INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR
Mid-Spring Semester Examination 2022-2023

Date of Examination: 23/2/23 Session:AN Duration: 2 hours Full Marks: 60
Subject No.: CS60010 Deep Learning
Department: Computer Science & Engineering

Special Instructions: Answer the questions in the boxes provided in the question cum answer booklet only. use the last pages for Rough work. Supplementary sheets used if any must be tied to the booklet.

---

1. Write the update rules for a parameter w_i in a neural network with loss L when the following algorithms are used: (8)

   (a) Stochastic gradient descent using one sample (x_1, y_1)

   Solution:

   $$w_i \longleftarrow w_i - \eta \frac{\partial L(f(x;\theta), y)}{\partial w_i}$$

   (b) Batch gradient descent using m samples $(x_1, y_1), \cdots, (x_m, y_m)$

   Solution:

   $$w_i \longleftarrow w_i - \eta \frac{\partial}{\partial w_i} \sum_i L(f(x_i;\theta), y_i)$$

   (c) Stochastic gradient descent with momentum using 1 sample.

   Solution:

   $$\begin{array}{l|l} v_{t+1} \longleftarrow \rho v_t + \nabla f(w_t) & v_{t+1} \longleftarrow \rho v_t - \eta g \\ w_{t+1} \longleftarrow w_t - \eta v_{t+1} & w_{t+1} \longleftarrow w_t + v_{t+1} \end{array}$$

   $$g = \nabla_{w_i} L(f(x;\theta), y)$$

   (d) Stochastic gradient descent with Nesterov accelerated momentum using 1 sample

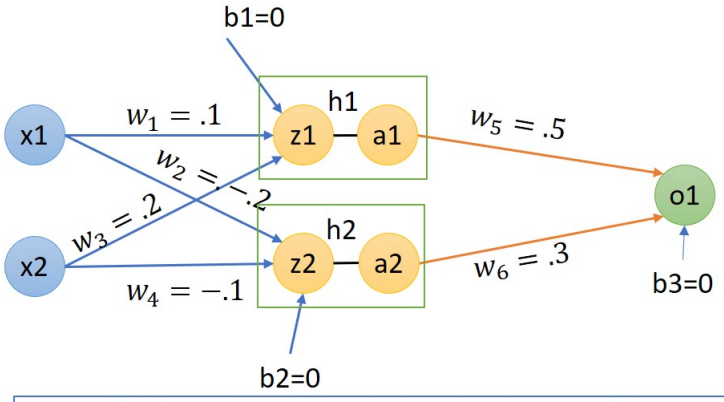
   Solution:

   $$\begin{array}{c|c} \tilde{\theta} = \theta_t + \eta v_t & v_{t+1} = \rho v_t - \alpha \nabla f(x_t + \rho v_t) \\ g_{NAG} = \nabla_{w_i} L(f(x;\tilde{\theta}, y) & x_{t+1} = x_t + v_{t+1} \\ v_{t+1} \longleftarrow \rho v_t - \eta g_{NAG} & \\ w_{t+1} \longleftarrow w_t + v_{t+1} & \end{array}$$

2. Consider the simple neural network drawn below for a regression task with two input units, one hidden layer with two neurons and RELU as the activation function, and a single output which is a linear unit. Use the MSE loss function defined as $Loss(y, \hat{y}) = \frac{1}{2} * (y - \hat{y})^2$. (15)

   The initial weight values associated with the network are $w_1 = .1, w_2 = .2, w_3 = -.2, w_4 = -.1, w_5 = .5, w_6 = .3, b1 = b2 = b3 = 0$. Recall, RELU$(z)$= max$(0, z)$

   Show how stochastic gradient descent using `Instance1`= $\{x_1 = 1, x2 = 1, y = .5\}$ will update the network weights when the learning rate $\eta = 0.5$ by filling up the steps in the answer box.



