THE EGGENBERGER-PÓLYA URN PROCESS: PROBABILITIES OF REVISITED BALL RATIOS

RICHARD J. MATHAR

ABSTRACT. The Eggenberger-Pólya urn process starts with an urn that contains balls of two colors. At each step a ball in the urn is randomly chosen and a ball of the same color added to the urn. The probabilities of finding specified numbers of balls of the two colors later on can be visualized as a non-isotropic walk of U(p) and R(ight) steps on a square lattice. We discuss some numerical aspects of the probability that a ratio of the ball numbers of the two colors reappears later on during the process.

1. Urn model

The Eggenberger-Pólya urn model starts with an urn initially filled with m balls of one color and n balls of another color [2]. At each step one of the $n+m$ balls in the urn is selected randomly with equal probability, and a ball of the same color is added to the urn. The number of balls increases by one at each step; since the probability to pick the majority color is higher (if $m \neq n$), the probabilities have a tendency to keep the majority color—the bias known as 'the rich get richer.'

2. Walk model

One can visualize the contents of the urn by a point at coordinate (m, n) in a square grid given the initial numbers of the two colors [11].

The states of the urn are a random walk on the upper right quadrant of a square grid with anisotropic probability of $m/(m+n)$ of a step to the right (increasing m by 1) or probability $n/(n+m)$ for a step up (increasing n by 1). Figure 1 shows two examples of walks that start with 5 balls of the majority color and 3 balls of the minority color.

The number of walks from (m, n) to (m', n') with $t = m' + n' - m - n$ steps is $\binom{t}{m'-m}$ —A result of the stars and bars considering of where the horizontal steps are placed. For each of these walks one needs to build the product of these probabilities along the intermediate points, assigning a probability to each edge.

Definition 1. *(Transition probability)* $P(m, n \to m', n')$ *is the probability that a walk starting at position* (m, n) *on the square grid passes through* (m', n')*,* $m' \geq m$*,* $n' \geq n$*.*

Date: February 14, 2025.
2020 *Mathematics Subject Classification.* Primary 60C05; Secondary 60G50.
Key words and phrases. Pólya Urn, Probability.

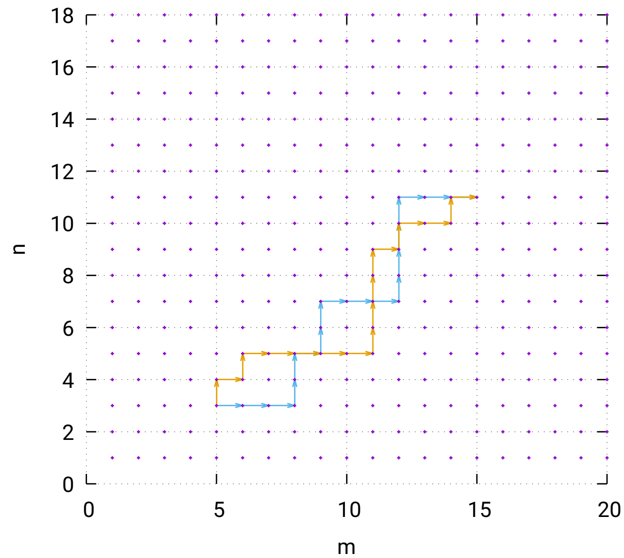

FIGURE 1. Two different—occasionally crossing—UR walks on the square grid from $(m, n) = (5, 3)$ to $(15, 11)$.

This can for example be calculated recursively from the most recent possibilities of visiting a point with a U or an R step:

(1)
$$P(m, n \to m', n') = \frac{m'-1}{m'+n'-1} P(m, n \to m'-1, n') + \frac{n'-1}{m'+n'-1} P(m, n \to m', n'-1)$$

with the boundary conditions

$$P(m, n \to m', n') = \begin{cases} 0, & \text{if } m' < m \vee n' < n, \\ 1, & \text{if } m' = m \wedge n' = n. \end{cases} \quad (2)$$

All walks from (m, n) to (m', n') they all pick up the denominators $(m+n)(m+n+1)\cdots(m'+n'-1)$ [with $m'-m+n'-n$ factors] and the numerators $m(m+1)\cdots(m'-1)$ at the horizontal steps and the numerators $n(n+1)\cdots(n'-1)$ at the vertical steps in varying permutations. The number of the permutations is $\binom{m'+n'-m-n}{m'-m}$ [14][7, (10.3)].

(3)
$$P(m, n \to m', n') = \binom{m'+n'-m-n}{m'-m} \frac{m(m+1)\cdots(m'-1)n(n+1)\cdots(n'-1)}{(m+n)(m+n+1)\cdots(m'+n'-1)}$$
$$= \binom{\Delta m + \Delta n}{\Delta m} \frac{m(m+1)\cdots(m+\Delta m-1)n(n+1)\cdots(n+\Delta n-1)}{(m+n)(m+n+1)\cdots(m+\Delta m+n+\Delta n-1)}.$$

$\Delta m = m' - m$ and $\Delta n = n' - n$ are the horizontal and vertical distances of points during a walk.

Example 1. *The probability in Figure 1 that a walk starting at* $(5, 3)$ *passes through* $(15, 11)$ *is* $P(5, 3 \to 15, 11) = 819/8740$.

