Adv. Appl. Prob. **35,** 363–376 (2003)
Printed in Northern Ireland

AN ALMOST SURE RESULT FOR PATH LENGTHS IN BINARY SEARCH TREES

F. M. DEKKING * AND
L. E. MEESTER,* ** *Thomas Stieltjes Institute for Mathematics and Delft University of Technology*

Abstract

This paper studies path lengths in random binary search trees under the random permutation model. It is known that the total path length, when properly normalized, converges almost surely to a nondegenerate random variable Z. The limit distribution is commonly referred to as the 'quicksort distribution'. For the class $\mathcal{A}_m$ of finite binary trees with at most m nodes we partition the external nodes of the binary search tree according to the largest tree that each external node belongs to. Thus, the external path length is divided into parts, each part associated with a tree in $\mathcal{A}_m$. We show that the vector of these path lengths, after normalization, converges almost surely to a constant vector times Z.

Keywords: Binary search tree; random permutation model; path length; almost sure convergence; quicksort distribution

AMS 2000 Subject Classification: Primary 68P05; 60C05
Secondary 60F15; 68Q25

1. Introduction

Consider the growth of a binary search tree under the random permutation model, i.e. the tree is generated from an independent and identically distributed (i.i.d.) sequence of keys from a continuous distribution; let $\mathcal{T}_n$ denote the tree after n keys have been inserted. The candidate nodes for the next key to be inserted are called the *external* nodes of $\mathcal{T}_n$. There are $n + 1$ of them, each of which has probability $1/(n + 1)$ of receiving the next key.

On closer inspection this symmetry disappears: the external nodes can be distinguished by the nature of their sibling, which may be internal or external. These were called, respectively, arm nodes and foot nodes in [2]. From an algorithmic point of view, arm nodes are bad and foot nodes are good: well-balanced trees have very few arm nodes. We recall the close connection between the process of growing binary search trees by inserting keys and sorting these keys using the quicksort algorithm [6]. The number of comparisons needed for the sorting is closely related to the external path length of the corresponding tree. In quicksort, arm nodes mark an inefficient splitting of the list. For this reason, we would like to know the proportion of arm nodes, but also the proportion of the total path length that arm nodes contribute. Since it seems obvious that arm nodes occur closer to the root than foot nodes (see [2] for a proof), we are inclined to think that their total path length contribution will be (proportionally) smaller than that of the foot nodes. However, as our main result shows, these distances to the root do not play a role in the limit: the contribution of the arm nodes to the total path length is solely determined

Received 7 December 2001; revision received 16 January 2003.
* Postal address: Department of ITS, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands.
** Email address: l.e.meester@its.tudelft.nl

by their proportion. It turns out that the same phenomenon occurs if we generalize to the case where the external nodes are characterized by more elaborate local information than just the status of their sibling node.

The technique we use to obtain these results is a bookkeeping of the different types of external nodes as the tree grows. The way we do this is closely related to the urn process modelling described by Aldous *et al.* [1]. Since we are not only interested in *numbers* of external nodes but also in the associated *path lengths*, our analysis is somewhat more involved. In particular, we need more information on the spectrum of the 'growth matrix' $\boldsymbol{G}$ describing the bookkeeping than necessary in [1], and also more than in Smythe's work [13] on the analysis of binary search trees with extended Pólya urn models. For example, the eigenanalysis as applied to the generator matrix in Section 1 of [13] does not suffice for our growth matrix $\boldsymbol{G}$: from the positive regularity of $\boldsymbol{G} + m\boldsymbol{I}$ and the fact that the maximal eigenvalue of $\boldsymbol{G}$ is 1, it follows that (the real parts of) the other eigenvalues are less than 1, whereas our proof requires that they be less than $\frac{1}{2}$.

Other connections to previous work are found in papers on so-called 'fringe heuristics', modifications of the quicksort algorithm aiming at a more balanced tree. An example is the median-of-three rule and, more generally, the median-of-$(2k+1)$. In [10] and [8], matrices are encountered which are reminiscent of the growth matrix $\boldsymbol{G}$, as well as recurrences of the type found in this paper and their solution. Poblete and Munro [10] use an eigenanalysis to solve a generating function recurrence similar to (14) below.

2. External node patterns

Let $\mathcal{C}_k$ denote the class of (fixed) trees with k nodes for $k \geq 1$, and let $\mathcal{C}_0 = \{\varnothing\}$, where $\varnothing$ denotes the empty tree, equivalently seen as one external node. Let m be an arbitrary positive integer. Let $\mathcal{A}_m = \bigcup_{k=0}^{m} \mathcal{C}_k$, the class of all trees of at most m internal nodes. For example, the trees in $\mathcal{A}_2$ are depicted in Figure 1. Here, the dots represent (internal) nodes and the boxes external nodes; a_0 is the empty tree. In most of the sequel, m is fixed, so where possible we omit explicit reference to m, and write $\mathcal{A}$ for $\mathcal{A}_m$.

As usual, we denote the distribution of $\mathcal{T}_k$ by

$$\mathrm{P}(\mathcal{T}_k = a) = \lambda(a) \quad \text{for } a \in \mathcal{C}_k.$$

In this paper, the vector $\boldsymbol{\lambda} = (\lambda(a))_{a\in\mathcal{A}}$ plays an important role. This is not a probability vector, but a simple computation (see also Section 6) shows that $\boldsymbol{\pi} = (\pi(a))_{a\in\mathcal{A}}$ defined by

$$\pi(a) = \frac{2}{(m+1)(m+2)} n_a \lambda(a) \quad \text{for } a \in \mathcal{A}$$

satisfies $\sum_{a\in\mathcal{A}} \pi(a) = 1$. Here, n_a denotes the number of external nodes of a, that is, $n_a = k+1$ if $a \in \mathcal{C}_k$. We call $\boldsymbol{\pi}$ the *aggregate distribution* of the trees in $\mathcal{A}$.

Let R_n be the external path length of $\mathcal{T}_n$, i.e. the sum of the distances of the external nodes to the root of $\mathcal{T}_n$. It has been known for some time (see [11] and [12]) that R_n satisfies

$$\frac{R_n - \mathrm{E}\, R_n}{n+1} \to Z \quad \text{almost surely as } n \to \infty, \tag{1}$$

where Z has the so-called 'quicksort distribution'. Our goal is to establish a multivariate version of this result.

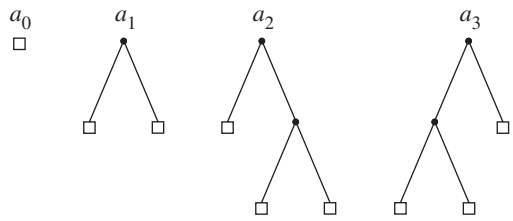

FIGURE 1: The trees of the class $\mathcal{A}_2$.

