COMBINATORICS OF GEOMETRICALLY DISTRIBUTED RANDOM VARIABLES: LEFT-TO-RIGHT MAXIMA

Helmut Prodinger

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

Abstract. Assume that the numbers $x_1, \dots, x_n$ are the output of n independent geometrically distributed random variables. The number x_i is a left-to-right maximum if it is greater (or equal, for a variation,) than $x_1, \dots, x_{i-1}$. A precise average case analysis is performed for the parameter 'number of left-to-right maxima'. The methods include generating functions and a technique from complex analysis, called Rice's method. Some additional results are also given.

1. Introduction

Let X denote a geometrically distributed random variable, i.e. $\mathbb{P}\{X = k\} = pq^{k-1}$ for $k \in \mathbb{N}$ and $q = 1 - p$.

The combinatorics of n geometrically distributed independent variables $X_1, \dots, X_n$ becomes more and more important, especially because of applications in Computer science. We just mention two areas: The **skiplist** ([2], [14], [15]) and **probabilistic counting** ([4], [8], [9], [10]).

> The skip list is a data structure for searching. To each of the n elements that are stored there will be some pointer fields, and the number of those is chosen according to a geometric random variable. Furthermore, the 'horizontal search costs' of a particular element are just the number of left-to-right maxima of the truncated and reversed sequence.
>
> In the case of probabilistic counting we think about $p = q = \frac{1}{2}$. The parameter of interest is then the smallest natural number that does not appear as an output of any of the X_i's.

We will cite one particular result, since we can use the corresponding asymptotic formula in the sequel for our own findings.

Theorem 1. [Szpankowski and Rego, [17]] *The expected value E_n of $\max\{X_1, \dots, X_n\}$ is given by*

$$E_n = \sum_{k \ge 0} \Big[1 - \big(1 - q^k\big)^n\Big] = \log_Q n + \frac{\gamma}{L} + \frac{1}{2} - \delta\left(\log_Q n\right) + O\Big(\frac{1}{n}\Big). \tag{1.1}$$

Here and in the whole paper, $Q = q^{-1}$ and $L = \log Q$; γ is Euler's constant and $\delta(x)$ is a periodic function of period 1 and mean 0 which is given by the Fourier series

$$\delta(x) = \frac{1}{L} \sum_{k \ne 0} \Gamma\left(-\chi_k\right) e^{2k\pi i x}. \tag{1.2}$$

The complex numbers χ_k are given by $\chi_k = 2k\pi i/L$.

It should be mentioned that the asymptotic evaluation of the series in (1.1) is by now a "folklore theorem"; it appears e.g. in Knuth's book [12] or in [6].

In this paper we want to concentrate on the *number of left-to-right maxima*. This is a well studied parameter in the context of random permutations; practically every text dealing with the analysis of algorithms devotes a section to it ([11], [7], [6], [16]).

As is easy to guess by the name, this parameter counts how often we meet a number that is larger than all the elements to the left.

The results for random permutations are: The harmonic number H_n for the expectation and $H_n - H_n^{(2)}$ for the variance; here, $H_n^{(2)} = 1 + \frac{1}{4} + \cdots + \frac{1}{n^2}$ denotes the n-th harmonic number of second order .

There are 2 meaningful ways to define left-to-right maxima: A number is a left-to-right maximum if it is strictly larger than the elements to left (the "strict" sense) or a number is a left-to-right maximum if it is larger or equal than the elements to left (the "loose" sense). On might call the first variant "without repetitions" and the second "with repetitions"; with respect to the computations involved the terms "easy" and "not quite easy" might also be in order. (Or "strong" versus "weak".) Of course, for random permutations this distinction does not make a difference.

The next two sections deal with the expectations; they are followed by two sections on the variances. In the next two sections we study a related model (again with 2 variants), namely: The random variable X takes the values $1, \dots, M$, each with probability $\frac{1}{M}$. (The uniform distribution.)

Finally, we come back to the original model of geometrically distributed variables and consider left-to-right *minima*. This makes no difference for random permutations and equidistribution, but here it is different! However, the computations for the *maxima* require "higher mathematics" (Mellin transforms or Rice's method), and some periodicity phenomena are encountered, whereas the calculations for the *minima* are quite elementary.

It is interesting to note that the concept of left-to-right maxima appears also as "records", compare[1].

To obtain the leading terms of the asymptotic expansions in Theorems 2 and 3 (namely $p \log_Q n$ and $\frac{p}{q} \log_Q n$) is not a big deal; however it quite surprising (and non trivial) that the lower order terms contain "fluctuations" (see below for details).

Since we are considering, assuming different models, the analogous quantities over and over, we decided to stay with the same notations. So the reader is warned that the same notations might have different meanings in different sections.

2. The expectation in the strict model

Our approach will be by generating functions; to find them, we set up a certain "language" $\mathcal{L}$ and translate it to enumerating generating functions. We give two references for such a procedure: [13] and [6].

To avoid confusion, we denote the "letters" by $\mathbf{1}, \mathbf{2}, \dots$. We decompose all sequences $x_1 x_2 \dots$ in a canonical way as follows: We combine each left-to-right maximum $\mathbf{k}$ with the following (smaller or equal) elements. Such a part is decribed by

$$\mathcal{A}_k := \mathbf{k}\{\mathbf{1}, \dots, \mathbf{k}\}^*. \tag{2.1}$$

Such a group may be present or not. This observation gives the desired "language", where ε denotes the "empty word":

$$\mathcal{L} := (\mathcal{A}_1 + \varepsilon) \cdot (\mathcal{A}_2 + \varepsilon) \cdot (\mathcal{A}_3 + \varepsilon) \dots \tag{2.2}$$

It is important not to confuse ε with something like a letter $\mathbf{0}$ (not being present here).

