The New Periodicity Lemma Revisited

Haoyue Bai[1], Frantisek Franek[1], William F. Smyth[1,2,3]

[1] *Department of Computing and Software*
McMaster University, Hamilton, Ontario, Canada
[2] *School of Engineering & Information Technology,*
Murdoch University, Perth, Western Australia
[3] *School of Computer Science & Software Engineering*
University of Western Australia, Perth, Western Australia

Abstract

In 2006, the *New Periodicity Lemma* (NPL) was published, showing that the occurrence of two squares starting at a position i in a string necessarily precludes the occurrence of other squares of specified period in a specified neighbourhood of i. The proof of this lemma was complex, breaking down into 14 subcases, and requiring that the shorter of the two squares be *regular*. In this paper we significantly relax the conditions required by the NPL and removing the need for regularity altogether, and we establish a more precise result using a simpler proof based on lemmas that expose new combinatorial structures in a string, in particular a *canonical factorization* for any two squares that start at the same position.

Keywords: string, square, canonical factorization, double square, New Periodicity Lemma

1. Introduction

In 1995 Crochemore and Rytter [3] considered three distinct squares, all prefixes of a given string $\boldsymbol{x}$, and proved the *Three Squares Lemma*, stating that, subject to certain restrictions, the largest of the three was at least the length of the sum of the other two. In 2006 Fan *et al.* [5] considered two squares that were prefixes of $\boldsymbol{x}$ with the third square offset some distance

Email addresses: baih3@mcmaster.ca (Haoyue Bai[1]), franek@mcmaster.ca (Frantisek Franek[1]), smyth5@mcmaster.ca (William F. Smyth[1,2,3])

to the right; they proved a *New Periodicity Lemma* (NPL), describing conditions under which the third square could not exist. Since that time there has been considerable work done [2, 7, 8, 10] in an effort to specify more precisely the combinatorial structure of the string in the neighbournood of such squares.

In this paper we first discuss a *canonical factorization*, a unique breakdown into primitive strings of what we call a *double square*, i.e. a pair of two squares starting at the same position and being of "comparable" lengths. A weaker form of the factorization was instrumental in the improved upper bound for the number of distinct squares [4]; weaker in the sense that it only applied to FS-double squares, i.e. two squares that start at the same position and both are rightmost occurrences. Note that our notion of double squares is weaker and hence every FS-double square is a double square, but not the other way around. The canonical factorization indicates that double squares indeed have an intricate highly periodic intrinsic structure. This structure has two factors that are unique in their occurrences within the double square. They were introduced in [4] and referred to as *inversion factors* due to their structure. In [12], Thierry discusses the *core of the period interrupt*, a very similar concept to the one we introduce here as RIS (Right Inversion Subfactor) and LIS (Left Inversion Subfactor). RIS has only two occurrences in the double square, and so does LIS. The usage of RIS, respective LIS, is straightforward as it significantly limits the size and the placement of a possible third square: let $\boldsymbol{u}^2$ be a prefix of $\boldsymbol{v}^2$ and consider a third square $\boldsymbol{w}^2$; if it contains RIS in the first $\boldsymbol{w}$, it must contain RIS in the second $\boldsymbol{w}$ and vice-versa, and hence $\boldsymbol{w}$ has the same size as $\boldsymbol{v}$. So the only other possibilities are that either $\boldsymbol{w}^2$ is "too small" that it does not contain RIS, or "too big" that it contains both RIS in the first $\boldsymbol{w}$. The restrictions imposed by RIS or LIS allow us to prove a new version of the NPL that is much more general in its application while at the same time being more precise in its result.

The paper is structured as follows: in Section 2 we discuss the basic facts and notations. In Section 3 we present and prove the *Two Squares Factorization Lemma* giving what we refer to as the canonical factorization of a double square. In Section 4 we discuss the inversion factors and their refinements RIS and LIS. The new formulation of the NPL is then presented and proved in Section 5. Finally, Section 6 presents a brief conclusion of the research described.

2. Preliminaries

In this section we introduce the basic notation and develop the combinatorial tools that will be used to determine a canonical factorization for a double square. Chief among these are the Synchronization Principle (Lemma 2) and the Common Factor Lemma (Lemma 3), that lead to the Two Squares Factorization Lemma (Lemma 5).

