CLASSIFICATION OF URN MODELS WITH MULTIPLE DRAWINGS

MARKUS KUBA

Abstract. We consider multicolor urn models with multiple drawings. An urn model is called linear if the conditional expected value of the urn composition at time $n$ is a linear function of the composition at time $n-1$. For four different sampling schemes - ordered and unordered samples with or without replacement - we classify urns into linear and non-linear models. We also discuss representations of the expected value and the covariance for linear models.

1. Introduction

Urn models are simple, useful mathematical tools for describing many evolutionary processes in diverse fields of application such as analysis of algorithms and data structures, statistics and genetics.

The dynamics of the standard Pólya-Eggenberger urn models in the case of two types of colors and sample size $m=1$ can be described as follows. At the beginning, the urn contains $W_0$ white and $B_0$ black balls. At every step, we choose a ball at random from the urn, examine its color and put it back into the urn and then add/remove balls according to its color by the following rules: if the ball is white, then we put $a$ white and $b$ black balls into the urn, while if the ball is black, then $\gamma$ white balls and $\delta$ black balls are put into the urn. The values $a, b, c, d \in \mathbb{Z}$ are fixed integer values and the urn model is specified by the ball transition matrix $M = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$. Models with $r \geq 2$ types of colors can be described in an analogous way and are specified by an $r \times r$ ball transition matrix. One usually considers so-called tenable urn models where the process of sampling and replacing balls never stops. Quantities of interest are the number of white balls $W_n$ after $n$ draws, and the number of black balls $B_n$ after $n$ draws in the case of $r = 2$ colors; in the general case $r \geq 2$ one is interested in the distribution of the random vector $X_n = (X_n^{(1)}, \ldots, X_n^{(r)})$, where $X_n^{(i)}$ denotes the number of balls colored $i$ after $n$ draws and the initial composition of the urn model at time $n=0$ is given by the (non-random) vector $X_0 = (X_0^{(1)}, \ldots, X_0^{(r)})$. In the classic version of Pólya urns with $r \geq 2$ different colors a single ball is sampled at each unit of discrete time.

Due to their importance in applications, there is a huge literature on the stochastic behavior of urn models. The earliest contributions are the classical work of Ehrenfest and Ehrenfest [8] and the article of Polya and Eggenberger [7]. We also refer to the classic surveys of Johnson and Kotz [13, 19], the book of Mahmoud [23] and the references therein. The recent works of Chauvin et al. [2, 3], Janson [10, 11, 12], Neininger and Knape [17], Pouyanne [29], Mailler [25], Müller and Neininger [27], Müller [28], are all devoted to urn models where only a single ball is sampled at each step.

In this work we are concerned with generalizations of so-called Pólya-Eggenberger urn models. We study an $r$-color Pólya urn model, where multiple balls are drawn at each discrete time step. We assume that the $r \geq 2$ different colors are in a fixed order and we thus speak of balls coloured $i$, with $1 \leq i \leq r$. Their colors are inspected, then the sample is reinserted in the urn.

Additions and deletions take place according to the drawn sample of fixed size $m$ and we refer to the positive integer $m$ as the sample size; the sample is either a multiset of size $m$ or a
sequence of length \( m \). Such urn models recently received attention in the literature. Mahmoud and Tsukiji [31] used an urn model with multiple drawings to study the distribution of random circuits, Mahmoud studied urn models with sample size two [24]; we also refer the reader to the survey [14]. Chen and Wei [6] generalized the original two-color Polya urn model to multiple drawings, see also Chen and Kuba [5]. A generalization of the Friedman urn model was an discussed in [20]. We also refer the reader to the general works of Moler et al. [26], Higueras et al. [9], Renlund [30], as well as the recent very general work of Lasmar, Mailler and Selmi [15].

Urn models with multiple drawings and sample size \( m > 1 \) are usually more difficult to analyze compared to the ordinary urn models with sample size \( m = 1 \). The standard techniques - moment methods, analytic combinatorics and generating functions, embedding into continuous stochastic processes, and the contraction method - are not easily applicable. In particular, the expected values of continuous stochastic processes, and the contraction method - are not easily applicable. In particular, the expected values of \( X_n^{(i)} \) usually depend on the higher moments, making explicit and also asymptotic computations more complicated when the sample size is larger than one. It is however possible to obtain central limit theorems using the (RobbinsMonro) stochastic approximations techniques, see [26, 9, 25], obtaining quite general limit theorems, also for unbalanced urn models. A small drawback of these techniques is that expressions for the (positive integer) moments, in particular expectation and variance, moment convergence, as well as more precise information about the limit laws for so-called large-index urn models and also triangular urn models are at present elusive. Thus, it is of interest to study classes of urn models with multiple drawings which generalize the existing very precise results for the standard case of sample size \( m = 1 \).

In the two-color case \( X_n = (X_n^{(1)}, X_n^{(2)}) = (W_n, B_n) \) with black and white balls it turned out [21] that a class of balanced urn models with multiple drawings and linear affine expected value contained the special cases treated before [5, 6, 14, 20, 24, 26, 31]. This class was characterized by a condition on the first column entries of the ball replacement matrix:

\[
a_k = (m - 1)(a_{m-1} - a_m) + a_m, \quad \text{for } 0 \leq k \leq m, \tag{1}
\]

and the beforehand mentioned conditional expectation of the number of white balls \( W_n \) after \( n \) draws with an (affine) linear structure of the form

\[
E[W_n | \mathcal{F}_{n-1}] = \alpha_n W_{n-1} + \beta_n, \quad n \geq 1. \tag{2}
\]

Here, \( \alpha_n, \beta_n \) denote certain sequences depending only on the number of draws \( n \), \( a_{m-1}, a_m \), the total balance \( \sigma \), and \( \mathcal{F}_n \) denotes the sigma-algebra generated by the first \( n \) draws from the urn:

\[
\alpha_n = \frac{T_{n-1} + m(a_{m-1} - a_m)}{T_{n-1}}, \quad \beta_n = a_m, \quad n \geq 1.
\]

For sample size \( m = 1 \) the linear affine class reduces to ordinary balanced two-color urn models. The condition (1) and Equation (2) allowed to obtain exact and asymptotic expressions for the expected value and the variance generalizing the previously known results for the case \( m = 1 \), also leading to very precise limit laws [21, 22], laws of the iterated logarithm. Moreover it was proven that the martingale limits exhibit densities, bounded under suitable assumptions, and exponentially decaying tails.

