CSC 321H5 S 2020 Midterm
Duration — 50 minutes
Aids allowed: none

Student Number:

0	0	0	0	0	0	0	0	0	0
1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2
3	3	3	3	3	3	3	3	3	3
4	4	4	4	4	4	4	4	4	4
5	5	5	5	5	5	5	5	5	5
6	6	6	6	6	6	6	6	6	6
7	7	7	7	7	7	7	7	7	7
8	8	8	8	8	8	8	8	8	8
9	9	9	9	9	9	9	9	9	9

UTORid: _____

Last Name: _____ **First Name:** _____

Lecture Section: L0101 Test Version: A Instructor: Pouria Fewzee

---

*Do **not** turn this page until you have received the signal to start.*
(Please fill out the identification section above, and read the instructions below.)
Good Luck!

This test consists of 5 questions on 7 pages (including this page). *When you receive the signal to start, please make sure that your copy is complete.*

Fill in the identification section above and bubble in your student number in either pen or pencil. Answer each question directly on the test paper, in the space provided, using either a blue or black pen or a pencil. If you need more space for one of your solutions, use the extra pages at the end of the test paper and indicate clearly the part of your work that should be marked.

Write up your solutions carefully! Marks cannot be awarded for solutions that are not understandable by the grader, and may be deducted if you make false assertions. If you are giving only one part of an answer, indicate clearly what you are doing. Part marks might be given for incomplete solutions.

1: _____/ 7

2: _____/ 8

3: _____/ 7

4: _____/ 4

5: _____/ 4

TOTAL: _____/30

Total Pages = 7

Question 1. [7 MARKS]

Circle the best answer for each of the questions below. Do not circle more than one answer per question.

Part (a) [1 MARK]

Which of the following is an issue preventing neural networks from being more widely used?

(A) They are not easily interpretable.
(B) They don't perform well in general.
(C) They don't perform well on image-based tasks.
(D) They don't perform well on sequence-based tasks.
(E) They are not universal approximators.

Part (b) [1 MARK]

Which of the following about the k-Nearest Neighbour model is true?

(A) It is always better to choose a smaller k.
(B) A k-Nearest Neighbour model has exactly k trainable parameters.
(C) The training accuracy of a 3-Nearest Neighbour model is usually higher than a 1-Nearest Neighbour model.
(D) The decision boundary of a k-Nearest Neighbour is linear.
(E) As k increases, the bias (in the sense of bias-variance decomposition) increases.

Part (c) [1 MARK]

Which of the following about logistic regression is **true**?

(A) If every training example is classified correctly, then the training cost is zero.
(B) If every training example is classified incorrectly, then the training cost is infinite.
(C) The training cost is always positive.
(D) The training cost is always negative.
(E) None of the above are true.

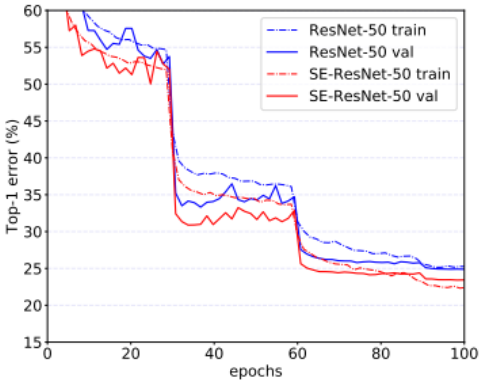
Part (d) [1 MARK]

Gradient descent and backpropagation are used to update the following:

(A) The activations only.
(B) The layers and activations.
(C) The layers only.
(D) The parameters and activations.
(E) The parameters only.

Part (e) [1 MARK]

This learning curve shows the epoch count on the x-axis, and the error rate (100% minus accuracy rate) on the y-axis. What is the most likely setting change in epochs 25 and 60 that could have produced this learning curve shape?



(A) The learning rate was decreased.
(B) The learning rate was increased.
(C) The batch size was increased.
(D) The batch size was decreased.
(E) Weight decay was introduced.

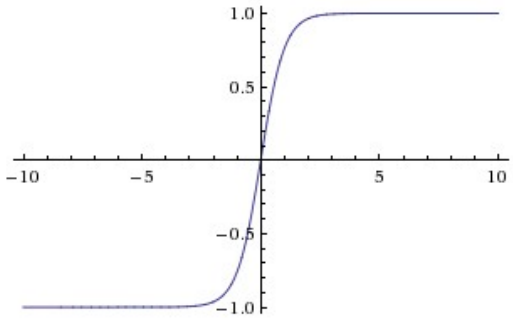
Part (f) [1 MARK]

Which of the following does **not** help prevent overfitting?

