Introduction to Information and Coding Theory — June 23, 2019

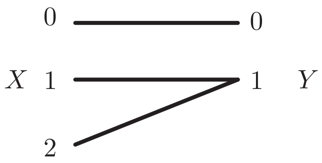
(Prof. Permuter Haim, Mr. Oron Sabag and Mr. Ben Marinberg)

Final Exam - Moed A

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) **Optimal outputs distribution (33 Points)**: Consider the following deterministic channel:



   a) (4 Points) Find the capacity .
   b) (5 Points) **True/False** The optimal input distribution $P(x)$ unique (there is only one such distribution).
   c) (6 Points) **True/False** The optimal output distribution $P(y)$ unique.

   We will now show the phenomena above for general memoryless channels.
   Define the joint distribution $P(\theta, x, y) = P(\theta)P(x|\theta)P(y|x)$.

   d) (3 Points) Put sign between

$$I(X;Y) \;\; ?? \;\; I(X;Y|\theta).$$

   e) (5 Points) Find a sufficient and necessary condition for equality above.
   f) (10 Points) **True/False** For memoryless channels, the optimal outputs distribution is unique.

2) **Divergence between Markov processes with the same transition probability, True or False (25 Points)** Prove each relation or provide counter example.
   a) (6 Points) Given are two input distributions P_X and Q_X that are transmitted on the same channel given by $P_{Y|X}$. The corresponding output distributions are denoted by P_Y and Q_Y.

$$D(P_X||Q_X) \geq D(P_Y||Q_Y). \tag{1}$$

   b) (12 Points) Define a Markov transition probability that, given x_1, generates $X_2, X_3, \ldots$ using $P_{X+|X}$. For example, at time $t = 4$, the joint probability given x_1 is:

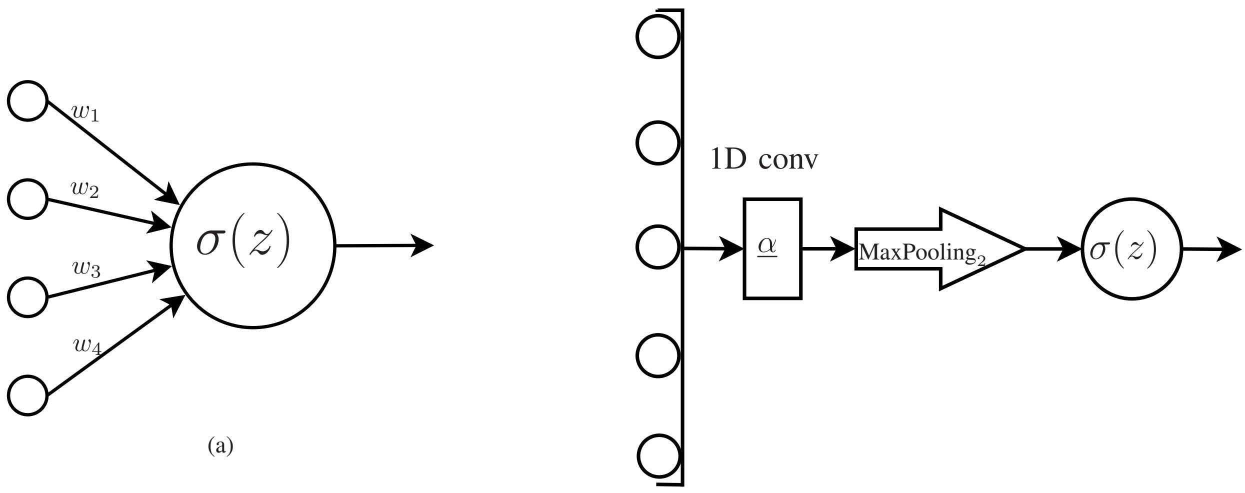
$$P(x_4, x_3, x_2|x_1) = P_{X+|X}(x_4|x_3)P_{X+|X}(x_3|x_2)P_{X+|X}(x_2|x_1)$$

   There are two distributions P_{X_1} and Q_{X_1} that serve as the initial distribution on X_1. Is the following true,

$$D(P_{X_i}||Q_{X_i}) \geq D(P_{X_j}||Q_{X_j}) \quad \text{for all } i \leq j. \tag{2}$$

   c) (7 Points) Let $f(t)$ and $g(t)$ be convex functions over $t \in [0, 1]$. Is the function $h(t) = \max(f(t), g(t))$ convex?

