APPLICATIONS OF STEIN'S METHOD IN THE ANALYSIS OF RANDOM BINARY SEARCH TREES

Luc Devroye
luc@cs.mcgill.ca

ABSTRACT. Under certain conditions, sums of functions of subtrees of a random binary search tree are asymptotically normal. We show how Stein's method can be applied to study these random trees, and survey methods for obtaining limit laws for such functions of subtrees.

KEYWORDS AND PHRASES. Binary search tree, data structures, probabilistic analysis, limit law, convergence, toll functions, Stein's method, random trees, contraction method, Berry-Esseen theorem.

CR CATEGORIES: 3.74, 5.25, 5.5.

Research of the first author was sponsored by NSERC Grant A3456 and by FCAR Grant 90-ER-0291. Address: School of Computer Science, McGill University, 3480 University Street, Montreal, Canada H3A 2K6.

§1. Random binary search trees

Binary search trees are almost as old as computer science: they are trees that we frequently use to store data in situations when dictionary operations such as INSERT, DELETE, SEARCH, and SORT are often needed. Knuth (1973) gives a detailed account of the research that has been done with regard to binary search trees. The purpose of this note is twofold: first, we give a survey, with proofs, of the main results for many random variables defined on random binary search trees. Based upon a simple representation given in (4), we can obtain limit laws for sums of functions of subtrees using Stein's theorem. The results given here improve on those found in devroye (2002), and this was our second aim. There are sums of functions of subtree sizes that cannot be dealt with by Stein's method, most notably because the limit laws are not normal. A convenient way of treating those is by the fixed-point method, which we survey at the end of the paper.

Formally, a binary search tree for a sequence of real numbers $x_1, \ldots, x_n$ is a Cartesian tree for the pairs $(1, x_1), (2, x_2), \ldots, (n, x_n)$. The Cartesian tree for pairs $(t_1, x_1), \ldots, (t_n, x_n)$ (Françon, Viennot and Vuillemin (1978) and Vuillemin (1978)) is recursively defined by letting

$$s = \arg\min\{t_i : 1 \le i \le n\},$$

making (t_s, x_s) the root (with key x_s), attaching as a left subtree the Cartesian tree for $(t_i, x_i) : t_i > t_s, x_i < x_s$, and as a right subtree the Cartesian tree for $(t_i, x_i) : t_i > t_s, x_i > x_s$. We assume that all t_i's and x_i's are distinct. Note that the t_i's play the role of time stamps, and that the Cartesian tree is a heap with respect to the t_i's: along any path from the root down, the t_i's increase. It is a search tree with respect to the keys, the x_i's. Observe also that the Cartesian tree is unique, and invariant to permutations of the pairs (t_i, x_i).

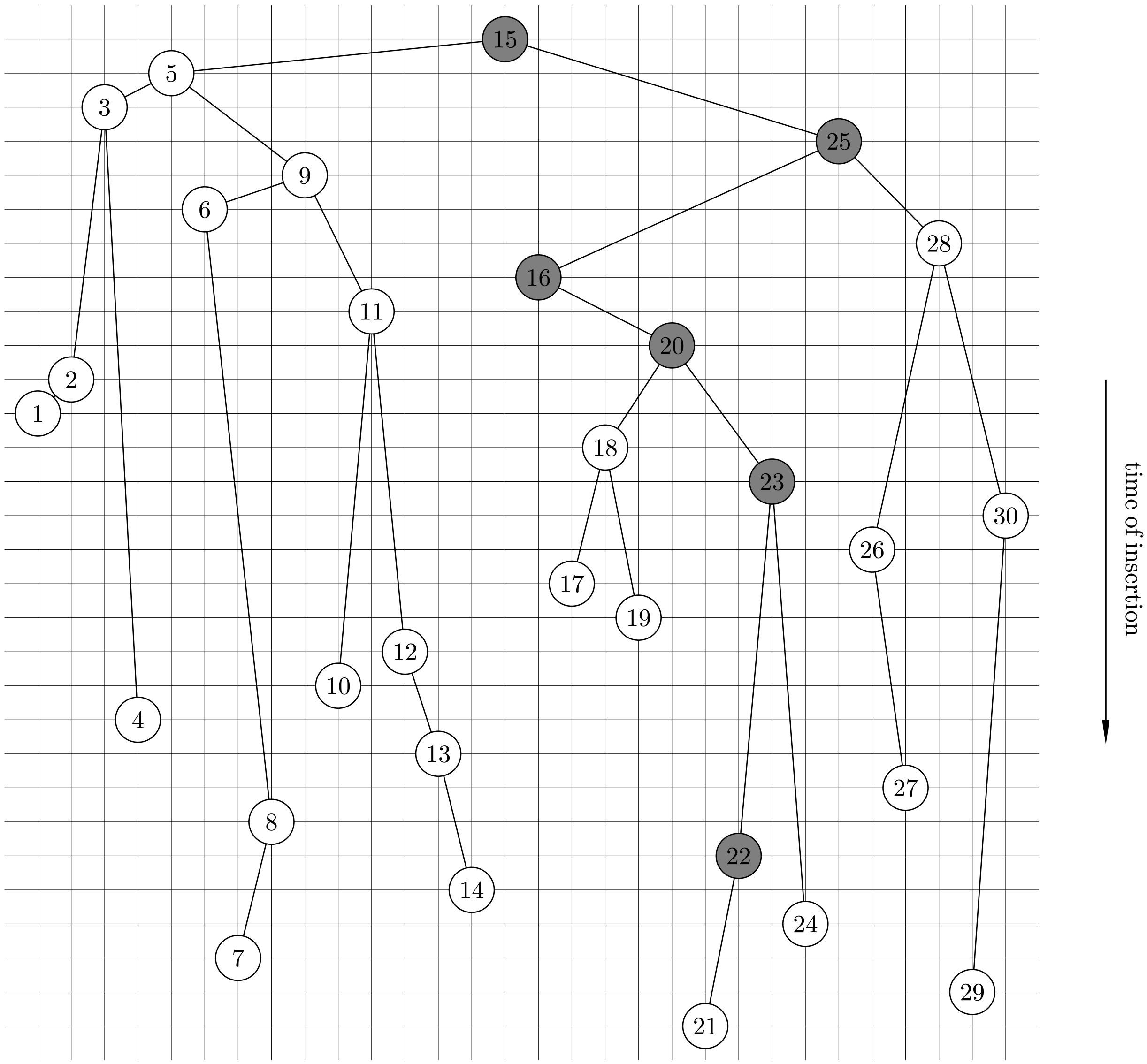

Figure 1. The binary search tree for the sequence 15, 5, 3, 25, 9, 6, 28, 16, 11, 20, 2, 1, 18, 23, 30, 26, 17, 19, 12, 10, 4, 13, 27, 8, 22, 14, 24, 7, 29, 21. The y-coordinates are the time stamps (the y-axis is pointing down) and the x-coordinates are the keys.

The keys are fixed in a deterministic manner. When searching for an element, we travel from the root down until the element is located. The complexity is equal to the length of the path between the element and the root, also called the depth of a node or element. The height of a tree is the maximal depth. Most papers on the subject deal with the choice of the t_i's to make the height or average depth acceptably small. This can trivially be done in time $O(n \log n)$ by sorting the x_i's, but one often wants to achieve this dynamically, by adding the

x_i's, one at a time, insisting that the n-th addition can be performed in time $O(\log n)$. One popular way of constructing a binary search tree is through randomization: first randomly and uniformly permute the x_i's, using the permutation $(\sigma_1, \ldots, \sigma_n)$, obtaining $(x_{\sigma_1}, \ldots, x_{\sigma_n})$, and then construct the Cartesian tree for

$$(1, x_{\sigma_1}), (2, x_{\sigma_2}), \ldots, (n, x_{\sigma_n}) \tag{1}$$

incrementally. This is the classical random binary search tree. This tree is identical to

$$(t_1, x_1), (t_2, x_2), \ldots, (t_n, x_n) \tag{2}$$

where $(t_1, t_2, \ldots, t_n)$ is the inverse random permutation, that is, $t_i = j$ if and only if $\sigma_j = i$. Since $(t_1, t_2, \ldots, t_n)$ itself is a uniform random permutation, we obtain yet another identically distributed tree if we had considered

$$(U_1, x_1), (U_2, x_2), \ldots, (U_n, x_n) \tag{3}$$

where the U_i's are i.i.d. uniform $[0, 1]$ random variables. Proposal (3) avoids the (small) problem of having to generate a random permutation, provided that we can easily add (U_n, x_n) to a Cartesian tree of size $n - 1$. This is the methodology adopted in the treaps (Aragon and Seidel, 1989, 1996), in which the U_i's are permanently attached to the x_i's.

But since the Cartesian tree is invariant to permutations of the pairs, we can assume without loss of generality in all three representations above that

$$x_1 < x_2 < \cdots < x_n,$$

and even that $x_i = i$. We will see that one of the simplest representations (or embeddings) is the one that uses the model

$$(U_1, 1), (U_2, 2), \ldots, (U_n, n). \tag{4}$$

Observe that neighboring pairs (U_i, i) and $(U_{i+1}, i+1)$ are necessarily related as ancestor and descendant, for if they are not, then there exists some node with key between i and $i + 1$ that is a common ancestor, but such a node does not exist. We write $j \sim i$ if (U_j, j) is an ancestor of (U_i, i). Finally, note that in (1), only the relative ranks of the x_{σ_i} variables matter, so that we also obtain a random binary search tree as the Cartesian tree for

$$(1, U_1), (2, U_2), \ldots, (n, U_n). \tag{5}$$

§2. The depths of the nodes

Consider the depth D_i of the node of rank i (i.e., of (U_i, i)). We have

$$D_i = \sum_{j<i} \mathbb{1}_{[j\sim i]} + \sum_{j>i} \mathbb{1}_{[j\sim i]}.$$

But for $j < i$, $j \sim i$ if and only if

$$U_j = \min(U_j, U_{j+1}, \ldots, U_i).$$

This happens with probability $1/(i-j+1)$.

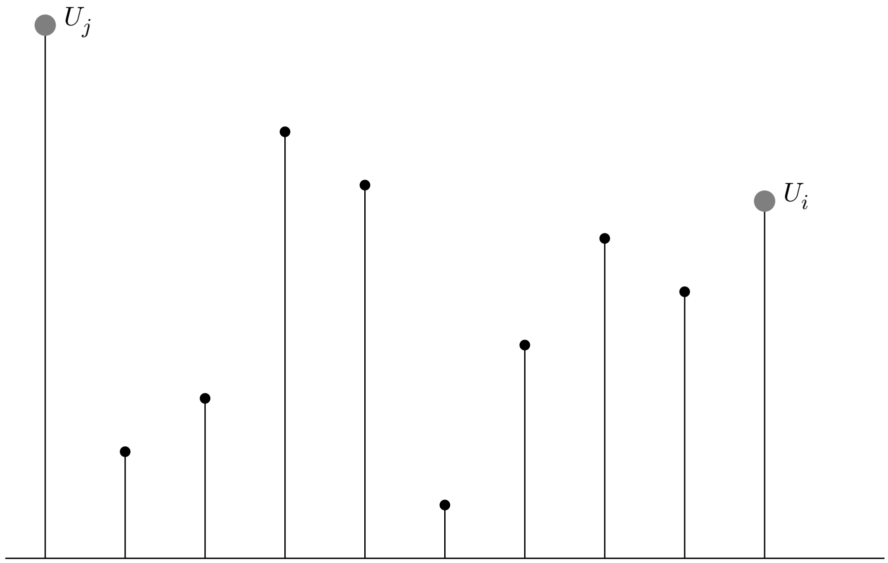

Figure 2. We show the time stamps $U_j, U_{j+1}, \ldots, U_i$ with the axis pointing down. Note that in this set-up, the j-th node is an ancestor of the i-th node, as it has the smallest time stamp among $U_j, U_{j+1}, \ldots, U_i$.

Therefore,

$$\begin{aligned}
\mathbb{E}\{D_i\} &= \sum_{j<i} \frac{1}{i-j+1} + \sum_{j>i} \frac{1}{j-i+1} \\
&= \sum_{j=2}^{i} \frac{1}{j} + \sum_{j=2}^{n-i+1} \frac{1}{j} \\
&= H_i + H_{n-i+1} - 2
\end{aligned}$$

where $H_n = \sum_{i=1}^{n}(1/i)$ is the n-th harmonic number. Thus,

$$\sup_{1\le i\le n} |\mathbb{E}\{D_i\} - \log(i(n-i+1))| \le 2.$$

We expect the binary search tree to hang deeper in the middle. With $\alpha \in (0,1)$, we have $\mathbb{E}\{D_{\alpha n}\} = 2\log n + \log(\alpha(1-\alpha)) + O(1)$.

Let us dig a little deeper and derive the limit law for D_i. It is convenient to work with the Cartesian tree in which U_i is replaced by 1. Then for $j < i$, $j \sim i$ if and only if

$$U_j = \min(U_j, U_{j+1}, \ldots, U_{i-1}),$$

which has probability $1/(i-j)$. Denoting the new depth by D'_i, we have obviously $D'_i \geq D_i$. More importantly, $\sum_{j<i} \mathbb{1}_{[j\sim i]}$ is the number of records in an i.i.d. sequence of $i-1$ uniform $[0,1]$ random variables. It is well-known that the summands are independent (exercise!) and therefore, we have

$$D'_i = \sum_{k=1}^{i-1} B(1/k) + \sum_{k=1}^{n-i} B'(1/k)$$

where $B(p)$ and $B'(p)$ are Bernoulli (p) random variables, and all Bernoulli variables in the sum are independent. We therefore have, first using linearity of expectation, and then independence,

$$\mathbb{E}\{D'_i\} = H_{i-1} + H_{n-i},$$

$$\begin{aligned}\mathbb{V}\{D'_i\} &= \sum_{k=1}^{i-1} \frac{1}{k}\left(1-\frac{1}{k}\right) + \sum_{k=1}^{n-i} \frac{1}{k}\left(1-\frac{1}{k}\right) \\ &= H_{i-1} + H_{n-i} - \theta\end{aligned}$$

where $0 \leq \theta \leq \pi^2/3$. It is amusing to note that

$$0 \leq \mathbb{E}\{D'_i - D_i\} = 2 + H_{i-1} + H_{n-i} - H_i - H_{n-i+1} \leq 2,$$

so the two random variables are not very different. In fact, by Markov's inequality, for $t > 0$,

$$\mathbb{P}\{D'_i - D_i \geq t\} \leq \frac{\mathbb{E}\{D'_i - D_i\}}{t} \leq \frac{2}{t}.$$

The Lindeberg-Feller version of the central limit theorem (Feller, 1968) applies to sums of independent but not necessarily identically distributed random variables X_i. A particular form of it, known as Lyapunov's central limit theorem, states that for bounded random variables ($\sup_i |X_i| \leq c < \infty$), provided that $\sum_{i=1}^n \mathbb{V}\{X_i\} \to \infty$,

$$\frac{\sum_{i=1}^n (X_i - \mathbb{E}\{X_i\})}{\sqrt{\sum_{i=1}^n \mathbb{V}\{X_i\}}} \xrightarrow{\mathcal{L}} \text{normal}(0,1)$$

as $n \to \infty$, where $Z_n \xrightarrow{\mathcal{L}} Z$ denotes convergence in distribution of the random variables Z_n to the random variable Z, that is, for all points of continuity x of the distribution function F of Z,

$$\lim_{n\to\infty} \mathbb{P}\{Z_n \leq x\} = F(x) = \mathbb{P}\{Z \leq x\}.$$

The normal distribution function is denoted by Φ. Applied to D'_i, we have to a bit careful, because the dependence of i on n needs to be specified. While one can easily treat specific cases of such dependence, it is perhaps more informative to treat all cases together! This can be done by making use of the Berry-Esseen inequality (Esseen, 1945, see also Feller, 1968).

LEMMA 1 (BERRY-ESSEEN INEQUALITY). *Let $X_1, X_2, \ldots, X_n$ be independent random variables with zero means and finite variances. Put*

$$S_n = \sum_{i=1}^{n} X_i, B_n = \sqrt{\sum_{i=1}^{n} \mathbb{E}\{X_i^2\}}.$$

If $\mathbb{E}\{|X_i|^3\} < \infty$, then

$$\sup_x \left| \mathbb{P}\left\{ \frac{S_n}{B_n} \leq x \right\} - \Phi(x) \right| \leq \frac{C \sum_{i=1}^{n} \mathbb{E}\{|X_i|^3\}}{B_n^3}$$

where $C = 0.7655$.

