*The Annals of Statistics*
2011, Vol. 39, No. 1, 643–671
DOI: 10.1214/10-AOS851


# IMMIGRATED URN MODELS—THEORETICAL PROPERTIES AND APPLICATIONS

BY LI-XIN ZHANG[1], FEIFANG HU[2], SIU HUNG CHEUNG[3] AND WAI SUM CHAN

*Zhejiang University, University of Virginia, The Chinese University of Hong Kong and The Chinese University of Hong Kong*

Urn models have been widely studied and applied in both scientific and social science disciplines. In clinical studies, the adoption of urn models in treatment allocation schemes has proved to be beneficial to researchers, by providing more efficient clinical trials, and to patients, by increasing the likelihood of receiving the better treatment. In this paper, we propose a new and general class of immigrated urn (IMU) models that incorporates the immigration mechanism into the urn process. Theoretical properties are developed and the advantages of the IMU models are discussed. In general, the IMU models have smaller variabilities than the classical urn models, yielding more powerful statistical inferences in applications. Illustrative examples are presented to demonstrate the wide applicability of the IMU models. The proposed IMU framework, including many popular classical urn models, not only offers a unify perspective for us to comprehend the urn process, but also enables us to generate several novel urn models with desirable properties.

## 1. Introduction.

1.1. *Urn models and their applications.* Urn models have long been considered powerful mathematical instruments in many areas, including the physical sciences, biological sciences, social sciences and engineering [Johnson and Kotz (1977), Kotz and Balakrishnan (1997)]. For example, in medical science, Knoblauch, Neitz and Neitz (2006) apply an urn model to study cone ratios in human and macaque retinas. In population genetics, Hoppe (1984) and Donnelly and Kurtz (1996) employ a Pólya-like urn model to study Ewen's sampling distribution in neutral genetics models. Benaïm, Schreiber and Tarrès (2004) also make use of

---

Received March 2010; revised September 2010.

[1]Supported by Grants from the National Natural Science Foundation of China (No. 11071214), Natural Science Foundation of Zhejiang Province (No. R6100119) and Fundamental Research Funds for the Central University (No. 2010QNA3032).

[2]Supported by the NSF Grants DMS-03-49048 and DMS-09-07297.

[3]Supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project No. CUHK400608).

*AMS 2000 subject classifications.* Primary 60F15, 62G10; secondary 60F05, 60F10.

*Key words and phrases.* Adaptive designs, asymptotic normality, clinical trial, urn model, branching process with immigration, birth and death urn, drop-the-loser rule.

a class of generalized Pólya urn models to scrutinize evolutionary processes. In economics, Beggs (2005) uses the models to capture the mechanism of reinforcement learning. In addition, numerous examples of applications of urn models in the areas of physics, communication theory and computer science are provided by Milenkovic and Compton (2004).

In statistics, an important application of urn models is to randomize treatments to patients in a clinical trial [Hu and Rosenberger (2006)]. Consider an urn containing balls of $K$ types, representing $K$ treatments. Patients normally arrive sequentially, and treatment assignment based on urn models is usually an adaptive scheme that depends on the urn composition and previous treatment outcomes. The urn composition is also continuously revised according to treatment outcomes.

Early studies of urn models in statistics include the generalized Pólya urn models (GPU) of Athreya and Karlin (1968), Wei and Durham (1978) and Wei (1979). Another renowned variation of the Pólya urn is the randomized Pólya urn (RPU) proposed by Durham, Flournoy and Li (1998). These classic urn models have a number of drawbacks. (i) They are usually proposed for binary (multinomial) responses. (ii) The urn process has a predetermined limit of urn proportions that does not have any connection with formal optimal properties [Hu and Rosenberger (2006)]. (iii) The urn process usually has higher variability than other types of procedures [Hu and Rosenberger (2003)] and is thus less powerful in statistical inferences. (iv) The formulation of the asymptotic variability is usually quite complex, and it is intricate to derive a reasonable estimate. For instance, the asymptotic variabilities of the Pólya urn-type models are related to the variance of a complicated Gaussian process. In particular, for the multi-treatment case, to derive the variability requires extremely complicated calculations of matrices [cf. Smythe (1996), Janson (2004), Bai and Hu (2005), Zhang, Hu and Cheung (2006), Higueras et al. (2006)]. (v) The models are designed mainly for the comparison of two treatments, so there is a shortage of methodology to handle cases with multiple treatments.

By embedding the urn process in a continuous-time birth and death process [Ivanova et al. (2000), Ivanova and Flournoy (2001), Ivanova (2006)], Ivanova (2003) formulates the drop-the-loser (DL) rule for a clinical trial with two treatments. The DL rule utilizes the idea of immigration and has been shown to yield a smaller variability among various urn models [Hu and Rosenberger (2003)]. The DL rule is generalized by Zhang et al. (2007) to provide more flexible urn models. However, these recent proposals are fragmented, offering only a partial solution to the aforementioned drawbacks of the classic urn models. To supply a complete resolution, we seek to provide a comprehensive paradigm through which one will be able to connect existing urn models, develop useful theoretic results, and compare merits of different classes of urn models.

1.2. *Objectives and organization of the paper.* In this paper, we propose the IMU framework that encompasses a wide spectrum of urn models and incorporates the immigration process, offering a greater flexibility in the choice of appropriate urn models in applications. This framework includes many urn models in the literature and provides a basis for us to derive several new urn models, together with their desirable properties. These new urn models are found to be capable of solving the aforementioned problems of classic urn models.

In the literature, the asymptotic properties of urn models are usually obtained by using Athreya and Ney's (1972) technique of embedding the urn process in a continuous-time branching process. However, this technique relies on the assumption that the transition of urn composition is governed by the adding rules, which are identical and nonrandom (homogeneous). This assumption is no longer valid for the IMU models in general due to the possibility that the urn composition may be generated by a nonhomogeneous immigration process. Hence, alternative mathematical approaches have to be utilized. Another major theoretical intricacy regarding the IMU process is that it depends on both the immigration rates and the adding rules (refer to Section 2.1 for details). To overcome these mathematical difficulties, we put forward a feasible solution. First, the IMU process is approximated by using martingales, which can handle both immigration rates and adding rules simultaneously; then, the IMU process is approximated by the Wiener process. Based on the Wiener process, we will be able to obtain the asymptotic properties of the IMU.

To summarize, the major contributions of this paper are as follows.

(a) It formulates a general framework of urn models (IMU models) that not only encompasses most existing urn models for adaptive designs in the literature, but also enables us to derive new urn models with desirable properties such as the freedom to design an urn process according to pre-specified optimality requirements.

(b) The paper derives asymptotic properties of the IMU models, including strong consistency and asymptotic normality of treatment allocation proportions. These asymptotic properties cover many existing asymptotic properties of urn models as special cases and form the basis for comparisons of different IMU models.

(c) The paper proposes and discusses several new IMU models that are useful in clinical trial applications.

The general IMU models and their asymptotic properties are provided in Section 2. In addition, several popular urn models that are members of the IMU class are discussed. In Section 3, new IMU models are developed and their applications are given. Concluding remarks are presented in Section 4. Finally, technical proofs are provided in the Appendix.

## 2. The immigrated urn model.

2.1. *The basic IMU framework.* In a clinical trial, suppose that subjects arrive sequentially to be randomized to one the $K$ available treatments, and responses are obtained immediately after treatment. An IMU model is defined as follows. Consider an urn that contains balls of $K+1$ types. Balls of types $1,\dots,K$ represent treatments, and balls of type 0 are the immigration balls. The urn allows negative and fractional number of balls.

Initially, there are $Z_{0,i}(\geq 0)$ balls of type $i$, $i=0,\dots,K$. Let $\mathbf{Z}_0=(Z_{0,0},\dots, Z_{0,K})$ be the initial urn composition. Immediately before the $m$th ($m>0$) subject arrives to be randomized to a treatment, let the urn composition be $\mathbf{Z}_{m-1}=(Z_{m-1,0},\dots,Z_{m-1,K})$. To avoid a negative likelihood of selecting a treatment, we adopt a slight adjustment to $\mathbf{Z}_{m-1}$ and let $Z^+_{m-1,i}=\max(0,Z_{m-1,i})$, $i=1,\dots,K$, and $\mathbf{Z}^+_{m-1}=(Z^+_{m-1,0},\dots,Z^+_{m-1,K})$.

To randomize the $m$th subject, a ball is drawn at random without replacement. The probability of selecting a ball of type $i$ is $Z^+_{m-1,i}/|\mathbf{Z}^+_{m-1}|$, $i=0, 1,\dots,K$. Here, $|\mathbf{Z}^+_{m-1}|=\sum_{j=0}^K Z^+_{m-1,j}$, and $\mathbf{Z}^+_{m-1}/|\mathbf{Z}^+_{m-1}|$ is defined to be $(0,1/K,\dots,1/K)$ if $|\mathbf{Z}^+_{m-1}|=0$. Hence, the balls with negative values in $\mathbf{Z}_{m-1}$ will have no chance of being selected unless all $Z^+_{m-1,k}$, $k=0,\dots,K$, are zeros, and when $|\mathbf{Z}^+_{m-1}|=0$ (only for the particular case where the IMU model has no immigration ball), a treatment ball is drawn with an equal probability of $1/K$. Now, consider the following two possibilities.

(a) If the selected ball is of type 0 (i.e., an immigration ball), no treatment is assigned and the ball is returned to the urn. $A_{m-1}=a_{m-1,1}+\cdots+a_{m-1,K}$ additional balls, $a_{m-1,k}(\geq 0)$ of treatment type $k$, $k=1,\dots,K$ are added to the urn. Then, a ball is drawn from this updated urn again until a treatment ball is drawn. If the immigration ball is selected $l$ times before a treatment ball is drawn, the urn composition $\mathbf{Z}_{m-1}$ is updated to $(Z_{m-1,0},Z_{m-1,1}+la_{m-1,1},\dots,Z_{m-1,K}+la_{m-1,K})$ and the $\mathbf{Z}^+_{m-1}$ is updated to $(Z_{m-1,0},(Z_{m-1,1}+la_{m-1,1})^+,\dots,(Z_{m-1,K}+la_{m-1,K})^+)$.

(b) If a treatment ball is drawn (say, of type $k$, $k=1,\dots,K$), the $m$th subject is given treatment $k$ and the treatment outcome (response) $\xi_{m,k}$ of this subject on treatment $k$ is observed. The ball is not replaced. Instead, $D_{m,kj}=D_{kj}(\xi_{m,k})$ balls of type $j$ are added to the urn, $j=1,\dots,K$. $D_{m,kj}<0$ signifies the removal of balls.

With the IMU, the number of immigration balls remains unchanged and a treatment ball is dropped when it is drawn. The number of treatment balls that is added to the urn depends on:

(a) the value of $a_{m,k}$ when an immigration ball is drawn from the urn and
(b) the value of $D_{m,kj}$ when a ball of treatment type $k$ is selected.

Here, $a_{m,k}$'s represent the immigration rates and $D_{m,kj}$'s represent the adding rules. Both $a_{m,k}$ and $D_{m,kj}$ allow fractional values which enable us to define a design in a flexible manner for application. The IMU models unify many existing urn models in the literature. Classic urn models, mainly designed for binary responses, are members of the IMU family. Here we list a few popular models.

(1) The randomized play-the-winner (RPW) rule [Wei and Durham (1978)]. When $K = 2$, $Z_{0,0} = 0$ or $a_{m,k} = 0$ for all $m$ and $k$. Further, $D_{m,kk} = 2$ if the response of the $m$th subject on treatment $k$ is a success and $D_{m,kk} = D_{m,kj} = 1$ ($j \neq k$) otherwise.

(2) Generalized Pólya urn models [Athreya and Karlin (1968), also called the generalized Friedman's urn]. When $a_{m,k} = 0$, we obtain the GPU models if one chooses the adding rule $D_{m,kj}$ as in Section 4.1 in Hu and Rosenberger (2006). If $D_{m,kj}$ is nonhomogeneous, we obtain the nonhomogeneous GPU models discussed by Bai and Hu (1999, 2005).

(3) The birth and death urn (BDU) [Ivanova et al. (2000)]. Suppose that $a_{m,k} \equiv 1$, $D_{m,kj} = 0$ for $j \neq k$. In addition, $D_{m,kk} = 2$ if the response of the $m$th subject on treatment $k$ is a success and $D_{m,kk} = 0$ otherwise. When $K = 2$, we obtain the birth and death urn (BDU) [Ivanova et al. (2000)]. When $K > 2$, we obtain generalized birth and death urn (BDU) for $K$ treatments.

(4) The drop-the-loser (DL) rule [Ivanova (2003)]. Suppose that $a_{m,k} \equiv 1$, $D_{m,kj} = 0$ for $j \neq k$. In addition, $D_{m,kk} = 1$ if the response of the $m$th subject on treatment $k$ is a success and $D_{m,kk} = 0$ otherwise. When $K = 2$, we obtain the DL rule [Ivanova (2003)]. When $K > 2$, we obtain DL rule for $K$ treatments.

(5) The generalized drop-the-loser (GDL) rule [Zhang et al. (2007)]. Suppose that $a_{m,k} = a_k$ (does not depend on $m$) are constants and $D_{m,kj} = 0$ for $j \neq k$. When $K = 2$, we obtain the GDL rule. When $K > 2$, we obtain GDL rules for $K$ treatments.

(6) Sequential estimated urn (SEU) models [Zhang, Hu and Cheung (2006)]. When $Z_{0,0} = 0$ or $a_{m,k} = 0$ for all $m$ and $k$, and $D_{m,kj}$ depends on estimation, we obtain the SEU models proposed by Zhang, Hu and Cheung (2006) and the urn models in Bai, Hu and Shen (2002).

