STRUCTURAL FACTORIZATION OF SQUARES IN STRINGS

Haoyue Bai

STRUCTURAL FACTORIZATION OF SQUARES IN STRINGS

By Haoyue BAI

A Thesis Proposal Submitted to the Department of Computing and Software of McMaster University in Partial Fulfilment of the Requierement for the Degree of Doctor of Philosophy

McMaster University © Copyright by Haoyue Bai, May 2017

Ph.D. Thesis — McMaster University

Department Computing and Software — Hamilton, Ontario, Canada

TITLE:	Structural Factorization of Squares in Strings
AUTHOR:	Haoye Bai
	M.Eng. McMaster University
	Bachelor. Beijing University of Post and Telecommunications
SUPERVISOR:	Dr. Antoine Deza & Dr. Frantisek Franek
NUMBER OF PAGES:	V,73,III

Abstract

A balanced double square in a string $\boldsymbol{x}$ consists of two squares starting in the same position and of comparable lengths. We present a unique factorization of the longer square into primitive components refereed to as the canonical factorization and analyze its properties. In particular, we examine the inversion factors and the right and left inversion subfactors. All three substrings are collectively referred to as rare factors as they occur only twice in a significant portion of the larger square. The inversion factors were essential for determining the classification of mutual configurations of double squares and thus providing the best-to-date upper bound of $11n/6$ for the number of distinct squares in a string of length n in [8] by Deza, Franek, and Thierry. The right and left inversion subfactors have the advantage of being half the length of the inversion factors, thus providing a stronger discrimination property for a possible third square. This part of the thesis was published in [2] by Bai, Franek, and Smyth.

The canonical factorization and the right and left inversion subfactors are used to formulate and prove a significantly stronger version of the New Peri-

odicity Lemma [9] by Fan, Puglisi, Smyth, and Turpin, 2006, that basically restricts what kind of a third square can exists in a balanced double square. This part of the thesis was published in [3] by Bai, Franek, and Smyth.

The canonical factorization and the inversion factors are applied to formulate and prove a stronger version of the Three Squares Lemma [7] by Crochemore and Rytter. This part of the thesis was published in [1] by Bai, Deza, and Franek.

Acknowledgments

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.

My special thanks go to the members of the supervisory and defence committees: Dr. Antoine Deza, Dr. Frantisek Franek, Dr. Kai Huang, Dr. Fred Hoppe, Dr. Dalibor Froncek.

I appreciated the help and moral support from all my colleagues including the members of the Advanced Optimization Laboratory and the members of the school of Computational Science and Engineering.

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

To my family and friends.

Table of Contents

List of Definition

Symbol	Meaning	Page
$\boldsymbol{x}$	a *string* $\boldsymbol{x}$ over an alphabet Σ	8
$\boldsymbol{\varepsilon}$	empty string	8
$\boldsymbol{x}[1..n]$	a finite string $\boldsymbol{x}$ with $\boldsymbol{n}$ symbols	8
$\lvert\boldsymbol{x}\rvert$	the length of string $\boldsymbol{x}$	8
$\boldsymbol{xy}$	concatenation of two strings $\boldsymbol{x}$ and $\boldsymbol{y}$	9
$\prec$	an *order* of an alphabet Σ	10
$\boldsymbol{u}^k$	concatenation of k copies of $\boldsymbol{u}$	11
$\boldsymbol{u}_{[i]}$	denotes the i-th occurrence of $\boldsymbol{u}$ in $\boldsymbol{u}^k$, where $1 \leq i \leq k$	11
$R_k(\boldsymbol{x})$	is a *right k-rotation* of a string $\boldsymbol{x}$	16
$\boldsymbol{u} \sim \boldsymbol{v}$	strings $\boldsymbol{u}$ and $\boldsymbol{v}$ are *conjugates*	17
$\mathrm{lcp}(\boldsymbol{u}, \boldsymbol{v})$	the length of the *longest common prefix*	18
$\mathrm{lcs}(\boldsymbol{u}, \boldsymbol{v})$	the length of the *longest common suffix*	18
$DS(\boldsymbol{u}, \boldsymbol{v})$	a double square that consist of two squares $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$ start at the same position and that $\lvert\boldsymbol{u}\rvert < \lvert\boldsymbol{v}\rvert$.	27
$DS(\boldsymbol{u},\ \boldsymbol{v}\ :$	$\boldsymbol{u}_1, \boldsymbol{u}_2, e_1, e_2)$ the *canonical factorization* of the double square $DS(\boldsymbol{u}, \boldsymbol{v})$	28
$\mathbb{IF}$	inversion factor	37
$\mathbb{IF}_1$	the first occurrence of $\mathbb{IF}$ in a double square	37
$\mathbb{IF}_2$	the second occurrence of $\mathbb{IF}$ in a double square	37
$\mathbb{RIS}$	the right inversion subfactor	41
$\mathbb{LIS}$	the left inversion subfactor	41
$\boldsymbol{R}_1$	the maximal right cyclic shift of the left half of $\mathbb{IF}_1$	45
$\boldsymbol{R}_2$	the maximal right cyclic shift of the left half of $\mathbb{IF}_2$	45
$\boldsymbol{L}_1$	the maximal left cyclic shift of the right half of $\mathbb{IF}_1$	45
$\boldsymbol{L}_2$	the maximal left cyclic shift of the right half of $\mathbb{IF}_2$	45

Chapter 1

Introduction

1.1 Background of the research

The structural properties of several squares starting in the same proximity have been studied by researchers interested in two major problems concerning periodicities in strings: the *maximum number of runs* and the *maximum number of distinct squares* in a string. In 1995, Crochemore and Rytter [7] described the relationship among the lengths of three distinct squares starting at the same position in a string in the *Three Squares Lemma*:

> *Let* $\boldsymbol{u}^2 \neq \boldsymbol{v}^2$ *be proper prefixes of* $\boldsymbol{w}^2$ *and let* $\boldsymbol{u}$, $\boldsymbol{v}$, *and* $\boldsymbol{w}$ *be primitive, then* $|\boldsymbol{u}|+|\boldsymbol{v}| < |\boldsymbol{w}|$.

In [10], Fraenkel and Simpson observed that in fact a slightly different version was proven as only the primitivness of $\boldsymbol{u}$ was actually used in the proof, they also gave a counterexample for the sharp inequality: $\boldsymbol{u} = aba$, $\boldsymbol{v} = abaab$

and $\boldsymbol{w} = abaababa$. Thus, they used the lemma in this form:

> *Let* $\boldsymbol{u}^2$ *and* $\boldsymbol{v}^2$ *be prefixes of* $\boldsymbol{w}^2$ *so that* $\boldsymbol{u}$ *is primitive and* $|\boldsymbol{u}| < |\boldsymbol{v}| < |\boldsymbol{w}|$. *Then* $|\boldsymbol{u}| + |\boldsymbol{v}| \leq |\boldsymbol{w}|$.

and used it for obtaining a bound of $2|\boldsymbol{x}|$ for the number of distinct squares in a string $\boldsymbol{x}$. We will discuss the Three Squares Lemma in more details in Chapter 4.

In 2006, Fan *et al.* [9] considered a case of two squares starting at the same position with a third square possibly offset some distance to the right; they presented the *New Periodicity Lemma* describing conditions under which the third square could not exist. Since that time there has been considerable work done [4, 11, 14, 16] in an effort to specify more precisely the combinatorial structure of a string in the neighbourhood of such two squares. We will discuss the New Periodicity Lemma in more details in Chapter 3.

In [8], the canonical factorization of two rightmost squares starting at the same position, which is referred there as an FS-double square, proved essential for improving the Frankel-Simpson upper bound for the maximum number of distinct squares in a string $\boldsymbol{x}$ to $\frac{11}{6}|\boldsymbol{x}|$.

This was the motivation of the research described here and it lead to three publications that anchor this thesis: *Two squares canonical factorization* [2], *The New Periodicity Lemma Revisited* [3], and *On a lemma of Crochemore and Rytter* [1].

In the first of the publications, *Two squares canonical factorization* [2], the canonical factorization for all balanced double squares is analyzed. In

a sense, it was a surprising result as no strong assumptions such as primitiveness of the squares or being the rightmost occurrences, are needed. This result is described in Chapter 2.

The realization that the inversion factors of the canonical factorization are such prominent structures as to completely remove the need for using the Three Squares Lemma for improving the upper bound for the maximum number of distinct squares in [8], lead to a renewed interest in the structure of three squares starting at the same proximity.

The New Periodicity Lemma [9] has a complicated proof as the combinatorial structure of three squares two of which start at the same position is quite forbidding. However, since when the two squares starting at the same position form a balanced double square can be dealt with via the inversion factors imposed by the canonical factorization, we embarked on the research. It required a refinement of the notion of the inversion factor, but lead to a new theorem with significantly reduced assumptions about the squares: originally, the smallest square was required to be regular – a very stringent requirement, and the bigger square was required to be primitive. This research and results lead to the publication *The New Periodicity Lemma Revisited* [3] and it is described in Chapter 3.

The canonical factorization also lead to a more straightforward and streamlined proof of the Three Squares Lemma with weaker assumptions as it deals simply with the cases when two smallest or the two biggest squares form a balanced double square. This research and results lead to the publication

On a lemma of Crochemore and Rytter [1] and are described in Chapter 4.

The thesis is concluded with Chapter 5 containing a conclusion and ideas for future research.

1.2 Preliminaries

Definition 1.1. *An* alphabet Σ *is a non-empty set of elements referred to as* symbols *or* letters*. Frequently, it is required to be finite, but it can also be infinite. A* string $\boldsymbol{x}$ *over an alphabet* Σ *is a contiguous sequence of symbols drawn from* Σ*. A string without any symbol is called the* empty string *and denoted by* $\boldsymbol{\varepsilon}$*. A string is also often called a* word*.*

Definition 1.2. *The* alphabet *of a string is the set of all of the symbols that occur in that string.*

string	Example	Non-example
$aabba$	$\{a, b\}$	$\{a, c\}$
12445	$\{1, 2, 4, 5\}$	$\{3\}$
$abcca$	$\{a, b, c\}$	$\{a, b\}$
ab^+	$\{a, b\}$	$\{a\}$
123323	$\{1, 2, 3\}$	$\{1, 2, 3, 4, ...\}$

Table 1.1: Examples and Non-examples of alphabet of a string

Definition 1.3. *The* length *of a finite string* $\boldsymbol{x}$ *is the number of symbols it is composed of, denoted as* $|\boldsymbol{x}|$*. The empty string has length zero. A finite string* $\boldsymbol{x}$ *with* $\boldsymbol{n}$ *symbols can be represented as an array* $\boldsymbol{x}[1..n]$ *or as a sequence* $\boldsymbol{x}[1]\boldsymbol{x}[2]...\boldsymbol{x}[n]$*.*

one-sided infinite string	bi-infinite string
has a first element, no final element	neither a first nor a final element
aab...	*...aab...*
...bbbb	*...bbbbb...*

Table 1.2: Examples of infinite strings

Definition 1.4. Concatenation *is a basic operation of strings that joins two strings* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *together into one, which is denoted in the order that the two strings are concatenated as* $\boldsymbol{xy}$*. For instance,* $\boldsymbol{xy}$ *represents a string that is the concatenation of string* $\boldsymbol{x}$ *followed by string* $\boldsymbol{y}$*; that is,* $\boldsymbol{xy} = \boldsymbol{x}[1][2]...\boldsymbol{x}[i]\boldsymbol{y}[1]\boldsymbol{y}[2]...\boldsymbol{y}[j]$*, where* $\boldsymbol{x} = \boldsymbol{x}[1...i]$ *and* $\boldsymbol{y} = \boldsymbol{y}[1...j]$ *and* $i,\ j \geq 1$*.*

Definition 1.5. *If a string* $\boldsymbol{x} = \boldsymbol{uvw}$*, where* $\boldsymbol{u}$*,* $\boldsymbol{v}$*, and* $\boldsymbol{w}$ *are strings, then* $\boldsymbol{u}$ *(respectively,* $\boldsymbol{v}$*,* $\boldsymbol{w}$*) is said to be a* prefix (*respectively,* substring*,* suffix) *of* $\boldsymbol{x}$ *if* $0 \leq |\boldsymbol{u}| \leq |\boldsymbol{x}|$ (*respectively,* $0 \leq |\boldsymbol{v}| \leq |\boldsymbol{x}|$*,* $0 \leq |\boldsymbol{w}| \leq |\boldsymbol{x}|$*);*
a proper prefix (*respectively,* proper substring*,* proper suffix) *if* $0 < |\boldsymbol{u}| < |\boldsymbol{x}|$ (*respectively,* $0 < |\boldsymbol{v}| < |\boldsymbol{x}|$*,* $0 < |\boldsymbol{w}| < |\boldsymbol{x}|$*). A substring is also called a* factor *or* subword*.*

string	Example	Non-example
aabba	*a, aa, aab, aabb, ab, abb, abba, b, bb, bba, ba*	ε
12445	1, 12, 124, 1244, 2, 24, 244, 2445, 4, 44, 445, 45, 5	12445
abcca	*a, ab, abc, abcc, b, bc, bcc, bcca, c, cc, cca, ca*	*ccc*
abbb	*a, ab, abb, b, bb, bbb*	*aba*
323	3, 32, 2, 23	3234

Table 1.3: Examples and Non-examples of proper substrings of a string

Definition 1.6. *The symbol* Σ^* *denotes the set of all finite strings over the alphabet* Σ*, including the empty string* ε*.*

Definition 1.7. *An* order (*or* ordering) $\prec$ *of an alphabet* Σ *is a transitive, reflexive, and antisymmetric relation* $\prec$ *over the alphabet that is total, i.e. that for every pair of distinct symbols* $x, y \in \Sigma$*, either* $x \prec y$ *or* $y \prec x$*. Thus, the alphabet can be diagrammed as a line of symbols.*

Definition 1.8. *The* lexicographic ordering $\prec$ *is an ordering of* Σ^* *induced by the ordering* $\prec$ *of the alphabet* Σ*: For every pair of strings* $x = x_0x_1...x_{k-1}$ *and* $y = y_0y_1...y_{l-1}$ *in* Σ^**,* $x \prec y$ *if there is an index* $i \geq 0$ *such that* $i < k$ *and* $i < l$ *and* $x_i \prec y_i$ *and* $x_j = y_j$ *for all* $0 \leq j < i$ *; or if* $l > k$ *and for all* $0 \leq j < k$ *we have* $x_j = y_j$*.*

order	lexicographic order
$a \prec b \prec c$	$ab \prec abb \prec abbc \prec abc \prec acbc \prec cbc$
$a \prec c \prec b$	$acbc \prec ab \prec abc \prec abb \prec abbc \prec cbc$
$b \prec a \prec c$	$ab \prec abb \prec abbc \prec abc \prec acbc \prec cbc$
$b \prec c \prec a$	$cbc \prec ab \prec abb \prec abbc \prec abc \prec acbc$
$c \prec a \prec b$	$cbc \prec acbc \prec ab \prec abc \prec abb \prec abbc$
$c \prec b \prec a$	$cbc \prec acbc \prec ab \prec abc \prec abb \prec abbc$

Table 1.4: lexicographic order of strings *ab*, *abb*, *abc*, *cbc abbc*, *acbc*, over alphabet $\{a, b, c\}$

Observation 1.9. *For an alphabet of size* n*, there are* $C_n^1 \cdot C_{n-1}^1 \cdot C_{n-2}^1 \cdot C_{n-3}^1 \ ... \cdot 1 = n \cdot (n-1) \cdot (n-2) \cdot (n-3) \cdot \ ... \ \cdot 3 \cdot 2 \cdot 1 = n!$ *different orderings. I.e. there are as many different orderings of* Σ^* *as there are permutations of* $|\Sigma|$*.*

Definition 1.10. *A string* $\boldsymbol{x}$ *is* primitive *if there are no string* $\boldsymbol{u}$ *and an integer* $k \geq 2$ *so that* $\boldsymbol{x} = \boldsymbol{u}^k = \underbrace{\boldsymbol{u}\boldsymbol{u}\ldots\boldsymbol{u}}_{k \text{ times}}$ *. For every string* $\boldsymbol{x}$ *there is a unique smallest string* $\boldsymbol{u}$ *so that* $\boldsymbol{x} = \boldsymbol{u}^k$ *for some integer* $k \geq 1$*;* $\boldsymbol{u}$ *is necessarily primitive and is called the* primitive root *of* $\boldsymbol{x}$*.*

Definition 1.11. *For a concatenation* $\boldsymbol{u}^k$*,* $k \geq 2$*,* $\boldsymbol{u}_{[1]}$ *denotes the first occurrence of* $\boldsymbol{u}$ *in* $\boldsymbol{u}^k$*,* $\boldsymbol{u}_{[2]}$ *denotes the second occurrence of* $\boldsymbol{u}$ *in* $\boldsymbol{u}^k$*, ...,* $\boldsymbol{u}_{[k]}$ *denotes the k-th occurrence of* $\boldsymbol{u}$ *in* $\boldsymbol{u}^k$*.*

Example	non-Example
ab	*abab*
a	*aaaa*
aab	*aabaab*
abac	*acac*
aabca	*abaabaaba*

Table 1.5: Examples and Non-examples of primitive strings

Definition 1.12. *A* repeat *is a collection of identical substrings of* $\boldsymbol{x}[1..n]$ *described by* $\boldsymbol{u} = \boldsymbol{x}[i_1..i_1 + p - 1] = \boldsymbol{x}[i_2..i_2 + p - 1] = ... = \boldsymbol{x}[i_q..i_q + p - 1]$*, where* $1 \leq i_1 < i_2 < ... < i_q \leq n,\ q \geq 2$*, and* $p \geq 1$*.* $\boldsymbol{u}$ *is called the* generator (*or the* root) *of the repeat.*

Definition 1.13. *A* repetition *is a tandem repeat; that is, for every two consecutive identical substrings in the repeat, the gap between their starting positions is equal to the size of the generator. A repetition with generator* $\boldsymbol{u}$ *repeating* q *times* (q *an integer*) *can be presented as* $\boldsymbol{u}^q$*, where* $\boldsymbol{u}$ *is a non-empty string, and* $q \geq 2$*. The integer value* $|\boldsymbol{u}|$ *is called the* period *and* q *the* exponent *or* power *of the repetition.*

Example	Non-example
abab	*ab*
aaaa	*a*
abaaba	*abaabaa*
aabaab	*baabaab*
abcabcabc	*cabcabcabcab*

Table 1.6: Examples and Non-examples of repetitions

Definition 1.14. *A* square *is a repetition with power of* 2. *A* cube *is a repetition with power* 3.

Example	Non-example
abab	*ab*
aaaa	*a*
abaaba	*aabaaba*
aa	*aaa*
aabaaaba	*aabab*

Table 1.7: Examples and Non-examples of squares

Example	Non-example
ababab	*abab*
aaaaaa	*aa*
bbb	*abaaba*
abcabcabc	*abc*
aabaabaab	*aabab*

Table 1.8: Examples and Non-examples of cubes

Definition 1.15. *If* $\boldsymbol{u}$ *is primitive, then the square* $\boldsymbol{u}^2$ *is called* primitively rooted.