The constant C given in Lemma 1 is due to Shiganov (1986), who improved on Van Beek's constant 0.7975 (1972). If we apply this Lemma to

$$D'_i - \mathbb{E}\{D'_i\} = \sum_{k=1}^{i-1} (B(1/k) - (1/k)) + \sum_{k=1}^{n-i} (B'(1/k) - (1/k)),$$

then the the "B_n" in Lemma 1 is nothing but $\sqrt{\mathbb{V}\{D'_i\}}$, which we computed earlier: it is $\sqrt{\log(i(n-i+1))} + O(1)$ uniformly over all i. Its minimal value over all i is $\sqrt{\log n} - O(1)$. Finally, recalling that

$$\mathbb{E}\{|B(p) - p|^3\} \leq \mathbb{E}\{(B(p) - p)^2\} = p(1-p) \leq p,$$

we note that

$$\begin{aligned} \sup_{x,i} \left| \mathbb{P}\left\{ \frac{D'_i - \mathbb{E}\{D'_i\}}{\sqrt{\mathbb{V}\{D'_i\}}} \leq x \right\} - \Phi(x) \right| &\leq \sup_i \frac{C \times 2H_n}{(\mathbb{V}\{D'_i\})^{3/2}} \\ &= \frac{C \times 2H_n}{(\sqrt{\log n} - O(1))^3} \\ &= O(1/\sqrt{\log n}). \end{aligned}$$

The uniformity both with respect to x and i is quite striking.

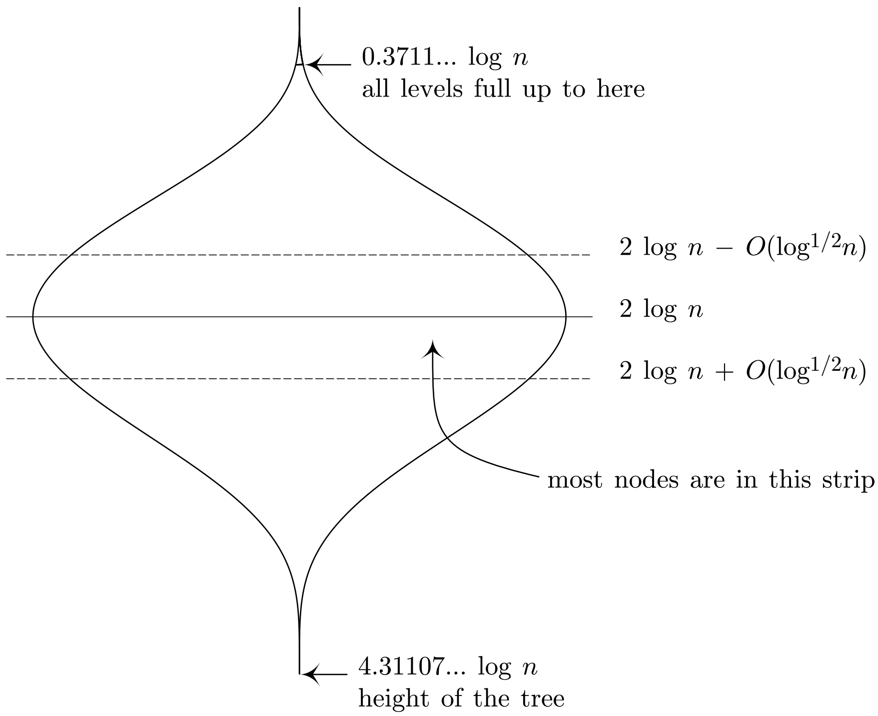

Figure 3. This figure illustrates the shape of a random binary search tree. The root is at the top, while the horizontal width represents the number of nodes at each level, also called the profile of a tree. Up to $0.3711\ldots \log n$, all levels are full with high probability. Also, the height (maximal depth) is $4.31107\ldots \log n + O(\log\log n)$ with high probability. The limit law mentioned above implies that nearly all nodes are in a strip of width $O(\sqrt{\log n})$ about $2\log n$. Finally, the profile around the peak of the belly is gaussian.

Having obtained a limit law for D_i', it should now be possible to to the same for D_i, which of course has dependent summands. We observe that for fixed t,

$$\begin{aligned}
&\mathbb{P}\left\{D_i - \mathbb{E}\{D_i\} \le x\sqrt{\mathbb{V}\{D_i'\}}\right\} \\
&\quad \le \mathbb{P}\left\{D_i' - \mathbb{E}\{D_i'\} \le x\sqrt{\mathbb{V}\{D_i'\}} + t + 2\right\} + \mathbb{P}\left\{|D_i' - \mathbb{E}\{D_i'\} - (D_i - \mathbb{E}\{D_i\})| \ge t + 2\right\} \\
&\quad \le \Phi(x) + o(1) + \mathbb{P}\{D_i' - D_i \ge t\} \\
&\quad \le \Phi(x) + o(1) + \frac{2}{t},
\end{aligned}$$

which is as close to $\Phi(x)$ as desired by choice of t. A similar lower bound can be derived, and therefore,

$$\lim_{n\to\infty} \sup_{x,i} \left| \mathbb{P}\left\{\frac{D_i - \mathbb{E}\{D_i\}}{\sqrt{\mathbb{V}\{D_i'\}}} \le x\right\} - \Phi(x)\right| = 0.$$

One can even obtain a rate by choice of t that is marginally worse than $O(1/\sqrt{\log n})$.

The depth of the last node added (the one with the maximal time stamp) in the classical random binary search tree (1) is distributed as D'_N, where N is a uniform random index from $1, 2, \ldots, n$, is also asymptotically normal:

$$\lim_{n\to\infty} \mathbb{P}\left\{D'_N - \mathbb{E}\{D'_N\} \le x\sqrt{\mathbb{V}\{D'_N\}}\right\} = \Phi(x).$$

This follows from the estimates above with a little extra work (see Devroye (1988) for an alternative proof). It is also known that $\mathbb{E}\{D'_N\} = 2(H_n - 1)$ and that $\mathbb{V}\{D'_N\} = \mathbb{E}\{D'_N\} - O(1)$. We note here that the exact distribution of D'_N has been known for some time (Lynch, 1965; Knuth, 1973):

$$\mathbb{P}\{D'_N = k\} = \frac{1}{n!}\begin{bmatrix} n-1 \\ k \end{bmatrix} 2^k \ , \ 1 \le k < n \ ,$$

where [.] denotes the Stirling number of the first kind. In these references, we also find the weak law of large numbers: $D'_N/(2\log n) \to 1$ in probability. The same law of large numbers also holds for D_N, the depth of a randomly selected node, and the limit law for D_N is like that for D'_N (Louchard, 1987).

§3. The number of leaves

Some parameters of the random binary search tree are particularly easy to study. One of those is the number of leaves, L_n. Using embedding (4) to study this, we note that (U_i, i) is a leaf if and only if $(U_{i-1}, i-1)$ and $(U_{i+1}, i+1)$ are ancestors, that is, if and only if

$$U_i = \max(U_{i-1}, U_i, U_{i+1}).$$

Thus,

$$L_n \stackrel{\mathcal{L}}{=} \mathbb{1}_{[U_1 > U_2]} + \sum_{i=2}^{n-1} \mathbb{1}_{[U_i > \max(U_{i-1}, U_{i+1})]} + \mathbb{1}_{[U_n > U_{n-1}]} \stackrel{\text{def}}{=} \mathbb{1}_{[U_1 > U_2]} + L'_n + \mathbb{1}_{[U_n > U_{n-1}]}.$$

This is an n-term sum of random variables that are 2-dependent, where we say that a sequence $X_1, X_2, \ldots, X_n$ is k-dependent if $X_1, \ldots, X_\ell$ is independent of $X_{\ell+k+1}, \ldots, X_n$ for all ℓ. 0-dependence corresponds to independence.

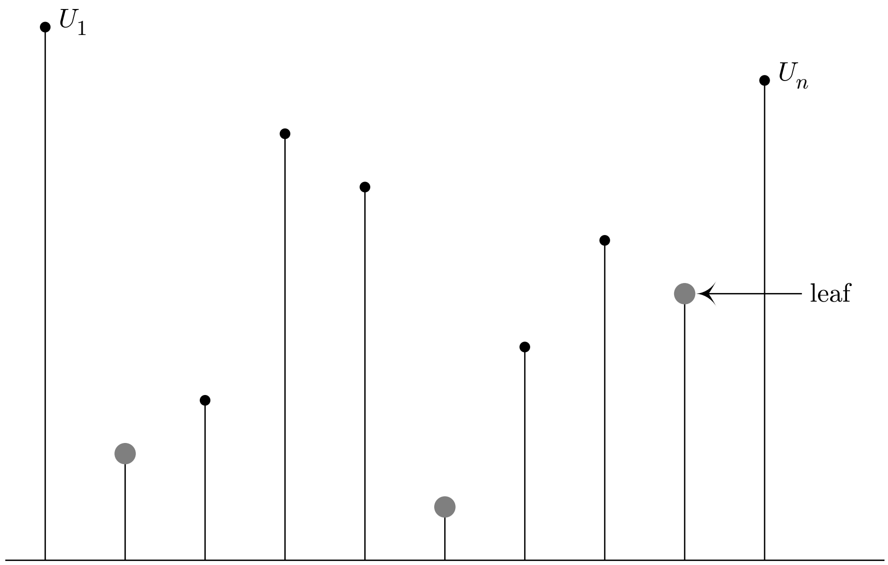

Figure 4. For a tree with these time stamps (small on top), the leaves (marked) are those nodes whose stamps are larger than those of their neighbors.

Let $\mathcal{N}(0,\sigma^2)$ denote the normal distribution with mean 0 and variance σ^2. We will use a simple version of the central limit theorem for k-dependent stationary sequences due to Hoeffding and Robbins (1949):

LEMMA 2. *Let $Z_1,\ldots,Z_n,\ldots$ be a stationary sequence of random variables (i.e., for any ℓ, the distribution of $(Z_i,\ldots,Z_{i+\ell})$ does not depend upon i), and let it also be k–dependent with k held fixed. Then, if $\mathbb{E}|Z_1|^3<\infty$, the random variable*

$$\frac{\sum_{i=1}^n (Z_i-\mathbb{E}\{Z_i\})}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,\sigma^2)$$

where

$$\sigma^2=\mathbb{V}\{Z_1\}+2\sum_{i=2}^{k+1}(\mathbb{E}\{Z_1Z_i\}-\mathbb{E}\{Z_1\}\mathbb{E}\{Z_i\}).$$

The standard central limit theorem for independent (or 0-dependent) random variables is obtained as a special case. Subsequent generalizations of Lemma 1 were obtained by Brown (1971), Dvoretzky (1972), McLeish (1974), Ibragimov (1975), Chen (1978), Hall and Heyde (1980) and Bradley (1981), to name just a few.

Clearly,

$$\mathbb{E}\{L_n\}=\frac{2}{2}+\frac{n-2}{3}=\frac{n+1}{3},$$

a result first obtained by Mahmoud (1986). We study L'_n as $|L_n - L'_n| \le 2$. Its summands are identically distributed and of mean 1/3. Thus,

$$\frac{L'_n - (n-2)/3}{\sqrt{n-2}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

where

$$\sigma^2 = \mathbb{V}\{Z_2\} + 2\sum_{i=3}^{4} (\mathbb{E}\{Z_2 Z_i\} - \mathbb{E}\{Z_2\}\mathbb{E}\{Z_i\}).$$

Here $Z_i = \mathbb{1}_{[U_i]} = \max(U_{i-1}, U_i, U_{i+1})$. As Z_i is Bernoulli (1/3), we have $\mathbb{V}\{Z_i\} = (1/3) \cdot (2/3) = 2/9$. Furthermore, $Z_2 Z_3 = 0$ as neighboring nodes cannot both be leaves. Thus,

$$\sigma^2 = 2(\mathbb{E}\{Z_2 Z_4\} - \mathbb{E}\{Z_2\}\mathbb{E}\{Z_4\}).$$

We have

$$\mathbb{E}\{Z_2 Z_4\} = \frac{2}{15}$$

by a quick combinatorial argument. Thus, $\sigma^2 = 2/45$. It is then trivial to verify that

$$\frac{L_n - n/3}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}\left(0, \frac{2}{45}\right).$$

Let O_n denote the number of nodes with one child and let T_n be the number of nodes with two children. The quantities L_n, O_n and T_n are closely related, since

$$L_n + O_n + T_n = n, L_n = T_n + 1.$$

This implies that $O_n = n + 1 - 2L_n$, $T_n = L_n - 1$. Thus, $\mathbb{E}L_n \sim n/3$ implies the same thing for $\mathbb{E}O_n$ and $\mathbb{E}L_n$. Furthermore, $\mathbb{V}(O_n) \sim 4\mathbb{V}(L_n) \sim \mathbb{V}(T_n)$. Therefore, T_n follows the same limit laws as L_n, while $(O_n - n/3)/\sqrt{n}$ converges in distribution to a normal distribution with zero mean and variance 8/45.

§4. Local counters

Next we consider parameters that describe the number of nodes having a certain "local" property. The purpose of this section is to generalize the method of the previous section towards all local parameters of a tree, also called local counters. Besides L_n, we may consider V_{kn}, the number of nodes with k proper descendants, or L_{kn}: the number of nodes with k proper left descendants. Aldous (1990) gives a general methodology based upon urn models and branching processes for obtaining the first order behavior of the local quantities; his methods apply to a wide variety of trees; for the binary search tree, he has shown, among other things, that $V_{kn}/n \to 2/(k+2)(k+3)$ in probability as $n \to \infty$. We will give a short proof of this, and obtain the limit law for V_{kn} as well.

We call a random variable N_n defined on a random binary search tree a local counter of order k if in embedding (4) it can be written in the form

$$N_n = \sum_{i=1}^{n} f(U_{i-k}, \ldots, U_i, \ldots, U_{i+k}),$$

where k is a fixed constant, $U_i = 0$ if $i \leq 0$ or $i > n$, and f is a $\{0,1\}$-valued function that is invariant under transformations of the U_i's that keep the relative order of the arguments intact. Clearly, L_n is a local counter of order one. Local counters have two key properties:

A. The i-th and j-th terms in the definition of N_n are independent whenever $|i-j| > 2k$.

B. The distribution of the i-th term is the same for all $i \in \{k+1, \ldots, n-k\}$. Thus, we have the representation $N_n = A_n + \sum_{i=k+1}^{n-k} Z_i$, where $0 \leq A_n \leq 2k$, and where the Z_i's are identically distributed and $2k$-dependent.

As a corollary of Lemma 2, we have the following limit law.

Theorem 1 (devroye, 1991). *Let N_n be a local counter for a random binary search tree, with fixed parameter k. Define $Z_i = f(U_i, \ldots, U_{i+2k})$, where $U_1, U_2, \ldots$ is a sequence of i.i.d. uniform $[0,1]$ random variables. Then*

$$\frac{N_n - n\mathbb{E}\{Z_1\}}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2),$$

where

$$\sigma^2 = \mathbb{V}\{Z_1\} + 2\sum_{i=2}^{2k+1} (\mathbb{E}\{Z_1 Z_i\} - \mathbb{E}\{Z_1\}\mathbb{E}\{Z_i\}).$$

If $\mathbb{E}\{Z_1\} \neq 0$, then $N_n/n \to \mathbb{E}\{Z_1\}$ in probability and in the mean as $n \to \infty$.

Proof. The random variable $N_n - 2k$ is distributed as $\sum_{i=1}^{n-2k} Z_i$, and satisfies the conditions of Lemma 2. Thus,

$$\frac{N_n - 2k - (n-2k)\mathbb{E}\{Z_1\}}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2).$$

Here we used the fact that the Z_i's are $2k$-dependent. But

$$\left| \frac{N_n - n\mathbb{E}\{Z_1\}}{\sqrt{n}} - \frac{N_n - 2k - (n-2k)\mathbb{E}\{Z_1\}}{\sqrt{n}} \right| \leq \frac{4k}{\sqrt{n}} = o(1),$$

so that the first statement of Theorem 1 follows without work. The second statement follows from the first one. □

Let k be fixed, independent of n. Simple considerations show that V_{kn}, the number of nodes with precisely k descendants, is indeed a local counter. Note that all the proper descendants of a node (U_i, i) are found by finding the largest $0 \le j < i$ with $U_j < U_i$, and the smallest ℓ greater than i and no more than n such that $U_\ell < U_i$. All the nodes $(U_{j+1}, j+1), \ldots, (U_{\ell-1}, \ell-1)$, (U_i, i) excluded, are proper descendants of (U_i, i). Thus, to decide whether (U_i, i) has exactly k descendants, it suffices to look at $(U_{i-k-1}, i-k-1), \ldots, (U_{i+k+1}, i+k+1)$, so that V_{kn} is a local counter with parameter $k+1$.

