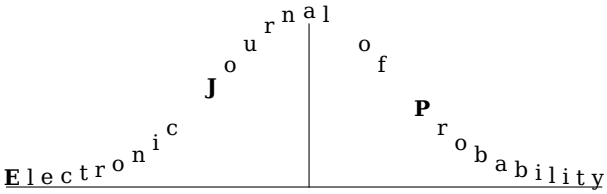

Electron. J. Probab. **20** (2015), no. 37, 1–22.
ISSN: 1083-6489 DOI: 10.1214/EJP.v20-3832

Random recursive trees: a boundary theory approach*

Rudolf Grübel† Igor Michailow‡

Abstract

We show that an algorithmic construction of sequences of recursive trees leads to a direct proof of the convergence of random recursive trees in an associated Doob-Martin compactification; it also gives a representation of the limit in terms of the input sequence of the algorithm. We further show that this approach can be used to obtain strong limit theorems for various tree functionals, such as path length or the Wiener index.

Keywords: Doob-Martin compactification; Markov chains; path length; random trees; Harris trees; Wiener index.
AMS MSC 2010: Primary 60C05, Secondary 05C05; 60J50; 68Q87.
Submitted to EJP on October 1, 2014, final version accepted on March 25, 2015.
Supersedes arXiv:1406.7614.

1 Introduction

A tree with node set $[n] := \{1, \dots, n\}$ is recursive if the node numbers along the unique path from 1 to j increase for $j = 2, \dots, n$. Trees with this property may be encoded by a sequence $(j_1, \dots, j_{n-1})$, where $j_k \in [k]$ denotes the direct ancestor of $k+1$ (next node on the way to the 'root', which is at 1). Such a sequence also gives a recipe for growing the corresponding tree: Starting with the unique recursive tree of size (number of nodes) 1, which consists of the root node only, we obtain the respective next tree by joining node k to node j_{k-1}, $k = 2, \dots, n$. Choosing the ancestor of the next node uniformly at random among the nodes of the current tree we obtain a sequence $Y_1, Y_2, \dots$ of random recursive trees, which we collect into a stochastic process $Y = (Y_n)_{n \in \mathbb{N}}$.

A survey of random recursive trees and their applications is given in [23]; for a more recent reference see [4, Chapter 6]. Various functionals of these structures have been considered by different authors, a representative but not exhaustive list being node degrees [25, 9, 11], height [20], path length [15, 2], profiles [8], spectra [1], and various 'topological' indices, such as the Wiener and the Zagreb indices [17, 6]. Often the results are limit theorems, with (strong) convergence of the random variables or convergence of their distributions as $n \to \infty$. This, in the authors' view, naturally raises

*Supported by the Deutsche Forschungsgemeinschaft (DFG).
†Leibniz Universität Hannover, Germany. E-mail: rgrubel@stochastik.uni-hannover.de
‡Leibniz Universität Hannover, Germany. E-mail: michail@stochastik.uni-hannover.de

the question of convergence of the trees themselves, with the aim of developing a systematic approach to the strong asymptotics of tree functionals. The Doob-Martin compactification, initiated by the fundamental paper [3], is a general tool that can be used in this context; see [26] for a recent textbook introduction. In particular, using concepts from discrete potential theory it provides an enlargement of the state space of a Markov chain such that the variables converge almost surely. This approach has been used in [5] to obtain convergence results for a class of randomly growing discrete structures that includes various random trees.

If we use the encoding explained in the opening paragraph then it is possible to retrace the full sequence $Y_1, \dots, Y_{n-1}$ of previous trees from the current tree Y_n. In such a case the discrete potential theory approach leads to concept of convergence that is of little help for proving convergence of functionals—informally, in such a situation 'the sequence is the limit'; see [5, Section 9] for details. Noting that the functionals of interest are often invariant under relabelling (a phrase that has to be made precise) we therefore choose a model that is coarser in the sense that it 'forgets the labels' but retains the Markov property. This partial loss of information turns the sequence Y into a sequence $X = (X_n)_{n\in\mathbb{N}}$ of randomly growing subsets of a fixed infinite tree. For this chain, the Doob-Martin compactification has been determined in [5]. The first of our aims here is to show that the convergence result provided by the general theory, i.e. the fact that a (transient) Markov chain converges a.s. in its own Doob-Martin topology, can be obtained more directly by using a suitable algorithmic construction, and that this approach has the advantage of leading to a description of the limit X_∞ in terms of the input sequence of the algorithm. The representation serves as the basis for the analysis of tree functionals such as different notions of path length and the Wiener index; indeed, our second objective is to obtain strong limit theorems for such functionals.

The present paper continues the research presented in [5] and [10]. In the earlier article the boundary theory approach was applied to Markov chains that are nested counting processes. These models contain binary search trees and the trees in present paper as special cases. The associated compactifications were worked out by going back to the general definition of the Doob-Martin compactification via an extension of the Martin kernel. Many Markov chains of randomly growing discrete structures arise in connection with a sequential algorithm with random input. It was noted in [10] that in the case of binary search trees the underlying algorithm can be used to obtain almost sure convergence in the Doob-Martin topology (in fact, even in stronger topologies), and that such convergence results on the level of the discrete structures themselves can be used to unify and, in some cases, amplify known results on the convergence of functionals of the structures. In the present paper we follow the strategy in [10] to obtain similar representations and limit theorems in the case of recursive trees. As will become apparent below, the implementation offers some challenges, resulting, for example, from the fact that the underlying tree is no longer locally finite.

In the next section we first take care of a variety of formal details, including some terminology and notation, and then give a new 'constructive' proof of the basic limit result. In Section 3 we discuss various tree functionals and comment on the connections to related work.

2 The limit tree and its distribution

We introduce Harris trees and the Harris chain generated by the RRT process; in view of its confounding potential we spell out the details of the transition from recursive to Harris trees. From the RRT sequence the Harris chain inherits a useful decomposition property; see Section 2.2. Next, we recall from [5] the Doob-Martin compactification

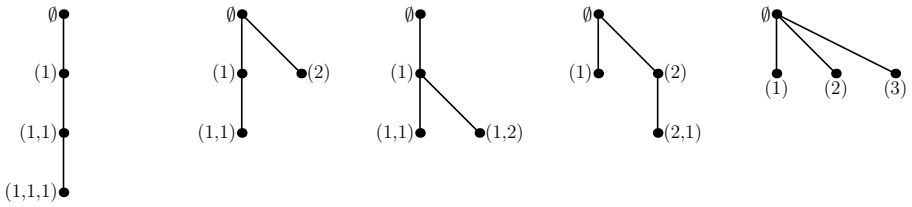

Figure 1: The Harris trees with four nodes.

of the Harris chain. As mentioned in the introduction it is one of the main points of the present paper that an algorithmic construction provides an alternative approach: We explain this algorithm and then use it to obtain a new proof of that part of [5, Theorem 6.1] that is relevant for our present purposes, together with a representation of the limit. Finally, we collect some auxiliary results on the distribution of the limit that will be useful in the next section when we analyze tree functionals.

2.1 From recursive trees to Harris trees

We regard the set $\mathbb{V} = \mathbb{N}^\star$ of finite sequences of natural numbers as the set of potential tree nodes and write $u{:}v = (u_1, \dots, u_k, v_1, \dots, v_l)$ for the concatenation of the nodes $u = (u_1, \dots, u_k)$ and $v = (v_1, \dots, v_l)$, abbreviating $u{:}(i)$ to ui, $i \in \mathbb{N}$. By a *Harris tree* we mean a finite subset x of $\mathbb{V}$ with the properties

(H1) if $u{:}v \in x$, then $u \in x$,
(H2) if $ui \in x$ with $i > 1$, then $uj \in x$ for $j = 1, \dots, i-1$.

Condition (H1) is prefix stability if we regard nodes as words with letters from the alphabet $\mathbb{N}$. In a family tree interpretation, condition (H2) means that a non-root node must either be the first child of its ancestor node or that it must have earlier-born siblings. Harris trees are also known as Ulam-Harris trees; they may be seen as rooted planar trees with a specific labelling of nodes.

We write $\mathbb{H}$ for the set of Harris trees and $\mathbb{H}_n$ for the subset of those trees that have n nodes. In order to relate Harris trees to recursive trees we map the nodes j of a recursive tree to words $u(j) = (u_1(j), \dots, u_k(j)) \in \mathbb{V}$ as follows: The length k of the word is the distance to the root of (the node labelled) j, and $u_k(j)$ is the number of nodes $i \in [j]$ that have the same direct ancestor as j. The prefix sequences similarly encode the nodes from the root to j. This corresponds to an embedding of recursive trees into the plane where new nodes are placed to the right of their siblings.

Clearly, there are $(n-1)!$ possibilities for the encoding sequences for recursive trees with n nodes, hence this is also the number of recursive trees with n nodes. Figure 1 shows the five elements of $\mathbb{H}_4$. Of the $(4-1)! = 6$ recursive trees with four nodes, encoded by $(1,2,3)$, $(1,1,2)$, $(1,2,1)$, $(1,2,2)$, $(1,1,3)$ and $(1,1,1)$ respectively, the second and third are mapped to the same Harris tree. The figure also offers an opportunity to comment on the informal expression of 'forgetting the labels' that we used above and that often appears in the literature: It is tempting to regard this as passing from graphs to isomorphism classes, but this is not what is happening here—indeed, the second and the fourth Harris tree in Figure 1 are isomorphic as rooted trees. A compatible notion of equivalence and isomorphism in the present situation can be obtained on the basis of the above planar embedding of recursive trees.

Writing Ψ for the function that maps recursive trees to Harris trees, we define the Harris chain $X = (X_n)_{n\in\mathbb{N}}$ by $X_n := \Psi(Y_n)$ for all $n \in \mathbb{N}$, where Y is the RRT chain

introduced in Section 1. Informally, instead of passing to the graph isomorphim classes of the recursive trees we still keep track of the ordering in time of the descendants of the individual nodes: The respective components in the representation of the nodes as elements of $\mathbb{N}^\star$ reflect the order of the siblings, with 1 being the firstborn and so on.

