Generating sequences

The *generating sequence* of a set of words X is the sequence $u_X = (u_n)_{n\geq 0}$ with

$$u_n = \#\{\text{words of length } n \text{ in } X\}$$

with g.s.

$$u_X(z) = \sum_{n\geq 0} u_n z^n$$

General problem: Given a family $\mathcal{F}$ of sets of words, characterize the generating sequences of the elements of $\mathcal{F}$.

Example: The generating sequences of prefix codes on k-symbols are the sequences satisfying Kraft's inequality

$$\sum_{n\geq 0} u_n k^{-n} \leq 1$$

i.e. $u(1/k) \leq 1$.

Regular sequences

A set of words X is *regular* if it can be recognized by a finite automaton.

A sequence (u_n) of integers is *regular* if there exists a finite graph G and two vertices i, t such that

$$u_n = \#\{\text{paths of length } n \text{ from } i \text{ to } t\}$$

Theorem The generating sequences of regular sets are the regular sequences.

Example: If $X = a^*b$, then

$$u_X(z) = \frac{z}{1-z}.$$

Regular vs. rational

Any regular sequence of non-negative integers is rational, i.e. $u(z) = p(z)/q(z)$. The converse is not true.

Theorem[Berstel] *The poles of minimal modulus of a regular sequence are of the $\rho\varepsilon$ where ρ is the radius of convergence and $\varepsilon^p = 1$ for some $p \geq 1$.*

In particular, the radius of convergence is a pole.

Example: Let $0 < \theta < \pi/2$ be such that $\cos\theta = a/c$ with $0 < a < c$ and $c \neq 2a$. The sequence

$$u_n = c^{2n} \cos^2 n\theta$$

is rational but not regular (roots: $1, e^{2i\theta}, e^{-2i\theta}$).

Soittola's theorem

A *dominating pole* of a rational sequence $s(z) = p(z)/q(z)$ is a pole r such that $|r'| > r$ for any other root r' of $q(z)$.

A sequence u is an *interleaving* of sequences

$$u^{(0)}, \ldots, u^{(p-1)}$$

if for $n \geq 0, 0 \leq i < p$,

$$u_{pn+i} = u_n^{(i)}.$$

Theorem[Soittola, 1976]*A sequence of non-negative integers is regular iff it is an interleaving of rational sequences with a dominating pole.*

Example:

$$1, 2, 1, 4, 2, 8, 3, 16, 5, \ldots$$

A finite-state version of the Kraft-McMillan theorem

Theorem [F. Bassino, M.P. Béal, D.P.,*SIAM Computing*, 2000] *A sequence of integers is the generating sequence of a rational prefix code iff*

(i) it is regular.

(ii) it satisfies Kraft's inequality.

Examples: Let $u(z) = 3z^2/(1-z^2)$. Then $u(1/2) = 1$. Solution on $A = \{a, b\}$:

$$X = (aa)^*(ab + ba + bb)$$

Much more difficult:

$$u(z) = \frac{z^2}{1-z^2} + \frac{z^2}{1-5z^3}.$$

Solution:

$$u(z) = z^2 z^{6*}(2+z^2+2z^3+z^4+3z^5(1+3z^3)(3z^3)^*).$$

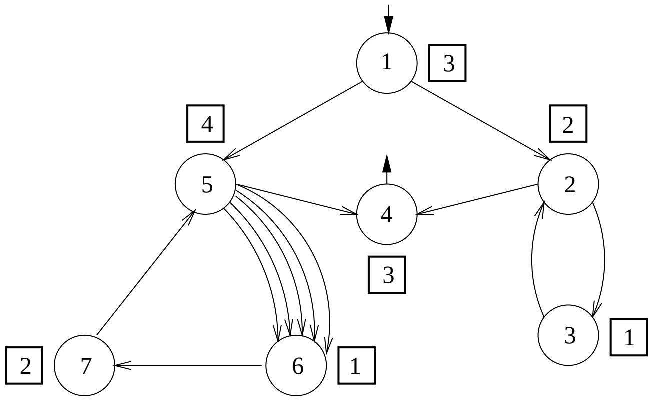


Figure 1: A normalized representation of u

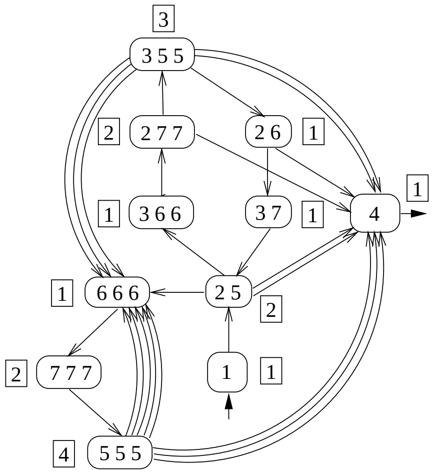


Figure 2: An admissible extension H.

