12. Some open problems

As a conclusion, let us emphasize the underlying arithmetic structure and the interaction between the geometric and symbolic nature of the dynamical systems we have considered throughout this book. We are mainly interested in the two following problems: first, finding geometric interpretations of various symbolic dynamical systems including those generated by substitutions, and secondly, developing multidimensional continued fraction algorithms reflecting the dynamics of the systems.

12.1 The S-adic conjecture

Let us start with an algorithmic approach. For more details on the subject, the reader is referred to [167, 169].

A sequence is said to have an *at most linear complexity* if there exists a constant C such that for every positive integer n, $p(n) \leq Cn$, or in other words, if $p(n) = O(n)$. Many sequences that we have encountered in this book have at most linear complexity, including fixed points of primitive substitutions, automatic sequences, Sturmian sequences, Arnoux-Rauzy sequences, and so on. On the other hand, numerous combinatorial, ergodic or arithmetic properties can be deduced from this indication on the growth-order of the complexity function.

Let us start with a purely combinatorial result.

Theorem 12.1.1 (Cassaigne [106]). *A sequence has at most linear complexity if and only if the first difference of the complexity $p(n+1) - p(n)$ is bounded.*

Furthermore, an upper bound on the first difference can be explicitly given. Let us note that the equivalent result does not hold in the case of sequences with at most quadratic complexity: in [167], one can find an example of a sequence with a quadratic complexity function and unbounded second differences $p(n+2) + p(n) - 2p(n+1)$.

[1] This chapter has been written by P. Arnoux and V. Berthé

12.1.1 S-adic expansions

The following result [167] can be deduced from Theorem 12.1.1 and considerations on the graph of words. The complexity function of a symbolic dynamical system is defined as the function which counts the number of distinct factors of the *language* of this system, i.e., the union of the sets of factors of the sequences of the system.

Theorem 12.1.2 (Ferenczi [167]). *Let X be a minimal symbolic system on a finite alphabet $\mathcal{A}$ such that its complexity function $p_X(n)$ is at most linear, or, equivalently (following Theorem 12.1.1), such that $p_X(n+1) - p_X(n)$ is bounded; then there exists a finite set of substitutions $\mathcal{S}$ over an alphabet $\mathcal{D} = \{0, ..., d-1\}$, a substitution φ from $\mathcal{D}^\star$ to $\mathcal{A}^\star$, and an infinite sequence of substitutions $(\sigma_n)_{n\geq 1}$ with values in $\mathcal{S}$ such that $|\sigma_1\sigma_2...\sigma_n(r)| \to +\infty$ when $n \to +\infty$, for any letter $r \in \mathcal{D}$, and any word of the language of the system is a factor of $\varphi(\sigma_1\sigma_2...\sigma_n)(0)$ for some n.*

The above proposition can be read as follows: *minimal systems with at most linear complexity are generated by a finite number of substitutions.* Using a variation of the Vershik terminology, we propose to call such systems *S-adic systems*, and the pair $(\varphi, (\sigma_n))_{n\geq 1})$ is called an *S-adic expansion* of X. We similarly call a sequence *S-adic* if the symbolic dynamical system generated by the sequence is itself S-adic, and the sequence $(\varphi(\sigma_n))_{n\geq 1}$ is again called an *S-adic expansion* of the sequence x. Such expansions appear for instance in Chap. 6.

The fact that the lengths of the words tend to infinity, which generalizes the notion of everywhere growing substitutions (i.e., substitutions such that $\forall r$, $\exists n \in \mathbb{N}$, $|\sigma^n(r)| \geq 2$), is necessary to make Theorem 12.1.2 nonempty. Furthermore, it can be seen in the proof of [167] that this prevents us from getting a universal bound on the number d of letters of the alphabet $\mathcal{D}$. However, in an important particular case generalizing the Arnoux-Rauzy sequences [49], we do have a universal upper bound:

Proposition 12.1.3. *For minimal systems over a three-letter alphabet such that $p_X(n+1) - p_X(n) \leq 2$ for every n large enough, Theorem 12.1.2 is satisfied with $d \leq 3$ and $\mathrm{Card}\, S \leq 3^{27}$.*

Such universal upper bounds exist (but are very large) as soon as one has an upper bound on $p(n+1) - p(n)$.

A measure-theoretic consequence of Theorem 12.1.2 is that a minimal and uniquely ergodic system of at most linear complexity cannot be strongly mixing [167].

