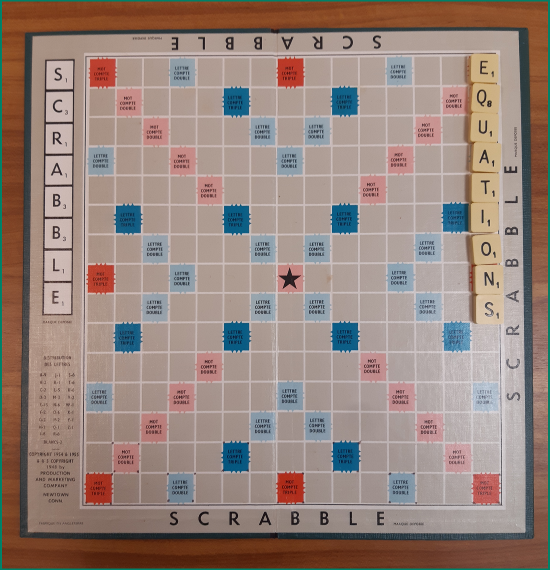
SCRABBLE: String Correlations Recurrences Analysis By Border's Language Equations

Julien (Clément | Courtiel), Matthieu Dien, Pablo Rotondo

December 8, 2021

Goal of this talk



Outline

1 Borders of words

2 Generating functions and Mahler equations

3 The (soon to be a theorems) conjectures

Definition

Definition

Let $w = w_0 \dots w_{n-1}$ be a word over an alphabet $\Sigma = \{a, c, b, d, e, \dots\}$.

- A **border** b of w is is both **proper prefix** and **proper suffix** of w. Note that the empty word ϵ is always a border of w.
- The **maximal** border of w is the longest border of w.

For convenience, we will use the letters

Definition

Definition

Let $w = w_0 \dots w_{n-1}$ be a word over an alphabet $\Sigma = \{a, c, b, d, e, \dots\}$.

- A **border** b of w is is both **proper prefix** and **proper suffix** of w. Note that the empty word ϵ is always a border of w.
- The **maximal** border of w is the longest border of w.

For convenience, we will use the letters

Examples

$$\mathrm{Borders}(acab) =$$

Definition

Let $w = w_0 \dots w_{n-1}$ be a word over an alphabet $\Sigma = \{a, c, b, d, e, \dots\}$.

- A **border** b of w is is both **proper prefix** and **proper suffix** of w. Note that the empty word ϵ is always a border of w.
- The **maximal** border of w is the longest border of w.

For convenience, we will use the letters

Examples

$$\text{Borders}(acab) = \{\epsilon\}$$

$$\text{Borders}(abracdabra) =$$

Definition

Let $w = w_0 \dots w_{n-1}$ be a word over an alphabet $\Sigma = \{a, c, b, d, e, \dots\}$.

- A **border** b of w is is both **proper prefix** and **proper suffix** of w. Note that the empty word ϵ is always a border of w.
- The **maximal** border of w is the longest border of w.

For convenience, we will use the letters

Examples

$$\text{Borders}(acab) = \{\epsilon\}$$

$$\text{Borders}(abracdabra) = \{\epsilon, a, abra\}$$

$$\text{Borders}(nations) =$$

Definition

Let $w = w_0 \dots w_{n-1}$ be a word over an alphabet $\Sigma = \{a, c, b, d, e, \dots\}$.

- A **border** b of w is is both **proper prefix** and **proper suffix** of w. Note that the empty word ϵ is always a border of w.
- The **maximal** border of w is the longest border of w.

For convenience, we will use the letters

Examples

$$\text{Borders}(acab) = \{\epsilon\}$$

$$\text{Borders}(abracdabra) = \{\epsilon, a, abra\}$$

$$\text{Borders}(nations) = \{\epsilon\}$$

$$\text{Borders}(popopo) =$$

Definition

Let $w = w_0 \ldots w_{n-1}$ be a word over an alphabet $\Sigma = \{a, c, b, d, e, \ldots\}$.

- A **border** b of w is is both **proper prefix** and **proper suffix** of w. Note that the empty word ϵ is always a border of w.
- The **maximal** border of w is the longest border of w.

For convenience, we will use the letters

Examples

$$\text{Borders}(acab) = \{\epsilon\}$$
$$\text{Borders}(abracdabra) = \{\epsilon, a, abra\}$$
$$\text{Borders}(nations) = \{\epsilon\}$$
$$\text{Borders}(popopo) = \{\epsilon, po, popo\}$$

Some notations

Given a word w we note :

- $|w|$ the length of w
- $w[i:j]$ the subword $w_i \cdots w_{j-1}$
- $w[:k]$ the prefix of length k of w
- $w[-k:]$ the suffix of length k of w
- $Border(w)$ its maximal border

The borders' table: definition

Question

How to compute the borders ?

