Graded Assignment 1

The due date for submitting this assignment has passed.
Due on 2023-06-18, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

You have last submitted on: 2023-06-18, 23:29 IST

1) Which of the following threshold values of MP neuron implements AND Boolean function? Assume that the number of inputs to the neuron is 10 and the neuron does not have any inhibitory inputs. *1 point*

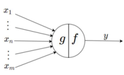
- [ ] 1
- [x] 10
- [ ] 11
- [ ] 12

Yes, the answer is correct.
Score: 1

Accepted Answers:

10

2) The diagram below shows an MP neuron. *1 point*



Suppose that we define the function g as follows,

$$g(x_1, x_2, \cdots, x_m) = \left(\sum_{i=1}^{n} x_i\right) \cdot \prod_{j=n+1}^{m} (1 - x_j).$$

The output from the neuron is given by,

$$f(g(\mathbf{x})) \begin{cases} 1, & \text{if } g(\mathbf{x}) \geq \theta \\ 0, & g(\mathbf{x}) < \theta \end{cases}$$

How many inputs are inhibitory in the neuron? Assume θ is positive valued.

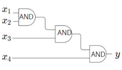
- [ ] 0
- [ ] m
- [ ] n
- [x] m-n
- [ ] m-n+1

Yes, the answer is correct.
Score: 1

Accepted Answers:

m-n

3) The diagram below shows an implementation of a Boolean function using three logical AND gates. What is the minimum number of MP neurons required to implement the same Boolean function? Assume that all inputs are excitatory. *1 point*



- [ ] 3
- [ ] 2
- [x] 1
- [ ] 4

Yes, the answer is correct.
Score: 1

Accepted Answers:

1

4) We know that a Boolean function maps $\{0,1\}^n \rightarrow \{0,1\}$. How many Boolean functions are there when n = 0? Assume that the function allows an empty element as an input. *1 point*

- [ ] 0
- [ ] 1
- [x] 2
- [ ] 4

Yes, the answer is correct.
Score: 1

Accepted Answers:

2

5) An input to the perceptron model is, $\mathbf{x} = [x_0, x_1, \cdots, x_n]^T \mathbb{R}^{n+1}$ and the weight vector is $\mathbf{w} = [w_0, w_1, \cdots, w_n]^T$. The output from the perceptron is given by *1 point*

$$f(g(\mathbf{x})) \begin{cases} 1, & \text{if } g(\mathbf{x}) \geq 0 \\ 0, & g(\mathbf{x}) < 0 \end{cases}$$

where $g(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, then the decision boundary (i.e, the hyper-plane that separates the data points) passes through the origin in?

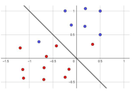
- [ ] n^{th} dimension
- [x] $n + 1^{th}$ dimension
- [ ] $n - 1^{th}$ dimension
- [ ] It is not possible to decide whether it passes through the origin in any of the dimension

Yes, the answer is correct.
Score: 1

Accepted Answers:

$n + 1^{th}$ dimension

6) In the diagram shown below, the blue data points belong to positive class ($\mathbf{w}^T\mathbf{x} \geq 0$) and red data points belong to negative class ($\mathbf{w}^T\mathbf{x} < 0$). The number of misclassified data points according to the decision line as shown in the figure is? *1 point*



- [ ] 0
- [ ] 1
- [ ] 2
- [x] Insufficient information

Yes, the answer is correct.
Score: 1

Accepted Answers:

Insufficient information

7) The table below shows the truth table for NAND gate (a Boolean function). Suppose that the perceptron model is to be used for learning the decision line such that it separates all the data points with zero-classification error *1 point*

\begin{array} \hline x_1 & x_2 & y \\ \hline 0 & 0 & 1 \\ \hline 0 & 1 & 1 \\ \hline 1 & 0 & 1 \\ \hline 1 & 1 & 0 \\ \hline \end{array}

The value of the bias term (w_0) should strictly be

- [x] Positive
- [ ] Non-negative
- [ ] negative
- [ ] Depends on the other weight(w_1, w_2) values

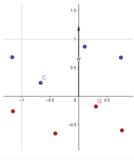
Yes, the answer is correct.
Score: 1

Accepted Answers:

Positive

Common data for question 8,9 and 10

In the figure shown below, the blue points belong to class 1 (positive class) and the red points belong to class 0 (negative class). Suppose that we use a perceptron model, with the weight vector $\mathbf{w}$ as shown in the figure, to separate these data points. We define the point belongs to class 1 if $\mathbf{w}^T\mathbf{x} \geq 0$ else it belongs to class 0.



8) The point G and C will be classified as? *1 point*

Note: the notation $(G, 0)$ denotes the point G will be classified as class-0 and $(C, 1)$ denotes the point C will be classified as class-1

- [ ] (G, 0), (C, 0)
- [ ] (G, 1), (C, 0)
- [x] (G, 0), (C, 1)
- [ ] (G, 1), (C, 1)

Yes, the answer is correct.
Score: 1

Accepted Answers:

(G, 0), (C, 1)

9) The statement that "there exists more than one decision lines that could separate these data points with zero error " is, *1 point*

- [x] True
- [ ] False

Yes, the answer is correct.
Score: 1

Accepted Answers:

True

10) Suppose that we multiply the weight vector w by -1. Then the same points G and C will be classified as? *1 point*

- [ ] (G, 0), (C, 0)
- [x] (G, 1), (C, 0)
- [ ] (G, 0), (C, 1)
- [ ] (G, 1), (C, 1)

Yes, the answer is correct.
Score: 1

Accepted Answers:

(G, 1), (C, 0)

Graded Assignment 2

The due date for submitting this assignment has passed.
Due on 2023-06-18, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

You have last submitted on: 2023-06-18, 23:30 IST

1) Assume that we have an absolutely linearly separable data points $\mathbf{x}_i \in \mathbb{R}^d, i = 1, 2, \cdots N$. Each data point belongs to either class +1 or class -1 (i.e., $y \in \{+1, -1\}$). The perceptron outputs 1 if $(\mathbf{w}^T\mathbf{x}) \geq 0$ else it outputs -1. Suppose we use the perceptron learning algorithm to separate the data points. Suppose further that the perceptron found $\mathbf{w}^*$ that separates the data points. Choose which of the following equations is (are) true. *1 point*

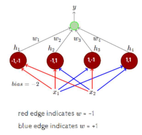
- (●) $y \cdot (\mathbf{w}^{*T}\mathbf{x}) \geq 0$ for all $\mathbf{x}$
- ( ) $y \cdot (\mathbf{w}^{*T}\mathbf{x}) \leq 0$ for all $\mathbf{x}$
- ( ) $\mathbf{w}^{*T}\mathbf{x} \geq 0$
- ( ) $\mathbf{w}^{*T}\mathbf{x} \leq 0$ for all $\mathbf{x}$

Yes, the answer is correct.
Score: 1

Accepted Answers:

$y \cdot (\mathbf{w}^{*T}\mathbf{x}) \geq 0$ for all $\mathbf{x}$

2) Suppose that we implement the XOR Boolean function using the network shown below. Consider the statement that "A hidden layer with two neurons is suffice to implement XOR". The statement is *2 points*



. The statement is

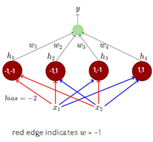
- (●) True
- ( ) False

Yes, the answer is correct.
Score: 2

Accepted Answers:

True

3) Consider a Multilayer Perceptron network shown below. *1 point*



Suppose we wish the network to implement a Boolean function which has the following truth table,

x_1	x_2	y
−1	−1	1
−1	1	−1
1	−1	−1
1	1	−1

Which of the following w_0 (bias of the perceptron in the output layer) is (are) valid?
Note: $h_i \in \{0, 1\}$ and $y \in \{-1, 1\}$. Assume a suitable threshold function.

- [x] 0
- [x] 1
- [x] -1
- [x] -2

Yes, the answer is correct.
Score: 1

Accepted Answers:

0
1
-1
-2

4) Consider a sigmoid function $\frac{1}{1+e^{-(wx+b)}}$. Assume that w is a positive quantity. Select all the correct statements about the function *1 point*

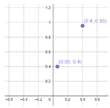
- [x] Increasing the value of b shifts the sigmoid function to the left (i.e., towards negative infinity)
- [ ] Increasing the value of b shifts the sigmoid function to the right (i.e., towards positive infinity)
- [x] Increasing the value of w increases the slope of the sigmoid function
- [ ] Increasing the value of w decreases the slope of the sigmoid function

Yes, the answer is correct.
Score: 1

Accepted Answers:

Increasing the value of b shifts the sigmoid function to the left (i.e., towards negative infinity)
Increasing the value of w increases the slope of the sigmoid function

5) Consider the data points as shown in the figure below, *1 point*



Suppose that the sigmoid function given below is used to fit these data points.

$$\frac{1}{1+e^{-(20x+1)}}$$

Compute the Mean Square Error (MSE) loss $\mathcal{L}(w, b)$

- (●) 0.126
- ( ) 0
- ( ) 0.216
- ( ) 1

Yes, the answer is correct.
Score: 1

Accepted Answers:

0.126

6) Consider the dataset shown below. The points x are sampled from a sigmoid function of the form *1 point*

$$\sigma(x) = \frac{1}{1+e^{-(wx+b)}}$$

x	y
−1	0.5
0.2	0.97

Suppose that we run the gradient descent algorithm with the parameters w and b initialized to $w_0 = 2$ and $b_0 = 2$. Select the right sequence that represents the sequence of updates correspond to w and b

Note: You are welcome to write a small piece of code to find the answer.

