Random generation of Finite Sturmian Words [‡]

Jean Berstel[*] Michel Pocchiola[†]

March 18, 1993

Abstract

We present a bijection between the set of factors of given length of Sturmian words and some set of triples of nonnegative integers. This bijection and its inverse are both computable in linear time. Its applications are : a bijective proof of Mignosi's formula for counting Sturmian words, a linear probabilistic algorithm for generating finite Sturmian word at random, and, using similar techniques, a linear on-line algorithm for computing the longest Sturmian prefix of a given word.

The construction of the bijection relies on concepts from combinatorial geometry.

1 Introduction

A doubly-infinite sequence $w = \cdots w_{-2}w_{-1}w_0w_1w_2\cdots$ over the alphabet $\{0,1\}$ is called *Sturmian* if there exist real numbers α, β with $0 \leq \alpha, \beta < 1$ such that

$$w_n = \lfloor \alpha(n+1) + \beta \rfloor - \lfloor \alpha n + \beta \rfloor \qquad (1)$$

for all integers n. Sturmian words have a long history and appear under a great variety of denominations. A clear exposition of early work

[*] Laboratoire d'Informatique Théorique et de Programmation, Institut Blaise Pascal, 4 place Jussieu, 75252 Paris Cédex 05, France. (berstel@litp.ibp.fr)

[†] Département de Mathématiques et d'Informatique, Ecole normale supérieure, ura cnrs 1327, 45 rue d'Ulm, 75230 Paris Cédex 05, France. (pocchiola@dmi.ens.fr)

[‡] Partially supported by PRC "Mathématiques et Informatique"

by J. Bernoulli, Christoffel, and A. A. Markov is given in the book by Venkov [23]. The term "Sturmian" has been used by Hedlund and Morse in their development of symbolic dynamics [9, 10, 11]. The same objects are known as "characteristic" sequences, "cutting" sequences, "Beatty" sequences, "nonhomogeneous spectra", "billiard" trajectories and others.

There is a large literature about properties of these sequences (see for example Series [21], Fraenkel *et al.* [7], Stolarsky [22]). From a combinatorial point of view, they have been considered by Rauzy [16, 17, 18], Brown [4], Ito, Yasutomi [12] in particular in relation with iterated morphisms, and by Séébold [20], Mignosi [14]. Sturmian words appear in ergodic theory [15], in computer graphics [3], and in crystallography [13]. Dulucq and Gouyou-Beauchamps [5] considered the set of all finite words that are factors of some Sturmian word. They proved that the complement, say C, of this set is a context-free language, and they conjectured that C is inherently ambiguous. To show this, they in fact conjectured a formula for the number of factors of length m of Sturmian words. Since the generating series of these numbers is transcendental, the Chomsky-Schützenberger theorem would prove inherent ambiguity (see Flajolet [6] for a systematic exposition). This formula was proved later by Mignosi [14] and will be obtained in the present paper as an easy corollary of our main result.

The aim of this paper is to present a bijection between the set $\mathcal{S}$ of factors of length m of Sturmian words and some set $\mathcal{T}$ of triples (a, p, q) of nonnegative integers bounded by m. This bijection and its inverse are both computable in linear time. The derivation of the bijection involves the analysis of the partition of lines induced by the lattice points and some elementary concepts of combinatorial geometry.

The bijection has several interesting applications. First we obtain a straightforward proof of Mignosi's enumeration formula for factors of Sturmian words. A second consequence is an algorithm for random generation of Sturmian words of given length which may have some application in computer graphics. More precisely it is a probabilistic algorithm in the sense that it successfully terminates only in the average. A third application is of interest in pattern recognition: we give a a linear time algorithm to compute the longest prefix of a word that is Sturmian. It is on-line and tests in constant time whether the given prefix can be extented by the next input symbol to a Sturmian word. Our algorithm is simpler and more general than [2].

2 Results

We call a factor of length m of a Sturmian word, a *Sturmian m-factor.* It is easily verified that the sequence

$$(\lfloor \alpha(n+1)+\beta \rfloor - \lfloor \alpha n + \beta \rfloor)_{0 \leq n < m} \tag{2}$$

ranges over all Sturmian m-factors when α and β range over $[0,1[$. We denote by $\mathcal{S}$ the set of Sturmian m-factors.

