UNIVERSITY OF BRISTOL

January 2024 Examination Period

School of Engineering, Mathematics and Technology

Third Year Examination for the Degree of
Bachelor of Science and Master of Engineering

EXAM PAPER CODE: EMAT-31530J
UNIT CODE: EMAT-31530

Introduction to AI

TIME ALLOWED:
2 Hours

Answers to EMAT-31530: Introduction to AI

Answer all 15 questions.
All questions should be answered on the MCQ sheet.
Each question has exactly *one* correct answer.
All answers will be used for assessment.
The maximum for this paper is *100 marks*.
The exam is closed-book (so no additional materials are allowed).
You may write workings out on the exam paper, and blank pages are provided at the end for this purpose. These workings out will not be collected or marked. You must enter your answers on the provided answer sheet only.

Other Instructions:

You may use a calculator.
Only non-programmable calculators may be used.

Using the computer marked sheets:

All questions in this examination will be computer marked. It is crucial that you follow the instructions below and fill in the red coloured answer sheets carefully. Use a PENCIL (not pen). Use a pencil eraser to correct any errors you make.

Insert your candidate name, title of the examination, unit code and date onto the answer sheets in the relevant boxes. Fill in your student number:

TURN OVER ONLY WHEN TOLD TO START WRITING

Help Formulas:

2D Convolution:

$$a_{x,y} = \sum_{\delta_x=-1}^{1} \sum_{\delta_y=-1}^{1} h_{x+\delta_x, y+\delta_y} W_{\delta_x, \delta_y}$$

Optimal weights for linear regression:

$$\mathbf{w}^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Matrix inversion:

$$\begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}.$$

Matrix Determinant:

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc$$

Multi-class cross-entropy loss:

$$\text{cross entropy}(\boldsymbol{\ell}, y) = -\ell_y + \log \sum_c \exp(\ell_c). \tag{1}$$

Binary cross-entropy loss:

$$\text{cross entropy}(\ell, y) = \begin{cases} \log(1 + e^{-\ell}) & \text{if } y = 1 \\ \log(1 + e^{\ell}) & \text{if } y = 0 \end{cases} \tag{2}$$

ReLU:

$$\text{relu}(x) = \begin{cases} x & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Q1. Consider the following code:

```
import torch as t

A = t.randn(2,3)
B = t.randn(1,4,1)
C = A+B
```

What is the shape of `C`?

A. `t.Size([4,2,3])`

B. `t.Size([2,4,3])`

C. `t.Size([2,3,4])`

D. `t.Size([3,2,4])`

E. We don't get to `C`: there's a shape error.

[6 marks]

Q2. Given the following code:

```
import torch as t

A = t.randn(2,3,4)
b = t.randn(4)
C = A*B
D = C.mean(1)
```

which is the equivalent mathematical expression for computing D?

A. $D_{jk} = \frac{1}{2}\sum_{i=1}^{2} A_{ijk}b_i$

B. $D_{ij} = \frac{1}{4}\sum_{k=1}^{4} A_{ijk}b_k$

C. $D_{ik} = \frac{1}{3}\sum_{j=1}^{3} A_{ijk}b_i$

D. $D_{ik} = \frac{1}{3}\sum_{j=1}^{3} A_{ijk}b_k$

E. $D_{ik} = \frac{1}{3}\sum_{j=1}^{3} A_{ijk}b_j$

[6 marks]

Solution:

```
A.shape = t.Size([2,3,4])
b.shape = t.Size([4])
C.shape = t.Size([2,3,4])
C.mean(1).shape = t.Size([2,4])
```

Final line takes the mean over the middle dimension.

