A hooray for Poisson approximation

Rudolf Grübel[1]

[1] *Institut für Mathematische Stochastik, Universität Hannover, Postfach 60 09, D-30060 Hannover, Germany*

We give several examples for Poisson approximation of quantities of interest in the analysis of algorithms: the distribution of node depth in a binary search tree, the distribution of the number of losers in an election algorithm and the discounted profile of a binary search tree. A simple and well-known upper bound for the total variation distance between the distribution of a sum of independent Bernoulli variables and the Poisson distribution with the same mean turns out to be very useful in all three cases.

Keywords: binary search tree, multiplicity of maxima, tree profile

1 Introduction

One of the truly classical topics of applied probability is the 'law of small numbers', the approximation of the distribution $\mathcal{L}(S_n)$ of the number S_n of successes in n repetitions with success probability p by the Poisson distribution $\mathrm{Po}(\lambda)$ with mean $\lambda = np \,(= ES_n)$. A standard reference in this area is the book 'Poisson Approximation' by Barbour et al. (1992), which contains the following result,

$$d_{\mathrm{TV}}\Big(\mathop{\bigstar}_{i=1}^{n} \mathrm{Ber}(p_i)\,,\ \mathrm{Po}\Big(\sum_{i=1}^{n} p_i\Big)\Big) \ \le\ \frac{\sum_{i=1}^{n} p_i^2}{\sum_{i=1}^{n} p_i}\,. \tag{1}$$

Here we wrote $\mathrm{Ber}(p)$ for the Bernoulli distribution with parameter p, so that $\mathcal{L}(X) = \mathrm{Ber}(p)$ means $P(X=1) = p = 1 - P(X=0)$, and d_{TV} for the total variation distance of probability measures,

$$d_{\mathrm{TV}}(\mu,\nu) \ :=\ \sup_B \big|\mu(B) - \nu(B)\big|,$$

which for μ, ν concentrated on some countable set A can be written as

$$d_{\mathrm{TV}}(\mu,\nu) \ =\ \frac{1}{2}\sum_{k\in A} \big|\mu(\{k\}) - \nu(\{k\})\big|.$$

Finally, '$\star$' denotes convolution. This is the distributional form; in the language of random variables, where $S_n = I_1 + \cdots + I_n$ with independent indicator variables $I_1, \ldots, I_n$ and Y_n is Poisson distributed with the same mean as S_n, (1) becomes

$$\sup_{A\subset\mathbb{N}_0} \big|P(S_n\in A) - P(Y_n\in A)\big| \ \le\ \frac{\sum_{i=1}^{n} P(I_i=1)^2}{\sum_{i=1}^{n} P(I_i=1)}\,.$$

Of course, Poisson approximation is a familiar topic in the analysis of algorithms, see e.g. p.198ff in Sedgewick and Flajolet (1996). Below we give three examples where direct use can be made of (1). The first of these is known as 'unsuccessful searching' in a binary search tree; see Section 2.4 in Mahmoud (1992). A variant of this problem, where we ask for the insertion depth of an item with a specified label, has recently received a more detailed look. In the second example we consider distributed leader election; see Section 7.8 in Sedgewick and Flajolet (1996). We show that (1) leads to a simple proof of a two-dimensional distributional limit theorem, including a rate result. The third example is more elaborate, it deals with the profile of binary search trees where we hope that our approach might provide a new angle.

2 Node depth in binary search trees

A random binary tree results if we apply the BST algorithm to a random permutation of the set $\{1, \ldots, n\}$; see Section 2.1 in Mahmoud (1992) or Section 5.5 in Sedgewick and Flajolet (1996) for an explanation of the algorithm. Let X_n denote the insertion depth of the last item, to avoid distracting trivialities we assume that $n \geq 2$. It is well known that

$$\mathcal{L}(X_n) \;=\; \mathop{\bigstar}_{i=2}^{n} \mathrm{Ber}(2/i),$$

Devroye (1988) gave a beautiful proof, based on the relation to records. Let

$$H_n := \sum_{i=1}^{n} \frac{1}{i}, \quad H_n^{(2)} := \sum_{i=1}^{n} \frac{1}{i^2},$$

be the harmonic numbers of the first and second kind. The following is now an immediate consequence of (1), as has already been noted by Dobrow and Smythe (1996):

$$d_{\mathrm{TV}}\big(\mathcal{L}(X_n),\, \mathrm{Po}(2H_n - 2)\big) \;\leq\; \frac{2(H_n^{(2)} - 1)}{H_n - 1} \;\leq\; \frac{\pi^2 - 6}{3(\log n - 1)}.$$

Together with the familiar asymptotics for Poisson distributions this can be used to obtain other results such as asymptotic normality of X_n, including Berry-Esséen style bounds. Based on an analysis of subtree dependence, Grübel and Stefanoski (2005) recently obtained an analogous result for the insertion depth X_{nl} of the item with label l,

$$d_{\mathrm{TV}}\big(\mathcal{L}(X_{nl})\,,\ \mathrm{Po}(EX_{nl})\big) \;\leq\; \frac{28 + \pi^2}{\log n} \quad \text{for } l = 1, \ldots, n,$$

they also discuss mixed Poisson approximation in connection with the Wasserstein distance as an alternative to the total variation distance.

3 Selecting a loser

To motivate our second application we consider the following situation, somewhat related to the author's professional life: A maths department has to select a chairperson from its professors. These simultaneously throw coins; those that obtain 'head' may leave; those with 'tails' continue into the next round. A tie results if all remaining candidates throw 'head'. What is the probability that this happens, if the department has n professors and the coins show 'head' with probability p? This problem, of somewhat playful appearance, has attracted a surprising multitude of researchers. It is also a good example for the variety of tools that can be brought to bear, see Kirschenhofer and Prodinger (1996) for an analytic approach and Bruss and Grübel (2003) for an approach based on the Sukhatme-Rényi representation of exponential order statistics, a familiar tool in mathematical statistics.