Now we want to mark each letter by a "z" and each left-to-right maximum by a "y". The probability pq^{k-1} for a letter $\mathbf{k}$ should of course not being forgotten. $\{\mathbf{1},\ldots,\mathbf{k}\}$ maps into $z(1-q^k)$ and its star $\{\mathbf{1},\ldots,\mathbf{k}\}^*$ into $1/\big(1-z(1-q^k)\big)$. So we obtain the generating function $F(z,y)$ as an infinite product:

$$F(z,y)=\prod_{k\ge1}\left(1+\frac{yzpq^{k-1}}{1-z(1-q^k)}\right) \tag{2.3}$$

To be explicit, the coefficient of z^ny^k in $F(z,y)$ is the probability that n random variables have k left-to-right maxima. Observe that, as it is to be expected, $F(z,1)=\frac{1}{1-z}$, as it is then a telescoping product. Let $f(z)=\frac{\partial F(z,y)}{\partial y}\big|_{y=1}$. It is the generating function for the expected values E_n, i.e. the $E_n=[z^n]f(z)$. Performing this differentiation we obtain

$$f(z)=\frac{pz}{1-z}\sum_{k\ge0}\frac{q^k}{1-z(1-q^k)}, \tag{2.4}$$

which is also, by partial fraction decomposition,

$$f(z)=p\sum_{k\ge0}\left[\frac{1}{1-z}-\frac{1}{1-z(1-q^k)}\right]. \tag{2.5}$$

From this the coefficients E_n are easy to see, because there are only geometric series:

$$E_n=[z^n]f(z)=p\sum_{k\ge0}\Big[1-\big(1-q^k\big)^n\Big] \tag{2.6}$$

But, as announced earlier, apart from the factor p, the asymptotic evaluation of this is well known, and we have obtained the following result.

Theorem 2. *The average number E_n of left-to-right maxima (strict model) in the context of n independently distributed geometric random variables has the asymptotic expansion*

$$E_n=p\left[\log_Q n+\frac{\gamma}{L}+\frac12-\delta\left(\log_Q n\right)\right]+O\big(\frac1n\big) \tag{2.7}$$

with the periodic function $\delta(x)$ from (1.2).

Observe that the factor of the leading term, $p/\log Q=(q-1)/\log q$, goes monotonically from 0 to 1 as q varies between 0 and 1. So the logarithmic term disappears for $q\to0$, which is intuitively clear, since the result of the "random variables" is then the sequence $\mathbf{111}\ldots\mathbf{1}$, with only one left-to-right maximum.

3. The expectation in the loose model

Again, we are defining an appropriate "language" $\mathcal{L}$ from which a bivariate generating function $F(z,y)$ can be derived. Set $\mathcal{A}_k:=\mathbf{k}\{\mathbf{1},\ldots,\mathbf{k-1}\}^*$, then $\mathcal{L}:=\mathcal{A}_1^*\cdot\mathcal{A}_2^*\cdot\mathcal{A}_3^*\ldots$, and

$$F(z,y)=\prod_{k\ge1}\frac{1}{1-\frac{yzpq^{k-1}}{1-z(1-q^{k-1})}}=\prod_{k\ge0}\frac{1-z(1-q^k)}{1-z+zq^k(1-py)}. \tag{3.1}$$

Therefore

$$f(z)=\frac{\partial F(z,y)}{\partial y}\bigg|_{y=1}=\frac{pz}{1-z}\sum_{k\ge0}\frac{q^k}{1-z(1-q^{k+1})}=\frac{p}{q}\sum_{k\ge1}\left[\frac{1}{1-z}-\frac{1}{1-z(1-q^k)}\right] \tag{3.2}$$

and

$$E_n=[z^n]f(z)=\frac{p}{q}\sum_{k\ge1}\left[1-\left(1-q^k\right)^n\right]. \tag{3.3}$$

By virtue of Theorem 1 we therefore obtain

Theorem 3. *The average number E_n of left-to-right maxima (loose model) in the context of n independently distributed geometric random variables has the asymptotic expansion*

$$E_n=\frac{p}{q}\left[\log_Q n+\frac{\gamma}{L}-\frac{1}{2}-\delta\left(\log_Q n\right)\right]+O\big(\frac{1}{n}\big) \tag{3.4}$$

with $\delta(x)$ from Theorem 1 (and Theorem 2).

The function $q\to p/q\log Q$ (the factor of the leading coefficient $\log n$) is monotone decreasing from infinity to zero as q varies from 0 to 1.

4. The variance in the strict model

We start from $F(z,y)$ from equation (2.3) and observe (following e.g. [11]) that the variance V_n may be obtained by

$$V_n=[z^n]\frac{\partial^2F(z,y)}{\partial y^2}\bigg|_{y=1}+E_n-E_n^2. \tag{4.1}$$

Therefore, we consider

$$\frac{\partial^2F(z,y)}{\partial y^2}\bigg|_{y=1}=g(z)+h(z). \tag{4.2}$$

We use Leibniz' formula to differentiate a product; then $g(z)$ comprises all terms where different factors are differentiated once and $h(z)$ comprises all terms where one factor is differentiated twice. In this (easier) instance, $h(z)=0$.

$$\begin{aligned} g(z)&=\frac{2}{1-z}\sum_{1\le i<j}\frac{zpq^{i-1}}{1-z(1-q^{i-1})}\cdot\frac{zpq^{j-1}}{1-z(1-q^{j-1})}\\ &=2p^2\sum_{0\le i<j}\left[\frac{1}{1-z}+\frac{1}{q^{i-j}-1}\cdot\frac{1}{1-z(1-q^i)}+\frac{1}{q^{j-i}-1}\cdot\frac{1}{1-z(1-q^j)}\right] \end{aligned} \tag{4.3}$$

and

$$g_n:=[z^n]g(z)=2p^2\sum_{0\le i<j}\left[1+\frac{1}{q^{i-j}-1}\cdot\left(1-q^i\right)^n+\frac{1}{q^{j-i}-1}\cdot\left(1-q^j\right)^n\right]. \tag{4.4}$$

This time we cannot resort to earlier results from the literature, whence we have to derive the asymptotic equivalent for g_n from scratch.

