Final Exam - Moed A

Total time for the exam: 3 hours!

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

Important: For **True / False** questions, copy the statement to your notebook and write clearly true or false. You should prove the statement if true, or disprove it, e.g. by providing a counter-example, otherwise.

1) **True or False (16 Points)**:
   a) For any two random variables X, Y and any $a \in \mathbb{R}$: $H(Y|aX) = H(Y|X)$.
   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.
   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.
   d) The divergence is symmetric, i.e.,
   $$D(P_X||Q_X) = D(Q_X||P_X) \tag{1}$$
   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 (1) holds.

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 $H(W, X, Y)$?
   b) What is the condition under which this **maximum** is achieved?
   c) What is the **minimum** possible value of $H(W, X, Y)$?
   d) Give a specific example achieving this **minimum**.

3) **Cascaded BSCs (32 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) (6 points) 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).
   b) (4 points) Assume that decoding and encoding is allowed at the intermediate points. What is the capacity of this channel as a function of α, β, and k?
   c) (4 points) What is the capacity of each of the above settings when $k \to \infty$?
   d) (6 points) 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.
   e) (12 points) We now cascade k (even integer) identical and independent $\mathrm{BSC}(\alpha)$, with $\alpha \in (0, 0.5)$. Assume that decoding and encoding 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: $\alpha \oplus \beta = \alpha(1 - \beta) + (1 - \alpha)\beta$.

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.

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

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.

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.

i) Always select model 1.
ii) Always select model 2.
iii) First select model 1. Then, for larger N, select model 2.
iv) First select model 2. Then, for larger N, select model 1.

c) (6 points) Provide an explanation for your last answer.

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.

5) **Variational Inference (28 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 exponential assignment of x_i, which, as taught in class, can be encoded into a one-hot vector. The apriori distribution of it is $c_i \sim \mathsf{Unif}(K)$.

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) (4 points) What is z^m in our question? what is the size of m?

b) (6 points) Write $P(x^n, z^m)$ as explicit as you can by filling the following qualities in your notebook:

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

c) (8 points) Write explicitly the ELBO for our case, assume mean-field approximation.
Reminder: ELBO $= \mathbb{E}\left[\log P(z^m)\right] + \mathbb{E}\left[\log P(x^n|z^m)\right] - \mathbb{E}\left[\log q(z^m)\right]$.

d) (6 points) 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.

e) (4 points) 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?
Reminder: The probability density of the exponential distribution with parameter λ is given by $f_X(x) = \lambda \exp(-\lambda x)$ for $x > 0$.

Good Luck!