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

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

Expand

Correct

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

1 / 1 point

- ( ) Mini-Batch Gradient Descent with mini-batch size $m/2$.
- (•) Batch Gradient Descent
- ( ) Stochastic Gradient Descent, Batch Gradient Descent, and Mini-Batch Gradient Descent all make equal use of vectorization.
- ( ) Stochastic Gradient Descent

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. Why is the best mini-batch size usually not 1 and not m, but instead something in-between? Check all that are true.

1 / 1 point

- [x] If the mini-batch size is 1, you lose the benefits of vectorization across examples in the mini-batch.

  Correct

- [ ] If the mini-batch size is 1, you end up having to process the entire training set before making any progress.
- [x] 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.

  Correct

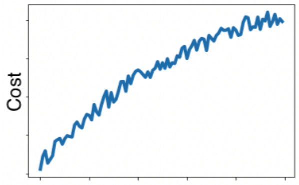
- [ ] If the mini-batch size is m, you end up with stochastic gradient descent, which is usually slower than mini-batch gradient descent.

Expand

Correct
Great, you got all the right answers.

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:

0 / 1 point



Which of the following do you agree with?

- ( ) No matter if using mini-batch gradient descent or batch gradient descent something is wrong.
- ( ) If you are using batch gradient descent, this looks acceptable. But if you're using mini-

batch gradient descent, something is wrong.

- ( ) If you are using mini-batch gradient descent or batch gradient descent this looks acceptable.
- (•) If you are using mini-batch gradient descent, this looks acceptable. But if you're using batch gradient descent, something is wrong.

Expand

Incorrect
No. The cost is larger than when the process started, this is not right at all.

5. Suppose the temperature in Casablanca over the first two days of March are the following:

1 / 1 point

March 1st: $\theta_1 = 10^\circ\ \mathrm{C}$

March 2nd: $\theta_2 = 25^\circ\ \mathrm{C}$

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 $v_2^{\text{corrected}}$ is the value you compute with bias correction. What are these values?

- (•) $v_2 = 15$, $v_2^{\text{corrected}} = 20$.
- ( ) $v_2 = 20$, $v_2^{\text{corrected}} = 20$

Expand

Correct
Correct. $v_2 = \beta v_{t-1} + (1-\beta)\,\theta_t$ thus $v_1 = 5, v_2 = 15$. Using the bias correction $\frac{v_t}{1-\beta^t}$ we get $\frac{15}{1-(0.5)^2} = 20$.

6. Which of the following is true about learning rate decay?

1 / 1 point

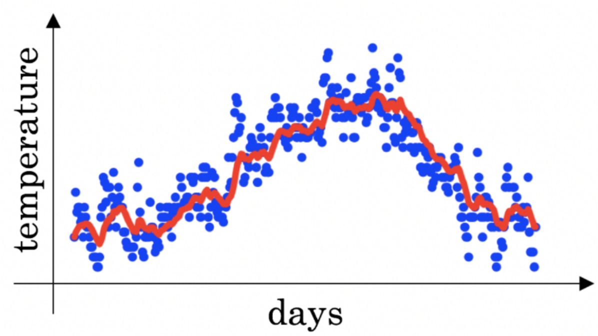
- ( ) It helps to reduce the variance of a model.
- ( ) The intuition behind it is that for later epochs our parameters are closer to a minimum thus it is more convenient to take larger steps to accelerate the convergence.
- (•) The intuition behind it is that for later epochs our parameters are closer 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.

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.
- [x] Increasing β will shift the red line slightly to the right.

green line $\beta = 0.98$ that is slightly shifted to the right.

- [x] Decreasing β 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 lot of oscillations.

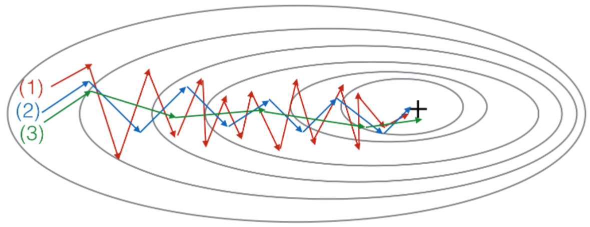
- [ ] Increasing

β

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

β), (2) is gradient descent with momentum (small

β

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

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) 1 / 1 point

- [x] 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.

- [x] 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.

- [x] 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.

- [ ] Add more data to the training set.

Expand

Correct
Great, you got all the right 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 combines the advantages of RMSProp and momentum.
- ( ) Adam automatically tunes the hyperparameter α.

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.