Decompositions of the double sums must be done with some care because of convergence problems. The second sum is easy:

$$\sum_{0\le i<j}\frac{1}{q^{i-j}-1}\cdot\left(1-q^i\right)^n=\sum_{i\ge0}\left(1-q^i\right)^n\cdot\sum_{h\ge1}\frac{1}{q^{-h}-1}=\alpha\cdot\sum_{i\ge0}\left(1-q^i\right)^n \tag{4.5}$$

with the traditional abbreviation $\alpha = \alpha_Q = \sum_{h\geq 1} \dfrac{1}{Q^h - 1}$. Now consider the other part:

$$\sum_{0\leq i<j} \left[1 + \frac{1}{q^{j-i}-1} \cdot \left(1-q^j\right)^n \right] = \sum_{j\geq 1} \left[j + \sum_{h=1}^{j} \frac{1}{q^h - 1} \left(1-q^j\right)^n \right]$$

$$= \sum_{j\geq 1} \left[j - \left(j + \sum_{h=1}^{j} \frac{1}{Q^h-1} \right) \left(1-q^j\right)^n \right] = \sum_{j\geq 1} j \left[1 - \left(1-q^j\right)^n \right] - \sum_{j\geq 1} \sum_{h=1}^{j} \frac{1}{Q^h-1} \left(1-q^j\right)^n \tag{4.6}$$

Collecting the terms we find

$$\frac{g_n}{2p^2} = \sum_{j\geq 1} j \left[1 - \left(1-q^j\right)^n \right] + \sum_{j\geq 1} \sum_{h>j} \frac{1}{Q^h-1} \left(1-q^j\right)^n . \tag{4.7}$$

We want to apply *Rice's method*, which we cite as follows: (compare, e.g. [5], [6])

Lemma. *Let $\mathcal{C}$ be a curve surrounding the points $1, 2, \dots, n$ in the complex plane and let $f(z)$ be analytic inside $\mathcal{C}$. Then*

$$\sum_{k=1}^{n} \binom{n}{k} (-1)^k f(k) = -\frac{1}{2\pi i} \int_{\mathcal{C}} [n;z] f(z) dz, \tag{4.8}$$

where

$$[n;z] = \frac{(-1)^{n-1} n!}{z(z-1)\dots(z-n)} = \frac{\Gamma(n+1)\Gamma(-z)}{\Gamma(n+1-z)}. \quad \square \tag{4.9}$$

Extending the contour of integration it turns out that under suitable growth conditions on $f(z)$ (compare [5]) the asymptotic expansion of the alternating sum is given by

$$\sum \operatorname{Res}\big([n;z]f(z)\big) + \text{smaller order terms} \tag{4.10}$$

where the sum is taken over all poles z_0 different from $1, \dots, n$.

The range $1, \dots, n$ for the summation is not sacred; if we sum, for example, over $k = 2, \dots, n$, the contour must encircle $2, \dots, n$, etc.

Therefore we rewrite the terms of (4.7) as alternating sums.

$$\sum_{j\geq 1} j \left[1 - \left(1-q^j\right)^n \right] = -\sum_{k=1}^{n} \binom{n}{k} (-1)^k \sum_{j\geq 1} j q^{jk} = \sum_{k=1}^{n} \binom{n}{k} (-1)^k \varphi(z) \tag{4.11}$$

with

$$\varphi(z) = -\frac{Q^z}{(Q^z-1)^2}. \tag{4.12}$$

The other sum from (4.7) is

$$\sum_{j\geq 1} \sum_{h>j} \frac{1}{Q^h-1} \left(1-q^j\right)^n = \sum_{k=0}^{n} \binom{n}{k} (-1)^k \sum_{h>j\geq 1} \frac{1}{Q^h-1} Q^{-jk} = \sum_{k=0}^{n} \binom{n}{k} (-1)^k \psi(k), \tag{4.13}$$

with

$$\psi(z)=\sum_{m\geq 1}\sum_{h>j\geq 1}Q^{-hm-jz}=\sum_{m,s,j\geq 1}Q^{-sm-jm-jz}=\sum_{m\geq 1}\frac{1}{Q^m-1}\cdot\frac{1}{Q^{m+z}-1}. \tag{4.14}$$

Now we turn to the computation of the residues and concentrate first on (4.11). There is a triple pole at $z=0$ and double poles at $z=2k\pi i/L$, $k\in\mathbb{Z}$, $k\neq 0$.

An easy computation gives the local expansion at $z=0$;

$$-[n;z]\frac{Q^z}{(Q^z-1)^2}\sim\frac{1}{L^2z^3}\left(1-\frac{L^2z^2}{12}\right)\left(1+zH_n+z^2\frac{H_n^2+H_n^{(2)}}{2}\right) \tag{4.15}$$

and thus the residue is

$$\frac{1}{L^2}\left(-\frac{L^2}{12}+\frac{H_n^2}{2}+\frac{H_n^{(2)}}{2}\right). \tag{4.16}$$

The computations for the double poles $z=\chi_k$ are similar and we just give the results that are collected into the 2 periodic functions $\delta_1(x)$ and $\delta_2(x)$.

The contribution of the other sum is only $O(n^{-1})$. Using the expansion $H_n\sim\log n+\gamma$, we have

$$g_n\sim p^2\left(\log_Q^2 n+\frac{2\gamma}{L}\log_Q n+\frac{\gamma^2}{L^2}+\frac{\pi^2}{6L^2}-\frac{1}{6}+\log_Q n\cdot\delta_1\left(\log_Q n\right)+\delta_2\left(\log_Q n\right)\right), \tag{4.17}$$

with $\quad\delta_1(x)=-2\delta(x)\quad$ and $\quad\delta_2(x)=\dfrac{2}{L^2}\displaystyle\sum_{k\neq 0}\Gamma'(-\chi_k)\,e^{2k\pi ix}$.

