2005 International Conference on Analysis of Algorithms DMTCS proc. **AD**, 2005, 231–256

The number of distinct values of some multiplicity in sequences of geometrically distributed random variables

Guy Louchard[1] and Helmut Prodinger[2†] and Mark Daniel Ward[3]

[1] *Université Libre de Bruxelles, Département d'Informatique, CP 212, Boulevard du Triomphe, B-1050 Bruxelles, Belgium.* louchard@ulb.ac.be

[2] *Stellenbosch University, Department of Mathematics, 7602 Stellenbosch, South Africa.* hproding@sun.ac.za

[3] *Purdue University, Department of Mathematics, 150 North University Street, West Lafayette, IN 47907–2067, USA.* mward@math.purdue.edu

We consider a sequence of n geometric random variables and interpret the outcome as an urn model. For a given parameter m, we treat several parameters like what is the largest urn containing at least (or exactly) m balls, or how many urns contain at least m balls, etc. Many of these questions have their origin in some computer science problems. Identifying the underlying distributions as (variations of) the extreme value distribution, we are able to derive asymptotic equivalents for all (centered or uncentered) moments in a fairly automatic way.

Keywords: Distinct values, geometric random variables, extreme value distribution

1 Introduction

Let us consider a sequence of n random variables (RV), $Y_1, \dots, Y_n$, distributed (independently) according to the geometric distribution Geom(p). Set $q := 1 - p$, then $\mathbb{P}(Y = j) = pq^{j-1}$. If we neglect the order in which the n items arrive, we can think about an urn model, with urns labeled $1, 2, \dots$, the probability of each ball falling into urn j being given by pq^{j-1}.

Various questions arise about the distribution of the n balls into these urns. There is a large number of interesting parameters that were studied in the literature, often because of a computer science application. We will give a few examples. The number of the largest nonempty urn ("the maximum" or "the height") was analysed in [18], see also [9]; it is related to a data structure called skip list (see [15]). This is a list-based data structure that one may use instead of search trees.

Another parameter that appears in *probabilistic counting* [4] is the smallest index of a nonempty urn minus 1, or the length of the largest sequence of nonempty urns (starting with urn 1). And, clearly, a parameter that is between those two, is simply the number of nonempty urns. There are also several generalisations around, involving a parameter m (sometimes denoted b or d); like "how many urns are there that contain at least m balls." The instance $m = 1$ refers then to the number of nonempty urns.

We would also like to emphasize that *integer compositions* are closely related to the instance $p = q = \frac{1}{2}$; the probability that digit 1 occurs is $\frac{1}{2}$, that digit 2 occurs is $\frac{1}{4}$, etc. The difference is that the sum of the digits must be n, whereas normally we are interested in n balls (or digits in this case). However, the differences are minor, and we refer to [6] and the references therein.

In the present paper, we extend, generalise, and rederive many known (and unknown) results, using a procedure that we will describe in a minute. As applications, we deal with the 3 parameters described (or variants thereof), under the following assumptions. a) we deal with nonempty urns; b) we deal with urns that contain *exactly* m balls and c) we deal with urns that contain *at least* m balls (which is a generalisation of a). The intuition is as follows, say, for $p = q = \frac{1}{2}$: About $\frac{n}{2}$ balls will go into urn 1, about $\frac{n}{4}$ into urn 2, etc. For a while, every urn will be nonempty (or contain $\geq m$ balls), then there is a sharp transition, and

†This material is based upon work supported by the National Research Foundation under grant number 2053748

then the urns will be empty. So, in the instance b) (exactly m balls in the urn), we can expect to see such urns only in this (small) transition range.

It is that special situation with the sharp transition that makes the analysis of this paper possible. To be more precise, we are dealing here with the *extreme value distribution* and variants thereof. Once a few technical conditions have been checked, the machinery developed in [6] applies, and we get asymptotic forms of all the moments, as well as the centered moments and asymptotic distributions. As it often happens in these type of problems, there are periodic fluctuations (oscillations) involved. The approximations obtained from the extreme value distribution, together with the Mellin transform, establish the fluctuations in the form of Fourier series. After several preliminaries have been discussed, what remains is to a large extent mechanical, and here computer algebra (Maple) comes in.

Several subsections where derivations and reasonings are similar to others, are brief and sketchy, in order not to make this already long paper longer than necessary.

Note that, in [8], Karlin obtained some interesting results on similar topics, including some non-geometric RV.

Here is the plan of the paper: Section 2 sets up the general framework; Section 3 is a continuation of it, dealing with fluctuations. Then we come to the discussion of multiplicities: Section 4 deals with multiplicity at least 1, Section 5 with the number of distinct values (number of urns). Section 6 is concerned with multiplicity at least m, and Section 7 contains a few final remarks.

2 The general setting

We will use the following notations:

$$\begin{aligned}
\sim &:= \text{asymptotically}, \quad n \to \infty, \\
m &:= \text{the fixed multiplicity (an integer value)}, \\
n^* &:= np/q, \\
Q &:= 1/q, \\
L &:= \ln Q, \\
\log &:= \log_Q, \\
\widetilde{\alpha} &:= \alpha/L, \\
B(v,i) &:= \binom{n}{v}(pq^{i-1})^v(1-pq^{i-1})^{n-v}, \\
T(j) &:= \sum_{i=0}^{m-1} B(i,j), \\
g(v,\eta) &:= \exp(-e^{-L\eta})\frac{e^{-Lv\eta}}{v!}, \\
R(j,n) &:= \sum_{i=0}^{m-1} \frac{1}{i!}(n^*q^j)^i \exp(-n^*q^j), \\
g(\eta) &:= \sum_{i=0}^{m-1} g(i,\eta), \\
V(u) &:= \frac{p^u q^{\binom{u}{2}}}{(q)_u}, \\
(q)_l &:= (1-q)(1-q^2)\dots(1-q^l), \\
K &:= (q)_\infty.
\end{aligned}$$

The following facts will be frequently used:

$$\begin{aligned}
&(1-u)^n < e^{-nu}, \quad u \in]0,1[\,, \\
&(1-u)^n = e^{-nu}\left[1 - nu^2/2 + \mathcal{O}(nu^3)\right], \quad u \in]0,1[\,, \\
&(1-u)^n = \left[1 - nu + n(n-1)/2u^2 + \mathcal{O}(nu^3)\right], \quad u \in]0,1/n[\,,
\end{aligned}$$

$$\binom{n}{i}u^i(1-u)^{n-i} = (nu)^i\frac{e^{-nu}}{i!}\left[1+\mathcal{O}(1/n)+\mathcal{O}(u)+\mathcal{O}(nu^2)\right], u\in]0,1[, \quad i \text{ fixed};$$

this is the Poisson approximation.

For all discrete RVs Y_n analyzed in this paper, we set

$$p(j) = \mathbb{P}(Y_n = j), \qquad P(j) := \mathbb{P}(Y_n \le j).$$

We will either set $\eta = j - \log n$ or $\eta = j - \log n^*$, depending on the situation. After all, there is not much difference; only a shift by a constant amount $\log(p/q)$. We will first compute f and F such that

$$p(j) \sim f(\eta), \quad P(j) \sim F(\eta), \quad n \to \infty,$$

and, of course,

$$f(\eta) = F(\eta) - F(\eta - 1).$$

Asymptotically, the distribution will be a periodic function of the fractional part of $\log n$. The distribution $P(j)$ does not converge in the weak sense; it does, however, converge in distribution along subsequences n_m for which the fractional part of $\log n_m$ is constant. For instance such subsequences exist if $Q = n_1^{1/n_2}$, n_1, n_2 integers.

Next, we must show that

$$\mathbb{E}\big(Y_n^k\big) = \sum_{j=1}^{\infty} j^k p(j) \sim \sum_{j=1}^{\infty}(\eta + \log n^*)^k[F(\eta) - F(\eta-1)], \tag{2.1}$$

by computing a suitable rate of convergence. This is related to a uniform integrability condition (see Loève [11, Sec.11.4]).

Finally we will use the following result from Hitczenko and Louchard [6] related to the dominant part of the moments (the '$\widetilde{\ }$' sign is related to the moments of the discrete RV Y_n).

Lemma 2.1 *Let a (discrete) RV Y_n be such that $\mathbb{P}(Y_n - \log n^* \le \eta) \sim F(\eta)$, where $F(\eta)$ is the distribution function of a continuous RV Z with mean m_1, second moment m_2, variance σ^2 and centered moments μ_k. Assume that $F(\eta)$ is either an extreme-value distribution function or a convergent series of such and that (2.1) is satisfied. Let*

$$\varphi(\alpha) = \mathbb{E}(e^{\alpha Z}) = 1 + \sum_{k=1}^{\infty}\frac{\alpha^k}{k!}m_k = e^{\alpha m_1}\lambda(\alpha),$$

say, with

$$\lambda(\alpha) = 1 + \frac{\alpha^2}{2}\sigma^2 + \sum_{k=3}^{\infty}\frac{\alpha^k}{k!}\mu_k.$$

Let w, κ's (with or without subscripts) denote periodic functions of $\log n^*$, with period 1 and with small (usually of order no more than 10^{-6}) mean and amplitude. Actually, these functions depend on the fractional part of $\log n^*$: $\{\log n^*\}$.

Then the mean of Y_n is given by

$$\begin{aligned}\mathbb{E}(Y_n - \log n^*) &\sim \int_{-\infty}^{+\infty} x[F(x) - F(x-1)]dx + w_1\\ &= \widetilde{m}_1 + w_1, \quad \textit{with} \quad \widetilde{m}_1 = m_1 + \tfrac{1}{2}.\end{aligned}$$

More generally, the centered moments of Y_n are asymptotically given by $\widetilde{\mu}_i + \kappa_i$, where

$$\Theta(\alpha) := 1 + \sum_{k=2}^{\infty}\frac{\alpha^k}{k!}\widetilde{\mu}_k = \frac{2}{\alpha}\sinh(\tfrac{\alpha}{2})\lambda(\alpha).$$

The neglected part is of order $1/n^{\beta}$ with $0 < \beta < 1$.

