Phases in the Two Color Tenable Zero-Balanced Pólya Process

by Joshua H. Sparks

B.S. in Mathematics, May 2010, Eastern Kentucky University, Richmond, Kentucky
B.A. in Sociology, May 2010, Eastern Kentucky University, Richmond, Kentucky
M.A. in Sociology, November 2012, University of Toronto, Toronto, Ontario

A Thesis submitted to

The Faculty of
The Columbian College of Arts and Sciences
of The George Washington University
in partial fulfillment of the requirements
for the degree of Master of Science

August 31, 2013

Thesis directed by

Hosam M. Mahmoud
Professor of Statistics

Dedication

The author wishes to dedicate this to Juby for being the driving force behind everything that he does.

Acknowledgment

The author wishes to acknowledge the Department of Statistics at The George Washington University and would like to express his gratitude for their support and financial assistance through the Jerome Cornfield Award and their graduate fellowship.

He would also like to thank Professor Hosam Mahmoud for his astounding support throughout the author's graduate tenure. the author knows he would not have made it through the progam without Professor Mahmoud's knowledge and guidance.

Abstract of Thesis

Phases in the Two Color Tenable Zero-Balanced Pólya Process

The Pólya process is obtained by embedding the usual (discrete-time) Pólya urn scheme in continuous time. We study the class of tenable Pólya processes of white and blue balls with zero balance (no change in n, the total number of balls, over time). This class includes the (continuous-time) Ehrenfest process and the (continuous-time) Coupon Collector's process. We look at the composition of the urn at time t_n (dependent on n). We identify a critical phase of t_n at the edges of which phase transitions occur. In the subcritical phase, under proper scaling the number of white balls is concentrated around a constant. In the critical phase, we have sufficient variability for an asymptotic normal distribution to be in effect. In this phase, the influence of the initial conditions is still somewhat pronounced. Beyond the critical phase, the urn is very well mixed with an asymptotic normal distribution, in which all initial conditions whither away. The results are obtained by an analytic approach utilizing partial differential equations.

Table of Contents

List of Tables

Chapter 1: General and Historical Perspectives

The fields of applied probability and stochastic processes have taken shape over the past centuries primarily as methods to explain the natural phenomena among us. Areas such as statistical physics and evolutionary biology have utilized models created by probabilists to make sense of fluid flow and population growth, and it is through these efforts that inspired the Dimer model and Durrett's probability models for DNA sequence evolution (2010). The purpose of this thesis is to take one of these famous probability models − the urn model problem − and generate a specialized case in order to derive its asymptotic behavior in infinite time.

The construction of the urn model problem stems from a very simple concept. In a two-color urn model, we start with an urn and fill it with two types of balls, say W_0 white balls and B_0 blue balls. Then, whenever we randomly draw a ball from the urn, we reach one of two results. If we select a white ball, then we adjust the number of white and blue balls within the urn by adding or removing a specified number of balls from each respective color. If we choose a blue ball, then we again adjust the ball counts, although typically we do so with different adjustment measures from that of the white draw. With each consecutive draw, we therefore construct a path for the distribution of balls that exist inside the urn. We can take this dynamical structure and represent it by using a replacement matrix of the form:

$$\begin{pmatrix} U & V \\ X & Y \end{pmatrix}; U, V, X, Y \in \mathbb{Z}. \tag{1.1}$$

The matrix above is beneficial in that it summarizes our actions to the urn after each draw. Each row represents the resulting adjustment of white and blue balls inside the urn following a draw of a specific color. For example, we may associate the first and second columns to be the adjustment counts of white and blue balls,

respectively, and the first row to be the event that a white ball was drawn (the second row determines that a blue ball was drawn). Thus, if the white ball is drawn, we add U white balls to the urn as well as V blue balls. As U and V may be any number out of the integers, we should note that a command of adding U white balls could instead mean to remove white balls from the urn (if $U < 0$) or simply to leave the white ball count the same (if $U = 0$). The equivalent routine can also be assigned to randomly selecting a blue ball from the urn and instead adding X white balls and Y blue balls to the container. Note that each of these entries may be deterministic (U is always the same number) or it may be randomized (V could be a c $\times$ Ber(p), a constant multiplied by a Bernoulli random variable with parameter p).

Jacob Bernoulli's work, *Ars Conjectandi* (1713), has been considered to be one of the initial studies of the inverse probability problem, a concept strongly associated to the urn model. His analysis consisted of determining the proportions of different colored pebbles within an urn, given the number of pebbles drawn from the container. Bernoulli's research on the subject attracted the attention of numerous scholars, including Thomas Bayes and thus lending its hand in Bayesian inference. To date, these models have been perpetuated by various disciplines to analyze occurances such as contagion, random walks, fluid diffusion, occupancy problems, and coupon collection.

In 1923, Eggenberger and Pólya first introduced the original Pólya-Eggenherger urn scheme (later simplified to the Pólya urn) to study the spreading of railway deaths contagious disease, and the spreading of plant genetics. In this model, the following replacement matrix was used:

$$\begin{pmatrix} U & 0 \\ 0 & U \end{pmatrix}; U \in \mathbb{N}. \tag{1.2}$$

Let us picture this model with respect to the plant scenario. Let there exist two

types of flowers in a specific garden, say daisies and petunias, both competing to reproduce within the garden. For any given plot of land, only one plant may grow at a time, as it will absorb the nutrients from other competing flowers. Also, for each plot of unoccupied land, we will assume that every plant in the garden offers one seed to germinate, and if the seed succeeds, it will produce U more flowers of the given plant. Finally, if we claim that this process happens once per week, then we may determine the eventual proportion of these two flowers through the urn model. If the daisy (the white ball) is selected from the garden (our urn), we add U daisies to that plot of land. However, if the petunia wins the germination battle, U more petunias will occupy the garden that week.

The Ehrenfest urn scheme (1907) shifted the mode of thinking about the urn models, creating a replacement matrix that left the urn with the same number of balls regardless of the number of draws:

$$\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}. \tag{1.3}$$

Such a scheme can be envisioned to model the mixing of two gases inside two separate chambers. As one gas particle leaves its chamber, it enters the opposing chamber, adjusting the number of particles in each compartment. The total number of particles stays the same; however, the proportions of particles in each chamber change with each draw. Schrödinger and Kohlrausch (1926) showed that this process can also be used to model a random walk on a bounded segment (for example, our plane lies within [-R, R]). We can see that for every step in one direction, we have one fewer unit of space from the boundary, and once we reach the boundary, we must begin the trip in the opposite direction. In both cases above, the replacement matrix possesses equal row sums, denoted as K, providing what is called a *balanced urn*

scheme.

In 1949, Bernard Friedman encountered Eggenberger and Pólya's idea and with the use of Schröedinger and Kohlrausch's work, he modified the replacement matrix for the urn to provide adjustments to both colors:

$$\begin{pmatrix} U & V \\ V & U \end{pmatrix}; U, V \in \mathbb{N}. \tag{1.4}$$

His paper weaves his generalized urn model and difference-differential equations into previously modeled phenomena as well as a study of accident counts and the number of safety campaigns associated with them. Feller (1968) continued this study and used the process to model branching processes and other well-established statistical problems.

Overall, two-color urn schemes evolving in discrete time have drawn a fair deal of attention over the last century due to their potential applications in many fields. Economists have employed urn models to visualize the resource movement of scarce items (oil, commerce, services) among competing alternatives. Arthur, Ermoliev, and Kaniovski (1984) introduces nonlinear Pólya urn models to study path-dependent processes that applied to industrial location theory and the emergence of technological structures in the economy. In 2004, Windrum developed a coupled urn model to simulate the browser war between Microsoft and Netscape and demonstrate the ways Microsoft was exploiting regulations.

Motivated by evolutionary processes in information technology, Samuel (1998) used the original Pólya urn model to assess microdata disclosure risk. In turn, Fienberg and Makov (2001) modified Samuel's Pólya urn scheme to study disclosure risk. In cognitive science, Navarro, Griffiths, Steyvers, and Lee (2005) modeled individual differences in personality by applying a Pólya urn to an infinite group of people us-

ing the stochastic Dirichlet process. Xue (2005) used Pólya urn models to study the conformity of an individuals behavior within a specific social structure.

In informatics, urns have been used to model recursive trees structure. Najock and Heyde (1982) created a model to construct the trees of preserved copies of ancient manuscripts. Gastwirth and Bhattacharya (1984) used them to model the social phenomenon of chain letters and pyramid schemes. Martin and Ho (2002) further applied a Pólya urn scheme to discuss gambling games. For additional literature and results on recursive trees, along with its connection to the Pólya urn, Mahmoud and Smythe's (1991) survey of the subject provides interesting examples of this discrete concept.

In population genetics, Benaim, Schreiber, and Tarres (2004) used Pólya urn models to describe the dynamics of finite populations of interacting genotypes. Mahmoud (2008) discusses the use of Pólya schemes in various bioscience applications. Adaptive design in clinical trials is also one of the classical applications of urn models. Early advances in adaptive design have been introduced by Thompson (1933) and Robbins (1952). For general background on urn models see Johnson and Kotz (1977), Kotz and Balakrishnan (1997) or Mahmoud (2008).

