1989

On the Height of Digital Trees and Related Problems

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

Report Number:
88-816

ON THE HEIGHT OF DIGITAL TREES
AND RELATED PROBLEMS

Wojciech Szpankowski

CSD-TR-816
November 1988
Revised May 1989

ON THE HEIGHT OF DIGITAL TREES AND RELATED PROBLEMS

Wojciech Szpankowski*

Department of Computer Science
Purdue University
West Lafayette, IN 47907, USA

Abstract

This paper studies in a probabilistic framework some topics concerning the way words (strings) can overlap, and relationship of this to the height of digital trees associated with this set of words. A word is defined as a random sequence of (possibly infinite) symbols over a finite alphabet. A key notion of an *alignment* matrix $\{C_{ij}\}_{i,j=1}^{n}$ is introduced where C_{ij} is the length of the longest string that is a prefix of the i-th and the j-th word. It is proved that the height of an associated digital tree is simply related to the alignment matrix through some order statistics. In particular, using this observation and proving some inequalities for order statistics, we establish that the height of a *digital trie* under an *independent model* (i.e., all words are statistically independent) is asymptotically equal to $2\log_{\alpha}n$ where n is the number of words stored in the trie and α is a parameter of the probabilistic model. This result is generalized in three directions, namely we consider b-tries, *Markovian model* (i.e., dependency among letters in a word), and a *dependent model* (i.e., dependency among words) In particular, when consecutive letters in a word are Markov dependent (Markovian model), then we demonstrate that the height converges in probability to $2\cdot\log_{\theta}n$ where θ is a parameter of the underlying Markov chain. On the other hand, for suffix trees which fall into the dependent model, we show that the height does not exceed $2\log_{\kappa}n$, where κ is a parameter of the probabilistic model. These results find plenty of applications in the analysis of data structures built over digital words.

* This research was supported by NSF grants NCR-8702115 and CCR-8900305, and in part by grant R01 LM05118 from National Library of Medicine, and by AFOSR grant 89NM407.

1. INTRODUCTION

Correlation on words are often studied through some associated data structures such as digital trees built over these words (e.g., radix tries, subword trees, suffix trees, etc. [1,2,3]). Digital trees are important in their own right due to many applications in computer science (e.g., searching and sorting [1,2], dynamic hashing [4,5], pattern matching algorithms [1,3], etc.) and telecommunications (e.g., coding, conflict resolution algorithms for broadcast communications [6,7,8], etc.). In this paper, we investigate the height of digital trees under different probabilistic models and show that the height is simply related to the longest common prefix of any two words stored in the tree. The key notion of an *alignment matrix* $\mathbf{C} = \{C_{ij}\}_{i,j=1}^{n}$ is introduced, where n is the number of words (keys, strings) and C_{ij} measures the overlap on the first symbols in the i-th and the j-th words. We shall study properties of the alignment C_{ij} in a probabilistic framework, that is, we assume that words (keys) form a random sequence of (possible infinite) symbols over a finite alphabet. The symbols occur independently or Markov dependently in a word, and in addition words might be statistically dependent (see Section 2).

By proving some theorems on *order statistics* (i.e, maximum) of *dependent* random variables (that is, alignments C_{ij}), we shall establish in this paper a new methodology to study the height of digital trees and some other related problems (e.g., the longest prefix of any pair of words, the longest substring that can be fully recopied, testing for square-free words, memory requirements in the extendible hashing [5, 16], optimization problems [23], and so forth [27]). In particular, we prove that for large n, the height H_n of a digital trie with independent keys is equal to $2 \log_\alpha n$ *in probability* where α is a parameter of the probabilistic model. This result is generalized in four directions. At first, we drop the assumption that the fixed number of keys (words) are stored in the trie, and we prove that under a Poisson distribution with parameter μ of keys the average height EH_μ is asymptotically equal to $2 \log_\alpha \mu$. Secondly, for digital tries that can store up to b words in external nodes (i.e., b-tries) we establish that the height H_n is asymptotically equal to $(1+1/b)\log_\beta n$, where β is a parameter of the model depending upon b. Then, we assume Markov dependency among consecutive letters, and establish that the height behaves asymptotically like $\log_\theta n$ where θ is reciprocal of the largest eigenvalue of the Schur product of the transition matrix for the underlying Markov chain. Finally, we consider a dependent model, that is, the case when keys (words) are statistically dependent (e.g., suffix tree [1,3]). We prove that the height in this case does not exceed $2\log_\kappa n$ for some κ.

The height of digital trees has been previously investigated in [2,5, 9-15]. In [5], Flajolet studied an independent model of binary symmetric b-tries. Based on some classical counting results in occupancy problems, Flajolet derived the asymptotic distribution of the height. Using complex analysis (e.g., Cauchy integral formula) he also found the average height of a trie. Jacquet and Regnier [9], extended Flajolet's result to binary asymmetric (i.e., symbols occur with different probabilities) tries. They have made extensive use of the Mellin transform technique. Devroye [10] analyzed binary symmetric tries (independent model again), and based on the occupancy problem he derived some inequalities on the asymptotic distribution of the height. The most general results were obtained by Pittel [11] (see also [12]), where general asymmetric tries (i.e., dependency among letters are allowed but not among words) with $b=1$ were investigated (in [12] $b>1$ was discussed but only under independent model). Unfortunately, the proofs in [11] are not constructive and the results are well hidden. For some more results, see also [13] and [14]. We note here that all results discussed so far have been established for independent models, that is, for statistically independent keys. To the best of our knowledge, the dependent models were only studied by Szpankowski [15], and Apostolico and Szpankowski [16]. In [16] the authors investigate the height of suffix trees.

Our approach to compute the height of digital trees is quite different in comparison with the ones established in [2,5, 9-14]. In contrast to the previous analyses, we use here some novel results from order statistics, and therefore avoid explicit computation of the height distribution.

In addition, the purpose of this paper is to establish solid methodology which can be applied to analyze different algorithms and data structures built over digital words. Therefore, we do not restrict ourselves to a particular data structure or algorithm, and rather focus on methodological aspects of the problem.

The paper is organized as follows. In the next section, we present our probabilistic framework. Section 3, the heart of this paper, presents our contribution to the analysis of some order statistics of dependent random variables, and contains our main results. Finally, Section 4 provides some generalizations of the results from Section 3, namely it presents the analysis of b-tries, Markovian model, and dependent models.

2. MODEL FORMULATION

In this section we build our probabilistic framework, which sets up a stochastic model for our studies. Let $\mathcal{A} = \{\omega_1, \omega_2, \ldots, \omega_V\}$ be an alphabet of V symbols, and let $\mathcal{S} = \{X_1, X_2, \ldots, X_n\}$ be a set of n (possibly infinite) strings (keys, words, sequences) over the alphabet $\mathcal{A}$. To characterize the stochastic model, we need to describe the probabilistic features of the set $\mathcal{S}$. In our basic probabilistic model, we assume:

(i) A word $X_k = x_k^1 x_k^2 \cdots$, is an infinite sequence of symbols from $\mathcal{A}$ such that it forms an independent sequence of Bernoulli trials with probability of sampling symbol ω_i equal to p_i, where $\sum_{i=1}^{V} p_i = 1$, that is, $p_i = Pr\{x_k^j = \omega_i\}$ for any k and j. If $p_1 = p_2 = \cdots = p_V = 1/V$, then the model is called *symmetric,* otherwise it is *asymmetric.*

(ii) The words $X_1, X_2, \ldots, X_n$ are statistically independent.

(iii) The number of words is fixed and equal to n.

These three assumptions form our basic probabilistic framework called the *Bernoulli* *model.* Some modifications of this basic model might be considered (see Section 4). For example, one can replace (iii) by more general assumption

(iii′) The number of keys is a random variable N with a probability distribution function $p(n) = Pr\{N = n\}$.

