A Note on the Height of Suffix Trees

Luc Devroye

Wojciech Szpankowski
Purdue University, spa@cs.purdue.edu

Bonita Rais

Report Number:
89-905
A NOTE ON THE HEIGHT OF SUFFIX TREES

Luc Devroye
Wojciech Szpankowski
Bonita Rais

CSD-TR-905
September 1989
A NOTE ON THE HEIGHT OF SUFFIX TREES

Luc Devroye
School of Computer Science
McGill University

and

Wojciech Szpankowski and Bonita Rais
Department of Computer Science
Purdue University

Abstract

Consider a word in which the individual symbols are independent integers occurring with probabilities $p_i$, and let $H_n$ be the height of the suffix tree constructed from the first $n$ suffixes of this word. We show that $H_n$ is asymptotically close to $2 \log n / \log(1/\sum p_i^2)$ in many respects: the difference is $O(\log \log n)$ in probability, and the ratio tends to one in probability and in the mean.

Keywords and phrases: Suffix tree, height, trie hashing, analysis of algorithms, weak convergence.

CR Categories: 3.74, 5.25, 5.5.
1. Introduction.

Tries are efficient data structures that were developed and modified by Fredkin (1960), Knuth (1973), Larson (1978), Fagin, Nievergelt, Pippenger and Strong (1979), Litwin (1981, 1985), Aho, Hopcroft and Ullman (1983) and others. Multidimensional generalizations were given in Nievergelt, Hinterberger and Sevcik (1984) and Régnier (1985). One kind of trie, the suffix tree, is of particular utility in a variety of algorithms on strings (Aho, Hopcroft and Ullman (1975), McCreight (1976), Apostolico (1985)). However, except for the results in Apostolico and Szpankowski (1987), who give an upper bound on the expected height (see also Szpankowski (1988)), very little is known about the expected behavior of suffix trees. Also noteworthy is a result by Blumer, Ehrenfeucht and Haussler (1989) who obtained asymptotics for the expected size of the suffix tree under an equal probability model. The difficulty arises from the interdependence between the keys, which are suffixes of one string. In this note, we study the height of the suffix tree. The results of our analysis find applications in many areas (Aho, Hopcroft and Ullman (1975), Apostolico (1985)). For example, suffix trees are used as a unifying framework for a class of linear time sequential data compression, and they are employed to decide whether a word contains a square subword, to spot all such squares, to allocate statistics without overlap of all subwords in a textstring, and so forth. Consequences of our findings for an efficient design of algorithms are extensively discussed in Apostolico and Szpankowski (1988).

We consider an i.i.d. sequence $X_1, X_2, \ldots$ of integer-valued nonnegative random variables with $P(X_i = i) = p_i$ for $i = 0, 1, 2, \ldots$ and $\sum_i p_i = 1$. The $X_i$'s should be considered as symbols in some alphabet. We do not assume that the alphabet is finite, but we will assume that no $p_i$ is one, for otherwise all the symbols are identical with probability one. The suffixes $Y_i$ are obtained by forming the sequences $Y_i = \{X_i, X_{i+1}, \ldots\}$. The suffix tree based upon $Y_1, \ldots, Y_n$ is nothing but the trie obtained when the $Y_i$'s are used as words (for a definition of tries, see Knuth (1973); for a survey of recent results, see Szpankowski (1988)). Note however that we do not compress the trie as in a PATRICIA trie, i.e. no substrings are collapsed into one node.

In this note we study the height $H_n$ of the suffix tree, which is nothing but

$$H_n = \max_{i,j,1 \leq i,j \leq n} C_{ij},$$

where $C_{ij}$ is the length of the common prefix of $Y_i$ and $Y_j$, i.e. $C_{ij} = k$ if $(X_i, \ldots, X_{i+k-1}) = (X_j, \ldots, X_{j+k-1})$ and $X_{i+k} \neq X_{j+k}$. In the announcement and derivation of the results, we will need the metrics $||p||_r = (\sum_i p_i^r)^{1/r}, \quad 0 < r < \infty$ and $||p||_\infty = \max_i p_i$. Finally, to alleviate the notation, we define $Q = 1/||p||_2$.

**Theorem.**

For the suffix tree, $H_n/\log n \to 1/\log Q$ in probability. Also, for all $m \geq 1$, $E H_n^m \sim \log^m n/\log^m Q$.

