Homework 4

10-704 Information Processing and Learning

Instructor: Aarti Singh

The HW is worth 40 pts and is **due on April 20 at noon**. Hand in to: Michelle Martin GHC 8001. If she is not around, note down the time on your HW sheet and slide it under her door.

1. [14 pts] **Minimum Description Length**
   In this problem you will use the idea of minimum description length to select the best model from the class of dyadic density trees. Akin to dyadic decision trees, dyadic density trees for a one-dimensional distribution split the domain into intervals of finer and finer granularity by splitting at the midpoint of the coarser intervals. However, the difference is that density trees assign the fraction of points in a leaf to be the "label" of the leaf.

   (a) [7 pts] Consider Q_γ as the class of all density trees of depth γ. We can encode the data using a specific dyadic density tree model $q \in Q_\gamma$ by the shannon information content or -ve log loss $L_q(x^n) = \log 1/q(x^n)$. Now we also need to encode the decision tree model i.e. specify $L(q)$. To encode a decision tree model, you need to i) encode the tree structure (hint: prefix codes) and ii) encode the value at each leaf node. Propose a strategy for encoding the decision tree model.

   (b) [3 pts] State the two-stage optimization procedure for selecting the best model in Q_γ.

   (c) [4 pts] Propose an efficient bottom-up strategy to solve the two-stage MDL optimization for trees.

2. [12 pts] **Channel Capacity**

   (a) [5 pts] Let C_1 and C_2 be the capacities of two DMC's with transition matrices P_1 and P_2, respectively, and let C be the capacity of the DMC with transition matrix P_1P_2. Prove that $C \leq \min(C_1; C_2)$.

   (b) [7 pts] The Z channel has binary input and output alphabets and transition probabilities $p(y|x)$ given by the following table:

   $$P = \begin{bmatrix} 1 & 0 \\ 1/2 & 1/2 \end{bmatrix}$$

   Find the capacity of the Z channel and the maximizing input probability distribution.

3. [14 pts] **Rate-Distortion**

(a) [6 pts] Suppose we want to compress a sequence x^n of n independent draws from a Gaussian random variable with zero mean and variance σ^2 while incuring a squared error distortion of D per symbol. Lets do an intuitive argument to justify that the minimum number of bit needed to describe the sequence up to distortion D a.k.a. the rate distortion function is

$$R(D) = \frac{1}{2}\log\frac{\sigma^2}{D} \quad \text{if} \quad 0 \leq \mathrm{D} \leq \sigma^2.$$

i. What is the radius of the sphere in which the sequence x^n is expected to lie?

ii. Since a squared distortion of D is allowed, sequences in a radius of $\sqrt{nD}$ can be mapped to one codeword. What is the minimum number of such codewords needed to cover the sphere of all sequences of length n?

iii. What is the fewest number of bits needed to describe the sequence to distortion D?

Remark: This is a sphere-covering argument akin to the sphere-packing argument needed for capacity of Gaussian channel.

(b) [8 pts] Suppose we want to compress K independent Gaussian random variables with zero mean and variances $\{\sigma_k^2\}_{k=1}^K$ while incuring a total squared error distortion of D. Compute the rate-distortion function and argue that it is equivalent to describing the random variables with large variances, but using no bits to describe the random variables with small variances.
Remark: This is essentially a reverse water-filling argument, inverse of what we did for distributing power amongst K independent Gaussian channels - there we allocated less power to channels with large variances and more power to channels with small variances.