Electronic Journal of Probability

Vol. 10 (2005), Paper no. 12, pages 420-435.

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

# Martingales and Profile of Binary Search Trees

B. Chauvin, T. Klein, J-F. Marckert , A. Rouault,

Université de Versailles
45 Avenue des Etats Unis
78035 Versailles Cedex
France

**Abstract**

We are interested in the asymptotic analysis of the binary search tree (BST) under the random permutation model. Via an embedding in a continuous time model, we get new results, in particular the asymptotic behavior of the profile.

**Keywords.** Binary search tree, branching random walk, convergence of martingales **AMS 2000**

**Subject Classification.** 60J25, 60J80, 68W40, 60C05, 60G42, 60G44.

Submitted to EJP on October 19, 2003. Final version accepted on January 20, 2005.

# 1 Introduction

For a convenient definition of trees we are going to work with, let us first define

$$\mathbb{U} = \{\emptyset\} \cup \bigcup_{n \geq 1} \{0,1\}^n$$

the set of finite words on the alphabet $\{0,1\}$ (with $\emptyset$ for the empty word). For $u$ and $v$ in $\mathbb{U}$, denote by $uv$ the concatenation of the word $u$ with the word $v$ (by convention we set, for any $u \in \mathbb{U}$, $\emptyset u = u$). If $v \neq \emptyset$, we say that $uv$ is a descendant of $u$ and $u$ is an ancestor of $uv$. Moreover $u0$ (resp. $u1$) is called left (resp. right) child of $u$.

A *complete binary tree* $T$ is a finite subset of $\mathbb{U}$ such that

$$\begin{cases} \emptyset \in T \\ \text{if } uv \in T \text{ then } u \in T\,, \\ u1 \in T \Leftrightarrow u0 \in T\,. \end{cases}$$

The elements of $T$ are called *nodes*, and $\emptyset$ is called the *root* ; $|u|$, the number of letters in $u$, is the *depth* of $u$ (with $|\emptyset| = 0$). Write **BinTree** for the set of complete binary trees.

A tree $T \in$ **BinTree** can be described by giving the set $\partial T$ of its *leaves*, that is, the nodes that are in $T$ but with no descendant in $T$. The nodes of $T \backslash \partial T$ are called *internal* nodes.

### Models

This paper deals with two classical models of binary tree processes: the binary search tree process and the Yule tree process.

□ A binary search tree (BST) is a structure used in computer science to store totally ordered data (the monograph of Mahmoud [22] gives an overview of the state of the art). Each unit of time, an item is inserted in the tree, inducing a sequence of labeled binary trees. Under suitable assumptions on the data (the so-called random permutation model), the process of constructed unlabeled binary trees $(\mathcal{T}_n)_{n\geq 0}$ is a Markov chain on **BinTree** that has the following representation:
• $\mathcal{T}_0 = \{\emptyset\}$
• $\mathcal{T}_{n+1} = \mathcal{T}_n \cup \{D_n0, D_n1\}$ where $D_n$, the random node where the $n+1$-st key is inserted, has the following distribution

$$P(D_n = u \,|\, \mathcal{T}_n) = (n+1)^{-1}, \qquad u \in \partial\mathcal{T}_n\,.$$

In other words, each unit of time, a leaf is chosen equally likely among the leaves of the current tree, and it is replaced by an internal node with two children.

□ The Yule tree process $(\mathbb{T}_t)_{t\geq 0}$ is a continuous time binary tree process in which each leaf behaves independently from the other ones.
• At time 0, one individual is alive and the tree $\mathbb{T}_0$ is reduced to a leaf: $\mathbb{T}_0 = \{\emptyset\}$.
• Each individual has an **Exp**(1) distributed lifetime (independent of the other ones). At his death, he disappears and is replaced by two children. The set of individuals alive at time $t$ is $\partial\mathbb{T}_t$. We call Yule tree process, the **BinTree**-valued random process $(\mathbb{T}_t)_{t\geq 0}$. The process $(\mathbb{T}_t)_{t\geq 0}$ is a pure jump Markov process.

### Embedding

The counting process $(N_t)_{t\geq 0}$ giving the number of leaves in $\mathbb{T}_t$,

$$N_t := \#\partial\mathbb{T}_t, \tag{1}$$

is the classical Yule (or binary fission) process (Athreya-Ney [2]).

Let $0 = \tau_0 < \tau_1 < \tau_2 < ...$ be the successive jump times of $\mathbb{T}_.$,

$$\tau_n = \inf\{t : N_t = n+1\}\,. \tag{2}$$

The jump time intervals $(\tau_n - \tau_{n-1})_{n\geq 1}$ are independent and satisfy[1] $\tau_n - \tau_{n-1} \sim \mathbf{Exp}(n)$ for any $n \geq 1$. Moreover, the processes $(\tau_n)_{n\geq 1}$ and $\left(\mathbb{T}_{\tau_n}\right)_{n\geq 1}$ are independent, since the jump chain and jump times are independent.

Due to the lack of memory of the exponential distribution, in the Yule tree process, each leaf is equally likely the first one to produce children. Hence, the two processes $(\mathbb{T}_{\tau_n})_{n\geq 0}$ and $(\mathcal{T}_n)_{n\geq 0}$ have the same law. From now, we consider the Yule tree process and the BST process on the same probability space on which

$$\left(\mathbb{T}_{\tau_n}\right)_{n\geq 0} = \left(\mathcal{T}_n\right)_{n\geq 0}. \tag{3}$$

We say that the BST process is embedded in the Yule tree process. This observation was also made in Aldous-Shields [1] section 1, see also Kingman [17] p.237 and Tavaré [30] p.164 in other contexts. Various embeddings are also mentioned in Devroye [10], in particular those due to Pittel [25], and Biggins [6, 7].

We define $(\mathcal{F}_t)_{t\geq 0}$ by $\mathcal{F}_t = \sigma(\mathbb{T}_s, s \leq t)$ and $\mathcal{F}_{(n)}$ by $\sigma(\mathcal{T}_1, \ldots, \mathcal{T}_n)$, the natural filtrations of $(\mathbb{T}_t)_{t\geq 0}$ and of $(\mathcal{T}_n)_{n\geq 0}$.

