



# 1 Analysis of Lempel-Ziv'78 for Markov Sources

2 Philippe Jacquet 

3 INRIA, Paris, France

4 philippe.jacquet@inria.fr

5 Wojciech Szpankowski 

6 Center for Science of Information, Department of Computer Science, Purdue University, West

7 Lafayette, IN, USA

8 spa@cs.purdue.edu

## 9 — Abstract —

---

10 Lempel-Ziv'78 is one of the most popular data compression algorithms. Over the last few decades  
11 fascinating properties of LZ'78 were uncovered. Among others, in 1995 we settled the Ziv conjecture  
12 by proving that for a *memoryless source* the number of LZ'78 phrases satisfies the Central Limit  
13 Theorem (CLT). Since then the quest commenced to extend it to Markov sources. However, despite  
14 several attempts this problem is still open. The 1995 proof of the Ziv conjecture was based on two  
15 models: In the DST-model, the associated digital search tree (DST) is built over  $m$  *independent*  
16 strings. In the LZ-model a *single* string of length  $n$  is partitioned into variable length phrases such  
17 that the next phrase is not seen in the past as a phrase. The Ziv conjecture for memoryless source  
18 was settled by proving that both DST-model and the LZ-model are asymptotically equivalent. The  
19 main result of this paper shows that this is not the case for the LZ78 algorithm over Markov sources.  
20 In addition, we develop here a large deviation for the number of phrases in the LZ78 and give a  
21 *precise* asymptotic expression for the redundancy which is the excess of LZ78 code over the entropy  
22 of the source. We establish these findings using a combination of combinatorial and analytic tools.  
23 In particular, to handle the strong dependency between Markov phrases, we introduce and precisely  
24 analyze the so called *tail symbol* which is the first symbol of the next phrase in the LZ'78 parsing.

25 **2012 ACM Subject Classification** General and reference → General literature; General and reference

26 **Keywords and phrases** Lempel-Ziv algorithm, digital search trees, depoissonization, analytic com-  
27 binatorics, large deviations

28 **Digital Object Identifier** 10.4230/LIPIcs...

29 **Funding** *Wojciech Szpankowski*: This work was supported by NSF Center for Science of Information  
30 (CSoI) Grant CCF-0939370, and in addition by NSF Grant CCF-1524312.



© Philippe Jacquet and Wojciech Szpankowski;  
licensed under Creative Commons License CC-BY

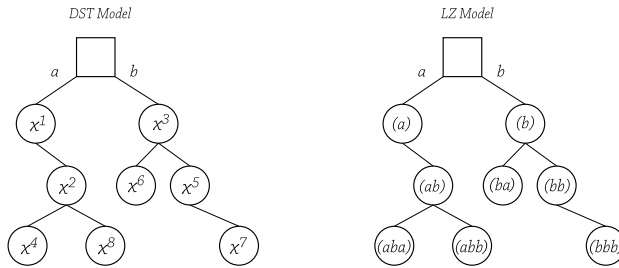
Leibniz International Proceedings in Informatics

LIPICs Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

31 **1** Introduction

32 The Lempel-Ziv compression algorithm [17] is a universal compression scheme. It partitions  
 33 the text to be compressed into consecutive phrases such that the next phrase is the unique  
 34 shortest prefix (of the uncompressed text) not seen before as a phrase. For example,  
 35  $aababbababb$  is parsed as  $()(a)(ab)(abb)(aba)(b)(bb)$ . The LZ'78 compression code consists of  
 36 a pointer to the previous phrase and the last symbol of the current phrase. The distribution  
 37 of the number of phrases and other related quantities (such as redundancy and code length)  
 38 are known for memoryless sources [10; 15] but research over the past 40 years has failed to  
 39 produce any significant progress for Markov sources. In this paper, we resolve the central  
 40 question and present novel large deviations and precise redundancy results that had been  
 41 wanting since the algorithm inception, as well as some surprising findings regarding the  
 42 difference between the memoryless case and the Markov case.

43 It is convenient to organize phrases (dictionary) of the Lempel-Ziv scheme in a *digital*  
 44 *search tree* (DST) [7] which represents a parsing tree. We assume throughout that  $\mathcal{A} = \{a, b\}$ .  
 45 Then the root contains an empty phrase. The first phrase is the first symbol, say " $a \in \mathcal{A}$ "  
 46 which is stored in a node appended to the root. The next phrase is either  $(aa) \in \mathcal{A}^2$  stored  
 47 in another node that branches out from the node containing the first phrase " $a$ " or  $(ab)$  that  
 48 is stored in a node attached to the root. This process repeats recursively until the text is  
 parsed into full phrases (see Figure 1). A detailed description can be found in [3; 7; 8].



49 **Figure 1** The DST-model vs LZ-model. In the DST-model we inserted eight (infinite) strings:  
 $X^1 = ab\mathbf{b}\dots$ ,  $X^2 = ab\mathbf{b}\dots$ ,  $X^3 = bb\mathbf{b}a\dots$ ,  $X^4 = abaa\dots$ ,  $X^5 = bb\mathbf{a}a\dots$ ,  $X^6 = ba\mathbf{a}a\dots$ ,  
 $X^7 = bb\mathbf{b}a\dots$  and  $X^8 = abbb\dots$ , where bold symbols denote DST tail symbols. In the LZ-model  
 we parsed one string  $X = ()(a)(\mathbf{a}b)(\mathbf{b})(\mathbf{a}aba)(\mathbf{b}b)(\mathbf{b}bb)(\mathbf{a}bb)$  with bold denoting LZ tail symbols.

50 We consider two models called the DST-model and the LZ-model. In the DST-model we  
 51 insert *independent strings* although each string may be generated by a source with memory  
 52 like a Markov source. In the LZ-model we parse a *single* string as shown in Figure 1. We  
 53 distinguish two types of DST and LZ models. To define them we need to introduce the path  
 54 length  $L$  as the sum of all depths in the digital search tree or the sum of all phrases in the  
 55 LZ model. In the " $m$ "-DST model we insert  $m$  independent strings into a digital search  
 56 tree – leading to a variable path length denoted as  $L_m$  – while the " $n$ "-DST model is built  
 57 over a random number of independent strings such that the total path length is equal to  $n$ .  
 58 Similarly, we have " $m$ "-LZ and " $n$ "-LZ models: In the former we construct  $m$  LZ phrases  
 59 to form a string of (variable) length denoted as  $\mathcal{L}_m$  while in the " $n$ "-LZ model we parse a  
 60 string of length  $n$  into a variable number of phrases that we denote as  $M_n$ . Throughout,  $m$   
 61 will denote number of strings or phrases while  $n$  will stand for the length of a string.

62 There is a simple relation between  $M_n$  and  $\mathcal{L}_m$  called the *renewal equation* which asserts

$$63 \quad P(M_n > m) = P(\mathcal{L}_m < n). \quad (1)$$

64 Finally, observe that the code length of the LZ78 algorithm is  $C_n = \sum_{k=1}^{M_n} \lceil \log_2(k) \rceil +$   
 65  $\lceil \log_2(|\mathcal{A}|) \rceil$  since the pointer to the  $k$ th node requires at most  $\lceil \log_2 k \rceil$  bits, while the next  
 66 symbol costs  $\lceil \log_2 |\mathcal{A}| \rceil$  bits. For binary alphabet  $\mathcal{A} = \{a, b\}$  we simplify the code length to  
 67  $C_n = M_n (\log_2 M_n + 1)$ .

68 To understand LZ'78 behavior one must analyze the limiting distribution of  $M_n$  and/or  
 69  $\mathcal{L}_m$  connected through the renewal equation (1). For *memoryless* sources we benefited from  
 70 the fact the random variable  $L_m$  and  $\mathcal{L}_m$  are *probabilistically equivalent* as shown in 1995  
 71 paper [3]. Unfortunately, this equivalence breaks for sources with memory such as Markov  
 72 sources. To capture this dependency we introduce the notion of the *tail symbol*. In the  
 73 DST-model the tail symbol of an inserted string is the first non-inserted symbol of that string,  
 74 as shown in Figure 1. In the LZ-model the tail symbol of a phrase is the first symbol of the  
 75 next phrase (see Figure 1). Furthermore, in the Markov case there is additional complication,  
 76 even for the DST-model. In the DST-model we need to consider two digital search trees: one  
 77 built over all (independent) strings starting with symbol  $a \in \mathcal{A}$ , and the second one built  
 78 over all strings that start with  $b \in \mathcal{A}$ . At the end we construct a cumulative knowledge by  
 79 weighting over the initial symbols (see [6]).

80 In this conference paper, we present large deviation results for the number of phrases  $M_n$   
 81 in “ $n$ ”-LZ model and the average length of a LZ (Markov) string built over  $m$  phrases in the  
 82 “ $m$ ”-LZ model.<sup>1</sup> In the memoryless case we could read the number of phrases  $M_n$  directly  
 83 from the path length  $L_m$  of the  $m$ -DST model. It is *not* the case in the Markov model but  
 84 through the tail symbol distribution we will connect both quantities. Recall that  $\mathcal{L}_m$  is the  
 85 length of a string generated by a Markov source which is parsed by the LZ78 scheme until  
 86 we see  $m$  phrases (our  $m$ -LZ model). This should be compared to the total path length  $L_m$   
 87 (notice roman font for  $L$ ) in the the  $m$ -DST model. In the memoryless case, we proved in  
 88 [3; 5] that the expected value of  $L_m$  and the expected value of the length of a string built  
 89 from  $m$  phrases,  $\mathcal{L}_m$ , are the same. Somewhat surprisingly it is not the case for the Markov  
 90 case. We will prove in Theorem 5 that  $\mathbf{E}[L_m] - \mathbf{E}[\mathcal{L}_m] = \Theta(m)$ .

91 Let us now briefly review literature on LZ'78 and DST analysis. The goal is to prove  
 92 the Central Limit Theorem (CLT) for the number of phrases and establish precise rate of  
 93 decay of the LZ'78 code redundancy for Markov sources. For memoryless sources, CLT was  
 94 already proved in [3] while the average redundancy was presented in [10; 15]. It should  
 95 be pointed out that since 1995 [3] no simpler, in fact, no new proof of CLT was presented  
 96 except the one by Neininger and Rüschemdorf [14] but only for *unbiased* memoryless sources  
 97 (as in [1]). The only known to us analysis of LZ'78 for Markov sources is presented in [6],  
 98 but the authors restricted their attention to a single phrase. We should point out that for  
 99 another Lempel-Ziv scheme known as LZ'77 algorithm, Fayolle and Ward [2] analyzed an  
 100 associated suffix tree built over a Markov string and obtained the distribution of the depth,  
 101 which allows us to conclude the limiting distribution of a phrase in the LZ'77 scheme (see  
 102 also [11; 12]). Regarding analysis of digital search trees, and in general digital trees, more  
 103 is known [8; 7; 16]. Digital trees for memoryless sources were analyzed in [1; 10; 7] while  
 104 digital trees under Markovian models were studied in [6; 9; 2]. This information is surveyed  
 105 in detail in [7].

---

<sup>1</sup> From now on we drop the quotes around  $m$  and  $n$  to simplify the presentation.