The S-adic expansion is known in an explicit way for the symbolic dynamical systems generated by Sturmian sequences, by the Arnoux-Rauzy sequences [49], for the systems generated by some binary codings of rotations [3, 146] (i.e., for codings with respect to a partition of the unit circle into intervals the lengths of which are larger than or equal to the angle of the rotation), and for systems generated by irrational interval exchanges (see in particular [2, 171, 270]). In the Sturmian case (see Sec. 6.4.4 in Chap. 6), one gets a more precise result: one can expand a given Sturmian sequence as an infinite composition of two substitutions; the rules for the iteration of these substitutions follow the Ostrowski expansion with respect to the angle of the initial point whose orbit is coded (for more details, see [41]). In some other cases (Arnoux-Rauzy sequences, binary codings, interval exchanges) one uses multidimensional continued fraction expansions. See also [1, 201] for a detailed sudy of the behaviour of sequences for which $p(n+1) - p(n) \in \{1, 2\}$ through the use of graphs of words. See also for more results on S-adicity [446].

12.1.2 The conjecture

The converse of Theorem 12.1.2 is clearly false. To produce a counter-example, it is sufficient to consider a non-primitive substitution [320] with a fixed point of complexity function satisfying $\Theta(n \log\log n)$ as $a \mapsto aba$, $b \mapsto bb$, or $\Theta(n \log n)$ as $a \mapsto aaba$, $b \mapsto bb$. Nevertheless, in the case where the initial sequence u is minimal, the fact that it is a fixed point of a substitution guarantees that the complexity is at most linear. This is not true any more if the sequence u is generated by two substitutions, even if u is minimal. Such an example was proposed by Cassaigne: one uses a positive substitution which appears infinitely often in the iteration, and a substitution having a fixed point of quadratic complexity which appears in long ranges in the iteration.

Question. Let u be a sequence generated by the iteration of a finite number of substitutions; which restrictions should one add to these substitutions, so that the sequence u has at most linear complexity?

We still have to find a *stronger* form of S-adicity which would be equivalent to at most linear complexity. This is the *S-adic conjecture* stating that minimal systems have at most linear complexity if and only if they are *strongly* S-adic.

12.1.3 Linear recurrence

Durand gives a sufficient condition in [155] for a sequence to have at most linear complexity. Namely, let u be a given recurrent sequence and let W be a factor of the sequence u. Let us recall that a *return word* over W is a word V such that VW is a factor of the sequence u, W is a prefix of VW and W has exactly two occurrences in VW. A sequence is *linearly recurrent* if there exists a constant $C > 0$ such that for every factor W, the length of every return word V of W satisfies $|V| \leq C|W|$. Such a sequence always has at most linear complexity [156]. Unfortunately, this condition is strictly stronger than having at most linear complexity. Durand shows that a sequence is linearly recurrent if and only if it is proper S-adic with bounded partial quotients, i.e., every substitution comes back with bounded gaps in the S-adic expansion [155]. In particular, a Sturmian sequence is linearly recurrent if and only if the partial quotients in the continued fraction expansion of its angle are bounded, which means that not every sequence with at most linear complexity is linearly recurrent.

12.1.4 Periodic case and substitutive sequences

Given an S-adic sequence, one can ask whether this sequence is substitutive, that is, whether it is a letter-to-letter projection of a fixed point of a substitution. Substitutive Sturmian sequences correspond to quadratic angles (for more details, see [126, 465] and Chap. 9). This can be deduced from Durand's characterization of substitutive sequences based on return words (the notion of derived sequence can be geometrically seen using the induction). Hence, this result can be considered as a generalization of Galois' theorem for continued fraction expansions (see Sec. 6.5). See also [86] for a connected result: if all the parameters of an interval exchange belong to the same quadratic extension, the sequence of induced interval exchanges (by performing always the same induction process) is ultimately periodic.

12.2 Multidimensional continued fraction expansions

12.2.1 Arithmetics and S-adicity

One of the main interests of the S-adic expansion lies in the fact that it provides an arithmetic description of the sequences we consider. For instance, the sequence (σ_n) (in Theorem 12.1.2) is governed by the continued fraction expansion of the angle in the Sturmian case. More generally, in numerous cases, the "partial quotients" (i.e., the gaps between successive occurrences of runs of the same substitution which appear in the iteration in Theorem 12.1.2) provide a generalized continued fraction expansion which describes the combinatorial properties of the sequence: one can find such examples of continued fraction expansions in [2, 114, 146, 170, 171, 172, 270, 349, 358, 467]. The techniques which are usually used in these problems are, on the one hand, the use of the graphs of words (see for instance [73]), and, on the other hand, the induction process (see Chaps. 5 and 6).

