SPANNING TREE SIZE IN RANDOM BINARY SEARCH TREES

ALOIS PANHOLZER† AND HELMUT PRODINGER‡

ABSTRACT. This paper deals with the size of the spanning tree of p randomly chosen nodes in a binary search tree. It is shown via generating functions methods, that for fixed p, the (normalized) spanning tree size converges in law to the Normal distribution. The special case $p = 2$ reproves the recent result (obtained by the contraction method in [7]), that the distribution of distances in random binary search trees has a Gaussian limit law. In the proof we use the fact that the spanning tree size is closely related to the number of passes in Multiple Quickselect. This parameter, in particular its first two moments were studied earlier in [9]. Here we show also that this normalized parameter has for fixed p order statistics a Gaussian limit law. For $p = 1$ this gives the well known result that the depth of a randomly selected node in a random binary search tree converges in law to the Normal distribution.

1. INTRODUCTION

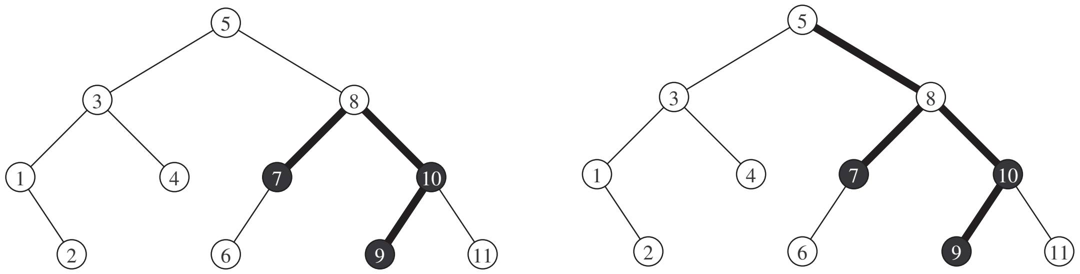
In the papers [7] and [1] the distances between nodes in random search trees resp. random recursive trees were studied. It was proven in [7] that the (edge) distances Δ_n between two randomly selected nodes in random binary search trees of size n are asymptotically normally (Gaussian) distributed, where the so called random permutation model was used as the model of randomness for the trees. This means that every permutation of length n is assumed to be equally likely when generating a binary search tree; furthermore, for selecting nodes, all $\binom{n}{2}$ pairs of nodes are assumed to be equally likely.

In [1] it was shown that the distribution of the distance $\Delta(i,n)$ between a fixed node i and the node n in a random recursive tree of size n is (for a fixed ratio $\rho := \frac{i}{n}$ with $0 < \rho < 1$) asymptotically normally distributed. A related parameter to the distance between two randomly selected nodes is the *Wiener index* of a graph, which is defined to be the sum of all distances between pairs of nodes in the considered graph. The Wiener index was studied for certain families of graphs, and although the scaled mean of this parameter must coincide with the mean distance of two randomly selected nodes, it turned out that the Wiener index was asymptotically *not* normally distributed for random recursive trees and random binary search trees (see [8, 5]).

In this paper we concentrate on random binary search trees and study a natural extension of the distance between two randomly selected nodes, namely the size of the spanning tree of p randomly chosen nodes in the tree. Again, we use the random permutation model for the generation of the binary search trees and also that all $\binom{n}{p}$ possibilities to select p nodes in the tree of size n are equally likely. The random variable $Y_{n,p}$, which counts the size of the spanning tree of p selected nodes in a random binary search tree of size n, is then a direct extension of Δ_n, since the edge distance between two nodes is nothing else than the size of the spanning tree of these two nodes minus one and thus $\Delta_n \stackrel{\mathcal{L}}{=} Y_{n,2} - 1$, where $\stackrel{\mathcal{L}}{=}$ denotes equality in distribution.

In the mathematical analysis of $Y_{n,p}$ we use the fact that it is closely related to the random variable $X_{n,p}$, which counts the number of passes required in the Multiple Quickselect algorithm to find a random p–order statistic in a data file of length n (see [9] and the references cited therein for a description of this divide and conquer algorithm); the natural probability model for the data is, that their ranks form a random permutation of $\{1, \ldots, n\}$ and we assume further that all $\binom{n}{p}$ sets of p order statistics $\{1 \leq i_1 < \ldots < i_p \leq n\}$ are equally likely. Then by well known relations between binary search trees and Quicksort-like algorithms, $X_{n,p}$ is equal to the number of ascendants of p randomly chosen nodes in a random binary search tree of size n or, equivalently, to the size of the spanning tree, spanned by the root and p randomly chosen nodes (where of course the root could have been chosen as well) in a random binary search tree of size n. See Figure 1 for a comparison of both parameters.

FIGURE 1. A binary search tree with the two parameters under consideration.



Spanning tree of the nodes 7, 9 and 10 is of size 4.

5 passes of the Multiple Quickselect algorithm to find the ranks 7, 9 and 10.

The parameter $X_{n,p}$ was studied already in [9], where exact formulæ for the expectation and the variance were given. Here we show additionally that $X_{n,p}$ is, for fixed $p \geq 1$, asymptotically normally distributed.[1]

For $Y_{n,p}$ we also derive exact formulæ for the expectation and the variance and show, that $Y_{n,p}$ is, for fixed $p \geq 2$, asymptotically normally distributed, where the special case $p = 2$ reproves that the distances Δ_n between two randomly selected nodes in random binary search trees of size n are asymptotically normally distributed. Our approach uses generating functions, singularity analysis and a central limit theorem for combinatorial structures due to Hwang and avoids the difficulties which occur in [7] when showing the asymptotic normality for Δ_n using the contraction method which arises due to the degenerate nature of the distributional limit equation for $X_{n,1}$ (that was studied there to obtain the result for Δ_n).

