Analysis of the average depth in a suffix tree under a Markov model

Julien Fayolle[1] and Mark Daniel Ward[2]

[1]*Projet Algorithmes, INRIA, F-78153 Rocquencourt, France.* `julien.fayolle@inria.fr`
[2]*Department of Mathematics, Purdue University, West Lafayette, IN, USA.* `mward@math.purdue.edu`

In this report, we prove that under a Markovian model of order one, the average depth of suffix trees of index n is asymptotically similar to the average depth of tries (a.k.a. digital trees) built on n independent strings. This leads to an asymptotic behavior of $(\log n)/h + C$ for the average of the depth of the suffix tree, where h is the entropy of the Markov model and C is constant. Our proof compares the generating functions for the average depth in tries and in suffix trees; the difference between these generating functions is shown to be asymptotically small. We conclude by using the asymptotic behavior of the average depth in a trie under the Markov model found by Jacquet and Szpankowski ([4]).

Keywords: Suffix trees, depth, average analysis, asymptotics, analytic methods

1 Introduction

The suffix tree is a data structure that lies at the core of pattern matching, used for example in the lossless data compression algorithm of Lempel and Ziv [7]. Suffix trees are also utilized in bio-informatics to track "significant" patterns. A thorough survey of the use of suffix trees in computer science can be found in Apostolico ([1]). The average depth in a suffix tree of index $n+1$ is the average time (number of letters read) necessary to insert a new suffix into a tree of index n; an alternate interpretation is the average time to discriminate between the current suffix and any previous one.

A device called a *source* emits letters randomly, and independently of their emission date. In this report, we focus on Markov sources of order one: the letter emitted at a given time is a random variable obeying a Markov dependency of order one i.e., the distribution of each letter depends only on the letter actually emitted immediately beforehand). Increasing the order introduces no new technical challenges—computations only become more intricate—and our proof for a Markovian dependency of order 1 can easily be extended to any source with greater Markovian dependency. We assume that the source is stationary throughout this report. We denote the stationary probability vector as $\pi = (\pi_i)$, the transition probability matrix as $P = (p_{i,j})$, and the probability the source starts emitting a text with the word w as $\mathbf{P}(w)$. We also introduce the stationary probability matrix Π whose rows are the stationary probability vector π. We make the usual assumptions that the underlying Markov chain is irreducible and aperiodic.

This paper focuses on the asymptotic behavior of the average depth in suffix trees. There are two challenging parts in this study: first the suffixes on which the suffix tree is built are mutually dependent. For example, if we just found the pattern $w = \mathtt{0000}$ in the text, then it suffices to have the letter $\mathtt{0}$ next, in order to find w once more in the text. The trie (a.k.a. digital tree) is much easier to analyze than the suffix tree, because the strings in a trie are independent from each other. The second challenge lies in the probabilistic dependency between symbols (Markov model). The probability that a pattern occurs in the text depends not only on the pattern itself but also on what was previously seen in the text.

In Jacquet and Szpankowski ([4]), inclusion-exclusion was used to obtain the asymptotics of the average depth in an (independent) trie with underlying Markovian source. Therefore, it suffices for us to compare the asymptotic average depths in suffix trees against that of independent tries (where the probability model is Markovian in both cases). We prove that, for a Markovian source of order one, the average depth of a suffix tree of index n has a similar asymptotic behavior as a trie built on n strings.

In section 2, we present the main results of this paper. We give the precise asymptotics for the average depth in a suffix tree with an underlying Markovian source. Afterwards, we give a sketch of the proof. In section 3 we prove that the autocorrelation polynomial $S_w(z)$ is approximately 1 with high probability (for w of large length). To prove that suffix trees and independent tries have similar average depths, we first derive bivariate generating functions for D_n and D_n^t in section 4. Then in subsections 5.1 and 5.2 we analyze the difference between the generating functions for D_n and D_n^t by utilizing complex asymptotics. Ultimately, we conclude in subsection 5.3 that the depth in a suffix tree has the same average up to the constant term as the depth in a trie. Our method is motivated by the analytic approach of Jacquet and Szpankowski ([3]) for the suffix tree under a memoryless (a.k.a. Bernoulli) source model. Our results are proved in the more general context of Markovian sources.

2 Main Results

Consider a tree (say it's binary and it shall be so throughout this paper) with n leaves numbered from 1 to n. The depth $D_n(i)$ of the leaf number i is defined as the length of the path from the root to this leaf. Pick one of the n leaves at random (uniformly); the *typical depth* D_n of the tree is the depth of the picked leaf. The random variable D_n informs us about the typical profile of the tree, rather than its height (i.e., the maximum depth).

