On the “Heaviest” Polya Urn

Jeremy Chen

Department of Decision Sciences, National University of Singapore Business School
15 Kent Ridge Drive, Singapore 119245

Abstract

In the classical Polya urn problem, one begins with $d$ bins, each containing one ball. Additional balls arrive one at a time, and the probability that an arriving ball is placed in a given bin is proportional to $m^\gamma$, where $m$ is the number of balls in that bin. In this note, we consider the case of $\gamma = 1$, which corresponds to a process of “proportional preferential attachment” and is a critical point with respect to the limit distribution of the fraction of balls in each bin. It is well known that for $\gamma < 1$ the fraction of balls in the “heaviest” bin (the bin with the most balls) tends to $1/d$, and for $\gamma > 1$ the fraction of balls in the “heaviest” bin tends to 1. To partially fill in the gap for $\gamma = 1$, we characterize the limit distribution of the fraction of balls in the “heaviest” bin for $\gamma = 1$ by providing explicit analytical expressions for all its moments.

Keywords: Polya urn, Proportional Preferential Attachment, Limit distribution

1. Preliminaries

The Polya urn problem describes a well-studied family of random processes that have been fruitfully applied in diverse fields ranging from telecommunications to understanding self-organizing processes like network formation and herd behavior. In the classical Polya urn problem, one begins with $d$ bins, each containing one ball. Additional balls arrive one at a time, and the probability that an arriving ball is placed in a given bin is proportional to $m^\gamma$, where $m$ is the number of balls in that bin.

In this note, we consider the case of $\gamma = 1$, which corresponds to a process of “proportional preferential attachment” and is a critical point with respect to the limit distribution of the fraction of balls in each bin. It is well known that for $\gamma < 1$ the fraction of balls in the “heaviest” bin (the bin with the most balls) tends to $1/d$, and for $\gamma > 1$ the fraction of balls in the “heaviest” bin tends to 1. (See, for instance, [1].)

To partially fill in the gap for $\gamma = 1$, we characterize the limit distribution of the fraction of balls in the “heaviest” bin for $\gamma = 1$ by providing explicit analytical expressions for all its moments. But before proceeding to prove that, we reproduce, for completeness, the following well-known result:

**Lemma 1.1.** Suppose there are $d$ urns. The probability distribution of the number of balls in the $d$ urns after $m$ additional balls are added is uniform on

$$S_{(d,m)} := \{v \in \mathbb{Z}^d : v \geq e, v^T e = d + m\}. \quad (1)$$
Furthermore,
\[ |S_{(d,m)}| = \frac{(m + d - 1)!}{m!(d - 1)!} = \frac{(m + d - 1)}{d - 1}. \] (2)

**Proof of Lemma 1.1**: Let the probability that \( v_k \) balls are in the \( k \)-th urn (of \( d \)) after \( m \) additional balls are added be \( \pi_{(d,m)}(v) \). Clearly, \( \pi_{(d,0)}(v) = 1 \).

Suppose \( \pi_{(d,m)}(v) = \eta_{(d,m)} \) for some constant \( \eta_{(d,m)} \) for all \( v \in S_{(d,m)} \). Then, for all \( v \in S_{(d,m+1)} \),
\[
\pi_{(d,m+1)}(v) = \sum_{v > 1} \eta_{(d,m)} \frac{v - 1}{m + d} = \sum_{k=1}^{d} \eta_{(d,m)} \frac{v_k - 1}{m + d} \]
\[ = \eta_{(d,m)} \frac{m + 1}{m + d} = \eta_{(d,m+1)} \] (3) (4) (5)

where the first equality follows from the dynamics of preferential attachment. This completes the first part of the proof.

Now, \( |S_{(d,m)}| = 1/\eta_{(d,m)} \), and \( |S_{(d,0)}| = 1 \). Therefore, by Equation (3),
\[ S_{(d,m)} = 1 \cdot \frac{d}{1} \cdot \frac{d + 1}{2} \cdots \frac{m + d - 1}{m} \] (6)
and the proof is complete. \( \blacksquare \)

We will leverage Lemma 1.1 and a simple partitioning of the set of possible outcomes to characterize the limiting distribution of the fraction of balls in the “heaviest” urn. In the proofs that follow, we will also make use of the easily verifiable fact that:

**Lemma 1.2.** For integral \( a, b > 0 \) and \( c > 0 \),
\[ \int_0^\infty x^a(c-x)^b \, dx = \frac{a!b!}{(a+b+1)!} e^{a+b+1}. \] (7)

**2. The Main Result**

Denote the number of balls in the “heaviest” urn (when there are \( d \) urns) after a total of \( n - d \geq 0 \) balls are added as \( H_d(n) \). We will first provide a result for the limiting first moment (the mean) of \( H_d(n)/n \) to outline the key proof ideas, and then develop the general result for all the higher moments.

