Distribution Converter using Prefix Trees

Hiroyoshi Morita, Kingo Kobayashi, and Mamoru Hoshi

The Graduate School of Information Systems, University of Electro-Communications, Chofu, Tokyo 182, Japan

Abstract — **We consider a problem to encode a source sequence into codewords using a prefix code and to decode them using another prefix code. The average number of the output symbols has been evaluated for an *i.i.d.* source with a finite alphabet.**

I. Notation and Terminology

A D-ary tree is either an external node or an internal node connected to an ordered D tuple of D-ary trees for an integer $D \geq 2$. Let $\mathcal{T}_D$ be the set of all D-ary trees. The path of a node in $T \in \mathcal{T}_D$ is defined as a sequence of links from the root to the node. By labelling links from each internal node, from left to right, with integers 0 to $D-1$, each path can be represented as a D-ary string, called a path string. The height of T, denoted by $h(T)$, is defined as the maximum path length of nodes in T. If a D-ary tree T has K external nodes ($K = 1 (\mathrm{mod}\ D-1)$), then T is called a D-ary tree of size K. Suppose that each of external nodes of T is also labelled, from left to right, with symbols in $\mathcal{A}_K = \{a_1, a_2, \ldots, a_K\}$. In what follows, we will think of an external node and its label as being the same. The set of D-ary trees of size K is denoted by $\mathcal{T}_D^{(K)}$.

A probability distribution is called D-*adic* if each of the probabilities is equal to D^{-n} for some n. Given a D-ary tree $T \in \mathcal{T}_D^{(K)}$, the set of external nodes of T, that is, $\mathcal{A}_K$, uniquely determines a D-adic probability distribution $P_T = \{D^{-\pi(\alpha)}, \alpha \in \mathcal{A}_K\}$ where $\pi(\alpha)$ denotes the path length of a node α. An *i.i.d.* source is called D-adic associated with a D-ary tree T if P_T is the probability distribution of the source.

Let $L(T)$ be $\sum_{\alpha \in \mathcal{A}_K} D^{-\pi(\alpha)} \pi(\alpha)$. Moreover, for a probability distribution over $\mathcal{A}_K$ and $T \in \mathcal{T}_D^{(K)}$, $L(P,T)$ is defined as $\sum_{\alpha \in \mathcal{A}_K} P(\alpha)\pi(\alpha)$.

II. Tree Conversion

For two integers $K_1 > 1$ and $K_2 > 1$, let us take a pair of two D-ary trees $T_1 \in \mathcal{T}_D^{(K_1)}$ and $T_2 \in \mathcal{T}_D^{(K_2)}$, respectively. Note that these trees are not always the same. For a pair (T_1, T_2), Algorithm DC transforms an infinite sequence $\boldsymbol{x} = x_1 x_2 \cdots$ over $\mathcal{A}_{K_1}$ into an infinite sequence $\boldsymbol{z} = z_1 z_2 \cdots$ over $\mathcal{A}_{K_2}$:

Algorithm DC

1) Each symbol $x_i = \alpha \in \mathcal{A}_{K_1}$ of $\boldsymbol{x}$ is mapped into the corresponding external node labelled with α in T_1. Then output its path string from the root τ_1 to α as the codeword $W(x_i, T_1)$. Let $\boldsymbol{y} = y_1 y_2 \ldots$ be an infinite D-ary sequence concatenating $W(x_i, T_1)$ $(i = 1, 2, \ldots)$.

2) Starting at the root τ_2 of T_2 and the beginning of $\boldsymbol{y}$, travel through T_2 as directed by $\boldsymbol{y}$, and each time an external node is encountered, output the codeword associated with the label of the external node and restart at the root.

This work was partly supported by MESSC, under grant number C2-08680354

III. Main Results

Suppose that $\boldsymbol{X} = X_1 X_2 \ldots$ be a random K_1-ary sequence generated by an *i.i.d.* source with probability distribution P where X_i takes a value $\alpha \in \mathcal{A}_{K_1}$ with the probability $P(\alpha)$. By replacing $\boldsymbol{x}$ with $\boldsymbol{X}$ in Algorithm DC, each of $\boldsymbol{Y}$ and $\boldsymbol{Z}$, the counter parts of $\boldsymbol{y}$ and $\boldsymbol{z}$, also becomes a sequence of random variables. However, they are not necessarily independent nor identically distributed.

We are particularly interested in the stochastic behavior of $\boldsymbol{Z}$ as a function of $\boldsymbol{X}$ and the expected length of the subsequence of $\boldsymbol{Z}$ corresponding to a finite subsequence of $\boldsymbol{X}$. However, to avoid an unnecessary difficulty, we consider a fixed-length subsequence $\boldsymbol{Y}^n$ of the intermediate D-ary sequence $\boldsymbol{Y}$, instead of a fixed-length subsequence of $\boldsymbol{X}$.

Given $T \in \mathcal{T}_D$, let us parse $\boldsymbol{Y}^n$ into codewords as in step 2) in Algorithm DC although it may terminate at some internal node of T. With respect to $\boldsymbol{Y}^n$, we denote the number of the parsed codewords by $N(\boldsymbol{Y}^n, T)$. In case of D-adic sources, we obtained the following theorems:

Theorem 1 *Given D-ary trees T_1 and T_2, Algorithm DC converts a D-adic source associated with T_1 into another D-adic source associated with T_2.*

Theorem 2 *Given T_1, T_2, and a D-adic source associated with T_1,*

$$\lim_{n \to \infty} \frac{N(\boldsymbol{Y}^n, T_2)}{N(\boldsymbol{Y}^n, T_1)} = \frac{L(T_1)}{L(T_2)} \qquad \textit{with probability one.}$$

To analyze more general cases, first we will assign a specific conditional probability distribution with each internal node of T_1, which is induced by the probability distribution of external nodes. Next, we introduce a graph product $T_1 \times T_2$ of trees T_1 and T_2 as follows:

- For each tree, make a link from a parent p of each external node e to the root. And assign the same symbol as the label of the link from p to e with that link. The modified T_1 and T_2 are denoted by T_1' and T_2', respectively.
- The node set of $T_1 \times T_2$ is the set of pairs (a, b) where a and b are internal nodes of T_1 and T_2, respectively.
- There exists a link from (a, b) to (c, d) in $T_1 \times T_2$ if and only if there are two links from a to c in T_1' and from b to d in T_2' which are labelled with the same symbol.

Let $A_{T_1 \times T_2}$ be the adjacency matrix of $T_1 \times T_2$. Moreover, let $P_{T_1 \times T_2}$ be the transition probability matrix obtained by replacing 1-elements of $A_{T_1 \times T_2}$ with the conditional probabilities assigned with internal nodes of T_1.

Conjecture 1 *Given T_1, T_2, and an i.i.d. source with alphabet $\mathcal{A}_{K_1}$,*

$$\lim_{n \to \infty} \frac{N(\boldsymbol{Y}^n, T_2)}{N(\boldsymbol{Y}^n, T_1)} = \frac{\Pr(\tau_2)}{\Pr(\tau_1)} \qquad \textit{with probability one}$$

where $\Pr(\tau_i)$ $(i = 1, 2)$ *is the marginal probability corresponding to the root τ_i of the stationary distribution of $P_{T_1 \times T_2}$.*