RuFiDiM 2014, Petrozavodsk — September 18th, 2014

Decomposition of a language of factors into sets of bounded complexity

Julien Cassaigne

Institut de mathématiques de Marseille, France
julien.cassaigne@math.cnrs.fr

Joint work with Anna Frid (Marseille and Novosibirsk),
Svetlana Puzynina (Turku and Novosibirsk),
and Luca Zamboni (Lyon and Turku).

Decomposition of a language of factors into sets of bounded complexity

- Examples, definitions
- Linear factor complexity
- Higher factor complexity
- Open problems

Motivation

Let $\mathbf{u} \in A^{\mathbb{N}}$ be an infinite word.
Language of factors: $L = F(\mathbf{u})$.
Factor complexity: $p_L(n) = \#(L \cap A^n)$.

Question:
Is it possible to express elements of L using a smaller language S?

Thue-Morse

Let $t_0 = a$, $\bar{t}_0 = b$, $t_{k+1} = t_k \bar{t}_k$, $\bar{t}_{k+1} = \bar{t}_k t_k$.
Then $t_1 = ab$, $\bar{t}_1 = ba$, $t_2 = abba$, $t_3 = abbabaab$, etc.
Thue-Morse word: $\mathbf{t} = \lim t_k = abbabaabbaababbabaababba\ldots$

If $w \in L = F(\mathbf{t})$ with $|w| \geq 2$, then there is k such that:
w is a factor of t_{k+1} or $\bar{t}_{k+1}$, but neither of t_k nor of $\bar{t}_k$.

Then $w = sp$, where:
s is a suffix of t_k or $\bar{t}_k$, and
p is a prefix of t_k or $\bar{t}_k$.

Let S be the language of those prefixes and suffixes:
then $L \subseteq S.S$, with $p_S(n) \leq 4$, while $p_L(n) = \Theta(n)$.