In general, we can select suitable $a_{m,k}$ and $D_{m,kj}$ to obtain the desirable IMU model for both binary and continuous responses (see examples in Section 3).

In clinical trials, let $N_{n,k}$ be the number of subjects who have been assigned to treatment $k$, $k = 1, \dots, K$. Denote $\mathbf{N}_n = (N_{n,1}, \dots, N_{n,K})$. In clinical studies, the proportions $N_{n,k}/n$, $k = 1, \dots, K$, of patients being assigned to various treatments are useful statistics. In fact, for urn model applications, there are several important statistics, including:

(a) the urn proportion $Z_{n,k}/\sum_{k=1}^{K} Z_{n,k}$;
(b) the allocation proportion $N_{n,k}/n$ and
(c) the estimation of the unknown parameters in the model.

It is worthwhile noting that both $a_{m,k}$ and $D_{m,kj}$ depend on $m$. This allows both the immigration rates and the adding rules to be expressed as functions of all previous responses thus far in the clinical trial. Then, we are able to construct desirable IMU models that can be used to suit pre-specified allocation proportion targets. To reiterate, as both $a_{m,k}$ and $D_{m,kj}$ depend on $m$, it is impossible to use

Athreya and Ney's (1972) technique of embedding the urn process in a continuous-time branching process.

It is also worth noting that Hoppe's urn [Hoppe (1984)] and its extensions [see, e.g., Donnelly and Kurtz (1996)] are not members of the IMU models. For Hoppe's urn, the number of ball types is increasing and random, but for an IMU model the number of ball types is fixed ($K + 1$).

2.2. *Notation and assumptions.* Before the discussion of major asymptotic results regarding the IMU models, we introduce some basic notation and the necessary assumptions. Suppose that $\xi_{m,k}$ ($k = 1, \dots, K, m = 1, 2, 3, \dots$) is the random variable representing the response of the $m$th subject on treatment $k$. In practice, we only observe one $\xi_{m,k}$ for each $m$. Without loss of generality, we assume that the unknown parameter $\theta_k$ is the mean of the outcome $\xi_{m,k}$ and take the sample mean as its estimate. Write $\boldsymbol{\xi}_m = (\xi_{m,1}, \dots, \xi_{m,K})$. For the adding rules, let $\mathbf{D}_m = (D_{m,kj}; k, j = 1, \dots, K)$, $\mathbf{D}_m^{(k)} = (D_{m,k1}, \dots, D_{m,kK})$, $k = 1, \dots, K$, and $\mathbf{H}_m = (h_{kj}(m)) = \mathsf{E}\mathbf{D}_m$. Let $\widehat{\theta}_{m-1,k}$ be the sample mean of the responses

$$\widehat{\theta}_{m-1,k} = \frac{c_1 + S_{m-1,k}}{c_2 + N_{m-1,k}}, \tag{2.1}$$

where $S_{m-1,k}$ is the sum of the responses on treatment $k$ of all the previous $m-1$ subjects. Here, $c_1, c_2 > 0$ are used to avoid the nonsense case of $0/0$. These two constants play a minor role, only in the earlier stages of the clinical trial when accumulated observations of the treatments are still very small. In general, many estimators, such as the MLE, can be written in the form of (2.1) with $S_{m-1,k}$ being replaced by a sum of functions of the responses plus a negligible remainder [see Hu and Zhang (2004a) for a detailed discussion].

As discussed in Section 2.1, the immigration rate $a_{m,k}$ plays an important role in the IMU models. Its significance will be illustrated in the later part of this section when the theoretical properties of the IMU models are being reviewed. In clinical trials, optimal allocation proportions usually depend on the unknown parameters $\boldsymbol{\theta}$ [see Rosenberger et al. (2001) and Tymofyeyev, Rosenberger and Hu (2007)]. To achieve these proportions, one can select the immigration rates $a_{m,k}$ as functions of $\boldsymbol{\theta}$. In practice, as $\boldsymbol{\theta}$ is unknown, one can use $a_{m-1,k} = a_k(\widehat{\boldsymbol{\theta}}_{m-1})$ as the immigration rates. The guidelines for the selection of the function $a_k$ will be given in Section 3. In most applications, the adding rules $\mathbf{D}_m = (D_{m,kj}; k, j = 1, \dots, K)$ normally depend on the response $\boldsymbol{\xi}_m$, similar to those in the GPU models. Hence, we need the following assumptions.

ASSUMPTION 2.1. Functions $a_k(\cdot) > 0$ are continuous and twice differentiable at $\boldsymbol{\theta}$.

ASSUMPTION 2.2. $\{(\xi_{m,k}, D_{m,k1}, \dots, D_{m,kK}); m \geq 1\}$, $k = 1, \dots, K$, are $K$ sequences of i.i.d. random variables with

$$\sup_m \mathsf{E}|D_{m,kj}|^{2+\delta} < \infty \quad \text{and} \quad \sup_m \mathsf{E}|\xi_{m,k}|^{2+\delta} < \infty$$

for some $0 < \delta \leq 2$, $k = 1, \dots, K$. Hence, let $\mathbf{H}_m = \mathbf{H}$, which does not depend on $m$. Further assume that $D_{m,kk} \geq -C$ for some $C$, $k = 1, \dots, K$, and also $D_{m,kj} \geq 0$ for $k \neq j$.

The continuity of $a_k(\cdot)$ in Assumption 2.1 is needed to show that $0 < \min_{m,k} a_{m,k} \leq \max_{m,k} a_{m,k} < \infty$ as given in Lemma A.5. The differentiability of the function is required for the Taylor expansion. The moment condition in Assumption 2.2 is useful for applying the limit theorems and the approximation of related martingales. Finally, the lower bound of $D_{m,kj}$ implies that when a ball is drawn, the maximum number of balls of that treatment type which can be removed is $C + 1$. This condition is used to derive the lower bound of $Z_{n,k}$, as given in Lemma A.3.

2.3. *Main asymptotic results.* We now discuss the asymptotic properties related to urn proportions and model parameter estimators. Asymptotic results are classified into one of following three possible cases, according to the expectation of the adding rules.

1. $\mathbf{H1}' < \mathbf{1}'$ where $\mathbf{1} = (1, \dots, 1)$. Hence $\sum_{j=1}^K h_{kj} < 1$ for all $k = 1, \dots, K$. The urn composition is mainly updated by the immigration balls because, on average, the number of added balls in each step according to the outcome of a treatment is less than the number of dropped balls, which is 1. The derivation of asymptotic results for this case is of the utmost importance and plays a crucial role in this paper.
2. $\mathbf{H1}' > \mathbf{1}'$. The total number of balls in the urn gradually increases to infinity. Hence, the probability of drawing an immigration ball drops to zero. For this case, we will prove that the IMU model is asymptotically equivalent to the generalized Pólya urn model without immigration (refer to Theorem 2.1).
3. $\mathbf{H1}' = \mathbf{1}'$. This is the borderline case in which both the treatment balls and the immigration ball retain their roles in the urn updating process.

These three cases lead to very different asymptotic results. Let us first consider the case of $\mathbf{H1}' > \mathbf{1}'$. The following theorem ensures that the IMU model behaves asymptotically, the same as the generalized Pólya urn model, when $\mathbf{H1}' > \mathbf{1}'$. The proof is given in the Appendix. Based on this theorem, we can obtain the asymptotic properties, including the strong consistency, asymptotic normality and Gaussian approximation, of the generalized Pólya urn model as discussed by Janson (2004), Bai and Hu (2005), Zhang, Hu and Cheung (2006) and Zhang and Hu (2009), among others.

THEOREM 2.1. *Suppose that Assumption 2.2 is satisfied*, $\mathbf{H}\mathbf{1}' = \gamma\mathbf{1}'$ *with* $\gamma > 1$ *and* $0 \le a_{m,k} \le Cm^{1/2-\delta_0}$ *for some* $\delta_0 > 0$ *and all* $m, k$. *Let* $\mathbf{v} = (v_1, \dots, v_K)$ *be the left eigenvalue vector of* $\mathbf{H}$ *that corresponds to the largest eigenvalue* $\gamma$ *and satisfies* $v_1 + \cdots + v_K = 1$, *and denote* $\widetilde{\mathbf{H}} = \frac{\mathbf{H}-\mathbf{I}}{\gamma-1} - \mathbf{1}'\mathbf{v}$. *Further, let* $\lambda_2, \dots, \lambda_K$ *be the other* $K-1$ *eigenvalues of* $\mathbf{H}$ *and* $\lambda = \max\{\mathrm{Re}(\lambda_2), \dots, \mathrm{Re}(\lambda_K)\}$. *Assume that* $\lambda - 1 < (\gamma-1)/2$. *Then, there exist two independent standard* $K$*-dimensional Wiener processes* $\mathbf{B}_{t1}$ *and* $\mathbf{B}_{t2}$ *such that*

$$(N_{n,1}, \dots, N_{n,K}) - n\mathbf{v} = \mathbf{G}_{n1} + \frac{1}{\gamma-1}\int_0^t \frac{\mathbf{G}_{x2}}{x}\,dx(\mathbf{I} - \mathbf{1}'\mathbf{v}) \\ + o(n^{1/2-\varepsilon}) \qquad a.s.,$$

$$(Z_{n,1}, \dots, Z_{n,K}) - (\gamma-1)n\mathbf{v} = (\gamma-1)\mathbf{G}_{n1}\widetilde{\mathbf{H}} + \mathbf{G}_{n2} + o(n^{1/2-\varepsilon}) \qquad a.s.$$

*for some* $\varepsilon > 0$, *where* $\mathbf{G}_{ti}$ *is the solution of the equation*

$$\mathbf{G}_{ti} = \mathbf{B}_{ti}\mathbf{\Lambda}_i^{1/2} + \int_0^t \frac{\mathbf{G}_{xi}}{x}\,dx\,\widetilde{\mathbf{H}}$$

*with* $\mathbf{\Lambda}_1 = \mathrm{diag}(\mathbf{v}) - \mathbf{v}'\mathbf{v}$ *and* $\mathbf{\Lambda}_2 = \sum_{k=1}^K v_k \mathsf{Var}\{\mathbf{D}_1^{(k)}\}$. *In particular*,

$$\frac{Z_{n,0}}{Z_{n,0} + \cdots + Z_{n,K}} \to 0 \qquad a.s., \qquad \frac{Z_{n,k}}{Z_{n,0} + \cdots + Z_{n,K}} \to v_k \qquad a.s.,$$

$$\frac{N_{n,k}}{n} \to v_k \qquad a.s.,$$

$k = 1, \dots, K$, *and*

$$n^{1/2}\left(\frac{Z_{n,1}}{(\gamma-1)n} - v_1, \dots, \frac{Z_{n,K}}{(\gamma-1)n} - v_K\right) \xrightarrow{\mathscr{D}} N(\mathbf{0}, \mathbf{\Gamma}_1),$$

$$n^{1/2}\left(\frac{N_{n,1}}{n} - v_1, \dots, \frac{N_{n,K}}{n} - v_K\right) \xrightarrow{\mathscr{D}} N(\mathbf{0}, \mathbf{\Gamma}_2).$$

*Here, the variance–covariance matrices* $\mathbf{\Gamma}_1$ *and* $\mathbf{\Gamma}_2$ *can be specified in line with Bai and Hu* (2005) *and Zhang and Hu* (2009) *with* $\frac{\mathbf{D}_m-\mathbf{I}}{\gamma-1}$ *and* $\frac{\mathbf{H}-\mathbf{I}}{\gamma-1}$ *replacing* $\mathbf{D}_m$ *and* $\mathbf{H}$, *respectively. For details, one can refer to Proposition* 3.4 *of Zhang and Hu* (2009).

Now we consider the case in which $\mathbf{H}\mathbf{1}' < \mathbf{1}'$. Different from the case when $\mathbf{H}\mathbf{1}' > \mathbf{1}'$ in which the urn proportion and the sample allocation proportion have the same limit, the urn proportion may not have a limit in this case. For the immigration rates, write $a_k = a_k(\boldsymbol{\theta})$. Let $\mathbf{a} = (a_1, \dots, a_K)$, $\mathbf{u} = \mathbf{a}(\mathbf{I} - \mathbf{H})^{-1}$, $s = \mathbf{a}(\mathbf{I} - \mathbf{H})^{-1}\mathbf{1}' = \sum_{k=1}^K u_k$ and $\mathbf{v} = \mathbf{u}/s$. Further, denote $\mathbf{\Sigma}_k = \mathsf{Var}\{\mathbf{D}_1^{(k)}\}$, $\mathbf{\Sigma}_{11} = \sum_{k=1}^K v_k \mathbf{\Sigma}_k$, $\mathbf{\Sigma}_{12} = (\mathsf{Cov}\{D_{1,kj}, \xi_k\}; j, k = 1, \dots, K)$, $\mathbf{\Sigma}_{22} = \mathrm{diag}(\mathsf{Var}\{\xi_{1,1}\}, \dots, \mathsf{Var}\{\xi_{1,K}\})$ and

$$\mathbf{\Lambda} = \begin{pmatrix} \mathbf{\Lambda}_{11} & \mathbf{\Lambda}_{12} \\ \mathbf{\Lambda}_{12}' & \mathbf{\Lambda}_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12}\,\mathrm{diag}(\mathbf{v}) \\ \mathrm{diag}(\mathbf{v})\mathbf{\Sigma}_{12}' & \mathbf{\Sigma}_{22}\,\mathrm{diag}(\mathbf{v}) \end{pmatrix}. \tag{2.2}$$

THEOREM 2.2. *Suppose that Assumptions 2.1 and 2.2 are satisfied*, $\mathbf{H1}' < \mathbf{1}'$ *and* $Z_{0,0} > 0$. *Then*

$$Z_{n,k} = o(n^{1/2-\varepsilon}) \qquad a.s., \qquad k = 1, \dots, K,$$

*for some* $\varepsilon > 0$, *and, one can define a* $2K$*-dimensional Wiener processes* $(\mathbf{W}(t), \mathbf{B}(t))$ *such that*

$$\mathsf{Var}\{(\mathbf{W}(t), \mathbf{B}(t))\} = t\boldsymbol{\Lambda} \tag{2.3}$$

*and*

$$\mathbf{N}_n - n\mathbf{v} = \mathbf{W}(n)\mathbf{A} + \int_0^n \frac{\mathbf{B}(x)}{x}\,dx\,\mathrm{diag}\left(\frac{1}{\mathbf{v}}\right)\frac{\partial \mathbf{v}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + o(n^{1/2-\varepsilon}) \qquad a.s. \tag{2.4}$$

*for some* $\varepsilon > 0$, *where* $\mathbf{A} = (\mathbf{I} - \mathbf{H})^{-1}(\mathbf{I} - \mathbf{1}'\mathbf{v})$,

$$\mathbf{v} = \mathbf{v}(\boldsymbol{\theta}) = \frac{a(\boldsymbol{\theta})(\mathbf{I} - \mathbf{H})^{-1}}{a(\boldsymbol{\theta})(\mathbf{I} - \mathbf{H})^{-1}\mathbf{1}'} \quad and \quad \frac{\partial \mathbf{v}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial v_k(\boldsymbol{\theta})}{\partial \theta_j}; j, k = 1, \dots, K\right).$$

*Here*, $1/\mathbf{v} = (1/v_1, \dots, 1/v_K)$.

