NANYANG TECHNOLOGICAL UNIVERSITY

SEMESTER 1 EXAMINATION 2022-2023

CE4042/CZ4042 – NEURAL NETWORKS AND DEEP LEARNING

Nov/Dec 2022 Time Allowed: 2 hours

INSTRUCTIONS

1. This paper contains 4 questions and comprises 7 pages.

2. Answer **ALL** questions.

3. This is an open-book examination.

4. All questions carry equal marks.

---

1. (a) Briefly state the following. Each part carries 2 marks.

   (i) The difference between gradient descent (GD) and stochastic gradient descent (SGD) learning algorithms.

   (ii) How the time to weight update varies with batch size in semi-batch SGD learning.

   (iii) How one could use a linear neuron to learn a given nonlinear equation.

   (iv) How a discrete perceptron is able to perform linear classification.

   (v) Two limitations when a logistic regression neuron is trained with gradient descent learning to perform classification.

   (vi) Two ways to initialize weights of a network to improve convergence.

   (12 marks)

Note: Question No. 1 continues on Page 2

(b) A two-layer discrete perceptron network receives 2-dimensional inputs $(x_1, x_2)^T \in \boldsymbol{R^2}$ and has one output neuron. The shaded region of Figure Q1 shows the input space for which the output of the network, $y = 1$. Draw the network clearly indicating the values of the weights and biases of the neurons.

(13 marks)

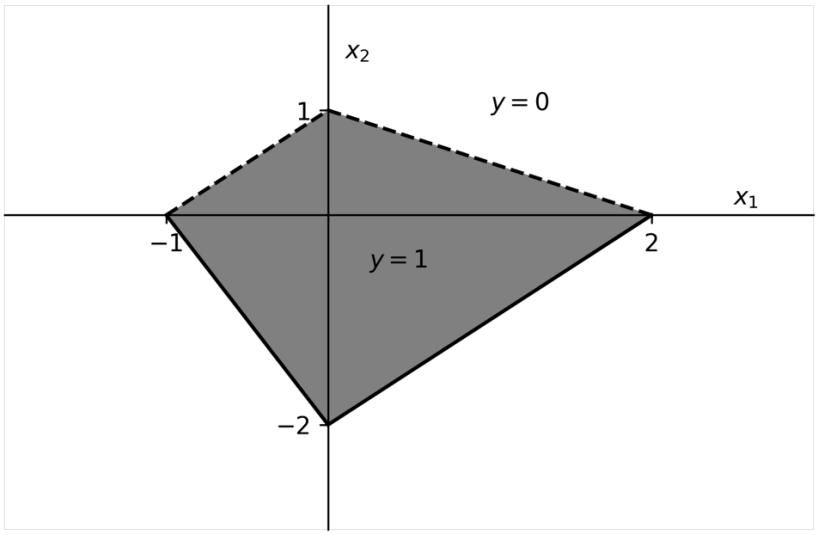

Figure Q1

2. A two-layer feedforward neural network shown in Figure Q2 receives two-dimensional inputs $(x_1, x_2) \in \boldsymbol{R^2}$ and produces two-dimensional outputs (y_1, y_2). The hidden layer consists of three perceptrons with activation functions $f(u) = \frac{2}{1+e^{-u}}$ and the output layer is a linear layer with two neurons. The weights of the network are initialized as indicated in the figure and all the biases are initialized to 0.1 (not shown).

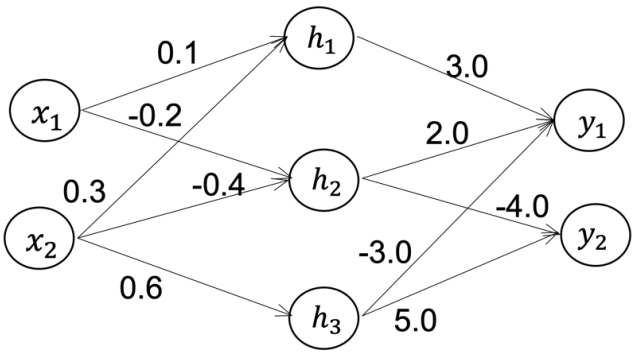

Figure Q2

Note: Question No. 2 continues on Page 3

The network is trained to produce a desired output $\boldsymbol{d} = \begin{pmatrix} -1.0 \\ 1.0 \end{pmatrix}$ for an input $\boldsymbol{x} = \begin{pmatrix} 1.5 \\ 2.0 \end{pmatrix}$. You are to perform one iteration of gradient descent learning with the example $(\boldsymbol{x}, \boldsymbol{d})$. The learning factor $\alpha = 0.1$. Give your answers up to two decimal places.