We consider the joint distribution of the number of rounds and the number of losers. To be precise we start with a sequence $(X_n)_{n\in\mathbb{N}}$ of independent random variables with $\mathcal{L}(X_i) = \mathrm{Geo}(p)$, i.e.

$$P(X_i = k) \;=\; q^{k-1}p \quad \text{for all } i, k \in \mathbb{N},$$

and put

$$M_n := \max\{X_1, \ldots, X_n\}, \quad W_n := \#\{1 \leq i \leq n :\ X_i = M_n\}.$$

The event that the maximum of the first n variables is equal to k and that exactly j of these have this value is equivalent to the event that j of the variables $X_1, \ldots, X_n$ are equal to k and the other $n - j$ values are at most $k - 1$, hence

$$P(M_n = k, W_n = j) \;=\; \binom{n}{j}(pq^{k-1})^j(1 - q^{k-1})^{n-j}, \quad k \in \mathbb{N},\, j = 1, \ldots, k.$$

If interest is in one of the marginal variables only, then the natural next step is to sum out the other index. Looking at the joint distribution, however, we recognize an almost binomial pattern. Guided by Poisson approximation we therefore introduce the distributions Q_n, $n \in \mathbb{N}$, on $A = \{(0,0)\} \cup \mathbb{N} \times \mathbb{N}$ by

$$Q_n(0,0) := e^{-n}, \quad Q_n(k,j) \;:=\; \frac{(npq^{k-1})^j}{j!}\, e^{-nq^{k-1}}, \quad k, j \in \mathbb{N}\,,$$

where we have written $Q_n(k,j)$ instead of $Q_n\big(\{(k,j)\}\big)$; checking $\sum_{(k,j)\in A} Q_n(k,j) = 1$ is easy. Writing $\mathrm{Bin}(n,p)$ for the binomial distribution with parameters n and p we then obtain, for arbitrary $k \in \mathbb{N}$,

$$\begin{aligned}
\sum_{j=1}^{n} \big|P(M_n = k, W_n = j) - Q_n(k,j)\big| &= \sum_{j=1}^{n} p^j \left| \binom{n}{j} q^{j(k-1)}(1-q^{k-1})^{n-j} - e^{-nq^{k-1}} \frac{(nq^{k-1})^j}{j!} \right| \\
&\le \sum_{j=1}^{n} \left| \binom{n}{j} q^{j(k-1)}(1-q^{k-1})^{n-j} - e^{-nq^{k-1}} \frac{(nq^{k-1})^j}{j!} \right| \qquad (2) \\
&\le d_{\mathrm{TV}}\big(\mathrm{Bin}(n,q^{k-1}), \mathrm{Po}(nq^{k-1})\big) \\
&\le q^{k-1},
\end{aligned}$$

with (1) used in the last step. For $\epsilon > 0$ fixed let

$$K(n) := \big\{k \in \mathbb{N} : -(1-\epsilon)\log_q n \le k \le -2\log_q n\big\}.$$

Then, for all $n \in \mathbb{N}$,

$$d_{\mathrm{TV}}\big(\mathcal{L}(M_n, W_n), Q_n\big) \le P\big(M_n \notin K(n)\big) + Q_n\big(\{(k,j) \in A : k \notin K(n)\}\big) + \sum_{k \in K(n)} q^{k-1}. \qquad (3)$$

Standard procedures give the rate $O(n^{-1})$ for the first two terms, the third is of order $O(n^{-1+\epsilon})$. Hence we have the following result.

Theorem 1 *With M_n, W_n and Q_n as defined above and $n \to \infty$,*

$$d_{\mathrm{TV}}\big(\mathcal{L}(M_n, W_n), Q_n\big) = o(n^{-\gamma}) \quad \textit{for all } \gamma < 1.$$

The theorem implies that $(M_n - \lfloor -\log_q n \rfloor, W_n)$ converges in distribution, with limit law $\tilde{Q}_\eta$ given by

$$\tilde{Q}_\eta(k,j) := \frac{p^j q^{j(k-\eta-1)}}{j!} e^{-q^{k-\eta-1}}, \quad k \in \mathbb{Z},\ j \in \mathbb{N},$$

along subsequences $(n_l)_{l\in\mathbb{N}}$ that satisfy $\lim_{l\to\infty}(-\log_q n_l - \lfloor -\log_q n_l \rfloor) = \eta$. It may be interesting to note that, in the language of the motivating example, the dependence between the number of rounds required and the number of losers does not vanish asymptotically as the number of participants grows to infinity.

The total variation distance does not increase if we apply a function to the random variables in question; formally,

$$d_{\mathrm{TV}}\big(\mathcal{L}(\phi(X)), \mathcal{L}(\phi(Y))\big) \le d_{\mathrm{TV}}\big(\mathcal{L}(X), \mathcal{L}(Y)\big).$$

Hence Theorem 1 immediately yields an asymptotic distributional approximation for the multiplicity W_n of the maximum. Such a result has been obtained in Bruss and Grübel (2003), who used a variant of the total variation distance incorporating the weight function $j \mapsto \gamma^j$, $\gamma < 1/p$. In this context it may be interesting to note that we have been rather generous when simply dropping p^j in (2) above. Bruss and Grübel (2003) obtained the rate $O(n^{-1})$. The present simple argument, based on (1) and with $o(n^{-\gamma})$ for all $\gamma < 1$ only, comes close. It should be noted that the joint distribution provides additional information and could for example be used to relate the well-known small periodic fluctuations of the distributions of M_n and W_n to each other. Finally, (3) can be used to obtain (non-asymptotic) upper bounds for the total variation distance between $\mathcal{L}(M_n, W_n)$ and Q_n.

4 The BST profile

As in Section 2 we consider the random binary tree T_n obtained by applying the BST algorithm to a uniformly distributed random permutation of $\{1,\dots,n\}$. Let U_{nk} be the number of external nodes at level k of T_n (the root node has level 0 and is the only external node of the empty tree T_0). The stochastic dynamics of the sequence of random sequences $(U_{nk})_{k\in\mathbb{N}_0}$, $n \in \mathbb{N}_0$, can be described with the help of an infinite urn model: We think of U_{nk} as the number of balls in urn k at time n. Initially, with $n = 0$, there is one ball in urn 0, all other urns are empty. At time n we choose one of the (then necessarily $n+1$) balls uniformly at random, remove it from its urn, which has label k, say, and add two balls to the urn with number $k+1$. It is easy to see that the *discounted profile*

$$\Psi(T_n) = \bigl(\Psi_k(T_n)\bigr)_{k\in\mathbb{N}_0}, \quad \Psi_k(T_n) := 2^{-k} U_{nk} \ \text{ for all } k \in \mathbb{N}_0,$$

then constitutes a random probability measure on $\mathbb{N}_0$.

