Elect. Comm. in Probab. **12** (2007), 28–35

ELECTRONIC COMMUNICATIONS in PROBABILITY

A PROBABILISTIC PROOF OF A WEAK LIMIT LAW FOR THE NUMBER OF CUTS NEEDED TO ISOLATE THE ROOT OF A RANDOM RECURSIVE TREE

ALEX IKSANOV
National Taras Shevchenko University of Kiev, 60, Volodymyrska, 01033 Kiev, Ukraine
email: iksan@unicyb.kiev.ua

MARTIN MÖHLE
Mathematical Institute, University of Tübingen, Auf der Morgenstelle 10, 72076 Tübingen, Germany
email: martin.moehle@uni-tuebingen.de

Submitted November 24, 2006, *accepted in final form* February 21, 2007

AMS 2000 Subject classification: Primary: 60F05, 60G50; Secondary: 05C05, 60E07
Keywords: coupling, random recursive tree, random walk, stable limit

Abstract

We present a short probabilistic proof of a weak convergence result for the number of cuts needed to isolate the root of a random recursive tree. The proof is based on a coupling related to a certain random walk.

1 Introduction and main result

Meir and Moon [9] introduced a procedure to isolate a root of a random recursive tree T with n vertices, by the following successive deletions (cuts) of edges. One starts with choosing one of the $n-1$ edges at random and cuts this edge. This edge-removing procedure is iterated with the subtree containing the root, and the procedure stops as soon as the root has been isolated. For more information on (random) recursive trees and related edge-removal procedures we refer to Janson [7, 8] and Panholzer [11].

For $n \in \mathbb{N} := \{1, 2, \ldots\}$ let X_n denote the number of cuts needed to isolate the root of a random recursive tree with n vertices. It is known [3] that the sequence $(X_n)_{n\in\mathbb{N}}$ satisfies the distributional recursion $X_1 = 0$ and

$$X_n \stackrel{d}{=} 1 + X_{n-D_n}, \quad n = 2, 3, \ldots, \tag{1}$$

where D_n is a random variable independent of $X_2, \ldots, X_n$ with distribution

$$P(D_n = k) \;=\; \frac{n}{(n-1)k(k+1)}, \quad k \in \{1, \ldots, n-1\}.$$

Note that (1) is equivalent to

$$P(X_n = j) \; = \; \frac{n}{n-1} \sum_{k=1}^{n-1} \frac{P(X_{n-k} = j-1)}{k(k+1)}, \quad j, n \in \mathbb{N}, j < n. \tag{2}$$

Equation (2) can be viewed as a recursion on $n \in \mathbb{N}$ with initial values $P(X_1 = j) = \delta_{j0}$, $j \in \mathbb{N}_0 := \{0, 1, 2, \ldots\}$, or, alternatively, as a recursion on $j \in \mathbb{N}_0$ with initial values $P(X_n = 0) = \delta_{n1}$, $n \in \mathbb{N}$, where δ denotes the Kronecker symbol.
As already pointed out in [2], there are important other interpretations for X_n. For example, in the language of coalescent processes, X_n is the number of collision events that take place in a Bolthausen-Sznitman n-coalescent [1] until there is just a single block. As a consequence, X_n can be also interpreted as the absorption time of the Markov chain, which counts the number of ancestors in a Bolthausen-Sznitman n-coalescent.
We are interested in the asymptotic behavior of X_n as n tends to infinity. The question of finding a weak limit law for X_n was motivated by the work of Meir and Moon [9]. This problem was unsolved for many years and, during that period, readdressed by several authors (see, for example, Panholzer [11, p. 269]). A first proof of the following weak limit law for X_n appeared in [3].

Theorem 1. *As n tends to infinity,*

$$\frac{(\log n)^2}{n} X_n - \log n - \log \log n$$

converges in distribution to a stable random variable X with characteristic function $\mathrm{E}(e^{itX}) = \exp(it \log |t| - \frac{\pi}{2}|t|)$, $t \in \mathbb{R}$.

Note that the same normalization as in Theorem 1 appears in [7] to derive an asymptotic result for the number of cuts needed to isolate the root of a complete binary tree.
The proof of Theorem 1 in [3] works with analytic methods and is based on a singular analysis of generating functions. It is natural and, in our opinion, important to understand the probabilistic structure behind Theorem 1. We therefore present a completely different, purely probabilistic proof of Theorem 1. Our proof is considerably shorter than that given in [3] and provides more insight in the probabilistic mechanisms behind the convergence. The key idea is to replace X_n by a suitable random variable M_n with $M_n \stackrel{d}{=} X_n$. We construct M_n in such a way that it is related (coupled) to the first passage time of a certain random walk and, hence, easier to handle than X_n. We also mention that our coupling method is quite general and might be useful to derive weak limit laws for other sequences $(X_n)_{n \in \mathbb{N}}$ as well, but we do not try to generalize our results here. We start with the coupling in Section 2, and prove Theorem 1 in Section 3.