106 The paper is organized as follows. In the next section we present our main results  
 107 regarding the LZ and DST models including the mean, variance and distribution of the  
 108 number of tail symbols in the DST model (see Theorem 2–4), and large deviations as well  
 109 as precise redundancy for the LZ model (see Theorems 5–6). We prove these findings in  
 110 Section 3 (DST model) and in Section 4 (LZ model), with most details delayed till the  
 111 appendix. Throughout we use combinatorics on words and analytic tools such as generating  
 112 functions, Poisson transform, analytic depoissonization, and Mellin transform.

## 113 2 Main Results

114 We consider a stationary ergodic Markov source generating a sequence of symbols drawn  
 115 from a finite alphabet  $\mathcal{A}$ . In this conference paper we study only a Markovian process of  
 116 order 1 with the transition matrix  $\mathbf{P} = [P(c|d)]_{c,d \in \mathcal{A}}$  where  $\mathcal{A} = \{a, b\}$ . In this section we  
 117 present our main results with proof delayed till Sections 3–4 and appendix. However, first  
 118 we present a road map of our methodology and findings.

119 Our main goal is to analyze the Lempel-Ziv'78 scheme for Markovian input. However,  
 120 as discussed before, we first consider an auxiliary model named DST-model built over  $m$   
 121 independent Markov strings, also called the  $m$ -DST model. However, for Markov sources  
 122 we need to construct two *conditional* digital search trees: one built over  $m$  Markov strings  
 123 all starting with symbol  $a \in \mathcal{A}$  and the other DST built over  $m$  strings starting with  $b \in \mathcal{A}$ .  
 124 We write  $c \in \mathcal{A}$  for a generic symbol from  $\mathcal{A}$ , that is, either  $c = a$  or  $c = b$ . For a given  
 125  $c \in \mathcal{A}$ , we consider  $m$  independent Markov strings all starting with  $c$  and build an  $m$ -DST  
 126 tree. For such a tree we analyze two quantities, namely the total path length denoted as  
 127  $L_m^c$ , and the number  $T_m^c(a)$  of inserted strings (all starting with  $c$ ) with the tail symbol  $a$ ,  
 128 that is, among  $m$  Markov strings there are  $T_m^c(a)$  strings with the tail symbol  $a$ . Clearly,  
 129  $T_m^c(a) + T_m^c(b) = m$ . Throughout, we also assume that the tail symbol is always  $a$  so we just  
 130 write  $T_m^c := T_m^c(a)$ . In Theorems 2-3 we summarize our new results regarding  $T_m^c$ , while in  
 131 Theorem 4 we present large deviation results for both  $T_m^c$  and  $L_m^c$ .

132 Second, we consider the  $m$ -LZ model (in which we run LZ78 algorithm on a single  
 133 string until we see  $m$  phrases) and tie it up to the  $m$ -DST model just discussed. Here we  
 134 use combinatorial approach. For a given sequence  $\mathbf{s}$  over  $\mathcal{A}$  of length  $m$  we compare in  
 135 Lemmas 10-11 two probabilities: (i) the probability that in the  $m$ -LZ model (constructed  
 136 from  $m$  LZ phrases) we end up with a LZ sequence of length  $n$  having all tail symbols equal  
 137 to  $\mathbf{s}$ ; and (ii) the probability that in the  $m$ -DST model (built over  $m$  independent Markov  
 138 strings) the resulting digital search tree has path length equal to  $n$  and all tail symbols are  
 139 equal to  $\mathbf{s}$ . Using this, we present in Theorem 5 our large deviations for the  $m$ -LZ model  
 140 and using the renewal equation (1) in Theorem 6 we establish large deviations for the  $n$ -LZ  
 141 model. In Corollary 7 we find a *precise* expression for the redundancy of LZ78 for Markov  
 142 sources.

143 Finally, when comparing the average path length  $L_m^c$  in the  $m$ -DST model with the  
 144 length  $\mathcal{L}_m^c$  in the  $m$ -LZ model we shall use the following simple fact.

145 ► **Proposition 1.** For  $\delta < 1$  let there exist  $B, C > 0$  such that for a discrete random variable  
 146  $X_m$  the following holds uniformly

$$147 \quad P(X_m = k) \leq B \exp(-Cm^{-\delta}|k - A_m|). \quad (2)$$

148 Then

$$149 \quad \mathbf{E}[X_m] = A_m + \Theta(m^\delta). \quad (3)$$

150 **Proof.** Define  $B_m = m^\delta(\log B)/C \leq |k - A_m|$ . Then it is easy to see that  $EX_m =$   
 151  $\sum_k kP(X_m = k) = A_m + \sum_k (k - A_m)P(X_m = k)$ , and the latter term can be estimated by  
 152 the integral  $2B \int_0^\infty \exp(-Cm^{-\delta x})(x + 1)dx = O(m^\delta)$ . This complete the proof. ◀

## 153 2.1 Results on DST

154 In this section we summarize our results for the  $m$ -DST model: We first focus on the number  
 155 of times,  $T_m^c := T_m^c(a)$ , the tail symbol is  $a$  when all  $m$  Markov sequences start with  $c \in \mathcal{A}$ .  
 156 Then we study the path length  $L_m^c$  in the  $m$ -DST model when all sequences start with  $c$ .  
 157 Finally, we present large deviations for both  $T_m^c$  and  $L_m^c$ .

158 For  $c \in \mathcal{A}$ , let  $D_{m,k}^c = P(T_m^c = k)$  and  $D_m^c(u) = E[u^{T_m^c}]$  be the probability generating  
 159 function of  $T_m^c$  defined for a complex variable  $u$ . We have the recursion:

$$160 \quad D_{m+1}^c(u) = (P(a|c)u + 1 - P(a|c)) \sum_k \binom{m}{k} P(a|c)^k P(b|c)^{m-k} D_k^a(u) D_{m-k}^b(u) \quad (4)$$

161 subject to  $D_0^c(u) = 1$  and  $D_1^c(u) = P(a|c)u + 1 - P(a|c)$ . Furthermore, define the bivariate  
 162 Poisson transform  $D_c(z, u) = \sum_{m \geq 0} \mathbf{E}[u^{T_m^c}] \frac{z^m}{m!} e^{-z}$ . From above we easily find the following  
 163 differential-functional equation

$$164 \quad \partial_z D_c(z, u) + D_c(z, u) = D_1^c(u) D_a(P(a|c)z, u) \cdot D_b(P(b|c)z, u) \quad (5)$$

165 with  $D_c(z, 1) = 1$  where  $\partial_z$  is the partial derivative with respect to variable  $z$ .

We now focus on the first Poisson moment  $X_c(z) = \partial_u D_c(z, 1)$  where  $\partial_u$  is the derivative  
 with respect to variable  $u$ . We also study the Poisson variance  $V_c(z) = \partial_u^2 D_c(z, 1) + X_c(z) -$   
 $(X_c(z))^2$ , and the limiting distribution of  $T_m^c$ . After finding asymptotic behavior of the  
 Poisson mean  $X_c(z)$  and variance  $V_c(z)$  for large  $z \rightarrow \infty$  we invoke the depoissonization  
 lemma of [4] to extract the original mean and variance:

$$\mathbf{E}[T_m^c] = X_c(m) - \frac{1}{2} m \partial_z X_c(m) + O(X_c(m)/m), \quad \text{Var}[T_m^c] \sim V_c(m) - m[\partial_z X_c(m)]^2.$$

166 Let us start with the Poisson mean  $X_c(z)$ . Taking the derivative of (5) with respect to  $u$   
 167 and setting  $u = 1$  we find

$$168 \quad \partial_z X_c(z) + X_c(z) = P(a|c) + X_a(P(a|c)z) + X_b(P(b|c)z). \quad (6)$$

169 To complete this equation we need to calculate the initial values of  $\mathbf{E}[T_m^c]$ . It is easy to see  
 170 that

$$171 \quad \mathbf{E}[T_0^c] = 0, \quad \mathbf{E}[T_1^c] = P(a|c), \quad \mathbf{E}[T_2^c] = P(a|c) + P(a|c)P(a|a) + P(b|c)P(a|b). \quad (7)$$

172 In a similar fashion we can derive the differential-functional equation for the Poisson  
 173 variance. After some tedious algebra we arrive at

$$174 \quad \partial_z V_c(z) + V_c(z) = P(a|c) - P^2(a|c) + [\partial_z X_c(z)]^2 + V_a(P(a|c)z) + V_b(P(b|c)z). \quad (8)$$

175 Both differential-functional system of equations (5) and (7) can be solved using complicated  
 176 Mellin transform approach [16]. We will provide details of our approach in the Appendix.  
 177 For now we need to introduce some extra notation to present our main results. For complex  
 178  $s$  define

$$179 \quad \mathbf{P}(s) = \begin{bmatrix} P(a|a)^{-s} & P(b|a)^{-s} \\ P(a|b)^{-s} & P(b|b)^{-s} \end{bmatrix}. \quad (9)$$

For such  $\mathbf{P}(s)$  we denote by  $\lambda(s)$  the main eigenvalue and  $\boldsymbol{\pi}(s)$  the main eigenvector. We notice that  $\boldsymbol{\pi}(-1)$  is the stationary vector of the Markov process. We also need another matrix

$$\mathbf{Q}(s) = \prod_{i \geq 1} (\mathbf{I} - \mathbf{P}(s - i))^{-1} \prod_{j=-\infty}^{j=-2} (\mathbf{I} - \mathbf{P}(j))$$

180 defined for  $\Re(s) \in (-2, 0)$ . Furthermore,  $\langle \mathbf{x}, \mathbf{y} \rangle$  is the scalar product of vectors  $\mathbf{x}$  and  $\mathbf{y}$ .

181 Now we are in the position to formulate our main result.

182 ► **Theorem 2.** *Consider a digital search tree built over  $m$  independent sequences ( $m$ -DST) generated by a Markov source. We have  $\mathbf{E}[T_m^c] = \tau_c(m)m$  and  $\mathbf{E}[L_m^c] = m \log m/h + m + \mu_c(m)m$  such that:*

- 183 ■  $\tau_c(m+1) - \tau_c(m) = O(1/m)$  and  $\mu_c(m+1) - \mu_c(m) = O(1/m)$
- 184 ■  $\forall (c, d) \in \mathcal{A}^2 \tau_c(m) - \tau_d(m) = O(1/m)$  and  $\mu_c(m) - \mu_d(m) = O(1/m)$ .

185 Thus  $\tau_c(m) = \tau(m) + O(1/m)$  where  $\tau(m)$  does not depend on initial symbol  $c$ . In fact,  $\tau(m)$  depends on the tail symbol, but since throughout the paper we assume the tail symbol is always  $a$ , we drop this dependency on  $a$  in  $\tau(m)$ . We present precise formula on  $\tau(m)$  in the next theorem.

186 Similarly we have  $\mu_c(m) = \mu(m) + O(1/m)$ . The function  $\mu(m)$  for Markov sources is given in Theorem 1 of [6]. For the memoryless source, it is  $\frac{h_2}{h} + \gamma - 1 + \alpha$  and the average path length is  $m \log m/h + m\mu(m)$ , as discussed in [3].

187 To complete our analysis of the tail symbol, we present now precise behaviour of  $\tau(m)$ . We give a detailed proof in the Appendix.

