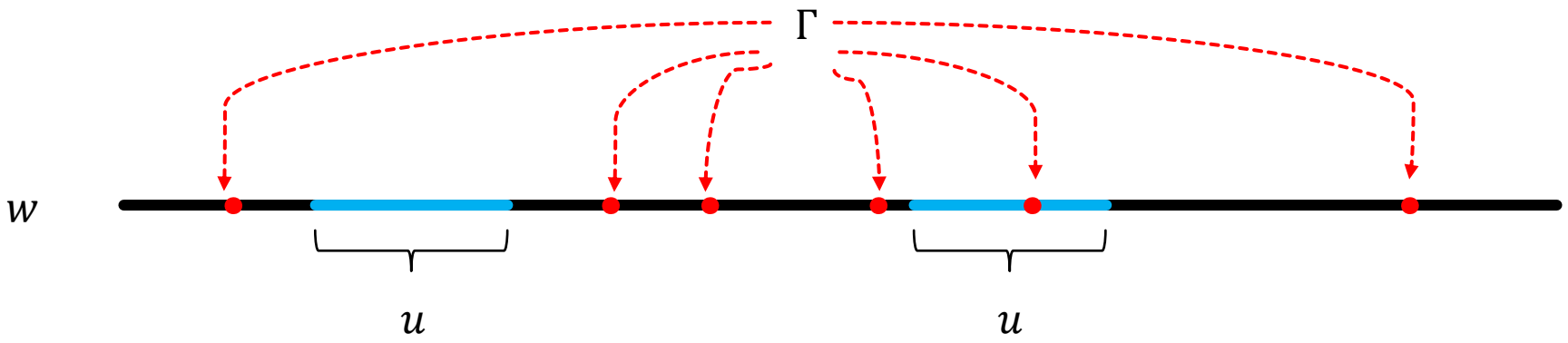
String Attractor:
from data compression to Combinatorics on words

Giuseppe Romana – Università degli Studi di Palermo

One World Combinatorics on Words Seminar – 17/10/2022

String Attractor

- **Definition** [Kempa and Prezza, STOC 2018]
- A **string attractor** Γ of a word $w \in \Sigma^n$ is a set of γ positions such that every distinct factor of w has at least an occurrence ***crossing*** a position in Γ.



String Attractor: example

- w = 1 2 3 4 5 6 7 8 9 10 11 12
  a d c **b** a **a** d **c** b a **d** c
- $\Gamma = \{1, 4, 6, 8, 11\}$
- **Note**: $\Gamma^* = \{4, 6, 8, 11\}$ is a string attractor too

List of all factors with occurrences not crossing a position in Γ

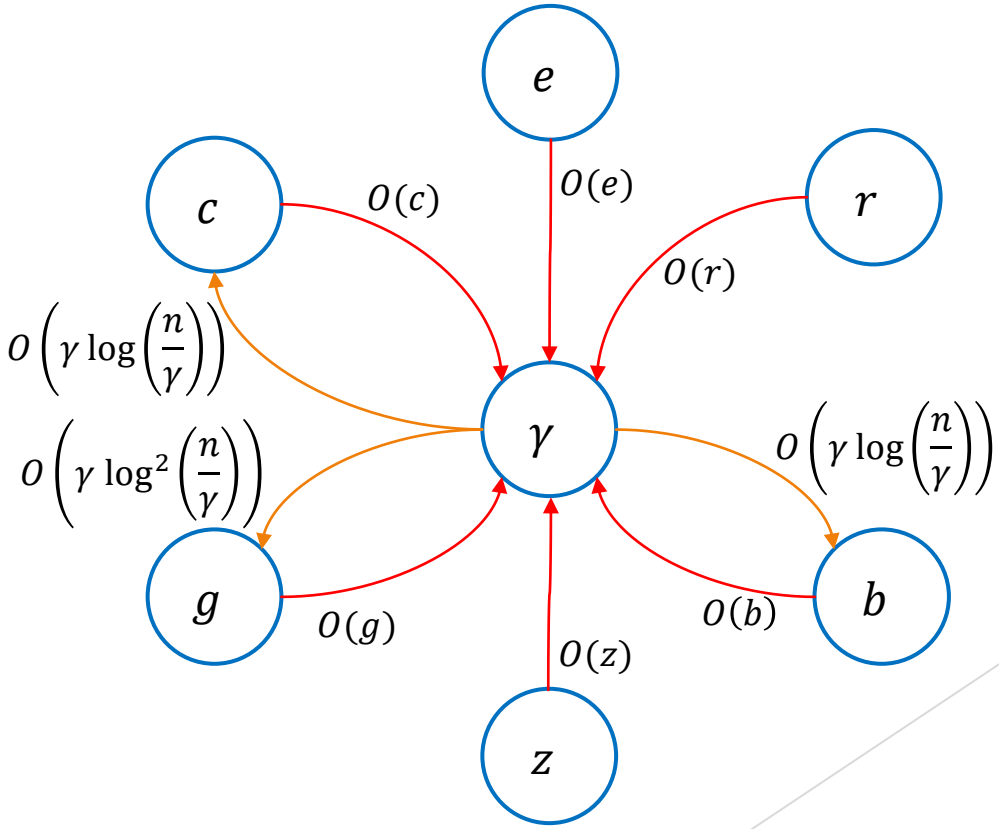
- d
- c
- a
- b
- dc
- ba

- $|\Gamma^*|$ is minimum, since any string attractor must have size $\gamma^* \geq |\Sigma|$
- We denote by γ^* the size of a string attractor of **minimum size**
- Computing the size γ^* for a word w is an NP-complete problem [Kempa and Prezza, STOC 2018]

String Attractor and Data Compression

- String attractors can be considered as unifying frameworks for different compression schemes based on repetitions [Kempa and Prezza, STOC 2018]

- z: LZ77-parse size
- r: BWT runs
- g: SLP size
- b: macro scheme size
- c: collage system size
- e: CDAWG size



String Attractor and repetitiveness

- Dictionary-based compressors exploit **repetitions** to compress and index data
  - Efficient on highly repetitive datasets (DNA sequences, astronomical observations, ...)
  - Relationships among compression schemes and repetitiveness measures of current interest [Navarro, ACM Comput. Surv. 2021]

