(Prof. Permuter Haim, Mr. Eli Shmuel, Mr. Dor Tsur, Mr. Bashar Huleihel and Mr. Omer Luxembourg)

Final Exam - Moed B

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 (24 Points)**:
   a) For a source X, the average length of a non-singular code must be **greater than or equal** its entropy, $H(X)$.
   b) All typical sequences in $\mathcal{A}_\epsilon^{(n)}$ have the same probability up to a negligible factor.
   c) If $W \perp\!\!\!\perp X$, $X \perp\!\!\!\perp Y$ and $W \perp\!\!\!\perp Y$ then (W, X, Y) are mutually (jointly) independent.
   d) Let $P(y|x)$ characterize a discrete memoryless channel (DMC) with input and output alphabets $\mathcal{X} = \{1, 2, \ldots, m\}, \mathcal{Y} = \{1, 2, \ldots, n\}$, respectively. Assume for all $y \in \mathcal{Y}$ that $P(y|X = 1) = P(y|X = 2) = \cdots = P(y|X = m)$. The capacity of this DMC is 0.

2) **Channels with dependence between letters (32 points)**: Consider a channel over a binary alphabet that takes in two bit symbols and produces a two bit output, as determined by the following mapping: $00 \to 01$, $01 \to 10$, $10 \to 11$, and $11 \to 00$. Thus if the two bit sequence 01 is the input to the channel, the output is 10 with probability 1. Let X_1, X_2 denote the two input symbols and Y_1, Y_2 denote the corresponding output symbols.
   a) Find $I(X_1, X_2; Y_1, Y_2)$ as a function of the input distribution on the four possible pairs of inputs.
   Remark: For convenience, denote the input distribution on the four possible pairs by p_{00}, p_{01}, p_{10}, and p_{11}.
   b) What is the capacity of a pair of transmissions on this channel?
   c) Calculate $I(X_1; Y_1)$ under the maximizing input distribution.
   d) Does the maximizing input distribution that you found necessarily maximize the mutual information between the individual symbols, and their corresponding outputs? Explain your answer.

3) **Classifier (18 points)** A classifier can be written as $H(x) = sign(F(x))$, where $H(x) : \mathbb{R}^d \to \{-1, 1\}$ and $F(x) : \mathbb{R}^d \to \mathbb{R}$. Alternatively, for a given x we want to label it 1 or -1 (instead of 1 or 0 as we saw in class).

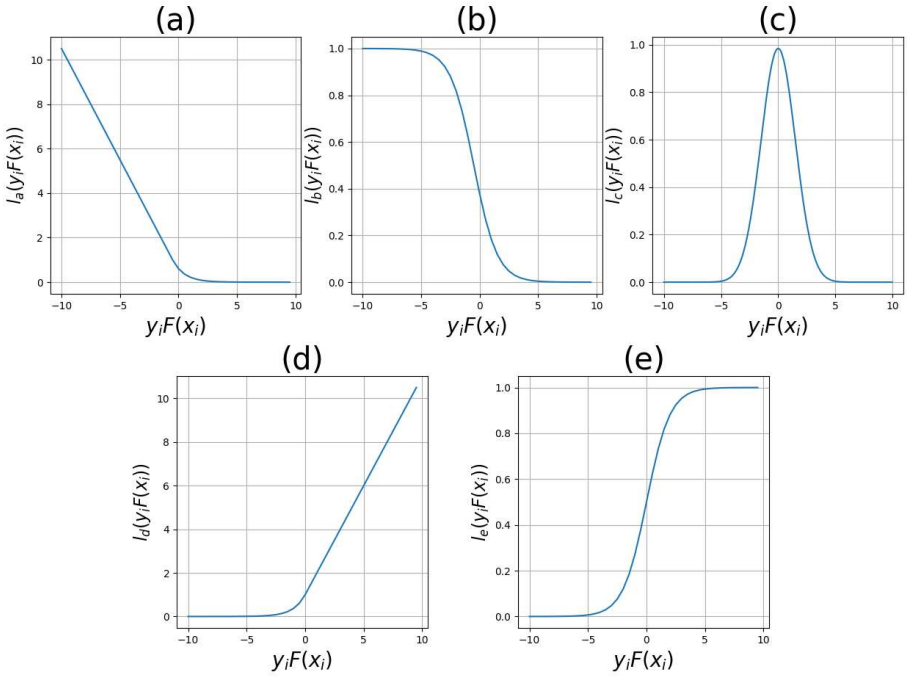

Fig. 1: Loss functions: The x axis is $y_iF(x_i)$, and the y axis is $l(y_iF(x_i))$.

