Approximate Counting via the Poisson-Laplace-Mellin Method

Michael Fuchs[1†], Chung-Kuei Lee[1†] and Helmut Prodinger[2‡]

[1] *Department of Applied Mathematics, National Chiao Tung University, Hsinchu, 300, Taiwan*
[2] *Department of Mathematics, University of Stellenbosch, Stellenbosch, 7602, South Africa*

Approximate counting is an algorithm that provides a count of a huge number of objects within an error tolerance. The first detailed analysis of this algorithm was given by Flajolet. In this paper, we propose a new analysis via the Poisson-Laplace-Mellin approach, a method devised for analyzing shape parameters of digital search trees in a recent paper of Hwang et al. Our approach yields a different and more compact expression for the periodic function from the asymptotic expansion of the variance. We show directly that our expression coincides with the one obtained by Flajolet. Moreover, we apply our method to variations of approximate counting, too.

Keywords: approximate counting, digital search tree, JS-admissibility, Laplace transform, Mellin transform

1 Introduction and Results

Approximate counting, an algorithm proposed by Morris [22] in 1978, is used for counting within a certain error tolerance a huge amount of objects with very limited space. The algorithm has found many applications such as in the analysis of the Webgraph, monitoring network traffic, finding patterns in protein and DNA sequencing, computing frequency moments of data streams, data storage in flash memory, and many variants and improvements have been proposed; see Csűrös [7], Mitchell and Day [21], Gronemeier and Sauerhoff [12], Aspnes and Censor [2], Cichoń and Macyna [4] and references therein.

Here, we are going to revisit the analysis of the classical algorithm which is described as follows: a counter C_n is maintained with initial value $C_0 = 0$. After "counting n objects", a random decision based only on the current content of the counter determines whether or not the counter should be increased when "counting the $(n+1)$-st object". More precisely, the counter obeys the following rule

$$C_{n+1} = \begin{cases} C_n + 1, & \text{with probability } q^{C_n}; \\ C_n, & \text{with probability } 1 - q^{C_n}, \end{cases} \tag{1}$$

where $0 < q < 1$ is fixed. Hence, $(C_n)_{n\geq 0}$ is a Markov chain describing a pure birth process. The same chain was also encountered in a couple of other problems: width of greedy decomposition of random

[†] Partially supported by National Science Council of ROC under the grant NSC-99-2115-M-009-007-MY2.
[‡] Partially supported by an incentive grant from the NRF (South Africa).

acyclic digraphs into node-disjoint paths (see Simon [30]), size of greedy independent set and greedy clique in random graphs (see Simon [30]) and length of the leftmost path in digital search trees (see below).

We mention in passing that many variants of the above Markov chain have been investigated as well; e.g. see Crippa and Simon [6], Louchard and Prodinger [20], Bertoin, Biane and Yor [3] and Guillemin, Robert and Zwart [13]. Applications range from Computer Science over Particle Physics to Molecular Biology; see the detailed discussion in [6].

As for the classical chain C_n, the first detailed analysis was given by Flajolet in [8] who used Mellin transform (see also Prodinger [24] for a similar analysis). Other approaches have been given by Kirschenhofer and Prodinger [17] via Rice method, Prodinger [25] via Euler transform, Louchard and Prodinger [19] via analysis of extreme value distributions, Rosenkrantz [29] via martingale theory and Robert [28] via probabilistic tools. In this paper, we are going to propose a new approach which will be based on the connection with digital search trees and the new method (nicknamed Poisson-Laplace-Mellin method) for analyzing shape parameters in digital search trees proposed in Hwang, Fuchs and Zacharovas [14].

We next explain the connection between approximate counting and digital search trees (DSTs). Therefore, we start with the definition of DSTs which are a fundamental data structure in computer science and were proposed by Coffman and Eve in [5]. Consider a set of n keys which are infinite 0-1 strings. The digital search tree is built from these n keys as follows: the first key is placed in the root; all other keys are directed to the left or right subtree according to whether the first bit is 0 or 1, respectively; finally, the first bits are removed and the subtrees are built recursively according to the same principle; see Figure 1 for an example.

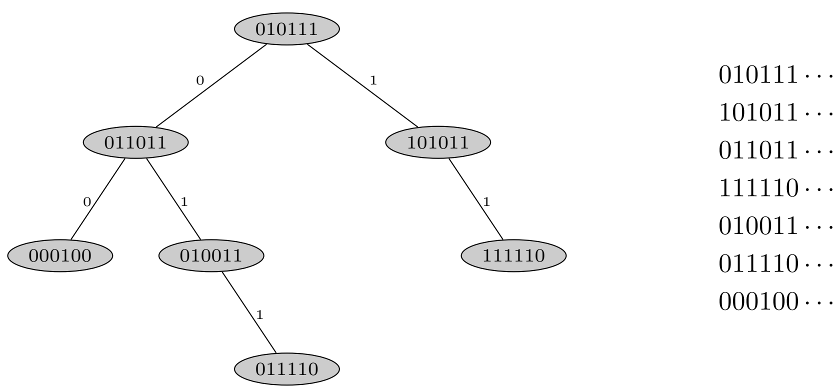

Fig. 1: A DST built from 7 keys with length of leftmost path equal to 3.

Shape parameters in random DSTs have been analyzed in many papers; see [14] and references therein. The standard random model used in such an analysis is the *Bernoulli model*. Here, the bits of the keys are assumed to be i.i.d. Bernoulli random variables with the probability of 0 being q. Shape parameters of random DSTs of size n then become random variables. One example of such a shape parameter is the length of the leftmost path which is defined as the number of vertices on the leftmost path from the root to the leftmost leaf. We denote this length in a random digital search tree of size n by X_n. Obviously, X_n

satisfies the following distributional recurrence

$$X_{n+1} \stackrel{d}{=} X_{B_n} + 1, \qquad (n \geq 0) \tag{2}$$

with $X_0 = 0$ and $B_n \stackrel{d}{=} \text{Binom}(n, q)$. This recurrence just reflects the trivial fact that X_n can be computed by starting from the root (which counts as 1) and then moving on to the left subtree (which has size B_n) where the same procedure is repeated. Now, a moment's reflection reveals that C_n is related to X_n as

$$C_n \stackrel{d}{=} X_n.$$

This relation will be the starting point of our analysis. We will use it to derive asymptotic expansions for mean and variance of C_n.