In the original process, Y_n is uniformly distributed on its range, but X_n is not uniformly distributed on $\mathbb{H}_n$ as explained above for $n = 4$. In this new process, it is no longer possible to 'trace back' to previous values. As Ψ does not change the number of nodes it is adapted to the combinatorial family $\mathbb{H}$ in the sense that $P(X_n \in \mathbb{H}_n) = 1$ for all $n \in \mathbb{N}$. To see that it retains the Markov property and to obtain the corresponding transition probabilities we argue as follows: Let y_n, y'_n be recursive trees with n nodes and let y_{n+1} be a recursive tree with $n+1$ nodes. Suppose that $\Psi(y_n) = \Psi(y'_n) =: x_n$ and let $x_{n+1} := \Psi(y_{n+1})$. If $x_n \subset x_{n+1}$ then there is a unique recursive tree z_{n+1} such that $\Psi(z_{n+1}) = x_{n+1}$ and

$$P(X_{n+1} = x_{n+1} | Y_n = y_n) \;=\; P(Y_{n+1} = z_{n+1} | Y_n = y_n) \;=\; \frac{1}{n},$$

and similarly there is a z'_{n+1} with the same property for y'_n. Clearly, if $x_n \not\subset x_{n+1}$, then these probabilities will be 0. This shows that

$$P(X_{n+1} = x_{n+1} | Y_n = y_n) \;=\; P(X_{n+1} = x_{n+1} | Y_n = y'_n)$$

whenever $\Psi(y_n) = \Psi(y'_n)$, and further that

$$P(X_{n+1} = x_{n+1} | X_n = x_n) \;=\; \begin{cases} 1/n, & x_n \subset x_{n+1}, \\ 0, & \text{otherwise}, \end{cases}$$

for $x_n \in \mathbb{H}_n$, $x_{n+1} \in \mathbb{H}_{n+1}$. By [14, Lemma 2.5] the first of these implies that X is a Markov chain; the second shows that, as with Y, we select the ancestor for the new node uniformly at random in the step from X_n to X_{n+1}.

2.2 A tree decomposition

We associate with a node $u = (u_1, \ldots, u_k) \in \mathbb{V}$ its 'flat' and 'raised' version

$$u^\flat := \; (1, u_1, \ldots, u_k), \quad u^\sharp := \begin{cases} (1 + u_1, u_2, \ldots, u_k) & \text{if } u \neq \emptyset, \\ \emptyset, & \text{if } u = \emptyset, \end{cases}$$

and lift this to trees $x \in \mathbb{H}$ via

$$x^\flat := \{u \in \mathbb{V} : \; u^\flat \in x\}, \quad x^\sharp := \; \{u \in \mathbb{V} : \; u^\sharp \in x\}.$$

These are the subtree of x rooted at (1) and the shifted tree that remains if this subtree is taken out; see Figure 2 for an illustration.

It is well known that the random variables $X_n^\flat$ and $X_n^\sharp$ are conditionally independent given $K_n := \#X_n^\flat$, that K_n is uniformly distributed on $[n-1]$, and that, conditionally on $K_n = k$, $X_n^\flat$ and $X_n^\sharp$ have the same distribution as X_k and X_{n-k} respectively. An interesting combinatorial proof of the corresponding statement for the Y process, based on a bijection between permutations and random recursive trees, is given in [2]. An alternative proof can be obtained on using the algorithmic background to be given in Section 2.4 below.

2.3 The Doob-Martin compactification of the Harris chain

The paths of the stochastic process X are sequences of growing subsets of the set $\mathbb{V}$ of all potential nodes. We may regard $\mathbb{V}$ itself as the infinite Harris tree (note that

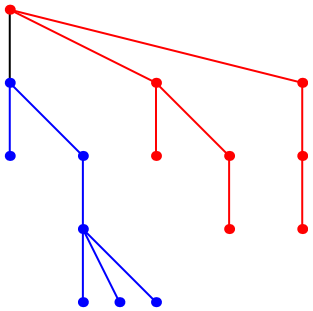

Figure 2: A Harris tree $x \in \mathbb{H}_{15}$ and its decomposition into $x^\flat \in \mathbb{H}_7$ (blue) and $x^\sharp \in \mathbb{H}_8$ (red); the black edge disappears.

this tree is not locally finite). It can be shown that, in the X-sequence, every potential node will eventually be an element of the infinite Harris tree. Hence, if we embed $\mathbb{H}$ into $\{0,1\}^{\mathbb{V}}$ via the node indicators,

$$x \;\mapsto\; \big(u \mapsto 1_x(u)\big),$$

then X_n converges almost surely to this infinite tree, which is represented by the function that is constant 1. This, however, does not capture the 'true' asymptotics of X. In contrast, Markov chain boundary theory provides a state space completion (compactification) $\bar{\mathbb{H}}$ of $\mathbb{H}$, the Doob-Martin compactification, which has in particular the following properties,

(L) $X_n \to X_\infty \in \partial\mathbb{H} := \bar{\mathbb{H}} \setminus \mathbb{H}$ with probability 1 as $n \to \infty$,

(T) X_∞ generates the tail σ-field associated with X, up to null sets.

Here $\mathbb{H}$ itself is endowed with the discrete topology (which in turn derives from the metric that assigns distance 1 to each pair of distinct points). For (T) we require that the range of X_n is disjoint from the range of X_m if $n \neq m$. For the Harris sequence this is the case, so (T) implies that the Doob-Martin compactification captures the persisting randomness of the sequence, whereas for any one-point compactification the σ-field generated by the limit will always be trivial in the sense that only 0 and 1 arise as probabilities of its elements.

The Doob-Martin compactification $\bar{\mathbb{H}}$ of $\mathbb{H}$ with respect to the Harris chain has been identified in [5]. Let

$$\bar{\mathbb{V}} \;:=\; \mathbb{N}^\star \sqcup \mathbb{N}^\infty \sqcup \bigsqcup_{k=0}^{\infty} \big(\mathbb{N}^k \times \{\infty\}^\infty\big).$$

In words: $\bar{\mathbb{V}}$ consists of all finite and infinite sequences of natural numbers, plus all infinite sequences $u = (u_i)_{i\in\mathbb{N}} \subset \mathbb{N} \sqcup \{\infty\}$ with the property that, for some $k \in \mathbb{N}$, $u_i \in \mathbb{N}$ for $i < k$ and $u_i = \infty$ for $i \geq k$. For $u \in \mathbb{V}$ and $v \in \bar{\mathbb{V}}$ write $u \leq v$ if u is a prefix of v and put

$$A_u := \{v \in \bar{\mathbb{V}} :\; u \leq v\}, \quad u \in \mathbb{V}.$$

Let $\mathcal{V}$ be the σ-field on $\bar{\mathbb{V}}$ generated by the sets A_u, $u \in \mathbb{V}$, let $\bar{\mathbb{H}}$ be the set of probability measures μ on $(\mathbb{V}, \mathcal{V})$, and endow $\bar{\mathbb{H}}$ with the coarsest topology that makes the functions $\mu \mapsto \mu(A_u)$, $u \in \mathbb{V}$, continuous. Finally, embed $\mathbb{H}$ into $\bar{\mathbb{H}}$ by identifying $x \in \mathbb{H}$ with the uniform distribution on x as a subset of $\mathbb{V}$. Then $\bar{\mathbb{H}}$ is the Doob-Martin compactification of $\mathbb{H}$ induced by the chain X, up to homeomorphism.

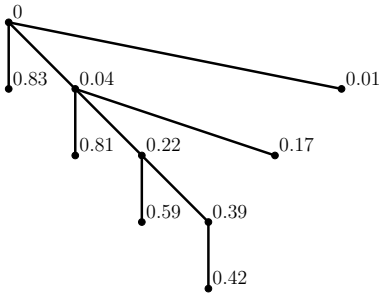

Figure 3: The tree obtained from $t = (.83, .04, .81, .22, .59, .01, .39, .42, .17, \ldots)$.

2.4 The algorithmic construction

For $x \in \mathbb{H}$ and $u \in \mathbb{V}$ let $x(u) := \{v \in \mathbb{V} \colon u\!:\!v \in x\}$ be the subtree of x rooted at u. Then the embedding of $\mathbb{H}$ into $\bar{\mathbb{H}}$ may be written as

$$x \;\mapsto\; \big(u \mapsto \#x(u)/\#x\big),$$

and we can restate the convergence in the Doob-Martin topology of a sequence $(x_n)_{n\in\mathbb{N}}$ with $x_n \in \mathbb{H}_n$ for all $n \in \mathbb{N}$ to $\mu \in \partial\mathbb{H}$ as

$$\lim_{n\to\infty} \frac{1}{n}\, \#x_n(u) = \mu(A_u) \quad \text{for all } u \in \mathbb{V}.$$

Our plan is to prove the almost sure convergence of the Harris chain X in this topology by using an algorithm that generates X if the input is chosen appropriately.

The *recursive tree algorithm* maps an input sequence $t = (t_n)_{n\in\mathbb{N}}$ of pairwise distinct positive real numbers to an output sequence $(x_n, \phi_n)_{n\in\mathbb{N}}$ of labelled trees, with $x_n \in \mathbb{H}_n$ and $\phi_n : x_n \to \mathbb{R}$. The algorithm works sequentially, starting with $x_1 = \{\emptyset\}$ and the label $\phi_1(\emptyset) = t_0 := 0$ for the root node. As explained in Section 1 and at the end of Section 2.1, we need to specify the direct ancestor of the new node v to be added in the step from x_n to x_{n+1}: We attach v as a next (resp. the first) child to the node with label $\max\{t_j : j = 0, \ldots, n-1,\, t_j < t_n\}$ and then label v by t_n. Figure 3 shows an example where new children are positioned to the right of their older siblings. By $\mathrm{RT}(t)$ we mean the sequence $(x_n)_{n\in\mathbb{N}}$, i.e. we ignore the labels.

Clearly, if the trees converge, then the limit must be a function of the input sequence. In order to be able to specify this relationship we need some more notation: Given an increasing sequence $(x_n)_{n\in\mathbb{N}}$ of Harris trees, let

$$\tau(u) \;:=\; \inf\{n \in \mathbb{N} : \, u \in x_{n+1}\}$$

(note that x_n is built from $t_1, \ldots, t_{n-1}$). Further, for any sequence $(t_n)_{n\in\mathbb{N}}$ of pairwise distinct elements of the open unit interval let

$$0 =: t_0^{(n)} < t_1^{(n)} < t_2^{(n)} < \cdots < t_n^{(n)} < t_{n+1}^{(n)} := 1$$

be the augmented increasing order statistics associated with the first n values $t_1, \ldots, t_n$, and let

$$\kappa(u) := \#\{1 \le i \le \tau(u) : \, t_i \le t_{\tau(u)}\}$$

be the rank of $t_{\tau(u)}$ in $t_1, \ldots, t_{\tau(u)}$, so that $t_{\kappa(u)}^{(\tau(u))} = t_{\tau(u)}$. In Figure 3 for example, the node $u = (2, 2)$ has $\tau(u) = 4$, $\kappa(u) = 2$ and $t_{\tau(u)} = 0.22$.

