ISIT'2013 1569718041

# Classification of Markov Sources Through Joint String Complexity: Theory and Experiments

Philippe Jacquet and Dimitris Milioris
Bell Labs, Alcatel-Lucent
Centre de Villarceaux
91620, Nozay, France
Email: {philippe.jacquet, dimitrios.milioris}@alcatel-lucent.com

Wojciech Szpankowski
Department of Computer Science
Purdue University
West Lafayette, IN 47907-2066 U.S.A.
Email: spa@cs.purdue.edu

***Abstract*—We propose a classification test to discriminate Markov sources based on the joint string complexity. String complexity is defined as the cardinality of a set of all distinct words (factors) of a given string. For two strings, we define the *joint string complexity* as the cardinality of the set of words which both strings have in common. In this paper we analyze the average joint complexity when both strings are generated by two Markov sources. We provide fast converging asymptotic expansions and present some experimental results showing usefulness of the joint complexity to text discrimination.**

## I. Introduction

In the last decades, several attempts have been made to capture mathematically the concept of "complexity" of a sequence, *i.e.* the number of different factors contained in a sequence. In other words, if $X$ is a sequence and $I(X)$ its set of factors (distinct subwords), then the cardinality $|I(X)|$ is the complexity of the sequence. For example, if $X = aabaa$ then $I(X) = \{\nu, a, b, aa, ab, ba, aab, aba, baa, aaba, abaa, aabaa\}$ and $|I(X)| = 12$ ($\nu$ denotes the empty string). Sometimes the complexity of a string is called the $I$-complexity [4]. The notion is connected with quite deep mathematical properties, including rather elusive concept of randomness in a string (see e.g., [2], [10], [11]).

In general, information contained in a string cannot be measured in absolute and a reference string is required. To this end we introduced in [3] the concept of the *joint* complexity, or $J$-complexity, of two strings. The $J$-complexity is the number of common distinct factors in two sequences. In other words, the $J$-complexity of sequences $X$ and $Y$ is equal to $J(X, Y) = |I(X) \cap I(Y)|$. We denote by $J_{n,m}$ the *average* value of $J(X, Y)$ when $X$ is of length $n$ and $Y$ is of length $m$. In this paper, we study the joint string complexity for Markov sources when $n = m$.

The $J$-complexity is an efficient way of estimating similarity degree of two strings. For example, genome sequences of two dogs will contain more common words than genome sequences of a dog and a cat. Similarly, two texts written in the same language have more words in common than texts written in very different languages. Also, the $J$-complexity is larger when languages are close (*e.g.* French and Italian), and smaller when languages are different (*e.g.* English and Polish). Furthermore, texts in the same language but on different topics (*e.g.* law and cooking) have smaller $J$-complexity than texts on the same topic (*e.g.* medicine).

In this paper we offer a precise analysis of the joint complexity, (see also [7]) together with some experimental results (cf. Figures 1 and 2) confirming usefulness of the joint string complexity for text discrimination.

In [3] is proved that the $J$-complexity of two texts generated by two *different* binary memoryless sources grows as

$$\gamma \frac{n^\kappa}{\sqrt{\alpha \log n}}$$

for some $\kappa < 1$ and $\gamma, \alpha > 0$ depending on the parameters of the sources. When the sources are identical, then the $J$-complexity growth is $O(n)$, hence $\kappa = 1$. When the texts are identical (i.e, $X = Y$), then the $J$-complexity is identical to the $I$-complexity and it grows as $\frac{n^2}{2}$ [8]. Indeed, the presence of a common factor of length $O(n)$ inflates the $J$-complexity to $O(n^2)$.

We should point out that our experiments indicate a very slow convergence of the complexity estimates for memoryless sources. Furthermore, memoryless sources are not appropriate for modeling many sources, e.g., natural languages. In this paper and [7] we extend the $J$-complexity estimates to Markov sources of any order for a finite alphabet. Although Markov models are no more realistic in some applications than memoryless sources, they seem to be fairly good approximation for text generation.

Joint string complexity has a variety of applications, such as the detection of similarity degree of two sequences, for example "copy-paste" in texts or documents. It could also be used in analysis of social networks (e.g. tweets that are limited to 140 characters) and classification. Therefore it could be a pertinent tool for automated monitoring of social networks. However, real time search in blogs, tweets and other social media must balance quality and relevance of the content, which – due to short but frequent posts – is still an unsolved problem.

