Your grade: 100%

Your latest: **100%** • Your highest: **100%**
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

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

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

- ◉ One iteration of mini-batch gradient descent (computing on a single mini-batch) is faster than one iteration of batch gradient descent.
- ○ 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).

Correct

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.

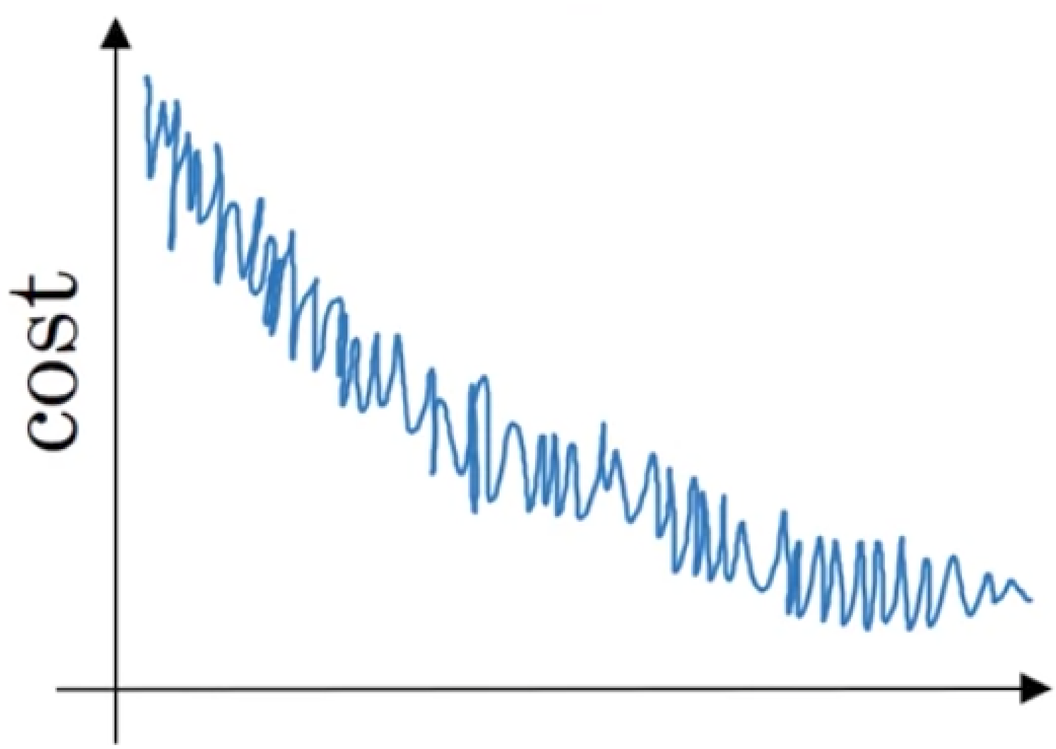
1 / 1 point

- False
- True

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. 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?

- ◯ 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.
- ◯ 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.

Correct

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

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

March 2nd: $\theta_2 = 25^\circ$ 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?

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

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

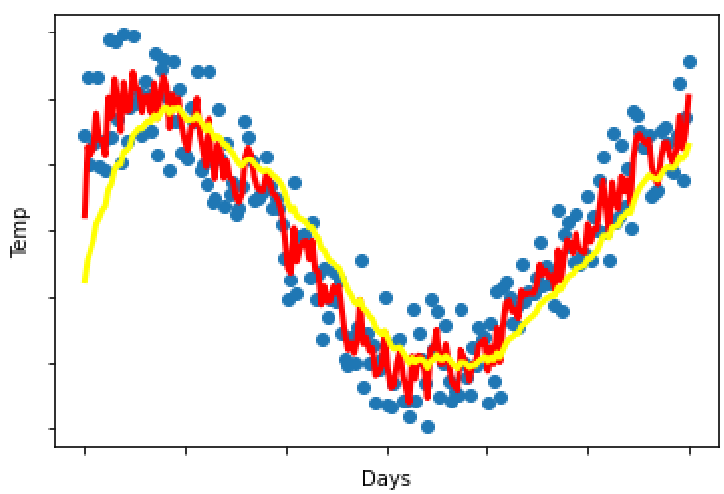
1 / 1 point

- ( ) $\alpha = 0.95^t \alpha_0$
- ( ) $\alpha = \frac{1}{\sqrt{t}} \alpha_0$
- ( ) $\alpha = \frac{1}{1+2*t} \alpha_0$
- (•) $\alpha = e^t \alpha_0$

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?

1 / 1 point



- [x] $\beta_1 > \beta_2$.
- [ ] $\beta_1 = 0, \beta_2 > 0$.
- [ ] $\beta_1 = \beta_2$.
- [ ] $\beta_1 < \beta_2$.

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

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

1 / 1 point

- [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.

- [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] Increasing the hyperparameter β smooths out the process of gradient descent.

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

- [ ] It decreases the learning rate as the number of epochs increases.

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 $\mathrm{J}\,(W^{[1]}, b^{[1]}, ..., W^{[L]}, b^{[L]})$. Which of the following techniques could help find parameter values that attain a small value for$\mathrm{J}$? (Check all that apply)

1 / 1 point

- [x] Try tuning the learning rate α

Correct

- [x] Try using Adam

Correct

- [ ] Try initializing all the weights to zero

- [x] Try better random initialization for the weights

Correct

- [x] Try mini-batch gradient descent

Correct

10. Which of the following are true about Adam?

- ○ The most important hyperparameter on Adam is ϵ and should be carefully tuned.
- ◉ Adam combines the advantages of RMSProp and momentum.
- ○ Adam can only be used with batch gradient descent and not with mini-batch gradient descent.
- ○ Adam automatically tunes the hyperparameter α.

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.