Feedback — Lecture 6 Quiz

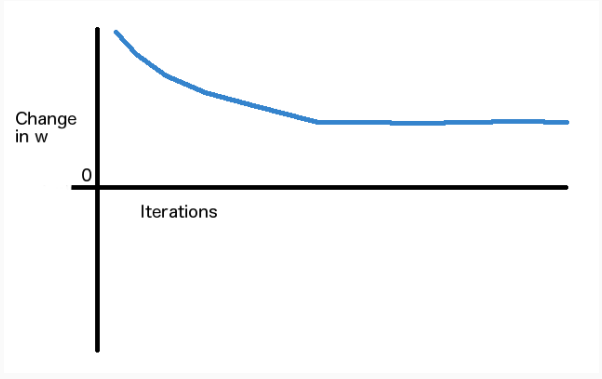
Help Center

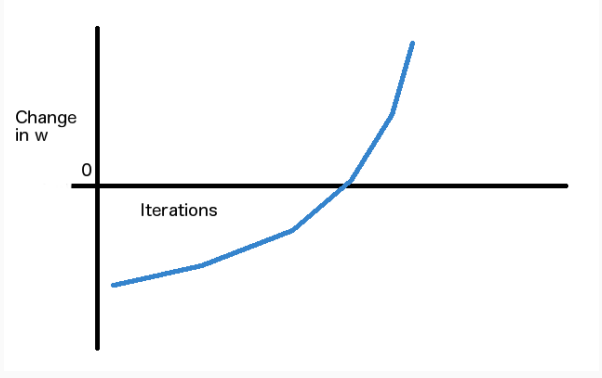
You submitted this quiz on **Thu 30 Jun 2016 1:29 PM CEST**. You got a score of **6.00** out of **8.00**. However, you will not get credit for it, since it was submitted past the deadline.

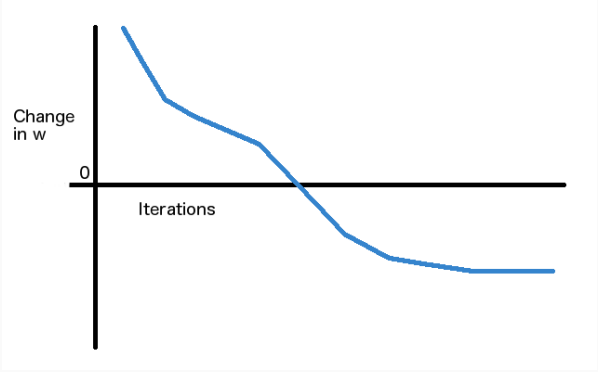
Question 1

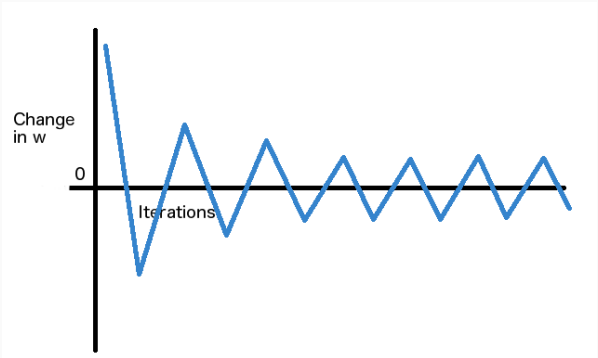
Suppose w is the weight on some connection in a neural network. The network is trained using gradient descent until the learning **converges**. However, the dataset consists of two mini-batches, which differ from each other somewhat. As usual, we alternate between the mini-batches for our gradient calculations, and that has implications for what happens after convergence. 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*.

Your Answer	Score	Explanation









✔ 1.00

Total 1.00 / 1.00

Question Explanation

If the optimization has converged, w must converge to, or oscillate around, a point. So the change in w must converge to (or oscillate around) zero. The two mini-batches, however, will produce different gradients, one the opposite of the other. That causes w to go back and forth.

Question 2

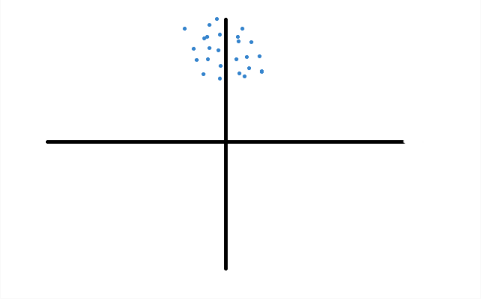
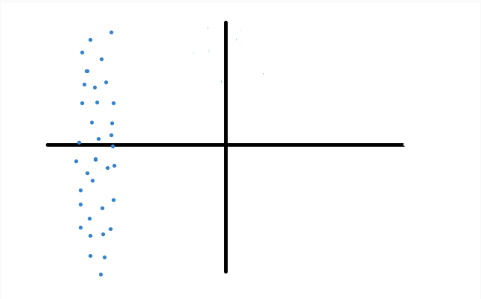
Suppose you are using mini-batch gradient descent for training some neural net on a large dataset. You have to decide on the learning rate, weight initializations, preprocess the inputs etc. You try some values for these and find that the value of the objective function on the training set decreases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