Instance1: $x1 = 1, x2 = 1, y = .5$

h1 and h2 have RELU as activation function. o1 is linear.
$a1 = \text{relu}(z1), a2 = \text{relu}(z2)$

Solution: Instance1 : Input $x1 = 1, x2 = 1$; Output $y = 0.5$

$z1 = .3$ $a1 = .3$
$z2 = -.3$ $a2 = 0$
$o1 = .15$

$$L = \frac{1}{2} * (y - o1)^2$$

Loss, $L = .06125$

$$\frac{\partial L}{\partial o1} = -(y - o1) = -.35$$

$$\frac{\partial L}{\partial w_5} = \frac{\partial L}{\partial o1}\frac{\partial o1}{\partial w_5} = -.35 * a1 = -.35 * .3 = -.105$$

$$\frac{\partial L}{\partial w_6} = \frac{\partial L}{\partial o1}\frac{\partial o1}{\partial w_6} = -.35 * a2 = -.35 * 0 = 0$$

$$\frac{\partial L}{\partial b3} = \frac{\partial L}{\partial o1}\frac{\partial o1}{\partial b3} = -.35 * 1 = -.35$$

Solution:

$$\frac{\partial L}{\partial a1} = \frac{\partial L}{\partial o1}\frac{\partial o1}{\partial a1} = -.35 * w5 = -.175$$
$$\frac{\partial L}{\partial a2} = \frac{\partial L}{\partial o1}\frac{\partial o1}{\partial a2} = -.35 * w6 = -.105$$

$$\frac{\partial L}{\partial z1} = \frac{\partial L}{\partial a1}\frac{\partial a1}{\partial z1} = -.175 * x1 = -.175$$
$$\frac{\partial L}{\partial z2} = \frac{\partial L}{\partial a2}\frac{\partial a2}{\partial z2} = 0$$

$$\frac{\partial L}{\partial w_1} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial w_1} = -.175 * 1 = -.175$$
$$\frac{\partial L}{\partial w_3} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial w_3} = -.175 * x2 = -.175$$

$$\frac{\partial L}{\partial w_2} = \frac{\partial L}{\partial z_2}\frac{\partial z_2}{\partial w_2} = 0$$
$$\frac{\partial L}{\partial w_4} = \frac{\partial L}{\partial z_2}\frac{\partial z_2}{\partial w_4} = 0$$

$$\frac{\partial L}{\partial b1} = \frac{\partial L}{\partial z_1}\frac{\partial z_1}{\partial b_1} = -.175 * 1 = -.175$$
$$\frac{\partial L}{\partial b2} = \frac{\partial L}{\partial z_2}\frac{\partial z_2}{\partial b_2} = 0$$

$$w_1 = w_1 - \eta\frac{\partial L}{\partial w_1} = .1 + .5 * .175 = .1875$$
$$w_2 = w_2 - \eta\frac{\partial L}{\partial w_2} = -.2 + 0 = -.2$$
$$w_3 = w_3 - \eta\frac{\partial L}{\partial w_3} = .2 + 0.5 * .175 = .2875$$
$$w_4 = w_4 - \eta\frac{\partial L}{\partial w_4} = -.1 + 0 = -.1$$
$$w_5 = w_5 - \eta\frac{\partial L}{\partial w_5} = .5 + .5 * .105 = .5525$$
$$w_6 = w_6 - \eta\frac{\partial L}{\partial w_6} = .3 + 0 = .3$$
$$b1 = b1 - \eta\frac{\partial L}{\partial b1} = 0 + .5 * .175 = .0875$$
$$b2 = b2 - \eta\frac{\partial L}{\partial b2} = 0 + 0 = 0$$
$$b3 = b3 - \eta\frac{\partial L}{\partial b3} = 0 + .5 * .35 - .175$$

3. Which of the following techniques can be used to reduce model overfitting? Justify. (4)

√ Dropout

◯ Adam for gradient descent

Solution: Dropout Justification:

4. (a) Suppose that you are training a neural network for classification, but you notice that the training loss is much lower than the validation loss. Which of the following can be used to address the issue (select all that apply) (2)

- √ Use a network with fewer layers
- ◯ Decrease dropout probability
- √ Increase L2 regularization weight
- ◯ Increase the size of each hidden layer

(b) Which of the following best describes the purpose of batch normalization in a neural network? (2)

- √ Regularizing the model to prevent overfitting
- √ Speeding up the training process by reducing internal covariate shift
- ◯ Introducing noise to the input data to increase robustness
- ◯ Ensuring the model outputs are in the same range as the labels

(c) Which of the following best describes how the Adagrad optimizer adjusts the learning rate during training? (2)

- √ Maintains a running average of the gradients and scales the learning rate by the inverse of the average
- √ Adapts the learning rate for each weight based on the history of the weight gradients
- ◯ Uses the momentum of the weight updates to modify the learning rate for each weight
- ◯ Modifies the learning rate based on the ratio of the current gradient to the previous gradient.

