On the number of *t*-ary trees with a given path length

Gadiel Seroussi
Information Theory Research
HP Laboratories Palo Alto
HPL-2004-127
July 26, 2004*

E-mail: gadiel.seroussi@hp.com

binary trees, *t*-ary trees, path length, universal types

We show that the number of t-ary trees with path length equal to p is $t^{\frac{t h(t^{-1}) p}{\log_2 p}(1+o(1))}$ where $h(x) = -x\log_2 x-(1-x)\log_2(1-x)$ is the binary entropy function. Besides its intrinsic combinatorial interest, the question recently arose in the context of information theory, where the number of t-ary trees with path length p estimates the number of universal types, or, equivalently, the number of different possible Lempel-Ziv'78 dictionaries for sequences of length p over an alphabet of size t.

"Some of the most instructive applications of the mathematical theory of trees to the analysis of algorithms are connected with formulas for counting how many different trees there are of various kinds." D. E. Knuth, [7, p. 386].

On the number of t-ary trees with a given path length

Gadiel Seroussi*

July 23, 2004

Abstract

We show that the number of t-ary trees with path length equal to p is $t^{h(t^{-1})\frac{tp}{\log_2 p}(1+o(1))}$, where $h(x)=-x\log_2 x-(1-x)\log_2(1-x)$ is the binary entropy function. Besides its intrinsic combinatorial interest, the question recently arose in the context of information theory, where the number of t-ary trees with path length p estimates the number of universal types, or, equivalently, the number of different possible Lempel-Ziv'78 dictionaries for sequences of length p over an alphabet of size t.

Keywords: binary trees, t-ary trees, path length, universal types

"Some of the most instructive applications of the mathematical theory of trees to the analysis of algorithms are connected with formulas for counting how many different trees there are of various kinds." D. E. Knuth, [7, p. 386].

1 Introduction

Fix an integer $t \geq 2$. A t-ary tree T is defined recursively as either being empty or consisting of a *root node* r and the nodes of t disjoint, ordered, t-ary (sub-)trees $T_1, T_2, \ldots, T_t$, any number of which may be empty [7, Sec. 2.3.4.5]. When T_i is not empty, we say that there is an *edge* from r to the root r' of T_i, and that r' is a *child* of r. The total number of nodes of T is zero if T is empty, or $n_T = 1 + \sum_{i=1}^{t} n_{T_i}$ otherwise. A node of T is called a *leaf* if it has no children. The *depth* of a node $v \in T$ is defined as the number of edges traversed to get

*Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, USA; gadiel.seroussi@hp.com.

from the root r to v. We denote by $D_j^{(T)}$, $j \geq 0$, the number of nodes at depth j in T. The sequence $\{D_j^{(T)}\}$ is called the *profile* of T; we only consider finite trees, so $\{D_j^{(T)}\}$ has finite support. The *path length* of a non-empty tree T, denoted by p_T, is the sum of the depths of all the nodes in T, namely

$$p_T = \sum_{j \geq 1} j D_j^{(T)}$$

(the subscript T in n_T and p_T will be omitted when the tree being discussed is clear from the context). We call a t-ary tree with n nodes a $[t, n]$ *tree*. A $[t, n]$ tree with path length equal to p will be called a $[t, n, p]$ *tree*, and a t-ary tree with path length equal to p and an unspecified number of nodes will be referred to as a $[t, \cdot, p]$ *tree*.

Path length is an important global parameter of a tree that arises in various computational contexts, where it often relates to execution time [7, Sec. 2.3.4.5]. For example, when the access time of the data stored in a tree node is proportional to the depth of the node, the path length, after normalization, represents the average access time for a uniformly distributed random node of the tree.