- In recent works ([Kociumaka et al., LATIN 2020], [Kociumaka et al., LATIN 2022], …), it has been investigated the $\boldsymbol{\delta}$-*measure* of a finite word w

$$\delta(w) = \max_{1 \le k \le |w|} \frac{\left|F(w) \cap \Sigma^k\right|}{k}$$

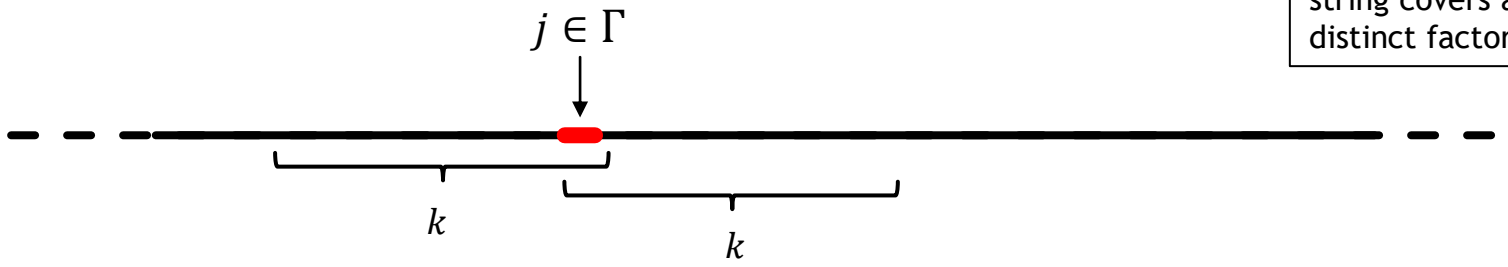
where $F(w)$ denotes the set of all distinct factors in the word w

String Attractor: some lower bounds

- **Proposition** [Kempa and Prezza, STOC 2018]
- Let Γ be a string attractor for a word w of size γ. Then

$$\delta(w) \leq \gamma(w)$$

Note: any position in a string covers at most k distinct factors of length k



- In [Kempa and Prezza, STOC 2018] it is also defined another lower bound of γ^* related to the length l of the **longest repeated factor**:

$$\gamma^*(w) \geq \frac{|w| - l}{l + 1}$$

String Attractors and Finite Words: Combinatorial properties

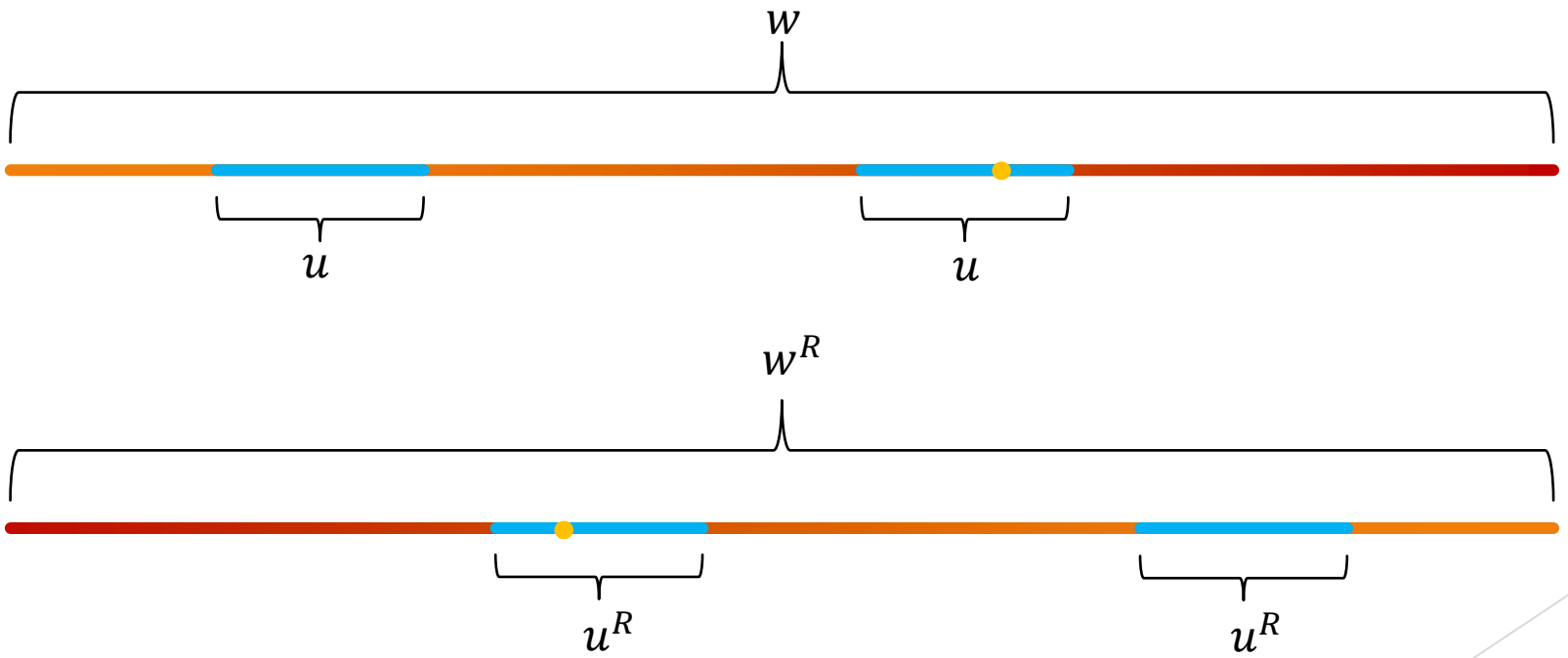
A combinatorial view on String Attractor

