Proceedings of the 27th International Conference on Probabilistic, Combinatorial and Asymptotic Methods for the Analysis of Algorithms
Kraków, Poland, 4-8 July 2016

Average Size of a Suffix Tree for Markov Sources

Philippe Jacquet[1], Wojciech Szpankowski[2†]

[1]*Bell Labs, Alcatel-Lucent, France.*
[2]*Department of Computer Science, Purdue University, USA*

We study a suffix tree built from a sequence generated by a Markovian source. Such sources are more realistic probabilistic models for text generation, data compression, molecular applications, and so forth. We prove that the average size of such a suffix tree is asymptotically equivalent to the average size of a trie built over n independent sequences from the same Markovian source. This equivalence is only known for memoryless sources. We then derive a formula for the size of a trie under Markovian model to complete the analysis for suffix trees. We accomplish our goal by applying techniques of analytic combinatorics on words also known as analytic pattern matching.

Keywords: Suffix tree, Markov sources, digital trees, size, pattern matching, number of occurrences.

1 Introduction

Suffix trees are the most popular data structures on words. They find myriad of applications in computer science and telecommunications, most notably in algorithms on strings, data compressions (Lempel-Ziv'77 scheme), and codes. Despite this, little is still known about their typical behaviors for general probabilistic models (see [2, 8, 4]).

A suffix tree is a *trie* (a digital tree; see [12]) built from the suffixes of a single string. In Figure 1 we show the suffix tree constructed for the first four suffixes of the string $X = 0101101110$. More precisely, we actually build a suffix tree on the first n *infinite* suffixes of a string X as shown in Figure 1. We shall call it simply a suffix tree which we study in this paper. Such a tree consists of internal (branching) nodes and external nodes storing the suffixes. Our goal is to analyze the number of internal nodes called also the *size* of a suffix tree built from a sequence X generated by a Markov source. We accomplish it by employing powerful techniques of analytic combinatorics on words known also as *analytic pattern matching* [12].

In recent years there has been a resurgence of interest in algorithmic and combinatorial problems on words due to a number of novel applications in computer science, telecommunications, and most notably in molecular biology. The reader is referred to our recent book [12] for more details. Here we present

†W. Szpankowski is also with the Faculty of Electronics, Telecommunications and Informatics, Gdańsk University of Technology, Poland. His work was supported by NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, and NIH Grant 1U01CA198941-01, and the NCN grant, grant UMO-2013/09/B/ST6/02258.

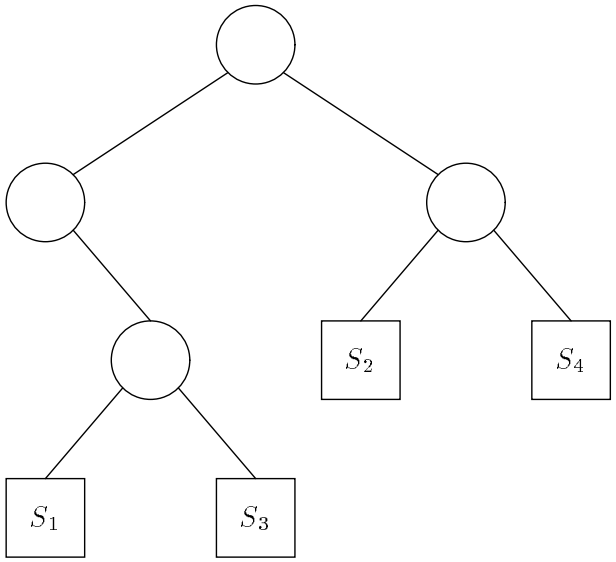

Fig. 1: Suffix tree built from the first five suffixes of $X = 0101101110$, i.e. 0101101110, 101101110, 01101110, 1101110.

only a brief discussion. In computer science and molecular biology many algorithms depend on a solution to the following problem: given a word X and a set of arbitrary $b+1$ suffixes S_1, ... , S_{b+1} of X, what is the longest common prefix of these suffixes. In coding theory (e.g., prefix codes) one asks for the shortest prefix of a suffix S_i which is not a prefix of any other suffixes S_j, $1 \leq j \leq n$ of a given sequence X (cf. [20]). In data compression schemes, the following problem is of prime interest: for a given "data base" sequence of length n, find the longest prefix of the $(n+1)$st suffix S_{n+1} which is not a prefix of any other suffixes S_i ($1 \leq i \leq n$) of the data base sequence. And last but not least, in molecular sequences comparison (e.g., finding homology between DNA sequences), one may search for the longest run of a given motif, a unique sequence, the longest alignment, and the number of common subwords [12]. These, and several other problems on words, can be efficiently solved and analyzed by a clever manipulation of *suffix trees*. In literature other names have been also coined for this structure, and among these we mention here position trees, subword trees, directed acyclic graphs, *etc*.

The extension of suffix tree analysis to Markov sources is quite significant, especially when the suffix tree is used for natural languages. Indeed, Markov sources of finite memory approximate very well realistic texts. For example, the following quote is generated by a memoryless source with the letter statistic of the *Declaration of Independence*:

> esdehTe,a; psseCed vcenseusirh vra f uetaiapgnuev n cosb mgffgfL itbahhr nijue n S ueef,ru
> s,k smodpztrnno.eeteespfg mtet tr i aur oiyr

which should be compared to the following quote generated by a Markov source of order 3 trained on the same text:

> We hat Government of Governments long that their right of abuses are these rights, it, and or themselves and are disposed according Men, der.

