Urn Models:
Towards Applications in Biology

Master Thesis in Biostatistics (STA495)

by

Michael Hediger
s11714185

supervised by

Prof. Dr. Jean Bertoin

Zürich, December 3, 2019

Urn Models:
Towards Applications in Biology

Michael Hediger

Zürich, December 3, 2019

Contents

Chapter 1

Introduction

An urn model, in its general form, refers to a system of one ore more urns (a collection of urns), containing balls of different colors. The system is subject to time, that is, a given collection of urns evolves over time and its appearance, at a given point in time, is governed by certain drawing and returning rules. An urn model can be seen as an abstract tool, capable to reflect probabilistic phenomenons. For instance, Bernoulli (1768) introduced an urn model where at a given, discrete, point in time, one considers a collection of two urns, $\mathcal{A}$ and $\mathcal{B}$. Associated to the collection are balls of two colors, white and black. At the beginning, the collection is in balance and completely separated: Both urns start out with an equal number of balls and either of the two urns starts all black and the other all withe. The evolution of the two urns is described by exchanging balls between the urns, one at a time: For a given point in time, one draws a ball (at random, each ball equally likely) from, say, urn $\mathcal{A}$ and the withdrawn ball is put into urn $\mathcal{B}$. Similarly, associated to the same step in time, a ball from urn $\mathcal{B}$ is drawn and placed into urn $\mathcal{A}$. Bernoulli was interested in the following question: What is the expected number of white balls, in urn $\mathcal{A}$, after a certain number of balls have been exchanged? The above system of urns is known as an Ehrenfest urn model – consider also Johnson and Kotz (1977, chap. 4.8) for a detailed introduction. Such systems of urns have applications as models for the mixing of particals in two connected gas chambers. As we will see later, the above model can be formalized with a single urn, instead of two.

The Ehrenfest urn models belong to the class of so called Pólya urn models (or just Pólya urns), which were popularized by the work of Eggenberger and Pólya (1923). Pólya urns are built upon a single urn, containing balls of different colors, which evolves, in discrete or continuous time, under specific drawing and returning rules: At a given drawing step, a single ball (or a bag of balls, a multiset of balls), is sampled (uniformly at random) from the urn and depending on the observed color (or depending on the observed colors within the bag of balls), balls of various colors are returned to the urn. Pólya urns, which, at a given drawing step, are characterized by single ball draws, are called simple Pólya urns. Tightened to Pólya urns is the so-called replacement scheme (the Pólya urn scheme), which characterizes the drawing and returning rules. The scheme is represented by a matrix with integer coefficients, where the row index represents the colors of the ball (or the colors of the balls associated to the given bags of balls) drawn from the urn and the column index represents the colors of the balls added to the urn.

In their original work, Eggenberger and Pólya (1923) have assumed that the urn can contain balls of two different colors. At each given sampling step, the chosen ball, of given color, is returned to the urn, together with a certain number of balls of the same color. This process is repeated throughout the evolution of the urn. An urn, which evolves according to the above sheme is called a Pólya-Eggenberger urn. Tightened to the Pólya-Eggenberger urn is the Pólya distribution: It is the distribution of the random variable, which describes, after a certain number of ball draws, the number of times a ball of a certain color is chosen. The Pólya-Eggenberger urn was introduced as a modal of contagion. A comprehensive treatment is given in Johnson and Kotz (1977, chap. 4).

On the level of the individual colors, associated to general Pólya urns, one central interest is on the abundance of a given color after a long period of sampling sessions. In the case of simple Pólya urns, Smythe (1996) has established a law of large numbers and a central limit theorem result for the number of balls of a certain color under certain conditions on the given urn sheme. These conditions comprise the eigenvalues and eigenvectors, associated to the replacement matrix. Actually, for simple Pólya urns, law of large numbers and central limit theorem results go back to the work of Athreya and Karlin (1968). These early results were established using poissonization-depoissonization, whereas Smythe (1996) has introduced the class of extended Pólya urn schemes and followed a martingale approach for the establishment of the urn asymptotics. For more general Pólya urns, that is, for systems of urns, where at a given sampling point in the evolution of the urn, multiple ball draws are allowed, similar general asymptotic results are not yet established (see also Mahmoud, 2013 and Konzem and Mahmoud, 2016 for some recent work on more general Pólya urns).

The present thesis aims to present urn models as a motivational tool to understand probabilistic problems. The problems might arise in several different contexts, such as, for example, particle systems in physics, social networks in sociology, or population genetics in biology. A decent introduction to urn models and their applications is given in the book by Johnson and Kotz (1977), or more recently in Mahmoud (2008). In the following chapter, I will summarize some main ideas, results, and applications tightened to urn models. The chapter should serve as a motivation which attempts to demonstrate the significance of urn models in probability theory. By the end of chapter 2, I will introduce the so-called Hoppe's urn model (Hoppe, 1984), which has its motivation in Ewens sampling formula (Ewens, 1972), and its application in population genetics. Finally, in chapter 3, I take up the context of genetics and introduce a non-simple, 3-color, Pólya urn which describes, in a simplified setting, the genotypic composition of diploid organisms. As a main result of chapter 3, I will provide an explicit formula for the expected difference of the number of homozygous wild-type and homozygous mutant individuals, after a given number of mating sessions.

Chapter 2

Motivational Results

2.1 The Martingale central limit Theorem

In this first part of the chapter, we will consider a martingale point of view on the classical central limit theorem (CLT). In the classical setting, one assumes an i.i.d. sequence of random variables Y_j, $j = 1, 2, \dots$, on a probability space $(\Omega,\, \mathcal{A},\, P)$, and considers the sum $\sum_{j=1}^{n} Y_j$ for $n \geq 1$. Then, if we assume that $Y_1 \in L^2(\Omega)$ (that is $\mathrm{E}(|Y_1|) < \infty$ and $\mathrm{Var}(Y_1) = \sigma^2 < \infty$) and further write $\mathrm{E}(Y_1) = \mu$, we arrive at the classical result:

$$\frac{\sum_{j=1}^{n} Y_j - n\mu}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, \sigma^2),$$

where $\xrightarrow{d}$ indicates that the convergence is in distribution. In the following, we will go through some objects and assumptions which will be part of the main conditions which are subject to the martingale central limit theorem. My elaborations in the present section are mostly based on chapter 3 in Hall and Heyde (1980).

***Martingale Arrays*:** Here, we shall discuss a limit theorem for the sum of martingale differences. Explicitly, consider a probability space $(\Omega,\, \mathcal{F},\, P)$ and a collection of sub-σ-fields $\{\mathcal{F}_{n,i},\, 1 \leq i \leq k_n,\, n \geq 1\}$, where $k_n \uparrow \infty$, as $n \to \infty$. On the given probability space we will assume that for any given $n \geq 1$, $\mathcal{F}_{n,i} \subseteq \mathcal{F}_{n,i+1}$. Hence, for a given $n \geq 1$, the sub-σ-fields act as a filtration on the probability space introduced above. Now, on the same probability space, let, for any given $n \geq 1$, $\{S_{n,i} = \sum_{j=1}^{i} X_{n,j},\, \mathcal{F}_{n,i},\, 1 \leq i \leq k_n\}$ be a zero-mean, square-integrable, martingale with martingale differences $X_{n,i}$ ($S_{n,0} = 0$). That is, for any given $n \geq 1$,

$$\mathrm{E}(|X_{n,i}|) < \infty, \quad \text{and} \quad \mathrm{E}(X_{n,i}|\mathcal{F}_{n,i-1}) = 0 \text{ a.s.}, \quad 1 \leq i \leq k_n.$$

In conclusion, the collection $\{S_{n,i} = \sum_{j=1}^{i} X_{n,j},\, \mathcal{F}_{n,i},\, 1 \leq i \leq k_n,\, n \geq 1\}$, as introduced above, is called a martingale array. Notice that in the classical CLT setting, as given earlier, where one has, for any given $n \geq 1$ and for any instance $1 \leq i \leq n$, a collection of i.i.d. random variables $\{Y_j,\, 1 \leq j \leq i\}$, we can set $\mathcal{F}_{n,i} = \mathcal{A}_i$, where $\mathcal{A}_i$ is the σ-field generated by the random variables $Y_1,\, Y_2,\, \dots,\, Y_i$, and treat the collection $\{n^{-1/2}(\sum_{j=1}^{i} Y_j - i\mu),\, \mathcal{A}_i,\, 1 \leq i \leq n,\, n \geq 1\}$ as a martingale array with $k_n = n$ and martingale differences $n^{-1/2}(Y_i - i\mu)$.

***The Conditional Lindeberg Condition*:** As above, let us assume a zero-mean, square-integrable, martingale array $\{S_{n,i} = \sum_{j=1}^{i} X_{n,j},\, \mathcal{F}_{n,i},\, 1 \leq i \leq k_n,\, n \geq 1\}$ with martingale differences $X_{n,i}$. Further, write $V_{n,i}^2 = \sum_{j=1}^{i} \mathrm{E}(X_{n,j}^2|\mathcal{F}_{n,j-1})$, as the conditional variance for the martingale $S_{n,i}$ and call $U_{n,i}^2 = \sum_{j=1}^{i} X_{n,j}^2$ the squared variation term of $S_{n,i}$. The first sufficient condition for establishing the martingale CLT is named after the Finnish mathematician Jarl Waldemar Lindeberg and is called the conditional Lindeberg condition. It reads as follows:

$$\text{For all } \epsilon > 0, \quad \sum_{i=1}^{k_n} \mathrm{E}(X_{n,i}^2 \mathbb{1}_{\{|X_{n,i}|>\epsilon\}} | \mathcal{F}_{n,i-1}) \xrightarrow{P} 0, \quad \text{as } n \to \infty \tag{2.1}$$

where $\xrightarrow{P}$ denotes the convergence in probability. As we will see later, under some conditions on the conditional variance, (2.1) is actually equivalent to the so called weak Lindeberg condition:

$$\text{For all } \epsilon > 0, \quad \sum_{i=1}^{k_n} X_{n,i}^2 \mathbb{1}_{\{|X_{n,i}|>\epsilon\}} \xrightarrow{P} 0, \quad \text{as } n \to \infty. \tag{2.2}$$

But, since we have $P(\max_{1\le i\le k_n} |X_{n,i}| > \epsilon) = P(\sum_{i=1}^{k_n} X_{n,i}^2 \mathbb{1}_{\{|X_{n,i}|>\epsilon\}} > \epsilon^2)$, (2.2) is actually equivalent to

$$\max_{1\le i\le k_n} |X_{n,i}| \xrightarrow{P} 0, \quad \text{as } n \to \infty, \tag{2.3}$$

which can be seen as an assumption of asymptotic negligibility.

***The conditional Variance condition*:** In the context of martingale differences, and with the same notation as introduced above, the conditional variance can be introduced by the Doob decomposition (Doob, 1953, p. 297) of the submartingale array $\{S_{n,i}^2, \mathcal{F}_{n,i}, 1 \le i \le k_n, n \ge 1\}$. We can write the unique decomposition as follows: Fix $n \ge 1$ and write,

$$S_{n,i}^2 = M_{n,i} + A_{n,i}, \quad 1 \le i \le k_n$$

where the collection $\{M_{n,i}, \mathcal{F}_{n,i}, 1 \le i \le k_n\}$ is a regular martingale and $\{A_{n,i}, 1 \le i \le k_n\}$ is a non-decreasing (non-decreasing in $1 \le i \le k_n$) sequence of non-negative random variables such that for each $n \ge 1$, $A_{n,i}$ is $\mathcal{F}_{n,i-1}$ measurable. But then, since we have,

$$A_{n,i} - A_{n,i-1} = \mathrm{E}(X_{n,i}^2 | \mathcal{F}_{n,i-1}),$$

we can conclude that $V_{n,i}^2 = A_{n,i}$. Hence, we can view, for a given $n \ge 1$, the conditional variance as the non-negative, non-decreasing, process in the decomposition of the submartingale $\{S_{n,i}^2, \mathcal{F}_{n,i}, 1 \le i \le k_n\}$. The condition, imposed on the conditional variance $V_{n,i}^2$, sufficient for the martingale CLT, can be given as follows:

$$V_{n,k_n}^2 \xrightarrow{P} \sigma^2, \quad \text{as } n \to \infty, \tag{2.4}$$

where σ^2 is some non-negative constant.

***The relationship between the conditional variance and the squared variation*:** The following theorem gives a statement about the asymptotic relationship between $V_{n,i}^2$ and $U_{n,i}^2$. It is part of a larger theorem, proved in Hall and Heyde (1980, Theorem 2.23, p. 44).

Theorem 2.1. Let, as above, $\{S_{n,i}, \mathcal{F}_{n,i}, 1 \le i \le k_n, n \ge 1\}$ be a zero-mean, square-integrable, martingale array with martingale differences $X_{n,i}$. Then let us suppose that the conditional variances, V_{n,k_n}^2, are tight, that is:

$$\sup_{n\ge 1} P(V_{n,k_n}^2 > \lambda) \to 0 \quad \text{as } \lambda \to \infty \text{ and } n \to \infty. \tag{2.5}$$

Additionally, let us assume that the conditional Lindeberg condition, condition (2.1), is satisfied. Then:

$$\max_{1\le i\le k_n} |U_{n,i}^2 - V_{n,i}^2| \xrightarrow{P} 0, \quad \text{as } n \to \infty. \tag{2.6}$$

As a follow-up of the above thoughts, we are now ready to state the martingale CLT:

Theorem 2.2 (*The Martingale CLT*)**.** Let $\{S_{n,i},\, \mathcal{F}_{n,i},\, 1 \leq i \leq k_n,\, n \geq 1\}$ be a zero-mean, square-integrable, martingale array with martingale differences $X_{n,i}$, and let σ^2 be some positive constant. Assume that the conditional Lindeberg condition,

$$\text{For all } \epsilon > 0, \quad \sum_{i=1}^{k_n} \mathrm{E}(X_{n,i}^2 \mathbb{1}_{\{|X_{n,i}|>\epsilon\}} | \mathcal{F}_{n,i-1}) \xrightarrow{P} 0, \quad \text{as } n \to \infty,$$

as well as the conditional variance condition

$$V_{n,k_n}^2 \xrightarrow{P} \sigma^2, \quad \text{as } n \to \infty,$$

are met. Additionally, let us also assume that

$$\mathrm{E}(\max\nolimits_{1\leq i\leq k_n} X_{n,i}^2) \quad \text{is bounded in } n, \tag{2.7}$$

and that the sub-σ-fields are nested, that is:

$$\mathcal{F}_{n,i} \subseteq \mathcal{F}_{n+1,i} \quad \text{for } 1 \leq i \leq k_n,\, n \geq 1. \tag{2.8}$$

Then we can conclude that,

$$S_{n,k_n} = \sum_{i=1}^{k_n} X_{n,i} \xrightarrow{d} \mathcal{N}(0, \sigma^2), \quad \text{as } n \to \infty.$$

Remark 2.1. The above statement, theorem 2.2, should be considered as a possible formulation of the martingale CLT – it is stated as a corollary in Hall and Heyde (1980, Corallary 3.1, p. 58). Theorem 2.2 is a reformulation of a theorem which is referred to as the martingale CLT in Hall and Heyde (1980, Theorem 3.2, p. 58) – it depends on the assumption of asymptotic negligibility (2.3) and a convergence in probability for the squared variation term, $U_{n,i}^2$. The result reads as follows:

Theorem 2.3. Consider a zero-mean, square-integrable, martingale array $\{S_{n,i},\, \mathcal{F}_{n,i},\, 1 \leq i \leq k_n,\, n \geq 1\}$ with martingale differences $X_{n,i}$ and further let σ^2 be some positive constant. Suppose that

$$U_{n,k_n}^2 = \sum_{j=1}^{k_n} X_i^2 \xrightarrow{P} \sigma^2, \quad \text{as } n \to \infty, \tag{2.9}$$

$$\mathrm{E}(\max\nolimits_{1\leq i\leq k_n} X_{n,i}^2) \quad \text{is bounded in } n, \tag{2.10}$$

and $\{X_{n,i},\, 1 \leq i \leq k_n,\, n \geq 1\}$ is asymptotic negligible, that is (2.3) is verified, and the sub-σ-fields are nested in $n \geq 1$, hence also condition (2.8) is met. Then one arrives, again, at the normal limit – hence one gets that $S_{n,k_n} \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, as $n \to \infty$.

We can actually quickly sketch some arguments which lead to the conclusion that our conditions, stated in theorem 2.2, lead to the conclusion of theorem 2.3: First of all, under the assumption of theorem 2.2 we have that

$$V_{n,k_n}^2 \xrightarrow{P} \sigma^2, \quad \text{as } n \to \infty,$$

and hence we can verify that the conditional variances, V_{n,k_n}^2, are tight. Now, since we know that (2.1) is fulfilled, we actually meet the assumptions of theorem 2.1. As a conclusion, we have that

$$\max_{1\leq i\leq k_n} |U_{n,i}^2 - V_{n,i}^2| \xrightarrow{P} 0, \quad \text{as } n \to \infty,$$

and hence we verify (2.9). To verify (2.3) we can notice that, both together, the conditional Lindeberg condition and the convergence in probability, of V_{n,k_n}^2 to σ^2, give the weak Lindeberg condition which is actually equivalent to the assumption of asymptotic negligibility:

$$\max_{1\leq i\leq k_n} |X_{n,i}| \xrightarrow{P} 0, \quad \text{as } n \to \infty, \tag{2.11}$$

and hence (2.3).

2.2 Urn Models

An urn model is built upon a collection of urns, where the urns can contain balls of different colors. One further assumes that each given ball (or each given multiset of balls, bag of balls) within a given urn is chosen equally likely. In many applications, the urn model is considered as a mathematical model which governs the workflow of a certain phenomenon. Of course the phenomenon is specified within the given context and the range of different contexts is far-reaching. For a general introduction to urn models and their applications one can consider the book by Johnson and Kotz (1977). Given a certain context, the workflow of the phenomenon of interest is represented by the application of drawing-and-adding operations, associated with balls or bags of balls, to the the urns (or to the urn). The drawing and addition of balls are done according to certain rules and these rules are determined by the given context and phenomenon one aims to describe. Here, in the present work, we are aiming to describe the usage of urn models within the context of biology. Before we do so, we will get to know some classical, historical urn problems, as for example the occupancy problem, or the coupon collector's problem. Further, as the main object of this work, we will get to know some specific classes of urns, called Pólya urns, which will be important for our language used in the context of biology.

2.2.1 Some classical Urn Problems

An overview of some classical urn problems is found in chapter 2 of Mahmoud (2008). These classical urn problems can be brought into different kind of contexts and play an important role in a broad kind of applications. To get a feeling for the usage of urn models, I will give a short review here. For a rigorous justification of the given results, I refer to the source just given above.

