COMPUTER SCIENCE TRIPOS Part II – 2026 – Paper 8

8 Information Theory (rkh23)

This question concerns arithmetic coding of messages.

(*a*) You are designing an arithmetic code for a source that emits messages of random length N, where the expected length is $L = E[N]$. You implement a termination symbol ϕ such that $P(\phi) = \epsilon$, scaling the probabilities of the data symbols by $(1 - \epsilon)$.

(*i*) Explain why adding a termination symbol in this way increases the overall message length by more than the cost of emitting the termination symbol itself. [1 mark]

(*ii*) Derive an expression for the total increase in Shannon Information for a message of length N caused by the addition of the terminating character. Explain how this relates to the number of additional bits needed to encode the message. [4 marks]

(*iii*) Derive the optimal value of ϵ that minimises the total overhead for a message of average length L. You may use the approximation $\ln(1 - \alpha) \approx -\alpha$ for small α without proof. [4 marks]

(*b*) An alternative approach to termination symbols is to use a Large Language Model (LLM) as the probability model for the arithmetic coder. LLMs use a special token EoS (End of Sequence) to terminate generation. This EoS symbol can be used as the termination symbol in the coder.

(*i*) Give three disadvantages of using an LLM as the probability model for an arithmetic coder. [3 marks]

(*ii*) Explain how an LLM used in this way can result in shorter encoded messages. [3 marks]

(*iii*) Assume you have an LLM trained on long sequences ($L \sim 1000$) but you need to send messages of various lengths N, including some for which $N \leq 400$. You find that the LLM is assigning $P(\text{EoS}) = 10^{-7}$ for these shorter messages, which makes the termination symbol costly. Your colleague suggests assigning $P(\text{EoS}) = 0.5$ (neutral) for messages with $N \leq 400$. Calculate the message length at which this neutral approach adds a larger overhead than the default LLM approach. [5 marks]