Huffman-like optimal codes and search codes for infinite alphabets

Akiko Kato[1]
Dept. of Mathematical Engineering and Information Physics,
University of Tokyo, Tokyo 113-8656, JAPAN
Email: akiko@misojiro.t.u-tokyo.ac.jp

We consider any probability distribution $P = \{p_i\}_{i=0}^{\infty}$ on the nonnegative integers $\mathcal{X} = Z \cup \{0\}$, and assume the monotonicity $p_0 \geq p_1 \geq p_2 \geq \cdots$ and the positivity $p_i > 0$ for all $i \in \mathcal{X}$.

For any integer $D \geq 2$, let $\mathcal{D} = \{0, 1, \cdots, D-1\}$, and let $\mathcal{D}^*$ be the set of finite sequences of $\mathcal{D}$. A subset of $\mathcal{D}^*$ is said to be a *prefix set* if no element in it is a prefix of another. A mapping $C : \mathcal{X} \to \mathcal{D}^*$ is called a *D-ary prefix code for $\mathcal{X}$* if the image $\mathrm{Im}(C)$ is a prefix set. $C(i)$ is the codeword for $i \in \mathcal{X}$ in C. The *expected codeword length* $l(C)$ of a code C is defined as $\sum_{i=0}^{\infty} p_i |C(i)|$, where $|\cdot|$ denotes the codeword length. Let $H_D(P)$ be the entropy of P with the logarithm of base D: $H_D(P) = -\sum_{i=0}^{\infty} p_i \log_D p_i$. It is well known that if $H_D(P)$ converges then $H_D(P) \leq l(C)$ for any D-ary prefix code C, and there exists a D-ary prefix code C such that $l(C) < H_D(P) + 1$.

For any probability distribution on a *finite* set, a D-ary prefix code for the finite set is said to be *optimal* if no other D-ary prefix code for the set has less expected codeword length than it has, and Huffman algorithm gives an optimal code [2]. If the entropy converges, optimal codes for probability distributions on $\mathcal{X}$ are defined as above. However, for divergent entropies, it gets meaningless (since every prefix code has infinite expected codeword length), and thus we introduce a new definition of optimal codes. Let C be a D-ary prefix code for $\mathcal{X}$, and let $T(C)$ be the code tree of C. A *node* refers to either an internal node or a leaf. For any node v of the code tree $T(C)$ of C, let

$$E_v = \{i \in \mathcal{X} \mid \text{the leaf corresponding to } i \text{ is a successor of } v\}.$$

We adopt the convention that v is a successor of v itself. A finite partition $\mathcal{A} = \{A^{(1)}, \cdots, A^{(N)}\}$ of $\mathcal{X}$ is said to be *compatible with C* if there exists a set $\{v^{(1)}, \cdots, v^{(N)}\}$ of nodes of $T(C)$ so that $A^{(1)} = E_{v^{(1)}}, \cdots, A^{(N)} = E_{v^{(N)}}$. Note that no $v^{(i)}$ is a successor of $v^{(j)}$, $1 \leq i < j \leq n$. For a finite partition $\mathcal{A}$ compatible with C, $T_{\mathcal{A}}(C)$ is defined as the subtree of $T(C)$ whose leaves are $v^{(1)}, \cdots, v^{(N)}$, and clearly, $T_{\mathcal{A}}(C)$ is a code tree of a D-ary prefix code $C_{\mathcal{A}}$ for $\mathcal{A}$: $T_{\mathcal{A}}(C) = T(C_{\mathcal{A}})$. Also, a probability distribution $P_{\mathcal{A}} = \{p(A^{(j)})\}_{j=1}^{N}$ on $\mathcal{A}$ is defined as $p(A^{(j)}) = \sum_{i \in A^{(j)}} p_i$ for all $j = 1, 2, \ldots, N$. An infinite sequence $(\mathcal{A}_1, \mathcal{A}_2, \cdots)$ of finite partitions of $\mathcal{X}$ is said to be *complete with C* if: (a) $\mathcal{A}_k$ is compatible with C for each $k \in \{1, 2, \ldots\}$; (b) $\mathcal{A}_{k+1}$ is a refinement of $\mathcal{A}_k$ for each $k \in \{1, 2, \ldots\}$; and (c) For every $i \in \mathcal{X}$ there exists a unique $k \geq 1$ such that $\{i\} \in \mathcal{A}_k \setminus \mathcal{A}_{k-1}$, where we set $\mathcal{A}_0 = \{\mathcal{X}\}$ for conveniences' sake. Now we give a new definition of optimal codes. A D-ary prefix code C for the probability distribution P on $\mathcal{X}$ is said to be *optimal* if there exists an infinite sequence $(\mathcal{A}_1, \mathcal{A}_2, \ldots)$ of partitions complete with C such that the code $C_{\mathcal{A}_k}$ is optimal for $P_{\mathcal{A}_k}$ for every $k = 1, 2, \ldots$. It was shown in [3] that the new definition is equivalent to the previous one for convergent entropies.

