On the Internal Path Length of d–dimensional Quad Trees

Ralph Neininger and Ludger Rüschendorf
University of Freiburg

Abstract. It is proved that the internal path length of a d–dimensional quad tree after normalization converges in distribution. The limiting distribution is characterized as a fixed point of a random affine operator. We obtain convergence of all moments and of the Laplace transforms. The moments of the limiting distribution can be evaluated from the recursion and lead to first order asymptotics for the moments of the internal path lengths. The analysis is based on the contraction method. In the final part of the paper we state similar results for general split tree models if the expectation of the path length has a similar expansion as in the case of quad trees. This applies in particular to the m-ary search trees.

Key words: quad trees, contraction method, multidimensional data structure, analysis of algorithms, split trees, m-ary search trees

AMS subject classification: 68Q25, 60F05

1 Introduction

Quad trees are a classical data structure introduced by Finkel and Bentley [7] to store and retrieve data from some multidimensional data space that extends the familiar binary search tree for one dimensional data. Several characteristics of quad trees have been analysed in the standard random model which assumes that the data points are independent and identically distributed.

The mean depth has been found in Flajolet, Labelle, Laforest and Salvy [9];

$$\mathbb{E}D_n = \lambda_d \ln n + \mu_d + o(1) \tag{1}$$

with $\lambda_d = 2/d$ and μ_d has an explicit (more complicated) form which is e.g. for $d = 2$ given by $0.4105\ldots$ numerically. The first order asymptotic had been given before independently by Flajolet, Gonnet, Puech and Robson [8] and Devroye and Laforest [5]. The variance of D_n is of order $(2/d^2)\ln n$ (see Devroye and Laforest [5] for $d = 2$ and Flajolet and Lafforgue [10] for $d \geq 2$). In the last mentioned paper also asymptotic normality of D_n has been proved.

The expected height H_n still is of logarithmic order (see Devroye [2], [3])

$$\mathbb{E}H_n \sim \frac{c}{d}\ln n, \qquad c = 4.31107\ldots \tag{2}$$

and $H_n/\ln n \xrightarrow{P} c/d$ (where $\xrightarrow{P}$ denotes convergence in probability). The asymptotic variance and distribution are still unknown. For a presentation of basic results and an introduction to alternative data structures we refer to Mahmoud [12].

In this paper we investigate the asymptotics of the internal path length Y_n, i.e. of the sum of the levels of each node of the quad tree built from n random data. The asymptotics of the mean $\mathbb{E}Y_n$ results from (1)

$$\mathbb{E}Y_n = \frac{2}{d} n \ln n + \mu_d n + R_n \tag{3}$$

where $\lim R_n/n = 0$.

We will prove that the variance of Y_n is asymptotically of the form $\sim v_d n^2$ and obtain an explicit formula for v_d. We, therefore, introduce the normalized internal path length

$$X_n = \frac{Y_n - \mathbb{E}Y_n}{n}. \tag{4}$$

Our main result in this paper states that X_n converges weakly to a random variable X which is characterized as the fixed point of a random affine operator T. We also obtain convergence of all moments and of the Laplace

transform; X has exponential tails. It's moments are determined (in principle) by a recursion. We calculate the second moment of X which gives the asymptotic variance v_d of Y_n.

For the proof we extend the contraction method introduced by Rösler [16] for the analysis of Quicksort to the analysis of the internal path length of quad trees. It seems difficult to extend the martingale method which was used in Régnier [15] for the analysis of Quicksort. The contraction method has been further developed independently in Rösler [17] and Rachev and Rüschendorf [14]. The essential ingredient our proof uses is the recursion satisfied by Y_n plus the second order asymptotics of the first moments as in (3). We do not need a priori asymptotics of the second moments. The explanation of this behavior from the point of view of the contraction method is the fact, that the limiting operator T of the recursion of the Y_n (which exists by the asymptotics of $\mathbb{E}Y_n!$) has contraction properties w.r.t. the l_2–metric. This implies that only control of the first moments is necessary as is well known from the theory of probability metrics.

In the final part of the paper we consider general split tree models as introduced in Devroye [4]. The limit theorem for the quad trees extends to this class of split trees if the first moment of the path length has a similar expansion of the form $cn \ln n + dn + o(n)$, as in the case of quad trees. This method thus yields a limit theorem also for m-ary trees.

The authors would like to thank P. Flajolet for an essential hint to the limiting distribution.

2 Quad trees and the internal path length

A quad tree is constructed similarly to a binary search tree. Given a data vector $(p^{(1)}, \ldots, p^{(n)}), p^{(i)} \in \mathbb{R}^d$, the points $p^{(i)}$ build up successively the quad tree. Without loss of generality $p^{(i)} \in [0,1]^d$. Then the i–th data point $p^{(i)} \in [0,1]^d$ partitions the rectangle it belongs to into 2^d quadrants and thus creates 2^d new rectangles (quadrants) each having $p^{(i)}$ as a vertex. See Flajolet, Gonnet, Puech and Robson [8] or Mahmoud [12] for details of this contruction. We use a special ordering for the insertion of a new key $p \in [0,1]^d$ as in Flajolet et al. [8]. When comparing p with a node $w \in [0,1]^d$ of the quad tree we determine the number of the subtree in which p is inserted in the following way: The node w partitions the quadrant it belongs to into

2^d subquadrants. Let the number of a subquadrant be given by

$$\sum_{k=1}^{d} 2^{k-1} \mathbf{1}_{\{w_k \le s_k\}}, \qquad w = (w_i), s = (s_i) \tag{5}$$

if s is a point in this subquadrant. Now, p is inserted in the i–th subtree if it belongs to the i–th subquadrant. For the binary representation of $0 \le i \le 2^d - 1$

$$i = \sum_{k=1}^{d} a_k 2^{k-1}, \qquad a_k = a_k(i) \in \{0, 1\} \tag{6}$$

let

$$E(i) := \{k \in \{1 \ldots, d\} \,|\, a_k(i) = 1\},$$
$$N(i) := \{k \in \{1 \ldots, d\} \,|\, a_k(i) = 0\}.$$

Then equivalently, p is inserted in the i–th subtree of a node w if $p_k \ge w_k$ for all $k \in E(i)$ and $p_k < w_k$ for all $k \in N(i)$.

