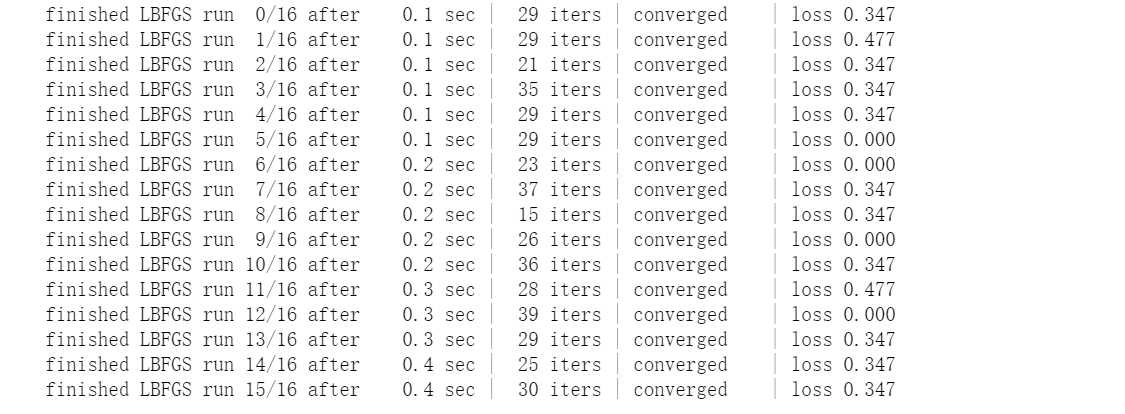
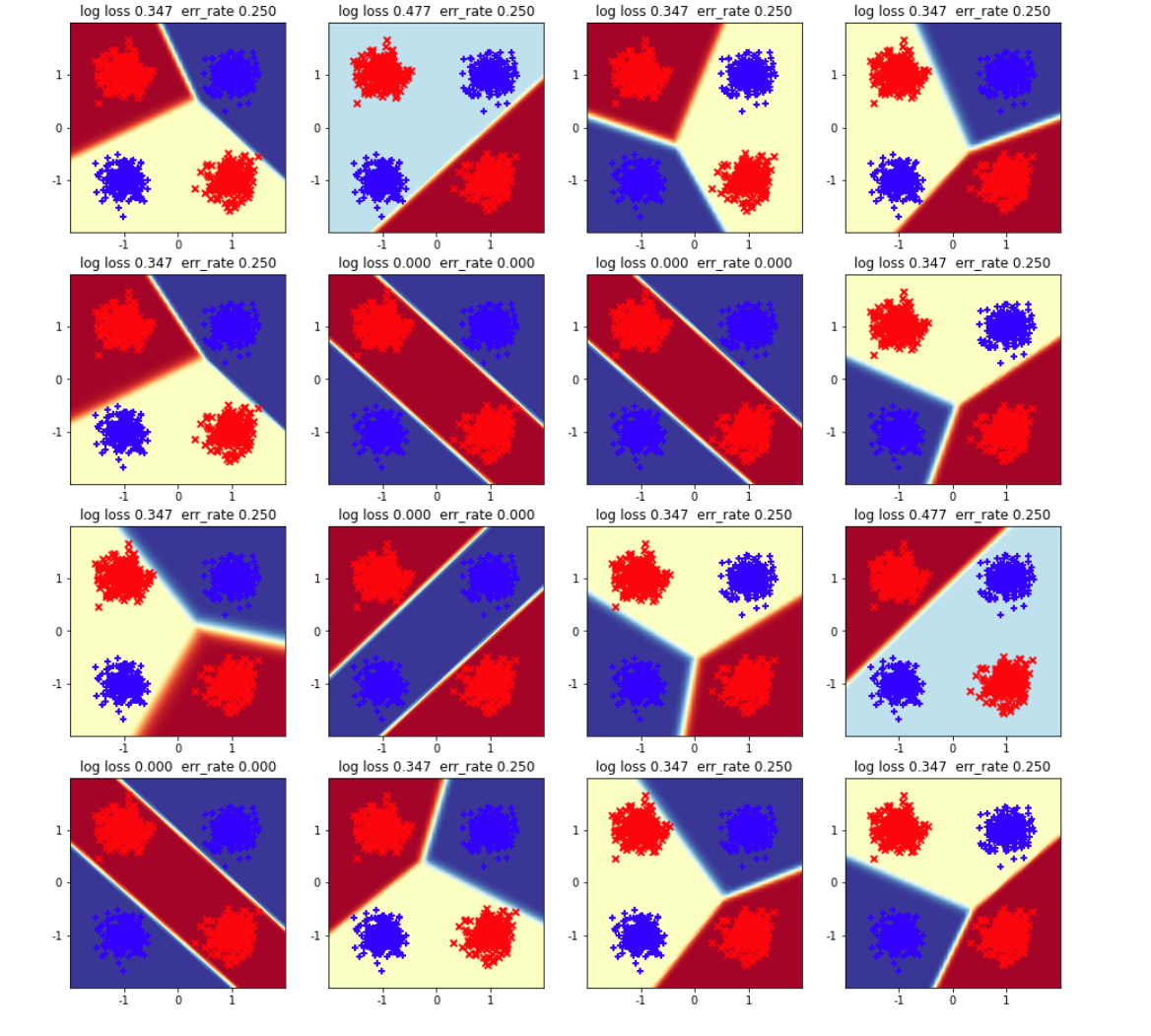
**Question 1**



