Your grade: 80%

Your latest: **80%** • Your highest: **80%** • To pass you need at least 80%. We keep your highest score.

Next item →

1. Using the notation for mini-batch gradient descent. To what of the following does $a^{[2]\{4\}(3)}$ correspond? **0 / 1 point**

- (●) The activation of the second layer when the input is the fourth example of the third mini-batch.
- ( ) The activation of the fourth layer when the input is the second example of the third mini-batch.
- ( ) The activation of the second layer when the input is the third example of the fourth mini-batch.
- ( ) The activation of the third layer when the input is the fourth example of the second mini-batch.

Expand

Incorrect

No. 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. Suppose you don't face any memory-related problems. Which of the following make more use of vectorization. 1 / 1 point

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

Expand

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. Which of the following is true about batch gradient descent? 1 / 1 point

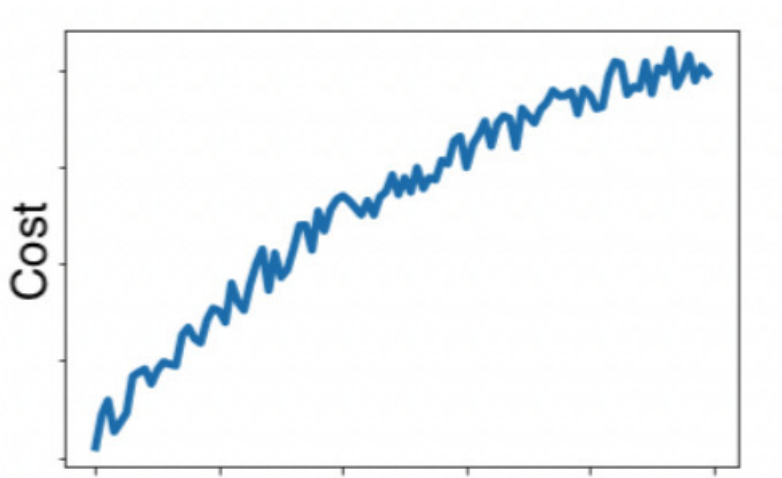
- (●) It is the same as the mini-batch gradient descent when the mini-batch size is the same as the size of the training set.
- ( ) It has as many mini-batches as examples in the training set.
- ( ) It is the same as stochastic gradient descent, but we don't use random elements.

Expand

Correct

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

Which of the following do you agree with?

- ○ 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.
- ○ 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.

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: **1 / 1 point**

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

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

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?

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

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 $\frac{15}{1-0.5^2} = 20$

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. Which of these is NOT a good learning rate decay scheme? Here, t is the epoch number. 1 / 1 point

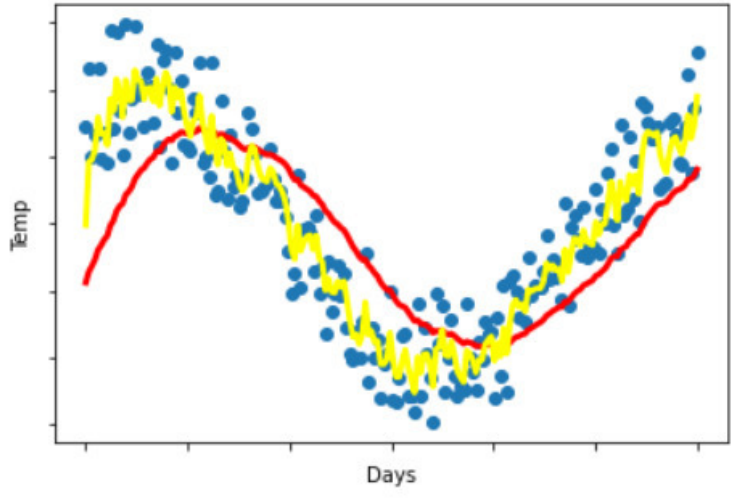
- (selected) $\alpha = e^t \alpha_0$
- $\alpha = \frac{1}{\sqrt{t}} \alpha_0$
- $\alpha = \frac{1}{1 + 2 * t} \alpha_0$
- $\alpha = 0.95^t \alpha_0$

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 yellow and red lines were computed using values $beta_1$ and $beta_2$ respectively. Which of the following are true?