For integer \( m \geq 0 \) and integer \( d \geq 1 \), let
\[ M_d^{(m)} := \lim_{n \to \infty} \mathbb{E} \left[ \left( \frac{H_d(n)}{n} \right)^m \right]. \] (8)

For notational convenience, let \( M_0^{(0)} := 1 \).

**Proposition 2.1 (Limiting Moments of the Fraction of Balls in the “Heaviest Urn”).** For integer \( m \geq 1 \) and integer \( d \geq 1 \),
\[ M_d^{(m)} = \sum_{k=0}^{m} \frac{d - 1}{d^k} \frac{m!}{(m-k)!} \frac{(m + d - k - 2)!}{(m + d - 1)!} M_{d-1}^{(m-k)}. \] (9)
Proof of Proposition 2.3. Clearly, \( M_d^{(0)} = 1 \) and \( M_d^{(m)} = 1 \). For cases where \( m > 0 \) and/or \( d > 1 \), the result will be proven by constructing an expression for the above expectation and then taking the limit as \( n \to \infty \).

First, note that the set \( S_{(d,(\alpha-1)d)} \) (for \( \alpha \in \mathbb{N} \)) as defined in Lemma 1.1 may be expressed as the following disjoint union:

\[
S_{(d,(\alpha-1)d)} = \{ \alpha e \} \cup \bigcup_{\mu=1}^{\alpha-1} \bigcup_{\tau=1}^{d-1} T_{(\alpha,d,\mu,\tau)}
\]

(10)

where

\[
T_{(\alpha,d,\mu,\tau)} := \{ v \in \mathbb{Z}^d : v \geq \mu e, v^T e = \alpha d, \gamma(v,\mu) = \tau \}
\]

(11)

and \( \gamma(v, x) := ||k : v_k = x|| \) is the number of entries of the vector \( v \) with the value \( x \). This is because there is a single vector in \( S_{(d,(\alpha-1)d)} \) where all entries take the same value \( (\alpha e) \) and \( \alpha - 1 \) other possible values of the smallest entry of vectors in \( S_{(d,(\alpha-1)d)} \) (specifically, 1, 2, \ldots, \( \alpha - 1 \)). In each of the latter cases, the number of entries taking on the minimum value may range from 1 to \( d - 1 \).

Noting that vectors in \( T_{(\alpha,d,\mu,\tau)} \) each have \( \tau \) entries taking value \( \mu \), and \( d - \tau \) entries taking values strictly larger than \( \mu \), the cardinality of \( T_{(\alpha,d,\mu,\tau)} \) must be \( |S_{(d-\tau,(\alpha-\mu)d-(d-\tau))}| \) multiplied by the number of ways to pick the \( \tau \) entries taking value \( \mu \). Therefore, using Lemma 1.1 one may deduce that

\[
|T_{(\alpha,d,\mu,\tau)}| = |S_{(d-\tau,(\alpha-\mu)d-(d-\tau))}| \binom{d}{\tau} \binom{d-\tau-1}{\tau}.
\]

(12)

Thus one obtains the identity

Corollary 2.2.

\[
\binom{ad-1}{d-1} = 1 + \sum_{\mu=1}^{\alpha-1} \sum_{\tau=1}^{d-1} \binom{\alpha-\mu}{d-\tau-1} \binom{d}{\tau}
\]

(13)

which, itself, may be proven directly by induction.

Now, equation (8) may be written equivalently as

\[
E[H_d(n)^m] = M_d^{(m)} n^m + o(n^m)
\]

(14)

for integer \( m \geq 1 \). Furthermore, equation (13) clearly holds for \( d = 1 \).

