1994

An Analytic Approach for the Asymptotic Distribution of the Height of an Incomplete Digital Tree

Hosam Mahmoud

Wojciech Szpankowski
Purdue University, spa@cs.purdue.edu

Report Number:
94-062

AN ANALYTIC APPROACH FOR THE ASYMPTOTIC DISTRIBUTION OF THE HEIGHT OF AN INCOMPLETE DIGITAL TREE

Hosam Mahmoud
Wojciech Szpankowski

CSD-TR-94-062
September 1994

AN ANALYTIC APPROACH FOR THE ASYMPTOTIC DISTRIBUTION OF THE HEIGHT OF AN INCOMPLETE DIGITAL TREE

Hosam Mahmoud*† Wojciech Szpankowski**‡

*Department of Statistics/Statistical Computing
The George Washington University
Washington, D.C. 20052, U.S.A.

**Department of Computer Science
Purdue University
West Lafayette, IN 47907, U.S.A.

Abstract. We investigate the duration of an elimination process for identifying a loser by coin tossing, or equivalently the height of a random incomplete trie. Via Poissonization and de-Poissonization, we obtain a simple expression for the asymptotic discrete distribution of the height by an analytic approach that is not widely known in the context of random tree height problems. The expression includes a periodic function which does not admit a limit distribution.

AMS 1991 subject classifications. Primary 05C05, 60E05; secondary 05C80, 60G70.

Key words and phrases. Random trees, height, de-Poissonization, asymptotic distribution, extreme value.

† This author's research is partially supported by NSA grant MDA904–92–H3086.
‡ This author's research is supported by NSF grants NCR-9206315 and CCR-9201078.

1. Introduction

We investigate the distribution of the height of a random incomplete trie, the discrete structure that underlies the process of identifying a loser by coin tossing. Height distributions of random digital trees have usually been attacked by purely probabilistic methods which only identify the leading terms [cf. Devroye (1992), Flajolet (1983), Mendelson (1982), and Pittel (1985,1986)]. Our approach here is analytic and provides a bound on the rate of convergence.

The following elimination process has several applications, such as the election of a leader in a distributed or parallel system (a practise exercised when a token is lost or when synchronization is lost in a token-passing ring-connected computer network). A group of n people wishes to identify a loser by tossing fair coins. All n people who throw heads are winners, those who throw tails are candidate losers and should flip their coins again. The process is repeated among candidate losers until one loser is identified. If at any stage all remaining candidate losers throw heads, the tosses are considered inconclusive and they all participate again as candidate losers in the next round of coin tossing.

A binary tree structure underlies the above elimination process. At the root of the tree we have one node labelled with all participants. After all the participants flip their coins for the first time, winners (if any) are placed in a *leaf* node that is attached to the root as a right child, and all candidate losers are placed in a node that is attached as a left child. Leaf nodes are terminal nodes that are not developed any further. The process repeats recursively on every left child until a single loser is identified. The node containing the loser is also considered as a leaf as it is terminal. Figure 1 illustrates the discrete structure underlying the elimination process. In Figure 1, leaf nodes are represented as rectangles, all the other nodes of the tree have an oval shape. An edges of the tree in Figure 1 leading from a parent node to its child is labelled with H (head) or T (tail) according to the throw obtained by the group within the child node.