We introduce a generalization of the condition (1) to the \( r \geq 2 \) color case for several different sampling models and the notion of linear urn models with multiple drawings.

**Definition 1.** A \( r \)-color urn model with multiple drawing and sample size \( m \geq 1 \) is called linear if the random vector \( X_n = (X_n^{(1)}, \ldots, X_n^{(r)}) \), specifying the composition of the urn after \( n \) draws, satisfies

\[
E[X_n | \mathcal{F}_{n-1}] = X_{n-1} \cdot C_n,
\]

for certain matrices \( \{C_k\}_{k \in \mathbb{N}} \subseteq \mathbb{R}^{r \times r} \).

**Example 1** (Sample size \( m = 1 \)). In the case of sample size \( m = 1 \) all models are by definition linear. There, \( C_k = I + \frac{1}{k-1} M \) and \( M \) denoting the \( r \times r \) ball transition matrix.
The definition above extends to unbalanced urn models. There, the individual row sums of the ball replacement matrix \( M \) are at least once different and the total number of balls \( T_n \) is itself a random variable. As a consequence, the matrices \( (C_k)_{k \in \mathbb{N}} \) are random. In this work we focus on the ball sampling point of view. However, it is possible to extend the definition to (generalized) Polya urn processes (compare with the discussions in \cite{10, 11, 29}).

We consider four different sampling models and will determine all linear \( r \)-color urn models with multiple drawings and will show in Theorems 1 and 2 that the resulting matrices \( C_n \) are given by

\[
C_n = I + \frac{1}{T_n-1} \cdot A,
\]

where \( I \) denotes the \( r \times r \) identity matrix and \( A \) a certain \( r \times r \) reduced ball transition matrix, depending on the transition matrix \( M \) of the urn model. We discuss the expected values \( \mu_n = \mathbb{E}[X_n] \) and the (co)-variances \( \Sigma_n = \mathbb{E}[(X_n - \Sigma)^T(X_n - \Sigma)] \) of linear classes. In the final section we briefly comment on limit laws.

1.1. Notation. Throughout this work we use boldface letters \( a, k, \) etc. to denote row vectors. Given a vector \( k = k_1 e_1 + \cdots + k_r e_r = (k_1, \ldots, k_r) \) we frequently use the shorthand notation \( (k_1 e_1) = (k_1, \ldots, k_r) \) for the multinomial coefficients. Moreover, given a vector \( x = (x_1, \ldots, x_r) \) we use the notation \( x^k := x_1^{k_1} \cdot x_2^{k_2} \cdots \cdot x_r^{k_r} \). We denote with \( x^k \) the falling factorials \( x^k = x(x-1) \cdots (x-(k-1)) \), and use the abbreviation \( x^k := x_1^{k_1} \cdot x_2^{k_2} \cdots \cdot x_r^{k_r} \).

2. Preliminaries - Sampling schemes

In the following we discuss in detail the four different sampling schemes and corresponding tenability assumptions. In the two-color case and \( m = 1 \) tenability assumptions were classified in \cite{1}. In multiple drawings schemes, sufficient conditions for tenability under sampling without replacement, unordered samples, were formulated by Konzem and Mahmoud \cite{18}.

2.1. Sampling schemes: unordered samples. We denote with \( \{X^{(1)}\}^{k_1}, \ldots, \{X^{(r)}\}^{k_r} \} \) an unordered sample of size \( m \) containing \( k_i \) balls of color \( i \), \( 1 \leq i \leq r \). The vector \( a = (a_1, \ldots, a_r) \in \mathbb{Z}^r \) with index \( k = k_1 e_1 + \cdots + k_r e_r \) specifies that when drawing a multiset \( \{X^{(1)}\}^{k_1}, \ldots, \{X^{(r)}\}^{k_r} \} \) of consisting of \( k_i \) balls colored \( i \), then we add/subtract \( a_i \) balls colored \( i \), \( 1 \leq i \leq r \), with \( \sum_{i=1}^r k_i = m \). The replacement of balls is specified by the the ball replacement matrix \( M \) of dimension \( (r+m-1)^m \times m \). It consists of \( (r+m-1)^m \) row vectors \( a = a_k \). The index \( k \) is contained in the discrete simplex

\[
\Delta = \{k = k_1 e_1 + \cdots + k_r e_r = (k_1, \ldots, k_r) \mid k_i \geq 0, \sum_{i=1}^r k_i = m\};
\]

here, \( e_i \) denotes the \( i \)th unit vector, such that

\[
M = (a_k)_{k \in \Delta}. \tag{4}
\]

Throughout this work we consider balanced urn models (and briefly only comment on extensions to unbalanced models) such that overall number of added/removed balls is a positive integer constant \( \sigma > 0 \), independent of the composition of the sample; Equivalently, we assume that the replacement matrix \( M \) has constant row sum such that for all row vectors \( a_k = (a_1, \ldots, a_r) \) of \( M \) we have

\[
a_k \cdot 1^T = \sum_{i=1}^r a_i = \sigma, \tag{5}
\]

with total balance \( \sigma > 0 \). Here \( 1 = \sum_{i=1}^r e_i = (1, \ldots, 1) \). As a consequence, the total number \( T_n = \sum_{i=1}^r X_n^{(i)} \) of ball after \( n \) draws is given by the deterministic quantity

\[
T_n = n \cdot \sigma + T_0, \quad \text{with} \quad T_0 = X_0^{(1)} + \cdots + X_0^{(r)} = X_0 \cdot 1^T.
\]
In the two-color case $r = 2$ with sample size $r = 2$ the ball replacement matrix $M$ is a $3 \times 2$ matrix $M = \begin{pmatrix} a_{01} & a_{02} \\ a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$; more commonly, the colors are addressed as white and black and $M$ is written as

\[
W = \begin{pmatrix} W_{WW} \\ W_{WB} \\ B \end{pmatrix} = \begin{pmatrix} a_{00} & b_{0} \\ a_{10} & b_{1} \\ a_{20} & b_{2} \end{pmatrix}.
\]

