Quiz 4

- Due No due date
- Points 10
- Questions 10
- Available Feb 6 at 6pm - Feb 8 at 11:59pm
- Time Limit None
- Allowed Attempts 3

Take the Quiz Again

Attempt History

	Attempt	Time	Score
LATEST	Attempt 1	62 minutes	9 out of 10

ⓘ Correct answers are hidden.

Score for this attempt: 9 out of 10
Submitted Feb 8 at 11:20pm
This attempt took 62 minutes.
Incorrect

Question 1
0 / 1 pts

You are trying to minimize the cross-entropy loss of the logistic function $y = \frac{1}{1+exp(0.5w)}$ with respect to parameter w, when the target output is 1.0. Note that this corresponds to optimizing the logistic function $y = \frac{1}{1+exp(-wx)}$, given only the training input (x, d(x)) = (-0.5, 1). You are using Nestorov's updates. Your current estimate (in the k-th step) is $w^{(k)} = 0$. The last step you took was $\Delta w^{(k)} = 0.5$. Using the notation from class, you use $\beta = 0.9$ and $\eta = 0.1$. What is the value of $w^{(k+1)}$ when using Nestorov's update? Truncate the answer to three decimals (**do not round up**).

Hint: The cross entropy loss is identical to the KL divergence (lec8, "choices for divergence"), when the target output is binary (or, more generally, one hot).

0.456

[REFER TO THE APPROPRIATE SLIDES FROM LECT 4 INSTEAD] when the target output is binary (or, more generally, one hot).

Question 2
1 / 1 pts

Several researchers separately decide to estimate an unknown function d(x), for a variable x. Although they do not know the function, they do have access to an oracle who does know d(x). Upon demand the oracle will randomly draw a value x from a ***uniform*** probability distribution P(x)=1, 0≤x≤1 and return (x,d(x)), i.e. a random value of x and the value of the function d(x) at that x. Each of the researchers independently obtains 1000 training pairs from the oracle, and begins to estimate d(x) from them. They begin with an initial estimate of d(x) as y(x)=0 (where y(x) is the estimate of d(x)They do not update y(x) during this exercise). They plan to process their training data using stochastic gradient descent. So each of them computes the L2 divergence (as defined in lecture 4) between the estimated output and the desired output for each of their training instances. They do not update y(x) during the exercise.

In order to get a better handle on their problem, the researchers then get together and pool their divergences over all of their combined training instances, and compute the statistics of the collection of divergences.

Unknown to them (but known to the oracle), $d(x) = \sqrt{x}$
What would you expect the average of the set of divergences to be?
Truncate the answer to 2 decimals (and write both, even if the answer has the form *.00)

Hint: Lec 7, slide 56-96. The L2 divergence is as defined in Lec 4, “Examples of divergence functions”, i.e ½ times the squared Euclidean error (Slides 88-90).

0.25

Question 3
1 / 1 pts

Several researchers separately decide to estimate an unknown function d(x), for a variable x. Although they do not know the function, they do have access to an oracle who does know d(x). Upon demand the oracle will randomly draw a value x from a ***Exponential*** probability distribution $P(x) = Exponential(\lambda = 0.1)$ and return (x,d(x)), i.e. a random value of x and the value of the function d(x) at that x. Each of the researchers independently obtains 1000 training pairs from the oracle, and begins to estimate d(x) from them. They begin with an initial estimate of d(x) as y(x)=0 (where y(x) is the estimate of d(x). They do not update y(x) during this exercise)). They plan to process their training data using stochastic gradient descent. So each of them computes the L2 divergence (as defined in lecture 4) between the estimated output and the desired output for each of their training instances. They do not update y(x) during this exercise.

In order to get a better handle on their problem, the researchers then get together and pool their

divergences over all of their combined training instances, and compute the statistics of the collection of divergences.

Unknown to them (but known to the oracle), d(x)=$\sqrt{x}$
What would you expect the variance of the set of divergences to be?

Note: Recall that a random variable $X \sim Exponential(\lambda)$ if the pdf is $p_X(x) = \lambda exp(-\lambda x)$ for x belongs to [0,infinity). The expectation of the random variable is given by E[X] =1/λ and the variance of the random variable is given by Var(X) = 1/λ^2.