REMARK 2.1. Note that $h_{ij} \geq 0$ for $i \neq j$. The existence of $(\mathbf{I} - \mathbf{H})^{-1}$ is implied by the assumption that $\mathbf{H1}' < \mathbf{1}'$. This assumption can be replaced by a more general assumption in which there is a vector $\mathbf{e} = (e_1, \dots, e_K)$ such that $\mathbf{He}' < \mathbf{e}'$ and $e_i > 0$, $i = 1, \dots, K$.

Based on Theorem 2.2, we can see that the urn composition $(\sqrt{n})^{-1}Z_{n,k}$ converges to 0 almost surely. It is because when $\mathbf{H1}' < \mathbf{1}'$, there will be a net loss of balls from the urn on average if a treatment ball is drawn. The proof of Theorem 2.2 is given in the Appendix. The consistency and asymptotic normality of $\mathbf{N}_n$ can be derived by using (2.4) as follows.

COROLLARY 2.1. *Under the assumptions in Theorem 2.2*,

$$\mathbf{N}_n - n\mathbf{v} = O\big(\sqrt{n \log\log n}\big) \qquad a.s. \quad and \quad \sqrt{n}\left(\frac{\mathbf{N}_n}{n} - \mathbf{v}\right) \xrightarrow{\mathscr{D}} N(\mathbf{0}, \boldsymbol{\Sigma}), \tag{2.5}$$

*where* $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_D + 2\boldsymbol{\Sigma}_\xi + \boldsymbol{\Sigma}_{D\xi} + \boldsymbol{\Sigma}'_{D\xi}$ *and*

$$\boldsymbol{\Sigma}_D = \mathbf{A}'\boldsymbol{\Sigma}_{11}\mathbf{A}, \qquad \boldsymbol{\Sigma}_{D\xi} = \mathbf{A}'\boldsymbol{\Sigma}_{12}\frac{\partial \mathbf{v}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}},$$
$$\boldsymbol{\Sigma}_\xi = \left(\frac{\partial \mathbf{v}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right)' \mathrm{diag}\left(\frac{\mathsf{Var}\{\xi_{1,1}\}}{v_1}, \dots, \frac{\mathsf{Var}\{\xi_{1,K}\}}{v_K}\right)\frac{\partial \mathbf{v}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}.$$

*In particular, if* $\mathbf{D}_m \equiv \mathrm{const}$, *then*

$$\sqrt{n}\left(\frac{\mathbf{N}_n}{n} - \mathbf{v}\right) \xrightarrow{\mathscr{D}} N(\mathbf{0}, 2\boldsymbol{\Sigma}_\xi)$$

*and if* $a_{m,k} \equiv a_k$, $k = 1, \dots, K$, *do not depend on the estimates*, *then*

$$\sqrt{n}\left(\frac{\mathbf{N}_n}{n} - \mathbf{v}\right) \xrightarrow{\mathscr{D}} N(\mathbf{0}, \mathbf{\Sigma}_D).$$

PROOF. Note that $(\mathbf{W}(n), \int_0^n \frac{\mathbf{B}(x)}{x}\, dx)$ is a centered Gaussian vector with

$$\begin{aligned}
\mathbf{W}(n) &= O\left(\sqrt{n \log\log n}\right) \qquad \text{a.s.},\\
\int_0^n \frac{\mathbf{B}(x)}{x}\, dx &= O(1) + \int_e^n \frac{O(\sqrt{x \log\log x})}{x}\, dx\\
&= O\left(\sqrt{n \log\log n}\right) \qquad \text{a.s.},\\
\mathsf{Var}\{\mathbf{W}(n)\} &= n\mathbf{\Sigma}_{11},\\
\mathsf{Var}\left\{\int_0^n \frac{\mathbf{B}(x)}{x}\, dx\right\} &= \mathbf{\Sigma}_{22}\,\mathrm{diag}(\mathbf{v}) \int_0^n \int_0^n \frac{x \wedge y}{xy}\, dx\, dy = 2n\mathbf{\Sigma}_{22}\,\mathrm{diag}(\mathbf{v})
\end{aligned}$$

and

$$\mathsf{Cov}\left\{\mathbf{W}(n), \int_0^n \frac{\mathbf{B}(x)}{x}\, dx\right\} = \mathbf{\Sigma}_{12}\,\mathrm{diag}(\mathbf{v}) \int_0^n \frac{x \wedge n}{x}\, dx = n\mathbf{\Sigma}_{12}\,\mathrm{diag}(\mathbf{v}).$$

Equation (2.5) follows from (2.4) immediately. □

REMARK 2.2. In practice, the responses in clinical trials are frequently not available immediately before the treatment allocation of the next subject (delayed response). The parameters can be estimated and the urn can be updated only by using all available observed responses. In the delayed response situation, let $\mu_k(m, l)$ be the probability that the response of the $m$th subject on treatment $k$ occurs after at least another $l$ subjects arrive. If $\mu_k(m, l) \le Cl^{-\gamma}$ for some $\gamma \ge 2$, then we can show that the total sum of unobserved outcomes up to the $n$th assignment is with a high order of $\sqrt{n}$ and thus the conclusion in Theorem 2.2 remains true. It has been shown that the delay mechanism does not effect the asymptotic properties for many response-adaptive designs if the delay decays with a power rate [cf. Bai, Hu and Rosenberger (2002), Hu and Zhang (2004b), Zhang et al. (2007)].

In many IMU models [such as special cases (3), (4) and (5) in Section 2], the additional rule, $\mathbf{D}_m$, is a diagonal matrix ($D_{m,kj} = 0$, $j \ne k$). For this special case, we have the following corollary that helps us to obtain the asymptotical limits and covariance matrix of $\mathbf{N}_n$ easily.

COROLLARY 2.2. *Suppose that Assumptions* 2.1 *and* 2.2 *are satisfied*, $D_{m,kj} = 0$ *for* $j \ne k$ *and* $h_k = 1 - \mathsf{E}D_{1,kk} > 0$. *Write* $\mathbf{h} = (h_1, \dots, h_K)$,

$$v_k(\boldsymbol{\theta}, \mathbf{h}) = \frac{a_k(\boldsymbol{\theta})/h_k}{\sum_{j=1}^K a_j(\boldsymbol{\theta})/h_j}, \qquad k = 1, \dots, K,$$

$\mathbf{v} = \mathbf{v}(\boldsymbol{\theta}, \mathbf{h}) = (v_1(\boldsymbol{\theta}, \mathbf{h}), \dots, v_K(\boldsymbol{\theta}, \mathbf{h}))$ *and*

$$\frac{\partial \mathbf{v}(\boldsymbol{\theta}, \mathbf{h})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial v_k(\boldsymbol{\theta}, \mathbf{h})}{\partial \theta_j}; j, k = 1, \dots, K\right),$$
$$\frac{\partial \mathbf{v}(\boldsymbol{\theta}, \mathbf{h})}{\partial \mathbf{h}} = \left(\frac{\partial v_k(\boldsymbol{\theta}, \mathbf{h})}{\partial h_j}; j, k = 1, \dots, K\right).$$

*Then*,

$$\frac{\mathbf{N}_n}{n} \to \mathbf{v} \quad a.s. \quad and \quad \sqrt{n}\left(\frac{\mathbf{N}_n}{n} - \mathbf{v}\right) \xrightarrow{\mathscr{D}} N(\mathbf{0}, \boldsymbol{\Sigma}), \tag{2.6}$$

*where* $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_D + 2\boldsymbol{\Sigma}_\xi + \boldsymbol{\Sigma}_{D\xi} + \boldsymbol{\Sigma}'_{D\xi}$,

$$\boldsymbol{\Sigma}_D = \left(\frac{\partial \mathbf{v}(\boldsymbol{\theta}, \mathbf{h})}{\partial \mathbf{h}}\right)' \operatorname{diag}\left(\frac{\sigma^2_{D1}}{v_1}, \dots, \frac{\sigma^2_{DK}}{v_K}\right) \frac{\partial \mathbf{v}(\boldsymbol{\theta}, \mathbf{h})}{\partial \mathbf{h}},$$
$$\boldsymbol{\Sigma}_\xi = \left(\frac{\partial \mathbf{v}(\boldsymbol{\theta}, \mathbf{h})}{\partial \boldsymbol{\theta}}\right)' \operatorname{diag}\left(\frac{\sigma^2_{\xi 1}}{v_1}, \dots, \frac{\sigma^2_{\xi K}}{v_K}\right) \frac{\partial \mathbf{v}(\boldsymbol{\theta}, \mathbf{h})}{\partial \boldsymbol{\theta}},$$
$$\boldsymbol{\Sigma}_{D\xi} = -\left(\frac{\partial \mathbf{v}(\boldsymbol{\theta}, \mathbf{h})}{\partial \mathbf{h}}\right)' \operatorname{diag}\left(\frac{\sigma_{D\xi 1}}{v_1}, \dots, \frac{\sigma_{D\xi K}}{v_K}\right) \frac{\partial \mathbf{v}(\boldsymbol{\theta}, \mathbf{h})}{\partial \boldsymbol{\theta}}$$

*and* $\sigma^2_{Dk} = \mathsf{Var}\{D_{1,kk}\}$, $\sigma^2_{\xi k} = \mathsf{Var}\{\xi_{1,k}\}$, $\sigma_{\xi Dk} = \mathsf{Cov}\{D_{1,kk}, \xi_{k,1}\}$, $k = 1, 2, \dots, K$.

PROOF. It is easy to check that

$$\boldsymbol{\Sigma}_{11} = \operatorname{diag}(\sigma^2_{D1} v_1, \dots, \sigma^2_{DK} v_K), \qquad \boldsymbol{\Sigma}_{12} = \operatorname{diag}(\sigma_{\xi D1}, \dots, \sigma_{\xi DK}),$$
$$\boldsymbol{\Sigma}_{22} = \operatorname{diag}(\sigma^2_{\xi 1}, \dots, \sigma^2_{\xi K}), \qquad \mathbf{A} = \operatorname{diag}(1/\mathbf{h})(\mathbf{I} - \mathbf{1}'\mathbf{v})$$

and $\partial \mathbf{v}(\boldsymbol{\theta}, \mathbf{h})/\partial \mathbf{h} = -\operatorname{diag}(\mathbf{v})\mathbf{A}$. Then, the results follow from Corollary 2.1 directly. $\square$

To improve statistical efficiency, a suitable response adaptive randomization procedure should be adopted because of variability [Hu and Rosenberger (2003)]. Hu, Rosenberger and Zhang (2006) studied the variability of a randomization procedure that targets any given allocation proportion. They obtained a lower bound of the variability. For a large class of the IMU models in this paper, the lower bound of the variability is attained. When the variance of IMU model attains the lower bound, we can use the Cramér–Rao formula to compute the variance. In general, we have the following theorem.