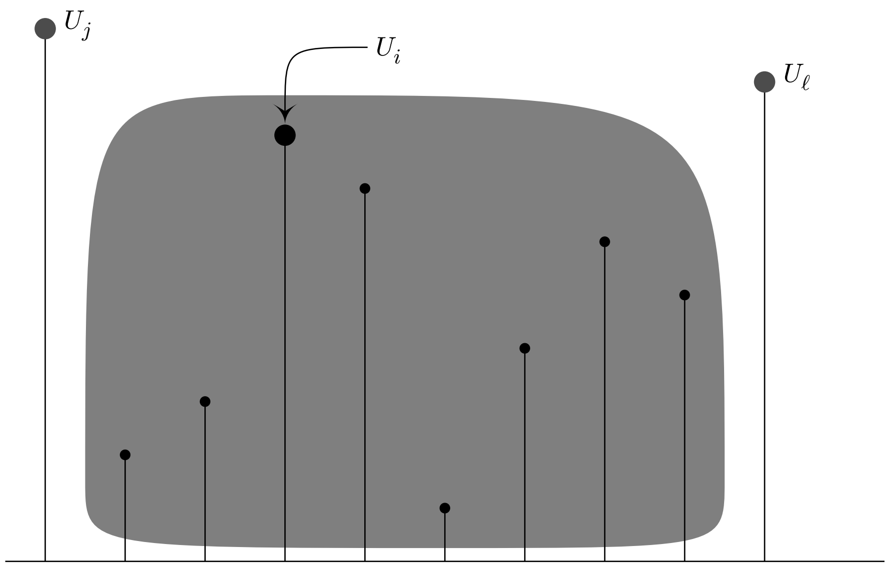

Figure 5. The subtree rooted at the i-th node consists of all the nodes in the shaded area. To the left and right, the subtree is "guarded" by nodes with smaller time stamps, U_j and U_ℓ.

Theorem 1 above implies the following (Devroye, 1991):

THEOREM 2. *Let V_{kn} be the number of nodes with k proper descendants. Then*

$$\frac{V_{kn}}{n} \longrightarrow p_k \stackrel{\text{def}}{=} \frac{2}{(k+3)(k+2)} \quad \textit{in probability}$$

and

$$\frac{V_{kn} - np_k}{\sqrt{n}} \longrightarrow \mathcal{N}(0, \sigma_k^2) \quad \textit{in distribution}$$

as $n \to \infty$, where

$$\sigma_k^2 \stackrel{\text{def}}{=} p_k(1-p_k) + 2(k+1)^2 \rho_k - 2(k+2)p_k^2$$

and

$$\rho_k \stackrel{\text{def}}{=} \frac{5k+8}{(k+1)^2(k+2)^2(2k+5)(2k+3)} \ .$$

Remark. The first part of this theorem is also implicit in Aldous (1990). □

Remark. When $k=0$, we get $p_0=1/3$, $\rho_0=2/15$, $\sigma_0^2=2/45$. For $k=1$, we obtain $p_1=1/6$, $\rho_1=13/1260$ and $\sigma_1^2=23/420$. □

Proof. We have the representation

$$V_{kn}=\sum_{i=1}^{n} Z_i \ ,$$

where

$$Z_i \stackrel{\text{def}}{=} \sum_{j=0}^{k} Z_i(j,k-j) \ ,$$

and $Z_i(j,\ell)$ is the indicator of the event that $(X_{(i)},U_i)$ has j left descendants and ℓ right descendants. Assume throughout that $1\le i-k-1$, $i+k+1\le n$ when we discuss Z_i. The values $Z_1,\ldots,Z_{k+1}$ and $Z_{n-k},\ldots,Z_n$ are all zero or one, and affect V_{kn} jointly by at most $2k+2$ (which is a constant). We also have the representation, for $1\le i-k-1$, $i+\ell+1\le n$:

$$\begin{aligned} Z_i(j,\ell) = \mathbb{1}&_{[U_{i-j-1}<U_i<\min(U_{i-1},\ldots,U_{i-j})]} \\ \times \mathbb{1}&_{[U_{i+\ell+1}<U_i<\min(U_{i+1},\ldots,U_{i+\ell})]} \ . \end{aligned}$$

A simple argument shows that for i,j,ℓ as restricted above,

$$\mathbb{E}\{Z_i(j,\ell)\}=\frac{2(j+\ell)!}{(j+\ell+3)!}=\frac{2}{(j+\ell+3)(j+\ell+2)(j+\ell+1)} \ .$$

Thus, for $1\le i-k-1$, $i+k+1\le n$,

$$\mathbb{E}\{Z_i(j,k-j)\}=\frac{2}{(k+3)(k+2)(k+1)} \stackrel{\text{def}}{=} q_k \ ,$$

and

$$\mathbb{E}\{Z_i\}=\sum_{j=0}^{k}\mathbb{E}\{Z_i(j,k-j)\}=\sum_{j=0}^{k} q_k=\frac{2}{(k+3)(k+2)} \ .$$

It is clear that V_{kn} is a local counter for a random binary search tree, so we may apply Theorem 1. To do so, we need to study $\mathbb{E}\{Z_iZ_{i+r}\}$, where $1\le i-k-1$, $i+r+k+1\le n$, $1\le i\le n$, $1\le r$. For $0\le j\le k$, $0\le \ell\le n$, we claim the following:

$$\mathbb{E}\{Z_i(j,k-j)Z_{i+r}(\ell,k-\ell)\}=\begin{cases} 0 \ , & \text{if } r<k-j+\ell+2; \\ \rho_k \ , & \text{if } r=k-j+\ell+2; \\ \mathbb{E}\{Z_i(j,k-j)\}\mathbb{E}\{Z_{i+r}(\ell,k-\ell)\} \ , & \text{if } r>k-j+\ell+2, \end{cases}$$

where

$$\rho_k \stackrel{\text{def}}{=} \frac{k!^2}{(2k+5)!}\left\{\binom{2k+4}{k+2} + 2\binom{2k+3}{k+2} + 2\binom{2k+2}{k+1}\right\}$$
$$= \frac{5k+8}{(k+1)^2(k+2)^2(2k+5)(2k+3)} \ .$$

The last expression is obtained by noting that of the $(2k+5)!$ possible permutations of $U_{i-j-1}, \dots, U_{i+r+k-\ell+1}$, with $r = k - j + \ell + 2$, only $\rho_k(2k+5)!$ are such that $Z_i(j, k-j)Z_{i+r}(\ell, k-\ell) = 1$. The three terms in the expression of ρ_k are obtained by considering

A. $U_{i+k-j+1}$ is smaller than both U_{i-j-1} and $U_{i+r+k-\ell+1}$.

B. $U_{i+k-j+1}$ is smaller than one of U_{i-j-1} and $U_{i+r+k-\ell+1}$.

C. $U_{i+k-j+1}$ is larger than both U_{i-j-1} and $U_{i+r+k-\ell+1}$.

If $r > 2k+2$, then Z_i and Z_{i+r} are independent. Thus, we need only consider the case $1 \le r \le 2k+2$. Let L, J be independent random variables uniformly distributed on $\{0, \dots, k\}$.

$$\mathbb{E}\{Z_i Z_{i+r}\} = \mathbb{E}\left\{\sum_{j=0}^{k} Z_i(j, k-j) \sum_{\ell=0}^{k} Z_{i+r}(\ell, k-\ell)\right\}$$
$$= \sum_{j=0}^{k}\sum_{\ell=0}^{k} \rho_k \mathbb{1}_{[r = k-j+\ell+2]} + \sum_{j=0}^{k}\sum_{\ell=0}^{k} q_k^2 \mathbb{1}_{[r > k-j+\ell+2]}$$
$$= (k+1)^2 \rho_k \mathbb{P}\{r = k - J + L + 2\} + (k+1)^2 q_k^2 \mathbb{P}\{r > k - J + L + 2\} \ .$$

Summing this gives

$$\sum_{r=1}^{2k+2} (\mathbb{E}\{Z_i Z_{i+r}\} - \mathbb{E}\{Z_i\}\mathbb{E}\{Z_{i+r}\})$$
$$= (k+1)^2 \rho_k \sum_{r=1}^{2k+2} \mathbb{P}\{J - L = k + 2 - r\}$$
$$+ (k+1)^2 q_k^2 \sum_{r=1}^{2k+2} \mathbb{P}\{J - L > k + 2 - r\} - (2k+2)p_k^2$$
$$= (k+1)^2 \rho_k + p_k^2 \sum_{r=1}^{k+1} \mathbb{P}\{J - L > k + 2 - r\}$$
$$+ p_k^2 \sum_{r=k+2}^{2k+2} \mathbb{P}\{J - L > k + 2 - r\} - (2k+2)p_k^2$$
$$= (k+1)^2 \rho_k + p_k^2 \sum_{r=1}^{k+1} \mathbb{P}\{J - L > r\}$$

$$
\begin{aligned}
&+ p_k^2 \sum_{r=0}^{k} \mathbb{P}\{J - L > -r\} - (2k+2)p_k^2 \\
&= (k+1)^2 \rho_k + p_k^2 \sum_{r=2}^{k+2} \mathbb{P}\{J - L \geq r\} \\
&+ p_k^2 \sum_{r=0}^{k} (1 - \mathbb{P}\{J - L \geq r\}) - (2k+2)p_k^2 \\
&= (k+1)^2 \rho_k - p_k^2 \sum_{r=0}^{1} \mathbb{P}\{J - L \geq r\} + p_k^2(k+1) - (2k+2)p_k^2 \\
&= (k+1)^2 \rho_k - p_k^2 + p_k^2(k+1) - (2k+2)p_k^2 \\
&= (k+1)^2 \rho_k - (k+2)p_k^2 \ .
\end{aligned}
$$

By Lemma 1, $V_{kn}/n \to p_k$ in probability as $n \to \infty$ and

$$
\frac{V_{kn} - np_k}{\sqrt{n}} \longrightarrow \mathcal{N}(0, \sigma_k^2) \quad \text{in distribution,}
$$

where

$$
\sigma_k^2 = p_k(1 - p_k) + 2(k+1)^2 \rho_k - 2(k+2)p_k^2 \ . \ \square
$$

§5. Urn models

The limit law for L_n can be obtained by several methods. For example, Poblete and Munro (1985) use the properties of Pólya-Eggenberger urn models for the analysis of search trees. Bagchi and Pal (1985) developed a limit law for general urn models and applied it in the analysis of random 2-3 trees.

In a binary search tree with n nodes, let W_n be the number of external nodes with another sibling external node, and let B_n count the remaining external nodes. Clearly, $W_n + B_n = n+1$, $W_0 = 0$ and $B_0 = 1$. When a random binary search tree is grown, each external node is picked independently with equal probability (see, e.g., Knuth and Schönhage, 1978). Thus, upon insertion of node $n+1$, we have:

$$
(W_{n+1}, B_{n+1}) = (W_n, B_n) + \begin{cases} (0,1) & \text{with probability } \frac{W_n}{W_n + B_n}; \\ (2,-1) & \text{with probability } \frac{B_n}{W_n + B_n}. \end{cases}
$$

This is known as a generalized Pólya-Eggenberger urn model. The model is defined by the matrix

$$
\begin{pmatrix} a & b \\ c & d \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 2 & -1 \end{pmatrix}
$$

For general values of a, b, c, d, the asymptotic behavior of W_n is governed by the following Lemma (Bagchi and Pal, 1985) (for a special case, see e.g. Athreya and Ney, 1972):

LEMMA 3. *Consider an urn model in which* $a+b=c+d \stackrel{\text{def}}{=} s \geq 1$, $W_0+B_0 \geq 1$, $0 \leq W_0$, $0 \leq B_0$, $a \neq c$, $b, c > 0$, $a-c \leq s/2$, *and, if* $a<0$, *then* a *divides both* c *and* W_0, *and if* $d<0$, *then* d *divides both* b *and* B_0. *Then*

$$\frac{W_n}{W_n+B_n} \to \frac{c}{b+c} \quad \text{almost surely},$$

and

$$\frac{W_n - \mathbb{E}\{W_n\}}{\sqrt{n}} \longrightarrow \mathcal{N}(0,\sigma^2) \quad \textit{in distribution},$$

where

$$\sigma^2 = \frac{bc}{(b+c)^2}\frac{(s-b-c)^2}{2b+2c-s}.$$

For L_n, we have $\sigma^2 = 8/45$. Since $L_n = W_n/2$, the variance of L_n is one fourth that of W_n, so Theorem 1 follows immediately from Lemma 3 as well. Additionally, Lemma 3 implies that $W_n/n \to 1/3$ almost surely for our way of growing the tree.

§6. Berry-Esseen laws for k-dependent sequences

For local counters in which k grows with n, a normal limit law can be obtained by any of a number of generalizations of the Berry-Esseen inequality to k-dependent sequences. One of the most versatile ones, given for k-dependent random fields, is due to Chen and Shao (2004). We say that a sequence of random variables $X_1,\ldots,X_n$ is k-dependent if $\{X_i : i \in A\}$ is independent of $\{X_i : i \in B\}$ whenever $\min\{|i-j| : i \in A, j \in B\} > k$.

LEMMA 4. *If* $X_1, X_2, \ldots, X_n$ *is a* k*-dependent sequence with zero means and* $\mathbb{E}\{|X_i|^3\} < \infty$ *for all* i, *then, setting*

$$S_n = \sum_{i=1}^n X_i, B_n = \sqrt{\mathbb{V}\{S_n\}},$$

we have

$$\sup_x \left| \mathbb{P}\left\{\frac{S_n}{B_n} \leq x\right\} - \Phi(x) \right| \leq \frac{60(k+1)^2 \sum_{i=1}^n \mathbb{E}\{|X_i|^3\}}{B_n^3}.$$

Shergin (1979) had previously obtained this result with a non-explicit constant. The paper by Chen and Shao (2004) also offers a non-uniform bound in which the factor 60 is replaced by $C/(1+|x|)^3$. Lemma 4 can be used to obtain limit laws for counters that have $k \to \infty$. In a further section, we give a related limit law, which, like Lemma 4, can be derived from Stein's method. We will describe its applications there.

§7. Quicksort

Quicksort is a fast and simple sorting method in which we take a random element from a set of n elements that need to be sorted, make it the pivot, and split the set of elements about this pivot (see figure below) using $n-1$ comparisons.

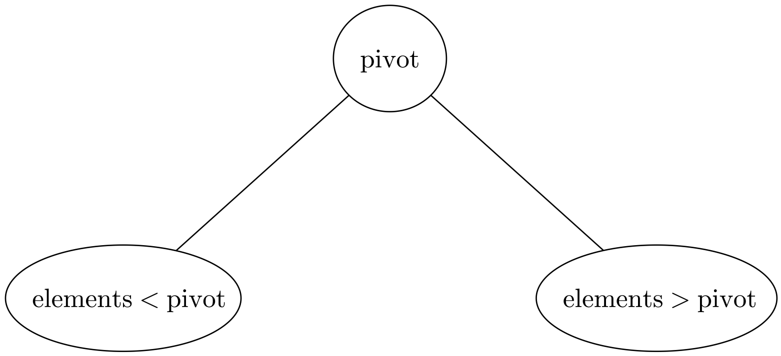

Figure 6. Basic quicksort step: the pivot about which a set of elements is split is the first element of the set.

It is well-known that the standard version of quicksort takes $\sim 2n \log n$ comparisons on the average. A particular execution of quicksort can be pictured as a binary tree: the pivot determines the root, and we apply the data splits recursively as in the construction of a binary search tree from the same data. It is easy to see that the binary search tree that explains quicksort is in fact a random binary search tree. We can count the number of comparisons by counting for each element how often it is involved in a comparison as a non-pivot element. For an element at depth D_i in the binary search tree, this happens precisely D_i times. Thus, the total number of comparisons (C_n) is given by

$$C_n = \sum_{i=1}^{n} D_i \ ,$$

where $D_1, \ldots, D_n$ is as for a random binary search tree. Therefore,

$$\begin{aligned} \mathbb{E}\{C_n\} &= \sum_{i=1}^{n} \mathbb{E}\{D_i\} \\ &= \sum_{i=1}^{n} (H_i + H_{n-i+1} - 2) \\ &= \sum_{i=1}^{n} 2(H_i - 1) \\ &= 2nH_n + 2H_n - 4n \\ &\sim 2n \log n \ . \end{aligned}$$

A problem occurs when we try looking at $\mathbb{V}\{C_n\}$ because the D_i's are very dependent. Thus, the variance of the sum no longer is the sum of the variances. If it were, we would obtain that $\mathbb{V}\{C_n\} \sim 2n \log n$. But other studies (e.g., Sedgewick, 1983; Sedgewick and Flajolet, 1986) tell us that

$$\mathbb{V}\{C_n\} \sim \left(7 - \frac{2\pi^2}{3}\right) n^2 \ .$$

Furthermore, the asymptotic distribution of $(C_n - 2n \log n)/n$ is not normal as one would expect after seeing that C_n can be written as a simple sum. In other words, the study of C_n requires another approach. We will tackle it by the contraction method in a future section, and show that it is just beyond the area of application of Stein's method. It is noteworthy that

$$C_n = \sum_{i=1}^{n} (N_i - 1),$$

where N_i is the size of the subtree rooted at the i-th node in the random binary search tree. In other words, C_n is a special case of a toll function, a sum of functions of subtree sizes. These parameters will be studied in the next few sections.