If $p(n)$ is Poisson distributed, then the model (i), (ii) and (iii′) is called the *Poisson model* (see Remark (ii) in Section 3). The next extension concerns assumption (i) since in some circumstances this assumption is too unrealistic. For example, if the alphabet $\mathcal{A}$ consists of English letters or $\mathcal{A}$ contains either four nucleotides or twenty amino acids (for DNA and proteins analysis, respectively [25, 26]), then there is a dependency between the occurrence of two consecutive symbols. In a more elaborate random model, the assumption (i) is replaced by

(i′) There is a *Markovian dependency* between neighboring symbols in a word $X_k = x_k^1 x_k^2 \cdots$, that is, the probability $p_{ij} = Pr\{x_k^l = \omega_j \mid x_k^{l+1} = \omega_i\}$, prescribes the conditional probability of sampling symbol ω_j following symbol ω_i.

The model (i′), (ii), (iii) or (iii′) is called *Markovian model.* A more sophisticated dependency may occur (see [11,12]). Note that the models discussed so far are very suitable for the analysis of digital search tries, since it is reasonable to assume that keys are independent (assumption (ii)). This is not the case, however, for suffix trees [1, 3] because the keys $X_2, X_3, \ldots, X_n$ are suffixes of the first key, hence strongly dependent. Therefore, we modify the assumption (ii) as follows.

(ii′) The keys $X_1, X_2, \ldots, X_n$ are dependent.

A probabilistic model containing assumption (ii′) is called *dependent model* in contrast to *independent model* when assumption (ii) is adopted.

The most popular data structure associated with a set of (digital) words (keys) is a *digital tree* [1,2]. Such a tree is built in a fairly natural way, that is, edges are labeled by symbols from the alphabet $\mathcal{A}$ and leaves (external nodes) contain the keys. The access path from the root to a leaf is a minimal prefix of information contained in the leaf. A brute force construction of such a tree is simple, that is, on the k-th level of the tree, we look at the k-th symbol, and if it is ω_1 we "go left" in the tree, if it is ω_2 then we "go next to the left", and so on. This process continues until all words $X_1, X_2, \ldots, X_n$ can be separated (distinguished) and the words are stored in external nodes. The following three examples present different types of digital trees.

EXAMPLE 2.1. *Radix tries*

Figure 1 shows $V = 3$-ary trie (see [1,2] for detailed definition of tries) built over alphabet $\mathcal{A} = \{0,1,2\}$ with $n = 6$ records (keys, words, strings) $A, B, \ldots, F$. The internal nodes

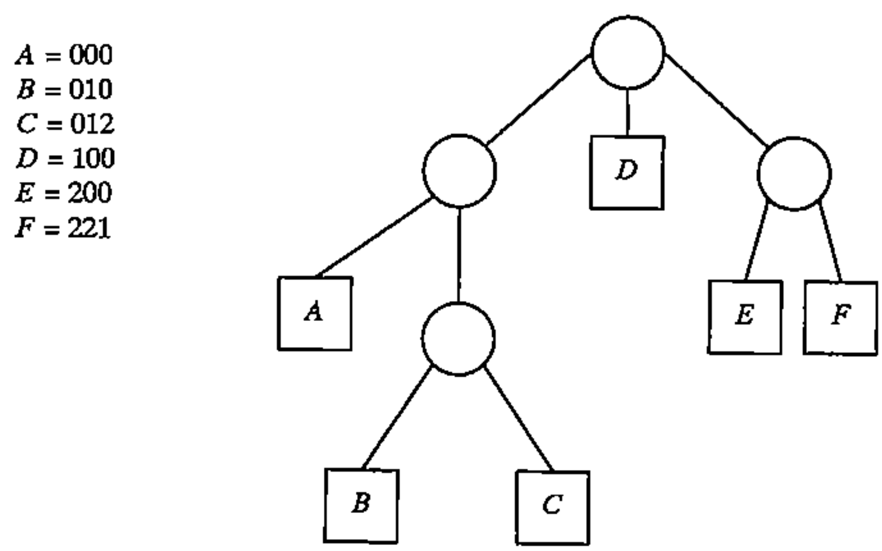

Figure 1. Example of a 3-ary digital trie with $n = 6$.

(circles in Figure 1) are used to branch keys, while external nodes (squares in the figure) contain the words.

EXAMPLE 2.2. *Suffix tree*

The purpose of this example is to present a digital tree illustrating the dependent model. We concentrate on the *suffix tree* [1,3], which is a data structure relatively often used in combinatorial algorithms on words [3]. Let $\mathcal{A} = \{a,b\}$ be a binary alphabet, and $X = abbabaa...$ a string. We build five suffixes of X, that is, $X_1 = X$, $X_2 = bbabaa...$, $X_3 = babaa...$ and so on (see Figure 2). The suffix tree constructed from the first five suffixes of X is shown in Figure 2.

$X_1 = abbabaa...$
$X_2 = bbabaa...$
$X_3 = babaa...$
$X_4 = abaa...$
$X_5 = baa...$

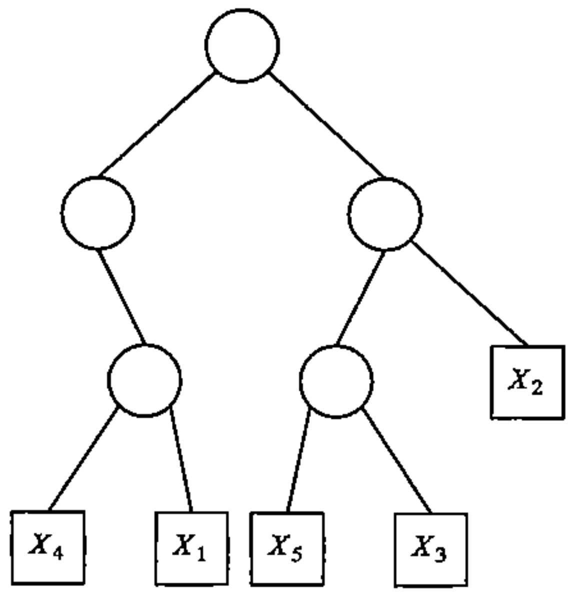

Figure 2. A suffix tree built from the first five suffixes of $X = abbabaa \cdots$.

EXAMPLE 2.3. *b-tries*

For keys $A, B, \ldots, F$ as in Example 2.1 we build a trie, but now we allow to store up to b keys in an external node. Such a digital tree is called b-trie.

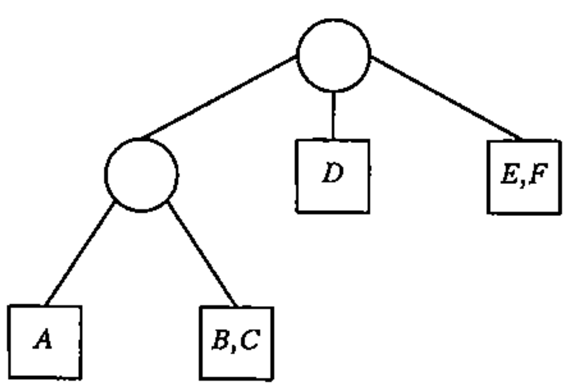

Figure 3. Example of a 3-ary digital 2-trie with $n = 6$.

In Figure 3 we show a 2-trie. Note that the average searching time for a key decreases in comparison to the standard trie shown in Figure 1, however, for searching one needs additionally to look up a linear list in an external node. □

Parameters of interest for digital trees are: depth of a leaf D_n, external path length L_n, height of the tree H_n and the shortest path h_n. We first introduce the depth of the i-th leaf $D_n^{(i)}$ which counts the number of edges from the root to the i-th leaf. Then, the above parameters are defined as follows

$$D_n = \frac{1}{n} \sum_{i=1}^{n} D_n^{(i)} \tag{2.1a}$$

$$L_n = \sum_{i=1}^{n} D_n^{(i)} \tag{2.1b}$$

$$H_n = \max_{1 \le i \le n} \{D_n^{(i)}\} \tag{2.1c}$$

$$h_n = \min_{1 \le i \le n} \{D_n^{(i)}\} \tag{2.1d}$$

The height H_n could be the most useful parameter in the analysis of algorithms since by definition it upper bounds other parameters (for L_n one must consider nH_n). Moreover, it is reasonable to believe that H_n, D_n and h_n have the same order of magnitude, whence the height is worth studying. We note, however, that the height is not a good measure of balancing property for trees (see [17] for more details). In this paper, we concentrate on establishing asymptotics for the height H_n. For b-tries, the depth D_n was extensively studied by Szpankowski in [17], external path length by Knuth [2], Kirschenhofer, Prodinger and Szpankowski [18] and the shortest path by Pittel [12].