1 / 1 point



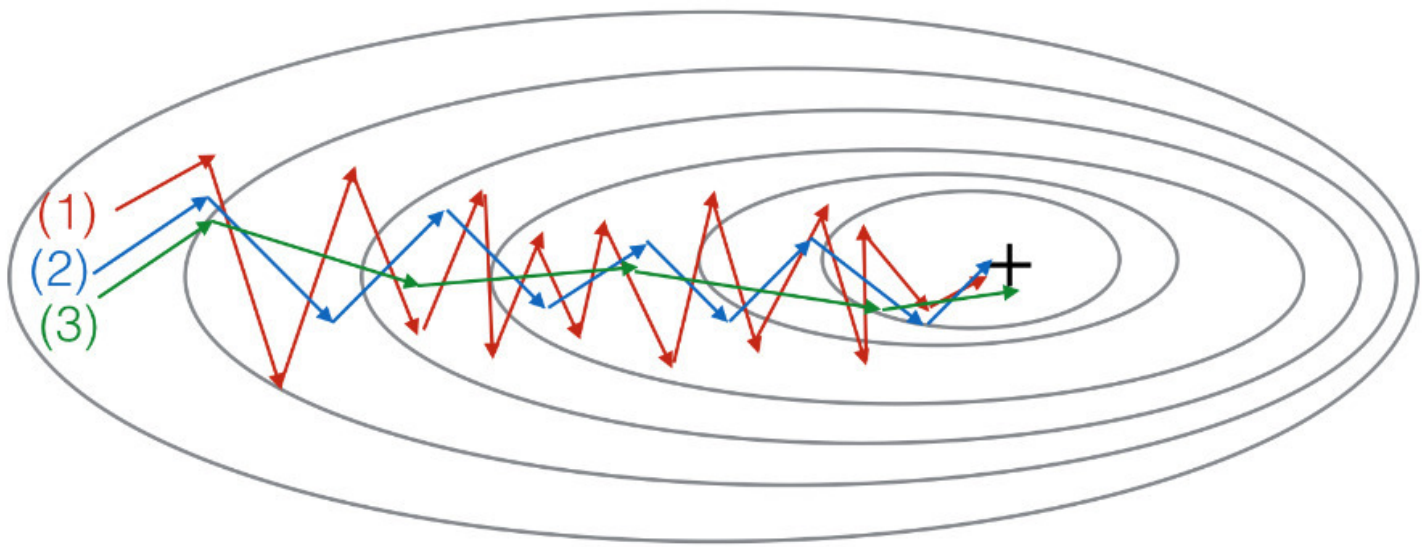
- $\beta_1 > \beta_2$.
- $\beta_1 = 0, \beta_2 > 0$.
- $\beta_1 = \beta_2$.
- $\beta_1 < \beta_2$.

Expand

Correct

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

8. Consider this figure: **1 / 1 point**



These plots were generated with gradient descent; with gradient descent with momentum (β = 0.5); and gradient descent with momentum (β = 0.9). Which curve corresponds to which algorithm?

These plots were generated with gradient descent; with gradient descent with momentum ($\beta = 0.5$); and gradient descent with momentum ($\beta = 0.9$). Which curve corresponds to which algorithm?

- (1) is gradient descent. (2) is gradient descent with momentum (small β). (3) is gradient descent with momentum (large β)
- (1) is gradient descent. (2) is gradient descent with momentum (large β). (3) is gradient descent with momentum (small β)
- (1) is gradient descent with momentum (small β), (2) is gradient descent with momentum (small β), (3) is gradient descent
- (1) is gradient descent with momentum (small β). (2) is gradient descent. (3) is gradient descent with momentum (large β)

Expand

Correct

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

- [ ] Try initializing the weight at zero.
- [x] Normalize the input data.

  ✓ **Correct**

  Yes. In some cases, if the scale of the features is very different, normalizing the input data will speed up the training process.

- [x] Try using Adam.

  ✓ **Correct**

  Yes. Adam combines the advantages of other methods to accelerate the convergence of the gradient descent.

- [x] Try mini-batch gradient descent.

  ✓ **Correct**

  Yes. Mini-batch gradient descent is faster than batch gradient descent.

Expand

Correct
Great, you got all the right answers.

10. Which of the following are true about Adam? **1 / 1 point**

- Adam can only be used with batch gradient descent and not with mini-batch gradient descent.
- The most important hyperparameter on Adam is ϵ and should be carefully tuned.
- Adam automatically tunes the hyperparameter α.
- Adam combines the advantages of RMSProp and momentum.

Expand

Correct
True. Precisely Adam combines the features of RMSProp and momentum that is why we use two-parameter β_1 and β_2, besides ϵ.