Truncate the answer to 1 decimals (and write the number after the decimal, even if the answer has the form *.0)

Extra Hint: Lec 7, slide 50-80. The L2 divergence is as defined in lec 8, “choices for divergence”, i.e ½ times the squared Euclidean error.

25.0

Question 4
1 / 1 pts

Which of the following statements is true? **[select all that apply]**

- [x] ADAM accounts for variations in both the first and second moments of the derivatives
- [ ] SGD is slow, but will never get stuck in saddle points
- [ ] Nestorov’s method directly accounts for the second moments of the derivatives in different directions
- [x] RMSProp normalizes the step size by the inverse root-mean-squared value of the derivative

Question 5
1 / 1 pts

Compared to batch gradient descent, SGD is often faster (takes less time) because:

- ( ) it is not faster
- ( ) it needs about the same number of iterations, but they are faster to compute
- ( ) its iterations take longer to compute, but it converges in fewer iterations
- (•) we get many more updates in a single pass through the training data

Question 6
1 / 1 pts

The derivative of a loss function for a network with respect to a specific parameter w is upper bounded by $\frac{dL}{dw} \leq 0.5$. You use SGD with the simple gradient update rule to learn the network (no momentum, or any higher order optimization is employed). The initial estimate of w is at a distance of 5.0 from the optimum (i.e. $|w^* - w_0| = 5.0$ where w^* is the optimal value of w and w_0 is its initial value).

Your SGD uses a learning rate schedule where the learning rate at the i-th iteration is η_i. What is the maximum value L for the sum of the sequence of learning rates, in your learning rate schedule (i.e. for $\sum_{i=1}^{\infty} \eta_i$) such that for $\sum_{i=1}^{\infty} \eta_i < L$ you will definitely never arrive at the optimum.