Now, for the variance V_n, we must collect $g_n+E_n-E_n^2$. After some simplifications we obtain

Theorem 4. *The variance V_n of the number of left-to-right maxima (strict model) in the context of n independently distributed geometric random variables has the asymptotic expansion for $n\to\infty$*

$$V_n=pq\log_Q n+p^2\left(-\frac{5}{12}+\frac{\pi^2}{6L^2}-\frac{\gamma}{L}-\left[\delta^2\right]_0\right)+p\left(\frac{\gamma}{L}+\frac{1}{2}\right)+\delta_3(\log_Q n)+O\big(\frac{1}{n}\big) \tag{4.18}$$

Here, $\left[\delta^2\right]_0$ is the mean of the square of $\delta^2(x)$, a very small quantity that can be neglected for numerical purposes. Furthermore, $\delta_3(x)$ is a periodic function with mean 0; its Fourier coefficients could be described if needed.

The function $q\to pq/\log Q$ goes monotonically from 0 to 1 as q varies between 0 and 1.

5. The variance in the loose model

We start from $F(z,y)$ from equation (3.3) and observe that the variance V_n may be obtained by

$$V_n=[z^n]\frac{\partial^2F(z,y)}{\partial y^2}\bigg|_{y=1}+E_n-E_n^2. \tag{5.1}$$

Therefore we consider

$$\frac{\partial^2F(z,y)}{\partial y^2}\bigg|_{y=1}=g(z)+h(z). \tag{5.2}$$

$g(z)$ comprises all terms where different factors are differentiated once and $h(z)$ comprises all terms where one factor is differentiated twice. This gives

$$g(z)=\frac{2p^2}{q^2}\frac{z^2}{1-z}\sum_{1\le i<j}\frac{q^{i+j}}{\big(1-z(1-q^i)\big)\big(1-z(1-q^j)\big)}; \tag{5.3}$$

if we write it as $g(z)=g_1(z)-g_2(z)$ with

$$g_1(z)=\frac{2p^2}{q^2}\frac{z^2}{1-z}\sum_{0\le i<j}\frac{q^{i+j}}{\big(1-z(1-q^i)\big)\big(1-z(1-q^j)\big)} \tag{5.4}$$

and

$$g_2(z)=\frac{2p^2}{q^2}\frac{z^2}{1-z}\sum_{j\ge 1}\frac{q^j}{1-z(1-q^j)} \tag{5.5}$$

we see that $g_1(z)$ was already studied in Section 4 (compare (4.3)) and can concentrate on $g_2(z)$,

$$g_2(z)=\frac{2p^2}{q^2}\sum_{j\ge 1}\left[\frac{q^j}{1-q^j}+\frac{1}{1-z}-\frac{1}{1-q^j}\frac{1}{1-z(1-q^j)}\right]. \tag{5.6}$$

For $n\ge 1$ we have therefore

$$\begin{aligned}
[z^n]g_2(z)&=\frac{2p^2}{q^2}\sum_{j\ge 1}\left[1-\frac{1}{1-q^j}\left(1-q^j\right)^n\right]=\frac{2p^2}{q^2}\sum_{j\ge 1}\left[1-\left(1-q^j\right)^{n-1}\right]\\
&=\frac{2p^2}{q^2}\left[\log_Q(n-1)+\frac{\gamma}{L}-\frac{1}{2}-\delta\big(\log_Q(n-1)\big)\right]+O\big(\frac{1}{n}\big)\\
&=\frac{2p^2}{q^2}\left[\log_Q n+\frac{\gamma}{L}-\frac{1}{2}-\delta(\log_Q n)\right]+O\big(\frac{1}{n}\big)
\end{aligned} \tag{5.7}$$

We can collect;

$$\begin{aligned}
g_n=&\frac{p^2}{q^2}\bigg(\log_Q^2 n+\frac{2\gamma}{L}\log_Q n+\frac{\gamma^2}{L^2}+\frac{\pi^2}{6L^2}-\frac{1}{6}+\log_Q n\cdot\delta_1\left(\log_Q n\right)+\delta_2\left(\log_Q n\right)\bigg)\\
&-\frac{2p^2}{q^2}\bigg(\log_Q n+\frac{\gamma}{L}-\frac{1}{2}-\delta(\log_Q n)\bigg)+O\big(\frac{1}{n}\big).
\end{aligned} \tag{5.8}$$

We also find

$$\begin{aligned}
h(z)&=\frac{2p^2}{q^2}\frac{z^2}{1-z}\sum_{k\ge 1}\frac{q^{2k}}{\big(1-z(1-q^k)\big)^2}\\
&=\frac{2p^2}{q^2}\sum_{k\ge 1}\left[\frac{1}{1-z}-\frac{q^k}{1-q^k}\frac{1}{\big(1-z(1-q^k)\big)^2}+\frac{2q^k-1}{1-q^k}\frac{1}{1-z(1-q^k)}\right]
\end{aligned} \tag{5.9}$$

and therefore

$$\begin{aligned}
h_n&=\frac{2p^2}{q^2}\sum_{k\ge 1}\left[1-\frac{q^k}{1-q^k}(n+1)\left(1-q^k\right)^n+\frac{2q^k-1}{1-q^k}\left(1-q^k\right)^n\right]\\
&=\frac{2p^2}{q^2}\sum_{k\ge 1}\left[1-\left(1-q^k\right)^n-n\frac{q^k}{1-q^k}\left(1-q^k\right)^n\right]
\end{aligned} \tag{5.10}$$

This involves a well known quantity and a new one:

$$\begin{aligned} l_n &:= -n\sum_{j\ge1}\frac{q^j}{1-q^j}\left(1-q^j\right)^n = -n\sum_{j\ge1}q^j\left(1-q^j\right)^{n-1} = -n\sum_{k=0}^{n-1}\binom{n-1}{k}(-1)^k\sum_{j\ge1}q^{j(k+1)} \\ &= \sum_{k=0}^{n-1}\binom{n}{k+1}(k+1)(-1)^{k+1}\frac{1}{Q^{k+1}-1} = \sum_{k=1}^{n}\binom{n}{k}(-1)^k\frac{k}{Q^k-1} \end{aligned} \tag{5.11}$$

