×

Lecture 6 Quiz

返回第 6 周课程

✓ **5/5** 得分（100%）

未认证 本次测验未认证。您必须以认证方式重新完成测验，才有资格获得课程证书。

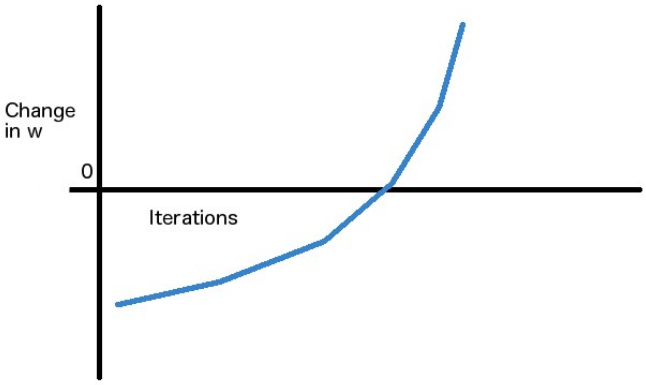
✓ 1 / 1 分

1.

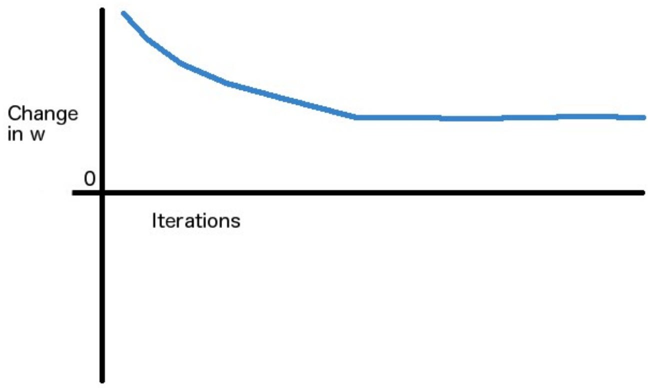
Suppose w is the weight on some connection in a neural network. The network is trained using gradient descent until the learning *converges*. We plot the change of w as training progresses. Which of the following scenarios shows that convergence has occurred? **Notice that we're plotting the change in w , as opposed to w itself.**

Note that in the plots below, each *iteration* refers to a single *step* of steepest descent on a *single minibatch*.

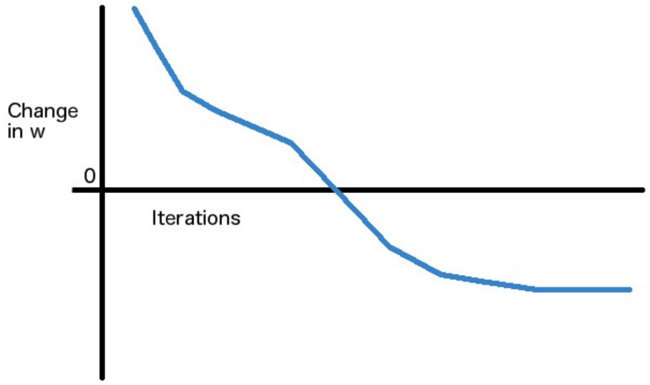
○



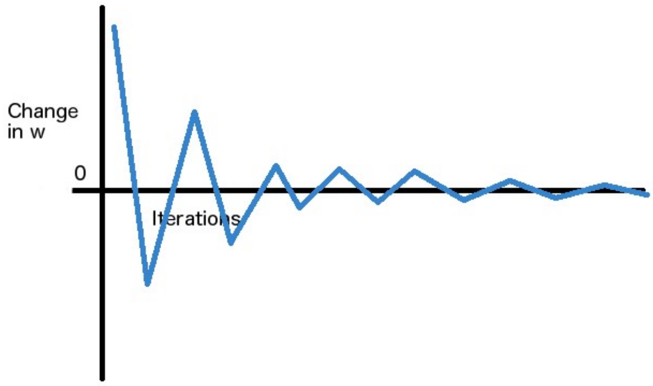
○



○



●



正确回答

If the optimization has converged, w must converge to (or at most oscillate around) a point. So the change in w must converge to (or oscillate around) zero.

