Your grade: 100%

Your latest: **100%** • Your highest: **100%** • To pass you need at least 80%. We keep your highest score.

Next item →

1. Using the notation for mini-batch gradient descent. To what of the following does $a^{[2]\{4\}(3)}$ correspond? **1 / 1 point**

- (●) The activation of the second layer when the input is the third example of the fourth mini-batch.
- ( ) The activation of the second layer when the input is the fourth example of the third mini-batch.
- ( ) The activation of the third layer when the input is the fourth example of the second mini-batch.
- ( ) The activation of the fourth layer when the input is the second example of the third mini-batch.

Correct
Yes. In general $a^{[l]\{t\}(k)}$ denotes the activation of the layer l when the input is the example k from the mini-batch t.

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

- ( ) 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.
- (●) When the mini-batch size is the same as the training size, mini-batch gradient descent is equivalent to batch gradient descent.

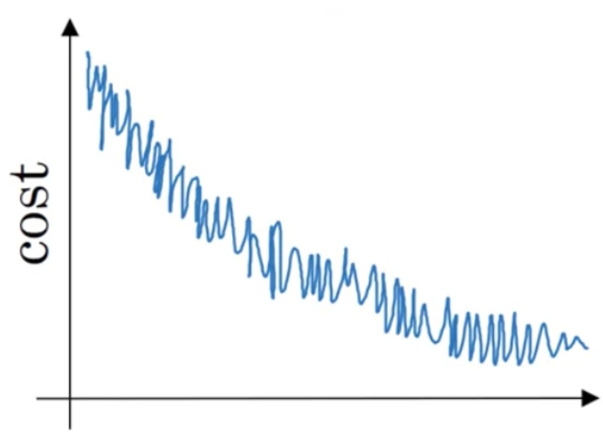
Correct
Correct. Batch gradient descent uses all the examples at each iteration, this is equivalent to having only one mini-batch of the size of the complete training set in mini-batch gradient descent.

3. 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 has as many mini-batches as examples in the training set.
- (●) 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.

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

4. Suppose your learning algorithm's cost J, plotted as a function of the number of iterations, looks like this: **1 / 1 point**



Which of the following do you agree with?

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

◉ If you're using mini-batch gradient descent, this looks acceptable. But if you're using batch gradient descent, something is wrong.

Correct

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

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

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

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?

○ $v_2 = 20, v_2^{\text{corrected}} = 15.$

○ $v_2 = 15, v_2^{\text{corrected}} = 15.$

○ $v_2 = 20, v_2^{\text{corrected}} = 20.$

◉ $v_2 = 15, v_2^{\text{corrected}} = 20.$

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 $\frac{15}{1-(0.5)^2} = 20.$

1 / 1 point

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

○ $\alpha = \frac{\alpha_0}{1+3t}$

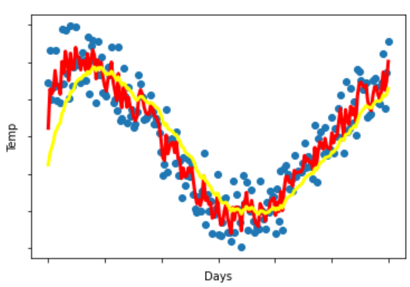
○ $\alpha = e^{-0.01\,t}\alpha_0.$

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

◉ $\alpha = 1.01^t\,\alpha_0$

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**



◉ $\beta_1 > \beta_2.$

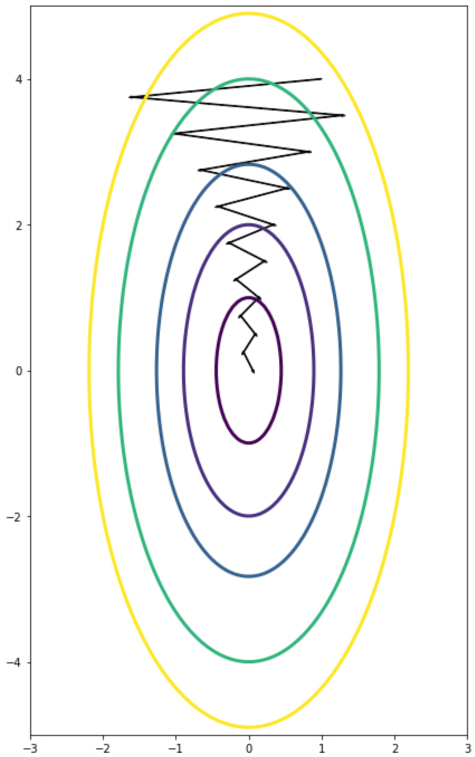
○ $\beta_1 = 0, \beta_2 > 0.$

○ $\beta_1 < \beta_2.$

○ $\beta_1 = \beta_2.$

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

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



Suppose this plot was generated with gradient descent with momentum $\beta = 0.01$. What happens if we increase the value of β to 0.1?

- ( ) The gradient descent process moves more in the horizontal and the vertical axis.
- ( ) The gradient descent process starts oscillating in the vertical direction.
- ( ) The gradient descent process starts moving more in the horizontal direction and less in the vertical.
- (•) The gradient descent process moves less in the horizontal direction and more in the vertical direction.

Correct
Yes. The use of a greater value of β causes a more efficient process thus reducing the oscillation in the horizontal direction and moving the steps more in the vertical direction.

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 $\mathcal{J}(W^{[1]}, b^{[1]}, \ldots, W^{[L]}, b^{[L]})$. 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

- [x] Try better random initialization for the weights

Correct

- [x] Try tuning the learning rate α

Correct

- [ ] Try initializing all the weights to zero
- [x] Try mini-batch gradient descent

Correct

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 ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$)

- ○ We usually use "default" values for the hyperparameters β_1, β_2 and ε in Adam ($\beta_1 = 0.9, \beta_2 = 0.999, \varepsilon = 10$)
- ○ Adam combines the advantages of RMSProp and momentum

Correct