Theory of Computing Systems **40** (2007), 205–223

DISTRIBUTION OF ADDITIVE FUNCTIONS WITH RESPECT TO NUMERATION SYSTEMS ON REGULAR LANGUAGES

PETER J. GRABNER† AND MICHEL RIGO

Abstract. We study the distribution of values of additive functions related to numeration systems defined via regular languages.

1. Introduction

Additive numeration systems and the corresponding additive arithmetic functions have been studied from various points of view since the seminal papers of H. Delange [7, 8], where such functions were investigated for the usual k-adic numeration system. Later more exotic systems of numeration, such as general linear numeration systems [16, 17], especially such systems defined by linear recurring sequences were considered. Different aspects of such representations of the integers were studied: dynamics of corresponding adding machine ("odometer") [18], topological dynamics of the odometer [1], asymptotic properties of summatory functions of additive functions such as the "sum-of-digits" function [11, 12, 15, 20, 21], local and global versions of central limit theorems for the values of additive functions [9, 10, 13], existence of distribution functions of additive functions [2].

In [22] numeration systems based on regular languages have been introduced. Let $\mathcal{L}$ be a regular language on the ordered alphabet Σ, then $\mathcal{L}$ is equipped with the genealogical ordering defined in Section 2. The positive integer n is then represented as the $n+1$-st word in $\mathcal{L}$ with respect to this ordering. The usual k-adic numeration systems, recursion based numeration systems as studied in [10, 21], and numeration systems related to substitutions on finite alphabets (cf. [11, 13]) are all special cases of this general concept. The notion of the "odometer" has been extended to this type of numeration system in [4].

In the present paper we continue investigations initiated in [19], where the asymptotic behaviour of summatory functions of additive functions was studied. We will analyze the distribution of the values of additive functions related to regular languages. In particular, our results can be used to derive information on the distribution of the letters in a given regular language. Precise results (local and global limit theorems) of that type for languages with primitive adjacency matrix have been obtained in [5]. It will turn out that in general there is no central limit theorem for the values of such additive functions, which is in

Date: February 14, 2008.

2000 *Mathematics Subject Classification.* Primary: 11A67, Secondary: 68Q45, 11K65, 60F05.

Key words and phrases. regular languages, numeration systems, additive functions, Gaussian limit distribution.

† This author is supported by the START project Y96-MAT of the Austrian Science Fund.

striking contrast to the usual situation (cf. [10, 13]). We make precise the conditions for such a limit theorem to hold.

2. Preliminaries

Let Σ be a finite alphabet. The free monoid generated by Σ with identity ε is Σ^*. If w is a word over Σ, $|w|$ denotes its length. We assume that the reader is familiar with classical notions of formal languages theory like (minimal) automaton or regular language (see for instance [14]).

Describing an infinite regular language $\mathcal{L}$ over a totally ordered alphabet $(\Sigma, <)$ with respect to the genealogical ordering (cf. [25]) gives a one-to-one and onto increasing mapping between $\mathbb{N}$ and $\mathcal{L}$. If w is the n-th word of the genealogically ordered language $\mathcal{L}$, $n \in \mathbb{N} \setminus \{0\}$, then we denote by $\mathrm{val} : \mathcal{L} \to \mathbb{N}$ the application mapping w onto $n-1$. The integer $\mathrm{val}(w)$ is said to be the *numerical value* of w. So each non-negative integer n is represented by a unique word $\mathrm{val}^{-1}(n) \in \mathcal{L}$ and this leads to the notion of *numeration system on a regular language*. These systems have been introduced in [22] and generalize classical numeration systems like the k-adic systems, the Fibonacci system and the linear numeration systems whose characteristic polynomial is the minimal polynomial of a Pisot number (for the properties of these latter systems we refer to [6]).

In this paper, $\mathcal{L}$ always refers to an infinite regular language having $\mathcal{M}_{\mathcal{L}} = (Q, \Sigma, s, \delta, F)$ as trimmed minimal automaton (to obtain unambiguous constructions, we only consider minimal automata; in order to relate the size of the language with the eigenvalues of the incidence matrix, we assume the automaton to be trimmed) and we represent integers using the numeration system built upon $\mathcal{L}$ for a given ordering of the alphabet. We often write $q.w$ as a shorthand for $\delta(q, w)$, $q \in Q$, $w \in \Sigma^*$. Recall that $\mathcal{M}_{\mathcal{L}}$ is said to be *trimmed* if it is *accessible*, i.e., for all $q \in Q$, there exists $w \in \Sigma^*$ such that $s.w = q$ and *coaccessible*, i.e., for all $q \in Q$, there exists $w \in \Sigma^*$ such that $q.w \in F$. The incidence matrix A of $\mathcal{M}_{\mathcal{L}}$ is defined by $A_{p,q} = \#\{\sigma \in \Sigma \mid p.\sigma = q\}$.

For each state $q \in Q$, we define the language

$$\mathcal{L}_q = \{w \in \Sigma^* \mid \delta(q, w) \in F\}$$

of the words accepted by $\mathcal{M}_{\mathcal{L}}$ from q. In particular, $\mathcal{L} = \mathcal{L}_s$. Since $\mathcal{L}_q$ is a regular language, it can be genealogically ordered and this leads to a new numeration system on $\mathcal{L}_q$ where the function mapping the words of $\mathcal{L}_q$ onto their corresponding numerical value is simply denoted $\mathrm{val}_q : \mathcal{L}_q \to \mathbb{N}$. In particular, $\mathrm{val}_s = \mathrm{val}$. (If $\mathcal{L}_q$ is finite then the domain of val_q is finite and its image is restricted to $\{0, \ldots, \#\mathcal{L}_q - 1\}$.)

For each state $q \in Q$, we define two functions $u_q(n)$ and $v_q(n)$ counting the number of words in $\mathcal{L}_q$ respectively of length n and of length less or equal to n,

$$u_q(n) = \#(\mathcal{L}_q \cap \Sigma^n) \quad \text{and} \quad v_q(n) = \#(\mathcal{L}_q \cap \Sigma^{\leq n}).$$

Using the definition of the genealogical ordering, a formula for computing numerical values can be derived.

