Discrete Applied Mathematics 212 (2016) 30–36

Contents lists available at ScienceDirect

Discrete Applied Mathematics

journal homepage: www.elsevier.com/locate/dam

The New Periodicity Lemma revisited

Haoyue Bai [a], Frantisek Franek [a,*], William F. Smyth [a,b,c]

[a] Department of Computing and Software, McMaster University, Hamilton, Ontario, Canada
[b] School of Engineering & Information Technology, Murdoch University, Perth, Western Australia, Australia
[c] School of Computer Science & Software Engineering, University of Western Australia, Perth, Western Australia, Australia

ARTICLE INFO

Article history:
Received 30 September 2014
Received in revised form 20 January 2016
Accepted 2 May 2016
Available online 1 June 2016

Keywords:
String
Square
Canonical factorization
Double square
New Periodicity Lemma

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.

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 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 neighbourhood 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

* Corresponding author.
E-mail addresses: baih3@mcmaster.ca (H. Bai), franek@mcmaster.ca (F. Franek), bill@arg.cas.mcmaster.ca (W.F. Smyth).

http://dx.doi.org/10.1016/j.dam.2016.05.003
0166-218X/

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 $|\boldsymbol{u}|$th ***rotation*** of $\boldsymbol{x} = \boldsymbol{uv}$, written $R_{|\boldsymbol{u}|}(\boldsymbol{x})$, or the $-|\boldsymbol{v}|$th ***rotation*** of $\boldsymbol{x}$, written $R_{-|\boldsymbol{v}|}(\boldsymbol{x})$, while $R_0(\boldsymbol{x}) = \boldsymbol{x}$. 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 $\boldsymbol{x_2 x^p x_1}$, 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 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 $q \geq 2$, $\boldsymbol{x_2 x^p x_1}$ 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 $\boldsymbol{x_2 x^p x_1}$ and $\boldsymbol{y_2 y^q y_1}$. Then $\boldsymbol{x_2 x^p x_1} = \boldsymbol{ufu'}$ for some $\boldsymbol{u}$ and $\boldsymbol{u'}$. If $|\boldsymbol{u}| \geq |\boldsymbol{x}|$, then $\boldsymbol{f}$ is a factor of $\boldsymbol{x_1 x^{p-1} x_2}$, 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 $\tilde{\boldsymbol{x}}^p$.