Now l_n can be analyzed by Rice's method. At $z=0$ we have $[n;z]\frac{z}{Q^z-1}\sim-\frac{1}{Lz}$; at $z=\chi_k$ we have $[n;z]\frac{z}{Q^z-1}\sim n^{\chi_k}\Gamma(-\chi_k)\frac{\chi_k}{L}\frac{1}{z-\chi_k}$. Therefore

$$l_n = -\frac{1}{L}-\delta_4(\log_Q n)+O\big(\frac{1}{n}\big) \quad \text{with} \quad \delta_4(x)=\frac{1}{L}\sum_{k\ne0}\Gamma(1-\chi_k)e^{2k\pi ix}. \tag{5.12}$$

We can collect the expansions (1.1) and (5.12) to get

$$h_n = \frac{2p^2}{q^2}\left[\log_Q n+\frac{\gamma}{L}-\frac{1}{2}-\frac{1}{L}-\delta(\log_Q n)-\delta_4(\log_Q n)\right]+O\big(\frac{1}{n}\big). \tag{5.13}$$

Now we can sum up for the variance $V_n=g_n+h_n+E_n-E_n^2$. After some simplications we find

Theorem 5. *The variance V_n of the number of left-to-right maxima (loose model) in the context of n independently distributed geometric random variables has the asymptotic expansion for $n\to\infty$*

$$V_n = \frac{p}{q^2}\log_Q n+\frac{p^2}{q^2}\left(-\frac{5}{12}+\frac{\pi^2}{6L^2}+\frac{\gamma}{L}-\frac{2}{L}-\left[\delta^2\right]_0\right)+\frac{p}{q}\left(\frac{\gamma}{L}-\frac{1}{2}\right)+\delta_5(\log_Q n)+O\big(\frac{1}{n}\big). \tag{5.14}$$

Here, $\left[\delta^2\right]_0$ is the mean of the square of $\delta^2(x)$, a very small quantity that can be neglected for numerical purposes. Furthermore, $\delta_5(x)$ is a periodic function with mean 0; its Fourier coefficients could be described if needed.

The function $q\to p/q^2\log Q$ goes monotonically from infinity to 0 as q varies between 0 and 1.

6. Uniform distribution: Left-to-right minima in the strict model

The ideas from Section 2 apply here *mutatis mutandis*. There are the letters $\mathbf{1},\dots,\mathbf{M}$, and each one can occur with probability $\frac{1}{M}$. Thus, let for $k=1,\dots,M$

$$\mathcal{A}_k := \mathbf{k}\{\mathbf{1},\dots,\mathbf{k}\}^*. \tag{6.1}$$

Then

$$\mathcal{L} := \left(\mathcal{A}_1+\varepsilon\right)\cdot\left(\mathcal{A}_2+\varepsilon\right)\dots\left(\mathcal{A}_M+\varepsilon\right) \tag{6.2}$$

is the desired language. Translating it gives the generating function

$$F(z,y)=\prod_{k=1}^{M}\left(1+\frac{\frac{yz}{M}}{1-\frac{zk}{M}}\right). \tag{6.3}$$

Let $f(z) = \left.\dfrac{\partial F(z,y)}{\partial y}\right|_{y=1}$ be generating function for the expected values E_n. We find

$$f(z) = \frac{1}{M}\frac{z}{1-z}\sum_{k=0}^{M-1}\frac{1}{1-\frac{kz}{M}} \tag{6.4}$$

or, by partial fraction decomposition,

$$f(z) = \sum_{k=0}^{M-1}\left[\frac{1}{M-k}\frac{1}{1-z} - \frac{1}{M-k}\frac{1}{1-\frac{kz}{M}}\right]. \tag{6.5}$$

Therefore

$$E_n = [z^n]f(z) = \sum_{k=0}^{M-1}\frac{1}{M-k}\left[1-\left(\frac{k}{M}\right)^n\right] = \sum_{k=1}^{M}\frac{1}{k}\left[1-\left(1-\frac{k}{M}\right)^n\right] = H_M + O\left(\left(\frac{M-1}{M}\right)^n\right). \tag{6.6}$$

Theorem 6. *The expected number E_n of left-to-right maxima (strict model) where each element $1,\dots,M$ can occur with probability $\frac{1}{M}$ is for $n\to\infty$ given by*

$$E_n = H_M + O\left(\left(\frac{M-1}{M}\right)^n\right). \tag{6.7}$$

From this computation we have seen that it is not necessary to determine all terms in the partial fraction decomposition explicitely, since we are not interested to get precise expressions for the exponentially small terms. This principle will save a few unpleasant computations in the sequel. We have

$$\begin{aligned} g(z) = \left.\frac{\partial^2 F(z,y)}{\partial y^2}\right|_{y=1} &= \frac{2}{M^2}\frac{z^2}{1-z}\sum_{0\le i<j<M}\frac{1}{1-\frac{iz}{M}}\cdot\frac{1}{1-\frac{jz}{M}} \\ &= 2\sum_{0\le i<j<M}\left[\frac{1}{(M-i)(M-j)}\frac{1}{1-z}+\dots\right] \end{aligned} \tag{6.8}$$

and therefore

$$g_n = [z^n]g(z) = 2\sum_{0\le i<j<M}\left[\frac{1}{(M-i)(M-j)}+\dots\right]. \tag{6.9}$$

The series may be computed as follows:

$$2\sum_{0\le i<j<M}\frac{1}{(M-i)(M-j)} = 2\sum_{1\le l<k\le M}\frac{1}{kl} = H_M^2 - H_M^{(2)} \tag{6.10}$$

For the variance we must collect the terms $H_M^2 - H_M^{(2)} + H_M - H_M^2$. Hence we have obtained

Theorem 7. *The variance V_n of the number of left-to-right maxima (strict model) where each element $1,\dots,M$ can occur with probability $\frac{1}{M}$ is for $n\to\infty$ given by*

$$V_n = H_M - H_M^{(2)} + O\left(\left(\frac{M-1}{M}\right)^n\right). \tag{6.11}$$

Interestingly enough, for random permutations the corresponding result is $V_n = H_n - H_n^{(2)}$.

