CONTRIBUTIONS TO GENERAL ALGEBRA 6
Verlag Hölder-Pichler-Tempsky, Wien 1988 - Verlag B. G. Teubner, Stuttgart

Peter Kirschenhofer - Helmut Prodinger

B-TRIES: A PARADIGM FOR THE USE OF NUMBERTHEORETIC METHODS IN THE ANALYSIS OF ALGORITHMS

Dedicated to the memory of Prof. W. Nöbauer

Abstract

In this paper the variance of the size of an important data structure (called *b*-tries) is studied using transformation results from the theory of modular functions. This continues research work due to Knuth, Flajolet, Szpankowski and the authors.

1. Introduction

Analysis of algorithms is a rapidly developing area in Theoretical Computer Science: The cost of the performance of algorithms, including topics as the storage requirements of data structures and the execution time of certain subroutines, is usually described in terms of worst case behaviour and average case behaviour.

Sophisticated methods have been applied in order to optimize the worst-case behaviour of algorithms; nevertheless for practical purposes the question of the *average-case performance of data structures and algorithms* is considered more and more to be of great importance.

From the mathematician's point of view, average-case analysis of algorithms is an area asking for the use of methods from very different fields of mathematics, such as *combinatorics, probability theory, complex variable theory* and, last but not least, methods usually applied in *analytic number theory*. The purpose of this paper is to demonstrate the latter via the analysis of a data structure which is of considerable importance in Computer Science:

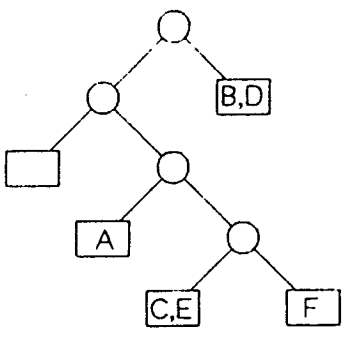
Digital searching is a familiar technique for the storage and retrieval of information using the lexicographic (digital) structure of records. The most important algorithms are described in Knuth's famous book [9]. A (finite) set of records represented by keys over some alphabet (e.g. 0-1-sequences) is represented by a tree called *trie*, where the end-nodes (or "leaves") contain the keys and the edges are labelled by letters from the

alphabet. The path from the root to a leaf is a minimal prefix of the key stored in the leaf. An important variant of tries is obtained using a sequential storage algorithm for subtrees with a size less than or equal to a fixed bound b: this enables to improve the storage utilization by reducing the number of pointers used. Such a trie is called a *b-trie* (compare [3], [5], [13], [14]). In other words: Given a finite set of keys, which we assume to be given as 0-1-sequences, the corresponding (*binary*) *b-trie* may be constructed as follows:
If there are less than or equal to b keys all keys are stored in a single leaf. Otherwise divide the set of keys into 2 (possibly empty) classes according to the first bit and proceed with each of the classes and the following bit in the same way until each class has less than or equal to b elements: These classes are stored in the leaves determined by their minimal common prefix.

Example. The binary 2-trie created by the keys

A = 01011...
B = 10111...
C = 01101...
D = 10011...
E = 01100...
F = 01110...

is



Important parameters of this data-structure have been studied by Knuth [9], Flajolet [5] and Szpankowski [13], [14]. For practical purposes the question of storage utilization, i.e. the number of internal nodes of the created b-trie, is of great importance.

Under the assumption that all 0-1-sequences (of infinite length) are equally probable as keys Knuth [9] gives for the average number l_N of internal nodes of a b-trie generated from N keys the asymptotic formula

$$l_N = \frac{N}{\log 2}\left(\frac{1}{b} + \delta_1(\log_2 N)\right) + O(1), \tag{1.1}$$

where $\delta_1(x)$ is a continuous, periodic function of period 1 with known Fourier expansion and small amplitude. (Compare Theorem 4; it should be noticed that in Knuth's solution the Fourier coefficients are slightly erroneous!)

For practical purposes it is of great importance to gain more insight into the distribution of the random variable in discussion. For this reason we aim to establish an asymptotic formula for the *variance* V_N: Interestingly enough it turns out that the variance is quite small, i.e. of order N. The mathematics to achieve this result relies on the application of methods which come from the area of analytic number theory, especially deep transformation results for *modular functions*.

In the following we *sketch our method* in short: In the first step we set up recursions for the probability generating functions. By differentiation we get recurrence relations

for the factorial moments which may be solved explicitly via the solution of corresponding functional equations. The explicit expressions are always of the type

$$\sum_{k \geq b+1} \binom{N}{k} (-1)^k f(k) ,$$

where f may be extended into the complex plane. To obtain asymptotic information we use the following lemma which Knuth attributes to S. O. Rice [9; ex. 5.2.2.-54].

