A RESULT IN ORDER STATISTICS RELATED TO PROBABILISTIC COUNTING

Peter Kirschenhofer and Helmut Prodinger

Department of Algebra and Discrete Mathematics
Technical University of Vienna, Austria

Abstract. Considering geometrically distributed random variables the d-maximum of these events is investigated, i.e. the d-th largest element (with repetitions allowed). The quantitative behaviour of expectation and variance is analyzed thoroughly. In particular the asymptotics of the variance for d getting large is established by means of nontrivial techniques from combinatorial analysis and complex variable theory. These results apply to probabilistic counting algorithms, where the cardinalities of large sets are estimated.

Zusammenfassung. Bezüglich geometrisch verteilter zufälliger Veränderlicher wird das d-Maximum solcher Ereignisse studiert, also das d-größte Element, wobei Wiederholungen erlaubt sind. Das quantitative Verhalten von Erwartungswert und Varianz wird ausgiebig analysiert. Insbesondere wird das Verhalten der Varianz für große d mit Hilfe nichttrivialer Techniken aus Kombinatorik und komplexer Analysis untersucht. Diese Resultate haben Anwendungen bei probabilistischen Zählalgorithmen, die zur Schätzung der Kardinalitäten großer Mengen verwendet werden.

1. Introduction

Szpankowski and Rego [14] have considered n independently and identically distributed geometric random variables $X_1, \dots, X_n$. Their parameter of interest was the maximum, $\max\{X_1, \dots, X_n\}$.

The aim of this paper is threefold:

- A slightly more direct approach will be used to this problem; *Szpankowski and Rego* set up recurrences for the expectation and the second moment. These recurrences are solved by *generating function techniques*. This is somewhat superfluous as we can directly compute the probabilities for these *max*–statistics.
- In the main part the following generalization will be considered: Let $d \geq 1$ be an integer. We consider the d–maximum, i.e. the d–th largest element (with repetitions allowed). ($d = 1$ is the maximum.) To be more precise, if there are numbers $x_1, \dots, x_n$ and we sort them in descending order as $y_1 \geq \cdots \geq y_n$, then the d–maximum $\max_d\{x_1, \dots, x_n\}$ is y_d. It will turn out, that, by choosing d to be large, the variance can be made quite small. We will give quantitative asymptotic results. This is a nontrivial task since

the expectation contains periodic fluctuations which give an essential contribution to the variance.

- An important observation is the *application* of these results to *probabilistic counting*, [2]. Without going into details, n data values produce an infinite $0,1$–string with only finitely many 1's, looking like

$$1\,1\,1\ldots 1\,\mathbf{0}\,\sigma\,\sigma\,\sigma\ldots\sigma\,\mathbf{1}\,0\,0\,0\ldots,$$

where $\sigma\in\{0,1\}$. The region between the first 0 and the last 1 is called the *fringe*. The parameter R_n to estimate $\log_2 n$, used by *Flajolet and Martin*, is the *index of the first 0*. This parameter may by analyzed according to very delicate observations, compare [2], [9] and [11]. The index of the last 1 is just Szpankowski and Rego's parameter minus 1 (assuming equal probabilities for 0 and 1). (One has to subtract 1 since the indices in the string start with 0 and the geometric random variable takes only values ≥ 1.) It is less accurate than the original one used by *Flajolet and Martin*, but with the parameter d, e.g. for $d=3$, we can beat R_n with respect to the variance. However, in [2] there are also other variations reported in order to make the variance small. Nevertheless we mention that even an arbitrarily small variance cannot compensate the effect of approximating a discrete quantity by a continuous one. Furthermore it should be stated explicitely that the applicability is somehow limited in the following sense. The original algorithm of Flajolet and Martin can be used to compute the cardinality of a multiset. This is especially useful in database applications when the number of different elements in the union of several databases is of interest. For $d\geq 2$ this doesn't work well, because the d-maximum is sensitive (though not much) to multiple appearences of the same element.

