Quiz 2, Fall 23
Section A

NAME, EID: ________________________
For Multiple Choice Questions, circle the best answer.

Q1. (2 points) If we apply a modeling approach to different datasets of the same size obtained i.i.d. from the same underlying probability distribution and then observe that these trained models make accurate and similar predictions on a large test dataset, then your modeling approach has :

a. high bias, low variance
b. low bias, high variance
c. **low bias, low variance**
d. high bias, high variance

Q2. (2 points) Taking larger mini-batches in stochastic gradient descent:

a. reduces the variance, at no additional computational expense
b. reduces the variance, but increases the computational cost of a single iteration
c. reduces the bias, reduces the variance
d. does not change the variance, but improves computational costs

Q3. (1+1 = 2 points) Suppose you fit a regularized non-linear regression model with some training data. What will you expect to happen?
3.a) To model variance, if you increase the amount of regularization

a. Decrease
b. Increase
c. No change
d. Can't say

3.b) To model bias, if you double the size of the training data

a. Increase
b. Decrease
c. **No change**
d. Can't say

Q4. (2 points) What is the key difference between stochastic gradient descent (SGD) and regular gradient descent (GD) in terms of their update mechanisms? Write the weight update equations for both regular GD and SGD.
SGD - weight updates after computing the error and the gradient on one random data point (w_{t+1} = w_t - eta * gradient(Error(x_i))); GD - weight updates after computing the error and the gradient on all of the data (w_{t+1} = w_t - eta * gradient(average error)).

Q5. (2 points) Mention any two advantages SGD has over regular GD.
SGD is computationally faster; can avoid local minimas; low memory footprint.

Quiz 2, Fall 23
Section B
NAME, EID: ______________________
For Multiple Choice Questions, circle the best answer.

Q1) (2 points) Suppose you fit a regularized non-linear regression model with some training data.
What will you expect to happen to model bias, if you add more learnable parameters in the model?

a. **Decrease or no change**
b. Increase
c. No change
d. Always decrease

Q2. (2 points) Which of the following statements is not always true regarding Multi-layer Perceptron (MLP)?

a. **The more hidden-layer units an MLP has, the better its generalization error after a given number of epochs.**
b. MLPs that have one hidden layer with hyperbolic tangent activation functions and an output layer employing linear activation functions are universal approximators.
c. **The backpropagation learning algorithm performs stochastic gradient-descent to learn weights in MLP.**
d. An MLP in which all activation functions are linear cannot be a universal approximator.

Q3. (2 points) Which of the following is false?

a. **Gradient descent optimization always attains the global minimum of a function if learning rate is chosen appropriately.**
b. SGD optimization sometimes leads to local minima
c. SGD gradient updates are noisy
d. SGD is generally faster than Gradient descent

Q4. (2 points) Briefly explain why “backpropagation of error” is considered an efficient algorithm for MLPs.
Backpropagation uses chain rule from calculus to compute the gradients of the error with respect to each parameter layer by layer. At every layer, the upstream gradient is propagated and just multiplied by the local gradient which makes it an efficient way to compute the gradients for networks containing many hidden layers.

Q5. (2 points) What makes SGD “stochastic”? What happens to the amount of stochasticity if the size of “mini-batches” used for training is increased?
The random selection of a datapoint for the gradient update introduces stochasticity in SGD. The stochasticity decreases as the size of mini-batches is increased.