Lazy Ostrowski Numeration and Sturmian Words

Jeffrey Shallit
School of Computer Science, University of Waterloo
Waterloo, Ontario N2L 3G1, Canada
shallit@uwaterloo.ca
https://cs.uwaterloo.ca/~shallit
My co-authors

Daniel Gabric

Narad Rampersad
An integer $p$, with $1 \leq p \leq |x|$, is called a **period** of a finite word $x$ if $x[i] = x[i + p]$ for $1 \leq i \leq |x| - p$.

Example: alfalfa has period 3.

A period $p$ of $x$ is **nontrivial** if $p < |x|$.

The least period of a word $x$ is called the **period**, and is written \( \text{per}(x) \).

The number of nontrivial periods of a word $x$ is denoted \( \text{nnp}(x) \). For example, \( \text{nnp}(adoradora) = 2 \).
The **exponent** of a finite nonempty word $x$ is defined to be

$$\exp(x) := \frac{|x|}{\per(x)}.$$  

For example, $\exp(\text{entente}) = 7/3$.

The **critical exponent** $ce(x)$ of a finite or infinite word $x$ is defined to be

$$ce(x) := \sup\{\exp(p) : p \text{ is a nonempty factor of } x\}.$$
The original motivation for this research was to answer the following question:

*When does a word have lots of periods?*

Obviously, one way a word can have lots of periods is if it is periodic: $0^n$ has $n$ periods. So a word with high exponent will have lots of periods.

On the other hand, $0^n 1^{n^2} 0^n$ has lots of periods, but very small exponent $(n^2 + 2n)/(n^2 + n) \approx 1 + 1/n$. So exponent alone can’t be the whole story. Maybe critical exponent?

No! A word like $01^n 0$ has only one period, but has high critical exponent.

So what should we do?
Instead we’ll consider the initial critical exponent.

The *initial critical exponent* $\text{ice}(x)$ of a finite or infinite word $x$ is defined to be

$$
\text{ice}(x) := \sup\{\exp(p) : p \text{ is a nonempty prefix of } x\}.
$$

For example, $\text{ice}(\text{phosphorus}) = 7/4$.

This concept was (essentially) introduced by Berthé, Holton, and Zamboni in 2006.
A word $w$ is a border of a word $x$ if $w$ is both a prefix and suffix of $x$.

For example, ionization has the border ion.

Borders are allowed to overlap, but we generally rule out borders $w$ where $w = \epsilon$ or $w = x$.

A border $w$ of $x$ is short if $|w| < |x|/2$.

**Basic observation:** A word has a nontrivial period $t$ iff it has a border of length $n - t$.

Example: abracadabra has nontrivial periods 7 and 10, and borders of length 4 and 1.
Now, back to counting periods. Here is our main result #1, relating periods to ice:

**Theorem.** Let $x$ be a bordered word of length $n \geq 1$. Let $e = \text{ice}(x)$. Then

$$\text{nnp}(x) \leq \frac{e}{2} + 1 + \frac{\ln(n/2)}{\ln(e/(e - 1))}.$$

**Proof.**
Break the bound up into two pieces, by considering the periods of size $\leq n/2$ and $> n/2$. Call these the *short* and *long* periods.
Proof of the period inequality

Let $p = \text{per}(x)$, the shortest period of $x$.

If $p$ is short, then $x$ has short periods $p, 2p, 3p, \ldots, \lfloor n/(2p) \rfloor p$.

Clearly $\text{ice}(x) \geq n/p$, so we get at most $e/2$ short periods from this list.

To see that there are no other short periods, let $q$ be some short period not on this list. Then $p < q \leq n/2$ by assumption.

By the Fine-Wilf theorem, if a word of length $n$ has two periods $p, q$ with $n \geq p + q - \gcd(p, q)$, then it also has period $\gcd(p, q)$.

Since $\gcd(p, q) \leq p$, either $\gcd(p, q) < p$, which is a contradiction, or $\gcd(p, q) = p$, which means $q$ is a multiple of $p$, another contradiction.
Next, let’s consider the long periods or, alternatively, the short borders (those of length $< \frac{n}{2}$).

Suppose $x$ has borders $y, z$ of length $q$ and $r$ respectively, with $q < r < \frac{n}{2}$.

