2007 Conference on Analysis of Algorithms, AofA 07 DMTCS proc. **AH**, 2007, 303–312

One-sided Variations on Tries: Path Imbalance, Climbing, and Key Sampling

Costas A. Christophi[1] and Hosam M. Mahmoud [2]

[1] *Cyprus International Institute for the Environment and Public Health in association with Harvard School of Public Health, 1105, Nicosia, Cyprus; email: cchristophi@cyprusinstitute.org*
[2] *Department of Statistics, The George Washington University, Washington, D.C. 20052, U.S.A.; email: hosam@gwu.edu*

received 26th February 2007, *revised 23rd January 2008*, *accepted tomorrow*.

One-sided variations on path length in a trie (a sort of digital trees) are investigated: They include imbalance factors, climbing under different strategies, and key sampling. For the imbalance factor accurate asymptotics for the mean are derived for a randomly chosen key in the trie via poissonization and the Mellin transform, and the inverse of the two operations. It is also shown from an analysis of the moving poles of the Mellin transform of the poissonized moment generating function that the imbalance factor (under appropriate centering and scaling) follows a Gaussian limit law. The method extends to several variations of sampling keys from a trie and we sketch results of climbing under different strategies. The exact probability distribution is computed in one case, to demonstrate that such calculations can be done, at least in principle.

Keywords: Random trees, digital trees, recurrence, Mellin transform, poissonization, depoissonization, singularity analysis.

1 Introduction

This extended abstract summarizes results in Mahmoud (2007+) and Christophi and Mahmoud (2007+), where one-sided measures relating to imbalance and trie climbing are treated. Imbalance in random binary trees has come into the spotlight when Donald Knuth posed it as a question of interest in his keynote address to the 2004 Workshop on Analysis of Algorithms, which convened at Berkeley, and the random climbing of tries has been a subject that authors revisit from time to time. It was considered in Moon (1970) and in Meir and Moon (1975, 1978).

Kuba and Panholzer (2007+) analyze the basic binary search models for imbalance. We carry out this program a step further and study the imbalance of keys in one flavor of digital trees.

The subject of tree climbing has been revisited recently in Panholzer (2005), who considered several classes of random trees, including simply generated families and Pólya trees. In these investigations, a class of trees is considered, and a type of random walk on it is exercised. Starting at the root, certain nodes are accessed, and at each node a randomly selected edge emanating from it is chosen *at random* (all edges coming out of a node being equally likely). The process is perpetuated until it is no longer possible to proceed. When the process is stopped the path inscribed in the tree by climbing reaches a leaf. We shall study the climbing path length in tries, under this and a few other strategies.

The *binary trie* is a basic digital tree structure well known to this community. For properties and uses see Knuth (1998) or Mahmoud (1992). Suppose we have $n \ge 0$ keys given in their dyadic representation. We assume that the keys in the trie are independent and of infinite precision and within each the bits are independent with p probability of a bit being 1, and q probability of being 0 ($p + q = 1$). Thus, each key can be viewed as an infinite sequence of Bernoulli trials. This model is often called the *Bernoulli model*.

2 The imbalance factor

The *imbalance factor* of a key is the number of right-going edges minus the number of left-going edges on the path from the root of the tree to that key. We shall study Δ_n, the imbalance factor of a randomly chosen key in a digital tree constructed from n keys.

Let L_n and R_n be respectively the number of keys residing in the left and right subtrees, among the n keys stored in the tree (so, $L_n + R_n = n$). In view of the Bernoulli model, L_n is distributed like a binomial random variable on n independent trials with rate of success q in each. Owing to the independence of the keys and the bits within, the recursion of the insertion algorithm preserves the probabilistic structure in the subtrees of the trie.

Given L_n, Δ_n can be Δ_{L_n} minus one with probability L_n/n when the randomly chosen key is from the left subtree, or Δ_{R_n} plus one with probability R_n/n when the randomly chosen key is from the right subtree. We thus have

$$\Delta_n \mid L_n = \begin{cases} \Delta_{L_n} - 1, & \text{with probability } \frac{L_n}{n}; \\ \\ \Delta_{R_n} + 1, & \text{with probability } \frac{R_n}{n}, \end{cases} \tag{1}$$

with boundary conditions $\Delta_0 = \Delta_1 = 0$.