This embedding allows to use independence properties between subtrees in the Yule tree process (it is a kind of Poissonization).

### Contents

Many functionals of the BST can be derived using known results on the Yule tree. An interesting quantity is the profile of $\mathcal{T}_n$: it is the sequence $(U_k(n))_{k\geq 0}$ where $U_k(n)$ is the number of leaves of $\mathcal{T}_n$ at level $k$. Here, in (17), the martingale family $(\mathcal{M}_n(z), \mathcal{F}_{(n)})_{n\geq 0}$ which encodes the profile of $(\mathcal{T}_n)_{n\geq 0}$ is shown to be strongly related to the martingale family $(M(t,z), \mathcal{F}_t)_{t\geq 0}$ that encodes the profile of $(\mathbb{T}_t)_{t\geq 0}$. We call the martingale $(\mathcal{M}_n(z), \mathcal{F}_{(n)})_{n\geq 0}$, introduced by Jabbour-Hattab [16], the BST martingale.

It turns out that $M(\tau_n, z) = \mathcal{M}_n(z)\mathrm{C}_n(z)$ where $(\mathrm{C}_n(z), \sigma(\tau_1, \cdots, \tau_n))_{n\geq 0}$ is a martingale independent of $(\mathcal{T}_n)_{n\geq 0}$ (Proposition 2.1). This allows to revisit the study of $(\mathcal{M}_n(z))_{n\geq 0}$ using the embedding. In Section 2.4, we study the convergence, as $n \to \infty$, of the BST martingale $\mathcal{M}_n(z)$. For $z > 0$, we recover very quickly the behavior of the limit $\mathcal{M}_\infty(z)$: positive when $z \in (z_c^-, z_c^+)$, zero when $z \notin [z_c^-, z_c^+]$. In the critical cases $z = z_c^\pm$ the behavior was unknown. We prove that $\mathcal{M}_\infty(z_c^\pm) = 0$ a.s (Theorem 2.5). Moreover we study in Theorem 2.6 the convergence of the family of martingales $\mathcal{M}'_n(z) = \frac{d}{dz}\mathcal{M}_n(z)$. The limits $\mathcal{M}'_\infty(z)$ and $\mathcal{M}_\infty(z)$ satisfy a splitting formula (35)

[1] $\mathbf{Exp}(\lambda)$ is the exponential distribution of parameter $\lambda$. We recall that the minimum of $n$ independent random variables $\mathbf{Exp}(1)$-distributed is $\mathbf{Exp}(n)$-distributed

which, for $z = 1$ gives the Quicksort equation (Corollary 2.7). See the companion paper [9] for complements.

In Section 3 we describe the asymptotic behavior of the profile $U_k(n)$ when $k \simeq 2z \log n$ in the whole range $z \in (z_c^-, z_c^+)$. Previously, the result was known only on a sub-domain where the $L^2$ method works [8].

# 2 Martingales

## 2.1 The BST martingale

Among the known results about the evolution of BST, the saturation level $h_n$ and the height $H_n$,

$$h_n = \min\{|u| : u \in \partial \mathcal{T}_n\} \quad , \quad H_n = \max\{|u| : u \in \partial \mathcal{T}_n\} \tag{4}$$

grow logarithmically (see for instance Devroye [10])

$$\text{a.s.} \quad \lim_{n\to\infty} \frac{h_n}{\log n} = c' = 0.3733... \qquad \lim_{n\to\infty} \frac{H_n}{\log n} = c = 4.31107...\,; \tag{5}$$

the constants $c'$ and $c$ are the two solutions of the equation $\eta_2(x) = 1$ where

$$\eta_\lambda(x) := x \log \frac{x}{\lambda} - x + \lambda, \qquad x \geq 0\,, \tag{6}$$

is the Cramer transform of the Poisson distribution of parameter $\lambda$. Function $\eta_2$ reaches its minimum at $x = 2$. It corresponds to the rate of propagation of the depth of insertion: $\frac{d_n}{2\log n} \xrightarrow{P} 1$. More precise asymptotics for $H_n$ can be found in [12, 26, 27, 20].

To get asymptotic results on the profile, we encode it by the so-called polynomial level $\sum_k U_k(n) z^k$, whose degree is $H_n$. Jabbour-Hattab [8, 16] proved a martingale property for these random polynomials. More precisely, for $z \notin \frac{1}{2}\mathbb{Z}^- = \{0, -1/2, -1, -3/2, \cdots\}$ and $n \geq 0$, let

$$\mathcal{M}_n(z) := \frac{1}{C_n(z)} \sum_{k\geq 0} U_k(n) z^k = \frac{1}{C_n(z)} \sum_{u \in \partial \mathcal{T}_n} z^{|u|}\,, \tag{7}$$

where $C_0(z) = 1$ and for $n \geq 1$,

$$C_n(z) := \prod_{k=0}^{n-1} \frac{k + 2z}{k+1} = (-1)^n \binom{-2z}{n}\,. \tag{8}$$

Then $(\mathcal{M}_n(z), \mathcal{F}_{(n)})_{n\geq 0}$ is a martingale, the BST martingale. If $z > 0$, this positive martingale is a.s. convergent; the limit $\mathcal{M}_\infty(z)$ is positive a.s. if $z \in (z_c^-, z_c^+)$, with

$$z_c^- = c'/2 = 0.186..., \quad z_c^+ = c/2 = 2.155... \tag{9}$$

and $\mathcal{M}_\infty(z) = 0$ for $z \notin [z_c^-, z_c^+]$ (Jabbour-Hattab [16]). This martingale is also the main tool to prove that, properly rescaled around $2\log n$, the profile has a Gaussian limiting shape (see Theorem 1 in [8]).

## 2.2 Martingales and connection

The measure valued process $(\rho_t)_{t\geq 0}$ defined by

$$\rho_t = \sum_{u\in\partial\mathbb{T}_t} \delta_{|u|}\,, \tag{10}$$

can be seen as a continuous time branching random walk. The set of positions is $\mathbb{N}_0 = \{0, 1, 2, \cdots\}$. At time 0, an ancestor is at position 0. Each individual lives during an $\mathbf{Exp}(1)$ distributed time and does not move. At his death, he disappears and is replaced by two children, whose both positions are their parent's position shifted by 1. The set of individuals alive at time $t$ is $\partial\mathbb{T}_t$ and the position of individual $u$ is simply $|u|$.