A trie is defined recursively on a finite set S of infinite words on $\mathcal{A} = \{\mathtt{0}, \mathtt{1}\}$ as

$$\operatorname{trie}(S) = \begin{cases} \emptyset & \text{if } |S| = 0, \\ \bullet & \text{if } |S| = 1, \\ \langle \bullet, \operatorname{trie}(S\backslash\mathtt{0}), \operatorname{trie}(S\backslash\mathtt{1})\rangle & \text{else,} \end{cases}$$

where $S\backslash\alpha$ represents the set of words starting with the letter α whose first letter has been removed.

In order to build a suffix tree, one needs as input an infinite string T on $\mathcal{A}$ called *text* and an integer n. We write $T = T_1T_2T_3\ldots$ and then $T^{(i)} = T_iT_{i+1}T_{i+2}\ldots$ denotes the ith suffix of T (for example, $T = T^{(1)}$, namely the entire string itself, is the first suffix). The suffix tree of index n built on T is defined as the trie built on the set of the n first suffixes of T (namely, $T^{(1)}, \ldots, T^{(n)}$). In the context of suffix trees, $D_n(i)$ is interpreted as the largest value of k such that there exists $j \neq i$ with $1 \leq j \leq n$ for which $T_i^{i+k-1} = T_j^{j+k-1}$.

Our discussion in this paper primarily concerns the comparison of the average depths in suffix trees versus independent tries. Throughout this discussion, D_n always denotes the typical depth in a *suffix* tree. In a trie built over n *independent* strings, we let D_n^t denote the typical depth.

In [4], the asymptotic behavior of the average depth for a trie built on n independent strings under a *Markovian source of order 1* was established using inclusion-exclusion. In this paper, we prove the analogous result for suffix trees. Namely, we establish the asymptotic average depth for a *suffix* tree with an underlying Markovian source. Our proof technique consists of showing that D_n and D_n^t have similar asymptotic distributions. To do this, we estimate the difference of their probability generating functions, and show that the difference is asymptotically small.

In order to prove that D_n and D_n^t have asymptotically similar averages, we first discuss the autocorrelation of a string w. Since a string w can overlap with itself, the bivariate generating function $D(z, u)$ for D_n, where u marks the depth and z the size, includes the autocorrelation polynomial $S_w(z)$. Fortunately, with high probability, a random string w has very little overlap with itself. Therefore the autocorrelation polynomial of w is close to 1 with high probability.

After discussing autocorrelation, we present the bivariate generating functions for both D_n and D_n^t, which we denote by $D(z, u)$ and $D^t(z, u)$, respectively. We have

$$D(z,u) = \frac{1-u}{u} \sum_{w \in \mathcal{A}^*} (zu)^{|w|} \frac{\mathbf{P}(w)}{\mathfrak{D}_w(z)^2} \qquad \text{and} \qquad D^t(z,u) = \frac{1-u}{u} \sum_{w \in \mathcal{A}^*} u^{|w|} \frac{z\mathbf{P}(w)}{(1-z+z\mathbf{P}(w))^2}, \tag{1}$$

where $\mathfrak{D}(z)$ is a generating function.

Our goal is to prove that the two generating functions are asymptotically very similar; it follows from there that D_n^t and D_n have the same average asymptotically up to the constant term.

In order to determine the asymptotics of the difference $D(z, u) - D^t(z, u)$, we utilize complex analysis. From (1), we see that $D^t(z, u)$ has exactly one pole of order 2 for each w. Using Rouché's theorem, we prove that, for $|z| \leq \rho$ (with $\rho > 1$) and for sufficiently large $|w|$, the generating function $\mathfrak{D}_w(z)$ has exactly one dominant root of order 2 for each w. Therefore $D(z, u)$ has exactly one dominant pole of order 2 for each w. We next use Cauchy's theorem and singularity analysis to quantify the contribution from each of these poles to the difference $Q_n(u) := u(1-u)^{-1}[z^n](D(z,u) - D^t(z,u))$. Our analysis of the difference $Q_n(u)$ ultimately relies on the Mellin transform.

We conclude from there that the averages of the depths D_n (in suffix trees) and D_n^t (in independent tries) are asymptotically similar. Therefore, D_n has mean $\frac{1}{h}\log n$ plus some fluctuations. Specifically,

Theorem 1 *For a suffix tree built on the first n suffixes of a text produced by a source under the Markovian model of order one, the average typical depth is asymptotically*

$$\mathbb{E}(D_n) = \frac{1}{h}\left(\log n + \gamma + \frac{h_2}{2h} - H + P_1(\log n)\right) + O\left(n^{-\epsilon}\right), \tag{2}$$

for some positive ϵ, where γ is the Euler constant, h is the entropy of the Markov source, h_2 is the second order entropy, H is the stationary entropy and P_1 is a function fluctuating around zero with a very small amplitude. In particular

$$h := -\sum_{i,j\in\mathcal{A}^2} \pi_i p_{i,j} \log p_{i,j} \quad \text{and} \quad H := -\sum_{i\in\mathcal{A}^2} \pi_i \log \pi_i. \tag{3}$$