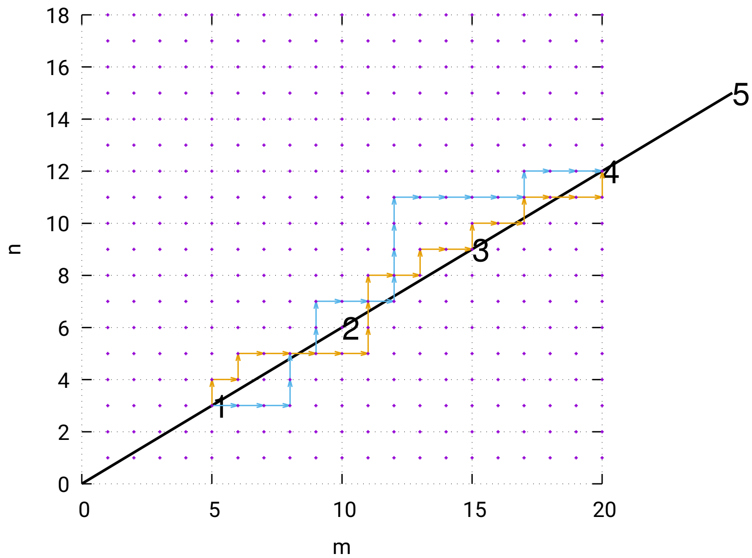

FIGURE 2. A reference line with slope $\bar{n}/\bar{m} = 3/5$ and the labels 1, 2, ... for the lattice points on that line. Two walks from $i = 1$ to $j = 4$ are sequences of arrows, both missing $k = 2$, the brown walk passing through $k = 3$ and the blue walk missing it.

3. RECURRING BALL RATIO

The interest of this manuscript is a discussion of chances that the initial ball ratio m/n is found again later on the walk, registering/triggering on events of visiting points (m', n') that have a common angle towards the horizontal, $m'/n' = m/n$.

Example 2. *If $(m, n) = (15, 3)$, the lattice points with the same ratio are $(20, 4)$, $(25, 5)$, $(30, 6)$, $(35, 7)$ and so on, which are reachable after $t = 6, 12, 18, 24, \ldots$ steps, the multiples of 6.*

If $d = \gcd(m, n)$ is the greatest common divisor of m and n, and $m = d\bar{m}$ and $n = d\bar{n}$ with coprime $\bar{m}$ and $\bar{n}$, distances Δm, Δn that preserve the ratio obey

$$\frac{m + \Delta m}{n + \Delta n} = \frac{m}{n} = \frac{\bar{m}}{\bar{n}},$$

The projected distances are therefore

$$\frac{\Delta m}{\Delta n} = \frac{m}{n} = \frac{\bar{m}}{\bar{n}}. \tag{4}$$

A fixed coprime pair $\gcd(\bar{m}, \bar{n}) = 1$ defines a reference line through the origin for sets of points of fixed ratio m/n, $m = i\bar{m}$, $n = i\bar{n}$, $m' = j\bar{m}$, $n' = j\bar{n}$. A simplified notation with one-dimensional indices i and j suffices to specify the start and end vertex of the walk on such a line if we are only concerned with the events that the walk starting at i later passes through j or misses it. Figure 2 shows such a bold reference line and a labeling of these one-dimensional indices with a larger font. With that 1-dimensional indexing (3) is

$$
\begin{aligned}
(5)\quad P(i\to j) &= P(i\bar m, i\bar n \to j\bar m, j\bar n)\\
&= \binom{(j-i)(\bar m+\bar n)}{(j-i)\bar m}\frac{\Gamma(m+\Delta m)n(n+1)\cdots(n+\Delta n-1)}{\Gamma(m)(m+n)(m+n+1)\cdots(m+\Delta m+n+\Delta n-1)}\\
&= \binom{(j-i)(\bar m+\bar n)}{(j-i)\bar m}\frac{\Gamma(m+\Delta m)\Gamma(n+\Delta n)}{\Gamma(m)\Gamma(n)(m+n)(m+n+1)\cdots(m+\Delta m+n+\Delta n-1)}\\
&= \binom{(j-i)(\bar m+\bar n)}{(j-i)\bar m}\frac{\Gamma(m+\Delta m)\Gamma(n+\Delta n)\Gamma(m+n)}{\Gamma(m)\Gamma(n)\Gamma(m+\Delta m+n+\Delta n)}\\
&= \binom{(j-i)(\bar m+\bar n)}{(j-i)\bar m}\frac{\Gamma(j\bar m)\Gamma(j\bar n)\Gamma(i(\bar m+\bar n))}{\Gamma(i\bar m)\Gamma(i\bar n)\Gamma(j(\bar m+\bar n))}.
\end{aligned}
$$

With the notation

Definition 2. *(Beta-function)* [6, 8.384.1][1, 6.2.2]

$$B(x,y)\equiv\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \tag{6}$$

the probabilities are a β-distribution

$$P(i\to j)=\binom{(j-i)(\bar m+\bar n)}{(j-i)\bar m}\frac{B(j\bar m,j\bar n)}{B(i\bar m,i\bar n)}. \tag{7}$$

Example 3. *The probability in Figure 2 that a walk starting at* $(5,3)$ *does* not *pass through* $(10,6)$ *is with* $(\bar m,\bar n)=(5,3)$ *equal to* $1-P(1\to 2)=115/143$ *.*

Definition 3. *(Generating function of revisiting probabilities) The probabilities that the walk starting at i visits j defines a generating function concerned with the coarse label distance $j-i$ on the reference line:*

$$\hat p_{i,\bar m,\bar n}(z)=\sum_{j\ge i}P(i\to j)z^{j-i}. \tag{8}$$

Inserting (5), the embedded Γ-functions are converted to Pochhammer-Symbols [10][1, 6.1.22]