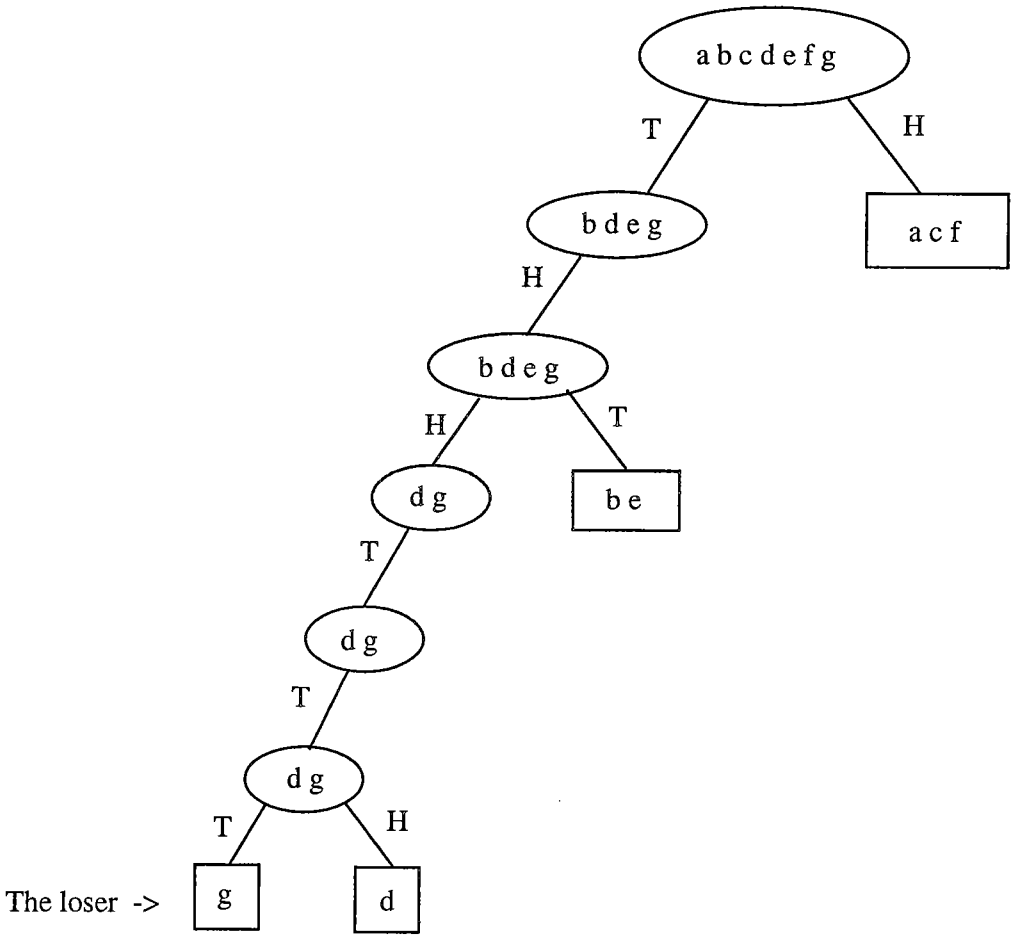

Figure 1. An incomplete trie for the elimination process starting with 7 people.

This random discrete structure is similar in some aspects to the random trie structure, a classical data structure for digital data [see Knuth (1973), Mahmoud (1992)]. The difference between the discrete structure of the elimination process and the standard trie is that in the trie the nodes which are right children of their parent are further developed if they contain more than two data items so that each datum is eventually in a node by itself. Thus, in a sense, the tree structure underlying the elimination process is an *incomplete trie* and will be called so in this paper. This terminology was coined in Prodinger (1993) who introduced this tree structure and found the average behavior of several of its characteristic properties. Grabner (1993) generalized the process to identify several losers instead of only one. Grabner (1993) finds the average behavior of some of the characteristic properties of

this more general incomplete trie.

The elimination process to identify a single or several losers also has the spirit of a class of problems posed by Réyni (1961) in his lecture series at Michigan State University. Pittel and Rubin (1992) find connections between one of Réyni's interesting questions and PATRICIA trees, a kind of tries with path compression for faster data retrieval [Knuth (1973)].

The *height* of an incomplete trie is the length of the path from the root to the loser, which is the longest root-to-leaf path in the tree. We shall denote the height of an incomplete trie underlying the elimination process beginning with n people by H_n. This quantity is the number of elimination rounds until the loser is identified, which is a measure of the time duration of the elimination process, if all the coin tosses at any stage are carried out simultaneously.

In this paper we investigate the asymptotic distribution of H_n. This random variable has a very wide range. It can assume values between 1 and ∞. We shall find the asymptotic distribution function of a centered version of H_n. With the proper centering, we shall see that no limit distribution exists. However, the distribution function of the centered H_n is sandwiched between well-defined extremes. Periodic fluctuations appear as n increases causing the distribution to gradually shift from one extreme to the other. More specifically, the periodic function[1]

$$\alpha(n) \stackrel{\text{def}}{=} \log n - \lfloor \log n \rfloor.$$

appears in the distribution of $H_n - \log n$. For values of n that are proper powers of 2, the left extreme is a discrete distribution function that coincides at the integer points with the continuous density

$$\frac{2^{-x}}{\exp(2^{-x}) - 1}.$$

[1] In this paper all logarithms with an unspecified base refer to the logarithm with base 2; the natural logarithm is denoted, as usual, by ln.

As n gradually increases, at any integer point the periodic effect of $\alpha(n)$ on the distribution is to lower the staircase discrete distribution till it comes very close to the other extreme discrete distribution function which is also a staircase that coincides (at the integer points) with the continuous density

$$\frac{2^{1-x}}{\exp(2^{1-x})-1}.$$

For various values of n the distribution function of $H_n - \log n$ is a staircase function lying between the two extreme discrete distributions. The periodic nature of $\alpha(n)$ makes the distribution function of $H_n - \log n$ "wrap around" every time n goes up from an integer preceding a natural power of 2 to become that power of two. This behavior will be formally expressed as a corollary to the main theorem.

We naturally chose the centering factor $\log n$ because this is the leading asymptotic term in the average height, as shown by Prodinger (1993) who found this fact by differentiating a functional equation for the probability generating function. In this paper we shall solve Prodinger's functional equation asymptotically.

The plan of this paper is as follows. Our starting point is Prodinger's functional equation for the probability generating function [Prodinger (1993)]. We take up this functional in Section 2. Fixed-population digital tree problems are notoriously hard and not amenable to direct attacks. However, these problems are known to be somewhat tractable under Poissonization, that is, when the number of objects (the number of people in our case) is assumed to follow a Poisson distribution instead of being a fixed number. Following the Poissonization path, we obtain in Section 2 an asymptotic expression for the moment generating function. The harmonic sum appearing in this moment generating function is approximated via the Mellin transform and its inverse. In Section 3 we de-Poissonize the problem back to the case of fixed population, thus solving Prodinger's functional equation asymptotically. We obtain in Section 3 the following main result.

Theorem. *For any integer k,*

$$\text{Prob}\{H_n \leq \log n + k\} = \frac{2^{\alpha(n)-k}}{\exp(2^{\alpha(n)-k}) - 1} + O\Big(\frac{1}{\sqrt{n}}\Big),$$

as $n \to \infty$. ∎

The function $\alpha(n)$ is dense on the interval $[0, 1)$ but not uniformly dense [see Kuipers and Niederreiter (1974)]. This consideration leads us to conclude the following.

