INFORMATIQUE THÉORIQUE ET APPLICATIONS

G. LOUCHARD

Exact and asymptotic distributions in digital and binary search trees

Informatique théorique et applications, tome 21, n° 4 (1987), p. 479-495.

<http://www.numdam.org/item?id=ITA_1987__21_4_479_0>

Informatique théorique et Applications/Theoretical Informatics and Applications
(vol. 21, n° 4, 1987, p. 479 à 496)

EXACT AND ASYMPTOTIC DISTRIBUTIONS IN DIGITAL AND BINARY SEARCH TREES *

by G. LOUCHARD (1)

Communicated by P. FLAJOLET

Abstract. – *Combinatorial relations and classical analysis are used to derive exact and asymptotic distributions for the number of steps during a successful search in digital and binary search trees.*

Résumé. – *Diverses relations combinatoires et des méthodes d'analyse classique sont utilisées afin de déterminer les distributions exactes et asymptotiques du coût d'une recherche avec succès dans un arbre digital et dans un arbre binaire de recherche.*

1. INTRODUCTION

A *digital search tree* (DST) is a data structure (binary tree) where keys are represented by binary numbers (*see* Knuth [13], § 6.3). Their bits are used to govern searching to the left or right branch at each insertion (or search).

When the keys are uniform [0, 1] random variables, the asymptotic mean and variance of *JD*, the number of steps in a successful search, are known (Knuth [13], Ex. 6.3/27, Flajolet and Sedgewick [5], Kirschenhofer and Prodinger [11]).

A *binary search tree* (BST) (Knuth [13], § 6.2.2) also stores the keys in nodes but now, upon an insertion, the key is compared to the key present at the node to govern the choice of the branch. The exact mean, variance and distribution of *JB* (number of steps in a successful search) are known [Knuth [13], Ex. 6.2.2/6, Brown and Shubert [2] eq. (4.5)] when all permutations of keys have equal probability.

(*) Received July 1986, revised May 1987.
(1) Laboratoire d'Informatique Théorique, CP 212, Université Libre de Bruxelles, boulevard du Triomphe, B-1050 Bruxelles, Belgique.

Informatique théorique et Applications/Theoretical Informatics and Applications
0296-1598 87/04 479 18/$ 3.80/

The purpose of this paper is to obtain the asymptotic density and distribution function of *JD* and *JB* when the size of the tree is large. Our first intention was to use a technique based on the Brownian Motion or on the Poisson process, a technique that we applied to other complexity problems in [14] and [15]. We soon discovered that this path was useless (*see* Section 3), and turned instead to the direct approach, based on exact distributions. This leads indeed to the solution of our problem: only combinatorial relations and classical analysis are necessary.

The paper is organized as follows: in Section 2 we summarize basic notations and known results. Section 3 deals with exact distributions for *JD*. Section 4 is devoted to asymptotic results for *JD*. Section 5 and 6 deal with *JB*. Section 7 concludes the paper.

An appendix collects the combinatorial relations we need in the text.

A third classical tree, the *trie structure*, will not be considered here: It has been analysed, for instance, in Louchard [15], Pittel [16], Jacquet and Regnier [9], Devroye [3].

2. BASIC NOTATIONS AND KNOWN RESULTS

Throughout the paper, n denotes the number of keys in the DST or the BST. The searched key is choosen at random among these n keys. Quantities of interest are *JD* and *JB* that represent the number of steps necessary to retrieve a key in a DST or a BST (the searched key is always assumed to be in the tree). We let $PD(j) := \Pr[JD = j]$ and $PB(j) := \Pr[JB = j]$ denote their probability distribution (probability of success at step j). Note that

$$PD(j) = 0, \qquad \forall j > n. \tag{1}$$

The following results are known:

THEOREM A: *The asymptotic mean and variance of the number of steps JD in a DST are given by*

$$E(JD) = \sum_{j=1}^{\infty} j\, PD(j) \sim \log_2 n + (\gamma - 1)/\ln 2 + \frac{3}{2} - \alpha + \delta(\log_2 n) + O(\log_2 n/n) \tag{2}$$

where

- $\gamma :=$ *the classical Euler constant*
- $\alpha := \sum_{j=1}^{\infty} 1/(2^j-1) = 1+1/3+1/7+\ldots = 1.60669\ldots$
- $\delta(.)$ *is a (small) periodic function.*

The proof may be found in Knuth [13], Ex. 6.3/27, where α is given with more precision, and in Flajolet and Sedgewick [5].

$$\mathrm{VAR}(JD) \sim \frac{1}{12} + \frac{\pi^2}{6(\ln 2)^2} + \frac{1}{(\ln 2)^2} - \alpha - \beta + \omega(\log_2 n)$$

where

- $\beta := \sum_{j=1}^{\infty} 1/(2^j-1)^2$
- $\omega(.)$ *is a (small) periodic function.*

The proof may be found in Kirschenhofer and Prodinger [11]. ■

THEOREM B: *The mean, variance and probability distribution of the number of steps JB in a BST are given by*

$$E(JB) = 2\left(1+\frac{1}{n}\right)H_n - 3 \quad (\text{Knuth [13], § 6.2.2}).$$

where $H_n := \sum_{k=1}^{n} (1/k) =$ *the n-th Harmonic number,* $H_n \sim \ln n + \gamma + O(1/n)$.

