Two Squares Canonical Factorization*

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

[1] Department of Computing and Software
McMaster University, Hamilton, Ontario, Canada
{baih3,franek,smyth}@mcmaster.ca
[2] School of Computer Science & Software Engineering
University of Western Australia

Abstract. We present a new combinatorial structure in a string: a canonical factorization for any two squares that occur at the same position and satisfy some size restrictions. We believe that this canonical factorization will have application to related problems such as the New Periodicity Lemma, Crochemore-Rytter Three Squares Lemma, and ultimately the maximum-number-of-runs conjecture.

Keywords: string, primitive string, square, double square, factorization

1 Introduction

In 1995 Crochemore and Rytter [2] 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.* [4] considered a special case of such two squares prefixes of $\boldsymbol{x}$ with a third square possibly offset some distance to the right; they proved a *New Periodicity Lemma* describing conditions under which the third square could not exist. Since that time there has been considerable work done [1,5,6,8] in an effort to specify more precisely the combinatorial structure of the string in the neighbourhood of such two squares.

In this paper we present a unique *canonical factorization* into primitive strings of what we call *double squares* – i.e. two squares starting at the same position and satisfying some size restrictions. The notion of double squares and their unique factorization can be traced to Lam [7]. A version of the factorization for more specific double squares was presented in [3]. Here we present it in full generality. In conclusion we indicate how this result can be applied to the proof of New Periodicity Lemma.

2 Preliminaries

In this section we develop the basic combinatorial tools that will be used to determine a canonical factorization for a double square. Chief among these are the Synchronization Principle (see Lemma 2), and the Common Factor Lemma (see Lemma 3), that lead to the main result, the Two Squares Factorization Lemma (see Lemma 6).

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

* This work was supported by the *Natural Sciences and Engineering Research Council of Canada*

be a ***prefix*** (respectively, ***substring***, ***suffix***) of $\boldsymbol{x}$; a ***proper prefix*** (respectively, ***proper substring***, ***proper suffix***) if $|\boldsymbol{u}| < |\boldsymbol{x}|$ (respectively, $|\boldsymbol{v}| < |\boldsymbol{x}|$, $|\boldsymbol{w}| < |\boldsymbol{x}|$). 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 $\boldsymbol{x} = \boldsymbol{v}^2$ has a proper prefix $\boldsymbol{u}^2$, $|su| < |\boldsymbol{v}| < 2|\boldsymbol{u}|$, 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 $\boldsymbol{x}[i{+}k..j{+}k]$ 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.

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 $\boldsymbol{x}$, written $R_{|\boldsymbol{u}|}(\boldsymbol{x})$, or the $-|\boldsymbol{v}|^{\text{th}}$ ***rotation*** of $\boldsymbol{x}$, written $R_{-|\boldsymbol{v}|}(\boldsymbol{x})$, while $R_0(\boldsymbol{x}) = R_{-|\boldsymbol{x}|}(\boldsymbol{x}) = \boldsymbol{x}$ 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.

In the following lemma, the symbol $\mid$ denotes *divisibility*, i.e. $a \mid b$ means that a is divisible by b.

Lemma 1 [9, Lemma 1.4.2] *Let $\boldsymbol{x}$ be a string of length n and minimum period $\pi \leq n$, and let $j = 1, \ldots, n{-}1$ be an integer. Then $R_j(\boldsymbol{x}) = \boldsymbol{x}$ if and only if $\boldsymbol{x}$ is not primitive ($\pi < n$, $\pi \mid n$) and $j \mid \pi$.*

The following results (Lemmas 2–6) are based on the development given in [3]. Though Lemmas 2 and 3 are folklore, we include their proofs.

Lemma 2 (Synchronization Principle) *The primitive string $\boldsymbol{x}$ occurs exactly p times in $\boldsymbol{x_2 x^p x_1}$, 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 rotation $R_j(\boldsymbol{x})$ 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 $\boldsymbol{x}^p$. $\square$

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 nonnegative integers p and q, $\boldsymbol{x_2 x^p x_1}$ and $\boldsymbol{y_2 y^q y_1}$ 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 $\boldsymbol{f} = \boldsymbol{yx}$, 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 $\boldsymbol{x_2x^px_1}$ and $\boldsymbol{y_2y^qy_1}$. Then $\boldsymbol{x_2x^px_1} = \boldsymbol{ufu'}$ for some $\boldsymbol{u}$ and $\boldsymbol{u'}$. If $|\boldsymbol{u}| \geq |\boldsymbol{x}|$, then $\boldsymbol{f}$ is a factor of $\boldsymbol{x_1x^{p-1}x_2}$, 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 $\boldsymbol{y_2y^qy_1} = \boldsymbol{vfv'}$, 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. □

Note that Lemma 3 could be equivalently stated in a more general form:

Lemma 4 *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, $\boldsymbol{x_2x^px_1}$ and $\boldsymbol{y_2y^qy_1}$ have a common factor of length $|\boldsymbol{x}|{+}|\boldsymbol{y}|$, then the primitive root $\overline{\boldsymbol{x}}$ of $\boldsymbol{x}$ and the primitive root $\overline{\boldsymbol{y}}$ of $\boldsymbol{y}$ are conjugates.*

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

Lemma 5 *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 $\boldsymbol{x^px_1} = \boldsymbol{y^qy_1}$ 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 $\boldsymbol{x^px_1} = \boldsymbol{y^qy_1}$ has a common factor of length at least $|\boldsymbol{x}|{+}|\boldsymbol{y}|$, hence the result. □

Note that in Lemma 5(b) the requirement $p \geq 2,\ q \geq 2$ is essential. For instance, $\boldsymbol{x} = aabb,\ \boldsymbol{x_1} = aa$ and $p = 2$ yields $\boldsymbol{x^px_1} = aabbaabbaa$, identical to $\boldsymbol{y^qy_1}$ produced by $\boldsymbol{y} = aabbaabba,\ \boldsymbol{y_1} = a$ and $q = 1$ — but of course $\boldsymbol{x} \neq \boldsymbol{y}$.

3 Main Result – Two Squares Factorization Lemma

The next lemma specifies the structure imposed by the occurrence of two squares at the same position in a string. This structure has been described before, see [3,4,5,6,7], but not as precisely and with more assumptions required; above all, Lemma 6 establishes the uniqueness of the breakdown.

Lemma 6 (Two Squares Factorization Lemma) *For a double square* $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v})$, *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* $e_1 > e_2 \geq 1$*;*
(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. If we have $\boldsymbol{u}^k$, $k \geq 2$, we refer to the first copy of $\boldsymbol{u}$ as $\boldsymbol{u}_{[1]}$, to the second copy of $\boldsymbol{u}$ as $\boldsymbol{u}_{[2]}$ etc.

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 suffix $\boldsymbol{z'}$ 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:

(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 5(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 5(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 $\boldsymbol{u} = \boldsymbol{u_1u_2} = \boldsymbol{w_1}^{f_1}\boldsymbol{w_2}$ and $\boldsymbol{v} = \boldsymbol{u_1u_2u_1} = \boldsymbol{w_1}^{f_1}\boldsymbol{w_2}\boldsymbol{w_1}^{f_2}$, 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 $\boldsymbol{u_1u_2} = \boldsymbol{w_1}^{f_1}\boldsymbol{w_2} = \boldsymbol{u_1}^{f_1}\boldsymbol{w_2}$, 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 Common Factor Lemma 3 to conclude that $\boldsymbol{w} \sim \boldsymbol{u_1}$, thus by (1) that $\boldsymbol{w} = \boldsymbol{u_1}$. But (1) then requires that the primitive string $\boldsymbol{u_1} = \boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ aligns with $\boldsymbol{u_2u_1}$, and so $\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ 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 6, $\boldsymbol{w}^{2k}$ 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 Common Factor Lemma 3, 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 6(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 $\boldsymbol{u_1}$ is *not* primitive.

Lemma 6 gives credence to the following definition of terminology and notation:

Definition 7 *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 6, 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 6 also gives rise to a number of important observations:

Observation 8 *In Lemma 6,* $|\boldsymbol{u_2}| > 0$ *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 6, $|\boldsymbol{u_2}| > 0$ 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 8(f), consider the double square $\mathrm{DS}(\boldsymbol{u},\boldsymbol{v})$ 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.

Now, following [3], consider the case $|\boldsymbol{u_2}| > 0$ of Lemma 6 and set $\boldsymbol{u_1} = \boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$. Thus $\boldsymbol{v}^2$ becomes

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

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 9 *Consider a double square* $\mathrm{DS}(\boldsymbol{u},\boldsymbol{v}) = (\boldsymbol{u_1},\boldsymbol{u_2},e_1,e_2)$ *with a non-empty* $\boldsymbol{u_2}$. *Then the inversion factor* IF *have exactly two occurrences in* $\boldsymbol{v}^2$ *exactly a distance of* $|\boldsymbol{v}|$ *apart as shown in* (2).

Proof. If IF occurs elsewhere in $\boldsymbol{v}^2$, by the Synchronization principle its subfactor $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ must align with an occurrence of $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ as it is primitive. Thus, its subfactor $\overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2}$ must align with $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$, contradicting the primitiveness of $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$, see Lemma 1. □

The quantity lcs$(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$ gives the maximal 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 lcp$(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$ gives the maximal number of positions the structures $(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1}$ and $(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})^{e_1+e_2}$ can be cyclically shifted to the right. In [3], the following lemma limiting the size of lcs$(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$+lcp$(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$ was given.