3 Autocorrelation

The depth of the ith leaf in a suffix tree can be construed in the language of pattern matching as "the length of the longest factor starting at a position i in the text that can be seen at least once more in the text". Hence $D_n(i) \geq k$ means that there is at least another pattern in the text T starting with T_i^{i+k-1}

While looking for the longest pattern in T that matches the text starting in position i, there is a possibility that two occurrences of this longest pattern overlap; this possible overlap of the pattern w with itself causes complications in the enumeration of occurrences of w in a text. The phenomenon exhibited when w overlaps with itself is called *autocorrelation*.

To account for this overlapping, we use the probabilistic version of the autocorrelation polynomial. For a pattern w of size k, the polynomial is defined as:

$$S_w(z) := \sum_{i=0}^{k-1} [\![w_1^{k-i} = w_{i+1}^k]\!] \mathbf{P}(w_{k-i+1}^k | w_{k-i}) z^i. \tag{4}$$

For simplicity, we introduce $c_i = [\![w_1^{k-i} = w_{i+1}^k]\!]$ where $[\![.]\!]$ is Iverson's bracket notation for the indicator function. We say there is an overlap of size $k-i$ if $c_i = 1$. This means that the suffix and the prefix of size $k-i$ of w coincide. Graphically, an overlap of a pattern (white rectangles) looks like this:

,

where the two black boxes are the matching suffix and prefix. For example w=`001001001` has overlaps of sizes 3, 6 and 9; note that $|w|$ (here $|w| = 9$) is always a valid overlap since the pattern w always matches itself. We define a *period* of a pattern w as any integer i for which $c_i = 1$ (so a pattern w often has several periods), the minimal period $m(w)$ of w is the smallest positive period, if one exists (if w has no positive period for w, we define $m(w) = k$).

We now formulate precisely the intuition that, for most patterns of size k, the autocorrelation polynomial is very close to 1. It stems from the fact that the sum of the probabilities $\mathbf{P}(w)$ over all loosely correlated patterns (patterns with a large minimal period) of a given size is very close to 1.

Lemma 1 *There exist $\delta < 1$, $\rho > 1$ with $\rho\delta < 1$, and $\theta > 0$, such that for any integer k*

$$\sum_{w\in\mathcal{A}^k} [\![|S_w(\rho) - 1| \leq (\rho\delta)^k\theta]\!] \mathbf{P}(w) \geq 1 - \theta\delta^k. \tag{5}$$

Proof: Note that $S_w(z) - 1$ has a term of degree i with $1 \leq i \leq k-1$ if and only if $c_i = 1$. If $c_i = 1$ the prefix of size i of w will repeat itself fully in w as many times as there is i in k, hence knowing $c_i = 1$ and the first i letters of w allows us to describe fully the word w. Therefore, given a fixed $w_1 \dots w_i$, there is *exactly one word* $w_{i+1}\dots w_k$ such that the polynomial $S_w(z)-1$ has minimal degree $i \le k$. We let $\tilde{p}_{i,j} = p_{w_i,w_j}$, $\tilde{\pi}_i = \pi_{w_i}$, and $p = \max_{1\le i,j\le 2}(\tilde{p}_{i,j}, \tilde{\pi}_i)$. Then, for fixed j and k,

$$\begin{aligned}\sum_{i=1}^{j}\sum_{w\in\mathcal{A}^k}[\![m(w)=i]\!]\mathbf{P}(w) &= \sum_{i=1}^{j}\sum_{w_1,\dots,w_i\in\mathcal{A}^i}\tilde{\pi}_1\tilde{p}_{1,2}\cdots\tilde{p}_{i-1,i}\sum_{w_{i+1},\dots,w_k\in\mathcal{A}^{k-i}}[\![m(w)=i]\!]\tilde{p}_{i,i+1}\cdots\tilde{p}_{k-1,k}\\ &\le \sum_{i=1}^{j}\sum_{w_1,\dots,w_i\in\mathcal{A}^i}\tilde{\pi}_1\tilde{p}_{1,2}\tilde{p}_{2,3}\cdots\tilde{p}_{i-1,j}p^{k-i},\end{aligned} \tag{6}$$

but we can factor p^{k-i} outside the inner sum, since it does not depend on $w_1,\dots,w_i$. Next, we observe that $\sum_{w_1,\dots,w_i\in\mathcal{A}^i}\tilde{\pi}_1\tilde{p}_{1,2}\tilde{p}_{2,3}\cdots\tilde{p}_{i-1,i} = \Pi P^{i-1}\mathbf{1} = 1$, so

$$\sum_{w\in\mathcal{A}^k}[\![S_w(z)-1 \text{ has minimal degree} \le j]\!]\mathbf{P}(w) \le \sum_{i=1}^{j} p^{k-i} \le \frac{p^{k-j}}{1-p},$$

and this holds when $j = \lfloor k/2 \rfloor$. So

$$\sum_{w\in\mathcal{A}^k}[\![\text{all terms of } S_w(z)-1 \text{ have degree} > \lfloor k/2\rfloor]\!]\mathbf{P}(w) \ge 1 - \frac{p^{\lceil k/2\rceil}}{1-p} = 1-\theta\delta^k. \tag{7}$$

