A UNIFIED PRESENTATION OF SOME URN MODELS

MICHAEL DRMOTA*, DANIÈLE GARDY**, AND BERNHARD GITTENBERGER*

ABSTRACT. For a sequence of m urns we investigate how the number of urns satisfying a certain condition (e.g. being empty) evolves in time when after each time unit a ball is thrown. We show for a variety of urn models that this process (suitably normalized) converges weakly to a Gaussian process.

1. Introduction

Consider a sequence of m urns into which we throw balls according to some rules. The balls are thrown one at a time and independently. Moreover, we assume that the balls are usually undistinguishable.

Assign to each urn U a valuation $Y(U)$ that is a real valued random variable and is additive, i.e. when we allocate balls in batches the final value of $Y(U)$ is the sum of the values for each batch considered separately. Furthermore, let $\mathcal{E}$ be a subset of the set of possible values of $Y(U)$. We are interested in the random variable X equal to the number of urns U such that $Y(U) \in \mathcal{E}$:

$$X = \sum_{i=1}^{m} 1_{Y(U_i)\in\mathcal{E}}.$$

We will deal with several urn models which are covered by the following two cases:

- If we are interested in the number of urns having a specified number of balls, then $Y(U)$ is the number of balls in the urn U, the set $\mathcal{E}$ is the set of the required numbers for a single urn, and is a subset of the natural integers. For example, empty urns (which have been studied in [11]) correspond to $\mathcal{E} = \{0\}$, urns with exactly r balls to $\mathcal{E} = \{r\}$, and urns with at most r balls to $\mathcal{E} = [0 \dots r]$.
- When we allocate balls of two colors (say red and blue), and consider the urns having a specified balance, $Y(U)$ is the balance of the urn, i.e. the difference between the number of blue balls and the number of red balls, and the set $\mathcal{E}$ of required balances is a subset of the set of relative integers. For example, balanced urns are obtained for $\mathcal{E} = \{0\}$, urns with balance r for $\mathcal{E} = \{r\}$, and urns with positive balance for $\mathcal{E} = N$. For previous work dealing with this case see [3].

We shall prove in this paper that, when the balls are thrown at each unit time the process associated to the number of urns with a specified number of balls or a specified balance converges weakly towards a Gaussian process, whose covariance matrix can be explicitely computed.

2. A General Model

We are interested in the stochastic process $X_m(n)$, $n = 0, 1, 2, \dots$, defined by the value of X at the time when exactly n balls have been thrown into the set of m urns. We will study the behavior of this process as $m \to \infty$. Consider for a moment the case where $Y(U)$ equals the number of balls and $\mathcal{E}$ has the shape $\mathcal{E} = \{r, r+1, r+2, \dots\}$. If at some time t_1 an urn satisfies $Y(U) \in \mathcal{E}$, it will satisfy this condition for all the times $t_2 \geq t_1$. In such cases (we expect that) the limiting

Date: April 21, 1999.

* Department of Algebra and Discrete Mathematics, Technische Universität Wien, Wiedner Hauptstraße 8-10/118, A-1040 Wien, Austria.

These authors' work was supported by the Austrian Science Foundation FWF, grant P10187-MAT.

** PRISM, UMR 8636 CNRS and Université de Versailles Saint-Quentin, 78035 Versailles Cedex, France. This author was also supported by the Austrian-French project AMADEUS No. 97049.

process (as $m \to \infty$) will be a Markov process (compare with Section 2.3). In the other cases this does not hold: An urn may satisfy $Y(U) \in \mathcal{E}$ at some time t_1, but not at a further time $t_2 > t_1$. This is obvious for the models with two types of balls, and this also holds for the number of urns with exactly or at most r balls. In these cases the limiting process will (probably) be non-Markov (compare with Section 2.3).

In this section we will first describe a generating function approach to the problem of empty urns which was studied by Kolchin et al.[11] (In fact they also studied the more general case of urns with exactly r balls). Afterwards we will present a generalization of this model and some examples covered by the general model.

2.1. **The number of empty urns.** We will apply the generating function technique for combinatorial enumeration (for an introduction to this method see e.g. [5, 9]). We have undistinguishable balls and distinguishable urns and thus we will use generating functions which are exponential w.r.t. the balls and ordinary w.r.t. the urns. As there is only one way to throw n balls into a single urn, the generating function of one urn is e^z and the generating function of a set of m urns is given by e^{mz} where z marks the balls. We introduce the variable x in order to mark the empty urns. This leads to the generating function

$$\Phi_1(x,z) = (e^z + x - 1)^m .$$

In this setup we have

$$P\{X_m(n) = k\} = \frac{[z^n x^k]\Phi_1(z,x)}{[z^n]\Phi_1(z,1)} \tag{2.1}$$

We are interested in asymptotic distributional properties in the *central domain*, i.e. when the ratio of the number of balls n and the number of urns m either tends to a constant or belongs to a compact set of $]0,+\infty[$. From (2.1) we get

$$EX_m(n) = \frac{[z^n]\frac{\partial}{\partial x}\Phi_1(z,1)}{[z^n]\Phi(z,1)} = m\left(1-\frac{1}{m}\right)^n \sim me^{-\theta},$$

for $m \to \infty$ and $n/m \to \theta > 0$. The variance is

$$\begin{aligned} VarX_m(n) &= \frac{[z^n]\frac{\partial^2}{\partial x^2}\Phi_1(z,1)}{[z^n]\Phi(z,1)} \\ &= m(m-1)\left(1-\frac{2}{m}\right)^n + m\left(1-\frac{1}{m}\right)^n - m^2\left(1-\frac{1}{m}\right)^{2n} \\ &\sim me^{-\theta}(1-(1+\theta)e^{-\theta}). \end{aligned}$$

The generating function for the bi-dimensional distribution is

$$\Phi_2(x_1,x_2,z_1,z_2) = ((e^{z_1}-1)e^{z_2} + x_1(e^{z_2}-1) + x_1x_2)^m$$

and here we have

$$P\{X_m(n_1) = k_1, X_m(n_1+n_2) = k_2\} = \frac{[z_1^{n_1}z_2^{n_2}x_1^{k_1}x_2^{k_2}]\Phi_2(z_1,z_2,x_1,x_2)}{[z_1^{n_1}z_2^{n_2}]\Phi_2(z_1,z_2,1,1)}.$$

The asymptotic covariance at (normalized) times $\theta_1 m$ and $\theta_2 m$ is $me^{-\theta_2}\,(1-(1+\theta_1)e^{-\theta_1})$; it is factorized w.r.t. θ_1 and θ_2, which means that the limiting process is Markovian. (For a proof of the existence of the limiting process see [11, Ch. IV]; for a relationship between the existence of a factorized form and the Markov property see [12])
The function marking the urns whose state has changed between the times t_1 and t_2 is

$$\Phi(x,z_1,z_2) = \left(e^{z_1+z_2} + (x-1)(e^{z_2}-1)\right)^m$$

and the g.f. describing the multivariate (d-dimensional) distributions is

$$\Phi_d(z_1,\dots,z_d,x_1,\dots,x_d) = \left(\sum_{i=0}^{d}(e^{z_{i+1}}-1)\exp\left(\sum_{j=i+2}^{d} z_j\right)\prod_{j=1}^{i} x_j\right)^m \tag{2.2}$$

2.2. **General g.f.-model.** We will now generalize the model described in the previous section. Let $g(z)$ be the generating function enumerating the allocation of balls into a single urn, and $f(z)$ the function enumerating those allocations such that $Y \in \mathcal{E}$. Furthermore, let us assume that $g(z)$ and $f(z)$ are entire functions. As before we mark the balls by z and the urns satisfying $Y(U) \in \mathcal{E}$ by x. Then the generating function describing the allocations of balls in the m urns is

$$\Phi_{\mathcal{E},1}(x,z) = (g(z) + (x-1)f(z))^m.$$

As can be seen from (2.1), the idea is to extract the coefficients $[z^n]\Phi_{\mathcal{E},1}(x,z)$, $[z^n x^k]\Phi_{\mathcal{E},1}(x,z)$, and their multivariate analoga. $\Phi_{\mathcal{E},1}(x,z)$ is analytic w.r.t. z and we can use a saddle point approximation. The shape of $\Phi_{\mathcal{E},1}(x,z)$ allows a straight forward application of the results of Bender and Richmond ([1]), which gives directly the convergence towards a Gaussian distribution. The asymptotic mean is obtained by a saddle point approximation: We have

$$\begin{aligned} EX_m(n) &= m\,\frac{[z^n]f(z)g(z)^{m-1}}{[z^n]g(z)^m} \\ &\sim m\frac{f(\rho)}{g(\rho)} \end{aligned}$$

where ρ is the solution of the saddle point equation $zg'(z)/g(z) = n/m$. The normalizing factor which keeps appearing in the sequel is given by

$$[z^n]g(z)^m = \frac{g(\rho)^m}{\rho^n\sqrt{2\pi m\, s^2}}(1+o(1)),$$

with $s^2 = \rho^2 g''(\rho)/g(\rho) - (\rho g'(\rho)/g(\rho))^2 + \rho g'(\rho)/g(\rho)$. In the frequent case where $g(z) = e^z$ we get $\rho = n/m$ and $s^2 = \rho$ and thus the above equation transforms to

$$[z^n]\{e^{mz}\} \sim \frac{m^n}{\sqrt{2\pi n}\,(n/e)^n},$$

and we reobtain Stirling's approximation of $n!$.

To cope with the multivariate distribution define $\phi_{\mathcal{E},d}(x_1,\ldots,x_d; z_1,\ldots,z_d)$ as the generating function enumerating all the possible allocations in a single urn, where we use the variables $z_1,\ldots,z_d$ to mark the balls allocated before the time θ_1, then between θ_1 and θ_2, etc., and the variables $x_1,\ldots,x_d$ to mark the urns U such that $Y(U) \in \mathcal{E}$ at the times $\theta_1,\ldots,\theta_d$. Of course the generating function relative to the system of m urns will be $\Phi_{\mathcal{E},d} = \phi^m_{\mathcal{E},d}$. Now we can get a recurrence equation on the $\phi_{\mathcal{E},d}$ by a "renewal" argument as follows.

Consider a sequence of times $\theta_1,\ldots,\theta_d$, and partition the allocations into an urn U according to the first time θ_l when $Y(U) \in \mathcal{E}$ $(1 \le l \le d)$, i.e., $l = \min\{i : Y(U,\theta_i) \in \mathcal{E}$ where $Y(U,\theta_i)$ equals to $Y(U)$ evaluated at time θ_i.

For the case where $Y(U,\cdot)$ never belongs to $\mathcal{E}$, i.e., for $l = \infty$, define $K_{d+1}(z_1,\ldots,z_d)$ as the function enumerating those allocations. Of course, there is no occurrence of any x_i.

If $Y(U,\cdot) \in \mathcal{E}$ does not hold for $\theta_1,\ldots,\theta_{l-1}$, but holds for θ_l, then we enumerate the allocations up to and including time θ_l by a function $x_l K_l(z_1,\ldots,z_l)$. Note, that only z_i for $i \le l$ appear, since we stop counting at time θ_l. Hence we must also enumerate the allocations after the time θ_l. For $i > l$, define $Z_i(U) := Y(U,\theta_i) - Y(U,\theta_l)$, which can due to additivity be interpreted as the value of Y if the allocations between θ_l (excluded) and θ_i had been done into an empty urn. Assume that $Y(U,\theta_l) = r$ for some $r \in \mathcal{E}$ and define

$$\mathcal{E} - r := \{y : y + r \in \mathcal{E}\}.$$

Then $Y(U,\theta_i) \in \mathcal{E}$ if and only if $Z(U) \in \mathcal{E} - r$. Hence the allocations after the time θ_l, knowing that Y has value r at the time θ_l, are described by the function $\phi_{\mathcal{E}-r,d-l}(x_{l+1},\ldots,x_d; z_{l+1},\ldots,z_d)$.