$$
\begin{aligned}
(9)\quad \hat p &= \sum_{j\ge i}\binom{(j-i)(\bar m+\bar n)}{(j-i)\bar m}\frac{\Gamma(j\bar m)\Gamma(j\bar n)\Gamma(i(\bar m+\bar n))}{\Gamma(i\bar m)\Gamma(i\bar n)\Gamma(j(\bar m+\bar n))}z^{j-i}\\
&= \frac{\Gamma(i(\bar m+\bar n))}{\Gamma(i\bar m)\Gamma(i\bar n)}\sum_{j\ge i}\binom{(j-i)(\bar m+\bar n)}{(j-i)\bar m}\frac{\Gamma((j-i+i)\bar m)\Gamma((j-i+i)\bar n)}{\Gamma((j-i+i)(\bar m+\bar n))}z^{j-i}\\
&= \frac{\Gamma(i(\bar m+\bar n))}{\Gamma(i\bar m)\Gamma(i\bar n)}\sum_{l\ge 0}\binom{l(\bar m+\bar n)}{l\bar m}\frac{\Gamma((l+i)\bar m)\Gamma((l+i)\bar n)}{\Gamma((l+i)(\bar m+\bar n))}z^{l}\\
&= \frac{\Gamma(i(\bar m+\bar n))}{\Gamma(i\bar m)\Gamma(i\bar n)}\sum_{l\ge 0}\binom{l(\bar m+\bar n)}{l\bar m}\frac{\Gamma(l\bar m+i\bar m)\Gamma(l\bar n+i\bar n)}{\Gamma(l(\bar m+\bar n)+i(\bar m+\bar n))}z^{l}\\
&= \frac{\Gamma(i(\bar m+\bar n))}{\Gamma(i\bar m)\Gamma(i\bar n)}\sum_{l\ge 0}\frac{(1)_{l(\bar m+\bar n)}}{(1)_{l\bar m}(1)_{l\bar n}}\frac{\Gamma(l\bar m+i\bar m)\Gamma(l\bar n+i\bar n)}{\Gamma(l(\bar m+\bar n)+i(\bar m+\bar n))}z^{l}\\
&= \sum_{l\ge 0}\frac{(1)_{l(\bar m+\bar n)}}{(1)_{l\bar m}(1)_{l\bar n}}\frac{(i\bar m)_{l\bar m}(i\bar n)_{l\bar n}}{(i(\bar m+\bar n))_{l(\bar m+\bar n)}}z^{l};
\end{aligned}
$$

the multiplication formula of the Pochhammer Symbols is applied to all 6 Pochhammer symbols [12, I.26], common factors in upper and and lower parameters are removed. The generating functions are generalized hypergeometric functions:

$$
\begin{aligned}
(10)\quad \hat{p}_{i,\bar m,\bar n}(z) &= \sum_{l\ge 0} \frac{(\frac{1}{\bar m+\bar n})_l(\frac{2}{\bar m+\bar n})_l\cdots(\frac{\bar m+\bar n}{\bar m+\bar n})_l(\bar m+\bar n)^{l(\bar m+\bar n)}}{(\frac{1}{\bar m})_l(\frac{2}{\bar m})_l\cdots(\frac{\bar m}{\bar m})_l(\bar m)^{l\bar m}(\frac{1}{\bar n})_l(\frac{2}{\bar n})_l\cdots(\frac{\bar n}{\bar n})_l(\bar n)^{l\bar n}} \\
&\times \frac{(\frac{i\bar m}{\bar m})_l(\frac{1+i\bar m}{\bar m})_l\cdots(\frac{i\bar m+\bar m-1}{\bar m})_l\bar m^{l\bar m}(\frac{i\bar n}{\bar n})_l(\frac{1+i\bar n}{\bar n})_l\cdots(\frac{i\bar n+\bar n-1}{\bar n})_l\bar n^{l\bar n}}{(\frac{i(\bar m+\bar n)}{\bar m+\bar n})_l(\frac{1+i(\bar m+\bar n)}{\bar m+\bar n})_l\cdots(\frac{i(\bar m+\bar n)+\bar m+\bar n-1}{\bar m+\bar n})_l(\bar m+\bar n)^{l(\bar m+\bar n)}} z^l \\
&= \sum_{l\ge 0} \frac{(\frac{1}{\bar m+\bar n})_l(\frac{2}{\bar m+\bar n})_l\cdots(\frac{\bar m+\bar n}{\bar m+\bar n})_l}{(\frac{1}{\bar m})_l(\frac{2}{\bar m})_l\cdots(\frac{\bar m}{\bar m})_l(\bar m)^{l\bar m}(\frac{1}{\bar n})_l(\frac{2}{\bar n})_l\cdots(\frac{\bar n}{\bar n})_l(\bar n)^{l\bar n}} \\
&\times \frac{(\frac{i\bar m}{\bar m})_l(\frac{1+i\bar m}{\bar m})_l\cdots(\frac{i\bar m+\bar m-1}{\bar m})_l\bar m^{l\bar m}(\frac{i\bar n}{\bar n})_l(\frac{1+i\bar n}{\bar n})_l\cdots(\frac{i\bar n+\bar n-1}{\bar n})_l\bar n^{l\bar n}}{(\frac{i(\bar m+\bar n)}{\bar m+\bar n})_l(\frac{1+i(\bar m+\bar n)}{\bar m+\bar n})_l\cdots(\frac{i(\bar m+\bar n)+\bar m+\bar n-1}{\bar m+\bar n})_l} z^l \\
&= \sum_{l\ge 0} \frac{(\frac{1}{\bar m+\bar n})_l(\frac{2}{\bar m+\bar n})_l\cdots(\frac{\bar m+\bar n}{\bar m+\bar n})_l}{(\frac{1}{\bar m})_l(\frac{2}{\bar m})_l\cdots(\frac{\bar m}{\bar m})_l(\frac{1}{\bar n})_l(\frac{2}{\bar n})_l\cdots(\frac{\bar n}{\bar n})_l} \\
&\times \frac{(\frac{i\bar m}{\bar m})_l(\frac{1+i\bar m}{\bar m})_l\cdots(\frac{i\bar m+\bar m-1}{\bar m})_l(\frac{i\bar n}{\bar n})_l(\frac{1+i\bar n}{\bar n})_l\cdots(\frac{i\bar n+\bar n-1}{\bar n})_l}{(\frac{i(\bar m+\bar n)}{\bar m+\bar n})_l(\frac{1+i(\bar m+\bar n)}{\bar m+\bar n})_l\cdots(\frac{i(\bar m+\bar n)+\bar m+\bar n-1}{\bar m+\bar n})_l} z^l \\
&= \sum_{l\ge 0} \frac{(\frac{1}{\bar m+\bar n})_l(\frac{2}{\bar m+\bar n})_l\cdots(\frac{\bar m+\bar n-1}{\bar m+\bar n})_l}{(\frac{1}{\bar m})_l(\frac{2}{\bar m})_l\cdots(\frac{\bar m-1}{\bar m})_l(\frac{1}{\bar n})_l(\frac{2}{\bar n})_l\cdots(\frac{\bar n-1}{\bar n})_l} \\
&\times \frac{(\frac{i\bar m}{\bar m})_l(\frac{1+i\bar m}{\bar m})_l\cdots(\frac{i\bar m+\bar m-1}{\bar m})_l(\frac{i\bar n}{\bar n})_l(\frac{1+i\bar n}{\bar n})_l\cdots(\frac{i\bar n+\bar n-1}{\bar n})_l}{(\frac{i(\bar m+\bar n)}{\bar m+\bar n})_l(\frac{1+i(\bar m+\bar n)}{\bar m+\bar n})_l\cdots(\frac{i(\bar m+\bar n)+\bar m+\bar n-1}{\bar m+\bar n})_l} \frac{z^l}{l!} \\
&= {}_{2(\bar m+\bar n)-1}F_{2(\bar m+\bar n)-2}\Big(\\
&\begin{matrix} \frac{1}{\bar m+\bar n},\frac{2}{\bar m+\bar n},\cdots,\frac{\bar m+\bar n-1}{\bar m+\bar n},i,i+\frac{1}{\bar m},i+\frac{2}{\bar m},\cdots,i+\frac{\bar m-1}{\bar m},i,i+\frac{1}{\bar n},i+\frac{2}{\bar n},\cdots,i+\frac{\bar n-1}{\bar n} \\ \frac{1}{\bar m},\frac{2}{\bar m},\cdots,\frac{\bar m-1}{\bar m},\frac{1}{\bar n},\frac{2}{\bar n},\cdots,\frac{\bar n-1}{\bar n},i,i+\frac{1}{\bar m+\bar n},i+\frac{2}{\bar m+\bar n},\cdots,i+\frac{\bar m+\bar n-1}{\bar m+\bar n}\end{matrix} \mid z\Big) \\
&= {}_{2(\bar m+\bar n)-2}F_{2(\bar m+\bar n)-3}\Big(\\
&\begin{matrix} \frac{1}{\bar m+\bar n},\frac{2}{\bar m+\bar n},\cdots,\frac{\bar m+\bar n-1}{\bar m+\bar n},i,i+\frac{1}{\bar m},i+\frac{2}{\bar m},\cdots,i+\frac{\bar m-1}{\bar m},i+\frac{1}{\bar n},i+\frac{2}{\bar n},\cdots,i+\frac{\bar n-1}{\bar n} \\ \frac{1}{\bar m},\frac{2}{\bar m},\cdots,\frac{\bar m-1}{\bar m},\frac{1}{\bar n},\frac{2}{\bar n},\cdots,\frac{\bar n-1}{\bar n},i+\frac{1}{\bar m+\bar n},i+\frac{2}{\bar m+\bar n},\cdots,i+\frac{\bar m+\bar n-1}{\bar m+\bar n}\end{matrix} \mid z\Big).
\end{aligned}
$$