✓ 1 / 1 分

2.

Suppose you are using mini-batch gradient descent for training some neural net on a large dataset. You have to decide on the learning rate, weight initializations, preprocess the inputs etc. You try some values for these and find that the value of the objective function on the training set decreases smoothly but very slowly. What could be causing this? Check all that apply.

☑ The inputs might have a very large scale (hint: think of what this would do to the logistic hidden units).

正确回答
Large values of inputs may saturate the hidden units. Their derivatives would become small (be on a "plateau") and learning would get slowed down.

□ The minibatch size is too small.

正确回答
Small mini-batches will cause noisy gradients which will show up as erratic changes in the error function. So the convergence will not be smooth.

☑ The learning rate may be too small.

正确回答
A small learning rate leads to small changes in the parameters, and to slow convergence.

☑ The weights might have been initialized to very large values (hint: think of what this would do to the logistic hidden units).

正确回答
Large values of weights may saturate the hidden units. Their derivatives would become small (be on a "plateau") and learning would get slowed down.

✓ 1 / 1 分

3.

Full-batch gradient descent can be used to minimize an objective function if the dataset is not too large. Which statement regarding full-batch gradient descent is **false**?

○ Using momentum can be useful for full batch gradient descent.

○ For some setting of the learning rate, it is possible that the objective function increases in some iteration.

○ Adaptive learning rate methods perform well for full-batch (or large mini-batch) gradient descent.

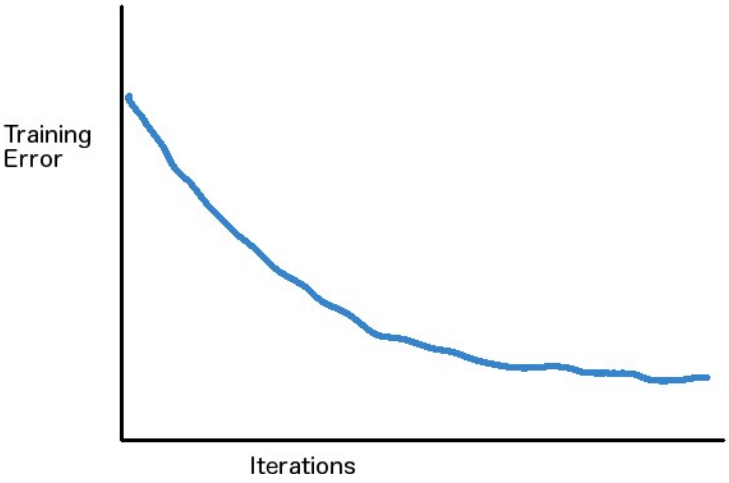
● Full batch gradient descent is guaranteed to find a better local minimum than mini-batch gradient descent.

正确回答
This is not necessarily true because mini-batch gradient descent can search through weight space due to noise and potentially escape bad local minima.

✓ 1 / 1 分

4.

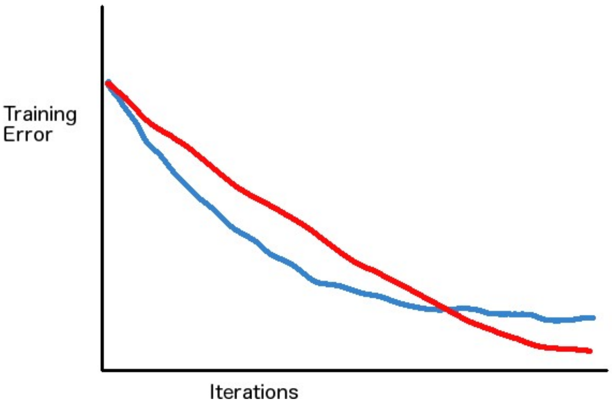
Claire is training a neural net using mini-batch gradient descent. She chose a particular learning rate and found that the training error decreased as more iterations of training were performed as shown here in blue



She was not sure if this was the best she could do. So she tried a **smaller** learning rate. Which of the following error curves (shown in red) might she observe now? Select the two most likely plots.