2.1 Functional equations

We derive a functional equation for $\phi_n(t)$, the moment generating function of Δ_n, from the basic conditional recurrence (1):

$$\mathbf{E}\big[e^{\Delta_n t} \mid L_n\big] = e^{-t} e^{\Delta_{L_n} t} \times \frac{L_n}{n} + e^t e^{\Delta_{R_n} t} \times \frac{R_n}{n},$$

with an unconditional expectation satisfying

$$n\phi_n(t) := n\mathbf{E}\big[e^{\Delta_n t}\big] = e^{-t}\mathbf{E}\big[L_n e^{\Delta_{L_n} t}\big] + e^t \mathbf{E}\big[R_n e^{\Delta_{R_n} t}\big], \tag{2}$$

valid for $n \ge 2$. It is not straightforward to solve this recurrence. However, a poissonized version of the problem is amenable to the Mellin transform (see Flajolet, Gourdon and Dumas, 1995). Subsequently, we suppose that instead of fixed n, the number of keys to be stored in the tree, is first determined by a random draw from a Poisson distribution with parameter z. Let N_z be such a random number.

We introduce the generating function $\Phi(t, z) = e^{-z} \sum_{n=0}^{\infty} n\phi_n(t) \frac{z^n}{n!}$, so that $\Phi(t, z) = \mathbf{E}\big[N_z \phi_{N_z}(t)\big]$ is the Poisson transform of the sequence $n\phi_n(t)$. We multiply both sides of (2) by $z^n e^{-z}/n!$, and sum over $n \geq 2$. We do the calculations on the right-hand side by conditioning on L_n, and get

$$\begin{aligned}\Phi(t,z) - ze^{-z} &= e^{-t}e^{-z}\sum_{n=2}^{\infty}\frac{z^n}{n!}\sum_{\ell=0}^{n}\ell\phi_\ell(t)q^\ell p^{n-\ell}\binom{n}{\ell} \\ &\qquad + e^{t}e^{-z}\sum_{n=2}^{\infty}\frac{z^n}{n!}\sum_{r=0}^{n}r\phi_r(t)p^r q^{n-r}\binom{n}{r} \\ &= e^t\Phi(t,pz) + e^{-t}\Phi(t,qz) - ze^{-z}(pe^t + qe^{-t}).\end{aligned}$$

The function $\Phi(t,z)$ does not have a Mellin transform. However, the shifted function $\Psi(t,z) := \Phi(t,z) - ze^{-z}$ does. The functional equation for Ψ is

$$\Psi(t,z) = e^t\Psi(t,pz) + e^{-t}\Psi(t,qz) + pze^t(e^{-pz} - e^{-z}) + qze^{-t}(e^{-qz} - e^{-z}).$$

For the Mellin transform of a term like $e^{-pz} - e^{-z}$, we first write it as $(e^{-pz} - 1) - (e^{-z} - 1)$ to facilitate the computation in the desired existence strip. Subsequently, for small $|t|$ we get

$$\Psi^*(t,s) = \frac{\Gamma(s+1)[e^t(p^{-s} - p) + e^{-t}(q^{-s} - q)]}{1 - e^t p^{-s} - e^{-t}q^{-s}}, \tag{3}$$

in the strip $\langle -2, -s_0(t)\rangle$, where $s_0(t)$ is the unique real solution of the *bivariate characteristic* equation

$$1 - e^t p^{-s} - e^{-t}q^{-s} = 0. \tag{4}$$

We restrict t to be in an interval around 0, so that the right edge of the strip intersects the real line at a point contained in a small neighborhood of $s = -1$.

We shall see in Section 2.3 a shortcut to the calculation of the leading terms in the asymptotic expansion of the mean and variance. We carry out in the next subsection a more detailed asymptotic expansion to capture such refined details as an oscillating behavior in the mean.

2.2 The Mean

We can find all the poissonized (and ultimately depoissonized) moments from the Mellin transform of the bivariate moment generating function. The kth derivative of (3) with respect to t, when evaluated at $t = 0$, yields a functional equation for the Mellin transform of the poissonized kth moment of Δ_n. An accurate asymptotic expansion for the kth moment can then be recovered by the inverse Mellin transform then depoissonization. We shall carry out this program on the first moment, keeping in mind that it can be extended, only at the expense of increasing computational effort, to higher moments.