Then $x = yy'y = zz'z$ for words $y'$ and $z'$. Hence $z = yt = t'y$ for some nonempty words $t$ and $t'$.

Then by the Lyndon-Schützenberger theorem we know there exist words $u, v$ with $u$ nonempty, and an integer $d \geq 0$, such that $t' = uv$, $t = vu$, and $y = (uv)^d u$.

Hence $x$ has the prefix $z = yt = (uv)^{d+1} u$, which means $e = \text{ice}(x) \geq \frac{|z|}{|uv|} = \frac{r}{r - q}$. 
Proof of the period inequality

The inequality $r/(r - q) \leq e$ is equivalent to $r/q \geq e/(e - 1)$.

If $b_1 < b_2 < \cdots < b_t$ are the lengths of all the short borders of $x$ then

\[
\begin{align*}
b_1 & \geq 1 \\
b_2 & \geq (e/(e - 1))b_1 \geq e/(e - 1),
\end{align*}
\]

and so forth, and hence $b_t \geq (e/(e - 1))^{t-1}$.

All these borders are of length at most $n/2$, so $n/2 > b_t \geq (e/(e - 1))^{t-1}$.

Hence

\[
t \leq 1 + \frac{\ln(n/2)}{\ln(e/(e - 1))},
\]

and the result follows. ■
**Theorem.** Let $k \geq 2$. Over a $k$-letter alphabet, the expected number of borders (equivalently, the number of nontrivial periods) of a length-$n$ word is $k^{-1} + k^{-2} + \cdots + k^{1-n} \leq \frac{1}{k-1}$.

**Proof.** By the linearity of expectation, the expected number of borders is the sum, from $i = 1$ to $n - 1$, of the expected value of the indicator random variable $B_i$ taking the value 1 if there is a border of length $i$, and 0 otherwise.

Once the left border of length $i$ is chosen arbitrarily, the $i$ bits of the right border are fixed, and so there are $n - i$ free choices of symbols.

This means that $E[B_i] = k^{n-i}/k^n = k^{-i}$. 
**Theorem.** The expected value of $\text{ice}(x)$, for finite or infinite words $x$, is $\Theta(1)$.

**Proof.** Let’s count the fraction $H_j$ of words having at least a $j$’th power prefix. Count the number of words having a $j$’th power prefix with period 1, 2, 3, etc. This double counts, but shows that 

$$H_j \leq k^{1-j} + k^{2(1-j)} + \cdots = 1/(k^{j-1} - 1) \text{ for } j \geq 2.$$ 

Clearly $H_1 = 1$. Then $H_{j-1} - H_j$ is the fraction of words having a $(j - 1)$th power prefix but no $j$th power prefix. These words will have an ice at most $j$. So the expected value of ice is bounded above by

$$2(H_1 - H_2) + 3(H_2 - H_3) + 4(H_3 - H_4) + \cdots$$

$$= 2H_1 + H_2 + H_3 + H_4 + \cdots = 2 + H_2 + H_3 + H_4 + \cdots$$

$$= 2 + \sum_{j \geq 2} 1/(k^{j-1} - 1) = 2 + \sum_{j \geq 1} 1/(k^{j} - 1).$$
Let $0 < \alpha < 1$ be an irrational real number with continued fraction expansion $[0, a_1, a_2, \ldots]$.

The *characteristic Sturmian word* $x_\alpha$ is an infinite word

$$x_1 x_2 x_3 \cdots$$

defined by

$$x_i = \lfloor (i + 1)\alpha \rfloor - \lfloor i\alpha \rfloor.$$ 

For example, for $\alpha = \sqrt{2} - 1$ the characteristic Sturmian word $x_\alpha$ is

$$010100101001010100101001010100 \cdots.$$
You were waiting patiently for the numeration systems. Here they are.

With every real irrational $\alpha$, $0 < \alpha < 1$, we associate a numeration system based on the continued fraction expansion $\alpha = [0, a_1, a_2, a_3, \ldots]$ This is called the Ostrowski $\alpha$-numeration system.

