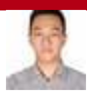
Skip to Main Content

SEARCH IN COURSE

Search

- 1

- 

  Quang Minh Vũ

1. Improving Deep Neural Networks: Hyperparameter Tuning, Regularization and Optimization
2. Week 2
3. Optimization Algorithms

PreviousNext

- **Optimization Algorithms**
- **Lecture Notes (Optional)**
- **Quiz**
  - **Quiz:** Optimization Algorithms

    10 questions
- **Programming Assignment**
- **Heroes of Deep Learning (Optional)**

Optimization Algorithms

Quiz20 minutes • 20 min

Submit your assignment

Due April 25, 1:59 PM +07Apr 25, 1:59 PM +07

Attempts 3 every 8 hours

Try again

Receive grade

To Pass 80% or higher

Your grade

82.50%

View Feedback

We keep your highest score

Like

Dislike

Report an issue

Back

Optimization Algorithms

Graded Quiz • 20 min

DueApr 25, 1:59 PM +07

Congratulations! You passed!

Grade received 82.50%

To pass 80% or higher

Go to next item

Optimization Algorithms

Latest Submission Grade 82.5%

1.

Question 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]\lbrace 7 \rbrace (4)}$a^{[3]\{7\}(4)}$

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

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

Correct

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

2.

Question 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

Mini-Batch Gradient Descent with mini-batch size m/2$m/2$.

Batch Gradient Descent

Stochastic Gradient Descent, Batch Gradient Descent, and Mini-Batch Gradient Descent all make equal use of vectorization.

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.

Question 3

Which of the following is true about batch gradient descent?

1 / 1 point

It has as many mini-batches as examples in the training set.

It is the same as stochastic gradient descent, but we don't use random elements.

It is the same as the mini-batch gradient descent when the mini-batch size is the same as the size of the training set.

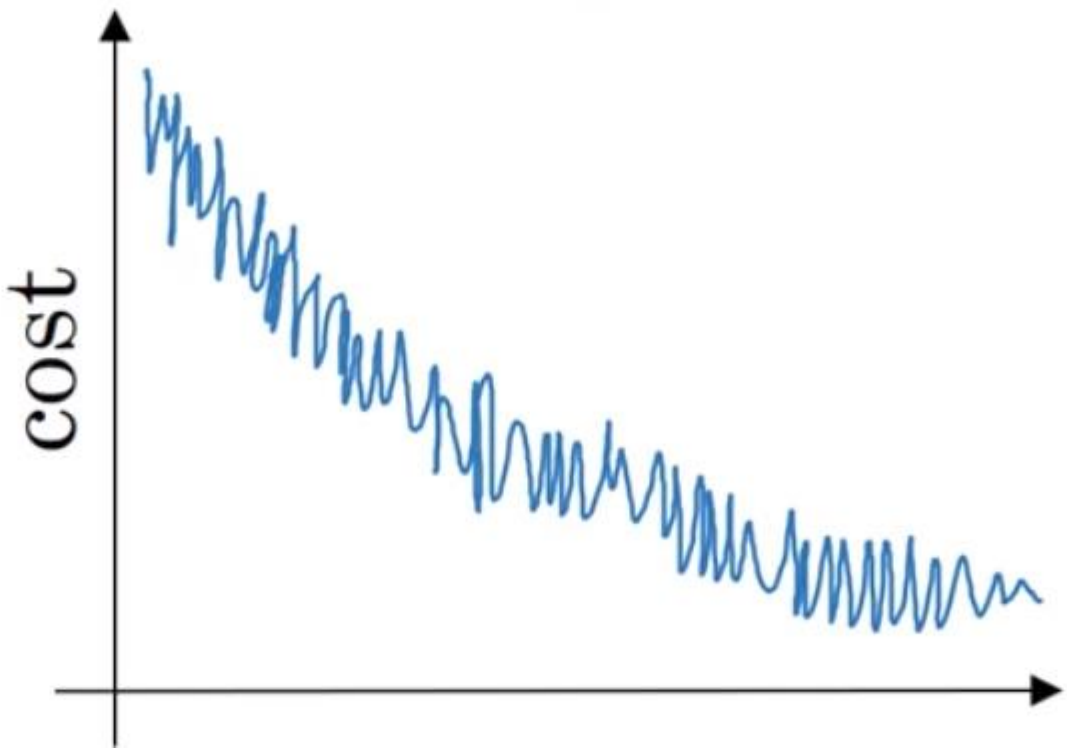
Correct

Correct. When using batch gradient descent there is only one mini-batch thus it is equivalent to batch gradient descent.

