← Back Optimization Algorithms
Graded Quiz • 50 min

English Due Sep 1, 11:59 PM +06

Congratulations! You passed!

Grade received 90% **Latest Submission Grade** 90% **To pass** 80% or higher

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

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

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

- ( ) 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).

Expand

Correct

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

- [ ] 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

- [x] If the mini-batch size is 1, you lose the benefits of vectorization across examples in the mini-batch.

  ✓ Correct

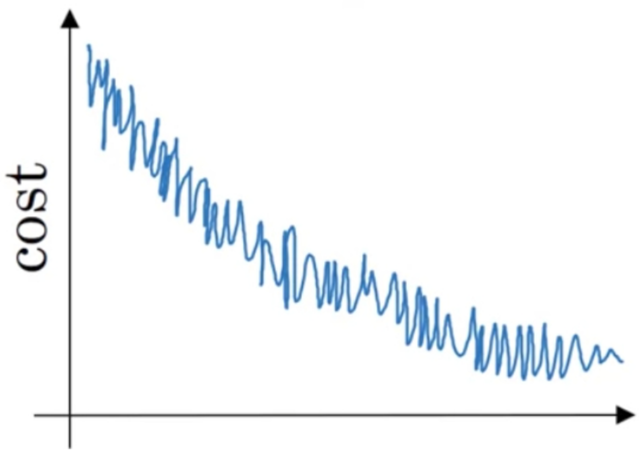
Expand

Correct
Great, you got all the right answers.

Great, you got all the right answers.

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.
- Whether you're using batch gradient descent or mini-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, something is wrong.

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.)

- $v_2 = 10$, $v_2^{corrected} = 7.5$
- $v_2 = 7.5$, $v_2^{corrected} = 7.5$
- $v_2 = 10$, $v_2^{corrected} = 10$
- $v_2 = 7.5$, $v_2^{corrected} = 10$

Expand

Correct

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

1 / 1 point

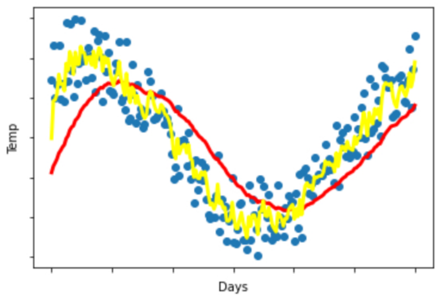
- $\alpha = \frac{1}{\sqrt{t}}\alpha_0$
- $\alpha = 0.95^t\alpha_0$
- $\alpha = \frac{1}{1+2*t}\alpha_0$
- $\alpha = e^t\alpha_0$

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 yellow and red lines were computed using values $beta_1$ and $beta_2$ respectively. Which of the following are true?

1 / 1 point



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

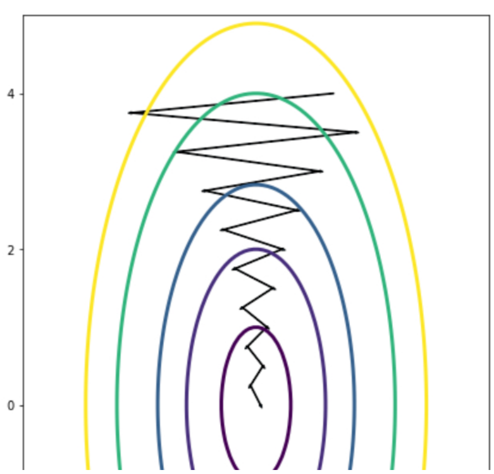
Expand

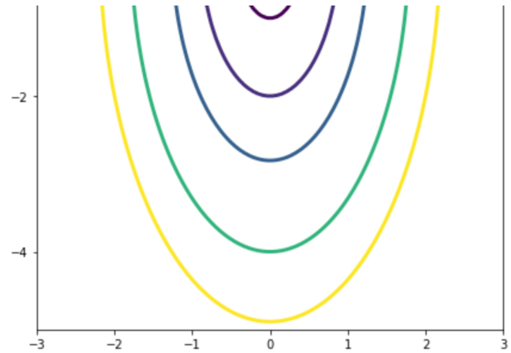
Correct

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

8. Consider the figure:

0 / 1 point





Suppose this plot was generated with gradient descent with momentum $\beta = 0.01$. What happens if we increase the value of β to 0.1?

- ( ) The gradient descent process moves less in the horizontal direction and more in the vertical direction.
- ( ) The gradient descent process moves more in the horizontal and the vertical axis.
- (•) The gradient descent process starts moving more in the horizontal direction and less in the vertical.
- ( ) The gradient descent process starts oscillating in the vertical direction.

Expand

Incorrect
No. The use of a greater value of β causes a more efficient process thus reducing the oscillation in the horizontal direction and moving the steps more in the vertical direction.

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

- [x] Try using Adam

  Correct

- [ ] Try initializing all the weights to zero
- [x] Try tuning the learning rate α

  Correct

- [x] Try mini-batch gradient descent

  Correct

- [x] Try better random initialization for the weights

  Correct

Expand

Correct
Great, you got all the right answers.

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

1 / 1 point

- Adam combines the advantages of RMSProp and momentum
- Adam should be used with batch gradient computations, not with mini-batches.
- 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 ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$)

Expand

Correct