- ( ) w_1=[1.9,1.19,0.39], b_1 =[2.04,2.20,2.31]
- (●) w_1=[2.008,2.19,2.39], b_1 =[2.04,2.20,2.31]
- ( ) w_1=[2.008,2.19,2.39], b_1 =[1.9,1.19,0.39]
- ( ) w_1=[-2.008,-2.19,-2.39], b_1 =[1.9,1.19,0.39]

Yes, the answer is correct.
Score: 1

Accepted Answers:

w_1=[2.008,2.19,2.39], b_1 =[2.04,2.20,2.31]

7) *1 point*

Suppose that the input to the logistic function is a $n-$dimensional vector $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, then the gradient $\nabla \mathbf{w}$ is given as $\nabla \mathbf{w} = (f(\mathbf{x}) - y) * f(\mathbf{x})(1 - f(\mathbf{x}) * \mathbf{x}$. The claim is

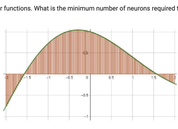
- (●) True if y is a scalar
- ( ) True if y is a vector
- ( ) False

Yes, the answer is correct.
Score: 1

Accepted Answers:

True if y is a scalar

8) The function $f(x)$ show below is approximated using 150 tower functions. What is the minimum number of neurons required to construct the network that approximates the function? *2 points*



- ( ) 450
- ( ) 151
- (●) 301
- ( ) 300
- ( ) 451
- ( ) 150

Yes, the answer is correct.
Score: 2

Accepted Answers:

301

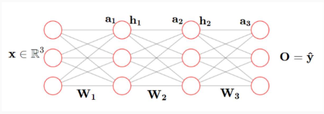
Graded Assignment 3

The due date for submitting this assignment has passed.
Due on 2023-07-02, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

You have last submitted on: 2023-07-02, 23:29 IST

The diagram below shows a neural network used for a classification problem. The network contains two hidden layers and one output layer. The input to the network is a column vector $\mathbf{x} \in \mathbb{R}^3$. The first hidden layer contains 3 neurons, the second hidden layer contains 3 neurons and the output layer contains 3 neurons. Each neuron in the l^{th} layer is connected to all the neurons in the $(l+1)^{th}$ layer. Each neuron has a bias connected to it (not explicitly shown in the figure).



The weights and biases are initialized as given below,

$$\mathbf{W_1} = \begin{bmatrix} 0.5488135 & 0.71518937 & 0.60276338 \\ 0.54488318 & 0.4236548 & 0.64589411 \\ 0.43758721 & 0.891773 & 0.96366276 \end{bmatrix}$$

$$\mathbf{W_2} = \begin{bmatrix} 0.56804456 & 0.92559664 & 0.07103606 \\ 0.0871293 & 0.0202184 & 0.83261985 \\ 0.77815675 & 0.87001215 & 0.97861834 \end{bmatrix}$$

$$\mathbf{W_3} = \begin{bmatrix} 0.11827443 & 0.63992102 & 0.14335329 \\ 0.94466892 & 0.52184832 & 0.41466194 \\ 0.26455561 & 0.77423369 & 0.45615033 \end{bmatrix}$$

$$\mathbf{b_1} = \begin{bmatrix} 0.38344152 \\ 0.79172504 \\ 0.52889492 \end{bmatrix}$$

$$\mathbf{b_2} = \begin{bmatrix} 0.79915856 \\ 0.46147936 \\ 0.78052918 \end{bmatrix}$$

$$\mathbf{b_3} = \begin{bmatrix} 0.56843395 \\ 0.0187898 \\ 0.6176355 \end{bmatrix}$$

The weights that connects outputs from neurons in the previous $(i-1)$ layer to a neuron in the present i^{th} layer correspond to a row in the weight matrix.The input to the network

$$\mathbf{x} = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

and the corresponding label $y = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$.

All the neurons in the hidden layers use Sigmoid activation function and the neurons in the output layer uses Softmax function. Assume that the network uses the cross entropy loss (use natural log).

You are advised to use the **Numpy** package to compute matrix vector multiplications. You can download the initial weights HERE.You can load the weights using

files = np.load('parameters.npz')

and use, say, *file.get('W1')* to get $\mathbf{W_1}$ weight matrix.
Important: Do Not truncate or round off any elements of the parameters. Use them as is

1) How many (learnable) parameters are there in the network?

36

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 35.9,36.1

1 point

2) What is the sum of the elments of output $\mathbf{a_1}$? (Choose the nearest option to your answer) *1 point*

- 3.7
- 5.44
- 4.16
- 17.59

Yes, the answer is correct.
Score: 1

Accepted Answers:

5.44

3) What is the sum of the elements of output $\mathbf{h_1}$? (Choose the nearest option to your answer) *1 point*

- 2.57
- 3.46
- 0.67
- 0.42

Yes, the answer is correct.
Score: 1

Accepted Answers:

2.57

4) The sum of the elements of $[\mathbf{a_2}, \mathbf{h_2}, \mathbf{a_3}]$, respectively, are $[6.4, 2.63, 4.87]$. What is the loss value?

1.23

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 0.83,0.87
(Type: Range) 1.1,1.3

1 point

5) Choose the vector that corresponds to $\nabla_{a_3}\mathscr{L}(\theta)$. *1 point*

- $\begin{bmatrix} 0 \\ 0 \\ -2.35 \end{bmatrix}$
- $\begin{bmatrix} 0 \\ 0 \\ 2.35 \end{bmatrix}$
- $\begin{bmatrix} 0.23 \\ 0.33 \\ -2.35 \end{bmatrix}$
- $\begin{bmatrix} 0.23 \\ 0.33 \\ 0.42 \end{bmatrix}$
- $\begin{bmatrix} 0.23 \\ 0.33 \\ -0.57 \end{bmatrix}$
- $\begin{bmatrix} -0.23 \\ -0.33 \\ 0.57 \end{bmatrix}$

Yes, the answer is correct.
Score: 1

Accepted Answers:

$\begin{bmatrix} 0.23 \\ 0.33 \\ -0.57 \end{bmatrix}$

6) We know that after computing gradients, we update the values of $\mathbf{b_2}$ by subtracting its gradient, as shown below, *2 points*

$$\mathbf{b_2} - \eta\nabla_{b_2}\mathscr{L}(\theta).$$

Which of the following is the gradient vector of $\mathbf{b_2}$?

- $\begin{bmatrix} 0.018 \\ -0.019 \\ -0.003 \end{bmatrix}$
- $\begin{bmatrix} 0.0 \\ 0.0 \\ 0.0 \end{bmatrix}$
- $\begin{bmatrix} 0.003 \\ 0.01 \\ -0.18 \end{bmatrix}$
- $\begin{bmatrix} 0.23 \\ 0.33 \\ 0.42 \end{bmatrix}$

Yes, the answer is correct.
Score: 2

Accepted Answers:

$\begin{bmatrix} 0.018 \\ -0.019 \\ -0.003 \end{bmatrix}$

7) Update all the parameters with the calculated gradients. Forward propagate the input through the network. What is the new loss value ? (Take $\eta = 1$)

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) 0.07,0.08
(Type: Range) 0.1,0.2

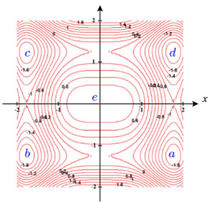
3 points

Graded Assignment 4

The due date for submitting this assignment has passed.
Due on 2023-07-09, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

1) The figure below shows the contours of a 3D surface. Some points in the figure are marked with the labels (a, b, c, d, e) that correspond to either maxima or minima. ***2 points***



Select all true statements about the map

- [ ] The function has four minima
- [ ] The function has 4 maxima
- [ ] The function has only one maximum
- [ ] The function has only one minimum
- [ ] It is not possible to decide whether the labels correspond to a minima or maxima without color coding the map

No, the answer is incorrect.
Score: 0

Accepted Answers:

The function has four minima
The function has only one maximum

2) Consider a loss function $L(w) = 0.1w^2$. The gradient descent rule is used to update the weight value as given below $w_{t+1} = w_t - \eta \nabla w_t$ where, ∇w_t is the gradient of the loss function with respect to w_t. Suppose that you need to choose between two learning rates $\eta \in \{1, 10\}$ so that the algorithm converges to zero loss quickly. Select the correct statements ***1 point***

- [ ] $\eta = 1$ is a good choice if w is initialized to 1 (that is, $w_0 = 1$)
- [ ] $\eta = 10$ is a good choice if w is initialized to 1 (that is, $w_0 = 1$)
- [ ] $\eta = 1$ is a good choice if w is initialized to 10 (that is, $w_0 = 10$
- [ ] $\eta = 10$ is a good choice if w is initialized to 10 (that is, $w_0 = 10$)

No, the answer is incorrect.
Score: 0

Accepted Answers:

$\eta = 1$ is a good choice if w is initialized to 1 (that is, $w_0 = 1$)
$\eta = 1$ is a good choice if w is initialized to 10 (that is, $w_0 = 10$

3) Consider the same loss function used in the question above, $L(w) = 0.1w^2$. Suppose that we prefer to use momentum based gradient descent rule to update the weight value as given below

$u_t = \beta u_{t-1} + \nabla w_t$

$w_{t+1} = w_t - \eta u_t$

where, ∇w_t is the gradient of the loss function with respect to w_t. What is the loss value after 100 iterations? Take,$w_0 = 1, \beta = 0.9$ and $\eta = 10$.

(Writing a small piece of code helps :-)).