A random d–dimensional quad tree is a quad tree built up by inserting a sequence of points $(p^{(1)}, \ldots, p^{(n)})$ independent and identically distributed on $[0, 1]^d$. Note that the insertion depends only on the relative ranks of the components and so the results for random quad trees hold true for general multivariate distributions with continuous marginals and independent components.

For a random quad tree with n nodes let $I^{(n)} = (I_0^{(n)}, \ldots, I_{2^d-1}^{(n)})$ denote the number of nodes in the 2^d subtrees. Let $U = (U_1, \ldots, U_d)$ denote the first key to be inserted, i.e. $U_1, \ldots, U_d$ are independent, uniformly distributed on $[0, 1]$. Given $U = u = (u_1, \ldots, u_d)$, the volume of the i–th quadrant (in the above numbering) generated by U is given by

$$\langle u \rangle_i := \prod_{k \in N(i)} u_k \prod_{k \in E(i)} (1 - u_k) \quad (i = 0, \ldots, 2^d - 1). \tag{7}$$

Let $\langle u \rangle := (\langle u \rangle_0, \ldots, \langle u \rangle_{2^d-1})$ denote the vector of the generated volumes, then $I^{(n)}$ is conditionally given $U = u$ multinomial $M(n-1, \langle u \rangle)$ distributed:

$$I\!P^{I^{(n)} \,|\, U=u} = M(n - 1, \langle u \rangle). \tag{8}$$

In the following we denote convergence in probability and convergence in distribution by $\xrightarrow{P}$ and $\xrightarrow{\mathcal{D}}$ respectively and write $\stackrel{\mathcal{D}}{=}$ for equality in distribution

when either two random variables are compared or a random variable and a probability distribution are compared.

The conditional distribution in (8) implies the week law of large numbers for $I^{(n)}$.

Lemma 2.1 *The vector $I^{(n)}$ of the sizes of the subtrees satisfies*

$$I\!E\left(\frac{I_k^{(n)}}{n} - \langle U\rangle_k\right)^2 = \frac{I\!E\langle U\rangle_k(1-\langle U\rangle_k)}{n} \tag{9}$$

and

$$\frac{I^{(n)}}{n} \xrightarrow{P} \langle U\rangle = (\langle U\rangle_0, \ldots, \langle U\rangle_{2^d-1}) \tag{10}$$

where U is uniformly distributed on $[0,1]^d$.

From a geometrical point of view Lemma 2.1 says that the limiting distribution of $I^{(n)}/n$ is concentrated on a d–dimensional smooth surface embedded in a (2^d-1)–dimensional simplex in $I\!R^{2^d}$. In particular, for $d=2$ this surface is a hyperbolic paraboloid.

Since $0 \le \|I^{(n)}/n\| \le 1$, Lemma 2.1 implies the convergence of all moments. We shall need second moments in the following.

Corollary 2.2 *The asymptotic of the second moments of $I^{(n)}/n$ is given by*

$$\lim_{n\to\infty} I\!E\left(\frac{I_k^{(n)}}{n}\right)^2 = (1/3)^d. \tag{11}$$

Proof:

$$\begin{aligned}\lim_{n\to\infty} I\!E\left(\frac{I_k^{(n)}}{n}\right)^2 &= I\!E\langle U\rangle_k^2 = I\!E\langle U\rangle_0^2 = I\!E(U_1\cdot\ldots\cdot U_d)^2\\ &= (I\!EU_1^2)^d = (1/3)^d\end{aligned}$$

since $\langle U\rangle_k \stackrel{\mathcal{D}}{=} \langle U\rangle_0$ and the components are independent. ■

Let Y_n denote the internal path length of the random d–dimensional quad tree, i.e. Y_n is the sum of the depths of the nodes in the quad tree where

the depth of the root is defined to be one; then $Y_1 = 1$, $Y_2 = 3, \ldots$ Since the subtrees of a random quad tree are again random quad trees the following recursion for the internal path length holds in distribution

$$Y_n \stackrel{\mathcal{D}}{=} \sum_{k=0}^{2^d-1} Y^{(k)}_{I^{(n)}_k} + n \tag{12}$$

where $(Y_i^{(k)})$ are independent copies of Y_i and $\{(Y_i^{(k)}),\, k = 0, \ldots, 2^d - 1\}, I^{(n)}$ are independent. We define $Y_0 := 0$. The expectation of the internal path length Y_n is given in (3). The normalized version X_n of Y_n given by

$$X_n := \frac{Y_n - I\!E Y_n}{n} \tag{13}$$

satisfies the modified recursion

$$X_n \stackrel{\mathcal{D}}{=} \sum_{k=0}^{2^d-1} \frac{I^{(n)}_k}{n} X^{(k)}_{I^{(n)}_k} + C_n(I^{(n)}) \tag{14}$$

where $(X_i^{(k)})$ are independent copies of X_i, further $\{(X_i^{(k)}),\, k = 0, \ldots, 2^d - 1\}, I^{(n)}$ are independent and

$$C_n(i) := 1 + \frac{1}{n}\left(\sum_{k=0}^{2^d-1} I\!E Y_{i_k} - I\!E Y_n\right) \tag{15}$$

for $i = (i_0, \ldots, i_{2^d-1})$ with $\sum i_k = n - 1$.

