Introduction to Information Theory and Machine Learning July 28, 2021.
(Prof. Permuter Haim, Mr. Eli Shmuel, Mr. Dor Tsur, Mr. Bashar Huleihel and Mr. Omer Luxembourg)

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) **True or False (16 Points)**:
   a) For any two random variables X, Y and any $a \in \mathbb{R}$: $H(Y|aX) = H(Y|X)$.
   Solution:
   False. Notice that for $a = 0$ we have $H(Y)$ which is not equal to $H(Y|X)$ in general. Only for $a \neq 0$ the equality would hold, because the conditional entropy is label invariant, which follows from $H(Y|X) = H(X,Y) - H(X)$ and the fact that the entropy of a random variable is label invariant.
   b) Assume that the Markov chain $Y_0 - Y_1 - \cdots - Y_n$ holds. Then $H(Y_0|Y_n)$ is non-decreasing with n.
   Solution:
   True.
$$H(Y_0) - H(Y_0|Y_{n-1}) = I(Y_0; Y_{n-1}) \geq I(Y_0; Y_n) = H(Y_0) - H(Y_0|Y_n), \tag{1}$$
   where the inequality follows from the given Markov chain and the data processing theorem. Hence, $H(Y_0|Y_{n-1}) \leq H(Y_0|Y_n)$, i.e., $H(Y_0|Y_n)$ is non-decreasing with n.
   c) There exists a discrete memoryless channel (DMC) with the following input and output alphabets: $\mathcal{X} = \{0,1\}$ and $\mathcal{Y} = \{2,3,4,5\}$, respectively, with capacity $C = 2$ bits / channel use.
   Solution:
   False.
$$I(X;Y) = H(X) - H(X|Y) \leq H(X) \leq \log|\mathcal{X}| = 1. \tag{2}$$
   d) The divergence is symmetric, i.e., $D(P_X||Q_X) = D(Q_X||P_X)$ for any P_X and Q_X.
   If true, prove it. **Otherwise**, in this section only, you do not need to provide a counter-example, but give an example of two different distributions P_X and Q_X ($P_X \neq Q_X$), such that an equality holds.
   Solution:
   False. Clearly, in the general case
$$D(P_X||Q_X) = \sum_x P(x) \log \frac{P(x)}{Q(x)} \neq \sum_x Q(x) \log \frac{Q(x)}{P(x)} = D(Q_X||P_X),$$
   and you can think of easy counter-examples. Hence, we are asked to give an example for which the equality holds.
   Example: Assume a binary alphabet and $P(X=0) = p, Q(X=0) = q$ where $q = 1-p$. Hence,
$$D(P_X||Q_X) = p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} = q \log \frac{q}{p} + (1-q) \log \frac{1-q}{1-p} = D(Q_X||P_X).$$

2) **Entropy of 3 pairwise independent random variables (12 Points)**: Let W, X, Y be 3 random variables distributed each Bernoulli (0.5) that are pairwise independent, i.e., $I(W;X) = I(X;Y) = I(W;Y) = 0$.
   a) What is the **maximum** possible value of $H(W,X,Y)$?
   Solution:
   From the chain rule for entropy and the fact that W and X are independent,
$$H(W,X,Y) = H(W) + H(X) + H(Y|W,X) \leq H(W) + H(X) + H(Y),$$
   where the inequality follows since conditioning reduces entropy. $H(W) = H(X) = H(Y) = 1$, thus $H(W,X,Y) \leq 3$.
   b) What is the condition under which this **maximum** is achieved?
   Solution:
   Notice that the maximum is achieved if $H(Y|X,W) = H(Y)$, i.e., Y is independent of the pair (X,W) (or similarly W is independent of (X,Y), or X is independent of (W,Y)).
   c) What is the **minimum** possible value of $H(W,X,Y)$?
   Solution:
   On the other hand,
$$H(W,X,Y) = H(W) + H(X) + H(Y|W,X) \geq H(W) + H(X), \tag{3}$$
   where the inequality follows from the non-negativity of the conditional entropy $H(Y|W,X)$. Thus, $H(W,X,Y) \geq 2$.
   d) Give a specific example achieving this **minimum**.
   Solution:

If Y is a deterministic function of both (W, X), the inequality is achieved. Notice that it cannot be a deterministic function of just one of them since it contradicts the assumption of the question. Let, for instance, $Y = W \oplus X$, where $\oplus$ denotes the addition modulo 2 (i.e., XOR).