While the discrete Pólya scheme has shown to be a reliable model for numerous circumstances, there are events in which a continuous-time model would be desired. In certain models, the draws may not be evenly distributed: two disjoint time intervals may have an unequal number of draws, making the discrete case impractical. One instance is when a virus is spread to a host's cells. In this case, numerous cells could be infected over a period of time, while antibodies could instead slow down the infection and prevent this spread. Here, the discrete model would not be as desirable as it could not properly represent the infection times of this virus.

We may also wish to investigate the distribution of the urn in finite time. For example, we may wish to study the mixing patterns of gas particles or heat between two separated chambers. In applications like cooking and air-conditioning, one would prefer to know the amount of gas or heat that resides within a given space after a period of hours rather than over an infinite time span. These scenarios make the discrete Pólya urn scheme impractical as the time intervals between mixings are not all equal. It is when we begin to vary the time intervals between particle mixings that we abandon the discrete system in favor of a continuous one.

What distinguishes continuous-time Pólya processes from the usual discrete-time random urn schemes is the timing of the ball draws. In the discrete-time scheme, balls are picked at time intervals equally spaced over a designated era. With our plant model, for example, there was the assumption that the flower growth was decided once a week, thereby spacing our draw over a period of seven days. In the continuous-time process, each ball is endowed with a clock, and the point in which each clock rings is distributed identically to an exponential random variable with mean 1. The clocks are independent of one another as well as all other previous random variables. When a ball's clock rings − a *renewal point* or *epoch* in the Pólya process − we take it that the ball has been picked from the urn and we instantaneously execute the rules associated with its color.

With each new renewal process, all new balls added come endowed with their own independent clocks. The collective process enjoys a memoryless property as it is induced by independent clocks with exponential ringing times. Thus, once we execute an epoch, all clocks attached to their repective balls are reset as renewed Exp(1) random variables, causing our race to begin anew. The Pólya process therefore is Markovian, providing no memory of the past. However, the rates may change

depending on the number of balls added. In general, one should note that Markovian processes may consist of inhomogeneous jump processes occurring at each of the epochs.

As a result, there is a process associated with a Pólya urn scheme that is obtained by embedding the model in continuous time. This embedded stochastic process was named the Pólya process in Balaji and Mahmoud (2006). Furthermore, for readability and later reference, we will distinguish growth in discrete time to be a scheme, while a growth in continuous time will be considered as a "process."

The technique of transforming the discrete-time structure into a continuous-time model was introduced in Athreya and Karlin (1968) as a mathematical transform (Poissonization) to understand urns evolving in discrete-time. In this article, Athreya and Karlin report difficulty in the de-Poissonization to obtain the inverse transform. This is primarily due to the construction of the total number of balls, called τ_n, after some amount of time. In the discrete-time scheme, we are able to determine the number of draws that have been conducted, thereby deriving the number of possible outcomes that could result in this urn structure. However, in the continuous framework, there is no guarantee on the number of draws that have been conducted. The bell for the continuous-time structure may ring at any moment, and so there may be a large − and possibly infinite − number of draws over any given time interval (though with small probability).

Since τ_n is vital to determining the asymptotic behavior of these models, some restrictions have been made in our study of continuous-time Pólya processes in order to construct the appropriate analytical tools. One such constraint is that each Pólya process is *tenable*; a Pólya process is tenable if it is always possible to draw balls and apply our replacement matrix to every stochastic path. To illustrate what it means

to be tenable, suppose we have an urn of white and blue balls with the following replacement matrix:

$$\begin{pmatrix} -3 & 3 \\ 3 & -3 \end{pmatrix}. \tag{1.5}$$

If we begin with two white balls and three blue balls inside our urn, we develop a potential hazard: if our first draw were to be a white ball, we would be unable to legitimately remove three white balls from the urn. Therefore, in order to keep this stochastic process working correctly, we should affirm that there the number of white balls and the number of blue balls are both a multiple of three. This way, we will always be allowed to remove the required number of balls as directed by our replacement matrix.

Furthermore, we wish to deal with only classes of two-color tenable Pólya processes which are zero-balanced. This implies that when we sum the entries in each row of our replacement matrix, we end up with 0. In summary, every tenable zero-balanced Pólya process retains the same number of balls inside our urn − there is no need to worry about eventually having to find a bigger container! Instead, what matters within these urns are the proportions of balls of each color after each draw. For the purpose of this thesis we shall designate n to be the number of balls in our zero-balanced tenable Pólya process.

Kholfi and Mahmoud (2012b) have recently shown that the only possible replacement matrices for tenable zero-balanced urns are in the form:

$$c\begin{pmatrix} -\text{Ber}(p) & \text{Ber}(p) \\ \text{Ber}(r) & -\text{Ber}(r) \end{pmatrix}, \tag{1.6}$$

for a positive integer c that divides n, and $\text{Ber}(s)$ is a Bernoulli random variable with success probability $s \in [0, 1]$. Like Kholfi and Mahmoud (2012b), we also will

avoid situations in which $s = 0$, leading our replacement matrix to be reducible and changing the problem to resemble more of a coupon-collecting scenario. We may note that when the number of balls does not change, the Pólya process becomes a Markov chain with a well-understood stationary distribution.

While the study of the distribution of balls in phases of drawing in discrete Pólya urn schemes has recently received attention (see Balaji et al. (2010), Mahmoud (2010), Smythe (2011), Kholfi and Mahmoud (2012a; 2012b), Mahmoud and Smythe (2012)), this has yet to be done for continuous-time models. The study of phases is also of interest in related areas, such as the evolution of random hypergraphs (see Bender et al. (1997)) and occupancy problems (see Vatutin and Mikhailov (1982)). The book Kolchin et al. (1978) also gives a broad coverage of occupancy problems. Therefore, it is the goal of this thesis to investigate the phases within the two-color tenable zero-balanced Pólya process. Along the way, we shall determine its stationary distribution under each of these phases and find approximations for the probability distribution in the various phases.

Chapter 2: Methods and Theorem

2.1 Technical setup

The purpose of this section is to provide the backgound for our two-color urn model. Once sufficient preparation has been established, we can begin to delve into the theorem at hand. We shall begin with the construction of our Pólya urn as seen above:

$$\begin{pmatrix} U & V \\ X & Y \end{pmatrix}; U, V, X, Y \in \mathbb{Z}. \tag{2.1}$$

We will assume that this urn is zero-balanced and tenable, Let $W(t) = W_n(t)$ and $B(t) = B_n(t)$ be, respectively, the number of white and blue balls at time t in a tenable zero-balanced Pólya process on a total of n white and blue balls. Formally speaking, the two-color Pólya process is the two-component vector $\begin{pmatrix} W(t) \\ B(t) \end{pmatrix}$ which we may track at any value of t. Let

$$\phi_n(t, u, v) := \mathbf{E}[e^{uW(t)+vB(t)}] \tag{2.2}$$

be the joint moment generating function of $\begin{pmatrix} W(t) \\ B(t) \end{pmatrix}$. Mahmoud (2008) discovered that this moment generating function satifies the partial differential equation

$$\begin{aligned} \frac{\partial \phi_n(t,u,v)}{\partial t} + \Big(1 - \eta_{U,V}(u,v)\Big)\frac{\partial \phi_n(t,u,v)}{\partial u} \\ + \Big(1 - \eta_{X,Y}(u,v)\Big)\frac{\partial \phi_n(t,u,v)}{\partial v} = 0, \end{aligned} \tag{2.3}$$

where $\eta_{U,V}(u, v)$ is the joint moment generating function of U and V, and $\eta_{X,Y}(u, v)$ is the joint moment generating function of X and Y. As this functional equation is quite new, it has only been solved in a very limited number of cases. Some notable solutions include the case of forward and backward diagonal processes (Balaji and Mahmoud (2006)) and the Ehrenfest process (Balaji et al. (2006)). However, we would like to

note that the partial differential equation is valid for general tenable urns and not just those which are zero-balanced. Earlier forms of this partial differential equation appear in Balaji and Mahmoud (2006), where the functional equation is derived for a deterministic replacement matrix.

