1}{3}|T| \geq \frac{1}{2}f+\frac{1}{3}(e_1+1)|u_1| \geq \frac{1}{2}f+\frac{1}{2}f = f$.

Case when $\mathcal{V}$ is a super-ε-mate of $\mathcal{U}$. By Lemma 3.10, there are two subcases:

Subcase (a). $G \geq (2e_1+e_2-3)|u_1|$ and $T \geq (e_1+e_2-2)|u_1|$. Then $\frac{1}{2}|G|+\frac{1}{3}|T| \geq \frac{6e_1+3e_2-9+2e_1+2e_2-4}{6}|u_1| = \frac{8e_1+5e_2-13}{6}|u_1| = \frac{4e_1+4e_1+5e_2-13}{6}|u_1|$. Since $e_1 \geq 3$ and $e_2 \geq 1$, $\frac{1}{2}|G|+\frac{1}{3}|T| \geq \frac{4e_1+12+5-13}{6}|u_1| = \frac{4e_1+4}{6}|u_1| \geq f$.
Subcase (b). $G \geq e_1|u_1|$ and $T \geq (e_1+e_2-1)|u_1|$, then $\frac{1}{2}|G|+\frac{1}{3}|T| \geq \frac{3e_1+2e_1+2e_2-2}{6}|u_1| = \frac{4e_1+e_1+2e_2-2}{6}|u_1| \geq \frac{4e_1+3}{6}|u_1| = \frac{2e_1+12}{6}|u_1| > \frac{2e_1+2}{16}|u_1| \geq f$, since $e_1 \geq 4$ and $e_2 \geq 1$. □

4.2 At most $\lfloor 11n/6 \rfloor$ distinct squares

Chapter 2 discusses double squares and their periodic properties via the canonical factorization. Chapter 3 analyzes the mate relations for FS-double squares. Chapter 4 discusses how to handle α-families, (α+β)-family, and (α+β+γ)-families and how to estimate their sizes. Combining these ingredients, allows to prove the main result by induction.

Theorem 4.12. *The number of FS-double squares in a string of length n is at most $\lfloor 5n/6 \rfloor$.*

Proof. We prove by induction a slightly stronger statement: $\Delta(x) \leq \frac{5}{6}|x| -$

$\frac{1}{3}|u|$ for $|x| \geq 10$ where u is the root of the short square of the first, i.e. leftmost, FS-double square of x. A string of length at most 9 contains no FS-double square, and string of length 10 contains at most one FS-double square. Consider a FS-double square $\mathcal{U} := (u_1, u_2, e_1, e_2)$, then $|UU| = 2(e_1 + e_2)|u_1| + 2e_2|u_2| \geq 2 \cdot 2 \cdot 2 + 2 \cdot 1 \cdot 1 = 10$. Thus, the statement holds if $|x| \leq 10$. Assuming the statement is true for all $|x| \leq n$, we shall prove it holds for all $|x| \leq n{+}1$. If $x = x[1, \ldots, n{+}1]$ does not start with a FS-double square, then $\Delta(x) = \Delta(x[2..n{+}1]) \leq \frac{5}{6}|x[2..n{+}1]| - \frac{1}{3}|u| \leq \frac{5}{6}|x[1..n{+}1]| - \frac{1}{3}|u|$. Thus, we can assume that x starts with a FS-double square $\mathcal{U}$. If $\mathcal{U}$ is the only FS-double square of x, then $|x| \geq 2|u|$, thus the statement holds. Therefore, we can assume that x starts with a FS-double square $\mathcal{U}$ and $\Delta(x) \geq 2$.

Case when x starts with an α-family of $\mathcal{U}$. If there is no further FS-double square in x, the statement holds by Observation 4.5. Otherwise, we carry out the induction step by Observation 4.6.

Case whent x starts with an $(\alpha{+}\beta)$-family of $\mathcal{U}$. If there is no further FS-double square in x, the statement holds by Observation 4.7. Otherwise, we carry out the induction step by Observation 4.8.

Case when x starts with an $(\alpha{+}\beta{+}\gamma)$-family of $\mathcal{U}$. If there is no further FS-double square in x, the statement holds by Observation 4.10. Otherwise, we carry out the induction step by Observation 4.11. □

Corollary 4.13. *The number of distinct squares in a string of length n is at most $\lfloor 11n/6 \rfloor$.*

Proof. The number of distinct squares in a string is the sum of the number of FS-double squares plus the number of single rightmost squares. Since, for a string of length n, the number of FS-double squares is at most $\lfloor 5n/6 \rfloor$, the number of distinct squares is at most $\lfloor (2 \cdot 5/6 + 1/6)n \rfloor = \lfloor 11n/6 \rfloor$. □

Chapter 5

Conclusion and further work

We presented the proof that the number of distinct squares in a string of length n at most $\lfloor 11n/6 \rfloor$, and thus strengthened the universal bound of $2n$ established by Fraenkel and Simpson in 1998. The combinatorial structure yielding this strengthening might help investigating and generalizing results such as the Three Squares Lemma and the New Periodicity Lemma.