Let $C_t(n)$ denote the number of $[t, n]$ trees, and $L_t(p)$ the number of $[t, \cdot, p]$ trees. It is well known [7, p. 589] that

$$C_t(n) = \frac{1}{(t-1)n+1} \binom{tn}{n}. \tag{1}$$

In the binary case ($t = 2$), these are the well known *Catalan numbers* that arise in many combinatorial contexts. The determination of $L_t(p)$, on the other hand, has remained elusive, even for $t = 2$. Consider the bivariate generating function $B(w, z)$ defined so that the coefficient of $w^p z^n$ in $B(w, z)$ counts the number of $[2, n, p]$ trees. $B(w, z)$ satisfies the functional equation [7, p. 595]

$$zB(w, wz)^2 = B(w, z) - 1.$$

However, solving this equation for the generating function $B(w, 1)$ of the numbers $L_2(p)$ appears quite challenging. Nevertheless, the equation, and others of similar structure, has been studied in the literature. In particular, the limiting distribution of the path length for a given number of nodes is related to the area under a Brownian excursion [11, 12, 13], which is also known as an *Airy distribution*. This distribution occurs in many combinatorial problems of theoretical and practical interest (cf. [4] and references therein).

These studies, however, have not yielded explicit asymptotic estimates for the numbers $L_t(p)$. The numbers recently arose in an information-theoretic context, in connection with the notion of *universal type* [9, 10], based on the incremental parsing of Ziv and Lempel (LZ78) [14]. When applied to a t-ary sequence, the LZ78 parsing produces a dictionary of strings that is best represented by a t-ary tree whose path length corresponds to the length of the sequence. Two sequences are said to be of the same universal type if they yield the same t-ary parsing tree. Sequences of the same universal type are, in a sense, statistically indistinguishable, as their empirical probability distributions of any finite order converge in the limit [9, 10]. Universal types generalize the notion underlying the classical *method of types*, which has lead to important theoretical results in information theory [3]. Of great interest in this context is the estimation of the number of different types for sequences of a given length p. For universal types, this translates to the number of different LZ78 dictionaries, or trees with a given path length, namely, $L_t(p)$.

Let $\lg x = \log_2 x$, and let $h(x) = -x \lg x - (1-x) \lg(1-x)$ denote the binary entropy function. The main result of this paper is the following asymptotic estimate of $L_t(p)$.

Theorem 1 *Let $\alpha = t\, h(t^{-1})$. Then, $L_t(p) = t^{\frac{\alpha p}{\lg p}(1+o(1))}$.*

The theorem is derived by proving matching upper and lower bounds on $L_t(p)$. The proof is presented in Section 2.

We remark that Knessl and Szpankowski [6] have recently applied the WKB heuristic [1] to obtain an asymptotic expansion of $\lg L_2(p)$ using tools of complex analysis. The heuristic makes certain assumptions on the form of asymptotic expansions, and is often considered a practically effective albeit non-rigorous method. The proofs in this paper, on the other hand, use mostly combinatorial arguments. The main term in the expansion of [6] is consistent with Theorem 1 for $t = 2$.

2 Proof of the main result

In the following lemma, we list some elementary properties of t-ary trees that will be referred to in the proof of Theorem 1. For a discussion of these properties, see [7, Sec. 2.3.4.5].[1]

Lemma 1 *(i) Let ℓ be a positive integer, and let T be a $[t, n, p]$ tree achieving minimal path length among all t-ary trees with ℓ leaves. Then,*

$$n = \ell - \left\lceil \frac{\ell - 1}{t - 1} \right\rceil \tag{2}$$

and the profile of T is given by

$$D_j^{(T)} = \begin{cases} t^j, & 0 \le j \le m-1, \\ \ell_1, & j = m, \\ 0, & j > m, \end{cases} \tag{3}$$

[1] A slight change of terminology is required: nodes of t-ary trees in our terminology correspond to *internal* nodes of *extended* t-ary trees in [7].

where

$$m = \lceil \log_t \ell \,\rceil, \tag{4}$$

and

$$\ell_1 = \ell - \left\lfloor \frac{t^m - \ell}{t-1} \right\rfloor . \tag{5}$$

In particular, all the leaves of T are either at depth m or $m-1$.

(ii) A $[t, n, p]$ tree with minimal path length satisfies

$$p = p_{\min} = \left(n + \frac{1}{t-1}\right)\mu - \frac{t(t^\mu - 1)}{(t-1)^2} = n \log_t n - O(n), \tag{6}$$

where $\mu = m$ whenever $n \not\equiv 2 \bmod t$, or $\mu = m+1$ otherwise, with m defined in (4) for the number of leaves, ℓ, of the tree. In particular, the tree of (i) satisfies (6) with $\mu = m$.