In a tenable zero-balanced urn on a total of n white and blue balls, the number of balls inside the urn never changes. Thus, we gain a valuable condition on this class of urn models: the property that

$$W(t) + B(t) = n \tag{2.4}$$

for all times $t \geq 0$. For this class of urns, the partial differential equation in (2.3) reduces to a simpler form, one in terms of either only $W(t)$ or $B(t)$. Since we are given the initial quantities of the urn, we can in fact determine the value of n and find either $W(t)$ or $B(t)$ in terms of the opposite color. In view of (2.4), it suffices to find one marginal distribution at time t to completely determine the urn status at time t. So, we shall choose to only study the distribution of $W(t)$. Let

$$\psi_n(t, u) = \phi_n(t, u, 0) = \mathbf{E}[e^{uW(t)}] \tag{2.5}$$

be the moment generating function of $W(t)$. Since the moment generating function is continuous with respect to v at 0 and with respect to u at 0, we can now use Leibniz integral rule and interchange the integral and derivative operators to manipulate our equation in (2.3) to find a partial differential equation for $\psi_n(t, u)$. Observe the following partial derivatives:

$$\begin{aligned}
\frac{\partial}{\partial t}\phi_n(t, u, v)\Big|_{v=0} &= \frac{\partial}{\partial t}\psi_n(t, u); \\
\frac{\partial}{\partial u}\phi_n(t, u, v)\Big|_{v=0} &= \frac{\partial}{\partial u}\psi_n(t, u); \\
\frac{\partial}{\partial v}\phi_n(t, u, v)\Big|_{v=0} &= \mathbf{E}[B(t)e^{uW(t)}]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbf{E}\Big[\Big(n - W(t)\Big)e^{uW(t)}\Big] \\
&= n\psi_n(t,u) - \frac{\partial}{\partial u}\psi_n(t,u).
\end{aligned}
$$

A two-color tenable zero-balanced Pólya process has the replacement matrix (1.6). Once we set $v = 0$, the functional equation (2.3) transforms into

$$
\frac{\partial \psi_n(t,u)}{\partial t} + (1 - \mathbf{E}[e^{-uc\,\mathrm{Ber}(p)}])\frac{\partial \psi_n(t,u)}{\partial u} + (1 - \mathbf{E}[e^{uc\,\mathrm{Ber}(r)}])\Big(n\psi_n(t,u) - \frac{\partial}{\partial u}\psi_n(t,u)\Big) = 0.
$$

After rearranging the above expression in terms of our partial differential equation, this becomes

$$
\frac{\partial \psi_n(t,u)}{\partial t} + (p - r + re^{cu} - pe^{-cu})\frac{\partial \psi_n(t,u)}{\partial u} + rn(1 - e^{cu})\psi_n(t,u) = 0.
$$

Through this rearrangement, we may solve the partial differential equation. With the assistance of Maple, we discover the solution to be the following equation:

$$
\begin{aligned}
\psi_n(t,u) &= \frac{e^{-nt(p+r)}}{(p+r)^{\frac{n}{c}}}\Big(\frac{re^{ct(p+r)+cu} + pe^{ct(p+r)} - p + pe^{cu}}{re^{ct(p+r)+cu} + pe^{ct(p+r)} + r - re^{cu}}\Big)^{\frac{W(0)}{c}} \\
&\quad \times \Big(re^{ct(p+r)+cu} + pe^{ct(p+r)} + r - re^{cu}\Big)^{\frac{n}{c}}. \qquad (2.6)
\end{aligned}
$$

While it is proper to solve the equation as is, it may be beneficial to break the equation up into separate pieces in order to evaluate each section asymptotically. In order to facilitate subsequent asymptotic analysis, we shall take the above solution and re-write this as

$$
\begin{aligned}
\psi_n(t,u) &= \frac{e^{-nt(p+r)}}{(p+r)^{\frac{n}{c}}}(re^{ct(p+r)+cu} + pe^{ct(p+r)})^{\frac{n}{c}} A_1 A_2 \\
&= \Big(\frac{re^{cu} + p}{p+r}\Big)^{\frac{n}{c}} A_1 A_2, \qquad (2.7)
\end{aligned}
$$

where

$$
A_1 := A_1(n,t,u) = \Big(1 + \frac{(p+r)(e^{cu} - 1)}{re^{ct(p+r)+cu} + pe^{ct(p+r)} + r - re^{cu}}\Big)^{\frac{W(0)}{c}}, \qquad (2.8)
$$

and

$$A_2 := A_2(n,t,u) \;=\; \Big(1 + \frac{r - re^{cu}}{re^{ct(p+r)+cu} + pe^{ct(p+r)}}\Big)^{\frac{n}{c}}. \tag{2.9}$$

2.2 Scope of main result and organization

Since our study of the Pólya process is governed by the replacement matrix (1.6), one factor which should be noted is the range in which we wish to evaluate the parameters p and r. In the case of $p = r = 0$, we have a trivial urn which never changes. So, we shall instead assume that $p + r > 0$, assuring that at least one of the two probabilities is positive. As previously mentioned, for tenable zero-balanced Pólya processes, $W_n(t)$ is not merely a ball count, but also a Markov chain over time t. When we fix n and allow t to approach infinity, it is straightforward to demonstrate that $W_n(t)$ possesses a binomial stationary distribution $\mathrm{Bin}(\frac{n}{c}, \frac{r}{p+r})$, which enumerates the number of successes in $\frac{n}{c}$ independent trials with rate of success $\frac{r}{p+r}$.

We will focus our attention on the cases in which we deal with very large values of n. These situations arise naturally in many applications, including those discussed in Chapter 1. For example, according to Avogadro's number, a mole of gas contains about 6.022×10^{23} particles. So, when we consider real applications involving gas diffusion between two chambers, we are dealing with billions of gas particles. The assumption that n grows to ∞ gives a realistic approximation of the situation, as we can determine a suitable estimate of the distribution of the number of balls in the urn. For large n, the asymptotic behavior of $\psi_n(t,u)$ depends on how t relates to n. From this point on, we will let $t = t_n$ be a function of n. This is because when t grows at different rates, unique asymptotic distributions formulate under what we consider as "phases". Of the numerous patterns available, there exist three signficant growth

phases in which we wish to study:

(i) The subcritical phase, when $t_n \to \infty$, but $t_n = o(\ln n)$, or when $t_n \sim K \ln n$, for some $K < \frac{1}{2c(p+r)}$;

(ii) The critical phase, when $t_n \sim \frac{1}{2c(p+r)} \ln(b_n n)$, for some positive b_n that is bounded away from 0 and infinity, i.e. $M_1 \leq b_n \leq M_2$, for some positive M_1 and M_2.

(iii) The supercritical phase, when $t_n \sim K \ln n$, for some $K > \frac{1}{2c(p+r)}$, or when $\ln n = o(t_n)$. One will notice that this phase has the form of the critical phase, except in this phase $b_n \to \infty$.

Given our value of n, we may determine how time moves forward based on one of these three phases. One may note that these phases depend strongly on the rate of logarithmic growth of t_n in relation to n. Originally, the research advocated for the phases to be named based on their logarithmic growth pattern. However, as seen above, it was discovered that the boundaries between the "logarithmic" and "sublogarithmic, as well as the "logarithmic" and "superlogarithmic", growth phases are blurred depending on both the presence of the factor K and whether the series b_n is bounded within the finite real numbers. Thus, it was decided to adjust the name to reflect the asymptotic boundaries and to denote the middle phase as our "critical" phase and define the subcritical and supercritical phases accordingly.

We should note, however, that these are not the only time growths that can be studied. One can study, for example, the evolution across various values of n of the Pólya process for the sequence $b_n = \{1, \frac{1}{2}, 2, \frac{1}{3}, 3, \frac{1}{4}, 4, ...\}$. In such a case, we can note that $\liminf_{n\to\infty} b_n = 0$ and $\limsup_{n\to\infty} b_n = \infty$, and this sequence does not fall in any of the three phases above. However, certain subsequences can be approached using a

combination of these three phases. In the example above, we can split the sequence into $\{b_{2n}\}_{n=1}^{\infty}$ and $\{b_{2n-1}\}_{n=1}^{\infty}$ and identify each sequence as one of the above three possible phases.

In Section 3.1, we will solve for both the exact and asymptotic mean and variance of $W_n(t)$ under each of the three phases above. These central moments hold great importance to the asymptotic distribution of $W_n(t)$ as they guide us to capture the structure of these phases. We will now provide the main result of this thesis with the following theorem.

Theorem 2.2.1. *Suppose we have a tenable, zero-balanced Pólya process on n white and blue balls in total, governed by (1.6) and starting with $W(0) = \alpha_n n + o(n)$ white balls, where $0 \leq \alpha_n \leq 1$. Let $W(t)$ be the number of white balls which reside in our urn at time t_n. Then,*

(a) In the subcritical phase, when $t_n = o(\ln n)$, or $t_n \sim K \ln n$, for $K < \frac{1}{2c(p+r)}$, we have

$$\frac{W(t_n) - \dfrac{r}{p+r}n}{ne^{-c(p+r)t_n}} - \alpha_n \xrightarrow{\mathcal{P}} -\frac{r}{p+r}.$$

(b) Let b_n be a positive function of n with $\liminf_{n\to\infty} b_n > 0$. In the critical ($b_n$ bounded from above) and supercritical ($b_n \to \infty$) phases, when $t_n = \frac{1}{2c(p+r)} \ln(b_n n)$, for $r > 0$ we have

$$\frac{W(t_n) - \dfrac{r}{p+r}\, n - \Big(\alpha_n - \dfrac{r}{p+r}\Big)\sqrt{\dfrac{n}{b_n}}}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{prc}{(p+r)^2}\Big),$$

and if $r = 0$ we have

$$\frac{W(t_n) - \alpha_n\sqrt{\dfrac{n}{b_n}}}{\sqrt{n}} \xrightarrow{\mathcal{P}} 0.$$

The proof of this result appears in Section **??**. We shall use equation (2.7) to asymptotically configure each of the three parts of the partial differential equation and use this to find the approximate proportion of white balls within the urn. As it is often useful to see the application of the theorem in everyday life, in Chapter 4 we shall provide some examples with remarks and interpretations in order to create a valid picture for how this theorem works.

Chapter 3: Proof of Theorem

The purpose of this chapter is to prove our claim and to explore *why* exactly our urn distribution varies, depending on the time phrase in which we exist. We begin with the calculation of our exact moments as derived from the moment generating function found in Chapter 2. Those calculations will allow us to construct an asymptotic solution to the approximate limiting distribution for $W_n(t)$. Afterward, we approach the moment generating function to provide a proof to Theorem 2.2.1. Finally, we examine the asymptotics of functions $A_1(n,t,u)$ and $A_2(n,t,u)$ and combine these functions to asymptotically analyze our expression (2.7) as n approaches ∞ under our different phases and complete our proof.