Corollary. *A limit distribution cannot exist for $H_n - \log n$. However,*

$$\liminf_{n\to\infty} \text{Prob}\{H_n \leq \log n + k\} = \frac{2^{-k+1}}{\exp(2^{-k+1}) - 1},$$
$$\limsup_{n\to\infty} \text{Prob}\{H_n \leq \log n + k\} = \frac{2^{-k}}{\exp(2^{-k}) - 1},$$

at any integer k. ∎

The distribution in the main theorem can be expanded as a Taylor series involving doubly exponential terms which resemble the extreme value distribution (i.e., $e^{-e^{-x}}$) that often appears as the limiting distribution of the maximum of n *continuous* i.i.d. random variables [see Galambos (1987)]. However, already in 1970 Anderson observed that such a limiting distribution of n *discrete* i.i.d. random variables may not exist. This phenomenon occurs in the height of incomplete tries since, as we shall see from our concluding remarks, the height H_n can be represented as an extreme statistic. Actually, others problems on the standard digital trees led to similar conclusions [cf. Jacquet and Szpankowski (1991), Pittel (1986)], and we discuss this in a more detailed manner in Section 4.

2. Poissonization

Let

$$G_n(z) \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} \text{Prob}\{H_n = k\} z^k$$

be the probability generating of the fixed-population height H_n. We take as our starting point Prodinger's functional equation [Prodinger (1993)]:

$$G_n(z) = \frac{z}{2^n} \sum_{k=0}^{\infty} \binom{n}{k} G_k(z) - \frac{z}{2^n} + \frac{zG_n(z)}{2^n}, \tag{1}$$

which is valid for $n \geq 2$, with the boundary values $G_0(z) = G_1(z) = 0$.

It is difficult to solve Prodinger's functional equation directly. Poissonization however proved to be a fruitful avenue in digital problems [see Aldous (1989), Rais, Jacquet and Szpankowski (1993), etc.]. Possonization is carried out as follows. Suppose that instead of a fixed population, we first determine the number of people participating in the elimination contest by a draw from a Poisson distribution with parameter w. We allow w to be any positive real number. In fact later on, when we de-Poissonize the problem, we shall even allow w to be complex to be able to manipulate the resulting generating function by considering its analytic continuation to the w complex plane. Eventually we shall take $w = n$, when we de-Poissonize the problem.

We shall call the tree constructed under Poissonization the *Poissonized incomplete trie*. We shall also call its properties Poissonized; in particular we shall call the height of the Poissonized incomplete trie the *Poissonized height* and we simply denote it by H_N, where N, the Poissonized number of participants, is distributed according to a Poisson distribution with mean w. Introduce the generating function

$$g(w, z) = \sum_{n=0}^{\infty} \frac{G_n(z) w^n e^{-w}}{n!}.$$

This bivariate generating function has the following interpretation as a probability generating function for the Poissonized height:

$$\begin{aligned} \mathbf{E}\left[z^{H_N}\right] &= \sum_{k=0}^{\infty} \mathbf{E}\left[z^{H_N} \mid N = k\right] \frac{w^k e^{-w}}{k!} \\ &= \sum_{k=0}^{\infty} \mathbf{E}\left[z^{H_k}\right] \frac{w^k e^{-w}}{k!} \\ &= g(w, z). \end{aligned}$$

Multiplying both sides of (1) by $e^{-w}w^n/n!$ and summing over the range of validity of the recurrence (i.e. $n \geq 2$), then adjusting for the boundary cases $n = 0$ and $n = 1$, we obtain

$$g(w,z) = z(1+e^{-w/2})g\Big(\frac{w}{2},z\Big) + e^{-w}\big[(1+w)(1-z) - ze^{w/2}\big]. \tag{2}$$

To handle this latter recurrence, introduce

$$h(w,z) = \frac{g(w,z)}{1-e^{-w}},$$

which transforms the recurrence into

$$h(w,z) = zh\Big(\frac{w}{2},z\Big) + A(w,z),$$

where

$$A(w,z) \stackrel{\text{def}}{=} \frac{(1+w)(1-z) - ze^{w/2}}{e^w - 1}.$$

The last recurrence can now be solved by direct iteration [see Szpankowski (1987) for a general solution of this type of equations]. This gives

$$h(w,z) = \sum_{k=0}^{\infty} A\Big(\frac{w}{2^k},z\Big)z^k + \lim_{k\to\infty} A\Big(\frac{w}{2^k},z\Big)z^k. \tag{3}$$

The limit can easily be shown to be 0 if $|z| < 1/2$. For the rest of this work $|z|$ will be assumed to belong to the interior of the circle $|z| = 1/2$ in the z complex plane. The sum in (3) is a harmonic sum and can be accurately approximated by the Mellin transform method [see Flajolet (1988) or Mahmoud (1992)]. We denote the Mellin transform of a function $f(w,z)$, with respect to w, by $f^*(s,z)$, that is,

$$f^*(s,z) \stackrel{\text{def}}{=} \int_0^{\infty} f(w,z)w^{s-1}\,dw.$$

So, formally

$$h^*(s,z) = A^*(s,z)\sum_{k=0}^{\infty}(2^s z)^k.$$

This transform exists if:

(i) The sum $\sum_{k=0}^{\infty}(2^s z)^k$ converges, which happens if $\Re(s) < -\log|z|$.