Now, suppose that equation (14) is true for 1, 2, \ldots, \( d - 1 \) urns. Then noting that

\[
E[(H_d(n) + \mu)^m] = E \left[ \sum_{k=0}^{m} \binom{m}{k} H_{d}(n)^{m-k} \mu^k \right],
\]

the conditional on realizations being in \( T_{(\alpha,d,\mu,\tau)} \), the \( m \)-th moment of the number of balls in the “heaviest” urn may be shown to be

\[
E[H_d(n)^m|T_{(\alpha,d,\mu,\tau)}] = E[(H_d-\tau((\alpha-\mu)d + \mu)^m)]
\]

\[
= \sum_{k=0}^{m} \binom{m}{k} M_{d-\tau}^{(m-k)} ((\alpha-\mu)d)^{m-k} \mu^k + o(\alpha^m)
\]

(15)

with an application of Equation (14). Therefore,

\[
E\left( \frac{H_d(ad)}{ad} \right)^m = \frac{1}{(ad)^m} \sum_{\mu=1}^{\alpha-1} \sum_{\tau=1}^{d-1} \left[ \sum_{k=0}^{m} \binom{m}{k} M_{d-\tau}^{(m-k)} ((\alpha-\mu)d)^{m-k} \mu^k + o(\alpha^m) \right] \binom{\alpha-\mu}{d-\tau-1} \binom{d}{\tau}
\]

(16)
and taking limits,

\[
\lim_{a \to \infty} \mathbb{E} \left[ \left( \frac{H_d(ad)}{ad} \right)^m \right] = \lim_{a \to \infty} \frac{1}{(ad)^m} \left[ \frac{\alpha + \sum_{\mu=1}^{a-1} \sum_{\tau=1}^{d-1} \left( \sum_{k=0}^{m} \frac{m!}{k!} \right) M_{d-\tau}^{(m-k)}((\alpha - \mu)d)^{m-k} \mu^k + o(\alpha^m)}{\left( \frac{ad - 1}{d - 1} \right)^{(d-1)!}} \right]
\]

\[
= \lim_{a \to \infty} \frac{d - 1}{d^{m+ad-1}} \sum_{\mu=1}^{a-1} \sum_{k=0}^{m} \left( \frac{m!}{k!} \right) M_{d-1}^{(m-k)}((\alpha - \mu)d)^{m-k} \mu^k + o(\alpha^m) \left( \frac{\alpha - \mu}{d} \right)^{d-2}
\]

\[
= \lim_{a \to \infty} \frac{d - 1}{d^{m+ad-1}} \left[ \sum_{k=0}^{m} \left( \frac{m!}{k!} \right) M_{d-1}^{(m-k)}((\alpha - \mu)^{m+d-k-2} \mu^{d-k} + o(\alpha^{m+d-2})) \right]
\]

\[
= \lim_{a \to \infty} \frac{d - 1}{d^{m+ad-1}} \left[ \sum_{k=0}^{m} \left( \frac{m!}{k!} \right) M_{d-1}^{(m-k)}((m + d) - k)! \frac{1}{(m + d - 1)!} \mu^{d-k} + o(\alpha^{m+d-1}) \right]
\]

\[
= \sum_{k=0}^{m} \frac{d - 1}{d^k} \left( \frac{m!}{(m - k)!} \right) \frac{(m + d - k - 2)!}{(m + d - 1)!} M_{d-1}^{(m-k)} + o(1)
\]

The second and third equalities hold because in taking the limit \(\alpha \to \infty\), it suffices to consider only the highest order terms. The sixth equality arises by an application of Lemma 10 to evaluate the integral.

To complete the proof, we argue that the limit

\[
\lim_{r \to \infty} \mathbb{E} \left[ \left( \frac{H_d(n_r)}{n_r} \right)^m \right] = \sum_{k=0}^{m} \frac{d - 1}{d^k} \left( \frac{m!}{(m - k)!} \right) \frac{(m + d - k - 2)!}{(m + d - 1)!} M_{d-1}^{(m-k)}
\]

also arises for all increasing positive integer sequences \(\{n_r\}_{r=1}^{\infty}\) whose elements are greater than \(d\) but are not all necessarily integral multiples of \(d\).

Note that given any \(n\), for \(\beta_{n,d} := d \lceil x/d \rceil\), we have \(\beta_{n,d} \leq n \leq (\beta_{n,d} + 1)d\). Since, in the right-hand-size of equation (16), both the numerator and denominator are increasing in \(\alpha\), one can construct upper and lower bounds by using either \(\beta_{n,d}\) or \(\beta_{n,d} + 1\) accordingly in place of \(\alpha\) in the numerator and denominator. However, for both upper and lower bounds, the same leading order terms arise and hence the same limits. Therefore, with this sandwiching, the proof is complete.

Following from Proposition 12, a convenient expression arises for the limiting mean by solving the resulting recurrence.

