SIAM J. COMPUT.
Vol. 19, No. 5, pp. 821-832, October 1990

004

AN ANALYSIS OF RANDOM d-DIMENSIONAL QUAD TREES*

LUC DEVROYE† AND LOUISE LAFOREST‡

Abstract. It is shown that the depth of the last node inserted in a random quad tree constructed from independent uniform $[0, 1]^d$ random vectors is in probability asymptotic to $(2/d) \log n$, where log denotes the natural logarithm. In addition, for $d = 2$, exact values are obtained for all the moments of the depth of the last node.

Key words. average time analysis, probability ineq· alities, random quad tree, multidimensional data structures, search tree, expected behavior, analysis of a ;orithms

AMS(MOS) subject classifications. 68P05, 68Q25

C.R. categories. 3.74, 5.25, 5.5

1. Introduction. Various data structures have been proposed for retrieval on composite keys (or associative retrieval) such as $k-d$ trees (Bentley (1975)), multidimensional trees (Rivest (1974); Orenstein (1982)); grids with variable-sized cells (Tamminen (1981), (1982)); quad trees (Finkel and Bentley (1974)); $k-d-b$ trees (Robinson (1981)); quintary trees (Lee and Wong (1981)); and multipaging structures (Merrett and Otoo (1981)). A partial survey of these structures can be found in Tamminen (1981) or Gonnet (1984). In this paper, we analyze random quad trees. These trees have been used with a great deal of success in computer graphics (see Woodwark (1982) and the references found there) and image processing (Hunter and Steiglitz (1979)). Detailed discussions of some common operations on quad trees, and possible improvements, can be found in Bentley, Stanat, and Williams (1977), and Samet (1980). See also the survey article by Samet (1984). The quad trees considered here are known as point quad trees since they are used to store points. Many applications require region quad trees for storing screenfuls of pixels. Random region quad trees were analyzed for example by Puech and Yahia (1985).

A **quad tree** is constructed as a binary search tree. When a key X_i occupies a node, it partitions the rectangle it belongs to orthogonally into 2^d parts (called quadrants), and thus creates 2^d new rectangles, each having X_i as a vertex. We should note here that the traversal of one node requires d comparisons. A **random quad tree** is constructed by inserting $X_1, \cdots, X_n$, independently and identically distributed uniform $[0, 1]^d$ random vectors, in the standard manner into an initially empty quad tree. We will look at D_n, the **depth** of X_n after it is inserted into the tree, where, by convention, the depth of the root is zero. The level L_n of a node is equal to its depth plus one. Other important quantities are the **average depth** $A_n = (1/n) \sum_{i=1}^n D_i$ and the **height** $\max_{1 \le i \le n} D_i$. The height is in probability asymptotic to $(c/d) \log n$, where $c = 4.31107 \cdots$ is the unique solution greater than two of the equation $c \log (2e/c) = 1$ (Devroye (1987)). However, unsuccessful search times are in most cases appropriately

* Received by the editors April 27, 1988; accepted for publication (in revised form) September 7, 1989. This research was sponsored by Natural Sciences and Engineering Research Council of Canada grant A3456 and by FCAR grant EQ-1678.

† School of Computer Science, McGill University, 805 Sherbrooke Street West, Montréal, Canada H3A 2K6.

‡ Département de Mathématique et d'Informatique, Université du Québec à Montréal, Montréal, Québec Canada H3C 3P8.

measured by D_n, the depth of the last node added to the tree. Our main result is the following.

THEOREM M1. *$D_n/\log n$ tends in probability to $2/d$ as $n \to \infty$. Also, $\mathbf{E}D_n \sim \mathbf{E}A_n \sim (2/d)\log n$ as $n \to \infty$.*

For $d = 1$, the quad tree reduces to the binary search tree, and the random quad tree coincides with a binary search tree constructed from a random equiprobable permutation. Its properties, including the law of large numbers given in Theorem M1, have been obtained in a series of papers by Lynch (1965), Knuth (1973), Robson (1979), Sedgewick (1983), Pittel (1984), Mahmoud and Pittel (1984), Brown and Shubert (1984), Louchard (1987), and Devroye (1986), (1987), (1988).

In § 3, we will derive large deviation inequalities for D_n. In effect, we will prove the following theorem.

THEOREM M2. *For every $\delta > 0$, there exist positive constants a, b such that*

$$\mathbf{P}\left(\left|\frac{D_n}{(2/d)\log n} - 1\right| > \delta\right) \leqq an^{-b}.$$

The extra material needed to prove this is presented in § 2. In § 4, for the planar case ($d = 2$), we obtain exact values for $\mathbf{E}D_n$ and $\mathrm{Var}(D_n)$. Both are of the order of $\log n$. Chebyshev's inequality then gives

$$\mathbf{P}\left(\left|\frac{D_n}{(2/d)\log n} - 1\right| > \delta\right) \leqq \frac{a}{\log n},$$

which is weaker than the bound obtained in Theorem M2. The exact expressions are obtained by solving some recurrences. It should be noted that the mean was obtained independently by Flajolet et al. (1988), based upon an analysis that involves computing the generating function.