VAR (*JB*) is given in Knuth [13], ex. 6.2.2/6. With misprint corrections, it yields

$$\mathrm{VAR}(JB) = (2+10/n)H_n - 4(1+1/n)(H_n^2/n + H_n^{(2)}) + 4,$$

hence $\mathrm{VAR}(JB) \sim 2\ln n$

$$PB(j) = \frac{2^{j-1}}{n} \sum_{k=j}^{n} \begin{bmatrix} n \\ k \end{bmatrix} \Big/ n! \tag{3}$$

where $\begin{bmatrix} n \\ i \end{bmatrix} :=$ the Stirling number of the first kind (unsigned version). The proof is given in Brown and Shubert [2], eq. (4.5). ■

For DST, the following quantities are of interest:

– $B(i, j) := (i+j)!/(i!\, j!\, 2^{i+j})$ i. e. the probability that among $i+j$ (fair) coin tosses, there are i success. It is the classical binomial distribution with $p \equiv q \equiv 1/2$.

– $Q(j) := \prod_{k=1}^{j} (1-(1/2^k))$, $Q(0) := 1$.

$Q(\infty) = 0.28879\ldots$ (*see* Knuth [13], Ex. 6.3/26 for more precision).

– $R(j) := (-1)^{j+1} \prod_{k=1}^{j} 1/(2^k-1)$, $R(0) := -1$.

Note that $R(j) \equiv (-1)^{j+1} 2^{-j(j+1)/2}/Q(j)$ but we prefer to keep separate notations to simplify some formulas

– $\{x\} := x - \lfloor x \rfloor$.

For BST, the following quantities will be used:

– $\overline{\begin{bmatrix} n \\ i \end{bmatrix}} := \begin{bmatrix} n \\ i \end{bmatrix} \Big/ (n-1)!$

– $G_n(z) := z(1+z)(1+z/2)\ldots(1+z/(n-1))$

– $\mathcal{N}(m, v) :=$ normal (Gaussian) random variable with mean m and variance v.

– $\zeta(z) :=$ the classical Riemann zeta function.

All our analyses are relative to $n \to \infty$ and $\sim$ is assumed to mean: asymptotic to, for $n \to \infty$.

3. EXACT DISTRIBUTIONS FOR DIGITAL SEARCH TREES

We will firstly analyse the exact distribution $PD(j)$. This is given by:

THEOREM 1: *The probability distribution of the cost JD of a search in a DST is given by*

$$PD(j+1) = \frac{2^j}{n}\left[1 + \sum_{k=1}^{j} \frac{R(j-k)}{Q(k-1)}\left(1-\frac{1}{2^k}\right)^{n-1}\right], \qquad j \geqq 1$$
$$PD(1) = 1/n. \tag{4}$$

Proof: The proof will be divided into two steps.

(i) Obviously $PD(1)=1/n$. Let $PD^*(j)=PD$ [success at step j | failure at step 1]. Of course $PD(j)=(1-(1/n))\,PD^*(j)$. $PD^*(2)$ is given by:

$$B(0,\,n-1)\times\frac{1}{n-1}+\sum_{i=1}^{n-2}B(i,n-1-i)\left[\frac{i}{n-1}\cdot\frac{1}{i}+\frac{n-1-i}{n-1}\cdot\frac{1}{n-1-i}\right]$$
$$+B(n-1,\,0)\times\frac{1}{n-1}\quad(5)$$

the binomial yields the left-right subtree decomposition, each bracket term gives the probability for the searched key of falling into a left (or right) subtree multiplied by the probability of success at step 1 in this subtree.

After elementary manipulations this leads to

$$PD^*(2)=\frac{2}{n-1}\left(1-\frac{1}{2^{n-1}}\right),$$

hence

$$PD(2)=\left(1-\frac{1}{n}\right)PD^*(2)=\frac{2}{n}\left(1-\frac{1}{2^{n-1}}\right).$$

We immediately see why the Brownian approach is doomed to failure: it cannot take into account the term $1/2^{n-1}$ which is crucial in the following steps.