(ii) The transform $A^*(s,z)$ exists. Using the geometric series representation

$$\frac{1}{e^w - 1} = \sum_{k=1}^{\infty} e^{-wk},$$

and the harmonic sum formula for Mellin transform [Flajolet (1988) or Mahmoud (1992)], we find after some simple algebra

$$A^*(s,z) = (1 - z2^s)\,\Gamma(s)\,\zeta(s) + (1-z)\,\Gamma(s+1)\,\zeta(s+1),$$

and both the gamma function $\Gamma(\cdot)$ and Riemann's zeta function $\zeta(\cdot)$ exist if $\Re(s) > 1$ [see Abramowitz and Stegun (1972)].

So, the Mellin transform $h^*(s,z)$ exists in a domain of s complex plane satisfying:

$$1 < \Re(s) < -\log|z|.$$

Observe that this fundamental strip is non-empty for $|z| < 1/2$, which is also sufficient to annihilate the limit in (3). Within this fundamental strip, the transform is given by

$$h^*(s,z) = \frac{(1 - z2^s)\,\Gamma(s)\,\zeta(s) + (1-z)\,\Gamma(s+1)\,\zeta(s+1)}{1 - 2^s z}.$$

Note that $h^*(s,0) = \Gamma(s)\zeta(s) + \Gamma(s+1)\zeta(s+1)$, for $\Re(s) > 1$, which has the simple inversion $(1+w)/(e^w - 1)$, leading to

$$g(w,0) = e^{-w}(1+w).$$

This can also be seen from direct probability considerations, as $H_0 = H_1 = 0$ and is also clear from the functional equation (2) when applied at $z = 0$.

For any fixed $z \neq 0$, but within the disc $|z| < 1/2$, we shall recover the asymptotic function $h(w,z)$ from its transform, for $w \to \infty$. Let γ be a real number between 1 and $-\log|z|$, so that the inverse Mellin transform is given by the integral

$$h(w,z) = \frac{1}{2\pi i}\int_{\gamma - i\infty}^{\gamma + i\infty} h^*(s,z) w^{-s}\,ds,$$

which is a line integral over a vertical line completely contained within the domain of existence of the Mellin transform $h^*(s,z)$. The line integral can be evaluated by a routine method involving shifting the line integral to the right to a vertical line past $-\log|z|$ (say at distance $d_z > -\log|z|$ from the vertical axis). We then compensate for the poles between the two lines by taking their residues. This shifting introduces only a small error. The method works because one can form a closed rectangular contour with the two vertical lines as its two sides; and a top and a bottom short horizontal segments at $i\infty$ and $-i\infty$ completing the rectangle, and the line integral is replaced by a contour integral. This can be done in a rigorous way in a limiting sense; one takes the horizontal line segments at $+iM$ and $-iM$, with M chosen in such a way that the two segments avoid the poles of $h^*(s,z)w^{-s}$, we then let $M \to \infty$. The line integrals over the horizontal segments diminish very quickly owing to the fast exponential decay in the magnitude of the $\Gamma(\cdot)$ function along a vertical line. As $M \to \infty$ the contribution of the top and bottom horizontal line segments approaches 0. The integral on the right line introduces an $O(w^{-d_z})$ term. We omit many details of the standard tool kit of the inverse Mellin transform. One finally finds [see Flajolet (1988) or Mahmoud (1992)]

$$h(w,z) = \sum \text{Residues} + O(w^{-d_z}),$$

where the sum is taken over all the residues along the line $-\log|z|$, and in the error term $d_z > -\log|z| > 1$.

The Poissonized problem has now been reduced to a computational problem involving residue calculation. We sketch this calculation next. The poles of $h^*(s,z)w^{-s}$ are the roots of the equation $2^s z = 1$ for $z \neq 0$. These are the simple roots

$$s_k = s_k(z) = -\log z + \frac{2\pi i k}{\ln 2}, \qquad k = 0, \pm 1, \pm 2 \ldots .$$

The residue at the pole s_0 is

$$-\frac{w^{\log z}}{\ln 2}(1-z)\,\Gamma(1-\log z)\,\zeta(1-\log z).$$

Collectively all the other roots contribute

$$-\frac{w^{\log z}}{\ln 2}\rho(w,z),$$

where

$$\rho(w,z) \stackrel{\text{def}}{=} (1-z)\sum_{\substack{k=-\infty\\ k\neq 0}}^{\infty} \Gamma\Big(1-\log z+\frac{2\pi i k}{\ln 2}\Big)\,\zeta\Big(1-\log z+\frac{2\pi i k}{\ln 2}\Big)e^{-2\pi i k\log w}\,. \qquad (4)$$

Collecting the contribution of all the poles and adding the error introduced by shifting the line integral, we reconstruct $h(w,z)$. We thus arrive at an asymptotic solution to a Poissonized version of (1). For $w\to\infty$, and fixed $z\neq 0$ from the disc $|z|<1/2$, the asymptotic Poissonized probability generating function is

$$\begin{aligned}\mathbf{E}\big[z^{H_N}\big] &= g(w,z)\\ &= (1-e^{-w})h(w,z)\\ &= -\frac{w^{\log z}}{\ln 2}\Big[(1-z)\,\Gamma(1-\log z)\,\zeta(1-\log z)+\rho(w,z)\Big]\big(1-e^{-w}\big)+O\big(w^{-d_z}\big),\end{aligned}$$

with $d_z > -\log|z| > 1$.