2. Passes in Multiple Quickselect and spanning tree size in binary search trees

First we want to translate the close relation between $X_{n,p}$ and $Y_{n,p}$ into an equation for suitable generating functions as described below.

Here we denote with $\varphi_{n,p,m} := \mathbb{P}\{X_{n,p} = m\}$ the probability that exactly m passes of the Multiple Quickselect algorithm are required, in order to find a random set of p order statistics in a data file of length n, and with $F_{n,p,m} := \mathbb{P}\{Y_{n,p} = m\}$ the probability that the size of the spanning tree of p randomly chosen nodes in a binary search tree of size n is exactly m. Using the recursive structure of the search trees, we obtain for the generating functions $\phi_p(z,v) = \sum_{n,m\geq 0} \binom{n}{p} \varphi_{n,p,m} z^n v^m$ resp. $F_p(z,v) = \sum_{n,m\geq 0} \binom{n}{p} F_{n,p,m} z^n v^m$ for

[1]Philippe Flajolet mentioned that to H. P. in 1998 without working out the details.

$p \geq 1$ the recurrences

$$\frac{\partial}{\partial z}\phi_p(z,v) = v\sum_{i=0}^{p}\phi_i(z,v)\phi_{p-i}(z,v) + v\sum_{i=0}^{p-1}\phi_i(z,v)\phi_{p-1-i}(z,v) \tag{1}$$

and

$$\frac{\partial}{\partial z}F_p(z,v) = v\sum_{i=1}^{p-1}\phi_i(z,v)\phi_{p-i}(z,v) + v\sum_{i=0}^{p-1}\phi_i(z,v)\phi_{p-1-i}(z,v) + 2F_0(z,v)F_p(z,v), \tag{2}$$

with the initial functions $\phi_0(z,v) = F_0(z,v) = \frac{1}{1-z}$. The difference in the above recurrences reflects the difference between both parameters coming from the instance where the root is not selected and also the left ($i = 0$) resp. right ($i = p$) subtree of the root does not contain a selected node.

Introducing the trivariate generating functions $\Phi(z,u,v) = \sum_{p\geq 0}\phi_p(z,v)u^p$ and $F(z,u,v) = \sum_{p\geq 0}F_p(z,v)u^p$, we obtain first from (1) a Riccati differential equation:

$$\frac{\partial}{\partial z}\Phi(z,u,v) = v(1+u)\Phi^2(z,u,v) + \frac{1-v}{(1-z)^2}, \tag{3}$$

with the initial value $\Phi(0,u,v) = 1$. The solution of this equation is already given in [9]:

$$\Phi(z,u,v) = \frac{\Omega + 1 - 2v + (1-z)^{\Omega}(\Omega - 1 + 2v)}{\big(\Omega + 1 - 2v(1+u) + (1-z)^{\Omega}(\Omega - 1 + 2v(1+u))\big)(1-z)}, \tag{4}$$

with

$$\Omega = \sqrt{1 - 4(1+u)v(1-v)}. \tag{5}$$

For $F(z,u,v)$ we get from (2) the differential equation

$$\frac{\partial}{\partial z}F(z,u,v) = v(1+u)\Phi^2(z,u,v) - \frac{2v}{1-z}\Phi(z,u,v) + \frac{2}{1-z}F(z,u,v) + \frac{v-1}{(1-z)^2}$$

or

$$\frac{\partial}{\partial z}F(z,u,v) = \frac{\partial}{\partial z}\Phi(z,u,v) - \frac{2v}{1-z}\Phi(z,u,v) + \frac{2}{1-z}F(z,u,v) + \frac{2(v-1)}{(1-z)^2},$$

with $F(0,u,v) = 1$.

This equation has then the solution

$$F(z,u,v) = \frac{1+2z(v-1)}{(1-z)^2} + \frac{1}{(1-z)^2}\int_0^z \left[\frac{\partial}{\partial t}\Phi(t,u,v) - \frac{2v}{1-t}\Phi(t,u,v)\right](1-t)^2 dt, \tag{6}$$

with $\Phi(z,u,v)$ given by (4).

3. Expectation and variance of the spanning tree size

From (6) it is easy to obtain exact formulæ for the expectation

$$E_{n,p} = \mathbb{E}(Y_{n,p}) = \frac{1}{\binom{n}{p}}[z^n u^p]\left.\frac{\partial}{\partial v}F(z,u,v)\right|_{v=1}$$

and the second factorial moment

$$M_{n,p}^{(2)} = \mathbb{E}\big(Y_{n,p}(Y_{n,p}-1)\big) = \frac{1}{\binom{n}{p}}[z^n u^p]\left.\frac{\partial^2}{\partial v^2}F(z,u,v)\right|_{v=1}$$

(and thus also for the variance $V_{n,p}$) of $Y_{n,p}$, the size of the spanning tree of p randomly selected nodes in a random binary search tree of size n.

Differentiating (6) with respect to v and evaluating at $v=1$ gives the following equation for $E(z,u) := \frac{\partial}{\partial v} F(z,u,v)\big|_{v=1}$:

$$E(z,u) = \frac{2z}{(1-z)^2} + \frac{1}{(1-z)^2} \int_0^z \left[\frac{4(1-t)(1+u)u^2}{(1-t(1+u))^3} \log \frac{1}{1-t} + \frac{X}{(1-t(1+u))^3} \right] dt, \quad (7)$$

with $X = (-2+u) + (6+3u-3u^2-4u^3)t + (1+u)(2u^2-3u-6)t^2 + (2+u)(1+u)^2 t^3$.