In this paper we analyze the average number of internal nodes (size) of a suffix tree built from n (infinite) suffixes of a string generated by a Markov source. We first prove in Theorem 1 that the average size of a suffix tree under Markovian model is asymptotically equivalent to the size of a *trie* that is built from n *independently* generated strings, each string emitted by the corresponding Markovian source. To accomplish this, we study another quantity, namely the number of occurrences of a given pattern w in a string of length n generated by a Markovian source. We use its properties to establish our asymptotic equivalence between suffix trees and tries. Finally, we compare the average size of suffix trees to trie size under Markovian model (see Theorem 2), which – to the best of our knowledge – is only partially known [3].

There is an extensive literature on tries and scarce one on suffix trees [12]. Knuth [13] initiated in early 1970's work on tries for unbiased memoryless sources followed by further extensive and in-depth studies of digital trees for biased memoryless sources [5, 7, 16, 18, 19, 22]. Perhaps the first analysis of tries for Markovian sources was the study of depth presented in [10]. Suffix trees were analyzed only much later, with the initial study of the height in [1] (see also [21]). A rigorous analysis of the depth of suffix tree was first presented in [8] for memoryless sources, and then extended in [4] to Markov sources. In this paper we follow the approach of [4], however, we should point out that depth grows like $O(\log n)$ which makes the analysis manageable. In fact, height and fillup level for suffix tree – which are also of logarithmic growth – were analyzed in [21] (see also [2, 20]). But the average size grows like $O(n)$ and is harder to analyze. For memoryless sources the size was discussed in [16] for tries and in [8] for suffix trees. We also know that some parameters of suffix trees (e.g., profile) cannot be inferred from tries, see [6]. Markov sources add additional level of complications in the analysis of suffix trees as well documented in [2]. In fact, the average size of tries under general dynamic sources was already analyzed in [3], however, specifications to Markov sources requires extra care, especially for the so called rational Markov sources (i.e, periodic case). This is the reason why we have to present a brief proof here of the average size for tries for Markovian model.

This plan for this paper is as follows. In the next section we present our main results, in particular the asymptotic equivalence of random tries and suffix trees (Theorem 1) followed by a precise statement of the average size for tries (see Theorem 2). In the same section we also sketch the main idea of our proofs. Technical lemmas are established in the last section.

2 Main Results

We consider a stationary source generating a sequence of symbols drawn from a finite alphabet $\mathcal{A}$.

We first derive a formula for the average size of a suffix tree in terms of the number of pattern occurrences. Let w be a word over $\mathcal{A}$. We denote by $O_n(w)$ the number of occurrences of word w in a sequence of length n. We observe [8] that the average size s_n of a suffix tree built over a sequence of length n is

$$s_n = \sum_{w \in \mathcal{A}^*} P(O_n(w) \geq 2). \tag{1}$$

irrespective of the underlying probabilistic source. We compare it to the average size t_n of trie built over n *independent* sequences. If $N_n(w)$ is the number of words that begin with w (in a trie built from n words), then we have

$$t_n = \sum_{w \in \mathcal{A}^*} P(N_n(w) \geq 2). \tag{2}$$

Let $P(w)$ be the probability of observing w in a sequence, then $N_n(w)$ is Bernoulli distributed $(n, P(w))$ and the average trie size t_n can be written as

$$t_n = \sum_{w \in \mathcal{A}^*} 1 - (1 - P(w))^n - nP(w)(1 - P(w))^{n-1}. \tag{3}$$

We specifically consider a Markovian source. We assume that the source is stationary and ergodic. We will consider a Markovian process of order 1 with the transition matrix $\mathbf{P} = [P(a|b)]_{a,b \in \mathcal{A}}$. Extensions to higher order Markov is possible since a Markovian source of order r is simply a Markovian source of order 1 over the alphabet $\mathcal{A}^r$. Notice that unlike previous analyses we do not assume that $P(a|b) > 0$ for all $(a, b) \in \mathcal{A}^2$.

Our main result of the paper is formulated next.

Theorem 1 *Consider a suffix tree built over n suffixes of a sequence of length n generated by a Markov source with the transition matrix $\mathbf{P}$. There exists $\varepsilon > 0$ such that*

$$s_n - t_n = O(n^{1-\varepsilon}) \tag{4}$$

for large n.

In order to apply Theorem 1 one needs to estimate the average size of a trie under Markovian model. This seems to be unknown except for some general dynamic sources [3]. In fact, analysis of tries under Markovian sources is quite challenging (see [9]). But we can offer the following result for the average size of a trie under Markovian assumptions. In order to formulate it succinctly we recall that a set of real numbers are *commensurable* (also known as "rationally related") when their ratios are rational numbers. A sketch of the proof is presented in Section 4.