§8. Random binary search tree parameters

Most shape-related quantities of the random binary search tree have been well–studied, including the expected depth and the exact distribution of the depth of the last inserted node (Knuth, 1973; Lynch, 1965), the limit theory for the depth (Mahmoud and Pittel, 1984, Devroye, 1988), the first two moments of the internal path length (Sedgewick, 1983), the limit theory for the height of the tree (Pittel, 1984; Devroye, 1986, 1987), and various connections with the theory of random permutations (Sedgewick, 1983) and the theory of records (Devroye, 1988). Surveys of known results can be found in Vitter and Flajolet (1990), Mahmoud (1992) and Gonnet (1984).

Consider tree parameters for random binary search trees that can be written as

$$X_n = \sum_u f(S(u)),$$

where f be a mapping from the space of all permutations to the real line, $S(u)$ is the random permutation associated with the subtree rooted at node u in the random binary search tree, which is considered in format (4), and the summation is with respect to all nodes u in the tree.

The root of the binary search tree contains that pair (U_i, i) with smallest U_i value, the left subtree contains all pairs (U_j, j) with $j < i$, and the right subtree contains those pairs with $j > i$. Each node u can thus (recursively) be associated with a subset $S(u)$ of $(U_1, 1), \ldots, (U_n, n)$, namely that subset that correponds to nodes in its subtree.

With this embedding and representation, X_n is a sum over all nodes of a certain function of the permutation associated with each node. As each permutation uniquely determines subtree shape, a special case includes the functions of subtree shapes.

EXAMPLE 1: THE TOLL FUNCTIONS. In the first class of applications, we let $N(u)$ be the size of the subtree rooted at u, and set $f(S(u)) = g(|S(u)|)$:

$$X_n = \sum_u g(N(u)) \ .$$

Examples of such tree parameters abound:

A. If $g(n) \equiv 1$ for $n > 0$, then $X_n = n$.

B. If $g(n) = \mathbb{1}_{[n=k]}$ for fixed $k > 0$, then X_n counts the number of subtrees of size k.

C. If $g(n) = \mathbb{1}_{[n=1]}$, then X_n counts the number of leaves.

D. If $g(n) = n - 1$ for $n > 1$, then X_n counts the number of comparisons in classical quicksort.

E. If $g(n) = \log_2 n$ for $n > 0$, then X_n is the logarithm base two of the product of all subtree sizes.

F. If $g(n) = \mathbb{1}_{[n=1]} - \mathbb{1}_{[n=2]}$ for $n > 0$, then X_n counts the number of nodes in the tree that have two children, one of which is a leaf.

EXAMPLE 2: TREE PATTERNS. Fix a tree T. We write $S(u) \approx T$ if the subtree at u defined by the permutation $S(u)$ is equal to T, where equality of trees refers to shape only, not node labeling. Note that at least one, and possibly many permutations with $|S(u)| = |T|$, may give rise to T. If we set

$$X_n = \sum_u \mathbb{1}_{[S(u) \approx T]}$$

then X_n counts the number of subtrees precisely equal to T. Note that these subtrees are necessarily disjoint. We are tempted to call them suffix tree patterns, as they hug the bottom of the binary search tree.

Example 3: Prefix tree patterns. Fix a tree T. We write $S(u) \supset T$ if the subtree at u defined by the permutation $S(u)$ consists of T (rooted now at u) and possibly other nodes obtained by replacing all external nodes of T by new subtrees. Define

$$X_n = \sum_u \mathbb{1}_{[S(u) \supset T]}$$

For example, if T is a single node, then X_n counts the number of nodes, n. If T is a complete subtree of size $2^{k+1}-1$ and height k, then X_n counts the number of occurrences of this complete subtree pattern (as if we try and count by sliding the complete tree to all nodes in turn to find a match). Matching complete subtrees can possibly overlap. If T consists of a single node and a right child, then X_n counts the number of nodes in the tree with just one right child.

Example 4: Imabalance parameters. If we set $f(S(u))$ equal to 1 if and only if the sizes of the left and right subtrees of u are equal, then X_n counts the number of nodes at which we achieve a complete balance.

To study X_n, we first derive the mean and variance. This is followed by a weak law of large numbers for X_n/n. Several interesting examples illustrate this universal law. A general central limit theorem with normal limit is obtained for X_n using Stein's method. Several specific laws are obtained for particular choices of f. For example, for toll functions g as in Example 1, with $g(n)$ growing at a rate inferior to $\sqrt{n}$, a universal central limit theorem is established in Theorem 3.

§9. Representation (4) again

We replace the sum over all nodes u in a random tree in the definition of X_n by a sum over a deterministic set of index pairs, thereby greatly facilitating systematic analysis. We denote by $\sigma(i,k)$ the subset $(i,U_i),\ldots,(i+k-1,U_{i+k-1})$, so that $|\sigma(i,k)| = k$. We define $\sigma^*(i,k) = \sigma(i-1,k+1)$, with the convention that $(0,U_0) = (0,0)$ and $(n+1,U_{n+1}) = (n+1,0)$. Define the event

$$A_{i,k} = [\sigma(i,k) \text{ defines a subtree}] \ .$$

This event depends only on $\sigma^*(i,k)$, as $A_{i,k}$ happens if and only if among $U_{i-1},\ldots,U_{i+k}$, U_{i-1} and U_{i+k} are the two smallest values. We will call these the "guards": they cut off and protect the subtree from the rest of the tree. We set $Y_{i,k} = \mathbb{1}_{A_{i,k}}$. Note that if we include 0 and $n+1$ in the summation, then $\sum_{i<k} Y_{i,k} = n$, as we have exactly n subtrees. Rewrite our tree parameter as follows:

$$X_n = \sum_u f(S(u)) = \sum_{i=1}^{n} \sum_{k=1}^{n-i+1} Y_{i,k} f(\sigma(i,k)) \ .$$

For example, in the example with toll function g, this yields

$$X_n = \sum_u g(|S(u)|) = \sum_{i=1}^{n} \sum_{k=1}^{n-i+1} Y_{i,k} g(k) \ .$$

The interest in this formula is that the $Y_{i,k}$ are Bernoulli random variables with known mean:

$$\mathbb{E}\{Y_{i,k}\} = \begin{cases} 1 & \text{if } i = 1 \text{ and } i + k = n + 1; \\ \frac{1}{k+1} & \text{if } i = 1 \text{ or } i + k = n + 1 \text{ but not both;} \\ \frac{2}{(k+2)(k+1)} & \text{otherwise.} \end{cases}$$

There are about n^2 of them, and they are dependent, but at least, all the randomness is now tightly controlled.

§10. Mean and variance for toll functions

Let σ be a uniform random permutation of size k. Then define

$$\mu_k = \mathbb{E}\{f(\sigma)\} \ ,$$

$$\tau_k^2 = \mathbb{E}\{f^2(\sigma)\} \ ,$$

and

$$M_k = \sup_{\sigma : |\sigma| = k} |f(\sigma)| \ .$$

Note that $|\mu_k| \le \tau_k \le M_k$. In the toll function example, we have $\mu_k = g(k)$ and $\tau_k = M_k = |g(k)|$. We opt to develop the theory below in terms of these parameters.

Lemma 5. *Assume $|\mu_k| < \infty$ for all k, $\mu_k = o(k)$, and*

$$\sum_{k=1}^{\infty} \frac{|\mu_k|}{k^2} < \infty \ .$$

Define

$$\mu = \sum_{k=1}^{\infty} \frac{2\mu_k}{(k+2)(k+1)} \ .$$

Then

$$\lim_{n \to \infty} \frac{\mathbb{E}\{X_n\}}{n} = \mu \ .$$

If also $|\mu_k| = O(\sqrt{k}/\log k)$, then $\mathbb{E}\{X_n\} - \mu n = o(\sqrt{n})$.

PROOF. We have

$$\begin{aligned}
\mathbb{E}\{X_n\} &= \sum_{i=1}^{n}\sum_{k=1}^{n-i+1}\mathbb{E}\{Y_{i,k}\}\mu_k \\
&= \sum_{i=2}^{n}\sum_{k=1}^{n-i}\frac{2}{(k+2)(k+1)}\mu_k + \sum_{k=1}^{n-1}\frac{1}{k+1}\mu_k + \sum_{i=1}^{n}\frac{1}{n-i+2}\mu_{n-i+1} + \mu_n .
\end{aligned}$$

It is trivial to conclude the first part of Lemma 5. For the last part, we have:

$$\begin{aligned}
&|\mathbb{E}\{X_n - \mu n\}| \\
&\le \sum_{k=1}^{\infty}\frac{2}{(k+2)(k+1)}|\mu_k| + \sum_{i=2}^{n}\sum_{k=n-i+1}^{\infty}\frac{2}{(k+2)(k+1)}|\mu_k| + 2\sum_{k=1}^{n}\frac{|\mu_k|}{k+1} + |\mu_n| \\
&\le O(1) + \sum_{k=1}^{\infty}\frac{2\min(k,n)|\mu_k|}{(k+2)(k+1)} + 2\sum_{k=1}^{n}\frac{|\mu_k|}{k+1} + |\mu_n| \\
&\le O(1) + 4\sum_{k=1}^{n}\frac{|\mu_k|}{k+1} + n\sum_{k=n+1}^{\infty}\frac{|\mu_k|}{(k+2)(k+1)} + |\mu_n| \; . \square
\end{aligned}$$

The expression for μ is quite natural, as the probability of guards at positions $i-1$ and $i+k$ is $\mathbb{E}\{Y_{i,k}\} = 2/((k+2)(k+1))$. The following technical Lemma is due to Devroye (2002).

LEMMA 6. *Assume that $M_n < \infty$ for all n and that $f \ge 0$. Assume that for some $b \ge c \ge a > 0$, we have $\mu_n = O(n^a)$, $\tau_n = O(n^c)$, and $M_n = O(n^b)$. If $a+b < 2$, $c < 1$, then $\mathbb{V}\{X_n\} = o(n^2)$. If $a+b < 1$, $c < 1/2$, then $\mathbb{V}\{X_n\} = O(n)$. If f is a toll function and $M_n = O(n^b)$, then $\mathbb{V}\{X_n\} = o(n^2)$ if $b < 1$ and $\mathbb{V}\{X_n\} = O(n)$ if $b < 1/2$.*

PROOF. Let $Z_\alpha, \alpha \in A$, be a finite collection of random variables with finite second moments. Let E denote the collection of all pairs (α, β) from A^2 with $\alpha \ne \beta$ and Z_α not independent of Z_β. If $S = \sum_{\alpha \in A} Z_\alpha$, then

$$\mathbb{V}\{S\} = \sum_{\alpha\in A}\mathbb{V}\{Z_\alpha\} + \sum_{(\alpha,\beta)\in E}\left(\mathbb{E}\{Z_\alpha Z_\beta\} - \mathbb{E}\{Z_\alpha\}\mathbb{E}\{Z_\beta\}\right) .$$

We apply this fact with A being the collection of all pairs (i,k), with $1 \le i \le n$ and $1 \le k \le n-i+1$. Let our collection of random variables be the products $Y_{i,k} f(\sigma(i,k))$, $(i,k) \in V$. Note that E consists only of pairs $((i,k),(j,\ell))$ from A^2 with $i+k \ge j-1$ and $j+\ell \ge i$. This means that the intervals $[i, i+k-1]$ and $[j, j+\ell-1]$ correspond to an element of E if and only if they overlap or are disjoint and separated by exactly zero or one integer m. But to bound $\mathbb{V}\{X_n\}$ from above, since $f \geq 0$, we have

$$\mathbb{V}\{X_n\} \leq \sum_{(i,k)\in A} \mathbb{V}\{Y_{i,k} f(\sigma(i,k))\} + \sum_{((i,k),(j,\ell))\in E} \mathbb{E}\{Y_{i,k} f(\sigma(i,k)) Y_{j,\ell} f(\sigma(j,\ell))\} = I + II \ .$$

By the independence of $Y_{i,k}$ and $f(\sigma(i,k))$, we have

$$\begin{aligned} \mathbb{V}\{Y_{i,k} f(\sigma(i,k))\} &= \mathbb{V}\{Y_{i,k}\}\mathbb{E}\{f^2(\sigma(i,k))\} + (\mathbb{E}\{Y_{i,k}\})^2 \mathbb{V}\{f(\sigma(i,k))\} \\ &= \mathbb{V}\{Y_{i,k}\}\tau_k^2 + (\mathbb{E}\{Y_{i,k}\})^2(\tau_k^2 - \mu_k^2) \\ &\leq \mathbb{E}\{Y_{i,k}\}\tau_k^2 \end{aligned}$$

and thus $I = O(n)$ if $\tau_n^2 = O(n)$, $\sum_{k=1}^n \tau_k^2/k = O(n)$ and $\sum_k \tau_k^2/k^2 < \infty$. These conditions hold if $c < 1/2$. We have $I = o(n^2)$ if $c < 1$.

In II, we have $Y_{i,k}Y_{j,\ell} = 0$ unless the intervals $[i, i+k-1]$ and $[j, j+\ell-1]$ are disjoint and precisely one integer apart, or nested. For disjoint intervals, we note the independence of $Y_{i,k}Y_{j,\ell}$, $f(\sigma(i,k))$ and $f(\sigma(j,\ell))$, so that

$$\mathbb{E}\{Y_{i,k} f(\sigma(i,k)) Y_{j,\ell} f(\sigma(j,\ell))\} = \mathbb{E}\{Y_{i,k}Y_{j,\ell}\}\mu_k\mu_\ell \ .$$

If none of the intervals contains 1 or n, then a brief argument shows that

$$\mathbb{E}\{Y_{i,k}Y_{j,\ell}\} \leq \frac{4}{(k+\ell+3)(k+1)(\ell+1)} \ .$$

If one interval covers 1 and the other n, then $k+\ell = n-1$, and $\mathbb{E}\{Y_{i,k}Y_{j,\ell}\} = 1/n$. In the other cases, the expected value is bounded by $2/(k+\ell+2)(k+1)$ or $2/(k+\ell+2)(\ell+1)$, depending upon which interval covers 1 or n. Thus, the sum in II limited to disjoint intervals is bounded by

$$\begin{aligned} & n\sum_{k=1}^n \sum_{\ell=1}^n \frac{4\mu_k\mu_\ell}{(k+\ell+3)(k+1)(\ell+1)} + 1 + \sum_{k=1}^n \sum_{\ell=1}^n \frac{4\mu_k\mu_\ell}{(k+\ell+2)(k+1)} \\ & \leq 2n\sum_{k=1}^n \sum_{\ell=1}^k \frac{4\mu_k\mu_\ell}{(k+3)(k+1)(\ell+1)} + 1 + 2\sum_{k=1}^n \sum_{\ell=1}^k \frac{4\mu_k\mu_\ell}{(k+2)(k+1)} \ . \end{aligned}$$

If $\mu_n = O(n^a)$ for $a > 0$, then it is easy to see that the three sums taken together are $O(n^{2a})$.

We next consider nested intervals. For properly nested intervals, with $[i, i+k-1]$ being the bigger one, we have

$$\begin{aligned} \mathbb{E}\{Y_{i,k} f(\sigma(i,k)) Y_{j,\ell} f(\sigma(j,\ell))\} &= \mathbb{E}\{Y_{i,k}\}\mathbb{E}\{f(\sigma(i,k)) Y_{j,\ell} f(\sigma(j,\ell))\} \\ &\leq \frac{2M_k\mathbb{E}\{Y_{i,k}\}\mu_\ell}{(\ell+2)(\ell+1)} \ . \end{aligned}$$

Summed over all allowable pairs $(i,k),(j,\ell)$ with the outer interval not covering 1 or n, and noting that in all cases considered, $a<1$, this yields a quantity not exceeding

$$n\sum_{k=1}^{n} k \sum_{\ell=1}^{k} \frac{4M_k\mu_\ell}{(\ell+2)(\ell+1)(k+2)(k+1)} \le n\sum_{k=1}^{n} M_k O(k^{a-2}) = \begin{cases} O(n^{b+a}) & \text{if } b+a\neq 1\ , \\ O(n\log n) & \text{if } b+a=1. \end{cases}$$

The contribution of the border effect is of the same order. This is $o(n^2)$ if $a+b<2$. It is $O(n)$ if $a+b\le 1$.