The square $\boldsymbol{u}^2$ *is* regular *if no prefix of* $\boldsymbol{u}$ *is a square.*

Example	Non-example
$(ab)ab$	$(abab)abab$
$(a)a$	$(aa)aa$
$(abc)abc$	$(abcabc)abcabc$
$(aba)aba$	$(aaa)aaa$
$(aab)aab$	$(bbbb)bbbb$

Table 1.9: Examples and Non-examples of primitively rooted squares – '()' shows the root of the square

Definition 1.16. *If a repetition* $\boldsymbol{u}^q = \boldsymbol{x}[s..s+qp-1]$ *where* $p = |\boldsymbol{u}|$*, can be extended by another copy of* $\boldsymbol{u}$ *to the left of* $\boldsymbol{x}$*, that is,* $\boldsymbol{x}[s-p..s-1] = \boldsymbol{u}$*, then we say the repetition can be extended to the left. A repetition is called a* left-maximal repetition *if it cannot be extended to the left. Similarly, if we cannot extend by another copy of* $\boldsymbol{u}$ *to the right, then the repetition is a* right-maximal repetition *. A* maximal repetition *refers to a repetition that can be extended neither to the left nor to the right.*

string	Example
$\boldsymbol{x} = cccababccc$	$ccc(abab)ccc$
$\boldsymbol{x} = ababaaaababa$	$abab(aaaa)baba$
$\boldsymbol{x} = bbabaababb$	$bb(abaaba)bb$
$\boldsymbol{x} = abaaabaabbab$	$aba(aabaab)bab$
$\boldsymbol{x} = abbabcabccba$	$abb(abcabc)cba$

Table 1.10: Examples of maximal repetitions in a string $\boldsymbol{x}$ – '()' shows the maximal repetition

Non-example
left extendable: $cccb(abab)cccc$
ight extendable: $abab(aaaa)abab$
two-side extendable : $aaa(abaaba)aaa$
right extendable: $aba(aabaab)aba$
two-side extendable : $cbc(abcabc)aba$

Table 1.11: Non-examples of maximal repetitions – '()' shows the repetition

Definition 1.17. *A* run *in a string* $\boldsymbol{x} = \boldsymbol{x}[1..n]$ *starting at a position* s*, ending at a position* e *with a period* p *is a substring* $\boldsymbol{x}[s..e]$ *so that*

1. $\boldsymbol{x}[s..e] = \boldsymbol{x}[s..s+p-1]^r\boldsymbol{t}$ *where the integer* $r \geq 2$ *and* $\boldsymbol{t}$ *called* tail *is a proper prefix of the root* $\boldsymbol{x}[s..s+p-1]$*;*
2. *the root* $\boldsymbol{x}[s..s+p-1]$ *is primitive;*
3. *the root* $\boldsymbol{x}[s..s+p-1]$ *cannot be shifted left, i.e. either* $s = 1$ *or* $\boldsymbol{x}[s-1] \neq \boldsymbol{x}[s+p-1]$*;*
4. $\boldsymbol{x}[e-p+1..e]$ *cannot be shifted right, i.e. either* $e = n$ *or* $\boldsymbol{x}[e+1] \neq \boldsymbol{x}[e-p+1]$.

A run can thus be encoded by a triple $r = (s, e, p)$ where s, e, and p specify the starting position, the ending position, and the period of the run, respectively. Note that the run without its tail, e.i. $\boldsymbol{x}[s..s+p-1]^r$, is a maximal repetition, but not every maximal repetition is a run.

string	Example	Non-example
$ccababccc$	$cc(abab)ccc$	$ccabab(cc)c$
$baaaab$	$b(aaaa)b$	$ba(aa)ab$
$aaaabaabaaaa$	$aa(aabaabaa)aa$	$(aaa)abaabaaaa$
$abaabbabbab$	$ab(aa)bbabbab$	$abaa(bbabba)b$
$acaacaacabcabcac$	$acaacaa(cabcabca)c$	$(acaacaac)abcabcac$

Table 1.12: Examples and Non-examples of runs in a string $\boldsymbol{x}$, '()' shows the run and non-run

Definition 1.18. *The* maximum-number-of-runs *problem counts the maximum number of runs in a string, i.e. the occurrences of the runs are counted rather than their types.*

Definition 1.19. *The* maximum-number-of-distinct-squares *problem counts the maximum number of distinct squares in a string, i.e. the types of the squares in the string are counted rather than their occurrences.*

string	square-type	run-occurrence
$ababcabab$	$(ab)^2 \times 1$	$(ab)^2 \times 2$
$abababab$	$(ab)^2 \times 1,\ (ba)^2 \times 1$	$(ab)^4 \times 1$
$aaaaaa$	$a^2 \times 1$	$a^6 \times 1$

Table 1.13: Examples of maximum-number-of-distinct-squares *vs* maximum-number-of-runs when we only count distinct primitively rooted squares

string	square-type	run-occurrence
$aaaabaacaabaaaa$	$a^2 \times 1, (aa)^2 \times 1$	$a^4 \times 2, a^2 \times 2$
$aaaaaa$	$a^2 \times 1,\ (aa)^2 \times 1,\ (aaa)^2 \times 1$	$a^6 \times 1$
$aaaa$	$a^2 \times 1,\ (aa)^2 \times 1$	$a^4 \times 1$

Table 1.14: Examples of maximum-number-of-distinct-squares *vs* maximum-number-of-runs when we count distinct squares no matter whether their roots are primitive or not

Definition 1.20. *The* factorization *or* decomposition *of a string is a collection of disjoint substrings such that their concatenation gives the whole string.*

Definition 1.21. [Lampel-Ziv factorization]
Let $\boldsymbol{x} = \boldsymbol{x}[1..n]$ *be a string of length* $\boldsymbol{n}$ *over an alphabet* Σ*. The Lampel-Ziv factorization of* $\boldsymbol{x}$ *is a factorization* $\boldsymbol{x} = \boldsymbol{w}_1\boldsymbol{w}_2...\boldsymbol{w}_m$ *such that each* $\boldsymbol{w}_k,\ 1 \leq k \leq m$*, is either:*

- *a letter* $c \in \Sigma$ *that does not occur in* $\boldsymbol{w}_1\boldsymbol{w}_2...\boldsymbol{w}_{k-1}$*, or*
- *the longest substring of* $\boldsymbol{x}$ *that occurs at least twice in* $\boldsymbol{w}_1\boldsymbol{w}_2...\boldsymbol{w}_k$*.*

string	L-Z factorization	Non-example
$aaaa$	a, aaa	aa, aa
$abaaaba$	a, b, a, aa, ba	a, b, aaa, ba
$aabaab$	a, a, b, aab	a, a, ba, a, b
$aaabcac$	a, aa, b, c, a, c	a, a, a, b, c, a, c
$abababa$	$a, b, ababa$	a, b, ab, ab, a

Table 1.15: Examples and Non-examples of Lampel-Ziv factorization

Definition 1.22. *The notion of* right rotation *is defined by induction:*

- *A string* $\boldsymbol{x}$ *is a* right 1-rotation *of a string* $\boldsymbol{y}$ *iff* $\boldsymbol{y} = \boldsymbol{y}[1..n]$ *and* $\boldsymbol{x} = \boldsymbol{y}[2..n]\boldsymbol{y}[1]$*,* $\boldsymbol{x}$ *is also denoted as* $R_1(\boldsymbol{y})$*.*
- *A string* $\boldsymbol{x}$ *is a* right k-rotation*,* $k \geq 2$*, of a string* $\boldsymbol{y}$ *iff* $\boldsymbol{x}$ *is a right 1-rotation of* $\boldsymbol{z}$ *which is a right* $(k-1)$*-rotation of* $\boldsymbol{y}$*,* $\boldsymbol{x}$ *is also denoted as* $R_k(\boldsymbol{y})$*.*

Similarly for left rotation*.*

A string $\boldsymbol{x}$ *is referred to as a* trivial rotation *of itself. A string* $\boldsymbol{x}$ *is a rotation of a string* $\boldsymbol{y}$ *if it is a left or right* k*-rotation of* $\boldsymbol{y}$ *for some* $k \geq 1$ *or if* $\boldsymbol{x} = \boldsymbol{y}$*. If* $\boldsymbol{x}$ *is a rotation of* $\boldsymbol{y}$*, we may also say that* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *are* conjugates *or that* $\boldsymbol{x}$ *is a* conjugate *of* $\boldsymbol{y}$*.*

Definition 1.23. *For* $\boldsymbol{x} = \boldsymbol{x}[1..n]$*,* $1 \leq i < j \leq j+k \leq n$*, the string* $\boldsymbol{x}[i+k..j+k]$ *is a* right cyclic shift *by* k *positions of* $\boldsymbol{x}[i..j]$ *if* $\boldsymbol{x}[i] = \boldsymbol{x}[j+1]$*,* $\ldots$*,* $\boldsymbol{x}[i+k-1] = \boldsymbol{x}[j+k]$*. Equivalently, we can say that* $\boldsymbol{x}[i..j]$ *is a* left cyclic shift *by* k *positions of* $\boldsymbol{x}[i+k..j+k]$*. When it is clear from the context, we may leave out the number of positions and just speak of a cyclic shift.*

Note a run can be described as a maximal right cyclic shift of the generator which itself cannot be left cyclic shifted; the length of the right cyclic shift must be at least twice the length of the generator. Similarly, a ran can be described as a maximal right shift of the leading square of the run which itself cannot be left cyclic shifted; the right cyclic shift can be of any length, including 0. The leading square of a run is its leftmost square.

Definition 1.24. *Strings* $\boldsymbol{uv}$ *and* $\boldsymbol{vu}$ *are* conjugates*, written* $\boldsymbol{uv} \sim \boldsymbol{vu}$*. We also say that* $\boldsymbol{vu}$ *is the* $|\boldsymbol{u}|^{th}$ rotation *of* $\boldsymbol{x}$*, written* $R_{|\boldsymbol{u}|}(\boldsymbol{x})$*, or the* $-|\boldsymbol{v}|^{th}$ rotation *of* $\boldsymbol{x}$*, written* $R_{-|\boldsymbol{v}|}(\boldsymbol{x})$*, while* $R_0(\boldsymbol{x}) = R_{-|\boldsymbol{x}|}(\boldsymbol{x}) = \boldsymbol{x}$ *is a* primitive rotation*. Similarly as for the cyclic shift, when it is clear from the context, we may leave out the number of rotations and just speak of a rotation. Note that all cyclic shifts are conjugates, but not the other way around.*

Example

In a string $\boldsymbol{x} = abbaababbbaabb$:

$bbaa$ is the cyclic shift of $abba$, and $bba|a \sim a|bba$,

while $abb|a \sim a|abb$, but $aabb$ is not a cyclic shift of $abba$

Definition 1.25. *Given strings* $\boldsymbol{u}$ *and* $\boldsymbol{v}$, $lcp(\boldsymbol{u}, \boldsymbol{v})$ $\big($*respectively,* $lcs(\boldsymbol{u}, \boldsymbol{v})\big)$ *denotes the length of the* longest common prefix (*respectively,* longest common suffix) *of* $\boldsymbol{u}$ *and* $\boldsymbol{v}$.

Definition 1.26. *A string* $\boldsymbol{x}$ *over an ordered alphabet* $(\mathcal{A}, \prec)$ *is* Lyndon *w.r.t. (with respect to)* $\prec$ *iff* $\boldsymbol{x} \prec \boldsymbol{y}$ *for any non-trivial rotation* $\boldsymbol{y}$ *of* $\boldsymbol{x}$.
A string of length one is thus Lyndon, and it is referred to as trivial *Lyndon word.*

Observation 1.27. *Let* $\boldsymbol{x} = \boldsymbol{x}[1..n], n > 1$ *be a string over an ordered alphabet* $(\mathcal{A}, \prec)$. *Then the following are equivalent:*

1. $\boldsymbol{x}$ *is a non-trivial Lyndon word w.r.t.* $\prec$.
2. $\boldsymbol{x}[1..j-1] \prec \boldsymbol{x}[j..n]$ *for any* $1 < j \leq n$.
3. $\boldsymbol{x}[1..n] \prec \boldsymbol{x}[j..n]$ *for any* $1 < j \leq n$.
4. *there exists* $1 < j \leq n$ *so that both* $\boldsymbol{x}[1..j-1]$ *and* $\boldsymbol{x}[j..n]$ *are Lyndon.*

Item 4 is the basis of the definition of the standard factorization *of a Lyndon word* $\boldsymbol{x}$*: it is a pair of Lyndon words* $\boldsymbol{u}$, $\boldsymbol{v}$ *so that* $\boldsymbol{x} = \boldsymbol{uv}$ *where* $\boldsymbol{v}$ *is maximal such (or, equivalently* $\boldsymbol{u}$ *is minimal such).*

Definition 1.28. *A Lyndon substring* $\boldsymbol{x}[i..j]$ *of a string* $\boldsymbol{x} = \boldsymbol{x}[1..n]$ *is* maximal Lyndon substring *of* $\boldsymbol{x}$ *if either* $j = n$ *or* $\boldsymbol{x}[i..k]$ *is not Lyndon for any* $j < k \leq n$.

A Lyndon substring $\boldsymbol{x}[i..j]$ *of* $\boldsymbol{x}$ *is a* non-extensible Lyndon substring $\boldsymbol{x}$ *if* $\boldsymbol{x}[i..j]$ *is a maximal Lyndon substring of* $\boldsymbol{xy}$ *for any* $\boldsymbol{y}$.

string	Example of Lyndon factorization	Non-example
$abbcc$	$a, bbcc$	ab, acc
aab	a, ab	aa, b
$aaabab$	$a, aabab$	$aa, abab$
$abacc$	ab, acc	aba, cc
$aababb$	$a, ababb$	aab, abb

Table 1.16: Examples and Non-examples of Standard factorization of a Lyndon word

The following theorem is due to the Chen, Fox, and Lyndon, [5], though it is not stated there in this form:

Theorem 1.29. [Lyndon factorization]

For any string $\boldsymbol{x}$ *over an ordered alphabet* $(\mathcal{A}, \prec)$*, there is a unique factorization of* $\boldsymbol{x}$ *into* m *factors,* $\boldsymbol{x} = \boldsymbol{x}_1\boldsymbol{x}_2...\boldsymbol{x}_m$*, such that each factor* $\boldsymbol{x}_i$*,* $1 \leq i \leq m$*, is a maximal Lyndon word, and* $\boldsymbol{x}_1 \succeq \boldsymbol{x}_2 \succeq ... \succeq \boldsymbol{x}_m$.

string	Example of Lyndon factorization	Non-example
$aaaa$	a, a, a, a	aa, aa
aba	ab, a	a, ba
$aabaab$	aab, aab	$a, abaab$
$abaaba$	ab, aab, a	aba, aba
$abccba$	$abccb, a$	abc, cba

Table 1.17: Examples and Non-examples of Lyndon factorization

The Synchronization Principle and Common Factor Lemma are important tools for investigation of canonical factorizations, so we state them here. Lemma 1.30 will be used in the proof of the Synchronization Principle. The symbol | represents *divides*, i.e. $a|b$ means that a divides b.

Lemma 1.30 ([17],Lemma 1.4.2)**.** *Let $\boldsymbol{x}$ be a string of length n and minimum period $\pi \leq n$, let j be an integer and $1 \leq j < n$. Then $R_j(\boldsymbol{x}) = \boldsymbol{x}$ if and only if $\boldsymbol{x}$ is not primitive ($\pi < n,\ \pi \mid n$) and $\pi \mid j$.*

string	n	π	j	$R_j(\boldsymbol{x})$
bababa	6	2	4	*ba*$\underline{ba}$*ba*
aaa	3	1	3	$\underline{aaa}$

Table 1.18: Example of strings with $R_j(\boldsymbol{x}) = \boldsymbol{x}$, where underlined letters are those that were rotated

string	n	π	j	$R_j(\boldsymbol{x})$	comment
ababa	5	5	1	*baba*$\underline{a}$	$\boldsymbol{x}$ is primitive, $\pi \nmid j$, $R_j(\boldsymbol{x}) \neq \boldsymbol{x}$
ababab	6	2	3	*bab*$\underline{aba}$	$\boldsymbol{x}$ is not primitive but $\pi \nmid j$, $R_j(\boldsymbol{x}) \neq \boldsymbol{x}$

Table 1.19: Examples of strings with $R_j(\boldsymbol{x}) \neq \boldsymbol{x}$, where underlined letters are those that were rotated. Moreover, when $\boldsymbol{x}$ is primitive, $n = \pi$ and $j < n$, π can not divide j.

Lemma 1.31 (Synchronization Principle)**.** *The primitive string $\boldsymbol{x}$ occurs exactly p times in $\boldsymbol{x}_2\boldsymbol{x}^p\boldsymbol{x}_1$, where p is a nonnegative integer and $\boldsymbol{x}_1$ (respectively, $\boldsymbol{x}_2$) is a proper prefix (respectively, proper suffix) of $\boldsymbol{x}$.*

Proof. Suppose there exists another occurrence of $\boldsymbol{x}$ except those occurrences determined by $\boldsymbol{x}^p$, then it must start at the same position with a rotation of

$\boldsymbol{x}$, which means $\boldsymbol{x} = R_j(\boldsymbol{x})$ somewhere in the string. Thus $\boldsymbol{x}$ is not primitive by Lemma 1.30, contradiction. □

string	$\boldsymbol{x}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{p}$	comment
$c\underline{abc}a$	abc	a	c	1	$\boldsymbol{x}$ can't shift
$ab\underline{abbababbababbab}abb$	$abbab$	abb	ab	3	$\boldsymbol{x}$ can't shift

Table 1.20: Illustrations of Synchronization Principle holding for primitive string $\boldsymbol{x}$, the underlined letters are occurrences of $\boldsymbol{x}$

string	$\boldsymbol{x}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{p}$	comment
$\underline{babbabbabbabbabbab}$	$babbab$	bab	bab	2	$\boldsymbol{x}$ occurs 4 times $> \boldsymbol{p}$
$b\underline{ababababab}$	$abab$	ab	b	2	$\boldsymbol{x}$ occurs 5 times $> \boldsymbol{p}$

Table 1.21: Illustrations of Synchronization Principle not holding for non-primitive string $\boldsymbol{x}$, the underlined letters are occurrences of $\boldsymbol{x}$

Lemma 1.32 (Common Factor Lemma)**.** *Suppose that $\boldsymbol{x}$ and $\boldsymbol{y}$ are primitive strings, where $\boldsymbol{x}_1$ (respectively, $\boldsymbol{y}_1$) is a proper prefix and $\boldsymbol{x}_2$ (respectively, $\boldsymbol{y}_2$) a proper suffix of $\boldsymbol{x}$ (respectively, $\boldsymbol{y}$). If for nonnegative integers p and q, $\boldsymbol{x}_2\boldsymbol{x}^p\boldsymbol{x}_1$ and $\boldsymbol{y}_2\boldsymbol{y}^q\boldsymbol{y}_1$ have a common factor of length $|\boldsymbol{x}| + |\boldsymbol{y}|$, then $\boldsymbol{x} \sim \boldsymbol{y}$.*

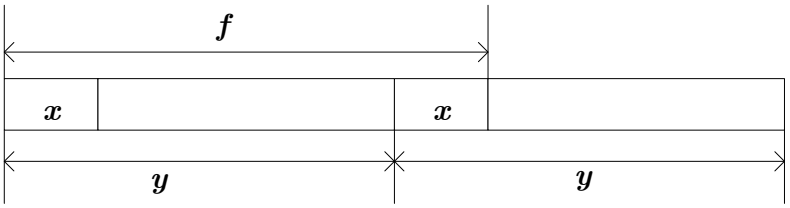
Proof. First consider the special case $\boldsymbol{x}_1 = \boldsymbol{x}_2 = \boldsymbol{y}_1 = \boldsymbol{y}_2 = \boldsymbol{\varepsilon}$, where $\boldsymbol{x}^p$, $\boldsymbol{y}^q$ have a common *prefix* $\boldsymbol{f}$ of length $|\boldsymbol{x}| + |\boldsymbol{y}|$. We show that in this case $\boldsymbol{x} = \boldsymbol{y}$.

Observe that $\boldsymbol{f}$ has prefixes $\boldsymbol{x}$ and $\boldsymbol{y}$, so that if $|\boldsymbol{x}| = |\boldsymbol{y}|$, then $\boldsymbol{x} = \boldsymbol{y}$, as required. Therefore suppose without loss of generality that $|\boldsymbol{x}| < |\boldsymbol{y}|$. Note that $\boldsymbol{y} \neq \boldsymbol{x}^k$ for any integer $k \geq 2$, since otherwise $\boldsymbol{y}$ would not be primitive, contradicting the hypothesis of the lemma.

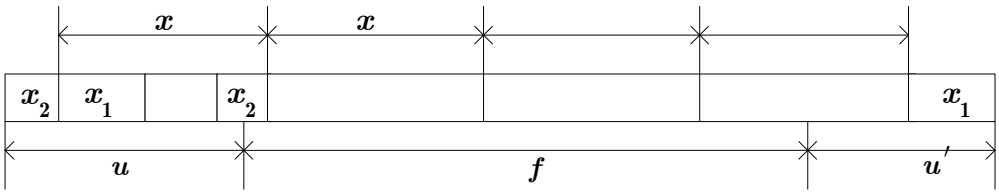
Hence there exists $k \geq 1$ such that $k|\boldsymbol{x}| < |\boldsymbol{y}|$ and $(k+1)|\boldsymbol{x}| > |\boldsymbol{y}|$. But since $\boldsymbol{f} = \boldsymbol{y}\boldsymbol{x}$, it follows that

$$R_{|\boldsymbol{y}|-k|\boldsymbol{x}|}(\boldsymbol{x}) = \boldsymbol{x},$$

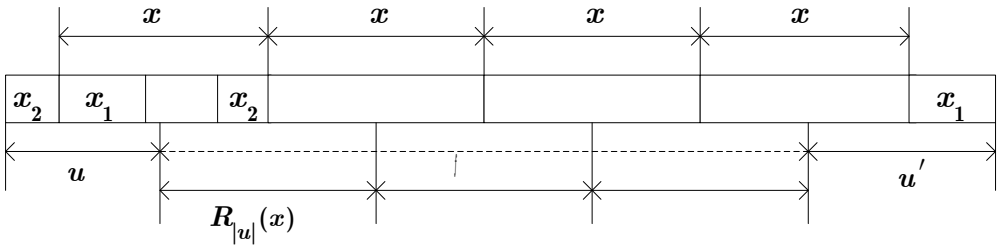
again by Lemma 1.30 contrary to the assumption that $\boldsymbol{x}$ is primitive. We conclude that $|\boldsymbol{x}| \not< |\boldsymbol{y}|$, hence that $|\boldsymbol{x}| = |\boldsymbol{y}|$ and $\boldsymbol{x} = \boldsymbol{y}$, as required.



