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.

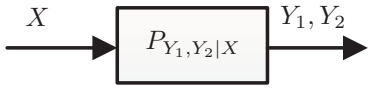

Fig. 1: Channel with a single input and two outputs.

a) (8 points) What is the capacity of this channel? Write explicitly the joint distribution of (X, Y_1, Y_2).
Solution: This is a memoryless channel, so that the capacity is $C_A = \max_{P_X} I(X; Y_1, Y_2)$. The joint distribution is $P_X P_{Y_1,Y_2|X}$.

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).
Solution: This is also a memoryless channel, but the channel is $Q_{Y_1,Y_2|X} = P_{Y_1|X} P_{Y_2|X}$. The capacity is also $C_B = \max_{P_X} I(X; Y_1, Y_2)$, but the joint distribution is $P_X Q_{Y_1,Y_2|X} = P_X P_{Y_1|X} P_{Y_2|X}$.

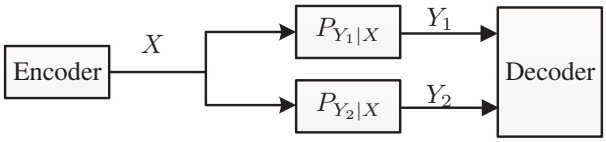

Fig. 2: Channel with a single input and two outputs according to marginals.

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 $Q_{X,Y,Z} = P_X P_{Y|X} P_{Z|X}$. Is it true that Z is a deterministic function of Y under the new distribution Q?
Solution: False. Take $X = \emptyset$ and $Z = Y$ in the P distribution. Clearly, under the Q distribution Z and Y are independent, and therefore, are not a function of each other.

d) (5 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).**

Solution: To be precise, we denote $I_P(X; Y_1, Y_2)$ as the mutual information that is computed with respect to $P_X P_{Y|X_1,X_2}$, and the same for Q. consider the following chain of inequalities:

$$\begin{aligned} C_A &= \max_{P_X} I_P(X; Y_1, Y_2) \\ &= \max_{P_X} I_P(X; Y_1) \\ &= \max_{P_X} I_Q(X; Y_1) \\ &\leq \max_{P_X} I_Q(X; Y_1, Y_2) \\ &= C_B, \end{aligned}$$

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 $C_A < C_B$.
Solution: Take X and Y_1 be the input and output of a BSC with transition probability p. Also, take $Y_2 = Y_1$. When $p = 0.5$, we have $C_A = C_B = 0$. For $p = 0.25$, one can show that $C_B > C_A$. Intuitively, in C_A, the decoder observes a single BSC, and in C_B, the decoder observes two BSC with single input so it has more information.

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.
Solution: False. Counterexample, let X and θ be two independent random variables, each distributed according to Bernoulli(0.5). Also, let $Y = X \oplus \theta$. One can check that $H(X|Y) \neq H(X|Y,\theta)$.

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

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

Solution: Consider the following chain of inequalities:

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &\overset{(a)}{=} H(X|\theta) - H(X|Y) \\ &\overset{(b)}{\leq} H(X|\theta) - H(X|Y,\theta) \\ &= I(X;Y|\theta), \end{aligned}$$

where (a) follows from the independence of X and θ, and (b) follows from conditioning reduces entropy. Therefore, $I(X;Y) \leq I(X;Y|\theta)$.

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!
Solution: We showed in class that mutual information is convex. Define:

$$\begin{aligned} P(\theta) &\sim Bern(\lambda) \\ P_{Y|X,\theta=0} &= P^1_{Y|X} \\ P_{Y|X,\theta=1} &= P^2_{Y|X} \end{aligned} \tag{1}$$

, where $\lambda \in [0,1]$, and $P^i_{Y|X}$ are two conditional distributions. From the previous question, we have $I(X;Y) \leq I(X;Y|\theta)$ and substituting (1) into this result shows the desired convexity.

