An average-case analysis of basic parameters of the suffix tree

Julien Fayolle

ABSTRACT: *The* **LZ'77** *algorithm offers one of the best available rates for lossless data compression. It is based on the suffix tree structure. Our aim is to obtain the asymptotics of the mean size and external path length of a suffix tree by comparing them to those of a trie or digital tree. The core problem lies within the set on which we build the suffix tree. This set is correlated, so we cannot use the methods that have proved efficient for the trie. The proof relies on combinatorics, generating functions, and complex analysis.*

1 Introduction

The *trie* or digital tree data structure [7, 8, 9] manages efficiently dictionaries. Queries for an existing word or insertion of a new word in the dictionary can be performed in expected logarithmic time in the number of items stored in the dictionary.

While the **LZ'78** data compression algorithm is based on digital search trees, we will focus on *suffix trees*, a particular kind of trie which lies at the very heart of the popular and efficient **LZ'77** [11] lossless compression algorithm. This algorithm is behind the `gzip` software.

In a groundbreaking article, Jacquet and Szpankowski [6] have developped a sophisticated "string ruler" approach to obtain results on the asymptotics of parameters of suffix trees. This paper uses Jacquet and Szpankowski's lead idea: asymptotically the mean size and external path length for a trie built on n words and for a suffix tree built on n suffixes are very close.

This paper's aim is threefold: first to provide simpler proofs of Jacquet and Szpankowski's results than they do, second to obtain more accurate results, and third to lay the groundwork for a study of suffix trees under a broader model.

1.1 Tries

We first define recursively a trie on a set X of infinite words on the m-ary alphabet $\mathcal{A} = \{a_1, \cdots, a_m\}$ as

$$\text{trie}(X) = \begin{cases} \emptyset & \text{if } |X| = 0, \\ \bullet & \text{if } |X| = 1, \\ \langle \bullet, \text{trie}(X \backslash a_1), ..., \text{trie}(X \backslash a_m) \rangle & \text{else,} \end{cases}$$

where $X \backslash \alpha$ is defined as the set of words starting with the letter α whose first letter is removed.

From now on the alphabet will be binary $\mathcal{A} = \{\mathtt{0}, \mathtt{1}\}$, a choice that entails no loss of generality.

1.2 Source model

How do we obtain the infinite words constituing the set X? By a device, called *source*, producing randomly symbols from the binary alphabet regularly in time.

The type of source we are dealing with has two main characteristics: it is *probabilised*, symbols are emitted with probabilities; and *memoryless*, the emission of a symbol at a given time is independent of the symbols already emitted. In this work, the probability of occurrence of a symbol is independent of when it is emitted. This specifies the memoryless source model a.k.a. Bernoulli model.

Definition 1.1 *The probability that the source emits a sequence of symbols starting with the pattern w is noted p_w, and called occurrence probability. For a memoryless source, p_w is the product of the probabilities of the letters composing w.*

We note p the probability of emitting the symbol `0` and q the probability of emitting `1`. A source is said to be *symmetric* if $p = q = 1/2$ and *biased* otherwise. For convenience, we adopt the convention that p is the largest of the two probabilities.

1.3 Parameters under the spotlights

For each internal node of a trie, the successive left (encoded by 0) or right (encoded by 1) steps taken to go from the root to the node encode the prefix associated to this node. An internal node exists within the trie (relatively to an infinite complete binary tree) if there are at least two words in X starting by the prefix associated to this node.

For a pattern w, $N_w(X)$ is introduced as the number of words of X starting by w. Sometimes, we note N_w for $N_w(X)$.

Let the parameters S and P denote the size and the external path length. Both can be rewritten in terms of N_w for a trie built on the set X by

$$S(X) := \sum_{w \in \mathcal{A}^\star} [\![N_w(X) \geq 2]\!],$$
$$P(X) := \sum_{w \in \mathcal{A}^\star} N_w [\![N_w(X) \geq 2]\!],$$

where $[\![P]\!] = 1$ if P is true, 0 else. This is Iverson's bracket notation.

1.4 Suffix trees

Let y be an infinite word on the alphabet $\mathcal{A}$ and Y_n the set of the first n suffixes of y (we consider y to be its own suffix). The suffix tree of index n based on y is nothing but the trie built on Y_n (this operation is valid since Y_n is a set of infinite words over the alphabet $\mathcal{A}$).