In this paper, we derive a second order asymptotics for $J$-complexity for Markov sources of the following form

$$\gamma \frac{n^\kappa}{\sqrt{\alpha \log n + \beta}}$$

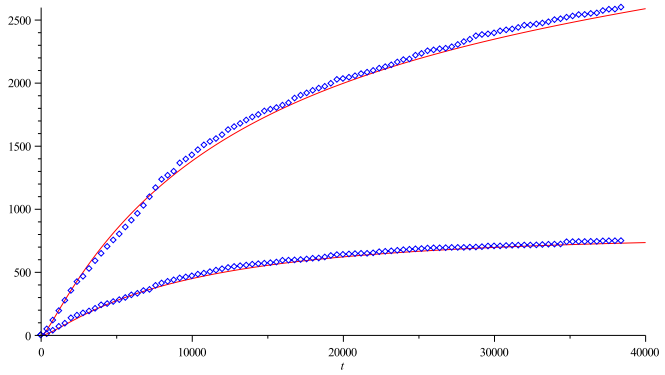


Fig. 1. Joint complexity of actual simulated texts (3rd Markov order) of English, vs French (top), Polish (bottom) languages, versus average theoretical (plain).

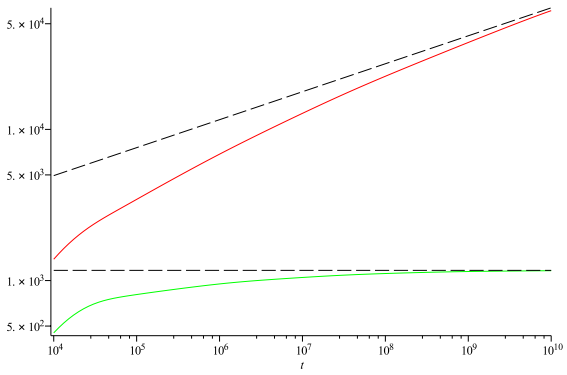


Fig. 2. Average theoretical Joint complexity of 3rd Markov order text of English, vs French (top), and vs Polish (bottom) languages, versus order estimate $O(n^\kappa)$.

for some $\beta > 0$. This new estimate converges faster, although for text length of order $n \approx 10^2$ one needs to compute additional terms. In fact, for some Markov sources our analysis indicates that $J$-complexity oscillates with $n$. This is manifestated by appearing a periodic function $Q(\log n)$ in the leading term of our asymptotics. Surprisingly, this additional term even further improves the convergence for small values of $n$.

In view of these facts, we propose to use the $J$-complexity to discriminate between two identical/non-identical Markov sources [15]. We introduce the following discriminant function

$$d(X,Y) = 1 - \frac{1}{\log n} \log J(X,Y)$$

for two sequences $X$ and $Y$ of length $n$. This discriminant allows us to determine whether $X$ and $Y$ are generated by the same Markov source or not by verifying whether $d(X,Y) = O(1/\log n) \to 0$ or $d(X,Y) = 1 - \kappa + O(\log\log n/\log n) > 0$, respectively. In this conference paper we mainly concentrate on the analysis of $J$-complexity leaving further analysis of the discriminant $d(X,Y)$ to a forthcoming full paper (see also [7]). However, we present some experimental evidence of usage of our discriminant in real texts.

In Figure 1 we compare the joint complexity of a simulated English text to the same length texts simulated in French and in Polish. We also compare it to our theoretical results. In the simulation we use a Markov model of order 3. It is easy to see that even for texts of lengths smaller than a thousand one can discriminate between these languages. In fact, computations show that for English versus French we have $\kappa = 0.18$; versus Greek: $\kappa = 0$. (cf. Theorems 4 and 2, respectively; and versus Polish: $\kappa = 0.01$, The theoretical curve is computed via the iterative resolution of functional equations (7) and (9) (although the curve for Polish vs English should be flat). Figure 2 shows the continuation of our theoretical estimates up to $n = 10^{10}$ and compared with the theoretical estimate $O(n^\kappa)$. Furthermore, computations show that a Markov model of order 3 English text has entropy (per symbol): 0.944; French: 0.934; Polish: 0.665. The joint complexity of such texts grows like $O(n)$ as predicted by theory but due to space limitation it is not displayed here.