The main result of the paper is a description of a natural bijection between the set $\mathcal{S}$ and a subset $\mathcal{T}$ of $\{0,1,\ldots,m\}^3$. The set $\mathcal{T}$ is defined by

$$\mathcal{T} = \{(m,1,1)\} \cup \{\, (a,p,q) \mid 1 \leq q \leq p \leq a+p \leq m,\ \gcd(p,q)=1\,\} \tag{3}$$

Let B be the mapping which associates to a triple $(a,p,q) \in \mathcal{T}$ the sequence

$$B(a,p,q) = u_1 - u_0, u_2 - u_1, \ldots, u_m - u_{m-1} \tag{4}$$

where

$$u_n = \begin{cases} \lfloor \alpha n + \beta \rfloor, & \text{for } n = 0, \ldots, a; \\ \lceil \alpha n + \beta \rceil - 1, & \text{for } n = a+1, \ldots, m, \end{cases} \tag{5}$$

with $\alpha = q/p$ and $\beta = \lceil \alpha a \rceil - \alpha a$.

The sequence $B(a,p,q)$ is a $\{0,1\}$-sequence because α is less than 1.

Theorem 1 *The mapping B is a bijection from $\mathcal{T}$ onto the set $\mathcal{S}$ of Sturmian m-factors.*

Observe that B is computable in linear time. The fact that the inverse mapping B^{-1} is also computable in linear time will be a consequence of Proposition 3 below.

The proof is based on a geometric encoding of Sturmian words. Sturmian m-factors are defined by straight lines but many different straight lines may defined the same Sturmian m-factor. Each equivalence class is a convex polygon in the space of lines. As we shall see each polygon can be characterized by choosing a suitable edge. In some sense that will be made precise this is equivalent to representing a class by a pair of lines. This representation is specially convenient for computation because it allows a simply derivation the associated triple in $\mathcal{T}$. The construction will be detailed in the next section.

We now come to the applications.

Mignosi [14] proved the following result.

Proposition 1 *The number of factors of length m of Sturmian words is given by the sum*

$$1+\sum_{1}^{m}(m-i+1)\phi(i), \tag{6}$$

where ϕ is the Euler function, i.e., $\phi(n)$ is the number of natural integers less than n and coprime to n.

The authors of the present paper gave another proof in [1]. In view of the theorem above this proposition becomes straightforward. Indeed, the formula (6) counts precisely the number of elements in set the $\mathcal{T}$.

The next application concerns random generation of Sturmian m-factors. For this it suffices in view of Theorem 1 to generate triples (a,p,q) in the set $\mathcal{T}$ with uniform distribution. This is done by the following algorithm.

Algorithm 1

```
(1)  Repeat
         Generate uniformly a triple $(a,p,q) \in \{0,\ldots,m\}^3$
     until $(a,p,q)$ is in $\mathcal{T}$;
(2)  Compute $B(a,p,q)$.
```

Step 1 is the usual rejection algorithm : we generate triples (a,p,q) in $\{0,1,\ldots,m\}^3$ uniformly and reject those which are not in $\mathcal{T}$. Since the size of $\mathcal{T}$ is asymptotically equal to m^3/π^2 the expected number of samplings is asymptotically π^2, and therefore is independent of m. Thus we have.

Proposition 2 *Algorithm 1 generates factors of length m of Sturmian words in linear expected time.* □

Finally we consider the problem of testing if a given word is Sturmian. More precisely we shall prove the following result.

Proposition 3 *Given a word w in $\{0,1\}^*$, its maximal Sturmian prefix z can be computed on-line in time proportional to the length of z.*

Briefly, the algorithm runs as follow. For each Sturmian prefix we maintain the polygon (in the space of lines) of all lines defining this Sturmian factor. We incrementally transform this polygon by adding the geometric constraints defined by a new letter. The main point is that this can be done in constant time because the current polygon always has at most 4 edges. Our algorithm is on-line, this means that it is not necessary (contrary to the algorithm in [2]) to read all the word w before starting the decision process. Details will be given in the last section.

