A Combinatorial Approach to Golomb Trees

Mordecai J. Golin
Hong Kong University

September 22, 1997

[summary by Philippe Dumas and Michèle Soria]

Abstract

Given a set of weights, the problem of finding the binary-tree with minimum weighted external path length is very well understood. It can be solved using Huffman encoding. The problem of finding such an (infinite) tree, with minimal path-length for an infinite set of weights, is not nearly as well studied. Twenty years ago Gallager and Van Voorhis described such trees for the case in which the infinite set of weights is a geometric series. These trees are now known as Golomb trees. Here, the problem is handled with a combinatorial approach.

Let F be an alphabet equipped with a probability distribution. The problem is to encode the alphabet into a language on the binary alphabet $\{0,1\}$, in such a way that the codeword length mean value is minimal. Such a code is said to be optimal. For a finite alphabet, the problem is known to be solved by Huffman encoding [3]: a tree is built in which each leaf is associated to a (prefix-free) codeword. Hence the path-length of the tree is the codeword length.

When the alphabet is infinite, the problem is solved only for the geometric case, that is the case when the set F is an infinite sequence $a_0, a_1, \ldots$ and the probability of letter a_i occurring in a message is $(1-p)p^i$ with $0 < p < 1$. For example, suppose that we have a string of x's and y's in which each character occurs independently of every other one, x's occurring with probability p, and y's occurring with probability $1-p$. Every infinite message can be uniquely written as the concatenation of words $a_i = x^i y$, each a_i occurring with probability $(1-p)p^i$. The geometric case was studied by Gallager and Van Voorhis [1], who exhibited an optimal tree. Their technique is to construct the Huffman tree for each finite case $\{a_0, a_1, \ldots, a_n\}$, and take the limit in some sense when n goes to infinity. They show that the infinite limit tree is an optimal tree.

Golin's approach is based on combinatorial transformations of trees, which preserve optimality. For his purpose, the important combinatorial feature of the tree is not the whole topological structure, but only its profile, that is the number of internal nodes at each level. He extends the problem to d-ary trees. Considering the number $p \in]0,1[$ and the integer $d \geq 2$, there is a unique positive integer m which satisfies

$$p^m + p^{m+1} \leq 1 < p^m + p^{m-1}.$$

Define α_k to be the unique positive root of equation

$$1 - \alpha = \alpha^{k(d-1)}(1 - \alpha^d),$$

with the particular case $\alpha_0 = 0$. Using this notation, Golin's result can be stated in the following way.

Theorem 1. *If $\alpha_{m-1} < p < \alpha_m$, then there is a unique optimal tree profile: the first levels from the root are $1, d, d^2, \ldots$ as long as the powers of d are smaller than m, and all the next levels are equal to m.*

If $p = \alpha_m$, then there is an infinite set of optimal tree profiles. They all begin as in the previous case, but after the transition each level is either m or $m+1$.

Notice that this result extends the work of Gallager and Van Voorhis, who did not study the uniqueness of the solution.

The key point is that the geometric character of the distribution entails that the width of an optimal tree at each level is bounded. The proof is valid only for the geometric distribution, since it strongly uses the fact that the shift from a level to the next one translates into a multiplication of the weights p_i's. It does not extend to other types of distributions.

Bibliography

[1] Gallager (Robert G.) and Van Voorhis (David C.). – Optimal Source Codes for Geometrically Distributed Integer Alphabets. *IEEE Transactions on Information Theory*, March 1975, pp. 228–230.

[2] Golin (Mordecai J.). – A Combinatorial Approach to Golomb Forests. – September 1997. Preprint.

[3] Huffman (D. A.). – A Method for the Construction of Minimum-Redundancy Codes. *Proceedings of the IRE*, n° 40, September 1952, pp. 1098–1101.