

# Quiz 3



Quiz submitted



- Due No due date
- Points 10
- Questions 10
- Available Jan 30 at 6pm - Feb 1 at 11:59pm
- Time Limit None
- Allowed Attempts 3

## Instructions

This quiz primarily covers lectures 5-6, but you are expected to be familiar with concepts from previous lectures as well.

Several of the questions refer to hidden slides that were not presented in class.

Some of the questions also require you to read additional material, links to which are posted in the quiz questions.

[Take the Quiz Again](#)

## Attempt History

	Attempt	Time	Score
LATEST	<a href="#">Attempt 1</a>	89 minutes	6.75 out of 10

**!** Correct answers are hidden.

Score for this attempt: 6.75 out of 10

Submitted Feb 1 at 10:49pm

This attempt took 89 minutes.



Question 1

1 / 1 pts

For this question, please read the paper: [Rumelhart, Hinton and Williams \(1986\)](#) ↗

[\(<http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf>\)](http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf) ↗

[\(<http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf>\)](http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf).

[Can be found at: <http://www.cs.toronto.edu/~hinton/absps/naturebp.pdf>]

One version of gradient descent changes each weight by an amount proportional to the accumulated  $\delta E/\delta w$ .  Quiz submitted

$$\Delta w = -\epsilon \frac{\delta E}{\delta w}$$

Select all that are true about this method:

It cannot be implemented by local computations in parallel hardware.

It's simpler than methods that use second derivatives.

"This method does not converge as rapidly as methods which make use of the second derivatives, but it is much simpler [...]" p535

This method converges as rapidly as methods that make use of second derivatives.

It can be improved without sacrificing simplicity and locality.

"It can be significantly improved, without sacrificing the simplicity and locality, [...]" p535

Partial



Question 2

0.75 / 1 pts

(Select all that apply) As discussed in lecture, which of the following is true for the backpropagation algorithm?

Hint: Lecture 5, starting at "training by backprop". Lec 5 (Pages 58 - 92)

It computes the derivative of the divergence between the true and desired outputs of the network for a training input

It can be used to compute the derivative of the divergence with respect to the input of the network

It computes the derivative of the average divergence for a batch of inputs

It cannot be performed without first doing a feed-forward pass of the input(s) through the network

It is used to compute derivatives that are required for the gradient descent algorithm that trains the network



Question 3

1 / 1 pts

Consider a perceptron in a network that has the following vector activation:

$$y_j = \prod_{j \neq i} z_i$$

Where  $y_j$  is the j-th component of column vector  $y$ , and  $z_i$  is the i-th component of column vector  $z$ . Using the notation from lecture, which of the following is true of the derivative of  $y$  w.r.t.  $z$ ? (select all that are true)

Hint: Vector Calculus Notes 1 (lecture 5, slide 135 and beyond)

- It is a matrix where all components are equal.
- Quiz submitted
- It will be a matrix.

- It is a column vector whose  $i$ -th component is given by  $\prod_{j \neq i} z_j$
- It is a row vector whose  $i$ -th component is given by  $\prod_{j \neq i} z_j$
- It is a matrix whose  $(i, j)$ th component where  $i \neq j$  is given by  $\prod_{k \neq i, k \neq j} z_k$

Incorrect



#### Question 4

0 / 1 pts

Let  $d$  be a scalar-valued function with multivariate input,  $f$  be a vector-valued function with multivariate input, and  $X$  be a vector such that  $y = d(f(X))$ . Using the lecture's notation, assuming the output of  $f$  to be a column vector, the derivative  $\nabla_f y$  of  $y$  with respect to  $f(X)$  is...

Hint: (Lecture 4 and) Lecture 5, Vector calculus, Notes 1.

- A row vector
- Composed of the partial derivatives of  $y$  w.r.t the components of  $X$
- A matrix
- A column vector

#### Question 5

1 / 1 pts

Which of the following is true given the Hessian of a scalar function with multivariate inputs?

Hint: Lec 4 "Unconstrained minimization of a function". Also note that an eigen value of 0 indicates that the function is flat (to within the second derivative) along the direction of the corresponding Hessian Eigenvector.

- At a local minima, the Hessian matrix has eigenvalues that are non-negative, at least one eigenvalue might be 0.
- The eigenvalues are all strictly positive at global minima, but not at local minima.
- If the eigenvalues of a hessian matrix are all strictly negative, then the function is at a local maximum.
- If the eigenvalues of a hessian matrix are all strictly positive, then the function is at a local minimum.

