Pólya Urn Models in Random Trees

Hosam M. Mahmoud
The George Washington University

October 20, 1997

[summary by Marko R. Riedel]

1. Examples and previous results

Consider an urn that contains balls of k different colors $1, 2, \ldots, k$. There is a set of evolution rules: *(i)* a ball is chosen at random from the urn, where all balls are equally likely; *(ii)* that ball's color or type is noted, and the ball is returned to the urn; *(iii)* if the ball had color i, α_{ij} balls of color j are added to the urn.

Question of interest. What is the composition of the urn after n draws?

The model is encoded in the *addition matrix* $A = [\alpha_{ij}]$, $1 \le i, j \le k$. The α_{ij} may themselves be random, but this talk is concerned exclusively with deterministic α_{ij}, i.e., the case of a fixed addition matrix A.

Pólya and Eggenberger (1923) investigated the two-color problem, with $A = sI$ and s a positive integer. Suppose the two colors are red and blue, and let R_n and B_n be the number of red and blue balls after n picks.

Example. Set $s = 2$ and start with two red balls and one blue ball. One of eight possible length-3 runs is: Pick blue (probability 1/3), the composition of the urn is now $R_1 = 2$ and $B_1 = 3$; Pick blue (probability 3/5), $R_2 = 2$ and $B_2 = 5$; Pick red (probability 2/7), $R_3 = 4$ and $B_3 = 7$.

What is the typical behavior of R_n and B_n? Bernard Friedman (1949) studied a more general urn, the addition matrix now being $A = \left[\begin{smallmatrix} \alpha & \beta \\ \beta & \alpha \end{smallmatrix}\right]$. Freedman (1965) showed that

$$R_n^* \xrightarrow{D} N(0,1), \quad B_n^* \xrightarrow{D} N(0,1), \qquad \text{where} \qquad R_n^* = \frac{R_n - E[R_n]}{\sqrt{V[R_n]}}$$

and B_n^* is defined similarly.

2. The connection to random trees

Recall the random permutation model for binary trees, where n keys are inserted into a binary tree such that the root of any subtree is larger than all left and less than all right descendants. We have a uniform distribution on the $n!$ possible key orderings and wish to compute tree statistics associated to this model.

The Poblete-Munro (1985) heuristic suggests that we can obtain a more balanced tree with little extra work: we require that all subtrees on the fringe and of size at most three be balanced. This means that we rebalance size 3 subtrees on the fringe, if necessary. This process yields shorter trees; in fact $E[D_n] = (12/7)\ln n$ (compare with $2\ln n$ for standard RBSTs).

2.1. **Balanced trees.** Work by Yao (1978) on 2-3 trees, Baeza-Yates, Gonnet and Ziviani on other tree statistics S_n shows that we can study $E[S_n]$ by studying fringe configurations of RBSTs, to obtain bounds of the type $f_1(n) \leq E[S_n] \leq f_2(n)$. Fringe analysis is based on exact counting of all (sub)trees less than or equal to a given height. The results improve in accuracy as the height is increased.

Mahmoud (1998) has used Pólya urn models to study the Poblete-Munro heuristic. We map fringe configurations to colors. The growth of the tree is modeled by a 3×3 urn. Suppose an incoming node is placed on one of the four leaves of a balanced subtree on three nodes. It is inserted without rebalancing the tree. Its sibling is a leaf. Suppose the next node is placed at that leaf. No rebalancing is required. Finally suppose that the next node is not placed at the sibling leaf, but rather at a leaf of the previous node. The tree must be rebalanced. We distinguish these three configurations by assigning different colors to the leaves concerned: color 1 to the leaves of any terminal node whose sibling is not a leaf, color 3 to the leaves of any terminal node whose sibling is a leaf, and color 2 to all such leaves. The leaves correspond to balls in a Pólya urn. The complexity measure of an insertion is the number of rotations, call it R_n. The addition matrix of the Pólya urn becomes

$$A = \begin{bmatrix} -2 & 1 & 2 \\ 4 & -1 & -2 \\ 4 & -1 & -2 \end{bmatrix}.$$

E.g., if we replace a leaf of color 1, we lose that leaf and recolor its sibling with color 2. The new leaves have color 3. R_n is therefore the number of picks of color 3. The row sums of the addition matrix A form the vector $S = [1,1,1]^T$, which reflects the fact that every BST on n nodes has $n+1$ leaves.