The proof uses the following combinatorial lemma. This lemma is also used in a similar context by Adler et al. and Marcus. It is actually presented in (Aigner,Ziegler 1998) as a nice variant of the pigeon-hole principle.

Lemma *Let $w_1, w_2, \ldots, w_m$ be positive integers. Then there is a non-empty subset $S \subset \{1, 2, \ldots, m\}$ such that $\sum_{q \in S} w_q$ is divisible by m.*

Bifix codes

A *bifix code* is a set X of words which is both a prefix and a suffix code.

Open problem: *what are the generating sequences of bifix codes?*

Example: $u = (1, 2)$ is not realizable on a binary alphabet although $u(1/2) = 1$.

Theorem [Ahlswede, Balkenhol, Khachatrian] *For any integer sequence u such that*

$$u(1/2) \leq 1/2,$$

there is a bifix code X s.t. $u = u_X$.

Proof: if $\sum_{n=1}^{N} u_n 2^{-n} < 1/2$, then

$$2 \sum_{n=1}^{N} u_n 2^{N-n} = 2^{N+1} (\sum u_n 2^{-n}) < 2^N$$

and there is still room for one more word of length N.

Conjecture: true if $u(1/2) \leq 3/4$?

Subshifts of finite type

A subshift of finite type is a set S of biinfinite words avoiding a finite set of words.

Example: words on $\{a, b\}$ without a block bb.

Let

$$s_n = \#\{\text{words of period } n \text{ in } S\}.$$

The *zeta function* of S is the series

$$\zeta(S) = \exp \sum_{n \geq 1} s_n z^n$$

For S as above

$$\zeta(S) = \frac{1}{1 - z - z^2}.$$

Circular codes

A *circular code* is a set X of words such that the factorization of words written on a circle is unique.

Example:

$a + aba$ is a circular code.

$ab + ba$ is not.

Strong connexion with *subshifts of finite type*:

- if X is a finite circular code, the infinite concatenations of words of X form a subshift of finite type.
- if S is a subshift of finite type, the set of first returns to a given vertex is a circular code.

If X is a circular code, let

$$\sum_{n\geq 1} \frac{s_n}{n} z^n = \log \frac{1}{1-u_X}$$

Then s_n is the number of words of length n with a circular factorization in words of X.

Let $s_n = \sum_{d|n} d l_d$. Then l_n is the number of nonperiodic necklaces with a factorization.

Cyclotomic identity

$$1 - kz = \prod_{n \geq 1} (1 - z^n)^{l_n(k)}.$$

Tabulation of the Witt numbers:

n	1	2	3	4	5	6
$l_n(2)$	2	1	2	3	6	9
$l_n(3)$	3	3	8	18	48	116
$l_n(4)$	4	3	20			

The value $l_3(4) = 20$ is famous because of the genetic code: there are precisely 20 amino-acids coded by words of length 3 over a 4-symbol alphabet A,C,G,U.

Witt vectors

For any sequence $a(a_n)_{n\geq 1}$ let

$$s_n = \sum_d d a_d^{n/d}$$

The pair (a, s) is called a *Witt vector.* The numbers s_n are the *ghost components.*

$$\exp \sum_{n\geq 1} \frac{s_n}{n} z^n = \prod_{n\geq 1} (1 - a_n z^n)^{-1}$$

For a circular code X, one has

$$\begin{aligned} 1 - u_X(z) &= \prod_{n\geq 1} (1 - z^n)^{l_n} \\ &= \prod_{n\geq 1} (1 - a_n z^n) \end{aligned}$$

Theorem (Schützenberger, 1965) : *The generating sequences of circular codes on k symbols are the sequences u s.t.*

$$l_n(u) \leq l_n(k) \qquad (n \geq 1)$$

Sequence of inequalities :

$$\begin{aligned} u_1 &\leq k \\ u_2 + \frac{1}{2}(u_1^2 - u_1) &\leq \frac{1}{2}(k^2 - k) \\ u_3 + u_1 u_2 + \frac{1}{3}(u_1^3 - u_1) &\leq \frac{1}{3}(k^3 - k) \\ \cdots &\leq \frac{1}{4}(k^4 - k^2) \end{aligned}$$

- can be deduced of Krieger's embedding theorem
- Complements in Bassino (1999).

