J. Appl. Prob. **50,** 228–238 (2013)
Printed in England

ASYMPTOTIC ANALYSIS OF HOPPE TREES

KEVIN LECKEY * AND
RALPH NEININGER,* ** *Goethe University Frankfurt*

Abstract

We introduce and analyze a random tree model associated to Hoppe's urn. The tree is built successively by adding nodes to the existing tree when starting with the single root node. In each step a node is added to the tree as a child of an existing node, where these parent nodes are chosen randomly with probabilities proportional to their weights. The root node has weight $\vartheta > 0$, a given fixed parameter, all other nodes have weight 1. This resembles the stochastic dynamic of Hoppe's urn. For $\vartheta = 1$, the resulting tree is the well-studied random recursive tree. We analyze the height, internal path length, and number of leaves of the Hoppe tree with n nodes as well as the depth of the last inserted node asymptotically as $n \to \infty$. Mainly expectations, variances, and asymptotic distributions of these parameters are derived.

Keywords: Hoppe urn; random tree; weak convergence; martingale; combinatorial probability

2010 Mathematics Subject Classification: Primary 60F05; 60C05
Secondary 60G42; 68R05

1. Introduction

We consider a random tree model associated and derived from Hoppe's urn. In Hoppe's urn, see [9], there is initially one red ball. In each step one of the balls is drawn from the urn independently with a probability proportional to the weight of the ball. The red ball has weight $\vartheta > 0$, all other balls have weight 1. Here the parameter $\vartheta > 0$ is given and fixed throughout the evolution of the urn. When a ball is drawn, it is returned to the urn together with a ball of the same color unless the ball drawn is the red ball. In this case the red ball is returned to the urn together with a ball of a color not yet present in the urn. This model has been introduced to derive and interpret the Ewens sampling formula and is related to the infinite alleles model in population genetics, with the parameter $\vartheta > 0$ modeling the mutation rate. The decomposition of the balls into groups of the same color (neglecting the red ball) leads to a Chinese restaurant process, the $(0, \vartheta)$ seating plan; see [13, p. 61].

A random tree model, which we subsequently call the Hoppe tree, is associated to the Hoppe urn as follows. The balls in the urn are represented by nodes in the tree. Each node v is a child of node w in the tree if the ball corresponding to v was placed first in the urn together with the ball corresponding to w when the w-ball was drawn. In other words, the tree grows successively. In each step a node is chosen independently and with probability proportional to the weight of the node (the root having weight ϑ and all other nodes having weight 1) and a new node is added as a child of the chosen node. For $\vartheta = 1$, this is a well-known and well-studied random tree model, the random recursive tree; see, e.g. [15].

Received 23 February 2012; revision received 5 July 2012.
* Postal address: Institute for Mathematics, Goethe University, 60054 Frankfurt am Main, Germany.
** Email address: neiningr@math.uni-frankfurt.de

The aim of the present paper, which is based on the first author's masters thesis [10], is to study asymptotic properties of the Hoppe tree as its size n tends to ∞. In particular, we are interested in the deviation from the random recursive tree model caused by the perturbation of the root weight from $\vartheta = 1$ to $\vartheta \neq 1$. As characteristics of the tree, we study the depth $D_n^{(\vartheta)}$ of the nth inserted node in the tree, defined as its distance to the root of the tree. Furthermore, we study the tree's height $H_n^{(\vartheta)}$, which is the maximal depth $\max_{1\le i\le n} D_i^{(\vartheta)}$, its internal path length $I_n^{(\vartheta)} = \sum_{1\le i\le n} D_i^{(\vartheta)}$, and the number of leaves of the tree. A node is a leaf if it has no child in the tree. Our results show that the perturbation of the root weight does not typically affect the first-order behavior of the quantities, an exception being the variance and limit law of the internal path length. Hence, we give second-order expansions to reveal the asymptotic dependence on ϑ.

The paper is organized as follows. In Section 2 we state the results on the four quantities mentioned above. The proofs are collected in Section 3.

2. Results

In this section the results on the depth, height, internal path length, and number of leaves are stated. Throughout, the parameter $\vartheta > 0$ is arbitrary and fixed. All asymptotic statements as well as the use of the Bachmann–Landau symbols are understood as n, the number of nodes in the Hoppe tree, tends to ∞. Moreover, we use the digamma and trigamma functions $\Psi = \mathrm{d}\log\Gamma/\mathrm{d}x$ and $\Psi_1 = \mathrm{d}^2\log\Gamma/\mathrm{d}x^2$, respectively. By the properties of the digamma and trigamma functions, see, e.g. [1, Sections 6.3 and 6.4], we have

$$\sum_{i=1}^{n-2} \frac{1}{\vartheta + i} = \Psi(\vartheta + n - 1) - \Psi(\vartheta + 1) = \log n - \Psi(\vartheta + 1) + o(1),$$

$$\sum_{k=1}^{\infty} \left(\frac{1}{\vartheta + k}\right)^2 = \Psi'(\vartheta + 1) = \Psi_1(\vartheta + 1).$$

Depth of a node. For the depth $D_n^{(\vartheta)}$, we have a distributional representation as the sum of independent Bernoulli variables.

Theorem 2.1. *For the depth $D_n^{(\vartheta)}$ of the nth node in a Hoppe tree, we have, for all $n \ge 2$,*

$$D_n^{(\vartheta)} \stackrel{\mathrm{D}}{=} 1 + \sum_{i=1}^{n-2} B_i,$$

where $B_1, \dots, B_{n-2}$ are independent and $\mathbb{P}(B_i = 1) = 1 - \mathbb{P}(B_i = 0) = 1/(\vartheta + i)$ for $i = 1, \dots, n-2$.