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) 0.000004,0.000006

1 point

4) Select the true statements about the factor β used in the momentum based gradient descent algorithm ***1 point***

- [ ] Setting $\beta = 0.1$ allows the algorithm to move faster than vanilla (plain) gradient descent algorithm
- [ ] Setting $\beta = 1$ makes it equivalent to vanilla gradient descent algorithm
- [ ] Setting $\beta = 0$ makes it equivalent to vanilla gradient descent algorithm
- [ ] Oscillation around the minimum will be less if we set $\beta = 0.1$ than setting $\beta = 0.99$

No, the answer is incorrect.
Score: 0

Accepted Answers:

Setting $\beta = 0.1$ allows the algorithm to move faster than vanilla (plain) gradient descent algorithm
Setting $\beta = 0$ makes it equivalent to vanilla gradient descent algorithm
Oscillation around the minimum will be less if we set $\beta = 0.1$ than setting $\beta = 0.99$

5) Consider a Sigmoid neuron with a single input. The value of input $x = 0.5$ and the true output value $y = 1$. The weight and bias of the neuron are initialized to $w_0 = 5$ and $b_0 = -5$. The model uses Nesterov Accelerated gradient descent algorithm with $\beta = 0.9$ and $\eta = 0.1$. Also,it uses Mean Square Error (MSE) loss of the form $L(w) = \frac{1}{2}(\hat{y} - y)^2$.

Suppose that the model has been trained for 10 iterations (iteration starts from zero). The state of the parameters at $10^{th} (t = 9)$ iteration is as follows, $u_8 = 0.8, w_9 = 2.5, b_9 = 0$

Calculate the gradient of look-ahead value for the next weight update.

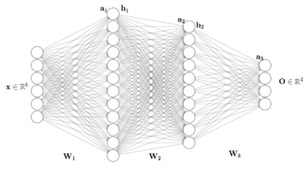
No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) -0.059,-0.052

2 points

6) Suppose that we implement the momentum based gradient descent to a feed forward neural network shown below. ***2 points***



The parameter update rule for the weight matrix $\mathbf{W_3}$ is written as follows,

$$u_3^t = \beta u_3^{t-1} + \nabla a_3^t (h_2^t)^T$$
$$W_3^{t+1} = W_3^t - \eta u_3^t$$

Choose the correct statements about the update rule

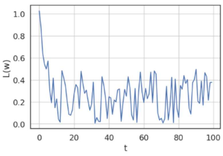
- [ ] At any given time step u_3^t is a vector
- [ ] At any given time step u_3^t is a matrix
- [ ] The update rule for u_3^t is wrong, the second term should have been written as ∇W_3^t
- [ ] β is a scalar

No, the answer is incorrect.
Score: 0

Accepted Answers:

At any given time step u_3^t is a matrix
β is a scalar

7) The figure below shows the change in loss value over iterations ***1 point***



The oscillation in the loss value might be due to

- [ ] Stochastic gradient descent algorithm used for parameter updates
- [ ] Mini-batch gradient descent algorithm used for parameter updates
- [ ] Batch gradient descent with constant learning rate algorithm used for parameter updates
- [ ] Batch gradient descent with line search algorithm used for parameter updates

No, the answer is incorrect.
Score: 0

Accepted Answers:

Stochastic gradient descent algorithm used for parameter updates
Mini-batch gradient descent algorithm used for parameter updates

Graded Assignment 5

The due date for submitting this assignment has passed.
Due on 2023-07-24, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

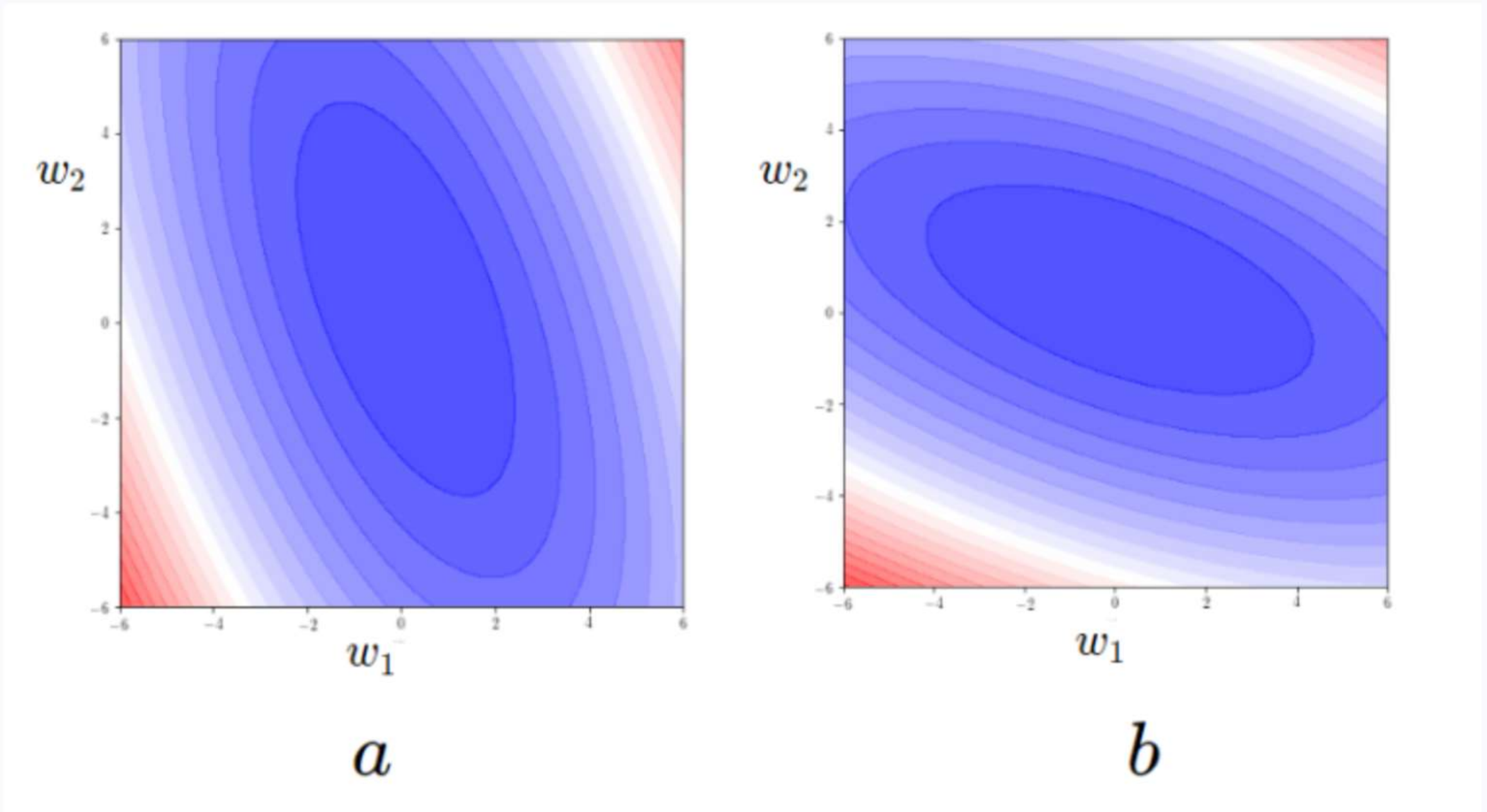
You have last submitted on: 2023-07-24, 23:45 IST

Common data for Q1 and Q2

The figures below show contours of a loss surface generated by the following model under two different settings,

$\hat{y} = w_1x_1 + w_2x_2$
$\mathscr{L}(w_1, w_2) = \frac{1}{2m}\sum_{i=1}^{m}(\hat{y} - y)^2$
where, m is a number of training samples, x_1 and x_2 are sparse feature and y is the corresponding ground truth.

Assume that both features x_1 and x_2 are normalized to be in the same range.



1) Which of the following statement(s) is (are) true? *2 points*

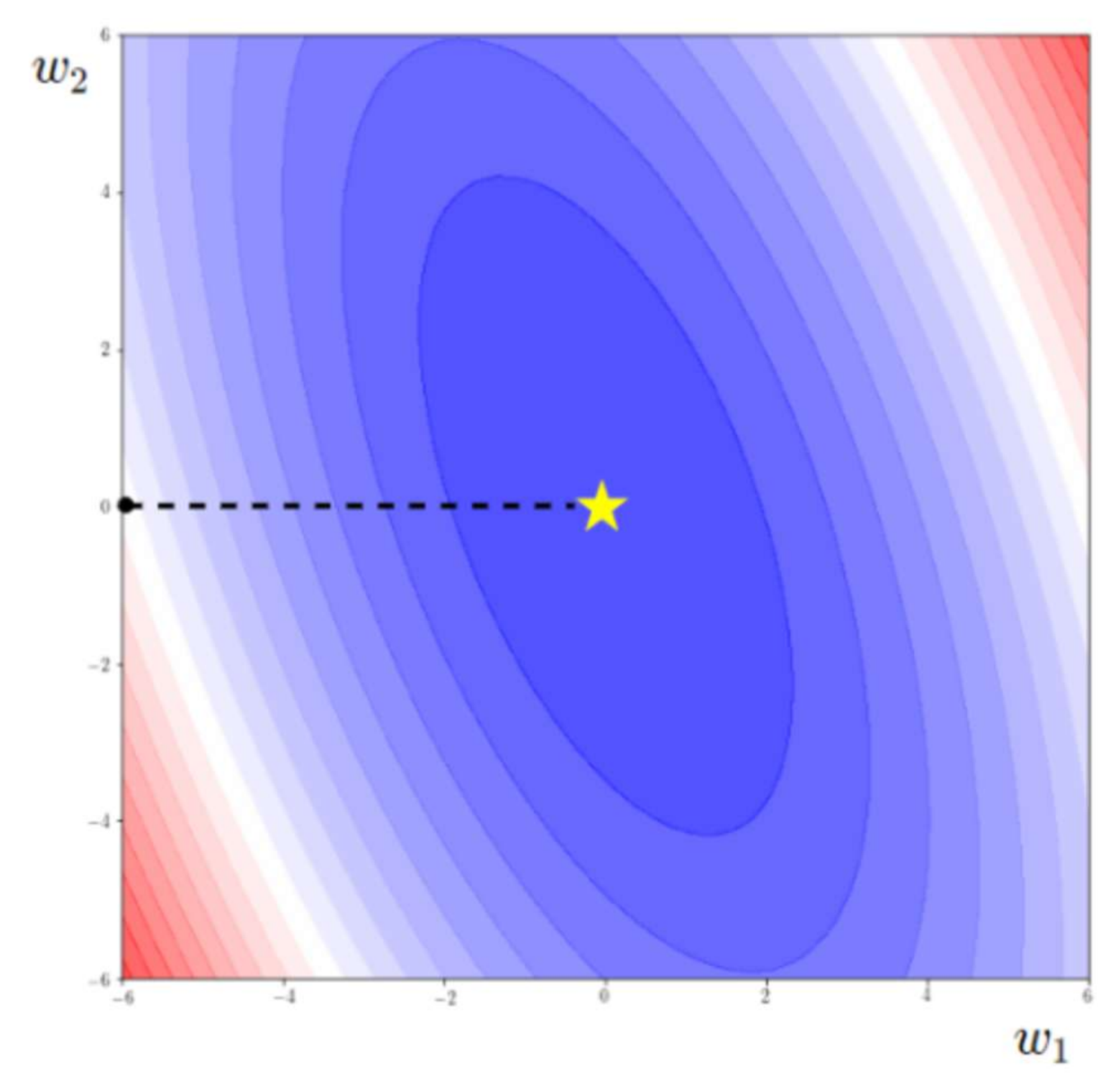
- [ ] In figure **a**, the feature x_1 is sparse and the feature x_2 is dense
- [x] In figure **a**, the feature x_1 is dense and the feature x_2 is sparse
- [ ] In figure **b**, the feature x_1 is dense and the feature x_2 is sparse
- [x] In figure **b**, the feature x_1 is sparse and the feature x_2 is dense