7. Uniform distribution: Left-to-right minima in the loose model

Let, for $k = 1, \dots, M$, $\mathcal{A}_k := \mathbf{k}\{\mathbf{1}, \dots, \mathbf{k-1}\}^*$, then $\mathcal{L} := \mathcal{A}_1^* \cdot \mathcal{A}_2^* \dots \mathcal{A}_M^*$ and

$$F(z,y) = \prod_{k=1}^{M} \frac{1}{1 - \frac{\frac{yz}{M}}{1 - z\frac{k-1}{M}}} = \prod_{k=0}^{M-1} \frac{1 - \frac{zk}{M}}{1 - \frac{zk}{M} - \frac{yz}{M}}. \tag{7.1}$$

Let $f(z) = \left.\frac{\partial F(z,y)}{\partial y}\right|_{y=1}$ be the generating function for the expected values E_n. We find

$$f(z) = \frac{1}{M}\frac{z}{1-z}\sum_{k=1}^{M}\frac{1}{1-\frac{kz}{M}} = \frac{1}{M}\frac{z}{(1-z)^2} + \sum_{k=1}^{M-1}\left[\frac{1}{M-k}\frac{1}{1-z} + \dots\right]. \tag{7.2}$$

This gives

$$E_n = \frac{n}{M} + \sum_{k=1}^{M-1}\frac{1}{M-k} = \frac{n}{M} + H_{M-1}. \tag{7.3}$$

Theorem 8. *The expected number E_n of left-to-right maxima (loose model) where each element $1, \dots, M$ can occur with probability $\frac{1}{M}$ is for $n \to \infty$ given by*

$$E_n = \frac{n}{M} + H_{M-1} + O\left(\left(\frac{M-1}{M}\right)^n\right). \tag{7.4}$$

The term $\frac{n}{M}$ has a natural interpretation, as we might expect the appearence of the largest element M about $\frac{n}{M}$ times, and each time we count it. The term H_{M-1} has a similar explanation.

For the variance, we write traditionally $\left.\frac{\partial^2 F(z,y)}{\partial y^2}\right|_{y=1} = g(z) + h(z)$ with

$$g(z) = \frac{2}{M^2}\frac{z^2}{1-z}\sum_{1\le i<j\le M}\frac{1}{1-\frac{iz}{M}}\cdot\frac{1}{1-\frac{jz}{M}} \quad \text{and} \quad h(z) = \frac{2}{1-z}\sum_{k=0}^{M-1}\left(\frac{\frac{z}{M}}{1-\frac{z(k+1)}{M}}\right)^2. \tag{7.5}$$

We must write $g(z) = g_1(z) + g_2(z)$, where the terms for $j = M$ are collected into $g_2(z)$. We have

$$g_1(z) = \sum_{1\le i<j<M}\left[\frac{2}{(M-j)(M-i)}\frac{1}{1-z} + \dots\right] \tag{7.6}$$

and therefore

$$[z^n]g_1(z) = \sum_{1\le i<j<M}\left[\frac{2}{(M-j)(M-i)} + \dots\right] = H_{M-1}^2 - H_{M-1}^{(2)} + \dots. \tag{7.7}$$

Furthermore

$$g_2(z) = \frac{2}{M^2}\frac{z^2}{(1-z)^2}\sum_{1\le i<M}\frac{1}{1-\frac{iz}{M}} = \sum_{1\le i<M}\left[\frac{2}{M(M-i)}\frac{1}{(1-z)^2} - \frac{2(2M-i)}{M(M-i)^2}\frac{1}{1-z}\dots\right], \tag{7.8}$$

and therefore

$$[z^n]g_2(z) = \sum_{1\le i<M}\left[\frac{2}{M(M-i)}(n+1) - \frac{2(2M-i)}{M(M-i)^2}\dots\right] = \frac{2(n+1)H_{M-1}}{M} - \frac{2}{M}\sum_{i=1}^{M-1}\frac{M+i}{i^2}$$
$$= \frac{2(n+1)H_{M-1}}{M} - \frac{2}{M}\left(MH_{M-1}^{(2)} + H_{M-1}\right) = \frac{2nH_{M-1}}{M} - 2H_{M-1}^{(2)}. \tag{7.9}$$

Collecting (7.7) and (7.9) we get

$$g_n = [z^n]g(z) = H_{M-1}^2 - H_{M-1}^{(2)} + \frac{2nH_{M-1}}{M} - 2H_{M-1}^{(2)} = H_{M-1}^2 - 3H_{M-1}^{(2)} + \frac{2nH_{M-1}}{M}. \tag{7.10}$$

Now we turn to $h(z) = \dfrac{2z^2}{M^2(1-z)^3} + 2\sum_{k=1}^{M-1}\left[\dfrac{1}{(M-k)^2}\dfrac{1}{1-z} + \dots\right]$ and thus

$$h_n = [z^n]h(z) = \frac{n(n-1)}{M^2} + 2H_{M-1}^{(2)}. \tag{7.11}$$

Therefore the variance is (up to the exponentially small terms)

$$V_n = H_{M-1}^2 - 3H_{M-1}^{(2)} + \frac{2nH_{M-1}}{M} + \frac{n(n-1)}{M^2} + 2H_{M-1}^{(2)} + \frac{n}{M} + H_{M-1} - \left(\frac{n}{M} + H_{M-1}\right)^2. \tag{7.12}$$

A simplification of this expression leads to

Theorem 9. *The variance V_n of the number of left-to-right maxima (loose model) where each element $1, \dots, M$ can occur with probability $\frac{1}{M}$ is for $n \to \infty$ given by*

$$V_n = n\frac{M-1}{M^2} + H_{M-1} - H_{M-1}^{(2)} + O\left(\left(\frac{M-1}{M}\right)^n\right). \tag{7.13}$$

8. Left-to-right minima in the strict model

In this and the next section we come back to the model of geometric random variables.