Single string complexity was studied extensively in the past. The literature is reviewed in [8] where precise analysis of string complexity is discussed for strings generated by unbiased memoryless sources. Another analysis of the same situation was also proposed in [3] where for the first time the joint string complexity for memoryless sources was presented. It was evident from [3] that precise analysis of the joint complexity is quite challenging due to intricate singularity analysis and infinite number of saddle points. In this paper we deal with the joint string complexity for Markov sources. To the best of our knowledge this problem was never tackled before except in our recent paper [7]. As expected, its analysis is very sophisticated but at the same time quite rewarding. It requires generalized (two-dimensional) dePoissonization and generalized (two-dimensional) Mellin transforms.

## II. Main Results

### A. Models and notations

We begin by introducing some general notation. Let $\omega$ and $\sigma$ be two strings over alphabet $\mathcal{A}$. We denote by $|\omega|_\sigma$ the number of times $\sigma$ occurs in $\omega$ (e.g., $|abbba|_{bb} = 2$). By convention $|\omega|_\nu = |\omega| + 1$, where $\nu$ is the empty string.

Throughout we denote by $X$ a string (text) whose complexity we plan to study. We also assume that its length $|X|$ is equal to $n$. Then the string complexity is $I(X) = |\{\omega : \ |X|_\omega \geq 1\}|$. Observe that

$$|I(X)| = \sum_{\sigma \in \mathcal{A}^*} 1_{|X|_\sigma \geq 1},$$

where $1_A$ is the indicator function of a Boolean $A$. Notice that $|I(X)|$ is equal to the number of nodes in the associated suffix tree of $X$ [8], [14] (see also [5]).

Now, let $X$ and $Y$ be two sequences (not necessarily of the same length). We define the *joint complexity* as the cardinality of the set $J(X,Y) = I(X) \cap I(Y)$. We have

$$|J(X,Y)| = \sum_{\sigma \in \mathcal{A}^*} 1_{|X|_\sigma \geq 1} \times 1_{|Y|_\sigma \geq 1} \ .$$

We now assume that both strings $X$ and $Y$ are generated by two *independent Markov sources* of order $r$ (we will only deal here with Markov of order 1, but extension to arbitrary order is straightforward). We assume that source $i$, for $i \in \{1,2\}$ has the transition probabilities $P_i(a|b)$ from state $b$ to state $a$, where $(a,b) \in \mathcal{A}^r$. We denote by $\mathbf{P}_1$ (resp. $\mathbf{P}_2$) the transition matrix of Markov source 1 (resp. source 2). The stationary distributions are respectively denoted by $\boldsymbol{\pi}_1(a)$ and $\boldsymbol{\pi}_2(a)$ for $a \in \mathcal{A}^r$.

Let $X_n$ and $Y_m$ be two strings of respective lengths $n$ and $m$, generated by Markov source 1 and Markov source 2, respectively. We write $J_{n,m} = \mathbf{E}(|J(X_n, Y_m)|) - 1$ for the joint complexity, *i.e.* omitting the empty string.

### B. Summary of Main Results

We say that a matrix $\mathbf{M} = [m_{ab}]_{(a,b)\in\mathcal{A}^2}$ is *rationally balanced* if $\forall (a,b,c) \in \mathcal{A}^3$: $m_{ab} + m_{ca} - m_{cb} \in \mathbb{Z}$, where $\mathbb{Z}$ is the set of integers. We say that a positive matrix $\mathbf{M} = [m_{ab}]$ is *logarithmically rationally balanced* when the matrix $\log^*(\mathbf{M}) = [l_{ab}]$ is rationally balanced, where $\ell_{ab} = \log(m_{ab})$ when $m_{ab} > 0$ and $l_{ab} = 0$ otherwise. Furthermore, we say that two matrices $\mathbf{M} = [m_{ab}]_{(a,b)\in\mathcal{A}^2}$ and $\mathbf{M}' = [m'_{ab}]$ are *logarithmically commensurable* when matrices $\log^*(\mathbf{M})$ and $\log^*(\mathbf{M}')$ are commensurable, that is, there exist a nonzero pair of reals $(x,y)$ such that $x\log^*(\mathbf{M}) + y\log^*(\mathbf{M}')$ is logarithmically rationally balanced.

We now present our main theoretical results in a series of theorems each treating different cases of Markov sources.

