J. Appl. Prob. **52,** 180–190 (2015)
Printed in England

MAXIMAL PERCENTAGES IN PÓLYA'S URN

ERNST SCHULTE-GEERS,* *Bundesamt für Sicherheit in der Informationstechnik*
WOLFGANG STADJE,** *University of Osnabrück*

Abstract

We show that the supremum of the successive percentages of red balls in Pólya's urn model is almost surely rational, give the set of values that are taken with positive probability, and derive several exact distributional results for the all-time maximal percentage.

Keywords: Pólya's urn; binomial random walk; all-time maximal percentage; exact distribution

2010 Mathematics Subject Classification: Primary 60C05

1. Introduction

The classical urn of Eggenberger–Pólya [3] contains initially $r \geq 1$ red balls and $b \geq 1$ black balls. In the course of the drawing process, in each draw one ball is taken from the urn (where each ball in the urn has the same chance of being drawn), and this ball and another $d \geq 1$ balls of the same color are put into the urn. The theory of Pólya urn schemes is expounded in [8], where an extensive bibliography can be found.

Let R_n, $n \geq 1$, denote the number of red balls in the urn after the nth draw and set $R_0 = r$. The ratio $Z_n = R_n/(nd + r + b)$ gives the percentage of red balls among all balls in the urn at 'time' n. Let

$$S_{r,b} = \sup_{n \geq 0} Z_n$$

be the supremum of all successive percentages of red balls during the entire drawing process.

In this paper we show that $S_{r,b}$ is attained almost surely (a.s.) and that $\mathbb{P}\{S_{r,b} \text{ is rational}\} = 1$, thereby settling in the affirmative a conjecture of Knuth (posed in answer to problem 88 of prefascicle 5a to volume 4B of *The Art of Computer Programming* [7]). The support of $S_{r,b}$ (i.e. the set of values that are taken with positive probability) is also given. Moreover, we derive exact equations for certain values of the distribution function of $S_{r,b}$, mainly in the case $d = 1$. For general $d \geq 1$ we show that

$$\mathbb{P}\left\{S_{r,b} > \frac{t-1}{t}\right\} = \sum_{n=0}^{\infty} \frac{a+1}{n(t-1)+a+1} \binom{nt+a}{n} \frac{B(n(t-1)+a+1+(r/d), n+(b/d))}{B(r/d, b/d)},$$

Received 14 January 2014; revision received 11 April 2014.
* Postal address: Bundesamt für Sicherheit in der Informationstechnik (BSI), Godesberger Allee 185–189, 53175 Bonn, Germany. Email address: ernst.schulte-geers@bsi.bund.de
** Postal address: Institute of Mathematics, University of Osnabrück, 49069 Osnabrück, Germany. Email address: wstadje@uos.de

where $(r+b)/r \le t \in \mathbb{N}$, $a = \lfloor (b(t-1)-r)/d \rfloor$ and $B(\alpha, \beta)$ is the Beta function. For $d = 1$ we have

$$\mathbb{P}\left\{S_{1,t-1} \le \frac{1}{t}\right\} = \begin{cases} 1 - \ln 2 & \text{for } t = 2, \\ \dfrac{(t-1)(1-1/t)^{t-2} - 1}{t-2} & \text{for } t > 2 \end{cases}$$

and

$$\mathbb{P}\left\{S_{1,1} = \frac{t-1}{t}\right\} = \frac{2t-3}{t} H\left(1 - \frac{1}{t}\right) - \frac{t-2}{t} H\left(1 - \frac{2}{t}\right) - \frac{t-2}{t-1},$$
$$\mathbb{P}\left\{S_{1,1} \le \frac{t-1}{t}\right\} = \left(1 - \frac{1}{t}\right) H\left(1 - \frac{1}{t}\right),$$

where $H(x) = \Psi(x+1) + \gamma$. Here $\Psi = \Gamma'/\Gamma$ denotes the Digamma function and γ is Euler's constant. Another exact result in the form of an infinite series is (again for $d = 1$)

$$\mathbb{P}\left\{S_{t-1,1} > \frac{t-1}{t}\right\} = (t-1) \sum_{n=0}^{\infty} \frac{(nt)!}{(nt-n+1)!} \frac{((n+1)(t-1))!}{((n+1)t)!}.$$

In the course of our derivations we obtain a remarkable *equidistribution property* of the supremum $M(p) = \sup_{n \in \mathbb{N}} n^{-1}(B_1 + \cdots + B_n)$ of the binomial random walk generated by independent and identically distributed (i.i.d.) Bernoulli variables B_i (i.e. $\mathbb{P}\{B_i = 1\} = 1 - \mathbb{P}\{B_i = 0\} = p \in (0, 1)$). For $t \in \mathbb{N}$ with $pt \le 1$ we show that

$$\mathbb{P}\left\{M(p) \in \left(\frac{1}{k+1}, \frac{1}{k}\right]\right\} = \frac{p}{1-p} \quad \text{for every } k \in \{1, \ldots, t-1\},$$
$$\mathbb{P}\left\{M(p) \in \left(p, \frac{1}{t}\right]\right\} = \frac{1-tp}{1-p}.$$

2. Used facts and related work

In Pólya's urn scheme let $X_n = 1$ if a red ball is drawn in the nth draw and $X_n = 0$ otherwise, and let $X = (X_1, X_2, \ldots)$ be the full sequence of these red ball indicators.