Example 2.1 (*Ballot Problems*)**.** The analogy in this example is the progress of an election or a vote. In a simple example, one can suppose that two candidates, $\mathcal{A}$ and $\mathcal{B}$, are running against each other. Now, the scenario is the following: We fix the number of total votes for $\mathcal{A}$ and $\mathcal{B}$ to be m and n respectively. Further we suppose that $m \geq n$ (hence $\mathcal{A}$ has more votes or is tied). This fixes the situation at the end of the election and of course, one could also assume the mirrored outcome – $\mathcal{B}$ wins or is tied. During the election, we can imagine an urn which will be, by the end of the vote, filled with a total of $m+n$ balls of two different colors. While the vote is in progress, the votes for $\mathcal{A}$ and $\mathcal{B}$ are counted. In the language of the urn: While the urn is filled with balls of different colors, the urn is depleted ball by ball and the respective colors are noted. The question associated to the ballot problem is the following: *What is the probability that $\mathcal{A}$ stays ahead of $\mathcal{B}$ throughout the counting?* The answer to the question can be obtained by conditioning on who receives the last vote and writing down a simple recurrence relationship – In conclusion, we have that

$$P(\{\mathcal{A} \text{ stays ahead of } \mathcal{B} \text{ throughout the counting}\}) = \frac{m-n}{m+n}.$$

Example 2.2 (*Occupancy Problems*)**.** In this example, the urn analogy goes as follows: One has a set of m urns and a supply of balls, arriving one after each other. Each ball is assigned, independently and uniformly at random, to an urn. After a supply of $n \geq 1$ balls one could ask the following question: *What is the probability that no urn will be empty?* An answer is obtained by, first considering the probability that a particular urn $i \in \{1, 2, \ldots, m\}$ stays empty, and making use of the independence assumption. Then, one can extend the argument to any of the m urns and again, after applying independence, one can arrive at the probability that not a single ball hits any of the m urns after n arrivals of balls. In the end, one can consider the complementary event and arrives at:

$$P(\{\text{Not a single urn will be empty after an arrival of } n \text{ balls}\}) = \sum_{i=1}^{m} (-1)^i \binom{m}{i} \left(1 - \frac{i}{m}\right)^n .$$

For a more detailed discussion on occupancy problems and further interesting applications and examples I would like to refer to Johnson and Kotz (1977, chap. 3).

Example 2.3 (*Coupon Collector's Problems*)**.** This problem is associated with coupon collection. In the following we will assume a set of n different coupons. A simple urn analogy goes as follows: The n coupons are distinguished as n types of balls of different colors in one single urn. From this urn, one samples, at random, with replacement. The sampling continues until one has observed all the n different colors. This actually reflects the question of interest in this scenario: *How soon does a certain collector get all different coupons?* If one defines X_n as the random variable which gives the number of draws form the urn until a set of balls with n different colors is observed, one can get a result for the expected number of draws until all the coupons are collected – hence an expression for $\mathrm{E}(X_n)$. The expression is obtained as follows: Write,

$$X_n = Y_1 + Y_2 + \cdots + Y_n,$$

where $Y_i \sim \text{Geo}\left({}^{(n-i+1)}/_{n}\right)$, hence the collectors attempt to get the ith coupon follows a geometric random variable. The probability of success reflects the fact that up to the ith coupon the collector has already obtained $i-1$ different coupons and his chance to get the ith among the n coupons is given by ${}^{(n-i+1)}/_{n}$. Now, we get that

$$\mathrm{E}(X_n) = \mathrm{E}(Y_1) + \mathrm{E}(Y_2) + \cdots + \mathrm{E}(Y_n) = n \sum_{k=1}^{n} \frac{1}{k},$$

which shows the asymptotic equivalence (as $n \to \infty$), of $\mathrm{E}(X_n)$, to $n \log(n)$. Actually, there is a connection to the occupancy problem: The event $\{X_n \leq x\}$ indicates that all the n different coupons have appeared before x draws. But this is just as one would have an arrival of x balls to n urns and not a single urn will be empty. Hence, we have

$$P(\{X_n \leq x\}) = \sum_{i=1}^{n} (-1)^i \binom{n}{i} \left(1 - \frac{i}{n}\right)^x, \qquad x = n, n+1, \ldots$$

The above examples should serve as a demonstration of how several interesting far-reaching problems can be brought into analogy with urn models. We will now consider a certain subclass of urn models – the Pólya urns.

2.2.2 Pólya Urn Models

The Pólya urn models are a subclass of urn models in the following sense: One considers the evolution of a *single urn*, containing balls of different colors. At each stage, in the evolution of the urn, a general method of ball drawing and replacement is applied. In the following I will first present the definition of the simple k-color Pólya urn, whereby simple refers to the fact that one draws, at each drawing step, one ball at a time. Generalizations, where at each step a bag of ball is drawn, are considered as a follow-up.

Definition 2.1 (*The simple k-Color Pólya Urn*)**.** A simple k-color Pólya urn is an urn containing up to k different colors. The urn starts with a certain number of balls of each color – the initial composition of colors. This initial condition can be represented by a vector of initial colors $\mathbf{X_0}$ with entries representing the initial numbers of balls for each color $X_0^{(1)}$, $X_0^{(2)}$, $\ldots$, $X_0^{(k)}$. The urn evolves in discrete time steps and the points in time where a ball is withdrawn are called *epochs*. At each epoch a *single* ball is chosen uniformly at random and the color of the withdrawn ball is observed. The chosen ball is returned to the urn (sample with replacement). If the color of the drawn ball was i, $i \in \{1,2,\ldots,k\}$, then A_{ij} balls of color j, $j \in \{1,2,\ldots,k\}$, are placed in the urn, where the entries A_{ij} can be deterministic (positive or negative) or random. In case of random entries, the A_{ij}'s are considered as discrete random variables on the set of integers – when we pick a ball of color i, we add to the urn an independent copy of A_{i1} balls of color 1, A_{i2} balls of color 2, and so on, up to the addition of A_{ik} balls of color k (by independence, we mean, independent of the given status of the urn and the past history of the urn). Following the initial conditions, the number of balls of color r, $r \in \{1,2,\ldots,k\}$, at time point n can be written as $X_n^{(r)}$.

Definition 2.2 (*The simple k-Color Pólya Urn Scheme*)**.** The simple k-color Pólya urn scheme is the $k \times k$ matrix

$$\begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,k} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ A_{k,1} & A_{k,2} & \cdots & A_{k,k} \end{pmatrix},$$

with entries A_{ij}, $i,j \in \{1,2,\ldots,k\}$, which represents the ball drawing and replacement method for a simple k-color Pólya Urn with initial conditions $\mathbf{X_0}$. The row-index specifies the color of the withdrawn ball and the column-index specifies the color of the ball which is added to the urn. The above matrix is often referred to as the ball addition matrix, generator, or just the urn scheme.

Remark 2.2 (*k-Color Pólya Urn Schemes*)**.** Generalizations are given in terms of the nature of the time epochs (allowing for continuous time evolution of the urn) and also the number of balls withdrawn at each time epoch (allowing for multiset ball draws). In such cases, if at each time epoch, the number of balls to draw from the urn is s (again, bags of size s are uniformly at random, and with exception of the nature of the time epochs and the number of balls to draw, the notation and thoughts associated with definition 2.1 remain the same), there are $\binom{k+s-1}{s}$ possible drawing configurations of size s. Hence, together with initial conditions, one can reformulate a general drawing and replacement matrix (not necessarily a square matrix) as follows:

$$\mathbf{A} = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,k} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ A_{\binom{k+s-1}{s},1} & A_{\binom{k+s-1}{s},2} & \cdots & A_{\binom{k+s-1}{s},k} \end{pmatrix}.$$

An introduction to the case, where $s = 2$ is given in Mahmoud (2013) and the general case is introduced in Konzem and Mahmoud (2016). In general, if I refer to a Pólya urn scheme (or a Pólya urn with associated urn scheme), I have in mind a general, tenable (consider the upcoming remark), scheme as given with $\mathbf{A}$, where, as in the simple case, the row-index specifies the color of the withdrawn ball and the column-index specifies the color of the ball which is added to the urn.

Remark 2.3 (Tenability)**.** In the above description of Pólya urns, we have assumed that the ball drawing and replacement scheme is always applicable, otherwise further restrictions would be necessary. In this scenario, "always" means that the urn scheme is applicable on any possible stochastic path. Such an urn is called a tenable urn. In simple words: A tenable urn is an urn where one can always draw and replace balls according to the rules given by the urn scheme and one will never end up in a scenario where one tries to subtract more balls of a color than are left in the urn. Hence, the k-color Pólya urns are considered as a class of tenable urns. For future references: If one presents a certain urn scheme $\mathbf{A}$, and one aims to classify this scheme as a Pólya urn scheme, one has to assume that the given entries of $\mathbf{A}$ are such that the associated urn model is tenable given the initial condition. Tenability is of particular importance when one wants to develop asymptotic theories – the ball drawing and replacing workflow should persist over time. It is clear that, under the assumption of non-negative entries A_{ij}, the urn will always be tenable, given any possible starting condition. Tenability issues arise when the ball replacement matrix contains negative entries. For an illustration, let us consider two examples:

Example 2.4. The urn which underlies the ballot problem (see example 2.1) is one instance of an untenable urn. In the given scenario, we have two players $\mathcal{A}$ and $\mathcal{B}$, running against each other with a total number of m and n votes, respectively. In this case, we can think of a 2-color urn, containing an initial number of m white (W) balls (votes for $\mathcal{A}$) and an initial number of n blue (B) balls (votes for $\mathcal{B}$) – hence an urn containing a total number of $n + m$ balls. Now, during the count, the balls are taken out, one ball at a time (without replacement) and the color of the ball is recorded – a scheme is given by,

$$\begin{array}{c c} & \begin{array}{cc} W & B \end{array} \\ \begin{array}{c} W \\ B \end{array} & \begin{pmatrix} -1 & 0 \\ 0 & -1 \end{pmatrix}, \end{array}$$

where, as in definition 2.2, the labeling of rows and colors specify, respectively, which color to draw from, or to add to, the urn. Now, since the urn is depleted after $m + n$ draws, the urn is not tenable under any given starting condition.

Example 2.5. For an example of an not necessarily tenable urn scheme, associated to an urn containing white (W) and blue (B) balls, let us consider the following scheme,

$$\begin{array}{c c} & \begin{array}{cc} W & B \end{array} \\ \begin{array}{c} W \\ B \end{array} & \begin{pmatrix} -1 & -X \\ 2 & 3 \end{pmatrix}, \end{array}$$

where X is a Bernoulli random variable which translates the outcome (head or tail) of a fair coin flip:

$$X = \begin{cases} 1 & \text{if the coin shows head} \\ 0 & \text{otherwise} \end{cases}.$$

As in the Pólya urn case, one aims to sample balls with replacement. If one supposes that the initial conditions are such that there are 2 white balls and 1 blue ball in the urn, the urn gets stuck on the stochastic path where one draws twice a white ball. Hence, given the initial conditions above, drawing and replacing balls according to the above scheme is not feasible on any stochastic path.

Sufficient conditions for tenability in 2-color, simple, deterministic, urn schemes as

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix},$$

which operate as in definition 2.2, that is, the row-index specifies the color of the withdrawn ball and the column-index specifies the color of the ball which is added to the urn, are formulated in Mahmoud (2008, chap. 3). There, it can be seen that the tenability conditions can be categorized according to the number of negative entries in the given urn scheme. In the more general setting, of k-color Pólya urns, sufficient conditions for tenability, are formulated in Konzem and Mahmoud (2016).

2.2.3 Some classical Pólya Urn Schemes

Having discussed the appearance of Pólya urns it is helpful to consider some examples. One early treatment of the subject was done by Eggenberger and Pólya (1923). Their studies have main interest in the 2×2 scheme known as the Pólya and Eggenberger urn. Several other 2-color Pólya urns, as for example the *Bernard Friedman's Urn*, the *Bagchi-Pal Urn* or the *Ehrenfest Urn* were introduced later. Here, I would like to summarize some of the results – for a more comprehensive reading, I would like to refer to Mahmoud (2008, chap. 3). In the upcoming examples of 2-color, simple, Pólya urn schemes, I will assume, for illustration purposes, that the urn is populated with white (W) and blue (B) balls.

Example 2.6 (*The Ehrenfest Urn*)**.** Associated to the Ehrenfest urn model, composed of white and blue balls, is the scheme

$$\begin{array}{cc} & \begin{array}{cc} W & B \end{array} \\ \begin{array}{c} W \\ B \end{array} & \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} \end{array}. \tag{2.12}$$

The specialty of this urn scheme is that the total number of balls within the urn remains constant throughout time. Whenever a ball, of a given color, is removed from the urn, a ball of the opposite color joins the urn – one talks about a zero-balanced urn. This urn model can be used as an analogy to describe the mixing of particles in two connected gas chambers, $\mathcal{A}$ and $\mathcal{B}$. The phenomenon can be pictured as follows: At the initial state we have a number of particles in each of the two chambers. The chambers are disconnected by some partition. At time zero, the partition is removed, creating one connected environment for both populations of particles. The mixing of the particles is now described by the act of switching from one chamber to the other. This switch can be characterized by a ball replacement, that is a change of color: We can think about an initial population of white and blue balls, where white and blue balls represent, respectively, particles from chambers $\mathcal{A}$ and $\mathcal{B}$. A change of color, is represented by a particle which moves from chamber $\mathcal{A}$ to chamber $\mathcal{B}$. In short terms, going back to the language of balls in an urn: The total number of balls, call it τ, remains constant throughout time but the number of balls of one particular color, say white, can take values in $\{0, 1, 2 \dots, \tau\}$. For illustration purposes let us focus on the number of white balls. Suppose that the initial population size of white balls, call it W_0, is represented as $W_0 = \text{Bin}(\tau, 1/2)$ and the initial population of blue balls is given by $B_0 = \tau - W_0$. Then if we let W_n to be the number of white balls in the urn (connected chambers) after n draws we arrive at

$$W_n \xrightarrow{d} \text{Bin}\left(\tau, \frac{1}{2}\right).$$

The above statement is stated and proved (as theorem 3.6) in Mahmoud (2008, p. 63). I can provide a quick sketch of the arguments: Notice that the process $\{W_n,\, n \geq 1\}$ can be regarded as a irreducible Markov chain on the finite state space $\{0,\, 1,\, 2\ldots,\tau\}$. Hence, we are working with a positive recurrent Markov chain with transition matrix given by

$$\begin{cases} \mathbf{M}(k,\, k+1) = \frac{\tau-k}{\tau} & 0 \leq k \leq \tau-1 \\ \mathbf{M}(k,\, k-1) = \frac{k}{\tau} & 1 \leq k \leq \tau \end{cases}.$$

As the Markov chain is positive recurrent, we are aware of a unique, stationary distribution, $\mu = \mu\mathbf{M}$, associated to the Markov chain. The measure μ, is reversible, and hence stationary, if and only if it satisfies

$$\mu(k)\frac{\tau-k}{\tau} = \mu(k+1)\frac{k+1}{\tau}$$

for any $0 \leq k \leq \tau - 1$. Hence, the unique stationary distribution on the state space satisfies

$$\mu(k+1) = \frac{\tau-k}{k+1}\mu(k),$$

which becomes iterated to

$$\mu(k) = \frac{1}{2^\tau}\binom{\tau}{k}.$$

As the Markov chain starts out with the above stationary measure, that is W_0 is distributed as μ, we have with $\mu = \mu\mathbf{M}^n$, that for any given $n \geq 1$,

$$P(W_n = k) = \frac{1}{2^\tau}\binom{\tau}{k},$$

and hence, the distribution of W_n is given by μ throughout the evolution of the urn.

Remark 2.4. By embedding the urn scheme above into continuous time – explicitly, by modeling the ball drawing and replacement by a *Poisson Process* one can actually further generalize the above result. Such an embedding into continuous time is called poissonization (consider Mahmoud (2008, chap. 4) for a good overview). For a general discussion on the Ehrenfest process (embedded in continuous time), I refer to Balaji *et al.* (2006).

Remark 2.5. Under the given conditions above, hence under the assumption that $W_0 = \mathrm{Bin}(\tau, 1/2)$ and $B_0 = \tau - W_0$, the Ehrenfest urn is a tenable urn. Notation wise, let us write A_{11}, A_{12}, A_{21}, and A_{22} for the entries associated with the Ehrenfest scheme given in (2.12). Now, the scheme has two negative entries and the conditions sufficient for tenability are given as follows:

- W_0 and the entry A_{21} are both multiples of $|A_{11}|$.
- B_0 and the entry A_{12} are both multiples of $|A_{22}|$.
- Both, A_{12} and A_{21} are positive.

For a reference one can consider Mahmoud (2008, p. 49).

Example 2.7 (*The Pólya-Eggenberger Urn*)**.** Pólya and Eggenberger were interested in an urn model with 2-color, deterministic, scheme,

$$\begin{array}{c|cc} & W & B \\ \hline W & s & 0 \\ B & 0 & s \end{array},$$

where s is some positive integer, and again, it is assumed that the urn is populated with white and blue balls. Let us further write τ_n for the total number of balls in the urn after n draws and again use W_n and B_n for the number of white and blue balls, respectively, in the urn after n draws. To complete the notation set-up, let us introduce the notation of rising factorials.

Definition 2.3 (*Rising Factorials of Order k*)**.** For a complex number x, the rising factorial of order k is defined as

$$\langle x \rangle_k = x(x+1)(x+2)\cdot\ldots\cdot(x+k-1) = \prod_{i=1}^{k}(x+i-1) \tag{2.13}$$

The upcoming results go back to the original work done in Eggenberger and Pólya (1923), a comprehensive treatment is given in Mahmoud (2008, chap. 3).

Proposition 2.1. Let $\widetilde{W}_n$ be the number of white ball draws in the Pólya and Eggenberger urn after n draws. Then,

$$P(\widetilde{W}_n = k) = \binom{n}{k}\frac{\langle {}^{W_0}\!/_{s} \rangle_k \langle {}^{B_0}\!/_{s} \rangle_{n-k}}{\langle {}^{\tau_0}\!/_{s} \rangle_n}, \qquad 0 \le k \le n.$$

Proposition 2.2. Let $\widetilde{W}_n$ be the number of white ball draws in the Pólya and Eggenberger urn after n draws. Then,

$$\frac{\widetilde{W}_n}{n} \xrightarrow{d} \text{Beta}\left(\frac{W_0}{s}, \frac{W_0}{s}\right).$$

Example 2.8 (*The Bagchi-Pal Urn*)**.** A more general deterministic 2-color Pólya urn scheme is considered by Bagchi and Pal (1985):

$$\begin{array}{c|cc} & W & B \\ \hline W & a & b \\ B & c & d \end{array}, \tag{2.14}$$

given any choice of integers a, b, c, and d which are tenable given the initial conditions (we want a proper Pólya urn scheme) and which further meets the following assumptions:

- One wants a balanced urn condition, that is constant row sums: $a + b = c + d = K$.
- The cases, $b = c = 0$ (Pólya-Eggenberger Urn), $a = c$ (no randomness), and the case where one minor diagonal element is zero are considered as degenerate cases and not considered.

Bagchi and Pal (1985) have developed asymptotic normality for the number of balls of a certain color in an non-degenerate Bagchi-Pal urn with constant row sums:

Theorem 2.4. Let W_n be the number of white balls after n draws from an non-degenerate Bagchi-Pal urn with constant row sums given by K. Then, if we assume that $a-c<{}^{K}\!/_{2}$ we get:

$$\frac{W_n - \frac{cK}{b+c}n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0,\sigma^2),$$

where

$$\mu = \frac{cK}{b+c} \quad \text{and} \quad \sigma^2 = \frac{bcK(a-c)^2}{(b-c)^2(K-2)(a-c)},$$

give the asymptotic mean and variance respectively.

We will actually further discuss the Bagchi-Pal urn in the next section. It turns out that under certain eigenvalue conditions on the Bagchi-Pal scheme given in (2.14), the above result can be established using a martingale approach. For an approach which is based on the convergence of moments one can consider Mahmoud (2008, chap. 3).