Mantaci, Restivo, **R**, Rosone, Sciortino

Theoret. Comput. Sci. 2021

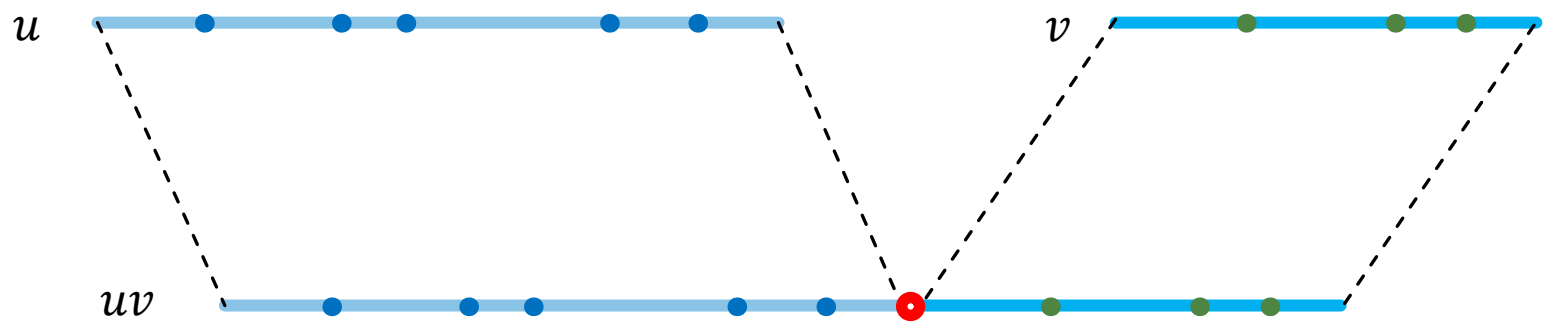
String Attractor on reverse of a word

- **Proposition**
- Let w^R be the reverse of the finite word w. Then, $\gamma^*(w) = \gamma^*(w^R)$.



String Attractor on concatenation: upper-bound

- **Proposition**
- Let u and v be two finite words and $\gamma^*(u)$ and $\gamma^*(v)$ the sizes of their respective smallest string attractor. Then, $\gamma^*(uv) \leq \gamma^*(u) + \gamma^*(v) + 1$.

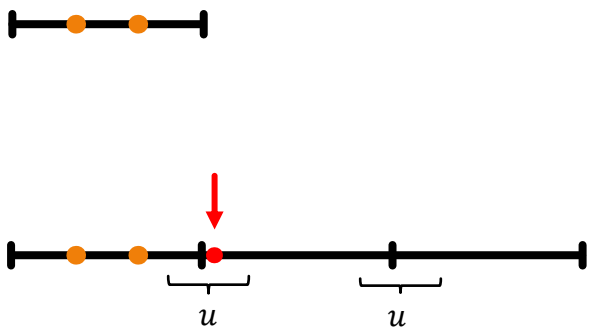


$$\gamma^*(uv) \leq |\Gamma_{uv}| = |\Gamma^*(u) \cup \{|u| + 1\} \cup \{p + |u| \text{ for every } p \in \Gamma^*(v)\}|$$

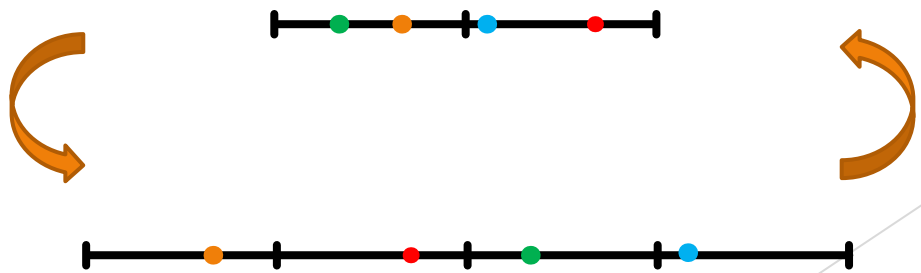
String attractor on powers ≥2

- **Proposition**
- Let w be a word over the alphabet Σ. Then, for every $n \geq 2$:
  1. $\gamma^*(w^n) \leq \gamma^*(w) + 1$
  2. $\gamma^*(w^n) = \gamma^*(w^2)$

1.



2.



Monotonicity of γ^*

- **Problem**: In [Kociumaka et al., LATIN 2020] the authors posed the question whether or not the measure γ^* of the smallest string attractor for a word w is monotonic

- In other terms, is $\gamma^*(w) \leq \gamma^*(wu)$ for all words w and u?

- **Answer**: The measure γ^* is not monotone

$\boldsymbol{w}$ = 1 2 3 4 5 6 7 8
a b b b a a a b

$\gamma^*(w) = 3$

$\boldsymbol{w \cdot b}$ = 1 2 3 4 5 6 7 8 9 10
a b b b a a a b b b

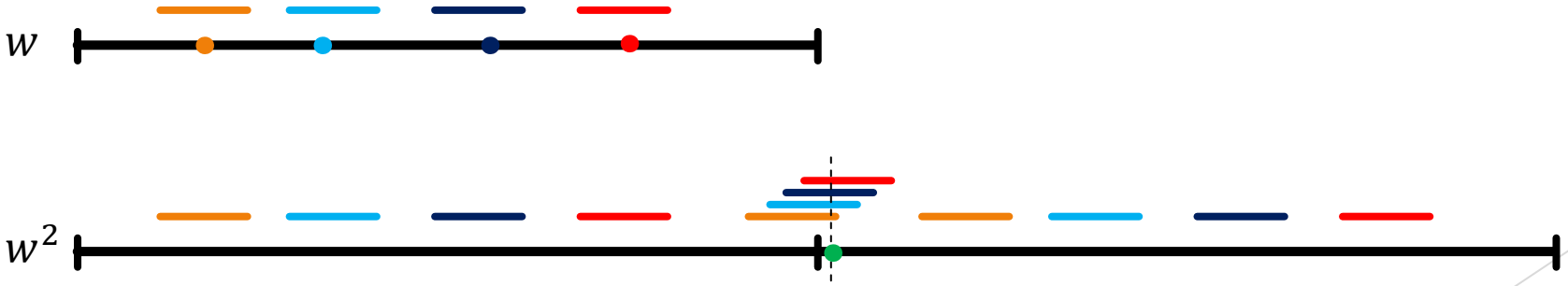
$\gamma^*(w \cdot \boldsymbol{b}) = 2$

Is $\gamma^*(w^n) \geq \gamma^*(w)$?

- **Question**: does monotonicity of γ^* holds from w to w^n?

