Sets Represented as the Length-n Factors of a Word

Shuo Tan and Jeffrey Shallit
School of Computer Science,
University of Waterloo,
Waterloo, Ontario
N2L 3G1 Canada
s22tan@uwaterloo.ca shallit@cs.uwaterloo.ca

September 14, 2013

Definitions

- Let $F_n(w)$ denote the set of length-n factors of a word w.
- Let $\mathrm{CF}_n(w)$ denote the set of length-n cyclic factors of a word w.

Definition

A word w **witnesses** (resp., **cyclically witnesses**) a subset S of Σ^n if $F_n(w) = S$ (resp., $\mathrm{CF}_n(w) = S$).

A subset S of Σ^n is **representable** (resp. **cyclically representable**) if there exists a word w that witnesses (cyclically witnesses) S.

Examples

Example (cyclic representable set)

The word $w = 000011$ cyclically witnesses the set $\{000, 001, 011, 110, 100\}$. Thus the set $\{000, 001, 011, 110, 100\}$ is cyclically representable.

Example (representable set)

$w = 00001$. $F_3(w) = \{000, 001\}$. Thus, the set $\{000, 001\}$ is representable. The set $\{000, 001\}$ is not cyclically representable.

Example (unrepresentable set)

The set $\{00, 11\}$ is not representable.

Definition

A binary **De Bruijn sequence** B_n of order n, is a word w over Σ such that every possible word of length n appears exactly once as a cyclic factor of w.

Example (De Bruijn sequences)

n	B_n
2	0011
3	00010111
4	0000100110101111
5	00000111011010111111001100010100 1

binary De Bruijn sequence

Theorem

For any order $n > 0$, a binary De Bruijn sequence B_n exists.

Corollary

The set Σ^n is cyclically representable.

Problems

Problem (1)

How many subsets of Σ^n are (cyclically) representable?

Problem (2)

For how many subsets S of Σ^n does there exist a word w of length t that witnesses S?

Numerical results for Problem 1

n	non-cyclic case	cyclic case
1	3	3
2	14	6
3	121	27
4	5921	972
5	20020315	2466131

Bounds for Problem 1

Theorem

A lower bound on the number of cyclically representable subsets is $2^{2^{n-1}}$.

Theorem

An upper bound on the number of cyclically representable subsets is $10^{2^{n-2}}$.

The number of subsets of Σ^n is $2^{2^n} = 16^{2^{n-2}}$.

Problem 2

Let $T(n, t)$ denote the number of subsets S of Σ^n such that there exists a word w of length t that witnesses S.

Problem (2)

$T(n, t) = ?$

For this problem, we have some results on the non-cyclic case.

Example

We fix $n = 2$ and $t = 3$.

word of length 3	witnesses
000	$\{00\}$
001	$\{00, 01\}$
010	$\{01, 10\}$
011	$\{01, 11\}$
100	$\{10, 00\}$
101	$\{01, 10\}$
110	$\{11, 10\}$
111	$\{11\}$

Note that 010 and 101 'coincide'. Thus $T(2, 3) = 7$.

Problem 2

In order to compute $T(n, t)$, we compute the number of coincidences; namely, we consider the number of words that witness the same subset of Σ^n.

Suppose $S \subseteq \Sigma^n$. Let $C_t(S)$ denote the number of words of length t that witness S. Then we have

$$T(n, t) = 2^t - \sum_{S \in \Sigma^n, C_t(S) > 1} (C_t(S) - 1)$$

.

Some easy cases

1. $T(n, n) = 2^n$ for $n > 0$.
2. $T(n, n+1) = 2^{n+1} - 1$ for $n \geq 1$.
3. $T(n, n+2) = 2^{n+2} - 5$ for $n \geq 2$.
4. $T(n, n+3) = 2^{n+3} - 14$ for $n \geq 3$.

Period

Definition

A positive integer p is a **period** of a (finite) word w, if for any $1 \leq i \leq |w| - p$ we have $w[i] = w[i + p]$.

Let $\pi(w)$ denote the minimal period of w.

Example

Both 3 and 5 are periods of the word $w = 010010$. The minimal period of w is 3.

Definition

A **root** $r(w)$ of a word w is the prefix of w with length $\pi(w)$. We say that two words w and w' are **root-conjugate** if $r(w)$ and $r(w')$ are conjugate.

Example

The words $w_1 = 01001$ and $w_2 = 10010$ are root-conjugate since $r(w_1) = 010$ and $r(w_2) = 100$ are conjugate.