Yes, the answer is correct.
Score: 2

Accepted Answers:

In figure **a**, the feature x_1 is dense and the feature x_2 is sparse
In figure **b**, the feature x_1 is sparse and the feature x_2 is dense

2) Suppose that we use the Gradient Descent algorithm with suitable learning rate (so that it won't move past the minimum). If the weights are initialized to $w_1 = -6, w_2 = 0$, then the algorithm follows the trajectory as shown in the figure below. *2 points*



The claim is

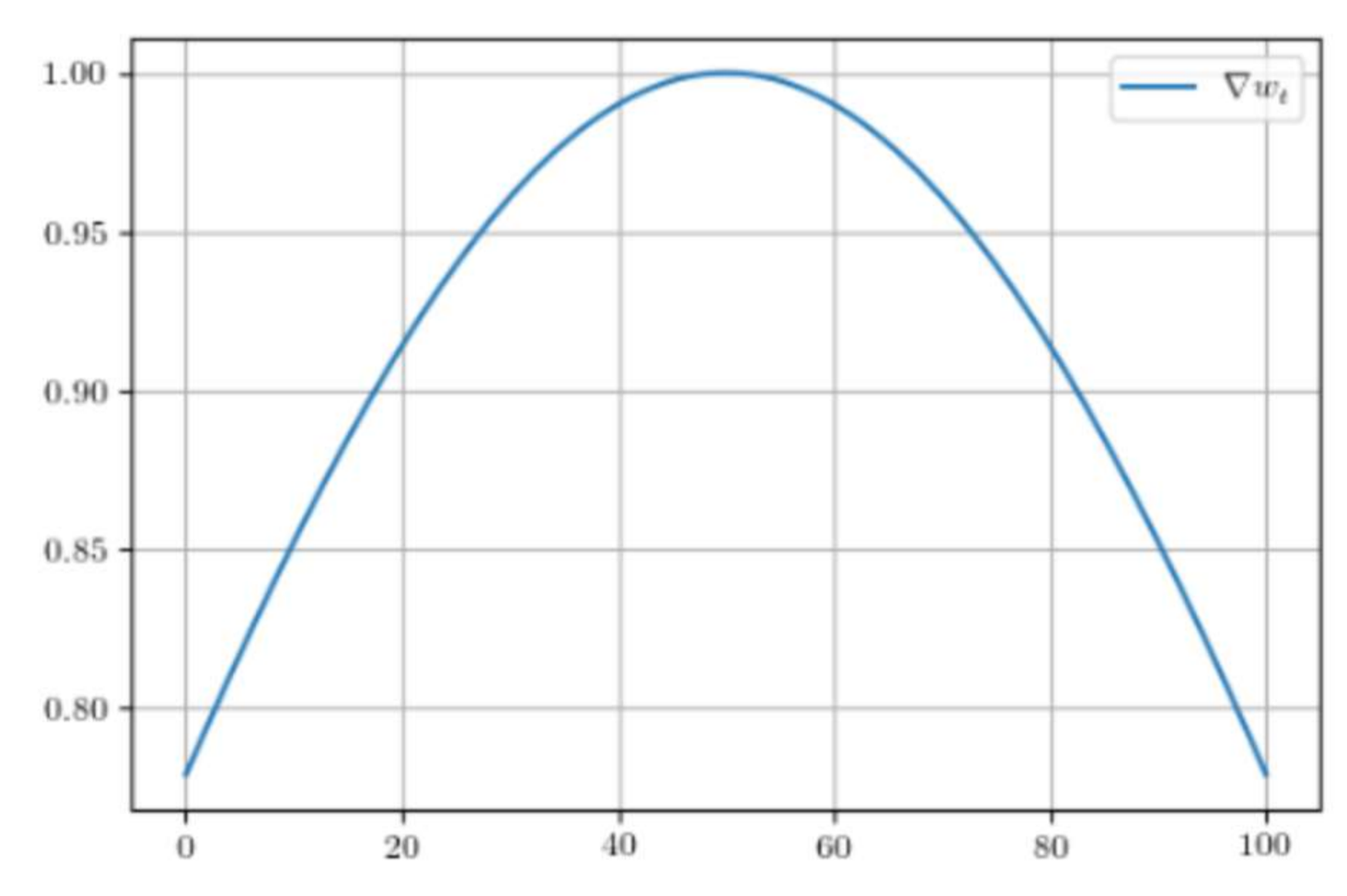
- ( ) True
- (•) False
- ( ) It is not possible to decide whether the claim is true or false without running the algorithm

