Poisson approximations for functionals of random trees

by Robert P. Dobrow and Robert T. Smythe[1]

Northeast Missouri State University and
George Washington University

Abstract

We use Poisson approximation techniques for sums of indicator random variables to derive explicit error bounds and central limit theorems for several functionals of random trees. In particular, we consider (i) the number of comparisons for successful and unsuccessful search in a binary search tree and (ii) internode distances in increasing trees. The Poisson approximation setting is shown to be a natural and fairly simple framework for deriving asymptotic results.

Keywords: Poisson approximation, binary search trees, increasing trees, recursive trees

[1]Research was carried out while the first author was a postdoctoral research associate at the National Institute of Standards and Technology, Statistical Engineering Division. The first author's institution will change its name to Truman State University in July, 1996.

1 Introduction and brief sketch of Poisson approximation

Many quantities of interest in the study of random trees can be naturally represented as sums of indicator random variables. In fortunate cases, these indicators are independent, but often they are not. Our interest here focuses on two different types of random trees: binary search trees and increasing trees. Binary search trees are described in detail by Mahmoud [14]; the class of increasing trees treated here includes, among others, recursive trees, plane-oriented recursive trees, and increasing binary trees. We use Poisson approximation techniques to study several quantities of algorithmic interest associated with these trees. The method not only gives explicit bounds for the error approximation, but also leads immediately to central limit theorems.

The Chen-Stein approach to Poisson approximation ([6], [7], [18]) provides a powerful tool for approximating probabilities by a Poisson distribution. A particular advantage of this approach is that it can often be used to approximate the distribution of sums of dependent indicator variables, when the dependence is local or otherwise sufficiently weak. Excellent accounts of the goals and methods of Poisson approximation may be found in the review paper by Arratia, Goldstein, and Gordon [1] and in the book by Barbour, Holst and Janson [4].

The total variation distance d_{TV} on probability measures P and Q over $\mathbf{Z}^+$ is defined by:

$$d_{\mathrm{TV}}(P,Q) := \sup_{A\subseteq\mathbf{Z}^+} |P(A)-Q(A)| = \frac{1}{2}\sum_{j\geq 0}|P(\{j\})-Q(\{j\})|.$$

Let $\mathrm{Po}(\lambda)$ denote the probability distribution of a Poisson random variable with parameter λ. Let $\mathrm{Po}(\lambda;A)$ denote the corresponding measure of event A. For a random variable X, let $\mathcal{L}(X)$ denote the distribution (law) of X. An important fact for our purposes ([4], p. 17) is that if W_n is a sequence of random variables for which

$$d_{\mathrm{TV}}(\mathcal{L}(W_n),\mathrm{Po}(\lambda_n)) \to 0 \ \text{ and } \lambda_n \to \infty,$$

then

$$\frac{W_n-\lambda_n}{\sqrt{\lambda_n}} \xrightarrow{d} \mathrm{N}(0,1),$$

where $\xrightarrow{d}$ denotes convergence in distribution and $\mathrm{N}(0,1)$ is a standard normal random variable.

Following [1], we describe our general setup. Given a finite or countable index set I, for each $\alpha \in I$, let X_α be a Bernoulli random variable with $p_\alpha := P(X_\alpha = 1) > 0$. Let

$$W := \sum_{\alpha \in I} X_\alpha, \qquad \lambda := \mathrm{E}[W] = \sum_{\alpha \in I} p_\alpha.$$

For each $\alpha \in I$, choose $B_\alpha \subseteq I$ with $\alpha \in B_\alpha$. We think of B_α as a "neighborhood" of α such that if $\beta \notin B_\alpha$, then X_α and X_β are almost (or exactly) independent. Define

$$b_1 := \sum_{\alpha \in I} \sum_{\beta \in B_\alpha} p_\alpha p_\beta,$$

$$b_2 := \sum_{\alpha \in I} \sum_{\substack{\beta \in B_\alpha \\ \beta \neq \alpha}} p_{\alpha\beta}, \quad \text{where } p_{\alpha\beta} := \mathrm{E}(X_\alpha X_\beta), \quad \text{and}$$

$$b_3 := \sum_{\alpha \in I} \mathrm{E}[|\mathrm{E}[X_\alpha - p_\alpha | X_\beta : \beta \notin B_\alpha]|].$$

The basic result we use is:

Theorem 1 (Arratia, et al. [1])

$$\begin{aligned} d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\lambda)) &\leq 2\left[(b_1 + b_2)\frac{1 - e^{-\lambda}}{\lambda} + b_3 \min\left(1, 1.4\lambda^{-1/2}\right)\right] \\ &\leq 2(b_1 + b_2 + b_3). \end{aligned}$$

When the X_α are independent, the factor of 2 in these inequalities may be removed ([4], p. 26), and by taking $B_\alpha = \{\alpha\}$, we have $b_1 = \sum_{\alpha \in I} p_\alpha^2$ and $b_2 = b_3 = 0$.