Remark that, if all terms of $S_w(z)-1$ have degree $> \lfloor k/2 \rfloor$, then

$$|S_w(\rho)-1| \le \sum_{i=\lfloor k/2\rfloor}^{k}(\rho p)^i \le \rho^k\frac{p^{\lfloor k/2\rfloor+1}}{1-p} = (\rho\delta)^k\theta. \tag{8}$$

We select $\delta = \sqrt{p}$, $\theta = (1-p)^{-1}$ and some $\rho > 1$ with $\delta\rho < 1$ to complete the proof of the lemma. □

The next lemma proves that, for $|w|$ sufficiently large and for some radius $\rho > 1$, the autocorrelation polynomial does not vanish on the disk of radius ρ.

Lemma 2 *There exist K, $\rho' > 1$ with $p\rho' < 1$, and $\alpha > 0$ such that for any pattern w of size larger than K and z in a disk of radius ρ', we have*

$$|S_w(z)| \ge \alpha.$$

Proof: Like for the previous lemma, we split the proof into two cases, according to the index i of the minimal period of the pattern w of size k. Since the autocorrelation polynomial always has $c_0 = 1$, we write

$$S_w(z) = 1 + \sum_{j=i}^{k-1} c_j\mathbf{P}(w_{k-j+1}^k|w_{k-j})z^j. \tag{9}$$

We introduce $\rho' > 1$ such that $p\rho' < 1$. Therefore, if $i > \lfloor k/2 \rfloor$, then

$$|S_w(z)| \ge 1 - \left|\sum_{j=i}^{k-1} c_j\mathbf{P}(w_{k-j+1}^k|w_{k-j})z^j\right| \ge 1 - \frac{(p\rho')^i}{1-p\rho'}, \tag{10}$$

in $|z| \le \rho'$. But since $i > \lfloor k/2 \rfloor$ and $p\rho' < 1$, we get

$$|S_w(z)| \ge 1 - \frac{(p\rho')^{k/2}}{1-p\rho'}. \tag{11}$$

We observe that, for some K_1 sufficiently large, any pattern w of size larger than K_1 satifies $|S_w(z)| \ge \alpha$ and this lower bound α is positive.

We recall that if $c_i = 1$, the prefix u of size i of w will repeat itself fully in w as many times as there is i in k i.e., $q := \lfloor k/i \rfloor$ times, the remainder will be the prefix v of u of size $r := k - \lfloor k/i \rfloor i$, hence $w = u^{\lfloor k/i \rfloor}v$. We also introduce the word v' such that $vv' = u$ (of length $t := i - r = i - k + \lfloor k/i \rfloor i$).

If $i \le \lfloor k/2 \rfloor$, we make the autocorrelation polynomial explicit:

$$S_w(z) = 1 + \mathbf{P}(v'v|w_k)z^i + \mathbf{P}(v'uv|w_k)z^{2i} + \cdots + \mathbf{P}(v'u^{q-2}v|w_k)z^{i(q-1)}S_{uv}(z). \tag{12}$$

Overall, we can write $\mathbf{P}(v'u^jv|w_k) = A^{j+1}$ where $A = p_{w_k,v_1'}p_{v_1',v_2'}\cdots p_{v_t',v_1}\cdots p_{v_{r-1},v_r}$ is the product of i transition probabilities, but since $w_k = v_r$ we obtain

$$S_w(z) = 1 + Az^i + (Az^i)^2 + \cdots + (Az^i)^{q-1}S_{uv}(z) = \frac{1-(Az^i)^{q-1}}{1-Az^i} + (Az^i)^{q-1}S_{uv}(z). \tag{13}$$

Then we provide a lower-bound for $|S_w(z)|$:

$$\begin{aligned}|S_w(z)| &\ge \left|\frac{1-(Az^i)^{q-1}}{1-Az^i}\right| - \left|(Az^i)^{q-1}S_{uv}(z)\right| \ge \frac{1-(p\rho')^{i(q-1)}}{1+(p\rho')^i} - (p\rho')^{i(q-1)}|S_{uv}(z)| \\ &\ge \frac{1-(p\rho')^{i(q-1)}}{1+(p\rho')^i} - \frac{(p\rho')^{i(q-1)}}{1-p\rho'}.\end{aligned}$$