The first derivative of $\Psi(t,z)$ is $A_1(z) := \frac{\partial}{\partial t}\Psi(t,z)\Big|_{t=0} = \mathbf{E}\big[N_z\Delta_{N_z}\big]$, with the Mellin transform,

$$A_1^*(s) = -\frac{(p-q)\Gamma(s+1)}{1 - p^{-s} - q^{-s}},$$

existing in $\langle -2, -1\rangle$. The poissonized average is retrieved by the inversion

$$A_1(z) = \frac{1}{2\pi i}\int_{-\frac{3}{2}-i\infty}^{-\frac{3}{2}+i\infty} A_1^*(s)z^{-s}\,ds.$$

We evaluate this integral by "the method of closing the box" (see Szpankowski, 2001):

$$A_1(z) = -\sum \text{Residues of poles in } \Big\langle -\frac{3}{2}, \theta \Big\rangle + O(z^{-\theta}), \tag{5}$$

for some $\theta > 0$.

This leaves us with residue calculation at the poles of the Gamma function, and the roots of the *characteristic equation* $1 - p^{-s} - q^{-s} = 0$ which have been studied before (see the refined exposition in Drmota, Reznik, Savari and Szpankowski (2007+)).

The results are conveniently expressed in terms of the entropy-like functions $h_p = -p \ln p - q \ln q$, and $\tilde{h}_p = p \ln^2 p + q \ln^2 q$. Let $a = \inf_{\Re s_k > -1} \Re s_k$ be the smallest of any real part of a root lying to the right of the fundamental strip $\langle -2, -1 \rangle$. As θ is arbitrarily chosen, we can take $\theta > a$. We next collect contributions of the residues of the poles to the right of the fundamental strip, as required in (5). Note that the contribution of all the poles to the right of $s_0 = -1$ is $O(z^{-a})$. Putting all residues together we get

$$\mathbf{E}[N_z \Delta_{N_z}] \;=\; (p-q)\Big(\frac{z \ln z}{h_p} + \Big(\frac{\gamma}{h_p} + \frac{\tilde{h}_p}{2h_p^2} + \beta_1(z)\Big) z + O(z^{-a})\Big),$$

where, γ is Euler's constant, and $\beta_1(.)$ is the function

$$\beta_1(z) = \begin{cases} \dfrac{1}{h_p} \displaystyle\sum_{\substack{k=-\infty \\ k \neq 0}}^{\infty} \Gamma\Big(\frac{2\pi i k r}{\ln p}\Big) z^{-\frac{2\pi i k r}{\ln p}}; & \text{if } \frac{\ln p}{\ln q} = \frac{r}{m}, \text{ rational with } \gcd(r, m) = 1, \\ 0; & \text{otherwise.} \end{cases} \tag{6}$$

In all cases, $\beta_1(.)$ is a small absolutely bounded function.

All conditions for depoissonization (see Jacquet and Szpankowski (1998)) are satisfied.

Proposition 1 *In a trie of n random keys following the Bernoulli model, the average imbalance factor of a randomly selected key is*

$$\mathbf{E}[\Delta_n] \;=\; (p-q)\Big(\frac{\ln n}{h_p} + \frac{\gamma}{h_p} + \frac{\tilde{h}_p}{2h_p^2} + \beta_1(n) + O\Big(\frac{1}{n^{\min\{a+1, 0.99999\}}}\Big)\Big),$$

where $\beta_1(n)$ is the oscillating function given in (6), and a is the minimum among the real parts of characteristic roots to the right of -1.

Remark: *In principle, one could continue pumping the higher moments in this manner, but the computation becomes too involved.*

2.3 Limit distribution

The *bivariate characteristic* equation (4) has an infinitely countable number of roots ($s_k(t)$, for $k = 0, \pm 1, \pm 2, \ldots$) that depend on t. If t is chosen from a small interval around 0, and $\ln p / \ln q$ is an irrational number, the equation has one real root $s_0(t)$, and $\Re(s_k) > s_0(t)$, for all $k \neq 0$. But if $\ln p / \ln q = r/m$ is rational (with r and m being two integers with $\gcd(r, m) = 1$) the equation has an infinite number of equispaced roots lined up on the vertical line located at $s_0(t)$. These roots are of the form $\tilde{s}_k(t) =$

