Properties of k-colourings of Random Increasing Trees

Alex Francis, Helmut H. Pitters

Department of Statistics
University of California, Berkeley

Abstract

In graph theory, graph coloring refers to the problem of labeling vertices such that no connected vertices have the same color. The problem has recently become prominent in the Computer Science community, especially among artificial intelligence researchers studying constraint satisfaction problems, or CSPs. In this paper, we consider a simpler problem called tree coloring, in which the graphs are restricted to be trees. We consider two types of random tree structures, the *random recursive tree* and the *d-ary tree*. We explore the limiting distribution of colors in the entire tree using urns as a model, and then, in the RRT, turn to the limiting distribution of colors at height H using Markov Chains. Finally, for the RRT, we construct a classification algorithm for the color of the root given only the frequency distribution at height H of the tree, and explore the properties of such an algorithm.

Contents

1. Random Recursive Trees

1.1. Introduction

A random recursive tree of size n is a non-planar rooted tree with labels on the integers $\{1, \ldots, n\}$. Random recursive trees are increasing trees, meaning that the label of any parent node must exceed the label of its child. RRTs, as we shall henceforth refer to them, are distinguished from other increasing trees by its construction rule, which is quite simple, and occurs in two steps [1].

1. Choose a node from the tree uniformly at random.
2. Give that node a child node, and give it label $n+1$ (assuming the current tree has n vertices).

1.2. Limiting Frequencies of Colors: $k = 2$

Consider a random recursive tree T_n of size n. A k-colouring of a tree $t = (V, E)$ is an assignment $V^{[k]}$ of one of k colours (labeled $\{1, \ldots, n\}$) to each vertex $v \in V$ of t such that no two adjacent vertices have the same colour, i.e. $\{v, w\} \in E$ implies that v and w are assigned different colours. The number $C_n^{(k)}$ of k-colourings of a tree of size n is given by $C_1^{(k)} = k$, and

$$C_n^{(k)} = C_{n-1}^{(n-1)}(n-1)(k-1) = (n-1)!(k-1)^{n-1}k \qquad (n \in \mathbb{N}).$$

Let $C_n^{(k)}(i)$, $1 \leq i \leq k$, denote the number of nodes on T_n of color i, after T_n was coloured with a k-colouring chosen uniformly at random.

If we restrict ourselves to the case of $k = 2$ colours (also known as the hardcore model where $C_n^{(k)}$ counts the number of independent sets), then the random vectors recording the relative frequencies $\frac{1}{n}\langle C_n^{(2)}(1), C_n^{(2)}(1)\rangle$ of the two colours converge in distribution

$$\frac{1}{n}\langle C_n^{(2)}(1), C_n^{(2)}(1)\rangle \to \langle C^{(2)}(1), C^{(2)}(2)\rangle$$

as $n \to \infty$. Why? Because, if we restrict our focus only to the colours of the vertices with labels $\leq n$ (and ignore the tree structure), we have the special case of Bernard Friedmann's urn, started with two balls of different colours and ball-replacement matrix [4],

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

In words: At each step a ball is chosen from the urn at random and returned with another ball of the opposite colour.

It is classical that [4, Theorem 3.4],

$$\frac{C_n^{(2)}(1) - \frac{1}{2}n}{\sqrt{n}} \to_d \mathcal{N}(0, \frac{1}{12}).$$

1.3. Limiting Frequencies of Colors: k arbitrary

Now consider k colours for some arbitrary positive integer k. Focus on the external nodes (vertices that are not currently in the tree, but may be placed in the tree on the next iteration; in an RRT, there are n external nodes, since there is no degree limit on individual vertices) in T_n, and say that an external node has colour c, if its parent (which necessarily is an internal node) has color c. The ball-replacement matrix for the urn of n balls on k colours is,

$$\begin{bmatrix} 0 & B_1 & B_2 & \cdots & B_{k-1} \\ B_1 & 0 & B_2 & \cdots & B_{k-1} \\ & & \vdots & & \\ B_1 & B_2 & \cdots & B_{k-1} & 0 \end{bmatrix},$$

where $\langle B_1, \ldots, B_{k-1} \rangle$ is an exchangeable random vector of Bernoulli $\{0,1\}$ random variables such that $B_1 + B_2 + \cdots + B_{k-1} = 1$. Clearly, this implies $\mathbb{E}[B_i] = 1/(k-1) \forall i \in [k-1]$. Mahmoud gives us an explicit formula for computing the limiting distribution of multichromatic extended urns, like the one above. The theorem is repeated below, and is due to Smythe [6, Lemma 2.2] [4, Theorem 6.4].

THEOREM 1.3.1. Suppose a k-color extended Polya urn scheme has an average generator with principal eigenvalue $\lambda_1 > 0$, and the corresponding (ℓ_1 normalized) principal left eigenvector is $\mathbf{v} = \langle v_1 \ldots v_k \rangle$. Let $X_{j,n}$ be the number of balls of color j after n draws, for $j = 1, \ldots, k$. For each component, $j = 1, \ldots, k$

$$\frac{X_{j,n}}{n} \to \lambda_1 v_j$$

Note that an *average generator* is simply the expected value of the ball-replacement matrix. Now, it seems that the only way to illuminate these results is to compute an explicit formula for the maximum eigenvalue of the ball-replacement matrix above. Luckily, Mahmoud tells us that the principal eigenvalue of an extended multichromatic urn is the row sum of the average generator, in this case 1 (for a challenge, one can verify this result using Rayleigh quotients to obtain a simple optimization problem and invoking strong duality to solve it!) [4]. Knowing this, we can compute the principal left eigenvector by noting that such a vector has the form,

$$v^\top B = \lambda_1 v = v$$

One can use Gaussian elimination to uncover $v_1 = v_2 = \ldots = v_k$. Since the eigenvector in Mahmouds formulation is ℓ_1-normalized, the explicit form of v is,

$$v = \frac{1}{k}\mathbf{1}$$

Where $\mathbf{1}$ is the vector of 1s. This tells us that the frequency distribution of colors in an RRT converges to uniformity as the number of vertices in the tree is driven to infinity.

1.4. Distribution of Nodes at Height H

The result above is somewhat unintuitive. Since the balls in the urn are picked based on the color of the root node, it would seem like the color root node would impact the distribution of colors in the limiting distribution. This raises another research question: how large does the tree have to be in order for the root node to be abstracted away? In order to pursue this question further, consider utilizing the urns framework for analysis on the nodes at a particular *depth* in the tree, say, height $H \leq \max(\text{depth})$.

