# Machine Learning for Communication Systems
# MALCOM Spring 2020
# EURECOM

Final Examination (2 hours)

- This exam is open books and documents. Use of internet resources, web search, and any sort of communication are strictly forbidden during the exam.
- In the short questions and the exercises, partial credit will be given for good explanations even if you have not provided the complete correct answer. Therefore, please explain your idea, reasoning, derivations, calculations, etc., even if you are unsure of your answers.

**Examination Honor Code:** This take-home exam has an Honor Code. You may not consult or collaborate with anyone about the questions. Such collaboration is a violation of the Honor Code.

I attest on my honor that I have not given or received any unauthorized assistance on this examination.

Signature: _____

The exam contains **8** pages including this cover page.

## Multiple Choice Questions (20 points)

*For each of the following questions, circle the letter of your choice. There is only ONE correct choice unless explicitly mentioned. No explanation is required. There is no penalty for a wrong answer.*
*Alternatively, you can enter and send your answers in the separated word file provided together with the exam sheet.*

**Question 1**: Which of the following models can be used in unsupervised learning? (Check all that apply)

(i) k-means
(ii) SVM
(iii) Linear regression
(iv) Autoencoder

**Question 2**: The gradient estimated during a step of mini-batch gradient descent has on average a lower bias when the data is i.i.d. (independent and identically distributed).

(i) True
(ii) False

**Question 3**: You are doing full batch gradient descent using the entire training set (not stochastic gradient descent). It is necessary to shuffle the training data.

(i) True
(ii) False

**Question 4**: Consider a trained logistic regression for a wireless communication system. Its weight vector is *W* and its test accuracy on a given data set is *C*. Assuming there is no bias, dividing *W* by 2 will change the test accuracy.

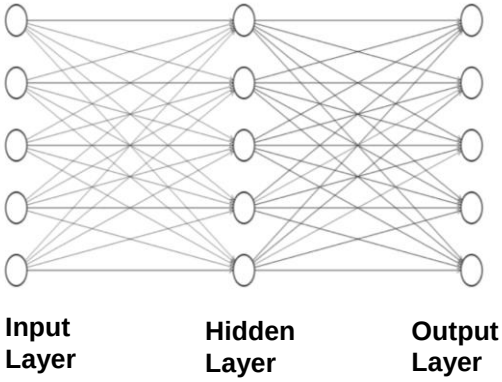
(i) True
(ii) False

**Question 5**: Which of the following is true about the vanishing gradient problem? (Check all that apply)

(i) Tanh is usually preferred over sigmoid because it does not suffer from vanishing gradients.
(ii) Leaky ReLU is less likely to suffer from vanishing gradients than sigmoid.

(iii) Vanishing gradient causes deeper layers to learn more slowly than earlier layers.
(iv) Weight initialization (e.g. Xavier, He) can help prevent the vanishing gradient problem.
(v) None of the above.

**Question 6**: A 2-layer neural network with 5 neurons in each layer has a total of 70 parameters (i.e. weights and biases)




(i) True
(ii) False

**Question 7**: Which of the following propositions are true about a convolutional (CONV) layer? (Check all that apply.)
(i) The number of weights depends on the depth of the input volume.
(ii) The number of biases is equal to the number of filters.
(iii) The total number of parameters depends on the stride.
(iv) The total number of parameters depends on the padding.

**Question 8**: Consider a Generative Adversarial Network (GAN) that successfully produces images of lions. Which of the following propositions is false?
(i) The generator aims to learn the distribution of lion images.
(ii) The discriminator can be used to classify images as lion vs. non-lion.
(iii) After training the GAN, the discriminator loss eventually reaches a constant value.
(iv) The generator can produce unseen images of lions.