Now consider the general case, where $\boldsymbol{f}$ of length $|\boldsymbol{x}| + |\boldsymbol{y}|$ is a common factor of $\boldsymbol{x}_2\boldsymbol{x}^p\boldsymbol{x}_1$ and $\boldsymbol{y}_2\boldsymbol{y}^q\boldsymbol{y}_1$. Then $\boldsymbol{x}_2\boldsymbol{x}^p\boldsymbol{x}_1 = \boldsymbol{u}\boldsymbol{f}\boldsymbol{u'}$ for some $\boldsymbol{u}$ and $\boldsymbol{u'}$. If $|\boldsymbol{u}| \geq |\boldsymbol{x}|$, then $\boldsymbol{f}$ is a factor of $\boldsymbol{x}_2\boldsymbol{x}^{p-1}\boldsymbol{x}_1$.



And so we can assume without loss of generality that $|\boldsymbol{u}| < |\boldsymbol{x}|$. Setting $\tilde{\boldsymbol{x}} = R_{|\boldsymbol{u}|}(\boldsymbol{x})$, we see that $\boldsymbol{f}$ is a prefix of $\tilde{\boldsymbol{x}}^p$.



Similarly, by setting $\boldsymbol{y}_2\boldsymbol{y}^q\boldsymbol{y}_1 = \boldsymbol{v}\boldsymbol{f}\boldsymbol{v}'$, we can assume that $|\boldsymbol{v}| < |\boldsymbol{y}|$, hence that $\boldsymbol{f}$ is also a prefix of $\tilde{\boldsymbol{y}}^q$ for $\tilde{\boldsymbol{y}} = R_{|\boldsymbol{v}|}(\boldsymbol{y})$. But this is just the special case considered above, for which $\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{y}}$. Since $\boldsymbol{x} \sim \tilde{\boldsymbol{x}}$ and $\boldsymbol{y} \sim \tilde{\boldsymbol{y}}$, the result follows. □

string	$\boldsymbol{x}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{p}$	length
aa<u>bbaabbaa</u>bbaab	*abba*	*ab*	*a*	3	4
string	$\boldsymbol{y}$	$\boldsymbol{y}_1$	$\boldsymbol{y}_2$	$\boldsymbol{p}$	length
<u>abbaabbaa</u>bbaabbaabbaa	*baab*	*baa*	*ab*	4	4

Table 1.22: Example of $\boldsymbol{x}_2\boldsymbol{x}^p\boldsymbol{x}_1$ and $\boldsymbol{y}_2\boldsymbol{y}^q\boldsymbol{y}_1$ having a common factor of length $|\boldsymbol{x}| + |\boldsymbol{y}|$, then $\boldsymbol{x} \sim \boldsymbol{y}$. Underlined letters are the common factor.

string	$\boldsymbol{x}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{p}$	length
bc<u>abcabc</u>a	*abc*	*a*	*bc*	2	3
string	$\boldsymbol{y}$	$\boldsymbol{y}_1$	$\boldsymbol{y}_2$	$\boldsymbol{p}$	length
<u>cabcabc</u>abcabc	*bca*	*bc*	*ca*	3	3

Table 1.23: Example of $\boldsymbol{x}_2\boldsymbol{x}^p\boldsymbol{x}_1$ and $\boldsymbol{y}_2\boldsymbol{y}^q\boldsymbol{y}_1$ having a common factor of length $|\boldsymbol{x}| + |\boldsymbol{y}|$, then $\boldsymbol{x} \sim \boldsymbol{y}$. Underlined letters are the common factor.

Note that Lemma 1.32 could be equivalently stated in a more general form:

Lemma 1.33. *Suppose that $\boldsymbol{x}$ and $\boldsymbol{y}$ are strings where $\boldsymbol{x}_1$ (respectively, $\boldsymbol{y}_1$) is a proper prefix and $\boldsymbol{x}_2$ (respectively, $\boldsymbol{y}_2$) a proper suffix of $\boldsymbol{x}$ (respectively,*

$\boldsymbol{y}$). *If for nonnegative integers* p *and* q, $\boldsymbol{x}_2\boldsymbol{x}^p\boldsymbol{x}_1$ *and* $\boldsymbol{y}_2\boldsymbol{y}^q\boldsymbol{y}_1$ *have a common factor of length* $|\boldsymbol{x}| + |\boldsymbol{y}|$, *then the primitive root* $\bar{\boldsymbol{x}}$ *of* $\boldsymbol{x}$ *and the primitive root* $\bar{\boldsymbol{y}}$ *of* $\boldsymbol{y}$ *are conjugates.*

string	$\boldsymbol{x}$	$\bar{\boldsymbol{x}}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{p}$	length
bcabcabcabcabcabcab	*abcabc*	*abc*	*ab*	*bcabc*	2	6
string	$\boldsymbol{y}$	$\bar{\boldsymbol{y}}$	$\boldsymbol{y}_1$	$\boldsymbol{y}_2$	$\boldsymbol{p}$	length
abcabcabcabcabc	*cabcab*	*cab*	*c*	*ab*	2	6

Table 1.24: Example of the primitive root $\bar{\boldsymbol{x}}$ of $\boldsymbol{x}$ and the primitive root $\bar{\boldsymbol{y}}$ of $\boldsymbol{y}$ being conjugates. Underlined letters are the common factor.

The Common Factor Lemma gives rise to the following useful corollary:

Lemma 1.34. *Suppose that* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *are primitive strings, and that* p *and* q *are positive integers.*

(*a*) *If* $\boldsymbol{x}^p = \boldsymbol{y}^q$, *then* $\boldsymbol{x} = \boldsymbol{y}$ *and* $p = q$.

(*b*) *If* $\boldsymbol{x}_1$ (*respectively,* $\boldsymbol{y}_1$) *is a proper prefix of* $\boldsymbol{x}$ (*respectively,* $\boldsymbol{y}$) *and* $\boldsymbol{x}^p\boldsymbol{x}_1 = \boldsymbol{y}^q\boldsymbol{y}_1$ *for* $p \geq 2,\ q \geq 2$, *then* $\boldsymbol{x} = \boldsymbol{y},\ \boldsymbol{x}_1 = \boldsymbol{y}_1$ *and* $p = q$.

Proof. For (a), first consider $p = 1$, thus $\boldsymbol{x} = \boldsymbol{y}^q$. Since $\boldsymbol{x}$ is primitive, therefore $q = 1$ and $\boldsymbol{x} = \boldsymbol{y}$, as required. Similarly for $q = 1$. Suppose then that $p, q \geq 2$. This means that $\boldsymbol{x}^p$ and $\boldsymbol{y}^q = \boldsymbol{x}^p$ have a common factor of length $p|\boldsymbol{x}| = q|\boldsymbol{y}| \geq |\boldsymbol{x}|+|\boldsymbol{y}|$, so that by Lemma 1.32 $\boldsymbol{x} \sim \boldsymbol{y}$. Hence $|\boldsymbol{x}| = |\boldsymbol{y}|$ and so $\boldsymbol{x} = \boldsymbol{y}$.

For (b), since again $p \geq 2,\ q \geq 2$, it follows as in (a) that $\boldsymbol{x}^p\boldsymbol{x}_1 = \boldsymbol{y}^q\boldsymbol{y}_1$ has a common factor of length at least $|\boldsymbol{x}| + |\boldsymbol{y}|$, hence the result. □

Note that in Lemma 1.34(b) the requirement $p \geq 2,\ q \geq 2$ is essential. For instance, $\boldsymbol{x} = aabb,\ \boldsymbol{x}_1 = aa$ and $p = 2$ yields $\boldsymbol{x}^p\boldsymbol{x}_1 = aabbaabbaa$, identical to $\boldsymbol{y}^q\boldsymbol{y}_1$ produced by $\boldsymbol{y} = aabbaabba,\ \boldsymbol{y}_1 = a$ and $q = 1$ — but of course $\boldsymbol{x} \neq \boldsymbol{y}$.

Chapter 2

Canonical factorization of double squares

In this chapter based on [2], we discuss the most general form of canonical factorization of double squares. But first we need to defined precisely what is meant by a *double square*.

The term *double square* is due to Lam, [15] and he defined it as two rightmost occurrences of a square $\boldsymbol{u}^2$ and a square $\boldsymbol{v}^2$ starting at the same position of a string $\boldsymbol{x}$. Following on work of Fraenkel and Simpson [10], he tried to obtain a better bound for the number of distinct squares by bounding the number of double squares. In [8], such double squares are referred to as *FS-double squares*, to acknowledge the pioneering work of Fraenkel and Simpson in this field and to prevent a confusion with configurations of two squares starting at the same position. The bounding of FS-double squares

carried in [8] gave an improved upper bound of $\frac{11}{6}|\boldsymbol{x}|$ for the number of distinct squares in $\boldsymbol{x}$.

From now on in this thesis, we will use the term *FS-double square* for two rightmost occurrences of squares starting at the same position, while the term *double square* for us is any two squares starting at the same position. We use the following notation: $DS(\boldsymbol{u}, \boldsymbol{v})$ indicates that the squares $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$ start at the same position and that $|\boldsymbol{u}| < |\boldsymbol{v}|$. For a double square $DS(\boldsymbol{u}, \boldsymbol{v})$, we say that the squares $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$ are *proportional* if $|\boldsymbol{u}| < |\boldsymbol{v}| < 2|\boldsymbol{u}| < 2|\boldsymbol{v}|$. If the two squares of a double square are proportional, then the double square is called *balanced*.

Lam, [15], tried to provide a taxonomy of mutual relationships of FS-double squares, and for that purpose he described a factorization of an FS-double square into primitive components. In [8], some important properties of such factorization were discussed and the factorization was shown to be unique, i.e. it is shown that every FS-double square $(\boldsymbol{u}, \boldsymbol{v})$ has a unique factorization $\boldsymbol{u}_1$, $\boldsymbol{u}_2$, e_1 and e_2 so that $\boldsymbol{u}^2 = (\boldsymbol{u}_1\boldsymbol{u}_2)^2$ and $\boldsymbol{v}^2 = \boldsymbol{u}_1^{e_1}\boldsymbol{u}_2\boldsymbol{u}_1^{e_1+e_2}\boldsymbol{u}_2\boldsymbol{u}_1^{e_2}$ where $\boldsymbol{u}_1$ is primitive, $\boldsymbol{u}_2$ is a non-trivial proper prefix of $\boldsymbol{u}_1$, and e_1, e_2 are integers so that $e_1 \geq e_2 \geq 1$. The term *canonical factorization* was used there.

The main result of [2] is the fact that any balanced double square admits the canonical factorization. In a sense, it was a big surprise that no additional conditions are required. For the Three Squares Lemma, the squares were required to be primitive, or as Fraenkel and Simpson showed, the smallest

one must be primitive. For the New Periodicity Lemma, the regularity of the smallest square was required, even stronger property than primitivness. The canonical factorization for FS-double squares was not surprising as being the rightmost occurrence is again stronger than being primitive.

The main result of this chapter is Lemma 2.1, *Two Squares Factorization Lemma*, which specifies the unique factorization imposed by the occurrence of two proportional squares at the same position in a string.

2.1 Two Squares Factorization

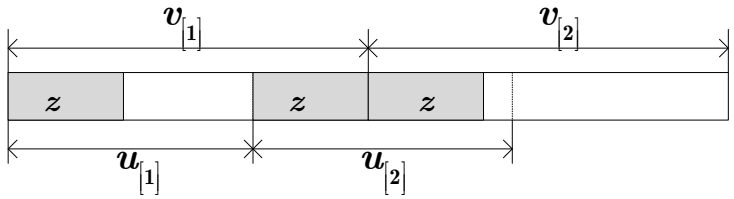
Lemma 2.1 (Two Squares Factorization Lemma)**.** *For a balanced double square $DS(\boldsymbol{u}, \boldsymbol{v})$, there exists a unique primitive string $\boldsymbol{u}_1$ such that $\boldsymbol{u} = \boldsymbol{u}_1^{e_1}\boldsymbol{u}_2$ and $\boldsymbol{v} = \boldsymbol{u}_1^{e_1}\boldsymbol{u}_2\boldsymbol{u}_1^{e_2}$, where $\boldsymbol{u}_2$ is a possibly empty proper prefix of $\boldsymbol{u}_1$ and e_1, e_2 are integers such that $e_1 \geq e_2 \geq 1$. Moreover,*

(a) *if $|\boldsymbol{u}_2| = 0$, then $e_1 > e_2 \geq 1$;*

(b) *if $|\boldsymbol{u}_2| > 0$, then $\boldsymbol{v}$ is primitive, and if in addition $e_1 \geq 2$, then $\boldsymbol{u}$ is also primitive.*

In both cases, the factorization is unique.

Proof. Let $\boldsymbol{z}$ be the nonempty proper prefix of $\boldsymbol{u}_{[2]}$ that is in addition a suffix $\boldsymbol{z}$ of $\boldsymbol{v}_{[1]}$. But then $\boldsymbol{z}$ is also a prefix of $\boldsymbol{v}_{[1]}$, hence of $\boldsymbol{v}_{[2]}$; thus if $|\boldsymbol{u}| \geq 2|\boldsymbol{z}|$, it follows that $\boldsymbol{z}^2$ is a prefix of $\boldsymbol{u}$.

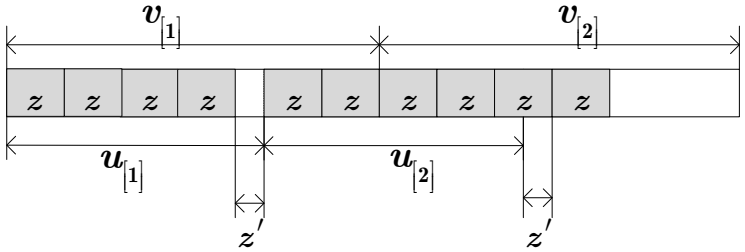


In general, there exists an integer $k = \lfloor |\boldsymbol{u}|/|\boldsymbol{z}| \rfloor \geq 1$ such that $\boldsymbol{u} = \boldsymbol{z}^k \boldsymbol{z'}$ for some proper prefix $\boldsymbol{z'}$ of $\boldsymbol{z}$. Let $\boldsymbol{u}_1$ be the primitive root of $\boldsymbol{z}$, so that $\boldsymbol{z} = \boldsymbol{u}_1^{e_2}$ for some integer $e_2 \geq 1$.

$$\begin{aligned} \boldsymbol{u} &= \boldsymbol{z}^k \boldsymbol{z'} \\ &= \boldsymbol{u}_1^{e_2 k} \boldsymbol{z'} \\ &= \boldsymbol{u}_1^{e_1} \boldsymbol{u}_2 \end{aligned}$$

$$\begin{aligned} \boldsymbol{v} &= \boldsymbol{u}\boldsymbol{z} \\ &= \boldsymbol{u}_1^{e_1} \boldsymbol{u}_2 \boldsymbol{u}_1^{e_2} \end{aligned}$$

Therefore, for some $e_1 \geq e_2 k$ and some prefix $\boldsymbol{u}_2$ of $\boldsymbol{u}_1$, $\boldsymbol{u} = \boldsymbol{u}_1^{e_1} \boldsymbol{u}_2$ and $\boldsymbol{v} = \boldsymbol{u}\boldsymbol{z} = \boldsymbol{u}_1^{e_1} \boldsymbol{u}_2 \boldsymbol{u}_1{}^{e_2}$, as required.



To prove the uniqueness of $\boldsymbol{u}_1$ we consider two cases:

(i) $|\boldsymbol{u}_2| = 0$

Here $\boldsymbol{u} = \boldsymbol{u}_1^{e_1}$ and $\boldsymbol{v} = \boldsymbol{u}_1^{e_1+e_2}$, so that $\boldsymbol{x} = \boldsymbol{u}_1^{2(e_1+e_2)}$. Since $|\boldsymbol{v}| < 2|\boldsymbol{u}|$ and $e_1 \geq e_2$, it follows that $e_1 > e_2$. The uniqueness of $\boldsymbol{u}_1$ is a consequence of Lemma 1.34(a).

(ii) $|\boldsymbol{u}_2| > 0$

Suppose the choice of $\boldsymbol{u}_1$ is not unique. Then there exists some primitive string $\boldsymbol{w}_1$ with proper prefix $\boldsymbol{w}_2$, together with integers $f_1 \geq f_2 \geq 1$, such that $\boldsymbol{u} = \boldsymbol{w}_1{}^{f_1}\boldsymbol{w}_2$ and $\boldsymbol{v} = \boldsymbol{w}_1{}^{f_1}\boldsymbol{w}_2\boldsymbol{w}_1{}^{f_2}$. If both $e_1 \geq 2$ and $f_1 \geq 2$, it follows from Lemma 1.34(b) that $\boldsymbol{u}_1 = \boldsymbol{w}_1$ and $e_1 = f_1$. If $e_1 = f_1 = 1$, we observe that $\boldsymbol{v} = \boldsymbol{u}\boldsymbol{u}_1 = \boldsymbol{u}\boldsymbol{w}_1$, so that again $\boldsymbol{u}_1 = \boldsymbol{w}_1$. In the only remaining case, exactly one of e_1, f_1 equals 1: therefore suppose without loss of generality that $f_1 > e_1 = 1$. Then $\boldsymbol{u} = \boldsymbol{u}_1\boldsymbol{u}_2 = \boldsymbol{w}_1{}^{f_1}\boldsymbol{w}_2$ and $\boldsymbol{v} = \boldsymbol{u}_1\boldsymbol{u}_2\boldsymbol{u}_1 = \boldsymbol{w}_1{}^{f_1}\boldsymbol{w}_2\boldsymbol{w}_1{}^{f_2}$, so that $\boldsymbol{u}_1 = \boldsymbol{w}_1{}^{f_2}$. But since $\boldsymbol{u}_1$ is primitive, this forces $f_2 = 1$ and $\boldsymbol{u}_1 = \boldsymbol{w}_1$, which, since $\boldsymbol{u}_1\boldsymbol{u}_2 = \boldsymbol{w}_1{}^{f_1}\boldsymbol{w}_2 = \boldsymbol{u}_1^{f_1}\boldsymbol{w}_2$, implies that $f_1 = 1$, a contradiction. Thus all cases have been considered, and $\boldsymbol{u}_1$ is unique.

We now show that $\boldsymbol{v}$ is primitive. Suppose the contrary, so there exists some primitive $\boldsymbol{w}$ and an integer $k \geq 2$ such that $\boldsymbol{v} = \boldsymbol{w}^k$. It follows that $|\boldsymbol{w}| \leq |\boldsymbol{v}|/2 \leq |\boldsymbol{u}_1^{e_1}| + |\boldsymbol{u}_2|$. Note that

$$\boldsymbol{w}^{2k} = \boldsymbol{v}^2 = \boldsymbol{u}_1^{e_1}\boldsymbol{u}_2\boldsymbol{u}_1{}^{e_1+e_2}\boldsymbol{u}_2\boldsymbol{u}_1{}^{e_2}, \tag{2.1}$$

so that $\boldsymbol{w}^{2k}$ and $\boldsymbol{u}_1^{e_1+e_2}\boldsymbol{u}_2$ have a common factor $\boldsymbol{u}_1^{e_1+e_2}\boldsymbol{u}_2$ of length

$$(|\boldsymbol{u}_1^{e_1}| + |\boldsymbol{u}_2|) + |\boldsymbol{u}_1^{e_2}| \geq |\boldsymbol{w}| + |\boldsymbol{u}_1|.$$

Thus we can apply Common Factor Lemma 1.32 to conclude that $\boldsymbol{w}$ and $\boldsymbol{u}_1$ are conjugates, thus by (2.1) that $\boldsymbol{w} = \boldsymbol{u}_1$. But (2.1) then requires that the primitive string $\boldsymbol{u}_1 = \boldsymbol{u}_2\overline{\boldsymbol{u}}_2$ aligns with $\boldsymbol{u}_2\boldsymbol{u}_1$, and so $\overline{\boldsymbol{u}}_2$ is a prefix of $\boldsymbol{u}_1$, which means $\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = \overline{\boldsymbol{u}}_2\boldsymbol{u}_2$ then $\boldsymbol{u}_1$ is non-primitive by Lemma 1.30. This is a contradiction to 2.1 which requires $\boldsymbol{u}_1$ to be primitive. Thus we conclude that $\boldsymbol{v}$ is primitive.