Finally, we consider nested intervals with $i=j$ and $\ell<k$. Then

$$\mathbb{E}\{Y_{i,k}f(\sigma(i,k))Y_{j,\ell}f(\sigma(j,\ell))\} \le \mathbb{E}\{Y_{i,k}\}M_k\frac{\mu_\ell}{\ell+1}\ .$$

Summed over all appropriate (i,k,ℓ) such that the outer interval does not cover 1 or n, we obtain a bound of

$$n\sum_{k=1}^{n}\sum_{\ell=1}^{k}\frac{2M_k\mu_\ell}{(k+2)(k+1)(\ell+1)} = O(n^{a+b}+\mathbb{1}_{[a+b=1]}n\log n)\ .$$

The border cases do not alter this bound. Thus, the contribution to II for these nested intervals is $o(n^2)$ if $a+b<2$ and is $O(n)$ if $a+b<1$. □

§11. A law of large numbers

The estimates of the previous section permit us to obtain a law of large numbers for X_n, which was defined at the top of section 8.

Theorem 3. *Assume that $M_n<\infty$ for all n and that $f\ge 0$. Assume that for some $b\ge c\ge a>0$, we have $\mu_n=O(n^a)$, $\tau_n=O(n^c)$, and $M_n=O(n^b)$. If $a+b<2$, $c<1$, then*

$$\frac{X_n}{n}\to\mu$$

in probability. If f is a toll function and $M_n=O(n^b)$, then $X_n/n\to\mu$ in probability when $b<1$.

Proof. Note that $a < 1$. By Lemma 1, we have $\mathbb{E}\{X_n\}/n \to \mu$. Choose $\epsilon > 0$. By Chebyshev's inequality and Lemma 1,

$$\mathbb{P}\{|X_n - \mathbb{E}\{X_n\}| > \epsilon n\} \leq \frac{\mathbb{V}\{X_n\}}{\epsilon^2 n^2} = o(1) \ .$$

Thus, $X_n/n - \mathbb{E}\{X_n\}/n \to 0$ in probability. □

The result does not apply to the number of comparisons in quicksort as $a = b = 1$, a border case just outside the conditions. In fact, for quicksort, we have $X_n/(2n \log n) \to 1$ in probability. For nearly all "smaller" toll functions and tree parameters, Theorem 3 is applicable. Three examples will illustrate this.

Example 1. We let f be the indicator function of anything, and note that the law of large numbers holds. For example, let $\mathcal{T}$ be a possibly infinite collection of possible tree patterns, and let X_n count the number of subtrees in a random binary search tree that match a tree from $\mathcal{T}$. Then, as shown below, the law of large numbers holds. There is no inherent limitation to $\mathcal{T}$, which, in fact, might be the collection of all trees whose size is a perfect square and whose height is a prime number at the same time. Let X_n be the number of subtrees in a random binary search tree that match a given prefix tree pattern T, with $|T| = k$ fixed.

Theorem 4. *For any non-empty tree pattern collection $\mathcal{T}$, we have*

$$\frac{X_n}{n} \to \mu$$

in probability, and $\mathbb{E}\{X_n\}/n \to \mu$, where

$$\mu = \sum_{n=1}^{\infty} \frac{2\mu_n}{(n+2)(n+1)}$$

and μ_n is the probability that a random binary search tree of size n matches an element of $\mathcal{T}$.

Proof. Theorem 3 applies since f is an indicator function. By Lemma 5, we obtain the limit μ for $\mathbb{E}\{X_n\}/n$. □

Note that Theorem 4 remains valid if we replace the phrase "matches an element of $\mathcal{T}$" by the phrase "matches an element of $\mathcal{T}$ at its root", so that $\mathcal{T}$ is a collection of what we called earlier prefix tree patterns.

Example 2. Perhaps more instructive is the example of the sumheight $\mathcal{S}_n$, the sum of the heights of all subtrees in a random binary search tree on n nodes.

Theorem 5. *For a random binary search tree, the sumheight satisfies*

$$\frac{\mathcal{S}_n}{\mathbb{E}\{\mathcal{S}_n\}} \to 1$$

in probability. Here

$$\mathbb{E}\{\mathcal{S}_n\} \sim n \sum_{k=1}^{\infty} \frac{2h_k}{(k+2)(k+1)} ,$$

where h_k is the expected height of a random binary search tree on k nodes.

Proof. The statement about the expected height follows from Lemma 5 without work. As the height of a subtree of size k is at most $k-1$, we see that we may apply Theorem 3 with $M_k = k-1$. By well-known results (Robson, 1977; Pittel, 1984; Devroye, 1986, 1987), we have $\mathbb{E}\{H_n^2\} = O(\log^2 n)$ where H_n is the height of a random binary search tree. Thus, we may formally take a and c arbitrarily small but positive, and $b = 1$. □

Example 3. Consider $X_n = \sum_u (N(u))^\beta$, $0 < \beta < 1$. Recall that $\sum_u N(u)$ is the number of comparisons in quicksort, plus n. Thus, X_n is a discounted parameter with respect to the number of quicksort comparisons. Clearly, Theorem 3 applies with $a = b = c = \beta$, and thus, $X_n/n \to \mu$ in probability, and $\mathbb{E}\{X_n\}/n$ tends to the same constant μ. In a sense, this application is near the limit of the range for Theorem 3. For example, it is known that with $X_n = \sum_u (N(u))^{1+\epsilon}$, $\epsilon > 0$, there is no asymptotic concentration, and thus, $X_n/g(n)$ does not converge to a constant for any choice of $g(n)$. Also, for $X_n = \sum_u N(u)$, the quicksort example, we have $X_n/2n\log n \to 1$ in probability (Sedgewick, 1983), so that once again Theorem 3 is not applicable.

§12. Dependence graph

We will require the notion of a dependency graph for a collection of random variables $(Z_\alpha)_{\alpha \in V}$, where V is a set of vertices. Let the edge set E be such that for all disjoint subsets A and B of V that are not connected by an edge, $(Z_\alpha)_{\alpha \in A}$ and $(Z_\alpha)_{\alpha \in B}$ are mutually independent.

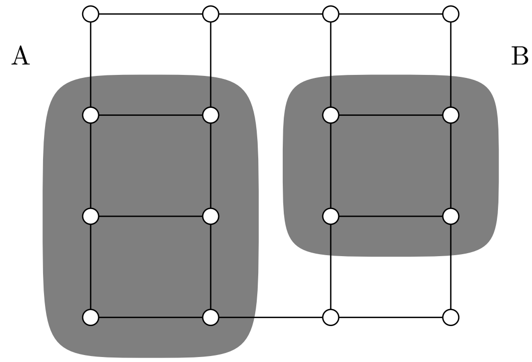

Figure 7. The dependence graph is such that any disjoint subsets that have no edges connecting them correspond to collections of random variables that are independent.

Clearly, the complete graph is a dependency graph for any set of random variables, but this is useless. One usually takes the minimal graph (V, E) that has the above property, or one tries to keep $|E|$ as small as possible. Note that necessarily, Z_α and Z_β are independent if $(\alpha, \beta) \notin E$, but to have a dependency graph requires much more than just checking pairwise independence. We call the neighborhood of $N(\alpha)$ of vertex $\alpha \in V$ the collection of vertices β such that $(\alpha, \beta) \in E$ or $\alpha = \beta$. We define the neighborhood $N(\alpha_1, \ldots, \alpha_r)$ as $\cup_{j=1}^r N(\alpha_j)$.

A. Consider now for V the pairs (i, k) with $1 \le i \le n$ and $1 \le k \le n - i + 1$. Let our collection of random variables be the permutations $\sigma(i, k)$, $(i, k) \in V$. Let us connect (i, k) to (j, ℓ) when $i + k \ge j - 1$ and $j + \ell \ge i$. This means that the intervals $(i, i + k - 1)$ and $(j, j + \ell - 1)$ correspond to an edge in E if and only if they overlap or are disjoint and separated by exactly zero or one integer m. We claim that (V, E) is a dependency graph. Indeed, if we consider disjoint subsets A and B of vertices with no edges between them, then these vertices correspond to intervals that are pairwise separated by at least two integers, and thus, $(\sigma(i, k))_{(i,k)\in A}$ and $(\sigma(j, \ell))_{(j,\ell)\in B}$ are mutually independent.

B. Consider next the collection of random variables $Y_{i,k} g(k)$. For this collection, we can make a smaller dependency graph. Eliminate all edges from the graph of the previous paragraph if the intervals defined by the endpoints of the edges are properly nested. For example, if $i < j < j + \ell - 1 < i + k$, then the edge between (i, k) and (j, ℓ) is removed. The graph thus obtained is still a dependency graph. This observation repeatedly uses the fact that if one considers a sequence $Z_1, \ldots, Z_n$ of i.i.d. random variables with a uniform $[0, 1]$ distribution, then given that Z_1 and Z_n are the two largest values, then $Z_2, \ldots, Z_{n-1}$ are i.i.d. and uniform on $[0, \min(Z_1, Z_n)]$. the permutation of $Z_2, \ldots, Z_{n-1}$

are all independent. Thus, for properly nested intervals as above, $Y_{i,k}g(k)$ is independent of $Y_{j,\ell}g(\ell)$.

§13. Stein's method

Stein's method (Stein, 1972) allows one to deduce a normal limit law for certain sums of random variables while only computing first and second order moments and verifying a certain dependence condition. Many variants have seen the light of day in recent years, and we will simply employ the following version derived in Janson, Łuczak and Ruciński (2000, Theorem 6.21):

LEMMA 7. *Suppose that $(S_n)_1^\infty$ is a sequence of random variables such that $S_n = \sum_{\alpha \in V_n} Z_{n\alpha}$, where for each n, $\{Z_{n\alpha}\}_\alpha$ is a family of random variables with dependency graph (V_n, E_n). Let $N(.)$ denote the neighborhood of a vertex or vertices. Suppose further that there exist numbers M_n and Q_n such that*

$$\sum_{\alpha \in V_n} \mathbb{E}\{|Z_{n\alpha}|\} \leq M_n,$$

and for every n and $r \geq 1$,

$$\sup_{\alpha_1, \alpha_2, \ldots, \alpha_r \in V_n} \sum_{\beta \in N(\alpha_1, \alpha_2, \ldots, \alpha_r)} \mathbb{E}\{|Z_{n\beta}| | Z_{n\alpha_1}, Z_{n\alpha_2}, \ldots, Z_{n\alpha_r}\} \leq B_r Q_n$$

for some number B_r depending upon r only. Let $\sigma_n^2 = \mathbb{V}\{S_n\}$. Then

$$\frac{S_n - \mathbb{E}\{S_n\}}{\sqrt{\mathbb{V}\{S_n\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

if for some real $s > 2$,

$$\lim_{n \to \infty} \frac{M_n Q_n^{s-1}}{\sigma_n^s} = 0 \ .$$

Lemma 8 (devroye, 2002). *Define* $X_n = \sum_u g(|S(u)|)$ *for a random binary search tree on* n *nodes. The following statements are equivalent:*

A. $\mathbb{V}\{X_n\} = \Omega(n)$, *i.e., there are positive numbers* a, N, *such at* $\mathbb{V}\{X_n\} \geq an$ *for all* $n \geq N$.

B. $\mathbb{V}\{X_k\} > 0$ *for some* $k > 0$.

C. The function g *is not constant on* $\{1, 2, \ldots\}$.

To study

$$X_n = \sum_u g(|S(u)|)$$

we define $G(n) = \max_{1 \leq i \leq n} |g(i)|$. The next theorem imporoves on the result of Devroye (2002).

Theorem 6. *Assume that* g *is not constant on* $\{1, 2, \ldots\}$. *If* $G(n) = O(n^{1/2-\epsilon})$ *for some* $\epsilon > 0$, *then*

$$\frac{X_n - \mathbb{E}\{X_n\}}{\sqrt{\mathbb{V}\{X_n\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \ .$$

Proof. We apply Lemma 7 to the basic collection of random variables $Y_{i,k} g(k)$, $(i,k) \in V_n$, where V_n is the set $\{(i,k) : 1 \leq i \leq n, 1 \leq k \leq n-i+1\}$, and $Y_{i,k}$ are the indicator variables of section 9. Let E_n be the edges in the dependency graph L_n defined by connecting (i,k) to (j,ℓ) if the respective intervals are overlapping without being properly nested, or if the respective intervals are disjoint with zero or one integers separating them. We note that

$$\sum_{(i,k) \in V_n} \mathbb{E}\{Y_{i,k} |g(k)|\} \leq (\nu + o(1))n$$

by computations not unlike those for the mean done in Lemma 5, where

$$\nu = \sum_{n=1}^{\infty} \frac{2|g(n)|}{(n+2)(n+1)} \ .$$

To apply Lemma 7, we note that we may thus take $M_n = O(n)$. We also note that $\sigma_n^2 = \Omega(n)$, by Lemma 8, since g is not constant on $\{1, 2, \ldots\}$. Since we can choose s in Stein's theorem after having seen ϵ, it suffices thus to show that $Q_n = O(n^{1/2-\epsilon})$ for some $\epsilon > 0$. Note that conditioning on $Y_{i,k} g(k)$ is equivalent to conditioning on $Y_{i,k}$. For the bound on Q_n, we need to consider any finite collection of random variables $Y_{i,k}$. We will give the bound in the case of just two such random variables, and leave it to the reader to verify the easy extension to any finite number, noting that the quantity B_r in Lemma 7 can absorb the bigger proportional constant one would obtain then. We may bound Q_n by

$$Q_n \le G(n) \sup_{(i,k),(j,\ell)\in V_n} \sum_{(p,r)\in N((i,k),(j,\ell))} \mathbb{E}\{Y_{p,r}|Y_{i,k},Y_{j,\ell}\} \ .$$

We show that the sum above is uniformly bounded over all choices of $(i,k),(j,\ell)$ by $O(\log^2 n)$.

Consider the set $S=\{0,1,\ldots,n,n+1\}$ and mark $0,n+1,i-1,i+k,j-1,j+\ell$ (where duplications may occur). The last four marked points are neighbors of the intervals represented by (i,k) and (j,ℓ). Mark also all integers in S that are neighbors of these marked numbers. The total number of marked places does not exceed $3\times 4+2\times 2=16$. The set S, when traversed from small to large, can be described by consecutive intervals of marked and unmarked integers. The number of unmarked integer intervals is at most five. We call these intervals $H_1,\ldots,H_5$, from left to right, with some of these possibly empty.

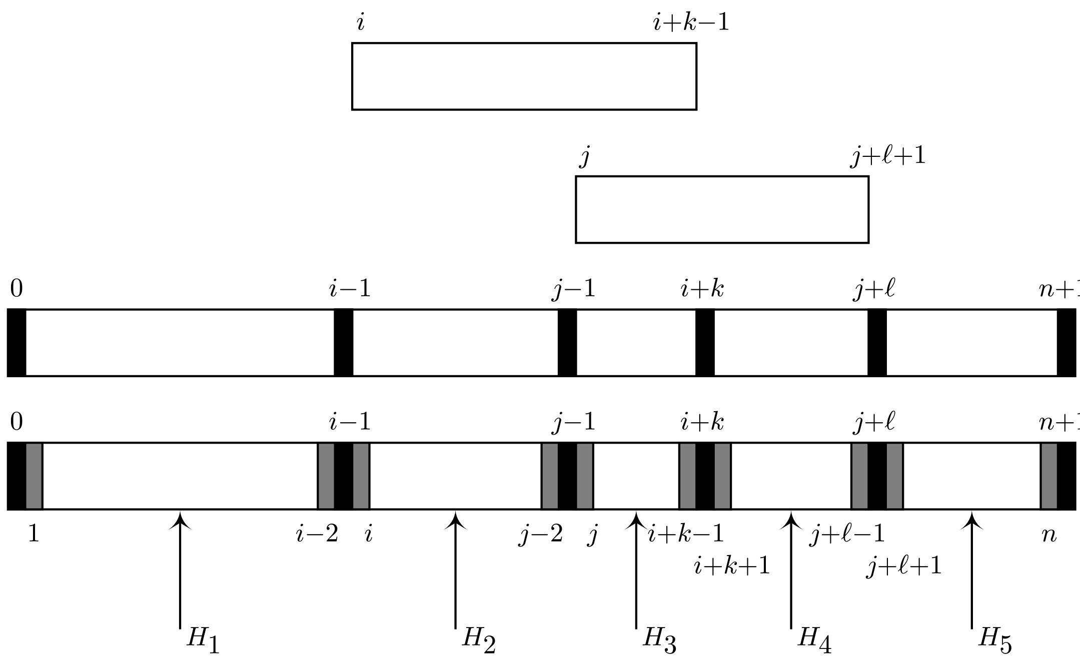

Figure 8. The marking of cells is illustrated. The top two lines show the original intervals. In the first round, we mark (in black) the immediate neighbors of these intervals, as well as cells 0 and $n+1$. In the second round, all neighbors (shown in gray) of already marked cells are also marked. The intervals of contiguous unmarked cells are called H_i. There are not more than five of these.