A ***string*** $\boldsymbol{x}$ is a finite sequence of symbols, called ***letters***, drawn from a (finite or infinite) set Σ, called the ***alphabet***. The length of the sequence is called the ***length*** of $\boldsymbol{x}$, denoted $|\boldsymbol{x}|$. Sometimes for convenience we represent a string $\boldsymbol{x}$ of length n as an array $\boldsymbol{x}[1..n]$. The string of length zero is called the ***empty string***, denoted $\boldsymbol{\varepsilon}$. If a string $\boldsymbol{x} = \boldsymbol{uvw}$, where $\boldsymbol{u}$,$\boldsymbol{v}$,$\boldsymbol{w}$ are strings, then $\boldsymbol{u}$ (respectively, $\boldsymbol{v}$,$\boldsymbol{w}$) is said to be a ***prefix*** (respectively, ***substring***, ***suffix***) of $\boldsymbol{x}$; a ***proper prefix*** (respectively, ***proper substring***, ***proper suffix***) if $|\boldsymbol{u}| < |\boldsymbol{x}|$ (respectively, $|\boldsymbol{v}| < |\boldsymbol{x}|$, $|\boldsymbol{w}| < |\boldsymbol{x}|$). An empty prefix or suffix is called ***trivial***. A substring is also called a ***factor***. Given strings $\boldsymbol{u}$ and $\boldsymbol{v}$, $\mathrm{lcp}(\boldsymbol{u}, \boldsymbol{v})$ (respectively, $\mathrm{lcs}(\boldsymbol{u}, \boldsymbol{v})$) is the ***longest common prefix*** (respectively, ***longest common suffix***) of $\boldsymbol{u}$ and $\boldsymbol{v}$.

If $\boldsymbol{x}$ is a concatenation of $k \geq 2$ copies of a nonempty string $\boldsymbol{u}$, we write $\boldsymbol{x} = \boldsymbol{u}^k$ and say that $\boldsymbol{x}$ is a ***repetition***; if $k = 2$, we say that $\boldsymbol{x} = \boldsymbol{u}^2$ is a ***square***; if there exist no such integer k and no such $\boldsymbol{u}$, we say that $\boldsymbol{x}$ is ***primitive***. If $\boldsymbol{x} = \boldsymbol{u}^k$, $k \geq 1$, and $\boldsymbol{u}$ is primitive, we call $\boldsymbol{u}$ the ***primitive root*** of $\boldsymbol{x}$. If $\boldsymbol{x} = \boldsymbol{v}^2$ has a proper prefix $\boldsymbol{u}^2$, $|\boldsymbol{u}| < |\boldsymbol{v}| < 2|\boldsymbol{u}|$, we say that $\boldsymbol{x}$ is a ***double square*** and write $\boldsymbol{x} = \mathrm{DS}(\boldsymbol{u}, \boldsymbol{v})$. A square $\boldsymbol{u}^2$ such that $\boldsymbol{u}$ has no square prefix is said to be ***regular***.

For $\boldsymbol{x} = \boldsymbol{x}[1..n]$, $1 \leq i < j \leq j+k \leq n$, the string $\boldsymbol{x}[i+1..j+1]$ is a ***right cyclic shift*** of $\boldsymbol{x}[i..j]$ by 1 position if $\boldsymbol{x}[i] = \boldsymbol{x}[j+1]$; the string $\boldsymbol{x}[i+k..j+k]$ is a right cyclic shift of $\boldsymbol{x}[i..j]$ by k positions if $\boldsymbol{x}[i+k-1..j+k-1]$ is a right cyclic shift of $\boldsymbol{x}[i..j]$ by $k-1$ positions and $\boldsymbol{x}[i+k..j+k]$ is a right cyclic shift of $\boldsymbol{x}[i+k-1..j+k-1]$ of 1 position. 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 left or right cyclic shift.

Strings $\boldsymbol{uv}$ and $\boldsymbol{vu}$ are ***conjugates***, written $\boldsymbol{uv} \sim \boldsymbol{vu}$. We also say that $\boldsymbol{vu}$ is the $|\boldsymbol{u}|^{\text{th}}$ ***rotation*** of $\boldsymbol{x} = \boldsymbol{uv}$, written $R_{|\boldsymbol{u}|}(\boldsymbol{x})$, or the $-|\boldsymbol{v}|^{\text{th}}$ ***rotation*** of $\boldsymbol{x}$, written $R_{-|\boldsymbol{v}|}(\boldsymbol{x})$, while $R_0(\boldsymbol{x}) = \boldsymbol{x}$. 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 whenever $\boldsymbol{x}[i+k..j+k]$ is a cyclic

shift of $\boldsymbol{x}[i..j]$, the two substrings must therefore be conjugates; however, for $k > j-i+1$, the converse does not hold (for example, in $\boldsymbol{x} = abacbaa$, $\boldsymbol{x}[5..7] = baa$ is a conjugate, but not a cyclic shift, of $\boldsymbol{x}[1..3] = aba$).