Theorem 2 *Consider a trie built over n independent sequences generated by a Markov source. For $(a, b, c) \in \mathcal{A}^3$ define*

$$\alpha_{abc} = \log \left[\frac{P(a|b)P(c|a)}{P(c|b)} \right]. \tag{5}$$

(i) [Aperiodic case] *If not all $\{\alpha_{abc}\}$ are commensurable, then*

$$t_n = \frac{n}{h} + o(n)$$

where $h = -\sum_{a,b} \pi_a P(b|a) \log P(b|a)$ is the entropy rate of the underlying Markov source with π_a, $a \in \mathcal{A}$, denoting the stationary probability.

(ii) [Periodic case] *If all $\{\alpha_{abc}\}$ are commensurable (rationally related) then*

$$t_n = \frac{n}{h}(1 + Q(n)) + O(n^{1-\varepsilon})$$

where $Q(n)$ is a periodic function and some $\varepsilon > 0$.

Remark We observe that if for all $(a,b) \in \mathcal{A}^2$ the α_{abc} are commensurable (rationally related) for one $c \in \mathcal{A}$, then α_{abc} are commensurable for all values of c. Furthermore in the aperiodic case the $o(n)$ term can have a growth rate arbitrary close to order n, depending on source settings as shown in [5] for the memoryless case.

In the rest of this section, we present a road map of the proof of (4). For this we will make use of ordinary generating functions. Let $w \in \mathcal{A}^k$ be a word of length k. We also define $N_0(z,w) = \sum_{n>0} P(O_n(w)=0)z^n$ and $N_1(z,w) = \sum_{n>0} P(O_n(w)=1)z^n$ for $z \in \mathbb{C}$. We know from [12] that

$$\begin{aligned} N_0(z,w) &= \frac{S_w(z)}{D_w(z)}, \\ N_1(z,w) &= \frac{z^k P(w)}{D_w^2(z)}, \end{aligned}$$

where $S_w(z)$ is the autocorrelation polynomial (see [12] for a definition) of word w and $D_w(z)$ is defined as follows

$$D_w(z) = S_w(z)(1-z) + z^k P(w)\left(1 + F_w(z)(1-z)\right), \tag{6}$$

where $F_w(z)$ is defined below in (7). Notice that for memoryless sources $F_w(z) = 0$. The addition of a non zero $F_w(z)$ significantly changes the analysis. In fact it captures the correlations between characters in the sequence and leads to non trivial developments. Here $F_w(z)$ for $w \in \mathcal{A}^* - \{\varepsilon\}$ is a function that depends on the Markov parameters of the source. It also depends only on the first and last character of w, say respectively a and b for $(a,b) \in \mathcal{A}^2$ as described below.

Let $\mathbf{P}$ be the transition matrix of the Markov source and $\boldsymbol{\pi}$ be its stationary vector with π_a its coefficient at symbol $a \in \mathcal{A}$. The vector $\mathbf{1}$ is the vector with all coefficients equal to 1 and $\mathbf{I}$ is the identity matrix. Assuming that $a \in \mathcal{A}$ (resp. b) is the first (resp. last) symbol of w, we have [17, 12]

$$F_w(z) = \frac{1}{\pi_a}\left[\left(\mathbf{P} - \boldsymbol{\pi}\otimes\mathbf{1}\right)\left(\mathbf{I} - z(\mathbf{P} + \boldsymbol{\pi}\otimes\mathbf{1})\right)^{-1}\right]_{b,a} \tag{7}$$

where $[\mathbf{A}]_{a,b}$ indicates the (a,b) coefficient of the matrix $\mathbf{A}$, and $\otimes$ represents the tensor product. An alternative way to express $F_w(z)$ is

$$F_w(z) = \frac{1}{\pi_a}\langle \mathbf{e}_a, \left(\mathbf{P} - \boldsymbol{\pi}\otimes\mathbf{1}\right)\left(\mathbf{I} - z(\mathbf{P} + \boldsymbol{\pi}\otimes\mathbf{1})\right)^{-1}\mathbf{e}_b\rangle \tag{8}$$

where $\mathbf{e}_c$ for $c \in \mathcal{A}$ is the vector with a 1 at the position corresponding to symbol c and all other coefficients are 0. Here $\langle \mathbf{x}, \mathbf{y}\rangle$ represents the scalar product of $\mathbf{x}$ and $\mathbf{y}$.

Let us define two important quantities:

$$\begin{aligned} d_n(w) &= P(O_n(w)=0) - (1-P(w))^n, \\ q_n(w) &= P(O_n(w)=1) - nP(w)(1-P(w))^{n-1}, \end{aligned}$$

and their corresponding generating functions

$$\begin{aligned} \Delta_w(z) &= \sum_{n>0} d_n(w) z^n \\ Q_w(z) &= \sum_{n>0} q_n(w) z^n. \end{aligned}$$

Observe that $t_n - s_n = \sum_{w\in\mathcal{A}^*} d_n(w) + q_n(w)$. Thus we need to estimate $d_n(w)$ and $q_n(w)$ for all $w \in \mathcal{A}^*$.

Remark The proof presented below – showing that the average size of a suffix tree is close to the average size of a trie – borrows many fundamental elements of the depth analysis from [4] (e.g., we both use $q_n(w)$). But the extension of the depth analysis to the size analysis requires to analyze the new term $d_n(w)$ which has non trivial properties.