We consider two different sampling schemes for drawing an unordered sample of size $m$ at each step: model $M$ and model $\mathcal{R}^{1}$. In model $M$ we draw the $m$ balls without replacement. The $m$ balls are drawn at once and their colors are examined. After the sample is collected, we put the entire sample back in the urn and execute the replacement rules according to the counts of colors observed. The tenability assumption\(^2\) for $0 \leq k \leq m$. implies that for model $M$ the coefficients $a_{i}$ of the vector $a = a_{i} = \{a_{1}, \ldots, a_{r}\}_{i}$, with index $k = k_{1}e_{1} + \cdots + k_{r}e_{r}$ satisfy

\[
a_{i} \geq -k_{i}, \quad 1 \leq i \leq r.
\]

We are never forced to remove more balls of color $i$ than previously drawn in the sample and the process of drawing and replacing balls can be continued ad infinitum. Assume that an urn contains $c_{i}$ balls of color $i$, $1 \leq i \leq r$, with $c_{i} > 0$. The probability $\mathbb{P}(\{(X(1))^{k_{1}}, \ldots, (X(r))^{k_{r}}\})$ of drawing $k_{i}$ balls of color $i$ is given by

\[
\mathbb{P}(\{(X(1))^{k_{1}}, \ldots, (X(r))^{k_{r}}\}) = \begin{pmatrix} m \\ k_{1}, \ldots, k_{r} \end{pmatrix} \frac{k_{1}! \cdots k_{r}!}{(c_{1} + \cdots + c_{r})^{m}} = \frac{(c_{1})_{k_{1}} \cdots (c_{r})_{k_{r}}}{(c_{1} + \cdots + c_{r})^{m}},
\]

with $k_{i} \geq 0$ and $\sum_{i=1}^{r} k_{i} = m$. The sample follows a multivariate hypergeometric distribution with $m$ draws and numbers $c_{1}, \ldots, c_{r}$.

In model $\mathcal{R}$, we draw the $m$ balls with replacement. The $m$ balls are drawn one at a time. After a ball is drawn, its color is observed, and is inserted in the urn, and thus it might reappear in the sampling of one multiset. After $m$ balls are collected in this way (and they are all back in the urn), we execute the replacement rules according to the counts of colors observed. By the tenability assumption the coefficients $a_{i}$ of the vector $a = a_{i} = \{a_{1}, \ldots, a_{r}\}_{i}$, with index $k = k_{1}e_{1} + \cdots + k_{r}e_{r}$ satisfy

\[
a_{i} \geq -1, \quad \text{for } k_{i} > 0, \quad \text{and } a_{i} \geq 0 \quad \text{for } k_{i} = 0,
\]

$1 \leq i \leq r$. The probability $\mathbb{P}(\{(X(1))^{k_{1}}, \ldots, (X(r))^{k_{r}}\})$ of drawing $k_{i}$ balls of color $i$ is given by

\[
\mathbb{P}(\{(X(1))^{k_{1}}, \ldots, (X(r))^{k_{r}}\}) = \begin{pmatrix} m \\ k_{1}, \ldots, k_{r} \end{pmatrix} \frac{c_{1}^{k_{1}} \cdots c_{r}^{k_{r}}}{(c_{1} + \cdots + c_{r})^{m}},
\]

with $k_{i} \geq 0$ and $\sum_{i=1}^{r} k_{i} = m$. Thus, the sample follows a multinomial distribution with $m$ trials and probabilities $p_{i} = c_{i}/(c_{1} + \cdots + c_{r})$, $1 \leq i \leq m$.

Conditioning on the outcome of the $n$th draw, we obtain a distributional equation for the random vector $X_{n}$. The number of balls after $n$ draws is the number of balls after $n - 1$ draws, plus the contribution of $n$th draw:

\[
X_{n} = X_{n-1} + \sum_{k \in \Delta} a_{k} \cdot \mathbb{I}_{n}(\{(X(1))^{k_{1}}, \ldots, (X(r))^{k_{r}}\}),
\]

(6)

\(^{1}\)The name of the models stems from the original works of Chen et al. [5, 6]

\(^{2}\)These assumptions can be relaxed a little bit, if the initial values are adapted to the entries in the ball replacement matrix. E.g., for $m = 1$ the urn model with ball replacement matrix $\begin{pmatrix} 3 & 8 \\ 6 & -1 \end{pmatrix}$ is still tenable because $W_{0}$ is a multiple of 3 and $B_{0}$ a multiple of 4.
for \( n \geq 1 \) with (non-random) initial composition \( X_0 \). The indicators variables satisfy

\[
P\left( \mathcal{I}_n\left( (X^{(1)})^{k_1}, \ldots, (X^{(r)})^{k_r} \right) = 1 | \mathcal{F}_{n-1} \right) = \frac{\left( \begin{array}{c} X^{(1)}_{n-1} \\ \vdots \\ X^{(r)}_{n-1} \end{array} \right)_{\mathcal{I}_n} \cdot \left( \begin{array}{c} k_1 \\ \vdots \\ k_r \end{array} \right)}{m^{k_1, \ldots, k_r}}
\]

for model \( M \), and

\[
P\left( \mathcal{I}_n\left( (X^{(1)})^{k_1}, \ldots, (X^{(r)})^{k_r} \right) = 1 | \mathcal{F}_{n-1} \right) = \frac{\left( \begin{array}{c} m \\ k_1, \ldots, k_r \end{array} \right) \left( \begin{array}{c} X^{(1)}_{n-1} \\ \vdots \\ X^{(r)}_{n-1} \end{array} \right)_{\mathcal{I}_n}^{k_1, \ldots, k_r}}{m^{k_1, \ldots, k_r}}
\]

for model \( R \).