- **Answer:** Monotonicity does not hold for power of a word

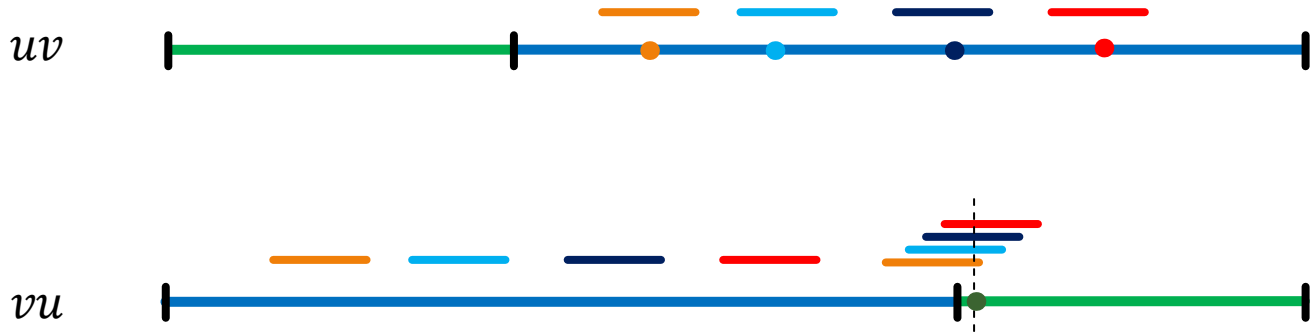
- **Proposition**
- For each $t > 0$, there exists an alphabet Σ_t and a word $w_t \in \Sigma_t^*$ such that $\gamma^*(w_t) - \gamma^*(w_t^n) > t$, for each $n \geq 2$



String attractors of conjugate words

- A similar result can be deduced for the minimum string attractors of two **conjugate words**

- **Corollary**
- For each $t > 0$, there exists an alphabet Σ_t and a word $w_t = uv \in \Sigma_t^*$ such that $\gamma^*(uv) - \gamma^*(vu) > t$



String Attractor and Infinite Words

String Attractor and Infinite Words
Restivo, **R**, Sciortino
LATIN 2022 (to appear)

Characteristic Sturmian Words

- Let $q_0, q_1, \dots, q_n, \dots$ be any sequence of natural integers such that $q_0 \geq 0$ and $q_i > 0$ $(i = 1, \dots, n, \dots)$, called **directive sequence.**

- The sequence $\{s_n\}_{n\geq 0}$ can be defined inductively as follows:
  - $s_0 = b$
  - $s_1 = a$
  - $s_{i+1} = (s_i)^{q_{i-1}} \cdot s_{i-1}$, for any $i \geq 1$

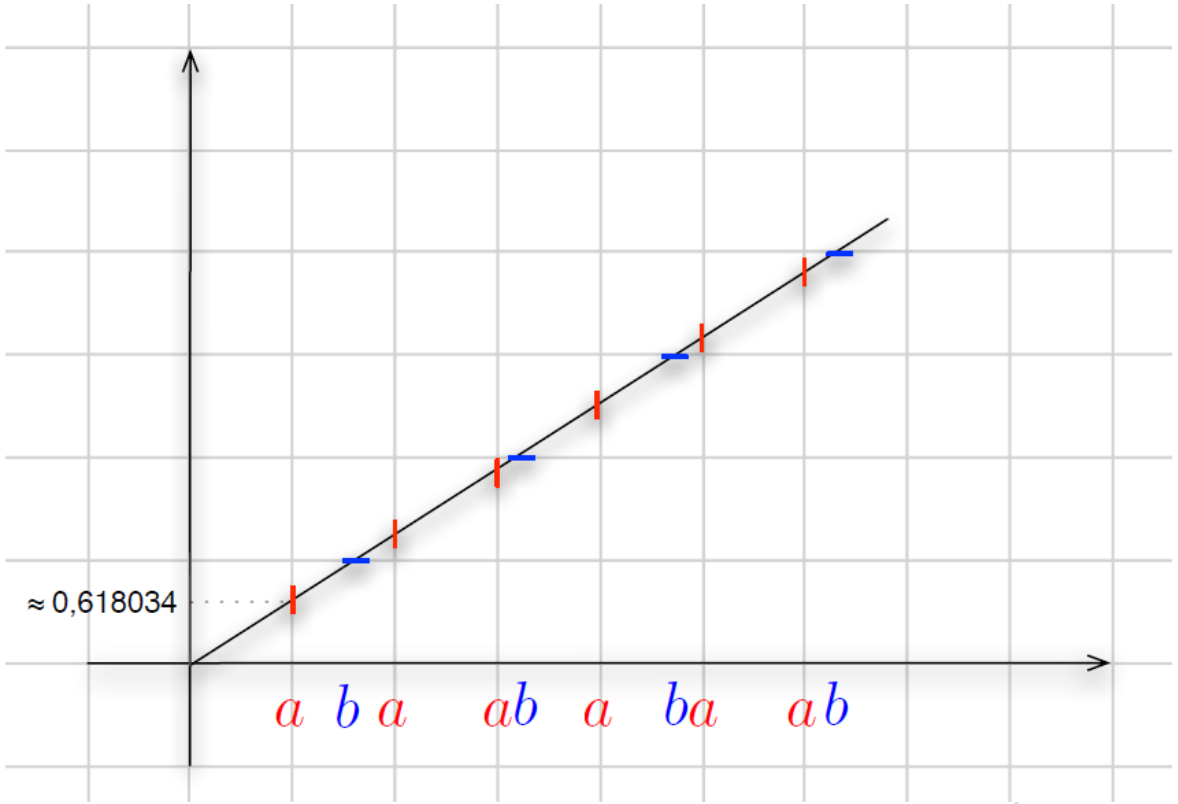
- All words s_n obtained from any directive sequence of integers are called **standard Sturmian words.**

- The infinite word $\lim_{i\to\infty} s_i$ is called **characteristic Sturmian word.**

Example: Fibonacci word

- Given the directive sequence 1,1,1,1,… consider the corresponding standard Sturmian words

- $s_0 = b$
- $s_1 = a$
- $s_2 = ab$
- $s_3 = aba$
- $s_4 = abaab$
- $s_5 = abaababa$
- …



String Attractor for standard Sturmian words

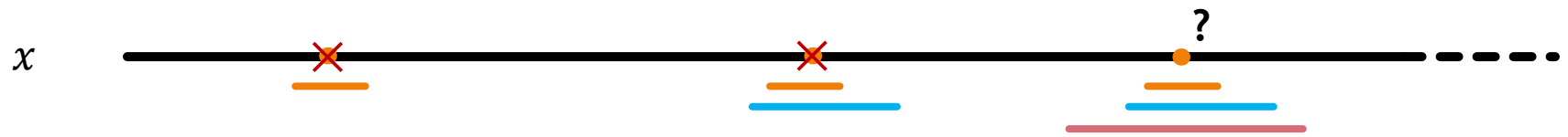
- **Theorem**

- For each standard Sturmian word s with $|s| \geq 2$, let η be the length of the longest palindromic proper prefix of $s[1, |s| - 2]$.
- Then, the set $\Gamma_1 = \{\eta + 1,\ \eta + 2\}$ or the set $\Gamma_2 = \{|s| - \eta - 3,\ |w| - \eta - 2\}$ is a smallest string attractor for s.