In this paper we propose a novel approach (and some new results) to evaluate the height H_n of digital trees under different models discussed above. The key notion is the *alignment* matrix $\mathbf{C} = \{C_{ij}\}_{i,j=1}$. For every pair (i,j), $i \neq j$, $i,j = 1,2,\ldots,n$, we define alignment C_{ij} as the length of the longest string that is a prefix of both X_i and X_j. Thus, $C_{ij} = k$ iff X_i and X_j agree exactly on their first k symbols, but differ on their $(k+1)$-st. Then, the height H_n, the external path length L_n, and the shortest path h_n can be alternatively defined as (cf. (2.1)),

$$H_n = \max_{1 \le i < j \le n} \{C_{ij}\} + 1 \qquad (2.2a)$$

$$L_n = \sum_{i=1}^{n} \max_{1 \le j \le n} \{C_{ij}\} + n \qquad (2.2b)$$

$$h_n = \min_{1 \le i \le n} \{ \max_{1 \le j \le n} \{C_{ij}\}\} + 1 \qquad (2.2c)$$

Hereafter, we concentrate only on the height H_n (for applications of definition (2.2b) and (2.2c), see [16]). At first, however, we illustrate the new definitions by an example.

EXAMPLE 2.4. *Alignment matrix*

Let us reconsider the suffix tree from Example 2.2 (see also Figure 2). Then the corresponding alignment matrix $\mathbf{C} = \{C_{ij}\}$ is as follows:

$$\mathbf{C} = \begin{bmatrix} * & 0 & 0 & 2 & 0 \\ 0 & * & 1 & 0 & 1 \\ 0 & 1 & * & 0 & 2 \\ 2 & 0 & 0 & * & 0 \\ 0 & 1 & 2 & 0 & * \end{bmatrix}$$

From $\mathbf{C}$ and the expressions (2.2), we obtain $H_n = 3$, $h_n = 2$, $D_n = 14/5$ and $L_n = 14$. $\square$

In order to evaluate H_n, we note that by definition (2.2a) we need to estimate the maximum of $m = n(n-1)/2$ *dependent* random variables C_{ij}, $i < j = 1,2, \ldots, n$. The "maximum" is an example of an *order statistic* [19,20], and has been investigated vigorously over the last twenty years, however, most results concern *independent* random variables [19]. In the next section, we propose how to deal with dependent random variables C_{ij} (see also [27]), and we derive asymptotics for the height H_n.

3. MAIN RESULTS

In this section we derive various results concerning asymptotic behavior of the height H_n of a regular trie ($b = 1$) under our basic model assumptions (i)-(iii). In fact, as a side effect, we present also a fairly general approach to investigate asymptotic behavior of some order statistics for a class of *dependent* random variables.

By definition (2.2a), the height H_n of a digital trie is one plus a maximum of n^2 dependent random variables (alignments) C_{ij}. In fact, since $C_{ij} = C_{ji}$, we can reduce n^2 to $m = n(n-1)/2$ different alignments. It is relatively easy to evaluate the distribution function $F(k) = Pr\{C_{ij} \le k\}$ of the alignments C_{ij}. Note that all alignments C_{ij} are identically distributed, whence we drop indices i and j in the notation of the distribution function $F(k)$. Indeed, let us adopt our basic stochastic model consisting of assumptions (i)-(iii). In particular, assumptions (i) and (ii) immediately imply that C_{ij} is geometrically distributed with parameter $P = \sum_{i=1}^{V} p_i^2$, that is,

$$1 - F(k) = P^{k+1} \qquad k = 0,1, \ldots , \tag{3.1}$$

If alignments C_{ij} were independent random variables, then the knowledge of the distribution function $F(k)$ alone would be enough to compute the order statistics $\max_{1 \le i < j \le n} \{C_{ij}\}$ [19,20,21]. Otherwise, for computing the distribution of the maximum (whence the average, variance and so on), we normally need joint distributions. Fortunately, in some cases, to estimate *asymptotic behavior* of $\max\{C_{ij}\}$, the marginal distribution (3.1) is *almost* enough (see Lemma 2 and Lemma 3 below for more specific conditions). Using these methods we prove in this section our main results.

THEOREM. Suppose assumptions (i)-(iii) of our basic probabilistic model hold.

(i) Let $R = -\log P = -\log \sum_{i=1}^{V} p_i^2$, where log is the natural logarithm. Then

$$\lim_{n\to\infty} \frac{H_n}{\log n} = \frac{2}{R} \qquad \textit{in probability} \quad (pr.) \tag{3.2}$$

that is, for every $\varepsilon > 0$ the following holds $\lim_{n\to\infty} \Pr\{(1-\varepsilon)\cdot 2\log n/R \le H_n \le (1+\varepsilon)\cdot 2\log n/R\} = 1$.

In another notation, this means that $H_n = (1 + o(1))\cdot \log n^2/R$ (*pr.*).

(ii) The r-th moment EH_n^r of the height H_n for large n satisfies the following relationship

$$EH_n^r \sim (\frac{2}{R}\cdot\log n)^r \tag{3.3a}$$

where $\sim$ means asymptotically equivalent. In particular, the variance var H_n is

$$\operatorname{var} H_n = o(1)\log^2 n = o(\log^2 n) \tag{3.3b}$$

Another analysis that concentrates on proving convergence of H_n in distribution (see for example [12]), can lead to a better estimate of the variance, namely, it can be proved that $\operatorname{var} H_n \approx \pi^2/(6R) + 1/12$. ■

We prove the theorem in two steps by deriving an upper bound and then a lower bound on $\max\{C_{ij}\}$. One needs to notice that the alignments C_{ij} are dependent random variables. More precisely, C_{12} depends on $2n$ alignments C_{kl} where either k or l is equal to one or two, and C_{12} is independent for the rest $n^2/2-2n$ alignments C_{kl} with $k,l>2$. This observation suggests that we must compute some order statistics for *dependent* random variables. In the next three lemmas we suggest fairly general methods for establishing upper and lower bounds for asymptotic behavior of some order statistics. In Section 4, which deals with some generalization of the above model, we shall appreciate this general approach.

We start with an upper bound for some order statistics. Let $Y_1, Y_2, \ldots, Y_m$ be identically distributed random variables with the distribution function $F(\cdot)$. We assume that $F(\cdot)$ satisfies the following two conditions.

$$F(y) < 1 \quad \text{for all} \quad y < \infty \tag{3.4a}$$

$$\lim_{y\to\infty} \frac{1-F(cy)}{1-F(y)} = 0 \qquad \text{for} \qquad c>1 \tag{3.4b}$$

Let also a_m be the smallest root of the following equations

$$m[1-F(a_m)] = 1 \tag{3.5}$$

The next lemma establishes an upper bound for the maximum M_m of the random variables

$Y_1, \ldots, Y_m$, i.e., $M_m = \max\{Y_1, Y_2, \ldots, Y_m\}$.

Lemma 1. Let conditions (3.4) hold for a sequence $Y_1, Y_2, \ldots, Y_m$ of identically distributed random variables. Then, the maximum M_m satisfies

$$\lim_{n \to \infty} \frac{M_m}{a_m} \le 1 \qquad \textit{in probability} \tag{3.6}$$

that is, $\lim_{m \to \infty} \Pr\{M_m > (1+\varepsilon) a_m\} = 1$, where a_m is the root of equation (3.5).