### 2.2. Sampling schemes: ordered samples

We use the notation \( (X^{(d_1)}, \ldots, X^{(d_m)}) \) to refer to an ordered sample of size \( m \) such that at the \( i \)th draw for the ordered sample we got a ball colored \( d_i, 1 \leq d_i \leq r, 1 \leq i \leq m \). The vector \( a = (a_1, \ldots, a_r) \in \mathbb{Z}^r \) with index \( d = d_1 \mathbf{e}_1 + \cdots + d_m \mathbf{e}_m \) specifies that when drawing \( (X^{(d_1)}, \ldots, X^{(d_m)}) \) then we add/subtract \( a_i \) balls colored \( i, 1 \leq i \leq r \). We consider all possible sequences of outcomes of length \( m \):

\[
S = \{d = (d_1, \ldots, d_m) | 1 \leq d_i \leq m, 1 \leq i \leq m\}.
\]

The ball replacement matrix \( M \) of the urn model is given by

\[
M = (a_d)_{d \in S}
\]

and has dimension \( rm \times r \). For example in the case \( m = 2 \) and \( r = 2 \) the ball replacement matrix \( M \) is a \( 4 \times 2 \) matrix

\[
\begin{pmatrix}
W & B \\
(W, W) & (W, B) \\
(W, B) & (B, W) \\
(B, B) & (B, B)
\end{pmatrix}
\]

more commonly written as

\[
M = \begin{pmatrix}
W \\
B
\end{pmatrix}
\begin{pmatrix}
a_0 & b_0 \\
a_1 & b_1 \\
a_2 & b_2 \\
a_3 & b_3
\end{pmatrix}
\]

We consider again two different sampling schemes for drawing an ordered sample of size \( m \) at each step: sampling without replacement in model \( M_{\text{SEQ}} \) and sampling with replacement in model \( R_{\text{SEQ}} \). The tenability assumptions are identical to the cases of unordered samples. Given \( d \in S \) let \( j(d) = (j_1(d), \ldots, j_r(d)) \in \Delta \), the discrete simplex \( \Delta \) as defined in (3), denote the vector of multiplicities:

\[
j_{\ell}(d) = \sum_{i=1}^{m} \mathbb{I}(d_i = \ell), \quad 1 \leq \ell \leq r.
\]

Note that, given a replacement matrix \( M = (a_k)_{k \in \Delta} \) associated to an unordered sampling scheme model \( M \) or model \( R \), there apparently exists an embedding into ordered sampling schemes. The corresponding replacement matrix \( \hat{M} = (\hat{a}_d)_{d \in S} \) can be defined as follows:

\[
\forall d \in S: \hat{a}_d = a_{j(d)}.
\]

Conditioning on the outcome of the \( n \)th draw, we obtain a distributional equation for the random vector \( X_n \). The indicators variables satisfy

\[
P\left( \mathcal{I}_n(\mathcal{S}^{d_1}, \ldots, \mathcal{S}^{d_m}) \right) = 1 | \mathcal{F}_{n-1} = \frac{\mathcal{S}^{d_1} \ldots \mathcal{S}^{d_m}}{m^{|S|}}
\]

for model \( M_{\text{SEQ}} \), and

\[
P\left( \mathcal{I}_n(\mathcal{S}^{d_1}, \ldots, \mathcal{S}^{d_m}) \right) = 1 | \mathcal{F}_{n-1} = \frac{\mathcal{S}^{d_1} \ldots \mathcal{S}^{d_m}}{m^{|S|}}
\]
for model $\mathcal{R}_{\text{SEQ}}$. The distributional equation for the random vector $X_n$ is given by

$$X_n = X_{n-1} + \sum_{d \in s} a_d \cdot I_n((X^{(d_1)}, \ldots, X^{(d_m)})). \quad (9)$$

### 3. Classification of linear models

#### 3.1. Unordered samples.

In the following we present a sufficient condition for (affine) linearity of the conditional expected values of the $r$-color urn model with multiple drawings for both unordered samples and ordered samples. We state first the result for unordered samples, generalizing the previous result of [21].

**Theorem 1.** A balanced $r$-color urn model with multiple drawings consisting of unordered samples of size $m$ is for both sampling schemes model $M$ and model $\mathcal{R}$ linear if and only if the vectors $a_k$ of the ball replacement matrix $M = (a_k)_{k \in \Delta}$ are given by affine combinations of the vectors $a_{m \cdot e_i}$:

$$a_k = \sum_{i=1}^r \frac{k_i}{m} a_{m \cdot e_i}.$$  

The conditional expected values is given by

$$E[X_n | \mathcal{F}_{n-1}] = X_{n-1} \cdot (I + \frac{1}{m} \cdot A),$$

with $A = (a_{m \cdot e_i})_{1 \leq i \leq m}$ denoting the reduced $r \times r$ ball replacement matrix.

**Remark 1.** We assume above that $M$ is balanced with a certain total balance $\sigma > 0$. Hence, we only can choose freely $r-1$ values in each row of the matrix with rows $a_{m \cdot e_i}, \; 1 \leq i \leq r$, due to the balance condition for the rows: $a_{m \cdot e_i} \cdot 1^T = \sigma$. We call linear urn models **triangular** if the reduced $r \times r$ ball replacement matrix $A$ is triangular.

**Remark 2.** The condition above also naturally arises using an entirely different approach based on analytic combinatorics and generating functions. It turns out that the condition above ensures a system of ordinary differential equations for the expected value of $X_n$ in contrast to the general case leading to (higher order) partial differential equations; this will be discussed elsewhere.

**Example 2 (Two-color case).** In the two-color case $r = 2$ and arbitrary $m \geq 1$ we re-obtain the previously derived condition for two colors white and black. Let $k = (k, m - k)$. We get

$$a_k = \frac{1}{m} (ka_{me_j} + (m-k)a_{me_j}).$$

Let $a_{me_j} = (a_m, b_m)$, $a_{me_j} = (a_0, b_0)$ and in general $a_k = (a_k, b_k)$, $0 \leq k \leq m$. Concerning the entries for the white balls we obtain the condition

$$a_k = \frac{1}{m} (ka_m - (m-k)a_0) = a_0 + k \cdot \frac{a_m - a_0}{m}, \quad 0 \leq k \leq m,$$

with $b_k = \sigma - a_k$. Rewriting the condition in terms of $a_{m-1}$ and $a_m$ instead of $a_0$ and $a_m$ we obtain

$$a_k = (m-k)(a_{m-1} - a_m) + a_m,$$

as stated in [21]. The special case $m = 2$ implies the condition of [24].

