J. Appl. Prob. **44,** 661–669 (2007)
Printed in England

GENERALIZED PÓLYA URN DESIGNS WITH NULL BALANCE

ALESSANDRO BALDI ANTOGNINI * AND
SIMONE GIANNERINI,* ** *University of Bologna*

Abstract

In this paper we propose a class of sequential urn designs based on generalized Pólya urn (GPU) models for balancing the allocations of two treatments in sequential clinical trials. In particular, we consider a GPU model characterized by a 2×2 random addition matrix with null balance (i.e. null row sums) and replacement rule depending upon the urn composition. Under this scheme, the urn process has a Markovian structure and can be regarded as a random extension of the classical Ehrenfest model. We establish almost sure convergence and asymptotic normality for the frequency of treatment allocations and show that in some peculiar cases the asymptotic variance of the design admits a natural representation based on the set of orthogonal polynomials associated with the corresponding Markov process.

Keywords: Markov chain; sequential design; generalized Pólya urn model; Ehrenfest process

2000 Mathematics Subject Classification: Primary 62L05
Secondary 60F05; 60J20

1. Introduction

Assume that we want to carry out a sequential experiment for comparing the efficacy of two competing treatments. The peculiar feature of sequential designs is that they employ the data collected in the past of the experiment in order to choose subsequent allocations, called *design points*. These procedures are essential in clinical trials and in all those situations for which the entrance of statistical units is naturally sequential.

In designing clinical trials, a component of randomization in the assignments is commonly required to protect against various types of bias and it is often considered to be a fundamental tool for correct inferential procedures. As is well known, a large family of sequential randomized designs is based on *urn models*, which are classical tools as a randomization device [4], [6], [9], [10], [22], [23], [24], [26], [27], [28], [29] and the most basic urn model is the generalized (or extended) Pólya urn (GPU) model [12], [20].

Sequential clinical trials for the comparison of two treatments based on the GPU model can be described briefly as follows. There is an urn containing balls of two different types or colours, say *type* 1 (white) and *type* 2 (black), standing for treatment T_1 and T_2, respectively. The initial urn composition is denoted by $\boldsymbol{W}_0 = (W_{0,1}; W_{0,2})$, where $W_{0,i}$ represents the number of balls of type i at the beginning. In several applications $\boldsymbol{W}_0$ is assumed to be a deterministic quantity and in this paper we assume that $\boldsymbol{W}_0 = (w; w)$. Patients arrive sequentially and are randomly allocated to the treatments. At each step k, a ball is drawn at random from the urn; if the ball

Received 4 January 2007; revision received 11 May 2007.
* Postal address: Dipartimento di Scienze Statistiche, Via delle Belle Arti 41, Bologna 40126, Italy.
** Email address: giannerini@stat.unibo.it

is of type i, then treatment T_i is assigned to the kth subject ($i = 1, 2$) and the outcome Z_k is observed. Furthermore, the selected ball is replaced in the urn together with an additional $R_k(i, j)$ balls of type j (with $j = 1, 2$), where negative values are allowed and correspond to removals of balls. Thus, at each stage k the urn updating scheme is governed by the *addition matrix*

$$\boldsymbol{R}_k = \begin{pmatrix} R_k(1,1) & R_k(1,2) \\ R_k(2,1) & R_k(2,2) \end{pmatrix}, \tag{1}$$

and if the row sums are constant the urn is said to be *balanced*. After n steps, $\boldsymbol{W}_n = (W_{n,1}; W_{n,2})$ denotes the urn composition, where $W_{n,i} \geq 0$ represents the number of balls of type i present in the urn. Let $\boldsymbol{\delta}_{n+1} = (\delta_{n+1,1}; \delta_{n+1,2})$ represent the allocation of the treatment to the next subject, with $\delta_{n+1,i} = 1$ if the $(n+1)$th unit is assigned to T_i and 0 otherwise, then, for any $n \in \mathbb{N}$,

$$\mathrm{P}(\delta_{n+1,i} = 1 \mid \boldsymbol{W}_n) = \frac{W_{n,i}}{W_{n,1} + W_{n,2}} \quad \text{for } i = 1, 2. \tag{2}$$

