Durham University

EXAMINATION PAPER

Examination Session:	Year:	Exam Code:
May/June	2025	MATH4267-WE01

Title: Deep Learning and Artificial Intelligence

Time:	2 hours	
Additional Material provided:		
Materials Permitted:		
Calculators Permitted:	No	Models Permitted: Use of electronic calculators is forbidden.

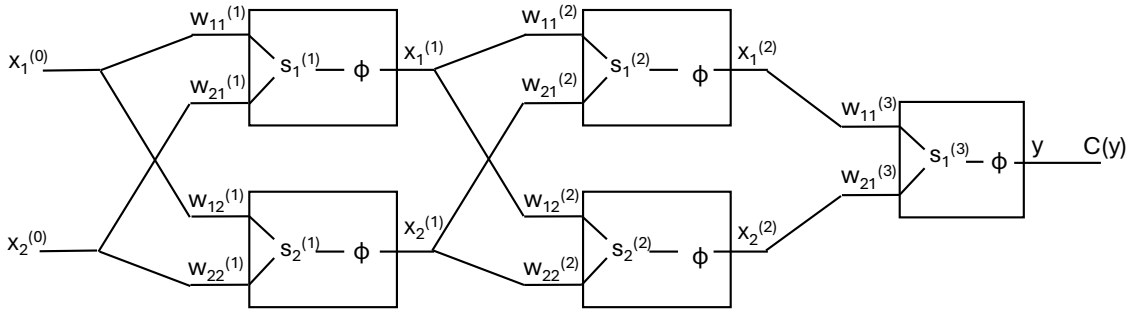
Instructions to Candidates:	Answer all questions. Section A is worth 40% and Section B is worth 60%. Within each section, all questions carry equal marks. Write your answer in the white-covered answer booklet with barcodes. Begin your answer to each question on a new page.

Revision:	

ED01/2025

SECTION A

Q1 Suppose we have the following neural network:



with the following specifications:

- All neurons have a common differentiable activation function ϕ
- The weight $w_{ij}^{(\ell)}$ indicates the weight for the ith input for the jth neuron in the ℓth layer. All neurons have zero bias
- The value $x_j^{(\ell)}$ is the output of the jth neuron in the ℓth layer, except for the final layer, for which the single output is y, and $x_1^{(0)}$, $x_2^{(0)}$, which are the inputs for the network.
- The value $s_j^{(\ell)}$ is the weighted sum of inputs for neuron j in layer ℓ, before the activation function is applied; that is $s_j^{\ell} = w_{1j}^{(\ell)} x_1^{(\ell-1)} + w_{2j}^{(\ell)} x_2^{(\ell-1)}$
- The loss function for the output y of the neural network is a differentiable function $C(y)$.

Answer the following questions:

(a) Let $\delta_i^{(\ell)} = \frac{\partial C}{\partial s_i^{\ell}}$, for $i \in \{1, 2\}$ and $\ell \in \{1, 2, 3\}$. Show that:

$$\delta_1^{(1)} = \phi'(s_i^{(1)}) \left(\delta_1^{(2)} w_{11}^{(2)} + \delta_2^{(2)} w_{12}^{(2)} \right).$$

(b) Show that:

$$\frac{\partial C}{\partial w_{12}^{(1)}} = \delta_2^{(1)} x_1^{(0)}.$$

(c) Suppose that we modify the overall cost to a new function L by adding a penalty term as follows (with $c > 0$ constant):

$$L = C(y) + c \sum_{\text{All weights } w} w^2.$$

Why might we want to do this? Name another method we could use for the same purpose.

Q2 Suppose we are given n discrete random variables X_1, X_2, ..., X_n, where $n \geq 3$. The mutual information I of all of these variables is defined as:

$$I(X_1, X_2, \ldots, X_n) = -H(X_1, X_2, \ldots, X_n) + \sum_{k=1}^{n} H(X_k),$$

where H is entropy.

(a) Show that for discrete random variables X, Y, Z we have:

$$I(X, Y, Z) = I(X, Y) + I((X, Y), Z).$$

(b) Let $S_1 = X_1$, $S_2 = (X_1, X_2)$, and $S_i = (X_1, X_2, \ldots, X_i)$. Show that:

$$I(X_1, X_2, \ldots, X_n) = \sum_{k=2}^{n} I(S_{i-1}, X_i).$$

Hint: you may wish to begin with an analogy of the identity in the previous part, and proceed by induction.

In this question, you may use any result from lectures, but must state when you do so.

SECTION B

Q3 We wish to design a variational autoencoder (VAE). We have some data, which we presume is generated by the following latent process:

- Sample a value $z \sim MVN(0, I_m)$ for some dimension m.
- Take a data sample X with distribution given by

$$(X|Z = z) \sim MVN\left(f(z, \theta^{(d)}), cI_n\right)$$

where $f(\cdot, \theta^{(d)}) : \mathbb{R}^m \to \mathbb{R}^n$ is some smooth function, $c > 0$ is a constant, $n > m$, and $\theta^{(d)}$ is the set of parameters of our decoder.