Remark. The last asymptotic expression for $g(w,z)$ is valid for any fixed $z\neq 0$ within the disc $|z|<1/2$. For such z, the O term is asymptotically negligible (as $|w|\to\infty$). A closer look at the constant of O reveals that the constant depends on z and increases superexponentially fast as $|z|$ approaches 0. (In fact, d_z grows like $\Gamma(1-\log|z|)$, as $z\to 0$.) On the other hand, at $z=0$ we know that $g(w,0)=e^{-w}(1+w)$. By the uniform continuity of the probability generating function $g(w,z)$, the two expressions must be close to each other near $z=0$, which must indeed be the case owing to proper cancellation of terms inside the O term (which are very large near $z=0$) and the other terms.

3. De-Poissonization

We use the De-Poissonization Lemma of Rais *et. al.* (1993) to extract the fixed-population probability generating function from the Poissonized probability generating function. That is, we shall find an asymptotic expression for $G_n(z)$ (as $n \to \infty$) from the asymptotic development of Section 2 for $g(w,z)$ (as $w \to \infty$). We restate next the De-Poissonization Lemma of Rais *et. al.* (1993).

Lemma. *For $\theta < \pi/2$, let S_θ be the cone $\{w : |\arg w| < \theta\}$. Suppose z belongs to a compact set in the vicinity of zero in the z complex plane. If, as $w \to \infty$, there exist real positive constants β_1, β_2, and $0 < a < 1$, such that:*

(i) For $w \in S_\theta$,

$$|g(w,z)| < \beta_1 |w|^\varepsilon,$$

(ii) For $w \notin S_\theta$,

$$|g(w,z)e^w| < \beta_2 |w|^\varepsilon e^{a|w|},$$

then for large n,

$$G_n(z) = g(n,z) + O\Big(\frac{1}{n^{1/2-\varepsilon}}\Big)$$

for any $\varepsilon < 1/2$. ■

The De-Poissonization Lemma says that if conditions (i) and (ii) are satisfied, the fixed-population probability generating function is essentially the same as the Poissonized probability generating function (with n replacing w) with an ignorable correction. So we need to verify (i) and (ii). We shall do this by considering the functional equation (2) as it is in a form suitable for induction.

We first need to introduce some notation. Suppose $1 < \lambda < 2$ is a given real number. Let

$$D_m = D_m(B) = \{w : B \le |w| < B\lambda^{m+1}\},$$

Note that if $w \in D_{m+1} - D_m$, then $w/2 \in D_m$, for $m = 0, 1, \ldots$.

Verification of the conditions of the De-Poissonization Lemma can now proceed by induction on m with respect to the domains D_m. Fix $\theta \in (0, \pi/2)$. The choices $\varepsilon = 0$, and $a = \cos\theta$ will suffice and henceforth will be assumed. The function xe^{-ax} is monotone decreasing if $x > 1/a$. We shall take advantage of this and repeatedly use

$$|w|e^{-a|w|} \leq Be^{-aB}, \qquad \text{for all } |w| > B > \frac{1}{a},$$

so we shall require $B > 1/a$. In the verification of (i) and (ii) below, a few other additional conditions will be required on B. So, B will be really the maximum that satisfies all the conditions simultaneously. Common to the verification of both items (i) and (ii) is the inequality

$$|g(w,z)| \leq \frac{1}{2}(1 + |e^{-w/2}|)\left|g\left(\frac{w}{2}, z\right)\right| + |e^{-w}|\left[\frac{3}{2}(1 + |w|) + \frac{1}{2}|e^{w/2}|\right], \tag{5}$$

which follows from (2) and the restriction of z within the disc $|z| < 1/2$.

Verification of (i):

We consider $\tilde{D}_m = D_m \bigcap S_\theta$. The continuous function $g(w,z)$ is bounded on the closure of $\tilde{D}_0$. So there must exist a positive constant β_1, such that $|g(w,z)| < \beta_1$, for every $w \in \tilde{D}_0$. Assume now that (i) holds for some $m \geq 0$ and $\varepsilon = 0$. That is, assume for $w \in \tilde{D}_m$ we have

$$|g(w,z)| < \beta_1.$$

Now suppose $w \in \tilde{D}_{m+1}$. If w is also in $\tilde{D}_m$, we are done. But if $w \in \tilde{D}_{m+1} - \tilde{D}_m$, then $w/2 \in \tilde{D}_m$. For the induction to work, B will have to be taken large enough to satisfy the following condition. Let δ be any fixed number in $(0, 1/2)$. We choose B so large that

$$B \geq \frac{2}{a} \ln \frac{1}{1 - 2\delta},$$

and

$$\frac{3}{2}e^{-aB}(1 + B) + \frac{1}{2}e^{-aB/2} \leq \delta\beta_1.$$

Then from (5), the induction hypothesis, monotonicity of $|w|e^{-a|w|}$, and our choices for B, we have

$$\begin{aligned}
|g(w,z)| &< \frac{1}{2}\big(1+e^{-a|w|/2}\big)\beta_1 + \frac{3}{2}e^{-a|w|}\big(1+|w|\big) + \frac{1}{2}e^{-a|w|/2} \\
&\leq \frac{1}{2}\big(1+e^{-aB/2}\big)\beta_1 + \frac{3}{2}e^{-aB}(1+B) + \frac{1}{2}e^{-aB/2} \\
&\leq (1-\delta)\beta_1 + \delta\beta_1 \\
&= \beta_1.
\end{aligned}$$