For instance, we derive

$$\widetilde{\mu}_2 = \widetilde{\sigma}^2 = \mu_2 + \frac{1}{12},$$
$$\widetilde{\mu}_3 = \mu_3,$$
$$\widetilde{\mu}_4 = \mu_4 + \frac{\sigma^2}{2} + \frac{1}{80},$$
$$\widetilde{\mu}_5 = \mu_5 + \frac{5}{6}\mu_3.$$

The moments of $Y_n - \log n^*$ are asymptotically given by $\widetilde{m}_i + w_i$, where the generating function of $\widetilde{m}_i$ is given by

$$\phi(\alpha) := \int_{-\infty}^{\infty} e^{\alpha\eta} f(\eta) d\eta = 1 + \sum_{i=1}^{\infty} \frac{\alpha^i}{i!}\widetilde{m}_i = \varphi(\alpha)\frac{e^\alpha - 1}{\alpha}. \tag{2.2}$$

This leads to

$$\widetilde{m}_1 = m_1 + \frac{1}{2},$$
$$\widetilde{m}_2 = m_2 + m_1 + \frac{1}{3},$$
$$\widetilde{m}_3 = m_3 + \frac{3}{2}m_2 + m_1 + \frac{1}{4};$$

w_i and κ_i will be analyzed in the next section. Note that

$$\Theta(\alpha) = \phi(\alpha)e^{-\alpha\widetilde{m}_1}.$$

This leads to

$$\widetilde{\mu}_2 = \widetilde{m}_2 - \widetilde{m}_1^2,$$
$$\widetilde{\mu}_3 = \widetilde{m}_3 + 2\widetilde{m}_1^3 - 3\widetilde{m}_2\widetilde{m}_1.$$

3 The fluctuating components in the moments of $Y_n - \log n^*$

To analyze the periodic component w_i to be added to the moments $\widetilde{m}_i$ we proceed as in Louchard and Prodinger [13]. For instance,

$$\mathbb{E}(Y_n - \log n^*) \sim E^{(1)}(n) = \sum_{j=1}^{\infty}[F(j - \log n^*) - F(j - \log n^* - 1)][j - \log n^*]. \tag{3.1}$$

Set $y = Q^{-x}$ and $G(y) = F(x)$. Equation (3.1) becomes

$$E^{(1)}(n) := \sum_{j=1}^{\infty}[G(n/Q^j) - G(n/Q^{j+1})][-\log(n/Q^j)],$$

the Mellin transform of which is (for a good reference on Mellin transforms, see Flajolet et al. [3] or Szpankowski [17])

$$\frac{Q^s}{1 - Q^s}\Upsilon_1^*(s), \tag{3.2}$$

and

$$\Upsilon_1^*(s) = \int_0^{\infty} y^{s-1}[G(y) - G(y/Q)][-\log y]dy = \int_{-\infty}^{\infty} Q^{-sx}[F(x) - F(x-1)]xLdx.$$

Then

$$\Upsilon_1^*(s) = L\,\phi'(\alpha)\big|_{\alpha=-Ls}. \tag{3.3}$$

The fundamental strip of (3.2) is usually of the form $s \in \langle -C_1, 0\rangle,\ C_1 > 0$. Set also

$$\Upsilon_0^*(s) = L\,\phi(\alpha)\big|_{\alpha=-Ls}, \quad \Upsilon_0^*(0) = L.$$

We assume now that all poles of $\frac{Q^s}{1-Q^s}\Upsilon_1^*(s)$ are simple poles, which will be the case in all our examples, and given by $s=0$, $s=\chi_l$, with $\chi_l := 2l\pi i/L$, $l \in \mathbb{Z}\setminus\{0\}$. Using

$$E^{(1)}(n) = \frac{1}{2\pi i}\int_{C_2-i\infty}^{C_2+i\infty} \frac{Q^s}{1-Q^s}\Upsilon_1^*(s)n^{-s}ds, \quad -C_1 < C_2 < 0,$$

the asymptotic expression of $E^{(1)}(n)$ is obtained by moving the line of integration to the right, for instance to the line $\Re s = C_4 > 0$, taking residues into account (with a negative sign). This gives

$$E^{(1)}(n) = -\mathrm{Res}\Big[\frac{Q^s}{1-Q^s}\Upsilon_1^*(s)n^{-s}\Big]\Big|_{s=0} - \sum_{l\neq 0}\mathrm{Res}\Big[\frac{Q^s}{1-Q^s}\Upsilon_1^*(s)n^{-s}\Big]\Big|_{s=\chi_l} + \mathcal{O}(n^{-C_4}).$$

The residue at $s=0$ gives of course

$$\widetilde{m}_1 = \frac{\Upsilon_1^*(0)}{L} = \phi'(0).$$

The other residues lead to

$$w_1 = \frac{1}{L}\sum_{l\neq 0}\Upsilon_1^*(\chi_l)e^{-2l\pi i \log n}. \tag{3.4}$$

More generally,

$$\mathbb{E}(Y_n - \log n^*)^k \sim \widetilde{m}_k + w_k,$$

with

$$w_k = \frac{1}{L}\sum_{l\neq 0}\Upsilon_k^*(\chi_l)e^{-2l\pi i \log n},$$

and

$$\Upsilon_k^*(s) = L\,\phi^{(k)}(\alpha)\Big|_{\alpha=-Ls}.$$

The residue analysis is similar to the previous one.

To compute the periodic component κ_i to be added to the centered moments $\widetilde{\mu}_i$, we first set

$$\mathfrak{m}_1 := \widetilde{m}_1 + w_1.$$

We start from

$$\phi(\alpha) := 1 + \sum_{k=1}^{\infty}\frac{\alpha^k}{k!}\widetilde{m}_k = \varphi(\alpha)\frac{e^\alpha - 1}{\alpha}.$$

We replace $\widetilde{m}_k$ by $\widetilde{m}_k + w_k$, leading to

$$\phi_p(\alpha) = \phi(\alpha) + \sum_{k=1}^{\infty}\frac{\alpha^k}{k!}w_k.$$

But since $\phi(2l\pi i) = 0$ for all $l \in \mathbb{Z}$, we have

$$\sum_{l\neq 0}\phi(-L\chi_l)e^{-2l\pi i \log n} = 0,$$

and so

$$\begin{aligned}\phi_p(\alpha) &= \phi(\alpha) + \sum_{k=0}^{\infty}\sum_{l\neq 0}\phi^{(k)}(\alpha)\Big|_{\alpha=-L\chi_l}e^{-2l\pi i \log n}\frac{\alpha^k}{k!}\\ &= \phi(\alpha) + \sum_{l\neq 0}\phi(\alpha - L\chi_l)e^{-2l\pi i \log n}\\ &= \sum_{l\in\mathbb{Z}}\phi(\alpha - L\chi_l)e^{-2l\pi i \log n}.\end{aligned} \tag{3.5}$$

Finally, we compute

$$\Theta_p(\alpha) = \phi_p(\alpha)e^{-\alpha \mathfrak{m}_1} = 1 + \sum_{k=2}^{\infty} \frac{\alpha^k}{k!}(\widetilde{\mu}_k + \kappa_k) = \Theta(\alpha) + \sum_{k=2}^{\infty} \frac{\alpha^k}{k!}\kappa_k, \tag{3.6}$$

leading to the (exponential) generating function (GF) of κ_k. This leads to

$$\begin{aligned}
\kappa_2 &= w_2 - w_1^2 - 2\widetilde{m}_1 w_1, \\
\kappa_3 &= 6\widetilde{m}_1^2 w_1 + 6\widetilde{m}_1 w_1^2 + 2w_1^3 - 3\widetilde{m}_2 w_1 - 3\widetilde{m}_1 w_2 - 3w_1 w_2 + w_3.
\end{aligned}$$

All algebraic manipulations of this paper will be mechanically performed by Maple. We will give explicit expressions for $\widetilde{\mu}_2$, κ_2, $\widetilde{\mu}_3$ and κ_3.

It will appear that $\Upsilon_k^*(s)$ are analytic functions (in some domain), depending on classical functions such as Euler's Γ function. But we know that $\Gamma(s)$ decreases exponentially towards $\pm i\infty$:

$$|\Gamma(\sigma + it)| \sim \sqrt{2\pi}|t|^{\sigma-1/2}e^{-\pi|t|/2}. \tag{3.7}$$

and all our functions will also decrease exponentially towards $\pm i\infty$.

4 Multiplicity at least 1

As in Hitczenko and Louchard [6] (where the case $p = 1/2$ is analyzed), we can check that, asymptotically, the urns become independent.

Set the indicator RV (in the sequel we drop the n-specification to simplify the notations):[‡]

$$X_i := [\![\text{value } i \text{ appears among the } n \text{ RVs}]\!].$$

4.1 Maximal non-empty urn

The maximal full urn index

$$M := \sup\{i : X_i = 1\}$$

is such that

$$P(j) := \mathbb{P}[M \le j] = (1 - q^j)^n \sim e^{-nq^j}.$$

With $\eta = j - \log n$, we obtain

$$P(j) \sim F(\eta),$$

with

$$F(\eta) = \exp(-e^{-L\eta}).$$

This is exactly the same behaviour as in the *trie case*, analyzed in Louchard and Prodinger [13, Section 4.1], where the rate of convergence is already computed. We note that the distribution is concentrated on the range $\eta = \mathcal{O}(1)$, i. e., in the concentration domain $j = \log n + \mathcal{O}(1)$.

From [13, Section 5.1], we derive the moments of $M - \log n$:

$$\begin{aligned}
\widetilde{m}_1 &= \frac{\gamma}{L} + \frac{1}{2}, \\
\widetilde{\mu}_2 &= \frac{\pi^2}{6L^2} + \frac{1}{12}, \\
\widetilde{\mu}_3 &= \frac{2\zeta(3)}{L^3}.
\end{aligned}$$

where γ is Euler's gamma constant. Let us now turn to the fluctuating components. We have here $\varphi(\alpha) = \Gamma(1 - \widetilde{\alpha})$. The fundamental strip for s is $\Re(s) \in \langle -1, 0\rangle$. First of all, (3.3) and (3.4) lead to

$$w_1 = -\frac{1}{L}\sum_{l \ne 0} \Gamma(\chi_l)e^{-2l\pi i \log n}.$$

[‡] Here we use the indicator function ('Iverson's notation') proposed by Knuth et al. [5].

Next we obtain

$$\kappa_2 = -2\frac{\gamma w_1}{L} - w_1^2 + \frac{2}{L^2}\sum_{l\neq 0}\Gamma(\chi_l)\psi(\chi_l)e^{-2l\pi i\log n},$$

$$\kappa_3 = -6\Big(\frac{w_1}{L^2}+\frac{\gamma}{L^3}\Big)\sum_{l\neq 0}\Gamma(\chi_l)\psi(\chi_l)e^{-2l\pi i\log n} - \frac{3}{L^3}\sum_{l\neq 0}\Gamma(\chi_l)\psi^2(\chi_l)e^{-2l\pi i\log n}$$
$$-\frac{3}{L^3}\sum_{l\neq 0}\Gamma(\chi_l)\psi(1,\chi_l)e^{-2l\pi i\log n} + 2w_1^3 + \frac{(6\gamma^2-\pi^2)w_1}{2L^2} + \frac{6\gamma w_1^2}{L},$$

where ψ is the digamma function (logarithmic derivative of the Γ function) and $\psi(1,x)$ is the trigamma function.

4.2 Number of distinct values

This is actually a measure of distinctness. Set $X := \sum_{i=1}^{\infty} X_i$.

We can now proceed as in Louchard and Prodinger [13, Section 5.8]. We don't consider the rate of convergence here: this will be computed in Section 6.1, in a more general setting. Note that

$$\mathbb{P}(X_i = 0) = (1-pq^{i-1})^n \sim e^{-npq^{i-1}} = e^{-n^* q^i}$$

if we set, as always, $n^* = np/q$.

This leads, for the moments of $X - \log n^*$, to

$$\begin{aligned}\widetilde{m}_1 &= \frac{\gamma}{L} - \frac{1}{2},\\ w_1 &= \beta_{1,1},\\ \widetilde{\mu}_2 &= \log 2,\\ \widetilde{\mu}_3 &= -3\log 2 + 2\log 3,\\ \kappa_2 &= \beta_{1,2} - \beta_{1,1},\\ \kappa_3 &= 2\beta_{1,3} - 3\beta_{1,2} + \beta_{1,1},\end{aligned}$$

with

$$\beta_{1,k} = -\frac{1}{L}\sum_{l\neq 0}\Gamma(\chi_l)e^{-2l\pi i\log(n^* k)}.$$

Note the presence of k in the exponent. Note also that the variance has here a periodic component, contrariwise to the case $p = \frac{1}{2}$: We have that $\kappa_2(x) = \beta_{1,1}(x+\log 2) - \beta_{1,1}(x)$, and this is zero for $Q = 2$, because of the periodicity 1. The first two moments are given in Archibald, Knopfmacher and Prodinger [1]; the cancellation for $p = q = \frac{1}{2}$ was noticed therein, see also [14]. In [8], Karlin mentions that the mean of X "could oscillate irregularly," but does not give an expression, even in the geometric case. In his Theorem $1'$, he provides the $\log n$ dominant term of $\mathbb{E}(X)$, and in Section 6.III, he gives $\widetilde{\mu}_2$, $\widetilde{\mu}_3$, mentioning that "the distribution of X is difficult to identify."