Before stating our result, we explain what is known about C_n. Flajolet in [8] showed that, as $n \to \infty$,

$$\mathbb{E}(C_n) \sim \log_{1/q} n + F_C(\log_{1/q} n),$$

where $F_C(z) = \sum_k f_k e^{2k\pi iz}$ is a 1-periodic function with Fourier coefficients

$$f_0 = \frac{\gamma}{L} + \frac{1}{2} - \alpha, \qquad f_k = -\frac{\Gamma(-\chi_k)}{L} \quad (k \neq 0).$$

Here, γ is Euler's constant, $\alpha = \sum_{l\geq 1} q^l/(1-q^l)$, $L = \log(1/q)$ and $\chi_k = 2k\pi i/L$. As for the variance, he showed that, as $n \to \infty$,

$$\text{Var}(C_n) \sim G_C(\log_{1/q} n),$$

where $G_C(z) = \sum_k g_k e^{2k\pi iz}$ is again a 1-periodic function with computable Fourier coefficients. Moreover, he gave the following expression for the average value of $G_C(z)$

$$g_0 = \frac{\pi^2}{6L^2} - \alpha - \beta + \frac{1}{12} - \frac{1}{L}\sum_{l\geq 1} \frac{1}{l\sinh(2l\pi^2/L)},$$

where $\beta = \sum_{l\geq 1} q^{2l}/(1-q^l)^2$.

In the next section, we will use the above connection to DSTs to re-derive these results. Our approach will in particular yield a different and more compact expression for all Fourier coefficients of $G_C(z)$.

Theorem 1 *For the variance of approximate counting, we have, as $n \to \infty$,*

$$\text{Var}(C_n) \sim G_C(\log_{1/q} n),$$

where $G_C(z) = \sum_k g_k e^{2k\pi iz}$ is a 1-periodic function with

$$g_k = \frac{Q_\infty}{L\Gamma(1+\chi_k)} \sum_{h,l,j\geq 0} \frac{(-1)^j q^{h+l+\binom{j+1}{2}}}{Q_h Q_l Q_j} \varphi(\chi_k, q^{h+j} + q^{l+j}).$$

Here, $Q_j = \prod_{l=1}^{j}(1-q^l)$, $Q_\infty = \lim_{j\to\infty} Q_j$ and

$$\varphi(\chi; x) := \begin{cases} \pi(x^\chi - 1)/(\sin(\pi\chi)(x-1)), & \text{if } x \neq 1, \\ \pi\chi/\sin(\pi\chi), & \text{if } x = 1. \end{cases}$$

Comparing with the above result, we obtain the following identity for which we will provide a direct proof in Section 3.

Corollary 1 *We have,*

$$\frac{Q_\infty}{L}\sum_{h,l,j\geq 0}\frac{(-1)^j q^{h+l+\binom{j+1}{2}}}{Q_hQ_lQ_j}\psi(q^{h+j}+q^{l+j})=\frac{\pi^2}{6L^2}-\alpha-\beta+\frac{1}{12}-\frac{1}{L}\sum_{l\geq 1}\frac{1}{l\sinh(2l\pi^2/L)},$$

where

$$\psi(x):=\begin{cases}\log x/(x-1), & \textit{if } x\neq 1;\\ 1, & \textit{if } x=1.\end{cases}$$

Finally, in Section 4, we will discuss extensions and variations of approximate counting. One such extension was proposed in [4] where instead of one counter m counters $C_n^{(1)},\dots,C_n^{(m)}$ were used (m fixed). Then, when "counting the $(n+1)$-st object", one of the counters is chosen uniformly at random and increased according to the stochastic rule (1). In Prodinger [26], mean and variance of $D_n := C_n^{(1)}+\cdots+C_n^{(m)}$ were derived. We will show that our approach greatly simplifies the analysis since the case of m counters can be reduced to the case of one counter.

Theorem 2 *For approximate counting with m counters, we have, as $n\to\infty$,*

$$\mathbb{E}(D_n)\sim m\log_{1/q}(n/m)+mF_C(\log_{1/q}(n/m)),$$
$$\mathrm{Var}(D_n)\sim mG_C(\log_{1/q}(n/m)),$$

where $F_C(z)$ and $G_C(z)$ are the periodic functions above.

Moreover, again in Section 4, we will show that similar simplifications can also be achieved for shape parameters in m-DSTs trees recently introduced by Prodinger in [27].

2 Analysis of Approximate Counting

Here, we are going to prove Theorem 1. Therefore, we will apply the Poisson-Laplace-Mellin method from [14]. We will first summarize the main steps of this method; for a more detailed discussion together with comparisons with other approaches, the reader is referred to [14] (in particular, see Figure 7 on page 131 in [14] which gives a comparison of the Poisson-Laplace-Mellin approach with a closely related approach of Flajolet and Richmond [10]). The main steps of our method are as follows.

- We first introduce poissonized mean and variance of X_n (or equivalently C_n) and show that they satisfy differential-functional equations of the same type;
- we use Jacquet and Szpankowski's theory of depoissonization [16] to show that it suffices to prove our claimed result for poissonized mean and variance;
- in order to get rid of the differential operator, we use Laplace transform which then only satisfies a functional equation;
- we use a normalization factor to simplify the functional equation for the Laplace transform;

- applying Mellin transform will allow us to solve the normalized functional equation for the Laplace transform (for an excellent survey on the Mellin transform see Flajolet, Gourdon and Dumas [9]);
- finally, we first use inverse Mellin transform and then inverse Laplace transform to obtain asymptotic expansions for poissonized mean and variance.

After completing these steps, which will work in a similar fashion for both mean and variance, the compact form of the Fourier coefficients for the variance are obtained by some straightforward simplifications.