Asymptotic results can therefore be easily obtained, as we show in Corollary 2.1 below. We denote by $\Pi(\lambda)$ the Poisson distribution with parameter $\lambda > 0$, by d_{TV} the total variation distance between probability measures, by '$\stackrel{\mathrm{D}}{\to}$' convergence in distribution, and by $\mathcal{N}(0, 1)$ a real random variable with the standard normal distribution.

Corollary 2.1. *The depth $D_n^{(\vartheta)}$ of the nth node in a Hoppe tree satisfies*

$$\mathbb{E}[D_n^{(\vartheta)}] = 1 + \sum_{i=1}^{n-2} \frac{1}{\vartheta + i} = \log n - \Psi(\vartheta + 1) + 1 + o(1),$$

https://doi.org/10.1239/jap/1363784435 Published online by Cambridge University Press

$$\operatorname{var}(D_n^{(\vartheta)}) = \sum_{i=1}^{n-2} \frac{1}{\vartheta+i} - \sum_{i=1}^{n-2} \left(\frac{1}{\vartheta+i}\right)^2 = \log n - \Psi(\vartheta+1) - \Psi_1(\vartheta+1) + o(1),$$

$$\frac{D_n^{(\vartheta)} - \mathbb{E}[D_n^{(\vartheta)}]}{\sqrt{\operatorname{var}(D_n^{(\vartheta)})}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1), \tag{2.1}$$

$$d_{\mathrm{TV}}(\mathcal{L}(D_n^{(\vartheta)}), \Pi(\mathbb{E}[D_n^{(\vartheta)}])) = \mathcal{O}\left(\frac{1}{\log n}\right).$$

Height of the Hoppe tree. The height $H_n^{(\vartheta)}$ of the Hoppe tree can be analyzed using results on the height of random recursive trees; see Addario-Berry and Ford [2], who, in particular, showed that

$$M_n := \mathbb{E}[H_n^{(1)}] = \mathrm{e}\log n - \tfrac{3}{2}\log\log n + \mathcal{O}(1) \tag{2.2}$$

as $n \to \infty$. We transfer their results to arbitrary $\vartheta > 0$.

Theorem 2.2. *For the height $H_n^{(\vartheta)}$ of a Hoppe tree with n nodes, we have, for all $\alpha < 1/3\mathrm{e}$ and $\beta < 1/2\mathrm{e}$, there exist constants $C_\alpha, C_\beta > 0$ such that, for all $t > 0$,*

$$\mathbb{P}(H_n^{(\vartheta)} - M_n \geq t) \leq C_\beta \mathrm{e}^{-\beta t}, \qquad \mathbb{P}(H_n^{(\vartheta)} - M_n \leq -t) \leq C_\alpha \mathrm{e}^{-\alpha t}.$$

The constant C_β can be chosen independently of ϑ.

Corollary 2.2. *The height $H_n^{(\vartheta)}$ of a Hoppe tree with n nodes satisfies*

$$\mathbb{E}[H_n^{(\vartheta)}] = \mathrm{e}\log n - \tfrac{3}{2}\log\log n + \mathcal{O}(1), \qquad \operatorname{var}(H_n^{(\vartheta)}) = \mathcal{O}(1).$$

Number of leaves. The number of leaves in a Hoppe tree is related to a two-color urn model.

Theorem 2.3. *Let $L_n^{(\vartheta)}$ be the number of leaves in a Hoppe tree with $n \geq 2$ nodes. Then*

$$\mathbb{E}[L_n^{(\vartheta)}] = \frac{n}{2} + \frac{\vartheta-1}{2} + \mathcal{O}\left(\frac{1}{n}\right),$$

$$\operatorname{var}(L_n^{(\vartheta)}) = \frac{n}{12} + \frac{\vartheta-1}{12} + \mathcal{O}\left(\frac{1}{n}\right),$$

$$\mathbb{P}(|L_n - \mathbb{E}[L_n]| \geq t) \leq 2\exp\left(-\frac{6t^2}{n+\vartheta+1}\right) \quad \textit{for all } t > 0,\ n \geq 1, \tag{2.3}$$

$$\frac{L_n^{(\vartheta)} - \mathbb{E}[L_n^{(\vartheta)}]}{\sqrt{\operatorname{var}(L_n^{(\vartheta)})}} \xrightarrow{\mathrm{D}} \mathcal{N}(0,1).$$

Internal path length. Moments of the internal path length can be obtained from our results on the depths of the nodes.