We denote by $\mathcal{B}_k$ the set of words of length k that do not overlap with themselves over more than $k/2$ symbols (see [12, 8, 4] for more precise definition). To be precise $w \in \mathcal{A}^k - \mathcal{B}_k$ if there exist $j > k/2$ and $v \in \mathcal{A}^j$ and $(u_1, u_2) \in \mathcal{A}^{k-j}$ such that $w = u_1 v = v u_2$. This set plays fundamental role in the analysis. It is shown in [4] that

$$\sum_{w\in\mathcal{A}^k-\mathcal{B}_k} P(w) = O(\delta^k)$$

where δ is the largest probability in the Markovian transition matrix $\mathbf{P}$. Since the authors of [4] considered strictly positive matrix $\mathbf{P}$, then clearly $\delta < 1$. However, in the present paper we allow transition probabilities to be equal to 1 or 0, as long as the source is ergodic. Therefore δ may be equal to 1. To cope with this minor problem we define

$$\begin{aligned} p &= \exp\left(\lim_k \sup_{w\in\mathcal{A}^k} \frac{\log P(w)}{k}\right) < 1, \\ q &= \exp\left(\lim_k \inf_{w\in\mathcal{A}^k, P(w)\neq 0} \frac{\log P(w)}{k}\right) > 0 \end{aligned}$$

where the inequalities above follow from [15]. These two parameters are related to the Renyí's entropies of order $+\infty$ and $-\infty$ [22].

From now we set $\delta = \sqrt{p} < 1$. The following two crucial lemmas, proved in the next section, imply Theorem 1.

Lemma 1 *There exist $\varepsilon < 1$ such that $\sum_{w\in\mathcal{A}^*} q_n(w) = O(n^\varepsilon)$.*

Lemma 2 *There exists a sequence $R_n(w)$, for $w \in \mathcal{A}^*$ such for all $1 > \varepsilon > 0$ we have*

- *(i) for $w \in \mathcal{B}_k$: $d_n(w) = O((nP(w))^\varepsilon k\delta^k) + R_n(w)$;*
- *(ii) for $w \in \mathcal{A}^k - \mathcal{B}_k$: $d_n(w) = O((nP(w))^\varepsilon) + R_n(w)$,*

where $R_n(w)$ is such that $\sum_{w\in\mathcal{A}^*} R_n(w) = O(1)$.

Remark: The sequence $d_n(w)$ is the main new element in our analysis when compared to [4]. The sequence $R_n(w)$ reflects the impact of the Markovian source on the analysis. In particular, it is a consequence of the introduction of a non zero function $F_w(z)$.

Proof of Theorem 1: We already know via Lemma 1 that there exists $\varepsilon < 1$ such that $\sum_{w\in\mathcal{A}^*} q_n(w) = O(n^\varepsilon)$. Let now $d_n^{(1)} = \sum_k \sum_{w\in\mathcal{B}_k} (d_n(w) - R_n(w))$ and for all $\varepsilon > 0$ observe that

$$d_n^{(1)} = \sum_k \sum_{w\in\mathcal{B}_k} O(n^\varepsilon P^\varepsilon(w) k\delta^k) = \sum_k O(n^\varepsilon k(p^\varepsilon\delta)^k).$$

Hence it converges for all $\varepsilon > 0$. Also let $d_n^{(2)} = \sum_k \sum_{w \in \mathcal{A}^k - \mathcal{B}_k} (d_n(w) - R_n(w))$. Observe that

$$\begin{aligned} d_n^{(2)} &= \sum_k \sum_{w \in \mathcal{A}^k - \mathcal{B}_k} O(n^\varepsilon P^{\varepsilon-1}(w) P(w)) \\ &= \sum_k \sum_{w \in \mathcal{A}^k - \mathcal{B}_k} O(n^\varepsilon q^{(\varepsilon-1)k} P(w)) \\ &= \sum_k O(n^\varepsilon (\delta q^{\varepsilon-1})^k), \end{aligned}$$

which converges for all ε such that $\delta q^{\varepsilon-1} < 1$ (take $\varepsilon < 1$ close enough to 1) and is $O(n^\varepsilon)$. Finally $d_n^{(1)} + d_n^{(2)} + \sum_{w \in \mathcal{A}^*} R_n(w)$ is also $O(n^\varepsilon)$ for $\varepsilon > 0$ since $\sum_{w \in \mathcal{A}^*} R_n(w)$ is finitely bounded. This completes the proof of Theorem 1. □

3 Proof of Lemmas

In this section we prove Lemma 1 and Lemma 2. In the proof of Lemma 1 we shall use some facts from [4], however, our proof follows the pattern matching approach developed in [12].

3.1 Proof of Lemma 1

The result is in fact already proven in [4]. Define

$$Q_w(z) = P(w) \left(\frac{z^k}{D_w^2(z)} - \frac{z}{(1 - (1 - P(w))z)^2} \right). \tag{9}$$

In [4] one defines $Q_n(1) = \frac{1}{n} \sum_{w \in \mathcal{A}^*} q_n(w)$ and it is proven there that $Q_n(1) = O(n^{-\varepsilon})$ for some $\varepsilon > 0$.

3.2 Proof of Lemma 2

First we have the following simple lemma. The largest eigenvalue of $\mathbf{P}$ is 1, let $\lambda_1, \lambda_2, \ldots$ be a sequence of other eigenvalues in the decreasing order of their modulus.