(ii) To proceed by induction, we observe that an operator of type (5) (call it TD), when applied to a term $((1/n)\,\rho^{n-1})$, again yields

$$TD\left(\frac{\rho^{n-1}}{n}\right):=B(0,n-1)\times\frac{1}{n-1}\rho^{n-2}$$
$$+\sum_{i=1}^{n-2}B(i,n-1-i)\times\left[\frac{i}{n-1}\cdot\frac{\rho^{i-1}}{i}+\frac{n-1-i}{n-1}\cdot\frac{\rho^{n-2-i}}{n-1-i}\right]$$
$$+B(n-1,0)\times\frac{1}{n-1}\rho^{n-2}=\frac{2}{\rho(n-1)}\left[-\frac{1}{2^{n-1}}+\left(\frac{1+\rho}{2}\right)^{n-1}\right].$$

For instance

$$PD^*(2)=TD\left(\frac{1^{n-1}}{n}\right)=\frac{2}{n-1}\left(-\frac{1}{2^{n-1}}+1\right).$$

$$PD^*(3)=TD\,(PD\,(2))$$
$$=2\left[TD\left(\frac{1^{n-1}}{n}\right)-TD\left(\frac{(1/2)^{n-1}}{n}\right)\right]=\frac{4}{n-1}\left[1+\frac{1}{2^{n-1}}-\frac{1}{1/2}\left(\frac{3}{4}\right)^{n-1}\right].$$

Now, to obtain $PD(j+1)$, we see that application of TD to each term of $PD(j)$ gives:

– $2^j/n$: arising from $2^{j-1}/n$

– $R(j-k)/Q(k-1)\,(1-(1/2^k))^{n-1}$, $k=2\ldots j$: arising from the corresponding term in $PD(j)$

– a term in $\lambda/2^{n-1}$: arising from $2^{j-1}/n$ and *all terms* $k=1\,..\,j-1$ in $PD(j)$.

Direct computation of the coefficient λ of $1/2^{n-1}$ is complicated but fortunately we can use (1); from (4) it follows that

$$1+\lambda+\sum_{k=2}^{j}\frac{R(j-k)}{Q(k-1)}=0,$$

hence $\lambda=R(j-1)$ by (A.8). ■

The distribution function $FD(j)=\sum_{i=1}^{j} PD(i)$ is given by:

THEOREM 2: *The distribution function of the cost JD of a search in a DST is given by*

$$FD(j+1)=\frac{1}{n}\left[(2^{j+1}-1)+\sum_{k=1}^{j}2^k\frac{R(j-k)}{Q(k-1)}\left(1-\frac{1}{2^k}\right)^{n-1}\right],\qquad j\geqq 1. \quad (6)$$

Proof: From (4), the first term is given by $(1/n)\sum_{0}^{j}2^i$.

Again from (4), the coefficient of $1/Q(k-1)(1-(1/2^k))^{n-1}$ becomes $\sum_{i=0}^{j-k}2^{j-i}R(j-k-i)$ which gives $2^kR(j-k)$ by simple induction. ■

Remark: the expression (4) for $PD(j+1)$ is curiously related to the probability $PC(j)$ derived by Flajolet in [4] for the approximate counting algorithm. In our notation, Flajolet obtains

$$PC(j)=2^j\left[\sum_{k=1}^{j}\frac{-R(j-k)}{2^kQ(k-1)}\left(1-\frac{1}{2^k}\right)^n\right].$$

The asymptotic corresponding mean is given by

$$\log_2 n+\gamma/\ln 2+\frac{1}{2}-\alpha+\omega(\log_2 n)+O\,(1/n^{0.98})$$

where $\omega(.)$ is a (small) periodic function. This result is proved in [4] by Mellin transform techniques.

4. ASYMPTOTIC DISTRIBUTIONS FOR DIGITAL SEARCH TREES

The asymptotic distributions are given by:

THEOREM 3: *The asymptotic distribution function and probability distribution of the cost JD of a search in a DST are given by the following equations: set* $\eta:=j-\log_2 n$, *with* $\eta=O\,(1)$.

$$FD\,(j)\sim GD\,(\eta) \tag{7}$$

where

$$GD\,(\eta):=2^{\eta}\left(1+\frac{1}{2\,Q\,(\infty)}\sum_{i=0}^{\infty}\frac{R\,(i)}{2^i}e^{-2^{-(\eta-1-i)}}\right).$$

$$PD\,(j)=FD\,(j)-FD\,(j-1)\sim GD\,(\eta)-GD\,(\eta-1)$$

$$=2^{\eta-1}\left(1+\frac{1}{Q\,(\infty)}\sum_{i=0}^{\infty}R\,(i)\,e^{-2^{-(\eta-1-i)}}\right).$$

Letting $\eta=j-\lfloor\log_2 n\rfloor-\{\log_2 n\}$, *asymptotically, the distributions are periodic functions of* $\log_2 n$.

Proof: Letting $n\to\infty$ in (6) [for $FD\,(j)$], we readily obtain (7), as $Q\,(i)\underset{i\to\infty}{\longrightarrow}Q\,(\infty)$. We have no problem of passage to the limit as the series $\sum_0^{\infty}|R\,(i)|$ obviously converges.