Lemma 1. [22] *If σy belongs to $\mathcal{L}_q$, $y \in \Sigma^+ = \Sigma^* \setminus \{\varepsilon\}$, $\sigma \in \Sigma$ then*

$$\mathrm{val}_q(\sigma y) = \mathrm{val}_{q.\sigma}(y) + v_q(|y|) - v_{q.\sigma}(|y|-1) + \sum_{\sigma'<\sigma} u_{q.\sigma'}(|y|).$$

Iterating this formula and reordering the summands we obtain a formula analogous to the classical decomposition of an integer in the k-adic system. Here the powers of k are replaced by the different sequences $u_q(n)$'s. Let $w = w_1 \cdots w_n \in \mathcal{L}$, we have

$$\mathrm{val}(w) = \sum_{q\in Q} \sum_{i=1}^{|w|} \beta_{q,i}(w) u_q(|w|-i) \tag{2.1}$$

where

$$\beta_{q,i}(w) := \#\{\sigma < w_i \mid s.w_1 \cdots w_{i-1}\sigma = q\} + \delta_{q,s} \quad \text{for } i = 1, \ldots, |w| \tag{2.2}$$

Observe that these coefficients are bounded :

$$0 \le \sum_{q\in Q} \beta_{q,i}(w) \le \#\Sigma.$$

In the following we will shortly discuss how to decompose the automaton $\mathcal{M}_\mathcal{L}$ into irreducible components. We will follow the description given in [23, § 4.4]. We say that two states p and q *communicate*, if there exist words $w_1, w_2 \in \Sigma^*$ such that $p.w_1 = q$ and $q.w_2 = p$. Clearly, this defines an equivalence relation on Q. The equivalence classes are called *communicating classes*. The automaton $\mathcal{M}_\mathcal{L}$ induces a graph $\mathcal{G}_\mathcal{L}$ on the set of communicating classes in the following way: let $\mathcal{C}_1$ and $\mathcal{C}_2$ be classes, then there is an oriented edge $\mathcal{C}_1 \to \mathcal{C}_2$, if there exist states $q_1 \in \mathcal{C}_1$ and $q_2 \in \mathcal{C}_2$ and a word $w \in \Sigma^+$ such that $q_1.w = q_2$. This graph does not contain oriented cycles. Furthermore, there exist classes, which have no outgoing edges. Because, if every class were to have an outgoing edge then we would obtain an oriented cycle. Such classes will be called *sink classes*. Proceeding as in [23] the communicating classes and the corresponding states can be ordered in such a way that the incidence matrix A of $\mathcal{M}_\mathcal{L}$ has a block triangular form

$$A = \begin{pmatrix} A_1 & * & * & \ldots & * \\ 0 & A_2 & * & \ldots & * \\ 0 & 0 & A_3 & \ldots & * \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & A_k \end{pmatrix},$$

where the matrices A_i are irreducible and $*$ stands for possibly non-zero matrices. From now on we will assume that the following hypothesis holds for the language $\mathcal{L}$.

Hypothesis 1. The adjacency matrices A_i of the communicating classes are all primitive.

Remark 1. Recall that a nonnegative square matrix M is *primitive* if there exists an integer n such that $M^n > 0$ (the inequality being interpreted element-wise). In this case, the Perron-Frobenius theorem holds and M has a unique dominating real eigenvalue λ_M referred as the *Perron-Frobenius eigenvalue* of M (cf. [26]).

We will also make use of the same assumption as in [19]

Hypothesis 2. The language $\mathcal{L}$ is exponential and the adjacency matrix of $\mathcal{M}_{\mathcal{L}}$ has one dominating eigenvalue $\lambda > 1$. Otherwise stated, $u_s(n) = P(n)\lambda^n + o(\lambda^n)$ with P a non-zero polynomial.

Taking these two hypotheses into account, a communicating class will be said to be *essential* if its Perron-Frobenius eigenvalue is equal to the dominating eigenvalue λ of $\mathcal{L}$.

A function $f : \Sigma^* \to \mathbb{R}$ is called *completely additive*, if

$$f(w) = \sum_{\ell=1}^{k} f(\sigma_\ell)$$

holds for any word $w = \sigma_1\sigma_2\cdots\sigma_k \in \Sigma^*$. For a fixed language $\mathcal{L}$ we write $f(n)$ for $f(\mathrm{val}^{-1}(n))$. In [19] the summatory function

$$\sum_{n<N} f(n)$$

was studied.

3. Automata and matrices

In this section, our aim is to define a matrix $\tilde{A}$ related to $\mathcal{M}_{\mathcal{L}}$. We will refer to $\tilde{A}$ as the extended adjacency matrix of $\mathcal{L}$. At the end of the section, the reader can find an explicit example of the construction of $\tilde{A}$.

Consider the graph $\mathcal{G}_{\mathcal{L}}$ defined in Section 2. Let us denote by $\mathcal{C}_s$ the communicating class containing the initial state s of $\mathcal{M}_{\mathcal{L}}$. We say that a communicating class $\mathcal{C}$ is *final* if a final state of $\mathcal{M}_{\mathcal{L}}$ belongs to $\mathcal{C}$. Due to the minimality of the trimmed automaton $\mathcal{M}_{\mathcal{L}}$, it is clear that each sink class is final.