The usual continued fraction algorithm provides the best rational approximations of a real number α; it is fundamentally connected to the toral rotation of $\mathbb{T}$, $x \mapsto x + \alpha$. It describes in a natural way the combinatorial properties of Sturmian sequences, as the properties of their geometric representation as a coding of a rotation, as well as those of their associated symbolic dynamical system. What can be said now when considering the two-dimensional torus $\mathbb{T}^2$, or when trying to get simultaneous approximation results?

There exist several unidimensional or multidimensional symbolic objects which generalize Sturmian sequences in a natural way and provide multidimensional continued fraction algorithms. One can for instance code a rotation over $\mathbb{T}$ with respect to a partition into two intervals or more, the lengths of which do not depend on the angle α of the rotation. Such sequences have been studied for instance in [8, 75, 146, 363].

The following three particular generalizations have been the subject of recent studies: interval exchanges, the Arnoux-Rauzy sequences and codings of the $\mathbb{Z}^2$-action of two rotations on the unit circle.

12.2.2 Interval exchanges

Three-interval exchanges are fundamentally connected (via an induction process) to binary codings of rotations, that is, to codings of irrational rotations on $\mathbb{T}$ with respect to a two-interval partition. Ferenczi, Holton, and Zamboni develop in [170, 171, 172] a multidimensional algorithm which generates the orbits of the discontinuity points and opens new ways in the study of the ergodic and spectral properties of interval exchanges. This algorithm is produced in a combinatorial way by catenative rules of production of bispecial factors, or in a geometric way, by the observation of the evolution of the intervals corresponding to the bispecial factors. It can also be expressed in terms of a new induction process which does not correspond to the usual ones. See [2, 270] and also [269] for an S-adic expansion.

12.2.3 Arnoux-Rauzy sequences

Let us consider now the *Arnoux-Rauzy sequences.* It will be recalled that these are recurrent sequences defined over a three-letter alphabet with the following extra combinatorial property [49]: for every n, there is exactly one right special factor and one left special factor of length n, and these special factors can be extended in three different way. It will be recalled that a factor W of the sequence u is called right special (respectively left special) if W is a prefix (respectively suffix) of at least two words of length $|W|+1$ which are factors of the sequence u (see also Chap. 6). Let us note that they can be similarly defined over any alphabet of larger size, say d; one thus obtains sequences of complexity $(d-1)n+1$. Arnoux-Rauzy sequences are a natural generalization of Sturmian sequences, since they share with Sturmian sequences the fundamental combinatorial property of the unicity of the right (and left) special factors of given length. Contrary to the Sturmian case, these sequences are not characterized by their complexity function any more.

One knows perfectly well the S-adic expansion of the Arnoux-Rauzy sequences [49], and more precisely, the S-adic expansion of the associated dynamical system and the S-adic expansion of the sequence which has as prefixes the left special factors; following the terminology in the Sturmian case, this sequence is called the *left special sequence* or the *characteristic sequence*. The combinatorial properties of the Arnoux-Rauzy sequences are well-understood and are perfectly described by a two-dimensional continued fraction algorithm defined over a subset of zero measure of the simplex introduced in [358, 467] and in [116]. See also [114] for the connections with a generalization of the Fine and Wilf's theorem for three periods. By using this algorithm, one can express in an explicit way the frequencies of factors of given length [461], one can count the number of all the factors of the Arnoux-Rauzy sequences [297], or prove that the associated dynamical system has always simple spectrum [116].

These sequences can also be described as an exchange of six intervals of the unit circle [49]. Let us recall that the *Tribonacci sequence* (i.e., the fixed point beginning with a of the *Rauzy substitution*: $a \mapsto ab$, $b \mapsto ac$, $c \mapsto a$) is an Arnoux-Rauzy sequence. In this case, one obtains good approximation results [117]. The dynamical system generated by this sequence is isomorphic to a rotation of the torus $\mathbb{T}^2$ (for more details, see Chaps. 7 and 8). More generally, in the periodic case, i.e., in the case where the S-adic expansion of the special sequence is purely periodic, this sequence is a fixed point of a unimodular substitution of Pisot type (and conversely). All the results of Chaps. 7 and 8 apply. In particular, the dynamical systems generated by such sequences are obtained as toral rotations (the sufficient condition of Theorem 8.4.1 holds).

It was believed that all Arnoux-Rauzy sequences originated from toral rotations, and more precisely, that they were natural codings of rotations over $\mathbb{T}^2$. We say that a sequence u is a *natural coding* of a rotation if there exists a measurable dynamical system (X, T, μ), which is itself measure-theoretically isomorphic to a rotation, a generating partition $\mathcal{P} = \{P_1, P_2, P_3\}$ of X, and a point $x \in X$ such that

$$\forall n \in \mathbb{N},\ u_n = i \text{ if and only if } T^n(x) \in P_i;$$

furthermore, the map T is supposed to be a piecewise translation, the translation vector taking exactly one value on each P_i. Namely, in all the examples, we know that the rotation is constructed as an exchange of domains in $\mathbb{R}^2$, the pieces of the exchange being isomorphic to the cylinders (see Chaps. 7 and 8).