$s_0(t) + 2\pi i k r/\ln p$, for $k = 0 \pm 1, \pm 2, \ldots$, and move with t. The rest of the roots fall to the right of $s_0(t)$. The functions $s_k(t)$ are continuous and infinitely differentiable. In particular the essential root $s_0(t)$ has a Taylor series expansion $s_0(t) = s_0(0) + s_0'(0)t + \frac{1}{2}s_0''(0)t^2 + O(t^3)$. The bivariate characteristic equation (4) immediately gives $s_0(0) = -1$. The first derivative of (4) gives us $s_0'(0) = -\frac{p-q}{h_p}$ and the second one gives

$$-s_0''(0) = \frac{1}{h_p}\Big(1 + 2s_0'(0)(q\ln q - p\ln p) + \big(s_0'(0)\big)^2\tilde{h}_p\Big), \sigma_p^2 = \frac{pq\ln^2(pq)}{h_p^3}.$$

We shall consider the roots for a small range of t around 0. The inverse Mellin transform is $\Psi(t,z) = \frac{1}{2\pi i}\int_{-\frac{3}{2}-i\infty}^{-\frac{3}{2}+i\infty} \Psi^*(t,s) z^{-s}\, ds$. Evaluating this integral by closing the box we have

$$\Psi(t,z) = \Phi(t,z) - ze^{-z} \sim -\operatorname*{Res}_{s=s_0(t)} \frac{z^{-s}\Gamma(s+1)[e^t(p^{-s}-p) + e^{-t}(q^{-s}-q)]}{1 - e^t p^{-s} - e^{-t}q^{-s}}.$$

Theorem 1 *Let Δ_n be the imbalance factor of a randomly chosen key in a random trie constructed from n keys from a biased Bernoulli model. Then*

$$\frac{\Delta_n - \frac{p-q}{h_p}\ln n}{\sqrt{\ln n}} \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \frac{pq\ln^2(pq)}{h_p^3}\Big).$$

Proof. We restrict t to a range such that $s_0(t)$ falls in a small interval around -1. As $t \to 0$, the Gamma function valuation at $s_0(t)+1$ becomes very large, whereas, for $k \neq 0$, the valuation at $s_k(t)+1$ remains bounded. Thus, if we let $t \to 0$, as we shall, the essential contribution comes from $s_0(t)$. After depoissonization we have

$$n\mathbf{E}\big[e^{\Delta_n t}\big] \sim \frac{n^{-s_0(t)}\Gamma(s_0(t)+1)[e^t(p^{-s_0(t)}-p) + e^{-t}(q^{-s_0(t)}-q)]}{e^t p^{-s_0(t)}\ln p + e^{-t}q^{-s_0(t)}\ln q}.$$

We take $t = v/\sqrt{\ln n}$, for fixed v, and let n be very large. Using the expansions $c^{-x} - c \sim -(c\ln c)(x+1)$, and $\Gamma(x+1) \sim \frac{1}{x+1}$, as $x \to -1$, we obtain

$$\begin{aligned}\mathbf{E}\big[e^{\Delta_n \frac{v}{\sqrt{\ln n}}}\big] &\sim n^{-s_0(\frac{v}{\sqrt{\ln n}})-1} \\ &\sim e^{-s_0'(0)\frac{v\ln n}{\sqrt{\ln n}} - s_0''(0)\frac{v^2}{2}}.\end{aligned}$$

So, we can write $\mathbf{E}\big[e^{(\Delta_n + s_0'(0)\ln n)\frac{v}{\sqrt{\ln n}}}\big] \to e^{-s_0''(0)\frac{v^2}{2}}$. The right-hand side in the latter relation is the moment generating function of $\mathcal{N}\big(0, -s_0''(0)\big)$, the normal random variate with mean 0 and variance $-s_0''(0)$. It follows from Lévy's continuity theorem that $\frac{\Delta_n + s'(0)\ln n}{\sqrt{\ln n}} \xrightarrow{\mathcal{L}} \mathcal{N}\big(0, -s''(0)\big)$. □

3 Climbing a trie

We consider the climbing of tries under different strategies. We change the "model of randomness" from the usual uniform choice of edges to a climbing model that conforms with the manner in which these tries are randomly generated. The trie emerges from keys that are taken from a data generator that emits bits of data with 1's having probability p, and 0's having probability $q = 1 - p$. The process may not necessarily end on a leaf, as it may terminate at a null node, but it always generates a key (not necessarily in the trie).