It is not hard to establish an appropriate "language", as in Section 2. Let

$$\mathcal{A}_k := \mathbf{k}\{\mathbf{k}, \mathbf{k+1}, \mathbf{k+2}, \dots\}^*, \quad \text{then} \quad \mathcal{L} := \dots(\mathcal{A}_3 + \varepsilon)\cdot(\mathcal{A}_2 + \varepsilon)\cdot(\mathcal{A}_1 + \varepsilon) \tag{8.1}$$

is our language, whence

$$F(z,y) = \prod_{k\ge1}\left(1 + \frac{yzpq^{k-1}}{1-zq^{k-1}}\right) = \prod_{k\ge0}\frac{1-zq^k(1-py)}{1-zq^k}. \tag{8.2}$$

Therefore

$$f(z) = \frac{\partial F(z,y)}{\partial y}\bigg|_{y=1} = \frac{1}{1-z}\sum_{k\ge0}\frac{zpq^k}{1-zq^{k+1}} = \frac{p}{q}\sum_{k\ge1}\frac{q^k}{1-q^k}\left[\frac{1}{1-z} - \frac{1}{1-zq^k}\right] \tag{8.3}$$

and so

$$E_n = [z^n]f(z) = \frac{p}{q}\sum_{k\ge1}\frac{1}{Q^k-1}\left[1 - Q^{-kn}\right]. \tag{8.4}$$

Since the second part of the sum only produces an exponentially small contribution, we have

Theorem 10. *The average number E_n of left-to-right minima in the strict sense is for $n \to \infty$*

$$E_n = \frac{p}{q}\alpha + O(Q^{-n}), \tag{8.5}$$

where, as before, $\alpha = \alpha_Q = \sum_{k\geq 1} \frac{1}{Q^k - 1}$.

Observe that the function $q \to \frac{p}{q}\alpha$ goes monotonically from 1 to infinity as q varies between 0 and 1.

Now let us attack $g(z) = \left.\frac{\partial^2 F(z,y)}{\partial y^2}\right|_{y=1}$. It is easily seen to be

$$g(z) = \frac{2p^2}{q^2}\frac{z^2}{1-z}\sum_{1\leq i<j}\frac{q^{i+j}}{(1-zq^i)(1-zq^j)}. \tag{8.6}$$

Now, since we do not intend to compute the exponentially small terms explicitely, we confine ourselves with the main term $1/(1-z)$ in the partial fraction decomposition.

$$g(z) = \frac{2p^2}{q^2}\sum_{1\leq i<j}\left[\frac{1}{(Q^i-1)(Q^j-1)}\cdot\frac{1}{1-z} + \dots\right] \tag{8.7}$$

and

$$g_n = [z^n]g(z) = \frac{2p^2}{q^2}\sum_{1\leq i<j}\left[\frac{1}{(Q^i-1)(Q^j-1)}\dots\right]. \tag{8.8}$$

The series is easily evaluated and gives, by symmetry, $\frac{\alpha^2-\beta}{2}$, with the usual notation $\beta = \beta_Q = \sum_{k\geq 1}\frac{1}{(Q^k-1)^2}$. So the main term in the variance V_n is $\left(\frac{p}{q}\right)^2\cdot(\alpha^2-\beta) + \frac{p}{q}\alpha - \left(\frac{p}{q}\alpha\right)^2 = \frac{p}{q}\alpha - \left(\frac{p}{q}\right)^2\beta$, and we have

Theorem 11. *The variance V_n of left-to-right minima in the strict sense is for $n \to \infty$*

$$V_n = \frac{p}{q}\alpha - \left(\frac{p}{q}\right)^2\beta + O(Q^{-n}). \tag{8.9}$$

Observe that $q \to \frac{p}{q}\alpha - \left(\frac{p}{q}\right)^2\beta$ goes monotonically from 0 to infinity as q varies between 0 and 1. The following alternative representation for $\frac{p}{q}\alpha - \left(\frac{p}{q}\right)^2\beta$ might be of interest:

$$\sum_{k\geq 0}\frac{1}{1+Q+\dots+Q^k} - \sum_{k\geq 0}\frac{1}{\left(1+Q+\dots+Q^k\right)^2} \tag{8.10}$$

9. Left-to-right minima in the loose model

This time, let $\mathcal{A}_k := \mathbf{k}\{\mathbf{k+1},\mathbf{k+2},\dots\}^*$, then $\mathcal{L} := \dots\mathcal{A}_3^*\cdot\mathcal{A}_2^*\cdot\mathcal{A}_1^*$ and

$$F(z,y) = \prod_{k\geq 1}\frac{1}{1-\frac{yzpq^{k-1}}{1-zq^k}} = \prod_{k\geq 1}\frac{1-zq^k}{1-zq^{k-1}(py+q)}. \tag{9.1}$$

We obtain

$$
\begin{aligned}
f(z) = \frac{\partial F(z,y)}{\partial y}\bigg|_{y=1} &= \frac{1}{1-z}\sum_{k\geq 1}\frac{zpq^{k-1}}{1-zq^{k-1}} \\
&= p\frac{z}{1-z}\sum_{k\geq 0}\frac{q^k}{1-zq^k} = p\frac{z}{(1-z)^2} + p\sum_{k\geq 1}\left[\frac{1}{Q^k-1}\cdot\frac{1}{1-z}+\dots\right],
\end{aligned}
\tag{9.2}
$$

and

$$
E_n = [z^n]f(z) = pn + p\alpha + \dots. \tag{9.3}
$$

Theorem 12. *The average number E_n of left-to-right minima in the loose sense is for $n\to\infty$*

$$
E_n = p(n+\alpha) + O(Q^{-n}). \tag{9.4}
$$

Now for the variance we write, as before, $\displaystyle \frac{\partial^2 F(z,y)}{\partial y^2}\bigg|_{y=1} = g(z)+h(z),$ with

