Quiz 3

- 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
KEPT	Attempt 2	22 minutes	6.75 out of 10
LATEST	Attempt 2	22 minutes	6.75 out of 10
	Attempt 1	89 minutes	6.75 out of 10

Correct answers are hidden.

Score for this attempt: 6.75 out of 10
Submitted Feb 1 at 11:11pm
This attempt took 22 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) .

[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 δE/δw.

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

Select all that are true about this method:

- [ ] This method converges as rapidly as methods that make use of second derivatives.
- [x] 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

- [ ] It cannot be implemented by local computations in parallel hardware.
- [x] 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)

- [x] It cannot be performed without first doing a feed-forward pass of the input(s) through the network
- [ ] 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
- [x] It is used to compute derivatives that are required for the gradient descent algorithm that trains the network
- [x] It computes the derivative of the divergence between the true and desired outputs of the network for a training input

Question 3

1 / 1 pts

(Select all that apply) At any point, the gradient of a scalar function with multivariate inputs…

Hint: Lecture 4, "Gradient of a scalar function of a vector" and "properties of a gradient" Lec 4 (Pages 10 - 21) .

- [x] Is the vector of local partial derivatives w.r.t. all the inputs
- [ ] Is parallel to equal-value contours of the function

- [ ] Is in the direction of steepest descent
- [x] Is in the direction of steepest ascent

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.

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

Question 5
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)

- [x] It is a matrix whose (i, j)th component where $i \neq j$ is given by $\prod_{k \neq i, k \neq j} z_k$
- [ ] 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 is given by $z_i z_j$
- [x] It will be a matrix whose diagonal entries are all 0.
- [ ] It is a column vector whose i-th component is given by $\prod_{j \neq i} z_j$

Incorrect

Question 6
0 / 1 pts

The KL divergence between the output of a multi-class network with softmax output $\boldsymbol{y} = [y_1 \ldots y_K]$ and *desired* output $\boldsymbol{d} = [d_1 \ldots d_K]$ is defined as $\boldsymbol{KL} = \sum_i d_i \log d_i - \sum_i d_i \log y_i$. The first term on the right hand side is the entropy of $\boldsymbol{d}$, and the second term is the *Cross-entropy* between $\boldsymbol{d}$ and $\boldsymbol{y}$, which we will represent as $\boldsymbol{Xent(y,d)}$. Minimizing the KL divergence is strictly equivalent to minimizing the cross entropy, since $\sum_i d_i \log d_i$ is not a parameter of network parameters. When we do this, we refer to $\boldsymbol{Xent(y,d)}$ as the cross-entropy loss.

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

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

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), 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 != 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 != i) depends on the signs of z_i and z_j, and cannot be predicted without knowing them
- [ ] The derivative of SR(z_i) with respect to z_j (for j != i) will be negative
- [x] The derivative of SR(z_i) with respect to z_i will be positive
- [x] The sign of the derivative of SR(z_i) with respect to z_j (for j != i) depends on 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

Gradient descent yields a solution that is not sensitive to how a network's weights are initialized.

Hint: Basic gradient descent from lecture 5 - slide 5

- ( ) True
- (•) False

Question 9
1 / 1 pts

What are the challenges of using Newton's method with neural networks?

Hint: Lecture 6, "Issues 1"

- [x] Its memory usage scales with the square of the number of weights
- [x] It can produce unstable updates if the optimized function is not strictly convex
- [ ] It cannot find the minimum in any quadratic function
- [ ] It has a very large Jacobian
- [x] It is difficult to compute the inverse of the Hessian

Question 10
1 / 1 pts

Let $f(.)$ be a scalar-valued function with multivariate input and $x = [x_1, x_2]$ be a two-component vector such that $y = f(x)$. y is being minimized using RProp from lecture. In the k-th iteration, the derivative of y with respect to x_1 is $\frac{dy}{dx_1} = 2$, the derivative of y with respect to x_2 is $\frac{dy}{dx_2} = -1$. As a result, x_1 has a step size of $\Delta x_1^{(k)} = 1$ and x_2 has a step size of $\Delta x_2^{(k)} = 1$. At the (k+1)-th iteration,

the derivative of y with respect to x_1 is $\frac{dy}{dx_1} = 0.5$ and the derivative of y with respect to x_2 is $\frac{dy}{dx_2} = 1$. Which of the following is true about the step size at the (k+1)-th iteration?

Hint: Lecture 6, RProp Slide: 112-122

- $\Delta x_1^{(k+1)} < 1$ and $\Delta x_2^{(k+1)} < 1$
- $\Delta x_1^{(k+1)} < 1$ and $\Delta x_2^{(k+1)} > 1$
- $\Delta x_1^{(k+1)} > 1$ and $\Delta x_2^{(k+1)} > 1$
- $\Delta x_1^{(k+1)} > 1$ and $\Delta x_2^{(k+1)} < 1$

Quiz Score: 6.75 out of 10