The (discrete) asymptotic density for $PD\,(j)$ is immediate from (7) [and could also be derived from (4)].

We could refine on the validity intervals of (7), as Flajolet did for approximate counting in [4]. Propositions 2, 3 and 4. We will not give the details here. ■

Note that by (A.4), (7) can also be rewritten as

$$GD(\eta)=\frac{2^{\eta}}{2Q(\infty)}\sum_{i=0}^{\infty}\frac{R(i)}{2^{i}}(e^{-2^{-(\eta-1-i)}}-1). \tag{8}$$

Of course $GD(\eta)\xrightarrow[\eta\to-\infty]{}0$. It is easy to check that $GD(\eta)\xrightarrow[\eta\to\infty]{}1$.

Indeed $2^{\eta}[e^{-2^{-(\eta-1-i)}}-1]\xrightarrow[\eta\to\infty]{}-2^{i+1}$ and

$$GD(\eta)\to-\frac{1}{Q(\infty)}\sum_{i=0}^{\infty}R(i)=1$$

by (A.3). Obviously, the asymptotic non-periodic term in the moments of *JD* is given by:

THEOREM 4: *The constant term* $\bar{E}$ *in the Fourier expansion (in* $\log_2 n$*) of the moments of JD is asymptotically given by*

$$\bar{E}[JD-\log_2 n]^i\sim\int_{-\infty}^{+\infty}\eta^i[GD(\eta)-GD(\eta-1)]\,d\eta. \ \blacksquare$$

For instance the first terms of (2) can be easily rederived as follows. Let us firstly simplify our expressions.

Set

$$\varphi(x):=\int_{-\infty}^{x}GD(\eta)\,d\eta,\ \psi(x):=\int_{x}^{\infty}[GD(\eta)-1]\,d\eta,$$

$$\bar{\varphi}(x):=\int_{-\infty}^{x}GD(\eta)\,\eta\,d\eta,$$

$$\bar{\psi}(x):=\int_{x}^{\infty}[GD(\eta)-1]\,\eta\,d\eta.$$

We obtain

$$\begin{aligned}\int_{-\infty}^{+\infty}&[GD(\eta)-GD(\eta-1)]\,\eta\,d\eta\\ &=\int_{-\infty}^{0}[GD(\eta)\,\eta-GD(\eta-1)(\eta-1)]\,d\eta-\int_{-\infty}^{0}GD(\eta-1)\,d\eta\\ &+\int_{0}^{\infty}[[GD(\eta)-1]\,\eta-[GD(\eta-1)-1](\eta-1)]\,d\eta-\int_{0}^{\infty}[GD(\eta-1)-1]\,d\eta\\ &=\bar{\varphi}(0)-\bar{\varphi}(-1)+\bar{\psi}(0)-\bar{\psi}(-1)-\varphi(-1)-\psi(-1)\\ &=-[\varphi(0)+\psi(0)]+1-\int_{-1}^{0}(-\eta)\,d\eta=\frac{1}{2}-[\varphi(0)+\psi(0)].\end{aligned} \tag{9}$$

It is well known (*see* Johnson and Kotz [10], p. 272) that the extreme-value distribution function $e^{-e^{-x}}$ has mean γ and variance $\pi^2/6$. From this, we deduce, from (8), after a few elementary manipulations, that

$$\varphi(0)+\psi(0)=\frac{1}{Q_\infty \ln 2}\sum_{i=0}^{\infty} R(i)[\gamma-1+(i+1)\ln 2]. \tag{10}$$

From (A.3) the first two terms of (10) yield $-(\gamma-1)/(\ln 2)$.

By (A.5) the last term gives $-(1-\alpha)$. (9) now becomes $(3/2)-\alpha+(\gamma-1)/(\ln 2)$, which is exactly the dominant non-periodic part of (2).

As we known the variance of the distribution $e^{-e^{-x}}$, the dominant terms of the variance of *JD* could be derived by similar (but tedious) computation.

The Mellin transform techniques used in [4] allow the derivation of complete asymptotic forms for $E(JD)$ and $\mathrm{VAR}(JD)$ from Theorem 3. We will not pursue this matter here.

5. EXACT DISTRIBUTIONS FOR BINARY SEARCH TREES

We shall prove another form for the distribution. The proof is direct and avoids the preliminary analysis of the unsuccessful search done by Brown and Shubert [2] to prove (3).

