Arithmetic Coding: Example 2

Suppose we have an alphabet $\{1, 2, 3\}$ and a sequence $X_1X_2X_3X_4$ of four variables drawn from the alphabet. Assume a first order Markov model with conditional probability:

$$P(X_{j+1} \mid X_j) = \frac{1}{8}\begin{bmatrix} 3 & 6 & 2 \\ 3 & 1 & 4 \\ 2 & 1 & 2 \end{bmatrix}$$

and marginal probability of the first symbol

$$P(X_1) = \frac{1}{4}\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}$$

For this case, the arithmetic coding "induction equations" are:

$$\begin{aligned} l_{k+1} &= l_k + (u_k - l_k)\, F(pred(i_{k+1}) \mid i_k) \\ u_{k+1} &= l_k + (u_k - l_k)\, F(i_{k+1} \mid i_k) \end{aligned}$$

where the cumulative distribution of the conditional is:

$$F(X_{j+1} \mid X_j) = \frac{1}{8}\begin{bmatrix} 3 & 6 & 2 \\ 6 & 7 & 6 \\ 8 & 8 & 8 \end{bmatrix}$$

and cumulative of the marginal distribution of the first symbol

$$F(X_1) = \frac{1}{4}\begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}$$

Suppose the sequence is $\vec{x} = (2, 3, 1, 2)$. What is the codeword $C(\vec{x})$? Since $n = 4$, the encoder needs to calculate l_k and u_k.

k	X_k	l_k	u_k
1	2	$\frac{1}{4}$	$\frac{1}{2}$
2	3	$\frac{15}{32}$	$\frac{16}{32}$
3	1	$\frac{60}{128}$	$\frac{61}{128}$
4	2	$\frac{483}{1024}$	$\frac{486}{1024}$

Thus the tag is

$$T(\vec{x}) = \frac{l_4 + u_4}{2} = \frac{969}{2048}$$

which in binary is .01111001001.

Note that it is very easy to make mistakes when doing the l_k and u_k distributions by hand. One way to check if you made a mistake is to compute the probability of the sequence, and verify that it is the same as the difference $u_n - l_n$.

$$p(\vec{x}) = p(2,3,1,2) = u_4 - l_4 = \frac{3}{1024}$$

You should doublecheck that this probability is correct:

$$p(2,3,1,2) \;=\; p(X_1 = 2)\, p(X_2 = 3|X_1 = 2)\, p(X_3 = 1|X_2 = 3)\, p(X_4 = 2|X_3 = 1) = \frac{1}{4}\cdot\frac{1}{8}\cdot\frac{1}{4}\cdot\frac{3}{8} = \frac{3}{1024}$$

Next, how many bits in the codeword?

$$\lambda(\vec{x}) = \lceil \log \frac{2}{p(\vec{x})} \rceil = \lceil \log \frac{2048}{3} \rceil = 10 \text{ bits}$$

Hence, truncating the tag to ten bits gives:

$$C(\vec{x}) = 0111100100\ .$$

Decoder

Let's now consider the decoder. Suppose the decoder is given that $n = 4$ and it is given the alphabet and the probabilities. The decoder needs to infer what is the sequence $\vec{x} = (i_1, i_2, i_3, i_4)$ such that $C(\vec{x}) = 011110010$. Writing this number in binary (.01111001) we get that the tag truncated to nine bits is $\frac{121}{256}$.
The decoder wants to find a sequence l_k and u_k such that $\frac{121}{256} \in [l_k, u_k)$ for all k.

$$\begin{aligned}
\text{Try } i_1 = 1, \quad l_1 &= \frac{1}{4} = \frac{64}{256} \\
u_1 &= \frac{1}{2} = \frac{128}{256} \qquad \text{Yes, since } \frac{64}{256} < \frac{121}{256} < \frac{128}{256} \\
\text{Try } i_2 = 1, \quad l_2 &= \frac{1}{4} + \frac{1}{4}\cdot 0 = \frac{1}{4} \\
u_2 &= \frac{1}{4} + \frac{1}{4}\cdot\frac{3}{8} = \frac{9}{32} = \frac{72}{256} \qquad \text{, No, since } 72 < 121 \\
\text{Try } i_2 = 2, \quad u_2 &= \frac{1}{4} + \frac{1}{4}\cdot\frac{6}{8} = \frac{14}{32} = \frac{112}{256} \qquad \text{, No, since } 112 < 121
\end{aligned}$$

Thus, $i_2 = 3$ since this is the only remaining possibility.

We then continue on using similar reasoning to get i_3 and i_4.