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

○ $a^{[8]\{7\}(3)}$

○ $a^{[8]\{3\}(7)}$

○ $a^{[3]\{7\}(8)}$

● $a^{[3]\{8\}(7)}$

✓ **Correct**

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

○ Training one epoch (one pass through the training set) using mini-batch gradient descent is faster than training one epoch using batch gradient descent.

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

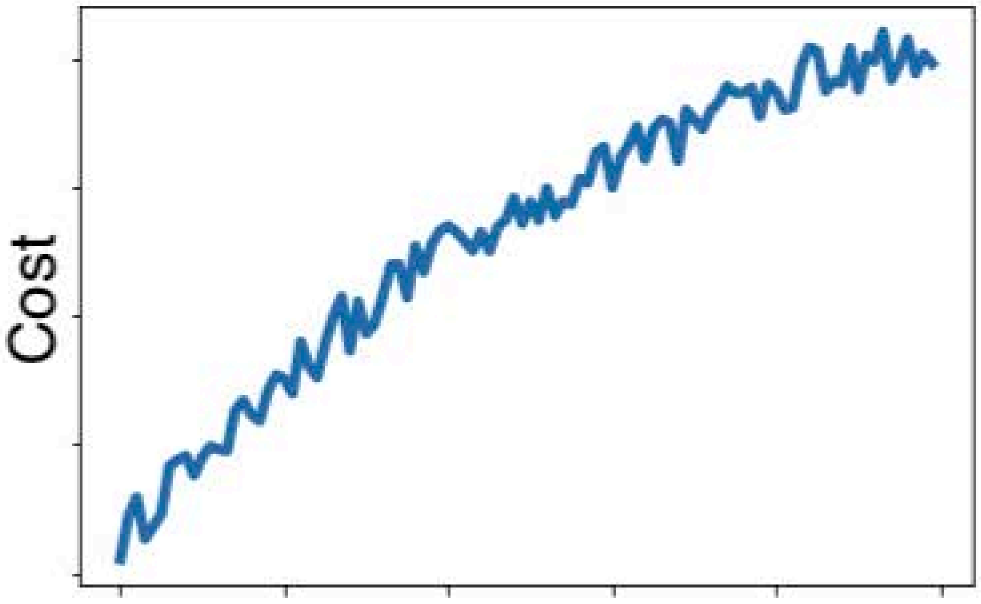
✓ **Correct**

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. 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

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

○ If you are using mini-batch gradient descent or batch gradient descent this looks acceptable.

✓ **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 March are the following:

1 / 1 point

March 1st: $\theta_1 = 30^\circ \text{ C}$

March 2nd: $\theta_2 = 15^\circ \text{ C}$

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?

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

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

1 / 1 point

March 1st: $\theta_1 = 30^\circ \text{ C}$

March 2nd: $\theta_2 = 15^\circ \text{ C}$

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?

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

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

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

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

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.

1 / 1 point

● $\alpha = e^t \alpha_0$

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

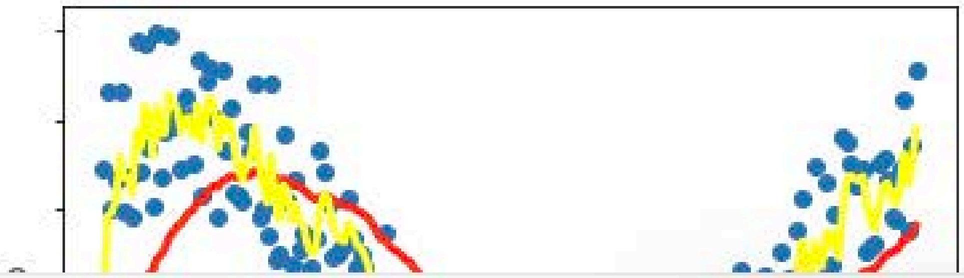
○ $\alpha = \frac{1}{1+2*t} \alpha_0$

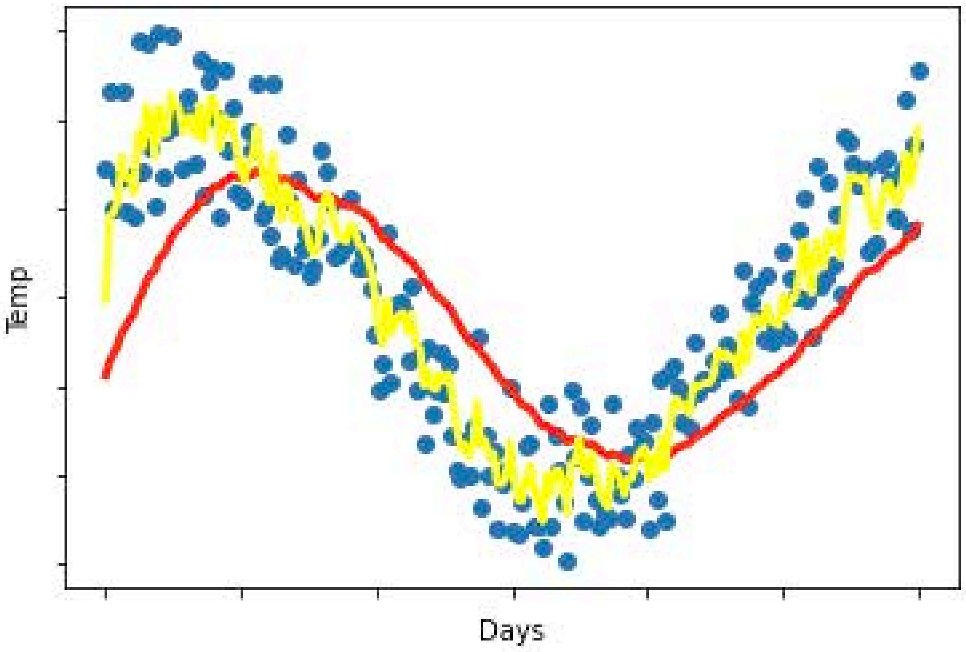
○ $\alpha = 0.95^t \alpha_0$

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 $beta_1$ and $beta_2$ respectively. Which of the following are true?

1 / 1 point





○ $\beta_1 > \beta_2$.

● $\beta_1 < \beta_2$.

○ $\beta_1 = \beta_2$.

○ $\beta_1 = 0, \beta_2 > 0$.

Correct
Correct. $\beta_1 < \beta_2$ since the yellow curve is noisier.

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

1 / 1 point

☐ It decreases the learning rate as the number of epochs increases.

☑ 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.

☑ 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 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.

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

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 better random initialization for the weights

Correct
Yes. As seen in previous lectures this can help the gradient descent process to prevent vanishing gradients.

☑ Try using gradient descent with momentum.

Correct
Yes. The use of momentum can improve the speed of the training. Although other methods might give better results, such as Adam.

☑ 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.

□ Add more data to the training set.

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

1 / 1 point

○ 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 combines the advantages of RMSProp and momentum

○ The learning rate hyperparameter α in Adam usually needs to be tuned.

● Adam should be used with batch gradient computations, not with mini-batches.

Correct