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) Parallel marginal channels. (34 Points)

Consider the channel that is given in Fig. 1. The channel input is \( X \) and it has two outputs \( (Y_1, Y_2) \). The channel law is given by \( P_{Y_1,Y_2|X} \). We denote the capacity of this channel as \( C_A \).

![Fig. 1: Channel with a single input and two outputs.]

\[
X \xrightarrow{P_{Y_1,Y_2|X}} Y_1,Y_2
\]

a) (8 points) What is the capacity of this channel? Write explicitly the joint distribution of \((X, Y_1, Y_2)\).

b) (8 points) We now use the marginal version of this channel in Fig. 2. Specifically, the input is \( X \) and the outputs are \((Y_1, Y_2)\), but are generated according to the marginals distribution \( P_{Y_1|X} \) and \( P_{Y_2|X} \). That said, \( P_{Y_1|X} \) and \( P_{Y_2|X} \) are the marginals distributions of the original distribution \( P_{Y_1,Y_2|X} \). We denote the capacity of this channel as \( C_B \). What is the capacity of this channel? Write explicitly the joint distribution of \((X, Y_1, Y_2)\).

![Fig. 2: Channel with a single input and two outputs according to marginals.]

\[
\text{Encoder} \xrightarrow{P_{Y_1|X}} Y_1, \xrightarrow{P_{Y_2|X}} Y_2 \text{ Decoder}
\]

c) (8 points) True/False This sub-question is not related to the any of the above. For a joint distribution, \( P_{X,Y,Z} \), it is given that \( Z \) is a deterministic function of \( Y \). Define a new distribution \( Q_{X,Y,Z} = P_X P_{Y|X} P_{Z|Y} \). Is it true that \( Z \) is a deterministic function of \( Y \) under the new distribution \( Q \)??

d) (8 points) We now want to compare the capacities of the two settings above. Consider the special case where \( Y_2 \) is a function of \( Y_1 \) in the original distribution \( P_{Y_1,Y_2|X} \) (Fig. 1). Write \( \leq, =, \geq \) between \( C_A \) and \( C_B \), prove your answer.

   Hint: you can use the conclusion from c).

e) (5 points) Demonstrate the result you proved in the previous question by providing specific examples. For instance, if you proved that \( C_A \leq C_B \), then you should give one example for a channel with \( C_A = C_B \) and another example where \( C_A < C_B \).

2) True/False (27 Points):

a) Properties of mutual information: A joint distribution is given by \( P(x, \theta, y) = P(x)P(\theta)P(y|x, \theta) \). Answer the following three questions:

   i) (4 points) True/False: Is it true that there is a Markov chain \( X - Y - \theta \)? Prove or provide a counter example.

   ii) (4 points) Inequalities: Fill (and prove) one of the relations \( \leq, =, \geq \) between the following expressions:

\[
I(X; Y) \quad ?? \quad I(X; Y|\theta).
\]

iii) (3 points) Convex/Concave: Determine whether the mutual information, \( I(X_1; X_2) \) is convex OR concave function of \( P(x_2|x_1) \) for a fixed \( P(x_1) \). Hint: You can use your answers from the previous questions. You can not use the results we showed in class!

b) Machine learning:

   i) (4 points) True/False: In Tree Distribution lecture we conclude that the criteria for an optimal tree is \( \max_{\text{All Trees}} \sum_{i=1}^{n} I(x_i, x_{j(i)}) \), where \( x_i \) is the \( i^{th} \) feature, \( x_{j(i)} \) is the parent of the \( i^{th} \) feature, and \( I \) is the mutual information between both features. This criteria is equivalent to the Maximum-Likelihood criteria.

   ii) (3 points) True/False: In distribution tree a node can have more than two ‘sons’.

   iii) (9 points) True/False: In Fig. 3 the vertical axis represent the log-likelihood of some data, and the horizontal axis corresponds to the number of iterations. Copy each figure number and write True/False if this is a valid learning curve of an EM algorithm over GMM model.

   *iteration = E-step + M-step
3) **Neural Network Cost Function (30 Points):** Consider a standard Neural net with L layers. Denote $w^{l+1}$ as the matrix transformation from layer $l$ to layer $l + 1$, and $z^l$, $a^l$ as the pre-activation and post-activation neurons, i.e. $a^l = \sigma(z^l)$, $z^{l+1} = w^{l+1}a^l$ where $\sigma$ is the activation function. We define $a_1 \triangleq x$, i.e. the inputs, and $a \triangleq a^L$, i.e. the outputs.

The back-propagation equations that we learned in class are

\[
\delta^L = \nabla_C \otimes \sigma'(z^L) \\
\delta^l = ((w^{l+1})^T \delta^{l+1}) \otimes \sigma'(z^l) \\
\frac{\partial c}{\partial b_j} = \delta^1_j \\
\frac{\partial c}{\partial w^l_{j,k}} = a^{l-1}_{k} \delta^l_j
\]

where $\otimes$ is an element-wise multiplication.

a) (5 Points) Set $\sigma(z) = z$, i.e. identity function, and the cost function to be $c = \frac{1}{2}(y - a)^2$. Calculate $\delta^L$ in terms of $a$ and $y$.

From now on we denote $\delta^{(1)}$ as the $\delta^L$ calculated in (a).

b) (5 Points) Set $\sigma(x) = \frac{1}{1+e^{-x}}$, i.e. the sigmoid function, and the cost function to be $c = \frac{1}{2}(y - a)^2$. Calculate $\delta^L$ in terms of $\delta^{(1)}$ and $a$.

c) (8 Points) Set $\sigma(x) = \frac{1}{1+e^{-x}}$, i.e. the sigmoid function. Find new cost functions (there is more than one) that gives $\delta^L = \delta^{(1)}$.

d) (4 Points) Insight: What is the benefit of using one of the cost functions you found in (c) instead of using $c = \frac{1}{2}(y - a)^2$?

e) (8 Points) Set $\sigma(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$, i.e. the hyperbolic tangent function. Find new cost functions (there is more than one) that gives $\delta^L = \delta^{(1)}$. 

Fig. 3: Learning curves.
*Reminder: partial fraction decomposition example

\[
\frac{x + 9}{(x + 2)(x - 5)} = \frac{A}{x + 2} + \frac{B}{x - 5}
\]

\[A(x - 5) + B(x + 2) = x + 9\]

\[A + B = 1\]

\[-5A + 2B = 9\]

\[A = -1\]

\[B = 2.\]

4) **Decision trees (24 Points):** You wish to generate a model to predict if a mushroom is poisonous or not. In order to do so, you decide to use a decision tree and build it using the *ID3* algorithm with information gain. You have some empirical data:

<table>
<thead>
<tr>
<th>Example</th>
<th>Is heavy</th>
<th>Is smelly</th>
<th>Is spotted</th>
<th>Is smooth</th>
<th>Is poisonous</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>B</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>D</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>G</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>H</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>U</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>V</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>?</td>
</tr>
<tr>
<td>W</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
</tbody>
</table>

a) (5 Points) What is the empirical entropy of 'Is poisonous'?  
b) (9 Points) Which feature should you choose as the root of the decision tree? What is its information gain? *Hint: You can figure this out by looking at the data without explicitly computing the information gain of all four features.*  
c) (10 Points) Build the entire tree and use it to predict whether U,V,W are poisonous.

Good Luck!