Support and density of the limit $m$-ary search trees distribution

Brigitte Chauvin$^1$, Quansheng Liu$^2$ and Nicolas Pouyanne$^1$

$^1$Université de Versailles-St-Quentin, Laboratoire de Mathématiques de Versailles, CNRS, UMR 8100
$^2$Université de Bretagne Sud, Laboratoire de Mathématiques de Bretagne Atlantique, CNRS, UMR 6205

The space requirements of an $m$-ary search tree satisfy a well-known phase transition: when $m \leq 26$, the second order asymptotics is Gaussian. When $m \geq 27$, it is not Gaussian any longer and a limit $W$ of a complex-valued martingale arises. We show that the distribution of $W$ has a square integrable density on the complex plane, that its support is the whole complex plane, and that it has finite exponential moments. The proofs are based on the study of the distributional equation $W = \sum_{k=1}^{m} V_k W_k$, where $V_1, \ldots, V_m$ are the spacings of $(m-1)$ independent random variables uniformly distributed on $[0,1]$, $W_1, \ldots, W_m$ are independent copies of $W$ which are also independent of $(V_1, \ldots, V_m)$ and $\lambda$ is a complex number.


1 Introduction

Search trees are fundamental data structures in computer science used in searching and sorting. For integers $m \geq 2$, $m$-ary search trees generalize the binary search tree. The quantity $m$ is called the branching factor.

A random $m$-ary search tree is an $m$-ary tree in which each node has the capacity to contain $(m-1)$ elements called the data or the keys. The keys can be considered as i.i.d. random variables $x_i$, $i \geq 1$, with any absolutely continuous distribution on the interval $[0,1]$.

The tree $T_n, n \geq 0$, is recursively defined as follows: $T_0$ has an empty node-root; $T_1$ has a node-root which contains $x_1$; $T_2$ has a node-root which contains $x_1$ and $x_2$; $\ldots$ $T_{m-1}$ still has one node-root, containing $x_1, \ldots, x_{m-1}$. As soon as the $(m-1)$-th key is inserted in the root, $m$ empty subtrees of the root are created, corresponding from left to right to the $m$ ordered intervals $I_1 = [0, x_{(1)}], \ldots, I_m = [x_{(m-1)}, 1]$, where $0 < x_{(1)} < \cdots < x_{(m-1)} < 1$ are the ordered first $(m-1)$ keys. Each following key $x_m, x_{m+1}, \ldots$ is recursively inserted in the subtree corresponding to the unique interval $I_j$ to which it belongs. As soon as a node is saturated, $m$ empty subtrees of this node are created.
For each $i = \{1, \ldots, m-1\}$ and $n \geq 1$, $X^{(i)}_n$ is the number of nodes in $T_n$ which contain $(i-1)$ keys (and $i$ gaps or free places) after insertion of the $n$-th key; such nodes are named nodes of type $i$. We only take into consideration the external nodes and not the internal nodes which are the saturated nodes. The vector $X_n$ is called the composition vector of the $m$-ary search tree. It provides a model for the space requirement of the algorithm. By spreading the input data in $m$ directions instead of only 2, as is the case for a binary search tree, one seeks to have shorter path lengths and thus quicker searches. One can refer to Mahmoud’s book [Mah92] for further details on search trees.

The following figure is an example of 4-ary search tree obtained by insertion of the successive numbers 0.3, 0.1, 0.4, 0.15, 0.9, 0.2, 0.6, 0.5, 0.35, 0.8, 0.97, 0.93, 0.23, 0.84, 0.62, 0.64, 0.33, 0.83. The corresponding composition vector is $X_{18} = (9, 2, 2)$.

A vast literature is devoted to the asymptotic behavior of this composition vector. A famous phase transition appears. When $m \leq 26$, the random vector admits a central limit theorem with convergence in distribution to a Gaussian vector: see also Mahmoud’s book [Mah92] or Janson [Jan04] for a vectorial treatment.

When $m \geq 27$, an almost sure asymptotics for the composition vector has been obtained in [CP04]:

$$X_n = nv_1 + \Re(n^{\lambda_2}Wv_2) + o(n^{\sigma_2}) \quad a.s.$$

(throughout the paper we write $\Re(z)$ or $\Re z$ for the real part of $z$, and $\Im(z)$ or $\Im z$ for the imaginal part of $z$), where $\lambda_2 = \sigma_2 + i\tau_2$ is the root of the polynomial

$$\prod_{k=1}^{m-1} (z + k) - m!,$$
having the second largest real part $\sigma_2$ and a positive imaginary part $\tau_2$, $v_1$ and $v_2$ are two deterministic vectors, and $W$ is the limit of a complex-valued martingale that admits moments of all positive orders.