Lemma 1. [11, Lemma 1.4.2] *Let $\boldsymbol{x}$ be a string of length n and minimum period $\pi \leq n$, and let $j \in 1..n-1$ be an integer. Then $R_j(\boldsymbol{x}) = \boldsymbol{x}$ if and only if $\boldsymbol{x}$ is not primitive and π is divisible by j.*

The following results (Lemmas 2–3) were first stated in [4] without proof as they all follow from the Periodicity Lemma of Fine and Wilf, [6]. Although proofs were later provided in [1], we repeat them here for completeness.

Lemma 2 (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. From Lemma 1 a cyclic shift $R_j(\boldsymbol{x})$ of $\boldsymbol{x}$ can equal $\boldsymbol{x}$ only if $\boldsymbol{x}$ is not primitive. Since here $\boldsymbol{x}$ is primitive, the only occurrences of $\boldsymbol{x}$ are exactly those determined by $\boldsymbol{x}^p$. □

Lemma 3 (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 integers $p \geq 2$ and $q \geq 2$, $\boldsymbol{x_2}\boldsymbol{x}^p\boldsymbol{x_1}$ 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 WLOG 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 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 $\boldsymbol{x_1}\boldsymbol{x}^{p-1}\boldsymbol{x_2}$, and so we can assume WLOG 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 $\boldsymbol{y_2 y^q y_1} = \boldsymbol{v f v'}$, 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. □

The Common Factor Lemma gives rise to the following corollary useful for showing the uniqueness of the canonical factorization of a double square presented in Section 3:

Lemma 4 ([4])**.** *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 3 $\boldsymbol{x} \sim \boldsymbol{y}$. 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 4(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 $\boldsymbol{x} \neq \boldsymbol{y}$.

3. Canonical Factorization of Double Squares

The most general unique factorization of any double square DS($\boldsymbol{u}$,$\boldsymbol{v}$) discussed in this section was presented in [1]. For the sake of completeness, we not only quote the results from [1], but include the proofs as well.

The uniqueness of the factorization allows us to speak of the *canonical factorization of* DS($\boldsymbol{u}$,$\boldsymbol{v}$). This structure has been described before [4, 5, 8, 7, 9], but not as precisely and with more assumptions required or in a weaker form; above all, Lemma 5 establishes the uniqueness of the breakdown with no additional assumptions.

Lemma 5 ([1], Two Squares Factorization Lemma)**.** *If* $\boldsymbol{x} = \mathrm{DS}(\boldsymbol{u}, \boldsymbol{v})$*, 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* $e_1 > e_2$ *(thus* $e_1 \geq 2$*);*

(b) if $|\boldsymbol{u_2}| > 0$*, then* $\boldsymbol{v}$ *is primitive, and if in addition* $e_1 \geq 2$*, then* $\boldsymbol{u}$ *also is primitive.*

In both cases, the factorization is unique.

Proof. In this proof, for a tandem repeat $\boldsymbol{ww}$, we use $\boldsymbol{w}_{[1]}$ to refer to the first $\boldsymbol{w}$, while $\boldsymbol{w}_{[2]}$ to refer to the second $\boldsymbol{w}$.

Let $\boldsymbol{z}$ be the nonempty proper prefix of $\boldsymbol{u}_{[2]}$ that is in addition a suffix 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 $k = \left\lfloor |\boldsymbol{u}|/|\boldsymbol{z}| \right\rfloor \geq 1$ 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 $e_2 \geq 1$. 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 $\boldsymbol{v} = \boldsymbol{uz} = \boldsymbol{u_1}^{e_1}\boldsymbol{u_2}\boldsymbol{u_1}^{e_2}$, as required. To prove uniqueness we consider two cases:

$|\boldsymbol{u_2}| = 0$: 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 4(a).

$|\boldsymbol{u_2}| > 0$: 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 4(b) that $\boldsymbol{u_1} = \boldsymbol{w_1}$ and $e_1 = f_1$. If $e_1 = f_1 = 1$, we observe that $\boldsymbol{v} = \boldsymbol{uu_1} = \boldsymbol{uw_1}$, 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 WLOG 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

$$\boldsymbol{w}^{2k} = \boldsymbol{v}^2 = \boldsymbol{u_1}^{e_1}\boldsymbol{u_2}\boldsymbol{u_1}^{e_1+e_2}\boldsymbol{u_2}\boldsymbol{u_1}^{e_2}, \tag{1}$$

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 Lemma 3 with variables

$$(\boldsymbol{x}, \boldsymbol{y}, p, q) \equiv (\boldsymbol{w}, \boldsymbol{u_1}, 2k, e_1+e_2)$$

to conclude that $\boldsymbol{w} \sim \boldsymbol{u_1}$, thus by (1) that $\boldsymbol{w} = \boldsymbol{u_1}$. Let $\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ be a suffix of $\boldsymbol{u_1}$ so that $\boldsymbol{u_1} = \boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$. By the Synchronization Principle (Lemma 2), (1) implies that $\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ is a prefix of $\boldsymbol{u_1}$, in contradiction to Lemma 1. We conclude that $\boldsymbol{v}$ is primitive.

