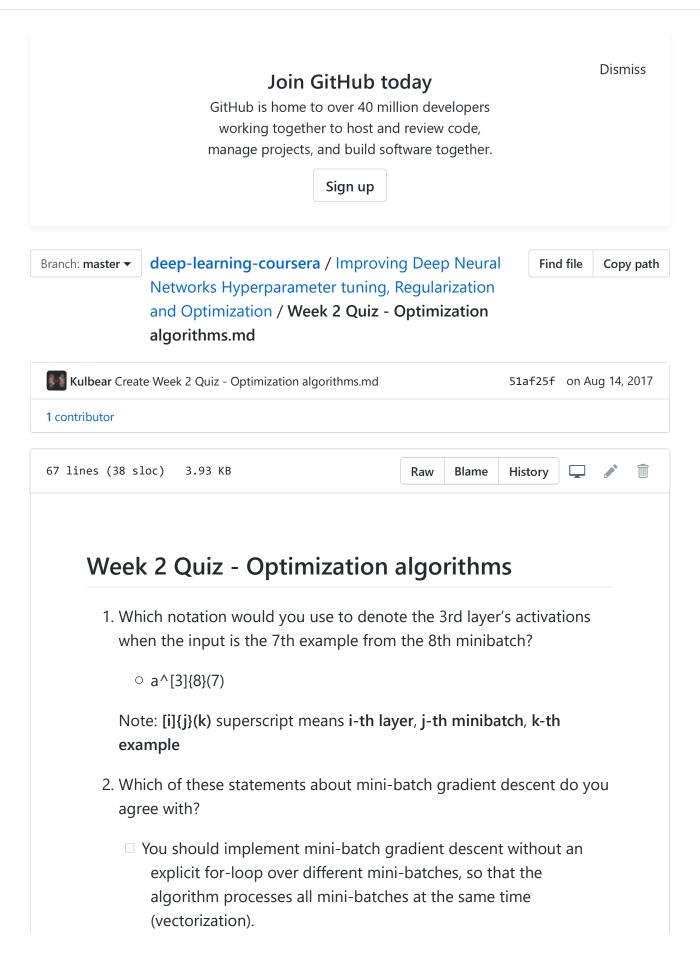
## This repository has been archived by the owner. It is now read-only.



1 of 4 12/27/2019, 10:59 AM

- ☐ Training one epoch (one pass through the training set) using minibatch gradient descent is faster than training one epoch using batch gradient descent.
- One iteration of mini-batch gradient descent (computing on a single mini-batch) is faster than one iteration of batch gradient descent.

Note: Vectorization is not for computing several mini-batches in the same time.

- 3. Why is the best mini-batch size usually not 1 and not m, but instead something in-between?
  - If the mini-batch size is 1, you lose the benefits of vectorization across examples in the mini-batch.
  - If the mini-batch size is m, you end up with batch gradient descent, which has to process the whole training set before making progress.
- 4. Suppose your learning algorithm's cost *J*, plotted as a function of the number of iterations, looks like this:
  - If you're using mini-batch gradient descent, this looks acceptable. But if you're using batch gradient descent, something is wrong.

Note: There will be some oscillations when you're using mini-batch gradient descent since there could be some noisy data example in batches. However batch gradient descent always guarantees a lower J before reaching the optimal.

5. Suppose the temperature in Casablanca over the first three days of January are the same:

Jan 1st: 
$$\theta_{1} = 10$$

Say you use an exponentially weighted average with  $\beta = 0.5$  to track the temperature:  $v_0 = 0$ ,  $v_t = \beta v_t - 1 + (1 - \beta)\theta_t$ . If  $v_2$  is the value computed after day 2 without bias correction, and v^corrected\_2 is the value you compute with bias correction. What are these values?

$$\circ$$
 v\_2 = 7.5, v^corrected\_2 = 10

2 of 4 12/27/2019, 10:59 AM 6. Which of these is NOT a good learning rate decay scheme? Here, t is the epoch number.

$$\circ \alpha = e^t * \alpha_0$$

Note: This will explode the learning rate rather than decay it.

- 7. You use an exponentially weighted average on the London temperature dataset. You use the following to track the temperature:  $v_t = \beta v_t - 1 + (1 - \beta)\theta_t$ . The red line below was computed using  $\beta = 0$ 0.9. What would happen to your red curve as you vary  $\beta$ ? (Check the two that apply)
  - $\circ$  Increasing  $\beta$  will shift the red line slightly to the right.
  - $\circ$  Decreasing  $\beta$  will create more oscillation within the red line.
- 8. Consider this figure:

These plots were generated with gradient descent; with gradient descent with momentum ( $\beta = 0.5$ ) and gradient descent with momentum ( $\beta$  = 0.9). Which curve corresponds to which algorithm?

- (1) is gradient descent. (2) is gradient descent with momentum (small  $\beta$ ). (3) is gradient descent with momentum (large  $\beta$ )
- 9. Suppose batch gradient descent in a deep network is taking excessively long to find a value of the parameters that achieves a small value for the cost function J(W[1],b[1],...,W[L],b[L]). Which of the following techniques could help find parameter values that attain a small value forJ? (Check all that apply)
  - ☑ Try using Adam
  - ☑ Try better random initialization for the weights
  - $\square$  Try tuning the learning rate  $\alpha$
  - Try mini-batch gradient descent
  - Try initializing all the weights to zero
- 10. Which of the following statements about Adam is False?
  - Adam should be used with batch gradient computations, not with mini-batches.

Note: Adam could be used with both.

3 of 4

deep-learning-coursera/Week 2 Quiz - Optimization algorithms.md at ma	https://github.com/Kulbear/deep-learning-coursera/blob/master/Improvi

4 of 4