Since tries and suffix trees are based on the same recursive decomposition, the expression for the parameters S and P under consideration are identical. For a trie on a set X, N_w means the number of words starting with the pattern w, so for a trie on the set Y_n (suffix tree on y) it coincides with the number of suffixes (amongst the n first of them) of the word y for which w is a prefix, or equivalently, the number of occurrences of the pattern w in the n first positions of the word y. We introduce $\widehat{N}_w(y; n)$ as the number of occurrences of the pattern w in the first

n positions of y, and we express the size S and the external path length P of a suffix tree as

$$S(y;n) := \sum_{w\in\mathcal{A}^\star} [\![\widehat{N}_w(y;n) \geq 2]\!], \tag{1}$$

$$P(y;n) := \sum_{w\in\mathcal{A}^\star} \widehat{N}_w(y;n)[\![\widehat{N}_w(y;n) \geq 2]\!] \tag{2}$$

1.5 Plan

We recall the results obtained by Knuth [7] on the mean size and external path length for a trie built on n strings:

$$\mathbb{E}_n^{\mathrm{t}}(S) = \frac{n}{h}(1+\epsilon'(n)) + O(\log n), \tag{3}$$

$$\mathbb{E}_n^{\mathrm{t}}(P) = \frac{n\log n}{h} + (K+\epsilon(n))n + O(\log n), \tag{4}$$

where ϵ and ϵ' are oscillating function of very small amplitude around 0. These results were proved in a less intricate way in [1].

The purpose of the next section is to obtain, *via* Guibas and Odlyzko's work and complex analysis, an asymptotic expression of the mean size, $\mathbb{E}_n(S)$, and mean external path length, $\mathbb{E}_n(P)$, for suffix trees built on n suffixes. There are two probabilistic models on tries: one can build them on a set of size n (fixed-size model a.k.a. *Bernoulli*) or on a set which size follows a *Poisson* law of parameter z (Poisson model). The difference between mean size for tries under Poisson of parameter n and Bernoulli of parameter n models is fairly small (of order $\log n$), this is also true for mean external path length. Section 3 is dedicated to studying the difference Δ between mean path length for tries under the Poisson model of parameter n, $\mathbb{E}^{\mathrm{t}}_{\mathcal{P}(n)}(P)$, and for a suffix tree built on n suffixes, $\mathbb{E}_n(P)$. Our aim is to show that Δ is small. Section 3 focuses on the external path length but the same techniques can be used for size.

2 Asymptotics

Since we know from the previous part the expression for the size S and the external path length P of a suffix tree, we write down the mean over suffix trees built with n suffixes for both parameters:

$$\begin{aligned}\mathbb{E}_n(S) &= \sum_{w\in\mathcal{M}^\star} \mathbb{E}_n([\![\widehat{N}_w \geq 2]\!]) = \sum_{w\in\mathcal{M}^\star} \mathbb{P}_n(\widehat{N}_w \geq 2)\\ &= \sum_{w\in\mathcal{M}^\star} 1 - \mathbb{P}_n(\widehat{N}_w = 0) - \mathbb{P}_n(\widehat{N}_w = 1),\end{aligned} \tag{5}$$

$$\mathbb{E}_n(P) = \sum_{w\in\mathcal{M}^\star} \mathbb{E}_n(\widehat{N}_w[\![\widehat{N}_w \geq 2]\!]) = \sum_{w\in\mathcal{M}^\star} \mathbb{E}_n(\widehat{N}_w) - \mathbb{P}_n(\widehat{N}_w = 1). \tag{6}$$

This part is dedicated to finding the asymptotic value for the two probabilities appearing in the formulæ.

2.1 Combinatorics

The expressions for the mean size and external path length of a suffix tree make use of the probabilities $\mathbb{P}_n(\widehat{N}_w = 1)$ and $\mathbb{P}_n(\widehat{N}_w = 0)$, that is the probabilities to obtain one (resp. zero) occurrence of the pattern w in the first $n + |w| - 1$ letters of an infinite word (*i.e.*, an occurrence of w starting at one of the first n letters of the infinite word).

In order to obtain these probabilities, we introduce ordinary generating functions counting the number of texts with a given number of occurrences of the pattern w according to their size: let $O_w(z) = \sum_{n\geq 0} o_n z^n$ be the ordinary generating function (ogf) counting texts with w occurring only once according to their size and $N_w(z)$ the ogf counting texts with no occurrence of the pattern w.