This can be simplified to

$$\begin{aligned} E(z,u) &= \frac{2u(1+u)}{(1-z(1+u))^2} \log \frac{1}{1-z} - \frac{2u}{(1-z)^2(1+u)^2} \log \frac{1}{1-z(1+u)} \\ &+ \frac{zu(1-2z-3u+z^2+uz-2u^2+2uz^2+3u^2z+u^2z^2)}{(1-z)^2(1+u)(1-z(1+u))^2}. \end{aligned} \quad (8)$$

To extract coefficients we use here and in the sequel the general formulæ (see e. g. [3])

$$\begin{aligned} [z^n] \frac{1}{(1-z)^{m+1}} \log \frac{1}{1-z} &= \binom{n+m}{n} (H_{n+m} - H_m), \\ [z^n] \frac{1}{(1-z)^{m+1}} \log^2 \frac{1}{1-z} &= \binom{n+m}{n} \big((H_{n+m} - H_m)^2 - (H_{n+m}^{(2)} - H_m^{(2)}) \big), \end{aligned}$$

where $H_n = \sum_{k=1}^n \frac{1}{k}$ and $H_n^{(2)} = \sum_{k=1}^n \frac{1}{k^2}$ denote the first and second order harmonic numbers.

By lengthy, but routine calculations, we finally get for $E_{n,p} = \frac{1}{\binom{n}{p}} [z^n u^p] E(z,u)$ an exact formula, which is given in the next lemma:

Lemma 1. *The expectation $E_{n,p} = \mathbb{E}(Y_{n,p})$ of the size of the spanning tree of p randomly chosen nodes in a random binary search tree of size n is for $p \geq 1$ given by*

$$\begin{aligned} E_{n,p} = &\frac{2p(n+1)^2}{(n+2-p)(n+1-p)} (H_n - H_p) + \frac{2(2p-1)(n+1)}{(n+2-p)(n+1-p)} + 3 + 2p - \frac{2pn}{n+1-p} \\ &+ \frac{2p(n+1)(-1)^p}{\binom{n}{p}} H_n + \frac{2p(n+1)(-1)^p}{\binom{n}{p}} \sum_{k=1}^{p-1} \frac{(-1)^k}{k} \binom{n}{k}, \end{aligned}$$

and asymptotically for fixed $p \geq 2$ by

$$E_{n,p} = 2p \log n + 2p\gamma - 2pH_p + 3 - 2p - \frac{2p}{p-1} + \mathcal{O}\Big(\frac{\log n}{n}\Big).$$

For $p=1$ the formula simplifies to $E_{n,1} = 1$ as it should.

We remark that

$$H_n = \sum_{k=1}^n \frac{(-1)^{k-1}}{k} \binom{n}{k},$$

and so one can give the alternative formula

$$\begin{aligned} E_{n,p} = &\frac{2p(n+1)^2}{(n+2-p)(n+1-p)} (H_n - H_p) + \frac{2(2p-1)(n+1)}{(n+2-p)(n+1-p)} + 3 + 2p - \frac{2pn}{n+1-p} \\ &+ \frac{2p(n+1)(-1)^p}{\binom{n}{p}} \sum_{k=p}^{n} \frac{(-1)^{k-1}}{k} \binom{n}{k}. \end{aligned}$$

When we differentiate equation (6) twice with respect to v and evaluate at $v = 1$, we finally obtain the following formula for $M_2(z,u) := \frac{\partial^2}{\partial v^2} F(z,u,v)\Big|_{v=1}$:

$$
\begin{aligned}
M_2(z,u) = &-\frac{8u}{(1-z)^2}\int_0^z \frac{1}{1-t(1+u)}\log\frac{1}{1-t}dt + \frac{4u(1+u)^2(1-z+2u-uz)}{(1-z(1+u))^3}\log^2\frac{1}{1-z} \\
&+ \frac{12u}{(1-z)^2(1+u)^2}\log\frac{1}{1-z(1+u)} + \frac{4u\Psi_1}{(1-z)(1-z(1+u))^3}\log\frac{1}{1-z} \\
&+ \frac{2u^2z\Psi_2}{(1-z)^2(1-z(1+u))^3(1+u)},
\end{aligned} \tag{9}
$$

with the abbreviations

$$
\begin{aligned}
\Psi_1 &= -z^2u^2 - 3z^2 - 5z^2u + u^3z^2 + 6z + 9zu + 2u^2z - u^3z - 3u^2 - 4u - 3, \\
\Psi_2 &= -z^3u^3 + 3u^3z^2 - 2u^3z - 19u^2z + 22z^2u^2 + 6u^2 - 3z^3u^2 + 14u - 3z^3u - 46zu \\
&\quad + 35z^2u + 14 - 29z - z^3 + 16z^2.
\end{aligned}
$$

Extracting coefficients gives after a somewhat lengthy calculation an exact formula for the second factorial moment $M_{n,p}^{(2)} = \frac{1}{\binom{n}{p}}[z^n u^p]M_2(z,u)$ and we get via $V_{n,p} = M_{n,p}^{(2)} + E_{n,p} - E_{n,p}^2$ the following result:

Lemma 2. *The variance $V_{n,p} = \mathbb{V}(Y_{n,p})$ of the size of the spanning tree of p chosen nodes in a random binary search tree of size n is for $p \geq 2$ given by*

$$
\begin{aligned}
V_{n,p} = &\frac{4(-1)^p(n+1)(2pH_n - 2pH_p + 2 - 3p^2)}{p\binom{n}{p}}\sum_{k=1}^{p-1}\frac{(-1)^k}{k}\binom{n}{k} \\
&- \frac{8(-1)^p(n+1)}{\binom{n}{p}}\sum_{k=1}^{p-1}\frac{(-1)^k}{k^2}\binom{n}{k} + \frac{4(-1)^p(n+1)}{\binom{n}{p}}(H_n^2 - H_n^{(2)} - 2H_pH_n) \\
&+ \frac{4(-1)^p(2-3p^2)(n+1)}{p\binom{n}{p}}H_n - \frac{4\Psi_3}{(n+4-p)^{\underline{4}}}(H_n - H_p) \\
&+ \frac{4p(n+2)(n+1)^3(np+2+p)}{(n+4-p)^{\underline{4}}}\big((H_n-H_p)^2 - (H_n^{(2)} - H_p^{(2)})\big) + \frac{2\Psi_4}{(n+4-p)^{\underline{4}}} \\
&+ E_{n,p} - E_{n,p}^2,
\end{aligned}
$$

with

$$
\begin{aligned}
\Psi_3 = &-2p^4 - 6n^2p^3 + 16p^3 - 2n^3p^3 - 45p^2n - 58p^2 - 4n^2p^2 + 2p^2n^4 + 7n^3p^2 + 56p \\
&+ 78np + 6n^3p + 41pn^2 - n^4p - 8 - 20n - 16n^2 - 4n^3, \\
\Psi_4 = &-144 - 6p^5 - 3n^4 - 152p^3n + 2p^4n^2 + 25p^4n - 234n + 78p + 10np - 5n^4p - 39n^2p^3 \\
&- 4n^3p^3 + 250p^2n + 119n^2p^2 + 2p^2n^4 + 25n^3p^2 - 22n^3p - 35pn^2 + 155p^2 - 173p^3 + 58p^4 \\
&- 153n^2 - 42n^3.
\end{aligned}
$$

Further we have $V_{n,1} = 0$ and the following asymptotic expansion for $n \to \infty$ and fixed $p \geq 2$:

$$
V_{n,p} = 2p\log n - 2p(H_p - \gamma) - 4p^2\Big(\frac{\pi^2}{6} - H_p^{(2)}\Big) + \frac{2(-2+7p-5p^2+2p^3)}{(1-p)^2} + \mathcal{O}\Big(\frac{\log^2 n}{n}\Big).
$$

Here we used the abbreviation $x^{\underline{m}} := x(x-1)\cdots(x-m+1)$ for the falling factorials.

We remark again that an alternative representation of the variance would be possible using the additional formula

$$\frac{1}{2}\Big(H_n^2 + H_n^{(2)}\Big) = \sum_{k=1}^{n} \frac{(-1)^{k-1}}{k^2}\binom{n}{k}.$$

4. The limiting distribution of the number of passes in Multiple Quickselect

We will show that both random variables $X_{n,p}$ and $Y_{n,p}$ satisfy, for fixed p, a Gaussian limit law. To do this, we will expand the coefficients at u^p (for fixed p) of the trivariate generating functions $\Phi(z,u,v)$ resp. $F(z,u,v)$ around their dominant singularity $z = 1$, where the expansion holds uniformly for $|v-1| < \tau$, for $\tau > 0$. Singularity analysis (see [2]) of generating functions allows then to translate these expansions into an asymptotic expansion of the moment generating function (the Laplace transform) of the considered random variables. Then we can apply the so called Quasi power theorem (see [4]) to establish the weak convergence of the random variables to the normal distribution with certain convergence rates.

In this section we will treat the random variable $X_{n,p}$. As described above, we are interested in an asymptotic expansion of $\frac{1}{\binom{n}{p}}[z^n u^p]\Phi(z,u,v)$ for $n \to \infty$ and fixed p uniformly for $|v-1| \le \tau$, where $\Phi(z,u,v)$ is given by the exact formula (4).

To expand $\Phi(z,u,v)$ we will use some auxiliary expansions of

$$f(u) = \Omega + 1 - 2v + (1-z)^{\Omega}(\Omega - 1 + 2v), \tag{10}$$

$$g(u) = \Omega + 1 - 2v(1+u) + (1-z)^{\Omega}(\Omega - 1 + 2v(1+u)) \tag{11}$$

with Ω given by (5). All $\mathcal{O}$–terms in the expansions given below are uniform for $|v-1| \le \tau$, as required. In the sequel, we will use the notations D_u for the differential operator w. r. t. u and N_u for the evaluation operator at $u = 0$.

Since

$$\Omega = (2v-1)\sum_{k\ge 0}\binom{\frac{1}{2}}{k}\Big(-\frac{4v(1-v)}{1-4v(1-v)}\Big)^k u^k,$$

we get

$$(1-z)^{\Omega} = e^{\Omega\log(1-z)} = (1-z)^{2v-1}\exp\Big[(2v-1)\log(1-z)\sum_{k\ge 1}\binom{\frac{1}{2}}{k}\Big(-\frac{4v(1-v)}{1-4v(1-v)}\Big)^k u^k\Big],$$

and thus

$$N_u D_u^p (1-z)^{\Omega} = \mathcal{O}\Big((1-z)^{2v-1}\log^p(1-z)\Big). \tag{12}$$