Since, $\mathcal{G}_{\mathcal{L}}$ does not contain any oriented cycle, the number N_p of paths in $\mathcal{G}_{\mathcal{L}}$ starting in $\mathcal{C}_s$ and ending in a sink class is finite. Let $\mathfrak{p}_j$ be the jth path of this kind, $1 \le j \le N_p$,

$$\mathfrak{p}_j : \mathcal{C}_{j,1} = \mathcal{C}_s \to \mathcal{C}_{j,2} \to \cdots \to \mathcal{C}_{j,\ell_j}.$$

Let $S_j \subseteq \{1, \ldots, \ell_j\}$ be the set of indices defined by

$$i \in S_j \Leftrightarrow \mathcal{C}_{j,i} \text{ is final.}$$

In particular, ℓ_j belongs to S_j. For each $i \in S_j$, we consider the sub-path

$$\mathfrak{p}_{j,i} : \mathcal{C}_{j,1} \to \cdots \to \mathcal{C}_{j,i}$$

of length $i-1$ of the path $\mathfrak{p}_j$. In particular, $\mathfrak{p}_j = \mathfrak{p}_{j,\ell_j}$.

Remark 2. Considering this construction for all the possible values of j, at the end of the process, it is possible to obtain multiple copies of the same path, i.e., namely one could possibly find $j \neq j'$ such that $\mathfrak{p}_{j,i} = \mathfrak{p}_{j',i}$. Indeed, assume for instance that we have an automaton $\mathcal{M}_{\mathcal{L}}$ leading to four communicating classes $\mathcal{C}_i$, $i = 1, \ldots, 4$, all being final with

$\mathcal{C}_1 = \mathcal{C}_s$ and such that $\mathcal{C}_1 \to \mathcal{C}_2$, $\mathcal{C}_2 \to \mathcal{C}_3$ and $\mathcal{C}_2 \to \mathcal{C}_4$. The sink classes are $\mathcal{C}_3$ and $\mathcal{C}_4$. With our notation $N_p = 2$, $S_1 = S_2 = \{1, 2, 3\}$ and we obtain the paths

$$\mathfrak{p}_{1,1} : \mathcal{C}_1, \quad \mathfrak{p}_{1,2} : \mathcal{C}_1 \to \mathcal{C}_2, \quad \mathfrak{p}_{1,3} : \mathcal{C}_1 \to \mathcal{C}_2 \to \mathcal{C}_3$$

and

$$\mathfrak{p}_{2,1} : \mathcal{C}_1, \quad \mathfrak{p}_{2,2} : \mathcal{C}_1 \to \mathcal{C}_2, \quad \mathfrak{p}_{2,3} : \mathcal{C}_1 \to \mathcal{C}_2 \to \mathcal{C}_4.$$

Here, we have $\mathfrak{p}_{1,1} = \mathfrak{p}_{2,1}$ and $\mathfrak{p}_{1,2} = \mathfrak{p}_{2,2}$. For counting reasons, we do not want to repeat the same path more than once and we will therefore modify the sets S_j's accordingly. In this particular example, one can take $S_1 = \{1, 2, 3\}$ and $S_2 = \{3\}$. In what follows, we will always make this assumption and we still write S_j for the new restricted sets of indices.

In the statement of our theorems, we will attach a particular importance to paths of a special form: a path $\mathfrak{p}_{j,i}$ is said to be *essential* if it contains a maximal number of essential classes.

For $j \in \{1, \dots, N_p\}$ and $i \in S_j$, let $E_{j,i}$ be the adjacency matrix of the sub-automaton of $\mathcal{M}_{\mathcal{L}}$ restricted to the states belonging to the classes $\mathcal{C}_{j,1}, \dots, \mathcal{C}_{j,i}$. Proceeding as in Section 2, we order the communicating classes and the corresponding states with respect to the order induced by the path $\mathfrak{p}_{j,i}$. Therefore the matrix $E_{j,i}$ has again a block triangular form.

Definition 1. The *extended adjacency matrix* of $\mathcal{M}_{\mathcal{L}}$ is a block diagonal matrix $\tilde{A}$ whose diagonal blocks are exactly the $E_{j,i}$'s for all $j \in \{1, \dots, N_p\}$ and $i \in S_j$.

Remark 3. The entries of $\tilde{A}$ correspond to states of $\mathcal{M}_{\mathcal{L}}$ and it is clear that a single state can be associated to more than one element of $\tilde{A}$. For instance, the initial state s appears in every path and therefore at least one entry of every matrix $E_{j,i}$ corresponds to s. Conversely, any state q corresponds to at most one entry in $E_{j,i}$.

Assume that $\tilde{A}$ has the form

$$\tilde{A} = \operatorname{diag}\left(E_{1,i_{1,1}}, \dots, E_{1,i_{1,n_1}}, \dots, E_{t,i_{t,1}}, \dots, E_{t,i_{t,n_t}}\right),$$

where each $E_{k,\ell}$ is a square block triangular matrix. Corresponding to this matrix $\tilde{A}$, we define for each $q \in Q$ an horizontal vector

$$V_{1,q} = \left(v^{(q)}_{1,i_{1,1}}, \dots, v^{(q)}_{1,i_{1,n_1}}, \dots, v^{(q)}_{t,i_{t,1}}, \dots, v^{(q)}_{t,i_{t,n_t}}\right)$$

where $v^{(q)}_{k,\ell}$ has the same dimension as the corresponding $E_{k,\ell}$. This is a null vector if q does not correspond to any entry of $E_{k,\ell}$. Otherwise, $v^{(q)}_{k,\ell}$ contains exactly a one in the position corresponding to q in the matrix $E_{k,\ell}$. We also define a vertical vector V_2 having the same structure

$$V_2^T = \left(r_{1,i_{1,1}}, \dots, r_{1,i_{1,n_1}}, \dots, r_{t,i_{t,1}}, \dots, r_{t,i_{t,n_t}}\right)$$

where $r_{k,\ell}$ is defined in the following way. From the above discussion, $E_{k,\ell}$ comes from a path $\mathfrak{p}_{k,\ell} : \mathcal{C}_{k,1} \to \cdots \to \mathcal{C}_{k,\ell}$. All the components of $r_{k,\ell}$ corresponding to final states appearing in $\mathcal{C}_{k,\ell}$ are set to one. The other entries are set to zero.