Verification of (ii):

We define $\hat{D}_m = D_m \cap \bar{S}_\theta$, where $\bar{S}_\theta$ is the complement of S_θ. Note that if $w \in \hat{D}_m$, then

$$|e^w| = e^{\Re(w)} \leq e^{a|w|}.$$

Again we employ induction on m. As in the verification of (i), at the basis of induction ($m = 0$) a constant $\beta_2 > 0$ must exist such that $|g(w,z)e^w| < \beta_2 e^{a|w|}$, over $\hat{D}_0$. Assume (ii) holds for $\hat{D}_m$, i.e. if $w \in \hat{D}_m$, then $|g(w,z)e^w| < \beta_2 e^{a|w|}$. Let $w \in \hat{D}_{m+1}$. Again, if $w \in \hat{D}_m$, we are done. If $w \in \hat{D}_{m+1} - \hat{D}_m$, then $w/2 \in \hat{D}_m$. As in the verification of (i), it follows from (5) and the induction hypothesis that

$$|g(w,z)e^w| \leq \frac{1}{2}e^{a|w|/2}\big(e^{\Re(w/2)}+1\big)\beta_2 + \frac{3}{2}\big(1+|w|\big) + \frac{1}{2}e^{\Re(w/2)}.$$

In view of the monotonicity of $|w|e^{-a|w|}$, we have

$$|g(w,z)e^w| < e^{a|w|}\Big[\frac{\beta_2}{2} + \frac{1}{2}(\beta_2+1)e^{-aB/2} + \frac{3}{2}e^{-aB}(1+B)\Big] < \beta_2 e^{a|w|},$$

with the rightmost inequality holding if B is chosen to satisfy

$$\frac{1}{2}(\beta_2+1)e^{-aB/2} + \frac{3}{2}e^{-aB}(1+B) < \frac{\beta_2}{2}.$$

This completes the verification of (i) and (ii) of the de-Poissonization Lemma.

So we finally arrive at the following asymptotic solution to Prodinger's functional equation (1):

$$\begin{aligned}\mathbf{E}\big[z^{H_n}\big] &= G_n(z)\\ &= \big(1-e^{-n}\big)h(n,z)+O\Big(\frac{1}{\sqrt{n}}\Big)\\ &= -\frac{n^{\log z}}{\ln 2}\Big[(1-z)\,\Gamma(1-\log z)\,\zeta(1-\log z)+\rho(n,z)\Big]+O\Big(\frac{1}{\sqrt{n}}\Big).\end{aligned}$$

We can now manipulate $G_n(z)$ to find the asymptotic distribution function of $H_n - \log n$. We have

$$\begin{aligned}\sum_{n=0}^{\infty}\mathrm{Prob}\{H_n\le j\}z^j &= \frac{G_n(z)}{1-z}\\ &= -\frac{n^{\log z}}{\ln 2}\Big[\Gamma(1-\log z)\,\zeta(1-\log z)+\frac{\rho(n,z)}{1-z}\Big]+O\Big(\frac{1}{\sqrt{n}}\Big).\end{aligned}$$

We use Cauchy's formula to extract the distribution function:

$$\begin{aligned}\mathrm{Prob}\{H_n\le j\} = -\frac{1}{2\pi i\ln 2}\Big[\oint\frac{n^{\log z}\,\Gamma(1-\log z)\,\zeta(1-\log z)}{z^{j+1}}\,dz\\ +\oint\frac{n^{\log z}\rho(n,z)}{(1-z)z^{j+1}}\,dz\Big]+\oint\frac{O(n^{-1/2})}{z^{j+1}}\,dz;\end{aligned}$$

These contour integrations are taken along a circle of radius $|z| < 1/2$ and centered at the origin. The last integral is clearly equal to 0. So, if we now choose $j = \lfloor\log n + k\rfloor = \lfloor\log n\rfloor + k$, for any integer k, then

$$\begin{aligned}\mathrm{Prob}\{H_n\le \log n+k\} &= \mathrm{Prob}\{H_n\le\lfloor\log n+k\rfloor\}\\ &= -\frac{1}{2\pi i\ln 2}\Big[\oint z^{\alpha(n)-k-1}\Gamma(1-\log z)\,\zeta(1-\log z)\,dz\\ &\qquad+\oint\frac{z^{\alpha(n)-k-1}\rho(n,z)}{(1-z)}\,dz\Big].\end{aligned}$$

The above integrals can be replaced by the residues of the poles of the integrands outside the contour as the integral at an infinite circle is 0. For the second integral we obtain from (4)

$$\oint\frac{z^{\alpha(n)-k-1}\rho(n,z)}{(1-z)}\,dz = 2\pi i\operatorname*{Res}_{z=1}\frac{z^{\alpha(n)-k-1}\rho(n,z)}{1-z}\equiv 0.$$