The following facts about Pólya's urn are well known [2], [4].

(1) Z_n converges a.s. to a random variable Z, which has a Beta$(r/d, b/d)$ distribution on $(0, 1)$. (Here and in the sequel Beta(α, β) denotes the Beta distribution with parameters $\alpha, \beta > 0$.)

(2) Conditionally on $Z = z$, $X_1, X_2, \ldots$, are independent, $\{0, 1\}$-valued random variables with $\mathbb{P}\{X_i = 1\} = z$.

Thus, the percentage Z_n of red balls eventually tends to a random value Z, and if Z_n 'exits' to $Z = z$, the red ball indicators behave like independent $0-1$ variables with 'success probability' z. One can therefore try to average known results for the Bernoulli sequence to obtain results for Pólya's urn.

For a sequence of $\{-1, 1\}$-valued Bernoulli variables $U_1, U_2, \ldots$, with $\mathbb{P}\{U_i = 1\} = 1 - \mathbb{P}\{U_i = -1\} = p \in (0, 1)$ the supremum,

$$M = \sup_{n \ge 1} n^{-1}(U_1 + \cdots + U_n),$$

https://doi.org/10.1239/jap/1429282614 Published online by Cambridge University Press

of the averages of their partial sums was considered in [9]. It was shown that this supremum is attained and that

(1) $\mathbb{P}\{M \in (2p-1, 1] \cap \mathbb{Q}\} = 1$.

(2) $\mathbb{P}\{M = x\} > 0$ for each $x \in (2p-1, 1] \cap \mathbb{Q}$.

Moreover, explicit equations for the distribution of M were derived. Note that M is one of the rare examples of a naturally occurring random variable taking every rational number in some interval with positive probability.

In the sequel we combine these results in order to study the all-time maximal percentage of red balls in Pólya's urn.

3. Existence and possible values of maxima

Let X be the Pólya red ball indicator sequence. We have already remarked that one may view this sequence as a randomized Bernoulli sequence. To make this paper self-contained we include a short proof.

Theorem 3.1. *Let Z and $Y = (Y_1, Y_2, \dots,)$ be defined on some probability space such that*

(a) *Z is $Beta(r/d, b/d)$-distributed;*

(b) *conditionally on $Z = z$, $(Y_1, Y_2, \dots,)$ is an i.i.d. sequence of $\{0, 1\}$-valued random variables with $\mathbb{P}\{Y_i = 1\} = 1 - \mathbb{P}\{Y_i = 0\} = z$.*

Then $Y \stackrel{\mathrm{D}}{=} X$.

Proof. Let $(y_1, \dots, y_n) \in \{0, 1\}^n$, $s_n = y_1 + \cdots + y_n$. Then

$$\mathbb{P}\{Y_1 = y_1, \dots, Y_n = y_n \mid Z = z\} = z^{s_n}(1-z)^{n-s_n}.$$

Therefore,

$$\begin{aligned}\mathbb{P}\{Y_1 = y_1, \dots, Y_n = y_n\} &= \frac{1}{B(r/d, b/d)} \int_0^1 z^{(r/d)-1+s_n}(1-z)^{n+(b/d)-1-s_n}\,\mathrm{d}z \\ &= \frac{B((r/d)+s_n, n+(b/d)-s_n)}{B(r/d, b/d)}.\end{aligned}$$

Hence, a straightforward calculation (using the elementary properties of the Beta function) yields $\mathbb{P}\{Y_1 = 1\} = r/(r+b)$ and

$$\mathbb{P}\{Y_{n+1} = 1 \mid Y_1 = y_1, \dots, Y_n = y_n\} = \frac{r + s_n d}{nd + r + b},$$

so that the finite-dimensional distributions of Y coincide with those of X. Thus, $Y \stackrel{\mathrm{D}}{=} X$.

Since the red ball indicators in Pólya's urn scheme are a randomized Bernoulli sequence they have similar properties.

Proposition 3.1. *In the situation of Theorem 3.1 let $Z_n = (r + dS_n)/(nd + r + b)$ and $S_n = \sum_{i=1}^n Y_i$. Then $\lim_{n\to\infty} Z_n = Z$ a.s. and $\limsup_{n\to\infty}(S_n - nZ) = \infty$ a.s.*

https://doi.org/10.1239/jap/1429282614 Published online by Cambridge University Press

Proof. Let $z \in (0, 1)$. We need to prove only the assertion conditionally on $Z = z$. Since Y is, conditionally on $Z = z$, a sequence of i.i.d. $0-1$ variables with probability z for the value 1, we have

$$\mathbb{P}\Big\{Z_n \to z,\ \limsup_{n\to\infty}(S_n - nz) = \infty \ \Big|\ Z = z\Big\} = 1.$$

Note that $\mathbb{P}\{Z_n \to z \mid Z = z\} = 1$ follows from the strong law of large numbers, and $\mathbb{P}\{\limsup_{n\to\infty}(S_n - nz) = \infty \mid Z = z\} = 1$ follows (e.g.) from the law of the iterated logarithm.