…	…
$\eta = 1$	$s_3 = a\boldsymbol{ba}ab$
$\eta = 3$	$s_4 = aba\boldsymbol{ab}aba$
$\eta = 6$	$s_5 = abaaba\boldsymbol{ba}abaab$
$\eta = 11$	$s_6 = abaababaaba\boldsymbol{ab}abaababa$
…	…

String Attractor profile function

- The notion of string attractor is not immediately extendible to infinite words



- On the other hand, standard Sturmian words can be seen as sequences of prefixes of characteristic Sturmian words

- **Definition** [Schaeffer & Shallit, arXiv 2021]
- Given an infinite word x, the **string attractor profile** function s_x is defined as follows

$$s_x(n) = \gamma^*(x[0..n-1])$$

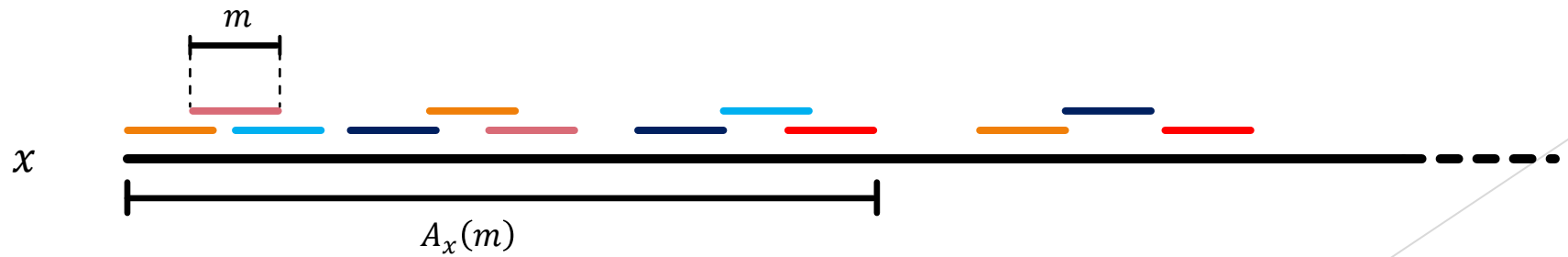
Factor complexity & Appearance function

- Let x be an infinite word

- **Factor complexity function p_x:** for each length m, it counts the number of distinct factors of length m that occur in x

$$p_x(m) = |F(x) \cap \Sigma^m|$$

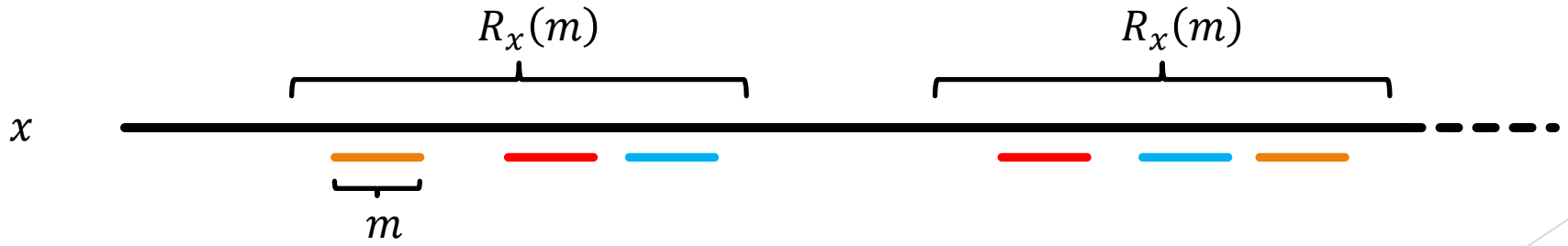
- **Appearance function A_x:** for each length m, it returns the length of the shortest prefix of x which contains all factors of x of length m



Recurrence of a word

- Let x be an infinite word
- x is called **recurrent** if every factor of x occurs infinitely often

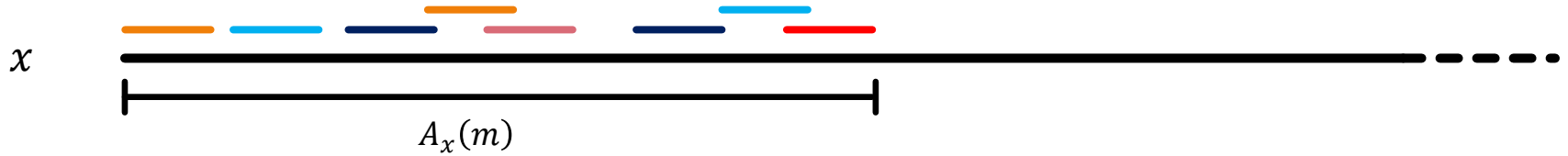
- x is called **uniformly recurrent** if there exists a function $R_x(m)$ (the **recurrence function**) such that every factor of x of length $R_x(m)$ contains at least an occurrence of each factor of x of length m



- Moreover, if $R_x(m)$ is linear, x is called **linearly recurrent**

Relationship between s_x and p_x

- **Theorem**
- Let x be an infinite word.
- For all $m > 0$, one has $p_x(m) \leq m \cdot s_x(A_x(m))$



- **Corollary**
- If there exists k such that $s_x(n) < k$ for each $n > 0$, then $p_x(n) \leq n \cdot k$
  - In other words, if s_x is bounded by a constant, then x has **at most** linear factor complexity

String Attractors for uniformly recurrent words

- **Theorem** [Schaeffer and Shallit, arXiv 2021]
- Let x be an infinite word.
- If x is linearly recurrent, then, $s_x(n) = O(1)$.

$$\begin{cases} p_x(m) = O(m) \\ R_x(m) = O(m) \end{cases}$$

- However, not all infinite words with a constant bound on the function s_x are linearly recurrent
  - Every Sturmian word is uniformly recurrent, but not all are linearly recurrent

- **Open question**: Let x be a uniformly recurrent word such that p_x is linear. Is $s_x(n)$ bounded by a constant value?