3.1 Exact moments

The intention of this section is to derive our central moments from the moment generating function given by equation (2.6). We wish to obtain these expressions in order to derive a central limit theorem and obtain our approximate distribution.

Recall the exact expression from (**??**):

$$\begin{aligned}\psi_n(t,u) &= \frac{e^{-nt(p+r)}}{(p+r)^{\frac{n}{c}}}\left(\frac{re^{ct(p+r)+cu}+pe^{ct(p+r)}-p+pe^{cu}}{re^{ct(p+r)+cu}+pe^{ct(p+r)}+r-re^{cu}}\right)^{\frac{W(0)}{c}} \\ &\quad \times\left(re^{ct(p+r)+cu}+pe^{ct(p+r)}+r-re^{cu}\right)^{\frac{n}{c}}. \qquad (3.1)\end{aligned}$$

We can use the equation above to find the moments of the generating function of $W_n(t)$ in exact form. The k^{th} moment is obtained by taking the k^{th} derivative with respect to u and setting u to 0. Taking the first partial derivative with respect to u, we get the following equation:

$$\frac{\partial}{\partial u}\psi_n(t,u) = \left(\frac{1}{c}\right)e^{-n(p+r)t}n(p+r)^{-\frac{n}{c}}\left(\frac{-p+e^{c(p+r)t}p+e^{cu}p+e^{c(p+r)t+cu}r}{e^{c(p+r)t}p+r-e^{cu}r+e^{c(p+r)t+cu}r}\right)^{\frac{W_n(0)}{c}}$$

$$
\begin{aligned}
&\times (e^{c(p+r)t}p + r - e^{cu}r + e^{c(p+r)t+cu}r)^{-1+\frac{n}{c}}(-ce^{cu}r + ce^{c(p+r)t+cu}r) \\
&+\Big(\frac{1}{c}\Big)e^{-n(p+r)t}(p+r)^{-\frac{n}{c}}\Big(\frac{-p + e^{c(p+r)t}p + e^{cu}p + e^{c(p+r)t+cu}r}{e^{c(p+r)t}p + r - e^{cu}r + e^{c(p+r)t+cu}r}\Big)^{-1+\frac{W_n(0)}{c}} \\
&\times (e^{c(p+r)t}p + r - e^{cu}r + e^{c(p+r)t+cu}r)^{\frac{n}{c}} \\
&\times W_n(0)\Big(\frac{ce^{cu}p + ce^{c(p+r)t+cu}r}{e^{c(p+r)t}p + r - e^{cu}r + e^{c(p+r)t+cu}r} \\
&\quad - \frac{(-p + e^{c(p+r)t}p + e^{cu}p + e^{c(p+r)t+cu}r)(-ce^{cu}r + ce^{c(p+r)t+cu}r)}{(e^{c(p+r)t}p + r - e^{cu}r + e^{c(p+r)t+cu}r)^2}\Big).
\end{aligned}
$$

By setting $u = 0$, we are able to reduce our moment generating function to the following expression:

$$
\begin{aligned}
\frac{\partial}{\partial u}\psi_n(t,0) &= \Big(\frac{1}{c}\Big)e^{-n(p+r)t}n(p+r)^{-\frac{n}{c}}\Big(\frac{e^{c(p+r)t}(p+r)}{e^{c(p+r)t}(p+r)}\Big)^{\frac{W_n(0)}{c}} \\
&\quad \times (e^{c(p+r)t}(p+r))^{-1+\frac{n}{c}}(e^{-ct(p+r)})(cr)(1-e^{-ct(p+r)}) \\
&\quad +\Big(\frac{1}{c}\Big)e^{-n(p+r)t}(p+r)^{-\frac{n}{c}}\Big(\frac{e^{c(p+r)t}(p+r)}{e^{c(p+r)t}(p+r)}\Big)^{-1+\frac{W_n(0)}{c}}(e^{c(p+r)t}(p+r))^{\frac{n}{c}} \\
&\quad \times (c)W_n(0)\Big(\frac{p + e^{c(p+r)t}r}{e^{c(p+r)t}(p+r)} - \frac{(r)e^{c(p+r)t}(p+r)(e^{c(p+r)t}-1)}{(e^{c(p+r)t}(p+r))^2}\Big) \\
&= \frac{nr}{p+r}\Big(1 - e^{-ct(p+r)}\Big) + \frac{W_n(0)}{e^{ct(p+r)}}.
\end{aligned}
$$

Now that we have reduced our moment generating function, we can simplify the expression and provide the expected value below:

$$
\mathbf{E}[W_n(t)] = \frac{r}{p+r}\,n + \Big(W_n(0) - \frac{rn}{r+p}\Big)e^{-ct(p+r)},
$$

which is valid for all $t \geq 0$.

One can note that asymptotically, as $n \to \infty$, the expected value function experiences multiple phase transitions. Recalling that $W_n(0) = \alpha_n n + o(n)$, we may view the expected value through three separate phases: the subcritical, the critical, and the supercritical phases.

At the subcritical phase, when $t_n = o\Big(\ln(n)\Big)$, or $t_n \sim K \ln n$ for some positive

$K < \frac{1}{(2c(r+p))}$, we may simplify our expression as:

$$\begin{aligned}
\mathbf{E}[W(t_n)] &= \frac{r}{p+r}\, n + \Big((\alpha_n n + o(n)) - \frac{rn}{r+p}\Big) e^{-ct_n(p+r)} \\
&= \frac{r}{p+r}\, n + \Big(\alpha_n n - \frac{rn}{r+p}\Big) e^{-ct_n(p+r)} + o(n) e^{-ct_n(p+r)} \\
&= \Big(\frac{r}{p+r} + \Big(\alpha_n - \frac{r}{p+r}\Big)\, e^{-c(p+r)t_n}\Big)\, n + o(n e^{-c(p+r)t_n}).
\end{aligned}$$

At the critical phase, when $t_n \sim \frac{1}{2c(p+r)} \ln(b_n n)$, for some positive b_n that is bounded away from 0 and infinity, i.e. $M_1 \leq b_n \leq M_2$, for some positive M_1 and M_2, we may reduce our expression to be:

$$\begin{aligned}
\mathbf{E}[W(t_n)] &= \frac{r}{p+r}\, n + \Big((\alpha_n n + o(n)) - \frac{rn}{r+p}\Big) e^{-c(p+r)(\frac{1}{2c(p+r)}) ln(b_n n)} \\
&= \frac{r}{p+r}\, n + \Big((\alpha_n - \frac{r}{r+p}) n + o(n)\Big) \sqrt{\frac{1}{b_n n}} \\
&= \frac{r}{p+r}\, n + \Big(\alpha_n - \frac{r}{p+r}\Big) \sqrt{\frac{n}{b_n}} + o\Big(\sqrt{\frac{n}{b_n}}\,\Big).
\end{aligned}$$

If we let $b_n \to \infty$, we enter the supercritical phase, either when $t_n \sim K \ln(n)$ for some $K > \frac{1}{2c(p+r)}$ or when $\ln(n) = o(t_n)$, and obtain the value:

$$\mathbf{E}[W(t_n)] = \frac{r}{p+r}\, n + o(\sqrt{n}\,).$$

To summarize, we can determine the expected value of $W_n(t)$ based upon which of our three possible growth phases our urn experiences:

$$\mathbf{E}[W(t_n)] = \begin{cases} \Big(\frac{r}{p+r} + \Big(\alpha_n - \frac{r}{p+r}\Big)\, e^{-c(p+r)t_n}\Big)\, n + o(n e^{-c(p+r)t_n}), \\ \qquad\qquad \text{in the subcritical phase;} \\ \frac{r}{p+r}\, n + \Big(\alpha_n - \frac{r}{p+r}\Big) \sqrt{\frac{n}{b_n}} + o\Big(\sqrt{\frac{n}{b_n}}\,\Big), \\ \qquad\qquad \text{in the critical phase;} \\ \frac{r}{p+r}\, n + o(\sqrt{n}\,), \\ \qquad\qquad \text{in the supercritical phase.} \end{cases}$$

In similar fashion, we are able to now take the second derivative of (**??**) with respect to u and set u to 0. This operation provides the second moment for the

moment generating function for $W_n(t)$, and from this we obtain the second central moment, our variance, as follows:

$$\begin{aligned}
\mathbf{Var}[W_n(t)] &= -e^{-2(c+n)(p+r)t}(p+r)^{-\frac{2(c+n)}{c}}\Big(e^{c(p+r)t}(p+r)\Big)^{\frac{2n}{c}} \\
&\quad \times \Big((-1+e^{c(p+r)t})nr + (p+r)W_n(0)\Big)^2 \\
&\quad + e^{-2(c-n)(p+r)t}(p+r)^{\frac{2c-n}{c}}\Big(e^{c(p+r)t}(p+r)\Big)^{\frac{-4c+n}{c}} \\
&\quad \times \Big(\Big((-1+e^{c(p+r)t})nr + (p+r)W_n(0)\Big)^2 \\
&\quad + c(e^{c(p+r)t}-1)(e^{c(p+r)t}npr + nr^2 + (p^2-r^2)W_n(0))\Big) \\
&= e^{-2(c-n)(p+r)t}(p+r)^{\frac{2c-n}{c}}\Big(e^{c(p+r)t}(p+r)\Big)^{\frac{-4c+n}{c}} \\
&\quad \times c(e^{c(p+r)t}-1)(e^{c(p+r)t}npr + nr^2 + (p^2-r^2)W_n(0))\Big) \\
&= \frac{c}{(p+r)^2}\Big(prn + \Big((p^2-r^2)W_n(0) + rn(r-p)\Big)e^{-ct(p+r)} \\
&\qquad - \Big((p^2-r^2)W_n(0) + r^2n\Big)e^{-2ct(p+r)}\Big),
\end{aligned}$$

which is valid for all t ≥ 0.