Q3. For the data in the table, fit a model of the form $\hat{y} = w_1 + w_2 x$

x	y
0	3.2
1	1.9
2	1.2
3	-0.1
4	-0.9

A. $w_1 = 3.03$ and $w_2 = -1.02$

B. $w_1 = 3.03$ and $w_2 = -0.98$

C. $w_1 = 3.10$ **and** $w_2 = -1.02$

D. $w_1 = 3.10$ and $w_2 = -0.98$

E. None of the above

[6 marks]

Q4. Consider the following choices of loss function. Which of the following is **not** a sensible loss function? By "sensible loss function", we mean that the loss should always go down as the predictions, $\hat{y}(x_i)$ get closer to the data, y_i.

A. $\mathcal{L} = \sum_i |\hat{y}(x_i) - y_i|$

B. $\mathcal{L} = \sum_i (\hat{y}(x_i) - y_i)^2$

C. $\mathcal{L} = \sum_i (\hat{y}(x_i) - y_i)^3$

D. $\mathcal{L} = \sum_i (\hat{y}(x_i) - y_i)^4$

E. $\mathcal{L} = -\sum_i \frac{1}{1+(\hat{y}(x_i)-y_i)^2}$

Solution: $\mathcal{L} = \sum_i (\hat{y}(x_i) - y_i)^3$

To minimize this loss, we should just make $\hat{y}(x_i)$ as big and negative as possible. Which will give nonsensical predictions.

Q5. Compute the cross entropy loss for binary classification, where the logits is,

$$\text{logits}(x) = -3 + x + x^2 \tag{3}$$

with data,

x	y
0	0
1	0
2	0
3	1
4	1

[6 marks]

A. 3.38

B. 3.41

C. 3.45

D. 3.51

E. 3.67

Q6. Consider the following model for a discrete random variable, x, that can take on three values: 1, 2 or 3.

$$P(x=1) = \theta$$
$$P(x=2) = \tfrac{1}{2} - \tfrac{1}{2}\theta$$
$$P(x=3) = \tfrac{1}{2} - \tfrac{1}{2}\theta$$

Consider the following dataset:

(cont.)

value of x	frequency
1	4
2	3
3	2

What is the gradient of $\log P(x_1, \ldots, x_9)$ wrt θ?

A. $\frac{5}{\theta} - \frac{4}{1-\theta}$

B. $\frac{5}{\theta} + \frac{4}{1-\theta}$

C. $\frac{4}{\theta} - \frac{5}{1-\theta}$

D. $\frac{4}{\theta} + \frac{5}{1-\theta}$

E. None of the above.

Solution:

$$\log P(\mathbf{x}) = 4 \log P(x=1) + 3 \log P(x=2) + 3 \log P(x=3) \tag{4}$$
$$\log P(\mathbf{x}) = 4 \log \theta + 3 \log(\tfrac{1}{2} - \tfrac{1}{2}\theta) + 2 \log(\tfrac{1}{2} - \tfrac{1}{2}\theta) \tag{5}$$
$$\log P(\mathbf{x}) = 4 \log \theta + 5 \log(\tfrac{1}{2} - \tfrac{1}{2}\theta) \tag{6}$$

To compute the derivative, we use the chain rule with,

$$u = \tfrac{1}{2} - \tfrac{1}{2}\theta \tag{7}$$

Thus,

$$\frac{\partial u}{\partial \theta} = -\tfrac{1}{2} \tag{8}$$
$$\frac{\partial \log P(\mathbf{x})}{\partial \theta} = 4\tfrac{1}{\theta} + 5\frac{\partial \log u}{\partial u}\frac{\partial u}{\partial \theta} \tag{9}$$
$$\frac{\partial \log P(\mathbf{x})}{\partial \theta} = 4\tfrac{1}{\theta} + 5\tfrac{1}{u}(-\tfrac{1}{2}) \tag{10}$$
$$\frac{\partial \log P(\mathbf{x})}{\partial \theta} = 4\tfrac{1}{\theta} - \frac{5}{2(\frac{1}{2} - \frac{1}{2}\theta)} \tag{11}$$
$$\frac{\partial \log P(\mathbf{x})}{\partial \theta} = 4\tfrac{1}{\theta} - \frac{5}{1-\theta} \tag{12}$$
$$\tag{13}$$