Actually, the asymptotic distribution of X can be adapted from Hitczenko and Louchard [6]. We obtain the following result:

Theorem 4.1 *Set $\eta := j - \log n^*$ and*

$$\Psi_1(\eta) := e^{-e^{-L\eta}}\prod_{i=1}^{\infty}\Big[1 - e^{-e^{-L(\eta-i)}}\Big].$$

Then, with $j \in \mathbb{Z}$ and $\eta = \mathcal{O}(1)$,

$$\mathbb{P}(X = j) \sim f(\eta) = \sum_{u=0}^{\infty}\Psi_1(\eta-u+1)e^{-e^{-L(\eta+1-u)}/(Q-1)}\sum_{\substack{r_1<\cdots<r_u\\ r_j\geq 2-u}}\prod_{i=1}^{u}\frac{1-e^{-e^{-L(\eta+r_i)}}}{e^{-e^{-L(\eta+r_i)}}},$$

$$\mathbb{P}(X \leq j) \sim F(\eta), \qquad \text{with} \qquad F(\eta) := \sum_{i=0}^{\infty} f(\eta - i).$$

4.3 First empty urn

Set $E := \inf\{i : X_i = 0\}$.

Again, we start from [13, Section 4.8]. Setting

$$A_1(j) := \prod_{i=1}^{j}[1-(1-pq^{i-1})^n] \le 1,$$

we obtain $P(j) \sim 1 - A_1(j)$. We have

$$\begin{aligned}
\eta &= j - \log n^*,\\
p(j) &\sim \left(1-pq^{j-1}\right)^n A_1(j-1),\\
\Psi_2(\eta) &:= \prod_{k=0}^{\infty}\left[1-\exp(-e^{-L(\eta-k)})\right],\\
F(\eta) &:= 1-\Psi_2(\eta),\\
f(\eta) &:= \Psi_1(\eta).
\end{aligned}$$

The rate of convergence is fully analyzed in [13] in the case $p = 1/2$. The analysis is similar here. Also, in this case $p = 1/2$, from [13, Section 5.9.1], we first define the entire function $N(s)$ which is the analytic continuation of

$$\sum_{j\ge 1}\frac{(-1)^{\nu(j)}}{j^s},$$

where $\nu(j)$ denotes the number of ones in the binary representation of j. This gives

$$\begin{aligned}
N(0) &= -1,\\
N'(0) &= -.4874506\ldots,\\
N''(0) &= .8433214\ldots,\\
N'''(0) &= -.8683385\ldots.
\end{aligned}$$

We obtain the moments of $E - \log n^*$ for $p = 1/2$:

$$\begin{aligned}
\widetilde{m}_1 &= \frac{\gamma+N'(0)}{L}+\frac{1}{2},\\
\widetilde{\mu}_2 &= \frac{1}{6L^2}\big(-6N'(0)^2+\pi^2-6N''(0)\big)+\frac{1}{12},\\
\widetilde{\mu}_3 &= (2N'(0)^3+3N''(0)N'(0)+N'''(0)+\frac{2\zeta(3)}{L^3}.
\end{aligned}$$

Let us now turn to the fluctuating components:

$$\begin{aligned}
w_1 &= \frac{1}{L}\sum_{l\ne 0}N(\chi_l)\Gamma(\chi_l)e^{-2l\pi i\log n^*},\\
\kappa_2 &= -w_1^2-\frac{2}{L^2}\sum_{l\ne 0}\big[N(\chi_l)(\gamma+N'(0))+N'(\chi_l)+N(\chi_l)\psi(\chi_l)\big]\Gamma(\chi_l)e^{-2l\pi i\log n^*}.
\end{aligned}$$

Next we obtain

$$\begin{aligned}
\kappa_3 = \sum_{l\ne 0}\Big\{&3\psi(1,\chi_l)N(\chi_l)/L^3+\psi(\chi_l)[6N(\chi_l)w_1/L^2+6N'(\chi_l)/L^3+6N(\chi_l)(\gamma+N'(0))/L^3]\\
&+3N(\chi_l)\psi^2(\chi_l)/L^3+3\big(2(N'(0)+\gamma)N'(\chi_l)+N''(\chi_l)\big)/L^3\Big\}\Gamma(\chi_l)e^{-2l\pi i\log n^*}\\
&+6w_1\sum_{l\ne 0}N'(\chi_l)\Gamma(\chi_l)e^{-2l\pi i\log n^*}/L^2
\end{aligned}$$

$$-\frac{w_1}{2L^2}\big(6\gamma^2+12\gamma N'(0)-6N''(0)+L^2+\pi^2\big)-\frac{6w_1^2}{L}\big(\gamma+N'(0)\big)-4w_1^3.$$

In the case $p\neq 1/2$, we follow the lines of Sections 2,3, and we define (we have no explicit form here)

$$\varphi(\alpha):=\int_{-\infty}^{\infty}e^{\alpha\eta}F'(\eta)d\eta=-\alpha\int_{-\infty}^{\infty}e^{\alpha\eta}F(\eta)d\eta.$$

This leads to

$$\begin{aligned}
\widetilde{m}_1&=\frac{1}{2}+\varphi'(0),\\
w_1&=-\sum_{l\neq 0}\frac{\varphi(-L\chi_l)}{L\chi_l}e^{-2l\pi i\log n^*},\\
\widetilde{\mu}_2&=\frac{1}{12}-\varphi'(0)^2+\varphi''(0),\\
\widetilde{\mu}_3&=2\varphi'(0)^3-3\varphi''(0)\varphi'(0)+\varphi'''(0),\\
w_2&=-\sum_{l\neq 0}\left[2\frac{\varphi'(-L\chi_l)}{L\chi_l}+\frac{\varphi(-L\chi_l)}{L\chi_l}+2\frac{\varphi(-L\chi_l)}{L^2\chi_l^2}\right]e^{-2l\pi i\log n^*},\\
w_3&=-\sum_{l\neq 0}\left[3\frac{\varphi''(-L\chi_l)}{L\chi_l}+3\frac{\varphi'(-L\chi_l)}{L\chi_l}+6\frac{\varphi'(-L\chi_l)}{L^2\chi_l^2}\right.\\
&\qquad\left.+\frac{\varphi(-L\chi_l)}{L\chi_l}+3\frac{\varphi(-L\chi_l)}{L^2\chi_l^2}+6\frac{\varphi(-L\chi_l)}{L^3\chi_l^3}\right]e^{-2l\pi i\log n^*},\\
\kappa_2&=-w_1-2\varphi'(0)w_1-w_1^2+w_2,\\
\kappa_3&=\tfrac{1}{2}w_1-\tfrac{3}{2}w_2+w_3+3\varphi'(0)w_1+3w_1^2+6\varphi'(0)^2w_1\\
&\qquad+6\varphi'(0)w_1^2-3\varphi''(0)w_1-3\varphi'(0)w_2+2w_1^3-3w_1w_2.
\end{aligned}$$

Alternatively, we could start from

$$\phi(\alpha):=\int_{-\infty}^{\infty}e^{\alpha\eta}f(\eta)d\eta.$$

5 Multiplicity exactly m

We consider fixed $m=\mathcal{O}(1)$. Set

$$X_i(n):=[\![\text{value } i \text{ appears among the } n \text{ GEOM}(p) \text{ RVs with multiplicity } m]\!].$$

Then $\mathbb{P}[X_j(n)=1]=B(m,j)$, with

$$B(m,j):=\binom{n}{m}(pq^{j-1})^m(1-pq^{j-1})^{n-m}.$$

We immediately see that the dominant range is given by $j=\log n+\mathcal{O}(1)$. To the left and the right of this range, $\mathbb{P}[X_j(n)=1]\sim 0$. Within and to the right of this range, $\mathbb{P}[X_j(n)=1]$ is asymptotically equivalent to a Poisson distribution:

$$\mathbb{P}[X_j(n)=1]\sim\frac{1}{m!}(n^*q^j)^m\exp(-n^*q^j).\tag{5.1}$$

Setting again $\eta:=j-\log n^*$, we derive $\mathbb{P}[X_j(n)=1]\sim g(m,\eta)$, with

$$g(m,\eta):=\exp(-e^{-L\eta})\frac{e^{-Lm\eta}}{m!}.$$

5.1 Number of distinct values

Set $X(n) := \sum_{i=1}^{\infty} X_i(n)$. We must first check the asymptotic independency of the urns. Let us consider $\Pi_n(z) = \mathbb{E}(z^{X(n)})$. We are interested in the behaviour of $\Pi_n(z)$ for complex $z \in \overline{D}_\epsilon(1) = \{t \mid |t-1| \leq \epsilon\}$, where ϵ is a small fixed positive real number. We choose $\epsilon > 0$ such that $c := \log(1+\epsilon) < 1$.

Theorem 5.1 *We have*

$$\Pi_n(z) = \prod_{l=1}^{\infty} \left[\left(1 - \frac{1}{m!}(n^* q^l)^m e^{-n^* q^l}\right) + z \frac{1}{m!}(n^* q^l)^m e^{-n^* q^l} \right] + O(n^{c-1}), \quad n \to \infty,$$

uniformly for $z \in \overline{D}_\epsilon(1)$*, where* $0 < c = \log(1+\epsilon) < 1$.

Proof
We use an urn model, as in Sevastyanov and Chistyakov [16] and Chistyakov [2], and the Poissonization method (see, for instance Jacquet and Szpankowski [7] for a general survey). In the above formulation, we have a *fixed* number n of geometric random variables, each corresponding to a ball. The value of each RV denotes the bin into which the ball is placed. For instance, if $Y_1 = 3$, then the first ball is placed into the third bin.

In order to utilize the Poissonization method, instead of using a *fixed* number of balls, we use N balls, where N is a Poisson random variable with $\mathbb{E}(N) = \tau$. It follows that the urns are *independent*, and the number of balls in urn l is a Poisson random variable with parameter $\tau^* q^l$. We use a "$\widetilde{\ }$" to denote that we are working in the Poissonized model. For instance, $\widetilde{X}_l(\tau)$ denotes the lth GEOM(p) RV in the Poissonized model, i.e., $\widetilde{X}_l(\tau)$ corresponds to $X_l(n)$. It follows that urn l has exactly m balls with probability $\frac{1}{m!}(\tau^* q^l)^m e^{-\tau^* q^l}$. So the generating function of $\widetilde{X}_l(\tau)$ is

$$\mathbb{E}(z^{\widetilde{X}_l(\tau)}) = \left(1 - \frac{1}{m!}(\tau^* q^l)^m e^{-\tau^* q^l}\right) + z \frac{1}{m!}(\tau^* q^l)^m e^{-\tau^* q^l} = 1 + (z-1)\frac{1}{m!}(\tau^* q^l)^m e^{-\tau^* q^l}.$$

We have $\widetilde{X}(\tau) = \sum_{l=1}^{\infty} \widetilde{X}_l(\tau)$, and thus

$$G(\tau, z) := \mathbb{E}(z^{\widetilde{X}(\tau)}) = \mathbb{E}(z^{\sum \widetilde{X}_l(\tau)}).$$

Since the urns are independent in the Poissonized model, then $\mathbb{E}(z^{\sum \widetilde{X}_l(\tau)}) = \prod_{l=1}^{\infty} \mathbb{E}(z^{\widetilde{X}_l(\tau)})$. Thus

$$G(\tau, z) = \prod_{l=1}^{\infty} \left[1 + (z-1)\frac{1}{m!}(\tau^* q^l)^m e^{-\tau^* q^l}\right]. \tag{5.2}$$

We write $\tau = Re^{it}$ for real $R \geq 0$ and $-\pi < t \leq \pi$. Thus $|\tau| = R$. We denote the linear cone containing all τ with $-\pi/4 \leq t \leq \pi/4$ as $\mathcal{S}_{\pi/4} = \{\tau = Re^{it} \mid -\pi/4 \leq t \leq \pi/4\}$. Now we derive asymptotics about the growth of $|G(\tau, z)|$ for $\tau \in \mathcal{S}_{\pi/4}$. Our estimates are valid *uniformly* for $z \in \overline{D}_\epsilon(1) = \{t \mid |t-1| \leq \epsilon\}$. We encapsulate our results in the following lemma.