The following result relates the algorithm and the limit object. Let $\mathrm{unif}(0, 1)$ be the uniform distribution on the unit interval. We write $\mathcal{L}(Y)$ for the distribution (law) of the random quantity Y and sometimes use $Y \sim \mu$ instead of $\mathcal{L}(Y) = \mu$.

Theorem 2.1. *Let η_i, $i \in \mathbb{N}$, be independent random variables, with $\eta_i \sim \mathrm{unif}(0,1)$ for all $i \in \mathbb{N}$.*

(a) *The algorithm* RT *generates the RRT chain in the sense that* $\mathrm{RT}(\eta)$ *and X are identical in distribution.*

(b) *Suppose that $X = \mathrm{RT}(\eta)$. Then X_n converges almost surely to X_∞ in the Doob-Martin topology as $n \to \infty$, where on a set of probability 1 the limit X_∞ is given by*

$$X_\infty(A_u) = \eta_{\kappa(u)+1}^{(\tau(u))} - \eta_{\tau(u)} \quad \text{for all } u \in \mathbb{V}. \tag{2.1}$$

Proof. Part (a) belongs to the folklore of the subject. Due to its importance for the present paper we recall for the proof that the rank of η_n in $\eta_1, \dots, \eta_n$ is uniformly distributed on $\{1, \dots, n\}$, and that rank i means that η_n is attached to the node with label $\eta_{i-1}^{(n)}$ (which is the root if $i = 1$).

With each node u we associate the interval $I(u) = (\eta_{\tau(u)}, \eta_{\kappa(u)+1}^{(\tau(u))})$. From the definition of the RT algorithm, nodes added to the tree at a time $n > \tau(u)$ will have prefix u if and only if $\eta_n \in I(u)$. The random variables $\eta_{\tau(u)+n}$, $n \in \mathbb{N}$, are independent and uniformly distributed on the unit interval, hence (2.1) follows with the Glivenko-Cantelli theorem. □

Theorem 2.1 can be related to the corresponding result [10, Theorem 1] for binary search trees via the natural or rotation correspondence between Harris trees and binary trees [13, Section 2.3.2] [7, p.73]; details will be given in [16].

In addition to the convergence of the trees we also obtain the distribution of the limit X_∞, which takes its values in the set of probability measures μ on $(\bar{\mathbb{V}}, \mathcal{V})$. As a preliminary step we extend the tree decomposition introduced in Section 2.2 to $\bar{\mathbb{H}}$ as follows: For $\mu \in \partial\mathbb{H}$ with $0 < \mu(A_{(1)}) < 1$ we define $\mu^\flat, \mu^\sharp \in \partial\mathbb{H}$ by

$$\mu^\flat(A_u) = \frac{\mu(A_{u^\flat})}{\mu(A_{(1)})}, \quad \mu^\sharp(A_u) = \frac{\mu(A_{u^\sharp})}{1 - \mu(A_{(1)})}$$

for all $u \in \mathbb{V} \setminus \{\emptyset\}$, and $\mu^\flat(A_u) = \mu^\sharp(A_u) = 1$ if $u = \emptyset$.

Proposition 2.2. *Let X_∞ be as in Theorem 2.1. Then the random variables $\eta := X_\infty(A_{(1)})$, $X_\infty^\sharp$ and $X_\infty^\flat$ are independent. Further, $\eta \sim \mathrm{unif}(0,1)$, and $X_\infty^\sharp$ and $X_\infty^\flat$ have the same distribution as X_∞.*

Proof. Let $\eta = (\eta_i)_{i \in \mathbb{N}}$ be a sequence of independent, $\mathrm{unif}(0,1)$-distributed random variables. We define two new sequences $\eta^\flat = (\eta_i^\flat)_{i \in \mathbb{N}}$ and $\eta^\sharp = (\eta_i^\sharp)_{i \in \mathbb{N}}$ by successively transforming the η_i's with $\eta_i > \eta_1$ into $\eta_j^\flat = (\eta_i - \eta_1)/(1 - \eta_1)$ and the η_i's with $\eta_i < \eta_1$ into $\eta_j^\sharp = \eta_i/\eta_1$. Clearly, η_1, $\eta^\flat$ and $\eta^\sharp$ are independent, and $\eta^\flat$ and $\eta^\sharp$ are again sequences of independent, $\mathrm{unif}(0,1)$-distributed random variables. From this, the statement of the theorem follows in view of $X_\infty(A_{(1)}) = 1 - \eta_1$, $X^\flat = \mathrm{RT}(\eta^\flat)$, and $X^\sharp = \mathrm{RT}(\eta^\sharp)$. □

We call μ atom-free and diffuse if

$$\mu(\{u\}) = 0 \quad \text{and} \quad \mu(A_u) > 0 \quad \text{for all } u \in \mathbb{V}.$$

Let $\Sigma_\infty \subset [0,1]^\infty$ be the infinite-dimensional probability simplex, that is, the set of all sequences $(\rho_i)_{i \in \mathbb{N}}$ with $\rho_i \geq 0$ for all $i \in \mathbb{N}$ and $\sum_{i=1}^\infty \rho_i = 1$. An atom-free and diffuse μ associates with each $u \in \mathbb{V}$ an element $\rho(\mu, u) = (\rho_i(\mu, u))_{i \in \mathbb{N}}$ of Σ_∞ via

$$\rho_i(\mu, u) := \frac{\mu(A_{ui})}{\mu(A_u)} \quad \text{for all } i \in \mathbb{N}.$$

For later use we note that, for such μ,

$$\rho(\mu, u^\sharp) = \rho(\mu^\sharp, u), \quad \rho(\mu, u^\flat) = \rho(\mu^\flat, u) \quad \text{for all } u \in \mathbb{V}.$$

Clearly, μ can be reconstructed from $\rho(\mu, u)$, $u \in \mathbb{V}$. In fact,

$$\mu(A_u) = \prod_{i=1}^{k} \rho_{u_i}\big(\mu, (u_1, \dots, u_{i-1})\big) \quad \text{for all } u = (u_1, \dots, u_k) \in \mathbb{V}. \tag{2.2}$$

Of course, for a random input both the τ- and the κ-values will be random, as will be μ.

We say that a random variable $\xi = (\xi_i)_{i\in\mathbb{N}}$ with values in Σ_∞ has the (standard) GEM (Griffiths-Engen-McCloskey) distribution if its components can be written as

$$\xi_1 = \zeta_1, \quad \xi_i = \zeta_i \prod_{j=1}^{i-1}(1-\zeta_j) \ \text{ for } i > 1, \tag{2.3}$$

with ζ_i, $i \in \mathbb{N}$, independent and $\zeta_i \sim \mathrm{unif}(0,1)$ for all $i \in \mathbb{N}$.

At each level k, the sets A_u with $|u| = k$ provide a partition of $\bar{\mathbb{V}} \setminus \mathbb{N}^{k-1}$. The corresponding values $X_\infty(A_u)$ are related to the kth nested decomposition of the unit interval into descending records of the input sequence. This interpretation suggests the following result, which gives a description of the distribution of X_∞.

Theorem 2.3. *Let X_∞ be as in Theorem 2.1. Then the random variables $\rho(X_\infty, u)$, $u \in \mathbb{V}$, are independent and GEM distributed.*

Proof. By Proposition 2.2, $\rho_1(X_\infty, \emptyset) = 1 - \eta_1 = X_\infty(A_{(1)})$, $X_\infty^\flat$ and $X_\infty^\sharp$ are independent. Repeating the decomposition with the respective raised part, we obtain that the variables

$$\frac{\rho_i(X_\infty, \emptyset)}{\rho_{i-1}(X_\infty, \emptyset)}, \quad i \in \mathbb{N},$$

with $\rho_o(X_\infty, \emptyset) := 1$, are independent and $\mathrm{unif}(0,1)$-distributed. Moreover, they are independent of the random probability measures $X_{\infty,i}$, $i \in \mathbb{N}$, defined by

$$X_{\infty,i}(A_u) := \frac{X_\infty(A_{(i):u})}{X_\infty(A_{(i)})}, \quad u \in \mathbb{V}.$$

Finally, these measures are independent and identical in distribution to X_∞. (It is easy to see that $X_{\infty,i}$ arises as the $\flat$-part of the ith iteration of the decomposition). In particular, $\rho(X_\infty, \emptyset) \sim$ GEM. Taken together, this proves the case $k = 0$ of the following statement:

(i) $\rho(X_\infty, u) \sim$ GEM for all $u \in \mathbb{V}$ with $|u| \le k$,
(ii) the random sequences $\rho(X_\infty, u)$, $u \in \mathbb{V}$, $|u| \le k$, are independent,
(iii) the random measures $X_{\infty,v}$, $v \in \mathbb{V}$, $|v| = k+1$, given by

$$X_{\infty,v}(A_u) := \frac{X_\infty(A_{v:u})}{X_\infty(A_v)}, \ u \in \mathbb{V},$$

are independent and identical in distribution to X_∞,
(iv) $\{\rho(X_\infty, u) : u \in \mathbb{V}, |u| \le k\}$ and $\{X_{\infty,v} : v \in \mathbb{V}, |v| = k+1\}$ are independent.

We can apply the same reasoning used for $k = 0$ separately to each of the nodes at level $k+1$ to obtain the induction step from k to $k+1$.

This shows that the above compound statement holds for all $k \in \mathbb{N}$; clearly, (i) and (ii) imply the assertion of the theorem. □

In view of the fact that $X_\infty(A_u)$ is a function of the variables $\rho(X_\infty, v)$ with $|v| < |u|$ we obtain that $X_\infty(A_u)$ and $\rho(X_\infty, u)$ are independent, for all $u \in \mathbb{V}$.

2.5 Conditional distributions

In order to be able to use the general limit theorem for the analysis of tree functionals in the next section we need the conditional distribution of X_∞ given X_n. For this we rely on the results in [5, Section 6]; we also need some more notation.