Lemma 3 *Uniformly for all $w \in \mathcal{A}^*$ we find $F_w(z) = O(\frac{1}{1-|\lambda_1 z|})$.*

Proof: By spectral representation of $\mathbf{P}$ we know that

$$\mathbf{P} = \boldsymbol{\pi} \otimes \mathbf{1} + \sum_{i>0} \lambda_i \mathbf{u}_i \otimes \boldsymbol{\zeta}_i,$$

where $\mathbf{u}_i$ (resp. $\boldsymbol{\zeta}_i$) are the corresponding right (resp. left) eigenvectors. In fact we can introduce the matrices $\mathbf{D} = \boldsymbol{\pi} \otimes \mathbf{1}$ and $\mathbf{R} = \sum_{i>0} \lambda_i \mathbf{u}_i \otimes \boldsymbol{\zeta}_i$ whose spectral radius is $|\lambda_1|$ and satisfies the orthogonal property: $\mathbf{RD} = \mathbf{DR} = 0$.

Let

$$\mathbf{M}(z) = \mathbf{P} - \boldsymbol{\pi} \otimes \mathbf{1}) \left(\mathbf{I} - z(\mathbf{P} + \boldsymbol{\pi} \otimes \mathbf{1}) \right)^{-1}.$$

We have

$$\mathbf{M}(z) = \mathbf{R}(\mathbf{1} - z\mathbf{R})^{-1}.$$

Since

$$\mathbf{R}^k = O(|\lambda_1|z)\mathbf{R}(\mathbf{I} - z\mathbf{R})^{-1}$$

is defined for all z such that $|z| < \frac{1}{|\lambda_1|}$ and is $O(\frac{1}{1-|\lambda_1 z|})$, and so is $F_w(z) = [\mathbf{M}(z)]_{a,b}$. □

The next lemma is important.

Lemma 4 *For z such that $|\lambda_1 z| < 1$ we have for all integers k*

$$\sum_{w\in\mathcal{A}^{k+1}} P(w)F_w(z) = O(\lambda_1^k). \tag{10}$$

Proof: The function $F_w(z)$ depends only on the first and last symbol of w. Considering a pair of symbols $(a, b) \in \mathcal{A}^2$ the sum of the probabilities of the words of length $k + 1$ starting with a and ending with b, $\sum_{awb\in\mathcal{A}^{k+1}} P(w)$, equals $\pi_a\langle \mathbf{e}_b\mathbf{P}^k\mathbf{e}_a\rangle$. Easy algebra leads to

$$\begin{aligned}\sum_{w\in\mathcal{A}^{k+1}} P(w)F_w(z) &= \sum_{(a,b)\in\mathcal{A}^2} \langle \mathbf{e}_a\mathbf{M}(z)\mathbf{e}_b\rangle\langle \mathbf{e}_b\mathbf{P}^k\mathbf{e}_a\rangle &(11)\\ &= \text{trace}\left(\mathbf{M}(z)\mathbf{P}^k\right). &(12)\end{aligned}$$

But since $\mathbf{P}^k = \mathbf{D} + \mathbf{R}^k$ and $\mathbf{M}(z)\mathbf{D} = 0$ and $\mathbf{R}^k = O(|\lambda_1|^k)$, this concludes the proof. □

We now follow the approach developed in [4] and in [8, 12].
The generating function $\Delta_w(z) = \sum_{n\geq 0} d_n(w)z^n$ becomes

$$\Delta_w(z) = \frac{P(w)z}{1-z}\left(\frac{1+(1-z)F_w(z)}{D_w(z)} - \frac{1}{1-z+P(w)z}\right). \tag{13}$$

We have

$$d_n(w) = \frac{1}{2i\pi}\oint \Delta_w(z)\frac{dz}{z^{n+1}},$$

integrated on any loop encircling the origin in the definition domain of $d_w(z)$. Extending the result from [8], the authors of [4] showed that there exists $\rho > 1$ such that the function $D_w(z)$ has a single root in the disk of radius ρ. Let A_w be such a root. We have via the residue formula

$$d_n(w) = \text{Res}(\Delta_w(z), A_w)A_w^{-n} - (1 - P(w))^n + d_n(w, \rho), \tag{14}$$

where $\text{Res}(f(z), A)$ denotes the residue of function $f(z)$ on a complex number A. Then

$$d_n(w, \rho) = \frac{1}{2i\pi}\oint_{|z|=\rho} \Delta_w(z)\frac{dz}{z^{n+1}}. \tag{15}$$

We have

$$\text{Res}(\Delta_w(z), A_w) = \frac{P(w)\left(1 + (1 - A_w)F_w(A_w)\right)}{(1 - A_w)C_w} \tag{16}$$

where $C_w = D'_w(A_w)$. But since $D_w(A_w) = 0$ we can write

$$\text{Res}(\Delta_w(z), A_w) = -\frac{A_w^{-k} S_w(A_w)}{C_w} \tag{17}$$

We now consider asymptotic expansion of A_w and C_w as it is described in [12], in Lemma 8.1.8 and Theorem 8.2.2. Although the expansion presented in [12] was for memoryless case, we easily extend it to Markov sources by simply replacing $S_w(1)$ by $S_w(1) + P(w)F_w(1)$. We find

$$\begin{aligned} A_w &= 1 + \tfrac{P(w)}{S_w(1)}, \\ &\quad + P(w)^2 \left(\tfrac{k - F_w(1)}{S_w^2(1)} - \tfrac{S'_w(1)}{S_w^3(1)} \right) + O(P(w)^3), \\ C_w &= -S_w(1) + P(w) \left(k - F_w(1) - 2\tfrac{S'_w(1)}{S_w(1)} \right) \\ &\quad + O(P(w)^2). \end{aligned} \tag{18}$$

Notice that these expansions in the Markov model first appeared in [4].