3) **Cascaded BSCs (xx Points)**: Consider two binary symmetric channels, $\mathrm{BSC}(\alpha)$ and $\mathrm{BSC}(\beta)$, with crossover probabilities $\alpha, \beta \in (0, 1)$. Let k be an even integer. Then, we cascade independent k such channels, where half of them are (identical) $\mathrm{BSC}(\alpha)$, and the other half are (identical) $\mathrm{BSC}(\beta)$. We cascade those channels *alternately*, namely, the first channel is $\mathrm{BSC}(\alpha)$, the second channel is $\mathrm{BSC}(\beta)$, the third channel is $\mathrm{BSC}(\alpha)$, etc.

a) Assume that neither encoding nor decoding is allowed at the intermediate terminals. What is the capacity of this cascaded channel as a function of α, β, and k? You are allowed to express your result in terms of convolutions (see remark below).
Solution:
Recall that cascading BSCs result in a new BSC with a new crossover probability γ. Thus, the capacity is given by $C = 1 - H_2(\gamma)$ where γ can be written as

$$\gamma = (\bar{\alpha}\beta + \alpha\bar{\beta})^{\oplus \frac{k}{2}}.$$

More explicitly, γ can be found as:

$$\begin{aligned}\gamma &= \sum_{\{i \leq \frac{k}{2} : i \text{ is odd}\}} \binom{\frac{k}{2}}{i} (\alpha \oplus \beta)^i (1 - \alpha \oplus \beta)^{\frac{k}{2} - i} \\ &= 0.5 \left(1 - \left(1 - 2(\bar{\alpha}\beta + \alpha\bar{\beta})\right)^{\frac{k}{2}}\right).\end{aligned}$$

b) Assume that encoding and decoding is allowed at the intermediate points. What is the capacity of this channel as a function of α, β, and k?
Solution:
In the HW you have shown that when encoding and deconding is allowed at the intermediate points then the capacity of the cascaded channels is given by $C = \min C_i$. Since there are only two types of channels, i.e., BSCs with crossover probabilities α and β, then the capacity is given by:

$$C = \min\{1 - H_2(\alpha), 1 - H_2(\beta)\}.$$

c) What is the capacity of each of the above settings when $k \to \infty$?
Solution:
It is easy to verify that $\left(1 - 2(\bar{\alpha}\beta + \alpha\bar{\beta})\right) < 1$. Accordingly, $\gamma \to 0.5$ as $k \to \infty$, which implies that for the first item $C_a \to 0$. For the second item, the number of cascaded channels does not change the capacity which remains $C_b = \min\{1 - H_2(\alpha), 1 - H_2(\beta)\}$.

d) Suppose now that the channels are not cascaded alternately but in some random order. Will the capacity remain the same in section (a) and in section (b)? Explain your answers.
Solution:
For section (a), the capacity will remain the same since the crossover probability of the cascaded channels will not be changed. Further, for section (b), there are still the same two types of channels, and therefore, here too, the capacity remains the same.

e) We now cascade k (even integer) identical and independent $\mathrm{BSC}(\alpha)$, with $\alpha \in (0, 0.5)$. Assume that encoding and decoding is allowed only at one intermediate point $1 < m < k$. What is the capacity of this channel as a function of α, k and m? Also, which m maximizes the capacity, i.e., where is the optimal position of the intermediate point? Prove your answer.
Remark: You can express your result in terms of convolutions (for your convenience you may use $\alpha^{\oplus k}$ to designate the convolution of α with itself k times). The convolution between α and β is defined as: $\alpha \oplus \beta = \alpha(1 - \beta) + (1 - \alpha)\beta$.
Solution:
In this case the capacity is given by:

$$C = \min\left\{1 - H_2\left(\alpha^{\oplus m}\right), 1 - H_2\left(\alpha^{\oplus (k-m)}\right)\right\}.$$