Theorem 2.4. *The internal path length $I_n^{(\vartheta)}$ of a Hoppe tree with n nodes satisfies*

$$\mathbb{E}[I_n^{(\vartheta)}] = (\vartheta+n-1)\sum_{i=1}^{n-1}\frac{1}{\vartheta+i} = n\log n - \Psi(\vartheta+1)n + o(n),$$

$$\operatorname{var}(I_n^{(\vartheta)}) = \left(\frac{2}{\vartheta+1} - \Psi_1(\vartheta+1)\right)n^2 + o(n^2).$$

https://doi.org/10.1239/jap/1363784435 Published online by Cambridge University Press

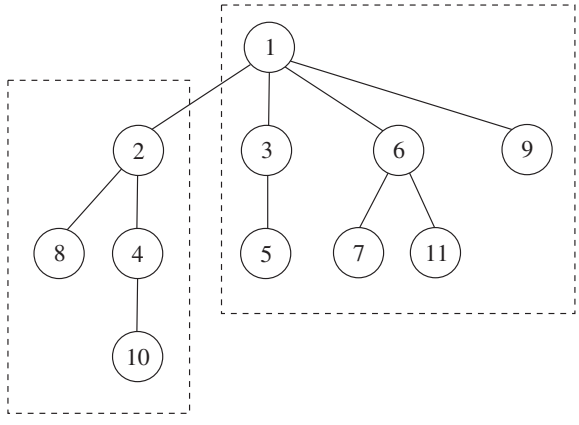

FIGURE 1: A Hoppe tree with 11 nodes. The decomposition into the subtree rooted at the node labelled 2 and the remaining part of the tree is indicated using dashed boxes.

Moreover,

$$\left(\frac{I_n^{(\vartheta)} - \mathbb{E}[I_n^{(\vartheta)}]}{\vartheta + n - 1}\right)_{n\geq 1}$$

is a zero-mean martingale.

The internal path length can be analyzed either via martingale methods or the recursive distributional decomposition explained in Figure 1, which allows us to apply the contraction method.

Theorem 2.5. *The internal path length $I_n^{(\vartheta)}$ of a Hoppe tree with n nodes satisfies*

$$\frac{I_n^{(\vartheta)} - n\log n}{n} \to X^{(\vartheta)}$$

for a nondegenerate random variable $X^{(\vartheta)}$, where the convergence holds almost surely and in L_2. The distribution $\mathcal{L}(X^{(\vartheta)})$ is the only integrable solution of the distributional fixed point equation

$$X^{(\vartheta)} \stackrel{\mathrm{D}}{=} (1-B)X^{(\vartheta)} + B\tilde{X}^{(1)} + B\log(B) + (1-B)\log(1-B) + B, \tag{2.4}$$

where $X^{(\vartheta)}$, $\tilde{X}^{(1)}$, and B are independent, B has the beta$(1,\vartheta)$ *distribution, and $\tilde{X}^{(1)}$ is distributed as $X^{(1)}$. For $\vartheta \neq 1$, the solution of (2.4) is unique even without the integrability assumption.*

Theorem 2.6. *The limit distribution $\mathcal{L}(X^{(\vartheta)})$ in Theorem 2.5 has a Lebesgue density f_ϑ, which is in the Schwartz space on $\mathbb{R}$, i.e. f_ϑ is infinitely differentiable and, together with all its derivatives, rapidly decreasing.*

3. Proofs

In the analysis of the tree below the random decomposition of the Hoppe tree shown in Figure 1 is used. The tree is decomposed into the subtree of the second inserted node (left dashed box of Figure 1) and the remaining part of the tree (right dashed box of Figure 1).

https://doi.org/10.1239/jap/1363784435 Published online by Cambridge University Press

The stochastic dynamic of the Hoppe tree with parameter ϑ implies that, conditioned on the size N_n of the subtree of the second inserted node, this subtree is a random recursive tree, whereas the remaining part is a Hoppe tree with parameter ϑ and size $n - N_n$. Moreover, conditional on N_n, these two trees are independent. We have the asymptotic behavior

$$\frac{N_n}{n} \to B \quad \text{almost surely as } n \to \infty, \tag{3.1}$$

where B has the beta$(1, \vartheta)$ distribution with Lebesgue density $x \mapsto \vartheta(1-x)^{\vartheta-1}$, $x \in [0, 1]$; see [6].

Proof of Theorem 2.1. We calculate the depth of a node by counting its ancestors in the tree. We have $D_n^{(\vartheta)} = \sum_{i=1}^{n-1} \mathbf{1}_{A_{i,n}}$, where $A_{i,j}$ denotes the event that node i is an ancestor of node j, $i < j$. Clearly, $\mathbb{P}(A_{1,n}) = 1$. Moreover, $\mathbb{P}(A_{i,i+1}) = 1/(\vartheta + i - 1)$ for $i \geq 2$ by definition of the Hoppe tree. For general $i < n$, let $\xi_{i,n}$ be the number of descendants of node i in a Hoppe tree with n nodes, i.e. the size of the subtree rooted at i minus 1. By the dynamics of the Hoppe tree we have

$$\mathbb{P}(A_{i,n} \mid \xi_{i,n-1}) = \frac{1 + \xi_{i,n-1}}{\vartheta + n - 2}. \tag{3.2}$$

We calculate $\mathbb{E}[\xi_{i,n-1}]$ by the recursion

$$\mathbb{E}[\xi_{i,n-1}] = \mathbb{E}[\xi_{i,n-2} + \mathbf{1}_{A_{i,n-1}}] = \mathbb{E}[\xi_{i,n-2}] + \frac{1 + \mathbb{E}[\xi_{i,n-2}]}{\vartheta + n - 3}.$$