**Example 3 (Generalized Polya urn model).** Chen and Wei [6] introduced a multicolor generalization of the classical Polya urn for model $\mathcal{M}$, which was then also considered under model $\mathcal{R}$ [5]: “An urn contains balls of $r$ different colors. For each color $i$, we initially have $X_0^{(i)}$ balls. At each step we draw $m$ balls at random, say $k_i$ balls of color $i$, and their colors are noted. These balls are returned to the urn plus $c \cdot k_i$ balls of color $i$.” According to this description we have for all $k \in \Delta$: $a_k = \sum_{i=1}^r c \cdot k_i \cdot e_i$ and consequently the condition $a_k = \sum_{i=1}^r \frac{k_i}{m} a_{m \cdot e_i}$ is satisfied.
Proof. We obtain from the distributional equation the equation
\[ \mathbb{E}[X_n | F_{n-1}] = X_{n-1} + \sum_{k \in \Delta} a_k \cdot \mathbb{E}[I_n((X_1)^k, \ldots, (X^r)^{k^r})]. \]

For model \( R \) we get
\[ \mathbb{E}[X_n | F_{n-1}] = X_{n-1} + \sum_{k \in \Delta} a_k \frac{m^k}{C_{n-1}}. \]

On the other hand the linearity implies that \( \mathbb{E}[X_n | F_{n-1}] = X_{n-1} \cdot C_n \), such that
\[ X_{n-1} \cdot (C_n - 1) = \sum_{k \in \Delta} a_k \frac{m^k}{C_{n-1}}. \]

Multiplication with \( T_{n-1}^m \) gives
\[ X_{n-1} \cdot T_{n-1}^m \cdot (C_n - 1) = \sum_{k \in \Delta} a_k \left( \frac{m^k}{k} \right) X_{n-1}^k. \]

In order to match powers we use the simple but important fact that
\[ T_{n-1} = \sum_{i=1}^{r} X_{n-1}^{(i)}. \] (10)

We observe that unless \( C_n = 1 + \frac{1}{n} B_n \), with \( B_n \in \mathbb{R}^{r \times r} \), the powers of \( X_{n-1}^{k} \) do not match on both sides of the equation. Thus,
\[ X_{n-1} \cdot \left( \sum_{i=1}^{r} X_{n-1}^{(i)} \right)^{m-1} \cdot B_n = \sum_{k \in \Delta} a_k \left( \frac{m}{k} \right) X_{n-1}^k. \]

Set \( B_n = (b_{n,1}, \ldots, b_{n,r}) \), we get further
\[ \left( \sum_{i=1}^{r} X_{n-1}^{(i)} \right)^{m-1} \cdot \sum_{i=1}^{r} X_{n-1}^{(i)} \cdot b_{n,i} = \sum_{k \in \Delta} a_k \left( \frac{m}{k} \right) X_{n-1}^k. \]

The multinomial theorem gives
\[ \sum_{i=1, \ldots, t_i = m-1}^{r} \left( \sum_{i=1}^{r} X_{n-1}^{(i)} \right)^{m-1} \cdot \sum_{i=1}^{r} X_{n-1}^{(i)} \cdot b_{n,i} = \sum_{k \in \Delta} a_k \left( \frac{m}{k} \right) X_{n-1}^k. \]

Comparison of coefficients of \( X_{n-1}^k \) for every \( k \in \Delta \) gives the equation
\[ \sum_{i=1}^{r} \left( \frac{m-1}{k - e_i} \right) b_{n,i} = a_k \left( \frac{m}{k} \right), \]

such that for \( k = m \cdot e_i \) we obtain \( b_{n,i} = a_{m e_i} \), independent of \( n \). Consequently,
\[ \sum_{i=1}^{r} \left( \frac{m-1}{k - e_i} \right) a_{m e_i} = a_k \left( \frac{m}{k} \right), \]

and further
\[ a_k = \frac{1}{m} \sum_{i=1}^{r} \left( \frac{m-1}{k - e_i} \right) a_{m e_i}. \]

Since
\[ \frac{(m-1)!}{(m-1-k)!} = \frac{m!}{m! k_1! \ldots k_i! \ldots k_r!}, \]

we obtain the stated result.
For model $M$ we can proceed in similar way. We get first

$$\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] = X_{n-1} + \sum_{k \in \Delta} a_k \binom{m}{k} \frac{X_n^{k+1}}{T_n^{k+1}}.$$ 

Consequently, using the definition of the linearity

$$X_{n-1} \cdot T_{n-1}^{m} \cdot (C_n - 1) = \sum_{k \in \Delta} a_k \binom{m}{k} X_{n-1}^{k}.$$ 

We use again the fact that $T_{n-1} = \sum_{j=1}^{r} X_{n-1}^{(j)}$ and the multinomial theorem for the falling factorials. We observe that unless

$$C_n - I = \left( \frac{1}{T_{n-1} - j_1} b_{n,1}, \ldots, \frac{1}{T_{n-1} - j_r} b_{n,r} \right), \quad 0 \leq j_i \leq m - 1,$$

the coefficients do not match since the left hand side has powers larger than on the right. Assume that for $1 \leq i \leq r$ we have $j_i = 0$ such that

$$C_n - I = \frac{1}{T_{n-1}} \cdot B_n,$$  \tag{11}$$

similar to model $R$. Then, we can distribute the summand -1 in $(T_{n-1} - 1)^{m-1}$ according to the value of $i$ to form $X_{n-1}^{(i)} = 1$ and get

$$(T_{n-1} - 1)^{m-1} = \sum_{\ell_1 + \cdots + \ell_r = m-1} \binom{m-1}{1} \left( (X_{n-1}^{(i)} - 1) \sum_{j=1}^{r} (X_{n-1}^{(j)})^{\ell_i} \right)$$  \tag{12}$$

valid for any $1 \leq i \leq r$. Thus, we get further

$$\sum_{i=1}^{r} \sum_{\ell_1 + \cdots + \ell_r = m-1} \binom{m-1}{1} X_{n-1}^{1+\ell_1} b_{n,i} = \sum_{k \in \Delta} a_k \binom{m}{k} X_{n-1}^{k}.$$  \tag{13}$$

Comparison of coefficients of $X_{n-1}^{k}$ for every $k \in \Delta$ gives the equations

$$\sum_{i=1}^{r} \binom{m-1}{k-e_i} b_{n,i} = a_k \binom{m}{k},$$

identical to before, leading to the stated result. It remains to justify (11). Assume that $j_1$ is nonzero: we have

$$\frac{T_{n-1}^{m}}{T_{n-1} - j_1} = \frac{T_{n-1}^{1}}{T_{n-1} - j_1} \cdot (T_{n-1} - j_1 - 1)^{m-1-j_1}.$$ 