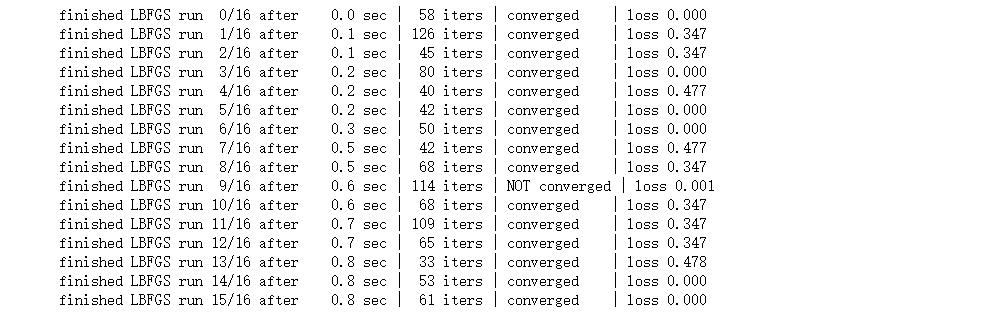
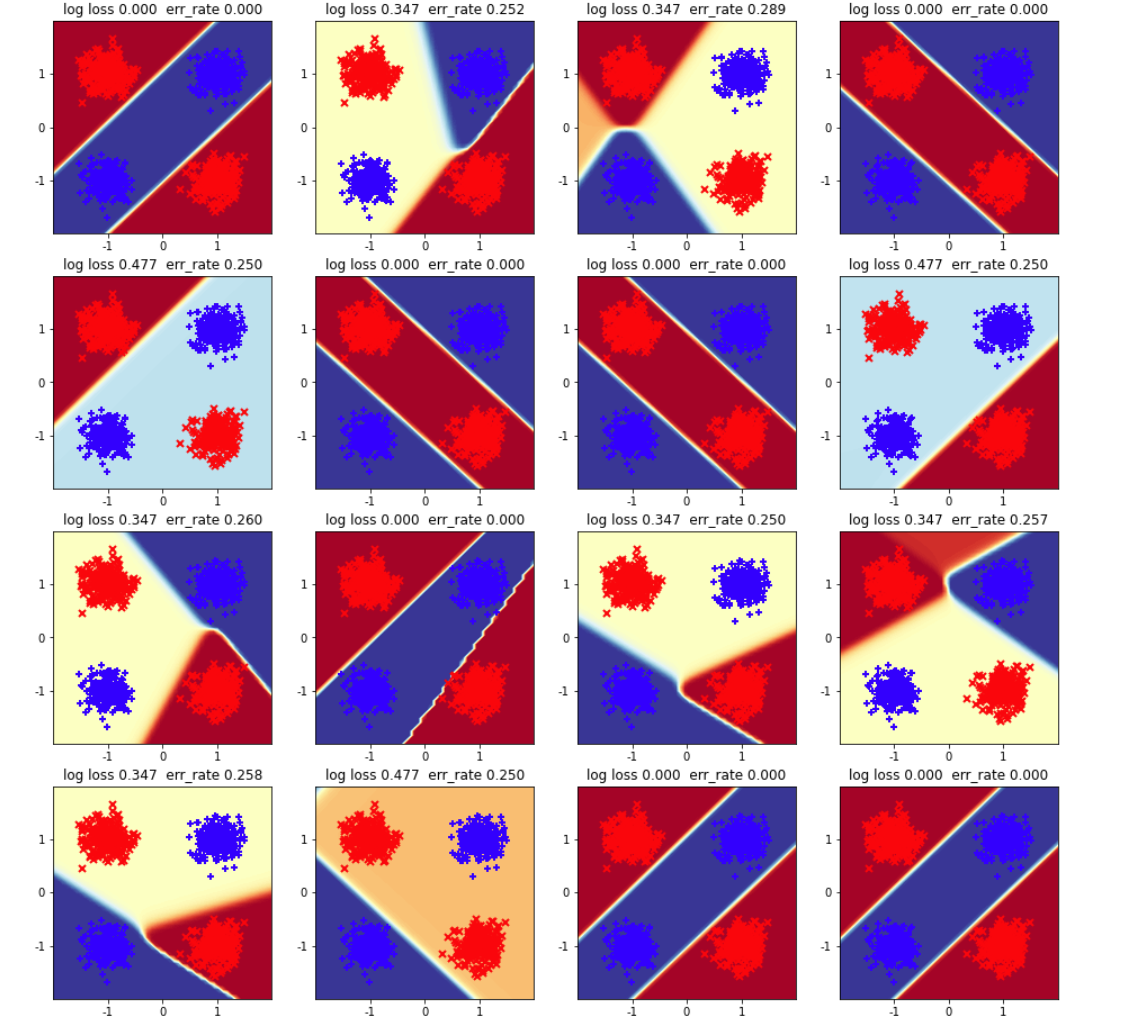