(a) Write initial weight matrix $\boldsymbol{W}$ and bias vector $\boldsymbol{b}$ of the hidden layer, and initial weight matrix $\boldsymbol{V}$ and bias vector $\boldsymbol{c}$ of the output layer.

(2 marks)

(b) Find synaptic input $\boldsymbol{z}$ and output $\boldsymbol{h}$ of the hidden layer and the output $\boldsymbol{y}$ of the output layer.

(6 marks)

(c) Find the square error at the output.

(2 marks)

(d) Find the derivative $f'(\boldsymbol{z})$ at the hidden layer.

(3 marks)

(e) Find gradients $\nabla_{\boldsymbol{u}} J$ and $\nabla_{\boldsymbol{z}} J$ of the cost J with respect to $\boldsymbol{u}$ and $\boldsymbol{z}$, respectively.

(6 marks)

(f) Find gradients $\nabla_{\boldsymbol{V}} J$, $\nabla_{\boldsymbol{c}} J$, $\nabla_{\boldsymbol{W}} J$, and $\nabla_{\boldsymbol{b}} J$ of the cost J with respect to $\boldsymbol{V}$, $\boldsymbol{c}$, $\boldsymbol{W}$, and $\boldsymbol{b}$, respectively.

(4 marks)

(g) Find the updated values of $\boldsymbol{V}$, $\boldsymbol{c}$, $\boldsymbol{W}$, and $\boldsymbol{b}$.

(2 marks)

3. In the following questions, the size of input or output volume is represented as $D \times H \times W$, where D is the number of channels, and $H \times W$ is the spatial size.

(a) Given an input volume of size $3 \times 512 \times 512$, we have 64 convolution filters each with a size of $3 \times 5 \times 5$, stride $= 1$.

(i) What is the output volume size if we use a padding size of 2?

(3 marks)

(ii) Give a reason why one would use padding in a convolution layer.

(2 mark)

(iii) What is the total number of parameters in this layer? Be reminded to account for the bias terms.

(3 marks)

(b) (i) Calculate the FLOPs of a standard convolution layer. Assume the filter size is 3×3, the input volume is $3 \times 45 \times 45$, and the output volume is $128 \times 43 \times 43$. Be reminded to account for the bias terms.

(3 marks)

(ii) Compute the reduction rate of FLOPs if we replace the standard convolution with a depthwise separable convolution (i.e., "depthwise + pointwise" convolutions). Assume the filter size of the depthwise convolution is 3×3, the input volume is still $3 \times 45 \times 45$, and the output volume is $128 \times 43 \times 43$. Be reminded to account for the bias terms.

(3 marks)

(iii) Figure Q3 on page 5 depicts a block that consists of three convolutional layers. The input volume has a size of $256 \times 32 \times 32$ and the second layer has 32 convolution filters each with a size of $64 \times 3 \times 3$, stride = 1 and padding = 1.

Provide the values of n_1, d_1, F_1, n_2, d_2, and F_2 to form a valid block. Explain your design.