Upon first glance, at level H of the tree, we must also include level $H-1$ of the tree in the urn model, since this impacts the distribution of noes on level h. But level $H-1$ is impacted by the level $H-2$ distribution, and so on, inductively, all the way up to the root node. The levels $> H+1$ are also necessary, for technical reasons concerning the formulation of the extended multichromatic urn in Mahmoud. Therefore, let's frame the problem an urn with $(H+2) \times k$ distinctly colored balls. The ball picked at an individual step in the construction algorithm is colored by height and color. Call the index (i, ℓ). The replacement rule, in words, is the following:

1. Replace the ball with another ball of identical height and color, (i, ℓ). Since we could perform the exact same update once again, this internal node must have remaining representation in the urn.
2. Add an additional ball at $(i+1, [k]\backslash\ell)$, where $[k]\backslash\ell$ is the complement ℓ in the set of colors. This adds an internal node to the tree of a different color, at the next depth.

The ball-replacement matrix is challenging to visualize. Note that we can formalize the language above by noting that any row and column must be indexed by *both* height and by color. Let height be the "outer index" and let color be the "inner index." Concretely, the ball replacement matrix, U, will have the form,

$$\begin{array}{ccccccccc}
\multicolumn{3}{c}{\overbrace{\qquad\qquad\qquad\qquad\qquad}^{\text{h = 0}}} & \multicolumn{3}{c}{\overbrace{\qquad\qquad\qquad\qquad\qquad}^{\text{h = 1}}} & & & \\
\end{array}$$
$$\begin{bmatrix}
U_{(1,1),(1,1)} & \cdots & U_{(1,1),(1,k)} & U_{(1,1),(2,1)} & \cdots & U_{(1,1),(2,k)} & \cdots & \cdots & U_{(1,1),(H+1,k)} \\
U_{(1,2),(1,1)} & \cdots & U_{(1,2),(1,k)} & U_{(1,2),(2,1)} & \cdots & U_{(1,2),(2,k)} & \cdots & \cdots & U_{(1,2),(H+1,k)} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \vdots \\
U_{(1,k),(1,1)} & \cdots & U_{(1,k),(1,k)} & U_{(1,k),(2,1)} & \cdots & U_{(1,k),(2,k)} & \cdots & \cdots & U_{(1,k),(H+1,k)} \\
U_{(2,1),(1,1)} & \cdots & U_{(2,1),(1,k)} & U_{(2,1),(2,1)} & \cdots & U_{(2,1),(2,k)} & \cdots & \cdots & U_{(2,1),(H+1,k)} \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \vdots \\
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \ddots & \vdots \\
U_{(H+1,k),(1,1)} & \cdots & U_{(H+1,k),(1,k)} & U_{(H+1,k),(2,1)} & \cdots & U_{(H+1,k),(2,k)} & \cdots & \cdots & U_{(H+1,k),(H+1,k)}
\end{bmatrix}$$

Then, for row (i, ℓ) in the matrix, all elements are zero, except for the (i, ℓ)th column element, which is one (bullet point (1)), and all elements in columns $(i+1, [k]\backslash\ell)$, which is an exchangeable random vector $\langle B_1, \ldots, B_{k-1}\rangle$ of Bernoulli $\{0,1\}$ random variables such that $B_1 + B_2 + \cdots + B_{k-1} = 1$ (bullet point (2)). As before, $\mathbb{E}[B_j] = 1/(k-1) \forall j \in [k-1]$.

Applying this formulation in order to obtain the ball-replacement matrix explicitly yields a convenient representation. Denote by 0_k the zero matrix with k rows and k columns, and denote B as the random $k \times k$ matrix,

$$B = \begin{bmatrix}
0 & B_1 & B_2 & \cdots & B_{k-1} \\
B_1 & 0 & B_2 & \cdots & B_{k-1} \\
 & & \vdots & & \\
B_1 & B_2 & \cdots & B_{k-1} & 0
\end{bmatrix}$$

The reader should verify that the explicit form of the ball-replacement matrix can therefore be represented as an upper-triangular matrix of blocks,

$$U = \begin{bmatrix} 0_k & B & 0_k & 0_k & \cdots & 0_k \\ 0_k & 0_k & B & 0_k & \cdots & 0_k \\ 0_k & 0_k & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0_k & B \\ 0_k & 0_k & \cdots & \cdots & \cdots & B \end{bmatrix}$$

From Mahmoud,

$$C_n^{(k)}(i) \to_P \lambda_1 v_i$$

Where v_i is ith component of the eigenvector corresponding to the principal eigenvalue of the average generator [4]. The average generator for U is,

$$\mathbb{E}[U] = \begin{bmatrix} 0_k & \mathbb{E}[B] & 0_k & 0_k & \cdots & 0_k \\ 0_k & 0_k & \mathbb{E}[B] & 0_k & \cdots & 0_k \\ 0_k & 0_k & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0_k & \mathbb{E}[B] \\ 0_k & 0_k & \cdots & \cdots & \cdots & \mathbb{E}[B] \end{bmatrix}$$

$$\mathbb{E}[B] = \begin{bmatrix} 0 & 1/(k-1) & 1/(k-1) & \cdots & 1/(k-1) \\ 1/(k-1) & 0 & 1/(k-1) & \cdots & 1/(k-1) \\ & & \vdots & & \\ 1/(k-1) & 1/(k-1) & \cdots & 1/(k-1) & 0 \end{bmatrix}.$$

In order to ascertain the final eigendecomposition, consider another definition. For a matrix A, λ is an eigenvalue, and v is the corresponding *left* eigenvector if,

$$v^\top A = \lambda v^\top$$

The principal left eigenvalue can be computed by cleverly manipulating Laplace's determinantal expansion and using the result on upper triangular matrices proven above. However, it is easier to notice that the maximum eigenvalue is simply the row sum in the average generator, which is 1 [4]. Then, the left eigenvector is computed by solving the system of equations,

$$\begin{bmatrix} v_1 & v_2 & \dots & v_{(H+2)k} \end{bmatrix} \begin{bmatrix} 0_k & \mathbb{E}[B] & 0_k & 0_k & \cdots & 0_k \\ 0_k & 0_k & \mathbb{E}[B] & 0_k & \cdots & 0_k \\ 0_k & 0_k & \ddots & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & 0_k & \mathbb{E}[B] \\ 0_k & 0_k & \cdots & \cdots & \cdots & \mathbb{E}[B] \end{bmatrix} = \begin{bmatrix} v_1 & v_2 & \dots & v_{(H+2)k} \end{bmatrix}$$

Then, $v_i = 0$ for $i \in \{0, \dots, (H+1)k\}$, and $v_j = 1$ for $j \in \{(H+1)k+1, \dots, (H+2)k\}$. By normalizing the principal left eigenvalue, we divide by the ℓ_1 norm, and therefore replace the v_j with $1/k$. This means, in words, the limiting frequency of nodes beyond the chosen level H dominates those of the first H levels of the tree, and that the distribution of colors among those nodes is uniform.[1]

[1] A nice visualization of this is available in the Appendix. The figures in the Appendix are drawn from experimental computer simulations. The code used to complete these experiments is publicly available at `https://github.com/afrancis13/random-tree-colorings`

1.5. Limiting Frequencies over n_H

Next, we turn to a discussion of something more directly related to the question posited in the previous section, namely the limiting frequencies of the colors of a height with respect to *only* the nodes at that height. This will prevent the nodes at level $> H$ from dominating the vertices at lower heights, as in the previous part.