This yields $\mathbb{E}[\xi_{i,n-1}] = (\vartheta + n - 2)/(\vartheta + i - 1) - 1$ and, therefore, by (3.2),

$$\mathbb{P}(A_{i,n}) = \frac{1}{\vartheta + i - 1}. \tag{3.3}$$

It remains to show that $A_{2,n}, \ldots, A_{n-1,n}$ are independent. Note that, for $i < j$, $A_{i,j}$ depends only on where the nodes $i + 1, \ldots, j$ are inserted. Therefore, we get, for all $2 \leq k \leq n - 2$ and $2 \leq i_1 < \cdots < i_k \leq n - 1$, independence of $A_{i_1,i_2}, A_{i_2,i_3}, \ldots, A_{i_k,n}$. Since $\bigcap_{j=1}^{k} A_{i_j,n}$ occurs if and only if i_j is an ancestor of i_{j+1} for every $j \leq k - 1$ and i_k is an ancestor of n, we have

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{j=1}^{k} A_{i_j,n}\right) &= \mathbb{P}(A_{i_1,i_2} \cap A_{i_2,i_3} \cap \cdots \cap A_{i_k,n}) \\
&= \mathbb{P}(A_{i_1,i_2})\mathbb{P}(A_{i_2,i_3}) \cdots \mathbb{P}(A_{i_k,n}) \\
&= \prod_{j=1}^{k} \mathbb{P}(A_{i_j,n}),
\end{aligned}$$

where (3.3) was used in the last equality. With $B_i = \mathbf{1}_{A_{i+1,n}}$ and $\mathbf{1}_{A_{1,n}} = 1$, this yields the assertion.

For related reasoning in the analysis of the depth in other random tree models, see [5].

Proof of Corollary 2.1. Theorem 2.1 implies the expectation and variance of $D_n^{(\vartheta)}$. Moreover, by Lindeberg's version of the central limit theorem (CLT) we obtain the CLT for $D_n^{(\vartheta)}$ in (2.1) and by [3, Equation (1.23)] we get $d_{\mathrm{TV}}(\mathcal{L}(D_n^{(\vartheta)}), \Pi(\mathbb{E}[D_n^{(\vartheta)}])) = \mathcal{O}(1/\log n)$.

https://doi.org/10.1239/jap/1363784435 Published online by Cambridge University Press

Proof of Theorem 2.2. Addario-Berry and Ford showed in [2, Corollary 1.3] that the expected height $M_n := \mathbb{E}[H^{(1)}]$ of a random recursive tree satisfies (2.2) and that, for all $c' < 1/2\mathrm{e}$, there exists a constant $C = C(c')$ such that, for all $n \geq 1$ and $t > 0$,

$$\mathbb{P}(|H_n^{(1)} - M_n| \geq t) \leq C\mathrm{e}^{-c't}.$$

Recall that in a Hoppe tree with $n \geq 1$ nodes and parameter $\vartheta > 0$, we denote by N_n the size of the subtree rooted in node 2 and that this subtree, conditioned on its size, is a random recursive tree.

By an obvious coupling argument between Hoppe trees for different parameters ϑ we have $H_n^{(\vartheta_1)} \preccurlyeq H_n^{(\vartheta_2)}$ for all $\vartheta_1 \geq \vartheta_2$, where '$\preccurlyeq$' denotes stochastic domination. In the extremal case $\vartheta = 0$ (for the definition of the tree, start with the root and one child) we obtain $H_n^{(\vartheta)} \preccurlyeq H_n^{(0)} \stackrel{\mathrm{D}}{=} 1 + H_{n-1}^{(1)} \preccurlyeq 1 + H_n^{(1)}$. Therefore, we get $\mathbb{P}(H_n^{(\vartheta)} - M_n \geq t) \leq \hat{C}\mathrm{e}^{-c't}$, $\hat{C} = C\mathrm{e}^{c'}$, using the result for random recursive trees.

In order to prove the left tail inequality, let $H_{N_n}^{(1)}$ be the height of the subtree rooted at node 2. From $H_n^{(\vartheta)} \geq H_{N_n}^{(1)}$ we obtain, for all $t > 0$ and $\alpha > 0$ (later we have to restrict to α as in the theorem),

$$\begin{aligned}\mathbb{P}(H_n^{(\vartheta)} - M_n \leq -t) &\leq \mathbb{P}(\{H_{N_n}^{(1)} - M_n \leq -t\} \cap \{N_n \geq \mathrm{e}^{-\alpha t}n\}) \\ &\quad + \mathbb{P}(\{H_{N_n}^{(1)} - M_n \leq -t\} \cap \{N_n < \mathrm{e}^{-\alpha t}n\}) \\ &\leq \mathbb{P}(H_{\lceil \mathrm{e}^{-\alpha t}n\rceil}^{(1)} - M_n \leq -t) + \mathbb{P}(N_n < \mathrm{e}^{-\alpha t}n).\end{aligned}$$

Again, by using the result for random recursive trees and $M_n - \mathbb{E}[H_{\lceil \mathrm{e}^{-\alpha t}n\rceil}^{(1)}] = \mathrm{e}\alpha t + \mathcal{O}(1)$, we obtain, for $\alpha = c'/(1 + \mathrm{e}c')$, a constant C_1 such that

$$\mathbb{P}(H_{\lceil \mathrm{e}^{-\alpha t}n\rceil}^{(1)} - M_n \leq -t) \leq C_1\mathrm{e}^{-c'(1-\mathrm{e}\alpha)t} = C_1\mathrm{e}^{-\alpha t}.$$

Hence, we have such an upper bound for all $\alpha < 1/3\mathrm{e}$. To obtain an upper bound for $\mathbb{P}(N_n < \mathrm{e}^{-\alpha t}n)$ note that, for all $1 \leq k \leq n-1$,

$$\mathbb{P}(N_n = k) = \binom{n-2}{k-1} \frac{\vartheta(\vartheta+1)\cdots(\vartheta+n-(k+2))(k-1)!}{(\vartheta+1)\cdots(\vartheta+n-2)}.$$

This yields, for all $\varepsilon \in (0, 1)$,

$$\mathbb{P}(N_n \leq \varepsilon n) \leq 3(\vartheta+1)\varepsilon.$$

Therefore,

$$\mathbb{P}(H_n^{(\vartheta)} - M_n \leq -t) \leq (C_1 + 3(\vartheta+1))\mathrm{e}^{-\alpha t}.$$

This implies the assertion.