Set $H=\cup_i H_i$. Define $H^c=S-H$. Consider $Y_{p,r}$ for $r\le n$ fixed. Let $s=p+r-1$ be the endpoint of the interval on which $Y_{p,r}$ sits. Note that $Y_{p,r}$ depends upon $\{U_i\}_{p-1\le i\le s+1}$. We note four situations:

A. If $p, s \in H_i$ for a given i, then $Y_{p,r}$ is clearly independent of $Y_{i,k}, Y_{j,\ell}$. In fact, then, $(p,r) \notin N((i,k),(j,\ell))$.

B. If $p, s \in H^c$, then we bound $\mathbb{E}\{Y_{p,r}|Y_{i,k}, Y_{j,\ell}\}$ by one.

C. If p or s is in H^c and the other endpoint is in H_i, then we bound as follows:

$$\mathbb{E}\{Y_{p,r}|Y_{i,k}, Y_{j,\ell}\} \leq \frac{1}{1 + |H_i \cap \{p, \ldots, s\}|}$$

because we can only be sure about the i.i.d. nature of the U_i's in $H_i \cap \{p, \ldots, s\}$ together with the two immediate neighbors of this set.

D. If $p \in H_i$, $s \in H_j$, $i < j$, then we argue as in case C twice, and obtain the following bound:

$$\mathbb{E}\{Y_{p,r}|Y_{i,k}, Y_{j,\ell}\} \leq \frac{1}{1 + |H_i \cap \{p, \ldots, s\}|} \times \frac{1}{1 + |H_j \cap \{p, \ldots, s\}|} \ .$$

The above considerations permit us to obtain a bound for Q_n by summing over all $(p,r) \in N((i,k),(j,\ell))$. The sum for all cases (A) is zero. The sum for case (B) is at most $16^2 = 256$. The sum over all (p,r) as in (C) is at most

$$2 \times 16 \times 5 \times \sum_{r=1}^{n} \frac{1}{r+1} \leq 160 \log(n+1) \ .$$

Finally, the sum over all (p,r) described by (D) is at most

$$\binom{5}{2} \left(\sum_{r=1}^{n} \frac{1}{r+1} \right)^2 \leq 10 \log^2(n+1) \ .$$

The grand total is $O(\log^2 n)$, as required. This argument can be repeated for conditioning on finite sets larger than two, as required by Stein's theorem. We leave that as an exercise (see also Devroye, 2002). □

§14. Sums of indicator functions

In this section, we take a simple example, in which

$$X_n = \sum_u \mathbb{1}_{[S(u) \in A_n]} \ ,$$

where A_n is a non-empty collection of permutations of length k, with k possibly depending upon n. We denote $p_{n,k} = |A_n|/k!$, the probability that a randomly picked permutation of length k is in the collection A_n. Particular examples include sets A_n that correspond to a particular tree pattern, in which case X_n counts the number of occurrences of a given tree pattern of size k (a "terminal pattern") in a random binary search tree. The interest here is in the case of varying k. As we will see below, for a central limit law, k has to be severely restricted.

THEOREM 7. *We have*

$$\mathbb{E}\{X_n\} = \frac{2np_{n,k}}{(k+2)(k+1)} + O(1)$$

regardless of how k varies with n. If $k = o(\log n/\log\log n)$, then $\mathbb{E}\{X_n\} \to \infty$, $X_n/\mathbb{E}\{X_n\} \to 1$ in probability, and

$$\frac{X_n - \mathbb{E}\{X_n\}}{\sqrt{\mathbb{V}\{X_n\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1) \ .$$

PROOF. Observe that

$$X_n = \sum_{i=1}^{n-k+1} Y_{i,k} Z_i \ .$$

where $Z_i = \mathbb{1}_{[\sigma(i,k)\in A_n]}$, Thus,

$$\mathbb{E}\{X_n\} = \sum_{i=2}^{n-k} \frac{2}{(k+2)(k+1)} \mathbb{E}\{Z_1\} + 2 \times \frac{1}{k+1} \mathbb{E}\{Z_1\} = \frac{2(n-k-1)p_{n,k}}{(k+2)(k+1)} + \frac{2p_{n,k}}{k+1} \ .$$

This proves the first part of the theorem.

The computation of the variance is slightly more involved. However, it is simplified by considering the variance of

$$Y_n = \sum_{i=2}^{n-k} Y_{i,k} Z_i \ ,$$

and noting that $|X_n - Y_n| \leq 2$. This eliminates the border effect. We note that $Y_{i,k}Y_{j,k} = 0$ if $i < j \leq i+k$. Thus,

$$\begin{aligned} \mathbb{E}\{Y_n^2\} &= \sum_{i=2}^{n-k} \mathbb{E}\{Y_{i,k}Z_i\} + 2 \sum_{2\leq i<j\leq n-k} \mathbb{E}\{Y_{i,k}Z_iY_{j,k}Z_j\} \\ &= \mathbb{E}\{Y_n\} + 2 \sum_{2\leq i,i+k+1\leq n-k} \mathbb{E}\{Y_{i,k}Z_iY_{i+k+1,k}Z_{i+k+1}\} + 2 \sum_{2\leq i,i+k+1<j\leq n-k} \mathbb{E}\{Y_{i,k}Z_i\}\mathbb{E}\{Y_{j,k}Z_j\} \\ &= (n-k-1)\beta + 2(n-2k-2)\alpha + (n-2k)^2\beta^2 + (10k+6-5n)\beta^2 \end{aligned}$$

where $\alpha = \mathbb{E}\{Y_{2,k}Z_2Y_{3+k,k}Z_{3+k}\}$, and $\beta = \mathbb{E}\{Y_{2,k}Z_2\}$. Also,

$$(\mathbb{E}\{Y_n\})^2 = ((n-k-1)\beta)^2 \ .$$

Thus,

$$\begin{aligned} \mathbb{V}\{Y_n\} &= 2(n-2k-2)\alpha + (n-k-1)\beta + \left((n-2k)^2 - (n-k-1)^2 + (10k+6-5n)\right)\beta^2 \\ &= n\left(2\alpha + \beta - (2k+3)\beta^2\right) + O(k\alpha + k\beta + k^2\beta^2) \ . \end{aligned}$$

We note that $\beta = 2p_{n,k}/(k+2)(k+1)$. To compute α, let A, B, C be the minimal values among $U_1, \ldots, U_{k+1}$; U_{k+2}, and $U_{k+3}, \ldots, U_{2k+3}$, respectively. Clearly,

$$\alpha = p_{n,k}^2 \mathbb{E}\{Y_{2,k} Y_{3+k,k}\}.$$

Considering all six permutations of A, B, C separately, one may compute the latter expected value as

$$\frac{2}{2k+3}\frac{1}{2k+2}\frac{1}{k+1} + \frac{1}{2k+3}\frac{1}{k+1}\frac{1}{k+1} + \frac{2}{2k+3}\frac{1}{2k+2}\frac{1}{2k+1} = \frac{5k+3}{(2k+3)(2k+1)(k+1)^2} \ .$$

Thus,

$$\alpha = \frac{(5k+3)p_{n,k}^2}{(2k+3)(2k+1)(k+1)^2} \ .$$

We have

$$\mathbb{V}\{Y_n\} = \\ = n\left(p_{n,k}^2 \frac{10k+6}{(2k+3)(2k+1)(k+1)^2} - p_{n,k}^2 \frac{8k+12}{(k+2)^2(k+1)^2} + p_{n,k}\frac{2}{(k+2)(k+1)}\right) + O(p_{n,k}/k) \ .$$

Note that regardless of the value of $p_{n,k}$, the coefficient of n is strictly positive. Indeed, the coefficient is at least

$$p_{n,k}^2 \left(\frac{(10k+6)(k+2)^2 - (8k-8)(2k+3)(2k+1) + 2(2k+3)(2k+1)(k+2)(k+1)}{(k+2)^2(k+1)^2(2k+3)(2k+1)}\right) \\ = p_{n,k}^2 \left(\frac{8k^4 + 18k^3 + 4k^2 - 6k}{(k+2)^2(k+1)^2(2k+3)(2k+1)}\right) \ .$$

Thus, there exist universal constants $c_1, c_2, c_3 > 0$ such that

$$\mathbb{V}\{Y_n\} \geq c_1 n p_{n,k}^2/k^2 - c_2 p_{n,k}/k \ ,$$

and

$$\mathbb{V}\{Y_n\} \leq c_3 n p_{n,k}/k^2 \ .$$

We have $Y_n/\mathbb{E}\{Y_n\} \to 1$ in probability if $\mathbb{V}\{Y_n\} = o(\mathbb{E}^2\{Y_n\})$, that is, if $k = o(\sqrt{np_{n,k}})$. Using $p_{n,k} \geq 1/k!$, we note that this condition holds if $k = o(\log n/\log\log n)$.

Finally, we turn to the normal limit law and note that

$$\frac{Y_n - \mathbb{E}\{Y_n\}}{\sqrt{\mathbb{V}\{Y_n\}}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

if $k = o(\log n/\log\log n)$. For the proof, we refer to Lemma 7 and its notation. Clearly, $M_n = O(n)$. Furthermore, with r as in Lemma 7, we have $Q_n = O(rk)$. Finally, we recall the lower bound on the variance from above. Assume that $np_{n,k}/k \to \infty$. The condition of Lemma 7, applied with $s = 3$, holds if

$$\lim_{n\to\infty} \frac{nk^2}{n^{3/2}p_{n,k}^3/k^3} = 0 \ .$$

Since $p_{n,k} \geq 1/k!$, this in turn holds if

$$\lim_{n\to\infty} \frac{k^5 k!^3}{\sqrt{n}} = 0.$$

Both conditions are satisfied if $k = o(\log n / \log\log n)$. □

The limitation on k in Theorem 7 cannot be lifted without further conditions: indeed, if we consider as a tree pattern the tree that consists of a right branch of length k only, then $\mathbb{E}\{X_n\} \to 0$ if $k > (1+\epsilon)\log n/\log\log n$ for any given fixed $\epsilon > 0$. As X_n is integer-valued, no meaningful limit laws can exist in such cases.

§15. Refinements and bibliographic remarks

Central limit theorems for slightly dependent random variables have been obtained by Brown (1971), Dvoretzky (1972), McLeish (1974), Ibragimov (1975), Chen (1978), Hall and Heyde (1980) and Bradley (1981), to name just a few. Stein's method (our Lemma 7, essentially) is one of the central limit theorems that is better equipped to deal with cases of considerable dependence. It is noteworthy that Berry-Esseen inequalities exist for various models of local dependence such as k-dependence (Shergin (1979), Baldi and Rinott (1989), Rinott (1994), Dembo and Rinott (1996)). The paper of Chen and Shao (2004) covers all of these and provides Berry-Esseen inequalities in terms of a generalized notion of dependence graph.

Stein's method offers short and intuitive proofs, but other methods may offer attractive alternatives. Using the contraction method and the method of moments, Hwang and Neininger (2002) showed that the central limit result of Theorem 6 holds with $G(n) = O(n^a)$, $a \leq 1/2$. Our result is weaker, but requires fewer analytic computations. The ranges of application of the results are not nested—in some situations, Stein's method is more useful, while in others the moment and contraction methods are preferable. The previous three section are based on Devroye (2002).

Aldous (1991) showed that the number $V_{k,n}$ of subtrees of size precisely k in a random binary search tree is in probability asymptotic to $2/(k+2)(k+1)$. The limit law for V_{kn} (Theorem 2) is due to Devroye (1991). The latter result follows also from the previous section if we take as toll function $f(\sigma) = \mathbb{1}_{[|\sigma|=k]}$.

Recently, there has been some interest in the logarithmic toll function $f(\sigma) = \log|\sigma|$ (Grabner and Prodinger, 2002) and the harmonic toll function $f(\sigma) = \sum_{i=1}^{|\sigma|} 1/i$ (Panholzer and Prodinger, 2002). The authors in these papers are mainly concerned with precise first and second moment asymptotics. Fill (1996) obtained the central limit theorem for the case $f(\sigma) = \log|\sigma|$.

Clearly, these examples fall entirely within the conditions of Lemma 7 or Theorem 6, with some room to spare.

Flajolet, Gourdon and Martinez (1997) obtained a normal limit law for the number of subtrees in a random binary search tree with fixed finite tree pattern. Clearly, this is a case in which (the indicator function) $f(\sigma)$ depends on σ in an intricate way, but $f = 0$ unless $|\sigma|$ equals the size of the tree pattern. The situation is covered by the law of large numbers of Theorem 4 and the central limit result of Theorem 7. Theorem 7 even allows tree patterns that change with n.

§16. The toll function ladder

The various types of limit behavior can best be illustrated by considering the toll function $g(n) = n^\alpha$. Define the random variable

$$X_n = \sum_u N(u)^\alpha$$

where $N(u)$ is the size of the subtree rooted at node u. For $\alpha = 1$, we have $X_n = n + L_n$, as can easily be verified. We can use (yet another!) representation of a random binary search tree in terms of the sizes of left and right subtrees of the root, $\lfloor nU \rfloor$, and $\lfloor n(1-U) \rfloor$, respectively, where U is uniform $[0,1]$. A different independent uniform random variable is thus associated with each node. For a first intuition, we can drop the truncation function.

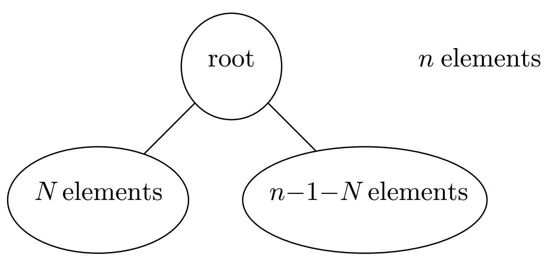

Figure 9. The root splits a random binary search tree into two subsets, of sizes N and $n-1-N$ respectively, where N is distributed as $\lfloor nU \rfloor$, and U is uniform $[0,1]$.

Call the (random) binary search tree T. Augment the tree T by associating with each node the size of the subtree rooted at that node, and call the augmented tree T'. The root of T' has value n. Since the rank of the root element of T is equally likely to be $1, \ldots, n$, the number N of nodes in the left subtree of the root of T is uniformly distributed on $\{0, 1, \ldots, n-1\}$. A moment's thought shows that

$$N \stackrel{\mathcal{L}}{=} \lfloor nU \rfloor \ ,$$

where U is uniformly distributed on $[0,1]$. Also, the size of the right subtree of the root of T is $n-1-N$, which is distributed as $\lfloor n(1-U) \rfloor$. All subsequent splits can be represented similarly by introducing independent uniform $[0,1]$ random variables. Once again, we have used an embedding argument: we have identified a new fictitious collection of random variables $U_1, U_2, \ldots$, and we can derive all the values of nodes in T' from it. This in turn determines T. However, the tree T obtained in this manner is only distributed as the T we are studying—it may be very different from the (random) instance of T presented to us. The rule is simply this: in an infinite binary tree, give the root the value n. Also, associate with each node an independent copy of U. If a node has value V, and its assigned copy of U is U' (say), then the value of the two children of the node are $\lfloor VU' \rfloor$ and $\lfloor V(1-U') \rfloor$ respectively. Thus, the value of any node at distance k from the root of T' is distributed as

$$\lfloor \cdots \lfloor \lfloor nU_1 \rfloor U_2 \rfloor \cdots U_k \rfloor ,$$

where $U_1, \ldots, U_k$ are i.i.d. uniform $[0,1]$.