Lemma 5.2 *For* $\tau \in \mathcal{S}_{\pi/4}$*, there exist reals* $B > 0$*,* $R_0 > 0$*, and* $0 < c < 1$*, such that if* $|\tau| = R > R_0$ *then*

$$|G(\tau, z)| \leq B|\tau|^c$$

uniformly for $z \in D_\epsilon(1)$.

Proof We first consider $l \geq 1 + \log R$. We have

$$\begin{aligned}
\prod_{l \geq 1+\log R} \left|\mathbb{E}(z^{\widetilde{X}_l(\tau)})\right| &= \prod_{l \geq 1+\log R} \left|1 + (z-1)\frac{1}{m!}(\tau^* q^l)^m e^{-\tau^* q^l}\right| \\
&\leq \prod_{l \geq 1+\log R} \left[1 + \epsilon \frac{1}{m!}(|\tau| p q^{l-1})^m e^{-\Re(\tau) p q^{l-1}}\right] \\
&= \prod_{l \geq \log R} \left[1 + \epsilon \frac{1}{m!}(|\tau| p q^l)^m e^{-\Re(\tau) p q^l}\right] \\
&= \exp\left(\sum_{l \geq \log R} \ln\left[1 + \epsilon \frac{1}{m!}(|\tau| p q^l)^m e^{-\Re(\tau) p q^l}\right]\right)
\end{aligned}$$

$$\leq \exp\left(\sum_{l\geq \log R}\left[\epsilon\frac{1}{m!}(|\tau|pq^l)^m e^{-\Re(\tau)pq^l}\right]\right) \tag{5.3}$$

where the inequality holds since $\ln(1+x)\leq x$ for real x. We note that $-\Re(\tau)<0$ since $\tau\in\mathcal{S}_{\pi/4}$. Thus $e^{-\Re(\tau)pq^l}\leq 1$. It follows that

$$\begin{aligned}
\prod_{l\geq 1+\log R}\left|\mathbb{E}(z^{\widetilde{X}_l(\tau)})\right| &\leq \exp\left(\epsilon\frac{1}{m!}(|\tau|p)^m\sum_{l\geq\log R}(q^m)^l\right)\\
&\leq \exp\left(\epsilon\frac{1}{m!}\frac{(|\tau|pq^{\log R})^m}{1-q^m}\right)\\
&\leq \exp\left(\epsilon\frac{1}{m!}\frac{p^m}{1-q^m}\right) \qquad \text{since } q^{\log R}=R^{-1}=|\tau|^{-1}\\
&=\mathcal{O}(1) \qquad\qquad (5.4)
\end{aligned}$$

Now we consider $l\leq\log R$. We have

$$\prod_{l\leq\log R}\left|\mathbb{E}(z^{\widetilde{X}_l(\tau)})\right|\leq\prod_{l\leq\log R}(1+\epsilon)\leq(1+\epsilon)^{\log R}=R^{\log(1+\epsilon)}=|\tau|^c.$$

Combining these results, we have $|G(\tau,z)|=\mathcal{O}(1)|\tau|^c=\mathcal{O}(|\tau|^c)$ uniformly for $z\in D_\epsilon(1)$. This completes the proof of the lemma. ■

We return to the proof of Theorem 5.1. The lemma we just completed shows that condition (I) holds for Theorem 10.3 of [17]. Now we prove that condition (O) of Theorem 10.3 of [17] holds too, namely: for $\tau\notin\mathcal{S}_{\pi/4}$, there exist A and $\alpha<1$ such that $|G(\tau,z)e^\tau|\leq A\exp(\alpha|\tau|)$ for $|\tau|>R_0$.

First consider $\tau\notin\mathcal{S}_{\pi/4}$ with $\Re(\tau)\geq 0$. Then the same proof given in the lemma above shows that $|G(\tau,z)|=\mathcal{O}(|\tau|^c)$ uniformly for $z\in D_\epsilon(1)$. Thus $|G(\tau,z)e^\tau|=\mathcal{O}(|\tau|^c e^{\Re(\tau)})$, and $\Re(\tau)\leq|\tau|/\sqrt{2}$ for these τ's, so by setting $\alpha=1/\sqrt{2}$, we conclude that condition (O) holds for $\tau\notin\mathcal{S}_{\pi/4}$ with $\Re(\tau)\geq 0$.

Now we consider τ with $\Re(\tau)<0$. By (5.3), we see that

$$\prod_{l\geq 1+\log R}\left|\mathbb{E}(z^{\widetilde{X}_l(\tau)})\right|\leq\exp\left(\sum_{l\geq\log R}\left[\epsilon\frac{1}{m!}(|\tau|pq^l)^m e^{-\Re(\tau)pq^l}\right]\right).$$

Note that $e^{-\Re(\tau)pq^l}\leq e^{-\Re(\tau)pR^{-1}}=e^{-p\Re(\tau)/|\tau|}\leq e^p$ for all τ with $\Re(\tau)<0$ and all l's with $l\geq 1+\log R$. So, proceding with reasoning similar to (5.4), we again see that

$$\prod_{l\geq 1+\log R}\left|\mathbb{E}(z^{\widetilde{X}_l(\tau)})\right|=O(1).$$

Also $\prod_{l\leq\log R}\left|\mathbb{E}(z^{\widetilde{X}_l(\tau)})\right|=|\tau|^c$ as before. So $|G(\tau,z)e^\tau|=\mathcal{O}(|\tau|^c e^{\Re(\tau)})=\mathcal{O}(|\tau|^c)$ since $\Re(\tau)<0$. Thus, any α with $0<\alpha<1$ is sufficient to satisfy condition (O) when $\Re(\tau)<0$.

We conclude that $\alpha=1/\sqrt{2}$ is sufficient to satisfy condition (O) when $\tau\notin\mathcal{S}_{\pi/4}$.

Therefore, conditions (I) and (O) of Theorem 10.3 of [17] are all satisfied, so we can depoissonize our results, i.e., $\Pi_n(z)$ and $G(\tau,z)$ have the same asymptotics. More precisely,

$$\Pi_n(z)=G(n,z)+O(n^{c-1}).$$

Substituting $\tau=n$ within (5.2), we see that

$$G(n,z)=\prod_{l=1}^{\infty}\left[\left(1-\frac{1}{m!}(n^*q^l)^m e^{-n^*q^l}\right)+z\frac{1}{m!}(n^*q^l)^m e^{-n^*q^l}\right],$$

and we note that $0<c<1$, so this completes the proof of Theorem 5.1. ■

Theorem 5.1 confirms the asymptotic independence assumption.

The moments can be derived as follows. We obtain, setting $z = e^s$,

$$\begin{aligned}\ln(\Pi_n) \sim S_2(s) &= \sum_{l=1}^{\infty} \ln\left[1 + (e^s - 1)B(m,l)\right] \\ &= \sum_{i=1}^{\infty} \frac{(-1)^{i+1}(e^s-1)^i V_i}{i}, \text{ with} \\ V_i &:= \sum_{l=1}^{\infty} \left[B(m,l)\right]^i .\end{aligned}$$

Let us first check that we can replace the Binomial by a Poisson distribution (see (5.1)) by computing a suitable rate of convergence. We will consider three ranges. Let $1/2 < \beta < 1$.

- For $j < \beta \log n^*$, $B(m,j)^k$ is small. Indeed

$$B(m,j)^k \leq \left[n^{*\,m} \exp(-n^{*\,1-\beta})/m!\right]^k .$$

- For $\beta \log n^* \leq j < 2 \log n^*$ we have

$$\begin{aligned}B(m,j)^k - g(m,\eta)^k &\sim \left[g(m,\eta)[1 + \mathcal{O}(1/n^*) + \mathcal{O}(1/Q^j) + \mathcal{O}(n^*/Q^{2j})]\right]^k - g(m,\eta)^k \\ &= \mathcal{O}(1/n^{*\,2\beta-1}).\end{aligned}$$

- For $j = 2\log n^* + x$, $x \geq 0$, we have

$$\begin{aligned}B(m,j)^k - g(m,\eta)^k &\sim \left[g(m,\eta)[1 + \mathcal{O}(1/n^*) + \mathcal{O}(1/n^{*\,2}) + \mathcal{O}(n^*/n^{*\,4})]\right]^k - g(m,\eta)^k \\ &= \mathcal{O}\left[1/(n^{*\,m}Q^{mx})\right]^k/n^*,\end{aligned}$$

  as

$$g(i,\eta) = \mathcal{O}[1/(n^*Q^x)^i].$$

Now we must bound

$$\Big|\sum_j [B(m,j)^k - g(m,\eta)^k]\Big|$$

which leads immediately to $\mathcal{O}(1/n^{*\,2\beta-1-\varepsilon})$, ε small > 0.

So V_i is given by a harmonic sum, which we will compute by the Mellin transform. Set $y = Q^{-\eta}$ and

$$g(y) := \left[g(m,\eta)\right]^k ,$$

the Mellin transform of which is

$$g^*(s) = \frac{\Gamma(mk+s)}{k^{mk+s}(m!)^k}.$$

This leads to

$$g^*(s)\frac{Q^s}{1-Q^s},$$

with fundamental strip $\Re(s) \in \langle -mk, 0\rangle$. We obtain, by residues,

$$V_i \sim B_i + \beta_i(\log n),$$

with

$$\begin{aligned}B_k &= \frac{(km-1)!}{m!^k L k^{km}}, \\ \beta_k &= \sum_{l\neq 0} \frac{\Gamma(\chi_l + mk)}{Lk^{mk}(m!)^k} e^{-2l\pi i \log(n^* k)}.\end{aligned}$$

Note again the presence of k in the exponent.

The centered moments of X can be obtained by analyzing

$$S_3(s) := \exp(S_2(s) - sV_1);$$

and finally, the moments are given by

$$\begin{aligned}
\widetilde{m}_1 &= B_1, \\
w_1 &= \beta_1, \\
\widetilde{\mu}_2 &= B_1 - B_2, \\
\widetilde{\mu}_3 &= B_1 - 3B_2 + 2B_3, \\
\kappa_2 &= \beta_1 - \beta_2, \\
\kappa_3 &= \beta_1 - 3\beta_2 + 2\beta_3.
\end{aligned}$$

The asymptotic distribution of X can be derived from Louchard [12]. This leads with

$$\begin{aligned}
\Psi_3(\eta) &= g(m,\eta) \prod_{j=1}^{\infty} \left[1 - g(m, \eta - j)\right], \\
\Psi_4(\eta) &= \prod_{j=0}^{\infty} \left[1 - g(m, \eta + j)\right]
\end{aligned}$$

to the following result.

Theorem 5.3 *Set $\psi(n^*) := \log n^* - \lfloor \log n^* \rfloor$, then*

$$\mathbb{P}(X = u+1) \sim \sum_{l=-\infty}^{\infty} \Psi_5\big(l - \psi(n^*)\big),$$

with

$$\Psi_5(\eta) = \Psi_3(\eta - 1)\Psi_4(\eta) \sum_{w_1 > w_2 > \cdots > w_u \geq 0} \prod_{i=1}^{u} \left\{ g(m, \eta + w_i) \Big/ \Big[1 - g(m, \eta + w_i)\Big] \right\}.$$

Note that, contrariwise to the previous section, the RV X is here $\mathcal{O}(1)$ in the sense that we do not have to normalize by $\log n^*$.