THEOREM M3. *Assume that $d = 2$. For $n \geqq 2$,*

$$\mathbf{E}D_n = H_n - \frac{1}{6} - \frac{2}{3n}$$

and

$$\mathrm{Var}(D_n) = H_n^{(2)} + \frac{1}{2}H_n + \frac{5}{9n} - \frac{4}{9n^2} - \frac{13}{6},$$

where

$$H_n \triangleq \sum_{i=1}^{n} 1/i \qquad \text{and} \qquad H_n^{(2)} \triangleq \sum_{i=1}^{n} (1/i^2).$$

In § 2, we obtain auxiliary results that allow us to prove Theorems M1 and M2. This will be done by reducing the d-dimensional problem to d one-dimensional problems for which we have ready solutions at hand. We consider the quad tree formed by consecutive insertions of $X_1, \cdots, X_{n+1}$, independently and identically distributed uniform $[0, 1]^d$ random vectors. The depth D_{n+1} of X_{n+1} is equal to the number of times the rectangle in the quad tree partition containing X_{n+1} gets "cut" by $X_1, \cdots, X_n$. We start with the full rectangle $[0, 1]^d$. The process of cutting can be summarized by a sequence of random variables (T_k, Z_k), $k \geqq 0$, where $T_0 = 0$, $Z_0 = 1$. T_k is a time counter, and Z_k is the size of the rectangle containing X_{n+1} after it has been cut precisely k times. Given (T_k, Z_k), X_{n+1} and $X_1, \cdots, X_{T_k}$, it is easy to see that $T_{k+1} - T_k$ is geometric with parameter Z_k, i.e., it takes the value i with probability $Z_k(1 - Z_k)^{i-1}$ for $i \geqq 1$. Furthermore, Z_{k+1} is distributed as the size of the rectangle containing X_{n+1}

after it has been cut precisely $k+1$ times. Note that D_{n+1} is equal to the maximal k for which $T_k \leqq n$. Thus, we have

$$\mathbf{P}(D_{n+1} \geqq k) = \mathbf{P}(T_k \leqq n).$$

We will exploit this duality and offer a study of the properties of the Z_k's in § 2.

2. Auxiliary results about spacings, records, and random cuts. Consider independently and identically distributed uniform [0, 1] random variables $U_1, \cdots, U_n$, and let S_{nx} be the size of the interval to which x, a fixed number from [0, 1], belongs.

LEMMA S1. *For any* $x \in [0, 1]$,

$$S_{nx} \overset{L}{=} \min(x, U_1, \cdots, U_n) + 1 - \max(x, U_1, \cdots, U_n),$$

where $\overset{L}{=}$ *denotes equality in distribution. If* $x = U$, *and* $U, U_1, \cdots, U_n$ *is an independently and identically distributed uniform* [0, 1] *sequence, then* S_{nU} *is distributed as the second smallest of* $U, U_1, \cdots, U_n$.

Proof of Lemma S1. We verify the distributional equality in three cases, according to the signs of $\min U_i - x$ and $\max U_i - x$. Consider first the case $\min U_i \leqq x$ and $\max U_i \geqq x$. Then, define

$$V_i = \begin{cases} x - U_i \text{ if } U_i < x, \\ 1 + x - U_i \text{ if } U_i \geqq x, \end{cases} \qquad 1 \leqq i \leqq n,$$

and note that

$$\begin{aligned} S_{nx} &= \left(x - \max_{i: U_i < x} U_i\right) + \left(\min_{i: U_i \geqq x} U_i - x\right) \\ &= \min_{i: U_i < x} V_i + 1 - \max_{i: U_i \geqq x} V_i \\ &= \min_i V_i + 1 - \max_i V_i. \end{aligned}$$

It is easy to verify the two other cases now. For example, if $\min U_i \geqq x$, then

$$S_{nx} = x + 1 - \max_{i: U_i \geqq x} V_i = \min(\min_i V_i, x) + 1 - \max_i V_i.$$

The second statement of the lemma follows from a property of uniform spacings (see, e.g., Pyke (1965), (1972) for a survey), which states that the sum of any k spacings is distributed as the sum of the first k spacings. □

LEMMA S2. *For* $t \in (0, 1)$,

$$\mathbf{P}(S_{nU} < t) = 1 - (1 + tn)(1 - t)^n \leqq (tn)^2\left(\frac{1}{2} + \frac{1}{n(1-t)}\right).$$

Also,

$$\mathbf{P}(S_{nU} > t) = (1 + tn)(1 - t)^n \leqq (1 + tn)\, e^{-tn} \leqq e^{-(tn)^2/(2(1+tn))}.$$

For $tn \geqq 1$, *the last upper bound is not greater than* $e^{-tn/4}$.

Proof of Lemma S2. Let Y be binomial $(n+1, t)$. Then, by Lemma S1,

$$\begin{aligned} \mathbf{P}(S_{nU} < t) &= \mathbf{P}(Y \geqq 2) = 1 - \mathbf{P}(Y = 0) - \mathbf{P}(Y = 1) \\ &= 1 - \binom{n+1}{0} t^0 (1-t)^{n+1} - \binom{n+1}{1} t^1 (1-t)^n \\ &= 1 - (1-t)^n (1 - t + (n+1)t) = 1 - (1-t)^n (1 + tn). \end{aligned}$$

Using $\log(1+\nu) \geqq \nu - \nu^2/2$, valid for $\nu \geqq 0$, and $\log(1-\nu) \geqq -\nu/(1-\nu)$, valid for $0 \leqq \nu < 1$, we see that

$$\mathbf{P}(S_{nU} < t) \leqq 1 - \exp\left[tn - \frac{(tn)^2}{2} - \frac{tn}{1-t}\right] \leqq -tn + \frac{(tn)^2}{2} + \frac{tn}{1-t} = (tn)^2\left(\frac{1}{2} + \frac{1}{n(1-t)}\right).$$