Next, it will be shown that $m = \frac{k}{2}$ maximizes the capacity. For convenience, let us denote $\gamma_m = \alpha^{\oplus m}$. It can be easily verified that the following recursive formula holds:

$$\gamma_m = 2\gamma_{m-1}(1 - \gamma_{m-1}).$$

Clearly, for $\alpha < 0.5$, this sequence is monotonically increasing. Also,

$$\begin{aligned}\lim_{m\to\infty} \gamma_m &= \lim_{m\to\infty} 0.5\left(1 - (1 - 2\alpha)^m\right) \\ &= 0.5.\end{aligned}$$

Accordingly, since $H(p)$ is a concave function that is maximized for $p = 0.5$, we can conclude that the optimal choice of m is achieved for $\frac{k}{2}$.

4) **Loss Functions for Logistic Regression Models (22 points)**: Given two models, we want to select the best model in terms of the loss function. Both of the models are a logistic regression model, but with a different architecture. The models are created by a function $g(\cdot) \to [0, 1]$ as follows:

$$\hat{P}(y|x; w) = g\left(w_0 + \sum_{n=1}^{M} w_n \phi_n(x)\right),$$

where $x \in \mathbb{R}$ is the input of the model.

$$\text{Model 1: } \phi_i^{(1)}(x) = \begin{cases} x^2 & i = 1 \\ 0 & i > 1 \end{cases}, \quad \text{Model 2: } \phi_i^{(2)}(x) = \begin{cases} x & i = 1 \\ cos(x) & i = 2 \\ 0 & i > 2 \end{cases}$$

a) **(4 points)** Given N training samples $(x_i, y_i), i \in \{1, 2, ...N\}$, we evaluate the MSE risk function score of the two models. Which is better in terms of the risk function score? Model 1, model 2 or neither? Explain your answer.
Solution:
There is no better model. The score is dependent on the training samples.

b) **(4 points)** Define the Baysian Information Criteria (BIC) as follows:

$$BIC = -2 \times LL(N) + \log(N) \times k,$$

where N is the number of samples, $LL(N)$ is the log-likelihood as a function of N, and k is the number of parameters in the model. This criterion measures the trade-off between model fit and complexity of the model.
Let $LL_1(N)$ be the log-probability of the labels that model 1 predicted to N training samples, where the probabilities are evaluated at the maximum likelihood setting of the parameters. Let $LL_2(N)$ be the corresponding log-probability for model 2. We assume here that $LL_1(N)$ and $LL_2(N)$ are evaluated on the basis of the first N training examples from a much larger set.
Our empirical studies has shown that these log-probabilities are related in the next way:

$$LL_2(N) - LL_1(N) \approx 0.001 \times N.$$

How will we select between the two models, when using the BIC score , as a function of the number of training examples? Choose the correct answer.
Solution:
First select model 1, then for larger N select model 2

c) **(6 points)** Provide an explanation for your last answer.
Solution:
For larger N we would select model 2 since it has a consistent (might be small) advantage. For smaller N's, however, we would choose model 1 due to smaller complexity penalty. To see this a bit more precisely, let's calculate BIC for each model: $BIC_1 = 2L_1(N) - d_1 \cdot log(N)$, where d_1 is the number of parameters in model 1.
Thus the difference between BIC scores is:

$$\begin{aligned} BIC_2 - BIC_1 &= -2\left(L_2(N) - L_1(N)\right) + (d_2 - d_1)log(N) \\ &= -0.002 \cdot N + (3-2)log(N) = -0.002 \cdot N + log(N) \end{aligned}$$

When N is small the complexity term dominates and $BIC_1 < BIC_2$ (difference between them is positive). For larger N the linear increase of the log-likelihood difference overcomes the logarithmic penalty and $BIC_1 > BIC_2$.

d) **(8 points)** This section does not depend on the previous ones. Let $g(\cdot)$ be the Sigmoid function and consider the Binary Cross-Entropy loss function, where the labels, $\{y_i\}_{i=1}^{N}$, are 0 or 1. Suppose you use gradient descent to obtain the optimal parameters $\{w_i\}_{i=0}^{M}$ for each model. Give the update rule to each parameter for the two models.
Solution:
To obtain the parameters in each model, we want to minimize $\frac{1}{N}\sum_i BCE(M_j(x_i), y_i)$ for model j.
Notation: $w_{j,i}$ is the weight w_i of model j. z is the argument inside the g function - $w_0 + \sum_{n=1}^{M} w_n \phi_n(x)$.

