A Multivariate View of Random Bucket Digital Search Trees

Friedrich Hubalek*
Institut für Finanz- und Versicherungsmathematik
Technische Universität Wien
1040 Vienna
Austria

Hsien-Kuei Hwang†
Institute of Statistical Science
Academia Sinica
Taipei 115
Taiwan

William Lew
Center for Technology & Engineering
The General Accounting Office
Washington, D. C. 20548
U. S. A.

Hosam Mahmoud‡
Department of Statistics
The George Washington University
Washington, D. C. 20052
U. S. A.

Helmut Prodinger§
School of Mathematics
University of the Witwatersrand
P. O. Wits, 2050 Johannesburg
South Africa

(October 25, 2001)

Abstract

We take a multivariate view of digital search trees by studying the number of nodes of different types that may coexist in a bucket digital search tree as it grows under an arbitrary memory management system. We obtain the mean of each type of node, as well as the entire covariance matrix between types, whereupon weak laws of large numbers follow from the orders of magnitude (the norming constants include oscillating functions). The result can be easily interpreted for practical systems like paging, heaps and UNIX's buddy system. The covariance results call for developing a Mellin convolution method, where convoluted numerical sequences are handled by convolutions of their Mellin transforms. Furthermore, we use a method of moments to show that the distribution is asymptotically normal. The method of proof is of some generality and is applicable to other parameters like path length and size in random tries and Patricia tries.

Key words and Phrases: Random trees, bucketing, searching, probabilistic analysis of algorithms, Mellin transform, Euler transform, singularity analysis, asymptotic normality.

---

*This author's research has been supported by Austrian National Bank Project No. 4995.

†The work of this author was done while he was visiting School of Computer Science, McGill University. He thanks the School for hospitality and support.

‡This author's research has been partially supported by NSA grant MDA904-92-H-3086.

§This author's research has been supported by The John Knopfmacher Centre of Applicable Analysis and Number Theory.

AMS 2000 Subject Classification: Primary 68W40, 05C05; secondary 60C05, 60F05.

1 Introduction

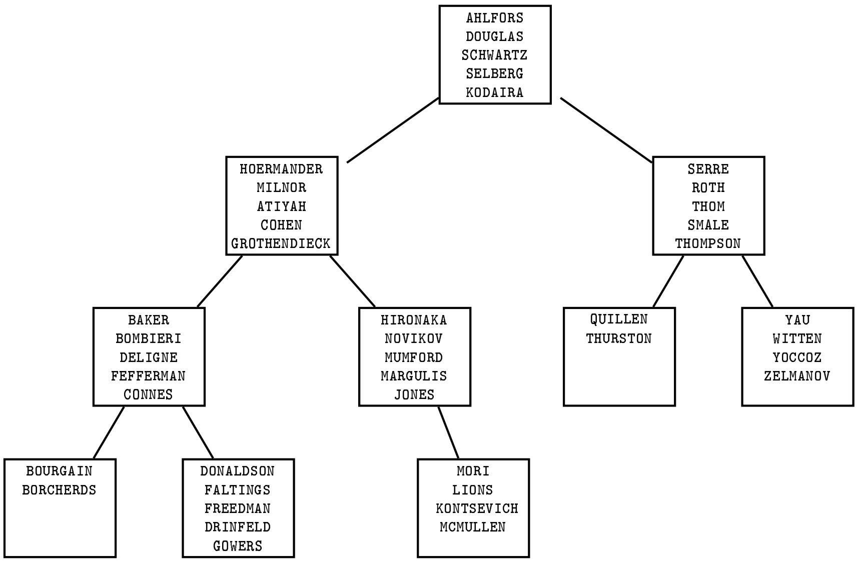
Tree structures may be designed to hold blocks of data in their nodes instead of just individual items. The advantage of such a "bucketing" scheme is multifold: First it helps reduce root-to-node path lengths in the tree, an important issue in communication with external devices like disks; second, it improves the storage utilization (especially when more balance properties are enforced), a key issue in database applications; third, the page faults are usually reduced for large bucket capacity. Bucketing has been an important tool in diverse algorithm design problems ranging from hashing to computational geometry; see, for example, [8], [16], [30], [35], [40]. In this paper, we study the probabilistic properties of space measures on random binary bucket digital search trees. Roughly, we show that the mean and the variance are both linear, and the limiting distribution is normal.

In binary bucket digital search trees (or generalized digital search trees; see [13]) with bucket capacity b, groups of up to b keys are stored in the nodes, and branching to the subtrees is based on a digital search method. These trees are used to store keys taken from some alphabet. We consider in this paper the binary alphabet, whose digits are most commonly used for representing files in a computer's internal memory. (Extension to larger alphabets does not lead to substantial complications.) Thus the keys considered will be viewed as sequences of binary digits from the set $\{0, 1\}$. In binary bucket digital search trees, the first b keys are stored in the root node; a subsequent key is guided to the left subtree if its first symbol is 0, or to the right if the first symbol is 1. The subtrees are recursively constructed by the same algorithm, but if the subtree root is at level ℓ, the $(\ell + 1)$st bit of the key is used for branching. See Figure 1 for an illustration, where a list of names is stored. Each name is first translated into a binary string by taking a juxtaposed transformation of its letters into bytes according to a standard interchange code.

The probability model usually assumed for the analysis of digital trees is the Bernoulli model. In this model the bits of a key are equiprobable and are independently taken from a stationary ergodic source. Thus each key is an infinite sequence of Bernoulli trials. Given n such independently sampled keys from the data model, we construct a bucket digital search tree with capacity $b \geq 1$. The resulting tree will be called in this paper a *random bucket digital search tree on n keys* or simply a *random tree*. Note that most of our analysis also carries over to the case $b = 0$, which essentially corresponds to tries in which only the leaf nodes hold the keys.

A nice property enjoyed by general random digital search trees like tries, PATRICIA trees, and digital search trees, etc. is that the average height is almost optimal, namely, asymptotic to $\log_2 n$; see [39]. This is mainly due to the high concentration of the binomial distribution near its mean. For large bucket capacity, [13] found out that on average these trees take $1/\log 2 \approx 1.44$ times as much space as the size of the data set itself when they are implemented under a fixed allocation scheme. We investigate in this paper the space utilization of random bucket digital search trees by looking at diverse methods of memory management including popular and practical methods such as heaps (a space control mechanism available in the PASCAL programming language) and the buddy system used in UNIX. We show that the mean and the variance of general weighted linear combinations of node types are both linear. We further derive a central limit theorem, indicating in some ways the robustness of the general storage scheme for bucket digital search trees under random input.

Our approach is roughly described as follows. The starting point is the recurrence relation for



Record No.	Key	Record	Record No.	Key	Record
1	00001 ...	AHLFORS	22	00110 ...	FEFFERMAN
2	00100 ...	DOUGLAS	23	01110 ...	MARGULIS
3	10110 ...	SCHWARTZ	24	10010 ...	QUILLEN
4	10110 ...	SELBERG	25	00011 ...	CONNES
5	01100 ...	KODAIRA	26	10111 ...	THURSTON
6	10110 ...	SERRE	27	11100 ...	YAU
7	10011 ...	ROTH	28	00100 ...	DONALDSON
8	10111 ...	THOM	29	00110 ...	FALTINGS
9	01000 ...	HOERMANDER	30	00110 ...	FREEDMAN
10	01110 ...	MILNOR	31	00100 ...	DRINFELD
11	00001 ...	ATIYAH	32	01011 ...	JONES
12	00011 ...	COHEN	33	01110 ...	MORI
13	00111 ...	GROTHENDIECK	34	11010 ...	WITTEN
14	10110 ...	SMALE	35	01101 ...	LIONS
15	00010 ...	BAKER	36	11100 ...	YOCCOZ
16	01000 ...	HIRONAKA	37	00010 ...	BOURGAIN
17	01111 ...	NOVIKOV	38	11101 ...	ZELMANOV
18	10111 ...	THOMPSON	39	00010 ...	BORCHERDS
19	00010 ...	BOMBIERI	40	00111 ...	GOWERS
20	01110 ...	MUMFORD	41	01100 ...	KONTSEVICH
21	00100 ...	DELIGNE	42	01110 ...	MCMULLEN

Figure 1: Illustration of a bucket digital search tree with node capacity 5 and 42 entries.

the moment generating function $M_n(y)$ of the weighted sum of node types, which is of the form

$$M_{n+b}(y) = \frac{e^{cy}}{2^n} \sum_{0\le k\le n} \binom{n}{k} M_k(y) M_{n-k}(y) \qquad (n \ge 1),$$

with suitable initial conditions, where c is a constant. Although this recurrence does not seem to be exactly solvable in terms of elementary functions, we can capture some major probabilistic properties by its moments, which satisfy the same recurrence of the type

$$\xi_{n+b} = \eta_{n+b} + 2^{1-n} \sum_{0\le k\le n} \binom{n}{k} \xi_k \qquad (n \ge 1),$$

with suitable initial conditions, where η_n is the "toll sequence". Such a recurrence can be solved either exactly or asymptotically. Roughly, the intuition is that if η_n is small (say $\ll n$), then ξ_n is of linear order (with periodic fluctuations for the constant); if η_n is linear, then ξ_n is of order $n \log n$; finally, if η_n is superlinear then ξ_n is of the same order as η_n. Such transfers from the asymptotics of the "toll sequence" to that of the total cost are the key steps needed to prove the asymptotic normality via the method of moments (see [4]). For more applications of this approach to recursive random structures, see [3], [5], [6], [22], [23]. We will also indicate more examples of applications of such an approach in Section 6. It should be mentioned that although the linearity of the mean and variance can in general be easily proved by elementary arguments, the precise determination of the leading constants (which are always periodic functions) is much more difficult.

The notion of increasing sizes for different type nodes of a linked data structure has been introduced before in the class of balanced trees where order statistics of the keys, instead of their digital composition, are used to guide searching. Some authors called these buckets *elastic*. [14] and [33] use elastic buckets for implementing increasing size nodes when they overflow instead of the usual splitting of nodes in B–trees and their derivatives. [2] analyzed the storage utilization of B^+–trees with elastic buckets. The notion of elastic buckets is also implicit in [17] who computed the average number of nodes of each type in a paged quadtree. [1] finds the optimal second size in the case when two page sizes are allowed for a multiway search tree. [31] studied the space of the multiway search trees under different memory management systems. The digital bucket trie and its close relative PATRICIA tree have been studied in [28], and in [37].

We study the space requirement for random bucket digital trees as they grow under different operating systems that can maintain different types of nodes. We first discuss in Section 2.1 some relevant practical issues. In Section 2.2, the associated problem of conversion between these different types is briefly discussed. The conversion problem is related to the overall construction cost of the tree. In Section 2.3, we outline a paradigm for multivariate analysis. We then proceed in Section 3 with the technical aspects of our study. It is natural here to formulate the problem in terms of exponential generating functions but the resolution of the resulting differential-functional equation is less transparent (see [34], [26]). The technique of [13], which switches back to the ordinary generating via Euler's transform, gives an exactly solvable functional equation for the ordinary generating, an equation that can be handled asymptotically by the Mellin transform and its inverse. We then apply the singularity analysis of [12] to transfer the asymptotic results for the ordinary generating function to that of the coefficients.

The mean is handled in Section 4. In the study of the variance (Section 5), technical subtleties are encountered in certain convolution terms. These are finessed by a recent technique devised by [18, 19], where the convoluted sums are handled by convolutions of their Mellin transform. We show

how to derive the asymptotic mean number of nodes of each type and the variance-covariance matrix. These results give us a weak law of large numbers for the amount of memory space needed for a growing random tree. The results are strengthened in Section 6 to a central limit theorem, with oscillating norming factors.

The general procedures of Section 3 are then used to carry out the calculation for some illustrating examples in Section 7 where paging, heaps, and the buddy system (UNIX's memory management scheme) are selected as examples because of their popularity in practice.

2 Practical Considerations

In practice, a random tree grows under the specifications of an operating system having a component that controls memory management.

2.1 Memory Management Schemes

One could view all the schemes of memory management as special cases of a general scheme where the computer system releases the space in arbitrarily predesignated block sizes. A particular tree implemented under this general scheme will then have to use the smallest system's block as the smallest node, the second smallest system's block as the next type of nodes and so on. In this system one may have p types of nodes, say (each of key capacity $\le b$), of which the first $p-1$ are pointerless terminal nodes or *leaves* and the p-th type has maximal key-capacity b and carries two pointers. In terms of keys, we assume the capacities of these nodes to be $0 = \beta_0 < \beta_1 < \beta_2 < \ldots < \beta_{p-1} \le \beta_p = b$. (Empty subtrees are thought of as an empty type of nodes with capacity $\beta_0 = 0$; they correspond to **nil** pointers.) All inequalities are strict, except possibly the last one—if $\beta_{p-1} = \beta_p$, the last two types carry the same number of nodes, but of course the last type contains some more bytes for the right and left pointers. We call the nodes with key-capacity β_j *type j nodes*, $j = 1, \ldots, p-1$. Nodes with pointers are of type p and are called *internal*.

