Uniformly balanced words with linear complexity and prescribed letter frequencies

V. Berthé
Laboratoire d'Informatique Algorithmique : Fondements et Applications
Université Paris Diderot
Paris 7 - Case 7014
F-75205 Paris Cedex 13, France
berthe@liafa.jussieu.fr

S. Labbé
Laboratoire de Combinatoire et d'Informatique Mathématique,
Université du Québec à Montréal,
C.P. 8888 Succursale "Centre-Ville", Montréal (QC), Canada H3C 3P8
labbe.sebastien@courrier.uqam.ca

We consider the following problem. Let us fix a finite alphabet $\mathscr{A} = \{1, 2, \cdots, d\}$; for any d-uple of letter frequencies $(f_1, \cdots, f_d) \in [0,1]^d$ with $\sum_{i=1}^d f_i = 1$, how to construct an infinite word u over the alphabet $\mathscr{A}$ satisfying the following conditions: u has linear complexity function, u is uniformly balanced, the letter frequencies in u are given by $(f_1, \cdots, f_d)$. This paper investigates a construction method for such words based on the use of mixed multidimensional continued fraction algorithms.

Keywords: balanced words, discrepancy, letter frequency, multidimensional continued fractions, discrete geometry

1 Introduction

We consider the following problem: let us fix a finite alphabet $\mathscr{A} = \{1, 2, \cdots, d\}$; for any d-uple of letter frequencies $(f_1, \cdots, f_d) \in [0,1]^d$ with $\sum_{i=1}^d f_i = 1$, how to construct an infinite word u over the alphabet $\mathscr{A}$ satisfying the following conditions:

1. u has linear complexity function
2. u is uniformly balanced
3. the letter frequencies in u are given by $(f_1, \cdots, f_d)$.

Let us first recall several definitions in order to clarify the previous statement. A word $u \in \mathscr{A}^{\mathbb{N}}$ is said to be *uniformly balanced* if there exists a constant $C > 0$ such that for any pair of factors of the same length v, w of u, and for any letter $i \in \mathscr{A}$,

$$||v|_i - |w|_i| \leq C,$$

where the notation $|x|_j$ stands for the number of occurrences of the letter j in the factor x. A word u has *linear complexity function* if there exists a constant $C' > 0$ such that the number of factors of u of length n is smaller than $C' \cdot n$, for every positive integer n. The *frequency* f_i of a letter $i \in \mathscr{A}$ in $u = (u_n)_{n \in \mathbb{N}}$ is defined as the limit (when n tends towards infinity), if it exists, of the number of occurrences of i in $u_0 u_1 \dots u_{n-1}$ divided by n.

Preliminary Report. Final version to appear in:
WORDS 2011

This problem has several motivations. The first one comes from discrete geometry: such an infinite word can be seen as a coding of a discrete line in $\mathbb{Z}^d$. Indeed one associates with any infinite word over the alphabet $\mathscr{A}$ a broken line obtained as a stair made of a union of segments of unit length directed according to the coordinate axes, whose vertices are obtained by replacing each of the letters of u by one of the canonical basis vectors and by concatenating these vectors. Let $\mathbf{l}: A^* \to \mathbb{N}^n,\ w \mapsto {}^t(|w|_{a_1}, \ldots, |w|_{a_n})$ stand for the *abelianisation map* or the *Parikh mapping*. More precisely, the set of vertices of this broken line is equal to $\{\mathbf{l}(u_0 \cdots u_{N-1}) \mid N \in \mathbb{N}\}$. The question is to know how well the line associated with the word u approximates the Euclidean line directed by the vector of letter frequencies of u, when they exist. There exist various strategies for defining and generating discrete lines in the three-dimensional space. With no claim for being exhaustive, let us quote e.g. [2, 8, 13, 22]. Nevertheless, they do not fulfill Condition 1. on the linear complexity. Note that the notion of discrete line defined in [2] corresponds to billiard words. Condition 1. means here that these discrete lines are "simple" in terms of number of local configurations.