Theorem 1 will be our main tool for the results on increasing trees in Section 3. For the binary search trees in Section 2, we will need to approximate a sum by a mixed Poisson distribution, that is, a Poisson distribution whose parameter Λ is a random variable. The relevant result here is provided by [4], p. 12:

Theorem 2 *Let W have the mixed Poisson distribution* $\mathrm{Po}(\Lambda)$. *For $\lambda > 0$,*

(i) $d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \leq \min(1, \lambda^{-1/2})\mathrm{E}[|\Lambda - \lambda|]$, *and*

(ii) $d_{\mathrm{TV}}(\mathcal{L}(W), \mathrm{Po}(\lambda)) \leq \dfrac{1 - e^{-\lambda}}{\lambda}\mathrm{Var}[\Lambda]$, *if* $\lambda = \mathrm{E}[\Lambda]$.

2 Unsuccessful and successful search in a binary search tree

A binary search tree on n nodes is a binary tree labeled with the elements of a set of keys, which, without loss of generality, we take to be $\{1, 2, \ldots, n\}$. We assume a random permutation model. In brief, the distribution of trees under this model is the distribution induced by building a binary search tree from a uniformly random permutation. (See [14] for details.) Let $H_k := \sum_{j=1}^{k} j^{-1}$ be the kth harmonic number.

2.1 Unsuccessful search

Consider the number of comparisons U_n in an unsuccessful attempt to find a key in a binary search tree with n keys. This is precisely the level (or depth) of insertion of the $(n+1)$-st key, when the root is taken at level 0. A result of Lynch [11] tells us that

$$P(U_n = k) = \frac{2^k}{(n+1)!} s(n,k),$$

where $s(n,k)$ is the signless Stirling number of the first kind. The probability generating function of U_n is thus given by

$$G_n(z) = \sum_{k \geq 0} z^k P(U_n = k) = \frac{1}{(n+1)!} \prod_{k=1}^{n} (2z + k - 1),$$

which is the probability generating function of a sum of independent Bernoulli random variables with parameters $p_k := 2/(k+1)$, $1 \leq k \leq n$. Thus by Theorem 1:

Theorem 3 *Let $\lambda_n := 2(H_{n+1} - 1)$. Then for $n \geq 1$,*

$$d_{\mathrm{TV}}(\mathcal{L}(U_n), \mathrm{Po}(\lambda_n)) \leq \frac{1}{\lambda_n} \sum_{i=1}^{n} p_i^2 \leq \frac{4}{\lambda_n} \left(\frac{\pi^2}{6} - 1 \right).$$

As a corollary we get that U_n satisfies a central limit theorem, a result due to Brown and Shubert [5].

Corollary 2.1

$$\frac{U_n - 2 \ln n}{\sqrt{2 \ln n}} \xrightarrow{d} N(0,1).$$

It turns out in this case that $(\ln n)^{-1}$ is exactly the correct rate of the error of Poisson approximation. By [3],

$$d_{\mathrm{TV}}(\mathcal{L}(U_n), \mathrm{Po}(\lambda_n)) \geq \frac{1}{8\lambda_n}\left(\frac{\pi^2}{6} - 1\right).$$

Remark: As pointed out by a referee, the distribution of U_n arises as a special case of the Ewens Sampling Formula (ESF). In that context, Theorem 3 is given in [2]. In the standard notation of the ESF, our example arises as a special case corresponding to $\theta = 2$.

2.2 Successful search

Let S_n be the number of comparisons needed to insert a key chosen uniformly at random from the first n keys. Let η denote a key chosen uniformly at random. Then $S_n = D_\eta$, where D_m is the depth of insertion of key m. By our earlier observation, $S_n = U_{\eta-1}$. With $U_0 = 0$, we have

$$P(U_{\eta-1} = k) = \sum_{j=1}^{n} P(U_{j-1} = k | \eta = j) P(\eta = j) = \frac{1}{n}\sum_{j=0}^{n-1} P(U_j = k).$$

For any $A \subseteq \mathbf{Z}^+$ and $j \geq 1$,

$$|P(U_j \in A) - \mathrm{Po}(\lambda_j; A)| \leq C(H_{j+1} - 1)^{-1},$$

by Theorem 3, where $C = (\pi^2 - 6)/3$. Let $\mathrm{Po}(\lambda_\eta)$ denote the distribution of a mixed Poisson random variable where

$$\lambda_\eta := 2\sum_{j=1}^{\eta} \frac{1}{j+1},$$

and η is distributed uniformly on $\{1, 2, \ldots, n\}$. Let $\lambda_0 := 0$ and $\mathrm{Po}(0)$ be point mass at 0. Then for $n \geq 2$,

$$\begin{aligned} |P(U_{\eta-1} \in A) - \mathrm{Po}(\lambda_{\eta-1}; A)| &\leq \frac{1}{n}\sum_{j=0}^{n-1} |P(U_j \in A) - \mathrm{Po}(\lambda_j; A)| \\ &\leq \frac{C}{n}\sum_{j=1}^{n-1} \frac{1}{H_{j+1} - 1} \leq 11C\lambda_n^{-1}. \end{aligned}$$