The second part of the lemma follows from the first one and the inequality $\log(1+\nu) - \nu \leqq -\nu^2/(2(1+\nu))$, valid for $\nu \geqq 0$. □

Let $U, U_1, \cdots, U_n$ be independently and identically uniform [0, 1] random variables, and define $[V_i, U]$ and $[U, W_i]$ as the spacings nearest to U after $U, U_1, \cdots, U_i$ have been considered, with the convent on that $V_0 = 0$, $W_0 = 1$. Let N_n be the number of indices i for which $(V_i, W_i) \neq (V_{i-1}, W_{i-1})$, $1 \leqq i \leqq n$. In Devroye (1988), it is shown that N_n is distributed as the sum of n independent Bernoulli random variables Y_i, i.e., $N_n = \sum_{i=1}^n Y_i$, where $\mathbf{E}Y_i = 2/(i+1)$. We will need to know more about the properties of N_n since N_n represents the number of times the spacing containing U is "cut" as we process the U_i's. In particular, we need solid tail bounds. These can be obtained by Chernoff's exponential bounding technique (Chernoff (1952)).

LEMMA S3. *Define* $\mu = 2(H_{n+1} - 1)$. *For* $k \geqq \mu$,

$$\mathbf{P}(N_n \geqq k) \leqq \exp\left(-\frac{(k-\mu)^2}{2k}\right),$$

and for $k \leqq \mu$,

$$\mathbf{P}(N_n \leqq k) \leqq \exp\left(-\frac{(\mu-k)^2}{2\mu}\right).$$

Proof of Lemma S3. By Jensen's inequality, for arbitrary $\lambda > 0$,

$$\begin{aligned}\mathbf{P}(N_n \geqq k) &= \mathbf{P}\left(\sum_{i=1}^n Y_i \geqq k\right) \leqq \mathbf{E}\exp\left(\lambda \sum_{i=1}^n Y_i - \lambda k\right) \\ &= e^{-\lambda k}\prod_{i=1}^n \left(1 - \frac{2}{i+1} + \frac{2e^\lambda}{i+1}\right) \\ &\leqq \exp\left[-\lambda k + 2(e^\lambda - 1)(H_{n+1} - 1)\right].\end{aligned}$$

The exponent is minimal for $e^\lambda = k/(2(H_{n+1} - 1))$. Resubstituting this value and using the notation $y = e^\lambda - 1 > 0$ gives the further upper bound

$$\exp\left[2(H_{n+1} - 1)(y - (1+y)\log(1+y))\right] \leqq \exp\left[-\frac{2(H_{n+1} - 1)y^2}{2(1+y)}\right],$$

where we used the fact that $y - (1+y)\log(1+y) \leqq -y^2/(2(1+y))$, which can be verified by using Taylor's series expansion with remainder. The last upper bound coincides with the first inequality in the statement of the lemma. To obtain the second inequality, we pick another $\lambda > 0$ and note that

$$\begin{aligned}\mathbf{P}(N_n \leqq k) &\leqq e^{\lambda k}\mathbf{E}(e^{-\lambda N_n}) = e^{\lambda k}\prod_{i=1}^n \left(1 - \frac{2}{i+1} + \frac{2e^{-\lambda}}{i+1}\right) \\ &\leqq \exp\left[\lambda k - 2(H_{n+1} - 1)(1 - e^{-\lambda})\right].\end{aligned}$$

The upper bound is minimal for $e^{-\lambda} = k/(2(H_{n+1} - 1))$. We define $y = 1 - e^{-\lambda}$, and resubstitute these values to obtain the upper bound

$$\exp\left[-2(H_{n+1} - 1)(y + (1-y)\log(1-y))\right] \leqq \exp\left[-(H_{n+1} - 1)y^2\right],$$

where once again we used Taylor's series expansion with remainder. This concludes the proof of Lemma S3. □

Lemma S3 shows very clearly that N_n is close to its expected value, $2(H_{n+1}-1)$. We are almost ready to get to the main lemma about uniform cuts. Consider an infinite sequence of independently and identically distributed uniform [0, 1] random variables $U, U_1, \cdots, U_n, \cdots$, and let Z_k be the size of the spacing to which U belongs after it has been "cut" or "hit" k times by members of the sequence $U_1, U_2, \cdots$. In notation introduced above, $Z_k = W_i - V_i$ where (V_i, W_i) is the kth pair not equal to its predecessor. Interestingly, Z_k, S_{nU}, and N_n are connected via the following inclusions of events:

LEMMA S4. *Let $k>0$ and $t\in(0,1)$ be fixed. Then, for any positive integer n,*

$$[Z_k < t] \subseteq [S_{nU} < t] \cup [N_n < k],$$

and

$$[Z_k \geqq t] \subseteq [S_{nU} \geqq t] \cup [N_n \geqq k].$$

Proof of Lemma S4. The proof is obvious. □

We can now announce our main lemma for the **uniform k-cuts** Z_k.