From now we follow the proof of Theorem 8.2.2 in [12]. We define the function

$$\delta_w(x) = \frac{A_w^{-k} S_w(A_w)}{C_w} A_w^{-x} - (1 - P(w))^x. \tag{19}$$

More precisely we define the function

$$\bar{\delta}_w(x) = \delta_w(x) - \delta_w(0) e^{-x}$$

which has a Mellin transform $\delta_w^*(s)\Gamma(s) = \int_0^\infty \bar{\delta}_w(x) x^{s-1} dx$ defined for all $\Re(s) \in (-1, 0)$ with

$$\delta_w^*(s) = \frac{A_w^{-k} S_w(A_w}{C_w} \left[(\log A_w)^{-s} - 1 \right] + 1 - \left[-\log(1 - P(w)) \right]^{-s}. \tag{20}$$

When $w \in \mathcal{B}_k$ with the expansion of A_w and since $S_w(1) = 1 + O(\delta^k)$ and $S'_w(1) = O(k\delta^k)$, we find that similarly as shown in [12]

$$\delta_w^*(s) = O(|s| k \delta^k) P(w)^{1-s}. \tag{21}$$

Therefore, by the reverse Mellin transform, for all $1 > \varepsilon > 0$:

$$\begin{aligned} \bar{\delta}(n, w) &= \frac{1}{2i\pi} \int_{-\varepsilon - i\infty}^{-\varepsilon + i\infty} \delta_w^*(s) \Gamma(s) n^{-s} ds \\ &= O(n^{1-\varepsilon} P(w)^{1-\varepsilon} k \delta^k). \end{aligned} \tag{22}$$

When $w \in \mathcal{A}^k - \mathcal{B}_k$ we don't have the $S_w(1) = 1 + O(\delta^k)$. But it is shown in [4] that there exists $\alpha > 0$ such that for all $w \in \mathcal{A}^*$: $S_w(z) > \alpha$ for all z such that $|z| \leq \rho$. Therefore we find

$$\bar{\delta}(n, w) = O(n^{1-\varepsilon} P(w)^{1-\varepsilon}).$$

We set

$$R_n(w) = d_w(0) e^{-n} + d_n(w, \rho). \tag{23}$$

We first investigate the quantity $d_w(0)$. We need to prove that $\sum_{w\in\mathcal{A}^*} d_w(0)$ converges. For this, noticing that

$$S_w(A_w) = S_w(1) + \frac{P(w)}{S_w(1)} S'_w(1) + O(P(w)^2)$$

we obtain

$$-\frac{A_w^{-k} S_w(A_w)}{C_w} = 1 - \frac{P(w)}{S_w(1)} \left(F_w(1) + \frac{S'_w(1)}{S_w(1)} \right) + O(P(w)^2). \tag{24}$$

Thus

$$d_w(0) = -\frac{P(w)}{S_w(1)} \left(F_w(1) + \frac{S'_w(1)}{S_w(1)} \right) + O(P(w))^2). \tag{25}$$

Without the term $F_w(1)$ we would have the same expression as in [12] whose sum over $w \in \mathcal{A}^*$ converges. Therefore we need to prove that the sum $\sum_{w\in\mathcal{A}^*} \frac{P(w)}{S_w(1)} F_w(1)$ converges. It is clear that the sum

$$\sum_k \sum_{w\in\mathcal{A}^k - \mathcal{B}_k} \frac{P(w)}{S_w(1)} F_w(1)$$

converges since

$$\sum_{w\in\mathcal{A}^k - \mathcal{B}_k} P(w) = O(\delta^k)$$

and $F_w(1)$ is uniformly bounded. Now we consider the other part

$$\sum_k \sum_{w\in\mathcal{B}_k} \frac{P(w)}{S_w(1)} F_w(1).$$

We know that $S_w(1) = 1 + O(\delta^k)$, therefore

$$\sum_{w\in\mathcal{B}_k} \frac{P(w)}{S_w(1)} F_w(1) = \sum_{w\in\mathcal{B}^k} P(w) F_w(1) + O(\delta^k). \tag{26}$$

But

$$\sum_{w\in\mathcal{B}^k} P(w) F_w(1) = \sum_{w\in\mathcal{A}^k} P(w) F_w(1) + O(\delta^k),$$

and we know by Lemma 4 that $\sum_{w\in\mathcal{A}^k} P(w) F_w(1) = O(\lambda_1^k)$. Thus

$$\sum_k \sum_{w\in\mathcal{A}^k} \frac{P(w)}{S_w(1)} F_w(1) < \infty$$

and this completes this part of the proof.

