← Back **Optimization Algorithms**
Graded Quiz • 50 min

Due Aug 14, 11:59 PM +03

Congratulations! You passed!

Grade received 90% **Latest Submission Grade** 90% **To pass** 80% or higher

Go to next item

1. Which notation would you use to denote the 3rd layer's activations when the input is the 7th example from the 8th minibatch? **1 / 1 point**

- ○ $a^{[3]\{7\}(8)}$
- ○ $a^{[8]\{3\}(7)}$
- ◉ $a^{[3]\{8\}(7)}$
- ○ $a^{[8]\{7\}(3)}$

Expand

Correct

2. Which of these statements about mini-batch gradient descent do you agree with? **1 / 1 point**

- ◉ One iteration of mini-batch gradient descent (computing on a single mini-batch) is faster than one iteration of batch gradient descent.
- ○ You should implement mini-batch gradient descent without an explicit for-loop over different mini-batches, so that the algorithm processes all mini-batches at the same time (vectorization).
- ○ Training one epoch (one pass through the training set) using mini-batch gradient descent is faster than training one epoch using batch gradient descent.

Expand

Correct

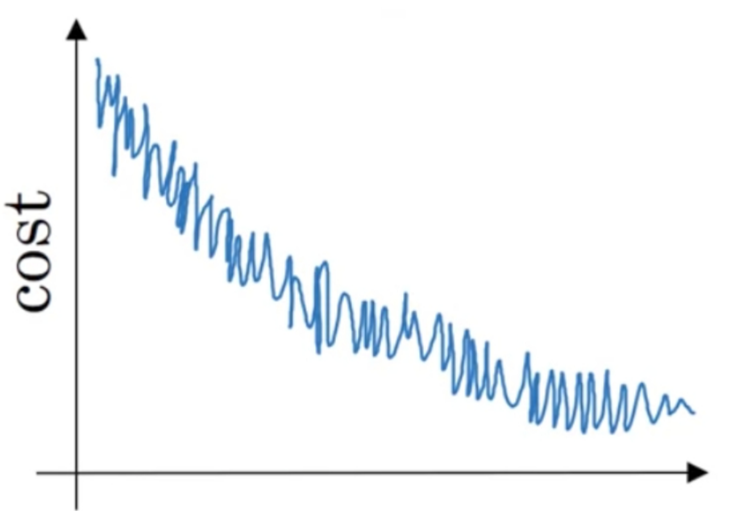
3. Which of the following is true about batch gradient descent? **1 / 1 point**

- ○ It is the same as stochastic gradient descent, but we don't use random elements.
- ◉ It is the same as the mini-batch gradient descent when the mini-batch size is the same as the size of the training set.
- ○ It has as many mini-batches as examples in the training set.

Expand

Correct
Correct. When using batch gradient descent there is only one mini-batch thus it is equivalent to batch gradient descent.

4. While using mini-batch gradient descent with a batch size larger than 1 but less than m, the plot of the cost function J looks like this: **1 / 1 point**



You notice that the value of J is not always decreasing. Which of the following is the most likely reason for that?

- ○ You are not implementing the moving averages correctly. Using moving averages will smooth the graph.
- ◉ In mini-batch gradient descent we calculate $J(\hat{y}^{\{t\}}, y^{\{t\}})$ thus with each batch we compute over a new set of data.
- ○ The algorithm is on a local minimum thus the noisy behavior.
- ○ A bad implementation of the backpropagation process, we should use gradient check to debug our implementation.

Loading [MathJax]/jax/output/CommonHTML/jax.js

Expand

Correct
Yes. Since at each iteration we work with a different set of data or batch the loss function doesn't have to be decreasing at each iteration.

5. Suppose the temperature in Casablanca over the first two days of January are the same: **1 / 1 point**

Jan 1st: $\theta_1 = 10^oC$

Jan 2nd: $\theta_2 = 10^oC$

(We used Fahrenheit in the lecture, so we will use Celsius here in honor of the metric world.)

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 $v_2^{corrected}$ is the value you compute with bias correction. What are these values? (You might be able to do this without a calculator, but you don't actually need one. Remember what bias correction is doing.)