Profiles of random binary trees, both in their various flavours, have been investigated by various researchers. From the many interesting papers on this subject we mention only two, since they were particularly stimulating for the point of view adopted below. Aldous and Shields (1988) considered a variant where a ball from urn k is chosen with probability proportional to c^{-k} for some fixed constant $c > 1$. They obtained a law of large numbers and a central limit theorem in the sense of a Gaussian diffusion approximation for the discounted profile. Roughly, $\Psi(T_n)$ keeps its shape and a fluctuation phenomenon (as in Section 3) occurs as $n \to \infty$. Chauvin et al. (2001) considered the 'raw' profile $k \mapsto U_{nk}/(n+1)$ and, among other results, showed that this random probability mass function can be approximated locally by the density of a normal distribution with mean and variance both equal to $2\log n$, hence the profile flattens out as $n \to \infty$. There is no periodicity at this level of detail, but see also the comments at the end of this section.

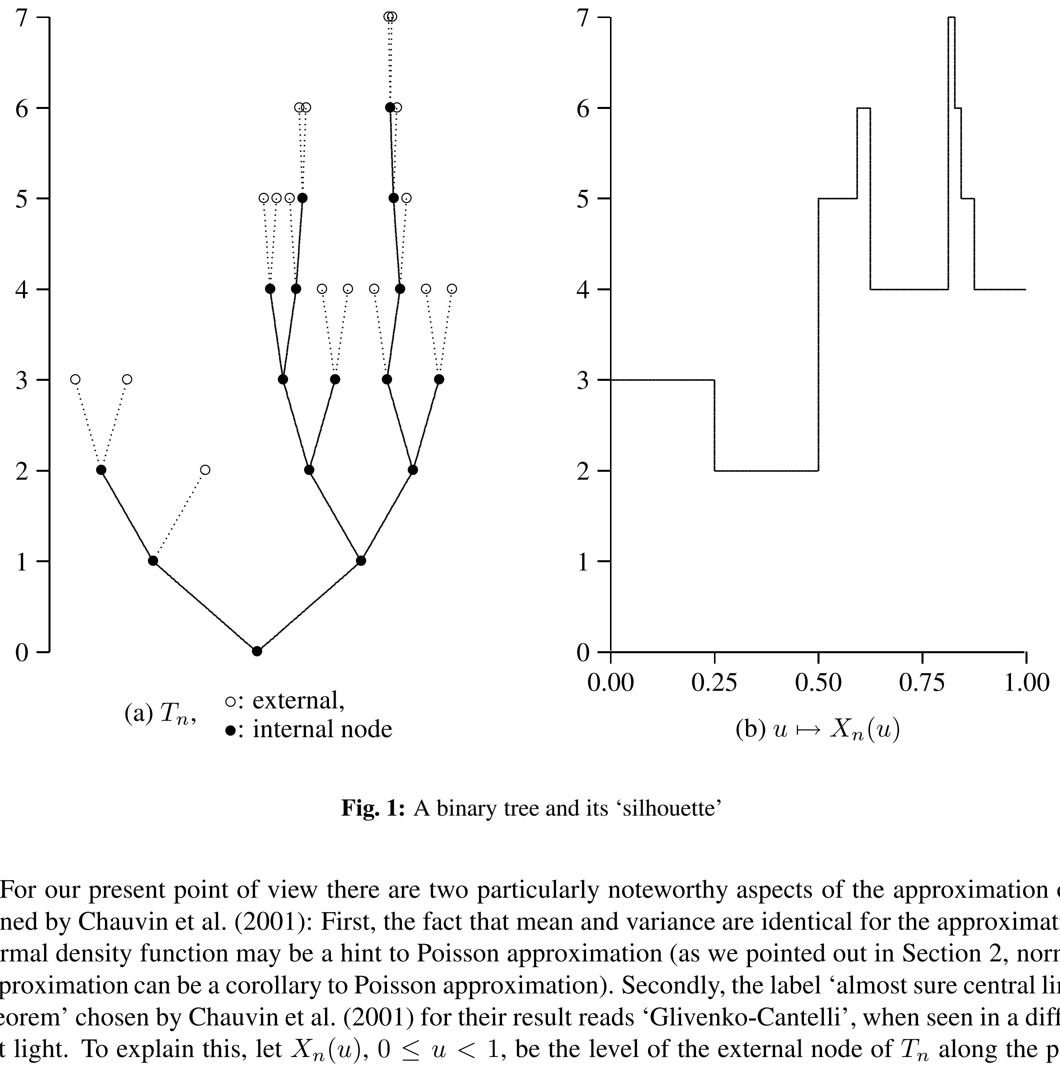

Fig. 1: A binary tree and its 'silhouette'

For our present point of view there are two particularly noteworthy aspects of the approximation obtained by Chauvin et al. (2001): First, the fact that mean and variance are identical for the approximating normal density function may be a hint to Poisson approximation (as we pointed out in Section 2, normal approximation can be a corollary to Poisson approximation). Secondly, the label 'almost sure central limit theorem' chosen by Chauvin et al. (2001) for their result reads 'Glivenko-Cantelli', when seen in a different light. To explain this, let $X_n(u)$, $0 \le u < 1$, be the level of the external node of T_n along the path obtained from the binary expansion $(u_1, u_2, u_3, \ldots)$ of u, where we interpret $u_k = 0$ as a move to the left in the k^{th} step and 1 as a move to the right. Figure 1 shows a tree together with its 'silhouette' $u \mapsto X_n(u)$. We can think of the values $X_n(u)$, n fixed and u varying over the unit interval, as a 'sample' (of admittedly unusual size) from the distribution of the height along one specific path. Of course, the sample values are not independent, but if dependence is not too strong, then we would expect that averaging over the whole sample should asymptotically reproduce the underlying distribution. Now, it is clear from the above urn description that we have

$$\mathcal{L}\big(X_n(u)\big) \;=\; \mathop{\bigstar}_{i=1}^{n} \mathrm{Ber}(1/i) \qquad \text{for all } u \in [0,1),$$

and (1) provides a bound for the distance of this underlying distribution to the Poisson distribution with mean $H_n \sim \log n$. Writing $\delta_k : \mathbb{N}_0 \to \{0,1\}$ for the Kronecker delta function, $\delta_k(k) = 1$, $\delta_k(i) = 0$ for $i \neq k$, we have the following basic connection between the discounted profile and the silhouette process:

$$\Psi_k(T_n) \;=\; \int_0^1 \delta_k\big(X_n(u)\big)\, du \qquad \text{for all } k, n \in \mathbb{N}_0. \tag{4}$$

