Combinatorics of the Interrupted Period

Adrien Thierry

Advanced Optimization Laboratory
McMaster University, Hamilton, Ontario, Canada
adrien.thierry@gmail.com

Abstract. This article is about discrete periodicities and their combinatorial structures. It presents and describes the unique structure caused by the alteration of a pattern in a repetition. Those alterations of a pattern arise in the context of double squares and were discovered while working on bounding the number of distinct squares in a string. Nevertheless, they can arise in other phenomena and are worth being presented on their own.

Keywords: string, period, primitive string, factorization

If x is a primitive word, and x_1 a prefix of x, the sequence $x^n x_1 x^m$ has a singularity: it has a periodic part of period x, an interruption, and a resumption of the pattern x. That interruption creates a different pattern, one that does not appear in x^n. The goal of this article is to unveil that pattern.

1 Preliminaries

In this section, we introduce the notations and present a simple property and two of its corollaries. These observations are not complicated, but their proofs introduce the technique used in the proof of the main theorem, Theorem 7, and allow for a clear understanding of the phenomenon described there.

We first fix some notations. An *alphabet* A is a finite set. We call *letters* the elements of A. If $|A| = 2$, the words are referred to as binary and are used in computers. Another well known example for $|A| = 4$ is DNA.
A vector of A^n is a *word* w of length $|w| = n$, which can also be presented under the form of an array $w[1 \dots n]$. Two words are *homographic* if they are equal to each other. If $x = x_1 x_2 x_3$ for non-empty words x_1, x_2 and x_3, then x_1 is a *prefix* of x, x_2 is a *factor* of x, and x_3 is a *suffix* of x (if both the prefix and the suffix are non empty, we refer to them as proper). We define *multiplication* as concatenation. In english, *breakfast* $=$ *break* $\cdot$ *fast*. In a traditional fashion, we define the n^{th} *power* of a word w as n time the multiplication of w with itself. A word x is *primitive* if x cannot be expressed as a non-trivial power of another word x'.
A word $\tilde{x}$ is a *conjugate* of x if $x = x_1 x_2$ and $\tilde{x} = x_2 x_1$ for non-empty words x_1 and x_2. The set of conjugates of x together with x form the conjugacy class of x which is denoted $Cl(x)$.
A factor $x, |x| = n$ of w has *period* p if $x[i] = x[i + |p|], \forall i \in [1, \dots, n - |p|]$.
The *number of occurrences* of a letter c in a word w is denoted $n_c(w)$, the *longest common prefix* of x and y as $lcp(x, y)$, while $lcs(x, y)$ denotes the *longest common suffix* of x and y (note that $\mathrm{lcs}(x, y)$ and $\mathrm{lcp}(x, y)$ are words).

The properties presented next rely on a simple counting argument. If the proofs are not interesting in themselves, they still allow for meaningful results.

Adrien Thierry: Combinatorics of the Interrupted Period, pp. 17–21.
Proceedings of PSC 2015, Jan Holub and Jan Žďárek (Eds.), ISBN 978-80-01-05787-2

Proposition 1 *A word w and all of its conjugates have the same number of occurrences for all of their letters, i.e. $\forall \tilde{w} \in Cl(w), \forall a \in A,\ n_a(w) = n_a(\tilde{w})$.*

Proof. Note that $\forall \tilde{w} \in Cl(w), \exists w_1, w_2$, such that $w = w_1w_2, \tilde{w} = w_2w_1$. Then, $\forall a \in A, n_a(w) = n_a(w_1) + n_a(w_2) = n_a(\tilde{w})$. □

The negation of Property 1 gives the following corollary:

Corollary 1. *If two words do not have the same number of occurrence for the same letter, they are not conjugates.*

Another important corollary of Property 1 is the following:

Corollary 2. *Let x be a word, $|x| \geq n+1$. If $u = x[1 \ldots n]$ and $v = x[2 \ldots n+1]$ are conjugates of each other, then $x[1] = x[n+1]$, i.e. v is a cyclic shift of u.*

Proof. Note that u and v have the factor $x[2 \ldots n]$ in common. Since u and v are conjugates, they have the same number of occurrences for all of their letters (Proposition 1). It follows that $n_{x[1]}(u) = n_{x[1]}(x[1 \ldots n]) = n_{x[1]}(x[2 \ldots n]) + 1 = n_{x[1]}(v) = n_{x[1]}(x[2 \ldots n]) + n_{x[1]}(x[n+1])$, hence $n_{x[1]}(x[n+1]) = 1$, i.e. $x[1] = x[n+1]$. □

2 Theorem

Discrete periods were described by N.J. Fine and H.S. Wilf in 1965 in the article "Uniqueness theorem for periodic functions" [1]. A corollary of that theorem, the synchronization principle, was proved by W. Smyth in [2] and L. Ilie in [3]:

Theorem 3. *If w is primitive, then, for all conjugates $\tilde{w}$ of $w, w \neq \tilde{w}$.*

Which is about the synchronization of patterns. The next theorem is about the impossible synchronization when a pattern is interrupted.