Putting all this together gives

$$\begin{aligned} \phi_{\mathcal{E},d}(x_1,\ldots,x_d; z_1,\ldots,z_d) &= K_{d+1}(z_1,\ldots,z_d) \\ &+ \sum_{l=1}^{d} x_l K_l(z_1,\ldots,z_l) \sum_{r\in\mathcal{E}} \phi_{\mathcal{E}-r,d-l}(x_{l+1},\ldots,x_d; z_{l+1},\ldots,z_d). \end{aligned} \tag{2.3}$$

For the two classes of examples mentioned in the introduction we obtain:

- For $Y(U)$ equal to the number of balls in the urn,
$$K_l(z_1,\dots,z_l) = \sum_{n_1,\dots,n_l} \frac{z_1^{n_1}}{n_1!}\cdots\frac{z_l^{n_l}}{n_l!} \qquad (l \le d)$$
where the sum is for $n_1,\dots,n_l$ such that $n_1 \notin \mathcal{E}, \dots, n_1+\dots+n_{l-1} \notin \mathcal{E}$ and $n_1+\dots+n_l \in \mathcal{E}$, and
$$K_{d+1}(z_1,\dots,z_d) = \sum_{n_1,\dots,n_d} \frac{z_1^{n_1}}{n_1!}\cdots\frac{z_l^{n_d}}{n_d!},$$
where the sum is for $n_1,\dots,n_d$ such that $n_1 \notin \mathcal{E}, \dots, n_1+\dots+n_d \notin \mathcal{E}$.
- For colored balls and $Y(U)$ equal to the balance of the urn,
$$K_l(z_1,\dots,z_l) = \sum_{n_1,\dots,n_l} I_{n_1}(2z_1)\dots I_{n_l}(2z_l),$$
where the sum is for $n_1,\dots,n_l$ such that $n_1 \notin \mathcal{E}, \dots, n_1+\dots+n_{l-1} \notin \mathcal{E}$ and $n_1+\dots+n_l \in \mathcal{E}$, and where the $I_n(z)$ are Bessel functions (details see [3]), and
$$K_{d+1}(z_1,\dots,z_d) = \sum_{n_1,\dots,n_d} I_{n_1}(2z_1)\dots I_{n_l}(2z_l),$$
where the sum is for $n_1,\dots,n_d$ such that $n_1 \notin \mathcal{E}, \dots, n_1+\dots+n_d \notin \mathcal{E}$.

We will show the following theorem:

Theorem 2.1. *Let $X_m(\lfloor mt \rfloor)$, $t \ge 0$, be the process associated to allocating balls into urns of a general urn model such that the generating functions describing the allocation process have the shape*
$$\Phi_{\mathcal{E},d}(x_1,\dots,x_d,z_1,\dots,z_d) = \phi_{\mathcal{E},d}(x_1,\dots,x_d,z_1,\dots,z_d)^m$$
where $\phi_{\mathcal{E},d}$ satisfies a recurrence relation of the form (2.3) with entire functions K_i. Then the following weak limit theorem holds:
$$Y_m(t) := \frac{X_m(\lfloor mt \rfloor) - EX_m(\lfloor mt \rfloor)}{\sqrt{m}} \xrightarrow{w} G(t)$$
where $G(t)$ is a centered Gaussian process with continuous sample paths. The covariance function $B_{s,t}$, $s,t \ge 0$, is given by
$$B_{s,t} = B_{t,s} = \left.\frac{\partial^2(\log \lambda_{s,t}(e^{u_1},e^{u_2}))}{\partial u_1 \partial u_2}\right|_{u_1=0,u_2=0} \tag{2.4}$$
if $s < t$ with
$$\lambda_{s,t}(x_1,x_2) = \frac{\phi_{\mathcal{E},2}(x_1,x_2,r_1,r_2)}{r_1^s r_2^{t-s}}$$
where $r_1 = r_1(x_1,x_2,s,t)$ and $r_2 = r_2(x_1,x_2,s,t)$ are the saddle points, which are defined by the equations in z_1 and z_2
$$z_1\,\partial\phi_{\mathcal{E},2}/\partial z_1 = s\phi_{\mathcal{E},2}; \tag{2.5}$$
$$z_2\,\partial\phi_{\mathcal{E},2}/\partial z_2 = (t-s)\phi_{\mathcal{E},2}, \tag{2.6}$$
and by
$$B_{s,s} = \left.\frac{\partial^2(\log \lambda_s(e^u))}{\partial^2 u}\right|_{u=0} \tag{2.7}$$
with
$$\lambda_s(x) = \frac{\phi_{\mathcal{E},1}(x,r)}{r^s}$$
where $r = r(x,s)$ is the saddle point defined by the equation in z
$$z\,\partial\phi_{\mathcal{E},1}/\partial z = s\phi_{\mathcal{E},1}. \tag{2.8}$$

Remark 1. Note that by the limiting process $G(t)$ is a Markov process if and only if the covariance can be factorized in the form $B_{s,t} = b_1(s)b_2(t)$ (see [12]).

Remark 2. It should be mentioned that the assumption of the above functions to be entire is not a necessity. We actually require that any saddle point considered during the evaluation of a Cauchy integral throughout the proof is closer to the origin than any singularity of the integrand.

Corollary. *Under the assumptions of Theorem 2.1, the limiting variance is*

$$VarX_m(n) \sim m\frac{f}{g}\left(1 - \frac{f}{g}\left[1 + \frac{sg'^2}{sgg'' - (s-1)g'^2}\left(1 - \frac{\rho f'}{sf}\right)^2\right]\right),$$

where the functions f, g and their derivatives are evaluated at the point ρ solution of the saddle point equation : $zg'(z) = sg(z)$ with $s = n/m$.

As a consequence, we have a general formula for the variance, for a fixed allocation scheme described by the function $g(z)$. We give below the variances for some examples we shall consider in the next subsection.

Corollary. *For the classical allocation scheme ($g(z) = e^z$),*

$$VarX_m(n) \sim mf(s)e^{-s}\left(1 - f(s)e^{-s}\left[1 + s\left(1 - \frac{f'(s)}{f(s)}\right)^2\right]\right).$$

For the allocation scheme on bounded urns ($g(z) = (1+z)^\delta$), and taking $\rho = s/(\delta - s)$,

$$VarX_m(n) \sim mf(\rho)\left(1 - \frac{s}{\delta}\right)^\delta\left(1 - f(\rho)\left(1 - \frac{s}{\delta}\right)^\delta\left[1 + \frac{\delta s}{\delta - s}\left(1 - \frac{f'(\rho)}{(\delta - s)f(\rho)}\right)^2\right]\right).$$

For the classical allocation scheme with colored balls ($g(z) = e^{2z}$),

$$VarX_m(n) \sim mf(s/2)e^{-s}\left(1 - f(s/2)e^{-s}\left[1 + s\left(1 - \frac{f'(s/2)}{2f(s/2)}\right)^2\right]\right).$$

For the allocation scheme on bounded urns and colored balls ($g(z) = (1+2z)^\delta$), and taking $\rho = s/2(\delta - s)$,

$$VarX_m(n) \sim mf(\rho)\left(1 - \frac{s}{\delta}\right)^\delta\left(1 - f(\rho)\left(1 - \frac{s}{\delta}\right)^\delta\left[1 + \frac{\delta s}{\delta - s}\left(1 - \frac{f'(\rho)}{2(\delta - s)f(\rho)}\right)^2\right]\right).$$

2.3. **Further examples.** We present in this part applications of our theorem to some problems relative to the number of urns satisfying some condition of the kind *specified number of balls* or *specified balance.* For each of them, we shall give the basic generating functions $g(z)$ and $f(z)$, the multivariate functions Φ_d (with an emphasis on Φ_1 and Φ_2, which determine the moments, hence the limiting Gaussian process), and either the exact formulae or the asymptotic expressions for the mean value, variance and covariance. Asymptotics for the mean value and variance are for $n/m \to \theta$; for the covariance the number of balls up to the (normalized) time θ_1 is n_1 such that $n_1/m \to \theta_1$ and the number of balls between the times θ_1 and θ_2 is n_2 such that $(n_1+n_2)/m \to \theta_2$. Some of the results presented below can be found in the literature, others, to the best of our knowledge, are new.

2.3.1. *Variations on the number of empty urns.* In the classical case, the balls are undistinguishable, $g(z) = e^z$, the valuation $Y(U)$ is equal to the number of balls in the urn, the set $\mathcal{E}$ is $\{0\}$, and $f(z) = 1$. A variation of this model occurs when studying some database parameter : the size of a projection in a relation without functional dependency [6]. Roughly speaking, the parameter of interest is the number of non-empty urns, when the balls are distinguishable and the urns have a bounded size δ. Such an urn can be seen as a sequence of δ distinguishable cells, each of which can receive at most one ball; allocating n balls into an urn is equivalent to choosing the n cells that receive a ball : $g_n = \binom{\delta}{n}$ and $g(z) = (1+z)^\delta$.

This leads us to consider the number of empty urns in a general case, when the allocation of balls into an urn is described by an ordinary or exponential function $g(z)$; this covers the classical case of empty urns ($g(z)$ exponential and equal to e^z) and the case of projections ($g(z)$ ordinary and equal to $(1+z)^\delta$). The set $\mathcal{E}$ is the same, and $f(z) = 1$. The generating function describing empty urns is

$$\Phi_1(x;z) = (g(z) + x - 1)^m .$$

The generating function for the bi-dimensional distribution is

$$\Phi_2(x_1, x_2; z_1, z_2) = ((g(z_1) - 1)g(z_2) + x_1(g(z_2) - 1) + x_1 x_2)^m .$$

The multivariate generating function associated to the finite-dimensional distribution is an extension of the function for the classical case :

$$\phi_d(x_1, \dots, x_d, z_1, \dots, z_d) = \sum_{0 \le j \le d} x_1 \dots x_j \; f_j(z_1, \dots, z_d); \tag{2.9}$$

$$f_j(z_1, \dots, z_d) = (g(z_{j+1}) - 1)\; g(z_{j+2}) \cdots g(z_d). \tag{2.10}$$

Let us define $\gamma(i,n) = [z^n]\{g(z)^{m-i}\}$. When the parameter n is proportional to m, we can get an asymptotic development of the coefficients $\gamma(i,n)$ for fixed $i \in \{0,1,2\}$ [4]. The mean value is

$$EX_m(n) = m\,\frac{\gamma(1,n)}{\gamma(0,n)} \sim \frac{m}{g(\rho)},$$

with ρ defined by the saddle point equation $zg'(z)/g(z) = \theta$. The variance is

$$VarX_m(N) = m(m-1)\frac{\gamma(2,n)}{\gamma(0,n)} + m\frac{\gamma(1,n)}{\gamma(0,n)} - \left(m\frac{\gamma(1,n)}{\gamma(0,n)}\right)^2$$

and the covariance at (normalized) times θ_1 and θ_2 is

$$\begin{aligned} Cov[X(n_1), X(n_1+n_2)] &= m^2 \left(\frac{\gamma(2,n_1)}{\gamma(0,n_1)}\,\frac{\gamma(1,n_2)}{\gamma(0,n_2)} - \frac{\gamma(1,n_1)}{\gamma(0,n_1)}\,\frac{\gamma(1,n_1+n_2)}{\gamma(0,n_1+n_2)}\right) \\ &\quad + m\left(\frac{\gamma(1,n_1)}{\gamma(0,n_1)} - \frac{\gamma(2,n_1)}{\gamma(0,n_1)}\right)\frac{\gamma(1,n_2)}{\gamma(0,n_2)}. \end{aligned}$$