Poissonization and depoissonization. Our starting point is (2). First, denote the *Poisson generating function* of $\mathbb{E}(e^{X_n y})$ by

$$\tilde{P}(y,z) = e^{-z} \sum_{n\geq 0} \mathbb{E}(e^{X_n y}) \frac{z^n}{n!}.$$

Then, from (2), we obtain

$$\tilde{P}(y,z) + \frac{\partial}{\partial z}\tilde{P}(y,z) = e^{y}\tilde{P}(y,qz)$$

with $\tilde{P}(y,0) = \exp(-z)$.

From this, by differentiation with respect to y and setting $y = 0$, we obtain for the Poisson generating functions of the first and second moment of X_n (denoted by $\tilde{f}_1(z)$ and $\tilde{f}_2(z)$, respectively)

$$\begin{aligned} \tilde{f}_1(z) + \tilde{f}_1'(z) &= \tilde{f}_1(qz) + 1, \\ \tilde{f}_2(z) + \tilde{f}_2'(z) &= \tilde{f}_2(qz) + 2\tilde{f}_1(qz) + 1, \end{aligned} \tag{3}$$

with $\tilde{f}_1(0) = \tilde{f}_2(0) = 0$. Moreover, define the *poissonized variance* as $\tilde{V}(z) := \tilde{f}_2(z) - \tilde{f}_1^2(z)$. Then, the above two relations in turn yield

$$\tilde{V}(z) + \tilde{V}'(z) = \tilde{V}(qz) + \tilde{f}_1'(z)^2 \tag{4}$$

with $\tilde{V}(0) = 0$.

We first show that in order to derive asymptotics of $\mathbb{E}(X_n)$ and $\mathrm{Var}(X_n)$, it suffices to analyze $\tilde{f}_1(z)$ and $\tilde{V}(z)$ as $z \to \infty$, respectively. Therefore, we use the theory of analytic depoissonization due to Jacquet and Szpankowski. Recall that a function $\tilde{f}(z)$ is called JS-admissible if:

(I) There exist $\alpha, \beta \in \mathbb{R}$ such that uniformly for $|\arg(z)| \leq \epsilon$

$$\tilde{f}(z) = \mathcal{O}\left(|z|^{\alpha}(\log_+ |z|)^{\beta}\right),$$

where $\log_+ x = \log(1+x)$.

(O) Uniformly for $\epsilon \leq \arg(z) \leq \pi$,

$$f(z) := e^{z}\tilde{f}(z) = \mathcal{O}\left(e^{(1-\epsilon)|z|}\right).$$

(Here and throughout the work, ϵ denotes a small constant whose value might be different from one occurrence to another).

JS-admissibility of given functions is easily checked due to closure properties; see Lemma 2.3 in [14]. Moreover, JS-admissibility of functions which are given by differential-functional equations of the type above is also easily checked due to the following result.

Proposition 1 *Let $\tilde{f}(z)$ and $\tilde{g}(z)$ be entire functions with*

$$\tilde{f}(z)+\tilde{f}'(z)=\tilde{f}(qz)+\tilde{g}(z),$$

where $\tilde{f}(0)=0$. Then,

$$\tilde{f}(z) \text{ is JS-admissible} \quad \Longleftrightarrow \quad \tilde{g}(z) \text{ is JS-admissible}.$$

Proof: Similar to the proof of Proposition 2.4 in [14]. □

Consequently, $\tilde{f}_1(z)$ and $\tilde{f}_2(z)$ are both JS-admissible. Depoissonization (see Proposition 2.2 in [14] and the discussion in the introduction) then yields, as $n\to\infty$,

$$\mathbb{E}(X_n)\sim\tilde{f}_1(n) \qquad \text{and} \qquad \mathrm{Var}(X_n)\sim\tilde{V}(n).$$

Thus, we only have to find asymptotics of $\tilde{f}_1(z)$ and $\tilde{V}(z)$.

Analysis of the Mean. Here, we analyze the mean, where we start from (3). Since $\tilde{f}_1(z)$ is JS-admissible, we may apply Laplace transform to get rid of the differential operator. This yields

$$(s+1)\mathscr{L}[\tilde{f}_1;s]=\frac{1}{q}\mathscr{L}[\tilde{f}_1;s/q]+1/s. \tag{5}$$

Next, we derive an exact expression for the mean. Therefore, we iterate the above functional equation and obtain

$$\mathscr{L}[\tilde{f}_1;s]=\frac{1}{s}\sum_{j\geq 0}\frac{1}{(s+1)(q^{-1}s+1)\cdots(q^{-j}s+1)}.$$

Here, we have used that $\mathscr{L}[\tilde{f}_1;s]=\mathcal{O}(1/s^2)$ as $s\to\infty$. Next, by partial fraction expansion

$$\frac{1}{(s+1)(q^{-1}s+1)\cdots(q^{-j}s+1)}=\sum_{0\leq l\leq j}\frac{(-1)^{j-l}q^{\binom{j-l+1}{2}}}{(q^{-l}s+1)Q_lQ_{j-l}}.$$

Plugging this into the expression above yields

$$\begin{aligned}\mathscr{L}[\tilde{f}_1;s]&=\frac{1}{s}\sum_{j\geq 0}\sum_{0\leq l\leq j}\frac{(-1)^{j-l}q^{\binom{j-l+1}{2}}}{(q^{-l}s+1)Q_lQ_{j-l}}\\&=\frac{1}{s}\sum_{l\geq 0}\frac{1}{Q_l(q^{-l}s+1)}\sum_{j\geq 0}\frac{(-1)^jq^{\binom{j+1}{2}}}{Q_j}\\&=\frac{Q_\infty}{s}\sum_{l\geq 0}\frac{1}{Q_l(q^{-l}s+1)},\end{aligned}$$

where Q_j and Q_∞ have been defined in the introduction and we used the well-known identity

$$\sum_{j\geq 0}\frac{(-1)^jq^{\binom{j+1}{2}}}{Q_j}=Q_\infty.$$

Now, by inverse Laplace transform,

$$\tilde{f}_1(z) = Q_\infty \sum_{l\geq 0} \frac{1}{Q_l}(1 - e^{-q^l z})$$

and hence

$$\mathbb{E}(X_n) = Q_\infty \sum_{l\geq 0} \frac{1}{Q_l}(1 - (1 - q^l)^n).$$

We record this result for future reference.

