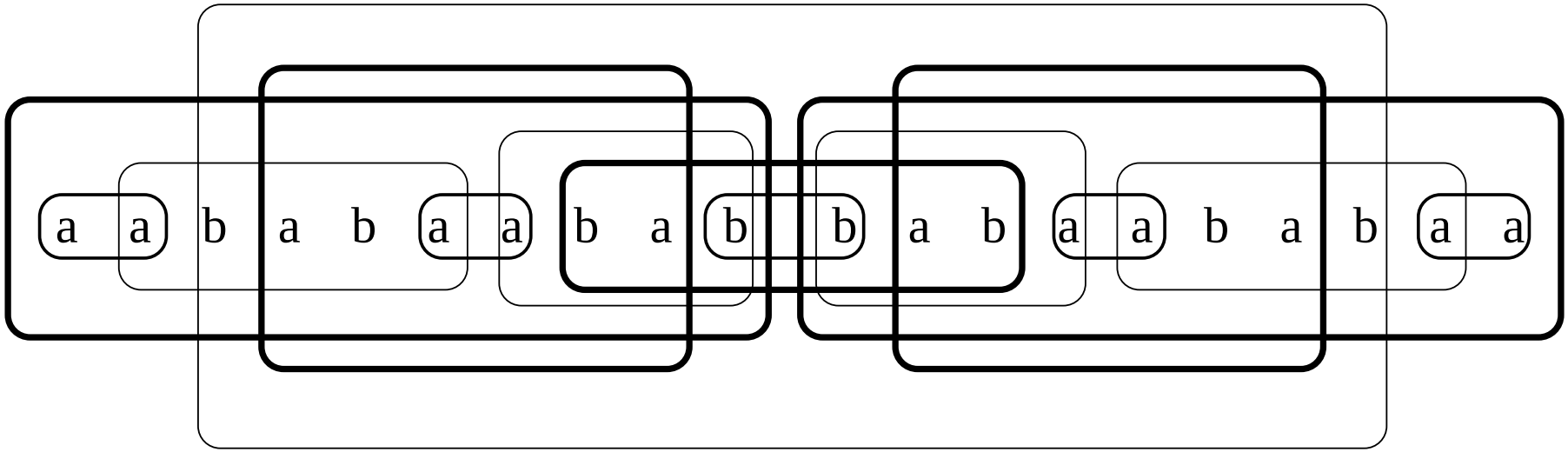
The Number of Runs in Sturmian Words

P. Baturo, M. Piatkowski, W. Rytter

University of Warsaw, Poland

and Copernicus University, Torun

The **runs** are maximal repetitions in a string.



The structure of $RUNS((aabab)^2(babaa)^2)$.

Denote by $\mathcal{S}$ the class of *standard Sturmian* words. It is a class of highly compressed words extensively studied in combinatorics of words, including the well known Fibonacci words.

The suffix automata for these words have a very particular structure. This implies a simple characterization (described in the paper by the Structural Lemma) of the periods of runs (maximal repetitions) in Sturmian words.

Using this characterization we derive an explicit formula for the number $\rho(w)$ of runs in words $w \in \mathcal{S}$, with respect to their *recurrences* (*directive sequences*). We show that

$$\frac{\rho(w)}{|w|} \leq \frac{4}{5} \quad \text{for each } w \in \mathcal{S},$$

There is an infinite sequence of strictly growing words $w_k \in \mathcal{S}$ such that

$$\lim_{k\to\infty} \frac{\rho(w_k)}{|w_k|} = \frac{4}{5}$$

The complete understanding of the function ρ for a large class $\mathcal{S}$ of complicated words is a step towards better understanding of the structure of runs in words. We also show how to compute the number of runs in a standard Sturmian word in linear time with respect to the size of its compressed representation (recurrences describing the word). This is an example of a very fast computation on texts given implicitly in terms of a special grammar-based compressed representation (usually of logarithmic size with respect to the explicit text).

The standard words are a generalization of Fibonacci words and, like Fibonacci words are described by recurrences.