Heated conjectures about the second order complex-valued limit distribution $W$ remain open (see [CP04], [Pou05], Chern and Hwang [CH01], Janson [Jan04]). Namely, what is the true order of magnitude of $o(n^{\sigma_2})$ in (1)? Can the $W$ distribution be expressed by means of usual distributions? How heavy are the tails of $W$? Can one identify the distributions of $|W|$ and Arg$(W)$?

A significant step is achieved by Fill and Kapur in [FK04], who establish that $W$ satisfies the following distributional equation called the smoothing equation:

$$W \overset{\text{d}}{=} \sum_{k=1}^{m} V_k^{\lambda_2} W_k,$$

where $V_1, \ldots, V_m$ are the spacings of $(m - 1)$ independent random variables uniformly distributed on $[0, 1]$, $W_1, \ldots, W_m$ are independent copies of $W$ which are also independent of $(V_1, \ldots, V_m)$. The precise definition of $V_k$ will be given hereunder. By a contraction method, Fill and Kapur prove that $W$ is the unique solution of Eq. (3) in the space $\mathcal{M}_2(C)$ of square integrable probability measures having $C = \mathbb{E}(W)$ as expectation. The present paper is based on this characterization of $W$.

It has been recently proved [CLP11] that the continuous-time embedding of the process $(X_n)_n$ has an analogous asymptotic behavior, with a second-order term which is a solution of some distributional equation (not the same one). Inspired by this study of the continuous-time case we prove the following theorem.

**Theorem 1** Let $W$ be the second order limit distribution of an $m$-ary search tree for $m \geq 27$, defined by (1).

(i) The support of $W$ is the whole complex plane.

(ii) The law of $W$ admits a continuous square integrable density on $\mathbb{C}$.

(iii) $\mathbb{E}e^{\delta |W|} < \infty$ for some $\delta > 0$. The exponential moment generating series of $W$ (thus) has a positive radius of convergence.

Thanks to Fill and Kapur results [FK04], these results are immediate corollaries of Theorems 3 and 6 proved in the next two sections.

In the sequel, let $V_1, \ldots, V_m$ be the spacings of $(m - 1)$ independent random variables uniformly distributed on $[0, 1]$. In other words, let $U_1, \ldots, U_{m-1}$ be independent random variables uniformly distributed on $[0, 1]$ and let $U_{(1)} \leq \cdots \leq U_{(m-1)}$ be their order statistics. Denote also $U_{(0)} := 0, U_{(m)} := 1$. For any $k \in \{1, \ldots, m\}$, the random variable $V_k$ is defined by

$$V_k := U_{(k)} - U_{(k-1)}.$$

The variables $V_k$ are Beta$(1, m-1)$-distributed and satisfy $\sum_{k=1}^{m} V_k = 1$ almost surely.

**Remark 2** Details about roots of (2) can be found in Hennequin [Hen91] and Mahmoud [Mah92]. Note that for $m = 2$, the polynomial (2) has the unique root $\lambda = 1$. For $m \geq 3$, it is known that if $\lambda_2$ is a root

\[ \sum_{k=1}^{m} V_k^{\lambda_2} W_k, \]
of the polynomial (2) having the second largest real part, then $\lambda_2$ is nonreal, $\Re \lambda_2 < 1$ for any $m \geq 3$, $\Re \lambda_2 > 0$ if and only if $m \geq 14$ (relate this to Hwang [Hwa03]), and

$$\Re \lambda_2 > \frac{1}{2} \iff m \geq 27.$$  

2 Support

The limit distribution $W$ satisfies (3). From now on, we consider the solutions of the distributional equation

$$Z \overset{\text{d}}{=} \sum_{k=1}^{m} V_k^\lambda Z_k, \quad (4)$$

where $V_1, \ldots, V_m$ are the spacings of $(m - 1)$ independent random variables uniformly distributed on $[0, 1]$, $Z_1, \ldots, Z_m$ are independent copies of $Z$ which are also independent of $(V_1, \ldots, V_m)$ and $\lambda$ is a nonreal complex number.

We assume that

$$\lambda \text{ is a nonreal root of (2) having a positive real part } \sigma. \quad (5)$$

Indeed, $V_1, \ldots, V_m$ are Beta$(1, m - 1)$-distributed and $\Re(\lambda) > 0$ guarantees that $\mathbb{E}|V_1^\lambda| < \infty$. Moreover, we are interested in solutions of (4) having a nonzero expectation and the existence of such solutions implies that $\lambda$ is a root of (2). Note that when $m \leq 13$, no $\lambda$ satisfies (5).

