McMaster University

Advanced Optimization Laboratory

Title:

On a Lemma of Crochemore and Rytter

Authors:

Haoyue Bai, Antoine Deza, and Frantisek Franek

AdvOL-Report No. 2014/1

July 2014, Hamilton, Ontario, Canada

On a Lemma of Crochemore and Rytter

Haoyue Bai, Antoine Deza, and Frantisek Franek

Abstract

Crochemore and Rytter introduced in 1995 a structural lemma on three squares starting at the same position. This influential lemma has been used by many researchers in the field of periodicities in strings. In particular, Fraenkel and Simpson used it in 1998 to obtain a universal upper bound for the maximum number of distinct squares occurring in a string. We present a generalization of Crochemore and Rytter's lemma by exploiting the combinatorics of two squares starting at the same position.

1 Introduction

Crochemore and Rytter [2] introduced in 1995 the following Lemma 1.

Lemma 1 ([2])**.** *Let $u^2 \neq v^2$ be proper prefixes of w^2 and let u, v, and w be primitive, then $|u|+|v| \leq |w|$.*

Lemma 1 has been used by many researchers including Kolpakov and Kucherov [9], Stoye and Gusfield [12], Fan, Puglisi, Smyth, and Turpin [5], Simpson [11]. Lemma 1 was essential for the 1998 result by Fraenkel and Simpson [7] giving a universal upper bound of $2n$ for the number of distinct squares in a string of length n. Note that for the problem of distinct squares, every type of square is only counted once, i.e. the types, rather than the occurrences, are counted. For illustration, $aabaab$ contains the following three underlined squares $\underline{aa}baab$, $\underline{aabaab}$ and $aab\underline{aa}b$ while the number of distinct squares is 2: aa and $aabaab$. Ilie [8] provided in 2005 an alternate proof of the main theorem of [7] not directly using Lemma 1. Noticing that the proof of Lemma 1 by Crochemore and Rytter only requires the primitiveness of the shortest square, Fraenkel and Simpson [7] proposed the following strengthening referred to as three-prefix-square Lemma in [3] where additional context and references can be found.

Lemma 2 ([7])**.** *Let $u^2 \neq v^2$ be proper prefixes of w^2 and let the shorter of the two strings u and v be primitive, then $|u|+|v| \leq |w|$.*

Fraenkel and Simpson illustrated the necessity of the primitiveness for the shortest square with the following example: $u = a^2$, $v = a^4$, and $w = a^5$. We present a further strengthening based on the recently investigated structural properties of two squares starting at the same position, see [1, 4]. The proof of Lemma 3 is given in Section 3 not to impede the clarity of the exposition.

Lemma 3. *Let $u^2 \neq v^2$ be proper prefixes of w^2, then $|u| + |v| \leq |w|$ unless $u = {v_1}^t$, $v = {v_1}^{p_1}v_2$, and $w = {v_1}^{p_1}v_2{v_1}^{p_2}$ where v_1 is primitive, v_2 a proper possibly empty prefix of v_1, $t > p_2$, and $p_1 \geq p_2 \geq 1$.*

Lemma 3 shows that the strings (u, v, w) violating $|u|+|v| \leq |w|$ consist of two types; one corresponding to the example given by Fraenkel and Simpson. Corollary 4 illustrates that Lemma 3 is a generalization of Lemma 2.

Corollary 4. *Let u^2 be a proper prefix of v^2 that is a proper prefixes of w^2 and let u be primitive, then $|u|+|v| \leq |w|$. Moreover, if $|u| < |v| < 2|u|$ and either v or w is primitive, then $|u|+|v| \leq |w|$.*

Proof. Let us assume by contradiction that $|u|+|v| > |w|$. Then by Lemma 3, $u = {v_1}^t$, $v = {v_1}^{p_1}v_2$ and $w = {v_1}^{p_1}v_2{v_1}^{p_2}$ for a primitive v_1, a proper possibly empty prefix v_2 of v_1, and $t > p_2$, $p_1 \geq p_2 \geq 1$. If u is primitive, $t = 1$ and so $t > p_2 \geq 1$ is a contradiction. If $|v| < 2|u|$, then ${v_1}^{p_1}v_2v_1$ is a prefix of ${v_1}^{2t}$, which can only be true when v_2 is empty due to Lemma 6. If v is primitive, then $p_1 = 1$ and so $p_2 = 1$ and so $u = {v_1}^t$, $t > 1$ and $v = v_1$ and $w = {v_1}^2$, and so $|u| \geq |w|$, a contradiction. If w is primitive, then $w = v_1$, and so $|w| = |v|$, a contradiction. □