This setting can be directly generalized to the case of several treatments (see, for instance [3]). Obviously, the treatment allocation process plays a fundamental role both from an ethical point of view and for inferential purposes (see, e.g. [18] and [10]). Let $\boldsymbol{N}_n = (N_{n,1}; N_{n,2})$ represent the number of allocations to the two treatments, where

$$N_{n,i} = \sum_{k=1}^{n} \delta_{k,i} \quad \text{and} \quad N_{n,1} + N_{n,2} = n.$$

Clearly, both the evolution of the GPU process $\{\boldsymbol{W}_n\}_{n\in\mathbb{N}}$ and the limiting behaviour of the associated urn design $\{\boldsymbol{N}_n\}_{n\in\mathbb{N}}$ depend on the nature of the sequence of rules $\{\boldsymbol{R}_n\}_{n\in\mathbb{N}}$.

1.1. The deterministic addition matrix

There is vast probabilistic literature regarding the evolution of the urn process $\{\boldsymbol{W}_n\}_{n\in\mathbb{N}}$ when the addition matrix is deterministic. In particular, if $\boldsymbol{R}_n$ has constant entries at each step we have

$$\boldsymbol{R}_n = \begin{pmatrix} \tau & \upsilon \\ \gamma & \varphi \end{pmatrix} \quad \text{for any } n \in \mathbb{N}; \tag{3}$$

thanks to its flexibility such a case has attracted the continued attention of several authors (see, e.g. [5], [12], and [14]). Historically, the constant and deterministic adding rule has been investigated in the context of sequential procedures for obtaining a balanced allocation of the treatments. Originally, Wei [26], [27] proposed a sequential GPU design based on rule (3) with $\tau = \varphi \geq 0$ and $\upsilon = \gamma \geq 0$; subsequently, Schouten [24] analyzed this procedure in the case in which $\tau = \varphi = -1$, that is, assuming that the extractions from the urn are made without replacement. When the urn is balanced and the row sums of (3) are equal to 0, the process $\{\boldsymbol{W}_n\}_{n\in\mathbb{N}}$ is a time-homogeneous Markov chain. Within this framework, Chen [9] proposed the Ehrenfest urn design (ED), a sequential procedure based on the Ehrenfest process, which can be regarded as a special case of the deterministic GPU model in (3) with $\tau = \varphi = -1$ and $\upsilon = \gamma = 1$. Recently, Baldi Antognini [6] analyzed a generalization of the ED based upon the following choice of parameters: $-\tau = -\varphi = \upsilon = \gamma \geq 0$. Also, Chen [10] introduced a modified version of the ED which satisfies the central limit property.

1.2. The random addition matrix

In the most general formulation of the GPU model, at any step n each $R_n(i, j)$ ($i, j = 1, 2$) may be a function of the accrued information, so that the addition matrix $\boldsymbol{R}_n$ is random

(several authors refer to this case as randomly reinforced GPU models [22] or GPU in a random environment [16]). For instance, in response-adaptive experiments the adding rule depends at each step on the observed responses (see, e.g. [17], [28], and [29]). Note that, if $\boldsymbol{R}_n$ is random, a key quantity that governs the evolution of the process is the *generating matrix* $\boldsymbol{H}_n = (\mathrm{E}[R_n(ij) \mid \mathfrak{F}_{n-1}, \boldsymbol{\delta}_n, Z_n]; i, j = 1, 2)$, where $\mathfrak{F}_k$ is the sigma field generated by $\{\boldsymbol{W}_0, \boldsymbol{\delta}_1, Z_1, \boldsymbol{W}_1, \ldots, \boldsymbol{\delta}_k, Z_k, \boldsymbol{W}_k\}$.