Now suppose in addition that $e_2 \geq 2$, 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 5, $\boldsymbol{w}^{2k}$ and $\boldsymbol{u_1}^{e_1+e_2}$ have a common prefix $\boldsymbol{u_1}^{e_1}\boldsymbol{u_2}$. Note that $|\boldsymbol{u_1}^{e_1}\boldsymbol{u_2}| \geq |\boldsymbol{v}|+|\boldsymbol{u_1}|$, so that again applying Lemma 3 with variables $(\boldsymbol{x}, \boldsymbol{y}, p, q) \equiv (\boldsymbol{w}, \boldsymbol{u_1}, 2k, e_1+e_2)$, 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 5(b) requires that $e_1 \geq 2$. To see that this condition is not necessary, consider $\boldsymbol{x} = abaababaab$, 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 $\boldsymbol{x} = abaabaabaababaabaabaab$, 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.

Lemma 5 gives credence to the following definition of terminology and notation:

Definition 6. *For a double square* $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v})$ *we call the unique factorization* $\boldsymbol{v}^2 = \boldsymbol{u_1}^{e_1}\boldsymbol{u_2}\boldsymbol{u_1}^{e_1+e_2}\boldsymbol{u_2}\boldsymbol{u_1}^{e_2}$ *guaranteed by Lemma 5 the* ***canonical factorization*** *of* $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v})$ *and denote it by* $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u_1}, \boldsymbol{u_2}, e_1, e_2)$. *The symbol* $\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ *denotes the suffix of* $\boldsymbol{u_1}$ *such that* $\boldsymbol{u_1} = \boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$.

Lemma 5 gives rise to a number of important observations:

Observation 7. *In Lemma 5,* $|\boldsymbol{u_2}| > 0$ *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* $\boldsymbol{v}^2$*;*

(*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* $e_1 = e_2 = 1$.

Proof. (a) $|\boldsymbol{u_2}| = 0$ implies that $\boldsymbol{v} = \boldsymbol{u_1}^{e_1+e_2}$ and since $e_1+e_2 \geq 2$, it follows that $\boldsymbol{v}$ is not primitive.

(b) We show that $|\boldsymbol{u_2}| = 0$ implies $\boldsymbol{u}$ not primitive. First note that $u = u_1{}^{e_1}$. If $e_1 \geq 2$, then it follows directly that u is not primitive. If $e_1 = 1$, then $e_2 = 1$ and so $v = u_1{}^2$ and $u = u_1$ which is a contradiction as $|v| < 2|u|$.

(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 5, $|\boldsymbol{u_2}| > 0$ 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 $e_1 = e_2 = 1$.

□

In the context of Observation 7(f), consider the double square

$$\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v}) = aabaaaabaabaaaab,$$

with $\boldsymbol{u} = aabaa$, $\boldsymbol{v} = aabaaaab$. 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.

4. Rare Factors in Double Squares

In this section we consider a double square $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u_1}, \boldsymbol{u_2}, e_1, e_2)$ with nonempty $\boldsymbol{u_2}$. We present three kinds of factors of $\boldsymbol{v}^2$ that have highly restricted number of occurrences in $\boldsymbol{v}^2$. The first of these, the so-called *inversion factor* (IF for short) was introduced in [4]. The additional two factors introduced here, *right inversion subfactor* (RIS for short) and *left inversion subfactor* (LIS for short), are in fact subfactors of IF. Note that A. Thierry [12] examines rare factors in configurations $\boldsymbol{u_1}^{e_1}\boldsymbol{u_2}\boldsymbol{u_1}^{e_2}$.

Using the canonical factorization of DS($\boldsymbol{u}$,$\boldsymbol{v}$), we have

$$\begin{aligned} \boldsymbol{v}^2 &= (\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1}\boldsymbol{u_2}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1+e_2}\boldsymbol{u_2}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_2} \\ &= (\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1-1}\boldsymbol{u_2}(\mathrm{IF})(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1+e_2-2}\boldsymbol{u_2}(\mathrm{IF})(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_2-1} \qquad (2) \end{aligned}$$

where $\mathrm{IF} = \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}} = R_{|\boldsymbol{u_2}|}(\boldsymbol{u_1})\boldsymbol{u_1}$ is called the ***inversion factor***.

Lemma 8. *Consider a double square* $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u_1}, \boldsymbol{u_2}, e_1, e_2)$ *with nonempty* $\boldsymbol{u_2}$. *Then the inversion factor* IF *has exactly two occurrences in* $\boldsymbol{v}^2$ *that are distance* $|\boldsymbol{v}|$ *apart, as shown in* (2).

