Electronic Journal of Probability

Vol. 16 (2011), Paper no. 35, pages 1020–1047.

Journal URL
http://www.math.washington.edu/~ejpecp/

On the asymptotic internal path length and the asymptotic Wiener index of random split trees

Götz Olaf Munsonius

Institute of Mathematics, J.W. Goethe University
60054 Frankfurt a.M., Germany

munsonius@math.uni-frankfurt.de

Abstract

The random split tree introduced by Devroye (1999) is considered. We derive a second order expansion for the mean of its internal path length and furthermore obtain a limit law by the contraction method. As an assumption we need the splitter having a Lebesgue density and mass in every neighborhood of 1. We use properly stopped homogeneous Markov chains, for which limit results in total variation distance as well as renewal theory are used. Furthermore, we extend this method to obtain the corresponding results for the Wiener index.

Key words: random trees, probabilistic analysis of algorithms, internal path length, Wiener index.

AMS 2010 Subject Classification: Primary 60F05; 68P05; 05C05.

Submitted to EJP on January 13, 2011, final version accepted May 1, 2011.

1 Introduction

The random split tree introduced by Devroye (1999) is a general tree model which for special choices of its parameters covers various random trees that are fundamental in Computer Science for their use as data structures, e.g. binary search trees, quadtrees, m-ary search trees, simplex trees, tries etc. Many characteristic quantities of these trees such as node depths, height, path length or other distance measures between nodes describe the complexity of algorithms that make use of the trees. In the probabilistic analysis of algorithms the asymptotic behavior of such quantities is studied for this reason. Whereas often such characteristic quantities are studied one by one for each tree Devroye's idea was to derive universal results valid for the whole class of his split tree model.

We recall the definition of the split tree from Devroye (1999). Four parameters $b, s, s_0, s_1 \in \mathbb{N}_0$ are given where $b \geq 2$ is the branching factor, $s > 0$ is the vertex capacity and s_0 and s_1 satisfy the two conditions

$$0 \leq s_0 \leq s, \qquad 0 \leq bs_1 \leq s + 1 - s_0.$$

Furthermore, a random vector $\mathcal{V} = (V_1, \ldots, V_b) \in [0,1]^b$ with $\sum_{k=1}^{b} V_k = 1$ is given. The random split tree of size n is obtained by distributing n balls to the nodes of the infinite b-ary tree according to the following procedure. For a node u of the b-ary tree let $C(u)$ denote the number of balls already assigned to this node and $N(u)$ be the number of balls associated to any node in the subtree rooted at this node. For each node u take an independent copy $\mathcal{V}^{(u)} = (V_1^{(u)}, \ldots, V_b^{(u)})$ of the random vector $\mathcal{V}$. Initially, there are no balls (i.e. $C(u) = 0$ for all u) distributed. The balls are added to the tree sequentially. Adding a ball to a tree rooted at u proceeds as follows:

1. If u is not a leaf (i.e. $C(u) < N(u)$), choose child i with probability $V_i^{(u)}$, increment $N(u)$ by 1 and recursively add the ball to the subtree rooted at child i.

2. If u is a leaf and $C(u) = N(u) < s$, then add the ball to u and stop. $C(u)$ and $N(u)$ are incremented by 1.

3. If u is a leaf but $C(u) = N(u) = s$, we set $N(u) = s + 1$ and $C(u) = s_0$, place $s_0 \leq s$ randomly selected balls at u, give s_1 randomly selected balls to each of the b children of u and set $C(v) = s_1 = N(v)$ for all children v of u. After that, we add each of the remaining $s + 1 - s_0 - bs_1 \geq 0$ balls one by one randomly and independently to the subtree rooted at child i with probability $V_i^{(u)}$ by applying the procedure recursively.

Usually, one assumes that $V_i \stackrel{d}{=} V_1 =: V$ for all $i = 2, \ldots, b$ where V is called the splitter and its distribution is called the splitting distribution. By $\stackrel{d}{=}$ it is denoted that left and right hand side have identical distributions. Whenever the functional under consideration is independent of the tree ordering, this assumption does not mean any loss of generality. This can be seen by a random permutation argument, already stated in Devroye (1999). In this paper we need some additional assumption:

General assumption: Throughout this paper we assume that the distribution of V has a Lebesgue density f_V and that for the distribution function we have $F_V(x) < 1$ for all $x < 1$.

As mentioned in the beginning, the random split tree models many common random trees. For instance, choosing $s = s_0 = b - 1$ for some $b \geq 2$, $s_1 = 0$ and $V = \min\{U_1, \ldots, U_{b-1}\}$ where

$U_1, \dots, U_{b-1}$ are independent random variables uniformly distributed on $[0,1]$ one gets the random b-ary search tree. The random median-of-$(2k+1)$ binary search tree can be realized by setting $b = 2$, $s = 2k$, $s_0 = 1$, $s_1 = k$ and $V = \text{median}(U_1, \dots, U_{2k+1})$. Also some digital data structures are covered by the split tree model. For V uniformly distributed on the deterministic set $\{p_1, \dots, p_b\}$, $s = 1$ and $s_1 = 0$ one obtains in the case $s_0 = 0$ the trie and in the case $s_0 = 1$ the digital search tree. In Table 1 in Devroye (1999) more examples of important tree models are listed with the corresponding choices of the parameters.

The general assumption and with it the results of this paper hold true for many of these examples as random binary search trees, random b-ary search trees, random quadtrees, random median-of-$(2k+1)$ binary search trees, random simplex trees, (extended) AB trees and random m-grid trees. Whereas the results are not applicable to the common digital data structures as tries and digital search trees.

The depth of the n-th ball in a random split tree, denoted by D_n, is the number of edges on the path from the ball to the root of the tree. The internal path length of balls in the split tree is the sum of all depths of balls and is denoted by P_n for the tree with n balls. Thus, we have

$$P_n = \sum_{k=1}^{n} D_k.$$

The asymptotic expansion of the expectation of P_n was investigated for m-ary search trees in Mahmoud (1986), for random quadtrees by Flajolet et al. (1995) and for the median of $(2k+1)$-binary search tree by Chern and Hwang (2001) and Rösler (2001). In Holmgren (2010) the internal path length of random split trees is considered under the assumption that the splitting distribution is non-lattice. The first term and an upper bound of the second term of the asymptotic mean are derived using renewal theory.

Limit theorems for the distribution of the path length are proved for the random binary search tree in Régnier (1989) and Rösler (1991) and for the random recursive tree in Dobrow and Fill (1999).

Using the contraction method, Neininger and Rüschendorf (1999, Theorem 5.1) showed a universal limit theorem for the internal path length of random split trees under the assumption that the asymptotic expansion of the expectation of the internal path length is of the form

$$E[P_n] = d_1 n \log n + d_2 n + o(n) \tag{1}$$

as $n \to \infty$. Therefore, it is of interest to characterize all splitting distributions providing an asymptotic expectation of the form (1). The first result of this paper is the following.

Theorem 1.1. *Let P_n denote the internal path length in a random split tree of size n with branching factor b where the one-dimensional marginal distribution V of the splitting vector fulfills the general assumption. Then there exists a constant $c_p \in \mathbb{R}$ with*

$$E[P_n] = \frac{1}{\mu} n \log n + c_p n + o(n)$$

as $n \to \infty$ where $\mu = -bE[V \log V]$.

To state the result which follows from the combination of the limit theorem from Neininger and Rüschendorf (1999) with Theorem 1.1 we introduce some notation. By $\mathscr{M}_{0,2}$ we denote the set of centered probability measures on $\mathbb{R}$ with finite second moments. We denote the distribution of a random variable X by $\mathscr{L}(X)$ or P^X. The Wasserstein-metric ℓ_2 on $\mathscr{M}_{0,2}$ is defined by

$$\ell_2(\nu_1,\nu_2) := \inf\{\|X-Y\|_2 : \mathscr{L}(X)=\nu_1, \mathscr{L}(Y)=\nu_2\} \tag{2}$$

where the L_2-norm $\|\cdot\|_2$ is given by $\|X\|_2 = (E[\|X\|^2])^{1/2}$. For random variables X and Y we set $\ell_2(X,Y) := \ell_2(\mathscr{L}(X),\mathscr{L}(Y))$. It is well known that convergence with respect to the metric ℓ_2 (denoted by $\xrightarrow{\ell_2}$) is equivalent to weak convergence plus convergence of the second moments (see e.g. Bickel and Freedman (1981)).

Corollary 1.2. *Let P_n denote the internal path length in a random split tree of size n where the one-dimensional marginal distribution of the splitting vector $(V_1,\dots,V_b)$ fulfills the general assumption. Define $X_n := (P_n - E[P_n])/n$. Then the following holds true:*

1. *As $n\to\infty$ we have $\ell_2(X_n,X)\to 0$ where $\mathscr{L}(X)$ is the in $\mathscr{M}_{0,2}$ unique solution of the fixed point equation*

$$X \overset{d}{=} \sum_{k=1}^{b} V_k X^{(k)} + 1 + \frac{1}{\mu}\sum_{k=1}^{b} V_k \log V_k$$

   where $\mu := -bE[V_1\log V_1]$, $\mathscr{L}(X^{(k)}) = \mathscr{L}(X)$ for all $k = 1,\dots,b$ and $X, X^{(1)},\dots,X^{(b)},(V_1,\dots,V_b)$ are independent.

2. *In particular, the convergence in a) implies*

$$\mathrm{Var}(P_n) = \sigma^2 n^2 + o(n^2)$$

   with

$$\sigma^2 = \left(\frac{1}{\mu^2}E\left[\left(\sum_{k=1}^{b} V_k\log V_k\right)^2\right] - 1\right)\left(1-\sum_{k=1}^{b}E\left[V_k^2\right]\right)^{-1}.$$

3. *Exponential moments exist and converge,*

$$E[\exp(\lambda X_n)] \to E[\exp(\lambda X)], \qquad \lambda\in\mathbb{R}.$$

4. *For all $k\in\mathbb{N}$ we have as $n\to\infty$,*

$$P(|P_n - E[P_n]| \ge \varepsilon E[P_n]) = O(n^{-k}).$$

Remark 1.3. The tail bound given in d) is known not to be sharp in particular examples. McDiarmid and Hayward (1996) and Fill and Janson (2002) give a more precise bound for the random binary search tree.

The Wiener index of a random split tree is defined as the sum of the distances between all unordered pairs of balls, where the distance between two balls is given by the minimum number of edges connecting the nodes which are associated to the balls. For trees, the two dimensional vector consisting of the Wiener index and the internal path length suffices a recursion formula similar to

that of the latter one. Using this recursion formula, Neininger (2002) proved a limit theorem for the Wiener index of the random binary search tree and the random recursive tree by the use of the multivariate contraction theorem. In a final remark, Neininger (2002) mentioned that a limit theorem for the Wiener index of the general split tree can be proved in a similar way after determining the asymptotic expansion of its expectation sufficiently well.

We prove this asymptotic expansion and use the contraction method to obtain the limit theorem for the Wiener index of random split trees which fulfil the general assumption.

Theorem 1.4. *Let W_n denote the Wiener index in a random split tree of size n with branching factor b where the one-dimensional marginal distribution V of the splitting vector fulfills the general assumption. Then there exists a constant $c_w \in \mathbb{R}$ with*

$$E[W_n] = \frac{1}{\mu} n^2 \log n + c_w n^2 + o(n)$$

as $n \to \infty$ where $\mu = -bE[V \log V]$.