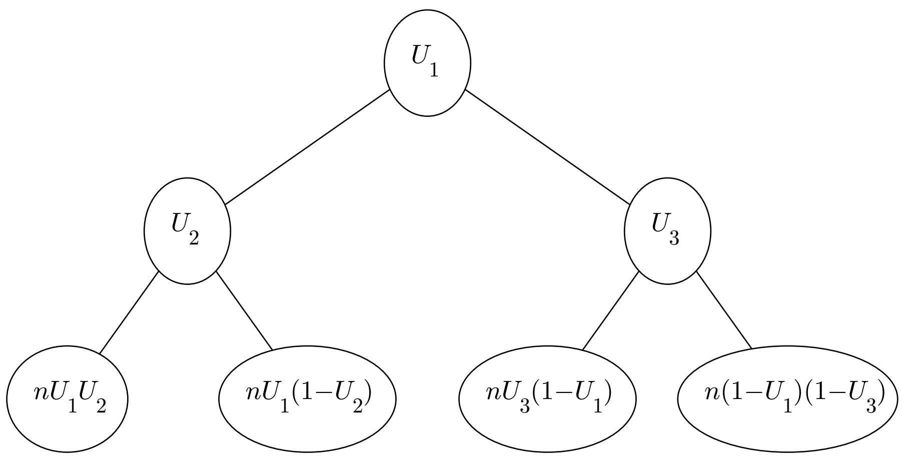

Figure 10. The augmented tree T': the values associated with the nodes are the sizes of the subtrees of corresponding nodes in T. Only the values of the nodes two levels away from the root are shown (modulo some floor functions). The random splits in the tree are controlled by independent uniform $[0,1]$ random variables $U_1, U_2, \ldots$.

For $\alpha > 1$, the total contribution to X_n from the root level is n^α. From the next level, we collect $n^\alpha(U^\alpha + (1-U)^\alpha)$, from the next level $n^\alpha(U^\alpha(U_1^\alpha + (1-U_1)^\alpha) + (1-U)^\alpha(U_2^\alpha + (1-U_2)^\alpha))$, and so forth, where all the U_i's are independent uniform $[0,1]$ random variables. Taking expectations, it is easy to see that the expected contribution from the i-th level is about

$$n^\alpha \left(\frac{2}{\alpha+1} \right)^i .$$

This decreases geometrically with i, the distance from the root. Therefore, the top few levels, which are all full, are of primary importance. It is rather easy to show that

$$\frac{X_n}{n^\alpha} \xrightarrow{\mathcal{L}} Y$$

where Y may be described either as

$$1 + U_1^\alpha + (1-U_1)^\alpha + U_1^\alpha(U_2^\alpha + (1-U_2)^\alpha) + (1-U_1)^\alpha(U_3^\alpha + (1-U_3)^\alpha) + \cdots,$$

or as the unique solution of the distributional identity

$$Y \stackrel{\mathcal{L}}{=} 1 + U^\alpha Y + (1-U)^\alpha Y'$$

where Y, Y', U are independent and $Y \stackrel{\mathcal{L}}{=} Y'$. Taking expectations in this recurrence, we see that $\mathbb{E}\{Y\} = (\alpha+1)/(\alpha-1)$, and that this tends to ∞ as $\alpha \downarrow 1$.

At $\alpha = 1$, we have a transition to the next behavior, because $X_n/n \to \infty$ in probability. The contributions from the first few full levels is n, so that we expect X_n to be of the order of $n \log n$. In fact, as we know, $X_n/(n \log n) \to 2$ in probability. To discover a limit law, we have to subtract the mean, and, as we now know,

$$\frac{X_n - \mathbb{E}\{X_n\}}{n} \xrightarrow{\mathcal{L}} Y$$

where Y has the quicksort limit distribution. See the next section for a description and derivation.

For $\alpha < 1/2$, we have seen (Theorem 6) that

$$\frac{X_n - \mathbb{E}\{X_n\}}{\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0, \sigma^2)$$

for a certain $\sigma^2 > 0$ depending upon α. The main contributions to X_n all come from the bottom levels, and normality is obtained because they are slightly dependent.

In the range $1/2 < \alpha < 1$, we have an intermediate behavior, with

$$\frac{X_n - \mathbb{E}\{X_n\}}{n^\alpha} \xrightarrow{\mathcal{L}} Y,$$

where Y satisfies the distributional identity given above. The details are given by Hwang and Neininger (2002).

§17. The contraction method

In some cases, a random variable X_n can be represented in a recursive manner as a function of similarly distributed random variables of smaller parameter. A limit law for X_n, properly normalized, can sometimes be obtained by the contraction method. This requires several steps. First, we need to settle on a metric on the space of distributions of interest. Then, we may attempt to normalize X_n in such a way that the limit distribution can be guessed as a stable point of a distributional identity (which one gets from the recursive description of X_n). Then one establishes that the recursion yields a contraction for the given metric, that is, a fractional reduction in the distance between two distributions. This implies in many cases that the stable point is unique. Finally, the most work is usually related to the proof that X_n or its normalized version tends to that unique limit law in the given metric.

We will illustrate this methodology on the number of comparisons in quicksort. For other examples, we refer to Rösler and Rüschendorf (2001) or Neininger and Rüschendorf (2004).

Let L_n denote the internal path length of a random binary search tree T_n. We recall that this is equal to the number of comparisons needed to sort a random permutation of $\{1, 2, \ldots, n\}$ by QUICKSORT. We write $\stackrel{\mathcal{L}}{=}$ to denote equality in distribution. The root recursion of T_n implies the following:

$$L_n \stackrel{\mathcal{L}}{=} L_{Z_n-1} + L'_{n-Z_n} + n - 1, n \geq 2,$$

where Z_n is uniformly distributed on $\{1, 2, \ldots, n\}$, $L_i \stackrel{\mathcal{L}}{=} L'_i$, and $Z_n, L_i, L'_i, 1 \leq i \leq n$, are independent. Furthermore, $L_0 = L_1 = 0$. We know that

$$\mathbb{E}\{L_n\} \sim 2n \log n$$

from earlier remarks, so some normalization is necessary to establish a limit law. We define

$$Y_n = \frac{L_n - \mathbb{E}\{L_n\}}{n}.$$

Régnier (1989) was the first one to prove by a martingale argument that $Y_n \stackrel{\mathcal{L}}{\to} Y$, where Y is a non-degenerate non-normal random variable, and $\stackrel{\mathcal{L}}{\to}$ denotes convergence in distribution. Rösler (1991, 1992) characterized the law of Y by a fixed-point equation: Y is the unique solution of the distributional identity

$$Y \stackrel{\mathcal{L}}{=} UY + (1-U)Y' + \varphi(U),$$

where U is uniform on $[0, 1]$, $Y \stackrel{\mathcal{L}}{=} Y'$, and U, Y, Y' are independent, and

$$\varphi(u) = 2u \log u + 2(1-u) \log(1-u) + 1.$$

The limit law has no simple explicit form, but we know that it has a density f (Tan and Hadjicostas, 1995), which is in C^∞ (Fill and Janson, 2000), and additional properties of that density are now well understood (Fill and Janson, 2001, 2002). We will give Rösler's proof, which has been used in many other applications, and is now referred to as the contraction method.

Step 1: a space and a metric. Let $\mathcal{D}$ denote the space of distribution functions with zero first moment and finite second moment. Then the Wasserstein metric d_2 is defined by

$$d_2(F,G) = \inf \sqrt{\mathbb{E}\{(X-Y)^2\}},$$

where the infimum is taken over all pairs (X,Y), X has distribution function F and Y has distribution function G. This metric thus uses the idea of a maximal coupling between X and Y. If F^{-1} denotes the inverse of F, then the maximal coupling for which the infimum in $d_2(F,G)$ is reached is given by setting

$$(X,Y) = \left(F^{-1}(U), G^{-1}(U)\right)$$

where U is uniform $[0,1]$. Using the same U, we can in thisd manner couple infinite families of random variables. We will call this a universal coupling of a collection of random variables.

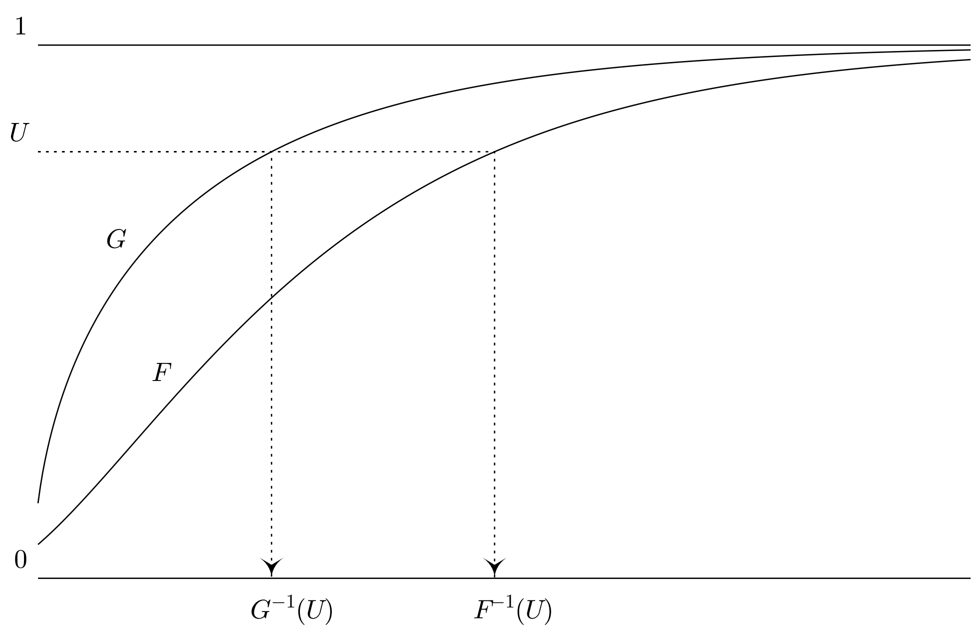

Figure 11. Illustration of optimal coupling for the Wasserstein metric.

If $d_2(F,G) = 0$, then there exists a coupling such that $X = Y$ with probability one, and thus, $F = G$. It is known that $(\mathcal{D}, d_2)$ is a complete metric space. Observe furthermore that a sequence F_n converges to F in $\mathcal{D}$ if and only if F_n converges weakly to F and if the second moments of F_n converge to the second moments of F.

Step 2: prove that there is a unique fixed point. We begin by showing that the key distributional identity has a unique solution with $\mathbb{E}\{Y\} = 0$.

LEMMA 1. *Define a map $M : \mathcal{D} \to \mathcal{D}$ as follows: if Y and Y' are identically distributed with distribution function F of zero mean, U is uniform $[0,1]$, and U, Y and Y' are independent, then $M(F)$ is the distribution function of*

$$UY + (1-U)Y' + \varphi(U).$$

Then we have a contraction with respect to the Wasserstein metric:

$$d_2(M(F), M(G)) \le \sqrt{\frac{2}{3}}\, d_2(F,G).$$

(In other words, $M(.)$ is a contraction, and thus, there is a unique fixed point $F \in \mathcal{D}$ with $M(F) = F$. Indeed, set $F_0 = F$, $F_{n+1} = M(F_n)$, and note that $d_2(F_{n+1}), F_n) \le \sqrt{2/3}\, d_2(F_n, F_{n-1}) \le (2/3)^n\, d_2(F_1, F_0) \to 0$ as $n \to \infty$. But then F_n converges in $(\mathcal{D}, d_2)$ to some distribution function F. There is at least one solution. There can't be two, as $d_2(M(F), M(G)) = d_2(F,G)$.)

PROOF. Let X, X' both have distribution function F, and let Y, Y' both have distribution function G, and assume that the means are zero. Furthermore, let U be as above, let $(X,Y), (X',Y'), U$ be independent. Then, as $M(F)$ is the distribution function of $UX + (1-U)X' + \varphi(U)$, and $M(G)$ is that of $UY + (1-U)Y' + \varphi(U)$,

$$\begin{aligned}
d_2^2(M(F), M(G)) &\le \mathbb{E}\{(UX + (1-U)X' - UY - (1-U)Y')^2\} \\
&= \mathbb{E}\{(U(X-Y) + (1-U)(X'-Y'))^2\} \\
&= \mathbb{E}\{(U(X-Y))^2\} + \mathbb{E}\{((1-U)(X'-Y'))^2\} \\
&\qquad \text{(this step uses the zero mean assumption)} \\
&= 2\mathbb{E}\{U^2\}\mathbb{E}\{(X-Y))^2\} \\
&= \frac{2}{3}\mathbb{E}\{(X-Y))^2\}.
\end{aligned}$$

Take the infimum with respect to all possible pairs (X,Y) and obtain

$$d_2^2(M(F), M(G)) \le \frac{2}{3}\, d_2^2(F,G). \square$$

Observe that this proof can also be used to show that for any $\alpha > 1/2$,

$$Y \stackrel{\mathcal{L}}{=} U^\alpha Y + (1-U)^\alpha Y' + \varphi(U)$$

is a contraction.

Step 3: prove convergence to the fixed point. So, we are done if we can show that $d_2(F_n, F) \to 0$, where F_n is the distribution function of Y_n and F is the distribution function of Y. This will be the first time we will need the actual form of φ. We note throughout that Y and Y_n have zero mean and that both have finite variances. From the recursive distributional identity of L_n, we deduce these distributional identities: $Y_0 = Y_1 = 0$, and furthermore, for $n \geq 2$,

$$\begin{aligned}
Y_n &\stackrel{\mathcal{L}}{=} \frac{L_n - \mathbb{E}\{L_n\}}{n} \\
&\stackrel{\mathcal{L}}{=} \frac{L_{Z_n-1} + L'_{n-Z_n} + n - 1}{n} \\
&\quad - \frac{\mathbb{E}\{L_{Z_n-1}|Z_n\} + \mathbb{E}\{L'_{n-Z_n}|Z_n\} + (n-1)}{n} \\
&\quad + \frac{\mathbb{E}\{L_{Z_n-1}|Z_n\} - \mathbb{E}\{L_{Z_n-1}\} + \mathbb{E}\{L'_{n-Z_n}|Z_n\} - \mathbb{E}\{L'_{n-Z_n}\}}{n} \\
&\stackrel{\mathcal{L}}{=} Y_{Z_n-1}\frac{Z_n - 1}{n} + Y'_{n-Z_n}\frac{n - Z_n}{n} + \varphi_n(Z_n)
\end{aligned}$$

where

$$\begin{aligned}
\varphi_n(j) &= \frac{\mathbb{E}\{L_{j-1}\} + \mathbb{E}\{L'_{n-j}\} - \big(\mathbb{E}\{L'_{n-Z_n}\} + \mathbb{E}\{L_{Z_n-1}\}\big)}{n} \\
&= \frac{\mathbb{E}\{L_{j-1}\} + \mathbb{E}\{L'_{n-j}\} - (\mathbb{E}\{L_n - (n-1)\})}{n}.
\end{aligned}$$

Let us do a quick back-of-the-envelope calculation, replacing Z_n by nU, $n - Z_n$ by $n(1-U)$ and $\mathbb{E}\{L_i\}$ by $2i \log i$. If the Y_n's indeed converge to Y in law, then we have that Y should be approximately distributed as

$$UY + (1-U)Y' + \varphi(U).$$

THEOREM 8.

$$\lim_{n\to\infty} d_2(F_n, F) = 0.$$

PROOF. We let (Y', Y'_n) be an independent copy of (Y, Y_n) for all n, and let these pairs be independent of U. It is critical to pick versions of (Y_j, Y) and $Y'_j, Y')$ such that

$$\mathbb{E}\{(Y_j - Y)^2\} = \mathbb{E}\{(Y'_j - Y')^2\} = d_2^2(F_j, F).$$

Also, the sequence of Y_j's is independent, and similarly for Y'_j. Since Z_n is uniform on $\{1, \ldots, n\}$, it is distributed as $\lceil nU \rceil$, where U is uniform $[0,1]$. Thus,

$$Y_n \stackrel{\mathcal{L}}{=} Y_{\lceil nU \rceil - 1}\frac{\lceil nU \rceil - 1}{n} + Y'_{n-\lceil nU \rceil}\frac{n - \lceil nU \rceil}{n} + \varphi(\lceil nU \rceil).$$