In classical continuous time branching random walks, the position of the particule $u$ is $X_u$, the parameter of exponential lifetime is $\beta$, and the offspring point process is $Z$. The classical family of "additive" martingales parameterized by $\theta$ in $\mathbb{R}$ (sometimes in $\mathbb{C}$) and indexed by $t \geq 0$, is given by

$$m(t, \theta) := \sum_{u\in\partial\mathbb{T}_t} \exp(\theta X_u - tL(\theta)),$$

where $L(\theta) = \beta(E \int e^{\theta x} Z(dx) - 1)$ (see [5, 21, 31]).

Here, $X_u = |u|$, $\beta = 1$ and $Z = 2\delta_1$. Then

$$L(\theta) = 2e^{\theta} - 1. \tag{11}$$

For easier use, we set $z = e^{\theta}$ and then consider the family of $(\mathcal{F}_t, t \geq 0)$-martingales

$$M(t, z) := m(t, \log z) = \sum_{u\in\partial\mathbb{T}_t} z^{|u|} e^{t(1-2z)}. \tag{12}$$

In particular $M(t, 1/2) = 1$ and $M(t, 1) = e^{-t}N_t$.

A classical result (see Athreya-Ney [2] or Devroye [10] 5.4) is that, a.s., $e^{-t}N_t$ converges when $t \to +\infty$, and

$$\xi := \lim_{t\to\infty} e^{-t}N_t \ \sim \mathbf{Exp}(1)\,. \tag{13}$$

Since $\lim_n \tau_n = \infty$ a.s. we get from (2) and (13),

$$\text{a.s. } \lim_n ne^{-\tau_n} = \xi\,. \tag{14}$$

The embedding formula (3) allows to connect the family of BST martingales $(\mathcal{M}_n(z), \mathcal{F}_{(n)})_{n\geq 0}$ to the family of Yule martingales $(M(t, z), \mathcal{F}_t)_{t\geq 0}$. If we observe the martingale $(M(., z))$ at the stopping times $(\tau_n)_{n\geq 1}$, we can "extract" (Proposition 2.1 below) the space component $\mathcal{M}_n(z)$ and the time component. Let, for $n \geq 0$,

$$\mathrm{C}_n(z) := e^{\tau_n(1-2z)} C_n(z)\,. \tag{15}$$

**Proposition 2.1 (martingale connection)** *Let us assume* $z \in \mathbb{C} \setminus \frac{1}{2}\mathbb{Z}^-$.

*1) The family* $\big(C_n(z), \sigma(\tau_1, \dots, \tau_n)\big)_{n\geq 0}$ *is a martingale with mean 1, and*

$$a.s.\ \lim_n\ C_n(z) = \frac{\xi^{2z-1}}{\Gamma(2z)}. \tag{16}$$

*Moreover, if* $\Re z$*, the real part of* $z$*, is positive, the convergence is in* $L^1$*.*

*2) The two martingales* $(C_n(z))_{n\geq 0}$ *and* $(\mathcal{M}_n(z))_{n\geq 0}$ *are independent and*

$$M(\tau_n, z) = \ C_n(z)\mathcal{M}_n(z)\,. \tag{17}$$

**Proof:** 1) The martingale property comes from the properties of the sequence $(\tau_n)_{n\geq 1}$. The Stirling formula gives the very useful estimate:

$$C_n(z) \underset{n}{\sim} \frac{n^{2z-1}}{\Gamma(2z)}\,, \tag{18}$$

which yields (16) owing to (14).

2) The second claim comes from (3) and (12); the independence comes from the independence between the jump chain and the jump times. ■

Proposition 2.1 allows us to transfer known results about the Yule martingales to BST martingales, thus giving very simple proofs of known results about the BST martingale and also getting much more. In particular, in Theorem 2.4 2), we give the answer to the question asked in [16], about critical values of $z$, with a straightforward argument.

## 2.3 Limiting proportions of nodes in a subtree

Let us study some meaningful random variables arising as a.s limits and playing an important role in the results of Section 2.4. These variables describe the evolution of relative sizes of subtrees in Yule trees (and in BST).

For every $u \in \mathbb{U}$, let $\tau^{(u)} = \inf\{t : u \in \mathbb{T}_t\}$ be the time (a.s. finite) when $u$ appears in the Yule tree, and for $t > 0$, let

$$\mathbb{T}_t^{(u)} = \{v \in \mathbb{U} : uv \in \mathbb{T}_{t+\tau^{(u)}}\}$$

be the tree process growing from $u$. In particular, set

$$N_t^{(u)} = \#\partial\mathbb{T}_t^{(u)}\,.$$

For $t > \tau^{(u)}$, the number of leaves at time $t$ in the subtree issued from node $u$ is $n_t^{(u)} := N_{t-\tau^{(u)}}^{(u)}$. The branching property and (13) give that a.s. for every $u \in \mathbb{U}$

$$\lim_{t\to\infty} e^{-t} N_t^{(u)} = \xi_u \quad , \quad \lim_{t\to\infty} e^{-t} n_t^{(u)} = \xi_u\, e^{-\tau^{(u)}}\,, \tag{19}$$

where $\xi_u$ is distributed as $\xi$ i.e. **Exp**(1). Moreover, if $u$ and $v$ are not in the same line of descent, the r.v. $\xi_u$ and $\xi_v$ are independent. Since, for $t > \tau^{(u0)}$,

$$n_t^{(u)} = n_t^{(u0)} + n_t^{(u1)} \quad \text{and} \quad \tau^{(u0)} = \tau^{(u1)}, \tag{20}$$

a small computation yields[2]

$$\frac{n_t^{(u0)}}{n_t^{(u)}} \xrightarrow{a.s.} U^{(u0)} := \frac{\xi_{u0}}{\xi_{u0}+\xi_{u1}}, \qquad \frac{n_t^{(u1)}}{n_t^{(u)}} \xrightarrow{a.s.} U^{(u1)} := 1 - U^{(u0)} = \frac{\xi_{u1}}{\xi_{u0}+\xi_{u1}}, \tag{21}$$

which allows to attach a $\mathcal{U}([0,1])$ r.v. to each node of $\mathbb{U}$. In particular we set

$$U := U^{(0)} = \frac{\xi_0}{\xi_0+\xi_1} \tag{22}$$

so that

$$\xi := \xi_\emptyset = e^{-\tau_1}(\xi_0+\xi_1) \quad , \quad \xi_0 = U\xi e^{\tau_1} \quad , \quad \xi_1 = (1-U)\xi e^{\tau_1} \,. \tag{23}$$

It is straightforward to see that, by embedding, the property of the above subsection holds true for limiting proportions of nodes in the BST, as $n \to \infty$.