Proof: We proceed as follows. Note first that Boole's inequality implies

$$\begin{aligned} Pr\{M_m > r\} = Pr\{Y_1 > r \ \text{ or } \ Y_2 > r \ \text{ or } \ \cdots \ \text{ or } \ Y_m > r\} = \\ \le m\, Pr\{Y_1 > r\} = m[1 - F(r)] \end{aligned}$$

that is,

$$Pr\{M_m > r\} \le \min\{1, m[1 - F(r)]\} \tag{3.7}$$

Let now $r = (1 + \varepsilon)a_m$ where ε is any positive number. Then quoting condition (3.4a), inequalities (3.7) becomes

$$Pr\{M_m > r\} \le m[1 - F((1 + \varepsilon)a_m)]$$

To complete the proof we must show that the RHS of of the above is $o(1)$ for large m. But, condition (3.4b) with $c = 1 + \varepsilon > 0$ and (3.5) imply

$$Pr\{M_m > (1 + \varepsilon)a_m\} \le m[1 - F((1 + \varepsilon)a_m)] = m \cdot o(1)[1 - F(a_m)] = o(1) \tag{3.8}$$

whence (3.6) follows. ■

The nice thing about Lemma 1 is that in order to establish an upper bound, we need only information about (marginal) distribution of Y's, and not the joint distribution $Pr\{Y_1 < r, Y_2 < r, \ldots, Y_m < r\}$. Unfortunately, this is not any longer true for lower bounds. The next two lemmas show how to establish lower bounds, but this time we need much more restrictive assumptions. For the next lemma, which is also called the *mixing condition approach*, we replace (3.4) by the following

$$\lim_{y \to \infty} \frac{1 - F(by)}{[1 - F(y)]^b} = \beta = \text{const} \qquad \text{for all} \quad b < 1 \tag{3.9}$$

In addition, we curb the joint distribution $Pr\{Y_1 < r, \ldots, Y_m < r\}$ by assuming existence of $\alpha(m) = O(m^{\kappa})$ for some constant κ such that

$$Pr\{Y_1 < r\ , Y_2 < r\ , \ldots,\ Y_m < r\} \le \alpha(m) \cdot [Pr\{Y_1 < r\}]^m = \alpha(m) \cdot F^m(r) \qquad (3.10)$$

Then, the following lemma can be proved.

Lemma 2. If condition (3.9) and (3.10) with $\alpha = O(m^{\kappa})$ hold, then

$$\lim_{m \to \infty} \inf \frac{M_m}{a_m} \ge 1 \qquad \textit{almost surely} \qquad (3.11)$$

where a_m is the smallest root of (3.5).

Proof. Let $r = (1-\varepsilon)a_m$ in (3.10), that is,

$$Pr\{M_m < (1-\varepsilon)a_m\} \le \alpha(m) F^m((1-\varepsilon)a_m) \qquad (3.12)$$

But, by (3.9) with $b = 1-\varepsilon$, one finds

$$1 - F((1-\varepsilon)a_m) = (1+o(1))\beta[1-F(a_m)]^{1-\varepsilon} = \frac{\beta(1+o(1))}{m^{1-\varepsilon}}$$

Substituting the above into (3.12), we show that

$$Pr\{M_m < (1-\varepsilon)a_m\} \le \alpha(m) \left[1 - \frac{\beta(1+o(1))m^{\varepsilon}}{m}\right]^m \le \alpha(m)\exp[-m^{\varepsilon}\beta(1+o(1))]$$

where the last inequality is the consequence of the fact that $(1-x/n)^n \le e^{-x}$ for $x/n \to 0$ as $n \to \infty$. Since $\alpha = O(m^{\kappa})$, then (3.11) follows from Borel-Cantelli Lemma [21]. ■

Before we leave this approach, we note that condition (3.10) in Lemma 2 can be replaced by a weaker one (but easier to prove), namely

$$\Pr\{Y_i < r\ , Y_j < r\} \le \alpha \cdot \Pr\{Y_i < r\}\Pr\{Y_j < r\} \qquad (3.10a)$$

for some $\alpha \le 1$.

The second method to establish a lower bound for M_m is based on the so-called *second moment method* [27,28]. We follow here the approach suggested in Aldous [27]. To recall, for a random variable $Z \ge 0$ such that $EZ^2 < \infty$, the following inequality is the basis for the second moment method

$$Pr\{Z > 0\} \geq \frac{(EZ)^2}{EZ^2} \tag{3.13}$$

Note that $Pr\{Z > 0\}$ tends to one, provided $(EZ)^2/EZ^2 \to 1$. This fact is used to derive the next lemma. Let us define for some sequence r_m the following quantity

$$\gamma(r_m) = \sum_{k=2}^{m} \frac{Pr\{Y_1 \geq r_m, Y_k \geq r_m\}}{m\, Pr^2\{Y_1 \geq r_m\}} \tag{3.14}$$

Then, the second moment method can be formulated as below.

Lemma 3. Suppose that $\lim_{m \to \infty} m[1 - F(r_m)] = \infty$ together with

$$\lim_{m \to \infty} \gamma(r_m) = 1 \tag{3.15}$$

Then, $\lim_{m \to \infty} Pr\{M_m \geq r_m\} = 1$ where $M_m = \max\{Y_1, Y_1, \ldots, Y_m\}$. In particular, if for every $\varepsilon > 0$, $r_m = (1 - \varepsilon)a_m$, where a_m is given in (3.5), and (3.15) holds, then $M_m/a_m \geq 1$ $(pr.)$ that is,

$$\lim_{m \to \infty} Pr\{M_m > (1 - \varepsilon)a_m\} = 1 \tag{3.16}$$

Proof: The proof follows immediately from Aldous [27], however, we present it for completeness. Define a set of events $\mathcal{B}_i = \{Y_i \geq r_m\}$, and consider $Z_m = \sum_{i=1}^{n} 1_{\mathcal{B}_i}$ where $1_{\mathcal{B}}$ is the indicator function of the event $\mathcal{B}$. To prove the lemma it suffices to note that $\{Z_m > 0\} = \{\bigcup_{i=1}^{m} \mathcal{B}_i\} = \{M_m \geq r_m\}$ and apply inequality (3.13). ■

Now we are ready to prove our Theorem. We note that the height H_n is maximum over $m = n(n-1)/2$ dependent random variables C_{ij}. By (3.1) we immediately find that the root a_n of (3.5) (we prefer to use here a_n instead of a_m, since $m \sim n^2$ and n is the original tree parameter) is

$$a_n = \frac{\log n(n-1)/2}{\log P^{-1}} = \frac{2 \cdot \log n}{\log P^{-1}} + O(1) = -2\log_P n + O(1) \tag{3.17}$$

To establish the upper bound for H_n, we just check that conditions (3.4a) and (3.4b) hold for the geometric distribution (3.1). This immediately proves that $H_n/\log n \le 2/R$ $(pr.)$

To prove the lower bound for H_n, we either use the mixing-condition approach (Lemma 2) or the second moment method (Lemma 3). In either case, we must compute the joint distribution of the alignments $\{C_{ij}\}$. In particular, one needs to evaluate $Pr\{C_{12} \ge r,\ C_{ij} \ge r\}$ for some $i,j \in \{1,2,\ldots,n\}$. We note that for $i,j > 2$, the above alignments are independent, that is, $Pr\{C_{12} \ge r\ ,\ C_{ij} \ge r\} = Pr\{C_{12} \ge r\} \cdot Pr\{C_{ij} \ge r\}$, provided $i,j > 2$. The dependency is among the first $2n$ random variables, that for $i=1$ or $j=1$. But, a simple probabilistic analysis reveals that

$$Pr\{C_{12} \ge r\ ,\ C_{1j} \ge r\} = (p^3+q^3)^r \tag{3.18}$$

(and the same holds for $j=1$). For symmetric case, i.e., $p=q=\frac{1}{2}$, we note that (3.18) implies $Pr\{C_{12} \ge r,\ C_{ij} \ge r\} = (1/4)^r = Pr\{C_{12} \ge r\} \cdot Pr\{C_{1j} \ge r\}$, hence Lemma 2 holds with $\alpha = 1$. The asymmetric case needs, however, a little different treatment. We appeal to Lemma 3. Set $m = n^2/2$ in (3.14), and by the above discussion, we split $\gamma(r_m)$ into two terms, namely

$$\gamma(r_m) = 2\sum_{k=3}^{n} \frac{Pr\{C_{12} \ge r_n\ ,\ C_{1k} \ge r_n\}}{n^2/2 \cdot \Pr\{C_{12} \ge r_n\}} + \frac{n^2/2 - 2n}{n^2/2} \tag{3.19}$$