Remark 1. *There are $\bar m+\bar n-1$ upper parameters in the hypergeometric function and $\bar m+\bar n-1$ lower parameters that depend on i. The generic contiguous equations permit to decrement upper or lower parameters to find connection coefficients for variable indices i [5, 8].*

The excess of this series (sum of the $2(\bar m+\bar n-1)$ upper parameters minus sum of the $2(\bar m+\bar n)-2$ lower parameters) is zero:

$$
(11)\quad \sum_{j=1}^{\bar m+\bar n-1}\frac{j}{\bar m+\bar n}+\sum_{j=0}^{\bar m-1}(i+\frac{j}{\bar m})+\sum_{j=1}^{\bar n-1}(i+\frac{j}{\bar n}) = \sum_{j=1}^{\bar m-1}\frac{j}{\bar m}+\sum_{j=1}^{\bar n-1}\frac{j}{\bar n}+\sum_{j=1}^{\bar n+\bar m-1}(i+\frac{j}{\bar m+\bar n}).
$$

As a side effect, the series (10) does not converge at $z=1$ [12, §2.2], which means that the sum of the probabilities $\sum_{j>i}P(i\to j)$ is ∞.

$\bar m$	$\bar n$	$f_{\bar m,\bar n}(z)$	[3]
1	1	$1+2z+6z^2+20z^3+70z^4+252z^5+\cdots$	A000984
1	2	$1+3z+15z^2+84z^3+495z^4+3003z^5+\cdots$	A005809
1	3	$1+4z+28z^2+220z^3+1820z^4+\cdots$	A005810
1	4	$1+5z+45z^2+455z^3+4845z^4+\cdots$	A001449
1	5	$1+6z+66z^2+816z^3+10626z^4+\cdots$	A004355
2	2	$1+6z+70z^2+924z^3+12870z^4+\cdots$	A001448
2	3	$1+10z+210z^2+5005z^3+125970z^4+\cdots$	A001450
3	3	$1+20z+924z^2+48620z^3+2704156z^4+\cdots$	A066802

Table 1. Examples of the binomial series (12). If $\bar m$ and $\bar n$ are not coprime, the coefficients are a multisection of the sequence for the smaller coprime series.

Remark 2. *As always for hypergeometric functions they are P-finite with 2 terms, i.e., $P(i\to j)/P(i\to j-1)$ is a rational polynomial of $j-i$; This is already obvious by building ratios of terms of (5).*

In $P(i \to j)$ in (7) let the binomial factor (that counts the paths) be called the multiplicity and the second factor (ratio of two B-functions) be called the drift factor.