2. Expectation and Variance

Let X be a random variable distributed according to the geometric distribution with parameter p; as usual, set $q=1-p$. It is convenient to define $Q=q^{-1}$ and $L=\log Q$. Then, $\mathrm{Prob}\{X\leq k\}=1-q^k$ and $\mathrm{Prob}\{X>k\}=q^k$. Therefore

$$\begin{aligned}
&\mathrm{Prob}\left\{\max\nolimits_d\{X_1,\ldots,X_n\}\leq k\right\}\\
&=\sum_{\lambda=0}^{d-1}\mathrm{Prob}\{\lambda \text{ elements are } >k \text{ and the others are } \leq k\}\\
&=\sum_{\lambda=0}^{d-1}\binom{n}{\lambda}q^{k\lambda}\left(1-q^k\right)^{n-\lambda}.
\end{aligned}$$

Let $E_n^{(d)}$ be the *expectation* of the d–maximum. We have

$$\begin{aligned}
E_n^{(d)}&=\sum_{k\geq 0}\mathrm{Prob}\left\{\max\nolimits_d\{X_1,\ldots,X_n\}>k\right\}\\
&=\sum_{k\geq 0}\left[1-\sum_{\lambda=0}^{d-1}\binom{n}{\lambda}q^{k\lambda}\left(1-q^k\right)^{n-\lambda}\right]
\end{aligned}$$

We expand $(1-q^k)^{n-\lambda}$ according to the binomial theorem, interchange the sums and combine the "1" with the instance $\lambda = 0$ to get

$$E_n^{(d)} = -\sum_{i=1}^{n} \binom{n}{i} (-1)^i \frac{1}{1-q^i} - \sum_{\lambda=1}^{d-1} \binom{n}{\lambda} \sum_{i=0}^{n-\lambda} \binom{n-\lambda}{i} (-1)^i \frac{1}{1-q^{\lambda+i}}.$$

Observe that the first sum is just $E_n^{(1)}$. Alternating sums of the type

$$\sum_{k=s}^{n} \binom{n}{k} (-1)^k f(k)$$

are most easily evaluated asymptotically by *Rice's method*, compare [5]:

Lemma 1. *Let $\mathcal{C}$ be a curve surrounding the points $s, s+1, \ldots, n$ $(s \in \mathbb{N})$ in the complex plane and let $f(z)$ be analytic inside $\mathcal{C}$. Then*

$$\sum_{k=s}^{n} \binom{n}{k} (-1)^k f(k) = -\frac{1}{2\pi i} \int_{\mathcal{C}} [n;z] f(z) dz,$$

where

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

This lemma is useful, because the integral may be asymptotically evaluated by collecting the residues of $[n;z]f(z)$ with $\Re z < s$. Consider the instance of Szpankowski and Rego,

$$E_n^{(1)} = -\sum_{i=1}^{n} \binom{n}{i} (-1)^i \frac{1}{1-Q^{-i}} = \sum_{i=1}^{n} \binom{n}{i} (-1)^i f(i)$$

with

$$f(z) = -\frac{1}{1-Q^{-z}}.$$

At $z = 0$, the local expansion looks like

$$f(z) \sim -\frac{1}{Lz}\left(1 + \frac{Lz}{2}\right).$$

The quantity $[n;z]$ has also a pole at $z = 0$;

$$[n;z] \sim -\frac{1}{z}\Big(1 + zH_n\Big)$$

with a *harmonic number* $H_n = 1 + \frac{1}{2} + \cdots + \frac{1}{n}$. The residue is[1]

$$[z^{-1}] \frac{1}{Lz}\left(1 + \frac{Lz}{2}\right) \frac{1}{z}\Big(1 + zH_n\Big) = \frac{1}{2} + \frac{H_n}{L}.$$

Observe that $H_n \sim \log n + \gamma$.

There are also simple poles at $z = \chi_k = \frac{2k\pi i}{L}$, $k \in \mathbb{Z} \setminus \{0\}$ with residue

$$-\frac{[n;\chi_k]}{L} \sim -\frac{1}{L}\Gamma(-\chi_k) n^{\chi_k} \qquad (n \to \infty).$$

Collecting all residues we obtain again the result from Szpankowski and Rego:

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

Theorem 2.

$$E_n^{(1)} \sim \log_Q n + \frac{\gamma}{L} + \frac{1}{2} + P_1\left(\log_Q n\right)$$

where $P_1(x)$ is a continuous periodic function of period 1, mean zero, small amplitude and Fourier expansion

$$P_1(x) = -\frac{1}{L}\sum_{k\neq 0}\Gamma\left(-\chi_k\right)e^{2\pi ikx}. \quad \Box$$