Lemma 1. (compare [10]) *Let C be a curve surrounding the points $b+1, \ldots, N$ and $f(z)$ be analytic within C. Then*

$$\sum_{k \geq b+1} \binom{N}{k} (-1)^k f(k) = -\frac{1}{2\pi i} \int_C [N; z] f(z)\, dz$$

with

$$[N; z] = \frac{(-1)^{N-1} N!}{z(z-1)\ldots(z-N)} .$$

In our application f is a meromorphic function and the asymptotic expansion of the factorial moments is obtained via

$$\sum \mathrm{Res}([N; z] f(z)),$$

where the sum is taken over all poles different from $b+1, \ldots, N$. Thus we first derive a more accurate asymptotic expansion for the expectation l_N (Theorem 4). The variance is computed via the formula

$$V_N = w_N + l_N - l_N^2 ,$$

where w_N denotes the second factorial moment. It turns out that formally V_N starts with order N^2, namely a term of the form

$$N^2 (A + \tau(\log_2 N) - \frac{1}{\log^2 2} \delta_1^2(\log_2 N)),$$

where A is a constant and τ (as well as δ_1) is a continuous periodic function of mean zero. The crucial point of the derivation is now to find a "simple" explicit expression for the zeroth Fourier coefficient $[\delta_1^2]_0$ of $\delta_1^2(x)$ (Note carefully that $\delta_1^2(x)$ does not have mean zero!). This is gained by the application of some *series transformation* results due to *Ramanujan*.

Surprisingly enough

$$A - \frac{1}{\log^2 2} [\delta_1^2]_0 = 0 ,$$

and, by a continuity argument and the nonnegativeness of the variance, we conclude that V_N is of order N (Proposition 9).

The exact form of the leading term follows from an accurate residue calculation (Theorem 10).

Some notational remarks: We will frequently use the following abbreviations:

1) $[z^n]f(z)$ denotes the n-th coefficient in the Laurent series $f(z)$.

2) $L = \log 2$.

3) $\chi_k = \frac{2k\pi i}{L}$.

2. Results

Let $F_N(z)$ be the probability generating function where N refers to the number of records and the coefficient of z^k is the probability that the b-trie has k internal nodes. Then we have

Lemma 2. $$F_i(z) = 1 \quad (i=0,1,\ldots,b),$$
$$F_N(z) = z\sum_{k=0}^{N} 2^{-N}\binom{N}{k}F_k(z)F_{N-k}(z) \qquad (N>b).$$

Proof. $2^{-N}\binom{N}{k}$ is the probability that k records start with 0 and $N-k$ records start with 1; the factor z reflects the root.

Now we get for the expectation $l_N = F_N'(1)$:

Lemma 3.
$$l_N = \sum_{k=b+1}^{N}\binom{N}{k}\frac{(-1)^{k+b+1}\binom{k}{b}}{1-2^{1-k}} \quad , \qquad (N>b),$$
$$l_0 = l_1 = \ldots = l_b = 0.$$

Proof. From Lemma 2 we have

$$l_N = 1 + 2^{1-N}\sum_{k=0}^{N}\binom{N}{k}l_k \qquad (N>b),\quad l_0 = l_1 = \ldots = l_b = 0.$$

Now let

$$L(z) = \sum_{N\geq 0} l_N \frac{z^N}{N!}$$

be the exponential generating function of the l_N's.

The recursion immediately translates into

$$L(z) = e^z - e_b(z) + 2L(\tfrac{z}{2})e^{z/2}$$

with the truncated exponential function

$$e_b(z) = 1 + z + \frac{z^2}{2!} + \ldots + \frac{z^b}{b!}.$$

The functional equation for $L(z)$ becomes easier by introducing

$$\tilde{L}(z) = e^{-z}L(z) = \sum_{N\geq 0}\tilde{l}_N\frac{z^N}{N!} \quad :$$

$$\tilde{\tilde{L}}(z) = 2\tilde{\tilde{L}}(\tfrac{z}{2}) + 1 - e_b(z)e^{-z}.$$

Taking coefficients we find

$$\tilde{\tilde{l}}_N = 2^{1-N}\tilde{\tilde{l}}_N + \delta_{N,0} - (-1)^N[1 - N + \binom{N}{2} - \binom{N}{3} + - \ldots + (-1)^b\binom{N}{b}].$$

Now we use the elementary identity (cf. [12])

$$\sum_{i=0}^{b}\binom{N}{i}(-1)^i = (-1)^b\binom{N-1}{b}$$

to get the result.