Use one step in the distributional identity of Y to note the following:

$$d_2^2(F_n, F) \leq \mathbb{E}\left\{\left(Y_{\lceil nU \rceil - 1}\frac{\lceil nU \rceil - 1}{n} - UY + Y'_{n-\lceil nU \rceil}\frac{n - \lceil nU \rceil}{n} - (1-U)Y' + \varphi(\lceil nU \rceil) - \varphi(U)\right)^2\right\}$$

where we picked the same Y and Y' in both distributional identities. By the fact that Y, Y', Y_n, Y_n' have all zero mean, and the independence of (Y, Y_n), (Y', Y_n') and U, the right-hand-side is equal to

$$2\mathbb{E}\left\{\left(Y_{\lceil nU\rceil-1}\frac{\lceil nU\rceil-1}{n}-UY\right)^2\right\}+\mathbb{E}\left\{(\varphi(\lceil nU\rceil)-\varphi(U))^2\right\}. \tag{5}$$

Here we used the symmetry in the first two terms. We treat each term in the upper bound separately, starting with the first one. Note that

$$\left|U-\frac{\lceil nU\rceil-1}{n}\right|\le\frac{1}{n},$$

so that, writing θ for any quantity in absolute value less than one, the first term is not more than

$$\begin{aligned}
&2\mathbb{E}\left\{\left((Y_{\lceil nU\rceil-1}-Y)\frac{\lceil nU\rceil-1}{n}+\frac{\theta Y}{n}\right)^2\right\}\\
&=\frac{1}{n}\sum_{j=1}^{n}2\mathbb{E}\left\{\left((Y_{j-1}-Y)\frac{j-1}{n}+\frac{\theta Y}{n}\right)^2\right\}\\
&\le\frac{1}{n}\sum_{j=1}^{n}2\left(\frac{j-1}{n}\right)^2\mathbb{E}\left\{(Y_{j-1}-Y)^2\right\}+\frac{1}{n}\sum_{j=1}^{n}4\frac{j-1}{n^2}\mathbb{E}\left\{\theta Y(Y_{j-1}-Y)\right\}+\frac{1}{n}\sum_{j=1}^{n}2\mathbb{E}\left\{\left(\frac{\theta Y}{n}\right)^2\right\}\\
&\le\frac{2}{n}\sum_{j=1}^{n}\left(\frac{j-1}{n}\right)^2 d_2^2(F_{j-1},F)+\frac{4}{n}\sup_j\mathbb{E}\left\{|Y||Y_{j-1}-Y|\right\}+\frac{2\mathbb{E}\left\{Y^2\right\}}{n^2}\\
&\le\frac{2}{3}\sup_{j<n}d_2^2(F_j,F)+\frac{4}{n}\sup_{j<n}d_2(F_j,F)\sqrt{\mathbb{E}\left\{Y^2\right\}}+\frac{2\mathbb{E}\left\{Y^2\right\}}{n^2}\\
&\quad\text{(by the Cauchy-Schwarz inequality)}\\
&\le\left(\frac{2}{3}+\frac{4\sqrt{\mathbb{E}\left\{Y^2\right\}}}{n}\right)\sup_{j<n}d_2^2(F_j,F)+\frac{4\sqrt{\mathbb{E}\left\{Y^2\right\}}}{n}+\frac{2\mathbb{E}\left\{Y^2\right\}}{n^2}.
\end{aligned}$$

The second term of (5) is bounded from above by the square of

$$\sup_{0\le u\le 1}\left|\varphi_n(\lceil nu\rceil)-\varphi(u)\right|.$$

We recall that, using H_n for the n-th harmonic number, and $H_n=\log n+\gamma+(1/2n)+O(1/n^2)$, where $\gamma=0.57721566490153286060\ldots$ is the Euler-Mascheroni constant,

$$\mathbb{E}\{L_n\}=2nH_n+2H_n-4n=2n\log n+n(2\gamma-4)+2\log n+2\gamma+1+O(1/n).$$

Thus, uniformly over $2\le j\le n-1$,

$$\begin{aligned}
\varphi_n(j)&=\frac{n-1+\mathbb{E}\{L_{j-1}\}+\mathbb{E}\{L_{n-j}\}-\mathbb{E}\{L_n\}}{n}\\
&=\frac{n+2j\log(j-1)+2(n-j+1)\log(n-j)-2(n+1)\log n+O(1)}{n}
\end{aligned}$$

$$= 1 + \frac{2j}{n}\log\left(\frac{j-1}{n}\right) + \frac{2(n-j+1)}{n}\log\left(\frac{n-j}{n}\right) + O\left(\frac{1}{n}\right).$$

Denoting by θ any number between 0 and 1, we have uniformly over $2 \le j \le n-1$, and over $(j-1)/n < u \le j/n$,

$$\begin{aligned}
&|\varphi_n(\lceil nu \rceil) - \varphi(u)| \\
&\quad = O\left(\frac{1}{n}\right) + \left| \frac{2j}{n}\log\left(\frac{j-1}{n}\right) + \frac{2(n-j+1)}{n}\log\left(\frac{n-j}{n}\right) \right. \\
&\qquad \left. - \frac{2(j-\theta)}{n}\log\left(\frac{j-\theta}{n}\right) - \frac{2(n-j+\theta)}{n}\log\left(\frac{n-j+\theta}{n}\right) \right| \\
&\quad \le O\left(\frac{1}{n}\right) + \left| \frac{2j}{n}\log\left(\frac{j-1}{n}\right) - \frac{2j}{n}\log\left(\frac{j-\theta}{n}\right) \right| \\
&\qquad + \left| \frac{2(n-j)}{n}\log\left(\frac{n-j}{n}\right) - \frac{2(n-j)}{n}\log\left(\frac{n-j+\theta}{n}\right) \right| \\
&\qquad + \frac{2}{n}\left|\log\left(\frac{n-j}{n}\right)\right| + \frac{2}{n}\left|\log\left(\frac{j-\theta}{n}\right)\right| + \frac{2}{n}\left|\log\left(\frac{n-j+\theta}{n}\right)\right| \\
&\quad \le O\left(\frac{1}{n}\right) + \left| \frac{2j}{n}\log\left(\frac{j-1}{j-\theta}\right) \right| \\
&\qquad + \left| \frac{2(n-j)}{n}\log\left(\frac{n-j}{n-j+\theta}\right) \right| + \frac{6\log n}{n} \\
&\quad = O\left(\frac{\log n}{n}\right).
\end{aligned}$$

For $j = 1$ and $j = n$,

$$\varphi_n(j) = \frac{n - 1 + \mathbb{E}\{L_{n-1}\} - \mathbb{E}\{L_n\}}{n} = 1 + O\left(\frac{\log n}{n}\right),$$

and for the same j, with $u = (j-\theta)/n$,

$$\sup_{0 \le \theta \le 1} |\varphi(u) - 1| = O\left(\frac{\log n}{n}\right).$$

Combining this shows that

$$\sup_{0 \le u \le 1} |\varphi_n(\lceil nu \rceil) - \varphi(u)| = O\left(\frac{\log n}{n}\right).$$

We conclude that for some universal constants C, D,

$$d_2^2(F_n, F) \le \left(\frac{2}{3} + \frac{D}{n}\right) \sup_{j<n} d_2^2(F_j, F) + \frac{C}{n}.$$

It is left as an easy exercise to show that this implies that

$$\lim_{n\to\infty} d_2^2(F_n, F) = O\left(\frac{1}{n}\right). \square$$

§18. Acknowledgment

The referee's commens are much appreciated.

§19. References

D. Aldous, "Asymptotic fringe distributions for general families of random trees," *The Annals of Applied Probability*, vol. 1, pp. 228–266, 1991.

C. Aragon and R. Seidel, "Randomized search trees," *Algorithmica*, vol. 16, pp. 464–497, 1996.

C. R. Aragon and R. G. Seidel, "Randomized search trees," in: *Proceedings of the 30th IEEE Symposium on the Foundations of Computer Science*, pp. 540–545, 1989.

A. Bagchi and A. K. Pal, "Asymptotic normality in the generalized Polya-Eggenberger unrn model, with an application to computer data structures," *SIAM Journal on Algeraic and Discrete Methods*, vol. 6, pp. 394–405, 1985.

P. Baldi and Y. Rinott, "On normal approximation of distributions in terms of dependency graph," *Annals of Probability*, vol. 17, pp. 1646–1650, 1989.

R. C. Bradley, "Central limit theorems under weak dependence," *Journal of Multivariate Analysis*, vol. 11, pp. 1–16, 1981.

B. M. Brown, "Martingale central limit theorems," *Annals of Mathematical Statistics*, vol. 42, pp. 59–66, 1971.

L. H. Y. Chen, "Two central limit problems for dependent random variables," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 43, pp. 223–243, 1978.

L. H. Y. Chen and Q.-M. Shao, "Normal approximation under local dependence," *Annals of Probability*, vol. 32, pp. 185–2028, 2004.

A. Dembo and Y. Rinott, "Some examples of normal approximations by Stein's method," in: *Random Discrete Structures*, (edited by D. Aldous and R. Pemantle), pp. 25–44, Springer-Verlag, New York, 1996.

L. Devroye, "A note on the height of binary search trees," *Journal of the ACM*, vol. 33, pp. 489–498, 1986.

L. Devroye, "Branching processes in the analysis of the heights of trees," *Acta Informatica*, vol. 24, pp. 277–298, 1987.

L. Devroye, "Applications of the theory of records in the study of random trees," *Acta Informatica*, vol. 26, pp. 123–130, 1988.

L. Devroye, "Limit laws for local counters in random binary search trees," *Random Structures and Algorithms*, vol. 2, pp. 303–316, 1991.

L. Devroye, "On random Cartesian trees," *Random Structures and Algorithms*, vol. 5, pp. 305–327, 1994.

L. Devroye, "Limit laws for sums of functions of subtrees of random binary search trees," *SIAM Journal on Computing*, vol. 32, pp. 152–171, 2002.

L. Devroye and R. Neininger, "Distances and finger search in random binary search trees," *SIAM Journal on Computing*, vol. 33, pp. 647–658, 2004.

M. Drmota, "Stochastic analysis of tree-like data structures," *Proceedings of the Royal Society: Mathematical, Physical and Engineering Sciences*, vol. 460, pp. 271–307, 2004.

A. Dvoretzky, "Central limit theorems for dependent random variables," in: *Proceedings of the Sizth Berkeley symposium on Mathematical statistics and Probability*, pp. 513–535, University of California Press, 1972.

C. G. Esseen, "Fourier analysis of distribution functions: a mathematical study of the Laplace-Gaussian law," *Acta Mathematica*, vol. 77, pp. 1–125, 1945.

W. Feller, "On the Berry-Esseen theorem," *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 10, pp. 261–268, 1968.

J. A. Fill, "On the distribution of binary search trees under the random permutation model," *Random Structures and Algorithms*, vol. 8(1) , pp. 1–25, 1996.

J. A. Fill and S. Janson, "Smoothness and decay properties of the limiting QUICKSORT density function," in: *Mathematics and Computer Science: Algorithms, Trees, Combinatorics and Probabilities*, (edited by D. Gardy and A. Mokkadem (eds)), pp. 53–64, Birkhäuser, Boston, 2000.

J. A. Fill and S. Janson, "Approximating the limiting QUICKSORT distribution," *Random Structures and Algorithms*, vol. 19, pp. 376–406, 2001.

J. A. Fill and S. Janson, "Quicksort asymptotics," *Journal of Algorithms*, vol. 44, pp. 4–28, 2002.

P. Flajolet and R. Sedgewick, "Digital search trees revisited," *SIAM Journal on Computing*, vol. 15, pp. 748–767, 1986.

P. Flajolet, X. Gourdon, and C. Martinez, "Patterns in random binary search trees," *Random Structures and Algorithms*, vol. 11(3), pp. 223–244, 1997.

J. Françon, G. Viennot, and J. Vuillemin, "Description and analysis of an efficient priority queue representation," in: *Proceedings of the 19th IEEE Symposium on the Foundations of Computer Science*, pp. 1–7, 1978.

G. H. Gonnet, *Handbook of Algorithms and Data Structures*, Addison-Wesley, Reading, Mass., 1984.

P. Grabner and H. Prodinger, "Sorting algorithms for broadcast communications: mathematical analysis," *Theoretical Computer Science*, vol. 289, pp. 51–67, 2002.

G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes*, Oxford University Press, 1992.

P. Hall and C. C. Heyde, *Martingale Limit Theory and its Applications*, Academic Press, New York, 1980.

W. Hoeffding and H. Robbins, "The central limit theorem for dependent random variables," *Annals of Mathematical Statistics*, vol. 20, pp. 773–780, 1949.

H.-K. Hwang and R. Neininger, "Phase change of limit laws in quicksort recurrence under varying toll functions," *SIAM Journal on Computing*, vol. 31, pp. 1687–1722, 2002.

I. A. Ibragimov, "A note on the central limit theorem for dependent random variables," *Theory of Probability and its Applications*, vol. 20, pp. 135–141, 1975.

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

D. E. Knuth, *The Art of Computer Programming, Vol. 3 : Sorting and Searching*, Addison-Wesley, Reading, Mass., 1973.

D. E. Knuth and A. Schönhage, "The expected linearity of a simple equivalence algorithm," *Theoretical Computer Science*, vol. 6, pp. 281–315, 1978.

G. Louchard, "Exact and asymptotic distributions in digital binary search tress," *RAIRO Theoretical Informatics and Applications*, vol. 13, pp. 253–276, 1987.

W. C. Lynch, "More combinatorial problems on certain trees," *Computer Journal*, vol. 7, pp. 299–302, 1965.

H. Mahmoud and B. Pittel, "On the most probable shape of a search tree grown from ," *SIAM Journal on Algebraic and Discrete Methods*, vol. 5, pp. 69–81, 1984.

H. M. Mahmoud, “The expected distribution of degrees in random binary search trees,” *The Computer Journal*, vol. 29, pp. 36–37, 1986.

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

D. L. McLeish, “Dependent central limit theorems and invariance principles,” *Annals of Probability*, vol. 2, pp. 620–628, 1974.

R. Neininger and L. Rüschendorf, “A general contraction theorem and asymptotic normality in combinatorial structures,” *Annals of Applied Probability*, vol. 14, pp. 378–418, 2004.

A. Panholzer and H. Prodinger, “Binary search tree recursions with harmonic toll functions,” *Journal of Computational and Applied Mathematics*, vol. 142, pp. 211–225, 2002. To appear.

B. Pittel, “On growing random binary trees,” *Journal of Mathematical Analysis and Applications*, vol. 103, pp. 461–480, 1984.

P. V. Poblete and J. I. Munro, “The analysis of a fringe heuristic for binary search trees,” *Journal of Algorithms*, vol. 6, pp. 336–350, 1985.

M. Régnier, “A limiting distribution for quicksort,” *RAIRO Informatique Théorique et Application*, vol. 23, pp. 335–343, 1989.

Y. Rinott, “On normal approximation rates for certain sums of dependent random variables,” *Journal of Computational and Applied Mathemtics*, vol. 55, pp. 135–143, 1994.

J. M. Robson, “The height of binary search trees,” *The Australian Computer Journal*, vol. 11, pp. 151–153, 1979.

U. Rösler, “A limit theorem for quicksort,” *Theoretical Informatics and Applications*, vol. 25, pp. 85–100, 1991.

U. Rösler, “A fixed point theorem for distributions,” *Stochastic Processes and their Applications*, vol. 42, pp. 195–214, 1992.

U. Rösler and L. Rüschendorf, “The contraction method for recursive algorithms,” *Algorithmica*, vol. 29, pp. 3–33, 2001.

R. Sedgewick, “Mathematical analysis of combinatorial algorithms,” in: *Probability Theory and Computer Science*, (edited by G. Louchard and G. Latouche), pp. 123–205, Academic Press, London, 1983.

A. A. Shergin, “On the convergence rate in the central limit theorem for m-dependent random variables,” *Theory of Probability and its Applications*, vol. 24, pp. 782–796, 1979.

I. S. Shiganov, “Refinement of the upper bound of the constant in the central limit theorem,” *Journal of Soviet Mathematics*, vol. 50, pp. 2545–2550, 1986.

C. Stein, “A bound for the error in the normal approximation to the distribution of a sum of dependent random variables,” *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, vol. 2, pp. 583–602, 1972.

K. H. Tan and P. Hadjicostas, “Some properties of a limiting distribution in quicksort,” *Statistics and Probability Letters*, vol. 25, pp. 87–94, 1995.

P. Van Beek, “An application of Fourier methods to the problem of sharpening the Berry-Esseen inequality,” *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 23, pp. 183–196, 1972.

J. S. Vitter and P. Flajolet, “Average-case analysis of algorithms and data structures,” in: *Handbook of Theoretical Computer Science, Volume A: Algorithms and Complexity*, (edited by J. van Leeuwen), pp. 431–524, MIT Press, Amsterdam, 1990.

J. Vuillemin, “A data structure for manipulating priority queues,” *Communications of the ACM*, vol. 21, pp. 309–314, 1978.