To perform the generalization to the general case we only have to consider the extra terms

$$-\sum_{\lambda=1}^{d-1}\binom{n}{\lambda}\sum_{i=0}^{n-\lambda}\binom{n-\lambda}{i}(-1)^i\frac{1}{1-Q^{-\lambda-i}}.$$

To this end, let $N := n-\lambda$ and consider

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

with

$$f(z) = \frac{1}{1-Q^{-\lambda-z}}.$$

Now there is only a simple pole at $z=-\lambda$. We find immediately that

$$\operatorname*{Res}_{z=-\lambda}[N;z]\frac{1}{1-Q^{-\lambda-z}} = \frac{1}{L}\frac{N!(\lambda-1)!}{(N+\lambda)!}.$$

Therefore the correction is (apart from the fluctuations) asymptotic to

$$-\frac{1}{L}\sum_{\lambda=1}^{d-1}\binom{n}{\lambda}\frac{(n-\lambda)!(\lambda-1)!}{n!} = -\frac{1}{L}H_{d-1}.$$

Hence we have

Theorem 3. *As $n\to\infty$, we have*

$$E_n^{(d)} \sim \log_Q n + \frac{\gamma}{L} + \frac{1}{2} - \frac{1}{L}H_{d-1} + P_1\left(\log_Q n\right)$$

where $P_1(x) = P_{1,d}(x)$ is a continuous periodic function of period 1, mean zero, small amplitude and Fourier expansion

$$P_1(x) = -\frac{1}{L}\sum_{k\neq 0}e^{2\pi ikx}\Gamma\left(-\chi_k\right)\left(1+\sum_{\lambda=1}^{d-1}\frac{(\lambda-1-\chi_k)_\lambda}{\lambda!}\right). \quad \Box$$

With regard to the variance we start with the second moment, which is easily computed as

$$\begin{aligned}
&\sum_{k\ge 0}(2k+1)\left[1-\sum_{\lambda=0}^{d-1}\binom{n}{\lambda}q^{k\lambda}\left(1-q^k\right)^{n-\lambda}\right]\\
&=-\sum_{i=1}^{n}\binom{n}{i}(-1)^i\frac{1+Q^{-i}}{\left(1-Q^{-i}\right)^2}\\
&-\sum_{\lambda=1}^{d-1}\binom{n}{\lambda}\sum_{i=0}^{n-\lambda}\binom{n-\lambda}{i}(-1)^i\frac{1+Q^{-\lambda-i}}{\left(1-Q^{-\lambda-i}\right)^2}.
\end{aligned}$$

Let's start with the main term

$$\sum_{i=1}^{n}\binom{n}{i}(-1)^i f(i)$$

with

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

At $z=0$ there is altogether a triple pole; we have

$$f(z)\sim-\frac{2}{L^2z^2}\left(1+\frac{Lz}{2}+\frac{L^2z^2}{6}\right).$$

Also,

$$[n;z]\sim-\frac{1}{z}\left(1+zH_n+z^2\frac{H_n^2+H_n^{(2)}}{2}\right),$$

(with $H_n^{(2)}=\sum_{k=1}^{n}\frac{1}{k^2}$, a harmonic number of order 2) and the residue of $[n;z]f(z)$ is

$$\begin{aligned}
&\frac{H_n^2}{L^2}+\frac{H_n^{(2)}}{L^2}+\frac{H_n}{L}+\frac{1}{3}\\
&\sim\log_Q^2 n+\frac{2\gamma\log_Q n}{L}+\frac{\gamma^2}{L^2}+\frac{\pi^2}{6L^2}+\log_Q n+\frac{\gamma}{L}+\frac{1}{3}.
\end{aligned}$$

For the correction, consider (with $N=n-\lambda$)

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

with

$$f(z)=\frac{1+Q^{-\lambda-z}}{\left(1-Q^{\lambda-z}\right)^2}.$$

At $z=-\lambda$ there is a double pole; with $w=\lambda+z\to 0$ we have

$$f(z)\sim\frac{2}{L^2w^2}\left(1+\frac{Lw}{2}\right).$$

Furthermore,

$$[N;z] \sim \frac{N!(\lambda-1)!}{(N+\lambda)!}\Big[1 + w\left(H_{N+\lambda} - H_{\lambda-1}\right)\Big],$$

and the residue of $[N;z]f(z)$ at $z = -\lambda$ is

$$\frac{N!(\lambda-1)!}{(N+\lambda)!}\frac{2}{L^2}\left(\frac{L}{2} + H_{N+\lambda} - H_{\lambda-1}\right).$$

