Theoretical Computer Science, 144, 1995

ASYMPTOTIC BEHAVIOR OF THE LEMPEL-ZIV PARSING SCHEME AND DIGITAL SEARCH TREES*

January 30, 1995

Philippe Jacquet[†]
INRIA
Rocquencourt
78153 Le Chesnay Cedex
France
jacquet@blagny.inria.fr

Wojciech Szpankowski[‡]
Department of Computer Science
Purdue University
W. Lafayette, IN 47907
U.S.A.
spa@cs.purdue.edu

Abstract

The Lempel-Ziv parsing scheme finds a wide range of applications, most notably in data compression and algorithms on words. It partitions a sequence of length n into variable phrases such that a new phrase is the shortest substring not seen in the past as a phrase. The parameter of interest is the number M_n of phrases that one can construct from a sequence of length n. In this paper, for the memoryless source with *unequal* probabilities of symbols generation we derive the limiting distribution of M_n which turns out to be normal. This proves a long standing open problem. In fact, to obtain this result we solved another open problem, namely, that of establishing the limiting distribution of the internal path length in a digital search tree. The latter is a consequence of an asymptotic solution of a multiplicative differential-functional equation often arising in the analysis of algorithms on words. Interestingly enough, our findings are proved by a combination of probabilistic techniques such as renewal equation and uniform integrability, and analytical techniques such as Mellin transform, differential-functional equations, de-Poissonization, and so forth. In concluding remarks we indicate a possibility of extending our results to Markovian models.

[*] Preliminary version of this paper was presented in *26-th ACM Symposium on the Theory of Computing*, Montreal, 1994.

[†] This work was partly supported by the Esprit Basic Research Action No. 7141 (Alcom II).

[‡] This research was primarily done while the author was visiting INRIA in Rocquencourt, France. The author wishes to thank INRIA (projects ALGO, MEVAL and REFLECS) for a generous support. In addition, support was provided by NSF Grants NCR-9206315, CCR-9201078 and INT-8912631, and in part by NATO Collaborative Grant 0057/89.

1. INTRODUCTION

The primary motivation for this work is the desire to understand the asymptotic behavior of the fundamental parsing algorithm on words due to Lempel and Ziv [27]. It partitions a word into phrases (blocks) of variable sizes such that a new block is the shortest subword not seen in the past as a phrase. For example, the string 110010100010001000 is parsed into (1)(10)(0)(101)(00)(01)(000)(100).

These parsing algorithms play a crucial rôle in universal data compression schemes and their numerous applications such as efficient transmission of data (cf. [26, 27]), estimation of entropy (cf. [25]), discriminating between information sources (cf. [7]), test of randomness, estimating the statistical model of individual sequences (cf. [16]), and so forth. The parameters of interest for these applications are: the number of phrases, the number of phrases of a given size, the size of a phrase, the length of a sequence built from a given number of phrases, etc. But, by all means the most important parameter is the number of phrases: This parameter is used to obtain the compression ratio in a universal data compression (cf. [3]), while its distribution is needed in the analysis of other parameters of the Lempel-Ziv scheme (e.g., redundancy rate [21], length of a phrase [14], and so forth).

In this paper, we study the distribution of the number of phrases M_n constructed form a word of a fixed length n in a probabilistic framework. We assume that the word is generated by a probabilistic memoryless binary source (extension to finite non-binary alphabet is simple). That is: *symbols are generated in an independent manner with "0" and "1" occurring respectively with probability p and $q = 1 - p$.* If $p = q = 0.5$, then such a probabilistic model will be further called the *symmetric Bernoulli* model; otherwise we refer to the *asymmetric Bernoulli* model.

Aldous and Shields [1] attested that obtaining the limiting distribution of the number of phrases is a difficult problem. They solved it only for the *symmetric* Bernoulli model. The authors of [1] wrote: "It is natural to conjecture that asymptotic normality holds for a larger class of processes ... . But in view of the difficulty of even the simplest case (i.e., the fair coin-tossing case we treat here) we are not optimistic about finding a general result. We believe the difficulty of our normality result is intrinsic ... ." We settle the conjecture of [1] in the affirmative for the *asymmetric* Bernoulli model, and in concluding remarks we indicate a possibility of extending our findings to Markovian models. Actually, we do more, and provide solutions to some other problems that have been open up-to-date, namely: the limiting distribution for internal path lengths in digital trees (cf. [1, 11, 15]), the number of parsings of given length built from a fixed number of words (cf. [7]), and probabilistic behavior of the Lempel-Ziv code redundancy (cf. [16, 21]).

All of these problems are solved in a uniform manner by a combination of probabilistic and analytical methods. We apply the *renewal equation* (cf. [2]) to reduce the problem of finding the number of phrases in the Lempel-Ziv scheme to another problem on digital search trees, namely that of finding the limiting distribution of the internal path length in a digital search tree built from a *fixed* number of independent words.

The reader is referred to [11, 15] for discussion and definition of the digital trees. However, for the reader's convenience we show in Figure 1 the digital search tree associated with the word mentioned at the beginning of this section. In particular, the root of the tree is empty (i.e., we start parsing with an empty phrase). All other phrases of the Lempel-Ziv parsing algorithm are stored in internal nodes. When a new phrase is created, the search starts at the root and proceeds down the tree as directed by the input symbols exactly in the same manner as in the digital tree construction, that is, symbol "0" in the input string means move to the right and "1" means proceed to the left. The search is completed when a branch is taken from an existing tree node to a new node that has not been visited before. Then, the edge and the new node are added to the tree. The phrases created in such a way are stored directly into the nodes of the tree. In passing, we note that for a word of fixed length, n, the size of the associated digital search tree is *random*, and this fact gives a new twist to the analysis of digital trees (cf. also [14]).

Second-order properties, such as limiting distributions and large deviation results of the Lempel-Ziv scheme, have been scarcely discussed in the past with a notable exception of the work of Aldous and Shields [1] who studied the symmetric model. Recently, Louchard and Szpankowski [14] obtained the limiting distribution of a randomly selected phrase length in the Lempel-Ziv scheme.

On the other hand, digital search trees (built from a *fixed* number of independent words!) have been quite thoroughly investigated in the past (cf. [4, 5, 10, 11, 13, 14, 22]). In particular, Knuth [11], and Flajolet and Sedgewick [4] introduced analytical methods in the analysis of digital search trees. This was continued by Flajolet and Richmond [5], Louchard [13], and Szpankowski [22]. None of these papers, however, deals with second order properties of the internal path length in digital search trees, which is the main object of our study. Only very recently, Kirschenhofer, Prodinger and Szpankowski [10] obtained an asymptotic expression for the variance of the internal path length in the *symmetric* Bernoulli model (in fact, this allowed to close the gap in the Aldous and Shields analysis by yielding the leading term in the variance of the number of phrases in the Lempel-Ziv parsing scheme). The authors of [10], however, did not extend their results to the asymmetric model. We not only provide such an