The integral in (4) corresponds to the averaging over the sample in an empirical process framework. In summary, we expect that the random probability distribution $\Psi(T_n)$ is close to the deterministic probability distribution $\mathrm{Po}(\log n)$ with high probability for n large. Note that $k \mapsto \Psi_k(T_n)$ can also be regarded as the probability mass function of the distribution of X_n, if we interpret $u \mapsto X_n(u)$ as a random variable on the standard probability space, i.e. the unit interval endowed with its Borel σ-field and the uniform distribution.

We now regard probability distributions on $\mathbb{N}_0$, random or not, as sequences. Apart from a factor 2, the l_1-distance $\|p - q\|_1 = \sum_{k=0}^{\infty} |p_k - q_k|$ of two such sequences $p = (p_k)_{k\in\mathbb{N}_0}$, $q = (q_k)_{k\in\mathbb{N}_0}$ is then equal to the total variation distance of the associated probability measures. Instead of l_1 and total variation distance we use below the (Hilbert) space

$$l_2(\mathbb{N}_0) \;=\; \big\{a = (a_k)_{k\in\mathbb{N}_0} \in \mathbb{R}^{\mathbb{N}_0} : \|a\|_2 < \infty\big\}, \quad \text{with } \|a\|_2^2 := \sum_{k=0}^{\infty} a_k^2 \,.$$

Within this framework the theorem below confirms the above conjecture. We require the following properties of the norms and distances of Poisson distributions.

Lemma 1 (a)

$$\lim_{\lambda\to\infty} \lambda^{1/2}\, \|\,\mathrm{Po}(\lambda)\|_2^2 = \frac{1}{2\pi^{1/2}}\,.$$

(b) *There exist finite constants λ_0 and C such that, for all $\lambda \geq \lambda_0$ and all $\eta > 0$,*

$$\|\,\mathrm{Po}(\lambda) - \mathrm{Po}(\eta)\|_2^2 \;\leq\; C\,\big(1 + (\lambda-\eta)^2\big)\,\lambda^{-3/2}.$$

Proof. Using the modified Bessel function

$$I_0(z) \;:=\; \sum_{k=0}^{\infty} \frac{1}{k!\,k!}\Big(\frac{z}{2}\Big)^{2k}$$

we can write the squared l_2-norm of the Poisson distribution with parameter λ as

$$\|\,\mathrm{Po}(\lambda)\|_2^2 \;=\; e^{-2\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{2k}}{k!\,k!} \;=\; e^{-2\lambda} I_0(2\lambda).$$

It is known that

$$e^{-t} I_0(t) \;=\; \frac{1}{\sqrt{2\pi t}} \;+\; O\big(t^{-3/2}\big) \tag{5}$$

as $t \to \infty$, see e.g. Formula 9.7.1 in Abramowitz and Stegun (1964). Part (a) now follows easily.

For the proof of (b) we first note that

$$\|\,\mathrm{Po}(\lambda) - \mathrm{Po}(\eta)\|_2^2 \;\leq\; \|\,\mathrm{Po}(\lambda)\|_2^2 + \|\,\mathrm{Po}(\eta)\|_2^2 \;=\; \sum_{k=0}^{\infty}\mathrm{Po}_k(\lambda)^2 + \sum_{k=0}^{\infty}\mathrm{Po}_k(\eta)^2 \;\leq\; 2,$$

so that it is enough to prove the statement for values of η satisfying the condition $|\lambda - \eta| \leq \lambda/2$. Using the modified Bessel function as in the proof of (a) we obtain

$$\|\,\mathrm{Po}(\lambda) - \mathrm{Po}(\eta)\|_2^2 \;=\; \sum_{k=0}^{\infty}\Big(e^{-\lambda}\frac{\lambda^k}{k!} - e^{-\eta}\frac{\eta^k}{k!}\Big)^2 \;=\; e^{-2\lambda}I_0(2\lambda) \;+\; e^{-2\eta}I_0(2\eta) \;-\; 2e^{-\lambda-\eta}I_0\big(2\sqrt{\lambda\eta}\big).$$

With $R(t) := e^{-t}I_0(t) - 1/\sqrt{2\pi t}$ this can be rewritten as

$$\|\operatorname{Po}(\lambda) - \operatorname{Po}(\eta)\|_2^2 \;=\; R(2\lambda) + R(2\eta) - 2e^{-\eta-\lambda+2\sqrt{\lambda\eta}}\,R\bigl(2\sqrt{\lambda\eta}\bigr) \tag{6}$$

$$+\frac{1}{2\sqrt{\pi}}\Bigl(\frac{1}{\sqrt{\lambda}} + \frac{1}{\sqrt{\eta}} - \frac{2}{\sqrt[4]{\lambda\eta}}e^{-\eta-\lambda+2\sqrt{\lambda\eta}}\Bigr). \tag{7}$$

Using (5), $0 \le \exp(-\eta-\lambda+2\sqrt{\lambda\eta}) = \exp(-(\sqrt{\lambda}-\sqrt{\eta})^2) \le 1$ and $\eta \ge \lambda/2$ we obtain the required rate $O(\lambda^{-3/2})$ for the terms on the right hand side of (6). The term in big brackets in (7) can be estimated from above by $R_1(\lambda,\eta) + R_2(\lambda,\eta)$, with

$$R_1(\lambda,\eta) := \frac{1}{\sqrt{\lambda\eta}}\bigl(\lambda^{1/4} - \eta^{1/4}\bigr)^2, \quad R_2(\lambda,\eta) := \frac{2}{\sqrt[4]{\lambda\eta}}\bigl(1 - e^{-(\sqrt{\lambda}-\sqrt{\eta})^2}\bigr)$$

and the proof will be complete once we have shown that, with suitable constants C_1, C_2 and λ_0,

$$\lambda^{3/2}R_i(\eta,\lambda) \;\le\; C_i\,(\lambda-\eta)^2 \quad \text{for all } \lambda > \lambda_0,\ \lambda/2 \le \eta \le 2\lambda \ \text{ and } i = 1,2. \tag{8}$$