Let us begin by considering the $k = 2$ case. The behavior here is trivial. Notice that, once the color of the root node is fixed, the colors of the nodes at height H are deterministic. Namely, denote the relative frequency of colors at height H, as $n \to \infty$, as $C_H^{(2)}(i)$, where $i \in \{1, 2\}$. We have the piecewise relation,

$$C_H^{(2)}(i) = \begin{cases} 1 & H \text{ is odd, root color is not } i \\ 1 & H \text{ is even, root color is } i \\ 0 & \text{otherwise} \end{cases}$$

One observation is that we can utilize the ball-replacement matrix to describe this process. Fix an initial state vector of probabilities, $v_0 = \begin{bmatrix} 1 & 0 \end{bmatrix}^\top$. This represents a root in which the color of the node is the first color. This color choice is arbitrary; we could have just as easily chosen the vector $v_0 = \begin{bmatrix} 0 & 1 \end{bmatrix}^\top$, and the argument is analogous for this case. Then, notice that we can describe the composition of colors on level H as a probability vector by the process,

$$v_H = B^H v_0$$

This is just a Markov Chain with B as the transition matrix. Since the Markov Chain is *not* aperiodic (consider, for example $B^2 = B$), the Markov Chain is not guaranteed to converge to a stationary distribution.

Let us try to utilize a Markov Chain for $k > 2$. The initial probability vector v_0 is once again the indicator vector, $v_0 = \mathbb{1}\{i = r\}$, where r is selected uniformly at random from $\{1, \dots, k\}$. Once again, the transition matrix is defined by,

$$B = \begin{bmatrix} 0 & 1/(k-1) & 1/(k-1) & \cdots & 1/(k-1) \\ 1/(k-1) & 0 & 1/(k-1) & \cdots & 1/(k-1) \\ & & \vdots & & \\ 1/(k-1) & 1/(k-1) & \cdots & 1/(k-1) & 0 \end{bmatrix}$$

Then, we can approximate the urn process, it seems, by multiplying the initial probability vector by the transition matrix. This process repeats until we reach the desired height, producing the formula for the relative frequencies of colors at a particular height in the RRT,

$$v_H = B^H v_0$$

It is worth noting that as $H \to \infty$, the relative frequencies converge to the uniform distribution. This can be seen by noting that this Markov Chain *is* aperiodic, positive-recurrent, and irreducible. The stationary distribution can be unearthed by plugging the vector of uniform probabilities into the definition of stationarity.

$$\begin{aligned}\lim_{H\to\infty} B^H v_0 &= v_H : Bv_H = v_H\\ B(1/k)\mathbf{1} &= \left(\frac{1}{k-1}\right)\left(\frac{1}{k}\right)(k-1)\mathbf{1}\\ &= \frac{1}{k}\mathbf{1}\end{aligned}$$

The motivation for using a Markov Chain approximation on the $k > 2$ case is, therefore quite compelling. But Markov Chains differ substantially from urn processes, and therefore seem unsuited for the problem. The theorem, and proof below, gives an unintuitive result. Let's first introduce a lemma, which will provide a useful result that will aid us in completing the proof of the theorem.

LEMMA 1.5.1 The number of children of any given node in the RRT diverges almost surely.

Proof. We take advantage of *Kolmogorov's Three-Series Theorem.* Denote the label of the chosen node as some finite $j \in \mathbb{N}$. Denote the number of children of node j as $Z_j = \sum_{i=j+1}^{\infty} Z_{j,i}$, where $Z_{j,i} \sim \text{Bernoulli}(1/i)$ (and are independent). We must prove that Z_j, the infinite series, diverges almost surely. The aforementioned three-series theorem states that a random series converges almost surely if and only if the following conditions are satisfied for some $A > 0$ [5].

1. $\sum_{i=1}^{\infty} P(|Z|_n \geq A)$ converges
2. Let $Y_n = Z_n \mathbb{1}_{\{|Z|_n \leq A\}}$, then $\sum_{n=1}^{\infty} \mathbb{E}[Y_n]$, the series of expected values of Y_n, converges
3. $\sum_{n=1}^{\infty} \text{Var}(Y_n)$ converges

It is immediately obvious that we cannot choose $A \leq 0$ or $A \geq 1$, since the former choice would violate condition (1) and the latter choice would violate condition (2) (Why? $\sum_{n=1}^{\infty} \mathbb{E}[Y_n] = \sum_{n=1}^{\infty} \mathbb{E}[Z_n] = \sum_{i=j}^{\infty} 1/i \to \infty$). Now consider some $A \in (0, 1)$. Note that, for any A in this range, $P(|Z_n| \geq A) = 1/i$, for some iteration i in the sum, which again yields a harmonic series, this time for condition (1). Therefore, the random series diverges almost surely. □

THEOREM 1.5.2. The distribution of colors at height H in a RRT described by the aforementioned urn process converges to the Markov Chain in the limit.

Proof. The proof technique is induction. First, consider the base case, that is, the transition from the root, which we will say has color C, to $H = 1$. Then, the distribution of nodes at height 0 is $v_0 = \mathbb{1}\{i = C\}$. Since the root label is 1, and the number of children of the root diverges almost surely as $n \to \infty$ (from the lemma), the distribution of nodes at height 1 converges exactly to,

$$v_1 = \frac{1}{k-1}\mathbf{1} - \frac{1}{k-1}\mathbb{1}\{i = C\} = B_{C_0^{(k)},*} = Bv_0$$

Where, in the above, $B_{C,*}$ denotes the Cth row of the ball-replacement matrix B. For the inductive step, we have from the inductive hypothesis that the distribution of colors at height H is the theoretical probability from the Markov Chain analysis, that is, that $v_H = Bv_{H-1}$. Now, ignore the levels $< H$, and consider tree level H as a set of n_H rooted trees. For each tree, the distribution of colors is the *corresponding row in B* (since, by the proof of the lemma, the number of children of each node grows without bound), that is, if the color of node i is C_i, the distribution of the children is $B_{C_i,*}$. Applying this argument over all n_H nodes at height H, we have,

$$v_{H+1} = \frac{1}{n_H} \sum_{j=1}^{k} n_H (v_H)_j B_{j,*} = \sum_{j=1}^{k} (v_H)_j B_{*,j} = B v_H$$

Where $(v_H)_j$ is the jth entry in v_H. The proof is complete. □

1.6. Classifying the Root

An interesting scientific question motivating this work has been: "how easy, or difficult, is it to recover the root at a particular level H?" At this point in the discussion, we have identified that the distribution of total colors in the tree, and below any particular finite height H in the tree, in the limit as the uniform distribution. We have also identified the distribution of colors at a particular height using a Markov process. Next, it is interesting to address the scientific question directly, even algorithmically. The following intermediary results are useful.