Q7. Which of the following statements about how neural networks are implemented in PyTorch is **FALSE**:

A. PyTorch typically represents neural network layers as classes.

B. PyTorch neural network layers live in the `torch.nn`.

C. PyTorch's neural netowrk layers almost all subclass `torch.nn.Module`.

D. PyTorch neural network layers can be called as-if they were functions.

E. To represent a neural network layer as a class in PyTorch, you must subclass `torch.nn.Module`.

Solution: You can define your own classes that behave like `torch.nn.Module`'s, and we do that in the course notebooks!

Q8. Compute the backward pass for the following nonlinearity,

$$h = \begin{cases} a^3 & \text{if } a > 0 \\ 0 & \text{otherwise} \end{cases}$$

i.e. compute $\frac{d\mathcal{L}}{da}$ in terms of $\frac{d\mathcal{L}}{dh}$.

A. $\frac{d\mathcal{L}}{da} = 3a^2 \frac{d\mathcal{L}}{dh}$

B. $\frac{d\mathcal{L}}{da} = 3h^2 \Theta(h) \frac{d\mathcal{L}}{dh}$

C. $\frac{d\mathcal{L}}{da} = 3h \Theta(h) \frac{d\mathcal{L}}{dh}$

D. $\frac{d\mathcal{L}}{da} = 3\Theta(a) a^2 \frac{d\mathcal{L}}{dh}$

E. $\frac{d\mathcal{L}}{da} = 3h \Theta(a) \frac{d\mathcal{L}}{dh}$

remember that $\Theta(a)$ is the Heaviside step function,

$$\Theta(a) = \begin{cases} 1 & \text{if } a > 0 \\ 0 & \text{otherwise} \end{cases}.$$

[7 marks]

Solution:

$$\frac{d\mathcal{L}}{da} = \frac{\partial h}{\partial a} \frac{d\mathcal{L}}{dh}$$

$$\frac{\partial h}{\partial a} = \begin{cases} 3a^2 & \text{if } a > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\frac{d\mathcal{L}}{da} = \begin{cases} 3a^2 \frac{d\mathcal{L}}{dh} & \text{if } a > 0 \\ 0 & \text{otherwise} \end{cases}$$

This can be rewritten in terms of the Heaviside step,

$$\frac{d\mathcal{L}}{da} = 3\Theta(a) a^2 \frac{d\mathcal{L}}{dh}$$

Q9. Compute the backward pass for the following operation,

$$Y_{ik} = \sum_j A_{ij} B_{jk} \tag{14}$$

i.e. compute $\frac{d\mathcal{L}}{dA_{\alpha\beta}}$ and $\frac{d\mathcal{L}}{dB_{\alpha\beta}}$ in terms of $\frac{d\mathcal{L}}{dY_{ik}}$.

A. $\frac{d\mathcal{L}}{dA_{\alpha\beta}} = \sum_k \frac{d\mathcal{L}}{dY_{\alpha k}} B_{\beta k}$ and $\frac{d\mathcal{L}}{dB_{\alpha\beta}} = \sum_i \frac{d\mathcal{L}}{dY_{i\beta}} A_{\alpha i}$

B. $\frac{d\mathcal{L}}{dA_{\alpha\beta}} = \sum_k \frac{d\mathcal{L}}{dY_{\alpha k}} B_{\beta k}$ **and** $\frac{d\mathcal{L}}{dB_{\alpha\beta}} = \sum_i A_{i\alpha} \frac{d\mathcal{L}}{dY_{i\beta}}$

C. $\frac{d\mathcal{L}}{dA_{\alpha\beta}} = \sum_k B_{\beta k} \frac{d\mathcal{L}}{dY_{k\alpha}}$ and $\frac{d\mathcal{L}}{dB_{\alpha\beta}} = \sum_i \frac{d\mathcal{L}}{dY_{i\beta}} A_{\alpha i}$