If an addition matrix A has the property that there exists an m such that all the entries of A^m are positive, we say that A is *regular.* In this particular example, A is not regular; nonetheless Mahmoud (1998) shows that

$$\frac{R_n - 2/7n}{\sqrt{n}} \xrightarrow{D} N\left(0, 66/637\right).$$

2.2. m**-ary search trees.** Under this model $m-1$ keys $k_1, k_2, \ldots, k_{m-1}$ are placed at the root of the tree. These keys partition the remaining keys into m intervals, $(-\infty, k_1), (k_1, k_2), \ldots, (k_{m-1}, +\infty)$, i.e., subtrees. The construction is recursive and the branch factor is m.

Example. Let $m = 3$ and consider the keys $9, 16, 4, 23, 11, 10, \ldots$ The first two keys are placed at the root, the key 4 is placed to the left of 9 and starts a new subtree, 23 is placed to the right of 16, also in a new subtree, and 11 and 10 fall between 9 and 16, starting a new subtree with root intervals $(9,10)$, $(10,11)$ and $(11,16)$. There are three types of nodes (or *blocks* in a hardware-oriented setting): leaves, nodes that contain a single key, and nodes that contain two keys.

More generally, we ask about S_n, the number of nodes after n insertions, where $S_n = \sum_j X_n^j$ and X_n^j counts the number of nodes that contain j keys. We construct an urn model by mapping gaps between keys at a node to balls whose color indicates the number of gaps at that node. We can recover the number of nodes of each type from the number of gaps of the corresponding color. For instance, consider a leaf that contains i keys and hence $i+1$ gaps. We map these gaps to balls of color $i+1$. Now suppose that $i < m-1$ and we insert a key at this leaf. We lose $i+1$ gaps of color $i+1$ and gain $i+2$ gaps of color $i+2$. The addition matrix associated to this model has the following shape:

$$A = \begin{bmatrix} & \ddots & & & & & \\ & \cdots & -r & r+1 & & & \\ & \cdots & \cdots & -(r+1) & r+2 & & \\ & & & & & \ddots & \\ m & & & & & & -1 \end{bmatrix}.$$

The eigenvalues of the addition matrix A. Order the eigenvalues according to their real part, letting λ_1 be the eigenvalue whose real part is the largest. Athreya and Nay (1972) showed that if the real part of λ_2 is less than half the real part of λ_1, a condition guaranteed if A is regular, then the colors have a normal $N(0,1)$ distribution.

In the urn model associated to m-ary trees, this property holds for $m < 27$, even though the associated urn is not regular. (This suggests that regularity is too strong a precondition for the results of Athreya and Nay.) When $m = 27$, there are two conjugate eigenvalues whose real part is larger than half the real part of λ_1. More precisely, Lew and Mahmoud (1994) showed that for any sequence $c_1, c_2, \ldots, c_k$, where c_j is the cost of a node that contains j keys, the vector of random variables $\mathbf{X}_n = [X_n^1, X_n^2, \ldots, X_n^k]^T$ converges to a multivariate normal, i.e.,

$$\frac{\mathbf{X}_n - E[\mathbf{X}_n]}{\sqrt{n}} \overset{D}{\rightarrow} \mathrm{MVN}(0, \Lambda) \qquad \text{for} \qquad m = 2, 3, \ldots, 26.$$

2.3. **Paged binary trees.** In this model every external node stores at most b keys, while internal nodes store a single key. Overflow on external nodes is processed by splitting the node according to some splitting rule, say by selecting the median and adding two subtrees whose roots store $b/2$ keys. The corresponding counting problem leads to a differential equation in $F(x,y)$, the super exponential generating function of paged binary trees:

$$\frac{\partial^{b-1} F(x,y)}{\partial x^{b-1}} = \left(\frac{\partial F(x,y)}{\partial x}\right)^2.$$

PBSTs have been considered by Flajolet, Mahmoud and Martínez Parra. Results that are based on Pólya urn models indicate that there is a phase transition at $b = 118$, when the real part of the second eigenvalue of the addition matrix becomes larger than half the real part of the first eigenvalue. Work in progress by Flajolet *et al.* seeks to construct an interpretation of this fact in the context of generating functions.

3. Case study: plane-oriented recursive trees

This type of tree models a recruiting process where the recruiting probability of a recruiting officer increases with the number of recruits attracted so far, or more generally, where the probability of a node to receive a new node is proportional to its degree, a scenario that Mahmoud (1991) calls "success breeds success." We use plane-oriented recursive trees as the underlying model[1], as proposed by Bergeron, Flajolet and Salvy (1992). Every node has outdegree $2k+1$, for some $k \geq 0$; k of its children are plane-embedded nodes, and $k+1$ leaves are placed in the gaps between adjacent nodes.