We denote by $\mathscr{M}_{0,2}^2$ the set of centered probability measures on $\mathbb{R}^2$ with finite second moments. The Wasserstein-metric ℓ_2 on the set $\mathscr{M}_{0,2}^2$ is defined similarly to the one-dimensional case.

Theorem 1.5. *Let (W_n, P_n) denote the vector consisting of the Wiener index and the internal path length of a random split tree of size n with branching factor b where the one-dimensional marginal distribution of the splitting vector $(V_1, \dots, V_b)$ fulfills the general assumption. Then the following holds true:*

1. *We have as $n \to \infty$,*

$$\ell_2\left(\left(\frac{W_n - E[W_n]}{n^2}, \frac{P_n - E[P_n]}{n}\right), (W, P)\right) \to 0$$

where (W, P) is the unique distributional fixed-point of the map $T : \mathscr{M}_{0,2}^2 \to \mathscr{M}_{0,2}^2$ given for $\nu \in \mathscr{M}_{0,2}^2$ by

$$T(\nu) := \mathscr{L}\left(\sum_{i=1}^{b} \begin{bmatrix} V_i^2 & V_i(1-V_i) \\ 0 & V_i \end{bmatrix} \begin{pmatrix} X_1^{(i)} \\ X_2^{(i)} \end{pmatrix} + \begin{pmatrix} b_1^* \\ b_2^* \end{pmatrix}\right)$$

with

$$\begin{pmatrix} b_1^* \\ b_2^* \end{pmatrix} = \frac{1}{\mu} \sum_{i=1}^{b} V_i \log V_i \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \begin{pmatrix} (1 + c_p - c_w)\left(1 - \sum_{i=1}^{b} V_i^2\right) \\ 1 \end{pmatrix}$$

where $\mathscr{L}(X^{(i)}) = \nu$ for $X^{(i)} := (X_1^{(i)}, X_2^{(i)})$, and $X^{(1)}, \dots, X^{(b)}, D, Z$ are independent.

2. *In particular, the convergence in a) implies*

$$\mathrm{Var}(W_n) = \sigma^2 n^4 + o(n^4)$$

with some constant $\sigma^2 > 0$.

Remark 1.6. The constant $\mu = -bE[V \log V]$ in the first order terms of the expectations of the internal path length and of the Wiener index appears already in the results about the height and depth in Devroye (1999). There, the explicit values of this constant for the individual splitting distributions are given in Table 2.

Remark 1.7. Besides the internal path length for the balls considered here, there is also the internal path length for the nodes where the depths of all nodes are summed up. Since there can be up to s balls in one node, these two path lengths may differ. In Holmgren (2010), the relation between the two versions is investigated. Let N_n denote the number of nodes in the random split tree with n balls. Assuming that the distribution of $-\log V$ is non-lattice, $P(V=1)=P(V=0)=0$ and

$$E[N_n] = \alpha n + O\left(\frac{n}{(\log n)^{1+\varepsilon}}\right) \tag{3}$$

for some constant $\alpha > 0$ and $\varepsilon > 0$, Holmgren (2010) showed that Theorem 1.1 implies the similar asymptotic behavior for the internal path length for the nodes in that random split tree. This finally yields the general limit theorem for the internal path length for the nodes in split trees which additionally fulfil equation (3). For instance, Mahmoud and Pittel (1989) showed the stronger result $E[N_n] = \alpha n + O(n^{1-\varepsilon})$ in the case of the b-ary search tree.

It seems that there are no results on the corresponding alternative version of the Wiener index in terms of the node-to-node distances.

The internal path length and the Wiener index have been considered also for random trees that do not belong to the class of split trees. A universal limit law for the path length of simply generated trees is proved in Janson (2003) where the limit distribution is given as a function of the Brownian excursion. Furthermore, the moments of the limit are derived. For the class of random increasing trees, which covers in particular the random recursive tree and the plane oriented recursive tree, the second order asymptotic of the expectation of the internal path length is derived in Bergeron et al. (1992). In Munsonius and Rüschendorf (2010) the asymptotic behavior of the expectation and a limit theorem for the internal path length of random b-ary trees with weighted edges is proved. By special choices of the edge weights, the analogous results are obtained for the class of random linear recursive trees, which encompasses in particular the random plane oriented recursive tree. Tail bounds for the Wiener index of random binary search trees have been considered by Ali Khan and Neininger (2007).

For a random split tree with n balls we denote by $I_n = (I_{n,1}, \dots, I_{n,b})$ the vector of the sizes of the subtrees, i.e. the number of balls assigned to nodes in the subtrees, rooted at the children of the root. By the construction of the split tree it follows that I_n is conditionally given $\mathcal{V}^{(\text{root})} = (v_1, \dots, v_b)$ multinomial distributed $M(n - s_0 - bs_1; v_1, \dots, v_b)$. Thus, under the assumption that $V_i \stackrel{d}{=} V_1 =: V$ for all $i = 2, \dots, b$ we obtain

$$P(I_{n,i} = k + s_1) = \int_0^1 \binom{\eta_n}{k} x^k (1-x)^{\eta_n - k} \mathrm{d}P^V(x), \tag{4}$$

where we set $\eta_n := n - s_0 - bs_1$. Throughout this paper, $\mathrm{Bin}(m, x)$ denotes a random variable with binomial distribution with parameters $m \in \mathbb{N}$ and $x \in [0,1]$.

The proofs of Theorem 1.1 and Theorem 1.4 are based on a method developed in Bruhn (1996) for recurrences where the toll function is bounded. In Section 2, we recall definitions and results of Bruhn (1996) and extend his method to the case of an unbounded toll function. We check the conditions of this method in the case of the random split tree in Section 3. Section 4 is devoted to the application in the case of the internal path length and the proof of Theorem 1.1. In Section 5 we give the proofs of Theorem 1.4 and Theorem 1.5 concerning the Wiener index.

Acknowledgement. The author is grateful to Ralph Neininger for several hints to literature and for comments to previous versions of this paper and to Nicolas Broutin for helpful discussions and making a preliminary manuscript of the paper Broutin and Holmgren (2011) on the internal path length of split trees available to him. Furthermore, he thanks an unknown referee for valuable suggestions for improvement of the paper.

2 The setting of Bruhn

Starting from recursion formulas of the form

$$H_n = \sum_{k=0}^{n-1} \nu_n(\{k\}) H_k + r(n)$$

where ν_n is a probability measure on $\{0,\dots,n-1\}$ for all $n \in \mathbb{N}$, the main idea of Bruhn (1996) is to define a homogeneous Markov chain $(S_t)_{t\in\mathbb{N}}$ with state space $\mathscr{E} = \{-\log n : n \in \mathbb{N}\} \cup \{1\}$ where the transition probabilities are given for $n > 0$ by

$$P(S_1 = x \mid S_0 = -\log n) = \begin{cases} \nu_n(\{e^{-x}\}), & \text{for } x \in \{-\log(n-1), \dots, -\log 1\} \\ \nu_n(\{0\}), & \text{for } x = 1 \end{cases}$$

and $P(S_1 = 1 \mid S_0 = 1) = 1$. Now, let $\sigma(n_1) := \inf\{t \mid S_t > -\log n_1\}$ be the stopping time when the Markov chain exceeds $-\log n_1$ for $n_1 \in \mathbb{N}$. Then, Bruhn proved the representation formula given in the following Lemma. (Since the PhD-thesis of Bruhn seems to be not available in English, the proofs of Bruhn (1996) are stated in Appendix B.)

We denote by $Y_t := S_t - S_{t-1}$ the increments of S. For $x \in \mathscr{E}$ we write $P_x(\cdot)$ in short for $P(\cdot \mid S_0 = x)$ and correspondingly $E_x[\cdot]$ for the expectation with respect to the measure P_x. We denote by F_x the distribution function of $P_x^{S_1 - x}$, i.e. $F_x(y) = P(S_1 - x \le y \mid S_0 = x)$.

Lemma 2.1. *Let H_n be a sequence of real numbers satisfying*

$$H_n = \sum_{k=0}^{n-1} \nu_n(\{k\}) H_k + r(n)$$

for some function r. Then it is for any $n_1 \in \mathbb{N}$ with the notations above

$$H_n = E_{-\log n} H_{\exp(-S_{\sigma(n_1)})} + E_{-\log n} \sum_{t=0}^{\sigma(n_1)-1} r(\exp(-S_t)). \tag{5}$$

To analyze the Markov chain $(S_t)_{t\in\mathbb{N}}$ we consider in the following a general state space $\mathscr{E} \subset \mathbb{R}$.

Definition 2.2. The Markov chain $(S_t)_{t\in\mathbb{N}_0}$ is said to be an *AR-process* (approximate renewal) if the state space $\mathscr{E}$ has no lower bound, the increments $Y_t := S_t - S_{t-1}$ are strictly positive, F_x converges in distribution as $x \to -\infty$ to a distribution function F, i.e. for all points t where F is continuous it is

$$\lim_{x\to-\infty} F_x(t) = F(t),$$

and $0 < \int t \, \mathrm{d}F(t) < \infty$.

For $a \in \mathbb{R}_-$ we define $\bar{F}_a : \mathbb{R} \to [0,1]$ by $\bar{F}_a(t) := \inf_{x \le a} F_x(t)$ and $\underline{F}_a : \mathbb{R} \to [0,1]$ by $\underline{F}_a(t) := \sup_{x \le a} F_x(t)$.

Definition 2.3. The set of distributions $\{F_x\}$ fulfills the *integrability condition* if

$$\lim_{a \to -\infty} \int x \, \mathrm{d}\bar{F}_a(x) = \int x \, \mathrm{d}F(x).$$

In the case of an AR-process, the theorem of dominated convergence implies that the integrability condition is equivalent to

$$\int x \, \mathrm{d}\bar{F}_a(x) < \infty \tag{6}$$

for some $a \in \mathbb{R}$.

The first summand in (5) can be handled by considering the distribution of $S_{\sigma(n_1)}$. The following key result is implicitly given in Rösler (2001) in a more general setting. The essential part of Rösler (2001) which gives the proof is stated in Appendix A in a self-contained way. For probability measures P and Q, let $d_{\mathrm{TV}}(P,Q)$ denote their total variation distance. Moreover, we define $\tau(d) := \inf\{t : S_t \ge d\}$.

Lemma 2.4. *Let $(S_t)_{t \in \mathbb{N}}$ be an AR-process which fulfills the integrability condition with a discrete state space $\mathscr{E}$. If there exist $\varepsilon > 0$, $x_0 \in \mathbb{R}_-$ and $K > 0$ such that for all $x, y \le x_0$ with $|x - y| \le K$ we have*

$$d_{\mathrm{TV}}\left(P_x^{S_1}, P_y^{S_1}\right) < 2(1-\varepsilon) \qquad \text{and} \qquad \lim_{x_0 \to -\infty} \inf_{z < y \le x_0} P_z(S_{\tau(y)} - y \le K) > 0, \tag{7}$$

then it holds for any $a \in \mathbb{R}_-$

$$\lim_{x_0 \to -\infty} \sup_{x,y \le x_0} d_{\mathrm{TV}}\left(P_x^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right) = 0.$$

The asymptotic behavior of the second summand in (5) can be analyzed by using the elementary renewal theorem. Since the Markov chain $(S_t)_{t \in \mathbb{N}}$ is not a renewal process, we couple it with three renewal processes using the functions F, $\bar{F}_a$ and $\underline{F}_a$. Because of the convergence $\lim_{x \to -\infty} F_x(t) = F(t)$, the functions $\bar{F}_a$ and $\underline{F}_a$ are again distribution functions.