The recurrence for a standard word is related to so called *directive sequence* - an integer sequence of the form

$\gamma = (\gamma_0, \gamma_1, ..., \gamma_n)$, where $\gamma_0 \geq 0$, $\gamma_i > 0$ for $0 < i \leq n$.

The standard word corresponding to γ, denoted by $S(\gamma) = x_{n+1}$, is defined by recurrences:

$$x_{-1} = b,\ x_0 = a,\ x_1 = x_0^{\gamma_0} x_{-1},\ x_2 = x_1^{\gamma_1} x_0, \tag{1}$$

$$x_3 = x_2^{\gamma_2} x_1,\ \ldots x_n = x_{n-1}^{\gamma_{n-1}} x_{n-2},\ x_{n+1} = x_n^{\gamma_n} x_{n-1} \tag{2}$$

For example the recurrence for the 4-th Fibonacci word is:

$$fib_{-1} = b,\ \ fib_0 = a,\ \ fib_1 = fib_0^1 b,\ \ fib_2 = fib_1^1 fib_0,$$

$$fib_3 = fib_2^1 fib_1,\ \ fib_4 = fib_3^1 fib_2.$$

Hence

$$fib_4 \ = \ abaababa \ = \ S(\gamma_0, \gamma_1, \gamma_2, \gamma_3)$$

where

$$(\gamma_0, \gamma_1, \gamma_2, \gamma_3) = \ (1, 1, 1, 1)$$

We consider here standard words starting with the letter a,

hence assume $\gamma_0 > 0$. The case $\gamma_0 = 0$ can be considered similarly.

For even $n > 0$ a word x_n has suffix ba, and for odd n it has suffix ab.

The number $N = |x_{n+1}|$ is the (real) size, while n can be thought of as the compressed size.

Consider more complicated example (used later to demonstrate counting of runs), let

