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

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

Expand

✓ **Correct**

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

1 / 1 point

- (•) When the mini-batch size is the same as the training size, mini-batch gradient descent is equivalent to 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**
Correct. Batch gradient descent uses all the examples at each iteration, this is equivalent to having only one mini-batch of the size of the complete training set in mini-batch gradient descent.

3. We usually choose a mini-batch size greater than 1 and less than m, because that way we make use of vectorization but not fall into the slower case of batch gradient descent.

1 / 1 point

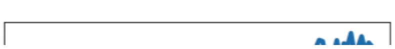
- (•) True
- ( ) False

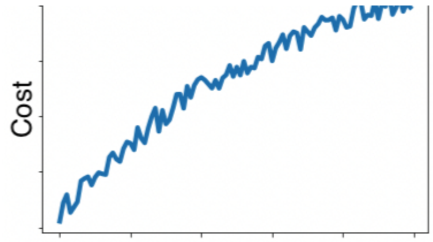
Expand

✓ **Correct**
Correct. Precisely by choosing a batch size greater than one we can use vectorization; but we choose a value less than m so we won't end up using 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





Which of the following do you agree with?

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

Expand

Correct
Yes. 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 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.)

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

Expand

Correct

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

0 / 1 point

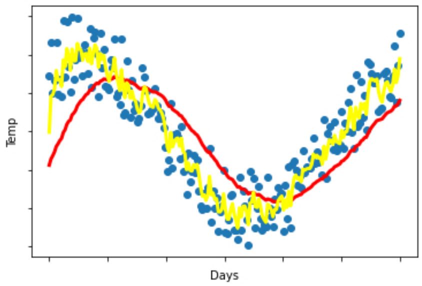
- ◉ $\alpha = \dfrac{\alpha_0}{1+3t}$
- ○ $\alpha = e^{-0.01\,t}\alpha_0.$
- ○ $\alpha = \dfrac{\alpha_0}{\sqrt{1+t}}.$
- ○ $\alpha = 1.01^t\,\alpha_0$

Expand

Incorrect
Incorrect. This is a good learning rate decay since it is a decreasing 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 $beta_1$ and $beta_2$ respectively. Which of the following are true?

1 / 1 point



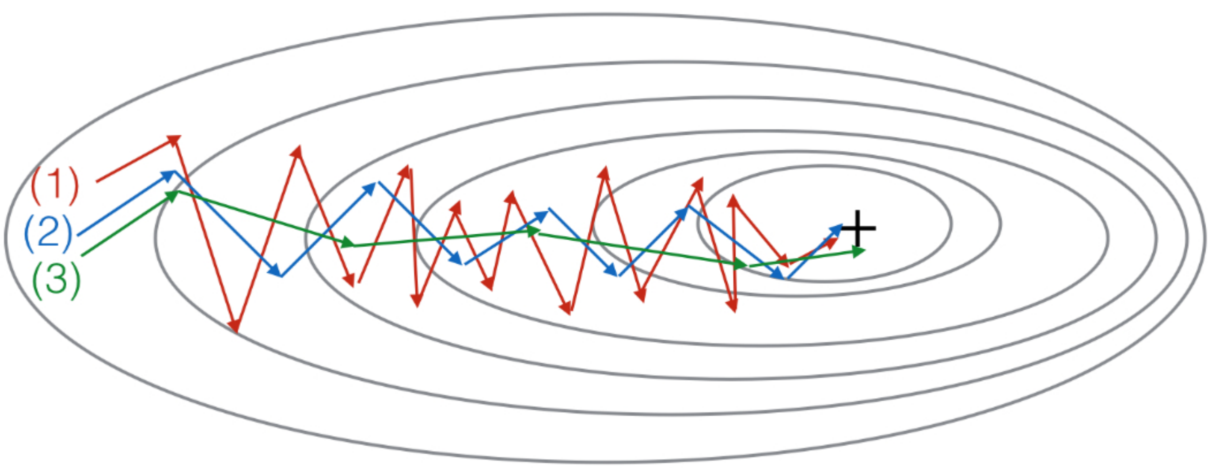
- ○ $\beta_1 = 0, \beta_2 > 0$.
- ◉ $\beta_1 < \beta_2$.
- ○ $\beta_1 = \beta_2$.
- ○ $\beta_1 > \beta_2$.

Expand

Correct
Correct. $$\beta_1 < \beta_2$$ since the yellow 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 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 (large β). (3) is gradient descent with momentum (small β)

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] Try using Adam.

  Correct
  Yes. Adam combines the advantages of other methods to accelerate the convergence of the gradient descent.

- [ ] Try initializing the weight at zero.
- [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 mini-batch gradient descent.

  Correct
  Yes. Mini-batch gradient descent is faster than batch gradient descent.

Expand

Correct
Great, you got all the right answers.

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

1 / 1 point

- ( ) Adam combines the advantages of RMSProp and momentum
- ( ) 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 should be used with batch gradient computations, not with mini-batches.
- ( ) The learning rate hyperparameter α in Adam usually needs to be tuned.

Expand

Correct