Combinatorics of the interrupted period

A. Thierry

Prague Stringology Conference

24 août 2015
An alphabet $A$ is a finite set. We call letters the elements of $A$. A vector of $A^n$ is a word $w$ of length $|w| = n$, which can also be presented under the form of an array $w[1, ..., n]$. A factor $x, |x| = n$ of $w$ has period $p \leq 2n$ if $x[i] = x[i + |p|], \forall i \in [1, ... n - |p|]$.

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 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 $x = x_1x_2$ and $\tilde{x} = x_2x_1$ 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 $Cl(x)$. 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 $y$. 
The core of the interrupt was discovered while studying the maximal number of distinct squares in a string. M. Crochemore and W. Rytter proved in [1] that no more than two squares can have their last occurrence starting at the same position.
The core of the interrupt was discovered while studying the maximal number of distinct squares in a string. M. Crochemore and W. Rytter proved in [1] that no more than two squares can have their last occurrence starting at the same position. If two squares $uu$ and $UU$ have their last occurrences starting at the same position, the double square $\mathcal{U}$, the set of the two squares $uu$ and $UU$, has a canonical factorization.
The canonical factorization of a double square is \( u_0^{e_1} u_1 u_0^{e_2} u_0^{e_1} u_1 u_0^{e_2} \)
for a primitive word \( u_0 \) and \( u_1 \) a proper prefix of \( u_0 \).
What we call an interrupted periodicity is a factor \( u_0^{e_1} u_1 u_0^{e_2} \) for a primitive word \( u_0 \) and a proper prefix \( u_1 \) of \( u_0 = u_1 u_2 \).
Albeit it was defined for studying double squares, interrupted periodicities can occur in other contexts.
If you rewrite the canonical factorization of a double square

\[ UU = u_0^{e_1} u_1 u_0^{e_2} u_0^{e_1} u_1 u_0^{e_2} \]
If you rewrite the canonical factorization of a double square

\[UU = u_0^{e_1} u_1 u_0^{e_2} u_0^{e_1} u_1 u_0^{e_2}\]

as

\[UU = u_0^{e_1-1} u_0 u_1 u_0 u_0^{e_1+e_2-2} u_0 u_1 u_0 u_0^{e_2-1}.\]
If you rewrite the canonical factorization of a double square

$$UU = u_0^{e_1} u_1 u_0^{e_2} u_0^{e_1} u_1 u_0^{e_2}$$

as

$$UU = u_0^{e_1-1} u_0 u_1 u_0 u_0^{e_1+e_2-2} u_0 u_1 u_0 u_0^{e_2-1}.$$

Since $u_1$ is a proper prefix of $u_0$, $u_0 = u_1 u_2$ for a proper suffix $u_2$ of $u_0$:

$$UU = u_0^{e_1-1} u_1 u_2 u_1 u_2 u_0^{e_1+e_2-2} u_1 u_2 u_1 u_2 u_0^{e_2-1}.$$
We can see two factors $u_2 u_1 u_1 u_2$ appear:

\[
\begin{align*}
&u_0^{e_1 - 1} u_1 u_2 u_1 u_1 u_2 u_0^{e_1 + e_2 - 2} u_1 u_2 u_1 u_1 u_2 u_0^{e_2 - 1} \\
&\text{here} \quad \text{and here.}
\end{align*}
\]

Note that everywhere else, we have a succession of $u_1 u_2$. 
One of Fine and Wilf’s famous periodicity lemma’s [3] corollary mentioned by Fraenkel and Simpson [4] tells us that no conjugates of $u_0$ are equal to $u_0$. Hence, $u_1 u_2$ only appears twice in $u_0^2$. 
One of Fine and Wilf’s famous periodicity lemma’s [3] corollary mentioned by Fraenkel and Simpson [4] tells us that no conjugates of $u_0$ are equal to $u_0$.

Hence, $u_1 u_2$ only appears twice in $u_0^2$.