We have $p\rho' < 1$ so that $(p\rho')^k$ tends to zero with k. Since $i(q-1)$ is close to k (at worse $k/3$ if $w = uuv$) for some K_2 sufficiently large and patterns of size larger than K_2, only the term $(1+(p\rho')^i)^{-1}$ remains. Finally, we set $K = \max\{K_1, K_2\}$. □

4 On the Generating Functions

The probabilistic model for the random variable D_n is the product of a Markov model of order one for the source generating the strings (one string in the case of the suffix tree, n for the trie), and a uniform model on $\{1,\ldots,n\}$ for choosing the leaf. Hence, if X is a random variable uniformly distributed over $\{1,\ldots,n\}$, and T a random text generated by the source, the typical depth is

$$D_n := \sum_{i=1}^{n} D_n(i)(T)[\![X = i]\!].$$

For the rest of the paper, the t exponent on a quantity will indicate its trie version.

Our aim is to compare asymptotically the probability generating functions of the depth for a suffix tree (namely, $D_n(u) := \sum_k \mathbf{P}(D_n = k)u^k$) and a trie (namely, $D_n^t(u) := \sum_k \mathbf{P}(D_n^t = k)u^k$). We provide in this section an explicit expression for these generating functions and their respective bivariate extensions $D(z,u) := \sum_n nD_n(u)z^n$ and $D^t(z,u) := \sum_n nD_n^t(u)z^n$.

We first derive $D_n^t(u)$. Each string from the set of strings S is associated to a unique leaf in the trie. By definition of the trie, the letters read on the path going from the root of the trie to a leaf form the smallest prefix distinguishing one string from the $n-1$ others. We choose uniformly a leaf among the n leaves of the trie. Let $w \in \mathcal{A}^k$ denote the prefix of length k of the string associated to this randomly selected leaf. We say that $D_n^t < k$ if and only if the other $n-1$ texts do not have w as a prefix. It follows immediately that

$$\mathbf{P}(D_n^t(i) < k) = \sum_{w\in\mathcal{A}^k} \mathbf{P}(w)(1-\mathbf{P}(w))^{n-1},$$

consequently

$$D_n^t(u) = \frac{1-u}{u}\sum_{w\in\mathcal{A}^*} u^{|w|}\mathbf{P}(w)(1-\mathbf{P}(w))^{n-1} \qquad \text{and} \qquad D^t(z,u) = \frac{1-u}{u}\sum_{w\in\mathcal{A}^*} u^{|w|}\frac{z\mathbf{P}(w)}{(1-z+z\mathbf{P}(w))^2}, \tag{14}$$

for $|u| < 1$ and $|z| < 1$.

The suffix tree generating function is known from [5] and [3]: for $|u| < 1$ and $|z| < 1$

$$D(z,u) = \frac{1-u}{u}\sum_{w\in\mathcal{A}^*} (zu)^{|w|}\frac{\mathbf{P}(w)}{\mathfrak{D}_w(z)^2}, \tag{15}$$

where $\mathfrak{D}_w(z) = (1-z)S_w(z) + z^{|w|}\mathbf{P}(w)(1+(1-z)F(z))$ and for $|z| < ||P-\Pi||^{-1}$,

$$F(z) := \frac{1}{\pi_{w_1}}\left[\sum_{n\ge0}(P-\Pi)^{n+1}z^n\right]_{w_m,w_1} = \frac{1}{\pi_{w_1}}[(P-\Pi)(I-(P-\Pi)z)^{-1}]_{w_m,w_1}. \tag{16}$$

5 Asymptotics

5.1 Isolating the dominant pole

We prove first that for a pattern w of size large enough there is a single dominant root to $\mathfrak{D}_w(z)$. Then we show that there is a disk of radius greater than 1 containing each single dominant root of the $\mathfrak{D}_w(z)$'s for any w of size big enough but no other root of the $\mathfrak{D}_w(z)$'s.

Lemma 3 *There exists a radius $\rho > 1$ and an integer K' such that for any w of size larger than K', $\mathfrak{D}_w(z)$ has only one root in the disk of radius ρ.*

Proof: Let w be a given pattern. We apply Rouché's Theorem to show the uniqueness of the smallest modulus root of $\mathfrak{D}_w(z)$. The main condition we need to fulfill is that, on a given circle $|z| = \rho$,

$$f(z) := |(1-z)S_w(z)| > |\mathbf{P}(w)z^k(1 + (1-z)F(z))| =: g(z). \tag{17}$$

The function f is analytic since it is a polynomial, F is analytic for $|z| < ||P-\Pi||^{-1}$ (where $||P-\Pi||^{-1} > 1$), so g is too. For patterns of a size large enough, $\mathbf{P}(w)z^k$ will be small enough to obtain the desired condition.