188 ► **Theorem 3.** *For  $(a, b, c) \in \mathcal{A}^3$  define*

$$189 \quad \alpha_{abc} = \log \left[ \frac{P(a|b)P(c|a)}{P(c|b)} \right]. \quad (10)$$

190 (i) [Aperiodic case] *If not all  $\{\alpha_{abc}\}$  are rational, then  $\tau(m) = \bar{\tau} + o(1)$  with*

$$191 \quad \bar{\tau} = \pi_a + \frac{1}{\lambda'(-1)} \langle (\boldsymbol{\pi}'(-1) + \boldsymbol{\pi}\mathbf{Q}'(-1)) (\mathbf{I} - \mathbf{P})\mathbf{P}\mathbf{e}_a \rangle, \quad (11)$$

192 where  $\pi_a$  is the stationary distribution of symbol  $a$ , and  $\mathbf{e}_a$  is the vector made of a single 1 at the position corresponding to symbol  $a$  and zero otherwise.

193 [Periodic case] *If all  $\{\alpha_{abc}\}$  are rationally related, then for some  $\varepsilon > 0$  we have  $\tau(m) = \bar{\tau}(m) + O(m^{-\varepsilon})$  with  $\bar{\tau}(m) = \bar{\tau} + Q_1(\log m)$ , where  $Q_1(\cdot)$  is a periodic function.*

194 (ii) [Variance] *The variance  $\text{Var}[T_m^c]$  grows linearly, that is  $\text{Var}[T_m^c] \sim m\omega_a(m)$ , where  $\omega_a(m) = \bar{\omega}_a$  for the aperiodic case and  $\omega_a(m) = \bar{\omega}_a + Q_2(m)$  for the periodic case, where  $\bar{\omega}_a$  is given explicitly in the Appendix in (A.17) of Theorem 14, and  $Q_2(m)$  is a nonzero periodic function for rationally related case, and zero otherwise.*

195 (iii) [Central Limit Theorem] *For any  $c \in \mathcal{A}$  we have*

$$\frac{T_m^c - \mathbf{E}[T_m^c]}{\text{Var}[T_m^c]} \rightarrow N(0, 1)$$

196 where  $N(0, 1)$  denotes the standard normal distribution.

197 Similarly we have the same behaviour for  $\mu(m)$  which is equal to  $\bar{\mu} + o(1)$  in the aperiodic case and, in the periodic case, is equal to  $\bar{\mu} + Q_3(\log m) + O(m^{-\varepsilon})$  whose expressions are in [3] and [6] where  $Q_3(\cdot)$  is a periodic function. For details the reader is referred to [6].

212 We notice that, unexpectedly, the number of tail symbols equal to  $a$  is *not* converging to  
 213  $n\pi_a$  as we should expect from a Markovian sequence. The reason is that the tail symbol is  
 214 not picked up at random in the sequence but occurs when the sequence path leaves the tree.

215 Finally, we present joint large deviations for both  $T_m^c$  and  $L_m^c$  which is a new result  
 216 needed to establish large deviations for the LZ model. We prove it in Section 3.

217 ► **Theorem 4.** *Consider a digital search tree (DST) built over  $m$  independent sequences*  
 218 *generated by a Markov source. For all  $\delta > 1/2$  there exist  $B, C$  and  $\beta$  strictly positive such*  
 219 *that for all  $x > 0$  uniformly in  $x$*

$$220 \quad P(|T_m^c - \mathbf{E}[T_m^c]| + |L_m^c - \mathbf{E}[L_m^c]| \geq xm^\delta) \leq Be^{-x Cm^\beta} \quad (12)$$

221 for large  $m$ .

## 222 2.2 Results for the LZ78 Model

223 Let us start with the  $m$ -LZ model. For a given  $m$ , let  $\mathcal{L}_m^c$  (note calligraphic  $\mathcal{L}$ ) be the length  
 224 of the LZ'78 string composed of  $m$  phrases when the first phrase starts with symbol  $c$ . For  
 225 memoryless sources, this quantity is equivalent to the path length  $L_m$  in the associated DST  
 226 built over  $m$  independent strings. However, it is not the case for Markov sources. In Section 4  
 227 we prove Theorem 5 presented below by showing that  $\mathbf{E}[L_m^c] - \mathbf{E}[\mathcal{L}_m^c] = \Theta(m)$ , unlike in the  
 228 memoryless case. Figure 2 compares the difference  $\mathbf{E}[L_m^c] - \mathbf{E}[\mathcal{L}_m^c]$  obtained by simulation  
 229 results confirming our theoretical findings.

230 ► **Theorem 5.** *For  $m$  given, let  $m^* := m^*(m)$  be the root of  $x - x\tau(x) - (m-x)\tau(m-x)$ .*  
 231 *(i) The average length  $\mathbf{E}[\mathcal{L}_m^c]$  of the LZ-sequence consisting of the first  $m$  phrases is (for the*  
 232 *aperiodic case)*

$$233 \quad \mathbf{E}[\mathcal{L}_m^c] = m \log m/h + \mu(m^*)m^* + \mu(m-m^*)(m-m^*) + m(1-H(m^*/m)/h) + O(m^\delta) \quad (13)$$

234 where  $H(x) = -x \log x - (1-x) \log(1-x)$  is the binary entropy.

235 (ii) For all  $\delta > 1/2$  there exist  $B, C, \beta > 0$ , and  $\gamma > 0$  such that uniformly for all  $x > 0$

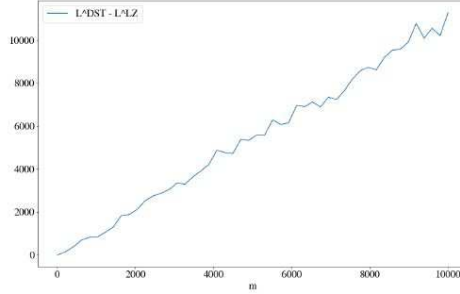
$$236 \quad P(|\mathcal{L}_m^c - \mathbf{E}[\mathcal{L}_m^c]| \geq xm^\delta) \leq Bm^\gamma e^{-x Cm^\beta} \quad (14)$$

237 for large  $m$ .

### 238 Remark

239 The property of function  $\tau(\cdot)$  implies that the equation  $x - x\tau(x) - (m-x)\tau(m-x)$  has a  
 240 single root as we will see in the proof of Section 4. Notice that  $m^*/m$  converges to  $\bar{\tau}$  in the  
 241 aperiodic case, and similarly  $\mu(m^*)m^* + \mu(m-m^*)(m-m^*)$  is asymptotically equivalent  
 242 to  $\bar{\mu}m$ . In the periodic case there will be small periodic contributions (contained in  $\tau(m)$   
 243 and  $\mu(m)$ ) as shown in Theorem 3. Notice that  $H(m^*/m)$  is the tail symbol entropy, which  
 244 is equal to  $h$  when the source is memoryless.

245 Our next goal is to present large deviation for the number of LZ phrases in the  $n$ -LZ  
 246 model. Let  $M_n^c$  be the number of phrases obtained by parsing a Markovian sequence of length  
 247  $n$  starting with symbol  $c$ . By the renewal equation (1) we have  $P(M_n^c > m) = P(\mathcal{L}_m^c < n)$   
 248 for all legitimate  $m$  and  $n$ . This allows us to read large deviation of  $M_n^c$  from Theorem 5.  
 249 Following the footsteps of Theorem 2 of [5] we arrive at our next main result.



■ **Figure 2** The difference  $\mathbf{E}[L_m^c] - \mathbf{E}[C_m^c]$  by simulation confirming that it grows linearly with  $m$ .

► **Theorem 6.** For all  $\delta > 1/2$  there exist  $B, C, \beta,$  and  $\gamma$  all strictly positive such that

$$P(|M_n^c - \ell_c^{-1}(n)| \geq xn^\delta) \leq Bn^\gamma e^{-x C n^\beta}$$

where  $\ell_c^{-1}(\cdot)$  is the inverse function of  $\ell_c(m) = \ell(m) + o(1)$  defined as  $\ell(m) = \frac{m}{h} (\log m + \beta(m))$  with

$$\beta(m) = h\mu(m^*)m^*/m + h\mu(m - m^*)(m - m^*)/m - h + H(m^*/m)$$

250 where  $\mu(m)$  contains the extra fluctuating function in the periodic case.

251 Using Theorem 6 we can find a precise estimate on the LZ'78 redundancy. Indeed,  
 252 a good approximation for the LZ'78 code length is  $C_n^c = M_n^c (\log M_n^c + 1)$ . The average  
 253 conditional redundancy is defined as  $r_n^c := \mathbf{E}[C_n^c]/n - h$ , while the total average redundancy  
 254 is  $r_n = \pi_a r_n^a + \pi_b r_n^b$ .

► **Corollary 7.** The average redundancy rate  $r_n$  satisfies for all  $\frac{1}{2} < \delta < 1$ :

$$r_n = h \frac{1 - \beta(\ell^{-1}(n))}{\log \ell^{-1}(n) + \beta(\ell^{-1}(n))} + O(n^{\delta-1} \log n) \sim h \frac{1 - \beta(\ell^{-1}(n))}{\log n},$$

and more specifically in the aperiodic case we have

$$r_n \sim h \frac{1 - \bar{\mu}}{\log n} + \frac{H(\bar{\tau}) - h}{\log n}$$

255 for large  $n$ .

### 256 3 Proof of Theorem 4 for DST

Now we prove Theorem 4, that is, the joint large deviations for  $T_m^c$  and  $L_m^c$  in the  $m$ -DST model. We use Chernoff's bounds, so we need to introduce some bivariate generating functions. Define  $P_{m,k,\ell}^c = P(T_m^c = k \ \& \ L_m^c = \ell)$ ,  $P_m^c(u, v) = \mathbf{E}[u^{T_m^c} v^{L_m^c}] = \sum_{k,\ell} P_{m,k,\ell}^c u^k v^\ell$  and  $P_c(z, u, v)$  to be the Poisson generating function  $P_c(z, u, v) = \sum_m P_m^c(u, v) \frac{z^m}{m!} e^{-z}$ . The following partial differential equation for  $P_c(z, u, v)$  is easy to establish from (5)

$$\partial_z P_c(z, u, v) + P_c(z, u, v) = (uP(a|c) + P(b|c))P_a(P(a|c)zv, u, v)P_b(P(b|c)zv, u, v).$$

257 Lemma below is equivalent to Theorem 10 of [5] so we skip the proof in this conference  
 258 paper.



259 ► **Lemma 8.** For all real number  $\varepsilon' > 0$  and  $\varepsilon > 0$ , there exists  $0 < \vartheta < \pi/2$  and a complex  
 260 neighborhood  $\mathcal{U}(0)$  of 0 such that for  $(t_1, t_2) \in \mathcal{U}(0)^2$  and  $|\arg(z)| < \vartheta \log(P_c(z, e^{t_1|z|^{-\varepsilon'}}, e^{t_2|z|^{-\varepsilon'}}))$   
 261 exists and  $\log(P_c(z, e^{t_1|z|^{-\varepsilon'}}, e^{t_2|z|^{-\varepsilon'}})) = O(z^{1+\varepsilon})$  uniformly in  $(t_1, t_2) \in \mathcal{U}(0)^2$ .

262 To prove Theorem 4 we need the following property that will be established in the final  
 263 version of this paper.

