Your grade: 87.50%

Your latest: **87.50%** • Your highest: **87.50%** • 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?

0 / 1 point

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

Incorrect
No. 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. 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 $m/2$.
- ( ) Stochastic Gradient Descent, Batch Gradient Descent, and Mini-Batch Gradient Descent all make equal use of vectorization.
- (•) Batch Gradient Descent
- ( ) Stochastic Gradient Descent

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.

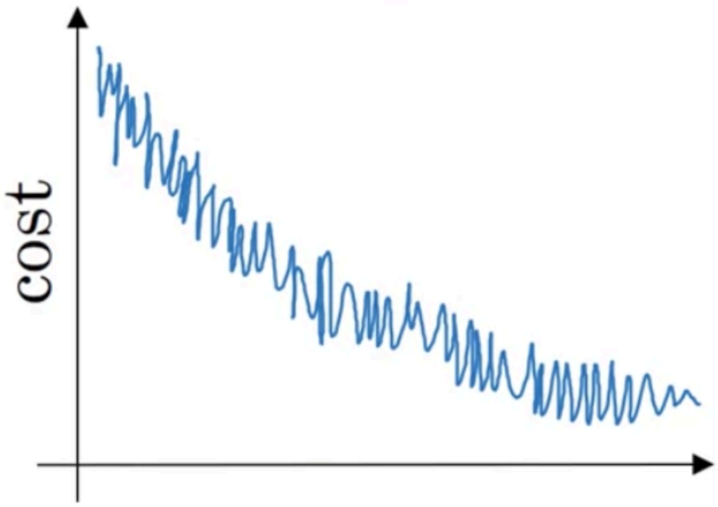
Correct

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

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.
- ◉ 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.
- ○ 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 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.)

- (selected) $v_2 = 7.5, v_2^{corrected} = 10$
- $v_2 = 10, v_2^{corrected} = 10$
- $v_2 = 10, v_2^{corrected} = 7.5$
- $v_2 = 7.5, v_2^{corrected} = 7.5$

Correct

1 / 1 point

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

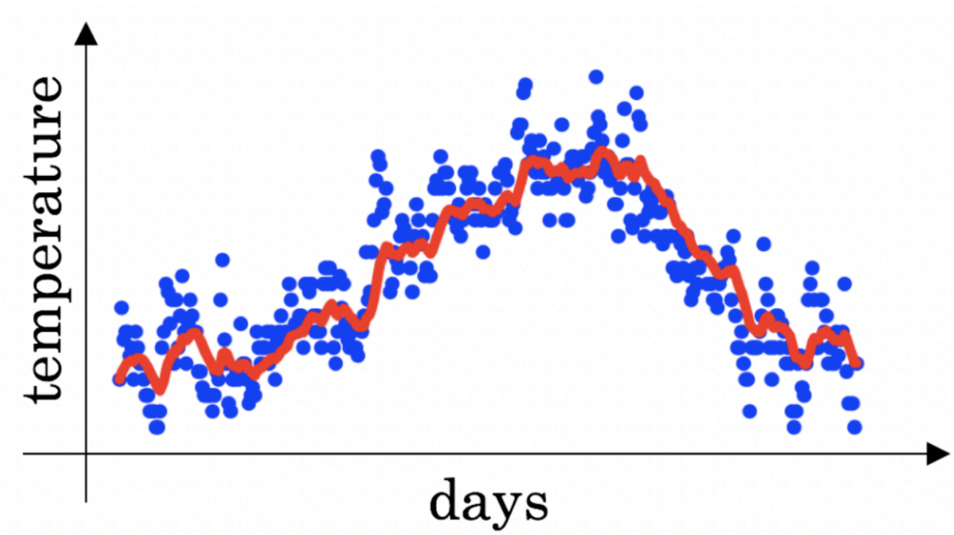
- $\alpha = e^{-0.01\,t}\alpha_0$.
- (selected) $\alpha = 1.01^t\,\alpha_0$
- $\alpha = \frac{\alpha_0}{1+3\,t}$
- $\alpha = \frac{\alpha_0}{\sqrt{1+t}}$.

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 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.

> **Correct**
> True, remember that the red line corresponds to $\beta = 0.9$. In the lecture we had a 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 will create more oscillations within the red line.

8. Which of the following are true about gradient descent with momentum?

0.75 / 1 point

- [x] Gradient descent with momentum makes use of moving averages.

Correct
Correct. Gradient descent with momentum makes use of moving averages, which smooths out the gradient descent process.

- [x] It generates faster learning by reducing the oscillation of the gradient descent process.

Correct
Correct. The use of momentum makes each step of the gradient descent more efficient by reducing oscillations.

- [ ] It decreases the learning rate as the number of epochs increases.
- [ ] Increasing the hyperparameter β smooths out the process of gradient descent.

You didn't select all the correct answers

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

- [ ] 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.

- [x] Try using Adam.

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

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

1 / 1 poi

- ○ Adam combines the advantages of RMSProp and momentum
- ◉ 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 ($\beta_1 = 0.9$, $\beta_2 = 0.999, \varepsilon = 10^{-8}$)
- ○ The learning rate hyperparameter α in Adam usually needs to be tuned.

Correct