Hence

$$d_{\mathrm{TV}}(\mathcal{L}(S_n), \mathrm{Po}(\lambda_{\eta-1})) \leq 11C\lambda_n^{-1}. \tag{1}$$

The next step is to compare $\mathrm{Po}(\lambda_{\eta-1})$ and $\mathrm{Po}(\lambda_n)$ using Theorem 2(ii). Instead of λ_n we take

$$\lambda_n^* := E[\lambda_{\eta-1}] = \frac{1}{n}\sum_{k=1}^{n-1}\lambda_k = \lambda_n + O(1). \tag{2}$$

With the help of Maple, $\mathrm{Var}[\lambda_{\eta-1}] = 8 + o(1)$ and, for $n \geq 2$, $\mathrm{Var}[\lambda_{\eta-1}] \leq 8$. Theorem 2 gives

$$d_{\mathrm{TV}}(\mathrm{Po}(\lambda_{\eta-1}), \mathrm{Po}(\lambda_n^*)) \leq 8(\lambda_n^*)^{-1} \leq 16\lambda_n^{-1}. \tag{3}$$

Combining (1) and (3) we have

Theorem 4

$$d_{\mathrm{TV}}(\mathcal{L}(S_n), \mathrm{Po}(\lambda_n^*)) \leq 32\lambda_n^{-1}.$$

From (2), $\lambda_n^* = 2\ln n + O(1)$. Louchard's central limit theorem for S_n ([10]; see also [14], p. 80) is a direct consequence.

3 Increasing trees

In this section we use Poisson approximation to obtain limiting distributions for internode distances in several classes of increasing trees. We also derive some general results for increasing trees which may be of independent interest.

An *increasing tree* of size n is a rooted tree labeled by distinct integers $\{1, 2, \ldots, n\}$ such that the sequence of labels along any root-to-leaf path is increasing. One can consider increasing trees as growing dynamically: At time n, tree T_n is formed by joining node n to tree T_{n-1}, that is, node n becomes a child of one of the nodes of T_{n-1}. There are numerous varieties of such trees. Below we introduce a parametrization which includes many well-studied examples, such as recursive trees [17] and binary trees.

In working with increasing trees it is convenient to consider an extension of these trees obtained by adding a different type of node called *external* at each possible insertion point. Thus at time n, tree T_n is formed when one of the external nodes of T_{n-1} becomes an internal node (node n) and additional external nodes are created at node n and possibly at other nodes. The usual model of randomness on the space of increasing trees is the uniform model,

i.e., all trees are equally likely. Given a random tree T_{n-1} on $n-1$ nodes, one obtains a random tree on n nodes by choosing an external node uniformly at random among all the external nodes of T_{n-1}. The selected external node becomes an internal node n in T_n and additional external nodes are created depending on the growth rule of the tree.

The distance $D_{i,j}$ between nodes i and j in a random recursive tree of order n was studied by Moon [15] who found the expectation and variance of $D_{i,j}$. Dobrow [9] gives exact and asymptotic formulas for the distribution of $D_{i,j}$. Mahmoud [12], Devroye [8], and others have studied the height of the last node inserted in T_n, that is, $D_{1,n}$.

Let T_n denote a random increasing tree on n nodes. The trees we consider are governed by a deterministic growth rule: At time n, a fixed number of new external nodes are created at node n and possibly at the parent of node n. Let α denote the number of new external nodes created at the parent of node n. Let β denote the number of external nodes created at node n. The number of external nodes in T_n, which we denote by X_n, is given by

$$X_n = \begin{cases} \beta, & n = 1 \\ X_{n-1} + \alpha + \beta - 1, & n \geq 2. \end{cases}$$

Hence

$$X_n = \beta + (n-1)(\alpha + \beta - 1), \quad n \geq 1. \tag{4}$$

Table 1 identifies the above parameters for several increasing tree models. Plane-oriented trees were first introduced by Szymański [19] and further studied by Mahmoud [13]. For such trees a new node is added with probability proportional to 1 plus the outdegree of its parent. Pittel [16] considered a generalization of such trees where a new node is added with probability proportional to 1 plus a constant times the outdegree of its parent. Our parametrization includes such trees, which we call m-oriented trees.

Table 1.

Tree	α	β	X_n
Recursive	1	1	n
Plane-oriented	2	1	$2n-1$
m-oriented	m	1	$m(n-1)+1$
Binary	0	2	$n+1$
m-ary	0	m	$(m-1)n+1$

Remark: In the examples we treat in this paper, the parameters α and β are both constant. However, similar results can be obtained by letting them depend on n. That is, the growth rule at time n depends on the *label* of the node to which node n is joined. For instance, letting $\alpha = 0$ and $\beta_k \equiv k$ gives a tree where node k can have at most k children. It is interesting that the internode distances for such a tree behave like those of an increasing binary tree. In particular, the expected distance between the root and node n is about $2 \ln n$.