Applying this to $g(z) = (1+z)^\delta$, we find again (cf. [6, 8]) that the projection size converges weakly to a non-Markovian process with mean value and variance

$$\begin{aligned} EX_m(n) &= m\left(1 - \frac{\binom{(m-1)\delta}{n}}{\binom{m\delta}{n}}\right) \sim m\left[1 - \left(1 - \frac{\theta}{\delta}^{\delta}\right)\right]; \\ VarX_m(n) &= m^2\left[\frac{\binom{(m-2)\delta}{n}}{\binom{m\delta}{n}} - \frac{\binom{(m-1)\delta}{n}^2}{\binom{m\delta}{n}^2}\right] + m\left[\frac{\binom{(m-2)\delta}{n}}{\binom{m\delta}{n}} + \frac{\binom{(m-1)\delta}{n}}{\binom{m\delta}{n}}\right] \\ &\sim m\left(1 - \frac{\theta}{\delta}\right)^\delta \left[1 - \left(1 - \frac{\theta}{\delta}\right)^\delta \left(1 + \frac{\delta\theta}{\delta - \theta}\right)\right]. \end{aligned}$$

Its covariance is

$$\begin{aligned} Cov\quad [X_m(n_1), X_m(n_1+n_2)] &= m^2\,\frac{\binom{(m-1)\delta}{n_1+n_2}}{\binom{m\delta}{n_1+n_2}}\left(1 - \frac{\binom{(m-1)\delta}{n_1}}{\binom{m\delta}{n_1}}\right) \\ &\quad - m(m-1)\,\frac{\binom{(m-1)\delta}{n_1}}{\binom{m\delta}{n_1}}\left(\frac{\binom{(m-1)\delta}{n_1}}{\binom{m\delta}{n_1}} - \frac{\binom{(m-2)\delta}{n_1}}{\binom{m\delta}{n_1}}\right) \\ \sim\ m\left(1 - \frac{\theta_1}{\delta}\right)^\delta &\left(1 - \frac{\theta_2 - \theta_1}{\delta}\right)^\delta \left[1 - \left(1 - \frac{\theta_1}{\delta}\right)^\delta \left(1 + \frac{\delta\theta_1}{\delta - \theta_1}\right)\right]. \end{aligned}$$

2.3.2. *Number of urns when the number of balls satisfies some condition.* We begin by studying the *number of urns with exactly r balls* (see [11] for a different presentation of some of these results); of course this includes the number of empty urns. Here again, the balls are undistinguishable, with $g(z) = e^z$, and $Y(U)$ is the number of balls in the urn. The set $\mathcal{E}$ is $\{r\}$, and $f(z) = z^r/r!$.

$$\Phi_1(x; z) = \left(e^z + (x-1)\frac{z^r}{r!}\right)^m .$$

The mean value is

$$EX_m(n) = \binom{n}{r}\frac{(m-1)^{n-r}}{m^n} \sim m\frac{\theta^r}{r!}e^{-\theta}\left(1 + \frac{(2r-\theta)\,\theta - r(r-1)}{2m\theta}\right).$$

The variance is

$$\begin{aligned} VarX_m(n) &= m(m-1)\binom{n}{r,r}\frac{(m-2)^{n-2r}}{m^n} + \binom{n}{r}\frac{(m-1)^{n-r}}{m^n} \\ &\quad - \left(\binom{n}{r}\frac{(m-1)^{n-r}}{m^n}\right)^2 \\ &\sim m\frac{\theta^r}{r!}e^{-\theta}\left(1 - \frac{\theta^r}{r!}e^{-\theta}\left(1 + \frac{(r-\theta)^2}{\theta}\right)\right). \end{aligned}$$

Define $h(z_1, z_2) := \sum_{0 \le k < r} (z_1^k/k!)\,(z_2^{r-k}/(r-k)!)$: this function enumerates the allocations that put strictly less than r balls into an urn at the time t_1, and exactly r balls at the time t_2. Then

$$\begin{aligned} &\Phi_2(x_1, x_2; z_1, z_2) = \\ &\left(\left(e^{z_1} - \frac{z_1^r}{r!}\right)e^{z_2} - h(z_1, z_2) + x_1\frac{z_1^r}{r!}(e^{z_2} - 1) + x_2 h(z_1, z_2) + x_1 x_2 \frac{z_1^r}{r!}\right)^m . \end{aligned}$$

For any urn, the condition $Y(U) = r$ is not satisfied at the beginning, and may never hold; if at some point it is satisfied, after some time it will cease to hold. The functions describing the finite-dimensional distributions are

$$\phi_d = e^{z_1 + \cdots z_d} + \sum_{i=1}^{d} f_i(z_1, \ldots z_i)\left(x_i \nu_{d-i}(z_{i+1}, \ldots, z_d; x_{i+1}, \ldots, x_d) - e^{z_{i+1} + \cdots + z_d}\right),$$

where f_i is defined by

$$f_{i+1} = \sum_{l_1 + \cdots + l_i < r} \frac{z_1^{l_1}}{l_1!} \cdots \frac{z_i^{l_i}}{l_i!}\,\frac{z_{i+1}^{r-(l_1+\cdots+l_i)}}{(r - l_1 - \cdots l_i)!},$$

and where ν_d describes the allocation of d batches into a single urn, and is the function we met while studying empty urns : ν_d is the function ϕ_d of Equation (2.2). The covariance is [11, p. 181]

$$\begin{aligned} &Cov[X(n_1), X(n_1 + n_2)] = \\ &m\binom{n_1}{r}\left(1 - \frac{1}{m}\right)^{n_2}\left(\frac{1}{m}\right)^r . \\ &\left(\left(1 - \frac{1}{m}\right)^{n_1 - r} - \left(1 - \frac{2}{m}\right)^{n_1 - r} + m\left(\left(1 - \frac{2}{m}\right)^{n_1 - r} - \binom{n_1 + n_2}{r}\left(1 - \frac{1}{m}\right)^{2n_1 - r}\right)\right) \\ &\sim m\,\frac{\theta_1^r}{r!}e^{-\theta_2}\left(1 - \frac{\theta_2^r}{r!}e^{-\theta_1}\left(1 + \frac{(r-\theta_1)(r-\theta_2)}{\theta_2}\right)\right). \end{aligned}$$

When $r = 0$, we get back the expression for empty urns. For $r \neq 0$, the covariance cannot be factored w.r.t. θ_1 and θ_2, and the limiting process is not Markovian.

We can extend these results when the allocation of balls into a single urn is described by a function $g(z) = \sum_i g_i z^i$. The function $f(z)$ is now $g_r z^r$, and

$$\Phi_1(x; z) = (g(z) + (x-1) g_r z^r)^m .$$

The mean value is

$$EX_m(n) = mg_r \frac{[z^{n-r}]g(z)^{m-1}}{[z^n]g(z)^m} \sim m\, \frac{g_r\, \rho(\theta)^r}{g(\rho(\theta))},$$

with $\rho(\theta)$ defined, as usual, by the saddle-point equation : $zg^{'}(z)/g(z) = \theta$. The variance is

$$\begin{aligned} VarX_m(n) &= m(m-1)g_r^2 \frac{[z^{n-2r}]g(z)^{m-2}}{[z^n]g(z)^m} + mg_r \frac{[z^{n-r}]g(z)^{m-i}}{[z^n]g(z)^m} \\ &\quad -m^2 g_r^2 \left(\frac{[z^{n-r}]g(z)^{m-i}}{[z^n]g(z)^m} \right)^2 . \end{aligned}$$

The function enumerating the allocations that put strictly less than r balls into an urn at the time t_1, and a total of exactly r balls at the time t_2, is $h(z_1, z_2) := \sum_{0 \le k < r} g_k\, g_{r-k}\; z_1^k\, z_2^{r-k}$. Then

$$\begin{aligned} &\Phi_2(x_1, x_2; z_1, z_2) = \\ &\left((g(z_1) - g_r z_1^r)\, g(z_2) - h(z_1, z_2) + x_1 g_r z_1^r (g(z_2) - 1) + x_2 h(z_1, z_2) + x_1 x_2 g_r z_1^r \right)^m . \end{aligned}$$

The functions describing the finite-dimensional distributions are $\Phi_d = \phi_d^m$, with

$$\begin{aligned} \phi_d &= g(z_1) \cdots g(z_d) \\ &\quad + \sum_{i=1}^{d} f_i(z_1, \dots z_i) \left(x_i \nu_{d-i}(z_{i+1}, \dots, z_d; x_{i+1}, \dots, x_d) - g(z_{i+1}) \cdots g(z_d) \right), \end{aligned}$$

where ν_d is again the function ϕ_d associated with empty urns, and is defined here by the equation (2.9), and where the f_i are such that

$$f_{i+1} = \sum_{l_1 + \cdots + l_i < r} g_{l_1} z_1^{l_1} \dots g_{l_i} z_i^{l_i}\, g_{(r - l_1 - \cdots l_i)} z_{i+1}^{r-(l_1+\cdots+l_i)}.$$

The covariance can be expressed as a function of the coefficients $\gamma(i, n) := [z^n]g(z)^{m-i}$:

$$\begin{aligned} Cov(X_m(n_1), X_m(n_1+n_2)) &= mg_r \frac{\gamma(1, n_2)}{\gamma(0, n_2)} . \left(\frac{\gamma(1, n_1 - r)}{\gamma(0, n_1)} + (m-1) \frac{\gamma(2, n_1 - r)}{\gamma(0, n_1)} \right) \\ &\quad -m^2 g_r^2 \frac{\gamma(1, n_1 + n_2 - r)}{\gamma(0, n_1 + n_2)}\; \frac{\gamma(1, n_1 - r)}{\gamma(0, n_1)} . \end{aligned}$$

For example, if we consider bounded urns, we get

$$\begin{aligned} EX_m(n) &\sim m \binom{\delta}{r} \left(\frac{\theta}{\delta - \theta} \right)^r \left(1 - \frac{\theta}{\delta} \right)^\delta ; \\ VarX_m(n) &\sim m \binom{\delta}{r} \left(\frac{\theta}{\delta - \theta} \right)^r \left(1 - \frac{\theta}{\delta} \right)^\delta \left[1 - \binom{\delta}{r} \left(\frac{\theta}{\delta - \theta} \right)^r \left(1 - \frac{\theta}{\delta} \right)^\delta \left(1 + \frac{\delta (\theta - r)^2}{\theta (\delta - \theta)} \right) \right] . \end{aligned}$$

As regards the asymptotic covariance, we can compute its asymptotic value for a given r.