This conjecture was disproved in [110], where an example of a totally unbalanced Arnoux-Rauzy sequence is constructed (a sequence is said to be *totally unbalanced* if there exists a letter a such that for each positive integer n, there exist two factors of the sequence with equal length, with one having at least n more occurrences of the letter a than the other). Namely, a natural coding of a rotation cannot be totally unbalanced following [165]. Let us note that the partial quotients (in the sense of the S-adic expansion) are unbounded in the above example of totally unbalanced Arnoux-Rauzy sequence.

The following questions are thus natural.

Questions. How to characterize the Arnoux-Rauzy sequences which are natural codings of rotations? In a weaker form, how to characterize the Arnoux-Rauzy sequences which are measure-theoretically isomorphic to rotations? Is linear recurrence a sufficient condition for either question? Are there Arnoux-Rauzy sequences which are weakly mixing? What are the eigenvalues of the Arnoux-Rauzy sequences? Do they admit a rotation as a factor?

12.2.4 Codings of two rotations

A third way of generalizing the Sturmian case consists in introducing a second parameter. One thus gets two dual approaches: one can either code a rotation of angle (α, β) in the two-dimensional torus $\mathbb{T}^2$, or a $\mathbb{Z}^2$-action by two irrational rotations of angle α and β on $\mathbb{T}$. In the first case, one gets a unidimensional sequence, in the second, a two-dimensional sequence. We discussed the first case by considering Arnoux-Rauzy sequences and their connections with rotations over $\mathbb{T}^2$ in the previous section. Consider now the second approach.

Let $(\alpha, \beta) \in \mathbb{R}^2$, such that $1, \alpha, \beta$ are rationally independent, and let $\rho \in \mathbb{R}$. Let u be the two-dimensional sequence defined over $\mathbb{Z}^2$ with values in $\{0, 1\}$ by

$$\forall (m,n) \in \mathbb{Z}^2,\ (u(m,n) = 0 \Longleftrightarrow m\alpha + n\beta + \rho \in [0, \alpha[\text{ modulo } 1).$$

Such a two-dimensional sequence has $mn+n$ rectangular factors of size (m,n) and is *uniformly recurrent* (i.e., for every positive integer n, there exists an integer N such that every square factor of size (N,N) contains every square factor of size (n,n)). The function which counts the number of rectangular factors of given size is called the *rectangle complexity function* (for more details, see Sec. 12.3). Conversely, every two-dimensional uniformly recurrent sequence with complexity $mn + n$ admits such a geometric description [77].

Let us see why these sequences can be considered as a generalization of Sturmian sequences.

They have the smallest complexity function known among two-dimensional sequences which are uniformly recurrent and not periodic. They are obtained as a letter-to-letter projection of two-dimensional sequences defined over a three-letter alphabet which code a plane approximation (let us recall that Sturmian sequences code and describe discrete lines [66, 69]): one can approximate a plane with irrational normal by square faces oriented along the three coordinates planes; this approximation is called a *discrete plane* or a *stepped surface* as in Chap. 8; after projection on the plane $x+y+z=0$, along $(1,1,1)$, one obtains a tiling of the plane with three kinds of diamonds, namely the projections of the three possible faces (for more details, see [441, 78], see also [356, 177] and Chap. 8). One can code this projection over $\mathbb{Z}^2$ by associating with each diamond the name of the projected face. One thus gets a sequence with values in a three-letter alphabet. These sequences are shown to code a $\mathbb{Z}^2$-action over the unit circle $\mathbb{T}$ [78].

These latter sequences are generated by two-dimensional substitutions governed by the Jacobi-Perron algorithm. Namely, a geometric interpretation of the Jacobi–Perron algorithm is given in [40]: in the same way as classical continued fractions can be interpreted in terms of induction of rotations, this algorithm can be considered as an induction algorithm for a group of rotations operating on the unidimensional torus; it is not trivial that this group can be induced on a subinterval to obtain a new group of rotations, but by inducing on a suitable interval, the induction of the $\mathbb{Z}^2$-action we consider is again generated by a $\mathbb{Z}^2$-action through a pair of rotations. However, unlike the classical $\mathbb{Z}$-action, the generators of the $\mathbb{Z}^2$-action are not canonically defined (since we can find an infinite number of bases for the lattice $\mathbb{Z}^2$). This can be related to the fact that there seems to be no way of defining a "best" two-dimensional continued fraction algorithm. Using this induction process, a sequence of two-dimensional substitutions is defined in [40] associated with the Jacobi–Perron algorithm (see also Chap. 8) which generates the two-dimensional Sturmian sequences mentioned above over a three-letter alphabet. We shall allude again to this notion of two-dimensional substitution in Sec. 12.3.

