Block Arithmetic Coding for Markov Sources

Charles G. Boncelet Jr.
Department of Electrical Engineering
University of Delaware
Newark, DE 19716

1 Introduction

In a recent paper submitted to IEEE Transactions on Information Theory [1], we introduced BAC. BAC is a variable to fixed block coder in that the input is parsed into variable length substrings which are encoded with fixed length output strings. Assume the input is taken from an alphabet with m symbols and the codebook has K codewords. With each input symbol, the encoder splits the set of codewords into m disjoint, nonempty subsets. The recursion continues until fewer than m codewords remain. One of these is transmitted, and the encoder reinitialized. The encoding process is described in Figure 1.

```
K = K_S;
A = 1;
while ((l = getinput()) ≠ EOF){
        Compute K_1, K_2, ..., K_m;
        A = A + Σ_{j=1}^{l-1} K_j;
        K = K_l;
        if (K < m){
                Output code(A);
                A = 1;
                K = K_S;
        }
}
doeof(A, K);
```

Figure 1: Basic BAC Encoder.

The K_j satisfy the following: 1) $K_j = 1 + L_j * (m-1)$ for $L = 0, 1, 2, \ldots$ and 2) $\sum_{j=1}^{m} K_j = K$. The first condition assures two things: that $K_j > 0$ and that K_j equals the number in a complete and proper set. Let $N(K)$ denote the expected number of codewords encoded with K codewords. For i.i.d. inputs, $N(K)$ satisfies

$$N(K) = 1 + \sum_{j=1}^{m} p_j N(K_j) \tag{1}$$

The principal question is how are the K_j determined. In [1], we offered several methods. Firstly, K_j can be chosen optimally by dynamic programming. Secondly, K_j can be chosen by an arithmetic coding heuristic: $K_j = [p_j K]$, where $[p_j K]$ is a quantization such that conditions 1) and 2) above are satisfied. Thirdly, if we imagine that we can ignore the necessity that K_j be integer, then take $K_j = p_j K$. This solution results in a hypothetical entropy coder.

Denote the expected number number of input symbols encoded with K codewords by $N_o(K)$, $N_h(K)$, and $N_e(K)$, respectively. Then, for i.i.d. inputs, for all K and for some constant C, we showed the following:

$$\frac{\log K}{H(X)} = N_e(K) \geq N_o(K) \geq N_h(K) \geq \frac{\log K}{H(X)} - C \tag{2}$$

For Markov sources, let $p(i|l) = \Pr(x_j = a_i | x_{j-1} = a_l)$. Then the recursion for $N(K)$ splits into two parts. The first is for the first input symbol; the second is for all other input symbols:

$$\begin{aligned} N(K) &= 1 + \sum_{l=1}^{m} p_l N(K_l|l) && (3) \\ N(K|l) &= 1 + \sum_{i=1}^{m} p(i|l) N(K_i|i), && (4) \\ & && (5) \end{aligned}$$

where $N(K|l)$ is the number of input symbols encoded using K codewords given that the current input is a_l. ($N(K)$ does not include the current input, a_l.) The heuristic is as follows: Choose $K_i = [p(i|l)K]$. The optimal K_j can again be chosen by dynamic programming. We can state the following theorem:

Theorem: If the Markov chain is time-invariant, ergodic, and symmetrical in the following way,

$$H(X|l) = -\sum_{j=1}^{m} p(j|l) \log p(j|l) \tag{6}$$

is independent of l, then, for all l,

$$N_h(K|l) \geq \frac{\log K}{H(X|l)} - C \tag{7}$$

and

$$N(K) \geq \frac{\log K}{H(X|l)} - C' \tag{8}$$

Proof (sketched): The symmetry condition allows that the entropy solution, $K_j = p_j K$, satisfies (4). The proof that $N_h(K|l)$ satisfies (7) follows almost identically from [1]. (8) follows directly from (7).

If the encoder starts anew with each block, then some loss of efficiency occurs since the first symbol of each block is encoded with its stationary probability, nots its Markov one conditioned on the previous symbol. However, encoding blocks separately yields greater resistance to channel errors.

To get a feeling for the magnitude of C, we computed $N(K)$ for two situations. The first is a binary symmetric Markov chain with crossover probability equal to 0.05. For 65536 codewords (16 bits), $N_e = 53.4$, $N_o = 51.0$, and $N_h = 50.5$. The second is a binary asymmetric Markov chain with crossover probabilities equal to 0.05 and 0.50. Again for 16 bit codewords, $N_e = 45.3$, $N_o = 45.2$, and $N_h = 44.4$.

References

[1] C. G. Boncelet Jr. Block arithmetic coding for source compression. *IEEE Trans. on Info. Theory*, 1993. Submitted Sept. 1991.