Proof of Corollary 2.2. By Theorem 2.2 we have

$$\mathbb{E}[|H_n^{(\vartheta)} - M_n|] = \mathcal{O}(1).$$

Consequently, $\mathbb{E}[H_n^{(\vartheta)}] = M_n + \mathcal{O}(1) = \mathrm{e}\log n - \frac{3}{2}\log\log n + \mathcal{O}(1)$.

Moreover, the tail bound from Theorem 2.2 implies that

$$\mathrm{var}(H_n^{(\vartheta)}) \leq \mathbb{E}[(H_n^{(\vartheta)} - M_n)^2] = \mathcal{O}(1).$$

https://doi.org/10.1239/jap/1363784435 Published online by Cambridge University Press

For the proof of the tail bound in Theorem 2.3, we use the following version of Azuma–Hoeffding's inequality with conditional ranges.

Proposition 3.1. *Let $W_1, \dots, W_n$ be a martingale difference sequence with respect to a filtration $(\mathcal{F}_i)_{0\le i\le n}$ with $\mathcal{F}_0 = \{\varnothing, \Omega\}$. Suppose that, for every $1 \le i \le n$, there exists a constant $c_i \ge 0$ and an $\mathcal{F}_{i-1}$-measurable random variable Z_i such that $Z_i \le W_i \le Z_i + c_i$ almost surely. Then we have, for all $t > 0$,*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} W_i\right| \ge t\right) \le 2\exp\left(-\frac{2t^2}{\sum_{i=1}^{n} c_i^2}\right).$$

Proof of Theorem 2.3. We have $L_n^{(\vartheta)} = L_{n-1}^{(\vartheta)} + Y_n$, where

$$Y_n = \begin{cases} 1 & \text{if the parent of node } n \text{ was not a leaf at time } n-1, \\ 0 & \text{otherwise.} \end{cases}$$

Therefore, for $n \ge 2$, almost surely,

$$\mathbb{E}[L_{n+1}^{(\vartheta)} \mid L_1^{(\vartheta)}, \dots, L_n^{(\vartheta)}] = L_n^{(\vartheta)} + 1 - \frac{L_n^{(\vartheta)}}{\vartheta + n - 1} = \frac{\vartheta + n - 2}{\vartheta + n - 1} L_n^{(\vartheta)} + 1.$$

With

$$X_n = (\vartheta + n - 2)\left(L_n^{(\vartheta)} - \left(\frac{n-1}{2} + \frac{\vartheta(n-1)}{2(\vartheta + n - 2)}\right)\right), \tag{3.4}$$

the sequence $(X_n)_{n\ge 2}$ is a zero-mean martingale and

$$\mathbb{E}[L_n^{(\vartheta)}] = \frac{n-1}{2} + \frac{\vartheta(n-1)}{2(\vartheta + n - 2)} = \frac{\vartheta + n - 1}{2} + \mathcal{O}\left(\frac{1}{n}\right).$$

With the representation

$$X_i - X_{i-1} = (\vartheta + i - 2)(Y_i - \mathbb{E}[Y_i]) + L_{i-1}^{(\vartheta)} - \mathbb{E}[L_{i-1}^{(\vartheta)}], \qquad i \ge 3,$$

we have $Z_i \le X_i - X_{i-1} \le Z_i + \vartheta + i - 2$, where $Z_i = L_{i-1}^{(\vartheta)} - \mathbb{E}[L_{i-1}^{(\vartheta)}] - (\vartheta + i - 2)\mathbb{E}[Y_i]$. By Proposition 3.1 we have, for all $t > 0$,

$$\mathbb{P}(|X_n| \ge t) \le 2\exp\left(-\frac{2t^2}{\sum_{i=3}^{n}(i + \vartheta - 2)^2}\right).$$

Using the fact that the sum in the denominator of the latter exponent is bounded by $(n + \vartheta - 2)^3/3 + (n + \vartheta - 2)^2$ and the scaling in (3.4), we obtain (2.3).

In order to compute $\operatorname{var}(L_n^{(\vartheta)})$, we have $X_n = (\vartheta + n - 2)X_{n-1}/(\vartheta + n - 3) + (\vartheta + n - 2)(Y_n - \mathbb{E}[Y_n])$. Hence,

$$\begin{aligned}\mathbb{E}[X_n^2] = \left(\frac{\vartheta + n - 2}{\vartheta + n - 3}\right)^2 \mathbb{E}[X_{n-1}^2] + 2\frac{(\vartheta + n - 2)^2}{\vartheta + n - 3}\mathbb{E}[X_{n-1}(Y_n - \mathbb{E}[Y_n])] \\ + (\vartheta + n - 2)^2 \operatorname{var}(Y_n).\end{aligned} \tag{3.5}$$

https://doi.org/10.1239/jap/1363784435 Published online by Cambridge University Press