Using $T_{n-1}^{1} = (T_{n-1} - 1)^{1} + j_1 T_{n-1}^{j_1}$ we get

$$\frac{T_{n-1}^{m}}{T_{n-1} - j_1} = (T_{n-1} - 1)^{m} + j_1 \frac{T_{n-1}^{j_1}}{T_{n-1} - j_1} - (T_{n-1} - j_1 - 1)^{m-1-j_1}.$$ 

From the first summands we can argue exactly the same way as for $j_1 = 0$ leading to $a_k = \sum_{i=1}^{r} \binom{m}{k} X_{n-1}^{(i)}$. We observe that an additional contribution appears due to the second summand leading to additional powers in (13) non-existent on the right hand side. This is a contradiction, such that $j_1 = 0$. Alternatively, by the properties of the multivariate hypergeometric distribution we get

$$\mathbb{E}[X_n \mid \mathcal{F}_{n-1}] = X_{n-1} + \sum_{k \in \Delta} a_k \binom{m}{k} \frac{X_n^{k+1}}{T_n^{k+1}} = X_{n-1} + \sum_{i=1}^{r} X_{n-1}^{(i)} a_{m-e_i},$$

such that we have a second rationale for $j_1 = 0$. \hfill $\Box$
3.2. Ordered samples. For ordered samples we obtain the following counterpart of Theorem 1.

**Theorem 2.** A balanced $r$-color urn model with multiple drawings consisting of ordered samples of size $m$ is linear if and only if the vectors $a_d$ of $M$ satisfy for all $k \in \Delta$ the condition:

$$\sum_{d \in S} a_d = \sum_{i=1}^{r} \binom{m-1}{k-e_i} a_{i,1}.$$

Then for both sampling schemes with or without replacement the identity

$$\mathbb{E}[X_n | \mathcal{F}_{n-1}] = X_{n-1} \cdot (I + \frac{1}{T_{n-1}}) \cdot A$$

holds, with $A = \{a_{i,1}\}_{1 \leq i \leq m}$ denoting the reduced $r \times r$ matrix and $1 = \sum_{j=1}^{m} e_j$.

**Example 4** (Two-color case). We consider the balanced two-color case $r = 2$ with sample size $m = 2$. Let $a_{1,1} = a_{(1,1)} = (a_0, b_0), a_{1,2} = (a_1, b_1), a_{2,1} = (a_2, b_2)$ and $a_{2,1} = a_{(2,2)} = (a_3, b_3)$. We get the balancing condition $a_1 + a_2 = a_0 + a_3$ or equivalently $b_1 + b_2 = b_0 + b_3$.

**Example 5** (Embedding of $m = 1$). Given a quadratic replacement matrix $C \in \mathbb{R}^{r \times r}$ with row vectors $e_i, 1 \leq i \leq r$, associated to sample size $m = 1$. We can embed this case using a suitable setup of the replacement matrix $(7)$ $M = \{a_d\}_{d \in S}$ such that $\forall d \in S : a_d \in \{e_1, \ldots, e_r\}$. Given $k \in \Delta$ we choose $\binom{m-1}{k}$ out of the $\binom{m}{k}$ sequences $d \in S$ with $j(d) = k$ and define $a_d = e_i$. Consequently, $a_{i,1} = e_i$ and further

$$\sum_{d \in S} a_d = \sum_{i=1}^{r} \binom{m-1}{k-e_i} e_i = \sum_{i=1}^{r} \binom{m-1}{k-e_i} a_{i,1}.$$

Due to the specific structure of $M$ the indicators can be grouped together and the distributional equation 9 can be simplified mirroring essentially the case $m = 1$.

**Proof.** Since the proof is similar to the proof of Theorem 1 we will only discuss the main steps. We obtain from (9)

$$\frac{1}{T_{n-1}} X_n (C_n - I) = \sum_{d \in S} a_d \mathbb{E} [\mathbb{I}_n \{ (X_{(d_1)}, \ldots, X_{(d_m)}) \} | \mathcal{F}_{n-1}] .$$

For model $R_{SEQ}$ we observe that $C_n - I = \frac{1}{T_{n-1}} B_n$, with $B_n \in \mathbb{R}^{r \times r}$, such that

$$T_{n-1}^{m-1} X_n B_n = \sum_{d \in S} a_d \cdot X_{n-1}^{[j(d)]},$$

with $j(d)$ denoting the vector of multiplicities (8). Expansion of $T_{n-1}$ as stated in (10) and extraction of coefficients from $X_{n-1}^{k}$, with $k \in \Delta$, leads to the stated result. For model $M_{SEQ}$ we observe again that $C_n - I = \frac{1}{T_{n-1}} B_n$, with $B_n \in \mathbb{R}^{r \times r}$, obtain an equation similar to the previous one:

$$T_{n-1}^{m-1} X_n B_n = \sum_{d \in \mathcal{D}} a_d \cdot X_{n-1}^{[j(d)]}.$$

We use (12) and compare coefficients from $X_{n-1}^{k}$, with $k \in \Delta$, to get the stated result. \hfill \Box

4. Properties of linear models

In the following we obtain the expected value and the covariance of linear models. First we turn to exact representations, then we discuss asymptotic expansions.
4.1. Expected value and covariance. We collect a few properties of linear models.

We readily obtain a recurrence relation for the expected value from Theorem 1:

$$\mu_n = \mu_{n-1}(1 + \frac{1}{t_n^{-1}}A), \quad n \geq 1, \quad \mu_0 = X_0.$$ 

Iteration of the recurrence relation gives the explicit result for $\mu_n$.

**Proposition 1** (Expected value for linear models). For a linear r-color urn model with multiple drawings the expected value $\mu_n = \mathbb{E}[X_n]$ is given by

$$\mu_n = X_0(1 + \frac{1}{t_0}A) \ldots (1 + \frac{1}{t_{n-1}}A),$$

where $A$ denotes the reduced ball replacement matrix of the linear model.

Note that we can decompose the initial composition $X_0$ in terms of a base of (generalized) left-eigenvectors of the ball replacement matrix $A$. Due to the balance condition the vector $(1, \ldots, 1)$ is always an eigenvector corresponding to the largest eigenvalue $\sigma$. Specific assumptions on the algebraic multiplicity of the eigenvalues of $A$ readily lead to asymptotic expansions of the expected value.