Consider the most detailed multivariate problem of the number of leaves holding j keys, $j = 1, 2, \ldots, b$, under a fixed allocation scheme (only one node type of key capacity b, i.e. $p = 1$). Let $X_n^{(i)}$ be the number of leaves holding i keys in a random tree, $1 \le i \le b$, after n insertions; let $X_n^{(b+1)}$ denote the number of internal nodes after n insertions. Further, define

$$\mathbf{X}_n = \left(X_n^{(1)}, X_n^{(2)}, \ldots, X_n^{(b+1)}\right)^T$$

(the superscript T stands for the transpose operation). If one could characterize the probabilistic behavior of this most detailed problem, one would also be able to characterize the probabilistic behavior of the problem under any other scheme by lumping nodes as follows. By implementing all the nodes holding $\beta_{i-1}+1, \beta_{i-1}+2, \ldots, \beta_i$ keys under fixed allocation as type i in the general scheme, one can compute $Y_n^{(i)}$, the number of type i nodes in the general scheme after n insertions, from the relation

$$Y_n^{(i)} = X_n^{(\beta_{i-1}+1)} + \cdots + X_n^{(\beta_i)}.$$

Thus the vector $\mathbf{Y}_n = \left(Y_n^{(1)}, \ldots, Y_n^{(p)}\right)^T$ is obtained from $\mathbf{X}_n$ via an obvious matrix multiplication and distributional properties of $\mathbf{Y}_n$ will follow from those of $\mathbf{X}_n$.

In a practical implementation of multitype-node trees, some extra space must be reserved within each node to serve as a *hidden tag*. These hidden bytes may hold information like the size of the bucket—most algorithms manipulating a multitype structure would demand to know the type of a node before it is processed. In most memory allocation schemes, the hidden tag has a uniform length in all nonempty buckets regardless of their size.

Some particular practical instances are described as follows:

(i) Paging: This is a fixed block allocation scheme with only one type of non-empty nodes ($p = 1$) with key-capacity $\beta_1 = b$ keys. A node (page) carries two pointers as well. No hidden tag is needed for this scheme.

(ii) Exact fit: This scheme can be implemented in computer systems that provide access to a "heap" of storage where the user can request an exact number of bytes. Under exact fit the first b nodes are pointerless and have key-capacities $\beta_j = j$ for $j = 1, \ldots, b$, and type $b+1$ with key-capacity $\beta_{b+1} = b$ is the only type that carries the pointers. A hidden tag of fixed length is needed in each node.

(iii) The buddy system: This is a choice in the UNIX operating system. This scheme allocates only blocks of sizes 2^k words of memory for some positive integer k. The maximal block capacity w (in memory words) must then be chosen as a proper-power-of-2 word. With 1–word hidden tag, the capacities of the blocks (in terms of words corresponding to the pointers to the actual keys) are $2^j - 1$, for $j = 0, \ldots, \log_2 w$. In this case we may have a situation intermediate between exact-fit schemes and uniform allocation schemes. Partial filling may take place but it may happen in nodes of size less than the maximum capacity chosen, still achieving some degree of saving in space.

Of course, in practice, digital keys are finite. However, in many applications they are rather long. For example, in DNA sequencing problems, a typical key may be several thousand, or even several hundred thousand bits long. Thus, the Bernoulli model that assumes infinitely long keys is only a theoretical device toward understanding the true behavior of digital search trees. Furthermore, practical data keys may not be all of uniform length. Their length is even sometimes unpredictable. In this case, a filled data block or tree node may actually hold a variable number of variable length keys, and programming the tree becomes quite difficult. Therefore, digital data in real indexed database applications are usually assumed to reside in a sequential data file stored somewhere in memory (or external memory) outside the digital tree, and each key is clearly demarcated by boundary characters. The digital tree itself is then programmed to serve as an index storing uniform length pointers to the keys, rather than storing the keys themselves, in the tree nodes. This greatly simplifies updating algorithms on digital trees. Under this latter assumption, the number of keys (even if they have variable length) associated with a filled tree node is our constant b.

2.2 Conversion Between Types

Conversion between types is necessary when a tree with elastic buckets is growing; repeated insertions in a node eventually overflow that node and the node is stretched or "converted" into a node of the next larger type, which explains the choice of the name elastic bucket.

One could study the total number of conversions. But these conversions do not have the same cost. In the general scheme a conversion from type $i-1$ into type i involves the relocation of β_{i-1} keys. The more detailed *conversion vector*

$$\mathbf{C}_n = \left(C_n^{(1)}, \ldots, C_n^{(p)} \right)^T,$$

with $C_n^{(i)}$ being the number of conversions from type $i-1$ into type i, after n key insertions have been made, serves as a more reasonable measure of the construction cost of the whole tree ($C_n^{(1)}$ is the cost of converting a **nil** pointer into a pointer to a new node with capacity β_1 keys and the cost of allocating that new node).

The joint distribution of $\mathbf{C}_n$ may be obtained from that of $\mathbf{Y}_n$ since the components of these two vectors are related as follows:

$$C_n^{(i)} = Y_n^{(i)} + \cdots + Y_n^{(p)};$$

this can be seen from the fact that every node of type i or higher has passed at some point in its history via type i, and thus has called for a conversion from type $i-1$ into type i. So, again, $\mathbf{C}_n$ and $\mathbf{Y}_n$ are related by a simple matrix multiplication and distributional information on either vector will furnish distributional information on the other.

2.3 A Paradigm for Multivariate Analysis

As discussed in Sections 2.1–2.2, all the desired multivariate problems are well encapsulated in considering

$$Z_n = a_1 X_n^{(1)} + \cdots + a_{b+1} X_n^{(b+1)} = \mathbf{a}^T \mathbf{X}_n,$$

where a_j for $j = 1, \ldots, b+1$ are arbitrary real numbers and $\mathbf{a}^T = (a_1, \ldots, a_{b+1})^T$. There are several advantages in studying this random variable:

(i) One avoids duplicating the effort for each type; if some moment of Z_n is known, then the same moment for a node of a certain type are obtained by the proper settings for the a_i's as illustrated next for the mean and variance. Setting $a_i = 1$, $a_j = 0$, for $j \neq i$, $i = 1, \ldots, b+1$, simply reduces Z_n to $X_n^{(i)}$. With this setting, the average and variance of $X_n^{(i)}$ are just the average and variance of Z_n. The setting $a_i = a_j = 1$, for $j \neq i$, $i = 1, \ldots, b+1$, in Z_n reduces its variance formula to

$$\mathbf{Var}[Z_n] = \mathbf{Var}\big[X_n^{(i)} + X_n^{(j)}\big].$$

Having computed all individual variances for the different components of $\mathbf{X}_n$, we can find the covariance between $X_n^{(i)}$ and $X_n^{(j)}$ from the elementary formula

$$\mathbf{Var}[Z_n] = \mathbf{Var}\big[X_n^{(i)}\big] + \mathbf{Var}\big[X_n^{(j)}\big] + 2\mathbf{Cov}\big[X_n^{(i)}, X_n^{(j)}\big].$$

(ii) It is usually the case that Z_n satisfies recurrences similar to those satisfied by any particular type.

(iii) The vector $\mathbf{Y}_n$ is obtained from $\mathbf{X}_n$ via the relation

$$\mathbf{Y}_n = \mathbf{K}\mathbf{X}_n,$$

where $\mathbf{K} = [k_{ij}]$ is the $p \times (b+1)$ matrix whose ith row has 1's at positions $\beta_{i-1}+1, \beta_{i-1}+2, \ldots, \beta_i$; the rest of the positions are 0. Thus the mean and covariance matrix for $\mathbf{Y}_n$ are given by:

$$\mathbf{E}[\mathbf{Y}_n] = \mathbf{K}\,\mathbf{E}[\mathbf{X}_n],$$

and

$$\mathbf{Cov}[\mathbf{Y}_n] = \mathbf{K}^T \mathbf{Cov}[\mathbf{X}_n]\,\mathbf{K}.$$

(iv) The conversion vector $\mathbf{C}_n$ is obtained from $\mathbf{Y}_n$ via the multiplication

$$\mathbf{C}_n = \mathbf{M}\mathbf{Y}_n,$$

where $\mathbf{M} = [m_{ij}]$ is the $p \times p$ upper diagonal matrix with components

$$m_{ij} = \begin{cases} 1, & \text{if } i \geq j; \\ 0, & \text{otherwise.} \end{cases}$$

And so,

$$\mathbf{C}_n = \mathbf{M}\,\mathbf{K}\,\mathbf{X}_n;$$

it follows immediately that

$$\mathbf{E}[\mathbf{C}_n] = \mathbf{M}\,\mathbf{K}\,\mathbf{E}[\mathbf{X}_n],$$

and

$$\mathbf{Cov}[\mathbf{C}_n] = \mathbf{K}^T\,\mathbf{M}^T\,\mathbf{Cov}[\mathbf{X}_n]\,\mathbf{M}\,\mathbf{K}.$$

(v) The technical results of the next section will show that, for any $\mathbf{a}$, both mean and variance are of the exact order n. An application of Chebychev's inequality then reveals that, for any $\varepsilon > 0$,

$$\mathrm{P}\left\{\left|\frac{Z_n}{\mathbf{E}[Z_n]} - 1\right| \geq \varepsilon\right\} \leq \frac{\mathbf{Var}[Z_n]}{\varepsilon^2\mathbf{E}^2[Z_n]} = O\left(n^{-1}\right) \to 0.$$

Thus $Z_n/\mathbf{E}[Z_n] \to 1$ in probability. We strengthen the result by proving that the limiting distribution of Z_n is normal.

3 Preliminaries

3.1 Moment generating function

Let $X_n^{(i)}$ and the linear combination Z_n be defined as in Section 2. We first derive a recurrence relation for the moment generating function

$$M_n(y) := \mathbf{E}\left[e^{Z_n y}\right].$$

Lemma 1. *The moment generating function $M_n(y)$ of Z_n satisfies $M_0(y) = 1$, $M_n(y) = e^{a_n y}$ for $1 \leq n \leq b$, and*

$$M_{n+b}(y) = \frac{e^{a_{b+1}y}}{2^n} \sum_{0 \leq k \leq n} \binom{n}{k} M_k(y) M_{n-k}(y) \qquad (n \geq 1). \tag{1}$$

Proof. By our assumptions on the probability model, the number of keys that end up in the left subtree $\mathcal{L}_n$ has a binomial distribution of parameters $n-b$ and $1/2$. Given that k keys end up in the left subtree, the individual types $1, 2, \ldots, b$ satisfy the *conditional* recurrence

$$X_{n+b}^{(i)} = X_k^{(i)} + \tilde{X}_{n-k}^{(i)} \qquad (n \geq 1),$$

where $\tilde{X}_n^{(i)} \stackrel{d}{=} X_n^{(i)}$, and is independent of $X_k^{(\ell)}$ for all feasible i, k, ℓ and n. This follows from the behavior of the construction algorithm (once it decides to continue the search within one subtree, it never looks at the structure of the other subtree, and within a subtree it behaves in a recursive manner giving rise to the afore-mentioned identical distribution). As the root of the tree is of type $b+1$ when $n > b$, type $b+1$ satisfies a slightly different recurrence:

$$X_{n+b}^{(i)} = X_k^{(i)} + \tilde{X}_{n-k}^{(i)} + 1 \qquad (n \geq 1),$$

where $\tilde{X}_n^{(b+1)} \stackrel{d}{=} X_n^{(b+1)}$, and is independent of $X_k^{(i)}$ for all i, k, n. Taking a linear combination of these conditional recurrences, with a_k the coefficient of $X_n^{(k)}$, we obtain the conditional recurrence (given $\mathcal{L}_{n+b} = k$):

$$Z_{n+b} = Z_k + \tilde{Z}_{n-k} + a_{b+1},$$

where $\tilde{Z}_k$ is distributed like Z_k, and is independent of Z_i for all i and k. Thus,

$$\begin{aligned}
\mathbf{E}\left[e^{Z_{n+b}y}\right] &= \sum_{0\leq k\leq n} \mathbf{E}\left[e^{Z_{n+b}y} \middle| \mathcal{L}_{n+b} = k\right] \times \mathbf{P}\{\mathcal{L}_{n+b} = k\} \\
&= 2^{-n} e^{a_{b+1}y} \sum_{0\leq k\leq n} \mathbf{E}\left[e^{(Z_k+\tilde{Z}_{n-k})y}\right] \times \binom{n}{k}.
\end{aligned}$$

Using the independence and identical distribution assumptions the lemma follows. □

3.2 Euler transform

Given a sequence $\{\xi_n\}_{n=0}^{\infty}$, let $F_\xi(z) := \sum_{n\geq 0} \xi_n z^n$ be the ordinary generating function of ξ_n. Define the Euler transform of ξ_n by

$$G_\xi(z) := \frac{1}{1+z} F_\xi\left(\frac{z}{1+z}\right) = \sum_{n\geq 0} \xi_n \frac{z^n}{(1+z)^{n+1}}.$$

Then

$$F_\xi(z) = \frac{1}{1-z} G_\xi\left(\frac{z}{1-z}\right). \tag{2}$$

Lemma 2. *Assume that ξ_n is a sequence defined recursively by $\xi_0 = 0$ and*

$$\xi_n = \eta_n + 2^{1-n+b} \sum_{0\leq k\leq n-b} \binom{n-b}{k} \xi_k \qquad (n \geq 1), \tag{3}$$

where $\{\eta_n\}_{n\geq 1}$ *is any given sequence and the sum on the right hand side is zero for* $n \leq b$. *Then* $G_\xi(z)$ *satisfies*

$$G_\xi(z) = \sum_{k\geq 0} 2^k G_\eta(z/2^k) \prod_{0\leq j<k} \left(\frac{z/2^j}{1+z/2^j}\right)^b, \tag{4}$$

where the right-hand side is taken as a formal power series expansion when it diverges.