Remark 4. For all $q \in Q$ and $n \in \mathbb{N}$, we have

$$V_{1,q}(\tilde{A})^n V_2 = u_q(n). \tag{3.1}$$

Indeed, $v_{k,\ell}^{(q)}(E_{k,\ell})^n (r_{k,\ell})^T$ counts the number of paths of length n starting in q and ending in a final state of $\mathcal{C}_{k,\ell}$. By definition of $V_{1,q}, V_2$ and $\tilde{A}$, each path of length n from q to a final state is counted exactly once.

Remark 5. For any path $\mathfrak{p}_{j,i} : \mathcal{C}_{j,1} \to \mathcal{C}_{j,2} \to \cdots \to \mathcal{C}_{j,i}$ in the graph $\mathcal{G}_{\mathcal{L}}$ connecting communicating classes, the number $W_{j,i}(n)$ of words of length n corresponding to this path can be expressed in terms of the dominating eigenvalues α_k of the components $\mathcal{C}_{j,k}$ $(k = 1, \ldots, i)$. Since the number of words of length n originating from the component $\mathcal{C}_{j,k}$ is $\sim C_k \alpha_k^n$ we have

$$W_{j,i}(n) \sim C_{j,1} \cdots C_{j,i} \sum_{k_1 + \cdots + k_i = n} \alpha_1^{k_1} \cdots \alpha_i^{k_i}.$$

Let $\theta = \max(\alpha_1, \ldots, \alpha_i)$ be the dominating eigenvalue of the path and assume that it occurs for d components. Then

$$W_{j,i}(n) \asymp n^{d-1} \theta^n.$$

From this it follows that Jordan-decomposition of the corresponding matrix $E_{j,i}$ contains a Jordan block of size d for the eigenvalue θ. In particular, if an essential path contains m essential classes then $u_s(n) \asymp n^{m-1} \lambda^n$.

Example 1. Consider the language $\mathcal{L}$ over the alphabet $\{a, b, c\}$ having the automaton depicted in Figure 1 as trimmed minimal automaton (the initial state is indicated by an unlabelled arrow and the final states are represented with double circles). We have five

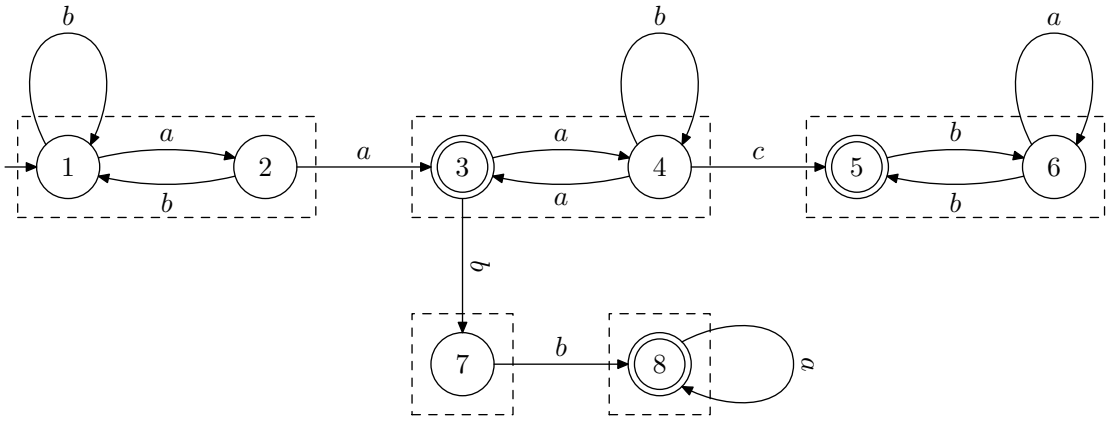

FIGURE 1. A trimmed minimal automaton.

communicating classes partitioning the set of states: $\mathcal{C}_1 = \{1, 2\}$, $\mathcal{C}_2 = \{3, 4\}$, $\mathcal{C}_3 = \{5, 6\}$, $\mathcal{C}_4 = \{7\}$ and $\mathcal{C}_5 = \{8\}$ where $\mathcal{C}_3, \mathcal{C}_5$ are the sink classes and $\mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_5$ are the final ones.

We have to consider exactly three paths, $\mathfrak{p}_{1,2} : \mathcal{C}_1 \to \mathcal{C}_2$, $\mathfrak{p}_{1,3} : \mathcal{C}_1 \to \mathcal{C}_2 \to \mathcal{C}_3$ and $\mathfrak{p}_{2,4} : \mathcal{C}_1 \to \mathcal{C}_2 \to \mathcal{C}_4 \to \mathcal{C}_5$. The matrices corresponding to these paths are

$$E_{1,2} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}, E_{1,3} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 1 \end{pmatrix}, E_{2,4} = \begin{pmatrix} 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

These are the adjacency matrices (with block triangular form) of the sub-automata having respectively $\{1,2,3,4\}$, $\{1,\dots,6\}$ and $\{1,2,3,4,7,8\}$ as set of states. The extended adjacency matrix of $\mathcal{L}$ is the block diagonal matrix $\tilde{A}$ having $E_{1,2}$, $E_{1,3}$ and $E_{2,4}$ as diagonal blocks. The corresponding vectors $V_{1,q}$'s are given by

$$\begin{array}{l} V_{1,1} = \left(\begin{array}{cccc|cccccc|cccccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{array} \right) \\ V_{1,2} = \left(\begin{array}{cccc|cccccc|cccccc} 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{array} \right) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \vdots \\ V_{1,8} = \left(\begin{array}{cccc|cccccc|cccccc} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right) \end{array}$$

and the vector V_2 by

$$V_2^T = \left(\begin{array}{cccc|cccccc|cccccc} 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{array} \right).$$

Notice that the 7th component of V_2 corresponding to the state 3 is set to zero because for the ith components ($i = 5,\dots,10$) we have to consider the final states appearing only in $\mathcal{C}_3$. (These components of V_2 correspond to the path $\mathfrak{p}_{1,3}$.)