**Corollary 2.3 (Limiting Mean Fraction of Balls in the “Heaviest Urn”):**

\[
M^{(1)}_d = \frac{1}{d} \sum_{k=1}^{d} \frac{1}{k}
\]

(17)
Proof of Corollary 2.3: Equation (17) informs us that for \( d > 1 \)
\[
M_d^{(1)} = \frac{d - 1}{d} M_{d - 1}^{(1)} + \frac{1}{d^2} M_{d - 1}^{(0)}
\]
Noting that \( M_{d - 1}^{(0)} = 1 \) and \( M_d^{(1)} = 1 \), one may easily verify the truth of the corollary.

Furthermore, it is readily apparent that
\[
\frac{1}{d}(\log d) \leq M_d^{(1)} \leq \frac{1}{d}(\log d + 1).
\]

3. Numerical Experiments and Approximating the Expectation of Smooth Functions of \( H_d(n)/n \)

Simulations show that those limiting values are good approximations even for systems with just thousands of balls added. Illustrations of Proposition 2.3 may be found in Figures 1 thru 8 along with simulated average fractions of balls in the “heaviest” urn and quantiles obtained from simulation (for \( n \) ranging from 100 to 20,000). Note the similarity of the figures over the large range of \( n \).

Using Taylor’s Theorem, we obtain the following error estimate for a smooth function \( f \) that is \( m + 1 \) times continuously differentiable:

\[
\left| \mathbb{E} \left\{ f \left( \frac{H_d(n)}{n} \right) \right\} - \sum_{k=0}^{m} \frac{f^{(k)}(0)}{k!} M_d^{(k)} \right| \leq \max_{\tau \in [0,1]} \left( \frac{f^{(m+1)}(\tau)}{(m+1)!} \right) \mathbb{E} \left\{ \left( \frac{H_d(n)}{n} \right)^{m+1} \right\} + \sum_{k=0}^{m} \frac{f^{(k)}(0)}{k!} \mathbb{E} \left\{ \left( \frac{H_d(n)}{n} \right)^k \right\} - M_d^{(k)} \right|.
\]

Note also that since the fraction of balls in the “heaviest” urn is a (non-degenerate) random variable taking values in \((0, 1)\),
\[
1 > \mathbb{E} \left[ \frac{H_d(n)}{n} \right] > \mathbb{E} \left[ \left( \frac{H_d(n)}{n} \right)^2 \right] > \ldots > \mathbb{E} \left[ \left( \frac{H_d(n)}{n} \right)^m \right] > \mathbb{E} \left[ \left( \frac{H_d(n)}{n} \right)^{m+1} \right] > \ldots
\]
and
\[
1 > M_d^{(1)} > M_d^{(2)} > \ldots > M_d^{(m)} > M_d^{(m+1)} > \ldots
\]
for \( d > 1 \) (equality holds throughout for \( d = 1 \)). All this indicates that a truncated Taylor series as in equation (18) may provide a good approximation when the number of balls is sizable. (Simulating the outcome of the Polya’s Urn process is equivalent to sampling a uniform permutation of \([1, 2, \ldots, n]\), which is relatively expensive, requiring \( O(n) \) time and space.)

References

[1] F. Chung, S. Handjani, D. Jungreis, Generalizations of Polya’s urn Problem, Annals of Combinatorics 7 (2) (2003) 141–153.
Figure 1. Maximum Fraction of Balls in the “Heaviest” Urn: The Limiting Mean and Simulated Data ($n = 100; 10,000$ samples)

Figure 2. Maximum Fraction of Balls in the “Heaviest” Urn: The Limiting Mean and Simulated Data ($n = 250; 10,000$ samples)

Figure 3. Maximum Fraction of Balls in the “Heaviest” Urn: The Limiting Mean and Simulated Data ($n = 500; 10,000$ samples)
Figure 4. Maximum Fraction of Balls in the “Heaviest” Urn: The Limiting Mean and Simulated Data ($n = 1,000; 10,000$ samples)

Figure 5. Maximum Fraction of Balls in the “Heaviest” Urn: The Limiting Mean and Simulated Data ($n = 2,000; 10,000$ samples)

Figure 6. Maximum Fraction of Balls in the “Heaviest” Urn: The Limiting Mean and Simulated Data ($n = 5,000; 10,000$ samples)
Figure 7. Maximum Fraction of Balls in the “Heaviest” Urn: The Limiting Mean and Simulated Data ($n = 10,000$; 10,000 samples)

Figure 8. Maximum Fraction of Balls in the “Heaviest” Urn: The Limiting Mean and Simulated Data ($n = 20,000$; 10,000 samples)