Optimization Algorithms

1.
Question 1

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

1 / 1 point

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.
Question 2

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

1 / 1 point

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.
Question 3

Why is the best mini-batch size usually not 1 and not m, but instead something in-between? Check all that are true.

0 / 1 point

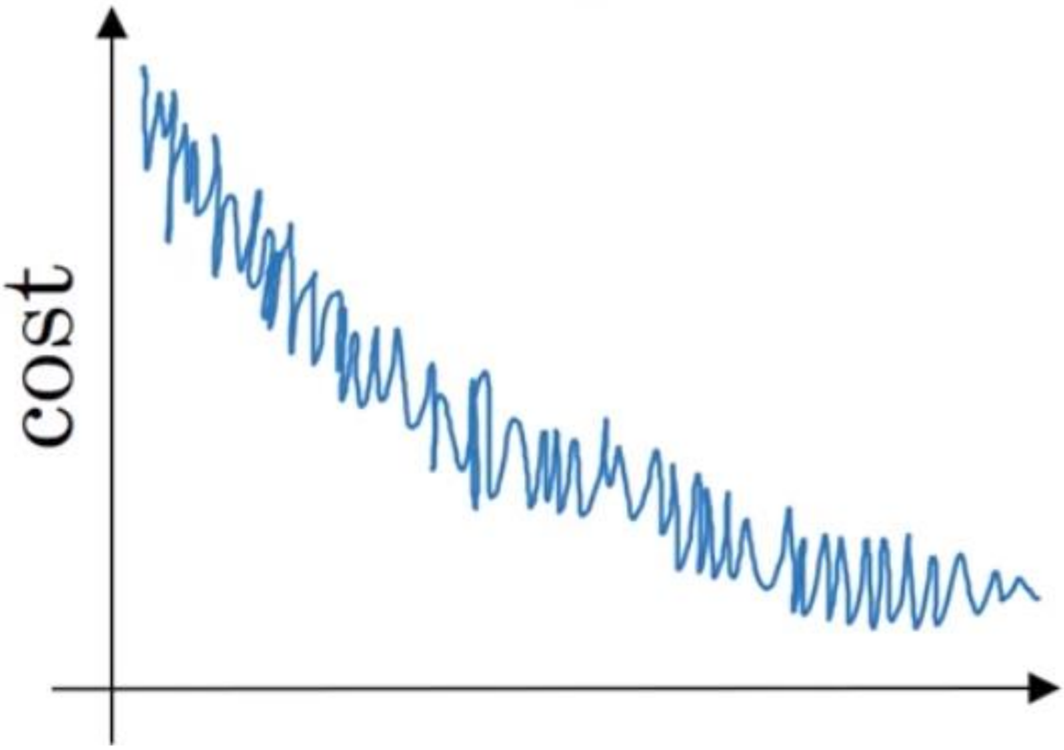
Expand

Incorrect

You didn't select all the correct answers

4.
Question 4

Suppose your learning algorithm's cost J, plotted as a function of the number of iterations, looks like this:



Which of the following do you agree with?

1 / 1 point

Expand

Correct

5.
Question 5

Suppose the temperature in Casablanca over the first two days of March are the following:

March 1st: $\theta_1 = 10^\circ\ C$

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

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 $v_2^{corrected}$ is the value you compute with bias correction. What are these values?

1 / 1 point

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.
Question 6

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

1 / 1 point

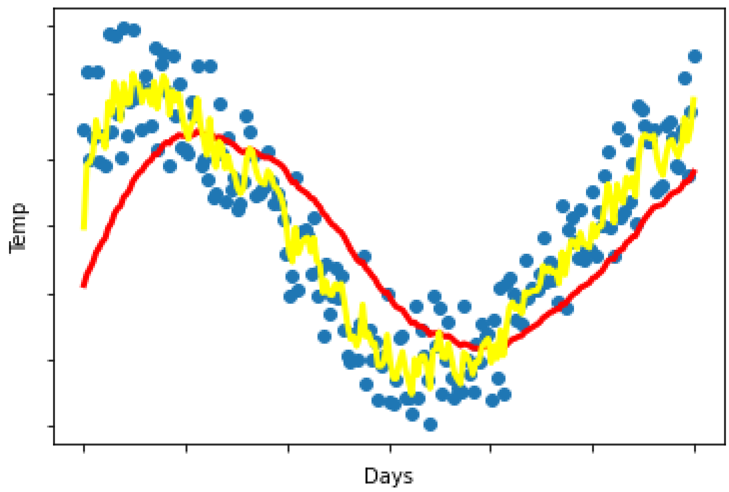
Expand

Correct

Correct. This is not a good learning rate decay since it is an increasing function of �t.

7.
Question 7

You use an exponentially weighted average on the London temperature dataset. You use the following to track the temperature: ��=���−1+(1−�)��$v_t=\beta v_{t-1}+(1-\beta)\theta_t$. The yellow and red lines were computed using values ����1$beta_1$ and ����2$beta_2$ respectively. Which of the following are true?