## 2.4 Convergence of martingales

In this section are given the main results about the asymptotic behaviors of the Yule and BST martingales. The martingale connection (Proposition 2.1) allows to express the links between the limits.

Theorem 2.2 gives an answer to a natural question asked in [8] about the domain *in the complex plane* where the BST martingale is $L^1-$convergent and uniformly convergent. Theorem 2.5 gives the optimal $L^1$ domain on $\mathbb{R}$.

**Theorem 2.2** *For* $1 < q < 2$, *let* $\mathcal{V}_q := \{z : \sup_t \mathbb{E}|M(t,z)|^q < \infty\}$. *Then* $\mathcal{V}_q = \{z : \ f(z,q) > 0\}$ *with*

$$f(z,q) := 1 + q(2\Re z - 1) - 2|z|^q \,. \tag{24}$$

*If we denote* $\mathcal{V} := \cup_{1<q<2}\mathcal{V}_q$, *we have:*

*a) As* $t \to \infty$, $\{M(t,z)\}$ *converges, a.s. and in* $L^1$, *uniformly on every compact* $C$ *of* $\mathcal{V}$.

*b) As* $n \to \infty$, $\{\mathcal{M}_n(z)\}$ *converges, a.s. and in* $L^1$, *uniformly on every compact* $C$ *of* $\mathcal{V}$.

**Proof:** a) is proved in [5] Theorem 6 (see also [4]).

b) We prove

$$\lim_N \sup_{n \geq N} \mathbb{E} \sup_{z \in C} |\mathcal{M}_n(z) - \mathcal{M}_N(z)| = 0 \,, \tag{25}$$

which implies the uniform $L^1$ convergence and, since $(\sup_{z\in C} |\mathcal{M}_n(z) - \mathcal{M}_N(z)|)_{n \geq N}$ is a submartingale, this will imply also the a.s. uniform convergence[3]. From the martingale connection (Proposition 2.1), we have

$$\mathcal{M}_n(z) - \mathcal{M}_N(z) = \mathbb{E}[M(\tau_n, z) - M(\tau_N, z)|\mathcal{F}_{(n)}]$$

[2] If $\xi_a$ and $\xi_b$ are two independent, **Exp**(1)-distributed random variables then $\xi_a/(\xi_a+\xi_b) \sim \mathcal{U}[0,1]$, the uniform distribution on [0,1].

[3] For the uniform a.s. convergence, it is possible to give a proof directly from [5]

so that taking supremum and expectation we get

$$\mathbb{E}\sup_{z\in C}|\mathcal{M}_n(z)-\mathcal{M}_N(z)|\leq \mathbb{E}\left(\sup_{z\in C}|M(\tau_n,z)-M(\tau_N,z)|\right).$$

Taking again the supremum in $n$ we get

$$\sup_{n\geq N}\mathbb{E}\sup_{z\in C}|\mathcal{M}_n(z)-\mathcal{M}_N(z)|\leq \mathbb{E}\sup_{n\geq N}\left(\sup_{z\in C}|M(\tau_n,z)-M(\tau_N,z)|\right)\leq \mathbb{E}\Delta_n\,, \tag{26}$$

where we have set $\Delta_n := \sup_{T\geq\tau_n}\left(\sup_{z\in C}|M(T,z)-M(\tau_n,z)|\right)$. Since $M(t,z)$ converges a.s. uniformly, we have a.s. $\lim_n \Delta_n = 0$. Moreover, by the triangle inequality $\Delta_n \leq 2\Delta_0$, and by the proof of Proposition 1 in [4], $\Delta_0$ is integrable. The dominated convergence theorem gives $\lim_n \mathbb{E}\Delta_n = 0$ and (25) holds, which ends the proof of Theorem 2.2. ■

**Remark** 2.3 *As usual the $L^1$ convergence in a) of the above theorem comes from a $L^q$ bound (for some $1 < q \leq 2$); more precisely, following the steps in [3] section 2.4, the quantity*

$$\beta_t(\lambda) := (M(t,z)-1)\,e^{t(2z-1)}$$

*satisfies*

$$E|\beta_t(z)|^q \ \leq e^{tq(2\Re z-1)}\int_0^t \exp\left(-sf(z,q)\right)\,ds \qquad \text{for } 1<q\leq 2. \tag{27}$$

**Theorem 2.4** *Let us assume $z\in(z_c^-, z_c^+)$.*

1) *We have the* **limit martingale connection***:*

$$a.s.\quad M(\infty,z)=\frac{\xi^{2z-1}}{\Gamma(2z)}\,\mathcal{M}_\infty(z)\,, \tag{28}$$

*where the exponential variable $\xi$ is defined in (13).*

2) *We have the following two splitting formulas:*

   a) *for the Yule process,*

$$M(\infty,z)=ze^{(1-2z)\tau_1}\left(M_0(\infty,z)+M_1(\infty,z)\right) \quad a.s. \tag{29}$$

   *where $M_0(\infty,z)$ and $M_1(\infty,z)$ are independent, distributed as $M(\infty,z)$ and independent of $\tau_1$.*

   b) *for the BST,*

$$\mathcal{M}_\infty(z)=z\left(U^{2z-1}\mathcal{M}_{\infty,(0)}(z)+(1-U)^{2z-1}\mathcal{M}_{\infty,(1)}(z)\right) \tag{30}$$