Proof. The recurrence (3) is translated into the functional equation

$$G_\xi(z) = G_\eta(z) + \frac{2z^b}{(1+z)^b} G_\xi(z/2).$$

The required expression follows from iterating this equation. □

Binomial transform. The Euler transform of a sequence $\{\xi_n\}$ is the ordinary generating function of the binomial transform $\hat{\xi}_n := \sum_k \binom{n}{k}(-1)^{n-k}\xi_k$ of ξ_n.

3.3 Mellin transforms

Given a function $A(x)$ defined on the positive real axis, the Mellin transform of A, denoted by A^*, is defined by

$$A^*(s) := \int_0^\infty A(x)x^{s-1}dx,$$

whenever the integral is convergent. Typically the integral converges in a vertical strip in the s complex plane $\sigma_- < \Re s < \sigma_+$, and defines there an analytic function. Following [11], we use the notation $\langle \sigma_-, \sigma_+ \rangle$ to denote this strip.

We also use consistently another notation of [11] for the singular (or Laurent) expansion of a Mellin transform:

If $A(z)$ has the Laurent series expansion $A_j(z)$ for z near z_j, then we write

$$A(z) \asymp \sum_{j\in\mathcal{J}} [A_j(z)]_{z=z_j},$$

where $\mathcal{J}$ is some index set. For example (see [11])

$$\frac{1}{s^2(s^2-1)} \asymp \frac{1}{s+1} + \left[\frac{1}{s^2} - \frac{1}{s}\right]_{s=0} + \frac{1}{2(s-1)}.$$

where we drop the brackets and subscripts in the case when the singularity is a simple pole.

3.4 Singularity Analysis

The singularity analysis of [12] is useful in deriving the asymptotic behavior of coefficients of nicely behaved analytic functions.

We briefly state a special form for our purposes; see [12] for more precise expansions.

If $\xi(z) = \sum_n \xi_n z^n$ is analytic in the unit circle, and analytically continuable to the region

$$\mathcal{D} := \{z \,:\, |z| \leq 1+\varepsilon, |\arg(1-z)| \geq \varepsilon'\} \setminus \{1\} \qquad (0 < \varepsilon' < \pi/2),$$

where it satisfies

$$\xi(z) = c(1-z)^{-\alpha}\left(1 + O\left(|1-z|^{\kappa}\right)\right) \qquad (\alpha, \kappa > 0),$$

then

$$\xi_n = \frac{c}{\Gamma(\alpha)} n^{\alpha-1}\left(1 + O(n^{\max\{-\kappa,-1\}}\right).$$

In particular, if $c = 0$, then $\xi_n = o(n^{\alpha-1})$.

3.5 Flajolet and Richmond's Approach

The method of proof we use for the first two moments follows the analytic approach introduced by [13], and is sketched as follows.

To derive the asymptotic expansions for the moments, we need to solve asymptotically the recurrence (3). We consider first the Euler transform $G_\xi(z)$ of ξ_n, which satisfies a special form that can be explicitly expressed as a series. In particular, the series is a harmonic sum (see [11]). We then apply the Mellin transform to derive asymptotic approximations to the Euler transform $G_\xi(z)$, as $|z| \to \infty$ and $|\arg z| \leq \pi/2 + \varepsilon$, $0 < \varepsilon < \pi/2$. The same asymptotic result is translated to an asymptotic formula for the ordinary generating function $F_\xi(z)$, as $z \to 1$ in $\mathcal{D}$. Then we apply the singularity analysis (see [12]) to complete the asymptotic analysis.

4 Mean of Z_n

Recurrence. Let $\mu_n := M_n'(0) = \mathbf{E}[Z_n]$. For brevity, we define $a_n := a_{b+1}$, for $n > b+1$. By (1), we have $\mu_0 = 0$, $\mu_j = a_j$ for $1 \leq j \leq b$ and

$$\mu_{n+b} = a_{n+b} + 2^{1-n} \sum_{0 \leq k \leq n} \binom{n}{k} \mu_k \qquad (n \geq 1). \tag{5}$$

Euler transform. Applying (4) gives

$$G_\mu(z) = \sum_{k \geq 0} 2^k \left[\frac{z(z/2)\ldots(z/2^{k-1})}{(1+z)(1+z/2)\ldots(1+z/2^{k-1})}\right]^b G_a\left(\frac{z}{2^k}\right),$$

where

$$\begin{aligned} G_a(z) &= \frac{1}{1+z}\sum_{n\geq 0} a_n \left(\frac{z}{1+z}\right)^n \\ &= a_1 \frac{z}{(1+z)^2} + a_2 \frac{z^2}{(1+z)^3} + \cdots + a_b \frac{z^b}{(1+z)^{b+1}} + a_{b+1}\frac{z^{b+1}}{(1+z)^{b+1}}. \end{aligned}$$

Harmonic sum. It is slightly more convenient to work with the function $\bar{G}_\mu(z) = G_\mu(1/z)$. Following [13], we introduce $Q(z) := \prod_{j\geq 0}(1+z/2^j)$. Then we have

$$\bar{G}_\mu(z) = Q^b\left(\frac{z}{2}\right) H_\mu(z),$$

where

$$H_\mu(z) = \sum_{k\geq 0} 2^k h_\mu(2^k z),$$

is a harmonic sum (see [11]) with base function

$$\begin{aligned} h_\mu(z) &= \frac{\bar{G}_a(z)}{Q^b(z)}(1+z)^b \\ &= \frac{1}{Q^b(z)}\left(\sum_{1\leq j\leq b} a_j z(1+z)^{b-j-1} + \frac{a_{b+1}}{1+z}\right). \end{aligned} \tag{6}$$

Mellin transform of H_μ. The Mellin transform of H_μ satisfies

$$H_\mu^*(s) = \frac{h_\mu^*(s)}{1-2^{1-s}}.$$

By (6), $h_\mu(z) = O(1)$ as $z \to 0^+$ and $h_\mu(z) = O(z^{-K})$, for any $K > 0$, as $z \to +\infty$ (see [13]). It follows that the fundamental strip of $H_\mu^*(s)$ is $\langle 1, \infty\rangle$. We need more analytic properties of $H_\mu^*(s)$ in order to apply the inversion formula.

Singular expansion of $H_\mu^*(s)$. By (6),

$$h_\mu(z) = a_{b+1} + q_1 z + O(|z|^2) \qquad (z \to 0),$$

where $q_1 := \sum_{1\leq j\leq b} a_j - (2b+1)a_{b+1}$; thus the Mellin transform of $h_\mu(z)$ can be meromorphically continued to $\Re s > -2$ (see [11]) and has the expansion

$$h_\mu^*(s) \asymp \frac{a_{b+1}}{s} + \frac{q_1}{s+1}.$$

Consequently, we obtain the singular expansion of $H_\mu^*(s)$

$$H_\mu^*(s) \asymp \frac{h_\mu^*(1)}{L(s-1)} + \sum_{k\neq 0} \frac{h_\mu^*(1+\chi_k)}{L(s-1-\chi_k)} - \frac{a_{b+1}}{s} - \frac{q_1}{3(s+1)}, \tag{7}$$

for $\Re s > -2$, where

$$\chi_k = \frac{2k\pi i}{L}, \quad L := \log 2.$$

Growth order of $H^*_\mu(s)$ at $\pm i\infty$. We use the representation

$$
\begin{aligned}
h^*_\mu(s) &= \int_0^\infty \frac{\sum_{2\le j\le b-1} a_j z(1+z)^{b-j-1}}{Q^b(z)} z^{s-1}dz + \int_0^\infty \frac{a_b z^s + a_{b+1} z^{s-1}}{(1+z)Q^b(z)}\,dz \\
&= \sum_{2\le j\le b-1} a_j \sum_{0\le k\le b-j-1} \binom{b-j-1}{k} I^*(k+s-1) + \int_0^\infty \frac{a_b z^s + a_{b+1} z^{s-1}}{(1+z)Q^b(z)}\,dz,
\end{aligned}
$$

valid for $\Re s > 0$, where $I^*(s)$ denotes the Mellin transform of $I(z) := 1/Q^b(z)$. Analytic properties of $I^*(s)$ were studied in detail in [13], in [18, 19], and in [32]. In particular, we have, arguing similarly as in [13],

$$
h^*_\mu(\sigma + it) = O\left(e^{-(\pi-\varepsilon)|t|}\right) \qquad (0 < \varepsilon < \pi/2, \sigma > -2),
$$

as $|t| \to \infty$. This is sufficient for our purposes.

Asymptotics of $\bar{G}_\mu(z)$. By the Mellin inversion formula (see [11]), we deduce that

$$
H_\mu(z) = \frac{1}{L} h^*_\mu(1) z^{-1} + \sum_{k\neq 0} \frac{1}{L} h^*_\mu(1+\chi_k) z^{-1-\chi_k} - a_{b+1} + O(z), \tag{8}
$$

where the O-term is uniform as $z \to 0$ in $|\arg z| \le \pi/2 + \varepsilon$, $0 < \varepsilon < \pi/2$.

Since $Q^b(z/2) = 1 + bz + O(z^2)$, as $z \to 0$, we have

$$
\bar{G}_\mu(z) = \rho_1 z^{-1} + \sum_{k\neq 0} \rho_{1,k} z^{-1-\chi_k} + \rho_0 + \sum_{k\neq 0} \rho_{0,k} z^{-\chi_k} + O(z), \tag{9}
$$

where

$$
\begin{aligned}
\rho_1 &= \frac{1}{L} h^*_\mu(1), & \rho_{1,k} &= \frac{1}{L} h^*_\mu(1+\chi_k), \\
\rho_0 &= \frac{b}{L} h^*_\mu(1) - a_{b+1}, & \rho_{0,k} &= \frac{b}{L} h^*_\mu(1+\chi_k).
\end{aligned} \tag{10}
$$

Singularity analysis. By (2) and (9)

$$
\begin{aligned}
F_\mu(z) &= \rho_1 \frac{z}{(1-z)^2} + \sum_{k\neq 0} \rho_{1,k} \frac{z^{1+\chi_k}}{(1-z)^{2+\chi_k}} \\
&\quad + \rho_0 \frac{1}{1-z} + \sum_{k\neq 0} \rho_{0,k} \frac{z^{\chi_k}}{(1-z)^{1+\chi_k}} + O(z^{-1}),
\end{aligned}
$$

as $z \to 1$ in $\mathcal{D}$.