5.2 Maximal non-empty urn

We derive, by asymptotic urn independence,

$$\begin{aligned}
p(j) := \mathbb{P}(M = j) &\sim B(m,j) \prod_{i=j+1}^{\infty} [1 - B(m,i)], \\
P(j) &\sim \prod_{i=j+1}^{\infty} [1 - B(m,i)].
\end{aligned}$$

This leads to

$$\begin{aligned}
p(j) &\sim f(\eta) = g(m,\eta)\Psi_4(\eta+1), \\
P(j) &\sim F(\eta) = \Psi_4(\eta+1).
\end{aligned}$$

We have here product forms: the rate of convergence for this kind of asymptotics is fully detailed in Louchard and Prodinger [13]. We can now proceed as in Section 4.3.

5.3 First full urn

Set $E := \inf\{i : X_i = 1\}$.

Note the difference with Section 4.3, where we were concerned by the first empty urn; that question would not make sense here since the first 'empty' urn ($\neq m$ elements) would be urn 1 with very high probability.

We obtain

$$p(j) \sim B(m,j) \prod_{i=1}^{j-1} [1 - B(m,i)],$$

$$P(j) \sim 1 - \prod_{i=1}^{j} [1 - B(m,i)].$$

This leads to

$$p(j) \sim f(\eta) = \Psi_3(\eta),$$

$$P(j) \sim F(\eta) = 1 - \prod_{j=0}^{\infty} [1 - g(m, \eta - j)] .$$

We can now proceed as in Section 4.3. We don't give more details here.

6 Multiplicity at least m

We again consider fixed $m = \mathcal{O}(1)$. Set

$$X_i(n) := [\![\text{value } i \text{ appears among the } n \text{ GEOM}(p) \text{ RV with multiplicity at least } m]\!].$$

We have

$$\mathbb{P}[X_j(n) = 0] = T(j) := \sum_{i=0}^{m-1} B(i,j).$$

Again, in the range given by $j \geq \log n^*$ we can use the Poisson approximation:

$$\mathbb{P}[X_j(n) = 1] \sim 1 - R(j,n), \tag{6.1}$$

with

$$R(j,n) := \sum_{i=0}^{m-1} \frac{1}{i!} (n^* q^j)^i \exp(-n^* q^j).$$

Setting again $\eta := j - \log n^*$, we derive $\mathbb{P}[X_j(n) = 1] \sim 1 - g(\eta)$ with

$$g(\eta) := \sum_{i=0}^{m-1} g(i, \eta).$$

6.1 Number of distinct values

Set $X(n) := \sum\limits_{i=1}^{\infty} X_i(n)$. We must first check the asymptotic independency of the urns. Let us consider $\Pi_n(z) = \mathbb{E}(z^{X(n)})$. We are again interested in the behaviour of $\Pi_n(z)$ for complex $z \in \overline{D}_\epsilon(1) = \{t \mid |t - 1| \leq \epsilon\}$, where ϵ is a small fixed positive real number. We choose $\epsilon > 0$ such that $c := \log(1 + \epsilon) < 1$.

Theorem 6.1 *We have*

$$\Pi_n(z) \sim \prod_{l=1}^{\infty} \left[R(l,n) + z(1 - R(l,n)) \right], \quad n \to \infty.$$

uniformly for $z \in \overline{D}_\epsilon(1)$, where $0 < c = \log(1 + \epsilon) < 1$.

Proof
We again use the urn model. As before, we replace the *fixed* number n of balls with N balls, where N is a Poisson random variable with $\mathbb{E}(N) = \tau$. Thus, the urns are *independent*, and the number of balls in urn l is a Poisson random variable with parameter $\tau^* q^l$. Again we use a "$\sim$" to denote the Poissonized model. It follows that urn l has exactly i balls with probability $\frac{1}{i!}(\tau^* q^l)^i e^{-\tau^* q^l}$. So the generating function of $\widetilde{X}_l(\tau)$ is

$$\mathbb{E}(z^{\widetilde{X}_l(\tau)}) = R(l,\tau) + z(1 - R(l,\tau)) = 1 + (z-1)(1 - R(l,\tau)).$$

We have $\widetilde{X}(\tau) = \sum_{l=1}^{\infty} \widetilde{X}_l(\tau)$, and thus

$$G(\tau, z) := \mathbb{E}(z^{\widetilde{X}(\tau)}) = \mathbb{E}(z^{\sum \widetilde{X}_l(\tau)}).$$

Since the urns are independent in the Poissonized model, then $\mathbb{E}(z^{\sum \widetilde{X}_l(\tau)}) = \prod_{l=1}^{\infty} \mathbb{E}(z^{\widetilde{X}_l(\tau)})$. Thus

$$G(\tau, z) = \prod_{l=1}^{\infty} \big[1 + (z-1)(1 - R(l, \tau))\big]. \tag{6.2}$$

We again write $\tau = Re^{it}$ for real $R \geq 0$ and $-\pi < t \leq \pi$. Thus $|\tau| = R$. We again consider the linear cone $\mathcal{S}_{\pi/4} = \{\tau = Re^{it} \mid -\pi/4 \leq t \leq \pi/4\}$. Now we derive asymptotics about the growth of $|G(\tau, z)|$ for $\tau \in \mathcal{S}_{\pi/4}$. Our estimates are valid *uniformly* for $z \in \overline{D}_\epsilon(1) = \{t \mid |t-1| \leq \epsilon\}$. We encapsulate our results in the following lemma.

Lemma 6.2 *For $\tau \in \mathcal{S}_{\pi/4}$, there exist reals $B > 0$, $R_0 > 0$, and $0 < c < 1$, such that if $|\tau| = R > R_0$ then*

$$|G(\tau, z)| \leq B|\tau|^c$$

uniformly for $z \in D_\epsilon(1)$.

Proof We first consider $l \geq 1 + \log R$. We have

$$\begin{aligned}
\prod_{l \geq 1+\log R} \big|\mathbb{E}(z^{\widetilde{X}_l(\tau)})\big| &= \prod_{l \geq 1+\log R} \big|1 + (z-1)(1 - R(l, \tau))\big| \\
&= \prod_{l \geq 1+\log R} \left|1 + (z-1)\sum_{i=m}^{\infty} \frac{1}{i!}(\tau^* q^l)^i e^{-\tau^* q^l}\right| \\
&\leq \prod_{l \geq 1+\log R} \left[1 + \epsilon \sum_{i=m}^{\infty} \frac{1}{i!}(|\tau| p q^{l-1})^i e^{-\Re(\tau) p q^{l-1}}\right] \\
&= \prod_{l \geq \log R} \left[1 + \epsilon \sum_{i=m}^{\infty} \frac{1}{i!}(|\tau| p q^{l})^i e^{-\Re(\tau) p q^{l}}\right] \\
&= \exp\left(\sum_{l \geq \log R} \ln\left[1 + \epsilon \sum_{i=m}^{\infty} \frac{1}{i!}(|\tau| p q^{l})^i e^{-\Re(\tau) p q^{l}}\right]\right) \\
&\leq \exp\left(\sum_{l \geq \log R} \left[\epsilon \sum_{i=m}^{\infty} \frac{1}{i!}(|\tau| p q^{l})^i e^{-\Re(\tau) p q^{l}}\right]\right)
\end{aligned}$$

where the inequality holds since $\ln(1+x) \leq x$ for real x. We note that $-\Re(\tau) < 0$ since $\tau \in \mathcal{S}_{\pi/4}$. Thus $e^{-\Re(\tau) p q^l} \leq 1$. It follows that

$$\begin{aligned}
\prod_{l \geq 1+\log R} \big|\mathbb{E}(z^{\widetilde{X}_l(\tau)})\big| &\leq \exp\left(\epsilon \sum_{i=m}^{\infty} \frac{1}{i!}(|\tau| p)^i \sum_{l \geq \log R} (q^i)^l\right) \\
&\leq \exp\left(\epsilon \sum_{i=m}^{\infty} \frac{1}{i!} \frac{(|\tau| p q^{\log R})^i}{1 - q^i}\right) \\
&\leq \exp\left(\epsilon \sum_{i=m}^{\infty} \frac{1}{i!} \frac{p^i}{1 - q^i}\right) \qquad \text{since } q^{\log R} = R^{-1} = |\tau|^{-1} \\
&\leq \exp\left(\frac{\epsilon}{1-q} \sum_{i=m}^{\infty} \frac{1}{i!} p^i\right) \qquad \text{since } 1/(1-q^i) \leq 1/(1-q) \\
&\leq \exp\left(\frac{\epsilon}{1-q} e^p\right) \\
&= \mathcal{O}(1).
\end{aligned}$$

Now we consider $l \le \log R$. We have

$$\prod_{l \le \log R} \left|\mathbb{E}(z^{\widetilde{X}_l(\tau)})\right| \le \prod_{l \le \log R} (1+\epsilon) \le (1+\epsilon)^{\log R} = R^{\log(1+\epsilon)} = |\tau|^c.$$

Combining these results, we have $|G(\tau, z)| = \mathcal{O}(1)|\tau|^c = \mathcal{O}(|\tau|^c)$ uniformly for $z \in D_\epsilon(1)$. This completes the proof of the lemma. ■

Now we return to the proof of Theorem 6.1. The lemma we just completed shows that condition (I) holds for Theorem 10.3 of [17]. Similar reasoning as in Theorem 5.1 shows that condition (O) holds too, namely: for $\tau \notin \mathcal{S}_{\pi/4}$, there exists A and $\alpha < 1$ such that $|G(\tau, z)e^\tau| \le A \exp(\alpha|\tau|)$ for $|\tau| > R_0$.

So the assumptions of Theorem 10.3 of [17] are all satisfied; therefore, we can depoissonize our results. In other words, $\Pi_n(z)$ and $G(\tau, z)$ have the same asymptotics. More precisely,

$$\Pi_n(z) = G(n, z) + \mathcal{O}(n^{c-1}).$$

Substituting $\tau = n$ within (6.2), we see that

$$G(n, z) = \prod_{l=1}^{\infty} \left[R(l, n) + z(1 - R(l, n))\right],$$

and we note that $0 < c < 1$, so this completes the proof of Theorem 6.1. ■

Theorem 6.1 confirms the asymptotic independence assumption.

Now with asymptotic independence of the urns representing each integer,

$$\mathbb{E}(e^{\alpha X}) \sim \exp\left[\sum_{j=1}^{\infty} \ln\left(1 + (e^\alpha - 1)(1 - T(j))\right)\right] = \exp\left[\sum_{l=1}^{\infty} \frac{(-1)^{l+1}}{l}(e^\alpha - 1)^l V_l\right],$$

with

$$V_l := \sum_{j=1}^{\infty} \big(1 - T(j)\big)^l.$$

We obtain

$$\begin{aligned}
V_l &= \sum_{j=1}^{\infty} \left\{ \sum_{k=0}^{l} (-1)^k \binom{l}{k} T(j)^k \right\} \\
&= \sum_{j=1}^{\infty} \left\{ \sum_{k=0}^{l} (-1)^k \binom{l}{k} T(j)^k - \sum_{k=0}^{l} \binom{l}{k} (-1)^k \right\} \\
&= \sum_{k=1}^{l} \binom{l}{k} (-1)^{k+1} S_k, \text{ with} \\
S_k &:= \sum_{j=1}^{\infty} \left(1 - T(j)^k\right).
\end{aligned}$$

First of all, let us check that, for large n, $T(j)$, as a function of j, is an honest distribution function in the sense that it is monotonous in j. Considering j as a continuous variable, we obtain

$$T'(j) = -L \sum_{i=0}^{m-1} B(i, j)(i - n^* q^j)/(1 - pq^{j-1}).$$

But to the left of the concentration domain, $n^* q^j \gg m$, so that $T'(j) > 0$. In and to the right of the concentration domain, the Poisson approximation leads, with $\lambda := e^{-L\eta}$, to

$$\sum_{i=0}^{m-1} e^{-\lambda} \lambda^i / i!(i - \lambda) = -e^{-\lambda} \lambda^m / (m-1)! < 0$$

and again, $T'(j) > 0$. Setting $\eta = j - \log(n^*)$, this leads to

$$T(j)^k \sim G(\eta) := g(\eta)^k,$$

and S_k is the mean of the RV with distribution function $T(j)^k$, minus 1 (as the sum starts here at $j = 1$).