(iii) The number of nodes of a $[t, n, p]$ tree satisfies

$$n \leq \frac{p}{\log_t p - O(\log \log p)} = \frac{p}{\log_t p}(1 + o(1)). \tag{7}$$

(iv) The maximal path length of a $[t, n]$ tree is achieved by a tree in which each internal node has exactly one child (and, hence, there is exactly one leaf). The path length of such a tree is

$$p_{\max} = \frac{n(n-1)}{2} . \tag{8}$$

(v) There is a $[t, n, p]$ tree for each p in the range $p_{\min} \leq p \leq p_{\max}$.

Proof. Items (i),(ii), and (iv) follow immediately from the discussion in [7, Sec. 2.3.4.5]. For convenience in the proof of Theorem 1, we characterize, in Item (i), trees with minimal path length for a given number of *leaves*, while the discussion in [7] does so for trees with a given number of *nodes*. The two characterizations coincide, except for values of n such that $n \equiv\ 2 \bmod t$, which never occur in (2). In that case, a tree with $n-1$ nodes would have the same number of leaves and a shorter path length. A tree that has minimal path length for its number of leaves, on the other hand, always has minimal path length also for its number of nodes (given in (2)).

Item (iii) follows from (ii) by solving for n in an equation of the form $p = n \log_t n - O(n)$. Solutions of equations of this form are related to the *Lambert W* function, a detailed discussion of which can be found in [2].

To prove the claim of Item (v), consider a $[t, n, p]$ tree T such that $D_j^{(T)} > 1$ for some integer j. Let j_T be the largest such integer for the tree T. It follows from these assumptions that T must have nodes u and v at depth j_T, such that u is a leaf, $v \neq u$, and v has at most one child. Thus, we can transform T by deleting u and adding a child to v, and obtain a $[t, n, p+1]$ tree. Starting with a $[t, n, p_{\min}]$ tree, the transformation can be applied repeatedly to obtain a sequence of trees with consecutive values of p, as long as the transformed tree has at least two leaves. When this condition ceases to hold, we have the tree of Item (iv), which has path length $p_{\max}$. $\square$

We will also rely on an estimate of $C_t(n)$, which is derived from (1) using Stirling's approximation to express a binomial coefficient in terms of the binary entropy function (see, e.g., [8, Ch. 10]). Specifically, for positive real numbers c_1 and c_2, which depend on t but not on n, we have

$$c_1 n^{-\frac{3}{2}} 2^{\alpha n} \leq C_t(n) \leq c_2 n^{-\frac{3}{2}} 2^{\alpha n}, \tag{9}$$

where, as before, $\alpha = t\, h(t^{-1})$.

Proof of Theorem 1.

(a) Upper bound: $L_t(p) \leq t^{\frac{\alpha p}{\lg p}(1+o(1))}$.

Let T be a $[t, n, p]$ tree. T can be completely determined by specifying n and the the index of T in an exhaustive enumeration of all $[t, n]$ trees. Thus, given p, but without other prior assumptions on n, T can be described in $K = \lg p + \lg C_t(n) + O(1)$ bits. Using the estimate (9), we can write $K = \alpha n - \frac{3}{2} \lg n + \lg p + O(1)$, and, applying (7), it follows that $K \leq K_{\max} = \frac{\alpha p}{\log_t p}(1 + o(1))$. Thus, every $[t, \cdot, p]$ tree can be completely specified using at most $K_{\max}$ bits, and, hence, we must have $L_t(p) \leq 2^{K_{\max}}$, from which the desired upper bound follows. The asymptotic error term $o(1)$ in the upper bound is, by (7), of the form $O(\log \log p / \log p)$.

(b) Lower bound: $L_t(p) \geq t^{\frac{\alpha p}{\lg p}(1+o(1))}$.

We prove the lower bound by constructing a sufficiently large class of $[t, \cdot, p]$ trees.