   *where $U\sim\mathcal{U}([0,1])$ is defined in (22), $\mathcal{M}_{\infty,(0)}(z),\mathcal{M}_{\infty,(1)}(z)$ are independent (and independent of $U$) and distributed as $\mathcal{M}_\infty(z)$.*

**Proof:** 1) is a consequence of (16) and the martingale connection (17).

2) a) For $t > \tau_1$ we have the decomposition

$$M(t,z) = ze^{(1-2z)\tau_1}\left[M^{(0)}(t-\tau_1,z) + M^{(1)}(t-\tau_1,z)\right] \tag{31}$$

where for $i = 0, 1$

$$M^{(i)}(s,z) = \sum_{u\in\partial\mathbb{T}_s^{(i)}} z^{|u|}e^{s(1-2z)}\,,$$

and $\mathbb{T}^{(i)}$ is defined in Section 2.3.

b) Take $t = \tau_n$ in (31), condition on the first splitting time $\tau_1$, apply the branching property, let $n \to \infty$ and apply the limit martingale connection (28) to get

$$\frac{\xi^{2z-1}}{\Gamma(2z)}\mathcal{M}_\infty(z) = ze^{(1-2z)\tau_1}\left(\frac{\xi_0^{2z-1}}{\Gamma(2z)}\mathcal{M}_{\infty,(0)}(z) + \frac{\xi_1^{2z-1}}{\Gamma(2z)}\mathcal{M}_{\infty,(1)}(z)\right) \tag{32}$$

where $\xi_0$ and $\xi_1$ come from section 2.3, which yields b) with the help of (23). ■

The following theorem gives the behavior in the remaining cases

**Theorem 2.5** *For $z \in (0,\infty)\setminus(z_c^-, z_c^+)$, then a.s. $\lim_t M(t,z) = 0$ and $\lim_n \mathcal{M}_n(z) = 0$.*

**Proof:** The continuous time result is in [5] (see also [4]); it remains to use again the martingale connection (17). ■

## 2.5 Derivative martingales

From the above section, we deduce that the derivatives

$$M'(t,z) := \frac{d}{dz}M(t,z),\quad \mathcal{M}'_n(z) := \frac{d}{dz}\mathcal{M}_n(z) \tag{33}$$

are martingales which are no longer positive. They are called the derivative martingales. Their behaviors are ruled by the following theorem.

**Theorem 2.6** *1) For $z \in (z_c^-, z_c^+)$, the martingales $(M'(t,z))_{t\geq 0}$ and $(\mathcal{M}'_n(z))_{n\geq 0}$ are convergent a.s.. Let us call $M'(\infty,z)$ and $\mathcal{M}'_\infty(z)$ their limits.*

*a) We have the (derivative martingale) connection:*

$$M'(\infty,z) = \frac{\xi^{2z-1}}{\Gamma(2z)}\left(\mathcal{M}'_\infty(z) + 2\left(\log\xi - \frac{\Gamma'(2z)}{\Gamma(2z)}\right)\mathcal{M}_\infty(z)\right) \qquad a.s. \tag{34}$$

*where $\xi \sim$ **Exp**$(1)$ is defined in (13) and is independent of $\mathcal{M}_\infty(z)$ and $\mathcal{M}'_\infty(z)$.*

*b) We have the splitting formula:*

$$\begin{aligned}\mathcal{M}'_\infty(z) &= zU^{2z-1}\mathcal{M}'_{\infty,(0)}(z) + z(1-U)^{2z-1}\mathcal{M}'_{\infty,(1)}(z) \\ &+ 2z\left(U^{2z-1}\log U\right)\mathcal{M}_{\infty,(0)}(z) + 2z\left((1-U)^{2z-1}\log(1-U)\right)\mathcal{M}_{\infty,(1)}(z) \\ &+ z^{-1}\mathcal{M}_\infty(z) \qquad (35)\end{aligned}$$

*where $U \sim \mathcal{U}([0,1])$ is defined in (22), and the r.v. $\mathcal{M}'_{\infty,(0)}(z)$ and $\mathcal{M}'_{\infty,(1)}(z)$ are independent (and independent of $U$) and distributed as $\mathcal{M}'_\infty(z)$.*

*2) a) The martingales $(M'(t,z_c^-))_{t\geq 0}$ and $(\mathcal{M}'_n(z_c^-))_{n\geq 0}$ (resp. $(M'(t,z_c^+))_{t\geq 0}$ and $(\mathcal{M}'_n(z_c^+))_{n\geq 0}$) are convergent a.s.. Their limits denoted by $M'(\infty,z_c^-)$ and $\mathcal{M}'_\infty(z_c^-)$ (resp. $M'(\infty,z_c^+)$ and $\mathcal{M}'_\infty(z_c^+)$) are positive (resp. negative) and satisfy*

$$\mathbb{E}\big(M'(\infty,z_c^-)\big) = \mathbb{E}(\mathcal{M}_\infty(z_c^-)) = +\infty, \qquad (36)$$

$$\mathbb{E}\big(M'(\infty,z_c^+)\big) = \mathbb{E}(\mathcal{M}_\infty(z_c^+)) = -\infty. \qquad (37)$$

*b) $M'(\infty,z_c^\pm)$ and $\mathcal{M}'_\infty(z_c^\pm)$ satisfy equations similar to (28), (29) and (30):*

$$M'(\infty,z_c^\pm) = \frac{\xi^{2z_c^\pm-1}}{\Gamma(2z_c^\pm)}\,\mathcal{M}'_\infty(z_c^\pm) \qquad (38)$$

$$M'(\infty,z_c^\pm) = z_c^\pm e^{(1-2z_c^\pm)\tau_1}\left(M'_0(\infty,z_c^\pm) + M'_1(\infty,z_c^\pm)\right) \qquad (39)$$

$$\mathcal{M}'_\infty(z_c^\pm) = z_c^\pm\Big(U^{2z_c^\pm-1}\mathcal{M}'_{\infty,(0)}(z_c^\pm) + (1-U)^{2z_c^\pm-1}\mathcal{M}'_{\infty,(1)}(z_c^\pm)\Big) \quad a.s.. \qquad (40)$$