3 Limit theorem for the internal path length

In order to obtain a limiting form of the recursion (14) we introduce the simplex

$$T_{2^d-1} := \{x \in [0,1]^{2^d} \mid \sum_{i=0}^{2^d-1} x_i = 1\} \tag{16}$$

and the entropy functional

$$C : T_{2^d-1} \to I\!R, \;\; C(x) := 1 + \frac{2}{d} \sum_{i=0}^{2^d-1} x_i \ln x_i \tag{17}$$

where $x \ln x$ is defined to be 0 for $x = 0$. Let

$$G_n := \{(i_0, \ldots, i_{2^d-1}) \in I\!N^{2^d} \mid \sum_{k=0}^{2^d-1} i_k = n-1\}$$

denote the domain of C_n, then as in Rösler [16] C approximates C_n in the following sense:

Lemma 3.1 *Let $(z^{(n)})$ be a sequence, $z^{(n)} \in G_n$ such that $z^{(n)}/n \to z \in (0,1]^{2^d}$, then*

$$\lim_{n\to\infty} C_n(z^{(n)}) = C(z). \tag{18}$$

Furthermore

$$\sup_{n\in I\!N} \|C_n\|_\infty < \infty. \tag{19}$$

Proof: Let $(z^{(n)})$ be a sequence, $(z^{(n)}) \in G_n$ such that $z^{(n)}/n \to z \in (0,1]^{2^d}$. Using the expansion (3) of the expectation of Y_n,

$$I\!E Y_n = \frac{2}{d} n \ln n + \mu_d n + R_n \quad \text{with} \quad R_n/n = o(1),$$

we obtain

$$\begin{aligned} C_n(z^{(n)}) &= 1 + \frac{1}{n}\left(\sum_{k=0}^{2^d-1} I\!E Y_{z_k^{(n)}} - I\!E Y_n\right) \qquad (20)\\ &= 1 + \frac{1}{n}\left(\sum_{k=0}^{2^d-1} \left(\frac{2}{d} z_k^{(n)} \ln z_k^{(n)} + \mu_d z_k^{(n)} + R_{z_k^{(n)}}\right) - \frac{2}{d} n \ln n - \mu_d n - R_n\right). \end{aligned}$$

Since $\sum_{k=0}^{2^d-1} z_k^{(n)} = n-1$ by definition of G_n,

$$\sum_{k=0}^{2^d-1} \frac{2}{d} z_k^{(n)} \ln z_k^{(n)} - \frac{2}{d} n \ln n = \sum_{k=0}^{2^d-1} \frac{2}{d} z_k^{(n)} \ln \frac{z_k^{(n)}}{n} - \frac{2}{d} \ln n. \tag{21}$$

With (20) and (21) we derive

$$\begin{aligned} C_n(z^{(n)}) &= 1 + \frac{1}{n}\left(\sum_{k=0}^{2^d-1} \left(\frac{2}{d} z_k^{(n)} \ln \frac{z_k^{(n)}}{n} + R_{z_k^{(n)}}\right) - \frac{2}{d} \ln n - \mu_d - R_n\right)\\ &= C(z^{(n)}/n) + \frac{1}{n} \sum_{k=0}^{2^d-1} R_{z_k^{(n)}} - \frac{1}{n}\left(\frac{2}{d} \ln n - \mu_d - R_n\right). \qquad (22) \end{aligned}$$

Now observe that

$$2\alpha := \min_{0\le k\le 2^d-1} z_k > 0$$

since $z \in (0,1]^{2^d}$. This implies $z_k^{(n)} \ge \alpha n$ for $0 \le k \le 2^d-1$ and n sufficiently large. Let $\bar{R}_n := \sup_{i\ge n} |R_i|$. Then $(\bar{R}_n)_{n\in I\!N}$ is decreasing and $\bar{R}_n/n \to 0$. For n sufficiently large it follows

$$\begin{aligned}\left|\frac{1}{n}\sum_{k=0}^{2^d-1} R_{z_k^{(n)}}\right| &\le \frac{1}{n}\sum_{k=0}^{2^d-1} \bar{R}_{z_k^{(n)}} \\ &\le \frac{1}{n}\sum_{k=0}^{2^d-1} \bar{R}_{\lfloor \alpha n\rfloor} \\ &\le 2^d\alpha\frac{1}{\lfloor \alpha n\rfloor}\bar{R}_{\lfloor \alpha n\rfloor} \to 0 \quad \text{for} \quad n\to\infty. \end{aligned} \tag{23}$$

The third summand of (22) also tends to zero. So (22) implies

$$C_n(z^{(n)}) = C(z^{(n)}/n) + o(1).$$

Therefore, continuity of C and the triangle inequality imply

$$\left|C_n(z^{(n)}) - C(z)\right| \le \left|C_n(z^{(n)}) - C(z^{(n)}/n)\right| + \left|C(z^{(n)}/n) - C(z)\right| \longrightarrow 0.$$

In order to get an estimate for $C_n(z^{(n)})$ uniformly in $z^{(n)} \in G_n$ let

$$L := \sup_{n\in I\!N} |R_n/n| < \infty.$$

Then (22) implies

$$\begin{aligned}|C_n(z^{(n)})| &\le |C(z^{(n)}/n)| + 2^d L + o(1) \\ &\le \|C\|_\infty + 2^d L + o(1). \end{aligned} \tag{24}$$

The second claim follows. ■

Lemmas 2.1, 3.1 suggest that a limit X of (X_n) is a solution of the limiting equation

$$X \stackrel{\mathcal{D}}{=} \sum_{k=0}^{2^d-1} \langle U\rangle_k X^{(k)} + C(\langle U\rangle) \tag{25}$$

where $X^{(k)}$ are iid copies of X and $\{X^{(k)}, k = 0, \dots 2^d - 1\}, U$ are independent, U uniformly distributed on $[0,1]^d$.

Define

$$M_{0,2} := \{\, \mu \in M^1(I\!R^1, \mathcal{B}^1) \,|\, I\!E\mu = 0,\ \mathrm{Var}\,\mu < \infty\} \tag{26}$$

where $I\!E\mu, \mathrm{Var}\,\mu$ are defined as expectation respectively variance of a corresponding random variable and $M^1(I\!R^1, \mathcal{B}^1)$ denotes the space of probability measures on the real line. Define the random affine operator

$$T : M^1(I\!R^1, \mathcal{B}^1) \to M^1(I\!R^1, \mathcal{B}^1), \quad T(\mu) \stackrel{\mathcal{D}}{=} \sum_{k=0}^{2^d-1} \langle U\rangle_k Z^{(k)} + C(\langle U\rangle) \tag{27}$$

where $(Z^{(k)}), U$ are independent, $Z^{(k)} \stackrel{\mathcal{D}}{=} \mu$ and U is uniformly distributed on $[0,1]^d$.