Indeed: Using the elementary inequalities

$$\bigl|(1+x)^{1/4} - 1\bigr| \le |x|,\ \bigl|(1+x)^{1/2} - 1\bigr| \le |x| \ \text{ for } x \ge -1/2, \quad 1 - e^{-x} \le x \ \text{ for } x \ge 0,$$

we obtain with $\lambda_0 := 1$ and $\eta \ge \lambda/2$

$$\lambda^{3/2}R_1(\eta,\lambda) \;=\; \Bigl(\frac{\lambda}{\eta}\Bigr)^{1/2}\lambda\Bigl(1 - \Bigl(1 + \frac{\eta-\lambda}{\lambda}\Bigr)^{1/4}\Bigr)^2 \;\le\; \sqrt{2}\,\lambda\frac{(\eta-\lambda)^2}{\lambda^2} \;\le\; \sqrt{2}(\eta-\lambda)^2,$$

$$\begin{aligned}\lambda^{3/2}R_2(\eta,\lambda) \;\le\; 2\lambda\Bigl(\frac{\lambda}{\eta}\Bigr)^{1/4}\bigl(1 - e^{-(\sqrt{\lambda}-\sqrt{\eta})^2}\bigr) \;\le\; 2^{5/4}\lambda\bigl(\sqrt{\lambda}-\sqrt{\eta}\bigr)^2 \;&\le\; 2^{5/4}\lambda^2\Bigl(1 - \Bigl(1 + \frac{\eta-\lambda}{\lambda}\Bigr)^{1/2}\Bigr)^2\\ &\le\; 2^{5/4}(\eta-\lambda)^2,\end{aligned}$$

which proves (8). □

We also need bounds for the tails of various random variables; the standard approach via moment generating functions and Markov's inequality is enough for our purposes. For sums of independent indicator variables we will use the following variant, which is Theorem 2.8 in Janson et al. (2000).

Lemma 2 *Suppose that $\mathcal{L}(X) = \bigstar_{i=1}^n \operatorname{Ber}(p_i)$ and let $\lambda := \sum_{i=1}^n p_i$. Then, with $\phi(x) := (1+x)\log(1+x) - x$,*

$$P\bigl(X \ge (1+\alpha)\lambda\bigr) \le \exp\bigl(-\lambda\phi(\alpha)\bigr),\ P\bigl(X \le (1-\alpha)\lambda\bigr) \le \exp\bigl(-\lambda\phi(-\alpha)\bigr) \quad \textit{for all } \alpha > 0.$$

As a typical application we consider the time $V_l(u)$ at which the node with distance l from the root along some specific path u first becomes an external node in the sequence $(T_n)_{n\in\mathbb{N}}$. Using

$$P\bigl(V_l(u) \ge k\bigr) \;=\; P\bigl(X_k(u) \le l\bigr) \tag{9}$$

and $\mathcal{L}\bigl(X_k(u)\bigr) = \bigstar_{i=1}^k \operatorname{Ber}(1/i)$ we obtain from Lemma 2 that, for $l = l(n) = O(\log\log n)$ and $k = k(n) \ge n^\kappa$ for some $\kappa > 0$,

$$P\bigl(V_{l(n)}(u) \ge k(n)\bigr) \;=\; O\bigl((\log n)^{-\gamma}\bigr) \quad \text{for all } \gamma > 0. \tag{10}$$

A sequence of probability mass functions with supremum tending to 0 will converge to 0 in l_2-norm. Part (a) of Lemma 1 shows that this happens with rate $(\log n)^{-1/4}$ for $\operatorname{Po}(\log n)$ as $n \to \infty$, hence the approximation in the following theorem makes sense.

Theorem 2 *With $(T_n)_{n\in\mathbb{N}}$ and Ψ as above,*

$$E\bigl\|\Psi(T_n) - \operatorname{Po}(\log n)\bigr\|_2^2 \;=\; O\bigl((\log n)^{-3/2}\bigr) \qquad \textit{as } n \to \infty.$$

Proof. We use Pythagoras' theorem to split the squared l_2-distance into a squared bias term and a variance term:

$$E\big\|\Psi(T_n) - \mathrm{Po}(\log n)\big\|_2^2 = \sum_{k=0}^{\infty}\big(E\Psi_k(T_n) - \mathrm{Po}(\log n)(\{k\})\big)^2 \;+\; \sum_{k=0}^{\infty} E\big(\Psi_k(T_n) - E\Psi_k(T_n)\big)^2 .$$

Using (4) and Fubini's Theorem we see that the expected profile is the mass function associated with $\mathcal{L}(X_n(0)) = \bigstar_{i=1}^{n} \mathrm{Ber}(1/i)$. Hence (1) implies, with H_n as in Section 2,

$$\big\|E\Psi(T_n) - \mathrm{Po}(H_n)\big\|_1 \;=\; O\big((\log n)^{-1}\big).$$

Since $\sum_{k=0}^{\infty} a_k^2 \;\le\; \big(\sum_{k=0}^{\infty}|a_k|\big)^2$ this in turn yields the rate $O\big((\log n)^{-1}\big)$ for the l_2-distance between $E\Psi(T_n)$ and $\mathrm{Po}(H_n)$. Lemma 1 (b) provides the rate $O\big((\log n)^{-3/4}\big)$ for the l_2-distance between $\mathrm{Po}(H_n)$ and $\mathrm{Po}(\log n)$. The desired rate for the squared bias term now follows with the triangle inequality.