Let $\lambda = \frac{1+\sqrt{5}}{2}$ be the dominating eigenvalue of $\mathcal{L}$. The three components $\mathcal{C}_i$ ($i = 1,2,3$) are essential, i.e., they all have λ as dominating eigenvalue. Consequently, the path $\mathfrak{p}_{1,3}$ is essential. Using the same reasoning as in Remark 5, the number of words of length n starting in 1 and ending in 3 (resp. in 5, 8) is of order $n\lambda^n$ (resp. $n^2\lambda^n$, $n\lambda^n$). In the Jordan-decomposition of $E_{1,2}$ (resp. $E_{1,3}$, $E_{2,4}$) appears exactly one Jordan block of size 2 (resp. 3, 2) for the eigenvalue λ.

4. Multiplicative functions

In the course of our discussion of the distribution behaviour of additive functions we will make use of multiplicative functions. A function $g : \mathcal{L} \to \mathbb{C}$ is called *multiplicative*, if

$$g(w_1 w_2 \dots w_k) = g(w_1) \cdots g(w_k)$$

holds for any word $w_1 w_2 \dots w_k \in \mathcal{L}$.

Lemma 2. *Let $g : \mathcal{L} \to \mathbb{C}$ be a multiplicative function on the language $\mathcal{L}$. Then the following formula holds for $W = W_1 W_2 \dots W_n$*

$$\sum_{\substack{w < W \\ w \in \mathcal{L}}} g(w) = \sum_{q \in Q} \sum_{k=1}^{|W|} \gamma_{q,k}(W) G_q(|W| - k), \tag{4.1}$$

where

$$G_q(\ell) := \sum_{\substack{w\in\mathcal{L}_q\\ |w|=\ell}} g(w)$$

and

$$\gamma_{q,k}(W) := g(W_1)\cdots g(W_{k-1}) \sum_{\substack{\sigma<W_k\\ s.W_1\ldots W_{k-1}\sigma=q}} g(\sigma) + \delta_{q,s}.$$

Proof. The proof is similar to the proof of (2.1) given in [22]. For the sake of completeness, we recall the details

$$\begin{aligned}
\sum_{\substack{w<W\\ w\in\mathcal{L}}} g(w) &= \sum_{\substack{|w|<|W|\\ w\in\mathcal{L}}} g(w) + \sum_{\substack{|w|=|W|\\ w<W, w\in\mathcal{L}}} g(w)\\
&= \sum_{k=0}^{|W|-1} G_s(k) + \sum_{k=1}^{|W|}\sum_{\sigma<W_k}\sum_{\substack{|w|=|W|-k\\ w\in\mathcal{L}_{s.W_1\ldots W_{k-1}\sigma}}} g(W_1\ldots W_{k-1}\sigma w)\\
&= \sum_{k=0}^{|W|-1} G_s(k) + \sum_{k=1}^{|W|} g(W_1)\cdots g(W_{k-1}) \sum_{\sigma<W_k} g(\sigma) \underbrace{\sum_{\substack{|w|=|W|-k\\ w\in\mathcal{L}_{s.W_1\ldots W_{k-1}\sigma}}} g(w)}_{=G_{s.W_1\ldots W_{k-1}\sigma}(|W|-k)}
\end{aligned}$$

where we have used the multiplicativity of g in the last line. The conclusion follows from the definition of $\gamma_{q,k}$. □

The function $G_q(\ell)$ can be given in terms of a matrix product similar to the formula (3.1) given for $u_q(\ell)$ in Section 3: let $B_{p,q} = \sum_{p.\sigma=q} g(\sigma)$, then

$$G_q(n) = (B^n V)_q$$

where V is vertical vector such that $V_q = 1$ if and only if $q \in F$.

Clearly, for an additive function f, $g(w,t) = \exp(itf(w))$ is a multiplicative function.

We apply the same construction as in Section 3 to the matrix B associated to the multiplicative function $\exp(itf(w))$ to obtain the matrix $\tilde{B}(t)$. For $t = 0$ the occurring matrix is the extended adjacency matrix $\tilde{A}$ from Section 3. Since all the diagonal blocks of $\tilde{A}$ have a dominating eigenvalue and all the coefficients of B depend on t holomorphically, there corresponds a function $\lambda_k(t)$ to every component $\mathcal{C}_k$ of the automaton such that $\lambda_k(0) = \lambda_k$ (the Perron-Frobenius eigenvalue of the component), cf. [3]. Furthermore, we have $|\lambda_k(t)| \le \lambda_k(0)$ for $t \in \mathbb{R}$ with equality for $t = 0$.

Let TJT^{-1} be the Jordan-decomposition of $\tilde{A}$. Then there exist matrices $T(t)$ and $J(t)$ in some open interval I around 0 such that $T(0) = T$, $J(0) = J$, and $\tilde{B}(t) = T(t)J(t)T(t)^{-1}$.

Furthermore, the matrix function $J(t)$ has the same block structure as J, i.e., the block

$$\begin{pmatrix} \lambda & 1 & 0 & \dots & 0 \\ 0 & \lambda & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \dots & & \lambda \end{pmatrix}$$

of J corresponds to a block

$$J_j(t) = \begin{pmatrix} \lambda_{k_{1,j}}(t) & 1 & 0 & \dots & 0 \\ 0 & \lambda_{k_{2,j}}(t) & 1 & \dots & 0 \\ \vdots & \vdots & \ddots & & \vdots \\ 0 & 0 & \dots & & \lambda_{k_{\ell_j,j}}(t) \end{pmatrix}$$

of $J(t)$. Then

$$G_q(n,t) = V_{1,q}\tilde{B}(t)^n V_2 \tag{4.2}$$

and it remains to study the powers of $J_j(t)$.