Theorem 4. *The expectation l_N of the size of b-tries from N records fulfills asymptotically*

$$l_N = \frac{N}{L}(\frac{1}{b} + \delta_1(\log_2 N)) - 1 - \frac{1}{2L}\delta_2(\log_2 N)$$

with

$$\delta_1(x) = (-1)^{b+1}\sum_{k\neq 0} e^{2k\pi i x}\,\Gamma(-1-\chi_k)\binom{\chi_k}{b}$$

and

$$\delta_2(x) = (-1)^{b+1}\sum_{k\neq 0} e^{2k\pi i x}\,\Gamma(1-\chi_k)\binom{\chi_k}{b}.$$

Proof. According to the Introduction we may apply Lemma 1 with

$$f(z) = \frac{(-1)^{b+1}}{1-2^{1-z}}\cdot\binom{z-1}{b}.$$

We find the following residues:

$$\mathrm{Res}([N;z]f(z)\,;\, z=1) = \frac{N}{bL}$$

$$\mathrm{Res}([N;z]f(z)\,;\, z=1+\chi_k) = (-1)^{b+1}\frac{1}{L}N^{1+\chi_k}\Gamma(-1-\chi_k)\Big(1-\frac{\chi_k(\chi_k+1)}{2N}\Big)\binom{\chi_k}{b}$$

$$\mathrm{Res}([N;z]f(z)\,;\, z=0) = -1.$$

Adding up these values the result is obtained.

Now we turn to the second factorial moment $w_N = F_N''(1)$; from this we will find the variance V_N by

$$V_N = w_N + l_N - l_N^2.$$

Lemma 5.

$$w_N = \sum_{k=b+1}^{N}\binom{N}{k}(-1)^k\,\tilde{w}_k$$

with

$$\tilde{w}_N = \frac{1}{1-2^{1-N}}\Bigg[\frac{2^{1-N}}{1-2^{2-N}}\cdot\Bigg[(-1)^b\binom{N-1}{b}(1-\frac{2}{1-2^{1-N}}) + \sum_{k=b+1}^{N}\binom{N}{k}\binom{k-1}{b}\binom{N-k-1}{b}(1+\frac{2}{2^{k-1}-1})\Bigg] + (-1)^{b+1}\frac{2^{2-N}}{1-2^{1-N}}\binom{N-1}{b}\Bigg].$$

Proof. From Lemma 2 we get by twofold differentiation

$$w_N = 2^{2-N}\sum_{k=0}^{N}\binom{N}{k}l_k + 2^{1-N}\sum_{k=0}^{N}\binom{N}{k}w_k + 2^{1-N}\sum_{k=0}^{N}\binom{N}{k}l_k l_{N-k}\,.$$

or with

$$W(z) = \sum_{N\geq 0} w_N \frac{z^N}{N!}\,:$$

$$W(z) = 2(L(\tfrac{z}{2}))^2 + 4e^{z/2}L(\tfrac{z}{2}) + 2e^{z/2}W(\tfrac{z}{2})\,.$$

Again, let

$$\tilde{W}(z) = e^{-z}W(z) = \sum_{N\geq 0}\tilde{w}_N\frac{z^N}{N!}$$

and

$$\hat{L}(z) = (\tilde{L}(z))^2 = \sum_{N\geq 0}\hat{l}_N\frac{z^N}{N!}\,.$$

Then the functional equation can be simplified:

$$\tilde{W}(z) = 2\hat{L}(\tfrac{z}{2}) + 4\tilde{L}(\tfrac{z}{2}) + 2\tilde{W}(\tfrac{z}{2}),$$

and by taking coefficients

$$\tilde{w}_N = \frac{1}{1-2^{1-N}}\left[2^{1-N}\hat{l}_N + 2^{2-N}\tilde{l}_N\right] \text{ for } N\geq 0\,.$$

Now we need an expression for $\hat{l}_N$ which can be extended for complex values for N.

From

$$\tilde{L}(z) - 1 + e_b(z)e^{-z} = 2\tilde{L}(\tfrac{z}{2})$$

we obtain by squaring

$$\hat{L}(z) + 1 + e_b^2(z)e^{-2z} - 2\tilde{L}(z) + 2e_b(z)e^{-z}\tilde{L}(z) - 2e_b(z)e^{-z} = 4\hat{L}(\tfrac{z}{2})$$

or, by extracting coefficients, for $N>b$

$$\hat{l}_N - 2\tilde{l}_N + [\tfrac{z^N}{N!}]e_b^2(z)e^{-2z} + 2[\tfrac{z^N}{N!}]e_b(z)e^{-z}\tilde{L}(z) - 2\sum_{k=0}^{b}\binom{N}{k}(-1)^{N-k} = 2^{2-N}\hat{l}_N\,.$$