Proposition 2 *We have*

$$\mathbb{E}(X_n) = Q_\infty \sum_{l\geq 0} \frac{1}{Q_l}(1 - (1 - q^l)^n).$$

Next, we derive an asymptotic expansion. This will be done by using the Mellin transform. Therefore, set $\bar{\mathscr{L}}[\tilde{f}_1; s] = \mathscr{L}[\tilde{f}_1; s]/Q(-s)$, where

$$Q(-s) = \prod_{i=1}^{\infty} \left(1 + q^i s\right).$$

Then, by dividing (5) by $Q(-s/q)$,

$$\bar{\mathscr{L}}[\tilde{f}_1; s] = \frac{1}{q}\bar{\mathscr{L}}[\tilde{f}_1; s/q] + \frac{1}{sQ(-s/q)}.$$

Now, from the fact that $\tilde{f}_1(z)$ is JS-admissible and well-known growth properties of $Q(-s/q)$ (see page 127 in [14]), we obtain suitable polynomial bounds for $\bar{\mathscr{L}}[\tilde{f}_1; s]$ as s tends both to zero and ∞. This ensures the existence of the Mellin transform of $\bar{\mathscr{L}}[\tilde{f}_1; s]$ in a non-trivial strip. Thus, we may apply Mellin transform and obtain

$$\mathscr{M}[\bar{\mathscr{L}}; \omega] = \frac{M_1(\omega)}{1 - q^{\omega - 1}}, \qquad (\Re(\omega) > 1),$$

where

$$M_1(\omega) = \int_0^\infty \frac{s^{\omega-2}}{Q(-s/q)} \mathrm{d}s = \frac{Q(q^{1-\omega})}{Q_\infty}\Gamma(\omega + 1)\Gamma(-\omega).$$

Note that the latter function is meromorphic for $\Re(\omega) > 0$ with a simple pole at $\omega = 1$. Moreover, due to rapid decay of the Γ function along vertical lines, we have $M_1(c + it) = O(e^{-\pi|t|})$ for $c > 0$ and $|t|$ large. Hence, inverse Mellin transform implies for $|\arg(s)| \leq \pi - \epsilon$ and $|s| \to 0$,

$$\bar{\mathscr{L}}[\tilde{f}_1; s] \sim \frac{1}{s}\log_{1/q}\frac{1}{s} + \frac{1}{s}\left(\frac{1}{2} - \alpha + \frac{1}{L}\sum_{k\neq 0} M_1(1 + \chi_k)s^{-\chi_k}\right),$$

where notations are as in the introduction. Since $Q(-s/q) = 1 + \mathcal{O}(s)$ for $|\arg(s)| \leq \pi - \epsilon$ and $|s| \to 0$, the above in turn yields

$$\mathscr{L}[\tilde{f}_1; s] \sim \frac{1}{s}\log_{1/q}\frac{1}{s} + \frac{1}{s}\left(\frac{1}{2} - \alpha + \frac{1}{L}\sum_{k\neq 0} M_1(1 + \chi_k)s^{-\chi_k}\right).$$

By Proposition 2.6 of [14], we may apply inverse Laplace transform and obtain for $|\arg(z)| \leq \frac{\pi}{2} - \epsilon$ and $|z| \to \infty$,

$$\begin{aligned}\tilde{f}_1(z) &\sim \log_{1/q} z + \frac{\gamma}{L} + \frac{1}{2} - \alpha + \frac{1}{L}\sum_{k\neq 0} \frac{M_1(1+\chi_k)}{\Gamma(1+\chi_k)} z^{\chi_k}\\ &= \log_{1/q} z + \frac{\gamma}{L} + \frac{1}{2} - \alpha - \frac{1}{L}\sum_{k\neq 0} \Gamma(-\chi_k) z^{\chi_k}.\end{aligned}$$

The same asymptotic expansion also holds for $\mathbb{E}(X_n)$ by depoissonization.

Analysis of the Variance. For an asymptotic expansion of the variance, we start from (4) and proceed by the same method as above. First note that due to the above analysis and Ritt's Theorem (Theorem 4.2 of [23]), we have uniformly for $|\arg(z)| \leq \frac{\pi}{2} - \epsilon$

$$\tilde{f}_1'(z)^2 = \begin{cases} \mathcal{O}(1), & \text{if } |z| \to 0;\\ \mathcal{O}(|z|^{-2}), & \text{if } |z| \to \infty.\end{cases} \tag{6}$$

This in turn yields the following rough bounds for $\tilde{V}(z)$

$$\tilde{V}(z) = \begin{cases} \mathcal{O}(z), & \text{if } z \to 0+;\\ \mathcal{O}(z^{\epsilon}), & \text{if } z \to \infty.\end{cases} \tag{7}$$

Therefore, we may apply Laplace transform. Hence,

$$(s+1)\mathscr{L}[\tilde{V}; s] = \frac{1}{q}\mathscr{L}[\tilde{V}; s/q] + \tilde{g}(s),$$

where $\tilde{g}(s) = \mathscr{L}[\tilde{f}_1'^2; s]$. Next, set $\bar{\mathscr{L}}[\tilde{V}; s] = \mathscr{L}[\tilde{V}; s]/Q(-s)$. Dividing by $Q(-s/q)$ yields

$$\bar{\mathscr{L}}[\tilde{V}; s] = \frac{1}{q}\bar{\mathscr{L}}[\tilde{V}; s/q] + \tilde{g}(s)/Q(-s/q).$$

Now, from (7) and growth properties of $Q(-s)$, we have

$$\bar{\mathscr{L}}[\tilde{V}; s] = \begin{cases} \mathcal{O}(1/s^{1+\epsilon}), & \text{if } s \to 0+;\\ \mathcal{O}(1/s^{b}), & \text{if } s \to \infty,\end{cases}$$

where $b > 0$ is an arbitrary large constant. Hence, the Mellin transform of $\bar{\mathscr{L}}[\tilde{V}; s]$ exists for $\Re(\omega) > 1$. Consequently,

$$\mathscr{M}[\bar{\mathscr{L}}; \omega] = \frac{M_2(\omega)}{1 - q^{\omega-1}}, \qquad (\Re(\omega) > 1),$$

where

$$M_2(\omega) = \int_0^\infty \frac{s^{\omega-1}}{Q(-s/q)} \int_0^\infty e^{-zs} \tilde{f}_1'(z)^2 \mathrm{d}z \mathrm{d}s.$$