Now suppose in addition that $e_1 \geq 2$, but that $\boldsymbol{u}$ is not primitive. Then there exists some primitive $\boldsymbol{w}$ and some integer $k \geq 2$ such that $\boldsymbol{u} = \boldsymbol{w}^k$. Hence $|\boldsymbol{w}| \leq |\boldsymbol{u}|/2 = (|\boldsymbol{u}_1^{e_1}| + |\boldsymbol{u}_2|)/2 < |\boldsymbol{u}_1^{e_1-1}| + |\boldsymbol{u}_2|$, since $e_1 \geq 2$ and $|\boldsymbol{u}_2| > 0$. Therefore, since $\boldsymbol{u}_1^{e_1}\boldsymbol{u}_2$ is a prefix of $\boldsymbol{u}^2 = \boldsymbol{w}^{2k}$, and since $e_2 \geq 1$ by Lemma 2.1, $\boldsymbol{w}^{2k}$ and $\boldsymbol{u}_1^{e_1+e_2}$ have a common prefix $\boldsymbol{u}_1^{e_1}\boldsymbol{u}_2$. Note that $|\boldsymbol{u}_1^{e_1}\boldsymbol{u}_2| \geq |\boldsymbol{w}| + |\boldsymbol{u}_1|$, so that again applying Common Factor Lemma 1.32, we conclude that $\boldsymbol{u}_1 = \boldsymbol{w}$. This in turn implies $\boldsymbol{u} = \boldsymbol{u}_1^{e_1}\boldsymbol{u}_2 = \boldsymbol{u}_1^k$, impossible since $0 < |\boldsymbol{u}_2| < |\boldsymbol{u}_1|$. Therefore $\boldsymbol{u}$ is primitive, as required.

Finally we remark that since $\boldsymbol{u}_1$ is a uniquely determined primitive string, therefore $\boldsymbol{u}_2$, e_1 and e_2 are also uniquely determined. □

The following examples show that the statement of the lemma is sharp:

(a) The second part of Lemma 2.1(b) requires that $e_1 \geq 2$. To see that this

condition is not necessary, consider $\boldsymbol{v}^2 = abaababaab$, where $\boldsymbol{u} = (ab)a$, $\boldsymbol{v} = (ab)a(ab)$, so that $\boldsymbol{u}_1 = ab$, $\boldsymbol{u}_2 = a$, $e_1 = e_2 = 1$, but $\boldsymbol{u}$ is primitive.

(b) On the other hand, consider $\boldsymbol{v}^2 = abaabaabaababaabaabaab$, where $\boldsymbol{u} = (aba)^2 = (abaab)a$, $\boldsymbol{v}$= $(abaab)a(abaab)$, so that $\boldsymbol{u}_1 = abaab$, $\boldsymbol{u}_2 = a$, $e_1 = e_2 = 1$, where now $\boldsymbol{u}$ is *not* primitive.

$\boldsymbol{u}^2$	$\boldsymbol{v}^2$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$
$abaabaaba \cdot abaabaaba$	$abaabaabaaba \cdot abaabaabaaba$	aba	$\lvert\boldsymbol{u}_2\rvert = 0$
$bbabba \cdot bbabba$	$bbabbabba \cdot bbabbabba$	bba	$\lvert\boldsymbol{u}_2\rvert = 0$

Table 2.1: Illustrations of Lemma 2.1 case (a),'$\cdot$' divides two roots of $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$, in the first example, $e_1 = 3,\ e_2 = 1$, and in the second example, $e_1 = 2,\ e_2 = 1$

$\boldsymbol{u}^2$	$\boldsymbol{v}^2$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$
$abcaabcaabc \cdot abcaabcaabc$	$abcaabcaabcabca \cdot abcaabcaabcabca$	$abca$	abc
$bcbcbcb \cdot bcbcbcb$	$bcbcbcbbcbc \cdot bcbcbcbbcbc$	bc	b

Table 2.2: Illustrations of Lemma 2.1 case (b),'$\cdot$' divides two roots of $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$, in the first example, $e_1 = 2,\ e_2 = 1$, and in the second example, $e_1 = 3,\ e_2 = 2$

$\boldsymbol{u}^2$	$\boldsymbol{v}^2$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$
$abccaabc \cdot abccaabc$	$abccaabcabcca \cdot abccaabcabcca$	$abcca$	abc
$aba \cdot aba$	$abaab \cdot abaab$	ab	a
$abaaba \cdot abaaba$	$abaabaabaab \cdot abaabaabaab$	$abaab$	a
$abaaabaa \cdot abaaabaa$	$abaaabaaabaaaba \cdot abaaabaaabaaaba$	$abaaaba$	a

Table 2.3: Example that the condition of Lemma 2.1 case (b) is not necessary : when $e_1 = e_2 = 1$, $\boldsymbol{u}$ can be either primitive or non-primitive, the first two examples are the case $\boldsymbol{u}$ is not primitive, the last two examples are the case$\boldsymbol{u}$ is primitive, '$\cdot$' divides two roots of $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$, in all examples $e_1 = e_2 = 1$

Lemma 2.1 gives credence to the following definition of terminology and notation:

Definition 2.2. *For a balanced double square $DS(\boldsymbol{u}, \boldsymbol{v})$, we call the unique factorization guaranteed by Lemma 2.1, the* canonical factorization *of the double square and denote it by $DS(\boldsymbol{u}, \boldsymbol{v} \ : \ \boldsymbol{u}_1, \boldsymbol{u}_2, e_1, e_2)$. The symbol $\overline{\boldsymbol{u}}_2$ denotes the suffix of $\boldsymbol{u}_1$ such that $\boldsymbol{u}_1 = \boldsymbol{u}_2\overline{\boldsymbol{u}}_2$.*

$\boldsymbol{u}^2$	$bbab \cdot bbab$
$\boldsymbol{v}^2$	$bbabbba \cdot bbabbba$
$\boldsymbol{u}_1$	bba
$\boldsymbol{u}_2$	b
$\overline{\boldsymbol{u}}_2$	ba
e_1	1
e_2	1

Table 2.4: An example of canonical factorization, '$\cdot$' divides two roots of $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$

Lemma 2.1 also gives rise to a number of important observations:

Observation 2.3. *In Lemma 2.1, $|\boldsymbol{u}_2| > 0$ if any one of the following conditions holds:*

(a) $\boldsymbol{v}$ *is primitive;*

(b) $\boldsymbol{u}$ *is primitive;*

(c) *there is no other occurrence of $\boldsymbol{u}^2$ farther to the right in $\boldsymbol{v}^2$ ($\boldsymbol{u}^2$ is rightmost);*

(d) $\boldsymbol{u}^2$ *is regular.*

Moreover:

(e) $|\boldsymbol{u}_2| > 0$ *if and only if* $\boldsymbol{v}$ *is primitive;*

(f) *If* $\boldsymbol{u}^2$ *is regular, then* $e_1 = e_2 = 1$ *and* $\boldsymbol{u}_1$ *is regular.*

Proof. (a) $|\boldsymbol{u}_2| = 0$ implies $\boldsymbol{v}$ not primitive.

(b) $|\boldsymbol{u}_2| = 0$ implies $\boldsymbol{u}$ not primitive.

(c) $|\boldsymbol{u}_2| = 0$ implies $\boldsymbol{u}^2 = \boldsymbol{u}_1^{2e_1}$, which occurs twice in $\boldsymbol{v}^2 = \boldsymbol{u}_1^{2(e_1+e_2)}$, in particular as a suffix.

(d) Since $\boldsymbol{u}^2$ is regular, therefore $\boldsymbol{u}$ is primitive, so that by (b), $|\boldsymbol{u}_2| > 0$.

(e) By (a), primitive $\boldsymbol{v}$ implies $|\boldsymbol{u}_2| > 0$; by Lemma 2.1, $|\boldsymbol{u}_2| > 0$ implies that $\boldsymbol{v}$ is primitive.

(f) By (d), regular $\boldsymbol{u}^2$ implies $|\boldsymbol{u}_2| > 0$, so that $\boldsymbol{u} = \boldsymbol{u}_1^{e_1}\boldsymbol{u}_2$, which is regular only if $e_1 = e_2 = 1$ and $\boldsymbol{u}_1$ is regular.

□

In the context of Observation 2.3(f), consider the double square $DS(\boldsymbol{u}, \boldsymbol{v} : \boldsymbol{u}_1, \boldsymbol{u}_2, e_1, e_2)$ where $\boldsymbol{u} = aabaa,\ \boldsymbol{v} = aabaaaab$. In this case, we find $\boldsymbol{u}_1 = aab,\ \boldsymbol{u}_2 = aa,\ e_1 = e_2 = 1$, but observe that $\boldsymbol{u}$ has prefix a^2, so $\boldsymbol{u}^2$ is not regular. Thus the condition $e_1 = 1$ is more general than the requirement that $\boldsymbol{u}^2$ be regular.

The following tables illustrate some of the cases discussed in this chapter.

$\boldsymbol{v}$	$\boldsymbol{u}$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	e_1	e_2
$abaababa$: $\boldsymbol{v}$ is primitive	$abaab$	aba	ab : $\lvert\boldsymbol{u}_2\rvert > 0$	1	1
$ababab$: $\boldsymbol{v}$ is non-primitive	$abab$	ab	$\lvert\boldsymbol{u}_2\rvert = 0$	2	1

Table 2.5: Example of Observation 2.3 case (a)

$\boldsymbol{v}$	$\boldsymbol{u}$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$
$abaab$	aba : $\boldsymbol{u}$ is primitive	ab	a : $\lvert\boldsymbol{u}_2\rvert > 0$
$abcaabcaabcabca$	$abcaabcaabc$: $\boldsymbol{u}$ is primitive	$abca$	abc : $\lvert\boldsymbol{u}_2\rvert > 0$

Table 2.6: Example of Observation 2.3 case (b), in the first example, $e_1 = e_2 = 1$, and in the second example, $e_1 = 2,\ e_2 = 1$

$\boldsymbol{v}$	$\boldsymbol{u}$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	e_1	e_2
$abaabaabaab$	$abaaba$	$abaab$	a	1	1
$aaaaaaa$	$aaaa$	aaa	a	1	1

Table 2.7: Example of the unnecessity of the condition in Observation 2.3 case (b): $\boldsymbol{u}$ is not primitive , but $\lvert\boldsymbol{u}_2\rvert > 0$

$\boldsymbol{v}^2$	$\boldsymbol{u}^2$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	e_1	e_2
$\underline{baababaa \cdot ba}ababaa$	$baaba \cdot baaba$	baa	ba	1	1
$\underline{abaab \cdot a}baab$	$aba \cdot aba$	ab	a	1	1
$abbabb\underline{abb \cdot abbabbabb}$	$abb \cdot abb$	abb	$\lvert\boldsymbol{u}_2\rvert = 0$	2	1

Table 2.8: Example of Observation 2.3 case (c), in the third example, there exists other occurrence of of $\boldsymbol{u}^2$ farther to the right in $\boldsymbol{v}^2$, underlined letters are the right-most occurrence of $\boldsymbol{u}^2$ in $\boldsymbol{v}^2$,'$\cdot$' divides two roots of $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$

$\boldsymbol{v}^2$	$\boldsymbol{u}^2$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	e_1	e_2
$aaaaaa\underline{aa \cdot aaaaaaaa}$	$aaaaa\lvert aaaaa$	aaa	aa	1	1

Table 2.9: Example of the unnecessity of the condition in Observation 2.3 case (c): $\boldsymbol{u}^2$ is not the rightmost in $\boldsymbol{v}^2$, but $\lvert\boldsymbol{u}_2\rvert > 0$, '$\cdot$' divides two roots of $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$

$\boldsymbol{v}$	$\boldsymbol{u}$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	e_1	e_2
$abaaaba \cdot abaaaba$	$abaa \cdot abaa$	aba	a	1	1
$abcababc \cdot abcababc$	$abcab \cdot abcab$	abc	ab	1	1

Table 2.10: Example of Observation 2.3 case (d), '$\cdot$' divides two roots of $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$

$\boldsymbol{v}$	$\boldsymbol{u}$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	e_1	e_2
$aabaabaaab \cdot aabaabaaab$	$\underline{aa}baaba \cdot aabaaba$	aab	a	2	1
$ababaaababa \cdot ababaaababa$	$\underline{ababa}a \cdot ababaa$	$ababa$	a	1	1

Table 2.11: Example of the unnecessity of the condition in Observation 2.3 case (d): $\boldsymbol{u}^2$ is not regular, but $|\boldsymbol{u}_2| > 0$, underlined letters are the square prefix of $\boldsymbol{u}^2$, '$\cdot$' divides two roots of $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$

$\boldsymbol{v}$	$\boldsymbol{u}$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	e_1	e_2
$abcababc \cdot abcababc$	$abcab \cdot abcab$	abc	ab	1	1
$abbaabb \cdot abbaabb$	$abba \cdot abba$	abb	a	1	1

Table 2.12: Illustrations of Observation 2.3 case (f),'$\cdot$' divides two roots of $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$

$\boldsymbol{v}$	$\boldsymbol{u}$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	e_1	e_2
$abaabaabaab \cdot abaabaabaab$	$\underline{abaaba} \cdot abaaba$	$abaab$	a	1	1

Table 2.13: Example of the unnecessity of the condition in Observation 2.3 case (f): $\boldsymbol{u}^2$ is not regular, but $e_1 = e_2 = 1$ and $\boldsymbol{u}_1$ is regular, underlined letters are the square prefix of $\boldsymbol{u}^2$,'$\cdot$' divides two roots of $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$

2.2 Inversion Factors in a Balanced Double Square

Inversion factors were introduced in [8] and used in FS double squares, here we extend the concept to all double squares that admit special canonical factorization: consider a double square $DS(\boldsymbol{u}, \boldsymbol{v} : \boldsymbol{u}_1, \boldsymbol{u}_2, e_1, e_2)$ where $|\boldsymbol{u}_2| > 0$, and let $\boldsymbol{u}_1 = \boldsymbol{u}_2\overline{\boldsymbol{u}}_2$. Thus $\boldsymbol{v}^2$ becomes

$$
\begin{aligned}
\boldsymbol{v}^2 &= (\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1}\boldsymbol{u}_2(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1+e_2}\boldsymbol{u}_2(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_2} \\
&= (\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1-1}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)\boldsymbol{u}_2(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1+e_2-2}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)\boldsymbol{u}_2(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_2-1} \\
&= (\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1-1}\boldsymbol{u}_2(\overline{\boldsymbol{u}}_2\boldsymbol{u}_2\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1+e_2-2}\boldsymbol{u}_2(\overline{\boldsymbol{u}}_2\boldsymbol{u}_2\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_2-1} \\
&= (\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1-1}\boldsymbol{u}_2(\mathbb{IF})(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1+e_2-2}\boldsymbol{u}_2(\mathbb{IF})(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_2-1} \qquad (2.2)
\end{aligned}
$$

where $\mathbb{IF} = \overline{\boldsymbol{u}}_2\boldsymbol{u}_2\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = R_{|\boldsymbol{u}_2|}(\boldsymbol{u}_1)\boldsymbol{u}_1$ is called the ***inversion factor***.

$\boldsymbol{v}^2$	$\boldsymbol{u}^2$	$\boldsymbol{u}_1$	$\boldsymbol{u}_2$	$\overline{\boldsymbol{u}_2}$	$\mathbb{IF}$
$\underline{abbaaabba} \cdot a\underline{bbaaabba}$	$abbaa \cdot abbaa$	$abba$	a	bba	$bbaaabba$
$abbab\underline{ababb} \cdot abbab\underline{ababb}$	$abbabbab \cdot abbabbab$	abb	ab	b	$bababb$

Table 2.14: underlined letters emphasize the inversion factors,'$\cdot$' divides two roots of $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$, in the first example, $e_1 = e_2 = 1$, and in the second example, $e_1 = 2,\ e_2 = 1$

Lemma 2.4. *Consider a double square* $DS(\boldsymbol{u}, \boldsymbol{v} : \boldsymbol{u}_1, \boldsymbol{u}_2, e_1, e_2)$ *with a non-empty* $\boldsymbol{u}_2$*. Then the inversion factor* $\mathbb{IF}$ *has exactly two occurrences in* $\boldsymbol{v}^2$ *exactly a distance of* $|\boldsymbol{v}|$ *apart as shown in* (2.2)*.*

Proof. Let $\mathbb{IF}_1$ denote the first inversion factor in a double square , $\mathbb{IF}_2$ denote the second inversion factor in a double square .
$\boldsymbol{v}^2 = (\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1-1}\boldsymbol{u}_2(\mathbb{IF}_1)(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1+e_2-2}\boldsymbol{u}_2(\mathbb{IF}_2)(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_2-1}$
The distance between $\mathbb{IF}_1$ and $\mathbb{IF}_2$ is $|(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1+e_2}|+|\boldsymbol{u}_2| = |\boldsymbol{u}_1^{e_1}\boldsymbol{u}_2\boldsymbol{u}_1^{e_2}| = |\boldsymbol{v}|$. Suppose there exists an inversion factor occurring somewhere else than the positions of $\mathbb{IF}_1$ and $\mathbb{IF}_2$ in $\boldsymbol{v}^2$. According to the Synchronization principle Lemma 1.31, the subfactor $\boldsymbol{u}_2\overline{\boldsymbol{u}}_2$ of the inversion factor must align with one of the three occurrences of $\boldsymbol{u}_2\overline{\boldsymbol{u}}_2$ in $\boldsymbol{v}^2$, as $\boldsymbol{u}_1 = \boldsymbol{u}_2\overline{\boldsymbol{u}}_2$ is primitive. On the other hand, another subfactor $\overline{\boldsymbol{u}}_2\boldsymbol{u}_2$ always aligns with $\boldsymbol{u}_2\overline{\boldsymbol{u}}_2$. Thus $\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = \overline{\boldsymbol{u}}_2\boldsymbol{u}_2$, which means $\boldsymbol{u}_1 = \boldsymbol{u}_2\overline{\boldsymbol{u}}_2$ is not primitive by the Lemma 1.30 , contradicting the primitiveness of $\boldsymbol{u}_1$. □

The quantity $\mathrm{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$ gives the maximal number of positions the structures $(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1+e_2}$ and $(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_2}$ can be cyclically shifted to the left in $\boldsymbol{v}^2$, while $\mathrm{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$ gives the maximal number of positions the structures $(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1}$ and $(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)^{e_1+e_2}$ can be cyclically shifted to the right. In [8], the following lemma limiting the size of $\mathrm{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) + \mathrm{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$ was given.