Like the mean, as $n \to \infty$, the variance asymptotically undergoes separate phase transitions. For the critical and supercritical phases, we can determine the asymptotic variance to be:

$$\begin{aligned}
\mathbf{Var}[W(t_n)] &= \frac{c}{(p+r)^2}\Big(prn + \Big((p^2-r^2)(\alpha_n n + o(n)) + rn(r-p)\Big)e^{(\frac{-c(p+r)}{2c(p+r)})ln(b_n n)} \\
&\quad -\Big((p^2-r^2)(\alpha_n n + o(n)) + r^2 n\Big)e^{(\frac{-2c(p+r)}{2c(p+r)})ln(b_n n)}\Big) \\
&= \frac{c}{(p+r)^2}\Big(n(pr) + \frac{n(p^2\alpha_n + r^2 - pr - r^2\alpha_n n) + o(n)}{\sqrt{b_n n}} \\
&\qquad + \frac{n(r^2\alpha_n - p^2\alpha_n - r^2) + o(n)}{b_n n}\Big) \\
&= \frac{c}{(p+r)^2}\Big(n(pr) + O(\sqrt{n}\,) + o(\sqrt{n}\,)\Big) \\
&= \frac{cpr}{(p+r)^2}n + O(\sqrt{n}\,).
\end{aligned}$$

Looking at the subcritical phase, we reveal that our variance follows as:

$$\begin{aligned}
\mathbf{Var}[W(t_n)] &= \frac{c}{(p+r)^2}\Big(n(pr) + \frac{n(p^2\alpha_n + r^2 - pr - r^2\alpha_n n) + o(n)}{e^{c(p+r)t_n}} \\
&\qquad + \frac{n(r^2\alpha_n - p^2\alpha_n - r^2) + o(n)}{e^{2c(p+r)t_n}}\Big) \\
&= \frac{c}{(p+r)^2}\Big(n(pr) + \frac{n(p^2\alpha_n + r^2 - pr - r^2\alpha_n n)}{e^{c(p+r)t_n}} \\
&\qquad + \frac{n(r^2\alpha_n - p^2\alpha_n - r^2)}{e^{2c(p+r)t_n}} + o(ne^{-c(p+r)t_n})\Big) \\
&= \frac{cn}{(p+r)^2}\Big(pr + \frac{(p^2\alpha_n + r^2 - pr - r^2\alpha_n n)}{e^{c(p+r)t_n}} \\
&\qquad + \frac{(r^2\alpha_n - p^2\alpha_n - r^2)}{e^{2c(p+r)t_n}}\Big) + o(ne^{-c(p+r)t_n}) \\
&= \frac{cn}{(p+r)^2}\Big(pr + o(e^{-c(p+r)t_n})\Big) + o(ne^{-c(p+r)t_n}) \\
&= \frac{cpr}{(p+r)^2}n + o(ne^{-c(p+r)t_n}).
\end{aligned}$$

In summary, we have determined the variance for $W_n(t)$ for each of our three phases:

$$\mathbf{Var}[W(t_n)] = \begin{cases} \dfrac{cpr}{(p+r)^2}\, n + o(n\, e^{-c(p+r)t_n}), & \text{in the subcritical phase;} \\ \dfrac{cpr}{(p+r)^2}\, n + O(\sqrt{n}\,), & \text{in the critical and} \\ & \text{supercritical phases.} \end{cases}$$

With these central moments alone, we can approximate the distibution of $W_n(t)$ based upon which growth phase it resides. It becomes useful to take our original equations for our expected value and variance and construct their asymptotic forms. With each phase, we can illustrate what will happen once we rescale the variable $W_n(t)$ by the factors described in Theorem 2.2.1. However, just knowing the central moments is not enough for our academic tastes, and we wish to venture further! Our next section will provide a formal proof for the distribution of $W_n(t)$ for our three phases.

3.2 Proof of the main result

While the previous section gave us the exact moments of the moment generating function for $W_n(t)$, we wish to show that the distribution for $W_n(t)$ does in fact converge to either a normal distribution or a degenerate constant. Hence, we turn to mathematical proof in this section − using the concept of asymptotics, probability, and mathematical analysis, starting with the subcritical phase in Subsection **??**. In Subsection **??**, the proofs for the critical and supercritical phases are combined to simplify our result and prove that the difference in our variances is inconsequential between these two phases.

3.2.1 The subcritical phase

In the subcritical phase we have $t_n = o(\ln(n))$, or $t_n \sim K\ln(n)$, for some positive $K < \frac{1}{2c(r+p)}$. For the sake of simplicity, let us set $s_n = ne^{-c(p+r)t_n}$. In the solution (**??**), we chose to reduce the moment generating function to $\psi_n(t,u) = \Big[\Big(\frac{re^{cu}+p}{p+r}\Big)^{\frac{n}{c}} A_1 A_2\Big]$ in order to break down our asymptotic analysis and easily complete our task.

Let us begin with part A_1 (cf. (**??**)). Asymptotically, we have the following simplification for $A_1(n,t,\frac{u}{s_n})$:

$$\begin{aligned}
A_1\Big(n,t,\frac{u}{s_n}\Big) &= \Big(1+\frac{(p+r)(e^{\frac{cu}{s_n}}-1)}{e^{c(p+r)t_n}(re^{\frac{cu}{s_n}}+p)-r(e^{\frac{cu}{s_n}}-1)}\Big)^{\frac{W_0(n)}{c}} \\
&= \Big(1+\frac{(p+r)(\frac{cu}{s_n}+O(\frac{1}{s_n^2}))}{e^{c(p+r)t_n}[p+r+O(\frac{1}{s_n})]+O(\frac{1}{s_n})}\Big)^{\frac{\alpha_n}{c}n+o(n)} \\
&\sim \Big(1+\frac{(p+r+O(\frac{1}{s_n}))(\frac{cu}{s_n}+O(\frac{1}{s_n^2}))e^{-c(p+r)t_n}}{(p+r+O(\frac{1}{s_n}))(1+O(1))}\Big)^{\frac{\alpha_n}{c}n+o(n)} \\
&\sim \Big(1+\Big(\frac{cu}{s_n}+O(\frac{1}{s_n^2})\Big)e^{-c(p+r)t_n}\Big)^{\frac{n}{c}(\alpha_n+o(1))} \\
&\sim \Big(1+\Big(\frac{1}{n/c}\Big)\Big[u+O\Big(\frac{1}{s_n}\Big)\Big]\Big)^{\frac{n}{c}(\alpha_n+o(1))} \\
&\sim e^{\big((\alpha_n+o(1))\big)\big(u+O(1/s_n)\big)}
\end{aligned}$$

$$\sim \quad e^{\alpha_n u}.$$

Likewise, we can perform similar methods for the function $A_2(n,t,\frac{u}{s_n})$ (cf. (**??**)) to obtain the following asymptotic relation below:

$$\begin{aligned}
A_2\Big(n,t,\frac{u}{s_n}\Big) &= \Big(1+\frac{r-re^{\frac{cu}{s_n}}}{e^{c(p+r)t_n}\big(re^{\frac{cu}{s_n}}+p\big)}\Big)^{\frac{n}{c}} \\
&= \Big(1-\frac{\frac{cu}{s_n}r+O(\frac{1}{s_n^2})}{e^{c(p+r)t_n}[p+r+O(\frac{1}{s_n})]}\Big)^{\frac{n}{c}} \\
&\sim \Big(1-\frac{\frac{cur}{n}+O(\frac{1}{n\,s_n})}{p+r+O(\frac{1}{s_n})}\Big)^{\frac{n}{c}} \\
&\sim \Big(1+\Big(\frac{1}{n/c}\Big)\frac{-ur-O(\frac{1}{s_n})}{p+r+O(\frac{1}{s_n})}\Big)^{\frac{n}{c}} \\
&\sim \Big(1+\Big(\frac{1}{n/c}\Big)\Big(\frac{-ur}{p+r}\Big)\Big)^{\frac{n}{c}} \\
&\sim e^{-\frac{r}{p+r}u}.
\end{aligned}$$