Our aim is to show that T is the limiting operator of the recursive sequence (X_n) in (14). Supply $M_{0,2} \subset M^1(I\!R^1, \mathcal{B}^1)$ with the minimal l_2–metric

$$l_2(\mu,\nu) = \inf\{(I\!E|X-Y|^2)^{1/2} \,:\, X \stackrel{\mathcal{D}}{=} \mu, Y \stackrel{\mathcal{D}}{=} \nu\}. \tag{28}$$

For random variables X, Y we use synonymously $l_2(X,Y) = l_2(I\!P^X, I\!P^Y)$. Then $(M_{0,2}, l_2)$ is a complete metric space and $l_2(\mu_n, \mu) \to 0$ is equivalent to

$$\mu_n \stackrel{\mathcal{D}}{\to} \mu \quad \text{and} \quad \int x^2 \,\mathrm{d}\mu_n(x) \to \int x^2 \,\mathrm{d}\mu(x). \tag{29}$$

(see Rachev [13])

Lemma 3.2 *$T : M_{0,2} \to M_{0,2}$ is a contraction w.r.t. l_2:*

$$l_2(T(\mu), T(\nu)) \le \left(\frac{2}{3}\right)^{d/2} l_2(\mu,\nu) \qquad \textit{for all } \mu,\nu \in M_{0,2}. \tag{30}$$

Proof: Obviously Var $(T(\mu)) < \infty$ and

$$\begin{aligned} I\!ET(\mu) &= 1 + \frac{2}{d} \sum_{k=0}^{2^d-1} I\!E\,[\langle U\rangle_k \ln\langle U\rangle_k] \\ &= 1 + \frac{2}{d}\, 2^d\, I\!E\,[\langle U\rangle_0 \ln\langle U\rangle_0] \\ &= 1 + \frac{2}{d}\, 2^d \int_{[0,1]^d} u_1 \cdots u_d \ln(u_1 \cdots u_d)\ \mathrm{d}\lambda^d(u) \end{aligned}$$

$$
\begin{aligned}
&= 1 + 2^{d+1} \left(\int_0^1 u \, \mathrm{d}u \right)^{d-1} \int_0^1 u \ln u \, \mathrm{d}u \\
&= 1 + 4(-1/4) = 0 \qquad (31)
\end{aligned}
$$

so T is a well defined mapping $T : M_{0,2} \to M_{0,2}$.

To prove contractivity let $\mu, \nu \in M_{0,2}$ and let $(V^{(k)}, W^{(k)}), U$ be independent, U uniformly distributed on $[0,1]^d$. Let $(V^{(k)}, W^{(k)})$ be optimal l_2–couplings of (μ, ν), i.e. $V^{(k)} \stackrel{\mathcal{D}}{=} \mu, W^{(k)} \stackrel{\mathcal{D}}{=} \nu$ and $l_2^2(\mu,\nu) = I\!E(V^{(k)} - W^{(k)})^2$. Then using the independence properties and $I\!EV^{(k)} = I\!EW^{(k)} = 0$

$$
\begin{aligned}
l_2^2(T(\mu), T(\nu)) &= l_2^2 \left(\sum_{k=0}^{2^d-1} \langle U \rangle_k V^{(k)} + C(\langle U \rangle), \sum_{k=0}^{2^d-1} \langle U \rangle_k W^{(k)} + C(\langle U \rangle) \right) \\
&\leq I\!E \left(\sum_{k=0}^{2^d-1} \langle U \rangle_k (V^{(k)} - W^{(k)}) \right)^2 \\
&= \sum_{k=0}^{2^d-1} I\!E \left[\langle U \rangle_k^2 (V^{(k)} - W^{(k)})^2 \right] \\
&= 2^d \cdot I\!E \langle U \rangle_0^2 \cdot l_2^2(\mu,\nu) \\
&= \left(\frac{2}{3} \right)^d l_2^2(\mu,\nu) \qquad (32)
\end{aligned}
$$

so T is a contraction on $M_{0,2}$. ■

By Banach's fixed point theorem T has a unique fixed point ρ in $M_{0,2}$ and

$$
l_2(T^n(\mu), \rho) \to 0 \qquad (33)
$$

exponentially fast for any $\mu \in M_{0,2}$.

We call a random variable X with distribution ρ also a fixed point of T. (compare equation (25))

Theorem 3.3 (Limit theorem for the internal path length) *The normalized internal path length X_n of a random quad tree converges w.r.t. l_2 to the unique fixed point X in $M_{0,2}$ of the limiting operator T, i.e.*

$$
l_2(X_n, X) \to 0. \qquad (34)
$$

Proof: Let $X_n^{(k)} \stackrel{\mathcal{D}}{=} X_n, X^{(k)} \stackrel{\mathcal{D}}{=} X, 0 \leq i \leq 2^d - 1$ such that $(X_n^{(k)}, X^{(k)})$ are optimal couplings of X_n, X, i.e. $l_2^2(X_n, X) = I\!E(X_n^{(k)} - X^{(k)})^2$. Furthermore let $I^{(n)}$ be conditionally given $U = u$ multinomial $M(n-1, \langle u \rangle)$ distributed and by Lemma 2.1 assume w.l.o.g. that $I^{(n)}/n \to \langle U \rangle$ a.s., U uniformly distributed on $[0,1]^d$. Finally assume that $((X_n^{(0)})_{n \in I\!N}, X^{(0)}), \ldots, ((X_n^{(2^d-1)})_{n \in I\!N}, X^{(2^d-1)}), (I^{(n)}/n, U)$ are independent. Then using the independence properties and that $I\!EX^{(k)} = I\!EX_n^{(k)} = 0$ we obtain

$$\begin{aligned}
l_2^2(X_n, X) &= l_2^2 \left(\sum_{k=0}^{2^d-1} \frac{I_k^{(n)}}{n} X^{(k)}_{I_k^{(n)}} + C_n(I^{(n)}), \sum_{k=0}^{2^d-1} \langle U \rangle_k X^{(k)} + C(\langle U \rangle) \right) \\
&\leq I\!E \left(\sum_{k=0}^{2^d-1} \left(\frac{I_k^{(n)}}{n} X^{(k)}_{I_k^{(n)}} - \langle U \rangle_k X^{(k)} \right) + C_n(I^{(n)}) - C(\langle U \rangle) \right)^2 \\
&= I\!E \left[\sum_{k=0}^{2^d-1} \left(\frac{I_k^{(n)}}{n} X^{(k)}_{I_k^{(n)}} - \langle U \rangle_k X^{(k)} \right)^2 + \Big(C_n(I^{(n)}) - C(\langle U \rangle) \Big)^2 \right] \\
&= \sum_{k=0}^{2^d-1} I\!E \left(\frac{I_k^{(n)}}{n} X^{(k)}_{I_k^{(n)}} - \langle U \rangle_k X^{(k)} \right)^2 + I\!E \Big(C_n(I^{(n)}) - C(\langle U \rangle) \Big)^2. \quad (35)
\end{aligned}$$