Let ℓ be a positive integer. We start with a t-ary tree T with ℓ leaves and shortest possible path length, as characterized in Lemma 1(i). Let q be the integer satisfying

$$C_t(q-1) < \ell - 1 \leq C_t(q), \tag{10}$$

and let $\tau_1, \tau_2, \ldots, \tau_{\ell-1}$ be the first $\ell - 1$ $[t, q]$ trees when these trees are arranged in increasing order of path length. Additionally, let τ_{F} be a tree with βq nodes, for some positive constant β to be specified later. Finally, let π be a permutation on $\{1, 2, \ldots, \ell - 1\}$. We construct a tree T_π by attaching the trees $\tau_1, \tau_2, \ldots, \tau_{\ell-1}$ and τ_{F} to the leaves of T, so that the i-th leaf (taken in some fixed order) becomes the root of a copy of $\tau_{\pi(i)}$, $1 \leq i < \ell$, with τ_{F} attached to the last leaf of T, which is assumed to be at (the maximal) depth m. The construction is illustrated in Figure 1.

Next, we compute the path length, p, of T_π. By Lemma 1(i), all the leaves of T are either at depth $m = \lceil \log_t \ell \rceil$ or at depth $m-1$. Assume τ_i, $1 \leq i \leq \ell - 1$, is attached to a leaf of

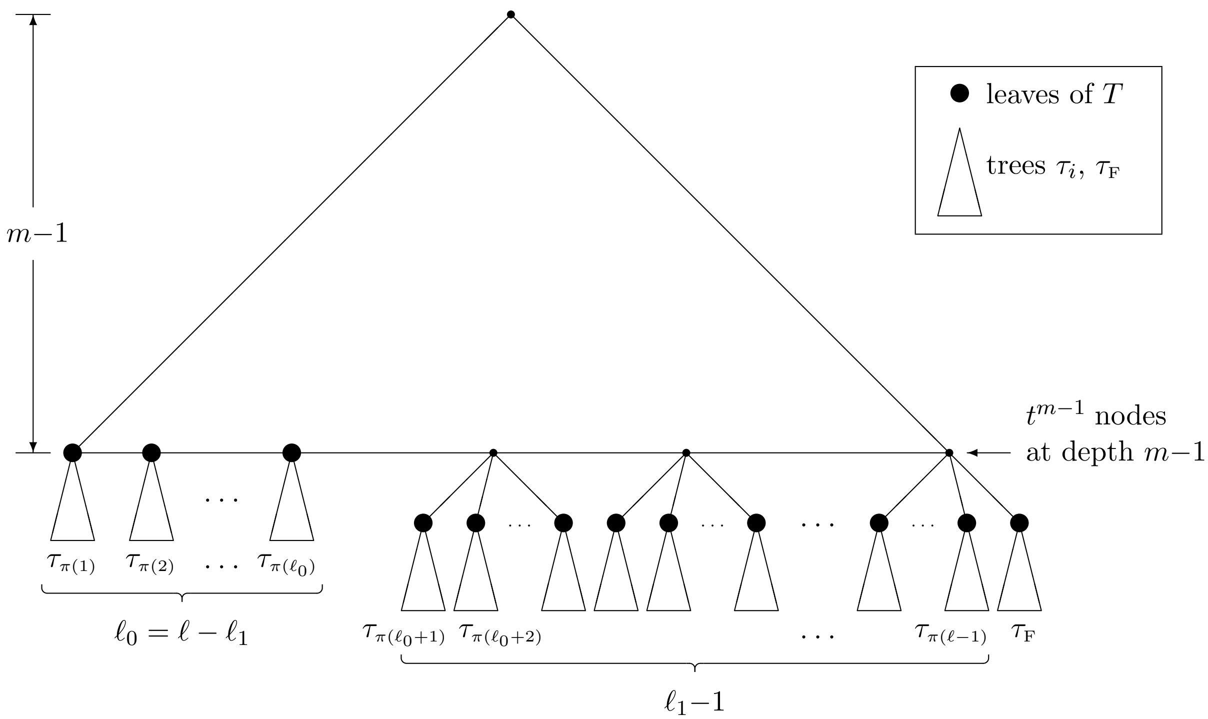

Figure 1: Tree T_π

depth $m-1+\epsilon_i$, $\epsilon_i \in \{0,1\}$, of T. The contribution of τ_i (excluding its root) to p is