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.
Solution: True.
We showed in class that

$$P_t(x,y,z,w) = 2^{-n(H(P_{x^n,y^n,z^n,w^n}) + D(P_{x^n,y^n,z^n,w^n}||P_t))}$$

since the empirical entropy $H(P_{x^n,y^n,z^n,w^n})$ does not depend on the selected tree but only on samples, then to obtain the maximum probability (likelihood), we should look for the minimum of the divergence $D(P_{x_n,y_n,z_n,w_n}||P_t)$.
As we saw in the lecture,

$$D(P||P_t) \overset{(b)}{=} const - \sum_{a \in \mathcal{X}} I(x_i, x_{j(i)}).$$

As we can see, to minimize the divergence, we need to maximize sum of the mutual information between parents and sons.

ii) (3 points) **True/False:** In distribution tree a node can have more than two 'sons'.
Solution: True.
By the definition of the tree structure.

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

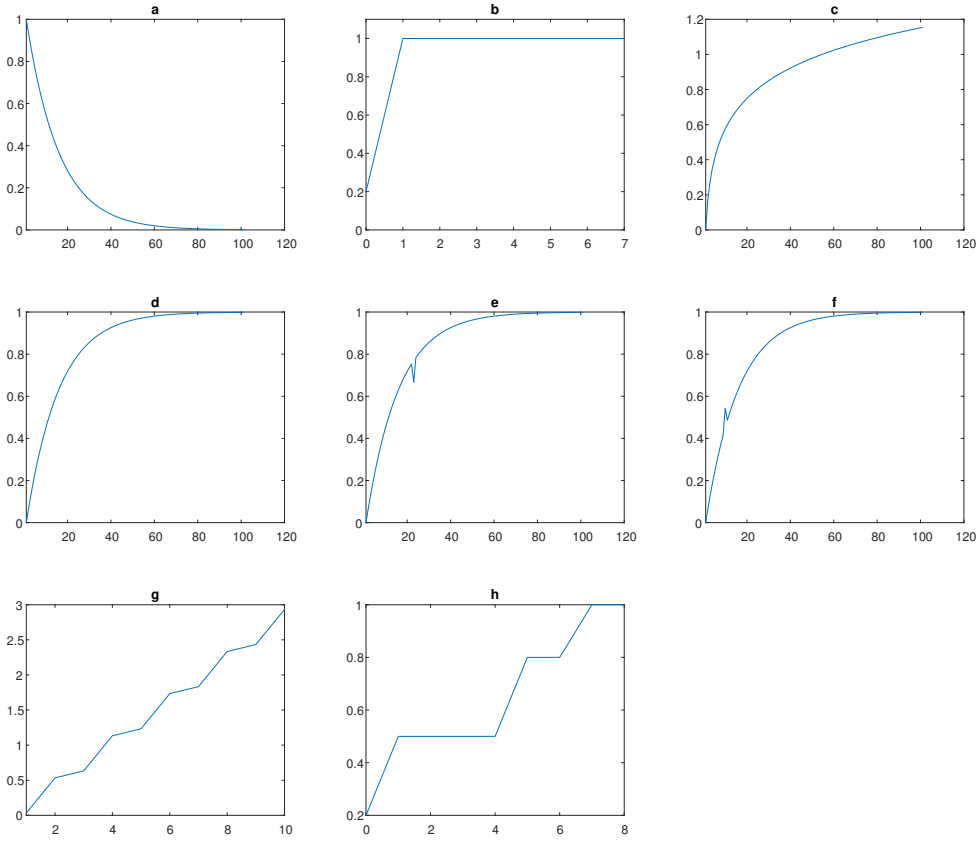

Fig. 3: Learning curves.

