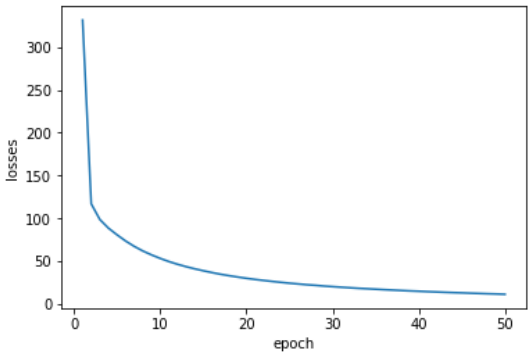
Name: Junzhe Wu NetID: junzhew3

1. Test accuracy: 0.9772
2. I use one hidden layer and the number of hidden unit is 128.
3. The learning rate is 0.1 and it does not decrease, the number of epoch is 50, and the batch size is 200. The figure of losses change with epoch is shown below.



4. The W1 and W2 matrix are randomly initialized with a normalization constant 0.01. The b1 and b2 matrix are initialized to 0.
5. I use the minibatch stochastic gradient descent method and the batch size is 200. I shuffle the data each time.
6. I have written code for a really similar problem in the CS440 last semester, so I use this code as an example. Of course I make some changes since the method of gradient descent and the number of layers are different.