- ○ $v_2 = 10$, $v_2^{corrected} = 7.5$
- ○ $v_2 = 10$, $v_2^{corrected} = 10$
- ○ $v_2 = 7.5$

,

$v^{corrected} = 7.5$

Expand

Correct

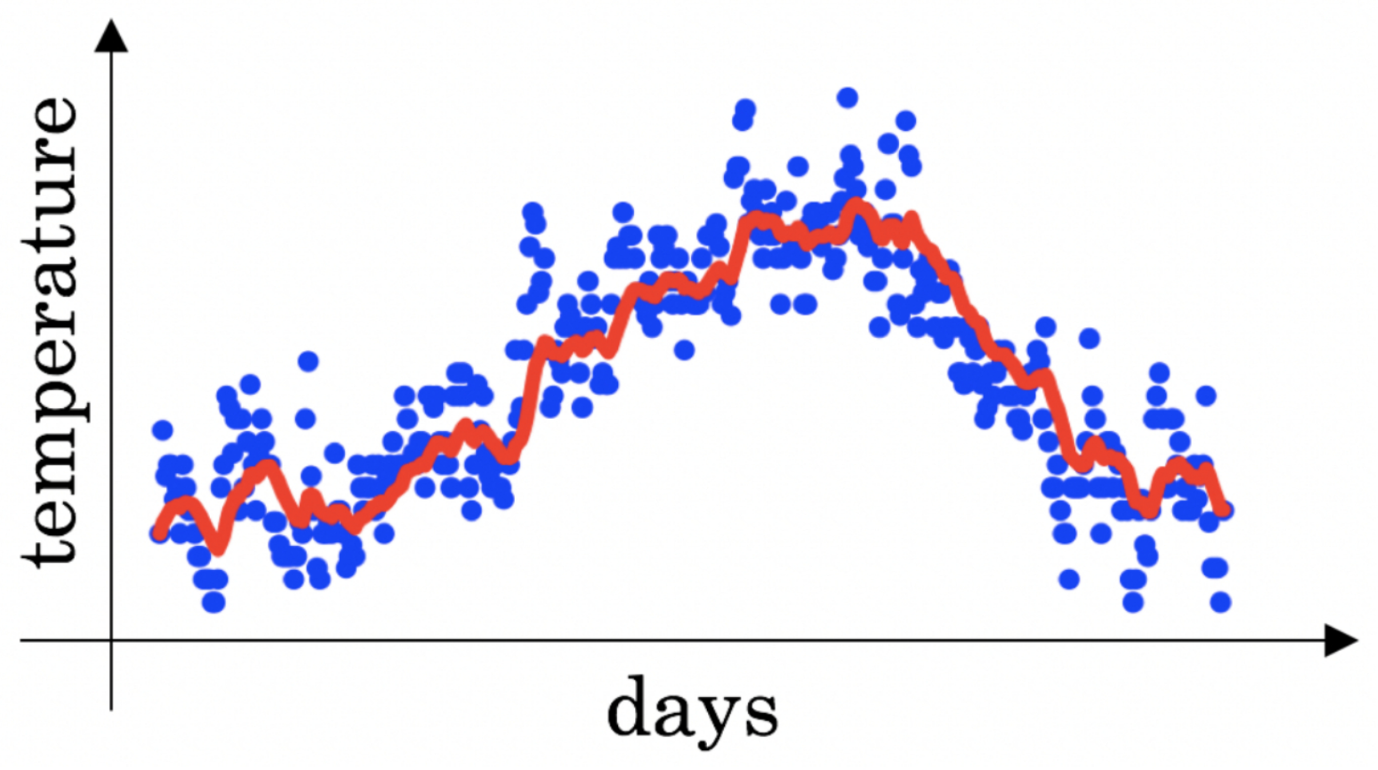
6. Which of the following is true about learning rate decay? **1 / 1 point**

- ◉ The intuition behind it is that for later epochs our parameters are closer to a minimum thus it is more convenient to take smaller steps to prevent large oscillations.
- ○ We use it to increase the size of the steps taken in each mini-batch iteration.
- ○ It helps to reduce the variance of a model.
- ○ The intuition behind it is that for later epochs our parameters are closer to a minimum thus it is more convenient to take larger steps to accelerate the convergence.

Expand

Correct
Correct. Reducing the learning rate with time reduces the oscillation around a minimum.

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 $\beta = 0.9$. What would happen to your red curve as you vary β? (Check the two that apply) **1 / 1 point**



- ☐ Decreasing β will shift the red line slightly to the right.
- ☑ Increasing β will shift the red line slightly to the right.

  ✓ **Correct**
  True, remember that the red line corresponds to $\beta = 0.9$. In the lecture we had a green line $\beta = 0.98$ that is slightly shifted to the right.

- ☑ Decreasing $$\beta$$ will create more oscillation within the red line.