$$p_i = \sum_{j\geq 1}(m-1+\epsilon_i+j)D_j^{(\tau_i)} = (m-1+\epsilon_i)\sum_{j\geq 1} D_j^{(\tau_i)} + \sum_{j\geq 1} jD_j^{(\tau_i)} = (m-1+\epsilon_i)(q-1)+\nu_i,$$

where ν_i denotes the path length of τ_i. Similarly, denoting by ν_{F} the path length of τ_{F}, the contribution of this tree to p is $p_{\mathrm{F}} = m(\beta q-1)+\nu_{\mathrm{F}}$. Considering also the contribution of T according to its profile (3), we obtain

$$p = \sum_{i=1}^{\ell-1}(m-1+\epsilon_i)(q-1) + \sum_{i=1}^{\ell-1}\nu_i + m(\beta q-1) + \nu_{\mathrm{F}} + \sum_{j=1}^{m-1} jt^j + \ell_1 m\,. \tag{11}$$

Further, observing that $\sum_{i=1}^{\ell-1}\epsilon_i = \ell_1$, and defining $\overline{\nu} = (\ell-1)^{-1}\sum_{i=1}^{\ell-1}\nu_i$, we obtain

$$p = \left((\ell-1)(m-1)+\ell_1\right)(q-1) + (\ell-1)\overline{\nu} + m(\beta q-1) + \nu_{\mathrm{F}} + \sum_{j=1}^{m-1} jt^j + \ell_1 m\,. \tag{12}$$

Recall that the τ_i were selected preferring shorter path lengths, so their average path length $\overline{\nu}$ is at most as large as the average path length of *all* $[t,q]$ trees. The latter average is known to be $O(q^{3/2})$ (this follows from the results of [5]; see also [7, Sec. 2.3.4.5] for $t=2$). Observe also that, by (4),(9), and (10), we have

$$q = \frac{\lg t}{\alpha} m + O(\log m). \tag{13}$$

Recalling now that $n_{\tau_F} = \beta q$, and, hence, $\nu_{\text{F}} = O(q^2)$, it follows, after standard algebraic manipulations, that (12) can be rewritten as

$$p = \frac{\lg t}{\alpha} m^2 \ell + O(m^{3/2}\ell). \tag{14}$$

It also follows from (12) that p is independent of the choice of permutation π. Moreover, by construction, each permutation π defines a different tree T_π. Therefore, we have

$$L_t(p) \geq (\ell-1)!\,. \tag{15}$$

Now, from (14), (4) and (15), using Stirling's approximation, we obtain

$$\frac{\log_t L_t(p)}{p} \geq \frac{\log_t((\ell-1)!)}{p} = \frac{\ell \log_t \ell - O(\ell)}{p} = \frac{m\ell - O(\ell)}{\alpha^{-1}\,(\lg t)\, m^2\ell + O(m^{3/2}\ell)}\,. \tag{16}$$

Also, from (14) and (4) we have $\lg p = \lg \ell + O(\log m) = m \lg t + O(\log m)$. Combining with (16), and simplifying asymptotic expressions, we obtain

$$\frac{\log_t L_t(p)}{p} \geq \frac{\alpha}{\lg p}(1 - o(1)), \tag{17}$$

from which the desired lower bound follows. The $o(1)$ term in (17) is $O((\log p)^{-\frac{1}{2}})$.

The above construction yields large classes of trees of path length p for a sparse sequence of values of p, controlled by the parameter ℓ. Next, we show how the gaps in the sparse sequence can be filled, yielding constructions, and validating the lower bound, for all (sufficiently large) integer values of p. In the following discussion, to emphasize the dependency of m, ℓ_1, q, and p on ℓ, we use the notations $m(\ell), \ell_1(\ell), q(\ell)$, and $p(\ell)$, respectively. Also, for any such function $f(\ell)$, we denote by Δf the difference $f(\ell+1) - f(\ell)$. We start by estimating Δp.