We have

$$f(0) = g(0) = (1-z)^{2v-1} 2(2v-1) \tag{13}$$

and we get further

$$N_u D_u f(u) = -\frac{4v(1-v)}{2(2v-1)} + \mathcal{O}\Big((1-z)^{2v-1}\log(1-z)\Big), \tag{14a}$$

$$N_u D_u g(u) = -\frac{2v^2}{2v-1} + \mathcal{O}\Big((1-z)^{2v-1}\log(1-z)\Big), \tag{14b}$$

$$N_u D_u^p f(u) = (2v-1)p!\binom{\frac{1}{2}}{p}\Big(-\frac{4v(1-v)}{1-4v(1-v)}\Big)^p + \mathcal{O}\big((1-z)^{2v-1}\log^p(1-z)\big), \tag{14c}$$

$$N_u D_u^p g(u) = (2v-1)p!\binom{\frac{1}{2}}{p}\Big(-\frac{4v(1-v)}{1-4v(1-v)}\Big)^p + \mathcal{O}\big((1-z)^{2v-1}\log^p(1-z)\big), \tag{14d}$$

for $p \geq 2$. Furthermore, we want to expand $D_u^p(g(u))^{-1}$ (for $p \geq 1$) in terms of falling powers of $(g(u))^{-1}$, which gives

$$\begin{aligned} D_u^p(g(u))^{-1} = (-1)^p p!(g(u))^{-p-1}(g'(u))^p + \frac{(-1)^{p-1}(p-1)p!}{2}(g(u))^{-p}(g'(u))^{p-2}g''(u) \\ + \mathcal{O}((g(u))^{-p+1}) \end{aligned}$$

and hence we obtain the expansion

$$\begin{aligned} N_u D_u^p(g(u))^{-1} &= (-1)^p p!\frac{\big(-\frac{2v^2}{2v-1} + \mathcal{O}\big((1-z)^{2v-1}\log(1-z)\big)\big)^p}{(1-z)^{(p+1)(2v-1)}2^{p+1}(2v-1)^{p+1}} + \mathcal{O}\Big(\frac{1}{(1-z)^{p(2v-1)}}\Big) \\ &= \frac{(-1)^p p!\big(-\frac{2v^2}{2v-1}\big)^p}{(1-z)^{(p+1)(2v-1)}2^{p+1}(2v-1)^{p+1}} + \mathcal{O}\Big(\frac{\log(1-z)}{(1-z)^{p(2v-1)}}\Big). \end{aligned} \tag{15}$$

This finally gives

$$\begin{aligned} &N_u D_u^p \Phi(z,u,v) = N_u D_u^p \frac{f(u)}{(1-z)g(u)} \\ &= \frac{1}{1-z}f(0)N_u D_u^p(g(u))^{-1} + \frac{1}{1-z}pf'(0)N_u D_u^{p-1}(g(u))^{-1} + \frac{1}{1-z}\mathcal{O}\Big(\frac{1}{(1-z)^{(p-1)(2v-1)}}\Big) \\ &= \frac{(1-z)^{2v-1}2(2v-1)(-1)^p p!\big(-\frac{2v^2}{2v-1}\big)^p}{(1-z)^{(p+1)(2v-1)+1}2^{p+1}(2v-1)^{p+1}} + \frac{p\big(-\frac{4v(1-v)}{2(2v-1)}\big)(-1)^{p-1}(p-1)!\big(-\frac{2v^2}{2v-1}\big)^{p-1}}{(1-z)^{p(2v-1)+1}2^p(2v-1)^p} \\ &\quad + \mathcal{O}\Big(\frac{\log(1-z)}{(1-z)^{(p-1)(2v-1)+1}}\Big) \\ &= \frac{p!\big(\frac{v}{2v-1}\big)^{2p-1}}{(1-z)^{p(2v-1)+1}} + \mathcal{O}\Big(\frac{\log(1-z)}{(1-z)^{(p-1)(2v-1)+1}}\Big). \end{aligned} \tag{16}$$

Singularity analysis leads then directly to

$$\begin{aligned} [z^n]N_u D_u^p\Phi(z,u,v) &= p!\big(\frac{v}{2v-1}\big)^{2p-1}\frac{n^{p(2v-1)}}{\Gamma(p(2v-1)+1)}\Big(1+\mathcal{O}\big(\frac{1}{n}\big)\Big) + \mathcal{O}\big((\log n)\, n^{(p-1)(2v-1)}\big) \\ &= \frac{p!\big(\frac{v}{2v-1}\big)^{2p-1}n^{p(2v-1)}}{\Gamma(p(2v-1)+1)}\Big(1+\mathcal{O}\big(\frac{1}{n}\big)\Big)\Big(1+\mathcal{O}\big(\frac{\log n}{n^{2v-1}}\big)\Big) \\ &= \frac{p!\big(\frac{v}{2v-1}\big)^{2p-1}n^{p(2v-1)}}{\Gamma(p(2v-1)+1)}\Big(1+\mathcal{O}\big(\frac{1}{n^{1-\epsilon}}\big)\Big), \end{aligned} \tag{17}$$

uniformly for $\epsilon > 0$ and $|v-1| \leq \tau := \frac{\epsilon}{3}$ and also to the following expansion, which is valid for fixed $p \geq 1$:

$$\begin{aligned} \frac{1}{\binom{n}{p}}[z^n u^p]\Phi(z,u,v) &= \frac{p!}{n^p}[z^n u^p]\Phi(z,u,v)\Big(1+\mathcal{O}\big(\frac{1}{n}\big)\Big) \\ &= \frac{1}{n^p}[z^n]N_u D_u^p\Phi(z,u,v)\Big(1+\mathcal{O}\big(\frac{1}{n}\big)\Big) \\ &= \frac{p!\big(\frac{v}{2v-1}\big)^{2p-1}n^{p(2v-2)}}{\Gamma(p(2v-1)+1)}\Big(1+\mathcal{O}\big(\frac{1}{n^{1-\epsilon}}\big)\Big) \end{aligned} \tag{18}$$

$$= \exp\left[p(2v-2)\log n + \log\left(\frac{p!\left(\frac{v}{2v-1}\right)^{2p-1}}{\Gamma(p(2v-1)+1)}\right)\right]\Big(1+\mathcal{O}\big(\frac{1}{n^{1-\epsilon}}\big)\Big). \tag{19}$$