**Question 9**: in which distributed system architecture, I can have model consistency at each training epoch/iteration. (Choose all that apply)

(i) Parameter server (PS) with synchronous SGD
(ii) PS with asynchronous SGD
(iii) All reduce
(iv) Gossip/decentralized
(v) None of the above

**Question 10**: In 1-bit gradient quantization, ….

(i) the quantization function is unbiased
(ii) a gradient descent method will always converge
(iii) none of the above

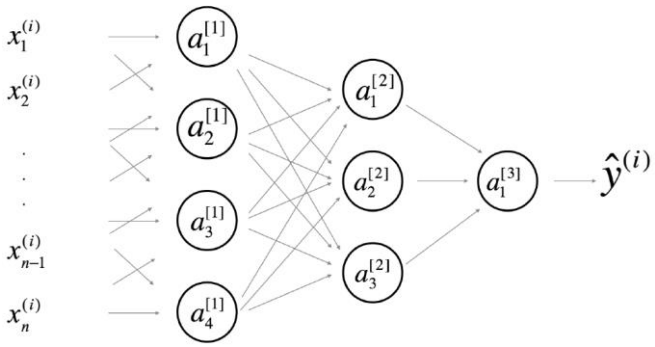
**Question 11**: You are designing a model to predict the presence (labeled 1) or absence (labeled 0) of a pedestrian, to prevent accidents in an autonomous car. Which of the following evaluation metrics would you choose to use?

(i) Loss function value
(ii) $\text{metric A} = \frac{\text{True positive examples}}{\text{Total positive examples}}$
(iii) $\text{metric B} = \frac{\text{True positive examples}}{\text{Total predicted positive examples}}$
(iv) Accuracy

**Question 12**: Which of the following is true, given the optimal learning rate?

(i) Batch gradient descent is always guaranteed to converge to the global optimum of a loss function.
(ii) Stochastic gradient descent is always guaranteed to converge to the global optimum of a loss function.
(iii) For convex loss functions, stochastic gradient descent is guaranteed to eventually converge to the global optimum while batch gradient descent is not.
(iv) For convex loss functions, both stochastic gradient descent and batch gradient descent will eventually converge to the global optimum.
(v) For convex loss functions, neither stochastic gradient descent nor batch gradient descent are guaranteed to converge to the global optimum.
(vi) For convex loss functions, batch gradient descent is guaranteed to eventually converge to the global optimum while stochastic gradient descent is not.

**Question 13**: You design the following 2-layer fully connected neural network. All activations are sigmoids and your optimizer is stochastic gradient descent. You initialize all the weights and biases to zero and forward propagate an input $x \in \mathbb{R}^{n\times 1}$ in the network. What is the output $\hat{y}$?




(i) -1
(ii) 0
(iii) 0.5
(iv) 1

**Question 14**: Mini-batch gradient descent is a better optimizer than full-batch gradient descent to avoid getting stuck in saddle points.

(i) True
(ii) False

## Short Answers Questions (40 points)

*For the questions in this section, please be concise and provide 2-3 sentences in your responses.*

**Question 1**: Why do the layers in a deep neural network architecture need to be non-linear?

**Question 2**: You're solving a binary classification task for a WiFi modulation signal problem. The final two layers in your network are a ReLU activation followed by a sigmoid activation. What will happen?

**Question 3:** Softmax takes in an *n*-dimensional vector $x$ and outputs another *n*-dimensional vector $y$ (scores):

$$y_i = \frac{e^{x_i}}{\sum_k e^{x_k}}$$

The objective of this question is to compute the gradient of $y$ with respect to $x$.
Let $\delta_{ij} = \frac{\partial y_i}{\partial x_j}$. Derive an expression for (1) $\delta_{ii}$, and (2) for $\delta_{ij}$ for $i \neq j$.

**Question 4**: Let $s_k$ be the score for a specific class *k* and $\theta$ is a constant that we substract from all scores of a sample. Show that $softmax(s_k)$ is equal to $softmax(s_k - \theta)$.
Explain what this property of the softmax function implies and comment on why this property is useful when training neural networks.

**Question 5**: You design a fully connected neural network architecture for an end-to-end communication system, where all activation functions are sigmoids. You initialize the weights with large positive numbers. Is this a good idea? Explain your answer.

**Question 6**: We would like to implement a deep wireless transmitter by training a fully-connected neural network with 5 hidden layers, each with 10 hidden units. The input is a 20-dimensional vector and the output is a scalar. What is the total number of trainable parameters in your network?