THEOREM 2.3. *If each* $D_{m,kj}$ *is a linear function of a random* $\eta_{m,k}$, $j = 1, \dots, K$, *where* $\eta_{m,k}$ *may be a function of* $\xi_{m,k}$ *and for each* $k$, $\eta_{m,k}$, $m = 1, 2, \dots,$

*are i.i.d. random variables with finite variances, then we have*

$$\boldsymbol{\Sigma}_D = \left(\frac{\partial \mathbf{v}}{\partial \mathbf{d}}\right)' \operatorname{diag}\left(\frac{\mathsf{Var}\{\eta_{1,1}\}}{v_1}, \dots, \frac{\mathsf{Var}\{\eta_{1,K}\}}{v_K}\right) \frac{\partial \mathbf{v}}{\partial \mathbf{d}}, \tag{2.7}$$

$$\boldsymbol{\Sigma}_{D\xi} = \left(\frac{\partial \mathbf{v}}{\partial \mathbf{d}}\right)' \operatorname{diag}\left(\frac{\mathsf{Cov}\{\eta_{1,1}, \xi_{1,1}\}}{v_1}, \dots, \frac{\mathsf{Cov}\{\eta_{1,K}, \xi_{1,K}\}}{v_K}\right) \frac{\partial \mathbf{v}}{\partial \boldsymbol{\theta}}, \tag{2.8}$$

*where* $\mathbf{d} = (d_1, \dots, d_K) = (\mathsf{E}\eta_{1,1}, \dots, \mathsf{E}\eta_{1,K})$. *Further, if* $\mathbf{a}(\cdot) = \text{const}$ *and* $\mathsf{Var}\{\eta_{1,k}\}$ *is the inverse of the Fisher information of* $d_k$, *then the asymptotic variance–covariance matrix of* $\mathbf{N}_n/\sqrt{n}$ *attains the following lower bound*:

$$\left(\frac{\partial \mathbf{v}}{\partial \mathbf{d}}\right)' \operatorname{diag}((v_1 I_1)^{-1}, \dots, (v_K I_K)^{-1}) \left(\frac{\partial \mathbf{v}}{\partial \mathbf{d}}\right), \tag{2.9}$$

*where* $I_k$ *is the Fisher information function of parameter* $d_k$.

PROOF. If we write $\mathbf{D}_1^{(k)} = \boldsymbol{\alpha}_k + \boldsymbol{\beta}_k \eta_{1,k}$ and $\mathbf{K} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_K \end{pmatrix}$, then

$$\begin{aligned} \boldsymbol{\Lambda}_{11} &= \sum_{k=1}^{K} v_k \mathsf{Var}\{\eta_{1,k}\} \boldsymbol{\beta}_k' \boldsymbol{\beta}_k \\ &= (\operatorname{diag}(\mathbf{v})\mathbf{K})' \operatorname{diag}\left(\frac{\mathsf{Var}\{\eta_{1,1}\}}{v_1}, \dots, \frac{\mathsf{Var}\{\eta_{1,K}\}}{v_K}\right) \operatorname{diag}(\mathbf{v})\mathbf{K} \end{aligned}$$

and

$$\boldsymbol{\Sigma}_{12} = (\operatorname{diag}(\mathbf{v})\mathbf{K})' \operatorname{diag}\left(\frac{\mathsf{Cov}\{\eta_{1,1}, \xi_{1,1}\}}{v_1}, \dots, \frac{\mathsf{Cov}\{\eta_{1,K}, \xi_{1,K}\}}{v_K}\right).$$

However, $\partial \mathbf{H}/\partial d_k = \operatorname{diag}(\mathbf{1}_k)\mathbf{K}$, where $\mathbf{1}_k$ has zero elements except the $k$th one which is 1. In addition,

$$\frac{\partial (\mathbf{I} - \mathbf{H})^{-1}}{\partial d_k} = (\mathbf{I} - \mathbf{H})^{-1} \frac{\partial \mathbf{H}}{\partial d_k} (\mathbf{I} - \mathbf{H})^{-1} = (\mathbf{I} - \mathbf{H})^{-1} \operatorname{diag}(\mathbf{1}_k)\mathbf{K}(\mathbf{I} - \mathbf{H})^{-1}.$$

It follows that

$$\begin{aligned} \frac{\partial \mathbf{v}}{\partial d_k} &= \frac{\partial \mathbf{a}(\mathbf{I} - \mathbf{H})^{-1}/\partial d_k}{\mathbf{a}(\mathbf{I} - \mathbf{H})^{-1}\mathbf{1}'} - \frac{\partial \mathbf{a}(\mathbf{I} - \mathbf{H})^{-1}/\partial d_k}{(\mathbf{a}(\mathbf{I} - \mathbf{H})^{-1}\mathbf{1}')^2} \mathbf{1}'\mathbf{a}(\mathbf{I} - \mathbf{H})^{-1} \\ &= \mathbf{v} \operatorname{diag}(\mathbf{1}_k)\mathbf{K}(\mathbf{I} - \mathbf{H})^{-1}(\mathbf{I} - \mathbf{1}'\mathbf{v}) = \mathbf{v} \operatorname{diag}(\mathbf{1}_k)\mathbf{K}\mathbf{A}, \end{aligned}$$

that is, $\partial \mathbf{v}/\partial \mathbf{d} = \operatorname{diag}(\mathbf{v})\mathbf{K}\mathbf{A}$. Hence, (2.7) and (2.8) are proved by Corollary 2.1. $\square$

Corollary 2.2 and Theorem 2.3 are useful for deriving the asymptotic variance. We will illustrate this idea by introducing several interesting examples in the next section.

REMARK 2.3. In Theorem 2.3, for simplicity of notation we assume that the parameter $d_k$ is a one-dimensional parameter that corresponds to treatment $k$. The theorem is still valid if reformulated using a vector parameter $d_k$, without extra assumptions.

Finally, we consider the case when $\mathbf{H1}' = \mathbf{1}'$. The following theorem, with proof given in the Appendix, can be used to yield the consistency property of the allocation proportion. However, it is still unknown whether $\mathbf{N}_n$ is asymptotically normal.

THEOREM 2.4. *Suppose that Assumptions 2.1 and 2.2 are satisfied, and* $\mathbf{H1}' = \mathbf{1}'$, $Z_{0,0} > 0$. *Suppose further that* 1 *is a single eigenvalue of* $\mathbf{H}$. *Then*

$$\mathbf{N}_n - n\mathbf{v} = O\big(\sqrt{n \log\log n}\big) \qquad a.s. \quad and \quad \mathbf{N}_n - n\mathbf{v} = O_P\big(\sqrt{n}\big),$$

*where* $\mathbf{v}$ *is the left eigenvalue vector of* $\mathbf{H}$ *that corresponds to the eigenvalue* 1 *and satisfies* $v_1 + \cdots + v_K = 1$.

These theorems and corollaries are related to the sample allocation proportion $\mathbf{N}_n/n$. Regarding the estimator $\widehat{\boldsymbol{\theta}}_n$, we have the following theorem.

THEOREM 2.5. *Suppose that the assumptions in Theorems 2.1 or 2.2 or 2.4 are satisfied. We have*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \to N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{2.10}$$

*where*

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}} = \operatorname{diag}\left(\frac{\mathsf{Var}\{\xi_{1,1}\}}{v_1}, \ldots, \frac{\mathsf{Var}\{\xi_{1,K}\}}{v_K}\right).$$

Note that $\mathbf{N}_n/n \to \mathbf{v}$ a.s. according to Theorems 2.1 or 2.2 or 2.4, so the proof of this theorem is the same as that of Lemma 1 of Hu, Rosenberger and Zhang (2006) and is thus omitted here.

**3. Examples and applications.** In this section, we apply the general asymptotic results in Section 3 to selected IMU models for illustrative purposes. In Section 2.1, we listed several classic families of urn models as special cases of IMU models. We can apply directly the theoretical results in Section 3 to these special cases and obtain their asymptotic properties for both $K = 2$ (available in the literature) and for the general value of $K \geq 3$. In this section, we focus on the generation of new families of urn models from the IMU framework and discuss their corresponding properties. Several illustrative examples are given. First, we consider continuous-type responses that are frequently encountered in clinical studies, even though there has been a lack of related studies in the literature.

EXAMPLE 1 (Two treatments with continuous responses). Suppose that $\xi_{m,1}$ ($m = 1, 2, 3, \ldots$) are i.i.d. random variables from $N(\mu_1, \sigma_1^2)$, and $\xi_{m,2}$ ($m = 1, 2, 3, \ldots$) are i.i.d. random variables from $N(\mu_2, \sigma_2^2)$. Without the loss of generality, assume that the smaller the value of the response, the better the treatment. We now introduce four IMU models.

(1.A) Let $a_{m,k} \equiv 1$, $D_{m,kj} = 0$ for $j \neq k$. Let $C$ be a constant such that $D_{m,kk} = 1$ if the response of the $m$th subject on treatment $k$, $\xi_{m,k}$, is less than $C$ and $D_{m,kk} = 0$ otherwise.

(1.B) Suppose that there are two critical values $C_1 < C_2$ and if it is very desirable to have the value of the response fall between $C_1$ and $C_2$, then the following IMU model is appropriate. Take $a_{m,k} \equiv 1$, $D_{m,kj} = 0$ for $j \neq k$. Further, let $D_{m,kk} = 1$ if $\xi_{m,k} < C_1$, $D_{m,kk} = 0$ if $\xi_{m,k} > C_2$ and else $D_{m,kk} = 1/2$.

(1.C) If the power of statistical inferences is an important concern, the Neyman allocation $\sigma_1/(\sigma_1 + \sigma_2)$ can be adopted to maximize the power of testing. Then, consider the following IMU model. Let $a_{m,k} = \widehat{\sigma}_k$, $D_{m,kj} = 0$ for all $j, k$. Here, $\widehat{\sigma}_k^2$ is the current sample variance of the responses on treatment $k$, $k = 1, 2$, and can be used as estimates in the Neyman allocation rule.

(1.D) If the aim is to lower the proportion of subjects being assigned to the inferior treatments for ethical reasons, the allocation target $\sqrt{\mu_2}\sigma_1/(\sqrt{\mu_2}\sigma_1 + \sqrt{\mu_1}\sigma_2)$ where $\mu_1, \mu_2 > 0$ [Zhang and Rosenberger (2006)] is an option. Let $a_{m,1} = \sqrt{\widehat{\mu}_2}\widehat{\sigma}_1$, $a_{m,2} = \sqrt{\widehat{\mu}_1}\widehat{\sigma}_2$, $D_{m,kj} = 0$ for all $j, k$. Here, $\widehat{\mu}_k$, $\widehat{\sigma}_k^2$ are the current sample mean and sample variance of the responses on treatment $k$, respectively, $k = 1, 2$. To avoid the situation of $\widehat{\mu}_k \leq 0$, simply replace $\widehat{\mu}_2$ by $1/m$ when such an occasion arises.

Designs (1.A) and (1.B) cover a wide spectrum of potential applications. Note that design (1.A) is equivalent to the DL rule for binary response if the critical value $C$ is used to classify responses into two categories. Designs (1.C) and (1.D) incorporate pre-specified objectives of a clinical trial, depending on whether the objective is to increase the testing power [as in (1.C)], or reduce the number of patients being assigned to the inferior treatments [as in (1.D)]. Further, it would not be difficult to generalize these four designs to studies with $K > 2$ treatments.

The asymptotic properties of the four designs can be obtained using Theorem 2.2. For illustrative purposes, we discuss asymptotic normalities for design (1.C). It is easy to verify that

$$\begin{aligned}
\widehat{\sigma}_k^2 =: \widehat{\sigma}_{m,k}^2 &= \frac{1}{N_{m,k}} \sum_{j=1}^{m} X_{j,k}(\xi_{j,k} - \mu_k)^2 - (\widehat{\mu}_k - \mu_k)^2 \\
&= \frac{1}{N_{m,k}} \sum_{j=1}^{m} X_{j,k}(\xi_{j,k} - \mu_k)^2 + O\left(\frac{\log\log N_{m,k}}{N_{m,k}}\right) \qquad \text{a.s.}
\end{aligned}$$

By Corollary 2.1,

$$\frac{N_{n,1}}{n} \to v_1 \qquad \text{a.s.} \quad \text{and} \quad n^{1/2}\left(\frac{N_{n,1}}{n} - v_1\right) \xrightarrow{\mathscr{D}} N(0, \sigma^2),$$

where $v_1 = \sigma_1/(\sigma_1 + \sigma_2)$ and $\sigma^2$ equals to

$$2\left(\frac{\partial v_1}{\partial(\sigma_1^2)}, \frac{\partial v_1}{\partial(\sigma_2^2)}\right) \operatorname{diag}\left(\frac{\mathsf{Var}\{(\xi_{1,1} - \mu_1)^2\}}{v_1}, \frac{\mathsf{Var}\{(\xi_{1,2} - \mu_2)^2\}}{1 - v_1}\right)\left(\frac{\partial v_1}{\partial(\sigma_1^2)}, \frac{\partial v_1}{\partial(\sigma_2^2)}\right)'.$$

After simplification, we have $\sigma^2 = \sigma_1\sigma_2/(\sigma_1 + \sigma_2)^2$. One can also use Theorem 2.5 to derive the asymptotic distribution of the estimators of the unknown parameters. For example, in design (1.C), $\sqrt{n}(\widehat{\sigma}_{n,k}^2 - \sigma^2) \xrightarrow{\mathscr{D}} N(0, 2\sigma_k^4/v_k)$.

EXAMPLE 2 [Modified DL (MDL) rule]. We propose the MDL rule, which is a modification of the DL rule. The procedure is similar to the DL rule in that when a treatment ball is drawn, this ball is replaced only when the response is a success. However, when an immigration ball is drawn, instead of adding an equal number of treatment balls to the urn, we add $C\widehat{p}_k$ ($C > 0$) balls of type $k$, $k = 1, \dots, K$, where $\widehat{p}_k$ is the current estimate of the successful probability $p_k$ of treatment $k$, and $C$ is a constant. With this model, more balls are immigrated to treatments with higher success rates, and subsequently, the limit proportions will be higher for better treatments.