3 The bijection

Recall that a *Sturmian m-factor* is a factor of length m of a Sturmian word, and that the sequence

$$(\lfloor\alpha(n+1)+\beta\rfloor-\lfloor\alpha n+\beta\rfloor)_{0\leq n<m} \tag{7}$$

ranges over all Sturmian m-factors when α and β range over $[0,1[$. We denote by $\mathcal{S}$ the set of Sturmian m-factors.

In what follows we consider the Sturmian m-factor (7) as a function of the straight line represented in Cartesian coordinates (x,y) by the linear equation

$$y=\alpha x+\beta. \tag{8}$$

The real α is the *slope* of the straight line and the real β may be called the *intercept*. For a straight line ℓ with equation (8) we denote by ℓ^+ the closed half-plane $y\leq\alpha x+\beta$.

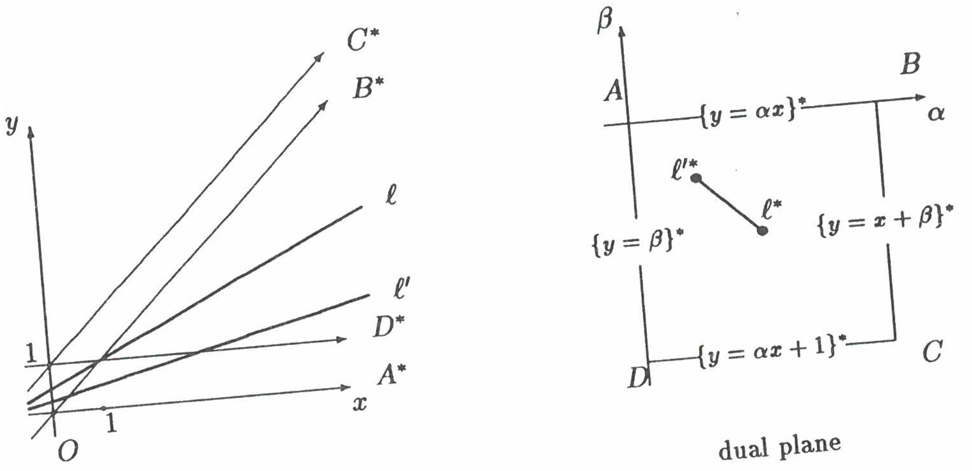
Let $\mathcal{H}$ be the set of straight lines with slope and intercept in the closed interval $[0,1]$ and let $\mathcal{L}$ be the subset of those straight lines whose slope and intercept lie in the semi-open interval $[0,1[$. For ℓ and ℓ' in $\mathcal{H}$, the *segment* $[\ell,\ell']$ is the set of straight lines with equation

$$y=(u\alpha+(1-u)\alpha')x+(u\beta+(1-u)\beta')$$

where u ranges over $[0,1]$ and where α,α' and β,β' are the slopes and intercepts of ℓ and ℓ', respectively. Related notions, such as *open segment, polygon, convexity*, etc. are defined as usual and freely used in the sequel.

To represent sets of lines by sets of points we use a *duality* transform $x\mapsto x^*$. Duality maps the line $\ell\in\mathcal{H}$ with equation $y=\alpha x+\beta$ to the point ℓ^* with coordinates $(\alpha,-\beta)$ and the point p with coordinates (α,β) to the line p^* with equation $y=\alpha x-\beta$. It can be easily verified that the duality transform is an involution, and that it preserves the incidence relation i.e.,

$$p\in\ell\Leftrightarrow\ell^*\in p^*,\qquad p\in\ell^+\Leftrightarrow\ell^*\in(p^*)^+. \tag{9}$$



primal plane

dual plane

Figure 1: The duality transform.

The (x, y)-plane and (α, β)-plane are called the *primal* and *dual* plane, respectively. The segment $[\ell, \ell']$ is then represented in the dual plane by the segment $[\ell^*, \ell'^*]$ (see Figure 1).

We define the *upper closure* of a subset X of the plane to be the set of points (x, y) of the topological closure of X such that $(x, y - \epsilon) \in X$ for every sufficiently small positive ϵ. We define then the *upper closure* of a set of lines by duality.