[1] If we were using the terminology of combinatorial analysis, we would refer to these trees as increasing trees; more precisely, as R-enriched increasing trees, where R is the *list* structure.

Consider a chain letter scheme where the acquisition price of a letter is 100F, and copies of the letter are sold at 40F. Given that there are n participants in the scheme, we ask how many of them have just broken even, i.e., sold three letters. Let blue represent insertion slots at nodes that have bought, but not sold a single letter; red, nodes that bought and sold one copy of the letter, green, two copies, and white, three, i.e., broken even, and let B_n, R_n, G_n and W_n be the corresponding RVs. (We start with a single participant, i.e., $B_0 = 1$, $R_0 = G_0 = W_0 = 0$.) Finally, assume that the success probability of a participant is proportional to the number of letters sold (other models are possible and even reasonable). The addition matrix is easily seen to be

$$A = \begin{bmatrix} 0 & 2 & 0 & 0 \\ 1 & -2 & 3 & 0 \\ 1 & 0 & -3 & 4 \\ 1 & 0 & 0 & 1 \end{bmatrix}.$$

E.g., if a participant has sold two letters and sells another, the three green insertion slots at the corresponding node are replaced by four white ones, and a new participant who has not sold any copy of his letter yet must be accounted for. Note that A is not regular. It can be shown that B_n is the number of leaves in a random tree of size $n+1$.

Mahmoud, Smythe and Szymański (1993) show that

$$\mathrm{E}\begin{pmatrix} B_n \\ R_n \\ G_n \end{pmatrix} \sim \begin{pmatrix} 1/3 \\ 1/6 \\ 1/10 \end{pmatrix} (2n+1),$$

and that the covariance matrix is

$$\mathrm{Cov}(B_n, R_n, G_n) \sim \begin{pmatrix} 1/9 & -8/45 & -1/15 \\ -8/45 & 23/45 & -11/105 \\ -1/15 & -11/105 & -179/350 \end{pmatrix} n.$$

We sketch the proof of this result. Introduce the indicator variables $I_n^{(B)}, I_n^{(R)}, I_n^{(G)}$, $I_n^{(W)}$ so that $I_n^{(B)} + I_n^{(R)} + I_n^{(G)} + I_n^{(W)} = 1$. We now have e.g., $R_n = R_{n-1} + 2I_n^{(B)} - 2I_n^{(R)}$, and hence

$$E[R_n] = E[R_{n-1}] + 2E[I_n^{(B)}] - 2E[I_n^{(R)}].$$

The expectations of the indicator variables are obtained by conditioning on the $n-1$ picks that lead to a particular urn (call this σ-field T_{n-1}), so that

$$E[I_n^{(B)} \mid T_{n-1}] = \frac{B_{n-1}(T_{n-1})}{2n-1} \quad \text{and} \quad E[I_n^{(B)}] = \frac{E[B_n]}{2n-1}.$$

Substitute to get

$$E[R_n] = E[R_{n-1}] + 2\frac{E[B_n]}{2n-1} - 2\frac{E[R_n]}{2n-1}.$$

Similar computations for $E[B_n]$, $E[G_n]$ and $E[W_n]$ yield a system of recurrences of the form

$$[B_n, R_n, G_n, W_n]^T = F(n)[B_{n-1}, R_{n-1}, G_{n-1}, W_{n-1}]^T$$

where $F(n)$ is a matrix that depends on n. This system may be solved asymptotically. (Note that we have made critical use of the fact that the total number of balls in the urn is a function of n, namely $2n+1$.)

The computation of the covariance is more involved. Start from $R_n = R_{n-1} + 2I_n^{(B)} - 2I_n^{(R)}$ as before, square both sides and use simple properties of mutually exclusive indicator variables to get

$$R_n^2 = R_{n-1}^2 + 4I_n^{(B)}R_{n-1} - 4I_n^{(R)}R_{n-1} + 4I_n^{(B)} + 4I_n^{(R)}.$$

Binary cross products of the four RVs appear on taking expectations, i.e., we develop recurrences for $E[R_n^2] = E[R_n R_n]$ and these recurrences involve terms like

$$\begin{aligned} E[I_n^{(B)} R_{n-1}] &= E\Big[E[I_n^{(B)} R_{n-1} \mid T_{n-1}]\Big] \\ &= E\Big[R_{n-1} E[I_n^{(B)} \mid T_{n-1}]\Big] = E\left[R_{n-1}\frac{B_{n-1}}{2n-1}\right] = \frac{E[R_{n-1}B_{n-1}]}{2n-1} \end{aligned}$$

The result is a system of recurrences in all binary cross products that yields the desired asymptotics.