If we consider now the *number of urns with at most r balls,* in the classical case $g(z) = e^z$, the set $\mathcal{E}$ is $[0...r]$ and $f(z) = e_r(z)$, with $e_r(z) = \sum_{0 \le i \le r} z^i / i!$.

$$\Phi_1(x; z) = \left(e^z - e_r(z) + x e_r(z) \right)^m .$$

The average number of urns with at most r balls is

$$EX_m(n) = m \frac{n!}{m^n} [z^n] \{ e_r(z)\, e^{(m-1)z} \} = m \sum_{i=0}^{r} \binom{n}{i} (1 - 1/m)^{n-i}\, m^{-i} \sim m e_r(\theta) e^{-\theta}.$$

The variance can be obtained explicitely; its asymptotic value is

$$VarX_m(n) \sim m\, e^{-\theta} \left(e_r(\theta) - e^{-\theta} \left[e_r^2(\theta) + \theta^{2r+1} / r!^2 \right] \right).$$

The generating function marking the urns that have at most r balls, after throwing the balls in two batches, is $\Phi_2 = \phi_2^m$, with

$$\phi_2(x_1, x_2; z_1, z_2) =$$
$$[e^{z_1} - e_r(z_1)]\, e^{z_2} + x_1 \left[e_r(z_1)\, e^{z_2} - \sum_{0 \le i \le r} \frac{z_1^i}{i!}\, e_{r-i}(z_2) \right] + x_1 x_2 \sum_{0 \le i \le r} \frac{z_1^i}{i!}\, e_{r-i}(z_2).$$

The finite-dimensional distribution is described by the function

$$\phi_d = e^{z_1 + \cdots + z_l} + \sum_{j=1}^{l-1} x_1 \ldots x_j \ (e^{z_{j+1}} f_j - f_{j+1})\ e^{z_{j+2} + \cdots + z_d} + x_1 \ldots x_l\ f_l,$$

where the functions f_j are defined, for $j \ge 1$, as

$$f_j(z_1, \ldots, z_j) = \sum_{i_1 + \cdots + i_j \le r} \frac{z_1^{i_1}}{i_1!} \cdots \frac{z_j^{i_j}}{i_j!}$$

The covariance is

$$\begin{aligned}
Cov\quad & [X_m(n_1), X_m(n_1+n_2)] = m \sum_{0 \le k \le r} \binom{n_1+n_2}{k} \left(1 - \frac{1}{m}\right)^{n_1+n_2-k} \left(\frac{1}{m}\right)^k \\
& + m(m-1) \sum_{0 \le i \le i+j \le r; i \le p \le i+r} \binom{n_1}{p}\binom{n_2}{j}\binom{p}{i} \left(1 - \frac{2}{m}\right)^{n_1-p} \left(1 - \frac{1}{m}\right)^{n_2-j} \left(\frac{1}{m}\right)^{p+j} \\
& - m^2 \left(\sum_{0 \le i \le r} \binom{n_1}{i} \left(1 - \frac{1}{m}\right)^{n_1-i} \left(\frac{1}{m}\right)^i \right) \left(\sum_{0 \le k \le r} \binom{n_1+n_2}{k} \left(1 - \frac{1}{m}\right)^{n_1+n_2-k} \left(\frac{1}{m}\right)^k \right) \\
\sim\quad & m\, e^{-\theta_2} \left(e_r(\theta_2) - e^{-\theta_1} \left(e_r(\theta_1)\, e_r(\theta_2) + \frac{\theta_1^{r+1} \theta_2^r}{r!} \right) \right).
\end{aligned}$$

In the general case, where the function describing the allocation of balls into an urn is no longer e^z, but any function $g(z) = \sum_i g_i z^i$ with suitable coefficients, we can get a similar expression for the functions ϕ_d. We shall use the functions $h_r(z) := \sum_{i \le r} g_i z^i$ and $k(z_1, z_2) := \sum_{i,r: i<r} g_i z_1^i h_{r-i}(z_2)$; then

$$\begin{aligned}
\Phi_1(x; z) &= (g(z) + (x-1) h_r(z))^m\,; \\
\Phi_2(x_1, x_2; z_1, z_2) &= (g(z_1) g(z_2) + (x_1 - 1) h_r(z_1) g(z_2) + x_1 (x_2 - 1) k(z_1, z_2))^m\,.
\end{aligned}$$

The mean value is asymptotically

$$EX_m(n) \sim m \frac{h_r(\rho)}{g(\rho)},$$

with ρ defined, as usual, by the equation $zg^{'}(z)/g(z) = \theta$. Now the multivariate functions are given by

$$\begin{aligned}
\phi_d &= g(z_1) \cdots g(z_d) + \\
& \quad \sum_{j=1}^{d-1} x_1 \ldots x_j \ (g(z_{j+1}) f_j - f_{j+1})\ g(z_{j+2}) \cdots g(z_d) + x_1 \ldots x_d\ f_d; \\
with \qquad f_j &= \sum_{i_1 + \cdots + i_j \le r} g_{i_1} \ldots g_{i_j}\, z_1^{i_1} \ldots z_j^{i_j}.
\end{aligned}$$

2.3.3. *Urns with colored balls.* Here the balls can have two colors, let's say red and blue, and the state of an urn is defined by its *balance,* i.e. by the difference *number of blue balls minus number of red balls*, which is a relative integer : the valuation $Y(U)$ is the balance of the urn. We begin by the *number of balanced urns.* We have undistinguishable balls : $g(z) = e^{2z}$; the condition to be satisfied is : *Either the urn is empty, or the numbers of blue and red balls are equal,* which corresponds to a set $\mathcal{E} = \{0\}$, and the function describing the allocations leading to a balanced urn is a Bessel coefficient : $f(z) = I_0(2z) = \sum_{n \geq 0} z^{2n}/(n!)^2$. We have

$$\Phi_1(x; z) = \left(e^{2z} + (x-1)I_0(2z)\right)^m .$$

The average number of balanced urns is

$$EX_m(n) = m \sum_p \binom{n}{p, p} \left(1 - \frac{1}{m}\right)^{n-2p} \left(\frac{1}{2m}\right)^{2p} \sim m I_0(\theta)\, e^{-\theta}.$$

The exact variance has a complicated expression, and is given in [3]; its asymptotic value is

$$VarX_m(n) \sim m e^{-\theta} \left(I_0(\theta) - e^{-\theta} I_0^2(\theta) - \theta e^{-\theta}[I_0(\theta) - I_1(\theta)]^2\right).$$

The generating function marking the urns that are balanced at two different times is $\Phi_2 = \phi_2^m$, with

$$\begin{aligned}
\phi_2(x_1, x_2; z_1, z_2) = \quad & e^{2z_1+2z_2} - I_0(2z_1)e^{2z_2} - I_0(2(z_1+z_2)) + I_0(2z_1)I_0(2z_2) \\
& + x_1 I_0(2z_1)[e^{2z_2} - I_0(2z_2)] \\
& + x_2[I_0(2(z_1+z_2)) - I_0(2z_1)\, I_0(2z_2))] \\
& + x_1 x_2 I_0(2z_1)\, I_0(2z_2).
\end{aligned}$$

The asymptotic covariance is

$$m\, e^{-\theta_2}[I_0(\theta_1)I_0(\theta_2-\theta_1) - e^{-\theta_1}[\theta_1(I_0(\theta_1) - I_1(\theta_1))(I_0(\theta_2) - I_1(\theta_2)) + I_0(\theta_1)I_0(\theta_2)]].$$

When we throw the balls in d batches, we obtain

$$K_l(z_1, ..., z_l) := \sum_{q_1, \cdots, q_l} I_{q_1}(2z_1)...I_{q_l}(2z_l), \tag{2.11}$$

where the summation is on $q_1, ..., q_l$ such that $q_1 \neq 0$, $q_1 + q_2 \neq 0$, ..., $q_1 + ... + q_{l-1} \neq 0$ (the urn is not balanced at $\theta_1, ..., \theta_{l-1}$), but $q_1 + ... + q_l = 0$ (the urn is balanced at the time θ_l). This formula extends to $l = d+1$, but here the summation is on $q_1, ..., q_{d+1}$ such that $q_1 \neq 0$, $q_1 + q_2 \neq 0$, ..., and $q_1 + ... + q_{d+1} \neq 0$. For a given l, it is possible to simplify further the functions K_l, using the following property of the Bessel coefficients, where the summation is for relative integers q_i such that $q_1 + ... + q_l = n$:

$$I_n(z_1 + ... + z_l) = \sum_{q_1, \cdots, q_l} I_{q_1}(z_1)...I_{q_l}(z_l).$$

Such a transformation was already applied to obtain the expression of Φ_2 given above; for $d = 3$ for example we obtain

$$\begin{aligned}
K_3(z_1, z_2, z_3) \quad = \quad & I_0(2z_1 + 2z_2 + 2z_3) - I_0(2z_1 + 2z_2)I_0(z_3) \\
& - I_0(2z_1)I_0(2z_2 + 2z_3)) + I_0(2z_1)I_0(2z_2)I_0(2z_3).
\end{aligned}$$

We can extend the preceding results for a general allocation scheme described by a function $g(z)$. We shall use the functions f_q enumerating the allocations into an urn that lead to a balance q in this urn :

$$f_q(y) \quad = \quad [z^q] g\left(y\left(z + \frac{1}{z}\right)\right) = \sum_p \binom{q+2p}{p} g_{q+2p}\, y^{q+2p}.$$

For $q = 0$ we obtain $f_0(y) = \sum_p \binom{2p}{p} g_{2p} y^{2p}$, and

$$\Phi_1(x; z) = \left(g(2z) + (x-1)f_0(z)\right)^m .$$

The average number of balanced urns is

$$EX_m(n) = m\frac{[z^n]\{f_0(z)g(2z)^{m-1}\}}{[z^n]g(2z)^m} \sim mf_0(\rho)/g(2\rho),$$

with ρ defined by the equation $2zg'(2z)/g(2z) = \theta$. When we throw the balls in d batches, the only difference with the classical case is in the definition of the functions K_l, which will not simplify as much as when dealing with Bessel coefficients. The equations defining these functions are

$$K_l(z_1, ..., z_l) \quad = \quad \sum_{q_1,\cdots,q_l} f_{q_1}(z_1)...f_{q_l}(z_l) \qquad (1 \le l \le d+1)$$

where the summations are the same as the ones for the equation (2.11).

Consider now the *number of urns with balance q.* We begin with the classical case : $g(z) = e^z$. The set $\mathcal{E}$ is $\{q\}$, and the function describing this is $f(z) = I_q(2z)$, with $I_q(z) = \sum_{n:n+q\ge 0}(z/2)^{2n+q}/n!(n+q)!$ a Bessel coefficient :

$$\Phi_1(x;z) = \left(e^{2z} + (x-1)I_q(2z)\right)^m.$$

The average number of urns with balance q is

$$EX_m(n) = m\sum_p \binom{n}{p, p+q}\left(1-\frac{1}{m}\right)^{n-2p-q}\left(\frac{1}{2m}\right)^{2p+q} \sim mI_q(\theta)\, e^{-\theta},$$

and its asymptotic variance is

$$VarX_m(n) \sim me^{-\theta}\left(I_q(\theta) - e^{-\theta}I_q^2(\theta) - \frac{1}{4}\,\theta e^{-\theta}(I_{q-1}(\theta) - 2I_q(\theta) + I_{q+1}(\theta))^2\right).$$

What about the multivariate function ϕ_d associated to d batches? For $d = 2$, we have

$$\begin{aligned}\phi_2(x_1, x_2; z_1, z_2) \quad = \quad & e^{2z_1+2z_2} - [I_q(2z_1+2z_2) - I_q(2z_1)(e^{2z_2} - I_0(2z_2))] \\ & +x_1 I_q(2z_1)(e^{2z_2} - I_0(2z_2)) \\ & +x_2[I_q(2z_1+2z_2) - I_q(2z_1)I_0(2z_2)] \\ & +x_1x_2I_q(2z_1)I_0(2z_2).\end{aligned}$$

The asymptotic covariance is

$$m\, e^{-\theta_2}(I_q(\theta_1)I_0(\theta_2-\theta_1) - e^{-\theta_1}F_q(\theta_1,\theta_2)),$$

with

$$\begin{aligned}F_q(\theta_1,\theta_2) \quad := \quad & \theta_1(I_q(\theta_1) - I_{q+1}(\theta_1))(I_q(\theta_2)(1-q/\theta_2) - I_{q+1}(\theta_2)) \\ & -qI_q(\theta_1)(I_q(\theta_2) - I_{q+1}(\theta_2)) + I_q(\theta_1)I_q(\theta_2)(1+q^2/\theta_2).\end{aligned}$$

For general d, the multivariate functions are defined from

$$K_l(z_1, ..., z_l) = \sum_{q_1,\cdots,q_l} I_{q_1}(2z_1)...I_{q_l}(2z_l),$$

where the summation for $1 \le l \le d$ is on $q_1, ..., q_l$ such that $q_1 \ne p$, $q_1+q_2 \ne p$, ..., $q_1+...+q_{l-1} \ne p$ but $q_1 + ... + q_l = p$; for $l = d+1$ the summation is on $q_1, ..., q_{d+1}$ such that $q_1 \ne p$, $q_1 + q_2 \ne p$, ..., and $q_1 + ... + q_{d+1} \ne p$. Of course, for $q = 0$ all the results of this part are simply those relative to the number of balanced urns.