2 Preliminaries and Notations

For a string x, we use the indexing from 1, i.e. $x[1]$ refers to the first symbol of x, $x[2]$ the second symbol of x etc. The string $x = x[1\,..\,n]$ is a sequence of n symbols and the length, also called size, of a string x is denoted by $|x|$. The same range notation is used for *substring*, also called *factor*, i.e. $x[i\,..\,j]$ refers to the string consisting of $x[i]x[i{+}1]\dots\ x[j]$. The string of length 0 is called the *empty string*. Given a string $x = x[1\,..\,n]$ and $1 \leq i \leq n$, the substring $x[1\,..\,i]$, respectively $x[i\,..\,n]$, is called a *prefix*, respectively *suffix*, of x and we speak of a *proper prefix*, respectively *proper suffix*, if $i \neq n$, respectively $i \neq 1$. For an integer $n \geq 2$, the nth *power* of a string x, denoted x^n, is a concatenation of n copies of x. In particular, x^2 is referred to as a *square*. A string x is *primitive* if it is not a power of at least 2 of some non-empty string. For a string x, the unique shortest primitive u so that $x = u^k$ for some integer $k \geq 1$ is called the *primitive root* of x. For two substrings y and z of x, $lcs(y, z)$ refers to the length of the longest common suffix of y and z, while $lcp(y, z)$ refers to the length of the longest common prefix of y and z.

A right shift by one position of a substring $x[i\,..\,j]$ is the substring $x[i{+}1\,..\,j{+}1]$. The shift is referred to as *cyclic*, if $x[i] = x[j{+}1]$. In such case, we say that the substring $x[i\,..\,j]$ can be *cyclically shifted one position to the right* or *right cyclically shifted by one position*. A substring $x[i\,..\,j]$ can be *cyclically shifted right by k positions* if each of the substrings $x[i\,..\,j]$, $\dots$, $x[i{+}k{-}1\,..\,j{+}k{-}1]$ can be cyclically shifted by one position to the right. For instance, for $x[1\,..\,5] = abaaa$, the substring $x[1\,..\,2] = ab$ can be cyclically shifted right by 1 position, but not by 2 positions; similarly $x[1\,..\,3] = aba$ can be cyclically shifted right by 1 position, but not by 2 positions; if $x[1\,..\,5] = aabaa$, then $x[1\,..\,3] = aab$ can be cyclically shifted by 2 positions to the right while $x[1\,..\,2]$ cannot be cyclically shifted by 3 positions.

Similarly, a left shift by one position of a substring $x[i\,..\,j]$ is the substring $x[i-1\,..\,j-1]$. The shift is referred to as *cyclic*, if $x[i-1] = x[j]$. In such case we say that the substring $x[i\,..\,j]$ can be *cyclically shifted one position to the left* or *left cyclically shifted by one position.* A substring $x[i\,..\,j]$ can be *cyclically shifted left by k positions* if each of the substrings $x[i\,..\,j]$, …, $x[i-k+1\,..\,j-k+1]$ can be cyclically shifted by one position to the left. Strings x and y are *conjugates* if $x = uv$ for some strings u and v and $y = vu$. Equivalently, x is a *rotation* of y or that y is a *rotation* of x. If either $|u| = 0$ or $|v| = 0$, we speak of a *trivial rotation.* Note that a left cyclic shift of $x[i\,..\,j]$ is a rotation of $x[i\,..\,j]$, i.e. they are conjugates, similarly for a right cyclic shift.

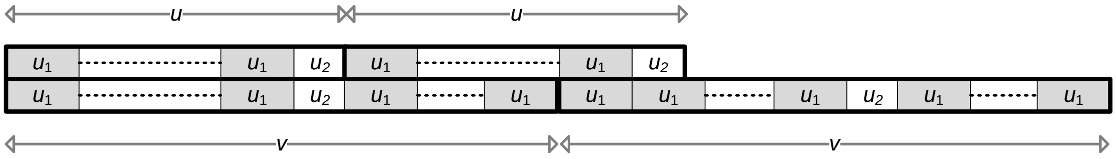
The notion of double squares and their factorization can be traced to Lam [10], and was further investigated and generalized by Deza, Franek, and Thierry [4] and by Bai, Franek, and Smyth [1].