There are three different cases for the behaviour of the powers of $J_j(t)$ depending on the power series expansions of the corresponding eigenvalues around $t = 0$. Let $J_j(t)$ have the diagonal entries $\mu_1(t), \dots, \mu_\ell(t)$ and define $a_1, \dots, a_\ell$ and $b_1, \dots, b_\ell$ by

$$\mu_m(t) = \lambda \exp\left(i a_m t - b_m \frac{t^2}{2} + \mathcal{O}(t^3) \right).$$

First notice that an easy induction shows that for any r, s such that $1 \le r < s \le \ell$

$$(J_j(t)^n)_{r,s} = \sum_{\substack{\alpha_r + \dots + \alpha_s = n-s+r \\ \alpha_m \ge 0}} \mu_r(t)^{\alpha_r} \cdots \mu_s(t)^{\alpha_s}$$

case 1: Not all a_m's are equal: in this case we consider the matrices $J_j(t/n)^n$. By definition we have

$$(J_j(t/n)^n)_{1,\ell} = \lambda^{n-\ell+1} n^{\ell-1} \sum_{\substack{\alpha_1 + \dots + \alpha_\ell = n-\ell+1 \\ \alpha_m \ge 0}} \exp\left(it \sum_{m=1}^{\ell} \frac{\alpha_m}{n} a_m + \mathcal{O}\left(\frac{t^2}{n}\right) \right) n^{-(\ell-1)}.$$

The sum can be interpreted as a Riemann sum for the integral

$$P_1(a_1, \dots, a_\ell, t) = \int_{\substack{x_1 + \dots + x_{\ell-1} \le 1 \\ x_m \ge 0}} \cdots \int \exp\left(it \left(\sum_{m=1}^{\ell-1} a_m x_m + a_\ell (1 - x_1 - \cdots - x_{\ell-1}) \right) \right) dx_1 \cdots dx_{\ell-1}.$$

Since the integrand is differentiable with respect to all its variables, the difference between the Riemann sum and the integral is $\mathcal{O}(1/n)$. Thus we have

$$(J_j(t/n)^n)_{1,\ell} = \lambda^{n-\ell+1} n^{\ell-1} \left(P_1(a_1, \dots, a_\ell, t) + \mathcal{O}\left(\frac{t^2}{n}\right) \right). \tag{4.3}$$

From this and (4.2) we get $G_q(n,t/n) = u_q(n)C_q(t) + \mathcal{O}(t^2 u_q(n) n^{-1})$, where C_q is a linear combination of functions P_1.

case 2: $a_1 = \cdots = a_\ell = a$, but not all b_m's are equal: in this case we consider $J_j(t/\sqrt{n})^n$ and proceed as in the first case to obtain

$$\left(J_j(t/\sqrt{n})^n\right)_{1,\ell} = \lambda^{n-\ell+1} n^{\ell-1} e^{iat\sqrt{n}} \left(P_2(b_1,\ldots,b_\ell,t) + \mathcal{O}\left(\frac{t^3}{\sqrt{n}}\right)\right), \tag{4.4}$$

where

$$P_2(b_1,\ldots,b_\ell,t) = \int\limits_{\substack{x_1+\cdots+x_{\ell-1}\le 1\\ x_m\ge 0}} \cdots \int \exp\left(-\frac{t^2}{2}\left(\sum_{m=1}^{\ell-1} b_m x_m + b_\ell(1-x_1-\cdots-x_{\ell-1})\right)\right) dx_1\cdots dx_{\ell-1}.$$

From this we get $G_q(n,t/n) = u_q(n)e^{iat} + \mathcal{O}(t^2 u_q(n) n^{-1})$ and $G_q(n,t/\sqrt{n}) = u_q(n) e^{iat\sqrt{n}} D_q(t) + \mathcal{O}(t^3 u_q(n) n^{-\frac{1}{2}})$, where D_q is a linear combination of functions P_2.

case 3: $a_1 = \cdots = a_\ell = a$ and $b_1 = \cdots = b_\ell = b$: again we consider $J_j(t/\sqrt{n})^n$ and obtain

$$\left(J_j(t/\sqrt{n})^n\right)_{1,\ell} = \lambda^{n-\ell+1} n^{\ell-1} e^{iat\sqrt{n}} e^{-\frac{b}{2}t^2} \left(1 + \mathcal{O}\left(\frac{t^3}{\sqrt{n}}\right)\right). \tag{4.5}$$

From this we get $G_q(n,t/\sqrt{n}) = e^{iat\sqrt{n}} e^{-\frac{b}{2}t^2} u_q(n) + \mathcal{O}(t^3 u_q(n) n^{-\frac{1}{2}})$.

5. Distribution of additive functions

Theorem 1. *Let $\mathcal{L}$ be a regular language given by its trimmed minimal automaton $\mathcal{M}_\mathcal{L}$. Assume further that all its communicating classes are primitive and that the adjacency matrix of $\mathcal{M}_\mathcal{L}$ has a unique dominant eigenvalue λ. Let $f : \mathcal{L} \to \mathbb{R}$ be an additive function and define $\lambda_k(t)$ as the continuous function giving the eigenvalue of the marked adjacency matrix $\tilde{B}$ introduced in Section 4 and corresponding to the blocks arising from essential paths. Define real numbers $a_1,\ldots,a_\ell$ by $\lambda_m(t) = \lambda \exp(ia_m t + \mathcal{O}(t^2))$. If not all a_m's are equal then*

$$\lim_{n\to\infty} \frac{1}{u_s(n)} \#\{w \in \mathcal{L} \mid |w| = n, f(w) < nx\} = F(x)$$

exists for all $x \in \mathbb{R}$ with the possible exception of finitely many points.

Remark 6. Theorem 1 explains why in general the asymptotic main term of $\sum_{w<W, w\in\mathcal{L}} f(w)$ involves a fluctuating function (cf. [19, Theorem 1]).