*Theorem 1:* Consider the average joint complexity of two texts of length $n$ generated by the same general stationary Markov source, that is, $\mathbf{P} := \mathbf{P}_1 = \mathbf{P}_2$.
(i) [*Noncommensurable Case.*] Assume that $\mathbf{P}$ is not logarithmically rationally balanced. Then

$$J_{n,n} = \frac{2\log 2}{h} n + o(1) \tag{1}$$

where $h$ is the entropy rate of the source.
(ii) [*Commensurable Case.*] Assume that $\mathbf{P}$ is logarithmically rationally balanced. Then there is $\epsilon > 0$ such that:

$$J_{n,n} = \frac{2\log 2}{h}(1 + Q_0(\log n)) + O(n^{-\epsilon})$$

where $Q_0(.)$ is a periodic function of small amplitude.

Now we consider sources that are not the same and have respective transition matrices $\mathbf{P}_1$ and $\mathbf{P}_2$. The transition matrices are on $\mathcal{A}^r \times \mathcal{A}^r$. If $(a,b) \in \mathcal{A}^r \times \mathcal{A}^r$, we denote by $P_i(a|b)$ the $(a,b)$-th coefficient of matrix $\mathbf{P}_i$. For a tuple of complex numbers $(s_1, s_2)$ we write $\mathbf{P}(s_1,s_2)$ for a matrix whose $(a,b)$-th coefficient is $(\mathbf{P}_1(a|b))^{-s_1}(\mathbf{P}_2(a|b))^{-s_2}$.

We first consider the case when matrix $\mathbf{P}(s_1,s_2)$ is nilpotent [9], that is, for some $K$ the matrix $\mathbf{P}^K(s_1,s_2)$ is a null matrix.

*Theorem 2:* If $\mathbf{P}(s_1,s_2)$ is nilpotent, then there exists $\gamma_0$ such that $\lim_{n\to\infty} J_{n,n} = \gamma_0 := \langle \mathbf{1}(\mathbf{I} - \mathbf{P}(0,0))^{-1} | \mathbf{1}\rangle$ where $\mathbf{1}$ is the unit vector and $\langle\cdot|\cdot\rangle$ is the scalar product.

This result is not surprising and rather trivial since the common factors can only occur in a finite window at the beginning of the strings. It turns out that $\gamma_0 = 1168$ for 3rd order Markov model of English versus Polish languages used in our experiments.

Throughout, now we assume that $\mathbf{P}(s_1,s_2)$ is not nilpotent. We denote by $\mathcal{K}$ the set of real tuple $(s_1,s_2)$ such that $\mathbf{P}(s_1,s_2)$ has the main eigenvalue $\lambda(s_1,s_2) = 1$. Let

$$\begin{aligned} \kappa &= \min_{(s_1,s_2)\in\mathcal{K}} \{-s_1 - s_2\} \\ (c_1,c_2) &= \arg\min_{(s_1,s_2)\in\mathcal{K}} \{-s_1 - s_2\}. \end{aligned}$$

Easy algebra proves that $\kappa < 1$.

*Theorem 3:* Assume $\mathbf{P}(s_1,s_2)$ is not nilpotent and either $c_1 > 0$ or $c_2 > 0$.
(i) [*Noncommensurable Case.*] We assume that $\mathbf{P}_2$ is not logarithmically balanced. Let $c_0 < 0$ such that $(c_0, 0) \in \mathcal{K}$. There exist $\gamma_1$ and $\epsilon > 0$ such that

$$J_{n,n} = \gamma_1 n^{-c_0}(1 + O(n^{-\epsilon})) \tag{2}$$

(ii) [*Commensurable Case.*] Let now $\mathbf{P}_2$ be logarithmically rationally balanced. There exists a periodic function $Q_1(.)$ of small amplitude such that

$$J_{n,n} = \gamma_1 n^{-c_0}(1 + Q_1(\log n) + O(n^{-\epsilon}).$$

The case when both $c_1$ and $c_2$ are between $-1$ and $0$ is the most intricate.

*Theorem 4:* Assume that $c_1$ and $c_2$ are between $-1$ and $0$.
(i) [*Noncommensurable Case.*] When $\mathbf{P}_1$ and $\mathbf{P}_2$ are not logarithmically commensurable matrices, then there exist $\alpha_2$, $\beta_2$ and $\gamma_2$ such that

$$J_{n,n} = \frac{\gamma_2 n^\kappa}{\sqrt{\alpha_2 \log n + \beta_2}}(1 + o(1)) \ . \tag{3}$$

(ii) [*Commensurable Case.*] Let $\mathbf{P}_1$ and $\mathbf{P}_2$ be logarithmically commensurable matrices. Then there exists a double periodic function $Q_2(.)$ of small amplitude such that

$$J_{n,n} = \frac{\gamma_2 n^\kappa}{\sqrt{\alpha_2 \log n + \beta_2}}(1 + Q_2(\log n) + o(1)).$$

## III. Theoretical Analysis

In this section we present a sketchy proof of our main results.