The second term of the above is the consequence of the independence of C_{ij} and C_{12} for $i,j > 2$. To verify (3.15) we need only to prove, that the first term of (3.19), say $\gamma_1(r_n)$ tends to zero for appropriately chosen r_n. Now, as in (3.16) we assume $r_n = (1-\varepsilon)a_n$ where $a_n = -2\log_P n$ as in (3.17). To prove $\gamma_1(r_n) \to 0$ as $n \to \infty$, we need an upper bound for the joint distribution in the numerator of $\gamma_1(r_n)$. But, the following inequality can be easily proved

$$(p^3+q^3)^{\frac{1}{3}} \le (p^2+q^2)^{\frac{1}{2}} \tag{3.20}$$

Indeed, it is enough to note that the function $f(x) = (p^x + q^x)^{1/x}$, $p+q=1$, is decreasing for $x \ge 1$. Then, (3.18) and (3.20) imply

$$Pr\{C_{12} \ge (1-\varepsilon)a_n,\ \ C_{1k} \ge (1-\varepsilon)a_n\} \le n^{1-\varepsilon}\, Pr^2\{C_{12} \ge (1-\varepsilon)a_n\}, \tag{3.21a}$$

so

$$\gamma_1(1-\varepsilon)a_n) \le n^{2-\varepsilon}/n^2 \to 0 \quad \text{as} \quad n \to \infty. \tag{3.21b}$$

This proves the lower bound of H_n by appealing to Lemma 3, and it completes the proof of our Theorem (i). To establish the convergence in mean presented in Theorem (ii), one needs to show uniformly integrability of $\{H_n^r/(\log n)^r\}$. But this directly follows from the proof of Theorem 5 in [21] by noting that the alignments C_{ij} are geometrically distributed, (hence (3.4b) holds as needed in [21]). Finally, regarding our comments of the variance of H_n, that is, $\text{var}\, H_n \approx \pi^2/(6R)+1/12 = 1.6445/R+1/12$. This is a consequence of the limiting distribution of H_n which can be proved is equal to $Pr\{H_n < x\} = \approx \exp[-1/2n(n-1)P^x]$ (the proof of this fact is beyond the scope of this paper, and the reader is referred to [10] and [12]). The term 1/12 comes from a uniform correction.

Remarks

(i) *Second-order asymptotic approximation.* Our main Theorem of this section establishes first-order asymptotics (i.e., leading term) for the height H_n. A natural question arises, namely what are the next terms of the asymptotic approximation of H_n. Although our approach presented in Lemmas 1 to 3 limits the asymptotics to the leading factor, we may, however, comment on the other terms. Let us concentrate on the average height EH_n. In the next section, we prove (repeating arguments from Lai and Robbins [21]), the following bound (see also Section 4.3, Lemma 4)

$$EH_n \le a_n + \frac{1}{2}n(n-1)\sum_{k=a_n}^{\infty}[1-F(k)] \tag{3.22}$$

where a_n is given in (3.17). Using it and (3.1), we find

$$EH_n \le \frac{2}{R}\log n + 1 + \frac{1-\log 2}{R} + O(n^{-1}) \tag{3.23}$$

where as in Theorem $R = -\log\sum_{i=1}^{V} p_i^2$. How tight is this bound ? For binary symmetric tries

($R = \log 2$) Devroye [10] proved that

$$EH_n \le 2\log_2 n+1 + \frac{\gamma - \log 2}{\log 2} \tag{3.24}$$

hence the upper bound (3.23) is greater than (3.24) by 0.61. On the other hand, Flajolet [5] demonstrated that for binary symmetric tries

$$EH_n = 2\log_2 n + \frac{\gamma - \log 2}{\log 2} + P(\log n) + o(1) \tag{3.25}$$

where $P(\log n)$ is a periodic function with very small amplitude. The derivation of (3.24) and (3.25) require, however, much more advanced techniques. In both cases, the average H_n was obtained through the analysis of limiting distribution functions of H_n.

(ii) *Poisson model.* We replace assumption (iii) by (iii′), that is, we assume that the number of words (records) N stored in a trie is a random variable distributed according to Poisson with parameter μ. Let H_μ, H_n denote the heights in the Poisson and Bernoulli models, respectively. Restricting our analysis to r-th moments EH_μ^r of the height H_μ, we find out that

$$EH_\mu^r = \sum_{n=0}^{\infty} EH_n^r \frac{\mu^n}{n!} e^{-\mu} \tag{3.26}$$

where EH_n^r for Bernoulli model is discussed in our Theorem. In particular, for $r = 1$ we obtain

$$EH_\mu \le \frac{2}{R} e^{-\mu} \sum_{n=1}^{\infty} \log n \frac{\mu^n}{n!} + 1 + \frac{1 - \log 2}{R} \tag{3.27}$$

where in the above we explicitly used the upper bound (3.23). To evaluate the series in (3.27), we use the inequality $\log n \le \mathcal{H}_n$, where $\mathcal{H}_n$ is the n-th Harmonic number. Then, after some algebra and using some properties of the Harmonic numbers [24, p.79, Ex. 20] we prove

$$EH_\mu \le \frac{2}{R}\log\mu + \frac{E_1(\mu) + \gamma + 1 - \log 2}{R} + 1$$

where $E_1(\mu)$ is the *exponential integral* defined as $E_1(x) = \int_x^{\infty} e^{-t} t^{-1} dt$ ($|\arg x| < \pi$). A stronger result is obviously available. Referring to (3.3) in our Theorem and the above, one can easily prove that $EH_\mu \sim 2 \cdot \log \mu / R$. Finally, we note that this asymptotic approximation can be extended to some other distributions of keys.

(iii) *Almost sure convergence.* Using our approach we can prove some stronger results, namely that the convergence *in probability* of the height H_n can be replaced in our Theorem by *almost sure* convergence. According to Borel-Cantelli lemma, we need only to prove that

$$\sum_{n=1}^{\infty} Pr\{|H_n - a_n| > \varepsilon\} < \infty \tag{3.28}$$

where $a_n = 2/R \cdot \log n$. Proofs of our Theorem and Lemma 2 (cf. (3.8)) imply that

$$Pr\{|H_n - a_n| > \varepsilon\} < n^{-\varepsilon}. \tag{3.29}$$

Naturally, this bound by itself is not yet enough to show (3.28). But, selecting an appropriate subsequence of n in (3.28) will do the trick. Indeed, if we replace n in (3.28) by a subsequence $s(k) = m\,2^k$ for all $m \ge 1$ and note that H_n is a nondecreasing function of n, then one immediately proves (3.28). This is the main idea behind the proof of the *almost sure* convergence for H_n, and details can be found in Kingman [30, Sec. 3.1].

(iv) *More applications.* In the next section, we present some generalization of our theorems to more sophisticated digital trees. This, of course, does not limit the applications of our general approach expressed in Lemma 1 to 3. In fact, the results can be easily applied to analyze maximum queue length, traveling salesman problems, spanning tree problems, assignment problems and so on (for details see [23]). As mentioned in the introduction, we rather focus in this paper on methodology needed to establish the height of some digital trees (i.e., maximum of some dependent random variables). Therefore, we do not elaborate more on these applications.

4. GENERALIZATION

In this section we generalize our Theorem in three different directions by extending assumptions (i)-(iii) in our basic probabilistic model. At first, we shall investigate generalization of tries to b-tries (see Example 2.3). Then, we focus on the Markovian model (assumption (i')), and finally dependent models are considered (assumption (iii')). In particular, we present some preliminary results for suffix trees.