Proof. If IF occurs elsewhere in $\boldsymbol{v}^2$, then by the Synchronization Principle (Lemma 2) its primitive subfactor $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ can only align with another occurrence of $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$. Therefore its other subfactor $\overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}$ must align with $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$, thus by Lemma 1 contradicting the primitiveness of $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$. □

The quantity $\mathrm{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$ gives the maximum number of positions the structures $(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1+e_2}$ and $(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_2}$ can be cyclically shifted to the left in $\boldsymbol{v}^2$, while $\mathrm{lcp}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$ gives the maximum number of positions $(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1}$ and $(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1+e_2}$ can be cyclically shifted to the right. In [4], the following lemma limiting the size of $\mathrm{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})+\mathrm{lcp}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$ was proved:

Lemma 9 ([4])**.** *Consider* $\boldsymbol{u_1}^{e_1}\boldsymbol{u_2}\boldsymbol{u_1}^{e_1+e_2}\boldsymbol{u_2}\boldsymbol{u_1}^{e_2}$*, where* $\boldsymbol{u_1}$ *is primitive and* $\boldsymbol{u_2}$ *is a nonempty proper prefix of* $\boldsymbol{u_1}$*,* $e_1 \geq e_2 \geq 1$*, and* $\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ *a suffix of* $\boldsymbol{u_1}$ *so that* $\boldsymbol{u_1} = \boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$*. Then* $lcs(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})+lcp(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}) \leq |\boldsymbol{u_1}|-2$*.*

In fact, in [4] 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 $|\overline{\boldsymbol{w}}| = |\overline{\boldsymbol{u}}_{\boldsymbol{2}}|$; then a stronger result is given (re-phrased in the terminology of this paper):

Lemma 10 ([4])**.** *Consider a double square* $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u_1}, \boldsymbol{u_2}, e_1, e_2)$ *with nonempty* $\boldsymbol{u_2}$*, and let* $p = lcp(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$*,* $s = lcs(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$*. Then any inversion factor in* $\boldsymbol{v}^2$ *is either* $R_i(\mathrm{IF})$ *or* $R_{-j}(\mathrm{IF})$ *for some* $i \in 0..p$ *or some* $j \in 0..s$*. Moreover, for every* $i \in 0..p$ *(respectively,* $j \in 0..s$*), every* $R_i(\mathrm{IF})$ *(respectively,* $R_{-j}(\mathrm{IF})$*) occurs exactly twice in* $\boldsymbol{v}^2$ *with occurrences distance exactly* $|\boldsymbol{v}|$ *apart.*

The following simple lemma will be used for determining a different type of rare factor in a 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 11. *If the substrings* $\boldsymbol{x}[1..n]$ *and* $\boldsymbol{x}[2..n+1]$ *of* $\boldsymbol{x} = \boldsymbol{x}[1..n+1]$ *are conjugates, then* $\boldsymbol{x}[1] = \boldsymbol{x}[n+1]$*.*

Proof. (Due to A. Thierry) Since $\boldsymbol{x}[1..n]$ and $\boldsymbol{x}[2..n+1]$ are conjugates, the frequency of the alphabet symbols in both must be the same. Let $\boldsymbol{x}[1] = a$. Then $\boldsymbol{x}[1..n]$ must have the same number of a's as $\boldsymbol{x}[2..n+1]$, and so $x[n+1] = a$. □

Note that Lemma 11 does not hold if $\boldsymbol{u}$ and $\boldsymbol{u}'$ overlap less. For

$$\underbrace{abbbb}_{\boldsymbol{u}}ba\ldots \text{ and } ab\underbrace{bbbba}_{\boldsymbol{u}'}\ldots,$$

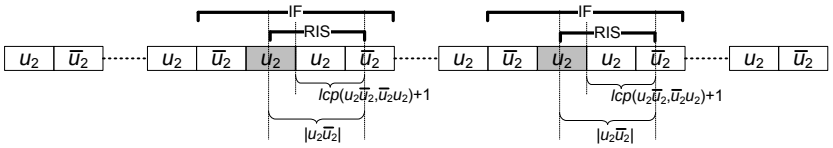
$\boldsymbol{u}$ cannot be cyclically shifted to the right nor $\boldsymbol{u}'$ to the left, yet $\boldsymbol{u}'$ is a rotation of $\boldsymbol{u}$.