### A. Equivalence of Suffixes Tress and Independent Tries

We have the identity:

$$J_{n,m} = \sum_{w\in\mathcal{A}^* - \{\nu\}} P(w \in I(X_n)| \geq 1) \cdot P(w \in I(Y_n) \geq 1) \ . \tag{4}$$

In [13] the generating function of $P(w \in I(X_n)| \geq 1)$ for Markov sources is derived. It involves the *autocorrelation* polynomial of word $w$. However, to make our analysis tractable we notice that $w \in I(X_n)$ is equivalent to the fact that $w$ is a prefix of at least one of the $n$ suffixes of $X_n$. But this is not sufficient to push forward our analysis. We need a second much deeper observation that replaces *dependent suffixes* with *independent strings* to shift analysis from suffix trees to tries, as already observed in [5]. In order to accomplish

it, let's define $I_1(n)$ (resp. $I_2(n)$) to be the set of prefixes of $n$ independent strings generated by source 1 (resp. 2). Define

$$C_{n,m} = \sum_{w \in \mathcal{A}^* - \{\nu\}} P(w \in I_1(n)) P(w \in I_2(n)) \ .$$

The following holds.

*Lemma 1:* For some $\epsilon > 0$

$$J_{n,n} = C_{n,n} \left(1 + O(n^{-\epsilon})\right) + O(1). \tag{5}$$

A proof of this lemma follows from [5], [12], and will be given in the journal version of this paper.

### B. Functional Equations

Let $a \in \mathcal{A}$. Let

$$C_{a,m,n} = \sum_{w \in a\mathcal{A}^*} P(w \in I_1(n)) P(w \in I_2(n))$$

where $w \in a\mathcal{A}^*$ means that $w$ starts with an $a \in \mathcal{A}$. Notice that $C_{a,m,n} = 0$ when $n = 0$ or $m = 0$. Using Markov nature of the string generation, the quantity $C_{a,n,m}$ for $n, m \geq 1$ satisfies the following recurrence for all $a, b \in \mathcal{A}$

$$\begin{aligned} C_{b,n,m} &= 1 + \sum_{a \in \mathcal{A}} \sum_{n_a, m_a} \binom{n}{n_a} \binom{m}{m_a} \\ &\quad \times (P_1(a|b))^{n_a} (1 - P_1(a|b))^{n - n_a} \\ &\quad \times (P_2(a|b))^{m_a} (1 - P_2(a|b))^{m - m_a} C_{a,n_a,m_a} \ , \end{aligned}$$

where $n_a$ (resp. $m_a$) denotes the number of strings among $n$ (resp. $m$) independent strings from source 1 (resp. 2) that have symbol $a$ followed by symbol $b$. The *unconditional* average $C_{n,m}$ satisfies for $n, m \geq 2$

$$\begin{aligned} C_{n,m} &= 1 + \sum_{a \in \mathcal{A}} \sum_{n_a, m_a} \binom{n}{n_a} \binom{m}{m_a} \pi_1^{n_a}(a) (1 - \pi_1(a))^{n - n_a} \\ &\quad \times \pi_2^{m_a}(a) (1 - \pi_2(a))^{m - m_a} C_{a,n,m}. \end{aligned}$$

We introduce the double Poisson transform of $C_{a,n,m}$ as

$$C_a(z_1, z_2) = \sum_{n,m \geq 0} C_{a,n,m} \frac{z_1^n z_2^m}{n! m!} e^{-z_1 - z_2} \tag{6}$$

that translates the above recurrence into the following functional equation:

$$\begin{aligned} C_b(z_1, z_2) &= (1 - e^{-z_1})(1 - e^{-z_2}) \\ &\quad + \sum_{a \in \mathcal{A}} C_a \left(P_1(a|b) z_1, P_2(a|b) z_2\right). \end{aligned} \tag{7}$$

Furthermore, the cumulative double Poisson transform

$$C(z_1, z_2) = \sum_{n,m \geq 0} T_{n,m} \frac{z_1^n z_2^m}{n! m!} e^{-z_1 - z_2} \tag{8}$$

satisfies

$$\begin{aligned} C(z_1, z_2) &= (1 - e^{-z_1})(1 - e^{-z_2}) \\ &\quad + \sum_{a \in \mathcal{A}} C_a(\pi_1(a) z_1, \pi_2(a) z_2) \ . \end{aligned} \tag{9}$$

### C. DePoissonization

Using [6], [7], [14] we prove the following lemma.