2 A coupling

Let $(\xi_i)_{i \in \mathbb{N}}$ be a sequence of independent copies of a random variable ξ with values in $\mathbb{N}$. For arbitrary but fixed $n \in \mathbb{N}$, define a two-dimensional (coupled) process $(R_i, S_i)_{i \in \mathbb{N}_0}$ recursively via $(R_0, S_0) := (0, 0)$ and, for $i \in \mathbb{N}$,

$$(R_i, S_i) \; := \; (R_{i-1}, S_{i-1}) + \begin{cases} (\xi_i, \xi_i) & \text{if } \xi_i < n - R_{i-1}, \\ (0, \xi_i) & \text{else.} \end{cases}$$

The process $(R_i, S_i)_{i\in\mathbb{N}_0}$ depends on the parameter n. For convenience, we do not indicate this dependence in our notation. The process $(S_i)_{i\in\mathbb{N}_0}$ is a zero-delayed random walk ($S_i = \xi_1 + \cdots + \xi_i$, $i \in \mathbb{N}_0$) and does not depend on n. The process $(R_i)_{i\in\mathbb{N}_0}$ has non-decreasing paths, starts in $R_0 = 0$ and satisfies $R_i < n$ for all $i \in \mathbb{N}_0$. By induction on i it follows that $R_i \le S_i$, $i \in \mathbb{N}_0$.

Let $M_n := \#\{i \in \mathbb{N} \mid R_{i-1} \neq R_i\} = \sum_{l\ge 0} 1_{\{R_l+\xi_{l+1}<n\}}$ denote the total number of jumps of the process $(R_i)_{i\in\mathbb{N}_0}$. Note that $0 \le M_n \le n-1$, $n \in \mathbb{N}$. Let $N_n := \inf\{i \in \mathbb{N} \mid S_i \ge n\}$ denote the number of steps the random walk $(S_i)_{i\in\mathbb{N}_0}$ needs to reach a state larger than or equal to n. Note that $1 \le N_n \le n$, $n \in \mathbb{N}$.

We have $R_i = S_i < n$ for $i \in \{0, \ldots, N_n - 1\}$. Therefore, the process $(R_i)_{i\in\mathbb{N}_0}$ has at least $N_n - 1$ jumps, i.e., $M_n \ge N_n - 1$, $n \in \mathbb{N}$. We are interested in the asymptotics of M_n and N_n as $n \to \infty$. The following lemma provides a recursion for the distributions of the marginals M_n, $n \in \mathbb{N}$.

Lemma 1. *(Recursion for the distribution of M_n)*
The distribution of M_n satisfies the recursion $P(M_1 = \cdots = M_{n_0} = 0) = 1$ and, for $n > n_0$,

$$P(M_n = j) \;=\; \frac{1}{1-q_n}\sum_{k=1}^{n-1} p_k P(M_{n-k} = j-1), \qquad j \in \{1, \ldots, n-1\},$$

where $p_k := P(\xi = k)$, $q_k := P(\xi \ge k)$, $k \in \mathbb{N}$, and $n_0 := \sup\{k \in \mathbb{N} \mid q_k = 1\} \in \mathbb{N}$.

Proof. Obviously $P(\xi \ge n_0) = 1$. Hence, for fixed $n \le n_0$, the process $(R_i)_i$ is almost surely constant equal to zero, and, therefore, $M_1 = \cdots = M_{n_0} = 0$ almost surely. Now fix $n > n_0$ and let $I := \inf\{i \in \mathbb{N} \mid R_i > 0\}$ denote the first jump time of the process $(R_i)_i$. At this time, the process $(R_i)_i$ will reach a state $k \in \{1, \ldots, n-1\}$. Thus, for $j \in \{1, \ldots, n-1\}$,

$$\begin{aligned} P(M_n = j) &= \sum_{i\ge 1}\sum_{k=1}^{n-1} P(I = i, R_I = k, M_n = j) \\ &= \sum_{i\ge 1}\sum_{k=1}^{n-1} P(\xi_1 \ge n, \ldots, \xi_{i-1} \ge n, \xi_i = k, M_n = j). \end{aligned}$$

