(1)

(a) In stochastic gradient descent (SGD), we don't go through the whole dataset in one update step. In other words, the loss function should not have a sum over N, but rather over a mini-batch of 16, 32, 64, 128 data points. That's because with big datasets and complex models, computing gradients in that way would be computationally expensive and even not feasible.

(b) The code and intermediate results for this part are given in the next page.

(c) The code and intermediate results are given in the next pages.

→ Also Adagrad helps in this algorithm. We can see that by comparing losses.