LEMMA S5. *For $k\geqq 3$ and $\delta>0$, we have*

$$\mathbf{P}\left(Z_k < \exp\left[-\frac{k-1}{2}(1+2\delta)\right]\right) \leqq 6\exp[-\delta(k-1)] + \exp\left[-\frac{\delta^2(k-1)}{2(1+\delta)}\right].$$

Also, if $\delta\in(0,\frac{1}{2})$, $\delta\geqq 3/k$, and $k\geqq 2/(1-\delta)$, we have

$$\mathbf{P}\left(Z_k \geqq \exp\left[-\frac{k}{2}(1-2\delta)\right]\right) \leqq \exp\left[\frac{1}{2} - \left(\frac{1}{4}\right)e^{k\delta/2}\right] + (2e)^{\delta^2/(1-\delta)}\exp\left[-\frac{k\delta^2}{2}\right].$$

Proof of Lemma S5. From Lemmas S2, S3, and S4 we recall that for some n to be picked further on,

$$\begin{aligned}\mathbf{P}(Z_k < t) &\leqq \mathbf{P}(S_{nU} < t) + \mathbf{P}(N_n < k)\\ &\leqq (tn)^2\left(\frac{1}{2} + \frac{1}{n(1-t)}\right) + \exp\left[-(H_{n+1}-1)\left(1-\frac{k-1}{2(H_{n+1}-1)}\right)^2\right],\end{aligned}$$

valid for $k-1\leqq 2(H_{n+1}-1)$. Consider a constant $\delta\in(0,\frac{1}{2})$, and define

$$n = \left\lfloor 2\exp\left[\frac{k-1}{2}(1+\delta)\right]\right\rfloor.$$

We note that

$$H_{n+1} - 1 = \sum_{i=2}^{n+1}\frac{1}{i} \geqq \int_2^{n+2}\frac{1}{x}\,dx = \log\left(\frac{n+2}{2}\right) \geqq \frac{k-1}{2}(1+\delta).$$

This implies that $k-1\leqq 2(H_{n+1}-1)$, as required. Using the fact that the function $(y-a)^2/y$ is increasing for $y>a\geqq 0$, that $n\geqq 2$ (by definition), and that $t\leqq\frac{1}{2}$ (by assumption), we see that

$$\begin{aligned}\mathbf{P}(Z_k < t) &\leqq 4t^2\exp[(k-1)(1+\delta)]\left(\frac{1}{2} + \frac{1}{n(1-t)}\right) + \exp\left[-\frac{\delta^2(k-1)}{2(1+\delta)}\right]\\ &\leqq 6t^2\exp[(k-1)(1+\delta)] + \exp\left[-\frac{\delta^2(k-1)}{2(1+\delta)}\right].\end{aligned}$$

We obtain the first half of the lemma by setting $t = \exp[((k-1)/2)(1+2\delta)]$. The condition $t \leqq 1/2$ is fulfilled when $(k-1)(1+2\delta) \geqq 2$, which is certainly the case whenever $k \geqq 3$.

Consider now the second half of the lemma. Assume that n is such that $k \geqq 2(H_{n+1}-1)$. Assume that $tn \geqq 1$. From Lemmas S2, S3, and S4, we retain that

$$\mathbf{P}(Z_k > t) \leqq \mathbf{P}(S_{nU} > t) + \mathbf{P}(N_n \geqq k) \leqq e^{-tn/4} + \exp\left[-\frac{(k-2(H_{n+1}-1))^2}{2k}\right].$$

We now choose $\delta \in (0, 1/2)$, and define

$$n = \lfloor e^{(k/2)(1-\delta)} \rfloor - 1.$$

This value of n is at least one if $k(1-\delta) \geqq 2$. It is easy to verify that $H_{n+1} - 1 \leqq k(1-\delta)/2$ so that the condition relating n and k is indeed satisfied. If we set $y = k/(2(H_{n+1}-1))$ (which, as we have seen, is at least equal to $1/(1-\delta)$), then the inequality reads

$$\mathbf{P}(Z_k > t) \leqq e^{-tn/4} + \exp\left[-(H_{n+1}-1)\frac{(y-1)^2}{y}\right] \leqq e^{-tn/4} + \exp\left[-(H_{n+1}-1)\frac{\delta^2}{1-\delta}\right]$$

$$\leqq e^{-tn/4} + (2e)^{\delta^2/(1-\delta)}\, e^{-k\delta^2/2},$$

where we noted that

$$H_{n+1} - 1 \geqq \log\frac{n+2}{2} - 1 \geqq \log(e^{k(1-\delta)/2}) - \log(2e) = \frac{k(1-\delta)}{2} - \log(2e).$$

For $t = e^{-(k/2)(1-2\delta)}$, we have $tn \geqq e^{k\delta/2} - 2$, so that the upper bound becomes

$$\mathbf{P}(Z_k > t) \leqq e^{1/2-(1/4)e^{k\delta/2}} + (2e)^{\delta^2/(1-\delta)}\, e^{-k\delta^2/2}.$$

Also, the condition $tn \geqq 1$ is fulfilled if $k\delta \geqq 3$. □

3. A law of large numbers for quad trees. The purpose of this section is to prove Theorem M1.

THEOREM M1. *$D_n/\log n$ tends in probability to $2/d$ as $n \to \infty$. Also, $\mathbf{E}D_n \sim \mathbf{E}A_n \sim (2/d)\log n$ as $n \to \infty$.*

Recall the definition of T_k and Z_k from the Introduction. We have

$$\mathbf{P}(D_{n+1} \geqq k) = \mathbf{P}(T_k \leqq n) \leqq \mathbf{P}(T_k - T_{k-1} \leqq n).$$