(A) Decreasing the number of hidden units in a multi-layer perceptron.
(B) Using a larger training set.
(C) Using a smaller batch size.
(D) Using stochastic gradient descent with momentum.
(E) Both (C) and (D).

Part (g) [1 MARK]

What activation function is this?



(A) ReLU
(B) tanh
(C) sigmoid
(D) softmax
(E) None of the above.

Question 2. [8 MARKS]

Suppose we somehow know that the weights for a two-dimensional regression problem should sum up to 1, so that $w_1 = m$ and $w_2 = 1 - m$. That is, the output y should be a weighted average of the inputs x_1 and x_2. The model and loss function are as follows:

$$\begin{aligned} w_1 &= m \\ w_2 &= 1 - m \\ y &= w_1 x_1 + w_2 x_2 \\ \mathcal{L} &= \frac{1}{2}(y - t)^2 \end{aligned}$$

Part (a) [3 MARKS]

Determine the backprop update rules which let you compute the derivative $\frac{\partial \mathcal{L}}{\partial m}$. Your equations should refer to previously computed values (e.g. your formula for $\overline{w_1}$ should refer to $\overline{y}$). You do not need to show your work, but it may help you get partial credit.

$$\overline{\mathcal{L}} = 1$$

$$\overline{y} =$$

$$\overline{w_1} =$$

$$\overline{w_2} =$$

$$\overline{m} =$$

Part (b) [1 MARK]

Suppose we would like to penalize values of m far away from $\frac{1}{2}$, so our regularized loss function becomes $\mathcal{L} = \frac{1}{2}(y - t)^2 + \lambda(\frac{1}{2} - m)^2$. How would this modification change your answers from part (a)?

Part (c) [4 MARKS]

We would like to use (full batch) gradient descent **with momentum** to optimize m. Explain how we can do so, assuming that there is a function you can call to compute $\frac{\partial \mathcal{E}}{\partial m}$. Write enough detail so that someone can code an algorithm based on your explanation.

Question 3. [7 MARKS]

For this question, we will work with the following 3×3 convolutional kernel.

0.5	1.0	1.0
-0.5	0.0	0.5
0.5	0.0	-0.5

Part (a) [3 MARKS]

What is the output if we apply the above convolution on the following input, **with the stride set to 2 and padding set to 1**? Assume that the bias is 0, and that the number of input and output channels are both 1. Draw the output feature map, clearly showing the shape of the output feature map, and each unit's value.

1	1	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	0	0
0	0	1	1	1

Part (b) [2 MARKS]

How many parameters are in the PyTorch layer:
`nn.Conv2d(in_channels=5, out_channels=10, kernel_size=3, stride=2, padding=0)`?

Part (c) [1 MARK]

How many parameters are in the PyTorch layer `nn.MaxPool2d(kernel_size=2, stride=2)`?

Part (d) [1 MARK]

How many parameters are in the PyTorch layer `nn.Linear(in_features=3, out_features=7)`?

Question 4. [4 MARKS]

Part (a) [2 MARKS]

Consider the sign function $sign(x) = \frac{x}{|x|}$. Explain why we can't use this function as an activation function to train a multi-layer perceptron.

Part (b) [2 MARKS]

What does the term "checkpointing" mean in the context of neural network training? How does "checkpointing" relate to the prevention of overfitting?

Question 5. [4 MARKS]

Part (a) [1 MARK]

What is a saddle point?

Part (b) [3 MARKS]

Show that in a multi-layer perceptron, the origin in the parameter-space is a saddle point. In other words, show that the point in the parameter space where all our weights are biases are zero is a saddle point of the cost function.

Assume the ReLU activations are used, so that each layer's computation is as follows:

$$\mathbf{z}^{(k)} = W^{(k)}\mathbf{h}^{(k-1)} + b^{(k)}$$
$$\mathbf{h}^{(k)} = \text{ReLU}(\mathbf{z}^{(k)})$$

[Use the space below for rough work. This page will not be marked unless you clearly indicate the part of your work that you want us to mark.]