Enumerative combinatorics on words

Dominique Perrin

Enumerative combinatorics on words

Dominique Perrin
Institut Gaspard Monge, Université de Marne-la-Vallée,
77454 Marne-la-Vallée Cedex 2 France.
perrin@univ-mlv.fr.

Abstract

We present the state of the art in the field of generating series for formal languages. The emphasis is on regular languages and rational series. The paper covers aspects including regular trees and the Kraft-McMillan inequality as well as necklaces and zeta functions.

Contents

1 Introduction

Generating series, also called generating functions play an important role in combinatorial mathematics. Many enumeration problems can be solved by transferring the basic operations on sets into algebraic operations on formal series leading to a solution of an enumeration problem. The famous paper by Doubilet, Rota and Stanley 'The idea of generating function' [41], places the subject in a general mathematical frame allowing to present in a unified way the diverse sorts of generating functions from the ordinary ones to the exponential or even Dirichlet ones.

Their place within the field of combinatorics on words is particular. It was indeed M. P. Schützenberger's point of view that sets of words can be considered as series in several non-commutative variables. The generating series of the set appears then as a the image of the non-commutative series through an homomorphism. This gives rise to a rich domain in which an interplay between classical commutative algebra and combinatorics on words is present.

In these lectures, I will survey on several aspects of these generating functions on words. The emphasis is on the most elementary case corresponding to sets of words which can be defined using a finite automaton, usually called regular. The corresponding series are actually rational. Two special cases will be considered in turn. The first one is the case of sets of wodrs corresponding to leaves in a tree and usually called prefix codes. A recent result due to Frédérique Bassino, Marie-Pierre Béal and myself [10] is presented. It completely characterizes the generating series of regular prefix codes. The second one is the case of sets of words considered up to a cyclic permutation, often called necklaces. The corresponding generating series are the zeta functions of symbolic dynamics.

A word on the terminology used here. We constantly use the term *regular* where a richer terminology is often used. In particular, what we call here a regular sequence is, in Eilenberg's terminology, an $\mathbb{N}$-rational sequence (see [22], [42] or [18]).

2 Regular sequences and automata

We consider the set A^* of all words on a given alphabet A. A subset of A^* is often called a *formal language*. For sets $X, Y \subset A^*$, we denote

$$\begin{aligned} X + Y &= X \cup Y, \\ XY &= \{xy \mid x \in X, y \in Y\}, \\ X^* &= \{x_1 x_2 \cdots x_n \mid x_i \in X, n \geq 0\} \end{aligned}$$

We say that the pair (X, Y) is unambiguous if for each $z \in XY$ there is at most one pair $(x, y) \in X \times Y$ such that $z = xy$.

We say that a set of nonempty words X is a *code* if for each $x \in X^*$ there is at most one sequence $(x_1, x_2, \ldots, x_n)$ with $x_i \in X$ such that $x = x_1 x_2 \cdots x_n$ (one also says that X is uniquely decipherable). A particular case of a code is a *prefix code*. It is a set of words X such that no element of X is a prefix of another one. It is easy to see that such a set is either reduced to the empty word or does not contain the empty word and is then a code.

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

$$u_n = \mathrm{Card}(X \cap A^n).$$

We denote by u_X the formal series

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

which is the ordinary generating series of the sequence u_X.

For example, the length distribution of $X = A^*$ is $u(z) = \frac{1}{1-kz}$ where $k = \mathrm{Card}(A)$.

The *entropy* of a formal language X is

$$h(X) = \log(1/\rho),$$

where ρ is the radius of convergence of the series $u_X(z)$. It is well defined provided X is infinite and thus ρ is finite. If the alphabet A has k elements, we have $h(X) \leq \log k$.

The following result relates the basic operations on sets with operations on series.

PROPOSITION 1 *The following properties hold for any subsets X, Y of A^*.*

(i) If $X \cap Y = \emptyset$, then $u_{X+Y} = u_X + u_Y$.

(ii) If the pair (X, Y) is unambiguous, then $u_{XY} = u_X u_Y$.

(iii) If X is a code, then $u_{X^*} = 1/(1 - u_X)$.

Proof. The first two formulae are clear. If X is a code, every word in X^* has a unique decomposition as a product of words in X. This implies that

$$u_{X^n} = (u_X)^n$$

and thus,

$$u_{X^*} = 1 + u_X + \cdots + u_{X^n} + \cdots = 1/(1 - u_X)\,.$$

□

EXAMPLE 1 The set $X = \{b, ab\}$ is a prefix code. The series u_{X^*} is

$$u_{X^*}(z) = \frac{1}{1 - z - z^2}.$$

Let $(F_n)_{n\geq 0}$ be the sequence of Fibonacci numbers defined by $F_0 = 0$, $F_1 = 1$, and $F_{n+2} = F_{n+1} + F_n$. It follows from the recurrence relation that

$$\frac{z}{1 - z - z^2} = \sum_{n\geq 0} F_n z^n.$$

Consequently, $u_{X^*}(z) = \sum_{n\geq 0} F_{n+1} z^n$. It can also be proved by a combinatorial argument that the number of words of length n in X^* is F_{n+1}.

There are several variants of the generating series considered above. One may first define

$$p_X(z) = \sum_{n\geq 0} \frac{u_n}{k^n} z^n,$$

where $k = \mathrm{Card}(A)$. The coefficients of z^n in $p_X(z)$ is the probability for a word of length n to be in the set X. The relation between u_X and p_X is simple since $p_X(z) = u_X(z/k)$. Another variant of the generating series is the *exponential generating series* of the sequence $(u_n)_{n\geq 0}$ defined as

$$e(z) = \sum_{n\geq 0} \frac{u_n}{n!} z^n.$$

We will also use the zeta function of a sequence $(u_n)_{n\geq 1}$ defined as

$$\zeta(z) = \exp \sum_{n\geq 1} \frac{u_n}{n} z^n.$$

2.1 Regular sequences

We consider sequences of natural integers $s = (s_n)_{n\geq 0}$. We shall not distinguish between such a sequence and the formal series $s(z) = \sum_{n\geq 0} s_n z^n$.

We usually denote a vector indexed by elements of a set Q, also called a Q-vector, with boldface symbols. For $\mathbf{v} = (v_q)_{q\in Q}$ we say that v is nonnegative, denoted $\mathbf{v} \geq 0$, (resp. positive, denoted $\mathbf{v} > 0$) if $v_q \geq 0$ (resp. $v_q > 0$) for all $q \in Q$. The same conventions are used for matrices. A nonnegative $Q \times Q$-matrix M is said to be *irreducible* if, for all indices p, q, there is an integer m such that $(M^m)_{p,q} > 0$. The matrix is *primitive* if there is an integer m such that $M^m > 0$.

The *adjacency matrix* of a graph $G = (Q, E)$ is the $Q \times Q$-matrix M such that for each $p, q \in Q$, the integer $M_{p,q}$ is the number of edges from p to q. The adjacency matrix of a graph G is irreducible iff the graph is strongly connected. It is primitive if, moreover, the $g.c.d$ of lengths of cycles in G is 1.

Let G be a finite graph and let I, T be two sets of vertices. For each $n \geq 0$, let s_n be the number of distinct paths of length n from a vertex of I to a vertex of T. The sequence $s = (s_n)_{n\geq 0}$ is called the sequence *recognized* by (G, I, T) or also by G if I and T are already specified. When $I = \{i\}$ and $T = \{t\}$, we simply denote (G, i, t) instead of $(G, \{i\}, \{t\})$.

A sequence $s = (s_n)_{n\geq 0}$ of nonnegative integers is said to be *regular* if it is recognized by such a triple (G, I, T), where G is finite. We say that the triple (G, I, T) is a *representation* of the sequence s. The vertices of I are called *initial* and those of T *terminal*. Two representations are said to be *equivalent* if they recognize the same sequence.

A representation (G, I, T) is said to be *trim* if every vertex of G is on some path from I to T. It is clear that any representation is equivalent to a trim one.

A well known result in theory of finite automata allows one to use a particular representation of any regular sequence s such that $s_0 = 0$. One can always choose in this case a representation (G, i, t) of s with a unique initial vertex i, a unique final vertex $t \neq i$ such that no edge is entering vertex i and no edge is going out of vertex t. Such a representation is called a *normalized representation* (see for example [37] page 14).

Let (G, i, t) be a trim normalized representation. If we merge the initial vertex i and the final vertex t in a single vertex still denoted by i, we obtain a new graph denoted by $\overline{G}$, which is strongly connected. The triple $(\overline{G}, i, i)$ is called the *closure* of (G, i, t).

Let s be a regular sequence such that $s_0 = 0$. The *star* s^* of the sequence

s is defined by

$$s^*(z) = \frac{1}{1 - s(z)}.$$

PROPOSITION 2 *If (G, i, t) is a normalized representation of s, its closure $(\overline{G}, i, i)$ recognizes the sequence s^*.*

Proof. The sequence s is the length distribution of the paths of first returns to vertex i in $\overline{G}$, that is of finite paths going from i to i without going through vertex i. The length distribution of the set of all returns to i is thus $1 + s(z) + s^2(z) + \ldots = 1/(1 - s(z))$. □

An equivalent definition of regular sequences uses vectors instead of sets I, F. Let $\mathbf{i}$ be a Q-row vector of nonnegative integers and let $\mathbf{t}$ be a Q-column vector of nonnegative integers. We say that $(G, \mathbf{i}, \mathbf{t})$ recognizes the sequence $s = (s_n)_{n \geq 0}$ if for each integer $n \geq 0$

$$s_n = \mathbf{i} M^n \mathbf{t},$$

where M is the adjacency matrix of G. The proof that both definitions are equivalent follows from the fact that the family of regular sequences is closed under addition (see [22]). A triple $(G, \mathbf{i}, \mathbf{t})$ recognizing a sequence s is also called a representation of s and two representations are called equivalent if they recognize the same sequence.

A sequence $s = (s_n)_{n \geq 0}$ of nonnegative integers is *rational* if it satisfies a recurrence relation with integral coefficients. Equivalently, s is rational if there exist two polynomials $p(z), q(z)$ with integral coefficients and with $q(0) = 1$ such that

$$s(z) = \frac{p(z)}{q(z)}.$$

Figure 1: The Fibonacci graph.

For example, the sequence s defined by $s(z) = \frac{z}{1-z-z^2}$ is the sequence of Fibonacci numbers also defined by $s_0 = 0, s_1 = 1$ and $s_{n+1} = s_n + s_{n-1}$. It is recognized by the graph of Figure 1 with $I = \{1\}$ and $T = \{2\}$.

Any regular sequence is rational. The converse is however not true (see Section 3.6).

A theorem of Soittola [42], also found independently in [27] characterizes those rational sequences which are regular. We say that a rational sequence has a *dominating root*, either if it is a polynomial or if it has a real positive pole which is strictly smaller than the modulus of any other one. A sequence r is a *merge* of the sequences r_i if there is an integer p such that

$$r(z) = \sum_{i=0}^{p-1} z^i r_i(z^p).$$

THEOREM 1 (SOITTOLA) *A sequence of nonnegative integers $r = (r_n)_{n\geq 0}$ is regular if and only if it is a merge of rational sequences having a dominating root.*

This result shows that it is decidable if a rational series is regular (see [42]). In the positive case, there is an algorithm computing a representation of the sequence.