The distribution $\mathrm{Beta}(\alpha,\beta)$ with parameters $\alpha,\beta>0$ is given by its density

$$f(t|\alpha,\beta) \;=\; \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, t^{\alpha-1}(1-t)^{\beta-1}, \quad 0<t<1. \tag{2.4}$$

For later use we recall that

$$E\xi=\frac{\alpha}{\alpha+\beta}, \quad E\xi^2=\frac{\alpha(\alpha+1)}{(\alpha+\beta)(\alpha+\beta+1)} \quad \text{if } \xi\sim\mathrm{Beta}(\alpha,\beta), \tag{2.5}$$

and, clearly, $\mathrm{Beta}(1,1)=\mathrm{unif}(0,1)$. For $a=(a_1,\dots,a_k)\in\mathbb{N}^\star$ we write $\mathrm{GEM}(a)$ for the distribution of the Σ_∞-valued random sequence $\xi=(\xi_i)_{i\in\mathbb{N}}$ given by

$$\xi_1=\zeta_1, \quad \xi_i=\zeta_i\prod_{j=1}^{i-1}(1-\zeta_j) \quad \text{for } i>1, \tag{2.6}$$

where ζ_i, $i\in\mathbb{N}$, are independent and

$$\mathcal{L}(\zeta_i)=\begin{cases}\mathrm{Beta}\bigl(a_i,1+\sum_{j=i+1}^k a_j\bigr), & \text{for } i<k,\\ \mathrm{Beta}(a_k,1), & \text{for } i=k,\\ \mathrm{Beta}(1,1), & \text{for } i>k.\end{cases} \tag{2.7}$$

Interestingly, the marginals of such random sequences are again beta distributed (of course, they are no longer independent).

Lemma 2.4. *If $\xi=(\xi_i)_{i\in\mathbb{N}}\sim\mathrm{GEM}(a)$ with $a=(a_1,\dots,a_k)\in\mathbb{N}^\star$, then, with $b:=\sum_{i=1}^k a_i$,*

$$\xi_i\sim\mathrm{Beta}(a_i,1+b-a_i) \;\; \textit{for} \;\; i=1,\dots,k.$$

Moreover, with $(\zeta_i)_{i\in\mathbb{N}}$ as in (2.6) *and* (2.7)

$$1-\sum_{j=1}^k\xi_j \;=\; \prod_{j=1}^k(1-\zeta_j)\sim\mathrm{Beta}(1,b).$$

Proof. This follows with the known rule for products of independent beta-distributed random variables, see e.g. [12, p.378, Exercise 11.8]. □

Recall that the distribution of X_∞ is specified by the (joint) distribution of the Σ_∞-valued quantities $\rho(X_\infty,u)$, $u\in\mathbb{V}$, and that $\#x(ui)>0$ implies $\#x(uj)>0$ for $j=1,\dots,i-1$ by property (H2) of Harris trees; see also (2.2).

Theorem 2.5. *The conditional distribution of $\rho(X_\infty,u)$ given X_n is $\mathrm{GEM}(a)$, where $a=(a_1,\dots,a_k)$ with*

$$k=\max\{i\in\mathbb{N}:\,\#X_n(ui)>0\}, \quad a_i=\#X_n(ui) \;\; \textit{for } i=1,\dots,k. \tag{2.8}$$

Further, the random sequences $\rho(X_\infty,u)$, $u\in\mathbb{V}$, are conditionally independent given X_n.

Proof. By the general theory of Markov chain boundaries the conditional distribution Q_2 of X_∞ given $X_n=x\in\mathbb{H}_n$ has density $K(x,\cdot)$ with respect to the (unconditional) distribution Q_1 of X_∞, where K denotes the extended Martin kernel. Note that Q_1 and

Q_2 are probability measures on the set $\bar{\mathbb{H}}$ of probability measures μ on $(\bar{\mathbb{V}}, \mathcal{V})$, where the σ-field on $\bar{\mathbb{H}}$ is the one generated by the evaluation maps $\mu \mapsto \mu(A_u)$, $u \in \mathbb{V}$. The extended Martin kernel has been determined in [5]: It can be written as the product of 'local extended kernels',

$$K(x, \mu) = \prod_{u \in \mathbb{V}} K_u\big(u(x), \rho(\mu, u)\big), \tag{2.9}$$

where $u(x) = (a_1, \dots, a_k) \in \mathbb{N}^\star$ is given by (2.8) with x instead of X_n, and

$$K_u(x, s) \;=\; \frac{\big(\sum_{i=1}^k a_i\big)!}{\prod_{i=1}^k (a_i - 1)!} \prod_{i=1}^k s_i^{a_i - 1} \prod_{i=1}^{k-1} \Big(1 - \sum_{j=1}^i s_i\Big) \tag{2.10}$$

for all $s = (s_i)_{i \in \mathbb{N}} \in \Sigma_\infty$. Also, $K_u(x, \cdot)$ is the conditional density of the distribution $Q_{2,u}$ of $\rho(X_\infty, u)$ given $X_n = x$ with respect to its corresponding unconditional counterpart $Q_{1,u}$, which we know to be the GEM distribution. The product form (2.9) implies that the independence of the sequences $\rho(X_\infty, u)$, $u \in \mathbb{V}$, remains intact in the transition from Q_1 to Q_2. This proves the second part of the theorem.

Now let $T : \Sigma_\infty \to [0,1]^\infty$ be given by

$$(s_i)_{i \in \mathbb{N}} \mapsto (t_i)_{i \in \mathbb{N}}, \quad t_i := \frac{s_i}{1 - s_1 - \dots - s_{i-1}} \text{ for all } i \in \mathbb{N}.$$

This is the inverse of the transition from ζ to ξ in (2.6). We know that the push-forward $Q_{1,u}^T$ of $Q_{1,u}$ under T is the infinite product of uniforms. The first part of the theorem refers to the push-forward $Q_{2,u}^T$ of $Q_{2,u}$ under T; it asserts that a density of $Q_{2,u}^T$ with respect to $Q_{1,u}^T$ is given by

$$g(t) \;=\; \prod_{i=1}^{k-1} f\Big(t_i \Big| a_i, \sum_{j=i+1}^k a_j\Big) \cdot f(t_k | a_k, 1)$$

for almost all $t = (t_i)_{i \in \mathbb{N}} \in [0,1]^\infty$, with f as in (2.4). With all this notation in place it remains to check that $g \circ T = K_u(x, \cdot)$, with K_u as in (2.10). This, however, is a bookkeeping task. □

The embedding of $\mathbb{H}$ into $\bar{\mathbb{H}}$, which maps X_n to the uniform distribution on its nodes, leads to an interpretation of X_n as a real-valued random function on $\mathbb{V}$ via $u \mapsto \#X_n(u)/n$. Similarly, the limit X_∞ can be seen as the random function $u \mapsto X_\infty(A_u)$ on $\mathbb{V}$. Obviously, all these functions are bounded and, if we endow $\mathbb{V}$ with the discrete topology, they are continuous. This displays X_n, $n \in \mathbb{N}$, and X_∞ as random elements of an infinite-dimensional separable Banach space.

Corollary 2.6. *Let $(\mathcal{F}_n)_{n \in \mathbb{N}}$, with $\mathcal{F}_n := \sigma(X_1, \dots, X_n)$, $n \in \mathbb{N}$, be the natural filtration of the Harris chain $(X_n)_{n \in \mathbb{N}}$. Then*

$$X_n = E[X_\infty | \mathcal{F}_n] \quad \textit{for all } n \in \mathbb{N}.$$

In particular, $(X_n, \mathcal{F}_n)_{n \in \mathbb{N}}$ is a martingale.

Proof. We have $E[X_\infty | \mathcal{F}_n] = E[X_\infty | X_n]$ due to the Markov property. Further, the notion of infinite-dimensional martingale, see e.g. [18, Section-V.2], in the present context means that we have to check that

$$E[X_\infty(A_u) | X_n] = \frac{1}{n} \#X_n(u) \quad \text{for all } u \in \mathbb{V}.$$

Let $u = (u_1, \dots, u_k) \in \mathbb{V}$ be given and let

$$\xi_i := \rho_{u_i}\big(X_\infty, (u_1, \dots, u_{i-1})\big), \quad i = 1, \dots, k.$$

From (2.2) we obtain $X_\infty(A_u) = \prod_{i=1}^k \xi_i$, and by Theorem 2.5 the factors are conditionally independent given X_n. Hence, using Lemma 2.4 and (2.5),

$$\begin{aligned} E[X_\infty(A_u)|X_n] &= \prod_{i=1}^k E[\xi_i|X_n] \\ &= \prod_{i=1}^k \frac{\#X_n\big((u_1, \dots, u_i)\big)}{1 + \sum_{j=1}^\infty \#X_n\big((u_1, \dots, u_{i-1}, j)\big)} \\ &= \prod_{i=1}^k \frac{\#X_n\big((u_1, \dots, u_i)\big)}{\#X_n\big((u_1, \dots, u_{i-1})\big)} \quad = \frac{1}{n}\#X_n(u). \end{aligned}$$ □

This result may be seen as a consequence of the general Doob-Martin construction.

3 Tree functionals

Let $Y = (Y_n)_{n\in\mathbb{N}}$ be the RRT chain and let $X = (X_n)_{n\in\mathbb{N}}$, with $X_n = \Psi(Y_n)$ for all $n \in \mathbb{N}$, be the associated Harris chain. In this section we consider functionals of the recursive trees that are invariant under Ψ and hence can be written as functions $V_n = \Phi(X_n)$ of the X-variables. A typical example is the total path length, which is the sum of the depth of all nodes in the tree. The methods discussed below can be applied to fairly general functions Φ, but here we will restrict ourselves to the real-valued case.

There are two main probabilistic methods to obtain distributional or even strong limit results for suitably standardized versions of the V-variables. In the first of these, we try to find a suitable martingale and then apply a martingale limit theorem. In the second, we use the internal structure of the X-variables to find a recursion for the V-variables and then apply Banach's fixed point theorem with a suitably chosen metric space of probability measures. The prototypical example is the number of comparisons needed by the Quicksort algorithm, which can be related to the total path length of binary search trees: The martingale approach is carried out in [21], whereas [22] employed the second approach, which since then has come to be known as the contraction method. The two methods may fruitfully be combined, as exemplified by [2] in connection with the total path length of random recursive trees.

On its own the martingale method does not say anything about the limit, and the contraction method may miss the fact that the random variables converge almost surely or in L^p and hence in a stronger mode than convergence in distribution. The method suggested in the present paper and in [10] needs some additional investment in connection with proving the convergence of the discrete structures themselves but then provides a unifying approach: In view of the fact that X_∞ generates the tail σ-field associated with the Harris chain, see property (T) in Section 2.3, any almost sure limit Y_∞ must be a functional $Y_\infty = \Psi(X_\infty)$ of X_∞, up to null sets. Projecting Y_∞ on the natural filtration we obtain a convergent martingale, which often turns out to be a simple transformation of the variables V_n. Below we carry this out for two versions of the total path length and for the Wiener index. In the final subsection we consider a new functional that combines the two versions of the pathlength into a quantity that can serve as a measure of complexity for the algorithm presented in Section 2.4.