String attractor profile function for infinite words

x infinite word		Construction	Recurrent	$p_x(n)$	$s_x(n)$	
s Ch. Sturmian	$abaababaabaababaababa\cdots$	- dir. seq. $d_0, d_1, d_2, \ldots$ - $s_0 = b$, $s_1 = a$ - $s_{i+1} = s_i^{d_{i-1}} s_{i-1}$	Uniformly	$\Theta(n)$	2	[Restivo et al., LATIN 2022]
pd Period-doubling	$10111010101110111011110\cdots$	$\rho: \begin{cases} 1 \mapsto 10 \\ 0 \mapsto 11 \end{cases}$	Linearly	$\Theta(n)$	2	[Schaeffer and Shallit, arXiv 2021]
t Thue-Morse	$01101001100101101001011\cdots$	$\tau: \begin{cases} 1 \mapsto 10 \\ 0 \mapsto 01 \end{cases}$	Linearly	$\Theta(n)$	4	[Kutsukake et al., SPIRE 2020] [Schaeffer and Shallit, arXiv 2021]
z (5,3)-Toeplitz	$12121122111222112112121\cdots$	$12???$	Uniformly	$\Theta\left(n^{\frac{\log 5}{\log 5 - \log 3}}\right)$	unbounded	[Restivo et al., LATIN 2022]
c Powers of 2	$11010001000000010000000\cdots$	- $c[i] = 1$ if $i = 2^k$ - $c[i] = 0$ otherwise	No	$\Theta(n)$	$\Theta(\log n)$	[Kociumaka et al., LATIN 2020] [Schaeffer and Shallit, arXiv 2021]

On ultimately periodic and ω-power free words

- **Proposition**
- Let x be an infinite word.
- If x is ultimately periodic, then $s_x(n) = \Theta(1)$.
  - $s_x(n) = \gamma^*(x[0, n-1]) = \gamma^*(uv^\ell) \leq \gamma^*(u) + \gamma^*(v^\ell) + 1 \leq \gamma^*(u) + \gamma^*(v) + 2$

- **Proposition**
- Let x be an infinite word.
- If $s_x(n) = \Theta(1)$, then either x is ultimately periodic or ω-power free.

For each $u \in F(x)$ there exists k such that $u^k \notin F(x)$

New String Attractor-based complexities

Words with $s_x = \Theta(1)$: examples

$(ab)^\omega =$ $ababababababababababababababababababababababababababab \ldots$

$f =$ $abaababaabaababababaabaabaababaabaababaabaababaaba \ldots$

$pd =$ $10111010101110111011101010111010101110101110111 \ldots$

$s_{(ab)^\omega}(n) = s_f(n) = s_{pd}(n)$, for all $n > 0$

Span and Leftmost string attractor

- Instead of just focusing on s_x, we have considered some structural properties of string attractors

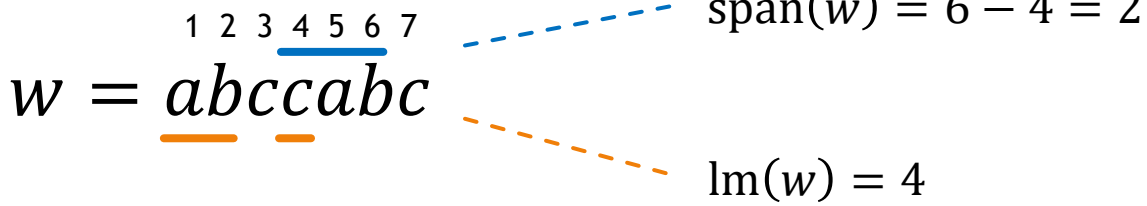
- Let $\mathcal{G}$ be the set of all suitable string attractors for a given finite word w

- **Span**

$$\mathrm{span}(w) = \min_{\Gamma \in \mathcal{G}}\{\max(\Gamma) - \min(\Gamma)\}$$

- **Leftmost string attractor**

$$\mathrm{lm}(w) = \min_{\Gamma \in \mathcal{G}}\{\max(\Gamma)\}$$



Span Complexity and Leftmost Complexity

- Analogously, we define the span and the leftmost complexity for infinite words

Span complexity

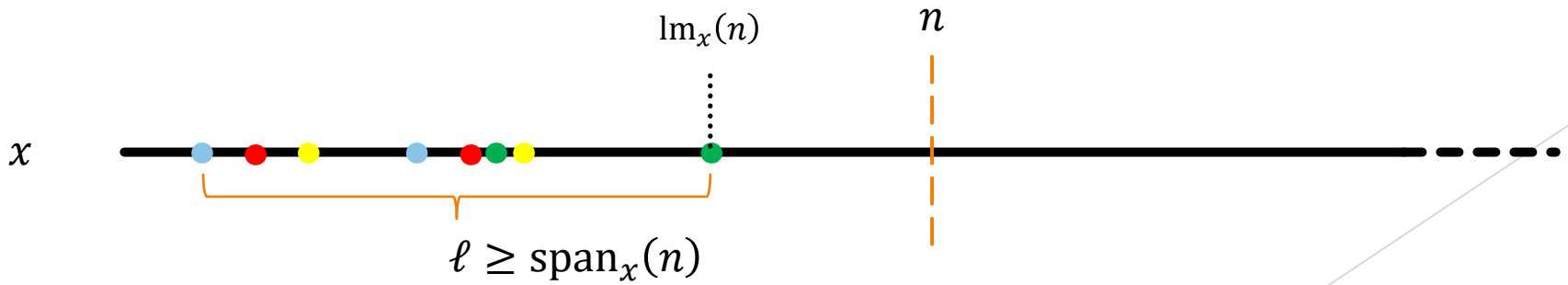
$$\mathrm{span}_x(n) = \mathrm{span}(x[0, n-1])$$

Leftmost complexity

$$\mathrm{lm}_x(n) = \mathrm{lm}(x[0, n-1])$$

Proposition

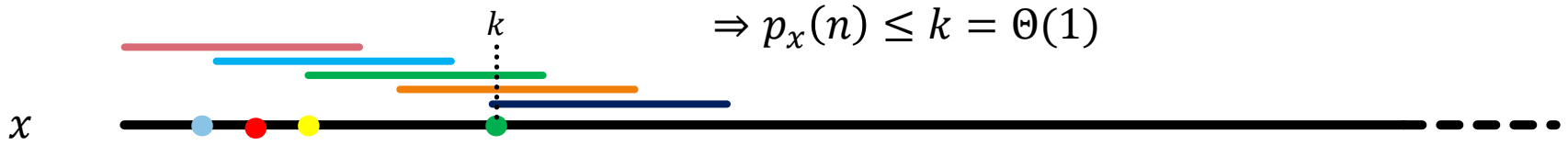
- Let x be an infinite word. Then $s_x(n) - 1 \leq \mathrm{span}_x(n) \leq \mathrm{lm}_x(n)$.