Theorem 2. *Under the hypotheses of Theorem 1 define real numbers $b_1,\ldots,b_\ell$ by $\lambda_m(t) = \lambda \exp(ia_m t - b_m \frac{t^2}{2} + \mathcal{O}(t^3))$ and assume that $a_1 = \cdots = a_\ell = a$ but not all b_m's are equal. Then*

$$\lim_{\mathrm{val}(W)\to\infty} \frac{1}{|W|\,\mathrm{val}(W)} \sum_{\substack{w<W\\ w\in\mathcal{L}}} f(w)$$

exists and equals a. Furthermore,

$$\lim_{n\to\infty}\frac{1}{u_s(n)}\#\left\{w\in\mathcal{L}\mid |w|=n, f(w)-na<x\sqrt{n}\right\}=G(x)$$

exists for all $x\in\mathbb{R}$. The function $G(x)$ is infinitely differentiable.

Remark 7. Theorem 2 gives a more transparent and less technical condition for the existence of an asymptotic formula for $\sum_{w<W} f(w)$ than in [19, Corollary 3].

Theorem 3. *Under the hypotheses of Theorem 2 assume that $a_1=\cdots=a_\ell=a$ and $b_1=\cdots=b_\ell=b$. Then*

$$\lim_{\mathrm{val}(W)\to\infty}\frac{1}{\mathrm{val}(W)}\#\left\{w\in\mathcal{L}\mid w<W, f(w)-na<x\sqrt{b|W|}\right\}=\Phi(x),$$

where $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}e^{-t^2/2}\,dt$ is the normal distribution function.

Proof of Theorem 1. The hypothesis of Theorem 1 are exactly covered by the situation of **case 1** in Section 4. Thus we have

$$\lim_{n\to\infty}\frac{1}{u_s(n)}\sum_{\substack{w\in\mathcal{L}\\|w|=n}}\exp\left(it\frac{f(w)}{n}\right)=C_s(t),$$

where $C_s(t)$ is a linear combination of functions of the form $P_1(a_1,\dots,a_\ell,t)$. Since P_1 is continuous at $t=0$ Levy's continuity theorem (cf. [24]) implies the assertion. □

Proof of Theorem 2. The hypothesis of Theorem 2 are exactly covered by the situation of **case 2** in Section 4. From (4.1) applied to the function $g(w)=\exp(itf(w)/|W|)$ and the asymptotic expression for $G_q(k)$ derived after (4.4) we get

$$\begin{aligned}\sum_{\substack{w<W\\w\in\mathcal{L}}}e^{it\frac{f(w)}{|W|}}&=\sum_{k=1}^{|W|}\sum_{q\in Q}\gamma_{q,k}\left(W,\frac{t}{|W|}\right)G_q\left(|W|-k,\frac{t}{|W|}\right)\\&=e^{ita}\,\mathrm{val}(W)+\mathcal{O}(t\,\mathrm{val}(W)/|W|)+\mathcal{O}(t^2\,\mathrm{val}(W)),\end{aligned}$$

where we have used that $\gamma_{q,k}(W,\frac{t}{|W|})=\beta_{q,k}(W)+\mathcal{O}(t/|W|)$ and the expansion for $G_q(k)$. Differentiating with respect to t and setting $t=0$ gives the first assertion.

The proof of the second assertion runs along the same lines as the proof of Theorem 1. The differentiability of the function $G(x)$ follows from the fact that $P_2(b_1,\dots,b_\ell,t)$ decays like $\exp(-Ct^2)$ for $|t|\to\infty$. □

Proof of Theorem 3. The hypothesis of Theorem 3 are exactly covered by the situation of **case 3** in Section 4. An argument similar to the proof of the first assertion of Theorem 2 yields

$$\lim_{\mathrm{val}(W)\to\infty}\frac{1}{\mathrm{val}(W)}\sum_{\substack{w<W\\w\in\mathcal{L}}}\exp\left(it\frac{f(w)-a|W|}{\sqrt{b|W|}}\right)=e^{-\frac{t^2}{2}},$$

which again by Levy's continuity theorem (cf. [24]) implies the assertion. □

6. Examples

In this section we exhibit a number of examples which show that our theorems cannot be improved. We will always assume that the occurring letters are ordered alphabetically.

Example 2. Let $\mathcal{L} = \{a,b\}^* \cup \{c,d\}^*$ and $f(a) = f(c) = 1$, $f(b) = -1$, $f(d) = 0$. Then the corresponding matrix B is of the form

$$B = \begin{pmatrix} 0 & 2\cos t & 1+e^{it} \\ 0 & 2\cos t & 0 \\ 0 & 0 & 1+e^{it} \end{pmatrix}$$

so, $\lambda_1(t) = 2\exp(-\frac{t^2}{2} + \mathcal{O}(t^3))$ and $\lambda_2(t) = 2\exp(i\frac{t}{2} - \frac{t^2}{8} + \mathcal{O}(t^3))$. Notice that for $n > 0$, $u_s(n) = 2^{n+1}$ and $\#\{w \in \mathcal{L} \mid w < c^n\} = 1 + \sum_{k=1}^{n-1} u_s(k) + 2^n = 3 \cdot 2^n - 1$. We have

$$\lim_{n\to\infty} \frac{1}{2^{n+1}} \#\left\{w \in \mathcal{L} \mid |w| = n, f(w) < nx\right\} = \begin{cases} 0 & \text{for } x < 0 \\ \frac{1}{2} & \text{for } 0 < x < \frac{1}{2} \\ 1 & \text{for } x > \frac{1}{2}. \end{cases}$$

Furthermore,

$$\lim_{n\to\infty} \frac{1}{3\cdot 2^n} \#\{w < c^n \mid f(w) < nx\} = \begin{cases} 0 & \text{for } x < 0 \\ \frac{2}{3} & \text{for } 0 < x < \frac{1}{2} \\ 1 & \text{for } x > \frac{1}{2}, \end{cases}$$

which shows that

$$\frac{1}{\operatorname{val}(W)} \#\{w \in \mathcal{L} \mid w < W, f(w) < |W|x\}$$

does not converge for $\operatorname{val}(W) \to \infty$.