3.1 Total path length

This is simply the sum of all node depths and can be written in terms of subtree sizes as

$$\mathrm{TPL}(x) \;:=\; \sum_{u\in x}|u| \;=\; \sum_{u\in x}\#x(u)-\#x, \quad x\in\mathbb{H}.$$

Here we have written $|u|$ for the length (or depth) k of $u=(u_1,\dots,u_k)\in\mathbb{V}$. We need the auxiliary function

$$C:\Sigma_\infty\to[-\infty,1], \quad (s_i)_{i\in\mathbb{N}}\mapsto 1+\sum_{i=1}^{\infty}s_i\log s_i.$$

The harmonic numbers

$$H_0:=1, \quad H_n:=\sum_{k=1}^{n}\frac{1}{k} \;\text{ for all } n\in\mathbb{N},$$

will appear repeatedly; we will write $H(n)$ instead of H_n whenever this is typographically more convenient. We collect some auxiliary statements.

Lemma 3.1. (a) *If $\xi\sim\mathrm{GEM}(a)$ for some $a\in\mathbb{N}^\star$ then $\|C(\xi)\|_p<\infty$ for all $p\ge 1$.*

(b) *If $\xi\sim\mathrm{GEM}(a)$ with $a=(a_1,\dots,a_k)\in\mathbb{N}^\star$, then*

$$EC(\xi) \;=\; 1+\frac{\sum_{i=1}^k a_iH(a_i)}{1+\sum_{i=1}^k a_i} \;-\; H\Big(1+\sum_{i=1}^k a_i\Big). \tag{3.1}$$

In particular, $EC(\xi)=0$ if $\xi\sim\mathrm{GEM}$.

Proof. For the proof of the first part we assume that $a=\emptyset$ and use the representation of ξ by a sequence $(\zeta_i)_{i\in\mathbb{N}}$ of independent random variables with distribution $\mathrm{unif}(0,1)$, see (2.6). Then, for each $i\in\mathbb{N}$,

$$\begin{aligned}
\|\xi_i\log\xi_i\|_p \;&=\; \Big\|\Big(\zeta_i\prod_{j=1}^{i-1}(1-\zeta_j)\Big)\Big(\log\zeta_i+\sum_{k=1}^{i-1}\log(1-\zeta_k)\Big)\Big\|_p\\
&\le\; \Big\|\zeta_i(\log\zeta_i)\prod_{j=1}^{i-1}(1-\zeta_j)\Big\|_p \;+\; \sum_{k=1}^{i-1}\Big\|\zeta_i\log(1-\zeta_k)\prod_{j=1}^{i-1}(1-\zeta_j)\Big\|_p\\
&=\; \|\zeta_i\log\zeta_i\|_p\prod_{j=1}^{i-1}\|1-\zeta_j\|_p\\
&\qquad\qquad +\sum_{k=1}^{i-1}\|\zeta_i\|_p\|(1-\zeta_k)\log(1-\zeta_k)\|_p\prod_{j\in[i-1]\setminus\{k\}}\|1-\zeta_j\|_p\\
&=\; \|\zeta_1\log\zeta_1\|_p\|\zeta_1\|_p^{i-1} \;+\; (i-1)\|\zeta_1\|_p\|\zeta_1\log\zeta_1\|_p\|\zeta_1\|_p^{i-2},
\end{aligned}$$

where we have used independence and $\mathcal{L}(\zeta_i)=\mathcal{L}(1-\zeta_i)=\mathcal{L}(\zeta_1)$. In view of

$$\int_0^1|t^p(\log t)^p|\,dt<\infty, \qquad \|\zeta_1\|_p<1,$$

this shows that $\|\xi_i\log\xi_i\|_p$ decreases at an exponential rate as $i\to\infty$. The generalization to an arbitrary $a\in\mathbb{N}^\star$ is straightforward.

For the proof of (b) we first note that, for $\zeta\sim\mathrm{Beta}(i,j)$ with $i,j\in\mathbb{N}$,

$$E\big(\zeta\log(\zeta)\big) \;=\; \frac{i}{i+j}\,\big(H_i-H_{i+j}\big). \tag{3.2}$$

Suppose now that $\xi \sim \mathrm{GEM}(a)$ with $a = (a_1, \dots, a_k) \in \mathbb{N}^\star$ and let $b := \sum_{j=1}^k a_j$. We have $\xi_i \sim \mathrm{Beta}(a_i, 1 + b - a_i)$ for $j = 1, \dots, k$ by Lemma 2.4, hence

$$E\xi_i \log \xi_i = \frac{a_i}{1+b}\big(H(a_i) - H(1+b)\big), \quad i = 1, \dots, k.$$

Using the second part of Lemma 2.4 we see that for $i > k$ we may write $\xi_i = \alpha_i\beta_i$ with α_i and β_i independent, $\alpha_i \sim \mathrm{Beta}(1, b)$ and β_i the product of $i - k$ independent $\mathrm{unif}(0, 1)$-distributed random variables. This gives, using (3.2) again,

$$\begin{aligned} E\xi_i \log \xi_i \;&=\; E\alpha_i \, E\beta_i \log \beta_i \;+\; E\beta_i \, E\alpha_i \log \alpha_i \\ &=\; \frac{1}{1+b}\,\frac{i-k}{2^{i-k-1}}\,\frac{(-1)}{4} \;+\; \frac{1}{1+b}\big(H(1) - H(1+b)\big)\,\frac{1}{2^{i-k}}, \end{aligned}$$

so that, after some elementary manipulations,

$$\sum_{i=k+1}^{\infty} E\xi_i \log \xi_i \;=\; -\,\frac{H(1+b)}{1+b}.$$

Putting pieces together we finally arrive at (3.1). □

Let

$$|u|_1 := \sum_{i=1}^k u_i \quad \text{for all } u = (u_1, \dots, u_k) \in \mathbb{V}. \tag{3.3}$$

Concatenating ('padding') a finite sequence of non-negative integers with an infinite sequence of 0's gives a natural embedding of $\mathbb{N}^\star$ into the space ℓ^1 of summable sequences of real numbers; in this extension $|u|_1$ is simply the associated ℓ^1-norm. In a family tree interpretation $|u|_1$ is the sum of all birth orders in the line from u to the root. Informally, this notion enables us to cope with the fact that the infinite tree underlying the Harris chain is not locally finite, in contrast to the situation with binary search trees.

Lemma 3.2. *For $u \in \mathbb{V}$ with $|u|_1 = k$,*

$$X_\infty(A_u) \;=_{\mathrm{d}}\; \prod_{i=1}^k \zeta_i,$$

with $\zeta_1, \dots, \zeta_k$ independent, $\zeta_i \sim \mathrm{unif}(0, 1)$ for $i = 1, \dots, k$.

Proof. With each $u = (u_1, \dots, u_l) \in \mathbb{V}$ we associate its direct predecessor respectively direct elder sibling $\bar{u}$ by

$$\bar{u} := \begin{cases} (u_1, \dots, u_{l-1}), & \text{if } u_l = 1, \\ (u_1, \dots, u_{l-1}, u_l - 1), & \text{if } u_l > 1. \end{cases}$$

We may then connect the root $\emptyset =: u[0]$ to $u[k] := u$ with nodes $u[i]$, $i = 1, \dots, k - l$ in such a way that $u[i-1] = \bar{u}[i]$ for $i = 1, \dots, k$. The ratios $X_\infty(A_{u[i]})/X_\infty(A_{u[i-1]})$ are independent and $\mathrm{unif}(0, 1)$-distributed, by (2.3) for a step to the right and by Theorem 2.3 for a down-step. □

The transition $u \mapsto \bar{u}$ in the proof corresponds to the transition to the direct ancestor (next node on the path to the root) in the infinite binary tree $\{0, 1\}^\star$ associated with $\mathbb{V}$ by the natural correspondence mentioned after the proof of Theorem 2.1.

Lemma 3.3. *The sequence* $(Y_{\infty,k})_{k\in\mathbb{N}}$ *with*

$$Y_{\infty,k} := \sum_{u\in\mathbb{V},|u|_1\le k} X_\infty(A_u)\, C\big(\rho(X_\infty,u)\big) \quad \textit{for all } k\in\mathbb{N},$$

converges in L^p *for all* $p\ge 1$.

Proof. Let $p>1$. We introduce the local abbreviations

$$\mathbb{V}[k] := \{u\in\mathbb{V}:\ |u|_1=k\},\quad \mathcal{H}_k := \sigma(\{X_\infty(A_u):\ u\in\mathbb{V}[k]\}).$$

Then

$$Y_{\infty,k}-Y_{\infty,k-1} \;=\; \sum_{u\in\mathbb{V}[k]} X_\infty(A_u)\, C(\rho(X_\infty,u)).$$

Lemma 3.2 yields

$$E\big(X_\infty(A_u)\big)^p \;=\; \frac{1}{(1+p)^k} \tag{3.4}$$

for all $u\in\mathbb{V}[k]$. On $X_\infty(A_u)=\alpha(u)$, $u\in\mathbb{V}[k]$, we have

$$\mathcal{L}\big(Y_{\infty,k}-Y_{\infty,k-1}\big|\mathcal{H}_k\big) \;=\; \mathcal{L}\Big(\sum_{u\in\mathbb{V}[k]}\alpha(u)\,\zeta_u\Big),$$

with ζ_u, $u\in\mathbb{V}[k]$, independent and identically distributed; further, $E|\zeta_u|^p<\infty$ by part (a) of Lemma 3.1. Rosenthal's inequality, see e.g. [19, p.59], gives

$$E\Big|\sum_{u\in\mathbb{V}[k]}\alpha(u)\,\zeta_u\Big|^p \;\le\; c_p\left(\sum_{u\in\mathbb{V}[k]}E\big|\alpha(u)\,\zeta_u\big|^p \;+\; \Big(\sum_{u\in\mathbb{V}[k]}\mathrm{var}\big(\alpha(u)\,\zeta_u\big)\Big)^{p/2}\right)$$

with some constant that depends on p only. Unconditioning and (3.4) lead to upper bounds for both sums that decrease at an exponential rate κ^k for some $\kappa<1/2$. This offsets the cardinality 2^k of $\mathbb{V}[k]$, and we conclude that $(Y_{\infty,k})_{k\in\mathbb{N}}$ is a Cauchy sequence in L^p. □

Let

$$Y_\infty := \sum_{u\in\mathbb{V}} X_\infty(A_u)\, C\big(\rho(X_\infty,u)\big) \tag{3.5}$$

be the limit in Lemma 3.3.

Theorem 3.4. *As* $n\to\infty$,

$$\frac{1}{n}\,\mathrm{TPL}(X_n) - H_n + 1 \;\to\; Y_\infty,$$

almost surely and in L^p *for every* $p>0$.