The second motivation comes from symbolic dynamical systems and Diophantine approximation: is it possible to define a Rauzy fractal associated with any translation of the torus? More precisely, assume we are given a translation $x \mapsto x + (\alpha_1, \cdots \alpha_d)$ defined on $\mathbb{T}^d = \mathbb{R}^d/\mathbb{Z}^d$; the Rauzy fractal $\mathscr{R}$ associated with an infinite word u over the d-letter alphabet $\mathscr{A}$ is defined by projecting along the frequency vector of u on a transverse hyperplane the vertices of the broken line associated with u (such as described above) and then, by taking the closure. For more on Rauzy fractals, see e.g. [7]. The problem now becomes the following: is it possible to construct an infinite word u over the d-letter alphabet $\mathscr{A}$ such that $\mathscr{R}$ is a compact set that tiles periodically this transverse hyperplane and such that u has linear complexity? Let us explain in this context the requirement concerning linear complexity (Condition 1.): we would like to recover from the dynamical and combinatorial properties of the infinite word u arithmetical information on the parameters underlying the translation on the torus. This will be easier if u has low complexity function, i.e., a low numbers of factors. Let us quote as a further motivation uniform distribution and the so-called chairman assignment problem, see e.g. [21], and the references therein.

There exist families of words that satisfy Conditions 2. and 3. but not Condition 1. Billiard words are defined as codings of trajectories of billiards in a cube; they are shown to have quadratic complexity (see [4, 5]). They satisfy Conditions 2. and 3. Let us also quote the construction described in [11] which produces step by step a broken line whose vertices belong to $\mathbb{Z}^3$ that approximates a given direction by choosing at each step the closest point. It is proved in [11] that such a broken line can be obtained by selecting integer points by shifting a polygonal window along the line. The complexity is here again quadratic. The corresponding infinite words satisfy Conditions 2. and 3. Note that 1-balanced words over a higher-alphabet do not seem to be good candidates for describing discrete segments in the space: not all frequencies can be reached. Fraenkel's conjecture states that the possible frequencies for 1-balanced words are rational and uniquely determined, when they are assumed to be distinct [15]. In particular, when $k = 3$, the only possible 1-balanced word is $(1213121)^\infty$ (if frequencies are distinct), up to a permutation of letters and up to shifts. For the irrational case, see [17] and [16]. For more references on the subject, see also the survey [23]. Note also that Arnoux-Rauzy words (see e.g. [**?**, 10, 9]) are infinite words that do not satisfy Condition 2., such as proved in [10], but that do satisfy Conditions 1. and 3. Furthermore, they are not defined for every d-uple of letter frequencies, but only for a set of zero measure in $[0,1]^d$. For an illustration, see Figure 4.

2 Multidimensional continued fractions and frequencies

The strategy we consider here for constructing infinite words satisfying the three above mentioned conditions consists in applying a multidimensional continued fraction algorithm to the frequency vector $(f_1, \cdots, f_d)$, according to [6]. We then associate with the steps of the algorithm substitutions, that is, rules that replace letters by words, with these substitutions having the matrices produced by the continued fraction algorithm as incidence matrices. More precisely, a *substitution* σ over the alphabet $\mathscr{A}$ is an endomorphism of the free monoid $\mathscr{A}^*$, and the *incidence matrix* of the substitution σ is the square matrix M_σ with entries $m_{i,j} = |\sigma(j)|_i$ for all $i, j \in \mathscr{A}$.

Let us recall the most classical multidimensional continued fraction algorithms such as described e.g. in [19], and in [10, 9, 24] for Arnoux-Rauzy algorithm. For the sake of simplicity, we express them in dimension $d = 3$:

- Jacobi-Perron: let $0 \leq u_1, u_2 \leq u_3$,

$$(u_1, u_2, u_3) \mapsto (u_2 - [\frac{u_2}{u_1}]u_1, u_3 - [\frac{u_3}{u_1}]u_1, u_1).$$

- Brun: we subtract the second largest entry from the largest one; for instance, if $0 \leq u_1 \leq u_2 \leq u_3$,

$$(u_1, u_2, u_3) \mapsto (u_1, u_2, u_3 - u_2).$$

- Poincaré: we subtract the second largest entry to the largest one, and the smallest entry from the second largest one; for instance, if $0 \leq u_1 \leq u_2 \leq u_3$,

$$(u_1, u_2, u_3) \mapsto (u_1, u_2 - u_1, u_3 - u_2).$$

- Selmer: we subtract the smallest positive entry from the largest one; for instance, if $0 < u_1 \leq u_2 \leq u_3$,

$$(u_1, u_2, u_3) \mapsto (u_1, u_2, u_3 - u_1).$$

- Fully subtractive: we subtract the smallest positive entry from all the largest ones; for instance, if $0 < u_1 \leq u_2 \leq u_3$,

$$(u_1, u_2, u_3) \mapsto (u_1, u_2 - u_1, u_3 - u_1).$$

- Arnoux-Rauzy: let $0 \leq u_1 \leq u_2 \leq u_3$ with $u_3 \geq u_1 + u_2$,

$$(u_1, u_2, u_3) \mapsto (u_1, u_2, u_3 - u_1 - u_2).$$

  otherwise the algorithm stops.