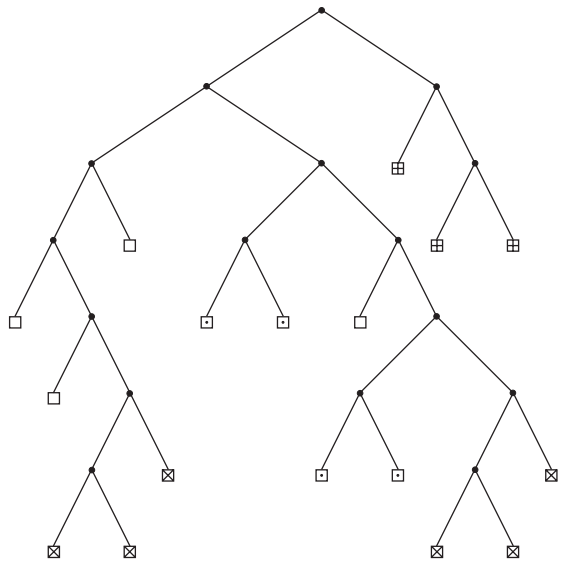

FIGURE 2: Partitioning external nodes using the trees in $\mathcal{A}_2$.

Partition the external nodes of the binary search tree according to the *largest* tree in $\mathcal{A}$ that each node belongs to. For $m = 2$, this is done for an example in Figure 2, marking the external nodes of a_3 with a $+$, those of a_2 with a $\times$, those of a_1 with a $\cdot$, and leaving the remaining external nodes, instances of a_0, unmarked. For example, the bottom right node in Figure 2 belongs to a subtree of type a_0 and to a subtree of type a_1, but the (unique) largest subtree of $\mathcal{A}$ it belongs to is a subtree of type a_3. Thus, the external path length, too, is divided into parts, each part associated with a tree in $\mathcal{A}$. In the example, total path lengths of 16, 20, 8, and 40 are associated with a_0, a_1, a_2, and a_3 respectively. Let X_n^a be the path length to external nodes of subtrees of type a. Note that the overall external path length is $\sum_{a\in\mathcal{A}} X_n^a = R_n$. Let $\boldsymbol{X}_n = (X_n^a)_{a\in\mathcal{A}}$, a *row* vector, and define

$$\boldsymbol{Z}_n = \frac{1}{n+1}(\boldsymbol{X}_n - \mathrm{E}\,\boldsymbol{X}_n).$$

Theorem 1. *Let m be a positive integer and let $\boldsymbol{\pi}$ be the aggregate distribution of the trees in $\mathcal{A}_m$. Then*

$$\boldsymbol{Z}_n \to \boldsymbol{\pi} Z \quad \textit{almost surely as } n \to \infty,$$

where Z is a random variable with the quicksort distribution.

While proving this theorem, we derive some results of independent interest. For example, we prove in Section 7 that the asymptotic formula of Flajolet *et al.* [5, Theorem 1], on means of occurrences of subtrees, after multiplication by $(n+1)/n$, provides the exact value for these means for n larger than the number of nodes in the subtree. We also provide an exact expression for $\mathrm{E}\, X_n^a$ (in (21) below), and it turns out that $\mathrm{E}\, X_n^a \sim \pi(a) 2n \ln n$ as $n \to \infty$. Combining this with the theorem above, we see that the aggregate distribution $\boldsymbol{\pi}$ plays a double role: $\boldsymbol{X}_n$ grows like a multiple of $\boldsymbol{\pi}$, but also (asymptotically) the variation around the mean $\mathrm{E}\, \boldsymbol{X}_n$ is in the direction of $\boldsymbol{\pi}$. The case $m = 1$ of the theorem was proved in [2], a paper that also contains results on the distribution of the depths of the two types of trees in $\mathcal{A}_1$.

3. Proof of Theorem 1

Let $\boldsymbol{w}_1$ denote a column vector of ones, then $\boldsymbol{X}_n \boldsymbol{w}_1 = R_n$, so $\boldsymbol{Z}_n \boldsymbol{w}_1 \to Z$ almost surely as $n \to \infty$ by (1). We shall show that $\boldsymbol{Z}_n \boldsymbol{w} \to 0$ almost surely for any $\boldsymbol{w}$ orthogonal to $\boldsymbol{w}_1$. Let $d = \#\mathcal{A}$ denote the cardinality of $\mathcal{A}$.

Two lemmas form the heart of the proof. In Sections 4–8 we shall prove the first of these.

Lemma 1. *There exist column vectors $\boldsymbol{w}_i$, $i = 2, 3, \ldots, d$, which, along with $\boldsymbol{w}_1$, form a linearly independent set such that, for $i \geq 2$,*

$$\operatorname{var}(\boldsymbol{Z}_n \boldsymbol{w}_i) = \mathcal{O}\left(\frac{\ln^2 n}{n}\right) \quad \text{as } n \to \infty.$$

By the first Borel–Cantelli lemma this implies that, for $i \geq 2$,

$$\boldsymbol{Z}_{n^2} \boldsymbol{w}_i \to 0 \quad \text{almost surely as } n \to \infty.$$

Define, for $i \geq 2$,

$$M_n^{(i)} := \sup_{1 \leq j \leq 2n} |(\boldsymbol{Z}_{n^2+j} - \boldsymbol{Z}_{n^2}) \boldsymbol{w}_i|.$$

In Section 9 we shall prove the second lemma.

Lemma 2. *For $i \geq 2$, $M_n^{(i)} \to 0$ almost surely as $n \to \infty$.*

So, for $i \geq 2$,

$$|\boldsymbol{Z}_n \boldsymbol{w}_i| \leq |\boldsymbol{Z}_{(\lfloor \sqrt{n} \rfloor)^2} \boldsymbol{w}_i| + M_{\lfloor \sqrt{n} \rfloor}^{(i)} \to 0 \qquad \text{almost surely as } n \to \infty.$$

Combined with (1) this proves that

$$\boldsymbol{Z}_n \boldsymbol{W} \to (Z, 0, 0, \ldots, 0, 0) \qquad \text{almost surely as } n \to \infty,$$

where $\boldsymbol{W}$ is a $d \times d$ matrix whose ith column is $\boldsymbol{w}_i$. Theorem 1 now follows from the nonsingularity of $\boldsymbol{W}$ and by checking that the first row of $\boldsymbol{W}^{-1}$ equals $\boldsymbol{\pi}$.

This shows the structure of the proof. It will take several sections to prove Lemma 1.

4. Growing and splitting trees

In our analysis we need to account for the number of subtrees of $\mathcal{T}_n$ for each tree in $\mathcal{A}$, as well as keep track of their average depth. For $a \in \mathcal{A}$, therefore, define $N_{n,k}^a$ to be the number of subtrees of $\mathcal{T}_n$ that are of type a and have their root at depth k; let $N_n^a = \sum_{k=0}^n N_{n,k}^a$ and

define $T_n^a = \sum_{k=0}^n k N_{n,k}^a$, the path length from the root of $\mathcal{T}_n$ to (the roots of) these subtrees. Recall that n_a denotes the number of external nodes of a; hence

$$\sum_{a \in \mathcal{A}} N_n^a n_a = n + 1. \tag{2}$$

For an arbitrary tree t, we denote its external path length by x_t. We shall analyse $\boldsymbol{N}_n = (N_n^a)_{a \in \mathcal{A}}$ and $\boldsymbol{T}_n = (T_n^a)_{a \in \mathcal{A}}$ and draw conclusions about $\boldsymbol{X}_n$ via

$$X_n^a = T_n^a n_a + N_n^a x_a.$$

As growth from $\mathcal{T}_n$ to $\mathcal{T}_{n+1}$ can be viewed as occurring through an insertion of a new node in one of the external nodes of $\mathcal{T}_n$, *one* subtree will grow by one node and possibly, if the size exceeds m, it will split at its root into two smaller trees.