264 ► **Lemma 9.** For all  $\delta > 1/2$  there exists  $B$  such that

$$265 \quad \left| P_m^c(e^{\tau_1 m^{-\delta}}, e^{\tau_2 m^{-\delta}}) \exp(-m^{-\delta}(\tau_1 \mathbf{E}[T_m^c] + \tau_2 \mathbf{E}[L_m^c])) \right| \leq B\sqrt{m}. \quad (15)$$

266 Now we proceed to prove Theorem 4. We apply Markov inequality for all  $\theta$  and for all  $x > 0$

$$\begin{aligned} P(|T_m^c - \mathbf{E}[T_m^c]| + |L_m^c - \mathbf{E}[L_m^c]| \geq 2xm^\delta) &\leq P(|T_m^c - \mathbf{E}[T_m^c]| \geq xm^\delta \vee (|L_m^c - \mathbf{E}[L_m^c]| \geq xm^\delta)) \leq \\ &= \left( P_m^c(e^\theta, 1)e^{-E[T_m^c]\theta} + P_m^c(e^{-\theta}, 1)e^{E[T_m^c]\theta} \right) e^{-x\theta m^\delta} \\ &\quad + \left( P_m^c(1, e^\theta)e^{-E[L_m^c]\theta} + P_m^c(1, e^{-\theta})e^{E[L_m^c]\theta} \right) e^{-x\theta m^\delta}. \end{aligned}$$

To complete the proof we will use (15) of Lemma 9. If we take  $\tau_1 = \pm C$  and  $\tau_2 = 0$  (and reverse) for some  $C > 0$  such that  $(\tau_1, \tau_2) \in \mathcal{U}(0)^2$ , and  $\theta = Cm^{-\delta'}$  for some  $\delta' < \delta$ , then we find  $e^{\theta m^\delta} = e^{-Cm^\beta}$  with  $\beta = \delta - \delta' > 0$ , and

$$P(|T_m^c - \mathbf{E}[T_m^c]| + |L_m^c - \mathbf{E}[L_m^c]| \geq 2xm^\delta) \leq 4\sqrt{m}Be^{-xCm^\beta}$$

267 which prove (12) of Theorem 4. We can readjust by taking  $0 < \beta' < \beta$  and the value of  $B$  to  
 268 omit the factor  $\sqrt{m}$ .

## 269 4 Proof of Theorem 5 for LZ

270 We now consider the LZ'78 algorithm over a single infinite sequence generated by a Markov  
 271 source, that is, the  $n$ -LZ model and connect it to the  $n$ -DST model in which the path length  
 272 is equal to  $n$  (over a variable number of independently inserted strings). In the  $m$ -LZ model  
 273 there are exactly  $m$  LZ phrases, each being a block carved in the Markovian sequence. The  
 274 blocks are *not* i.i.d Markovian sequences.

275 Let  $\mathcal{P}_{m,n}^c$  be the probability that the length of the first  $m$  LZ phrases is exactly  $n$  (when  
 276 the first symbol is  $c$ ), leading to the  $n$ -LZ model. Notice that not every pair  $(n, m)$  is feasible  
 277 in the LZ model since by adding another phrase the path length may “jump” by more than  
 278 one. We are interested in finding an asymptotic estimate of  $\mathcal{P}_{m,n}^c$ . We start by introducing  
 279 yet another model. Let  $\mathbf{s}$  be a sequence of  $m$  symbols, namely  $\mathbf{s} = (c_1, \dots, c_m) \in \mathcal{A}^m$ . For  
 280  $c \in \mathcal{A}$  we now compute the probability  $\mathcal{P}_{\mathbf{s},n}^c$  that an infinite Markovian sequence starting  
 281 with symbol  $c$  when parsed by LZ algorithm satisfies the following two properties: (i) the  
 282 first  $m$  blocks have tail symbols  $c_i \in \mathbf{s}$  for  $i \leq m$  so that  $c_i$  is the first symbol of block  $i + 1$ ;  
 283 (ii) the length of the first  $m$  LZ phrases is equal to  $n$ . If a string satisfies these two conditions,  
 284 then we say it is  $(\mathbf{s}, n)$  compatible and that it belongs to the  $(\mathbf{s}, n)$ -LZ model.

285 Given a string  $\mathbf{s}$  of tail symbols we denote by  $\mathbf{t}_c^a(\mathbf{s})$  (resp.  $\mathbf{t}_c^b(\mathbf{s})$ ) the subsequence of  $\mathbf{s}$   
 286 consisting of tail symbols of the LZ blocks starting with symbol  $a$  (resp. starting by symbol  
 287  $b$ ). Now, it is easy to see that given the initial symbol  $c$  we can deduce the sequence of  
 288 tails symbols and initial symbols of all phrases just by looking at the sequence  $\mathbf{s}$ , where the  
 289 initial symbol of the next phrase is the tail symbol of the previous phrase. For example, if  
 290  $\mathbf{s} = (a, b, a, b, b)$  and  $c = a$  we have the following tail symbol and initial symbol sequence  
 291 displayed in the following table:

block #	initial symbol	tail symbol
1	$a$	$a$
2	$a$	$b$
3	$b$	$a$
4	$a$	$b$
5	$b$	$b$

By taking the blocks (phrases) starting with  $c = a$  we find  $\mathbf{t}_a^a(\mathbf{s}) = (a, b, b)$  and the blocks starting with  $b$  yield  $\mathbf{t}_a^b(\mathbf{s}) = (a, b)$ .

Now we consider a sequence  $\mathbf{t}$  of  $m$  symbols and introduce a new  $n$ -DST model which we call  $(\mathbf{t}, n)$ -DST model. We define by  $P_{\mathbf{t}, n}^c$  the probability that  $m$  i.i.d. (independent) Markovian sequences all starting with  $c$  satisfy the following two conditions (notice that we use roman  $P$  for this probability and calligraphic  $\mathcal{P}$  for LZ model): (i) the tail symbol sequence follows the sequence  $\mathbf{t}$ ; (ii) the external path length of the DST is exactly  $n$ . We will say that such  $m$  strings are  $(\mathbf{t}, n)$ -fit if they satisfy the above conditions and call it  $(\mathbf{t}, n)$ -DST model. We also define

$$P_{m, k, n}^c = \sum_{\mathbf{t}: |\mathbf{t}|=m, |\mathbf{t}|_a=k} P_{\mathbf{t}, n}^c \quad (16)$$

with  $|\mathbf{t}|$  being the length of sequence  $\mathbf{t}$  and  $|\mathbf{t}|_a$  being the number of symbols equal to  $a$  in it.

We finally establish the following fundamental lemma that connects the above two parameters which also connects the LZ parsing over a single Markovian sequence and the DST made of independent Markovian sequences, that is,  $(\mathbf{s}, n)$ -LZ model and  $(\mathbf{t}, n)$ -DST model where  $\mathbf{t}$  is a function of  $\mathbf{s}$ .

► **Lemma 10.** *For any  $\mathbf{s} \in \mathcal{A}^m$  we have*

$$\mathcal{P}_{\mathbf{s}, n}^c = \sum_{n_a} P_{\mathbf{t}_c^a(\mathbf{s}), n_a}^a P_{\mathbf{t}_c^b(\mathbf{s}), n-n_a}^b \quad (17)$$

where  $n_c$  (equal either to  $n_a$  or  $n_b$ ) is the path length in  $n_c$ -DST model with all strings starting with  $c$ , and  $\mathbf{t}_c^a(\mathbf{s})$ ,  $\mathbf{t}_c^b(\mathbf{s})$  are substrings of  $\mathbf{s}$  as defined above.

**Proof.** In this conference paper, we give a proof using an example to ease the presentation. Let us consider  $X = aabbababab \dots$  which results in the following LZ blocks:  $(a)(ab)(b)(aba)(ba)(b \dots)$ . Or equivalently  $X = \mathbf{aabbababab} \dots$  where the initial block (phrase) symbols are displayed in bold. We notice that the first five blocks (excluding the initial empty block) accounts for a string of length 9. Thus the sequence  $X$  is  $(\mathbf{s}, 9)$  compatible with  $\mathbf{s} = (a, b, a, b, b)$ . Given that  $X$  starts with symbol  $a$  we have  $P(X) = P(\mathbf{a}|a)P(\mathbf{aa}|a)P(\mathbf{abb}|a)P(\mathbf{bab}|b)P(\mathbf{abab}|a)P(\mathbf{bab}|b)$ . Notice that we display in bold the tail symbol of each block (which is the initial symbol of the next block). We must incorporate  $P(X)$  into  $P_{\mathbf{s}, 9}^a$ . In fact  $X$  should be viewed as the set of (infinite) strings having  $aabbababab$  as the common prefix. We can rewrite  $P(X)$  by regrouping the terms with respect to the initial symbol of each block as:  $P(X) = [P(\mathbf{aa}|a)P(\mathbf{abb}|a)P(\mathbf{abab}|a)] \times [P(\mathbf{ba}|b)P(\mathbf{bab}|b)]$ . Observe that the sequence of strings  $(\mathbf{aa}, \mathbf{abb}, \mathbf{abab})$  are the prefixes of a set of tuples of independent infinite strings that are all  $(\mathbf{s}^a, 6)$  compatible with  $\mathbf{s}^a = \mathbf{t}_a^a(\mathbf{s}) = (a, b, b)$  under the condition that the strings start with symbol  $a$  (the path length in the DST excludes the tail symbols, thus we must remove one from the length of each prefix). The probability of such event is exactly  $P(\mathbf{aa}|a)P(\mathbf{abb}|a)P(\mathbf{abab}|a)$  and must be incorporated in  $P_{\mathbf{s}^a, 6}^a$ . Furthermore, these sequences are used to build one (left) part of the DST tree with independent Markov strings all starting with  $a$ . The same holds for the sequence of strings  $(\mathbf{ba}, \mathbf{bab})$  which is  $(\mathbf{s}^b, 3)$  compatible with  $\mathbf{s}^b = \mathbf{t}_a^b(\mathbf{s}) = (a, b)$  and used to build the other part (right) of the DST tree. This leads to (17). ◀

332 The next crucial lemma connects  $n$ -LZ and  $n$ -DST models.

333 ► **Lemma 11.** *The following holds*

$$334 \quad \mathcal{P}_{m,n}^c \leq \sum_{n_a} \sum_k \sum_{m_a} (P_{m_a,k,n_a}^a P_{m-m_a,m_a-k,n-n_a}^b) \quad (18)$$

$$335 \quad + P_{m_a,k,n_a}^a P_{m-m_a,m_a-k-1,n-n_a}^b + P_{m_a,k,n_a}^a P_{m-m_a,m_a-k+1,n-n_a}^b)$$

336 where  $n_a$  is the total path length of the first  $m_a$  phrases starting with an “a”.

337 **Proof.** We naturally have  $\mathcal{P}_{m,n}^c = \sum_{|\mathbf{s}|=m} \mathcal{P}_{\mathbf{s},n}^c$  where  $|\mathbf{s}|$  is the length of the sequence  $\mathbf{s}$ .  
 338 Similarly we have  $P_{m,k,n}^c = \sum_{\mathbf{t}, |\mathbf{t}|=m, |\mathbf{t}|_a=k} P_{\mathbf{t},n}^c$  with  $|\mathbf{t}|_a$  is the number of symbols identical  
 339 to  $a$  in  $\mathbf{t}$ . The rest follows from Lemma 10 but we need to take into account some boundary  
 340 effects.