Now we note that

$$\sum_{k=0}^{b}\binom{N}{k}(-1)^{N-k} = \binom{N-1}{b}(-1)^{N+b}$$

and

$$[\tfrac{z^N}{N!}]e_b^2(z)e^{-2z} = [\tfrac{z^N}{N!}]\Big(\sum_{N\geq 0}(-1)^{N+b}\binom{N-1}{b}\frac{z^N}{N!}\Big)^2 = (-1)^N\sum_{k=0}^{N}\binom{N}{k}\binom{k-1}{b}\binom{N-k-1}{b};$$

further

$$[\tfrac{z^N}{N!}]e_b(z)e^{-z}\tilde{L}(z) = \sum_{k=0}^{N}\binom{N}{k}(-1)^{N-k+b}\binom{N-k-1}{b}\tilde{l}_k\,.$$

Hence

$$(1-2^{2-N})\hat{l}_N = 2\tilde{l}_N + 2(-1)^{N+b}\binom{N-1}{b} + (-1)^{N+1}\sum_{k=0}^{N}\binom{N}{k}\binom{k-1}{b}\binom{N-k-1}{b}$$
$$+ 2(-1)^{N+1+b}\sum_{k=b+1}^{N}\binom{N}{k}(-1)^k\binom{N-k-1}{b}\tilde{l}_k\,.$$

Now we use that

$$\tilde{l}_N = \frac{(-1)^{N+1+b}}{1-2^{1-N}}\binom{N-1}{b}$$

and find

$$\hat{l}_N = \frac{1}{1-2^{2-N}}\left[(-1)^{N+b}\binom{N-1}{b}\left(1-\frac{2}{1-2^{1-N}}\right) + (-1)^N \sum_{k=b+1}^{N}\binom{N}{k}\binom{k-1}{b}\binom{N-k-1}{b}\left(1+\frac{2}{2^{k-1}-1}\right)\right]$$

which finishes the proof.

Interestingly enough the case $b=1$ is somehow different. Since this case was discussed at length in our paper [7] we concentrate in the remainder on $b \geq 2$.

Lemma 6.

$$w_N = \sum_{k=b+1}^{N}\binom{N}{k}(-1)^k f(k)$$

with

$$f(z) = \frac{2^{2-z}}{(1-2^{1-z})^2}(-1)^{b+1}\binom{z-1}{b} + \frac{2^{1-z}}{(1-2^{1-z})(1-2^{2-z})}\left[(-1)^{b+1}\binom{z-1}{b}\frac{2}{1-2^{1-z}}\right.$$

$$\left. + 2\binom{z-1}{b}(-1)^b\sum_{i=0}^{b}(-1)^i\binom{z-1-b}{i}2^{z-1-b-i} + 2\sum_{k\geq b+1}\binom{z}{k}\binom{k-1}{b}\binom{z-k-1}{b}\frac{1}{2^{k-1}-1}\right].$$

Proof. Note that by $\binom{N}{k} = \binom{N-1}{k} + \binom{N-1}{N-k}$ we have

$$\sum_{k=b+1}^{N-b-1}\binom{N}{k}\binom{k-1}{b}\binom{N-k-1}{b} =$$

$$= 2\binom{N-1}{b}\sum_{k}\binom{k-1}{b}\binom{N-1-b}{k}$$

$$= 2\binom{N-1}{b}\sum_{k}\binom{N-1-b}{k}(-1)^b\sum_{i=0}^{b}\binom{k}{i}(-1)^i$$

$$= 2\binom{N-1}{b}(-1)^b\sum_{i=0}^{b}(-1)^i\binom{N-1-b}{i}\sum_{k=b+1}^{N-b-1}\binom{N-1-b-i}{k-i}$$

$$= 2\binom{N-1}{b}(-1)^b\sum_{i=0}^{b}(-1)^i\binom{N-1-b}{i}\left[2^{N-1-b-i} - \sum_{k=i}^{b}\binom{N-1-b-i}{k-i}\right]$$

$$= 2\binom{N-1}{b}(-1)^b\left[\sum_{i=0}^{b}(-1)^i\binom{N-1-b}{i}2^{N-1-b-i} - 1\right],$$

since

$$\sum_{k=0}^{b}\sum_{i=0}^{k}(-1)^i\binom{N-1-b}{i}\binom{N-1-b-i}{k-i} = \sum_{k}\sum_{i}(-1)^i\binom{N-1-b}{N-1-b-k}\binom{k}{k-i}$$

$$= \sum_{k=0}^{b}\binom{N-1-b}{k}\delta_{k,0} = 1.$$

This completes the proof.