Lemma 10 ([3]) *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* $\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ *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 [3] 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 paper):

Lemma 11 ([3]) *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* $p = lcp(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$ *and* $s = lcs(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}, \overline{\boldsymbol{u}}_{\boldsymbol{2}}\boldsymbol{u_2})$*. Then any inversion factor in* $\boldsymbol{v}^2$ *is either* $R_i(\mathrm{IF})$ *or* $R_{-j}(\mathrm{IF})$ *for some* $i \in 0, \ldots, p$ *or some* $j \in 0, \ldots, s$*. Moreover, every* $R_i(\mathrm{IF})$ *or* $R_{-j}(\mathrm{IF})$ *appear exactly twice in* $\boldsymbol{v}^2$ *exactly a distance* $|\boldsymbol{v}|$ *apart for every* $i \in 0, \ldots, p$ *and every* $j \in 0, \ldots, s$*.*

4 Possible application to New Periodicity Lemma

Some years ago a New Periodicity Lemma was published [4], showing 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 was complex, breaking down into 14 subcases, and required a very strong condition that the shorter of the two squares be regular.

Lemma 12 ([4], 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 8, the requirement that v be primitive is redundant; the fact that u^2 is regular necessarily forces the primitivness of v. Also note that the regularity of u^2 necessarily implies that in the canonical factorization of $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u_1}, \boldsymbol{u_2}, e_1, e_2)$, $e_1 = e_2 = 1$.

Consider $\mathrm{DS}(\boldsymbol{u}, \boldsymbol{v}) = (\boldsymbol{u_1}, \boldsymbol{u_2}, 1, 1)$. Let $\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ be a suffix of $\boldsymbol{u_1}$ such that $\boldsymbol{u_1} = \boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$. The canonical factorization thus has the form

$(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})\boldsymbol{u_2}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})\boldsymbol{u_2}(\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}})$.

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.

If for instance $\boldsymbol{w}$ starts in the first $\boldsymbol{u_2}$ and ends in the fourth $\boldsymbol{u_2}$, then $\boldsymbol{w}$ contains fully the IF, so the second $\boldsymbol{w}$ has to as well, 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 IF, but still knowing that $\boldsymbol{u_2}\overline{\boldsymbol{u}}_{\boldsymbol{2}}$ is

primitive and also all its rotations are primitive, using the Synchronization principle 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, we believe, and it is our immediate goal for future research, that the canonical factorization will not only provide us with a significantly simplified proof of New Periodicity Lemma, but will also allow us to significantly reduce the conditions on $\boldsymbol{u}^2$ from $\boldsymbol{u}$ being regular to just being primitive. We also believe that the canonical factorization in the same way will not only provide a simpler proof of Crochemore-Rytter Three Squares Lemma, but will extend the applicability of the lemma to three squares when any of the squares is primitive (the original lemma requires that the smallest square be primitive).

5 Conclusion and future work

We presented a unique factorization of a double square, i.e. a configuration of two squares $\boldsymbol{u}^2$ and $\boldsymbol{v}^2$ starting at the same position and satisfying $|u| < |v| < 2|u|$. We call this factorization the *canonical factorization*. It has very strong combinatorial properties as it is an almost periodic repetition of a primitive string. We indicated that we would like to use this new insight into the structure of double squares in improving the New Periodicity Lemma [4] and Crochemore-Rytter's Three Squares Lemma [2] and simplifying their proofs. As of preparing this final version of the Prague Stringology Conference 2014 proceedings, we are happy to report that the canonical factorization presented here indeed greatly simplified and generalized both. The follow-up work will focus on presenting of these results in a near future.

References

1. W. Bland and W. F. Smyth: *Overlapping squares: the general case characterized & applications.* submitted for publication, 2014.
2. M. Crochemore and W. Rytter: *Squares, cubes, and time-space efficient string searching.* Algorithmica, 13 1995, pp. 405–425.
3. A. Deza, F. Franek, and A. Thierry: *How many double squares can a string contain?* submitted for publication, 2013.
4. K. Fan, S. Puglisi, W. F. Smyth, and A. Turpin: *A new periodicity lemma.* SIAM Journal on Discrete Mathematics, 20 2006, pp. 656–668.
5. F. Franek, R. C. G. Fuller, J. Simpson, and W. F. Smyth: *More results on overlapping squares.* Journal of Discrete Algorithms, 17 2012, pp. 2–8.
6. E. Kopylova and W. F. Smyth: *The three squares lemma revisited.* Journal of Discrete Algorithms, 11 2012, pp. 3–14.
7. N. H. Lam: *On the number of squares in a string. AdvOL-Report 2013/2, McMaster University,* 2013.
8. J. Simpson: *Intersecting periodic words.* Theoretical Computer Science, 374 2007, pp. 58–65.
9. B. Smyth: *Computing Patterns in Strings*, Pearson Addison-Wesley, 2003.