**Proof:** 1) For $z \in (z_c^-, z_c^+)$ the a.s. convergence of $M'(t,z)$ is a consequence of the uniform convergence of $M(t,z)$ (by Theorem 2.2) and of analyticity. Taking derivatives in the martingale connection (17) gives

$$M'(\tau_n,z) = \left[\frac{C'_n(z)}{C_n(z)} - 2\tau_n\right]\ \mathrm{C}_n(z)\mathcal{M}_n(z) + \ \mathrm{C}_n(z)\mathcal{M}'_n(z)\,. \qquad (41)$$

Using (14) again and

$$\frac{C'_n(z)}{C_n(z)} = \sum_{j=0}^{n-1}\frac{2}{j+2z} \quad , \quad \frac{\Gamma'(x)}{\Gamma(x)} = \lim_n\Big(\log n - \sum_{j=0}^{n-1}\frac{1}{x+j}\Big)\,,$$

we get

$$\text{a.s. } \lim_n\left[\frac{C'_n(z)}{C_n(z)} - 2\tau_n\right] \ = \ 2\left[-\frac{\Gamma'(2z)}{\Gamma(2z)} + \log\xi\right]\,. \qquad (42)$$

We conclude that $\mathcal{M}'_n(z)$ converges and that $\mathcal{M}'_\infty(z)$ satisfies (34) which proves a).

To prove b), we differentiate (31) with respect to $z$

$$M'(t,z) = (z^{-1} - 2\tau_1)M(t,z) + ze^{(1-2z)\tau_1}\Big[M^{(0)'}(t-\tau_1,z) + M^{(1)'}(t-\tau_1,z)\Big]\,,$$

and we use the same technique as above: take $t = \tau_n$, let $n \to \infty$, apply (34) and its analogs with $(M'^{(i)}, \mathcal{M}^{(i)}, \mathcal{M}'^{(i)}, \xi_i)_{i=0,1}$ instead of $(M', \mathcal{M}, \mathcal{M}', \xi)$, and use (23).

2) For $z = z_c^{\pm}$, the a.s. convergence of the martingales $M'(t, z)$ and the signs of the limits are proved in [4], and so is the relation

$$\mathbb{E}M'(\infty, z_c^-) = -\mathbb{E}M'(\infty, z_c^+) = \infty\,.$$

Relation (38) is a consequence of (41) and (42), since $\mathcal{M}_\infty(z_c^{\pm}) = 0$.

Formula (40) of 2) is straightforward from (35) since $\mathcal{M}_\infty(z_c^{\pm}) = 0$. Formula (38) is (34) for $z = z_c^{\pm}$. ■

An easy but interesting consequence of (35) is obtained in the following corollary, just taking $z = 1$ in (34) and (35) (remember that $\mathcal{M}_n(1) \equiv 1$). The distributional (weaker) version of (44) below is the subject of a broad literature (see for instance Fill, Janson, Devroye, Neininger, Rösler, Rüschendorf [13, 14, 11, 23, 29, 28]) and some properties of the distribution of $\mathcal{M}'_\infty(1)$ remain unknown.

**Corollary 2.7** *We have*

$$M'(\infty, 1) = \xi\left(\mathcal{M}'_\infty(1) + 2\left(\log \xi + \gamma - 1\right)\right) \quad a.s.\,, \tag{43}$$

*where $\gamma$ is the Euler constant, and $\mathcal{M}'_\infty(1)$ satisfies the a.s. version of the* **Quicksort** *equation:*

$$\mathcal{M}'_\infty(1) = U\mathcal{M}'_{\infty,(0)}(1) + (1-U)\mathcal{M}'_{\infty,(1)}(1) + 2U\log U + 2(1-U)\log(1-U) + 1\,, \tag{44}$$

*where as above, $\mathcal{M}'_{\infty,(0)}(1)$ and $\mathcal{M}'_{\infty,(1)}(1)$ are independent (and independent of $U$), distributed as $\mathcal{M}'_\infty(1)$ and $U \sim \mathcal{U}([0,1])$.*

# 3 Convergence of profiles

## 3.1 Results

According to (5), for every $\epsilon > 0$, there exists a.s. $n_0$ such that for $n \geq n_0$,

$$U_k(n) = 0 \quad \text{for} \quad k \notin [(c' - \epsilon)\log n, (c + \epsilon)\log n]\,.$$

It implies that the convenient scaling for $k$ is $(\log n)^{-1}$. We are interested in the asymptotic behavior of $U_k(n)$ for $k \cong x \log n$ and $x$ fixed in $(c', c)$. The mean profile is known $\mathbb{E}\big(U_k(n)\big) = 2^k S_n^{(k)}/n!$ where $S_n^{(k)}$ is the Stirling number of the first kind. Hwang ([15]) got an asymptotic estimate; for any $\ell > 0$ as $n \to \infty$ and $k \to \infty$ such that $r = k/\log n \leq \ell$ :

$$\mathbb{E}\, U_k(n) = \frac{(2\log n)^k}{k!\, n\, \Gamma(r)}(1 + o(1))\,, \tag{45}$$

from which we deduce that for any $\ell > 0$ :

$$\mathbb{E}\, U_k(n) = \frac{n^{1-\eta_2(\frac{k}{\log n})}}{\Gamma(\frac{k}{\log n})\sqrt{2\pi k}}(1 + o(1))\,, \tag{46}$$

where $o(1)$ is uniform for $k/\log n \leq \ell$ and $\eta_2$ was defined in (6). For related results, see Kolchin ([18], [19] chap.4).

The convergence of the profile is given by the following theorem.

**Theorem 3.1** *Almost surely, for any compact subset $K$ of $(c', c) = (0.373... , 4.311...)$*

$$\lim_n \sup_{k:(k/\log n)\in K} \Big( \frac{U_k(n)}{\mathbb{E}\big(U_k(n)\big)} - \mathcal{M}_\infty\big(\frac{k}{2\log n}\big)\Big) = 0\,. \tag{47}$$

This theorem improves the result of Jabbour-Hattab & al. [8] where the convergence was shown in [1.2 , 2.8].