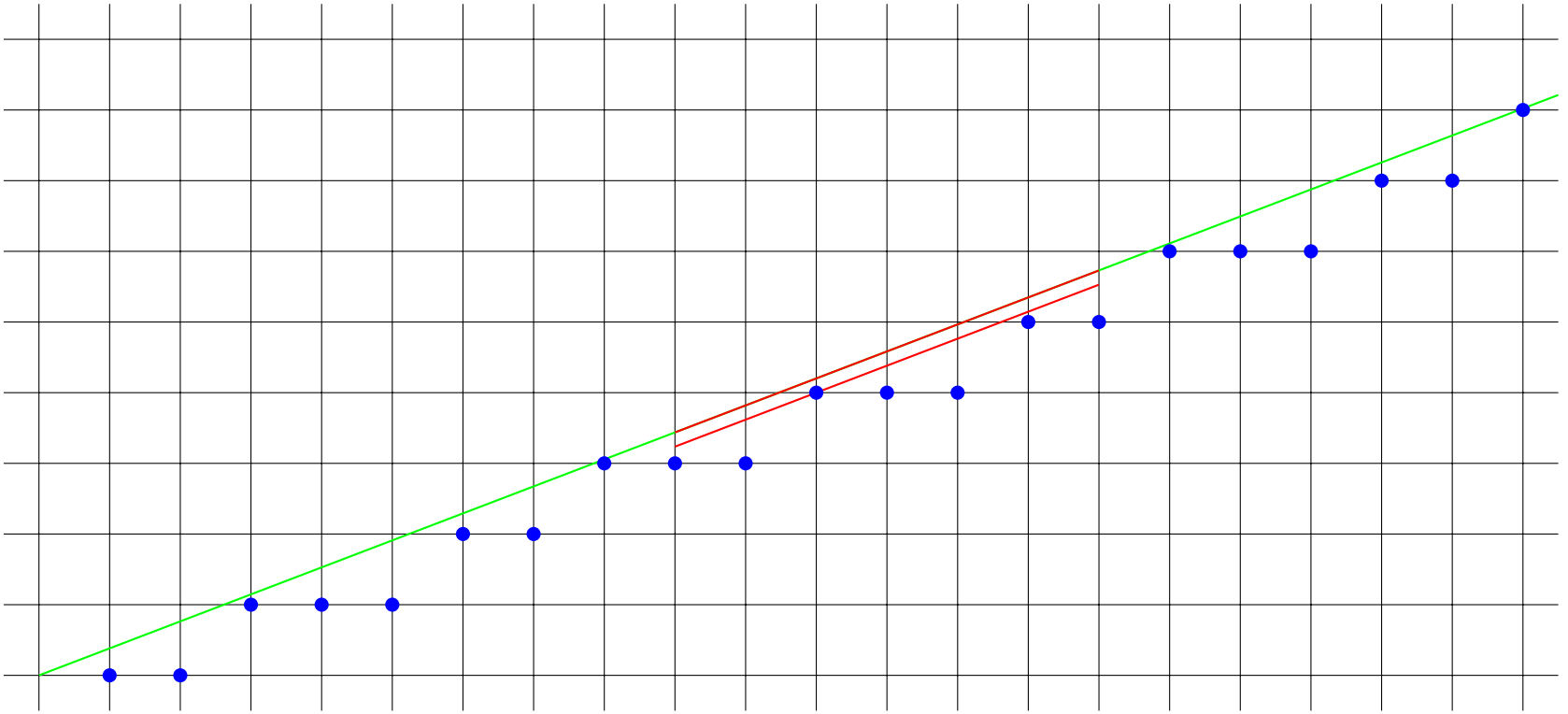
Sturmian words

Fix $\alpha \in [0,1] \setminus \mathbb{Q}$.
Standard Sturmian word of slope α:
$\mathbf{u} = u_1 u_2 \ldots$ with $u_n = \lfloor \alpha(n+1) \rfloor - \lfloor \alpha n \rfloor \in A = \{0,1\}$.

For each $w \in F(\mathbf{u}) \cup A^n$, there is $\rho \in \mathbb{R}$ such that
$w_i = \lfloor \alpha(i+1) + \rho \rfloor - \lfloor \alpha i + \rho \rfloor$ for $0 \le i < n$.
ρ can be adjusted so that $\alpha j + \rho \in \mathbb{Z}$ for some j, $0 \le j \le n$.
Then $w_0 \ldots w_{j-1} = \tilde{x}1$ (or ε if $j = 0$)
and $w_j \ldots w_{n-1} = 0y$ (or ε if $j = n$)
where x and y are prefixes of $\mathbf{u}$.

Here $p_S(n) \le 2$, while $p_L(n) = n + 1$.

Sturmian words



Fibonacci word

$\mathbf{u} = 0100101001001010010100100101001001010010100101\ldots$

$S = \{\varepsilon, 0, 00, 001, 0010, 00100, 001001, 0010010, 00100101, \ldots$
$1, 01, 101, 0101, 00101, 100101, 0100101, 10100101, \ldots\}.$

$F(\mathbf{u}) \cap A^8 = \{00100101, 00101001, 01001001, 01001010,$
$01010010, 10010010, 10010100, 10100100, 10100101\}.$

Definitions

$\mathcal{L}_1$: class of languages L for which p_L is bounded (slender languages).

$\mathcal{L}_k$: class of languages L for which there exist S_1, ..., S_k in $\mathcal{L}_1$ such that $L \subseteq S_1.S_2.\ldots.S_k$.

$\mathcal{W}_k$: class of infinite words $\mathbf{u}$ for which $F(\mathbf{u}) \in \mathcal{L}_k$.

$\mathcal{P}_\alpha$: class of infinite words $\mathbf{u}$ for which $p_{\mathbf{u}}(n) = O(n^\alpha)$.

We have seen that Thue-Morse and Sturmian words are in $\mathcal{W}_2$.

Decomposition and complexity

Question:
How are decomposition classes $\mathcal{W}_k$ related to complexity classes P_α?

Counting argument: $\mathcal{W}_k \subseteq \mathcal{P}_{k-1}$ for all $k \geq 1$.
Indeed, if p_{S_i} is bounded by C, then $S_1.S_2.\ldots.S_k$ contains at most $\binom{n+k-1}{k-1} C^k$ words of length n.

Trivially: $\mathcal{W}_1 = \mathcal{P}_0$.

Is it true in general that $\mathcal{W}_k = \mathcal{P}_{k-1}$?

Linear factor complexity

Our main result is:

Theorem.
$\mathcal{W}_2$ is exactly the class of infinite words with linear factor complexity.

It remains to prove that $\mathcal{P}_1 \subseteq \mathcal{W}_2$.

Let $\mathbf{u} \in \mathcal{P}_1$.
We have to design a way to factor any $v \in F(\mathbf{u})$ as $v = s_v t_v$,
so that $S = \{s_v | v \in F(\mathbf{u})\}$ and $T = \{t_v | v \in F(\mathbf{u})\}$ are slender.
For this we shall use markers.