Now we need a rate of convergence. This is computed as follows. We will consider three ranges. Let $1/2 < \beta < 1$.

- For $j < \beta \log n^*$, $T(j)^k$ is small. Indeed
$$T(j)^k \leq \left[mn^{*\,m}\exp(-n^{*\,1-\beta})\right]^k.$$
- For $\beta \log n^* \leq j < 2\log n^*$ we have
$$\begin{aligned} T(j)^k - g(\eta)^k &= \left[\sum_{i=0}^{m-1} B(i,j)\right]^k - \left[\sum_{i=0}^{m-1} g(i,\eta)\right]^k \\ &\sim \left[\sum_{i=0}^{m-1} g(i,\eta)[1+\mathcal{O}(1/n^*)+\mathcal{O}(1/Q^j)+\mathcal{O}(n^*/Q^{2j})]\right]^k - \left[\sum_{i=0}^{m-1} g(i,\eta)\right]^k \\ &= m^k\mathcal{O}\big(1/n^{*\,2\beta-1}\big). \end{aligned}$$
- For $j = 2\log n^* + x$, $x \geq 0$, we have
$$\begin{aligned} T(j)^k - g(\eta)^k &\sim \Bigg[g(0,\eta)[1+\mathcal{O}(n^*/(n^{*\,4}Q^{2x}))] \\ &\quad + \sum_{i=1}^{m-1} g(i,\eta)[1+\mathcal{O}(1/n^*)+\mathcal{O}(1/n^{*\,2})+\mathcal{O}(n^*/n^{*\,4})]\Bigg]^k - \left[\sum_{i=0}^{m-1} g(i,\eta)\right]^k \\ &\sim m^k\mathcal{O}\left[1/(n^*Q^x)\right]^k/n^* \end{aligned}$$
as
$$g(i,\eta) = \mathcal{O}\big[(1/(n^*Q^x)^i)\big].$$

We can then proceed as in Section 5.1.

Now we return to the main problem: compute the mean of the distribution function

$$T(j)^k \sim G(\eta) := g(\eta)^k.$$

- Let us first consider the case $k = 1$. This leads for $i = 0$ to
$$\varphi_1(\alpha) = \int_{-\infty}^{\infty} e^{\alpha x} g'(0,x)dx = \Gamma(1-\widetilde{\alpha}), \quad \Re(\alpha) < L.$$
Next, we derive
$$\varphi_2(\alpha) = \int_{-\infty}^{\infty} e^{\alpha x}\sum_{i=1}^{m-1} g'(i,x)dx = -\alpha M_1(\alpha)$$
with
$$M_1(\alpha) = \sum_{i=1}^{m-1}\frac{\Gamma(i-\widetilde{\alpha})}{Li!}, \quad \Re(\alpha) < L.$$
This leads to
$$\phi(\alpha) = -M_1(\alpha)(e^\alpha - 1) + \Gamma(1-\widetilde{\alpha})(e^\alpha-1)/\alpha.$$
Proceeding as in Section 3 and as in the trie case (see [13]) we obtain
$$S_1 \sim \log n^* + \frac{\gamma}{L} - \frac{1}{2} - M_1(0) + \beta_{1,1} + \mathcal{O}(1/n)$$
with $M_1(0) = H_{m-1}/L$ and
$$\beta_{1,1} = \frac{1}{L}\sum_{l\neq 0}\left[-\sum_{i=1}^{m-1}\frac{\Gamma(i+\chi_l)}{i!} - \Gamma(\chi_l)\right]e^{-2l\pi i\log n^*} = -\frac{1}{L}\sum_{l\neq 0}\frac{\Gamma(m+\chi_l)}{\chi_l(m-1)!}e^{-2l\pi i\log n^*},$$
by induction, which is exactly the expression given in [1].

- For $k=2$, we derive similarly

$$\varphi_1(\alpha) = 2^{\widetilde{\alpha}}\Gamma(1-\widetilde{\alpha}),$$
$$\varphi_2(\alpha) = -\alpha M_2(\alpha),$$
$$M_2(\alpha) = \frac{1}{L}\sum_{i=0}^{m-1}\sum_{v=0}^{m-1}[\![v+i\neq 0]\!]\frac{2^{\widetilde{\alpha}}\Gamma(i+v-\widetilde{\alpha})}{2^{i+v}i!v!}.$$

  This leads to

$$S_2 \sim \log n^* + \log 2 + \frac{\gamma}{L} - \frac{1}{2} - M_2(0) + \beta_{1,2} + \mathcal{O}(1/n),$$

  with

$$\beta_{1,2} = \sum_{l\neq 0}\left[-M_2(\alpha)|_{\widetilde{\alpha}=-\chi_l} - 2^{-\chi_l}\Gamma(\chi_l)/L\right]e^{-2l\pi i\log n^*}$$
$$= -\sum_{l\neq 0}\sum_{i=0}^{m-1}\sum_{v=0}^{m-1}\frac{\Gamma(i+v+\chi_l)}{2^{i+v}i!v!L}e^{-2l\pi i\log(2n^*)}.$$

- For general k we finally obtain

$$S_k \sim \log n^* + \log k + \frac{\gamma}{L} - \frac{1}{2} - M_k(0) + \beta_{1,k} + \mathcal{O}(1/n),$$

  with

$$M_k(\alpha) = \sum_{i_1=0}^{m-1}\cdots\sum_{i_k=0}^{m-1}[\![i_1+\cdots+i_k\neq 0]\!]\frac{k^{\widetilde{\alpha}}\Gamma(i_1+\ldots+i_k-\widetilde{\alpha})}{k^{i_1+\cdots+i_k}i_1!\ldots i_k!L},$$
$$\beta_{1,k} = -\sum_{l\neq 0}\sum_{i_1=0}^{m-1}\cdots\sum_{i_k=0}^{m-1}\frac{\Gamma(i_1+\cdots+i_k+\chi_l)}{k^{i_1+\cdots+i_k}i_1!\ldots i_k!L}e^{-2l\pi i\log(kn^*)}.$$

  Note again the presence of k in the exponent.

This gives

$$V_l \sim \log n^* - 1/2 + \gamma/L + B_l + C_l + \beta_l, \quad \text{with}$$
$$B_l := \sum_{k=2}^{l}\binom{l}{k}(-1)^{k+1}\log k,$$
$$\beta_l = \sum_{k=1}^{l}\binom{l}{k}(-1)^{k+1}\beta_{1,k},$$
$$C_l = \sum_{k=1}^{l}\binom{l}{k}(-1)^{k+1}(-M_k(0))$$

and, finally

$$\mathbb{E}(e^{\alpha X}) = \exp\Bigg[\alpha(\log n^* - 1/2 + \gamma/L) + \sum_{l=2}^{\infty}\frac{(-1)^{l+1}}{l}(e^\alpha-1)^l B_l + \sum_{l=1}^{\infty}\frac{(-1)^{l+1}}{l}(e^\alpha-1)^l\beta_l$$
$$+ \sum_{l=1}^{\infty}\frac{(-1)^{l+1}}{l}(e^\alpha-1)^l C_l + \mathcal{O}(1/n)\Bigg].$$

From this, we derive

$$\Theta_p(\alpha) = \exp\Bigg[\sum_{l=2}^{\infty}\frac{(-1)^{l+1}}{l}(e^\alpha-1)^l B_l + \sum_{l=1}^{\infty}\frac{(-1)^{l+1}}{l}(e^\alpha-1)^l\beta_l$$

$$+\sum_{l=1}^{\infty}\frac{(-1)^{l+1}}{l}(e^{\alpha}-1)^l C_l-\alpha(-M_1(0)+\beta_{1,1})\Big],$$

and the moments of $X-\log n^*$ are given by

$$\begin{aligned}
\widetilde{m}_1&=\frac{\gamma}{L}-\frac{1}{2}-M_1(0),\\
w_1&=\beta_{1,1},\\
\widetilde{\mu}_2&=\log 2+M_1(0)-M_2(0),\\
\widetilde{\mu}_3&=-3\log 2+2\log 3-M_1(0)+3M_2(0)-2M_3(0),\\
\kappa_2&=\beta_{1,2}-\beta_{1,1},\\
\kappa_3&=\beta_{1,1}-3\beta_{1,2}+2\beta_{1,3}.
\end{aligned}$$

The quantities $\widetilde{m}_1$, w_1, $\widetilde{\mu}_2$ are given in Archibald, Knopfmacher and Prodinger [1]. Since they look somehow different, here is a

Direct proof that the two expressions for the variance coincide.

What is denoted $\widetilde{\mu}_2$ here, comes out in [1] as

$$\log 2+\frac{2}{L}\sum_{i\ge1}\frac{(-1)^{i+m-1}}{i(Q^i-1)}\binom{i+m-1}{i}\binom{i-1}{m-1}-\frac{2}{L}\sum_{j=1}^{m-1}\frac{1}{2j}\binom{2j}{j}\sum_{h\ge0}\binom{-2j}{h}\frac{1}{Q^{h+j}-1}$$
$$+\frac{2}{L}\sum_{h\ge1}\frac{(-1)^{h-1}}{h(Q^h-1)}-\frac{1}{L}\sum_{j=1}^{m-1}\frac{1}{2j}\binom{2j}{j}2^{-2j}.$$