Using $\mathbb{E}[X_{n-1}] = 0$, we have

$$\begin{aligned}\mathbb{E}[X_{n-1}(Y_n - \mathbb{E}[Y_n])] &= \mathbb{E}[X_{n-1}\mathbb{E}[Y_n \mid L_1^{(\vartheta)}, \dots, L_{n-1}^{(\vartheta)}]] \\ &= \mathbb{E}\left[X_{n-1}\left(1 - \frac{L_{n-1}^{(\vartheta)}}{\vartheta + n - 2}\right)\right] \\ &= -\frac{1}{(\vartheta + n - 2)(\vartheta + n - 3)}\mathbb{E}[X_{n-1}^2].\end{aligned}$$

Moreover, $\mathbb{E}[Y_n] = 1 - \mathbb{E}[L_{n-1}^{(\vartheta)}]/(\vartheta + n - 2) = \frac{1}{2} + \mathcal{O}(1/n^2)$ and $\operatorname{var}(Y_n) = \frac{1}{4} + \mathcal{O}(1/n^2)$.

Solving (3.5) by the substitution $Q_n = (\vartheta + n - 3)\mathbb{E}[X_n^2]/(\vartheta + n - 2)$ yields

$$\operatorname{var}(L_n^{(\vartheta)}) = \frac{\vartheta + n - 1}{12} + \mathcal{O}\left(\frac{1}{n}\right).$$

To obtain the CLT for $L_n^{(\vartheta)}$, the representation

$$\frac{L_n^{(\vartheta)} - \mathbb{E}[L_n^{(\vartheta)}]}{\sqrt{\operatorname{var}(L_n^{(\vartheta)})}} = \frac{X_n}{\sqrt{\operatorname{var}(X_n)}}$$

allows us to apply a general martingale CLT; see, e.g. [8, Theorem 3.2]. It is sufficient to show that

$$\Delta_{n,i} := \frac{1}{\sqrt{\operatorname{var}(X_n)}}(X_i - X_{i-1}), \qquad n \geq 3,\ 3 \leq i \leq n,$$

satisfies

(a) $\max_{3\leq i\leq n} |\Delta_{n,i}| \xrightarrow{\mathbb{P}} 0$,

(b) $\sum_{3\leq i\leq n} \Delta_{n,i}^2 \xrightarrow{\mathbb{P}} 1$,

(c) $\max_{n\geq 3} \mathbb{E}[\max_{3\leq i\leq n} \Delta_{n,i}^2] < \infty$.

For (a) and (c), we have

$$|X_i - X_{i-1}| = |L_i^{(\vartheta)} - \mathbb{E}[L_i^{(\vartheta)}] + (\vartheta + i - 3)(Y_i - \mathbb{E}[Y_i])| \leq \vartheta + 2n + 3 \quad \text{for } i \leq n$$

and $\operatorname{var}(X_n) = (\vartheta + n - 1)^2 \operatorname{var}(L_n^{(\vartheta)}) \sim n^3/12$. Hence, $|\Delta_{n,i}| \leq (2n + \vartheta + 3)/\sqrt{\operatorname{var}(X_n)}$ almost surely, which reveals that $\max_i |\Delta_{n,i}| \xrightarrow{\mathbb{P}} 0$ and that $\mathbb{E}[\max_i \Delta_{n,i}^2]$ is bounded in n.

To compute $\sum_i \Delta_{n,i}^2$, note that, by (2.3) and the Borel–Cantelli lemma, we have $(L_n^{(\vartheta)} - \mathbb{E}[L_n^{(\vartheta)}])/n \to 0$ almost surely. Hence, for all $n \geq 3$,

$$\begin{aligned}\sum_{i=3}^{n} \Delta_{n,i}^2 &= \frac{1}{\operatorname{var}(X_n)} \sum_{i=3}^{n} (L_i^{(\vartheta)} - \mathbb{E}[L_i^{(\vartheta)}])^2 \\ &\quad + \frac{2}{\operatorname{var}(X_n)} \sum_{i=3}^{n} (L_i^{(\vartheta)} - \mathbb{E}[L_i^{(\vartheta)}])(\vartheta + i - 3)(Y_i - \mathbb{E}[Y_i]) \\ &\quad + \frac{1}{\operatorname{var}(X_n)} \sum_{i=3}^{n} (\vartheta + i - 3)^2 (Y_i - \mathbb{E}[Y_i])^2. \qquad (3.6)\end{aligned}$$

https://doi.org/10.1239/jap/1363784435 Published online by Cambridge University Press

By $(L_n^{(\vartheta)} - \mathbb{E}[L_n^{(\vartheta)}])/n \to 0$, $\operatorname{var}(X_n) \sim n^3/12$, and the Cesàro mean, we have

$$\frac{1}{\operatorname{var}(X_n)} \sum_{i=3}^{n} (L_i^{(\vartheta)} - \mathbb{E}[L_i^{(\vartheta)}])^2 \le \frac{n^3}{\operatorname{var}(X_n)} \frac{1}{n} \sum_{i=3}^{n} \left(\frac{L_i^{(\vartheta)} - \mathbb{E}[L_i^{(\vartheta)}]}{i} \right)^2 \to 0$$