By Lemma 2.1 respectively dominated convergence and Lemma 3.1 as $n \to \infty$

$$I\!E(I_k^{(n)}/n - \langle U \rangle_k)^2 \to 0 \qquad \text{and} \quad (36)$$

$$I\!E(C_n(I^{(n)}) - C(\langle U \rangle))^2 \to 0. \quad (37)$$

For the first term of (35) consider

$$\begin{aligned}
& I\!E \left(\frac{I_k^{(n)}}{n} X^{(k)}_{I_k^{(n)}} - \langle U \rangle_k X^{(k)} \right)^2 \\
= \quad & I\!E \left(\frac{I_k^{(n)}}{n} \left(X^{(k)}_{I_k^{(n)}} - X^{(k)} \right) + \left(\frac{I_k^{(n)}}{n} - \langle U \rangle_k \right) X^{(k)} \right)^2 \\
\leq \quad & I\!E \left(\frac{I_k^{(n)}}{n} \left(X^{(k)}_{I_k^{(n)}} - X^{(k)} \right) \right)^2 + I\!E \left(\left(\frac{I_k^{(n)}}{n} - \langle U \rangle_k \right) X^{(k)} \right)^2
\end{aligned}$$

$$+2 \cdot I\!E \left[\frac{I_k^{(n)}}{n} \left(X^{(k)}_{I_k^{(n)}} - X^{(k)} \right) \left(\frac{I_k^{(n)}}{n} - \langle U \rangle_k \right) X^{(k)} \right]. \tag{38}$$

By independence and (36) the second term in (38) converges to zero. With the Cauchy-Schwarz-inequality the third term in it's absolute value is estimated from above by

$$\begin{aligned} & 2I\!E \left[\left(\frac{I_k^{(n)}}{n} \right)^2 \left(\frac{I_k^{(n)}}{n} - \langle U \rangle_k \right)^2 \left(X^{(k)} \right)^2 \right] I\!E \left(X^{(k)}_{I_k^{(n)}} - X^{(k)} \right)^2 \\ = \;& o(1) I\!E \left(X^{(k)}_{I_k^{(n)}} - X^{(k)} \right)^2, \end{aligned} \tag{39}$$

where again (36) has been used. With (35)–(39) and denoting $a_n := l_2^2(X_n, X)$ we derive

$$\begin{aligned} a_n &\leq 2^d I\!E \left(\left(\frac{I_k^{(n)}}{n} + o(1) \right) \left(X^{(k)}_{I_k^{(n)}} - X^{(k)} \right) \right)^2 + b_n \\ &= 2^d \sum_{i=0}^{n-1} I\!P(\{(I_k^{(n)}/n) = (i/n)\}) \\ &\qquad \times ((i/n)^2 + o(1))\, I\!E(X_i^{(k)} - X^{(k)})^2 + b_n \end{aligned} \tag{40}$$

where $b_n \to 0$ for $n \to \infty$. From Corollary 2.2 we conclude

$$\begin{aligned} a_n &\leq 2^d \sum_{i=1}^{n-1} I\!P(\{(I_k^{(n)}/n) = (i/n)\})\, ((i/n)^2 + o(1)) \sup_{1 \leq i \leq n-1} a_i + b_n \\ &= ((2/3)^d + o(1)) \sup_{1 \leq i \leq n-1} a_i + b_n \end{aligned} \tag{41}$$

which implies that (a_n) is bounded. This implies as in Rösler [16] that for a given $\epsilon > 0$ there exists n_0 such that for $n \geq n_0$

$$a_n \leq a + \epsilon \qquad \text{with} \qquad a := \limsup_{n \to \infty} a_n$$

and the prefactor in (41) is uniformly less than a $\gamma < 1$. Therefore

$$a_n \leq 2^d \sum_{i=1}^{n_0-1} I\!P\left(\frac{I_k^{(n)}}{n} = \frac{i}{n} \right) \left(\left(\frac{i}{n} \right)^2 + o(1) \right) a_i$$

$$+2^d \sum_{i=n_0}^{n-1} I\!P\left(\frac{I_k^{(n)}}{n} = \frac{i}{n}\right)\left(\left(\frac{i}{n}\right)^2 + o(1)\right)(a+\epsilon) + b_n$$

$$\leq \quad \gamma(a+\epsilon) + o(1). \tag{42}$$

Then $0 \leq a = \limsup a_n \leq \gamma(a+\epsilon)$ which implies $a = 0$. ■

4 Moments and tail of the internal path length

Let X be the unique solution of the fixed point equation (25) for the internal path length in $M_{0,2}$

$$X \stackrel{\mathcal{D}}{=} \sum_{k=0}^{2^d-1} \langle U \rangle_k X^{(k)} + C(\langle U \rangle) \tag{43}$$

where $X^{(k)}$ are iid copies of X and $\{X^{(k)}, k = 0, \ldots 2^d-1\}, U$ are independent, U uniformly distributed on $[0,1]^d$. Then (43) implies recursive equations for the moments $I\!EX^k$ of X which can be solved (in principle) as soon as we know the existence of higher order moments. For the variance of X we obtain (note that $I\!EX = 0$)