We give here the Quasi power theorem as proven in [4], which we want to apply to our problem:

Theorem 3. [H. K. Hwang] *Let $\{\Omega_n\}_{n\ge 1}$ be a sequence of integral random variables. Suppose that the moment generating function satisfies the asymptotic expression*

$$M_n(s) := \mathbb{E}\big(e^{\Omega_n s}\big) = \sum_{m\ge 0} \mathbb{P}\{\Omega_n = m\} e^{ms} = e^{H_n(s)}\big(1+\mathcal{O}(\kappa_n^{-1})\big),$$

the $\mathcal{O}$–term being uniform for $|s|\le\tau$, $s\in\mathbb{C}$, $\tau>0$, where

(i) *$H_n(s) = u(s)\phi(n) + v(s)$, with $u(s)$ and $v(s)$ analytic for $|s|\le\tau$ and independent of n; $u''(0)\ne 0$,*
(ii) *$\phi(n)\to\infty$,*
(iii) *$\kappa_n\to\infty$.*

Under these assumptions, the distribution of Ω_n is asymptotically Gaussian:

$$\mathbb{P}\bigg\{\frac{\Omega_n - u'(0)\phi(n)}{\sqrt{u''(0)\phi(n)}} < x\bigg\} = \Phi(x) + \mathcal{O}\bigg(\frac{1}{\kappa_n} + \frac{1}{\sqrt{\phi(n)}}\bigg),$$

uniformly with respect to x, $x\in\mathbb{R}$. Here $\Phi(x)$ denotes the distribution function of the standard normal distribution $\mathcal{N}(0,1)$.

Moreover, the mean and the variance of Ω_n satisfy

$$\mathbb{E}(\Omega_n) = u'(0)\phi(n) + v'(0) + \mathcal{O}(\kappa_n^{-1}), \qquad \mathbb{V}(\Omega_n) = u''(0)\phi(n) + v''(0) + \mathcal{O}(\kappa_n^{-1}).$$

From (19) we get with the notations in Theorem 3:

$$u(s) = p(2e^s - 2), \quad v(s) = \log\left(\frac{p!\left(\frac{e^s}{2e^s-1}\right)^{2p-1}}{\Gamma(p(2e^s-1)+1)}\right), \quad \phi(n) = \log n, \quad \kappa_n = n^{1-\epsilon}.$$

We find

$$u'(0) = 2p, \qquad u''(0) = 2p, \tag{20}$$

and

$$\begin{aligned} v'(0) &= -2p + 1 - 2p\Psi(p+1) = -2pH_p + 2p\gamma + 1 - 2p,\\ v''(0) &= 2(2p-1) - 2p\Psi(p+1) - 4p^2\Psi'(p+1)\\ &= 2(2p-1) - 2pH_p + 2p\gamma - \frac{2}{3}\pi^2 p^2 + 4p^2 H_p^{(2)}, \end{aligned} \tag{21}$$

where $\Psi(x)$ denotes the digamma function: $\Psi(x) := \big(\log\Gamma(x)\big)'$.

Hence, with equations (20) and (21), we get from Theorem 3 the following result:

Theorem 4. *The distribution of the random variable $X_{n,p}$, which counts the number of passes in the Multiple Quickselect algorithm that are required to find a random order statistic of p elements in a data file of size n, is for fixed $p\ge 1$ asymptotically Gaussian, where the convergence rate is of order $\mathcal{O}\big(\frac{1}{\sqrt{\log n}}\big)$:*

$$\mathbb{P}\left\{\frac{X_{n,p} - 2p\log n}{\sqrt{2p\log n}}\right\} = \Phi(x) + \mathcal{O}\Big(\frac{1}{\sqrt{\log n}}\Big),$$

and the expectation $E_{n,p} = \mathbb{E}(X_{n,p})$ and the variance $V_{n,p} = \mathbb{V}(X_{n,p})$ satisfy

$$E_{n,p} = 2p\log n + 1 - 2p - 2pH_p + 2p\gamma + \mathcal{O}\big(\frac{1}{n^{1-\epsilon}}\big),$$

$$V_{n,p} = 2p\log n + 4p - 2 - 2pH_p + 2p\gamma + 4p^2 H_p^{(2)} - \frac{2}{3}\pi^2 p^2 + \mathcal{O}\big(\frac{1}{n^{1-\epsilon}}\big).$$

The result for $E_{n,p}$ and $V_{n,p}$ already appeared in [9], but unfortunately there was a typo in the formula for $V_{n,p}$.

For the case $p = 1$ we have that $X_{n,1}$ counts the number of comparisons encountered by a successful search in a random binary search tree and this is, up to an additive constant, the same as the depth D_n of a randomly selected node, $D_n \stackrel{\mathcal{L}}{=} X_{n,1} - 1$. The asymptotic normality of the distribution of $X_{n,1}$ is well known (see e. g. [6]), and the convergence rate was recently established in [7].

5. The limiting distribution of the spanning tree size in binary search trees

In this section we will show, that the normalized random variable $Y_{n,p}$, as defined in Section 4, has for fixed p a Gaussian limiting distribution. Hence we are interested in an asymptotic expansion of $\frac{1}{\binom{n}{p}}[z^n u^p]F(z,u,v)$ for $n \to \infty$ and fixed p uniformly for $|v-1| \le \tau$, where $F(z,u,v)$ is given by (6).