0 / 1 point

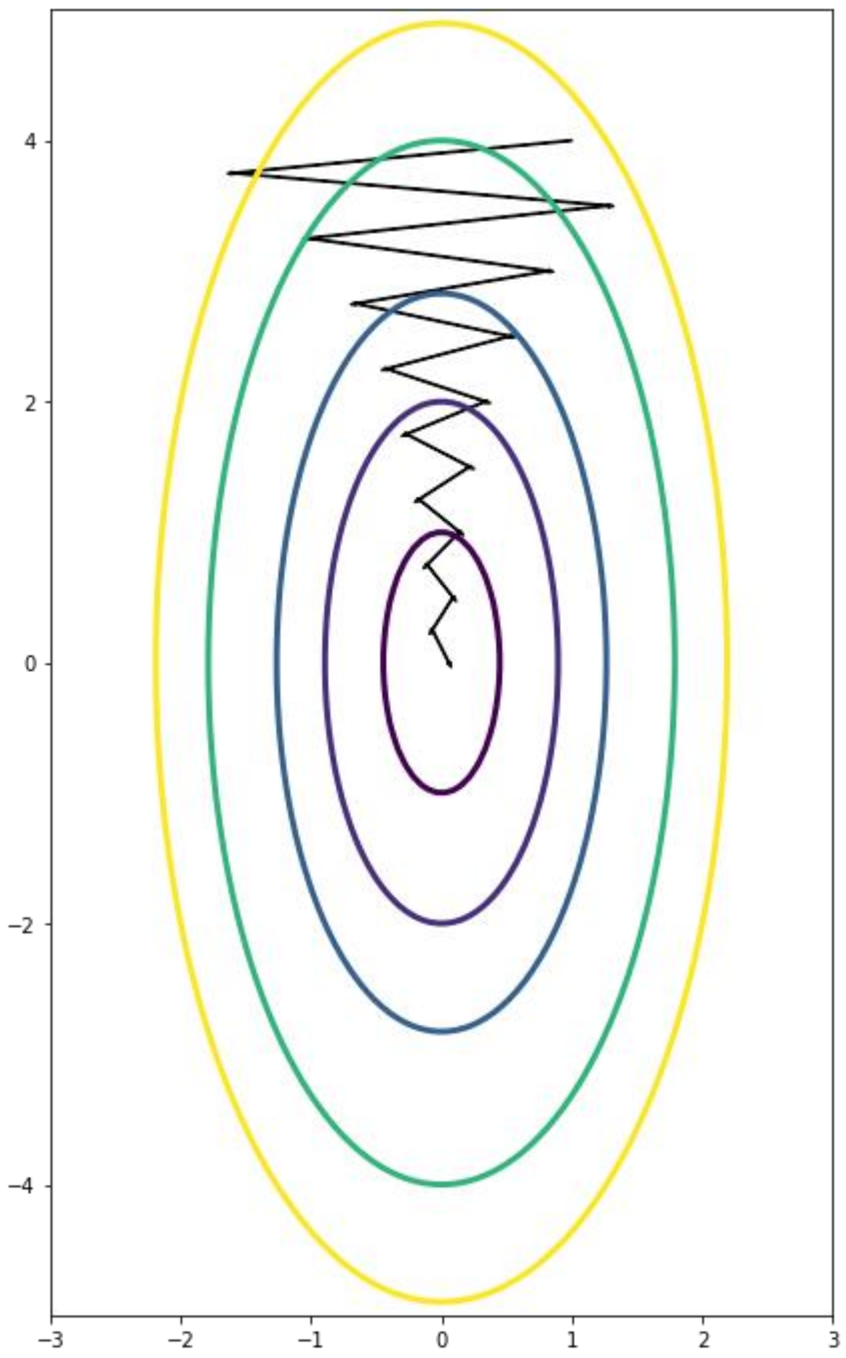
Expand

Incorrect

Incorrect. �1<�2$\beta_1<\beta_2$ since the yellow curve is noisier.

8.
Question 8

Consider the figure:



Suppose this plot was generated with gradient descent with momentum �=0.01$\beta=0.01$. What happens if we increase the value of �β to 0.1$0.1$?

0 / 1 point

Expand

Incorrect

No. 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.
Question 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
�(�[1],�[1],...,�[�],�[�])$J(W_{[1]},b_{[1]},\ldots,W_{[L]},b_{[L]})$. Which of the following techniques could help find parameter values that attain a small value for �J? (Check all that apply)

1 / 1 point

Expand

Correct

Great, you got all the right answers.

10.
Question 10

Which of the following are true about Adam?

0 / 1 point

Expand

Incorrect

False. The mechanics of Adam works the same with the complete batch or with mini-batches.