Assume first that ℓ is such that $\Delta q = 0$ and $\Delta m = 0$. Then, substituting $\ell+1$ for ℓ in (12), and subtracting the original equation, we obtain

$$\Delta p = (m - 1 + \Delta\ell_1)(q-1) + \nu_\ell + \Delta\ell_1\, m. \tag{18}$$

It follows from (5) that, with m fixed, we have $0 \leq \Delta\ell_1 \leq 2$. Also, by (8), we have $\nu_\ell < \frac{1}{2}q^2$. Hence, recalling (13), it follows from (18) that

$$\Delta p < \left(\frac{\alpha}{\lg t} + \frac{1}{2}\right) q^2 + O(q \log q). \tag{19}$$

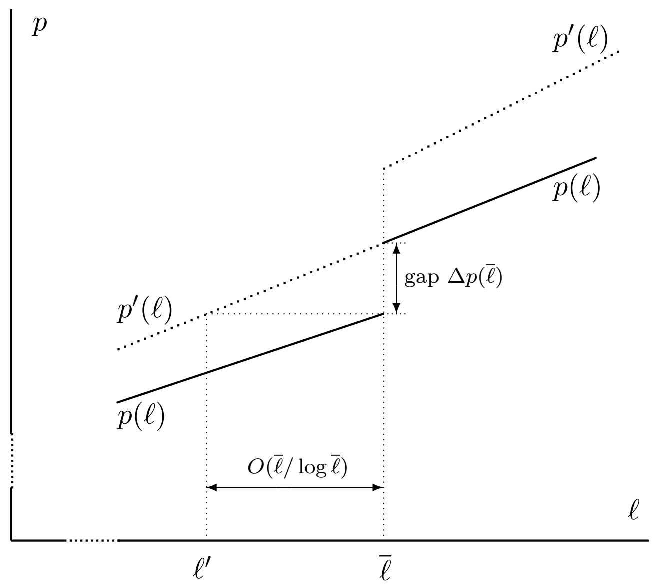

Figure 2: Bridging the gap in q-breaks

Notice that, in (12), with all other parameters of the construction staying fixed, any increment in ν_{F} produces an identical change in p. By Lemma 1(v), by an appropriate evolution of τ_{F}, we can make ν_{F} assume any value in the range $(\nu_{\mathrm{F}})_{\min} \leq \nu_{\mathrm{F}} \leq (\nu_{\mathrm{F}})_{\max}$, where $(\nu_{\mathrm{F}})_{\min} = O(\beta q \log_t q)$, and $(\nu_{\mathrm{F}})_{\max} = \frac{1}{2}\beta q(\beta q - 1)$. Choosing $\beta > \sqrt{2\alpha(\lg t)^{-1} + 1}$, this range of ν_{F} will make p span the gap between $p(\ell)$ and $p(\ell+1)$ as estimated in (19), for all sufficiently large ℓ satisfying the conditions of this case. Still, the variation in the value of p is asymptotically negligible and does not affect the validity of (17).

If $\Delta m = 1$, we must have $\ell = \ell_1(\ell) = t^m$, and $\ell_1(\ell+1) = 2$. In this case, using (12) again, we obtain

$$
\begin{aligned}
\Delta p &= (\ell m + 2)(q-1) + \beta q - 1 + \nu_\ell + m\ell + 2(m+1) - ((\ell-1)(m-1) + \ell)(q-1) - \ell m \\
&= (m+1)(q+1) + \nu_\ell + \beta q - 1,
\end{aligned}
$$

which admits the same asymptotic upper bound as Δp in (19). Thus, the gap between $p(\ell)$ and $p(\ell+1)$ is filled also in this case by tuning the structure of τ_{F}.

The above method cannot be applied directly when $\Delta q = 1$. We call a value of ℓ such that $q(\ell+1) = q(\ell)+1$ a *q-break*. At a q-break, Δp is exponential in q, and a tree τ_{F} of polynomial size cannot compensate for such a gap. However, we observe that the construction of T_π, and its analysis in (12)–(17) would also be valid if we chose $q' = q+1$, instead of q, as the size of the trees τ_i. This choice would produce a different sequence of path length values $p'(\ell)$, which would also satisfy (17) and would validate the lower bound of the theorem. It follows from (12) that $p'(\ell) > p(\ell)$. Equivalently, for any given (sufficiently large) value ℓ, there exists an integer $\ell' < \ell$ such that $p'(\ell') \leq p(\ell) \leq p'(\ell'+1)$.