Lemma 2.5 ([8]). *Considering $\boldsymbol{u}_1^{e_1}\boldsymbol{u}_2\boldsymbol{u}_1^{e_1+e_2}\boldsymbol{u}_2\boldsymbol{u}_1^{e_2}$, where $\boldsymbol{u}_1$ is primitive and $\boldsymbol{u}_2$ is a non-empty proper prefix of $\boldsymbol{u}_1$, $e_1 \geq e_2 \geq 1$, and $\overline{\boldsymbol{u}}_2$ a suffix of $\boldsymbol{u}_1$ so that $\boldsymbol{u}_1 = \boldsymbol{u}_2\overline{\boldsymbol{u}}_2$, then $lcs(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) + lcp(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) \leq |\boldsymbol{u}_1| - 2$.*

Example

$\boldsymbol{u} = aababa,\ \boldsymbol{u}_1 = aabab,\ \boldsymbol{u}_2 = a,\ \overline{\boldsymbol{u}}_{\boldsymbol{2}} = abab$

$\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = aabab,\ \overline{\boldsymbol{u}}_2\boldsymbol{u}_2 = ababa$

$\text{lcs} = 0,\ \text{lcp} = 1,\ lcs(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) + \text{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = 1 \leq |\boldsymbol{u}_1| - 2 = 3$

In fact, in [8] the inversion factor is defined more generally as any factor $\overline{\boldsymbol{w}}\boldsymbol{w}\boldsymbol{w}\overline{\boldsymbol{w}}$ of $\boldsymbol{v}^2$ such that $|\boldsymbol{w}| = |\boldsymbol{u}_2|$ and $|\overline{\boldsymbol{w}}| = |\overline{\boldsymbol{u}}_{\boldsymbol{2}}|$ and a stronger result is given (re-phrased in the terminology of this thesis):

Lemma 2.6 ([8]). *Consider a double square $DS(\boldsymbol{u}, \boldsymbol{v} : \boldsymbol{u}_1, \boldsymbol{u}_2, e_1, e_2)$ with a non-empty $\boldsymbol{u}_2$ and let $p = lcp(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$ and $s = lcs(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$. Then any inversion factor in $\boldsymbol{v}^2$ is either $R_i(\mathbb{IF})$ or $R_{-j}(\mathbb{IF})$ for some $i \in 0..p$ or some $j \in 0..s$. Moreover, every $R_i(\mathbb{IF})$ or $R_{-j}(\mathbb{IF})$ appear exactly twice in $\boldsymbol{v}^2$ exactly a distance $|\boldsymbol{v}|$ apart for every $i \in 0..p$ and every $j \in 0..s$.*

Example

$\boldsymbol{u} = aaabbaaabbaa,\ \boldsymbol{u}_1 = aaabb,\ \boldsymbol{u}_2 = aa,\ \overline{\boldsymbol{u}}_2 = abb,\ e_1 = 2,\ e_2 = 2$

$\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = aaabb,\ \overline{\boldsymbol{u}}_2\boldsymbol{u}_2 = abbaa$

$\mathbb{IF} = \overline{\boldsymbol{u}}_2\boldsymbol{u}_2\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = abbaaaaabb$

$s = \text{lcs} = 0,\ p = \text{lcp} = 1,\ R_1(\mathbb{IF}) = bbaaaaabba$

$\boldsymbol{v}^2 = (aaabbaaa\underline{bbaaaaabba}aabb)(aaabbaaa\underline{bbaaaaabba}aabb)$, where underlined letters are the two occurrences of $R_1(\mathbb{IF})$, the distance is $|aabbaaabbaaabbaaaaabba| = 22 = |\boldsymbol{v}|$

2.3 Rare Factors in Balanced Double Squares

In this section, we discuss other factors with limited number of occurrences which we generically refer to as *rare factors* or *rarely occurring factors*. In particular, we introduce *right inversion subfactor* ($\mathbb{RIS}$) and *left inversion subfactor* ($\mathbb{LIS}$) in balanced double squares. These rare factors are half the size of the inversion factors, and, since they are shorter, they provide a stronger restriction on the positions and lengths of a third square. We will use them significantly in the next chapter for the New Periodicity Lemma.

In [19], Thierry discusses the core of the period interrupt, a very similar concept to the one we introduce here as $\mathbb{RIS}$ and $\mathbb{LIS}$. For a double square $DS(\boldsymbol{u}, \boldsymbol{v} : \boldsymbol{u}_1, \boldsymbol{u}_2, e_1, e_2)$ with $|\boldsymbol{u}_2| > 0$, both $\mathbb{RIS}$ and $\mathbb{LIS}$ have only two occurrences in a significant portion of $\boldsymbol{v}^2$.

Lemma 2.7 will be used to prove rare occurrences of $\mathbb{RIS}$ and $\mathbb{LIS}$ in a balanced double square. It says that if a substring $\boldsymbol{u}$ of a string $\boldsymbol{x}$ and its rotation $\boldsymbol{u'}$ completely overlap except for one symbol, then $\boldsymbol{u}$ can be cyclically shifted one position to the right, or, equivalently, $\boldsymbol{u'}$ can be cyclically shifted one position to the left.

Lemma 2.7. *Consider a string* $\boldsymbol{x} = \boldsymbol{x}[1..n]$. *If the substrings* $\boldsymbol{x}[i..i+k]$ *and* $\boldsymbol{x}[i+1..i+1+k]$ *are conjugates, then* $\boldsymbol{x}[i] = \boldsymbol{x}[i+k+1]$.

Proof. Since $\boldsymbol{x}[i..i+k]$ and $\boldsymbol{x}[i+1..i+1+k]$ are conjugates, the frequency of the alphabet symbols in both must be the same. Let $\boldsymbol{x}[i] = a$. Then

$\boldsymbol{x}[i..i+k]$ must have the same number of a's as $\boldsymbol{x}[i+1..i+1+k]$, and so $\boldsymbol{x}[i+1+k] = a$. □

Example

$$\begin{aligned} \boldsymbol{x}[1..n+1] &= a\underbrace{babbab}a \\ \boldsymbol{x}[1..n] &= a\underbrace{babbab} \\ \boldsymbol{x}[2..n+1] &= \underbrace{babbab}a \end{aligned}$$

Note that Lemma 2.7 does not hold if $\boldsymbol{u}$ and $\boldsymbol{u'}$ overlap less. For example:

Example

$$\begin{aligned} \boldsymbol{x}[1..n+1] &= ab\underbrace{ababa}ba \\ \boldsymbol{x}[1..n-1] &= ab\underbrace{ababa} \\ \boldsymbol{x}[3..n+1] &= \underbrace{ababa}ba \end{aligned}$$

neither $\boldsymbol{x}[1..n-1]$ can be cyclically shifted to the right nor $\boldsymbol{x}[3..n+1]$ to the left, yet $\boldsymbol{x}[1..n-1]$ and $\boldsymbol{x}[3..n+1]$ are conjugates.

Note: From the definition below to the end of this chapter, we are going to index from 0 to keep consistent with the publication *The New Periodicity Lemma Revisited* [3].

Definition 2.8. *Consider a double square* $DS(\boldsymbol{u}, \boldsymbol{v} : \boldsymbol{u}_1, \boldsymbol{u}_2, e_1, e_2)$ *with nonempty* $\boldsymbol{u}_2$. *The* right inversion subfactors (*or* $\mathbb{RIS}$) *are two substrings*

$\boldsymbol{x}[i..j]$ *and* $\boldsymbol{x}[i+|\boldsymbol{v}|, j+|\boldsymbol{v}|]$*, both of length* $|\boldsymbol{u}_1|$*, where*

$$i = (e_1 - 1)|\boldsymbol{u}_1| + |\boldsymbol{u}_2| + 1 + lcp(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$$
$$j = e_1|\boldsymbol{u}_1| + |\boldsymbol{u}_2| + lcp(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$$

The left inversion subfactors (*or* $\mathbb{LIS}$) *are two substrings* $\boldsymbol{x}[i..j]$ *and* $\boldsymbol{x}[i+|\boldsymbol{v}|, j+|\boldsymbol{v}|]$*, both of length* $|\boldsymbol{u}_1|$*, where*

$$i = e_1|\boldsymbol{u}_1| + |\boldsymbol{u}_2| - 1 - lcs(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$$
$$j = (e_1 + 1)|\boldsymbol{u}_1| + |\boldsymbol{u}_2| - 2 - lcs(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$$

Here are two examples of $\mathbb{RIS}$ and $\mathbb{LIS}$:

Example 1

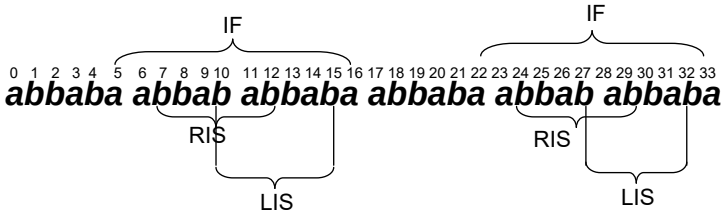
$\boldsymbol{u}_2 = abbab,\ \overline{\boldsymbol{u}}_2 = a,\ \boldsymbol{u}_1 = abbaba,\ e_1 = e_2 = 1$

$\mathrm{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = 1,\ \mathrm{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = 0$

$\boldsymbol{u} = \boldsymbol{u}_1\boldsymbol{u}_2 = abbabaabbab,\ \boldsymbol{v} = \boldsymbol{u}_1\boldsymbol{u}_2\boldsymbol{u}_1 = abbabaabbababbaba$

$\mathbb{IF} = \overline{\boldsymbol{u}}_2\boldsymbol{u}_2\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = a\,abbab\,abbab\,a$

$\mathbb{RIS} = bbabab$ and $\mathbb{LIS} = babbab$



$\mathbb{RIS}$:

$i = (e_1 - 1)|\boldsymbol{u}_1| + |\boldsymbol{u}_2| + 1 + \mathrm{lcp}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_\boldsymbol{2}, \overline{\boldsymbol{u}}_\boldsymbol{2}\boldsymbol{u_2}) = 0 + 5 + 1 + 1 = 7$ and

$i + |\boldsymbol{v}| = 7 + 17 = 24$

$j = e_1|\boldsymbol{u}_1| + |\boldsymbol{u}_2| + \text{lcp}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_\boldsymbol{2}, \overline{\boldsymbol{u}}_\boldsymbol{2}\boldsymbol{u_2}) = 1 \times 6 + 5 + 1 = 12$ and $j + |\boldsymbol{v}| = 12 + 17 = 29$

$\mathbb{LIS}$:

$i = e_1|\boldsymbol{u}_1| + |\boldsymbol{u}_2| - 1 - \text{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_\boldsymbol{2}, \overline{\boldsymbol{u}}_\boldsymbol{2}\boldsymbol{u_2}) = 1 \times 6 + 5 - 1 - 0 = 10$ and $i + |\boldsymbol{v}| = 10 + 17 = 27$

$j = (e_1 + 1)|\boldsymbol{u}_1| + |\boldsymbol{u}_2| - 2 - \text{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_\boldsymbol{2}, \overline{\boldsymbol{u}}_\boldsymbol{2}\boldsymbol{u_2}) = 2 \times 6 + 5 - 2 - 0 = 15$ and $j + |\boldsymbol{v}| = 15 + 17 = 32$

Example 2

$\boldsymbol{u}_2 = abca,\ \overline{\boldsymbol{u}}_2 = a,\ \boldsymbol{u}_1 = abcaa,\ e_1 = 2,\ e_2 = 1$

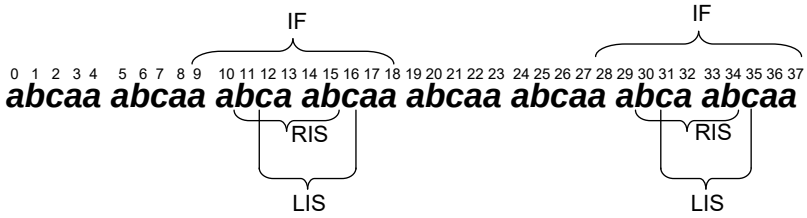
$\text{lcp}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_\boldsymbol{2}, \overline{\boldsymbol{u}}_\boldsymbol{2}\boldsymbol{u_2}) = 1,\ \text{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_\boldsymbol{2}, \overline{\boldsymbol{u}}_\boldsymbol{2}\boldsymbol{u_2}) = 1$

$\boldsymbol{u} = \boldsymbol{u}_1^2\boldsymbol{u}_2 = abcaaabcaaabca,$

$\boldsymbol{v} = \boldsymbol{u}_1^2\boldsymbol{u}_2\boldsymbol{u}_1 = abcaaabcaaabcaabcaa$

$\mathbb{IF} = \overline{\boldsymbol{u}}_2\boldsymbol{u}_2\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = a\,abca\,abca\,a$

$\mathbb{RIS} = bcaab$ and $\mathbb{LIS} = caabc$



$\mathbb{RIS}$:

$i = (e_1 - 1)|\boldsymbol{u}_1| + |\boldsymbol{u}_2| + 1 + \text{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = 1 \times 5 + 4 + 1 + 1 = 11$ and $i + |\boldsymbol{v}| = 11 + 19 = 30$

$j = e_1|\boldsymbol{u}_1| + |\boldsymbol{u}_2| + \text{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = 2 \times 5 + 4 + 1 = 15$ and $j + |\boldsymbol{v}| =$

$15 + 19 = 34$

$\mathbb{LIS}$:

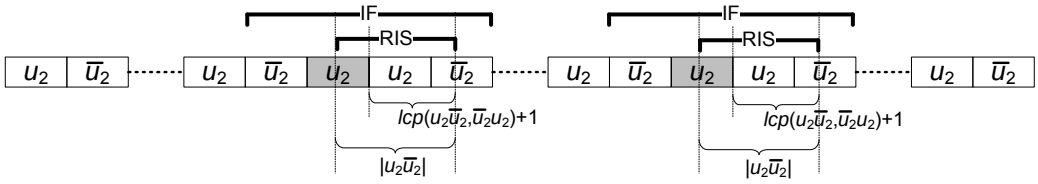
$i = e_1|\boldsymbol{u}_1| + |\boldsymbol{u}_2| - 1 - \text{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_\boldsymbol{2}, \overline{\boldsymbol{u}}_\boldsymbol{2}\boldsymbol{u_2}) = 2 \times 5 + 4 - 1 - 1 = 12$ and $i + |\boldsymbol{v}| = 12 + 19 = 31$

$j = (e_1 + 1)|\boldsymbol{u}_1| + |\boldsymbol{u}_2| - 2 - \text{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_\boldsymbol{2}, \overline{\boldsymbol{u}}_\boldsymbol{2}\boldsymbol{u_2}) = 3 \times 5 + 4 - 2 - 1 = 16$ and $j + |\boldsymbol{v}| = 16 + 19 = 35$

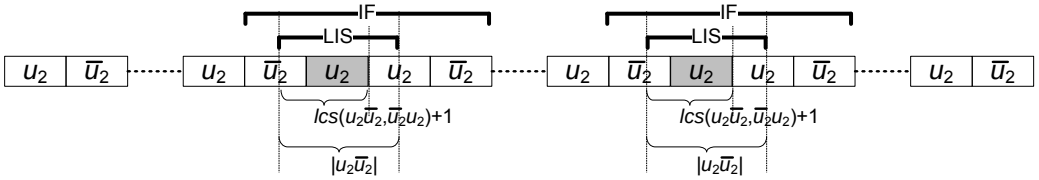
It is useful to view $\mathbb{RIS}$ in the following way: let $\boldsymbol{x}[i..j]$ be the maximum right cyclic shift of the left half of the first $\mathbb{IF}$, respectively of the left half of the second $\mathbb{IF}$. Then $\mathbb{RIS}$ is $\boldsymbol{x}[i+1, j+1]$. According to Lemma 2.7, if $\boldsymbol{x}[i, j]$ and $\boldsymbol{x}[i+1, j+1]$ were conjugate, $\boldsymbol{x}[i+1, j+1]$ would have to be a right cyclic shift of $\boldsymbol{x}[i, j]$, which is a contradiction. Thus $\boldsymbol{x}[i+1, j+1]$ cannot be conjugate with $\boldsymbol{x}[i, j]$ and hence a copy of it can occur neither in ${\boldsymbol{u}_1}^{e_1+e_2}$, nor in ${\boldsymbol{u}_1}^{e_2}$.

Similarly for $\mathbb{LIS}$: let $\boldsymbol{x}[i, j]$ be the maximum left cyclic shift of the right half of the first $\mathbb{IF}$, respectively of the right half of the second $\mathbb{IF}$. Then $\mathbb{LIS}$ is $\boldsymbol{x}[i-1, j-1]$. Again, according to Lemma 2.7, if $\boldsymbol{x}[i, j]$ and $\boldsymbol{x}[i-1, j-1]$ were conjugate, $\boldsymbol{x}[i-1, j-1]$ would have to be a left cyclic shift of $\boldsymbol{x}[i, j]$, which is a contradiction. Thus $\boldsymbol{x}[i-1, j-1]$ cannot be conjugate with $\boldsymbol{x}[i, j]$ and hence a copy of it can occur neither in ${\boldsymbol{u}_1}^{e_1}$, nor in ${\boldsymbol{u}_1}^{e_1+e_2}$.

For a better understanding, we illustrate the two natural occurrences of $\mathbb{RIS}$ in the following diagram:



and the two natural occurrences of $\mathbb{LIS}$ as follows:



The purpose of presenting $\mathbb{RIS}$ and $\mathbb{LIS}$ is to have some structures similar to $\mathbb{IF}$'s but shorter, as shorter factors provide a stronger constraint about where the third square can occur in a balanced double square. This will be used in the proof of a reformulated New Periodicity Lemma in Chapter 3.

Let us explain what we mean by "stronger constraint": when using just $\mathbb{IF}$'s to prove where we can have a third square in a balanced double square $DS(\boldsymbol{u}, \boldsymbol{v})$, the argument follows this line: suppose we have a third square $\boldsymbol{w^2}$ such that $\boldsymbol{w}_{[1]}$ contains the first $\mathbb{IF}$, then necessarily $\boldsymbol{w}_{[2]}$ must contain the second $\mathbb{IF}$ (and so $|\boldsymbol{w}| = |\boldsymbol{v}|$), or $\boldsymbol{w}_{[1]}$ must contain a part of the second $\mathbb{IF}$ (and so $|\boldsymbol{w}| > |\boldsymbol{v}|$); i.e. it forces the square $\boldsymbol{w}^2$ to be big and hence a smaller square cannot exist. However, this argument fails when $\boldsymbol{w}_{[1]}$ just contains a part of the first $\mathbb{IF}$. Nevertheless, $\boldsymbol{w}_{[1]}$ may still contain the first $\mathbb{RIS}$ or $\mathbb{LIS}$ and it would force the size of $\boldsymbol{w}^2$ in the same way.

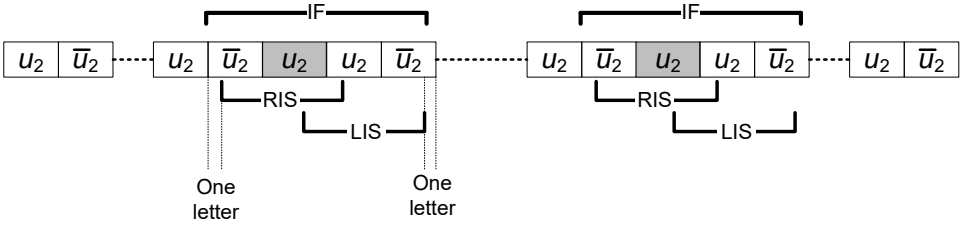
In the last part of this chapter, we will analyze the mutual positions of $\mathbb{RIS}$ and $\mathbb{LIS}$ by describing the extreme cases when $\mathbb{RIS}$ and $\mathbb{LIS}$ are almost

disjoint and when they completely overlap. Let $\boldsymbol{R}_1$ denote the maximal right cyclic shift of the left half of $\mathbb{IF}_1$, $\boldsymbol{R}_2$ denote the maximal right cyclic shift of the left half of $\mathbb{IF}_2$, while $\boldsymbol{L}_1$ denote the maximal left cyclic shift of the right half of $\mathbb{IF}_1$, and $\boldsymbol{L}_2$ denote the maximal left cyclic shift of the right half of $\mathbb{IF}_2$. We already established that the maximal right cyclic shift of the left half of $\mathbb{IF}$ is determined by $\text{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$ and the maximal left cyclic shift of the right half of $\mathbb{IF}$ is determined by $\text{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$. Lemma 2.5 bounds the size of the shifts as $\text{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) + \text{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) \leq |\boldsymbol{u}_1| - 2$.

Note: Since we get the first pair of $\mathbb{RIS}$ and $\mathbb{LIS}$ in the same way as the second pair of $\mathbb{RIS}$ and $\mathbb{LIS}$, we only discuss how we get the first pair in the following cases.

Case 1: $\text{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = \text{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = 0$

This is the case when $\mathbb{RIS}$ and $\mathbb{LIS}$ are almost disjoint. In this case, $\boldsymbol{R}_1$ is exactly the left half of $\mathbb{IF}_1$, and if we shift $\boldsymbol{R}_1$ one position to the right, then we get $\mathbb{RIS}$, and $\boldsymbol{L}_1$ is exactly the right half of $\mathbb{IF}_2$, and if we shift $\boldsymbol{L}_1$ one position to the left then, we get $\mathbb{LIS}$.