12.3 Combinatorics on two-dimensional words

When one works with multidimensional sequences, fundamental problems in the definition of the objects appear; for instance, how can one define a multidimensional complexity function? Consider two-dimensional sequences, i.e., sequences defined over $\mathbb{Z}^2$ with values in a finite alphabet. A possible notion of complexity consists in counting the rectangular factors of given size; we thus define the *rectangle complexity function*: $(m,n) \mapsto P(m,n)$. This notion is not completely satisfactory since it depends on the choice of a basis of the lattice $\mathbb{Z}^2$. For a more general definition of complexity, see [371]. The following questions are thus natural:

Questions. Can one characterize a two-dimensional sequence with respect to its rectangle complexity function? Which functions do exist as rectangle complexity functions?

A two-dimensional sequence is *periodic* if it is invariant under translation, i.e., if it admits a nonzero vector of periodicity. Note that the fact that the lattice of periodicity vectors has rank 2 is characterized by a bounded rectangle complexity function. There is no characterization of periodic sequences by means of the complexity function: one can construct two-dimensional sequences with a nonzero periodicity vector of very large complexity function;

namely, consider the sequence $(u_{m+n})_{(m,n)\in\mathbb{Z}^2}$, where the unidimensional sequence $(u_n)_{n\in\mathbb{Z}}$ has maximal complexity. Conversely Nivat has conjectured the following:

Conjecture 12.3.1. If there exists (m_0, n_0) such that $P(m_0, n_0) \leq m_0 n_0$, then the two-dimensional sequence u is periodic.

Let us note that the conjecture was proved for factors of size $(2, n)$ or $(n, 2)$ in [373]. A more general conjecture is given in [372, 373, 371]. Let us remark that in higher-dimensional cases, counter-examples to the conjecture can be produced [373].

It is interesting, with respect to this conjecture, to consider the "limit" case of sequences of rectangle complexity function $mn+1$. Such sequences are fully described in [109] and are all proved to be non-uniformly recurrent. The two-dimensional Sturmian sequences mentioned in Sec. 12.2.4 of complexity $mn+n$ are conjectured to be the uniformly recurrent sequences with smallest complexity function (it remains to give a more precise meaning to the term "smallest" for parameters belonging to $\mathbb{N}^2$).

Let us end this section with the problem of the definition of two-dimensional substitutions. Unlike the classical one-dimensional case, the notion of a two-dimensional substitution is not trivial. Consider a map which associates with a letter a pointed two-dimensional pattern. We first need more information to know where to place the image of a letter. Two natural problems then arise. First, it is not immediate to prove the consistency if one wants to apply the substitution to a finite pattern or to a double sequence. Secondly, how can one iterate such a process to generate a double sequence? Two notions of substitutions are introduced in [40]. *Pointed substitutions* are first defined: given the value of the initial sequence at the point x, one can deduce the value of the image sequence on a pointed pattern situated at a point y that can be computed from x and its value. This is however inconvenient for explicit computation, since one needs at each step global information. In particular it is difficult to iterate it in order to generate a double sequence. It is much more convenient to be able to use a local information, i.e., *local rules* (this is exactly what is done when one computes one-dimensional substitutions: one does not compute the exact position of a given pattern, but one only uses the fact that patterns follow each other). Roughly speaking, a local rule says how to place the image of a pointed letter with respect to the images of the letters belonging to a finite neighborhood. If we know the image of the initial point, we can compute the values of adjacent points by using a finite number of patterns, and in this way, compute the image of the complete sequence. See also for a study of rectangle substitutions [195].

Questions. Can one build other examples of substitutions endowed with local rules as the ones produced in [40]? Given a substitution with local rules, does the limit of the iterates cover $\mathbb{Z}^2$? Can one introduce a different notion of substitution which can be used for the generation of two-dimensional sequences?

12.4 Substitutions, rotations and toral automorphisms

The general question of the isomorphism between a substitutive dynamical system and a rotation brings many open questions. For more details, see Chap. 7.