We split the sum in the second (variance) term into the range $k \notin A(n)$ and $k \in A(n)$, with

$$A(n) \;:=\; \big\{k \in \mathbb{N}_0 :\; (\log n)/2 \;\le\; k \;\le\; 2\log n\big\}.$$

For the first sum we use

$$\sum_{k\notin A(n)} E\big(\Psi_k(T_n) - E\Psi_k(T_n)\big)^2 \;\le\; 2\sum_{k\notin A(n)} E\Psi_k(T_n) \;=\; 2\,P\big(X_n(0)\notin A(n)\big).$$

As $X_n(0)$ is the sum of independent Bernoulli random variables we can use Lemma 2 to obtain the required rate for $P\big(X_n(0)\notin A(n)\big)$. For the range $k\in A(n)$ we write

$$\begin{aligned}\mathrm{var}\big(\Psi_k(T_n)\big) \;&=\; E\bigg(\int_0^1 \Big(\delta_k\big(X_n(u)\big) - E\delta_k\big(X_n(u)\big)\Big)\,du\cdot\int_0^1\Big(\delta_k\big(X_n(s)\big) - E\delta_k\big(X_n(s)\big)\Big)\,ds\bigg)\\ &=\; \int_0^1\int_0^1 \mathrm{cov}\Big(\delta_k\big(X_n(u)\big),\delta_k\big(X_n(s)\big)\Big)\,du\,ds. \qquad (11)\end{aligned}$$

Let $\alpha(u,s)$ denote the level of the last common ancestor of the paths u and s. Formally,

$$\alpha(u,s) \;:=\; \max\big\{k\in\mathbb{N}_0 :\; j2^{-k}\le u,s<(j+1)2^{-k} \text{ for some } j\in\{0,\dots,2^k-1\}\big\}. \qquad (12)$$

We split the range of the double integral in (11) into $B(n)$ and $B(n)^{\mathrm{c}}$, with

$$B(n) \;:=\; \big\{(u,s)\in[0,1)^2 :\; \alpha(u,s)\ge 4\log\log n\big\}.$$

As a subset of the unit square $B(n)$ consists of the set of all pairs that have the same elements in their binary representation up to (at least) position $\lceil 4\log\log n\rceil$. Hence the area of $B(n)$ is of the order $O((\log n)^{-4\log 2})$, which suffices for this part of the integral as the integrand is bounded by 1 in absolute value and the number of terms in the sum over $k\in A(n)$ is bounded by $2\log n$.

We now fix $k\in A(n)$ and $(u,s)\notin B(n)$. For n large enough we have that $l := \alpha(u,s)$ is small in comparison with k. Let the random variable V denote the first time that $(u_1,u_2,\dots,u_l)$ becomes an external node (the dependence of V on u and s will not be displayed in the notation). For our fourth split we use the conditional covariance formula

$$\begin{aligned}&\mathrm{cov}\Big(\delta_k\big(X_n(u)\big),\delta_k\big(X_n(s)\big)\Big)\\ &\qquad = E\Big(\mathrm{cov}\big[\delta_k\big(X_n(u)\big),\delta_k\big(X_n(s)\big)\big|V\big]\Big) + \;\mathrm{cov}\Big(E\big[\delta_k\big(X_n(u)\big)\big|V\big],E\big[\delta_k\big(X_n(s)\big)\big|V\big]\Big),\end{aligned}$$

so that the proof of the theorem will be complete once we have shown that

$$\sum_{k\in A(n)}\int_0^1\int_0^1 1_{B(n)^{\mathrm{c}}}(u,s)\,E\Big(\mathrm{cov}\big[\delta_k\big(X_n(u)\big),\delta_k\big(X_n(s)\big)\big|V\big]\Big)\,du\,ds \;=\; O\big((\log n)^{-3/2}\big) \qquad (13)$$

and

$$\sum_{k\in A(n)} \int_0^1\!\!\int_0^1 1_{B(n)^c}(u,s)\, \operatorname{cov}\Big(E\big[\delta_k\big(X_n(u)\big)\big|V\big], E\big[\delta_k\big(X_n(s)\big)\big|V\big]\Big)\, du\, ds \;=\; O\big((\log n)^{-3/2}\big). \tag{14}$$

For (13) the idea is that the random variables $\delta_k\big(X_n(u)\big)$ and $\delta_k\big(X_n(s)\big)$ or, equivalently, the two events $\{X_n(u)=k\}$ and $\{X_n(s)=k\}$, are almost independent as k is large in comparison to l. The following computation is crucial for the success of our approach; in essence, it replaces the Poissonization-Depoissonization steps that are often used in connection with binary trees. The Poissonization makes subtrees independent—our calculation, using once again Poisson approximation for the distribution of the sum of independent indicator variables, proceeds in a more direct manner.

Under the condition that $V=m$ the event $\{X_n(u)=k\}\cap\{X_n(s)=k\}$ can be described as follows: Of the $n-m$ steps following the entrance into the last common ancestor of u and s, exactly $2(k-l)$ go in direction u or s, and of these, exactly $k-l$ proceed in the direction given by u. Writing U and S for the number of steps among those with time label $V+1,\dots,n$ taken from the ancestor in direction u or s respectively, it is clear that we have

$$\mathcal{L}(U|V)=\mathcal{L}(S|V)=\mathop{\bigstar}_{i=V+1}^{n}\ \operatorname{Ber}(1/i),\quad \mathcal{L}(U+S|V)=\mathop{\bigstar}_{i=V+1}^{n}\ \operatorname{Ber}(2/i)$$

and $\mathcal{L}(U|U+S,V)=\operatorname{Bin}(U+S,1/2)$. Hence

$$\begin{aligned} P\big(X_n(u)=k,X_n(s)=k\big|V=m\big) &= P\big(U=k-l,U+S=2(k-l)\big|V=m\big)\\ &= P\big(U=k-l\big|U+S=2(k-l)\big)\ P\big(U+S=2(k-l)\big|V=m\big)\\ &= \binom{2(k-l)}{k-l}2^{-2(k-l)}\ \mathop{\bigstar}_{i=m+1}^{n}\ \operatorname{Ber}(2/i)(\{2(k-l)\}). \end{aligned}$$