The labeling is done by the sign of a given scalar function $F(x)$, which does not matter for us.
To obtain the parameters in $F(x)$, we need to **minimize** the loss function averaged over the training set $\{x_i, y_i\}_{i=1}^N$, $Loss = \frac{1}{N}\sum_{i=1}^N \mathcal{L}(y_i, F(x_i))$, where $x_i \in \mathbb{R}^d$ is the features vector, $y_i \in \mathbb{R}$ is the label, and $\mathcal{L}$ is the loss function. The loss function is defined as follows:

$$\mathcal{L}(y_i, F(x_i)) = l\left(y_i \cdot F(x_i)\right), \quad l(\cdot) : \mathbb{R} \to \mathbb{R}$$

For **example**, for linear classifiers, $F(x_i) = w_0 + \sum_{j=1}^d w_j x_{i,j}$, and $y_i \cdot F(x_i) = y_i \cdot (w_0 + \sum_{j=1}^d w_j x_{i,j})$.
Notation - $x_{i,j}$ is the j element in the features vector x_i.

a) Which loss functions from Figure 1 are appropriate to use in classification? For the ones that are not appropriate, explain why. In general, what conditions does l have to satisfy in order to be an appropriate loss function?
b) Of the loss functions appropriate to use in classification (your previous section's answer), which one is the most immune to outliers? In other words, which loss function doesn't give large penalty for greater errors? Provide an explanation.
Remark: In the given model, x is outlier if $yF(x)$ is too negative or too positive.

4) **Auto-Encoders (36 points)** Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_x, \Sigma_x)$. We would like to create a generative model of $\mathbf{X}$.
a) (6 points) Assume we use a simple autoencoder (not variational) and X is some single-dimensional random variable (i.e., the encoder $F : \mathbb{R} \mapsto \mathbb{R}$ and so is the decoder $G : \mathbb{R} \mapsto \mathbb{R}$). What would be the optimal choice of F and G for MSE-minimizing?
b) (8 points) In this section we consider the linear case for a variational autoencoder and we will apply the reparametrization trick. We assume that $\boldsymbol{X}$ is m-dimensional. Let:

$$\boldsymbol{\mu_z} = \begin{pmatrix} w_1^\mathsf{T}\mathbf{X} \\ \vdots \\ w_d^\mathsf{T}\mathbf{X} \end{pmatrix} + b, \qquad \Sigma_Z = A \cdot \mathit{diag}(\mathbf{X}), \qquad \mathit{diag}(\begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_m \end{pmatrix}) = \begin{pmatrix} x_1 & 0 & \dots & \dots & 0 \\ 0 & x_2 & 0 & \dots & 0 \\ \vdots & 0 & x_3 & \dots & 0 \\ 0 & \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \dots & \dots & x_m \end{pmatrix}$$

with $w_i \in \mathbb{R}^m$ for $i = 1 \dots d$, $A \in \mathbb{R}^{d \times m}$ and $b \in \mathbb{R}^d$.
What is the distribution of $\boldsymbol{\mu_z}$? What is the distribution of Z given a realization $X = x$?
Reminder: if $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $Y = AX + b$ then $\mathbf{Y} \sim \mathcal{N}(A\boldsymbol{\mu} + b, A\Sigma A^\mathsf{T})$.
c) (5 points) We now focus on the one-dimensional case. In class we had defined the following loss for the variational-autoencoder:

$$\mathcal{L} := \underbrace{\mathbb{E}_{q(z)}\left[\frac{(x - f(z))^2}{2c}\right]}_{:=\mathsf{MSE}} + \underbrace{D_{KL}(\mathcal{N}(g(x), h^2(x)), \mathcal{N}(0,1))}_{:=\mathsf{D}} \tag{1}$$

where our goal is to apply $\min_{f,g,h} \mathcal{L}$. We define the functions as follows:

$$f(z) = z, \quad g(x) = \sum_{i=1}^m w_i x^i, \quad h(x) = \sum_{j=1}^m \exp(-u_j x)$$

Calculate $\frac{\partial \mathsf{MSE}}{\partial w_i}$ for some general i.
d) (8 points) The KL-divergence between two Gaussians $G_1 = \mathcal{N}(\mu_1, \sigma_1^2)$ and $G_2 = \mathcal{N}(\mu_2, \sigma_2^2)$ is given by:

$$D_{KL}(G_1, G_2) = \log\frac{\sigma_2}{\sigma_1} + \frac{\sigma_1^2 + (\mu_1 - \mu_2)^2}{2\sigma_2^2} - \frac{1}{2}.$$

Calculate $\frac{\partial \mathsf{D}}{\partial w_i}$ for some general i.
e) (3 points) Given some fixed learning rate μ, write down the SGD update rule for the models weights $\{w_i\}_{i=1}^m$.
f) (6 points) This section is independent of the previous ones. Assume we want to constraint our latent vector Z to have similar statistical characteristics as some other random vector Y. Propose a modification for the loss in (1) to obtain this request. Suggest a method to control how strongly we want to impose this constraint.

Good Luck!