Answer: the borders' table

$$\mathrm{BTable}(w) = \{\mathrm{Border}(w[: i]) \mid i \in [0, \ldots, |w|]\}$$

Examples

$$\mathrm{BTable}(\mathit{nations}) =$$

The borders' table: definition

Question

How to compute the borders ?

Answer: the borders' table

$$\mathrm{BTable}(w) = \{\mathrm{Border}(w[:i]) \mid i \in [0, \ldots, |w|]\}$$

Examples

$$\mathrm{BTable}(\mathit{nations}) = \{0, 0, 0, 0, 0, 1, 0\}$$

$$\mathrm{BTable}(\mathit{popopo}) =$$

The borders' table: definition

Question

How to compute the borders ?

Answer: the borders' table

$$\mathrm{BTable}(w) = \{\mathrm{Border}(w[:i]) \mid i \in [0, \ldots, |w|]\}$$

Examples

$$\mathrm{BTable}(\mathit{nations}) = \{0, 0, 0, 0, 0, 1, 0\}$$
$$\mathrm{BTable}(\mathit{popopo}) = \{0, 0, 1, 2, 3, 4\}$$
$$\mathrm{BTable}(\mathit{abcdab}) =$$

The borders' table: definition

Question

How to compute the borders ?

Answer: the borders' table

$$\mathrm{BTable}(w) = \{\mathrm{Border}(w[:i]) \mid i \in [0, \dots, |w|]\}$$

Examples

$$\mathrm{BTable}(\textit{nations}) = \{0, 0, 0, 0, 0, 1, 0\}$$
$$\mathrm{BTable}(\textit{popopo}) = \{0, 0, 1, 2, 3, 4\}$$
$$\mathrm{BTable}(\textit{abcdab}) = \{0, 0, 0, 0, 1, 2\}$$
$$\mathrm{BTable}(\textit{abbbab}) =$$

The borders' table: definition

Question

How to compute the borders ?

Answer: the borders' table

$$\mathrm{BTable}(w) = \{\mathrm{Border}(w[:i]) \mid i \in [0, \ldots, |w|]\}$$

Examples

$$\mathrm{BTable}(\textit{nations}) = \{0, 0, 0, 0, 0, 1, 0\}$$
$$\mathrm{BTable}(\textit{popopo}) = \{0, 0, 1, 2, 3, 4\}$$
$$\mathrm{BTable}(\textit{abcdab}) = \{0, 0, 0, 0, 1, 2\}$$
$$\mathrm{BTable}(\textit{abbbab}) = \{0, 0, 0, 0, 1, 2\}$$

Pattern matching in a text

The MP algorithm search a pattern w in *text* in $\mathcal{O}(|text| + |w|)$ character equality tests

Example

```
       w = abaababaabaa
btable(w)= 001123234564

0    5     11   16     23
|    |     |    |      |
abaababaababaabacabaabaababaabaa
abaababaabaa
     abaababaabaa
          abaabab
             abaa
               ab
                a
                 abaabab
                    abaababaabaa
```

Canonical words

Reminder

$$\mathrm{BTable}(abcdab) =$$

Canonical words

Reminder

$$\mathrm{BTable}(abcdab) = \{0, 0, 0, 0, 1, 2\}$$

$$\mathrm{BTable}(abbbab) =$$

Canonical words

Reminder

$$\mathrm{BTable}(abcdab) = \{0, 0, 0, 0, 1, 2\}$$
$$\mathrm{BTable}(abbbab) = \{0, 0, 0, 0, 1, 2\}$$

Canonical words

Reminder

$$\mathrm{BTable}(abcdab) = \{0,0,0,0,1,2\}$$
$$\mathrm{BTable}(abbbab) = \{0,0,0,0,1,2\}$$

Definition

A **canonical word** of borders' table *bt* is the smallest word wrt. the lexicographic order with this table.

Canonical words

Reminder

$$\text{BTable}(abcdab) = \{0,0,0,0,1,2\}$$
$$\text{BTable}(abbbab) = \{0,0,0,0,1,2\}$$

Definition

A **canonical word** of borders' table *bt* is the smallest word wrt. the lexicographic order with this table.