Characterization of ultimately periodic words

- **Proposition**

- Let x be an infinite word.
- x is ultimately periodic if and only if there exists $k > 0$ such that $\mathrm{lm}_x(n) \leq k$, for infinitely many $n > 0$.

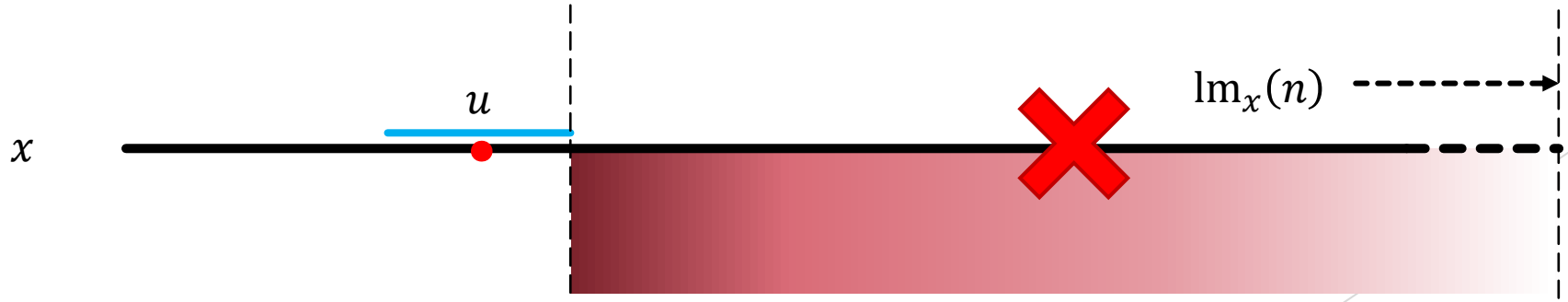


Span Complexity bounded by a constant

- **Proposition**
- Let x be an infinite word.
- If there exists $k > 0$ such that $\mathrm{span}_x(n) \leq k$ for infinitely many n, then x is recurrent or ultimately periodic.

  - x ultimately periodic ⇒ exists k such that $\mathrm{lm}_x(n) \leq k$ for each n (recall $\mathrm{span}_x(n) \leq \mathrm{lm}_x(n)$)
  - On the other hand, x aperiodic ⇒ for each $k > 0$ there exists n_0 such that $\mathrm{lm}_x(n) > k$, for all $n > n_0$.

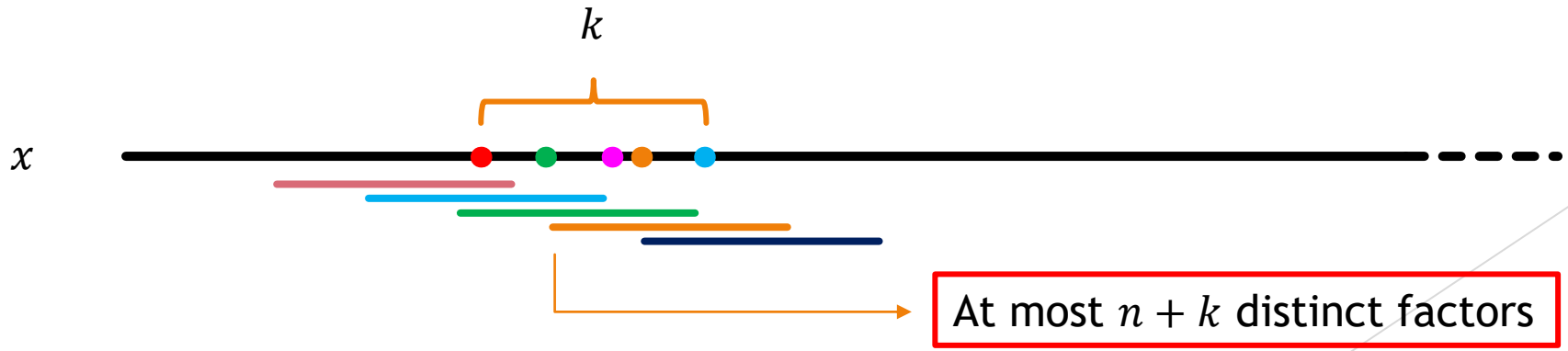
    Let us suppose x is not recurrent, i.e. exists $u \in F(x)$ that occurs only once



Relation between Span complexity and Factor complexity

- Actually, if the span complexity of an infinite word x is bounded by a constant k, a stronger result can be deduced

- **Lemma**
- Let w be a finite word.
- Then, for all $0 < n \leq |w|$, it holds that $|F(w) \cap \Sigma^n| \leq n + \mathrm{span}(w)$



Span complexity: a new characterization for Sturmian words (1)

- It is known that an infinite word is Sturmian iff $p_x(m) = m + 1$ for all m

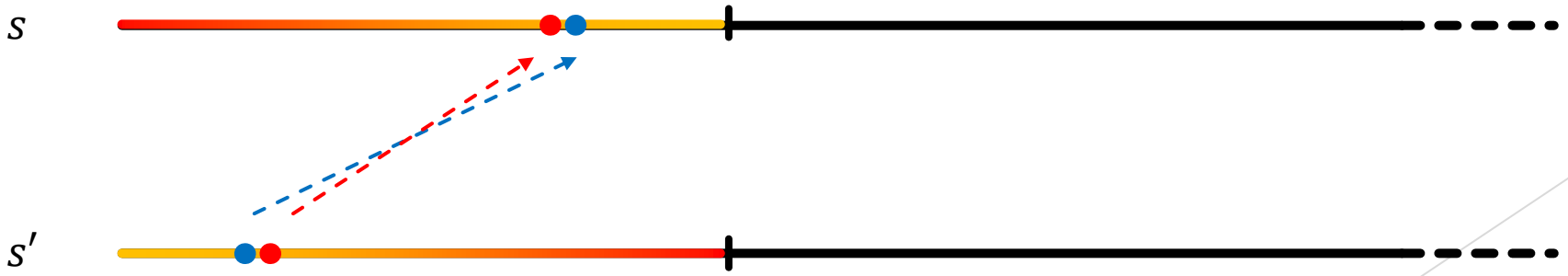
- **Theorem**
- Let x be an infinite aperiodic word.
- Then, x is Sturmian if and only if $\mathrm{span}_x(n) = 1$ for infinitely many $n > 0$.