Hint: Lec 7, explanation of SGD, and associated caveats

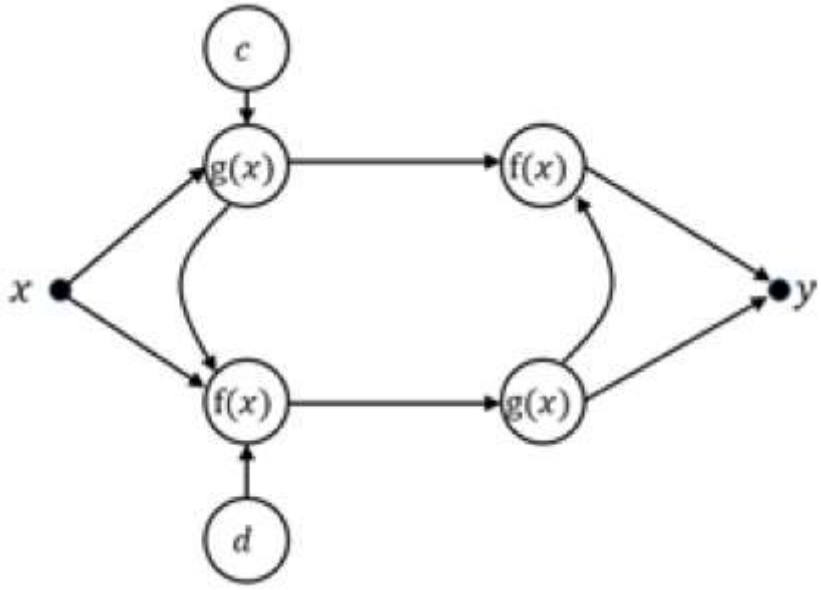
10

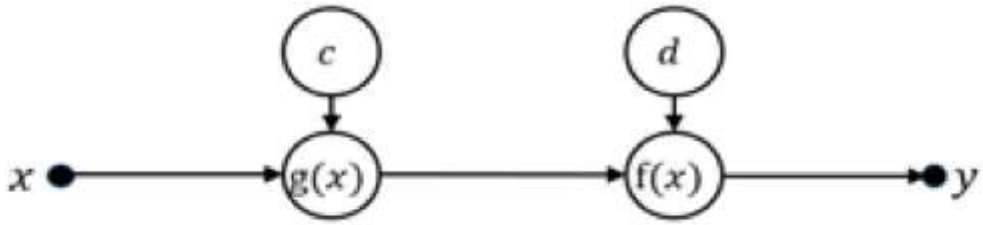
Question 7
1 / 1 pts

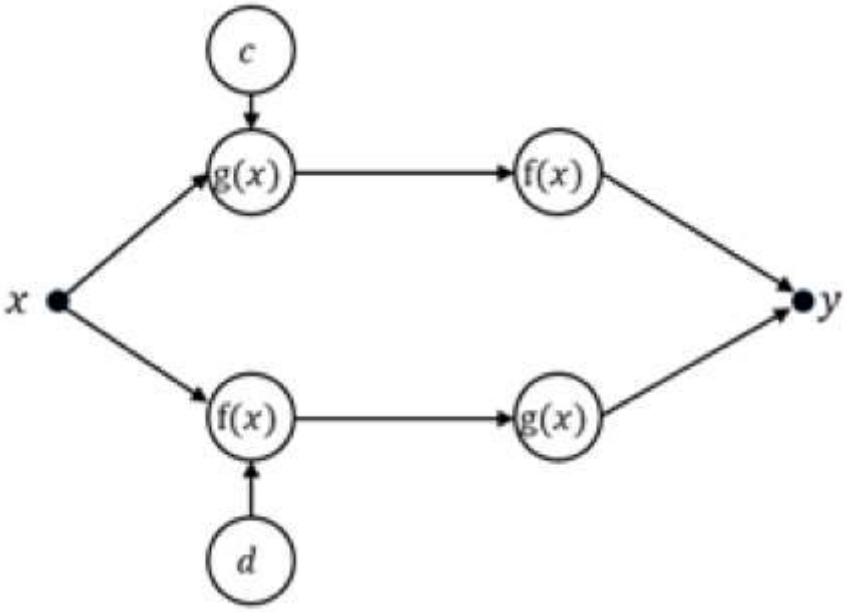
We are given the following relationship $y = f(x, g(x, c))g(x, f(x, d))$. Which of the following figures is the influence diagram for *y* as a function of **x.**

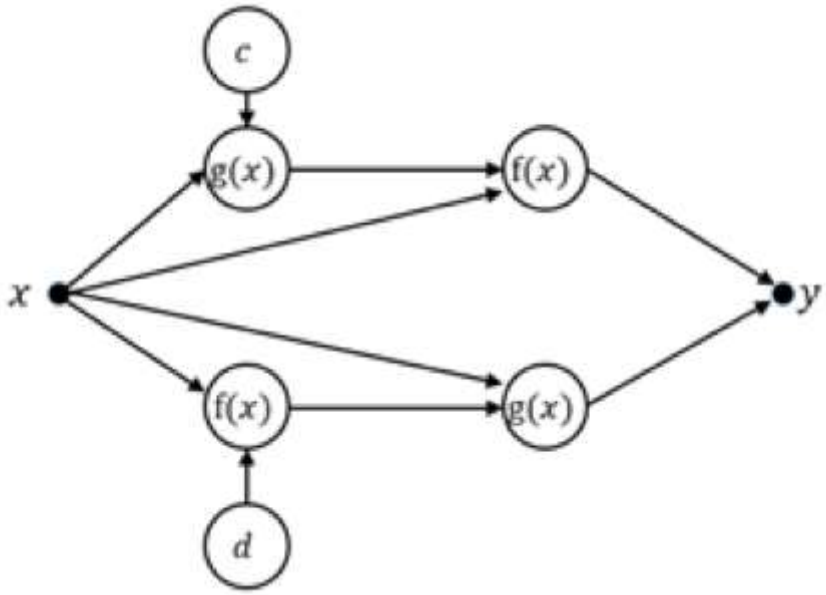
(In the figures below we have generically depicted the dependence between g() and x as g(x) as a shorthand notation. Similarly the dependence between f() and x is shown as f(x) as shorthand notation)