Markers

$M \subseteq A^n$ is a set of D-markers of length n for $\mathbf{u}$
if every $w \in F(\mathbf{u}) \cap A^{Dn}$ contains an element of M as a factor.

Lemma.
If $\mathbf{u} \in \mathcal{P}_1$, then there exist constants D and R such that, for any $n \in \mathbb{N}$, there is a set M of D-markers of length n for $\mathbf{u}$ with $\#M \leq R$.

Special factors

A word $w \in F(\mathbf{u})$ is a right special factor
if there exist letters $a \neq b$ such that $wa \in F(\mathbf{u})$ and $wb \in F(\mathbf{u})$.

The number of right special factors of length n
is at most $p_{\mathbf{u}}(n+1) - p_{\mathbf{u}}(n)$.

Theorem. [Cassaigne 1996]
If $p_{\mathbf{u}}(n) = O(n)$, then $p_{\mathbf{u}}(n+1) - p_{\mathbf{u}}(n)$ is bounded.
More precisely: if $\forall n \in \mathbb{N}$, $p_{\mathbf{u}}(n) \leq Cn + 1$,
then $\forall n \in \mathbb{N}$, $p_{\mathbf{u}}(n+1) - p_{\mathbf{u}}(n) \leq 2C(2C+1)^2$.

Proof of the lemma

Lemma.
If $\mathbf{u} \in \mathcal{P}_1$, then there exist constants D and R such that, for any $n \in \mathbb{N}$, there is a set M of D-markers of length n for $\mathbf{u}$ with $\#M \leq R$.

Assume $\mathbf{u}$ is not eventually periodic, with $p_{\mathbf{u}}(n) \leq Cn + 1$.
Take for M the set of right special factors of length n.
Let $D = C + 1$ and $R = 2C(2C + 1)^2$. Then $\#M$ is bounded by R.

If a factor w does not contain any right special factor of length n, then it cannot contain any repeated factor of length n, otherwise this would imply periodicity.
Then $|w| \leq Cn + n - 1 < Dn$.

Construction

For each $k \leq 1$, fix a set M_k of D-markers of length 2^k, with $\#M_k \leq R$.

Let $v \in F(\mathbf{u})$, and assume first $n = |v| \geq 2D$.
Choose an occurrence i of v in $\mathbf{u}$.
Let $m \in M_k$ be a marker that occurs in v, with k as large as possible.

We choose one occurrence j of m in v as follows.
Let π be the minimal period of m. If there are occurrences j such that m does not occur at position $i+j+\pi$ in $\mathbf{u}$, choose one (case 1).
Otherwise, let j be the first occurrence (case 2).

We have $v = xm_1m_2y$, with $|x| = j$ and $m_1 = m_2 = 2^{k-1}$.
Let $s_v = xm_1$ and $t_v = m_2y$. If $|v| < 2D$, let $s_v = v$ and $t_v = \varepsilon$.
Let $S = \{s_v | v \in F(\mathbf{u})\}$ and $T = \{t_v | v \in F(\mathbf{u})\}$. Then $F(\mathbf{u}) \subseteq ST$.

Fibonacci

$\mathbf{u} = 01001010010010100101001001010010010100101001010\ldots$

$D = 2$, $R = 1$, $M_1 = \{10\}$, $M_2 = \{0010\}$, $M_3 = \{01010010\}$,
$M_4 = \{0101001001010010\}$, ...

Take $v = 10010010100101001001010$. v occurs in $\mathbf{u}$ at $i = 6$.
v contains two overlapping occurrences of $m = 01010010$,
and no larger marker: $k = 3$.
We choose the second occurrence: $j = 10$,
since m does not occur in $\mathbf{u}$ at position $i + j + \pi = 21$.

Then $s_v = 10010010100101$ and $t_v = 0010010$.