The possible overlap of the pattern w with itself causes problem in the enumeration of occurrences of a pattern in a text. This phenomenon is called *autocorrelation*. The pattern w of size k has an overlap of size j if $1 \leq j \leq k$ and the prefix of size j, P_j, and the suffix of size j, S_j, of w coincide ($P_j = S_j$). Graphically, an overlap of a pattern (white rectangles) looks like this:

,

where the two black boxes are the prefix and suffix matching one above the other. For example w=001001001 has overlaps of size 3 and 6; $|w|$ is always a valid overlap.

The autocorrelation of a pattern is encoded by the autocorrelation polynomials: the combinatorial one is

$$c_w(z) = \sum_{i=0}^{k-1} c_i z^i$$

and, since our symbols are produced by a probabilistic source, the probabilistic version that we need is

$$\widehat{c}(z) = \sum_{i=0}^{k-1} c_i z^i \mathbb{P}(w_{k-i+1} \cdots w_k)$$

where $c_i = [\![S_{k-i} = P_{k-i}]\!]$, meaning there is an overlap of size $k - i$.

For the memoryless sources we are dealing with, the probabilistic autocorrelation polynomial satisfies a neat and useful relation:

Lemma 2.1

$$\sum_{w\in\mathcal{M}^k} \widehat{c}_w(1) = 2^k + k - 1.$$

Proof: On a binary alphabet there are 2^j patterns of size j for any $j < k$ and from each of these v, one can build a unique w of size k such that $c_{j,w} = 1$ and v is its suffix of size j. So there are *at most* 2^j patterns of size k with a given suffix of size j and satisfying $c_j = 1$. Furthermore, there is no way two different suffixes

of size j can create the same word of size k, so there are *exactly* 2^j patterns of size k satisfying $c_j = 1$ and this for every j between 1 and $k-1$. Hence,

$$\sum_{w\in\mathcal{M}^k} \widehat{c}_w(1) = \sum_{w\in\mathcal{M}^k} \sum_{j=0}^{k-1} c_{j,w}\mathbb{P}(w_{k-i+1}\cdots w_k)$$
$$= \sum_{j=0}^{k-1} \sum_{w\in\mathcal{M}^k:\, c_{j,w}=1} \mathbb{P}(w_{k-i+1}\cdots w_k),$$

but since we just proved that the suffix of size j is enough to obtain a word w of size k with $c_j = 1$ (and $c_{0,w} = 1$ for all $w \in \mathcal{M}^k$),

$$\sum_{w\in\mathcal{M}^k} \widehat{c}_w(1) = 2^k + \sum_{j=1}^{k-1} \sum_{v\in\mathcal{M}^j} \mathbb{P}(v_1\cdots v_j)$$
$$= 2^k + \sum_{j=1}^{k-1} 1 = 2^k + k - 1 \quad \blacksquare$$

In [4, 5], Guibas and Odlyzko devised a combinatorial method based on formal languages in order to obtain the generating functions counting the number of texts with a fixed number of occurrences of a given pattern w according to their size. It leads to the generating function $O_w(z)$ and $N_w(z)$

$$O_w(z) = \frac{z^k}{(c(z)(1-2z)+z^k)^2},$$
$$N_w(z) = \frac{c(z)}{c(z)(1-2z)+z^k},$$

where k is the length of w.

When no ambiguity arises, we abbreviate $O_w(z)$ and $N_w(z)$ by $O(z)$ and $N(z)$. Basing ourself on Guibas and Odlyzko's method, we obtain the probabilistic versions $\mathcal{O}(z)$ and $\mathcal{N}(z)$ of $O(z)$ and $N(z)$, these generating functions count texts with their probability, so

$$\mathcal{O}(z) := \sum_{n\geq 0} \mathbb{P}_n(\widehat{N}_w = 1)z^n = \frac{p_w z^k}{(\widehat{c}(z)(1-z)+p_w z^k)^2}, \tag{7}$$

and

$$\mathcal{N}(z) := \sum_{n\geq 0} \mathbb{P}_n(\widehat{N}_w = 0)z^n = \frac{\widehat{c}(z)}{\widehat{c}(z)(1-z)+p_w z^k}. \tag{8}$$

The next step will be to extract the coefficient of order n in these probabilised generating functions.