When considering urns with balance q in the general case, we obtain results similar to the preceding ones, deduced from them by substituting $g(z)$ for e^z, and general f_q for the I_q. We have to take care not to use any expression of the kind $I_p(z_1 + z_2 + ...)$, which is usually obtained from a summation on products of some functions I_q, by taking advantage of properties of Bessel coefficients, but otherwise the results translate nicely.

We finally turn to the *number of urns with (strictly) positive balance.* In the classical case, we have $g(z) = e^z$, and the set $\mathcal{E}$ becomes $\mathcal{E} = N$. The function enumerating the states satisfying $\mathcal{E}$ is

$$f(z) = \sum_{q>0} I_q(2z).$$

We have that

$$\Phi_1(x;z) = \left(e^{2z} + (x-1)f(z)\right)^m.$$

The average number of urns with positive balance is

$$EX_m(n) = m\frac{[z^n]\{f(z)g(2z)^{m-1}\}}{[z^n]\{g(2z)^m\}} \sim me^{-\theta} f(\theta/2) \sim me^{-\theta} \sum_{q>0} I_q(\theta).$$

To simplify the asymptotic variance, we use the relation $I_q^{'}(t) = (1/2)(I_{q-1}(t) + I_{q+1}(t))$, which gives $f^{'}(z) = 2f(z) + I_0(2z) - I_1(2z)$; we get

$$VarX_m(n) \sim me^{-\theta} \left(f(\theta/2) - e^{-\theta} f^2(\theta/2) - \theta e^{-\theta}[f(\theta/2) + I_0(\theta) - I_1(\theta)]^2\right).$$

The function marking the urns that have a positive balance at two different times is

$$\begin{aligned}\Phi_2 \quad & (x_1, x_2; z_1, z_2) = \\ & (e^{2z_1+2z_2} + (x_1-1)f(z_1)e^{2z_2} + (x_2-1)f(z_1,z_2) + (x_1-1)(x_2-1)S(z_1,z_2))^m,\end{aligned}$$

with the function $S(z_1, z_2)$ defined in [3] :

$$S(z_1,z_2) := \sum_{p_1>0, p_1+p_2>0} I_{p_1}(2z_1)\, I_{p_2}(2z_2).$$

We can generalize this to the *number of urns with balance greater than some bound.* We have here $g(z) = e^z$, and the set $\mathcal{E}$ becomes $[p... + \infty[$. The function enumerating such states is

$$f(z) = \sum_{q\geq p} I_q(2z).$$

We have that

$$\Phi_1(x;z) = \left(e^{2z} + (x-1)f(z)\right)^m.$$

The average number of urns with positive balance is

$$EX_m(n) = m\frac{[z^n]\{f(z)g(2z)^{m-1}\}}{[z^n]\{g(2z)^m\}} \sim me^{-\theta} f(\theta/2) = me^{-\theta} \sum_{q\geq p} I_q(\theta).$$

We have here $f^{'}(z) = 2f(z) + I_{p-1}(2z) - I_p(2z)$; we get

$$VarX_m(n) \sim me^{-\theta} \left(f(\theta/2) - e^{-\theta} f^2(\theta/2) - \theta e^{-\theta}[f(\theta/2) + I_{p-1}(\theta) - I_p(\theta)]^2\right).$$

We turn now to the functions describing what happens in a single urn when we allocate the balls in d batches. We have

$$K_l(z_1, ...z_l) = \sum_{q_1,\dots,q_l} I_{q_1}(z_1)...I_{q_l}(z_l),$$

with the summation being on the q_i such that $q_1 < p$, ..., $q_1 + ... + q_{l-1} < p$ but $q_1 + ... + q_l \geq p$ (or $q_1 + ... + q_l < p$ for K_{d+1}). Of course there is the usual possibility of extension to urns of a different type and a basic enumerating function $g(z)$.

3. Proof of Theorem 2.1

In order to prove Theorem 2.1 we first have to show that there exists a process with a.s. continuous sample paths the f.d.d.'s of which are characterized by the limiting f.d.d.'s of $Y_m(t)$. Afterwards we have to prove that this process is in fact the limit. As we are working in the space $C[0,\infty)$ this can be done via [2, Theorem 12.3]: We only have to show the weak convergence of the f.d.d.'s and that the sequence $Y_m(t)$ is tight. It follows immediately by [1] that the limiting distributions of the f.d.d.'s are centered Gaussian distributions with covariance matrices given by (2.4) and these are exactly the f.d.d.'s of $G(t)$ by construction.

By [13, Chap. I, Proposition 3.7] the existence of a centered Gaussian process having the same covariance matrices (and thus f.d.d.'s) as $G(t)$ is guaranteed by the convergence of the covariance matrices of $Y_m(t)$ to a limit determined by (2.4) which defines a positive semi-definite function. Thus we only have to show continuity of the sample paths. This can be done by means of Kolmogorov's criterion (see [13, Chap. I, Theorem 1.8], or more generally by the Kolmogorov-Čentsov theorem, see [10, Theorem 2.8]):

Theorem 3.1. *A real-valued process X for which there exist three constants α, β, $C > 0$ such that*

$$E[|X(t+h) - X(t)|^{\alpha}] \leq Ch^{1+\beta},$$

for every t and h, has a modification with a.s. continuous sample paths. The same holds on the space $C[0,T]$ with $t, t+h \leq T$.

The fact that $G(t)$ satisfies this criterion follows immediately by

Lemma 3.1. *We have*

$$E(G(t) - G(t+s))^4 = \mathcal{O}\left(s^2\right),$$

uniformly for $t = \mathcal{O}(1)$

Proof. For Gaussian processes we have

$$E(G(t) - G(t+s))^4 = \frac{1}{8}(B_{t,t} - 2B_{t,t+s} + B_{t+s,t+s})^2$$

Hence, it is sufficient to show

$$B_{t,t\pm s} - B_{t,t} = \mathcal{O}(s) \tag{3.1}$$

uniformly for $t = \mathcal{O}(1)$. In what follows we will only discuss the difference $B_{t,t+s} - B_{t,t}$ with $s > 0$. The remaining case can be managed in the same way.

First, let us consider $B_{t,t}$. We use the representation $\Phi = \Phi_{\mathcal{E},1}(x,r)) = g(r) + (x-1)k(r)$. Furthermore, for simplicity use $g' = g'(r)$ for the derivative with respect to r and the index-notation f_x for the derivative with respect to x. By definition

$$\begin{aligned} B_{t,t} &= \frac{g' r_x + k}{g} + \frac{g'' r_x^2 + 2k' r_x + g' r_{xx}}{g} - \frac{(g' r_x + k)^2}{g^2} \\ &\quad - t\left(\frac{r_x}{r} + \frac{r_{xx}}{r} - \frac{r_x^2}{r^2}\right) \\ &= \frac{k}{g} + \frac{g'' r_x^2 + 2k' r_x}{g} - \frac{2kg' r_x + k^2}{g^2}. \end{aligned}$$

r_x and r_{xx} (evaluate at $x = 1$) can be derived by implicit differentiation:

$$\begin{aligned} r_x &= \frac{tk - rk'}{g' + rg'' - tg'}, \\ r_{xx} &= \frac{r_x^2(rg''' + tg'') + r_x(rk'' - k + tk' - trk'')}{g' + rg'' - tg'} \end{aligned}$$

The definition of $B_{t,t+s}$ $(s > 0)$ is much more involved. Here we use the representation

$$\Phi_{\mathcal{E},2}(x_1, x_2, r_1, r_2) = f_1(r_1, r_2) + x_1 f_2(r_1, r_2) + x_2 f_3(r_1, r_2) + x_1 x_2 f_4(r_1, r_2).$$

Since $\Phi_{\mathcal{E},2}(x_1, 1, r_1, r_2) = \Phi_{\mathcal{E},1}(x_1, r_1) g(r_2)$ we have

$$f_2(r_1, r_2) + f_4(r_1, r_2) = k(r_1) g(r_2).$$

Furthermore, by construction $f_2(r_1, 0) = f_3(r_1, 0) = 0$, i.e. they contain a factor r_2. This can be easily seen in the following way: Since x_i marks urns such that $Y(U) \in \mathcal{E}$ at time θ_i, the function f_2 (resp. f_3) enumerates the allocations into an urn such that the condition $Y(U) \in \mathcal{E}$ is satisfied at time θ_1 (resp. θ_2) and not satisfied at time θ_2 (resp. θ_1). Since this can only happen if the urn receives at least one ball between θ_1 and θ_2 (and those balls are counted by r_2), f_2 and f_3 must contain a factor r_2.

For simplicity we use the notation $\phi = \Phi_{\mathcal{E},2}(x_1, x_2, r_1(x_1,x_2), r_2(x_1,x_2))$ and the index notation for partial derivatives. By definition

$$\begin{aligned} B_{t,t+s} &= \frac{\frac{\partial^2}{\partial x_1 \partial x_2}\phi}{\phi} - \frac{\frac{\partial}{\partial x_1}\phi}{\phi}\frac{\frac{\partial}{\partial x_2}\phi}{\phi} \\ &\quad -t\left(\frac{(r_1)_{x_1x_2}}{r_1} - \frac{(r_1)_{x_1}}{r_1}\frac{(r_1)_{x_2}}{r_1}\right) \\ &\quad -s\left(\frac{(r_2)_{x_1x_2}}{r_2} - \frac{(r_2)_{x_1}}{r_2}\frac{(r_2)_{x_2}}{r_2}\right). \end{aligned}$$

Now r_1, r_2 are determined by $r_1\phi_{r_1} = t\phi$, $r_2\phi_{r_2} = s\phi$. Since $\Phi(1,1,r_1,r_2) = g(r_1)g(r_2)$ we have

$$r_1 g'(r_1) = t g(r_1) \quad \text{and} \quad r_2 g'(r_2) = s g(r_2)$$

which gives $r_1 = r$ (from above) and $r_2 = \mathcal{O}(s)$.

Differentiation of the first (implicit) equation with respect to x_1 yields

$$\begin{aligned} &(r_1)_{x_1} g'(r_1)g(r_2) + r_1 g''(r_1)g(r_2)(r_1)_{x_1} + r_1 g'(r_1)g'(r_2)(r_2)_{x_1} + r_1(f_2+f_4)_{r_1} \\ &\quad = t\left(g'(r_1)g(r_2)(r_1)_{x_1} + g(r_1)g'(r_2)(r_2)_{x_1} + (f_2+f_4)\right) \end{aligned}$$

or (by applying $r_1 g'(r_1) = t g(r_1)$ and $f_2 + f_4 = k(r_1)g(r_2)$)

$$(r_1)_{x_1}\left(g'(r_1) + r_1 g''(r_1) - t g'(r_1)\right) = t k(r_1) - r_1 k'(r_1).$$

Hence $(r_1)_{x_1} = r_x$ (from above). Similarly we obtain $(r_2)_{x_1} = 0$.