Lemma 7.

$$w_N \sim -\binom{N}{2}\cdot\frac{1}{L}\left[-\frac{3}{b(b-1)}+\frac{2^{1-2b}}{b(b-1)}\binom{2b}{b}+\frac{2(-1)^{b+1}}{b}\binom{2b}{b}\sum_{l\ge b}(-1)^{l-1}\binom{l+b-1}{2b-1}\frac{1}{l^2-1}\cdot\frac{1}{2^l-1}\right]$$

$$-\frac{4N}{bL}+\frac{N}{bL}2^{-2b}\binom{2b}{b}+\frac{2N}{L}(-1)^b\binom{2b}{b}\sum_{l\ge b}(-1)^l\binom{l+b}{2b}\frac{1}{l(l+1)}\cdot\frac{1}{2^l-1}$$

$$+\frac{N^2}{L^2}\cdot\delta_3(\log_2 N)+\frac{N}{L}\cdot\delta_4(\log_2 N).$$

The periodic functions $\delta_3(x)$ and $\delta_4(x)$ have mean 0; their Fourier coefficients could be determined in principle.

Proof. We can use again Rice's method (Lemma 1) with the function $f(z)$ defined in Lemma 5. We see immediately that for $b\ge 3$:

$$f(3)=\ldots=f(b)=0,$$

therefore there are no residues of $[N;z]f(z)$ at $3,\ldots,b$. Now we compute $f(2)$:

$$f(2)=\frac{1}{L}\left[\frac{-4}{b(b-1)}+\frac{2}{b(b-1)}\sum_{i=0}^{b}\binom{b+i-2}{b-2}2^{1-b-i}\right.$$

$$\left.+4\sum_{k\ge b+1}\frac{(-1)^{k+1}}{k(k-1)(k-2)}\binom{k-1}{b}(-1)^b\binom{k+b-2}{b}\frac{1}{2^{k-1}-1}\right].$$

From the discussion of Banach's matchbox problem (see e.g. [4] or [12]) we know that

$$\sum_{i=0}^{b}\binom{b+i-2}{b-2}2^{1-b-i}=\frac{1}{2}+2^{-2b}\binom{2b}{b},$$

hence

$$f(2)=\frac{1}{L}\left[\frac{-3}{b(b-1)}+\frac{2^{1-2b}}{b(b-1)}\binom{2b}{b}+4(-1)^{b+1}\sum_{k\ge b+1}\frac{(-1)^k}{k(k-1)(k-2)}\binom{k-1}{b}\binom{k+b-2}{b}\frac{1}{2^{k-1}-1}\right]$$

and thus

$$\operatorname{Res}([N;z]f(z);z=2)=-\binom{N}{2}\cdot f(2).$$

The residues at $z=2+\chi_k$ will not be computed explicitly. We turn now to $f(1)$: A somewhat lengthy but elementary computation which uses again the identity from Banach's matchbox problem results in

$$\operatorname{Res}([N;z]f(z);z=1)=-\frac{4N}{bL}+\frac{N}{bL}\binom{2b}{b}2^{-2b}$$

$$+\frac{2N}{L}(-1)^{b-1}\binom{2b}{b}\sum_{k\ge b+1}(-1)^k\binom{k+b-1}{2b}\frac{1}{k(k-1)}\cdot\frac{1}{2^{k-1}-1}.$$

Thus the proof is finished.

Now we turn to the constant term in the Fourier expansion of $\delta_1^2(x)$.

Lemma 8. *The constant term* $[\delta_1^2]_0$ *in the Fourier expansion of* $\delta_1^2(x)$ *is*

$$[\delta_1^2]_0 = -\frac{1}{b^2} + \frac{3}{2}\frac{L}{b(b-1)} - \frac{L2^{-2b}}{b(b-1)}\binom{2b}{b} + \frac{L(-1)^b}{b}\binom{2b}{b}\sum_{l\geq b}(-1)^{l-1}\binom{l+b-1}{2b-1}\frac{1}{(l^2-1)(2^l-1)}.$$

Proof.

$$[\delta_1^2]_0 = \frac{1}{b!^2}\cdot 2\sum_{l\geq 1}\frac{|(\chi_l)_b|^2}{1+|\chi_l|^2}|\Gamma(\chi_l)|^2 = \frac{2}{b!^2}\sum_{l\geq 1}\sum_{d\geq 0}(-1)^d\frac{|\Gamma(\chi_l)|^2}{|\chi_l|^{2d+2}}\prod_{j=0}^{b-1}(|\chi_l|^2+j^2).$$

Now we define the coefficients $c_t(s)$ (as in [7]) by

$$\sum_{t=1}^{s-1}c_t(s)x^t = \prod_{j=0}^{s-2}(x+j^2)$$

and have with $s = b+1$

$$[\delta_1^2]_0 = \frac{L}{b!^2}\sum_{t=1}^{s-1}c_t(s)\sum_{d\geq 0}(-1)^d\left(\frac{4\pi^2}{L^2}\right)^{t-d-1}\cdot\sum_{l\geq 1}\frac{l^{2t-2d-3}}{\sinh\frac{2l\pi^2}{L}}.$$