Proposition 4.1 (Variance of X) *The variance of the limit X of the normalized internal path length of a d–dimensional random quad tree is given by*

$$v_d = \frac{21 - 2\pi^2}{9d(1-(2/3)^d)}. \tag{44}$$

In particular

$$v_1 = 0.4202\ldots, \quad v_2 = 0.1260\ldots, \quad v_3 = 0.0663\ldots$$

Proof: (43) and the independence properties imply

$$I\!EX^2 = \left(1 - \left(\frac{2}{3}\right)^d\right)^{-1} I\!EC^2(\langle U \rangle).$$

By calculation as in the proof of Lemma 3.2

$$\begin{aligned} I\!E C^2(\langle U\rangle) &= -1 + \frac{4}{d^2} I\!E \left(\sum_{i=0}^{2^d-1} \langle U\rangle_i \ln\langle U\rangle_i \right)^2 \\ &= -1 + \frac{4}{d^2} \sum_{i,j=0}^{2^d-1} I\!E[\langle U\rangle_i \langle U\rangle_j \ln\langle U\rangle_i \ln\langle U\rangle_j]. \end{aligned}$$

The distribution of the factors $\langle U\rangle_i \langle U\rangle_j \ln\langle U\rangle_i \ln\langle U\rangle_j$ only depends on the number of digits in which the dual representations of i and j differ (see (5), (6)). Therefore

$$I\!E \left(\sum_{i=0}^{2^d-1} \langle U\rangle_i \ln\langle U\rangle_i \right)^2 = \sum_{h=0}^{d} 2^d \binom{d}{h} l_h. \tag{45}$$

l_h can be calculated by first applying the functional equation of the logarithm. This yields d^2 terms of the form

$$\int_{[0,1]^d} \prod_{i=1}^{d-h} u_i^2 \prod_{i=d-h+1}^{d} (u_i(1-u_i)) \ln \tilde{u}_k \ln \tilde{u}_l \, \mathrm{d}\lambda^d(u)$$

with $\tilde{u}_k = u_k$ for $k \le d-h$ and $\tilde{u}_k = 1-u_k$ for $k > d-h$. Then distinguish the cases $1 \le k,l \le d-h$ and $d-h+1 \le k,l \le d$ for $k=l$ and $k \ne l$ and finally $1 \le k \le d-h < l \le d$. The arising integrals can be calculated elementary. This implies the representation

$$v_d = \left(1 - \left(\frac{2}{3}\right)^d\right)^{-1} \left[-1 + \frac{4}{d^2} \left(\frac{2}{3}\right)^d \sum_{h=0}^{d} \binom{d}{h} \left(\frac{1}{2}\right)^h s_h \right] \tag{46}$$

where

$$s_h = \left(\frac{d}{3} - \frac{h}{2}\right)^2 + \frac{d}{9} + \left(\frac{5}{4} - \frac{\pi^2}{6}\right) h.$$

Now a simplification with the help of Maple leads to the stated variance. ■

l_2-convergence implies convergence of second order moments. We obtain as Corollary the first order asymptotics of the variance of the internal path length Y_n.

Corollary 4.2

$$Var\,(Y_n) \sim v_d n^2 \tag{47}$$

with v_d given in (44).

In the Quicksort case $d = 1$ Rösler [16] showed finiteness of the Laplace transform of X and convergence of the Laplace transforms of X_n to that of X. In particular this implies finiteness and convergence of higher order moments. Röslers proof directly extends to the case $d \geq 1$. Lemma 4.1 in Rösler [16] holds in any dimension as follows.

Lemma 4.3 $\forall\, L > 0 :\ \exists\, K_L > 0 :\ \forall\, n \in I\!N$ *and* $\forall\, \lambda \in [-L, L]$ *holds*

$$I\!E \exp(\lambda X_n) \leq \exp(\lambda^2 K_L). \tag{48}$$

Proof: In place of the random variable U_n in Röslers proof use

$$V_n := \|I^{(n)}/n\|^2 - 1.$$

Then

a) $-1 \leq V_n < 0$ for all $n \in I\!N$

b) $\sup\limits_{n \in I\!N} I\!E V_n < 0$

c) $\sup\limits_{n \in I\!N} \|C_n\|_\infty < \infty$ by Lemma 3.1.

For the proof of b) note that $I\!E V_n < 0$ for all $n \in I\!N$ and $V_n \xrightarrow{\mathcal{D}} \|\langle U \rangle\|^2 - 1$ which implies by boundedness of V_n, $I\!E V_n \to I\!E(\|\langle U \rangle\|^2 - 1) < 0$. From a)–c) one obtains (48) as in Rösler [16]. ■

Theorem 4.4 (Convergence of Laplace transforms) *For the normalized internal path length X_n holds*

$$I\!E \exp(\lambda X_n) \to I\!E \exp(\lambda X), \qquad \lambda \in I\!R^1. \tag{49}$$

Proof: The exponential bound in (48) implies uniform integrability of $\exp(\lambda X_n)$ which by Theorem 3.3 yields (49). ■

Finally using the expansion of the mean $I\!E Y_n$ in (3) one obtains as in Corollary 4.3 of Rösler [16] the following bounds for (large) deviations.

Corollary 4.5 *Let Y_n denote the internal path length of a d–dimensional quad tree. Then for any $\lambda, \epsilon > 0$ there exists $C_{\lambda,\epsilon} > 0$ such that for all $n \in \mathbb{N}$*

$$\mathbb{P}(|Y_n - \mathbb{E}Y_n| \geq \epsilon \mathbb{E}Y_n) \leq C_{\lambda,\epsilon} n^{-(2\lambda\epsilon)/d}.$$

Equivalently

$$\mathbb{P}(|Y_n - \mathbb{E}Y_n| \geq \epsilon \mathbb{E}Y_n) = O(n^{-k})$$

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

5 Extension to a general split tree model

Several further random trees lead to a type of recursion for the internal path length similar to the recursion (12) for the random quad tree. For other characteristic quantities as the depth of insertion of a key or the height of a tree Devroye [4] gave a uniform treatment for a rather general model of a random tree which he calls the random split tree. This model contains many common trees, e.g. the random binary search tree, the m-ary search tree, the random quad tree ... A related model for a general class of random trees is discussed in Aldous [1]. Devroye's random split tree is determined by a fixed branch factor $b > 0$, the number $s_0 \geq 0$ of keys contained in an internal node (usually $s_0 = 1$, but for the m-ary tree we have $s_0 = m - 1$ keys in an internal node; in the following we assume $s_0 \geq 1$) and a split vector $\mathcal{V} = (V_1, \ldots, V_b)$ of random probabilities, $\sum V_k = 1, V_k \geq 0$ which controls the splitting process during the insertions of keys independently at each node together with some further parameters. For details see Devroye [4].