Next consider the vector X_i of centered RVs;

$$X_i = \begin{pmatrix} B_i^* \\ R_i^* \\ G_i^* \end{pmatrix} = \begin{pmatrix} B_i \\ R_i \\ G_i \end{pmatrix} - (2i+1)\begin{pmatrix} 1/3 \\ 1/6 \\ 1/10 \end{pmatrix}$$

Mahmoud, Smythe and Szymanski (1993) show that

$$\frac{X_i}{\sqrt{n}} \xrightarrow{D} \text{MVN}\left(0, \begin{pmatrix} 1/9 & -8/45 & -1/15 \\ -8/45 & 23/45 & -11/105 \\ -1/15 & -11/105 & -179/350 \end{pmatrix}\right).$$

The proof uses martingale techniques. Recall that a martingale is a sequence $Y_1, Y_2, Y_3, \ldots$ of random variables such that $E[Y_n \mid T_{n-1}] = Y_{n-1}$. E.g., consider a fair game ("win all or lose all with equal probability, i.e., 1/2"), which gives

$$E[Y_n \mid T_{n-1}] = 0 \cdot Y_{n-1}\frac{1}{2} + 2Y_{n-1}\frac{1}{2} = Y_{n-1}.$$

Note that $E[Y_n] = E[Y_{n-1}] = \cdots = E[Y_1] = 0$ in this example; this is known as the *martingale difference property*, because if $Y_1, Y_2, Y_3, \ldots$ is a martingale, then $E[Y_n - Y_{n-1} \mid T_{n-1}] = 0$. We can reconstruct the martingale from the sequence of first differences, i.e., via $\sum_{k=1}^{n} \triangle Y_k = Y_n$. More generally, we can construct a martingale from any sequence of random variables that has the martingale difference property; this was done e.g., by Régnier (1989) in the context of algorithms, who showed that the cost of Quicksort has a limit distribution.

If $E[\triangle Z_i \mid T_{i-1}] = 0$, then $A_n = \sum_{i=1}^{n} \triangle Z_i$ is a martingale, because

$$E[A_n \mid T_{n-1}] = E\left[\sum_{i=1}^{n} \triangle Z_i \mid T_{n-1}\right] = E\left[\triangle Z_n + \sum_{i=1}^{n-1} \triangle Z_i \mid T_{n-1}\right] = \sum_{i=1}^{n-1} \triangle Z_i = A_{n-1}.$$

By linearity, $E[\triangle Z_i \mid T_{i-1}] = 0$ implies $E[b_i \triangle Z_i \mid T_{i-1}] = 0$ for any sequence of constants $\{b_i\}$, and hence $\sum_{i=1}^{n} b_i \triangle Z_i$ is a martingale.

We return to the four-color urn of the chain letter scheme; consider the color blue.

$$E[B_i \mid T_{i-1}] = E\Big[B_{i-1} + \Big(1 - I_n^{(B)}\Big) \mid T_{i-1}\Big] = B_{i-1} + 1 - E[I_n^{(B)} \mid T_{i-1}] = B_{i-1} + 1 - \frac{1}{2i-1}B_{i-1}$$

Further manipulation yields

$$\begin{aligned} E[B_i - 1/3(2i+1) \mid T_{i-1}] = (B_{i-1} - 1/3(2i-1)) + 1/3(2i-1) - 1/3(2i+1) \\ + 1 - \frac{1}{2i-1}(B_{i-1} - 1/3(2i-1)) - 1/3 \end{aligned}$$

and hence $E[B_i^* \mid T_{i-1}] = B_{i-1}^* - B_{i-1}^*/(2i-1)$. B_i^* is not a martingale but

$$B_i^* - B_{i-1}^* + \frac{1}{2i-1}B_{i-1}^*$$

is a *martingale difference*, because $E\left[\triangle \mathrm{M}_i^B \mid T_{n-1}\right]=0$, where we have set

$$\triangle \mathrm{M}_i^B = B_i^* - B_{i-1}^* + \frac{1}{2i-1}B_{i-1}^*.$$

We can construct a martingale from $\triangle \mathrm{M}_i^B$, since$\sum_{i=1}^n b_i \triangle \mathrm{M}_i^B$ is a martingale for any sequence $\{b_i\}$.