We can sum this up as follows:

$$\begin{aligned}
&-\sum_{\lambda=1}^{d-1}\binom{n}{\lambda}\frac{(n-\lambda)!(\lambda-1)!}{n!}\frac{2}{L^2}\left(\frac{L}{2} + H_n - H_{\lambda-1}\right)\\
&= -\sum_{\lambda=1}^{d-1}\frac{1}{\lambda}\left(\frac{1}{L} + \frac{2}{L^2}H_n - \frac{2}{L^2}H_{\lambda-1}\right)\\
&= -\frac{1}{L}H_{d-1} - \frac{2}{L^2}H_nH_{d-1} + \frac{H_{d-1}^2}{L^2} - \frac{H_{d-1}^{(2)}}{L^2}
\end{aligned}$$

Altogether we find for the second moment (apart from the fluctuations) the asymptotic equivalent

$$\begin{aligned}
&\log_Q^2 n + \frac{2\gamma\log_Q n}{L} + \frac{\gamma^2}{L^2} + \frac{\pi^2}{6L^2} + \log_Q n + \frac{\gamma}{L} + \frac{1}{3}\\
&-\frac{1}{L}H_{d-1} - \frac{2H_{d-1}}{L}\log_Q n - \frac{2H_{d-1}}{L^2}\gamma + \frac{H_{d-1}^2}{L^2} - \frac{H_{d-1}^{(2)}}{L^2}.
\end{aligned}$$

Now for the variance we must subtract the square of the expectation. There are many terms cancelling, and we end up with

Proposition 4. *The variance $V_n^{(d)}$ fulfills*

$$V_n^{(d)} \sim \frac{\pi^2}{6L^2} + \frac{1}{12} - \frac{H_{d-1}^{(2)}}{L^2} - \left[P_1^2\right]_0 + P_2\left(\log_Q n\right). \quad \square$$

where $P_2(x)$ is a continuous periodic function of period 1 and mean 0, and $\left[P_1^2\right]_0$ is the "mean" of the square of the function $P_1(x)$ from Theorem 3.

3. An Analysis of the Function $P_1(x)$

First, by the elementary formula

$$\sum_{k=0}^{r}\binom{k+a}{k} = \binom{r+a+1}{r},$$

we may rewrite

$$P_1(x) = -\frac{1}{L}\sum_{k\neq 0} e^{2\pi ikx}\Gamma\left(-\chi_k\right)\sum_{\lambda=0}^{d-1}\frac{(\lambda-1-\chi_k)_\lambda}{\lambda!}$$

as

$$P_1(x) = -\frac{1}{L}\sum_{k\neq 0} e^{2\pi i k x}\,\Gamma\left(-\chi_k\right)\binom{d-1-\chi_k}{d-1}.$$

Therefore the mean $\left[P_1^2\right]_0$ of $P_1^2(x)$ is

$$\left[P_1^2\right]_0 = \frac{1}{L^2}\sum_{k\neq 0}\left|\Gamma\left(-\chi_k\right)\right|^2\binom{\chi_k-1}{d-1}\binom{-\chi_k-1}{d-1};$$

it is this quantity that we are going to study thoroughly in this section.

Lemma 5. $\displaystyle\binom{a-1}{n}\binom{-a-1}{n} = \sum_{k=0}^{n}\binom{a}{k}\binom{-a}{k}$

Proof. We use Euler's first identity for hypergeometric series [7], [8]

$${}_2F_1(c-a,c-b;c;x) = {}_1F_0(c-a-b;x)\,{}_2F_1(a,b;c;x)$$

in the special instance $c=1$, $b=-a$ where it takes the form

$$\sum_{n\geq 0}\frac{(1-a)^{(n)}(1+a)^{(n)}}{n!^2}x^n = \frac{1}{1-x}\sum_{n\geq 0}\frac{a^{(n)}(-a)^{(n)}}{n!^2}x^n$$

with the Pochhammer symbol $a^{(n)} = a(a+1)\dots(a+n-1)$. Now comparing the coefficients of x^n and observing that

$$\frac{(1-a)^{(n)}(1+a)^{(n)}}{n!^2} = \binom{n-a}{n}\binom{n+a}{n} = \binom{-a-1}{n}\binom{a-1}{n}$$

resp.

$$\frac{a^{(k)}(-a)^{(k)}}{k!^2} = \binom{a+k-1}{k}\binom{-a+k-1}{k} = \binom{-a}{k}\binom{a}{k}$$

gives the desired identity. $\square$

Remark. Once the formula is *known*, it may of course also be proved by induction.