**B**

\*\*Answer\*\*: There is 1/4 reach the 0 trainning error. The remainings' error rate is all 0.25. All of them in the 16 iterations are converged and they are averaged converged in 0.025 seconds. Overall, it converged really fast.

I think the reason may be for the gradient, there are some local minimum and absolute minimum. Therefore, the error rate may vary, and when the gradient reaches the 0, this is the absolute minimum.

**Question 2**

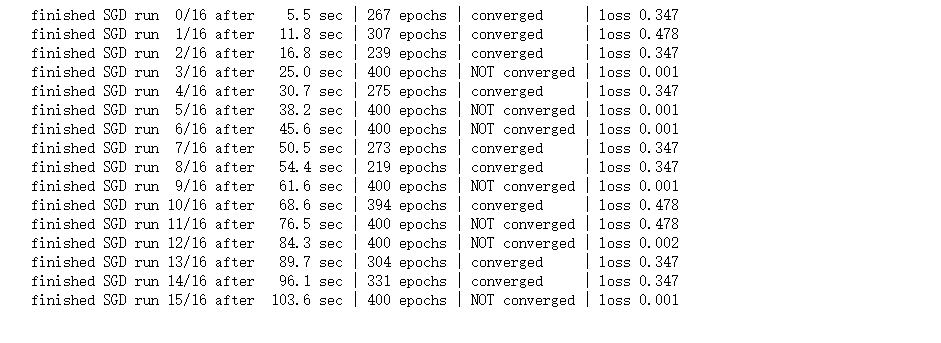
 

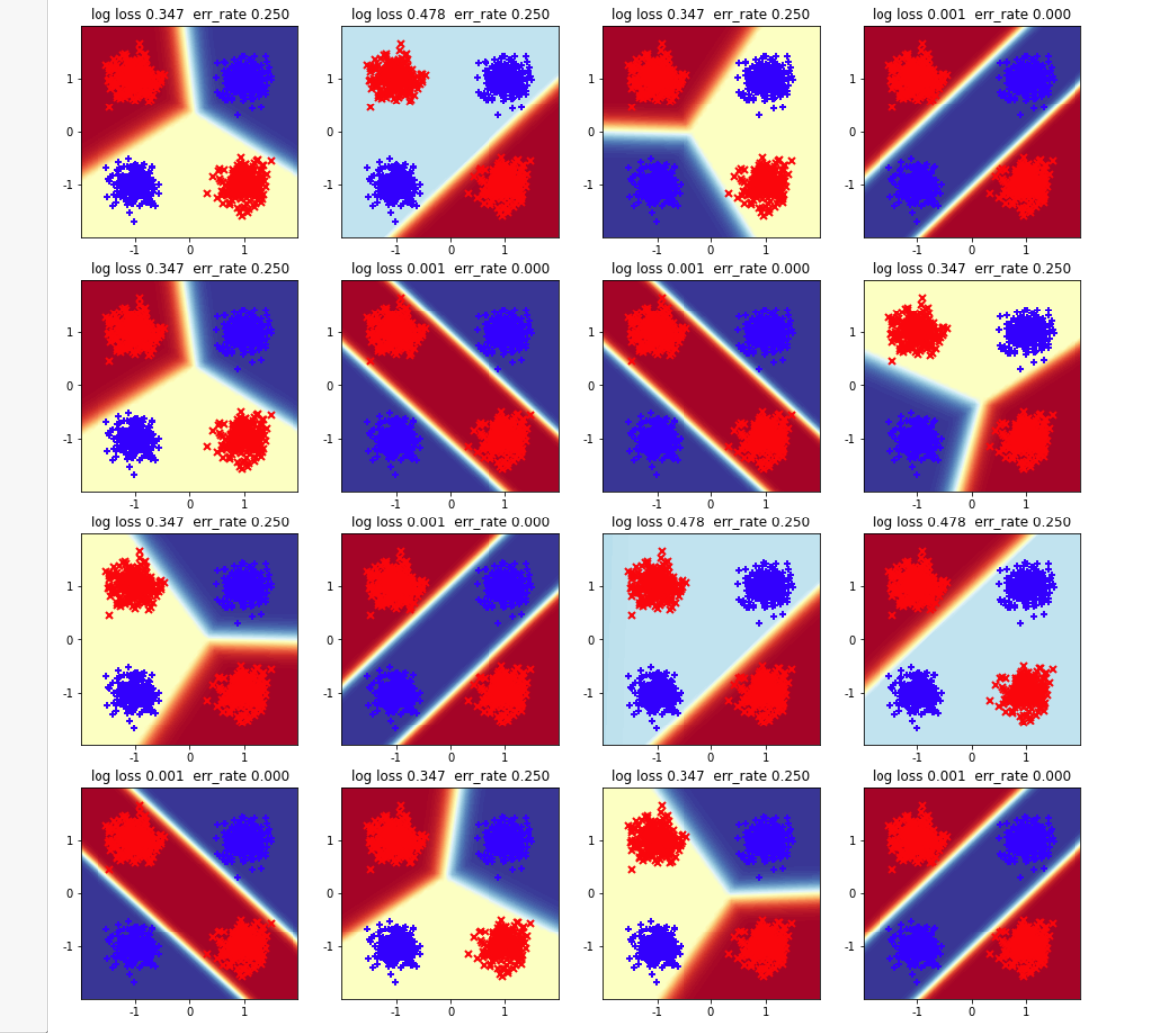
**B**

\*\*Answer\*\*: 7/16's runs reach 0 training error. Others are around 0.25 to 0.361. All of them in the 16 iterations are converged and they are averaged converged in 0.05 seconds. Overall, it converged really fast.