Example 1

$\boldsymbol{u}_2 = abc,\ \overline{\boldsymbol{u}}_2 = cb,\ \boldsymbol{u}_1 = abccb,\ e_1 = 2,\ e_2 = 1$

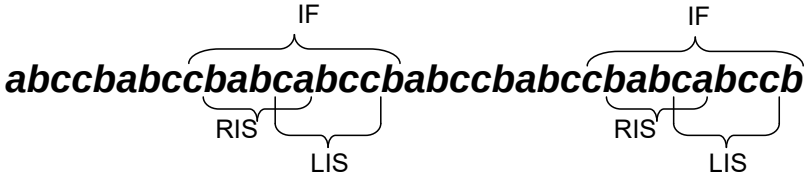
$\text{lcp}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = \text{lcp}(abccb, cbabc) = 0$

$\text{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = \text{lcs}(abccb, cbabc) = 0$

$\boldsymbol{u} = \boldsymbol{u}_1^2\boldsymbol{u}_2 = abccbabccbabc,\ \boldsymbol{v} = \boldsymbol{u}_1^2\boldsymbol{u}_2\boldsymbol{u}_1 = abccbabccbabcabccb$

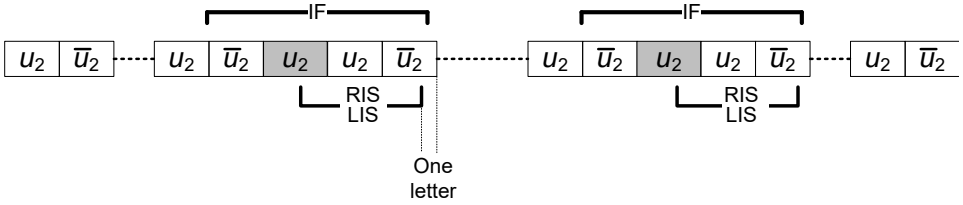
$\mathbb{IF} = \overline{\boldsymbol{u}}_2\boldsymbol{u}_2\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = cbabcabccb$

$\mathbb{RIS} = babca,\ \mathbb{LIS} = cabcc$



Case 2: $\textbf{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = |\boldsymbol{u}_1| - 2,\ \textbf{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = 0$

This is the case when $\mathbb{RIS}$ and $\mathbb{LIS}$ completely overlap. In this case, $\boldsymbol{L}_1$ is exactly the right half of $\mathbb{IF}_1$, and if we shift $\boldsymbol{L}_1$ one position to the left, then we get $\mathbb{LIS}$, and $\mathbb{RIS}$ is at the furthest position to the right as it can be.



Example 2

$\boldsymbol{u}_2 = ab,\ \overline{\boldsymbol{u}}_2 = aba,\ \boldsymbol{u}_1 = ababa,\ e_1 = 2,\ e_2 = 1$

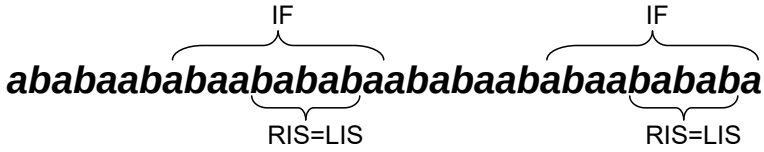
$\text{lcp}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}) = \text{lcp}(ababa, abaab) = 3$

$\text{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}) = \text{lcs}(ababa, abaab) = 0$

$\boldsymbol{u} = \boldsymbol{u}_1^2\boldsymbol{u}_2 = ababaababaab,\ \boldsymbol{v} = \boldsymbol{u}_1^2\boldsymbol{u}_2\boldsymbol{u}_1 = ababaababaababababa$

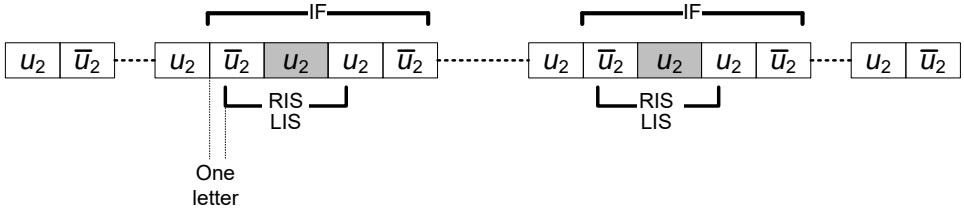
$\mathbb{IF} = \overline{\boldsymbol{u}}_2\boldsymbol{u}_2\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = abaababaaba$

$\mathbb{RIS} = \mathbb{LIS} = babab$



Case 3: $\mathbf{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = 0,\ \mathbf{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = |\boldsymbol{u}_1| - 2$

This, too, is a case when $\mathbb{RIS}$ and $\mathbb{LIS}$ completely overlap. In this case, $\boldsymbol{R}_1$ is exactly the left half of $\mathbb{IF}_1$, and if we shift $\boldsymbol{R}_1$ one position to the right then we get $\mathbb{RIS}$, and $\mathbb{LIS}$ is at the furthest position to the left as it can be.



Example 3

$\boldsymbol{u}_2 = baa,\ \overline{\boldsymbol{u}}_2 = a,\ \boldsymbol{u}_1 = baaa,\ e_1 = 2,\ e_2 = 2$

$\mathrm{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = \mathrm{lcp}(baaa, abaa) = 0$

$\mathrm{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = \mathrm{lcs}(baaa, abaa) = 2$

$\boldsymbol{u} = \boldsymbol{u}_1^2\boldsymbol{u}_2 = baaabaaabaa,\ \boldsymbol{v} = \boldsymbol{u}_1^2\boldsymbol{u}_2\boldsymbol{u}_1^2 = baaabaaabaabaaabaaa$

$\mathbb{IF} = \overline{\boldsymbol{u}}_2\boldsymbol{u}_2\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = abaabaaa$

$\mathbb{RIS} = \mathbb{LIS} = baab$

Case 4: $\mathbf{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) \neq 0$, $\mathbf{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) \neq 0$, $\mathbf{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) + \mathbf{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2) = |\boldsymbol{u}_1| - 2$

This, too, is a case when $\mathbb{RIS}$ and $\mathbb{LIS}$ completely overlap.

Example 4

$\boldsymbol{u}_2 = abb,\ \overline{\boldsymbol{u}}_2 = ab,\ \boldsymbol{u}_1 = abbab,\ e_1 = 2,\ e_2 = 2$

$\mathrm{lcp}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_\boldsymbol{2}, \overline{\boldsymbol{u}}_\boldsymbol{2}\boldsymbol{u_2}) = \mathrm{lcp}(abbab, ababb) = 2$

$\mathrm{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_\boldsymbol{2}, \overline{\boldsymbol{u}}_\boldsymbol{2}\boldsymbol{u_2}) = \mathrm{lcs}(abbab, ababb) = 1$

$\boldsymbol{u} = \boldsymbol{u}_1^2\boldsymbol{u}_2 = abbababbababb,\ \boldsymbol{v} = \boldsymbol{u}_1^2\boldsymbol{u}_2\boldsymbol{u}_1^2 = abbababbababbabbababbab$

$\mathbb{IF} = \overline{\boldsymbol{u}}_2\boldsymbol{u}_2\boldsymbol{u}_2\overline{\boldsymbol{u}}_2 = ababbabbab$

$\mathbb{RIS} = \mathbb{LIS} = bbabb$

Chapter 3

Application of canonical factorization to New Periodicity Lemma

In this chapter based on the paper [3], we discuss an application of the canonical factorization of balanced double squares to the New Periodicity Lemma, [9], which shows that the occurrence of two special squares at a position i in a string, necessarily precludes the occurrence of other squares of specific period in a specific neighbourhood of i. The proof of this lemma is complex, breaking down into 14 subcases, and requires very strong assumptions, namely that the shorter of the two squares be regular and the bigger one primitive.

Before proving a strengthen version of Lemma 3.1, we state the original

New Periodicity Lemma:

Lemma 3.1 ([9], New Periodicity Lemma)**.** *Let a regular* $\boldsymbol{u}^2$ *and a primitively rooted* $\boldsymbol{v}^2$ *be prefixes of a string* $\boldsymbol{x}$ *so that* $|\boldsymbol{u}| < |\boldsymbol{v}| < 2|\boldsymbol{u}|$*. Then for all integers* k *and* w *such that* $0 \leq k < |\boldsymbol{v}| - |\boldsymbol{u}| < w < |\boldsymbol{v}|$, $w \neq |\boldsymbol{u}|$, $\boldsymbol{x}[k{+}1..k{+}2w]$ *is not a square.*

Example 1

$\boldsymbol{u} = aba$, so $\boldsymbol{u}^2$ is regular

$\boldsymbol{v} = abaab$ is primitive, so $\boldsymbol{v}^2$ is primitively rooted.

$|u| = 3,\ |v| = 5,\ |v| - |u| = 2$

$\boldsymbol{x} = abaababaab\cdots$

$0 \leq k < 2 < w < 5$

e.g. when $k = 0$ and $w = 4$, then $\boldsymbol{x}[1..8] = aba|baba$ is not a square, or

e.g. when $k = 1$ and $w = 3$, then $\boldsymbol{x}[2..7] = baa|bab$ is not a square, and so on for all possible values of k and all possible values of w.

Example 2

$\boldsymbol{u} = abbaabb$ and so $\boldsymbol{u}^2$ is regular

$\boldsymbol{v} = abbaabbabba$ is primitive and so $\boldsymbol{v}^2$ is primitively rooted

$|u| = 7,\ |v| = 11,\ |v| - |u| = 4$

$\boldsymbol{x} = abbaabbabbaabbaabbabba\cdots$

$0 \leq k < 4 < w < 11$

e.g. when $k = 0$ and $w = 5$, then $\boldsymbol{x}[1..10] = abbaa|bbabb$ is not a square, or e.g. when $k = 0$ and $w = 6$, then $\boldsymbol{x}[1..12] = abbaab|babbaa$ is not a

square, or e.g. when $k = 1$ and $w = 5$, then $\boldsymbol{x}[2..11] = bbaab|babba$ is not a square, or e.g. when $k = 1$ and $w = 8$, then $\boldsymbol{x}[2..17] = bbaabbab|baabbaab$ is not a square, and so on for all possible values of k and all possible values of w.

Example 3

$\boldsymbol{u} = abcaabc$ and so $\boldsymbol{u}^2$ is regular

$\boldsymbol{v} = abcaabcabca$ is primitive and so $\boldsymbol{v}^2$ is primitively rooted

$|u| = 7,\ |v| = 11,\ |v| - |u| = 4$

$\boldsymbol{x} = abcaabcabcaabcaabcabca\cdots$

$0 \leq k < 4 < w < 11$

e.g. when $k = 0$ and $w = 5$, then $\boldsymbol{x}[1..10] = abcaa|bcabc$ is not a square, or e.g. when $k = 0$ and $w = 6$, then $\boldsymbol{x}[1..12] = abcaab|cabcaa$ is not a square, or e.g. when $k = 1$ and $w = 5$, then $\boldsymbol{x}[2..11] = bcaab|cabca$ is not a square, or e.g. when $k = 1$ and $w = 8$, then $\boldsymbol{x}[2..17] = bcaabcab|caabcaab$ is not a square, and so on for all possible values of k and all possible values of w.

In fact, the requirement that $\boldsymbol{u}^2$ be regular seems unnecessary. Here we list a few examples that satisfy the conclusion of Lemma 3.1 yet violate the regularity condition:

Example 4

$\boldsymbol{u} = abbababbababbababb$, $\boldsymbol{u}^2$ is not regular as for instance $(abbab)^2$ is a prefix of $\boldsymbol{u}$

$\boldsymbol{v} = abbababbababbababbababbababbab$ is primitive

$|u| = 18,\ |v| = 28,\ |v| - |u| = 10$

$\boldsymbol{x} = abbababbababbababbabbababbababbababbababbababbababbababbab\cdots$

$0 \leq k < 10 < w < 28$

e.g. when $k = 0$ and $w = 11$, then $\boldsymbol{x}[1..22] = abbababbaba|bbababbabba$ is not a square, or e.g. when $k = 0$ and $w = 12$, then $\boldsymbol{x}[1..24] = abbababbabab|bababbabbaba$ is not a square, and so on for all possible values of k and all possible values of w.

Example 5

$\boldsymbol{u} = aabaaab$, $\boldsymbol{u}^2$ is not regular as for instance a^2 is a prefix of $\boldsymbol{u}$

$\boldsymbol{v} = aabaaabaaba$ is primitive

$|u| = 7,\ |v| = 11,\ |v| - |u| = 4$

$\boldsymbol{x} = aabaaabaabaaabaaabaaba\cdots$

$0 \leq k < 4 < w < 11$

e.g. when $k = 3$ and $w = 6$, then $\boldsymbol{x}[4..15] = aaabaa|baaaba$ is not a square, or e.g. when $k = 3$ and $w = 7$, then $\boldsymbol{x}[4..17] = aaabaab|aaabaaa$ is not a square, and so on for all possible values of k and all possible values of w.

Example 6

$\boldsymbol{u} = aaba$, $\boldsymbol{u}^2$ is not regular as for instance a^2 is a prefix of $\boldsymbol{u}$

$\boldsymbol{v} = aabaaabaaba$ is primitive

$|u| = 7,\ |v| = 11,\ |v| - |u| = 4$

$\boldsymbol{x} = aabaaabaabaaabaaabaaba\cdots$

$0 \leq k < 4 < w < 11$

e.g. when $k = 3$ and $w = 6$, then $\boldsymbol{x}[4..15] = aaabaa|baaaba$ is not a square, or e.g. when $k = 3$ and $w = 7$, then $\boldsymbol{x}[4..17] = aaabaab|aaabaaa$ is not a square, and so on for all possible values of k and all possible values of w.

The Examples 4, 5, and 6 shows that the regularity of $\boldsymbol{u}^2$ is maybe an unnecessary assumption. Since $\boldsymbol{u}$ and $\boldsymbol{v}$ are proportional squares, they form a balanced double square. Thus, the required assumption of primitivness of $\boldsymbol{v}$ is redundant by Observation 2.3: the fact that $\boldsymbol{u}^2$ is regular already forces the primitivness of $\boldsymbol{v}$. Also note that the regularity of u^2 necessarily implies that in the canonical factorization of a double square $DS(\boldsymbol{u}, \boldsymbol{v} : \boldsymbol{u}_1, \boldsymbol{u}_2, e_1, e_2)$ the exponents are all equal 1, i.e. $e_1 = e_2 = 1$. Let us see what Lemma 3.1 says in the terms of the canonical factorization: so we have a double square $DS(\boldsymbol{u}, \boldsymbol{v} : \boldsymbol{u}_1, \boldsymbol{u}_2, 1, 1)$. As always, let $\overline{\boldsymbol{u}}_2$ be a suffix of $\boldsymbol{u}_1$ such that $\boldsymbol{u}_1 = \boldsymbol{u}_2\overline{\boldsymbol{u}}_2$. The canonical factorization thus gives

$$\boldsymbol{v}^2 = \overbrace{(\boldsymbol{u}_2 \underbrace{\overline{\boldsymbol{u}}_2)\boldsymbol{u}_2(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)}_{\mathbb{IF}_1}(\boldsymbol{u}_2 \underbrace{\overline{\boldsymbol{u}}_2)\boldsymbol{u}_2(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)}_{\mathbb{IF}_2}}^{\boldsymbol{u}_1}$$

Thus, $0 \leq k < |\boldsymbol{v}| - |\boldsymbol{u}| < w < |\boldsymbol{v}|$ gives $0 \leq k < |\boldsymbol{u}_1| < w < |\boldsymbol{v}|$. Therefore, Lemma 3.1 says that there is no square $\boldsymbol{w}^2$ that starts in the first $\boldsymbol{u}_1$ and such that $|\boldsymbol{u}_1| < |\boldsymbol{w}| < |\boldsymbol{v}|$ and $|\boldsymbol{w}| \neq |\boldsymbol{u}|$.

Now, let us consider a square $\boldsymbol{w}^2$ such that $|\boldsymbol{u}_1| < |\boldsymbol{w}| < |\boldsymbol{v}|$ and $|\boldsymbol{w}| \neq |\boldsymbol{u}|$. We want to show that this is not possible just from the properties of the canonical factorization. If for instance $\boldsymbol{w}$ starts in the first $\boldsymbol{u}_2$ and ends in

the fourth $\boldsymbol{u}_2$, then $\boldsymbol{w}_{[1]}$ contains fully the $\mathbb{IF}_1$, so $\boldsymbol{w}_{[2]}$ has to contain $\mathbb{IF}_2$, and so $|\boldsymbol{w}| \geq |\boldsymbol{v}|$, a contradiction. If $\boldsymbol{w}$ ends in the second $\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ we cannot argue using $\mathbb{IF}$, but still knowing that $\boldsymbol{u}_2\overline{\boldsymbol{u}}_2$ is primitive and also all its rotations are primitive, using the Synchronization principle Lemma 1.31 can be applied to obtain a contradiction.

Almost all possible cases for $\boldsymbol{w}^2$ except two can be easily shown impossible using only the properties of the canonical factorization. Thus, it was clear that the canonical factorization could not only provide a significantly simplified proof of Lemma 3.1, but also could allow to significantly reduce the assumption that $\boldsymbol{u}$ must be regular. This was the motivation for the research culminating in Theorem 3.2. To increase the power of the use of inversion factors for the use in the proof of Theorem 3.2, the right and left inversion subfactors $\mathbb{RI}\$$ and $\mathbb{LI}\$$ were studied, see Chapter 2. The main result in this chapter is the following Theorem 3.2.

Theorem 3.2. *Consider a balanced double square $DS(\boldsymbol{u}, \boldsymbol{v})$ and let $\boldsymbol{u'}$ be a suffix of $\boldsymbol{v}$ such that $\boldsymbol{v} = \boldsymbol{u}\boldsymbol{u'}$. Let $\boldsymbol{w}^2$ be any square that is a substring of $\boldsymbol{v}^2$. Then exactly one of the following mutually exclusive cases holds:*

(*a*) $\boldsymbol{w} = \boldsymbol{v}$, *or*

(*b*) $|\boldsymbol{w}| < |\boldsymbol{u}|$, *or*

(*c*) $|\boldsymbol{u}| \leq |\boldsymbol{w}| < |\boldsymbol{v}|$ *and the primitive root of $\boldsymbol{w}$ is a conjugate of the primitive root of $\boldsymbol{u'}$.*

Before we prove the theorem, let us discuss how it relates to the original Lemma 3.1. As mentioned above, for a very specific double square, the

lemma forbids squares starting in the first $\boldsymbol{u}_1$ of lengths bigger than $|\boldsymbol{u}_1|$ but smaller than $|\boldsymbol{v}|$ with a possible exception of length $|\boldsymbol{u}|$. Theorem 3.2 for such a double square forbids squares starting anywhere if their length is bigger than $|\boldsymbol{u}|$ and smaller than $|\boldsymbol{v}|$. So the "forbidding power" of the theorem is slightly less than that of the lemma with respect to the sizes of $\boldsymbol{w}^2$; however it covers a larger range of possible starts for "forbidden" squares (anywhere instead of in the first $\boldsymbol{u}_1$), and above all, it applies to **all balanced double squares without any additional conditions or constraints**. Now we can proceed with the proof of Theorem 3.2.

Proof. If $|w| \geq |v|$, then $w = v$ since w^2 is a substring of v^2, and thus case (a) holds. Hence, for the remainder of the proof we can assume that $|w| < |v|$. Since double square $DS(\boldsymbol{u}, \boldsymbol{v})$ is balanced, it admits the canonical factorization: $DS(\boldsymbol{u}, \boldsymbol{v} : \boldsymbol{u}_1, \boldsymbol{u}_2, e_1, e_2)$, by Lemma 2.1. Then $\boldsymbol{u'} = \boldsymbol{u}_1^{e_2}$ and the primitive root of $\boldsymbol{u'}$ is $\boldsymbol{u}_1$.

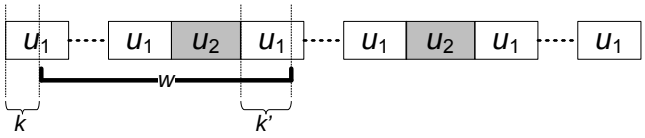
We first deal with the case $|\boldsymbol{u}_2| = 0$. By Lemma 2.1, $\boldsymbol{u}_1$ is the primitive root of $\boldsymbol{u} = \boldsymbol{u}_1^{e_1}$ and of $\boldsymbol{v} = \boldsymbol{u}_1^{e_1+e_2}$ with $e_1 > e_2 \geq 1$. If case (b) does not hold, we must have $|\boldsymbol{w}| \geq |\boldsymbol{u}| = |\boldsymbol{u}_1^{e_1}| > |\boldsymbol{u}_1|$. Thus $\boldsymbol{w}^2$ and $\boldsymbol{u}_1^{2e_1+2e_2}$ have a common factor of size $|\boldsymbol{u}_1|+|\boldsymbol{w}|$, so that by Common factor Lemma 1.32, the primitive root of $\boldsymbol{w}$ is a conjugate of $\boldsymbol{u}_1$, i.e. case (c) holds.

Now let us deal with case $|\boldsymbol{u}_2| > 0$. Suppose that (b) does not hold and so $|u| \leq |w| < |v|$.