Regarding the asymptotic variance, it is straightforward to show that $\mathbf{a} = (p_1C, \dots, p_KC)$ and $\mathbf{H} = \operatorname{diag}(p_1, \dots, p_K)$. The conditions in Corollary 2.2 are satisfied for all cases with $0 < p_k < 1$ and $k = 1, \dots, K$. Hence, the limit proportions are $v_k = (p_k/q_k)/(\sum_{j=1}^K p_j/q_j), k = 1, \dots, K$. The asymptotic variance–covariance can be derived by the formulae in Corollary 2.2, in which $\boldsymbol{\theta} = (p_1, \dots, p_K)$, $\mathbf{h} = (q_1, \dots, q_K)$ and $\sigma_{Dk}^2 = \sigma_{\xi k}^2 = \sigma_{D\xi k} = p_kq_k, k = 1, \dots, K$. For the two-treatment case,

$$\frac{N_{n,1}}{n} \to v_1 = \frac{p_1/q_1}{p_1/q_1 + p_2/q_2} \qquad \text{a.s.} \quad \text{and} \quad \sqrt{n}(N_{n,1}/n - v_1) \xrightarrow{\mathscr{D}} N(0, \sigma^2),$$

where $\sigma^2 = q_1q_2[p_1^2(1 + q_2^2) + p_2^2(1 + q_1^2)]/(p_2q_1 + p_1q_2)^3$. When the success probabilities $p_1$ and $p_2$ are both high, the variability $\sigma^2$ is close to the lower bound $q_1q_2(p_1^2 + p_2^2)/(p_2q_1 + p_1q_2)^3$.

Unlike the generalized Pólya urn models without immigration in which the asymptotic normality holds only when a very strict condition on eigenvalues of a generating matrix is satisfied [cf. Bai and Hu (2005), Janson (2004), Zhang, Hu and Cheung (2006)], the MDL rule allows asymptotic normality for all cases with $0 < p_k < 1, k = 1, \dots, K$.

In most IMU models, the adding rule $\mathbf{D}_m$ is a diagonal matrix. Here we give an example for the two-treatment case with dichotomous responses in which the adding rule $\mathbf{D}_m$ is not a diagonal matrix.

EXAMPLE 3 (Two treatments with dichotomous responses). Consider the two-treatment case with dichotomous responses, success or failure. Let $p_k$ be the success probability of treatment $k$ and $q_k = 1 - p_k$, $k = 1, 2$. We consider an immigrated urn in which $a_{m,1} = a_{m,2} \equiv 1$ and

$$\mathbf{D}_m = \begin{pmatrix} \beta\xi_{m,1} & \alpha(1 - \xi_{m,1}) \\ \alpha(1 - \xi_{m,2}) & \beta\xi_{m,2} \end{pmatrix},$$

where $\xi_{m,k} = 1$ if the outcome of the $m$th subject on treatment $k$ is a success, and 0 otherwise, $k = 1, 2$, $\alpha \geq 0$. In this design, the draw of an immigration ball generates a ball of each treatment type; when a treatment type ball is dropped, $\beta$ balls of the same treatment type are added if the outcome is a success and $\alpha$ balls of the alternate treatment type are added if the outcome is a failure. Hence,

$$\mathbf{H} = \begin{pmatrix} \beta p_1 & \alpha q_1 \\ \alpha q_2 & \beta p_2 \end{pmatrix}.$$

Based on Theorems 2.1–2.5 of Section 2, we can derive the asymptotic properties for the three cases: (i) $\mathbf{H1}' > \mathbf{1}'$; (ii) $\mathbf{H1}' < \mathbf{1}'$ and (iii) $\mathbf{H1}' = \mathbf{1}'$. The technical details are omitted here. Nevertheless, it is worth noting that different choices of $\alpha$ and $\beta$ generate various members of the IMU family.

REMARK 3.1. The GDL rule of Zhang et al. (2007) is a member of the IMU class with $D_{m,kj} = 0$, $j \neq k$. In practice, the values of $\theta_k, k = 1, \ldots, K$, are unknown and have to be estimated by sample statistics. The derivation of the asymptotic distributions of the treatment proportions $N_{n,k}$ is usually difficult and is not included by Zhang et al. (2007) if the estimates of $\theta_k, k = 1, \ldots, K$, are used. However, by applying Corollary 2.2, one can obtain the asymptotic properties of $N_{n,k}$ directly.

For example, if the optimal proportion $v_1 = \sqrt{p_1}/(\sqrt{p_1} + \sqrt{p_2})$ is used for comparing two treatments, we can select an IMU model with $\mathbf{D}_m \equiv 0$, $a_{m,k} = C\sqrt{\widehat{p}_k}$, where $\widehat{p}_k$ is the current estimate of the successful probability $p_k$ of treatment $k$, and $C$ is a constant, $k = 1, 2$. By Corollary 2.2, we have

$$\sqrt{n}(N_{n,1}/n - v_1) \xrightarrow{\mathscr{D}} N(0, \sigma^2) \qquad \text{where } \sigma^2 = \frac{1}{2(\sqrt{p_1} + \sqrt{p_2})^3}\left(\frac{p_2 q_1}{\sqrt{p_1}} + \frac{p_1 q_2}{\sqrt{p_2}}\right).$$

Zhang, Hu and Cheung (2006) proposed the use of a GPU without immigration to target this proportion (cf. their Example 2). The corresponding asymptotic variance is

$$\frac{\sqrt{p_1 p_2}}{(\sqrt{p_1} + \sqrt{p_2})^2} + \frac{3}{2(\sqrt{p_1} + \sqrt{p_2})^3}\left(\frac{p_2 q_1}{\sqrt{p_1}} + \frac{p_1 q_2}{\sqrt{p_2}}\right),$$

which is at least triple the variance of this IMU model.

The IMU models, such as those given in the foregoing examples, can be applied in clinical trials. We discuss the applications in three possible directions.

(i) There are numerous applications of urn models in clinical trials. One can apply the proposed IMU models with multiple objectives, such as ethical concerns and design efficiency. For instance, Tamura et al. (1994) discussed the application of the RPW rule, a member of the IMU family, to study the treatment of outpatients suffering from depressive disorder. Later, in a simulation study (using the same data), Bhattacharya (2008) showed that the DL rule, another member of the IMU family, has a smaller variability and yields higher power than the RPW rule. One can apply the asymptotics of the IMU model given in this paper to compare various urn allocation methods instead of using only the simulation results given by Bhattacharya (2008).

(ii) Urn models are also frequently employed in clinical studies to promote balance [see Matthews et al. (2010) and the references therein]. In such circumstances, IMU models should be considered as useful candidates. The introduction of the immigration urn will significantly improve these allocation schemes, mainly in relation to the variability of the urn proportions. Furthermore, asymptotic distributions of IMU models can be derived, leading to a more comprehensive understanding of these urn processes.

(iii) For comparing $K$ treatments, Tymofyeyev, Rosenberger and Hu (2007), Zhu and Hu (2009) obtained optimal allocation proportions for both binary and continuous responses. The IMU models are suitable choices due to their low variability and flexibility in targeting these optimal allocation proportions.

**4. Conclusions.** In this paper, we have proposed a general class of urn models that incorporates immigration. The IMU framework unifies many existing classes of urn models and provides crucial linkages among these models to enable us to have a more comprehensive understanding of different urn processes and their important properties. Further, this framework facilitates the generation of new urn models with desirable properties. Asymptotic properties of the IMU models, with widely satisfied conditions, are given in Section 2. These important results serve to connect existing asymptotic results about urn models. More importantly, the asymptotic normality formula in this article can be employed to evaluate and compare different urn models in terms of the distributions of treatment allocation proportions. Under very mild conditions, the suggested IMU models always yield relatively smaller asymptotic variances. In many cases, the asymptotic variance attains the lower bound. Thus, the IMU models have smaller variabilities than the corresponding generalized Pólya urn models.

In clinical trials, responses may not be available immediately after the patients have been treated. However, there are no logistical difficulties in incorporating delayed responses into the IMU framework. One can update the urn when responses become available. A moderate delay in response [see Hu and Zhang (2004b)] will not affect the asymptotic properties of the IMU. In fact, it is straightforward to modify the proof in the Appendix to incorporate delayed responses.

The discussion of clinical applications has been the main focus of this article because adaptive designs using urn models have received much attention in statistics. However, it is necessary to emphasize that our results are very general and should also play an important role in other areas as well. For example, in quantum mechanics, Niven and Grendar (2009) use the Pólya urn to understand the generalized probability distribution for Maxwell–Boltzmann, Bose–Einstein and Fermi–Dirac statistics. With different colors in the urn, a ball is sampled, recorded and returned to the urn. Then, $c$ balls of the same color are added to the urn. In their formulation, the choices of $c$ are $c > 0$, $c = 0$ and $c < 0$. As $c < 0$ implies a decrease of the number of balls in the urn, it would be interesting to explore the possibility of using the IMU framework to avoid the distinction of balls of a particular type.

## APPENDIX: PROOFS

The outline of the proofs is as follows. First, we prove Theorem 2.2, which is our main result, and then Theorem 2.4. Finally we give a sketch of the proof of Theorem 2.1.

Recall that $\mathbf{Z}_{m-1} = (Z_{m-1,0}, Z_{m-1,1}, \dots, Z_{m-1,K})$ represent the numbers of balls when the $m$th subject arrives to be randomized, $\mathbf{Z}^+_{m-1} = (Z^+_{m-1,0}, Z^+_{m-1,1}, \dots, Z^+_{m-1,K})$ are the nonnegative numbers and $|\mathbf{Z}^+_{m-1}| = Z^+_{m-1,0} + Z^+_{m-1,1} + \cdots + Z^+_{m-1,K}$. Write $\widetilde{\mathbf{Z}}_{m-1} = (Z_{m-1,1}, \dots, Z_{m-1,K})$. Because every immigration ball is replaced, $Z^+_{m-1,0} = Z_{m-1,0} = Z_{0,0}$ for all $m$. Let $\mathbf{X}_m$ be the result of the $m$th assignment, where $X_{m,k} = 1$ if the $m$th subject is assigned to treatment $k$ and 0 otherwise, $k = 1, \dots, K$. Then, $\mathbf{N}_n = (N_{n,1}, \dots, N_{n,K}) = \sum_{m=1}^n \mathbf{X}_m$. Further, we denote $\mathbf{a}_m = (a_{m,1}, \dots, a_{m,K})$, and $\nu_m$ to be the number of draws of type 0 balls between the $(m-1)$th assignment and the $m$th assignment.

Note that between the $(m-1)$th assignment and the $m$th assignment, we have drawn $\nu_m$ balls of type 0. Accordingly, we have added $a_{m-1,k}\nu_m$ balls of type $k$ to the urn. However, when a ball of type $k$ is drawn, it is not replaced and another $D_{m,kj}$ balls of type $j$ are added to the urn. Hence, the change in the number of balls after the $m$th assignment is

$$\widetilde{\mathbf{Z}}_m - \widetilde{\mathbf{Z}}_{m-1} = \mathbf{a}_{m-1}\nu_m + \mathbf{X}_m(\mathbf{D}_m - \mathbf{I}). \tag{A.1}$$

It follows that

$$\begin{aligned}
\widetilde{\mathbf{Z}}_n - \widetilde{\mathbf{Z}}_0 &= \sum_{m=1}^n \mathbf{a}_{m-1}\nu_m + \sum_{m=1}^n \mathbf{X}_m(\mathbf{D}_m - \mathbf{I}) \\
&= \sum_{m=1}^n \mathbf{a}_{m-1}\nu_m - \mathbf{N}_n(\mathbf{I} - \mathbf{H}) + \sum_{m=1}^n \mathbf{X}_m(\mathbf{D}_m - \mathsf{E}[\mathbf{D}_m]) \qquad \text{(A.2)} \\
&= \mathbf{a}N_{n,0} + \sum_{m=1}^n (\mathbf{a}_{m-1} - \mathbf{a})\nu_m - \mathbf{N}_n(\mathbf{I} - \mathbf{H}) + \mathbf{M}_n,
\end{aligned}$$

where $N_{n,0} = \sum_{m=1}^{n} \nu_m$ is total number of draws of type 0 balls after the $n$th assignment, and $\mathbf{M}_n = \sum_{m=1}^{n} \mathbf{X}_m(\mathbf{D}_m - \mathsf{E}[\mathbf{D}_m])$ is a martingale.

To prove Theorem 2.2 we need two lemmas. Their proofs will be given later.

LEMMA A.1. *Suppose that the assumptions in Theorem 2.2 are satisfied. Then, for* $0 < \delta_0 < \frac{1}{2} - \frac{1}{2+\delta}$,

$$Z_{n,k} = o(n^{1/2-\delta_0}) \qquad a.s., \qquad k = 1, \dots, K. \tag{A.3}$$

LEMMA A.2. *Suppose that the assumptions in Theorem 2.2 are satisfied. Then,*

$$N_{n,0} = n/s + O\big(\sqrt{n \log\log n}\big) \qquad a.s., \tag{A.4}$$

$$N_{n,k} = n v_k + O\big(\sqrt{n \log\log n}\big) \qquad a.s., \qquad k = 1, \dots, K, \tag{A.5}$$

*where* $s = \mathbf{a}(\mathbf{I} - \mathbf{H})^{-1}\mathbf{1}'$. *Also, for each* $k = 1, \dots, K$,

$$\widehat{\theta}_{n,k} \to \theta_k \qquad a.s. \tag{A.6}$$

*and*

$$\widehat{\theta}_{n,k} - \theta_k = \frac{Q_{n,k}}{n v_k} + o(n^{-1/2-\delta_0}) \qquad a.s., \tag{A.7}$$

*where* $Q_{n,k} = \sum_{m=1}^{n} X_{m,k}(\xi_{m,k} - \mathsf{E}\xi_{m,k})$ *is a martingale and* $\mathbf{Q}_n = (Q_{n,1}, \dots, Q_{n,K})$.