Proof. We project the prospective limit on the natural filtration introduced in Corollary 2.6. With $Y_{\infty,k}$ as in Lemma 3.3 and using the fact that $X\mapsto E[X|\mathcal{F}]$ is a contraction on L^p we get

$$E[Y_\infty|\mathcal{F}_n] \;=\; \lim_{k\to\infty}E[Y_{\infty,k}|\mathcal{F}_n] \;=\; \lim_{k\to\infty}\sum_{v\in\mathbb{V},|v|\le k}E[X_\infty(A_u)\,C(\rho(X_\infty,u))|\mathcal{F}_n],$$

again in L^p. For the conditional expectation of the product we use the conditional independence of the factors with respect to $\mathcal{F}_n$. For k greater than the height of X_n this

conditional expectation then evaluates to 0, so that we arrive at

$$\begin{aligned} &E[Y_\infty|\mathcal{F}_n] \\ &\quad= \sum_{u\in X_n} \frac{\#X_n(u)}{n}\Big(1+\frac{\sum_{i=1}^\infty \#X_n(ui)H(\#X_n(ui))}{\#X_n(u)} - H(\#X_n(u))\Big) \\ &\quad= \frac{1}{n}\big(\mathrm{TPL}(X_n)+n\big) \;-\; \frac{1}{n}\sum_{u\in X_n}\Big(\#X_n(u)H(\#X_n(u)) - \sum_{i=1}^\infty \#X_n(ui)H(\#X_n(ui))\Big) \\ &\quad= \frac{1}{n}\,\mathrm{TPL}(X_n) + 1 - H(n), \end{aligned}$$

where a telescope effect simplified the sums. The statement of the theorem now follows with the well-known martingale convergence theorems; see e.g. [18, Theorem IV-1-2, Proposition IV-2-7]. □

It is easy to see that $EY_\infty = 0$, hence it follows from the calculations in the proof that the mean of the total path length is given by $E\mathrm{TPL}(X_n) = nH_n - n$ for all $n \in \mathbb{N}$.

The formula for the mean and the almost sure and L^p-convergence, $p > 0$, of the standardized total path length of random recursive trees have already been obtained in [15] and [2] respectively; we augment this by the representation of the limit variable in terms of Doob-Martin limit X_∞. The technical difficulty in the proof of almost sure and L^2-convergence in [15], as in its analogue for search trees in [21], consists of showing that the respective martingales (which have to be found first) are bounded in L^2. Here we obtain the martingale as a projection of a variable with finite second (or pth) moment onto the natural filtration of the Harris chain, which implies the desired boundedness by Jensen's inequality for conditional expectations.

3.2 Horizontal total path length

We may regard the depth $|u|$ of a node u as its vertical position; it is the number of downward moves (if this is the direction of tree growth, from ancestor to child in familial terms) on the way from the root to u. The (vertical) total path length of a tree, considered in Section 3.1, is the sum of these positions, taken over all nodes in the tree. By the horizontal position of u we mean the number of moves to the right (if this is where new nodes are added to an existing family) on the way from the root to u. In the Harris encoding of nodes the horizontal position of the node $u = (u_1, \dots, u_k)$ is given by $|u|_1 - |u|$, and the horizontal total path length of a tree is the sum of these positions over all nodes of the tree,

$$\mathrm{HPL}(x) := \sum_{u\in x}\big(|u|_1 - |u|\big), \quad x \in \mathbb{H}.$$

The horizontal position of a node can be seen as a recursive tree analogue of the notion of vertical position in a binary tree; see [4, Chapter 5] for the latter. The total horizontal path length does not seem to have been considered before, but a close relative is the total path degree length investigated in [24].

We proceed as in the previous section, now using the auxiliary function

$$D : \Sigma_\infty \to [-2, \infty], \quad (s_i)_{i\in\mathbb{N}} \mapsto -2 + \sum_{i=1}^\infty i\, s_i.$$

For $\xi = (\xi_i)_{i\in\mathbb{N}} \sim \mathrm{GEM}$ the representation (2.3) implies

$$E\xi_i^p \;=\; E\zeta_i^p \prod_{j=1}^{i-1} E(1-\zeta_j)^p \;=\; \frac{1}{(1+p)^i},$$

hence

$$\|D(\xi)\|_p \;\le\; 2+\sum_{i=1}^{\infty} i\|\xi_i\|_p \;<\; \infty \quad \text{for all } p>1.$$

Using similar arguments as in the proof of Lemma 3.3 we obtain that the series

$$Z_\infty := \sum_{u\in\mathbb{V}} X_\infty(A_u)\, D\big(\rho(X_\infty,u)\big) \tag{3.6}$$

converges in L^p for all $p>1$. Further, for $\xi\sim \mathrm{GEM}(a)$ with $a=(a_1,\dots,a_k)\in\mathbb{N}^\star$ and $b:=a_1+\cdots+a_k$, Lemma 2.4 leads to

$$ED(\xi) \;=\; -2 \,+\, \frac{1}{1+b}\Big(\sum_{i=1}^{\infty} i a_i \,+\, k \,+\, 2\Big) \tag{3.7}$$

if $k>0$, and $ED(\xi)=0$ for $\xi\sim\mathrm{GEM}$.

Theorem 3.5. *As $n\to\infty$,*

$$\frac{1}{n}\,\mathrm{HPL}(X_n) - H_n + 2 \;\to\; Y_\infty + Z_\infty,$$

almost surely and in L^p for every $p>0$.

Proof. We project Z_∞ on the natural filtration. Using (3.7) and similar arguments as at the beginning of the proof of Theorem 3.4 we get

$$\begin{aligned}
&E[Z_\infty|\mathcal{F}_n]\\
&\quad= \sum_{u\in\mathbb{V}} E[X_\infty(A_u)|\mathcal{F}_n]\, E[D(\rho(X_\infty,u))|\mathcal{F}_n]\\
&\quad= \sum_{u\in X_n} \frac{\#X_n(u)}{n}\bigg(-2+\frac{1}{\#X_n(u)}\Big(\sum_{i=1}^{\infty} i\,\#X_n(ui) \,+\, \#\{i\in\mathbb{N}:\, ui\in X_n\}+2\Big)\bigg)\\
&\quad= -\frac{2}{n}\sum_{u\in X_n}\#X_n(u) \,+\, \frac{1}{n}\sum_{u\in X_n}\sum_{i=1}^{\infty} i\,\#X_n(ui) \,+\, \frac{1}{n}\sum_{u\in X_n}\#\{i\in\mathbb{N}:\, ui\in X_n\} \,+\, 2\\
&\quad= -\frac{2}{n}\,\mathrm{TPL}(X_n) \,+\, \frac{1}{n}\sum_{u\in X_n}|u|_1 \,+\, \frac{n-1}{n}\\
&\quad= -\frac{1}{n}\,\mathrm{TPL}(X_n) \,+\, \frac{1}{n}\,\mathrm{HPL}(X_n) \,+\, \frac{n-1}{n}.
\end{aligned}$$

Now we proceed as in the proof of Theorem 3.4. □

As in the vertical case, see the remark after the proof of Theorem 3.4, we may use the calculations in the proof to obtain an explicit formula for the mean horizontal path length,

$$E\mathrm{HPL}(X_n) \;=\; -(n-1)+n\,EZ_\infty + E\mathrm{TPL}(X_n) \;=\; nH_n-2n+1 \;\text{ for all } n\in\mathbb{N}. \tag{3.8}$$

3.3 The Wiener index

The chemist H. Wiener introduced

$$\mathrm{WI}(G) := \frac{1}{2}\sum_{(u,v)\in V\times V} d_\circ(u,v) \tag{3.9}$$

as a measure of spread of an arbitrary finite connected graph G with node set V. Here $d_\circ$ denotes the canonical graph distance, i.e. $d_\circ(u,v)$ is the minimum length of a path

connecting u and v in G. Let $u \wedge v$ be the longest common prefix of $u, v \in \mathbb{V}$. For trees we then have

$$d_\circ(u,v) = |u| + |v| - 2|u \wedge v|$$

and, as in the case of binary trees [10, eq.(34) corrected],

$$\sum_{(u,v)\in x\times x} |u \wedge v| \;=\; \sum_{u\in x} \#x(u)^2 \;-\; \#x\,^2,$$

so that we may rewrite the Wiener index for $x \in \mathbb{H}_n$ in terms of total path length and subtree sizes as

$$\mathrm{WI}(x) \;=\; n\,\mathrm{TPL}(x) + n^2 - \sum_{u\in x} \#x(u)^2.$$

Again, we will show that a suitably standardized version converges almost surely if we insert for x the random variables X_n of the Harris chain. In addition to Y_∞ as in (3.5) we need

$$W_\infty := \sum_{u\in\mathbb{V}} X_\infty(A_u)^2.$$

Arguments similar to those used for Y_∞ in the proof of Lemma 3.1 show that this series converges almost surely and that the limit has moments of all orders.

Theorem 3.6. *As $n \to \infty$,*

$$\frac{1}{n^2}\,\mathrm{WI}(X_n) - H_n + 1 \;\to\; Y_\infty - W_\infty,$$

almost surely and in L^p for every $p > 0$.

Proof. As in the proof of the corresponding results for the other tree functionals, we project the right hand side of the formula on the natural filtration. For Y_∞ this has been done in Section 3.1. For W_∞, we proceed as follows: For $u = (u_1, \dots, u_k)$ and $i = 1, \dots, k$ let $\xi_i := \rho_{u_i}\big(X_\infty, (u_1, \dots, u_{i-1})\big)$. Then, as in the proof of Corollary 2.6, $X_\infty(A_u)^2 = \prod_{i=1}^k \xi_i^2$, so that, using (2.5) and the conditional independence from Theorem 2.5,

$$\begin{aligned}
&E[X_\infty(A_u)^2|\mathcal{F}_n] \\
&\quad= \prod_{i=1}^k E[\xi_i^2|\mathcal{F}_n] \\
&\quad= \prod_{i=1}^k \frac{\#X_n((u_1,\dots,u_i))\big(1+\#X_n((u_1,\dots,u_i))\big)}{\big(1+\sum_{j=1}^\infty \#X_n((u_1,\dots,u_{i-1},j))\big)\big(2+\sum_{j=1}^\infty \#X_n((u_1,\dots,u_{i-1},j))\big)} \\
&\quad= \prod_{i=1}^k \frac{\#X_n((u_1,\dots,u_i))\big(1+\#X_n((u_1,\dots,u_i))\big)}{\#X_n((u_1,\dots,u_{i-1}))\big(1+\#X_n((u_1,\dots,u_{i-1}))\big)} \\
&\quad= \frac{\#X_n(u)\,(\#X_n(u)+1)}{n(n+1)}
\end{aligned}$$

whenever $u \in X_n$.