3.1 Exact distribution of distances

For $j > i$, write $\{j <_c i\}$ for the event that node j is a child of node i.

Theorem 5

$$P(j <_c i) = \frac{\beta}{X_i} \prod_{k=i+1}^{j-1} \left(1 - \frac{\beta}{X_k}\right) = P(i+1 <_c i) \prod_{k=i+1}^{j-1} P(k+1 \not<_c k). \quad (5)$$

Proof The second equality is immediate from the first. By conditioning on T_{j-1},

$$P(j <_c i) = \frac{\mathrm{E}[\beta_{i,j-1}]}{X_{j-1}}, \quad (6)$$

where $\beta_{r,s} :=$ the (random) number of external nodes of r in T_s $(1 \le r \le s)$. Consider $\beta_{i,j}$. If j is not a child of i, then $\beta_{i,j} = \beta_{i,j-1}$. If j is a child of i, then $\beta_{i,j} = \beta_{i,j-1} + \alpha - 1$. Thus

$$\begin{aligned} \mathrm{E}[\beta_{i,j}] &= \mathrm{E}[\beta_{i,j-1}]P(j \not<_c i) + (\mathrm{E}[\beta_{i,j-1}] + \alpha - 1)P(j <_c i) \\ &= \mathrm{E}[\beta_{i,j-1}] + (\alpha - 1)\frac{\mathrm{E}[\beta_{i,j-1}]}{X_{j-1}} \\ &= \left(\frac{X_j - \beta}{X_{j-1}}\right)\mathrm{E}[\beta_{i,j-1}] = \beta \prod_{k=i+1}^{j} \frac{X_k - \beta}{X_{k-1}}. \end{aligned}$$

With (6) the result follows. ∎

Corollary 3.1 *For* $1 \le i < j - 1$,

$$\frac{P(j <_c i)}{P(j-1 <_c i)} = P(j \not<_c j-1).$$

Our main result in this subsection, Theorem 6, is a characterization of the distribution of $D_{i,n}$ for arbitrary i. For independent random variables X and Y we write $X \oplus Y$ for the sum of X and Y. Let $\mathrm{Be}(p)$ denote a Bernoulli random variable with success probability p. Let $\mathbf{1}(A)$ denote the indicator of the event A.

Theorem 6

$$\mathcal{L}(D_{i,n}) = \mathcal{L}\left(D_{i,i+1} \oplus \oplus_{j=i+1}^{n-1} \mathrm{Be}(p_j)\right), \quad i < n, \tag{7}$$

where $p_j := P(j+1 <_c j) = \beta/X_j$.

Proof Without loss of generality, assume that $i \leq n-2$. By conditioning on the parent of node n,

$$\begin{aligned} P(D_{i,n} = d) &= \sum_{k=1}^{n-1} P(D_{i,k} = d-1 | n <_c k) P(n <_c k) \\ &= \sum_{k=1}^{n-2} P(D_{i,k} = d-1 | n <_c k) P(n <_c k) \\ &\quad + P(D_{i,n-1} = d-1 | n <_c n-1) P(n <_c n-1). \end{aligned}$$

Note that for $i, k \leq n-2$, the distribution of $D_{i,k}$ conditional on $\{n <_c k\}$ is equal to the distribution of $D_{i,k}$ conditional on $\{n-1 <_c k\}$. Also, for fixed $i \leq n-1$, the random variables $D_{i,n-1}$ and $\mathbf{1}(n <_c n-1)$ are independent. These observations together with Corollary 3.1 give

$$\begin{aligned} P(D_{i,n} = d) &= P(n \not<_c n-1) \sum_{k=1}^{n-2} P(D_{i,k} = d-1 | n-1 <_c k) P(n-1 <_c k) \\ &\quad + P(n <_c n-1) P(D_{i,n-1} = d-1) \\ &= P(n \not<_c n-1) P(D_{i,n-1} = d) \\ &\quad + P(n <_c n-1) P(D_{i,n-1} = d-1). \end{aligned}$$

Thus

$$\mathcal{L}(D_{i,n}) = \mathcal{L}(D_{i,n-1} \oplus \mathrm{Be}(p_{n-1}))$$

and the result follows. ■

Remark: In the case of recursive trees the random variables $D_{i,j}$ and $\mathbf{1}(n <_c j)$ are independent for $i, j < n$. This, however, is not true in general. Consider, for instance, an increasing binary tree. The event $\{4 <_c 1\}$ implies $\{D_{1,3} = 2\}$.