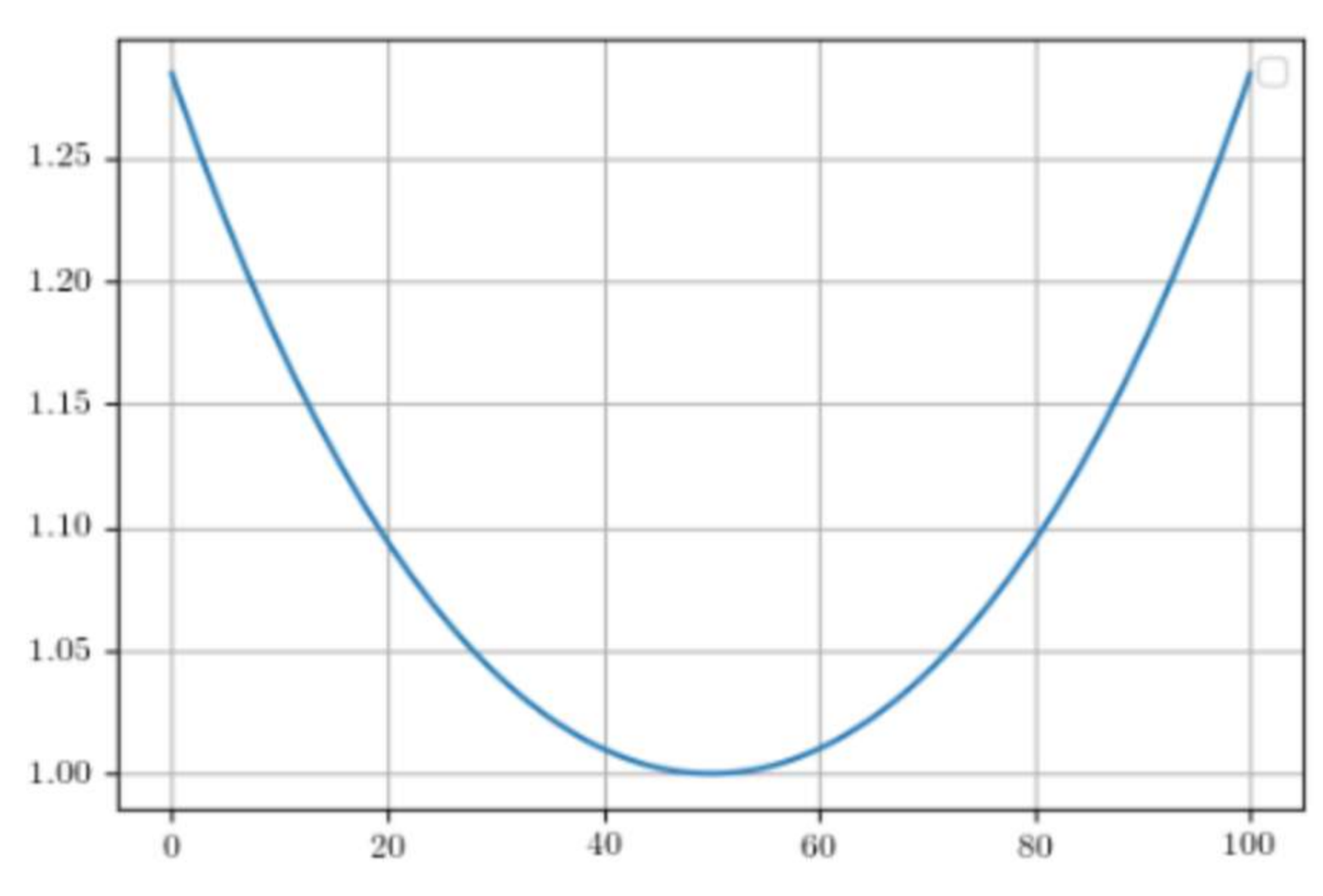
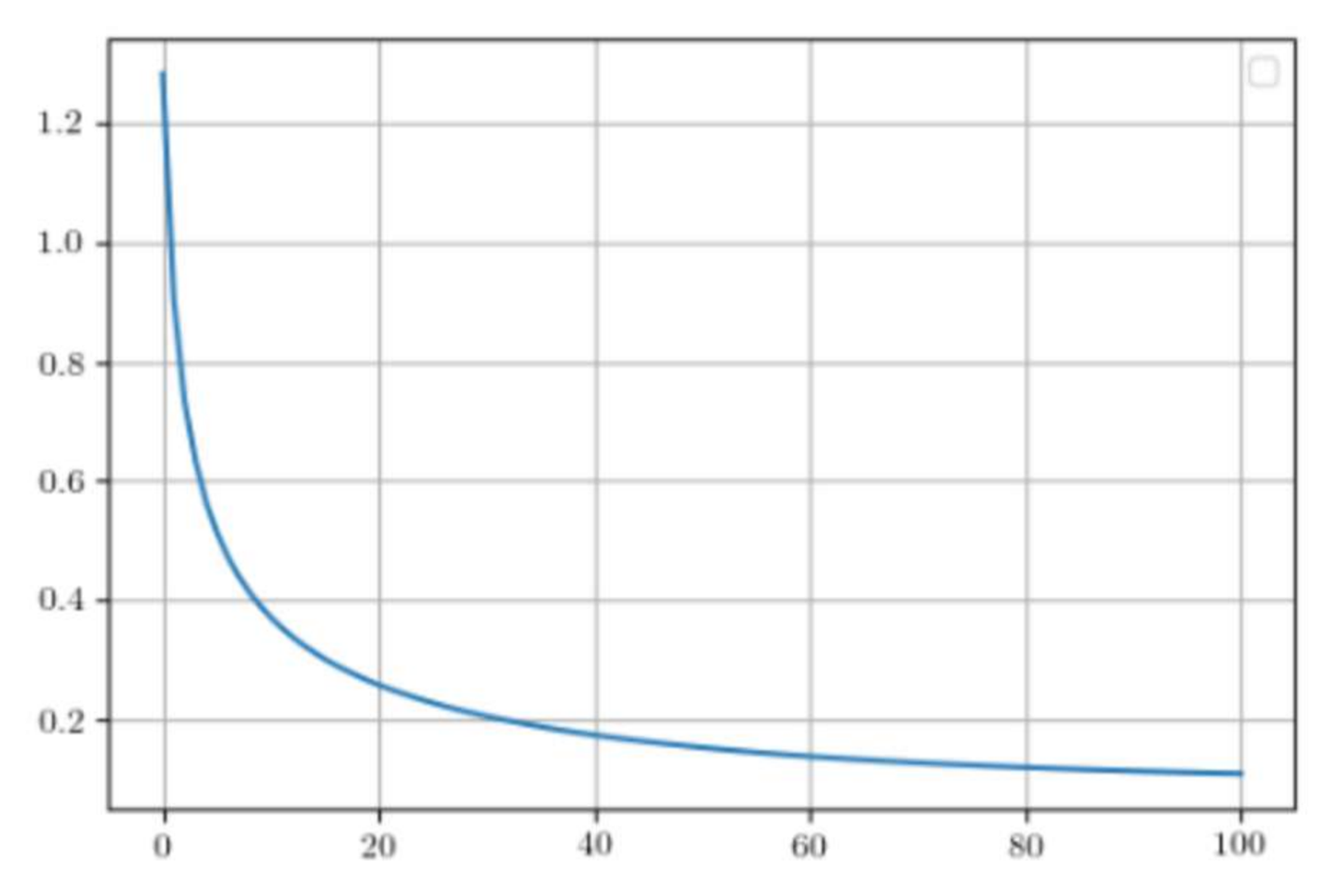
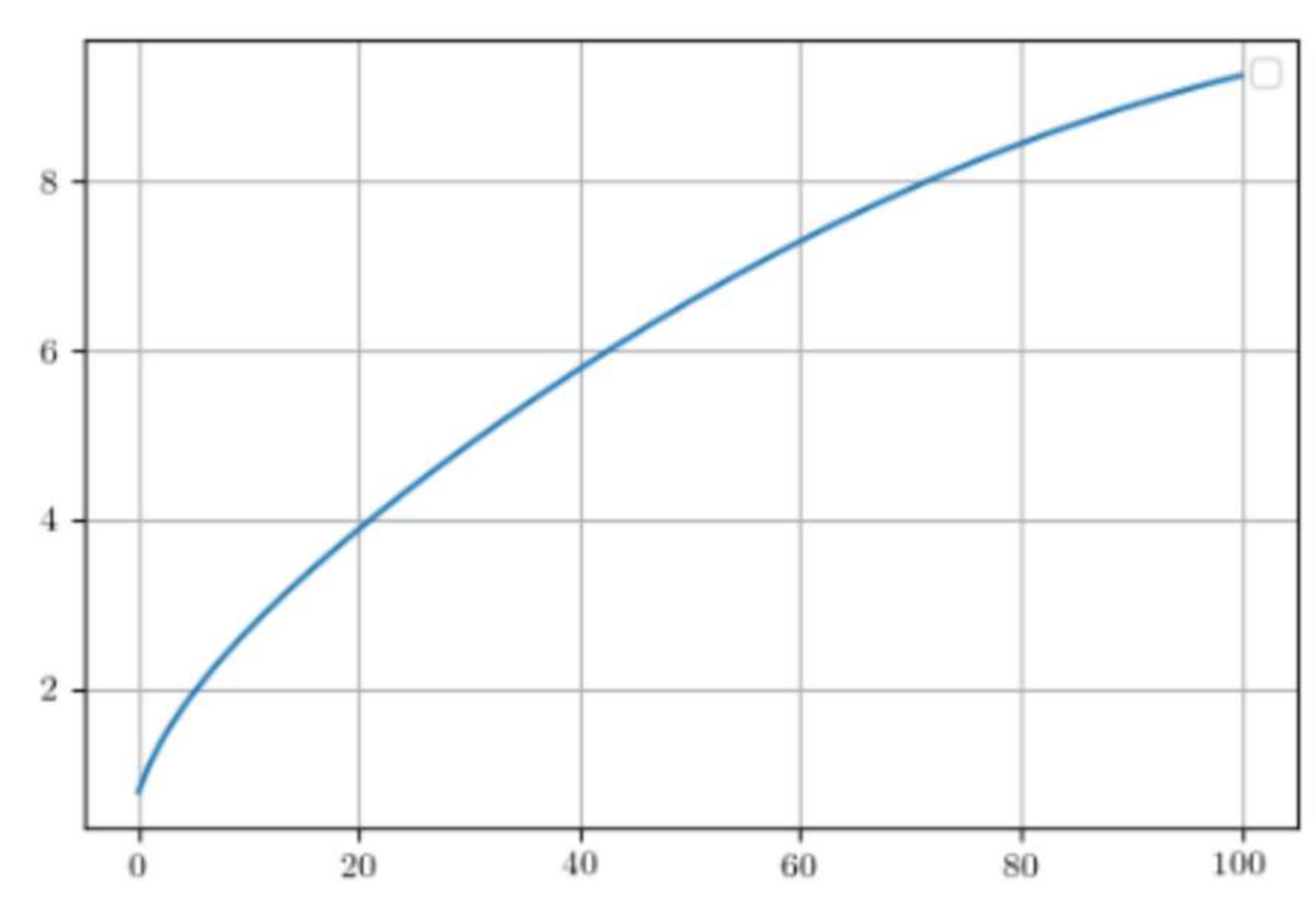
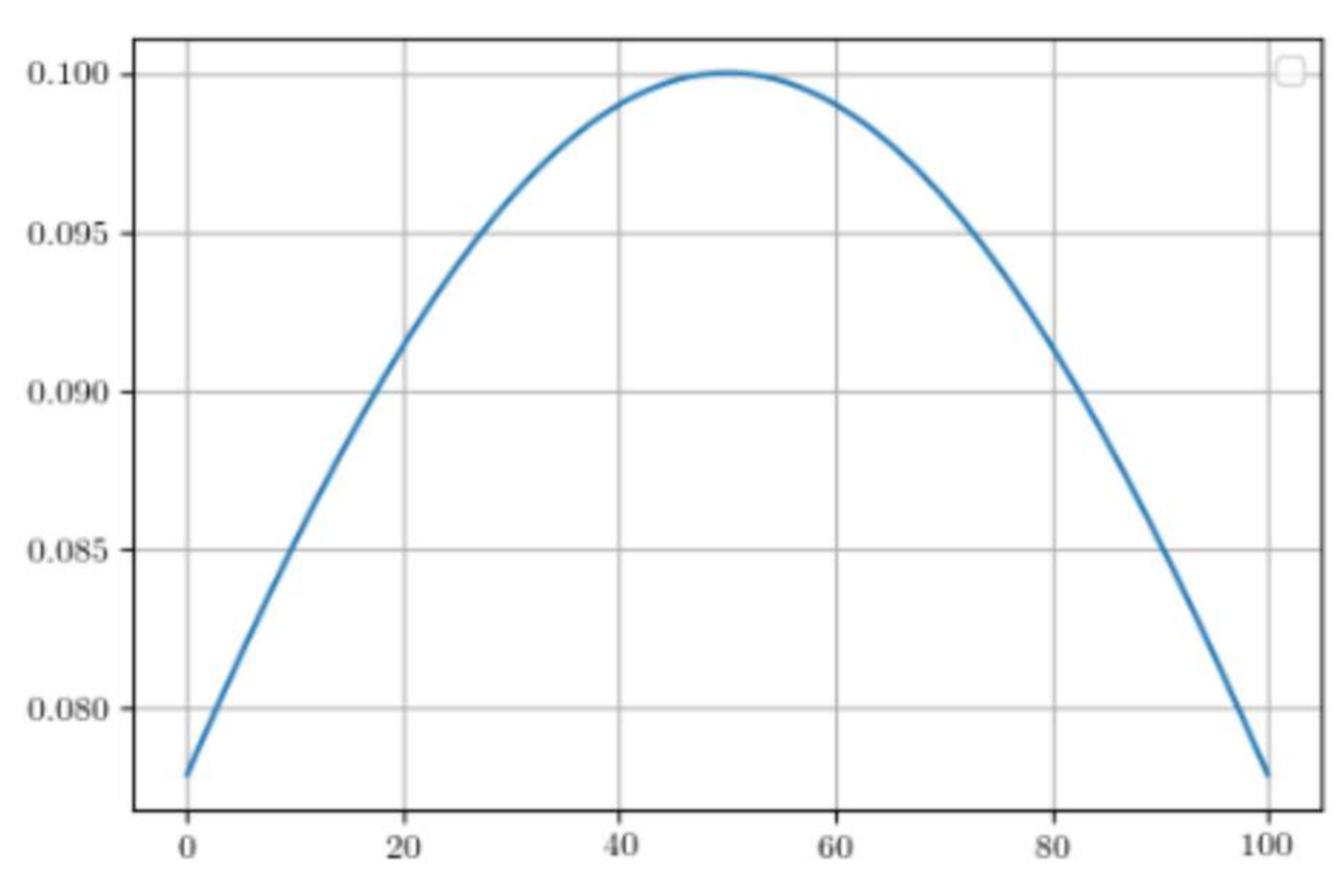
Yes, the answer is correct.
Score: 2