The interpretation of this climbing is that a typical key is being "sampled" from the data base. Hence, we call this strategy *typical climbing*. In the absence of knowledge of the key generating probability, we consider an alternative strategy called *uninformed climbing* in which we follow the right and left nexus with equal probability. We also consider the case of sampling extremal data.

3.1 Typical climbing

In typical sampling, we climb a trie by following an algorithm that emulates the natural frequency of bits. We start at the root and access nodes. At each node accessed we generate an independent Ber(p) random variable. If this variable yields 0, we follow the left edge if it exists (otherwise, the climbing is stopped), and if the value generated is 1, we follow the right edge if it exists (otherwise, the climbing is stopped).

Let S_n be the number of nodes on the path inscribed in the trie by the typical climbing, and let $\phi_n(t)$ be its moment generating function. Note that S_n can be linked to the depth D_n. If we are inserting the $(n+1)$st key this will follow the path of S_n. If the climbing terminates at an empty node, S_n and D_{n+1} are the same, but if the climbing terminates at a key, we need to insert a number of additional nodes. This number is geometrically distributed, but is dependent on S_n.

With L_n, R_n as before, the variable S_n satisfies a basic recurrence:

$$S_n \mid L_n = \begin{cases} 1 + S_{L_n}, & \text{with probability } q; \\ \\ 1 + S_{R_n}, & \text{with probability } p. \end{cases}$$

Let $\phi_n(t)$ be the moment generating function of S_n. Towards poissonization we construct the super generating function $A(z,t) = \sum_{n=0}^{\infty} \frac{\phi_n(t)}{n!} z^n$. With a development similar to what was done for imbalance we get

$$\begin{aligned} A(z,t) &= qe^t e^{pz} A(qz,t) + pe^t e^{qz} A(pz,t) \\ &\qquad + 1 - e^t + ze^t - (p^2+q^2)e^{2t} z - 2pqe^t z. \end{aligned}$$

To guarantee the existence of the Mellin transform, we deal with the shifted super moment generating function $B(z,t) = e^{-z}(A(z,t) - 1)$. This has the interpretation: $B(z,t) = \mathbf{E}\big[e^{S_{N(z)}t}\big] - e^{-z}$, where $N(z)$ is a Poisson random variable with parameter z. Whence, we obtain the functional equation

$$\begin{aligned} B(z,t) &= e^t \Big(pB(pz,t) + qB(qz,t) \\ &\qquad + p(e^{-pz} - e^{-z}) + q(e^{-qz} - e^{-z}) + (p^2+q^2) z e^{-z} \big(1 - e^t\big) \Big). \end{aligned}$$

The Mellin transform of this function is

$$B^*(s,t) = \frac{e^t \Gamma(s) \big(p^{-s+1} + q^{-s+1} - 1 + (p^2+q^2)(1-e^t)s\big)}{1 - e^t(p^{-s+1} + q^{-s+1})},$$

existing in the domain $-1 < \Re s < s_0(t)$, where $s_0(t)$ is the only real solution to

$$p^{-s+1} + q^{-s+1} = e^{-t}.$$

We shall keep $|t|$ small enough for the entire strip $\langle s_0(t), -s_0(t)\rangle$ to be contained in $\langle -\frac{1}{4}, \frac{1}{4}\rangle$.

Theorem 2 *Let S_n be the number of nodes on the path of typical climbing of a trie on n keys from the* Ber*(p) model. Then*

$$\begin{aligned}
\mathbf{E}[S_n] &= \frac{\ln n}{h_p} + \frac{1}{h_p}\big(\gamma - 1 - \ln p + 2pq - \ln q\big) \\
&\qquad -\frac{1}{2h_p^2}\big(p\ln^2 p + 2\ln p \ln q + q\ln^2 q\big) + \eta_1(\ln n) + o(1)\big), \\
\mathbf{Var}[S_n] &\sim \frac{pq(\ln p - \ln q)^2}{h_p^3}\ln n + o(\ln n),
\end{aligned}$$

where $\eta_1(.)$ is absolutely bounded by a very small number. The lower-order terms in the variance may also have small absolutely bounded oscillations.