The main issue is the bounding from above of $F(z)$ on the circle of radius ρ. We note $d = \min_{a\in\mathcal{A}} \pi_a$; this value is positive (otherwise a letter would never occur) and since $(P-\Pi)^{n+1} = P^{n+1} - \Pi$ (remember that $P\Pi = \Pi P = \Pi$ and $\Pi\Pi = \Pi$), we have:

$$|F(z)| \le \frac{1}{d}\left|\left[\sum_{n\ge 0}(P^{n+1}-\Pi)z^n\right]_{k,1}\right| \le \frac{1}{d}\sum_{n\ge 0}|[P^{n+1}]_{k,1} - [\Pi]_{k,1}||z|^n \tag{18}$$

$$\le \frac{1}{d}\sum_{n\ge 0} br^{n+1}\rho^n \le \frac{br}{d}\frac{1}{1-r\rho}, \tag{19}$$

where b and r are constants (independent of the pattern) with $0 < r < 1$ and ρ such that $r\rho < 1$.

Let K be an integer and ρ' some radius satisfying Lemma 2, we ask ρ to be smaller than ρ' so that $p\rho < 1$ and $|S_w(z)| \ge \alpha$ on $|z| = \rho$ and for $|w| \ge K$. There exists K'' large enough such that for any $k > K''$ we verify the condition

$$(p\rho)^k\left(1 + (1+\rho)\frac{br}{d}\frac{1}{1-r\rho}\right) < \alpha(\rho - 1). \tag{20}$$

On a disk of such radius ρ and for $k > \max\{K, K''\}$, the assumptions of Rouché's Theorem are satisfied, consequently $(f+g)(z) = \mathfrak{D}_w(z)$ has exactly as many zeros in the centered disk of radius ρ as $f(z)$, namely one zero, since $S_w(z)$ does not vanish.

Furthermore the assumptions of Rouché's Theorem are satisfied on $|z| = \rho$ for any pattern of size larger than $K' := \max\{K, K''\}$. So for any w with $|w| \ge K'$, $\mathfrak{D}_w(z)$ has exactly one root within the disk $|z| = \rho$. □

5.2 Computing residues

The use of Rouché's Theorem has established the existence of a single zero of smallest modulus for $\mathfrak{D}_w(z)$, we denote it A_w. We know by Pringsheim's Theorem that A_w is real positive. Let also B_w and C_w be the values of the first and second derivatives of $\mathfrak{D}_w(z)$ at $z = A_w$. We make use of a bootstrapping technique to find an expansion for A_w, B_w, and C_w along the powers of $\mathbf{P}(w)$ and we obtain

$$\begin{aligned} A_w &= 1 + \frac{\mathbf{P}(w)}{S_w(1)} + O(\mathbf{P}^2(w)),\\ B_w &= -S_w(1) + \mathbf{P}(w)\left[k - F(1) - 2\frac{S_w'(1)}{S_w(1)}\right] + O(\mathbf{P}^2(w)), \text{ and}\\ C_w &= -2S_w'(1) + \mathbf{P}(w)\left[-3\frac{S_w''(1)}{S_w(1)} + k(k-1) - 2F'(1) - 2kF(1)\right] + O(\mathbf{P}^2(w)). \end{aligned} \tag{21}$$

We now compare $D_n(u)$ and $D_n^t(u)$ to conclude that they are asymptotically close. We therefore introduce two new generating functions

$$Q_n(u) := \frac{u}{1-u}(D_n(u) - D_n^t(u)) \text{ and } Q(z,u) := \sum_{n\ge 0} nQ_n(u)z^n = \sum_{w\in\mathcal{A}^*} u^{|w|}\mathbf{P}(w)\left(\frac{z^{|w|}}{\mathfrak{D}_w^2(z)} - \frac{z}{(1-z(1-\mathbf{P}(w)))^2}\right).$$

We apply Cauchy's Theorem to $Q(z,u)$ with z running along the circle centered at the origin whose radius ρ was determined in 5.1. There are only three singularities within this contour: at $z = 0$, at A_w, and at $(1 - \mathbf{P}(w))^{-1}$. In order to justify the presence of the third singularity within the circle, we note that the condition (20) implies

$$\mathbf{P}(w)\rho < \mathbf{P}(w)\rho^k < (p\rho)^k \left(1 + (1+\rho)\frac{br}{d}\frac{1}{1-r\rho}\right) < \alpha(\rho-1) < (\rho-1), \tag{22}$$

since α is taken smaller than one. Thus $(1 - \mathbf{P}(w))^{-1}$ is smaller than the radius ρ.