$$\gamma \;=\; (1,2,1,3,1)$$

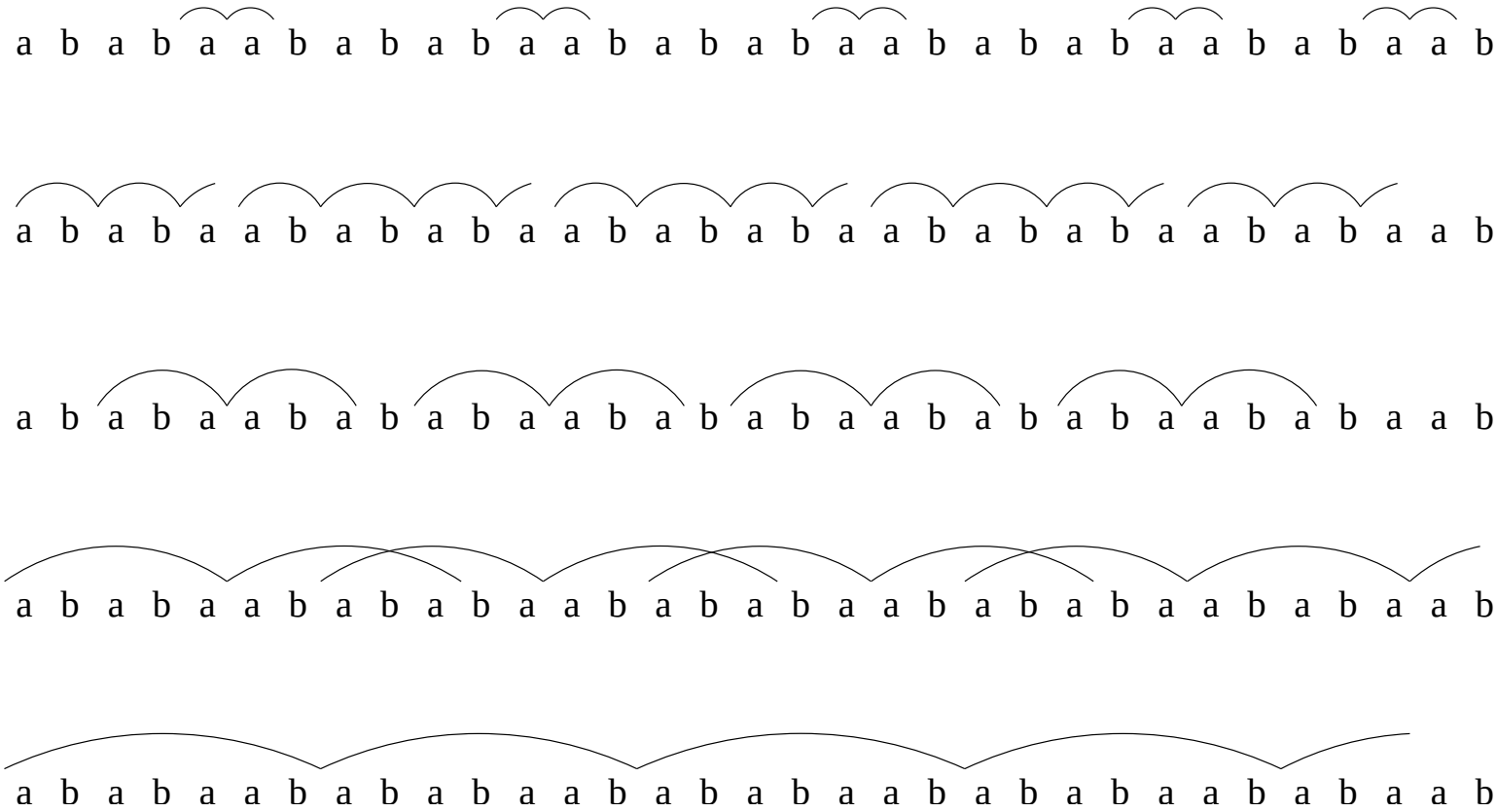
we have

$$S(\gamma) \;=\; ababaabababaabababaabababaababaab$$

The corresponding recurrence is:

$$x_{-1} = b;\;\; x_0 = a,\;\; x_1 = x_0^1 x_{-1},\;\; x_2 = x_1^2 x_0,$$

$$x_3 = x_2^1 x_1,\;\; x_4 = x_3^3 x_2,\;\; x_5 = x_4^1 x_3.$$



The structure of runs of $S(1,2,1,3,1)$. There are 5 runs with period $|a|$, 5 with period $|ab|$. We have 10 *short* runs (period of size at most $|x_1| = |ab|$), 8 medium (with period $|x_1| < p \leq |x_2| = 5$, and 1 large run. Consequently $\rho(1,2,1,3,1) = 19$.

The computation of runs in $S(\gamma_0, \gamma_1, \ldots \gamma_n)$ is reduced to a similar computation for $S(\gamma_1, \gamma_2, \ldots \gamma_n)$.
The relation between $S(\gamma_0, \gamma_1, \ldots \gamma_n)$ and $S(\gamma_1, \gamma_2, \ldots \gamma_n)$ is described in terms of morphisms transforming one of them to the other.

For $\gamma = (\gamma_0, \gamma_1, \ldots, \gamma_n)$ define the sequence of morphisms:

$$h_i(a) = a^{\gamma_i} b, \quad h_i(b) = a \ , \quad \text{for } 0 \leq i \leq n$$

FOr example for Fibonacci words we have:

$$h_i(a) = ab, \quad h_i(b) = a$$

Lemma 1

Assume $0 \leq i < n$.

We have

$$S(\gamma_n) = h_n(a),$$

$$S(\gamma_i, \gamma_{i+1} \ldots, \gamma_n) = h_i(S(\gamma_{i+1}, \gamma_{i+2} \ldots, \gamma_n)).$$