Next, we have to study properties of $M_2(\omega)$. Therefore, observe that from (6) and growth properties of $Q(-s)$, we have uniformly for $|\arg(s)| \le \pi - \epsilon$

$$\frac{\tilde{g}(s)}{Q(-s/q)} = \begin{cases} \mathcal{O}(s), & \text{if } s \to 0+; \\ \mathcal{O}(1/s^b), & \text{if } s \to \infty, \end{cases}$$

where $b > 0$ is again an arbitrary large constant. Hence, $M_2(\omega)$ is analytic for $\Re(\omega) > -1$. Moreover, from Proposition 5 in [9], we have $M_2(c + it) = O(e^{-(\pi-\epsilon)|t|})$ for $c > -1$ and $|t|$ large. Consequently, we can proceed as for the mean and obtain as $z \to \infty$,

$$\tilde{V}(z) \sim \frac{1}{L} \sum_{k \in \mathbb{Z}} \frac{M_2(1+\chi_k)}{\Gamma(1+\chi_k)} z^{\chi_k}.$$

The same then holds for $\mathrm{Var}(X_n)$ as well by depoissonization.

We conclude by simplifying $M_2(1 + \chi_k)$. For that, we use

$$\tilde{f}_1'(z) = Q_\infty \sum_{l \ge 0} \frac{1}{Q_l q^{-l}} e^{-zq^l}$$

and

$$\frac{1}{Q(-s/q)} = \frac{1}{Q_\infty} \sum_{j \ge 0} \frac{(-1)^j q^{\binom{j}{2}}}{Q_j (s + q^{-j})}.$$

Plugging this into the above integral yields

$$M_2(1+\chi_k) = Q_\infty \sum_{h,l,j \ge 0} \frac{(-1)^j q^{\binom{j}{2}}}{Q_h Q_l Q_j q^{-(l+h)}} \int_0^\infty \frac{s^{\chi_k}}{(s+q^{-j})(s+q^h+q^l)} \mathrm{d}s.$$

Denote by

$$\varphi(\chi; x) := \begin{cases} \pi(x^\chi - 1)/(\sin(\pi\chi)(x-1)), & \text{if } x \ne 1, \\ \pi\chi/\sin(\pi\chi), & \text{if } x = 1. \end{cases}$$

Then,

$$M_2(1+\chi_k) = Q_\infty \sum_{h,l,j \ge 0} \frac{(-1)^j q^{\binom{j+1}{2}}}{Q_h Q_l Q_j q^{-(l+h)}} \varphi(\chi_k, q^{h+j} + q^{l+j}).$$

3 Average Value of $G_C(z)$

Here, we are going to prove Corollary 1. We will use the abbreviation $Q = 1/q$. Furthermore, in order to be closer to the q-hypergeometric world and the identities of relevance (see the book of Andrews-Askey-Roy [1]), we use the classical notation $(q)_n$ instead of Q_n.

In [17], the alternative expression

$$\mathcal{P} := \frac{\log 2}{L} - \alpha - \beta + \frac{2}{L}\tau \quad \text{with} \quad \tau := \sum_{k \ge 1} \frac{(-1)^{k-1}}{k(Q^k - 1)}$$

was given for the constant in the variance, and we will show now the equality of this and

$$\mathcal{F} := \frac{(q)_\infty}{L} \sum_{j,l,h\ge 0} \frac{(-1)^j q^{\binom{j+1}{2}+l+h}}{(q)_j (q)_l (q)_h} \frac{\log(q^{h+j}+q^{l+j})}{q^{h+j}+q^{l+j}-1}.$$

In this expression, we have replaced the ψ function by what it is; in some exceptional cases a limit has to be taken.

We use the symmetry in l and h and set $l = h + d$ with $d \ge 0$; then we have to take the sum over $h, d \ge 0$ twice, and subtract the sum for $h \ge 0$ and $d = 0$. Therefore

$$\mathcal{F} = 2 \sum_{j,h,d\ge 0} \cdots - \sum_{j,h\ge 0,\ d=0} \cdots .$$

We think about d as being fixed, set $h = N - j$ and fix N as well: This leads to

$$\frac{(q)_\infty[-LN + \log(1+q^d)]}{L} \sum_{j=0}^{N} \frac{(-1)^j q^{\binom{j+1}{2}+2(N-j)+d}}{(q)_j (q)_{N-j+d} (q)_{N-j}} \frac{1}{q^N + q^{N+d} - 1}.$$

By automatic summation (q-Zeilberger's algorithm) we have the simplification

$$\sum_{j=0}^{N} \frac{(-1)^j q^{\binom{j+1}{2}+2(N-j)+d}}{(q)_j (q)_{N-j} (q)_{N+d-j}} \frac{1}{q^N + q^{N+d} - 1} = \frac{q^{N^2+dN}}{(q)_N (q)_{N+d}}.$$

Consequently,

$$\mathcal{F} = 2(q)_\infty \sum_{N,d\ge 0} \frac{-NL + \log(1+q^d)}{L} \frac{q^{N^2+dN}}{(q)_N (q)_{N+d}} + (q)_\infty \sum_{N\ge 0} \frac{NL - \log 2}{L} \frac{q^{N^2}}{(q)_N (q)_N}.$$

We will soon show that

$$(q)_\infty \sum_{N,d\ge 0} \frac{\log(1+q^d)}{L} \frac{q^{N^2+dN}}{(q)_N (q)_{N+d}} = \frac{\tau}{L} + \frac{\log 2}{L}, \tag{8}$$

which leaves us to prove that

$$2(q)_\infty \sum_{N,d\ge 0} \frac{N q^{N^2+dN}}{(q)_N (q)_{N+d}} - (q)_\infty \sum_{N\ge 0} \frac{(N - \frac{\log 2}{L}) q^{N^2}}{(q)_N (q)_N} = \frac{\log 2}{L} + \alpha + \beta.$$