Let us assume that there is a square $\boldsymbol{w}^2$ starting in $\boldsymbol{u}_1^{e_1}$ such that $|\boldsymbol{w}| > |\boldsymbol{u}|$. Since for $|\boldsymbol{w}| = |\boldsymbol{u}|$, $\boldsymbol{w}$ can only be a conjugate of $\boldsymbol{u}$, and hence the

primitive root of $\boldsymbol{w}$ must be a conjugate of the primitive root of $\boldsymbol{u}$, i.e. $\boldsymbol{u}_1$, so that (c) holds, we may suppose $|\boldsymbol{w}| > |\boldsymbol{u}|$. First note that due to the virtual left-right symmetry of the canonical factorization $\boldsymbol{u}_1^{e_1}\boldsymbol{u}_2\boldsymbol{u}_1^{e_1+e_2}\boldsymbol{u}_2\boldsymbol{u}_1^{e_2}$ (only the exponents e_1 and e_2 may differ), and to the fact that the arguments presented below can be applied either from the left or from the right, we therefore need only prove the assertion for $\boldsymbol{w}^2$ starting in $\boldsymbol{v}_{[1]}$. Several cases need to be discussed:

(1) $\boldsymbol{w}_{[1]}$ starts in the first $\boldsymbol{u}_1$ of $\boldsymbol{u}_1^{e_1}$ and ends in the first $\boldsymbol{u}_1$ of $\boldsymbol{u}_1^{e_1+e_2}$



Since $|\boldsymbol{w}| > |\boldsymbol{u}| = |\boldsymbol{u}_1^{e_1}|+|\boldsymbol{u}_2|$, $k < k'$.

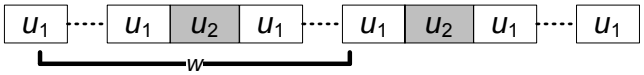
(i) $k' \leq \mathrm{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$

Then $\boldsymbol{w}_{[1]}$ has as a prefix a k-th rotation of $\boldsymbol{u}_1$ and $\boldsymbol{w}_{[2]}$ has as a prefix a k'-th rotation of $\boldsymbol{u}_1$. By the Synchronization principle Lemma 1.31, $k = k'$, a contradiction. This case is not possible.

(ii) $k' > \mathrm{lcp}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$

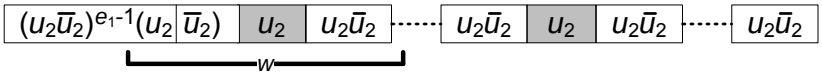
Here $\boldsymbol{w}_{[1]}$ contains the first $\mathbb{RIS}$, and so $\boldsymbol{w}_{[2]}$ must contain an occurrence of $\mathbb{RIS}$. Since $\boldsymbol{w}_{[2]}$ is a factor in $\boldsymbol{u}_1^{e_1+e_2}\boldsymbol{u}_2$, therefore by Lemma 2.8 $\boldsymbol{w}_{[2]}$ must contain the second $\mathbb{RIS}$ and so $|\boldsymbol{w}| \geq |\boldsymbol{v}|$, a contradiction. This case is not possible.

(2) $\boldsymbol{w}_{[1]}$ starts in the first $\boldsymbol{u}_1$ of $\boldsymbol{u}_1^{e_1}$ and ends past the first $\boldsymbol{u}_1$ of $\boldsymbol{u}_1^{e_1+e_2}$.



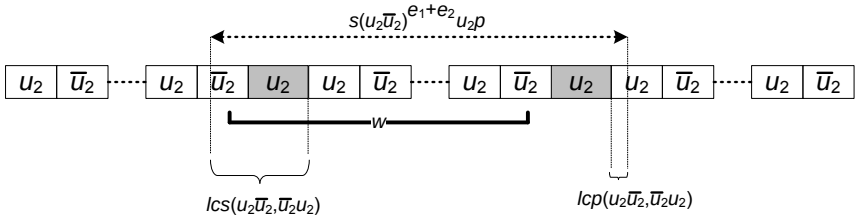
The same argument as in (1)(ii) gives $|\boldsymbol{w}| \geq |\boldsymbol{v}|$, a contradiction. This case is not possible.

(3) $\boldsymbol{w}_{[1]}$ starts in $\boldsymbol{u}_1^{e_1-1}\boldsymbol{u}_2$ but not in the first $\boldsymbol{u}_1$.



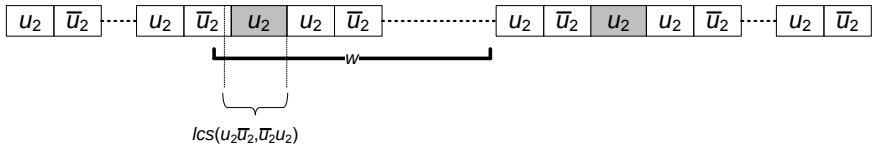
Then $e_1 > 1$ and since $|\boldsymbol{w}| > |\boldsymbol{u}| = |\boldsymbol{u}_1^{e_1}\boldsymbol{u}_2|$, $\boldsymbol{w}_{[1]}$ ends past the first $\boldsymbol{u}_1$ of $\boldsymbol{u}_1^{e_1+e_2}$. Therefore, $\boldsymbol{w}_{[1]}$ contains $\mathbb{RIS}$ and so $|\boldsymbol{w}| \geq |\boldsymbol{v}|$, i.e. this case is not possible.

(4) $\boldsymbol{w}_{[1]}$ starts in the suffix of $\overline{\boldsymbol{u}}_2\boldsymbol{u}_2$ of $\boldsymbol{u}$ whose length $\leq \text{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$.



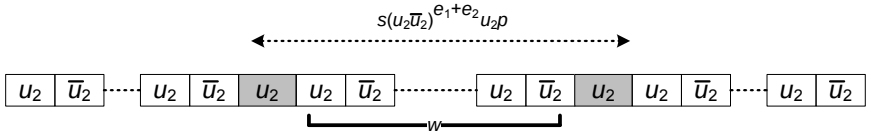
Here $\boldsymbol{w}_{[1]}$ is a factor in $\boldsymbol{s}\boldsymbol{u}_1{}^{e_1+e_2}\boldsymbol{u}_2\boldsymbol{p}$, where $\boldsymbol{s}$ is the maximal common suffix and $\boldsymbol{p}$ the maximal common prefix of $\boldsymbol{u}_2\overline{\boldsymbol{u}}_2$ and $\overline{\boldsymbol{u}}_2\boldsymbol{u}_2$. Thus $\boldsymbol{w}^2$ and $\boldsymbol{s}\boldsymbol{u}_1{}^{e_1+e_2}\boldsymbol{u}_2\boldsymbol{p}$ have a common factor of size $|\boldsymbol{u}_1+\boldsymbol{w}|$ and by the Common Factor Lemma 1.32, the primitive root of $\boldsymbol{w}$ is a conjugate of $\boldsymbol{u}_1$, i.e. case (c) holds.

(5) $\boldsymbol{w}_{[1]}$ starts in the suffix of $\overline{\boldsymbol{u}}_2\boldsymbol{u}_2$ of $\boldsymbol{u}$ whose length $> \text{lcs}(\boldsymbol{u}_2\overline{\boldsymbol{u}}_2, \overline{\boldsymbol{u}}_2\boldsymbol{u}_2)$.



Then $\boldsymbol{w}_{[1]}$ contains $\mathbb{L}\mathbb{I}\$$ and thus $\boldsymbol{w}_{[2]}$ must contain an occurrence of $\mathbb{L}\mathbb{I}\$$ and so $|\boldsymbol{w}| \geq |\boldsymbol{v}|$, and so this case is not possible.

(6) $\boldsymbol{w}_{[1]}$ starts past the first $\boldsymbol{u}$.



The same argument as in (4) shows that the primitive root of $\boldsymbol{w}$ is a conjugate of $\boldsymbol{u}_1$ and so case (c) holds.

□

In the following we present a few examples of the Theorem 3.2 illustrating various situations encountered in the proof.

Example 1

This is an example of a square $\boldsymbol{w}^2$ such that $|\boldsymbol{w}| < |\boldsymbol{u}_1|$

$\boldsymbol{u}_2 = abaa,\ \overline{\boldsymbol{u}}_2 = baa,\ \boldsymbol{u}_1 = abaabaa,\ e_1 = 1,\ e_2 = 1$

$\boldsymbol{u} = \boldsymbol{u}_1\boldsymbol{u}_2 = abaabaaabaa$

$\boldsymbol{v} = \boldsymbol{u}_1\boldsymbol{u}_2\boldsymbol{u}_1 = abaabaaabaaabaabaa$

$\boldsymbol{w^2} = abaaba$

$$\boldsymbol{v}^2 = \underbrace{abaaba}_{\boldsymbol{w}^2}aabaaabaabaaabaabaaabaaabaabaa$$

Example 2

This is an example of a square $\boldsymbol{w}^2$ such that $|\boldsymbol{w}| < |\boldsymbol{u}|$

$\boldsymbol{u}_2 = abb,\ \overline{\boldsymbol{u}}_2 = ab,\ \boldsymbol{u}_1 = abbab,\ e_1 = 2,\ e_2 = 1$

$\boldsymbol{u} = \boldsymbol{u}_1^2\boldsymbol{u}_2 = abbababbababb$

$\boldsymbol{v} = \boldsymbol{u}_1^2\boldsymbol{u}_2\boldsymbol{u}_1 = abbababbabbabbab$

$\boldsymbol{w^2} = abbababbab$

$\boldsymbol{v}^2 = \underbrace{abbababbab}_{\boldsymbol{w}^2} abbabbababbababbababbabbab$

Example 3

This is an example of a square $\boldsymbol{w}^2$ such that $|\boldsymbol{w}| < |\boldsymbol{u}|$

$\boldsymbol{u}_2 = abb,\ \overline{\boldsymbol{u}}_2 = ab,\ \boldsymbol{u}_1 = abbab,\ e_1 = 2,\ e_2 = 2$

$\boldsymbol{u} = \boldsymbol{u}_1^2\boldsymbol{u}_2 = abbababbabab$

$\boldsymbol{v} = \boldsymbol{u}_1^2\boldsymbol{u}_2\boldsymbol{u}_1^2 = abbababbababbabbababbab$

$\boldsymbol{w^2} = babbababbababbababba$

$\boldsymbol{v}^2 = abbababbabab\underbrace{babbababba\,babbababba}_{\boldsymbol{w}^2}babbabbababbab$

Example 4

This is an example of a square $\boldsymbol{w}^2$ such that $|\boldsymbol{w}| < |\boldsymbol{u}|$

$\boldsymbol{u}_2 = abaa,\ \overline{\boldsymbol{u}}_2 = a,\ \boldsymbol{u}_1 = abaaa,\ e_1 = 2,\ e_2 = 1$

$\boldsymbol{u} = \boldsymbol{u}_1^2\boldsymbol{u}_2 = abaaaabaaaabaa$

$\boldsymbol{v} = \boldsymbol{u}_1^2\boldsymbol{u}_2\boldsymbol{u}_1 = abaaaabaaaabaaabaaa$

$\boldsymbol{w^2} = aabaaaabaaaabaaaabaa$

$\boldsymbol{v}^2 = abaaaabaaaaba\underbrace{aabaaaabaa\,aabaaaabaa}_{\boldsymbol{w}^2}abaaa$

Example 5

This is an example of a square $\boldsymbol{w}^2$ such that $|\boldsymbol{w}| < |\boldsymbol{u}|$

$\boldsymbol{u}_2 = ab,\ \overline{\boldsymbol{u}}_2 = b,\ \boldsymbol{u}_1 = abb,\ e_1 = 2,\ e_2 = 2$

$\boldsymbol{u} = \boldsymbol{u}_1^2\boldsymbol{u}_2 = abbabbab$

$\boldsymbol{v} = \boldsymbol{u}_1^2\boldsymbol{u}_2\boldsymbol{u}_1^2 = abbabbababbabb$

$\boldsymbol{w^2} = bbabbabbabba$

$\boldsymbol{v}^2 = abbabbaba\underbrace{bbabbabbabba}_{\boldsymbol{w}^2}babbabb$

Example 6

This is an example of a square $\boldsymbol{w}^2$ such that $|\boldsymbol{w}| < |\boldsymbol{u}|$

$\boldsymbol{u}_2 = aabc,\ \overline{\boldsymbol{u}}_2 = bc,\ \boldsymbol{u}_1 = aabcbc,\ e_1 = 3,\ e_2 = 1$

$\boldsymbol{u} = \boldsymbol{u}_1^3\boldsymbol{u}_2 = aabcbcaabcbcaabcbcaabc$

$\boldsymbol{v} = \boldsymbol{u}_1^3\boldsymbol{u}_2\boldsymbol{u}_1 = aabcbcaabcbcaabcbcaabcaabcbc$

$\boldsymbol{w^2} = bcbcaabcbcaabcbcaabcbcaa$

$\boldsymbol{v}^2 = aabcbcaabcbcaabcbcaabcaa\underbrace{bcbcaabcbcaa\,bcbcaabcbcaa}_{\boldsymbol{w}^2}bcaabcbc$

Chapter 4

Application of canonical factorization to Three Squares Lemma

In this chapter based on the paper [1], we discuss an application of the canonical factorization of double squares to the Three Squares Lemma by Crochemore and Rytter introduced in 1995. It is a structural lemma on three squares starting at the same position. This influential lemma has been used by many researchers in the field of periodicities in strings. In particular, Fraenkel and Simpson used it in 1998 to obtain a universal upper bound for the maximum number of distinct squares occurring in a string. We present a generalization of Three Squares Lemma by exploiting the canonical factorization of balanced double squares discussed in Chapter 2. Let us first state

the original lemma.

Lemma 4.1 ([7], Three Squares Lemma)**.** *Let* $\boldsymbol{u}^2 \neq \boldsymbol{v}^2$ *be proper prefixes of* $\boldsymbol{w}^2$ *and let* $\boldsymbol{u}$*,* $\boldsymbol{v}$*, and* $\boldsymbol{w}$ *be primitive, then* $|\boldsymbol{u}|+|\boldsymbol{v}| < |\boldsymbol{w}|$*.*

Lemma 4.1 has been used by many researchers including Kolpakov and Kucherov [13], Stoye and Gusfield [18], Fan, Puglisi, Smyth, and Turpin [9], Simpson [16]. Lemma 4.1 was essential for the 1998 result by Fraenkel and Simpson [10] giving a universal upper bound of $2n$ for the number of distinct squares in a string of length n. Note that for the problem of distinct squares, every type of square is only counted once, i.e. the types, rather than the occurrences, are counted. For illustration, *aabaab* contains the following three underlined squares $\underline{aa}baab$, $\underline{aabaab}$ and $aab\underline{aa}b$ while the number of distinct squares is 2: *aa* and *aabaab*. Ilie [12] provided in 2005 an alternate proof of the main theorem of [10] not directly using Lemma 4.1. Noticing that the proof of Lemma 4.1 by Crochemore and Rytter only requires the primitiveness of the shortest square, Fraenkel and Simpson [10] proposed the following strengthening referred to as three-prefix-square Lemma in [6] where additional context and references can be found.

Lemma 4.2 ([10],Three Squares Lemma – Fraenkel and Simpson variant)**.** *Let* $\boldsymbol{u}^2 \neq \boldsymbol{v}^2$ *be proper prefixes of* $\boldsymbol{w}^2$ *and let the shorter of the two strings* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *be primitive, then* $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$*.*

Example 1

$\boldsymbol{w}^2 = aabaaababaabaaabab$

$\boldsymbol{w} = aabaaabab,\ \boldsymbol{v} = aaba,\ \boldsymbol{u} = a$

$|\boldsymbol{w}| = 9,\ |\boldsymbol{v}| = 4,\ |\boldsymbol{u}| = 1$, thus, $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$

Example 2

$\boldsymbol{w}^2 = abbabbaabbabbaaabbabbaabbabbaa$

$\boldsymbol{w} = abbabbaabbabbaa,\ \boldsymbol{v} = abbabba,\ \boldsymbol{u} = abb$

$|\boldsymbol{w}| = 15,\ |\boldsymbol{v}| = 7,\ |\boldsymbol{u}| = 3$, thus, $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$

Example 3

$\boldsymbol{w}^2 = abaabaabaaba$

$\boldsymbol{w} = abaabaabaaba,\ \boldsymbol{v} = abaaba,\ \boldsymbol{u} = aba$

$|\boldsymbol{w}| = 12,\ |\boldsymbol{v}| = 6,\ |\boldsymbol{u}| = 3$, thus, $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$

Example 4

$\boldsymbol{w}^2 = abababababab$

$\boldsymbol{w} = ababab,\ \boldsymbol{v} = abab,\ \boldsymbol{u} = ab$

$|\boldsymbol{w}| = 6,\ |\boldsymbol{v}| = 4,\ |\boldsymbol{u}| = 2$, thus, $|\boldsymbol{u}|+|\boldsymbol{v}| = |\boldsymbol{w}|$

Fraenkel and Simpson illustrated the necessity of the primitiveness for the shortest square with the following example: $\boldsymbol{u} = a^2$, $\boldsymbol{v} = a^4$, and $\boldsymbol{w} = a^5$. We present a further strengthening based on the recently investigated structural properties of two squares starting at the same position, see [2, 8].

Fraenkel and Simpson counterexample

$\boldsymbol{w}^2 = aaaaaaaaaa$

$\boldsymbol{w} = aaaaa,\ \boldsymbol{v} = aaaa,\ \boldsymbol{u} = aa$

$|\boldsymbol{w}| = 5,\ |\boldsymbol{v}| = 4,\ |\boldsymbol{u}| = 2$, thus, $|\boldsymbol{u}|+|\boldsymbol{v}| > |\boldsymbol{w}|$

In the paper [1], our strengthened version of the Three Sqaure Lemma was formulated in the following way

Lemma 4.3. *Let $\boldsymbol{u}^2 \neq \boldsymbol{v}^2$ be proper prefixes of $\boldsymbol{w}^2$, then $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$ unless $\boldsymbol{u}, \boldsymbol{v}$, and $\boldsymbol{w}$ have the same primitive root.*

After the publication of [1], a fellow AdvOL Ph.D. student Adrien Thierry came with a counterexample for the lemma:

$\boldsymbol{u} = (ab)^2$, $\boldsymbol{v} = (ab)^3a$, and $\boldsymbol{w} = (ab)^3a(ab)$, clearly $|\boldsymbol{u}| + |\boldsymbol{v}| = 4+7 = 11$ while $|\boldsymbol{w}| = 9$, yet $\boldsymbol{ww} = \overbrace{\underbrace{abab}_{\boldsymbol{u}}abaababababa}^{\boldsymbol{v}}\underbrace{baab}_{\boldsymbol{u}}$ and $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ do not have the same primitive root.

Thierry's counterexample pointed out a little oversight in the proof of Lemma 4.3. The proof was corrected and it weakened the conclusion of the lemma slightly. Here we present the corrected Lemma 4.4 and its proof. The corrigendum was published soon after the error was detected, see [1].

Lemma 4.4. *Let $\boldsymbol{u}^2 \neq \boldsymbol{v}^2$ be proper prefixes of $\boldsymbol{w}^2$, then $|\boldsymbol{u}| + |\boldsymbol{v}| \leq |\boldsymbol{w}|$ unless $\boldsymbol{u} = {\boldsymbol{v}_1}^t$, $\boldsymbol{v} = {\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2$, and $\boldsymbol{w} = {\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2{\boldsymbol{v}_1}^{p_2}$ where $\boldsymbol{v}_1$ is primitive, $\boldsymbol{v}_2$ a proper possibly empty prefix of $\boldsymbol{v}_1$, $t > p_2$, and $p_1 \geq p_2 \geq 1$.*

Lemma 4.4 shows that the squares $\boldsymbol{u}^2$, $\boldsymbol{v}^2$, and $\boldsymbol{w}^2$ violating $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$ consist of two types; one corresponding to the counterexample given by Fraenkel and Simpson (when $\boldsymbol{v}_2 = \varepsilon$) and one corresponding to the coun-

terexample of Thierry (when $\boldsymbol{v}_2 \neq \varepsilon$). Bellow is essentially Fraenkel-Simpsons counterexample:

$\boldsymbol{w}^2 = abababababababab$

$\boldsymbol{w} = abababab,\ \boldsymbol{v} = ababab,\ \boldsymbol{v} = abab,$

$|\boldsymbol{w}| = 8,\ |\boldsymbol{v}| = 6,\ |\boldsymbol{u}| = 4$, thus, $|\boldsymbol{u}|+|\boldsymbol{v}| > |\boldsymbol{w}|$

Corollary 4.5 illustrates that Lemma 4.4 is a true generalization of Lemma 4.2.