THEOREM 5: *The probability distribution of the cost JB of a search in a BST is given by*

$$PB(j+1)=\frac{2^j}{n}\left[1-\frac{1}{n}\sum_{k=1}^{j}\left\lceil\frac{n}{k}\right\rceil\right], \qquad j\geq 1$$
$$PB(1)=\frac{1}{n}. \tag{11}$$

Proof: (i) Clearly $PB(1)=1/n$. Let $PB^*(j)=PB$ [success at step j | failure at step 1], with $PB(j)=(1-(1/n))\,PB^*(j)$. $PB^*(2)$ is given by:

$$\frac{1}{n}\left[\frac{1}{n-1}+\sum_{i=1}^{n-2}\left[\frac{i}{n-1}\cdot\frac{1}{i}+\frac{n-1-i}{n-1}\cdot\frac{1}{n-1-i}\right]+\frac{1}{n-1}\right], \tag{12}$$

this is similar to (5) with $1/n$ yielding the probability of left-right subtree decomposition. This leads to

$$PB^*(2)=\frac{2}{n} \qquad \text{and} \qquad PB(2)=\frac{2}{n}\left(1-\frac{1}{n}\right).$$

(ii) we must now study the operator TB, as given by (12).

On $1/n^2$, this gives

$$TB\left(\frac{1}{n^2}\right)=\frac{1}{n}\left[\frac{1}{(n-1)^2}+\sum_{i=1}^{n-2}\left[\frac{i}{n-1}\left(\frac{1}{i}\right)^2+\frac{n-1-i}{n-1}\frac{1}{(n-1-i)^2}\right]\right.$$
$$\left.+\frac{1}{(n-1)^2}\right]=\frac{2}{n(n-1)}H_{n-1}=\frac{2}{n(n-1)}\left[\overline{\begin{matrix}n\\2\end{matrix}}\right]$$

by Knuth [12], Ex. 1.2.7/6.

Hence

$$PB(3)=\frac{4}{n}\left[1-\frac{1}{n}-\frac{1}{n}\left[\overline{\begin{matrix}n\\2\end{matrix}}\right]\right].$$

More generally, for $TB\left[\left[\overline{\begin{matrix}n\\j\end{matrix}}\right]\Big/n^2\right]$, we obtain

$$\frac{1}{n}\left\{\left[\overline{\begin{matrix}n-1\\j\end{matrix}}\right]\Big/(n-1)^2\right.$$
$$+\sum_{i=1}^{n-2}\left[\frac{i}{n-1}\left[\overline{\begin{matrix}i\\j\end{matrix}}\right]\Big/i^2+\frac{n-1-i}{n-1}\left[\overline{\begin{matrix}n-1-i\\j\end{matrix}}\right]\Big/(n-1-i)^2\right]$$
$$\left.+\left[\overline{\begin{matrix}n-1\\j\end{matrix}}\right]\Big/(n-1)^2\right\}=\frac{2}{n(n-1)}\left[\overline{\begin{matrix}n\\j+1\end{matrix}}\right]$$

by (A.9). This proves (11). The form (3) given in Theorem B is now immediate from (11) and (A.11). ■

The distribution function $FB(j) = \sum_{i=1}^{j} PB(i)$ and the generating function are given by:

THEOREM 6: *The distribution function of the cost JB of a search in a BST is given by*

$$FB(j+1) = \frac{1}{n}\left[(2^{j+1}-1) - \frac{1}{n}\sum_{k=1}^{j} 2^k (2^{j-k+1}-1)\left[\overline{\begin{matrix} n \\ k \end{matrix}}\right]\right], \qquad j \geqq 1$$

$$= -\frac{1}{n} + 2PB(j+1) + \frac{1}{n^2}\left[\sum_{k=1}^{j} 2^k \left[\overline{\begin{matrix} n \\ k \end{matrix}}\right]\right].$$

The probability generating function is given by

$$GB(z) := \sum_{1}^{\infty} z^j PB(j) = \frac{1}{n^2}\left[\frac{z}{2z-1} G_n(2z) - \frac{nz}{2z-1}\right]. \tag{13}$$

Proof: FB is obtained by direct summation on (11). The generating function $GB(z)$ is easily derived as follows from (3):

$$GB(z) = \sum_{1}^{n} z^j \frac{2^{j-1}}{n^2} \sum_{k=j}^{n} \left[\overline{\begin{matrix} n \\ k \end{matrix}}\right]$$

$$= \frac{1}{n^2}\sum_{k=1}^{n}\left[\overline{\begin{matrix} n \\ k \end{matrix}}\right] 2z \frac{(2z)^k - 1}{2(2z-1)} = \frac{1}{n^2}\left[\frac{z}{2z-1} G_n(2z) - \frac{nz}{2z-1}\right]$$

by (A.10). ■

All moments of JB can be derived from (13), which may also be written as

$$\frac{z}{n(2z-1)}[(n+1) g_{n+1}(z) - 1], \tag{14}$$

where

$$g_n(z) := \prod_{i=0}^{n-2} (2z+i)/n! \tag{15}$$

is the generating function of the number of comparisons in an unsuccessful search for a $(n-1)$ keys BST (*see* Knuth [13], Ex. 6.2.2/6).

