Maximal number of subword occurrences in a word

Wenjie Fang
LIGM, Université Gustave Eiffel
arXiv:2406.02971

18 December 2024, Seminar of Combinatorics and Graph Theory,
National University of Singapore

Subword occurrences

A word: $w = (w_1, \ldots, w_\ell)$ with w_i in a finite alphabet $\mathcal{A}$.

Two notions of patterns: **subword** (scattered) and factor (consecutive)

Example: 01 occurs 5 times as subword in 011001, but twice as factor

Counting pattern occurrences: harder for subwords, easier for factors

Occurrence of u in w: a subset of positions in w that gives u

$\mathrm{occ}(w, u)$ or $\binom{w}{u}$: number of occurrences of u as subword of w

Flajolet, Szpankowski, Vallée (2006): normal limit law and large deviation of $\mathrm{occ}(w, u)$ for fixed u and $w \sim \mathrm{Unif}(A^n)$, $n \to \infty$.

Quite some research in many directions! Difficulty from self-correlation.

Subword entropy

Given $w \in \mathcal{A}^*$, what are its **most frequent subwords**?

Related to data-mining for finding patterns appearing frequently.

Surprisingly difficult! Complexity unknown.

$\mathrm{maxocc}(w) := \max_u \mathrm{occ}(w, u)$: maximal number of subword occurrences

Subword entropy: $S_{\mathrm{sw}}(w) := \log_2 \mathrm{maxocc}(w)$:

Easy to maximize: $\mathrm{maxocc}(0^n) = \binom{n}{\lfloor n/2 \rfloor}$, $S_{\mathrm{sw}}(0^n) = n + O(\log_2 n)$.

Minimal subword entropy for $|\mathcal{A}| = k$, length n:

$$\mathrm{min}S_{\mathrm{sw}}^{(k)}(n) := \min_{u \in \mathcal{A}^n} S_{\mathrm{sw}}(w).$$

The first bounds for $\min S_{\mathrm{sw}}^{(k)}(n)$

Trivial upper bound (from 0^n): for some constant c,

$$\min S_{\mathrm{sw}}^{(k)}(n) \leq n - \frac{1}{2}\log_2 n + c.$$

Easy lower bound: for some constant c',

$$\min S_{\mathrm{sw}}^{(k)}(n) \geq \log_2(1 + k^{-1})n - \frac{1}{2}\log_2 n + c'.$$

Reasoning: For any fixed w, take random word u of length αn. Then

$$S_{\mathrm{sw}}(w) \geq \log_2 \mathbb{E}[\mathrm{occ}(w, u)] = \log_2 \left(\binom{n}{\alpha n} k^{-\alpha n} \right).$$

Maximized at $\alpha = (k+1)^{-1}$. Holds for all w.

Super-additivity

Proposition (Super-additivity of $\min S_{\mathrm{sw}}$)

Given $k \geq 2$, for $n, m \geq 1$,

$$\min S_{\mathrm{sw}}^{(k)}(n+m) \geq \min S_{\mathrm{sw}}^{(k)}(n) + \min S_{\mathrm{sw}}^{(k)}(m).$$

Not difficult, but a little twist!

Lemma (Fekete's lemma)

For (g_n) super-additive, when $n \to +\infty$, then g_n/n either tends to $+\infty$, or converges to some limit L.

Corollary

The minimal subword entropy per letter $\min S_{\mathrm{sw}}^{(k)}(n)/n$ has a limit L_k:

$$\log_2(1+k^{-1}) \leq L_k \leq 1.$$

Better bounds?

Binary words with minimal entropy

When no idea, brute force!