D. $\frac{d\mathcal{L}}{dA_{\alpha\beta}} = \sum_k B_{\beta k} \frac{d\mathcal{L}}{dY_{k\alpha}}$ and $\frac{d\mathcal{L}}{dB_{\alpha\beta}} = \sum_i A_{i\alpha} \frac{d\mathcal{L}}{dY_{i\beta}}$

E. None of the above

[7 marks]

Solution:

$$\frac{d\mathcal{L}}{dA_{\alpha\beta}} = \sum_{ik} \frac{\partial Y_{ik}}{\partial A_{\alpha\beta}} \frac{d\mathcal{L}}{dY_{ik}} \tag{15}$$

$$\frac{d\mathcal{L}}{dA_{\alpha\beta}} = \sum_{ik} \frac{\partial}{\partial A_{\alpha\beta}} [\sum_j A_{ij} B_{jk}] \frac{d\mathcal{L}}{dY_{ik}} \tag{16}$$

$$\frac{d\mathcal{L}}{dA_{\alpha\beta}} = \sum_{ijk} \frac{\partial}{\partial A_{\alpha\beta}} [A_{ij} B_{jk}] \frac{d\mathcal{L}}{dY_{ik}} \tag{17}$$

$$\frac{d\mathcal{L}}{dA_{\alpha\beta}} = \sum_{ijk} \delta_{i\alpha} \delta_{j\beta} B_{jk} \frac{d\mathcal{L}}{dY_{ik}} \tag{18}$$

$$\frac{d\mathcal{L}}{dA_{\alpha\beta}} = \sum_k \frac{d\mathcal{L}}{dY_{\alpha k}} B_{\beta k} \tag{19}$$

$$\frac{d\mathcal{L}}{dB_{\alpha\beta}} = \sum_{ik} \frac{\partial Y_{ik}}{\partial B_{\alpha\beta}} \frac{d\mathcal{L}}{dY_{ik}} \tag{20}$$

$$\frac{d\mathcal{L}}{dB_{\alpha\beta}} = \sum_{ik} \frac{\partial}{\partial B_{\alpha\beta}} [\sum_j A_{ij} B_{jk}] \frac{d\mathcal{L}}{dY_{ik}} \tag{21}$$

$$\frac{d\mathcal{L}}{dB_{\alpha\beta}} = \sum_{ijk} \frac{\partial}{\partial B_{\alpha\beta}} [A_{ij} B_{jk}] \frac{d\mathcal{L}}{dY_{ik}} \tag{22}$$

$$\frac{d\mathcal{L}}{dB_{\alpha\beta}} = \sum_{ijk} A_{ij} \delta_{j\alpha} \delta_{k\beta} \frac{d\mathcal{L}}{dY_{ik}} \tag{23}$$

$$\frac{d\mathcal{L}}{dB_{\alpha\beta}} = \sum_i A_{i\alpha} \frac{d\mathcal{L}}{dY_{i\beta}} \tag{24}$$

Q10. Which statement about backprop is **FALSE**:

A. Ultimately, the goal of backprop in NNs is to compute the gradients of the loss wrt the parameters.

B. You could compute these gradients by explicitly computing Jacobians for each operation in the compute graph.

C. But Jacobians can be very large (i.e. they can take lots of memory to store), which results in an inefficient algorithm.

D. Instead, in backprop, we directly compute $\frac{d\mathcal{L}}{d\text{inputs}}$ as a function of $\frac{d\mathcal{L}}{d\text{outputs}}$ for each operation.

E. None of the above

Q11. Which statement about momentum based optimization is **FALSE**:

A. You want a high learning rate in low-gradient directions, so you learn something in those directions.

B. You want a low learning rate in high-gradient directions, so to avoid instability.

C. Thus, a learning rate that is stable in the high-gradient directions is going to give slow learning in the low-gradient directions.