The integrand in the first integral has simple poles at

$$z_j = 2^j, \qquad j = 1, 2, \ldots,$$

which are contributed by the $\Gamma(\cdot)$ function, and one additional pole at $z_0 = 1$ coming from Riemann's zeta function. The residue at z_0 is $-\ln 2$. The residue at z_j, for $j = 1, 2, \ldots$, is

$$\frac{(-1)^j \ln 2}{(j-1)!} 2^{j(\alpha(n)-k)} \zeta(1-j).$$

The special values of Riemann's zeta function appearing in the last expression are related to the Bernoulli numbers B_r, $r = 0, 1, 2 \ldots$, as follows: $\zeta(0) = -1/2$, $\zeta(1-2k) = -B_{2k}/(2k)$, for $k = 1, 2, 3, \ldots$, and $B_{-2k} = 0$, for $k = 1, 2, \ldots$ [cf. Abramowitz and Stegun (1972)]. Subsequently we arrive at

$$\text{Prob}\{\tilde{H}_n \leq \log n + k\} = \Big[1 - 2^{\alpha(n)-k-1} + \sum_{j=1}^{\infty} 2^{j(\alpha(n)-k)} \frac{B_{2j}}{(2j)!} + O\Big(\frac{1}{\sqrt{n}}\Big)\Big].$$

The remaining sum is expressible in terms of the classical generating function of the Bernoulli numbers [cf. Abramowitz and Stegun (1972)]

$$\frac{y}{e^y - 1} = \sum_{j=0}^{\infty} \frac{B_j y^j}{j!},$$

and this yields our main theorem.

4. Concluding Remarks

We have established the asymptotic discrete distribution of the elimination process to identify a loser, or equivalently the height of a random incomplete trie. The expression for the asymptotic distribution includes a periodic function that does not admit a limiting distribution. However, as discussed in the introduction, the distribution lies between two well defined extremes. Each extreme is a discretized version of a simple continuous distribution function. That is, each extreme is a staircase that rises at the integer points to agree with the corresponding continuous density. The left extreme distribution $2^{-x}/(e^{2^{-x}} - 1)$ is sketched in Figure 2 and it sis compared with the double exponential distribution $e^{-2^{-x}}$.

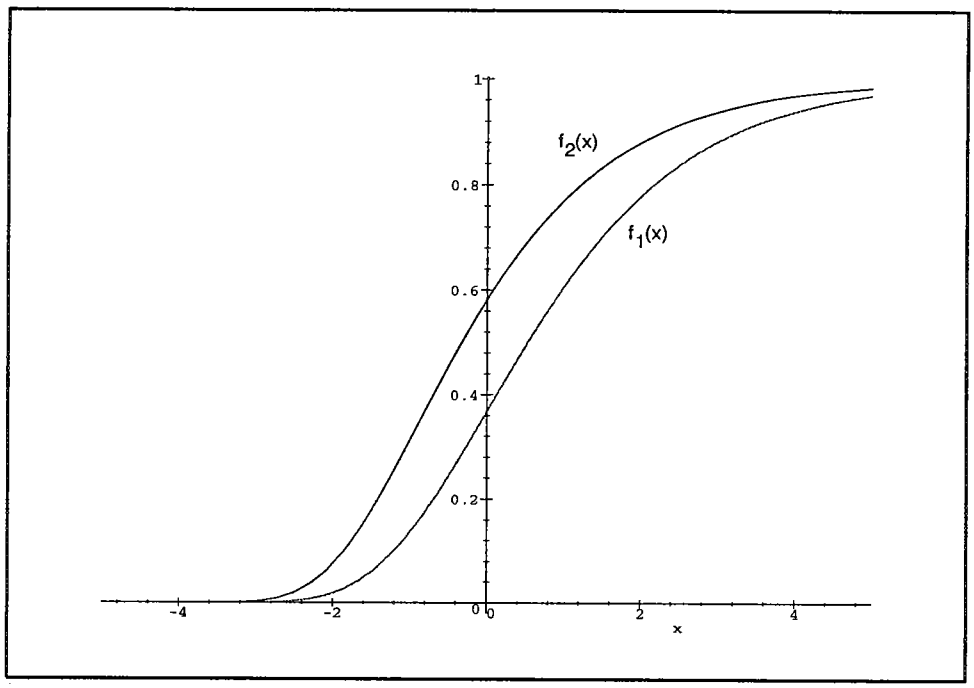

Figure 2. The extreme continuous distribution functions: $f_1(x) = e^{-2^{-x}}$ and $f_2(x) = 2^{-x}/(e^{2^{-x}} - 1)$.

As already mentioned in Section 1, the limiting distribution of the height resembles the double exponential distribution (i.e., $e^{-2^{-x}}$), which occurs in digital tree analyses more commonly than the standard extreme value distribution function $e^{-e^{-x}}$ [cf. Flajolet (1983), Devroye (1992), Pittel (1986)]. This is not a coincidence, as we try to explain below, and relate it to some other extreme distributions occurring in tries.