Questions

- Given a finite alphabet, how many canonical words of length n are they ?
- What's the shape of a *big* canonical word ?
- What does look like the language of canonical words ? (rational, algebraic, etc.)

Warm-up: binary canonical word

Examples

$$\mathrm{BTable}(aba) = \{0, 0, 1\}$$

Warm-up: binary canonical word

Examples

$$\mathrm{BTable}(aba) = \{0, 0, 1\}$$
$$\mathrm{BTable}(bab) = \{0, 0, 1\}$$

Warm-up: binary canonical word

Examples

$$\text{BTable}(\textit{aba}) = \{0, 0, 1\}$$
$$\text{BTable}(\textit{bab}) = \{0, 0, 1\}$$
$$\text{BTable}(\textit{baabaa}) = \{0, 0, 0, 1, 2, 3\}$$

Warm-up: binary canonical word

Examples

$$\mathrm{BTable}(aba) = \{0, 0, 1\}$$
$$\mathrm{BTable}(bab) = \{0, 0, 1\}$$
$$\mathrm{BTable}(baabaa) = \{0, 0, 0, 1, 2, 3\}$$
$$\mathrm{BTable}(abbabb) = \{0, 0, 0, 1, 2, 3\}$$

Warm-up: binary canonical word

Examples

$$\mathrm{BTable}(aba) = \{0, 0, 1\}$$
$$\mathrm{BTable}(bab) = \{0, 0, 1\}$$
$$\mathrm{BTable}(baabaa) = \{0, 0, 0, 1, 2, 3\}$$
$$\mathrm{BTable}(abbabb) = \{0, 0, 0, 1, 2, 3\}$$

Answers

- Binary canonical words are the binary words beginning by a a and so, their language is rational.
- There are 2^{n-1} binary canonical words of length n.

Heels to buttocks: (almost) canonical words with 3 letters (1)

Definitions

- $\Sigma = \{a, b\}$
- $\mathcal{D}_3$: canonical words over $\Sigma \cup \{c\}$ and such toverline the c only appears in last position
  $\mathcal{D}_3 = \{abac, abaac, abbac, abaaac, ababac, abbaac, abbbac, abaac, \ldots\}$
- $\overline{\mathcal{D}_3}c = \mathcal{D}_3$: the words of $\mathcal{D}_3$ without the c

Idea

A word of $\mathcal{D}_3$ is like

$$\textit{bigBorder } b \textit{ whatEver bigBorder } c.$$

Let's decompose them !

Heels to buttocks: (almost) canonical words with 3 letters (2)

Result

$$\sum_{u\in\Sigma^*} aub\Sigma^* auc = \overline{\mathcal{D}_3}c + \sum_{w\in\overline{\mathcal{D}_3}} wb\Sigma^* wc.$$

Each word of $\mathcal{D}_3$ can be decomposed in several ways.

Example

$$\begin{aligned} a\ ba\ b\ abbaabab\ a\ ba\ c &= abababbaabababa\ c \\ &= aba\ b\ abbaabab\ aba\ c \\ &= ababa\ b\ baab\ ababa\ c \end{aligned}$$

Generating series: a reminder

Definition

Given a language $\mathcal{L}$, its generating series $L(z)$ is defined as

$$L(z) = \sum_{w \in \mathcal{L}} z^{|w|} = \sum_{n \geq 0} \sharp\{w \in \mathcal{L} \mid |w| = n\} z^n.$$

Regular expressions to generating series

$$\begin{aligned}
a, b, c, \ldots \in \Sigma &\mapsto z \\
\mathcal{L}_1 \cdot \mathcal{L}_2 &\mapsto L_1(z) \cdot L_2(z) \\
\mathcal{L}_1 \cup \mathcal{L}_2 &\mapsto L_1(z) + L_2(z) \\
\mathcal{L}^* &\mapsto \frac{1}{1 - L(z)} \\
\mathcal{L}^2 = \{ww \mid w \in \mathcal{L}\} &\mapsto L(z^2)
\end{aligned}$$

Mahler equation for $\mathcal{D}_3$ (1)

Result

The equation over the language $\overline{\mathcal{D}_3}$,

$$\sum_{u \in \Sigma^*} aub\Sigma^* auc = \overline{\mathcal{D}_3}c + \sum_{w \in \overline{\mathcal{D}_3}} wb\Sigma^* wc,$$

translates to a Mahler equation over the generating function $D_3(z) = \overline{D_3}(z) \cdot z$:

$$\frac{z^4}{(1-2z)(1-2z^2)} = D_3(z) + \frac{1}{1-2z} D_3(z^2).$$

Mahler equation for $\mathcal{D}_3$ (2)