Definition 12. *Consider a double square* $\boldsymbol{x} = \mathrm{DS}(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u_1}, \boldsymbol{u_2}, e_1, e_2)$ *with nonempty* $\boldsymbol{u_2}$*. The* ***right inversion subfactor*** *(or* RIS*) is defined to be a factor* $\boldsymbol{x}[i..j]$ *of length* $|\boldsymbol{u_1}|$ *where* $i = (e_1{-}1)|\boldsymbol{u_1}|{+}|\boldsymbol{u_2}|{+}lcp(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$ *and* $j = e_1|\boldsymbol{u_1}|{+}|\boldsymbol{u_2}|{+}lcp(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}){+}1$*. Similarly, the* ***left inversion subfactor*** *(or* LIS*) is a factor* $\boldsymbol{x}[i..j]$ *of length* $|\boldsymbol{u_1}|$ *where* $i = e_1|\boldsymbol{u_1}|{+}|\boldsymbol{u_2}|{-}lcs(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$ *and* $j = (e_1{+}1)|\boldsymbol{u_1}|{+}|\boldsymbol{u_2}|{-}lcs(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}){-}1$*.*

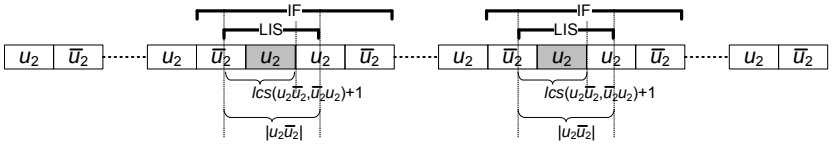
Note that another RIS also naturally occurs at position $(e_1-1)|\boldsymbol{u_1}| + |\boldsymbol{u_2}| + \mathrm{lcp}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}) + |\boldsymbol{v}|$, and another LIS at position $e_1|\boldsymbol{u_1}| + |\boldsymbol{u_2}| - \mathrm{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}) + |\boldsymbol{v}|$.

It is possible to view RIS in this way: let $\boldsymbol{x}[i..j]$ be the maximum right cyclic shift of the rightmost $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ of $(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1}$, respectively of $(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1+e_2}$. Then RIS is $\boldsymbol{x}[i{+}1, j{+}1]$. Similarly, let $\boldsymbol{x}[i, j]$ be the maximum left cyclic shift of the first $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ of $(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1+e_2}$, respectively of $(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_2}$. Then LIS is $\boldsymbol{x}[i{-}1, j{-}1]$.

For a better understanding, we illustrate the two natural occurrences of RIS in the following diagram:



and the two natural occurrences of LIS as follows:



Lemma 13. *Consider a double square* $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u_1}, \boldsymbol{u_2}, e_1, e_2)$ *with nonempty* $\boldsymbol{u_2}$*. Let* $\boldsymbol{p}$ *be the longest common prefix and* $\boldsymbol{s}$ *the longest common suffix of* $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ *and* $\overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}$*. Then there is no occurrence of* RIS*, respectively* LIS*, that is a factor of* $\boldsymbol{s}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1+e_2}\boldsymbol{u_2}\boldsymbol{p}$*.*

Proof. Let us assume that there is an occurrence of RIS that is a factor of $\boldsymbol{s}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1+e_2}\boldsymbol{u_2}\boldsymbol{p}$. Therefore RIS must be a rotation of $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ since its length is $|\boldsymbol{u_1}|$. Consider $\boldsymbol{w}$, the maximum right cyclic shift of the rightmost $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ of $(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1}$: $\boldsymbol{w}$ overlaps with RIS except for one symbol and RIS is its rotation,

so we can apply Lemma 11 to conclude that $\boldsymbol{w}$ can be right cyclically shifted one more position, a contradiction.

The proof for LIS follows the same line of reasoning. □

5. New Periodicity Lemma Revisited

Let us first state the original New Periodicity Lemma (NPL) in terms used in this paper:

Lemma 14 ([5], New Periodicity Lemma)**.** *Let* $\boldsymbol{x} = \mathrm{DS}(\boldsymbol{u}, \boldsymbol{v})$*, where we require that* $\boldsymbol{u}^2$ *be regular and that* $\boldsymbol{v}$ *be primitive. Then for all integers* k *and* w *such that* $0 \leq k < |\boldsymbol{v}| - |\boldsymbol{u}|$ *and* $|\boldsymbol{v}| - |\boldsymbol{u}| < w < |\boldsymbol{v}|$*,* $w \neq |\boldsymbol{u}|$*,* $\boldsymbol{x}[k{+}1..\,k{+}2w]$ *is not a square.*

First note that by Observation 7(d)-(e), the requirement that $\boldsymbol{v}$ be primitive is redundant: the fact that $\boldsymbol{u}^2$ is regular forces $|\boldsymbol{u_2}| > 0$ as well as the primitiveness of $\boldsymbol{v}$. Also by Observation 7(f), the regularity of $\boldsymbol{u}^2$ implies that in the canonical factorization $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u_1}, \boldsymbol{u_2}, e_1, e_2)$, $e_1 = e_2 = 1$; in other words, NPL applies only to a small subset of possible double squares. Therefore $\boldsymbol{u} = \boldsymbol{u_1 u_2}$, $\boldsymbol{v} = \boldsymbol{u_1 u_2 u_1}$, and $|\boldsymbol{v}| - |\boldsymbol{u}| = |\boldsymbol{u_1}|$. Thus in loose terms, NPL forbids a square $\boldsymbol{w}^2$ starting in $\boldsymbol{u_1}$ with size $|\boldsymbol{u_1}| < |\boldsymbol{w}| < |\boldsymbol{v}|$, with the possible exception of size $|\boldsymbol{u}|$.