By singularity analysis (see [12]), we thus obtain the following result:

$$
\mu_n = (\rho_1 + \delta_1(\log_2 n))\, n + (\rho_0 + \delta_0(\log_2 n)) + O\left(\frac{1}{n}\right), \tag{11}
$$

where δ_1 and δ_0 are 1-periodic, infinitely differentiable, mean-zero functions with Fourier expansions

$$
\delta_1(x) = \sum_{k\neq 0} \frac{\rho_{1,k}}{\Gamma(2+\chi_k)} e^{2k\pi i x}, \qquad \delta_0(x) = \sum_{k\neq 0} \left(b - \frac{\chi_k}{2}\right) \frac{\rho_{1,k}}{\Gamma(1+\chi_k)} e^{2k\pi i x}.
$$

Theorem 1. *The mean of Z_n satisfies*

$$\mathbf{E}[Z_n] = P(\log_2 n)n + Q(\log_2 n) + O(n^{-1}),$$

where $P(x)$ and $Q(x)$ are 1-periodic, infinitely differentiable functions whose Fourier expansions are given by

$$\begin{aligned} P(x) &= \frac{1}{L}h_\mu^*(1) + \sum_{k\neq 0} \frac{h_\mu^*(1+\chi_k)}{L\Gamma(2+\chi_k)} e^{2k\pi i x}, \\ Q(x) &= \frac{b}{L}h_\mu^*(1) - a_{b+1} + \sum_{k\neq 0} \left\{ \frac{b}{L} - \frac{\chi_k}{2L} \right\} \frac{h_\mu^*(1+\chi_k)}{\Gamma(1+\chi_k)} e^{2k\pi i x}. \end{aligned}$$

Note that by definition the constant $h_\mu^*(1)$ is

$$h_\mu^*(1) = \int_0^\infty Q^{-b}(x) \left(\sum_{2\leq j\leq b} a_j x(1+x)^{b-j-1} + \frac{a_{b+1}}{1+x} \right) dx.$$

5 Variance of Z_n

Recurrence. Let $\tau_n := M_n''(0) = \mathbf{E}[Z_n^2]$. Then $\mathbf{Var}[Z_n] = \tau_n - \mu_n^2$. By (1), τ_n satisfies $\tau_0 = 0$, $\tau_j = a_j^2$ for $1 \leq j \leq b$ and

$$\tau_{n+b} = a_{b+1} 2^{2-n} \sum_{0\leq k\leq n} \binom{n}{k} \mu_k + a_{b+1}^2 + 2\nu_n + 2^{1-n} \sum_{0\leq k\leq n} \binom{n}{k} \tau_k \quad (n \geq 1),$$

where

$$\nu_n := 2^{-n} \sum_{0\leq k\leq n} \binom{n}{k} \mu_k \mu_{n-k} \quad (n \geq 0)$$

is the *binomial convolution* of μ_n.

Euler transform. Applying (4) gives

$$G_\tau(z) = \sum_{k\geq 0} 2^k \left[\frac{z(z/2)\dots(z/2^{k-1})}{(1+z)(1+z/2)\dots(1+z/2^{k-1})} \right]^b G_\lambda\left(\frac{z}{2^k}\right),$$

with

$$G_\lambda(z) = 4a_{b+1} \frac{z^b}{(1+z)^b} G_\mu\left(\frac{z}{2}\right) + G_{a^2}(z) + 2\frac{z^b}{(1+z)^b} G_\nu(z).$$

Harmonic sum. As for the mean, we introduce the more convenient $\bar{G}_\tau(z) = G_\tau(1/z)$ and get

$$\bar{G}_\tau(z) = Q^b\left(\frac{z}{2}\right) H_\tau(z),$$

where

$$H_\tau(z) = \sum_{j\geq 0} 2^j h_\tau(2^j z)$$

is the harmonic sum with base function

$$h_\tau(z) = 4a_{b+1}H_\mu(2z) + H_{a^2}(z) + 2N(z).$$

Here H_{a^2} is the function obtained from h_μ by substituting $a_i \mapsto a_i^2$, and $N(z) := \bar{G}_\nu(z)I(z)$.

Mellin transform. The Mellin transform of $H_\tau(z)$ is

$$H_\tau^*(s) = \frac{h_\tau^*(s)}{1 - 2^{1-s}},$$

where

$$h_\tau^*(s) = a_{b+1}2^{2-s}H_\mu^*(s) + H_{a^2}^*(s) + 2N^*(s). \tag{12}$$

The fundamental strip of $\bar{G}_\mu^*$ is $\langle 1, \infty\rangle$. The fundamental strip of $H_{a^2}^*$ is easily seen to be contained in $\langle 1, \infty\rangle$. It remains to determine the fundamental strip of N^*. From $\nu_0 = \nu_1 = 0$ we deduce that

$$\bar{G}_\nu(x) = x\sum_{n\geq 2} \nu_n(1+x)^{-n-1} = O(x^{-2}) \qquad (x \to \infty);$$

on the other hand, by Theorem 1, $\nu_n = O(n^2)$, so that

$$\bar{G}_\nu(x) = O\left(x\sum_n n^2(1+x)^{-n-1}\right) = O(x^{-2}) \qquad (x \to 0).$$

Since $I(z) = O(1)$ as $z \to 0$ and $I(z) = O(z^{-K})$ for any $K > 0$ as $z \to \infty$, it follows that the fundamental strip of N^*, and thus that of h_τ^*, is $\langle 2, \infty\rangle$.

Mellin transform of $\bar{G}_\mu$. To derive more analytical properties of H_τ^*, we use the Mellin transform $\bar{G}_\mu^*$. Here we encounter a minor technical difficulty: If $a_1 \neq 0$, then, by Theorem 1, $\mu_n = \Theta(n)$, and thus $\bar{G}_\mu^*$ has an empty fundamental strip. In other words, the Mellin transform $\bar{G}_\mu^*$ would not exist. Fortunately we can use the property that the total number of records in various nodes in the bucket digital tree is identically n. The corresponding equation

$$X_n^{(1)} + 2X_n^{(2)} + \cdots + bX_n^{(b)} + bX_n^{(b+1)} = n$$

shows that $X_n^{(1)}$ can be obtained as an affine expression in $X_n^{(2)}, \ldots, X_n^{(b+1)}$. Therefore we assume for the following technical analysis $a_1 = 0$, implying that the fundamental strip of $\bar{G}_\mu^*$ is contained in $\langle 1, 2\rangle$. Moreover, using the asymptotic expansion (9), it can be continued to $\langle -1, 2\rangle$ with singular expansion

$$\bar{G}_\mu^*(s) \asymp \frac{\rho_1}{s-1} + \sum_{k\neq 0} \frac{\rho_{1,k}}{s-1-\chi_k} + \frac{\rho_0}{s} + \sum_{k\neq 0} \frac{\rho_{0,k}}{s-\chi_k}. \tag{13}$$

Smallness of $\bar{G}_\mu^*$ at $\pm i\infty$. A quantitative version of the *Mellin reciprocity* says that if a function is analytic in a cone containing the positive real axis and has power growth in this cone, then its Mellin transform is analytic in a strip and has exponential decay in this strip and vice versa. For a more precise statement that goes back to old work by Mellin himself (see [9], in particular, Ch. 11, §2 for the basic reciprocity, §3 for convolutions, and [10], Ch. 5, §3 for extension to the meromorphic continuations). Actually we need an extension of those results to the case when there are infinitely many poles on a *vertical* line.

We have shown that $H^*(s)$ decays exponentially in a strip when the imaginary part goes to infinity, therefore $H(z)$ satisfies the estimate (8) in a cone. A similar power estimate holds by (9) for the analytic function $\bar{G}_\mu(z)$ in this cone. Thus $\bar{G}_\mu^*$ also satisfies $\bar{G}_\mu^*(x+iy) = O(e^{-(\pi-\epsilon)|y|})$ as $|y| \to \infty$, where $\epsilon \in (0,\pi)$. A different argument that gives the sharper estimate $\bar{G}_\mu^*(x+iy) = O(|y|^b e^{-\pi|y|})$ is to use the series representation for $\bar{G}_\mu^*$ given below in (22).

Mellin transform of $\bar{G}_\nu$. The next step is the computation of the Mellin transform $\bar{G}_\nu^*$. We use the integral formula (see [18])

$$\bar{G}_\nu^*(s) = \frac{1}{2\pi i} \int_{\varsigma - i\infty}^{\varsigma + i\infty} 2^{-s} \binom{s}{t} \bar{G}_\mu^*(t) \bar{G}_\mu^*(s-t)\,dt \qquad (s \in \langle 2,4 \rangle), \tag{14}$$

with appropriate $\varsigma \in (1,2)$. Shifting the line of integration to the left, we deduce that

$$\bar{G}_\nu^*(s) = \rho_1 s 2^{-s} \bar{G}_\mu^*(s-1) + \Sigma_2(s) + \rho_0 2^{-s} \bar{G}_\mu^*(s) + \Sigma_1(s) + \bar{G}_{\nu,1}^*(s), \tag{15}$$

for $s \in \langle 0,2 \rangle$, where

$$\begin{aligned} \Sigma_2(s) &= 2^{-s} \sum_{k \neq 0} \binom{s}{1+\chi_k} \rho_{1,k} \bar{G}_\mu^*(s-1-\chi_k), \\ \Sigma_1(s) &= 2^{-s} \sum_{k \neq 0} \binom{s}{\chi_k} \rho_{0,k} \bar{G}_\mu^*(s-\chi_k), \end{aligned}$$

and

$$\bar{G}_{\nu,1}^*(s) = \frac{1}{2\pi i} \int_{\varsigma_1 - i\infty}^{\varsigma_1 + i\infty} 2^{-s} \binom{s}{t} \bar{G}_\mu^*(t) \bar{G}_\mu^*(s-t)\,dt, \tag{16}$$

with appropriate $\varsigma_1 \in (-1,0)$.

By using estimates for the function $\bar{G}_\mu^*$ and the coefficients $\rho_{1,k}$ and $\rho_{0,k}$, we can show that the series for Σ_1 and Σ_2 converge uniformly on compact sets avoiding the poles inherited from $\bar{G}_\mu^*$. By (13), we deduce the singular expansion

$$\bar{G}_\nu^*(s) \asymp \frac{V_2}{s-2} + \sum_{k \neq 0} \frac{V_{2,k}}{s-2-\chi_k} + \frac{V_1}{s-1} + \sum_{k \neq 0} \frac{V_{1,k}}{s-1-\chi_k}, \tag{17}$$

for $s \in \langle 0, 4 \rangle$, where

$$V_2 = \frac{1}{2}\rho_1^2 + \frac{1}{4}\sum_{k\neq 0}\binom{2}{1+\chi_k}\rho_{1,k}\rho_{1,-k},$$

$$V_{2,k} = \left(1+\frac{\chi_k}{2}\right)\rho_1\rho_{1,k} + \frac{1}{4}\sum_{\ell\neq 0,k}\binom{2+\chi_k}{1+\chi_\ell}\rho_{1,\ell}\rho_{1,k-\ell},$$

$$V_1 = \rho_1\rho_0 + \frac{1}{2}\sum_{k\neq 0}\binom{1}{1+\chi_k}\rho_{1,k}\rho_{0,-k} + \frac{1}{2}\sum_{k\neq 0}\binom{1}{\chi_k}\rho_{0,k}\rho_{1,-k},$$

$$V_{1,k} = (1+\chi_k)\rho_1\rho_{0,k} + \rho_0\rho_{1,k} + \frac{1}{2}\sum_{\ell\neq 0,k}\binom{1+\chi_k}{1+\chi_\ell}\rho_{1,\ell}\rho_{0,k-\ell} + \frac{1}{2}\sum_{\ell\neq 0,k}\binom{1+\chi_k}{\chi_\ell}\rho_{0,\ell}\rho_{1,k-\ell}.$$

This can be simplified to

$$V_2 = \frac{1}{2}\rho_1^2 + \frac{1}{2}\sum_{k\neq 0}\frac{\rho_{1,k}\rho_{1,-k}}{\Gamma(2+\chi_k)\Gamma(2-\chi_k)},$$

$$V_{2,k} = \left(1+\frac{\chi_k}{2}\right)\rho_1\rho_{1,k} + \frac{1}{4}\Gamma(3+\chi_k)\sum_{\ell\neq 0,k}\frac{\rho_{1,\ell}\rho_{1,k-\ell}}{\Gamma(2+\chi_\ell)\Gamma(2-\chi_{k-\ell})},$$

$$V_1 = b\rho_1^2 - \rho_1 a_{b+1} + b\sum_{k\neq 0}\frac{\rho_{1,k}\rho_{1,-k}}{\Gamma(2+\chi_k)\Gamma(2-\chi_k)},$$

$$V_{1,k} = (2+\chi_k)b\rho_1\rho_{1,k} - a_{b+1}\rho_{1,k} + \frac{b}{2}\Gamma(3+\chi_k)\sum_{\ell\neq 0,k}\frac{\rho_{1,\ell}\rho_{1,k-\ell}}{\Gamma(2+\chi_\ell)\Gamma(2-\chi_{k-\ell})}.$$

Smallness of $\bar{G}_\nu^*$ at $\pm i\infty$. Inserting the estimates for $\bar{G}_\mu(s)$ and well-known estimates for the gamma function in (14) and (16), respectively, we obtain $\bar{G}_\nu^*(s) = O(e^{-(\pi-\epsilon)|y|})$ in vertical strips of finite width for arbitrary $\epsilon \in (0, \pi)$.