We now use a renewal argument similar to those presented in [6]. For $i, m \in \mathbb{N}$ define $\widehat{R}_0^{(i,m)} := 0$ and

$$\widehat{R}_{k+1}^{(i,m)} \;:=\; \widehat{R}_k^{(i,m)} + \xi_{i+k+1}\, 1_{\{\widehat{R}_k^{(i,m)}+\xi_{i+k+1}<m\}}, \qquad k \in \mathbb{N}_0.$$

Then, $\widehat{M}_{i,m} := \sum_{l=0}^{\infty} 1_{\{\widehat{R}_l^{(i,m)}+\xi_{i+l+1}<m\}}$ is an independent copy of M_m, which is also indepen-

dent of $\xi_1, \ldots, \xi_i$. Therefore,

$$\begin{aligned}
P(M_n = j) &= \sum_{i\ge 1}\sum_{k=1}^{n-1} P\Big(\xi_1 \ge n, \ldots, \xi_{i-1} \ge n, \xi_i = k, \sum_{l=0}^{\infty} 1_{\{\widehat{R}_l^{(0,n)}+\xi_{l+1}<n\}} = j\Big)\\
&= \sum_{i\ge 1}\sum_{k=1}^{n-1} P\Big(\xi_1 \ge n, \ldots, \xi_{i-1} \ge n, \xi_i = k, \sum_{l=i}^{\infty} 1_{\{\widehat{R}_l^{(0,n)}+\xi_{l+1}<n\}} = j-1\Big)\\
&= \sum_{i\ge 1}\sum_{k=1}^{n-1} P(\xi_1 \ge n, \ldots, \xi_{i-1} \ge n, \xi_i = k, \widehat{M}_{i,n-k} = j-1)\\
&= \sum_{i\ge 1}\sum_{k=1}^{n-1} P(\xi_1 \ge n)\cdots P(\xi_{i-1} \ge n)P(\xi_i = k)P(M_{n-k} = j-1)\\
&= \sum_{i\ge 1} q_n^{i-1}\sum_{k=1}^{n-1} p_k\, P(M_{n-k} = j-1) \;=\; \frac{1}{1-q_n}\sum_{k=1}^{n-1} p_k\, P(M_{n-k} = j-1).
\end{aligned}$$

$\square$

Remark. An analogous renewal argument can be used to derive a recursion for the joint distribution of M_n and N_n, but we do not need to study the joint distribution of M_n and N_n in our further considerations. We will only need an appropriate upper bound for the difference $M_n - N_n$.

3 Proof of Theorem 1

The following probabilistic proof of Theorem 1 is based on the coupling presented in Section 2 in which

$$p_k \;:=\; P(\xi = k) \;:=\; \frac{1}{k(k+1)}, \qquad k \in \mathbb{N}.$$

Proposition 1. *For each $n \in \mathbb{N}$, the distribution of M_n coincides with the distribution of a random variable X_n introduced in Section 1.*

Proof. We have $q_n = P(\xi \ge n) = 1/n$, $n \in \mathbb{N}$. By Lemma 1, for $n \ge 2$,

$$P(M_n = j) \;=\; \frac{n}{n-1}\sum_{k=1}^{n-1} \frac{P(M_{n-k} = j-1)}{k(k+1)}, \quad j \in \{1, \ldots, n-1\}. \tag{3}$$

This recursion coincides with the recursion (2) for the distributions of the random variables X_n. Therefore, $M_n \stackrel{d}{=} X_n$ for all $n \in \mathbb{N}$. $\square$

Proposition 2. *As n tends to infinity,*

$$\frac{(\log n)^2}{n} N_n - \log n - \log\log n$$

converges in distribution to a stable random variable with characteristic function $t \mapsto \exp(-\frac{\pi}{2}|t| + it\log|t|)$, $t \in \mathbb{R}$.

Proof. According to the theory of stable distributions and their domain of attraction (see, for example, Theorem 1 and Remark 3 in [5]),

$$\frac{S_n}{n} - \log n \;=\; \frac{\xi_1 + \cdots + \xi_n}{n} - \log n \;\to\; Z$$

in distribution as n tends to infinity, where Z is a random variable with characteristic function $\mathrm{E}(e^{itZ}) = \exp(-\frac{\pi}{2}|t| - it\log|t|)$, $t \in \mathbb{R}$. In applications, the distribution of Z is sometimes called the continuous Luria-Delbrück distribution (see, for example, [10]). Let F denote the distribution function of Z. As F is continuous, we have uniform convergence