$$C_{BCE} = -\frac{1}{N}\sum_{(x,y)} \left(ylog(\sigma(z)) + (1-y)log(1-\sigma(z))\right)$$

$$\frac{\partial \sigma(z)}{\partial z} = \sigma(z)(1-\sigma(z))$$

$$\begin{aligned}
\frac{\partial C(w_i)}{\partial z} &= -\frac{1}{N}\sum_{(x,y)}\left(y\frac{\sigma'(z)}{\sigma(z)}-(1-y)\frac{\sigma'(z)}{1-\sigma(z)}\right)\\
&= -\frac{1}{N}\sum_{(x,y)}\left(y\frac{\sigma(z)(1-\sigma(z))}{\sigma(z)}-(1-y)\frac{\sigma(z)(1-\sigma(z))}{1-\sigma(z)}\right)\\
&= -\frac{1}{N}\sum_{(x,y)}\left(y(1-\sigma(z))-(1-y)\sigma(z)\right)\\
&= \frac{1}{N}\sum_{(x,y)}\left(\sigma(z)-y\right)
\end{aligned}$$

$$\begin{aligned}
\Delta w_{1,0} &= \frac{\partial C(w_{1,0})}{\partial z}\frac{\partial z}{\partial w_{1,0}} = \frac{1}{N}\sum_{(x,y)}\big(\sigma(w_{1,0}+w_{1,1}x^2)-y\big)\\
\Delta w_{1,1} &= \frac{\partial C(w_{1,1})}{\partial z}\frac{\partial z}{\partial w_{1,1}} = \frac{1}{N}\sum_{(x,y)}\big(\sigma(w_{1,0}+w_{1,1}x^2)-y\big)\cdot x^2\\
\Delta w_{2,0} &= \frac{\partial C(w_{2,0})}{\partial z}\frac{\partial z}{\partial w_{2,0}} = \frac{1}{N}\sum_{(x,y)}\left(\sigma(w_{2,0}+w_{2,1}x+w_{2,2}cos(x))-y\right)\\
\Delta w_{2,1} &= \frac{\partial C(w_{2,1})}{\partial z}\frac{\partial z}{\partial w_{2,1}} = \frac{1}{N}\sum_{(x,y)}\left(\sigma(w_{2,0}+w_{2,1}x+w_{2,2}cos(x))-y\right)\cdot x\\
\Delta w_{2,2} &= \frac{\partial C(w_{2,2})}{\partial z}\frac{\partial z}{\partial w_{2,2}} = \frac{1}{N}\sum_{(x,y)}\left(\sigma(w_{2,0}+w_{2,1}x+w_{2,2}cos(x))-y\right)\cdot cos(x).
\end{aligned}$$

5) **Variational Inference (xx points)** In class, we had learned how to apply the tools of variational inference to the Bayesian mixture of Gaussians model. In this question, we will apply them to approximate the mixture of exponential distributions. Assume the following setting: We consider a variant of the Bayesian mixture distribution taught in class: Our data is distributed as a mixture of K exponential distributions with the following parameters:

- The exponential distribution parameter is also a random variable, apriori distributed exponentially: $\mu_k \sim \exp(\lambda_k)$ for $k \in \{1...K\}$
- c_i is the probability that x_i belongs to the k^{th} exponential distribution for $k \in \{1\dots K\}$. The apriori distribution of it is $c_i \sim \mathsf{Unif}(K)$, which, as taught in class, can be encoded into a one-hot vector.

In this question we would like to approximate $P(z^m|x^n)$ from a set of n samples x^n using variational inference. Therefore, we will use the distribution $q(z^m)$ to do that. This distribution is defined using the parameters (φ, b^K) as follows:

- $\mu_k \sim \exp(b_k)$ for $k \in \{1\dots K\}$
- $q(c_i)$ is the categorical distribution $c_i \sim \boldsymbol{\varphi}_i$ for $i \in \{1\dots n\}$ with $\boldsymbol{\varphi}_i = \{\varphi_{i,1},\dots,\varphi_{i,k}\}$.

a) What is z^m in our question? what is the size of m?