Theorem

Let t, n, k be integers such that $t = n + k$, $n \geq k + 1$, and $k \geq 0$. For any distinct words w, w' of length t, we have $F_n(w) = F_n(w')$ if and only if w and w' are root-conjugate and $\pi(w) = \pi(w') \leq k + 1$.

Example

For example, let $t = 7$, $n = 4$, and $k = 3$. We consider the words $w_1 = 0110110$ and $w_2 = 1011011$ that are root-conjugate. We have $F_4(w_1) = F_4(w_2) = \{0110, 1101, 1011\}$.

Theorem

Let t, n, k be integers such that $t = n + k$, $n \geq k + 1$, and $k \geq 0$. For any distinct words w, w' of length n, we have $F_n(w) = F_n(w')$ if and only if w and w' are root-conjugate and $\pi(w) = \pi(w') \leq k + 1$.

Proof.

If w and w' are root-conjugate with period $p \leq k + 1$, then there are p places to begin, and considering consecutive factors of length $n + p - 1$ gives exactly p distinct length-n factors.

Proof (Cont.)

For the other direction, we give a proof by induction on k.

- The base case where $k = 0$ is easy. In this case $t = n$ and thus $F_n(w) = \{w\}$ and $F_n(w') = \{w'\}$. Thus $w = w'$, contradicting the fact that w and w' are distinct.
- We assume the result holds for $k - 1$ and we prove it for k.
- Let $p_i(w)$ denote the length-i prefix of w. Let $s_i(w)$ denote the length-i suffix of w.

Proof (Cont.)

We consider the following cases:

1.
   - $p_{n-1}(w)$ appears only once as a factor of w.
   - In this case we can prove that $p_n(w) = p_n(w')$.
   - Let $s = w[2..t]$ and $s' = w[2..t]$. We can show that $F_n(s) = F_n(s')$.
   - Let $t' = t - 1$ and $k' = k - 1$. Applying induction we obtain that s and s' are root-conjugate and $\pi(w) = \pi(w') \leq k' + 1 = k$.
   - w and w' are root-conjugate and $\pi(w) = \pi(w') \leq k + 1$.

2. Similarly we can handle the case where $s_{n-1}(w)$ appears only once as a factor of w.

3. The last case is where both $p_{n-1}(w)$ and $s_{n-1}(w)$ appear more than once. We again consider two sub-cases depending on whether $p_{n-1}(w) = s_{n-1}(w)$.

Corollary

For $t \leq 2n-1$, *we have* $T(n,t) = 2^t - \sum_{k=1}^{t-n+1} \frac{k-1}{k} \sum_{d|k} \mu(\frac{k}{d}) 2^d$,

where $\mu(\cdot)$ *is the Möbius function.*

1. $C_t(S) > 1$ iff there exists a word w that witnesses S with $\pi(w) \leq k+1$.
2. $C_t(S) = \pi(w)$. Correspond to a set of root-cojugate words that witnesses S. Represented by their lexicographically least roots (Lyndon words).

   As an example, let $t = 5$ and $n = 3$. The set $\{010, 100, 001\}$ is witnessed by $01001, 00100, 10010$. It can be represented by 001.

Thus we have

$$
\begin{aligned}
T(n,t) &= 2^t - \sum_{S\in\Sigma^n, C_t(S)>1} (C_t(S)-1) \\
&= 2^t - \sum_{\substack{w \text{ is a Lyndon word and} \\ \pi(w)\le t-n+1}} (\pi(w)-1) \\
&= 2^t - \sum_{i=1}^{t-n+1} (i-1)\cdot L(i)
\end{aligned}
$$

where $L(i) = \frac{1}{i}\sum_{d|i}\mu(\frac{i}{d})2^d$ is the number of Lyndon words of length i.

Open problems

1. Let $\mathring{R}_n$ denote the set of all non-empty circularly representable subsets of Σ^n. Does the limit $\lim_{n\to\infty} |\mathring{R}_n|^{\frac{1}{2^n}}$ exist?

2. Derive a formula for $T(n,t)$ where $t = 2n$.

   We conjecture that if x and y are distinct binary words of length $2n$ with $F_n(x) = F_n(y)$ then $\pi(x) = \pi(y)$ and furthermore x and y are root-conjugate. However, it is possible in this case that $\pi(x) > n+1$.

   Furthermore it seems that if $\pi(x) > n+1$, then $x = uv01vu$ and $y = uv10v^R u$ (or vice versa) for some nonempty words u, v where u is the longest palindrome prefix of uv and $\pi(x) = t - |v|$.