2.2 Finite automata

We present here a brief introduction to the concepts used in automata theory. For a general reference, see [38] or [22].

An *automaton* over the alphabet A is composed of a set Q of *states*, a set $E \subset Q \times A \times Q$ of *edges* or *transitions* and two sets $I, T \subset Q$ of *initial* and *terminal* states.

A *path* in the automaton $\mathcal{A}$ is a sequence

$$(p_1, a_1, p_2), (p_2, a_2, p_3), \ldots, (p_n, a_n, p_{n+1})$$

of consecutive edges. Its label is the word $x = a_1 a_2 \cdots a_n$. A path is *successful* if it starts in an initial state and ends in a terminal state. The set *recognized* by the automaton is the set of labels of its successful paths.

An automaton is *deterministic* if, for each state p and each letter a, there is at most one edge which starts at p and is labeled by a. The term *right resolving* is also used.

EXAMPLE 2 Let $\mathcal{A}$ be the automaton given in Figure 2 with 1 as unique initial and terminal state. It recognizes the set X^* where X is the prefix code $X = \{b, ab\}$.

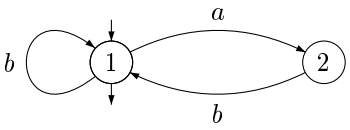

Figure 2: Golden mean automaton.

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

It is a classical result that a set of words is regular iff it can be obtained by a finite number of operations union, product and star, starting form the finite sets.

The following result is also classical (see [22] for example).

PROPOSITION 3 *Every regular set can be recognized by a finite deterministic automaton having a unique initial state.*

The following theorem is of fundamental importance. It belongs to the early folklore of automata theory.

THEOREM 2 *The length distributions of regular sets are the regular sequences.*

Proof. Let X be a regular set. By Proposition 3, it can be recognized by a deterministic automaton $\mathcal{A}$. Since $\mathcal{A}$ is deterministic, there is at most one path with given label, origin and end. Thus the number of paths of length n from the initial state to a terminal state is equal to the number u_n of words of X of length n.

Conversely, let u be a regular sequence enumerating the paths in a graph G from I to T. We consider the graph G as an automaton with all edges with distinct labels. Let X be the set of labels of paths from I to T. The sequence u is the length distribution of the set X. □

EXAMPLE 3 If $X = a^*b$, then

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

2.3 Beyond regular sequences

There are several natural classes of series beyond the rational ones. The algebraic series are those satisfying an algebraic equation. More generally,

the hypergeometric series are those such that the quotient of two successive terms is given by a rational fraction (see [26]).

The class of algebraic series is linked with the class of context-free sets (see [23]). A typical example of a context-free set is the set of words on the binary alphabet $\{a, b\}$ having as many a's as b's. We compute below its length distribution which is an algebraic series.

EXAMPLE 4 The set of words on $A = \{a, b\}$ having an equal number of occurrences of a and b is a submonoid of A^* generated by a prefix code D. Since any word of D^* of length $2n$ is obtained by choosing n positions among $2n$, we have

$$u_{D^*}(z) = \sum_{n \geq 0} \binom{2n}{n} z^{2n}.$$

By a simple application of the binomial formula, we obtain

$$u_{D^*}(z) = (1 - 4z^2)^{-\frac{1}{2}}.$$

This follows indeed, using the simple identity

$$\binom{-\frac{1}{2}}{n} = \frac{1}{(-4)^n} \binom{2n}{n}.$$

We have $u_D(z) = 1 - 1/u_{D^*}(z)$ and thus

$$u_D(z) = 1 - \sqrt{1 - 4z^2}.$$

Thus $u_D(z)$ is an algebraic series, solution of the equation

$$f^2 - 2f + 4z^2 = 0.$$

3 Enumeration on regular trees

We now turn to the study of generating sequences linked with trees. Actually, we do not enumerate trees but objects within a tree like the nodes or the leaves at each level. This is actually equivalent to the enumeration of particular sets of words, namely prefix-closed sets and prefix codes, as we shall see below (Section 4).

3.1 Graphs and trees

In this paper, we use directed multigraphs i.e. graphs with possibly several edges with the same origin and the same end. We simply call them graphs in all what follows. We denote $G = (Q, E)$ a graph with Q as set of vertices and E as set of edges. We also say that G is a graph on the set Q.

A *tree* T on a set of nodes N with a *root* $r \in N$ is a function $T : N - \{r\} \longrightarrow N$ which associates to each node distinct from the root its father $T(n)$, in such a way that, for each node n, there is a nonnegative integer h such that $T^h(n) = r$. The integer h is the *height* of the node n.

A tree is k-ary if each node has at most k children. A node without children is called a *leaf*. A node which is not a leaf is called *internal*. A node n is a *descendant* of a node m if $m = T^h(n)$ for some $h \geq 0$. A k-ary tree is *complete* if all internal nodes have exactly k children and have at least one descendant which is a leaf.

For each node n of a tree T, the *subtree* rooted at n, denoted T_n is the tree obtained by restricting the set of nodes to the descendants of n.

Two trees S, T are isomorphic, denoted $S \equiv T$, if there is a map which transforms S into T by permuting the children of each node. Equivalently, $S \equiv T$ if there is a bijective map $f : N \to M$ from the set of nodes of S onto the set of nodes of T such that $f \circ S = T \circ f$. Such a map f is called an isomorphism.

If T is a tree with N as set of nodes, the *quotient graph* of T is the graph $G = (Q, E)$ where Q and E are defined as follows. The set Q is the quotient of N by the equivalence $n \equiv m$ if $T_n \equiv T_m$. Let $\bar{m}$ denote the class of a node m. The number of edges from $\bar{m}$ to $\bar{n}$ is the number of children of m equivalent to n.

Conversely, the set of paths in a graph with given origin is a tree. Indeed, let $G = (Q, E)$ be a graph. Let $r \in Q$ be a particular vertex and let N be the set of paths in G starting at r. The tree T having N as set of nodes and such that $T(p_0, p_1, \ldots, p_n) = (p_0, p_1, \ldots, p_{n-1})$ is called the *covering tree* of G starting at r.

Both constructions are mutually inverse in the sense that any tree T is isomorphic to the covering tree of its quotient graph starting at the image of the root.

PROPOSITION 4 *Let T be a tree with root r. Let G be its quotient graph and let i be the vertex of G which is the class of the root of T. For each vertex q of G and for each $n \geq 0$, the number of paths of length n from i to q is equal to the number of nodes of T at height n in the class of q.*

A tree is said to be *regular* if it admits only a finite number of non-isomorphic subtrees, i.e. if its quotient graph is finite.

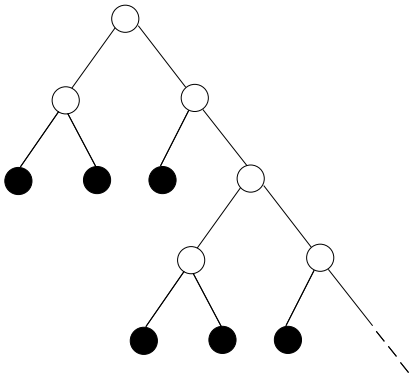

Figure 3: A regular tree.

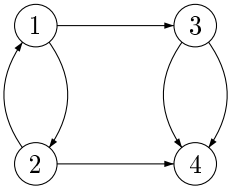

Figure 4: And its quotient graph.

For example, the infinite tree represented on Figure 3 is a regular tree. Its quotient graph is represented on Figure 4.

3.2 Regular sequences and trees

If T is a tree, its *generating sequence of leaves* is the sequence of numbers $s = (s_n)_{n\geq 0}$, where s_n is the number of leaves at height n. We also simply say that s is the *generating sequence* of T.

The following result is a direct consequence of the definitions.

THEOREM 3 *The generating sequence of a regular tree is a regular sequence.*

Proof. Let T be a regular tree and let G be its quotient graph. Since T is regular, G is finite. The leaves of T form an equivalence class t. By

Proposition 4, the generating sequence of T is recognized by (G, i, t) where i is the class of the root of T. $\square$

We say that a sequence $s = (s_n)_{n\geq 1}$ satisfies the Kraft inequality for the integer k if

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

i.e. using the formal series $s(z) = \sum_{n\geq 0} s_n z^n$, if

$$s(1/k) \leq 1.$$

We say that s satisfies the strict Kraft inequality for k if $s(1/k) < 1$. The following result is well-known (see [4] page 35 for example).

THEOREM 4 *A sequence s is the generating sequence of a k-ary tree iff it satisfies the Kraft inequality for the integer k.*

Let us consider the Kraft's equality case. If $s(1/k) = 1$, then any tree T having s as generating sequence is complete. The converse property is not true in general (see [22] p. 231). However, it is a classical result that when T is a complete regular tree, its generating sequence satisfies $s(1/k) = 1$ (see Proposition 8).

For the sake of a complete description of the construction described above in the proof of Theorem 4, we have to specify the choice made at each step among the leaves at height n. A possible policy is to choose to give as many children as possible to the nodes which are not leaves and of maximal height.

If we start with a finite sequence s satisfying Kraft's inequality, the above method builds a finite tree with generating sequence equal to s. It is not true that this incremental method gives a regular tree when we start with a regular sequence, as shown in the following example.

Let $s(z) = z^2/(1 - 2z^2)$. Since $s(1/2) = 1/2$, we may apply the Kraft construction to build a binary tree with length distribution s. The result is the tree $T(X)$ where X is the set of prefixes of the set

$$Y = \bigcup_{n\geq 0} 01^n 0\{0, 1\}^n.$$

which is not regular.

If s is a regular sequence such that $s_0 = 0$, there exists a regular tree T having s as generating sequence. Indeed, let (G, i, t) be a normalized representation of s. The generating sequence of the covering tree of G starting

at i is s. If s satisfies moreover the Kraft inequality for an integer k, it is however not true that the regular covering tree obtained is k-ary, as shown in the following example.

Let s be the regular sequence recognized by the graph of Figure 5 on the left with $i = 1$ and $t = 4$. We have $s(z) = 3z^2/(1 - z^2)$. Furthermore $s(1/2) = 1$ and thus s satisfies Kraft's equality for $k = 2$. However there are four edges going out of vertex 2 and its regular covering tree starting at 1 is 4-ary. A solution for this example is given by the graph of Figure 5 on the right. It recognizes s and its covering tree starting at 1 is the regular binary tree of Figure 3.

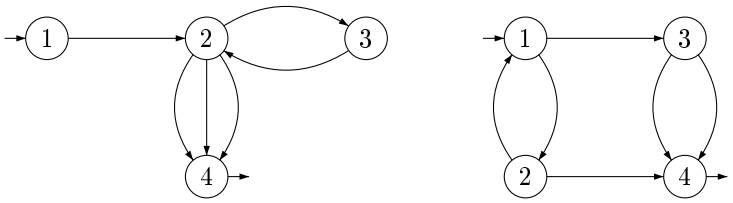

Figure 5: Graphs recognizing $s(z) = 3z^2/(1 - z^2)$.

The aim of Section 3.5 is to build from a regular sequence s that satisfies the Kraft inequality for an integer k a tree with generating sequence s which is both regular and k-ary.