Therefore, as $n\to\infty$, we can replace our asymptotic simplifcations of A_1 and A_2 in our moment generating function $\psi_n(t_n,\frac{u}{s_n})$ to derive the following:

$$\begin{aligned}
\psi_n\Big(t_n,\tfrac{u}{s_n}\Big) &\sim \Big(\frac{p}{p+r}+\frac{r}{p+r}\Big[1+\frac{cu}{s_n}+O\Big(\frac{1}{s_n^2}\Big)\Big]\Big)^{\frac{n}{c}}e^{\big(\alpha_n-\frac{r}{p+r}\big)u} \\
&\sim \Big(\frac{p+r}{p+r}+\frac{rc}{p+r}u+O\Big(\frac{1}{s_n^2}\Big)\Big)^{\frac{n}{c}}e^{\big(\alpha_n-\frac{r}{p+r}\big)u} \\
&\sim \Big(1+\Big[\frac{rn}{\frac{n}{c}(p+r)}u+O\Big(\frac{1}{s_n^2}\Big)\Big]\Big)^{\frac{n}{c}}e^{\big(\alpha_n-\frac{r}{p+r}\big)u} \\
&\sim e^{\frac{rn}{(p+r)s_n}u}e^{\big(\alpha_n-\frac{r}{p+r}\big)u}.
\end{aligned}$$

This derivation seals the deal as we can now rescale our random variable $W(t_n)$ in order to provide our adjusted moment generating function, as seen below:

$$\begin{aligned}
\mathbf{E}\Big[e^{\big(W(t_n)-\frac{r}{p+r}n\big)\frac{u}{s_n}}\Big] &= e^{\big(-\frac{r}{p+r}n\big)\frac{u}{s_n}}\times\mathbf{E}\Big[e^{W(t_n)\frac{u}{s_n}}\Big] \\
&= e^{\big(-\frac{r}{p+r}n\big)\frac{u}{s_n}}\times\psi_n\Big(t_n,\frac{u}{s_n}\Big) \\
&\sim e^{\big(-\frac{r}{p+r}n\big)\frac{u}{s_n}}\times\Big[e^{\frac{rn}{(p+r)s_n}u}e^{\big(\alpha_n-\frac{r}{p+r}\big)u}\Big] \\
&= e^{\big(\alpha_n-\frac{r}{p+r}\big)u}.
\end{aligned}$$

We recognize this as the moment generating function for a degenerate random variable. Since degenerate random variables have zero variance, we can raise our level of convergence of $W(t_n)$ from convergence in distribution to convergence in probability by applying Chebychev's inequality as a bound:

$$\mathbf{Prob}(|\frac{W(t_n)-\frac{r}{p+r}n}{s_n} - \alpha_n + \frac{r}{p+r}| > \epsilon) \leq \frac{\mathbf{Var}(\frac{W(t_n)-\frac{r}{p+r}n}{s_n})}{\epsilon^2}.$$

Since $\mathbf{Var}(\frac{W(t_n)-\frac{r}{p+r}n}{s_n}) \to 0$ as $n \to \infty$, then for all $\epsilon > 0$,

$$\lim_{n\to\infty} \mathbf{Prob}(|\frac{W(t_n)-\frac{r}{p+r}n}{s_n} - \alpha_n + \frac{r}{p+r}| > \epsilon) \leq \lim_{n\to\infty} \frac{\mathbf{Var}(\frac{W(t_n)-\frac{r}{p+r}n}{s_n})}{\epsilon^2} = 0.$$

Therefore, it follows that

$$\frac{W(t_n) - \frac{r}{p+r}n}{s_n} - \alpha_n \xrightarrow{\mathcal{P}} -\frac{r}{p+r}. \quad \square$$

3.2.2 The critical phase and beyond

In the critical phase, we are given that time grows at a rate such that $t_n = \frac{1}{2c(p+r)} \ln(b_n n)$, for some positive b_n that is bounded away from 0 and ∞; in other words, there are two positive numbers M_1 and M_2, such that $M_1 \leq b_n \leq M_2$, for all $n \geq 1$. In the supercritical phase, we assume that either the rate at which time passes will follow the asymptotic relation $t_n \sim K \ln(n)$ for $K > \frac{1}{2c(p+r)}$, or t_n is superlogarithmic altogether with $\ln(n) = o(t_n)$.

For both of these supercritical forms, we can instead write t_n as the expression $\frac{1}{2c(p+r)} \ln(b_n n)$, with b_n converging to infinity. In the first supercritical form, we can conclude that the sequence $\{b_n\}_{n=1}^{\infty}$ is defined such that $O(b_n) = n^{\delta}$ for some $\delta > 0$. In the second supercritical form, we have $O(b_n) = \exp\Big(n^{\frac{2c(p+r)-1}{2c(p+r)}}\Big)$. Thus, as $n \to \infty$, $b_n \to \infty$ as denoted above.

In the solution (**??**), we can take the part A_1 (cf. (**??**)) and asymptotically reduce it to the following expression:

$$\begin{aligned}
A_1\Big(n,t,\frac{u}{\sqrt{n}}\Big) &= \Big(1+\frac{(p+r)(e^{\frac{cu}{\sqrt{n}}}-1)}{e^{\frac{1}{2}\ln(b_n n)}(re^{\frac{cu}{\sqrt{n}}}+p)-r(e^{\frac{cu}{\sqrt{n}}}-1)}\Big)^{\frac{W_0(n)}{c}} \\
&\sim \Big(1+\frac{(p+r)(\frac{cu}{\sqrt{n}}+O(\frac{1}{n}))}{e^{\frac{1}{2}\ln(b_n n)}(re^{\frac{cu}{\sqrt{n}}}+p)+O(\frac{1}{\sqrt{n}})}\Big)^{\frac{\alpha_n}{c}n+o(n)} \\
&\sim \Big(1+\frac{(p+r)(\frac{cu}{\sqrt{n}}+O(\frac{1}{n}))}{\sqrt{b_n n}((r+p)+O(\frac{1}{\sqrt{n}}))+O(\frac{1}{\sqrt{n}})}\Big)^{\frac{\alpha_n}{c}n+o(n)} \\
&\sim \Big(1+\frac{(p+r)(\frac{cu}{\sqrt{n}}+O(\frac{1}{n}))}{(r+p)(\sqrt{b_n n}+O(\sqrt{b_n}))}\Big)^{\frac{\alpha_n}{c}n+o(n)} \\
&\sim \Big(1+\frac{\frac{cu}{\sqrt{b_n}}}{n}(1+O(\frac{1}{n}))\Big)^{(\frac{\alpha_n}{c})n+o(n)} \\
&\sim e^{\frac{\alpha_n u}{\sqrt{b_n}}}.
\end{aligned}$$

In similar fashion, we can simplify the function $A_2(n,t,\frac{u}{\sqrt{n}})$ (cf. (**??**)) by following its asymptotic relation:

$$\begin{aligned}
A_2\Big(n,t,\frac{u}{\sqrt{n}}\Big) &= \Big(1+\frac{r-re^{\frac{cu}{\sqrt{n}}}}{e^{\frac{1}{2}\ln(b_n n)}(re^{\frac{cu}{\sqrt{n}}}+p)}\Big)^{\frac{n}{c}} \\
&\sim \Big(1+(-r)\frac{\frac{cu}{\sqrt{n}}+O(\frac{1}{n})}{(r+p)(\sqrt{b_n n}+O(\sqrt{b_n}))}\Big)^{\frac{n}{c}} \\
&\sim \Big(1+\frac{\frac{-rcu}{(p+r)\sqrt{b_n}}}{n}(1+O(\frac{1}{n}))\Big)^{\frac{n}{c}} \\
&\sim e^{-\frac{r}{(p+r)\sqrt{b_n}}u}.
\end{aligned}$$

Therefore, as $n\to\infty$, we have

$$\begin{aligned}
\psi_n\Big(t_n,\frac{u}{\sqrt{n}}\Big) &= \Big(\frac{re^{cu}+p}{p+r}\Big)^{\frac{n}{c}}A_1A_2 \\
&\sim \Big(\frac{p}{p+r}+\frac{r}{p+r}e^{\frac{cu}{\sqrt{n}}}\Big)^{\frac{n}{c}}e^{\frac{1}{\sqrt{b_n}}(\alpha_n-\frac{r}{p+r})u}.
\end{aligned}$$

On the right-hand side of the above asymptotic relation, the parenthesized expression raised to the power $\frac{n}{c}$ is the moment generating function of $c\times\mathrm{Bin}(\frac{n}{c},\frac{r}{p+r})$,

evaluated at at the point $\frac{u}{\sqrt{n}}$. According to the standard normal approximation of the binomial distribution, we instead can estimate the moment generating function as one for a normal distribution for large n, and we have the following relation:

$$\begin{aligned}\psi_n\Big(t_n, \frac{u}{\sqrt{n}}\Big)e^{-(\frac{r}{p+r})\sqrt{n}\,u}e^{-\frac{1}{\sqrt{b_n}}(\alpha_n-\frac{r}{p+r})u} &\sim e^{-\left(\frac{r}{p+r}\right)\sqrt{n}\,u}\Big(\frac{p}{p+r}+\frac{r}{p+r}e^{\frac{cu}{\sqrt{n}}}\Big)^{\frac{n}{c}} \\ &\to e^{-(\frac{r}{p+r})\sqrt{n}\,u}\,e^{(\frac{cu}{\sqrt{n}})(\frac{n}{c})(\frac{r}{p+r})+(\frac{p}{p+r})(\frac{r}{p+r})(\frac{n}{c})(\frac{c^2u^2}{n})} \\ &= e^{\frac{prcu^2}{2(p+r)^2}}.\end{aligned}$$

Our end result is the moment generating function of the normal random variable $\mathcal{N}(0, \frac{prc}{(p+r)^2})$ with mean 0 and variance $\frac{prc}{(p+r)^2}$. Therefore, in the critical and supercritical phases we have a convergence in distribution for our random variable $W(t_n)$:

$$\frac{W(t_n) - \dfrac{r}{p+r}\,n - \Big(\alpha_n - \dfrac{r}{p+r}\Big)\sqrt{\dfrac{n}{b_n}}}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{prc}{(p+r)^2}\Big).$$

Furthermore, we can note that if $r = 0$ (and $p > 0$) − as in the cases such as the coupon collection problem − our variance suddenly vanishes, thereby causing the limiting normal distribution to degenerate into a probability mass concentration at 0. Hence, we can upgrade the convergence level for the critical and supercritical cases as converging in probability to 0. □

One can observe that the starting condition α_n (the initial proportion of white balls) has a somewhat pronounced influence in the asymptotic distribution in the critical phase, but not in the supercritical phase. This is due to the fact that this influence is attenuated as we get deeper in that phase, and as $\mathrm{b}_n \to \infty$ we can witness that the α_n term eventually dies out.