Now we begin the proof of Theorem 2.2. Consider the $2K$-dimensional martingale $\{(\mathbf{M}_n, \mathbf{Q}_n), \mathcal{A}_n; n \geq 1\}$, where $\mathcal{A}_n = \sigma(\mathbf{X}_1, \dots, \mathbf{X}_n, \boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_{n+1})$. According to (A.5) we have

$$\begin{aligned} \sum_{i=1}^{n} \mathsf{E}[(\Delta \mathbf{M}_i)' \Delta \mathbf{M}_i | \mathcal{A}_{i-1}] &= \sum_{k=1}^{K} N_{n,k} \boldsymbol{\Sigma}_k \\ &= n\boldsymbol{\Sigma}_{11} + O\big(\sqrt{n \log\log n}\big) \qquad \text{a.s.}, \end{aligned} \tag{A.8}$$

$$\begin{aligned} \sum_{i=1}^{n} \mathsf{E}[(\Delta \mathbf{Q}_i)' \Delta \mathbf{Q}_i | \mathcal{A}_{i-1}] &= \boldsymbol{\Sigma}_{22}\, \mathrm{diag}(\mathbf{N}_n) \\ &= n\boldsymbol{\Lambda}_{22} + O\big(\sqrt{n \log\log n}\big) \qquad \text{a.s.}, \end{aligned} \tag{A.9}$$

$$\begin{aligned} \sum_{i=1}^{n} \mathsf{E}[(\Delta \mathbf{M}_i)' \Delta \mathbf{Q}_i | \mathcal{A}_{i-1}] &= \boldsymbol{\Sigma}_{12}\, \mathrm{diag}(\mathbf{N}_n) \\ &= n\boldsymbol{\Lambda}_{12} + O\big(\sqrt{n \log\log n}\big) \qquad \text{a.s.} \end{aligned} \tag{A.10}$$

By Corollary 1.1 of Zhang (2004), we can define the $2K$-dimensional Wiener processes $(\mathbf{W}(t), \mathbf{B}(t))$ with variance–covariance matrix $\mathbf{\Lambda}$ such that for some $\varepsilon > 0$,

$$\mathbf{M}_n = \mathbf{W}(n) + o(n^{1/2-\varepsilon}) \quad \text{a.s.}, \qquad \mathbf{Q}_n = \mathbf{B}(n) + o(n^{1/2-\varepsilon}) \quad \text{a.s.} \tag{A.11}$$

Without loss of generality, we assume that $\varepsilon \le \delta_0$, where $\delta_0$ is defined as it is in Lemma A.1. Next, we need to show that $(\mathbf{W}(t), \mathbf{B}(t))$ satisfies (2.4). Combining (A.2) and (A.3) yields

$$\mathbf{N}_n(\mathbf{I} - \mathbf{H}) - \mathbf{a}N_{n,0} = \mathbf{M}_n + \sum_{m=1}^{n} (\mathbf{a}_{m-1} - \mathbf{a})\nu_m + o(n^{1/2-\delta_0}) \quad \text{a.s.} \tag{A.12}$$

Recall that $\mathbf{A} = (\mathbf{I} - \mathbf{H})^{-1}(\mathbf{I} - \mathbf{1}'\mathbf{v})$, $\mathbf{v} = \mathbf{a}(\mathbf{I} - \mathbf{H})^{-1}/(\mathbf{a}(\mathbf{I} - \mathbf{H})^{-1}\mathbf{1}')$ and note that $\mathbf{N}_n\mathbf{1}' = n$, $\mathbf{a}\mathbf{A} = s\mathbf{v}(\mathbf{I} - \mathbf{1}'\mathbf{v}) = \mathbf{0}$. According to (A.12),

$$\mathbf{N}_n - n\mathbf{v} = \left(\mathbf{M}_n + \sum_{m=1}^{n} (\mathbf{a}_{m-1} - \mathbf{a})\nu_m\right)\mathbf{A} + o(n^{1/2-\delta_0}) \quad \text{a.s.} \tag{A.13}$$

For $\mathbf{a}_m - \mathbf{a}$, due to (A.7) and (A.11),

$$\begin{aligned} \mathbf{a}_m - \mathbf{a} &= (\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}) \frac{\partial \mathbf{a}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + O(\|\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}\|^2) \\ &= \frac{\mathbf{Q}_m}{m} \operatorname{diag}\left(\frac{1}{\mathbf{v}}\right) \frac{\partial \mathbf{a}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + o(m^{-1/2-\delta_0}) \\ &= \frac{\mathbf{B}(m)}{m} \operatorname{diag}\left(\frac{1}{\mathbf{v}}\right) \frac{\partial \mathbf{a}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} + o(m^{-1/2-\varepsilon}). \end{aligned} \tag{A.14}$$

Note that immigration occurs only when a type 0 ball is drawn. Let $\tau_m$ be the total number of draws when the $m$th type 0 ball is drawn. At that time, $\tau_m - m$ subjects have been assigned and the $(\tau_m - m + 1)$th subject arrives to be randomized. Hence, we add $a_{(\tau_m - m+1)-1,k}$ balls of type $k$ to the urn, $k = 1, \dots, K$. It follows that

$$\sum_{j=1}^{n} a_{j-1,k} \cdot \nu_j = \sum_{m=1}^{N_{n,0}} a_{\tau_m - m,k},$$

that is,

$$\sum_{m=1}^{n} (\mathbf{a}_{m-1} - \mathbf{a})\nu_m = \sum_{m=1}^{N_{n,0}} (\mathbf{a}_{\tau_m - m} - \mathbf{a}).$$

It is easily seen that $\tau_m = \min\{n : N_{n,0} \ge m\} + m$. Due to (A.4),

$$\tau_m - m = \min\{n : N_{n,0} \ge m\} = sm + O\left(\sqrt{m \log\log m}\right) \quad \text{a.s.}$$

It follows that

$$\begin{aligned}\mathbf{a}_{\tau_m-m}-\mathbf{a} &= \frac{\mathbf{B}(\tau_m-m)}{\tau_m-m}\operatorname{diag}\left(\frac{1}{\mathbf{v}}\right)\frac{\partial\mathbf{a}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}+o(m^{-1/2-\varepsilon})\\ &= \frac{\mathbf{B}(sm)}{sm}\operatorname{diag}\left(\frac{1}{\mathbf{v}}\right)\frac{\partial\mathbf{a}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}+o(m^{-1/2-\varepsilon}) \qquad \text{a.s.}\end{aligned}$$

Using (A.4), we conclude that

$$\begin{aligned}&\sum_{m=1}^{n}(\mathbf{a}_{m-1}-\mathbf{a})\nu_m\\ &\quad=\sum_{m=1}^{N_{n,0}}\left(\frac{\mathbf{B}(sm)}{sm}\operatorname{diag}\left(\frac{1}{\mathbf{v}}\right)\frac{\partial\mathbf{a}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}+o(m^{-1/2-\varepsilon})\right)\\ &\quad=\int_0^{n/s}\frac{\mathbf{B}(sx)}{sx}\,dx\operatorname{diag}\left(\frac{1}{\mathbf{v}}\right)\frac{\partial\mathbf{a}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}+o(n^{1/2-\varepsilon})\\ &\quad=\int_0^{n}\frac{\mathbf{B}(x)}{x}\,dx\operatorname{diag}\left(\frac{1}{\mathbf{v}}\right)\frac{1}{s}\frac{\partial\mathbf{a}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}+o(n^{1/2-\varepsilon}) \qquad \text{a.s.}\end{aligned}\tag{A.15}$$

However, it is easily checked that

$$\frac{1}{s}\frac{\partial\mathbf{a}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\mathbf{A}=\frac{\partial\mathbf{v}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}.\tag{A.16}$$

Combining (A.11)–(A.16) the proof of (2.4) is complete.

Three more lemmas are needed before we prove Lemmas A.1 and A.2.

LEMMA A.3. *Under Assumption 2.2 and* $Z_{0,0}>0$, *we have*

$$Z^-_{n,k}=O(1) \qquad a.s., \qquad k=1,\dots,K.$$

PROOF. Note that $|\mathbf{Z}^+_m|\ge Z_{0,0}>0$ for all $m$ and so that the balls with negative numbers have no chance of being drawn. In addition, at most $C+1$ balls of each treatment type have the chance of being removed only when a ball of the same type is drawn because of the Assumption 2.2. It follows that $Z_{n,k}\ge -C-1$. □

LEMMA A.4. *Let* $\mathcal{F}_n=\sigma(\mathbf{X}_1,\dots,\mathbf{X}_n,\mathbf{Z}_1,\dots,\mathbf{Z}_n)$ *be the history sigma field, and* $A_m=\sum_{k=1}^K a_{m,k}$. *Suppose that Assumption 2.2 is satisfied. Then,* $\underline{A}:=\min_m A_m>0$ *implies*

$$\mathsf{E}[\nu_n^p|\mathcal{F}_{n-1}]\le c_p\left(\left(\sum_{k=1}^K Z_{n-1,k}\right)^-\Big/\underline{A}\right)^{p+1}\frac{Z_{0,0}}{|\mathbf{Z}^+_{n-1}|} \qquad a.s. \qquad \forall p\ge 1,\tag{A.17}$$

*where* $c_p>0$ *is a random variable that is a function of* $Z_{0,0}$ *and* $\min_m A_m$, *particularly,*

$$\min_m A_m>0 \quad \textit{implies} \quad \mathsf{E}[\nu_n^p|\mathcal{F}_{n-1}]=O(1) \qquad a.s.\tag{A.18}$$

PROOF. The event $\{\nu_n = l\}$ means that when the $n$th subject is assigned, we have drawn $l+1$ balls continuously in which the first $l$ balls is of type 0 and the last one is not. Hence, $\mathsf{P}(\nu_n = 0|\mathcal{F}_{n-1}) = 1 - Z_{0,0}/|\mathbf{Z}_{n-1}^+|$, and for $l = 1, 2, \dots,$

$$\begin{aligned}\mathsf{P}(\nu_n = l|\mathcal{F}_{n-1}) = \frac{Z_{0,0}}{|\mathbf{Z}_{n-1}^+|}\prod_{j=1}^{l-1}\frac{Z_{0,0}}{|(\mathbf{Z}_{n-1} + j\mathbf{a}_{n-1})^+|}\\ \times\left(1 - \frac{Z_{0,0}}{|(\mathbf{Z}_{n-1} + l\mathbf{a}_{n-1})^+|}\right).\end{aligned} \tag{A.19}$$

Obviously, $\mathsf{P}(\nu_n = l|\mathcal{F}_{n-1}) \le Z_{0,0}/|\mathbf{Z}_{n-1}^+|$, $l \ge 1$. Note that

$$|(\mathbf{Z}_{n-1} + \mathbf{j}\mathbf{a}_{n-1})^+| = Z_{0,0} + \sum_{k=1}^{K}(Z_{n-1,k} + ja_{n-1,k})^+ \ge Z_{0,0} + \sum_{k=1}^{K} Z_{n-1,k} + jA_{n-1}.$$

It follows that $\underline{A} > 0$ and $\sum_{k=1}^{K} Z_{n-1,k} \ge -L\underline{A}$ imply for $l \ge L$,

$$\mathsf{P}(\nu_n = l|\mathcal{F}_{n-1}) \le \frac{Z_{0,0}}{|\mathbf{Z}_{n-1}^+|}\prod_{j=L}^{l-1}\frac{Z_{0,0}}{Z_{0,0} + (j-L)\underline{A}} \le c_0\frac{Z_{0,0}}{|\mathbf{Z}_{n-1}^+|}e^{-2(l-L)}, \tag{A.20}$$

where $c_0 > 0$ depends only on $\underline{A}$ and $Z_{0,0}$. So

$$\mathsf{E}[\nu_n^p|\mathcal{F}_{n-1}] \le \sum_{l=1}^{L} l^p\frac{Z_{0,0}}{|\mathbf{Z}_{n-1}^+|} + \sum_{l=L+1}^{\infty} l^p c_0\frac{Z_{0,0}}{|\mathbf{Z}_{n-1}^+|}e^{-2(l-L)} \le c_p L^{p+1}\frac{Z_{0,0}}{|\mathbf{Z}_{n-1}^+|}.$$

Taking $L = [(\sum_{k=1}^{K} Z_{n-1,k})^-/\underline{A}] + 1$ completes the proof of (A.17). Equation (A.18) follows from (A.17) and Lemma A.3. □

LEMMA A.5. *Suppose that Assumptions 2.1 and 2.2 are satisfied. Then*

$$\min_{m,k} a_{m,k} > 0 \quad \textit{and} \quad \max_{m,k} a_{m,k} < \infty \qquad \textit{a.s.} \tag{A.21}$$

PROOF. By Lemma A.4 of Hu and Zhang (2004a), we have

$$N_{n,k} \to \infty \quad \text{implies} \quad \widehat{\theta}_{n,k} \to \theta_k \qquad \text{a.s.}, \qquad k = 1, \dots, K. \tag{A.22}$$

Then, $a_k(\mathbf{y}) > 0$ for any $\mathbf{y}$ on closure$\{\widehat{\boldsymbol{\theta}}_m; m = 1, 2, \dots\} = \bigotimes_{k=1}^{K}\{\theta_k, \widehat{\theta}_{m,k}; m = 1, 2, \dots\}$. By the continuity of $a_k(\cdot)$, (A.21) is satisfied. □

PROOF OF LEMMA A.1. By Lemma A.5,

$$\underline{A} =: \min_m A_m > 0 \quad \text{and} \quad \overline{A} =: \max_m A_m < \infty. \tag{A.23}$$