Let $M_{r,b} = \sup_{n\geq 1} Z_n$. We now use a variant of the proof of the corresponding property of M in [9] to show that this supremum is a.s. attained.

Proposition 3.2. (a) *The supremum in the definition of $M_{r,b}$ is a.s. attained.*

(b) $\mathbb{P}\{M_{r,b}$ *is rational*$\} = 1$.

Proof. Again we condition on $Z = z$. Since $\limsup_{n\to\infty}(S_n - nz) > b+1$ a.s. it follows that a.s. infinitely often $(r+dS_n)-(nd+r+b)z > d(b+1)+r(1-z)-bz \geq 1+(r+b)(1-z) > 0$, i.e. $Z_n > z$ for infinitely many n a.s., and, since $Z_n \to z$ a.s., the supremum is a.s. attained. It is then obviously rational.

Define

$$Q = \left\{ \max_{1\leq j\leq n} \frac{r + i_j d}{r + b + jd} \ \middle|\ n \in \mathbb{N},\ i_1, \ldots, i_n \in \mathbb{N} \cup \{0\},\right.$$
$$\left. i_j \leq j \text{ for all } j \in \{1, \ldots, n\},\ 0 \leq i_1 \leq i_2 \leq \cdots \leq i_n \right\}.$$

Clearly, Q is the set of *all* possible maximal percentages for finite sequences of draws from the urn. It follows from Proposition 3.2 that $\mathbb{P}\{M_{r,b} \in Q\} = 1$.

Proposition 3.3. *We have $\mathbb{P}\{M_{r,b} = q\} > 0$ for all $q \in Q$.*

Proof. Fix an element q of Q and correspondingly $m \in \mathbb{N}$ and nonnegative integers $i_1 \leq i_2 \leq \cdots \leq i_m$ such that $i_j \leq j$ for all $j \in \{1, \ldots, m\}$ and

$$q = \max_{1\leq j\leq m} \frac{r + i_j d}{r + b + jd}.$$

Let the maximum be attained at $n \in \{1, \ldots, m\}$. Then we have

$$q = \frac{r + i_n d}{r + b + nd} = \max_{1\leq j\leq n} \frac{r + i_j d}{r + b + jd}.$$

Let E be the event that $R_j = i_j d$ for $j = 1, \ldots, n$ and $M_{r,b} = q$ (thus, on this event the percentage after the nth draw is the largest among the first n ones). It remains to show that $\mathbb{P}\{E\} > 0$.

https://doi.org/10.1239/jap/1429282614 Published online by Cambridge University Press

Write $\mathbb{P}\{E \mid Z = z\}$ in the form

$$\begin{aligned}\mathbb{P}\{E \mid Z = z\} &= \mathbb{P}\Bigg\{R_j = i_j d \text{ for } j = 1, \dots, n \text{ and} \\ &\qquad \sup_{k\in\mathbb{N}} \frac{r + i_n d + (Y_1' + \cdots + Y_k')d}{r + b + nd + kd} \le \frac{r + i_n d}{r + b + nd} \;\Bigg|\; Z = z\Bigg\} \\ &= \mathbb{P}\{R_j = i_j d \text{ for } j = 1, \dots, n \mid Z = z\} \\ &\qquad \times \mathbb{P}\left\{\sup_{k\in\mathbb{N}} \frac{r + i_n d + (Y_1' + \cdots + Y_k')d}{r + b + nd + kd} \le \frac{r + i_n d}{r + b + nd}\right\},\end{aligned}$$

where $Y_1', Y_2', \dots,$ is an i.i.d. sequence with $\mathbb{P}\{Y_i' = 1\} = 1 - \mathbb{P}\{Y_i' = 0\} = z$ which is independent of Y and Z. The first factor on the right-hand side is obviously positive. Next, the inequality

$$\sup_{k\in\mathbb{N}} \frac{r + i_n d + (Y_1' + \cdots + Y_k')d}{r + b + nd + kd} \le \frac{r + i_n d}{r + b + nd}$$

holds if and only if

$$\sup_{k\in\mathbb{N}} \frac{Y_1' + \cdots + Y_k'}{k} \le \frac{r + i_n d}{r + b + nd}.$$

By [9], this latter inequality occurs with positive probability if

$$\mathbb{P}\{Y_1' = 1\} < \frac{r + i_n d}{r + b + nd}.$$

It follows that $\mathbb{P}\{E \mid Z = z\} > 0$ for $z \in (0, (r + i_n d)/(r + b + nd))$. Hence,

$$\mathbb{P}\{E\} \ge \int_0^{\min[1,(r+i_n d)/(r+b+nd)]} \mathbb{P}\{E \mid Z = z\}\beta_{r/d,b/d}(z)\,\mathrm{d}z > 0,$$

where $\beta_{r/b,b/d}(z)$ is the density of Beta$(r/b, b/d)$. The proof is complete.