1.
Question 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

Expand

Correct

2.
Question 2

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

1 / 1 point

Expand

Correct

3.
Question 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

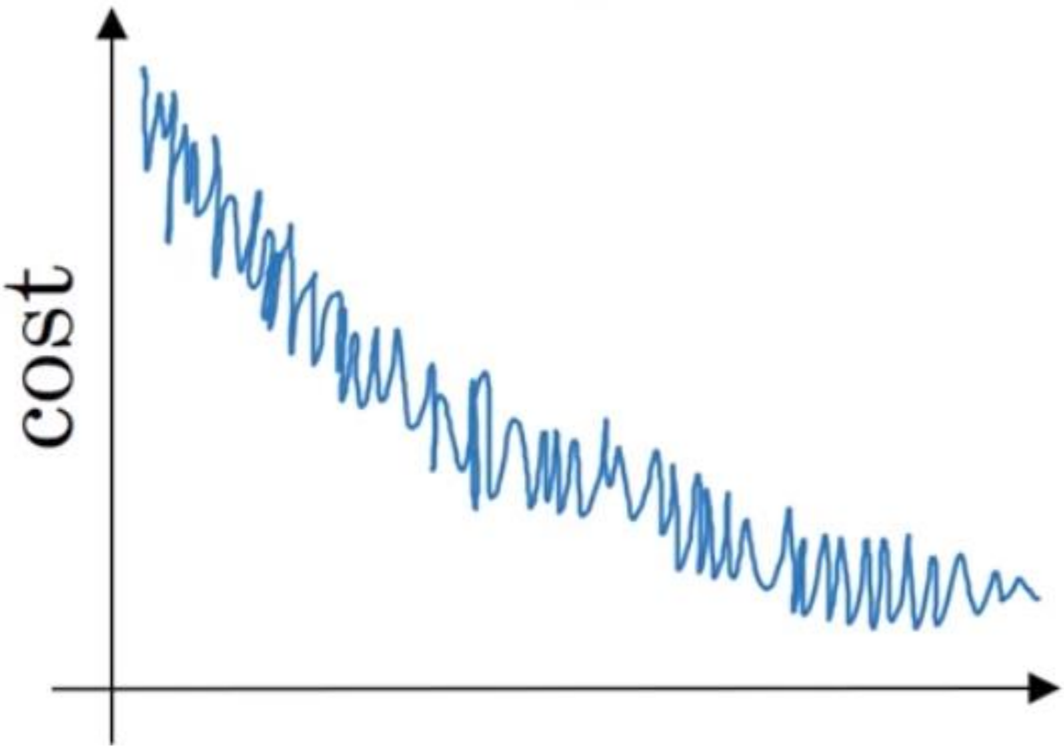
Expand

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.
Question 4

Suppose your learning algorithm's cost �J, plotted as a function of the number of iterations, looks like this:



Which of the following do you agree with?

1 / 1 point

Expand

Correct

5.
Question 5

Suppose the temperature in Casablanca over the first two days of January are the same:

Jan 1st: $\theta_1 = 10^o C$

Jan 2nd: $\theta_2 = 10^o C$

(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.)

1 / 1 point

Expand

Correct

6.
Question 6

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

1 / 1 point

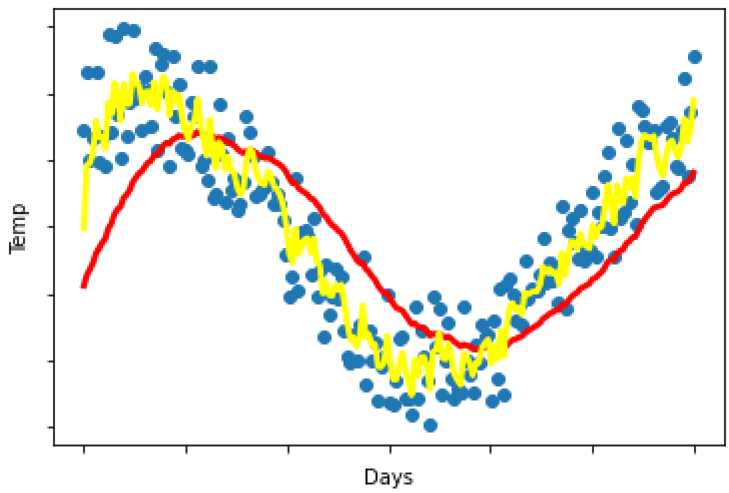
Expand

Correct

Correct. This is not a good learning rate decay since it is an increasing function of t.

7.
Question 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

Expand

Correct

Correct. �1<�2 $\beta_1<\beta_2$ since the yellow curve is noisier.

8.
Question 8

Which of the following are true about gradient descent with momentum?

1 / 1 point

Expand

Correct

Great, you got all the right answers.

9.
Question 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 �(�[1],�[1],...,�[�],�[�]) $J(W^{[1]},b^{[1]},...,W^{[L]},b^{[L]})$. Which of the following techniques could help find parameter values that attain a small value for � J? (Check all that apply)

1 / 1 point

Expand

Correct

Great, you got all the right answers.

10.
Question 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

Expand

Correct

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