The problem that we ask is what makes the factors $u_2 u_1 u_1 u_2$ “unique”.

Indeed, the factors $u_2 u_1 u_1 u_2$ in

\[ u_0^{e_1-1} u_1 u_2 u_1 u_1 u_2 u_0^{e_1+e_2-2} u_1 u_2 u_1 u_1 u_2 u_0^{e_2-1} \]

serve as notches which were used for alignment of double squares by Deza, Franek, T. in [2].
We try to understand what makes the factor $u_2u_1u_1u_2$ unique, and focus our attention on the word

$$w = u_0^{e_1} u_1 u_0^{e_2}.$$ 

Note that we are not studying double squares anymore but interrupted periodicities.
Deza, Franek, T., [2], showed, for a primitive $x$ and a conjugate $\tilde{x}$, that $|\text{lcp}(x, \tilde{x})| + |\text{lcs}(x, \tilde{x})| \leq |x| - 2$. 
Deza, Franek, T., [2], showed, for a primitive \( x \) and a conjugate \( \tilde{x} \), that

\[
|\text{lcp}(x, \tilde{x})| + |\text{lcs}(x, \tilde{x})| \leq |x| - 2.
\]

Set \( p = \text{lcp}(u_1 u_2, u_2 u_1) \), \( s = \text{lcs}(u_1 u_2, u_2 u_1) \).

We can write:

\[
\begin{align*}
    u_1 u_2 &= pr_p r r_s s \\
    u_2 u_1 &= pr'_p r'_r r'_s s
\end{align*}
\]

for the letters \( r_p \neq r'_p \), \( r_s \neq r'_s \) and the possibly empty and possibly homographic words \( r \) and \( r' \)
Write

\[ w = u_0^{e_1} u_1 u_0^{e_2}. \]

\[ w = u_0^{e_1-1} u_1 u_2 u_1 u_2 u_0^{e_2-1}. \]
Write

\[ w = u_1^{e_1} u_0^{e_2}. \]

\[ w = u_0^{e_1-1} u_1 u_2 u_1 u_2 u_0^{e_2-1}. \]

\[ w = u_0^{e_1-1} u_1 p r_p' r_s' s p r_p r r_s s u_0^{e_2-1}. \]
We see that the prefix of \( w \) ending at position \( u_{0}^{e_{1}-1}u_{1}p'r'r'sp \) has period \( |u_{0}| \). The same goes for the suffix that starts at position \( u_{0}^{e_{1}-1}u_{1}p'r'r's \).

\[ w = u_{0}^{e_{1}-1}u_{1}p'r'r'sp_{p}rrrsu_{0}^{e_{2}-1}. \]
We see that the prefix of $w$ ending at position $|u_0^{e_1-1}u_1pr'r'r'sp|$ has period $|u_0|$. The same goes for the suffix that starts at position $|u_0^{e_1-1}u_1pr'r'r'|$.

$$W = u_0^{e_1-1}u_1pr'r'r'spr'r'su_0^{e_2-1}. $$

We still haven’t defined what makes the factor $u_2u_1u_1u_2$ unique, but we can see that the factor $sp$ in bold must play an important role.
We define the core of the interrupt as the factor $r'_s spr_p$ of $w$.

$$w = u_0^{e_1-1} u_1 p r'_p r'_s spr_p r r_s s u_0^{e_2-1}.$$

*here*
We define the core of the interrupt as the factor $r'_s spr_p$ of $w$. 

$$w = u_0^{e_1-1} u_1 pr'_p r' \underbrace{r'_s spr_p}_{here} rr_s su_0^{e_2-1}.$$ 