Note that $\widetilde{\mathbf{Z}}_n\mathbf{1}' = \sum_{k=1}^K Z_{n-1,k}$. By (A.17) and Lemma A.3, $\mathsf{E}[\nu_n|\mathcal{F}_{n-1}] \le C_0 Z_{0,0}/|\mathbf{Z}_{n-1}^+|$. So, according to (A.1) or (A.2), we have

$$\begin{aligned}\widetilde{\mathbf{Z}}_n\mathbf{1}' &= \widetilde{\mathbf{Z}}_{n-1}\mathbf{1}' + \nu_n A_{n-1} - \mathbf{X}_n(\mathbf{I}-\mathbf{H})\mathbf{1}' + \Delta\mathbf{M}_n\mathbf{1}' \\ &\le \widetilde{\mathbf{Z}}_{n-1}\mathbf{1}' + A_{n-1}\mathsf{E}[\nu_n|\mathcal{F}_{n-1}] - \underline{h} \\ &\qquad + A_{n-1}(\nu_n - \mathsf{E}[\nu_n|\mathcal{F}_{n-1}]) + \Delta\mathbf{M}_n\mathbf{1}' \\ &\le \widetilde{\mathbf{Z}}_{n-1}\mathbf{1}' + C_0\overline{A}\frac{Z_{0,0}}{|\mathbf{Z}_{n-1}^+|} - \underline{h} + \Delta U_n \\ &\le \widetilde{\mathbf{Z}}_{n-1}\mathbf{1}' + \Delta U_n - \underline{h}/2 \qquad \text{if } \widetilde{\mathbf{Z}}_{n-1}\mathbf{1}' \ge 2C_0\overline{A}Z_{0,0}/\underline{h},\end{aligned} \tag{A.24}$$

where $\underline{h} = \min_k(1 - \sum_{j=1}^K h_{kj}) > 0$. Here, $U_n = \sum_{m=1}^n A_{m-1}(\nu_m - \mathsf{E}[\nu_m|\mathcal{F}_{n-1}]) + \mathbf{M}_n\mathbf{1}'$ is a real martingale. Let $S_n = \max\{1 \le j \le n : \widetilde{\mathbf{Z}}_j\mathbf{1}' < 2C_0\overline{A}Z_{0,0}/\underline{h}\}$, where $\max(\varnothing) = 0$. Then, according to (A.24),

$$\begin{aligned}\widetilde{\mathbf{Z}}_n\mathbf{1}' &\le \widetilde{\mathbf{Z}}_{n-1}\mathbf{1}' + \Delta U_n - \underline{h}/2 \le \cdots \\ &\le \widetilde{\mathbf{Z}}_{S_n}\mathbf{1}' + \Delta U_{S_n+1} + \cdots + \Delta U_n - (n - S_n)\underline{h}/2 \\ &\le |\mathbf{Z}_0| \vee (2C_0\overline{A}Z_{0,0}/\underline{h}) + U_n - U_{S_n} - (n - S_n)\underline{h}/2.\end{aligned} \tag{A.25}$$

For the martingale $\{U_n, \mathcal{F}_n; n = 1, 2, \ldots\}$, we have

$$\mathsf{E}[|\Delta U_n|^{2+\delta}|\mathcal{F}_{n-1}] \le C + C\max_j A_j^{2+\delta} = O(1)$$

due to Assumption 2.2 and (A.18). Accordingly, we can show that

$$U_n = O\big(\sqrt{n\log\log n}\big) \qquad \text{a.s.,} \tag{A.26}$$

$$\max_{m\le\sqrt{n\log n}}\big|U_{n-[\sqrt{n\log n}]+m} - U_{n-[\sqrt{n\log n}]}\big| = o\big(n^{1/(2+\delta)}\log n\big) \qquad \text{a.s.} \tag{A.27}$$

If $n - S_n \ge \sqrt{n\log n}$, then for $n$ large enough

$$U_n - U_{S_n} - (n - S_n)\underline{h}/2 \le O\big(\sqrt{n\log\log n}\big) - \underline{h}\sqrt{n\log n}/2 < 0$$

due to (A.26). Note that $n \ge S_n$. If $n - S_n < \sqrt{n\log n}$, then

$$\begin{aligned}U_n - U_{S_n} - (n - S_n)\underline{h}/2 &\le 2\max_{m\le\sqrt{n\log n}}\big|U_{n-[\sqrt{n\log n}]+m} - U_{n-[\sqrt{n\log n}]}\big| \\ &= o\big(n^{1/(2+\delta)}\log n\big) \qquad \text{a.s.}\end{aligned}$$

by (A.27). It follows that $\sum_{k=1}^K Z_{n,k} \le o(n^{1/2-\delta_0})$ a.s. due to (A.25). However, $Z_{n,k}^- = O(1)$ a.s. by Lemma A.3. Equation (A.3) is proved. $\square$

PROOF OF LEMMA A.2. Recall $Q_{n,k} = \sum_{m=1}^n X_{m,k}(\xi_{m,k} - \theta_k)$, $k = 1, \ldots, K$, and both $\{M_{n,k}, \mathcal{A}_n; n \ge 1\}$ and $\{Q_{n,k}, \mathcal{A}_n; n \ge 1\}$ are martingales. According to the law of the iterated logarithm for martingales, we have

$$M_{n,k} = O\big(\sqrt{n\log\log n}\big) \quad \text{and} \quad Q_{n,k} = O\big(\sqrt{n\log\log n}\big) \qquad \text{a.s.} \tag{A.28}$$

However, for each $k = 1, \dots, K$,

$$\widehat{\theta}_{n,k} - \theta_k = \frac{Q_{n,k} + O(1)}{N_{n,k} + c_2} \qquad \text{a.s.} \tag{A.29}$$

Equation (A.12) remains true by Lemma A.1. By (A.12) and (A.28) we have

$$\mathbf{N}_n(\mathbf{I} - \mathbf{H}) = \sum_{m=1}^{n} \mathbf{a}_{m-1} v_m + o(n) \qquad \text{a.s.} \tag{A.30}$$

Note that all elements of the vector $\sum_{m=1}^{n} \mathbf{a}_{m-1} v_m$ are between $\underline{a} N_{n,0}$ and $\overline{a} N_{n,0}$, where $\underline{a} = \min_{m,k} a_{m,k}$ and $\overline{a} = \max_{m,k} a_{m,k}$. Hence, it is obvious that $\liminf_{n\to\infty} N_{n,0}/n > 0$ a.s. because otherwise the limit of $\mathbf{N}_n/n$ may be $\mathbf{0}$ which contradicts to $\mathbf{N}_n \mathbf{1}' = n$. On the other hand, the $k$th element of $\mathbf{N}_n(\mathbf{I} - \mathbf{H})$ does no exceed $(1 - h_{kk}) N_{n,k}$. It follows that $\liminf_{n\to\infty} N_{n,k}/n > 0$ a.s. by (A.30), which, together with (A.29) and (A.28), implies

$$\widehat{\theta}_{n,k} - \theta_k = O\left(\frac{Q_{m,k} + O(1)}{n}\right) = O\left(\sqrt{\frac{\log\log n}{n}}\right) \to 0 \qquad \text{a.s.}$$

Equation (A.6) is proved and also

$$\begin{aligned} a_{m,k} - a_k &= a_k(\widehat{\boldsymbol{\theta}}_m) - a_k(\boldsymbol{\theta}) = O(\|\widehat{\boldsymbol{\theta}}_m - \boldsymbol{\theta}\|) \\ &= O\big(\sqrt{(\log\log m)/m}\big) \qquad \text{a.s.} \end{aligned} \tag{A.31}$$

Hence, by Theorem 2.18 of Hall and Heyde (1980) it is easy to check that $\sum_{m=1}^{n} (a_{m-1,k} - a_k)(v_m - \mathsf{E}[v_m|\mathcal{F}_{m-1}]) = o(\sqrt{n})$ a.s. It follows that

$$\begin{aligned} \sum_{m=1}^{n} (a_{m-1,k} - a_k) v_m &= \sum_{m=1}^{n} (a_{m-1,k} - a_k) \mathsf{E}[v_m|\mathcal{F}_{m-1}] + o(\sqrt{n}) \\ &= \sum_{m=1}^{n} O\left(\sqrt{\frac{\log\log m}{m}}\right) O(1) + o(\sqrt{n}) \\ &= O\big(\sqrt{n \log\log n}\big) \qquad \text{a.s.} \end{aligned} \tag{A.32}$$

by (A.18) and (A.31). Combining (A.12), (A.28) and (A.32) yields

$$\mathbf{N}_n - N_{n,0} \mathbf{a} (\mathbf{I} - \mathbf{H})^{-1} = O\big(\sqrt{n \log\log n}\big) \qquad \text{a.s.,}$$

which, together with $\mathbf{N}_n \mathbf{1}' = n$, implies (A.4) and (A.5). Then, combining (A.5), (A.28) and (A.29) yields

$$\widehat{\theta}_{n,k} - \theta_k = \frac{Q_{n,k} + O(1)}{n v_k + O(\sqrt{n \log\log n})} = \frac{Q_{n,k}}{n v_k} + o(n^{1/2 - \delta_0}) \qquad \text{a.s.}$$

Equation (A.7) is proved, and the proof of Theorem 2.2 is completed. $\square$

PROOF OF THEOREM 2.4. Note that Assumptions 2.1 and 2.2 are satisfied, and $Z_{0,0} > 0$, $\mathbf{H}\mathbf{1}' = \mathbf{1}'$. Similarly to (A.24),

$$\widetilde{\mathbf{Z}}_n \mathbf{1}' = \widetilde{\mathbf{Z}}_{n-1}\mathbf{1}' + \nu_n A_{n-1} + \Delta \mathbf{M}_n \mathbf{1}' \le \widetilde{\mathbf{Z}}_{n-1}\mathbf{1}' + C_0 \overline{A} \frac{Z_{0,0}}{|\mathbf{Z}_{n-1}^+|} + \Delta U_n$$
$$\le \widetilde{\mathbf{Z}}_{n-1}\mathbf{1}' + C_0 \overline{A}/\sqrt{n} + \Delta U_n \qquad \text{if } \widetilde{\mathbf{Z}}_{n-1}\mathbf{1}' \ge Z_{0,0}\sqrt{n}.$$

It follows that

$$\widetilde{\mathbf{Z}}_n \mathbf{1}' \le \widetilde{\mathbf{Z}}_{S_n}\mathbf{1}' + \Delta U_{S_n+1} + \cdots + \Delta U_n + C_0 \overline{A}(n - S_n)/\sqrt{n}$$
$$\le 2C_0 \overline{A}\sqrt{n} + U_n - U_{S_n} \le 2C_0 \overline{A}\sqrt{n} + 2 \max_{m \le n} |U_m|,$$

where $S_n = \max\{1 \le j \le n : \widetilde{\mathbf{Z}}_j \mathbf{1}' < Z_{0,0}/\sqrt{n}\}$ and $\max(\varnothing) = 0$. Hence,

$$\widetilde{\mathbf{Z}}_n = O\big(\sqrt{n \log\log n}\big) \qquad \text{a.s.} \quad \text{and} \quad = O_P\big(\sqrt{n}\big)$$

by the properties of a martingale and Lemma A.3. So, by (A.2) and the law of the iterated logarithm of martingales, it follows that

$$\mathbf{N}_n(\mathbf{I} - \mathbf{H}) = \mathbf{M}_n + \sum_{m=1}^{n} \mathbf{a}_{m-1}\nu_m - \widetilde{\mathbf{Z}}_n + \widetilde{\mathbf{Z}}_0 \sum_{m=1}^{n} \mathbf{a}_{m-1}\nu_m$$
$$+ O\big(\sqrt{n \log\log n}\big) \qquad \text{a.s.}$$

Multiplying by $\mathbf{1}'$ yields $\sum_{m=1}^{n} \nu_m A_{m-1} = O(\sqrt{n \log\log n})$ a.s., and then $N_{n,0} = O(\sqrt{n \log\log n})$ a.s. and $\sum_{m=1}^{n} \mathbf{a}_{m-1}\nu_m = O(\sqrt{n \log\log n})$ a.s. by (A.21). So,

$$(\mathbf{N}_n - n\mathbf{v})\big(\mathbf{I} - (\mathbf{H} - \mathbf{1}'\mathbf{v})\big) = \mathbf{N}_n(\mathbf{I} - \mathbf{H}) = O\big(\sqrt{n \log\log n}\big) \qquad \text{a.s.}$$

It follows that $\mathbf{N}_n - n\mathbf{v} = O(\sqrt{n \log\log n})$ a.s. because $(\mathbf{I} - (\mathbf{H} - \mathbf{1}'\mathbf{v}))$ is invertible. The proof of $\mathbf{N}_n - n\mathbf{v} = O_P(\sqrt{n})$ is similar. $\square$

PROOF OF THEOREM 2.1. Recall (A.2); we have

$$\widetilde{\mathbf{Z}}_n - \widetilde{\mathbf{Z}}_0 = \sum_{m=1}^{n} \mathbf{a}_{m-1}\nu_m + \mathbf{N}_n(\mathbf{H} - \mathbf{I}) + \mathbf{M}_n. \tag{A.33}$$

It follows that $|\widetilde{\mathbf{Z}}_n| = \widetilde{\mathbf{Z}}_n \mathbf{1}' \ge (\gamma - 1)n + \mathbf{M}_n \mathbf{1}'$ by noticing $\mathbf{H}\mathbf{1}' = \gamma \mathbf{1}'$. Hence,

$$\liminf_{n \to \infty} \frac{|\widetilde{\mathbf{Z}}_n^+|}{n} \ge \liminf_{n \to \infty} \frac{|\widetilde{\mathbf{Z}}_n|}{n} \ge \gamma - 1 > 0 \qquad \text{a.s.}$$

Without loss of generality we can thus assume that $|\widetilde{\mathbf{Z}}_n^+| \ge cn > 0$ for all $n$. Then, the conclusion of Lemma A.3 remains true. By Lemma A.3, $\widetilde{\mathbf{Z}}_m = \widetilde{\mathbf{Z}}_m^+ + O(1)$ a.s.