Proof. Similar to that of Proposition 1, and we omit it. □

Theorem 3 *Let S_n be the number of nodes on the path of typical climbing of a trie on n keys from a biased* Ber*(p) model. Then*

$$\frac{S_n - \frac{1}{h_p}\ln n}{\sqrt{\ln n}} \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \frac{pq}{h_p^3}(\ln p - \ln q)^2\Big).$$

Proof. Similar to the proof of Theorem 1, and we omit it. □

3.2 Climbing with the lack of knowledge of p

If one is uninformed about p, one may be inclined to plead ignorance and simply generate moves in the random walk to the right and left subtrees with equal probability, hoping that this will average good and bad cases achieving a sampling strategy that is not too much worse than typical climbing.

The result below indicates that the average speed of climbing is improved in uninformed climbing on average. Of course the two strategies coincide when $p = q = \frac{1}{2}$, but uninformed climbing requires less time than typical climbing as p gets away from $\frac{1}{2}$, and the uninformed strategy speeds up considerably near the extremal values $p = 0$ and $p = 1$. However, the improved performance in the uninformed search comes at the expense of the quality of sampling, as less probable keys are given more weight than their actual probability.

Let $\tilde{S}_n$ be the number of nodes on the path inscribed in the trie by the uninformed climbing. The length $\tilde{S}_n$ satisfies a basic recurrence:

$$\tilde{S}_n \,|\, L_n = \begin{cases} 1 + \tilde{S}_{L_n}, & \text{with probability } \frac{1}{2}; \\[2ex] 1 + \tilde{S}_{R_n}, & \text{with probability } \frac{1}{2}. \end{cases}$$

The techniques used and the derivations are similar to those already presented, and we only state the results without proof.

Theorem 4 *Let $\tilde S_n$ be the number of nodes on the path of uninformed climbing of a trie on n keys from the* Ber*(p) model. Then*

$$\begin{aligned}\mathbf{E}[\tilde S_n] &= 2\log_{\frac{1}{pq}} n + \frac{\ln^2 p + (1-2\gamma)\ln(pq) + \ln^2 q}{\ln^2(pq)} + \eta_2(\ln n) + o(1),\\ \mathbf{Var}[\tilde S_n] &= \frac{2(\ln p - \ln q)^2}{\ln^3 \frac{1}{pq}}\ln n + O(\ln n),\end{aligned}$$

where $\eta_2(.)$ is a small function given by a Fourier expansion. The $o(\ln n)$ term in the variance may also have small bounded oscillations. Moreover,

$$\frac{\tilde S_n - 2\log_{\frac{1}{pq}} n}{\sqrt{\ln n}} \xrightarrow{\mathcal{L}} \mathcal{N}\Big(0, \frac{2(\ln p - \ln q)^2}{\ln^3 \frac{1}{pq}}\Big).$$

3.3 Extremal sampling

To develop a sense for the extremes of the data present in the trie, a sampler may take after the extremal strategy of following leftmost (for smallest) or rightmost (for largest) edges. Of course the two strategies are symmetric with respect to the roles of p and q, and we only analyze one of them.

Let us introduce $\hat S_n$ as the number of nodes on the leftmost path. If the leftmost path reaches a null node we may augment the corresponding prefix of zeros with a 1 to construct a representative sample of the smallest data. We state the result without proof.

Theorem 5 *Let $\hat S_n$ be the number of nodes on the path of leftmost extremal climbing of a trie on n keys from the* Ber*(p) model. Then*

$$\begin{aligned}\mathbf{E}[\hat S_n] &= \log_{\frac{1}{q}} n + \frac{2q + \ln q - 2\gamma}{2\ln q} + \eta_3(\ln n) + o(1),\\ \mathbf{Var}[\hat S_n] &= \frac{1}{12} + \frac{\pi^2}{6\ln^2 q} + \frac{2q}{\ln q} - \frac{q^2}{\ln^2 q} + o(1),\end{aligned}$$

where $\eta_3(.)$ is absolutely bounded by a very small number. The lower order terms in the variance may add small bounded oscillations. Furthermore, $\hat S_n - \lfloor \log_{\frac{1}{q}} n \rfloor$ does not have a nontrivial limit in distribution under any scaling.

3.4 The exact distribution

Some of the exact distributions within the scope of this research may be amenable to direct combinatorial methods. We illustrate this for extremal climbing.

