1. Suppose you are training a logistic regression classifier using stochastic gradient descent. You find that the cost (say, $cost(\theta, (x^{(i)}, y^{(i)}))$, averaged over the last 500 examples), plotted as a function of the number of iterations, is slowly increasing over time. Which of the following changes are likely to help?

   1 point

   ⦿ Try halving (decreasing) the learning rate α, and see if that causes the cost to now consistently go down; and if not, keep halving it until it does.

   ◯ Try averaging the cost over a smaller number of examples (say 250 examples instead of 500) in the plot.

   ◯ Use fewer examples from your training set.

   ◯ This is not possible with stochastic gradient descent, as it is guaranteed to converge to the optimal parameters θ.

2. Which of the following statements about stochastic gradient descent are true? Check all that apply.

   1 point

   ☑ You can use the method of numerical gradient checking to verify that your stochastic gradient descent implementation is bug-free. (One step of stochastic gradient descent computes the partial derivative $\frac{\partial}{\partial \theta_j} cost(\theta, (x^{(i)}, y^{(i)}))$.)

   ☑ Before running stochastic gradient descent, you should randomly shuffle (reorder) the training set.

   ☐ Suppose you are using stochastic gradient descent to train a linear regression classifier. The cost function $J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$ is guaranteed to decrease after every iteration of the stochastic gradient descent algorithm.

   ☐ In order to make sure stochastic gradient descent is converging, we typically compute $J_{\text{train}}(\theta)$ after each iteration (and plot it) in order to make sure that the cost function is generally decreasing.

3. Which of the following statements about online learning are true? Check all that apply.

   1 point

   ☐ One of the advantages of online learning is that there is no need to pick a learning rate α.

   ☑ When using online learning, in each step we get a new example (x, y), perform one step of (essentially stochastic gradient descent) learning on that example, and then discard that example and move on to the next.

   ☑ In the approach to online learning discussed in the lecture video, we repeatedly get a single training example, take one step of stochastic gradient descent using that example, and then move on to the next example.

   ☐ One of the disadvantages of online learning is that it requires a large amount of computer memory/disk space to store all the training examples we have seen.

4. Assuming that you have a very large training set, which of the following algorithms do you think can be parallelized using map-reduce and splitting the training set across different machines? Check all that apply.

   1 point

   ☐ An online learning setting, where you repeatedly get a single example (x, y), and want to learn from that single example before moving on.

   ☐ Logistic regression trained using stochastic gradient descent.

   ☑ Linear regression trained using batch gradient descent.

   ☑ A neural network trained using batch gradient descent.

5. Which of the following statements about map-reduce are true? Check all that apply.

   1 point

   ☐ Running map-reduce over N computers requires that we split the training set into N^2 pieces.

   ☑ When using map-reduce with gradient descent, we usually use a single machine that accumulates the gradients from each of the map-reduce machines, in order to compute the parameter update for that iteration.

   ☑ In order to parallelize a learning algorithm using map-reduce, the first step is to figure out how to express the main work done by the algorithm as computing sums of functions of training examples.

   ☑ If you have just 1 computer, but your computer has multiple CPUs or multiple cores, then map-reduce might be a viable way to parallelize your learning algorithm.