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

- (selected) $a^{[3]\{8\}(7)}$
- $a^{[8]\{3\}(7)}$
- $a^{[8]\{7\}(3)}$
- $a^{[3]\{7\}(8)}$

Expand

Correct

2. Suppose you don't face any memory-related problems. Which of the following make more use of vectorization.

1 / 1 point

- ◯ Stochastic Gradient Descent, Batch Gradient Descent, and Mini-Batch Gradient Descent all make equal use of vectorization.
- ◯ Mini-Batch Gradient Descent with mini-batch size $m/2$.
- ◉ Batch Gradient Descent
- ◯ Stochastic Gradient Descent

Expand

Correct

Yes. If no memory problem is faced, batch gradient descent processes all of the training set in one pass, maximizing the use of vectorization.

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

- (selected) True
- False

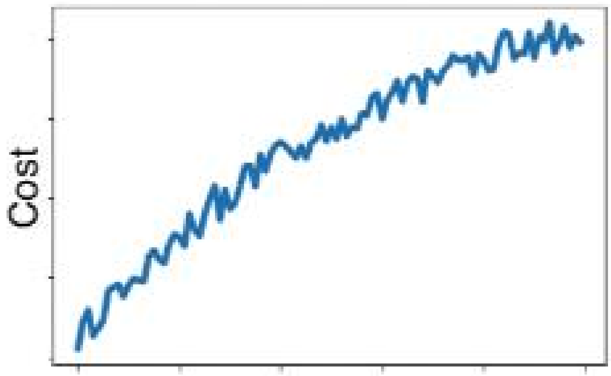
Expand

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?

- ○ 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.
- ○ No matter if using mini-batch gradient descent or 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 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.)

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

Expand

Correct

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

0 / 1 point

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

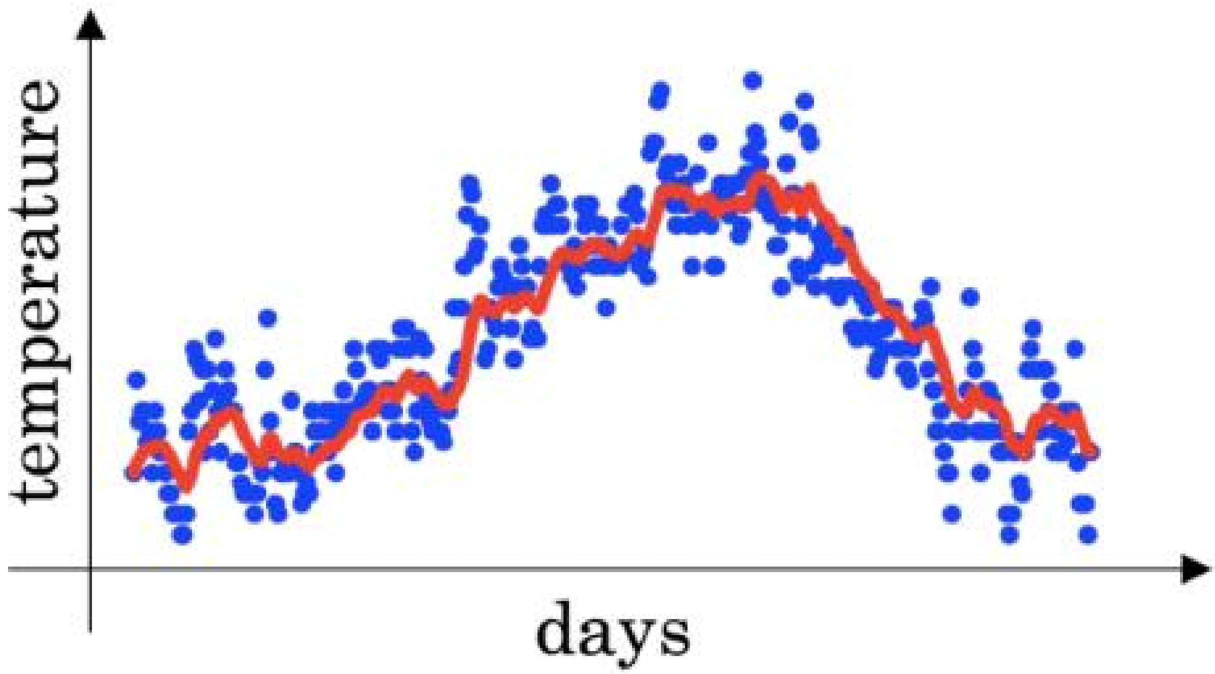
Expand

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 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



- [ ] Decreasing β will shift the red line slightly to the right.
- [x] Increasing β will shift the red line slightly to the right.

✓ Correct

True, remember that the red line corresponds to $\beta = 0.9$. In the lecture we had a green line $\beta = 0.98$ that is slightly shifted to the right.

- [x] Decreasing β will create more oscillation within the red line.

✓ Correct

True, remember that the red line corresponds to $\beta = 0.9$. In lecture we had a yellow line $\beta = 0.98$ that had a lot of oscillations.

- [ ] Increasing β will create more oscillations within the red line.

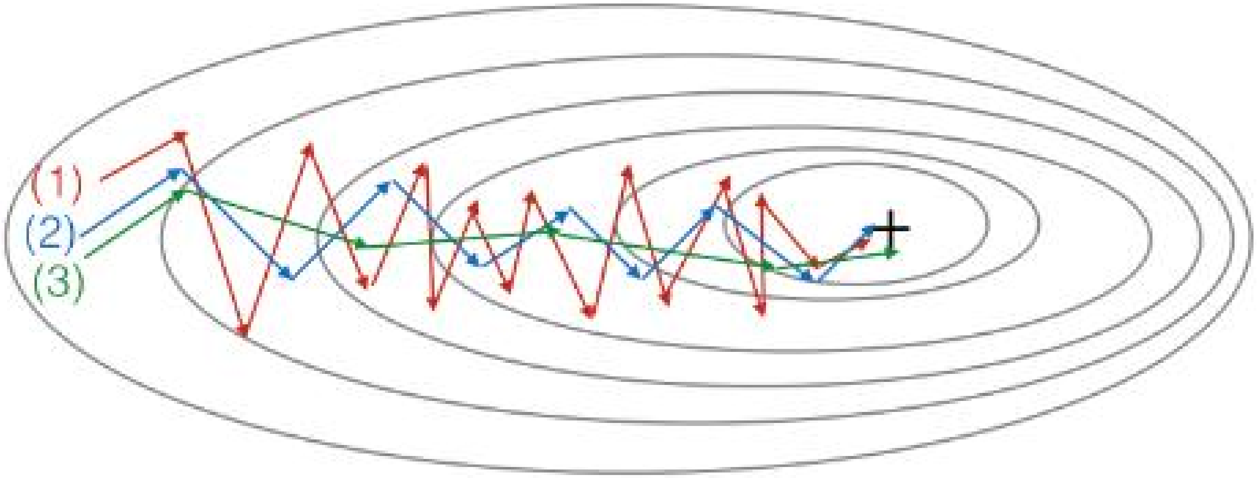
Expand

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?

- ○ (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 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 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 tuning the learning rate α

Correct

- [ ] Try initializing all the weights to zero
- [x] Try using Adam

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. In very high dimensional spaces it is most likely that the gradient descent process gives us a local minimum than a saddle point of the cost function. True/False?

1 / 1 point

- (●) False
- ( ) True

Expand

Correct

Correct. Due to the high number of dimensions it is much more likely to reach a saddle point, than a local minimum.