3.3 Summary

In this chapter, we varified Theorem 2.2.1, first by finding the exact moments to the moment generating function for $W(t_n)$ and then by presenting an argument that uses asymptotic relations in order to find an asymptotic equivalent for the moment generating function $\psi_n(t_n, \frac{u}{\sqrt{n}})$. By affirming its validity, we are now able to quickly find approximate distributions for the number of white balls in any two-color urn model that follows the replacement matrix (1.6), conditional on the amount of time which has passed within the urn. In our next chapter, we shall apply Theorem 2.2.1 to some examples in order to provide a clearer picture of the theorem in use.

Chapter 4: Applications of the Process

This is where the fun begins. In this chapter, we will explore two examples of Pólya processes and how the distribution of balls within these urns either inevitably converges based on the allotted time passed or can be approximated depending on the phase in which it exists. We shall first explore a basic example under specific time scenarios and use it to uncover various distributions over the three major time phases discussed in the previous chapter. In our second example, we will conduct a simulation over a given urn model and verify our solution to Theorem 2.2.1, providing readers an opportunity to see theory put into practice.

4.1 Example 1: The Ehrenfest Urn

To put our result in perspective, let us first see how it applies to a specific Pólya urn. Consider the replacement matrix $\begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$ of the Ehrenfest urn. In this situation, we are given that $p = r = 1$ and $c = 1$, providing us a very simple urn to explore. Now, let's also suppose that $W_n(0) = \lfloor \frac{2}{3}n + \frac{\sqrt{n}}{12} + 0.77\ln(n) \rfloor \sim \frac{2}{3}n$. Thanks to our simple asymptotic reduction, we can conclude therefore that $\alpha_n = \frac{2}{3}$. After time $t_n = \ln(\ln(n))$, according to Theorem **??** (a), we are able to deduce that

$$\frac{W\Big(\ln(\ln(n))\Big) - \frac{1}{2}n}{\frac{n}{\ln(\ln(n))}} \xrightarrow{\mathcal{P}} \frac{1}{6}.$$

This is a law of sharp concentration. Thus, during this subcritical phase, we should expect that the proportion of white balls in our specified urn should be approximately $\frac{1}{6}$. However, after a longer period of time, t_n eventually reaches the critical phase. In this situation, more variability appears, and this increased rate in variability gives us a central limit theorem.

For instance, consider when time exceeds past the subcritical phase and we observe the moment at which the time is $t_n = \frac{1}{4}\ln(n) + 5 + 3\cos(n) = \frac{1}{4}\ln(ne^{20+12\cos(n)})$. We can see that t_n indeed falls inside the critical phase with $b_n = e^{20+12\cos(n)}$. According to Theorem **??** (b), we can thus conclude that

$$\frac{W\Big(\frac{1}{4}\ln(n) + 5 + 3\cos(n)\Big) - \frac{1}{2}\,n - \frac{1}{6}\sqrt{\frac{n}{e^{20+12\cos(n)}}}}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{1}{4}\Big).$$

At this point in time, we should expect that the number of white balls inside of our theoretical urn runs approximately a normal distribution.

As the Pólya process carries on for an even longer period, such as a superlogarithmic amount of time, it enters in the supercritical phase. For example, suppose $t_n = \frac{6}{11}n^2$. We can rewrite this amount of time in the form $\frac{1}{4}\ln(n \times (\frac{1}{n}e^{\frac{24}{11}n^2}))$, implying that $b_n = \frac{1}{n}e^{\frac{24}{11}n^2}$. Note that b_n grows to infinity as $n \to \infty$. According to Theorem **??** (b), we again obtain a central limit theorem result:

$$\frac{W\Big(\frac{6}{11}n^2\Big) - \frac{1}{2}\,n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \frac{1}{4}\Big).$$

In general, when b_n is a fast-growing function, it annihilates the role of α_n. This is as we expect, since after the gas mixes for a very long period, the influence of the initial split is obliterated and the mixing has become so thorough that our urn begins to balance out. Hence, the likelihood of drawing a white ball eventually appears to become a Bin(n, $\frac{1}{2}$) random variable, which explains why the function $W(t_n)$ converges in distribution to the normal random variable, as seen above.

4.2 Example 2: Simulation

We present the final portion of research for this thesis with a simulation of our Pólya process. Let us begin with a Pólya urn with the replacement matrix

$$\begin{pmatrix} -\text{Ber}(\frac{1}{3}) & \text{Ber}(\frac{1}{3}) \\ \text{Ber}(\frac{1}{2}) & -\text{Ber}(\frac{1}{2}) \end{pmatrix}.$$

Here, our scalar multiple is $c = 1$ and our parameters for our Bernoulli random variables are $p = \frac{1}{3}$ and $r = \frac{1}{2}$, respectvely, meaning that $\frac{r}{p+r} = \frac{\frac{1}{2}}{\frac{1}{3}+\frac{1}{2}} = \frac{3}{5}$. Now, suppose we have $n = 1000$ balls inside our urn, and let $W_n(0) = \lfloor 0.37n \rfloor \sim 0.37n$, implying that $\alpha_n = 0.37$ and $W_{1000}(0) = 370$. We will now set our time t to range from 0 to $2\ln(1000) \approx 13.8155$. In theory, when t ranges from 0 to some buffered boundary about 4.14465, we are experiencing the mixing within the subcritical phase. According to Section 3.1, we should find that the number of white balls within the urn should be approximate to $W(t) \approx 600 - 230e^{-5t/6}$. Table 4.1 displays the distribution of the urn with observation points signalling draws at roughly half-second intervals. We can compare our theoretical points to the observations found at time t, and in fact each observation suprisingly falls within 3% of our expected value.

When we reach our critical phase, we should expect the proportion of white balls within our urn to be approximate to $W(t) \approx 600 + 2.3\sqrt{\frac{10}{b_n}}$. Based on observations within the interval $t \in (3.5, 5.4)$, we can estimate that this is roughly our critical phase, namely since this subsequence of values remains favorably above the threshold value 600. It could be argued that during this period, the values of $W_{1000}(t)$ average to around 615, which infers that our term $b_n \approx 0.235$.

Finally, once we reach the point $t = 5.6$, the proportion of white balls in our urn dips back under 0.6, leading us to believe that at some point earlier we have reached the urn's supercritical phase. We expect at this point for $W_{1000}(t)$ to fluctuate near

Table 4.1: Distribution of Balls at Time t

Observation Point	t	$W_{1000}(t)$	$B_{1000}(t)$	Theoretical $W_{1000}(t)$
1	0.001366436275	370	630	370
2	0.5009701032	444	556	448
3	1.000639636	507	493	500
4	1.500867173	549	451	534
5	2.000894922	567	433	557
6	2.502141400	562	438	571
7	3.001007869	568	432	581
8	3.501003021	605	395	615
9	4.001148041	607	393	615
10	4.500351631	632	368	615
11	5.001000078	609	391	615
12	5.500083287	606	394	600
13	6.000371487	598	402	600
14	6.500812559	594	406	600
15	7.000553490	594	406	600
16	7.500221587	594	406	600
17	8.000996224	599	401	600
18	8.500781218	595	405	600
19	9.000585699	608	392	600
20	9.500125629	600	400	600
21	10.00127250	617	383	600
22	10.50049903	602	398	600
23	11.00011012	598	402	600
24	11.50207723	582	418	600
25	12.00021261	583	417	600
26	12.50238834	596	404	600
27	13.00144345	596	406	600
28	13.50173381	602	398	600
29	13.81458712	582	418	600

600, and indeed observations show that the number of white balls in this urn never reaches below 570 nor above 615 after $t = 5.6$, fencing its range of white balls within 5% of the theoretical mean. Thus, the number of white balls in our simulation appears to support the theoretical values proved in Chapter 3. Therefore, during all three phases our simulation appears to have closely matched the conclusions derived by Theorem 2.2.1.

4.3 Summary

We discussed two practical examples using urn-model techniques, and in both instances we were able to take the conclusions from Theorem 2.2.1 and apply it to real-life phenomena. In our first example, we took the Ehrenfest model and used it to observe $W_n(t)$ under a nontraditional starting condition − one that included the floor function, the square root operation, and even a natural logarithm! Despite the unique initial setup, we quickly approximated the distribution of the urn by examining t_n over different phases, providing convergence in probability for the subcritical phase and convergence in distribution for the critical and supercritical phases.