For root-to-last node distances, we explicitly identify the Bernoulli random variables in (7):

Theorem 7

$$D_{1,n} = \sum_{k=1}^{n-1} \mathbf{1}(A_k),$$

where A_k is the event that node k is on the path from the root to node n. Furthermore,

$$P(A_k) = P(k+1 <_c k) = \frac{\beta}{X_k}, \quad 1 \le k \le n-1, \tag{8}$$

and the random variables $\mathbf{1}(A_1), \ldots, \mathbf{1}(A_{n-1})$ are independent.

Proof The first sentence of the theorem is clear. For $i < j$, let $\{j <_d i\}$ denote the event that node j is a descendant of node i. Then $A_k = \{n <_d k\}$. Independence follows from the fact that the location where a new node joins the tree is independent of any previous joins.

To show (8), we show $P(n <_d j) = P(j+1 <_c j)$ by induction on $n - j$. The basis case $n = j + 1$ is trivial. Assume $n > j + 1$. By considering the parent of node n,

$$\begin{aligned} P(n <_d j) &= P(n <_c j) + \sum_{k=j+1}^{n-1} P(n <_c k, k <_d j) \\ &= P(j+1 <_c j) \left(\frac{P(n <_c j)}{P(j+1 <_c j)} + \sum_{k=j+1}^{n-1} P(n <_c k) \right), \end{aligned}$$

where we have used independence and the induction hypothesis for the last equality. By Theorem 5,

$$\begin{aligned} \frac{P(n <_c j)}{P(j+1 <_c j)} &= \prod_{k=j+1}^{n-1} P(k+1 \not<_c k) \\ &= P(\cap_{k=j+1}^{n-1} \{k+1 \not<_c k\}). \end{aligned}$$

Also

$$\begin{aligned}\sum_{k=j+1}^{n-1} P(n <_c k) &= P(\cup_{k=j+1}^{n-1}\{n <_c k\}) \\ &= 1 - P(\cap_{k=j+1}^{n-1}\{n \not<_c k\}).\end{aligned}$$

It thus remains to show that

$$P(\cap_{k=j+1}^{n-1}\{k+1 \not<_c k\}) = P(\cap_{k=j+1}^{n-1}\{n \not<_c k\}).$$

We do this by (backward) induction on j. The case $j = n-2$ is trivial. For general j,

$$\begin{aligned}&P(\cap_{k=j+1}^{n-1}\{n \not<_c k\}) \\ &\quad = P(\cap_{k=j+2}^{n-1}\{n \not<_c k\}) - P(\{n <_c j+1\} \cap \cap_{k=j+2}^{n-1}\{n \not<_c k\}) \\ &\quad = P(\cap_{k=j+2}^{n-1}\{k+1 \not<_c k\}) - P(n <_c j+1),\end{aligned}$$

using the induction step and the fact that $\{n <_c j+1\} \subseteq \cap_{k=j+2}^{n-1}\{n \not<_c k\}$. Now use Theorem 5 to compute $P(n <_c j+1)$ and the result follows. ■

It is straightforward to derive the exact distribution and moments of $D_{1,n}$ for the increasing trees listed in Table 1. The distributions all involve the signless Stirling numbers of the first kind. To compute the exact distribution, take the probability generating function $\mathrm{E}[x^{D_{1,n}}]$ of $D_{1,n}$ and use the fact that for nonnegative integers n and k, $s(n,k)$ is the coefficient of x^k in the product $x(x+1)\cdots(x+n-1)$. We omit the details and collect results for root-to-last node distances in Tables 2 and 3. Results for recursive and plane-oriented recursive trees can also be found in [12], [13]. Let $H_k^{(2)} := \sum_{j=1}^{k} j^{-2}$. For real r and integer k define $(r)_k := r(r-1)\cdots(r-k+1)$.

Table 2.

Tree	$P(D_{1,n} = d)$
Recursive	$[(n-1)!]^{-1}s(n-1,d)$
Plane-oriented	$[2^d(n-3/2)_{n-1}]^{-1}s(n-1,d)$
m-oriented	$[m^d(n-(2m-1)/m)_{n-1}]^{-1}s(n-1,d)$
Binary	$(2^d/n!)s(n-1,d)$
m-ary	$((m/(m-1))^d[(n-(m-2)/(m-1))_{n-1}]^{-1}s(n-1,d)$

Table 3.

Tree	$\mathrm{E}[D_{1,n}]$	$\mathrm{Var}[D_{1,n}]$
Recursive	H_{n-1}	$H_{n-1} - H_{n-1}^{(2)}$
Plane-oriented	$H_{2n-3} - \frac{1}{2}H_{n-2}$	$H_{2n-3} - H_{2n-3}^{(2)} - \frac{1}{2}(H_{n-2} - H_{n-2}^{(2)})$
m-oriented	$\sim \frac{1}{m}\log n$	$\sim \frac{1}{m}\log n$
Binary	$2(H_n - 1)$	$2H_n - 4H_n^{(2)} + 2$
m-ary	$\sim \frac{m}{m-1}\log n$	$\sim \frac{m}{m-1}\log n$