The following theorem implies Theorem 1(i) because $W$ is integrable with $\mathbb{E}W = \frac{1}{\Gamma(1+\lambda_2)} \neq 0$ (see [Pou05] for instance).

**Theorem 3** Let $\lambda$ be a nonreal complex number having a positive real part. If $Z$ is a solution of (4) having a nonzero expectation, then the support of $Z$ is the whole complex plane.

The proofs of Theorem 3 and Theorem 6 make use of the complex-valued random variable

$$A := \sum_{k=1}^{m} V_k^\lambda. \quad (6)$$

Notice that the existence of an integrable solution $Z$ of (4) such that $\mathbb{E}(Z) \neq 0$ implies that

$$\mathbb{E}(A) = 1, \quad (7)$$

which just means that $\lambda$ is a root of the polynomial (2).

**Proof of Theorem 3.** For a complex valued random variable $X$, we denote its support by

$$\text{Supp}(X) = \{x \in \mathbb{C}, \ \forall \varepsilon > 0, \ P(|X - x| < \varepsilon) > 0\}.$$  

Let $Z$ be a solution of (4) having a nonzero expectation. We first prove that $\forall a \in \mathbb{C}$, $\forall z \in \mathbb{C}$,

$$[a \in \text{Supp}(A) \text{ and } z \in \text{Supp}(Z)] \implies az \in \text{Supp}(Z). \quad (8)$$
Indeed, let $\varepsilon > 0$, $a \in \text{Supp}(A)$ and $z \in \text{Supp}(Z)$. Let also $Z_1, \ldots, Z_m$ be i.i.d. copies of $Z$. Then, with positive probability, $|A - a| \leq \varepsilon$ and $|Z_k - z| \leq \varepsilon$ for any $k$. Therefore, with positive probability,

$$\left| \sum_{k=1}^{m} V_k^\lambda Z_k - az \right| = \left| \sum_{k=1}^{m} V_k^\lambda (Z_k - z) + z(A - a) \right| \leq (m + \varepsilon)\varepsilon + |z|\varepsilon.$$ 

Since $\varepsilon$ is arbitrary, this inequality shows that $az \in \text{Supp} \left( \sum_{k=1}^{m} V_k^\lambda Z_k \right)$ which implies that $az \in \text{Supp}(Z)$ because of (4).

Let $z \in \text{Supp}(Z) \setminus \{0\}$. Such $z$ exist because $\mathbb{E}(Z) \neq 0$. Iterating (8), any complex number of the form $a_1 \ldots a_n z$ where $a_1, \ldots, a_n \in \text{Supp}(A)$ belongs to $\text{Supp}(Z)$. Therefore, Lemmas 4 and 5 below imply that $\text{Supp}(Z)$ contains $\mathbb{C} \setminus \{0\}$ which suffices to get the conclusion since the support of a probability measure is a closed set. \hfill \square

**Lemma 4** There exist $c, c' \in \mathbb{C} \setminus \{0\}$ and respective open neighbourhoods $V$ and $V'$ of $c$ and $c'$ such that $|c| > 1$, $|c'| < 1$ and $V \cup V' \subseteq \text{Supp}(A)$.

**Proof.** Obviously,

\[
\text{Supp}(A) = \left\{ \sum_{k=1}^{m} t_k^\lambda, \ 0 \leq t_k \leq 1, \ \sum_{k=1}^{m} t_k = 1 \right\}.
\]

In particular, $\text{Supp}(A)$ contains the set $f \left( [0, 1]^2 \right)$ (the image of $[0, 1]^2$ by $f$), where $f$ is defined by

\[
f : \ [0, 1]^2 \rightarrow \mathbb{C} \quad (s, t) \mapsto (st)^\lambda + (s(1-t))^\lambda + (1-s)^\lambda.
\]

We show that there exist $(s_c, t_c)$ and $(s_{c'}, t_{c'})$ in $[0, 1]^2$ such that $c := f(s_c, t_c)$ and $c' := f(s_{c'}, t_{c'})$ satisfy $|c| > 1$, $0 < |c'| < 1$ and $f$ is a local diffeomorphism in some respective neighbourhoods of $(s_c, t_c)$ and $(s_{c'}, t_{c'})$, which implies the result.