The core of the interrupt is a peculiar factor, but as shown in the next slide, it doesn’t explain the uniqueness of $u_2 u_1 u_1 u_2$ in $w$. 
Consider $u_0 = aaabaaaaaabaaaa$, $u_1 = aaabaaaaaabaaa$ and $u_2 = a$. We have $|u_0| = 15$, and:

$$x.x_1.x = aaabaaaaaabaaaa.aaab \underbrace{aaaaa}_{w'}baaa.aaabaaaaaa ba$$

The core of the interrupt is presented in bold.
Consider $u_0 = aaabaaaaabaaaaa$, $u_1 = aaabaaaaabaaa$ and $u_2 = a$. We have $|u_0| = 15$, and :

$$x.x_1.x = aaabaaaaabaaaaa.aaab\underbrace{aaaaabaaa.aaabaaaaaa}_w baaa$$

The core of the interrupt is presented in bold. The factor $w' = aaaaaaabaaaaabaaaaaa$ of length $|u_0| + |\text{lcs}(u_1u_2, u_2u_1)| + |\text{lcp}(u_1u_2, u_2u_1)| - 1$ and which contains the core of the interrupt is a factor of $u_0^2$. 
The factors of length $|u_0|$ that starts and ends with the core of the interrupt are not factors of $u_0^2$. 
Write:

\[ w = u_0^{e_1-1} ru_1pr_p' r'_s spr_p rr_s su_0^{e_2-1}. \]
Write:

\[ w = u_0^{e_1-1} u_1 p r'_p r'_s s p r p r s u_0^{e_2-1}. \]

Let \( w_1 \) be the factor of length \( |u_0| \) that ends with the core of the interrupt, and \( w_2 \) be the factor of length \( |u_0| \) that starts with the interrupt.

\[ w = u_0^{e_1-1} u_1 p r'_p r'_s s p r p r s u_0^{e_2-1}. \]
Write:

\[ w = u_0^{e_1-1} u_1 pr' r' r'_s spr_p r s u_0^{e_2-1}. \]

Let \( w_1 \) be the factor of length \( |u_0| \) that ends with the core of the interrupt, and \( w_2 \) be the factor of length \( |u_0| \) that starts with the interrupt.

\[ w = u_0^{e_1-1} u_1 pr' r' r'_s spr_p r s u_0^{e_2-1}. \]

Hence \( w_1 = r' r'_s spr_p \) and \( w_2 = r'_s spr_p r \).
We have $w_1 = r'r'spr_p$, while $u_2u_1 = pr_p'r's$ hence $n_{rp}(w_1) \neq n_{rp}(u_2u_1)$ and $w_1$ is not a conjugate of $u_0$, hence doesn’t appear in $u_0^2$. 
We have \( w_1 = r'r'_s spr_p \), while \( u_2 u_1 = pr'_p r'r'_s s \) hence \( n_{r_p}(w_1) \neq n_{r_p}(u_2 u_1) \) and \( w_1 \) is not a conjugate of \( u_0 \), hence doesn’t appear in \( u_0^2 \).

Similarly, \( w_2 = r'_s spr_p r \), while \( u_1 u_2 = pr_p rr_s s \), \( n_{r_s}(w_2) \neq n_{r_s}(u_1 u_2) \) and \( w_2 \) is not a conjugate of \( u_0 \).
If we look at the previous example, where $u_0 = aaabaaaaaabaaaaa$,

$$w = aaabaaaaaaaaabaaaaaaa.aaabaaaaaabaaaaabaaabaabaaaaaaabaaabaaaaaaabaaaaabaaaaaaabaaabaaaaaaabaaabaaaaaaabaa,$$

the factors of length $|u_0| - 1$ that starts and ends with the inversion factor, $aaaaaabaaaaaab$ and $baaaaaabaaaaaa$, are both factors of $u_0^2$. In that regard, the result can be considered as tight.
Recall that Fine and Wilf’s periodicity lemma tells us that no conjugates of $u_0$ are equal to $u_0$. We showed that interrupting the periodicity gives rise to two factors that are not equal to any conjugates of $u_0$. 
Thank You!


Aviezri S Fraenkel and Jamie Simpson.  
How many squares can a string contain?  