For Fibonaci words we have:

$$S(1) = h(a) = ab,$$

$$S(1,1) = h(S(1)) = h(ab) = aba,$$

$$S(1,1,1) = h(S(1,1)) = h(aba) = abaab.$$

Let $|w|_r$ denote the number of occurrences of a letter $r \in \{a,\ b\}$ in the word w. Denote

$$N_\gamma(k) \ = \ |S(\gamma_k, \gamma_{k+1}, \ldots \gamma_n)|_a$$

$$M_\gamma(k) \ = \ |S(\gamma_k, \gamma_{k+1}, \ldots \gamma_n)|_b$$

The numbers $N_\gamma(k),\ M_\gamma(k)$ satisfy the equation:

$$N_\gamma(k) \ = \ \gamma_k \ N_\gamma(k+1) + N_\gamma(k+2); \quad M_\gamma(k) \ = \ N_\gamma(k+1) \qquad (3)$$

Example

In case of the directive sequence $(1, 1, \ldots 1)$ describing the Fibonacci word the numbers $N_\gamma(k)$ are Fibonacci numbers, since the number of letters a in fib_n equals the size of fib_{n-1}.

For the word

$$S(1,2,1,3,1) \;=\; ababaababababaababababaababababaababaab$$

from Figure 1 we have $\gamma \;=\; (1,2,1,3,1)$ and:

$$S(1) = ab,\;\; S(3,1) = aaaba,\;\; S(1,3,1) = (ab)^3 a\; ab,$$

$$S(2,1,3,1) = ((aaba)^3\; aab)\; aaba,\;\; S(1,2,1,3,1) = x_5$$

$$N_\gamma(2) = |S(1,3,1)|_a = 5,\;\; N_\gamma(1) = |S(2,1,3,1)|_a = 14,\;\; N_\gamma(0) = 19.$$

Lemma 2 *Let*

$A = N_\gamma(2),\ B = N_\gamma(3)$ *and* $w = S(\gamma_0, \gamma_1, \ldots \gamma_n)$. *Then*

$$|w| = ((\gamma_0 + 1)\,\gamma_1 + 1)\,A + (\gamma_0 + 1)\,B$$

For Fibonacci words we have

$$|S(\gamma)| = fib_{n+1},\ N\gamma(1) = fib_{n-1},$$
$$A = N_\gamma(2) = fib_{n-2},\ B = N_\gamma(3) = fib_{n-3}$$

We have

$$fib_{n+1} = (2+1)\,A + 2 \cdot B = 3 \cdot fib_{n-2} + 2 \cdot fib_{n-3}$$

For example

$$13 = 3 \cdot 3 + 2 \cdot 2$$

Theorem 1 **[Formula for the number of runs]**

Let $n \geq 3$. *The number of runs in* $S(\gamma_0, \ldots, \gamma_n)$ *equals:*

$$\rho(\gamma) = \begin{cases} 2\,A + 2\,B + \Delta(\gamma) - 1 & \textit{if} \quad \gamma_0 = \gamma_1 = 1 \\ (\gamma_1 + 2)\,A + B + \Delta(\gamma) - odd(n) & \textit{if} \quad \gamma_0 = 1;\ \gamma_1 > 1 \\ 2A + 3B + \Delta(\gamma) - even(n) & \textit{if} \quad \gamma_0 > 1;\ \gamma_1 = 1 \\ (2\,\gamma_1 + 1)\,A \ + \ 2\,B \ + \ \Delta(\gamma) & Otherwise \end{cases},$$

where:

$$\Delta(\gamma) = n - 1 - (\gamma_1 + \ldots + \gamma_n) - unary(\gamma_n).$$

$$A \ = \ N_\gamma(2) \ = \ |S(\gamma_2, \gamma_3 \ldots, \gamma_n)|_a,$$

$$B \ = \ N_\gamma(3) \ = \ |S(\gamma_3, \gamma_4 \ldots, \gamma_n)|_a$$

Example.

We show how to compute $\rho(1,2,1,3,1)$, using our formula. In this case

$$\gamma \;=\; (\gamma_0, \gamma_1\gamma_2, \gamma_3, \gamma_4) \;=\; (1,2,1,3,1) \;\text{ and }\; n = 4$$

$$A = N_\gamma(2) = 5,\;\; B = N_\gamma(3) = 4,\;\; \Delta = (4-1) - 7 = 4,\;\; even(n) = 1$$

Theorem 1 implies correctly:

$$\rho(\gamma) \;=\; (\gamma_1 + 2)A + B + \Delta - even(4)$$

$$= \;4A + B - 4 - 1 \;=\; 4 \cdot 5 + 4 - 4 - 1 \;=\; 19.$$

Example.