Solution:

As learned in class, z^m consists of the random variables forming the mixture distribution. Therefore, $z^m = (\mu^K, c^n)$, this means that $m = K + n$.

b) Write $P(x^n, z^m)$ as explicit as you can by filling the following equalities in your notebook:

$$\begin{aligned}
P(x^n, z^m) &= P(x^n, \mu^K, c^n)\\
&= ?\\
&= ?\\
&= ?
\end{aligned}$$

Solution:

The joint distribution is given by:

$$\begin{aligned}
P(x^n, z^m) &= P(x^n, \mu^K, c^n)\\
&= P(\mu^K)P(c^n, x^n|\mu^k)\\
&= \prod_{k=1}^{K} P(\mu_k)\prod_{i=1}^{n} P(c_i)P(x_i|c_i,\mu^K)\\
&= \prod_{k=1}^{K} \lambda_k \exp(-\lambda_k\mu_k)\prod_{i=1}^{n}\frac{1}{K}\mu_{c_i}\exp(-\mu_{c_i}x_i).
\end{aligned}$$

c) Write explicitly the ELBO for our case, assume mean-field approximation.

Solution:

$$\begin{aligned}\mathrm{ELBO}(\varphi, b^k) &= \mathbb{E}\left[\log P(z^m)\right] + \mathbb{E}\left[\log P(x^n|z^m)\right] - \mathbb{E}\left[\log q(z^m)\right] \\ &= \sum_{k=1}^{K} \mathbb{E}\left[\log P(\mu_k)\right] + \sum_{i=1}^{n} \mathbb{E}\left[\log P(c_i)\right] + \sum_{i=1}^{n} \mathbb{E}\left[\log P(x_i|c_i, \mu^K); \varphi_i, b^k\right] \\ &- \sum_{i=1}^{n} \mathbb{E}\left[\log q(c_i; \varphi_i)\right] - \sum_{k=1}^{K} \mathbb{E}\left[\log q(\mu_k; b_k)\right] \\ &= \sum_{k=1}^{K} \log(\lambda_k \exp(-\lambda_k \mu_k)) + \sum_{i=1}^{n} - \log K + \sum_{i=1}^{n} \mathbb{E}\left[\log(\mu_{c_i} \exp(-\mu_{c_i} x_i))\right] \\ &- \sum_{i=1}^{n} \mathbb{E}\left[\log q(c_i; \varphi_i)\right] - \sum_{k=1}^{K} \mathbb{E}\left[\log q(\mu_k; b_k)\right].\end{aligned}$$

The derivation for update equations (not in this question) requires to derivate with respect to $q(\cdot)$ so we can leave $q(c_i; \varphi_i)$ and $q(\mu_k; b_k)$ as is. Answers that explicitly showed them using the definition of $q(\cdot)$ were also accepted.

d) This section does not depend on the previous ones. The derivation taught in class for the Bayesian mixture of Gaussians arrives to an update of the form $\varphi_{i,k} \propto f(\mu_k, x_i, m_k, s_i^2)$ for some function f. What steps are required to derive a formula of the form $\varphi_{i,k} = g(\mu_k, x_i, m_k, s_i^2)$ for some function g? i.e., from just proportion to an equation.Express g as a function of f.

Solution:

We know that $\{\varphi_{i,1} \dots \varphi_{i,K}\}$ is a probability distribution. Therefore, we need to perform a normalization of the terms sum that $\sum_{k=1}^{K} \varphi_{i,k} = 1$.

e) This section does not depend on the previous ones. Assume we want to perform a maximization of some arbitrary function $g(x, y)$. We know how to maximize g over x when y is fixed, and over y when x is fixed. Suggest an algorithm for the maximization of g with respect to both variables. Under your suggested algorithm, are we guaranteed to converge to the global maximum of g?

Solution:

The proposed algorithm will be the coordinate ascent maximization algorithm. This algorithm will obtain the global maximum only if g is convex in both x and y.

Remark: This algorithm is a member of the family of alternating maximization algorithms. Alternating maximization algorithms are widely used in both the fields of information theory (you can read about the Blahut-Arimoto algorithm) and most machine learning algorithms that involve several models.

Good Luck!