4.1 Analysis of b-tries

In this section we are still within our basic probabilistic model (assumptions (i)-(iii)), however in addition we assume that an external node can store up to b keys (words) (see Figure 3 in Example 2.3). Our interest is to compute the height H_n in such a b-trie. We need a generalization of the alignments. Let $X_1, X_2, \ldots, X_n$ be the keys, and for $i_1, i_2, \ldots, i_{b+1} \in \{1,2,\ldots,n\}$ we denote $C_{i_1 i_2 \cdots i_{b+1}}$ the common prefix for $X_{i_1}, \ldots, X_{i_{b+1}}$, i.e., the number of digits that $X_{i_1}, \ldots, X_{i_{b+1}}$ agree. Note that we have $\binom{n}{b+1}$ random variables $C_{i_1 i_2, \ldots, i_{b+1}}$, and as in (2.2a) the height H_n can be represented as

$$H_n = 1 + \max_{1 \le i_1 < \cdots < i_{b+1} \le n} \{C_{i_1 i_2, \ldots, i_{b+1}}\}$$

To evaluate H_n, we apply Lemma 1 and Lemma 3 so we need the distribution function of the alignments $C_{i_1 i_2, \ldots, i_{b+1}}$. But arguing as in Section 3 (see Eq. (3.1)), we immediately obtain

$$Pr\{C_{i_1 i_2 \cdots i_{b+1}} \ge k\} = P_b^k \qquad k = 0, 1, \ldots, \tag{4.1}$$

where $P_b = \sum_{l=1}^{V} p_l^{\,b+1}$. Again (4.1) is geometrically distributed, so condition (3.4) required for Lemma 1 is satisfied. Then, a_n defined in (3.5) becomes

$$a_n = \frac{\log \binom{n}{b+1}}{R_b}$$

where $R_b = -\log P_b = -\log \sum_{l=1}^{V} p_l^{\,b+1}$. But,

$$\binom{n}{b+1} = \frac{n^{b+1}}{(b+1)!}\,(1 + O(n^{-1}))$$

so

$$a_n = \frac{b+1}{R_b} \log n + O(1) \tag{4.2}$$

Therefore, by Lemma 1 we conclude that $H_n/\log n \le 2/R_b$ *(pr.)*, and the upper bound for the height is established.

In order to derive a lower bound for H_n we apply the second moment method from Lemma 3. The derivation goes along the same line as in the proof of our main Theorem, so we would rather present only a sketch of the analysis. In particular, in order to verify (3.15) we must evaluate the joint distribution $Pr\{C_{1,2,\ldots,b+1} > r_n \ , \ C_{i_1 i_2,\ldots,i_{b+1}} > r_n\}$. This probability depends on the cardinality of the set $\mathcal{D} = \{1,2,\ldots,b+1\} \cap \{i_1,i_2,\ldots,i_{b+1}\}$. If $\mathcal{D} = \emptyset$ ($\emptyset$ means empty set), then the events $\{C_{1,2,\ldots,b+1} > r_n\}$ and $\{C_{i_1 i_2,\ldots,i_{b+1}} > r_k\}$ are independent, and as in the case $b=1$ the contribution of it to $\gamma(r_n)$ is $[n^b - O(n^b)]/n^b \to 1$ as $n \to \infty$. For $|\mathcal{D}| = k > 0$ (i.e., there are k common indices), we can easily find that

$$Pr\{C_{1,2,\ldots,b+1} \ge r_n \ , \ C_{i_1,i_2,\ldots,i_{b+1}} \ge r_n\} = (p^{b+1+k} + p^{b+1+k})^{r_n} \le (p^{b+2} + q^{b+2})^{r_n}$$

Using the following inequality $(p^{b+2} + q_{b+2})^{1/(b+2)} \le (p^{b+1} + q^{b+1})^{1/(b+1)}$ (see (3.20)) we show, as before, that for $r_n = (1-\varepsilon)a_n$, with a_n given in (4.2), the above joint distribution can be upper bounded as

$$Pr\{C_{1,2,\ldots,b+1} \ge (1-\varepsilon)a_n \ , \ C_{i_1,i_2,\ldots,i_{b+1}} \ge (1-\varepsilon)a_n\} \le n^{-b(1-\varepsilon)} \cdot Pr^2\{C_{1,2,\ldots,b+1} > (1-\varepsilon)a_n\}$$

This implies that the contribution $\gamma_1(r_n)$ of the dependent alignments is upper bounded by $\gamma_1(1-\varepsilon)a_n) \le n^{b(1-\varepsilon)}/n^b \to 0$ as $n \to \infty$, and this completes the verification of (3.15). Hence, by Lemma 3 $H_n/\log n \ge (b+1)/R_b$ *(pr.)*, and together with the upper bound proved above, we finally show that

$$\lim_{n\to\infty} \frac{H_n}{\log n} = \frac{b+1}{R_b} \quad (pr.) \tag{4.3}$$

The appropriate convergence *in mean* (see Eq. (3.3)) works too. In particular, for symmetric case we obtain from (4.3)

$$\lim_{n \to \infty} \frac{EH_n}{\log_V n} = \frac{b+1}{b}$$

which directly generalizes Flajolet's result [5] to V-ary b-tries.

4.2 Markovian Model

We again assume $b=1$ (for simplicity of further analysis), but we allow Markovian dependency among the consecutive letters as postulated in assumption (i′) which replaces assumption (i). In particular, we denote by $P = \{p_{ij}\}_{i,j=1}^{V}$, the transition matrix for the underlying Markov chain. The analysis in this case does not differ significantly from what we have seen in Section 3. The major problem lies in the evaluation of the distributions $Pr\{C_{ij} \ge k\}$ and $Pr\{C_{12} \ge k, C_{ij} \ge k\}$, but a literature (cf. [27, 29]) contains necessary mathematics.

We start with the upper bound, hence we need to evaluate $1 - F(k) = Pr\{C_{ij} \ge k\}$ for large k. Let $\pi = [\pi_1, \pi_2, \ldots, \pi_V]$ be the stationary vector associated with the Markov matrix $P = \{p_{ij}\}_{i,j=1}^{V}$. Then, one easily shows (cf. [29])

$$Pr\{C_{ij} \ge k\} = \sum_{\{j_1, j_2, \ldots, j_k\}} [\pi_{j_1} p_{j_1 j_2}, \ldots, p_{j_{k-1} j_k}]^2 \tag{4.4}$$

and the sum is over all $1 \le j_i \le V$. In short, (4.4) can be written as the inner product of $\pi^2 = [\pi_1^2, \ldots, \pi_V^2]$ and $P_{[2]}^{k-1}\mathbf{u}$ where $P_{[2]} = P \circ P$ is Schur power of the matrix P (that is, elementwise product), and $\mathbf{u} = (1, 1, \ldots, 1)$ (cf. [29]). This compact representation suggests to apply Perror-Frobenius theory [27] to $P_{[2]}$ in order to show that for large k [27,29]

$$Pr\{C_{ij} \ge k\} = 1 - F(k-1) \sim \beta \theta_{[2]}^{k} \tag{4.5}$$

where $\theta_{[2]}$ is the largest eigenvalue of $P_{[2]}$, and β is a constant. This asymptotics provide enough information to apply Lemma 1. In particular, solving (3.5) one proves that

$$a_n \sim 2 \log_{\theta_{[2]}} n^{-1} \tag{4.6}$$

and by Lemma 1, we obtain the following upper bound

$$H_n / 2 \log_{\theta_{[2]}} n^{-1} \le 1 \quad (pr.) \tag{4.7}$$

for the height H_n.

The lower bound, surprisingly, is not difficult to prove too, since most of our arguments from Section 3 can be adopted here. We apply the second moment method, so one needs to verify (3.15). As before, we split the sum $\gamma(r_n)$ into two terms as (3.19) shows. To prove $\gamma(r_n) \to 1$ for $r_n = (1-\varepsilon)a_n$ it suffices to show that the first term $\gamma_1(r_n)$ in (3.19) tends to zero for $n \to \infty$. We need to compute the joint distribution $Pr\{C_{12} \ge r_n,\ C_{ij} \ge r_n\}$.