An immediate consequence is the formula

$$\left[P_1^2\right]_0 = \frac{1}{L^2}\sum_{j=0}^{d-1}\frac{1}{j!^2}\sum_{k\neq 0}\left|\Gamma\left(j-\chi_k\right)\right|^2.$$

The following lemmata evaluate the inner sums. A proof for the special instance $Q=2$, using series transformation results due to Ramanujan, is given in [10]. In the Appendix, we sketch an alternative proof for general $Q>1$ using the Mellin transform combined with the residue calculus.

Lemma 6. *If $j \geq 1$,*

$$\sum_{k\neq 0} \Gamma(j+\chi_k)\Gamma(j-\chi_k) = 2L(2j-1)! \sum_{h\geq 0} \binom{-2j}{h} \frac{1}{Q^{h+j}-1} + L(2j-1)!2^{-2j} - (j-1)!^2.$$

Lemma 7.

$$\sum_{k\neq 0} \Gamma(\chi_k)\Gamma(-\chi_k) = \frac{\pi^2}{6} + \frac{L^2}{12} - L\log 2 - 2L \sum_{h\geq 1} \frac{(-1)^{h-1}}{h\,(Q^h-1)}$$

Now this gives us our final formula which we therefore state as a proposition.

Proposition 8.

$$\begin{aligned}
\left[P_1^2\right]_0 = {} & \frac{\pi^2}{6L^2} + \frac{1}{12} - \frac{\log 2}{L} - \frac{2}{L} \sum_{h\geq 1} \frac{(-1)^{h-1}}{h\,(Q^h-1)} \\
& + \frac{2}{L} \sum_{j=1}^{d-1} \frac{1}{2j} \binom{2j}{j} \sum_{h\geq 0} \binom{-2j}{h} \frac{1}{Q^{h+j}-1} \\
& + \frac{1}{L} \sum_{j=1}^{d-1} \frac{1}{2j} \binom{2j}{j} 2^{-2j} - \frac{H_{d-1}^{(2)}}{L^2}
\end{aligned}$$

4. Behaviour of the Variance for large d

Combining Propositions 4 and 8 we have proved that the main term of $V_n^{(d)}$ for $n \to \infty$ is given by

$$\begin{aligned}
& \frac{\log 2}{L} - \frac{1}{L} \sum_{j=1}^{d-1} \frac{1}{2j} \binom{2j}{j} 2^{-2j} + \frac{2}{L} \sum_{h\geq 1} \frac{(-1)^{h-1}}{h\,(Q^h-1)} \\
& - \frac{2}{L} \sum_{j=1}^{d-1} \frac{1}{2j} \binom{2j}{j} \sum_{h\geq 0} \binom{-2j}{h} \frac{1}{Q^{h+j}-1} + P_2(\log_2 n)
\end{aligned}$$

where $P_2(x)$ has mean zero.

In this section we want to analyze the nonfluctuating part of the above quantity for $d \to \infty$.

We start with the first sum and prove, as the main step,

Lemma 9.

$$4^{-j} \frac{1}{2j} \binom{2j}{j} = -[t^j] \log \frac{1+\sqrt{1-t}}{2}$$

Proof. We have

$$\begin{aligned}
[t^j]\log\frac{1+\sqrt{1-t}}{2} &= \frac{1}{j}[t^{j-1}]\left(\log\frac{1+\sqrt{1-t}}{2}\right)' \\
&= -\frac{1}{2}\frac{1}{j}[t^{j-1}]\frac{1}{\sqrt{1-t}\,\bigl(1+\sqrt{1-t}\,\bigr)}.
\end{aligned}$$