Let T be one of these algorithms applied to some vector $(f_1, f_2, f_3) \in [0,1]^3$. With each matrix M produced by T, we associate a substitution whose incidence matrix is given by M. We thus obtain a word by iterating these substitutions in an S-adic way. We recall that a word is said to be *S-adic* if it is generated by composing a finite number of substitutions. This covers various families of words with a rich dynamical behavior such as Sturmian sequences; for more on S-adic words, see e.g. [3, 12].

3 Fusion algorithms

We can also mix these algorithms by performing at each step one among these rules, and this still yields S-adic generated words. We call such algorithms *fusion algorithms*. We focus on fusion algorithms obtained by applying Arnoux-Rauzy algorithm when possible, and otherwise, consistently one algorithm among Brun, Poincaré, Selmer, or the Fully Subtractive algorithms. Indeed, experimental studies indicate that a combination of Arnoux-Rauzy steps with Brun steps, or with Poincaré steps produces good performances (see Table 1 and Figure 5 below), and even better performances than when performing only one algorithm. Furthermore, this allows us to exploit and extend the good mean behaviour of Arnoux-Rauzy algorithm to a larger set of parameters (compare Figure 4 and Figure 5).

The aim of this lecture is to study the properties of such fusion algorithms for both finite (rational frequencies) and infinite expansions (irrational frequencies). In particular, we will focus on the almost everywhere convergence properties and ergodic properties of these fusion algorithms when the frequency vector has irrational coordinates. The proof relies on classic techniques such as described e.g. in [19].

	Minimum	Mean	Maximum	Std
Arnoux-Rauzy	0.6000	0.9055	1.200	0.1006
Fully subtractive	0.6000	5.982	13.92	4.388
Fully subtractive as possible	0.6000	4.172	25.00	4.440
Selmer	0.5000	2.184	12.75	2.070
Brun	0.5000	1.114	2.000	0.2664
Brun Multiplicative	0.6000	1.117	2.000	0.2681
Poincaré	0.6000	2.527	11.13	2.261
Jacobi-Perron	0.6000	2.731	25.00	3.456
Random reduction	0.5000	2.426	24.99	2.779
Fusion of Arnoux-Rauzy and Fully subtractive	0.6000	1.095	2.800	0.3105
Fusion of Arnoux-Rauzy and Selmer	0.6000	0.9678	1.450	0.1438
Fusion of Arnoux-Rauzy and Brun Multiplicative	0.6000	0.9132	1.400	0.1143
Fusion of Arnoux-Rauzy and Poincaré	0.6000	0.8941	1.200	0.09733

Table 1: Statistics (minimum, mean, maximum, standard deviation) for the discrepancy for triplets of nonnegative rational vectors $(a_1/N, a_2/N, a_3/N)$ such that $a_1+a_2+a_3 = N$ with $N = 100$.

Consider now the case of rational frequencies. Table 1 displays some experimental results. We work here in dimension $d = 3$ with rational frequency vectors of the form $\mathbf{f} = (a_1/N, a_2/N, a_3/N)$, with $a_i \in \mathbb{N}$, $i = 1,2,3$, and with $a_1+a_2+a_3 = N$ being a positive integer. We apply a fusion algorithm to such a triplet, until we reach a vector whose entries are all equal to 0 but one. This produces a finite sequence of matrices, and thus, of substitutions, having these matrices as incidence matrices. Note that we have several choices for these substitutions, even if the incidence matrices have entries in $\{0,1\}$. Given a matrix M, we thus have to decide in which order letters will be chosen in the image of a letter by a substitution σ having M as incidence matrix. We choose as a convention to put the most frequent letter first. (This (partly) explains why the triangles obtained in Figure 1, 2, 3, 4, 5 are not perfectly symmetric.) Let us apply now to $\mathbf{f}$ a finite sequence of steps of a fusion algorithm together with a choice of substitutions associated with the produced matrices. One has $\mathbf{f} = M_1 \cdots M_n \mathbf{f}_n$, where the vector $\mathbf{f}_n$ has two coordinates equal to 0, and one non-zero coordinate of index, say $w_n \in \{1,2,3\}$. The associated substitutions are denoted by σ_k, for $1 \le k \le n$. The following diagram illustrates how we produce finite