3.2 Limiting distributions

For all of the increasing trees discussed in the previous subsection, the distribution of internode distances is asymptotically normal. This follows from estimating the discrepancy between the distributions of $D_{i,n}$ and a Poisson variate. Key quantities in estimating this discrepancy are the first two moments of $D_{i,i+1}$. For general increasing trees we do not know the distribution, or even the first moment, of $D_{i,i+1}$. Thus our estimate is crude, although we can still show asymptotic normality. Theorem 8 gives more precise results for the case of recursive trees. The proof of Theorem 8 will require three lemmas, which we state without proof. For a random recursive tree, Lemma 3.1, first shown by Moon [15], gives the first two moments of $D_{i,n}$. Lemma 3.2 characterizes the distribution of $D_{i,i+1}$ as a mixture of sums of independent Bernoulli random variables. Lemma 3.3, which gives a general bound for total variation distance of convolutions, is well-known.

Lemma 3.1 (Moon, [15]) *For a random recursive tree,*

$$\mathrm{E}[D_{i,n}] = H_i + H_{n-1} - 2 + (1/i).$$

$$\mathrm{Var}[D_{i,n}] = H_i + H_{n-1} - 3H_i^{(2)} - H_{n-1}^{(2)} + 4 - 4H_i/i + (3/i) - (1/i)^2.$$

Lemma 3.2 (Dobrow, [9]) *Let δ_k denote point mass at k.*

$$\mathcal{L}(D_{i,i+1}) = \frac{1}{i}\sum_{k=1}^{\min(i,2)}\delta_k + \sum_{j=0}^{i-3}\frac{2}{(i-j)(i-j-1)}\mathcal{L}\left(3 + \sum_{k=0}^{j-1}\mathrm{Be}\left(\frac{2}{i-k}\right)\right).$$

Lemma 3.3 *Let W, X, Y, Z be random variables such that W, X are independent and Y, Z are independent. Then*

$$d_{\mathrm{TV}}(\mathcal{L}(W+X), \mathcal{L}(Y+Z)) \le d_{\mathrm{TV}}(\mathcal{L}(W), \mathcal{L}(Y)) + d_{\mathrm{TV}}(\mathcal{L}(X), \mathcal{L}(Z)).$$

We proceed with an approximation for the distribution of $D_{i,n}$. It is of interest to know the behavior of $D_{i,n}$ when $i = i_n \to \infty$. We give two estimates to handle the full range of i_n.

Theorem 8 *For a recursive tree, let* $\lambda_n := \mathrm{E}[D_{i,n}] = H_{n-1} + H_i - 2 + (1/i)$. *Then for* $1 \le i < n$,

$$\text{(i) } d_{\mathrm{TV}}(\mathcal{L}(D_{i,n}), \mathrm{Po}(\lambda_n)) \le \frac{16(\ln i)^2 + 44}{\ln(n-1) + \ln i - 2}.$$

For $2 \le i < n$,

$$\text{(ii) } d_{\mathrm{TV}}(\mathcal{L}(D_{i,n}), \mathrm{Po}(\lambda_n)) \le \frac{H_{n-1}^{(2)} - H_i^{(2)}}{H_{n-1} - H_i} + \frac{11}{\ln i}.$$

Proof (i) Let A_k denote the event that node k is on the path from node i to node $i+1$. Write

$$D_{i,i+1} = \sum_{k=1}^{i} I(A_k).$$

By Theorem 6, it follows that

$$D_{i,n} \stackrel{d}{=} \sum_{j=1}^{i} Y_j + \sum_{j=i+1}^{n-1} Z_j, \tag{9}$$

where $Y_j \stackrel{d}{=} I(A_j)$ and the Y_j are independent of $\{Z_{i+1}, \ldots, Z_{n-1}\}$. Further, the Z_j are independent and $Z_j \stackrel{d}{=} \mathrm{Be}(p_j)$, where $p_j := 1/j$. Let $\pi_j := P(A_j)$ for $j = 1, \ldots, i$. Applying Theorem 1, let

$$B_j := \begin{cases} \{1, 2, \ldots, i\}, & \text{for } j = 1, 2, \ldots, i \\ \{j\}, & \text{for } j = i+1, \ldots, n-1. \end{cases}$$

We have

$$\begin{aligned} b_1 &= \sum_{j=1}^{i} \sum_{k=1}^{i} \pi_j \pi_k + \sum_{j=i+1}^{n-1} p_j^2 = (\mathrm{E}[D_{i,i+1}])^2 + \sum_{j=i+1}^{n-1} j^{-2} \\ &= 4H_i^2 - 8H_i + 4H_i/i + 4 - (4/i) + (1/i)^2 + H_{n-1}^{(2)} - H_i^{(2)}, \end{aligned}$$

$$b_2 = \mathrm{E}[D_{i,i+1}^2] - \mathrm{E}[D_{i,i+1}] = 4H_i^2 - 8H_i - 4H_i^{(2)} + 10 - (2/i),$$

and $b_3 = 0$. The first result follows, using $\ln i \le H_i \le \ln i + 1$.