For an arbitrary tree t, let t_0 denote its left subtree, t_1 the right subtree, and identify t and (t_0, t_1). Note that we obtain all trees of a fixed size recursively in this way:

$$\mathcal{C}_k = \bigcup_{j=0}^{k-1} \mathcal{C}_j \times \mathcal{C}_{k-1-j}, \tag{3}$$

starting from $\mathcal{C}_0$, which contains only the empty tree. It is well known (see e.g. [4, Lemma 2.1]) that the vector $\boldsymbol{\lambda}$ satisfies

$$\lambda(a_0)\lambda(a_1) = (n_a - 1)\lambda(a). \tag{4}$$

Let t^+ be the set of trees that can arise from t through one insertion; note that $\#t^+ = n_t$. It appears to be somewhat less well known that

$$\sum_{a : b \in a^+} \lambda(a) = (n_b - 1)\lambda(b). \tag{5}$$

This holds since the trees a in $\mathcal{C}_k$ either cannot produce b in $\mathcal{C}_{k+1}$ or there is a unique external node which produces b, and this has probability n_a^{-1} of occurring.

Let $\boldsymbol{e}_a = (e_a(b))_{b \in \mathcal{A}}$ denote the unit row vector corresponding to coordinate a: $e_a(b) = 1_{[b=a]}$ for $b \in \mathcal{A}$. For a uniformly random insertion in $a \in \mathcal{A}$, we define an 'offspring' vector $\boldsymbol{Y}_a$: for $a \in \mathcal{A}_{m-1}$,

$$\boldsymbol{Y}_a = \boldsymbol{e}_t \quad \text{with probability } \frac{1}{n_a} \text{ for } t \in a^+;$$

and, as the largest trees split, for $a \in \mathcal{C}_m$,

$$\boldsymbol{Y}_a = \boldsymbol{e}_{t_0} + \boldsymbol{e}_{t_1} \quad \text{with probability } \frac{1}{n_a} \text{ for } t = (t_0, t_1) \in a^+.$$

Define $\boldsymbol{K}_n$ by $\boldsymbol{N}_{n+1} = \boldsymbol{N}_n + \boldsymbol{K}_n$ and set $\mathcal{F}_n = \sigma(\mathcal{T}_0, \ldots, \mathcal{T}_n)$. Then

$$[\boldsymbol{K}_n \mid \mathcal{F}_n] \stackrel{\mathrm{D}}{=} \boldsymbol{Y}_a - \boldsymbol{e}_a \quad \text{with probability } \frac{n_a N_n^a}{n+1},$$

where $\stackrel{\mathrm{D}}{=}$ denotes equality in distribution and

$$\mathrm{E}(\boldsymbol{K}_n \mid \mathcal{F}_n) = \sum_{a \in \mathcal{A}} \frac{1}{n+1} (\mathrm{E}\, \boldsymbol{Y}_a - \boldsymbol{e}_a) n_a N_n^a = \frac{1}{n+1} \boldsymbol{N}_n \boldsymbol{G}, \tag{6}$$

where $\boldsymbol{G} = (G_{ab})_{a,b \in \mathcal{A}}$ has rows $\boldsymbol{g}_a$ defined by

$$\boldsymbol{g}_a = n_a (\mathrm{E}\, \boldsymbol{Y}_a - \boldsymbol{e}_a). \tag{7}$$

This growth model can also be viewed as an urn process belonging to the class described in [1]: each subtree type corresponds to a ball type, and in the selection process each type has a weight equal to its number of external nodes. The random vectors $\boldsymbol{Y}_a,\ a \in \mathcal{A}$, correspond to $(Z_j^{(i)},\ 1 \le j \le m),\ 1 \le i \le m$ in [1], and the matrix $\boldsymbol{G}$ plays the role of R. Theorem 2 in [1] implies that $\boldsymbol{N}_n/n \to \boldsymbol{x}/b$ almost surely as $n \to \infty$, where $\boldsymbol{x}$ is a left eigenvector of $\boldsymbol{G}$ corresponding to eigenvalue 1 and b is a normalizing constant.

We shall refer to $\boldsymbol{G}$ as the 'growth matrix' and explicitly determine this main eigenvector in Section 6. However, we also need a more complete eigenanalysis of $\boldsymbol{G}$, as this matrix plays a pivotal role in the recursive equations that are to be solved.

5. Eigenanalysis of the growth matrix

In this section, we use the convenient, but somewhat abusive notation

$$\boldsymbol{e}_{a^+} = \sum_{t \in a^+} \boldsymbol{e}_t.$$

Evaluating (7) leads to the following expression for the rows of $\boldsymbol{G}$:

$$\boldsymbol{g}_a = \begin{cases} \boldsymbol{e}_{a^+} - n_a \boldsymbol{e}_a & \text{if } a \in \mathcal{A}_{m-1}, \\ n_{a_1}\boldsymbol{e}_{a_0} + n_{a_0}\boldsymbol{e}_{a_1} + \boldsymbol{e}_{a_0^+} + \boldsymbol{e}_{a_1^+} - n_a \boldsymbol{e}_a & \text{if } a = (a_0, a_1) \in \mathcal{C}_m. \end{cases} \tag{8}$$

We will determine the spectrum of $\boldsymbol{G}$ by finding appropriate left and right eigenvectors of $\boldsymbol{G}$. We start with the eigenvalue $-m-1$.

Lemma 3. *Let $\boldsymbol{G}$ be the growth matrix of $\mathcal{A}$ and let $t = (t_0, t_1) \in \mathcal{C}_m$. Then*

$$\boldsymbol{u} = \boldsymbol{e}_{t_0} + \boldsymbol{e}_{t_1} - \boldsymbol{e}_t$$

is a left eigenvector for the eigenvalue $\mu = -m-1$.