(8 marks)

Note: Question No. 3 continues on Page 5

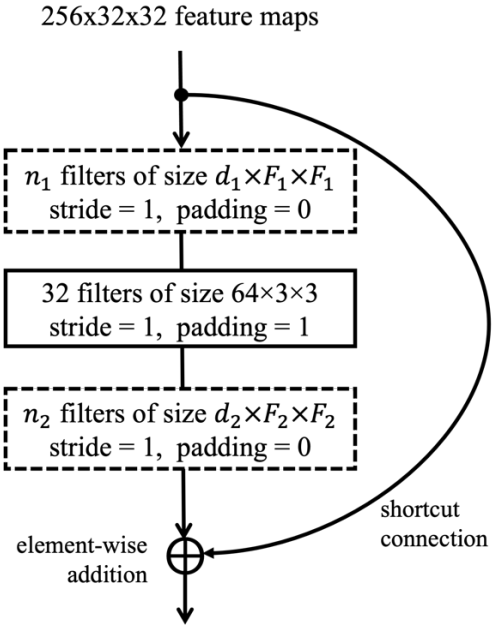

Figure Q3

(c) Select the correct option (A, B, C or D) for each question.

A. Both statements are TRUE.
B. Statement I is TRUE, but statement II is FALSE.
C. Statement I is FALSE, but statement II is TRUE.
D. Both statements are FALSE.

(i) Statement I: Autoencoders are a supervised learning technique.
Statement II. Autoencoder's output is exactly the same as the input.

(1 mark)

(ii) Statement I: One way to implement undercomplete autoencoder is to constrain the number of nodes present in hidden layer(s) of the neural network
Statement II. To train a denoising encoder, we use a loss between the original input and the reconstruction from a noisy version of the input.

(1 mark)

(iii) Statement I: Sparse autoencoders introduce information bottleneck by reducing the number of nodes at hidden layers.
Statement II. With the sparsity constraint, we will observe more neuron outputs that are close to zero.

(1 mark)

4. (a) Consider an Elman-type recurrent neural network (RNN) that receives 2-dimensional input patterns $\boldsymbol{x} \in \boldsymbol{R}^2$ and has one hidden layer. The RNN has five neurons in the hidden layer and two neurons in the output layer.

We denote the weight matrices connecting the input to the hidden layer as $\boldsymbol{U}$, the weight matrices connecting the previous hidden state to the next hidden state as $\boldsymbol{W}$, and the weight matrices connecting the hidden output to the output layer as $\boldsymbol{V}$.

(i) What is the dimension of $\boldsymbol{U}$, $\boldsymbol{W}$ and $\boldsymbol{V}$, respectively?

(3 marks)

(ii) If we change this RNN to a Jordan-type RNN, and keep the same number of input dimensions, hidden and output neurons, what is the dimension of the top-down recurrence weight matrix $\boldsymbol{W}$?

(1 mark)

(iii) Explain the reason of observing exploding gradients in RNN training. Describe a way to address this problem.

(5 marks)

(b) The statements below are all related to Transformers. Answer "TRUE" or "FALSE" to the following statements. Each part carries 1 mark.

(i) Divide the dot product of the query vector with the key vector by the square root of the dimension of the key vectors leads to vanishing gradients.

(ii) Multi-head attention alleviates the need for positional encoding.

(iii) The encoder-decoder attention layer of each decoder layer accepts the query and value matrices obtained from the output of the encoder stack as input.

(iv) In the same encoder layer, we apply the same feedforward network independently to each position.

(v) Transformer uses positional encoding to help learn better attention.

(vi) Positional encoding can either be pre-defined or made learnable.

(6 marks)

Note: Question No. 4 continues on Page 7

(c) Consider a self-attention layer in a Transformer, which receives the following key (**K**), query (**Q**), and value (**V**) matrices:

$$\mathbf{K} = \begin{pmatrix} \sqrt{2} & \sqrt{2} \\ \sqrt{2} & \sqrt{2} \end{pmatrix}, \ \mathbf{Q} = \begin{pmatrix} \ln 2 & \ln 3 \\ \ln 1 & \ln 4 \end{pmatrix} \text{ and } \mathbf{V} = \begin{pmatrix} 3 & 4 \\ 5 & 6 \end{pmatrix}$$

Compute the output of the scaled-dot product attention, Attention(**Q**, **K**, **V**).

(7 marks)

(d) Describe the changes that you need to make to turn an unconditional Generative Adversarial Network (GAN) to a conditional one that takes class labels. Explain one advantage of the conditional GAN in comparison to the unconditional GAN.

(3 marks)

END OF PAPER