The second and last effort concentrates on the term $d_n(w,\rho)$. We proceed as in the proof of Theorem 8.2.2 in [12]. We first have $d_n(w,\rho) = O(P(w)\rho^{-n})$ which is $O(n^\varepsilon P(w)^\varepsilon)$ without any condition on w.

The issue is now to work on $w \in \mathcal{B}_k$. In this case we have $S_w(z) = 1 + O(\delta^k)$ and therefore

$$\begin{aligned} d_n(w,\rho) &= \frac{1}{2i\pi}\oint \frac{P(w)}{1-z}\left(\frac{1}{D_w(z)} - \frac{1}{1-z+zP(w)}\right)\frac{dz}{z^{n+1}} \\ &\quad + \frac{1}{2i\pi}\oint P(w)\frac{F_w(z)}{D_w(z)}\frac{dz}{z^{n+1}}. \end{aligned} \tag{27}$$

We notice that the function

$$\frac{P(w)}{1-z}\left(\frac{1}{D_w(z)} - \frac{1}{1-z+zP(w)}\right)$$

is $O(P(w)\delta^k) + O(P(w)^2)$, therefore the first integral is $O(P(w)\delta^k\rho^{-n})$. Observe now that

$$P(w)\frac{F_w(z)}{D_w(z)} = P(w)F_w(z) + O(P(w)\delta^k).$$

We already know that $\sum_{w\in\mathcal{B}_k} P(w)F_w(z) = O(\lambda_1^k)$, thus the series converges and the lemma is proven.

4 Sketch of the Proof of Theorem 2

Let $a \in \mathcal{A}$. We denote by $t_{a,n}$ the average size of a trie over n independent Markovian sequences, all starting with the same symbol a. Then for $n \geq 2$

$$t_n = 1 + \sum_{a\in\mathcal{A}}\sum_{k=0}^{n}\binom{n}{k}\pi_a^k(1-\pi_a)^{n-k}t_{a,k}, \tag{28}$$

and similarly for $b \in \mathcal{A}$

$$t_{n,b} = 1 + \sum_{a\in\mathcal{A}}\sum_{k=0}^{n}\binom{n}{k}P(a|b)^k(1-P(a|b))^{n-k}t_{a,k}, \tag{29}$$

where we recall $P(a|b)$ is the (a,b) element of matrix $\mathbf{P}$. Let $T(z) = \sum_n t_n\frac{z^n}{n!}e^{-z}$ and $T_a(z) = \sum_n t_{a,n}\frac{z^n}{n!}e^{-z}$ be the familiar Poisson transforms. Using (28) and (29) we find

$$T(z) = 1 - (1+z)e^{-z} + \sum_{a\in\mathcal{A}} T_a(\pi_a z), \tag{30}$$

$$T_b(z) = 1 - (1+z)e^{-z} + \sum_{a\in\mathcal{A}} T_a(P(a|b)z). \tag{31}$$

Using dePoissonization arguments (see [11]) we shall obtain

$$t_n = T(n) + O(\frac{1}{n}T(n)).$$

Thus we need to study $T(z)$ for large z in a cone around the real axis. For this we apply the Mellin transform that we describe next. In fact the convergence between the quantities t_n and T_n could also be derived by the application of the Rice method on the Mellin transform, since the later as an explicit form.

Let now $\mathbf{T}(z)$ be the vector consisting of $T_a(z)$ for every $a \in \mathcal{A}$. It is not hard to see that its Mellin transform

$$\mathbf{T}^*(s) = \int_0^\infty \mathbf{T}(z) z^{s-1} dz$$

is defined for $-1 > \Re(s) > -2$ (since $\mathbf{T}(z) = O(z^2)$ when $z \to 0$), and

$$\mathbf{T}^*(s) = -(1+s)\Gamma(s)\mathbf{1} + \mathbf{P}(s)\mathbf{T}^*(s), \tag{32}$$

where $\mathbf{P}(s)$ is the matrix consisting of $P(a|b)^{-s}$ if $P(a|b) > 0$ and 0 otherwise. This identity leads to

$$\mathbf{T}^*(s) = -(1+s)\Gamma(s)(\mathbf{I} - \mathbf{P}(s))^{-1}\mathbf{1},$$

where $\mathbf{I}$ is the identity matrix. Similarly the Mellin transform $T^*(s)$ of $T(z)$ satisfies

$$T^*(s) = -(1+s)\Gamma(s) + \langle \boldsymbol{\pi}(s), \mathbf{T}^*(s) \rangle. \tag{33}$$

where $\boldsymbol{\pi}(s)$ is the vector composed of π_a^{-s}.

The inverse Mellin transform of $T^*(s)$ is defined as

$$T(n) = \frac{1}{2i\pi} \int_{c-i\infty}^{c+i\infty} T^*(s) n^{-s} ds, \quad -1 > c > -2. \tag{34}$$

In order to find asymptotic behavior of $T(z)$ as $z \to \infty$ we need to study the poles of $T^*(s)$ for $-2 < \Re(s)$. As discussed in [9, 12] this is equivalent to analyzing the poles of $\mathbf{T}^*(s)$. Since $(1+s)\Gamma(s)$ has no pole on $-2 < \Re(s) < 0$ we must consider poles of $(\mathbf{I} - \mathbf{P}(s))^{-1}$. In other words (see [9, 12]) we need to find s for which the eigenvalue of largest modulus $\lambda(s)$ of $\mathbf{P}(s)$ is equal to 1. It is easy to see that $\lambda(-1) = 1$ since $\mathbf{P}(-1) = \mathbf{P}$. The residue at $s = -1$ of $n^{-s}(\mathbf{I} - \mathbf{P}(s))^{-1}\mathbf{1}$ is equal to $\frac{n}{h}\mathbf{1}$ where h is the entropy rate of the Markovian source.