words w with frequency vector $\mathbf{f}$:

$$\mathbf{f} = \mathbf{f}_0 \xleftarrow{M_1} \mathbf{f}_1 \xleftarrow{M_2} \mathbf{f}_2 \xleftarrow{M_3} \cdots \xleftarrow{M_n} \mathbf{f}_n$$

$$w = w_0 \xleftarrow{\sigma_1} w_1 \xleftarrow{\sigma_2} w_2 \xleftarrow{\sigma_3} \cdots \xleftarrow{\sigma_n} w_n \in \{1,2,3\}$$

The experimental results of Table 1 indicate that the fusion algorithm obtained when applying Arnoux-Rauzy algorithm when possible, and otherwise, Poincaré algorithm, behaves in an efficient way with respect to the discrepancy. The *discrepancy* of a finite word $u_0 \cdots u_n \in \mathscr{A}^{n+1}$ is defined as

$$\max_{i \in \mathscr{A},\ 0 \le k \le n} |f_i \cdot k - |u_0 \cdots u_k|_i|.$$

This distance is considered e.g. in [21] and [1], and is intimately connected with the following balance measure. The *balance* of $u_0 \cdots u_n \in \mathscr{A}^{n+1}$ is defined as

$$\max_{i \in \mathscr{A},\ |v|=|w|} ||v|_i - |w|_i|,$$

(here v, w are factors of u of the same length $|v| = |w|$). We have chosen here to use the discrepancy for our numerical experiments in order to compare our results with the bound discussed in [21]. Indeed, in [21], an algorithm is given that produces, for any given frequency vector $(f_1, \cdots, f_d)$, an infinite word whose discrepancy is smaller than or equal to $3/4$. However, the factor complexity of such a word does not seem to be known. In the fusion algorithm obtained by combining Arnoux-Rauzy algorithm with Poincaré algorithm, one obtains a mean discrepancy equal to 0.8910 when $N = 100$. More generally, Figure 1, 2, 3, 4, 5 below illustrate the behaviour of the discrepancy for triplets of nonnegative rational vectors $(a_1/N, a_2/N, a_3/N)$ such that $a_1 + a_2 + a_3 = N$ for a given N.

min	0.6000
mean	1.484
max	3.000
std	0.6137

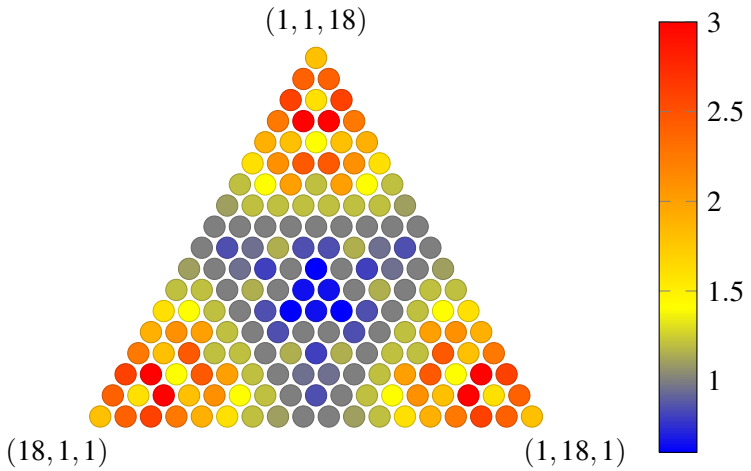


Figure 1: Discrepancy for triplets of nonnegative rational vectors $(a_1/N, a_2/N, a_3/N)$ such that $a_1 + a_2 + a_3 = N$ with $N = 20$ using Poincaré algorithm.