Definition A *double square* (u, v) consists of a square u^2 that is a proper prefix of a square v^2. A double square (u, v) is *balanced* if u and v are *proportional*; that is, if $|v| < 2|u|$.

Lemma 5 (Two-Square Factorization Lemma [1])**.** *Given a balanced double square (u, v), there is a unique primitive string u_1 such that $u = u_1^{e_1}u_2$ and $v = u_1^{e_1}u_2u_1^{e_2}$ where u_2 is a unique, possibly empty, proper prefix of u_1 and e_1, e_2 are unique integers such that $e_1 \geq e_2 \geq 1$. Moreover,*

(a) *if $|u_2| = 0$, then $e_1 > e_2$;*
(b) *$|u_2| > 0$ if and only if v is primitive;*
(c) *if u is primitive, then $|u_2| > 0$;*
(d) *if v^2 is a prefix of a string x and there is no other occurrence of u^2 in x, then $|u_2| > 0$.*

Given a balanced double square (u, v), the unique 4-tuple (u_1, u_2, e_1, e_2) yielded by Lemma 5 is referred to as the *canonical factorization* of the double square (u, v) and is denoted by $(u, v\ :\ u_1, u_2, e_1, e_2)$. See the following illustration of Lemma 5 and [1] for a proof.



Lemmas 6 and 7 are considered folklore and are both consequences of the Periodicity Lemma [6], and thus presented here without a proof. However, the interested reader may find their proofs in [1] and a more detailed treatment in [3]. Lemmas 6 and 7 are among the key tools to deal with the canonical factorizations of balanced double squares.

Lemma 6 (Synchronization Principle)**.** *The primitive string x occurs exactly p times in $x_2x^px_1$ where p is a non-negative integer and x_1 is a proper prefix of x and x_2 a proper suffix of x.*

Thus, a primitive string is its only conjugate and is not equal to any of its non-trivial rotations. In addition, any rotation or right or left cyclic shift of a primitive string is also primitive.

Lemma 7 (Common Factor Lemma)**.** *Consider strings x and y where x_1 is a proper prefix of x, x_2 a proper suffix of x, y_1 is a proper prefix of y, and y_2 a proper suffix of y. If for non-negative integers p and q, $x_2x^px_1$ and $y_2y^qy_1$ have a common factor of length $|x|+|y|$, then the primitive root of x and the primitive root of y are conjugates.*

The notion of inversion factor was introduced in [4]: let $(u, v\ :\ u_1, u_2, e_1, e_2)$ be a canonical factorization of a balanced double square (u, v) and let $\overline{u}_2$ denote the suffix of u_1 such that $u_1 = u_2\overline{u}_2$. The *inversion factor* is defined as $\overline{u}_2u_2u_2\overline{u}_2$. As shown in [4], the inversion factor has only two occurrences in v^2 as indicated in bold below:

$$v^2 = (u_2\overline{u}_2)^{e_1}u_2(u_2\overline{u}_2)^{e_1+e_2}u_2(u_2\overline{u}_2)^{e_2} =$$

$$(u_2\overline{u}_2)^{e_1-1}u_2\boldsymbol{\overline{u}_2u_2u_2\overline{u}_2}(u_2\overline{u}_2)^{e_1+e_2-2}u_2\boldsymbol{\overline{u}_2u_2u_2\overline{u}_2}(u_2\overline{u}_2)^{e_2-1}$$

Moreover, as shown in [4], for a balanced double square $(u, v\ :\ u_1, u_2, e_1, e_2)$

$$0 \le lcs(u_2\overline{u}_2, \overline{u}_2u_2) + lcp(u_2\overline{u}_2, \overline{u}_2u_2) \le |u_1|-2.$$

3 Proof of Lemma 3

Let $u \neq v$, and u^2 and v^2 be both proper prefixes of w^2. Lemma 3 states that $\{u = {v_1}^t, v = {v_1}^{p_1}v_2, w = {v_1}^{p_1}v_2{v_1}^{p_2}\ \}$ or $\{\ |u|+|v| \le |w|\ \}$. (S)

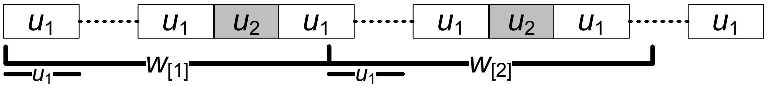
Without loss of generality, we can assume that $|u| < |v|$.