Proof. We have

$$\begin{aligned} \boldsymbol{u}\boldsymbol{G} &= \boldsymbol{g}_{t_0} + \boldsymbol{g}_{t_1} - \boldsymbol{g}_t \\ &= \boldsymbol{e}_{t_0^+} - n_{t_0}\boldsymbol{e}_{t_0} + \boldsymbol{e}_{t_1^+} - n_{t_1}\boldsymbol{e}_{t_1} - [n_{t_1}\boldsymbol{e}_{t_0} + n_{t_0}\boldsymbol{e}_{t_1} + \boldsymbol{e}_{t_0^+} + \boldsymbol{e}_{t_1^+} - n_t \boldsymbol{e}_t] \\ &= -(n_{t_0} + n_{t_1})\boldsymbol{e}_{t_0} - (n_{t_0} + n_{t_1})\boldsymbol{e}_{t_1} + n_t \boldsymbol{e}_t \\ &= -n_t(\boldsymbol{e}_{t_0} + \boldsymbol{e}_{t_1} - \boldsymbol{e}_t) \\ &= (-m-1)\boldsymbol{u}. \end{aligned}$$

Theorem 2. *Let $\boldsymbol{G}$ be the growth matrix of $\mathcal{A}_m$. Then $\boldsymbol{G}$ has $m+1$ real eigenvalues,* $\operatorname{spec}(\boldsymbol{G}) = \{1, -2, -3, \ldots, -m-1\}$, *with multiplicities equal to* $\#\mathcal{C}_0, \#\mathcal{C}_1, \ldots, \#\mathcal{C}_m$.

Proof. Let $\mu_1 = 1$ and $\mu_k = -k$ for $k = 2, \ldots, m+1$. We shall show that the dimension of the right eigenspace corresponding to μ_k is at least $\#\mathcal{C}_{k-1}$. Together with $\#\mathcal{C}_0 + \#\mathcal{C}_1 + \cdots + \#\mathcal{C}_m = d$ this proves the theorem. As the proof is by induction, we write $\boldsymbol{G}^{(m)}$ for the growth matrix of $\mathcal{A}_m$.

For $m = 1$,

$$\boldsymbol{G}^{(1)} = \begin{pmatrix} -1 & 1 \\ 2 & 0 \end{pmatrix}$$

has right eigenvectors $\boldsymbol{v}_1 = (1\ \ 2)$ and $\boldsymbol{v}_2 = (-1\ \ 1)$, with eigenvalues $\mu_1 = 1$ and $\mu_2 = -2$.

Suppose that the assertion of the theorem holds for $\boldsymbol{G}^{(m)}$. We first note that, by Lemma 3, there exists a collection of $\#\mathcal{C}_{m+1}$ left eigenvectors of $\boldsymbol{G}^{(m+1)}$, clearly independent, corresponding to the eigenvalue $\mu_{m+2} = -m-2$. Since left and right eigenspaces of a matrix have the same dimension, the right eigenspace corresponding to $\mu_{m+2} = -m-2$ has dimension at least $\#\mathcal{C}_{m+1}$. For the other eigenvalues, we use the right eigenvectors of $\boldsymbol{G}^{(m)}$ to construct those of $\boldsymbol{G}^{(m+1)}$. We claim that, if $\boldsymbol{v}_m$ is a right eigenvector of $\boldsymbol{G}^{(m)}$, then $\boldsymbol{v}_{m+1} = (v_{m+1}(a))_{a\in\mathcal{A}}$ defined by

$$v_{m+1}(a) = \begin{cases} v_m(a) & \text{if } a \in \mathcal{A}_m, \\ v_m(a_0) + v_m(a_1) & \text{if } a = (a_0, a_1) \in \mathcal{C}_{m+1}, \end{cases} \tag{9}$$

is an eigenvector of $\boldsymbol{G}^{(m+1)}$ with the same eigenvalue. Since this extension operation obviously preserves independence of the eigenvectors, this claim establishes the theorem.

Note that $\boldsymbol{G}^{(m)}\boldsymbol{v}_m = \mu\boldsymbol{v}_m$ implies that

$$\sum_{b\in a^+} v_m(b) = (\mu + n_a)v_m(a) \quad \text{if } a \in \mathcal{A}_{m-1}. \tag{10}$$

We split the proof that $\boldsymbol{G}^{(m+1)}\boldsymbol{v}_{m+1} = \mu\boldsymbol{v}_{m+1}$ into three cases.

Case 1: $a \in \mathcal{A}_{m-1}$. In this case, it follows directly from (10) that

$$\begin{aligned} \boldsymbol{g}_a^{(m+1)}\boldsymbol{v}_{m+1} &= \boldsymbol{e}_{a^+}\boldsymbol{v}_{m+1} - n_a\boldsymbol{e}_a\boldsymbol{v}_{m+1} \\ &= \sum_{b\in a^+} v_{m+1}(b) - n_a v_{m+1}(a) \\ &= \sum_{b\in a^+} v_m(b) - n_a v_m(a) \\ &= \mu v_m(a) = \mu v_{m+1}(a). \end{aligned}$$

Case 2: $a \in \mathcal{C}_m$. Then $a_0, a_1 \in \mathcal{A}_{m-1}$ and, by twice applying (9) with $m-1$ instead of m,

$$v_m(a) = v_{m-1}(a_0) + v_{m-1}(a_1) = v_m(a_0) + v_m(a_1). \tag{11}$$

As in Case 1, we have

$$\boldsymbol{g}_a^{(m+1)}\boldsymbol{v}_{m+1} = \sum_{b\in a^+} v_{m+1}(b) - n_a v_{m+1}(a). \tag{12}$$

Now

$$a^+ = \{(a_0, b) : b \in a_1^+\} \cup \{(b, a_1) : b \in a_0^+\} \subset \mathcal{C}_{m+1},$$

so

$$\begin{aligned} \sum_{b\in a^+} v_{m+1}(b) &= \sum_{b\in a^+} [v_m(b_0) + v_m(b_1)] \\ &= \sum_{b_1\in a_1^+} [v_m(a_0) + v_m(b_1)] + \sum_{b_0\in a_0^+} [v_m(b_0) + v_m(a_1)] \\ &= n_{a_1} v_m(a_0) + (\mu + n_{a_1})v_m(a_1) + n_{a_0} v_m(a_1) + (\mu + n_{a_0})v_m(a_0) \\ &= (\mu + n_a)v_m(a_0) + (\mu + n_a)v_m(a_1) \\ &= (\mu + n_a)v_m(a). \end{aligned}$$

Here we used (10) with a_0 and a_1 instead of a, and (11) in the last step. Substituting this result into (12), we obtain

$$\boldsymbol{g}_a^{(m+1)}\boldsymbol{v}_{m+1} = (\mu + n_a)v_m(a) - n_a v_{m+1}(a) = \mu v_{m+1}(a).$$

Case 3: $a \in \mathcal{C}_{m+1}$. This case is different from the first two because $\boldsymbol{g}_a$ is different, but also because it is now possible that $a_0 \in \mathcal{C}_m$. But note that, in fact, we *always* have

$$\boldsymbol{e}_{a_0^+}\boldsymbol{v}_{m+1} = (\mu + n_{a_0})v_m(a_0),$$

either by (10) if $a_0 \in \mathcal{A}_{m-1}$ or by the computation in Case 2 if $a_0 \in \mathcal{C}_m$. The same applies to a_1, and therefore

$$\begin{aligned}
\boldsymbol{g}_a^{(m+1)}\boldsymbol{v}_{m+1} &= [n_{a_1}\boldsymbol{e}_{a_0} + n_{a_0}\boldsymbol{e}_{a_1} + \boldsymbol{e}_{a_0^+} + \boldsymbol{e}_{a_1^+} - n_a\boldsymbol{e}_a]\boldsymbol{v}_{m+1} \\
&= n_{a_1}v_m(a_0) + n_{a_0}v_m(a_1) + (\mu + n_{a_0})v_m(a_0) + (\mu + n_{a_1})v_m(a_1) - n_a v_{m+1}(a) \\
&= (\mu + n_a)v_m(a_0) + (\mu + n_a)v_m(a_1) - n_a v_{m+1}(a) \\
&= (\mu + n_a)v_{m+1}(a) - n_a v_{m+1}(a) \\
&= \mu v_{m+1}(a).
\end{aligned}$$

6. The eigenvalue $\mu = 1$

In the analysis of Section 7, the eigenvalue $\mu = 1$ is the dominant one, so, in order to obtain explicit solutions, we now determine the corresponding eigenvector.