Remark 3.1. Proposition 3.3 can also be formulated in the following form. Let $\mathrm{supp}(U)$ denote the support of the random variable U. Let $q_{r,b}(i, n) = (r + id)/(r + b + nd)$ for $i, n \in \mathbb{N} \cup \{0\}$. Then

$$\mathrm{supp}(S_{r,b}) = \left\{q_{r,b}(i, n) \;\middle|\; 0 \le i \le n,\ q_{r,b}(i, n) \ge \frac{r}{r + b}\right\} \tag{3.1}$$

and

$$\mathrm{supp}(M_{r,b}) = \mathrm{supp}(S_{r,b+d}) \cup \mathrm{supp}(S_{r+d,b}). \tag{3.2}$$

Equation (3.2) is obvious, and (3.1) is proved in the same way as Proposition 3.3 (consider sequences of draws beginning with $X_1 = \cdots = X_{n-i} = 0,\ X_{n-i+1} = \cdots X_n = 1$).

In particular, for $d = 1$ we obtain $\mathrm{supp}(S_{r,b}) = [r/(r + b), 1) \cap \mathbb{Q}$.

4. Some closed-form results on maximal percentages for the binomial random walk

In this section we consider only the $d = 1$ case. We start with two exact results for the binomial random walk, which may be of independent interest. Although they are consequences of the t-ballot theorems they apparently have not been formulated in this form before.

https://doi.org/10.1239/jap/1429282614 Published online by Cambridge University Press

In the sequel fix $0 < p < 1$, $q = 1 - p$ and let $Y_1, Y_2, \ldots$, be i.i.d. $0-1$ random variables with $\mathbb{P}\{Y_i = 1\} = p$. We introduce their partial sums $S_n = Y_1 + \cdots + Y_n$ and define, for $r, b \in \mathbb{N} \cup \{0\}$,

$$M_{r,b}(p) = \sup_{n \geq 1} \frac{r + S_n}{r + b + n} \quad \text{and} \quad M(p) = M_{0,0}(p),$$

$$L_{r,b}(p) = \inf_{n \geq 1} \frac{r + S_n}{r + b + n} \quad \text{and} \quad L(p) = L_{0,0}(p).$$

Since $\mathbb{P}\{M_{r,b}(p) \leq x\} = \mathbb{P}\{L_{b,r}(q) \geq 1-x\}$ for $0 \leq x \leq 1$ the distribution function of $L_{b,r}(q)$ can be obtained from that of $M_{r,b}(p)$. Clearly, $M_{r,b}(p) \leq x$ if and only if $S_n - nx \leq bx - (1-x)r$ for all $n \in \mathbb{N}$. Thus, if $x \in \mathbb{Q} \cap (0, 1)$, say $x = s/t$ for positive integers s and t with greatest common divisor $\gcd(s, t) = 1$, the value of $\mathbb{P}\{M_{r,b}(p) \leq s/t\}$ depends on s, t only through the integer value $m = sb - (t - s)r$. For $m \in \mathbb{Z}_+$ we define $a_m = \mathbb{P}\{tS_n - ns \leq m \text{ for all } n \in \mathbb{N}\}$.

Remark 4.1. For two integers s, t satisfying $(s, t) = 1$, $p < s/t < 1$ the following can be shown (along the lines of [9]).

(a) The sequence $a_0, a_1, a_2, \ldots$, has a rational generating function; it is given by

$$\frac{q \sum_{j=0}^{s-1} a_j z^j}{pz^t - z^s + q}.$$

This function is regular for $|z| < 1$.

(b) The denominator $f(z) = pz^t - z^s + q$ has only simple roots: exactly $s - 1$ roots $z_1, \ldots, z_{s-1}$ inside the unit disk, $z_s = 1$, and exactly $t - s$ roots $z_{s+1}, \ldots, z_t$ outside the unit disk.

Thus, a_m can be written as a linear combination of 1 and the $(m + 1)$th negative powers of the roots of $f(z)$ outside the unit disk. However, explicit expressions are hard to come by.

Let $t \geq 2$ be an integer. In the sequel we will give expressions for the probabilities $\mathbb{P}\{M(p) \leq (t - 1)/t\}$ and $\mathbb{P}\{M(p) \leq 1/t\}$.

For the first case we need the $t-$ary tree function $T_t(z)$. Its power series is given by

$$T_t(z) = \sum_{n=0}^{\infty} \binom{nt}{n} \frac{z^n}{n(t-1)+1}.$$

This power series converges for $|z| < (1/t)(1 - 1/t)^{t-1}$ and for $z = (1/t)(1 - 1/t)^{t-1}$, and represents there the unique solution of the implicit equation $T_t(z) = 1 + zT_t(z)^t$. (Thus, $1/T_t$ is the inverse function of $y \mapsto y^{t-1}(1 - y)$ for $|1 - y| < 1/t$.) We define

$$R_t(p) = pT_t(qp^{t-1}).$$

We recall the following fact from path counting combinatorics ('t-ballot numbers'). (See, e.g. [6, Problem 26, Section 7.2.1.6] for an equivalent statement.)