*Lemma 2 (DePoissonization):* When $n$ and $m$ tend to infinity:

$$C_{n,m} = C(n, m)(1 + O(n^{-1}) + O(m^{-1})) \ .$$

This equivalence is obtained by proving some growth properties of $C(z_1, z_2)$ when $(z_1, z_2)$ are complex numbers.

### D. Same Markov sources

We first present a general result when the Markov sources are identical: $\mathbf{P}_1 = \mathbf{P}_2 = \mathbf{P}$. In this case (7) can be rewritten with $c_a(z) = C_a(z, z)$:

$$c_b(z) = (1 - e^{-z})^2 + \sum_{a \in \mathcal{A}} c_a \left(P(a|b) z\right). \tag{10}$$

This equation is directly solvable by the Mellin transform $c_a^*(s) = \int_0^\infty c_a(x) x^{s-1} dx$ defined for $-2 < \Re(s) < -1$. For all $b \in \mathcal{A}$ we find [14]

$$c_b^*(s) = (2^{-s} - 2)\Gamma(s) + \sum_{a \in \mathcal{A}} (P(a|b))^{-s} c_a^*(s) \ . \tag{11}$$

Then the Mellin transform $c^*(s)$ of $C(z, z)$ becomes

$$c^*(s) = (2^{-s} - 2)\Gamma(s) + \sum_{a \in \mathcal{A}} (\pi(a))^{-s} c_a^*(s) \ .$$

Thus

$$c^*(s) = (2^{-s} - 2)\Gamma(s) \left(1 + \langle \mathbf{1}(\mathbf{I} - \mathbf{P}(s))^{-1} | \boldsymbol{\pi}(s) \rangle\right) \tag{12}$$

where $\mathbf{1}$ is the vector of dimension $|\mathcal{A}|$ made of all 1's, $\mathbf{I}$ is the identity matrix, and $\mathbf{P}(s) = \mathbf{P}(s, 0) = \mathbf{P}(0, s)$, $\boldsymbol{\pi}(s)$ is the the vector made of coefficients $\pi(a)^{-s}$ and $\langle . | . \rangle$ denotes the inner product.

By applying the methodology of Flajolet [1], [14], the asymptotics of $c(z)$ for $|\arg(z)| < \theta$ is given by the residues of the function $c^*(s) z^{-s}$ occurring at $s = -1$ and $s = 0$. They are respectively equal to $\frac{2 \log 2}{h} z$ and $-1 - \langle \mathbf{1}(\mathbf{I} - \mathbf{P}(0,0))^{-1} \boldsymbol{\pi}(0) \rangle$. The first residues comes from the singularity of $(\mathbf{I} - \mathbf{P}(s))^{-1}$ at $s = -1$. This leads to Theorem 1(i). When $\mathbf{P}$ is logarithmically rationally balanced then there are additional poles on a countable set of complex numbers $s_k$ regularly spaced on the line $\Re(s_k) = -1$, and such that $\mathbf{P}(s_k)$ has eigenvalue 1. These poles contributes to the periodic terms of Theorem 1(ii).

### E. Different Markov Sources

In this section we establish Theorems 3 and 4. Since $\mathbf{P}_1 \neq \mathbf{P}_2$ we cannot obtain a functional equation for $C_a(z, z)$'s, and therefore we have to deal with two variables $z_1$ and $z_2$. We define the double Mellin transform $C_a^*(s_1, s_2) = \int_0^\infty \int_0^\infty C_a(z_1, z_2) z_1^{s_1 - 1} z_2^{s_2 - 1} dz_1 dz_2$ and similarly the double Mellin transform $C^*(s_1, s_2)$ of $C(z_1, z_2)$. We find

$$\begin{aligned} C_b^*(s_1, s_2) &= \Gamma(s_1)\Gamma(s_2) \\ &\quad + \sum_{a \in \mathcal{A}} (P_1(a|b))^{-s_1} (P_2(a|b))^{-s_2} C_a^*(s_1, s_2) \end{aligned} \tag{13}$$