for the first summand in (3.6) and

$$\begin{aligned} &\left| \frac{2}{\operatorname{var}(X_n)} \sum_{i=3}^{n} (L_i^{(\vartheta)} - \mathbb{E}[L_i^{(\vartheta)}])(\vartheta + i - 3)(Y_i - \mathbb{E}[Y_i]) \right| \\ &\quad \le \frac{2n^2(\vartheta + n + 3)}{\operatorname{var}(X_n)} \frac{1}{n} \sum_{i=3}^{n} \left| \frac{L_i^{(\vartheta)} - \mathbb{E}[L_i^{(\vartheta)}]}{i} \right| \\ &\quad \to 0 \end{aligned}$$

for the second summand in (3.6). Because $\mathbb{E}[Y_i] = \frac{1}{2} + \mathcal{O}(1/i^2)$ we have $(Y_i - \mathbb{E}[Y_i])^2 = \frac{1}{4} + \mathcal{O}(1/i^2)$ almost surely and, therefore,

$$\frac{1}{\operatorname{var}(X_n)} \sum_{i=3}^{n} (\vartheta + i - 3)^2 (Y_i - \mathbb{E}[Y_i])^2 \to 1 \quad \text{almost surely}$$

for the last summand in (3.6). This implies that $\sum_i \Delta_{n,i}^2 \xrightarrow{\mathbb{P}} 1$.

Proof of Theorem 2.4. For $j \ge 1$, let $\mathcal{F}_j = \sigma(D_1^{(\vartheta)}, \dots, D_j^{(\vartheta)})$. By the dynamics of the Hoppe tree we have, almost surely,

$$\begin{aligned} \mathbb{E}[D_n^{(\vartheta)} \mid \mathcal{F}_{n-1}] &= \frac{\vartheta}{\vartheta + n - 2}(D_1^{(\vartheta)} + 1) + \sum_{i=2}^{n-1} \frac{1}{\vartheta + n - 2}(D_i^{(\vartheta)} + 1) \\ &= 1 + \frac{1}{\vartheta + n - 2} I_{n-1}^{(\vartheta)}. \end{aligned}$$

Consequently,

$$\mathbb{E}[I_n^{(\vartheta)} \mid \mathcal{F}_{n-1}] = I_{n-1}^{(\vartheta)} + \mathbb{E}[D_n^{(\vartheta)} \mid \mathcal{F}_{n-1}] = \frac{\vartheta + n - 1}{\vartheta + n - 2} I_{n-1}^{(\vartheta)} + 1$$

almost surely. Therefore,

$$Z_n^{(\vartheta)} := \frac{1}{\vartheta + n - 1} I_n^{(\vartheta)} - \sum_{i=1}^{n-1} \frac{1}{\vartheta + i}$$

is a zero-mean martingale and $\mathbb{E}[I_n^{(\vartheta)}] = (\vartheta + n - 1) \sum_{i=1}^{n-1} 1/(\vartheta + i)$.

The calculations to obtain the expansion for the variance of $I_n^{(\vartheta)}$ are similar to those carried out in the proof of Theorem 2.3; for details, we refer the reader to [10].

Proof of Theorem 2.5. To apply a martingale convergence theorem, it is sufficient to have a bound on the variance of the martingale uniformly in n. Hence, our expansion of $\operatorname{var}(I_n^{(\vartheta)})$ in Theorem 2.4 is sufficient to imply almost-sure and L_2 convergence of the martingale there,

https://doi.org/10.1239/jap/1363784435 Published online by Cambridge University Press

which also applies to the slightly different scaling of $I_n^{(\vartheta)}$ in Theorem 2.5. By our decomposition of the Hoppe tree, see Figure 1, we obtain the recurrence

$$I_n^{(\vartheta)} \stackrel{\mathrm{D}}{=} I_{n-N_n}^{(\vartheta)} + \tilde{I}_{N_n}^{(1)} + N_n, \tag{3.7}$$

where $(I_j^{(\vartheta)})_{j\geq 1}$, $(\tilde{I}_j^{(1)})_{j\geq 1}$, and N_n are independent, and $(\tilde{I}_j^{(1)})_{j\geq 1}$ is distributed as $(I_j^{(1)})_{j\geq 1}$. For the scaling

$$X_n^{(\vartheta)} := \frac{I_n^{(\vartheta)} - n\log n}{n},$$

we obtain

$$\begin{aligned} X_n^{(\vartheta)} \stackrel{\mathrm{D}}{=} {} & \frac{n-N_n}{n} X_{n-N_n}^{(\vartheta)} + \frac{N_n}{n}\tilde{X}_{N_n}^{(1)} \\ & + \frac{1}{n}\left(N_n \log\left(\frac{N_n}{n}\right) + (n-N_n)\log\left(\frac{n-N_n}{n}\right) + N_n\right), \end{aligned} \tag{3.8}$$

with independence and distributional conditions as in (3.7). This suggests that the limit $X^{(\vartheta)}$ of $(X_n^{(\vartheta)})_{n\geq 1}$ should satisfy the recursive distributional equation

$$X^{(\vartheta)} \stackrel{\mathrm{D}}{=} (1-B)X^{(\vartheta)} + B\tilde{X}^{(1)} + B\log(B) + (1-B)\log(1-B) + B, \tag{3.9}$$

where $X^{(\vartheta)}$, $\tilde{X}^{(1)}$, and B are independent, and B has the beta$(1, \vartheta)$ distribution. Note that $\tilde{X}^{(1)}$ is the limit distribution of the internal path length of the random recursive tree, which has been obtained by martingale methods in [11] and by the contraction method in [4]. In particular, in [4] it was shown that $(X_n^{(1)})_{n\geq 1}$ converges to its limit $X^{(1)}$ in the minimal ℓ_2 metric, i.e. weakly and with second moments. This allows us to write recurrence (3.8) in the form

$$X_n^{(\vartheta)} \stackrel{\mathrm{D}}{=} A^{(n)} X_{n-N_n}^{(\vartheta)} + b^{(n)}$$

with coefficients

$$\begin{aligned} A^{(n)} &= \frac{n-N_n}{n}, \\ b^{(n)} &= \frac{N_n}{n}\tilde{X}_{N_n}^{(1)} + \frac{1}{n}\left(N_n \log\left(\frac{N_n}{n}\right) + (n-N_n)\log\left(\frac{n-N_n}{n}\right) + N_n\right). \end{aligned}$$