To do this, we will first study the behaviour near the singularity $z = 1$ of the expression

$$\tilde{\Phi}(z,u,v) = \left(\frac{\partial}{\partial z}\Phi(z,u,v) - \frac{2v}{1-z}\Phi(z,u,v)\right)(1-z)^2, \tag{22}$$

which we can write as

$$\tilde{\Phi}(z,u,v) = \frac{\tilde{f}(u)}{(g(u))^2}, \tag{23}$$

where the function $\tilde{f}(u)$ is defined by

$$\tilde{f}(u) = -\Omega(\Omega - 1 + 2v)(1-z)^{\Omega} g(u) + \Omega(\Omega - 1 + 2v(1+u))(1-z)^{\Omega} f(u) + (1-2v)f(u)g(u), \tag{24}$$

and $\Phi(z,u,v)$, Ω, $f(u)$ and $g(u)$ are given by equations (4), (5), (10) and (11), respectively.

The relevant expansions are now

$$\begin{aligned}
\tilde{f}(0) &= -4(2v-1)^3(1-z)^{4v-2},\\
\tilde{f}'(0) &= 8v^2(2v-1)(1-z)^{2v-1} + \mathcal{O}\big(\log(1-z)\,(1-z)^{4v-2}\big),\\
\tilde{f}''(0) &= \frac{8(v-1)v^3}{2v-1} + \mathcal{O}\big(\log(1-z)\,(1-z)^{2v-1}\big),
\end{aligned}$$

and

$$N_u D_u^p (g(u))^{-2} = (-1)^p (p+1)! \frac{(g'(0))^p}{(g(0))^{p+2}} + \mathcal{O}\Big(\frac{1}{(g(0))^{p+1}}\Big),$$

which leads for $p \ge 2$ eventually to

$$\begin{aligned}
N_u D_u^p \tilde{\Phi}(z,u,v) &= \tilde{f}(0) N_u D_u^p (g(u))^{-2} + p\tilde{f}'(0) N_u D_u^{p-1}(g(u))^{-2}\\
&\quad + \frac{p(p-1)}{2}\tilde{f}''(0) N_u D_u^{p-2}(g(u))^{-2} + \mathcal{O}\Big(\frac{1}{(1-z)^{(p-1)(2v-1)}}\Big)\\
&= \frac{(p-1)p!v\big(\frac{v}{2v-1}\big)^{2p-2}}{(1-z)^{p(2v-1)}} + \mathcal{O}\Big(\frac{\log(1-z)}{(1-z)^{(p-1)(2v-1)}}\Big).
\end{aligned} \tag{25}$$

This gives then

$$\frac{1}{\binom{n}{p}}[z^n u^p]F(z,u,v) = \frac{1}{\binom{n}{p}}[z^n u^p]\frac{1}{(1-z)^2}\int_{t=0}^{z}\tilde{\Phi}(t,u,v)dt$$

$$= \frac{1}{n^p}[z^n]\frac{1}{(1-z)^2}\int_{t=0}^{z} N_u D_u^p \tilde{\Phi}(t,u,v)dt\left(1+\mathcal{O}\big(\frac{1}{n}\big)\right) \tag{26}$$

$$= \frac{1}{n^p}[z^n]\frac{1}{(1-z)^2}\int_{t=0}^{z}\left[\frac{(p-1)p!v\big(\frac{v}{2v-1}\big)^{2p-2}}{(1-z)^{p(2v-1)}}+\mathcal{O}\Big(\frac{\log(1-z)}{(1-z)^{(p-1)(2v-1)}}\Big)\right]dt\left(1+\mathcal{O}\big(\frac{1}{n}\big)\right).$$

We get via singularity analysis

$$[z^n]\frac{1}{(1-z)^2}\int_{t=0}^{z}\frac{(p-1)p!v\big(\frac{v}{2v-1}\big)^{2p-2}}{(1-z)^{p(2v-1)}}dt = [z^n]\frac{(p-1)p!v\big(\frac{v}{2v-1}\big)^{2p-2}}{(p(2v-1)-1)(1-z)^{p(2v-1)+1}}$$

$$= \frac{(p-1)p!v\big(\frac{v}{2v-1}\big)^{2p-2}n^{p(2v-1)}}{(p(2v-1)-1)\Gamma(p(2v-1)+1)}\left(1+\mathcal{O}\big(\frac{1}{n}\big)\right) \tag{27}$$

and

$$\begin{aligned}
&[z^n]\frac{1}{(1-z)^2}\int_{t=0}^{z}\mathcal{O}\Big(\frac{\log(1-t)}{(1-t)^{(p-1)(2v-1)}}\Big)dt \\
&= \mathcal{O}\bigg(\sum_{k=1}^{n}[z^{n-k}]\frac{1}{(1-z)^2}[z^k]\int_{t=0}^{z}\frac{\log(1-t)}{(1-t)^{(p-1)(2v-1)}}dt\bigg) \\
&= \mathcal{O}\left(n\max_{1\le k\le n}[z^{n-k}]\frac{1}{(1-z)^2}\max_{1\le k\le n}[z^k]\int_{t=0}^{z}\frac{\log(1-t)}{(1-t)^{(p-1)(2v-1)}}dt\right) \\
&= \mathcal{O}\left(n^2[z^n]\int_{t=0}^{z}\frac{\log(1-t)}{(1-t)^{(p-1)(2v-1)}}dt\right) = \mathcal{O}\left((\log n)\, n^{(p-1)(2v-1)}\right) \\
&= \mathcal{O}\Big(n^{p(2v-1)}\frac{1}{n^{1-\epsilon}}\Big),
\end{aligned} \tag{28}$$

uniformly for $\epsilon > 0$ and $|v-1| \le \tau := \frac{\epsilon}{3}$.