Considering the AR-process (S_t) from above, there exists a sequence of independent random variables $(U_r)_{r \in \mathbb{N}}$ uniformly distributed on $[0,1]$ such that

$$Y_t = F^{-1}_{S_{t-1}} \circ U_t$$

for all $t \in \mathbb{N}$.

For $a \in \mathbb{R}$ we define three renewal processes $\bar{S}^{(a)}$, $\underline{S}^{(a)}$ and $\tilde{S}$ by $\bar{S}^{(a)}_0 = \underline{S}^{(a)}_0 = \tilde{S}_0 = S_0$ and the i.i.d. increments $\bar{Y}^{(a)}_r$, $\underline{Y}^{(a)}_r$ and $\tilde{Y}_r$ given by

$$\bar{Y}^{(a)}_t := \bar{F}^{-1}_a \circ U_t, \qquad \underline{Y}^{(a)}_t := \underline{F}^{-1}_a \circ U_t \qquad \text{and} \qquad \tilde{Y}_r := F^{-1} \circ U_t.$$

Thus, for all $t \in \mathbb{N}$ we have $\underline{Y}^{(a)}_t \le S_t - S_{t-1} \le \bar{Y}^{(a)}_t$ whenever $S_{t-1} \le a$.

Moreover, for each $t \in \mathbb{N}$ the sequence $\bar{Y}^{(a)}_t$ is decreasing and $\underline{Y}^{(a)}_t$ is increasing as $a \to -\infty$. Both sequences converge almost surely to $\tilde{Y}_r$.

Finally, we define the following stopping times for $a, d \in \mathbb{R}$:

$$\begin{aligned} \tau(d) &:= \inf\{t : S_t \geq d\}, & \gamma(d) &:= \inf\{t : S_t - S_0 \geq d\}, \\ \bar{\tau}^{(a)}(d) &:= \inf\{t : \bar{S}_t^{(a)} \geq d\}, & \bar{\gamma}^{(a)}(d) &:= \inf\{t : \bar{S}_t^{(a)} - \bar{S}_0^{(a)} \geq d\}, \\ \underline{\tau}^{(a)}(d) &:= \inf\{t : \underline{S}_t^{(a)} \geq d\}, & \underline{\gamma}^{(a)}(d) &:= \inf\{t : \underline{S}_t^{(a)} - \underline{S}_0^{(a)} \geq d\}, \\ \text{and} & & \tilde{\gamma}(d) &:= \inf\{t : \tilde{S}_t - \tilde{S}_0 \geq d\}. \end{aligned}$$

Using the renewal process $(\bar{S}_t)_{t\in\mathbb{N}}$, Bruhn (1996) shows the following result. (The proof is given in AppendixB.)

Lemma 2.5 (Bruhn (1996), Lemma 3.4)**.** *Consider an AR-process (S_t) with the notations above. Then there exist a real number a_* and a positive real number $\hat{u}(a_*)$ such that for all measurable functions $l : \mathbb{R} \to \mathbb{R}_+$, all real numbers y, z and all $x \in \mathscr{E}$ with $x < y < z < a_*$ we have*

$$E_x\left[\sum_{t=\tau(y)}^{\tau(z)-1} l(S_t)\right] \leq \hat{u}(a_*) \sum_{n=\lfloor y \rfloor}^{\lceil z \rceil} \sup_{t\in(n-1,n]} l(t).$$

To investigate also recurrences where the toll function r is not bounded as it is for example in the case of the Wiener index, we complete the results of Bruhn by the following lemma and corollary.

Lemma 2.6. *It holds for all decreasing continuous functions $l : \mathbb{R} \to \mathbb{R}_+$ and any $d \in \mathbb{R}_+$*

$$\begin{aligned} \lim_{a\to-\infty} E\left[\sum_{t=1}^{\bar{\gamma}^{(a)}(d)} l\left(\bar{S}_t^{(a)} - \bar{S}_0^{(a)}\right)\right] &= \lim_{a\to-\infty} E\left[\sum_{t=1}^{\underline{\gamma}^{(a)}(d)} l\left(\underline{S}_t^{(a)} - \underline{S}_0^{(a)}\right)\right] \\ &= E\left[\sum_{t=1}^{\tilde{\gamma}(d)} l\left(\tilde{S}_t - \tilde{S}_0\right)\right] < \infty. \end{aligned}$$

Proof. First, we consider the sequence $(\bar{S}_t^{(a)})$. By the construction we know that for each $s, t \in \mathbb{N}$ the mapping $a \mapsto \bar{Y}_s^{(a)}$ and thus the mapping $a \mapsto \bar{S}_t^{(a)} - \bar{S}_0^{(a)}$ are decreasing and converge almost surely to $\tilde{Y}_s$ and $\tilde{S}_t - \tilde{S}_0$ as $a \to -\infty$. This yields that for $d \in \mathbb{R}$ the mapping $a \mapsto \bar{\gamma}^{(a)}(d)$ is increasing and bounded from above by $\tilde{\gamma}(d)$. It is easy to see that $\bar{\gamma}^{(a)}(d) \to \tilde{\gamma}(d)$ almost surely as $a \to -\infty$. Since $\bar{\gamma}^{(a)}(d) \in \mathbb{N}$ for all $a \in \mathbb{R}$ and l is continuous, we obtain as $a \to -\infty$ almost surely

$$\sum_{t=1}^{\bar{\gamma}^{(a)}(d)} l\left(\bar{S}_t^{(a)} - \bar{S}_0^{(a)}\right) \to \sum_{t=1}^{\tilde{\gamma}(d)} l\left(\tilde{S}_t^{(a)} - \tilde{S}_0^{(a)}\right).$$

Furthermore, the left hand side is increasing as $a \to -\infty$ and

$$E\left[\sum_{t=1}^{\tilde{\gamma}(d)} l\left(\tilde{S}_t^{(a)} - \tilde{S}_0^{(a)}\right)\right] \leq l(0)E[\tilde{\gamma}(d)]$$

where we use that l is decreasing. The positivity of $\tilde{Y}_s$ ensures by Gut (1988, Chapter II, Theorem 3.1) that $E[\tilde{\gamma}(d)] < \infty$ and the claim follows for the first sum.

With the same arguments, we have

$$\sum_{t=1}^{\underline{\gamma}^{(a)}(d)} l\left(\underline{S}_t^{(a)} - \underline{S}_0^{(a)}\right) \to \sum_{t=1}^{\tilde{\gamma}(d)} l\left(\tilde{S}_t^{(a)} - \tilde{S}_0^{(a)}\right) \tag{8}$$

almost surely as $a \to -\infty$ and the left hand side is decreasing. It is

$$E\left[\sum_{t=1}^{\underline{\gamma}^{(a)}(d)} l\left(\underline{S}_t^{(a)} - \underline{S}_0^{(a)}\right)\right] \leq l(0) E[\underline{\gamma}^{(a)}(d)].$$

The monotone convergence theorem provides $\lim_{a\to-\infty} E[\underline{Y}_t^{(a)}] = E[\tilde{Y}_t] > 0$. Thus, $E[\underline{Y}_t^{(a)}] > 0$ for $a \in \mathbb{R}$ small enough and the elementary renewal theorem (see e.g. Gut 1988, Section II.4) implies $E[\underline{\gamma}^{(a)}(d)] < \infty$. So, the claim follows from (8) by the monotone convergence theorem. □

Choosing $l(x) = \exp(-\alpha x)$ with $\alpha > 0$ yields the following result.

Corollary 2.7. *For $\alpha, d > 0$ there exists a constant $c \in \mathbb{R}$ such that for each $\varepsilon > 0$ there exists $n_0 \in \mathbb{N}$ with*

$$\frac{1}{n^\alpha} E_{-\log n}\left[\sum_{t=0}^{\tau(-\log n + d)} \exp(-\alpha S_t)\right] \in (c - \varepsilon, c + \varepsilon)$$

for all $n \geq n_0$.

Proof. By construction we have for $-\log n + d \leq a$

$$\begin{aligned}\sum_{t=0}^{\bar{\gamma}^{(a)}(d)} \exp(-\alpha(\bar{S}_t^{(a)} - \bar{S}_0^{(a)})) &\leq \sum_{t=0}^{\gamma(d)} \exp(-\alpha(S_t - S_0) \\ &\leq \sum_{t=0}^{\underline{\gamma}(d)} \exp(-\alpha(\underline{S}_t^{(a)} - \underline{S}_0^{(a)})).\end{aligned}$$

For $\varepsilon > 0$, Lemma 2.6 provides $a_* \in \mathbb{R}$ such that for all $a < a_*$ we have

$$\left| E\left[\sum_{t=0}^{\bar{\gamma}(d)} \exp\left(-\alpha\left(\bar{S}_t^{(a)} - \bar{S}_0^{(a)}\right)\right)\right] - E\left[\sum_{t=0}^{\underline{\gamma}(d)} \exp\left(-\alpha\left(\underline{S}_t^{(a)} - \underline{S}_0^{(a)}\right)\right)\right] \right| < \varepsilon.$$

We choose n_0 such that $-\log n_0 + d \leq a_*$. Since we have for $n \geq n_0$

$$E_{-\log n}\left[\sum_{t=0}^{\tau(-\log n + d)} \exp(-\alpha S_t)\right] = n^\alpha E_{-\log n}\left[\sum_{t=0}^{\gamma(d)} \exp\left(-\alpha(S_t - S_0)\right)\right]$$

the claim follows using Lemma 2.6 once more. □

3 Recurrences for the random split tree

We consider a random split tree with the notation as introduced in Section 1 and set $\nu_n(\{k\}) := b\frac{k}{n}P(I_{n,1}=k)+\frac{s_0}{n}\mathbb{1}_{\{k=n-s_0\}}$. This function ν_n defines a probability measure on the set $\{0,\dots,n-s_0\}$. This is seen by summing up all values

$$\begin{aligned}\sum_{k=0}^{n-s_0}\nu_n(\{k\}) &= b\frac{1}{n}E[I_{n,1}]+\frac{s_0}{n}\\ &= \frac{n-s_0}{n}+\frac{s_0}{n}\\ &= 1.\end{aligned}$$

For the rest of the paper, we consider the Markov chain $(S_t)_{t\in\mathbb{N}}$ from Section 2 where the transition probabilities are given by this special choice of ν. In this section, we prove that for this choice the conditions of the Lemmata of the previous section are fulfilled.

3.1 The distribution of the subtreesize

When doing this, we frequently use the fact that the size of the first subtree rescaled properly converges.