3.3 Approximate eigenvector

Let M be the adjacency matrix of a graph G. By the Perron-Frobenius theorem (see [25], for a general presentation and [30], [28] or [11] for the link with graphs and regular sequences), the nonnegative matrix M has a nonnegative real eigenvalue of maximal modulus denoted by λ, also called the spectral radius of the matrix.

When G is strongly connected, the matrix is irreducible and the Perron-Frobenius theorem asserts that the dimension of the eigenspace of the matrix M corresponding to λ is equal to one, and that there is a positive eigenvector associated to λ.

Let k be an integer. A *k-approximate* eigenvector of a nonnegative matrix M is, by definition, an integral vector $\mathbf{v} \geq 0$ such that

$$M\mathbf{v} \leq k\mathbf{v}.$$

One has the following result (see [30] p. 152).

PROPOSITION 5 *An irreducible nonnegative matrix M with spectral radius λ admits a positive k-approximate eigenvector iff $k \geq \lambda$.*

For a proof, see [30] p. 152. When M is the adjacency matrix of a graph G, we also say that $\mathbf{v}$ is a k-approximate eigenvector of G. The computation of an approximate eigenvector can be obtained by the use of Franaszek's algorithm (see for example [30]). It can be shown that there exists a k-approximate eigenvector with elements bounded above by k^{2n} where n is the dimension of M [5]. Thus the size of the coefficients of a k-approximate eigenvector is bounded above by an exponential in n and can be in the worst case of this order of magnitude.

The following result is well-known. It links the radius of convergence of a sequence with the spectral radius of the associated matrix.

PROPOSITION 6 *Let s be a regular sequence recognized by a trim representation (G, I, T). Let M be the adjacency matrix of G. The radius of convergence of s is the inverse of the maximal eigenvalue of M.*

Proof. The maximal eigenvalue λ of M is $\lambda = \limsup_{n\geq 0} \sqrt[n]{\|M^n\|}$, where $\|\ \|$ is any of the equivalent matrix norms. Let ρ be the radius of convergence of s and, for each $p, q \in Q$, let ρ_{pq} be the radius of convergence of the sequence $u_{pq} = (M^n_{pq})_{n\geq 0}$. Then $1/\lambda = \min \rho_{pq}$. Since (G, I, T) is trim, we have $\rho_{pq} \geq \rho$ for all $p, q \in Q$. On the other hand, $\rho \geq \min \rho_{pq}$ since s is a sum of some of the sequences u_{pq}. Thus $\rho_s = \min \rho_{pq}$ which concludes the proof. □

As a consequence of this result, the radius of convergence ρ of a regular sequence s is a pole. Indeed, with the above notation, $s(z) = \mathbf{i}(1 - Mz)^{-1}\mathbf{t}$. Then $\det(I - Mz)$ is a denominator of the rational fraction s, the poles of s are among the inverses of the eigenvalues of M. And since $1/\lambda$ is the radius of convergence of s, it has to be a pole of s. In particular, s diverges for $z = \rho$.

The following result, due to Berstel, is also well-known. It allows one to compute the radius of convergence of the star of a sequence.

PROPOSITION 7 *Let s be a regular sequence. The radius of convergence of the series $s^*(z) = 1/(1-s(z))$ is the unique real number r such that $s(r) = 1$.*

For a proof, see [22] pp 211-214, [18] p. 82 or [11] p. 84. As a consequence, we obtain the following result.

PROPOSITION 8 *Let s be a regular sequence and let λ be the inverse of the radius of convergence of s^*. The sequence s satisfies the Kraft strict inequality $s(1/k) < 1$ (resp. equality $s(1/k) = 1$) if and only if $\lambda < k$ (resp. $\lambda = k$).*

We have thus proved the following result, which is the basis of the constructions of the next sections.

PROPOSITION 9 *Let s be a regular sequence satisfying Kraft's inequality $s(1/k) \leq 1$. Let (G, i, t) be a normalized representation of s and let $(\overline{G}, i, i)$ be the closure of (G, i, t). The adjacency matrix M of $\overline{G}$ admits a k-approximate eigenvector.*

Actually, under the hypothesis of Proposition 9, the graph G itself also admits a k-approximate eigenvector. Indeed, let $\overline{\mathbf{w}} = (\overline{w}_q)_{q \in Q-t}$ be a k-approximate eigenvector of $\overline{G}$. Then the vector $\mathbf{w} = (w_q)_{q \in Q}$ defined by $w_q = \overline{w}_q$ for $q \neq t$ and $w_t = \overline{w}_i$ is a k-approximate eigenvector of G. This is illustrated in the following example.

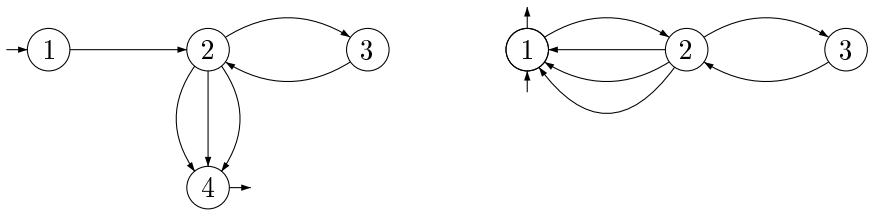

Figure 6: The graphs G and $\overline{G}$.

Let us for example consider again $s(z) = 3z^2/(1-z^2)$ (see Figure 5). The sequence s is recognized by the normalized representation $(G, 1, 4)$ where G is the graph represented on the left of Figure 6. The graph $\overline{G}$ is represented on the right. The vectors

$$\mathbf{w} = \begin{bmatrix} 3 \\ 2 \\ 1 \\ 3 \end{bmatrix}, \overline{\mathbf{w}} = \begin{bmatrix} 3 \\ 2 \\ 1 \end{bmatrix}$$

are 2-approximate eigenvectors of G and $\overline{G}$ respectively.

3.4 The multiset construction

In this section, we present the main construction used in this paper. It can be considered as a version with multiplicities of the subset construction used in automata theory to replace a finite automaton by an equivalent deterministic one. We use only unlabeled graphs but the construction can be easily generalized to graphs with edges labeled by symbols from an alphabet.

Our construction is also linked with one used by D. Lind to build a positive matrix with given spectral radius (see [30], especially Lemma 11.1.9).

We use for convenience the term *multiset* of elements of a set Q as a synonym of Q-vector. If $\mathbf{u} = (u_q)_{q\in Q}$ is such a multiset, the coefficient u_q is also called the *multiplicity* of q. The *degree* of $\mathbf{u}$ is the sum $\sum_{q\in Q} u_q$ of all multiplicities.

We start with a triple $(G, \mathbf{i}, \mathbf{t})$ where $G = (Q, E)$ is a finite graph and $\mathbf{i}$ (resp. $\mathbf{t}$) is a row (resp. column) Q-vector. We denote by M the adjacency matrix of G.

Let m be a positive integer. We define another triple $(H, \mathbf{J}, \mathbf{X})$ which is said to be obtained by the *multiset construction*. The graph H is called an *extension* of the graph G. The extension is not unique and depends as we shall see on some arbitrary choices. The set S of vertices of H is formed of multisets of elements of Q of total degree at most m. Thus, an element of S is a nonnegative vector $\mathbf{u} = (u_q)_{q\in Q}$ with indices in Q such that $\sum_{q\in Q} u_q \leq m$. This condition ensures that H is a finite graph.

We now describe the set of edges of the graph H by defining its adjacency matrix N. Let U be the $S \times Q$-matrix defined by $U_{\mathbf{u},q} = u_q$. Then N is any nonnegative $S \times S$-matrix which satisfies

$$NU = UM.$$

Equivalently, for all $\mathbf{u} \in S$,

$$\sum_{\mathbf{v}\in S} N_{\mathbf{u},\mathbf{v}}\mathbf{v} = \mathbf{u}M.$$

Let us comment informally the above formula. We can describe the construction of the graph H as a sequence of choices. If we reach a vertex $\mathbf{u}$ of H, we partition the multiset $\mathbf{u}M$ of vertices reachable from the vertices composing $\mathbf{u}$ into multisets of degree at most m to define the vertices reachable from $\mathbf{u}$ in H. The integer $N_{\mathbf{u},\mathbf{v}}$ is the multiplicity of $\mathbf{v}$ in the partition. The formula simply expresses the fact that the result is indeed a partition. In general, there are several possible partitions. The matrix U is called the *transfer matrix* of the extension.

We further define the S-row vector $\mathbf{J}$ and the S-column vector $\mathbf{X}$. Let $\mathbf{J}$ be the S-row vector such that $J_{\mathbf{i}} = 1$ and $J_{\mathbf{u}} = 0$ for $\mathbf{u} \neq \mathbf{i}$. Let $\mathbf{X}$ be the S-column vector such that $X_{\mathbf{u}} = \mathbf{u} \cdot \mathbf{t}$.

Thus

$$\mathbf{J}U = \mathbf{i}, \quad \mathbf{X} = U\mathbf{t}.$$

To avoid unnecessary complexity, we only keep in S the vertices reachable from $\mathbf{i}$. Thus, we replace the set S by the set of elements $\mathbf{u}$ of S such that there is a path from $\mathbf{i}$ to $\mathbf{u}$.

The number of multisets of degree at most m on a set Q with n elements is $\frac{n^{m+1}-1}{n-1}$. Thus the number of vertices of a multiset extension is of order n^m. It is polynomial in n if m is taken as a constant.

Figure 7: The graphs G and H.

Let for example G be the graph represented on Figure 7 on the left. The graph H represented on the right is a multiset extension of G with

$$\mathbf{i} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \quad \mathbf{j} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

The matrices M, N and U are

$$M = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}, N = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}, U = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}, \mathbf{J} = \begin{bmatrix} 1 & 0 \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

In this case, the matrix U is invertible and the matrices M, N are conjugate.

The basic property of an extension is the following one.

PROPOSITION 10 *Let H be an extension of G. The triple $(H, \mathbf{J}, \mathbf{X})$ is equivalent to $(G, \mathbf{i}, \mathbf{t})$.*

Proof. For each $n \geq 0$, we have

$$UM^n = N^n U.$$

Consequently, for each integer $n \geq 0$,

$$\begin{aligned} \mathbf{J}N^n\mathbf{X} &= \mathbf{J}N^nU\mathbf{t} \\ &= \mathbf{J}UM^n\mathbf{t} \\ &= \mathbf{i}M^n\mathbf{t}. \end{aligned}$$

This shows that $(H, \mathbf{J}, \mathbf{X})$ recognizes s. □

We will also make use of the following additional property of extensions.

PROPOSITION 11 *Let H be an extension of G. Let M (resp. N) be the adjacency matrix of G (resp. H) and let U be the transfer matrix. If $\mathbf{w}$ is a k-approximate eigenvector of M, the vector $\mathbf{W} = U\mathbf{w}$ is a k-approximate eigenvector of N. If $\mathbf{w}$ is positive, then $\mathbf{W}$ is also positive.*

Proof. We have

$$N\mathbf{W} = NU\mathbf{w} = UM\mathbf{w} \leq kU\mathbf{w} = k\mathbf{W}.$$

Since all rows of U are distinct from $\mathbf{0}$, the vector $\mathbf{W}$ is positive whenever $\mathbf{w}$ is positive. □

In the next section, we will choose a particular extension of the graph G called admissible and which is defined as follows. Let $\mathbf{w}$ be a positive Q-vector and let m be a positive integer. Let H be an extension of G, let U be the transfer matrix, and let $\mathbf{W} = U\mathbf{w}$. We say that H is *admissible* with respect to $\mathbf{w}$ and m if for each $\mathbf{u} \in S$, all but possibly one of the vertices $\mathbf{v}$ such that $(\mathbf{u}, \mathbf{v})$ is an edge of H satisfy $W_{\mathbf{v}} \equiv 0 \bmod m$.