As the next example derive the formula for the number of runs in Fibonacci word $fib_n = S(1,1,\ldots 1)$ (n ones) for $n \geq 3$.

Let F_n be the n-th Fibonacci number. In this case $N_\gamma(k) = F_{n-k-1}$. According to formula from Theorem 1 we have

$$\rho(fib_n) = 2N_\gamma(2) + 2N_\gamma(3) + n - 1 - n - 1 - 1$$

$$\boxed{\rho(fib_n) = 2\,F_{n-3} + 2\,F_{n-4} - 3 = 2\,F_{n-2} - 3.}$$

We have $\Delta(\gamma) < 0$,

we ignore Δ and estimate $\rho(\gamma)/|S(\gamma)|$ in terms of $A,\ B$.

Four cases should be considered, depending on

$$\gamma_0 > 1,\ \ \gamma_1 > 1$$

Tedious technical estimations give the following result

Theorem 2 [Upper Bound]

For each $w \in \mathcal{S}$:

$$\boxed{\rho(w) \leq \frac{4}{5}\ |w|}$$

Theorem 3 [Lower Bound]

For the class $\mathcal{S}$ of standard words we have

$$\boxed{\sup\ \{\ \frac{\rho(w)}{|w|}\ :\ w \in \mathcal{S}\ \}\ =\ 0.8.}$$

Proof: Let

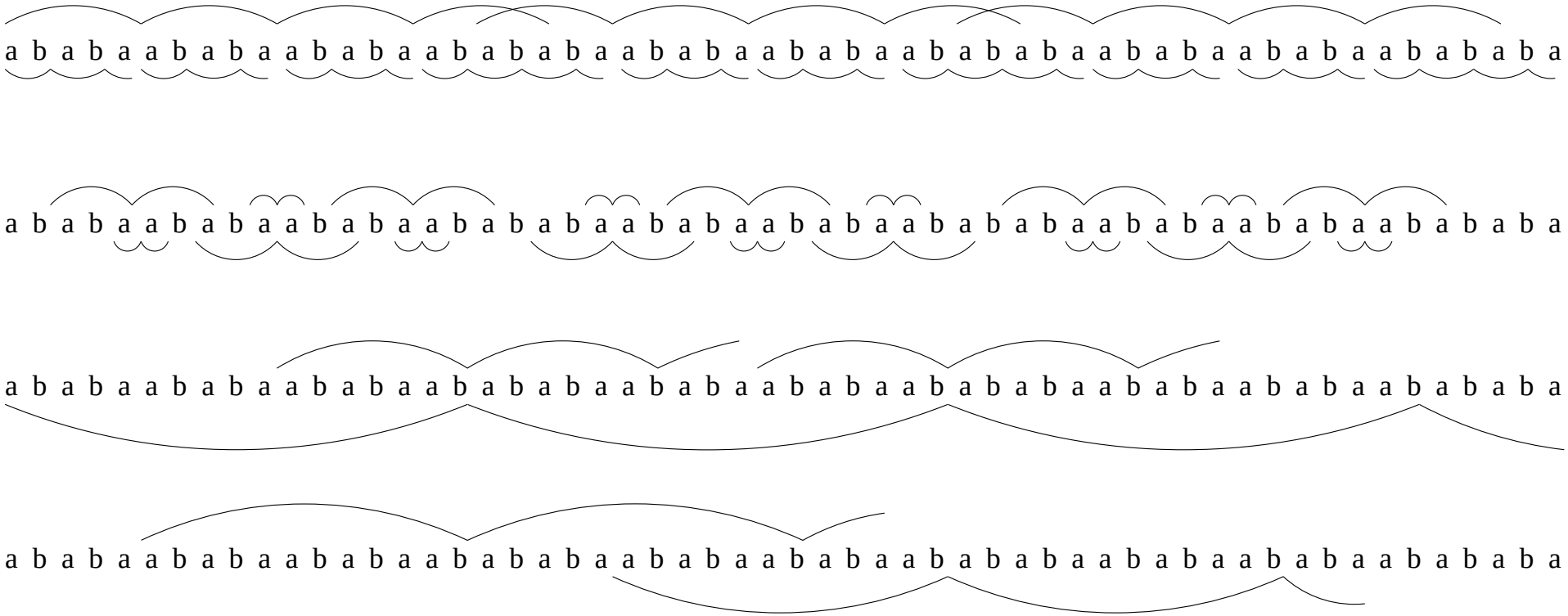
$$w_k = \mathsf{S}(1,2,k,k) = \Big((ababa)^k\ ab\Big)^k ababa,$$