Hence, we have convergence of the coefficients to the corresponding quantities in the recursive distributional equation (3.9) in ℓ_1 and ℓ_2, in fact in any ℓ_p, $p \geq 1$. This allows us to apply general convergence theorems in the framework of the contraction method; see [14, Theorem 3] and [12, Theorem 4.1]. In particular, one can first apply Theorem 4.1 of [12] with the choice $s = 1$. This implies convergence in distribution of $X_n^{(\vartheta)}$ to $X^{(\vartheta)}$, where $X^{(\vartheta)}$ is the unique integrable solution of (3.9), and convergence of the expectations. With this knowledge on the expectation, which, of course, is also covered by our explicit formula for $\mathbb{E}[I_n^{(\vartheta)}]$, one can apply either Theorem 4.1 of [12] with the choice $s = 2$ or Theorem 3 of [14] to also obtain convergence of the second moments.

Alternatively to applying the contraction method we could as well use the almost-sure convergence of $X_n^{(\vartheta)}$ from the martingale argument together with the almost-sure convergence of N_n/n in (3.1) to argue that the limit $X^{(\vartheta)}$ satisfies (3.9).

https://doi.org/10.1239/jap/1363784435 Published online by Cambridge University Press

Proof of Theorem 2.6. For the characteristic function $\varphi_\vartheta(t) := \mathbb{E}[\exp(\mathrm{i}tX^{(\vartheta)})]$ of $X^{(\vartheta)}$, the recursive distributional equation in Theorem 2.5 implies that

$$|\varphi_\vartheta(t)| \le \int_0^1 |\varphi_1(xt)||\varphi_\vartheta((1-x)t)|\vartheta(1-x)^{\vartheta-1}\,\mathrm{d}x, \qquad t \in \mathbb{R}.$$

We can apply the techniques of Fill and Janson [7] to show that this relation together with an initial bound on $|\varphi_\vartheta|$ allows us to show that $|\varphi_\vartheta|$ is rapidly decreasing. The details are carried out in [10]. Since the Fourier transform is an automorphism on the Schwartz space, this implies the assertion.

Acknowledgements

We thank Henning Sulzbach for comments on a draft of this paper and two anonymous referees for their careful reading.

References

[1] Abramowitz, M. and Stegun, I. A. (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables* (Natl. Bureau Standards Appl. Math. Ser. **55**). Government Printing Office, Washington, DC.
[2] Addario-Berry, L. and Ford, K. (2013). Poisson–Dirichlet branching random walks. *Ann. Appl. Prob.* **23,** 283–307.
[3] Barbour, A. D., Holst, L. and Janson, S. (1992). *Poisson Approximation* (Oxford Stud. Prob. **2**). The Clarendon Press, Oxford University Press, New York.
[4] Dobrow, R. P. and Fill, J. A. (1999). Total path length for random recursive trees. *Combin. Prob. Comput.* **8,** 317–333.
[5] Dobrow, R. P. and Smythe, R. T. (1996). Poisson approximations for functionals of random trees. *Random Structures Algorithms* **9,** 79–92.
[6] Donnelly, P. and Tavaré, S. (1986). The ages of alleles and a coalescent. *Adv. Appl. Prob.* **18,** 1–19. (Correction: **18** (1986), 1023.)
[7] Fill, J. A. and Janson, S. (2000). Smoothness and decay properties of the limiting Quicksort density function. In *Mathematics and Computer Science* (Versailles, 2000), Birkhäuser, Basel, pp. 53–64.
[8] Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.
[9] Hoppe, F. M. (1986). Size-biased filtering of Poisson–Dirichlet samples with an application to partition structures in genetics. *J. Appl. Prob.* **23,** 1008–1012.
[10] Leckey, K. (2011). Asymptotische Eigenschaften von Hoppe–Bäumen. Masters Thesis, Goethe Universität Frankfurt a.M. Available at http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/24214.
[11] Mahmoud, H. M. (1991). Limiting distributions for path lengths in recursive trees. *Prob. Eng. Inf. Sci.* **5,** 53–59.
[12] Neininger, R. and Rüschendorf, L. (2004). A general limit theorem for recursive algorithms and combinatorial structures. *Ann. Appl. Prob.* **14,** 378–418.
[13] Pitman, J. (2006). *Combinatorial Stochastic Processes* (Lecture Notes Math. **1875**). Springer, Berlin.
[14] Rösler, U. (2001). On the analysis of stochastic divide and conquer algorithms. *Algorithmica* **29,** 238–261.
[15] Smythe, R. T. and Mahmoud, H. M. (1994). A survey of recursive trees. *Teor. Ĭmovīr. Mat. Statist.* **51,** 1–29. English translation: *Theory Prob. Math. Statist.* **51** (1995), 1–27.