THEOREM 5 *For any graph G on Q, any positive Q-vector $\mathbf{w}$ and any integer $m > 0$, the graph G admits an admissible extension with respect to $\mathbf{w}$ and m.*

The proof relies on the following combinatorial lemma. This lemma is also used in a similar context by Adler et al. and Marcus [34],[1]. It is actually presented in [3] as a nice variant of the pigeon-hole principle.

LEMMA 1 *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.*

Proof. The partial sums $w_1, w_1 + w_2, w_1 + w_2 + w_3, \ldots, w_1 + w_2 + \cdots + w_m$ either are all distinct (mod m), or two are congruent (mod m). In the former

case, at least one partial sum must be congruent to 0 (mod m). In the latter, there are $1 \leq p < r \leq m$ such that

$$w_1 + w_2 + \cdots + w_p \equiv w_1 + w_2 + \cdots + w_r (\bmod m)$$

Hence $w_{p+1} + w_{p+2} + \cdots + w_r \equiv 0 \pmod{m}$. □

Proof. of Theorem 5. We build progressively the set of edges of H. Let $\mathbf{u}$ be an element of S. We prove by induction on the degree $d(\mathbf{u}M) = \sum_{q \in Q} (\mathbf{u}M)_q$ of $\mathbf{u}M$ that there exists $\mathbf{v}_1, \ldots, \mathbf{v}_n \in S$ such that $\mathbf{u}M = \sum_{i=1}^{n} \mathbf{v}_i$ and $W_{\mathbf{v}_i} \equiv 0 \bmod m$ for $1 \leq i \leq n-1$. If $\mathbf{u}M \in S$, i.e. if $d(\mathbf{u}M) \leq m$, we choose $n = 1$ and $\mathbf{v}_1 = \mathbf{u}M$. Otherwise, there exists a decomposition $\mathbf{u}M = \mathbf{v} + \mathbf{u}'$ such that $d(\mathbf{v}) = m$. Let $w_1, w_2, \ldots, w_m$ be the sequence of integers formed by the w_q repeated v_q times. By Lemma 1 applied to the sequence of integers w_i, there is a decomposition $\mathbf{v} = \mathbf{v}' + \mathbf{r}$ with $\mathbf{v}' \neq 0$ such that $W_{\mathbf{v}'} \equiv 0 \bmod m$. We have $\mathbf{u}M = \mathbf{v}' + \mathbf{w}'$ with $\mathbf{w}' = \mathbf{r} + \mathbf{u}'$. Since $d(\mathbf{w}') < d(\mathbf{u}M)$, we can apply the induction hypothesis to $\mathbf{w}'$, giving the desired result. □

For an S-vector $\mathbf{W}$, we denote by $\lceil \frac{\mathbf{W}}{m} \rceil$ the S-vector $\mathbf{Z}$ such that for each $\mathbf{u}$ in S,

$$Z_{\mathbf{u}} = \lceil \frac{W_{\mathbf{u}}}{m} \rceil.$$

Summing up the previous results, we obtain the following statement.

PROPOSITION 12 *Let H be an admissible extension of G with respect to $\mathbf{w}$ and m. Let M (resp. N) be the adjacency matrix of G (resp. H), let U be the transfer matrix and let $\mathbf{W} = U\mathbf{w}$. If $\mathbf{w}$ is a positive k-approximate eigenvector of M, then $\lceil \frac{\mathbf{W}}{m} \rceil$ is a positive k-approximate eigenvector of N.*

Proof. By Proposition 3.4, the vector $\mathbf{W}$ is a positive k-approximate eigenvector of N. Thus

$$N\mathbf{W} \leq k\mathbf{W}.$$

Let $\mathbf{u}$ be an element of S. We have $W_{\mathbf{v}} \equiv 0 \bmod m$ for all indices $\mathbf{v}$ such that $N_{\mathbf{u},\mathbf{v}} > 0$ except possibly for an index $\mathbf{v}_0$. The previous inequality implies that

$$\sum_{\mathbf{v} \in S - \{\mathbf{v}_0\}} N_{\mathbf{u},\mathbf{v}} \frac{W_{\mathbf{v}}}{m} + N_{\mathbf{u},\mathbf{v}_0} \frac{W_{\mathbf{v}_0}}{m} \leq k \, \frac{W_{\mathbf{u}}}{m}.$$

Since $\frac{W_{\mathbf{v}}}{m}$ is a nonnegative integer for $\mathbf{v} \in Q - \{\mathbf{v}_0\}$, we get

$$\sum_{\mathbf{v} \in S - \{\mathbf{v}_0\}} N_{\mathbf{u},\mathbf{v}} \frac{W_{\mathbf{v}}}{m} + N_{\mathbf{u},\mathbf{v}_0} \lceil \frac{W_{\mathbf{v}_0}}{m} \rceil \leq k \lceil \frac{W_{\mathbf{u}}}{m} \rceil.$$

This proves that

$$N \lceil \frac{\mathbf{W}}{m} \rceil \leq k \lceil \frac{\mathbf{W}}{m} \rceil.$$

□

3.5 Generating sequence of leaves

In what follows, we show how the multiset construction allows one to prove the main result of [10] concerning the generating sequences of regular trees. We begin with the following lemma, which is also used in the next section. We use the term leaf for a vertex of a graph without outgoing edges.

LEMMA 2 *Let G be a graph on a set Q of vertices. Let $i \in Q$ and $T \subset Q$. If G admits a k-approximate eigenvector $\mathbf{w}$, there is a graph G' and a set of vertices I' of G' such that*

1. *G' admits the k-approximate eigenvector $\mathbf{w}'$ with all components equal to 1.*
2. *the triple $(G, i, \mathbf{w})$ is equivalent to the triple $(G', I', \mathbf{w}')$;*
3. *If $w_p = 1$ for all $p \in T$, there is a set of vertices T' of G' such that the triple (G, i, T) is equivalent to the triple (G', I', T'). Moreover, if T is the set of leaves of G, we can choose for T' the set of leaves of G'.*

We now state the main result of [10].

THEOREM 6 *Let $s = (s_n)_{n \geq 0}$ be a regular sequence of nonnegative integers and let k be a positive integer such that $\sum_{n \geq 0} s_n k^{-n} \leq 1$. Then there is a k-ary rational tree having s as its generating sequence.*

Proof. Let us consider a regular sequence s and an integer k such that $\sum_{n \geq 0} s_n k^{-n} \leq 1$. Since the result holds trivially for $s(z) = 1$, we may suppose that $s_0 = 0$. Let (G, i, t) be a normalized representation of s and let $\overline{G}$ be the closure of G as defined at the beginning of Section 2.1. We denote by M (resp. $\overline{M}$) the adjacency matrix of G (resp. $\overline{G}$). Let $\overline{Q} = Q - \{t\}$

be the vertex set of $\overline{G}$. Let λ be the spectral radius of $\overline{M}$. By Proposition 8, the matrix $\overline{M}$ admits a positive k-approximate eigenvector $\overline{\mathbf{w}}$. By definition, we have $\overline{M}\overline{\mathbf{w}} \leq k\overline{\mathbf{w}}$.

Let $\mathbf{w}$ be the Q-vector defined by $w_q = \overline{w}_q$ for all $q \in \overline{Q}$ and $w_t = \overline{w}_i$. Then, since there is no edge going out of t in G, $\mathbf{w}$ is a positive k-approximate eigenvector of M. Let $\mathbf{t}$ be the Q-vector which is the characteristic vector of the vertex t. Let $m = w_i$.

By Theorem 5 there exists an admissible extension H of G with respect to $\mathbf{w}$ and m. Let U be the transfer matrix and let $\mathbf{W} = U\mathbf{w}$. Since $w_t \equiv 0 \bmod m$, we may choose H with the following additional property. For all $\mathbf{u} \in S$ either $u_t = 0$ or $\mathbf{u} = \mathbf{t}$.

According to Proposition 10, the sequence s is recognized by $(H, \mathbf{J}, \mathbf{X})$ where $\mathbf{J}$ is the characteristic row vector of $\mathbf{i}$ and $\mathbf{X}$ is the characteristic column vector of $\mathbf{t}$. This means that s is recognized by the normalized representation consisting in the graph H, the initial vertex $\mathbf{i}$, that we identify to i, and the terminal vertex $\mathbf{t}$, that we identify to t.

Let N be the adjacency matrix of H. By Proposition 12, the vector $\lceil \frac{\mathbf{W}}{m} \rceil$ is a positive k-approximate eigenvector of N. Remark that $\lceil \frac{\mathbf{W}}{m} \rceil_i = \lceil \frac{\mathbf{W}}{m} \rceil_t = 1$.

We may now apply Lemma 2 to construct a triple (H', I', T') equivalent to (H, i, t). The set T' is the set of leaves of H'. Since $\lceil \frac{\mathbf{W}}{m} \rceil_i = 1$, I' is reduced to one vertex i'. Since H' admits a k-approximate eigenvector with all components equal to one, the graph H' is of outdegree at most k. Finally s is the generating sequence of the covering tree of H' starting at i'. This tree is k-ary and regular. □

Let us consider the above constructions in the particular case of the equality in Kraft's inequality. In this case, the result is a complete k-ary tree. Indeed, by Proposition 8, the matrix $\overline{M}$ admits a positive integral eigenvector $\mathbf{w}$ for the eigenvalue k. We have for all $p \in \overline{Q}$,

$$\sum_{q \in \overline{Q}} M_{p,q} w_q = k w_p.$$

As a consequence, for any $\mathbf{u} \neq \mathbf{t}$, we have

$$\sum_{\mathbf{v} \in S} N_{\mathbf{u},\mathbf{v}} W_{\mathbf{v}} = k W_{\mathbf{u}}.$$

Then the graph constructed in Lemma 2 is of constant outdegree k. Thus the k-ary tree obtained is complete.

Let us consider the complexity of the construction used in the proof of Theorem 6. Let n be the number of vertices of the graph G giving a normalized representation of s. The size of the integer $m = w_i$ is exponential in n (see Section 3.3). Thus the number of vertices of the graph H is bounded by a double exponential in n. The final regular tree is the covering tree of a graph whose set of vertices has the same size in order of magnitude.

Let for example s be the sequence defined by

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

Since $s(1/2) = 1$, it satisfies the Kraft equality for $k = 2$. The sequence s is recognized by (G, i, t) where $G = (Q, E)$ is the graph given in Figure 3.5 with $Q = \{1, 2, 3, 4, 5, 6, 7\}$, $i = 1$, $t = 4$. The adjacency matrix of G admits the 2-approximate eigenvector represented on Figure 3.5, where the coefficients of $\mathbf{w}$ are represented in squares beside the vertices. Thus $m = 3$.