This theorem gives precise information on the profile. In particular, taking $k = 2\log n + \lambda\sqrt{\log n}$ and using $\mathcal{M}_\infty(k/(2\log n)) \to \mathcal{M}_\infty(1) = 1$, one can prove that the normalized profile

$$\lambda \to \frac{U_{2\log n + 2\lambda\sqrt{\log n}}\ (n)}{n/\sqrt{4\pi\log n}}$$

converges p.s., uniformly on all compacts, to the function $\lambda \to e^{-\lambda^2}$ (see also [8]).

## 3.2 Proof

The aim of this section is to prove Theorem 3.1.

Jabbour-Hattab in [16] introduced the random measure counting the levels of leaves in $\mathcal{T}_n$

$$r_n := \sum_k U_k(n)\delta_k = \sum_{u\in\partial\mathcal{T}_n} \delta_{|u|}\,.$$

He proved that for $x \in (2, c)$

$$\text{a.s.}\ \lim_{n\to\infty} \frac{1}{\log n}\ \log r_n((x\log n, \infty)) \ \ = \ \ 1 - \eta_2(x) \tag{48}$$

and that the same result holds for $x \in (c', 2)$, replacing $(x\log n, \infty)$ by $(0, x\log n)$ .

The random measure counting the levels of leaves in the Yule tree is

$$\rho_t = \sum_{u\in\partial\mathbb{T}_t} \delta_{|u|}\,.$$

With the notations of [31], the exponential rate of growing is ruled by the function

$$x \mapsto L^\star(x) := \sup_\theta \theta x - L(\theta) = \eta_2(x) - 1\,,$$

where the function $L$ is defined in (11) and $\eta_2$ in (6).

For $x \in (c', c)$, $\eta_2(x) < 1$, so there are in mean about $e^{(1-\eta_2(x))t}$ leaves at level $\simeq xt$. More precisely (Theorem 1' p. 909 [31]), for $x \in (c', c)$,

$$\lim_{t\to\infty} \sqrt{t}\ e^{tL^\star(x)}\rho_t([xt]) = \sqrt{\frac{(L^\star)''(x)}{2\pi}}\ M(\infty, x/2)\ \text{ a.s.}. \tag{49}$$

It is now tempting to replace $t$ by $\tau_n$ and $\rho_t([xt])$ by $\rho_{\tau_n}([x\log n]) = r_n([x\log n])$. To validate this, we need some uniformity in $x$ in (49). In [5], Biggins obtained such a result. However, it was in the non-lattice case, so we give in the next subsection a complete proof.

**Proof of Theorem 3.1**

In order to prove Theorem 3.1, we need the following lemma whose proof is postponed; it yields an asymptotic uniform behavior for $\rho_t(k)$, as $t \to \infty$.

**Lemma 3.2** *Almost surely, for any compact $C$ of $(z_c^-, z_c^+)$,*

$$\lim_{t\to\infty} \sup_{k\geq 1, z\in C} z^k \sqrt{t} e^{t(1-2z)} \left[\rho_t(k) - M(\infty, z) e^{-t} \frac{(2t)^k}{k!}\right] = 0 . \tag{50}$$

Let $C$ be a compact subset of $(z_c^-, z_c^+)$. From Lemma 3.2, we know that

$$\rho_t(k) = M(\infty, z) e^{-t} \frac{(2t)^k}{k!} + o(1) z^{-k} t^{-1/2} e^{-t(1-2z)} .$$

Recall that $o(1)$ is uniform in $k$ and in $z \in C$. If $\mathcal{P}^{(\lambda)}$ stands for the Poisson law with parameter $\lambda$, notice that a $\mathcal{P}^{(2t)}$ appears in the previous expression. Using a change of probability from $\mathcal{P}^{(2t)}$ to $\mathcal{P}^{(2tz)}$, we get

$$\rho_t(k) = z^{-k} t^{-1/2} e^{-t(1-2z)} \Big[t^{1/2} M(\infty, z) \mathcal{P}^{(2tz)}(k) + o(1)\Big] .$$

Using the local limit theorem [24], we have

$$\lim_{\lambda\to\infty} \sup_k \Big|\sqrt{2\pi\lambda}\, \mathcal{P}^{(\lambda)}(k) - \exp\Big(-\frac{(k-\lambda)^2}{2\lambda}\Big)\Big| = 0 .$$

Now, we set $\lambda = 2tz$ with $z \in C$ which yields

$$\lim_{t\to\infty} \sup_{z\in C} \sup_k \Big|\sqrt{4\pi t z}\, \mathcal{P}^{(2tz)}(k) - \exp\Big(-\frac{(k-2tz)^2}{4tz}\Big)\Big| = 0.$$

Hence,

$$\rho_t(k) = A_t(k, z) \Big[\exp\Big(-\frac{(k-2tz)^2}{4tz}\Big) M(\infty, z) + \Big((4\pi z)^{1/2} + M(\infty, z)\Big) o(1)\Big], \tag{51}$$

with

$$A_t(k, z) := \frac{e^{-t(1-2z)}}{z^k (4\pi t z)^{1/2}} .$$

Remembering that $U_k(n) = \rho_{\tau_n}(k)$, we take $t = \tau_n$ and $z = \dfrac{k}{2\log n}$ in (51). Using (14) again and the estimate (46), we get

$$\frac{A_{\tau_n}(k, z)}{[\mathbb{E}U_k(n)]\ \xi^{1-2z}\Gamma(2z)} = 1 + o(1) \quad , \quad \exp\Big(-\frac{(k-2\tau_n z)^2}{4\tau_n z}\Big) = 1 + o(1) .$$

Now we apply the limit martingale connection (28) and notice that

$$\sup_{z\in C} \Big((4\pi z)^{1/2} + M(\infty, z)\Big) < \infty$$

and we conclude

$$U_k(n) = [\mathbb{E}U_k(n)] \mathcal{M}_\infty(z)(1 + o(1)) ,$$

with $z = k/(2\log n)$ and $o(1)$ uniform in $z \in C$. ■

**Proof of Lemma 3.2**

We use the following lemma, which is the continuous time version of Lemma 5 in [5]. Its proof can be managed with the same arguments, replacing Lemma 6 there, by Remark 2.3. We omit the details.