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}} \Big| P\Big(\frac{S_n}{n} - \log n \le x\Big) - F(x) \Big| \;=\; 0.$$

Fix $x \in \mathbb{R}$. Let the integers n and k be functions of each other such that as $n \to \infty$ (or, equivalently, $k \to \infty$)

$$\frac{k}{n} - \log n \;\to\; x. \tag{4}$$

In the following it is assumed that all passages to the limit take place when $n \to \infty$ or $k \to \infty$. We have

$$P(N_k \le n) \;=\; P(S_n \ge k) \;=\; P\Big(\frac{S_n}{n} - \log n \ge \frac{k}{n} - \log n\Big) \;\to\; 1 - F(x).$$

Assume it is known that

$$\frac{n}{k}(\log k)^2 - \log k - \log\log k \to -x. \tag{5}$$

Then,

$$\begin{aligned} 1 - F(x) \;&\leftarrow\; P(N_k \le n) \\ &=\; P\Big(\frac{(\log k)^2}{k} N_k - \log k - \log\log k \le \frac{n}{k}(\log k)^2 - \log k - \log\log k\Big) \\ &\sim\; P\Big(\frac{(\log k)^2}{k} N_k - \log k - \log\log k \le -x\Big), \end{aligned}$$

which implies the result as $x \mapsto 1 - F(-x)$ is the distribution function of $-Z$, which has the desired characteristic function $\mathrm{E}(e^{-itZ}) = \exp(-\frac{\pi}{2}|t| + it\log|t|)$.
It remains to verify (5). From (4),

$$k \;\sim\; n\log n. \tag{6}$$

Therefore,

$$\log k - \log n - \log\log n \;\to\; 0, \tag{7}$$

which implies

$$\log k \;\sim\; \log n, \tag{8}$$

from which it follows

$$\log\log k - \log\log n \;\to\; 0. \tag{9}$$

From (6) and (8) we have

$$k \;\sim\; n\log k. \tag{10}$$

From (4) and (7)

$$\frac{k}{n} - \log k + \log\log n \ \to \ x.$$

By (9),

$$\log k \frac{n\log k - k}{n\log k} - \log\log k \ \to \ -x.$$

In view of (10),

$$\begin{aligned} A(n,k) &:= \log k \frac{n\log k - k}{k} - \frac{n\log k}{k}\log\log k \\ &= \log k \frac{n\log k - k}{k} - \Big(\frac{n\log k}{k} - 1\Big)\log\log k - \log\log k \\ &\to -x. \end{aligned} \tag{11}$$

Multiply $A(n,k)$ by $\log\log k/\log k$ to conclude that

$$\log\log k \frac{n\log k - k}{k} - \frac{n\log k}{k}\frac{(\log\log k)^2}{\log k} \ \to \ 0.$$

By (10), the second term tends to 0. Therefore,

$$\log\log k\Big(\frac{n\log k}{k} - 1\Big) \ \to \ 0.$$

Substituting this result into (11) gives (5). □

The following lemma presents a convergence result for the sequence of auxiliary random variables

$$Y_n \ := \ n - S_{\max\{i \,|\, S_i \le n\}} \ = \ n - S_{N_{n+1}-1}, \qquad n \in \mathbb{N}. \tag{12}$$

Lemma 2. *As n tends to infinity, $\log Y_n/\log n$ converges in distribution to Y, where Y is uniformly distributed on the unit interval $[0,1]$.*

Proof. This lemma is a direct consequence of Erickson [4, p. 287, Theorem 6]. In our situation, $P(\xi > x) = 1/([x]{+}1) \sim 1/x$, $x \to \infty$. It only remains to note that Erickson's Theorem 6 is still true when ξ is arithmetic, as mentioned at the beginning of Erickson's proof of his Theorem 6. By Erickson [4, p. 287, Remark 1], the 'truncated mean' function $m(t)$ in Theorem 6 is allowed to be replaced by $m_1(t) := \log t$, as $m(t) := \int_0^t P(\xi > x)\,dx = \sum_{i=1}^{[t]} 1/i + o(1) \sim \log t$. □

Lemma 2 allows to compare the random variables M_n and N_n as follows.

Proposition 3. *As n tends to infinity,*

$$\frac{(\log n)^2}{n}(M_n - N_n)$$

converges in probability to zero.