Idea of the proof

For $k = 3$ we consider the distribution

n	1	2	3	4
u_n	0	2	3	2
$l_n(u)$	0	2	3	3
$l_n(k)$	3	3	8	18

Lazard elimination algorithm:

$$a, b, c$$

Elimination of a:

$$b, c, ba, ca, baa, caa, baaa, caaa, \ldots$$

Elimination of b :

$$c, ba, ca, cb, baa, caa, cbb, bab, cab, \ldots$$

Elimination of c:

$$ba, ca, cb, baa, caa, cbb, bab, cab, bac, cac, cbc, \ldots$$

Elimination of ba :

$$ca, cb, baa, caa, cbb, bab, cab, bab, cac, cbc, \ldots$$

Link with Lie algebras

The free Lie algebra on A is formed by the linear combinations of elements generated from the symbols $a \in A$by the operation

$$[x, y] = xy - yx$$

One thus obtains:

$$a,\ b$$

$$[ba]$$

$$[[ba]a], [[ba]b]$$

$$[[[ba]a]a], [[[ba]a]b], [[[ba]b]b]$$

The dimension of the component of order n is $l_n(k)$ (Witt's formula).

Lazard's algorithm gives a basis.

Franaszek's code

Start :

$$a, b, c$$

Elimination of b :

$$a, c, ba$$

Elimination of c :

$$a, ba, ca, cba$$

Elimination of a :

$$ba, ca, aba, aca, cba, acba, aaca$$

Result: the Franaszek code (constraint $[2, 7]$ with $a = 00, b = 01, c = 10$.).

X	Y
ba	10
ca	11
aba	000
cba	010
aca	011
$acba$	0010
$aaca$	0011

The Adler & al. coding theorem

Theorem *If T is an sft s.t. $h(T) \geq \log k$, there exists an sft $S \subset T$ and a sliding block code $f : S \to A^Z$ from S onto the full shift on k symbols.*

Proof using circular codes in the case of $>$:

Let $G = (V, E)$ be a graph representing the transitive sft T.

Let X be the circular code formed by the set of first returns to some vertex of G.

Since $h(T) > \log k$, $u_X(1/k) > 1$. Let Y be a finite subset of X such that $u_Y(1/k) = 1$.

Let Z be a prefix code on k symbols s.t. $u_Z = u_Y$.

A one-to-one length preserving correspondance between Y and Z solves the problem.

Open problem

If the sequence u is $\mathbb{N}$-rational and satisfies the inequalities

$$l_n(u) \leq l_n(k) \qquad (n \geq 1)$$

does there exist a rational circular code such that $u = u_X$?

Example: Let $u = \frac{z^2}{1-2z^2}$. Then $u = u_X$ for

$$X = (aa + bb)^* ab$$

which is circular.

Fixed alphabet

What is the minimal possible alphabet for a given regular generating sequence? for example, the set $(a+b)^*c^+$ of words formed of a word on a, b followed by c have the same generating sequence as $(a+b)^*ab^*$.

Theorem[M.P. Béal, DP, *JACM*, 50, 955-980, 2003] *A sequence $(u_n)_{n\geq 0}$ is the generating sequence of a regular language on k symbols iff both $(u_n)_{n\geq 0}$ and $(k^n - u_n)_{n\geq 0}$ are regular sequences.*

Example

$$\begin{aligned} s_0 &= 1, \\ s_1 &= 2, \\ s_n &= 3s_{n-1} - 2s_{n-2}. \end{aligned}$$

$$\mathbf{i} = \begin{bmatrix} 1 & 0 \end{bmatrix}, M = \begin{bmatrix} 0 & 1 \\ -2 & 3 \end{bmatrix}, \mathbf{t} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

The characteristic polynomial is $(X-2)(X-1)$.

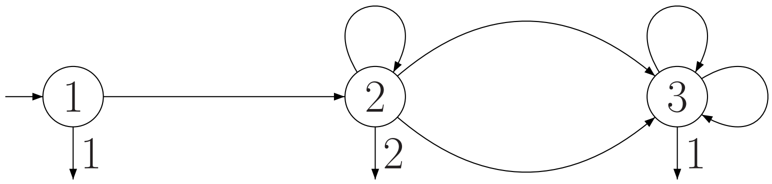
$$\begin{aligned} \mathbf{i}M^0 &= \begin{bmatrix} 1 & 0 \end{bmatrix}, \\ \mathbf{i}M^1 &= \begin{bmatrix} 0 & 1 \end{bmatrix}, \\ \mathbf{i}M^2 &= \begin{bmatrix} -2 & 3 \end{bmatrix} = \mathbf{i}M + 2\begin{bmatrix} -1 & 1 \end{bmatrix}. \end{aligned}$$

Let $\mathbf{u} = \begin{bmatrix} -1 & 1 \end{bmatrix}$. Then

$$\mathbf{u}M = 2\mathbf{u}.$$

We choose

$$U = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ -1 & 1 \end{bmatrix}.$$



$$(\mathbf{i}, M, \mathbf{t}) \xrightarrow[\mathbb{Z}]{U} (\mathbf{j}, N, \mathbf{x})$$

$$\mathbf{j} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}, N = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 2 \end{bmatrix}, \mathbf{x} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$