When, for any given pair of indexes $(i, j) \in \{1, 2\}^2$, $\{R_n(i, j)\}_{n\in\mathbb{N}}$ is an independent and identically distributed (i.i.d.) sequence of random variables, then $\boldsymbol{H}_n = \boldsymbol{H}$ for any n (i.e. the GPU model is said to be *homogeneous*) and Athreya and Karlin [1] and Smythe [25] derived the asymptotic behaviour of $\{\boldsymbol{W}_n\}_{n\in\mathbb{N}}$ and $\{\boldsymbol{N}_n\}_{n\in\mathbb{N}}$ under some conditions on the spectral structure of the generating matrix. In such a case, they assumed that the expected number of balls added at each step is a positive constant, namely that $\sum_{j=1}^{2} \mathrm{E}[R(ij)] = c > 0$ for any $i = 1, 2$. More recently, Janson [19] derived functional limit theorems for $\{\boldsymbol{W}_n\}_{n\in\mathbb{N}}$ and $\{\boldsymbol{N}_n\}_{n\in\mathbb{N}}$ for homogeneous GPU models under a set of assumptions [19, Assumptions A1–A6] which prevent the extinction of balls. Note that, as also shown by the author, nonextinction is also guaranteed when the process evolves as a Markov chain; however, in this instance condition A3 no longer holds and so the theoretical framework described in this paper cannot be exploited.

Other recent contributions include [2], [3], [4], [17], and [29], where the asymptotic normality of $\{\boldsymbol{W}_n\}_{n\in\mathbb{N}}$ and $\{\boldsymbol{N}_n\}_{n\in\mathbb{N}}$ is derived for nonhomogeneous GPU models by assuming that the total number of balls added at each step is a positive constant, namely that $\sum_{j=1}^{2} R(ij) = c > 0$ for any $i = 1, 2$.

In this paper we analyse sequential urn designs based upon a class of randomly reinforced GPU models with null balance, where the addition matrix depends on the urn composition. Such a class of Markovian processes can be regarded as a randomized extension of the Ehrenfest process and generalizes previous contributions on the topic (e.g. [9], [11], [13], and [21]). By exploiting martingale theory, we can identify sequential urn procedures that are asymptotically balanced and satisfy the central limit property. In particular, we focus our attention to the case in which any draw generates the birth or death of exactly one ball and we show that the asymptotic variance of the frequency of allocations admits a natural representation based on the eigenstructure of the transition matrix of the associated Markov chain.

2. Randomly reinforced GPU model with null balance for the balanced allocation of two treatments

In this section we consider a family of randomly reinforced GPU models, which allows us to define a class of sequential urn designs for the balanced allocation of two treatments. In particular, we analyse a specification of the GPU model in (1) with null balance, denoted by BN-GPU, in which any draw generates a random birth or death of balls depending only on the current number of balls already present in the urn.

From now on, assume that at each step n the random addition matrix is given by

$$\boldsymbol{R}_n = \begin{pmatrix} -A_n & A_n \\ B_n & -B_n \end{pmatrix}, \tag{4}$$

where A_n and B_n are nonnegative integer random variables. From (4), the total number of balls in the urn is kept constant (i.e. $W_{n,1} + W_{n,2} = 2w$ for any n), which allows us to restrict our attention to the univariate process $\{W_{n,1}\}_{n\in\mathbb{N}}$.

Clearly, the urn process (or, equivalently, the associated urn design) is completely determined by the specification of the sequences of random variables $\{A_n\}_{n\in\mathbb{N}}$ and $\{B_n\}_{n\in\mathbb{N}}$. Throughout

this paper, we define the BN-GPU model by assuming that, at each step n, given $W_{n-1,1} = x$, A_n and B_n satisfy the following conditions:

- for $x \in \{1, \ldots, 2w\}$, A_n has support $\mathfrak{A}_x \subseteq \{0, \ldots, x\}$ and probability distribution function (PDF) $f(\cdot; x)$, where

$$\mathrm{P}(A_n = a \mid \mathfrak{F}_{n-1}, \delta_{n,1}, Z_n) = \mathrm{P}(A_n = a \mid W_{n-1,1} = x) = f(a; x), \qquad a \in \mathfrak{A}_x;$$

- for $x \in \{0, \ldots, 2w-1\}$, B_n has support $\mathfrak{B}_x \subseteq \{0, \ldots, 2w-x\}$ and PDF $g(\cdot; x)$, where

$$\mathrm{P}(B_n = b \mid \mathfrak{F}_{n-1}, \delta_{n,1}, Z_n) = \mathrm{P}(B_n = b \mid W_{n-1,1} = x) = g(b; x), \qquad b \in \mathfrak{B}_x.$$

These conditions ensure the correct specification of the BN-GPU model, since (4) involves subtraction; furthermore, they characterize its probability structure. Note that, if $W_{n-1,1} = 0$ or $W_{n-1,1} = 2w$ we do not pose any restriction on A_n or B_n, respectively, since it is not relevant to the evolution of the process as the draw becomes deterministic.