Lemma 3.1. *For $\varepsilon>0$ we have*

$$P\left(\left|\frac{I_{n,1}}{n}-V\right|\geq\varepsilon\right)\leq 2\exp\left(-\frac{n\varepsilon^2}{4}\left(1+O\left(\frac{1}{n}\right)\right)\right).$$

In particular, this yields

$$E\left[\left|\frac{I_{n,1}}{n}-V\right|\right]=O\left(n^{-\frac{1}{3}}\right).$$

Proof. Starting from the distribution of $I_{n,1}$ given in (4) we obtain by Bernstein's inequality

$$\begin{aligned}P\left(\left|\frac{I_{n,1}}{n}-V\right|\geq\varepsilon\right) &= \int_0^1 P\left(\left|\mathrm{Bin}(\eta_n,x)-nx\right|\geq n\varepsilon\right)\mathrm{d}P^V(x)\\ &\leq 2\exp\left(-\frac{n\varepsilon^2}{4}\left(1+O\left(\frac{1}{n}\right)\right)\right).\end{aligned}$$

Since it is $|I_{n,1}/n-V|\leq 1$, this yields for the expectation

$$\begin{aligned}E\left[\left|\frac{I_{n,1}}{n}-V\right|\right] &= E\left[\left(\mathbb{1}_{\left\{\left|\frac{I_{n,1}}{n}-V\right|\leq n^{-\frac{1}{3}}\right\}}+\mathbb{1}_{\left\{\left|\frac{I_{n,1}}{n}-V\right|>n^{-\frac{1}{3}}\right\}}\right)\left|\frac{I_{n,1}}{n}-V\right|\right]\\ &\leq n^{-\frac{1}{3}}+2\exp\left(-\frac{n^{1/3}}{4}\left(1+O\left(\frac{1}{n}\right)\right)\right)\\ &= O\left(n^{-\frac{1}{3}}\right).\end{aligned}$$

□

At this point, we prove some asymptotic expansions needed later.

Lemma 3.2. *For the size of the first subtree $I_{n,1}$ in a random split tree with splitting distribution V it holds*

$$E[I_{n,1}^2] = E[V^2]n^2 + o(n^2),$$

$$E[I_{n,1} \log I_{n,1}] = \frac{1}{b} n \log n + E[V \log V]n + o(n)$$

and

$$E[I_{n,1}^2 \log I_{n,1}] = E[V^2]n^2 \log n + E[V^2 \log V]n^2 + o(n^2).$$

Proof. It is

$$\begin{aligned} E[I_{n,1}^2] &= \int_0^1 E[\mathrm{Bin}(\eta_n, x)^2] \mathrm{d}P^V(x) \\ &= \int_0^1 (\eta_n x(1-x) + \eta_n^2 x^2) \mathrm{d}P^V(x) \\ &= E[V^2]n^2 + o(n^2). \end{aligned} \tag{9}$$

Furthermore, we have by Lemma 3.1 $I_{n,1}/n \to V$ in probability. Since $x \mapsto x^k \log x$ is bounded on the interval $[0,1]$, we obtain for $k = 1, 2$

$$E\left[\frac{I_{n,1}^k}{n^k} \log \frac{I_{n,1}}{n}\right] \to E[V^k \log V].$$

This implies

$$E\left[I_{n,1}^k \log \frac{I_{n,1}}{n}\right] = E[V^k \log V]n^k + o(n^k).$$

On the other hand we have

$$E\left[I_{n,1}^k \log \frac{I_{n,1}}{n}\right] = E\left[I_{n,1}^k \log I_{n,1}\right] - E\left[I_{n,1}^k\right] \log n.$$

The claims follow with result (9) since we have $E[I_{n,1}] = (n - s_0)/b$. □

3.2 The Markov chain for the random split tree

Now, we consider the Markov chain from Section 2 with the transition probabilities $\nu_n(\{k\}) = b\frac{k}{n} P(I_{n,1} = k) + \frac{s_0}{n} \mathbb{1}_{\{k=n-s_0\}}$.

Lemma 3.3. *The process $(S_t)_{t \in \mathbb{N}_0}$ is an AR-process and the corresponding set of distributions $\{F_x\}$ fulfills the integrability condition.*

Proof. Since ν_n is a probability measure on the set $\{0,\dots,n-s_0\}$ we have $Y_t > 0$ for all t. For $x = -\log n$ we have by dominated convergence and Lemma 3.1 for any $y \in \mathbb{R}$

$$\begin{aligned}
F_x(y) &= P(Y_1 \le y \mid S_0 = x) \\
&= \sum_{k\in\mathbb{N}:-\log\frac{k}{n}\le y} \nu_n(\{k\}) \\
&= \sum_{k\in\mathbb{N}:-\log\frac{k}{n}\le y} b\frac{k}{n}P(I_{n,1}=k) + \frac{s_0}{n}\mathbb{1}_{\{n-s_0\ge e^{-y}n\}} \\
&= bE\left[\frac{I_{n,1}}{n}\mathbb{1}_{\{-\log(I_{n,1}/n)\le y\}}\right] + \frac{s_0}{n}\mathbb{1}_{\{n-s_0\ge e^{-y}n\}} \\
&\xrightarrow{n\to\infty} bE[V\mathbb{1}_{\{-\log V\le y\}}] =: F(y).
\end{aligned}$$

Moreover, we obtain with Fubini's Theorem

$$\begin{aligned}
\int_0^\infty t\,\mathrm{d}F(t) &= \int_0^\infty (1-F(t))\,\mathrm{d}t \\
&= \int_0^\infty bE\left[V\mathbb{1}_{\{-\log V>t\}}\right]\mathrm{d}t \\
&= -bE[V\log V].
\end{aligned}$$

This yields $0 < \int t\,\mathrm{d}F(t) < \infty$.
It remains to show the integrability condition, which means

$$\int t\,\mathrm{d}\bar{F}_a(t) < \infty$$

for an $a \in \mathbb{R}$ and $\bar{F}_a(t) := \inf_{x\le a} F_x(t)$. Using again Fubini's Theorem we obtain

$$\begin{aligned}
\int t\,\mathrm{d}\bar{F}_a(t) &= \int\int_0^\infty \mathbb{1}_{[0,t]}(y)\,\mathrm{d}y\,\mathrm{d}\bar{F}_a(t) \\
&= \int_0^\infty \int \mathbb{1}_{[y,\infty)}(t)\,\mathrm{d}\bar{F}_a(t)\mathrm{d}y.
\end{aligned}$$

Since

$$\int \mathbb{1}_{[y,\infty)}(t)\,\mathrm{d}\bar{F}_a(t) = \lim_{z\to\infty}\bar{F}_a(z) - \bar{F}_a(y) \le 1 - \bar{F}_a(y)$$

it follows for $a = -\log m$

$$\begin{aligned}
\int t\,\mathrm{d}\bar{F}_a(t) &\le \int_0^\infty \sup_{x\le a}(1-F_x(y))\,\mathrm{d}y \\
&\le \int_0^\infty b\sup_{n\ge m}\underbrace{E\left[\frac{I_{n,1}}{n}\mathbb{1}_{\{-\log(I_{n,1}/n)>y\}}\right]}_{\le e^{-y}}\mathrm{d}y
\end{aligned}$$

$$\leq \int_0^\infty b e^{-y}\,\mathrm{d}y$$
$$< \infty.$$

□

Lemma 3.4. *The process $(S_t)_{t\in\mathbb{N}}$ fulfills the assumptions of Lemma 2.4.*

Proof. In the previous proof we have already shown that $(S_t)_{t\in\mathbb{N}}$ is an AR-process, which fulfills the integrability condition. The state space $\mathscr{E} = \{-\log n \mid n \in \mathbb{N}\} \cup \{1\}$ is discrete. It remains to show conditions (7). Let $x = -\log n$ and $y = -\log m$ with $m < n$. It is

$$d_{\mathrm{TV}}\left(P_x^{S_1}, P_y^{S_1}\right) = 2 - 2\sum_{z\in E} \min\{P_x(S_1 = z), P_y(S_1 = z)\}. \tag{10}$$

We will show that there exists $0 < \tilde{\alpha} < \tilde{\beta} < 1$ such that for n large enough

$$0 < \sum_{k=\lceil \tilde{\alpha} n\rceil + s_1}^{\lfloor \tilde{\beta} n\rfloor + s_1} \min\left\{ \int_0^1 \binom{\eta_l - 1}{k - s_1 - 1} z^{k-s_1}(1-z)^{\eta_l - k + s_1}\mathrm{d}P^V(z) \mid l = n, m \right\}. \tag{11}$$

For $k = cn + o(n)$ with $c \in (0,1)$ and $n \to \infty$ we have

$$\begin{aligned}
&P_x(S_1 = -\log k)\\
&= b\frac{k}{n}P(I_{n,1} = k) + \frac{s_0}{n}\mathbb{1}_{\{k = n - s_0\}}\\
&= b\frac{k}{k-s_1}\frac{\eta_n}{n}\int_0^1 \frac{k-s_1}{\eta_n} P(\mathrm{Bin}(\eta_n, z) = k - s_1)\mathrm{d}P^V(z) + \frac{s_0}{n}\mathbb{1}_{\{k=n-s_0\}}\\
&= (1 + o(1))b\int_0^1 \binom{\eta_n - 1}{k - s_1 - 1} z^{k-s_1}(1-z)^{\eta_n - k + s_1}\mathrm{d}P^V(z) + o(1).
\end{aligned}$$

Hence, inequality (11) and equation (10) will imply

$$d_{\mathrm{TV}}\left(P_x^{S_1}, P_y^{S_1}\right) < 2 - 2\varepsilon$$

for some $\varepsilon > 0$. The condition $|x - y| \leq K$ is equivalent to $m \geq e^{-K}n$.

By the general assumption, the distribution of V has a Lebesgue density f_V. Thus, there exists $\tilde{z} \in (0,1)$ with $f_V(\tilde{z}) > 0$. Theorem 3 in Section 1.7.2 of Evans and Gariepy (1992) (which is a Corollary from the Lebesgue-Besicovitch Differentiation Theorem) implies that we can find a non-empty interval $(\alpha, \beta) \subset (0,1)$ and $\varepsilon_1 > 0$ such that $\lambda(\{z \in (\alpha,\beta) \mid f_V(z) < \varepsilon_1\}) = 0$ with λ the Lebesgue measure. Now, we can choose some $\varepsilon_2 > 0$ and $K > 0$ with $\tilde{\alpha} := \alpha + \varepsilon_2 < e^{-K}(\beta - \varepsilon_2) =: \tilde{\beta}$. We will show that for n large enough, for all $k \in [\tilde{\alpha}n + s_1, \tilde{\beta}n + s_1] \cap \mathbb{N}$ and for all $l \in [e^{-K}n, n] \cap \mathbb{N}$ it holds

$$\int_0^1 \binom{\eta_l - 1}{k - s_1 - 1} z^{k-s_1}(1-z)^{\eta_l - k + s_1}\mathrm{d}P^V(z) \geq \frac{1}{2}\varepsilon_1 \frac{1}{n+1}.$$

First, we consider the function $g : z \mapsto z^{k-s_1}(1-z)^{\eta_l-k+s_1}$. Integration by parts yields

$$\int_0^1 z^{k-s_1}(1-z)^{\eta_l-k+s_1}\mathrm{d}z = \frac{k-s_1}{(\eta_l+1)\eta_l}\binom{\eta_l-1}{k-s_1-1}^{-1}. \tag{12}$$