Mellin transform of N. Finally we use the convolution integral

$$N^*(s) = \frac{1}{2\pi i}\int_{\varsigma - i\infty}^{\varsigma + i\infty} I^*(t)\bar{G}_\nu^*(s-t)dt$$

for the Mellin transform of a product.

If the line of integration $\Re t = \varsigma$ is chosen in the fundamental strip of I^*, then the integral defines an analytic function for $s-t$ in the fundamental strip of $\bar{G}_\nu^*$, which is $\langle 2, 4\rangle$.

Thinking of s as fixed temporarily, we shift the line of integration to the left, use the singular expansion

$$I^*(s) \asymp \frac{1}{s} - \frac{2b}{s+1} \quad s \in \langle -2, \infty\rangle,$$

and obtain

$$N^*(s) = \bar{G}_\nu^*(s) - 2b\bar{G}_\nu^*(s+1) + N_1(s), \tag{18}$$

where

$$N_1^*(s) = \frac{1}{2\pi i} \int_{\varsigma_1 - i\infty}^{\varsigma_1 + i\infty} I^*(t) \bar{G}_\nu^*(s-t) dt \tag{19}$$

with appropriate $\varsigma_1 \in (-2, -1)$. Combining the singular expansions yields

$$N^*(s) \asymp \frac{V_2}{s-2} + \sum_{k \neq 0} \frac{V_{2,k}}{s-2-\chi_k} - \frac{a_{b+1}\rho_1}{s-1} - \sum_{k \neq 0} \frac{a_{b+1}\rho_{1,k}}{s-1-\chi_k} \tag{20}$$

for $s \in \langle 0, 4 \rangle$.

Smallness of N^* at $\pm i\infty$. Again we can use the quantitative version of the Mellin reciprocity. Since the function $\bar{G}_\nu^*$ has an exponential decay in a strip, $\bar{G}_\nu$ has a power growth in a cone. Multiplying by I gives an analytic function that again has at most a power growth in this cone, thus a Mellin transform with exponential decay $O(e^{-(\pi-\epsilon)|y|})$ in a strip.

Singular expansion of H_τ^*. The singular expansions of H_μ^* and N^* reveal that h_τ^* has an analytic continuation to $\langle 0, \infty \rangle$, except for simple poles at $s = 2$ and $s = 2 + \chi_k$. The singularities at $s = 1$ and $s = 1 + \chi_k$ cancel out.

Consequently H_τ^* has the fundamental strip $\langle 2, \infty \rangle$, with a meromorphic continuation to $\langle 0, \infty \rangle$, and the singular expansion

$$H_\tau^*(s) \asymp \frac{4V_2}{s-2} + \sum_{k \neq 0} \frac{4V_{2,k}}{s-2-\chi_k} + \frac{h_\tau^*(1)}{L(s-1)} + \sum_{k \neq 0} \frac{h_\tau^*(1+\chi_k)}{L(s-1-\chi_k)},$$

for $s \in \langle 0, \infty \rangle$.

Asymptotic expansion of $F_\tau(z)$. From the singular expansion of $H_\tau^*(s)$, we obtain by the Mellin inversion formula the asymptotic expansion of $H_\tau(z)$ as $z \to 0$ in a cone. Then we deduce, by elementary substitutions and expansions, that

$$F_\tau(z) = \frac{4V_2}{(1-z)^3} + \sum_{k \neq 0} \frac{4V_{2,k}}{(1-z)^{3+\chi_k}} + \frac{\theta_1}{(1-z)^2} + \sum_{k \neq 0} \frac{\theta_{1,k}}{(1-z)^{2+\chi_k}} + O\left(\frac{1}{(1-z)^{1+\epsilon}}\right),$$

as $z \to 1$ in the indented disk $\mathcal{D}$. Here $\theta_1 = 4(b-2)V_2 + h_\tau^*(1)/L$ and $\theta_{1,k} = 4(b-2)V_{2k} + h_\tau^*(1+\chi_k)/L$. By singularity analysis, we finally obtain

$$\begin{aligned} \tau_n \quad = \quad & 2V_2 n^2 + \sum_{k \neq 0} \frac{4V_{2,k}}{\Gamma(3+\chi_k)} n^{2+\chi_k} + (6V_2 + \theta_1)\, n \\ & + \sum_{k \neq 0} \left((3+\chi_k) \frac{2V_{2,k}}{\Gamma(2+\chi_k)} + \frac{\theta_{1,k}}{\Gamma(2+\chi_k)} \right) n^{1+\chi_k} + O(n^\epsilon). \end{aligned}$$

Indeed, more calculations show that the error term can be refined to $O(1)$.

Collecting Contributions. Finally, we compute the difference $\mathbf{Var}[Z_n] = \tau_n - \mu_n^2$. By the relations

$$2V_2 = \rho_1^2 + \sum_{k\neq 0} \frac{\rho_{1,k}\rho_{1,-k}}{\Gamma(2+\chi_k)\Gamma(2-\chi_k)},$$

and

$$\frac{4V_{2,k}}{\Gamma(3+\chi_k)} = 2\rho_1 \frac{\rho_{1,k}}{\Gamma(2+\chi_k)} + \sum_{\ell\neq 0,k} \frac{\rho_{1,\ell}\rho_{1,k-\ell}}{\Gamma(2+\chi_\ell)\Gamma(2-\chi_{k-\ell})}$$

we observe that the terms of order n^2 cancel. This concludes the proof for the asymptotic expansion of the variance.

Theorem 2. *The variance of Z_n satisfies*

$$\mathbf{Var}[Z_n] = \varpi(\log_2 n)n + O(n^{\varepsilon}),$$

where $\varpi(x)$ is a 1-periodic, infinitely differentiable function with Fourier expansion

$$\varpi(x) = \varpi_0 + \sum_{k\neq 0} \varpi_k e^{2k\pi i x}$$

given by the coefficients

$$\varpi_0 = (4b-2)V_2 + \frac{1}{L}h_\tau^*(1) - 2[PQ]_0$$

and

$$\varpi_k = (3+\chi_k)\frac{2V_{2,k}}{\Gamma(2+\chi_k)} + \frac{4(b-2)V_{2k} + h_\tau^*(1+\chi_k)/L}{\Gamma(2+\chi_k)} - 2[PQ]_k.$$

Remarks on numerical evaluation of the coefficients. All functions like h_μ^* and h_τ^* and their analytic continuations can be analytically constructed. In theory, we can use standard procedures to compute the coefficients ϖ_k. But a naive application of the integral formulas used in the proof would lead to a 5-dimensional complex integral for $h_\tau(1)$, where the highly oscillating integrand is given in terms of several infinite series and products. This is computationally less tractable. We now sketch a better approach that has been employed to obtain the numerical illustrations for the variance-covariance matrix given in Section 7 below. Details can be found in [20].

Flajolet and Richmond [13] show that

$$I^*(s) = \frac{\pi}{\sin \pi s} \sum_{0\le r\le b-1} \binom{s-1}{r} \left(\sum_{j\ge 0} \iota_j(r) 2^{js} \right), \tag{21}$$

where the infinite series represents an entire function. By a similar argument, we obtain

$$\bar{G}_\mu^*(s) = \frac{\pi}{\sin \pi s} \sum_{0\le r\le b} \binom{s-1}{r} \left(\sum_{j\ge 0} \epsilon_j(r) 2^{j(1-s)} \right), \tag{22}$$

where the series converge for $\Re(s) > 1$. Both formulas can be evaluated fairly fast and accurately, though care is required for the first few terms of the series. Then we require the analytic continuation of $\bar{G}_\mu^*$ to $\Re s > -1$. This can be done either by deriving a *functional equation* for $\bar{G}_\mu^*$ or by asymptotic expansion of $\epsilon_j(r)$ as $j \to \infty$. These representations can be used for numerical quadrature. Convolution integral evaluations for (16) are done by using QUADPACK (see [38]); on the other hand, we use CUBPACK (see [7]) to compute the double convolution integrals for $N_1(1)$ and $N_1(1+\chi_k)$ required to compute $h_\tau(1)$ and $h_\tau(1+\chi_k)$ from (19). An alternative approach is to use series representation by lengthy residue calculations and expansions that may be more suitable for high precision evaluations. Since the result is a triple infinite series with each term consisting of about b^3 components, we refrain from pursuing this approach further. The interested reader should consult [29], especially equation (2.16), which corresponds to $b = 1$ to get a flavor of such results.

Apart from practical considerations, there seems to be intricate cancelations involved: Formulas (21) and (22) involve polynomials of degree $b-1$ and b, respectively, and the coefficients $\iota_j(r)$ and $\epsilon_j(r)$ grow also considerably in b. On the other hand the result seems to decrease with b; see also [27], where the corresponding result is proved for digital bucket *tries* ($b = 0$).

6 Asymptotic normality of Z_n

We prove the asymptotic normality of Z_n in this section using the method of moments (see [4]) and asymptotic transfers.

6.1 Shifting the mean

Define $\zeta_n = \mathbf{E}[Z_n]$. Let

$$f_n(y) := \mathbf{E}(e^{(Z_n-\zeta_n)y}) = e^{-\zeta_n y} M_n(y).$$

Then $f_n(y)$ satisfies $f_0(y) = \cdots = f_b(y) = 1$, and

$$f_{n+b}(y) = 2^{-n} \sum_{0\le k\le n} \binom{n}{k} f_k(y) f_{n-k}(y) e^{\Delta_{n,k} y} \qquad (n \ge 1),$$

where

$$\Delta_{n,k} := \zeta_k + \zeta_{n-k} - \zeta_n + a_{b+1}.$$

Write $f_{n,m} := f_n^{(m)}(0) = \mathbf{E}\big[(X_n - \zeta_n)^m\big]$. Then $f_{n,1} = 0$, and for $m \ge 2$:

$$f_{n+b,m} = 2^{1-n} \sum_{0\le k\le n} \binom{n}{k} f_{k,m} + R_{n,m},$$

where

$$R_{n,m} = \sum_{\substack{p+q+r=m\\ p,q<m}} \binom{m}{p,q,r} 2^{-n} \sum_{0\le k\le n} \binom{n}{k} f_{k,p} f_{n-k,q} \Delta_{n,k}^r.$$

In particular,

$$R_{n,2} = 2^{-n} \sum_{0\le k\le n} \binom{n}{k} \Delta_{n,k}^2.$$

We need the expansion (see Theorem 1)

$$\zeta_n = P(\log_2 n)n + Q(\log_2 n) + O(n^{-1}), \tag{23}$$

where $P(u)$ and $Q(u)$ are periodic functions that are infinitely differentiable. The two periodic functions are related by the formula

$$Q(u) = -a_{b+1} + b\left(P(u) + \frac{P'(u)}{L}\right) - \frac{1}{2}\left(\frac{P'(u)}{L} + \frac{P''(u)}{L^2}\right), \tag{24}$$

which can be proved using the explicit expressions of the Fourier expansions of the two functions given in Theorem 1.

From (23) and (24), we deduce the following approximation.

Lemma 3. *For $k = n/2 + x\sqrt{n}/2$,*

$$\Delta_{n,k} = \frac{1}{2}(x^2 - 1)\left(\frac{P'(\log_2 n)}{L} + \frac{P''(\log_2 n)}{L^2}\right) + O\left((1 + x^4)n^{-1}\right), \tag{25}$$

uniformly for $x = o(n^{1/4})$.

6.2 Estimates of the binomial distribution

We need the local limit theorem for the binomial distribution, as well as estimates of large-deviations type.

Lemma 4 (Local limit theorem). *For $k = pn + x\sqrt{p(1-p)n}$,*

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{e^{-x^2/2}}{\sqrt{2\pi p(1-p)n}}\left(1 + O\left((1 + |x|^3)n^{-1/2}\right)\right), \tag{26}$$

uniformly for $x = o\left((p(1-p)n)^{1/6}\right)$.

This result holds by Stirling's approximation so long as $p(1-p)n \to \infty$ (and p may depend on n).