Under this scheme, the addition matrix $\boldsymbol{R}_n$ depends on stage n only through the number of balls in the urn, $W_{n-1,1}$. For instance, given $W_{n-1,1} = x$, if the ball that has been drawn is of type 1 then, with probability $f(a; x)$, a balls of type 1 will be removed and a balls of type 2 will be added.

This scheme is quite flexible, since the PDFs can be chosen ad hoc in order to model several kinds of replacements, and generalizes some proposals in the literature (see [9], [10], [11], [13], and [21]). Observe that, if $A_n = B_n = 1$ almost surely (a.s.) for any n, the ball that has been drawn is placed in the other urn with probability 1 and $\{W_{n,1}\}_{n\in\mathbb{N}}$ becomes the classical Ehrenfest urn process (so that the associated design is the ED proposed by Chen [9]).

Under the BN-GPU model, at each step n the urn process satisfies the recursive relation

$$W_{n,1} = W_{n-1,1} - \delta_{n,1} A_n + (1 - \delta_{n,1}) B_n, \tag{5}$$

and the sequence $\{W_{n,1}\}_{n\in\mathbb{N}}$ is a time-homogeneous Markov chain on the state space $\mathcal{X} = \{0, \ldots, 2w\}$ with transition matrix $\boldsymbol{P} = (p_{x,y})$ given by

$$p_{x,y} = \begin{cases} g(y-x; x)\left(1 - \dfrac{x}{2w}\right), & y > x, \\ f(0; x)\dfrac{x}{2w} + g(0; x)\left(1 - \dfrac{x}{2w}\right), & y = x, \\ f(x-y; x)\dfrac{x}{2w}, & y < x, \end{cases}$$

for $x = 1, \ldots, 2w-1$ and

$$p_{0,y} = g(y; 0) \quad \text{and} \quad p_{2w,y} = f(2w-y; 2w), \qquad y \in \mathcal{X}.$$

Clearly, the properties of the chain vary with $\{f(\cdot; x)\}_{x=1,\ldots,2w}$ and $\{g(\cdot; x)\}_{x=0,\ldots,2w-1}$, and throughout this paper we assume that A_n and B_n are nondegenerate random variables and that the following condition holds.

(C1) The families of PDFs $\{f(\cdot; x)\}_{x=1,\ldots,2w}$ and $\{g(\cdot; x)\}_{x=0,\ldots,2w-1}$ are chosen such that the Markov chain $\{W_{n,1}\}_{n\in\mathbb{N}}$ is irreducible.

Under assumption (C1), the urn process $\{W_{n,1}\}_{n\in\mathbb{N}}$ is ergodic (i.e. irreducible and positive recurrent), since the state space $\mathcal{X}$ is finite. Thus, the stationary distribution $\boldsymbol{\pi} = (\pi(0), \ldots, \pi(2w))$ exists and is unique and from now on let E_π denote the ergodic average, namely

$$E_\pi = \sum_{x=0}^{2w} x\pi(x).$$

2.1. Some general properties of BN-GPU designs

In this section we analyse the properties of sequential urn procedures generated by the BN-GPU model in (4). From (2) and (5), at each step n the treatment allocations are governed by

$$\mathrm{P}(\delta_{n+1,1} = 1 \mid W_{n,1}) = \frac{W_{n,1}}{2w} = \frac{1}{2} - \sum_{j=1}^{n}\left\{\frac{\delta_{j,1}A_j - (1-\delta_{j,1})B_j}{2w}\right\}.$$

Clearly, the properties of the BN-GPU designs vary on the basis of the specification of the sequences of random variables $\{A_n\}_{n\in\mathbb{N}}$ and $\{B_n\}_{n\in\mathbb{N}}$. The following theorem highlights the asymptotic relationship between the frequency of treatment allocations and the corresponding urn process.