Theorem 3. *Let* $\boldsymbol{\lambda} = (\lambda(a))_{a\in\mathcal{A}_m}$ *be the aggregate distribution vector of the trees in* $\mathcal{A}_m$. *Then* $\boldsymbol{\lambda}$ *is a left eigenvector of the growth matrix* $\boldsymbol{G}$ *with eigenvalue* 1.

Proof. We decompose the eigenvalue equation as follows:

$$\begin{aligned}
\sum_{a\in\mathcal{A}_m} \lambda(a)G_{ab} &= \sum_{a\in\mathcal{A}_{m-1}} \lambda(a)G_{ab} + \sum_{a\in\mathcal{C}_m} \lambda(a)G_{ab} \\
&=: \Sigma^{(m-1)} + \Sigma^{(m)}.
\end{aligned}$$

Here, $\Sigma^{(m-1)}$ is the easier to evaluate. First we consider the case $b \in \mathcal{A}_{m-1}$. To make the following work for $b = \varnothing$, we define $\mathcal{C}_{-1} = \varnothing$. From (8) and (5) we find that

$$\begin{aligned}
\Sigma^{(m-1)} = \sum_{a\in\mathcal{A}_{m-1}} \lambda(a)G_{ab} &= \sum_{k=0}^{m-1}\sum_{a\in\mathcal{C}_k} \lambda(a)[\boldsymbol{e}_{a^+}(b) - n_a\boldsymbol{e}_a(b)] \\
&= \sum_{a:b\in a^+} \lambda(a) - n_b\lambda(b) \\
&= (n_b - 1)\lambda(b) - n_b\lambda(b) \\
&= -\lambda(b).
\end{aligned}$$

In the case $b \in \mathcal{C}_m$, necessarily $e_a(b) = 0$ for all $a \in \mathcal{A}_{m-1}$, so then

$$\Sigma^{(m-1)} = \sum_{a\in\mathcal{A}_{m-1}} \lambda(a)G_{ab} = (n_b - 1)\lambda(b) = m\lambda(b).$$

We now turn to the second sum $\Sigma^{(m)}$, starting again with the case $b \in \mathcal{C}_j$ for some j with $0 \le j \le m-1$. Then $e_a(b) = 0$ for all $a \in \mathcal{C}_m$, so using (3) and (5) we find that

$$\begin{aligned}
\Sigma^{(m)} &= \sum_{a\in\mathcal{C}_m} \lambda(a) G_{ab} \\
&= \sum_{k=0}^{m-1} \sum_{a_0\in\mathcal{C}_k} \sum_{a_1\in\mathcal{C}_{m-1-k}} \frac{\lambda(a_0)\lambda(a_1)}{m} \bigl[n_{a_1} e_{a_0}(b) + n_{a_0} e_{a_1}(b) + e_{a_0^+}(b) + e_{a_1^+}(b)\bigr] \\
&= \frac{1}{m}\Biggl[\sum_{a_1\in\mathcal{C}_{m-1-j}} \lambda(b)\lambda(a_1) n_{a_1} + \sum_{a_0\in\mathcal{C}_{m-1-j}} \lambda(a_0)\lambda(b) n_{a_0} \\
&\qquad + \sum_{a_0\in\mathcal{C}_{j-1}} \sum_{a_1\in\mathcal{C}_{m-j}} \lambda(a_0)\lambda(a_1) e_{a_0^+}(b) + \sum_{a_0\in\mathcal{C}_{m-j}} \sum_{a_1\in\mathcal{C}_{j-1}} \lambda(a_0)\lambda(a_1) e_{a_1^+}(b) \Biggr] \\
&= \frac{1}{m}[2\lambda(b)(m-j) + 2j\lambda(b)] \\
&= 2\lambda(b),
\end{aligned}$$

where, in the last step but one, we used the fact that $\sum_{a\in\mathcal{C}_k} \lambda(a) = 1$ four times and the fact that $\sum_{a:b\in a^+} \lambda(a) = j\lambda(b)$ twice.

Finally, we consider $\Sigma^{(m)}$ in the case $b \in \mathcal{C}_m$. Now $e_a(b) = 1$ for some a; thus, relative to the previous case, this adds a term

$$-n_b\lambda(b) = -(m+1)\lambda(b)$$

to $\sum_{a\in\mathcal{C}_m} \lambda(a) G_{ab}$, so here

$$\Sigma^{(m)} = 2\lambda(b) - (m+1)\lambda(b) = (-m+1)\lambda(b).$$

In conclusion,

$$\sum_{a\in\mathcal{A}_m} \lambda(a) G_{ab} = \Sigma^{(m-1)} + \Sigma^{(m)} = \begin{cases} -\lambda(b) + 2\lambda(b) = \lambda(b) & \text{if } b \in \mathcal{A}_{m-1}, \\ m\lambda(b) + (-m+1)\lambda(b) = \lambda(b) & \text{if } b \in \mathcal{C}_m, \end{cases}$$

and we have obtained that $\boldsymbol{\lambda G} = \boldsymbol{\lambda}$.

A *right* eigenvector is more easily found: $\boldsymbol{G v} = \boldsymbol{v}$ holds for the column vector $\boldsymbol{v}$ defined by $v(b) = n_b$. This can be verified directly from (8), but it is quicker to apply induction as in the proof of Theorem 2: using $n_{a_0} + n_{a_1} = n_a$ in (9) immediately shows that $\boldsymbol{v}$ is an eigenvector with eigenvalue 1.

From now on, we denote this eigenvector by $\boldsymbol{v}_1$. We look for the left eigenvector $\boldsymbol{u}_1$ such that $\boldsymbol{u}_1\boldsymbol{v}_1 = 1$. Note that

$$\begin{aligned}
\boldsymbol{\lambda v}_1 &= \sum_{a\in\mathcal{A}_m} \lambda(a) n_a = \sum_{k=0}^{m} (k+1) \sum_{a\in\mathcal{C}_k} \lambda(a) = \sum_{k=0}^{m} (k+1) \\
&= \frac{(m+1)(m+2)}{2}.
\end{aligned}$$

We therefore define $\boldsymbol{u}_1$ by

$$\boldsymbol{u}_1 = \frac{2}{(m+1)(m+2)}\boldsymbol{\lambda}.$$

By Theorem 2, we can add eigenvectors $\boldsymbol{u}_2, \ldots, \boldsymbol{u}_d$ to $\boldsymbol{u}_1$ and obtain a basis. Let $\boldsymbol{U}$ be the matrix with these vectors as rows and let $\boldsymbol{V} = \boldsymbol{U}^{-1}$. The columns of $\boldsymbol{V}$ form a basis of right eigenvectors. Note that the first column equals $\boldsymbol{v}_1$ (since $\boldsymbol{u}_1\boldsymbol{v}_1 = 1$ and the eigenspace of $\mu = 1$ is one dimensional).