Example 2.9 (*The Chen-Wei Urn*)**.** In the following, we are going to consider an example of a balanced, multiset-draw urn scheme, which was introduced by Chen and Wei (2005). The 3×2 scheme,

$$\begin{array}{c c} & \begin{array}{cc} W & B \end{array} \\ \begin{array}{c} \{W,\,W\} \\ \{W,\,B\} \\ \{B,\,B\} \end{array} & \begin{pmatrix} 2C & 0 \\ C & C \\ 0 & 2C \end{pmatrix}, \end{array}$$

with fixed, integer valued, constant $C \geq 1$, is associated to an urn composed of two colors, white (W) and blue (B) and at any epoch, one draws two balls at a time. The three sets $\{W,\,W\}$, $\{W,\,B\}$, and $\{B,\,B\}$ refer to the three possible multiset-draws. If, at any epoch, the pair $\{W,\,W\}$ is drawn, we replace the two balls into the urn and add $2C$ white balls and zero blue balls. Similarly, we add, respectively, C white balls and C blue balls as a result of sampling the pair $\{W,\,B\}$ and we add, respectively, zero white balls and $2C$ blue balls after a draw of $\{B,\,B\}$. The urn scheme is balanced: At any drawing step, we add a total of $2C$ balls to the urn – if at the beginning, the urn is composed of τ_0 white and blue balls, after n draws, there are a total of $\tau_n = \tau_0 + n2C$ balls in the urn. Towards an asymptotic description of the Chen-Wei scheme above, let, respectively, W_n and B_n be the number of white and blue balls after n draws. In this scenario, a draw refers to a one step drawing of two balls at a time, where all possible distinct pairs of balls are equally likely. With respect to the two colors, white and blue, we can define two indicator variables which reflect the multiset-drawing workflow:

$$\mathbb{1}_n^{\{W,\,W\}}, \quad \mathbb{1}_n^{\{W,\,B\}}, \quad \text{and} \quad \mathbb{1}_n^{\{B,\,B\}}.$$

Generally, write $\mathbb{1}_n^{\{X,\,Y\}}$, where $X,\,Y \in \{W,\,B\}$. Then $\mathbb{1}_n^{\{X,\,Y\}}$ indicates the drawing of the multiset $\{X,\,Y\}$ at the nth step. Having set up the above indicators, we have a recurrence for W_n,

$$W_n = W_{n-1} + 2C\mathbb{1}_n^{\{W,\,W\}} + C\mathbb{1}_n^{\{W,\,B\}}.$$

Take $\mathcal{F}_n$, $n \geq 0$, to be the sigma algebra generated by the last n draws. By conditioning on the previous draws, we obtain

$$\mathrm{E}(W_n|\mathcal{F}_{n-1}) = W_{n-1} + 2C\,\mathrm{E}(\mathbb{1}_n^{\{W,\,W\}}|\mathcal{F}_{n-1}) + C\,\mathrm{E}(\mathbb{1}_n^{\{W,\,B\}}|\mathcal{F}_{n-1}). \tag{2.15}$$

But, according to the multiset-drawing, we have

$$\mathrm{E}(\mathbb{1}_n^{\{W,\, W\}} | \mathcal{F}_{n-1}) = \frac{\binom{W_{n-1}}{2}}{\binom{\tau_{n-1}}{2}}, \quad \text{and} \quad \mathrm{E}(\mathbb{1}_n^{\{W,\, B\}} | \mathcal{F}_{n-1}) = \frac{W_{n-1}B_{n-1}}{\binom{\tau_{n-1}}{2}}. \tag{2.16}$$

Plugging both of the relations in (2.16) into (2.15), one obtains, by developing the binomial coefficients and taking expectation,

$$\begin{aligned}\mathrm{E}(W_n) = \mathrm{E}(W_{n-1}) &+ \frac{2C}{\tau_{n-1}(\tau_{n-1}-1)}\,\mathrm{E}(W_{n-1}(W_{n-1}-1)) \\ &+ \frac{2C}{\tau_{n-1}(\tau_{n-1}-1)}\,\mathrm{E}(W_{n-1}B_{n-1}).\end{aligned} \tag{2.17}$$

Now, we can already see the advantage of the structure associated to the Chen-Wei urn: The entry in the first column of the last row is zero, which has, with respect to W_n, saved us some work with nasty quadratic terms coupled to B_n. Further, the second row is set up such that, after setting up a recurrence as above, there is hope to eliminate the quadratic term associated to W_n. One accomplishes this, by using the fact that the urn is also balanced: Replace $B_{n-1} = \tau_{n-1} - W_{n-1}$ in (2.17), to obtain the simple recurrence

$$\mathrm{E}(W_n) = \frac{2Cn + \tau_0}{2Cn + \tau_0 - 2C}\,\mathrm{E}(W_{n-1}).$$

Following the recurrence relation, one finds

$$\begin{aligned}\mathrm{E}(W_n) &= \left(\frac{2Cn+\tau_0}{2Cn+\tau_0-2C}\right)\left(\frac{2Cn+\tau_0-2C}{2Cn+\tau_0-4C}\right)\left(\frac{2Cn+\tau_0-4C}{2Cn+\tau_0-6C}\right)\cdot\ldots\cdot\left(\frac{2C+\tau_0}{\tau_0}\right)\mathrm{E}(W_0) \\ &= \frac{2CW_0}{\tau_0}n + W_0,\end{aligned}$$

and hence:

$$\frac{\mathrm{E}(W_n)}{n} \to \frac{2CW_0}{\tau_0}, \qquad \text{as } n \to \infty.$$

The above thoughts could be applied to a general 3 × 2, multiset-draw urn scheme

$$\begin{array}{c} \\ \{1,\,1\} \\ \{1,\,2\} \\ \{2,\,2\} \end{array} \begin{array}{c} \begin{array}{cc} 1 & 2 \end{array} \\ \begin{pmatrix} a & b \\ c & d \\ e & f \end{pmatrix} \end{array},$$

of two colors indexed as 1 and 2, where the entries could be either constant, or given by integer valued random variables. Notice that the appearance of the Chen-Wei urn was advantageous – in general, the structure of the urn scheme might be less beneficial, which results in more complicated recurrence relations. Further, in the general case, as discussed in remark 2.3, tenability issues have to be taken into account – a discussion is given in Mahmoud (2013).

2.3 Simple Pólya Urns – Eigenvalue and Martingale Theory

As we have seen in Section 2.2.2, a simple Pólya urn is a tenable urn, populated with a set of k different colors where at each time epoch a single ball is drawn from the urn. Here, in this section, we will discuss results on the number of balls of a certain color after having applied a given, simple, urn scheme over and over again – explicitly, if $X_n^{(j)}$ $(j \in \{1,\,2,\,\ldots,\,k\})$ is the number of balls, of color j, after n draws in a simple Pólya urn with given scheme $\mathbf{A}$, we will discuss results on $X_n^{(j)}$ as $n \to \infty$. As we will see, a law of large numbers (LLN) or a central limit theorem for the number of balls of a certain color after n draws is established under certain conditions on the eigenvalues and eigenvectors of the given urn scheme (or the expected urn scheme, in the case of random scheme entries). A basic tool, in the establishment of LLN or CLT asymptotic, is the theory of martingales. We will first consider the case of 2-color, simple, schemes with non-random entries (as the Bagchi-Pal urn) and then we will further generalize to k-colors with simple random generators. I would like to point out that the eigenvalue and martingale theory presented in this section is given in the framework of simple Pólya urn models, where the Pólya urn scheme arises as a square matrix. In the general case (remark 2.2), where one allows for multiset-ball draws, asymptotic results, as introduced in this section, are not yet established. So far, asymptotics, for more general Pólya urns, are discussed under some restricted conditions on either the number of balls to draw, or the general appearance of the non-square urn (for a discussion, consider Mahmoud, 2013 or Konzem and Mahmoud, 2016). We have already encountered a general 2-color Pólya urn, the Chen-Wei urn, in example 2.9, where at each epoch a bag of two balls is drawn. There, the structure of the urn was fortunate such that results on the expected number of balls, of a certain certain color, could be established by means of conditioning on past draws and solving a linear system of recurrence. Nevertheless, the simple Pólya urn scenario and the given eigenvalue and martingale theory, as introduced next, is of importance for many applications (see also Mahmoud (2008, chap. 8 and chap. 9) for an overview of applications in informatics and biology) and can be considered as a motivational baseline in the study of more general Pólya urn models.

2.3.1 2-Color, simple, deterministic Urn Schemes – LLN

In the following, let us assume a deterministic urn scheme $\mathbf{A}$, represented by the 2×2 matrix

$$\begin{array}{cc} & \begin{array}{cc} W & B \end{array} \\ \begin{array}{c} W \\ B \end{array} & \begin{pmatrix} a & b \\ c & d \end{pmatrix} \end{array}, \tag{2.18}$$

associated with a simple Pólya urn consisting of white and blue balls. Hence, by assuming a Pólya urn framework, we implicitly assume that the urn scheme is tenable given the initial conditions. As with the Bagchi-Pal urn, let us further assume that one operates in the balanced case of constant row sums, that is:

$$a + b = c + d = K.$$

To guarantee tenability, we will have to assume that $K \geq 0$, hence a non-negative number of balls is added at each epoch. Further, let us exclude the Ehrenfest urn scenario, that is, let us assume that $K > 0$. One verifies quickly, by remembering that $\operatorname{tr} \mathbf{A} = \lambda_1 + \lambda_2$, where λ_1 and λ_2 are the eigenvalues associated to the scheme $\mathbf{A}$ and tr denotes the trace of the matrix $\mathbf{A}$, that under the assumption of constant row sums, we have:

$$\lambda_1 = K, \quad \text{and} \quad \lambda_2 = a - c.$$

As the case $a = c$ (hence $\lambda_2 = 0$) is associated with no randomness (one always adds a white balls and $K - a$ blue balls after each draw), we will exclude this case. Hence, similar to 2.8, we will assume a balanced urn condition and exclude cases of non existing randomness. Let us further assume that the components of the left (row) eigenvector corresponding to the eigenvalue λ_1 are all strictly positive. Thus, we may assume that the left (row) eigenvector corresponding to λ_1 is normalized such that the components add up to 1. In terms of notation, denote the left (row), normalized eigenvector corresponding to the eigenvalues λ_1 as $\mathbf{v} = \begin{pmatrix} v_1 & v_2 \end{pmatrix}$. The right (row) eigenvector, corresponding to the eigenvalue λ_2, shall be given by $\mathbf{u} = \begin{pmatrix} u_1 & u_2 \end{pmatrix}$. Notice that, in the 2-color, deterministic urn case, we have a principal eigenvector, associated to $\lambda_1 = K$, given as $\begin{pmatrix} 1 & 1 \end{pmatrix}$. In summary, our urn scheme $\mathbf{A}$ satisfies the following properties:

- The scheme is tenable given the initial condition of the urn.
- The scheme has constant row sum given by K.
- The scheme has a principal, real, strictly positive, eigenvalue which is equal to K.
- The components of the principal left (row) eigenvector are all strictly positive.

If one adds the condition $\lambda_2 < {}^{\lambda_1}/_2$, the urn scheme $\mathbf{A}$ is called extended – one talks about an extended, 2-color, deterministic Pólya urn scheme. In conclusion, our urn scheme $\mathbf{A}$, introduced above, is one special case of an extended urn scheme. We will have a look at the class of general extended urns in section 2.3.3. For now, we will focus on the 2-color, deterministic, tenable, balanced ($K > 0$), urn scheme $\mathbf{A}$, as introduced above, with two distinct eigenvalues that satisfy $\lambda_2 < {}^{\lambda_1}/_2$ and where cases with no randomness are excluded. The given notation as well as the way of reasoning is mostly adapted from Mahmoud (2008, chap. 6), where the results presented are due to the work of Athreya and Karlin (1968) and Smythe (1996). In the following, I will present a summary of the main results and at some instances, I will give a sketch of the arguments used in the establishment of the results – for more details, I refer to Mahmoud (2008, chap. 6).

***Establishing a Sum of Martingale Differences*:** Let W_n and B_n represent, respectively, the number of white and blue balls after n draws in an urn with scheme $\mathbf{A}$ as above. If we consider τ_n as the total number of balls in the urn after n draws, that is (under the assumption of constant row sum) $\tau_n = \tau_0 + Kn$, we get that

$$\frac{\tau_n}{n} \xrightarrow{\text{a.s.}} \lambda_1, \tag{2.19}$$

where a.s. indicates that the convergence is an almost sure convergence. Then, towards the establishment of a sum of martingale differences, let us define

$$\mathbf{Z}_n = \begin{pmatrix} W_n \\ B_n \end{pmatrix},$$

and

$$X_n = \mathbf{u}\mathbf{Z}_n = u_1 W_n + u_2 B_n.$$

Recall that $\mathbf{u}$ represents the right (row) eigenvector corresponding to the eigenvalue λ_2. If one specifies, for each $j \geq 1$, $\mathcal{F}_j$ to be the sigma algebra generated by W_j one can establish a martingale difference as follows: Calculate,

$$\begin{aligned}
\mathrm{E}(X_n - X_{n-1}|\mathcal{F}_{n-1}) &= \mathrm{E}(X_n|\mathcal{F}_{n-1}) - X_{n-1}\\
&= \mathrm{E}(\mathbf{u}\mathbf{Z}_n|\mathcal{F}_{n-1}) - X_{n-1}\\
&= \mathrm{E}(u_1 W_n + u_2 B_n|\mathcal{F}_{n-1}) - X_{n-1}\\
&= u_1\left(W_{n-1} + a\frac{W_{n-1}}{\tau_{n-1}} + c\frac{B_{n-1}}{\tau_{n-1}}\right)\\
&\quad + u_2\left(B_{n-1} + b\frac{W_{n-1}}{\tau_{n-1}} + d\frac{B_{n-1}}{\tau_{n-1}}\right) - X_{n-1}\\
&= (u_1 W_{n-1} + u_2 B_{n-1}) + \frac{1}{\tau_{n-1}}\mathbf{u}\mathbf{A}^T\mathbf{Z}_{n-1} - X_{n-1}\\
&= X_{n-1} + \frac{\lambda_2}{\tau_{n-1}}\mathbf{u}\mathbf{Z}_{n-1} - X_{n-1}\\
&= \frac{\lambda_2}{\tau_{n-1}}X_{n-1}.
\end{aligned}$$

Hence, if we define $\triangledown X_n \coloneqq X_n - X_{n-1}$ we have that

$$\mathrm{E}\left(\triangledown X_n - \frac{\lambda_2}{\tau_{n-1}}X_{n-1}|\mathcal{F}_{n-1}\right) = 0,$$

and therefore if we set

$$M_n \coloneqq \triangledown X_n - \frac{\lambda_2}{\tau_{n-1}}X_{n-1}, \tag{2.20}$$

the set $\{M_n, \mathcal{F}_n, n \geq 1\}$ is a set of martingale differences. Now, towards large n behavior, we would like to write X_n as the sum of the above martingale differences such that we can infer the asymptotic of X_n from the asymptotics of the sum of the martingale differences. Explicitly, we would like to come up with an expression

$$S_n = \sum_{j=1}^{n} \beta_{jn} M_j,$$

where for each $j \in \{1, 2, \ldots, n\}$ the coefficients of the above expression, β_{jn}, are such that

$$S_n = X_n + \epsilon_n,$$

where ϵ_n is small when n is large, that is $\epsilon_n \to 0$ as $n \to \infty$. Notice, by the fact that $\mathrm{E}(M_j) = 0$ for each $1 \leq j \leq n$, we have that $\mathrm{E}(S_n) = 0$. Hence, S_n is a zero mean martingale. The above strategy becomes established by matching the coefficients. A detailed argument is found in Mahmoud (2008, p. 104), here I will stick to the result:

$$\beta_{nn} = 1, \quad \text{and} \quad \beta_{jn} = \prod_{k=j}^{n-1}\left(1 + \frac{\lambda_2}{\tau_k}\right) \quad \text{for } 1 \leq j \leq n-1.$$

By applying Stirlings approximation for the ratio of gamma functions, that is, for real valued x, r, and s:

$$\frac{\Gamma(x+r)}{\Gamma(x+s)} = x^{r-s} + \mathrm{O}\left(x^{r-s-1}\right), \quad \text{as } x \to \infty,$$

one can deduce asymptotic properties for the given coefficients and the respective error term. We have that

$$\beta_{jn} = \left(\frac{n}{j}\right)^{\lambda_2/\lambda_1} + \mathrm{O}\left(n^{\lambda_2/\lambda_1 - 1}\right), \quad \text{as } n \to \infty,$$

and,

$$\epsilon_n = \mathrm{O}\left(n^{\lambda_2/\lambda_1}\right), \qquad \text{as } n \to \infty,$$

By having set up the appearance of the coefficients belonging to the sum of martingale differences one can now deduce a weak law of large numbers result under urn scheme conditions as given above.

***Weak Law of large Numbers*:** The key result in establishing a weak law of large numbers for the number of white (or blue) balls after n draws associated to the 2-color, deterministic, tenable, balanced, urn scheme $\mathbf{A}$, as introduced in (2.18), is the following lemma:

Lemma 2.1. If one assumes that $\lambda_2 < \lambda_1/2$, hence our urn scheme $\mathbf{A}$ is an example of an extended urn scheme, we have that

$$\frac{X_n}{n} \xrightarrow{P} 0.$$

A detailed proof of the above lemma is given in Mahmoud (2008, p. 106). The main idea behind the above statement is to conclude that $\mathrm{Var}(S_n)$ is linearly bounded by n. That is, one wants a linear bound of the form,

$$\mathrm{Var}(S_n) \leq Cn, \qquad \text{as } n \to \infty,$$

where C is some constant not dependent on n. As a consequence, one can apply Chebyshev's inequality to conclude that

$$\frac{S_n}{n} \xrightarrow{P} 0.$$

Finally, by using the asymptotic relationship between X_n and S_n, the lemma becomes established. Thus, the main work is in the establishment of a linear bound for $\mathrm{Var}(S_n)$. But, by knowing that the sum, S_n, is a zero mean martingale one can simplify the calculation of the variance and write:

$$\mathrm{Var}(S_n) = \sum_{j=1}^{n} \beta_{jn}^2 \, \mathrm{E}(M_j^2).$$

By writing out M_j, as given in (2.20), one can deduce, by using that the urn scheme is balanced ($\tau_j = \lambda_1 j + \tau_0$, $j \in \{1, 2, \ldots, n\}$, $n \geq 1$), that for any given $j \geq 1$, $|X_j|$ is linearly bounded by j. As an intermediate step, one can show that the terms $\{\mathrm{E}(M_j^2),\ 1 \leq j \leq n,\ n \geq 1\}$, in the above sum, are bounded by some constant M, not depending on n. Therefore, as a first result, one can estimate the above variance as

$$\mathrm{Var}(S_n) \leq M \sum_{j=1}^{n} \beta_{jn}^2.$$

At this point, it becomes clear how the assumption, $\lambda_2 < \lambda_1/2$, comes into play. Remember that the significant terms, as n becomes large, for the coefficients $\{\beta_{jn},\ 1 \leq j \leq n-1\}$, are given by $\left(n/j\right)^{\lambda_2/\lambda_1}$. Thus, by taking the square, one arrives at significant terms of the form $\left(n/j\right)^{2\lambda_2/\lambda_1}$. In conclusion, by using the assumption of the lemma, $\lambda_2 < \lambda_1/2$, one verifies that, for large n, $\mathrm{Var}(S_n)$ is linearly bounded by n. Having set up lemma 2.1 and by considering (2.19) one can now arrive at the weak law of large numbers for the number of balls of color white (resp. blue) after n draws:

Theorem 2.5 (*Weak Law of Large Numbers*)**.** Let W_n and B_n, respectively, be the number of white and blue balls in the 2-color, simple, deterministic, tenable, balanced urn scheme

$$\begin{array}{c|cc} & W & B \\ \hline W & a & b \\ B & c & d \end{array},$$

as in (2.18), with two distinct eigenvalues that satisfy $\lambda_2 < \lambda_1/2$ and left (row), normalized eigenvector $\mathbf{v} = \begin{pmatrix} v_1 & v_2 \end{pmatrix}$ associated to the principal eigenvalue λ_1. Then,

$$\frac{W_n}{n} \xrightarrow{P} \lambda_1 v_1,$$

$$\frac{B_n}{n} \xrightarrow{P} \lambda_1 v_2.$$

2.3.2 2-Color, simple, deterministic Urn Schemes – CLT

As in the preceding section, we are dealing with a 2-color, simple, deterministic urn scheme $\mathbf{A}$, as in (2.18), associated to a Pólya urn composed of white and blue colors – in summary, we are going to adapt the same notation, and conditions on $\mathbf{A}$, as given in section 2.3.1. To establish a central limit theorem result one again relies on martingale techniques and the main idea is to establish a sum of martingale differences such that one can apply the central limit theorem for martingales (see theorem 2.2). In the following, I will sketch the ideas for the 2-color, deterministic, balanced urn case as they are presented in Mahmoud (2008, chap. 6), where the arguments are due to the work of Smythe (1996). Let us recall that we have $X_n = u_1 W_n + u_2 B_n$, where $\mathbf{u} = \begin{pmatrix} u_1 & u_2 \end{pmatrix}$ is a non-principal right (row) eigenvector, corresponding to the eigenvalue $\lambda_2 = a - c$. Now, the procedure is similar to section 2.3.1:

***Establishing a Sum of Martingale Differences*:** First, one defines

$$W_n^* = W_n - v_1 \tau_n.$$

As we can see from (2.19) and theorem 2.5, we have that

$$\frac{W_n^*}{n} = \frac{W_n}{n} - v_1 \frac{\tau_n}{n} \xrightarrow{P} 0, \quad \text{as } n \to \infty,$$

hence W_n^* is asymptotically centered. As in section 2.3.1 one seeks a sum of martingale differences which is asymptotically equivalent to W_n^*. That is one wants to find coefficients β_{jn}^* ($j \in \{1, 2, \ldots, n\}$, $n \geq 1$) such that we have the representation

$$S_n^* = \sum_{j=1}^{n} \beta_{jn}^* M_j^* = W_n^* - \epsilon_n^*,$$

where the set $\{M_j^*, \mathcal{F}_j, j \geq 1\}$ is a set of martingale differences and ϵ_n^* is of small magnitude for large n. To establish the martingale differences for $j \geq 1$, we shall define $\mathbb{1}_j^W$ to be the indicator of the event of picking a white ball in the nth draw and consider

$$\begin{aligned}
\mathrm{E}(W_j | \mathcal{F}_{j-1}) &= W_{j-1} + aP(\mathbb{1}_j^W = 1 | F_{j-1}) + cP(\mathbb{1}_j^W = 0 | F_{j-1}) \\
&= W_{j-1} + a\frac{W_{j-1}}{\tau_{j-1}} + c\frac{B_{j-1}}{\tau_{j-1}} \\
&= W_{j-1} + a\frac{W_{j-1}}{\tau_{j-1}} + c\frac{\tau_{j-1} - W_{j-1}}{\tau_{j-1}}.
\end{aligned}$$

Then by noting that $W_j = W_j^* + v_1\tau_j$, we have that

$$\begin{aligned}\mathrm{E}(W_j^* + v_1\tau_j|\mathcal{F}_{j-1}) &= W_{j-1}^* + v_1\tau_{j-1} + a\frac{W_{j-1}^* + v_1\tau_{j-1}}{\tau_{j-1}}\\ &+ c\frac{\tau_{j-1} - W_{j-1}^* - v_1\tau_{j-1}}{\tau_{j-1}}.\end{aligned}$$

But now we can recall that $\lambda_1 = K$ and $\lambda_2 = a - c$ and we further have that

$$v_1 = \frac{c}{K+c-a}, \quad \text{and,} \quad v_2 = \frac{K-a}{K+c-a},$$

and therefore, for the increments, we arrive at

$$\begin{aligned}\mathrm{E}(W_j^* - W_{j-1}^*|\mathcal{F}_{j-1}) &= v_1(\tau_{j-1} - \tau_j) + (a-c)\frac{W_{j-1}^*}{\tau_{j-1}} + (a-c)v_1 + c\\ &= \lambda_2\frac{W_{j-1}^*}{\tau_{j-1}}.\end{aligned}$$

In conclusion, we have that, for $j \geq 1$, the increments $\bigtriangledown W_j^* = W_j^* - W_{j-1}^*$, satisfy

$$\mathrm{E}(\bigtriangledown W_j^* - W_{j-1}^*\frac{\lambda_2}{\tau_{j-1}}|\mathcal{F}_{j-1}) = 0,$$

and hence if we define

$$M_j^* := \bigtriangledown W_j^* - W_{j-1}^*\frac{\lambda_2}{\tau_{j-1}},$$

the collection $\{M_j^*, \mathcal{F}_j, j \geq 1\}$, is a collection of martingale differences. Now exactly as in section 2.3.1 one can go for a matching of coefficients to find the same representation for β_{jn}^* $(1 \leq j \leq n)$ and ϵ_n^* as in the establishment of the weak law of large numbers – we have that:

$$\beta_{nn}^* = 1, \quad \text{and} \quad \beta_{jn}^* = \prod_{k=j}^{n-1}\left(1 + \frac{\lambda_2}{\tau_k}\right) \quad \text{for } 1 \leq j \leq n-1,$$

with asymptotic properties,

$$\beta_{jn}^* = \left(\frac{n}{j}\right)^{\lambda_2/\lambda_1} + \mathrm{O}\left(n^{\lambda_2/\lambda_1 - 1}\right), \quad \text{as } n \to \infty,$$

and,

$$\epsilon_n = \mathrm{O}\left(n^{\lambda_2/\lambda_1}\right), \quad \text{as } n \to \infty,$$

which gives us, for $n \geq 1$, a representation for the sum of martingale differences S_n^*. As in the establishment of a LLN result, the coefficients associated to the martingale S_n^* have fortunate asymptotic properties under the assumption of an extended urn scheme.

***Asymptotic Normality*:** The scaling necessary for the CLT result is $\sqrt{n}$. Hence, at this stage, one wants to verify the condition necessary for the martingale CLT by considering the zero-mean, ordinary, martingale $\{S_n^*/\sqrt{n},\ \mathcal{F}_n,\ n \geq 1\}$. Notice that in the present case, one can derive a martingale array, as introduced in section 2.1, from the given ordinary martingale, by setting, for $n \geq 1$ and $1 \leq i \leq n$, $k_n = n$, $\mathcal{F}_{n,i} = \mathcal{F}_i$, and $S_{n,i} = n^{-1/2}(\sum_{j=1}^{i} \beta_{ji}^* M_j^*)$. Notice also, that in the present case, of an ordinary martingale, the sub-σ-fields $\{\mathcal{F}_{n,i},\ 1 \leq i \leq n,\ n \geq 1\}$, are clearly nested in n. As a conclusion, having in mind Theorem 2.2, we need to verify the following conditions:

(P1) $S_n^*/\sqrt{n}$ is a zero mean square integrable martingale.

(P2) $S_n^*/\sqrt{n}$ satisfies the conditional Lindenberg condition, that is:
$\forall \epsilon > 0,\ \frac{1}{n}\sum_{j=1}^{n} \mathrm{E}\left((\beta_{jn}^* M_j^*)^2 \mathbb{1}_{\{|n^{-1/2}(\beta_{jn}^* M_j^*)| > \epsilon\}} | \mathcal{F}_{j-1}\right) \xrightarrow{P} 0$, as $n \to \infty$.

(P3) $S_n^*/\sqrt{n}$ satisfies the conditional variance condition, that is:
$\frac{1}{n}\sum_{j=1}^{n} \mathrm{E}\left((\beta_{jn}^* M_j^*)^2 | \mathcal{F}_{j-1}\right) \xrightarrow{P} \sigma^2$, as $n \to \infty$, for some constant σ^2.

(P4) $\mathrm{E}(\max_{1 \leq j \leq n} \frac{1}{n}(\beta_{jn}^* M_j^*)^2 | \mathcal{F}_{j-1})$ is bounded in n.

The above properties are discussed and verified in Mahmoud (2008, Section 6.3.2). Here, I will try to capture the main ideas and central thoughts which govern the verification of the above properties and hence lead to a CLT result. Let us first consider property (P4). The martingale differences, $\{M_j^*,\ \mathcal{F}_j,\ j \geq 1\}$, are uniformly bounded. One can see this by the following estimation:

$$\begin{aligned} |M_j^*| &\leq |\bigtriangledown W_j^*| + \frac{|\lambda_2|}{\tau_{j-1}} |W_{j-1}^*| \\ &= |\bigtriangledown W_j - v_1 \bigtriangledown \tau_j| + \frac{|\lambda_2|}{\tau_{j-1}} |W_{j-1} - v_1 \tau_{j-1}| \\ &\leq |\bigtriangledown W_j| + v_1 \bigtriangledown \tau_j + \frac{|\lambda_2|}{\tau_{j-1}} (W_{j-1} + v_1 \tau_{j-1}) \\ &\leq \max(|a|,\ |c|) + \lambda_1 v_1 + |\lambda_2|(1 + v_1). \end{aligned}$$

Additionally, similar to the LLN case, as the significant terms of the squared coefficients $\{(\beta_{jn}^*)^2,\ 1 \leq j \leq n-1\}$, are given by

$$\left(\frac{n}{j}\right)^{2\lambda_2/\lambda_1}, \quad \text{as } n \to \infty,$$

we have, together with the uniform bound for the martingale differences, $\{M_j^*,\ \mathcal{F}_j,\ j \geq 1\}$ and the assumption $\lambda_2 < \lambda_1/2$, an uniform bound, in n, for the set

$$\left\{\frac{(\beta_{jn}^* M_j^*)^2}{n},\ 1 \leq j \leq n\right\},$$

which establishes property (P4). Actually, as the martingale $\{M_j^*,\ \mathcal{F}_j,\ j \geq 1\}$ is uniformly bounded and under the assumption $\lambda_2 < \lambda_1/2$, we can see that for a given $1 \leq j \leq n$, the sequence $\{(\beta_{jn}^* M_j^*)^2/n,\ n \geq 1\}$, converges to zero as $n \to \infty$, which gives property (P2). For a detailed discussion, consider also Mahmoud (2008, p. 112). Property (P1) becomes established by using a similar argument as in section 2.3.1, where one has established the variance of the sum of the martingale differences. Here, if one computes $\mathrm{E}\left((S_n^*/\sqrt{n})^2\right)$, one can notice again, by the fact that one has a sum of martingale differences, that the cross terms do vanish and hence one is left with

$$\mathrm{E}\left(\sum_{j=1}^{n}\frac{(\beta^*_{jn}M^*_j)^2}{n}\right).$$

The above expectation is finite, since, as in the LLN case,

$$\sum_{j=1}^{n}(\beta^*_{jn})^2\,\mathrm{E}\left((M^*_j)^2\right),$$

is linearly bounded in n, see also Mahmoud (2008, p. 106) for transparency. For property (P3), the crucial argument is that

$$\mathrm{E}\left((M^*_j)^2|\mathcal{F}_{j-1}\right)\xrightarrow{P} av_1(a-2\lambda_1v_1)+cv_2(c-2\lambda_1v_1)+\lambda_1^2v_1^2.$$

A full proof of the above statement is given in Mahmoud (2008, p. 110). Together with the above convergence in probability, and the fact that

$$\frac{1}{n}\sum_{j=1}^{n}(\beta^*_{jn})^2\to\int_0^1\left(\frac{1}{x}\right)^{\frac{2\lambda_2}{\lambda_1}}\mathrm{d}x=\frac{1}{1-\frac{2\lambda_2}{\lambda_1}},\qquad\text{as } n\to\infty$$

one can establish,

$$\frac{1}{n}\sum_{j=1}^{n}\mathrm{E}\left((\beta^*_{jn}M^*_j)^2|\mathcal{F}_{j-1}\right)\xrightarrow{P}\frac{av_1(a-2\lambda_1v_1)+cv_2(c-2\lambda_1v_1)+\lambda_1^2v_1^2}{1-\frac{2\lambda_2}{\lambda_1}},$$

which gives the conditional variance condition, hence property (P3). All together, as a result of (P1), (P2), (P3), and (P4) one arrives at the desired result:

$$\frac{S^*_n}{\sqrt{n}}=\frac{W^*_n-\epsilon^*_n}{\sqrt{n}}\xrightarrow{d}\mathcal{N}(0,\sigma^2),\tag{2.21}$$

where, after reorganization,

$$\sigma^2=\frac{av_1(a-2\lambda_1v_1)+cv_2(c-2\lambda_1v_1)+\lambda_1^2v_1^2}{1-\frac{2\lambda_2}{\lambda_1}}=\frac{bcK(a-c)^2}{(b-c)^2(K-2)(a-c)}.$$

But then, since $\lambda_2<{}^{\lambda_1}\!/_2$, we have that

$$-\frac{\epsilon^*_n}{\sqrt{n}}=\frac{\mathrm{O}\left(n^{\lambda_2/\lambda_1}\right)}{\sqrt{n}}\xrightarrow{\text{a.s}}0,$$

hence, by an application of Slutsky's theorem we can combine the convergence in distribution of (2.21) with the above almost sure convergence and get:

$$\frac{W_n-v_1\tau_n}{\sqrt{n}}\xrightarrow{d}\mathcal{N}(0,\sigma^2).$$

Finally, with $\tau_n=\tau_0+\lambda_1n$ and a second application of Slutsky's theorem, we arrive at the asymptotic normality for the number of white balls after n draws:

$$\frac{W_n-\lambda_1v_1n}{\sqrt{n}}\xrightarrow{d}\mathcal{N}(0,\sigma^2).$$

Notice that the variance term coincides with the asymptotic variance term in the Bagchi-Pal urn (consider theorem 2.4). In summary, one can summarize the above thoughts for the 2-color, deterministic urn case as follows:

Theorem 2.6. Let W_n and B_n, respectively, be the number of white and blue balls in the 2-color, simple, deterministic, tenable, balanced urn scheme

$$\begin{array}{c c} & \begin{array}{cc} W & B \end{array} \\ \begin{array}{c} W \\ B \end{array} & \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \end{array}$$

as given in (2.18), with two distinct eigenvalues that satisfy $\lambda_2 < {}^{\lambda_1}/_2$ and left (row), normalized eigenvector $\mathbf{v} = \begin{pmatrix} v_1 & v_2 \end{pmatrix}$. Then,

$$\frac{W_n - \lambda_1 v_1 n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, {\sigma_W}^2)$$

$$\frac{B_n - \lambda_1 v_2 n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, {\sigma_B}^2),$$

where ${\sigma_W}^2$ and ${\sigma_B}^2$ reflect the asymptotic variance for the number of white and blue balls respectively.

In a next step, we would like to extend the above results to the case of random, simple, urn schemes with, possibly, more than 2 colors. First, we will have a look at a general definition of an extended, simple Pólya urn scheme. This, as justified by Smythe (1996), gives the sufficient framework for weak law of large numbers and a central limit theorem results for the k-color, simple urn schemes.

2.3.3 Extended, k-Color, simple, Urn Schemes with random Entries

Up to now, by considering the notation of the previous two sections, we have considered asymptotics for the number of white or blue balls of a 2-color, deterministic, simple, tenable, balanced urn scheme

$$\begin{array}{c c} & \begin{array}{cc} W & B \end{array} \\ \begin{array}{c} W \\ B \end{array} & \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \end{array}$$

with two distinct eigenvalues that satisfy $\lambda_2 < {}^{\lambda_1}/_2$. Here, we will give a general definition of the class of k-color, extended, simple urn schemes and as we will see, the conditions which make an urn scheme extended are the conditions sufficient for the establishment of asymptotics (weak law of large numbers and asymptotic normality) for the number of balls of a certain color after n draws.

Definition 2.4 (*Extended, simple, Urn schemes*)**.** Consider a k-color, simple, Pólya urn scheme

$$\mathbf{A} = \begin{pmatrix} A_{1,1} & A_{1,2} & \cdots & A_{1,k} \\ A_{2,1} & A_{2,2} & \cdots & A_{2,k} \\ \vdots & \vdots & \ddots & \vdots \\ A_{k,1} & A_{k,2} & \cdots & A_{k,k} \end{pmatrix},$$

where the entries in the scheme $\mathbf{A}$ are integer-valued random variables. Further, let us arrange the k, possibly complex, eigenvalues according to their decreasing real parts:

$$\mathrm{Re}(\lambda_1) \geq \mathrm{Re}(\lambda_2) \geq \ldots \geq \mathrm{Re}(\lambda_k),$$

where λ_1 is called the principal eigenvalue with corresponding principal eigenvector. The urn scheme $\mathbf{A}$ is called extended, if it satisfies the following properties:

(E1) The urn scheme is tenable given the initial conditions

(E2) All the entries in the urn scheme have finite variances

(E3) The average generator, $\mathrm{E}(\mathbf{A})$, has constant and positive row sums

(E4) The principal eigenvalue of $\mathrm{E}(\mathbf{A})$ is equal to the constant row sum and has positive left eigenvector

(E5) All the eigenvalues of $\mathrm{E}(\mathbf{A})$ are assumed to be distinct

(E6) For any nonprincipal eigenvalue λ we have $\mathrm{Re}(\lambda) < \lambda_1/2$

Such general classes of urns were considered by Athreya and Karlin (1968) and Smythe (1996). The following two results are proven in Smythe (1996), where the arguments are mostly the same as in the 2-color, extended, case, as treated in sections 2.3.1 and 2.3.2. For a discussion, I will refer to Mahmoud (2008, chap. 6.2, 6.4, and 6.5).

Theorem 2.7. Suppose a k-color, extended, simple Pólya urn scheme with average generator that has a principal eigenvalue $\lambda_1 > 0$ and corresponding left (row), normalized eigenvector $\mathbf{v} = \begin{pmatrix} v_1 & v_2 & \cdots & v_k \end{pmatrix}$. Let $X_n^{(j)}$ be the number of balls of color j after n draws, for $j \in \{1, 2, \ldots, k\}$. Then for each j, we have

$$\frac{X_n^{(j)}}{n} \xrightarrow{P} \lambda_1 v_j.$$

Theorem 2.8. Suppose a k-color, extended, simple Pólya urn scheme with average generator that has a principal eigenvalue $\lambda_1 > 0$ and corresponding left (row), normalized eigenvector $\mathbf{v} = \begin{pmatrix} v_1 & v_2 & \cdots & v_k \end{pmatrix}$. Let $X_n^{(j)}$ be the number of balls of color j after n draws, for $j \in \{1, 2, \ldots, k\}$. Then, for each $j \in \{1, 2, \ldots, k\}$, we have that

$$\frac{X_n^{(j)} - \lambda_1 v_j n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, {\sigma_j}^2),$$

where ${\sigma_j}^2$ gives the asymptotic variance of the given color $j \in \{1, 2, \ldots, k\}$.