Theorem 1. *For any sequential urn design based on the BN-GPU model in (4) which satisfies assumption (C1), then*

$$\lim_{n\to\infty}\frac{N_{n,1}}{n} = \frac{E_\pi}{2w} \quad a.s. \tag{6}$$

and

$$\sqrt{n}\left(\frac{N_{n,1}}{n} - \frac{E_\pi}{2w}\right) \xrightarrow{\mathrm{D}} N\left(0; \frac{\sigma^2}{(2w)^2}\right) \quad as\ n \to \infty,$$

where

$$\sigma^2 = \lim_{n\to\infty}\frac{1}{n}V\left(\sum_{k=0}^{n-1}W_{k,1}\right) < \infty \tag{7}$$

(the special case in which $\sigma^2 = 0$ corresponds to the convergence in probability to 0).

Proof. At each step n, consider the martingale process $\{M_n, \mathfrak{F}_n\}$, where

$$M_n = N_{n,1} - \frac{1}{2w}\sum_{k=0}^{n-1}W_{k,1} = \sum_{k=1}^{n}\left(\delta_{k,1} - \frac{W_{k-1,1}}{2w}\right) = \sum_{k=1}^{n}Q_k.$$

Let $\mathfrak{F}_0$ be the trivial σ-field, since

$$\mathrm{E}[Q_{n+1}^2 \mid \mathfrak{F}_n] = \frac{W_{n,1}}{2w}\left(1 - \frac{W_{n,1}}{2w}\right) \le \frac{1}{4} \quad \text{a.s. for any } n \in \mathbb{N}.$$

Then $\sum_{k=1}^{\infty}k^{-2}\,\mathrm{E}[Q_k^2 \mid \mathfrak{F}_{k-1}] < \infty$ and, hence, from Theorem 2.18 of [15], we have

$$\lim_{n\to\infty}\frac{M_n}{n} = \lim_{n\to\infty}\left[\frac{N_{n,1}}{n} - \frac{1}{2w}\frac{1}{n}\sum_{k=0}^{n-1}W_{k,1}\right] = 0 \quad \text{a.s.} \tag{8}$$

Thus, the almost sure convergence in (6) follows from the strong law of large numbers for ergodic Markov chains, since

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} W_{k,1} = E_\pi \quad \text{a.s.} \tag{9}$$

Furthermore, by (6), (8), and (9) it follows that

$$\lim_{n\to\infty} \frac{N_{n,1}}{\sum_{k=0}^{n-1} W_{k,1}} = \frac{1}{2w} \quad \text{a.s.} \tag{10}$$

and, from the central limit theorem for ergodic Markov chains (see [8]),

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{1}{n} \sum_{k=0}^{n-1} W_{k,1}\right) = W^* \quad \text{in law,} \tag{11}$$

where $W^* \sim N(E_\pi; \sigma^2)$ and σ^2 is as defined in (7). Thus, by (10) and (11), we have

$$\lim_{n\to\infty} \sqrt{n}\left(\frac{N_{n,1}}{n}\right) = \frac{W^*}{2w} \quad \text{in law.}$$

3. The BN^1-GPU designs

Consider now the special case of the BN-GPU model, denoted by BN^1-GPU, under which any draw generates the birth or the death of exactly one ball. Formally, we assume that at each step n and for any given $W_{n-1,1} = x$, $\{A_n\}_{n\in\mathbb{N}}$ and $\{B_n\}_{n\in\mathbb{N}}$ are sequences of Bernoulli random variables such that

$$\begin{aligned} \mathrm{P}(A_n = 1 \mid W_{n-1,1} = x) &= f(1; x) = f(x), \\ \mathrm{P}(B_n = 1 \mid W_{n-1,1} = x) &= g(1; x) = g(x), \end{aligned} \tag{12}$$

with $0 < f(x), g(x) < 1$ for any $x \in \mathcal{X}$. From (12), the sequence $\{W_{n,1}\}$ is a time-homogeneous Markov chain on $\mathcal{X} = \{0, \ldots, 2w\}$ with tridiagonal (or Jacobian) transition matrix $\boldsymbol{P} = (p_{x,y})$ given by