Thus we obtain by combining the results (26), (27) and (28) for $p \ge 2$ the asymptotic expansion

$$\frac{1}{\binom{n}{p}}[z^n u^p]F(z,u,v) = \frac{(p-1)p!v\big(\frac{v}{2v-1}\big)^{2p-2}n^{p(2v-2)}}{(p(2v-1)-1)\Gamma(p(2v-1)+1)}\Big(1+\mathcal{O}\big(\frac{1}{n^{1-\epsilon}}\big)\Big). \tag{29}$$

To apply the Quasi power theorem, we write (29) as

$$\frac{1}{\binom{n}{p}}[z^n u^p]F(z,u,v) = \exp\left[p(2v-2)\log n + \log\left(\frac{(p-1)p!v(\frac{v}{2v-1})^{2p-2}}{(p(2v-1)-1)\Gamma(p(2v-1)+1)}\right)\right] \times\left(1+\mathcal{O}\big(\frac{1}{n^{1-\epsilon}}\big)\right) \tag{30}$$

and then get, with the notations used in Theorem 3,

$$u(s) = p(2e^s-2), \quad v(s) = \log\left(\frac{(p-1)p!e^s\big(\frac{e^s}{2e^s-1}\big)^{2p-2}}{(p(2e^s-1)-1)\Gamma(p(2e^s-1)+1)}\right),$$
$$\phi(n) = \log n, \quad \kappa_n = n^{1-\epsilon}.$$

We have

$$u'(0) = 2p, \quad u''(0) = 2p, \tag{31}$$

and

$$v'(0) = -2p\Psi(p+1) + 3 - 2p - \frac{2p}{p-1} = -2pH_p + 2p\gamma + 3 - 2p - \frac{2p}{p-1},$$

$$v''(0) = -2p\Psi(p+1) - 4p^2\Psi'(p+1) + \frac{2(2p^3-5p^2+7p-2)}{(p-1)^2} \tag{32}$$
$$= -2pH_p + 2p\gamma - \frac{2}{3}\pi^2p^2 + 4p^2H_p^{(2)} + \frac{2(2p^3-5p^2+7p-2)}{(p-1)^2},$$

which leads now to the following result:

Theorem 5. *The distribution of the random variable $Y_{n,p}$, which counts the size of the spanning tree of p randomly chosen nodes in a binary search tree of size n is for fixed $p \geq 2$ asymptotically Gaussian, where the convergence rate is of order $\mathcal{O}\big(\frac{1}{\sqrt{\log n}}\big)$:*

$$\mathbb{P}\left\{\frac{Y_{n,p} - 2p\log n}{\sqrt{2p\log n}}\right\} = \Phi(x) + \mathcal{O}\Big(\frac{1}{\sqrt{\log n}}\Big),$$

and the expectation $E_{n,p} = \mathbb{E}(Y_{n,p})$ and the variance $V_{n,p} = \mathbb{E}(Y_{n,p})$ satisfy

$$E_{n,p} = 2p\log n - 2pH_p + 2p\gamma + 3 - 2p - \frac{2p}{p-1} + \mathcal{O}\big(\frac{1}{n^{1-\epsilon}}\big),$$
$$V_{n,p} = 2p\log n - 2pH_p + 2p\gamma - \frac{2}{3}\pi^2p^2 + 4p^2H_p^{(2)} + \frac{2(2p^3-5p^2+7p-2)}{(p-1)^2} + \mathcal{O}\big(\frac{1}{n^{1-\epsilon}}\big).$$

Of course, the case $p=1$ is trivial, since we have $\mathbb{P}\{Y_{n,1} = 1\} = 1$ due to the fact that the spanning tree of a single node is the node itself.

The case $p=2$ is of particular interest, since $Y_{n,2}$ is as described earlier, up to an additive constant, the distance Δ_n between two randomly selected nodes in a binary search tree of size n, viz. $\Delta_n \stackrel{\mathcal{L}}{=} Y_{n,2} - 1$. This parameter was studied already in [7], where the asymptotic normality of the distribution was shown by means of a refined contraction method.

References

[1] R. Dobrow, On the distribution of distances in recursive trees, *Journal of Applied Probability* 33, 749–757, 1996.

[2] Ph. Flajolet and A. Odlyzko, Singularity Analysis of Generating Functions, *SIAM Journal on Discrete Mathematics* 3, 216–240, 1990.

[3] D. Greene and D. Knuth, *Mathematics for the Analysis of Algorithms*, 2nd Ed., Birkhäuser, Boston, 1982.

[4] H.-K. Hwang, On Convergence Rates in the Central Limit Theorems for Combinatorial Structures, *European Journal of Combinatorics* 19, 329–343, 1998.

[5] S. Janson, The Wiener index of simply generated random trees, preprint, 2001.

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

[7] H. Mahmoud and R. Neininger, Distribution of Distances in Random Binary Search Trees, *The Annals of Applied Probability*, accepted for publication, 2002.

[8] R. Neininger, Wiener index of random trees, *Combinatorics, Probability & Computing*, accepted for publication, 2002.

[9] A. Panholzer and H. Prodinger, A Generating Functions Approach for the Analysis of Grand Averages for Multiple Quickselect, *Random Structures and Algorithms* 13, 189–209, 1998.

† Institut für Algebra und Computermathematik, Technische Universität Wien, Wiedner Hauptstrasse 8–10, A- 1040 Wien, Austria.

E-mail address: `Alois.Panholzer@tuwien.ac.at`

‡ The John Knopfmacher Centre for Applicable Analysis and Number Theory, School of Mathematics, University of the Witwatersrand, P.O. Wits, 2050 Johannesburg, South Africa.

E-mail address: `helmut@staff.ms.wits.ac.za`

WWW-address: `http://www.wits.ac.za/helmut/index.htm`