We will prove this result using only elementary probability theoretical tools, such as the second moment method. Nevertheless, we will in fact be able to show that...
\( P\left( |H_n - \frac{\log n}{\log Q}| > (1+\varepsilon) \log \log n \right) \to 0 \)

for all \( \varepsilon > 0 \). Thus, the variations of \( H_n \) are at best of the order of \( \log \log n \).

It is interesting to note that the first asymptotic term \( (\log n / \log Q) \) is of the same order of magnitude as for the asymmetric trie obtained if the words \( Y_1, \ldots, Y_n \) had been i.i.d. (Pittel, 1985, 1986; Szpankowski, 1988). In 1985, Pittel showed that \( H_n / \log n \to 1 / \log Q \) almost surely, and in 1986, he showed that \( H_n - \log n / \log Q = \mathcal{O}(1) \) in probability. Other properties of the height of a trie under the independent model can be found in Yao (1980), Régnier (1981), Flajolet (1983), Devroye (1984), Pittel (1985, 1986), Jacquet and Régnier (1986), and Szpankowski (1988), who presents a survey of recent results. The reader is also referred to some other related papers such as Kirschenhofer and Prodinger (1986), Flajolet and Puech (1986), Flajolet and Sedgewick (1986) and Szpankowski (1988).

2. Preliminary results.

We present four simple lemmata. The first two are trivial. The third one is due to Apostolico and Szpankowski (1987).

**Lemma 1.**
\[
\|p\|_2^2 \leq \|p\|_2 \leq \|p\|_\infty .
\]

**Lemma 2.** For every \( r \geq 2 \) the following holds
\[
\|p\|_r \leq \|p\|_2
\]

**Proof of Lemma 2.**
Let us consider a function \( f(x) = \left( \sum p_i x \right)^{1/x} \) for \( x > 0 \). Then, it is easy to show that the first derivative of \( f(x) \) is negative for all \( x > 0 \). This completes the proof. For more details see Szpankowski (1988), and Karlin and Ost (1985).

**Lemma 3.**
For \( 0 < |i-j| = d < k \), we have
\[
P(C_{ij} \geq k) = \left( \sum p_i^{l+2} \right)^r \left( \sum p_i^{l+1} \right)^d - r,
\]
where \( l = \lfloor k/d \rfloor \) (\( \lfloor . \rfloor \) denotes the integer fraction), and \( r = k - dl = k \mod d \). Also, for \( |i-j| \geq k \), we have \( P(C_{ij} \geq k) = \|p\|_2^k \).

**Lemma 4.**
For \( 0 < |i-j| = d < k \), we have \( P(C_{ij} \geq k) \leq \|p\|_2^{k+d} \).
Proof of Lemma 4.

In the notation of Lemma 3, and using Lemma 2 we immediately obtain

\[ P(C_{ij} \geq k) = \left( \sum_i p_i^{i+2} \right)^{d-\gamma} \left( \sum_i p_i^{i+1} \right)^{d-\gamma} \leq \|p\|_2^{(i+2)\gamma+(i+1)(d-\gamma)} = \|p\|_2^{k+d}. \]

3. Proof of the Theorem.

We prove our theorem by showing two tight bounds for the height \( H_n \). Roughly speaking, we shall show that for every \( \varepsilon > 0 \) and large \( n \) the following holds: \( P(H_n > (1+\varepsilon)\log Qn) \to 0 \) as \( n \to \infty \) (upper bound), and \( P(H_n < (1-\varepsilon)\log Qn) \to 1 \) as \( n \to \infty \) (lower bound).

We start with an easier part of our proof, namely the upper bound. Assume that \( 2 \leq k \leq n-1 \). We have from Lemmata 2 and 4, and Bonferroni's inequality

\[
P(\max_{i \neq j} C_{ij} \geq k) \leq 2n \left( \sum_{d=1}^{k-1} P(C_{1,1+d} \geq k) + \sum_{d=k}^{n-1} P(C_{1,1+d} \geq k) \right) \\
\leq 2n \left( \sum_{d=1}^{k-1} \|p\|_2^{k+1} + \sum_{d=k}^{n-1} \|p\|_2^{2k} \right) \\
\leq 2n \left( \frac{\|p\|_2^{k+1}}{1-\|p\|_2} + n \|p\|_2^{2k} \right). \quad (1)
\]