Lemma 4.1. *Let $a \in \mathbb{N} \cup \{0\}$. The number of paths, with steps in $\{(0, 1), (1, 0)\}$, from $(0, 0)$ to $(n, n(t - 1) + a)$ whose points lie on or below the line $y = a + x(t - 1)$ is the coefficient $[z^n]T_t(z)^{a+1}$. We have*

$$[z^n]T_t(z)^{a+1} = \frac{a+1}{n(t-1)+a+1} \binom{nt+a}{n}.$$

https://doi.org/10.1239/jap/1429282614 Published online by Cambridge University Press

Proposition 4.1. *If $m = b(t-1) - r \geq 0$, we have*

$$\mathbb{P}\left\{M_{r,b}(p) \leq \frac{t-1}{t}\right\} = 1 - R_t(p)^{m+1}, \tag{4.1}$$

$$\mathbb{P}\left\{M_{r,b}(p) = \frac{t-1}{t}\right\} = \begin{cases} p - R_t(p) + q\,R_t(p)^{t-1} & \text{for } m = 0, \\ (1 - R_t(p))R_t(p)^m & \text{for } m \geq 1. \end{cases}$$

Proof. (1) Consider $\mathbb{P}\{M_{r,b}(p) > (t-1)/t\}$. Start a lattice path $R_n = r + S_n$, $B_n = b + n - S_n$ in (r, b) by stepping $(1, 0)$ if $Y_i = 1$ or $(0, 1)$ if $Y_i = 0$. Since $M_{r,b}(p) > (t-1)/t$, there must be a smallest $j \geq b$ such that $R_\ell = j(t-1) + 1$, $B_\ell = j$, and $(t-1)B_i \geq R_i$ for $i \leq \ell = jt - r - b + 1$.

Equivalently, $j - b$ is the first position where the lattice path $(n - S_n, S_n)$ starting at $(0, 0)$ steps at time ℓ for the first time above the line $y = (t-1)x + m$. Call this event A_j. Clearly, the last step is $Y_\ell = 1$, appended to a path from $(0, 0)$ to $(j - b, j(t-1) - r) = (j - b, (j-b)(t-1) + (t-1)b - r)$ of the type considered in the lemma above. Since each path from $(0, 0)$ to $(j - b, j(t-1) - r + 1)$ has the same probability $p^{j(t-1)-r+1}q^{j-b}$, we obtain

$$\mathbb{P}\{A_j\} = [z^{j-b}]T_t(z)^{m+1}p^{j(t-1)-r+1}q^{j-b} = p^{m+1}[z^{j-b}]T_t(z)^{m+1}p^{(j-b)(t-1)}q^{j-b}.$$

Since $\{M_{r,b}(p) > (t-1)/t\}$ is the disjoint union of the A_j, the first assertion follows.

(2) For $k \geq 0$ let $b_k = \mathbb{P}\{tS_n < n(t-1) + k\}$ for all $n \in \mathbb{N}$. Conditioning with respect to Y_1 shows that $b_0 = qa_{t-2}$ and $b_j = a_{j-1}$ for $j \geq 1$. Since $\mathbb{P}\{M_{r,b}(p) = (t-1)/t\} = a_m - b_m$ the second assertion follows from the result in (1).

Remarks 4.2. (a) The path counting argument above is due to Knuth, who used it to determine $\mathbb{P}\{S_{1,1} > (t-1)/t\}$ in Pólya's urn (see Section 5). Alternatively, the results could be obtained using the theorem from [9] given below, but in a more laborious way.

(b) In the derivation no restrictions on p (other than $0 < p < 1$) were imposed, the equations above are, thus, also valid for $p \geq (t-1)/t$. In fact, by the remark before Lemma 4.1 we have $R_t(p) = 1$ for $p \geq (t-1)/t$ and the corresponding probabilities in Proposition 4.1 evaluate to 0 (as they must do in view of the strong law of large numbers).

(c) We have $R_t(p) \to p$ for $t \to \infty$. More precisely, $\lim_{t\to\infty} q^{-1}p^{-t}[R_t(p) - p] = 1$.

Example 4.1. (a) Let $t = 2$, $p < \frac{1}{2}$. Then $R_2(p) = \min(1, p/q)$ and we recover the well-known facts that for $m = 0$

$$\mathbb{P}\{M(p) > \tfrac{1}{2}\} = \frac{p}{q} \quad \text{and} \quad \mathbb{P}\{M(p) = \tfrac{1}{2}\} = \frac{(q-p)p}{q},$$

and that for $m \geq 1$

$$\mathbb{P}\{M_{r,b}(p) \geq \tfrac{1}{2}\} = \left(\frac{p}{q}\right)^m \quad \text{and} \quad \mathbb{P}\{M_{r,b}(p) = \tfrac{1}{2}\} = \left(\frac{p}{q}\right)^m \frac{q-p}{q}.$$

(b) Let $p = \frac{1}{2}$. Then $R_2(p) = 1$ and

$$R_3(p) \approx 0.618\,034, \quad R_4(p) \approx 0.543\,689, \quad R_5(p) \approx 0.518\,790,$$
$$R_6(p) \approx 0.508\,66, \quad R_7(p) \approx 0.504\,138.$$

https://doi.org/10.1239/jap/1429282614 Published online by Cambridge University Press

For example, we have

$$\mathbb{P}\left\{\sup_{n\geq 1}\frac{S_n}{n} > \frac{6}{7}\right\} \approx 0.504\,138 \quad \text{and} \quad \mathbb{P}\left\{\sup_{n\geq 1}\frac{S_n}{n+1} > \frac{2}{3}\right\} \approx 0.381\,937.$$

Let us now turn to the second probability $\mathbb{P}\{M(p) \leq 1/t\}$. We use the following result of [9].