**Question 7**: We would like to design an end-to-end communication system using an autoencoder. For that, we try to find a useful representation $\boldsymbol{r} \in \mathbb{R}^n$ of the input $\boldsymbol{s} \in \mathbb{R}^k$ at some intermediate layer through learning to reproduce the input at the output. If $k = 10$, how much $n$ should be?

**Question 8**: Consider a distributed learning system using a parameter server. If we compare *T* iterations of 1-sample SGD with *T*/*B* iterations of mini-batch (B-sample) SGD, B-sample SGD has better progress than 1-sample SGD. True or False? Please comment your answer.

**Question 9**: Consider a distributed learning system using a parameter server. We would like to train a useful signal vs. useless signal (noise) classifier using mini-batch gradient descent. We have already split your dataset into train and test sets and the classes are balanced. The server takes the training set, split it into batches and send each batch to the workers in a serial way (one batch after the other). For example, for *K* workers, worker 1 receives the first 1/*K* examples, worker 2 gets the second 1/*K* examples, etc. Within the training set, all signal examples are ordered in such a way that all useful signals come first and all noise signals come after. Explain what will happen in the performance of this distributed learning system. How we can improve its performance?

**Question 10**: A binary classification problem could be solved with the two approaches described below:

*Approach 1:* Simple Logistic Regression (one neuron)
Your output will be $\hat{y} = \sigma(W_1 x + b_1)$
Classify as 0 if $\hat{y} \leq 0.5$ and 1 otherwise.

*Approach 2:* simple softmax regression (two neurons)
Your output will be $\hat{y} = softmax(W_2 x + b_2) = [\hat{y}_1, \hat{y}_2]^T$
Classify as 0 if $\hat{y}_1 \geq \hat{y}_2$ and 1 otherwise.

Approach 2 involves twice as many parameters as approach 1.
Can approach 2 learn more complex models than approach 1?
If yes, give the parameters $(W_2, b_2)$ of a function that can be modeled by approach 2 but not by approach 1. If no, show that $(W_2, b_2)$ can always be written in terms of $(W_1, b_1)$.

## A Practical Problem (10 points)

You are the project manager for implementing ML-based telecom products in a famous company and you need to decide on the budget investment. Your company is about to launch two products:

- ML-based transceiver "MLC1.0", which is based on radio signal binary classification.
- ML-based transceiver "MLC2.0", which is based on variational autoencoders.

The hourly cost for labeling data is 10 euros/engineer and the hourly cost for testing/validation is 20 euros/engineer for MLC1.0 and 30 euros/engineer for MLC2.0.
For achieving the same performance (e.g., accuracy) in both products, we need to perform 30 hours of labeling and *H* hours of validation.

Your company allows you to hire up to 2 engineers. How many hours *H* you need to spend on validation so that MLC1.0 production costs less than MLC2.0?

## Exercise – Federated Learning (30 points)

Consider a federated learning system at the wireless edge, in which the data distribution of the edge devices is imbalanced (non i.i.d.), i.e., the local data follows different distributions. This will introduce biases in the model training and cause a decrease in accuracy of federated learning applications.
To improve the performance of the system, we introduce a controller (scheduler) that selects which edge device is allowed to send its updated model back to the server.

We assume that there are two possible outcomes, $x = 0$ and $x = 1$, (probability space $\mathcal{X} = \{0,1\}$). The data of device 1 follows a binomial distribution with $n = 1$ and $p = 0.2$, whereas the data of device 2 follows a binomial distribution with $n = 1$ and $p = 0.4$.

The scheduler will select the edge device whose Kullback-Leibler divergence from the uniform distribution to its own distribution is smaller. Which of the two edge devices will not send its model back to the server?

*<u>Reminder</u>:*

- *The binomial distribution with parameters n and p is a discrete probability distribution with pmf (probability mass function)*
$$f(k,n,p) = \mathbb{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}.$$
- *For discrete probability distributions P and Q defined on the same probability space $\mathcal{X}$, the Kullback–Leibler divergence from Q to P is defined to be*
$$D_{KL}(P||Q) = \sum_{x \in \mathcal{X}} P(x)\, log\left(\frac{P(x)}{Q(x)}\right).$$