In the same way we obtain by differentiation of the first equation with respect to x_2

$$\begin{aligned} (r_1)_{x_2} &= \frac{t\frac{f_3+f_4}{g(r_2)} - r_1\frac{(f_3+f_4)_{r_1}}{g(r_2)}}{g'(r_1) + r_1 g''(r_1) - t g'(r_1)} \\ &= (r_1)_{x_1} + \frac{t\frac{f_3-f_2}{g(r_2)} - r_1\frac{(f_3-f_2)_{r_1}}{g(r_2)}}{g'(r_1) + r_1 g''(r_1) - t g'(r_1)} \\ &= (r_1)_{x_1} + \mathcal{O}(r_2) \\ &= (r_1)_{x_1} + \mathcal{O}(s). \end{aligned}$$

In the same way we get $(r_2)_{x_2} = \mathcal{O}(s)$, $(r_1)_{x_1x_2} = r_{xx} + \mathcal{O}(s)$, and $(r_2)_{x_1x_2} = \mathcal{O}(s)$.

Applying this to the derivatives of ϕ we have

$$\begin{aligned} \frac{\partial}{\partial x_1}\phi &= g'(r_1)g(r_2)(r_1)_{x_1} + g(r_1)g'(r_2)(r_2)_{x_1} + (f_2+f_4) \\ &= g'(r)g(r_2)r_x + k(r)g(r_2) + \mathcal{O}(s), \\ \frac{\partial}{\partial x_2}\phi &= g'(r_1)g(r_2)(r_1)_{x_2} + g(r_1)g'(r_2)(r_2)_{x_2} + (f_3+f_4) \\ &= k(r)g(r_2) + \mathcal{O}(s), \\ \frac{\partial^2}{\partial x_1 \partial x_2}\phi &= g''(r_1)g(r_2)(r_1)_{x_1}(r_1)_{x_2} + g'(r_1)g'(r_2)\left((r_1)_{x_1}(r_2)_{x_2} + (r_1)_{x_2}(r_2)_{x_1}\right) \\ &\quad + g(r_1)g''(r_2)(r_2)_{x_1}(r_2)_{x_2} \\ &\quad + (f_3+f_4)_{r_1}(r_1)_{x_1} + (f_3+f_4)_{r_2}(r_2)_{x_1} + (f_2+f_4)_{r_1}(r_1)_{x_2} + (f_2+f_4)_{r_2}(r_2)_{x_2} \\ &\quad + f_4 \\ &= g''(r_1)g(r_2)r_x^2 + 2k'(r_1)g(r_2)r_x + k(r_1)g(r_2) + \mathcal{O}(s) \end{aligned}$$

Hence, it follows that $B_{t,t+s} = B_{t,t} + \mathcal{O}(s)$. □

In order to prove tightness let us first show the following

Lemma 3.2. *There exist constants $C_1, C_2 > 0$ such that*

$$\begin{aligned} E(X_m(n) - EX_m(n))^2 &\le C_1 n \\ E(X_m(n) - EX_m(n))^4 &\le C_2 n^2 \end{aligned} \tag{3.2}$$

for $m \to \infty$ *and* $n = \mathcal{O}(m)$.

Proof. Let us first deal with the second inequality. Set

$$c_{n,i} := [z^n]\frac{\partial^i}{\partial x^i}\Phi_{\mathcal{E},1}(z,1) \quad \text{and} \quad A_i := E\prod_{j=0}^{i-1}(X_m(n)-j) = \frac{c_{n,i}}{c_{n,0}}.$$

The fourth moment occurring in (3.2) can now be expressed by

$$E(X_m(n)-EX_m(n))^4 = A_4 - 4A_1A_3 + 6A_1^2A_2 - 3A_1^4 + 6A_3 - 12A_1A_2 + 6A_1^3 + 7A_2 - 4A_1^2 + A_1 \tag{3.3}$$

Hence we have to compute $c_{n,i}$. Let us start with $c_{n,0} = [z^n]g(z)^m$. Note that due to the fact that we allow an urn to be empty or to contain one ball we have $g(0) = g_0 \neq 0$ and $g'(0) = g_1 \neq 0$. Define $\kappa_j(z)$ by

$$\kappa_1(z) := \frac{zg'(z)}{g(z)} = \frac{g_1}{g_0}z + \left(\frac{2g_2}{g_0} - \frac{g_1^2}{g_0^2}\right)z^2 + \mathcal{O}\left(z^3\right)$$

and

$$\kappa_{j+1}(z) := z\kappa_j'(z), \quad j \geq 1.$$

Observe that $\kappa_j(z) = \mathcal{O}(z)$ for $z \to 0$ and by Taylor's theorem we have for real z and for any fixed $k \geq 1$ and $z_0 > 0$

$$g\left(ze^{i\theta}\right) = g(z)\exp\left(\sum_{j=1}^{k}\frac{(i\theta)^j}{j!}\kappa_j(z) + \mathcal{O}\left(|\theta^{k+1}|z\right)\right) \tag{3.4}$$

uniformly for $0 < z \leq z_0$ and $\theta \in [-\theta_0, \theta_0]$ where $\theta_0 > 0$ sufficiently small. Furthermore, in presence of the fact that there exist no r, d such that $g_n \neq 0$ if and only if $g_n \equiv r \mod d$ we have

$$|g(ze^{i\theta})| \leq g(z)e^{-c\theta^2} \tag{3.5}$$

for some positive constant c. In order to extract the desired coefficient we will use Cauchy's integral formula and the saddle point method. Let μ denote the inverse function of κ_1. Then we have

$$\mu(t) = \frac{g_0}{g_1}t + \left(\frac{g_0}{g_1} - \frac{2g_0^2g_2}{g_1^3}\right)t^2 + \mathcal{O}\left(t^3\right).$$

The saddle point of $g(z)^m z^{-n}$ is given by

$$\rho = \mu\left(\frac{n}{m}\right) = \frac{g_0}{g_1}\frac{n}{m}\left(1 + \mathcal{O}\left(\frac{n}{m}\right)\right).$$

By applying the saddle point method and using (3.4) and (3.5) we obtain

$$\begin{aligned}
[z^n]g(z)^m =& \frac{1}{2\pi i}\oint_{|z|=\rho} g(z)^m \frac{dz}{z^{n+1}} \\
=& \frac{1}{2\pi\rho^n}\left(\int_{|\theta|\leq(m\rho)^{-1/2+\varepsilon}} + \int_{(m\rho)^{-1/2+\varepsilon}\leq|\theta|\leq\theta_0} + \int_{\theta_0\leq|\theta|\leq\pi}\right) g(\rho e^{i\theta})e^{-in\theta}\,d\theta \\
=& \frac{g(\rho)^m}{2\pi\rho^n}\int_{|\theta|\leq(m\rho)^{-1/2+\varepsilon}} \exp\left(-\frac{\theta^2}{2}m\kappa_2(\rho) - i\frac{\theta^3}{3!}m\kappa_3(\rho) + \mathcal{O}\left(\theta^4 m\rho\right)\right)\,d\theta \\
&+ \mathcal{O}\left(g(\rho)^m\frac{e^{-(m\rho)^{2\varepsilon}c_1}}{m}\right) + \mathcal{O}\left(g(\rho)^m\frac{e^{-m\rho\theta_0^2 c}}{m}\right)
\end{aligned} \tag{3.6}$$

where $c_1 > 0$ is a suitable constant. Note that

$$m\kappa_j(\rho) = m\mu\left(\frac{n}{m}\right)\kappa_{j-1}'\left(\mu\left(\frac{n}{m}\right)\right) = n\bar{\kappa}_j\left(\frac{n}{m}\right)$$

where $\bar\kappa_j(t)$ are analytic functions with $\bar\kappa_j(0)=1$. Hence

$$\begin{aligned}
[z^n]g(z)^m &= \frac{g(\rho)^m}{2\pi\rho^n}\int\limits_{|\theta|\le n^{-1/2+\varepsilon}} \exp\left(-\frac{\theta^2}{2}n\bar\kappa_2\left(\frac{n}{m}\right)\right)\left(1-i\frac{\theta^3}{3!}n\bar\kappa_3\left(\frac{n}{m}\right)+\mathcal{O}\left(\theta^4 n\right)\right)\,d\theta\\
&= \frac{g(\rho)^m}{\rho^n}\left(\frac{1}{\sqrt{2\pi n\bar\kappa_2(n/m)}}+\mathcal{O}\left(n^{-3/2}\right)\right).
\end{aligned}$$

Using more terms we directly obtain an asymptotic series expansion of the form

$$[z^n]g(z)^m \sim \frac{g(\mu(n/m))^m}{\sqrt{2\pi n}\mu(n/m)^n}\left(\sum_{j\ge0}a_j\left(\frac{n}{m}\right)\frac{1}{n^j}\right)$$

where $a_j(t)$ are analytic functions that can be determined explicitly, especially $a_0(t)=\bar\kappa_2(t)^{-1/2}$.

Now let us investigate how the situation changes for $c_{n,\alpha}$ with $\alpha>0$. For technical convenience let us assume that $f(z)$ contains a factor z. So the g.f. under consideration has the form

$$\Phi_{\mathcal{E},1}(x,z)=(g(z)+(x-1)zk(z))^m.$$

Note that this is no restriction for the present purpose: If we have an urn model where $f(z)$ does not meet this constraint, then let us use the process $\tilde X_m = m-X_m$ instead. Since this does not change the fourth moment (3.3), the assumption is justified. We have

$$c_{n,\alpha}=[z^n]m(m-1)\cdots(m-\alpha+1)z^\alpha K(z)^\alpha g(z)^m$$

where $K(z)=k(z)/g(z)$. As above we get

$$K\left(ze^{i\theta}\right)=K(z)\exp\left(\sum_{j=1}^k\frac{(i\theta)^j}{j!}\lambda_j(z)+\mathcal{O}\left(z\theta^{k+1}\right)\right)$$

uniformly for $0<z<z_0$ and $|\theta|\le\theta_0$, where

$$\lambda_0=z\frac{K'(z)}{K(z)},\quad \lambda_{j+1}(z)=z\lambda_j'(z).$$

Furthermore, note that $|k(ze^{i\theta})|\le k(z)$ because of the positivity of the coefficients of $k(z)$. This in conjunction with (3.4) guarantees that the estimates for the remainder integral in (3.6) still hold in this case. Therefore, applying again the saddle point method gives

$$\begin{aligned}
&[z^n]m(m-1)\cdots(m-\alpha+1)z^\alpha K(z)^\alpha g(z)^m = \frac{m(m-1)\cdots(m-\alpha+1)\rho^\alpha g(\rho)^m K(\rho)^\alpha}{2\pi\rho^n}\\
&\quad\times\int\limits_{|\theta|\le(m\rho)^{-1/2+\varepsilon}}\exp\left(-\frac{\theta^2}{2}m\kappa_2(\rho)+\sum_{j=3}^k\frac{(i\theta)^j}{j!}m\kappa_j(\rho)+\alpha\sum_{j=1}^k\frac{(i\theta)^j}{j!}\lambda_j(\rho)\right.\\
&\quad\left.+\mathcal{O}\left(m\rho\theta^{k+1}\right)\right)d\theta\\
&= \frac{m(m-1)\cdots(m-\alpha+1)\rho^\alpha g(\rho)^m K(\rho)^\alpha}{2\pi\rho^n}\int\limits_{|\theta|\le(m\rho)^{-1/2+\varepsilon}}\exp\left(-\frac{\theta^2}{2}n\bar\kappa_2\left(\frac{n}{m}\right)\right.\\
&\quad\left.+\sum_{j=3}^k\frac{(i\theta)^j}{j!}n\bar\kappa_j\left(\frac{n}{m}\right)+\alpha\sum_{j=1}^k\frac{(i\theta)^j}{j!}\frac{n}{m}\bar\lambda_j\left(\frac{n}{m}\right)+\mathcal{O}\left(n\theta^{k+1}\right)\right)d\theta
\end{aligned}$$

