Congratulations! You passed!

Grade received 80%

Latest Submission Grade 80%

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

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

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?

0 / 1 point

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

Expand

Incorrect

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

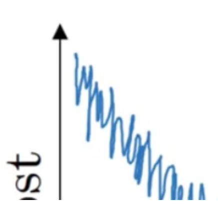
- False
- (selected) True

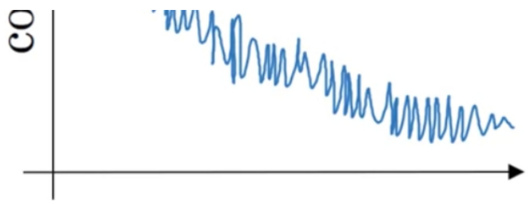
Expand

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

Expand

Correct

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

1 / 1 point

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

March 2nd: $\theta_2 = 15^\circ$ 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?

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

Expand

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

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

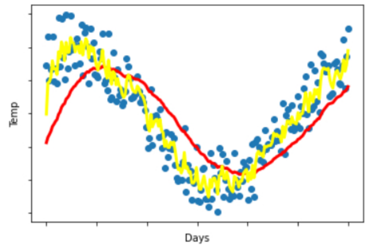
Expand

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

1 / 1 point



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

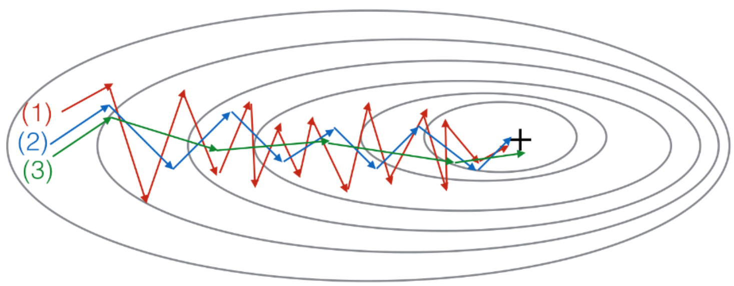
Expand

Correct

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

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?

- [ ] (1) is gradient descent with momentum (small β), (2) is gradient descent with momentum (small β), (3) is gradient descent
- [ ] (1) is gradient descent. (2) is gradient descent with momentum (large β). (3) is gradient descent with momentum (small β)
- [ ] (1) is gradient descent with momentum (small β). (2) is gradient descent. (3) is gradient descent with momentum (large β)
- [x] (1) is gradient descent. (2) is gradient descent with momentum (small β). (3) is gradient descent with momentum (large β)

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)

0 / 1 point

- [ ] Try using Adam
- [ ] Try better random initialization for the weights
- [ ] Try initializing all the weights to zero
- [x] Try tuning the learning rate α

Correct

- [x] Try mini-batch gradient descent

✓ Correct

Expand

Incorrect
You didn't select all the correct answers

10. Which of the following are true about Adam?

1 / 1 point

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

Expand

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.