which leads to

$$C^*(s_1,s_2)=\Gamma(s_1)\Gamma(s_2)\left(1+\langle \mathbf{1}(\mathbf{I}-\mathbf{P}(s_1,s_2))^{-1}|\boldsymbol{\pi}(s_1,s_2)\rangle\right) \tag{14}$$

where $\boldsymbol{\pi}(s_1,s_2)$ denotes the vector composed of $\pi_1(a)^{-s_1}\pi_2(a)^{-s_2}$. In fact to define the Mellin transform we need to apply it to $C(z_1,z_2)-\frac{\partial}{\partial z_1}C(0,z_2)z_1e^{-z_1}-\frac{\partial}{\partial z_2}C(z_1,0)z_2e^{-z_2}$ but we omit this technical detail. The inverse Mellin transform is

$$C(z,z)=\frac{1}{(2i\pi)^2}\int_{\Re(s_1)=\rho_1}\int_{\Re(s_2)=\rho_2}C^*(s_1,s_2)z^{-s_1-s_2}ds_1ds_2 \tag{15}$$

where $(\rho_1,\rho_2)$ belongs to the fundamental strip of $C^*(s_1,s_2)$.

Let $L(s)$ be the function of complex $s$ such that $\mathbf{P}(s,L(s))$ has eigenvalue 1 or where $(\mathbf{I}-\mathbf{P}(s_1,s_2))^{-1}$ cease to exist. The function $L(s)$ is meromorphic and has several branches; one branches describes the set $\mathcal{K}$ when $s$ is real. Now to evaluate the double integral (15) we move the line of integration with respect to $s_2$ from $\rho_2$ to some $M>1$ collecting on the way all residues. In particular, the dominant residue at $s_2=L(s)$ contributes

$$\begin{aligned}C(z,z) &= \frac{1}{2i\pi}\int_{\Re(s_1)=\rho_1}\mu(s_1)\Gamma(s_1)\Gamma(L(s_1))z^{-s_1-L(s_1)}ds_1\\ &\quad +O(z^{\rho_1-M})\end{aligned} \tag{16}$$

where $\mu(s)$ is the residue of $\langle \mathbf{1}(\mathbf{I}-\mathbf{P}(s,s_2))^{-1}\boldsymbol{\pi}(s_1,s_2)\rangle$ at point $(s,L(s))$, that is,

$$\mu(s_1)=\frac{1}{\frac{\partial}{\partial s_2}\lambda(s_1,s_2)}\langle \mathbf{1}|\boldsymbol{\zeta}(s_1,s_2)\rangle\langle \mathbf{u}(s_1,s_2)|\boldsymbol{\pi}(s_1,s_2)\rangle\,\big|_{s_2=L(s_1)}.$$

Above $\lambda(s_1,s_2)$ is the eigenvalue which has value 1 at $(s,L(s))$ and $\mathbf{u}(s_1,s_2)$ and $\boldsymbol{\zeta}(s_1,s_2)$ are the respective left and right eigenvectors with the convention that $\langle\boldsymbol{\zeta}(s_1,s_2)|\mathbf{u}(s_1,s_2)\rangle=1$.

The above expression is implicitly a sum since the function $L(s)$ is meromorphic, but we retain only the branch where $\lambda(s_1,s_2)$ is the main eigenvalue of $\mathbf{P}(s_1,s_2)$ that contributes to the leading term in the expansion of $C(z,z)$. For more details see [7] where detailed analysis is presented in the case when $\mathbf{P}_2$ corresponds to the uniform memoryless case, *i.e.* $\mathbf{P}_2=\frac{1}{|\mathcal{A}|}\mathbf{1}\otimes\mathbf{1}$.

Next, we evaluate (16) by moving the integration line for $s_1$ from $\rho_1$ to $c_1$ which corresponds to the position where function $-s_1-L(s_1)$ (actually $\kappa$) attains the minimum value. We only consider the case when $L(c_1)=c_2<0$ (the other case is similar). The poles are due to the function $\Gamma(.)$. The first encountered pole is $s_1=-1$ but this pole cancels out.