"Explicit" formula

$$D_3(z) = z^4 + \dots$$

$$D_3(z) = \frac{z^8}{2\,z-1} + \frac{z^4}{(2\,z^2-1)(2\,z-1)} + \dots$$

$$D_3(z) = \frac{z^4}{(2\,z^2-1)(2\,z-1)} + \frac{\frac{z^{16}}{2\,z^2-1} + \frac{z^8}{(2\,z^4-1)(2\,z^2-1)}}{2\,z-1} + \dots$$

$$D_3(z) = \frac{z^4}{(2\,z^2-1)(2\,z-1)} + \frac{\frac{z^8}{(2\,z^4-1)(2\,z^2-1)} + \frac{\frac{z^{32}}{2\,z^4-1} + \frac{z^{16}}{(2\,z^8-1)(2\,z^4-1)}}{2\,z^2-1}}{2\,z-1} + \dots$$

Close formula

$$D_3(z) = \sum_{k=2}^{\infty} (-1)^k z^{2^k} \prod_{i=0}^{k-1} \frac{1}{1-2z^{2^i}}.$$

The effort : words with longest border u

Definitions

- u: a canonical word of length m over Σ ($|\Sigma| = \sigma$)
- $\mathcal{F}_u$: language of words with longest border u
- $\mathcal{C}_u = \mathrm{Borders}(u) \cup \{u\}$
- $\overline{\mathcal{C}_u}\mathcal{L} = \{w \mid v \in \mathcal{C}_u, vw \in \mathcal{L}\}$: the left quotient of language $\mathcal{L}$ by the set $\mathcal{C}_u$

Result

$$u + \mathcal{F}_u + \sum_{w \in \mathcal{F}_u} w \left(\overline{\mathcal{C}_u} + \Sigma^* \right) w = u \left(\overline{\mathcal{C}_u} + \Sigma^* \right) u.$$

The union sign $+$ denotes the disjoint union.

The effort : words with longest border u

Definitions

- u: a canonical word of length m over Σ ($|\Sigma| = \sigma$)
- $\mathcal{F}_u$: language of words with longest border u
- $\mathcal{C}_u = \mathrm{Borders}(u) \cup \{u\}$
- $\overline{\mathcal{C}_u}\mathcal{L} = \{w \mid v \in \mathcal{C}_u, vw \in \mathcal{L}\}$: the left quotient of language $\mathcal{L}$ by the set $\mathcal{C}_u$

Result

$$u + \mathcal{F}_u + \sum_{w \in \mathcal{F}_u} w\left(\overline{\mathcal{C}_u} + \Sigma^*\right) w = u\left(\overline{\mathcal{C}_u} + \Sigma^*\right) u.$$

The union sign + denotes the disjoint union. Just a reformulation of Guibas ans Odlyzko ?

The effort : words with longest border u

Definitions

- u: a canonical word of length m over Σ ($|\Sigma| = \sigma$)
- $\mathcal{F}_u$: language of words with longest border u
- $\mathcal{C}_u = \mathrm{Borders}(u) \cup \{u\}$
- $\overline{\mathcal{C}_u}\mathcal{L} = \{w \mid v \in \mathcal{C}_u, vw \in \mathcal{L}\}$: the left quotient of language $\mathcal{L}$ by the set $\mathcal{C}_u$

Result

$$u + \mathcal{F}_u + \sum_{w \in \mathcal{F}_u} w \left(\overline{\mathcal{C}_u} + \Sigma^*\right) w = u \left(\overline{\mathcal{C}_u} + \Sigma^*\right) u.$$

The union sign $+$ denotes the disjoint union. Just a reformulation of Guibas ans Odlyzko ?

What's new ?

- Equations over languages $\Rightarrow$ Generating functions !
- Now we can put weights over the letters of Σ $\Rightarrow$ a bit more general

Examples

$$m = abaabaabaa$$
$$u = \mathrm{Border}(m) = abaabaa$$
$$\mathcal{C}_u = \{\epsilon, a, abaa, abaabaa, abaabaabaa\}$$

$$u \quad \left(\overline{\mathcal{C}_u} + \Sigma^*\right) \quad u \quad = \quad u \quad + \quad \mathcal{F}_u \quad + \quad \sum_{w \in \mathcal{F}_u} w\left(\overline{\mathcal{C}_u} + \Sigma^*\right) w.$$
$$abaabaa \quad \overline{abaa} \quad abaabaa \quad = \quad abaabaabaa$$

Mahler à vous !