We have $|w_k| = 5k^2 + 2k + 5$.

Theorem 1 implies that $|\rho(1,2,k,k)| = 4k^2 - k + 3$. Consequently

$$\lim_{k\to\infty} \frac{\rho(w_k)}{|w_k|} = \lim_{k\to\infty} \frac{4k^2 - k + 3}{5k^2 + 2k + 5} = 0.8$$

□



The structure of runs of $S(1, 2, k, k)$
for $k = 3$, there are $4k^2 - k + 3 \;=\; 36$ runs.

Theorem 4

We can count number of runs in standard word $S(\gamma_0, \ldots, \gamma_n)$ in time $O(n)$.

Proof: We need only to compute in $O(n)$ time the numbers $N_\gamma(k)$ for $k = 1, 2, 3$. We can compute it iterating Equation 2.

Algorithm Compute $N_\gamma(k)$;
$x := \gamma_{n-1}$; $y := 1$;
for $i := n - 2$ **downto** k **do**
$\quad (x, y) := (\gamma_i \cdot x + y,\ x)$
return x;

Now we apply the formulas from Theorem 1 and Lemmas 1,2. □

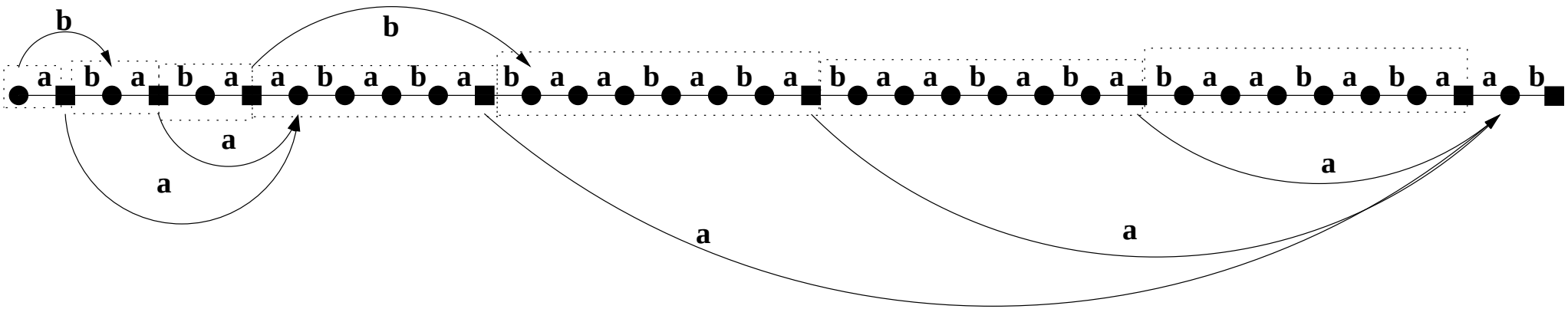
Assume that x_i' are as given by recurrences described in Equations 1,2.

The structure of subword graphs for standard Sturmian words is very special, in particular it implies the following fact.