First, we need to formalize what we call an interruption of the pattern. Let x be a primitive word and x_1 be a proper prefix of x, i.e. $x_1 \neq x$. Write $x = x_1x_2$ for some suffix x_2 of x.

Let $W = x^{e_1}x_1x^{e_2}$ with $e_1 \geq 1, e_2 \geq 1, e_1 + e_2 \geq 3$.

We see that W has a repetition of a pattern x as a prefix: $x^{e_1}x_1$, and then the repetition is interrupted at position $|x^{e_1}x_1|$, before starting again in the suffix x^{e_2}. We need one more definition (albeit that definition is not necessary, it is presented here for better understanding) before introducing the two factors that we claim have very restricted occurrences in W.

Definition 4. *Let $\tilde{p}$ be the prefix of length $|lcp(x_1x_2, x_2x_1)| + 1$ of x_1x_2 and $\tilde{s}$ the suffix of length $|lcs(x_1x_2, x_2x_1)| + 1$ of x_2x_1. The factor $\tilde{s}\tilde{p}$ starting at position $|x^{e_1}| + |x_1| - |lcs(x_1x_2, x_2x_1)| - 1$ is the* core of the interrupt *of W.*

If W and its interrupt are clear from the context, we will just speak of the core (of the interrupt).

Example 5. Consider $x = aaabaaaaaabaaaa$ and $x_1 = aaabaaaaaabaaa$, then xx_1x^2 has $xx_1x = aaabaaaaaabaaaa\mathbf{aaabaaaaaabaaa}aaabaaaaaabaaaa$ as a prefix and $x_2 = a$. It follows that $\mathrm{lcp}(x_1x_2, x_2x_1) = aaa$, and $\tilde{p} = aaab, \mathrm{lcs}(x_1x_2, x_2x_1) = aaa$, and $\tilde{s} = baaa$. The core of the interrupt, $\tilde{s}\tilde{p}$, is the underlined in:

$$xx_1x = aaabaaaaaabaaaaaaabaaaaaa\underbrace{baaaaaab}_{\tilde{s}\tilde{p}}aaaaaabaaaa.$$

The factors that were previously known to have very restricted occurrences in W, to the best of the author's knowledge, were the inversion factors defined by A. Deza, F. Franek and A. Thierry in [4]:

Definition 6. *Let $W = x^{e_1}x_1x^{e_2}$ with $x = x_1x_2$ a primitive word and$e_1 \geq 1, e_2 \geq 1, e_1 + e_2 \geq 3$. An* inversion factor *of W is a factor that starts at position i and for which:*

- $W[i + j] = W[i + j + |x| + |x_1|]$ *for* $0 \leq j < |x_1|$, *and*
- $W[i + j] = W[i + j + |x_1|]$ *for* $|x_1| \leq j \leq |x| + |x_1|$.

Those inversion factors, which have the structure of $x_2x_1x_1x_2 = \tilde{x}x$, and which length are twice the length of x, were used as two notches that forces a certain synchronization of certain squares in the problem of the maximal number of squares in a word, and allowed to offer a new bound to that problem. The main anticipated application of the next result is an improvement of that bound, though the technique has already proved useful in the improvement of M. Crochemore and W. Rytter's three squares lemma, [5], by H. Bay, A. Deza and F. Franek, [6], and in the proof of the New Periodicity Lemma by H. Bay, F. Franek and W. Smyth [7].

Now, let w_1 be the factor of length $|x|$ of W that has the core of the interrupt of W as a suffix, and let w_2 be the factor of length $|x|$ that has the core of the interrupt of W as a prefix. We will show that both w_1 and w_2 have very restricted occurrences in W.

Theorem 7. *Let x be a primitive word, x_1 a proper prefix of x and $W = x^{e_1}x_1x^{e_2}$ with $e_1 \geq 1, e_2 \geq 1, e_1 + e_2 \geq 3$. Let w_1 be the factor of length $|x|$ of W ending with the core of the interrupt of W, and let w_2 be the factor of length $|x|$ starting with the core of the interrupt of W. The words w_1 and w_2 are not in the conjugacy class of x.*

Proof. Define $p = \mathrm{lcp}(x_1x_2, x_2x_1)$ and $s = \mathrm{lcs}(x_1x_2, x_2x_1)$ (note that p and s can be empty).
Deza, Franek, and Thierry showed that $|\mathrm{lcs}(x_1x_2, x_2x_1)| + |\mathrm{lcp}(x_1x_2, x_2x_1)| \leq |x_1x_2| - 2$ when x_1x_2 is primitive (see [4]). Note that in the case $|\mathrm{lcs}(x_1x_2, x_2x_1)| + |\mathrm{lcp}(x_1x_2, x_2x_1)| = |x| - 2$, w_1 w_2 are the same factor.
Write $x = pr_prr_ss$ and $\tilde{x} = pr'_pr'r'_ss$ for the letters r_p, r'_p, r_s, r'_s, $r_p \neq r'_p, r_s \neq r'_s$ (by maximality of the longest common prefix and suffix) and the possibly empty and possibly homographic words r and r'.
We have, by construction, $w_1 = r'r'_sspr_p$ and $w_2 = r'_sspr_pr$.
Note that $n_{r_p}(w_1) = n_{r_p}(\tilde{x}) + 1$ and that $n_{r'_p}(\tilde{x}) = n_{r'_p}(w_1) + 1$ and, by Corollary 1,