7. Means and variances of subtree occurrences

These means and variances are needed because they appear in the recursive equations for means and variances of the path lengths.

From (6), we have

$$\mathrm{E}(\boldsymbol{N}_{n+1} \mid \mathcal{F}_n) = \boldsymbol{N}_n\left(\boldsymbol{I} + \frac{1}{n+1}\boldsymbol{G}\right), \tag{13}$$

where $\boldsymbol{I}$ is the identity matrix, and so

$$\mathrm{E}\,\boldsymbol{N}_{n+1} = \mathrm{E}\,\boldsymbol{N}_n\left(\boldsymbol{I} + \frac{1}{n+1}\boldsymbol{G}\right). \tag{14}$$

To solve for $\mathrm{E}\,\boldsymbol{N}_n$, change basis to the one given by the rows of $\boldsymbol{U}$. The ith coordinate of $\mathrm{E}\,\boldsymbol{N}_n$ with respect to this basis is $\mathrm{E}\,\boldsymbol{N}_n\boldsymbol{v}_i$. Define $\varphi_n = \mathrm{E}\,\boldsymbol{N}_n\boldsymbol{v}_i$, then

$$\varphi_{n+1} = \left(1 + \frac{\mu}{n+1}\right)\varphi_n,$$

where μ is the eigenvalue corresponding to $\boldsymbol{v}_i$. The solution to this recursion is, for $\mu = 1$,

$$\varphi_n = (n+1)\varphi_0 \quad \text{for } n \geq 0;$$

and for $\mu = -k$, where k is a positive integer,

$$\varphi_n = \begin{cases} \dfrac{(n-k)\cdots(1-k)}{n!}\varphi_0 & \text{for } 0 \leq n < k, \\ 0 & \text{for } n \geq k. \end{cases}$$

As general solution, taking initial conditions into account, we find that

$$\mathrm{E}\,\boldsymbol{N}_n = (n+1)\boldsymbol{u}_1 \quad \text{for } n \geq m+1. \tag{15}$$

This gives an exact expression for the first moments of occurrences of subtrees (cf. Theorem 1 in [5], where an asymptotic expression is given): for $a \in \mathcal{C}_m$ and $n \geq m+1$,

$$\mathrm{E}\,N_n^a = \frac{2}{(m+1)(m+2)}\lambda(a)(n+1). \tag{16}$$

Since, by conditioning on $\mathcal{F}_n$ and using (6),

$$\begin{aligned} \mathrm{cov}(\boldsymbol{N}_n, \boldsymbol{K}_n) &= \mathrm{E}\,\mathrm{E}((\boldsymbol{N}_n - \mathrm{E}\,\boldsymbol{N}_n)(\boldsymbol{K}_n - \mathrm{E}\,\boldsymbol{K}_n) \mid \mathcal{F}_n) \\ &= \mathrm{E}((\boldsymbol{N}_n - \mathrm{E}\,\boldsymbol{N}_n)\,\mathrm{E}((\boldsymbol{K}_n - \mathrm{E}\,\boldsymbol{K}_n) \mid \mathcal{F}_n)) \\ &= \frac{1}{n+1}\,\mathrm{var}(\boldsymbol{N}_n)\boldsymbol{G} \end{aligned}$$

and

$$\operatorname{var}(\boldsymbol{N}_{n+1}) = \operatorname{var}(\boldsymbol{N}_n) + \operatorname{cov}(\boldsymbol{N}_n, \boldsymbol{K}_n) + \operatorname{cov}(\boldsymbol{K}_n, \boldsymbol{N}_n) + \operatorname{var}(\boldsymbol{K}_n),$$

we find, for $\sigma_n = \operatorname{var}(\boldsymbol{N}_n \boldsymbol{v})$,

$$\sigma_{n+1} = \left(1 + \frac{2\mu}{n+1}\right)\sigma_n + \operatorname{var}(\boldsymbol{K}_n \boldsymbol{v}).$$

For $\mu = 1$ we have $\boldsymbol{N}_n \boldsymbol{v}_1 = n + 1$ by (2), and $\operatorname{var}(\boldsymbol{N}_n \boldsymbol{v}_1) = 0$ for all n. Since the components of $\boldsymbol{K}_n$ are bounded by 2, the variance $\operatorname{var}(\boldsymbol{K}_n \boldsymbol{v})$ is bounded as well (in fact, it is constant for $n \geq m + 1$). Hence, as the other eigenvalues satisfy $\mu < \frac{1}{2}$, for any other eigenvector $\boldsymbol{v}$,

$$\operatorname{var}(\boldsymbol{N}_n \boldsymbol{v}) = \mathcal{O}(n). \tag{17}$$

See Theorem 1 of [5] for an asymptotic expression for $\operatorname{var}(N_n^a)$.

8. Means and variances of subtree path lengths

We follow a scheme similar to that in the previous section. Set $\boldsymbol{T}_{n+1} = \boldsymbol{T}_n + \boldsymbol{W}_n$. Then, for $a \in \mathcal{A}$,

$$[\boldsymbol{W}_n \mid \mathcal{F}_n] \stackrel{\mathrm{D}}{=} k(\boldsymbol{Y}_a - \boldsymbol{e}_a) + 1_{[a \in \mathcal{C}_m]} \boldsymbol{Y}_a \quad \text{with probability } \frac{n_a N_{n,k}^a}{n+1},$$

where the second term is for the case $a \in \mathcal{C}_m$: the new subtrees have their roots one level deeper in the tree than their 'parent'. So

$$\begin{aligned}
\mathrm{E}(\boldsymbol{W}_n \mid \mathcal{F}_n) &= \frac{1}{n+1} \sum_{a \in \mathcal{A}} n_a \sum_{k=1}^{n} \{k(\mathrm{E}\, \boldsymbol{Y}_a - \boldsymbol{e}_a) + 1_{[a \in \mathcal{C}_m]} \mathrm{E}\, \boldsymbol{Y}_a\} N_{n,k}^a \\
&= \frac{1}{n+1} \sum_{a \in \mathcal{A}} n_a \{T_n^a (\mathrm{E}\, \boldsymbol{Y}_a - \boldsymbol{e}_a) + N_n^a 1_{[a \in \mathcal{C}_m]} \mathrm{E}\, \boldsymbol{Y}_a\} \\
&= \frac{1}{n+1} (\boldsymbol{T}_n \boldsymbol{G} + \boldsymbol{N}_n \boldsymbol{B}),
\end{aligned} \tag{18}$$

where $\boldsymbol{B} = (B_{ab})_{a,b \in \mathcal{A}}$, with rows

$$n_a 1_{[a \in \mathcal{C}_m]} \mathrm{E}\, \boldsymbol{Y}_a = \begin{cases} 0 & \text{if } a \in \mathcal{A}_{m-1}, \\ n_{a_1} \boldsymbol{e}_{a_0} + n_{a_0} \boldsymbol{e}_{a_1} + \boldsymbol{e}_{a_0^+} + \boldsymbol{e}_{a_1^+} & \text{if } a \in \mathcal{C}_m. \end{cases}$$