Lemma 5 (Tail estimates). *If $k < p(n+1)$, then*

$$\sum_{0 \le j \le k} \binom{n}{j} p^j (1-p)^{n-j} \le \frac{1}{1 - r_-}\binom{n}{k} p^k (1-p)^{n-k}, \tag{27}$$

where

$$r_- := \frac{(1-p)m}{p(n-m+1)};$$

on the other hand, if $m > p(n+1)-1$, *then*

$$\sum_{k\le j\le n} \binom{n}{j} p^j (1-p)^{n-j} \le \frac{1}{1-r_+}\binom{n}{k} p^k (1-p)^{n-k}, \tag{28}$$

where

$$r_+ := \frac{p(n-m)}{(1-p)(m+1)}.$$

Proof. We have

$$\begin{aligned}
\sum_{0\le j\le k} \binom{n}{j} p^j (1-p)^{n-j} &= \binom{n}{k} p^k (1-p)^{n-k} \sum_{0\le j\le k} \frac{(j+1)\cdots k}{(n-k+1)\cdots(n-j)} \left(\frac{1-p}{p}\right)^{k-j} \\
&\le \binom{n}{k} p^k (1-p)^{n-k} \sum_{0\le j\le k} r_-^{k-j}.
\end{aligned}$$

When $k < p(n+1)$, $r_- < 1$; thus the partial sum is bounded above by extending the upper limit to infinity. This proves (27). The proof of the estimate (28) for the right tail is similar and is omitted. $\square$

For more precise estimates, see [21].

Variance. We prove that $\mathbf{Var}[Z_n]/n$ is bounded above and below by suitable positive constants.

Lemma 6. *There exist constants* $0 < c_1 < c_2 < \infty$ *such that*

$$c_1 n \le \mathbf{Var}[Z_n] \le c_2 n \qquad (n \ge 1). \tag{29}$$

Proof. Although the upper bound follows from Theorem 2, we give a self-contained proof. Take $x = n^{1/7}$ and define $k_\pm = n/2 \pm \sqrt{n}x/2$. Then, by (25), (26), (27) and (28),

$$\begin{aligned}
R_{n,2} &= \left(\sum_{k\le k_-} + \sum_{k_-<k<k_+} + \sum_{k_+\le k\le n}\right) 2^{-n}\binom{n}{k}\Delta_{n,k}^2 \\
&= O\left(n^{1+5/14} e^{-n^{2/7}/2} + \int_{-n^{1/7}}^{n^{1/7}} |x^2-1| e^{-x^2/2} dx\right) \\
&= O(1).
\end{aligned}$$

Thus $R_{n,2} \le c_3$, for some $c_3 > 0$. It follows, either by the bound $f_{n,2} = O(\mathbf{E}(Z_n))$ or by induction, that $f_{n,2} = O(n)$. This proves the upper bound of (29).

For the lower bound, since the a_j's are not all zeros, there exists a constant $c_4 > 0$ such that

$$R_{n,2} \ge c_4 \qquad (n \ge n_0 \ge b+1).$$

Thus $\mathbf{Var}[Z_n] = f_{n,2} \ge v_n$, where $v_n := 0$ for $n < n_0$ and

$$v_{n+b} = c_4 + 2^{1-n} \sum_{0\le k\le n} \binom{n}{k} v_k \qquad (n \ge n_0).$$

Take

$$c_1 := \min\left\{\frac{c_4}{n_0}, \frac{c_4}{b+c_5}\right\},$$

where

$$c_5 := \max_{n\ge 1}\left\{2^{1-n}\sum_{0\le k<n_0}\binom{n}{k}k\right\} < \infty.$$

We prove by induction that $v_n \ge c_1 n$ for $n \ge n_0$. The case $n = n_0$ follows from our choice of c_1. Now, by induction,

$$\begin{aligned} v_{n+b} &\ge c_4 + 2^{1-n}c_1\left(\sum_{0\le k\le n}\binom{n}{k}k - \sum_{0\le k<n_0}\binom{n}{k}k\right) \\ &= c_4 + c_1 n - 2^{1-n}c_1\sum_{0\le k<n_0}\binom{n}{k}k \\ &\ge c_4 + c_1 n - c_1 c_5 \\ &\ge c_1(n+b). \end{aligned}$$

This completes the induction. □

Higher moments. We prove that for $m \ge 1$

$$f_{n,2m} \sim \nu_m \varpi(\log_2 n)^m n^m, \tag{30}$$

and

$$f_{n,2m-1} = o(n^{m-1/2}), \tag{31}$$

where $\varpi(u)$ is a 1-periodic function (see Theorem 2) and ν_m will be determined. The case $m = 1$ has already been proved in the paper, we consider $m \ge 2$.

Observe first that, by (25), (26) and induction, the terms with $r \ge 1$ in $R_{n,m}$ are bounded above by

$$\sum_{\substack{p+q+r=m\\ 0\le p,q<m\\ 1\le r\le m}}\binom{m}{p,q,r}n^{p/2+q/2}2^{-n}\sum_{0\le k\le n}\binom{n}{k}|\Delta_{n,k}|^r = o(n^{m/2}), \tag{32}$$

where the inner sum is treated by the same argument for $R_{n,2}$.

Odd moments. Consider now $R_{n,2m-1}$, which is estimated as follows. By (32)

$$R_{n,2m-1} = \sum_{1\le p\le 2m-2}\binom{2m-1}{p}2^{-n}\sum_{0\le k\le n}\binom{n}{k}f_{k,p}f_{n-k,2m-1-p} + o(n^{m-1/2}).$$

Observe that one of p and $2m-1-p$ is odd. Thus by induction

$$R_{n,2m-1} = o\left(\sum_{1\le p\le 2m-2}\binom{2m-1}{p}n^{m-1/2}\right) = o(n^{m-1/2}).$$

We need an asymptotic transfer for the recurrence (3) when $\eta_n = o(n^\alpha)$.

Proposition 1 (Asymptotic transfer). *Let $b \geq 1$ and $\alpha > 1$. Assume ξ_n satisfies (3). (i) If $\eta_n = O(n^\alpha)$, then $\xi_n = O(n^\alpha)$; (ii) if $\eta_n = o(n^\alpha)$, then $\xi_n = o(n^\alpha)$.*

The same results also hold for $b = 0$ by (36) below.

We will derive a more precise asymptotic transfer needed for even moments.

Lemma 7. *Assume ξ_n satisfies (3) with $\xi_0 = \eta_0 = 0$. Define $\xi(z) := \sum_{n\geq 1} \xi_n z^n$ and $\eta(z) := \sum_{n\geq 1} \eta_n z^n$. Then, for $b \geq 1$,*

$$\xi(z) = \sum_{k\geq 0} \frac{4^k}{2^k - (2^k-1)z} \eta\left(\frac{z}{2^k - (2^k-1)z}\right) \prod_{0\leq j<k} \left(\frac{z}{2^j - (2^j-1)z}\right)^b. \tag{33}$$

Proof. By (3),

$$\xi(z) = \eta(z) + \frac{4z^b}{2-z} \xi\left(\frac{z}{2-z}\right).$$

Iterating this equation yields (33). □

Corollary 1. *Assume ξ_n satisfies (3) with $\xi_0 = \eta_0 = 0$. Then*

$$\xi_n = [z^n] \sum_{k\geq 0} \frac{2^k z^{kb}}{1-(1-2^{-k})z} \eta\left(\frac{z/2^k}{1-(1-2^{-k})z}\right) \prod_{1\leq j<k} \left(2^j - (2^j-1)z\right)^{-b},$$

where $[z^n]F(z)$ denotes the coefficient of z^n in the Taylor expansion of $F(z)$.

Corollary 2. *If $\eta(z) = c(1-z)^{-\alpha-1} - c$, where $\alpha > 1$, then $\xi_n = O(n^\alpha)$.*

Proof. By the above corollary,

$$\xi_n = [z^n] c(1-z)^{-\alpha-1} Y(z) + O(1),$$

where

$$Y(z) := \sum_{k\geq 0} 2^k z^{bk} \left(1-(1-2^{-k})z\right)^\alpha \prod_{1\leq j<k} \left(2^j - (2^j-1)z\right)^{-b}.$$

Note that Y is analytic for $|z| \leq 1$ and that

$$Y(1) = \frac{1}{1-2^{1-\alpha}}.$$

Thus

$$\begin{aligned} \xi_n &= c \sum_{0\leq k\leq n} \binom{n-k+\alpha}{n-k} [z^k] Y(z) + O(1) \\ &= O\left(n^\alpha \sum_{k\geq 0} \left|[z^k]Y(z)\right|\right) + O(1) \\ &= O(n^\alpha). \end{aligned}$$

This proves the corollary. $\square$

Indeed, the implied constant in Corollary 2 is given by $c(1-2^{1-\alpha})^{-1}\Gamma(\alpha+1)^{-1}$; this is a special case of (35) below.

Proof of Proposition 1. We may always assume that $\xi_0 = 0$; the case when $\xi_0 > 0$ adds at most $O(n)$ to ξ_n by induction. Since $\eta_n = O(n^\alpha)$, there exists a large enough constant $K > 0$ such that, for all $n \geq 1$,

$$\eta_n \leq K\binom{n+\alpha}{n}.$$

Thus $\xi_n \leq \xi_n^{(1)}$, where $\xi_n^{(1)} := K\binom{n+\alpha}{n}$ for $1 \leq n \leq b$ and

$$\xi_{n+b}^{(1)} = K\binom{n+b+\alpha}{n+b} + 2^{1-n}\sum_{1\leq k\leq n}\binom{n}{k}\xi_k^{(1)} \qquad (n \geq 1).$$

Applying Corollary 2 gives

$$\xi_n \leq \xi_n^{(1)} = O(n^\alpha).$$

This proves (i). Changing K to $\varepsilon > 0$ and arguing similarly, we obtain (ii). $\square$

Since $m - 1/2 > 1$ for $m \geq 2$, we apply the proposition, giving

$$f_{n,2m-1} = o(n^{m-1/2}),$$

for $m \geq 2$.

Even moments. For even moments, we have, again by (32),

$$\begin{aligned} R_{n,2m} &= R_{n,2m}^{(1)} + \sum_{0\leq p<m}\binom{2m}{2p+1}2^{-n}\sum_{0\leq k\leq n}\binom{n}{k}f_{n,2p+1}f_{n-k,2m-2p-1} + o(n^m) \\ &= R_{n,2m}^{(1)} + o(n^m), \end{aligned}$$

where

$$R_{n,2m}^{(1)} := \sum_{1\leq p<m}\binom{2m}{2p}2^{-n}\sum_{0\leq k\leq n}\binom{n}{k}f_{n,2p}f_{n-k,2m-2p}.$$

By (26) and (30), we deduce, by the same argument used for $R_{n,2}$, that

$$R_{n,2m}^{(1)} \sim n^m 2^{-m}\varpi(\log_2 n)^m \sum_{1\leq p<m}\binom{2m}{2p}\nu_p\nu_{m-p}. \tag{34}$$

We need another asymptotic transfer.

Proposition 2 (Asymptotic transfer). *Assume that ξ_n satisfies (3). If $\eta_n \sim n^\alpha\Pi(\log_2 n)$, where $\alpha > 1$ and $\Pi(u) > 0$ is a periodic function of period 1 with continuous first derivative, then*

$$\xi_n \sim \frac{n^\alpha\Pi(\log_2 n)}{1-2^{1-\alpha}}. \tag{35}$$

The intuition is as follows. Since the binomial distribution on n identically distributed independent experiments with rate of success $\frac{1}{2}$ in each is highly concentrated around $n/2$, we have

$$\xi_n \approx 2\xi_{\lfloor n/2 \rfloor} + \eta_n;$$

iterating the formula yields

$$\xi_n \approx \sum_{2^k \le n} 2^k \eta_{\lfloor n/2^k \rfloor} \approx n^\alpha \Pi(\log_2 n) \sum_{2^k \le n} 2^{k(1-\alpha)},$$

and thus (35) follows.