Accepted Answers:

False

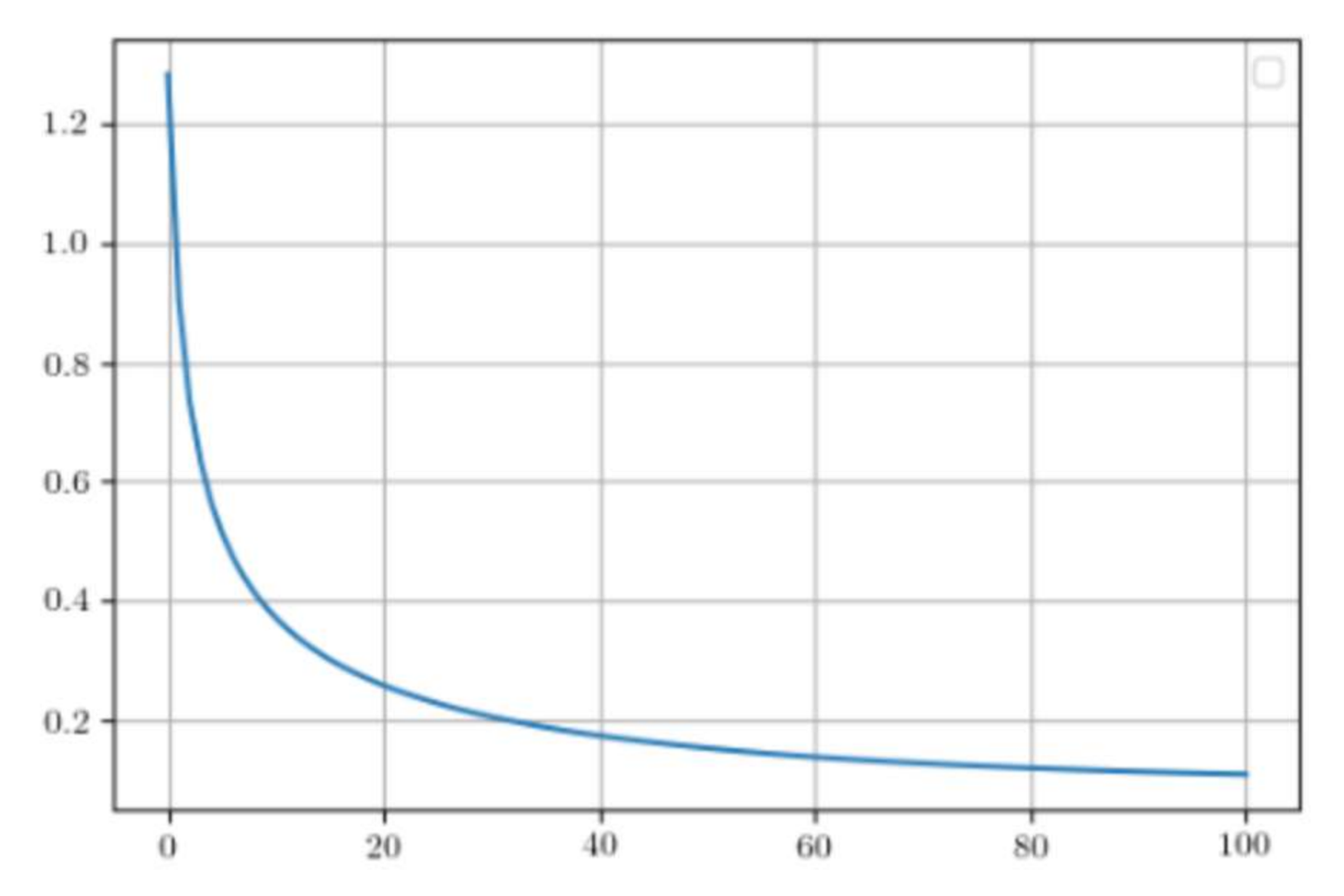
3) The figure below shows a gradient profile. Suppose we use the AdaGrad algorithm to compute v_t. Then which of the following figures show the effective learning rate of the algorithm *1 point*



- ( ) 
- (•) 
- ( ) 
- ( ) 

Yes, the answer is correct.
Score: 1

Accepted Answers:



Common Data Q4-Q6

Consider a gradient profile $\nabla w_t = [1, 0.9, 0.6, 0.01, 0.1, 0.2, 0.5, 0.55, 0.56]$.

Assume $v_{-1} = 0, \epsilon = 0, \beta = 0.9$ and the initial learning rate, $\eta_{-1} = 0.1$

4) Suppose that we use the AdaGrad algorithm. What is the value of η_6?

(enter the answer to 2 decimal points

0.063

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 0.05,0.07

1 point

5) Suppose that we compute η_{100}.Then the statement that the value of η_{100} will be greater than η_6 is *1 point*

- ( ) True
- (•) False
- ( ) Not possible to decide

Yes, the answer is correct.
Score: 1

Accepted Answers:

False

6) suppose that we use RMSProp algorithm. What is the value of η_6?

0.2548

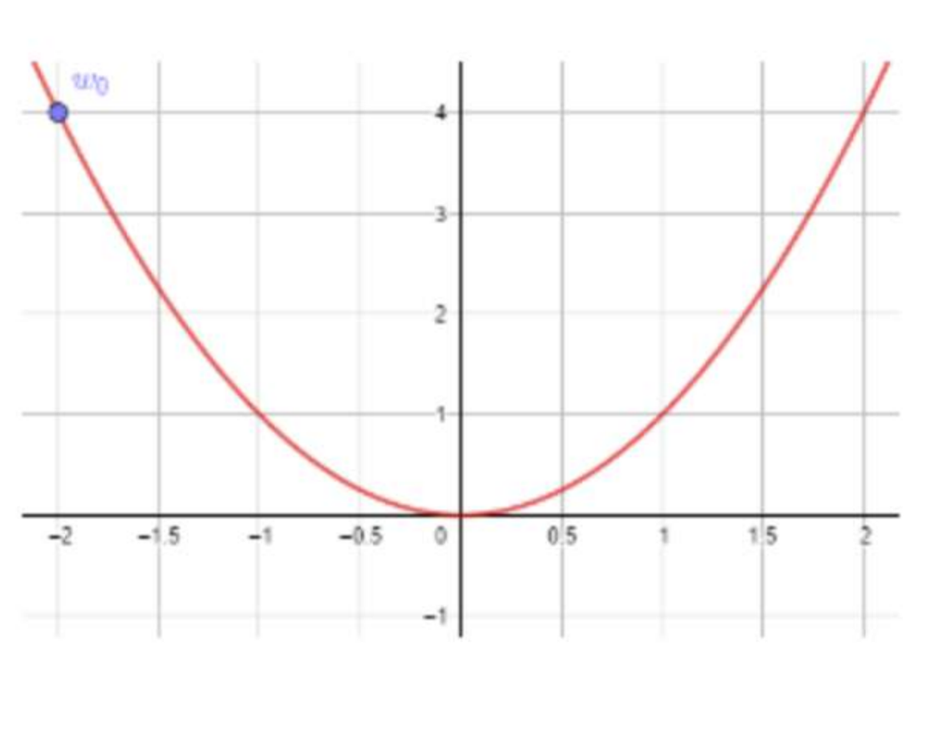
Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 0.23,0.27

1 point

7) The diagram below shows a loss curve $L(w)$. Suppose we decided to use Adam and Momentum based Gradient descent (MGD) algorithms to minimize the loss function. Suppose further that weights of both the algorithms are initialized at, $w_0 = -2$ with the initial learning rate set to 0.1. *2 points*



Suppose you run the algorithm for 10 iterations with the default values of β, that is, $\beta_1 = 0.9,\ \beta_2 = 0.999, \beta1_{-1} = \beta2_{-1} = 0$.

Which of the following statement(s) is (are) true?