$$p_{x,y} = \begin{cases} g(x)\left(1 - \dfrac{x}{2w}\right), & y = x + 1, \\ 1 - g(x)\left(1 - \dfrac{x}{2w}\right) - f(x)\dfrac{x}{2w}, & y = x, \\ f(x)\dfrac{x}{2w}, & y = x - 1, \end{cases}$$

together with the boundary condition $p_{0,-1} = p_{2w,2w+1} = 0$. It can be shown that the chain $\{W_{n,1}\}_{n\in\mathbb{N}}$ is ergodic and aperiodic, with stationary distribution $\boldsymbol{\pi}$ given by the equilibrium equations

$$\begin{aligned} \pi(x) &= \pi(x-1)\xi_x, \qquad x = 1, \ldots, 2w, \\ \pi(0) &= \left[1 + \sum_{j=1}^{2w} \prod_{x=1}^{j} \xi_x\right]^{-1}, \end{aligned} \tag{13}$$

where, for any $x = 1, \dots, 2w$,

$$\xi_x = \frac{g(x-1)(1-(x-1)/2w)}{f(x)(x/2w)} = \left(\frac{2w-x+1}{x}\right)\frac{g(x-1)}{f(x)}; \tag{14}$$

also, from (13), the ergodic average becomes

$$E_\pi = \pi(0)\sum_{x=1}^{2w} x \prod_{k=1}^{x} \xi_k. \tag{15}$$

Under the BN^1-GPU model, the chain $\{W_{n,1}\}_{n\in\mathbb{N}}$ is time-reversible, so that, letting $\lambda_0, \dots, \lambda_{2w}$ be the set of eigenvalues of $\boldsymbol{P}$, each λ_x is real and with suitable ordering we have $1 = \lambda_0 > \lambda_1 \geq \cdots \geq \lambda_{2w} \geq -1$. Also, for any $x = 0, \dots, 2w$, let $\boldsymbol{v}_x = (v_x(0), \dots, v_x(2w))$ be the right eigenvector associated with λ_x, where $\boldsymbol{v}_0 = \mathbf{1}$. Since eigenvectors are determined up to multiplication by a nonnull scalar, from now on we set

$$\sum_{j=0}^{2w} v_x^2(j)\pi(j) = 1 \quad \text{for any } x = 0, \dots, 2w,$$

i.e. $\boldsymbol{v}_0, \dots, \boldsymbol{v}_{2w}$ is the family of right eigenvectors of $\boldsymbol{P}$ which are orthonormal with respect to the stationary distribution $\boldsymbol{\pi}$.

Proposition 1. *Under any BN^1-GPU design, if the functions $f(\cdot)$ and $g(\cdot)$ in (12) satisfy the symmetric condition*

$$f(x) = g(2w-x) \quad \textit{for any } x = 0, \dots, 2w, \tag{16}$$

then the corresponding design is asymptotically balanced, namely

$$\lim_{n\to\infty} \frac{N_{n,1}}{n} = \frac{1}{2} \quad \textit{a.s.}$$

and furthermore, as $n \to \infty$,

$$\sqrt{n}\left(\frac{N_{n,1}}{n} - \frac{1}{2}\right) \xrightarrow{\text{D}} N\left(0; \frac{\sigma^2}{(2w)^2}\right), \tag{17}$$

where

$$\sigma^2 = \sum_{x=1}^{2w} \frac{1+\lambda_x}{1-\lambda_x}\left(\sum_{j=1}^{2w} j v_x(j)\pi(j)\right)^2. \tag{18}$$

Proof. Given assumption (16), from (13) and (14) we can write

$$\begin{aligned}\pi(x) &= \pi(0)\frac{g(0)\cdots g(x-1)}{g(2w-1)\cdots g(2w-x)}\prod_{k=1}^{x}\left(\frac{2w-k+1}{k}\right)\\ &= \pi(0)\frac{g(0)\cdots g(x-1)}{g(2w-1)\cdots g(2w-x)}\binom{2w}{x};\end{aligned}$$

so that

$$\pi(2w) = \pi(0)\frac{g(0)\cdots g(2w-1)}{g(2w-1)\cdots g(0)}\binom{2w}{2w} = \pi(0)$$

and, in general, $\pi(x) = \pi(2w - x)$ for any $x = 0, \ldots, 2w$, where

$$\pi(0) = \left[1 + \sum_{j=1}^{2w} \prod_{k=1}^{j} \binom{2w}{j} \frac{g(0) \cdots g(k-1)}{g(2w-1) \cdots g(2w-k)}\right]^{-1}.$$