In our notation, the probability corresponding to (15) is given by Knuth as $\begin{bmatrix} n-1 \\ k \end{bmatrix} 2^k/n(n+1)$. See also Brown and Shubert [2], § 4, and Françon [6] and [7].

Remark 2: (14) could also be derived in another way. Indeed, Knuth [13], Ex. 6.2.1/25 gives a relation between internal and external nodes generating functions. Translated into probabilities, this gives (14).

6. ASYMPTOTIC DISTRIBUTION FOR *JB*

We shall prove:

THEOREM 7: *The probability distribution of the cost JB of a search in a BST tends to a Gaussian:*

$$\frac{JB-2\ln n}{(2\ln n)^{1/2}} \sim \mathcal{N}(0,1).$$

Proof: (13) is not simple enough to take the immediate limit (as for instance Knuth [12] proceeds in Ex. 1.2.10/13 for Goncharov's theorem), or use the Central limit theorem (as done by Brown and Shubert [2] for the unsuccessful search). But, by Cauchy's theorem, we obtain

$$PB(j)=\frac{1}{2\pi i}\int_\Gamma \frac{z}{z^{j+1}n(2z-1)}\left[\frac{G_n(2z)}{n}-1\right]dz \tag{16}$$

where Γ is inside the analyticity domain of the integrand and encircles the origin (it is easy to check that $z=1/2$ is not a pole of this integrand). Set $j-2\ln n=\mu\sigma$, with

$$\sigma:=(2\ln n)^{1/2}. \tag{17}$$

We must let $n\to\infty$ in (16). We will use the saddle point method (*see* Greene and Knuth [8], p. 74 for this type of technique).

(i) Firstly we must study $G_n(z)$ as $n \to \infty$. From (A.10) G_n can also be written as

$$G_n(z) = z \exp\left[\sum_{i=1}^{n-1} \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k} \left(\frac{z}{i}\right)^k \right]$$

$$= z \exp\left[z H_{n-1} + \sum_{k=2}^{\infty} (-1)^{k+1} \frac{z^k}{k} \left[\zeta(k) + O\left(\frac{1}{n^{k-1}}\right) \right] \right]$$

$$= z \exp\left[z \ln n + \gamma z + \sum_{k=2}^{\infty} (-1)^{k+1} \frac{z^k \zeta(k)}{k} + (\ln(1+z)+z) O\left(\frac{1}{n}\right) \right]$$

$$= z \frac{n^z}{\Gamma(1+z)} \exp\left[(\ln(1+z)+z) O\left(\frac{1}{n}\right) \right]$$

by Euler's classical product formula (*see* Knuth [12], Ex. 1.2.7/24). (16) gives now, asymptotically:

$$PB[j] \sim \frac{1}{2\pi i} \int_{\Gamma} \frac{1}{n(2z-1)\Gamma(1+2z)} \exp[h(z)]\, dz$$

with

$$h(z) := \ln[2 z n^{2z-1} - \Gamma(1+2z)] - j \ln z.$$

(ii) We must now find the root of $h'(s)=0$. Let $z = 1+\varepsilon$. (This choice will be justified later on). We have by Abramowitz and Stegun [1] eq. (6.1.33) and (6.3.14):

$$\Gamma(1+2z) = \Gamma(3+2\varepsilon)$$

$$\sim (2+2\varepsilon)(1+2\varepsilon) \exp[-\ln(1+2\varepsilon) + 2\varepsilon(1-\gamma) + O(\varepsilon^2)]$$

$$[\Gamma(1+2z)]' = 2\Gamma(1+2z)\psi(1+2z) = 2\Gamma(1+2z)\left[\psi(2z-1) + \frac{1}{2z-1} + \frac{1}{2z} \right]$$

$$\sim 2(2+2\varepsilon)(1+2\varepsilon) e^{-2\gamma\varepsilon} \left[-\gamma + 2\zeta(2)\varepsilon + \frac{1}{1+2\varepsilon} + \frac{1}{2+2\varepsilon} + O(\varepsilon^2) \right].$$

Now

$$h'(z)=\frac{2n^{2z-1}+4z(\ln n)n^{2z-1}-[\Gamma(1+2z)]'}{2zn^{2z-1}-\Gamma(1+2z)}-\frac{j}{z} \text{ and, with (17), } h'(z)$$

$$\sim \frac{\left\{\begin{array}{c} 2n^{1+2\varepsilon}+4(1+\varepsilon)(\ln n)n^{1+2\varepsilon}-4(1+\varepsilon)(1+2\varepsilon) \\ e^{-2\gamma\varepsilon}[-\gamma+2\zeta(2)\varepsilon+(1-2\varepsilon)+(1/2)(1-\varepsilon)+O(\varepsilon^2)] \end{array}\right\}}{2(1+\varepsilon)n^{1+2\varepsilon}-2(1+\varepsilon)(1+2\varepsilon)e^{-2\gamma\varepsilon}[1+O(\varepsilon^2)]}-\frac{2(\ln n)+\mu\sigma}{1+\varepsilon}$$

$$\sim 2\ln n+1+O\left(\frac{\ln n}{n}\right)+O(\varepsilon)-\frac{2(\ln n)+\mu\sigma}{1+\varepsilon} \quad (18)$$

$h'(s)=0$ leads to: $s=1+\mu/\sigma+0\ (1/\ln n)$, which justifies the previous choice for z.