4.

Question 4

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



Which of the following do you agree with?

1 / 1 point

○

Whether you’re using batch gradient descent or mini-batch gradient descent, this looks acceptable.

○

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

○

Whether you’re using batch gradient descent or mini-batch gradient descent, something is wrong.

◉

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

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} \text{ C }$\theta_1=10^{\circ}$ C

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

Say you use an exponentially weighted average with \beta = 0.5$\beta=0.5$ to track the temperature: v_0 = 0$v_0=0$, v_t = \beta v_{t-1} + (1- \beta) \, \theta_t$v_t=\beta v_{t-1}+(1-\beta)\theta_t$. If v_2$v_2$ is the value computed after day 2 without bias correction, and v_2^{\text{corrected}}$v_2^{\text{corrected}}$ is the value you compute with bias correction. What are these values?

1 / 1 point

( )

v_2 = 20 $v_2=20$, v_2^{\text{corrected}} = 20 $v_2^{\text{corrected}}=20$.

( )

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

( )

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

(x)

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

Correct

Correct. v_2 = \beta v_{t-1} + (1- \beta) \, \theta_t $v_2=\beta v_{t-1}+(1-\beta)\theta_t$ thus v_1 = 5 $v_1=5$, v_2 = 15 $v_2=15$. Using the bias correction \frac{v_t}{1 - \beta^{t}} $\frac{v_t}{1-\beta^t}$ we get \frac{15}{1 - (0.5)^2} = 20 $\frac{15}{1-(0.5)^2}=20$.

6.

Question 6

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

1 / 1 point

( )

\alpha = 0.95^t \alpha_0 $\alpha=0.95^t\alpha_0$

(x)

\alpha = e^t \alpha_0 $\alpha=e^t\alpha_0$

( )

\alpha = \frac{1}{\sqrt{t}} \alpha_0 $\alpha=\frac{1}{\sqrt{t}}\alpha_0$

( )

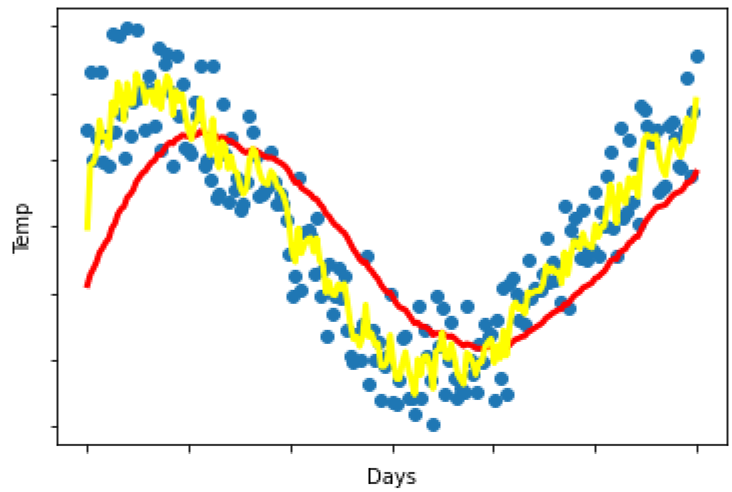
\alpha = \frac{1}{1+2*t} \alpha_0 $\alpha=\frac{1}{1+2*t}\alpha_0$

Correct

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 $v_t=\beta v_{t-1}+(1-\beta)\theta_t$. The yellow and red lines were computed using values beta_1 $beta_1$ and beta_2 $beta_2$ respectively. Which of the following are true?



1 / 1 point

- [ ] $\beta_1 = 0$, $\beta_2 > 0$.
- [ ] $\beta_1 = \beta_2$.
- [ ] $\beta_1 > \beta_2$.
- [x] $\beta_1 < \beta_2$.