Solution: The graph should be non-decreasing. Also, once there was no improvement in an iteration, it means that no further improvement can occur in the following iterations(explanation at (h)).
Therefore,
b, c, d, g: True
a: False, the log-likelihood should be monotonic non-decreasing.
e: False, the log-likelihood should be monotonic non-decreasing, it can not be a "spike" in the log-likelihood.
f: False, the log-likelihood should be monotonic non-decreasing, it can not be a "spike" in the log-likelihood.
h: False, we see that at the second step the likelihood remains the same. Therefore, the parameters haven't change at this step, i.e. the maximization hadn't change the parameters. It means that the expectation step at the following iteration will produce the same weights as before and therefore, once again, the parameters won't change at the maximization step. in this graph, the likelihood suddenly changes at the fifth iteration.

*Note - the values at vertical axis are $L = \log(\prod_{i=1}^{n} p(x_i;\theta)) = \sum_{i=1}^{n} \log p(x_i;\theta)$. $p(x_i;\theta)$ is a density function and therefore can have values greater then 1 (for example, a single Gaussian with a very small variance).

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 σ 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

$$\begin{aligned}
\delta^L &= \nabla_a C \odot \sigma'(z^L) \\
\delta^l &= ((w^{(l+1)})^T \delta^{l+1}) \odot \sigma'(z^l) \\
\frac{\partial c}{\partial b_j^l} &= \delta_j^l \\
\frac{\partial c}{\partial w_{j,k}^l} &= a_k^{l-1}\delta_j^l,
\end{aligned}$$

where $\odot$ 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 δ^L in terms of a and y.

From now on we denote Δ as the δ^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 δ^L in terms of Δ 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 $\delta^L = \Delta$.

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 $\delta^L = \Delta$.

*Reminder: partial fraction decomposition example

$$\begin{aligned} \frac{x+9}{(x+2)(x-5)} &= \frac{A}{x+2} + \frac{B}{x-5} \\ A(x-5) + B(x+2) &= x+9 \\ A + B &= 1 \\ -5A + 2B &= 9 \\ A &= -1 \\ B &= 2. \end{aligned}$$

Solution:

a)

$$\begin{aligned} \delta^L &= \nabla_a C \odot \sigma'(z^L) \\ &= \frac{\partial C}{\partial a} 1 \\ &= \frac{\partial(\frac{1}{2}(y-a)^2)}{\partial a} \\ &= a - y. \end{aligned}$$

We notice that when the difference between the real target and the net's output is great, Δ is high with respect to it.

b) We know that the derivative of Sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ is $\sigma'(x) = \sigma(x)(1-\sigma(x))$. Additionally, $\sigma(z^L)$ is given as a. Hence:

$$\begin{aligned} \delta^L &= \nabla_a C \odot \sigma'(z^L) \\ &= \frac{\partial C}{\partial a} \sigma(z^L)(1 - \sigma(z^L)) \\ &= (a-y)(1-a)a \\ &= \Delta(1-a)a. \end{aligned}$$

c) We want that Δ will be equal to $\frac{\partial C}{\partial a}(1-a)a$ when $\Delta = a - y$ as we found at (a). We get

$$\begin{aligned} \frac{\partial C}{\partial a} &= \frac{a-y}{(1-a)a} \\ &= \frac{1}{1-a} - \frac{y}{(1-a)a}. \end{aligned}$$

Therefore:

$$\frac{\partial C}{\partial a} = \frac{1-y}{1-a} - \frac{y}{a}.$$

Integrate both sides of the equation with respect to a :

$$C(a) = -(yln(a) + (1-y)ln(1-a)) + Constant.$$

d) We can identify the cost function we've got on (c) as the Cross Entropy function (when the constant is 0). We can see in the figure below that the Sigmoid function goes into saturation when its values are close to 0 or 1, i.e. gradients values

are close to zero. Therefore, We conclude that δ^L in (b) might be very small even if the difference between the real target and the net's output is high (For example, look at the case where the true label is 1, but the output is close to 0). We can see from the equations of Back-Propagation that the gradients of all of the parameters are multiplied by δ^L. This means that if the output neurons are saturated then all of the parameter's gradient values will go to zero. Therefore, using cross-entropy helps to deal with the vanishing gradient issue, and speeds up training.