Very hard... Start with the binary case.

n	Words w with min. subword entropy	$\mathrm{maxocc}(w)$	Symmetry
1	0	1	P
2	01	1	A
3	001	2	
	010		P
4	0110	2	P
5	01110	3	P
6	011001	5	A
7	0110001	6	
8	01110001	9	A
9	011000110	16	P
10	0110001110	22	
11	01110001110	33	P
12	011000111001	52	A
13	0111001001110	72	P
14	01100010111001	108	A

P: palindromic, A: anti-palindromic

Binary words with minimal entropy (cont'd)

Interesting, some more!

n	Words w with min. subword entropy	$\mathrm{maxocc}(w)$	Symmetry
15	0110001011110001	162	
16	01110001011110001	252	A
17	011000111111000110	390	P
18	0111001001011110001	588	
19	0110001011101000110	900	P
	0110001110110001110		
20	01110001011011000110	1320	
21	011100011011010001110	2049	
22	0110001110101000111001	2958	A
23	01110001011011010001110	4473	P
24	011000111010101000111001	6979	A
25	0111000101101101000111001	10602	
26	01110001011011001000111001	15962	
27	011100010101110101000111001	24150	
28	0110001111010010010111000110	36450	
	0111000101110101000101110001		A
29	01100011101010001010111000110	53671	P
30	011000111001100010101111000110	83862	

Binary words with minimal entropy (cont'd 2)

Confusing... A last push!

n	Words w with min. subword entropy	$\mathrm{maxocc}(w)$	Symmetry
31	0110001110101000101011110001110	127998	
32	01100011101010001010111010001110	189131	
33	011000111101010001011011010001110	288900	
34	0110001110101000101011101001001110	442386	
35	01110001011011001000110111001001110	681966	
36	011100010111010100010110111001001110	1047330	
37	0111000101101011000011011011010001110	1581150	
38	01110001011011011000100111011001001110	2387054	
39	011000110110010011101100010010111000110	3626580	
40	0110001110101000101011101010001110010110	5500610	

The last line took 6 days on a server with 32 cores.

Naïve complexity: $O(4^n n^2)$. A lot of optimizations needed.

Observations

n	Words w with min. subword entropy	$\text{maxocc}(w)$	Symmetry
31	0110001110101000101011110001110	127998	
32	01100011101010001010111010001110	189131	
33	011000111101010001011011010001110	288900	
34	0110001110101000101011101001001110	442386	
35	01110001011011001000110111001001110	681966	
36	011100010111010100010110111001001110	1047330	
37	0111000101101011000011011011010001110	1581150	
38	01110001011011011000100111011001001110	2387054	
39	011000110110010011101100010010111000110	3626580	
40	0110001110101000101011101010001110010110	5500610	

- For larger n, symmetry runs out.
- Average run length 1.6–2, mostly $1, 2, 3$, but length 4 and 5 exist.
- Growth rate slightly larger than 1.5 given by lower bound of L_2.

Idea: Find words like them, but analyzable.

Three families inspired by experiments

Average run length slightly less than 2. Most runs have length $1, 2, 3$.

Candidates: $(01)^m$, $(0011)^m$, $(000111)^m$.

Proposition

The following words has a most frequent subword of the form

- $(01)^m$: *subword* $(01)^r$;
- $(0011)^m$: *subword* $(01)^r$;
- $(000111)^m$: *subword* $(0011)^r$.

With local analysis in subword.

Key result for analysis, as most frequent subwords are hard to compute!

Experimentally, periodic words have periodic most frequent subwords.

But no proof!

Generating functions of some periodic subword occurrences

Occurrence generating function: $f_{w,u}(x,y) = \sum_{m,r\geq 0} \mathrm{occ}(w^m, u^r) x^m y^r$

Proposition

$$f_{01,01} = \frac{1-x}{(1-x)^2 - xy},$$

$$f_{0011,01} = \frac{1-x}{(1-x)^2 - 4xy},$$

$$f_{000111,0011} = \frac{(1-x)^3}{(1-x)^4 - 9x(1+2x)^2 y}.$$

$\mathrm{maxocc}(w^m) = \max_r [x^m y^r] f_{w,u}$ for these families.

Can be computed manually, or using (automated) ACSV or saddle-point on large powers.