Theorem 6 *Let $\hat{S}_n$ be the number of nodes on the path of leftmost extremal climbing of a trie on $n \geq 2$ keys from the* Ber*(p) model. Then, for $k \geq 2$,*

$$\mathbf{P}(\hat{S}_n = k) = nq^{k-1}\big(q(1-q^{k-1})^{n-1} - (1-q^{k-2})^{n-1}\big) + (1-q^k)^n - (1-q^{k-1})^n,$$

and $\mathbf{P}(\hat{S}_n = 0) = 0$*, and* $\mathbf{P}(\hat{S}_n = 1) = p^n$.

Proof. The boundary cases $\mathbf{P}(\hat{S}_n = k)$, for $k = 1, 2$ are trivial. We develop the result in terms of the number of edges $S'_n = \hat{S}_n - 1$. Letting $k \geq 2$ we dissect the event $\{S'_n = k\}$ into two disjoint subsets. One of them, A_1, corresponds to the case where the tree goes down the left path k edges and then turns right, with all the keys having a string of k zeros as prefix and 1 at position $k+1$ (there must be at least two such keys). This construction leaves a null node dangling at the leftmost position in the tree. This can occur by having r keys, $r = 2, \ldots, n$, in the subtree the root of which is a sibling of the leftmost null node; the probability for any specific r to have this key structure is $(q^k p)^r$. The rest of the $n - r$ keys are not allowed to have a prefix of k 0's, otherwise they would disturb the pattern. The probability for these keys not to have the forbidden prefix is $(1-q^k)^{n-r}$. The r keys can be chosen in $\binom{n}{r}$ ways. Hence, $\mathbf{P}(A_1) = \sum_{r=2}^{n} \binom{n}{r}(pq^k)^r(1-q^k)^{n-r}$. The second event, A_2, corresponds to having exactly one key at the end of a leftmost path with k internal vertices on it. By combinatorial arguments similar to that for $\mathbf{P}(A_1)$ we see that $\mathbf{P}(A_2) = \sum_{r=1}^{n-1}(r+1)\binom{n}{r+1}(pq^{k-1})^r q^k (1-q^{k-1})^{n-r-1}$. Hence,

$$\begin{aligned}\mathbf{P}(S'_n = k) &= \mathbf{P}(A_1 \cup A_2)\\ &= \sum_{r=2}^{n} \binom{n}{r}(pq^k)^r(1-q^k)^{n-r}\\ &\qquad + \sum_{r=1}^{n-1}(r+1)\binom{n}{r+1}(pq^{k-1})^r q^k(1-q^{k-1})^{n-r-1}.\end{aligned}$$

The sums can be reduced via the binomial theorem. □

References

[1] Christophi, C. and Mahmoud, H. (2007+). On Climbing tries (manuscript).

[2] Drmota, M., Reznik, Y. Savari, S. and Wojciech Szpankowski, W. (2007+) Analysis of variable-to-fixed length codes. (manuscript).

[3] Flajolet, P., Gourdon, X. and Dumas, P. (1995). Mellin transform and asymptotic harmonic sums. *Theoretical Computer Science*, **144**, 3–58.

[4] Jacquet, P. and Szpankowski, W. (1998). Analytical depoissonization and its applications. *Theoretical Computer Science*, **201**, 1–62.

[5] Knuth, D. (1998). *The Art of Computer Programming*, **Vol. 3**: *Sorting and Searching*, 2nd ed. Addison-Wesley, Reading, Massachusetts.

[6] Kuba, M. and Panholzer, A. (2007+). The left-right-imbalance of binary search trees (manuscript).

[7] Mahmoud, H. (1992). *Evolution of Random Search Trees*. Wiley, New York.

[8] Mahmoud, H. (2000). *Sorting: A Distribution Theory*. Wiley, New York.

[9] Mahmoud, H. (2007+). Imbalance in random digital trees. *Methodology and Computing in Applied Probability* (submitted).

[10] Meir, A. and Moon, J. (1975). Climbing certain types of rooted trees I. *Proceedings of the Fifth British Combinatorial Conference*, 461–469.

[11] Meir, A. and Moon, J. (1978). Climbing certain types of rooted trees II. *Acta Mathematica Academia Scientiarum Hungaricae*, **31**, 43–54.

[12] Moon, J. (1970). Climbing random trees. *Aequationes Mathematicae*, **5**, 68-74.

[13] Panholzer, A. (2005). The climbing depth of random trees. *Random Structures and Algorithms*, **26**, 84–109.

[14] Szpankowski, W. (2001). *Average Case Analysis of Algorithms on Sequences*. Wiley, New York.