For any w of a size larger than K', we have

$$\begin{aligned} I_w(\rho,u) &:= \frac{1}{2i\pi}\int_{|z|=\rho} u^{|w|}\mathbf{P}(w)\frac{1}{z^{n+1}}\left(\frac{z^{|w|}}{\mathfrak{D}_w^2(z)} - \frac{z}{(1-z(1-\mathbf{P}(w)))^2}\right)dz = u^{|w|}\mathbf{P}(w)f(w) \\ &= \operatorname{Res}(f(z);0) + \operatorname{Res}(f(z);A_w) + \operatorname{Res}\left(f(z);\frac{1}{1-\mathbf{P}(w)}\right) \\ &= nQ_n(u) + u^{|w|}\mathbf{P}(w)\left(\operatorname{Res}\left(\frac{z^{|w|}}{z^{n+1}\mathfrak{D}_w^2(z)};A_w\right) + \operatorname{Res}\left(\frac{z}{z^{n+1}(1-z(1-\mathbf{P}(w)))^2};\frac{1}{1-\mathbf{P}(w)}\right)\right). \end{aligned} \tag{23}$$

Since $z/(1-z(1-\mathbf{P}(w))^2$ is analytic at $z = A_w$ it does not contribute to the residue of $f(z)$ at A_w and can be discarded in the computation of the residue. The same is true for the $1/\mathfrak{D}_w^2(z)$ part in the residue of $f(z)$ at $z = 1/(1-\mathbf{P}(w))$.

We compute the residue at A_w using the expansion we found for B_w and C_w through bootstrapping. We set $k = |w|$ to simplify the notation, and by a Taylor expansion near A_w we obtain

$$\operatorname{Res}\left(\frac{z^{k-(n+1)}}{\mathfrak{D}_w^2(z)};A_w\right) = A_w^{|w|-n-1}\left(\frac{|w|-(n+1)}{B_w^2A_w} - \frac{C_w}{B_w^3}\right). \tag{24}$$

In order to compute the residue at $z = (1-\mathbf{P}(w))^{-1}$, we use the general formula $[z^n](1-az)^{-2} = (n+1)a^n$ thus

$$[z^{-1}]\frac{1}{z^n(1-z(1-\mathbf{P}(w)))^2} = [z^{n-1}]\frac{1}{(1-z(1-\mathbf{P}(w)))^2} = n(1-\mathbf{P}(w))^{n-1}. \tag{25}$$

Before we proceed to get the asymptotic behavior of $Q_n(u)$, the following technical lemma is very useful.

Lemma 4 *For any function f defined over the patterns on $\mathcal{A}$, and any y we have*

$$\sum_{w\in\mathcal{A}^k}\mathbf{P}(w)f(w) \le y + f_{max}\mathbf{P}(\{w\in\mathcal{A}^k : f(w) > y\}),$$

where f_{max} is the maximum of f over all patterns of size k.

Proof: For the patterns of size k with $f(w) > y$, f_{max} is an upper bound of $f(w)$; the probability of the other patterns is smaller than 1. □

We also note that $S_w(\rho) \le (1-p\rho)^{-1}$ and $\mathfrak{D}_w(z) = O(\rho^k)$ for $|z| \le \rho$. Thus (details are omitted), we obtain the following bound for the sum of $I_w(\rho,u)$ over all patterns of size k:

$$\sum_{w\in\mathcal{A}^k} I_w(\rho,u) = O((\delta\rho u)^k\rho^{-n})$$

There are only finitely many patterns w with $|w| < K'$; these terms provide a contribution of at most $O(B^{-n})$ to $Q_n(u)$ for some $B > 1$.

We have just proved the following

Lemma 5 *For some $\beta > 1$, and for all $|u| \le \beta$, there exists $B > 1$ for which we have*

$$Q_n(u) = \frac{1}{n}\sum_{w\in\mathcal{A}^*} u^{|w|}\mathbf{P}(w)\left(A_w^{|w|-n-1}\left(\frac{(n+1)-|w|}{B_w^2A_w} - \frac{C_w}{B_w^3}\right) - n(1-\mathbf{P}(w))^{n-1}\right) + O(B^{-n}). \tag{26}$$

5.3 Asymptotic behavior of $Q_n(u)$

Lemma 6 *For all β such that $1 < \beta < \delta^{-1}$, there exists a positive ϵ such that $Q_n(u) = O(n^{-\epsilon})$ uniformly for all $|u| \leq \beta$.*

Proof: For n large enough, the dominant term in equation 26 is

$$Q_n(u) = \sum_{w \in \mathcal{A}^*} u^{|w|} \mathbf{P}(w) \left(\frac{A_w^{|w|-n-2}}{B_w^2} - (1 - \mathbf{P}(w))^{n-1} \right) + O(n^{-1}). \tag{27}$$