The dynamical system generated by the Tribonacci substitution is measure-theoretically isomorphic to a translation on the torus $\mathbb{T}^2$. The isomorphism is a continuous onto map from the orbit closure of the fixed point of the substitution to the torus. The images of the three basic cylinders corresponding to the letters are connected, and even simply connected domains of the torus. The three basic sets are bases for the three cylinders of a Markov partition for the automorphism of the torus $\mathbb{T}^3$ associated with the substitution. Do these properties extend to some other symbolic dynamical systems?

Let us recall (Theorem 7.5.18) that the symbolic dynamical system associated with a unimodular substitution of Pisot type over a two-letter alphabet $\{0, 1\}$ is measure-theoretically isomorphic to a rotation on $\mathbb{T}$ (for more details, see Chap. 7, and in particular Sec. 7.5.3). This is obtained by proving [54] that this class of substitution satisfies the following coincidence condition: there exist two integers k, n such that $\sigma^n(0)$ and $\sigma^n(1)$ have the same k-th letter, and the prefixes of length $k-1$ of $\sigma^n(0)$ and $\sigma^n(1)$ have the same number of occurrences of the letter 0.

It is conjectured that every unimodular substitution of Pisot type over three letters satisfies a coincidence condition on three letters (see Chap. 7). It is not known whether there exists a unimodular substitution of Pisot type that does not satisfy the coincidence condition.

Conjecture 12.4.1. If the Perron–Frobenius eigenvalue of the incidence matrix of a unimodular substitution is a Pisot number, then the dynamics of the substitution is measure-theoretically isomorphic to a rotation on the torus.

A connected problem is the following, addressed by Liardet in [188]: *is the odometer associated with the numeration scale related to a real number $\beta > 1$ isomorphic in measure to a translation on a compact group if β is a canonical Pisot number (i.e., β has a finite β-expansion and the associated companion polynomial is irreducible over $\mathbb{Q}$)?*

Other natural questions arise concerning the topological properties of the fractal domain X_σ associated with a substitution, such as the question of connectedness and simple connectedness : Canterini gives in [101] a sufficient condition for the connectivity of the exchange of domains associated with unimodular substitutions of Pisot type.

Question. Can one give a characterization of those substitutions for which the associated domains are connected or simply connected?

12.5 Arithmetics in $SL(d, \mathbb{N})$ and $SL(d, \mathbb{Z})$

We have considered here products of substitutions via the S-adic conjecture. This can also be expressed in terms of products of matrices. It is thus natural from an algebraic point of view to consider the properties of the monoid $SL(d, \mathbb{N})$ and of the group $SL(d, \mathbb{Z})$, for $d \geq 2$.

Definitions. For $d \geq 2$, let $SL(d, \mathbb{N})$ (respectively $SL(d, \mathbb{Z})$) denote the set of matrices of determinant 1 with nonnegative (respectively integer) coefficients. This set endowed with the multiplication is a monoid (respectively a group), whose identity element is the identity matrix $\mathbf{I}_d$.

A matrix $\mathbf{M}$ in $SL(d, \mathbb{N})$ is said to be a *unit* if it admits an inverse in $SL(d, \mathbb{N})$, i.e., if there exists a matrix $\mathbf{N} \in SL(d, \mathbb{N})$, such that $\mathbf{MN} = \mathbf{I}_d$. It is easy to check that the unit matrices are exactly the even permutation matrices; there is no nontrivial unit for $d = 2$, and there are 2 units different from the identity if $d = 3$, that is, the matrices:

$$\mathbf{P} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 1 \end{pmatrix}, \ \mathbf{P}^2 = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}.$$

A matrix $\mathbf{M}$ in $SL(d, \mathbb{N})$ is said to be *undecomposable* if it is not a unit matrix, and if, for any pair of matrices $\mathbf{A}, \mathbf{B}$ in $SL(d, \mathbb{N})$ such that $\mathbf{M} = \mathbf{AB}$, $\mathbf{A}$ or $\mathbf{B}$ is a unit.

Structure of $SL(2, \mathbb{N})$. The structure of the monoid $SL(2, \mathbb{N})$ is very simple; this is a free monoid with two generators (for more details, see Chap. 6)

$$\mathbf{A} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}, \ \mathbf{B} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

These two matrices $\mathbf{A}$ and $\mathbf{B}$ are the only undecomposable matrices in $SL(2, \mathbb{N})$, and any matrix admits a unique decomposition in terms of these two undecomposable elements. This decomposition is a matricial translation of the Euclidean algorithm and it corresponds to the continued fraction expansion.

Structure of $SL(3, \mathbb{N})$. The situation in $SL(3, \mathbb{N})$ is completely different. Note that this monoid is not free. Furthermore, we have the following theorem. For a proof of this result (by J. Rivat), see Appendix A.