[1] This work is supported in part by a JSPS Research Fellowship for Young Scientists.

For convergent entropies, it was shown in [5] that optimal D-ary prefix codes exists for any probability distribution on $\mathcal{X}$. We strengthen their result as follows. A D-ary prefix code C is said to be *exhaustive* if the code tree is D-ary exhaustive. Let $<$ be the ordinary order on $\mathcal{X}$, and let $\prec$ be the lexicographical order on the set $\mathcal{D}^*$. A D-ary prefix code $C : \mathcal{X} \longrightarrow \mathcal{D}^*$ is called *alphabetic* if $C(a) \prec C(b)$ whenever $a < b$. A *search code* is an alphabetic exhaustive code.

Theorem 1 *There exists an optimal D-ary prefix code for any probability distribution on the nonnegative integers among the search codes, whether the entropy converges or diverges.*

Theorem 2 below is a converse of Theorem 1 in a sense.

Theorem 2 *Given any probability distribution P on the nonnegative integers, for any optimal D-ary prefix code, there exists a D-ary search code such that all the individual codeword lengths coincide with those in the optimal code, whether the entropy converges or diverges.*

Recently, it is shown in [1] that there is no general algorithm to construct optimal codes for $\mathcal{X}$. However, for each $i \in \mathcal{X}$, a finite-step procedure exists that finds all the candidates for the individual codeword length in all the optimal codes.

Theorem 3 *For each $i \in \mathcal{X}$, any individual codeword length in any optimal code for $\mathcal{X}$ belongs to the finite set*

$$\mathcal{L}(i) = \left\{ \left| C_{\mathrm{opt}/s}^{((m+1)+j_m(D-1))}(i) \right| \;\middle|\; 0 \leq j_m \leq k_m - 1 \right\},$$

where $m \in \mathcal{X}$ is the least integer such that $m \geq i$ and $m \equiv -1 \pmod{D-1}$, k_m is the least positive integer satisfying $p_m \geq \sum_{i=m+1+k_m(D-1)}^{\infty} p_i$, and $C_{\mathrm{opt}/s}^{((m+1)+j_m(D-1))}$ is an optimal code for a finite ordered set $\mathcal{X}_{(m+1)/(D-1)+j_m}$; $\mathcal{X}_n = \{0, 1, \ldots, n(D-1)-1, \alpha = \{i \mid i \geq n(D-1)\}\}$ is the finite ordered set whose order $<$ is defined as $0 < 1 < \ldots < n(D-1)-1 < \alpha$, with the probability distribution $P^{(n)} = \{p_0^{(n)}, \ldots, p_{n(D-1)-1}^{(n)}, p_\alpha^{(n)}\}$ satisfying $p_i^{(n)} = p_i$ for $i \in \{0, 1, \ldots, n(D-1)-1\}$ and $p_\alpha^{(n)} = \sum_{i=n(D-1)}^{\infty} p_i$.

References

[1] Chow T., and M. Golin, "Convergence and construction of minimum-cost infinite trees," *preprint,* October 1997.

[2] D. A. Huffman, "A method for the construction of minimum-redundancy codes," *Proceedings of IRE,* vol. 40, no. 10, pp. 1098–1101, 1952.

[3] A. Kato, T. S. Han, and H. Nagaoka, "Huffman coding with an infinite alphabet," *IEEE Trans. Inform. Theory,* vol. IT-42, No. 3, pp. 977–984, 1996.

[4] A. Kato, "Huffman-like optimal prefix codes and search codes for infinite alphabets," *submitted to IEEE Trans. Inform. Theory,* January 1997.

[5] T. Linder, V. Tarokh, and K. Zeger, "Existence of optimal prefix codes for infinite source alphabets," *IEEE Trans. Inform. Theory,* vol. IT-43, no. 6, pp. 2026–2028, 1997.