Quiz-04

- Due Sep 22 at 11:59pm
- Points 10
- Questions 10
- Available Sep 20 at 6:30pm - Sep 22 at 11:59pm
- Time Limit None
- Allowed Attempts 3

Attempt History

	Attempt	Time	Score
KEPT	Attempt 3	37 minutes	8 out of 10
LATEST	Attempt 3	37 minutes	8 out of 10
	Attempt 2	271 minutes	3.5 out of 10
	Attempt 1	1,269 minutes	3.5 out of 10

ⓘ Correct answers are hidden.

Score for this attempt: 8 out of 10
Submitted Sep 22 at 6:25pm
This attempt took 37 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).

0.394

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 $P(x) = 1, 0 \leq x \leq 1$ 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). Then each of them computes the average L2 divergence (as defined in lecture 4) over their entire batch of training examples.

In order to get a better handle on their problem, they pool their divergences. Since each of them got an independent set of training samples, all their divergences are different. So they 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 their pooled set of divergences to be? Truncate the answer to 2 decimals (and write both digits, even if the answer has the form *.00)

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.

0.25

IncorrectQuestion 3
0 / 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 process their training data in minibatches of size 10. Each of them thus obtains 100 minibatches. 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). Then each of them computes the average L2 divergence (as defined in lecture 4) over each of their minibatches.

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

Unknown to them (but known to the oracle), d(x)=$\sqrt{x}$

What do you expect the variance of this 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 decimal (and write the digit 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.

100.0

Please check Lecture 7

Question 4

1 / 1 pts

What could happen if incremental gradient descent was not stochastic (i.e. if we selected the training points in a constant order) ?

- Overfitting
- The behavior would not change
- The loss value would converge very quickly
- A function that swings around instead of converging

Explanation: See lec 7 slides on order of presentation

See lec 7 slides on order of presentation

Question 5

1 / 1 pts

Methods such as ADAM and RMS Prop improve upon Momentum by

- Incorporating 2nd moments
- Using approximations of the higher order derivatives to estimate local minima
- Clipping the gradients to within 1 standard deviation of the gradient history
- Having fixed learning rates

Explanation: Lec 7 slides. ADAM considers both 1st and 2nd moment, RMS only second-moment. They don’t have fixed learning rates. They also don’t necessarily approximate higher order

derivatives; just moments on the 1st derivative. Neither also clip the gradient.

Lec 7 slides. ADAM considers both 1st and 2nd moment, RMS only second-moment. They don't have fixed learning rates. They also don't necessarily approximate higher order derivatives; just moments on the 1st derivative. Neither also clip the gradient.

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 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