Theorem 4.1. *Let $s, r \in \mathbb{Z} \setminus \{0\}$, $s > 0$, $|r| < s$. The polynomial*

$$pz^{2s} - z^{s+r} + q \tag{4.2}$$

has exactly $s - r$ simple roots $z_{s+r}, \ldots, z_{2s-1}$ outside the unit disk. If $p < (r+s)/2s$, we have

$$\mathbb{P}\left\{M(p) \leq \frac{r+s}{2s}\right\} = \prod_{i=r+s}^{2s-1}(1 - z_i^{-1}),$$

$$\mathbb{P}\left\{M(p) = \frac{r+s}{2s}\right\} = \prod_{i=r+s}^{2s-1}(1 - z_i^{-1}) + p\prod_{i=r+s}^{2s-1}(1 - z_i).$$

Remarks 4.3. (a) In [9] this theorem is proved under the additional assumption $(r, s) = 1$, but it remains valid if $(r, s) = k > 1$. To see this let $u = z^k$. As a function of u the polynomial (4.2) has $(s - r)/k$ simple roots u_i outside the unit disk, say $u_1, \ldots, u_{(s-r)/k}$, by the theorem above. If η is a primitive kth root of unity, then as a function of z the polynomial (4.2) has exactly the $s - r$ simple roots $z_{i,j} = \eta^j |u_i|^{1/k}$, $j = 0, \ldots, k-1$, and $i = 1, \ldots, (s-r)/k$, outside the unit disk, and since $\prod_{j=0}^{k-1}(1 - z_{i,j}^{-1}) = (1 - u_i^{-1})$ and $\prod_{j=0}^{k-1}(1 - z_{i,j}) = (1 - u_i)$ the equations above remain valid.

(b) In [9] equations for the corresponding a_m in terms of the roots of (4.2) were also provided, but we do not use them here.

Proposition 4.2. *Let $p < 1/t$. If $m = b - (t-1)r \in \{0, \ldots, t-1\}$, we have*

$$\mathbb{P}\left\{M_{r,b}(p) \leq \frac{1}{t}\right\} = (1 - tp)q^{-(m+1)}, \tag{4.3}$$

$$\mathbb{P}\left\{M_{r,b}(p) = \frac{1}{t}\right\} = (1 - tp)q^{-(m+1)}p. \tag{4.4}$$

For $m \geq t$ the a_m can be computed by the recursion

$$qa_k = a_{k-1} - pa_{k-t}.$$

For $p \geq 1/t$ the probabilities in (4.3) and (4.4) are 0.

Proof. (1) Let $m = 0$ and set $s = t, r = -(t-2)$ in Theorem 4.1. The polynomial $pz^{2t} - z^2 + q$ has $2t - 2$ roots $z_2, \ldots, z_{2t-1}$ outside the unit disk, and $a_0 = \prod_{i=2}^{2t-1}(1 - 1/z_i)$. Equivalently, since $+z_i, -z_i$ are roots, we have $a_0 = \prod_{i=1}^{t-1}(1 - 1/y_i)$ where $y_1, \ldots, y_{t-1}$ are the roots of $py^t - y + q$ outside the unit disk. Let $y = 1/(1-u)$. Then we can write $a_0 = \prod_{i=1}^{t-1} u_i$ where the nonzero roots of the polynomial $g(u) = p - (1-u)^{t-1} + q(1-u)^t$ are $u_1, \ldots, u_{t-1}$. Thus, $qa_0 = -g'(0) = 1 - pt$. Furthermore, since $\prod_{i=2}^{2t-2} z_i = -q/p$ we obtain, from the second formula in Theorem 4.1, $\mathbb{P}\{M(p) = 1/t\} = a_0 - qa_0 = pa_0$.

(2) By conditioning on Y_1 we find that $a_k = pa_{k-t+1} + qa_{k+1}$, and $a_k = qa_{k+1}$ for $k = 0, \ldots, t-2$. The rest is straightforward.

https://doi.org/10.1239/jap/1429282614 Published online by Cambridge University Press

Remarks 4.4. (a) In the case $m = 0$ and $p \le 1/t$, we find a curious equidistribution property: the maximum $M(p)$ lies in any of the intervals $(1/(k+1), 1/k]$ (for $1 \le k \le t-1$) with the same probability p/q (and it lies in the interval $(p, 1/t]$ with probability $1 - (t-1)p/q$).

(b) The result for a_0 can also be shown to follow from a path counting result (Barbier's theorem): the number of paths with step set $\{(0,1),(1,0)\}$ from $(0,0)$ to (k,n) with $n > tk$ that never touch the line $y = tx$ except at $(0,0)$ is equal to $(n - tk)\binom{n+k}{n}/(n+k)$.