w_1 is not a conjugate of $\tilde{x}$, nor of x. And because $|w_1| = |x|, w_1$ is neither a factor of $x^{e_1}x_1$ nor of x^{e_2}.
Similarly for w_2, $n_{r'_s}(w_2) = n_{r'_s}(x) + 1$ and $n_{r_s}(x) = n_{r_s}(w_2) + 1$ and, by corollary 1, w_2 is not a conjugate of x, and because $|w_2| = |x|, w_2$ is neither a factor of $x^{e_1}x_1$ nor of x^{e_2}. □

Example 8. Consider again $x = aaabaaaaaabaaaa$, $x_1 = aaabaaaaaabaaa$ and $x_2 = a$. We have $|x| = 15$, and:

$$xx_1x = aaabaaaaaabaaaaaaa\overbrace{baaaaaa\,\mathrlap{\underbrace{\phantom{\mathbf{baaaaaab}}aaaaaab}_{w_2}}\mathbf{baaaaaab}}^{w_1}aaaaaab\,aaaa$$

The core of the interrupt is presented in bold.
The two factors w_1 and $w_2 = w_1 = baaaaaabaaaaaab$ (note that w_2 needs not be equal to w_1), starting at different positions, are not factors of x^2. Yet, the factor $aaaaaabaaaaaabaaaaaa$ of length $|x| + |\mathrm{lcs}(x, \tilde{x})| + |\mathrm{lcp}(x, \tilde{x})|$ and which contains the core of the interrupt is a factor of x^2. The same goes for the factors of length $|x|-1$ that starts and ends with the core of the interrupt, $aaaaaabaaaaaab$ and $baaaaaabaaaaaa$: they both are factors of x^2. For those reasons, the theorem can be regarded as tight

3 Conclusion

The core of the interrupt was discovered while studying double squares. An important result in the study of that problem is M. Crochemore and W. Rytter's three squares lemma, [5], of which L. Ilie offers a shorter proof in [3]. We offer here a very short proof of that result which relies on the core of the interrupt.

Lemma 9. *In a word, no more that two squares can have their last occurrence starting at the same position.*

Proof. Suppose that three squares $u_1^2, u_2^2, u_3^2, |u_1| < |u_2| < |u_3|$ start at the same position. Because u_2^2 and u_3^2 start at the same position, we can write $u_2 = x_0^{e_1}x_1$, $u_3 = x_0^{e_1}x_1x_0^{e_2}$ for $x_0 = x_1x_2$ a primitive word, x_1 a proper prefix of x_0 and $e_1 \geq e_2 \geq 1$, hence u_3 contains a core of the interrupt. Now, by synchronization principle, Theorem 3, $u_1, |u_1| < |u_2|$, cannot end in the suffix $\mathrm{lcs}(x_1x_2, x_2x_1)$ of u_2 (since u_1 has x_0 as a prefix) and ends before the core of the interrupt of u_3, but if $|u_1^2| \geq |u_3|$, the second occurrence of u_1 contains the core of the interrupt and a word of length $|x_0|$ that starts with it, while the first occurrence doesn't: which, by Theorem 7, is a contradiction.

Thanks to my supervisors Antoine Deza and Franya Franek for helpful discussions and advices and to Alice Heliou for proof reading of a preliminary version of this article.

References

1. N. J. Fine and H. S. Wilf: *Uniqueness theorems for periodic functions*, in Proceedings of the American Mathematical Society, vol. 16, no. 1, 1965, pp. 109–114.
2. B. Smyth: *Computing Patterns in Strings*. ACM Press Bks, Pearson/Addison-Wesley, 2003.

3. L. Ilie: *A simple proof that a word of length n has at most $2n$ distinct squares.* Journal of Combinatorial Theory, Series A, vol. 112, no. 1, 2005, pp. 163–164.
4. A. Deza, F. Franek, and A. Thierry: *How many double squares can a string contain?* Discrete Applied Mathematics, vol. 180, 2015, pp. 52–69.
5. M. Crochemore and W. Rytter: *Squares, cubes, and time-space efficient string searching.* Algorithmica, vol. 13, no. 5, 1995, pp. 405–425.
6. H. Bay, A. Deza, and F. Franek: *On a Lemma of Crochemore and Rytter*, to appear in Journal of Discrete Algorithms.
7. H. Bay, F. Franek, and W. Smyth: *The New Periodicity Lemma Revisited*, to appear in Journal of Discrete Applied Mathematics.