Using the substitution $\theta = u/\sqrt{n\bar{\kappa}_2(n/m)}$ yields

$$
\begin{aligned}
&[z^n]m(m-1)\cdots(m-\alpha+1)\rho^\alpha K(z)^\alpha g(z)^m \\
&\quad= \frac{m(m-1)\cdots(m-\alpha+1)g(\rho)^{m-\alpha}k(\rho)^\alpha}{2\pi\rho^n\sqrt{n\bar{\kappa}_2(n/m)}} \\
&\qquad\times \int\limits_{|u|\le(m\rho)^\varepsilon\sqrt{(n/m)\cdot\bar{\kappa}_2(n/m)/\rho}} \exp\left(-\frac{u^2}{2}+\sum_{j=3}^{k}\frac{(iu)^j}{j!}n^{1-j/2}\bar{\kappa}_j(n/m)\bar{\kappa}_2(n/m)^{-j/2}\right. \\
&\qquad\left. +\alpha\sum_{j=1}^{k}\frac{(iu)^j}{j!}\frac{n^{1-j/2}}{m}\bar{\lambda}_j(n/m)\bar{\kappa}_2(n/m)^{-j/2}+\mathcal{O}\left(n\left(\frac{u}{\sqrt{n}}\right)^{k+1}\right)\right)\,d\theta.
\end{aligned}
$$

Set

$$
\begin{aligned}
\tilde{\kappa}_j(x) &= \bar{\kappa}_j(x)\bar{\kappa}_2(x)^{-j/2},\\
\tilde{\lambda}_j(x) &= \bar{\lambda}_j(x)\bar{\kappa}_2(x)^{-j/2}.
\end{aligned}
$$

We have

$$
\begin{aligned}
&\exp\left(\sum_{j=3}^{k}\frac{(iu)^j}{j!}n^{1-j/2}\tilde{\kappa}_j(n/m)+\alpha\sum_{j=1}^{k}\frac{(iu)^j}{j!}\frac{n^{1-j/2}}{m}\tilde{\lambda}_j(n/m)\right) \\
&\quad= 1+i\alpha\frac{n}{m}\frac{\tilde{\lambda}_1(n/m)}{\sqrt{n}}u-\alpha\frac{\tilde{\lambda}_2(n/m)\cdot(n/m)+\alpha\tilde{\lambda}_1(n/m)^2\cdot(n/m)^2}{2n}u^2-\frac{i}{6}\left(\frac{\tilde{\kappa}_3(n/m)}{\sqrt{n}}\right. \\
&\qquad\left.+\frac{\alpha(n/m)\tilde{\lambda}_3(n/m)+3\alpha^2(n/m)^2\tilde{\lambda}_1(n/m)\tilde{\lambda}_2(n/m)+\alpha^3(n/m)^3\tilde{\lambda}_1(n/m)^3}{\sqrt{n^3}}\right)u^3+\cdots
\end{aligned}
$$

The odd powers of u do not contribute to the integral. Hence, using

$$
\int_{-\infty}^{\infty} v^{2k}e^{-v^2/2}\,dv = \frac{(2k)!}{2^k k!}\sqrt{2\pi}
$$

and setting

$$
V(\alpha) = \frac{g(\rho)^m K(\rho)^\alpha}{\sqrt{2\pi}\rho^n\sqrt{n\bar{\kappa}_2(n/m)}}
$$

gives

$$
c_{n,0} = V(0)\left(1+\frac{1}{n}\left(\frac{\tilde{\kappa}_4(n/m)}{8}-\frac{5\tilde{\kappa}_3(n/m)^2}{24}\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\right)
$$

and

$$
\begin{aligned}
c_{n,\alpha} =& V(\alpha)m\cdots(m-\alpha+1)\rho^\alpha\sqrt{2\pi}\left(1-\frac{1}{n}\left(\frac{\tilde{\kappa}_4(n/m)}{8}-\frac{5\tilde{\kappa}_3(n/m)^2}{24}+\frac{\alpha}{2}\frac{n}{m}(\tilde{\kappa}_3(n/m)\tilde{\lambda}_1(n/m)\right.\right. \\
&\left.\left.-\ \lambda_2(n/m))-\frac{\alpha^2}{2}\left(\frac{n}{m}\right)^2\lambda_1(n/m)^2\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\right)
\end{aligned}
$$

and thus

$$
A_\alpha = K(\rho)^\alpha m\cdots(m-\alpha+1)\rho^\alpha\left(1+\frac{1}{n}\left(\frac{n}{m}\right)\left(\frac{\alpha}{2}(\tilde{\kappa}_3\tilde{\lambda}_1-\tilde{\lambda}_2))+\frac{\alpha^2}{2}\frac{n}{m}\tilde{\lambda}_1^2\right)+\mathcal{O}\left(\frac{1}{n^2}\right)\right)
$$

Now inserting this into (3.3) and keeping in mind that $\rho=\mu(n/m)=\mathcal{O}(n/m)$ and $n=\mathcal{O}(m)$ shows that $E(X_m(n)-EX_m(n))^4=\mathcal{O}(n^2)$ as desired. The first inequality is now an easy exercise. □

Proposition 3.1. *The sequence $Y_m(t)$ is tight.*

Proof. Due to [2, Theorem 12.3] it suffices to show

$$E\frac{(X_m(n_1+n_2)-X_m(n_1)-E(X_m(n_1+n_2)-X_m(n_1)))^4}{m^2} \le C\left(\frac{n_2}{m}\right)^2 \tag{3.7}$$

where C is a positive constant. In order to treat the difference $Z_m(n_1,n_2)=X_m(n_1+n_2)-X_m(n_1)$ we use the generating function that enumerates the urns whose state has changed. In the general model this function has the shape

$$\Phi(x,z_1,z_2)=\left(g(z_1)g(z_2)+(x-1)f_2(z_1,z_2)+\left(\frac{1}{x}-1\right)f_3(z_1,z_2)\right)^m.$$

Note that f_2 and f_3 contain a factor z_2 (which proves to be important in the sequel), since a state change occurs if and only if the urn receives at least one ball during the time period under consideration. For simplicity, let us assume that $f_3\equiv 0$. Then the generating function can be expressed in the form

$$\Phi(x,z_1,z_2)=(g(z_1)g(z_2)+(x-1)z_2k(z_1,z_2))^m$$

with an analytic function $k(z_1,z_2)$.

Set

$$c_{n_1n_2,i}:=[z_1^{n_1}z_2^{n_2}]\frac{\partial^i}{\partial x^i}\Phi(z_1,z_2,1) \quad\text{and}\quad A_i:=E\prod_{j=0}^{i-1}(Z_m(n_1,n_2)-j)=\frac{c_{n_1n_2,i}}{c_{n_1n_2,0}}.$$

In analogy to (3.3) the fourth moment occurring in (3.7) can be expressed by

$$EZ_m(n_1,n_2)^4=A_4-4A_1A_3+6A_1^2A_2-3A_1^4+6A_3-12A_1A_2+6A_1^3+7A_2-4A_1^2+A_1 \tag{3.8}$$

Hence we have to compute $c_{n_1n_2,i}$. Let us start with $c_{n_1n_2,0}$. This is rather easy since it factorizes nicely:

$$\begin{aligned} c_{n_1n_2,0} =&[z_1^{n_1}z_2^{n_2}]g(z_1)^m g(z_2)^m=[z_1^{n_1}]g(z_1)^m[z_2^{n_2}]g(z_2)^m\\ =&\sqrt{2\pi}\left(1+\frac{1}{n_1}\left(\frac{5\tilde\kappa_3^2(n_1/m)}{24}-\frac{\tilde\kappa_4(n_1/m)}{8}\right)+\frac{1}{n_2}\left(\frac{5\tilde\kappa_3^2(n_2/m)}{24}-\frac{\tilde\kappa_4(n_2/m)}{8}\right)\right.\\ &\left.+\mathcal{O}\left(\frac{1}{n_1^2}\right)+\mathcal{O}\left(\frac{1}{n_2^2}\right)\right) \end{aligned} \tag{3.9}$$

Now we investigate what happens if $\alpha>0$ where we do not have such a factorization as for $c_{n_1n_2,0}$. We have

$$c_{n_1n_2,\alpha}=[z_1^{n_1}z_2^{n_2}]m(m-1)\cdots(m-\alpha+1)z_2^\alpha K(z_1,z_2)^\alpha g(z_1)^m g(z_2)^m$$

where $K(z_1,z_2)=\frac{k(z_1,z_2)}{g(z_1)g(z_2)}$. With (3.4) and

$$K\left(z_1e^{i\theta_1},z_2e^{i\theta_2}\right)=K(z_1,z_2)\exp\left(\sum_{j_1+j_2>0}^{k}\frac{(i\theta_1)^{j_1}(i\theta_2)^{j_2}}{j_1!j_2!}\lambda_{j_1j_2}(z_1,z_2)+\mathcal{O}\left(z_1|\theta_1^{k+1}|\right)+\mathcal{O}\left(z_2|\theta_2^{k+1}|\right)\right)$$

where

$$\begin{aligned} \lambda_{10}&=z_1\frac{\frac{\partial}{\partial z_1}K(z_1,z_2)}{K(z_1,z_2)} & \lambda_{01}&=z_2\frac{\frac{\partial}{\partial z_2}K(z_1,z_2)}{K(z_1,z_2)}\\ \lambda_{j_1+1,j_2}&=z_1\frac{\partial}{\partial z_1}\lambda_{j_1j_2} & \lambda_{j_1,j_2+1}&=z_2\frac{\partial}{\partial z_1}\lambda_{j_1j_2} \end{aligned}$$

we have

$$[z_1^{n_1}z_2^{n_2}]m(m-1)\cdots(m-\alpha+1)\left(z_2K(z_1,z_2)\right)^\alpha g(z_1)^m g(z_2)^m = \frac{g(\rho_1)^m g(\rho_2)^m K(\rho_1,\rho_2)^\alpha}{2\pi\rho_1^{n_1}\rho_2^{n_2}}$$

$$\times\, m(m-1)\cdots(m-\alpha+1)\rho_2^\alpha \iint\limits_B \exp\left(-\frac{\theta_1^2}{2}m\kappa_2(\rho_1)-\frac{\theta_2^2}{2}m\kappa_2(\rho_2)+\sum_{j=3}^k\frac{(i\theta_1)^j}{j!}m\kappa_j(\rho_1)\right.$$

$$+\sum_{j=3}^k\frac{(i\theta_2)^j}{j!}m\kappa_j(\rho_2)+\alpha i\theta_2+\alpha\sum_{j_1+j_2\geq 1}^k\frac{(i\theta_1)^{j_1}(i\theta_2)^{j_2}}{j_1!j_2!}\lambda_{j_1j_2}(\rho_1,\rho_2)$$

$$\left.+\,\mathcal{O}\left(m\rho_1|\theta_1^{k+1}|\right)+\mathcal{O}\left(m\rho_2|\theta_2^{k+1}|\right)\right)d\theta_1\,d\theta_2$$

$$=\frac{g(\rho_1)^m g(\rho_2)^m K(\rho_1,\rho_2)^\alpha}{2\pi\rho_1^{n_1}\rho_2^{n_2}}m(m-1)\cdots(m-\alpha+1)\rho_2^\alpha$$

$$\times\iint\limits_B\exp\left(-\frac{\theta_1^2}{2}n_1\bar\kappa_2\left(\frac{n_1}{m}\right)-\frac{\theta_2^2}{2}n_2\bar\kappa_2\left(\frac{n_2}{m}\right)+\sum_{j=3}^k\frac{(i\theta_1)^j}{j!}n_1\bar\kappa_j\left(\frac{n_1}{m}\right)+\sum_{j=3}^k\frac{(i\theta_2)^j}{j!}n_2\bar\kappa_j\left(\frac{n_2}{m}\right)\right.$$

$$\left.+\alpha i\theta_2+\alpha\sum_{j_1+j_2\geq 1}^k\frac{(i\theta_1)^{j_1}(i\theta_2)^{j_2}}{j_1!j_2!}\frac{n_1n_2}{m^2}\bar\lambda_{j_1j_2}\left(\frac{n_1}{m},\frac{n_2}{m}\right)+\mathcal{O}\left(m\rho_1|\theta_1^{k+1}|\right)+\mathcal{O}\left(m\rho_2|\theta_2^{k+1}\right)|\right)d\theta_1\,d\theta_2$$

where as above $\bar\kappa_j(z)=\kappa_j(\mu(z))/z$, $\rho_1=\mu(n_1/m)$, $\rho_2=\mu(n_2/m)$,

$$\bar\lambda_{j_1j_2}(z_1,z_2)=\frac{\lambda_{j_1j_2}(\mu(z_1),\mu(z_2))}{z_1z_2},$$

and the integration domain B is given by

$$B=\{(\theta_1,\theta_2)|\,|\theta_1|\leq(m\rho_1)^{-1/2+\varepsilon}\text{ and }|\theta_2|\leq(m\rho_2)^{-1/2+\varepsilon}\}$$