5. Distributional results for Pólya's urn

Returning to Pólya's urn, recall that $S_{r,b} = \sup_{n\ge 0} Z_n = \max\{r/(r+b), M_{r,b}\}$ and let $I_{r,b} = \inf_{n\ge 0} Z_n = \min\{r/(r+b), L_{r,b}\}$. We now present some equations for the distribution function of $S_{r,b}$ and $I_{r,b}$ for special values.

We introduce the generalized harmonic number function

$$H(x) = \sum_{n\ge 1}\left(\frac{1}{n} - \frac{1}{n+x}\right) = \Psi(x+1) + \gamma, \qquad x \in \mathbb{R}\backslash\{-1,-2,-3,\dots,\},$$

where Ψ is the Digamma function (the logarithmic derivative of the Γ-function) and γ is Euler's constant. For positive integers p, q with $p < q$ Gauss has shown (see, e.g. [5, Problem 19, Section 1.2.9]) that

$$H(p/q) = \frac{q}{p} - \frac{\pi}{2}\cot\left(\frac{p}{q}\pi\right) - \ln(2q) + 2\sum_{1\le n<q/2}\cos\left(\frac{2pn}{q}\pi\right)\ln\sin\left(\frac{n}{q}\pi\right).$$

Thus, $H(p/q)$ can be expressed in terms of finitely many elementary functions.

Let us first consider the distribution function of $S_{r,b}$ at $(t-1)/t$. The obvious route to the results is to condition on Z, use the equations for the binomial random walk and integrate the power series of T_t^m term by term, using the Beta integrals. The general solution is given in the following proposition. Note that in the case $d > 1$ we have to determine the probabilities $a_{m,d} = \mathbb{P}\{tdS_n - nds \le m \text{ for all } n \in \mathbb{N}\}$. Since, clearly, $a_{m,d} = a_{\lfloor m/d\rfloor}$, (for $r \in \mathbb{R}$, $\lfloor r\rfloor$ denotes the largest integer not exceeding r) this can be reduced to the computation of the a_m, i.e. the $d = 1$ case, which we dealt with in Section 4.

Proposition 5.1. *Let $m = b(t-1) - r \ge 0$ and $a = \lfloor m/d\rfloor$. Then*

$$\mathbb{P}\left\{S_{r,b} > \frac{t-1}{t}\right\} = \sum_{n=0}^{\infty}\frac{a+1}{n(t-1)+a+1}\binom{nt+a}{n}\frac{B(n(t-1)+a+1+(r/d), n+(b/d))}{B(r/d, b/d)}.$$

For $d = 1$ we obtain a series of rational functions of n, which can (in principle) be evaluated in closed form in terms of the Digamma function Ψ. We give only two examples for the results of these calculations.

Proposition 5.2. *For $d = 1$ and $2 \le t \in \mathbb{N}$ we have*

$$\mathbb{P}\left\{S_{1,1} = \frac{t-1}{t}\right\} = \frac{2t-3}{t}H\left(1-\frac{1}{t}\right) - \frac{t-2}{t}H\left(1-\frac{2}{t}\right) - \frac{t-2}{t-1},$$

$$\mathbb{P}\left\{S_{1,1} \le \frac{t-1}{t}\right\} = \left(1-\frac{1}{t}\right)H\left(1-\frac{1}{t}\right).$$

Essentially, this has already been shown by Knuth [7, Problem 88].

https://doi.org/10.1239/jap/1429282614 Published online by Cambridge University Press

The situation for $b = 1, r = t - 1$ is of special interest because in this case $p_+(t) = \mathbb{P}\{I_{t-1,1} \geq (t-1)/t\}$ (the probability that the urn content process stays above the line $y = (t-1)x$) as well as $p_-(t) = \mathbb{P}\{S_{t-1,1} \leq (t-1)/t\}$ (the probability that it stays below the line $y = (t-1)x$) can both be considered.

Let $q_-(t) = 1 - p_-(t) = \mathbb{P}\{S_{t-1,1} > (t-1)/t\}$ in the following proposition.

Proposition 5.3. *For $d = 1$,*

$$q_-(t) = (t-1)\sum_{n=0}^{\infty} \frac{(nt)!}{(nt-n+1)!}\frac{((n+1)(t-1))!}{((n+1)t)!}.$$

Some special values are

- $q_-(2) = \ln 2 \approx 0.6931$,
- $q_-(3) = \frac{4}{27}\pi\sqrt{3} \approx 0.8061$,
- $q_-(4) = \frac{9}{32}\ln 2 + \frac{27}{128}\pi \approx 0.8576$,
- $q_-(5) \approx 0.8874$, $q_-(6) \approx 0.9068, \ldots, q_-(20) \approx 0.9726$.

Asymptotically, we obtain $\lim_{t\to\infty} p_-(t) = 0$, since, clearly, $q_-(t) \geq 1 - (1/t)$ (the urn content process steps from $(t-1, 1)$ to $(t, 1)$ with probability $1 - (1/t)$).