Example 2.10. Let us consider an urn, composed of white (W) and blue (B) balls, which evolves in discrete time under the random generator $\mathbf{G}$, given by the 2×2 scheme

$$\begin{array}{cc} & \begin{array}{cc} W & \quad B \end{array} \\ \begin{array}{c} W \\ B \end{array} & \begin{pmatrix} X & 1-X \\ 1-Y & Y \end{pmatrix}, \end{array}$$

where $X \sim \mathrm{Bernoulli}(1/4)$ and Y is an independent copy of X. Clearly, this urn scheme is tenable given any possible initial condition of the urn. Further we can calculate the average generator:

$$\mathrm{E}(\mathbf{G}) = \begin{pmatrix} 1/4 & 3/4 \\ 3/4 & 1/4 \end{pmatrix}.$$

The eigenvalues are $\lambda_1 = 1$ and $\lambda_2 = -(1/2)$. It follows that the given scheme is in an extended state, $\lambda_2 = -(1/2) < 1/2 = \lambda_1/2$. The above scheme represents an urn model where one adds exactly one ball at each epoch of time. By the system of equations

$$\begin{aligned}\frac{w_1}{4}+\frac{3w_2}{4}&=w_1\\ \frac{3w_1}{4}+\frac{w_2}{4}&=w_2,\end{aligned}$$

we can see that the left (row) principal eigenvector is given by $\mathbf{w}=\begin{pmatrix}1 & 1\end{pmatrix}$ and hence we get the left (row), normalized principal eigenvector as

$$\mathbf{v}=\frac{1}{1+1}\begin{pmatrix}1 & 1\end{pmatrix}.$$

Now, we can use theorem 2.8 to deduce a CLT result for the number of blue balls after n ($n\geq 1$) draws. Explicitly, if we write B_n for the number of blue balls after n draws, we can deduce that

$$\frac{B_n-\frac{1}{2}n}{\sqrt{n}}\xrightarrow{d}\mathcal{N}(0,{\sigma_B}^2),$$

where ${\sigma_B}^2$ gives the asymptotic variance for the number of blue balls. For a complete characterization, one would still need to fully work out the given asymptotic variance. This could be accomplished by a stochastic recurrence. For a similar example, I refer to Mahmoud (2008, Exercise 6.3).

We will now directly tie up with the above example and consider a first step towards applications in biology. We will see how we can use the urn scheme presented in example 2.10 in the context of clinical trials.

2.4 Application of Urns in Biology – Clinical Trials

The most classical form of a clinical trial is the following: Suppose one has two potentially different clinical treatments $\mathcal{A}$ and $\mathcal{B}$. Patients, which are sensible for both of the treatments, either receive $\mathcal{A}$ or $\mathcal{B}$. In the given context, sensible means that, for a given patient, for either $\mathcal{A}$ or $\mathcal{B}$, the treatment can have a positive or neutral effect on the given patient. A positive effect on the patient is considered as a treatment success, whereas a neutral effect is not. An important aspect in clinical trials is the strategy of how the treatment assignment should be accomplished – one possible strategy is to set up an urn. An introduction of the context is given in Zelen (1969). The goal of this section is to introduce and derive characteristics of urn schemes which are known as *Play-the-Winner Schemes*. These class of urn schemes belong to the class of simple Pólya urn schemes and as we will see, we can make use of the results provided in chapter 2. My notation is mostly adapted from Mahmoud (2008, Section 9.4).

With respect to a classical clinical trial of two treatments $\mathcal{A}$ and $\mathcal{B}$ one establishes an urn composed of, say, white (W) and blue (B) balls. The abundance of white balls correspond to the degree of success in treatment $\mathcal{A}$ and the abundance of blue balls correspond to the degree of success in treatment $\mathcal{B}$. For a quantification, one introduces, for any integer $n\geq 0$, the random variables W_n and B_n, giving, respectively, the number of white and blue balls after n draws. Similarly one defines, W_n^* and B_n^* to be, respectively, the number of times a white ball or a blue ball is drawn among the first n draws. One further assumes that the urn starts with an equal number of white and blue balls, that is $W_0=B_0$. Hence, we can assume that τ_0, the number of balls at the beginning, is given as $2W_0$. With regard to the "quality" of treatments $\mathcal{A}$ and $\mathcal{B}$ one assumes a certain, unknown, probability of success for both treatments and the result of the treatment (either successful or not) is assumed to be observed instantaneously. Notationwise, when treatment $\mathcal{A}$ is administered, it has a successful effect on the patient with probability $p_{\mathcal{A}}$

and it fails to succeed with probability $q_{\mathcal{A}} = 1 - p_{\mathcal{A}}$. Similarly, treatment $\mathcal{B}$, succeeds or fails, respectively, with probability $p_{\mathcal{B}}$ and $q_{\mathcal{B}} = 1 - p_{\mathcal{B}}$. Now, the Play-the-Winner scheme can be described as follows: For the administration of either $\mathcal{A}$ and $\mathcal{B}$ to a certain patient, the clinician picks at random, independently from previous draws, a ball from the urn. If the ball is white, it is replaced in the urn, treatment $\mathcal{A}$ is administered and the result is observed. If the treatment was a success one adds one additional white ball to the urn – one kind of rewards the positive effect of the given treatment by tilting the scale towards treatment $\mathcal{A}$. On the other hand, if treatment $\mathcal{A}$ is sampled, but it fails to succeed, one adds one blue ball to the urn to enhance the preference for treatment $\mathcal{B}$ in the upcoming draw. One follows the opposite strategy if a blue ball is drawn. The drawn ball is replaced and treatment $\mathcal{B}$ is given. If it succeeds, one adds an additional blue ball to the urn and if not, a ball of the opposite color is added to the urn.

The described urn workflow can be summarized with a simple, balanced, Pólya urn scheme. To do so, let $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$ be, respectively, two independent Bernoulli($p_{\mathcal{A}}$) and Bernoulli($p_{\mathcal{B}}$) random variables. The urn scheme, call it **A**, which describes the above treatment administration strategy is given by the 2×2 replacement matrix

$$\begin{array}{c c} & \begin{array}{cc} W & B \end{array} \\ \begin{array}{c} W \\ B \end{array} & \begin{pmatrix} X_{\mathcal{A}} & 1 - X_{\mathcal{A}} \\ 1 - X_{\mathcal{B}} & X_{\mathcal{B}} \end{pmatrix} \end{array}. \tag{2.22}$$

In the following, we are going to assume that the hidden success probability of both treatments satisfy

$$p_{\mathcal{A}} + p_{\mathcal{B}} < \frac{3}{2},$$

which might not be far from reality if one assumes that one of the treatments is a placebo treatment. As the urn progresses in discrete time, one wishes to maximize the number of successful administered treatments (either $\mathcal{A}$ or $\mathcal{B}$). One could say that the urn above is constructed in such a way that a patient's positive response to the treatment leads to a higher chance of the treatment to be applied again. On the contrary, a neutral act of the treatment on the patient, is punished by reducing the chance of the treatment to be applied again.

In conclusion, one can say that the urn learns via reward and punishment. But does this mean that such an urn, after running it for some time, ends up in a state where the chances of successful treatment administration are larger compared to a treatment administration system which does not learn from previous successes or failures? A *randomized clinical trial*, that is, a treatment administration system where there is no underlying urn and each time a patient is in demand of a treatment, a fair coin is flipped and either of the two treatments is administered, is one example of a administration system which does not learn from the past. In the following, we will have a look at results, for the total number of successful treatment administrations after many sampling sessions, for both, randomized clinical trials and Play-the-Winner schemes. Let us first discuss the randomized clinical trial.

***Randomized Clinical Trial*:** With the same notation as above, let $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$ be, respectively, the two independent Bernoulli($p_{\mathcal{A}}$) and Bernoulli($p_{\mathcal{B}}$) random variables which quantify the binary outcome of the given treatment (successful or not). If we define Y to be the indicator of success for a given patient, then we have that

$$Y = \begin{cases} X_{\mathcal{A}} & \text{with probability } 1/2 \\ X_{\mathcal{B}} & \text{with probability } 1/2. \end{cases}$$

If we condition on the result of the coin flip, either head (H) or tail (T), and associate H with treatment $\mathcal{A}$, we can deduce the expected value of Y. First, write:

$$Y = X_{\mathcal{A}}\mathbb{1}_H + X_{\mathcal{B}}\mathbb{1}_T.$$

Then, take expectation and use the fact that the coin flip is independent of the treatment outcome – we have,

$$\begin{aligned} \mathrm{E}(Y) &= \mathrm{E}(X_{\mathcal{A}}\mathbb{1}_H) + \mathrm{E}(X_{\mathcal{B}}\mathbb{1}_T) \\ &= \frac{1}{2}\big(\mathrm{E}(X_{\mathcal{A}}) + \mathrm{E}(X_{\mathcal{B}})\big) \\ &= \frac{p_{\mathcal{A}} + p_{\mathcal{B}}}{2}. \end{aligned}$$

Now, as the administration procedure does not depend on previously conducted treatment administrations, we can write the total number of successful treatment administrations, after n treatments, as the sum of n independent copies of Y. Explicitly, if we let S_n to be this total number of successes, we have

$$S_n = Y_1 + Y_2 + \cdots + Y_n,$$

where for each $n \geq 1$, for each $i \in \{1, 2, \ldots, n\}$, Y_i are independent copies of Y. Now, according to the weak law of large numbers, we have that

$$\frac{S_n}{n} \xrightarrow{P} \mathrm{E}(Y) = \frac{p_{\mathcal{A}} + p_{\mathcal{B}}}{2}.$$

In summary, after many treatment administrations, the randomized clinical trial attains, in probability, a success rate which is the average of the two individual treatment success rates. Notice that the above convergence result could be strengthen to an almost sure convergence.

***Play-the-Winner scheme*:** Let us now deduce a similar result for clinical trials which perform treatment administration via the Play-the-Winner urn scheme given in (2.22). The average generator is given by

$$\mathrm{E}(\mathbf{A}) = \begin{pmatrix} p_{\mathcal{A}} & 1 - p_{\mathcal{A}} \\ 1 - p_{\mathcal{B}} & p_{\mathcal{B}} \end{pmatrix}.$$

Now, the above expected urn scheme is balanced with constant row sum given by 1. It has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = p_{\mathcal{A}} + p_{\mathcal{B}} - 1$. The condition for an extended, expected, urn scheme (see definition 2.4), becomes translated as

$$p_{\mathcal{A}} + p_{\mathcal{B}} < \frac{3}{2},$$

which is given by assumption. The principal, left (row), eigenvector $\mathbf{w} = \begin{pmatrix} w_1 & w_2 \end{pmatrix}$ satisfies $\mathbf{w}\,\mathrm{E}(\mathbf{A}) = \mathbf{w}$ and is given by $\mathbf{w} = \begin{pmatrix} 1 - p_{\mathcal{B}} & 1 - p_{\mathcal{A}} \end{pmatrix}$. Hence, the principal, standardized, left, eigenvector is given by

$$\mathbf{v} = \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \frac{1}{2 - p_{\mathcal{A}} - p_{\mathcal{B}}} \begin{pmatrix} 1 - p_{\mathcal{B}} \\ 1 - p_{\mathcal{A}} \end{pmatrix}.$$

Since the row sum of $\mathrm{E}(\mathbf{A})$ is 1, each administered treatment, adds one ball to the urn – after $n-1$ administrations, there are $\tau_{n-1} = \tau_0 + n - 1$ balls in the urn. Similar to the randomized clinical trial, we can now define a indicator of success, for a given patient, at the nth administration, as

$$Y_n = \begin{cases} X_{\mathcal{A}}^{(n)} & \text{with probability } \frac{W_{n-1}}{\tau_{n-1}} \\ X_{\mathcal{B}}^{(n)} & \text{with probability } \frac{B_{n-1}}{\tau_{n-1}}, \end{cases}$$

where $X_{\mathcal{A}}^{(n)}$ and $X_{\mathcal{B}}^{(n)}$ are, respectively, independent copies of $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$. Notice how, in comparison to the randomized situation, Y_n, the indicator of success, depends on the history of the urn. We are at the position to apply theorem 2.7 and deduce that

$$\frac{W_n}{n} \xrightarrow{P} v_1 = \frac{1-p_{\mathcal{B}}}{2-p_{\mathcal{A}}-p_{\mathcal{B}}}, \quad \text{and} \quad \frac{B_n}{n} \xrightarrow{P} v_2 = \frac{1-p_{\mathcal{A}}}{2-p_{\mathcal{A}}-p_{\mathcal{B}}}.$$

Towards our goal, let S_n be the number of successes after n treatment administrations in the Play-the-Winner scheme, then, again,

$$S_n = Y_1 + Y_2 + \cdots + Y_n,$$

where now, for $i \in \{1, 2, \ldots, n\}$, Y_i, the Y_i are dependent. But, here, in the given situation, it is clever to represent the number of successes, S_n, by the number of times a white or blue ball is drawn after n draws. Explicitly, to break the dependence structure introduced above, we can represent the sum above with the help of W_n^* and B_n^*:

$$S_n = \big(X_{\mathcal{A}}^{(1)} + X_{\mathcal{A}}^{(2)} + \cdots + X_{\mathcal{A}}^{(W_n^*)}\big) + \big(X_{\mathcal{B}}^{(1)} + X_{\mathcal{B}}^{(2)} + \cdots + X_{\mathcal{B}}^{(B_n^*)}\big).$$

Using $W_n = W_0 + X_{\mathcal{A}}W_n^* + (1-X_{\mathcal{B}})B_n^*$ and $B_n^* = n - W_n^*$, we first have,

$$\frac{W_n^*}{n} \xrightarrow{P} v_1 = \frac{1-p_{\mathcal{B}}}{2-p_{\mathcal{A}}-p_{\mathcal{B}}}.$$

But then, we also have that $W_n^* \xrightarrow{a.s.} \infty$ and hence

$$\frac{X_{\mathcal{A}}^{(1)} + X_{\mathcal{A}}^{(2)} + \cdots + X_{\mathcal{A}}^{(W_n^*)}}{W_n^*} \xrightarrow{a.s.} p_{\mathcal{A}}.$$

In conclusion, we have that

$$\frac{X_{\mathcal{A}}^{(1)} + X_{\mathcal{A}}^{(2)} + \cdots + X_{\mathcal{A}}^{(W_n^*)}}{n} = \left(\frac{X_{\mathcal{A}}^{(1)} + X_{\mathcal{A}}^{(2)} + \cdots + X_{\mathcal{A}}^{(W_n^*)}}{W_n^*}\right)\left(\frac{W_n^*}{n}\right) \xrightarrow{P} p_{\mathcal{A}}\left(\frac{1-p_{\mathcal{B}}}{2-p_{\mathcal{A}}-p_{\mathcal{B}}}\right).$$

Similarly, for treatment $\mathcal{B}$:

$$\frac{X_{\mathcal{B}}^{(1)} + X_{\mathcal{B}}^{(2)} + \cdots + X_{\mathcal{A}}^{(B_n^*)}}{n} \xrightarrow{P} p_{\mathcal{B}}\left(\frac{1-p_{\mathcal{A}}}{2-p_{\mathcal{A}}-p_{\mathcal{B}}}\right).$$

As a conclusion, we arrive at the following result relative success,

$$\begin{aligned} S_n &= \frac{\big(X_{\mathcal{A}}^{(1)} + X_{\mathcal{A}}^{(2)} + \cdots + X_{\mathcal{A}}^{(W_n^*)}\big)}{n} + \frac{\big(X_{\mathcal{B}}^{(1)} + X_{\mathcal{B}}^{(2)} + \cdots + X_{\mathcal{B}}^{(B_n^*)}\big)}{n} \\ &\xrightarrow{P} \frac{p_{\mathcal{A}}(1-p_{\mathcal{B}}) + p_{\mathcal{B}}(1-p_{\mathcal{A}})}{2-p_{\mathcal{A}}-p_{\mathcal{B}}}. \end{aligned}$$

The comparison of both scenarios, the randomized clinical trial and the Play-the-Winner scheme, boils down to compare the two limiting functions

$$f(x,\, y) = \frac{1}{2}(x+y), \quad \text{and} \quad g(x,\, y) = \frac{x(1-y)+y(1-x)}{2-x-y},$$

for both, $x,\, y \in (0,\, 1)$. By using the fact that, for a given range of interest in the interval $(0,\, 1)$,

$$g(x,\, y) - f(x,\, y) = \frac{(x-y)^2}{2(2-x-y)} \geq 0,$$

one can see that the limiting success rate of the Play-the-Winner is better than, or as good as ($p_{\mathcal{A}} = p_{\mathcal{B}}$ scenario), the randomized clinical trial administration strategy.

In the final, upcoming section of this chapter about motivational results, we will leave the class of Pólya urn models and introduce an urn model which aims to describe the emergence of species, through mutations.

2.5 Application of Urns in Biology – The Evolution of Species

As introduced in Section 2.2.2, the class of Pólya urn models is characterized by a single urn, where the ball drawing and addition scheme (Pólya urn scheme) is defined via a finite, fixed set of admissible colors. However, when approaching a mathematical model, which aims to be a sensible representative of a biological process such as the evolution of species, one must allow to violate the assumption of a finite, fixed set of admissible alleles (different representations of a gene). The violation of this assumption finds its justification in our knowledge about Darwinian evolution: Over time, alleles undergo changes in composition, because at any given time, as a result of mutation, a new allele might appear. In conclusion, if one wants to aim for a mathematical model as described above, and desires to represent different alleles by different colors in an urn, one leaves, in the strict sense, the Pólya urn scenario and enters the situation where one deals with an infinite set of admissible colors. Such a mathematical model would do its service in the field of population genetics, where one reduces a population to its genetic diversity (given a certain gene, diversity is high, if there are many different representatives of that given gene – many different alleles of the given gene). In that case, one would like to allow for a possible increase or decrease in genetic diversity. Here in this section, I will introduce the so called Hoppe's urn model, named after Fred M. Hoppe, which, in contrast to the Pólya urns, allows for an infinite set of admissible colors. As we will see, the Hoppe's urn can be brought into context of population genetics. The Hoppe's urn model, as well as important properties and characteristics, are introduced in Mahmoud (2008, Section 9.1.4 and 9.1.5). For a detailed discussion, I will refer to the given source – here, I will give a summary of the main results and important thoughts.

Before introducing the Hoppe's urn, I would like to consider a historical perspective and introduce a class of urn models known as the *simple Wright-Fisher models*. In the following, I will demonstrate, in form of an example, the most simplest case. For a more general discussion I will refer to Mahmoud (2008, Section 9.1.1).

Example 2.11 (Simple Wright-Fisher Allelic Urn Model – The smallest Urn)**.** A simple Wright-Fisher Allelic Urn model is an urn model, which serves as a genetic model for a population of two alleles, where the different alleles are represented by two different colors, say white and blue. The term "simple" refers to the fact that one only considers two possible alleles and the potential arise of new alleles, by means of mutations, is ignored. At any instance of discrete time one considers an urn, filled with a total of m balls. The composition of balls, of the urn, at the given time, is such that i of the balls are white (representative for one allele) and $m - i$ are blue (representative for the other allele). The model is such that if one has an urn at time, say n (call it U_n, where n is a non-negative integer), composed of m balls, one samples, from U_n, balls, independently, with replacement, m times and the drawn samples are stored in an subsequent urn (call it U_{n+1}). Hence, in contrast to Pólya urn models, one considers an infinite set of urns over time and one is interested in the composition of U_n for large n. In the present example I will assume the smallest possible urn – that is for any integer $n \geq 0$, U_n has $m = 2$ balls which can be of color white (W) or blue (B).

In the given framework, for any integer $n \geq 0$, we can represent the transition from U_n to U_{n+1} by a Markov chain with three possible states. If we decide to number a state by the number of white balls in a given urn, one will, starting out in state 0, stay in state 0 with probability 1 and similarly, one will, starting out in state 2, stay in state 2 with probability 1. Randomness is associated with state 1, where a one step transition, starting out in state 1, is represented by a random variable $X \sim \text{Bin}(2, {}^1\!/_2)$, reflecting the independent sampling with replacement and the probability of drawing a white ball starting out in state 1. We can summarize the situation, by assuming that U_0 contains one white and one blue ball (state 1), with the according transition graph (see figure 2.1).