Let’s look at it in more details. By (17) and above we find

$$\mathcal{P}_{m,n}^c = \sum_{|\mathbf{s}|=m} \sum_{n_a} P_{\mathbf{t}_c^a(\mathbf{s}),n_a}^a P_{\mathbf{t}_c^b(\mathbf{s}),n-n_a}^b.$$

341 We now partition  $\mathcal{A}^m$  into four sets  $\mathcal{S}_0^c(m)$ ,  $\mathcal{S}_1^c(m)$ ,  $\mathcal{S}_2^c(m)$  and  $\mathcal{S}_3^c(m)$ :

- 342 ■  $\mathbf{s} \in \mathcal{S}_0^c(m)$ : if neither of the initial symbol  $c$  or the final symbol of  $\mathbf{s}$ , namely  $c_m$  is identical  
 343 to  $a$ . Thus the total number of tail symbols equal to  $a$ , namely  $|\mathbf{s}|_a$  is equal to  $|\mathbf{t}_c^a(\mathbf{s})|$ ;
- 344 ■  $\mathbf{s} \in \mathcal{S}_1^c(m)$ : if both the final symbol and  $c$  are equal to  $a$  so that the total number of tail  
 345 (and initial) symbols equal to  $a$  is  $|\mathbf{t}_c^a(\mathbf{s})|$ .
- 346 ■  $\mathbf{s} \in \mathcal{S}_2^c(m)$ : if  $c = a$  but  $c_m \neq a$  so that the number of tail symbols equal to  $a$  is  $|\mathbf{t}_c^a(\mathbf{s})| - 1$ .
- 347 ■  $\mathbf{s} \in \mathcal{S}_3^c(m)$ : if  $c \neq a$  but the final symbol  $c_m = a$ . Thus the number of tail symbols equal  
 348 to  $a$  is  $|\mathbf{t}_c^a(\mathbf{s})| + 1$ .

Regrouping we have

$$\mathcal{P}_{m,n}^c = \sum_{\mathbf{s} \in \mathcal{S}_0^c(m) \cup \mathcal{S}_1^c(m)} \mathcal{P}_{\mathbf{s},n}^c + \sum_{\mathbf{s} \in \mathcal{S}_2^c(m)} \mathcal{P}_{\mathbf{s},n}^c + \sum_{\mathbf{s} \in \mathcal{S}_3^c(m)} \mathcal{P}_{\mathbf{s},n}^c.$$

349 Now we have to deal with the right hand side of (18), that is, with the DST model. Let  
 350  $\mathcal{T}_1(m)$  be the set of pairs of arbitrary sequences denoted as  $(\mathbf{t}^a, \mathbf{t}^b)$  such that  $|\mathbf{t}^a| + |\mathbf{t}^b| = m$   
 351 and  $|\mathbf{t}^a|_a + |\mathbf{t}^b|_a = |\mathbf{t}^a|$ . We notice that for  $\mathbf{s} \in \mathcal{S}_1^c(m) \cup \mathcal{S}_2^c(m)$ :  $(\mathbf{t}_c^a(\mathbf{s}), \mathbf{t}_c^b(\mathbf{s})) \in \mathcal{T}_1(m)$ , hence

$$352 \quad \sum_{\mathbf{s} \in \mathcal{S}_0^c(m) \cup \mathcal{S}_1^c(m)} \mathcal{P}_{\mathbf{s},n}^c = \sum_{n_a} \sum_{\mathbf{s} \in \mathcal{S}_0^c(m) \cup \mathcal{S}_1^c(m)} P_{\mathbf{t}_c^a(\mathbf{s}),n_a}^a P_{\mathbf{t}_c^b(\mathbf{s}),n-n_a}^b \leq \sum_{n_a} \sum_{(\mathbf{t}^a, \mathbf{t}^b) \in \mathcal{T}_1(m)} P_{\mathbf{t}^a,n_a}^a P_{\mathbf{t}^b,n-n_a}^b.$$

353 Notice that we have an upper bound, since for some pair  $(\mathbf{t}^a, \mathbf{t}^b)$  in  $\mathcal{T}_1^c(m)$  there may not  
 354 exist  $\mathbf{s} \in \mathcal{S}_1^c(m) \cup \mathcal{S}_2^c(m)$  such that  $\mathbf{t}^a = \mathbf{t}_c^a(\mathbf{s})$  and  $\mathbf{t}^b = \mathbf{t}_c^b(\mathbf{s})$ . For example, let  $c = a$  and for  
 355  $m = 4$  we set  $\mathbf{t}^a = (a, b)$  and  $\mathbf{t}^b = (b, a)$ , so that  $|\mathbf{t}^a|_a + |\mathbf{t}^b|_a = |\mathbf{t}^a|$  but it is impossible to  
 356 find  $\mathbf{s}$  such that  $(\mathbf{t}_c^a(\mathbf{s}), \mathbf{t}_c^b(\mathbf{s})) = (\mathbf{t}^a, \mathbf{t}^b)$ .

Thanks to (16) we have  $\sum_{\mathbf{t}: |\mathbf{t}|=m, |\mathbf{t}|_a=k} P_{\mathbf{t},n}^c = P_{m,k,n}^c$  leading to

$$\sum_{(\mathbf{t}^a, \mathbf{t}^b) \in \mathcal{T}_1(m)} \sum_{n_a} P_{\mathbf{t}^a,n_a}^a P_{\mathbf{t}^b,n-n_a}^b = \sum_{m_a,k} P_{m_a,k,n_a}^a P_{m-m_a,m_a-k,n-n_a}^b.$$

This proves the first term in the right hand side of (18). To prove the other two terms  
 we introduce  $\mathcal{T}_2(m)$  as the set of pairs of sequence  $(\mathbf{t}^a, \mathbf{t}^b)$  such that  $|\mathbf{t}^a| + |\mathbf{t}^b| = m$  and  
 $|\mathbf{t}^a|_a + |\mathbf{t}^b|_a = |\mathbf{t}^a| - 1$ . In this case

$$\sum_{\mathbf{s} \in \mathcal{S}_2^c(m)} \mathcal{P}_{\mathbf{s},n}^c \leq \sum_{n_a} \sum_{(\mathbf{t}^a, \mathbf{t}^b) \in \mathcal{T}_2(m)} P_{\mathbf{t}^a,n_a}^a P_{\mathbf{t}^b,n-n_a}^b,$$

**XX:12 Lempel-Ziv'78 for Markov Sources**

357 and the second term of (18) is proved. And finally with  $\mathcal{T}_3(m)$  as the set of pairs of sequence  
 358  $(\mathbf{t}^a, \mathbf{t}^b)$  such that  $|\mathbf{t}^a| + |\mathbf{t}^b| = m$  and  $|\mathbf{t}^a|_a + |\mathbf{t}^b|_a = |\mathbf{t}^a| + 1$ , we establish the third term of  
 359 (18).  $\blacktriangleleft$

To finish the proof of Theorem 5 we now use previous lemmas to upper bound  $\mathcal{P}_{m,n}$ . Let  
 $\mathcal{P}_{m,n}^c \leq K_{m,n}^c(0) + K_{m,n}^c(1) + K_{m,n}^c(-1)$  with

$$K_{m,n}^c(i) = \sum_{m_a} \sum_{n_a} \sum_k P_{m_a, k, n_a}^a P_{m-m_a, m_a-k-i, n-n_a}^b.$$

To simplify our presentation we only study  $K_{m,n}^c(0)$ . First, we rewrite the bound in Theorem 4  
 for the DST model as follows: for  $\delta > 1/2$  there exist  $B$  and  $C$  strictly positive such that

$$P_{m,k,n}^c \leq B \exp[-Cm^{-\delta}|k - \mathbf{E}[T_m^c]| - Cm^{-\delta}|n - \mathbf{E}[L_m^c]|].$$

360 Thus

$$361 \quad K_{m,n}^c(0) \leq \sum_{m_a+m_b=m} \sum_{k \leq m_a} \sum_{n_a+n_b=n} B^2 \exp[-Cm_a^{-\delta}|k - \mathbf{E}[T_{m_a}^c]| - Cm_a^{-\delta}|n_a - \mathbf{E}[L_{m_a}^a]| \\ 362 \quad Cm_b^{-\delta}|m_a - k - \mathbf{E}[T_{m_b}^b]| - Cm_b^{-\delta}|n_b - \mathbf{E}[L_{m_b}^b]|].$$

From here we use  $m_a, m_b \leq m$  to find

$$\begin{aligned} Cm_a^{-\delta}|k - \mathbf{E}[T_{m_a}^c]| + Cm_a^{-\delta}|n_a - \mathbf{E}[L_{m_a}^a]| + Cm_b^{-\delta}|m_a - k - \mathbf{E}[T_{m_b}^b]| + Cm_b^{-\delta}|n_b - \mathbf{E}[L_{m_b}^b]| &\geq \\ Cm^{-\delta}|k - \mathbf{E}[T_{m_a}^c]| + Cm^{-\delta}|n_a - \mathbf{E}[L_{m_a}^a]| + Cm^{-\delta}|m_a - k - \mathbf{E}[T_{m_b}^b]| + Cm^{-\delta}|n_b - \mathbf{E}[L_{m_b}^b]| & \\ \geq Cm^{-\delta}|m_a - \mathbf{E}[T_{m_a}^a]| - \mathbf{E}[T_{m_b}^b]| + Cm^{-\delta}|n - \mathbf{E}[L_{m_a}^a]| - \mathbf{E}[L_{m_b}^b]|. & \end{aligned}$$

363 Replacing the  $\mathbf{E}[T_m^c]$  by  $\tau_c(m)m$  and  $\mathbf{E}[L_m^c]$  by  $m \log m/h + m + m\mu_c(m)$  we arrive at

$$364 \quad K_{m,n}^c(0) \leq B^2 m \sum_{m_a+m_b=m} \exp(-Cm^{-\delta}|m_a - m_a\tau_c(m_a) - m_b\tau_c(m_b)|) \\ 365 \quad \times \exp(-Cm^{-\delta}|n - m \log m/h + m(H(m_a/m)/h - 1) - m_a\mu_c(m_a) - m_b\mu_c(m_b)|).$$

366 Without changing the order of magnitude we further can replace  $\tau_c(m)$  by  $\tau(m)$  and  $\mu_c(m)$   
 367 by  $\mu(m)$ .