For a straight line ℓ with equation (8), we denote by $S(\ell)$ the Sturmian m-factor given by (7). This defines a mapping S from $\mathcal{L}$ onto $\mathcal{S}$. The mapping S induces a natural partition of $\mathcal{L}$ whose parts are the sets $S^{-1}(s)$ when s ranges over $\mathcal{S}$. The following proposition relates this partition to the lattice set P defined by

$$P = \{\,(x, y) \in \mathbf{N} \mid 0 \leq x \leq m\,\}. \tag{10}$$

Proposition 4 *Two lines ℓ and ℓ' in $\mathcal{L}$ define the same Sturmian m-factor, i.e. $S(\ell) = S(\ell')$, if and only if*

$$\ell^+ \cap P = \ell'^+ \cap P.$$

Proof. The set $\ell^+ \cap P$ is clearly defined by the lattice points $(n, \lfloor \alpha n + \beta \rfloor)$ for $n = 0, \ldots, m$, and consequently by the Sturmian m-factor defined by ℓ since $\beta \in [0, 1[$. □

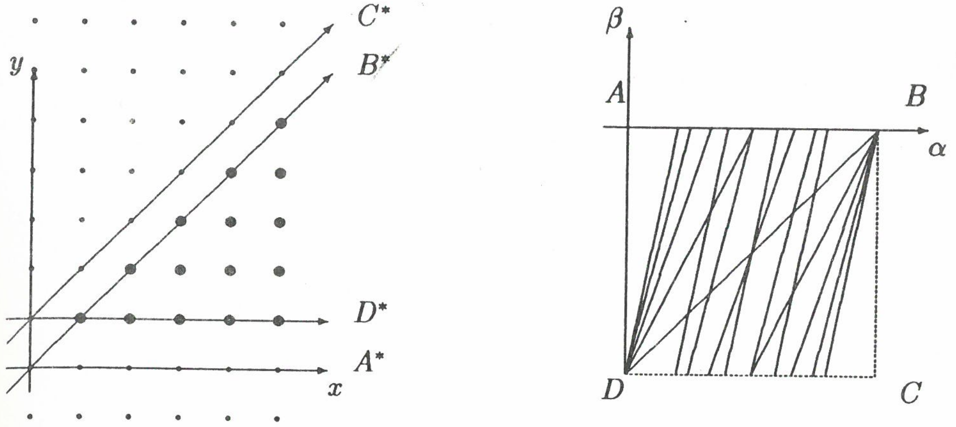


Figure 2: The partition $\mathcal{A}$ for $m = 5$

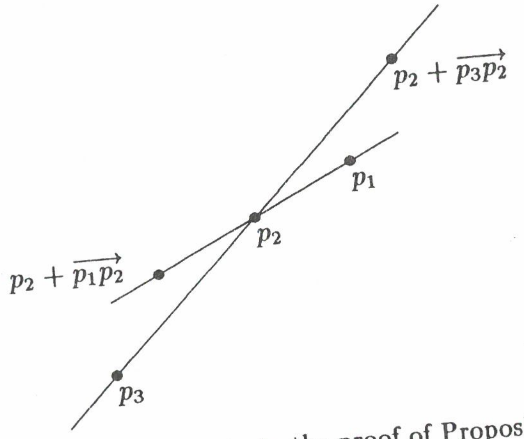
As shown by the previous proposition the Sturmian m-factors are closely related to the way lines separate the lattice set P or equivalently the way lattice points of P separate the set of lines $\mathcal{L}$. We make a brief account of this classical subject of combinatorial geometry.

Each lattice point p in P induces a partition of $\mathcal{L}$ in at most three parts, namely

$$\begin{array}{ll} \{\,\ell \in \mathcal{L} \mid p \in \ell\,\} & \text{the lines through } p \\ \{\,\ell \in \mathcal{L} \mid p \notin \ell^+\,\} & \text{the lines below } p \\ \{\,\ell \in \mathcal{L} \mid p \in \ell^+,\, p \notin \ell\,\} & \text{the lines above } p. \end{array}$$

We denote by $\mathcal{A}$ the intersection of these partitions as p ranges over P. For any line $\ell \in \mathcal{H}$, let $c(\ell)$ denote the number of lattice points in P lying on the line ℓ, i.e., $c(\ell) = \mathrm{Card}(\ell \cap P)$. It is clear that the function c is constant on each part of $\mathcal{A}$. One can easily verify that the cells (or parts) of $\mathcal{A}$ are relatively open convex polygons of $\mathcal{H}$ of dimension 2, 1 or 0 according to the value 0, 1 or ≥ 2 of the function c on the cell. The cells of dimension $0, 1$ and 2 are respectively called the vertices, edges and faces of the partition $\mathcal{A}$; their sets are denoted $\mathcal{V}$, $\mathcal{E}$ and $\mathcal{F}$, respectively.