Similarly, we have

$$P\big(X_n(u)=k\big|V=m\big) \;=\; P\big(U=k-l\big|V=m\big) \;=\; \mathop{\bigstar}_{i=m+1}^{n}\ \operatorname{Ber}(1/i)(\{k-l\}),$$

and for $P(X_n(s)=k|V=m)$ we obtain the same expression. Using the relation

$$\operatorname{Bin}(2j,1/2)(\{j\})\ \operatorname{Po}(2\lambda)(\{2j\}) \;=\; \big(\operatorname{Po}(\lambda)(\{j\})\big)^2$$

and some more notation,

$$\Delta_j(m,n,k) \;:=\; \Big|\mathop{\bigstar}_{i=m+1}^{n}\ \operatorname{Ber}\Big(\frac{j}{i}\Big)(\{k\}) \;-\; \operatorname{Po}\Big(\sum_{i=m+1}^{n}\frac{j}{i}\Big)(\{k\})\Big| \quad \text{for } j=1,2,$$

we obtain

$$\begin{aligned} &\Big|\operatorname{cov}\big[\delta_k\big(X_n(u)\big),\delta_k\big(X_n(s)\big)\big|V=m\big]\Big|\\ &\quad= \Big|P\big(X_n(u)=k,X_n(s)=k\big|V=m\big) - P\big(X_n(u)=k\big|V=m\big)\,P\big(X_n(s)=k\big|V=m\big)\Big|\\ &\quad\le \binom{2(k-l)}{k-l}2^{-2(k-l)}\,\Delta_2(m,n,2(k-l)) + 2\operatorname{Po}\Big(\sum_{m+1}^{n}\frac{1}{i}\Big)(\{k-l\})\,\Delta_1(m,n,k-l) \;+\; \Delta_1(m,n,k-l)^2. \end{aligned}$$

Summing over $k\ge(\log n)/2$, using

$$\binom{2n}{n}2^{-2n}\le(\pi n)^{-1/2},\qquad \sup_{k\in\mathbb{N}_0}\ \operatorname{Po}(\lambda)(\{k\})=O\big(\lambda^{-1/2}\big)\ \text{ as } \lambda\to\infty$$

(see Proposition A.2.7 in Barbour et al. (1992) for the second statement) and (1) twice we arrive at

$$\sum_{k\in A(n)}\Big|\operatorname{cov}\big[\delta_k\big(X_n(u)\big),\delta_k\big(X_n(s)\big)\big|V=m\big]\Big| \;=\; O\big((\log n)^{-3/2}\big), \tag{15}$$

uniformly in $m \le \sqrt{n}$ and $l \le \lceil 4 \log\log n \rceil$. To obtain (13) we now split the integral associated with the expected value into $\{V \le \sqrt{n}\}$ and $\{V > \sqrt{n}\}$. For the first part, the required bound follows because of (15). For the second part we use (10), the fact that the size of $A(n)$ is $O(\log n)$ and the bound 1 for the covariance of the indicator variables to obtain $O\big((\log n)^{-3/2}\big)$ for this part too.

For the proof of (14) we first note that, on $\alpha(u,s) = l$ with α as in (12),

$$E\big[\delta_k\big(X_n(u)\big)\big|V = m\big] \;=\; P\big(X_n(u) = k\big|V = m\big) \;=\; \psi(k,l,m,n),$$

where $k \mapsto \psi(k,l,m,n)$ is the mass function associated with $\delta_l \star \bigstar_{i=m+1}^{n} \mathrm{Ber}(1/i)$. For $E[\delta_k(X_n(s))|V = m]$ we obtain the same expression. Hence, using the fact that for real random variables ξ, $\mathrm{var}(\xi) \le E(\xi - a)^2$ for all $a \in \mathbb{R}$, we obtain that

$$\mathrm{cov}\Big(E\big[\delta_k\big(X_n(u)\big)\big|V\big], E\big[\delta_k\big(X_n(s)\big)\big|V\big]\Big) \;=\; \mathrm{var}\big(\psi(k,l,V,n)\big) \;\le\; E\big(\psi(k,l,V,n) - \mathrm{Po}_k(\log n)\big)^2.$$

Conditionally on $V = m$, the sum of this last expression over $k \in A(n)$ is therefore bounded by the squared l_2-distance between $\psi(\cdot,l,m,n)$ and $\mathrm{Po}(\log n)$. From (1) we get

$$\|\psi(\cdot,l,m,n) - \mathrm{Po}(l + H_n - H_m)\|_1 \;\le\; \frac{2l + \pi^2/3}{l + H_n - H_m}.$$

As in the proof of (13) we may assume that $V \le \sqrt{n}$, so that we have the upper bound

$$\sum_{k \in A(n)} \int_0^1\!\!\int_0^1 1_{B(n)^c}(u,s)\, 1_l(\alpha(u,s))\, E\big(\psi(k,l,V,n) - \mathrm{Po}_k(l + H_n - H_V)\big)^2\, du\, ds \;\le\; C\, l^2\, (\log n)^{-2},$$

where here and in the following C denotes a generic constant. Lemma 1 (b) provides

$$\sum_{k \in A(n)} \big(\mathrm{Po}_k(l + H_n - H_V) - \mathrm{Po}_k(\log n)\big)^2 \;\le\; C\big(1 + \big(l + 1 + \log(1+V)\big)^2\big)(\log n)^{-3/2}.$$

Note that V depends on u and s, but not on n. The proof of (14) will therefore be complete if we can show that

$$\int_0^1\!\!\int_0^1 1_{\{l\}}(\alpha(u,s))\, E\big(\log(1+V)\big)^2\, du\, ds \;\le\; C\, l^2 \tag{16}$$

and

$$\int_0^1\!\!\int_0^1 \alpha(u,s)^2\, du\, ds \;<\; \infty. \tag{17}$$

Using (9) we obtain that, again on $\alpha(u,s) = l$,

$$E\big(\log(1+V)^2\big) \;=\; 2\int_0^\infty t\, P(\log(1+V) \ge t)\, dt \;\le\; 2\sum_{k=0}^{\infty}(k+1) P(\xi_k \le l)$$

where the random variable ξ_k has distribution $\bigstar_{i=1}^{a_k} \mathrm{Ber}(1/i)$ with $a_k := \lceil e^k \rceil - 1$. The sum over $k \le 2l$ is bounded by $(2l+1)^2$. On $\{k > 2l\}$ we use Lemma 2 to obtain

$$P(\xi_k \le l) \;\le\; P\big(\xi_k \le 2(E\xi_k)/3\big) \;\le\; \exp\big(-(k-1)\phi(-1/3)\big),$$

which provides an upper bound for this part of the sum that does not depend on l. Taken together, this proves (16).