We now focus only on the aperiodic case and set  $\tau(m) = \bar{\tau}m$  and  $\mu(m) = \bar{\mu}m$ . (We know  
 that even in this case for small values of  $m$ , the  $\mu(m)$  and  $\tau(m)$  are not exactly linear in  $m$ ,  
 but we handle it later.) Thus our term  $K_{m,n}^c(0)$  is bounded by

$$K_{m,n}^c(0) \leq B^2 m \sum_{m_a \leq m} \exp[-Cm^{-\delta}|m_a - \bar{\tau}m|] \exp[-Cm^{-\delta}|n - m \log m/h - \bar{\mu}m + m(H(m_a/m)/h - 1)|].$$

If we take any  $\delta' > \delta$  we find

$$\begin{aligned} K_{m,n}^c(0) &\leq B^2 m \sum_{m_a \leq m} \exp[-Cm^{-\delta}|m_a - \bar{\tau}m|] \\ &\times \exp[-Cm^{-\delta'}|n - m \log m/h - \bar{\mu}m + m(H(m_a/m)/h - 1)|]. \end{aligned}$$

We observe that  $\exp[-Cm^{-\delta}|m_a - \bar{\tau}m|]$  attains its maximum at  $m_a = m^* = \bar{\tau}m$ . Thus

$$K_{m,n}^c(0) \leq B^2 \sum_{m_a \leq m^*} e^{Cm^{-\delta}(m-m^*)} \times \exp[-Cm^{-\delta'}|n - m \log m/h - \bar{\mu}m + m(H(m_a/m)/h - 1)|]$$

$$+B^2 \sum_{m_a \geq m^*} e^{Cm^{-\delta}(m^*-m)} \times \exp[-Cm^{-\delta'}|n - m \log m/h - \bar{\mu}m + m(H(m_a/m)/h - 1)|].$$

Notice that the terms  $e^{Cm^{-\delta}(m-m^*)}$  and  $e^{Cm^{-\delta}(m^*-m)}$  form a geometrically decreasing series with rate  $e^{-Cm^{-\delta}}$ . Since  $|mH((m_a+1)/m) - mH(m_a/m)| \leq \log m$ , the term

$$\exp[-Cm^{-\delta'}|n - m \log n/h - \bar{\mu}m + m(H(m_a/m)/h - 1)|]$$

is at most geometrically increasing with a rate  $e^{m^{-\delta'} \log m/h}$  which is smaller than  $e^{Cm^{-\delta}}$ . Therefore, the whole series has its maximum at  $m_a = m^*$  and

$$\begin{aligned} K_{m,n}^c(0) &\leq 2B^2 \sum_{k=0}^{\infty} e^{-Ck(m^{-\delta} - \log m/hm^{-\delta'})} \\ &\times \exp[-Cm^{-\delta'}|n - m \log n/h - \bar{\mu}m + m(H(m^*/m)/h - 1)|] \\ &= \frac{2B^2}{1 - e^{-(m^{-\delta} - \log m/hm^{-\delta'})C}} \\ &\times \exp[-Cm^{-\delta'}|n - m \log m/h - \bar{\mu}m + m(H(m^*/m)/h - 1)|] \\ &= O(2B^2 m^\delta) \exp[-Cm^{-\delta'}|n - m \log m/h - \bar{\mu}m + m(H(\bar{\tau})/h - 1)|]. \end{aligned}$$

Including all contributions, the final estimate for some  $B' > 0$  is

$$\mathcal{P}_{m,n}^c \leq B' m^{1+\delta} \exp[-Cm^{-\delta}|n - m \log m - \bar{\mu}m + m(H(\bar{\tau})/h - 1)|].$$

368 This gives the large deviation estimate and  $\mathbf{E}[\mathcal{L}_{m,n}^c] = m \log m/h + \bar{\mu}m - m(H(\bar{\tau})/h - 1) +$   
369  $O(m^\delta)$  by Fact 1. We recognize in  $H(\bar{\tau})$  the entropy of the tail symbol.

In fact the quantities  $\tau(m)$  and  $\mu(m)$  are not exactly  $\bar{\tau}m$  and  $m\bar{\mu}$ . To handle it we observe that due to their slowly varying properties, the function  $\exp(-Cm^{-\delta}|m_a - \tau(m_a)m_a - \tau(m - m_b)(m - m_a)|)$  attains the maximum for  $m^*$  such that

$$m^* = -\tau_a(m^*)m^* - \tau_b(m^*)(m - m^*).$$

370 Indeed the function  $m_a - \mathbf{E}[T_{m_a}^a] - \mathbf{E}[T_{m_b}^b]$  is a strictly increasing thus this value is unique.  
371 Then again  $\mathbf{E}[\mathcal{L}_m^c] = m \log m/h + m^* \mu(m^*) + (m - m^*)\mu(m - m^*) - m(H(m^*/m)/h - 1)$ ,  
372 and therefore  $\mathbf{E}[\mathcal{L}_m^c] + mH(m^*/m) + o(m)$ . The latter is equal to  $\mathbf{E}[\mathcal{L}_m^c] + mH(\bar{\tau}) + o(m)$   
373 in the aperiodic case. To complete the proof of Theorem 5 we just use Fact 1 applied to  $\mathcal{L}_m$ .

## 374 5 Conclusions

375 In this paper we analyze the Lempel-Ziv'78 algorithm for binary Markov sources, a problem  
376 left open since the algorithm inception. To handle the strong dependency between Markov  
377 phrases, we introduce and precisely analyze the so called tail symbol which is the first symbol  
378 of the next phrase in the LZ'78 parsing. We focus here on the large deviations for the number  
379 of phrases in the LZ78 and also give a precise asymptotic expression for the redundancy which  
380 is the excess of LZ78 code over the entropy of the source. In future work we plan to extend  
381 our analysis to non-binary Markov sources and present some bounds on the central limit  
382 theorem. Furthermore, we shall study LZ78 for Markov sources of higher order, however, this  
383 will require a new approach to the tail symbols which may span over consecutive phrases.

## 384 Acknowledgment

385 We thank Guillaume Duboc for simulation of LZ78 scheme resulting in Figure 2.

## 386 References

- 387 1 D. Aldous, and P. Shields, A Diffusion Limit for a Class of Random-Growing Binary  
388 Trees, *Probab. Th. Rel. Fields*, 79, 509–542, 1988.
- 389 2 J. Fayolle, M. D. Ward, Analysis of the average depth in a suffix tree under a Markov  
390 model, in *International Conference on Analysis of Algorithms DMTCs*, 95, 95-104, 2005.
- 391 3 P. Jacquet and W. Szpankowski, Asymptotic behavior of the Lempel-Ziv parsing scheme  
392 and digital search trees, *Theoretical Computer Science*, 144, 161–197, 1995.
- 393 4 P. Jacquet, W. Szpankowski, Analytical depoissonization and its applications. *Theoretical  
394 Computer Science*, 201(1), 1-62, 1998.
- 395 5 P. Jacquet and W. Szpankowski, On the Limiting Distribution of Lempel Ziv’78 Re-  
396 dundancy for Memoryless Sources, *IEEE Trans. Information Theory*, 60, 6917-6930,  
397 2014.
- 398 6 P. Jacquet, W. Szpankowski, and J. Tang, Average Profile of the Lempel-Ziv Parsing  
399 Scheme for a Markovian Source, *Algorithmica*, 31, 318-360, 2001.
- 400 7 P. Jacquet, W. Szpankowski, *Analytic Pattern Matching: From DNA to Twitter*. Cam-  
401 bridge University Press, Cambridge, 2015.
- 402 8 D. E. Knuth, *The Art of Computer Programming. Sorting and Searching*, Vol. 3, Second  
403 Edition, Addison-Wesley, Reading, MA, 1998.
- 404 9 K. Leckey, R. Neininger and W. Szpankowski, Towards More Realistic Probabilistic  
405 Models for Data Structures: The External Path Length in Tries under the Markov Model,  
406 SODA 2013, 877-886, New Orleans, 2013.
- 407 10 G Louchard, W Szpankowski, On the average redundancy rate of the Lempel-Ziv code.  
408 *IEEE Transactions on Information Theory*, 43, 2–8, 1997.
- 409 11 N. Merhav, Universal Coding with Minimum Probability of Codeword Length Overflow,  
410 *IEEE Trans. Information Theory*, 37, 556–563, 1991.
- 411 12 N. Merhav, and J. Ziv, On the Amount of Statistical Side Information Required for Lossy  
412 Data Compression, *IEEE Trans. Information Theory*,
- 413 13 N. Merhav and W. Szpankowski, Average Redundancy of the Shannon Code for Markov  
414 Sources, *IEEE Trans. Information Theory*, 59, 7186-7193, 2013.
- 415 14 R. Neininger and L. Rüschemdorf, A General Limit Theorem for Recursive Algorithms  
416 and Combinatorial Structures, *The Annals of Applied Probability*, 14, No. 1, 378-418,  
417 2004.
- 418 15 S. Savari, Redundancy of the Lempel-Ziv Incremental Parsing Rule, *IEEE Trans. In-  
419 formation Theory*, 43, 9–21, 1997.
- 420 16 W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley, 2001.
- 421 17 J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding,  
422 *IEEE Transactions on Information Theory*, 24, 530–536, 1978.

423 **Appendix: Proofs of Theorem 3(i)-(ii)**424 **A Proof of Theorem 3(i): Mean**

425 We first analyze asymptotically  $\mathbf{X}(z) = (X_a(z), X_b(z))$  that satisfies the system of differential-  
426 functional equations (6). We solve this system, and then apply Mellin transform and  
427 dePoissonization to prove Theorem 3(i).

428 Since for all integer  $m$ , we have  $T_m^c \leq m$ , we notice that the function  $X_c(z)$  is  $O(z)$  both  
429 when  $z \rightarrow \infty$  and when  $z \rightarrow 0$ . Thus the function  $\mathbf{X}(z)$  has no Mellin transform defined  
430 as  $X_c(s) = \int_0^\infty X_c(z)z^{s-1}dz$  (see [16] for more on the Mellin transform). To correct this  
431 we introduce  $\tilde{X}_c(z) = X_c(z) - G_c(z)$  with  $G_c(z) = (\mathbf{E}[T_1^c]z + \mathbf{E}[T_2^c]z^2/2)e^{-z}$  which is  $O(z^3)$   
432 when  $z \rightarrow 0$ , where  $\mathbf{E}[T_1^c]$  and  $\mathbf{E}[T_2^c]$  are defined in (7).

433 The Mellin transform  $X_c^*(s)$  of  $\tilde{X}_c(z)$  on the strip  $\Re(s) \in ]-3, -1[$  exists. The Mellin  
434 transform of  $\partial_z \tilde{X}_c(z)$  exists too on the strip  $\Re(s) \in ]-2, 0[$ . Thus the two Mellin transforms  
435 coexist on the strip  $\Re(s) \in ]-2, -1[$  and satisfies [16]

$$436 -(s-1)(X_c^*(s-1) + G_c^*(s)) + X_c^*(s) + G_c^*(s) = P(a|c)^{-s}(X_a^*(s) + G_a^*(s)) + P(b|c)^{-s}(X_b^*(s) + G_b^*(s))$$

437 where  $G_c^*(s)$  for  $c \in \mathcal{A}$  is the Mellin transform of  $G_c(z)$  and has the explicit expression  
438  $\mathbf{E}[T_1^c]\Gamma(1+s) + \mathbf{E}[T_2^c]\Gamma(s+2)/2$ . This expression is here for completeness.