Define

Definition 4. *(Generating function of the multiplicities)*

$$f_{\bar m,\bar n}(z) \equiv \sum_{i\ge 0} \binom{i(\bar m+\bar n)}{i\bar m} z^i \tag{12}$$

Translating all Γ-functions in that binomial formula to Pochhammer symbols [10] and using the reductions of these [12] shows that the $f(z)$ are generalized hypergeometric series:

$$f_{\bar m,\bar n}(z) = {}_{\bar m+\bar n-1}F_{\bar m+\bar n-2}\left(\begin{matrix} \frac{1}{\bar m+\bar n}, \frac{2}{\bar m+\bar n}, \ldots, \frac{\bar m+\bar n-1}{\bar m+\bar n} \\ \frac{1}{\bar m}, \frac{2}{\bar m}, \ldots, \frac{\bar m-1}{\bar m}, \frac{1}{\bar n}, \frac{2}{\bar n}, \ldots, \frac{\bar n-1}{\bar n} \end{matrix} \mid \frac{(\bar m+\bar n)^{\bar m+\bar n}}{\bar m^{\bar m}\bar n^{\bar n}} z \right). \tag{13}$$

Examples of these are in the Online Encyclopedia of Integer Sequences: Table 1.

4. First return to original ball ratio

There are walks that meet multiple points in the square lattice on the line of the reference angle. To consider the probability of meeting the original ratio for the first time, one must filter these probabilities with the inclusion-exclusion principle to avoid over-counting those walks that achieve the ratio more than once.

Definition 5. *We define $Q_{\bar m,\bar n}(i\to j)$ as the probability of starting the walk at $(i\bar m, i\bar n)$ and ending the walk at $(j\bar m, j\bar n)$* avoiding *all intermediate vertices $(k\bar m, k\bar n)$, $i<k<j$. This is the probability of a* first return *to the line of constant slope.*

The key aspect in our paper is: It can be derived from the probability $P(i \to j)$ by subtracting the probabilities that the walk has passed through the vertices k

$\bar{m}$	$\bar{n}$	i	j	$P(i \to j)$	$P(i \to j)$
1	1	1	2	1/3	0.3333333
1	1	1	3	1/5	0.2000000
1	1	1	4	1/7	0.1428571
1	1	2	3	2/5	0.4000000
1	1	2	4	9/35	0.2571429
1	1	2	5	4/21	0.1904762
1	1	3	4	3/7	0.4285714
1	1	3	5	2/7	0.2857143
1	1	3	6	50/231	0.2164502
2	1	1	2	3/10	0.3000000
2	1	1	3	5/28	0.1785714
2	1	1	4	7/55	0.1272727
2	1	2	3	5/14	0.3571429
2	1	2	4	5/22	0.2272727
2	1	2	5	24/143	0.1678322
2	1	3	4	21/55	0.3818182
2	1	3	5	36/143	0.2517483
2	1	3	6	42/221	0.1900452
3	1	1	2	2/7	0.2857143
3	1	1	3	28/165	0.1696970
3	1	1	4	11/91	0.1208791
3	1	2	3	56/165	0.3393939
3	1	2	4	14/65	0.2153846
3	1	2	5	154/969	0.1589267
3	1	3	4	33/91	0.3626374
3	1	3	5	77/323	0.2383901
3	1	3	6	550/3059	0.1797973
3	2	1	2	5/21	0.2380952
3	2	1	3	20/143	0.1398601
3	2	1	4	385/3876	0.0993292
3	2	2	3	40/143	0.2797203
3	2	2	4	735/4199	0.1750417
3	2	2	5	1911/14858	0.1286176
3	2	3	4	385/1292	0.2979876
3	2	3	5	5733/29716	0.1929264
3	2	3	6	11011/76038	0.1448092

TABLE 2. Revisiting probabilities $P(i \to j)$ for basic values of ratios $\bar{n}/\bar{m}$ and start indices i.

just before visiting j [4, (9)]:

$$Q_{\bar{m},\bar{n}}(i \to j) = P(i \to j) - \sum_{k=i+1}^{j-1} P(i \to k) Q_{\bar{m},\bar{n}}(k \to j), \tag{14}$$

where empty sums (upper limit smaller than lower limit) are zero.

Rephrasing this equation as

$$Q_{\bar m,\bar n}(i\to j)+\sum_{k=i+1}^{j-1}P(i\to k)Q_{\bar m,\bar n}(k\to j)=P(i\to j), \tag{15}$$

this is a linear system of equations with a lower triangular matrix of P-values,

$$\begin{pmatrix} 1 & 0 & 0 & 0 & \cdots & 0\\ P(j-2\to j-1) & 1 & 0 & 0 & \cdots & 0\\ P(j-3\to j-1) & P(j-3\to j-2) & 1 & 0 & \cdots & 0\\ \ldots & \ldots & \ldots & \ldots & \ddots & 0\\ P(i\to j-1) & P(i\to j-2) & P(i\to j-3) & P(i\to j-4) & \cdots & 1 \end{pmatrix} \cdot\begin{pmatrix} Q(j-1\to j)\\ Q(j-2\to j)\\ Q(j-3\to j)\\ \ldots\\ Q(i\to j)\end{pmatrix}=\begin{pmatrix} P(j-1\to j)\\ P(j-2\to j)\\ P(j-3\to j)\\ \ldots\\ P(i\to j)\end{pmatrix}. \tag{16}$$

Solving this system is equivalent to evaluating (14) recursively for $i=j$, $i=j-1$ etc for decreasing i starting at

$$Q_{\bar m,\bar n}(j\to j)=1,\quad Q_{\bar m,\bar n}(j-1\to j)=P(j-1\to j). \tag{17}$$

Alternatively the walks can be split into cases according to the *first* revisiting step on the reference line:

$$P(i\to j)=Q_{\bar m,\bar n}(i\to j)+\sum_{k=i+1}^{j-1}Q_{\bar m,\bar n}(i\to k)P(k\to j), \tag{18}$$

which is the linear system

$$\begin{pmatrix} 1 & P(j-1\to j) & P(j-2\to j) & P(j-3\to j) & \cdots & P(i+1\to j)\\ 0 & 1 & P(j-2\to j-1) & P(j-3\to j-1) & \cdots & P(i+1\to j-1)\\ 0 & 0 & 1 & P(j-3\to j-2) & \cdots & P(i+1\to j-2)\\ \ldots & \ldots & \ldots & \ldots & \ddots & \ldots\\ 0 & 0 & 0 & 0 & \cdots & 1 \end{pmatrix} \cdot\begin{pmatrix} Q(i\to j)\\ Q(i\to j-1)\\ Q(i\to j-2)\\ \ldots\\ Q(i\to i+1)\end{pmatrix}=\begin{pmatrix} P(i\to j)\\ P(i\to j-1)\\ P(i\to j-2)\\ \ldots\\ P(i\to i+1)\end{pmatrix}. \tag{19}$$

Solving this system is equivalent to evaluating (14) recursively for increasing j.