Let us concentrate for a moment on $Pr\{C_{12} \ge k,\ C_{1j} \ge k\}$. We note that the event $\{C_{12} \ge k,\ C_{1j} \ge k\}$ can be interpreted as the requirement that the common word (prefix) of the following three strings X_1, X_2 and X_j has length at least k. This falls exactly into the analysis of the longest common aligned word found in r sequences (in our case $r=3$) presented by Karlin and Ost in [29]. Naturally, a simple extension of (4.4) leads to

$$Pr\{C_{12} \ge k,\ C_{ij} \ge k\} = \sum_{\{j_1,\ldots,j_k\}} [\pi_{j_1} p_{j_1 j_2}, \ldots, p_{j_{k-1},j_k}]^3 \tag{4.8}$$

or in a compact representation

$$Pr\{C_{12} \ge k,\ \ C_{1j} \ge k\} = \langle \pi^3, P_{[3]}^{k-1}\mathbf{u}\rangle$$

where <x,y> is the inner product of $\mathbf{x}$ and $\mathbf{y}$. In particular, the above suggests that the largest eigenvalue $\theta_{[3]}$ of Schur product $P_{[3]} = P \circ P \circ P$ must be considered. Naturally, for large k

$$Pr\{C_{12} \ge k,\ \ C_{1j} \ge k\} \sim \beta' \theta_{[3]}^k$$

To complete our proof, we need to show that the first term in $\gamma(r_n)$, namely $\gamma_1(r_n) = \sum_{k=3}^{n} Pr\{C_{12} \ge r_n,\ \ C_{1k} \ge r_n\} / (n^2 \cdot Pr\{C_{12} \ge r_n)\} \sim\ \ \theta_{[3]}^{r_n} / (n\theta_{[2]}^{r_n})$ tends to zero for appropriately chosen r_n. Let $r_n = (1-\varepsilon)a_n$ where a_n is given in (4.5). In [29] it is proved that $(\theta_{[m]})^{1/m}$ is a decreasing function of m, hence $\theta_{[3]} \le \theta_{[2]}^{-1/2}\ \theta_{[2]}^2$ and finally

$$\gamma_1(r_n) \sim \frac{n^{2(1-\varepsilon)}}{n^2} \to 0 \text{ as } \ n \to \infty$$

as needed (see also (3.21b)). By Lemma 3, we prove that $H_n/2 \log_{\theta_{[2]}} n^{-1} \ge 1$ $(pr.)$, and together with (4.2) it gives our final result, namely

$$\lim_{n \to \infty} \frac{H_n}{\log_{\theta_{[3]}} n^{-1}} = 2 \quad (pr.) \tag{4.9}$$

Interestingly enough, this result can be extended to a more general dependency than Markovian. The crucial thing is to obtain the estimate suggested in (4.5). For more details, see [29].

4.3 Dependent model

In many applications keys (words) are statistically dependent, e.g., in DNA and RNA structures [25,26], in suffix tree [1,3], and so on. In this subsection, we relax assumption (ii) by adopting (ii′) and keeping the others unchanged (with $b=1$). We consider two examples. In the first, we assume only statistical dependency between directly aligned symbols in any two words. In the next (more realistic) example, we analyze suffix tree (see Example 2.2) in which keys are suffixes of a random word. We note also that in dependent models, the alignments are very rarely stationary (identically distributed), whence our Lemma 1 and 2 cannot be directly applied. In addition, analytical difficulties rapidly build up, so we restrict our interest to the average value of the height H_n.

Let us start with our first dependent model and let x_k^i, $x_\ell{}^i$ denote the i-th digits in the k-th and the ℓ-th keys. We assume that there is a dependency between x_k^i, $x_\ell{}^i$, which we express in terms of the joint distribution, that is,

$$p_{n,m}(k,\ell) = Pr\{x_k^i = \omega_n,\ \ x_\ell{}^i = \omega_m\} < 1 \tag{4.10}$$

where $k, \ell = 1,2,\ldots, n$, and $\omega_n, \omega_m \in \mathcal{A}$. Therefore, the alignment $C_{k\ell}$ is geometrically distributed with parameter $P_{k\ell} = \sum_{\ell-1}^{V} p_{ii}^2(k,\ell)$. Note, however, that this time the alignments $C_{k\ell}$ are not identically distributed, so Lemma 1 and Lemma 2 cannot be applied. We use the following result, which is a slight generalization of Lai and Robbins idea [21].

Lemma 4. Let $Y_1, Y_2, \ldots, Y_m$ be a sequence of random variables with distribution functions

$F_1(y), F_2(y), \ldots, F_m(y)$, respectively. Let $R_i(y) = Pr\{Y_i \ge y\}$ be the *complement* function of the distribution function $F_i(y)$ (function $R(\cdot)$ is sometimes called the *reliability* function). Finally, let $M_m = \max_{1 \le i \le m} Y_i$. Then if a_m is a solution of

$$\sum_{k=1}^{m} R_k(a_m) = 1, \tag{4.11}$$

then

$$EM_m \le a_m + \sum_{k=1}^{m} \sum_{j=a_m}^{\infty} R_k(j). \tag{4.12}$$

Proof: (i) Observe that, for any a (cf. [21])

$$M_m \le a + \sum_{k=1}^{m} [Y_k - a]^+ \tag{4.13}$$

where t^+ denotes $\max\{0, t\}$. Since $[Y_k - a]^+$ is a nonnegative random variable, then [22] $E[Y_k - a]^+ = \int_a^{\infty} R_k(y)dy$, so that (assuming for simplicity that Y_i is a continuous random variable) (4.13) implies

$$EM_m \le a + \sum_{k=1}^{m} \int_a^{\infty} R_k(x)dx \tag{4.14}$$

Minimizing the right-hand side (RHS) of (4.14) with respect to a, yields (4.11) and (4.12) with the optimal a_m given by (4.11). ■

To study the height H_n of a digital tree, we use our basic relationship between the height and the alignments, namely $H_n = \max_{1 \le k \le \ell \le n} \{C_{k\ell}\} + 1$, that is, H_n is maximum over $m \sim n^2$ (not necessary identically) distributed random variables. Let $F_{k\ell}(j)$ be the distribution function of $C_{k\ell}$ and our assumptions imply $F_{k\ell}(j) = 1 - P_{k\ell}^{j+1}$ where $P_{k\ell} = \sum_{\ell=1}^{V} p_{ii}^2(k, \ell)$. Then, by Lemma 4

$$EH_n \le a_n + 1 + \sum_{k,\ell=1}^{n} \sum_{j=a_n}^{\infty} [1 - F_{k\ell}(j)] \tag{4.15}$$

where $m = n(n-1)/2$. The RHS of (4.15) is minimized for such a_n that

$$\sum_{k=1}^{n} \sum_{\ell=k+1}^{n} R_{k\ell}(a_n) = 1 \tag{4.16}$$

For the geometric distribution with parameter $P_{k\ell}$ (4.16) becomes

$$\sum_{k=1}^{n} \sum_{\ell=k+1}^{n} P_{k\ell}^{a_n+1} = 1 \tag{4.17}$$

Let $P_{\max} = \max_{k,\ell} P_{k\ell}$, then one proves that

$$a_n \le \frac{\log m}{\log P_{\max}^{-1}}$$

where $m \sim n^2$. Showing that the contribution of the sum in (4.14) is O(1) we finally obtain

$$EH_n \le \frac{2}{R_{\min}} \log n + O(1) \tag{4.18}$$

where $R_{\min} = -\log P_{\max}$. We also point out that assumption $p_{n,m}(k,\ell) < 1$ is important. For example, if one builds a prefix tree (i.e., the k-th key is the prefix of the $(k+1)$-st key), then the height is obviously equal to n. But in this case $p_{n,m}(k,\ell)$ is either zero or one, so the restriction imposed in (4.10) is violated.

Finally, we consider one more sophisticated digital tree, namely a suffix tree [1, 3]. As shown in Example 2.2, a suffix tree is constructed from a random sequence X of symbols by taking the first n suffixes of X. Naturally, such a tree falls into the dependent model, and the i-th symbol in the k-th suffix depends on an j-th $(j < i)$ symbol in the ℓ-th suffix, $(\ell < k)$. To investigate the average height of the tree, we again apply Lemma 4. However, the major problem this time, is the computation of the distribution of the alignments C_{ij}. It is not difficult to observe that the distribution of C_{ij} varies with i and j in a way that depends on the differences $d = |j-i|$, rather than on the specific individual values of i and j. In other words, all random variables C_{ij} having the same value of $d = |j-i|$, have the same distribution. Thus, it is appropriate to reason in terms of the random variables C_d, where $d = 1, 2, \ldots, n-1$. For example, $C_{1,2}, C_{2,3}, \ldots, C_{n-1,n}$ have the same distribution, and are thus clustered in the new random variable C_1 (i.e., $d = 1$).

