Name: _____ Roll Number: _____

1. (a) (1 point) Gradient vector $\nabla_{\mathbf{w}} L(\mathbf{x}; \mathbf{w})$ lies in ...
   A. input domain $\mathbf{x}$.
   B. parameter space $\mathbf{w}$.
   C. output range of L.
   D. a direction orthogonal to L.

   (b) (1 point) Gradient vector points in the direction of fastest increase of a function.

   (c) (1 point) Describe the gradient descent method for minimizing a function.

Update weights by moving the negative direction of the gradient vector as $\mathbf{w}^{\text{new}} = \mathbf{w}^{\text{old}} - \eta \nabla_{\mathbf{w}} E$.

   (d) (1 point) What is the role of the learning rate η in gradient descent?

By decaying η, gradient descent converges to a local minimum.

   (e) Briefly describe the following.
   i. (1 point) Batch Gradient Descent.

Gradient is accumulated over the whole training set before updating the weights.

   ii. (1 point) Stochastic Gradient Descent.

Weights are updated using the gradient of each training example chosen randomly.

   iii. (1 point) Stochastic Gradient Descent using Mini-Batches.

Gradient is accumulated over a randomly selected mini-batch of training examples before updating the weights.

   (f) (2 points) How does stochastic gradient descent hep in avoiding local minima?

The negative of the true gradient vector accumulated over the whole training batch points towards the closest local minimum but an incomplete gradient vector from a single, stochastically selected training example does not necessarily point in that same direction. So negatives of stochastic gradients can point away from the closest local minima.

   (g) (1 point) Near a minimum, if the step size is large, the gradient vector will start oscillating. How does the idea of momentum help in such a situation?

Momentum refers to the running average of update vectors. Near a minimum, alternating update directions will cancel each other and not build momentum. This will make the update vector smaller and automatically prevent oscillations.