With $m = t-d-1$ this expression equals

$$= \frac{L}{b!^2}\sum_{t=1}^{s-1}c_t(s)(-1)^{t+1}\sum_{-\infty<m\leq t-1}(-1)^m\left(\frac{4\pi^2}{L^2}\right)^m\cdot\sum_{l\geq 1}\frac{l^{2m-1}}{\sinh\frac{2l\pi^2}{L}}.$$

The part of the sum concerning $-\infty<m\leq 0$ equals 0, since for $s\geq 3$

$$\sum_{t=1}^{s-1}c_t(s)(-1)^{t+1} = 0.$$

Treating the remaining sum ($m\geq 1$) we consider

$$\sum_{l\geq 1}\frac{l^{2m-1}}{\sinh\frac{2l\pi^2}{L}} = 2h_m\left(\frac{\pi^2}{L}\right) - 2h_m\left(\frac{2\pi^2}{L}\right) \tag{2.1}$$

with

$$h_m(x) = \sum_{l\geq 1}\frac{l^{2m-1}}{e^{2lx}-1}.$$

For these functions the following transformation formulae (due to Ramanujan) are known ([11] resp. [2]): For $x,y>0$ with $xy=\pi^2$

$$x\cdot h_1(x) + y\cdot h_1(y) = \frac{x+y}{24} - \frac{1}{4} \tag{2.2}$$

and for $m\geq 2$

$$x^m\cdot h_m(x) - (-y)^m\cdot h_m(y) = (x^m-(-y)^m)\frac{B_{2m}}{4m}, \tag{2.3}$$

where B_n indicates the n-th Bernoulli number defined by

$$\frac{z}{e^z-1} = \sum_{n\geq 0}B_n\frac{z^n}{n!}.$$

Thus (2.1) turns into

$$2\cdot(-1)^{m+1}\left(\frac{L^2}{4\pi^2}\right)^m\left[\sum_{l\geq 1}\frac{(-1)^{l-1}\,l^{2m-1}}{2^l-1}+\frac{B_{2m}}{4m}(2^{2m}-1)-\delta_{m,1}\cdot\frac{1}{2L}\right].$$

From this we have

$$[\delta_1^2]_0=\frac{1}{b!^2}(C_1+C_2+C_3)$$

with C_1, C_2, C_3 referring to the 3 terms in the above expression.

$$C_1=2L\sum_{t=1}^{s-1}c_t(s)(-1)^t\left[(t-1)+\sum_{l\geq 2}\frac{(-1)^{l-1}}{2^l-1}\cdot\frac{l((l^{2t-2}-1)}{l^2-1}\right].$$

From the definition of the constants $c_t(s)$ it follows by an easy computation that (for $s\geq 3$)

$$\sum_{t=1}^{s-1}c_t(s)(-1)^t(t-1)=\frac{1}{2}(s-1)!(s-3)!$$

and

$$\sum_{t=1}^{s-1}c_t(s)(-1)^t\,l^{2t}=(-1)^{s-1}\cdot l\cdot(l+s-2)_{2s-3}$$

(with $(x)_k=x(x-1)\ldots(x-k+1)$). Hence

$$C_1=L(s-1)!(s-3)!+2L(-1)^{s-1}\sum_{l\geq s-1}\frac{(-1)^{l-1}\,(l+s-2)_{2s-3}}{(l^2-1)(2^l-1)}$$

and so ($s=b+1$)

$$\frac{C_1}{b!^2}=\frac{L}{b(b-1)}+\frac{L(-1)^b}{b}\binom{2b}{b}\sum_{l\geq b}(-1)^{l-1}\binom{l+b-1}{2b-1}\frac{1}{(l^2-1)(2^l-1)}.$$

We proceed with C_3, since C_2 is more complicated:

$$C_3=-\sum_{t=2}^{s-1}c_t(s)(-1)^t=-c_1(s)=-(s-2)!^2$$

and therefore

$$\frac{C_3}{b!^2}=-\frac{1}{b^2}.$$

Now we turn to C_2:

$$C_2=C_2(s)=L\sum_{t=2}^{s-1}c_t(s)(-1)^t\sum_{m=1}^{t-1}\frac{B_{2m}}{2m}(2^{2m}-1).$$

Since

$$c_t(s+1)=(s-1)^2c_t(s)+c_{t-1}(s)$$

and

$$\sum_{t=1}^{s-1}c_t(s)(-1)^{t+1}\cdot\frac{B_{2t}}{2t}(2^{2t}-1)=(2s-3)!2^{2-2s}$$

(compare [7] for the proof of the last identity) we find the following recursion:

$$C_2(s+1)=s(s-2)C_2(s)+L\frac{(2s-3)!}{2^{2s-2}}\quad\text{for } s\geq 2,\ C_2(2)=0.$$