The distribution of C_d was evaluated by Apostolico and Szpankowski in [16]. In particular, they have proved that the complement function $R_d(\cdot)$ of the distribution function has the following form

$$Pr\{C_d \geq k\} = R_d(k) = \left\{\sum_{i=1}^{V} p_i^{\ell+2}\right\}^r \left\{\sum_{i=1}^{V} p_i^{\ell+1}\right\}^{d-r} \tag{4.19}$$

where k has a unique decomposition as $k = d\ell + r$ where $r < d$ and $\ell = 0, 1, \cdots$. Knowing $R_d(k)$ we can apply Lemma 4 to compute the height $H_n = \max\{C_{ij}\} + 1$ of a random suffix tree. In particular, we must solve (4.11) which in our case becomes

$$\sum_{d=1}^{n} (n-d) R_d(a_n) = 1 \tag{4.20}$$

Then, according to (4.12)

$$EH_n \leq a_n + \sum_{j=a_n}^{\infty} \sum_{d=1}^{n} (n-d) R_d(j) \tag{4.21}$$

It is not difficult to notice that (4.20) implies that the sum in (4.21) is $o(a_n)$. So we concentrate on computing a_n, and for simplicity we consider only binary case.

The asymptotic solution of (4.20) needs some work, however, a rude upper bound for a_n is immediately available. Indeed, noting that

$$R_d(k) \leq (p^{f+1} + q^{f+1})^d \tag{4.22}$$

where $f = \lfloor k/d \rfloor$ and $\lfloor \cdot \rfloor$ denotes the floor function, one shows after some simple algebra (cf. [16]) that

$$a_n \leq \frac{2}{\log p_{\max}^{-1}} \log n + O(1) \tag{4.23}$$

where $p_{\max} = \max_{1 \leq i \leq m} \{p_i\}$. To obtain more accurate estimate of a_n we first note that for $d > k$ the function $R_d(k)$ in (4.22) reduces to $R_d(k) = (p^2 + q^2)^k = P^k$, hence (4.20) can be rewritten as

$$1 \approx \sum_{d=1}^{\lfloor a_n \rfloor} (n-d) R_d(a_n) + \frac{n^2}{2} P^{a_n}$$

This can be easily solved asymptotically so that a_n becomes

$$a_n = \frac{2}{\log P^{-1}} \log n + O(\log n / n^{\delta}) \qquad (4.24)$$

for some positive δ. Details can be found in [16]. This and (4.21) establish a tight upper bound on the average height of a suffix tree built from a random string of characters.

A question arises whether a matching lower bound can be proved. Fortunately, Devroye, Szpankowski and Rais [31] have recently shown (using the second moment method) the matching lower bound, thus establishing the following remarkable result

$$\lim_{n \to \infty} \frac{H_n}{\log n} = \frac{2}{R} \qquad (pr.) \qquad (4.25)$$

Note that (4.25) proves that the suffix tree model is asymptotically equivalent to the independent model. We note, however, that the second leading factor for the suffix model is different than in the case of independent model (see Theorem (i)).

ACKNOWLEDGMENT

The author would like to thank David Aldous, University of California at Berkeley, for many fruitful discussions (by email) that led to better understanding of the behavior of order statistics for weakly dependent random variables. He also appreciates all comments of an anonymous referee that led to elimination of some slips in the previous versions of the paper, and to an improvement of the presentation. At last but not least the author thanks Alberto Apostolico, Purdue University for showing to him the beauty of the "stringology".

REFERENCES

[1] A.V. Aho, J.E. Hopcroft and J.D. Ullman, *The Design and Analysis of Computer Algorithms,* Addison-Wesley (1974).

[2] D. Knuth, *The Art of Computer Programming. Sorting and Searching,* vol. III, Addison-Wesley (1973).

[3] A. Apostolico, "The Myriad Virtues of Suffix Trees", *Combinatorial Algorithms on Words,* 85–96, Springer-Verlag, ASI F12 (1985).

[4] R. Fagin, J. Nievergelt, N. Pippenger and H. Strong, "Extendible Hashing: A Fast Access Method for Dynamic Files", *ACM TODS,* 4, 315–344 (1979).

[5] P. Flajolet, "On the Performance Evaluation of Extendible Hashing and Trie Searching", *Acta Informatica*, 20, 345–369 (1983).

[6] R. Gallager, *Information Theory and Reliable Communications*, John Wiley & Sons, New York (1968).

[7] J. Capetanakis, "Tree Algorithms for Packet Broadcast Channels", *IEEE Trans. on Information Theory*, IT-25, 505–525 (1979).

[8] *IEEE Transaction on Information Theory*, IT-31, 2 (1985).

[9] Ph. Jacquet and M. Regnier, "Trie Partitioning Process: Limiting Distributions", in *Lecture Notes in Computer Science*, vol. 214, pp. 196-210, Springer Verlag, New York 1986.

[10] L. Devroye, "A Probabilistic Analysis of the Height of Tries and of the Complexity of Trie Sort", *Acta Informatica*, 21, 229–232 (1984).

[11] B. Pittel, "Asymptotic Growth of a Class of Random Trees", *The Annals of Probability*, 13, 414–427 (1985).

[12] B. Pittel, "Path in a Random Digital Tree: Limiting Distributions", *Adv. Appl. Prob.*, 18, 139–155 (1986).

[13] M. Regnier, "On the Average Height of Trees in Digital Searching and Dynamic Hashing", *Inform. Processing Lett.*, 13, 64–66 (1981).

[14] A. Yao, "A Note on the Analysis of Extendible Hashing", *Inform. Processing Lett.*, 11, 84–86 (1980).

[15] W. Szpankowski, "On the Analysis of the Average Height of a Digital Trie: Another Approach", *Purdue University* CSD TR-646, (1986).

[16] A. Apostolico and W. Szpankowski, "Self-Alignments in Words and Their Applications", *Purdue University* CSD TR-732 (1987), submitted to a journal.

[17] W. Szpankowski, "Some Results on *V*-ary Asymmetric Tries", *Journal of Algorithms*, 9, 224–244 (1988).

[18] P. Kirschenhofer, H. Prodinger and W. Szpankowski, "On the Variance of the External Path Length in a Symmetric Digital Trie", *Discrete Applied Mathematics*, 25, 129-143 (1989).

[19] H. David, *Order Statistics*, John Wiley & Sons, New York (1980).

[20] J. Galambos, *The Asymptotic Theory of Extreme Order Statistics*, John Wiley & Sons, New York (1978).

[21] T. Lai and H. Robbins, "A Class of Dependent Random Variables and Their Maxima", *Z. Wahrscheinlichkeitscheorie*, 42, 89–111 (1978).

[22] P. Billingsley, *Probability and Measures*, John Wiley & Sons, New York (1986).

[23] W. Szpankowski, "(Probably) Optimal Solutions to Some Problems NOT Only on Graphs", *Purdue University* CSD TR 780, (1988); revision TR 872, (1989).

[24] D. Knuth, *Art of Computer Programming. Fundamental Algorithms*, vol. I., Addison-Wesley, Reading, Mass. (1973).

[25] D. Sankoff and J. Kruskal, (eds.), *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparisons*, Addison-Wesley, Reading, Mass. (1983).

[26] *Bulletin of Mathematical Biology*, 46, No. 4 (1984), (A special commemorative issue honoring Margonet O. Dayhoff).

[27] D. Aldous, *Probability Approximations via the Poisson Clumping Heristic*, Springer-Verlag, New York (1989).

[28] N. Kamarkar, R.M. Karp, G.S. Lueker and A.D. Odlyzko, "Probabilistic analysis of optimum partitioning", *J. Appl. Prob.*, 23, 626-645 (1986).

[29] S. Karlin, F. Ost, "Counts of long aligned word matches among random letter sequences", *Adv. Appl. Prob.*, 19, 293-351 (1987).

[30] J.F.C. Kingman, Subadditive processes, in *Ecole d'Eté de Probabilités de Saint-Flour V-1975,* Lecture Notes in Mathematics, 539, Springer-Verlag, Berlin 1976.

[31] L. Devroye, W. Szpankowski and B. Rais, A note on the height of suffix trees, *Purdue University,* CSD TR-905 (1989).