Incorrect



#### Question 6

0 / 1 pts

The KL divergence between the output of a multi-class network with softmax output  $y = [y_1 \dots y_K]$  and desired output  $d = [d_1 \dots d_K]$  is defined as  $KL = \sum_i d_i \log d_i - \sum_i d_i \log y_i$ . The first term on the right hand side is the entropy of  $d$ , and the second term is the Cross-entropy between  $d$  and  $y$ .

which we will represent as  $Xent(y, d)$ . Minimizing the KL divergence is strictly equivalent to minimizing the cross entropy.  Quiz submitted

we do this, we refer to  $Xent(y, d)$  as the cross-entropy loss.

Defined in this manner, which of the following is true of the cross-entropy loss  $Xent(y, d)$ ? Recall that in this setting both  $y$  and  $d$  may be viewed as probabilities (i.e. they satisfy the properties of a probability distribution).

- It's derivative with respect to  $y$  goes to zero at the minimum (when  $y$  is exactly equal to  $d$ )
- It goes to 0 when  $y$  equals  $d$
- It is always non-negative
- It only depends on the output value of the network for the correct class

If  $d$  is not one hot (e.g. when we use label smoothing), the cross entropy may not be 0 when  $d = y$ .

For one-hot  $d$ , we saw in class that the KL divergence is equal to the cross entropy. Also, in this case, at  $d=y$ , the gradient of the DL divergence (and therefore  $Xent(y,d)$ ) is not 0.

Incorrect



Question 7

0 / 1 pts

Tom decides to construct a new vector activation function based on the Softplus to output probabilities.

$$SR(z_i) = \frac{Softplus(z_i)}{\sum_j Softplus(z_j)}$$

Which of the following statements is true (multiple choice).

Hint: To understand the Softplus check [https://en.wikipedia.org/wiki/Rectifier\\_\(neural\\_networks\)](https://en.wikipedia.org/wiki/Rectifier_(neural_networks)). The above is also similar to the Softmax activation with Softplus replacing the exponential function (Which are similar: smooth and monotonically increasing). You can check the derivatives in lecture 5 slides 99 - 102

- The derivative of  $SR(z_i)$  with respect to  $z_j$  (for  $j \neq i$ ) will be positive
- The derivative of  $SR(z_i)$  with respect to  $z_i$  will be negative
- 

The sign of the derivative of  $SR(z_i)$  with respect to  $z_j$  (for  $j \neq i$ ) depends on  $z_i$  and  $z_j$ , and cannot be predicted without knowing them

- The derivative
- The derivative

Quiz submitted

The sign of the derivative of  $SR(z_i)$  with respect to  $z_j$  (for  $j \neq i$ ) depends on the signs of  $z_i$  and  $z_j$ , and cannot be predicted without knowing them

Softplus is smooth and monotonically increasing and derivative of a similar vector activation is shown on lecture 5 slides 99 - 102



### Question 8

1 / 1 pts

In order to maximize the possibility of escaping local minima and finding the global minimum of a generic function, the best strategy to manage step sizes during gradient descent is:

Hint: Lecture 6, "Issues 2"



To start with a large, divergent step size (e.g. greater than twice the optimal step size for a quadratic approximation at the initial location) and gradually decrease it over iterations

To keep the step size low throughout to prevent divergence into a local minima



To maintain a step size consistently close to the optimal step size (e.g. close to the inverse second derivative at the current estimate)



To start with a large, non-divergent step size (e.g. less than twice the optimal step size for a quadratic approximation at the initial location) and gradually decrease it over iterations



### Question 9

1 / 1 pts

Gradient descent with a fixed step size \_\_\_\_\_ for all convex functions (Fill in the blank)

Hint: Lecture 6

- Always converges to a local minimum
- Does not always converge
- Always converges to a global minimum
- Always converges to some point



### Question 10

1 / 1 pts

Let  $f$  be a quadrat



Quiz submitted

= 1. The

minimum has a value of  $x = \boxed{5}$  and a value of  $f(x) = \boxed{2}$ . (Truncate

your answer to 1 digit after the decimal point i.e. enter your answer in the format x.x, e.g. 4.5)

Hint: Lecture 6 "Convergence for quadratic surfaces"

**Answer 1:**

5

**Answer 2:**

2

Quiz Score: 6.75 out of 10