Theorem 1.6.1. The expected distribution of colors on level H of the tree is described by the limiting distribution. The covariance matrix at level H is described by,

$$(\Sigma_H)_{ij} = \begin{cases} \dfrac{(k-2)(n_H - n_{H,i})}{(n_H(k-1))^2} & i = j \\ -\dfrac{(n_H - n_{H,i} - n_{H,j})}{(n_H(k-1))^2} & i \neq j \end{cases}$$

Where n_H denotes the number of nodes at height H, and $n_{H,i}$ denotes the number of nodes at height H that have a *parent* of color $i \in [k]$.

Proof. The expected value follows from the fact that the transitions are defined by the ball-transition matrix defined earlier in this document, the expectation of which is the transition matrix used in the Markov chain formulation. This is weaker than the convergence condition we already derived.

The covariance matrices are more difficult to compute. It is fruitful to first consider the case $H = 1$. Notice that, once the root color is set, call it $C \in [k]$, the distribution of the colors of the nodes at height 1 is multinomial, with the parameter vector,

$$\langle p_1, \ldots, p_k \rangle = \frac{1}{k-1}(\mathbf{1} - \mathbb{1}\{i = C\})$$

It is a well-known fact of the multinomial distribution that covariance matrix of a vector of multinomial random variables is,

$$(\Sigma)_{ij} = \begin{cases} np_i(1-p_i) & i = j \\ -np_ip_j & i \neq j \end{cases}$$

Therefore, the covariance matrix for vector of color relative frequencies in the $H = 1$ case is,

$$(\Sigma)_{ij} = \begin{cases} 0 & i = C \text{ or } j = C \\ \dfrac{(k-2)}{n_1(k-1)^2} & i = j \\ -\dfrac{1}{n_1(k-1)^2} & i \neq j \end{cases}$$

Note that in both cases, we divide by n_1, in order to obtain the relative frequency, and that this term is squared in the variance/covariance calculation. This adheres to the claim in the theorem thus far.

The more general result follows from some algebraic manipulation. Let us attempt to compute the variance of the relative frequency of probabilities at height H. In order to do so, denote $(V_H)_i$ as the random variable representing the relative frequency of color i at height H. Denote $(Q_H)_i$ as the absolute frequency, of color i on height H (that is, $Q_H = V_H n_H$). Finally, denote $(Q_H)_i^j$ as the absolute frequency of color i on height H with parent $j \in [n_{H-1}]$, where j represents an index for the parent node. Let C_j be the color of this node. Note that $(Q_H)_i^j$ is once again a multinomial random variable, with parameter vector,

$$\langle p_1, \ldots, p_k \rangle = \frac{1}{k-1}(\mathbf{1} - \mathbb{1}\{i = C_j\})$$

Then, notice that,

$$(V_H)_i = \frac{1}{n_H} \sum_{j=1}^{n_{H-1}} (Q_H)_i^j$$

Computing variance, and denoting by n_H^j the number of children of parent with label j at height H (note that this is distinguished from $n_{H,j}$, the number of children with *color* j at height H), we have,

$$\begin{aligned}
\mathrm{Var}((V_H)_i) &= \mathrm{Var}\left(\frac{1}{n_H}\sum_{j=1}^{n_{H-1}} (Q_H)_i^j\right) \\
&= \frac{1}{n_H^2}\left(\sum_{j=1}^{n_{H-1}} \mathrm{Var}((Q_H)_i^j)\right) \\
&= \frac{1}{n_H^2}\left(\sum_{j=1}^{n_{H-1}} n_H^j \left(\frac{1}{k-1}\right)\left(\frac{k-2}{k-1}\right) \mathbb{1}\{i \neq C_j\}\right) \\
&= \frac{(k-2)}{(n_H(k-1))^2}\left(\sum_{j=1}^{n_{H-1}} n_H^j \mathbb{1}\{i \neq C_j\}\right) \\
&= \frac{(k-2)(n_H - n_{H,i})}{(n_H(k-1))^2}
\end{aligned}$$

Similarly, computing covariance,

$$\begin{aligned}
\mathrm{Cov}((V_H)_i, (V_H)_j) &= \mathrm{Cov}\left(\frac{1}{n_H}\sum_{k=1}^{n_{H-1}}(Q_H)_i^k, \frac{1}{n_H}\sum_{\ell=1}^{n_{H-1}}(Q_H)_j^\ell\right) \\
&= \frac{1}{n_H^2}\sum_{k=1}^{n_{H-1}}\sum_{\ell=1}^{n_{H-1}}\mathrm{Cov}((Q_H)_i^k, (Q_H)_j^\ell) \qquad (1) \\
&= \frac{1}{n_H^2}\sum_{k=1}^{n_{H-1}}\mathrm{Cov}((Q_H)_i^k, (Q_H)_j^k) \qquad (2) \\
&= \frac{1}{n_H^2}\sum_{k=1}^{n_{H-1}}\left(-\frac{n_H^k}{(k-1)^2}\mathbb{1}\{i \neq C_k \cup j \neq C_k\}\right) \\
&= -\frac{(n_H - n_{H,i} - n_{H,j})}{(n_H(k-1))^2}
\end{aligned}$$

From (1) to (2), we used the fact that the covariance between the relative frequency of colors stemming from one node is independent of the relative frequency of colors stemming from another node (this reduces the expression from a double summation to a single summation). □

The next step is to define a classification algorithm. The algorithm for classification is rather simple and elegant. The above theorem allows us to ascertain details about the properties of this algorithm. But first, we present the algorithm explicitly.

ALGORITHM 1.6.2. CLASSIFY-ROOT

```
Summary: Takes in a tree $T$, and a height $H$. Outputs a prediction for the color of the root,
based on the relative frequency of nodes at the stated height.

function CLASSIFY-ROOT($T$, $H$)
    $w_H \leftarrow$ GETFREQ($T, H$)                    ▷ Obtain frequency distribution at height $H$
    $\hat{v}_0 \leftarrow (B^H)^{-1} w_H$
    prediction $\leftarrow$ argmax$_i$ $((\hat{v}_0)_i)$
    return prediction
```

Next, let's prove two important properties of this algorithm, which will give us some idea about it's effectiveness in practice.

THEOREM 1.6.3. The algorithm produces an unbiased estimate for vector $\hat{v}_0$ for the true frequency distribution of the root node v_0.