For $k = c\eta_l + s_1$ the function g reaches its maximum at $\hat{z} = c$, is increasing on the interval $[0,c]$ and decreasing on $[c,1]$. Therefore, we have for any $\varepsilon_3 \in (0, c \wedge (1-c))$

$$\int_0^{c-\varepsilon_3} z^{c\eta_l}(1-z)^{(1-c)\eta_l}\mathrm{d}z \le \tilde{g}_c(\varepsilon_3)^{\eta_l}$$

and

$$\int_{c+\varepsilon_3}^1 z^{c\eta_l}(1-z)^{(1-c)\eta_l}\mathrm{d}z \le \tilde{g}_c(-\varepsilon_3)^{\eta_l}$$

where we set $\tilde{g}_c(\varepsilon_3) := (c-\varepsilon_3)^c(1-c+\varepsilon_3)^{(1-c)}$. Stirling's formula yields

$$\binom{\eta_l-1}{c\eta_l-1}^{-1} \sim \sqrt{2\pi c(1-c)}\frac{1}{c}((1-c)^{1-c}c^c)^{\eta_l}\sqrt{\eta_l} = \sqrt{2\pi\frac{1-c}{c}}\tilde{g}_c(0)^{\eta_l}\sqrt{\eta_l}.$$

Considering the derivative of $\tilde{g}_c$ in a neighborhood of 0, we obtain $\tilde{g}_c(x) < \tilde{g}(0) \le 1$ for all $x \neq 0$ with $|x|$ small enough. More precisely, for all $c \in [\tilde{\alpha}, \tilde{\beta}]$ and $\varepsilon_3 > 0$ small enough we have $\tilde{g}_c(\varepsilon_3)/\tilde{g}_c(0) \in (0, C)$ for some constant $C < 1$. Thus, for $\varepsilon_3 > 0$ small enough and l large enough we have

$$\int_0^{c-\varepsilon_3} z^{c\eta_l}(1-z)^{(1-c)\eta_l}\mathrm{d}z \le \frac{1}{4}\binom{\eta_l-1}{c\eta_l-1}^{-1}\frac{c}{\eta_l+1}$$

and

$$\int_{c+\varepsilon_3}^1 z^{c\eta_l}(1-z)^{(1-c)\eta_l}\mathrm{d}z \le \frac{1}{4}\binom{\eta_l-1}{c\eta_l-1}^{-1}\frac{c}{\eta_l+1}.$$

Together with (12), this implies for some $0 < \varepsilon_3 < \varepsilon_2$, l large enough and $c \in [\tilde{\alpha}, \tilde{\beta}]$ with $c\eta_l \in \mathbb{N}$

$$\int_{c-\varepsilon_3}^{c+\varepsilon_3}\binom{\eta_l-1}{c\eta_l-1}z^{c\eta_l}(1-z)^{(1-c)\eta_l}\mathrm{d}z \ge \frac{1}{2}\frac{c}{\eta_l+1}.$$

We obtain for any $k \in [\tilde{\alpha}n + s_1, \tilde{\beta}n + s_1] \cap \mathbb{N}$ and $l \in [e^{-K}n, n] \cap \mathbb{N}$ when n is large enough

$$\begin{aligned}
&\int_0^1 \binom{\eta_l-1}{k-s_1-1}z^{k-s_1}(1-z)^{\eta_l-k+s_1}\mathrm{d}P^V(z)\\
&\ge \varepsilon_1\int_\alpha^\beta \binom{\eta_l-1}{k-s_1-1}z^{k-s_1}(1-z)^{\eta_l-k+s_1}\mathrm{d}z\\
&\ge \frac{1}{2}\varepsilon_1\frac{\tilde{\alpha}}{\eta_l+1}\\
&\ge \frac{1}{2}\varepsilon_1\frac{\tilde{\alpha}}{n+1}.
\end{aligned}$$

This finally yields (11):

$$\begin{aligned}
&\sum_{k=\lceil\tilde{\alpha}n\rceil+s_1}^{\lfloor\tilde{\beta}n\rfloor+s_1} \min\left\{\int_0^1 \binom{\eta_l-1}{k-s_1-1} z^{k-s_1}(1-z)^{\eta_l-k+s_1} \mathrm{d}P^V(z) \mid l=n,m\right\} \\
&\quad \geq \frac{1}{2}\varepsilon_1\left(\tilde{\beta}-\tilde{\alpha}\right)\tilde{\alpha}+o(1) \\
&\quad > 0.
\end{aligned}$$

As in the proof of Lemma 3.3 we see that

$$\begin{aligned}
P_x(S_{\tau(y)}-y\leq K) &\geq \inf_{x<y} P_x(S_1-S_0\leq K) \\
&= \bar{F}_y(K) \\
&\xrightarrow{y\to-\infty} bE\left[V\mathbb{1}_{\{V\geq e^{-K}\}}\right].
\end{aligned}$$

Since $e^{-K}<1$ the general assumption $F_V(x)<1$ for all $x<1$ implies $bE\left[V\mathbb{1}_{\{V\geq e^{-K}\}}\right]>0$. This shows the second condition and the proof is finished. □

4 The internal path length

After these preliminaries, we are now able to prove Theorem 1.1. To show Theorem 1.1 we have to prove that the sequence

$$H_n := \frac{E[P_n]-\mu^{-1}n\log n}{n}$$

converges. The internal path length P_n suffices a recursive representation (see e.g. Neininger and Rüschendorf 1999, equation (50)) from where we get

$$E[P_n]=\sum_{k=0}^{n-s_0} bP(I_{n,1}=k)E[P_k]+n-s_0.$$

This recursion formula implies

$$H_n=\sum_{k=0}^{n-s_0}\nu_n(\{k\})H_k+t(n)-\frac{s_0}{n}H_{n-s_0}$$

with $t(n)=\frac{1}{n}(n-s_0-\mu^{-1}n\log n+b\mu^{-1}E[I_{n,1}\log I_{n,1}])$ and $\nu_n(\{k\})$ as in the previous section.
From the result about the mean of the depth in Devroye (1999) we know $H_n\leq C\log n$ for some constant $C>0$. Therefore, we have for any $\delta_1\in(0,1)$

$$\frac{s_0}{n}H_{n-s_0}\leq Cs_0\frac{\log n}{n}=O\left(\frac{1}{n^{\delta_1}}\right).$$

Furthermore, because of $n=bE[I_{n,1}]+s_0$, we have

$$t(n)=1-\frac{1}{E[V\log V]}E\left[\frac{I_{n,1}}{n}\log\frac{I_{n,1}}{n}\right]+O\left(\frac{1}{\sqrt{n}}\right).$$

The function $x \mapsto x \log x$ is Hölder continuous. Using this and considering the rate of convergence of $E[|\frac{I_{n,1}}{n} - V|]$ in Lemma 3.1 we obtain with Jensen's inequality $t(n) = O(n^{-\delta_2})$ for some $\delta_2 > 0$. Taking all this into account, we get

$$H_n = \sum_{k=0}^{n-s_0} \nu_n(\{k\}) H_k + r(n) \tag{13}$$

where $r(n) = O(n^{-\delta})$ for some $\delta \in (0,1]$.

Proof of Theorem 1.1. Equation (13) shows that the condition of Lemma 2.1 is fulfilled. Thus, we start with the representation of

$$H_n = \frac{E[P_n] - \mu^{-1} n \log n}{n}$$

from there and show that $(H_n)_{n\in\mathbb{N}}$ is a Cauchy sequence. Let $\varepsilon > 0$ be given.

For the second term in (5) we keep in mind that we have already shown $|r(n)| \le Cn^{-\delta}$ for some constant $0 < C < \infty$ and $\delta \in (0,1]$. We define $l : \mathbb{R} \to \mathbb{R}^+$ by $l(x) := \exp(\delta x)$. As in the proof of Theorem 4.2 in Bruhn (1996) we obtain with Lemma 2.5 for $n_1 \in \mathbb{N}$ with $-\log n_1 \le a_*$

$$\begin{aligned}
\left| E_{-\log n} \sum_{t=0}^{\sigma(n_1)-1} r(\exp(-S_t)) \right| &\le E_{-\log n} \sum_{t=0}^{\sigma(n_1)-1} C l(S_t) \\
&\le C\hat{u}(a_*) \sum_{n=-\infty}^{\lceil -\log n_1 \rceil} \sup_{t\in(n-1,n]} l(t) \\
&\le C\hat{u}(a_*) \int_{-\infty}^{\lceil -\log n_1 \rceil} l(t+1)\mathrm{d}t.
\end{aligned}$$

Since $\int_{-\infty}^{0} l(t)\mathrm{d}t < \infty$ we can choose $n_1 \in \mathbb{N}$ such that we have for all $n, m > n_1$,

$$\left| E_{-\log n} \left[\sum_{t=0}^{\sigma(n_1)-1} r(\exp(-S_t)) \right] \right| \le \frac{\varepsilon}{4}.$$

Considering the first term in (5), we set

$$a(n_1, n) := E_{-\log n} H_{\exp(-S_{\sigma(n_1)})}$$

and claim that there exists n_0 such that for all $n, m \ge n_0$ we have $|a(n_1,n) - a(n_1,m)| \le \varepsilon/2$. It is

$$\begin{aligned}
|a(n_1,n) - a(n_1,m)| &= \left| E_{-\log n} H_{\exp(-S_{\sigma(n_1)})} - E_{-\log m} H_{\exp(-S_{\sigma(n_1)})} \right| \\
&= \int H_{\exp(-x)} \left| P_{-\log n}^{S_{\sigma(n_1)}} - P_{-\log m}^{S_{\sigma(n_1)}} \right| (\mathrm{d}x) \\
&\le d_{\mathrm{TV}} \left(P_{-\log n}^{S_{\sigma(n_1)}}, P_{-\log m}^{S_{\sigma(n_1)}} \right) \sup_{k\in\{0,\dots,n_1\}} H_k.
\end{aligned}$$

Since n_1 is fixed we have $\sup_{k\in\{0,\dots,n_1\}} |H_k| \le C < \infty$ with some constant $C \in \mathbb{R}$. Lemma 2.4 in combination with Lemma 3.4 yields the claim.

