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

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

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

Incorrect

Question 2

0 / 1 pts

(**Select all that apply**) Which of the following is true of the vector and scalar versions of backpropagation?

Hint: Lecture 5

- [x] Both scalar backpropagation and vector backpropagation are optimization algorithms that are used to find parameters that minimize a loss function

- [x] Scalar backpropagation rules explicitly loop over the neurons in a layer to compute derivatives, while vector backpropagation computes derivative terms for all of them in a single matrix operation

- [ ] Scalar backpropagation and vector backpropagation only differ in their arithmetic notation and the implementation of their underlying arithmetic

- [ ] Scalar backpropagation is required for scalar activation functions, while vector backpropagation is essential for vector activation functions

Incorrect

Question 3

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.

- [x] 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.
- [x] The eigenvalues are all strictly negative at a local maximum.

Question 4

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 (and beyond)

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

- [ ] It is a row vector whose i-th component is given by $\prod_{j \neq i} z_j$

- [x] It is a matrix whose (i, j)th component where $i \neq j$ is given by $\prod_{k \neq i, k \neq j} z_k$

Question 5

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".

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

Question 6

1 / 1 pts

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

Hint: Lecture 5, slides 112-114.

- [x] $\nabla_{subgrad} RELU\,(0) = 0.5$

- [x] $\nabla_{subgrad} RELU\,(0) = 0$
- [x] $\nabla_{subgrad} RELU\,(0) = 1$
- [ ] $\nabla_{subgrad} RELU\,(0) = 1.5$
- [ ] $\nabla_{subgrad} RELU\,(0) = -0.5$

Incorrect

Question 7

0 / 1 pts

The KL divergence between the output of a multi-class network with softmax output $y = [y_1 \ldots y_K]$ and *desired* output $d = [d_1 \ldots d_K]$ 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, since $\sum_i d_i \log d_i$ is not a parameter of network parameters. When 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)
- [x] It is always non-negative
- [x] It only depends on the output value of the network for the correct class
- [ ] It goes to 0 when y equals d

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 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

- [x] 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

- [x] 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 will converge to the optimum while oscillating if the step size is less than the inverse of the second derivative of the function

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

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

Hint: Lecture 6

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

Question 10

1 / 1 pts

Let $f(.)$ be an affine function that you would like to optimize. At your current location, $x = 3$, $f(x) = 7$ and $f'(x) = 2$. After one iteration of gradient descent with a learning rate = 0.1, your new location has a value of x = 2.8 and a value of $f(x)$ = 6.6 .

(Truncate your answer to 1 digit after the decimal point, i.e. enter your answer in the format x.x, e.g. 4.5. If you use any other format canvas may mark your answer as being wrong)

Hint: Basic gradient descent from the lectures.

Answer 1:

2.8

Answer 2:

6.6

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.

Incorrect

Question 2

0 / 1 pts

(**Select all that apply**) Which of the following is true of the vector and scalar versions of backpropagation?

Hint: Lecture 5

- [x] Both scalar backpropagation and vector backpropagation are optimization algorithms that are used to find parameters that minimize a loss function
- [ ] Scalar backpropagation is required for scalar activation functions, while vector backpropagation is essential for vector activation functions
- [ ] Scalar backpropagation rules explicitly loop over the neurons in a layer to compute derivatives, while vector backpropagation computes derivative terms for all of them in a single matrix operation
- [ ] Scalar backpropagation and vector backpropagation only differ in their arithmetic notation and the implementation of their underlying arithmetic

Incorrect

Question 3

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.

- [x] The eigenvalues are all non-negative at local minima.
- [ ] The eigenvalues are all strictly negative at a local maximum.
- [ ] The eigenvalues are all strictly positive at a local minimum.
- [x] The eigenvalues are all strictly positive at global minima, but not at local minima.

Question 4

1 / 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.

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

Question 5

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".

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

Question 6

1 / 1 pts

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

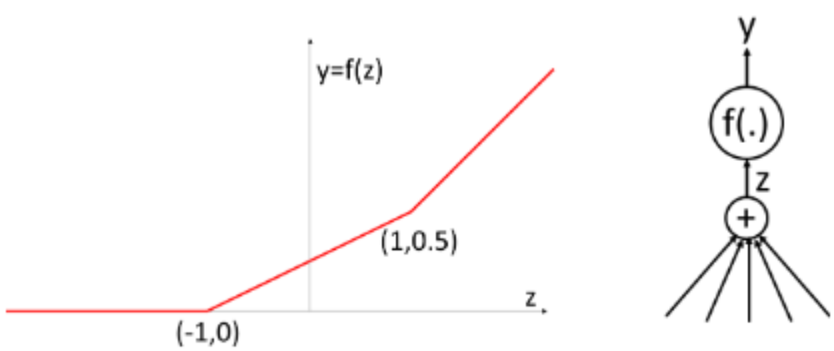
Hint: Lecture 5, slides 112-114.

- [ ] $\nabla_{subgrad} RELU\left(0\right) = 1.5$
- [x] $\nabla_{subgrad} RELU\left(0\right) = 1$
- [ ] $\nabla_{subgrad} RELU\left(0\right) = -0.5$
- [x] $\nabla_{subgrad} RELU\left(0\right) = 0$
- [x] $\nabla_{subgrad} RELU\left(0\right) = 0.5$

Question 7

1 / 1 pts

The following piecewise linear function with “hinges” at (-1,0) and (1,0.5) is used as an activation for a neuron. The slope of the last segment is 40 degrees with respect to the *z* axis (going anti-clockwise). Our objective is to find a z that minimizes the divergence div(y,d). Which of the following update rules is a valid subgradient descent update rule at z=1? Here η is the step size and is a positive number. The superscript on *z* represents the step index in an iterative estimate. The derivative $\frac{\partial div(y,\, d)}{\partial\, z}$ is computed at $z^k = 1$. The value of η must not factor into your answer (i.e. remember that η has only been included in the equations for completeness sake and do not argue with us that you can always adjust η to make any answer correct ☺)



Hint: Lecture 5, slides 112-114

- [x] $z^{k+1} = z^k - \eta 0.75 \frac{\partial div(y,d)}{\partial y}$
- [ ] $z^{k+1} = z^k - \eta \frac{\partial div(y,d)}{\partial y}$
- [x] $z^{k+1} = z^k - \eta 0.25 \frac{\partial div(y,d)}{\partial y}$
- [ ] $z^{k+1} = z^k - \eta 0.1 \frac{\partial div(y,d)}{\partial y}$
- [ ] $z^{k+1} = z^k + \eta \frac{\partial div(y,d)}{\partial y}$

Incorrect

Question 8

0 / 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

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 an affine function that you would like to optimize. At your current location, $x = 3$, $f(x) = 7$ and $f'(x) = 2$. After one iteration of gradient descent with a learning rate = 0.1, your new location has a value of x = 2.8 and a value of $f(x)$ = 6.6 .

(Truncate your answer to 1 digit after the decimal point, i.e. enter your answer in the format x.x, e.g. 4.5. If you use any other format canvas may mark your answer as being wrong)

Hint: Basic gradient descent from the lectures.

Answer 1:

2.8

Answer 2:

6.6