Thus, from (15), we have

$$E_\pi = \sum_{x=0}^{w-1} x\pi(x) + w\pi(w) + \sum_{r=0}^{w-1} (2w - r)\pi(2w - r) = w \sum_{x=0}^{2w} \pi(x) = w,$$

so that the proof follows directly from Theorem 1. The spectral representation of the asymptotic variance in (18) is derived through the central limit theorem for reversible Markov chains (see, for instance [7, pp. 232–235]).

The adoption of a random replacement rule renders the model quite general. However, note that this gain in flexibility produces an increase of variability in the asymptotic behaviour of treatment allocations. In fact, for instance, under Chen's ED

$$\lim_{n \to \infty} \sqrt{n}\left(\frac{N_{n,1}}{n} - \frac{1}{2}\right) = 0 \quad \text{in probability,}$$

whereas for the BN^1-GPU model (17) holds.

In the following we analyse some special cases of the BN^1-GPU model designed for balancing the treatment allocations.

Example 1. (*Constant expected replacement.*) Let us assume that $f(x) = t$ and $g(x) = s$, for any $x = 0, \ldots, 2w$ with $0 < t, s < 1$. Under this choice, $\{A_n\}_{n \in \mathbb{N}}$ and $\{B_n\}_{n \in \mathbb{N}}$ are sequences of i.i.d. Bernoulli random variables, $A_n \sim \text{Ber}(t)$ and $B_n \sim \text{Ber}(s)$, so that the adding rule does not depend on the number of balls already in the urn and the BN^1-GPU becomes homogeneous. From Proposition 1, in order to obtain a sequential procedure which is asymptotically balanced we set $s = t$. In such a case the stationary distribution of the ergodic chain $\{W_{n,1}\}_{n \in \mathbb{N}}$ is binomial with $\pi(\cdot) = \text{Bin}(2w, \frac{1}{2})$. Moreover, from the spectral representation given in [21] the transition matrix $\boldsymbol{P}$ has eigenvalues $\lambda_x = 1 - sx/w$ for $x = 0, \ldots, 2w$, with corresponding right eigenvectors given by the sequence of Krawtchouk polynomials (see, for instance [11]).

Example 2. (*Linear expected replacement.*) Consider now the nonhomogeneous BN^1-GPU model with linear replacement probabilities, namely

$$f(x) = \frac{x}{2w} \quad \text{and} \quad g(x) = 1 - \frac{x}{2w}, \quad \text{for any } x = 0, \ldots, 2w.$$

Observe that this choice of functions $f(\cdot)$ and $g(\cdot)$ satisfies (16), so that the corresponding urn procedure is asymptotically balanced. Furthermore, as shown by Dette [11], the stationary distribution of the chain $\{W_{n,1}\}_{n \in \mathbb{N}}$ is hypergeometric with

$$\pi(x) = \binom{4w}{2w}^{-1} \binom{2w}{x}^2, \qquad x = 0, \ldots, 2w.$$

The set of eigenvalues is given by $\lambda_x = 1 - x(4w + 1 - x)/4w^2$ for $x = 0, \ldots, 2w$ and the corresponding right eigenvectors are the so-called Hahn–Eberlein polynomials.

Acknowledgements

The authors wish to thank the anonymous referee and the Associate Editor for their constructive comments. The research has been supported by the PRIN 2005 project 'Statistical Design of Continuous Product Innovation'.