Recursive expansion translates Q into a sum over products of P with a sign determined by the number of intermediate stops, an alternating sign pattern reflecting the principle of inclusion-and-exclusion [9] [13, §3]:

$$Q_{\bar m,\bar n}(i\to j) = P(i\to j) - \sum_{i<k<j} P(i\to k)P(k\to j) + \sum_{i<k<l<j} P(i\to k)P(k\to l)P(l\to j) - \sum_{i<k<l<m<j} P(i\to k)P(k\to l)P(l\to m)P(m\to j) \pm\cdots \tag{20}$$

Remark 3. *The number of terms on the right hand side (RHS) which are products of s P's is the number of compositions of $j-i$ into s parts, which is $\binom{j-i-1}{s-1}$ [3, A007318][13, §1.2]. The total count of P-products on the RHS is $\sum_{s=1}^{j-i-1}\binom{j-i-1}{s-1} = 2^{j-i-1}$, $j-i\ge 1$.*

Remark 4. *In a formal way the RHS can also be written as the matrix inverse (again a lower triangular matrix of dimension $j-i$) of the LHS multiplied by a vector of P-values.*

The interesting point in the signed products of these terms for the compensation in the inclusion-exclusion is that the product of the drift factors is the same— actually the drift factor for the entire $P(i\to j)$— so the products of the multiplicities remain. Now these multiplicities depend only on the *relative* strides and no longer on the explicit start and stop, m, m' and n, n'.

The simplifying feature in the products of the P-factors on the right hand side of (20) is that the product of the drift factors is independent of the number of factors because the Γ-factors for intermediate points cancel.

Example 4. *The illustration for the case of two factors:*

$$\begin{aligned} P(i\to k)P(k\to j) &= \binom{(k-i)(\bar m+\bar n)}{(k-i)\bar m}\frac{B(k\bar m,k\bar n)}{B(i\bar m,i\bar n)} \times \binom{(j-k)(\bar m+\bar n)}{(j-k)\bar m}\frac{B(j\bar m,j\bar n)}{B(k\bar m,k\bar n)} \\ &= \binom{(k-i)(\bar m+\bar n)}{(k-i)\bar m}\binom{(j-k)(\bar m+\bar n)}{(j-k)\bar m}\frac{B(j\bar m,j\bar n)}{B(i\bar m,i\bar n)}, \end{aligned} \tag{21}$$

and the last factor, the Γ-ratio, is independent of k.

$$\begin{aligned} Q(i\to j) = \frac{B(j\bar m,j\bar n)}{B(i\bar m,i\bar n)}\Big[&\binom{(j-i)(\bar m+\bar n)}{(j-i)\bar m} - \sum_{i<k<j}\binom{(k-i)(\bar m+\bar n)}{(k-i)\bar m}\binom{(j-k)(\bar m+\bar n)}{(j-k)\bar m} \\ &+ \sum_{i<k<l<j}\binom{(k-i)(\bar m+\bar n)}{(k-i)\bar m}\binom{(l-k)(\bar m+\bar n)}{(l-k)\bar m}\binom{(j-l)(\bar m+\bar n)}{(j-l)\bar m} \\ &- \sum_{i<k<l<m<j}\binom{(k-i)(\bar m+\bar n)}{(k-i)\bar m}\binom{(l-k)(\bar m+\bar n)}{(l-k)\bar m}\binom{(m-l)(\bar m+\bar n)}{(m-l)\bar m}\binom{(j-m)(\bar m+\bar n)}{(j-m)\bar m}\pm\cdots\Big]. \end{aligned} \tag{22}$$

$\bar m$	$\bar n$	i	j	$Q(i\to j)$	$Q(i\to j)$
1	1	1	3	1/15	0.0666667
1	1	1	4	1/35	0.0285714
1	1	1	5	1/63	0.0158730
1	1	2	4	3/35	0.0857143
1	1	2	5	4/105	0.0380952
1	1	2	6	5/231	0.0216450
1	1	3	5	2/21	0.0952381
1	1	3	6	10/231	0.0432900
1	1	3	7	25/1001	0.0249750
2	1	1	3	1/14	0.0714286
2	1	1	4	7/220	0.0318182
2	1	1	5	18/1001	0.0179820
2	1	2	4	1/11	0.0909091
2	1	2	5	6/143	0.0419580
2	1	2	6	75/3094	0.0242405
2	1	3	5	72/715	0.1006993
2	1	3	6	21/442	0.0475113
2	1	3	7	9/323	0.0278638
3	1	1	3	4/55	0.0727273
3	1	1	4	3/91	0.0329670
3	1	1	5	91/4845	0.0187822
3	1	2	4	6/65	0.0923077
3	1	2	5	14/323	0.0433437
3	1	2	6	364/14421	0.0252410
3	1	3	5	33/323	0.1021672
3	1	3	6	150/3059	0.0490356
3	1	3	7	2/69	0.0289855
3	2	1	3	20/273	0.0732601
3	2	1	4	95/2652	0.0358220
3	2	1	5	183/8602	0.0212741
3	2	2	4	385/4199	0.0916885
3	2	2	5	399/8602	0.0463846
3	2	2	6	244/8671	0.0281398
3	2	3	5	3003/29716	0.1010567
3	2	3	6	209/4002	0.0522239
3	2	3	7	45201/1406036	0.0321478

TABLE 3. Revisiting probabilities $Q(i \to j)$ for basic values of ratios $\bar n/\bar m$ and start indices i. Because $Q(i\to i{+}1) = P(i \to i{+}1)$, the cases $j = i + 1$ are not repeated from Table 2.