The *Cramér-Wold* device can be used to prove convergence to a multivariate normal. Suppose we seek $[X_n^{(1)}, X_n^{(2)}, X_n^{(3)}, \ldots, X_n^{(j)}]^T \xrightarrow{D} \mathrm{MVN}(0,\Lambda)$. It suffices to prove that any linear combination of the $X_n^{(\cdot)}$ converges to a normal distribution, i.e.,

$$\alpha_1 X_n^{(1)} + \alpha_2 X_n^{(2)} + \cdots + \alpha_j X_n^{(j)} \xrightarrow{D} N\left(0, \sigma_{\alpha_1,\ldots,\alpha_j}\right),$$

$\alpha_1, \alpha_2, \ldots, \alpha_j$ arbitrary. Here $j=3$ and we study $W_n = \alpha_1 B_n^* + \alpha_2 R_n^* + \alpha_3 G_n^*$. Centering the remaining two variables, we have

$$\begin{aligned}
E[B_i^* - B_{i-1}^* \mid T_{i-1}] &= -\frac{1}{2i-1}B_{i-1}^* \\
E[R_i^* - R_{i-1}^* \mid T_{i-1}] &= +\frac{2}{2i-1}(B_{i-1}^* - R_{i-1}^*) \\
E[G_i^* - G_{i-1}^* \mid T_{i-1}] &= +\frac{3}{2i-1}(R_{i-1}^* - G_{i-1}^*).
\end{aligned}$$

Introduce the martingale differences

$$\begin{aligned}
\triangle \mathrm{M}_i^B &= B_i^* - B_{i-1}^* + \frac{1}{2i-1}B_{i-1}^* \\
\triangle \mathrm{M}_i^R &= R_i^* - R_{i-1}^* - \frac{2}{2i-1}(B_{i-1}^* - R_{i-1}^*) \\
\triangle \mathrm{M}_i^G &= G_i^* - G_{i-1}^* - \frac{3}{2i-1}(R_{i-1}^* - G_{i-1}^*)
\end{aligned}$$

and set $V_{nk} = \sum_{i=1}^{k}\left(b_{in}\triangle \mathrm{M}_i^B + c_{in}\triangle \mathrm{M}_i^R + d_{in}\triangle \mathrm{M}_i^G\right)$ for arbitrary $\{b_{in}\}, \{c_{in}\}, \{d_{in}\}$. It remains to choose $\{b_{in}\}, \{c_{in}\}$ and $\{d_{in}\}$. Expand V_{nn} to get

$$V_{nn} = b_{nn}\left(B_n^* - B_{n-1}^* + \frac{1}{2n-1}B_{n-1}^*\right) + \cdots + \left(B_{n-1}^* - B_{n-2}^* + \frac{1}{2n-3}B_{n-3}^*\right) + \cdots$$

To obtain the particular linear combination $\alpha_1 B_n^* + \alpha_2 R_n^* + \alpha_3 G_n^*$ we set $b_{nn} = \alpha_1$, so that the term in B_n^* is preserved, and choose the remaining $b_{n,i}$ to cancel B_{n-1}^*, B_{n-2}^* etc. This technique can be used to show that given α_1, α_2 and α_3, we may choose constants $\{b_{in}\}, \{c_{in}\}$ and $\{d_{in}\}$, so that

$$V_{nn} = (\alpha_1 B_n^* + \alpha_2 R_n^* + \alpha_3 G_n^*) - 3/2c_{1n} + 10/3d_{1n},$$

which by a martingale central limit theorem yields

$$\frac{\alpha_1 B_n^* + \alpha_2 R_n^* + \alpha_3 G_n^*}{\sqrt{n}} \sim \frac{V_{nn}}{\sqrt{n}}, \qquad \text{hence} \qquad \frac{W_n}{\sqrt{n}} \xrightarrow{D} N\left(0, \sigma_{\alpha_1,\ldots,\alpha_j}\right), \quad \begin{pmatrix} B_n^* \\ R_n^* \\ G_n^* \end{pmatrix} \xrightarrow{D} \mathrm{MVN}(0,\Lambda).$$

Concluding remark

It should be obvious from the highly restricted class of addition matrices that have been considered that an abundance of combinatorial problems and possible addition matrices remain to be analyzed; e.g., apparently simple instances such as $\left[\begin{smallmatrix} 2 & 0 \\ 3 & 4 \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 1 & 0 \\ 1 & 1 \end{smallmatrix}\right]$ have so far resisted attack.