min	0.6000
mean	5.982
max	13.92
std	4.388

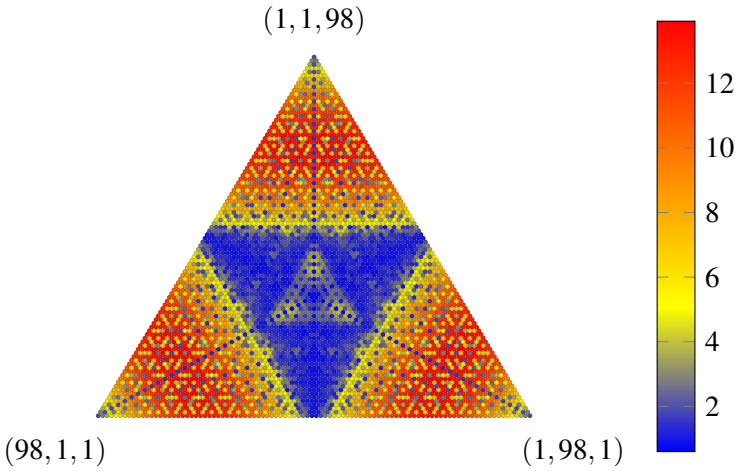


Figure 2: Discrepancy for triplets of nonnegative rational vectors $(a_1/N, a_2/N, a_3/N)$ such that $a_1+a_2+a_3=N$ with $N=100$ using Fully subtractive algorithm.

min	0.6000
mean	2.527
max	11.13
std	2.261

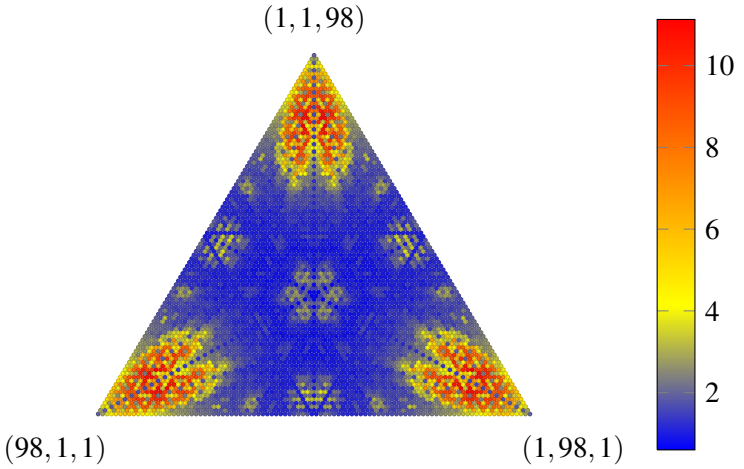
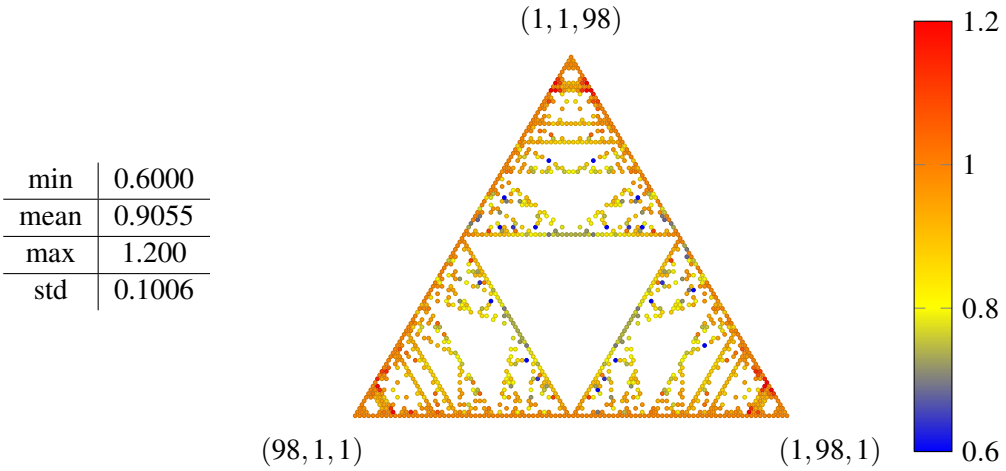
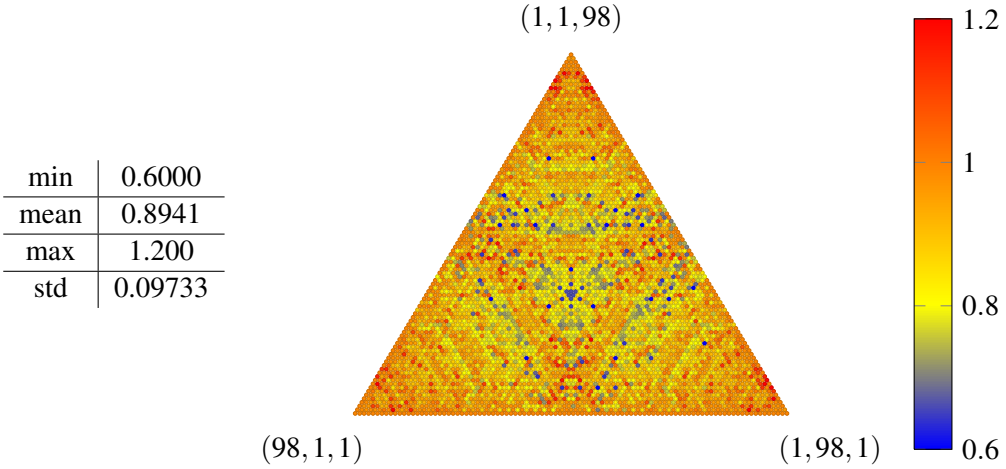