Lemma 8. *Define a sequence λ_n by $\lambda_0 = \lambda_1 = 0$ and*

$$\lambda_n = \eta_n + 2^{1-n} \sum_{0 \le k \le n} \binom{n}{k} \lambda_k \qquad (n \ge 2).$$

Then

$$\lambda_n = \sum_{k \ge 0} 2^k \sum_{2 \le j \le n} \binom{n}{j} 2^{-kj} \left(1 - 2^{-k}\right)^{n-j} \eta_j \qquad (n \ge 2). \tag{36}$$

Proof. This is obtained by considering the Poisson generating functions $\Lambda(z) := e^{-z} \sum_n \lambda_n z^n/n!$ of λ_n, which satisfies

$$\Lambda(z) = 2\Lambda(z/2) + e^{-z} \sum_{n \ge 2} \frac{\eta_n}{n!} z^n,$$

and then by iterating the functional equation. □

Proof of Proposition 2. Since we are mainly interested in fixed b, we use an elementary approach. First the difference $\nabla_n := \xi_{n+1} - \xi_n$ satisfies, by (3),

$$\nabla_{n+b} = \eta_{n+b+1} - \eta_{n+b} + 2^{-n} \sum_{0 \le k \le n} \binom{n}{k} \nabla_k,$$

with suitable initial conditions. Since, by assumption, $\eta_{n+b+1} - \eta_{n+b} = o(n^\alpha)$, we deduce, as in the proof of part (ii) of Proposition 1, that

$$\nabla_n = o(n^\alpha).$$

Thus

$$\xi_{n+b} = \xi_{n+b-1} + \nabla_{n+b-1} = \xi_{n+b-1} + o(n^\alpha) = \xi_n + o(n^\alpha).$$

Then the difference $\varepsilon_n := \xi_n - \lambda_n$ satisfies

$$\begin{aligned} \varepsilon_n &= \xi_{n+b} - \lambda_n + o(n^\alpha) \\ &= \eta_{n+b} - \eta_{n+b-1} + 2^{1-n} \sum_{0 \le k \le n} \binom{n}{k} \varepsilon_k + o(n^\alpha). \end{aligned}$$

Applying Proposition 1 gives $\varepsilon_n = o(n^\alpha)$, and thus

$$\xi_n = \lambda_n + o(n^\alpha).$$

We need only to prove that if η_n satisfies the assumptions of the proposition, then

$$\lambda_n \sim \frac{n^\alpha \Pi(\log_2 n)}{1-2^{1-\alpha}}.$$

Toward that purpose, we split the sum (in k) in (36) into three parts: $n/2^k \geq \log n$, $n/2^k \leq 1$ and $1 \leq n/2^k \leq \log n$.

If $n/2^k \geq \log n$, then, by using the same argument for $R_{n,2}$ together with the assumptions on η_n, we obtain

$$\sum_{2\leq j\leq n} \binom{n}{j} 2^{-kj}\left(1-2^{-k}\right)^{n-j}\eta_j \sim \left(\frac{n}{2^k}\right)^\alpha \Pi\left(\log_2 \frac{n}{2^k}\right) = n^\alpha 2^{-k\alpha}\Pi(\log_2 n).$$

If $n/2^k \leq 1$, then by Stirling formula, we have

$$\begin{aligned}\sum_{2\leq j\leq n} \binom{n}{j} 2^{-kj}\left(1-2^{-k}\right)^{n-j}\eta_j &= O\left(\sum_{j\geq 2}\frac{j^\alpha}{j!}\left(\frac{n}{2^k}\right)^j\right)\\ &= O\left(n^2 4^{-k}\right).\end{aligned}$$

If $1 \leq n/2^k \leq \log n$, then

$$\sum_{2\leq j\leq n} \binom{n}{j} 2^{-kj}\left(1-2^{-k}\right)^{n-j}\eta_j = O\left((1-2^{-k})^n \sum_{j\geq 2}\frac{j^\alpha}{j!}\left(\frac{n}{2^k-1}\right)^j\right).$$

By the saddlepoint method (or the local limit theorem for Poisson distribution), we have

$$\sum_{j\geq 0}\frac{j^\alpha}{j!}x^j = O(x^\alpha e^x) \qquad (x \geq 1).$$

Thus

$$\begin{aligned}\sum_{2\leq j\leq n} \binom{n}{j} 2^{-kj}\left(1-2^{-k}\right)^{n-j}\eta_j &= O\left(n^\alpha (2^k-1)^{-\alpha}\exp\left(-\frac{n}{2^k}+\frac{n}{2^k-1}\right)\right)\\ &= O\left(n^\alpha 2^{-k\alpha}\right).\end{aligned}$$

It follows that

$$\begin{aligned}\lambda_n &= n^\alpha \Pi(\log_2 n)\big(1+o(1)\big)\sum_{n/2^k>\log n} 2^{k(1-\alpha)} + \sum_{1\leq n/2^k\leq \log n} O\left(n^\alpha 2^{k(1-\alpha)}\right) + \sum_{n/2^k<1} O\left(n^2 2^{-k}\right)\\ &= \frac{n^\alpha \Pi(\log_2 n)}{1-2^{1-\alpha}} + O\left(n(\log n)^{\alpha-1}\right) + O(n). \quad \square\end{aligned}$$

By Theorem 2, (29) and (34), we can now apply the asymptotic transfer (35) and obtain

$$f_{n,2m} \sim \frac{n^m \varpi(\log_2 n)^m}{2^m-2}\sum_{1\leq p<m}\binom{2m}{2p}\nu_p \nu_{m-p}.$$

Thus we should take

$$\nu_m = \frac{(2m)!}{2^m m!},$$

so that ν_m satisfies

$$\nu_m = \frac{1}{2^m - 2} \sum_{1 \le p < m} \binom{2m}{2p} \nu_p \nu_{m-p} \qquad (m \ge 2),$$

with $\nu_1 = 1$ and the result is consistent. This proves (30) and the asymptotic normality of Z_n.

Theorem 3. *The distribution of* $(Z_n - \mathbf{E}[Z_n])/\sqrt{n\varpi(\log_2 n)}$ *is asymptotically normal*

$$\frac{Z_n - \mathbf{E}[Z_n]}{\sqrt{n\varpi(\log_2 n)}} \xrightarrow{d} N(0,1),$$

where the convergence holds also for all moments.

Total path length. The same approach applies *mutatis mutandis* to the total path length T_n of a random digital search tree, where the probability generating function $g_n(y)$ satisfies $g_0(y) = \cdots = g_{b-1}(y) = 1$ and

$$g_{n+b}(y) = y^n 2^{-n} \sum_{0 \le k \le n} \binom{n}{k} g_k(y) g_{n-k}(y) \qquad (n \ge 0).$$

We need, in particular, the precise expansion of the mean (see [19]):

$$\mathbf{E}[T_n] = n \log_2 n + P_1(\log_2 n)n + b \log_2 n + Q_1(\log_2 n) + O\left(n^{-1} \log n\right),$$

where

$$Q_1(u) = b\left(P_1(u) + \frac{P_1'(u)}{L}\right) - \frac{1}{2}\left(\frac{P_1'(u)}{L} + \frac{P_1''(u)}{L^2}\right) + 2b + \frac{b}{\log 2} - \frac{1}{2\log 2}.$$

The remaining proof of the central limit theorem is then almost identical to that for Z_n.

Theorem 4. *The distribution of the total path length* T_n *of a random bucket digital search tree satisfies*

$$\frac{T_n - \mathbf{E}[T_n]}{\sqrt{nD(\log_2 n)}} \xrightarrow{d} N(0,1),$$

where $D(u)$ *is a* 1*-periodic function; the convergence holds for all moments.*

Size and path length of tries. Another application of our method of moments is to the size (number of internal nodes) and internal path length of random tries. This corresponds to the case $b = 0$. Our proof requires very little modifications and the limiting distribution for both quantities are again normal. See [24, 25] or [36] for a different approach.

Path length of Patricia tries. Path length of random Patricia tries were studied by [28], where the mean and the variance were derived. The asymptotic normality can be derived by our approach.

7 Illustrating Examples

In this section we discuss the behavior of the random bucket digital tree with bucket capacity 5 under three examples of memory management schemes. The three systems we choose are the fixed-block allocation, heap allocation, and the buddy system. We choose one fixed bucket capacity across the different systems to create a context for meaningful comparisons of the space savings. In practice b may be much larger than 5 and the savings will be more pronounced. We begin with the most detailed model (that is the fixed block model with block capacity 5). Following the paradigm for our multivariate analysis, the other models follow by lumping the nodes of the fixed-block scheme.

According to the theory developed in Section 5, all calculations can be carried out to within $O(1/n)$-accuracy. However, this section is practically-minded. We only carry out the calculations for leading terms, and we also neglect the small periodic fluctuations in the leading terms.

(*a*) *Fixed-Bucket Scheme:* Under several operating systems, like IBM's OS/2 for personal computers, virtual memory is used where external storage devices can be thought of as extensions to the primary computer memory. Communication with the external devices takes place in *fixed* blocks called pages (a typical page size is .5–4K bytes, where 1K= 1024). For these operating systems, one page can be used for hundreds of keys. (Our miniature example with $b = 5$ is not likely to occur in practice; it is only meant for the sake of simple illustration.) This special case corresponds to $p = 1$, since we have only one type of nodes with capacity $\beta_1 = 5$.

Choose the settings of $\mathbf{a}$ as discussed in the paradigm of multivariate analysis. The procedures of Section 5 yield

$$\mathbf{E}[\mathbf{X}_n] \approx \begin{pmatrix} 0.036723 \\ 0.028003 \\ 0.021963 \\ 0.017636 \\ 0.014467 \\ 0.139699 \end{pmatrix} n.$$

and

$$\mathbf{Var}[\mathbf{X}_n] \approx \begin{pmatrix} 0.0295 & -0.0059 & -0.0060 & -0.0055 & -0.0054 & 0.0100 \\ -0.0059 & 0.0228 & -0.0046 & -0.0042 & -0.0038 & 0.0021 \\ -0.0060 & -0.0046 & 0.0181 & -0.0033 & -0.0030 & -0.0021 \\ -0.0055 & -0.0042 & -0.0033 & 0.0147 & -0.0025 & -0.0044 \\ -0.0054 & -0.0038 & -0.0030 & -0.0025 & 0.0122 & -0.0057 \\ 0.0100 & 0.0021 & -0.0021 & -0.0044 & -0.0057 & 0.0077 \end{pmatrix} n. \qquad (37)$$

In our miniature example, each data block contains four pointers to keys in the sequential files and two linkage pointers within the tree, a total of 6 pointers. Therefore, if the tree is implemented

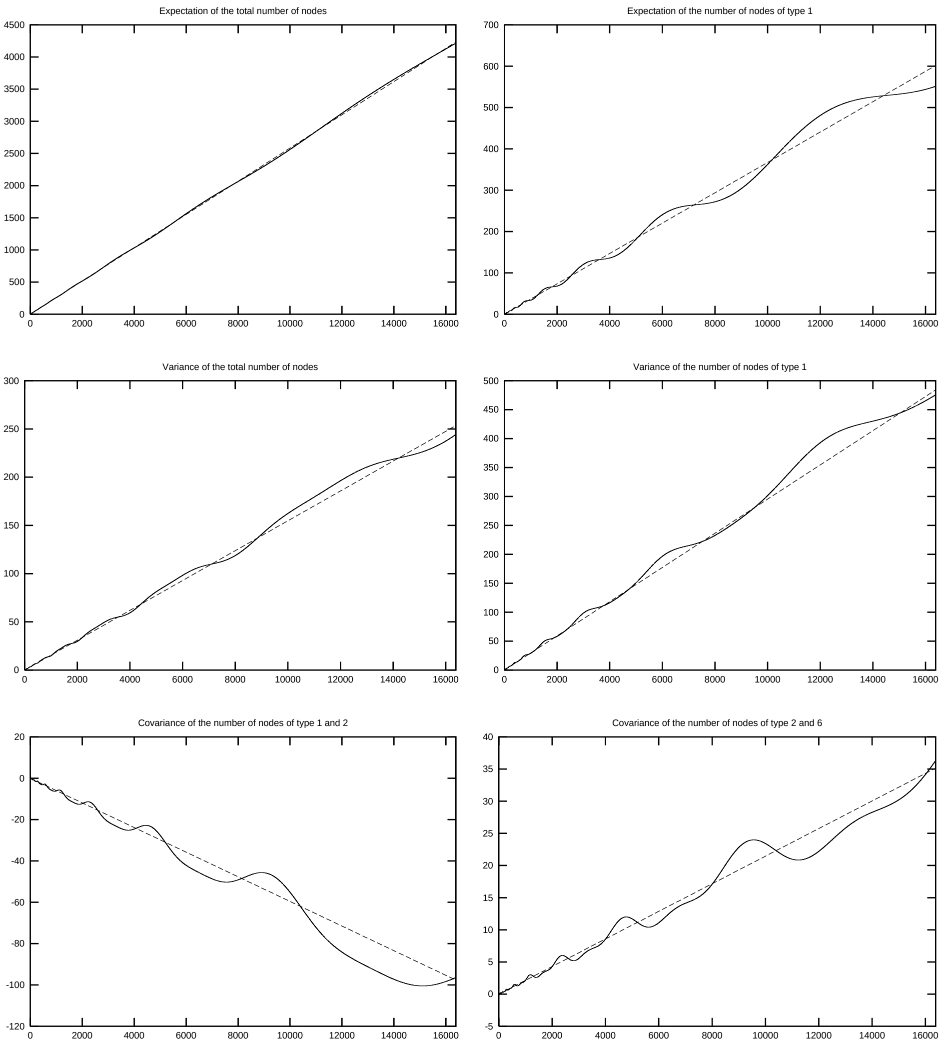


Figure 2: Graphics for recursions and simple approximations.

under an operating system that takes 4 bytes per pointer (a common choice in modern computers), then a node is a data block of 28 bytes. (Recall that the fixed-allocation scheme does not require a hidden tag.) Let S_n be the total actual space in bytes taken by the tree. Then

$$\mathbf{E}[S_n] = 28 \sum_{1 \le k \le 6} \mathbf{E}\left[X_n^{(i)}\right] \approx 7.237748n.$$

This result coincides with the average-case analysis in [13]. Let U_n be the number of useless bytes in the leaves (unused key slots and linkage pointers), that is

$$U_n = \sum_{1 \le k \le 5} \left[28 - 4(4-i)\right] X_n^{(i)}.$$

And so,

$$\mathbf{E}[U_n] = \sum_{1 \le k \le 5} (12 + 4i)\mathbf{E}\left[X_n^{(i)}\right] = 2.631492n;$$

only about 54% of the nodes make full use of their allocated block space, and the space utilization efficiency is about 64%. The cost of conversion in this model is not really meaningful—as we only have one node type, there are no conversions apart from the overhead associated with the allocation of each new block. We infer from the fifth step of the paradigm that

$$\frac{S_n}{(7.237748\ldots + \theta_1(n))\, n} \to 1, \qquad \frac{U_n}{(2.631492\ldots + \theta_2(n))\, n} \to 1,$$

in probability, where $\theta_1(n)$ and $\theta_2(n)$ are very small periodic fluctuations. We also have the central limit tendency of

$$(S_n - 7.237748\ldots\, n)/\sqrt{\theta_3(n)n}$$

to a normal distribution, for a periodic function θ_3.