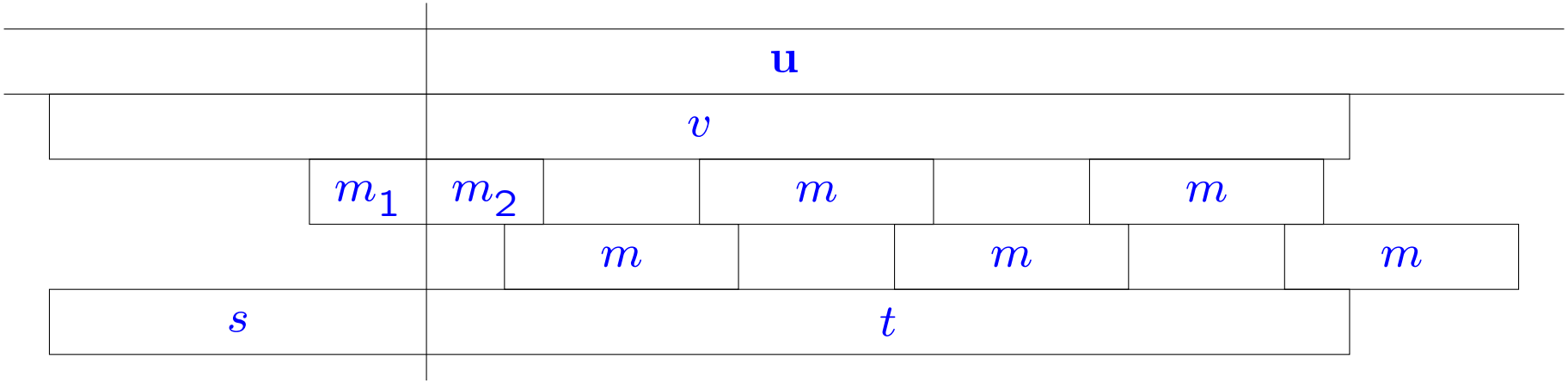
S and T are slender

Fix $\ell \geq 2D$.
Any $t \in T \cap A^\ell$ was obtained using a marker m.
As $m \in M_k$ with $2^{k-1} \leq \ell < D2^{k+1}$,
k may take one of at most $\lceil 2 + \log_2 D \rceil$ values,
so there are at most $R\lceil 2 + \log_2 D \rceil$ possible markers.

It now remains to prove that a particular marker m contributes
a bounded set $T_{m,\ell}$ to $T \cap A^\ell$ (and similarly to $S \cap A^\ell$).
Let $m = m_1 m_2$, with $|m_1| = |m_2| = 2^{k-1}$.
For each $t \in T_{m,\ell}$, we consider one particular occurrence
of a factor $v \in F(\mathbf{u})$ that was cut at an occurrence of m
and resulted in a decomposition $v = st$.
We distinguish cases 1 and 2 (and start with the easier case 2).

Case 2

In case 2, every position j where m occurs in v is such that there is also an occurrence of m at position $i + j + \pi$ in $\mathbf{u}$
(i.e. at position $j + \pi$ in v if it fits).
Therefore $m_1 t$ is periodic with period π.
There is only one such word of length $2^{k-1} + \ell$.



Case 1

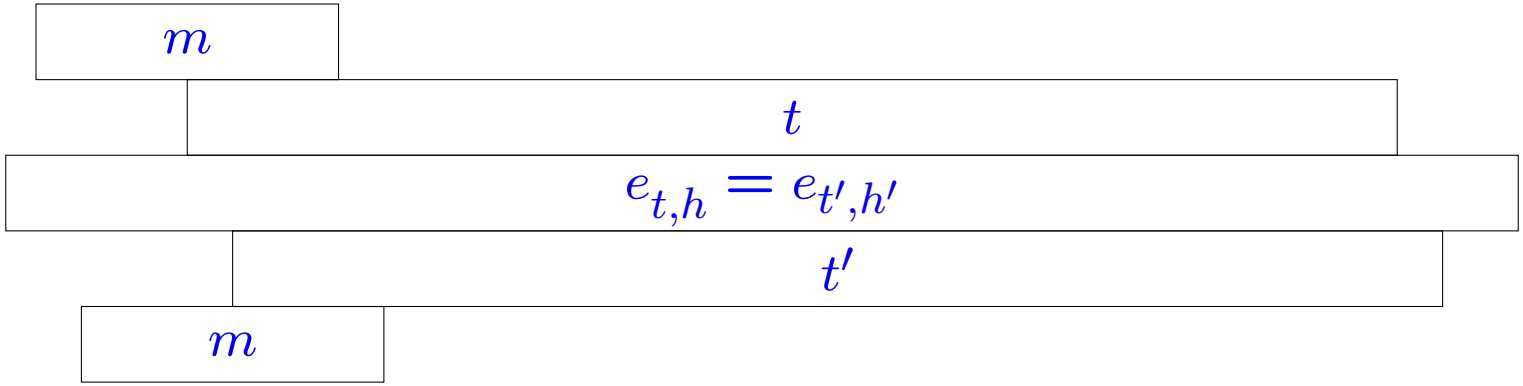
In case 1, we have chosen a position j where m occurs in v such that there is no occurrence of m at position $i+j+\pi$ in $\mathbf{u}$ (final occurrence).