On the other hand, by (A.19) we have

$$\mathsf{P}(\nu_m = 1|\mathcal{F}_{m-1}) = \frac{Z_{0,0}}{|Z_{m-1}^+|}\left(1 - \frac{Z_{0,0}}{|(Z_{m-1} + \mathbf{a}_{m-1})^+|}\right) \le c/m \qquad \text{a.s.},$$

$$\mathsf{P}(\nu_m \ge 2|\mathcal{F}_{m-1}) = \frac{Z_{0,0}}{|Z_{m-1}^+|}\frac{Z_{0,0}}{|(Z_{m-1} + \mathbf{a}_{m-1})^+|} \le \left(\frac{Z_{0,0}}{|Z_{m-1}^+|}\right)^2 \le c/m^2.$$

It follows that $\mathsf{P}(\nu_m \ge 2 \text{ i.o.}) = 0$ and $\sum_{m=1}^n I\{\nu_m = 1\} = O(\log^2 n)$ a.s. by Theorem 3.3.9(ii) of Stout (1974). So by the assumption stated in Theorem 2.1 that $0 \le a_{m,k} \le Cm^{1/2-\delta_0}$,

$$\sum_{m=1}^n \mathbf{a}_{m-1}\nu_m = O\Big(\max_{m\le n} A_{m-1}\Big)\left(\sum_{m=1}^n I\{\nu_m = 1\} + O(1)\right) = o(n^{1/2-\delta_0/2}) \qquad \text{a.s.},$$

which means that the immigrated balls can be neglected. In addition,

$$\begin{aligned}\mathsf{P}(X_{m,k} = 1|\mathcal{F}_{m-1}) &= \frac{\mathbf{Z}_{m-1,k}^+}{Z_{0,0} + |\widetilde{\mathbf{Z}}_{m-1}^+|}\left(1 - \frac{\mathbf{Z}_{0,0}^+}{Z_{0,0} + |\widetilde{\mathbf{Z}}_{m-1}^+|}\right) \\ &\quad + \mathsf{P}(X_{m,k} = 1, \nu_m \ge 1|\mathcal{F}_{m-1}) \\ &= \frac{\mathbf{Z}_{m-1,k}^+}{|\widetilde{\mathbf{Z}}_{m-1}^+|} + O\left(\frac{1}{m}\right) \qquad \text{a.s.}\end{aligned}$$

It follows that

$$\begin{aligned}\widetilde{\mathbf{Z}}_n^+ &= \widetilde{\mathbf{Z}}_n + O(1) = \mathbf{N}_n(\mathbf{H} - \mathbf{I}) + \mathbf{M}_n + o(n^{1/2-\delta_0/2}) \\ &= \sum_{m=1}^n (\mathbf{X}_m - \mathsf{E}[\mathbf{X}_m|\mathcal{F}_{m-1}])(\mathbf{H} - \mathbf{I}) + \mathbf{M}_n \\ &\quad + \sum_{m=1}^n \mathsf{E}[\mathbf{X}_m|\mathcal{F}_{m-1}](\mathbf{H} - \mathbf{I}) + o(n^{1/2-\delta_0/2}) \\ &= \sum_{m=0}^{n-1} (\mathbf{X}_m - \mathsf{E}[\mathbf{X}_m|\mathcal{F}_{m-1}])(\mathbf{H} - \mathbf{I}) + \mathbf{M}_n \\ &\quad + \sum_{m=1}^n \left[\frac{\widetilde{\mathbf{Z}}_{m-1}^+}{|\widetilde{\mathbf{Z}}_{m-1}^+|} + O\left(\frac{1}{m}\right)\right](\mathbf{H} - \mathbf{I}) + o(n^{1/2-\delta_0/2}) \\ &= (\gamma - 1)n\mathbf{v} + (\gamma - 1)\sum_{m=1}^n (\mathbf{X}_m - \mathsf{E}[\mathbf{X}_m|\mathcal{F}_{m-1}])\widetilde{\mathbf{H}} + \mathbf{M}_n \\ &\quad + \sum_{m=0}^{n-1} \frac{\widetilde{\mathbf{Z}}_m^+}{|\widetilde{\mathbf{Z}}_m^+|}(\gamma - 1)\widetilde{\mathbf{H}} + o(n^{1/2-\delta_0/2}) \qquad \text{a.s.}\end{aligned}$$

The expansion for $\widetilde{\mathbf{Z}}_n^+$ is similar to that for $\mathbf{Y}_n$ in (6.2) of Zhang and Hu (2009), pages 1324–1421. Hence, the rest of the proof is omitted. $\square$

**Acknowledgments.** Special thanks go to the anonymous referee and the Associate Editor for their constructive comments, which led to a much improved version of the paper.

## REFERENCES

ATHREYA, K. B. and KARLIN, S. (1968). Embedding of urn schemes into continuous time branching processes and related limit theorems. *Ann. Math. Statist.* **39** 1801–1817. MR0232455

ATHREYA, K. B. and NEY, P. E. (1972). *Branching Processes*. Springer, New York. MR0373040

BAI, Z. D. and HU, F. (1999). Asymptotic theorem for urn models with nonhomogeneous generating matrices. *Stochastic Process. Appl.* **80** 87–101. MR1670107

BAI, Z. D. and HU, F. (2005). Asymptotics in randomized urn models. *Ann. Appl. Probab.* **15** 914–940. MR2114994

BAI, Z. D., HU, F. and ROSENBERGER, W. F. (2002). Asymptotic properties of adaptive designs for clinical trials with delayed response. *Ann. Statist.* **30** 122–139. MR1892658

BAI, Z. D., HU, F. and SHEN, L. (2002). An adaptive design for multi-arm clinical trials. *J. Multivariate Anal.* **81** 1–18. MR1901202

BHATTACHARYA, R. (2008). Urn-based response adaptive procedures and optimality. *Drug Information Journal* **42** 441–448.

BEGGS, A. W. (2005). On the convergence of reinforcement learning. *J. Econom. Theory* **122** 1–36. MR2131871

BENAÏM, M., SCHREIBER, S. J. and TARRÈS, P. (2004). Generalized urn models of evolutionary processes. *Ann. Appl. Probab.* **14** 1455–1478. MR2071430

DONNELLY, P. and KURTZ, T. G. (1996). The asymptotic behavior of an urn model arising in population genetics. *Stochastic Process. Appl.* **64** 1–16. MR1419488

DURHAM, S. D., FLOURNOY, N. and LI, W. (1998). Sequential designs for maximizing the probability of a favorable response. *Canad. J. Statist.* **26** 479–495. MR1646698

HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and its Applications*. Academic Press, New York. MR0624435

HIGUERAS, I., MOLER, J., FLO, F. and SAN MIGUEL, M. (2006). Central limit theorems for generalized Pólya urn models. *J. Appl. Probab.* **43** 938–951. MR2274628

HOPPE, F. M. (1984). Pólya-like urns and the Ewens' sampling formula. *J. Math. Biol.* **20** 91–94. MR0758915

HU, F. and ROSENBERGER, W. F. (2003). Optimality, variability, power: Evaluating response-adaptive randomization procedures for treatment comparisons. *J. Amer. Statist. Assoc.* **98** 671–678. MR2011680

HU, F. and ROSENBERGER, W. F. (2006). *The Theory of Response-Adaptive Randomization in Clinical Trials*. Wiley, Hoboken, NJ. MR2245329

HU, F., ROSENBERGER, W. F. and ZHANG, L.-X. (2006). Asymptotically best response-adaptive randomization procedures. *J. Statist. Plann. Inference* **136** 1911–1922. MR2255603

HU, F. and ZHANG, L. X. (2004a). Asymptotic properties of doubly adaptive biased coin designs for multi-treatment clinical trials. *Ann. Statist.* **32** 268–301. MR2051008

HU, F. and ZHANG, L.-X. (2004b). Asymptotic normality of urn models for clinical trials with delayed response. *Bernoulli* **10** 447–463. MR2061440

IVANOVA, A. (2003). A play-the-winner type urn model with reduced variability. *Metrika* **58** 1–13. MR1999248

IVANOVA, A. (2006). Urn designs with immigration: Useful connection with continuous time stochastic processes. *J. Statist. Plann. Inference* **136** 1836–1844. MR2255599

IVANOVA, A. and FLOURNOY, N. (2001). A birth and death urn for ternary outcomes: Stochastic processes applied to urn models. In *Probability and Statistical Models with Applications* (C. A. Charalambides, M. V. Koutras and N. Balakrishnan, eds.) 583–600. Chapman and Hall/CRC Press, Boca Raton, FL.

IVANOVA, A., ROSENBERGER, W. F., DURHAM, S. D. and FLOURNOY, N. (2000). A birth and death urn for randomized clinical trials: Asymptotic methods. *Sankhyā Ser. B* **62** 104–118. MR1790832

JANSON, S. (2004). Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stochastic Process. Appl.* **110** 177–245. MR2040966

JOHNSON, N. L. and KOTZ, S. (1977). *Urn Models and Their Applications*. Wiley, New York. MR0488211

KOTZ, S. and BALAKRISHNAN, N. (1997). Advances in urn models during the past two decades. In *Advances in Combinatorial Methods and Applications to Probability and Statistics* (N. Balakrishnan, ed.) 203–257. Birkhäuser, Boston. MR1456736

KNOBLAUCH, K., NEITZ, M. and NEITZ, J. (2006). An urn model of the development of L/M cone ratios in human and macaque retinas. *Visual Neuroscience* **23** 387–394.

MATTHEWS, E. E., COOK, P. F., TERADA, M. and ALOIA, M. S. (2010). Randomizing research participants: Promoting balance and concealment in small samples. *Research in Nursing and Health* **33** 243–253.

MILENKOVIC, O. and COMPTON, K. J. (2004). Probabilistic transforms for combinatorial urn models. *Combin. Probab. Comput.* **13** 645–675. MR2095977

NIVEN, R. K. and GRENDAR, M (2009). Generalized classical, quantum and intermediate statistics and the Pólya urn model. *Phys. Lett. A* **373** 621–626. MR2493686

ROSENBERGER, W. F., STALLARD, N., IVANOVA, A., HARPER, C. N. and RICKS, M. L. (2001). Optimal adaptive designs for binary response trials. *Biometrics* **57** 909–913. MR1863454

SMYTHE, R. T. (1996). Central limit theorems for urn models. *Stochastic Process. Appl.* **65** 115–137. MR1422883

STOUT, W. F. (1974). *Almost Sure Convergence*. Academic Press, New York. MR0455094

TAMURA, R. N., FARIES, D. E., ANDERSEN, J. S. and HEILIGENSTEIN, J. H. (1994). A case study of an adaptive clinical trial in the treatment of out-patients with depressive disorder. *J. Amer. Statist. Assoc.* **89** 768–776.

TYMOFYEYEV, Y., ROSENBERGER, W. F. and HU, F. (2007). Implementing optimal allocation in sequential binary response experiments. *J. Amer. Statist. Assoc.* **102** 224–234. MR2345540

WEI, L. J. (1979). The generalized Pólya's urn design for sequential medical trials. *Ann. Statist.* **7** 291–296.

WEI, L. J. and DURHAM, S. D. (1978). The randomized play-the-winner rule in medical trials. *J. Amer. Statist. Assoc.* **73** 840–843.

ZHANG, L. J. and ROSENBERGER, W. F. (2006). Response-adaptive randomization for clinical trials with continuous outcomes. *Biometrics* **62** 562–569. MR2236838

ZHANG, L.-X. (2004). Strong approximations of martingale vectors and their applications in Markov-chain adaptive designs. *Acta Math. Appl. Sin. Engl. Ser.* **20** 337–352. MR2064011

ZHANG, L.-X. and HU, F. (2009). The Gaussian approximation for multi-color generalized Friedman's urn model. *Sci. China Ser. A* **52** 1305–1326. MR2520576

ZHANG, L.-X., CHAN, W. S., CHEUNG, S. H. and HU, F. (2007). A generalized urn model for clinical trials with delayed responses. *Statist. Sinica* **17** 387–409. MR2352516

ZHANG, L.-X., HU, F. and CHEUNG, S. H. (2006). Asymptotic theorems of sequential estimation-adjusted urn models. *Ann. Appl. Probab.* **16** 340–369. MR2209345

ZHU, H. and HU, F. (2009). Implementing optimal allocation in sequential continuous response experiments. *J. Statist. Plann. Inference* **139** 2420–2430. MR2508003

L.-X. ZHANG
DEPARTMENT OF MATHEMATICS
ZHEJIANG UNIVERSITY
HANGZHOU 310027
PEOPLE'S REPUBLIC OF CHINA
E-MAIL: stazlx@zju.edu.cn

F. HU
DEPARTMENT OF STATISTICS
UNIVERSITY OF VIRGINA
113 HALSEY HALL
CHARLOTTESVILLE, VIRGINIA 22904-4135
USA
E-MAIL: fh6e@virginia.edu

S. H. CHEUNG
DEPARTMENT OF STATISTICS
THE CHINESE UNIVERSITY OF HONG KONG
SHATIN, N.T., HONG KONG
PEOPLE'S REPUBLIC OF CHINA
E-MAIL: shcheung@sta.cuhk.edu.hk

W. S. CHAN
DEPARTMENT OF FINANCE
THE CHINESE UNIVERSITY OF HONG KONG
SHATIN, N.T., HONG KONG
PEOPLE'S REPUBLIC OF CHINA
E-MAIL: chanws@cuhk.edu.hk