Generating function

$$u + \mathcal{F}_u + \sum_{w \in \mathcal{F}_u} w \left(\overline{\mathcal{C}_u} + \Sigma^* \right) w = u \left(\overline{\mathcal{C}_u} + \Sigma^* \right) u$$

$$z^m + f_u(z) + f_u(z^2) \left(\overline{C_u}(z) + \tfrac{1}{1-\sigma z} \right) = z^{2m} \left(\overline{C_u}(z) + \tfrac{1}{1-\sigma z} \right)$$

As in the case of $\mathcal{D}_3$ the series can be bootstrapped to obtain a close formula:

$$f_u(z) = z^m \sum_{k=0}^{\infty} (-1)^k \left(z^{m2^k} \overline{B}(z^{2^k}) + \frac{z^{m2^k}}{1 - \sigma z^{2^k}} \right) \prod_{\ell=0}^{k-1} \left(1 + z^{m2^\ell} \overline{B}(z^{2^\ell}) + \frac{z^{m2^\ell}}{1 - \sigma z^{2^\ell}} \right).$$

The JC: lower bound

Context

Let $\mathcal{C}$ be the language of canonical words.
What is the number of words of length n in $\mathcal{C}$?

The JC: lower bound

Context

Let $\mathcal{C}$ be the language of canonical words.
What is the number of words of length n in $\mathcal{C}$?

Proposition

For all ϵ their exists a language $\mathcal{L}_\epsilon$ such that its generating function have radius of convergence $1 + \phi - \epsilon$, where $\phi = \frac{1+\sqrt{5}}{2}$.
$\mathcal{L}_\epsilon$ is sub-language of $\mathcal{C}$.

Proof

$\mathcal{L}_k = ab^k(ab^{<k}(\epsilon + cb^*))^*$ have radius of convergence $(1 + \phi - \epsilon_k)^{-1}$
where $\epsilon_k = \frac{1}{2}(1 + \frac{3}{5}\sqrt{5})\frac{1}{(1+\phi)^k}(1 + o(1))$.

The JC (not to be confused with JC) : upper bound (1)

Definitions

A p-list $p = [p_1, \ldots, p_k]$ is a sequence of integers corresponding to a word w such that:

$$w = w[: p_1]\ w[: p_2]\ \ldots\ w[: p_k],$$

where $w[: 0]$ (*i.e.* an empty prefix) means a new letter (not already in w).

Example

$$p = [0, 0, 1, 0, 3, 2]$$
$$w =$$

The JC (not to be confused with JC) : upper bound (1)

Definitions

A p-list $p = [p_1, \ldots, p_k]$ is a sequence of integers corresponding to a word w such that:

$$w = w[:p_1]\ w[:p_2]\ \ldots\ w[:p_k],$$

where $w[:0]$ (*i.e.* an empty prefix) means a new letter (not already in w).

Example

$$p = [0, 0, 1, 0, 3, 2]$$
$$w = a$$

The JC (not to be confused with JC) : upper bound (1)

Definitions

A p-list $p = [p_1, \ldots, p_k]$ is a sequence of integers corresponding to a word w such that:

$$w = w[:p_1]\ w[:p_2]\ \ldots\ w[:p_k],$$

where $w[:0]$ (*i.e.* an empty prefix) means a new letter (not already in w).

Example

$$p = [0, 0, 1, 0, 3, 2]$$
$$w = ab$$

The JC (not to be confused with JC) : upper bound (1)

Definitions

A p-list $p = [p_1, \ldots, p_k]$ is a sequence of integers corresponding to a word w such that:

$$w = w[: p_1] \; w[: p_2] \; \ldots \; w[: p_k],$$

where $w[: 0]$ (*i.e.* an empty prefix) means a new letter (not already in w).

Example

$$p = [0, 0, 1, 0, 3, 2]$$
$$w = aba$$

The JC (not to be confused with JC) : upper bound (1)

Definitions

A p-list $p = [p_1, \ldots, p_k]$ is a sequence of integers corresponding to a word w such that:

$$w = w[: p_1]\ w[: p_2]\ \ldots\ w[: p_k],$$

where $w[: 0]$ (*i.e.* an empty prefix) means a new letter (not already in w).