In our second example, we explored our Pólya urn via simulation, setting our scalar c to 1 and our Bernoulli parameters to $p = \frac{1}{3}$ and $r = \frac{1}{2}$. Via Table 4.1 we were able to demystify the theory and provide distinct values for $W_{1000}(t)$. We compared the theoretical values to our observations over each of our three phases and varified the theory up to a small degree of measurement error (3-5%). We were even able to roughly approximate the value b_n for the urn's critical phase. Hence, we can conclude that the results obtained by Theorem 2.2.1 do in fact read true when applying to real-world structures.

Chapter 5: Concluding Remarks

Over the course of this thesis, we have taken a simple urn with two different colors of balls and have turned it into a device by which we can explore numerous phenomena over continuous time. Through the Poissonization process, we were able to advance previous research in discrete time schemes and investigate the impact of different continuous-time phases on the distribution of the urn. We applied the partial differential equation found by Balaji and Mahmoud (2006) to the moment generating function for our two-color, zero-balanced Pólya process and used it to derive an exact solution $\psi_n(t, u)$. From inspection, we discovered three significant time phases − the subcritical, the critical, and the supercritical times phases − which modify the approximate central moments and the overall distributions depending on the phases at which the urn has entered. In our subcritical phase, we rescaled $W(t_n)$ to reveal that it converges in probability to a degenerate function. In our critical and supercritical phases, we again rescaled our random variable $W(t_n)$ to bring to surface a central limit theorem, implying convergence in distribution to a normal random variable.

In Chapter 4, we took Theorem 2.2.1 and applied it to real-life conditions to see how well the theory holds. In our first example, we derived values which adhere to our general knowledge of the Ehrenfest urn and we concluded that the theoretical distributions we hypothesized did fit with our natural concept of the urn. In our second example, we ran a simulation and tested how well its results aligned with the material in Chapters 2 and Chapter 3. Our results show to be promising, and our estimated critical phase even allowed us to roughly conjecture to the value b_n.

Recall our discussion in Chapter 2 of phases which do not fall into any of the three phases we have proven. While there are other ways in which one can examine these sequences, via subsequences or otherwise, the purpose of this thesis has decided to

focus upon the three phases found in Theorem 2.2.1. Further research within this field suggests numerous opportunities within these other types of growths, and it is with great hope that the reader will continue with this research and expand these discoveries to improve the knowledge within the field.

The purpose of this thesis was to not only reveal a new discovery concerning the field of urn models, but also to give the reader a fun alternative to examine the laws of probability. As discussed in Chapter 1, urns can be used to model a wide array of real-life situations by taking these events and structuring them into the form of different-colored balls within a container. The results in the previous chapters can be used to explain particle mixtures within containers, general heat dispersion within a given setting, and even viral infections, but there are numerous other ways for these constructs to be utilized. So, to the lucky readers of this thesis, we hope that this research has left you excited enough to take the above results and dive further into the area of urn models. However, we insist on one thing above all else: have a ball!

References

1 Arthur, W., Ermoliev, Y., and Kaniovski, Y. (1984). Strong Laws for a Class of Path-Dependent Stochastic Processes with Applications. *Proceedings of International Conference on Stochastic and Optimization*, Springer: Berlin, 287−300.

2 Athreya, K. and Karlin, S. (1968). Embedding of Urn Schemes into Continuous Time Markov Branching Process and Related Limit Theorems. *The Annals of Mathematical Statistics*, 39, 1801−1817.

3 Balaji, S. and Mahmoud, H. (2006). Exact and Limiting Distributions in Diagonal Pólya Processes. *Annals of the Institute of Statistical Mathematics*, 58, 171−185.

4 Balaji, S., Mahmoud, H. and Zhang, T. (2010). Phases in the Diffusion of Gases Via the Ehrenfest Urn Model. *Journal of Applied Probability*, 47, 841−855.

5 Benaim, M., Schreiber, S. and Tarres, P. (2004). Generalized Urn Models of Evolutionary Processes. *Annals of Applied Probability*, 14, 1455−1478.

6 Bender, E., Canfield, R., and McKay, B. (1997). The Asymptotic Number of Labeled Graphs with n Vertices, q Edges, and No Isolated Vertices. *Journal of Combinatorial Theory*, Series A 80, 124−150.

7 Bernoulli, J. (1713). *Arcs Conjectandi*. Reprinted in *Die Werke von Jakob Bernoulli*. (1975). Birkhäuser: Basel.

8 Durrett, R. (2010). *Probability Models for DNA Sequence Evolution*. Springer: Ithica, New York.

9 Eggenberger, F. and Pólya, G. (1923). Über Die Statistik Verketteter Vorgäge. *Zeitschrift für Angewandte Mathematik und Mechanik*, 3, 279−289.

10 Ehrenfest, P. and Ehrenfest, T. (1907). Über Zwei Bekannte Einwände Gegen das Boltzmannsche H-theorem. *Physikalische Zeitschrift*, 8, 311−314.

11 Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, Vol. I. Wiley: New York.

12 Fienberg, S. and Makov, U. (2001). Uniqueness, Urn Models and Disclosure Risk. *Research in Official Statistics*, 4, 23−40.

13 Friedman, B. (1949). A Simple Urn Model. *Communications of Pure and Applied Mathematics*, 2, 59−70.

14 Gastwirth, J. and Bhattacharya, P. (1984). Two Probability Models of Pyramid or Chain Letter Schemes Demonstrating that Their Promotional Claims are Unreliable. *Operations Research*, 32, 527−536.

15 Johnson, N. and Kotz, S. (1977). *Urn Models and Their Applications*: *An Approach to Modern Discrete Probability Theory*. Wiley: New York.

16 Kholfi, S. and Mahmoud, H. (2012a). The Class of Tenable Zero-Balanced Pólya Urns with an Initially Dominant Subset of Colors. *Statistics & Probability Letters*, 82, 49−57.

17 Kholfi, S. and Mahmoud, H. (2012a). The Class of Tenable Zero-Balanced Pólya Urn Schemes: Characterization and Gaussian Phases. *Advances in Applied Probability*, 44, 702−728.

18 Kolchin, V., Sevastianov, B., and Chistiakov, V. (1978). *Random Allocations*; translation. In: Balakrishnan, A.V. (Ed.), *Scripta Series in Mathematics*. V.H.Winston, Halsted Press: New York.

19 Kotz, S. and Balakrishnan, N. (1997). Advances in Urn Models During the Past Two Decades. In *Advances in Combinatorial Methods and Applications to Probability and Statistics*, 49, 203−257. Birkhäuser: Boston.

20 Mahmoud, H. (2008). *Pólya Urn Models*. Chapman-Hall: Boca Raton, Florida.

21 Mahmoud, H. (2010). Gaussian Phases in Generalized Coupon Collection. *Advances in Applied Probability*, 42, 994−1012.

22 Mahmoud, H. and Smythe, R. (1991). On the Distribution of Leaves in Rooted Subtrees of Recursive Trees. *The Annals of Applied Probability*, 1, 406−418.

23 Mahmoud, H. and Smythe, R. (2012). On the Joint Behavior of Types of Coupons in Generalized Coupon Collection. *Advances in Applied Probability*, 44:2, 429−451.

24 Martin, C. and Ho, Y. (2002). Value of Information in the Pólya Urn Process. *Information Science*, 147, 65−90.

25 Najock, D. and Heyde, C. (1982). On the Number of Terminal Vertices in Certain Random Trees with an Application to Stemma Construction in Philology. *Journal of Applied Probability*, 19, 675−680.

26 Navarro, D., Griffiths, T., Steyvers, M., and Lee, M. (2005). Modeling Individual Differences with Dirichlet Processes. *Proceedings of the* 27^{th} *Annual Conference of the Cognitive Science Society*, Stresa, Italy, 1594−1599.

27 Robbins, H. (1952). Some Aspects of the Sequential Design of Experiments. *Bulletin of the American Mathematical Society*, 58, 527−535.

28 Samuel, A. (1998). A Bayesian, Species-Sampling-Inspired Approach to the Uniques Problem in Microdata Disclosure Risk Assessment. *Journal of Official Statistics*, 14:4, 373−383.

29 Schrödinger, E. and Kohlrausch, F. (1926). Das Ehrenfestsche Modell der H-Kurve. *Physikalische Zeitschrift*, 27, 306−316.

30 Smythe, R. (2011). Generalized Coupon Collection: The Superlinear Case. *Journal of Applied Probability*, 48, 189−199.

31 Thompson, W. (1933). On the Likelihood that One Unknown Probability Exceeds Another in the View of the Evidence of the Two Samples. *Biometrika*, 25, 275−294.

32 Vatutin, V., and Mikhailov, V. (1982). Limit Theorems for the Number of Empty Cells in an Equiprobable Scheme for Group Allocation of Particles. *Theory of Probability and Its Applications*, 27, 684−692.

33 Windrum, P. (2004). Leveraging Technological Externalities in Complex Technologies: Microsofts Exploitation of Standards in the Browser Wars. *Research Policy*, 33, 385−394.

34 Xue, J. (2005). A Pólya Urn Model for Conformity. *Proceedings of the ESRC Game Theory Conference*, Essex, United Kingdom, 1−20.