2.2 Complex analysis

In order to isolate the dominant pole ρ of the generating functions $\mathcal{O}_w$ and $\mathcal{N}_w$, we use Rouché's theorem. The adequate contour is a circle $\mathcal{C}$ centered at the

origin with a radius R depending on the position of the first non-trivial 1 in w's autocorrelation polynomial.

For example, if $c_1 = 1$ we choose $R = 0.5(1 + 1/p)$, then the disc of radius R contains ρ as unique pole of the generating function. No general formula has yet been devised for the radius, but we are assured of its existence by Pringsheim's theorem.

The contour is also the one we use for an application of Cauchy's theorem. There is only one pole to the generating function $\mathcal{O}(z)$ (or $\mathcal{N}(z)$) inside the disc so that

$$\frac{1}{2i\pi}\int_{\mathcal{C}} \frac{\mathcal{O}(z)}{z^{n+1}}dz = \mathrm{Res}\left(\frac{\mathcal{O}(z)}{z^{n+1}}; 0\right) + \mathrm{Res}\left(\frac{\mathcal{O}(z)}{z^{n+1}}; \rho\right), \tag{9}$$

where $\mathrm{Res}(f, a)$ means the residue of the function f at a.

The modulus of $\mathcal{O}(z)$ can be bounded on the circle $\mathcal{C}$. Furthermore, the residue in 0 of $\mathcal{O}(z)/z^{n+1}$ is the coefficient of order n of $\mathcal{O}(z)$. By developping the generating function in Laurent serie near the pole ρ, we obtain the residue at ρ:

$$\mathrm{Res}\left(\frac{\mathcal{O}(z)}{z^{n+1}}; \rho\right) = \frac{2^{-k}}{F'(\rho)^3}\rho^{k-n-2}((k-(n+1))F'(\rho) - \rho F''(\rho)),$$

where $F(z) = \widehat{c}(z)(1-z) + p_w z^k$.

2.3 Approximation

This subsection is dedicated to finding a good approximation for the value of the dominant pole ρ. We recall that ρ is defined as the solution of smallest modulus of

$$F(z) = \widehat{c}(z)(1-z) + p_w z^k = 0. \tag{10}$$

Furthermore, due to Pringsheim's theorem [3], we know that ρ is positive real.

Since $F(1) = p_w$ is close to zero on all but a few patterns (the probability decreases exponentially with the size of the word w), and we are dealing with polynomials, ρ is greater than 1 but close enough to it.

Let's introduce α_1 such that $\rho = 1 + \alpha_1$. We know that α_1 is positive, and satisfies

$$\widehat{c}(1+\alpha_1)\alpha_1 + p_w(1+\alpha_1)^k = 0. \tag{11}$$

It is hard to solve this equation to get α_1, so we introduce α such that

$$\widehat{c}(1+\alpha)\alpha + p_w = 0.$$

α is close to α_1 since only small terms have been omitted from (11).

Using Rolle's theorem, we obtain

$$\widehat{c}(1+\alpha) = \widehat{c}(1) + \alpha\widehat{c}'(\beta) = \widehat{c}(1) + p_w\frac{\widehat{c}'(\beta)}{\widehat{c}(1+\alpha)},$$

for $\beta \in]1, 1+\alpha[$. Since the quantity $\widehat{c}'(\beta)/\widehat{c}(1+\alpha)$ is bounded by a constant, and p_w is very small compared to $\widehat{c}(1)$, one has

$$\widehat{c}(1+\alpha) \simeq \widehat{c}(1).$$

Finally, in the residue from the previous section, the ρ^{-n} term becomes

$$\rho^{-n} = (1+\alpha_1)^{-n} \simeq (1+\alpha)^{-n} \simeq \exp(-n\alpha) \simeq \exp\left(-\frac{np_w}{\widehat{c}(1)}\right).$$

3 Splitting in four

From now on, we only deal with the external path length parameter since the methods used for this parameter apply similarly to the size.