Theorem 12.5.1 (Rivat). *There exist infinitely many undecomposable matrices in $SL(3, \mathbb{N})$.*

Questions. Can one give a simple characterization of the undecomposable matrices in $SL(3,\mathbb{N})$? Can one find an explicit family of undecomposable generators of $SL(3,\mathbb{N})$?

Some motivations for this question are the following: first, a decomposition of a matrix as a product of undecomposable matrices can be seen as a multidimensional continued fraction algorithm; secondly, it can also provide a better understanding of the automorphisms of the torus.

Invertible substitutions. A substitution over the alphabet $\mathcal{A}$ is said to be *invertible* if it extends to a morphism of the free group generated by $\mathcal{A}$.

Invertible substitutions over a two-letter alphabet are completely characterized (see Sec. 9.2 and Chap. 6). In particular they are Sturmian, i.e., they preserve Sturmian words. In particular, the monoid of invertible substitutions is finitely generated. One is far from understanding the structure of the monoid of invertible substitutions over a larger size alphabet: indeed the monoid of invertible substitutions over a d-letter alphabet is no more finitely generated, for $d > 2$ (see [453]). Let us recall that $SL(3,\mathbb{N})$ is not finitely generated (Theorem 12.5.1). However, these two results, which could be thought nearly equivalent, are in fact completely unrelated. First, the substitutions of [453] have matrices which are decomposable as products of elementary matrices, and second, the abelianization map is not onto: there are positive matrices of determinant 1 that are not matrices of invertible substitutions for $d = 3$.

Question. Can one characterize invertible substitutions over a larger size alphabet?

Some decision problems. Some related decision problems can be considered, and most of them turn out to be undecidable in dimension at least 3, such as the mortality problem, that is, the presence of the zero matrix in a finitely generated subsemigorup.

In particular, is it possible to find an algorithm which decides whether the semigroup generated by two or more square matrices of dimension two over the nonnegative integers is free? The answer is negative in the three-dimensional case: namely, this problem is shown to be undecidable by reducing the Post correspondence problem to it [248]. It is shown in [111] that the above problem is still undecidable for square matrices of dimension 3 that are upper-triangular. Furthermore, in the case of two upper-triangular 2×2 matrices, sufficient conditions for freeness of the semigroup generated by these two matrices are given. Related undecidable problems for 2-generator matrix semigroups are the following (see [112]): given two square matrices, decide whether the semigroup that they generate contains the zero matrix, and whether it contains a matrix having a zero in the right upper corner.

A. Undecomposable matrices in dimension 3 (by J. Rivat)

Let us recall that $SL(d,\mathbb{N})$ denotes the monoid of square matrices of dimension d with nonnegative integer coefficients and determinant 1. A matrix $\mathbf{M}$ is a *unit* if there exists a matrix $\mathbf{N}$ such that $\mathbf{MN} = \mathbf{I}_d$. A matrix $\mathbf{M}$ of $SL(d,\mathbb{N})$ is *undecomposable* if it is not a unit, and if, for all pair of matrices $\mathbf{A}, \mathbf{B}$ in $SL(d,\mathbb{N})$ such that $\mathbf{M} = \mathbf{AB}$, $\mathbf{A}$ or $\mathbf{B}$ is a unit. The aim of this appendix is to answer to the natural question: does there exist an infinite number of undecomposable matrices in $SL(3,\mathbb{N})$? The following theorem gives a positive answer:

Theorem A.0.2. *For any integer $n \geq 3$, the following matrix is undecomposable:*

$$\mathbf{M}_n = \begin{pmatrix} 1 & 0 & n \\ 1 & n-1 & 0 \\ 1 & 1 & n-1 \end{pmatrix}.$$

Remark. The matrices $\mathbf{M}_1$ and $\mathbf{M}_2$ are not undecomposable, since

$$\mathbf{M}_1 = \begin{pmatrix} 1&0&1 \\ 1&0&0 \\ 1&1&0 \end{pmatrix} = \begin{pmatrix} 1&0&0 \\ 0&1&0 \\ 0&1&1 \end{pmatrix} \begin{pmatrix} 1&0&1 \\ 1&0&0 \\ 0&1&0 \end{pmatrix}, \quad \mathbf{M}_2 = \begin{pmatrix} 1&0&0 \\ 0&1&0 \\ 0&1&1 \end{pmatrix} \begin{pmatrix} 1&0&2 \\ 1&1&0 \\ 0&0&1 \end{pmatrix}.$$

Proof. We fix $n \geq 3$. We consider $\mathbf{A}, \mathbf{B} \in SL(3,\mathbb{N})$ such that $\mathbf{AB} = \mathbf{M}_n$, and we denote:

$$\mathbf{A} = \begin{pmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix}.$$