Observe that $Z_k = \prod_{i=1}^{d} Z_k(i)$, where the $Z_k(i)$'s are independently and identically distributed random variables distributed as the uniform k-cut dealt with in Lemma S5. We choose a small positive constant δ, and define $q = \exp(-(k-1)(1-2\delta)/2)$. Let A be the event that $\max_i Z_{k-1}(i) \leqq q$. By an obvious left tail bound for the geometric distribution and Lemma S5, we have

$$\begin{aligned}\mathbf{P}(D_{n+1} \geqq k) &\leqq \mathbf{P}(A,\, T_k - T_{k-1} \leqq n) + \mathbf{P}(A^c) \\ &\leqq nq^d + d \times \mathbf{P}(Z_{k-1}(1) > q), \qquad (1) \\ &\leqq nq^d + d(e^{1/2-(1/4)e^{(k-1)\delta/2}} + (2e)^{\delta^2/(1-\delta)}\, e^{(k-1)\delta^2/2})\end{aligned}$$

if $\delta \geqq 3/(k-1)$ and $k-1 \geqq 2/(1-\delta)$. As $k \to \infty$, the upper bound is $nq^d + o(1)$. If we now take

$$k - 1 = \left\lfloor \frac{2}{d(1-3\delta)} \log n \right\rfloor,$$

then it is easy to verify that $nq^d = o(1)$ as well. Hence, $\mathbf{P}(D_{n+1} \geqq k) = o(1)$, proving one half of the theorem (since δ is arbitrary). In fact, $\mathbf{P}(D_{n+1} \geqq k) = O(\log^{-R} n)$ for any positive constant R.

The second half is proved similarly. Because we obtained exponential inequalities in the lemmas of the previous section, we can actually get away with a very crude bounding technique. Let A be the event $\min_i Z_{k-1}(i) \geqq q$ where $q = \exp(((k-2)/2)(1+2\delta))$, let $k \geqq 3$, and assume that $\delta > 0$ is an arbitrary but small constant. Then,

$$\begin{aligned}
\mathbf{P}(D_{n+1} < k) = \mathbf{P}(T_k > n) &\leqq \mathbf{P}\left(\bigcup_{j=1}^{k} \left[T_j - T_{j-1} > \frac{n}{k}\right]\right) \\
&\leqq k\mathbf{P}\left(T_k - T_{k-1} > \frac{n}{k}\right) \\
&\leqq k\left(\mathbf{P}\left(A, T_k - T_{k-1} > \frac{n}{k}\right) + \mathbf{P}(A^c)\right) \\
&\leqq k\left(\sum_{i > n/k} q^d(1-q^d)^{i-1} + d \times \mathbf{P}(Z_{k-1}(1) < q)\right) \\
&\leqq k(1-q^d)^{(n/k)-1} + kd \times \mathbf{P}(Z_{k-1}(1) < q) \\
&\leqq k \exp\left[-\left(\frac{n}{k} - 1\right) q^d\right] + o(1)
\end{aligned}$$

whenever $k \to \infty$ (by Lemma S5 and our choice of q). If we take

$$k - 2 = \left\lfloor \frac{2}{d(1+3\delta)} \log n \right\rfloor,$$

it is a simple exercise to verify that $nq^d/k \to \infty$, which then shows that $\mathbf{P}(D_{n+1} < k) = o(1)$, as required. This concludes the proof of the weak convergence of D_n. This trivially implies that

$$\liminf_{n \to \infty} \frac{\mathbf{E}D_n}{\log n} \geqq \frac{2}{d}.$$

Also, for small $\delta > 0$, and arbitrary $M > 1$,

$$\begin{aligned}
\mathbf{E}D_n &= \int_0^n \mathbf{P}(D_n > t)\, dt \\
&\leqq 1 + \frac{2}{d(1-3\delta)} \log n + M \log n\, \mathbf{P}\left(D_n > 1 + \left\lfloor \frac{2}{d(1-3\delta)} \right\rfloor \log n\right) \\
&\quad + n\mathbf{P}(D_n > M \log n) \\
&= 1 + \frac{2}{d(1-3\delta)} \log n + o(1) + n\mathbf{P}(D_n > M \log n)
\end{aligned}$$

by the bounds obtained above. To conclude that $\mathbf{E}D_n \sim (2/d) \log n$, we need only establish that $\mathbf{P}(D_n > M \log n) = o(1/n)$ for some constant M. This follows by noting that the bound (1) with q as chosen there and $k = \lfloor M \log n \rfloor$ is $o(1/n)$ whenever $M > \max(2/\delta^2, 4/(d(1-2\delta)))$.

The statement about $\mathbf{E}A_n$ finally follows easily from the statement regarding $\mathbf{E}D_n$. □

4. Some recurrences related to quad trees. In this section, we only consider the case $d=2$. As above, we let D_n be the depth of the nth node in a tree of n nodes. We also define

$$p_{n,l} \triangleq \mathbf{P}(D_n = l),$$

and note that by convention $p_{1,0}=1$, i.e., the root node is at depth zero. We begin with the following recursion:

LEMMA R1. *Let N_j, $1 \leqq j \leqq 4$, be the cardinalities of the four subtrees of the root of a random quad tree in the plane. Then*

$$\mathbf{P}(N_j = i) = \frac{H_n - H_i}{n}, \qquad 0 \leqq i \leqq n-1,$$

when $n \geqq 2$. Also, for $n \geqq 2$,

$$p_{n,l} = \frac{4}{n(n-1)} \sum_{i=0}^{n-1} i(H_n - H_i) p_{i,l-1},$$

where $p_{i,l-1}=0$ for $0 \leqq i < l$.