To find the coefficient in this sum we use the *formal residue calculus*, as described in [6], with the substitution

$$t=\frac{4u}{(1+u)^2} \qquad \text{and} \qquad dt=\frac{4(1-u)}{(1+u)^3}du.$$

$$\begin{aligned}
&[t^{j-1}]\frac{1}{\sqrt{1-t}\,\bigl(1+\sqrt{1-t}\,\bigr)} \\
&= [t^{-1}]t^{-j}\frac{1}{\sqrt{1-t}\,\bigl(1+\sqrt{1-t}\,\bigr)} \\
&= [u^{-1}]4^{-j}u^{-j}(1+u)^{2j}\frac{1+u}{1-u}\frac{1+u}{2}\frac{4(1-u)}{(1+u)^3} \\
&= \frac{1}{2}4^{1-j}[u^{j-1}](1+u)^{2j-1}=\frac{1}{2}4^{1-j}\binom{2j-1}{j-1},
\end{aligned}$$

and this is clearly equivalent to the announced formula. $\square$

As an immediate consequence of the lemma we get

Proposition 10. *As $d\to\infty$,*

$$\frac{1}{L}\sum_{j=1}^{d-1}\frac{1}{2j}\binom{2j}{j}2^{-2j}=\frac{\log 2}{L}+\frac{1}{\sqrt{\pi}}d^{-1/2}+\mathcal{O}(d^{-3/2}).$$

Proof. We have

$$\begin{aligned}
\sum_{j=1}^{d-1}\frac{1}{2j}\binom{2j}{j}2^{-2j} &= -[t^{d-1}]\frac{1}{1-t}\log\frac{1+\sqrt{1-t}}{2} \\
&= [t^{d-1}]\frac{\log 2}{1-t}-[t^{d-1}]\frac{1}{1-t}\log\bigl(1+\sqrt{1-t}\,\bigr) \\
&= \log 2-[t^{d-1}]\frac{1}{1-t}\log\bigl(1+\sqrt{1-t}\,\bigr)
\end{aligned}$$

The latter generating function has its dominating singularity at $t=1$ and behaves as

$$-\frac{1}{1-t}\log\bigl(1+\sqrt{1-t}\,\bigr)=\frac{1}{\sqrt{1-t}}+\mathcal{O}\left((1-t)^{-3/2}\right) \qquad (t\to 1),$$

so that by *singularity analysis* [4] the coefficient of t^{d-1} behaves like

$$\frac{1}{\sqrt{\pi}}d^{-1/2} + \mathcal{O}(d^{-3/2}) \qquad (d \to \infty). \quad \square$$

It remains to treat the more complicated sum

$$\Sigma_{d-1} = \sum_{j=1}^{d-1} \frac{1}{2j}\binom{2j}{j} \sum_{h\geq 0} \binom{-2j}{h} \frac{1}{Q^{h+j}-1}.$$

Observe that

$$\begin{aligned}\sum_{h\geq 0}\binom{-2j}{h}\frac{1}{Q^{h+j}-1} &= \sum_{h\geq 0}\binom{-2j}{h}\frac{1}{Q^{h+j}}\sum_{m\geq 0} Q^{(-h-j)m} \\ &= \sum_{m\geq 1}\left(Q^{-m}\left(1+Q^{-m}\right)^{-2}\right)^{j}.\end{aligned}$$

Now it follows from a simple substitution in Lemma 9 that

$$\frac{1}{2j}\binom{2j}{j}\left(\frac{1}{Q^m(1+Q^{-m})^2}\right)^j = -[t^j]\log\left(\frac{1+\sqrt{1-\frac{4t}{Q^m(1+Q^{-m})^2}}}{2}\right).$$

Therefore we have

$$\Sigma_{d-1} = -[t^{d-1}]\frac{1}{1-t}\log\left(\frac{1+\sqrt{1-\frac{4t}{Q^m(1+Q^{-m})^2}}}{2}\right).$$

Observing that

$$\begin{aligned}\Sigma_\infty &= -\sum_{m\geq 1}\log\left(\frac{1+\sqrt{1-\frac{4t}{Q^m(1+Q^{-m})^2}}}{2}\right) \\ &= -\sum_{m\geq 1}\log\frac{1}{1+Q^{-m}} = \sum_{m\geq 1}\log\left(1+Q^{-m}\right)\end{aligned}$$

we find by singularity analysis

Proposition 11. *For any $\varepsilon > 0$, we have for $d \to \infty$*

$$\frac{2}{L}\Sigma_{d-1} = \frac{2}{L}\sum_{m\geq 1}\log\left(1+Q^{-m}\right) + \mathcal{O}\left(\left(\frac{4}{Q(1+Q^{-1})^2}\right)^{d-\varepsilon}\right)$$