Hence,

$$\mathrm{E}(\boldsymbol{T}_{n+1} \mid \mathcal{F}_n) = \boldsymbol{T}_n \left(\boldsymbol{I} + \frac{1}{n+1} \boldsymbol{G}\right) + \boldsymbol{N}_n \frac{1}{n+1} \boldsymbol{B}$$

and

$$\mathrm{E}\, \boldsymbol{T}_{n+1} = \mathrm{E}\, \boldsymbol{T}_n \left(\boldsymbol{I} + \frac{1}{n+1} \boldsymbol{G}\right) + \mathrm{E}\, \boldsymbol{N}_n \frac{1}{n+1} \boldsymbol{B}. \tag{19}$$

Consider the second term on the right-hand side. For $n \leq m - 1$, it is zero: $N_n^a = 0$ if $a \in \mathcal{C}_m$ and $B_{ab} = 0$ if $a \in \mathcal{A}_{m-1}$. For $n = m$,

$$\mathrm{E}\, N_n^a = \begin{cases} 0 & \text{if } a \in \mathcal{A}_{m-1}, \\ \lambda(a) & \text{if } a \in \mathcal{C}_m, \end{cases}$$

and therefore, for $b \in \mathcal{A}_m$,

$$\begin{aligned}(\mathrm{E}\, \boldsymbol{N}_n \boldsymbol{B})_b &= \sum_{a \in \mathcal{C}_m} \lambda(a) B_{ab} \\ &= \sum_{a \in \mathcal{C}_m} \lambda(a)(G_{ab} + n_a \boldsymbol{e}_a(b)) \\ &= \Sigma^{(m)} + (m+1)\lambda(b) 1_{[b \in \mathcal{C}_m]} \\ &= 2\lambda(b).\end{aligned}$$

For $n \geq m+1$, we find that $\mathrm{E}\, \boldsymbol{N}_n \boldsymbol{B} = (n+1)\boldsymbol{u}_1 \boldsymbol{B} = 2(n+1)\boldsymbol{u}_1$ by a similar derivation. Putting these results together,

$$\mathrm{E}\, \boldsymbol{N}_n \frac{1}{n+1} \boldsymbol{B} = \begin{cases} \boldsymbol{0} & \text{if } n \leq m-1, \\ (m+2)\boldsymbol{u}_1 & \text{if } n = m, \\ 2\boldsymbol{u}_1 & \text{if } n \geq m+1. \end{cases} \tag{20}$$

Therefore, for a right eigenvector $\boldsymbol{v}$ of $\boldsymbol{G}$ corresponding to eigenvalue $\mu \neq 1$, we find from (19) that

$$\mathrm{E}\, \boldsymbol{T}_{n+1} \boldsymbol{v} = \mathrm{E}\, \boldsymbol{T}_n \boldsymbol{v} \left(1 + \frac{\mu}{n+1}\right),$$

since $\boldsymbol{u}_1 \boldsymbol{v} = 0$. As before, for these $\boldsymbol{v}$ we find that $\mathrm{E}\, \boldsymbol{T}_n \boldsymbol{v} = 0$ for $n \geq m+1$. For $\mu = 1$, an explicit solution is found, using (20) and noting that $\boldsymbol{u}_1 \boldsymbol{v}_1 = 1$. Hence,

$$\mathrm{E}\, \boldsymbol{T}_n = (2H_{n+1} - 2H_{m+2} + 1)(n+1)\boldsymbol{u}_1 \quad \text{for } n \geq m+1,$$

where $H_k = \sum_{i=1}^{k} 1/i$. Combining this with (15), the total expected path length to external nodes of subtrees of type a is

$$\mathrm{E}\, X_n^a = \left(2H_{n+1} - 2H_{m+2} + 1 + \frac{x_a}{n_a}\right)(n+1)\pi(a) \quad \text{for } n \geq m+1. \tag{21}$$

Turning to variances again, via (18) we find that

$$\mathrm{cov}(\boldsymbol{T}_n, \boldsymbol{W}_n) = \frac{1}{n+1}[\mathrm{var}(\boldsymbol{T}_n)\boldsymbol{G} + \mathrm{cov}(\boldsymbol{T}_n, \boldsymbol{N}_n)\boldsymbol{B}]$$

and, for $\tau_n = \mathrm{var}(\boldsymbol{T}_n \boldsymbol{v})$,

$$\tau_{n+1} = \left(1 + \frac{2\mu}{n+1}\right)\tau_n + \frac{2}{n+1}\,\mathrm{cov}(\boldsymbol{T}_n \boldsymbol{v}, \boldsymbol{N}_n \boldsymbol{B} \boldsymbol{v}) + \mathrm{var}(\boldsymbol{W}_n \boldsymbol{v}).$$

By the Cauchy–Schwarz inequality, and since the entries in $\mathrm{var}(\boldsymbol{N}_n)$ are of order n, the covariance term is bounded by $C_1\sqrt{n\tau_n}$ for some positive constant C_1.

As the change in the total number of subtrees of a certain type is less than 2, and the depth of roots of subtrees is less than the insertion depth U_n of the new node, we have $|W_n^a| \leq 2U_n$ for all n and $a \in \mathcal{A}$. Since $\mathrm{E}\, U_n \sim 2\ln n$ and $\mathrm{var}(U_n) \sim 2\ln n$ (see [7]), $\mathrm{var}(\boldsymbol{W}_n \boldsymbol{v}) \leq C_2(\ln n)^2$ for some positive C_2. Hence, for every n,

$$\tau_{n+1} \leq \tau_n \left(1 + \frac{2\mu}{n+1}\right) + \frac{2}{n+1} C_1 \sqrt{n\tau_n} + C_2 \ln^2 n.$$

For $\mu < \frac{1}{2}$, we shall prove that this inequality implies that $\tau_n = \mathcal{O}(n \ln^2 n)$ by setting $\tau_n = \rho_n (n+1) \ln^2 n$ and showing that ρ_n is bounded. Set $\alpha = 1 - 2\mu$ and use

$$\sqrt{n \tau_n} \le (n+1)(1+\rho_n) \ln n$$

to obtain that

$$\begin{aligned}\rho_{n+1} &\le \rho_n \frac{(n+1)\ln^2 n}{(n+2)\ln^2(n+1)}\left(1 - \frac{\alpha}{n+2}\right) \\ &\quad + \frac{2C_1 \ln n}{(n+2)\ln^2(n+1)}(1+\rho_n) + \frac{C_2 \ln^2 n}{(n+2)\ln^2(n+1)} \\ &\le \rho_n \left(1 - \frac{\alpha}{n+2} + \frac{2C_1}{(n+2)\ln(n+1)}\right) + \frac{2C_1 + C_2 \ln n}{(n+2)\ln(n+1)}.\end{aligned}$$