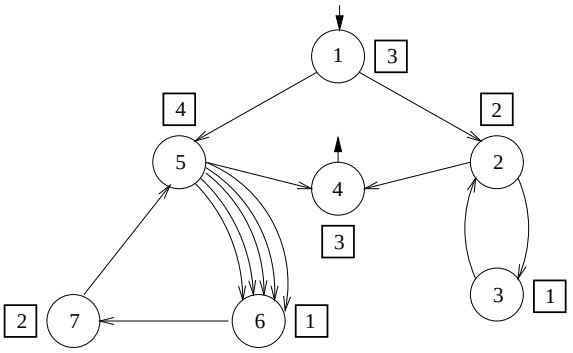

Figure 8: A normalized representation of s

An admissible extension H of G with respect to $\mathbf{w}$ and m is given in Figure 9. In this figure, each multiset of S is represented by a sequence of vertices with repetitions corresponding to the multiplicity. For example, the multiset $\mathbf{u} = (0, 0, 1, 0, 0, 2, 0)$ is represented by $(3, 6, 6)$. The sequence s is recognized by the normalized representation $(H, 1, 4)$, where the initial and final vertices are named as they appear on Figure 9. The coefficients of $\mathbf{W}$ are represented in squares beside the vertices.

A regular binary tree T having s as generating sequence of leaves, is given in Figure 10. In this figure, the nodes have been renumbered, with the children of a node with a given label represented only once. The leaves

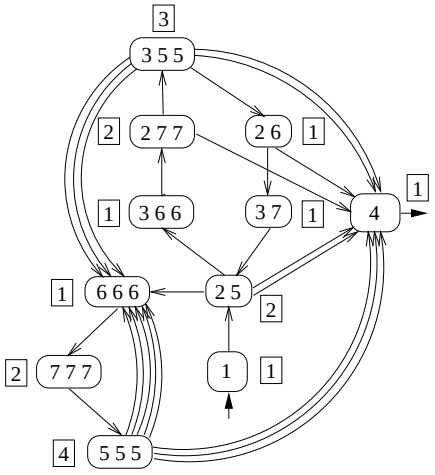

Figure 9: An admissible extension H.

of the tree are indicated by black boxes. The tree itself is obtained from the graph of Figure 9 by application of the construction of Lemma 2. For example, the vertex $(2, 5)$, which has coefficient 6 in $\mathbf{W}$, is split into two vertices named 2 and 3 in the tree.

This example was suggested to us by Christophe Reutenauer [39]. To check directly that the length distribution is equal to $s(z)$, one may compute from the graph the following regular expression of $s(z)$ and check by an elementary computation (possibly with the help of a symbolic computation system) that it is equal to $s(z)$.

$$s(z) = (z^6)^*(2z^2 + z^4 + 2z^5 + z^6 + (z^2 + 3z^5)(5z^3)^*3z^3). \qquad (1)$$

(note for a reader unfamiliar with regular expressions: the first factor $(z^6)^*$ corresponds to the vertex labeled 1 at level 6 of the tree. The term $2z^2 + z^4 + 2z^5 + z^6$ corresponds to the leaves reached by a path which does not use a vertex labeled 5. The factor $(z^2 + 3z^5)(5z^3)^*$ corresponds to the paths from the root to a vertex labeled 5. Finally, the factor $3z^3$ corresponds to the direct paths from 5 to a leaf.)

This example shows an interesting feature of this problem. In fact, from the point of view of regular expressions, the difficult operation in this problem is the sum. It would be a simple matter to build a rational tree for each term of the sum in the expression (1) (see the example of Figure 5). The difficulty would then be to merge these trees to obtain one corresponding to the sum.

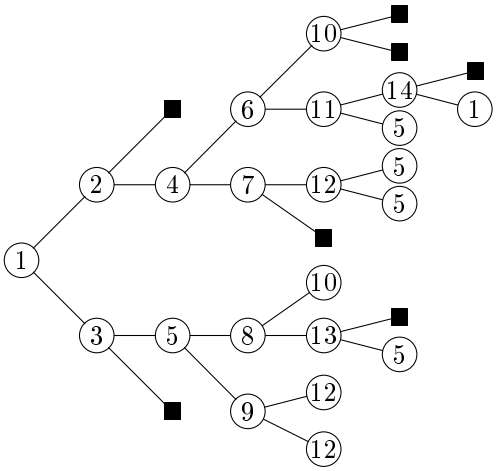

Figure 10: A regular binary tree with length distribution s.

A curious consequence of Theorem 6 is the following property of regular sequences.

COROLLARY 1 *Let $k \geq 2$ be an integer and let u be a regular sequence such that $u(1/k) \leq 1$ and $u(0) = 0$. Then there exist k regular sequences $u_1, \ldots, u_k$ such that $u_i(1/k) \leq 1$ and*

$$u(z) = \sum_{i=1}^{k} z u_i(z).$$

Proof. It is a simple consequence of Theorem 6. Indeed, if X is a regular prefix code on the k element alphabet A, then $X = \sum_{a \in A} aX_a$ where each X_a is a regular prefix code on the alphabet A. □

We don't know of a direct proof of this result.

3.6 Generating sequence of nodes

In this section, we consider the generating sequence of the set of all nodes in a tree instead of just the set of leaves. This is motivated by the fact that in search trees, the information can either be carried by the leaves or by all the nodes of the tree. We will see that the complete characterization

of the generating sequences of nodes in regular trees (Theorem 7) is more complicated than the one for leaves.

Soittola (see [42] p. 104) has characterized the series which are the generating sequences of nodes in a regular tree. We characterize the ones that correspond to k-ary trees (Theorem 7). We also give a more direct construction in a particular case (Theorem 8).

Let T be a tree. The generating sequence of nodes of the tree T is the sequence $t = (t_n)_{n\geq 0}$, where t_n is the number of nodes of T at height n. The sequence t satisfies $t_0 \leq 1$ and, moreover, if T is a k-ary tree, the condition

$$t_n \leq kt_{n-1}$$

for all $n \geq 1$. If T is a regular tree, then t is a regular sequence. We now completely characterize the regular sequences t that are the generating sequences of nodes of a k-ary regular tree.

THEOREM 7 *Let $t = (t_n)_{n\geq 0}$ be a regular sequence and let k be a positive integer. The sequence $(t_n)_{n\geq 0}$ is the generating sequence of nodes of a k-ary regular tree iff it satisfies the following conditions.*

(i) *the convergence radius of t is strictly greater than $1/k$,*

(ii) *the sequence $s(z) = t(z)(kz - 1) + 1$ is regular.*

Proof. Let us first show that the conditions are necessary. Let $\overline{T}$ be the complete k-ary tree obtained by adding i new leaves to each node that has $k - i$ children. Since T is a regular tree, $\overline{T}$ is also regular.

Let s be the generating sequence of leaves of $\overline{T}$. Since $\overline{T}$ is complete, $s(1/k) = 1$. Since $kt_n = s_{n+1} + t_{n+1}$ for all $n \geq 0$, we have

$$1 - s(z) = t(z)(1 - kz).$$

Since s is a regular sequence, its radius of convergence is strictly larger than $1/k$ (see Section 3.3). Since the value of the derivative of s at $z = 1/k$ is $kt(1/k)$, the same holds for t. This proves the necessity of the conditions.

Conversely, if t satisfies the conditions of the theorem, the regular series $s(z) = t(z)(kz - 1) + 1$ satisfies $s(1/k) = 1$. Thus, by Theorem 6, s is the generating sequence of leaves of a complete k-ary regular tree. The internal nodes of this tree form a k-ary regular tree whose generating sequence of nodes is t. □

The sequence s defined by condition (ii) is rational as soon as t is regular and therefore rational. Given a regular sequence t, condition (ii) is decidable in view of the theorem of Soittola (Theorem 1).

We may observe that condition (ii) of the theorem implies the nonnegativity of the coefficients of the series s and thus the inequality $\forall n \geq 1, t_n \leq kt_{n-1}$. It also implies that $t_0 \leq 1$.

We now show that there are regular sequences t satisfying $t_n \leq kt_{n-1}$ for all $n \geq 1$, and condition (i) of the theorem and such that the sequence $s(z) = t(z)(kz-1)+1$ is not regular. The example is based on an example of a rational sequence with nonnegative coefficients and which is not regular (see [18] page 95). Let

$$r_n = b^{2n} cos^2(n\theta)$$

with $cos(\theta) = \frac{a}{b}$ where the integers a, b are such that $b \neq 2a$ and $0 < a < b$. The sequence r is rational, has nonnegative integer coefficients and is not regular. Its poles are $\frac{1}{b^2}$, $\frac{1}{b^2}e^{2i\theta}$ and $\frac{1}{b^2}e^{-2i\theta}$. We now define the sequence t as follows:

$$\begin{aligned} t_{2h} &= k^h, \\ t_{2h+1} &= k^h + r_h. \end{aligned}$$

We also assume that $b^2 < k$. By Soittola's theorem, the sequence t is regular since it is a merge of rational sequences having a dominating root. The convergence radius of t is $\frac{1}{\sqrt{k}} > \frac{1}{k}$. Therefore the sequence t satisfies the first condition of Theorem 7. Let s be the sequence defined by $s(z) = t(z)(kz-1)+1$. If $h = 2p$ is even,

$$\begin{aligned} s_h &= kt_{h-1} - t_h \\ &= kk^{p-1} + kr_{p-1} - k^p + 1 = kr_{p-1} + 1. \end{aligned}$$

Thus the sequence s is not regular.

The above example does not work for the small values of k (the least value is $k = 10$). We do not know of similar examples for $2 \leq k \leq 9$.

We finally describe a particular case of Theorem 7 in which one has a relatively simple method, based on the multiset construction, to build the regular tree with a given generating sequence of nodes. This avoids the use of Soittola's characterization which leads to a method of higher complexity.

A *primitive* representation of a regular sequence s is a representation $(G, \mathbf{i}, \mathbf{t})$ such that the adjacency matrix of G is primitive. The following result is proved in [8] with a different proof using the state-splitting method of symbolic dynamics. The proof given in [10] relies on a simpler construction.

THEOREM 8 *Let $t = (t_n)_{n\geq 0}$ be a regular sequence and let k be a positive integer such that $t_0 = 1$, $t_n \leq kt_{n-1}$ for all $n \geq 1$ and such that*

(i) *the convergence radius of t is strictly greater than $1/k$,*

(ii) *t has a primitive representation.*

Then $(t_n)_{n\geq 0}$ is the generating sequence of nodes by height of a k-ary regular tree.

The proof of this theorem given in [10] uses the multiset construction. It relies on the following lemma.