(ii) We first derive approximations for $D_{i,i+1}$ and $\sum_{j=i+1}^{n-1} \mathrm{Be}(1/j)$. Then Theorem 6 and Lemma 3.3 will give the result.

Let $X_1, X_2, \ldots, X_i$ be independent Bernoulli random variables with

$$p_j := P(X_j = 1) = \begin{cases} 1, & j = 1, 2, 3 \\ 2/(i - j + 4), & j = 4, 5, \ldots, i. \end{cases}$$

For $1 \le k \le i$, let $W_k := \sum_{j=1}^{k} X_j$. Further, let η be a random variable, independent of the X_j's,with probability mass function

$$P(\eta = t) = \begin{cases} 1/i, & t = 1, 2 \\ 2/((i - t + 3)(i - t + 2)), & t = 3, 4, \ldots, i. \end{cases}$$

Then, according to Lemma 3.2,

$$\mathcal{L}(D_{i,i+1}) = \mathcal{L}(W_\eta).$$

Let $\tilde{\lambda}_i := E[W_i]$. Straightforward calculations give $\tilde{\lambda}_i = 2H_i - (2/3)$ and

$$\sum_{k=1}^{i} p_k^2 = 4H_i^{(2)} - (22/9).$$

Thus for $i \ge 2$,

$$d_{\mathrm{TV}}(\mathcal{L}(W_i), \mathrm{Po}(\tilde{\lambda}_i)) \le \frac{4H_i^{(2)} - (22/9)}{2H_i - (2/3)} \le C(\ln i)^{-1},$$

where $C = (3\pi^2 - 11)/9$. Let $\tilde{\lambda}_\eta := \sum_{k=1}^{\eta} p_k$. Then

$$\begin{aligned} & |P(W_\eta \in A) - \mathrm{Po}(\tilde{\lambda}_\eta; A)| \\ & \le \sum_{t=1}^{i} d_{\mathrm{TV}}(\mathcal{L}(W_t), \mathrm{Po}(\tilde{\lambda}_t)) P(\eta = t) \\ & \le C \left(\frac{\ln 2 + 1}{(\ln 2) i} + 2 \sum_{t=3}^{i} \frac{1}{(\ln t)(i - t + 3)(i - t + 2)} \right) \qquad (10) \\ & \le \frac{7C + 1}{2 \ln i}, \end{aligned}$$

where the last inequality is derived by splitting the sum in (10) into $\sum_{t=3}^{\lfloor i/2 \rfloor}$ and $\sum_{t=\lfloor i/2 \rfloor+1}^{i}$ and bounding the $1/(\ln t)$ factor by its maximum over the range of summation. We now have

$$d_{\mathrm{TV}}(\mathcal{L}(D_{i,i+1}), \mathrm{Po}(\tilde{\lambda}_\eta)) \le \frac{7C+1}{2\ln i}. \tag{11}$$

Now take $\lambda_i^* = E[\tilde{\lambda}_\eta] = 2H_i - 2 + (1/i)$. Direct calculation (which we omit) shows that for $n \ge 9$,

$$\mathrm{E}[\tilde{\lambda}_\eta^2] \le 4H_i^2 - 8H_i + 8$$

and thus $\mathrm{Var}[\tilde{\lambda}_\eta] \le 4$. By Theorem 2(ii), for $i \ge 2$,

$$d_{\mathrm{TV}}(\mathrm{Po}(\tilde{\lambda}_\eta), \mathrm{Po}(\lambda_i^*)) \le 4(2H_i - 2 + (1/i))^{-1} \le 2.5(\ln i)^{-1}.$$

Together with (11), we have

$$d_{\mathrm{TV}}(\mathcal{L}(D_{i,i+1}), \mathrm{Po}(\lambda_i^*)) \le \frac{7C+6}{2\ln i} \le \frac{11}{\ln i}.$$

Finally, by Poisson approximation for sums of independent indicators,

$$d_{\mathrm{TV}}(\mathcal{L}(\oplus_{j=i+1}^{n-1}\mathrm{Be}(1/j)), \mathrm{Po}(H_{n-1} - H_i)) \le \frac{H_{n-1}^{(2)} - H_i^{(2)}}{H_{n-1} - H_i}.$$

Applying Lemma 3.3 gives (ii). ∎

Remarks:

1. Theorem 8 implies that a central limit theorem holds for $D_{i,n}$ in the full range of i. Dobrow [9] also shows asymptotic normality for $D_{i,n}$ by using moment generating functions.

2. Consider the representation (9) of $D_{i,n}$. Roughly, all of the dependency is in $Y_1, \ldots, Y_i$. Thus it is reasonable that as i grows the accuracy of the approximation (i) will diminish. For i constant or $i_n = O(\ln n)$, (i) gives a better bound than (ii). But for $i_n = \Omega(n)$, say, (ii) is better.