(Of course, write a small piece of code)

- [x] MGD moves past the minimum because of added momentum
- [ ] Adam moves past the minimum because of added momentum
- [x] Adam doesn't move past the minimum because of adaptive learning rate
- [x] MGD is closer to the minimum than Adam, after 10 iterations
- [ ] Adam is closer to the minimum than MGD, after 10 iterations

Yes, the answer is correct.
Score: 2

Accepted Answers:

MGD moves past the minimum because of added momentum
Adam doesn't move past the minimum because of adaptive learning rate
MGD is closer to the minimum than Adam, after 10 iterations

Graded Assignment 6

The due date for submitting this assignment has passed.
Due on 2023-07-26, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

You have last submitted on: 2023-07-24, 23:51 IST

1) Consider two neural networks N_1 and N_2. The neural network N_1 contains K_1 hidden layers and the neural network N_2 contains K_2 hidden layers. Assume that the number of neurons in each hidden layer of both networks is n and $K_1 \gg K_2$. Which of these models has higher complexity? *1 point*

- (●) N_1
- ( ) N_2
- ( ) Randomly initializing the network parameters may lead to different weight values, leading to different complexity. Therefore, the given information is not sufficient.

Yes, the answer is correct.
Score: 1

Accepted Answers:

N_1

2) Which of the following statements are true according to the following equation used in slide 13 of the lecture, *1 point*

$E((y - \hat{f}(x))^2)$

- [x] The training error can be made zero with sufficiently complex models
- [ ] The training error cannot be made zero even with sufficiently complex models due to the presence of irreducible error
- [ ] The test error can be made zero with sufficiently complex models
- [x] The testing error cannot be made zero even with sufficiently complex models due to the presence of irreducible error

Yes, the answer is correct.
Score: 1

Accepted Answers:

The training error can be made zero with sufficiently complex models
The testing error cannot be made zero even with sufficiently complex models due to the presence of irreducible error

3) Suppose that a model produces zero training error. What happens if we use L_2 regularization, in general? *1 point*

- [x] It might increase training error
- [x] It might decrease test error
- [x] Reduce the complexity of the model by driving less important weights to close to zero
- [ ] It might decrease training error

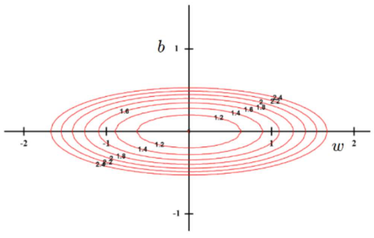
Yes, the answer is correct.
Score: 1

Accepted Answers:

It might increase training error
It might decrease test error
Reduce the complexity of the model by driving less important weights to close to zero

Common Data Q4-Q8

Consider a function $\mathscr{L}(w, b) = 0.5w^2 + 5b^2 + 1$ and its contour given below,



4) What is the value of $\mathscr{L}(w^*, b^*)$ where w^* and b^* are the values that minimize the function $\mathscr{L}(w, b)$?

1

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 0.9,1.1

1 point

5) What is the value of $\nabla\mathscr{L}(w^*, b^*)$?

0

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Numeric) 0

1 point

6) What is the determinant of $H\mathscr{L}(w^*, b^*)$, where H is the Hessian of the function?

10

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Numeric) 10

1 point

7) Compute the Eigenvalues and Eigenvectors of the Hessian. According to the eigenvalues of the Hessian, which parameter is the loss more sensitive to? *1 point*

- ( ) w
- (●) b

Yes, the answer is correct.
Score: 1

Accepted Answers:

b

8) Factorize the matrix using Eigenvalue decomposition. Then verify whether the following statement: ``applying regularization does not change the minimum of un-regularized function from (w^*, b^*) (that is, $(\tilde{w}, \tilde{b}) = (w^*, b^*)$' *1 point*

- (●) True
- ( ) False

Yes, the answer is correct.
Score: 1

Accepted Answers:

True

Graded Assignment 7

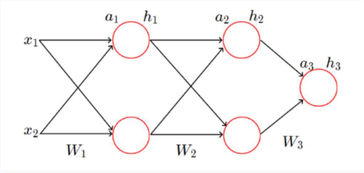
The due date for submitting this assignment has passed.
Due on 2023-07-31, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

You have last submitted on: 2023-07-31, 20:09 IST

Common Data for Q1 to Q4

Consider a neural network shown below.



The network uses the function

$(\frac{1}{2}\sum_{i=1}^{m}(\hat{y}-y)^2)$

to measure the loss. The network parameters are initialized as follows

$W_1 = W_2 = \begin{bmatrix} 10 & 0.1 \\ 0.25 & 10 \end{bmatrix}$

$W_3 = \begin{bmatrix} 10 & 0.1 \end{bmatrix}$

$b_1 = b_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$b_3 = 1.$

Also, the network parameters are updated using SGD algorithm with $\eta = 1$

1) Suppose that the input to the network is $(x_1 = 0, x_2 = 0)$ and the true label is $y = 0$. Assume that the Logistic activation function is used throughout the network. Do a forward propagation through the network and compute the loss value?

0.499

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 0.48,0.51

1 point

2) What is the value of $\frac{\partial L}{\partial b_3}$?

0.999

Yes, the answer is correct.
Score: 2

Accepted Answers:

(Type: Range) 0.9,1

2 points

3) What is the gradient of loss with respect to W_{31}, that is, $\nabla_{w_{31}}$? Round the answer to 2 decimal places

0

Yes, the answer is correct.
Score: 3

Accepted Answers:

(Type: Range) 0,0.01

3 points

4) Change the activation function of the output neuron to **ReLU**. Does the weight W_3 get updated by a larger value? *2 points*

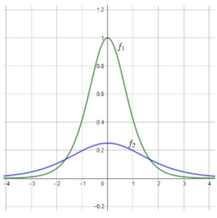
- (•) Yes
- ( ) No

Yes, the answer is correct.
Score: 2

Accepted Answers:

Yes

5) The figure below shows the derivate of two activation functions. Which one belongs to the logistic activation? *1 point*



- ( ) f_1
- (•) f_2
- ( ) None of these

Yes, the answer is correct.
Score: 1

Accepted Answers:

f_2

6) Consider a data point $(10, 0)$ where $x = 10$ and the respective label $y = 0$. Suppose that we use a logistic neuron with the w_0 and b_0 initialized to 10 and 0 respectively.Suppose further that we use MSE(Mean square error) for classification. The statement that `` If we run the algorithm for 1000 iterations, then the algorithm will converge with zero error" is *1 point*

- ( ) True
- (•) False

Yes, the answer is correct.
Score: 1

Accepted Answers:

False

7) Select the activation function that always gives zero mean and unit variance outputs. *1 point*

- ( ) ReLU
- ( ) ELU
- ( ) GELU
- (•) SELU

Yes, the answer is correct.
Score: 1

Accepted Answers:

SELU

Graded Assignment 8

The due date for submitting this assignment has passed.
Due on 2023-08-02, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

You have last submitted on: 2023-07-31, 20:12 IST

Common data:

Consider the input matrix X of shape 3×4 and the corresponding ground truth $y = 0.1$.

Suppose we use a kernel of size 1×2 and the kernel slides over the image with the stride $s = 2$.

Assume no zero padding.

$$X = \begin{bmatrix} 1 & -1 & 2 & -2 \\ 0 & 0 & 0 & 0 \\ 1 & -1 & 2 & -2 \end{bmatrix}$$

$$K = \begin{bmatrix} 0.6 & -0.2 \end{bmatrix}$$

Though you can answer many of the questions with a hand held calculator, you are advised to write a small piece of code to answer two questions at the end (Ya, backprop :-)).

1) Compute the convolution between X and K and store the output in the Matrix a_1. Choose the output dimension (that is, the dimension of the resulting feature map) ***1 point***

- 3×3
- 2×2
- 3×2
- 2×3

Yes, the answer is correct.
Score: 1

Accepted Answers:

2×2

2) Enter the sum of the elements in a_1

4.8

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 4.7,4.9

1 point

3) Now, pass the matrix a_1 through a sigmoid function and store the resultant matrix in h_1. What is the sum of elements in h_1?

3.02

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 3.00, 3.06

1 point

4) Flatten the matrix h_1 by concatenating rows (that is, $[h_{11}, h_{12}, h_{21}, h_{22}]$ to feed as an input to a sigmoid neuron as follows,

$$a_2 = h_{11} + h_{12} + h_{21} + h_{22}$$

note that this is simply a sum of all the inputs with the weight values fixed to 1 (not a learnable parameter).

$$\hat{y} = h_2 = \sigma(a_2)$$

where, $\hat{y}$ is the output of the network.

What is the value of $\hat{y}$?

0.95

Yes, the answer is correct.
Score: 2

Accepted Answers:

(Type: Range) 0.94,0.96

2 points

5) Compute the MSE loss $\mathscr{L} = \frac{1}{2}(\hat{y} - y)^2$.

0.36

Yes, the answer is correct.
Score: 3

Accepted Answers:

(Type: Range) 0.3,0.4

3 points

Now, do backpropagation to compute the gradients and update the weights $w_1 = 0.6$ and $w_2 = -0.2$. Assume $\eta = 1$.

6) Enter the value of $\frac{\partial L}{\partial a_2}$

0.04

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 0.030,0.050

1 point

7) Enter the updated weight value of w_1

0.45

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 0.42, 0.58

1 point

Graded Assignment 9

The due date for submitting this assignment has passed.
Due on 2023-08-11, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

You have last submitted on: 2023-08-11, 20:32 IST

Common Data for Q1 to Q7

A corpus contains the following sentences,
knowing the name of something is different from knowing something
knowing something about everything is alright

Note: Some calculations require coding. You can use NumPy.
Some useful functions are

np.outer()
np.linalg.svd()
np.transpose()
np.matmul() or @

1) Suppose that we construct a vocabulary that contains unique words in the corpus. What is the size of the vocabulary,$\mathbf{V}$? *1 point*

- 7
- 6
- 11
- 5

Yes, the answer is correct.
Score: 1

Accepted Answers:

11

2) Suppose that we use one hot encoding to convert the words in the sentences to vectors. Then each word in the sentences is converted to a vector of size?