Example

$$p = [0, 0, 1, 0, 3, 2]$$
$$w = abac$$

The JC (not to be confused with JC) : upper bound (1)

Definitions

A p-list $p = [p_1, \dots, p_k]$ is a sequence of integers corresponding to a word w such that:

$$w = w[: p_1]\ w[: p_2]\ \dots\ w[: p_k],$$

where $w[: 0]$ (*i.e.* an empty prefix) means a new letter (not already in w).

Example

$$p = [0, 0, 1, 0, 3, 2]$$
$$w = abacaba$$

The JC (not to be confused with JC) : upper bound (1)

Definitions

A p-list $p = [p_1, \ldots, p_k]$ is a sequence of integers corresponding to a word w such that:

$$w = w[:p_1]\ w[:p_2]\ \ldots\ w[:p_k],$$

where $w[:0]$ (*i.e.* an empty prefix) means a new letter (not already in w).

Example

$$p = [0, 0, 1, 0, 3, 2]$$
$$w = abacabaab$$

The JC (not to be confused with JC) : upper bound (1)

Definitions

A p-list $p = [p_1, \ldots, p_k]$ is a sequence of integers corresponding to a word w such that:

$$w = w[: p_1]\ w[: p_2]\ \ldots\ w[: p_k],$$

where $w[: 0]$ (*i.e.* an empty prefix) means a new letter (not already in w).

Example

$$p = [0, 0, 1, 0, 3, 2]$$
$$w = abacabaab$$

Proposition

Let $\mathcal{P} = 0.(\epsilon + (\geq 1)).0.(\geq 0)^*$ be the language of p-list, where $0, (\geq 0)$ and (≥ 1) are integers in unary notation.
$\mathcal{P}$ contains the language of border tables and so is an over-language of $\mathcal{C}$.

The JC (not to be confused with JC) : upper bound (2)

Proposition

The radius of convergence of P is $(1+\phi)^{-1}$.

The JC (not to be confused with JC) : upper bound (2)

Proposition

The radius of convergence of P is $(1+\phi)^{-1}$.

Corollary (of the JCs propositions)

The radius of convergence of $\mathcal{C}$ is $(1+\phi)^{-1}$

Experimentation 1

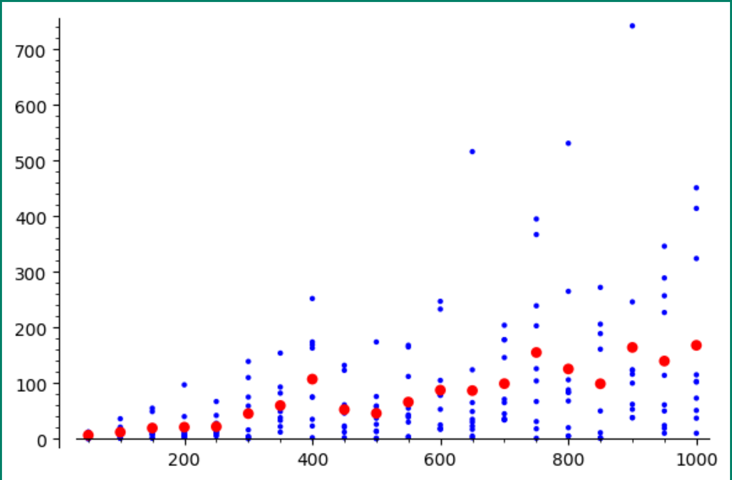
Question

What is the proportion of p-lists that are regular borders' tables ?

Experimentation 1

Question

What is the proportion of p-lists that are regular borders' tables ?



Experimentation 1

Question

What is the proportion of p-lists that are regular borders' tables ?

Conjecture 1

The proportion of length n p-lists that are borders' tables is $\frac{6}{n}(1 + o(1))$.

Experimentation 2

Question

How does look like a big canonical word ?

Experimentation 2

Question

How does look like a big canonical word ?

Conjecture 2

When n tends to ∞ a canonical word w of length n over an infinite alphabet Σ looks like:

$$w = \begin{cases} ab^{\Theta(\log n)} \left(ab^{\mathcal{O}(\log n)}(\epsilon + cb^*)\right)^* \\ ab^{\Theta(\log n)} \left(ab^{\mathcal{O}(\log n)}(\epsilon + cb^*)\right)^{\mathcal{O}\left(\frac{n}{\log n}\right)} d \left(ab^{\mathcal{O}(\log n)}(\epsilon + cb^*)\right)^* \end{cases}$$

With respective probabilities $\frac{\pi^2}{9\zeta(3)}$ and $1 - \frac{\pi^2}{9\zeta(3)}$.