Because of the identity [1, p. 567]

$$\sum_{N\ge 0} \frac{q^{N^2}}{(q)_N^2} = \frac{1}{(q)_\infty},$$

this leaves us with

$$2(q)_\infty \sum_{N,d\ge 0} \frac{N q^{N^2+dN}}{(q)_N (q)_{N+d}} - (q)_\infty \sum_{N\ge 0} \frac{N q^{N^2}}{(q)_N (q)_N} = \alpha + \beta. \tag{9}$$

Now, expanding $\log(1+q^d)$, (8) is proved once we can prove that

$$(q)_\infty \sum_{N\ge 0,\ d\ge 1} \frac{q^{N^2+dN+dk}}{(q)_N (q)_{N+d}} = \frac{1}{Q^k-1}.$$

But this follows from

$$\sum_{N\ge 0,\ d\ge 1} \frac{1}{(q)_d} \frac{q^{N^2+dN+dk}}{(q)_N (q^{d+1})_N} = \sum_{d\ge 1} \frac{q^{dk}}{(q)_d} \frac{1}{(q^{d+1})_\infty} = \frac{1}{(q)_\infty} \frac{1}{Q^k-1}.$$

We have used here the classical identity (Cauchy's identity) [1, p. 568]

$$\sum_{n\ge 0} \frac{x^n q^{n^2}}{(q)_n (xq)_n} = \frac{1}{(xq)_\infty}.$$

In order to prove (9), we will show that

$$-(q)_\infty \sum_{N\ge 0} \frac{N q^{N^2}}{(q)_N (q)_N} = \sum_{r\ge 1} \frac{(-1)^r q^{\binom{r+1}{2}}}{1-q^r}, \tag{10}$$

$$(q)_\infty \sum_{N,d\ge 0} \frac{N q^{N^2+dN}}{(q)_N (q)_{N+d}} = -\sum_{r\ge 1} \frac{(-1)^r q^{\binom{r+1}{2}}}{(1-q^r)^2}. \tag{11}$$

Since in [18, (3.16)], it was proved that

$$\sum_{r\ge 1} \frac{(-1)^r q^{\binom{r+1}{2}}}{1-q^r} - 2\sum_{r\ge 1} \frac{(-1)^r q^{\binom{r+1}{2}}}{(1-q^r)^2} = \alpha + \beta,$$

that would finish the proof. We start from

$$\sum_{n\ge 0} \frac{x^n q^{n^2}}{(q)_n (xq)_n} = \sum_{n\ge 0} \frac{x^n q^{n^2}}{(q)_n (xq)_\infty} (xq^{n+1})_\infty = \frac{1}{(xq)_\infty},$$

which is equivalent to

$$\sum_{n\ge 0} \frac{x^n q^{n^2}}{(q)_n} \sum_{k\ge 0} \frac{(-1)^k q^{\binom{k}{2}} x^k q^{(n+1)k}}{(q)_k} = 1.$$

Now differentiate this, and then set $x=1$:

$$\sum_{n\ge 0} \frac{n q^{n^2}}{(q)_n^2} + \frac{1}{(q)_\infty} \sum_{n\ge 0} \frac{q^{n^2}}{(q)_n} \sum_{k\ge 0} \frac{(-1)^k q^{\binom{k}{2}} k q^{(n+1)k}}{(q)_k} = 0.$$

Rearranging,

$$\sum_{n\ge 0} \frac{n q^{n^2}}{(q)_n^2} + \frac{1}{(q)_\infty} \sum_{N\ge 1} \sum_{n=0}^{N} \frac{q^{n^2}}{(q)_n} \frac{(-1)^{N-n} q^{\binom{N-n}{2}} (N-n) q^{(n+1)(N-n)}}{(q)_{N-n}} = 0,$$

and again by a mechanical proof,

$$\sum_{n\ge0}\frac{nq^{n^2}}{(q)_n^2}+\frac{1}{(q)_\infty}\sum_{N\ge1}\frac{(-1)^Nq^{\binom{N+1}{2}}}{1-q^N}=0.$$

This is (10). Now let us plug in $x=q^d$ after differentiation (instead of $x=1$, as before):

$$\sum_{n\ge0}\frac{nq^{d(n-1)}q^{n^2}}{(q)_n}\sum_{k\ge0}\frac{(-1)^kq^{\binom{k}{2}}q^{kd}q^{(n+1)k}}{(q)_k}+\sum_{n\ge0}\frac{q^{dn}q^{n^2}}{(q)_n}\sum_{k\ge0}\frac{(-1)^kq^{\binom{k}{2}}kq^{(k-1)d}q^{(n+1)k}}{(q)_k}=0.$$

After some simplifications (using Rothe's identity [1, p. 490]), this leads to

$$(q)_\infty\sum_{n\ge0}\frac{nq^{dn}q^{n^2}}{(q)_n(q)_{n+d}}+\sum_{N\ge1}\frac{(-1)^Nq^{\binom{N+1}{2}+dN}}{1-q^N}=0.$$

Now sum this on d:

$$(q)_\infty\sum_{n,d\ge0}\frac{nq^{dn}q^{n^2}}{(q)_n(q)_{n+d}}+\sum_{N\ge1}\frac{(-1)^Nq^{\binom{N+1}{2}}}{(1-q^N)^2}=0,$$

which is (11).

Remark. A direct proof that the Fourier coefficients, as computed here, agree with the ones given in [8], can be done in the same style.