So we are left to prove that

$$2\sum_{i\ge1}\frac{(-1)^{i+m-1}}{i(Q^i-1)}\binom{i+m-1}{i}\binom{i-1}{m-1}-2\sum_{j=1}^{m-1}\frac{1}{2j}\binom{2j}{j}\sum_{h\ge0}\binom{-2j}{h}\frac{1}{Q^{h+j}-1}$$
$$+2\sum_{h\ge1}\frac{(-1)^{h-1}}{h(Q^h-1)}-\sum_{j=1}^{m-1}\frac{1}{2j}\binom{2j}{j}2^{-2j}=-{\sum_{i,j=0}^{m-1}}'\frac{(i+j-1)!}{i!j!}2^{-i-j}+H_{m-1},$$

where the dashed sum means that the term $i=j=0$ has to be excluded. If we take the diagonal out of the sum with the dash, we are left to prove:

$$\sum_{i\ge1}\frac{(-1)^{i+m-1}}{i(Q^i-1)}\binom{i+m-1}{i}\binom{i-1}{m-1}-\sum_{j=1}^{m-1}\frac{1}{2j}\binom{2j}{j}\sum_{h\ge0}\binom{-2j}{h}\frac{1}{Q^{h+j}-1}$$
$$+\sum_{h\ge1}\frac{(-1)^{h-1}}{h(Q^h-1)}=-\sum_{0\le i<j<m}\frac{(i+j-1)!}{i!j!}2^{-i-j}+\frac{1}{2}H_{m-1}$$

or

$$\sum_{i\ge1}\frac{(-1)^i}{i(Q^i-1)}\binom{-i-1}{m-1}\binom{i-1}{m-1}-\sum_{i\ge1}\frac{(-1)^i}{Q^i-1}\sum_{j=1}^{\min(m-1,i)}\frac{(-1)^j(i+j-1)!}{j!j!(i-j)!}$$
$$-\sum_{i\ge1}\frac{(-1)^i}{i(Q^i-1)}=-\sum_{0\le i<j\le m-1}\frac{(i+j-1)!}{i!j!}2^{-i-j}+\frac{1}{2}H_{m-1};$$

notice that the right side does not depend on Q! Now we evaluate one appearing sum for $m>i\ge1$:

$$\begin{aligned}
\sum_{j=1}^{\min(m-1,i)}\frac{(-1)^j(i+j-1)!}{j!j!(i-j)!}&=\sum_{j=0}^{i}\frac{(-1)^j(i+j-1)!}{j!j!(i-j)!}-\frac{1}{i}\\
&=\frac{1}{i}\sum_{j=0}^{i}\binom{-i}{j}\binom{i}{i-j}-\frac{1}{i}
\end{aligned}$$

$$= \frac{1}{i}\binom{0}{i} - \frac{1}{i} = -\frac{1}{i},$$

by Vandermonde's convolution. Thus we are left to prove that

$$\sum_{i\ge m} \frac{(-1)^i}{i(Q^i-1)}\binom{-i-1}{m-1}\binom{i-1}{m-1} - \sum_{i\ge m}\frac{(-1)^i}{Q^i-1}\sum_{j=1}^{m-1}\frac{(-1)^j(i+j-1)!}{j!j!(i-j)!}$$
$$-\sum_{i\ge m}\frac{(-1)^i}{i(Q^i-1)} = -\sum_{0\le i<j<m}\frac{(i+j-1)!}{i!j!}2^{-i-j} + \frac{1}{2}H_{m-1}.$$

We will achieve that by proving that both sides are actually zero!

We treat the right side by induction on m, the instance $m = 1$ being clear. The induction step amounts to prove that

$$\sum_{0\le i<m}\frac{(i+m-1)!}{i!m!}2^{-i-m} = \frac{1}{2m},$$

or

$$\sum_{0\le i<m}\frac{(i+m-1)!}{i!(m-1)!}2^{-i} = 2^{m-1},$$

which is the "unexpected" sum (5.20) in [5].

Now we turn to the left side; we need to show that for $i \ge m$,

$$\binom{-i-1}{m-1}\binom{i-1}{m-1} - i\sum_{j=1}^{m-1}\frac{(-1)^j(i+j-1)!}{j!j!(i-j)!} - 1 = 0,$$

or

$$\binom{-i-1}{m-1}\binom{i-1}{m-1} = \sum_{j=0}^{m-1}\binom{-i}{j}\binom{i}{j}.$$

This follows from Euler's identity [5, ex. 28, p. 244], or can simply be proved by induction.

This finishes the proof.

Remark. We learn from this computation that the expression for $\widetilde{\mu}_2$ can *still* be simplified:

$$\widetilde{\mu}_2 = \log 2 - \frac{1}{L}\sum_{1\le i<m}\frac{(2i-1)!}{i!i!}2^{-2i}.$$

Now we continue after this intermezzo.— The asymptotic distribution of X is given by the following result:

Theorem 6.3 *Set* $\eta := j - \log n^*$ *and*

$$\Psi_6(\eta) := g(\eta)\prod_{i=1}^{\infty}[1 - g(\eta - i)].$$

Then, with j *integer and* $\eta = \mathcal{O}(1)$*,*

$$\mathbb{P}(X = j) \quad\sim\quad f(\eta) = \sum_{u=0}^{\infty}\Psi_6(\eta-u+1)\prod_{w=2-u}^{\infty}g(\eta+w)\sum_{\substack{r_1<\dots<r_u\\ r_j\ge 2-u}}\prod_{i=1}^{u}\frac{1-g(\eta+r_i)}{g(\eta+r_i)},$$

$$\mathbb{P}(X \le j) \quad\sim\quad F(\eta),\ \textit{with}\ F(\eta) := \sum_{i=0}^{\infty}f(\eta - i).$$

6.2 Maximal non-empty urn

6.2.1 General multiplicity m

We have here

$$p(j) := \mathbb{P}(M = j) \sim (1 - T(j)) \prod_{i=j+1}^{\infty} T(i),$$

$$P(j) \sim \prod_{i=j+1}^{\infty} T(i),$$

and this leads to

$$\begin{aligned} p(j) &\sim f(\eta) = (1 - g(\eta))\Psi_7(\eta), \\ P(j) &\sim F(\eta) = \Psi_7(\eta), \end{aligned} \tag{6.3}$$

with

$$\Psi_7(\eta) = \prod_{i=1}^{\infty} g(\eta + i).$$

Now we could proceed as in Section 4.3.

6.2.2 Particular case $m = 2$

In the following we use a different approach and work out the details for $m = 2$. The reason for this restriction is that the results are more appealing in this case; Euler's partition identity allows to expand a product into a sum, and there is nothing equivalent for $m > 2$. This can be compared with the analysis in [4] and the analysis in [10]; the latter does not have the nice explicit series $N(s)$.

We can compute $P(j)$ by noticing that there are some k elements which fall into urns numbered $> j$, but are alone in their urn, and the remaining $n - k$ elements which are in urns with numbers $\le j$, but no further restrictions. Thus

$$\begin{aligned} P(j) := \mathbb{P}[X \le j] &= \sum_{k=0}^{n} \binom{n}{k} (1 - q^j)^{n-k} k! \sum_{j < \lambda_1 < \cdots < \lambda_k} pq^{\lambda_1 - 1} \ldots pq^{\lambda_k - 1} \\ &= \sum_{k=0}^{n} \binom{n}{k} (1 - q^j)^{n-k} k! p^k q^{jk} \sum_{0 \le \lambda_1 < \cdots < \lambda_k} q^{\lambda_1 + \cdots + \lambda_k} \\ &= \sum_{k=0}^{n} \binom{n}{k} (1 - q^j)^{n-k} k! p^k q^{jk} [z^k] \prod_{l \ge 0} (1 + zq^l) \\ &= \sum_{k=0}^{n} \binom{n}{k} (1 - q^j)^{n-k} k! p^k q^{jk} \frac{q^{\binom{k}{2}}}{(q)_k}, \end{aligned}$$

by one of Euler's partition identities.

After these preliminaries, we consider the asymptotic form. We have

$$P(j) = \sum_{u=0}^{n} \frac{(pq^j)^u q^{\binom{u}{2}} n!}{(q)_u (n-u)!} (1 - q^j)^{n-u}. \tag{6.4}$$

Setting $\eta = j - \log n$, we obtain

$$P(j) \sim F(\eta),$$

with

$$F(\eta) = \sum_{u=0}^{\infty} \frac{p^u q^{\binom{u}{2}}}{(q)_u} e^{-Lu\eta} \exp(-e^{-L\eta}). \tag{6.5}$$

Let us first check the equivalence of (6.5) with $\Psi_7(\eta)$ given by

$$\Psi_7(\eta) = \prod_{k=1}^{\infty} \exp\left(-e^{-L(\widetilde{\eta}+k)}\right)\left[1+e^{-L(\widetilde{\eta}+k)}\right],$$

with

$$\widetilde{\eta} = j - \log n^* = \eta - \log\frac{p}{q}.$$

This leads to

$$\begin{aligned}\Psi_7(\eta) &= \prod_{k=1}^{\infty} \exp\left(-e^{-L\eta}pq^{k-1}\right)\left[1+e^{-L\eta}pq^{k-1}\right] \\ &= \exp\left(e^{-L\eta}\right)\prod_{k=1}^{\infty}\left[1+e^{-L\eta}pq^{k-1}\right].\end{aligned} \tag{6.6}$$

Now, again by Euler's identity, (6.5) gives

$$F(\eta) = \exp(-e^{-L\eta})\prod_{k=0}^{\infty}\left[1+e^{-L\eta}pq^{k}\right],$$

which is equivalent to (6.6).

Let us now compute the rate of convergence. We must bound $|P(j)-F(\eta)|$. Let $0<\beta<1$. We will consider three ranges

- For $j < \beta\log n$, $P(j)$ is small. Indeed

  $$P(j) \le \exp(-n^{1-\beta})/K\sum_{u=0}^{\infty}\frac{q^{u(u-1)/2}p^u}{(1-q^j)^u}q^{ju}n^u.$$

  The sum is bounded by

  $$\sum_{u=0}^{\infty} q^{u^2/2}n^u,$$

  which we can estimate by the Euler–Maclaurin formula (or by the Mellin transform). The sum is asymptotically given by

  $$\exp(L\log^2 n/2)\sqrt{2\pi/L}.$$

  Note that the maximum of the quadratic form in the exponent occurs at $u^* = \log n$.

- For $\beta\log n \le j < 2\log n$, we set $\delta := e^{-L\eta}$. Note that $1/n \le \delta \le n^{1-\beta}$.

  Now we use the "sum splitting technique." Set $r = n^{1/4}$.

  1. truncating the sum in (6.4) to r leads to an error E_1:

     $$E_1 \le \sum_{u=r}^{n}\frac{q^{u(u-1)/2}}{K}\delta^u e^{-\delta} \le e^{-Ln^{1/2}}E_{11}/K,$$

     where

     $E_{11} = \mathcal{O}[\exp(L(1-\beta)n^{1/4}\log n)\exp(-n^{1-\beta})],$ if $\delta = n^{1-\beta}$, as $r \gg u^* = (1-\beta)\log n$,
     $E_{11} = \mathcal{O}(1)$ if $\delta = 1$,
     $E_{11} = \mathcal{O}(\exp(-Ln^{1/4}\log n)),$ if $\delta = 1/n$.

  2. replacing $\frac{n!}{n^u(n-u)!}$ in the truncated sum by 1 leads to a relative error $-u^2/n$ (by Stirling), which leads to an error E_2:

     $$E_2 \le \sum_{u=1}^{r}\frac{q^{u(u-1)/2}}{K}\delta^u e^{-\delta}u^2/n.$$

This gives

$$E_2 \le \sum_{u=1}^{r} \frac{q^{u^2/2}}{K}(n^{1-\sigma})^u e^{-n^{1-\sigma}} u^2/n, \quad \text{if } \delta = n^{1-\sigma}.$$

Now we use the standard saddle point technique: the saddle point is

$$u^* = (1-\sigma)\log n + 2/(L(1-\sigma)\log n) + \mathcal{O}(1/\log^3 n),$$

and this leads to

$$E_2 \le \exp[L/2(1-\beta)^2(\log^2 n + \mathcal{O}(\log\log n))]e^{-n^{1-\beta}}/(Kn), \text{ if } \delta = n^{1-\beta},$$
$$E_2 = \mathcal{O}(1/n), \quad \text{if } \delta = 1,$$
$$E_2 = \mathcal{O}(1/n^2), \quad \text{if } \delta = 1/n.$$

3. replacing $(1-q^j)^{n-1}$ in the truncated sum by $\exp(-e^{-L\eta})$ leads to a relative error $nq^{2j} = \delta^2/n$ which gives an error E_3:

$$E_3 \le \sum_{u=0}^{r} \frac{q^{u(u-1)/2}}{K}\delta^u \delta^2 e^{-\delta}/n.$$

This gives

$$E_3 \le \sum_{u=0}^{r} \frac{q^{u^2/2}}{K} n^{(1-\sigma)(u+2)} e^{-n^{1-\sigma}}/n, \text{ if } \delta = n^{1-\sigma},$$

and this leads to

$$E_3 \le \exp[L/2(1-\beta)^2(\log^2 n + \mathcal{O}(\log n))]e^{-n^{1-\beta}}/(Kn), \text{ if } \delta = n^{1-\beta},$$
$$E_3 = \mathcal{O}(1/n), \text{ if } \delta = 1,$$
$$E_3 = \mathcal{O}(1/n^3), \text{ if } \delta = 1/n.$$

4. completing the sum in (6.5) leads to an error E_4:

$$E_4 \le \sum_{u=r}^{n} \frac{q^{u(u-1)/2}p^u}{K}\delta^u e^{-\delta},$$

which is analyzed as E_1.