439 An alternative but convenient way to see this equations is to consider the vector  $\mathbf{X}^*(s)$   
440 made of the quantities  $X_c^*(s)$ ,  $c \in \mathcal{A}$  which is also the Mellin transform of the vector  $\tilde{\mathbf{X}}(z)$   
441 made of the coefficients  $\tilde{X}_c(z)$ . This yields the linear equation

$$442 -(s-1)(\mathbf{X}^*(s-1) + \mathbf{G}^*(s-1)) + \mathbf{X}^*(s) + \mathbf{G}^*(s) = \mathbf{P}(s)(\mathbf{X}^*(s) + \mathbf{G}^*(s)) \quad (\text{A.1})$$

where  $\mathbf{G}^*(s)$  is the vector of the  $G_c^*(s)$ . It can be rewritten in

$$(s-1)(\mathbf{X}^*(s-1) + \mathbf{G}^*(s-1)) = (\mathbf{I} - \mathbf{P}(s))(\mathbf{X}^*(s) + \mathbf{G}^*(s)).$$

This kind of equation has been studied in [6] where we introduce a new function  $\mathbf{x}(s)$

$$\mathbf{X}^*(s) + \mathbf{G}^*(s) = \Gamma(s)\mathbf{x}(s).$$

443 Thus the equation becomes  $\mathbf{x}(s-1) = (\mathbf{I} - \mathbf{P}(s))\mathbf{x}(s)$ , which leads to  $\mathbf{x}(s) = \prod_{i \geq 0} (\mathbf{I} - \mathbf{P}(s -$   
444  $i))^{-1}\mathbf{K}$  where  $\mathbf{K}$  is a constant vector. Notice that the matrices very likely don't commute  
445 thus the product order is specified from the left to right. Indeed we have

$$446 \mathbf{K} = \left( \prod_{j \geq 2} (\mathbf{I} - \mathbf{P}(-j))^{-1} \right)^{-1} \mathbf{x}(-2) = \prod_{j=-\infty}^{j=2} (\mathbf{I} - \mathbf{P}(j))\mathbf{x}(-2). \quad (\text{A.2})$$

447 To handle it we need an explicit formula for  $\mathbf{x}(-2)$ . The following lemma from [6] is  
448 useful in this regard. We provide a proof for completeness.

449 **► Lemma 12.** *Let  $\{f_n\}_{n=0}^\infty$  be a sequence of real numbers having the Poisson transform*

$$450 \tilde{F}(z) = \sum_{n=0}^{\infty} \tilde{f}_n \frac{z^n}{n!} e^{-z} := \sum_{n=0}^{\infty} f_n \frac{z^n}{n!}, \quad (\text{A.3})$$

which is an entire function. Furthermore, let its Mellin transform  $F(s)$  have the following factorization

$$F(s) = \mathcal{M}[\tilde{F}(z); s] = \Gamma(s)\gamma(s).$$

451 Assume that  $F(s)$  exists for  $\Re(s) \in (-2, -1)$ , and that  $\gamma(s)$  is analytic for  $\Re(s) \in (-\infty, -1)$ .  
 452 Then

$$453 \quad \gamma(-n) = \sum_{k=0}^n \binom{n}{k} (-1)^k \tilde{f}_k = (-1)^n f_n, \quad \text{for } n \geq 2. \quad (\text{A.4})$$

**Proof.** Notice that  $f_n$  and  $\tilde{f}_n$  are related by [16]

$$\tilde{f}_n = \sum_{k=0}^n \binom{n}{k} (-1)^{n-k} f_k, \quad n \geq 0.$$

Define for some fixed  $M \geq 2$ , the function  $\tilde{F}_M(z) = \sum_{n=0}^{M-1} f_n \frac{z^n}{n!}$ . Due to our assumptions, we can continue  $F(s)$  analytically to the whole complex plane except  $s = -2, -3, \dots$ . In particular, for  $\Re(s) \in (-M, -M+1)$  we have  $F(s) = \mathcal{M}[\tilde{F}(z) - \tilde{F}_M(z); s]$ . As  $s \rightarrow -M$ , due to the factorization  $F(s) = \Gamma(s)\gamma(s)$ , we have

$$F(s) = \frac{1}{s+M} \frac{(-1)^M}{M!} \gamma(-M) + O(1);$$

454 thus by the inverse Mellin transform, we have

$$455 \quad \tilde{F}(z) - \tilde{F}_M(z) = \frac{(-1)^M}{M!} \gamma(-M) z^M + O(z^{M+1}) \quad \text{as } z \rightarrow 0. \quad (\text{A.5})$$

456 But

$$457 \quad \tilde{F}(z) - \tilde{F}_M(z) = \sum_{i=M}^{\infty} f_i \frac{z^i}{i!} = f_M \frac{z^M}{M!} + O(z^{M+1}). \quad (\text{A.6})$$

458 Comparing (A.5) and (A.6) shows that  $\gamma(-M) = (-1)^M f_M = \sum_{k=0}^M \binom{M}{k} (-1)^k \tilde{f}_k$ . ◀

459 Now we can compute  $\mathbf{x}(-2)$  using above and (7) leading to

$$460 \quad \mathbf{x}(-2) = \begin{bmatrix} T_2^a - 2P(a|a) \\ T_2^b - 2P(a|b) \end{bmatrix}. \quad (\text{A.7})$$

461 In another notation  $\mathbf{x}(-2) = (\mathbf{P}^2 - \mathbf{P})\mathbf{e}_a$ , where  $\mathbf{e}_a$  is the vector made of a single 1 at  $a$   
 462 position and zero otherwise.

Next, we notice that the vector

$$\Gamma(s) \prod_{i \geq 0} (\mathbf{I} - \mathbf{P}(s-i))^{-1} \prod_{j=-\infty}^{j=-2} (\mathbf{I} - \mathbf{P}(j)) \mathbf{x}(-2)$$

may have a double pole on  $s = -1$  since  $\Gamma(s)$  has a pole and also  $(\mathbf{I} - \mathbf{P}(s))^{-1}$  since  $\mathbf{I} - \mathbf{P}(-1) = \mathbf{I} - \mathbf{P}$  is singular. But in fact the pole multiplicity is reduced by one, as prove below. Let us also define

$$\mathbf{Q}(s) = \prod_{i \geq 1} (\mathbf{I} - \mathbf{P}(s-i))^{-1} \prod_{j=-\infty}^{j=-2} (\mathbf{I} - \mathbf{P}(j)).$$

463 Then  $\mathbf{x}(s) = (\mathbf{I} - \mathbf{P}(s))^{-1} \mathbf{Q}(s) \mathbf{x}(-2)$ .

464 We notice that when  $s \rightarrow -1$ , then  $\mathbf{Q}(s) = \mathbf{I} + (s+1)\mathbf{Q}'(-1) + O((s+1)^2)$ . Furthermore  
 465 let  $\lambda(s)$  be the main eigenvalue of matrix  $\mathbf{P}(s)$  and  $\mathbf{1}(s)$  and  $\boldsymbol{\pi}(s)$  be respectively the right



466 and left main eigenvectors. We have  $\lambda(-1) = 1$ ,  $\mathbf{1}(-1) = \mathbf{1}$  is all made of one's, and  $\boldsymbol{\pi}(-1)$   
467 is the stationary distribution of the Markov source.

468 From the matrix spectral representation [16] we have

$$469 \quad \mathbf{P}(s) = \lambda(s)\mathbf{1}(s) \otimes \boldsymbol{\pi}(s) + \mathbf{R}(s) = \lambda(s)\boldsymbol{\Pi}(s) + \mathbf{R}(s) \quad (\text{A.8})$$

470 where  $\mathbf{R}(s)$  is the automorphism of the eigenplan orthogonal to the main eigenvector and  
471  $\boldsymbol{\Pi}(s) = \mathbf{1}(s) \otimes \boldsymbol{\pi}(s)$  where  $\otimes$  is the tensor product. Note that  $\boldsymbol{\Pi} \cdot \mathbf{P} = \mathbf{P} \cdot \boldsymbol{\Pi} = \boldsymbol{\Pi}$ . Then

$$472 \quad (\mathbf{I} - \mathbf{P}(s))^{-1} = \frac{1}{1 - \lambda(s)} \mathbf{1}(-s) \otimes \boldsymbol{\pi}(s) \\ 473 \quad - \frac{1}{\lambda'(-1)} (\mathbf{1}'(-1) \otimes \boldsymbol{\pi}(-1) + \mathbf{1} \otimes \boldsymbol{\pi}'(-1)) + \mathbf{R}(-1)^{-1} + O(s+1).$$

474 Finally

$$475 \quad (\mathbf{I} - \mathbf{P}(s))^{-1} \mathbf{Q}(s) \mathbf{x}(-2) = \frac{\mathbf{1} \otimes \boldsymbol{\pi}(s) (\mathbf{I} - \mathbf{P}) \mathbf{e}_a}{1 - \lambda(s)} - \frac{1}{\lambda'(-1)} (\mathbf{1}'(-1) \otimes \boldsymbol{\pi} + \mathbf{1} \otimes \boldsymbol{\pi}'(-1)) \\ 476 \quad + \mathbf{R}^{-1}(-1) + \frac{(s+1)}{1 - \lambda(s)} \mathbf{1} \otimes \mathbf{Q}'(-1) + O(s+1).$$

Since

$$\frac{s+1}{1 - \lambda(s)} \rightarrow -\frac{1}{\lambda'(-1)}$$

477 when  $s \rightarrow -1$ , and  $\boldsymbol{\Pi} \mathbf{P} (\mathbf{I} - \mathbf{P}) \mathbf{e}_a = (\boldsymbol{\Pi} - \boldsymbol{\Pi}) \mathbf{e}_a = 0$ . Also

$$478 \quad \mathbf{R}^{-1}(-1) (\mathbf{I} - \mathbf{P}) \mathbf{P} \mathbf{e}_a = \mathbf{P} \mathbf{e}_a - \langle \boldsymbol{\pi} \mathbf{P} \mathbf{e}_a \rangle \mathbf{1} = \mathbf{P} \mathbf{e}_a - \langle \boldsymbol{\pi} \mathbf{e}_a \rangle \mathbf{1}. \quad (\text{A.9})$$

479 We finally have

$$480 \quad \lim_{s \rightarrow -1} \mathbf{x}(s) = \mathbf{P} \mathbf{e}_a - \pi_a \mathbf{1} - \frac{1}{\lambda'(-1)} \mathbf{1} \langle (\boldsymbol{\pi}'(-1) + \boldsymbol{\pi} \mathbf{Q}'(-1)) (\mathbf{I} - \mathbf{P}) \mathbf{P} \mathbf{e}_a \rangle, \quad (\text{A.10})$$

481 where  $\pi_a$  is the coefficient of the stationary distribution  $\boldsymbol{\pi}$  at symbol  $a$ .

482 Now we are in position to establish asymptotics of  $X_c(z)$  for large  $z$  and through  
483 depoissonization asymptotics of  $\mathbf{E}[T_m^c]$ . The inverse Mellin transform is

$$484 \quad \tilde{X}_c(z) = \frac{1}{2i\pi} \int_{x-i\infty}^{x+i\infty} X_c^*(s) z^{-s} ds \quad (\text{A.11})$$

485 valid for all  $x \in ]-2, -1[$ . Remembering that  $T_c(z) = \tilde{X}_c(z) + P(a|c)z$  we have indeed

$$486 \quad \tilde{\mathbf{X}}(z) = \frac{1}{2i\pi} \int_{x-i\infty}^{x+i\infty} \Gamma(s) \mathbf{x}(s) z^{-s} ds - \frac{1}{2i\pi} \int_{x-i\infty}^{x+i\infty} \mathbf{G}^*(s) z^{-s} ds \quad (\text{A.12})$$

487 We know that  $\mathbf{T}(z) - \tilde{\mathbf{X}}(z)$  is decaying exponentially fast when  $z \rightarrow \infty$ .

Moving the line of integration toward the right, we meet a single pole at  $s = -1$  of  
488  $\mathbf{G}^*(s) z^{-s}$  and its residues is  $-z \mathbf{P} \mathbf{e}_a$ . Then

$$\frac{1}{2i\pi} \int_{x-i\infty}^{x+i \inf ty} \mathbf{G}^*(s) z^{-s} ds = -\mathbf{P} \mathbf{e}_a + O(z^{-M})$$

488 for all  $M > 0$ .