From (18), $h''(s)\sim 2\ln n$ (and $h^k(s)=0(\ln n)$, $k\geqq 3$).

The steepest descent method leads now to (we omit the details)

$$PB(j)\sim \frac{1}{\sqrt{2\pi h''(s)}}\frac{e^{h(s)}}{n(2s-1)\Gamma(1+2s)}.$$

From (16), we finally obtain

$$PB(j)\sim \frac{1}{\sqrt{2\pi.2\ln n}(1+\mu/\sigma)^{2(\ln n)+\mu\sigma}n(1+2\mu/\sigma)}\left[\frac{2(1+\mu/\sigma)n^{2+2\mu/\sigma}}{n\Gamma(3+2\mu/\sigma)}-1\right].$$

But now, it is easily checked that $(1+\mu/\sigma)^{2(\ln n)+\mu\sigma}\sim e^{\mu\sigma}e^{\mu^2/2}$. We finally deduce

$$PB(j)\sim \frac{1}{\sqrt{2\pi.2\ln n}}e^{-\mu^2/2}$$

which proves the theorem. Each term in the summation part of $FB(j)$ (*see* Theorem 6) can be analysed in the same manner and leads of course to a Gaussian expression. ■

5. CONCLUSION

A direct approach has allowed us to derive the exact and asymptotic distributions of the number of steps in DST and BST. The first asymptotic distribution is related to the extreme-value classical distribution, the other one is simply Gaussian. We intend to pursue this direct approach on m-ary search tree analysis.

ACKNOWLEDGMENTS

The author is indebted to P. Flajolet for suggesting some problems which led to our interest into tree algorithms and to H. Prodinger who kindly supplied of copy of [11] before publication. The pertinent comments of a referee led to substantial improvement in the presentation.

APPENDIX

(i) Using two classical Euler identities, we will derive several sommation formulas we need in the text. These identities are:

$$\prod_{k=0}^{\infty} \frac{1}{1-q^k z} = 1 + \sum_{i=1}^{\infty} z^i \Big/ \prod_{k=1}^{i} (1-q^k) \tag{A.1}$$

$$\prod_{k=0}^{\infty} (1+q^k z) = 1 + \sum_{i=1}^{\infty} z^i q^{i(i-1)/2} \Big/ \prod_{k=1}^{i} (1-q^k) \tag{A.2}$$

(for a simple proof *see* Knuth [13], Ex. 5.1.1/16). As mentioned in Knuth [13], Ex. 6.3/26, it is easy to extract from (A.2) two expressions for $Q(\infty)$:

let first $z=-1/2$, $q=1/2$, this gives

$$Q(\infty) = -\sum_{i=0}^{\infty} R(i). \tag{A.3}$$

Letting now $z=-1/4$, $q=1/2$, we obtain

$$2Q(\infty) = -\sum_{i=0}^{\infty} 2^{-i} R(i). \tag{A.4}$$

Multiplying (A.2) by z and differentiating with respect to z gives

$$\prod_{k=0}^{\infty} (1+q^k z)\left[1 + z\sum_{i=0}^{\infty} \frac{q^i}{1+q^i z}\right] = 1 + \sum_{i=1}^{\infty} (i+1) z^i q^{i(i-1)/2} \Big/ \prod_{k=1}^{i} (1-q^k).$$

Letting $z=-1/2$, $q=1/2$ gives

$$Q(\infty)\left[1 - \sum_{i=1}^{\infty} \frac{1}{2^i-1}\right] = -\sum_{i=0}^{\infty} (i+1) R(i). \tag{A.5}$$

Set $z=-u/2$, $q=1/2$ in (A.2). This yields the generating function of $R(i)$:

$$\prod_{k=1}^{\infty}\left(1-\frac{u}{2^k}\right)=-\sum_{i=0}^{\infty}u^i R(i). \tag{A.6}$$

Set $z=u$, $q=1/2$ in (A.1). This gives

$$\prod_{k=0}^{\infty}\frac{1}{1-(u/2^k)}=\sum_{i=0}^{\infty}\frac{u^i}{Q(i)}. \tag{A.7}$$

(This identity is also used by Flajolet and Sedgewick in [5]).

