Congratulations! You passed!

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

Go to next item

1. Using the notation for mini-batch gradient descent. To what of the following does $a^{[2]\{4\}(3)}$ correspond?

1 / 1 point

- ○ The activation of the fourth layer when the input is the second example of the third mini-batch.
- ○ The activation of the third layer when the input is the fourth example of the second mini-batch.
- ○ The activation of the second layer when the input is the fourth example of the third mini-batch.
- ◉ The activation of the second layer when the input is the third example of the fourth mini-batch.

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

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. 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

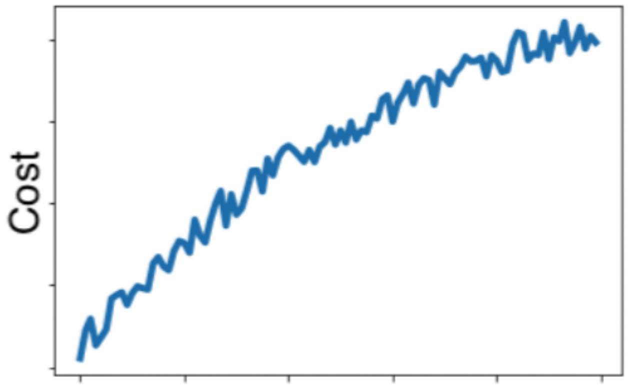
Expand

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. 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:

0 / 1 point



Which of the following do you agree with?

- [ ] No matter if using mini-batch gradient descent or batch gradient descent something is wrong.
- [x] If you are using mini-batch gradient descent or batch gradient descent this looks acceptable.
- [ ] 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, this looks acceptable. But if you're using batch gradient descent, something is wrong.

Expand

Incorrect

No. 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: $\theta_1 = 10^\circ \mathrm{C}$

March 2nd: $\theta_2 = 25^\circ \mathrm{C}$

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?

- ○ $v_2 = 20'\ v_2^{\text{corrected}} = 15^{\cdot}$
- ◉ $v_2 = 15'\ v_2^{\text{corrected}} = 20^{\cdot}$
- ○ $v_2 = 15'\ v_2^{\text{corrected}} = 15^{\cdot}$
- ○ $v_2 = 20'\ v_2^{\text{corrected}} = 20^{\cdot}$

Expand

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.

0 / 1 point

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

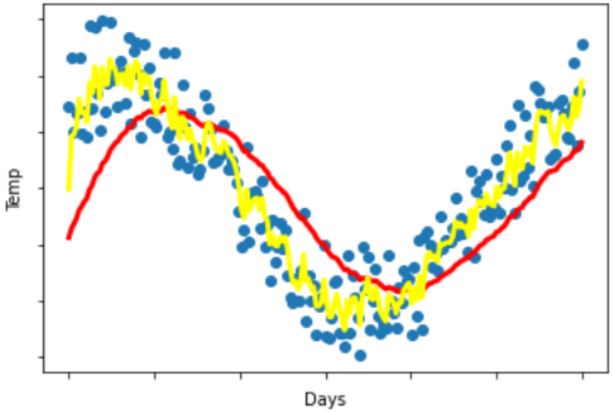
Expand

Incorrect

Incorrect. This is a good learning rate decay since it is a decreasing 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



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

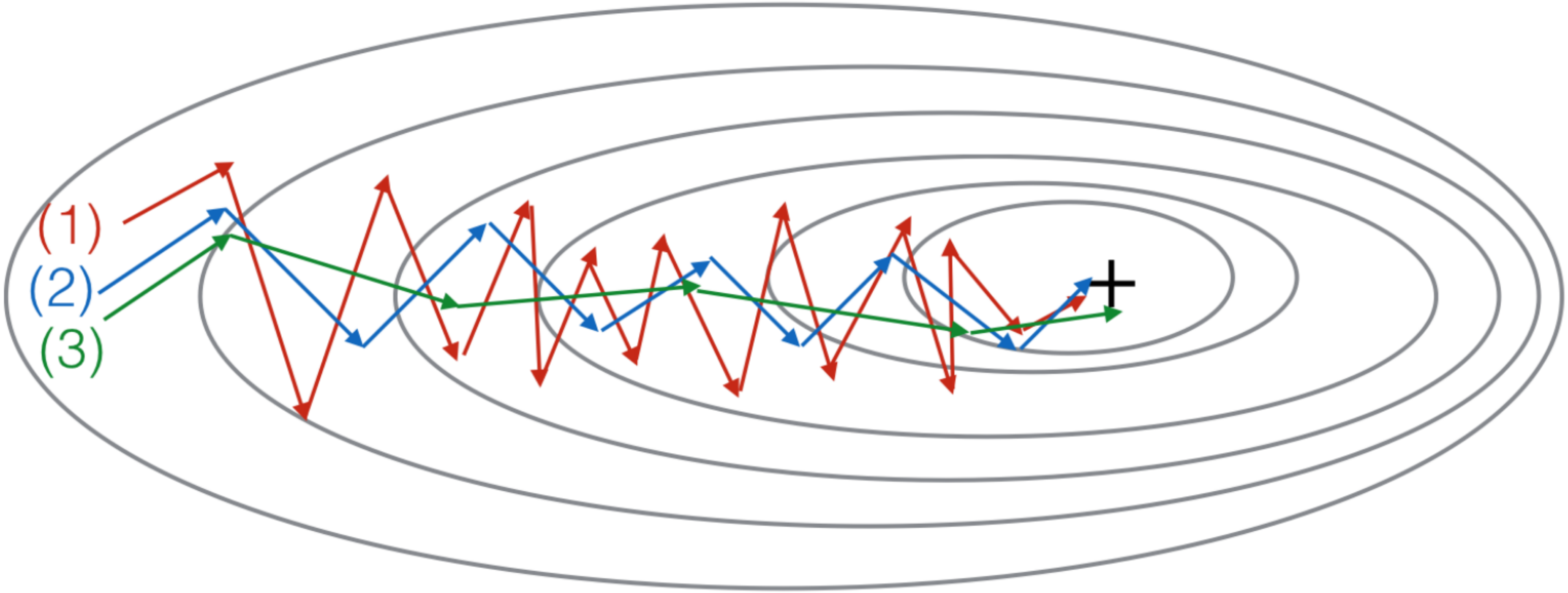
Expand

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 ($\beta = 0.5$); and gradient descent with momentum ($\beta = 0.9$). 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 (small β). (3) is gradient descent with momentum (large β)
- ○ (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)

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

- [x] 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.

- [x] 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.

- [x] 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.

Expand

Correct

Great, you got all the right 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 automatically tunes the hyperparameter α.
- ○ Adam can only be used with batch gradient descent and not with mini-batch gradient descent.
- ◉ Adam combines the advantages of RMSProp and momentum.

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.