Introduction to Information Theory, Fall 2019

Practice problem set #4

You do **not** have to hand in these exercises, they are for your practice only.

1. **Using the wrong symbol code** Supose we are given two probability distributions P and Q on the same set $\mathcal{A}_X$. We know that there exists a uniquely decodable symbol code with codeword lengths $l(C(x)) = \lceil \log \frac{1}{Q(x)} \rceil$ and it has expected length satisfying

$$H(Q) \leqslant L(C, Q) < H(Q) + 1.$$

Show that if we use this code for the 'wrong' distribution P the expected codelength will satisfy

$$H(P) + D(P \,\|\, Q) \leqslant L(C, P) < H(P) + D(P \,\|\, Q) + 1$$

where we recall that the relative entropy is defined as

$$D(P \,\|\, Q) = \sum_{x \in \mathcal{A}_X} P(x) \log(\frac{P(x)}{Q(x)}).$$

2. **Worst case analysis of Lempel-Ziv compression** This exercise is a bit harder, and it is optional. In the lecture we showed that the LZ algorithm performs well on average. That means that it necessarily makes some messages longer, but in this problem you will show that it will not make them too much longer (which is of course a nice property for a compression algorithm). To be precise, you will show that the worst case rate is $R \leqslant 1 + \mathcal{O}(\frac{1}{\log(N)})$. For simplicity we will assume that our set of symbols has only two elements.

   (a) Consider the string x_λ which is constructed by enumerating all phrases up to length λ, ordered from short to long, and concatenating them. For instance, if we enumerate all phrases up to length 2 we have $\{A, B, AA, AB, BA, BB\}$ and x_2 would be $ABAAABBABB$. Argue that x_λ gives

$$c_\lambda = \sum_{k=1}^{\lambda} 2^k = 2^{\lambda+1} - 2$$

   phrases in the LZ encoding (hint: geometric series) and that x_λ has length

$$N_\lambda = \sum_{k=1}^{\lambda} k2^k = (\lambda - 1)2^{\lambda+1} + 2$$

   (hint: induction for the second equality) and hence

$$c_\lambda \leqslant \frac{N_\lambda}{\lambda - 1}$$

   for $\lambda \geqslant 1$.

   (b) Argue that c_λ is the worst case number of phrases for strings of length N_λ in the LZ encoding.

(c) For the rest of the exercise, we will denote by N the string length, and by c (as a function of N) the worst case number of phrases for a message of length N in the LZ encoding. Argue that if $N_\lambda \leqslant N < N_{\lambda+1}$, then c satisfies

$$c \leqslant c_\lambda + \frac{N - N_\lambda}{\lambda + 1}.$$

(d) Using (a) and (c), show that c satisfies

$$c \leqslant \frac{N}{\lambda - 1}$$

for $\lambda > 1$ where λ satisfies

$$\lambda \geqslant \log(c) - 2.$$

(e) Deduce from (d) that

$$c \leqslant \frac{N}{\log(c) - 3}$$

and show that this implies that

$$c \leqslant \mathcal{O}(\frac{N}{\log(N)})$$

(hint: look at a similar argument in the lecture).

(f) Recall from the lecture that the length of the encoding l is bounded by

$$l \leqslant c \log(c) + 2c$$

and use this to show that

$$l \leqslant N + 5c \leqslant N + \mathcal{O}(\frac{N}{\log(N)}).$$

Conclude that the rate $R = \frac{l}{N}$ is bounded by

$$R \leqslant 1 + \mathcal{O}(\frac{1}{\log(N)}).$$