LEMMA 3 *Let M be a primitive matrix with spectral radius λ. Let $\mathbf{v}$ be a non-null and nonnegative integral vector and let k be an integer such that $\lambda < k$. Then there is a positive integer n such that $M^n\mathbf{v}$ is a positive k-approximate eigenvector of M.*

Proof. For a primitive matrix M with spectral radius λ, it is known that the sequence $((\frac{M}{\lambda})^n)_{n\geq 0}$ converges to $\mathbf{r}.\mathbf{l}$ where $\mathbf{r}$ is a positive right eigenvector and $\mathbf{l}$ a positive left eigenvector of M for the eigenvalue λ with $\mathbf{l}\cdot\mathbf{r} = 1$ (see for example [30] p. 130). Thus $(\frac{M^n}{\lambda^n}\mathbf{v})_{n\geq 0}$ converges to $\mathbf{r}.\mathbf{l}.\mathbf{v}$ which is equal to $\rho\mathbf{r}$ where ρ is a nonnegative real number. Since $M\mathbf{r} = \lambda\mathbf{r}$, we get, for a large enough integer n,

$$M\frac{M^n}{\lambda^n}\mathbf{v} \leq k\frac{M^n}{\lambda^n}\mathbf{v}$$

or equivalently $MM^n\mathbf{v} \leq kM^n\mathbf{v}$. If n is large enough, we moreover have $M^n\mathbf{v} > 0$ since M is primitive. □

The proof of Theorem 8 uses a shift of indices of the sequence to obtain a new sequence to which a simple application of the multiset construction can be applied. We illustrate it on an example.

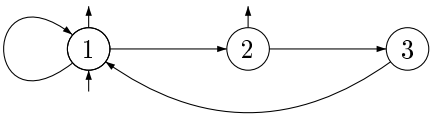

Figure 11: A primitive representation G of t.

Let t be the series recognized by the graph G of Figure 11 with

$$\mathbf{i} = \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} \text{ and } \mathbf{t} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}.$$

The adjacency matrix M of G is the primitive matrix

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

Its spectral radius is less than 2. The hypothesis of Theorem 8 are thus satisfied. We have

$$M^2\mathbf{t} = \begin{bmatrix} 2 \\ 1 \\ 2 \end{bmatrix} \text{ and } M^3\mathbf{t} = \begin{bmatrix} 3 \\ 2 \\ 2 \end{bmatrix}.$$

Since $M^3\mathbf{t} \leq 2M^2\mathbf{t}$, the vector $\mathbf{W} = M^2\mathbf{t}$ is an approximate eigenvector of M (the existence of such a vector is asserted by Lemma 3). Let $\mathbf{w} = M^2\mathbf{t}$.

Applying Lemma 2, we obtain from G the graph G' represented on the left side of Figure 12 . Moreover, $(G, \mathbf{i}, \mathbf{w})$ is equivalent to $(G', I', \mathbf{w}')$ where I' is the set of initial vertices indicated on Figure 12 and $\mathbf{w}$ is the vector with all components equal to 1. The covering trees $T_{1,1}$ and $T_{1,2}$ of G' starting at the vertices of I' give, with the appropriate shift of indices, the binary regular tree T represented on the right side of Figure 12 (the nodes of the tree have been renumbered).

4 Generating sequences of prefix codes

There is a close connexion between trees and prefix codes or prefix-closed sets of words. We present below the translation of some of the notions and results seen before in terms of prefix codes.

4.1 Trees and prefix codes

Let R be a set of words on the alphabet $A = \{0, 1, \ldots, k-1\}$. The set R is said to be *prefix-closed* if any prefix of an element of R is also in R. The set X of words which are not a proper prefix of a word in R is a prefix code, called the prefix code associated to R.

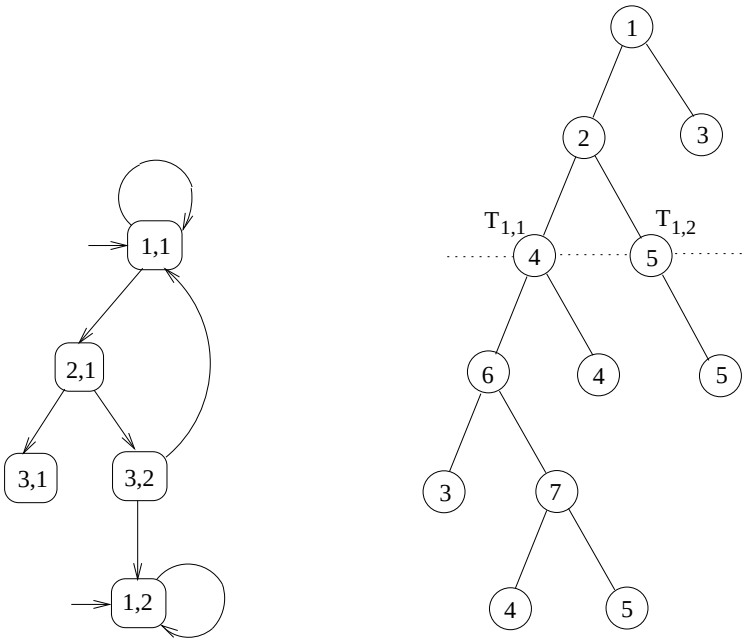

Figure 12: The graph G' and the tree T.

When R is prefix closed, we can build a tree $T(R)$ as follows. The set of nodes is R, the root is the empty word ϵ and $T(a_1a_2\cdots a_n) = a_1a_2\cdots a_{n-1}$. The leaves of T from a prefix code which is the prefix code associated to R. The generating sequence of T is the gerating sequence of X.

Let for example $R = \{\epsilon, 0, 1, 10, 11\}$. The tree $T(R)$ is represented on Figure 13. The associated prefix code is $X = \{0, 10, 11\}$.

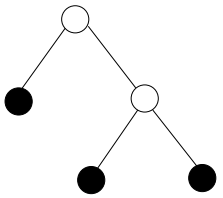

Figure 13: The tree $T(X)$.

Let X be a prefix code on an alphabet with k symbols. It is clear that

its length distribution $u = (u_n)_{n\geq 1}$ satisfies Kraft's inequality

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

or equivalently $u(1/k) \leq 1$. The number $u(1/k)$ can actually be interpreted as the probability that a long enough word has a prefix in X.

There is also a connexion with the notion of entropy. Actually, if X is a prefix code, the entropy of X^* is equal to $\log(1/\rho)$ where ρ is the solution of the equation $u_X(\rho) = 1$. Thus Kraft's inequality expresses the fact that $h(X^*) \leq \log k$.

Conversely, Kraft-McMillan's theorem states that for any such sequence $u = (u_n)_{n\geq 1}$, there exists a prefix code X on a k-symbol alphabet such that $u = u_X$.

The equality case in Kraft's inequality corresponds to a particular class of prefix codes often called *complete*. A prefix code X on the alphabet A is complete if any word on A has either a prefix in X or is a prefix of a word of X.

Theorem 6 shows that the generating sequences of regular prefix codes are exactly the regular sequences satisfying Kraft's inequality.

4.2 Bifix codes

We investigate here the length distributions of a particular class of prefix codes, called bifix. Several other classes of prefix codes could give rise to a similar study (for a description to these classes, see [21]).

The definition of a suffix code is symmetric to the definition of a prefix code. It is a set of words X such that no element of X is a suffix of another one. The notion of a complete suffix code is also symmetric. A *bifix code* is a set X of words which is both a prefix and a suffix code.

Any set of words of fixed length is obviously a bifix code but there are more complicated examples.

Example 5 The set

$$X = \{aaa, aaba, aabb, ab, baa, baba, babb, bba, bbb\}$$

is a complete prefix code pictured in Figure 14. It is also a complete suffix code as one may check by reading its words backwards.

Surprisingly, it is an open problem to characterize the length distributions of bifix codes. The following simple example shows that they are more constrained than those of prefix codes.

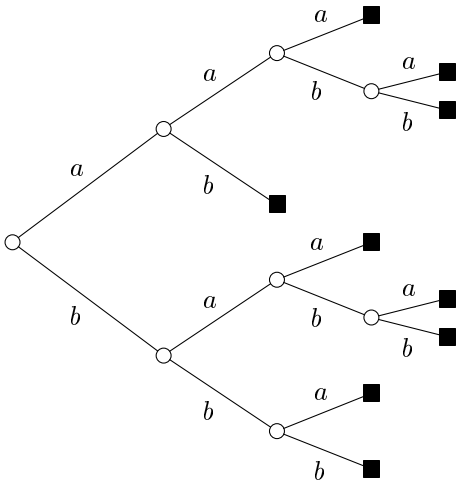

Figure 14: The bifix code X.

EXAMPLE 6 The sequence $u(z) = z + 2z^2$ is not realizable as the length distribution of a bifix code on a binary alphabet although $u(1/2) = 1$. Indeed, one of the symbols has to be in X, say a. Then bb is the only word of length 2 that can be added.

The following nice partial result is due to Ahlswede, Balkenhol and Khachatrian [2]. We state the result for a binary alphabet. It can be readily generalized to k symbols but it presents less interest.

THEOREM 9 *For any integer sequence u such that*

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

there is a bifix code X such that $u = u_X$.

Proof. The proof is by induction. We suppose that we have already built a bifix code X formed of words of length at most $n-1$ with length distribution $(u_1, u_2, \ldots, u_{n-1})$. We have

$$\sum_{i=1}^{n} u_i 2^{-i} \leq 1/2,$$

and thus

$$2\sum_{i=1}^{n} u_i 2^{n-i} \leq 2^n.$$

Finally, we obtain

$$u_n \leq 2^n - 2\sum_{i=1}^{n-1} u_i 2^{n-i}.$$

The expression of the right handside is at most equal to the number of elements of the set $A^n - XA^* - A^*X$. Thus, we can choose u_n words of length n which do not have a prefix or a suffix in X. This proves the result by induction. □

The authors of [2] formulate the interesting conjecture that Theorem 9 is still true if the hypothesis $u(1/2) \leq 1/2$ is replaced by $u(1/2) \leq 3/4$.

There are known additional conditions imposed on length distributions of bifix codes. For example, one has the following result, originally due to Schützenberger (see [16]).

THEOREM 10 *If X is a finite complete bifix code on k symbols, then $u_X(1/k) = 1$ and $\frac{1}{k}u'_X(1/k)$ is an integer.*

The number $\frac{1}{k}u'_X(1/k)$ can be interpreted as the average length of the words of X. Indeed

$$zu'_X(z) = \sum_{x \in X} |x| z^{|x|}.$$

EXAMPLE 7 For the bifix code of Example 5, we have

$$u_X(z) = z^2 + 4z^3 + 4z^4$$

and thus

$$u'_X(z) = 2z + 12z^2 + 16z^3.$$

Hence $\frac{1}{2}u'_X(1/2) = 3$.

The conditions of Theorem 10 show directly that the sequence of Example 6 is not realizable. Indeed, it satisfies the first condition but not the second one. The conditions of Theorem 10 are not sufficient. Indeed, if $u(z) = z + 4z^3$ we have $u(1/2) = 1$ and $u'(1/2) = 4$ although it is clearly impossible that $u = u_X$ for a bifix code X.

Recently, Ye and Yeung [45] have made some progress on this problem. They are in particular able to prove that Theorem 9 still holds when $u(1/2) \leq 5/8$.

5 Zeta functions, subshifts of finite type and circular codes

In this section, we present a number of results on interrelated objects which are connected with cyclic permutation of words. The link with enumerative combinatorics was developped in Lothaire's volume [31] and later in R. Stanley's book [44]. We begin with notions classical in symbolic dynamics (see [30] or [28] for a general reference; see [15] or [24] for the link with finite automata).