Proof. We know from THEOREM 1.2.2. that the expected distribution of colors on level H is described by the limiting distribution. That is, $\mathbb{E}[w_H] = v_H$. Then,

$$\mathbb{E}[\hat{v}_0] = \mathbb{E}[(B^H)^{-1} w_H] = (B^H)^{-1}\mathbb{E}[w_H] = (B^H)^{-1} v_H = v_0$$

□

Theorem 1.6.4. The covariance matrix of the estimated vector $\hat{v}_0$ is,

$$\mathrm{Cov}(\hat{v}_0) = \mathrm{Cov}((B^H)^{-1} w_H) = (B^H)^{-1} \Sigma ((B^H)^{-1})^\top$$

Where Σ is the covariance matrix defined in Theorem 1.2.2.[2]

Proof. It can be shown that $\mathrm{Cov}(A^\top X) = A^\top \Sigma_X A$, where $X \in \mathbb{R}^n$ is a random vector, and A is a conformable constant matrix. The result follows. □

Finally, based on these properties, it is possible to compute the confidence we have in our predictions. Recall that w_H, the sample proportion of colors at height H, is a linear transformation of multinomial random variables. Also, recall that the binomial distribution is well-approximated for large n by the normal density. An analogous approximation exists for the multinomial distribution and the multivariate normal density. More precisely, if n is large, and for a probability vector $\mathbf{p} = \langle p_1, \ldots, p_n \rangle$ [2],

$$\mathrm{Multi}(n, \mathbf{p}) \approx \mathcal{N}(n\mathbf{p}, \mathbf{P} - \mathbf{p}\mathbf{p}^\top)$$

Where $\mathbf{P} = \mathrm{diag}(p_1, \ldots, p_n)$. The above analysis (Theorem 1.6.1) simply aggregates all the multinomial random variables. Since the sum of independent multivariate normal random variables is itself a multivariate normal random variable, and since $\hat{v}_0$ is a linear transformation of w_H, we conclude that,

$$\hat{v}_0 \approx \mathcal{N}(v_0, (B^H)^{-1} \Sigma ((B^H)^{-1})^\top)$$

Then, to state the objective precisely, we would like to compute the quantity,

$$\mathrm{confidence} = P(\mathrm{argmax}_i((\hat{v}_0)_i) = k_0)$$

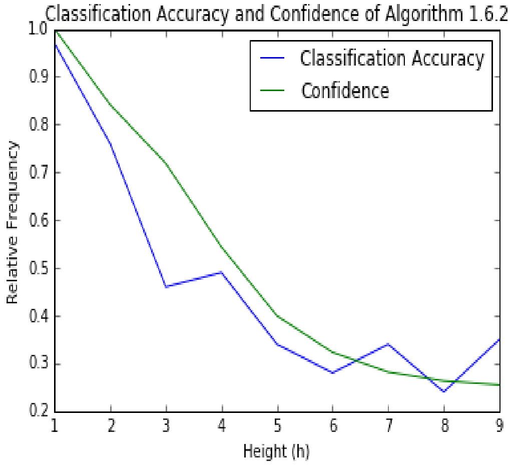
Where k_0 is the initial root color. While a closed-form solution is untenable, a simple bootstrap program in `Python` allows us to get estimates of confidence, which decrease as classification accuracy decreases. See the table below, and the Appendix for plotted results and more information about the code.

[2] For a more explicit covariance matrix formula, we note that,

$$(B^H)^{-1} = ((-I + \mathbf{1}\mathbf{1}^\top)/(k-1)^H)^{-1} = -(k-1)^H I + (k-1)^{H-1} \mathbf{1}\mathbf{1}^\top$$

This is computed by first noting that, $B^H = ((-I + \mathbf{1}\mathbf{1}^\top)/(k-1)^H)$, then evaluating the inverse. The process by which the inverse was attained is fairly mechanical, so I will not repeat it here. One should verify the result using the *Woodbury matrix identity*, which states that [3],

$$(A + UCV)^{-1} = A^{-1} - A^{-1} U (C^{-1} + V A^{-1} U)^{-1} V A^{-1}$$



k	H	Accuracy	Confidence
2	2	1.0	1.0
2	5	1.0	1.0
2	8	1.0	1.0
3	2	0.97	0.9937
3	5	0.70	0.9877
3	8	0.48	0.6983
4	2	0.74	0.8446
4	5	0.35	0.4028
4	8	0.24	0.2633
5	2	0.48	0.6176
5	5	0.26	0.2321
5	8	0.22	0.2015

2. *d*-ary Trees

2.1. Introduction

Consider a d-ary tree of size n. A k-coloring for $t = (V, E)$ is an assignment $V^{[k]}$ that is defined analogously to the random recursive tree considered in a previous section. The d-ary tree has the branching factor constraint that prevents us from directly reducing the tree to some version of the Polya's urn problem. However, by carefully divorcing *external nodes* and *internal nodes*, we can again rely on the theory of urns, albeit in a distinguished form from the random recursive tree.

An external node is defined as a node that has the potential to be added to the tree in the next step of the construction algorithm. Let the set of external nodes be defined by V_{EXT}. The set of internal nodes is defined by the relationship, $V_{\text{INT}} = V \backslash V_{\text{EXT}}$. An important well-known result about external nodes for general d-ary trees is now proved.

Claim 1. For any tree $t = (V, E)$ such that $|V| = n$, $V_{\text{EXT}} = n + 1$. That is, there are exactly $n(d-1)+1$ external nodes.

Proof. This follows by induction. The base case is the tree on 1 node. Since the outdegree of a node with no children in a binary tree is d, the claim is true for the base case. To see the inductive step argument, consider adding a node to a tree of size k. Since we fill one of the k available valid positions, but this creates d more valid positions for additional nodes by the definition of the structure, the tree on size $k+1$ has $k - 1 + d = k + (d-1)$ available places for an addition of a node, which is greater than the inductive hypothesis by the appropriate factor. □

In order to make the definitions and Claim 1 more palpable, consider the tree below.

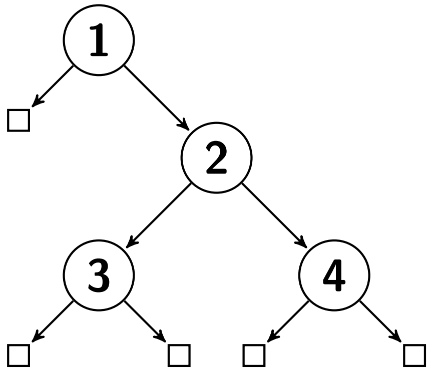


Figure 1: Binary tree $t = (V, E)$ (special case of d-ary tree, $d = 2$), with $|V| = 4$ and $V_{\text{EXT}} = 5$.

Consider the following construction procedure for a d-ary tree [1].

1. Start with a root node.
2. For a tree on $n - 1$ nodes t_{n-1}, there are $n(d-1)+1$ places at which we can choose to add an nth node, by CLAIM 1 above. Choose one node by the following probabilistic strategy: pick uniformly at random from the external nodes in the tree.