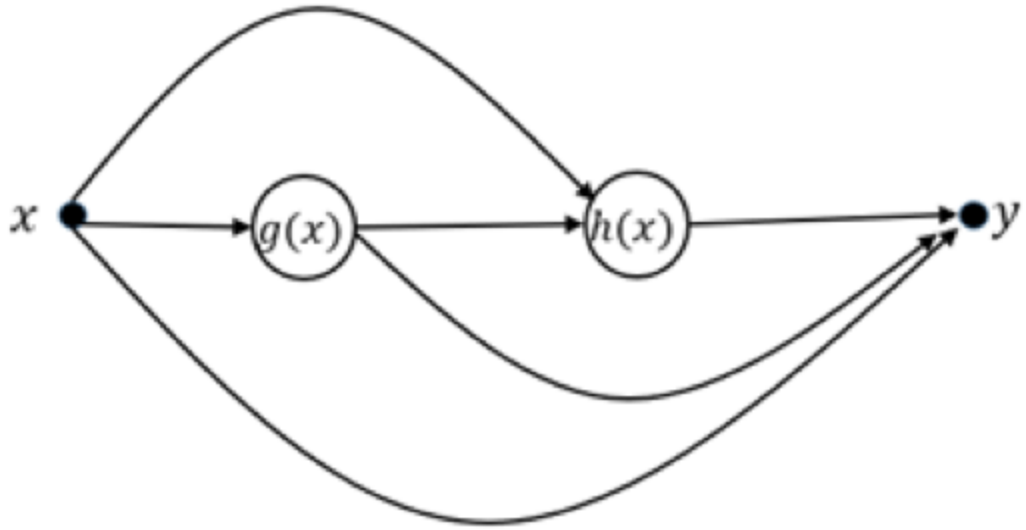
10

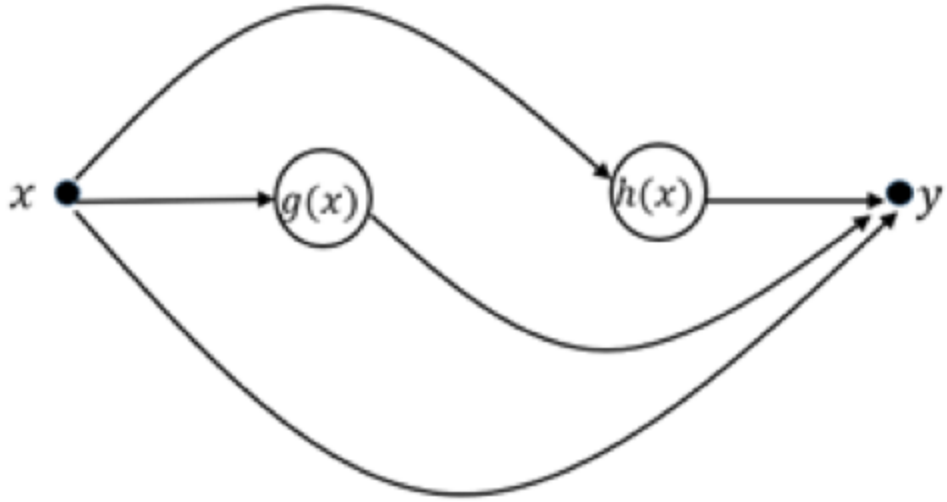
Question 7
1 / 1 pts

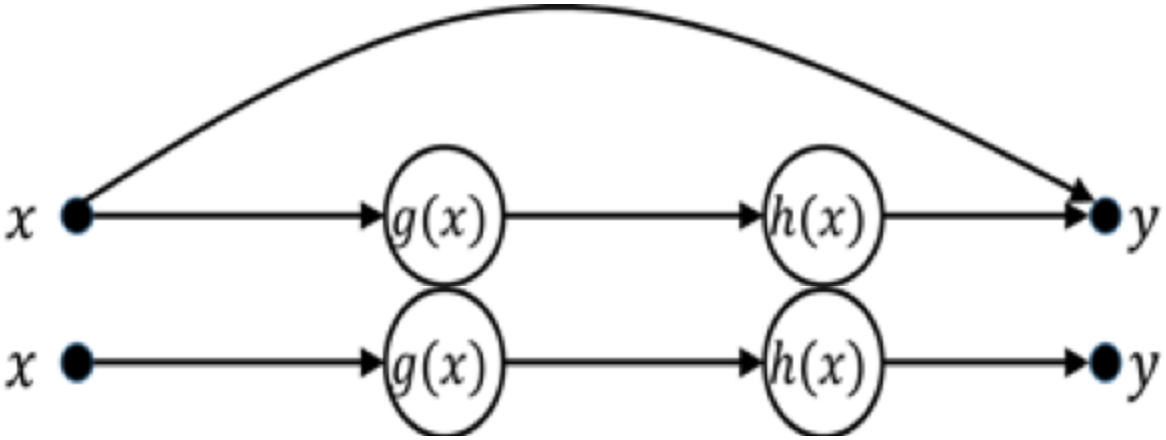
We are given the following relationship $y = f(x, g(x), h(x, g(x)))$. Which of the following figures is the influence diagram for *y* as a function of x.

(In the figures below we have generically depicted the dependence between h() and x as h(x) as a shorthand notation.)

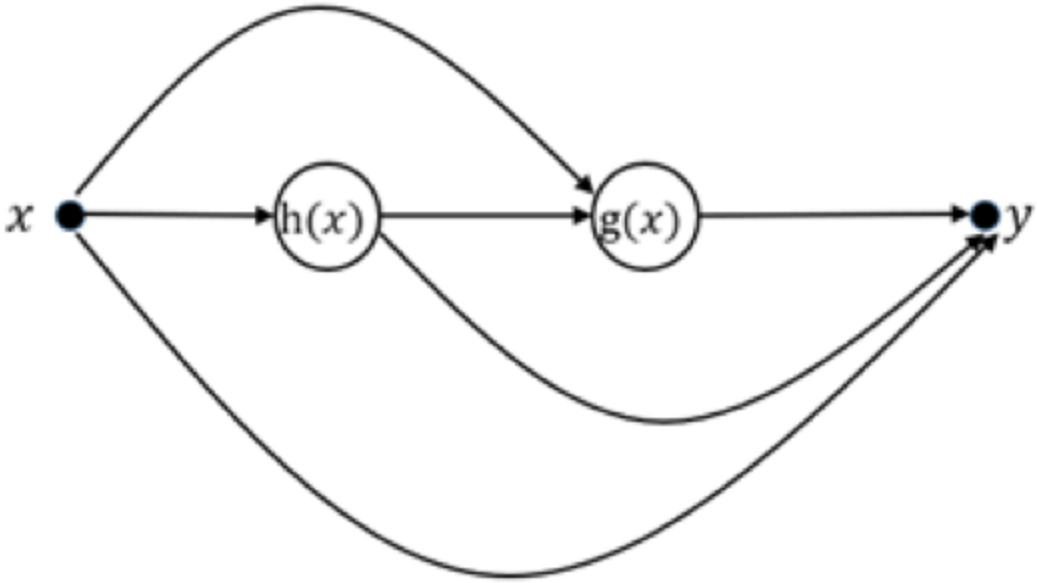
Hint: Lecture 5, Slides 11-23


x
g(x)
h(x)
y


x
g(x)
h(x)
y


x
g(x)
h(x)
y

x
g(x)
h(x)
y



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

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

Question 9
1 / 1 pts

Dropout on a network with N (non-output) neurons is often explained as bagging over all networks that may be composed by randomly selecting each of the non-output neurons with a probability. By this explanation, which of the following statements must be true (for test data)?

Hint: Lec 8, Dropout slides, slide 128

- It theoretically averages the prediction of 2^N separate networks
- It theoretically averages the prediction of N separate networks.
- It randomly selects αN separate networks and averages their predictions
- It theoretically averages the prediction of N^2 separate networks.

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: 8 out of 10