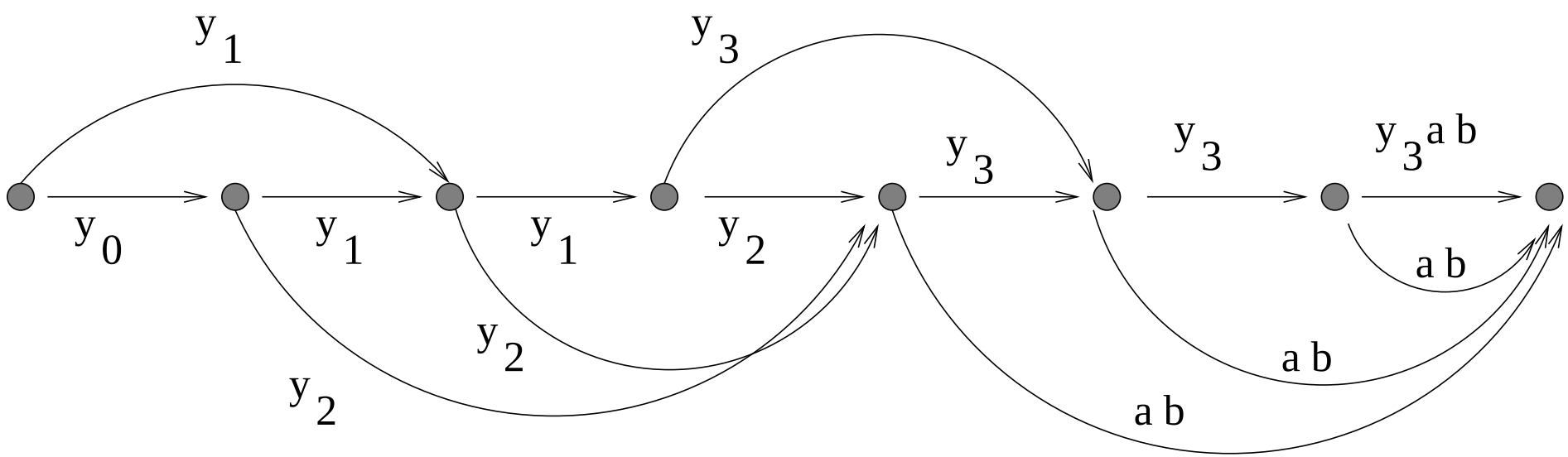
Lemma 3 [Structural Lemma]
The period of each run of $S(\gamma_0, \gamma_1, \ldots, \gamma_n)$ is of the form

$$\boxed{x_i^j x_{i-1}}$$

where $0 \leq j < \gamma_i$.



The structure of the subword graph of $Word(1, 2, 1, 3, 1)$.



The compacted version of the subword graph. We have on edges the reversed words x_k, denoted by y_k.

$$y_0 = a;\ \ y_1 = ba;\ \ y_2 = ababa;\ \ y_3 = baababa$$

$$Word(1,2,1,3,1) \ = \ a\ ba\ ba\ ababa\ baababa\ baababa\ baababa\ ab \ = \ y_0 y_1^2 y_2 y_3^3 \hat{y}_4.$$

We say that a run is

short:

$$period \leq |x_1|$$

large:

$$period > |x_2|$$

medium:

$$|x_2| < period \leq |x_2|.$$

Denote by

$$\rho_{med}(\gamma),\ \rho_{med}(\gamma),\ \rho_{large}(\gamma)$$

the number of short, medium and large runs in $S(\gamma)$

Example.

Let $=$ S(1,2,1,3.1).

W have 10 short runs with periods: a, ab,

8 medium runs with periods: aba, ababa,

1 large run with period : ababaab,

Next each type of runs is counted separately.