The results for the covariance matrix $\Sigma_n = \mathbb{E}[(X_n - \mu_n)^T (X_n - \mu_n)]$ are not anymore model-independent in contrast to the expected value.

**Corollary 1** (Covariance matrix for linear models). The covariance matrix $\Sigma_n = \mathbb{E}[(X_n - \mu_n)^T (X_n - \mu_n)]$ satisfies the following recurrence relations:

- **Unordered samples of size $m$:**
  
  $$\Sigma_n = (1 + \frac{1}{r_n^{-1}}A)^T \Sigma_{n-1}(1 + \frac{1}{r_n^{-1}}A) - c_n A^T \Sigma_{n-1} + \mu_n^T \mu_n A$$

  + \frac{1}{m} \sum_{i=1}^{r} \mathbb{E}[X_{n-1}^{(i)} a_{m_{ei}}^T a_{m_{ei}}]

  with $c_n$ given by $c_n = \frac{1}{m t_n^{-1}}$ for model $M$ and $c_n = \frac{1}{m} \left( \frac{1}{m} - \frac{1}{t_n^{-1}} \right)$ for model $R$.

- **Ordered samples of size $m$:**
  
  $$\Sigma_n = (1 + \frac{1}{r_n^{-1}}A)^T \Sigma_{n-1}(1 + \frac{1}{r_n^{-1}}A) - \frac{1}{t_n^{-1}} A^T \Sigma_{n-1} A + d_n \sum_{d \in S} \mathbb{E}[X^{(d)}] a_{m_{d}}^T a_{d},$$

  with $d_n$ given by $d_n = \frac{1}{m t_n^{-1}}$ for model $M_{SEQ}$ and $d_n = \frac{1}{m} \left( \frac{1}{m} - \frac{1}{t_n^{-1}} \right)$ for model $R_{SEQ}$.

**Remark 3.** We observe that a linear model with unordered samples allows an exact and asymptotic computation of the covariance matrix. In contrast, for unordered samples the second (mixed) moments depend on higher moments $\mathbb{E}[X^{(d)}]$.

**Proof.** Concerning the covariance matrix we study the mixed moments $\mathbb{E}[X_{n}^T X_n]$ and use then $\mathbb{E}[X_{n}^T X_n] = \Sigma - \mu_n^T \mu_n$. The distributional equation (6) for $X_n$ implies that

$$\mathbb{E}[X_{n}^T X_n | \mathcal{F}_{n-1}] = X_{n-1}^T X_{n-1} + \sum_{k \in \Delta} \left( a_k^T X_{n-1} + X_{n-1}^T a_k + a_k^T a_k \right) \times \mathbb{E}[\{X^{(1)}\} k_1, \ldots, \{X^{(r)}\} k_r | \mathcal{F}_{n-1}].$$

For model $R$ we get

$$\mathbb{E}[X_{n}^T X_n | \mathcal{F}_{n-1}] = X_{n-1}^T X_{n-1} + \sum_{k \in \Delta} \frac{k_1}{m} \left( a_{m_{ei}}^T X_{n-1} + X_{n-1}^T a_{m_{ei}} \right) \left( \frac{m}{m} \right)^{k_1-1} \left( \frac{m}{m} \right)^{X_{n-1}^T a_{m_{ei}}}.$$
By the properties of the multinomial distribution
\[
\sum_{k \in \Delta} k_i k_j \binom{m}{k} X_{n-1}^k \frac{m(m-1)}{T_{n-1}} = \begin{cases} 
  m(m-1) \frac{X_{n-1}^{(i)} X_{n-1}^{(i)}}{I_{n-1}^{(i)^2}}, & \text{for } i \neq j, \\
  m(m-1) \frac{X_{n-1}^{(i)}}{I_{n-1}^{(i)}} + m \frac{X_{n-1}^{(i)}}{I_{n-1}^{(i)}}, & \text{for } i = j.
\end{cases}
\]

Consequently,
\[
E[X_n^T X_n | T_{n-1}] = X_{n-1}^T X_{n-1} + \frac{1}{I_{n-1}} \left( X_{n-1}^T X_{n-1} - X_{n-1}^T X_{n-1} A \right)
+ \frac{m-1}{m} I_{n-1} \frac{1}{T_{n-1}} A^T X_{n-1}^T X_{n-1} A + \frac{m}{m} \frac{r}{T_{n-1}} a_{mei} a_{mei}.
\]

We obtain the stated result for model \( \mathcal{X} \) with \( C_n = \frac{1}{T_{n-1}} \) after rearranging the terms and using
\[
E[X_n^T X_n] = \Sigma - \mu_n^T \mu_n.
\]

For model \( \mathcal{M} \) we proceed in an identical way. The only difference is that we use now the properties of the multivariate hypergeometric distribution
\[
\sum_{k \in \Delta} k_i k_j \binom{m}{k} X_{n-1}^k \frac{m(m-1)}{T_{n-1}} = \begin{cases} 
  m(m-1) \frac{X_{n-1}^{(i)} X_{n-1}^{(i)}}{I_{n-1}^{(i)^2}}, & \text{for } i \neq j, \\
  m(m-1) \frac{X_{n-1}^{(i)}}{I_{n-1}^{(i)}} + m \frac{X_{n-1}^{(i)}}{I_{n-1}^{(i)}}, & \text{for } i = j.
\end{cases}
\]

For ordered samples we can proceed in a similar manner: from the distributional equation (9) for \( X_n \) we get
\[
E[X_n^T X_n | T_{n-1}] = X_{n-1}^T X_{n-1} + \sum_{d \in S} \left( a_d^T X_{n-1} + X_{n-1}^T a_d + a_d^T a_d \right) \times E[I_n((X^{(1)})^{k_1}, \ldots, (X^{(r)})^{k_r}) | T_{n-1}].
\]

By linearity of the models we can simplify the sum
\[
\sum_{d \in S} \left( a_d^T X_{n-1} + X_{n-1}^T a_d \right) \times E[I_n((X^{(1)})^{k_1}, \ldots, (X^{(r)})^{k_r}) | T_{n-1}],
\]

leading to the stated result. \( \square \)