Substituting $\theta_1=u_1/\sqrt{n_1\bar\kappa_2(n_1/m)}$, $\theta_2=u_2/\sqrt{n_2\bar\kappa_2(n_2/m)}$ yields

$$[z_1^{n_1}z_2^{n_2}]m(m-1)\cdots(m-\alpha+1)\left(z_2K(z_1,z_2)\right)^\alpha g(z_1)^m g(z_2)^m$$

$$=\frac{g(\rho_1)^m g(\rho_2)^m K(\rho_1,\rho_2)^\alpha}{2\pi\rho_1^{n_1}\rho_2^{n_2}}\frac{m(m-1)\cdots(m-\alpha+1)\rho_2^\alpha}{\sqrt{n_1n_2\bar\kappa_2(n_1/m)\bar\kappa_2(n_2/m)}}$$

$$\times\iint\limits_{\tilde B}\exp\left(-\frac{u_1^2}{2}-\frac{u_2^2}{2}+\sum_{j=3}^k\frac{(iu_1)^j}{j!}n_1^{1-j/2}\tilde\kappa_j\left(\frac{n_1}{m}\right)+\sum_{j=3}^k\frac{(iu_2)^j}{j!}n_2^{1-j/2}\tilde\kappa_j\left(\frac{n_2}{m}\right)\right.$$

$$+\,\alpha\frac{iu_2}{\sqrt{n_2}}\left(\frac{1}{\bar\kappa_2(n_2/m)}+\frac{n_1n_2}{m^2}\tau_{01}\left(\frac{n_1}{m},\frac{n_2}{m}\right)\right)+\alpha\frac{iu_1}{\sqrt{n_1}}\tau_{10}\left(\frac{n_1}{m},\frac{n_2}{m}\right)$$

$$+\,\alpha\sum_{j_1,j_2=1}^k\frac{(iu_1)^{j_1}(iu_2)^{j_2}}{j_1!j_2!}n_1^{1-j_1/2}n_2^{1-j_2/2}\tau_{j_1j_2}\left(\frac{n_1}{m},\frac{n_2}{m}\right)$$

$$\left.+\,\mathcal{O}\left(m\rho_1\left|\frac{u_1}{\sqrt{n_1}}\right|^{k+1}\right)+\mathcal{O}\left(m\rho_2\left|\frac{u_2}{\sqrt{n_2}}\right|^{k+1}\right)\right)d\theta_1\,d\theta_2$$

where

$$\tilde\kappa_j(x)=\bar\kappa_j(x)\bar\kappa_2(x)^{-j/2},$$
$$\tau_{j_1j_2}(x,y)=\bar\lambda_{j_1j_2}(x,y)\bar\kappa_2(x)^{-j_1/2}\bar\kappa_2(y)^{-j_2/2}.$$

Expanding the exp-term into a series, evaluating the integral, and setting

$$V(\alpha)=\frac{g(\rho_1)^m g(\rho_2)^m K(\rho_1,\rho_2)^\alpha m(m-1)\cdots(m-\alpha+1)\rho_2^\alpha}{\sqrt{2\pi}\rho_1^{n_1}\rho_2^{n_2}\sqrt{mn_2\kappa_2(\rho_1)\bar\kappa_2(n_2/m)}}$$

gives

$$\begin{aligned}
c_{n_1n_2,\alpha} =&V(\alpha)\left(1-\frac{1}{n_1}\left(\frac{\tilde\kappa_4(n_1/m)}{8}-\frac{5\tilde\kappa_3(n_1/m)}{24}+\frac{\alpha\tilde\kappa_3(n_1/m)\tau_{1,0}(n_1/m,n_2/m)}{2}\right.\right.\\
&\left.-\frac{\alpha^2\tau_{1,0}(n_1/m,n_2/m)^2}{2}\right)+\frac{1}{n_2}\left(\frac{\tilde\kappa_4(n_2/m)}{8}-\frac{5\tilde\kappa_3(n_2/m)}{24}+\frac{\alpha\tilde\kappa_3(n_1/m)}{2\sqrt{\tilde\kappa_2(n_2/m)}}-\frac{\alpha^2}{2\tilde\kappa_2(n_2/m)}\right)\\
&\left.+\mathcal{O}\left(\frac{1}{n_1^2}\right)+\mathcal{O}\left(\frac{1}{n_2^2}\right)\right).
\end{aligned}$$

Note that $\rho_2=\mu(n_2/m)=n_2/m(1+\mathcal{O}(n_2/m))$. Hence let $L:=K(\rho_1,\rho_2)m\mu(n_2/m)/n_2$ and we get

$$\begin{aligned}
A_\alpha =&(Ln_2)^\alpha\left(1+\frac{\alpha}{2n_1}(\kappa_3(n_1/m)\tau_{1,0}(n_1/m,n_2/m)-\alpha\tau_{1,0}(n_1/m,n_2/m)^2)\right.\\
&\left.\frac{\alpha}{2n_2}\left(\frac{\tilde\kappa_3(n_2/m)}{\sqrt{\tilde\kappa_2(n_2/m)}}-\frac{\alpha}{\tilde\kappa_2(n_2/m)}\right)+\mathcal{O}\left(\frac{1}{n_1^2}\right)+\mathcal{O}\left(\frac{1}{n_2^2}\right)\right).
\end{aligned}$$

Inserting this into (3.8) shows that the terms containing n_2^3 or n_2^4/n_1 cancel and thus by assuming $n_2=\mathcal{O}(n_1)$ we get (3.7). In the case where $n_2=\mathcal{O}(n_1)$ does not hold let us assume $n_2\ge n_1$. Then set $X_m^c(n):=X_m(n)-EX_m(n)$ and use the crude estimate

$$EZ_m(n_1,n_2)^4\le EX_m^c(n_1+n_2)^4+6EX_m^c(n_1+n_2)^2EX_m^c(n_1)^2+EX_m^c(n_1)^4$$

in conjunction with Lemma 3.2.

In the general case (i.e. where $f_3\not\equiv 0$) the formulae are much more involved. In fact we have

$$\begin{aligned}
c_{n_1n_2,1} =&mg(z_1)^{m-1}g(z_2)^{m-1}(f_2(z_1,z_2)-f_3(z_1,z_2))\\
c_{n_1n_2,2} =&m(m-1)g(z_1)^{m-2}g(z_2)^{m-2}(f_2(z_1,z_2)-f_3(z_1,z_2))^2+2mg(z_1)^{m-1}g(z_2)^{m-1}f_3(z_1,z_2)\\
c_{n_1n_2,3} =&m(m-1)(m-2)g(z_1)^{m-3}g(z_2)^{m-3}(f_2(z_1,z_2)-f_3(z_1,z_2))^3\\
&+6m(m-1)g(z_1)^{m-2}g(z_2)^{m-2}f_3(z_1,z_2)(f_2(z_1,z_2)-f_3(z_1,z_2))\\
&-6mg(z_1)^{m-1}g(z_2)^{m-1}f_3(z_1,z_2)\\
c_{n_1n_2,4} =&m(m-1)(m-2)(m-3)g(z_1)^{m-4}g(z_2)^{m-4}(f_2(z_1,z_2)-f_3(z_1,z_2))^4\\
&+12m(m-1)(m-2)g(z_1)^{m-3}g(z_2)^{m-3}f_3(z_1,z_2)(f_2(z_1,z_2)-f_3(z_1,z_2))^2\\
&+12m(m-1)g(z_1)^{m-2}g(z_2)^{m-2}(f_3(z_1,z_2)^2-2f_3(z_1,z_2)(f_2(z_1,z_2)-f_3(z_1,z_2)))\\
&+24mg(z_1)^{m-1}g(z_2)^{m-1}f_3(z_1,z_2)
\end{aligned}$$

These formulae can be treated in the same way as above and yield the desired result. □

4. Future perspectives

We have shown in this paper that a class of additive valuations on occupancy urn models leads to limiting Gaussian processes. One of us has worked on some database parameters (join sizes) that can be modelized by urn models [6, 7]. It requires us to use two types of balls, and to compute a valuation on each urn according to the number of balls of each type that fall into the urn. The global parameter is obtained by summing the valuations on each urn; it may be additive, or not (the semijoin size is additive, but the equijoin size is not); even when it is additive, the results of the present paper do not apply : We have assumed in the present work that *the total number of balls* is known, and have studied the number of urns satisfying some condition ($Y(U)\in\mathcal{E}$), whereas the natural assumption for the modelization of join sizes is that the number of balls *of each type* is known, and we are interested in the valuation $\sum_{U\ urn}Y(U)$. However, it should be possible to extend our approach to deal with such situations, and possibly to take into account some types of deletions as well; we hope to present both in a future paper.

References

[1] E. A. Bender and L. B. Richmond, Central and local limit theorems applied to asymptotic enumeration II: multivariate generating functions, *J. Combinatorial Theory, Ser. A* 34, 255–265, 1983.

[2] P. Billingsley, *Convergence of Probability Measures*, John Wiley & Sons, New York, 1968.

[3] S. Boucheron and D. Gardy, An urn model from learning theory, *Random Struct. Alg.* 10, 43–67, 1997.

[4] M. Drmota, A bivariate asymptotic expansion of coefficients of powers of generating functions, *European Journal of Combinatorics* 15, 1994, 139-152.

[5] P. Flajolet and J. S. Vitter, Average-Case Analysis of Algorithms and Data Structures, in *Handbook of Theoretical Computer Science,* J. van Leeuwen, Ed., vol. A: Algorithms and Complexity. North Holland, 1990, ch. 9, pp. 431–524.

[6] D. Gardy, Normal limiting distributions for projection and semijoin sizes, *SIAM Journal on Discrete Mathematics,* 5(2), 219-248, 1992.

[7] D. Gardy, Join sizes, urn models and normal limiting distributions, *Theoretical Computer Science (A),* 131, 375-414, 1994.

[8] D. Gardy and G. Louchard, Dynamic analysis of some relational data bases parameters, *Theoretical Computer Science (A), (special issue on mathematical analysis of algorithms),* 144(1-2), 125-159, 1995.

[9] I. P. Goulden and D. M. Jackson, *Combinatorial Enumeration*, Wiley, New York, 1983.

[10] I. Karatzas and S. E. Shreve, *Brownian Motion and Stochastic Calculus,* Springer, New York, 1988.

[11] V. F. Kolchin, B. Sevast'yanov, and V. Chistyakov, *Random Allocations,* Wiley, New York, 1978.

[12] J. Neveu, *Processus aléatoires gaussiens,* Presses de l'Université de Montréal, 1968.

[13] D. Revuz and M. Yor, *Continuous Martingales and Brownian Motion,* Springer, 1991.