489 The value  $-1$  is also a simple pole for  $z^{-s} \Gamma(s) \mathbf{x}(s)$ . We know that its residue is

$$490 \quad -z \left( \mathbf{P} \mathbf{e}_a - \pi_a \mathbf{1} - \frac{1}{\lambda'(-1)} \mathbf{1} \langle (\boldsymbol{\pi}'(-1) + \boldsymbol{\pi} \mathbf{Q}'(-1)) (\mathbf{I} - \mathbf{P}) \mathbf{P} \mathbf{e}_a \rangle \right). \quad (\text{A.13})$$

491 Therefore we have

$$492 \quad \mathbf{X}(z) = z \left( \pi_a + \frac{1}{\lambda'(-1)} \mathbf{1} \langle (\pi'(-1) + \pi \mathbf{Q}'(-1)) (\mathbf{I} - \mathbf{P}) \mathbf{P} \mathbf{e}_a \rangle \right) \mathbf{1} + o(z). \quad (\text{A.14})$$

493 For irrational case, we know that  $s = -1$  is the only pole on the line  $\Re(s) = -1$ , leading to  
 494 the error term  $o(z)$  coming from other poles of  $(\mathbf{I} - \mathbf{P}(s))^{-1}$  which may occur on the right  
 495 half plan of  $s = -1$ .

496 But in the rational case, there is the possibility of other poles regularly spaced on the  
 497 axis  $\Re(s) = -1$  with some specific matrices  $\mathbf{P}$  detailed in [6] where the coefficients  $\alpha_{abc}$  are  
 498 introduced. In these very specific cases (the uniform probability distribution on  $\mathcal{A}$  is one  
 499 of them) the  $o(z)$  term should be replaced by a term  $zQ_c(\log z) + O(z^{1-\epsilon})$ , where  $Q_c$  is a  
 500 periodic vector of very small amplitude and mean zero, and  $\epsilon > 0$  depends on the matrix  $\mathbf{P}$ .  
 501 This proves Theorem 3(i).

## 502 **B Proof of Theorem 3(ii): Variance**

We now analyze asymptotically  $\mathbf{V}(z) = (V_a(z), V_b(z))$  that satisfies the system of differential-  
 functional equations (8). In order to apply depoissonization, for  $\theta \in [0, \pi/2]$  we define  $\mathcal{C}(\theta)$   
 as the complex cone containing the complex number  $z$  such that  $|\arg(z)| \leq \theta$  on increasing  
 domains [16; 5]

$$\mathcal{C}_k(\theta) = \{z, z \in \mathcal{C}(\theta) \& |z| \leq \rho^k\}$$

503 with  $\rho = \min_c \{ \frac{1}{P(a|c)}, \frac{1}{P(b|c)} \}$ .

504 Our first goal is to prove that  $V_c(z) = O(z)$ . We shall use the increasing domain  
 505 approach [16] applied to (8) following the footsteps of the proof of Lemma 7A of [3]. From  
 506 Fact 1 of [3] we conclude that

$$507 \quad V_c(z) = V_c(\rho z) e^{-z(1-\rho)} + e^{-z} \int_{\rho z}^z e^x (V_a(P(a|c)x) + V_b(P(b|c)x) + g(x)) dx \quad (\text{A.15})$$

508 where  $g(z) = P(a|c) - P^2(a|c) + [X_z^c(z)]^2 = O(1)$ . Indeed, it follows from Fact 1 of [3] that  
 509 the differential equation like

$$510 \quad f'(z) = b(z) - a(z)f(z) \quad (\text{A.16})$$

satisfies

$$f(z) = f(z_0) e^{A(z_0) - A(z)} + \int_{z_0}^z b(x) e^{A(x) - A(z)} dx$$

511 where  $A(z) = \int a(z)$  is the primitive function of  $a(z)$ . Setting in (A.16)  $f(z) = V_c(z)$ ,  
 512  $b(z) = V_a(P(a|c)z) + V_b(P(b|c)z) + g(z)$  and  $a(z) = 1$  we obtain (A.15).

513 Now we apply induction over the increasing domains. In short, we assume that for  
 514  $z \in \mathcal{C}_k(\theta)$  we have  $|V_c(z)| \leq B_k |z|$  for some  $B_k$ . Using the induction of the increasing  
 515 domains we prove, as in the Appendix of [3] that  $B_k$  are bounded. This completes the proof,  
 516 after applying the depoissonization lemma of [4].

In order to find a precise estimate of the asymptotic development of  $\mathbf{V}(z)$  we denote  
 $\mathbf{V}^*(s)$  the Mellin transform of  $\mathbf{V}(z)$ . From (8) we arrive at

$$-(s-1)\mathbf{V}^*(s-1) + \mathbf{V}^*(s) = \mathbf{P}(s)\mathbf{V}^*(s) + \mathbf{g}^*(s),$$

where  $\mathbf{g}^*(s)$  is the Mellin transform of the vector made of the coefficients  $(\partial_z X_c(z))^2$ . Let  $\mathbf{V}^*(s) = \Gamma(s)\mathbf{B}(s)$  and  $\mathbf{g}^*(s) = \Gamma(s)\mathbf{G}(s)$ . Then

$$\mathbf{B}(s) = (\mathbf{I} - \mathbf{P}(s))^{-1} (\mathbf{B}(s-1) + \mathbf{G}(s)).$$

517 The quantity  $(\mathbf{I} - \mathbf{P}(s))^{-1}$  has a pole at  $s = -1$ . Together with  $\Gamma(s)$  it would give a double  
518 pole at  $s = -1$  which is not possible, as proved above. Indeed, notice that the coefficient at  
519 the double pole at  $s = 1$  is  $\mathbf{\Pi}(\mathbf{B}(-2) + \mathbf{G}(-1))$ . But  $\mathbf{G}(-1)$  is the the coefficient at  $z$  of  $\mathbf{g}(z)$   
520 and  $\mathbf{B}(-2)$  is the coefficient at  $z^2$  of  $\mathbf{V}(z)$ , as already proved in Lemma 12. Then we easily  
521 see that  $\mathbf{B}(-2) + \mathbf{G}(-1) = \mathbf{P}^2\mathbf{e}_a - \mathbf{P}\mathbf{e}_a$ , and consequently the coefficient at the double pole  
522 at  $s = 1-$  is equal to  $\mathbf{\Pi}(\mathbf{P}^2\mathbf{e}_a - \mathbf{P}\mathbf{e}_a) = (\mathbf{\Pi} - \mathbf{\Pi})\mathbf{e}_a = 0$ , as desired.

523 Therefore, the contribution of pole  $s = -1$  to the asymptotic of  $\mathbf{V}(z)$  is  $\mathbf{B}(-1)$  becomes

$$\begin{aligned} 524 \mathbf{B}(-1) &= \frac{1}{\lambda'(-1)} (\langle \pi'(-1)(\mathbf{B}(-2) + \mathbf{G}(-1)) \rangle + \langle \pi(\mathbf{B}'(-2) + \mathbf{G}'(-1)) \rangle) \mathbf{1} \\ 525 &+ (\mathbf{I} - \mathbf{R}(-1))^{-1} (\mathbf{B}(-2) + \mathbf{G}(-1)). \end{aligned}$$

526 Notice also that  $(\mathbf{I} - \mathbf{R}(-1))^{-1}(\mathbf{P}^2\mathbf{e}_a - \mathbf{P}\mathbf{e}_a) = \langle \pi\mathbf{P}\mathbf{e}_a \rangle \mathbf{1} - \mathbf{P}\mathbf{e}_a = \langle \pi\mathbf{e}_a \rangle \mathbf{1} - \mathbf{P}\mathbf{e}_a$ .

527 The real issue here is how to compute  $\mathbf{B}'(-2)$  and  $\mathbf{G}'(-1)$ , which we address next.

528 ► **Lemma 13.** *Let a function  $g(z) = \sum_{n \geq 1} \frac{a_n}{n!} z^n$  and  $f(z) = g(z)e^{-z} = \sum_{n \geq 1} \frac{b_n}{n!} z^n$ . Let*  
529 *also  $g_k(z) = \sum_{n \leq k} \frac{a_n}{n!} z^n$  and  $f_k(z) = f(z) - g_k(z)e^{-z}$  with  $f_k^*(s)$  being its Mellin transform*  
530 *defined for  $-k-1 < \Re(s) < 0$ . Then*

$$\begin{aligned} 531 \lim_{s \rightarrow -k} \left( \frac{f^*(s)}{\Gamma(s)} \right)' &= f_k^*(-k) \left( \frac{1}{\Gamma(s)} \right)'_{s=-k} + \sum_{n \leq k} \frac{a_n}{n!} \left( s^{(n)} \right)'_{s=-k} \\ 532 &= f_k^*(-k)(-1)^{n-1} n! + \sum_{n \leq k} \frac{a_n}{n!} \left( s^{(n)} \right)'_{s=-k} \end{aligned}$$

533 where  $s^{(n)} = \frac{\Gamma(s+n)}{\Gamma(s)} = (s+n-1) \times \dots \times s$ .

**Proof.** We start with a simple identity

$$\frac{f^*(s) - f_k^*(s)}{\Gamma(s)} = \sum_{n \leq k} \frac{a_n}{n!} s^{(n)}$$

which is easy to derive. But the Mellin transform of  $f_k(z)$  and  $f_k^*(s)$  are defined for  $-k-1 < \Re(s) < 0$ . The derivative of  $f_k^*(s)/\Gamma(s)$  at  $s = -k$  is equal to  $f_k^*(-k) (\Gamma^{-1}(s))'_{s=-k}$  since  $\Gamma^{-1}(-k) = 0$ . Finally we notice that [16]

$$\lim_{s \rightarrow -k} \left( \frac{1}{\Gamma(s)} \right)' = \lim_{s \rightarrow -k} \frac{\Psi(s)}{\Gamma(s)} = \lim_{s \rightarrow -k} \frac{(s+n)\Psi(s)}{(s+n)\Gamma(s)} = (-1)^{n-1} n!$$

534 where  $\Psi(s)$  is the psi function. ◀

535 In absence of specific properties on  $f_k(z)$  there is no other way than numerical computation  
536 to get an estimate of  $f_k^*(-k)$ . Finally, we can present a precise asymptotic expression for the  
537 variance.

538 ► **Theorem 14.** *We have  $\mathbf{V}(z) = \bar{\omega}_a \mathbf{1} z + o(z)$  in the aperiodic case, and in the periodic case*  
539  *$\mathbf{V}(z) = \bar{\omega}_a \mathbf{1} z + Q_2(\log z) z + O(z^{1-\epsilon})$  for some  $\epsilon > 0$  and  $Q_2(\cdot)$  being a periodic function of*  
540 *small amplitude and mean zero, where*

$$541 \bar{\omega}_a = \frac{1}{\lambda'(-1)} (\langle \pi'(-1)((\mathbf{P} - \mathbf{I})\mathbf{P}\mathbf{e}_a) \rangle + \langle \pi(\mathbf{B}'(-2) + \mathbf{G}'(-1)) \rangle) + \langle \pi\mathbf{e}_a \rangle. \quad (\text{A.17})$$

542 Notice that  $\omega = B(-1) + \mathbf{P}\mathbf{e}_a$ .