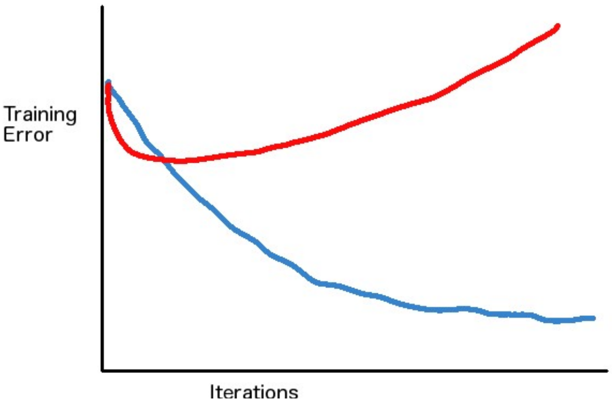
Note that in the plots below, each *iteration* refers to a single *step* of steepest descent on a *single minibatch*.

☑



正确回答
A smaller learning rate may lead to slower progress initially but result in a lower final error.

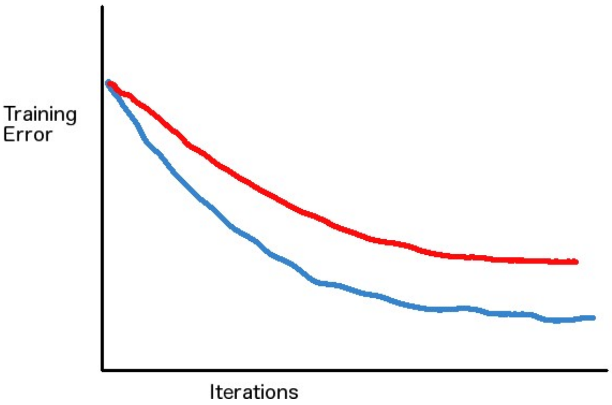
□



正确回答

The error is less likely to diverge if a smaller learning rate is used.

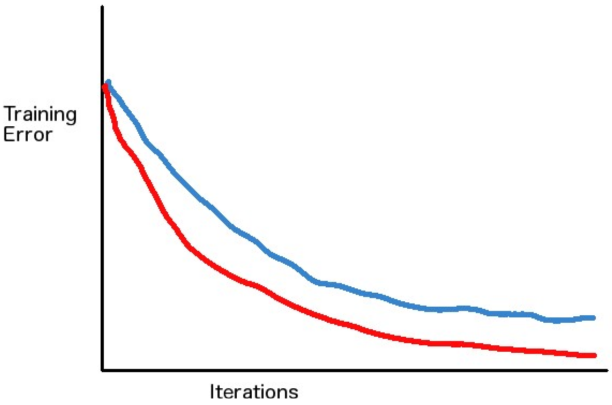
☑



正确回答

A smaller learning rate may lead to slower convergence.

☐



正确回答

The model will usually not make faster progress initially if a smaller learning rate is used.

✓ 1 / 1 分

5.

In the lectures, we discussed two kinds of gradient descent algorithms: mini-batch and full-batch. For which of the following problems is mini-batch gradient descent likely to be **a lot better** than full-batch gradient descent?

☑ Object detection: Identify which of 1000 categories an object image belongs to, given 10 million 256 X 256 pixel images.

正确回答

□ Sentiment Analysis: Decide whether a given movie review says that the movie is 'good' or 'bad'. The input consists of the word count in the review, for each of 50,000 words. The training set consists of 100 movie reviews written by experts for a newspaper.

正确回答

☑ Speech recognition: Identify which of 40 phonemes is being pronounced in a 10-millisecond window of speech sound. The training data consists of 50,000 hours of speech data (this is more than 10 billion 1800 dimensional training points)

正确回答

□ Disease prediction: Predict if a person will get cancer. The input consists of 1000 medical indicators (blood pressure, family cancer history, etc.); the training set consists of 100 patients who all suffered the same type of cancer, and 100 healthy patients.

正确回答

👍 👎 ⚑