In order to deal with the nodes not in X_n we use the operation $v \mapsto \bar v =: \phi(v)$ introduced in the proof of Lemma 3.2. Let

$$\partial X_n := \{v \notin X_n :\, \phi(v) \in X_n\}$$

be the set of external nodes of X_n and put

$$A_k(v) := \{w \in \mathbb{V} :\, \phi^k(w) = v\}, \quad k \in \mathbb{N}_0,$$

where $\phi^0(u) := u$. Clearly, $\#\partial X_n = n$, $\#A_k(v) = 2^k$ and $\mathbb{V} \setminus X_n = \sum_{v\in\partial X_n} \sum_{k=0}^{\infty} A_k(v)$. With $v = (v_1, \dots, v_k) \in \partial X_n$, $\xi_i := \rho_{v_i}\big(X_\infty, (v_1, \dots, v_{i-1})\big)$ and $\tilde v := (v_1, \dots, v_{k-1})$ we get

$$E[X_\infty(A_v)^2|\mathcal{F}_n] \;=\; \Big(\prod_{i=1}^{k-1} E[\xi_i^2|\mathcal{F}_n]\Big)\, E[\xi_k^2|\mathcal{F}_n] \;=\; \frac{\#X_n(\tilde v)\,(\#X_n(\tilde v)+1)}{n(n+1)}\, E[\xi_k^2|\mathcal{F}_n].$$

Conditionally on $\#X_n(\tilde v 1) = a_1, \dots, \#X_n(\tilde v j) = a_j$, $j := v_k - 1$, the distribution of ξ_k is equal to the distribution of YZ, with Y, Z independent and

$$Y \sim \mathrm{Beta}\Big(1, \sum_{i=1}^{j} a_i\Big), \quad Z \sim \mathrm{unif}(0,1).$$

In view of $1 + \sum_{i=1}^{j} a_i = \#X_n(\tilde v)$ we thus obtain, using (2.5) again,

$$E[\xi_k^2|\mathcal{F}_n] \;=\; \frac{2}{\#X_n(\tilde v)\,(\#X_n(\tilde v)+1)} \cdot \frac{1}{3},$$

and hence, for $w \in A_k(v)$,

$$E[X_\infty(A_w)^2|\mathcal{F}_n] \;=\; \frac{2}{n(n+1)} \Big(\frac{1}{3}\Big)^{k+1}.$$

For the contribution of the nodes not in X_n to the conditional expectation of W_∞ this gives

$$\begin{aligned}
\sum_{u\notin X_n} E[X_\infty(A_u)^2|\mathcal{F}_n] \;&=\; \sum_{v\in\partial X_n} \sum_{k=0}^{\infty} \sum_{w\in A_k(v)} E[X_\infty(A_w)^2|\mathcal{F}_n] \\
&=\; \sum_{v\in\partial X_n} \frac{2}{n(n+1)} \sum_{k=0}^{\infty} 2^k \Big(\frac{1}{3}\Big)^{k+1}. \\
&=\; \frac{2}{n+1}.
\end{aligned}$$

Putting pieces together we arrive at

$$E[W_\infty|\mathcal{F}_n] \;=\; \frac{1}{n(n+1)} \Big(n + \mathrm{TPL}(X_n) + \sum_{u\in X_n} \#X_n(u)^2\Big) \;+\; \frac{2}{n+1},$$

and we can now proceed as in the proof of Theorem 3.4. □

Again, we can use the proof to obtain expected values,

$$E\mathrm{WI}(X_n) \;=\; n(n+1)H_n - 2n^2 \quad \text{for all } n \in \mathbb{N}.$$

This agrees with Neininger's result [17, Theorem 1.2].

3.4 Distributional considerations

Let $X_{\infty,i}$, $i \in \mathbb{N}$, be as in the proof of Theorem 2.3. For the total path length the representation $Y_\infty = \Phi(X_\infty)$ in Section 3.1 of the limit Y_∞ in terms of X_∞ leads to

$$Y_\infty \;=\; C\big(\rho(X_\infty, \emptyset)\big) + \sum_{i=1}^{\infty} X_\infty(A_{(i)})\, Y_{\infty,i}, \tag{3.10}$$

with $Y_{\infty,i} := \Phi(X_{\infty,i})$. Note that this is an equality for random variables (strictly speaking, it refers to the underlying probability measure as we may have to discard a null set for X_∞ to be atom-free and diffuse). In terms of distributions this may be rewritten as

$$Y_\infty \;=_{\mathrm{d}}\; C(\rho) + \sum_{i=1}^{\infty} \rho_i \, Y_\infty^{(i)}, \tag{3.11}$$

with $\rho, Y_\infty^{(1)}, Y_\infty^{(2)}, \dots$ independent, and $\rho \sim \mathrm{GEM}$, $Y_\infty^{(i)} =_{\mathrm{d}} Y_\infty$ for all $i \in \mathbb{N}$. We recall that the 'toll function' $C : \Sigma_\infty : [-\infty, \infty)$ in this distributional fixed point equation is given by

$$C\big((s_i)_{i\in\mathbb{N}}\big) \;=\; 1 + \sum_{i=1}^{\infty} s_i \log s_i.$$

On the other hand, it is known [2] that the limiting total path length also satisfies

$$Y_\infty \;=_{\mathrm{d}}\; UY_\infty + (1-U)Y_\infty^\star + G(U), \tag{3.12}$$

with $G(u) := u + u\log u + (1-u)\log(1-u)$, $U, Y_\infty, Y_\infty^\star$ independent, $U \sim \mathrm{unif}(0,1)$, and $Y_\infty^\star =_{\mathrm{d}} Y_\infty$. What is the connection between the two equations?

Suppose that $\xi \sim \mathrm{GEM}$ and let $\zeta = (\zeta_i)_{i\in\mathbb{N}}$ be related to ξ as in (2.3). Consider the shifted sequence $\tilde\zeta = (\tilde\zeta_i)_{i\in\mathbb{N}}$ with $\tilde\zeta_i = \zeta_{i+1}$ for all $i \in \mathbb{N}$. Clearly, $\tilde\zeta$ is again a sequence of independent, $\mathrm{unif}(0,1)$-distributed random variables, and it is independent of ζ_1. This implies that the corresponding $\tilde\xi$ is GEM distributed, and we have

$$\begin{aligned} C(\xi) \;&=\; 1 + \zeta_1\log(\zeta_1) + (1-\zeta_1)\sum_{i=1}^{\infty}\tilde\xi_i\big(\log(1-\zeta_1) + \log\tilde\xi_i\big) \\ &=\; \zeta_1 + \zeta_1\log(\zeta_1) + (1-\zeta_1)\log(1-\zeta_1) + (1-\zeta_1)\,C(\tilde\xi). \end{aligned}$$

Using (3.10) we now get, with $\zeta_1 = \rho_1(X_\infty, \emptyset)$ and $\tilde\xi_i = \rho_{i+1}(X_\infty, \emptyset)$,

$$Y_\infty \;=\; G(\zeta_1) + \zeta_1 Y_{\infty,1} + (1-\zeta_1)Y_\infty^\star, \quad \text{with } \; Y_\infty^\star := C(\tilde\xi) + \sum_{i=1}^{\infty}\tilde\xi_i\, Y_{\infty,i+1}.$$

Together with (3.11) this leads to the distributional equation (3.12).

It is instructive to compare this with a proof of (3.12) that is based on the 'musical decomposition' in Section 2.2. The limit version of the decomposition given in Proposition 2.2 transforms X_∞ into independent components $\eta = X_\infty(A_{(1)})$, $X_\infty^\flat$ and $X_\infty^\sharp$ with the properties that

$$\begin{aligned} \rho(X_\infty, A_{u^\flat}) \;&=\; \rho(X_\infty^\flat, A_u) \;\text{ for all } u \in \mathbb{V}, \\ \rho(X_\infty, A_{u^\sharp}) \;&=\; \rho(X_\infty^\sharp, A_u) \;\text{ for all } u \in \mathbb{V},\, u \neq \emptyset, \end{aligned}$$

and with $\rho(X_\infty^\sharp, \emptyset) = \tilde\xi$, where $\tilde\xi$ is constructed from $\xi = \rho(X_\infty, \emptyset)$ as explained above. With this construction,

$$\begin{aligned} &\Phi(X_\infty) \\ &\quad= \sum_{u\in\mathbb{V}} X_\infty(A_u)\, C(\rho(X_\infty, u)) \\ &\quad= X_\infty(A_\emptyset)\, C(\rho(X_\infty,\emptyset)) + \sum_{u\in\mathbb{V}} X_\infty(A_{u^\flat})\, C(\rho(X_\infty, u^\flat)) + \sum_{u\in\mathbb{V}, u\neq\emptyset} X_\infty(A_{u^\sharp})\, C(\rho(X_\infty, u^\sharp)) \\ &\quad= C(\xi) + \eta \sum_{u\in\mathbb{V}} X_\infty^\flat(A_u)\, C(\rho(X_\infty^\flat, u)) + (1-\eta) \sum_{u\in\mathbb{V}, u\neq\emptyset} X_\infty^\sharp(A_u)\, C(\rho(X_\infty^\sharp, u)) \\ &\quad= C(\xi) - (1-\eta)\, C(\tilde\xi) \;+\; \eta\,\Phi(X_\infty^\flat) \;+\; (1-\eta)\,\Phi(X_\infty^\sharp), \end{aligned}$$

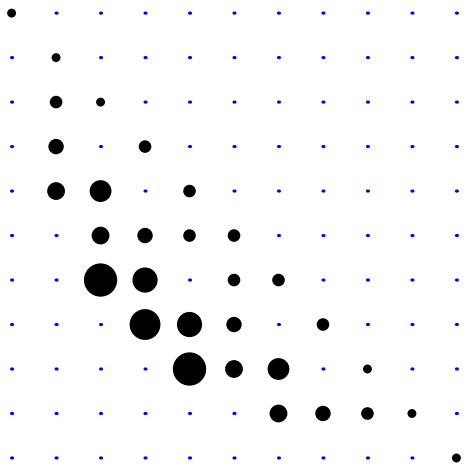

Figure 4: Joint distribution of the vertical and horizontal total path length for the trees in $\mathbb{H}_7$.

and it remains to make use of $C(\xi) - (1-\eta)C(\tilde{\xi}) = G(\eta)$, which we have proved above. Once again, we note that the decomposition takes place on the level of the random quantities themselves; there is no '$=_{\mathrm{d}}$'-sign.

