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

Please copy the following sentence and sign it: " I am respecting the rules of the exam: Signature:_________ "

1) **(7 points)** Assume $Y_1 - Y_2 - \cdots - Y_m$ forms a Markov chain. Simplify $I(Y_1; Y_2, Y_3, \ldots, Y_m)$ to its simplest form.
2) **(7 points)** Assume $X - Y - Z$ forms a Markov chain. Show that

$$I(X;Y) \geq I(X;Y|Z).$$

When does an equality hold?
Hint: Chain rule on $I(X;Y,Z)$.

3) **(7 points)** Let $f(y)$ be an arbitrary function defined for $y \geq 1$. Let X be a random variable taking values in $\mathcal{X} = \{x_1, x_2, ..., x_n\}$ with probability $p_i = \Pr(X = x_i), i = 1, 2, ..., n$. Define the f-entropy of X by

$$H_f(X) \triangleq \sum_{i=1}^{n} p_i f\left(\frac{1}{p_i}\right).$$

If $f(\cdot)$ is concave, show that the following inequality is always satisfied:

$$H_f(X) \leq f(n). \tag{1}$$

4) **(17 points)** Assume X is a random variable taking values in $\mathcal{X} = \{1, 2, 3, ...\}$ with $E[X] = M$.
   a) **(10 points)** Show: $H(X) \leq M$.
   b) **(7 points)** For $M = 2$, what distribution P_X achieves an equality?
5) **(12 points)** Consider a ternary channel with input X_i and output Y_i, i.e. $X_i, Y_i \in \{0, 1, 2\}$. Let $\oplus$ denote addition modulo-3. The channel law is given by

$$Y_i = X_i \oplus W_i$$

where noises $\{W_i\}$ are independent of $\{X_i\}$ and are distributed i.i.d. $\sim W$, $W_i \in \{0, 1, 2\}$.

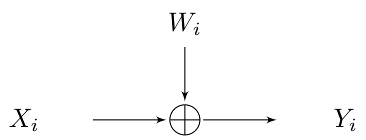

Fig. 1: An additive channel.

What is the capacity of this channel and what is the input distribution P_X that achieves the capacity?

6) **Neural networks Highway gate (28 pt)** Fig. 2 visualizes a simple Highway gated network. The network has three linear layers, the first two is followed by ReLU activation function (marked by σ). The Highway gate H and its complementary gate $\bar{H}$ are defined using a learnable parameter h as follows:

$$H(x) = x \cdot h, \tag{2}$$
$$\bar{H}(x) = x \cdot (1 - h). \tag{3}$$

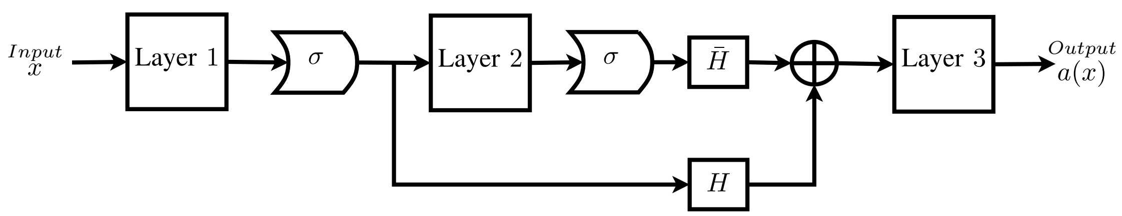

Fig. 2: A scheme of neural network with Highway gates

Initialize the network parameters as:

$$x = [0.1, 0.2, 0.6, 0.5]^T, y = 3, w^1 = \begin{bmatrix} 0.5 & 0.2 & 0.3 & -0.5 \\ 0.2 & -0.5 & 0.1 & 0.8 \\ -0.3 & 0.4 & 0.3 & -0.2 \end{bmatrix}, w^2 = \begin{bmatrix} 0.2 & 0.1 & 0.3 \\ 0.1 & -0.5 & 0.1 \\ 0 & 0.6 & -0.7 \end{bmatrix}, w^3 = [1.5, 1, 0.5]^T, h = 0.4.$$

a) **(10 points)** Calculate the derivatives $\frac{\partial C}{\partial w_{3,1}^2}, \frac{\partial C}{\partial h}$. Consider MSE cost function.
b) **(3 points)** Explain for what purpose one need to calculate the derivative in a).
c) **(8 points)** Calculate the derivative $\frac{\partial C}{\partial w_{2,2}^1}$ for $h = 0, h = 0.5$ and $h = 1$. In which case the derivative is largest?
d) **(7 points)** In feedforward neural networks with many layers, Highway gates are very common. Explain the motivation of using Highway gates in deep networks?

7) **Variant of MINE (32 pt)**
In this question we investigate an algorithm based on the mutual information neural estimator, using the following representation of mutual information:

$$I(X;Y) = H(X) + H(Y) - H(X,Y). \tag{4}$$

Let $X \sim P_X$, $Y \sim P_Y$ and denote the joint PMF of (X,Y) by P_{XY}. Let U_X be the PMF of the uniform discrete probability measure over $\mathcal{X}$, the alphabet of X (namely, $U_X(x) = \frac{1}{|\mathcal{X}|} \quad \forall x \in \mathcal{X}$).

a) **(5 points)** Prove the following equality:

$$H(X) = H(P_X, U_X) - D_{KL}(P_X \| U_X), \tag{5}$$

where $H(P_X, U_X)$ is the cross-entropy between P_X and U_X.
b) **(5 points)** If we replace the uniform PMF U_X by an arbitrary PMF V_X, does Eq. (5) still hold? Prove or disprove it.
c) **(5 points)** Based on the result of (a), prove the following equation:

$$I(X;Y) = D_{KL}(P_{XY} \| U_{XY}) - D_{KL}(P_X \| U_X) - D_{KL}(P_Y \| U_Y), \tag{6}$$

where U_Y and U_{XY} are defined in the same sense as U_X, on $\mathcal{Y}$ and $\mathcal{X} \times \mathcal{Y}$ respectively (assume that $|\mathcal{X} \times \mathcal{Y}| = |\mathcal{X}||\mathcal{Y}|$).
d) **(10 points)** Based on the KL divergence estimation method taught in class, propose an algorithm for the estimation of $I(X;Y)$ from a sample set $\{(x_i, y_i)\}_{i=1}^n \sim P_{XY}$, based on the equality proved in (b). Denote by $\widehat{I}_n^{(H)}(X,Y)$:
   i) Write the optimization objective
   ii) Give a block diagram of the proposed algorithm for estimating $\widehat{I}_n^{(H)}(X,Y)$. Assume the neural network consists of a single hidden layer with M units.
e) **(7 points)** We now wish to calculate the optimization objective $\widehat{I}_n^{(H)}(X,Y)$. For sufficiently large n, does the following hold? explain.

$$\widehat{I}_n^{(H)}(X,Y) \leq I(X;Y) \tag{7}$$

Good Luck!