As explained in [9] in the periodic case there are multiple values of s such that $\lambda(s) = 1$ and $\Re(s) = -1$. Since these poles are regularly spaced on the axis $\Re(s) = 0$, they contribute to the oscillating terms (function Q in Theorem 2) in the asymptotic expansion of t_n. Furthermore, the location of zeros of $\lambda(s) = 1$ in the periodic case tells us that there exists ε such that $(\mathbf{I} - \mathbf{P}(s))$ has no pole for $-1 < \Re(s) < -1 + \varepsilon$ leading to the error term $O(n^{1-\varepsilon})$.

In the aperiodic case there is only one pole on the line $\Re(s) = -1$, thus the oscillating term disappears. However, zeros of $\lambda(s) = 1$ can lie arbitrarily close to the line $\Re(s) = 1$, therefore the error term is just $o(n)$.

References

[1] A. Apostolico, and W. Szpankowski, Self-Alignments in Words and Their Applications, *J. Algorithms*, 13, 446–467, 1992.

[2] P. Cénac, B. Chauvin, F. Paccaut, and N. Pouyanne, Uncommon suffix tries, *Random Structures & Algorithms*, vol. 46, 117-141, 2015

[3] J. Clement, P. Flajolet, and B. Vallée, Dynamic Sources in Information Theory: A General Analysis of Trie Structures, *Algorithmica*, 29, 307-369, 2001.

[4] Fayolle, J., Ward, M. D. Analysis of the average depth in a suffix tree under a Markov model. In *International Conference on Analysis of Algorithms DMTCS*, proc. AD (Vol. 95, p. 104), 2005.

[5] P. Flajolet, N. Roux, M. and B. Vall'e, Digital trees and memoryless sources: from arithmetics to analysis. DMTCS Proceedings, (01), 2010.

[6] J. Geithner and M. Ward, Variance of the Profile in Suffix Trees, *Proc. 27th Int. Conference on Probabilistic, Combinatorial, and Asymptotic Methods for Analysis of Algorithms*, Kraków, 2016.

[7] HK Hwang, M. Fuchs, and V. Zacharovas, Asymptotic variance of random symmetric digital search trees, *Discrete Mathematics and Theoretical Computer Science* DMTCS vol. 12:2, 103-166, 2010.

[8] P. Jacquet, and W. Szpankowski, Autocorrelation on words and its applications: analysis of suffix trees by string-ruler approach. *J. Combinatorial Theory*, Series A, 66(2), 237-269, 1994.

[9] P. Jacquet, W. Szpankowski, and J. Tang, Average Profile of the Lempel-Ziv Parsing Scheme for a Markovian Source, *Algorithmica*, 2002.

[10] P. Jacquet, W. Szpankowski, Analysis of digital tries with Markovian dependency. *IEEE Transactions on Information Theory*, 37(5), 1470-1475, 1991.

[11] P. Jacquet, W. Szpankowski, Analytical depoissonization and its applications. *Theoretical Computer Science*, 201(1), 1-62, 1998.

[12] P. Jacquet, W. Szpankowski, *Analytic Pattern Matching: From DNA to Twitter*. Cambridge University Press, Cambridge, 2015.

[13] D. E. Knuth, *The Art of Computer Programming. Sorting and Searching*, Vol. 3, Second Edition, Addison-Wesley, Reading, MA, 1998.

[14] N. Merhav and W. Szpankowski, Average Redundancy of the Shannon Code for Markov Sources, *IEEE Trans. Information Theory*, 59, 7186-7193, 2013.

[15] B. Pittel, Asymptotic Growth of a Class of Random Trees, *Annals of Probability*, 18, 414–427, 1985.

[16] M. Regnier, and P. Jacquet, New Results on the Size of Tries, *IEEE Trans. Information Theory*, 35, 203–205, 1989.

[17] M. Régnier, and W. Szpankowski, On Pattern Frequency Occurrences in a Markovian Sequence, *Algorithmica*, 22, 631–649, 1998.

[18] W. Schachinger, On the Variance of a Class of Inductive Valuations of Data Structures for Digital Search, *Theoretical Computer Science*, 144, 251–275, 1995.

[19] R. Sedgewick, and P. Flajolet, *An Introduction to the Analysis of Algorithms*, Addison-Wesley, Reading, MA, 1995.

[20] P. Shields, *The Ergodic Theory of Discrete Sample Paths*, American Mathematical Society, Providence, 1996.

[21] W. Szpankowski, A Generalized Suffix Tree and Its (Un)Expected Asymptotic Behaviors, *SIAM J. Computing*, 22, 1176–1198, 1993.

[22] W. Szpankowski, *Average Case Analysis of Algorithms on Sequences*, John Wiley, New York, 2001.