Let $\sigma$ and $\tau$ be respectively the real part and the imaginary part of $\lambda$. By assumption, $0 < \sigma < 1$. We assume that $\tau > 0$; if not, replace $Z$ and $\lambda$ by their conjugates. For any integer $k \geq 1$, denote

\[u_k = \exp \left( -\frac{2k\pi}{\tau} \right) \quad \text{and} \quad u_k' = \exp \left( \frac{\pi - 2k\pi}{\tau} \right).
\]

Then, $u_k$ and $u_k'$ are real numbers in $[0, 1[$ that tend to 0 as $k$ tends to infinity, and they satisfy

\[u_k^\lambda = u_k'^\lambda \in [0, 1] \quad \text{and} \quad u_k'^\lambda = -u_k^\sigma \in ]-1, 0[.
\]

Denote moreover

\[
\begin{align*}
  s_k := u_k + u_k^2, \quad & t_k := \frac{1}{1 + u_k}, \\
  s_k' := u_k' + u_k'^2, \quad & t_k' := \frac{1}{1 + u_k'}.
\end{align*}
\]

As $0 < \sigma < 1$, we have

\[|f(s_k, t_k)| = |u_k^\lambda + u_k^{2\lambda} + (1 - u_k - u_k^2)^\lambda| = 1 + u_k^\sigma + O(u_k)
\]
We prove hereunder that $M \mod 2\ell$ open neighbourhoods of $M$. Then 0

Lemma 5

Let $V$ and $V'$ be respectively open neighbourhoods of $c \in \mathbb{C}$ and $c' \in \mathbb{C}$ with $|c| > 1$ and $0 < |c'| < 1$, which do not contain 0. Let

$$M := \{v_1v_2 \ldots v_n, \ n \geq 1, \ v_1, v_2, \ldots, v_n \in V \cup V'\}.$$

Then $M = \mathbb{C} \setminus \{0\}$.

Proof. Let $\ell$ and $\ell'$ be complex numbers such that $\Re \ell > 0$ and $\Re \ell' < 0$. Let $U$ and $U'$ be respectively open neighbourhoods of $\ell$ and $\ell'$. Denote by $\mathcal{M}$ the additive submonoid of $\mathbb{C}/2i\pi\mathbb{Z}$ generated by $U \cup U'$ mod $2i\pi$; it is the set of classes

$$\mathcal{M} := \{u_1 + u_2 + \cdots + u_n \mod 2i\pi, \ n \geq 1, \ u_1, u_2, \ldots, u_n \in U \cup U'\}.$$

We prove hereunder that $\mathcal{M} = \mathbb{C}/2i\pi\mathbb{Z}$. Taking the exponential, this suffices to prove the lemma.
Take an integer $p \geq 1$ large enough so that $pU$ contains a whole mesh of the lattice generated by $\ell$ and $\ell'$, i.e. such that $pU \supseteq p\ell + [0,1]\ell + [0,1]\ell'$. Then, $\mathcal{M}$ contains the classes mod $2i\pi$ of the sector $S := p\ell + \mathbb{R}_{\geq 0}\ell + \mathbb{R}_{\geq 0}\ell'$. Since $\mathbb{R}\ell \times \mathbb{R}\ell' < 0$, when $z$ is any complex number, there exists $q \in \mathbb{Z}$ such that $z - 2i\pi q \in S$, which proves the result. \hfill \square

## 3 Density and exponential moments

As in the beginning of Section 2, Theorem 1(ii) and (iii) are straightforward corollaries of the following theorem.

**Theorem 6** Let $\lambda$ be a nonreal complex number and $Z$ a solution of (4) having a nonzero expectation.

(i) If $\Re(\lambda) > 0$, then $Z$ admits a continuous square integrable density on $\mathbb{C}$.

(ii) If $\Re(\lambda) > \frac{1}{2}$, then $\mathbb{E}e^{\delta|Z|} < \infty$ for some $\delta > 0$. The exponential moment generating series of $Z$ (thus) has a positive radius of convergence.

**Proof.** It runs along the same lines as in [CLP11] and uses the Fourier transform $\varphi$ of $Z$, namely

$$\varphi(t) := \mathbb{E}\exp\{i(t,Z)\} = \mathbb{E}\exp\{i\Re(IZ)\}, \quad t \in \mathbb{C},$$

where $(x,y) = \Re(xy) = \Re(x)\Re(y) + \Im(x)\Im(y)$. In terms of Fourier transforms, Eq. (4) reads

$$\varphi(t) = \mathbb{E}\left(\prod_{k=1}^{m} \varphi \left( tV_k^X \right) \right). \quad (9)$$

- To get (i), we prove that $\varphi$ is in $L^2(\mathbb{C})$ because it is dominated by $|t|^{-\delta}$ for some $\delta > 1$ so that the inverse Fourier-Plancherel transform provides a square integrable density for $Z$. The guiding idea consists in adapting methods (developed in [Liu99] and [Liu01]) usually applied to positive real-valued
random variables to the present complex-valued case. For any $r \geq 0$, denote

$$\psi(r) := \max_{|t|=r} |\varphi(t)|. $$

Using Theorem 3, one can step by step mimick the proof of Theorem 7.17 in [CLP11] to get the result. We just give hereunder an overview of this proof, written as a succession of hints.