By the duality transform the partition $\mathcal{A}$ is represented by the arrangement of the square $\mathcal{H}^* = \{(\alpha, \beta) \mid 0 \leq \alpha, -\beta \leq 1\,\}$ induced by the dual lines in P^* which intersect the square $\mathcal{L}^*$, namely the $m(m+1)/2$ lines duals of the lattice points (x, y) such that $1 \leq y \leq x \leq m$. The partition $\mathcal{A}$, for $m = 5$, is represented in the dual plane in Figure 2: there are 17 vertices, 40



$p_2 + \overrightarrow{p_3p_2}$
p_1
p_2
$p_2 + \overrightarrow{p_1p_2}$
p_3

Figure 3: Illustrate the proof of Proposition 6

edges and 24 faces.

The following proposition states precisely the relation between the partitions $\mathcal{A}$ and the partition induced by the mapping S.

Proposition 5 *Two lines ℓ and ℓ' of $\mathcal{L}$ define the same Sturmian m-factor, i.e. $S(\ell) = S(\ell')$ if and only if ℓ and ℓ' belong to the upper closure of the same face of the partition $\mathcal{A}$.* □

Proof. Follows from Proposition 4 and the discussion above.

In particular the function S is constant on each cell of the partition $\mathcal{A}$.

We denote still by S the extension of S to $\mathcal{A}$. The restriction of S to $\mathcal{F}$ is a bijection from $\mathcal{F}$ onto $\mathcal{S}$. The encoding of the Sturmian words by the faces of the partition $\mathcal{A}$ will be done by selecting a distinguished edge in the upper boundary of each face.

Proposition 6 *The upper (lower) boundary of a face of the partition $\mathcal{A}$ contains one or two edges.*

Proof. Assume on the contrary that the upper boundary of some face contains a sequence of three consecutive edges e_1, e_2 and e_3. Let p_i be the common lattice point of lines in e_i for $i = 1, 2, 3$. The configuration is depicted in Figure 3. We assume that the edges are indexed in decreasing slopes in the dual plane, and consequently the lattice points p_i are indexed in decreasing first coordinate. By assumption the lines (p_3p_2) and (p_2p_1) define the same Sturmian m-factor, and consequently, according to Proposition 4, the cone of lines in the segment $[(p_3p_2)(p_2p_1)]$ contains no lattice point in P

excepted the p_i. Now consider the points $p_2 + \overrightarrow{p_3p_2}$ and $p_2 + \overrightarrow{p_1p_2}$; they clearly belong to the cone and at least one of these points belongs to the lattice set P; contradiction. A similar argument is used for the lower boundary. □

Corollary 1 *The restrictions of the function S to the sets $\mathcal{V}, \mathcal{E}, \mathcal{F}$ are injective (but not surjective), surjective (but not injective) and bijective, respectively.* □

For $e \in \mathcal{E}$ we denote by $\hat{e}$ the common lattice point of the straight lines in e, and we denote by $\sup e$ $(\inf e)$ the straight line through $\hat{e}$ whose slope is the upper bound (lower bound) of the slopes of the lines in e. Clearly $e =]\inf e, \sup e[$. Next, recall that the Farey sequence $\mathcal{F}_n$ of order n is the increasing sequence of irreducible fractions between 0 and 1 whose denominators do not exceed n. If we draw a ray through the lattice point $(0, 0)$ and rotate it round the origin in the counter-clockwise direction from initial position along the axis x, it will pass in turn through each lattice point (p, q) such that q/p is a Farey fraction (see [8, chap. III page 29]).

