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: L0102 Test Version: B Instructor: Lisa Zhang

---

*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 10 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

Question 1. [7 MARKS]

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

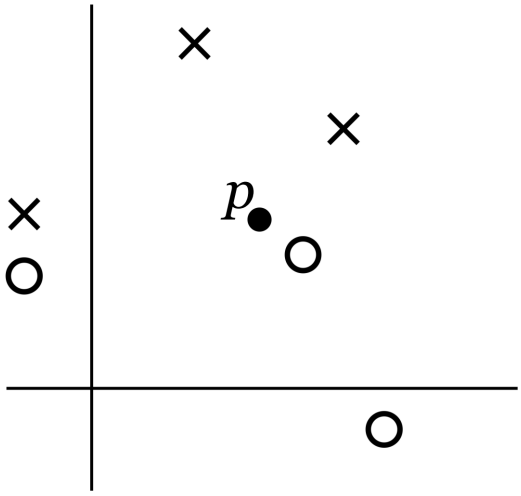
Part (a) [1 MARK]

Which of the following problems is most suited to solving using Machine Learning?

(A) Determine whether more than half of an image is green.
(B) Determine whether an array of integers is in sorted order.
(C) Determine whether a word is mispelled.
(D) Determine whether a sentence is about Queen Elizabeth.
(E) Both (C) and (D).

Part (b) [1 MARK]

We would like to use 1-Nearest Neighbour to classify point p using the data below. What is our prediction if we use cosine distance? Euclidean distance?



(A) Cosine distance: O, Euclidean distance: O
(B) Cosine distance: X, Euclidean distance: O
(C) Cosine distance: O, Euclidean distance: X
(D) Cosine distance: X, Euclidean distance: X

Part (c) [1 MARK]

Which of the following about multi-layer perceptrons is **false**?

(A) Multilayer feed-forward neural nets with ReLU activation functions are universal approximators.
(B) Two-layer feed-forward neural nets with ReLU activation functions are universal approximators.
(C) Multilayer feed-forward neural nets with linear activation functions are universal approximators.
(D) Multilayer feed-forward neural nets with sigmoid activation functions are universal approximators.
(E) None of the above are false.

Part (d) [1 MARK]

Assuming a learning rate that is not too large, which of the following statements about learning curves is true?

(A) If the batch size is the same as the size of the training set, then the training cost will always decrease in each iteration.

(B) A larger batch size is typically required for larger inputs (e.g. larger images).

(C) If the batch size is 1, then the training cost will always decrease in each iteration.

(D) Training a language model requires a larger batch size than training a supervised learning classifier with a similar number of parameters.

(E) If the training accuracy is not decreasing, then neither is the training cost.

Part (e) [1 MARK]

Which of the following about weight decay is true?

(A) Including weight decay generally reduces the training cost.

(B) Weight decay directly penalizes large activations.

(C) Weight decay should not be used with the optimizer Adam.

(D) Weight decay can help revive a "dead" or "saturated" neuron.

(E) Weight decay helps get out of saddle points.

Part (f) [1 MARK]

Which of the following about a high variance model (in the context of bias-variance tradeoff) is true, compared to a high bias model?

(A) A high variance model is more prone to underfitting.

(B) A high variance model requires more training data to train.

(C) A high variance model will have a higher training accuracy.

(D) A high variance model should be trained with a smaller batch size.

(E) Both (B) and (C) are true.

Part (g) [1 MARK]

What is the value of `softmax([2., 2., 1.])`?

(A) `[0.8808, 0.8808, 0.7311]`

(B) `[0.5, 0.5, 0.25]`

(C) `[0.4, 0.4, 0.2]`

(D) `[0.5, 0.5, 0.]`

(E) `[0.4223, 0.4223, 0.1554]`

Question 2. [8 MARKS]

Suppose we somehow know that the weights for a two-dimensional classification problem should be positive. Moreover, we should parameterize $w_1 = e^m$ and $w_2 = e^{-m}$. The model and loss function are as follows:

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

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 $\overline{z}$ should refer to $\overline{y}$). You do not need to show your work, but it may help you get partial credit.

You can also use the fact that $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ without justification.

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

$$\overline{y} = \overline{\mathcal{L}} \frac{y - t}{y(1-y)}$$

$$\overline{z} =$$

$$\overline{w_1} =$$

$$\overline{w_2} =$$

$$\overline{m} =$$

Part (b) [1 MARK]

Suppose we would like to penalize values of m far away from 0, so our regularized loss function becomes $\mathcal{L} = \frac{1}{2}(y-t)^2 + \lambda m^2$. How would this modification change your answers from part (a)?

Part (c) [2 MARKS]

We would like to use (full batch) gradient descent 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.

Part (d) [2 MARKS]

Suppose we have $m = 0$. Draw the decision boundary $y = 0.5$ in the input space, and annotated which side of the prediction will have positive prediction (predict $t = 1$) and negative prediction (predict $t = 0$).

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]

Explain why if we use batch normalization in PyTorch, we need to run the code `model.train()` before running backpropagation and `model.eval()` before computing the model accuracy.

Part (b) [2 MARKS]

In Project 1, we used gradient checking ensure that our gradients are computed correctly. Explain what "gradient checking" means, and how and why it works.

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.]