4 Approximate Counting with m Counters and m-DSTs

Approximate Counting with m Counters. I. Here, we consider approximate counting with m counters as discussed in the introduction. Recall that D_n denoted the sum of the counters after "counting n objects". Then, we have

$$D_n\stackrel{d}{=}C^{(1)}_{I_1}+\cdots+C^{(m)}_{I_m},$$

where $C^{(1)}_n,\dots,C^{(m)}_n$ are independent copies of C_n and $I_1,\dots,I_m$ are random variables with joint distribution

$$P(I_1=n_1,\dots,I_m=n_m)=\binom{n}{n_1,\dots,n_m}\frac{1}{m^n}$$

with $n_1+\cdots+n_m=n$. Now, set

$$\tilde Q(y,z)=e^{-z}\sum_{n\ge0}\mathbb{E}(e^{D_ny})\frac{z^n}{n!},\qquad \tilde P(y,z)=e^{-z}\sum_{n\ge0}\mathbb{E}(e^{C_ny})\frac{z^n}{n!}.$$

Then, by a straight-forward computation

$$\tilde Q(y,z)=\tilde P(y,z/m)^m.$$

From this, we can derive the following relations for the Poisson generating functions of the first and second moment of D_n and C_n (denoted by $\tilde{g}_1(z), \tilde{g}_2(z)$ for the former and as above for the latter)

$$\begin{aligned} \tilde{g}_1(z) &= m\tilde{f}_1(z/m), \\ \tilde{g}_2(z) &= m(m-1)\tilde{f}_1(z/m)^2 + m\tilde{f}_2(z/m). \end{aligned}$$

Moreover, again consider the poissonized variance $\tilde{W}(z) := \tilde{g}_2(z) - \tilde{g}_1(z)^2$. Then,

$$\tilde{W}(z) = m\tilde{V}(z/m).$$

Now, it follows from the closure properties of JS-admissibility (see Lemma 2.3 in [14]) that both $\tilde{g}_1(z)$ and $\tilde{g}_2(z)$ are JS-admissible. Hence, we only have to concentrate on the $\tilde{g}_1(z)$ and $\tilde{W}(z)$ whose asymptotic expansions, due to the above formulas, follow from the case $m = 1$.

m-DSTs. m-DSTs have been introduced in [27]. They are defined as follows: again we start with n keys, but they are now stored in m DSTs. For every key, one of the m DSTs is chosen uniformly and at random and the key is then stored in the chosen tree.

Clearly, the previous analysis also gives the sum of the lengths of the leftmost paths in m-DSTs. Similarly, one can consider other shape parameters in DST and extend them linearly to m-DSTs. Our method above can then be applied to such parameters as well and again the analysis will be reduced to the case $m = 1$.

We give two examples. The first example is the depth of a random node which was discussed in [27]. As a second example, consider the total path length T_n in a random digital search tree of size n which is the sum over all distances of nodes to the root. For this quantity, it was proved for $q = 1/2$ (see Kirschenhofer, Prodinger and Szpankowski [18] and [14]) that, as $n \to \infty$,

$$\mathbb{E}(T_n) \sim n \log_2 n + nF_T(\log_2 n)$$

and

$$\mathrm{Var}(T_n) \sim nG_T(\log_2 n),$$

where $F_T(z)$ and $G_T(z)$ are 1-periodic functions with computable Fourier coefficients (see below for a remark concerning the average value of $G_T(z)$). Similar results are known for the case $q \neq 1/2$ as well; see Jacquet and Szpankowski [15]. Now, denote by U_n the sum of all total path lengths in an m-DST. Then, with the same approach as above, we have the following result.

Theorem 3 *For the total path length in m-DSTs, we have, as $n \to \infty$,*

$$\begin{aligned} \mathbb{E}(U_n) &\sim (n/m)\log_2(n/m) + (n/m)F_T(\log_2(n/m)), \\ \mathrm{Var}(U_n) &\sim (n/m)G_T(\log_2(n/m)), \end{aligned}$$

where $F_T(z)$ and $G_T(z)$ are the periodic functions above.

The variance of the path-length. The constant in

$$\mathrm{Var}(T_n) \sim nG_T(\log_2 n)$$

was given in [14] as

$$\frac{(q)_\infty}{L}\sum_{j,h,l\ge 0}\frac{(-1)^j q^{\binom{j+1}{2}+h+l}}{(q)_j(q)_h(q)_l}\varphi(q^{j+h}+q^{j+l})\quad\text{with}\quad \varphi(x)=\frac{x-1-\log x}{(x-1)^2}.$$

(In some cases, limits have to be taken, and the notation $(q)_n$ is again used for Q_n.) This form is a huge improvement over the form provided in [18]. However, with the methods used earlier, even this form can be further improved, in the sense that no triple sums occur anymore. This is beneficial for the numerical evaluation of this constant. The result is

$$\frac{2\alpha}{L}+\frac{(q)_\infty}{L}\sum_{N\ge 1,\ d\ge 1}\frac{q^{N^2+dN}}{(q)_N(q)_{N+d}}\frac{NL-\log(1+q^d)}{q^N+q^{N+d}-1}$$
$$+2(q)_\infty\sum_{N\ge 2}\frac{q^{N^2}}{(q)_N^2}\frac{N-1}{q^{N-1}-1}-\frac{1}{L}-(q)_\infty+\frac{2}{L}\sum_{n\ge 0}\frac{(-1)^n q^{\binom{n+1}{2}}}{(q)_n}\sum_{k\ge 2}\frac{(-1)^k}{k}\frac{1}{2^{k+n-1}-1}.$$

Note that $q=\frac{1}{2}$ here. Details might appear elsewhere.

Approximate Counting with m Counters. II. Here, we again consider approximate counting with m counters, but this time we label them from 1 to m. Now, we proceed as follows: first, we use the first counter until it will be increased, then we use the second one until it will be increased, etc. until the last counter is increased then we return to the first one and repeat this procedure.

Let again D_n denote the sum of the m counters after "counting n objects". This clearly corresponds to the length of the leftmost path in random digital search trees, where every node can hold up to m keys (here, the length is the sum of all nodes on the leftmost path weighted by the number of keys contained in the nodes). Consequently, $D_n \stackrel{d}{=} X_n$, where X_n satisfies

$$X_{n+m}\stackrel{d}{=}X_{B_n}+m,\qquad (n\ge 0)$$

with $X_i=i, 0\le i\le m-1$. The Poisson-Laplace-Mellin approach can be applied to this sequence as well. We only sketch some details.