Solving this recursion we get

$$\frac{C_2(s)}{b!^2} = \frac{C_2(b+1)}{b!^2} = \frac{L}{b(b-1)} \sum_{k=2}^{b} \binom{2k-2}{k} \frac{2^{2-2k}}{2k-2}.$$

As a final step we note that

$$\begin{aligned}
\sum_{k=2}^{b} \binom{2k-2}{k} \frac{2^{2-2k}}{2k-2} &= \frac{1}{2} \sum_{k=0}^{b-1} \frac{1}{k+1} \binom{2k}{k} 2^{-2k} - \frac{1}{2} \\
&= [z^{b-1}] \frac{1}{2} \frac{1}{1-z} \sum_{k \geq 0} \frac{1}{k+1} \binom{2k}{k} 2^{-2k} z^k - \frac{1}{2} \\
&= [z^b] \left(\frac{1}{1-z} - \frac{1}{\sqrt{1-z}} \right) - \frac{1}{2} \\
&= \frac{1}{2} - 2^{-2b} \binom{2b}{b}.
\end{aligned}$$

So we obtain

$$\frac{C_2(s)}{b!^2} = \frac{L}{2b(b-1)} - \frac{L}{b(b-1)} \cdot 2^{-2b} \binom{2b}{b},$$

and thus the announced result.

As a consequence of Lemma 7 we find

Proposition 9. *The variance V_N fulfills*

$$V_N = O(N) \text{ for } N \to \infty.$$

Proof. We have

$$V_N = w_N + l_N - l_N^2.$$

Collecting the contributions of order N^2 to w_N and l_N^2 we find:

$$[N^2] V_N = \frac{1}{L^2} (\delta_3(\log_2 N) - 2b\delta_1(\log_2 N) - \delta_1^2(\log_2 N) + [\delta_1^2]_0) = \delta_5(\log_2 N),$$

where $\delta_5(x)$ is contiunous since its Fourier series is absolutely convergent, periodic with period 1 and mean zero. We claim that $\delta_5(x)$ vanishes identically:

If $\delta_5(x)$ would not do so we could find an $\varepsilon > 0$ and an interval, say $[a,b] \subseteq [0,1]$, such that $\delta_5(x) < -\varepsilon$ for $x \in [a,b]$. Since $\log_2 N$ is dense modulo 1, the variance V_N would be negative for an infinity of values N, an obvious contradiction.

Remark: It shall be noticed that from the last proof it follows that all Fourier coefficients of

$$\delta_3(x) - 2b\delta_1(x) - \delta_1^2(x) + [\delta_1^2]_0$$

are zero, which yields identities for convolutions of the Γ-function occuring in the Fourier coefficients of $\delta_1^2(x)$.

Theorem 10. *The variance V_N of the size of b-tries from N records fulfills asympotically*

$$V_N \sim \frac{N}{L}\left[\frac{1}{b\cdot 2^{2b+1}}\binom{2b}{b} + \frac{(-1)^{b+1}}{b}\binom{2b}{b}\sum_{l\geq b}(-1)^l \frac{l^2-b}{l(l+1)}\binom{l+b-1}{2b-1}\frac{1}{2^l-1}\right] + N\cdot\delta_6(\log_2 N),$$

where $\delta_6(x)$ is a continuous, periodic function with mean zero and small amplitude.

Proof. We have $V_N = w_N + l_N - l_N^2$. Collecting the contributions of order N we get (apart from the periodic fluctuations of mean zero):

$$\frac{1}{L}\left[\frac{1}{b\cdot 2^{2b}}\binom{2b}{b} - \frac{1}{Lb^2} + 2(-1)^b\binom{2b}{b}\sum_{l\geq b}(-1)^l\binom{l+b}{2b}\frac{1}{(l+1)l(2^l-1)} - \frac{1}{L}[\delta_1^2]_0 + \frac{1}{L}[\delta_1\delta_2]_0\right].$$

The expression $[\delta_1\delta_2]_0$ may now be treated in a similar manner as $[\delta_1^2]_0$ in the proof of Lemma 7 to get

$$[\delta_1\delta_2]_0 = [\delta_1^2]_0 + \frac{1}{b^2} - L\left[\frac{(-1)^b}{b}\binom{2b}{b}\sum_{l\geq b}(-1)^l\binom{l+b-1}{2b-1}\frac{1}{2^l-1} + \frac{1}{2b}\binom{2b}{b}2^{-2b}\right]$$

which gives the announced formula immediately.

Remarks: 1) The Fourier coefficients of $\delta_6(x)$ could be determined via the residues at $z=2+\chi_k$ and $z=1+\chi_k$. Since they are rather involved we omit explicit expressions here.