The main term coincides with

$$\frac{2}{L}\sum_{h\geq 1}\frac{(-1)^{h-1}}{h(Q^h-1)},$$

whereas, for $Q > 1$, the remainder term is exponentially small for $d \to \infty$ compared with the remainder term from Proposition 10. Altogether we have proved:

Theorem 12. *The nonfluctuating part of the main asymptotic term of the variance $V_n^{(d)}$ for $n \to \infty$ behaves for $d \to \infty$ as*

$$\frac{1}{\sqrt{\pi}} d^{-1/2} + \mathcal{O}\big(d^{-3/2}\big).$$

Remark. This analysis is of more than local interest; the periodic function $P_1(x)$ is not uncommon in the *Analysis of Algorithms*. We mention e.g. [13].

5. Application to Probabilistic Counting

To continue the discussion from the Introduction, let us just mention that (by means of a *hash function*) each data value produces exactly one bit, and it is in position $k \geq 0$ with probability 2^{-k-1}. The OR–composition of all the bits (1's) constitutes the string, as seen in the Introduction. Now this is exactly a *geometric probability*. The result of Flajolet and Martin states, that (apart from the fluctuations) the *average value* of R_n is asymptotic to $\log_2 n - 0.37$; the Szpankowski–Rego parameter -1 has an average of $\log_2 n + 0.33$, whence we deduce the average size of the fringe to be less than 1.

The *variance* in the Flajolet–Martin–instance is approximately 1.257 and in the Szpankowski–Rego–instance 3.507.

Below we give the limiting values of the variances $V_n^{(d)}$ for $N \to \infty$ and some values of d for the special instance $q = \frac{1}{2}$, i.e. $Q = 2$.

d	$\lim V_n^{(d)}$
1	3.5070
2	1.4256
3	0.9053
4	0.6740

Finally we mention that from the algorithmic point of view it is easy to maintain the parameter of interest. It is not even necessary to keep the whole actual bit string (the OR–composition). It is sufficient to store the actual d data values with highest entries.

Appendix

As announced in Section 3 we sketch here a derivation of the formula from Lemma 6 (Lemma 7 can be proved in a quite similar manner).

We start from the left hand side and interpret the sum as the collection of the residues at the poles χ_k, $k \neq 0$, of the function

$$L \frac{\Gamma(j+z)\Gamma(j-z)}{e^{Lz} - 1}.$$

Therefore we have

$$\sum_{k\neq 0} |\Gamma(j+\chi_k)|^2 = \frac{L}{2\pi i}\int_{(1/2)} \frac{\Gamma(j+z)\Gamma(j-z)}{e^{Lz}-1}dz - \frac{L}{2\pi i}\int_{(-1/2)} \frac{\Gamma(j+z)\Gamma(j-z)}{e^{Lz}-1}dz - \Gamma(j)^2.$$

($\Gamma(j)^2$ is the residue at $z=0$.)

Now we use the decomposition

$$\frac{1}{e^{Lz}-1} = -1 - \frac{1}{e^{-Lz}-1}$$

for the second integral and get

$$\begin{aligned} &-\frac{L}{2\pi i}\int_{(-1/2)} \frac{\Gamma(j+z)\Gamma(j-z)}{e^{Lz}-1}dz \\ &= \frac{L}{2\pi i}\int_{(-1/2)} \Gamma(j+z)\Gamma(j-z)dz + \frac{L}{2\pi i}\int_{(-1/2)} \frac{\Gamma(j+z)\Gamma(j-z)}{e^{-Lz}-1}dz \\ &= \frac{L}{2\pi i}\int_{(0)} \Gamma(j+z)\Gamma(j-z)dz + \frac{L}{2\pi i}\int_{(1/2)} \frac{\Gamma(j-z)\Gamma(j+z)}{e^{Lz}-1}dz. \end{aligned}$$

Therefore

$$\begin{aligned} &\sum_{k\neq 0} |\Gamma(j+\chi_k)|^2 \\ &= \frac{2L}{2\pi i}\int_{(1/2)} \frac{\Gamma(j+z)\Gamma(j-z)}{e^{Lz}-1}dz + \frac{L}{2\pi i}\int_{(0)} \Gamma(j+z)\Gamma(j-z)dz - \Gamma(j)^2 \\ &= I_1 + I_2 - \Gamma(j)^2. \end{aligned}$$

