.

Symbolic Dynamics and Finite Automata

Dominique Perrin
`perrin@univ-mlv.fr`

Length distributions

The *length distribution* 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 length distributions of the elements of $\mathcal{F}$.

Example: The length distributions 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. $u(1/k) \leq 1$.

$\mathbb{N}$-rational sequences

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

A sequence (u_n) of integers is $\mathbb{N}$-rational 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 length distributions of rational sets are the* $\mathbb{N}$*-rational sequences.*

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

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

A finite-state version of the Kraft-McMillan theorem

Theorem (F. Bassino, M.P. Béal, D.P.) *A sequence of integers is the length distribution of a rational prefix code iff*

(i) it is $\mathbb{N}$*-rational.*

(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:

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

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 length distributions of bifix codes?*

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

Theorem (Ahlswede, Balkenhol, Khachatrian, 1997) *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 *entropy* of S is

$$h(S) = \log 1/\rho$$

where ρ is the radius of convergence of $\sum_{n\geq 1} s_n z^n$.
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.

Examples:

$a + aba$ 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 $s_n = \sum_{d|n} l_d$. Then l_n is the number of nonperiodic words with a circular factorization.

Theorem (Schützenberger, 1965) : *The length distributions 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 $u = u_X$?

It is true if $u(1/k) < 1$ by Krieger's embedding theorem.