Show first that Theorem 3 implies that $\psi(r) < 1$ for any $r > 0$. Then, notice that

$$\psi(r) \leq E\left( \prod_{k=1}^{m} \psi(r|V_k|^2) \right).$$  \(10\)

By Fatou’s lemma, (10) implies that $\limsup_{r \to \infty} \psi(r) \in \{0, 1\}$. Iterating suitably inequality (10) leads to

$$\lim_{r \to \infty} \psi(r) = 0.$$  \(11\)

Finally, applying (10) again we can show that $\psi(r) = O(r^{-\delta})$ for some $\delta > 1$, so that $\varphi$ is square integrable on $\mathcal{C}$, which leads to the result.

• To get (ii), like in [CLP11], we use Mandelbrot’s cascades. Denote $V := (V_1, V_2, \ldots, V_m)$. Let $U$ be the set of finite sequences of positive integers between 1 and $m$, namely

$$U := \bigcup_{n \geq 1} \{1, 2, \ldots, m\}^n.$$  \(12\)

Elements of $U$ are denoted by concatenation. Let $V_u := (V_{u1}, V_{u2}, \ldots, V_{um})$, $u \in U$ be independent copies of $V$, indexed by all finite sequences of integers $u = u_1 \ldots u_n \in U$.

Introduce the martingale $(Y_n)_{n \geq 1}$ defined by

$$Y_n := \sum_{u_1, \ldots, u_n \in \{1, \ldots, m\}^n} V_{u_1}^{\lambda} V_{u_1 u_2}^{\lambda} \cdots V_{u_1 u_2 \ldots u_n}^{\lambda}.$$  \(13\)

By (7), $E(Y_n) = E(A) = 1$. It can be easily seen that

$$Y_{n+1} = \sum_{k=1}^{m} V_k^{\lambda} Y_{n,k},$$  \(14\)

where $Y_{n,k}$ for $1 \leq k \leq m$ are independent of each other and independent of the $V_k$ and each has the same distribution as $Y_n$. Besides, since $\sigma > \frac{1}{2}$, $mE V_1^{2\sigma} < 1$ and, by Cauchy-Schwarz inequality,

$$E|A|^2 \leq E\left( \sum_{k=1}^{m} |V_k| \right)^2 = E\left( \sum_{k=1}^{m} V_k^{2\sigma} \right)^2 \leq 2E \sum_{k=1}^{m} V_k^{2\sigma} = 2mE V_1^{2\sigma} < 2.$$  \(15\)

Therefore for $n \geq 1$, $Y_n$ is square integrable and

$$\text{Var} Y_{n+1} = (E|A|^2 - 1) + mE V_1^{2\sigma} \text{Var} Y_n,$$

where $\text{Var} X = E(|X - E|X|^2)$ denotes the variance of $X$. Thus, the martingale $(Y_n)_n$ is bounded in $L^2$, so that when $n \to +\infty$,

$$Y_n \to Y_\infty \text{ a.s. and in } L^2,$$
where $Y_\infty$ is a (complex-valued) random variable with variance

$$\text{Var}(Y_\infty) = \frac{E|A|^2 - 1}{1 - mE V_1^2}\sigma^2.$$ 

Passing to the limit in Eq. (11) shows that $Y_\infty$ is a solution of Eq. (4) and by unicity, (ii) in Theorem 6 holds as soon as it holds for $Y_\infty$.

This last fact comes from an adaptation of Lemma 8.29 in [CLP11], giving some constants $C > 0$ and $\varepsilon > 0$ such that for all $t \in C$ with $|t| \leq \varepsilon$, we have

$$\mathbb{E}e^{\langle t, Y_\infty \rangle} \leq e^{R(t) + C|t|^2}.$$ 

The adaptation relies on $\sum_{k=1}^m V_k^{2\sigma} < 1$ a.s. for $\sigma > \frac{1}{2}$. The last assertion implies that $\mathbb{E}e^{t|Y_\infty|} < \infty$ for $t > 0$ small enough, so that the exponential moment generating series of $Y_\infty$ has a positive radius of convergence. □

References