I_1 is evaluated by shifting the contour to the *right* and collecting the *negative* residues, which gives

$$I_1 = -2L\sum_{m\geq j} \frac{\Gamma(j+m)}{e^{Lm}-1}\frac{(-1)^{j-m+1}}{(m-j)!}$$

and with $m = h+j$

$$\begin{aligned} &= 2L\sum_{h\geq 0} \frac{(h+2j-1)!(-1)^h}{h!}\frac{1}{Q^{h+j}-1} \\ &= 2L(2j-1)!\sum_{h\geq 0} \binom{-2j}{h}\frac{1}{Q^{h+j}-1}. \end{aligned}$$

Integral I_2 is of interest for itself and appears already in early references to the *Mellin transform technique* as by Nielsen [12, p.224].

We start with the function

$$f(x) = \frac{x^j}{(1+x)^{2j}}$$

and perform its Mellin transform (see, e.g., [3] for definitions)

$$f^*(s) = \int_0^\infty f(x) x^{s-1} dx = B(j+s, j-s) = \frac{\Gamma(j+s)\Gamma(j-s)}{\Gamma(2j)}$$

with the Beta function $B(z, w)$ (compare [1]). The *fundamental strip* is $\langle -j, j \rangle$. Therefore the inversion formula for the Mellin transform gives

$$f(x) = \frac{1}{2\pi i} \int_{(0)} \frac{\Gamma(j+s)\Gamma(j-s)}{\Gamma(2j)} x^{-s} ds.$$

Now we may evaluate at $x = 1$ and get the formula

$$\frac{1}{2\pi i} \int_{(0)} \Gamma(j+s)\Gamma(j-s) ds = \Gamma(2j) 2^{-2j}.$$

This completes the proof of Lemma 6.

References

1. M.Abramowitz, I.A.Stegun, *Handbook of Mathematical Functions*, Dover, New York, 1970.
2. P.Flajolet and G.N.Martin, *Probabilistic Counting Algorithms for Data Base Applications*, J.Comput.Syst.Sci. **31** (1985), 182–209.
3. P.Flajolet, M.Régnier and R.Sedgewick, *Combinatorics on Words*, Springer NATO ASI Series F, Vol. 12, Berlin, 1985.
4. P.Flajolet and A.Odlyzko, *Singularity Analysis of Generating Functions*, SIAM J. Disc. Math. **3** (1990), 216–240.
5. P.Flajolet and R.Sedgewick, *Digital Search Trees Revisited*, SIAM J.Comput. **15** (1986), 748–767.
6. I.P.Goulden and D.M.Jackson, *Combinatorial Enumeration*, John Wiley & Sons, New York, 1983.
7. R.L.Graham, D.E.Knuth and O.Patashnik, *Concrete Mathematics*, Addison-Wesley, Reading, MA.
8. P.Henrici, *Applied and Computational Complex Analysis Volume 1*, John Wiley & Sons, New York, 1974.
9. P.Kirschenhofer and H.Prodinger, *On the Analysis of Probabilistic Counting*, Number Theoretic Analysis, E.Hlawka, R.F.Tichy (Eds.), Lecture Notes in Mathematics **1452** (1990), 117–120.
10. P.Kirschenhofer and H.Prodinger, *On Some Applications of Formulæof Ramanujan in the Analysis of Algorithms*, Mathematika **38** (1991), 14–33.
11. P.Kirschenhofer, H.Prodinger and W. Szpankowski, *How to Count Quickly and Accurately: A Unified Analysis of Probabilistic Counting and Other Related Problems*, ICALP92, W.Kuich (Ed.), Lecture Notes in Computer Science **623** (1992), 211–222.
12. N.Nielsen, *Handbuch der Theorie der Gammafunktion*, Teubner, Leipzig, 1906.

13. H.Prodinger, *Über längste 1-Teilfolgen in 0-1–Folgen*, in: Zahlentheoretische Analysis II, E.Hlawka, Ed., Lecture Notes in Mathematics **1262** (1987), 124–133.
14. W.Szpankowski and V.Rego, *Yet another Application of a Binomial Recurrence: Order Statistics*, Computing **43** (1990), 401–410.

Department of Algebra and Discrete Mathematics
TU Vienna
Wiedner Hauptstrasse 8-10/118
A-1040 Vienna
Austria