Consider a q-break $\bar{\ell}$. To construct large classes of trees for all values of p, proceed as follows (refer to Figure 2): use the original sequence of values $p(\ell)$, filling the gaps as described above, until $\ell = \bar{\ell}$. At that point, find the largest integer ℓ' such that $p'(\ell') \leq p(\bar{\ell})$, and "backtrack" to $\ell = \ell'$. Continue with the sequence $p'(\ell)$, $\ell = \ell', \ell' + 1, \ldots$, filling the gaps accordingly. Notice that $q'(\ell)$ to the left of $\bar{\ell}$ is the same as $q(\ell)$ to the right of that point. Thus, $p'(\ell)$ continues "smoothly" (i.e., with gaps Δp as in (19)) into $p(\ell)$ at $\ell = \bar{\ell}$. The process now rejoins the sequence $p(\ell)$ as before, until the next q-break point. By (12), since the function $m(\ell)$ remains the same for both p and p', we have, asymptotically, $\ell' \approx (1 - 1/q)\bar{\ell} \approx \bar{\ell} - c_3\bar{\ell}/\log\bar{\ell}$, for some positive constant c_3. Thus, for sufficiently large $\bar{\ell}$, although the difference between ℓ' and $\bar{\ell}$ is negligible with respect to $\bar{\ell}$, ℓ' is guaranteed to fall properly between q-breaks, and the number of sequence points $p'(\ell)$ used between ℓ' and $\bar{\ell}$ is unbounded. $\square$

Acknowledgment. Thanks to Wojciech Spankowski and Alfredo Viola for very useful discussions.

References

[1] C. Bender and S. Orszag, *Advanced Mathematical Methods for Scientists and Engineers*, Mc-Graw Hill, 1978. 3

[2] R. M. Corless, G. H. Gonnet, D. E. G. Hare, D. J. Jeffrey, and D. E. Knuth, *On the Lambert W function*, Adv. Comput. Math., 5 (1996), pp. 329–359. 4

[3] I. Csiszár, *The method of types*, IEEE Trans. Inform. Theory, IT-44 (1998), pp. 2505–2523. 3

[4] P. Flajolet and G. Louchard, *Analytic variations on the Airy distribution*, Algorithmica, 31 (2001), pp. 361–377. 2

[5] P. Flajolet and A. M. Odlyzko, *The average height of binary trees and other simple trees*, J. Comput. Syst. Sci., 25 (1982), pp. 171–213. 7

[6] C. Knessl and W. Szpankowski, *Enumeration of binary trees, Lempel-Ziv78 parsings, and universal types.* Preprint, 2004. 3

[7] D. E. Knuth, *The Art of Computer Programming. Fundamental Algorithms*, vol. 1, Addison-Wesley, Reading, MA, third ed., 1997. 1, 2, 3, 4, 7

[8] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error Correcting Codes*, North-Holland Publishing Co., Amsterdam, 1983. 5

[9] ——, *Universal types and simulation of individual sequences*, in LATIN 2004: Theoretical Informatics, M. Farach-Colton, ed., vol. LNCS 2976, Berlin, 2004, Springer-Verlag, pp. 312–321. 2, 3

[10] G. Seroussi, *On universal types*, in Proc. IEEE International Symp. Inform. Theory, Chicago, 2004, p. 10. Full paper in preparation. 2, 3

[11] L. Takács, *A Bernoulli excursion and its various applications*, Adv. Appl. Prob., 23 (1991), pp. 557–585. 2

[12] ——, *On a probability problem connected with railway traffic*, J. Applied Mathematics and Stochastic Analysis, 4 (1991), pp. 1–27. 2

[13] ——, *Conditional limit theorems for branching processes*, J. Applied Mathematics and Stochastic Analysis, 4 (1991), pp. 263–292. 2

[14] J. Ziv and A. Lempel, *Compression of individual sequences via variable-rate coding*, IEEE Trans. Inform. Theory, IT-24 (1978), pp. 530–536. 2