(*b*) *Exact Fit Scheme:* This is a multitype scheme with absolute control over memory. A request for a certain number of bytes is honored exactly. This kind of control may be achieved by accessing a "heap" of free space. When a bucket digital tree with bucket capacity 5 grows under this scheme, the tree has 6 different types ($p = 6$). Type-i bucket for $i = 1, \ldots, 5$ will have the exact space required for i pointers to keys in the sequential data file, and a hidden tag, meaning $\beta_i = i$ for $i = 1, \ldots, 5$; the hidden tag is usually only 1 word of memory (4 bytes on most mainframes), and as we assumed in the previous example, a pointer occupies 4 bytes. In terms of actual bytes, a type i node requires $4 + 4i$ bytes, for $i = 1, \ldots, 5$. A type 6 node will have the exact space for 5 pointers to keys in the sequential data file (20 bytes), two linkage pointers (8 bytes), and a hidden tag (4 bytes), a total of 32 bytes. Clearly, $X_n^{(i)}$ is the number of nodes of type i in the most detailed model, the one difference being that the nodes in the exact fit model are implemented differently—all extra unused key-pointer slots as well as the linkage pointers are eliminated from the leaves of the detailed model and replaced by more compact exact-fit blocks. Thus the probabilistic analysis is essentially the same. However, the *actual space* is different. The linear combination Z_n is exactly the same for this model as in the fixed-block model of the previous example. (In fact, here $\mathbf{K}$ is the identity matrix). Let $\tilde{S}_n$ be the total amount of space taken by a tree. The actual space required after n insertions is

$$\tilde{S}_n = 8X_n^{(1)} + 12X_n^{(2)} + 16X_n^{(3)} + 20X_n^{(4)} + 24X_n^{(5)} + 32X_n^{(6)}.$$

We find

$$\mathbf{E}[\tilde{S}_n] = 4 \sum_{1\le k\le 5} (k+1)\mathbf{E}\left[X_n^{(k)}\right] + 32\mathbf{E}\left[X_n^{(6)}\right] \approx 6.151524n,$$

which is only about 85% of the space needed under fixed-allocation, even though each node now carries a full word identifying tag. Accordingly, we have the central limit tendency of

$$(S_n - 6.151524\ldots\ n)/\sqrt{\theta_4(n)n}$$

to a normal distribution, for a periodic function θ_4.

(c) *The buddy system:* In this system blocks of sizes 2^k memory words can be allocated by the system. In practical implementations, however, doubling the size of a block is practiced up to a certain threshold, after which the increase follows a linear (instead of exponential) growth. For example, the UCBD 4.3 system can allocate blocks of sizes $2, 4, \ldots, 1024, 2048, 3072, 4096, 5120, \ldots$. If a user requests a block of size 27 words, say, the system will allocate one having 32 words, the smallest system's block that can accommodate the request. The buddy system reserves one word of memory for its own use. An internal node of a bucket digital tree with $b = 5$ uses a node with 8 words, of which five are for the keys (or pointers to them), two are linkage pointers and one is for the hidden tag. Thus, the growing tree can use the first three types, with key capacities $\beta_1 = 1$, $\beta_2 = 3$, and $\beta_3 = 5$. Here

$$\mathbf{K} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{pmatrix}, \quad \mathbf{M} = \begin{pmatrix} 1 & 1 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{pmatrix}.$$

Thus

$$\mathbf{E}[\mathbf{Y}_n] = \mathbf{K}\mathbf{E}[\mathbf{X}_n] \approx \begin{pmatrix} 0.036723 \\ 0.049966 \\ 0.171802 \end{pmatrix} n.$$

The total space allocated, $\hat{S}_n$ has an average:

$$\mathbf{E}[\hat{S}_n] = 8\mathbf{E}[Y_n^{(1)}] + 16\mathbf{E}[Y_n^{(2)}] + 32\mathbf{E}[Y_n^{(8)}] \approx 6.590904n;$$

about 91% of the average space under fixed allocation and 7% more than the average space for exact-fit allocation.

Acknowledgements.

Several authors conducted some of this research while visiting INRIA, Rocquencourt (France). We thank Philippe Flajolet for helpful discussions and support.

References

[1] Baeza-Yates, R. (1987). Some average measures in m–ary search trees. *Information Processing Letters*, **25**, 375–381.

[2] Baeza-Yates, R. and Larson, P. (1989). Performance of B^+–trees with partial expansions. *IEEE Transactions on Knowledge and Data Engineering*, **1**, 248–257.

[3] Bai, Z.-D., Hwang, H.-K., Liang, W.-Q. and Tsai, H.-T. (2001). Limit theorems for the number of maxima of random samples from planar regions. *Electronic Journal of Probability*, **6**, paper 3, 41 pages.

[4] Billingsley, P. (1995). *Probability and Measure.* Third edition. John Wiley & Sons, New York.

[5] Chen, W.-M. and Hwang, H.-K. (2001). Analysis of two randomized algorithms for finding the maximum in a broadcast communication model, submitted.

[6] Chern, H.-H. and Hwang, H.-K. (2000). Phase change in random m-ary search trees and generalized quicksort, *Random Structures and Algorithms*, in revision.

[7] Cools, R., Laurie, D. and Pluym, L. (1997). Algorithm 764: Cubpack ++: A $C++$ package for automatic two-dimensional cubature., *ACM Trans. Math. Softw.*, **23**, 1–15.

[8] Devroye, L. (1986). *Lecture Notes on Bucket Algorithms.* Birkhäuser Boston, Boston, MA.

[9] Doetsch, G. (1950). *Handbuch der Laplace-Transformation. Band I. Theorie der Laplace-Transformation.* Verlag Birkhäuser, Basel.

[10] Doetsch, G. (1955). *Handbuch der Laplace-Transformation. Bd II. Anwendungen der Laplace-Transformation. 1. Abteilung.* Birkhäuser Verlag, Basel und Stuttgart.

[11] Flajolet, P., Gourdon, X. and and Dumas, P. (1995). Mellin transforms and asymptotics: Harmonic sums. *Theoretical Computer Science*, **144**, 3–58.

[12] Flajolet, P. and Odlyzko, A. (1990). Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics*, **3**, 216–240.

[13] Flajolet, P. and Richmond, B. (1992). Generalized digital trees and their difference-differential equations. *Random Structures and Algorithms*, **3**, 305–320.

[14] Frederickson, G. (1979). Improving storage utilization in balanced trees. *Proc.* 17^{th} *Allerton Conf. Commun., Contr. and Comput.*, 255–264. Monticello, Illinois.

[15] Gaines, H. F. (1956). *Cryptanalysis.* Dover Publications Inc., New York.

[16] Gonnet, G. and Baeza-Yates, R. (1991). *Handbook of Algorithms and Data Structures.* Second Edition. Addison-Wesley, Reading, Mass.

[17] Hoshi, M. and Flajolet, P. (1992). Page usage in quadtree indexes. *BIT*, **32**, 384–402.

[18] Hubalek, F. (1993). *Beiträge zur Analyse verallgemeinerter digitaler Suchbäume*. Dissertation, Technical University of Vienna.

[19] Hubalek, F. (2000). On the variance of the internal path length of generalized digital trees—the Mellin convolution approach. *Theoretical Computer Science*, **242**, 143–168.

[20] Hubalek, F. (2001). Partial fractions and Mellin convolutions for the analysis of bucket digital search trees —- some numerical aspects, Technical University of Vienna, in preparation.

[21] Hwang, H.-K. (1997). Asymptotic estimates of elementary probability distributions. *Studies in Applied Mathematics*, **99**, 393–417.

[22] Hwang, H.-K. and Neininger, R. (2001). Phase change of limit laws in the quicksort recurrence under varying toll functions, submitted.

[23] Hwang, H.-K. and Tsai, H.-T. (2000) Quickselect and Dickman function, submitted.

[24] Jacquet, P. and Régnier, M. (1986) Trie partitioning process: limiting distributions. In *CAAP '86* (Nice, 1986), *Lecture Notes in Computer Science*, **214**, 196–210, Springer, Berlin.

[25] Jacquet, P. and Régnier, M. (1987). Normal limiting distribution of the size of tries. in *Performance '87* (Brussels, 1987), 209–223. North-Holland, Amsterdam, 1988.

[26] Jacquet, P., Szpankowski, W. and Tang, J. (2001). Average profile of the Lempel-Ziv parsing scheme for a Markovian source, *Algorithmica*, to appear.

[27] Kirschenhofer, P. and Prodinger, H. (1988). B-Tries: A paradigm for the use of number theoretic methods in the analysis of algorithms. In *Contributions to General Algebra 6*, 141–154. Hölder-Pichler-Tempsky, Vienna and Teubner, Stuttgart.

[28] Kirschenhofer, P., Prodinger, H. and Szpankowski, W. (1989). On the balance property of Patricia tries: external path length viewpoint. *Theoretical Computer Science*, **68**, 1–17.

[29] Kirschenhofer, P., Prodinger, H., and Szpankowski, W., Digital search trees again revisited: The internal path length perspective. *SIAM J. Comput.* 23, 598–616, 1994.

[30] Knuth, D. (1998). *The Art of Computer Programming: Sorting and Searching*, Vol 3. Second edition, Addison-Wesley, Reading, Mass.

[31] Lew, W. and Mahmoud, H. (1994). The joint distribution of elastic buckets in multiway search trees. *SIAM Journal on Computing*, **23**, 1050–1074.

[32] Lew, W. (1995). A probabilistic analysis of m-ary and digital search trees. Dissertation. The George Washington University.

[33] Lomet, D. (1987). Partial expansions for file organizations with an index. *ACM Transactions on Database Systems*, **12**, 65–84.

[34] Louchard, G., Szpankowski, W. and Tang, J. (1999) Average profile of the generalized digital search tree and the generalized Lempel-Ziv algorithm. *SIAM Journal on Computing*, **28**, 904–934.

[35] Mahmoud, H., Flajolet, P., Jacquet, P. and Régnier, M. (2000). Analytic variations on bucket selection and sorting. *Acta Informatica*, **36**, 735–760.

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

[37] Mahmoud, H. and Papadakis, T. (1992). A probabilistic analysis of fixed and elastic buckets in tries and PATRICIA trees. In *Proc.* 13^{th} *Allerton Conf. Commun., Contr. and Comput.*, 874–883. Monticello, Illinois.

[38] Piessens, R., de Doncker-Kapenga, E., Überhuber, C. W. and Kahaner, D.K. (1983). QUADPACK, A subroutine package for automatic integration, Springer-Verlag, Berlin.

[39] Pittel, B. (1986). Paths in a random digital tree: Limiting distributions. *Advances in Applied Probability*, **18**, 139–155.

[40] Preparata, F. P. and Shamos, M. I. (1985) *Computational Geometry*. Springer-Verlag, New York.