Multiplying (A.6) by (A.7) leads to

$$\frac{1}{1-u}=-\sum_{i=0}^{\infty}u^i\sum_{k=0}^{i}\frac{R(i-k)}{Q(k)}, \quad \text{hence} \quad -\sum_{k=0}^{i}\frac{R(i-k)}{Q(k)}=1. \tag{A.8}$$

(ii) A few relations on $\overline{\left[\begin{matrix}n\\k\end{matrix}\right]}$ will be now established.

We start from

$$H_{n-1}=\sum_{i=1}^{n-1}\frac{1}{i}=\overline{\left[\begin{matrix}n\\2\end{matrix}\right]}.$$

Define

$$\overline{\left[\begin{matrix}n\\3\end{matrix}\right]}:=\sum_{i=1}^{n-1}\frac{H_{i-1}}{i}=\sum_{i=1}^{n-1}\overline{\left[\begin{matrix}i\\2\end{matrix}\right]}\Big/i$$

and more generally

$$\overline{\left[\begin{matrix}n\\j+1\end{matrix}\right]}:=\sum_{i=1}^{n-1}\overline{\left[\begin{matrix}i\\j\end{matrix}\right]}\Big/i. \tag{A.9}$$

It is easily seen that the generating function of $\overline{\left[\begin{matrix}n\\j\end{matrix}\right]}$ is given by

$$\sum_{j=1}^{n}\overline{\left[\begin{matrix}n\\j\end{matrix}\right]}z^j=z(1+z)\left(1+\frac{z}{2}\right)\ldots\left(1+\frac{z}{n-1}\right)=G_n(z), \tag{A.10}$$

say and

$$G_n(1)=n. \tag{A.11}$$

We know (*see* Knuth [12], § 1.2.9 eq. (27) that

$$z(z+1)\dots(z+n-1)=\sum_{j=1}^{n}\left[\begin{matrix} n \\ j \end{matrix}\right] z^j, \quad \text{hence} \quad \overline{\left[\begin{matrix} n \\ j \end{matrix}\right]} \equiv \left[\begin{matrix} n \\ j \end{matrix}\right] \Big/ (n-1)!$$

Note that $\overline{\left[\begin{matrix} n \\ j \end{matrix}\right]} \Big/ n$ appears in another context: it is the probability of $(j-1)$ changes of the maximum in Algorithm M of Knuth [12], § 1.2.10.

REFERENCES

1. M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, 1965, Dover Publications.
2. G. G. Brown and B. O. Shubert, *On Random Binary Trees*, Math. Op. Res., Vol. 9, No. 1, 1984, pp. 43-65.
3. L. Devroye, *A Probabilistic Analysis of Height of Tries and of the Complexity of Triesort*, Acta Informatica, Vol. 21, 1984, pp. 229-237.
4. P. Flajolet, *Approximate Counting: a Detailed Analysis*, BIT, Vol. 25, 1985, pp. 113-134.
5. P. Flajolet and R. Sedgewick, *Digital Search Trees Revisited*, S.I.A.M. J. Comp., Vol. 15, No. 3, 1986, pp. 748-767.
6. J. Françon, *On the Analysis of Algorithms for Trees*, Th. Comp. Sc., Vol. 4, 1977, pp. 155-169.
7. J. Françon, *Arbres binaires de recherche : propriétés combinatoires et applications*, RAIRO, Inf. Th., Vol. 10, No. 12, 1986; pp. 35-50.
8. D. H. Greene and D. E. Knuth, *Mathematics for the analysis of algorithms*, 1981, Birkhäuser.
9. Ph. Jacquet and M. Regnier, *Limiting Distributions for Trie Parameters*, Proc. C.A.A.P. 86, Lecture Notes in Comp. Sc., Vol. 214, 1986, pp. 198-210.
10. N. L. Johnson and S. Kotz, *Distribution in statistics: continuous univariate distributions*, 1970, Wiley.
11. P. Kirschenhofer and H. Prodinger, *Some Further Results on Digital Search Trees*, Proc. I.C.A.L.P., 1986, Lect. Notes Comp. Sc., Vol. 226, pp. 177-185.
12. D. E. Knuth, *The Art of Computer Programming*, Vol. I, 1969, Addison-Wesley.
13. D. E. Knuth, *The Art of Computer Programming*, Vol. III, 1973, Addison-Wesley.
14. G. Louchard, *The Brownian Motion: a Neglected Tool for the Complexity Analysis of Sorted Tables Manipulations*, R.A.I.R.O., Inf. Th., Vol. 4, 1983, pp. 365-385.
15. G. Louchard, *Brownian Motion and Algorithms Complexity*, B.I.T., Vol. 26, 1986, pp. 17-34.
16. B. Pittel, *Paths in a Random Digital Tree: Limiting Distributions*, Adv. Appl. Prob., Vol. 18, 1986, pp. 139-155.