Corollary 4.5. *Let $\boldsymbol{u}^2$ be a proper prefix of $\boldsymbol{v}^2$ that is a proper prefixes of $\boldsymbol{w}^2$ and let $\boldsymbol{u}$ be primitive, then $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$. Moreover, if $|\boldsymbol{u}| < |\boldsymbol{v}| < 2|\boldsymbol{u}|$ and either $\boldsymbol{v}$ or $\boldsymbol{w}$ is primitive, then $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$.*

Proof. Let us assume by contradiction that $|\boldsymbol{u}|+|\boldsymbol{v}| > |\boldsymbol{w}|$. Then by Lemma 4.4, $\boldsymbol{u} = {\boldsymbol{v}_1}^t$, $\boldsymbol{v} = {\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2$ and $\boldsymbol{w} = {\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2{\boldsymbol{v}_1}^{p_2}$ for a primitive $\boldsymbol{v}_1$, a proper possibly empty prefix $\boldsymbol{v}_2$ of $\boldsymbol{v}_1$, and $t > p_2$, $p_1 \geq p_2 \geq 1$. If $\boldsymbol{u}$ is primitive, $t = 1$ and so $t > p_2 \geq 1$ is a contradiction. If $|\boldsymbol{v}| < 2|\boldsymbol{u}|$, then ${\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2$ is a prefix of ${\boldsymbol{v}_1}^{2t}$, which can only be true when $\boldsymbol{v}_2$ is empty due to Synchronization Principle Lemma 1.31. If $\boldsymbol{v}$ is primitive, then $p_1 = 1$ and so $p_2 = 1$ and so $\boldsymbol{u} = {\boldsymbol{v}_1}^t$, $t > 1$ and $\boldsymbol{v} = \boldsymbol{v}_1$ and $\boldsymbol{w} = {\boldsymbol{v}_1}^2$, and so $|\boldsymbol{u}| \geq |\boldsymbol{w}|$, a contradiction. If $\boldsymbol{w}$ is primitive, then $\boldsymbol{w} = \boldsymbol{v}_1$, and so $|\boldsymbol{w}| = |\boldsymbol{v}|$, a contradiction. □

Example 1 : when $\boldsymbol{u}$ is primitive

$\boldsymbol{w}^2 = abababababababab$

$\boldsymbol{w} = abababab,\ \boldsymbol{v} = abab,\ \boldsymbol{u} = ab$

$|\boldsymbol{w}| = 8,\ |\boldsymbol{v}| = 4,\ |\boldsymbol{u}| = 2$, thus, $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$

Example 2 : when $|\boldsymbol{u}| < |\boldsymbol{v}| < 2|\boldsymbol{u}|$ and $\boldsymbol{v}$ is primitive

$\boldsymbol{w}^2 = aabaabaabaab$

$\boldsymbol{w} = aabaab,\ \boldsymbol{v} = aab,\ \boldsymbol{u} = aa$

$|\boldsymbol{w}| = 6,\ |\boldsymbol{v}| = 3,\ |\boldsymbol{u}| = 2$, thus, $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$

Example 3 : when $|\boldsymbol{u}| < |\boldsymbol{v}| < 2|\boldsymbol{u}|$ and $\boldsymbol{w}$ is primitive

$\boldsymbol{w}^2 = aaaaaabaaaaaab$

$\boldsymbol{w} = aaaaaab,\ \boldsymbol{v} = aaa,\ \boldsymbol{u} = aa$

$|\boldsymbol{w}| = 7,\ |\boldsymbol{v}| = 3,\ |\boldsymbol{u}| = 2$, thus, $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$

Example 4 : when $\boldsymbol{v}$ is primitive, this example shows that conditions in Corollary 4.5 are not necessary to conclude that $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$, since $\boldsymbol{u}$ is not primitive and $|\boldsymbol{v}| > 2|\boldsymbol{u}|$.

$\boldsymbol{w}^2 = aaaabaaaabaaaabaaaab$

$\boldsymbol{w} = aaaabaaaab,\ \boldsymbol{v} = aaaab,\ \boldsymbol{u} = aa$

$|\boldsymbol{w}| = 10,\ |\boldsymbol{v}| = 5,\ |\boldsymbol{u}| = 2$, thus, $|\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|$

Proof of Lemma 4.4

Proof. Let $\boldsymbol{u} \neq \boldsymbol{v}$, and $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$ be both proper prefixes of $\boldsymbol{w}^2$. Lemma 4.4 states that

$\big\{\boldsymbol{u} = {\boldsymbol{v}_1}^t, \boldsymbol{v} = {\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2, \boldsymbol{w} = {\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2{\boldsymbol{v}_1}^{p_2}\ , t > p_2\ , p_1 \geq p_2 \geq 1\big\}$ or $\big\{\ |\boldsymbol{u}|+|\boldsymbol{v}| \leq |\boldsymbol{w}|\ \big\}$. (S)

Without loss of generality, we can assume that $|\boldsymbol{u}| < |\boldsymbol{v}|$.

If $2|\boldsymbol{v}| \leq |\boldsymbol{w}|$, then $|\boldsymbol{v}|+|\boldsymbol{u}| < |\boldsymbol{w}|$ as $|\boldsymbol{u}| < |\boldsymbol{v}|$, and thus (S) holds. Therefore,

we can assume that $|\boldsymbol{w}| < 2|\boldsymbol{v}|$; that is, $(\boldsymbol{v}, \boldsymbol{w})$ is a balanced double square and thus admits a canonical factorization $(\boldsymbol{v}, \boldsymbol{w} \ : \ \boldsymbol{v}_1, \boldsymbol{v}_2, p_1, p_2)$ by Lemma 2.1. We consider the following cases.

1. Case when $\boldsymbol{u}$ and $\boldsymbol{v}$ are not proportional, i.e. $2|\boldsymbol{u}| \leq |\boldsymbol{v}|$.

   If $|\boldsymbol{u}| < |\boldsymbol{v}_1|$, then

$$\begin{aligned} |\boldsymbol{u}| + |\boldsymbol{v}| &= |\boldsymbol{u}| + |\boldsymbol{v}_1^{p_1}\boldsymbol{v}_2| \\ &< |\boldsymbol{v}_1| + |\boldsymbol{v}_1^{p_1}\boldsymbol{v}_2| \\ &= |\boldsymbol{v}_1^{p_1+1}\boldsymbol{v}_2| \\ &\leq |\boldsymbol{v}_1^{p_1+p_2}\boldsymbol{v}_2| \\ &= |\boldsymbol{w}| \end{aligned}$$

   If $|\boldsymbol{u}| \geq |\boldsymbol{v}_1|$, since $\boldsymbol{u}^2$ is a prefix of $\boldsymbol{v} = {\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2$, then $\boldsymbol{u}^2$ and ${\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2$ have a common factor of length $|\boldsymbol{u}|+|\boldsymbol{v}_1|$, and by Lemma 1.32 , $\boldsymbol{u}$ and $\boldsymbol{v}_1$ have the same primitive root, and so $\boldsymbol{v}_1$ is the primitive root of $\boldsymbol{v}_1$. Thus $\boldsymbol{u} = {\boldsymbol{v}_1}^t$ for some $t \geq 1$, $\boldsymbol{v} = {\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2$, and $\boldsymbol{w} = {\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2{\boldsymbol{v}_1}^{p_2}$.

   Now we check the relation between $|{\boldsymbol{v}_1}^t + {\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2|$ and $|{\boldsymbol{v}_1}^{p_1}\boldsymbol{v}_2{\boldsymbol{v}_1}^{p_2}|$, which actually is the relation between $|{\boldsymbol{v}_1}^t|$ and $|{\boldsymbol{v}_1}^{p_2}|$.

   If $t \leq p_2$, then $|\boldsymbol{u}| + |\boldsymbol{v}| \leq |\boldsymbol{w}|$, but if $t > p_2$, then $|\boldsymbol{u}| + |\boldsymbol{v}| > |\boldsymbol{w}|$. In both cases , (S) holds.

2. Case when $\boldsymbol{u}$ and $\boldsymbol{v}$ are proportional, i.e. $|\boldsymbol{v}| < 2|\boldsymbol{u}|$.

Then $DS(\boldsymbol{u}, \boldsymbol{v})$ is a balanced double square and thus admits by the Two Square Factorization Lemma 2.1 a canonical factorization $DS(\boldsymbol{u}, \boldsymbol{v} : \boldsymbol{u}_1, \boldsymbol{u}_2, e_1, e_2)$. As a result, we have:

$$\boldsymbol{u} = \boldsymbol{u}_1^{e_1}\boldsymbol{u}_2$$

$$\boldsymbol{v} = \boldsymbol{u}_1^{e_1}\boldsymbol{u}_2\boldsymbol{u}_1^{e_2}$$

$$\boldsymbol{v}^2 = \boldsymbol{u}_1^{e_1}\boldsymbol{u}_2\boldsymbol{u}_1^{e_1+e_2}\boldsymbol{u}_2\boldsymbol{u}_1^{e_2}$$

and

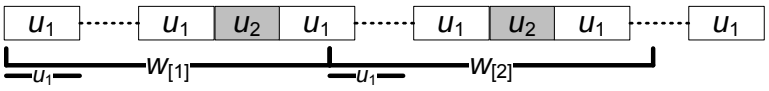
$$\boldsymbol{v} = \boldsymbol{v}_1^{p_1}\boldsymbol{v}_2$$

$$\boldsymbol{w} = \boldsymbol{v}_1^{p_1}\boldsymbol{v}_2\boldsymbol{v}_1^{p_2}$$

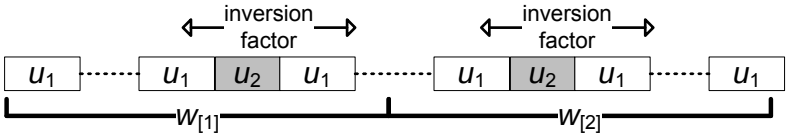
$$\boldsymbol{w}^2 = \boldsymbol{v}_1^{p_1}\boldsymbol{v}_2\boldsymbol{v}_1^{p_1+p_2}\boldsymbol{v}_2\boldsymbol{v}_1^{p_2}$$

(i) Case when $|\boldsymbol{u}_2| = 0$. Then $e_1 > e_2$, $\boldsymbol{u} = {\boldsymbol{u}_1}^{e_1}$, and $\boldsymbol{v} = {\boldsymbol{u}_1}^{e_1+e_2}$. Let us assume that $|\boldsymbol{w}| < |\boldsymbol{u}|+|\boldsymbol{v}| = (2e_1+e_2)|\boldsymbol{u}_1|$. Then $\boldsymbol{w}^2$ and ${\boldsymbol{u}_1}^{2e_1+2e_2}$ have a common factor of length $|\boldsymbol{w}|+|\boldsymbol{u}_1|$, and by the Common Factor Lemma 1.32 the primitive root of $\boldsymbol{w}$ is a conjugate of $\boldsymbol{u}_1$, i.e. equals $\boldsymbol{u}_1$. Thus, $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ all have the same primitive root, and thus (S) holds.

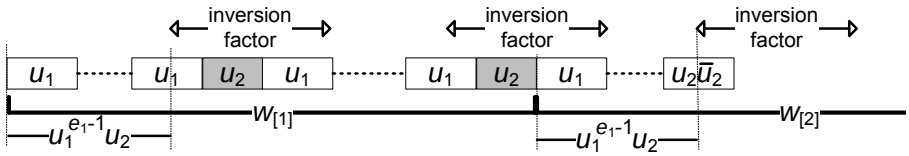
(ii) Case when $|\boldsymbol{u}_2| > 0$. Let $\boldsymbol{w}_{[1]}$ refer to the first occurrence of $\boldsymbol{w}$ and $\boldsymbol{w}_{[2]}$ to the second. First, we have to show that $\boldsymbol{w}_{[1]}$ does not end in the first $\boldsymbol{u}_1$ of ${\boldsymbol{u}_1}^{e_1+e_2}$. If it did, then it would contradict the Synchronization Principle Lemma 1.31 as $\boldsymbol{w}_{[1]}$ and hence $\boldsymbol{w}_{[2]}$ has the primitive $\boldsymbol{u}_1$ as a prefix as indicated by the following diagram:



Thus, $\boldsymbol{w}_{[1]}$ must end somewhere past the first $\boldsymbol{u}_1$ of $\boldsymbol{u}_1{}^{e_1+e_2}$:

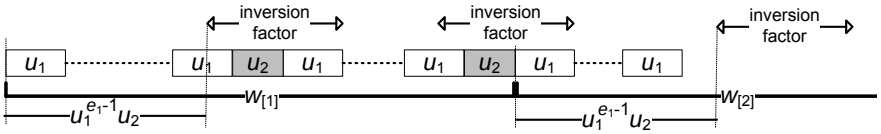


As a consequence, $\boldsymbol{w}s_{[1]}$ contains the first inversion factor of $\boldsymbol{u}_1{}^{e_1}\boldsymbol{u}_2\boldsymbol{u}_1{}^{e_1+e_2}\boldsymbol{u}_2\boldsymbol{u}_1{}^{e_2}$ exactly at a distance of $|\boldsymbol{u}_1{}^{e_1-1}\boldsymbol{u}_2|$ from the beginning. It follows that $\boldsymbol{w}_{[2]}$ must contain an occurrence of the inversion factor at exactly the same distance from the beginning of $\boldsymbol{w}_{[2]}$. If it were the second inversion factor, then the length of w would be exactly $|\boldsymbol{v}|$ as it is the distance between the two occurrences of the inversion factor in $\boldsymbol{w}^2$, a contradiction. Thus, it must be an occurrence of the inversion factor past the second one. The first possible start of another occurrence of the inversion factor is the suffix $\overline{\boldsymbol{u}}_2$ of $\boldsymbol{u}_1{}^{e_1}\boldsymbol{u}_2\boldsymbol{u}_1{}^{e_1+e_2}\boldsymbol{u}_2\boldsymbol{u}_1{}^{e_2}$. If $e_1 = e_2$, then it is the case that $\boldsymbol{w}_{[1]} = \boldsymbol{w}_{[2]} = \boldsymbol{u}_1{}^{e_1-1}(\overline{\boldsymbol{u}}_2\boldsymbol{u}_2\boldsymbol{u}_2\overline{\boldsymbol{u}}_2)\boldsymbol{u}_1{}^{e_1+e_2-1}\boldsymbol{u}_2$ (see below) and then

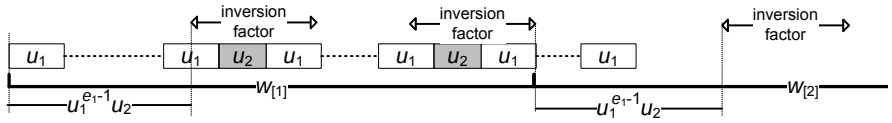


$|\boldsymbol{w}| = |\boldsymbol{u}|+|\boldsymbol{v}|$ as $|\boldsymbol{u}| = |\boldsymbol{u}_1{}^{e_1}\boldsymbol{u}_2|$ and $|\boldsymbol{v}| = |\boldsymbol{u}_1{}^{e_1+e_2}\boldsymbol{u}_2|$, thus (S) holds. If $e_1 > e_2$, then the occurrence of the inversion factor in $\boldsymbol{w}_{[2]}$ must be at a distance $|\boldsymbol{u}_1{}^{e_1-1}\boldsymbol{u}_2|$ from the beginning of $\boldsymbol{w}_{[2]}$. By the Synchronization Principle Lemma 1.31, the prefix $\boldsymbol{u}_1{}^{e_1-1}\boldsymbol{u}_2$ of

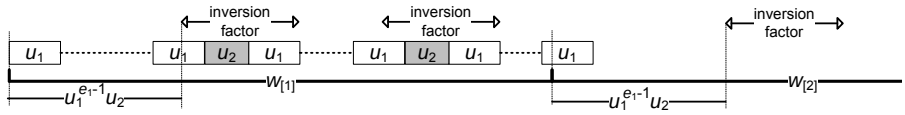
$\boldsymbol{w}_{[2]}$ must align with ${\boldsymbol{u}_1}^{e_2}$ or start in the last $\boldsymbol{u}_1$ of ${\boldsymbol{u}_1}^{e_2}$, and so $\boldsymbol{w}_{[1]}$ must have ${\boldsymbol{u}_1}^{e_1}\boldsymbol{u}_2{\boldsymbol{u}_1}^{e_1+e_2}\boldsymbol{u}_2$ as a prefix, again yielding $|\boldsymbol{w}| \geq |\boldsymbol{u}|+|\boldsymbol{v}|$ (see below), thus (S) holds.



or



or



This completes the proof. □

Chapter 5

Conclusion and Future work

We presented a unique canonical factorization of a balanced double square consisting of two squares starting in the same position and of comparable lengths. The interesting part is that this unique factorization is guaranteed without any additional conditions. We described and analyzed three kinds of factors occurring rarely in the longer square: inversion factors, right inversion subfactors and left inversion subfactors. We discussed how such rare factors can be used to constraint the sizes of any possible third square starting in proximity of the start of the double square. The right and left inversion subfactors provide stronger constraint in comparison to the inversion factors due to their length that is half of the length of the inversion factors. The right and left inversion subfactors were discovered and analyzed exactly for this stronger constraint for the purpose of proving the New Periodicity Lemma.

A stronger version of the New Periodicity Lemma was formulated with

a significantly weaker assumption using the canonical factorization and the right and left inversion subfactors. The range of possible starts for "forbidden" squares was enlarged without any additional constraints. We also gave a much simpler proof of the generalized New Periodicity Lemma in comparison to the proof of the original New Peridocity Lemma.

The canonical factorization and the inversion factors were applied to prove a stronger version of the Three Squares Lemma. In order to guarantee the conclusion of the original Three Squares Lemma, our generalization only needs one of the three squares to be primitive-rooted when the constraints of length are satisfied, while the original lemma requires the shortest square's root to be primitive.

The future work and research directions will focus on strengthening the ability of the rare factors to constrain the possible positions where the third square can occur in a balanced double square. This will require discovering some shorter factors that occur rarely in the double square. With shorter rare factors, both the New Periodicity Lemma and the Three Squares Lemma could be generalized and extended further. Since the rare factors were essential to proving the best-to-date upper bound of $11n/6$ for the number of distinct squares in a string of length n, it is quite promising that shorter rare factors will further improve this upper bound.

Bibliography

[1] H. Bai, A. Deza, and F. Franek. On a lemma of Crochemore and Rytter. *Journal of Discrete Algorithms*, 34:18–22, 2015. (Corrigendum, *Journal of Discrete Algorithms*, 38-41:50-51, 2016).

[2] H. Bai, F. Franek, and W. Smyth. Two squares canonical factorization. In *Proceedings of Prague Stringology Conference PSC 2014*, pages 52–58, 2014.

[3] H. Bai, F. Franek, and W. Smyth. The new periodicity lemma revisited. *Discrete Applied Mathematics*, 212(30):30–36, 2016.

[4] W. Bland and W. Smyth. Three overlapping squares. *Theor. Comput. Sci.*, 596(C):23–40, 2015.

[5] K. Chen, R. Fox, and R. Lyndon. Free differential calculus, IV. The quotient groups of the lower central series. *The Annals of Mathematics*, 68(1):81–95, 1958.

[6] M. Crochemore, C. Hancart, and T. Lecroq. *Algorithms on Strings*. Cambridge University Press, New York City, New York, 2007.

[7] M. Crochemore and W. Rytter. Squares, cubes, and time-space efficient string searching. *Algorithmica*, 13(5):405–425, 1995.

[8] A. Deza, F. Franek, and A. Thierry. How many double squares can a string contain? *Discrete Applied Mathematics*, 180(10):52–69, 2015.

[9] K. Fan, S. Puglisi, W. Smyth, and A. Turpin. A new periodicity lemma. *SIAM Journal on Discrete Mathematics*, 20(3):656–668, 2006.

[10] A. Fraenkel and J. Simpson. How many squares can a string contain? *Journal of Combinatorial Theory, Series A*, 82(1):112–120, 1998.

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

[12] L. Ilie. A simple proof that a word of length n has at most $2n$ distinct squares. *Journal of Combinatorial Theory, Series A*, 112(1):163–164, 2005.

[13] R. Kolpakov and G. Kucherov. Finding maximal repetitions in a word in linear time. In *Proceedings of 40th Annual Symposium on Foundations of Computer Science*, pages 596–604, 1999.

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

[15] N. Lam. On the number of squares in a string. Technical report, AdvOL-Report No. 2013/02, McMaster University, 2013.

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

[17] W. Smyth. *Computing Patterns in Strings*. Addison-Wesley, 2002.

[18] J. Stoye and D. Gusfield. Simple and flexible detection of contiguous repeats using a suffix tree. *Theoretical Computer Science*, 270(6):843–856, 2002.

[19] A. Thierry. Combinatorics of the interrupted period. In *Proceedings of Prague Stringology Conference PSC 2015*, pages 17–21, 2015.