Finally, we look at the argument value $1/t$. In this case we have obtained the representation

$$\mathbb{P}\left\{S_{1,t-1} \leq \frac{1}{t}\right\} = (t-1)\int_0^{1/t} (1-pt)q^{t-3}\,\mathrm{d}p,$$

which can be evaluated elementarily. The result is given in the following proposition.

Proposition 5.4. *For $d = 1$,*

$$\mathbb{P}\left\{I_{t-1,1} \geq \frac{t-1}{t}\right\} = \mathbb{P}\left\{S_{1,t-1} \leq \frac{1}{t}\right\} = \begin{cases} 1 - \ln 2 & \text{for } t = 2, \\ \dfrac{[1-1/t]^{t-2}(t-1)-1}{t-2} & \text{for } t > 2. \end{cases}$$

In particular, $p_+(t) \to \mathrm{e}^{-1}$ as $t \to \infty$.

It is also interesting to consider the other start positions $(a(t-1), a)$ $(2 \leq a \in \mathbb{N})$ on the line $y = (t-1)x$. Here we obtain (again for $d = 1$)

$$\mathbb{P}\left\{I_{a(t-1),a} \geq \frac{t-1}{t}\right\} = \mathbb{P}\left\{S_{a,a(t-1)} \leq \frac{1}{t}\right\} = \int_0^{1/t} (1-pt)q^{-1}\beta_{a,a(t-1)}(p)\,\mathrm{d}p.$$

A short calculation yields that the latter integral can be expressed in terms of the binomial distribution.

Proposition 5.5. *For $d = 1$ and $2 \leq a \in \mathbb{N}$,*

$$\mathbb{P}\left\{S_{a,a(t-1)} \leq \frac{1}{t}\right\} = \mathbb{P}\{X_{at-1,1/t} = a\} - \frac{1}{a(t-1)-1}\mathbb{P}\{X_{at-1,1/t} > a\},$$

https://doi.org/10.1239/jap/1429282614 Published online by Cambridge University Press

where $X_{n,p}$ is a random variable having the binomial distribution with parameters n and p. In particular,

$$\mathbb{P}\{S_{a,a} \le \tfrac{1}{2}\} = \left(1 + \frac{1}{a-1}\right)\binom{2a-1}{a}2^{-(2a-1)} - \frac{1}{2(a-1)}$$

and

$$\mathbb{P}\left\{S_{a(t-1),a} \le \frac{1}{t}\right\} = O(a^{-1/2}) \quad \text{as } a \to \infty.$$

In closing, we remark that for $t = 2$ the 'equalization probability' $\mathbb{P}\{S_{r,b} \ge \frac{1}{2}\}$ was already studied in [1] and [10], where results equivalent to the ones above were obtained in this special case. In particular, for $t = 2, m = b - r > 0$ we obtain from (4.1),

$$\mathbb{P}\{S_{r,b} \ge \tfrac{1}{2}\} = \int_0^1 \min\left(1, \frac{p}{q}\right)^{b-r} \beta_{r,b}(p)\,\mathrm{d}p = 2\int_0^{1/2} \beta_{b,r}(p)\,\mathrm{d}p,$$

yielding the identity

$$\mathbb{P}\{S_{r,b} \ge \tfrac{1}{2}\} = 2\mathbb{P}\{X_{b+r-1,1/2} \le r-1\},$$

which was shown in a different way in [10].

Acknowledgement

We would like to thank the referee for his/her very careful reading of the manuscript and for helpful comments.

References

[1] Antal, T., Ben-Naim, E. and Krapivsky, P. L. (2010). First-passage properties of the Pólya urn process. *J. Statist. Mech. Theory Exp.* **2010,** P07009.
[2] Blackwell, D. and Kendall, D. (1964). The Martin boundary for Pólya's urn scheme, and an application to stochastic population growth. *J. Appl. Prob.* **1,** 284–296.
[3] Eggenberger, F. and Pólya, G. (1923). Über die Statistik verketteter Vorgänge. *ZAMM* **3,** 279–289.
[4] Freedman, D. A. (1965). Bernard Friedman's urn. *Ann. Math. Statist.* **36,** 956–970.
[5] Knuth, D. E. (1997). *The Art of Computer Programming*, Vol. 1, *Fundamental Algorithms*, 3rd edn. Addison-Wesley, Reading, MA.
[6] Knuth, D. E. (2011). *The Art of Computer Programming*, Vol. 4a, Part 1, *Combinatorial Algorithms*. Addison-Wesley, Upper Saddle River, NJ.
[7] Knuth, D. E. (2013). *The Art of Computer Programming*, Vol. 4, Pre-fascicle 5a, *Mathematical Preliminaries Redux*. Available at http://www-cs-faculty.stanfor.edu/~uno/fasc5a.ps.gz.
[8] Mahmoud, H. M. (2009). *Pólya Urn Models*. CRC Press, Boca Raton, FL.
[9] Stadje, W. (2008). The maximum average gain in a sequence of Bernoulli games. *Amer. Math. Monthly* **115,** 902–910.
[10] Wallstrom, T. C. (2012). The equalization probability of the Pólya urn. *Amer. Math. Monthly* **119,** 516–518.