

Feedback — XVII. Large Scale Machine Learning

[Help Center](#)

You submitted this quiz on **Sun 12 Apr 2015 11:00 PM CEST**. You got a score of **5.00** out of **5.00**.

Question 1

Suppose you are training a logistic regression classifier using stochastic gradient descent. You find that the cost (say, $\text{cost}(\theta, (x^{(i)}, y^{(i)}))$, 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?

Your Answer	Score	Explanation
<input type="radio"/> Try averaging the cost over a smaller number of examples (say 250 examples instead of 500) in the plot.		
<input type="radio"/> Try averaging the cost over a larger number of examples (say 1000 examples instead of 500) in the plot.		
<input checked="" type="radio"/> 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.	✓ 1.00	Such a plot indicates that the algorithm is diverging. Decreasing the learning rate α means that each iteration of stochastic gradient descent will take a smaller step, thus it will likely converge instead of diverging.
<input type="radio"/> Use fewer examples from your training set.		
Total	1.00 / 1.00	

Question 2

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





Your Answer	Score	Explanation
<input checked="" type="checkbox"/> If you have a huge training set, then stochastic gradient descent may be much faster than batch gradient descent.	✓ 0.25	Because stochastic gradient descent can make progress after only a few examples, it can converge much more quickly than batch gradient descent.
<input type="checkbox"/> One of the advantages of stochastic gradient descent is that it uses parallelization and thus runs much faster than batch gradient descent.	✓ 0.25	Stochastic gradient descent still runs in series, one example at a time.
<input checked="" type="checkbox"/> One of the advantages of stochastic gradient descent is that it can start progress in improving the parameters θ after looking at just a single training example; in contrast, batch gradient descent needs to take a pass over the entire training set before it starts to make progress in improving the parameters' values.	✓ 0.25	This is true, since stochastic gradient descent updates the parameters for every training example, but batch gradient descent updates them based on an average over the entire training set.
<input type="checkbox"/> 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	✓ 0.25	We want to plot $\text{cost}(\theta, (x^{(i)}, y^{(i)}))$ at each iteration, as computing the full summation $J_{\text{train}}(\theta)$ is too expensive.

generally decreasing.

Total	1.00 /
	1.00

Question 3

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

Your Answer	Score	Explanation
<input checked="" type="checkbox"/> Online learning algorithms are usually best suited to problems where we have a continuous/non-stop stream of data that we want to learn from.	 0.25	Such a stream of data is well-suited to online learning because online learning does not save old training examples, but instead uses them once and then throws them out.
<input type="checkbox"/> One of the advantages of online learning is that there is no need to pick a learning rate α .	 0.25	One still must choose a learning rate to use online learning.
<input type="checkbox"/> 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.	 0.25	Since online learning algorithms do not save old examples, they can be very efficient in terms of computer memory and disk space.
<input checked="" type="checkbox"/> 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.	 0.25	This is one good approach to online learning discussed in the lecture video.
Total	1.00 /	
	1.00	

Question 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.

Your Answer	Score	Explanation
<input type="checkbox"/> Linear regression trained using stochastic gradient descent.	✓ 0.25	Since stochastic gradient descent processes one example at a time and updates the parameter values after each, it cannot be easily parallelized.
<input checked="" type="checkbox"/> Computing the average of all the features in your training set $\mu = \frac{1}{m} \sum_{i=1}^m x^{(i)}$ (say in order to perform mean normalization).	✓ 0.25	You can split the dataset into N smaller batches, compute the feature average of each smaller batch on one of N separate computers, and then average those results on a central computer to get the final result.
<input type="checkbox"/> A neural network trained using stochastic gradient descent.	✓ 0.25	Since stochastic gradient descent processes one example at a time and updates the parameter values after each, it cannot be easily parallelized.
<input checked="" type="checkbox"/> A neural network trained using batch gradient descent.	✓ 0.25	You can split the dataset into N smaller batches, compute the gradient for each smaller batch on one of N separate computers, and then average those gradients on a central computer to use for the gradient update.
Total	1.00 / 1.00	

Question 5

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

Your Answer	Score	Explanation
<input checked="" type="checkbox"/> Because of network latency and other overhead	✓ 0.25	The maximum speedup possible is N -fold, and it is unlikely you will get an N -fold speedup because of

associated with map-reduce, if we run map-reduce using N computers, we might get less than an N -fold speedup compared to using 1 computer.

the overhead.

☒ 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.

✓ 0.25

Such a setup allows us to use many computers to do the hard work of gradient computation while making the parameter update simple, as it occurs in one place.

☒ 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.

✓ 0.25

Treating each core as a separate computer makes map-reduce just as useful with multiple cores as with multiple computers.

☐ If we run map-reduce using N computers, then we will always get at least an N -fold speedup compared to using 1 computer.

✓ 0.25

The maximum speedup possible is N -fold, and it is unlikely you will get an N -fold speedup because of the overhead.

Total

1.00 /
1.00