I think the reason may be for the gradient, there are some local minimum and absolute minimum. Therefore, the error rate may vary, and when the gradient reaches the 0, this is the absolute minimum.

**Question 3**





**B**

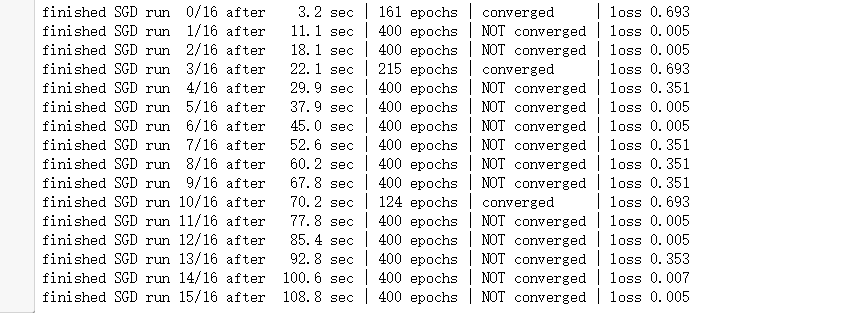
\*\*Answer\*\*: 3/8 of runs reach 0 training error, and others all have the same error rate, which is 0.25. In this time, only 9/16 of runs finally converged. For the converged one, the relatively fast converge is 4 seconds while the slow converge can be nearly 9 seconds. On the other hand, for the non-converged one, the speed is simliar.’