2.2. Limiting Frequencies of Colors: $k = 2$

Once again, we first consider the case of $k = 2$ colours, or the independent set problem. One approach that seems fruitful is to separate the external nodes into its own urn. This urn contains the external nodes, and is drawn from at each step. The replacement rule is as follows: replace the ball drawn by d of the *opposite color*. The ball-replacement matrix is therefore,

$$\begin{bmatrix} -1 & d \\ d & -1 \end{bmatrix}$$

These are modeled, once again, as balls, where the color of the ball is the color of the external node that may be added to the urn. This matrix is a tenable instance of the Bernard Friedman urn [4]. Using this reduction, and denoting by $\mathcal{E}_n^{(2)}(1)$ the number of external nodes of the first color,

$$\frac{\mathcal{E}_n^{(2)}(1) - \frac{1}{2}(d-1)n}{\sqrt{n}} \to_d \mathcal{N}\left(0, \frac{(d+1)^2}{4\left(1 + 2\left(\dfrac{d+1}{d-1}\right)\right)}\right)$$

The non-normalized version of this density will become useful rather shortly, so we compute this using elementary methods, and state the result.

$$\mathcal{E}_n^{(2)}(1) \to_d \mathcal{N}\left(\frac{1}{2}(d-1)n, \frac{n(d+1)^2}{4\left(1 + 2\left(\dfrac{d+1}{d-1}\right)\right)}\right)$$

While this is the distribution of the external nodes as $n \to \infty$, we would like to acquire a result that resembles those in the above sections by delineating the number of internal nodes. This can be stated more formally by defining the variable $C_n^{(2)}(1)$ as the number of balls of the first color that have been *drawn* after n steps. For this, a theorem and proof are necessary.

THEOREM 3.1. Consider the Bernard Friedman delineated by the ball-replacement matrix on the previous page. Denote by $\mathcal{E}_n^{(2)}(1)$ the number of balls of first color in the urn after n draws, and denote by $C_n^{(2)}(1)$ be the number of balls of first color *drawn* in n draws from the same urn. Then,

$$\frac{C_n^{(2)}(1) - \left(\frac{1}{2}(d-1)n + \dfrac{nd + C_0^{(2)}(1)}{d+1}\right)}{\sqrt{n}} \to_d \mathcal{N}\left(0, \frac{d+1}{4\left(1+2\left(\dfrac{d+1}{d-1}\right)\right)}\right)$$

Proof. The proof begins with a discussion of a transformation that maps $\mathcal{E}_n^{(2)}(1)$ to $C_n^{2}(1)$. It is intuitive that the former quantity is directly related to the latter quantity. A simple inductive argument gives that,

$$\begin{aligned} \mathcal{E}_n^{(2)}(1) &= C_0^{(2)}(1) - C_n^{(2)}(1) + C_n^{(2)}(2) \cdot d \\ &= C_0^{(2)}(1) - C_n^{(2)}(1) + (n - C_n^{(2)}(1)) \cdot d \end{aligned}$$

Where $C_0^{(2)}(1)$ is the "base case" — the number of balls of color one originally in the urn. Solving for $C_n^{(2)}(1)$, we obtain,

$$C_n^{(2)}(1) = \frac{nd + C_0^{(2)}(1) - \mathcal{E}_n^{(2)}(1)}{d+1}$$

This is an affine transformation of $\mathcal{E}_n^{(2)}$. Using the affine one-to-one change of variable formula for probability densities, and then normalizing the resulting density gives the result.[3] □

This is somewhat challenging to interpret as a standalone result, so another theorem is stated to provide more context on the result.

THEOREM 3.2. The proportion of internal nodes of first color is,

$$\frac{C_n^{(2)}(1)}{n} \to_P \frac{1}{2}$$

Proof. From (13),

$$\begin{aligned} \lim_{n\to\infty}\left(\frac{C_n^{(2)}(1)}{n}\right) &= \lim_{n\to\infty}\left(\frac{nd + C_0^{(2)}(1) - \mathcal{E}_n^{(2)}(1)}{n(d+1)}\right) \\ &= \lim_{n\to\infty}\left(\frac{nd}{n(d+1)}\right) + \lim_{n\to\infty}\left(\frac{C_0^{(2)}(1)}{n(d+1)}\right) - \lim_{n\to\infty}\left(\frac{\mathcal{E}_n^{(2)}(1)}{n(d+1)}\right) \\ &= \frac{d}{d+1} - \frac{1}{d+1}\lim_{n\to\infty}\left(\frac{\mathcal{E}_n^{(2)}(1)}{n}\right) \end{aligned}$$

[3] For $Y = aX + b$, $f_{aX+b}(y) = \frac{1}{|a|} f_X\left(\frac{y-b}{a}\right)$

The limit in this expression can be computed using a known result. Using a more general result on the Bagchi-Pal urn, of which the Bernard Friedman is a special case, it can be shown that [4],

$$\frac{\mathcal{E}_n^{(2)}(1)}{n} \to_P \frac{d-1}{2}$$

The proof is complete upon substitution. □

In words, the proportion of internal nodes of first color converges to 1/2 in the limit.

2.3. Limiting Frequencies of Colors: k arbitrary

Now, consider the $k > 2$ case. Once again, place all external nodes in the d-ary tree inside an urn. The ball-replacement matrix is,

$$\begin{bmatrix} -1 & dB_1 & dB_2 & \dots & dB_{k-1} \\ dB_1 & -1 & dB_2 & \dots & dB_{k-1} \\ & & \vdots & & \\ dB_1 & dB_2 & \dots & dB_{k-1} & -1 \end{bmatrix}$$

Intuitively, this means that we remove a ball which has the external node color from the urn, and replace it with d external nodes of one color from the remaining $(k-1)$ possibilities. It is implicit in the process that the internal node has the latter color. As with random recursive trees, $\langle B_1, \dots, B_{k-1} \rangle$ is an exchangeable random vector of Bernoulli $\{0, 1\}$ random variables such that $B_1 + B_2 + \cdots + B_{k-1} = 1$, and $\mathbb{E}[B_i] = 1/(k-1)$.

We know from the analysis on the RRT that the principal eigenvalue of the ball-replacement matrix is $(d-1)$. It can be shown that the eigenvector corresponding to the principal eigenvalue is $u = \vec{1}$, and therefore, $v = \frac{1}{k}\vec{1}$, since the principal eigenvector is ℓ_1 normalized in the formulation of Theorem 1.3.1. This allows us to obtain closed-form convergence and distributional results for the colors of the nodes in the tree as n grows. Denote by $\mathcal{E}_n^{(k)}(1)$ the number of external nodes in the urn of first color in the general k case. It is trivial to show that there exists a relation that resembles (12) in the $k > 2$ case.

$$\mathcal{E}_n^{(k)} \approx C_0^{(k)}(1) - C_n^{(k)}(1) + \frac{1}{k-1}\left(n - C_n^{(k)}(1)\right) d$$

Assuming that the color of the children of a particular node is selected uniformly at random from the $k-1$ possible colors, this approximation becomes exact in the limit. Reorganizing, we obtain,

$$C_n^{(k)}(1) \approx \frac{(k-1)C_0^{(k)}(1) + nd - (k-1)\mathcal{E}_n^{(k)}(1)}{k+d-1}$$

From (17), a result about proportions of internal nodes of a particular color, is derived in a similar fashion to the $k = 2$ case:

Theorem 3.5. The proportion of internal nodes of color i is,

$$\frac{C_n^{(k)}(i)}{n} \to_P \frac{1}{k}$$

Proof. Mahmoud shows that,

$$\frac{\mathcal{E}_n^{(k)}(i)}{n} \to \lambda_1 v_i$$

Where v_i is ith component of the eigenvector corresponding to the principal eigenvalue of the average generator [4]. From (17),

$$\begin{aligned}
\lim_{n\to\infty}\left(\frac{C_n^{(k)}(i)}{n}\right) &= \lim_{n\to\infty}\left(\frac{(k-1)C_0^{(k)}(i) + nd - (k-1)\mathcal{E}_n^{(k)}(i)}{n(k+d-1)}\right) \\
&= \lim_{n\to\infty}\left(\frac{(k-1)C_0^{(k)}(i)}{n(k+d-1)}\right) + \lim_{n\to\infty}\left(\frac{nd}{n(k+d-1)}\right) - \lim_{n\to\infty}\left(\frac{(k-1)\mathcal{E}_n^{(k)}(i)}{n(k+d-1)}\right) \\
&= \frac{d}{k+d-1} - \frac{k-1}{k+d-1}\lim_{n\to\infty}\left(\frac{\mathcal{E}_n^{(2)}(i)}{n}\right) \\
&= \frac{d}{k+d-1} + \frac{(d-1)(k-1)}{k(k+d-1)} \\
&= \frac{1}{k}
\end{aligned}$$

□

3. Appendix

3.1. Experimental Simulations

The code utilized for these simulations is available at `https://github.com/afrancis13/random-tree-colorings`. Each of the figures below belongs to a particular result in the paper. These were useful to confirm results and measure the performance of any algorithms defined in the document.

3.1.1. RRT

Figure 2 measures the limiting frequencies of colors in an RRT, and acts as empirical verification for *1.2, 1.3, 1.4*.

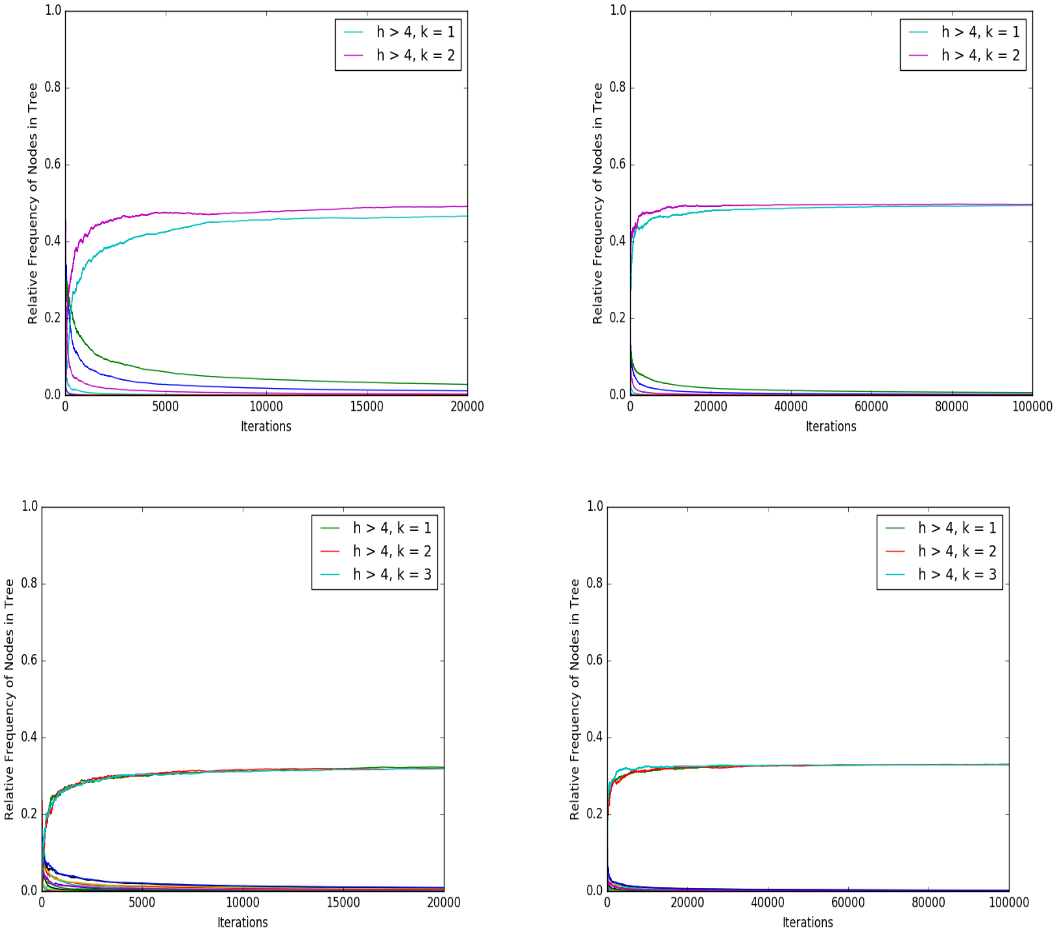


Figure 2: Simulation of Level $H = 4$, $k = 2$ for 20000 Iterations (Left) and 100000 Iterations (Right)

The following plot measures the *total variation distance*, a metric used to measure the distance between two probability distributions. This verifies the result in *1.5*.

Definition 3.1.1.1 **(Total Variation Distance)** For distributions P and Q on the same space, the total variation distance between P and Q is