Proof of Lemma R1. We note that $\mathbf{P}(N_1+N_2=i)=1/n$ for $0 \leqq i < n$. Given N_1+N_2, N_1 is again uniformly distributed on $0, \cdots, N_1+N_2$. Thus,

$$\mathbf{P}(N_1=i) = \sum_{j=i}^{n-1} \frac{1}{j+1} \mathbf{P}(N_1+N_2=j) = \frac{1}{n} \sum_{j=i}^{n-1} \frac{1}{j+1} = \frac{1}{n}(H_n - H_i).$$

For the second part of the lemma, we use the fact that given the N_i's, the last node ends up in the ith subtree with probability $N_i/(n-1)$. Thus,

$$p_{n,l} = \sum_{j=1}^{4} \sum_{i=0}^{n-1} \frac{i}{n-1} \mathbf{P}(N_j = i) p_{i,l-1},$$

from which we deduce our result by symmetry. □

The basic recurrence of Lemma R1 can now be used to obtain recurrences for the generating function and the moments of D_n. We define

$$\phi_n(t) = \mathbf{E}(e^{tD_n})$$

and

$$\mu_{n,m} = \mathbf{E}(D_n^m) = \phi_n^{(m)}(0).$$

LEMMA R2.

$$\phi_n(t) = \frac{4e^t}{n(n-1)} \sum_{i=1}^{n-1} i(H_n - H_i)\phi_i(t),$$

and, for $m>0$,

$$\mu_{n,m} = \frac{4}{n(n-1)} \sum_{i=1}^{n-1} i(H_n - H_i) \sum_{j=0}^{m} \binom{m}{j} \mu_{i,j},$$

where $\mu_{n,0}=1$.

Proof of Lemma R2. From Lemma R1,

$$\begin{aligned}
\phi_n(t) &= \sum_{l=1}^{n-1} p_{n,l} e^{tl} = \sum_{l=1}^{n-1} e^{tl} \frac{4}{n(n-1)} \sum_{i=l}^{n-1} i(H_n - H_i) p_{i,l-1} \\
&= \frac{4}{n(n-1)} \sum_{i=1}^{n-1} i(H_n - H_i) \sum_{l=1}^{i} e^{tl} p_{i,l-1} \\
&= \frac{4e^t}{n(n-1)} \sum_{i=1}^{n-1} i(H_n - H_i)\phi_i(t).
\end{aligned}$$

This proves the first recurrence of the lemma. Let us now take $f_n(t) = \sum_{i=1}^{n-1} i(H_n - H_i)\phi_i(t)$. We have

$$\phi_n(t) = \frac{4\,e^t}{n(n-1)} f_n(t),$$

and thus

$$\phi_n^{(m)}(t) = \frac{4\,e^t}{n(n-1)} \sum_{j=0}^{m} \binom{m}{j} f_n^{(j)}(t).$$

Thus,

$$\phi_n^{(m)}(0) = \frac{4}{n(n-1)} \sum_{j=0}^{m} \binom{m}{j} f_n^{(j)}(0) = \frac{4}{n(n-1)} \sum_{j=0}^{m} \binom{m}{j} \sum_{i=1}^{n-1} i(H_n - H_i)\phi_i^{(j)}(0),$$

which concludes the proof of the lemma. □

From Lemma R2, we can conclude, with a little work, the following lemmas:

LEMMA R3.

$$\mu_{n,m} = \sum_{j=0}^{m-1} \binom{m}{j} \mu_{n,j}(-1)^{m-1-j} + \frac{4}{n(n-1)} \sum_{i=1}^{n-1} i(H_n - H_i)\mu_{i,m}.$$

In particular, $\mu_{n,0} = 1$,

$$\mu_{n,1} = 1 + \frac{4}{n(n-1)} \sum_{i=1}^{n-1} i(H_n - H_i)\mu_{i,1},$$

and

$$\mu_{n,2} = 2\mu_{n,1} - 1 + \frac{4}{n(n-1)} \sum_{i=1}^{n-1} i(H_n - H_i)\mu_{i,2}.$$

The recurrences in Lemma R3 are of the following general form:

$$x_n = a_n + \frac{4}{n(n-1)} \sum_{i=1}^{n-1} i(H_n - H_i)x_i, \qquad n \geqq 2, \tag{2}$$

where

$$x_1 = 0, \qquad x_2 = a_2. \tag{3}$$

LEMMA R4. *The general solution of recurrences* (2) *and* (3) *is*

$$x_n = a_n + 4 \sum_{j=3}^{n} \frac{\sum_{i=1}^{j-1} i^2(i-1)a_i}{j^2(j-1)^2(j-2)}, \qquad n \geqq 3.$$

Proof of Lemma R4. The proof is omitted. □

Lemmas R3 and R4 can now be combined to obtain the moments of D_n. In particular, we have

THEOREM R1. *For* $n \geqq 2$,

$$\mu_{n,1} = H_n - \frac{1}{6} - \frac{2}{3n}$$

and

$$\mu_{n,2} = H_n^2 + H_n^{(2)} + \frac{H_n}{6} - \frac{4H_n}{3n} + \frac{7}{9n} - \frac{77}{36}.$$