Lemma 4 **[Short Runs]** *The number of short runs in* $S(\gamma)$ *is:*

$$\rho_{short}(\gamma) = \begin{cases} N_\gamma(2) + N_\gamma(3) - 1 & \textit{if} \quad \gamma_0 = \gamma 1 = 1 \\ 2\, N_\gamma(2) - odd(n) & \textit{if} \quad \gamma_0 = 1;\ \gamma 1 > 1 \\ N_\gamma(1) + N_\gamma(3) - even(n) & \textit{if} \quad \gamma_0 > 1;\ \gamma 1 = 1 \\ N_\gamma(1) + N_\gamma(2) & \textit{otherwise} \end{cases}$$

Lemma 5 **[Medium Runs, $\mathbf{n \geq 3}$]** *If* $n \geq 3$ *then*

$$\boxed{\rho_{med}(\gamma) \;=\; N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1}$$

Lemma 6 **[Medium Runs, n=2]** *If* $n = 2$ *then*

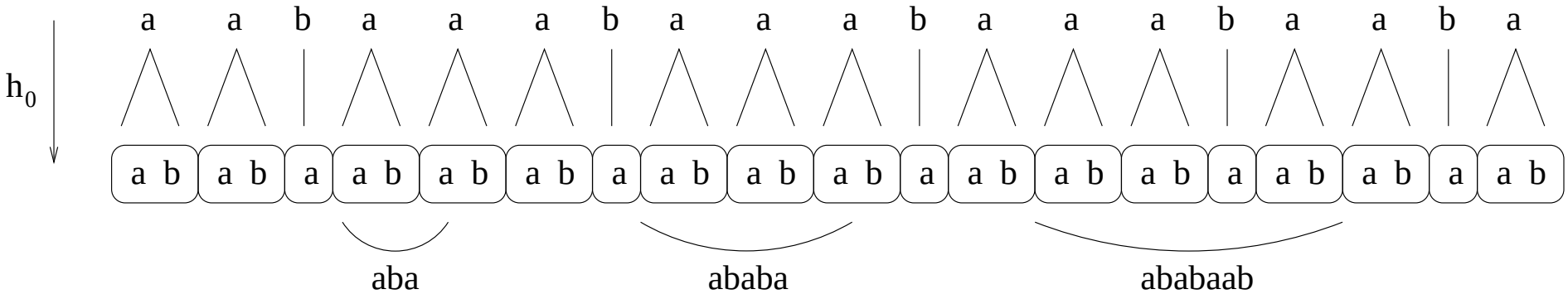
$$\boxed{\rho_{med}(\gamma) \;=\; N_\gamma(1) - N_\gamma(2) - \gamma_1 + 1 - unary(\gamma_n)}$$

We say that a subword w of x is **synchronized** with h in x iff each occurrence of w in x starts at the beginning of some h-block and ends at the end of some h-block.

Lemma 7 [Synchronization Lemma]
The large run-periods are synchronized with h_0 *in* $S(\gamma_0, \ldots, \gamma_n)$

The figure shows examples of synchronized and non-synchronized subwords with the morphism $h_0 \ : \ S(2,1,3,1) \rightarrow S(1,2,1,3,1)$ related to the morphic structure of $S(1,2,1,3,1)$. We have:

$$h_0(a) = a^{\gamma_0} b, \ \ h_0(b) = a$$



The medium run-periods $x_1 x_0 = aba$ and $x_2 = ababa$ do not synchronize with morphism h_0, while the large run-period $x_3 \ = \ ababaab$ is synchronized with h_0.

As a consequence of the Synchronization Lemma we have:

Lemma 8 **[Recurrence Lemma]**

$$\rho_{large}(\gamma_0, \gamma_1, \ldots \gamma_n) \;=\; \rho_{large}(\gamma_1, , \gamma_2, \ldots \gamma_n) \;+\; \rho_{med}(\gamma_1, \gamma_2, \ldots \gamma_n).$$

Lemma 9 **[Large Runs]**

$$\rho_{large} + \rho_{med} \;=\; N_\gamma(1) \;+\; n-1 \;-\; (\gamma_1 + ... + \gamma_n) \;- unary(\gamma_n)$$

Now the formula in Theorem 1 results by combining the formulas for ρ_{short} and for the sum $\rho_{large} + \rho_{med}$ using the equalities

$$\boxed{\rho(\gamma) \;=\; \rho_{short}(\gamma) + \rho_{med}(\gamma) + \rho_{large}(\gamma)}$$

and

$$N_\gamma(1) = \;\gamma_1 N_\gamma(2) + N\gamma(3).$$