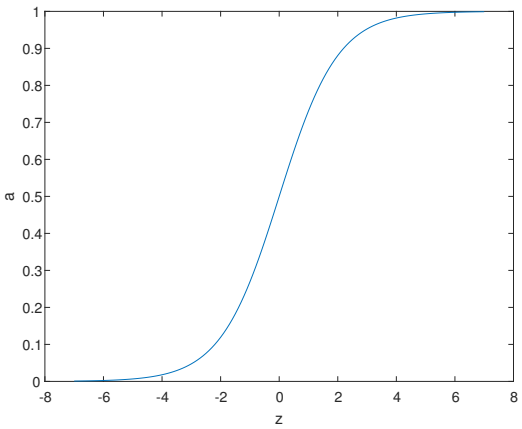

Fig. 4: Sigmoid cost function. Gradient goes to zero on saturation

You can read more here http://neuralnetworksanddeeplearning.com/chap3.html at section "The cross-entropy cost function".

e) Notice that

$$\begin{aligned}\sigma'(z) &= 1 - (\frac{e^z - e^{-z}}{e^z + e^{-z}})^2 \\ &= 1 - \sigma^2(z).\end{aligned}$$

Therefore:

$$\begin{aligned}\delta^L &= \frac{\partial C}{\partial a}(1 - \sigma^2(z^L)) \\ &= \frac{\partial C}{\partial a}(1 - a^2).\end{aligned}$$

For $\delta^L = \Delta = a - y$ we get:

$$\begin{aligned}\frac{\partial C}{\partial a} &= \frac{a-y}{1-a^2} \\ &= \frac{1}{2}\frac{1-y}{1-a} - \frac{1}{2}\frac{1+y}{1+a}.\end{aligned}$$

Integrate both sides of the equation as we did at (c):

$$C = -\frac{1}{2}((1+y)ln(1+a) + (1-y)ln(1-a)) + Constant.$$

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:

Example	Is heavy	Is smelly	Is spotted	Is smooth	Is poisonous
A	0	0	0	0	0
B	0	0	1	0	0
C	1	1	0	1	0
D	1	0	0	1	1
E	0	1	1	0	1
F	0	0	1	1	1
G	0	0	0	1	1
H	1	1	0	0	1
U	1	1	1	1	?
V	0	1	0	1	?
W	1	1	0	0	?

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.

Solution:

1) $h_b('Is\ poisonous') = h_b\left(\frac{3}{8}\right) = 0.954.$
2) It is possible to figure out that 'Is smooth' is dividing the data most i.e. yielding largest information gain. Alternately, it is very simple to calculate information gain for each feature:

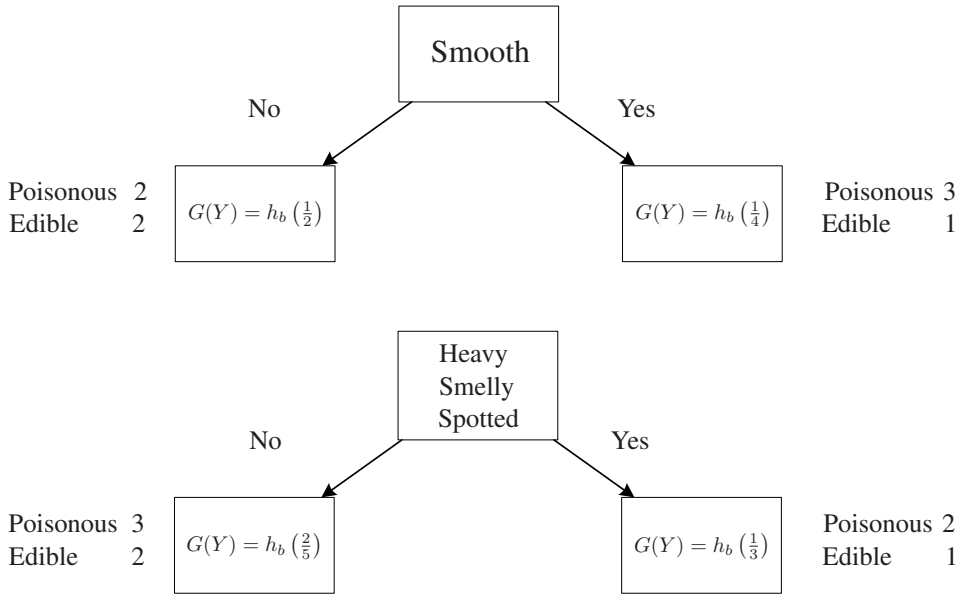

Fig. 5: Comparison of all features for the root

$$Information\ gain(Y;\Theta) = h_b(Y) - \sum_{\theta=0}^{1} P(\Theta=\theta)h_b(Y|\Theta=\theta)$$
$$Information\ gain(Y;\Theta=smooth) = 1 - \frac{1}{2}h_b(\frac{1}{2}) - \frac{1}{2}h_b(\frac{1}{4}) = 0.094$$
$$Information\ gain(Y;\Theta=heavy\backslash smelly\backslash spotted) = 1 - \frac{5}{8}h_b(\frac{2}{5}) - \frac{3}{8}h_b(\frac{1}{3}) = 0.048.$$

3) *ID3* algorithm's $argmax$ will pick 'Is smooth' as root node. Next steps (using recursion) will be called with the following data sets:
Data set for 'Is smooth'=yes:

Example	Is heavy	Is smelly	Is spotted	Is poisonous
C	1	1	0	0
D	1	0	0	1
F	0	0	1	1
G	0	0	0	1

And data set for 'Is smooth'=no:

Example	Is heavy	Is smelly	Is spotted	Is poisonous
A	0	0	0	0
B	0	0	1	0
E	0	1	1	1
H	1	1	0	1

The $argmax$ result here is also very visible. 'Is smelly' will be picked for maximizing the gain on both calls (splitting the data completely). The new recursion calls will mark nodes as leaves and label them accordingly(1st and 2nd stopping conditions).

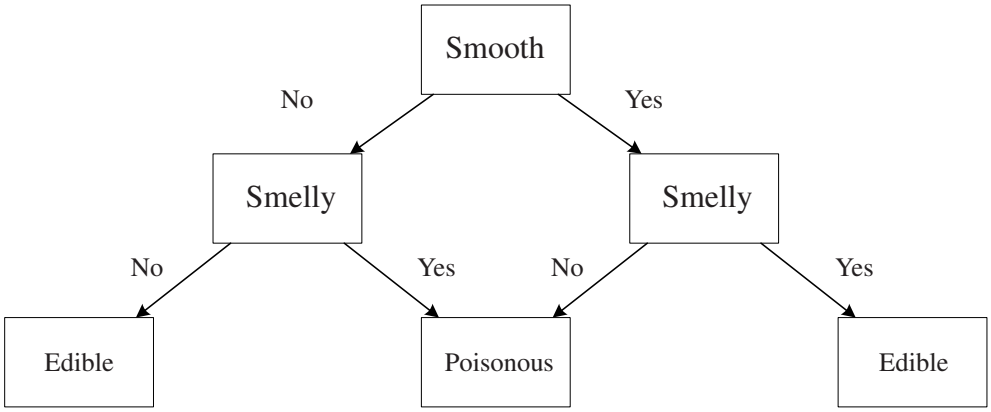

Fig. 6: Entire decision tree built by *ID3* with information gain

By the figure, U and V are edible while W is poisonous. Note that when you examine the tree carefully you see that the decision is made by exclusive or of 'Is smooth' and 'Is smelly' features.

Good Luck!