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

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

Expand

Correct
Yes. In general $a^{[l]\{k\}(i)}$ denotes the activation of the layer l when the input is the example k from the mini-batch t.

2. Suppose you don't face any memory-related problems. Which of the following make more use of vectorization. 0 / 1 point

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

Expand

Incorrect
No. For example, if mini-batch size is 1 (Stochastic Gradient Descent), you lose all the benefits of vectorization across examples in the mini-batch.

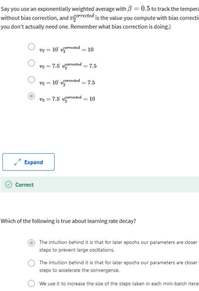
3. We usually choose a mini-batch size greater than 1 and less than m, because that way we make use of vectorization but not fall into the slower case of batch gradient descent. 1 / 1 point

- 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. 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?

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

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_t 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.)

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

Expand

Correct

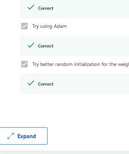
6. Which of the following is true about learning rate decay? 1 / 1 point

- The intuition behind it is that for later epochs our parameters are closer to a minimum thus it is more convenient to take smaller steps to prevent large oscillations.
- The intuition behind it is that for later epochs our parameters are closer to a minimum thus it is more convenient to take larger steps to accelerate the convergence.
- We use it to increase the size of the steps taken in each mini-batch iteration.
- It helps to reduce the variance of a model.

Expand

Correct
Correct. Reducing the learning rate with time reduces the oscillation around a minimum.

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 β_1 and β_2 respectively. Which of the following are true? 1 / 1 point

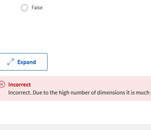


- $\beta_1 < \beta_2$
- $\beta_1 > \beta_2$
- $\beta_1 = 0$ $\beta_2 > 0$
- $\beta_1 = \beta_2$

Expand

Correct
Correct. $\beta_1 > \beta_2$ since the red curve is noisier.

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 moves more in the horizontal and the vertical axis.
- The gradient descent process moves less in the horizontal direction and more in the vertical direction.
- 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.

Expand

Correct
Yes. 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 tuning the learning rate α
  - Correct
- [ ] Try initializing all the weights to zero
- [x] Try mini-batch gradient descent
  - Correct
- [x] Try using Adam
  - 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? 0 / 1 point

- True
- False

Expand

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