Proposition 7 *Let $e \in \mathcal{E}$, let (a, b) be the coordinates of the lattice point $\hat{e}$, and let $y = \alpha x + \beta$ be the equation of the line $\sup e \in \mathcal{L}$ and $y = \alpha' x + \beta'$ be the equation of the line $\inf e \in \mathcal{H}$. Then*

(1) *α and α' are rationals, and $0 \leq \alpha' < \alpha \leq 1$*
(2) $0 \leq a \leq m$
(3) *α' and α are consecutive terms of the Farey serie $\mathcal{F}_{\max(m-a,a)}$.*
(4) $b = \lceil \alpha a \rceil$, $\beta = \lceil \alpha a \rceil - \alpha a$.
(5) *The Sturmian m-factor $S(e)$ is $B(a, p, q)$.*

Proof. Claims 1), 2), and 3) are obvious. Claim 4) follows from the relation $b = \alpha a + \beta$ with $0 \leq \beta < 1$. Finally claim 5) is obtained from the observation that $S(e) = S(\ell)$ where ℓ is any sufficiently small clockwise rotation of $\sup e$ around the point $\hat{e}$. □

Let $\alpha = q/p$ and $\alpha' = q'/p'$. An alternative way to compute $S(e)$ is to use the line in e whose slope is

$$\alpha'' = \frac{q + q'}{p + p'}.$$

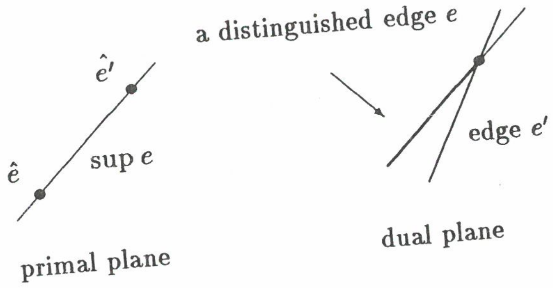


Figure 4: Distinguished edges

The computation of α' from α can be made in time $O(\log m)$ time [8, chap. III]. The use of α'' simplifies the practical implementation of B; indeed one gets

$$u_n = \lfloor \alpha'' n + \beta'' \rfloor$$

for all $n = 0, \ldots, m$. Thus, in practice, a call to a system routine (usualy called "drawline") suffices.

The mapping S is not injective on the set of edges $\mathcal{E}$. We therefore shall distinguish a special edge among all edges (in fact at most 2 according to Proposition 5) associated to some Sturmian m-factor. We distinguish the edge e for which the first coordinate of $\hat{e}$ is minimal or equivalently the edge of minimal slope in the dual plane. Let $\mathcal{E}'$ be the set of distinguished elements of $\mathcal{E}$.

It remains to identify the set of triples (a, p, q) associated to the edges of $\mathcal{E}'$. Here is the place where the set $\mathcal{T}$ appears.

Proposition 8 *The mapping $e \mapsto (a, p, q)$ which associates to each edge $e =]\inf e, \sup e[$ the triple (a, p, q) such that*

- *a is the first coordinate of the lattice point $\hat{e}$*
- *q/p (with p and q coprime) is the slope of $\sup e$*

is a bijection from $\mathcal{E}'$ onto the set $\mathcal{T}$.

Proof. An edge e is distinguished if and only if e is the rightmost edge (in the dual plane) of the upper boundary of some face of $\mathcal{A}$. In the primal plane this means that exists a lattice point in P on the line $\sup e$ whose first coordinate is greater than the first coordinate of $\hat{e}$ (see Figure 4). Straightforward calculations give then the result. □

The affiliation of Propositions 8, and 7 gives Theorem 1.

4 Recognition of Sturmian m-factors

In this section we consider the problem of testing whether a given word is Sturmian.

Let $s = v_1 v_2 \ldots v_m$ be any $\{0,1\}$-word of length m and let $u_0, \ldots, u_m$ be the integer sequence defined by

$$u_n = v_1 + v_2 + \cdots + v_n$$

(this is the "spectrum" of s in the sense of [2]).

The word s is a Sturmian m-factor if and only if there exists a line $\ell \in \mathcal{L}$ such that $S(\ell) = s$. Let α and β be the slope and the intercept of the line ℓ. The word s is a Sturmian m-factor if and only if the system of linear inequations

$$u_n \leq \alpha n + \beta < u_n + 1 \qquad n = 1, 2, \ldots, m \tag{11}$$

admits a solution. Observe that each letter in the word s adds two constraints to the above linear system.