- ($\Leftarrow$)
- x aperiodic $\Rightarrow p_x(m) \geq m + 1$
- $\mathrm{span}_x(n) = 1$ for infinitely many $n > 0 \Rightarrow p_x(m) \leq m + 1$
- Thus, $p_x(m) = m + 1$ and x is Sturmian

Span complexity: a new characterization for Sturmian words (2)

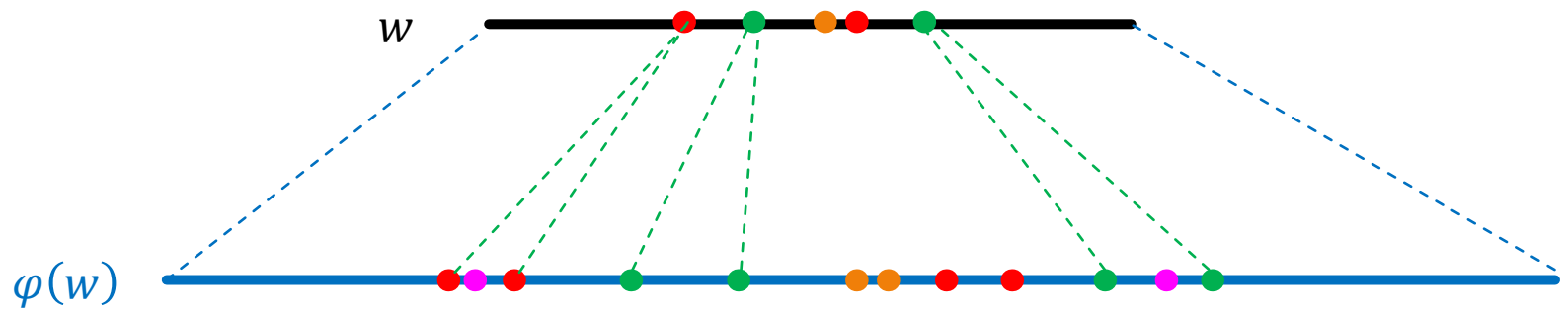
- ($\Rightarrow$)
- By using combinatorial arguments, we can prove that exists n_0 such that, for every characteristic Sturmian word s', $span_{s'}(n) = 1$ for every $n \geq n_0$
- Every Sturmian word, like every other **aperiodic AND recurrent** word, has an infinite number of **right special factors** as prefixes
- Further, for every right special factor u of a Sturmian word there exists a characteristic Sturmian word s' that has u^R as prefix [Lothaire - Algebraic Combinatorics on words, 2002]

$u \in F(x)$ is a right special factor if exist $a \neq b \in \Sigma$ such that $ua, ub \in F(x)$



Morphisms and string attractor-based measures

- **Proposition**
- Let $\varphi: \Sigma \mapsto \Sigma'$ be a morphism. Then there exists $K > 0$ which depends only from φ such that, for every $w \in \Sigma^*$:
  - $\gamma^*(\varphi(w)) \leq 2\gamma^*(w) + K$
  - $\mathrm{span}(\varphi(w)) \leq K \cdot \mathrm{span}(w)$
  - $\mathrm{lm}(\varphi(w)) \leq K \cdot \mathrm{lm}(w)$



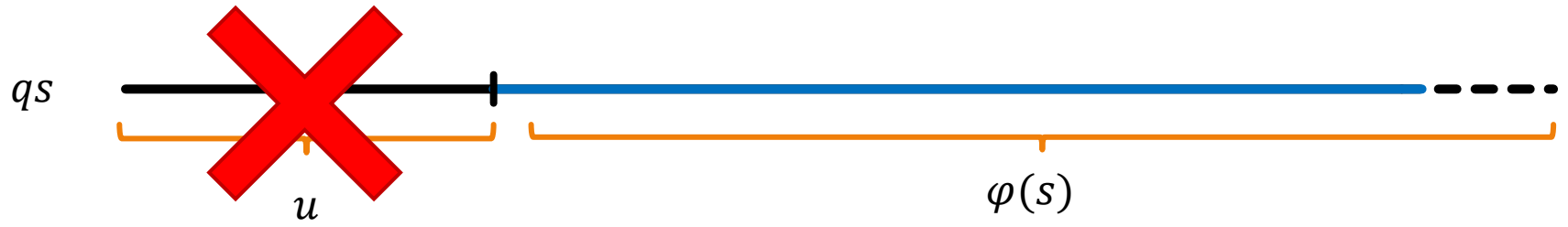
Quasi-Sturmian words

- $x \in \Sigma^\omega$ is Quasi-Sturmian if there exist n_0, k such that $p_x(n) = n + k$, for $n > n_0$

- **Proposition** [Cassaigne, DLT 1997]
- An infinite word $x \in \Sigma^\omega$ is quasi-Sturmian if and only if $x = u \cdot \varphi(s)$, where
  - $u \in \Sigma^*$ is a finite word
  - $s \in \{a, b\}^\omega$ is a Sturmian word
  - $\varphi: \{a, b\}^* \mapsto \Sigma^*$ is a morphism such that $\varphi(ab) \neq \varphi(ba)$

Characterization of Quasi-Sturmian words via span complexity

- **Theorem**
- An infinite aperiodic word $x \in \Sigma^\omega$ is Quasi-Sturmian if and only if there exist a suffix y of x and an integer $k > 0$ such that $\mathrm{span}_y(n) \leq k$ for infinitely many n.



- For every Sturmian s it holds that $\mathrm{span}_s(n) = 1$ for infinitely many n
- Further, there exists $k > 0$ such that $\mathrm{span}\big(\varphi(w)\big) \leq k \cdot \mathrm{span}(w)$ for every $w \in \Sigma^*$
- Thus, $\mathrm{span}_{\varphi(s)}(n) \leq k \cdot \mathrm{span}_s(n) = k$ for infinitely many n

Conclusions and open problems

- Notion of string attractor in between data compression and combinatorics
  - An NP-complete problem is solvable for infinite families of words by using combinatorial arguments
  - Chacterization of words via string attractor based complexities

- **Open question**: Are there other structural properties of string attractors that can be used to characterize infinite words?

- **Open question**: In general, for every finite word w, is there a set of string attractors that allow to uniquely recover w?

Thanks for your attention!