Taking everything into account, we obtain for all $n, m \ge \max\{n_0, n_1\}$

$$\begin{aligned}\left|H_n - H_m\right| &\le \left|a(n_1,n) - a(n_1,m)\right| + \left|E_{-\log n}\left[\sum_{t=0}^{\sigma(n_1)-1} r(\exp(-S_t))\right]\right| \\ &\quad + \left|E_{-\log m}\left[\sum_{t=0}^{\sigma(n_1)-1} r(\exp(-S_t))\right]\right| \\ &\le \varepsilon.\end{aligned}$$

This shows that $(H_n)_{n\in\mathbb{N}}$ is a Cauchy sequence and thus it converges. □

Proof of Corollary 1.2. Parts a), c) and d) of Corollary 1.2 are immediate consequences of Theorem 1.1 and Neininger and Rüschendorf (1999, Theorem 5.1). To prove part b), we use that convergence with respect to the ℓ_2-metric implies convergence of the second moments. Thus, we obtain as consequence of part a) $\lim_{n\to\infty} E[X_n^2] = E[X^2]$. Using the distributional fixed point equation characterizing X, we have

$$\begin{aligned}E[X^2] &= E\left[\left(\sum_{k=1}^{b} V_k X^{(k)} + 1 + \frac{1}{\mu}\sum_{k=1}^{b} V_k \log V_k\right)^2\right] \\ &= \sum_{k=1}^{b} E[V_k^2] E\left[\left(X^{(k)}\right)^2\right] + E\left[1 + \frac{2}{\mu}\sum_{k=1}^{b} V_k \log V_k + \frac{1}{\mu^2}\left(\sum_{k=1}^{b} V_k \log V_k\right)^2\right]\end{aligned}$$

where we used the independence between $(V_1,\dots,V_b)$ and $(X^{(1)},\dots,X^{(b)})$ as well as the fact that $E[X^{(k)}] = 0$ for all k. Since $\mu = -bE[V_i \log V_i]$ for all $i = 1,\dots,b$ and $E[X^2] = E[(X^{(k)})^2] =: \sigma^2$ the claim follows. □

5 The Wiener index

We now turn to the investigation of the Wiener index. To handle the Wiener index similarly to the internal path length, we first need a recursion formula for it. The Wiener index is the sum of the distances between all unordered pairs of balls in the tree. Let $\Delta_{k,l}$ denote the distance between the balls k and l. Then we have

$$W_n = \sum_{k<l} \Delta_{k,l}.$$

Subdividing the sum into the sum for all pairs, where both balls are located in the same subtree, and the sum for all other pairs, we obtain

$$W_n = \sum_{i=1}^{b} W_{I_{n,i}}^{(i)} + \sum_{i<j} \sum_{l\in T_{n,j}} \sum_{k\in T_{n,i}} \Delta_{k,l}$$

where $W^{(i)}_{I_{n,i}}$ denotes the Wiener index of the i-th subtree $T_{n,i}$ being of size $I_{n,i}$. For $k \in T_{n,i}$ and $l \in T_{n,j}$ with $i \neq j$ it is $\Delta_{k,l} = D^{(i)}_k + 1 + D^{(j)}_l + 1$ where $D^{(i)}_k$ is the depth of the ball k with respect to the subtree $T_{n,i}$. By symmetry of $\Delta_{k,l}$ we can sum up only the first part $D^{(i)}_k + 1$ but for all ordered pairs of balls and we obtain

$$\sum_{i<j} \sum_{l \in T_{n,j}} \sum_{k \in T_{n,i}} \Delta_{k,l} = \sum_{i \neq j} \sum_{l \in T_{n,j}} \sum_{k \in T_{n,i}} (D^{(i)}_k + 1).$$

The summation over $k \in T_{n,i}$ yields

$$\sum_{i \neq j} \sum_{l \in T_{n,j}} \sum_{k \in T_{n,i}} (D^{(i)}_k + 1) = \sum_{i \neq j} \sum_{l \in T_{n,j}} (P^{(i)}_{I_{n,i}} + I_{n,i})$$

where $P^{(i)}_{I_{n,i}}$ denotes the internal path length of the i-th subtree $T_{n,i}$. Since there are all together $n - I_{n,i}$ balls not lying in $T_{n,i}$, we finally obtain the recursion formula for the Wiener index of the random split tree with n balls:

$$W_n = \sum_{i=1}^{b} \left[W^{(i)}_{I_{n,i}} + (n - I_{n,i}) P^{(i)}_{I_{n,i}} + I_{n,i}(n - I_{n,i}) \right]. \tag{14}$$

Proof of Theorem 1.4. Starting from equation (14) and taking the expectation yields

$$E[W_n] = b \sum_{k=0}^{n-s_0} P(I_{n,1} = k) \left(E[W_k] + (n-k)E[P_k] + nk - k^2 \right) \tag{15}$$

because all subtrees are identically distributed. Theorem 1.1 implies $E[P_k] = \frac{1}{\mu} k \log k + c_p k + o(k)$. Substituting this in (15) yields with $E[I_{n,1}] = n/b + o(n)$,

$$\begin{aligned} E[W_n] = b \sum_{k=0}^{n-s_0} P(I_{n,1} = k) E[W_k] + \frac{1}{\mu} b \left(nE[I_{n,1} \log I_{n,1}] - E[I^2_{n,1} \log I_{n,1}] \right) \\ + (c_p + 1)n^2 - (c_p + 1) b E[I^2_{n,1}] + o(n^2). \end{aligned} \tag{16}$$

Substituting the results from Lemma 3.2 in (16) provides

$$\begin{aligned} E[W_n] = \sum_{k=0}^{n-s_0} b P(I_{n,1} = k) E[W_k] + \frac{1}{\mu}(1 - bE[V^2]) n^2 \log n \\ - \left(\frac{b}{\mu} E[V^2 \log V] + bE[V^2] - c_p(1 - bE[V^2]) \right) n^2 + o(n^2). \end{aligned} \tag{17}$$

We set

$$H_n := \frac{E[W_n] - \frac{1}{\mu} n^2 \log n}{n}.$$

To prove Theorem 1.4 it suffices to show that for each $\varepsilon > 0$ there exists a constant $c \in \mathbb{R}$ and $n_0 \in \mathbb{N}$ such that for all $n \geq n_0$

$$\frac{H_n}{n} \in (c - \varepsilon, c + \varepsilon).$$

So, let $\varepsilon > 0$ be given. Substituting H_n in (17) and using Lemma 3.2 yields

$$H_n = \sum_{k=0}^{n-s_0} \nu_n(\{k\}) H_k + r(n)$$

with

$$r(n) := -\left(bE[V^2] - c_p(1 - bE[V^2])\right) n + o(n).$$

We set $\tilde{d} := -bE[V^2] + c_p(1 - bE[V^2])$. As in the proof of Theorem 1.1 the conditions of Lemma 2.1 are fulfilled and we have the representation

$$H_n = E_{-\log n} H_{\exp(-S_{\sigma(n_1)})} + E_{-\log n} \sum_{t=0}^{\sigma(n_1)-1} r(\exp(-S_t)). \tag{18}$$

We start again with the second term and split it in the following way

$$\begin{aligned} E_{-\log n} \sum_{t=0}^{\sigma(n_1)-1} r(\exp(-S_t)) = E_{-\log n} &\sum_{t=0}^{\tau(-\log n+d)} r(\exp(-S_t)) \\ &+ E_{-\log n} \sum_{t=\tau(-\log n+d)+1}^{\sigma(n_1)-1} r(\exp(-S_t)). \end{aligned}$$

For the second summand we obtain by Lemma 2.5 with $l(x) := \tilde{d}\exp(-x)$ and n_1 large enough such that $-\log n_1 \le a_*$

$$\begin{aligned} 0 \le \left| E_{-\log n} \sum_{t=\tau(-\log n+d)+1}^{\sigma(n_1)-1} r(\exp(-S_t)) \right| &\le \hat{u}(a_*) \sum_{n=\lfloor -\log n+d \rfloor}^{\lceil -\log n_1 \rceil} \sup_{t\in(n-1,n]} |\tilde{d}| e^{-t} \\ &\le C \int_{-\log n+d-3}^{-\log n_1} e^{-x}\,\mathrm{d}x \\ &\le Cne^{-d+3} \end{aligned}$$

with some constant C. We choose d large enough, such that $Ce^{-d+3} < \varepsilon/3$. For this d Corollary 2.7 yields $\hat{n}_0 \in \mathbb{N}$ such that for all $n \ge \hat{n}_0$

$$\frac{1}{n} E_{-\log n} \sum_{t=0}^{\tau(-\log n+d)} r(\exp(-S_t)) \in \left(c - \frac{\varepsilon}{3}, c + \frac{\varepsilon}{3}\right) \tag{19}$$

for some constant c. As in the proof of Theorem 1.1 the first summand in (18) is a Cauchy sequence, i.e. there exists $\tilde{n}_0 \in \mathbb{N}$ such that for all $n \ge \tilde{n}_0$ we have

$$\left| \frac{1}{n} E_{-\log n}[H_{\exp(S_{\sigma(n_1)})}] \right| < \frac{\varepsilon}{3}.$$

Altogether, we have seen that for $n_1 \in \mathbb{N}$ with $-\log n_1 \le a_*$ there exists $n_0 \in \mathbb{N}$ such that for all $n \ge n_0$ we have

$$\frac{H_n}{n} = \frac{1}{n} E_{-\log n} H_{\exp(-S_{\sigma(n_1)})} + \frac{1}{n} E_{-\log n} \sum_{t=0}^{\sigma(n_1)-1} r(\exp(-S_t))$$

$$\in (c-\varepsilon, c+\varepsilon)$$

with the constant c in (19). Thus, the claim follows. □

Proof of Theorem 1.5. We define

$$w_n := E[W_n] = \frac{1}{\mu} n^2 \log n + c_w n^2 + o(n^2),$$
$$p_n := E[P_n] = \frac{1}{\mu} n \log n + c_p n + o(n)$$

and

$$X_n := \left(\frac{W_n - w_n}{n^2}, \frac{P_n - p_n}{n} \right)^T.$$

For $i \in \{1, \dots, b\}$ let $X_n^{(i)}$ be an independent copy of X_n. Since the subtrees of the random split tree are independent conditioned upon there sizes, we obtain from (14) for the standardized vector X_n the following recursion formula

$$X_n \stackrel{d}{=} \sum_{i=1}^{b} A_i^{(n)} X_{I_{n,i}}^{(i)} + b^{(n)}$$

with

$$A_i^{(n)} := \begin{bmatrix} \frac{1}{n^2} & 0 \\ 0 & \frac{1}{n} \end{bmatrix} \begin{bmatrix} 1 & n - I_{n,i} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} I_{n,i}^2 & 0 \\ 0 & I_{n,i} \end{bmatrix} = \begin{bmatrix} \frac{I_{n,i}^2}{n^2} & \frac{I_{n,i}(n-I_{n,i})}{n^2} \\ 0 & \frac{I_{n,i}}{n} \end{bmatrix}$$

and $b^{(n)} = \left(b_1^{(n)}, b_2^{(n)} \right)^T$ where

$$b_1^{(n)} = \frac{1}{n^2} \left\{ \sum_{i=1}^{b} I_{n,i} \left(n - I_{n,i} \right) - \frac{1}{\mu} n^2 \log n - c_w n^2 + o(n^2) \right.$$
$$\left. + \sum_{i=1}^{b} w_{I_{n,i}} + n \sum_{i=1}^{b} p_{I_{n,i}} - \sum_{i=1}^{b} I_{n,i} \, p_{I_{n,i}} \right\}$$

and

$$b_2^{(n)} := 1 - \frac{1}{\mu} \log n - c_p + o(1) + \frac{1}{n} \sum_{i=1}^{b} p_{I_{n,i}} + o(1).$$

Using $\sum_{i=1}^{b} I_{n,i} = n - s_0$ it follows

$$n \sum_{i=1}^{b} p_{I_{n,i}} - \frac{1}{\mu} n^2 \log n = n \frac{1}{\mu} \sum_{i=1}^{b} I_{n,i} \log \frac{I_{n,i}}{n} + c_p n (n - s_0) + o(n^2)$$

and

$$\sum_{i=1}^{b} w_{I_{n,i}} - \sum_{i=1}^{b} I_{n,i} \, p_{I_{n,i}} = (c_w - c_p) \sum_{i=1}^{b} I_{n,i}^2 + o(n^2).$$

This yields with $I_{n,i} = o(n^2)$

$$b_1^{(n)} = \frac{1}{\mu}\sum_{i=1}^{b}\frac{I_{n,i}}{n}\log\frac{I_{n,i}}{n} + (1 + c_p - c_w)\left(1 - \sum_{i=1}^{b}\frac{I_{n,i}^2}{n^2}\right) + o(1). \tag{20}$$

By similar arguments we have

$$b_2^{(n)} = \frac{1}{\mu}\sum_{i=1}^{b}\frac{I_{n,i}}{n}\log\frac{I_{n,i}}{n} + 1 + o(1). \tag{21}$$

In order to use the contraction method as in Neininger (2001, Theorem 4.1) it suffices to show that for $n \to \infty$

$$\left(A_1^{(n)}, \ldots, A_b^{(n)}, b^{(n)}\right) \xrightarrow{\ell_2} \left(A_1^*, \ldots, A_b^*, b^*\right), \tag{22}$$

$$E\left[\mathbb{1}_{\{I_{n,i}\le l\}\cup\{I_{n,i}=n\}}\left\|(A_i^{(n)})^T A_i^{(n)}\right\|_{\mathrm{op}}\right] \to 0 \tag{23}$$

for all $l \in \mathbb{N}$ and

$$\sum_{i=1}^{b} E\left\|(A_i^*)^T A_i^*\right\|_{\mathrm{op}} < 1 \tag{24}$$

where $\|\cdot\|_{\mathrm{op}}$ is the operator norm.