  ✓ **Correct**
  True, remember that the red line corresponds to $$\beta = 0.9$$. In lecture we had a yellow line $$\beta = 0.98$$ that had a

  $\beta = 0.9$

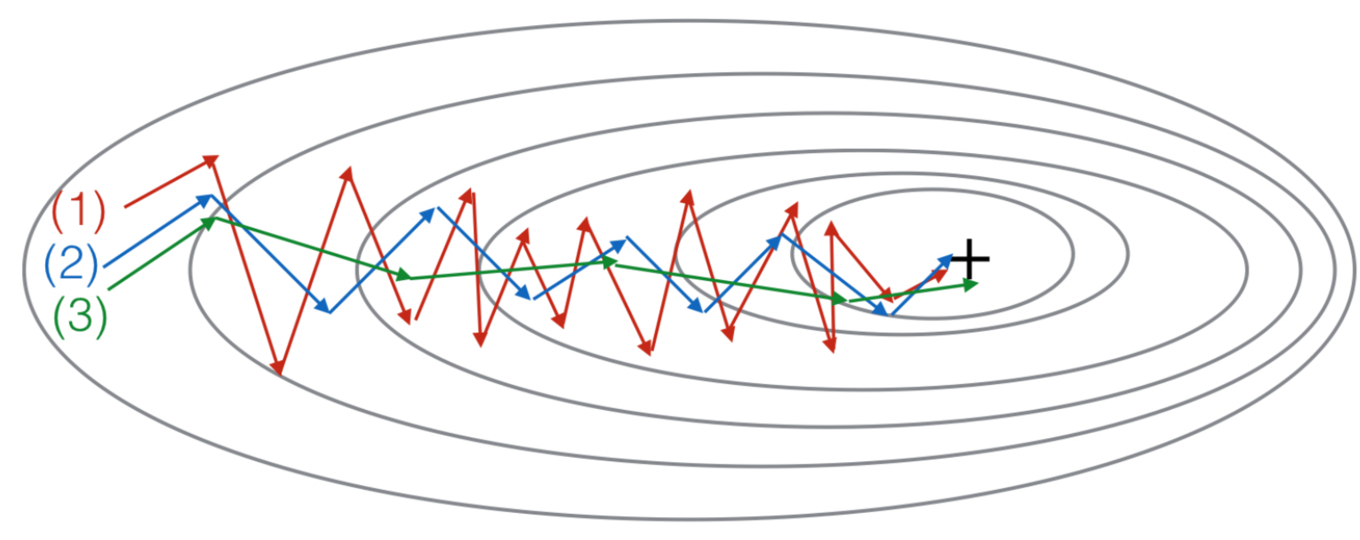
  . In lecture we had a yellow line

  $\beta = 0.98$

Expand

Correct
Great, you got all the right answers.

8. Consider this figure: **1 / 1 point**



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

- ○ (1) is gradient descent with momentum (small β). (2) is gradient descent. (3) is gradient descent with momentum (large β)
- ○ (1) is gradient descent with momentum (small β), (2) is gradient descent with momentum (small β), (3) is gradient descent
- ◉ (1) is gradient descent. (2) is gradient descent with momentum (small β). (3) is gradient descent with momentum (large β)
- ○ (1) is gradient descent. (2) is gradient descent with momentum (large $$\beta$$). (3) is gradient descent with momentum (small $$\beta$$)

Expand

Correct

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 $\mathcal{J}(W^{[1]}, b^{[1]}, ..., W^{[L]}, b^{[L]})$. Which of the following techniques could help find parameter values that attain a small value for $\mathcal{J}$? (Check all that apply) **0 / 1 point**

- ☑ Try using gradient descent with momentum.

  ✓ **Correct**
  Yes. The use of momentum can improve the speed of the training. Although other methods might give better results, such as Adam.

- ☐ Try better random initialization for the weights
- ☐ Normalize the input data.
- ☐ Add more data to the training set.

Expand

Incorrect
You didn't select all the correct answers

10. Which of the following are true about Adam? **1 / 1 point**

- ○ Adam can only be used with batch gradient descent and not with mini-batch gradient descent.
- ○ The most important hyperparameter on Adam is ϵ and should be carefully tuned.
- ○ Adam automatically tunes the hyperparameter α.
- ◉ Adam combines the advantages of RMSProp and momentum.

Expand

Correct
True. Precisely Adam combines the features of RMSProp and momentum that is why we use two-parameter β_1 and β_2, besides ϵ.