General result on periodic subword occurrences

In fact a universal and effective result!

Theorem

For any words $w, v \in \mathcal{A}^*$, the g.f. $f_{w,v}(x,y)$ is rational in x, y.

Proof.

- $g_{w,u}(x) = \sum_{m \geq 0} \operatorname{occ}(w^m, u) x^m$ is rational by looking at "clusters" of letters of u in the same copy of w.
- The same holds when fixing the occurrence of the first and the last letter of u.
- We consider variants of $f_{w,v}(x,y)$ fixing the first letter of u^r in w^m.
- We write a linear system of variants of $f_{w,v}(x,y)$ using variants of $g_{w,u}(x)$, by considering the last copy of u in w^m.
- The system is invertible, so the unique solution is rational. □

Problem is that we don't know the most frequent subwords...

Asymptotics and bounds on L_2

Proposition

Word w	*Subword*	*Max at*	$S_{\mathrm{sw}}(w)$
$(01)^m$	$(01)^r$	$r = \frac{m}{\sqrt{5}}$	$m \log_2 \frac{3+\sqrt{5}}{2} + \frac{\log_2 m}{2} + O(1)$
$(0011)^m$	$(01)^r$	$r = \frac{m}{\sqrt{2}}$	$m \log_2(3+2\sqrt{2}) + \frac{\log_2 m}{2} + O(1)$
$(000111)^m$	$(0011)^r$	$r = \alpha m$	$m\gamma - \frac{\log_2 m}{2} + O(1)$

Here, $\alpha \approx 0.66\ldots$ *is the pos. sol. of* $457\alpha^4 - 246\alpha^2 + 72\alpha - 27 = 0$, *and*

$$\gamma = \alpha \log_2 9 + 2\alpha \log_2 \frac{1+2\zeta}{(1-\zeta)^2} - (1-\alpha)\log_2 \zeta,$$

$$\zeta = \frac{1 - 9\alpha + \sqrt{73\alpha^2 - 18\alpha + 9}}{4+4\alpha}.$$

Upper bounds of L_2: $0.694\ldots$, $0.636\ldots$, $0.654\ldots$.

We have $0.585\ldots = \log_2(3/2) \le L_2 \le \frac{1}{2}\log_2(1+\sqrt{2}) = 0.636\ldots$.

Open problems

- Value of L_2? Value of other L_k?
- Better bounds? We should have $L_2 > \log_2(3/2)$.
- For occurrences of $(0011)^r$ in $(0001100111)^m$ (rotation of record for $n = 10$),

$$f_{0001100111,0011} = \frac{(1-x)^3 - x(9x^2+78x+13)y}{(1-x)^4 - 9x(1-6x)^2y^2 - x(9x+16)(21x+4)y},$$

  has a growth rate $0.63272\ldots$. Better bound of L_2 if indeed optimal subword.
- Does periodic word have a quasi-periodic most frequent subword?
- Can we reach L_2 with periodic words?
- Any structure on words almost realizing $\min S_{\mathrm{sw}}^{(k)}(n)$?

Difficult "minimal of maximal" structure, chaos in experimental data

Open problems (cont'd)

A lot of unknowns, even intuitive ones!

- Is $\min S_{\mathrm{sw}}^{(2)}(n)/n$ ultimately increasing?
- For any w, is every most frequent subword is of length $\leq \lceil |w|/2 \rceil$?
- What are the n's with multiple words realizing $\min S_{\mathrm{sw}}^{(2)}(n)$?
- What are the n's with optimal words containing runs > 3?

Open problems (cont'd)

A lot of unknowns, even intuitive ones!

- Is $\min S_{\mathrm{sw}}^{(2)}(n)/n$ ultimately increasing?
- For any w, is every most frequent subword is of length $\leq \lceil |w|/2 \rceil$?
- What are the n's with multiple words realizing $\min S_{\mathrm{sw}}^{(2)}(n)$?
- What are the n's with optimal words containing runs > 3?

Thank you for your attention!