This tends to zero provided that \( \|p\|_2 < 1 \) (this is always true) and that \( n \|p\|_2^{\frac{k}{2}} \to 0 \) (for this, it suffices that \( k = (\log n + \omega_n)/(-\log \|p\|_2) \), with \( \omega_n \to \infty \)). This establishes the weak upper bound of the Theorem. Note also that, by (1),

\[
E(\max_{i \neq j} C_{ij} \log(\frac{1}{\|p\|_2}) - \log n)^m \\
= \int_0^\infty P(\max_{i \neq j} C_{ij} \log(\frac{1}{\|p\|_2}) - \log n > u^{1/m}) \, du \\
\leq \int_0^\infty (\frac{2e^{-u^{1/m}}}{1-\|p\|_2} + \frac{2e^{-2u^{1/m}}}{\|p\|_2^{2}}) \, du < \infty.
\]

Thus, \( EH_n \leq (\log n + A(m,p))/\log Q \) for some finite constant \( A(m,p) \).

A matching lower bound is obtained by the second moment method. We will use a form due to Chung and Erdös (1952), which states that for events \( A_i \), we have

\[
P(\cup_i A_i) \geq \frac{(\sum_i P(A_i))^2}{\sum_i P(A_i) + \sum_{i \neq j} P(A_i \cap A_j)}.
\]

Let \( S \) be the collection of pairs of indices \((i,j)\) with \( 1 \leq i,j \leq n \), and \( |i-j| \geq k \). Let
$A_{ij} = [C_{ij} \geq k]$. Then

$$P(\max_{i,j} C_{ij} \geq k) \geq P(\bigcup_{(i,j) \in S} A_{ij})$$

$$\geq \frac{\left( \sum_{(i,j) \in S} P(A_{ij}) \right)^2}{\sum_{(i,j) \in S} P(A_{ij}) + \sum_{(i,j),(l,m) \in S:(i,j) \neq (l,m)} P(A_{ij} \cap A_{lm})}.$$  

To prove our lower bound it is enough to show that the probability on the RHS of the above tends to 1 for $k$ "slightly" larger than $\log_2 n$ ($k = \log_2 n + \omega(n)$). First we note that when $k = o(n)$, then

$$\sum_{(i,j) \in S} P(A_{ij}) = 4\|p\|_{\infty}^2 - n^2 \|p\|_2^2$$  

(Lemma 1).

We decompose the collection of pairs of pairs of indices $\{(i,j),(l,m)\}:(i,j) \in S \cap (l,m) \in S, (i,j) \neq (l,m)\}$ as follows into $I_1 \cap I_2 \cap I_3$: $I_1$ captures all members with $\min(1-i,l-1,1-j,l-1) \geq k$ and $\min(1-m-i,l-m-j,1-m-j) \geq k$. $I_2$ holds all members with either $\min(1-i,l-1,1-j,l-1) \geq k$ and $\min(1-m-i,l-m-j,1-m-j) < k$, or $\min(1-i,l-1,1-j,l-1) < k$ and $\min(1-m-i,l-m-j,1-m-j) \geq k$. Finally, $I_3$ collects all members with $\min(1-i,l-1,1-j,l-1) < k$ and $\min(1-m-i,l-m-j,1-m-j) < k$. By Lemmata 1 and 2,

$$\sum_{(i,j),(l,m) \in I_1} P(A_{ij} \cap A_{lm}) \leq n^4 \|p\|_2^4,$$

$$\sum_{(i,j),(l,m) \in I_2} P(A_{ij} \cap A_{lm}) \leq 8k n^3 \|p\|_2^k \|p\|_\infty \leq 8k n^3 \|p\|_2^3,$$

$$\sum_{(i,j),(l,m) \in I_3} P(A_{ij} \cap A_{lm}) \leq (4k)^2 n^2 \|p\|_2^2.$$  

If we choose $k$ such that $n \|p\|_2^k \rightarrow \infty$, then indeed

$$\sum_{(i,j),(l,m) \in S:(i,j) \neq (l,m)} P(A_{ij} \cap A_{lm}) \rightarrow n^4 \|p\|_2^4.$$  

Collecting all this shows that $P(H_n \geq k) \rightarrow 1$ when $n \rightarrow \infty$. Note that we can take $k = [(\log n - (1+\varepsilon)\log \log n)/(-\log \|p\|_2)]$ for $\varepsilon > 0$. Also,

$$EH_n \geq k P(H_n \geq k) - k$$

if $k$ is chosen as indicated. This concludes the proof of the lower bound and of the Theorem.  

4. References.