By Lemma 3.1 we know that I_n/n converges in probability to $V := (V_1, \ldots, V_b)$, which is the splitting vector. By equations (20) and (21) we have $b^{(n)} \to b^*$ in probability as $n \to \infty$ with

$$b^* = \frac{1}{\mu}\sum_{i=1}^{b} V_i \log V_i \begin{pmatrix}1\\1\end{pmatrix} + \begin{pmatrix}(1 + c_p - c_w)\left(1 - \sum_{i=1}^{b} V_i^2\right)\\ 1\end{pmatrix}.$$

By the boundedness of the function $x \mapsto x\log x$ on $[0,1]$ and as $I_{n,i}/n \in [0,1]$ there exists a constant C such that

$$\left|b_1^{(n)}\right| \le C \qquad \text{and} \qquad \left|b_2^{(n)}\right| \le C.$$

Thus, we get the uniform integrability of $(b_1^{(n)})^2$ and $(b_2^{(n)})^2$ and consequently the convergence of $b^{(n)}$ with respect to the ℓ_2-metric. Similar arguments yield the convergence of $A_i^{(n)}$ with respect to the ℓ_2-metric to

$$A_i^* = \begin{bmatrix} V_i^2 & V_i(1 - V_i)\\ 0 & V_i\end{bmatrix}.$$

This shows condition (22).

Condition (23) follows from the deterministic boundedness of $\|A_i^{(n)}\|_{\mathrm{op}}$ and from the fact that

$$\begin{aligned}&\lim_{n\to\infty} P\left(\{I_{n,i} \le l\} \cup \{I_{n,i} = n\}\right)\\ &= \lim_{n\to\infty}\int_0^1 P(\mathrm{Bin}(\eta_n, x) \le l - s_1)\mathrm{d}P^V(x)\end{aligned}$$

$$
\begin{aligned}
&\leq \lim_{n\to\infty} P\left(V \leq \left((l-s_1)/\eta_n\right)^{\frac{1}{3}}\right) \\
&\quad + \lim_{n\to\infty} \int_{\left(\frac{l-s_1}{\eta_n}\right)^{\frac{1}{3}}}^{1} \exp\left(-\frac{1}{4}\eta_n^{\frac{1}{3}}(l-s_1)^{\frac{2}{3}}\left(1-\left(\frac{l-s_1}{\eta_n}\right)^{\frac{2}{3}}\right)^2\right) \mathrm{d}P^V(x) \\
&= 0
\end{aligned}
$$

where we used Bernstein's inequality.

It remains to show (24). We observe that the eigenvalues of A_i^* are V_i^2 and V_i. Since V_i is bounded by 1 and non-negative, it is $\|A_i^*\|_{\mathrm{op}} = \|(A_i^*)^T\|_{\mathrm{op}} = V_i$. We use the inequality $\|AB\|_{\mathrm{op}} \leq \|A\|_{\mathrm{op}}\|B\|_{\mathrm{op}}$. With $\sum_{i=1}^b V_i^2 < 1$ almost surely we finally conclude

$$
E\left[\sum_{i=1}^{b} \left\|(A_i^*)^T A_i^*\right\|_{\mathrm{op}}\right] \leq E\left[\sum_{i=1}^{b} V_i^2\right] < 1. \tag{25}
$$

The claim for the asymptotic behavior of the variance of W_n follows directly from the first part, since convergence with respect to the ℓ_2-metric implies convergence of the second moments. □

References

T. Ali Khan and R. Neininger. Tail bounds for the Wiener index of random trees. In *2007 Conference on Analysis of Algorithms, AofA 07*, Discrete Math. Theor. Comput. Sci. Proc., AH, pages 279–289. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2007. MR2509528

F. Bergeron, P. Flajolet, and B. Salvy. Varieties of increasing trees. In *CAAP '92 (Rennes, 1992)*, volume 581 of *Lecture Notes in Comput. Sci.*, pages 24–48. Springer, Berlin, 1992. MR1251994

P. J. Bickel and D. A. Freedman. Some asymptotic theory for the bootstrap. *Ann. Statist.*, 9(6): 1196–1217, 1981. MR0630103

N. Broutin and C. Holmgren. The total path length of split trees. preprint. 2011. `http://arxiv.org/abs/1102.2541`

V. Bruhn. *Eine Methode zur asymptotischen Behandlung einer Klasse von Rekursionsgleichungen mit einer Anwendung in der stochastischen Analyse des Quicksort-Algorithmus*. PhD thesis, University of Kiel, Germany, 1996.

H.-H. Chern and H.-K. Hwang. Transitional behaviors of the average cost of Quicksort with median-of-$(2t+1)$. *Algorithmica*, 29(1–2):44–69, 2001. MR1887298

L. Devroye. Universal limit laws for depths in random trees. *SIAM J. Comput.*, 28(2):409–432 (electronic), 1999. MR1634354

R. P. Dobrow and J. A. Fill. Total path length for random recursive trees. *Combin. Probab. Comput.*, 8(4):317–333, 1999. MR1723646

L. C. Evans and R. F. Gariepy. *Measure theory and fine properties of functions*. CRC Press, Boca Raton, FL, 1992. MR1158660

J. A. Fill and S. Janson. Quicksort asymptotics. *J. Algorithms*, 44(1):4–28, 2002. MR1932675

P. Flajolet, G. Labelle, L. Laforest, and B. Salvy. Hypergeometrics and the cost structure of quadtrees. *Random Structures Algorithms*, 7(2):117–144, 1995. MR1369059

D. Griffeath. A maximal coupling for Markov chains. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete*, 31:95–106, 1974/75. MR0370771

A. Gut. *Stopped random walks*. Springer-Verlag, New York, 1988. MR0916870

C. Holmgren. Novel characteristics of split trees by use of renewal theory. preprint. 2010. MR2607374

S. Janson. The Wiener index of simply generated random trees. *Random Structures Algorithms*, 22 (4):337–358, 2003. MR1980963

H. M. Mahmoud. On the average internal path length of m-ary search trees. *Acta Inform.*, 23(1): 111–117, 1986. MR0845626

H. M. Mahmoud and B. Pittel. Analysis of the space of search trees under the random insertion algorithm. *J. Algorithms*, 10(1):52–75, 1989. MR0987097

C. J. H. McDiarmid and R. B. Hayward. Large deviations for Quicksort. *J. Algorithms*, 21(3): 476–507, 1996. MR1417660

G. O. Munsonius and L. Rüschendorf. Limit theorems for depths and distances in weighted random b-ary recursive trees. preprint. 2010.

R. Neininger. On a multivariate contraction method for random recursive structures with applications to Quicksort. *Random Structures Algorithms*, 19(3–4):498–524, 2001. MR1871564

R. Neininger. The Wiener index of random trees. *Combin. Probab. Comput.*, 11(6):587–597, 2002. MR1940122

R. Neininger and L. Rüschendorf. On the internal path length of d-dimensional quad trees. *Random Structures Algorithms*, 15(1):25–41, 1999. MR1698407

M. Régnier. A limiting distribution for quicksort. *RAIRO Inform. Théor. Appl.*, 23(3):335–343, 1989. MR1020478

U. Rösler. A limit theorem for "Quicksort". *RAIRO Inform. Théor. Appl.*, 25(1):85–100, 1991.

U. Rösler. On the analysis of stochastic divide and conquer algorithms. *Algorithmica*, 29(1–2): 238–261, 2001. MR1887306

A Proof of Lemma 2.4

We give the essential parts of Rösler (1991) which prove Lemma 2.4.

Proof of Lemma 2.4. Let $a \in \mathbb{R}_-$. We use the notation

$$\Delta(a) := \lim_{x_0 \to -\infty} \sup_{x,y \le x_0} d_{\mathrm{TV}}\left(P_x^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right).$$

Since the function

$$x_0 \quad \mapsto \quad \sup_{x,y \le x_0} d_{\mathrm{TV}}\left(P_x^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right)$$

is increasing and non-negative, the limit for $x_0 \to -\infty$ exists. We will show that $\Delta(a) \le (1 - \tilde{\varepsilon})\Delta(a) + \delta$ for some $\tilde{\varepsilon} > 0$ and all $\delta > 0$. Then the claim follows.

Let $\delta > 0$ be an arbitrary number. Since the process S fulfills the integrability condition and $S_{\tau(y)} - y \le S_{\tau(y)} - S_{\tau(y)-1}$, there exists $x_1 \in \mathbb{R}_-$ such that for all $x < y < x_1$

$$E_x[S_{\tau(y)} - y] \;\le\; \int z \, \mathrm{d}\bar{F}_y(z) \;\le\; C < \infty$$

for some constant C. Thus, there exists $K_1 \ge K$ such that for all $y < x_1$

$$P_x(S_{\tau(y)} - y > K_1) \le \frac{E_x[S_{\tau(y)} - y]}{K_1} \le \frac{\delta}{4}. \tag{26}$$

Furthermore, we have for this K_1

$$\sup_{y+K \le z \le y+K_1} d_{\mathrm{TV}}\left(P_z^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right) \le \sup_{u,v \le y+K_1} d_{\mathrm{TV}}\left(P_u^{S_{\tau(a)}}, P_v^{S_{\tau(a)}}\right). \tag{27}$$

The distribution of the Markov chain S on the state space $\mathscr{E}$ is given by the kernel

$$\kappa(x,A) := P(S_{t+1} \in A \mid S_t = x) \qquad \text{for all } t \in \mathbb{N}_0 \text{ and } A \subset \mathscr{E}.$$

Let $S^{(a)}$ be the process S stopped at the moment when it exceeds $a \in \mathscr{E}$. The kernel κ_a corresponding to the process $S^{(a)}$ is then given by $\kappa_a(x,A) = \kappa(x,A)$ for $x \le a$ and $\kappa_a(x,A) := \mathbb{1}_A(x)$ for $x > a$ and for all $A \subset \mathscr{E}$.