Now, the one step transition matrix, with entries p_{ij} (that is the probability to go from state

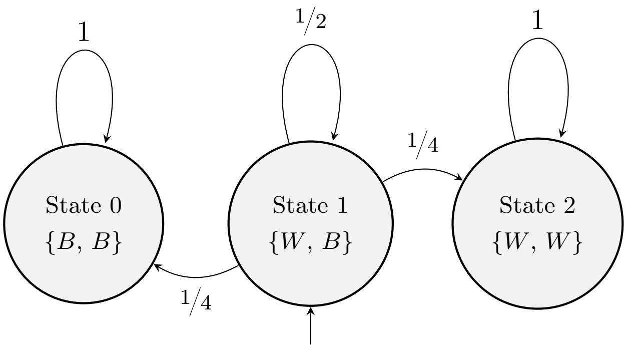


Figure 2.1: Transition graph for a simple Wright-Fisher model starting out with one white (W) and one blue (B) ball

$i \in \{0, 1, 2\}$ to state $j \in \{0, 1, 2\}$), is translated as follows:

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ 0 & 0 & 1 \end{pmatrix}.$$

From that, we can calculate the n-step ($n \geq 2$) transition matrix M^n with entries $p_{ij}^{(n)}$, where the entries reflect the probability to go from state i to state j in exactly n steps. By using the fact that the number of white balls in the nth urn is a Markov chain, we have that

$$p_{ij}^{(n)} = \sum_{k=0}^{2} p_{ik} p_{kj}^{(n-1)}, \tag{2.23}$$

and hence, by induction, we arrive at,

$$\mathbf{M}^n = \begin{pmatrix} 1 & 0 & 0 \\ \frac{2^n-1}{2^{n+1}} & \frac{2}{2^{n+1}} & \frac{2^n-1}{2^{n+1}} \\ 0 & 0 & 1 \end{pmatrix}.$$

Now if we define

$$\pi^{(n)} = \begin{pmatrix} \pi_0^n & \pi_1^n & \pi_2^n \end{pmatrix},$$

to be the probability state vector, that is, the entries π_k^n ($k \in \{0, 1, 2\}$) reflect the probabilities to arrive in state k after n steps, we have that $\pi^{(0)} = \begin{pmatrix} 0 & 1 & 0 \end{pmatrix}$ and hence we get, by rewriting (2.23), that

$$\begin{aligned} \pi_n &= \pi^{(0)} \mathbf{M}^n \\ &= \begin{pmatrix} \frac{2^n-1}{2^{n+1}} & \frac{2}{2^{n+1}} & \frac{2^n-1}{2^{n+1}} \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{2} & 0 & \frac{1}{2} \end{pmatrix} + \begin{pmatrix} -\frac{1}{2^{n+1}} & \frac{1}{2^n} & -\frac{1}{2^{n+1}} \end{pmatrix}. \end{aligned}$$

As a conclusion of the above thoughts, we can deduce that, for large n, the urn U_n is with high probability either in a state with all white or all blue balls.

Remark 2.6. The above example gives one instance of an urn model which describes the composition of two alleles after n generations, where the nth generation is represented by the urn U_n. I have considered the case where the population size is given by $m = 2$. For a population of size $m > 2$, the idea remains the same, and again, for large n, the urn (U_n) becomes either all white or all blue with high probability – for any possible start U_0, which is neither all blue nor all white, for any of the $m + 1$ states, there is a positive probability to end up in an absorbing state, that is a state where the urn becomes populated with only white or blue balls.

As said above, the simple Wright-Fisher Allelic Urn model considers a fixed number of 2 alleles. To capture a larger part of reality, we would like to allow for the emergence of new alleles (emergence of new colors). Such an adapted model, which allows for the rise of new alleles, could then be interpreted as a model which respects the process of gene mutations – hence, a model which incorporates a mechanism where new versions of a given gene might arise. One model, which respects such an adaptation is the Hoppe's urn.

The Hoppe's urn model is one solution to the problem of extending the number of admissible colors to a countable, infinite, set of possible colors. The urn was first discussed in Hoppe (1984), where the mathematical foundation was established earlier, by Ewens (1972). The set up is the following: A single color is considered as "special", and, in view of an infinite supply of colors, all the other colors are labeled with positive integers 1, 2, 3, 4, 5, ... – the special color does not account for any of the integer labeled colors. The urn progresses in discrete time steps and at the beginning, there are $\theta \geq 1$ balls of special color. At any discrete time step, a ball is sampled, at random, independently form the previous draws, from the urn (again, all balls, of any color, including the special color, are equally likely). If one samples a ball of non-special color, the ball is returned to the urn and a ball of the same color is added to the urn. If one draws a ball of special color, the ball is returned to the urn together with a ball of a new color. That is if, at time point n, there are colors labeled with integers 1, 2, $\ldots k$, and a ball of special color is sampled, one adds a ball of color $k + 1$ to the urn.

With regard to population genetics: At a given point in time, the urn reflects a population of different alleles (the non-special colors). The composition of alleles, that is the diversity of the given alleles can vary over time. The special colored balls serve as generators for new alleles, where the event of sampling a special color form the urn can be interpreted as a mutation and hence the emergence of a new allele, a new version of a particular gene arises. As a simple illustration, figure 2.2 gives the early sampling path from a Hoppe's urn starting out with one black ball (the special color is defined to be black).

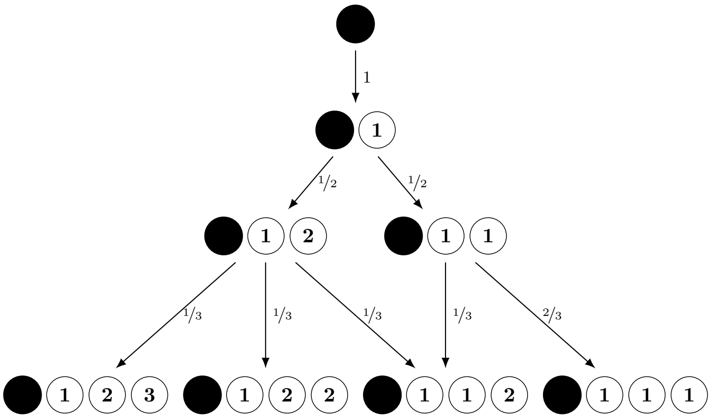


Figure 2.2: Early appearance of Hoppe's urn starting out with one special colored ball.

Having discussed the workflow of the Hoppe's urn, two natural questions arise: First, one could be concerned about the number of non-special colors, that is the number of different alleles, after a certain number of draws. Second, it could be interesting to ask about the composition of the urn after a particular number of samples from the urn. With regard to the latter, the composition of the urn at a particular point in time is given, by specifying the abundance of the different alleles at the given point. In the remaining part of this section on the Hoppe's urn I will discuss the answers to the above questions. Let us start with the number of non-special colors after a certain progression of the urn.

***The number of non-special colors in a Hoppe's urn*:** Let us develop the mean and variance for the number of non-special colors in a Hoppe's urn after n draws. To do so, let me first state some minor prerequisites:

Definition 2.5 (*The nth generalized harmonic Number of Order k:*)**.** Fix two integers, $n \geq 1$ and $k \geq 1$. The nth generalized harmonic number of order k, at some real $x \geq 1$, is given by the series

$$\mathcal{H}_n^{(k)}(x) = \frac{1}{x^k} + \frac{1}{(x+1)^k} + \cdots + \frac{1}{(x+n-1)^k}.$$

Remark 2.7 (*Asymptotic Properties of the generalized harmonic Number*)**.** With regard to definition 2.5, the standard harmonic number is given by $\mathcal{H}_n^{(1)}(1)$, where standard refers to the fact that x is chosen to be 1. The generalized harmonic number of order 1 have similar asymptotic to the standard harmonic numbers – Given any fixed, real, $x \geq 1$, we have that

$$\mathcal{H}_n^{(1)}(x) \sim \log(n), \qquad \text{as } n \to \infty. \tag{2.24}$$

For higher order, that is for fixed $k \geq 1$, the generalized harmonic number behaves, asymptotically, like a constant – for fixed, real $x \geq 1$, we arrive at

$$\mathcal{H}_n^{(k)}(x) \sim \mathrm{O}(1), \qquad \text{as } n \to \infty. \tag{2.25}$$

A proof of both properties is found in Mahmoud (2008, p. 174).

Now, with regard to the Hoppe's urn, let us define K_n to be the number of non-special colors (or number of different alleles) in the urn after n draws. We can now represent this random variable, K_n, as the sum of n independent random variables. Explicitly, in the language of the Hoppe's urn, with each conducted sample, we place one extra ball in the urn and hence, after i draws, there is a total of $\tau_i = \theta + i$ balls in the urn. The probability of picking a special colored ball in the ith draw is given by ${}^{\theta}/_{\tau_{i-1}}$. Now if we define the event

$$\Lambda_i := \{\text{The Ball in the } i\text{th Draw is Special}\},$$

and set

$$Y_i = \mathbb{1}_{\Lambda_i},$$

we have that $Y_i \sim \text{Bernoulli}({}^{\theta}/_{\tau_{i-1}})$. By the set up of the Hoppe's urn, the Bernoulli random variables are independent and have expectation and variance,

$$\mathrm{E}(Y_i) = \frac{\theta}{\tau_{i-1}}, \qquad \mathrm{Var}(Y_i) = \frac{\theta}{\tau_{i-1}}\left(1 - \frac{\theta}{\tau_{i-1}}\right).$$

In conclusion, we have the representation

$$K_n = Y_1 + Y_2 + \cdots + Y_n. \tag{2.26}$$

After having expressed the number of different colors after n draws as the sum of independent Bernoulli random variables, the expectation and variance of K_n are deduced naturally:

Proposition 2.3 (Ewens, 1972)**.** Let K_n be the number of non-special colors in a Hoppe's urn after n draws, starting out with $\theta \geq 1$ special colored balls. Then

$$\mathrm{E}(K_n) = \theta \mathcal{H}_n^{(1)}(\theta) \overset{n\to\infty}{\sim} \theta \log(n),$$

and

$$\mathrm{Var}(K_n) = \theta \mathcal{H}_n^{(1)}(\theta) - \theta^2 \mathcal{H}_n^{(2)}(\theta) \overset{n\to\infty}{\sim} \theta \log(n),$$

where $\overset{n\to\infty}{\sim}$ indicates the notion of asymptotic equivalence as $n \to \infty$.

Proof. If we take expectation of (2.26), we get

$$\mathrm{E}(K_n) = \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \cdots + \frac{\theta}{\theta+n-1} = \theta \mathcal{H}_n^{(1)}(\theta).$$

By independence, and the notation of (2.26), we can further deduce that

$$\begin{aligned}\mathrm{Var}(K_n) &= \mathrm{Var}(Y_1) + \mathrm{Var}(Y_1) + \cdots + \mathrm{Var}(Y_n) \\ &= \sum_{i=0}^{n-1} \frac{\theta}{\theta+i}\left(1 - \frac{\theta}{\theta+i}\right) \\ &= \theta \mathcal{H}_n^{(1)}(\theta) - \theta^2 \mathcal{H}_n^{(2)}(\theta).\end{aligned}$$

Now, the rest of the proposition follows from remark 2.7. □

We can now use the above asymptotics of expectation and variance to show that, as n becomes larger and larger, the probability that K_n deviates from its asymptotic expectation $\theta \log(n)$, becomes smaller and smaller – formally we have the following corollary:

Corollary 2.1.

$$\frac{K_n}{\log(n)} \xrightarrow{P} \theta.$$

Also here, a proof is established quickly. I will omit the details and refer to Mahmoud (2008, p. 176). The key idea is to first use Chebyshev's inequality, with $\epsilon > 0$, for K_n and replace ϵ with $\epsilon\,\mathrm{E}(K_n)$ to arrive at a convergence in probability of $K_n/\mathrm{E}(K_n)$ to 1. Combining this convergence in probability with a regular convergence for $\mathrm{E}(K_n)/\log(n)$ to θ, given in proposition 2.3, one arrives via Slutsky's theorem at the desired result.

In a next step, we will establish a closed form representation of a probability measure associated to K_n. First let us introduce the signless Stirling's numbers of the first kind.

Definition 2.6 (*Signless Stirling's Numbers of the first Kind*)**.** For a given complex number x, the rising factorial of order n writes as

$$\langle x \rangle_n = x(x+1)(x+2) \cdot \ldots \cdot (x+n-1). \tag{2.27}$$

For a given $1 \leq k \leq n$, the kth signless Stirling's number, of order n, of the first kind, associated to the polynomial $\langle x \rangle_n$, is given by the kth coefficient, belonging to x^k. It is denoted by $\left[{n \atop k}\right]$, for $1 \leq k \leq n$.

Proposition 2.4 (Ewens, 1972)**.** Let K_n be the number of non-special colors in a Hoppe's urn after n draws, starting out with $\theta \geq 1$ special colored balls. Then for $k = 1,\, 2,\, \ldots,\, n$,

$$P(K_n = k) = \frac{\theta^k}{\langle\theta\rangle_k} \begin{bmatrix} n \\ k \end{bmatrix},$$

where, as in definition 2.6, $\left[\begin{smallmatrix} n \\ k \end{smallmatrix}\right]$ refers to the k'th Stirling number of the first kind in the polynomial of degree n generated by the rising factorial of order n.

As the proof of the above proposition is presented without to much of an effort, by writing the probability generating function for (2.26), I will provide a full argument here:

Proof of Proposition 2.4. We start with the representation of K_n as the sum of n independent Bernoulli random variables and write down the probability generating function, $G_{K_n}(z)$, for the non-negative integer valued random variable K_n at complex numbers z with absolute value not greater than one. Explicitly, we get for complex numbers $|z| \leq 1$,

$$\begin{aligned} G_{K_n}(z) &= G_{Y_1+Y_2+\cdots+Y_n}(z) \\ &= G_{Y_1}(z)G_{Y_2}(z) \cdot \ldots \cdot G_{Y_n}(z) \\ &= \frac{\theta z}{\theta}\left(\left(1 - \frac{\theta}{\theta+1}\right) + \frac{\theta z}{\theta+1}\right) \cdot \ldots \cdot \left(\left(1 - \frac{\theta}{\theta+n-1}\right) + \frac{\theta z}{\theta+n-1}\right) \\ &= \frac{\theta z}{\theta}\left(\frac{\theta z+1}{\theta+1}\right) \cdot \ldots \cdot \left(\frac{\theta z+n-1}{\theta+n-1}\right) \\ &= \frac{1}{\langle\theta\rangle_n} \sum_{k=1}^{n} \begin{bmatrix} n \\ k \end{bmatrix} (\theta z)^k, \end{aligned} \tag{2.28}$$

where we have used the independence of the Bernoulli random variables and the fact that for each $1 \leq k \leq n$,

$$G_{Y_k}(z) = \left(1 - \frac{\theta}{\theta+k}\right) + \frac{\theta}{\theta+k} z.$$

If we extract the coefficient z^k in (2.28) we arrive at the desired result. □

One can give an approximation of the probabilities associated to K_n by means of establishing asymptotic normality for K_n via a central limit theorem result. The following proposition is proven as a corollary in Mahmoud (2008, p. 178, Corollary 9.2).

Proposition 2.5 (*Central Limit Result for the Number of non-special Colors*)**.** Let K_n be the number of non-special colors in a Hoppe's urn after n draws, starting out with $\theta \geq 1$ special colored balls. Then

$$\frac{K_n - \theta \log(n)}{\sqrt{\log(n)}} \xrightarrow{P} \mathcal{N}(0, \theta).$$

The above result is a direct consequence of a central limit theorem known as *Lyapunov's central limit theorem.* I will omit further details here and refer to Mahmoud (2008, p. 178, Theorem 9.2 and Corrollary 9.2) for a more comprehensive treatment.

***The composition of the Hoppe's urn*:** By considering the workflow of the Hoppe's urn it is clear that after n sampling periods there are n individuals forming a population which is composed of up to a maximum of n different alleles. If we stick to the population genetic interpretation of balls and colors, an individual is reduced to the gene of interest, where the gene of interest can have several allelic representations. What can we say about the diversity of individuals after n draws? In other terms, what can we say about the allele profile in the population of n individuals after n draws from the Hoppe's urn? One way of approaching this question is to reformulate it in a promising way. With regard to the allele profile, we could ask: How may alleles are represented by how many individuals? In the language of probability: Suppose, after n steps, there are A_1 alleles represented by 1 individual each, A_2 alleles represented by 2 individuals each, and so on up to A_n alleles which are represented by n individuals. By considering, for any $n \geq 1$, the sequence $A_1,\, A_2,\, \ldots,\, A_n$ as a sequence of random variables with possible realization $a_1,\, a_2,\, \ldots,\, a_n$, it is tempting to ask for a representation of the probability

$$P(A_1 = a_1,\, A_2 = a_2,\, \ldots,\, A_n = a_n).$$

With respect to the Hoppe's urn, not any of the n-tupel $a_1,\, a_2,\, \ldots,\, a_n$ is feasible and since there are n individuals and up to n different alleles after n draws, a feasible tuple must partition n:

$$n = a_1 1 + a_2 2 + \cdots + a_2 n.$$

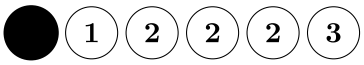
Hence, after n draws, a feasible realization of the sequence $A_1,\, A_2,\, \ldots,\, A_n$, partitions n into a_1 times 1 individuals which are composed of a_1 alleles, a_2 times 2 individuals which are composed of a_2 alleles, and so on. It is clear that the above probability is zero for any non-feasible tuple. For an illustrative example, one could consider the early sample path of the Hoppe's urn, starting out with one special color (black), depicted in figure 2.2. For instance, consider the third generation ($n = 3$). The right most configuration is partitioned as follows: There is one allele, the allele which is numbered with 1, which is represented by three individuals. There is no allele which is represented by two or one individuals – we have the configuration $a_1 = 0,\, a_2 = 0,\, a_3 = 1$. On the other hand, in the left most configuration, there are three alleles, the alleles numbered with 1, 2, and 3, which are represented by one individual each – the configuration is $a_1 = 3,\, a_2 = 0$, and $a_3 = 0$. With respect to the probability of interest, one would get

$$P(A_1 = 0,\, A_2 = 0,\, A_3 = 1) = \frac{2}{6}, \quad \text{and} \quad P(A_1 = 3,\, A_2 = 0,\, A_3 = 0) = \frac{1}{6},$$

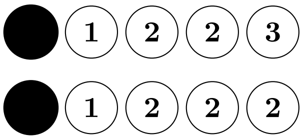
respectively.

Now, as the Hoppe's urn evolves, the complexity associated to possible configurations increases. Still, if we focus on a given generation, say generation $n + 1$, and a given feasible realized configuration $a_1,\, a_2,\, \ldots,\, a_{n+1}$, one can describe the given configuration at $n+1$, recursively, by the possible ways of reaching the given configuration at stage $n+1$ from the previous generation. As an example of the above thought, one can consider figure 2.3, where again, the special color is given by black. Figure 2.3a shows a possible configuration of the Hoppe's urn in the fifth generation. The configuration is given by $A_1 = 2,\, A_2 = 0,\, A_3 = 1,\, A_4 = 0,\, A_5 = 0$ – two alleles, alleles 1 and 3, are represented by one individual each and one allele is represented by 3 individuals. figure 2.3b gives the possible configurations in the fourth generation which might result in the configuration given in figure 2.3a. The bottom configuration of figure 2.3b evolves to the configuration in figure 2.3a if a black ball is drawn. The top configuration of figure 2.3b evolves to the configuration in figure 2.3a if a ball of color 2 is drawn in the nth generation.