Figure 3: Discrepancy for triplets of nonnegative rational vectors $(a_1/N, a_2/N, a_3/N)$ such that $a_1+a_2+a_3=N$ with $N=100$ using Poincaré algorithm.



min	0.6000
mean	0.9055
max	1.200
std	0.1006

Figure 4: Discrepancy for triplets of nonnegative rational vectors $(a_1/N, a_2/N, a_3/N)$ such that $a_1+a_2+a_3=N$ with $N=100$ using Arnoux-Rauzy algorithm. This algorithm is defined only for vectors whose largest entry is greater than or equal to the sum of the other two.



min	0.6000
mean	0.8941
max	1.200
std	0.09733

Figure 5: Discrepancy for triplets of nonnegative rational vectors $(a_1/N, a_2/N, a_3/N)$ such that $a_1+a_2+a_3=N$ with $N=100$ using a fusion of Arnoux-Rauzy and Poincaré algorithms.

Acknowledgements

We would like to thank warmly J. Shallit for pointing out reference [21]. This research was driven by computer exploration using the open-source mathematical software *Sage* [20] and its library on Combinatorics on words developed by the *Sage-Combinat* community, and in particular by the active developers: A. Blondin Massé, V. Delecroix, S. Labbé, T. Monteil and F. Saliola.

References

[1] B. Adamczewski (2003): *Balances for fixed points of primitive substitutions*. *Theoret. Comput. Sci.* 307(1), pp. 47–75, doi:10.1016/S0304-3975(03)00092-6. Available at `http://dx.doi.org/10.1016/`

S0304-3975(03)00092-6. Words.

[2] E. Andres (2003): *Discrete linear objects in dimension n: the standard model*. Graphical Models 65, pp. 92–111.

[3] P. Arnoux & V. Berthé (2002): *Some open problems*. In: Substitutions in dynamics, arithmetics and combinatorics, Lecture Notes in Math. 1794, Springer, Berlin, pp. 363–374.

[4] P. Arnoux, C. Mauduit, I. Shiokawa & J. i. Tamura (1994): *Complexity of sequences defined by billiard in the cube*. Bull. Soc. Math. France 122(1), pp. 1–12. Available at http://www.numdam.org/item?id=BSMF_1994__122_1_1_0.

[5] Yu. Baryshnikov (1995): *Complexity of trajectories in rectangular billiards*. Comm. Math. Phys. 174(1), pp. 43–56. Available at http://projecteuclid.org/getRecord?id=euclid.cmp/1104275093.

[6] V. Berthé & S. Labbé (2011): *An Arithmetic and Combinatorial Approach to Three-Dimensional Discrete Lines*. In I. Debled-Rennesson, E. Domenjoud, B. Kerautret & P. Even, editors: DGCI, Lecture Notes in Computer Science 6607, Springer, pp. 47–58. Available at http://dx.doi.org/10.1007/978-3-642-19867-0_4.

[7] V. Berthé & A. Siegel (2005): *Tilings associated with beta-numeration and substitutions*. Integers 5(3), pp. A2, 46.

