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. Which of these statements about mini-batch gradient descent do you agree with?

1 / 1 point

- (selected) One iteration of mini-batch gradient descent (computing on a single mini-batch) is faster than one iteration of 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

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

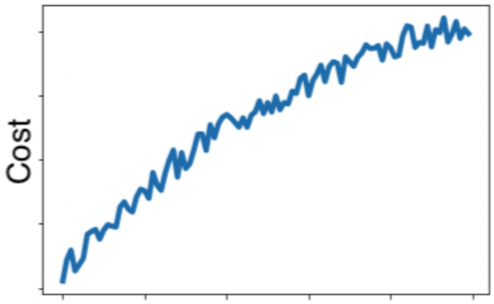
Correct

Correct. Precisely by choosing a batch size greater than one we can use vectorization; but we choose a value less than m so we

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, this looks acceptable. But if you're using 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.

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:

0 / 1 point

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

March 2nd: $\theta_2 = 15^\circ \text{ C}$

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?

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

  $v_2 = 20,$

  $v_2^{\text{corrected}} = 15$

  $v_2^{\text{corrected}} = 15$.

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

Expand

Incorrect

Incorrect. $$v_2 = \beta v_{t-1} + (1- \beta) \, \theta_t$$ 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

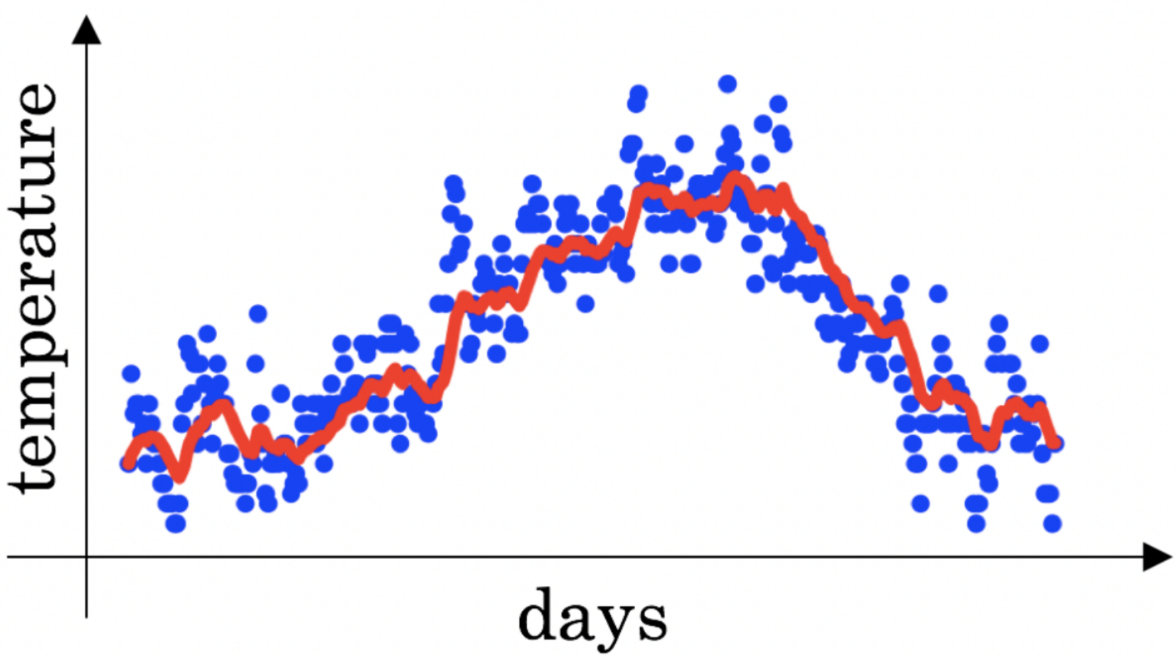
- [ ] $\alpha = \frac{1}{1+2*t}\alpha_0$
- [x] $\alpha = e^t \alpha_0$
- [ ] $$ \alpha = 0.95^t \alpha_0 $$
- [ ] $$\alpha = \frac{1}{\sqrt{t}} \alpha_0$$

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 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

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.

✓ **Correct**

True, remember that the red line corresponds to

$$\beta = 0.9$$

. In lecture we had a yellow line

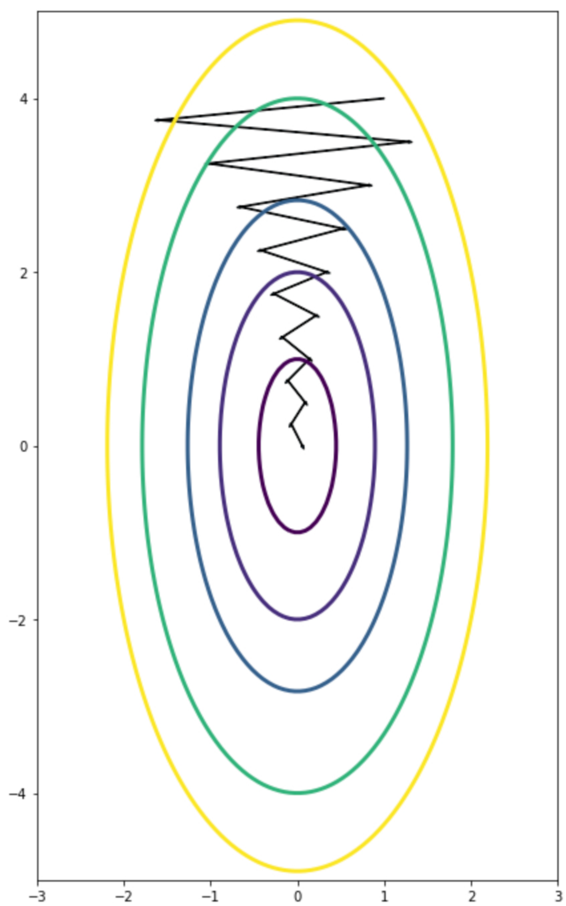
$$\beta = 0.98$$

Expand

Correct
Great, you got all the right answers.

8. Consider the figure:

1 / 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 starts oscillating in the vertical direction.
- ○ The gradient descent process starts moving more in the horizontal direction and less in the vertical.
- ◉ 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.

Expand

Correct
Yes. The use of a greater value of $$\beta$$ 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

- [x] Try better random initialization for the weights

  Correct

- [x] Try tuning the learning rate $$\alpha$$

  Correct

- [x] Try mini-batch gradient descent

  Correct

- [ ] Try initializing all the weights to zero

Loading [MathJax]/jax/output/CommonHTML/jax.js

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