Back Optimization Algorithms
Graded Quiz • 50 min

English Due Jul 7, 11:59 PM IST

Congratulations! You passed!

Grade received 100% **Latest Submission Grade** 100% **To pass** 80% or higher

Retake the assignment in **7h 54m**

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

○ $a^{[8]\{7\}(3)}$

● $a^{[3]\{8\}(7)}$

○ $a^{[3]\{7\}(8)}$

○ $a^{[8]\{3\}(7)}$

Expand

✓ Correct

2. Suppose you don't face any memory-related problems. Which of the following make more use of vectorization.

1 / 1 point

○ Stochastic Gradient Descent, Batch Gradient Descent, and Mini-Batch Gradient Descent all make equal use of vectorization.

○ Stochastic Gradient Descent

● Batch Gradient Descent

○ Mini-Batch Gradient Descent with mini-batch size $m/2$.

Expand

✓ Correct
Yes. If no memory problem is faced, batch gradient descent processes all of the training set in one pass, maximizing the use of vectorization.

3. Which of the following is true about batch gradient descent?

1 / 1 point

○ It has as many mini-batches as examples in the training set.

○ It is the same as stochastic gradient descent, but we don't use random elements.

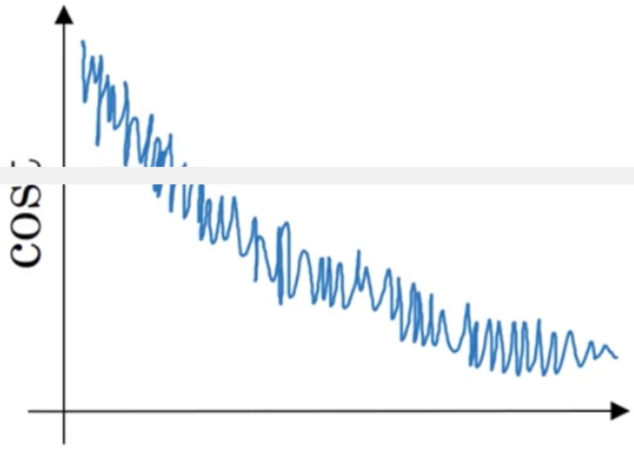
● It is the same as the mini-batch gradient descent when the mini-batch size is the same as the size of the training set.

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?

○ 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.

○ 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.

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.)

● $v_2 = 7.5$, $v_2^{corrected} = 10$

○ $v_2 = 10$, $v_2^{corrected} = 7.5$

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

○ $v_2 = 10$

$v_2 = 10$

Expand

Correct

6. Which of these is NOT a good learning rate decay scheme? Here, t is the epoch number.

1 / 1 point

○ $\alpha = \frac{\alpha_0}{1+3t}$

● $\alpha = 1.01^t \alpha_0$

○ $\alpha = \frac{\alpha_0}{\sqrt{1+t}}.$

○ $\alpha = e^{-0.01\,t}\alpha_0.$

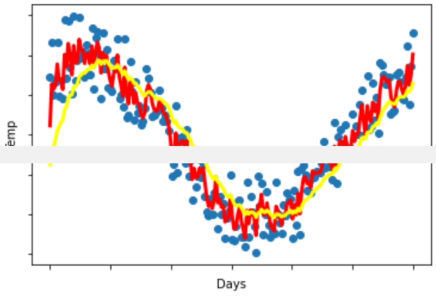
Expand

Correct

Correct. This is not a good learning rate decay since it is an increasing function of t.

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 yellow and red lines were computed using values β_1 and β_2 respectively. Which of the following are true?

1 / 1 point



○ $\beta_1 < \beta_2.$

● $\beta_1 > \beta_2.$

○ $\beta_1 = 0, \beta_2 > 0.$

○ $\beta_1 = \beta_2.$

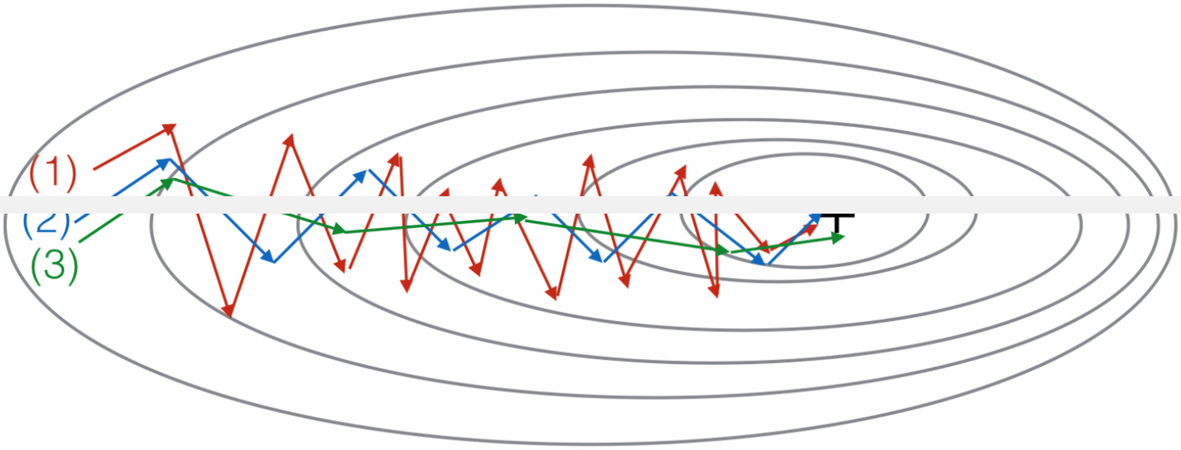
Expand

Correct

Correct. $\beta_1 > \beta_2$ since the red curve is noisier.

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. (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 β). (3) is gradient descent with momentum (small β)

○ (1) is gradient descent with momentum (small β), (2) is gradient descent with momentum (small β), (3) is gradient descent

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

1 / 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.

☑ Normalize the input data.

✓ **Correct**

Yes. In some cases, if the scale of the features is very different, normalizing the input data will speed up the training process.

☑ Try better random initialization for the weights

✓ **Correct**

Yes. As seen in previous lectures this can help the gradient descent process to prevent vanishing gradients.

□ Add more data to the training set.

Expand

Correct

Great, you got all the right answers.

10. Which of the following statements about Adam is ***False***?

1 / 1 point

○ The learning rate hyperparameter α in Adam usually needs to be tuned.

● Adam should be used with batch gradient computations, not with mini-batches.

○ We usually use "default" values for the hyperparameters β_1, β_2 and ε in Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$)

○ Adam combines the advantages of RMSProp and momentum

Expand

Correct