

❗ Correct answers will be available on Feb 6 at 12am.


Score for this attempt: 8 out of 10

Submitted Feb 5 at 2:45pm

This attempt took 72 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>).

[Can be found at:

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

One drawback of the learning procedure in the paper is that the error-surface may contain local minima so that gradient descent is not guaranteed to find a global minimum. This happens if the network has **more** than enough connections.

☐ True

☒ False

"Adding a few more connections creates extra dimensions in weight-space and these dimensions provide paths around the barriers that create poor local minima in the lower dimensional subspaces" -p535

Answer key: Happens if the network has *just* enough connections. The question here asks if the network has “more than enough” connections, which in that case, it will be able to create a path to go around this barrier.

Question 2

1 / 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 gradient descent".



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



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



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

Question 3

1 / 1 pts

Let \mathbf{d} be a scalar-valued function with multivariate input, \mathbf{f} be a vector-valued function with multivariate input, and \mathbf{X} be a vector such that $\mathbf{y} = \mathbf{d}(\mathbf{f}(\mathbf{X}))$. Further, $\mathbf{J}_{\mathbf{f}}(\mathbf{X})$ is the Jacobian of \mathbf{f} w.r.t \mathbf{X} . Using the lecture's notation, the derivative of \mathbf{y} w.r.t. \mathbf{X} is...

Hint: Lecture 5, Vector Calculus, Notes 1 and 2



Either a column vector given by $J_f(X) \nabla_f \text{y}$ or a row vector given by $\nabla_f \text{y} J_f(X)$



A matrix given by $\nabla_f \text{y} J_f(X)$



A column vector given by $J_f(X) \nabla_f \text{y}$



A row vector given by $\nabla_f \text{y} J_f(X)$

Question 4

1 / 1 pts

We are given a binary classification problem where the training data from both classes are linearly separable. We compare a perceptron, trained using the perceptron learning rule with a sigmoid-activation perceptron, trained using gradient descent that minimizes the L2 Loss. In both cases, we restrict the weights vector of the perceptron to have finite length. In all cases, we will say the algorithm has found a “correct” solution if the learned model is able to correctly classify the training data. Which of the following statements are true (select all that are true).

Hint: See slides 13-32, lecture 6



There are situations where the gradient-descent algorithm will not find the correct solution.



The perceptron algorithm will always find the correct solution.



The gradient-descent algorithm will always find the correct solution.



We cannot make any statement about the truth or falsity of the other options provided, based only on the information provided.

Incorrect

Question 5

0 / 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.

☒ The eigenvalues are all strictly negative at a local maximum.

☒ The eigenvalues are all strictly positive at a local minimum.

☒ The eigenvalues are all non-negative at local minima.



The eigenvalues are all strictly positive at global minima, but not at local minima.

Incorrect

Question 6

0 / 1 pts

The KL divergence between the output of a multi-class network with softmax output $\mathbf{y} = [y_1 \dots y_K]$ and *desired* output $\mathbf{d} = [d_1 \dots d_K]$ is defined as $KL = \sum_i d_i \log d_i - \sum_i d_i \log y_i$. The first term on the

entropy between \mathbf{d} and \mathbf{y} , which we will represent as $Xent(\mathbf{y}, \mathbf{d})$. 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 $Xent(\mathbf{y}, \mathbf{d})$ as the cross-entropy loss.

Defined in this manner, which of the following is true of the cross-entropy loss $Xent(\mathbf{y}, \mathbf{d})$? Recall that in this setting both \mathbf{y} and \mathbf{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.

Question 7

1 / 1 pts

Which of the following are valid subgradients of a RELU, given by $\text{Relu}(x)$, at $x = 0$? We will represent the subgradient as

$$\nabla_{\text{subgrad}} \text{RELU}(x)$$

Hint: Lecture 5, slides 112-114.



$$\nabla_{\text{subgrad}} \text{RELU}(0) = 0.5$$



$$\nabla_{\text{subgrad}} \text{RELU}(0) = 0$$



$$\nabla_{\text{subgrad}} \text{RELU}(0) = 1.5$$



$$\nabla_{\text{subgrad}} \text{RELU}(0) = 1$$



$$\nabla_{\text{subgrad}} \text{RELU}(0) = -0.5$$

Question 8

1 / 1 pts

Consider the class of twice differentiable convex functions (assume univariate scalar functions unless otherwise specified). Which of the following are true when minimizing a function using gradient descent? (select all that apply)

Hint: Lecture 6 - slide 46

☐

It will converge to the optimum while oscillating if the step size is less than the inverse of the second derivative of the function

☒

It will converge to the optimum monotonically and without oscillating if the step size is less than the inverse of the second derivative of the function

As explained in class

☒

It will diverge if the step size is more than twice the optimal step size for the quadratic approximation of the function at the current point.

As explained in class

☐

It will converge quickly if the step size is twice the inverse of the second derivative of the function at the current point.

☒

It has an optimal step size equal to the inverse of the second derivative of the function for the quadratic approximation of the function at the current point.

As explained in class

Question 9

1 / 1 pts

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

Hint: Lecture 6, "Issues 1"

☐ It has a very large Jacobian

☐ It cannot find the minimum in any quadratic function

☒ It is difficult to compute the inverse of the Hessian

☒ It can produce unstable updates if the optimized function is not strictly convex

☒ Its memory usage scales with the square of the number of weights

Question 10

1 / 1 pts

Let f be a quadratic function such that at $x = 1$, $f(x) = 10$, $f'(x) = -4$, and $f''(x) = 1$. The minimum has a value of $x =$

5

and a value of $f(x) =$

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: **8** out of 10