The key ingredients of the proof include the canonical factorization of double squares, and introducing and exploiting a combinatorial structure of double squares called *inversion factor*. The existence of the inversion factors and their unique occurrences yield a taxonomy of mutual relations of FS-double squares, called *mate relations*. The obtained taxonomy contains exactly 5 types.

A refinement of the inversion factor is presented in Chapter 2. The proof presented in the thesis could be carried using this refinement. Future work and research directions include to analyze and exploit how the refinement could help to further streamline the taxonomy and possible further strengthen the upper bound for the number of distinct squares.

Bibliography

[1] H. Bai, A. Deza, and F. Franek, *On a lemma of Crochemore and Rytter*, Journal of Discrete Algorithms, 34 (2015), pp. 18–22.

[2] H. Bai, F. Franek, and W. F. Smyth, *Two squares canonical factorization*, in Proceedings of the Prague Stringology Conference 2014, Czech Technical University in Prague, Czech Republic, 2014.

[3] ——, *The new periodicity lemma revisited*, Journal of Discrete Algorithms, (2015). to appear.

[4] J. Berstel, *Axel thue's papers on repetitions in words: translation*, unpublished, (1994).

[5] W. Bland and W. F. Smyth, *Three overlapping squares: the general case characterized & applications*, Theoretical Computer Science, 596-6 (2015), pp. 23–40.

[6] M. Crochemore and W. Rytter, *Squares, cubes, and time-space efficient string searching*, Algorithmica, 13 (1995), pp. 405–425.

[7] A. Deza, F. Franek, and M. Jiang, *A d-step approach for distinct squares in strings*, in Combinatorial Pattern Matching, Springer, 2011, pp. 77–89.

[8] A. Deza, F. Franek, and M. Jiang, *A computational framework for determining square-maximal strings*, in Proceedings of the Prague Stringology Conference 2012, J. Holub and J. Žďárek, eds., Czech Technical University in Prague, Czech Republic, 2012, pp. 111–119.

[9] A. Deza, F. Franek, and A. Thierry, *How many double squares can a string contain ?*, Discrete Applied Mathematics, 180 (2015), pp. 52–69.

[10] N. J. Fine and H. S. Wilf, *Uniqueness theorems for periodic functions*, Proceedings of the American Mathematical Society, 16 (1965), pp. 109–114.

[11] A. S. Fraenkel and J. Simpson, *How many squares can a string contain?*, Journal of Combinatorial Theory, Series A, 82 (1998), pp. 112–120.

[12] F. Franek, R. C. Fuller, J. Simpson, and W. F. Smyth, *More results on overlapping squares*, Journal of Discrete Algorithms, 17 (2012), pp. 2–8.

[13] Z. Galil and J. I. Seiferas, *Time-space-optimal string matching*, Journal of Computer and System Sciences, 26 (1983), pp. 280–294.

[14] L. Ilie, *A simple proof that a word of length n has at most $2n$ distinct squares*, Journal of Combinatorial Theory, Series A, 112 (2005), pp. 163 – 164.

[15] ——, *A note on the number of squares in a word*, Theoretical Computer Science, 380 (2007), pp. 373–376.

[16] F. Kangmin, S. J. Puglisi, W. F. Smyth, and A. Turpin, *A new periodicity lemma*, SIAM Journal on Discrete Mathematics, 20 (2006), pp. 656–668.

[17] E. Kopylova and W. F. Smyth, *The three squares lemma revisited*, Journal of Discrete Algorithms, 11 (2012), pp. 3–14.

[18] N. H. Lam, *On the number of squares in a string*, AdvOL-Report, 2013/2, McMaster University, 2013.

[19] M. Lothaire, ed., *Combinatorics on Words*, Cambridge University Press, second ed., 1997. Cambridge Books Online.

[20] J. Simpson, *Intersecting periodic words*, Theoretical Computer Science, 374 (2007), pp. 58–65.

[21] W. F. Smyth, *Computing Patterns in Strings*, ACM Press Bks, Pearson/Addison-Wesley, 2003.

[22] A. Thierry, *Combinatorics of the interrupted period*, in Prague Stringology Conference 2015, 2015, p. 17.

[23] A. Thue, *Über unendliche Zeichenreihen*, Skrifter udgivne af Videnskabsselskabet i Christiania, I. Math.-naturv. Klasse, 7, 1906.

[24] ——, *Über die gegenseitige Lage gleicher Teile gewisser Zeichenreihen*, Skrifter udgivne af Videnskabsselskabet i Christiania, I. Math.-naturv. Klasse, 1, 1912.