Also,

$$\mathrm{Var}\,(D_n) = H_n^{(2)} + \frac{1}{2}H_n + \frac{5}{9n} - \frac{4}{9n^2} - \frac{13}{6}.$$

Proof of Theorem R1. For $\mu_{n,1}$, we note that $a_n = 1$, so that simply

$$\begin{aligned}
\mu_{n,1} &= 1 + 4\sum_{j=3}^{n} \frac{\sum_{i=1}^{j-1} i^2(i-1)}{j^2(j-1)^2(j-2)} \\
&= 1 + \frac{1}{3}\sum_{j=3}^{n} \frac{3j-1}{j(j-1)} \\
&= 1 + \frac{1}{3}\sum_{j=3}^{n} \left(\frac{1}{j} + \frac{2}{j-1}\right) \\
&= 1 + \frac{1}{3}\left(H_n - 1 - \frac{1}{2} + 2(H_{n-1} - 1)\right) = H_n - \frac{1}{6} - \frac{2}{3n}.
\end{aligned}$$

From this, we see that in the computation of $\mu_{n,2}$, when we apply Lemma R4, a_n can be set equal to $2\mu_{n,1} - 1 = 2H_n - (4/3)((n+1)/n)$. From Lemmas R3 and R4, we then conclude that

$$\mu_{n,2} = 2H_n - \frac{4}{3}\frac{n+1}{n} + 4\sum_{j=3}^{n} \frac{b_j}{j^2(j-1)^2(j-2)}$$

where

$$\begin{aligned}
b_j &= \sum_{i=1}^{j-1} i^2(i-1)\left(2H_i - \frac{4}{3}\frac{i+1}{i}\right) \\
&= 2\left(\sum_{i=1}^{j} i^2(i-1)H_i - j^2(j-1)H_j\right) - \frac{4}{3}\sum_{i=1}^{j-1} i(i^2-1) \\
&= \frac{j(j-1)(j-2)}{72}(12(3j-1)H_j - (33j+29)).
\end{aligned}$$

Thus,

$$\begin{aligned}
\mu_{n,2} &= 2H_n - \frac{4}{3}\frac{n+1}{n} + \frac{1}{18}\sum_{j=3}^{n} \frac{12(3j-1)H_j - (33j+29)}{j(j-1)} \\
&= 2H_n - \frac{4}{3}\frac{n+1}{n} + \frac{2}{3}\sum_{j=3}^{n}\left(\frac{1}{j} + \frac{2}{j-1}\right)H_j - \frac{1}{18}\sum_{j=3}^{n}\left(\frac{62}{j-1} - \frac{29}{j}\right) \\
&= 2H_n - \frac{4}{3}\frac{n+1}{n} + \frac{2}{3}\left(\sum_{j=3}^{n}\frac{H_j}{j} + 2\sum_{j=3}^{n}\frac{H_{j-1} + 1/j}{j-1}\right) - \frac{1}{18}\left(33H_n - \frac{37}{2} - \frac{62}{n}\right) \\
&= \frac{1}{6}H_n - \frac{11}{36} + \frac{19}{9n} + \frac{2}{3}\left(\sum_{j=1}^{n}\frac{H_j}{j} - 1 - \frac{3}{4} + 2\sum_{j=2}^{n-1}\frac{H_j}{j} + 2\sum_{j=3}^{n}\frac{1}{j(j-1)}\right) \\
&= \frac{1}{6}H_n - \frac{53}{36} + \frac{19}{9n} - \frac{2}{3}(2H_n + 2) + 2\sum_{j=1}^{n}\frac{H_j}{j} + \frac{4}{3}\left(\sum_{j=2}^{n-1}\frac{1}{j} - \sum_{j=3}^{n}\frac{1}{j}\right) \\
&= \frac{1}{6}H_n - \frac{101}{36} + \frac{19}{9n} - \frac{4H_n}{3n} + H_n^2 + H_n^{(2)} + \frac{4}{3}\left(H_n - \frac{1}{n} - 1 - H_n + 1 + \frac{1}{2}\right) \\
&= \frac{1}{6}H_n - \frac{77}{36} + \frac{7}{9n} - \frac{4H_n}{3n} + H_n^2 + H_n^{(2)}.
\end{aligned}$$

In this derivation, we needed the following identities, some of which can be found in Knuth (1973, pp. 75-77):

$$\sum_{i=1}^{n} H_i = (n+1)H_n - n,$$

$$\sum_{i=1}^{n} iH_i = \frac{n(n+1)}{2} H_n - \frac{n(n-1)}{4},$$

$$\sum_{i=1}^{n} i^2 H_i = \frac{n(n+1)(2n+1)}{6} H_n - \frac{n(n-1)(4n+1)}{36},$$

$$\sum_{i=1}^{n} i^3 H_i = \frac{n^2(n+1)^2}{4} H_n - \frac{n(n^2-1)(3n-2)}{48},$$

$$\sum_{i=1}^{n} \frac{H_i}{i} = \frac{1}{2}(H_n^2 + H_n^{(2)}).$$

Finally, $\mathrm{Var}(D_n)$ is obtained as $\mu_{n,2} - \mu_{n,1}^2$. □

From Theorem R1, we conclude that $\mathbf{E}D_n = \log n + \gamma - \frac{1}{6} + o(1)$ and $\mathrm{Var}(D_n) = \frac{1}{2}\log n + \gamma/2 + \pi^2/6 - \frac{13}{6} + o(1)$, where γ is Euler's constant. Chebyshev's inequality now implies that $D_n/\log n \to 1$ in probability as $n \to \infty$. The resulting upper bound for $\mathbf{P}(D_n/\log n \notin (1-\varepsilon, 1+\varepsilon))$ drops off as $(1+o(1))(2\varepsilon^2 \log n)^{-1}$. The exponential bounding method for the previous section yields better tail bonds, however.

