COMPUTER SCIENCE TRIPOS Part II – 2023 – Paper 9

7 Information Theory (rkh23)

You are tasked with compressing a continuous textual data stream into a stream of binary codewords. You know the stream's alphabet contains three letters ('A', 'B', 'C') and three numbers ('1', '2', '3'), and that the stream alternates between letter and number. You have been provided with a sample of 100 consecutive characters and no more as follows:

```
A1A3A1C2B2
A3B2A2B2C2
C3C3C3C3A1
B1B2B1C1C1
A1C3A1A2A1
A1A2A1A2B3
A2B3A1A3B3
C2A3A3C2A1
C2C2A3A3C3
C2A3A3B2A3
```

(*a*) For each of the following source models find an encoding using Huffman codes and compare the average encoded character length to the entropy in bits per character.

(*i*) A pure character source, ignoring the alternating nature of the characters. [3 marks]

(*ii*) A mixture of two distinct sources, one for letters and one for numbers. [4 marks]

(*iii*) A stream of two-character symbols ('A1', 'A2', etc.). [5 marks]

(*b*) Explain conceptually the trend in entropy values you found in part (*a*). Which do you think is closer to the true entropy and why? [4 marks]

(*c*) Discuss whether it would be advantageous to model the stream as a stream of four-character symbols and apply Huffman coding. [2 marks]

(*d*) Give two advantages to using arithmetic coding instead of Huffman coding for this problem. [2 marks]