Finally, for the proof of (17), we regard α as random: The double integral $\int_0^1\!\int_0^1 \dots\, du\, ds$ means that we select u and s uniformly from the unit interval, and we have to show that the second moment of $l := \alpha(u,s)$ is bounded. This is, however, immediate from the fact that in this interpretation l has a geometric distribution with parameter $1/2$. $\square$

Similar to (3) we can write

$$\big\|\Psi(T_n) - \mathrm{Po}(\log n)\big\|_1 \;\le\; \sum_{k \in A(n)} \big|\Psi_k(T_n) - \mathrm{Po}_k(\log n)\big| + \sum_{k \notin A(n)} \Psi_k(T_n) + \sum_{k \notin A(n)} \mathrm{Po}_k(\log n)$$

for any subset $A(n)$ of $\mathbb{N}_0$. By the Cauchy-Schwarz inequality,

$$\sum_{k\in A(n)} \bigl|\Psi_k(T_n) - \mathrm{Po}_k(\log n)\bigr| \;\le\; \bigl|A(n)\bigr|^{1/2}\,\bigl\|\Psi(T_n) - \mathrm{Po}(\log n)\bigr\|_2.$$

Taking expectations and using $(E\xi)^2 \le E(\xi^2)$ we obtain

$$E\bigl\|\Psi(T_n) - \mathrm{Po}(\log n)\bigr\|_1 \;\le\; \bigl|A(n)\bigr|^{1/2}\Bigl(E\bigl\|\Psi(T_n) - \mathrm{Po}(\log n)\bigr\|_2^2\Bigr)^{1/2} + P\bigl(\xi_{1,n}\notin A(n)\bigr) + P\bigl(\xi_{2,n}\notin A(n)\bigr),$$

where $\mathcal{L}(\xi_{1,n}) = \bigstar_{i=1}^n\ \mathrm{Ber}(1/i)$ and $\mathcal{L}(\xi_{2,n}) = \mathrm{Po}(\log n)$. With $A(n) := \{k\in\mathbb{N}_0 : |k-\log n| \le (\log n)^\beta\}$, $\beta > 1/2$, and the appropriate tail inequalities we therefore obtain from Theorem 2 the following result on the expected total variation distance between the random probability measure $\Psi(T_n)$ and the (non-random) Poisson distribution with mean $\log n$.

Corollary 1 $E\bigl(d_{\mathrm{TV}}\bigl(\Psi(T_n),\,\mathrm{Po}(\log n)\bigr)\bigr) = o\bigl((\log n)^{-\gamma}\bigr)$ *for all* $\gamma < 1/2$.

In their recent investigation of BST profiles Drmota and Hwang (2004) found several phase transitions for the behaviour of the variance of the number of nodes at a particular level. Their analytic techniques, which are completely different from our approach, result in a variety of asymptotic expressions for $\mathrm{var}(U_{nk})$ (in our notation) and can therefore be used to replace some of the probabilistic arguments in the proof of Theorem 2.

In our approach, results such as Theorem 2 are essentially interpreted as a (functional version of the) law of large numbers. Distributional limit results can be regarded as the logical next step; see e.g. the recent paper by Fuchs et al. (2004). Interestingly, on the level of limit distributions, for example in connection with the distribution of $U_{n,k_n}/EU_{n,k_n}$ with $k_n - \log n = O(1)$, periodicities reappear.

Acknowledgements

The title was inspired by the title of a recent talk given by the father of the field 'Analysis of Algorithms'; see `http://www-cs-faculty.stanford.edu/~knuth/musings.html`. Some of the material was presented first during an Oberwolfach workshop in August 2005; I thank the organizers for the invitation and Professors Hsien-Kuei Hwang and Svante Janson for their comments and for providing me with valuable hints to the literature during that time. Finally, I am also grateful to the referees for their supportive remarks.

References

M. Abramowitz and I. A. Stegun. *Handbook of Mathematical Functions*. National Bureau of Standards, Washington, 1964.

D. Aldous and P. Shields. A diffusion limit for a class of randomly-growing binary trees. *Probab. Theory Related Fields*, 79:509–542, 1988.

A. D. Barbour, L. Holst, and S. Janson. *Poisson Approximation*. Oxford University Press, New York, 1992.

F. T. Bruss and R. Grübel. On the multiplicity of the maximum in a discrete random sample. *Ann. Appl. Probab.*, 13:1252–1263, 2003.

B. Chauvin, M. Drmota, and J. Jabbour-Hattab. The profile of binary search trees. *Ann. Appl. Probab.*, 11:1042–1062, 2001.

L. Devroye. Applications of the theory of records in the study of random trees. *Acta Inform.*, 26:123–130, 1988.

R. P. Dobrow and R. T. Smythe. Poisson approximations for functionals of random trees. In *Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)*, volume 9, pages 79–92, 1996.

M. Drmota and H.-K. Hwang. Bimodality and phase transitions in the profile variance of random binary search trees. Preprint, 2004.

M. Fuchs, H.-K. Hwang, and R. Neininger. Bimodality and phase transitions in the profile variance of random binary search trees. Preprint, 2004.

R. Grübel and N. Stefanoski. Mixed Poisson approximation of node depth distributions in random binary search trees. *Ann. Appl. Probab.*, 15:279–297, 2005.

S. Janson, T. Łuczak, and A. Rucinski. *Random Graphs*. Wiley, New York, 2000.

P. Kirschenhofer and H. Prodinger. The number of winners in a discrete geometrically distributed sample. *Ann. Appl. Probab.*, 6:687–694, 1996.

H. M. Mahmoud. *Evolution of Random Search Trees*. John Wiley & Sons Inc., New York, 1992.

R. Sedgewick and P. Flajolet. *An Introduction to the Analysis of Algorithms*. Addison-Wesley, Reading, 1996.