Let Δ be the difference between the mean of the external path length P for a trie under Poisson model of parameter n and for a suffix tree on n suffixes:

$$\Delta := \mathbb{E}^{\mathrm{t}}_{\mathcal{P}(n)}(P) - \mathbb{E}_n(P).$$

Collecting results from preceding sections, we obtain

$$\Delta = \sum_{w \in \mathcal{M}^\star} \mathbb{P}_n(\widehat{N}_w = 1) - \mathbb{P}^{\mathrm{t}}_{\mathcal{P}(n)}(N_w = 1),$$

where $\mathbb{P}^{\mathrm{t}}_{\mathcal{P}(n)}(N_w = 1)$ is the probability of the event $N_w = 1$ taken over all tries built on set of size z, where z follows the Poisson law $\mathcal{P}(n)$ of parameter n. The previous section has provided the asymptotic behavior of $\mathbb{P}_n(\widehat{N}_w = 1)$ and known results from poissonisation [2, 9] allow us to write informally

$$\Delta \simeq \sum_{w \in \mathcal{M}^\star} np_w \left(\exp\left(-\frac{np_w}{\widehat{c}(1)} \right) - \exp(-np_w) \right). \tag{12}$$

Remark: Since $\widehat{c}(1) \geq 1$, Δ is a positive quantity, hence asymptotically and on average the path length is longer for a trie than for a suffix tree.

The remainder of this section consists of a delicate subdivision of the set of all patterns in order to control the asymptotic growth that each of these subsets brings to the sum (12). In [6], it was shown that the sum Δ was of order $O(n^{1-\epsilon})$ for some unspecified $\epsilon > 0$, and we want to obtain an explicit bound.

The function $x \to x \exp(-x)$ dominates the behavior of the sum Δ. A perusal of this function's graph induces a three-part splitting of the set of all patterns on whether np_w tends to infinity, to zero or remains "almost constant"; the latter will also be cut into two according to how $\widehat{c}_w(1)$ is close to 1.

3.1 Small sizes

First, we focus on patterns of small sizes, which are in relatively small number. Bounding crudely their contribution to the sum Δ by the product of their number by the worse they grow will be sufficient to prove they do not contribute much in Δ's growth.

We define the small-sized words as those complying with

$$|w| = k \leq \frac{5}{6} \log_{1/q} n = \frac{5}{6} C_q \log n =: k_s(n),$$

for $C_q := (\log 1/q)^{-1}$.

Intuitively, a pattern of small size is one that satisfies $np_w \to \infty$; or in a more quantitative approach

$$np_w \geq n^{1/6}. \tag{13}$$

I call a *slice* of patterns the set of all patterns of a given length. It is more comfortable to handle slices, so the definition for small-sized patterns means those in slices satisfying (13).

We have a binary alphabet, thus the number of patterns of size smaller than $k_s(n)$ is of order $n^{5C_q/6}$ and for any small-sized pattern

$$\begin{aligned} np_w \left(\exp\left(-\frac{np_w}{\widehat{c}(1)} \right) - \exp(-np_w) \right) &\leq n^{1/6} \exp(-n^{1/6}/\widehat{c}(1)) \\ &\leq n^{1/6} \exp(-(1-p)n^{1/6}). \end{aligned}$$

Finally, the patterns of small size contribute to Δ less than

$$n^{5C_q/6} n^{1/6} \exp(-(1-p)n^{1/6}),$$

and due to the dominance of the exponential decrease, to $o(1)$ (this suffices for our goals).

3.2 Large sizes

This part deals with patterns of large size, defined as those whose respective slices satisfies the property $np_w \leq \frac{1}{\sqrt{n}}$. The intuition is to catch patterns with $np_w \to 0$, but we refine this condition quantitatively into $np_w \leq \frac{1}{\sqrt{n}}$ before resorting to slices.

These patterns indeed are of large sizes: for a symmetric source, for example, $np_w = n2^{-k} \leq 1/\sqrt{n}$, implies that k, the length of w, satisfies $k \geq 1.5 \log n$. This definition translates on the length of the patterns into

$$k \geq 1.5 \log_{1/p} n = 1.5 C_p \log n =: k_l(n).$$

With this definition, all large patterns obey $np_w \to 0$; so a Taylor expansion of the function $x \to x \exp(-x)$ near zero yields

$$\sum_{k \geq k_l(n)} \sum_{w \in \mathcal{M}^k} np_w \left(\exp\left(-\frac{np_w}{\widehat{c}(1)} \right) - \exp(-np_w) \right) \simeq \sum_{k \geq k_l(n)} \sum_{w \in \mathcal{M}^k} (np_w)^2 \left(1 - \frac{1}{\widehat{c}(1)} \right).$$

Obviously there are an infinite number of large patterns, which prevents us from using a brute-force majoration like for the small-sized patterns. However, one has

$$\begin{aligned} \sum_{w \in \mathcal{M}^k} p_w^2 &= \sum_{i=0}^{k} \binom{k}{i} (p^i q^{k-i})^2 \\ &= (p^2 + q^2)^k =: A_p^k, \end{aligned} \tag{14}$$

for a constant A_p smaller than 1 and depending only on p.