With the notation of partitions as row vectors, we can now further generalize the above example and consider the two possible boundary cases for partitions in the $n + 1$th generation:



(a) A possible Configuration of Hoppe's urn after 5 draws



(b) Feasible Configurations at step 4 in a Hoppe's urn that could end up in the configuration of 2.3a

Figure 2.3: Configuration transitions in Hoppe's urn

Case **1;** $a_{n+1} = 1$**:** In this scenario, one allelic type fully dominates the entire population at the $n+1$th generation. Then, the the partition after $n+1$ draws must be

$$\mathbf{A}_{n+1} = \begin{pmatrix} a_1 & a_2 & \ldots & a_{n+1} \end{pmatrix} = \begin{pmatrix} 0 & 0 & \ldots & 1 \end{pmatrix},$$

with one allele of abundance $n+1$. But this can only happen by a repetitive draw of allele type 1, for the first n draws. Such an event occurs with probability:

$$P\big(\mathbf{A}_{n+1} = \begin{pmatrix} 0 & 0 & \ldots & 1 \end{pmatrix}\big) = 1\frac{1}{\theta+1}\frac{2}{\theta+2}\frac{3}{\theta+3}\cdot\ldots\cdot\frac{n}{\theta+n}.$$

Case **2;** $a_{n+1} = 0$**:** For the $n+1$th generation, consider a feasible partition

$$\mathbf{A}_{n+1} = \begin{pmatrix} a_1 & a_2 & \ldots & a_{n+1} \end{pmatrix}.$$

Then one can distinguish between two sub-cases: First, if one assumes that a special colored ball is drawn in the nth generation – that is, the number of species of size 1, increases by exactly one. This case arises with probability ${}^{\theta}\!/_{\theta+n}$. A second sub-case is the case where no special colored ball is drawn in the nth generation but a ball of a color of size $i \in \{1, 2, 2, \ldots, n\}$ is sampled. If

$$\mathbf{B}_n = \mathbf{b}_n = \begin{pmatrix} b_1 & b_2 & \ldots & b_n \end{pmatrix},$$

is a previous partition, and a ball of color of size i is drawn, then $\begin{pmatrix} a_1 & a_2 & \ldots & a_{n+1} \end{pmatrix}$, is realized via $\begin{pmatrix} b_1 & b_2 & \ldots & b_n \end{pmatrix}$ through the process of adding a ball of a certain color to the set of balls of that color which are of size i and hence in the partition at the $n+1$th generation the number of colors which are of size i is equal to the number of colors which are of size i in the nth generation (that is b_i) minus one and the number of colors which are of size $i+1$ is equal to the number of colors which are of size $i+1$ in the nth generation (that is b_{i+1}) plus one – we have that

$$a_i = b_i - 1, \quad \text{and} \quad a_{i+1} = b_{i+1} + 1,$$

which gives a recursive formula for a realization of $\mathbf{B}_n$ as

$$\mathbf{b}_n = \begin{pmatrix} a_1 & a_2 & \ldots & a_{i-1} & a_i + 1 & a_{i+1} - 1 & a_{i+2} & \ldots & a_n \end{pmatrix}.$$

The probability associated to this case is given by the probability to draw a ball of a color which is of size i in the nth generation, which happens with probability ${}^{i(a_i+1)}\!/_{\theta+n}$.

The statement in the following theorem is known as *Ewens Sampling Formula* and it is basically established by working, inductively, with the above boundary cases:

Theorem 2.9 (Ewens, 1972, *Ewens Sampling Formula*:)**.** In a Hoppe's urn starting out with $\theta \geq 1$ balls of special color, let $\mathbf{A}_n = \begin{pmatrix} A_1 & A_2 & \dots & A_n \end{pmatrix}$ be a partition of the total population of all non-special colored balls in such a way that for each $i \in \{1, 2, \dots, n\}$, A_i gives the number of colors represented by i balls. Then if we consider a feasible realization $\mathbf{a}_n = \begin{pmatrix} a_1 & a_2 & \dots & a_n \end{pmatrix}$, we have that

$$P(\mathbf{A}_n = \mathbf{a}_n) = \frac{n!\theta^{a_1+a_2+\cdots+a_n}}{1^{a_1}2^{a_2} \cdot \ldots \cdot n^{a_n} a_1!a_2! \cdot \ldots \cdot a_n! \langle\theta\rangle_n}.$$

For a complete, rigorous argument, I refer to Mahmoud (2008, p. 181, Theorem 9.3).

Chapter 3

Genotypic Composition of diploid Organisms

A diploid organism is characterized by the fact that for any given cell, one finds two copies of each chromosome. In the language of genetics, one talks about a pair of homologous chromosomes, referring to the origin of the chromosome pair – maternal or paternal origin. Of course, for a given genetic locus, that is for a certain physical locus, on the chromosome, of a given gene, the genetic code of the maternal and paternal copies are not necessarily the same. In case of different copies, it is common to distinguish, for the given gene, on the given locus, between the *wild type* (call it WT) and the *mutant* (wt), non-standard, version of the gene – one talks about the wild type, or mutant, allele at the given locus.

In the following, we are going to fix a certain genetic locus of a certain chromosome in a diploid organism and assume that the given gene can appear as wild type or mutant – hence, a certain organism can have three possible states, referring to the genotypic composition at the given locus: A wild type and a mutant composition, two wild type copies, or two mutant copies. The first composition is defined to be a heterozygous composition, whereas the latter two are homozygous compositions. Notation wise, we can define the genetic locus via

$$\{WT, wt\}, \quad \{WT, WT\}, \quad \text{and} \quad \{wt, wt\}.$$

Towards an urn analogy, let me define three different colors, referring to the three different states above: The heterozygous composition is represented by a grey colored balls (G), the homozygous, double wild type, compositions is given by black colored balls (B), and the homozygous, double mutant, composition, is given by white colored balls (W). In conclusion, by reducing certain individuals, of a given species, to the given genetic locus, a population of individuals can be characterized by the three colors G, B, and W. If two individuals decide to mate, each of the (parent) individuals will undergo a process called meiosis, that is, the parents will halve their set of chromosomes, to produce, if everything goes well, offspring which is again diploid. With respect to our given locus of interest, we can introduce two independent Bernoulli random variables which indicate, in the case of a heterozygous state, whether the wild type or the mutant version of the gene is forwarded to the next generation.

Explicitly, in the heterozygous state, let us name the two mating individuals as individuals $\mathcal{A}$ and $\mathcal{B}$. Then, let $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$ be two independent copies of $X \sim \text{Bernoulli}(p)$ random variables which indicate, for either parent $\mathcal{A}$ or parent $\mathcal{B}$ whether the wild type version, WT, or the mutant version, wt, is forwarded. By assuming a population of one given species, the success rate p, giving the probability of successfully forwarding the wild type version of the gene, is assumed to be the same for both parents. One natural question to ask could be the following: Given an initial, genotyic, composition of individuals, what can we say about the genotypic composition of the individuals after a long period of mating? Covering the full complexity of the above question seems ambitious. In a simplified setting, where one assumes no depletion of the population, no environmental influences (no complex natural selection processes), and no gender specific mating, that is at any given stage of the population, one does not attribute sexual characteristics to the individuals and a successful mating is given by a random, uniform (each pair equally likely) draw of two individuals. Throughout the remaining, if not else given, we are going to assume such a simplified scenario.

Given such a setting, we could translate the above question into the language of an urn, where balls are individuals, colors specify the diploid genotype, and the process of successful mating is given by a multiset draw of the colors G, B, and W. Hence, the question becomes translated as follows: What can we say about the number of gray, black, and white colored balls after $n \geq 0$ draws? The Pólya urn scheme, reflecting the above scenario, can be represented by the balanced 6×3 multiset-draw scheme

$$\begin{array}{c} \\ \{G,\,G\} \\ \{G,\,B\} \\ \{G,\,W\} \\ \{B,\,B\} \\ \{B,\,W\} \\ \{W,\,W\} \end{array} \begin{array}{c} \begin{array}{ccc} \qquad G \qquad & \qquad B \qquad & \qquad W \qquad \end{array} \\ \begin{pmatrix} X_{\mathcal{A}}(1-X_{\mathcal{B}})+X_{\mathcal{B}}(1-X_{\mathcal{A}}) & X_{\mathcal{A}}X_{\mathcal{B}} & (1-X_{\mathcal{A}})(1-X_{\mathcal{B}}) \\ 1-X & X & 0 \\ X & 0 & 1-X \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} \end{array}, \tag{3.1}$$

where the Bernoulli random variables $X_{\mathcal{A}}$, $X_{\mathcal{B}}$ and X are operating as introduced earlier: If the multiset $\{G,\,G\}$ is drawn, reflecting a mating of two heterozygous parents, $\mathcal{A}$ and $\mathcal{B}$, we add an independent copy of, $X_{\mathcal{A}}(1-X_{\mathcal{B}})+X_{\mathcal{B}}(1-X_{\mathcal{A}})$ balls of color G (heterozygous offspring of genotype $\{WT,\,wt\}$), $X_{\mathcal{A}}X_{\mathcal{B}}$ balls of color B (homozygous offspring of genotype $\{WT,\,WT\}$), and $(1-X_{\mathcal{A}})(1-X_{\mathcal{B}})$ balls of color W (homozygous offspring of genotype $\{wt,\,wt\}$). If the drawn multiset is given by $\{G,\,B\}$, we add an independent copy of $1-X$ balls of color G (reflecting the fact that the heterozygous parent has given the mutant version of the gene), X balls of color B (the heterozygous parent has forwarded the wild type version of the gene), and zero balls of color W. Symmetrically, If the pair $\{G,\,W\}$ is drawn, we add an independent copy of X balls of color G, zero balls of color B, and $1-X$ balls of color W. Notice, given any of the rows, reflecting a $\{G,\,G\}$, $\{G,\,B\}$, or $\{G,\,W\}$ bag of balls draw, only one row entry is realized as 1 and the others are 0. Finally, if the drawn pair is such that either both parents are homozygous wildtype ($\{B,\,B\}$), of mixed form (one parent is homozygous wild type and the other is homozygous mutant, $\{B,\,W\}$) or both parents are homozygous mutant ($\{W,\,W\}$), we add, respectively, zero balls of color G and W and one ball of color B, zero balls of color B and W and one ball of color G, or, zero balls of color G and B and one ball of color W. Of course, as it is the nature of Pólya urn schemes, each time a pair is drawn, it is positioned back in the urn. Scheme (3.1) has constant row sum given by 1 and hence, if τ_0 gives the initial number of balls in the urn, after n draws according to scheme (3.1), we have a total of $\tau_n = \tau_0 + n$ balls in the urn.

Now, towards an asymptotic description of the urn, let us define G_n, B_n, and W_n, to be, respectively, the number of grey, black, or white colored balls after n draws form the scheme (3.1). Further, in the spirit of example 2.9, we have six indicator variables which reflect the multiset-drawing workflow:

$$\mathbb{1}_n^{\{G,\,G\}}, \quad \mathbb{1}_n^{\{G,\,B\}}, \quad \mathbb{1}_n^{\{G,\,W\}}, \quad \mathbb{1}_n^{\{B,\,B\}}, \quad \mathbb{1}_n^{\{B,\,W\}}, \quad \text{and} \quad \mathbb{1}_n^{\{W,\,W\}},$$

where, generally, $\mathbb{1}_n^{\{X,\,Y\}}$, $X,\,Y \in \{G,\,B,\,W\}$, indicates the drawing of the multiset $\{X,\,Y\}$ at the nth step. With respect to the three colors, we can set up three recurrence equations:

$$\begin{aligned}
G_n &= G_{n-1} + \big(X_{\mathcal{A}}(1-X_{\mathcal{B}}) + X_{\mathcal{B}}(1-X_{\mathcal{A}})\big)\mathbb{1}_n^{\{G,\,G\}} \\
&\quad + (1-X)\mathbb{1}_n^{\{G,\,B\}} + X\mathbb{1}_n^{\{G,\,W\}} + \mathbb{1}_n^{\{B,\,W\}} \\
B_n &= B_{n-1} + \big(X_{\mathcal{A}}X_{\mathcal{B}}\big)\mathbb{1}_n^{\{G,\,G\}} + X\mathbb{1}_n^{\{G,\,B\}} + \mathbb{1}_n^{\{B,\,B\}} \\
W_n &= W_{n-1} + \big((1-X_{\mathcal{A}})(1-X_{\mathcal{B}})\big)\mathbb{1}_n^{\{G,\,G\}} + (1-X)\mathbb{1}_n^{\{G,\,W\}} + \mathbb{1}_n^{\{W,\,W\}}.
\end{aligned}$$

If we let $\mathcal{F}_n$, $n \geq 0$, to be the sigma algebra generated by the last n draws, we obtain, by conditioning on the previous draw:

$$\begin{aligned}
\mathrm{E}(G_n|\mathcal{F}_{n-1}) &= G_{n-1} + \mathrm{E}\left(X_{\mathcal{A}}(1-X_{\mathcal{B}}) + X_{\mathcal{B}}(1-X_{\mathcal{A}})\right)\mathrm{E}(\mathbb{1}_n^{\{G,\,G\}}|\mathcal{F}_{n-1}) \\
&\quad + \mathrm{E}\left(1-X\right)\mathrm{E}(\mathbb{1}_n^{\{G,\,B\}}|\mathcal{F}_{n-1}) \\
&\quad + \mathrm{E}(X)\,\mathrm{E}(\mathbb{1}_n^{\{G,\,W\}}|\mathcal{F}_{n-1}) + \mathrm{E}(\mathbb{1}_n^{\{B,\,W\}}|\mathcal{F}_{n-1}) \\
\mathrm{E}(B_n|\mathcal{F}_{n-1}) &= B_{n-1} + \mathrm{E}\left(X_{\mathcal{A}}X_{\mathcal{B}}\right)\mathrm{E}(\mathbb{1}_n^{\{G,\,G\}}|\mathcal{F}_{n-1}) \\
&\quad + \mathrm{E}(X)\,\mathrm{E}(\mathbb{1}_n^{\{G,\,B\}}|\mathcal{F}_{n-1}) + \mathrm{E}(\mathbb{1}_n^{\{B,\,B\}}|\mathcal{F}_{n-1}) \\
\mathrm{E}(W_n|\mathcal{F}_{n-1}) &= W_{n-1} + \mathrm{E}\left((1-X_{\mathcal{A}})(1-X_{\mathcal{B}})\right)\mathrm{E}(\mathbb{1}_n^{\{G,\,G\}}|\mathcal{F}_{n-1}) \\
&\quad + \mathrm{E}\left(1-X\right)\mathrm{E}(\mathbb{1}_n^{\{G,\,W\}}|\mathcal{F}_{n-1}) + \mathrm{E}(\mathbb{1}_n^{\{W,\,W\}}|\mathcal{F}_{n-1}),
\end{aligned}$$

where I have used the fact that, given a pair which consists of at least one heterozygous parent, the indicator of giving either wild type or mutant is independent of, not only which pair is drawn, but also on the previous draw. In a more rigorous language, we have that the Bernoulli random variables, $X_{\mathcal{A}}$, $X_{\mathcal{B}}$, and X, are independent of $\sigma(\mathbb{1}_n^{\{X,\,Y\}}, \mathcal{F}_{n-1})$, for $X,\,Y \in \{G,\,B,\,W\}$. In conclusion, if we make use of the assumption that $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$ are independent Bernoulli random variables with parameter p and further use the relations,

$$\mathrm{E}(\mathbb{1}_n^{\{G,\,G\}}|\mathcal{F}_{n-1}) = \frac{\binom{G_{n-1}}{2}}{\binom{\tau_{n-1}}{2}}, \qquad \mathrm{E}(\mathbb{1}_n^{\{G,\,B\}}|\mathcal{F}_{n-1}) = \frac{G_{n-1}B_{n-1}}{\binom{\tau_{n-1}}{2}},$$

$$\mathrm{E}(\mathbb{1}_n^{\{G,\,W\}}|\mathcal{F}_{n-1}) = \frac{G_{n-1}W_{n-1}}{\binom{\tau_{n-1}}{2}}, \qquad \mathrm{E}(\mathbb{1}_n^{\{B,\,B\}}|\mathcal{F}_{n-1}) = \frac{\binom{B_{n-1}}{2}}{\binom{\tau_{n-1}}{2}},$$

$$\mathrm{E}(\mathbb{1}_n^{\{B,\,W\}}|\mathcal{F}_{n-1}) = \frac{B_{n-1}W_{n-1}}{\binom{\tau_{n-1}}{2}}, \qquad \mathrm{E}(\mathbb{1}_n^{\{W,\,W\}}|\mathcal{F}_{n-1}) = \frac{\binom{W_{n-1}}{2}}{\binom{\tau_{n-1}}{2}},$$

we arrive at the following recurrence system for the conditional expectation:

(S1) $\mathrm{E}(G_n|\mathcal{F}_{n-1}) = G_{n-1} + 2p(1-p)\frac{\binom{G_{n-1}}{2}}{\binom{\tau_{n-1}}{2}} + (1-p)\frac{G_{n-1}B_{n-1}}{\binom{\tau_{n-1}}{2}} + p\frac{G_{n-1}W_{n-1}}{\binom{\tau_{n-1}}{2}} + \frac{B_{n-1}W_{n-1}}{\binom{\tau_{n-1}}{2}}.$

(S2) $\mathrm{E}(B_n|\mathcal{F}_{n-1}) = B_{n-1} + p^2\frac{\binom{G_{n-1}}{2}}{\binom{\tau_{n-1}}{2}} + p\frac{G_{n-1}B_{n-1}}{\binom{\tau_{n-1}}{2}} + \frac{\binom{B_{n-1}}{2}}{\binom{\tau_{n-1}}{2}}.$

(S3) $\mathrm{E}(W_n|\mathcal{F}_{n-1}) = W_{n-1} + (1-p)^2\frac{\binom{G_{n-1}}{2}}{\binom{\tau_{n-1}}{2}} + (1-p)\frac{G_{n-1}W_{n-1}}{\binom{\tau_{n-1}}{2}} + \frac{\binom{W_{n-1}}{2}}{\binom{\tau_{n-1}}{2}}.$

If one defines the function f as the conditional expectation with respect to the last draw, the above equations form a system of equations, which is non-linear in f (applied to a certain color) and has dependencies up to second degree. Mahmoud (2013) has discussed solutions for the case of multiset drawings, of size two, when the number of different colors is two. The discussion had focus on a special case, where one is able to reduce the potential quadratic equation for a given color to a linear one (consider the Chen-Wei Urn in example 2.9). An analysis of $k \geq 1$ colors, in the case of zero-balanced urns (constant row sum of zero, as for example the Ehrenfest urn), is discussed in Konzem and Mahmoud (2016). With regard to the urn scheme 3.1, I can provide a result on the expected difference of black and white balls after n draws in the special case of $p = 1/2$. The statement can be summarized in a proposition:

Proposition 3.1. Let B_n and W_n be, respectively, the number of black and white balls after n draws, in an urn evolving according to the scheme 3.1 with success parameter $p = 1/2$. Then, the expected difference, between the number of white and black balls, after n draws, is given by

$$\mathrm{E}(W_n - B_n) = \frac{(W_0 - B_0)}{\tau_0} n + (W_0 - B_0).$$

Remark 3.1. Proposition 3.1 supports a natural intuition: If the urn starts out with $W_0 > B_0$ (or $W_0 < B_0$) balls, then at any instance of discrete time, the expected number of white balls is strictly greater (or strictly smaller) than the expected number of black balls. In the situation of $W_0 = B_0$, the equality remains throughout the sampling steps, that is given an equal composition of white and black balls in the initial urn, in expectation, the number of black and white balls after n draws are still the same.