11

Yes, the answer is correct.
Score: 1

Accepted Answers:

(Type: Range) 10.9,11.1

1 point

3) Construct a co-occurrence matrix using the vocabulary developed in question 1. However, drop the following words from the vocabulary (and hence from the sentences): of, the, alright, about, from. Arrange the words in the vocabulary in alphabetical order. Use a window of size 1, $k = 1$. How many non-zero entries are there in the matrix?

8

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Numeric) 16

1 point

4) By using the cooccurence matrix created in Q3, compute the cosine similarity between the words in the vocabulary. While computing cosine similarity, ensure that each word vector is normalized to have a unit magnitude. The word **knowing** is closest to which of the following words? *1 point*

- name
- something
- different
- everything
- is

No, the answer is incorrect.
Score: 0

Accepted Answers:

is

5) Compute the Pointwise Mutual Information (PMI) for the pair (knowing,something). Take $N = 9$ and $\log$ to the base 2. (Enter the answer upto 3 decimal places)

2.169

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Numeric) 1

1 point

6) Calculate the PPMI for Q5 and enter the value upto 3 decimal points

2.169

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Numeric) 1

1 point

7) Compute the SVD of the (normalized) co-occurrence matrix and take the rank-1 approximation (round the values in the matrix to 3 decimal points). Which of the following words are closer to the word knowing? We say the word is closer to the word \textbf{knowing} if its similarity score is greater than 0.5. *1 point*

- [ ] different
- [ ] everything
- [ ] name
- [x] something

No, the answer is incorrect.
Score: 0

Accepted Answers:

everything
name

8) Suppose that we use the continuous bag of words (CBOW) model to find vector representations of words. Suppose further that we use a context window of size 3 (that is, given the 3 context words, predict the target word $P(w_t|(w_i, w_j, w_k))$. The size of word vectors (vector representation of words) is chosen to be 100 and and the vocabulary contains 10000 words. The input to the network is the one-hot encoding (also called 1-of-V encoding) of word(s). How many parameters (weights), excluding bias, are there in W_{word}? Enter the answer in thousands. For example, if your answer is 50000, then just enter 50.

1000

Yes, the answer is correct.
Score: 2

Accepted Answers:

(Type: Numeric) 1000

2 points

Graded Assignment 10

The due date for submitting this assignment has passed.
Due on 2023-08-20, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

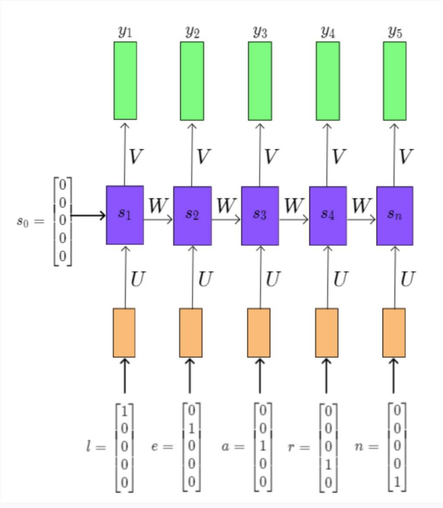
Common Data:

In this assignment, you will create an RNN model (using NumPy) that generates the sequence "learn" given the character "l" as input.

Each character in the word is represented using one hot encoding of size 5×1, for example

$$l = \begin{bmatrix}1\\0\\0\\0\\0\end{bmatrix}, \cdots, n = \begin{bmatrix}0\\0\\0\\0\\1\end{bmatrix}$$

The unrolled version of RNN architecture for the problem is shown below,



The shape of the variables involved is as follows,

$$U, W, V \in \mathbb{R}^{5\times 5}$$
$$s_t, y_t, x_t \in \mathbb{R}^{5\times 1}$$

where, y_t is a probability distribution over all 5 characters. Use cross entropy loss and gradient descent with $\eta = 1$

You can download the initial parameters for the network from here . Use the following code to load the weights

```
np.savez(`parameters',U=U,bu=bu,W=W,bw=bw,V=V,bv=bv)
parameters = np.load(`parameters.npz')
U = parameters.get(`U')
```

1) Suppose that we use the $\tanh(\cdot)$ activation function to compute the state vectors s_t. Do Forward propagation (that is feed in all characters one after the other and compute the loss for each time step) with the initialized parameters.

What is the loss value at the first time step $t = 1$, that is, $\mathscr{L}_1(\theta)$?

(Helper: the sum of elements in $s_1 = 0.18$)
Note: We haven't used the bias terms of W, U while computing s_t

---

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) 2.3,2.5

2 points

2) Suppose that we use $\tanh()$ activation to compute s_t. Do Forward propagation (that is feed in all characters one after the other and compute the loss for each time step) with the initialized parameters.Take, the true label for x_5 as a zero vector of size 5×1.

What is the total loss value , $\mathscr{L}(\theta)$?

(Helper: the sum of elements in $s_5 = -0.8$)

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) 12,12.5

2 points

3) Compute the grdient $\frac{\partial L}{\partial W}$ using BPTT algorithm and enter the sum of its elements.

(Helper: The maximum value of $\frac{\partial L}{\partial W}$ is at the index (5,5))

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) -0.05,-0.03

4 points

4) Update all the parameters with their respective gradients (that is, it completes one iteration by now). Now, run the model in ``Auto-regressive mode" (that is, the input to the next time step is the output from the previous time step). Feed the input ``l" to the model and choose the sequence of characters it generated for the next five consecutive steps. ***2 points***

- enlrn
- enrrn
- lnlnl
- llllll
- nelnr

No, the answer is incorrect.
Score: 0

Accepted Answers:

enlrn

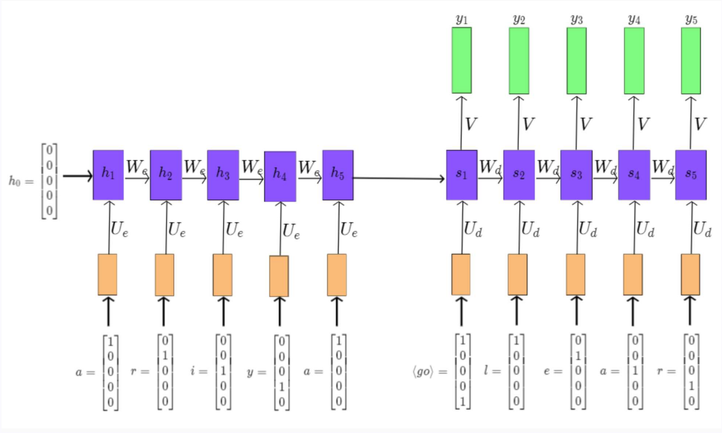
Graded Assignment 11

The due date for submitting this assignment has passed.
Due on 2023-08-31, 23:59 IST.

You may submit any number of times before the due date. The final submission will be considered for grading.

Common Data:

In this assignment, we extend the RNN model implemented in the previous assignment to translate the sequence of characters ``$x_{source}$=ariya" (in Tamil) to ``x_{target}=learn" (in English) using a simple encoder decoder architecture as show below. Each character is represented using one hot encoding of size 5×1 as shown in the figure



The shape of the variables is as follows,
$U_e, W_e, W_d, U_d, V \in \mathbb{R}^{5\times 5}$ $h_t, s_t, x_{source}, x_{target} \in \mathbb{R}^{5\times 1}$ y_t is a probability distribution over all 5 characters (in the vocabulary) at time step t. Let the initial state of the decoder is $s_0 = h_5$. Use cross entropy loss and gradient descent with $\eta = 1$.

You can download the initial parameters for the network from here. Use the following code to load the weights,

```
np.savez(`parameters_w11',U_e=U_e,W_e=W_e,W_d=W_d,U_d=U_d,V_d=V_d)
parameters = np.load(`parameters.npz')
U_e = parameters.get(`U_e')
```

1) Suppose that we use the $\tanh()$ activation function to compute the state vectors of encoder and decoder (h_t, s_t), respectively.

Do forward propagation with the initialized parameters. What is the total loss value $\mathscr{L}(\theta)$?

Assume the character `` n " is encoded as

$$n = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

(Helper: the sum of elements in $s_5 = -1.04$)

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) 17.1,17.6

3 points

2) Initialize the decoder with, $s_0 = h_5$. Run the decoder in Auto-regressive mode. The source sequence is translated into which of the following sequences? *1 point*

- rnnrn
- leenl
- rnlnn
- learn

No, the answer is incorrect.
Score: 0

Accepted Answers:

rnlnn

3) Train the model for 20 epochs. What is the total loss value now?

(Helper: the sum of elements in $s_5 = -4.84$)

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: Range) 0.04,0.06

3 points

4) With the updated parameters, compute h_5 of encoder and initialize the decoder with $s_0 = h_5$. Run the decoder(with the updated parameters) in Auto-regressive mode. The source sequence is translated into which of the following sequences? *1 point*

- rnnrn
- leenl
- rnlnn
- learn

No, the answer is incorrect.
Score: 0

Accepted Answers:

learn

5) Shuffle the characters in the source sequence and repeat Q4. What do you observe?

No, the answer is incorrect.
Score: 0

Accepted Answers:

(Type: String) h

0 points

6) Suppose that we use attention mechanism in the encoder and decoder model by computing the context vector as follows, *2 points*

$c_t = \sum_{j=1}^{T} \alpha_{jt} h_j$ what is the length of T?

- 5
- 4
- 10
- 8

No, the answer is incorrect.
Score: 0

Accepted Answers:

5