- For $j = 2\log n + x$, $x \ge 0$, we have $\delta = 1/(nQ^x)$. Set again $r = n^{1/4}$.

1. truncating the sum in (6.4) to r leads to an error E_1:

$$E_1 \le e^{-Ln^{1/2}/2} e^{-L(\log n + x)n^{1/4}}/K,$$

2. replacing $\frac{n!}{n^u(n-u)!}$ in the truncated sum by 1 leads to an error E_2:

$$E_2 = \mathcal{O}(1/(n^2 Q^x)),$$

3. replacing $(1-q^j)^{n-1}$ in the truncated sum by $\exp(-e^{-L\eta})$ leads to a an error E_3:

$$E_3 = \mathcal{O}[1/(n^3 Q^{2x})],$$

4. completing the sum in (6.5) leads to an error E_4:

$$E_4 \le \sum_{u=r}^{n} \frac{q^{u(u-1)/2}p^u}{K}\delta^u e^{-\delta},$$

which is analyzed as E_1.

Now we can bound the difference between the moments of X and the moments based on $F(\eta)$:

$$\begin{aligned}
&\Big|\sum_j j^k\Big([P(j)-P(j-1)]-[F(\eta)-F(\eta-1)]\Big)\Big| \\
&\quad\le 2\Big[\mathcal{O}\big((\beta\log n^*)^{k+1}\exp(-n^{1-\beta-\varepsilon})\big) \\
&\qquad + (2\log n^*)^{k+1}\mathcal{O}(1/n)+\mathcal{O}\Big(\sum_{x\ge0}(2\log n^*+x)^k Q^{-x}/n^2\Big)\Big] \\
&\quad=\mathcal{O}(1/n^{1-\varepsilon}),
\end{aligned}$$

where ε is any small positive real number. Now we turn to the moments. We obtain

$$\varphi(\alpha)=\Gamma(1-\tilde\alpha)-\alpha\sum_{u=1}^{\infty}V(u)\Gamma(u-\tilde\alpha)/L,$$

with

$$\tilde\alpha:=\alpha/L,\quad \Re(\alpha)<L,\quad V(u):=\frac{p^u q^{\binom{u}{2}}}{(q)_u}.$$

We recognize the trie expression in the first part. Note also that $\varphi(0)=1$ as it should. The second part of $\varphi(\alpha)$ leads to

$$\phi_2(\alpha)=-(e^{\alpha}-1)\sum_{u=1}^{\infty}V(u)\Gamma(u-\tilde\alpha)/L.$$

Set $\widetilde{\alpha}=-s$, $s=\sigma+it$, $\sigma\ge0$. Using (3.7), $|\phi_2(\alpha)|$ is bounded by

$$\mathcal{O}\Big(\sum_{u=1}^{\infty}\frac{q^{u(u-1)/2}p^u}{(q)_\infty}|t|^{u+\sigma-1/2}e^{-\pi|t|/2}\Big)=\mathcal{O}\left(e^{L\log^2(|t|)/2}\right)|t|^{\sigma-1/2}e^{-\pi|t|/2}$$

which is exponentially decreasing. Now we set

$$\begin{aligned}
C_1&:=\sum_{u=1}^{\infty}V(u)\Gamma(u),\\
C_2&:=\sum_{u=1}^{\infty}V(u)\Gamma(u)\psi(u),\\
C_3&:=\sum_{u=1}^{\infty}V(u)\Gamma(u)\psi(1,u),\\
C_4&:=\sum_{u=1}^{\infty}V(u)\Gamma(u)\psi^2(u).
\end{aligned}$$

This leads to

$$\begin{aligned}
m_1&=(\gamma-C_1)/L,\\
m_2&=(\pi^2/6+\gamma^2+2C_2)/L^2,\\
m_3&=(2\zeta(3)+\pi^2\gamma/2+\gamma^3-3C_3-3C_4)/L^3,\\
\tilde m_1&=m_1+1/2,\\
\tilde m_2&=m_1+1/3+(\pi^2/6+\gamma^2+2C_2)/L^2,\\
\tilde m_3&=m_1+1/4+(\pi^2/4+3\gamma^2/2+3C_2)/L^2+(2\zeta(3)+\pi^2\gamma/2+\gamma^3-3C_3-3C_4)/L^3,\\
\sigma^2&=(\pi^2/6+\gamma^2+2C_2)/L^2-m_1^2,\\
\mu_3&=2m_1^3+(-3m_1\gamma^2-m_1\pi^2/2-6m_1C_2)/L^2+(2\zeta(3)+\pi^2\gamma/2+\gamma^3-3C_3-3C_4)/L^3,
\end{aligned}$$

$\tilde{\mu}_2 = (\pi^2/6 + \gamma^2 + 2C_2)/L^2 - m_1^2 + 1/12,$
$\tilde{\mu}_3 = \mu_3.$

Let us now turn to the fluctuating components. The fundamental strip for s is $\Re(s) \in \langle -1, 0 \rangle$. First of all, (3.3) and (3.4) lead to

$$w_1 = -\sum_{l\neq 0}\Big[\Gamma(\chi_l) + \sum_{u=1}^{\infty} V(u)\Gamma(u+\chi_l)\Big]e^{-2l\pi i \log n}/L.$$

Equations (3.5) and (3.6) lead, after the usual simplifications necessary to help Maple, to

$$\begin{aligned}
\kappa_2 &= 2\sum_{l\neq 0}\Big[\Gamma(\chi_l)\psi(\chi_l) + \sum_{u=1}^{\infty} V(u)\Gamma(u+\chi_l)\psi(u+\chi_l)\Big]e^{-2l\pi i \log n}/L^2 - 2m_1 w_1 - w_1^2,\\
\kappa_3 &= \sum_{l\neq 0}\Big[-3\Gamma(\chi_l)\psi(1,\chi_l) - 3\Gamma(\chi_l)\psi^2(\chi_l) - 6\Gamma(\chi_l)\psi(\chi_l)L(w_1+m_1)\\
&\quad + \sum_{u=1}^{\infty}(-3V(u)\Gamma(u+\chi_l)\psi(1,u+\chi_l) - 6V(u)\Gamma(u+\chi_l)\psi(u+\chi_l)L(m_1+w_1)\\
&\quad - 3V(u)\Gamma(u+\chi_l)\psi^2(u+\chi_l))\Big]e^{-2l\pi i \log n}/L^3\\
&\quad + [4w_1^3L^2 + 12m_1^2w_1L^2 + 12m_1w_1^2L^2 - \pi^2 w_1 - 6\gamma^2 w_1 - 12C_2 w_1]/(2L^2).
\end{aligned}$$

6.3 First empty urn

Set $E := \inf\{i : X_i = 1\}$. We obtain

$$p(j) \sim T(j)\prod_{i=1}^{j-1}[1 - T(i)],$$
$$P(j) \sim 1 - \prod_{i=1}^{j}[1 - T(i)].$$

This leads to

$$p(j) \sim f(\eta) = \Psi_6(\eta),$$
$$P(j) \sim F(\eta) = 1 - \Psi_8(\eta),$$

with

$$\Psi_8(\eta) := \prod_{i=0}^{\infty}(1 - g(\eta - i)).$$

We proceed now exactly as in Section 4.3 and we derive all moments. We recognize here the splitting process arising in *probabilistic counting*: see Kirschenhofer, Prodinger and Szpankowski [10]. The quantities $\tilde{m}_1$, $\tilde{\mu}_2$ and w_1 are given in their paper. We don't give more details in this subsection.

7 Conclusion

If we compare the approach in this paper with other ones that appeared previously, then we can notice the following. Traditionally, one would stay with exact enumerations as long as possible, and only at a late stage move to asymptotics. Doing this, one would, in terms of asymptotics, carry many unimportant contributions around, which makes the computations quite heavy, especially when it comes to higher moments. Here, however, approximations are carried out as early as possible, and this allows for streamlined (and often automatic) computations of the higher moments.

One of the referees asked the question: can this work be extended to other distributions under conditions of exponentially decreasing tails? Indeed, this can be done, but at the expense of less explicit formulæ. Another interesting problems would be to consider Carlitz compositions (where two successive parts are different) and other Markov chains (see [13]). This will be the object of future work.

Acknowledgements

We want to thank A.V. Gnedin for pointing out a reference to Karlin's paper. We also thank the referees for pertinent comments.

References

[1] M. Archibald, A. Knopfmacher, and H. Prodinger. The number of distinct values in a geometrically distributed sample. *Discrete Mathematics, to appear*, 2003.

[2] V.P. Chistyakov. Discrete limit distributions in the problem of balls falling in cells with arbitrary probabilities. *Math. Notes*, 1:6–11, 1967.

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

[4] P. Flajolet and G.N. Martin. Probabilistic counting algorithms for data base applications. *Journal of Computer and System Sciences*, 31:182–209, 1985.

[5] R. L. Graham, D. E. Knuth, and O. Patashnik. *Concrete Mathematics (Second Edition)*. Addison Wesley, 1994.

[6] P. Hitczenko and G. Louchard. Distinctness of compositions of an integer: a probabilistic analysis. *Random Structures and Algorithms*, 19:407–437, 2001.

[7] P. Jacquet and W. Szpankowski. Analytic depoissonization and its applications. *Theoretical Computer Science*, 201:1–62, 1998.

[8] S. Karlin. Central limit theorems for certain infinite urn schemes. *Journal of Mathematics and Mechanics*, 17:373–401, 1967.

[9] P. Kirschenhofer and H. Prodinger. A result in order statistics related to probabilistic counting. *Computing*, 51:15–27, 1993.

[10] P. Kirschenhofer, H. Prodinger, and W. Szpankowski. Analysis of a splitting process arising in probabilistic counting and other related algorithms. *Random Structures and Algorithms*, 9:379–401, 1996.

[11] M. Loève. *Probability Theory*. D. Van Nostrand, 1963.

[12] G. Louchard. The number of distinct part sizes of some multiplicity in compositions of an integer. a probabilistic analysis. *Discrete Mathematics and Theoretical Computer Science*, AC:155–170, 2003.

[13] G. Louchard and H. Prodinger. The moments problem of extreme-value related distribution functions. *Algorithmica*, 2004. Submitted; see `http://www.ulb.ac.be/di/mcs/louchard/louchard.papers/mom7.ps`.

[14] H. Prodinger. Compositions and Patricia tries: no fluctuations in the variance! *SODA*, pages 1–5, 2004.

[15] W. Pugh. Skip lists: A probabilistic alternative to balanced trees. In *Algorithms and Data Structures*, volume 382 of *Lecture Notes in Computer Science*, pages 437–449, 1989.

[16] B.A. Sevastyanov and V.P. Chistyakov. Asymptotic normality in the classical problem of balls. *Theory of Probability and Applications*, 9:198–211, 1964.

[17] W. Szpankowski. *Average Case Analysis of Algorithms on Sequences*. Wiley, 2001.

[18] W. Szpankowski and V. Rego. Yet another application of a binomial recurrence. Order statistics. *Computing*, 43:401–410, 1990.