Proof of Proposition 3.1. First, by using the fact that $p^2 = (1-p)^2$, we can get rid of the second order terms involving G_n^2, by subtracting (S2) from (S3):

$$\begin{aligned}\mathrm{E}(W_n - B_n|\mathcal{F}_{n-1}) = W_{n-1} - B_{n-1} &+ p\frac{G_{n-1}W_{n-1}}{\binom{\tau_{n-1}}{2}} - p\frac{G_{n-1}B_{n-1}}{\binom{\tau_{n-1}}{2}} \\ &+ \frac{\binom{W_{n-1}}{2}}{\binom{\tau_{n-1}}{2}} - \frac{\binom{B_{n-1}}{2}}{\binom{\tau_{n-1}}{2}}.\end{aligned}$$

Then, in a next step, we can replace G_{n-1} by $\tau_{n-1} - W_{n-1} - B_{n-1}$, which gives:

$$\begin{aligned}\mathrm{E}(W_n - B_n|\mathcal{F}_{n-1}) = W_{n-1} - B_{n-1}& \\ &+ p\frac{(\tau_{n-1} - W_{n-1} - B_{n-1})W_{n-1}}{\binom{\tau_{n-1}}{2}} - p\frac{(\tau_{n-1} - W_{n-1} - B_{n-1})B_{n-1}}{\binom{\tau_{n-1}}{2}} \\ &+ \frac{\binom{W_{n-1}}{2}}{\binom{\tau_{n-1}}{2}} - \frac{\binom{B_{n-1}}{2}}{\binom{\tau_{n-1}}{2}}.\end{aligned}$$

Then by developing the above expression, we have:

$$\begin{aligned}\mathrm{E}(W_n - B_n|\mathcal{F}_{n-1}) = W_{n-1} - B_{n-1}& \\ &+ p\frac{\tau_{n-1}W_{n-1}}{\binom{\tau_{n-1}}{2}} - p\frac{W_{n-1}^2}{\binom{\tau_{n-1}}{2}} - p\frac{B_{n-1}W_{n-1}}{\binom{\tau_{n-1}}{2}} \\ &- p\frac{\tau_{n-1}B_{n-1}}{\binom{\tau_{n-1}}{2}} + p\frac{W_{n-1}B_{n-1}}{\binom{\tau_{n-1}}{2}} + p\frac{B_{n-1}^2}{\binom{\tau_{n-1}}{2}} \\ &+ \frac{\binom{W_{n-1}}{2}}{\binom{\tau_{n-1}}{2}} - \frac{\binom{B_{n-1}}{2}}{\binom{\tau_{n-1}}{2}},\end{aligned}$$

which simplifies to

$$\begin{aligned}\mathrm{E}(W_n - B_n|\mathcal{F}_{n-1}) = W_{n-1} - B_{n-1} &+ p\frac{\tau_{n-1}W_{n-1}}{\binom{\tau_{n-1}}{2}} - p\frac{\tau_{n-1}B_{n-1}}{\binom{\tau_{n-1}}{2}} \\ &- p\frac{W_{n-1}^2}{\binom{\tau_{n-1}}{2}} + p\frac{B_{n-1}^2}{\binom{\tau_{n-1}}{2}} + \frac{\binom{W_{n-1}}{2}}{\binom{\tau_{n-1}}{2}} - \frac{\binom{B_{n-1}}{2}}{\binom{\tau_{n-1}}{2}}.\end{aligned}$$

Now, under the assumption that $p = 1/2$, we have that

$$\frac{\binom{W_{n-1}}{2}}{\binom{\tau_{n-1}}{2}} = \frac{\frac{(W_{n-1})(W_{n-1}-1)}{2}}{\binom{\tau_{n-1}}{2}} = p\frac{W_{n-1}^2}{\binom{\tau_{n-1}}{2}} - p\frac{W_{n-1}}{\binom{\tau_{n-1}}{2}},$$

and

$$\frac{\binom{B_{n-1}}{2}}{\binom{\tau_{n-1}}{2}} = \frac{\frac{(B_{n-1})(B_{n-1}-1)}{2}}{\binom{\tau_{n-1}}{2}} = p\frac{B_{n-1}^2}{\binom{\tau_{n-1}}{2}} - p\frac{B_{n-1}}{\binom{\tau_{n-1}}{2}}.$$

Hence, by replacing the respective terms, we can further simplify to arrive at

$$\begin{aligned}\mathrm{E}(W_n - B_n|\mathcal{F}_{n-1}) = W_{n-1} - B_{n-1} &+ p\frac{\tau_{n-1}W_{n-1}}{\binom{\tau_{n-1}}{2}} - p\frac{\tau_{n-1}B_{n-1}}{\binom{\tau_{n-1}}{2}} \\ &- p\frac{W_{n-1}}{\binom{\tau_{n-1}}{2}} + p\frac{B_{n-1}}{\binom{\tau_{n-1}}{2}}.\end{aligned}$$

Rewriting the above recurrence, brings us to a linear system in $W_n - B_n$:

$$\mathrm{E}(W_n - B_n|\mathcal{F}_{n-1}) = \left(1 + p\frac{\tau_{n-1}}{\binom{\tau_{n-1}}{2}} - \frac{p}{\binom{\tau_{n-1}}{2}}\right)(W_{n-1} - B_{n-1}),$$

which becomes further simplified to

$$\mathrm{E}(W_n - B_n|\mathcal{F}_{n-1}) = \left(1 + p\frac{\tau_0 + (n-2)}{\binom{\tau_{n-1}}{2}}\right)(W_{n-1} - B_{n-1}).$$

But then, if we replace

$$\binom{\tau_{n-1}}{2} = \frac{(\tau_0 + (n-1))(\tau_0 + (n-2))}{2},$$

we have with $p = 1/2$

$$\mathrm{E}(W_n - B_n|\mathcal{F}_{n-1}) = \left(1 + \frac{1}{\tau_0 + (n-1)}\right)(W_{n-1} - B_{n-1}) = \left(\frac{\tau_0 + n}{\tau_0 + n - 1}\right)(W_{n-1} - B_{n-1}).$$

Taking expectation, we arrive at the recurrence

$$\mathrm{E}(W_n - B_n) = \left(\frac{\tau_0 + n}{\tau_0 + n - 1}\right) \mathrm{E}(W_{n-1} - B_{n-1}),$$

which is solved, in the same way as shown in example 2.9, to

$$\mathrm{E}(W_n - B_n) = \frac{(W_0 - B_0)}{\tau_0} n + (W_0 - B_0).$$

□

Finally, for developing a better intuition of the urn, with scheme (3.1), I have considered implementing the urn workflow in `R` (R Core Team, 2019, version 3.6.0). The results and a discussion are found in the Appendix. The appearance of the urn, gives one example of a, non-simple, Pólya urn which could be potentially interesting in the field of genetics. Initially, having in mind the structure of sheme (3.1), I have intended to come up with an explicit description of the expected number of homozygous mutant individuals after $n \geq 1$ mating steps. Although the urn is in a sparse and balanced condition, approaching the problem via the system of equations (S1), (S2), and (S3), seems rather demanding and one might consider a change of perspective.

Appendix: R Implementation

Towards a better understanding of the long time behavior of the urn, introduced in scheme (3.1) of chapter 3, I have considered a minimal simulation with the computer – The simulation was done with R (R Core Team, 2019, version 3.6.0). The results are depicted in figures A1 and A2.

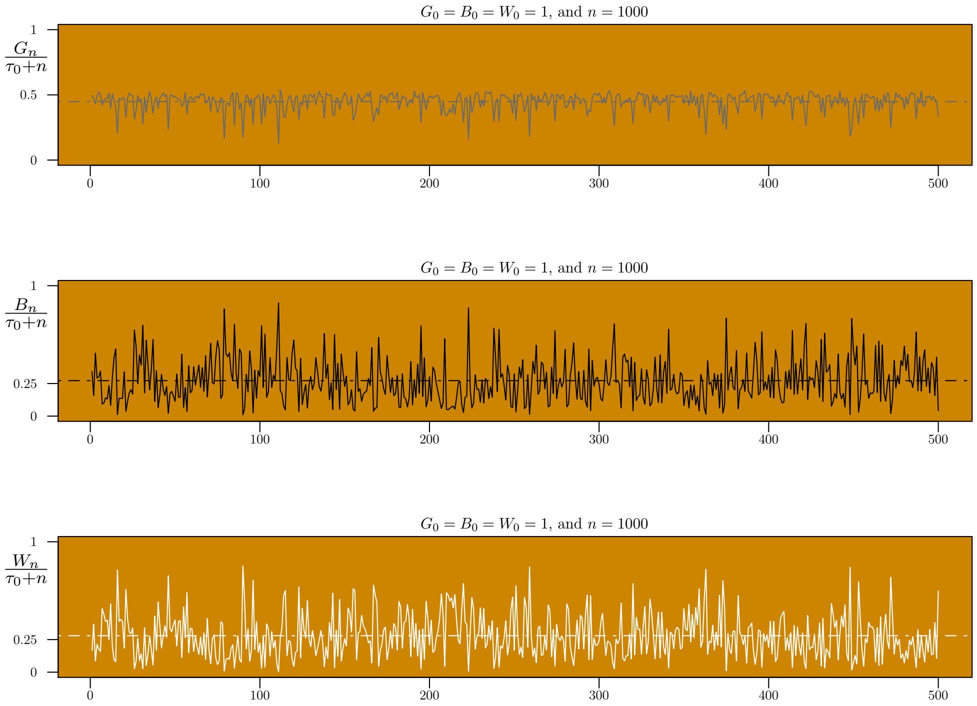


Figure A1: Simulation results for the composition of the urn, following scheme (3.1) of chapter 3: The initial composition of the urn is such that $G_0 = B_0 = W_0 = 1$ and hence $\tau_0 = 3$. For each panel, the horizontal axis labels the number of conducted experiments (1, 2, . . . , 500) and each given experiment represents a sequence of 1000 draws, following the initial urn, according to scheme (3.1). Depicted are, for each panel, for each genotype (each color, G, B, and W), the relative number of balls, given with a precision of 3 digits, after $n = 10000$ draws. The top panel shows the relative number of grey balls after 1000 draws for each of the 500 conducted simulation runs. The dashed line represents the sample average for the relative number of grey balls after 1000 draws over all the 500 experiments. The middle and lower panel show, respectively, the same results for black and white colored balls.

As a support of proposition 3.1, we can conclude the following: In the situation of a balanced starting composition of the urn (consider figure A1, with $G_0 = B_0 = W_0 = 1$), the relative number of both, homozygous wild type individuals and homozygous mutant individuals, fluctuate around approximately one quarter throughout the 500 experiments. On the other hand, in the

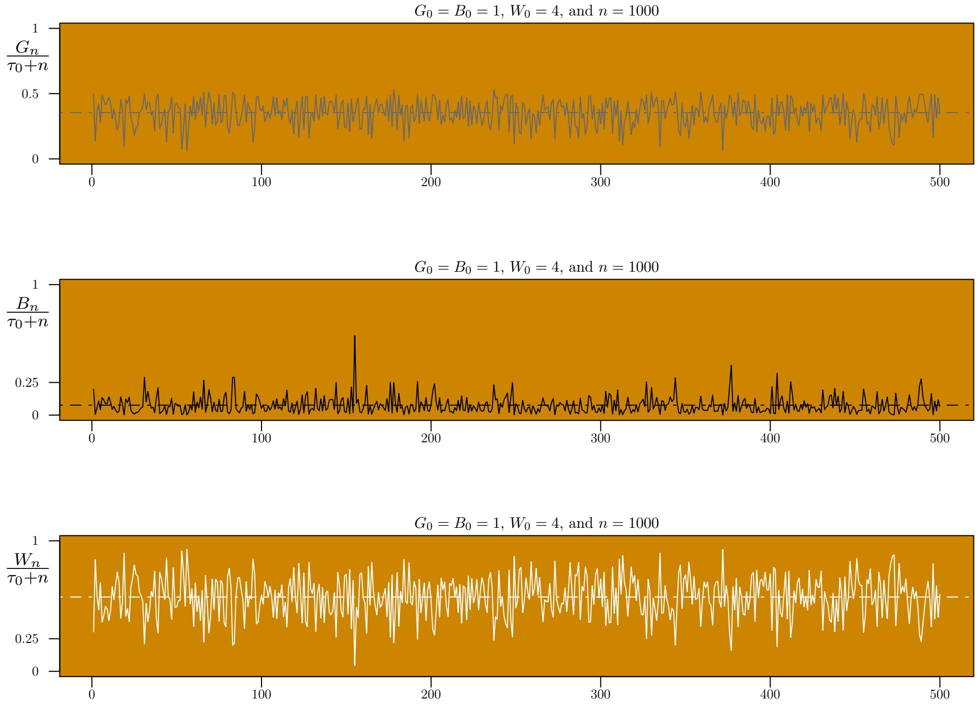


Figure A2: Simulation results for the composition of the urn, following scheme (3.1) of chapter 3: The initial composition of the urn is such that $G_0 = B_0 = 1$ and $W_0 = 4$ and hence $\tau_0 = 6$, with a head start for the homozygous mutant genotype. For each panel, the horizontal axis labels the number of conducted experiments (1, 2, ..., 500) and each given experiment represents a sequence of 1000 draws, following the initial urn, according to scheme (3.1). Depicted are, for each panel, for each genotype (each color, G, B, and W), the relative number of balls, given with a precision of 3 digits, after $n = 10000$ draws. The top panel shows the relative number of grey balls after 1000 draws for each of the 500 conducted simulation runs. The dashed line represents the sample average for the relative number of grey balls after 1000 draws over all the 500 experiments. The middle and lower panel show, respectively, the same results for black and white colored balls.

case of a head start for the number of homozygous mutant individuals (consider figure A2, with $G_0 = B_0 = 1$ and $W_0 = 4$), the relative number of homozygous mutant individuals remains, on average, greater than the relative number of homozygous wild type individuals throughout the 500 experiments. A respective example code, for implementing the urn, with two seeds which lead to the given results, is given bellow:

```r
library("compare")

#rm(list=ls())
#set.seed(1234): balanced start (1,1,1), 500 experiments, 1000 draws
#set.seed(12345): unbalanced start (1,1,4), 500 experiments, 1000 draws

number_of_experiments <- 500
number_of_draws <- 1000
colors <- c("grey", "black", "white")
initial_number_of_balls <- c(1,1,1)
```

```
color_dist_after_n <- replicate(number_of_experiments, {
  urn <- rep(colors, initial_number_of_balls)
  ball_counter <- as.list(initial_number_of_balls)
  names(ball_counter) <- colors

  for (i in 1:number_of_draws) {
    draw <- sample(urn, 2)
    if (as.logical(compare(draw, c("grey", "grey"),
                           ignoreOrder = TRUE)[1]) == TRUE) {
     X <- rbinom(1, size =1, prob = 0.5)
     Y <- rbinom(1, size =1, prob = 0.5)
     ball_counter$grey <- ball_counter$grey + (X*(1-Y) + Y*(1-X))
     ball_counter$black <- ball_counter$black + (X*Y)
     ball_counter$white <- ball_counter$white + ((1-X)*(1-Y))
     result <- data.frame(color = colors,
                          indi = c((X*(1-Y) + Y*(1-X)), (X*Y), ((1-X)*(1-Y))))
     color_result <- as.character(result[result$indi == 1, "color"])
     urn <- c(urn, color_result)
    } else if (as.logical(compare(draw, c("grey", "black"),
                                  ignoreOrder = TRUE)[1]) == TRUE) {
      X <- rbinom(1, size =1, prob = 0.5)
      Y <- rbinom(1, size =1, prob = 0.5)
      ball_counter$grey <- ball_counter$grey + (1-X)
      ball_counter$black <- ball_counter$black + X
      ball_counter$white <- ball_counter$white + 0
      result <- data.frame(color = colors, indi = c((1-X), X, 0))
      color_result <- as.character(result[result$indi == 1, "color"])
      urn <- c(urn, color_result)
    } else if (as.logical(compare(draw, c("grey", "white"),
                                  ignoreOrder = TRUE)[1]) == TRUE) {
      X <- rbinom(1, size =1, prob = 0.5)
      Y <- rbinom(1, size =1, prob = 0.5)
      ball_counter$grey <- ball_counter$grey + X
      ball_counter$black <- ball_counter$black + 0
      ball_counter$white <- ball_counter$white + (1-X)
      result <- data.frame(color = colors, indi = c(X, 0, (1-X)))
      color_result <- as.character(result[result$indi == 1, "color"])
      urn <- c(urn, color_result)
    } else if (as.logical(compare(draw, c("black", "black"),
                                  ignoreOrder = TRUE)[1]) == TRUE) {
      ball_counter$grey <- ball_counter$grey + 0
      ball_counter$black <- ball_counter$black + 1
      ball_counter$white <- ball_counter$white + 0
      urn <- c(urn, "black")
    } else if (as.logical(compare(draw, c("black", "white"),
                                  ignoreOrder = TRUE)[1]) == TRUE) {
      ball_counter$grey <- ball_counter$grey + 1
      ball_counter$black <- ball_counter$black + 0
      ball_counter$white <- ball_counter$white + 0
      urn <- c(urn, "grey")
    } else if (as.logical(compare(draw, c("white", "white"),
                                  ignoreOrder = TRUE)[1]) == TRUE) {
      ball_counter$grey <- ball_counter$grey + 0
      ball_counter$black <- ball_counter$black + 0
      ball_counter$white <- ball_counter$white + 1
      urn <- c(urn, "white")
    }
}
  dist_after_n <- lapply(ball_counter,
                         FUN = function(x){round(x/sum(as.numeric(ball_counter)),
                                                 digits = 3)})
})
```

Bibliography

Athreya, K. B. and Karlin, S. (1968). Embedding of urn schemes into continuous time markov branching processes and related limit theorems. *The Annals of Mathematical Statistics*, **39**, 1801–1817. 2, 16, 24

Bagchi, A. and Pal, A. (1985). Asymtotic normality in the generalized Pólya-Eggenberger urn model, with an application to computer data structures. *SIAM Journal on Algebraic Discrete Methods*, **6**, 394–405. 12

Balaji, S., Mahmoud, H., and Watanabe, O. (2006). Distributions in the Ehrenfest process. *Statistics and Probability Letters*, **76**, 666–674. 11

Bernoulli, D. (1768). De usu algorithmi infinitesimales in arte conjectandi specimen. *Novi Comment. Acad. Sci. Imp. Petropolitanae*, **12**, 87–98. 1

Chen, M.-R. and Wei, C.-Z. (2005). A new urn model. *Journal of Applied Probability*, **42**, 964–976. 13

Doob, J. (1953). *Stochastic Processes*. Wiley Publications in Statistics. Wiley, New York. 4

Eggenberger, F. and Pólya, G. (1923). Über die Statistik verketteter Vorgänge. *Zeitschrift für Angewandte Mathematik und Mechanik*, **3**, 279–289. 1, 10, 12

Ewens, W. (1972). The sampling theory of selectively neutral alleles. *Theoretical Population Biology*, **3**, 87–112. 2, 31, 33, 34, 37

Hall, P. and Heyde, C. (1980). *Martingale limit theory and its application*. Academic Press, New York. 3, 4, 5

Hoppe, F. (1984). Polya-like urns and the ewens sampling formula. *Journal of Mathematical Biology*, **20**, 91–94. 2, 31

Johnson, N. and Kotz, S. (1977). *Urn models and their application: An approach to modern discrete probability theory*. Wiley, New York. 1, 2, 6, 7

Konzem, S. R. and Mahmoud, H. M. (2016). Characterization and enumeration of certain classes of tenable Pólya urns grown by drawing multisets of balls. *Methodology and Computing in Applied Probability*, **18**, 359–375. 2, 8, 10, 15, 42

Mahmoud, H. (2008). *Polya Urn Models*. Chapman & Hall/CRC, 1 edition. 2, 6, 10, 11, 12, 13, 15, 16, 17, 18, 19, 21, 22, 24, 25, 29, 32, 33, 34, 37

Mahmoud, H. (2013). Drawing multisets of balls from tenable balanced linear urns. *Probability in the Engineering and Informational Sciences*, **27**, 147–162. 2, 8, 14, 15, 42

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 44, 45

Smythe, R. (1996). Central limit theorems for urn models. *Stochastic Processes and their Applications*, **65**, 115–137. 2, 16, 19, 23, 24

Zelen, M. (1969). Play the winner rule and the controlled clinical trial. *Journal of the American Statistical Association*, **64**, 131–146. 25