$$
\begin{aligned}
g(z) &= 2p^2\frac{z^2}{1-z}\sum_{0\leq i<j}\frac{q^{i+j}}{(1-zq^i)(1-zq^j)} \\
&= 2p^2\frac{z^2}{(1-z)^2}\sum_{j\geq 1}\frac{q^j}{1-zq^j} + 2p^2\frac{z^2}{1-z}\sum_{1\leq i<j}\frac{q^{i+j}}{(1-zq^i)(1-zq^j)} \\
&= 2p^2\sum_{j\geq 1}\left[\frac{1}{Q^j-1}\cdot\frac{1}{(1-z)^2} - \left(\frac{2}{Q^j-1}+\left(\frac{1}{Q^j-1}\right)^2\right)\frac{1}{1-z}+\dots\right] \\
&+ 2p^2\sum_{1\leq i<j}\left[\frac{1}{(Q^i-1)(Q^j-1)}\frac{1}{1-z}+\dots\right]
\end{aligned}
\tag{9.5}
$$

and

$$
h(z) = 2p^2\frac{z^2}{1-z}\sum_{k\geq 0}\frac{q^{2k}}{(1-zq^k)^2} = 2p^2\frac{z^2}{(1-z)^3} + 2p^2\sum_{k\geq 1}\left[\frac{1}{(Q^k-1)^2}\cdot\frac{1}{1-z}+\dots\right]. \tag{9.6}
$$

Therefore

$$
\begin{aligned}
g_n = [z^n]g(z) &= 2p^2\sum_{j\geq 1}\left[\frac{1}{Q^j-1}(n+1) - \frac{2}{Q^j-1} - \left(\frac{1}{Q^j-1}\right)^2\right] \\
&+ 2p^2\sum_{1\leq i<j}\frac{1}{(Q^i-1)(Q^j-1)} + O(Q^{-n}) = 2p^2\big[(n-1)\alpha-\beta\big] + 2p^2\frac{\alpha^2-\beta}{2} + O(Q^{-n})
\end{aligned}
\tag{9.7}
$$

and

$$
h_n = [z^n]h(z) = 2p^2\binom{n}{2} + 2p^2\beta + O(Q^{-n}). \tag{9.8}
$$

For the variance we must collect $g_n + h_n + E_n - E_n^2$, which is (up to exponentially small terms)

$$
p^2\big[2(n-1)\alpha - 2\beta + \alpha^2 - \beta + n(n-1) + 2\beta\big] + p\big[n+\alpha\big] - p^2\big[n+\alpha\big]^2. \tag{9.9}
$$

Simplifying this expression we find

Theorem 13. *The variance V_n of left-to-right minima in the loose sense is for $n\to\infty$*

$$
V_n = pqn + p\alpha - p^2(2\alpha+\beta) + O(Q^{-n}). \tag{9.10}
$$

This time, the function $q\to pq$ has a maximum at $p=q=\frac{1}{2}$.

10. Conclusions

We made some numerical experiments and obtained good agreement with the predicted theoretical results.

Without going into details, probabilistic counting resembles *coupon collecting* (compare [3]), and we think that there is some work to be done analyzing combinatorial parameters in the context of geometrically distributed random variables. We hope to report on some other problems and results in the future.

Acknowledgments. W. Szpankowski provided additional references. P. Kirschenhofer made interesting remarks. The symbolic computation system Maple was quite helpful.

References

1. L.Devroye, *Applications of the Theory of Records in the Study of Random Trees*, Acta Informatica **26** (1988), 123–130.
2. L.Devroye, *A limit theory for random skip lists*, Annals of Applied Probability **2** (1992), 597–609.
3. P.Flajolet, D.Gardy and L.Thimonier, *Probabilistic languages and random allocations*, Lecture Notes in Computer Science **317** (1988), 239–253.
4. P.Flajolet and G.N.Martin, *Probabilistic Counting Algorithms for Data Base Applications*, J. Comp. Syst. Sci. **31** (1985), 182–209.
5. P.Flajolet and R.Sedgewick, *Digital Search Trees Revisited*, SIAM J. Comput. **15** (1986), 748–767.
6. P.Flajolet and J.Vitter, *Average-Case Analysis of Algorithms and Data Structures*, Handbook of Theoretical Computer Science Vol. A "Algorithms and Complexity", North-Holland, 1990, pp. 431–524.
7. R.Kemp, *Fundamentals of the Average Case Analysis of Particular Algorithms*, Wiley-Teubner, Stuttgart, 1982.
8. P.Kirschenhofer and H.Prodinger, *On the Analysis of Probabilistic Counting*, Lecture Notes in Mathematics (E.Hlawka, R.F.Tichy, eds.), vol. 1452, 1990, pp. 117–120.
9. P.Kirschenhofer and H.Prodinger, *A Result in Order Statistics Related to Probabilistic Counting*, submitted (1992).
10. P.Kirschenhofer,H.Prodinger and W.Szpankowski, *How to Count Quickly and Accurately: A Unified Analysis of Probabilistic Counting and Other Related Problems*, Lecture Notes in Computer Science (W.Kuich, ed.), vol. 623, 1992, pp. 211–222.
11. D.E.Knuth, *The Art of Computer Science Vol.1*, Addison-Wesley, Reading, MA, 1968.
12. D.E.Knuth, *The Art of Computer Science Vol.3*, Addison-Wesley, Reading, MA, 1973.
13. M.Lothaire, *Combinatorics on Words*, Addison Wesley, Reading, MA, 1982.
14. T.Papadakis, I.Munro and P.Poblete, *Average search and update costs in skip lists*, BIT **32** (1992), 316–332.
15. W.Pugh, *Skip lists: a probabilistic alternative to balanced trees*, Communications of the ACM **33** (1990), 668–676.
16. R.Sedgewick, *Mathematical analysis of combinatorial algorithms*, Probability Theory and Computer Science (G.Louchard and G.Latouche, eds.), Academic Press, 1983, pp. 125–205.
17. W.Szpankowski and V.Rego, *Yet another Application of a Binomial Recurrence: Order Statistics*, Computing **43** (1990), 401–410.

TU Vienna
Wiedner Hauptstrasse 8–10
A-1040 Vienna
Austria
E-mail address: proding@email.tuwien.ac.at