3) **Derivatives of neural Network (32 Points)**:



(a) (b)

Fig. 1

a) (12 Points) Consider the neural network in Fig. (1a) with no bias. Calculate the derivative of w_2 for both MSE and cross entropy cost functions, i.e. $\frac{\partial C_{MSE}}{\partial w_2}$ and $\frac{\partial C_{CE}}{\partial w_2}$ given $x = [0.5, 0.6, 0.1, 0.4]^T$,$y = 0$, $w = [0.25, 0.5, 0.8, 0.3]$ and σ is the sigmoid function.
Reminder: The MSE cost is $C_{MSE}(y, a) = \frac{1}{2}(y - a)^2$. Binary cross entropy cost is $C_{BCE}\ (y, a) = -y\log(a) - (1 - y)\log(1 - a)$.

b) (15 Points) Now we introduce you to two new layers of neural networks, 1D convolution and MaxPooling$_\beta$. 1D convolution pass the input layer in a FIR filter of size M, without zero padding. For input $x = [x(1), x(2), .., x(N)]^T$ and filter coefficients $\alpha = [\alpha(0), \alpha(1), .., \alpha(M - 1)]$, the output y is defined as:

$$y(k) = \sum_{i=0}^{M-1} x(k + i) \cdot \alpha(i) \qquad k = 1, 2, .., N - M + 1. \tag{3}$$

MaxPooling$_\beta$ just pools the maximum value of β consecutive elements. For input $x = [x(1), x(2), .., x(N)]^T$ and some β, the output y is defined as:

$$y(k) = \max\{x(k), x(k + 1), .., x(k + \beta - 1)\} \qquad k = 1, 2, .., N - \beta + 1. \tag{4}$$

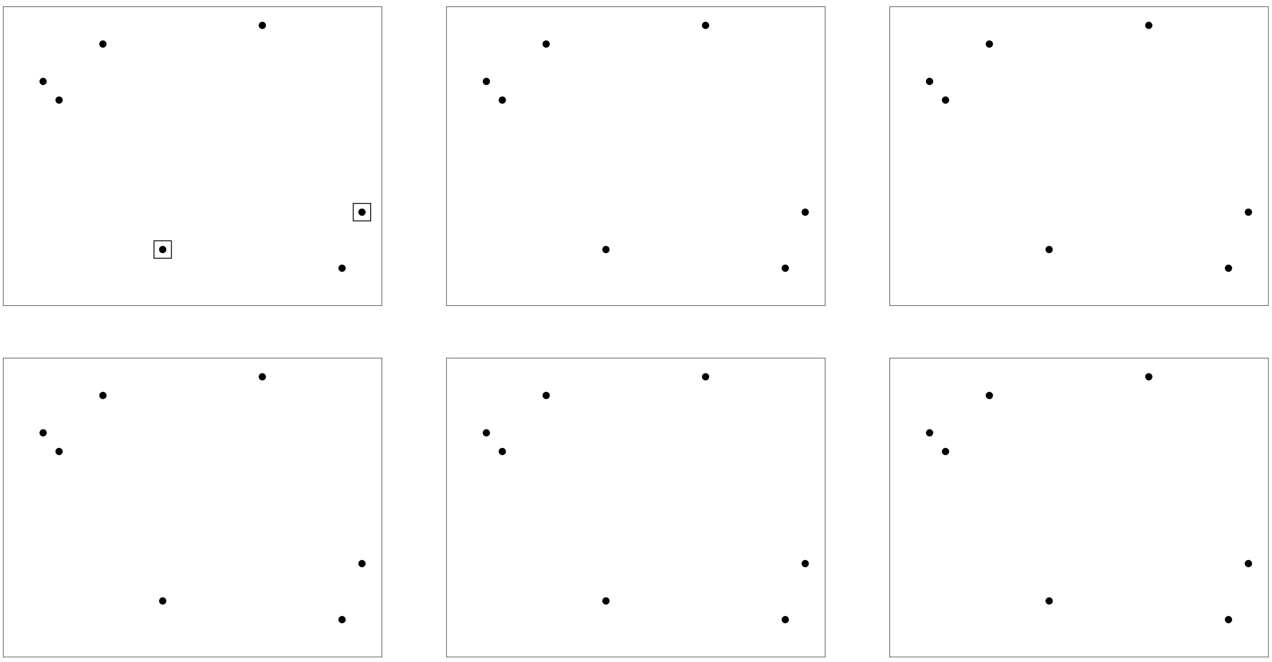
Consider the neural network in Fig. (1b) with MSE cost and no bias, calculate the derivative of w_2 , i.e. $\frac{\partial C_{MSE}}{\partial w_2}$, given $x = [0.5, 0.6, 1, 0.1, 0.4]^T$, $y = 0$, $\alpha = [0.25, 0.5, 0.25]$, $\beta = 2$ and $w = [0.25, 0.5]$.

c) (5 Points) A soft max with D input $z = [z(1), .., z(D)]$ and N outputs is

$$a(j) = \frac{e^{z(j)}}{\sum_{n=1}^{D} e^{z(n)}} \text{ for } j = 1, 2, ..., N. \tag{5}$$

Find $\frac{\partial a(i)}{\partial z(j)}$ and the sign of the derivative (positive or negative), i.e., sign $\left(\frac{\partial a(i)}{\partial z(j)}\right)$.

4) **K-means (15 Points)** Perform K-means iterations by hand on the dataset given below. Circles are data points and the two initial cluster centers are squares. Draw the cluster centers and the decision boundaries that define each cluster. If no points belong to a particular cluster, assume its center does not change. Use as many of the pictures as you need for convergence.



Good Luck!