Using Megiddo's algorithm [19, chap 15] the existence of a solution to (11) can be decided in linear time in the number of constraints i.e., in time $O(m)$. However we can do much better in our case.

Proposition 9 *Given a word in $\{0,1\}^*$, its maximal Sturmian prefix can be computed on-line in time proportional to its size.*

Proof. For each Sturmian prefix we maintain the minimal representation of the polygon (in the space of lines) of all lines defining this Sturmian factor. We incrementaly transform this polygon by adding the two geometric constraints defined by a new letter and by removing the superfluous constraints. Since, according to Proposition 6, the size of the current polygon is 3 or 4 this amounts to a constant number of operations. Each operation requires to compute the position of a lattice point with respect to a lattice line and thus involves only the computation of the sign of an integral determinant. □

Acknowledgement

We would like to thank Dominique Gouyou-Beauchamps for helpful discussions.

References

[1] J. Berstel and M. Pocchiola. A geometric proof of the enumeration formula for Sturmian words. *To appear in Intern. J. Alg. Comput.* Also available as TR LIENS-92-21, or by anonymous ftp on the machine spi.ens.fr in the directory pub/reports/liens.

[2] M. Boshernitzan and A. S. Fraenkel. A linear algorithm for nonhomogeneous spectra of numbers. *J. Algorithms*, 5:187–198, 1984.

[3] J. E. Bresenham. Algorithm for computer control of a digital plotter. *IBM Systems J.*, 4:25–30, 1965.

[4] T. C. Brown. A characterization of the quadratic irrationals. *Canad. Math. Bull.*, 34:36–41, 1991.

[5] S. Dulucq and D. Gouyou-Beauchamps. Sur les facteurs des suites de sturm. *Theoret. Comput. Sci.*, 71:381–400, 1991.

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

[7] A. S. Fraenkel, M. Mushkin, and U. Tassa. Determination of $\lfloor n\theta \rfloor$ by its sequence of differences. *Canad. Math. Bull.*, 21:441–446, 1978.

[8] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers.* Oxford Science Publications, 1979.

[9] G. Hedlund. Sturmian minimal sets. *Amer. J. Math.*, 66:605–620, 1944.

[10] G. Hedlund and M. Morse. Symbolic dynamics. *Amer. J. Math*, 60:815–866, 1938.

[11] G. Hedlund and M. Morse. Sturmian sequences. *Amer. J. Math*, 61:1–42, 1940.

[12] S. Ito and S. Yasutomi. On continued fractions, substitutions and characteristic sequences. *Japan. J. Math.*, 16:287–306, 1990.

[13] W. F. Lunnon and P. A. B. Pleasants. Quasicrystallographic tilings. *J. Math. Pures Appl.*, 66:217–263, 1987.

[14] F. Mignosi. On the number of factors of Sturmian words. *Theoret. Comput. Sci.*, 82:71–84, 1991.

[15] M. Queffélec. *Substitution Dynamical Systems-Spectral Analysis*, volume 1294 of *Lecture Notes Math.* Springer-Verlag, 1987.

[16] G. Rauzy. Suites à termes dans un alphabet fini. In *Sémin. Théorie des Nombres*, pages 25-01–25-16. Bordeaux, 1982–1983.

[17] G. Rauzy. Mots infinis en arithmétique. In D. Perrin, editor, *Automata on infinite words*, volume 192, pages 165–171. Lect. Notes Comp. Sci., Springer-Verlag, 1985.

[18] G. Rauzy. Sequences defined by iterated morphisms. In R. Capocelli, editor, *Workshop on Sequences.* Lect. Notes Comp. Sci., Springer-Verlag, 1991.

[19] A. Schrijver. *Theory of Linear and Integer Programming.* Wiley, 1986.

[20] P. Séébold. Fibonacci morphisms and Sturmian words. *Theoret. Comput. Sci.*, 88:367–384, 1991.

[21] C. Series. The geometry of Markoff numbers. *The Mathematical Intelligencer*, 7:20–29, 1985.

[22] K. B. Stolarsky. Beatty sequences, continued fractions, and certain shift operators. *Cand. Math. Bull.*, 19:473–482, 1976.

[23] B. A. Venkov. *Elementary Number Theory.* Wolters-Noordhoff, Groningen, 1970.