D. Momentum mitigates these issues by explicitly reducing the learning rate in high-gradient directions.

E. Adam combines explicit adaptive learning rates with momentum.

Solution: Momentum doesn't reduce the learning rate (that's Adam). Instead, momentum implicitly averages the gradient over multiple timesteps.

Q12. Consider the following loss function,

$$\mathcal{L} = \tfrac{1}{2}(y - wx)^2 \tag{25}$$

Do two steps of gradient descent for w with a learning rate of $\eta = 0.1$, with:

- The input fixed to $x = 2$.
- The output fixed to $y = 2$.
- The weight initialized to 0.

$$\Delta w = -\eta \frac{\partial \mathcal{L}}{\partial w} \tag{26}$$

A. 0.24

B. 0.4

(cont.)

C. 0.6

D. 0.64

E. 0.8

Solution: Apply the chain rule with, $u = y - wx$, so $\mathcal{L} = \frac{1}{2}u^2$,

$$\frac{\partial \mathcal{L}}{\partial w} = \frac{\partial \mathcal{L}}{\partial u}\frac{\partial u}{\partial w} \tag{27}$$

$$\frac{\partial \mathcal{L}}{\partial w} = -ux = -(y - wx)w \tag{28}$$

$$\Delta w = \eta x(y - wx) \tag{29}$$

At the first step, we have $x = 2, y = 2, w = 0, \eta = 0.1$, thus,

$$\Delta w = 0.1 \times 2 \times (2 - 2 \times 0) = 0.4 \tag{30}$$

Thus, after the first step, $w = 0.4$.

At the second step, we have $x = 2, y = 2, w = 0, \eta = 0.1$, thus,

$$\Delta w = 0.1 \times 2 \times (2 - 2 \times 0.4) = 0.2 \times 1.2 = 0.24 \tag{31}$$

Thus, after the second step, $w = 0.4 + 0.24 = 0.64$.

Q13. Consider doing gradient descent,

$$x_{\text{next}} = x - \eta \frac{\partial \mathcal{L}(x)}{\partial x},$$

on the the following objective,

$$\mathcal{L}(x) = x^2.$$

Select the fastest converging learning rate (you can get that by calculating the effect of a single gradient descent step with all the different learning rates, and selecting one that brings you closest to the optimum ($x = 0$),

A. $\eta = 0.25$

B. $\eta = 0.5$

C. $\eta = 1$

D. $\eta = 2$

E. None of the learning rates converge.

[7 marks]

Solution: Calculate the result of applying one gradient descent step, Alternatively, the threshold between stability and instability, is when $x \to -x$,

$$x_{\text{next}} = x_{\text{init}} - \eta \frac{\partial x^2}{\partial x} \tag{32}$$

$$x_{\text{next}} = x_{\text{init}} - 2\eta x_{\text{init}} \tag{33}$$

$$x_{\text{next}} = x_{\text{init}}(1 - 2\eta) \tag{34}$$

$\eta = 0.5$ converges in one step.

Q14. Which of these statements about cross-validation is FALSE:

A. Cross-validation can be used to assess overfitting.

B. Cross-validation reports performance on the training data used to fit the function.

C. Cross-validation can be used to choose the function class.

D. Cross-validation can be used to choose the amount of regularisation.

E. Cross-validation can be computationally expensive if we have more than one or two hyperparameters.

[7 marks]

Q15. Consider a 2D Convolution, with an input width=3, height=5. If we have kernel size=5, padding=1, stride=1, what is the output size?

(cont.)

A. output width=1 and output height=2

B. output width=1 and output height=3

C. output width=2 and output height=2

D. output width=2 and output height=3

E. None of the above.

[7 marks]

END OF PAPER

The following pages are left blank for your rough workings. They will not be collected or marked. You must enter your answers on the provided answer sheet only.