Your grade: 90%

Your latest: **90%** • Your highest: **90%** • To pass you need at least 80%. We keep your highest score.

Next item →

1. Which notation would you use to denote the 4th layer's activations when the input is the 7th example from the 3rd mini-batch?

1 / 1 point

- ○ $a^{[3]\{7\}(4)}$
- ◉ $a^{[4]\{3\}(7)}$
- ○ $a^{[7]\{3\}(4)}$

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?

- ○ Training one epoch (one pass through the training set) using mini-batch gradient descent is faster than training one epoch using 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).
- ◉ When the mini-batch size is the same as the training size, mini-batch gradient descent is equivalent to batch gradient descent.

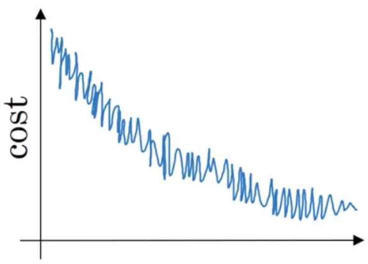
Correct
Correct. Batch gradient descent uses all the examples at each iteration, this is equivalent to having only one mini-batch of the size of the complete training set in mini-batch gradient descent.

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

- ○ 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.

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:



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

- ◉ 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.
- ○ You are not implementing the moving averages correctly. Using moving averages will smooth the graph.

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 March are the following: March 1st: $\theta_1 = 30^\circ$ C March 2nd: $\theta_2 = 15^\circ$ 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?

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

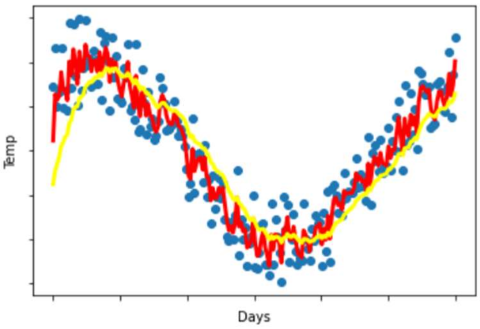
Correct
Correct. $v_2 = \beta v_{t-1} + (1-\beta)\,\theta_t$ thus $v_1 = 15, v_2 = 15$. Using the bias correction $\frac{v_t}{1-\beta^t}$ we get $\frac{15}{1-(0.5)^2} = 20.$

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

- ○ $\alpha = \frac{1}{1+2*t}\alpha_0$
- ○ $\alpha = 0.95^t\alpha_0$
- ○ $\alpha = \frac{1}{\sqrt{t}}\alpha_0$
- ◉ $\alpha = e^t\alpha_0$

Correct

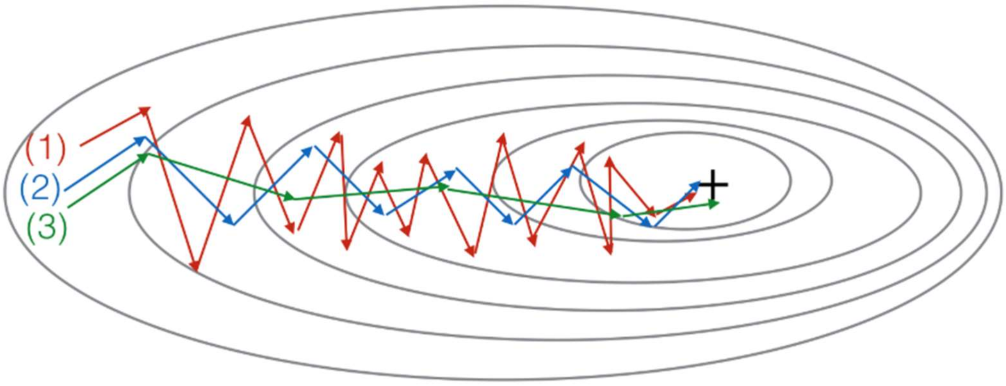
7. You use an exponentially weighted average on the London temperature dataset. You use the following to track the temperature: $v_t = \beta v_{t-1} + (1-\beta)\theta_t$. The yellow and red lines were computed using values β_1 and β_2 respectively. Which of the following are true?



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

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

8. Consider this figure:



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 (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
- ◉ (1) is gradient descent. (2) is gradient descent with momentum (small β). (3) is gradient descent with momentum (large β)

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)

- [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.

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

- ○ Adam combines the advantages of RMSProp and momentum
- ○ The learning rate hyperparameter α in Adam usually needs to be tuned.
- ○ 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.

Correct