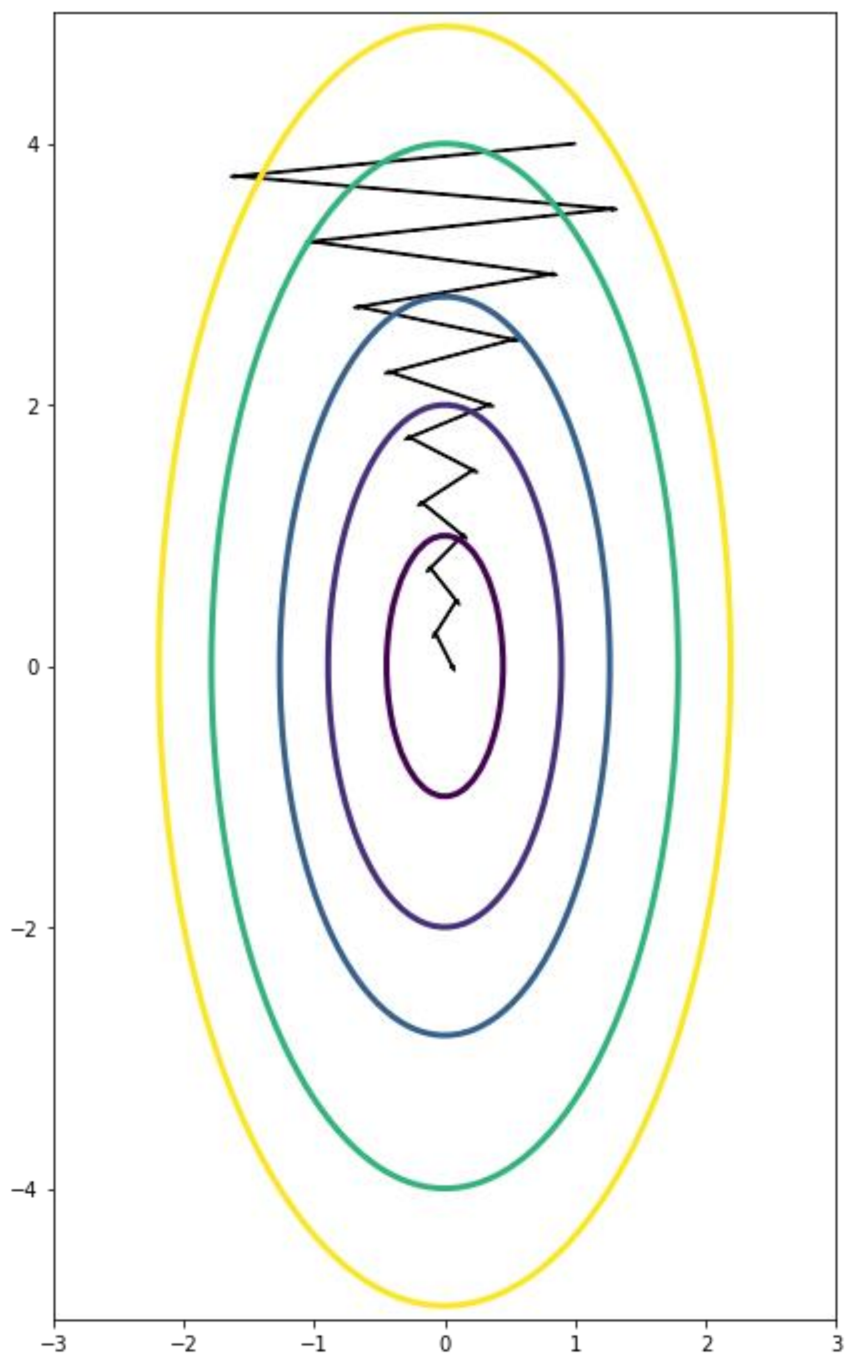
Correct

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

8.

Question 8

Consider the figure:



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

1 / 1 point

- ( ) The gradient descent process starts oscillating in the vertical direction.
- (•) The gradient descent process moves less in the horizontal direction and more 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 more in the horizontal and the vertical axis.

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.

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

0.25 / 1 point

- [x] Try using gradient descent with momentum.

Correct

Yes. The use of momentum can improve the speed of the training. Although other methods might give better results, such as Adam.

- [ ] Normalize the input data.
- [ ] Try better random initialization for the weights
- [x] Add more data to the training set.

This should not be selected

No. This might make the training process take longer.

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?

0 / 1 point

- ( ) False
- (•) True

Incorrect

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