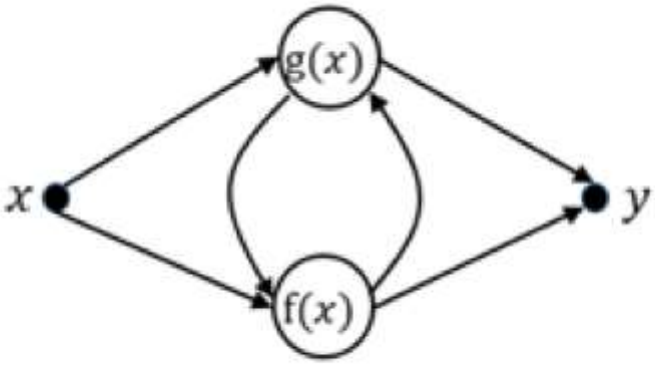
Hint: Lecture 5, Slides 11-23




c
d
x
g(x)
f(x)
y


c
g(x)
f(x)
x
y
f(x)
g(x)
d


c
g(x)
f(x)
x
y
f(x)
g(x)
d



Question 8
1 / 1 pts

We are given the relationship $y = f(x, g(x, c))g(x, f(x, d))$. Here **x** is a scalar. $f(.)$ and $g(.)$ are both scalar functions. Which of the following is the correct formula for the derivative of *y* w.r.t **x**? You may find it useful to draw the influence diagram. **(Select all that apply)**

Hint: Lecture 5, Slides 11-32. Also note the difference between full and partial derivatives

- [x] $dy/dx = g(x, f(x, d)) \times (\partial f/\partial x + \partial f/\partial g \times dg/dx) + f(x, g(x, c)) \times (\partial g/\partial x + \partial g/\partial f \times df/dx)$
- [ ] $dy/dx = \partial y/\partial x \times \partial f/\partial g \times dg/dx$
- [ ] $dy/dx = g(x, f(x, d)) \times (\partial f/\partial x) + f(x, g(x, c)) \times (\partial g/\partial x)$
- [ ] $dy/dx = \partial y/\partial x + \partial y/\partial f \times \partial f/\partial g \times \partial g/\partial x + \partial y/\partial g \times \partial g/\partial f \times \partial f/\partial x$
- [ ] $dy/dx = (\partial f/\partial x + \partial f/\partial g \times dg/dx) + (\partial g/\partial x + \partial g/\partial f \times df/dx)$

Question 9
1 / 1 pts

Dropout is a regularization technique, which theoretically emulates bagging, applied to a network with N (non-output) neurons. When we use Dropout, each (non-output) neuron in the network is selected (turned on) with a probability α. During inference, each of the neuron's activations is scaled by α to account for the dropout applied during training. Which of the following statements is true about this procedure?

Hint: Lec 8, Dropout slides, slide 133

It makes the approximation that

$$E[f(D_1 g_1(x), D_2 g_2(x), \ldots, D_N g_N(x))] \approx f(E[D_1 g_1(x)], E[D_2 g_2(x)], \ldots, E[D_N g_N(x)])$$

where $f()$ represents the network, $g_i(x)$ is the i-th neuron in the network, and D_i is the Bernoulli selector for the neuron, and $E[.]$ is the expectation operator. This is only an approximation.

It utilizes the fact that that

$$E[f(D_1 g_1(x), D_2 g_2(x), \ldots, D_N g_N(x))] = f(E[D_1 g_1(x)], E[D_2 g_2(x)], \ldots, E[D_N g_N(x)])$$

where $f()$ represents the network, $g_i(x)$ is the i-th neuron in the network, and D_i is the Bernoulli selector for the neuron, and $E[.]$ is the expectation operator. This is a provably correct equality.

It computes the output as

$$y = f(E[D_1 g_1(x)], E[D_2 g_2(x)], \ldots, E[D_N g_N(x)])$$

where $f()$ represents the network, $g_i(x)$ is the i-th neuron in the network, and D_i is the Bernoulli selector for the neuron, and $E[.]$ is the expectation operator. This is a theoretically precise computation of the output for bagging.

Question 10
1 / 1 pts

To answer this question, please read (**https://arxiv.org/abs/1502.03167** (**https://arxiv.org/abs/1502.03167**)).

As referred in the paper, in the BN with mini-batch algorithm, the normalized activations belong to a Gaussian distribution (assuming if we sample the elements of each mini-batch from the same distribution)

- (•) True
- ( ) False

Quiz Score: 9 out of 10