For such a general type of random split tree the internal path length Y_n satisfies the recursion

$$Y_n \stackrel{\mathcal{D}}{=} \sum_{k=1}^{b} Y^{(k)}_{I^{(n)}_k} + n \tag{50}$$

where $(Y_i^{(k)})$ are independent copies of Y_i, $\{(Y_i^{(k)}), k = 1, \ldots, b\}, I^{(n)}$ are independent and $I^{(n)}$ (the vector of the cardinalities of the subtrees) is conditionally given $\mathcal{V} = (v_1, \ldots, v_b)$ multinomial $\mathrm{M}(n - s_0, v_1, \ldots, v_b)$ distributed. Here $\mathcal{V}$ is the split vector controlling the splitting process at the root. Now

the question arises under which conditions on $\mathcal{V}$ a limit theorem for the internal path length of a random split tree holds. We can't solve this problem in general. Inspecting our proof for the case of the random quad tree from section 3 we will explain that similar limit theorems as in the case of quad trees hold true if the first moment $\mathbb{E}Y_n$ admits an expansion of the form

$$\mathbb{E}Y_n = cn \ln n + dn + o(n) \tag{51}$$

with $c > 0$ and $d \in \mathbb{R}$. In particular this type of expansion implies that the tree is well balanced like random binary trees. For an example which leads to a different order of expansion see Devroye [4] and the references given therein.

Assume (51) is valid for a random split tree with split vector $\mathcal{V} = (V_1, \ldots, V_b)$. Then the normalized internal path length

$$X_n := \frac{Y_n - \mathbb{E}Y_n}{n}$$

analogously to (14) satisfies the modified recursion

$$X_n \stackrel{\mathcal{D}}{=} \sum_{k=1}^{b} \frac{I_k^{(n)}}{n} X_{I_k^{(n)}}^{(k)} + C_n(I^{(b)}). \tag{52}$$

Here $(X_i^{(k)})$ are i.i.d. copies of X_i, further $\{(X_i^{(k)}), k = 1, \ldots, b\}, I^{(n)}$ are independent and

$$C_n(i) := 1 + \frac{1}{n}\left(\sum_{k=1}^{b} \mathbb{E}Y_{i_k} - \mathbb{E}Y_n\right) \tag{53}$$

for $i = (i_1, \ldots, i_b)$ with $\sum i_k = n - s_0$ (cf. (15)). The entropy functional

$$C : T_{b-1} \to \mathbb{R}, \qquad C(x) := 1 + c \sum_{k=1}^{b} x_i \ln x_i \tag{54}$$

approximates C_n in the sense of Lemma 3.1. Here the expansion (51) is used. The constant c in (54) is identical to the leading constant in (51). Therefore the limiting equation for the normalized internal path length is given by

$$X \stackrel{\mathcal{D}}{=} \sum_{k=1}^{b} V_k X^{(k)} + C(\mathcal{V}) \tag{55}$$

where $X^{(k)}$ are i.i.d. copies of X, $X^{(1)}, \ldots, X^{(b)}, \mathcal{V}$ are independent and $\mathcal{V}$ is a split vector. The associated random affine operator (cf. (27)) similarly to the proof of Lemma 3.2 turns out to be a contraction on $M_{0,2}$ w.r.t. l_2 with contraction factor

$$\left(I\!E \sum_{k=1}^{b} V_k^2 \right)^{1/2} =: \gamma^{1/2} < 1. \tag{56}$$

(By $\sum V_k = 1, V_k \geq 0$ we deduce $I\!E \sum V_k^2 \leq 1$. The case $I\!E \sum V_k^2 = 1$ corresponds to a degenerated tree contradicting (51).) Also the limit theorem corresponding to Theorem 3.3 can be established. Observe that the prefactor in (42) is given in general using an analogue of Corollary 2.2 by

$$\begin{aligned} I\!E \sum_{k=1}^{b} \left(\frac{I^{(n)}}{n} \right)^2 &= I\!E \sum_{k=1}^{b} V_k^2 + o(1) \\ &= \gamma + o(1) < 1 \end{aligned} \tag{57}$$

for n sufficiently large.

Further the results of section 4 concerning the Laplace transform, higher order moments and large deviation of the internal path length hold true in this general setting. Alltogether we can formulate the following limit theorem for general split tree models.

Theorem 5.1 (Limit theorem for the path length of split trees) *Let Y_n denote the internal path length of a general split tree model with split vector $\mathcal{V} = (V_1, \ldots, V_b)$. Assume that $I\!EY_n$ has the expansion*

$$I\!EY_n = cn \ln n + dn + o(n),$$

and define $X_n := (Y_n - I\!EY_n)/n$, then

(a) $l_2(X_n, X) \to 0$ *where X is the unique solution in $M_{0,2}$ of the fixed point equation*

$$X \stackrel{\mathcal{D}}{=} \sum_{k=1}^{b} V_k X^{(k)} + C(\mathcal{V}) \qquad \textit{(cp. (55))}$$

with C given in (54),

(b) *exponential moments exist and converge,*

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

(c) $\mathbb{P}(|Y_n - \mathbb{E}Y_n| \geq \epsilon \mathbb{E}Y_n) = O(n^{-k})$ *for all* $k \in \mathbb{N}$.

As a consequence we obtain as in Proposition 4.1, Corollary 4.2 an expansion of the variance of first order, i.e. Var $Y_n \sim vn^2$ as $n \to \infty$.

Therefore it is a challenging task to identify those split vectors $\mathcal{V} = (V_1, \ldots, V_b)$ which induce an expansion (51) for the mean of the internal path length.

