Week 2: Optimization Algorithms

1. Which notation would you use to denote the 3rd layer activations when the input is the 7th example from the 8th mini-batch?

   Ans: $a^{[3]8(7)}$. `[i]j(k)` superscript implies i-th layer, j-th minibatch, k-th example

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

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

3. Why is the best mini-batch size usually not 1 and not m, but instead something in-between?

   Ans:

   - If the mini-batch size is 1, you lose the benefits of vectorization across examples in the mini-batch.
   - If the mini-batch size is m, you end up with batch gradient descent, which has to process the whole training set before making progress.

4. Suppose your cost J is plotted as a function of the number of iterations

   Ans: If you are using mini-batch gradient descent, oscilations are acceptable. With batch gradient descent, something could be wrong

5. Suppose the temperature in Casablanca over the first three days of January are the same: Jan 1st: 10°C, Jan 2nd: 10°C Say you use an exponentially weighted average with $\beta = 0.5$ to track the temperature:

   $$v_0 = 0$$
   $$v_t = \beta v_{t-1} + (1-\beta)\theta_t$$

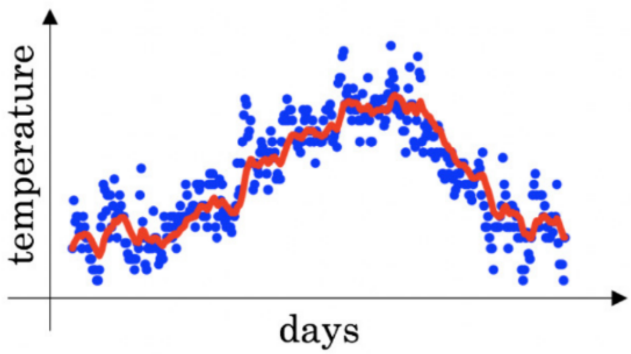
   If v_2 is the value computed after day-2 without bias correction, and $v_{corrected,2}$ is the value you compute with bias correction. What are these values?

   Ans: $v_2 = 7.5°\text{C}$, $v_{corrected,2} = 10°\text{C}$

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

   Ans: $\alpha = e^t \cdot \alpha_0$. This will explode the rate instead of decaying.

7. You use an exponentially weighted average on the London temperature dataset. You use the following to track the temperature with $\beta = 0.9$

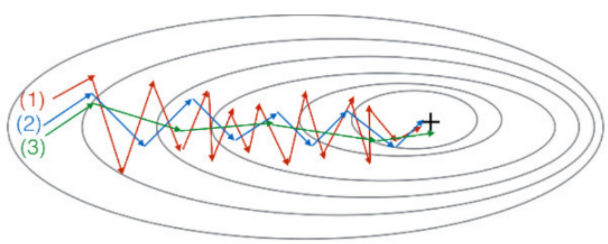
   

   $$v_t = \beta v_{t-1} + (1 - \beta)\theta_t$$

   What would happen to your moving average curve as you vary β? **Ans:**

   - Increasing β will shift the average line slightly to the right.
   - Decreasing β will create more oscillation within the average line.

8. Which curves correspond to which algorithm? **Ans:** Gradient descent has

   highest oscillation, gradient descent with small β for momentum will have some oscillations by few averaging, whereas, the least oscillation will come from gradient descent with large β. Hence,
   1: GD,
   2: GD + Momentum (small β),
   3: GD + Momentum (large β)

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 $J(W^{[1]}, b^{[1]}, ..., W^{[L]}, b^{[L]})$.
Which of the following techniques could help find parameter values that attain a small value forJ? (Check all that apply)

Ans:

- Try ADAM
- Try better random initialization for the weights
- Try tuning the learning rate α
- Try mini-batch gradient descent

10. Which of the following statements about ADAM is False?

    Ans: ADAM should be used with batch gradient computations, not with mini-batches.