5. Perspective and Acknowledgments

In this note we obtained a classification of linear balanced multicolor urn models with multiple drawings and sample size \( m \) greater or equal one. There are various directions for further investigations. It is highly likely that the results of Janson [12] for \( m = 1 \) concerning convergence of the expected value and the covariances can be extended to the linear models. Concerning limit laws for model model \( \mathcal{X} \) the general results of Moler et al. [26] are applicable to prove a central limit theorem when the second largest eigenvalue of the reduced ball replacement matrix is less than \( \frac{2}{3} \); moreover, for the remaining models the more general results of [15] seem to ne
applicable to all the other sampling models: model $M$, model $M_{SEQ}$ and model $R_{SEQ}$. Concerning unordered samples and both model $M$ and model $R$ we expect that the algebraic approach of Pouyanne [29] for so-called large-index and also triangular urns can be suitably adapted to the linear urn models. Moreover, we believe that methods of [21, 22] can be adapted to the multicolor case, compare with the results of Müller [28] for sample size $m = 1$.

The author warmly thanks Hosam M. Mahmoud for a great many valuable and interesting discussions concerning the generalization of the conditions in [24, 21] leading to the present work.

References

[1] A. Bagchi and A. K. Pal (1985). Asymptotic normality in the generalized Pólya-Eggenberger urn model, with an application to computer data structures. SIAM J. Algebraic Discrete Math. 6, 394–405.

[2] B. Chauvin, N. Pouyanne and R. Sahnoun (2011). Limit distributions for large Pólya urns. The Annals of Applied Probability, 21, 1–32

[3] B. Chauvin, N. Pouyanne, and C. Mailer (2015). Smoothing equations for large Pólya urns, J. Theoret. Probab. 28(3), 923–957.

[4] M.-R. Chen, S.-R. Hsiao and T.-H. Yau (2012). A New Two-Urn Model. Journal of Applied Probability (to appear).

[5] M.-R. Chen and M. Kuba (2013). On generalized Pólya urn models. Journal of Applied Probability 50, Number 4, 909–921.

[6] M.-R. Chen and C.-Z. Wei (2005). A New Urn Model, Journal of Applied Probability 42, 964–976, 2005.

[7] F. Eggenberger and G. Pólya (1923). Über die Statistik verketteter Vorgänge. Z. Angewandte Math. Mech. 1, 279–289.

[8] P. Ehrenfest and T. Ehrenfest (1907). Über zwei bekannte Einwände gegen das Boltzmannsche H-theorem. Physikalische Zeitschrift, 8, 311–314.

[9] I. Higuera, J. Moler, F. Plo and M. San Miguel (2006). Central Limit Theorems for generalized Pólya urn models. Journal of Applied Probability, 43, 938–951.

[10] S. Janson (2004). Functional limit theorems for multitype branching processes and generalized Pólya urns. Stochastic processes and applications, 110, 177–245.

[11] S. Janson (2006). Limit theorems for triangular urn schemes, Probability Theory and Related Fields 134, 417–452.

[12] S. Janson (2016). Mean and Variance of balanced Pólya urns. Submitted.

[13] N. L. Johnson and S. Kotz (1977). Urn Models and Their Application. John Wiley, New York.

[14] N. L. Johnson, S. Kotz, and H. Mahmoud (2004). Pólya-type urn models with multiple drawings. Journal of the Iranian Statistical Society, 3, 165–173.

[15] N. Lasmar, C. Mailier, O. Selmi Multiple drawing multi-colour urns by stochastic approximation. Available on the arXiv, https://arxiv.org/abs/1611.09090.

[16] N. Müller and R. Neininger (2016).

[17] M. Knape and R. Neininger (2014). Pólya urns via the contraction method. Combinatorics, Probability and Computing 23, 1148–1186.

[18] S. Konzem and H. Mahmoud (2014). Characterization and Enumeration of Certain Classes of Tenable Pólya Urns Grown by Drawing Multisets of Balls. Methodol. Comput. Appl. Probab., 1–17.

[19] S. Kotz and N. Balakrishnan (1997). Advances in urn models during the past two decades. Advances in Combinatorial Methods and Applications to Probability and Statistics, Birkhäuser, Boston, MA, pp. 203–257.

[20] M. Kuba, H. Mahmoud and A. Panholzer (2013). Analysis of a generalized Friedman’s urn with multiple drawings. Discrete Applied Mathematics, Volume 161, Issue 18, 2968–2984.

[21] M. Kuba and H. Mahmoud (2015). On urn models with multiple drawings. Submitted.

[22] M. Kuba and H. Sulzbach (2016). On martingale tail sums in affine two-color urn models with multiple drawings. To appear in Journal of Applied Probability.

[23] H. Mahmoud (2008). Pólya Urn Models, Chapman-Hall, Orlando

[24] H. Mahmoud (2013). Drawing multisets of balls from tenable balanced linear urns. Probability in the Engineering and Informational Sciences, 27, 147–162.

[25] C. Mailier (2014) Describing the asymptotic behaviour of multicolour Plya urns via smoothing systems analysis, submitted. Available on the arXiv, http://arxiv.org/abs/1407.2879.

[26] J. Moler, F. Plo and H. Urmeneta (2013). A generalized Pólya urn and limit laws for the number of outputs in a family of random circuits. TEST, 22, 46–61.

[27] N.S. Müller and R. Neininger (2016). The CLT Analogue for Cyclic Urns. Analytic Algorithmics and Combinatorics (ANALCO), 121-127.

[28] N.S. Müller (2016). Central Limit Theorem Analagoues for Multicolour Urn Models. Available on the arXiv, http: //arxiv.org/abs/1604.02964.

[29] N. Pouyanne (2008). An algebraic approach to Pólya processes. Annales de l’Institut Henri Poincaré, Vol. 44, No. 2, 293–323.
[30] H. Renlund (2010). Generalized Polya urns via stochastic approximation, available on the arXiv, http://arxiv.org/abs/1002.3716.

[31] T. Tsukiji and H. Mahmoud (2001). A limit law for outputs in random circuits, Algorithmica, 403–412.

Markus Kuba, Institute of Applied Mathematics and Natural Sciences, University of Applied Sciences – Technikum Wien, Höchstädtplatz 5, 1200 Wien

E-mail address: kuba@technikum-wien.at