Example 3. Consider again $\mathcal{L} = \{a,b\}^* \cup \{c,d\}^*$ but $f(a) = -f(b) = 1$, $f(c) = -f(d) = 2$. Then the corresponding matrix B is of the form

$$B = \begin{pmatrix} 0 & 2\cos t & 2\cos 2t \\ 0 & 2\cos t & 0 \\ 0 & 0 & 2\cos 2t \end{pmatrix}$$

so, $\lambda_1(t) = 2\exp(-\frac{t^2}{2} + \mathcal{O}(t^3))$ and $\lambda_2(t) = 2\exp(-2t^2 + \mathcal{O}(t^3))$. We have

$$\lim_{\operatorname{val}(W)\to\infty} \frac{1}{|W|\operatorname{val}(W)} \sum_{\substack{w<W \\ w\in\mathcal{L}}} f(w) = 0$$

and

$$\lim_{n\to\infty} \frac{1}{2^{n+1}} \#\left\{w \in \mathcal{L} \mid |w| = n, f(w) < x\sqrt{n}\right\} = \frac{1}{2}\left(\Phi(x) + \Phi(x/2)\right).$$

Furthermore,

$$\lim_{n\to\infty} \frac{1}{n2^n} \#\left\{w \in \mathcal{L} \mid w < c^n, f(w) < x\sqrt{n}\right\} = \frac{1}{3}\left(2\Phi(x) + \Phi(x/2)\right),$$

which shows that

$$\frac{1}{\mathrm{val}(W)}\#\{w \in \mathcal{L} \mid w < W, f(w) - a|W| < x\sqrt{|W|}\}$$

does not converge for $\mathrm{val}(W) \to \infty$.

References

1. G. Barat, T. Downarowicz, A. Iwanik, and P. Liardet, *Propriétés topologiques et combinatoires des échelles de numération*, Colloq. Math. **84/85** (2000), 285–306.
2. G. Barat and P. J. Grabner, *Distribution properties of G-additive functions*, J. Number Theory **60** (1996), 103–123.
3. H. Baumgärtel, *Endlichdimensionale analytische Störungstheorie*, Akademie-Verlag, Berlin, 1972, Mathematische Lehrbücher und Monographien, II. Abteilung. Mathematische Monographien, Band 28.
4. V. Berthé and M. Rigo, *Odometers on regular languages*, preprint, available at `http://www.discmath.ulg.ac.be/papers/vbmr.ps.gz`, 2004.
5. A. Bertoni, C. Choffrut, M. Goldwurm, and V. Lonati, *On the number of occurrences of a symbol in words of regular languages*, Theoret. Comput. Sci. **302** (2003), 431–456.
6. V. Bruyère and G. Hansel, *Bertrand numeration systems and recognizability*, Theor. Comput. Sci. **181** (1997), 17–43, Latin American Theoretical Informatics (Valparaíso, 1995).
7. H. Delange, *Sur les fonctions q-additive ou q-multiplicatives*, Acta Arith. **21** (1972), 285–298.
8. ———, *Sur la fonction sommatoire de la fonction "Somme des Chiffres"*, Enseign. Math. II. Ser. **21** (1975), 31–47.
9. M. Drmota, *q-Additive functions and well distribution*, Demonstr. Math. **30** (1997), 883–896.
10. M. Drmota and J. Gajdosik, *The distribution of the sum-of-digits function*, J. Théor. Nombres Bordx. **10** (1998), 17–32.
11. J.-M. Dumont and A. Thomas, *Digital sum problems and substitutions on a finite alphabet*, J. Number Theory **39** (1991), 351–366.
12. ———, *Digital sum moments and substitutions*, Acta Arith. **64** (1993), 205–225.
13. ———, *Gaussian asymptotic properties of the sum-of-digits function*, J. Number Theory **62** (1997), 19–38.
14. S. Eilenberg, *Automata, Languages and Machines*, vol. A, Academic Press, 1974.
15. P. Flajolet, P. J. Grabner, P. Kirschenhofer, H. Prodinger, and R. F. Tichy, *Mellin transforms and asymptotics: digital sums*, Theor. Comput. Sci. **123** (1994), 291–314.
16. C. Frougny, *Representation of numbers and finite automata*, Math. Syst. Theory **25** (1992), 37–60.
17. ———, *Numeration Systems*, ch. 7, in Lothaire [25], 2002.
18. P. J. Grabner, P. Liardet, and R. F. Tichy, *Odometers and systems of numeration*, Acta Arith. **70** (1995), 103–123.
19. P. J. Grabner and M. Rigo, *Additive functions with respect to numeration systems on regular languages*, Monatsh. Math. **139** (2003), 205–219.
20. P. J. Grabner and R. F. Tichy, *Contributions to digit expansions with respect to linear recurrences*, J. Number Theory **36** (1990), 160–169.
21. ———, *α-expansions, linear recurrences and the sum-of-digits function*, Manuscr. Math. **70** (1991), 311–324.
22. P. Lecomte and M. Rigo, *Numeration systems on a regular language*, Theory Comput. Syst. **34** (2001), 27–44.
23. D. Lind and B. Marcus, *An introduction to symbolic dynamics and coding*, Cambridge University Press, Cambridge, 1995.

24. M. Loève, *Probability theory*, Third edition, D. Van Nostrand Co., Inc., Princeton, N.J.-Toronto, Ont.-London, 1963.
25. M. Lothaire, *Algebraic combinatorics on words*, Encyclopedia of Mathematics and its Applications, no. 90, Cambridge University Press, Cambridge, 2002.
26. E. Seneta, *Non-Negative Matrices, An introduction to theory and applications*, Halsted Press [A division of John Wiley & Sons], New York, 1973.

(P. G.)
Institut für Mathematik A
Technische Universität Graz
Steyrergasse 30
8010 Graz
Austria
E-mail address: peter.grabner@tugraz.at

(M. R.)
Institut de Mathématique
Université de Liège
Grande Traverse 12 (B 37)
4000 Liège
Belgium
E-mail address: M.Rigo@ulg.ac.be