Let R_k be the number of candidate losers after k rounds of elimination. Further, suppose that the leaves are numbered $1, 2, \ldots, K$ from right to left. Observe that K, the number of leaves, is a random variable. Let C_j be the length of the path from the root to the common ancestor of the jth leaf and the leaf containing the loser. Clearly, $0 = C_1 \leq C_2 \leq \ldots \leq C_K$, and

$$H_n = 1 + \max\{j \geq 0 : R_j > 1\} = 1 + \max_{1 \leq j < K} \{C_j\} = 1 + C_{K-1}. \tag{6}$$

We can relate the above to some other properties of the standard trie (in the context of identification by coin tossing, even those that tossed heads continue the process until

one person is left in each leaf). In such a standard digital tree, let $\tilde{C}_{ij}$ be the length of a path from the root the common ancestor of the ith and jth leaves. We denote by $\tilde{D}_n(i)$ the length of a path from the root to the ith leaf, and by $\tilde{H}_n$ the longest root-to-leaf path in such a tree. It is an easy exercise [Jacquet and Szpankowski (1991)] to see that

$$\tilde{D}_n(i) = 1 + \max_{\substack{1\le j\le n\\ j\neq i}} \{\tilde{C}_{ij}\},$$

and

$$\tilde{H}_n = 1 + \max_{1\le i<j\le n} \{\tilde{C}_{ij}\}.$$

Observe that $\tilde{C}_{ij}$ is geometrically distributed.

Had $\tilde{C}_{ij}$ been i.i.d. random variables, then from the standard extreme distribution theory we would have immediately concluded that there exists a sequence $\tilde{a}_n = O(\log n)$ such that $\text{Prob}\{\tilde{H}_n - \tilde{a}_n \le k\}$ oscillates between $e^{-2^{-k-1}}$ and $e^{-2^{-k}}$. However, $\tilde{C}_{ij}$ are *not independent.* Nevertheless, for $\tilde{D}_n(i)$ and $\tilde{H}_n$, it can be proved that this is true [2] [cf. Jacquet and Szpankowski (1991), Pittel (1986)].

In view of this, it should not be a surprise that H_n in the incomplete trie as represented in (6) as an extreme statistic has an asymptotic distribution that resembles an extreme value distribution.

Acknowledgement. The first author wishes to thank Robert Smythe for some useful discussion.

[2] The depth $\tilde{D}_n(i)$ has stronger dependency than the height $\tilde{H}_n$, and in the case of unfair coins (i.e., non-equal probabilities of heads and tails in a coin toss), the depth has a *normal* limiting distribution!

References

1. Abramowitz, M. and Stegun, I., Eds. (1972). *Handbook of Mathematical Functions with Formulas, Graphs and Mathematical Tables.* Wiley, New York.
2. Aldous, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic.* Springer-Verlag, New York.
3. Anderson, C. (1970). Extreme value theory for a class of discrete distributions with applications to some stochastic processes. *J. Appl. Probability*, vol. 7, pp. 99–113.
4. Devroye, L. (1992). A study of trie-like structures under the density model. *Annals of Applied Probability*, vol. 2, pp. 402–434.
5. Flajolet, P. (1983). On the performance evaluation of extendible hashing and trie search. *Acta Informatica*, vol. 20, pp. 345–369.
6. Flajolet, P. (1988). Mathematical methods in the analysis of algorithms and data structures. In *Trends in Theoretical Computer Science*, pp. 225–304, Börger, E., Ed. Computer Science Press. Rockville, Maryland.
7. Galambos, H. (1987). *The Asymptotic Theory of Extreme Order Statistics.* Robert E. Krieger Publishing Company, Malabar, Florida.
8. Grabner, P. (1993). Searching for losers. *Random Structures and Algorithms*, vol. 4, pp. 99–110.
9. Jacquet, P. and Szpankowski, W. (1991). Analysis of tries with Markovian dependency. *IEEE Trans. Information Theory*, vol. 37, pp. 1470–1475.
10. Knuth, D. (1973). *The Art of Computer Programming, Vol. 3: Sorting and Searching.* Addison-Wesley, Reading, Massachusetts.
11. Kuipers, L. and Niederreiter, H. (1974). *Uniform Distribution of Sequences.* Wiley, New York.
12. Mahmoud, H., (1992). *Evolution of Random Search Trees.* Wiley, New York.
13. Mendelson, H. (1982). Analysis of extendible hashing. *IEEE Transactions on Software Engineering*, vol. 8, pp. 611–619.

14. Pittel, B. (1985). Asymptotical growth of a class of random trees. *Annals of Probability*, vol. 13, pp. 414–427.

15. Pittel, B. (1986). Paths in a random digital tree: Limiting distributions. *Adv. Appl. Prob.*, vol. 18, pp. 139–155.

16. Pittel, B. and Rubin, H. (1992). How many random questions are necessary to identify n distinct objects. *J. Combinatorial Theory, Ser. A*, vol. 55, pp. 292–312.

17. Prodinger, H. (1993). How to select a loser. *Discrete Math.*, vol. 120, pp. 149–159.

18. Rais, B., Jacquet, P. and Szpankowski, W. (1993). Limiting distribution for the depth in PATRICIA tries. *SIAM J. on Discrete Math.*, vol. 3, pp. 355–362.

19. Rényi, A. (1961). On random subsets of a finite set. *Mathematica (Cluj)*, vol. 3, pp. 355–362.

20. Szpankowski, W. (1987). Solution to a linear recurrence equation arising in the analysis of some algorithms. *SIAM J. Alg. Disc. Meth.*, vol. 8, pp. 233–250.