We introduce the function

$$f_w(x) = \left[\frac{A_w^{|w|-x-2}}{B_w^2} - (1 - \mathbf{P}(w))^{x-1} \right] - \left[\frac{1}{A_w^{2-|w|} B_w^2} - \frac{1}{1 - \mathbf{P}(w)} \right] \exp(-x),$$

and perform a Mellin transform (see Flajolet, Gourdon, and Dumas [2] for a thorough overview of this technique) on $\sum_w u^{|w|} \mathbf{P}(w) f_w(x)$. It would have seemed natural to define $f_w(x)$ by its first term but for simplicity's sake we force a $O(x)$ behavior for $f_w(x)$ near zero by subtracting some value. By an application of Lemma 4, with $y = (|u|\delta)^k$ for some $\delta < 1$, we prove that the sum is absolutely convergent for $|u| \leq \beta$. Hence the Mellin transform $f^*(s,u)$ of the sum is defined and we have

$$f_w^\star(s) = \Gamma(s) \left(\frac{(\log A_w)^{-s}}{A_w^{2-|w|} B_w^2} - \frac{(-\log(1 - \mathbf{P}(w)))^{-s}}{1 - \mathbf{P}(w)} \right) \text{ and } f^*(s,u) = \sum_w u^{|w|} \mathbf{P}(w) f_w^\star(s).$$

$f^*(s,u)$ is analytical within an open strip $-1 < \Re(s) < c$ for some positive c. In order to do so we split the sum over all patterns into two parts.

For the patterns with $\Im(s)\mathbf{P}(w)$ small, we use an expansion of $f_w^\star(s)$ and then apply Lemma 4. We make sure these patterns do not create any singularity. For any given s there are only finitely many patterns with $\Im(s)\mathbf{P}(w)$ large so their contribution to the sum $f^*(s,u)$, although each is individually large, do not create any singularity.

Therefore, we are able to take the inverse Mellin transform of $f^*(s,u)$, and since there is no pole in the strip we obtain the stated result. □

This last result is sufficient to prove our initial claim. By definition $D_n^t(1) = D_n(1) = 1$, thus we have

$$Q_n(u) = \frac{u}{1-u}(D_n(u) - D_n^t(u)) = u \left(\frac{D_n(u) - D_n(1)}{1-u} - \frac{D_n^t(u) - D_n^t(1)}{1-u} \right). \tag{28}$$

We have $D_n'(1) = \mathbb{E}(D_n)$ and $(D_n^t)'(1) = \mathbb{E}(D_n^t)$, therefore when u tends to 1 we obtain

$$\mathbb{E}(D_n^t) - \mathbb{E}(D_n) = O(n^{-\epsilon}). \tag{29}$$

This means that asymptotically the difference between the two averages is no larger than $O(n^{-\epsilon})$ for some positive ϵ.

6 Conclusion

We have shown that the average depths of tries and suffix trees behave asymptotically likewise for a Markov model of order one. This result can be extended to any order of the Markov model.

An extended analysis should yield analogous results for both the variance and the limiting distribution of typical depth. A normal distribution is then expected for suffix trees.

In the future, we also hope to extend our results to the more general probabilistic source model introduced by Vallée in [6].

Acknowledgements

The authors thank Philippe Flajolet and Wojciech Szpankowski for serving as hosts during the project. Both authors are thankful to Wojciech Szpankowski for introducing the problem, and also to Philippe Flajolet and Mireille Régnier for useful discussions.

References

[1] Alberto Apostolico. The myriad virtues of suffix trees. In A. Apostolico and Z. Galil, editors, *Combinatorial Algorithms on Words*, volume 12 of *NATO Advance Science Institute Series*. Series F: Computer and Systems Sciences, pages 85–96. Springer Verlag, 1985.

[2] Philippe Flajolet, Xavier Gourdon, and Philippe Dumas. Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, 144(1–2):3–58, June 1995.

[3] P. Jacquet and W. Szpankowski. Analytic approach to pattern matching. In M. Lothaire, editor, *Applied Combinatorics on Words*, chapter 7. Cambridge, 2005.

[4] Philippe Jacquet and Wojciech Szpankowski. Analysis of digital tries with Markovian dependency. *IEEE Transactions on Information Theory*, 37(5):1470–1475, 1991.

[5] Mireille Régnier and Wojciech Szpankowski. On pattern frequency occurrences in a Markovian sequence. *Algorithmica*, 22(4):631–649, 1998.

[6] Brigitte Vallée. Dynamical sources in information theory: Fundamental intervals and word prefixes. *Algorithmica*, 29(1/2):262–306, 2001.

[7] Jacob Ziv and Abraham Lempel. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, IT-23:337–343, 1977.