REFERENCES

J. L. BENTLEY (1975), *Multidimensional binary search trees used for associative searching*, Comm. ACM, 18, pp. 509-517.

J. L. BENTLEY, D. F. STANAT, AND E. H. WILLIAMS (1977), *The complexity of fixed-radius near neighbor searching*, Inform. Process. Lett., 6, pp. 209-212.

G. G. BROWN AND B. O. SHUBERT (1984), *On random binary trees*, Math. Oper. Res., 9, pp. 43-65.

H. CHERNOFF (1952), *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*, Annals of Mathematical Statistics, 23, pp. 493-507.

L. DEVROYE (1986), *A note on the height of binary search trees*, J. Assoc. Comput. Mach., 33, pp. 489-498.

——— (1987), *Branching processes in the analysis of the heights of trees*, Acta Inform., 24, pp. 277-298.

——— (1988), *Applications of the theory of records in the study of random trees*, Acta Inform., 26, pp. 123-130.

R. A. FINKEL AND J. L. BENTLEY (1974), *Quad trees: A data structure for retrieval on composite keys*, Acta Inform., 4, pp. 1-9.

P. FLAJOLET, G. GONNET, C. PUECH, AND M. ROBSON (1988), *Analytic variations on quadtrees*, Tech. Report, INRIA, Rocquencourt, France.

G. H. GONNET (1984), *A Handbook of Algorithms and Data Structures*, Addison-Wesley, Reading, MA.

G. M. HUNTER AND K. STEIGLITZ (1979), *Operations on images using quad trees*, IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-1, pp. 145-153.

D. E. KNUTH (1973), *The Art of Computer Programming, Vol. 1: Fundamental Algorithms*, 2nd ed., Addison-Wesley, Reading, MA.

——— (1973), *The Art of Computer Programming, Vol. 3: Sorting and Searching*, Addison-Wesley, Reading, MA.

D. T. LEE AND C. K. WONG (1981), *Quintary trees: A file structure for multidimensional data-base systems*, ACM Trans. Database Systems, 5, pp. 339-353.

G. LOUCHARD (1987), *Exact and asymptotic distributions in digital and binary search trees*, Theoretical Informatics and Applications, 21, pp. 479-496.

W. C. LYNCH (1965), *More combinatorial problems on certain trees*, Computer J., 7, pp. 299-302.

H. MAHMOUD AND B. PITTEL (1984), *On the most probable shape of a search tree grown from a random permutation*, SIAM J. Algebraic Discrete Methods, 5, pp. 69-81.

T. H. MERRETT AND E. OTOO (1981), *Multidimensional paging for associative searching*, Tech. Report SOCS 81-18, School of Computer Science, McGill University, Montréal.

J. A. ORENSTEIN (1982), *Multidimensional tries used for associative searching*, Tech. Report School of Computer Science, McGill University, Montréal.

B. PITTEL (1984), *On growing random binary trees*, J. Math. Anal. Appl., 103, pp. 461-480.

C. PUECH AND H. YAHIA (1985), *Quadtress, octrees, hyperoctrees: A unified analytical approach to tree data structures used in graphics, geometric modeling and image processing*, in Proc. Symposium on Computational Geometry, Association for Computing Machinery, New York, pp. 272-280.

R. PYKE (1965), *Spacings*, J. R. Statist. Soc. Ser. B, 7, pp. 395-445.

——— (1975), *Spacings revisited*, Proc. Sixth Berkeley Symposium, 1, pp. 417-427.

R. L. RIVEST (1974), *Analysis of associative retrieval algorithms*, Tech. Report STAN-CS-74-415, Computer Science Department, Stanford University, Stanford, CA.

J. T. ROBINSON (1981), *The* K-B-D *tree: A search structure for large multidimensional dynamic indexes*, Proc. ACM SIGMOD, pp. 10-18.

J. M. ROBSON (1979), *The height of binary search trees*, Austral. Comput. J., 11, pp. 151-153.

H. SAMET (1980), *Deletion in two-dimensional quad trees*, Comm. ACM, 23, pp. 703-710.

——— (1984), *The quadtree and related hierarchical data structures*, Comput. Surveys, 16, pp. 187-260.

R. SEDGEWICK (1985), *Mathematical analysis of combinatorial algorithms*, in Probability Theory and Computer Science, G. Louchard and G. Latouche, eds., Academic Press, London, pp. 123-205.

M. TAMMINEN (1981), *Order preserving extendible hashing and bucket tries*, BIT, 21, pp. 419-435.

——— (1981), *The EXCELL method for efficient geometric access to data*, Acta Polytech. Scand. Math. Comput. Sci. Ser. 34, Helsinki, Finland.

——— (1982), *The extendible cell method for closest point problems*, BIT, 22, pp. 27-41.

J. R. WOODWARK (1982), *The explicit quad tree as a structure for computer graphics*, Comput. J., 25, pp. 235-237.