2) For b getting large the dominating term in the expression of Theorem 10 is

$$\frac{1}{b\cdot 2^{2b+1}}\binom{2b}{b} \sim \frac{1}{2\sqrt{\pi}}b^{-3/2}.$$

This can be seen by a comparison of the series in the expression of Theorem 10 with

$$\frac{1}{b}\binom{2b}{b}\sum_{l\geq b}(-1)^{l+b}\binom{l+b-1}{2b-1}\frac{1}{2^l} = \frac{1}{b}\binom{2b}{b}\left(\frac{2}{9}\right)^b \sim \frac{1}{\sqrt{\pi}}b^{-3/2}\left(\frac{8}{9}\right)^b.$$

3) For the variance (as well as for the expectation) the amplitudes of the periodic fluctuations of mean zero are small compared with the dominating terms: For small values of b this follows from explicit estimates using

$$|\Gamma(iy)|^2 = \frac{\pi}{y\cdot\sinh(\pi y)}. \qquad ([1])$$

For b getting large one might proceed as sketched for $\delta_1(x)$ in the following lines:

Compare $\delta_1(x)$ with

$$\delta_7(x) = \frac{1}{b!}\sum_{k\neq 0} e^{2k\pi i x}\Gamma(-\chi_k)\chi_k^{b-1}$$

and estimate it as follows:

$$\begin{aligned}|\delta_7(x)| &\leq \frac{2}{b!}\sum_{k\geq 1} e^{-\frac{\pi}{2}\cdot\frac{2k\pi}{L}}\left(\frac{2k\pi}{L}\right)^{b-1}\\ &\sim \frac{2}{b!}\left(\frac{2\pi}{L}\right)^{b-1}\int_0^\infty e^{-\frac{\pi^2}{L}t}\cdot t^{b-1}\,dt\\ &= \frac{L}{b\pi}\cdot\left(\frac{2}{\pi}\right)^b.\end{aligned}$$

4) Even though we have proved Theorem 10 only for values $b \geq 2$, it is easily checked that the result of Theorem 10 coincides for $b=1$ with our previous result from [7].

However, it should be noticed that the proofs of both instances for b are significantly different.

We conclude with a small table of the constants appearing in Theorem 10: We have

$$V_N \sim C_b \cdot N$$

with

$C_1 = 0.845858\ldots$ $\quad$ $C_2 = 0.168054\ldots$

$C_3 = 0.070463\ldots$ $\quad$ $C_4 = 0.040147\ldots$.

References

[1] M. Abramowitz, I. A. Stegun: *Handbook of mathematical functions.* Dover, New York 1970.

[2] B. Berndt: *Modular transformations and generalizations of several formulae of Ramanujan.* Rocky Mountain J. Math. **7** (1977) 147-189.

[3] R. Fagin, J. Nievergelt, N. Pippenger, H. Strong: *Extendible hashing: A fast access method for dynamic files.* ACM TODS **4** (1979) 315-344.

[4] W. Feller: *An Introduction to Probability Theory and Its Applications, vol. 1.* Wiley, New York 1958.

[5] Ph. Flajolet: *On the performance evaluation of extendible hashing and trie searching.* Acta Informatica **20** (1983) 345-369.

[6] Ph. Flajolet, R. Sedgewick: *Digital search trees revisited.* SIAM J. Comput. **15** (1986) 748-767.

[7] P. Kirschenhofer, H. Prodinger: *On some applications of formulae of Ramanujan in the analysis of algorithms.* Preprint (1987), TU Vienna.

[8] P. Kirschenhofer, H. Prodinger, J. Schoissengeier: *Zur Auswertung gewisser Reihen mit Hilfe modularer Funktionen.* Lecture Notes in Math. **1262**, 108-110, Springer, Berlin 1987.

[9] D. E. Knuth: *The Art of Computer Programming, Vol. 3.* Addison Wesley, Reading Mass. 1973.

[10] N. E. Nörlund: *Vorlesungen über Differenzenrechnung.* Chelsea, New York 1954.

[11] S. Ramanujan: *Notebooks of Srinivasa Ramanujan (2 volumes).* Tata Institute of Fundamental Research, Bombay 1957.

[12] J. Riordan: *Combinatorial Identities.* Wiley, New York 1968.

[13] W. Szpankowski: *Some results on v-ary asymmetric tries.* Journal of Algorithms **9** (1988), in press.

[14] W. Szpankowski: *On the analysis of the average height of a digital trie: Another approach.* Preprint (1986), Purdue University.

Address of the authors

Peter Kirschenhofer, Helmut Prodinger
Institut für Algebra und Diskrete Mathematik
Technische Universität Wien
Wiedner Hauptstraße 8-10
A-1040 Wien
Österreich