First, for the Poisson generating functions of the mean and the poissonized variance (again denoted by $\tilde f_1(z)$ and $\tilde V(z)$, respectively), we have

$$\sum_{i=0}^{m}\binom{m}{i}\tilde f_1^{(i)}(z)=\tilde f_1(qz)+m$$

and

$$\sum_{i=0}^{m}\binom{m}{i}\tilde V^{(i)}(z)=\tilde V(qz)+\tilde g(z),$$

where $\tilde g(z)$ is of the form

$$\tilde g(z)=\left(\sum_{i=0}^{m}\binom{m}{i}\tilde f_1{}^{(i)}(z)\right)^2-\sum_{i=0}^{m}\binom{m}{i}\left(\tilde f_1(z)^2\right)^{(i)}.$$

Applying the Poisson-Laplace-Mellin method then yields asymptotic expansion of mean and variance. We content ourselves with stating the result for the variance.

Theorem 4 *For approximate counting with m-counters, where counters are chosen cyclically, we have, as $n \to \infty$,*

$$\mathrm{Var}(D_n) \sim G_D(\log_{1/q} n),$$

where $G_D(z) = \sum_k g_k e^{2k\pi i z}$ is a 1-periodic function with Fourier coefficients

$$g_k = \frac{1}{L\Gamma(1+\chi_k)} \int_0^\infty \frac{s^{\chi_k}}{Q(-s/q)^m} \left(p(s) + \int_0^\infty e^{-zs} \tilde{g}(z) \mathrm{d}z \right) \mathrm{d}s$$

and

$$p(s) = \frac{(s+1)^m - 1 - ms}{s^2}.$$

Acknowledgements

We thank the anonymous reviewers for helpful comments.

References

[1] G. E. Andrews, R. Askey, and R. Roy. *Special Functions*. Encyclopedia of Mathematics and Its Applications, The University Press, Cambridge, 1999.

[2] J. Aspnes and K. Censor (2010). Approximate shared-memory counting despite a strong adversary, *ACM Trans. Algorithms*, **6**, 23 pages.

[3] J. Bertoin, P. Biane, and M. Yor (2002). Poissonian exponential functionals, q-series, q-integrals, and the moment problem for log-normal distributions, Tech. Rep. PMA-705, Laboratoire de Probabilités et Modèles Aléatoires, Université Paris VI.

[4] J. Chichoń and W. Macyna (2011). Approximate counters for flash memory, In proceedings of the seventeenth IEEE international conference on embedded and real-time computing systems and applications, 185–189.

[5] E. G. Coffman Jr. and J. Eve (1970). File structures using hashing functions, *Commun. ACM*, **13**, 427–432.

[6] D. Crippa and K. Simon (1997). q-distributions and Markov processes, *Discrete Math.*, **170**, 81–98.

[7] M. Csűrös (2010). Approximate counting with a floating-point counter, Computing and Combinatorics, 16th Annual International Conference, Cocoon 2010, Lecture Notes in Computer Science, **6196**, 358–367.

[8] P. Flajolet (1985). Approximate counting: a detailed analysis, *BIT*, **25**, 113–134.

[9] P. Flajolet, X. Gourdon, and P. Dumas (1995). Mellin transforms and asymptotics: harmonic sums, *Theoret. Comput. Sci.*, **144**, 3–58.

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

[11] P. Flajolet and R. Sedgewick (1986). Digital search trees revisited, *SIAM J. Comput.*, **15**, 748–767.

[12] A. Gronemeier and M. Sauerhoff (2009). Applying approximate counting for computing the frequency moments of long data streams, *Theory Comput. Syst.*, **44**, 332–348.

[13] F. Guillemin, P. Robert, and B. Zwart (2004). AIMD algorithms and exponential functionals, *Ann. Appl. Probab.*, **14**, 90–117.

[14] H.-K. Hwang, M. Fuchs, and V. Zacharovas (2010). Asymptotic variance of random symmetric digital search trees, *Discrete Math. Theor. Comput. Sci.*, **12**, 103–166.

[15] P. Jacquet and W. Szpankowski (1995). Asymptotic behavior of the Lempel-Ziv parsing scheme and digital search trees, *Theoret. Comput. Sci.*, **144**, 161–197.

[16] P. Jacquet and W. Szpankowski (1998). Analytical de-Poissonization and its applications, *Theoret. Comput. Sci.*, **201**, 1–62.

[17] P. Kirschenhofer and H. Prodinger (1991). Approximate counting: an alternative approach, *RAIRO Inform. Théor. Appl.*, **25**, 43–48.

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

[19] G. Louchard and H. Prodinger (2006). Asymptotics of the moments of extreme-value related distribution functions, *Algorithmica*, **46**, 431–467.

[20] G. Louchard and H. Prodinger (2008). Generalized approximate counting revisited, *Theoret. Comput. Sci.*, **391**, 109–125.

[21] S. A. Mitchell and D. M. Day (2011). Flexible approximate counting, In 15th International Database Engineering & Applications Symposium, IDEAS 2011, 233–239.

[22] R. Morris (1978). Counting large numbers of events in small registers, *Comm. ACM*, **21**, 840–842.

[23] F. W. J. Olver. *Asymptotics and Special Functions*. Academic Press, 1974.

[24] H. Prodinger (1992). Hypothetic analyses: approximate counting in the style of Knuth, path length in the style of Flajolet, *Theoret. Comput. Sci.*, **100**, 243–251.

[25] H. Prodinger (1994). Approximate counting via Euler transform, *Math. Slovaka*, **44**, 569–574.

[26] H. Prodinger (2011). Digital search trees with m trees: level polynomials and insertion costs, *Discrete Math. Theor. Comput. Sci.*, **13**, 1–8.

[27] H. Prodinger (2012). Approximate counting with m counters: a detailed analysis, *Theoret. Comput. Sci.*, in press.

[28] P. Robert (2005). On the asymptotic behavior of some algorithms, *Random Structures Algorithms*, **27**, 235–250.

[29] W. A. Rosenkrantz (1987). Approximate counting: a martingale approach, *Stochastics*, **20**, 111–120.

[30] K. Simon (1988). Improved algorithm for transitive closure on acyclic digraphs, *Theoret. Comput. Sci.*, **58**, 325–346.