[8] V. E. Brimkov, R. P. Barneva & B. Brimkov (2009): *Minimal Offsets That Guarantee Maximal or Minimal Connectivity of Digital Curves in nD*. In S. Brlek, C. Reutenauer & X. Provençal, editors: DGCI, Lecture Notes in Computer Science 5810, Springer, pp. 337–349. Available at http://dx.doi.org/10.1007/978-3-642-04397-0_29.

[9] J. Cassaigne, S. Ferenczi & A. Messaoudi (2008): *Weak mixing and eigenvalues for Arnoux-Rauzy sequences*. Ann. Inst. Fourier (Grenoble) 58(6), pp. 1983–2005. Available at http://aif.cedram.org/item?id=AIF_2008__58_6_1983_0.

[10] J. Cassaigne, S. Ferenczi & L. Q. Zamboni (2000): *Imbalances in Arnoux-Rauzy sequences*. Ann. Inst. Fourier (Grenoble) 50(4), pp. 1265–1276. Available at http://www.numdam.org/item?id=AIF_2000__50_4_1265_0.

[11] N. Chevallier (2009): *Coding of a translation of the two-dimensional torus*. Monatsh. Math. 157(2), pp. 101–130, doi:10.1007/s00605-008-0074-y. Available at http://dx.doi.org/10.1007/s00605-008-0074-y.

[12] F. Durand (2003): *Corrigendum and addendum to: "Linearly recurrent subshifts have a finite number of non-periodic subshift factors" [Ergodic Theory Dynam. Systems* **20** *(2000) 1061–1078]*. Ergodic Theory Dynam. Systems 23(2), pp. 663–669, doi:10.1017/S0143385702001293. Available at http://dx.doi.org/10.1017/S0143385702001293.

[13] O. Figueiredo & J.-P. Reveillès (1996): *New results about 3D digital lines*. In: Proc. Internat. Conference Vision Geometry V , Proc. SPIE, 2826, pp. 98–108.

[14] R. Fischer & F. Schweiger (1975): *The number of steps in a finite Jacobi algorithm*. Manuscripta Math. 17(3), pp. 291–308.

[15] A. S. Fraenkel (1973): *Complementing and exactly covering sequences*. J. Combinatorial Theory Ser. A 14, pp. 8–20.

[16] R. L. Graham (1973): *Covering the positive integers by disjoint sets of the form* $\{[n\alpha+\beta] : n = 1, 2, \ldots\}$. J. Combinatorial Theory Ser. A 15, pp. 354–358.

[17] P. Hubert (2000): *Suites équilibrées*. Theoret. Comput. Sci. 242(1-2), pp. 91–108, doi:10.1016/S0304-3975(98)00202-3. Available at http://dx.doi.org/10.1016/S0304-3975(98)00202-3.

[18] R. Morikawa (1982/83): *On eventually covering families generated by the bracket function*. Bull. Fac. Liberal Arts Nagasaki Univ. 23(1), pp. 17–22.

[19] F. Schweiger (2000): *Multidimensinal Continued Fraction*. Oxford Univ. Press, New York.

[20] W. A. Stein et al. (2011): *Sage Mathematics Software (Version 4.7)*. The Sage Development Team. http://www.sagemath.org.

[21] R. Tijdeman (1980): *The chairman assignment problem*. *Discrete Math.* 32(3), pp. 323–330, doi:10.1016/0012-365X(80)90269-1. Available at `http://dx.doi.org/10.1016/0012-365X(80)90269-1`.

[22] J.-L. Toutant (2006): *Characterization of the Closest Discrete Approximation of a Line in the 3-Dimensional Space*. In: *ISVC (1)*, *Lecture Notes in Computer Science* 4291, Springer, pp. 618–627. Available at `http://dx.doi.org/10.1007/11919476_62`.

[23] L. Vuillon (2003): *Balanced words*. *Bull. Belg. Math. Soc. Simon Stevin* 10(suppl.), pp. 787–805. Available at `http://projecteuclid.org/getRecord?id=euclid.bbms/1074791332`.

[24] N. Wozny & L. Q. Zamboni (2001): *Frequencies of factors in Arnoux-Rauzy sequences*. *Acta Arith.* 96(3), pp. 261–278, doi:10.4064/aa96-3-6. Available at `http://dx.doi.org/10.4064/aa96-3-6`.