Define $p_i/q_i = [0, a_1, \ldots, a_i]$ to be the $i$’the convergent. In the (ordinary) Ostrowski $\alpha$-numeration system, we write

$$n = \sum_{0 \leq i \leq t} d_i q_i$$

where $d_t > 0$ and the $d_i$ satisfy certain inequalities.
But we’re going to be more concerned with the *lazy Ostrowski system* (Epifanio et al., 2012, 2016).

This representation is again defined through the sum \( n = \sum_{0 \leq i \leq t} d_i q_i \) but with slightly different conditions:

(a) \( 0 \leq d_0 < a_1 \);
(b) \( 0 \leq d_i \leq a_{i+1} \) for \( i \geq 1 \);
(c) For \( i \geq 2 \), if \( d_i = 0 \), then \( d_{i-1} = a_i \);
(d) If \( d_1 = 0 \), then \( d_0 = a_1 - 1 \).

By convention, we write it as a finite word \( d_t d_{t-1} \cdots d_1 d_0 \), starting with the most significant digit.
Here it is in words:

From the lazy Ostrowski $\alpha$-representation of $n$, one can directly read off all the periods of the length-$n$ prefix $X_n$ of the Sturmian characteristic word $x_\alpha$.

More precisely,
Let \( Y_n \) for \( n \geq 1 \) be the prefix of \( x_\alpha \) of length \( n \).

Let \( \text{PER}(n) \) denote the set of all periods of \( Y_n \) (including the trivial period \( n \)).

**Theorem.** (a) The number of periods of \( Y_n \) (including the trivial period \( n \)) is equal to the sum of the digits in the lazy Ostrowski representation of \( n \).

(b) Suppose the lazy Ostrowski representation of \( n \) is \( \sum_{0 \leq i \leq t} d_i q_i \).

Define

\[
A(n) = \left\{ eq_j + \sum_{j < i \leq t} d_i q_i : 1 \leq e \leq d_j \text{ and } 0 \leq j \leq t \right\}.
\]

Then \( \text{PER}(n) = A(n) \).
Example of the theorem

As an example of the theorem, suppose \( \alpha = \sqrt{2} - 1 \).

Write \( n = 23 \) in lazy Ostrowski: \( 12 + 2 \cdot 5 + 1 \).

Then the periods are
\[
12, 12 + 5 = 17, 12 + 5 + 5 = 22, 12 + 5 + 5 + 1 = 23.
\]

So the nonempty borders are size 11, 6, 1.

Take \( Y_{23} = 01010010100101010010100 \).

Here are the borders:

\[
010100101001010010100
01010010100101010010100
01010010100101010010100
010100101001010010100
\]
Brief sketch of the proof

Let $X_i = Y_{q_i}$.

Frid (2018) defined two kinds of Ostrowski representations.

A representation $n = \sum_{0 \leq i \leq t} d_i q_i$ is *legal* if $0 \leq d_i \leq a_{i+1}$.

A representation $n = \sum_{0 \leq i \leq t} d_i q_i$ is *valid* if $Y_n = X_t^{d_t} \cdots X_0^{d_0}$.

She proved the very nice result: every legal representation is valid.
Brief sketch of the proof

Let \( n = \sum_{0 \leq i \leq t} d_i q_i \) be the lazy Ostrowski representation of \( n \). It’s legal, hence valid, hence \( Y_n = X_t^{d_t} X_{t-1}^{d_{t-1}} \cdots X_0^{d_0} \).

What we want to show is that each of the following is a period of \( Y_n \):

\[
X_t, \ X_t^2, \ldots, \ X_t^{d_t}, \ X_t^{d_t} X_{t-1}, \ X_t^{d_t} X_{t-1}^2, \ldots, \ X_t^{d_t} X_{t-1}^{d_{t-1}}, \ldots, \ X_t^{d_t} X_{t-1}^{d_{t-1}} \cdots X_1^{d_1} X_0, \ X_t^{d_t} X_{t-1}^{d_{t-1}} \cdots X_1^{d_1} X_0^2, \ldots, \ X_t^{d_t} X_{t-1}^{d_{t-1}} \cdots X_1^{d_1} X_0^{d_0}. 
\]

To show \( A(n) \subseteq \text{PER}(n) \), we let \( U \) be one of the words above. Then by Frid’s theorem \( Y_n = U Y_{n'} \) for an appropriate \( n' \).