The square bracket on the RHS is an alternating sum of binomial products derived from all compositions of $j-i$. The basic algebraic structure of generating functions that produce such alternating sums is

$$\frac{g(z)}{1+g(z)} = g(z) - g(z)^2 + g(z)^3 - g(z)^4 + \dots . \tag{23}$$

$\bar m$	$\bar n$	$\bar f = 2 - 1/f$	[3]
1	1	$1+2z+2z^2+4z^3+10z^4+28z^5+\cdots$	A002420
1	2	$1+3z+6z^2+21z^3+90^4+429z^5+\cdots$	A024485
1	3	$1+4z+12z^2+60z^3+364z^4+\cdots$	A337291
1	4	$1+5z+20z^2+130z^3+1020z^4+\cdots$	A337292
2	2	$1+6z+34z^2+300z^3+3146z^4+\cdots$	A337350
2	3	$1+10z+110z^2+1805z^3+34770z^4+\cdots$	A337351
3	3	$1+20z+524z^2+19660z^3+854380z^4+\cdots$	A337352

TABLE 4. Examples of series inversions of $f_{\bar m,\bar n}$.

In the case of (22) the weak compositions (admitting parts equal to zero, contributions from walks that stay on a lattice point) need to be discarded; the coefficient $[z^0]f(z)$ is eliminated and the applicable alternating sign series is

$$f(z)-[f(z)-1]^2+[f(z)-1]^3-[f(z)-1]^4+\ldots = 1+[f(z)-1]\frac{1}{f(z)} = 2-\frac{1}{f(z)}. \tag{24}$$

Definition 6. *(Inverted binomial series)*

$$\bar f_{\bar m,\bar n}(z) \equiv 2 - \frac{1}{f_{\bar m,\bar n}(z)}. \tag{25}$$

The structure of (22) rewritten with the inverted series is

$$Q(i \to j) = \frac{B(j\bar m, j\bar n)}{B(i\bar m, i\bar n)}[z^{j-i}]\bar f_{\bar m,\bar n}(z). \tag{26}$$

$\bar f$ is a series inversion of f, and the terms can be computed recursively from the terms of f, the convolution $[\bar f_{\bar m,\bar n}(z)-2]f_{\bar m,\bar n}(z) = -1$ [1, 3.6.16][6, 0.313].

Remark 5. *This result can also be obtained from (18) by canceling a common factor $B(j\bar m, j\bar n)/B(i\bar m, i\bar n)$ on both sides, multiplying and summing over z^{j-i}, and converting the convolution on the right hand side into a product of generating functions.*

This is a transit from Table 1 to Table 4.

Conjecture 1. *For $\bar m = 1$ (equivalent to $\bar n = 1$ because $f_{i,\bar m,\bar n} = f_{i,\bar n,\bar m}$ is symmetric) consecutive coefficients $[z^k]$ and $[z^{k-1}]$ of $\bar f_{1,\bar n}$ also obey 2-term P-recurrences:*

$$\prod_{s=0}^{\bar n-1}(\bar n k - s)[z^k]\bar f_{1,\bar n}(z) = (\bar n+1)\prod_{s=2,s\neq \bar n+1}^{\bar n+2}((\bar n+1)k-s)[z^{k-1}]\bar f_{1,\bar n}(z), \quad k>1; \tag{27}$$

$$\binom{\bar n k}{\bar n}[z^k]\bar f_{1,\bar n}(z) = \frac{1}{k-1}\binom{(\bar n+1)k-2}{\bar n+1}[z^{k-1}]\bar f_{1,\bar n}(z), \quad k>1. \tag{28}$$

This does not translate straight into a generalized hypergeometric function for $\bar f_{1,\bar n}(z)$ because the recurrence does not start at $k=1$.

5. Expectation of First return

The expectation value for the number of draws of the urn (resp. steps of the walk) up to the first return on the reference line is is the product of the number of steps $k(\bar m+\bar n)$ needed times their probabilities,

(29)
$$E_{i,\bar m,\bar n} = \sum_{k\ge 1} k(\bar m+\bar n) Q(i\to i+k) = (\bar m+\bar n)\sum_{k\ge 1} k\frac{B((i+k)\bar m,(i+k)\bar n)}{B(i\bar m,i\bar n)}[z^k]\bar f_{\bar m,\bar n}(z).$$

Remark 6. *Extraction of the k-th coefficient of a generating function and multiplication with k may be relayed to the first derivative* [15, §2.2]

$$E_{i,\bar m,\bar n} = \frac{\bar m+\bar n}{B(i\bar m,i\bar n)}\sum_{k\ge 1} B((i+k)\bar m,(i+k)\bar n)[z^{k-1}]\bar f'_{\bar m,\bar n}(z). \tag{30}$$

I have not found a closed form for that expectation value. Numerical experimentation by accumulating the probabilities for $j=i+1$ to $i+\max k$ suggests that the series of the expectation value does *not* converge to finite values: Table 5.

The probability of return to the reference line in the last column seems to have definite limits obtained by extrapolation [16]:

$\bar m$	$\bar n$	i	$\sum_{k=1}^{\infty} Q(i\to i+k)$
1	1	1	≈ 0.5
1	1	2	≈ 0.625
1	1	3	≈ 0.6875
2	1	1	≈ 0.4858
2	1	2	≈ 0.60565
2	1	3	≈ 0.6671
3	1	1	≈ 0.47852

6. Summary

The probability that an initial ratio of $\bar n/\bar m$ of the numbers of two-colored balls in the Eggenberger-Pólya urn model appears for the first time after a finite number of steps is given by (26) in terms of the coefficients of an inverse of a generating function of binomial series. There is strong numerical evidence that the expectation value for the number of steps of recurrence to the initial ratio is infinite.

