Congratulations! You passed!

Grade received 100%

Latest Submission Grade 100%

To pass 80% or higher

Retake the assignment in **24h 57m**

Go to 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

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

Expand

✓ **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

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

Expand

✓ **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

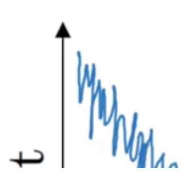
- ( ) It is the same as stochastic gradient descent, but we don't use random elements.
- (•) It is the same as the mini-batch gradient descent when the mini-batch size is the same as the size of the training set.
- ( ) It has as many mini-batches as examples in the training set.

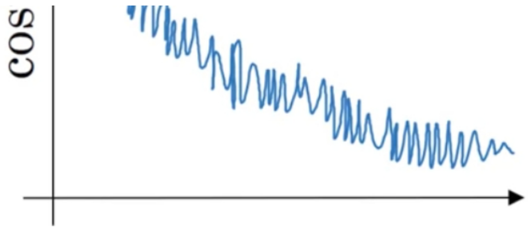
Expand

✓ **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?

Expand

Correct

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.)

Expand

Correct

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

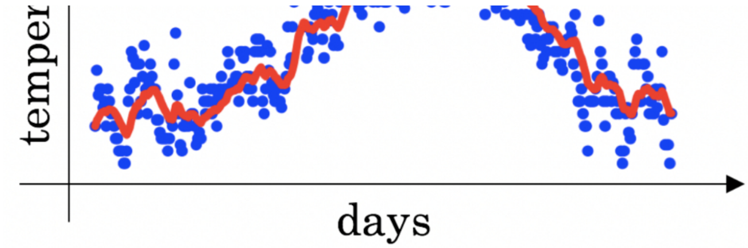
1 / 1 point

Expand

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 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

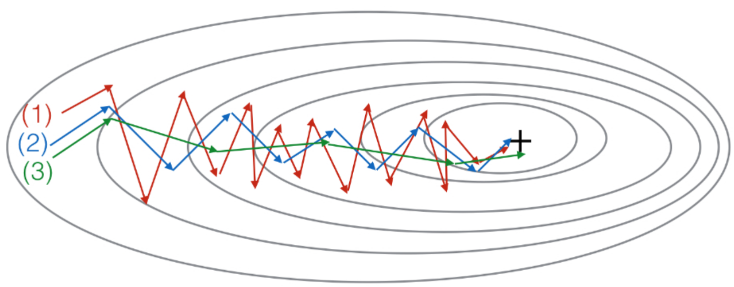


Expand

Correct
Great, you got all the right answers.

8. Consider this figure:

1 / 1 point



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?

Expand

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)

1 / 1 point

Expand

Correct
Great, you got all the right answers.

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

1 / 1 point

Expand

Correct