Here we present a theorem that extends the result to all possible double squares; the meaning of the suffix u' of v will be illuminated in the proof of the theorem.

Theorem 15. *Consider a double square* $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v})$ *and let* $\boldsymbol{u'}$ *be a suffix of* $\boldsymbol{v}$ *such that* $\boldsymbol{v} = \boldsymbol{u u'}$*. Let* $\boldsymbol{w}^2$ *be any square that is a factor 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 embark on the proof of the theorem, let us discuss how it relates to the original NPL. As mentioned above, for a very specific double square, NPL 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 15 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 NPL 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 double squares without any additional conditions or constraints — more than a fair trade-off in our opinion.

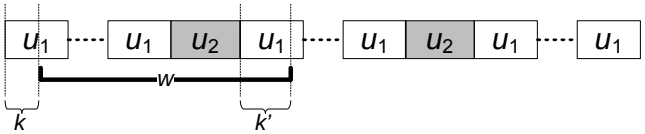
Proof. If $|w| \geq |v|$, then $w = v$ since w^2 is a factor of v^2, and thus case (a) holds. Hence, for the remainder of the proof we can assume that $|w| < |v|$. Let $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u_1}, \boldsymbol{u_2}, e_1, e_2)$. Then $\boldsymbol{u'} = \boldsymbol{u_1}^{e_2}$ and the primitive root of $\boldsymbol{u'}$ is $\boldsymbol{u_1}$.

We first deal with the case $|\boldsymbol{u_2}| = 0$. By Lemma 5, $\boldsymbol{u_1}$ 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 Lemma 3, 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 $|u| \leq |w| < |v|$. We employ the same notation as in the proof of Lemma 5: thus, for example, $\boldsymbol{w}_{[1]}$ refers to the first occurrence of $\boldsymbol{w}$ in $\boldsymbol{w}^k$, $\boldsymbol{w}_{[2]}$ to the second, etc.

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) $\boldsymbol{w}^2$ starts in the first $\boldsymbol{u_1}$ of $\boldsymbol{u_1}^{e_1}$ and ends in the first $\boldsymbol{u_1}$ of $\boldsymbol{u_1}^{e_1+e_2}$



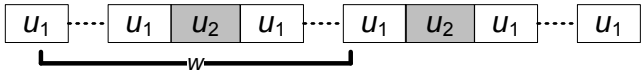
Since $|\boldsymbol{w}| > |\boldsymbol{u}| = |\boldsymbol{u_1}^{e_1}|+|\boldsymbol{u_2}|$, $k < k'$.

(i) $k' \leq \mathrm{lcp}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$
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, $k = k'$, a contradiction. This case is not possible.

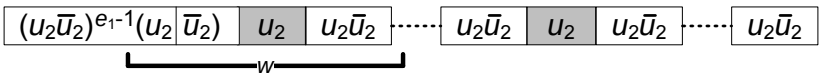
(ii) $k' > \text{lcp}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$
Here $\boldsymbol{w}_{[1]}$ contains the first RIS, and so $\boldsymbol{w}_{[2]}$ must contain an occurrence of RIS. Since $\boldsymbol{w}_{[2]}$ is a factor in $\boldsymbol{u_1}^{e_1+e_2}\boldsymbol{u_2}$, therefore by Lemma 13 $\boldsymbol{w}_{[2]}$ must contain the second RIS and so $|\boldsymbol{w}| \geq |\boldsymbol{v}|$, a contradiction. This case is not possible.

(2) $\boldsymbol{w}^2$ starts in the first $\boldsymbol{u_1}$ of $\boldsymbol{u_1}^{e_1}$ and ends past the first $\boldsymbol{u_1}$ of $\boldsymbol{u_1}^{e_1+e_2}$.