5.1 Subshifts of finite type

A *subshift* is a set of biinfinite words on a finite alphabet A which avoids a given set F of forbidden words. It is a topological space as a closed subset of the space $A^{\mathbb{Z}}$ of functions from $\mathbb{Z}$ into the set A. The *full shift* on A is the set of all biinfinite words on A. It corresponds to the case $F = \emptyset$.

A *sofic* subshift is the set of biinfinite labels of paths in a finite automaton. A sofic subshift is called *irreducible* if the automaton can be chosen strongly connected. A *subshift of finite type* is the set of biinfinite words avoiding a finite set of finite words. Any subshift of finite type is sofic but the converse is not true. The *edge shift* of a finite graph G is the set S_G of biinfinite paths in G (viewed as biinfinite sequences of edges). It is a subshift of finite type.

The *shift* σ is the function on a subshift S which maps a point x to the point $y = \sigma(x)$ whose ith coordinate is $y_i = x_{i+1}$.

A *morphism* from a subshift S into a subshift T is a function $f : S \to T$ which is continuous and invariant under the shift. A bijective morphism is called a *conjugacy*. Any subshift of finite type is conjugate to some edge shift.

The *entropy* $h(S)$ of a subshift S is the entropy of the formal language formed by the finite blocks occurring in words of S. It can be shown that the entropy is a topological invariant, in the sense that two conjugate subshifts have the same entropy.

While the entropy is a measure of number of forbidden words, it is possible to study the number of minimal forbidden words. It gives rise to another invariant of subshifts [13], [14].

An integer p is a *period* of a point $x = (a_n)_{n\in\mathbb{Z}}$ if $a_{n+p} = a_n$ for all $n \in \mathbb{Z}$. Equivalently, p is a period of x if $\sigma^p(x) = x$. The *zeta function* of a subshift

S, is defined as the series

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

where p_n is the number of words with period n in S. It is also a topological invariant, since a point of period n is mapped by a conjugacy on a point of the same period.

The following result due to Bowen and Lanford [19] is classical (see [30]).

PROPOSITION 13 *Let G be a finite graph and let M be the adjacency matrix of G. Then*

$$\zeta(S_G) = \det(I - Mz)^{-1}.$$

Proof. We first have for each $n \geq 1$

$$\mathrm{Tr}(M^n) = p_n$$

since the coefficient (i, j) of M^n is the number of paths from i to j. Thus

$$\begin{aligned} \zeta(S_G) &= \exp \sum_{n \geq 1} \frac{p_n}{n} z^n \\ &= \exp \sum_{n \geq 1} \frac{\mathrm{Tr}(M^n)}{n} z^n \\ &= \exp \mathrm{Tr}(\log(I - Mz)^{-1}) \\ &= \det(I - Mz)^{-1} \end{aligned}$$

since, by the formula of Jacobi, $\exp \mathrm{Tr} = \det \exp$. □

EXAMPLE 8 Let S be the edge shift of the graph G of Figure 15. We have

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

Consequently

$$\zeta(S) = \frac{1}{1 - z - z^3}.$$

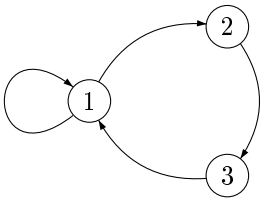

Figure 15: A subshift of finite type

Let S be a subshift of finite type and let p_n be the number of points with period n. Let q_n be the number of points with least period n. Since q_n is a multiple of n, we also denote $q_n = nl_n$. We have then the formula expressing the zeta function as an infinite product using the integers l_n as exponents.

$$\zeta(S) = \Pi_{n\geq 1}(1 - z^n)^{-l_n},$$

as one may verify using $p_n = \sum_{d|n} dl_d$ and the definition of $\zeta(S)$.

A classical result, related with what follows, is the following statement, known as Krieger's embedding theorem.

THEOREM 11 *Let S, T be two subshifts of finite type. There exists an injective morphism $f : S \to T$ with $f(S) \neq T$ iff*

1. *$h(S) < h(T)$*
2. *for each $n \geq 1$, $q_n(S) \leq q_n(T)$ where $q_n(S)$ (resp. $q_n(T)$) is the number of points of S (resp. T) of least period n.*

The following result is the basis of many applications of symbolic dynamics to coding. It is due to Adler, Coppersmith and Hassner [1].

THEOREM 12 *If S is an irreducible subshift of finite type such that $h(S) \geq \log k$, it is conjugate to a subshift of finite type S_G where the graph G has outdegree at least k.*

The proof is based on a state-splitting algorithm using approximate eigenvectors and Lemma 1. This result is part of a number of constructions leading to sliding block codes used in magnetic recording (see [35], [11] or [30]). It gives at the same time the following result.

THEOREM 13 *It S is a subshift of finite type such that $h(S) \leq \log k$, then there is a graph G of outdegree at most k such that S is conjugate to S_G.*

There is a connexion between this theorem and Theorem 6. Let indeed u be a regular sequence of integers such that $u(1/k) \leq 1$. Let G be a normalized graph recognizing u (in the sense of Section 2.1). Let $\bar{G}$ be the graph obtained by merging the initial and terminal vertex. Then $h(S_{\bar{G}}) \leq \log k$. We can apply Theorem 13 to obtain a graph H with outdegree at most k such that S_G and S_H are conjugate. This gives the conclusion of Theorem 6 provided the initial-terminal vertex did not split in the construction. The following examples show both cases (for details, see [7] and [8]).

EXAMPLE 9 Let G be the graph of Figure 5. The splitting of vertex 2 gives a graph of outdegree 2. A normalization gives the automaton on the right.

EXAMPLE 10 The sequence of the example given in Figure 6 is recognized by a graph G such that $\bar{G}$ has three cycles of length 2. The solution as a binary tree has only two cycles of length 2 and thus could not be obtained by state-splitting.

5.2 Circular codes

A *circular word*, or necklace, is the equivalence class of a word under cyclic permutation. For a word w, we denote by $\bar{w}$ the circular word represented by w.

Let X be a set of words and $w = x_1 x_2 \cdots x_n$ with $x_i \in X$. The set of cyclic permutations of the sequence $(x_1, x_2, \ldots, x_n)$ is called a factorization of the circular word $\bar{w}$.

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

EXAMPLE 11 The set $X = \{a, aba\}$ is a circular code. Indeed, the position of the symbols b determines uniquely the occurrences of aba.

EXAMPLE 12 The set $X = \{ab, ba\}$ is not a circular code. Indeed, the circular word $\bar{w}$ for $w = abab$ has two factorizations namely (ab, ab) and (ba, ba).

The following characterization is useful (see [16]).

PROPOSITION 14 *A set X is a circular code if and only if it is a code and for all $u, v \in A^*$,*

$$uv, vu \in X^* \Rightarrow u, v \in X^*$$

EXAMPLE 13 We obtain another way to prove that the set $X = \{ab, ba\}$ is not a circular code. Indeed, otherwise we would have $a, b \in X^*$ which is contradictory.

Let X be a finite code. The *flower automaton* of X, denoted $\mathcal{A}_X$, is the following automaton. The set of its states is

$$Q = \{(u, v) \in A^+ \times A^+ \mid uv \in X\} \cup (1, 1)$$

The transitions are of the form $(u, av) \xrightarrow{a} (ua, v)$ or $(1, 1) \xrightarrow{a} (a, v)$ or $(u, a) \xrightarrow{a} (1, 1)$. The unique initial and final state is $(1, 1)$.

EXAMPLE 14 The flower automaton of the circular code $\{a, aba\}$ is pictured in Figure 16.

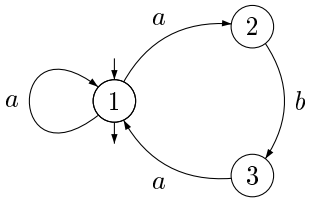

Figure 16: The flower automaton of $\{a, aba\}$.

The following result is easy to prove.

PROPOSITION 15 *The flower automaton $\mathcal{A}_X$ recognizes X^*. The code X is circular iff for each word w, there is at most one cycle with label w.*

We now study the length distributions of circular codes. Let X be a circular code and let $u_(z) = (u_n)_{n \geq 1}$ be its length distribution. For each $n \geq 1$, let p_n be the number of words w of length n such that $\bar{w}$ has a factorization in words of X.

PROPOSITION 16 *The sequences $(p_n)_{n \geq 1}$ and $(u_n)_{n \geq 1}$ are related by*

$$\exp \sum_{n \geq 1} \frac{p_n}{n} z^n = \frac{1}{1 - u(z)}. \tag{2}$$

Proof. Each (p_n) depends only on the first n terms of the sequence (u_n). It is therefore possible to suppose that the sequence (u_n) is finite, i.e. that the code X is finite. Let $\mathcal{A}$ be the flower automaton of X. Let S be the subshift of finite type associated with the graph of $\mathcal{A}$. Then p_n is the number of elements of period n in S. Indeed, each word w such that $\bar{w}$ has a factorization is counted exactly once as the label of a cycle in $\mathcal{A}$. We have also

$$\det(I - Mz) = 1 - u(z).$$

Thus, the result follows from Proposition 13. □

The explicit relation between the numbers u_n and p_n is the following. For each $i \geq 1$, let $u^{(i)} = (u_n^{(i)})_{n\geq 1}$ be the length distribution of X^i. Equivalently, $u_n^{(i)}$ is the coefficient of degree n of $u(z)^i$. Then for each $n \geq 1$

$$p_n = \sum_{i=1}^{n} \frac{n}{i} u_n^{(i)}.$$

We also have for each $n \geq 1$

$$p_n = nu_n + \sum_{i=1}^{n-1} p_i u_{n-i}. \tag{3}$$

This formula can be easily deduced from Formula (2) by taking the logarithmic derivative of each side of the formula. It shows directly that for any sequence $(u_n)_{n\geq 1}$ of nonnegative integers, the sequence p_n defined by Formula (2) is formed of nonnegative integers.

Formula (3) is known as Newton's formula in the field of symmetric functions. Actually, the numbers u_n can be considered, up to the sign, as elementary symmetric functions and the p_n as the sums of powers (see [32]). The link between Witt vectors and symmetric functions was established in [43].

Let $p_n = \sum_{d|n} dl_d$. Then l_n is the number of non-periodic circular words of length n with a factorization. In terms of generating series, we have

$$\exp \sum_{n\geq 1} \frac{p_n}{n} z^n = \prod_{n\geq 1} (1 - z^n)^{-l_n}. \tag{4}$$

Putting together Formulae (2) and (4), we obtain

$$\frac{1}{1 - u(z)} = \prod_{n\geq 1} (1 - z^n)^{-l_n}. \tag{5}$$

For any sequence $(u_n)_{n\geq 1}$ of nonnegative integers, the sequence $l = (l_n)_{n\geq 1}$ thus defined is formed of nonnegative integers. This can be proved either by a direct computation or by a combinatorial argument since any sequence u of nonnegative integers is the length distribution of a circular code on a large enough alphabet. We denote $l = \phi(u)$ and we say that l is the *ϕ-transform* of the sequence u.

We denote by $\varphi_n(k)$ the number of non-periodic circular words of length n on k symbols. The numbers $\varphi_n(k)$ are called the *Witt numbers*. It is clear that the sequence $(\varphi_n(k))_{n\geq 1}$ is the ϕ-transform of the sequence $(k^n)_{n\geq 1}$.