If $a_1 a_2 a_3 \neq 0$, using the relation $a_1 x_2 + a_2 y_2 + a_3 z_2 = 0$, we deduce that $x_2 = y_2 = z_2 = 0$, which is impossible for a matrix of determinant 1. Hence $a_1 a_2 a_3 = 0$, and, up to a cyclic permutation of the columns (multiplication by $\mathbf{P}$ or $\mathbf{P}^2$), one has either $a_1 \neq 0$, $a_2 = a_3 = 0$ or $a_1 \neq 0$, $a_2 = 0$, $a_3 \neq 0$ (the case $a_1 = a_2 = a_3 = 0$ is impossible since $\det(\mathbf{A}) = 1$).

[1] This appendix has been written by J. Rivat; Institut Elie Cartan de Nancy; Faculté des Sciences; B.P. 239; 54506 Vandoeuvre-lès-Nancy Cedex; FRANCE; rivat@iecn.u-nancy.fr

From the two null coefficients of $\mathbf{M}_n$, one obtains:

$$a_1x_2 = a_2y_2 = a_3z_2 = b_1x_3 = b_2y_3 = b_3z_3 = 0.$$

First case: Suppose $a_1 \neq 0$, $a_2 = a_3 = 0$, hence $x_2 = 0$.

Since $\det(\mathbf{A}) = a_1(b_2c_3 - b_3c_2) = 1$, one has $a_1 = 1$. Hence the first line of $\mathbf{A}$ is $(1\ 0\ 0)$, which implies that the first line of $\mathbf{B}$ is the same as the first line of $\mathbf{M}_n$.

From the equation above $b_1x_3 = 0$, and $x_3 = n$, one obtains $b_1 = 0$; from the equation $c_1x_3 + c_2y_3 + c_3z_3 = n - 1$ and $x_3 = n$, one obtains $c_1 = 0$, hence:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & b_2 & b_3 \\ 0 & c_2 & c_3 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & n \\ y_1 & y_2 & y_3 \\ z_1 & z_2 & z_3 \end{pmatrix}.$$

Since $\det(\mathbf{A}) = b_2c_3 - b_3c_2 = 1$, b_2 must be strictly positive; hence, from the equation above $b_2y_3 = 0$, one obtains $y_3 = 0$. From the equation $c_3z_3 = n - 1$, one obtains that z_3 is strictly positive, hence, since $b_3z_3 = 0$, one has $b_3 = 0$; using again $\det \mathbf{A} = 1$, we get $b_2 = c_3 = 1$. Since the second line of $\mathbf{A}$ is $(0\ 1\ 0)$, the second line of $\mathbf{B}$ must be that of $\mathbf{M}_n$, hence:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & c_2 & 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & n \\ 1 & n-1 & 0 \\ z_1 & z_2 & z_3 \end{pmatrix}.$$

This implies that $(n-1)c_2 + z_2 = 1$, hence $c_2 = 0$, and $\mathbf{A}$ is the identity matrix.

Second case: Suppose $a_1 \neq 0$, $a_2 = 0$, $a_3 \neq 0$.

We have then $x_2 = z_2 = 0$. Since $\det(\mathbf{B}) = y_2(x_1z_3 - x_3z_1) = 1$, one obtains $y_2 = 1$; this implies as above that the second column of $\mathbf{A}$ is that of $\mathbf{M}_n$. Hence:

$$\mathbf{A} = \begin{pmatrix} a_1 & 0 & a_3 \\ b_1 & n-1 & b_3 \\ c_1 & 1 & c_3 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} x_1 & 0 & x_3 \\ y_1 & 1 & y_3 \\ z_1 & 0 & z_3 \end{pmatrix}.$$

From the determinant of $\mathbf{B}$, one obtains that x_1 and z_3 are strictly positive, and from the equation $a_1x_1 + a_3z_1 = 1$, one obtains that $a_1 = x_1 = 1$, $z_1 = 0$, hence, again from the determinant of $\mathbf{B}$, $z_3 = 1$. Hence:

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & a_3 \\ b_1 & n-1 & b_3 \\ c_1 & 1 & c_3 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 1 & 0 & x_3 \\ y_1 & 1 & y_3 \\ 0 & 0 & 1 \end{pmatrix}.$$

From the equation $b_1 + (n-1)y_1 = 1$, one obtains $y_1 = 0$, $b_1 = 1$, and from the equation $b_1x_3 + (n-1)y_3 + b_3 = 0$, one obtains $x_3 = y_3 = 0$, hence $\mathbf{B}$ is the identity, which ends the proof. ∎