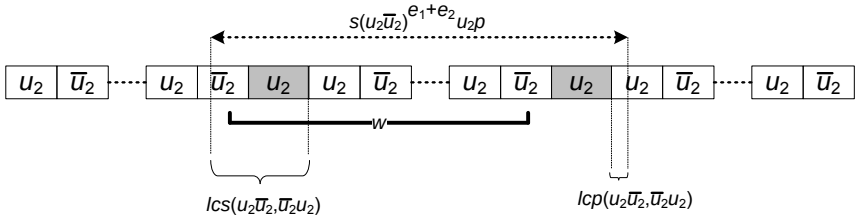
The same argument as in (1)(ii) gives $|\boldsymbol{w}| \geq |\boldsymbol{v}|$, a contradiction. This case is not possible.

(3) $\boldsymbol{w}^2$ starts in $\boldsymbol{u_1}^{e_1-1}\boldsymbol{u_2}$ but not in the first $\boldsymbol{u_1}$.



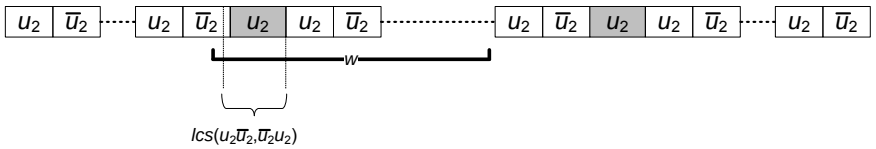
Then $e_1 > 1$ and since $|\boldsymbol{w}| > |\boldsymbol{u}| = |\boldsymbol{u_1}^{e_1}\boldsymbol{u_2}|$, $\boldsymbol{w}_{[1]}$ ends past the first $\boldsymbol{u_1}$ of $\boldsymbol{u_1}^{e_1+e_2}$. Therefore, $\boldsymbol{w}_{[1]}$ contains RIS and so $|\boldsymbol{w}| \geq |\boldsymbol{v}|$, i.e. this case is not possible.

(4) $\boldsymbol{w}^2$ starts in the suffix $\overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}$ of $\boldsymbol{u}$ whose length $\leq \text{lcs}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$.



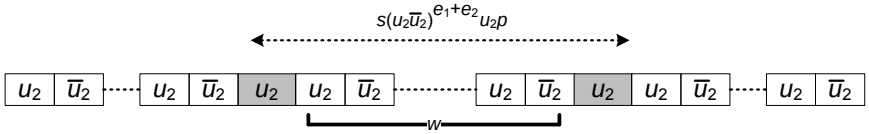
Here $\boldsymbol{w}_{[1]}$ is a factor in $\boldsymbol{s}\boldsymbol{u_1}^{e_1+e_2}\boldsymbol{u_2}\boldsymbol{p}$, where $\boldsymbol{s}$ is the maximal common suffix and $\boldsymbol{p}$ the maximal common prefix of $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ and $\overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}$. 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 (Lemma 3), the primitive root of $\boldsymbol{w}$ is a conjugate of $\boldsymbol{u_1}$, i.e. case (c) holds.

(5) $\boldsymbol{w}^2$ starts in the suffix $\overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u}_{\boldsymbol{2}}$ of $\boldsymbol{u}$ whose length $>$ lcs$(\boldsymbol{u}_{\boldsymbol{2}}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u}_{\boldsymbol{2}})$.



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

(6) $\boldsymbol{w}^2$ starts past the first $\boldsymbol{u}$.



The same argument as in (4) shows that the primitive root of $\boldsymbol{w}$ is a conjugate of $\boldsymbol{u}_{\boldsymbol{1}}$ and so case (c) holds.

□

6. Conclusion

We presented a unique factorization, referred to as the *canonical factorization*, of a double square; that is, a configuration of two squares starting at the same position of "comparable" lengths. Utilizing the canonical factorization we discussed so-called rare factors; that is, factors occurring in a few well-defined positions in a double square. The existence of the rare factors RIS and LIS is then used to establish existential limits for a third square (Theorem 15), greatly generalizing the New Periodicity Lemma. The theorem in comparison to NPL vastly extends the range of applicability from a very specific type of double square DS($\boldsymbol{u}$,$\boldsymbol{v}$) ($\boldsymbol{u}^2$ must be regular) to any type of double square.

References

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

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

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

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

[5] K. Fan, S.J. Puglisi, W.F. Smyth, and A. Turpin. A new periodicity lemma. *SIDMA*, 20-3:656–668, 2006.

[6] N. J. Fine and H. S. Wilf. Uniqueness theorems for periodic functions. *Proceedings of the American Mathematical Society*, 16(1):pp. 109–114, 1965.

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

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

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

[10] J. Simpson. Intersecting periodic words. *Theoretical Computer Science*, 374:58–65, 2007.

[11] B. Smyth. *Computing Patterns in Strings.* ACM Press Bks. Pearson/Addison-Wesley, 2003.

[12] A. Thierry. Combinatorics of the interrupted period. *Proceedings of the Prague Stringology Conference*, 2015.