Furthermore we have already seen that $1 \leq \widehat{c}(1) \leq 1/(1-p)$, hence

$$\sum_{k \geq k_l(n)} \sum_{w \in \mathcal{M}^k} np_w \left(\exp\left(-\frac{np_w}{\widehat{c}(1)} \right) - \exp(-np_w) \right) = O\left(\sum_{k \geq k_l(n)} n^2 A_p^k \right). \tag{15}$$

The largest $\sum_{k \geq k_l(n)} n^2 A_p^k$ can grow is in the case of a symmetric source. This case brings asymptotically an $O(\sqrt{n})$ contribution to the sum Δ; but for other letter occurrence probabilities, we can improve up to a $O(1/n)$ growth.

3.3 Periodic patterns

We introduce $B_k := \{w : |w| = k, \quad \widehat{c}(1) \geq 1 + 2^{-k/2}\}$ as the set of *periodic* patterns of size k. This part aims at patterns of intermediate size (neither small nor large) with the additional constraint they are periodic. We will abusively refer to these patterns as periodic.

A periodic pattern has the first non-trivial 1 in its autocorrelation polynomial for a small index j, and therefore w is formed of repetitions of its suffix of length j. For these patterns, the second term in $\widehat{c}(1)$ is the probability of the suffix of size j of w. But since j is small, the probability is large and $\widehat{c}(1)$ is relatively far from 1.

There are relatively few periodic patterns in B_k:

Lemma 3.1

$$\#B_k < k2^{k/2}. \tag{16}$$

Proof: We start by partitionning the patterns of size k into two:

$$\sum_{w\in\mathcal{M}^k} \widehat{c}(1) = \sum_{w\in B_k} \widehat{c}(1) + \sum_{w\notin B_k} \widehat{c}(1).$$

For the patterns in B_k, $\widehat{c}(1) \geq 1+2^{-k/2}$ and for the others, $\widehat{c}(1) \geq 1$. Using Lemma 1, we get

$$\sum_{w\in\mathcal{M}^k} \widehat{c}(1) = 2^k + k - 1 \geq \#B_k.(1 + 2^{-k/2}) + 1.(2^k - B_k) = \#B_k 2^{-k/2} + 2^k \quad \blacksquare$$

From there we bound the contribution of intermediate and periodic patterns

$$\begin{aligned}
\Delta_p &:= \sum_{k=k_s(n)}^{k_l(n)} \sum_{w\in B_k} np_w \left(\exp\left(-\frac{np_w}{\widehat{c}(1)}\right) - \exp(-np_w)\right) \\
&\leq \sum_{k=k_s(n)}^{k_l(n)} \#B_k \max_{w\in B_k} \left\{np_w \exp\left(-\frac{np_w}{\widehat{c}(1)}\right)\right\} \\
&< K \sum_{k=k_s(n)}^{k_l(n)} k2^{k/2},
\end{aligned}$$

where K is any upper bound on the function $x \to x\exp(-x/\widehat{c}(1))$ on positive reals.

We finally obtain a contribution of order $O(n^{0.75C_p}\log n)$ for the periodic patterns. Since we are looking for a sublinear contribution to Δ, this necessitates $0.75C_p < 1$, hence $p < 2^{-0.75} \simeq .5946035575$.

The $2^{-0.75}$ limiting value for p depends on the arbitrary (but smaller than 1) factor defining the bound $k_l(n)$ (here for example this factor is 1.5). We could extend the boundary value of p to $1/\sqrt{2}$ at the expense of a worse error term.

3.4 Aperiodic patterns

The *aperiodic* patterns are those remaining, they are of intermediate sizes (between $k_l(n)$ and $k_s(n)$) and not belonging to the set B_k. For these, $\widehat{c}(1)$ is very close to 1 hence the difference between $np_w\exp(-np_w)$ and $np_w\exp(-np_w/\widehat{c}(1))$ is small.