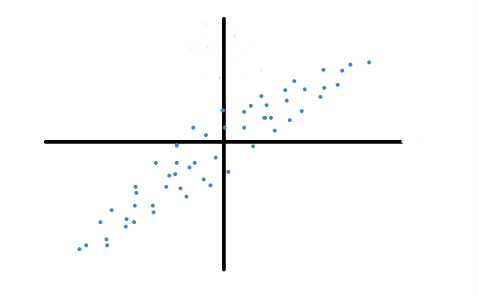
Your Answer		Score	Explanation
☐ The weights might have been initialized to very large values (hint: think of what this would do to the logistic hidden units).	✖	0.00	Large values of weights may saturate the hidden units. Their derivatives would become small (be on a "plateau") and learning would get slowed down.
☐ The minibatch size is too small.	✔	0.50	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.

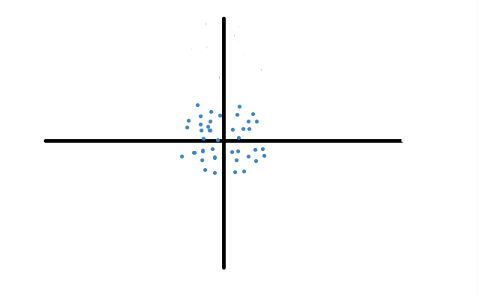
☐ The inputs might have a very large scale (hint: think of what this would do to the logistic hidden units).	✖	0.00	Large values of inputs may saturate the hidden units. Their derivatives would become small (be on a "plateau") and learning would get slowed down.
☑ The learning rate may be too small.	✔	0.50	A small learning rate leads to small changes in the parameters, and to slow convergence.
Total		1.00 / 2.00	

Question 3

Four datasets are shown below. Each dataset has two input values (plotted below) and a target value (not shown). Each point in the plots denotes one training case. Assume that we are solving a classification problem. Which of the following datasets would most likely be easiest to train using neural nets ?

Your Answer	Score	Explanation
○ 		
○ 		
○		





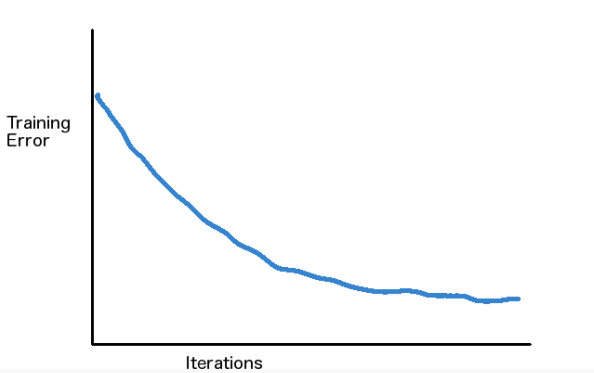
✔ 1.00	This dataset has mean zero and diagonal covariance.	
Total	1.00 / 1.00	

Question Explanation

It is preferable to have data sets with zero mean and diagonal covariance.

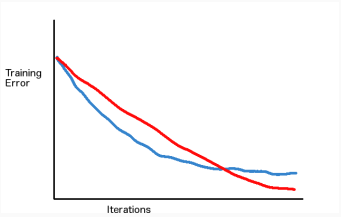
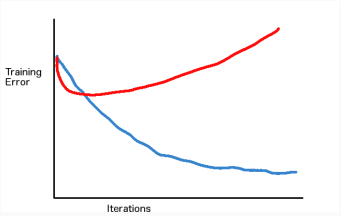
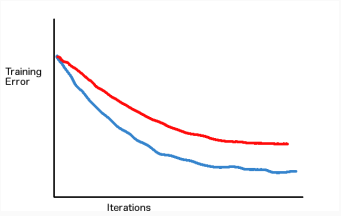
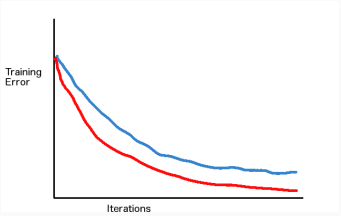
Question 4

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 **bigger** 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*.

Your Answer	Score	Explanation

☐ 	✔	0.50	A larger learning rate will usually not lead to slower convergence during the initial iterations of learning.
☑ 	✔	0.50	Large learning rate may lead to divergence and errors can blow up.
☐ 	✔	0.50	A larger learning rate will usually not lead to slower convergence during the initial iterations of learning.
☑ 	✔	0.50	Large learning rate may show better final performance.
Total		2.00 / 2.00	

Question 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?

Your Answer		Score	Explanation
☑ 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 1,000,000 movie reviews found on the internet.	✔	0.50	
☐ Language modeling: Predict the next word using the previous 3 words. The vocabulary consists of 100,000 words. The dataset	✖	0.00	

consists of all Wikipedia articles.		
☑ 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.	✖	0.00
☐ Predict if an experiment at the Large Hadron Collider is going to yield positive results. The input consists of 25 experiment parameters (energy level, types of particles, etc). The training set consists of the 200 experiments that have already been completed (some of those yielded positive results; some yielded only negative results).	✔	0.50
Total		1.00 / 2.00