References

1. Milton Abramowitz and Irene A. Stegun (eds.), *Handbook of mathematical functions*, 9th ed., Dover Publications, New York, 1972. MR 0167642
2. F. Eggenberger and G. Pólya, *über die statistik verketteter vorgänge*, Z. Angew. Math. Mech **3** (1923), no. 4, 279–289.
3. O. E. I. S. Foundation Inc., *The On-Line Encyclopedia Of Integer Sequences*, (2025), https://oeis.org/. MR 3822822
4. Anant P. Godbole, *On hypergeometric and related distributions of order k*, Comm. Stat. - Theory Meth. **19** (1990), no. 4, 1291–1301. MR 1073032
5. J. E. Gottschalk and E. N. Maslen, *Reduction formulae for generalised hypergeometric functions of one variable*, J. Phys. A: Math. Gen. **21** (1988), 1983–1998. MR 0952917
6. I. Gradshteyn and I. M. Ryzhik, *Table of integrals, series, and products*, 8 ed., Elsevier, Amsterdam, 2015. MR 3307944
7. Christian Krattenthaler, *Lattice path enumeration*, arXiv:1503.05930 [math.CO] (2017). MR 3409351

$\bar{m}$	$\bar{n}$	i	$\max k$	$\sum_{k=1}^{\max k} k(\bar{m}+\bar{n})Q(i \to i+k)$	$\sum_{k=1}^{\max k} Q(i \to i+k)$
1	1	1	20	1.9918683323157545	0.4878048780487805
1	1	1	40	2.3323805989835014	0.4938271604938272
1	1	1	80	2.6758771755290954	0.4968944099378882
1	1	1	160	3.0209003668228994	0.4984423676012461
1	1	1	320	3.3666957455590190	0.4992199687987520
1	1	2	20	2.6304570645541763	0.6069200226885990
1	1	2	40	3.1326628217238018	0.6157965194109772
1	1	2	80	3.6434312798123338	0.6203559044316580
1	1	2	160	4.1586761484519933	0.6226671681953647
1	1	2	320	4.6762109356377524	0.6238308630323440
1	1	3	20	2.9969009771287197	0.6651540934013991
1	1	3	40	3.6144924639476756	0.6760637320523926
1	1	3	80	4.2475059922354439	0.6817125259376526
1	1	3	160	4.8887387594688807	0.6845884534145869
1	1	3	320	5.5342194219253561	0.6860397125567109
2	1	1	20	3.1556230753119783	0.4714279613766105
2	1	1	40	3.7573861118445009	0.4785215005165067
2	1	1	80	4.3662232452531547	0.4821454796932558
2	1	1	160	4.9786790482884766	0.4839772398312585
2	1	1	320	5.5929652253052684	0.4848981225379886
2	1	2	20	4.1555270002464250	0.5845684268948829
2	1	2	40	5.0321139887802046	0.5948952097518738
2	1	2	80	5.9262672509687573	0.6002157082709455
2	1	2	160	6.8295967704791643	0.6029169607148711
2	1	2	320	7.7376187300102367	0.6042780661571367
2	1	3	20	4.7401071109541055	0.6411598692523401
2	1	3	40	5.8133417172720270	0.6537960189121131
2	1	3	80	6.9165828316706758	0.6603585535541516
2	1	3	160	8.0358011415177967	0.6637048178422400
2	1	3	320	9.1632732748648238	0.6653947272628833
3	1	1	20	4.2889101304396929	0.4631802004289054
3	1	1	40	5.1431762910960503	0.4707318580690347
3	1	1	80	6.0087616857744186	0.4745957979085818
3	1	1	160	6.8801404331774827	0.4765503589871076
3	1	1	320	7.7544503175901389	0.4775333595759256
3	1	2	20	5.6485377991679472	0.5742786301219858
3	1	2	40	6.8888260363838882	0.5852359548650560
3	1	2	80	8.1558286757784522	0.5908899291399151
3	1	2	160	9.4367849957367741	0.5937627050026980
3	1	2	320	10.7248763977563677	0.5952107995563793
3	1	3	20	6.4490716187428801	0.6303557399251043
3	1	3	40	7.9658791328112459	0.6437482714649499
3	1	3	80	9.5273766692313970	0.6507141915803493
3	1	3	160	11.1126619024529868	0.6542688781822494
3	1	3	320	12.7102338155505224	0.6560647397306205

TABLE 5. Evidence of non-convergence to a finite expectation value of the first-return to the reference line (4th column) for some small $\bar{m}$, $\bar{n}$ and i.

8. D. Rainville, *The contiguous function relations for $_pf_q$ with application to Bateman's $j_n^{u,\nu}$ and Rice's $h_n(\zeta, p, \nu)$*, Bull. Amer. Math. Soc. **51** (1945), no. 10, 714–723. MR 0012726
9. Gian-Carlo Rota, *On the foundations of combinatorial theory. I. theory of möbius functions*, Z. Wahrscheinlichkeitsth. u. Verw. Gebiete **2** (1964), no. 4, 340–368. MR 0174487
10. Ranjan Roy, *Binomial identities and hypergeometric series*, Amer. Math. Monthly **94** (1987), no. 1, 36–46. MR 0873603
11. Kanwar Sen, Manju Lata Agarwal, and Sonali Chakraborty, *Generalized polya-eggenberger model of order k via lattice path approach*, J. Stat. Plan. Inf. **102** (2002), no. 2, 467–476. MR 1897220
12. Lucy Joan Slater, *Generalized hypergeometric functions*, Cambridge University Press, 1966. MR 0201688
13. Richard P. Stanley, *Enumerative combinatorics*, 2 ed., vol. 1, Cambridge University Press, 2011. MR 1442260
14. Miloš Vacek, *Sur la loi de polya régissant les fait corrélatifs*, Aktuárské vědy **3** (1932), no. 1, 18–28.
15. Herbert S. Wilf, *Generatingfunctionology*, Academic Press, 2004. MR 2172781
16. P. Wynn, *On a device for computing the $e_m(s_n)$ transformation*, Math. Tabl. Aids Comput. **10** (1956), no. 54, 91–96. MR 0084056

URL: https://www.mpia-hd.mpg.de/~mathar

Max-Planck Institute for Astronomy, Königstuhl 17, 69117 Heidelberg, Germany