Since $w \notin B_k$, one has

$$\frac{1}{\widehat{c}(1)} \geq \frac{1}{1+2^{-k/2}} \geq 1 - 2^{-k/2},$$

so that

$$np_w \left(\exp\left(-\frac{np_w}{\widehat{c}(1)} \right) - \exp(-np_w) \right) \leq np_w e^{-np_w} \left(e^{np_w 2^{-k/2}} - 1 \right).$$

We are going to use a Taylor expansion of the exponential function near zero, but in order for the expansion to apply we need $np_w 2^{-k/2} \to 0$ for all aperiodic patterns; this leads to the condition $p < p_0 \simeq 0.5469205467$, where p_0 is the unique real solution to

$$\left(\frac{p}{\sqrt{2}} \right)^{5/6} + p - 1 = 0.$$

So, for $p < 0.54$, we can use a Taylor expansion and since $B_k < 2^k$, we derive

$$\begin{aligned}
\Delta_a &:= \sum_{k=k_s(n)}^{k_l(n)} \sum_{w \notin B_k} np_w \left(\exp\left(-\frac{np_w}{\widehat{c}(1)} \right) - \exp(-np_w) \right) \\
&\leq \sum_{k=k_s(n)}^{k_l(n)} 2^k (np_w)^2 e^{-np_w} 2^{-k/2} \\
&\leq \sum_{k=k_s(n)}^{k_l(n)} \beta 2^{k/2},
\end{aligned}$$

where $\beta = 4e^{-2}$ is the maximum value of $x \to x^2 \exp(-x)$ over the positive reals.

Hence the contribution of the aperiodic patterns to Δ is $O(n^{0.75 C_p})$. Similarly to the periodic case, we could increase the upper bound on p up to $2 - \sqrt{2} \simeq 0.5857$ at the expense of a less precise error term.

4 Conclusion

Each subset of patterns contributes less than $O(n^{0.85})$ to the difference Δ. Hence the asymptotic for the external path length (resp. size) of a trie and of a suffix tree only differ by a small quantity. Therefore we have obtained:

Theorem 4.1 *For a suffix tree built on the first n suffixes of a string produced by a memoryless (p,q)-source, and for $p < 0.54$, the mean of the external path length satisfies asymptotically*

$$\frac{n \log n}{h} + (K + \epsilon(n))n + O(n^{0.85}), \tag{17}$$

and the size

$$\frac{n}{h}(1 + \epsilon'(n)) + O(n^{0.85}). \tag{18}$$

where ϵ and ϵ' are oscillating functions of very small modulus centered in 0.

Future research related to this work includes: providing a larger range for the probability p (if possible the whole [0,1] interval); applying this method to other parameters of the suffix tree like the fill-up level, the profile or the height; finally, extending the source model to the powerful dynamical framework introduced by Vallée [10], as it has been done for tries in [2].

References

[1] Clément, J. *Arbres Digitaux et Sources Dynamiques.* Thèse de doctorat, Université de Caen, Sept. 2000.

[2] Clément, J., Flajolet, P., and Vallée, B. Dynamical sources in information theory: A general analysis of trie structures. *Algorithmica 29*, 1/2 (2001), 307–369.

[3] Flajolet, P., and Sedgewick, R. The average case analysis of algorithms: Complex asymptotics and generating functions. Research Report 2026, Institut National de Recherche en Informatique et en Automatique, 1993. 100 pages.

[4] Guibas, L. J., and Odlyzko, A. M. Periods in strings. *Journal of Combinatorial Theory,* Series A *30* (1981), 19–42.

[5] Guibas, L. J., and Odlyzko, A. M. String overlaps, pattern matching, and nontransitive games. *Journal of Combinatorial Theory. Series A 30*, 2 (1981), 183–208.

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

[7] Knuth, D. E. *The Art of Computer Programming*, vol. 3: Sorting and Searching. Addison-Wesley, 1973.

[8] Mahmoud, H. M. *Evolution of Random Search Trees.* John Wiley, New York, 1992.

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

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

[11] Ziv, J., and Lempel, A. A universal algorithm for sequential data compression. *IEEE Transactions on Information Theory*, IT-23 (1977), 337–343.

Julien Fayolle
INRIA, Projet ALGO
78153 Le Chesnay Cedex
France
julien.fayolle@inria.fr