Let's assume $c_1>0$. We meet the second pole at $s=0$ and the residue is equal to $\mu(0)\Gamma(c_0)z^{-c_0}$ since $L(0)=c_0$. This quantity turns out to be the leading term of $C(z,z)$ since the integration on $\Re(s_1)=c_1$ is $O(z^\kappa)$. This proves Theorem 3. When $\mathbf{P}_2$ is logarithmically balanced, there exists $\omega$ such that $\lambda(s,L(s)+ik\omega)=1$ for $k\in\mathbb{Z}$ and the terms $z^{c_0+ik\omega}$ lead to a periodic contribution.

The most tricky part is when $-1<c_1<0$. In this case, $C(z,z)=O(z^\kappa)$ but to find precise estimates one must use the saddle point methods at $s=c_1$ since (16) becomes

$$C(z,z)=\int_{\Re(s)=c_1}\mu(s)\exp(-(s+L(s))\log z)ds.$$

The integrand function above grows exponentially, thus the saddle point methods applies. We find

$$C(z,z)=\frac{e^{\kappa\log z}\mu(c_1)}{\sqrt{(\alpha_2\log z+\beta_2}}\left(1+O(\frac{1}{\sqrt{\log z}})\right)$$

with $\alpha_2=L''(c_1)$ and $\beta_2=\frac{\mu'(c_1)}{\mu(c_1)}$. In fact, the saddle point expansion is extendible to any order of $\frac{1}{\sqrt{\log n}}$. This proves Theorem 4 in the general case. However, in the case when $\mathbf{P}_1$ and $\mathbf{P}_2$ are logarithmically commensurable, the line $\Re(s_1)=c_1$ contains an infinite number of saddle points that contribute to the double periodic function $Q_2(\log n)$ (cf. [7] for more details).

## ACKNOWLEDGMENT

W. Szpankowski work was partially supported by the NSF Science and Technology Center for Science of Information Grant CCF-0939370, NSF Grants DMS-0800568 and CCF-0830140, and, NSA Grant H98230-11-1-0141. W. Szpankowski is also a Visiting Professor at ETI, Gdańsk University of Technology, Poland. D. Milioris work is also supported by INRIA and École Polytechnique ParisTech.

## REFERENCES

[1] P. Flajolet and R. Sedgewick, *Analytic Combinatorics*, Cambridge University Press, Cambridge, 2008.
[2] Ilie, L., Yu, S., and Zhang, K. Repetition Complexity of Words In *Proc. COCOON* 320–329, 2002.
[3] P. Jacquet, Common words between two random strings, *IEEE Intl. Symposium on Information Theory*, 1495-1499, 2007.
[4] V. Becher and P. A. Heiber, A better complexity of finite sequences, Abstracts of the 8th *Int. Conf. on Computability and Complexity in Analysis* and 6th *Int. Conf. on Computability, Complexity, and Randomness* , Cape Town, South Africa, January 31, February 4, 2011, p. 7.
[5] P. Jacquet, and W. Szpankowski, Autocorrelation on Words and Its Applications. Analysis of Suffix Trees by String-Ruler Approach, *J. Combinatorial Theory Ser. A*, 66, 237–269, 1994.
[6] P. Jacquet, and W. Szpankowski, Analytical DePoissonization and Its Applications, *Theoretical Computer Science*, 201, 1–62, 1998.
[7] P. Jacquet and W. Szpankowski, Joint String Complexity for Markov Sources, *23rd International Meeting on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms*, AofA'12, *DMTCS Proc.*, 303-322, Montreal, 2012.
[8] S. Janson, S. Lonardi and W. Szpankowski, On Average Sequence Complexity, *Theoretical Computer Science*, 326, 213-227, 2004.
[9] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.
[10] Li, M., and Vitanyi, P. *Introduction to Kolmogorov Complexity and its Applications.* Springer-Verlag, Berlin, Aug. 1993.
[11] Niederreiter, H., Some computable complexity measures for binary sequences, In *Sequences and Their Applications*, Eds. C. Ding, T. Hellseth and H. Niederreiter Springer Verlag, 67-78, 1999.
[12] J. Fayolle, M. War, Analysis of the average depth in a suffix tree under a Markov model DMTCS Proceedings of AofA 2005.
[13] M. Régnier and W. Szpankowski, On pattern frequency occurrences in a Markovian sequence, *Algorithmica*, 22, 631-649, 1998.
[14] W. Szpankowski, *Analysis of Algorithms on Sequences*, John Wiley, New York, 2001.
[15] J. Ziv, On classification with empirically observed statistics and universal data compression, *IEEE Trans. Information Theory*, 34, 278-286, 1988.