3. We omit the case $i = 1$ in Theorem 8 for technical reasons. In any case, for $i = 1$, the approximation of Theorem 9 below gives a better bound.

For general increasing trees we do not have a distributional representation for the distribution of $D_{i,i+1}$ similar to Lemma 3.2. However, in the following

theorem, Poisson approximation for sums of independent indicators affords an effortless result for root-to-last node distances. Devroye [8] proves a central limit theorem for $D_{1,n}$ for recursive trees by other means.

Theorem 9 *Let*

$$\lambda_n := \beta \sum_{j=1}^{n-1} \frac{1}{(\alpha+\beta-1)(j-1)+\beta} = \left(\frac{\beta}{\alpha+\beta-1}\right) \ln n + O(1).$$

Then

$$\begin{aligned} d_{\mathrm{TV}}(\mathcal{L}(D_{1,n}), \mathrm{Po}(\lambda_n)) &\leq \frac{\beta^2}{\lambda_n} \sum_{j=1}^{n-1} \frac{1}{((\alpha+\beta-1)(j-1)+\beta)^2} \\ &\leq \frac{1}{\lambda_n} + \left(\frac{\pi\beta}{\alpha+\beta-1}\right)^2 \frac{1}{6\lambda_n}. \end{aligned}$$

For fixed i, an argument similar to that given in Theorem 8 holds for general increasing trees. However, we must use the crude bounds

$$1 \leq D_{i,i+1} \leq D_{1,i} + D_{1,i+1}.$$

In all of the tree examples of the previous subsection

$$p_j = \frac{\beta}{(\alpha+\beta-1)(j-1)+\beta}$$

and thus

$$\sum_{j=i+1}^{n-1} p_j \approx \left(\frac{\beta}{\alpha+\beta-1}\right) (\ln n - \ln i).$$

This gives a bound in the Poisson approximation of order $(\ln n)^{-1}$.

4 Acknowledgements

The authors would like to thank the referees for helpful comments and suggestions and Jim Fill for a careful reading of the manuscript which led to the removal of several errors and an improved exposition.

References

[1] R. Arratia, L. Goldstein, and L. Gordon, Poisson approximation and the Chen-Stein method, *Stat. Sci.* **5**, 403–434 (1990).

[2] R. Arratia and S. Tavaré, Limit theorems for combinatorial structures via a discrete process approximations. *Rand. Struct. Alg.* **3**, 321–345 (1992).

[3] A. D. Barbour and P. Hall, On the rate of Poisson convergence, *Math. Proc. Camb. Phil. Soc.* **95**, 473–480 (1984).

[4] A. D. Barbour, L. Holst, and S. Janson, *Poisson Approximation,* Oxford University Press, Oxford, (1992).

[5] G. Brown and B. Shubert, On random binary trees, *Math. Oper. Res.* **9**, 43–65 (1984).

[6] L. H. Y. Chen, Poisson approximation for dependent trials, *Ann. Prob.*, **3**, 534–545 (1975).

[7] L. H. Y. Chen, An approximation theorem for sums of certain randomly selected indicators, *Z. Wahrsch. Verw. Geb.*, **33**, 223–243 (1975).

[8] L. Devroye, Applications of the theory of records in the study of random trees, *Acta Inf.* **26,** 123–130 (1988).

[9] R. P. Dobrow, On the distribution of distances in recursive trees, *J. Appl. Prob.,* to appear. (1996).

[10] G. Louchard, Exact and asymptotic distributions in digital and binary search trees, *Theo. Info. Appl.* **21**, 471–495 (1987).

[11] W. Lynch, More combinatorial properties of certain trees. *Comp. J.* **7**, 299–302 (1965).

[12] H. M. Mahmoud, Limiting distributions for path lengths in recursive trees. *Prob. Engr. Info. Sci.* **5,** 53–59 (1991).

[13] H. M. Mahmoud, Distances in plane-oriented recursive trees, *J. Comp. Appl. Math.* **41,** 237–245 (1992).

[14] H. M. Mahmoud, *Evolution of random search trees,* Wiley, New York, (1992).

[15] J. W. Moon, The distance between nodes in recursive trees, *London Math. Soc. Lecture Notes Ser.,* No. 13, Cambridge University Press, London, 125–132 (1974).

[16] B. Pittel, Note on the heights of random recursive trees and random m-ary search trees, *Rand. Struct. Alg.* **5** 337–347 (1994).

[17] R. T. Smythe and H. M. Mahmoud, A survey of recursive trees, *Theo. Prob. Math. Stat.*, to appear. (1995).

[18] C. M. Stein, A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, *Proc. 6th Berk. Symp. Math. Stat. Prob.* **2**, University California Press, Berkeley, Cal., 583–602 (1972).

[19] J. Szymański, On a nonuniform random recursive tree, *Ann. Disc. Math.* **33** 297–306 (1987).