For each integer h, $0 \leq h < 2^{k-1}$, consider the factor $e_{t,h}$ of $\mathbf{u}$ of length $\ell + 2^k$ starting at position $i + j - h$
(if this is negative, extend $\mathbf{u}$ to the left with a new letter z).

If we prove that all $e_{t,h}$ are distinct,
then $2^{k-1}(\#T_{m,\ell} - 1) \leq \#\{e_{t,h}\} \leq p_{\mathbf{u}}(\ell + 2^k) + 2^{k-1} - 1$
so that $\#T_{m,\ell} \leq 1 + C(\ell + 2^k)2^{1-k} + 1 < 4CD + 2C + 2$.

All $e_{t,h}$ are distinct

Assume that $e_{t,h} = e_{t',h'}$.
If $h = h'$, obviously also $t = t'$. Assume $h \neq h'$.



Observe that m has period $|h' - h| < 2^{k-1} = |m|/2$.
Then $|h' - h|$ is a multiple of π,
but then one of the occurrences of m is not final.

Quadratic complexity

Question:
Is it true in general that $\mathcal{W}_k = \mathcal{P}_{k-1}$?

The answer is no for $k = 3$.

Indeed, consider the word $\mathbf{u} = ababbabbbabbbbabbbbb\ldots$.
It is well-known that $\mathbf{u} \in \mathcal{P}_2$.
However, $\mathbf{u} \notin \mathcal{W}_3$, but $\mathbf{u} \in \mathcal{W}_4$.
(The proof of both facts is very technical, and omitted here.)

Higher complexity

If factor complexity is higher than quadratic, it is even worse.

Theorem.
For any unbounded nondecreasing positive integer function f, there exists an infinite word $\mathbf{u}$ such that $p_{\mathbf{u}}(n) = O(n^2 f(n))$ and $\mathbf{u} \notin \mathcal{W}_k$ for any k.

We can assume that $f(n) \leq n$. Let

$$\mathbf{u} = \prod_{p=1}^{\infty} \prod_{q=1}^{f(p)} (a^p b^q)^p.$$

$\mathbf{u} \notin \mathcal{W}_k$

Assume $F(\mathbf{u}) \subseteq S_1 \ldots . S_k$, where $S = S_1 \cup \ldots \cup S_k$ is slender, $p_S(n) \leq C$.
For every (p, q) such that $2k - 1 \leq p \leq \frac{n-3}{4k-2}$ and $q \leq f(p)$, there exists $s_{p,q} \in S$ of length less than n that contains $ba^p b^q a$ as a factor.
All the words $s_{p,q}$ are distinct, so their total number is

$$\sum_{p=2k-1}^{\lfloor \frac{n-3}{4k-2} \rfloor} f(p)$$

which is not bounded by Cn, a contradiction.

Complexity of $\mathbf{u}$

Factors of $\mathbf{u}$ are of the following types:
1. factors of the form a^i, b^j, a^ib^j, b^ja^k, $a^ib^ja^k$, $b^ja^kb^\ell$;
2. other factors of $(a^pb^q)^\omega$;
3. factors containing $ab^qa^pb^{q+1}$;
4. factors containing $ba^pb^{f(p)}a^{p+1}$ or $bba^{p+1}ba$.

The number of factors of type 1 is $O(n^2)$.
Factors of type 2 are determined by the value of p, the first occurrence of ab, both at most n, and the value of q, at most $f(n)$.
Factors of type 3 are determined by the value of p, the first occurrence of b^{q+1}, and the value of q.
Factors of type 4 are determined by the value of p and the first occurrence of a^{p+1}.
Therefore $p_{\mathbf{u}}(n) = O(n^2f(n))$.

Open problems

What is the largest real α_k such that $\mathcal{P}_{\alpha_k} \subseteq \mathcal{W}_k$?

What is the minimal possible complexity of a word not in any $\mathcal{W}_k$?

What is the minimal possible complexity of a uniformly recurrent word not in any $\mathcal{W}_k$?

Are all morphic words in some $\mathcal{W}_k$?