$$D_V(P,Q) = \frac{1}{2}\sum_{\sigma} |P(\sigma) - Q(\sigma)|$$

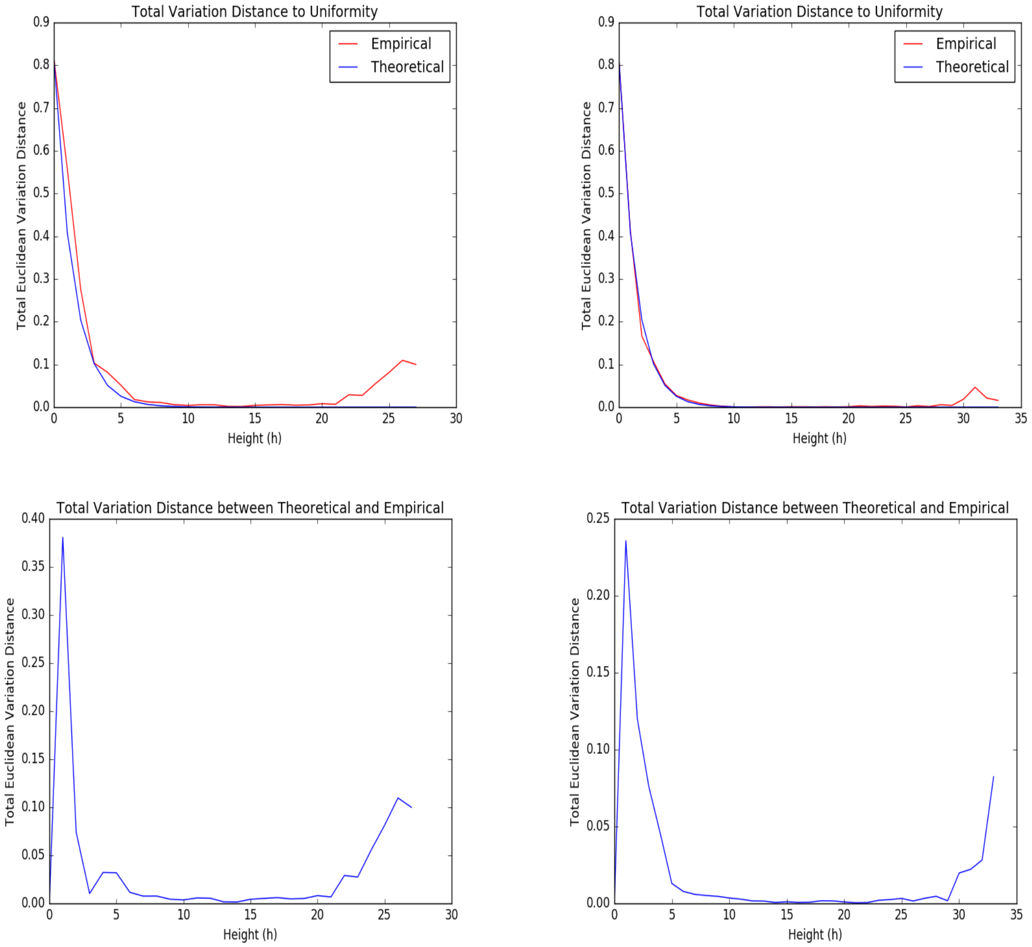


Figure 3: Simulations of Urn vs. Markov Process. The top panel plots the TVD between the empirical distribution and the uniform distribution in red and the TVD between the theoretical distribution from the Markov Chain and the uniform distribution in blue. The bottom panel plots the TVD between the empirical distribution and the theoretical distribution directly. The plots on the left side have simulation sample size $n = 10^6$ and the plots on the right have simulation sample size $n = 10^7$

Finally, the next plots are related to root classification accuracy and "confidence."

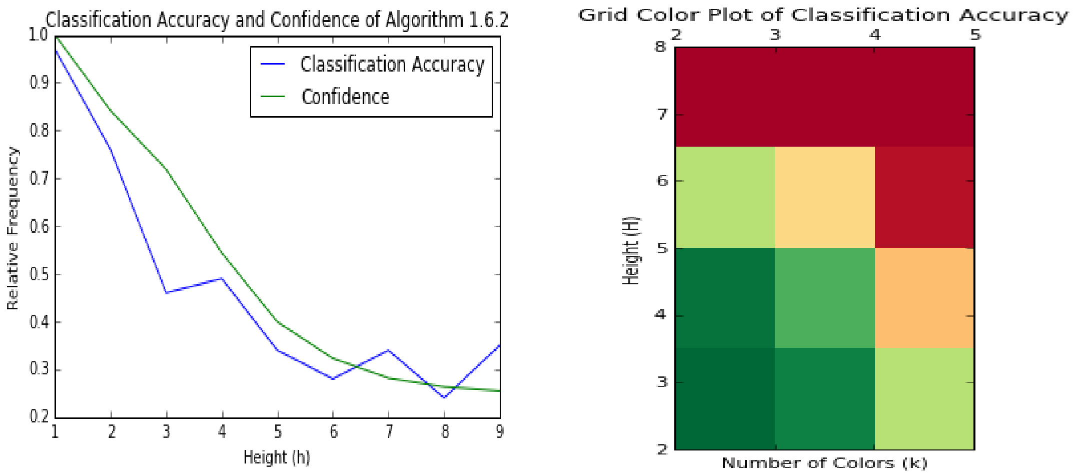


Figure 4: Decrease in confidence and classification accuracy as H increases for $k = 4$ (Left). Rectangular grid with two independent variables, H and k (Right). The right plot is colorized such that high accuracy points on the grid are green and low accuracy points on the grid are red. One can see that as k and H increase, the plot becomes more red.

3.1.2. d-ary Trees

There is only one plot included here, which delineates the convergence to color uniformity over all internal nodes. Additional total variation distance plots for specific H are included in the iPython notebook located at `https://github.com/afrancis13/random-tree-colorings`, but they are not included in this document, since this topic is not considered here.

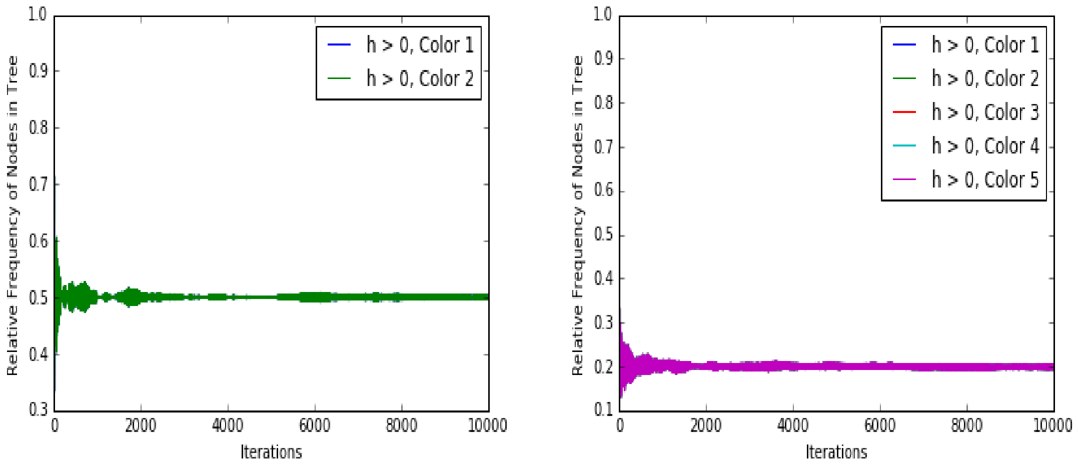


Figure 5: The limiting distribution of color relative frequencies converges to the uniform distribution for $d = 10, k = 2$ (Left) and $d = 2, k = 5$ (Right).

4. References

[1] Drmota, Michael. *Random Trees: An Interplay Between Combinatorics and Probability.* Wien: Springer, 2010.

[2] Geyer, C.J. *Stat 5101 Notes: Brand Name Distributions.* Reading. Minneapolis: Minnesota, United States of America.

[3] Higham, N.J. (2002). *Accuracy and Stability of Numerical Algorithms.* Philadelphia, PA: SIAM.

[4] Mahmoud, H. M. (2009). *Pólya urn models.* Boca Raton, Fla.: Chapman & Hall/CRC.

[5] Sun, R. *Lecture 4.* Reading, Singapore: National University of Singapore.

[6] Smythe, R.T. (1996). Central limit theorems for urn models. *Stochastic Processes and their Applications*, 65(1), 115137.