**Lemma 3.3** *For any* $z_0 \in (z_c^-, z_c^+)$ *there exists* $r > 0$ *for which* $z_c^- < z_0 - r < z_0 + r < z_c^+$ *and such that a.s.*

$$\lim_{t\to\infty} \sup_{z\in[z_0-r,z_0+r]} \int_{-\pi}^{\pi} \sqrt{t} \mid M(t, ze^{i\eta}) - M(\infty, z) \mid e^{-2tz(1-\cos\eta)} d\eta = 0 \,. \tag{52}$$

Write

$$M(t,z) = e^{t(1-2z)} \sum_k \rho_t(k) z^k \,,$$

and the Fourier inversion formula yields

$$\rho_t(k) = \frac{e^{-t(1-2z)} z^{-k}}{2\pi} \int_{-\pi}^{\pi} M(t, ze^{i\eta}) e^{-2tz(1-e^{i\eta})} e^{-ik\eta} d\eta$$

and, owing to Lemma 3.3

$$2\pi \rho_t(k) e^{t(1-2z)} z^k \sqrt{t} = M(\infty, z)\sqrt{t} \int_{-\pi}^{\pi} e^{-2zt(1-e^{i\eta})} e^{-ik\eta} d\eta + o(1)$$

with $o(1)$ uniform in $k$ and in $z$ in any compact subset of $(z_c^-, z_c^+)$. Now, from the Cauchy formula we get that

$$\int_{-\pi}^{\pi} e^{-2zt(1-e^{i\eta})} e^{-ik\eta} d\eta = 2\pi e^{-2zt} \frac{(2zt)^k}{k!} \,,$$

yielding (50), which ends the proof. ■

## References

[1] D. Aldous and P. Shields. A diffusion limit for a class of randomly-growing binary trees. *Probab. Theory Related Fields*, 79:509–542, 1988.

[2] K. B. Athreya and P. E. Ney. *Branching processes.* Springer-Verlag, New York, 1972.

[3] J. Bertoin. The asymptotic behavior of fragmentation processes. *J. Europ. Math. Soc.*, 5(4):395–416, 2003.

[4] J. Bertoin and A. Rouault. Discretization methods for homogeneous fragmentations. Preprint available at http://front.math.ucdavis.edu/math.PR/0409545, September 2004.

[5] J. D. Biggins. Uniform convergence of martingales in the branching random walk. *Ann. Probab.*, 20(1):137–151, 1992.

[6] J. D. Biggins. How fast does a general branching random walk spread? In *Classical and modern branching processes (Minneapolis, MN, 1994)*, volume 84 of *IMA Vol. Math. Appl.*, pages 19–39. Springer, New York, 1997.

[7] J. D. Biggins and D. R. Grey. A note on the growth of random trees. *Statist. Probab. Lett.*, 32(4):339–342, 1997.

[8] B. Chauvin, M. Drmota, and J. Jabbour-Hattab. The profile of binary search trees. *Ann. Appl. Prob.*, 11:1042–1062, 2001.

[9] B. Chauvin and A. Rouault. Connecting Yule process, bisection and binary search trees via martingales. *Journal of the Iranian Statistical Society*, 3(2):89–116, available at http://front.math.ucdavis.edu/math.PR/0410318, 2004.

[10] L. Devroye. Branching processes and their applications in the analysis of tree structures and tree algorithms. In M. Habib et al., editor, *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springer, 1998.

[11] L. Devroye, J.A. Fill, and R. Neininger. Perfect simulation from the quicksort limit distribution. *Electronic Communications in Probability*, 5:95–99, 2000.

[12] M. Drmota. Stochastic analysis of tree-like data structures. *Proc. R. Soc. Lond.*, A460(27):271–307, 2004.

[13] J.A. Fill and S. Janson. Approximating the limiting quicksort distribution. In *Special Issue of Analysis on Algorithms*, volume 19, pages 376–406, 2001.

[14] J.A. Fill and S. Janson. Quicksort asymptotics. In *Special Issue of Analysis on Algorithms*, volume 44, pages 4–28, 2002.

[15] H.K. Hwang. Asymptotic expansions for the Stirling numbers of the first kind. *J. Combin. Theory Ser. A*, 71(2):343–351, 1995.

[16] J. Jabbour-Hattab. Martingales and large deviations for binary search trees. *Random Structure and Algorithms*, 19:112–127, 2001.

[17] J.F.C. Kingman. The coalescent process. *Stochastic Process. Appl.*, 13:235–248, 1982.

[18] V. F. Kolchin. A problem of the allocation of particles in cells and cycles of random permutations. *Theor. Probab. Appl.*, XVI(1):74–90, 1971.

[19] V. F. Kolchin. *Random graphs.* Encyclopedia of Mathematics. Cambridge University Press, 1999.

[20] P.L. Krapivsky and S.T. Majumdar. Travelling waves, front selection, and exact nontrivial exponents in random fragmentation problem. *Phys. Review Letters*, 85(26):5492–5495, 2000.

[21] A. E. Kyprianou. A note on branching Lévy processes. *Stochastic Process. Appl.*, 82(1):1–14, 1999.

[22] H. Mahmoud. *Evolution of Random Search Trees.* John Wiley, New York, 1992.

[23] R. Neininger and L. Rüschendorf. A general limit theorem for recursive algorithms and combinatorial structures. *Annals of App. Probab.*, 14(1):378–418, 2004.

[24] V.V. Petrov. *Sums of independent random variables.* Springer Verlag, 1975.

[25] B. Pittel. On growing random binary trees. *J. Math. Anal. Appl.*, 103(2):461–480, 1984.

[26] B. Reed. The height of a random binary search tree. *Journal of the ACM*, 50(3):306–332, 2003.

[27] J.M. Robson. Constant bounds on the moments of the height of binary search trees. *Theor. Computer Sci.*, 276:435–444, 2002.

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

[29] U. Rösler. On the analysis of stochastic divide and conquer algorithms. *Algorithmica*, 29(1-2):238–261, 2001. Average-case analysis of algorithms (Princeton, NJ, 1998).

[30] S. Tavaré. The birth process with immigration, and the genealogical structure of large populations. *J. Math. Biol.*, 25(2):161–168, 1987.

[31] K. Uchiyama. Spatial growth of a branching process of particles living in $R^d$. *Ann. Probab.*, 10(4):896–918, 1982.