Proof. We have $R_i = S_i \le n-1$ for $i \in \{0,\ldots,N_n - 1\}$. Therefore,

$$\begin{aligned} M_n &= \#\{i \le N_n - 1 \,|\, R_{i-1} \ne R_i\} + \#\{i > N_n - 1 \,|\, R_{i-1} \ne R_i\} \\ &= N_n - 1 + \#\{i > N_n - 1 \,|\, R_{i-1} \ne R_i\}. \end{aligned}$$

From time $N_n - 1$ on, the process $(R_i)_i$ cannot have more than $(n-1) - R_{N_n-1}$ jumps, because otherwise this process would reach a state larger than $n-1$, which is impossible by construction. Therefore,

$$M_n \;\le\; N_n - 1 + (n-1) - R_{N_n-1} \;=\; N_n - 1 + (n-1) - S_{N_n-1},$$

or,

$$0 \le\; M_n - N_n + 1 \;\le n - 1 - S_{N_n-1} \;=\; Y_{n-1},$$

with Y_n defined via (12). It remains to verify that $Y_n/b_n \to 0$ in probability, where $b_n := n/(\log n)^2$. In order to see this fix $\varepsilon > 0$ and $\delta > 0$. From

$$\frac{\log(\varepsilon b_n)}{\log n} \;=\; \frac{\log\varepsilon + \log n - 2\log\log n}{\log n} \;\to\; 1, \quad n \to \infty$$

it follows that there exists a positive integer $n_0 = n_0(\varepsilon, \delta)$ such that

$$\frac{\log(\varepsilon b_n)}{\log n} \;\ge\; 1 - \delta \qquad \text{for all } n > n_0.$$

Therefore, for $n > n_0$,

$$P(Y_n > \varepsilon b_n) \;=\; P\Big(\frac{\log Y_n}{\log n} > \frac{\log(\varepsilon b_n)}{\log n}\Big) \;\le\; P\Big(\frac{\log Y_n}{\log n} > 1 - \delta\Big) \;\to\; \delta,$$

as $\log Y_n/\log n \to Y$ in distribution by Lemma 2, with Y uniformly distributed on $[0,1]$. But $\delta > 0$ can be chosen arbitrarily small, which shows that $P(Y_n > \varepsilon b_n) \to 0$. The convergence $Y_n/b_n \to 0$ in probability is established. □

Combining Proposition 1, 2, and 3 immediately yields that the convergence in Proposition 1 holds with N_n replaced by M_n or, alternatively, X_n. Thus we have found a probabilistic proof of Theorem 1, which was the aim of this study.

Acknowledgement. The authors would like to thank Gerold Alsmeyer for helpful comments. The work was done while A. Iksanov was visiting the Institute of Mathematical Statistics at the University of Münster in October and November 2006. He gratefully acknowledges financial support and hospitality.

References

[1] Bolthausen, E. and Sznitman, A.-S. (1998). On Ruelle's probability cascades and an abstract cavity method. *Comm. Math. Phys.* **197**, 247–276. MR1652734

[2] Drmota, M., Iksanov, A., Möhle, M., and Rösler, U. (2007). Asymptotic results about the total branch length of the Bolthausen-Sznitman coalescent. *Stoch. Process. Appl.* **117**, to appear

[3] Drmota, M., Iksanov, A., Möhle, M., and Rösler, U. (2006). A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree. Preprint, submitted to *Random Structures Algorithms*

[4] Erickson, K.B. (1970). Strong renewal theorems with infinite mean. *Trans. Amer. Math. Soc.* **151**, 263–291. MR0268976

[5] Geluk, J.L. and De Haan, L. (2000). Stable probability distributions and their domains of attraction: a direct approach. *Probab. Math. Statist.* **20**, 169–188. MR1785245

[6] Hinderer, K. and Walk, H. (1972). Anwendung von Erneuerungstheoremen und Taubersätzen für eine Verallgemeinerung der Erneuerungsprozesse. *Math. Z.* **126**, 95–115. MR0300354

[7] Janson, S. (2004). Random records and cuttings in complete binary trees. In: *Mathematics and Computer Science III*, Birkhäuser, Basel, pp. 241–253. MR2090513

[8] Janson, S. (2006). Random cutting and records in deterministic and random trees. *Random Structures Algorithms* **29**, 139–179. MR2245498

[9] Meir, A. and Moon, J.W. (1974). Cutting down recursive trees. *Math. Biosci.* **21**, 173–181. Math. Review number not available. Zbl 0288.05102

[10] Möhle, M. (2005). Convergence results for compound Poisson distributions and applications to the standard Luria-Delbrück distribution. *J. Appl. Probab.* **42**, 620–631. MR2157509

[11] Panholzer, A. (2004). Destruction of recursive trees. In: *Mathematics and Computer Science III*, Birkhäuser, Basel, pp. 267–280. MR2090518