Let $D := \{(x,x) \mid x \in \mathscr{E}\}$ denote the diagonal in $\mathscr{E}^2$. We define a kernel ϱ on $\mathscr{E}^2$ by the so called Wasserstein coupling (see e.g. Griffeath 1974/75), i.e. for $(x,y),(u,v) \in \mathscr{E}^2$ it is

$$\varrho((x,y),(u,v)) := \begin{cases} \min\{\kappa_a(x,u), \kappa_a(y,v)\}, & \text{if } u = v \\ \frac{(\kappa_a(x,u) - \kappa_a(y,u))^+ (\kappa_a(y,v) - \kappa_a(x,v))^+}{1 - \alpha(x,y)}, & \text{if } u \ne v \end{cases}$$

where $\alpha(x,y) := \sum_{z \in \mathscr{E}} \min\{\kappa_a(x,z), \kappa_a(y,z)\}$ and $r^+ = \max\{r, 0\}$ denotes the positive part of a real number r. Then the following properties hold:

1. $\varrho((x,y), A \times \mathscr{E}) = \kappa_a(x,A)$ and $\varrho((x,y), \mathscr{E} \times A) = \kappa_a(y,A)$ for all $x, y \in \mathscr{E}$ and $A \subset \mathscr{E}$
2. $\varrho((x,x), D) = 1$ for all $x \in \mathscr{E}$ and
3. $\varrho((x,y), D^c) \le 1 - \varepsilon$ for all $x, y \in \mathscr{E}$ with $|x - y| \le K$ and $x, y < x_0$.

The property c) follows from the assumption (7) and the fact that

$$\begin{aligned} d_{\mathrm{TV}}\left(P_x^{S_1}, P_y^{S_1}\right) &= \sum_{z\in E}\left|\kappa_a(x,z)-\kappa_a(y,z)\right| \\ &= 2\left(1-\sum_{z\in E}\min\{\kappa_a(x,z),\kappa_a(y,z)\}\right). \end{aligned}$$

For $(x,y)\in\mathscr{E}^2$ let $Z^{(x,y)}=(U^{(x,y)},V^{(x,y)})$ be the Markov chain generated by the kernel ϱ which starts in (x,y). We define the stopping time

$$\theta(a) := \inf\{t \mid Z_t^{(x,y)}\in(a,\infty)\times(a,\infty)\}.$$

Using this coupling we obtain for any $K_2>0$ and $z,y<a$

$$\begin{aligned} &d_{\mathrm{TV}}\left(P_z^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right) \\ &= \sum_{w\in\mathscr{E}}\left|P_z(S_{\tau(a)}=w)-P_y(S_{\tau(a)}=w)\right| \\ &= \sum_{w\in\mathscr{E}}\left|P\left(U_{\theta(a)}^{(z,y)}=w\right)-P\left(V_{\theta(a)}^{(z,y)}=w\right)\right| \\ &= \sum_{(u,v)\in\mathscr{E}^2}\sum_{w\in\mathscr{E}}P\left(Z_1^{(z,y)}=(u,v)\right) \\ &\qquad\times\Bigg|\underbrace{P\left(U_{\theta(a)}^{(z,y)}=w\mid Z_1^{(z,y)}=(u,v)\right)}_{=P_u(S_{\tau(a)}=w)}-\underbrace{P\left(V_{\theta(a)}^{(z,y)}=w\mid Z_1^{(z,y)}=(u,v)\right)}_{=P_v(S_{\tau(a)}=w)}\Bigg| \\ &\le \sup_{u,v\le y+K_2} d_{\mathrm{TV}}\left(P_u^{S_{\tau(a)}}, P_v^{S_{\tau(a)}}\right)\varrho((z,y),D^c)+2P\left(Z_1^{(z,y)}\notin(-\infty,y+K_2]^2\right). \end{aligned} \tag{28}$$

In the last step we used that $P_u(S_{\tau(a)}=w)-P_v(S_{\tau(a)}=w)=0$ for $u=v$. As seen in equation (26) and using property a) of the coupling, there exists by the integrability condition $K_2>K$ such that for all $y<x_1-K$ and $y<z<y+K$

$$\begin{aligned} P\left(Z_1^{(z,y)}\notin(-\infty,y+K_2]^2\right) &\le \kappa_a\left(z,(-\infty,y+K_2]^c\right)+\kappa_a\left(y,(-\infty,y+K_2]^c\right) \\ &\le \frac{\delta}{4}. \end{aligned} \tag{29}$$

After these preliminaries, we now turn to $\Delta(a)$. It is for $x<y<a-K$

$$\begin{aligned} d_{\mathrm{TV}}\left(P_x^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right) &= \int d_{\mathrm{TV}}\left(P_z^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right)\mathrm{d}P_x^{S_{\tau(y)}}(z) \\ &= \int_{[y,y+K]} d_{\mathrm{TV}}\left(P_z^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right)\mathrm{d}P_x^{S_{\tau(y)}}(z) \\ &\quad+\int_{(y+K,y+K_1]} d_{\mathrm{TV}}\left(P_z^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right)\mathrm{d}P_x^{S_{\tau(y)}}(z) \end{aligned}$$

$$
\begin{aligned}
&+ \int_{(y+K_1,\infty)} d_{\mathrm{TV}}\left(P_z^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right) \mathrm{d}P_x^{S_{\tau(y)}}(z) \\
&\leq P_x(S_{\tau(y)} - y \leq K) \sup_{y \leq z \leq y+K} d_{\mathrm{TV}}\left(P_z^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right) \\
&\quad + P_x(S_{\tau(y)} - y > K) \sup_{y+K \leq z \leq y+K_1} d_{\mathrm{TV}}\left(P_z^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right) \\
&\quad + 2P_x(S_{\tau(y)} - y \geq K_1).
\end{aligned}
$$

With the results in (26), (27), (28) and (29) as well as property c) of the kernel ϱ this finally yields

$$
\begin{aligned}
\Delta(a) &\leq \lim_{x_0 \to -\infty} \sup_{x < y \leq x_0} \Bigg[P_x(S_{\tau(y)} - y \leq K) \sup_{u,v \leq y+K_2} d_{\mathrm{TV}}\left(P_u^{S_{\tau(a)}}, P_v^{S_{\tau(a)}}\right)(1-\varepsilon) \\
&\quad + P_x(S_{\tau(y)} - y > K) \sup_{u,v \leq y+K_1} d_{\mathrm{TV}}\left(P_z^{S_{\tau(a)}}, P_y^{S_{\tau(a)}}\right) \\
&\quad + 2P\left(Z_1^{(z,y)} \notin (-\infty, y+K_2]^2\right)\Bigg] + \frac{\delta}{2} \\
&\leq \Delta(a) \lim_{x_0 \to -\infty} \sup_{x < y \leq x_0} \left(1 - \varepsilon P_x(S_{\tau(y)} - y \leq K)\right) + \delta \\
&\leq (1 - \tilde{\varepsilon})\,\Delta(a) + \delta
\end{aligned}
$$

where $\tilde{\varepsilon} = \varepsilon \lim_{x_0 \to -\infty} \inf_{x < y \leq x_0} P_x(S_{\tau(y)} - y \leq K) > 0$. □

B Proofs from Bruhn (1996)

Proof of Lemma 2.1. For $n \leq n_1$ the claim follows immediately since $\sigma(n_1) = 0$. For $n > n_1$ equation (5) follows by induction on n. It is with $H_{1/e} := H_0$

$$
\begin{aligned}
H_{n+1} &= \sum_{k=0}^{n} \nu_{n+1}(\{k\}) H_k + r(n+1) \\
&= \sum_{k=1}^{n} P_{-\log(n+1)}(S_1 = -\log k)\, E_{-\log k}\left[H_{\exp(-S_{\sigma(n_1)})} + \sum_{t=0}^{\sigma(n_1)-1} r(\exp(-S_t)) \right] \\
&\quad + E_{-\log(n+1)}[r(\exp(-S_0))] + P_{-\log(n+1)}(S_1 = 1) E_1[H_{\exp(-S_{\sigma(n_1)})}] \\
&= E_{-\log(n+1)} H_{\exp(-S_{\sigma(n_1)})} + E_{-\log(n+1)}\left[\sum_{t=0}^{\sigma(n_1)-1} r(\exp(-S_t)) \right]
\end{aligned}
$$

where we use the Kolmogorov-Chapman equation for Markov chains in the last step. □

Proof of Lemma 2.5. We use the notation from Section 2 and define for $x \in \mathbb{R}_-$ the function u_x by

$$
u_x(a) := E_x[|\{t : S_t \in (a, a+1]\}|].
$$

By the monotone convergence theorem we have $\lim_{a\to-\infty} E[\underline{Y}_t^{(a)}] = E[\tilde{Y}_t] > 0$. Thus, there exists $a_* \in \mathbb{R}$ such that for all $a < a_*$ it is $E[\underline{Y}_t^{(a)}] > 0$. For $x, n, a < a_*$ and $k \in \mathbb{N}$ it holds

$$\begin{aligned}
P_x(|\{t : S_t \in (n-1,n]\}| \geq k) &= \int_{(n-1,n]} P_y(S_{k-1} \leq n)\, \mathrm{d}P_x^{S_{\tau(n-1)}}(y) \\
&\leq \int_{(n-1,n]} P_y(\underline{S}_{k-1}^{(a)} \leq n)\, \mathrm{d}P_x^{S_{\tau(n-1)}}(y) \\
&\leq \int_{(n-1,n]} P_0(\underline{S}_{k-1}^{(a)} \leq 1)\, \mathrm{d}P_x^{S_{\tau(n-1)}}(y) \\
&\leq P_0(\underline{S}_{k-1}^{(a)} \leq 1) \\
&= P_0(|\{t : \underline{S}_t^{(a)} \in [0,1]\}| \geq k).
\end{aligned}$$

Thus, we have

$$\begin{aligned}
u_x(n-1) &\leq \sum_{k=1}^{\infty} P_0(|\{t : \underline{S}_t^{(a)} \in [0,1]\}| \geq k) \\
&= E_0[|\{t : \underline{S}_t^{(a)} \in [0,1]\}|] \\
&=: \hat{u}(a).
\end{aligned}$$

Since it is $E[\underline{Y}_t^{(a)}] > 0$ the elementary renewal theorem (see e.g. Gut 1988, Section II.4) provides $\hat{u}(a) < \infty$. Furthermore, the function $a \mapsto \hat{u}(a)$ is decreasing as $a \to -\infty$, i.e. $\hat{u}(a) \leq \hat{u}(a_*)$ for all $a < a_*$.

So we finally obtain for a function $l : \mathbb{R} \to \mathbb{R}_+$, $y, z \in \mathbb{R}$ and $x \in \mathscr{E}$ with $x < y < z < a_*$

$$\begin{aligned}
E_x\left[\sum_{t=\tau(y)}^{\tau(z)-1} l(S_t)\right] &\leq \sum_{n=\lceil y \rceil}^{\lceil z \rceil} u_x(n-1) \sup_{t\in(n-1,n]} l(t) \\
&\leq \hat{u}(a_*) \sum_{n=\lceil y \rceil}^{\lceil z \rceil} \sup_{t\in(n-1,n]} l(t).
\end{aligned}$$

□