As in the transition from Section 3.1 to Section 3.2 the detailed consideration of the vertical case now makes it easy to treat the horizontal path length. With $\Psi(X_\infty) = Y_\infty + Z_\infty$ the limit in Theorem 3.5 we just replace C by $C+D$ to obtain the decomposition

$$\Psi(X_\infty) \;=\; (C+D)(\xi) - (1-\eta)\,(C+D)(\tilde{\xi}) \;+\; \eta\,\Psi(X_\infty^{\flat}) \;+\; (1-\eta)\,\Psi(X_\infty^{\sharp}).$$

A straightforward computation gives $D(\xi) - (1-\eta)\,D(\tilde{\xi}) = 1-2\eta$, which leads to the horizontal analogue of (3.12) with $\tilde{G}(u) := 1 - u + u\log u + (1-u)\log(1-u)$ instead of G. Clearly, $\tilde{G}(\eta)$ and $G(\eta)$ are equal in distribution, which implies that the limit distributions arising in the vertical and horizontal case satisfy the same fixed point equation. It is straightforward to set up a metric space of probability distributions which contains these limit distributions and that turns the right hand side of (3.12) into a contraction, hence the limit distributions arising for the vertical and horizontal path length of random recursive trees are identical.

The above argument depends on the limit version of the decomposition. With some additional work the finite version in Section 2.2 can be used directly to obtain the convergence in distribution of the standardized path length; see [22] for the Quicksort situation. As pointed out at the beginning of this section, the contraction method may miss the fact that the random variables converge almost surely or in L^p. On the other hand, as the above path length example shows, the approach via a fixed point relation for the limit distribution may lead to the direct recognition of the equality of two limit distributions, which may not be apparent from the representation of the respective limit random variables in terms of the limit tree (indeed, the representations Y_∞ and $Y_\infty + Z_\infty$, given in Theorems 3.4 and 3.5 respectively, seem to suggest that the limit distributions are different).

Equality of the limit distributions naturally raises the question whether there is a relation between the respective distributions for finite trees. Figure 4 shows the pair (i,j) of values i for the vertical and j for the horizontal total path length for all $6! = 720$ recursive trees with 7 nodes, where the sizes of the black dots correspond to the multiplicities of the pairs and the blue dots represent pairs that do not appear. The

picture suggests that, up to a shift that is apparent from (3.8), the joint distribution of total vertical and total horizontal path length is symmetric. Clearly, this would imply that the limit distributions are the same.

We now define $T : \mathbb{V} \to \mathbb{V}$ by $T(\emptyset) = \emptyset$, $T((1)) = (1)$ and, if $u = (u_1, \dots, u_k)$ and $T(u) = v$ with $v = (v_1, \dots, v_j)$, by

$$\begin{aligned} T\big((u_1, \dots, u_k, 1)\big) &:= (v_1, \dots, v_{j-1}, v_j + 1), \\ T\big((u_1, \dots, u_{k-1}, u_k + 1)\big) &:= (v_1, \dots, v_j, 1). \end{aligned} \tag{3.13}$$

It is easy to see that T is bijective; in fact, $T^{-1} = T$ (T can be related to the natural correspondence mentioned after the proof of Theorem 2.1; see [16]). The recursive part (3.13) translates a move downwards into a move to the right and vice versa. Further, T is compatible with tree growth: If we add a node u to a tree x as a first child of $v \in x$, then $T(u)$ is the next next child to the parent of $T(u)$ and, again, vice versa. In particular, writing $T(x)$ for $\{T(u) : u \in x\}$, we may lift T to a bijective map on $\mathbb{H}$ with the property that $T(\mathbb{H}_n) = \mathbb{H}_n$ for all $n \in \mathbb{N}$. This construction proves

$$\mathcal{L}\big(\mathrm{TPL}(X_n) - (n-1)\big) \;=\; \mathcal{L}\big(\mathrm{HPL}(X_n)\big) \;\text{ for all } n \in \mathbb{N},\, n \geq 2,$$

if we can show that the distribution of the Harris chain $(X_n)_{n\in\mathbb{N}}$ is invariant under T and that

$$\mathrm{HPL}\big(T(x)\big) = \mathrm{TPL}\big(x\big) - 1 \;\text{ for all } x \in \mathbb{H},\, \#x > 1. \tag{3.14}$$

The first of these is an immediate consequence of the tree growth mechanism. To obtain (3.14) it is enough to show that

$$|T(u)|_1 - |T(u)| = |u| - 1 \;\text{ for all } u \in \mathbb{V},\, u \neq \emptyset.$$

This, however, can easily be proved by induction, considering the two cases in (3.13) separately.

In view of this simple bijective proof one may naturally wonder what the advantage of the boundary theory approach might be. The answer becomes clear as soon as we consider several functionals at the same time: Almost sure convergence of the standardized vertical and horizontal path lengths implies the convergence of any linear combinations, for example, whereas convergence in distribution does not 'vectorize' in this way. This is of interest in connection with the analysis of the recursive tree algorithm RT introduced in Section 2.4: The number C_n of comparisons needed to build the tree X_n for $n-1$ data is given by the sum of the horizontal and the vertical path length of X_n, hence

$$EC_n = 2nH_n - 3n - 1, \quad \frac{1}{n}\big(C_n - EC_n\big) \to 2Y_\infty + Z_\infty \text{ with probability } 1,$$

with Y_∞ and Z_∞ as in Sections 3.1 and 3.2. While the mean can be obtained from the symmetry and the individual results for the two versions of path length, we would need their joint distribution in order to obtain the limit result for the sum.

Acknowledgments. We thank the referees for their supportive and constructive comments.

References

[1] Shankar Bhamidi, Steven N. Evans, and Arnab Sen, *Spectra of large random trees*, J. Theoret. Probab. **25** (2012), no. 3, 613–654. MR-2956206

[2] Robert P. Dobrow and James Allen Fill, *Total path length for random recursive trees*, Combin. Probab. Comput. **8** (1999), no. 4, 317–333, Random graphs and combinatorial structures (Oberwolfach, 1997). MR-1723646

[3] J. L. Doob, *Discrete potential theory and boundaries*, J. Math. Mech. **8** (1959), 433–458; erratum 993. MR-0107098

[4] Michael Drmota, *Random Trees. An Interplay between Combinatorics and Probability*, Springer, Wien, 2009. MR-2484382

[5] Steven N. Evans, Rudolf Grübel, and Anton Wakolbinger, *Trickle-down processes and their boundaries*, Electron. J. Probab. **17** (2012), no. 1, 58. MR-2869248

[6] Qunqiang Feng and Zhishui Hu, *On the Zagreb index of random recursive trees*, J. Appl. Probab. **48** (2011), no. 4, 1189–1196. MR-2896676

[7] Philippe Flajolet and Robert Sedgewick, *Analytic Combinatorics*, Cambridge University Press, Cambridge, 2009. MR-2483235

[8] Michael Fuchs, Hsien-Kuei Hwang, and Ralph Neininger, *Profiles of random trees: limit theorems for random recursive trees and binary search trees*, Algorithmica **46** (2006), no. 3-4, 367–407. MR-2291961

[9] William Goh and Eric Schmutz, *Limit distribution for the maximum degree of a random recursive tree*, J. Comput. Appl. Math. **142** (2002), no. 1, 61–82, Probabilistic methods in combinatorics and combinatorial optimization. MR-1910519

[10] Rudolf Grübel, *Search trees: Metric aspects and strong limit theorems*, Ann. Appl. Probab. **24** (2014), no. 3, 1269–1297. MR-3199986

[11] Svante Janson, *Asymptotic degree distribution in random recursive trees*, Random Structures Algorithms **26** (2005), no. 1-2, 69–83. MR-2116576

[12] Maurice Kendall, Alan Stuart, and J. Keith Ord, *Kendall's Advanced Theory of Statistics. Vol. 1*, fifth ed., The Clarendon Press Oxford University Press, New York, 1987, Distribution Theory. MR-902361

[13] Donald E. Knuth, *The Art of Computer Programming. Vol. 1*, Addison-Wesley, Reading, MA, 1997, Fundamental algorithms, Third edition [of MR0286317]. MR-3077152

[14] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer, *Markov Chains and Mixing Times*, American Mathematical Society, Providence, RI, 2009. MR-2466937

[15] Hosam M. Mahmoud, *Limiting distributions for path lengths in recursive trees*, Probab. Engrg. Inform. Sci. **5** (1991), no. 1, 53–59. MR-1183165

[16] Igor Michailow, *Asymptotische Analyse zufälliger diskreter Strukturen mit Methoden der diskreten Potentialtheorie*, Ph.D. thesis, in preparation, Leibniz Universität Hannover, 2015.

[17] Ralph Neininger, *The Wiener index of random trees*, Combin. Probab. Comput. **11** (2002), no. 6, 587–597. MR-1940122

[18] J. Neveu, *Discrete-parameter Martingales*, revised ed., North-Holland, Amsterdam, 1975. MR-0402915

[19] Valentin V. Petrov, *Limit Theorems of Probability Theory*, Oxford Studies in Probability, vol. 4, The Clarendon Press, Oxford University Press, New York, 1995, Sequences of independent random variables, Oxford Science Publications. MR-1353441

[20] Boris Pittel, *Note on the heights of random recursive trees and random m-ary search trees*, Random Structures Algorithms **5** (1994), no. 2, 337–347. MR-1262983

[21] Mireille Régnier, *A limiting distribution for quicksort*, RAIRO Inform. Théor. Appl. **23** (1989), no. 3, 335–343. MR-1020478

[22] Uwe Rösler, *A limit theorem for "Quicksort"*, RAIRO Inform. Théor. Appl. **25** (1991), no. 1, 85–100. MR-1104413

[23] Robert T. Smythe and Hosam M. Mahmoud, *A survey of recursive trees*, Teor. Ĭmovīr. Mat. Stat. (1994), no. 51, 1–29. MR-1445048

[24] Jerzy Szymański, *On the complexity of algorithms on recursive trees*, Theoret. Comput. Sci. **74** (1990), no. 3, 355–361. MR-1073771

[25] _____, *On the maximum degree and the height of a random recursive tree*, Random graphs '87 (Poznań, 1987), Wiley, Chichester, 1990, pp. 313–324. MR-1094139

[26] Wolfgang Woess, *Denumerable Markov Chains. Generating Functions, Boundary Theory, Random Walks on Trees*, European Mathematical Society (EMS), Zürich, 2009. MR-2548569