But \( Y_{n'} \) is a prefix of \( Y_n \), so \( Y_n \) is a prefix of \( U Y_n \).

So \( U \) is a period of \( Y_n \), as desired. That proves one direction of our theorem. For the other direction, we use an induction.
Philipp Hieronymi and his group at Illinois have implemented a prover for Sturmian characteristic words.

With this prover they were able to prove our Main Result #2 above just by stating it in first-order logic!
In the special case of the Fibonacci word $f$, we have 
\[ \alpha = \left( \sqrt{5} - 1 \right) / 2. \]

To get the periods of the length-$n$ prefix $Y_n$ of $f$, write $n$ in “lazy Fibonacci” representation:

\[ n = F_{a_t} + F_{a_{t-1}} + \cdots + F_{a_1} \]

where $a_t > a_{t-1} > \cdots > a_1$.

Then the periods are

\[ F_{a_t}, \]
\[ F_{a_t} + F_{a_{t-1}}, \]
\[ \cdots, \]
\[ F_{a_t} + F_{a_{t-1}} + \cdots + F_{a_1}. \]
More results on the Fibonacci word:

The shortest prefix of $f$ having exactly $n$ periods (including the trivial period) is of length $F_{n+3} - 2$, for $n \geq 1$.

The longest prefix of $f$ having exactly $n$ periods (including the trivial period) is of length $F_{2n+2} - 1$, for $n \geq 1$.

The least period of $f[0..m-1]$ is $F_n$ for $F_{n+1} - 1 \leq m \leq F_{n+2} - 2$ and $n \geq 2$. 
Tightness of the inequality on periods

Let $g_s$, for $s \geq 1$, be the prefix of length $F_{s+2} - 2$ of $f$. Thus, for example, $g_1 = \epsilon$, $g_2 = 0$, $g_3 = 010$, $g_4 = 010010$, and so forth.

In our period inequality

$$\text{nnp}(x) \leq \frac{e}{2} + 1 + \frac{\ln(n/2)}{\ln(e/(e - 1))}$$

the bound is tight, up to an additive factor, for the words $g_s$.

Let $\tau = (1 + \sqrt{5})/2$, the golden ratio.

**Theorem.** Take $x = g_s$ for $s \geq 4$. Then the left-hand side of the inequality is $s - 2$, while the right-hand side is asymptotically $s + c$ for $c = 3 + \tau^2/2 - (\ln 2\sqrt{5})/(\ln \tau) \approx 1.19632$. 
What we have seen suggests exploring

\[ M(x) := \frac{\text{nnp}(x)}{\text{ice}(x) \ln |x|} \]

as a measure of periodicity for finite words \( x \). It also suggests studying the following measures of periodicity for infinite words \( x \).

For \( n \geq 2 \) let \( Y_n \) be the prefix of length \( n \) of \( x \). Then define

\[ P(x) := \limsup_{n \to \infty} M(Y_n) \]
\[ p(x) := \liminf_{n \to \infty} M(Y_n) \]

For the “typical” infinite word \( x \) we have \( P(x) = p(x) = 0 \).

Thus it is of interest to find words \( x \) where \( P(x) \) and \( p(x) \) are large.
The *period-doubling word* \( d \) is defined to be the fixed point of the morphism sending 1 \( \to \) 10 and 0 \( \to \) 11.

**Theorem.** \( P(d) = \frac{1}{2 \ln 2} \approx 0.7213 \) and \( p(d) = \frac{1}{4 \ln 2} \approx 0.36067 \).
An example: the period-doubling word

Proof. Let $r(n)$ denote the number of periods (including the trivial period) in the length-$n$ prefix of $d$. We can use the theorem-proving software Walnut to calculate the periods of prefixes of $d$.

We write a first-order logical formula $\text{pdp}(m, p)$ stating that the prefix of length $m \geq 1$ of $d$ has period $p$, $1 \leq p \leq m$:

$$\text{pdp}(m, p) := (1 \leq p \leq m) \land d[0..m−p−1] = d[p..m-1] = (1 \leq p \leq m) \land \forall t \ (0 \leq t < m − p) \implies d[t] = d[t + p].$$
An example: the period-doubling word

Such a formula can be automatically translated, using Walnut, to an automaton that recognizes the language

\[ \{(n, p)_2 : \text{the length}-n \text{ prefix of } d \text{ has period } p \}. \]
An example: the period-doubling word

Such an automaton can be automatically converted by Walnut to a linear representation for $r(n)$. This is a triple $(v, \rho, w)$ where $v, w$ are vectors, and $\rho$ is a matrix-valued morphism, such that $r(n) = v \cdot \rho(((n)_2)) \cdot w$.