If $2|v| \le |w|$, then $|v|+|u| < |w|$ as $|u| < |v|$, and thus (S) holds. Therefore, we can assume that $|w| < 2|v|$; that is, (v, w) is a balanced double square and thus admits a canonical factorization $(v, w\ :\ v_1, v_2, p_1, p_2)$ by Lemma 5. We consider the following cases.

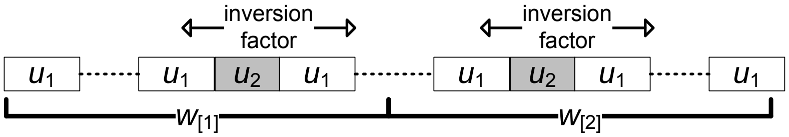
1. Case when u and v are not proportional, i.e. $2|u| \le |v|$.
   If $|u| < |v_1|$, then $|u| + |v| < |v_1|^{p_1+1} + |v_2| \le |v_1|^{p_1+p_2} + |v_2| = |w|$.
   If $|u| \ge |v_1|$, since u^2 is a prefix of $v = {v_1}^{p_1}v_2$, then u^2 and ${v_1}^{p_1}v_2$ have a common factor of length $|u|+|v_1|$, and by Lemma 7, u and v_1 have the same primitive root, and so v_1 is the primitive root of v_1. Thus $u = {v_1}^t$ for some $t \ge 1$, $v = {v_1}^{p_1}v_2$, and $w = {v_1}^{p_1}v_2{v_1}^{p_2}$.
   If $t \le p_2$, then $|u| + |v| \le |w|$, but if $t > p_2$,
2. Case when u and v are proportional, i.e. $|v| < 2|u|$. Then (u, v) is a balanced double square and thus admits by Lemma 5 a canonical factorization $(u, v\ :\ u_1, u_2, e_1, e_2)$.

   (i) Case when $|u_2| = 0$. Then $e_1 > e_2$, $u = {u_1}^{e_1}$, and $v = {u_1}^{e_1+e_2}$. Let us assume that $|w| < |u|+|v| = (2e_1+e_2)|u_1|$. Then w^2 and ${u_1}^{2e_1+2e_2}$ have a common factor of length $|w|+|u_1|$, and by Lemma 7 the primitive root of w is a conjugate of u_1, i.e. equals u_1. Thus, u, v, and w all have the same primitive root, and thus (S) holds.

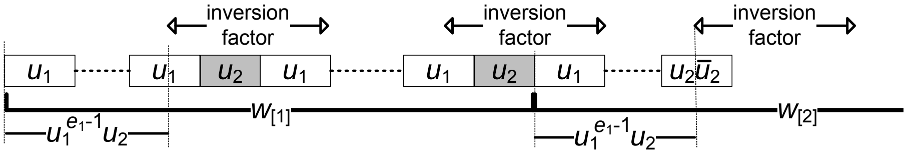
(ii) Case when $|u_2| > 0$. Let $w_{[1]}$ refer to the first occurrence of w and $w_{[2]}$ to the second. First, we have to show that $w_{[1]}$ does not end in the first u_1 of ${u_1}^{e_1+e_2}$. If it did, then it would contradict Lemma 6 as $w_{[1]}$ and hence $w_{[2]}$ has the primitive u_1 as a prefix as indicated by the following diagram:



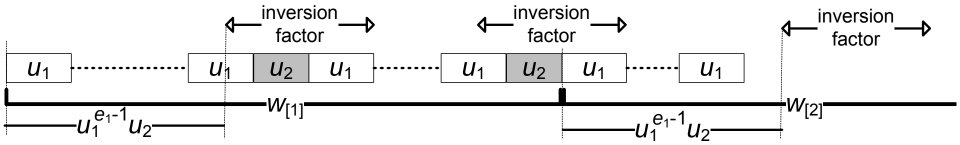
Thus, $w_{[1]}$ must end somewhere past the first u_1 of ${u_1}^{e_1+e_2}$:



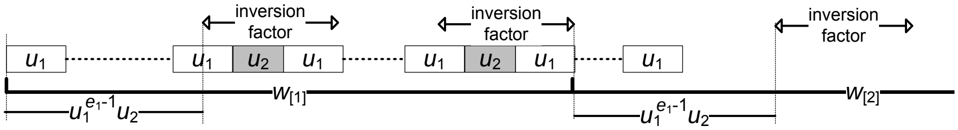
As a consequence, $w_{[1]}$ contains the first inversion factor of ${u_1}^{e_1}u_2{u_1}^{e_1+e_2}u_2{u_1}^{e_2}$ exactly at a distance of $|{u_1}^{e_1-1}u_2|$ from the beginning. It follows that $w_{[2]}$ must contain an occurrence of the inversion factor at exactly the same distance from the beginning. If it were the second inversion factor, then the length of w would be exactly $|v|$ as it is the distance between the two occurrences of the inversion factor in w^2, a contradiction. Thus, it must be an occurrence of the inversion factor past the second one. The first possible start of another occurrence of the inversion factor is the suffix $\overline{u}_2$ of ${u_1}^{e_1}u_2{u_1}^{e_1+e_2}u_2{u_1}^{e_2}$. If $e_1 = e_2$, then it is the case that $w_{[1]} = w_{[2]} = {u_1}^{e_1-1}(\boldsymbol{\overline{u}_2 u_2 u_2 \overline{u}_2}){u_2}^{e_1+e_2-1}u_2$ (see below) and then



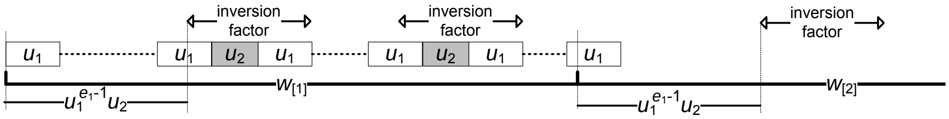
$|w| = |u|+|v|$ as $|u| = |{u_1}^{e_1}u_2|$ and $|v| = |{u_1}^{e_1+e_2}u_2|$, thus (S) holds. If $e_1 > e_2$, then the occurrence of the inversion factor in $w_{[2]}$ must be at a distance ${u_1}^{e_1-1}u_2$ from the beginning. By Lemma 6, the prefix ${u_1}^{e_1-1}u_2$ of $w_{[2]}$ must align with ${u_1}^{e_2}$ or start in the last u_1 of ${u_1}^{e_2}$, and so $w_{[1]}$ must have ${u_1}^{e_1}u_2{u_1}^{e_1+e_2}u_2$ as a prefix, again yielding $|w| \geq |u|+|v|$ (see below), thus (S) holds.



or



or



4 Conclusion

We showed that the conclusion of the Crochemore and Rytter's lemma on three squares starting at the same position also holds under alternative conditions. The proof is based on a novel insight into the combinatorics of double squares.

Acknowledgments. The authors would like to thank Adrien Thierry for pointing out an inconsistency in Lemma 3. This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada Discovery Grant program (Franek: RGPIN25112-2012, Deza: RGPIN-2015-06163), by the Digiteo Chair C&O program (Deza), and by the Canada Research Chairs program (Deza: CRC 950-213642).

References

[1] H. Bai, F. Franek, and W. F. Smyth. Two squares canonical factorization. In Jan Holub and Jan Žďárek, editors, *Proceedings of the Prague Stringology Conference 2014*, pages 52–58, Czech Technical University in Prague, Czech Republic, 2014.

[2] M. Crochemore, C. Hancart, and T. Lecroq. *Algorithms on strings.* Cambridge University Press, 2007.

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

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

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

[6] N.J. Fine and H.S. Wilf. Uniqueness theorems for periodic functions. *Proc. Amer. Math. Soc.*, 16:109–114, 1965.

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

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

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

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

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

[12] J. Stoye and D. Gusfield. Simple and flexible detection of contiguous repeats using a suffix tree. *Theoretical Computer Science*, 270:843–856, 2013.