The corresponding particular case of Identity (5)

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

is known as the *cyclotomic identity*.

The following arrays display a tabulation of the Witt numbers for small values of n and k.

n	$\varphi_n(2)$	$\varphi_n(3)$	$\varphi_n(4)$
1	2	3	4
2	1	3	6
3	2	8	20
4	3	18	60
5	6	48	204
6	9	116	670
7	18	312	2340
8	30	810	8160
9	56	2184	29120
10	99	5880	104754

The value $\varphi_3(4) = 20$ 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.

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

$$p_n = \sum_{d|n} d a_d^{n/d}.$$

The pair (a, p) is called a *Witt vector* (see [29] or [36]). The numbers p_n are the *ghost components*. In terms of generating series, one has

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

The following result is due to Schützenberger (see [16]).

THEOREM 14 *Let $u = (u_n)_{n\geq 1}$ be a sequence of nonnegative integers and let $l = (l_n)_{n\geq 1}$ be the ϕ-transform of u. The sequence $(u_n)_{n\geq 1}$ is the length distribution of a circular code on k symbols iff for all $(n \geq 1)$*

$$l_n \leq \varphi_n(k).$$

Several complements to Theorem 14 appear in [6]. In particular, the relation with Kraft's inequality is studied. The equality case in Kraft's inequality is characterized in terms of the sequence of inequalities above.

There is a connexion between Theorem 14 and Krieger's embedding theorem (Theorem 11), in the sense that Theorem 14 gives a simple proof of Theorem 11 in a particular case. Actually, let us consider the particular case of subshift of finite type, called a *renewal system.*

A renewal system S is the edge shift of a graph G made up of cycles sharing exactly one vertex. Such a graph is determined by the sequence $u = (u_i)_{1\leq i\leq n}$ where u_i is the number of loops with length i. Let T_k be the full shift on k symbols. Suppose that the pair formed by S and T_k satisfies the hypotheses of Krieger's theorem. The number $q_n(S)$ of points of least period n is nl_n where $l = (l_n)_{n\geq 1}$ is the ϕ-transform of the sequence u and $q_n(T_k) = n\varphi_n(k)$. Thus, the sequence u satisfies the hypotheses of Theorem 14. Consequently, there is circular code X such that $u_X = u$. The flower automaton of X defines an embedding of S_G into the full shift on k symbols. This gives an alternative proof of Krieger's theorem in this case.

It would be interesting to have a proof of Krieger's theorem along the same lines in the general case.

To close this section, we mention the following open problem: If the sequence u is regular and satisfies the inequalities

$$l_n \leq \varphi_n(k) \qquad (n \geq 1),$$

where $l = \phi(u)$, does there exist a rational circular code on k symbols such that $u = u_X$?

5.3 Zeta functions

Theorem 13 admits the following generalization due to Reutenauer [40].

THEOREM 15 *The zeta function of a sofic subshift is regular.*

We have seen already (Theorem 13) that the zeta function of a subshift of finite type is a rational fraction, and indeed the inverse of a polynomial. The stronger statement that it is regular follows from the following formula allowing to compute $\det(I - Mz)$ when M is the adjacency matrix of a $n \times n$ graph G. One has

$$\det(I - Mz) = (1 - v_1(z)) \cdots (1 - v_n(z)),$$

where $v_i(z)$ is the length distribution of the set of first returns to state i using only states $\{i, i+1, \ldots, n\}$ (see [12]).

The proof that the zeta function of a sofic subshift is rational is a result of Manning and Bowen [33], [20]. For an exposition, see [30] or [12]. A generalization appears in [17].

Acknowledgements

This paper is based on several texts written together with Frédérique Bassino and Marie-Pierre Béal (in particular [10] and [9]). I warmly thank them for agreeing the use of this material here.

The link between length distributions of circular codes and symmetric functions was disclosed to me by Jacques Désarménien and Jean-Yves Thibon. This connexion seems to open promising perspectives for the future.

References

[1] R. L. Adler, D. Coppersmith, and M. Hassner. Algorithms for sliding block codes. *IEEE Trans. Inform. Theory*, IT-29:5–22, 1983.

[2] R. Ahlswede, B. Balkenhol, and L. Khachatrian. Some properties of fix-free codes. Technical Report 039, University Bielefeld, 1997.

[3] Martin Aigner and Günter M. Ziegler. *Proofs from The Book.* Springer-Verlag, 1998.

[4] R. B. Ash. *Information Theory.* Dover Publications, Inc, New-York, 1990.

[5] Jonathan J. Ashley. A linear bound for sliding-block decoder window size. *IEEE Trans. Inform. Theory*, 1988.

[6] Frédérique Bassino. Generating functions of circular codes. *Adv. in Appl. Math*, 22(1):1–24, 1999.

[7] Frédérique Bassino, Marie-Pierre Béal, and Dominique Perrin. Enumerative sequences of leaves in rational trees. In *ICALP'97*, number 1256 in Lecture Notes in Computer Science, pages 76–86. Springer-Verlag, 1997.

[8] Frédérique Bassino, Marie-Pierre Béal, and Dominique Perrin. Enumerative sequences of leaves and nodes in rational trees. *Theoret. Comput. Sci.*, (221):41–60, 1999.

[9] Frédérique Bassino, Marie-Pierre Béal, and Dominique Perrin. Length distributions and regular sequences. In Joachim Rosenthal and Brian Marcus, editors, *Codes, Systems and Graphical Models*, IMA Volumes in Mathematics and its Applications. Springer-Verlag, 1999. To appear.

[10] Frédérique Bassino, Marie-Pierre Béal, and Dominique Perrin. A finite state version of the Kraft-McMillan theorem. *SIAM J. on Computing*, 2000. to appear.

[11] Marie-Pierre Béal. *Codage Symbolique*. Masson, 1993.

[12] Marie-Pierre Béal. Puissance extérieure d'un automate déterministe, application au calcul de la fonction fonction zêta d'un système sofique. *RAIRO Inform. Théor. Appl.*, 29:85–103, 1995.

[13] Marie-Pierre Béal, Filippo Mignosi, and Antonio Restivo. Minimal forbidden words and symbolic dynamics. In C. Puech and R. Reischuk, editors, *STACS'96*, volume 1046 of *Lecture Notes in Computer Science*, pages 555–566. Springer-Verlag, 1996.

[14] Marie-Pierre Béal, Filippo Mignosi, Antonio Restivo, and Marinella Sciortino. Forbidden words in symbolic dynamics. Technical Report 99-15, I.G.M., Université de Marne-la-Vallée, 1999. To appear in Adv. in Appl. Math.

[15] Marie-Pierre Béal and Dominique Perrin. Symbolic dynamics and finite automata. In G. Rosenberg and A. Salomaa, editors, *Handbook of Formal Languages*, volume 2, chapter 10. Springer-Verlag, 1997.

[16] J. Berstel and D. Perrin. *Theory of Codes*. Academic Press, 1985.

[17] Jean Berstel and Christophe Reutenauer. Zeta functions of formal languages. *Trans. Amer. Math. Soc.*, 321:533–546, 1990.

[18] Jean Berstel and Christophe Reutenauer. *Rational Series and their Languages*. Springer-Verlag, 1998.

[19] R. Bowen and O. E. Lanford. Zeta functions of restrictions of the shift transformation. In *Proc. Symp. Pure Math. AMS*, volume 14, pages 43–50, 1970.

[20] Rufus Bowen. On Axiom A diffeomorphisms. In *AMS-CBMS Reg. Conf.*, volume 35, Providence, 1978.

[21] Véronique Bruyère and Michel Latteux. Variable-length maximal codes. In F. Meyer and B. Monien, editors, *Proc. 23rd International Colloquium on Automata, Languages and Programming (ICALP'96)*, volume 1099, pages 24–47. Springer-Verlag, 1996.

[22] Samuel Eilenberg. *Automata,Languages and Machines*, volume A. Academic Press, 1974.

[23] Philippe Flajolet. Analytic models and ambiguity of context-free languages. *Theoret. Comput. Sci.*, 49:283–309, 1987.

[24] G. D. Forney, B. H. Marcus, N. T. Sindhushayana, and M. Trott. A multilingual dictionary: System theory, coding theory, symbolic dynamics and automata theory. In *Proceedings of Symposia in Applied Mathematics*, number 50, pages 109–138, 1995.

[25] F. R. Gantmatcher. *Matrix Theory, Volume II.* Chelsea Publishing Company, New-York, 1960.

[26] R. L. Graham, Donald Knuth, and O. Pataschnik. *Concrete Mathematics.* Addison Wesley, 1988.

[27] T. Katayama, M. Okamoto, and H. Enomoto. Characterization of the structure-generating of regular sets and DOL growth functions. *Information and Control*, 36:85–101, 1978.

[28] Bruce P. Kitchens. *Symbolic Dynamics.* Springer-Verlag, 1997.

[29] Serge Lang. *Algebra.* Addison Wesley, 1980.

[30] D. A. Lind and B. H. Marcus. *An Introduction to Symbolic Dynamics and Coding.* Cambridge, 1995.

[31] M. Lothaire. *Combinatorics on words.* Cambridge University Press, Cambridge, 1997.

[32] I. G. Macdonald. *Symmetric Functions and Hall Polynomials*. Oxford University Press, 1995.

[33] A. Manning. Axiom A difeomorphisms hava rational zeta functions. *Bull. London Math. Soc.*, 3:215–220, 1971.

[34] B. H. Marcus. Factors and extensions of full shifts. *Monats. Math*, 88:239–247, 1979.

[35] Brian H. Marcus, Ron M. Roth, and Paul H. Siegel. Constrained systems and coding for recording channels. In V. S. Pless and W. C. Huffman, editors, *Handbook of Coding Theory*, volume II, chapter 20, pages 1635–1764. North Holland, 1998.

[36] N. Metropolis and Gian-Carlo Rota. Witt vectors and the algebra of necklaces. *Advances in Math.*, 50:95–125, 1983.

[37] D. Perrin. Finite automata. In J. Van Leeuven, editor, *Handbook of Theoretical Computer Science*, volume B, chapter 1. Elsevier, 1990.

[38] D. Perrin. Finite automata. In Jan van Leeuwen, editor, *Handbook of Theoretical Computer Science*, volume B, chapter 1. Elsevier, 1990.

[39] Christophe Reutenauer. personal communication. 1997.

[40] Christophe Reutenauer. $\mathbb{N}$-rationality of zeta functions. *Adv. in Appl. Math.*, 29(1):1–17, 1997.

[41] Gian-Carlo Rota. *Finite Operator Calculus*. Academic Press, 1975.

[42] Arto Salomaa and Matti Soittola. *Automata Theoretic Properties of Formal Power Series*. Springer-Verlag, 1978.

[43] Thomas Scharf and Jean-Yves Thibon. On Witt vectors and symmetric functions. *Algebra Colloq.*, 3(3):231–238, 1996.

[44] Richard P. Stanley. *Enumerative combinatorics. Vol. 1.* Cambridge University Press, Cambridge, 1997.

[45] Chunxuan Ye and Raymond W. Yeung. Some basic properties of fix-free codes. 2000. submitted for publication.