Introduction to Information and Coding Theory July 16, 2019
(Prof. Permuter Haim, Mr. Oron Sabag and Mr. Ben Marinberg)

Final Exam - Moed B
Total time for the exam: 3 hours!

Important: For **True / False** questions, copy the statement to your notebook and write clearly true or false. You should prove the statement if true, and provide counterexample otherwise.

1) **Entropy and differential entropy of continuous random variable (28 Points)**: Let U be a uniform random variable on the interval $[0, \alpha]$, and let f_U be its density.
   a) (4 points) Draw the density f_U.
   b) (4 points) Differential entropy of U is defined as
   $$h(U) = \int_{-\infty}^{+\infty} -f_U(u) \log f_U(u) du.$$
   Compute the differential entropy of U.
   c) (4 points) The quantized version of U is given by X_N, where N is the number of the quantizations. Specifically, define $\Delta = \frac{\alpha}{N}$, and then
   $$X_N = i, \quad \text{if} \quad i\Delta \leq U < (i+1)\Delta.$$
   What is the PMF (probability mass function) of X_N, draw it and compute the entropy $H(X_N)$ as a function of N.
   d) (4 points) What is $\lim_{n\to\infty} H(X_N)$.
   e) (4 points) **True or False:** The entropy of a continuous random variable is infinite. (Explain your answer.)
   f) (4 points) What is $\lim_{n\to\infty}(H(X_N) + \log \Delta)$.
   g) (4 points) Suggest an algorithm to estimate differential entropy using entropy.
2) **Time - varying BSC (24 Points)** Consider a time-varying memoryless BSC: the probability for flipping a bit at time i is equal to p_i for $i = 1, \ldots, n$ as described in Fig. 1. The channel is memoryless and without feedback as assumed in class.

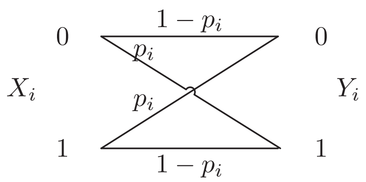

Fig. 1: Time-varying BSC.

   a) (6 points) Calculate the capacity that is given by $\max_{p(x_1,\ldots,x_n)} \frac{1}{n} I(X^n; Y^n)$.
   b) (6 points) Consider the time-invariant version of this channel with a normalized average parameter, i.e. $\bar{p} = \frac{1}{n}\sum_{i=1}^{n} p_i$. Calculate the capacity that is given by $\max_{p(x_1,\ldots,x_n)} \frac{1}{n} I(X^n; Y^n)$.
   c) (6 points) **True or False:** For a fixed $p(x)$ mutual information $I(X;Y)$ is convex in $p(y|x)$.
   d) (6 points) Is the time-varying or the time-invariant version has a greater capacity? Prove your answer.

3) **Neural network initialization (23 Points)**: Consider some neural network with L layers, where the first layer is input , denoted as $\mathbf{x}$ and last layer L is output $a(\mathbf{x})$. All activation functions are the sigmoid function. The network sizes are $[N_1, N_2, .., N_L]$. Assume that the initial weights are i.i.d. from some distribution $\mathcal{D}$, with an expectation and variance μ_w, σ_w.
   Reminder- following the lecture's notation, z_i^l is the input to the sigmoid of the i'th neuron of the l'th layer. The question is mostly on z_i^2 which is a linear function of the input $\mathbf{x}$.
   a) (6 Points) Assume input vector of N_1 ones, i.e. $\mathbf{x} = [\underbrace{1, 1, .., 1}_{N_1}]$. Calculate mean and variance of $z_i^2, \quad \forall i = 1, 2, .., N_2$, express your answers using μ_w, σ_w.
   b) (10 Points) What are the best values for the mean and variance of the activation inputs $z_i^2, \quad \forall i = 1, 2, .., N_2$, explain your answer.
   c) (7 Points) A student in the course suggested to initialize the weight according to $\mathcal{N}(0, \frac{1}{\sqrt{N_1}})$. Is he correct? Does it achieve the requirements you suggested?

4) **Linear regression (30 Points)**: Given training set $\{(x_i, y_i)\}_{i=1}^{N}$ where $(x_i, y_i) \in \mathbb{R}^2$.
   First, we use a linear regression method to model this data. To test our linear regressor, we choose at random some data records to be a training set, and choose at random some of the remaining records to be a test set. Now let us increase the training set size gradually.

a) (7 points) As the training set size increases, what do you expect will happen with the mean training error? explain your answer.

b) (7 points) As the training set size increases, what do you expect will happen with the mean test error? explain your answer.

Now we have prior knowledge of our dataset's distribution $y_i \sim \mathcal{N}(\log(wx_i), 1)$.

c) (10 points) Suppose you decide to do a maximum likelihood estimation of w. You figure out that w should satisfy one of the following equations. Which one? why?

i) $\sum_i x_i \log(wx_i) = \sum_i x_i y_i \log(wx_i)$

ii) $\sum_i x_i y_i = \sum_i x_i y_i \log(wx_i)$

iii) $\sum_i x_i y_i = \sum_i x_i \log(wx_i)$

iv) $\sum_i y_i = \sum_i \log(wx_i)$

d) (6 points) Provide a pseudocode (or a matlab code) for estimating w.

Good Luck!