A new and general approach to this problem was given recently by Rösler [18] using renewal theory. In particular Rösler derives an expansion (51) for the internal path length of the random median of $(2k+1)$-tree which leads to the limit law for this kind of tree. Another example which fits not exactly in the model of a random split tree but is of similar type is the random recursive tree. The recursion for the path length X_n of the random recursive tree is of the slightly modified form

$$X_n = X_K^{(1)} + X_{n-K}^{(2)} + K.$$

$(X_i^{(k)})$ are i.i.d. copies of X_i, $(X_i^{(1)}), (X_i^{(2)}), K$ are independent and K is uniformly distributed on $\{1, \ldots, n-1\}$. For this tree the limit law for X_n was proved by a similar method in Dobrow and Fill [6]. In this paper the authors also derive explicitly the higher moments of the limiting distribution in terms of the ζ-function.

Finally we remark that for the random m-ary search tree an expansion for the mean of the internal path length Y_n is known. In Mahmoud [11] the expansion

$$\mathbb{E}Y_n = \frac{1}{H_m - 1} H_n(n+1) + c_m n + O(n^\beta) \tag{58}$$

with $\beta < 1$ is given. Here H_n denotes the nth harmonic number, $H_n = \sum_{i=1}^n 1/i$. Substituting $H_n = \ln n + \gamma + O(1/n)$ in (58) with γ being Euler's constant $\mathbb{E}Y_n$ is of the form (51) with leading constant $c = 1/(H_m - 1)$. The split vector $\mathcal{V} = (V_1, \ldots, V_b)$ is given by the spacings of $m-1$ i.i.d.

random variables uniformly distributed on $[0,1]$. For $U_1, \ldots, U_{m-1}$ i.i.d. and uniformly distributed on $[0,1]$ denote by $U_{(1)}, \ldots, U_{(m-1)}$ the order statistics of $U_1, \ldots, U_{m-1}$. Then

$$\mathcal{V} \stackrel{\mathcal{D}}{=} (U_{(1)}, U_{(2)} - U_{(1)}, \ldots, U_{(m-2)} - U_{(m-1)}, 1 - U_{(m-1)}).$$

For the normalized internal path length $X_n := (Y_n - I\!EY_n)/n$ it follows:

Corollary 5.2 *The normalized internal path length X_n of a random m-ary search tree converges w.r.t. l_2 to the unique fixed point X in $M_{0,2}$ of the limiting equation*

$$X \stackrel{\mathcal{D}}{=} \sum_{k=1}^{m} V_k X^{(k)} + C(\mathcal{V}) \tag{59}$$

where $X^{(k)}$ are i.i.d. copies of X, $X^{(1)}, \ldots, X^{(b)}, \mathcal{V}$ are independent and $\mathcal{V}$ is the vector of spacings of $m-1$ independent random variables uniformly distributed on $[0,1]$. The entropy functional C in (59) is given by

$$C : T_{m-1} \to I\!R, \qquad C(x) := 1 + \frac{1}{H_m - 1} \sum_{k=1}^{m} x_i \ln x_i. \tag{60}$$

In principle higher moments can be calculated from the fixed point equation (59). The first order asymptotic for the second order moment of the path length of m-ary search trees has already been achieved by generating function methods (cf. Mahmoud [12, page 142]).

References

[1] Aldous, D. 1996 Probability Distributions on Cladograms. *Random Discrete Structures*, (D. Aldous and R. Pemantle, eds.) Springer (IMA Volumes Math. Appl. 76), 1-18.

[2] Devroye, L. 1986 A note on the expected height of binary search trees. *Journal of the ACM* **33**, 489-498.

[3] Devroye, L. 1987 Branching processes in the analysis of the heights of trees. *Acta Informatica* **24**, 277-298.

[4] Devroye, L. 1999 Universal limit laws for depths in random trees. *SIAM Journal on Computing* **28**, 409-432.

[5] Devroye, L. & Laforest, L. 1990 An analysis of random d–dimensional quad trees. *SIAM Journal on Computing* **19**, 821-832.

[6] Dobrow, R.P. & Fill, J.A. 1999 Total path length for random recursive trees. To appear in *Combinatorics, Probability and Computing*.

[7] Finkel, R.A. & Bentley, J.L. 1974 Quad trees, a data structure for retrieval on composite keys. *Acta informatica* **4**, 1-9.

[8] Flajolet, P., Gonnet, G., Puech, C. & Robson, J.M. 1993, Analytic Variations on Quadtrees. *Algorithmica* **10**, 473-500.

[9] Flajolet, P., Labelle, G., Laforest, L. & Salvy, B. 1995, Hypergeometrics and the cost structure of Quadtrees. *Random Struct. Algorithms* **7**, 117-144.

[10] Flajolet, P. & Lafforgue, T. 1994 Search costs in quadtrees and singularity perturbation asymptotics. *Discrete and Computational Geometry* **12**, 151-175.

[11] Mahmoud, H. 1986 On the Average Internal Path Length of m-ary Search Trees. *Acta Informatica* **23**, 111-117.

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

[13] Rachev, S.T. 1984 The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory Prob. Appl.* **29**, 647-676.

[14] Rachev, S.T. & Rüschendorf, L. 1995 Probability metrics and recursive algorithms. *Adv. Appl. Prob.* **27**, 770-799.

[15] Régnier, M. 1989 A limiting distribution for quicksort. *RAIRO, Theoretical Informatics and Applications* **23**, 335-343.

[16] Rösler, U. 1991 A limit theorem for "QUICKSORT". *RAIRO, Theoretical Informatics and Applications* **25**, 85-100.

[17] Rösler, U. 1992 A fixed point theorem for distributions. *Stoch. Proc. Appl.* **42**, 195-214.

[18] Rösler, U. 1998 On the analysis of stochastic divide and conquer algorithms. Preprint.

Authors' address:

Institut für Mathematische Stochastik, Universität Freiburg
Eckerstr. 1, D–79104 Freiburg
Germany
e-mail: rn@stochastik.uni-freiburg.de
ruschen@stochastik.uni-freiburg.de