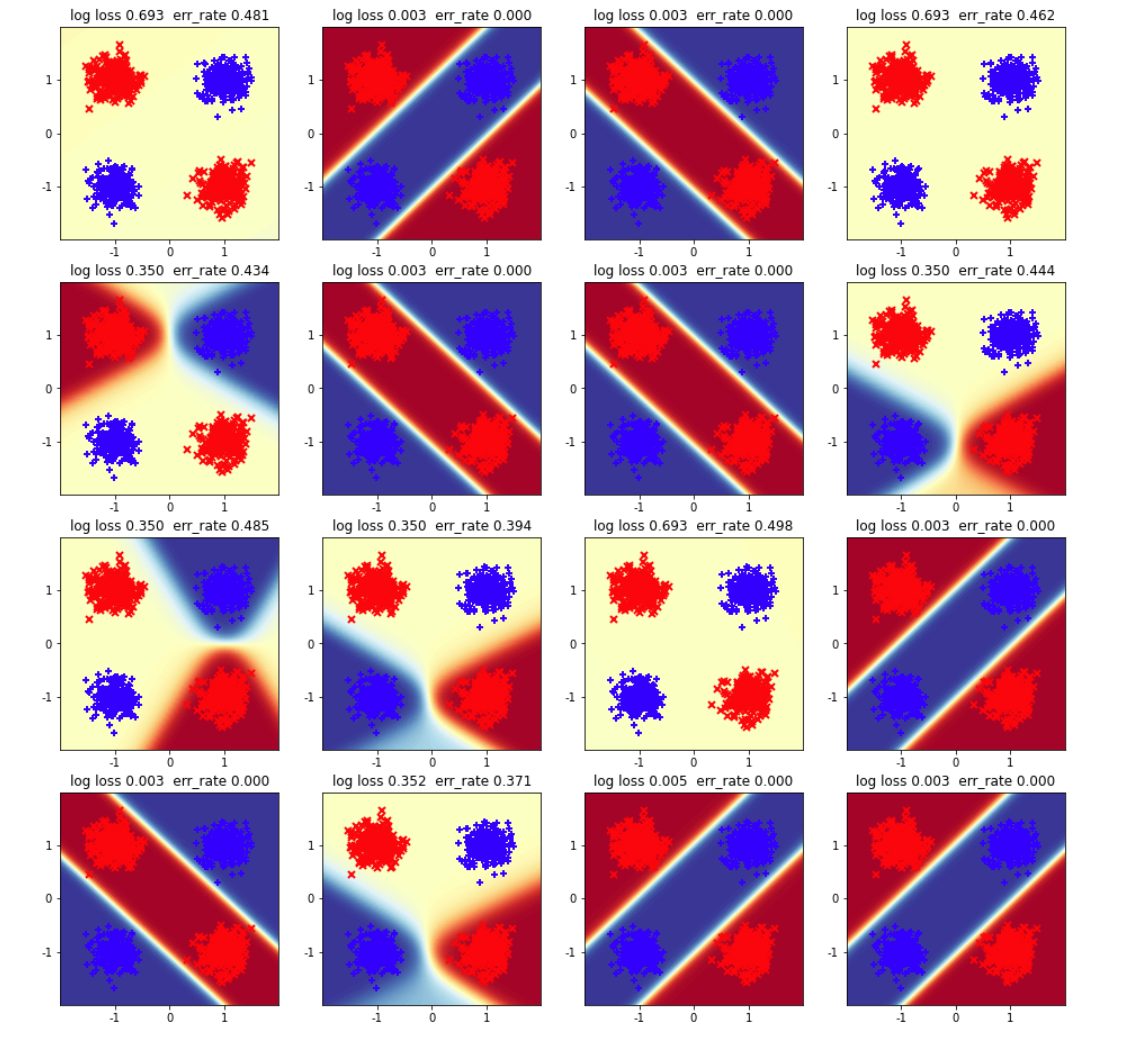
**C**

\*\*Answer\*\*: The SGD solver takes much longer time to calculate, and just about half of the runs successfully converge. I think this may because the SGD does not do the regularization to its input, which may cause it to be slow.

In addition, there is some local minimum, which makes the SGD think that it is already get the best solution. Therefore, the SGD cannot learn well.

**Question 4**





**B**

\*\*Answer\*\*: 1/2 of runs reach 0 training error, and the others are about 0.371 to 0.498. In this time, only 3/16 of runs finally converged. For the converged one, the relatively fast converge is 3 seconds while the slow converge can be nearly 6 seconds.

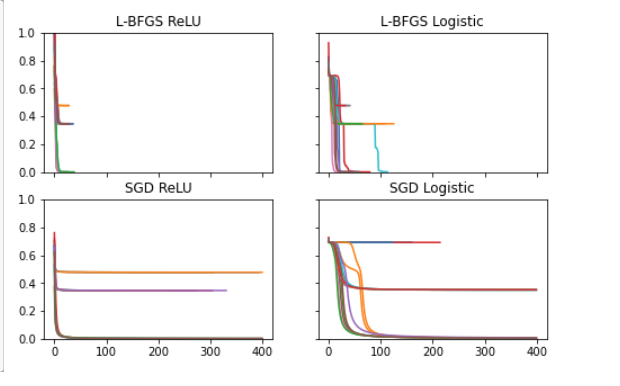
On the other hand, for the non-converged one, the speed is simliar.

**C**

\*\*Answer\*\*: The SGD solver takes much longer time to calculate, and just little runs successfully converge. I think this may because the SGD does not do the regularization to its input, which may cause it to be slow.

In addition, there is some local minimum, which makes the SGD think that it is already get the best solution. Therefore, the SGD cannot learn well.

**Question 5**



**B**

\*\*Answer\*\* All ReLU seems easier to optimize, and all Logistic Sigmoid requires more iterations.

**C**

\*\*Answer\*\*: 1. compare two functions with other different classfications

2. compare two functions with other different solvers

3.compare two functions with other different hypeparameters

(All do the compare experiment for the same data)