The values are given below:

\[
v = [1 \ 0 \ 0 \ 0 \ 0 \ 0] \quad \rho(0) = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad \rho(1) = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} \quad w = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}.
\]
An example: the period-doubling word

From this we can easily compute the relations

\[ r(0) = 0 \]
\[ r(2n + 1) = r(n) + 1, \quad n \geq 0 \]
\[ r(4n) = r(n) + 1, \quad n \geq 1 \]
\[ r(4n + 2) = r(n) + 1, \quad n \geq 0. \]

Reinterpreting this definition for \( r \), we see that \( r(n) \) is equal to the length of the (unique) factorization of \((n)_2\) into the factors 1, 00, and 10.

It now follows that

(a) The smallest \( m \) such that \( r(m) = n \) is \( m = 2^n - 1 \);
(b) The largest \( m \) such that \( r(m) = n \) is \( m = \lfloor 2^{2n+1}/3 \rfloor \), with \((m)_2 = (10)^n\).
An example: the period-doubling word

Similarly, we can use \texttt{Walnut} to determine the smallest period $p$ of every length-$n$ prefix of $d$. We use the predicate

$$ pdlp(n, p) := pdp(n, p) \land \forall q (1 \leq q < p) \implies pdp(n, q). $$

This gives the automaton

![Automaton Diagram]

Inspection of this automaton shows that least period of the prefix of length $n$ is, for $s \geq 2$, equal to $3 \cdot 2^{s-2}$ for $2^s \leq n < 5 \cdot 2^{s-2}$ and $2^s$ for $5 \cdot 2^{s-2} \leq n < 2^{s+1}$. So the ice of every length-$n$ prefix of $d$ for $2^t - 1 \leq n \leq 2^{t+1} - 2$, is $2 - 2^{1-t}$.

The result now follows.
Recall that an *overlap* is a word of the form $axaxa$, where $a$ is a single letter and $x$ is a (possibly empty) word. An example in English is the word *alfalfa*. We say a word is *overlap-free* if no finite factor is an overlap.

Define $f(p)$ to be the length of the shortest overlap-free binary word having $p$ nontrivial periods.

**Theorem.** We have $f(1) = 2$, $f(2) = 5$, and

$$f(p) \leq \frac{17}{6} \cdot 4^{p-2} + \frac{2}{3} \quad \text{for } p \geq 3.$$
Proof sketch. Define $\mu(0) = 01$ and $\mu(1) = 10$. If $w = axa$ for a single letter $a$, define $\gamma(w) = a^{-1}\mu^2(w)a^{-1}$. Furthermore define

$$A_n = \begin{cases} 
001001100100, & \text{if } n = 3; \\
\gamma(A_{n-1}), & \text{if } n \geq 4.
\end{cases}$$

Then we can prove by induction that $A_n$ is a overlap-free palindrome with $n$ nontrivial periods for $n \geq 3$. ■
Recall that a *square* is a word of the form $xx$, where $x$ is a nonempty word. An example in English is the word *murmur*. We say a word is *squarefree* if no finite factor is a square.

Define $g(p)$ to be the length of the shortest squarefree ternary word having $p$ nontrivial periods.

**Theorem.** We have $g(1) = 3$, $g(2) = 7$, and

$$g(p) \leq \frac{17}{12} \cdot 4^{p-1} + \frac{1}{3} \quad \text{for } p \geq 3.$$
Open problems

1. Prove that the bound for binary overlap-free words $f(p)$ obtained above is optimal.

2. For ternary squarefree words, determine the asymptotic behavior of $g(p)$.

3. Find an exact expression for the limit, as $n \to \infty$, of the expected value of ice of the length-$n$ words over a $k$-letter alphabet. For example, for $k = 2$, this seems to be about 2.494.