(d) Which of the following optimization methods modifies the momentum term to take into account the future gradient? (2)

- ◯ Momentum
- √ Nesterov accelerated momentum
- ◯ RMSprop
- ◯ Adam

(e) Which of the following optimization methods combines the benefits of both momentum and adaptive learning rates? (2)

- ◯ Momentum
- ◯ Nesterov accelerated momentum
- ◯ RMSprop
- √ Adam

(f) What is the role of Activation Functions in a deep neural network? (1)

- ◯ To convert the output of a Convolutional layer to a probability distribution.
- ◯ To add noise to the output of a Convolutional layer.
- √ To add non-linearity to the output of a Convolutional layer.
- ◯ To reduce the number of parameters in a Convolutional layer.

5. (a) Suppose you have a convolutional network with the following architecture: (3)

- The input is an RGB image of size 256×256.
- There are 2 consecutive convolution layers each with 32 feature maps and filters of size 3×3 with stride 1 and no pooling.
- They are followed by a pooling layer with a stride of 2 (so it reduces the size of each dimension by a factor of 2) and pooling groups of size 3×3.

What is the size of the receptive field for a single unit in the pooling layer. (i.e., the size of the region of the input image which influences the activation of that unit.)

Solution: 7×7

6. Which of the following statements do you agree with? (5)

◯ Visualization of the weights of a convolutional filter of a CNN provides useful insight for a filter at any layer of a CNN.

√ Visualization of the activation map of a convolutional filter of a CNN provides useful insight for a filter at any layer of a CNN.

◯ The gradient of the class score of an object detection CNN with respect to image pixels provides

√ The deeper layers of a neural network are typically computing more complex features of the input than the earlier layers.

◯ The earlier layers of a neural network are typically computing more complex features of the input than the deeper layers.

7. Consider a CNN based classifier for a 4-class classification problem to classify 4 different kinds of birds found in West Bengal ponds. The layers of the network are given in Column 1 of the table below. (14)

The notation follows the convention:

- CONV-K-N denotes a convolutional layer with N filters, each them of size K×K, Padding and stride parameters are always 0 and 1 respectively.
- POOL-K indicates a K × K pooling layer with stride K and padding 0.
- FC-N stands for a fully-connected layer with N neurons

Layer	Activation map dimensions	Number of weights	Number of biases
INPUT	$128 \times 128 \times 3$	0	0
CONV-7-32	$122 \times 122 \times 32$	$7 \times 7 \times 3 \times 32 = 4704$	32
POOL-2	$61 \times 61 \times 32$	0	0
CONV-5-64	$57 \times 57 \times 64$	$5 \times 5 \times 32 \times 64$	64
CONV-3-128	$55 \times 55 \times 128$	$3 \times 3 \times 64 \times 128$	128
POOL-2	$28 \times 28 \times 128$	0	0
FC-10	10	$28 \times 28 \times 128 \times 10$	10
OUTPUT	4	10×4	4

(a) For each layer, calculate the number of weights, number of biases and the size of the associated feature maps and fill the values in the table above.

(b) Following the last FC-10 layer of the network, what activation must be applied? Given a vector $a = [0.3, 0.3, 0.3, 0.3]$, what is the result of using your activation on this vector?

Solution: Softmax

$$[0.25, 0.25, 0.25, 0.25]$$

(c) Suppose this exact same network has already been trained by your friend on 1 million different images of birds of the Himalayas. Suggest a method to use this for your classification task. Use 1-2 sentences for your answer.

Solution: Transfer Learning

ROUGH WORK