The encoder of our VAE has parameters $\theta^{(e)}$, and we will aim to approximate the PDF $f_{Z|X=x}(z)$ by a distribution $q_x(z, \theta^{(e)})$. In this question:

- The expression $MVN(\mu, \Sigma)$ is the multivariate normal distribution with mean μ and covariance matrix Σ,
- If $x = (x_1, x_2, \ldots, x_k)$ then $||x||_2^2 = x_1^2 + x_2^2 + \cdots + x_k^2$,
- The expression I_k denotes the identity matrix of order k,
- The expression $f_Z(z)$ denotes the probability density function of Z at z, and $f_{X|Z=z}(x)$ denotes the probability density function of $(X|Z = z)$ at x.

(a) Sketch the architecture of a VAE, showing the encoder, the decoder, and any random sampling. Indicate the data space and latent space and their relative dimensions.

(b) Show that:

$$\arg\min_{\theta^{(e)}} D_{KL}\left(q_x(z, \theta^{(e)})||f_{Z|X=x}(z)\right) =$$
$$= \arg\min_{\theta^{(e)}} \left(\mathbb{E}_{q_x(z,\theta^{(e)})} \left\{ \frac{||x - f(z, \theta^{(d)})||_2^2}{2c} \right\} + D_{KL}\left(q_x(z, \theta^{(e)})||f_Z(z)\right) \right).$$

where $D_{KL}\left(f_1(z)||f_2(z)\right)$ denotes the Kullback-Leibler divergence between distributions with densities $f_1(z)$ and $f_2(z)$.

(c) Explain why we wish to resample the latent space with added noise, rather than encode our data to a latent representation and decode it directly.

(d) Briefly describe the operation of a generative adversarial network, and the problem of mode collapse. Why is a VAE less susceptible to mode collapse?

Q4 Suppose we wish to build a neural network which can approximate the function:

$$f(x) = \cos(2\pi x),$$

on the interval $x \in [0, p]$, where p is a positive integer. We want our approximation $g(x)$ to satisfy

$$|f(x) - g(x)| < \epsilon,$$

for all $x \in [0, p]$, for some $\epsilon > 0$.

Our network will have one input, some number of hidden layers, and an output layer consisting of a single neuron. Neurons in the hidden layers have ReLU activation functions, and the neuron in the final layer has an identity activation function (that is, $\phi(x) = x$).

(a) Show that there exists a function $k(\epsilon)$ such that if we only use one hidden layer, we need at least $k(\epsilon)p$ neurons in total to produce such an approximation.

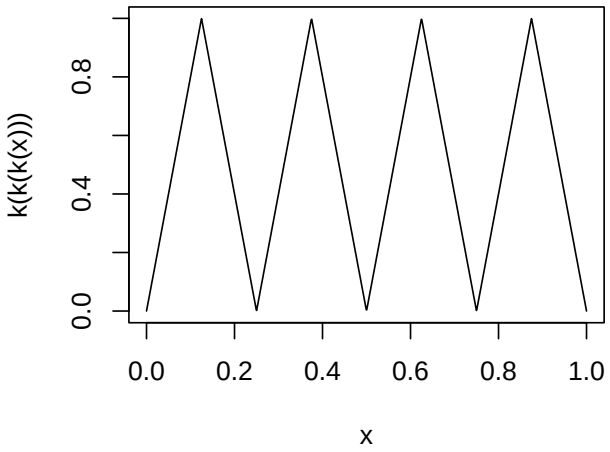
Hint: you may wish to consider the number of line segments needed to approximate $\cos(2\pi x)$ on $x \in [0, 1]$ to a maximum error of ϵ using a continuous piecewise linear function taking the value 1 at both endpoints. You do not need to find the function $k(\cdot)$, only show that it exists and is finite.

(b) Specify a network with more than one hidden layer which implements this function using at most $c\ln(p) + k(\epsilon)$ neurons, where c is a constant not depending on p or ϵ, and k is a function of ϵ only.

Hint: if

$$k(x) = ReLU\left(2ReLU(x) - 4ReLu\left(x - \frac{1}{2}\right)\right),$$

then $\overbrace{k(k(\ldots k(}^{r \text{ times}} x)))$ describes a 'sawtooth' function with 2^{r-1} 'teeth'; e.g.:



(c) Consider the set $A(p)$ of functions from $\{1, 2, \ldots, p\}$ to $\{0, 1\}$, and the set $A_m(p)$ of such functions which are periodic with period m. Describe the behaviour of the Kolmogorov complexity of typical members of $A(p)$ and $A_m(p)$ for large p and fixed m.