References

[1] Athreya, K. B. and Karlin, S. (1968). Embedding of urn schemes into continuous time Markov branching processes and related limit theorems. *Ann. Math. Statist.* **39,** 1801–1817.
[2] Bai, Z. and Hu, F. (1999). Asymptotic theorems for urn models with nonhomogeneous generating matrices. *Stoch. Process. Appl.* **80,** 87–101.
[3] Bai, Z. and Hu, F. (2005). Asymptotics in randomized urn models. *Ann. Appl. Prob.* **15,** 914–940.
[4] Bai, Z., Hu, F. and Zhang, L.-X. (2002). Gaussian approximation theorems for urn models and their applications. *Ann. Appl. Prob.* **12,** 1149–1173.
[5] Balaji, S., Mahmoud, H. and Watanabe, O. (2006). Distributions in the Ehrenfest process. *Statist. Prob. Lett.* **76,** 666–674.
[6] Baldi Antognini, A. (2005). On the speed of convergence of some urn designs for the balanced allocation of two treatments. *Metrika* **62,** 309–322.
[7] Brémaud, P. (1999). *Markov chains. Gibbs fields, Monte Carlo simulation, and Queues* (Texts Appl. Math. **31**). Springer, New York.
[8] Chan, K. and Geyer, C. (1994). Discussion: Markov chains for exploring posterior distributions. *Ann. Statist.* **22,** 1747–1758.
[9] Chen, Y.-P. (2000). Which design is better? Ehrenfest urn versus biased coin. *Adv. Appl. Prob.* **32,** 738–749.
[10] Chen, Y.-P. (2006). A central limit property under a modified Ehrenfest urn design. *J. Appl. Prob.* **43,** 409–420.
[11] Dette, H. (1994). On a generalization of the Ehrenfest urn model. *J. Appl. Prob.* **31,** 930–939.
[12] Flajolet, P., Gabarró, J. and Pekari, H. (2005). Analytic urns. *Ann. Prob.* **33,** 1200–1233.
[13] Garibaldi, U. and Penco, M. (2000). Ehrenfest's urn model generalized: an exact approach for market participation models. *Statistica Applicata* **12,** 249–272.
[14] Gouet, R. (1993). Martingale functional central limit theorems for a generalized Pólya urn. *Ann. Prob.* **21,** 1624–1639.
[15] Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.
[16] Higueras, I., Moler, J., Plo, F. and San Miguel, M. (2003). Urn models and differential algebraic equations. *J. Appl. Prob.* **40,** 401–412.
[17] Higueras, I., Moler, J., Plo, F. and San Miguel, M. (2006). Central limit theorems for generalized Pólya urn models. *J. Appl. Prob.* **43,** 938–951.
[18] Hu, F. and Rosenberger, W. (2003). Optimality, variability, power: evaluating response-adaptive randomization procedures for treatment comparisons. *J. Amer. Statist. Assoc.* **98,** 671–678.
[19] Janson, S. (2004). Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stoch. Process. Appl.* **110,** 177–245.
[20] Johnson, N. L. and Kotz, S. (1977). *Urn Models and Their Application*. John Wiley, New York.
[21] Krafft, O. and Schaefer, M. (1993). Mean passage times for tridiagonal transition matrices and a two-parameter Ehrenfest urn model. *J. Appl. Prob.* **30,** 964–970.
[22] Muliere, P., Paganoni, A. and Secchi, P. (2006). A randomly reinforced urn. *J. Statist. Planning Infer.* **136,** 1853–1874.
[23] Rosenberger, W. (2002). Randomized urn models and sequential designs. *Sequential Anal.* **21,** 1–28.
[24] Schouten, H. (1995). Adaptive biased urn randomization in small strata when blinding is impossible. *Biometrics* **51,** 1529–1535.
[25] Smythe, R. T. (1996). Central limit theorems for urn models. *Stoch. Process. Appl.* **65,** 115–137.
[26] Wei, L. (1977). A class of designs for sequential clinical trials. *J. Amer. Statist. Assoc.* **72,** 382–386.
[27] Wei, L. (1978). An application of an urn model to the design of sequential controlled clinical trials. *J. Amer. Statist. Assoc.* **73,** 559–563.
[28] Wei, L. (1979). The generalized Pólya's urn design for sequential medical trials. *Ann. Statist.* **7,** 291–296.
[29] Zhang, L.-X., Hu, F. and Cheung, S. H. (2006). Asymptotic theorems of sequential estimation-adjusted urn models. *Ann. Appl. Prob.* **16,** 340–369.