We can find an N_0 such that, for $n \ge N_0$, the last inequality may be replaced by

$$\rho_{n+1} \le \rho_n \left(1 - \frac{\alpha/2}{n+2}\right) + \frac{2C_2}{n+2}.$$

Now, set $K = \max(\rho_{N_0}, 4C_2/\alpha)$. The inequality implies that $\rho_n \le K$ for $n \ge N_0$, and we may conclude that

$$\operatorname{var}(\boldsymbol{T}_n \boldsymbol{v}) = \mathcal{O}(n \ln^2 n). \tag{22}$$

Collecting results, denoting by $\boldsymbol{D}$ and $\boldsymbol{F}$ diagonal matrices with diagonal entries $D_{aa} = n_a$ and $F_{aa} = x_a$, and letting $\boldsymbol{w} = \boldsymbol{D}^{-1}\boldsymbol{v}$, where $\boldsymbol{v}$ is a right eigenvector of $\boldsymbol{G}$ not corresponding to $\mu = 1$, we find that

$$\begin{aligned}\operatorname{var}(\boldsymbol{X}_n \boldsymbol{w}) &= \operatorname{var}((\boldsymbol{T}_n \boldsymbol{D} + \boldsymbol{N}_n \boldsymbol{F})\boldsymbol{w}) \\ &= \operatorname{var}(\boldsymbol{T}_n \boldsymbol{D}\boldsymbol{w}) + 2\operatorname{cov}(\boldsymbol{T}_n \boldsymbol{D}\boldsymbol{w}, \boldsymbol{N}_n \boldsymbol{F}\boldsymbol{w}) + \operatorname{var}(\boldsymbol{N}_n \boldsymbol{F}\boldsymbol{w}).\end{aligned}$$

By (22), the first term is $\mathcal{O}(n \ln^2 n)$; by the Cauchy–Schwarz inequality, (22), and (17), the second term is $\mathcal{O}(n \ln n)$; and the third is $\mathcal{O}(n)$, by (17). Therefore, $\operatorname{var}(\boldsymbol{Z}_n \boldsymbol{w}) = \mathcal{O}(\ln^2 n / n)$, and as this is the case no matter the choice of $\boldsymbol{v}$, provided that $\mu \ne 1$, this proves the assertion of Lemma 1. (Note that $\boldsymbol{w}_1 = \boldsymbol{D}^{-1}\boldsymbol{v}_1$ is the vector of ones.)

9. Proof of Lemma 2

From the definition,

$$M_n^{(i)} \le \sup_{a \in \mathcal{A}} \Big(\sup_{1 \le j \le 2n} |Z^a_{n^2+j} - Z^a_{n^2}| \Big) \sup_{a \in \mathcal{A}} |w_i(a)|$$

and

$$\begin{aligned}\sup_{1 \le j \le 2n} |Z^a_{n^2+j} - Z^a_{n^2}| &\le \sum_{k=n^2+1}^{n^2+2n} |Z^a_k - Z^a_{k-1}| \\ &\le \sum_{k=n^2+1}^{n^2+2n} \left(\left| \frac{X^a_k}{k+1} - \frac{X^a_{k-1}}{k} \right| + \left| \frac{\mathrm{E}\, X^a_k}{k+1} - \frac{\mathrm{E}\, X^a_{k-1}}{k} \right| \right).\end{aligned}$$

These terms can be bounded by considering the largest change in the number of subtrees (2), using the upper bound for the number of external nodes and the maximum path length D_k, as follows:

$$\begin{aligned}\left|\frac{X_k^a}{k+1}-\frac{X_{k-1}^a}{k}\right| &\le \frac{|X_k^a - X_{k-1}^a|}{k+1} + \frac{|X_{k-1}^a|}{k(k+1)} \\ &\le \frac{2(m+1)D_k}{k+1} + \frac{D_{k-1}}{k+1} \\ &\le 2(m+2)\frac{D_k}{k+1}.\end{aligned}$$

The bound for the second term in the sum follows immediately from the first, by pulling the expectation through the absolute value:

$$\left|\frac{\mathrm{E}\, X_k^a}{k+1}-\frac{\mathrm{E}\, X_{k-1}^a}{k}\right| \le 2(m+2)\frac{\mathrm{E}\, D_k}{k+1}.$$

Therefore,

$$\sup_{1\le j\le 2n} |Z_{n^2+j}^a - Z_{n^2}^a| \le 4(m+2)\frac{D_{n^2+2n} + \mathrm{E}\, D_{n^2+2n}}{n} \to 0$$

almost surely as $n \to \infty$, since $D_n/\ln n \to \gamma$ almost surely and in L_1 as $n \to \infty$ for some constant γ (see [9] and [3]).

Acknowledgement

We are grateful to the referee for comments that led to vast improvements to the introduction.

References

[1] Aldous, D. J., Flannery, B. and Palacios, J. L. (1988). Two applications of urn processes: the fringe analysis of search trees and the simulation of quasi-stationary distributions of Markov chains. *Prob. Eng. Inf. Sci.* **2,** 293–307.

[2] Dekking, F. M., De Graaf, S. and Meester, L. E. (2000). On the node structure of binary search trees. In *Mathematics and Computer Science* (Versailles, 2000), eds D. Gardy and A. Mokkadem, Birkhäuser, Basel, pp. 31–40.

[3] Devroye, L. (1986). A note on the height of binary search trees. *J. Assoc. Comput. Mach.* **33,** 489–498.

[4] Fill, J. A. (1996). On the distribution of binary search trees under the random permutation model. *Random Structures Algorithms* **8,** 1–25.

[5] Flajolet, P., Gourdon, X. and Martínez, C. (1997). Patterns in random binary search trees. *Random Structures Algorithms* **11,** 223–244.

[6] Knuth, D. E. (1973). *The Art of Computer Programming*, Vol. 3, Sorting and Searching. Addison-Wesley, Reading, MA.

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

[8] Mahmoud, H. M. (1998). On rotations in fringe-balanced binary trees. *Inf. Process. Lett.* **65,** 41–46.

[9] Pittel, B. (1985). Asymptotical growth of a class of random trees. *Ann. Prob.* **13,** 414–427.

[10] Poblete, P. V. and Munro, J. I. (1985). The analysis of a fringe heuristic for binary search trees. *J. Algorithms* **6,** 336–350.

[11] Régnier, M. (1989). A limiting distribution for quicksort. *RAIRO Inf. Théorique Appl.* **23,** 335–343.

[12] Rösler, U. (1991). A limit theorem for 'quicksort'. *RAIRO Inf. Théorique Appl.* **25,** 85–100.

[13] Smythe, R. T. (1996). Central limit theorems for urn models. *Stoch. Process. Appl.* **65,** 115–137.