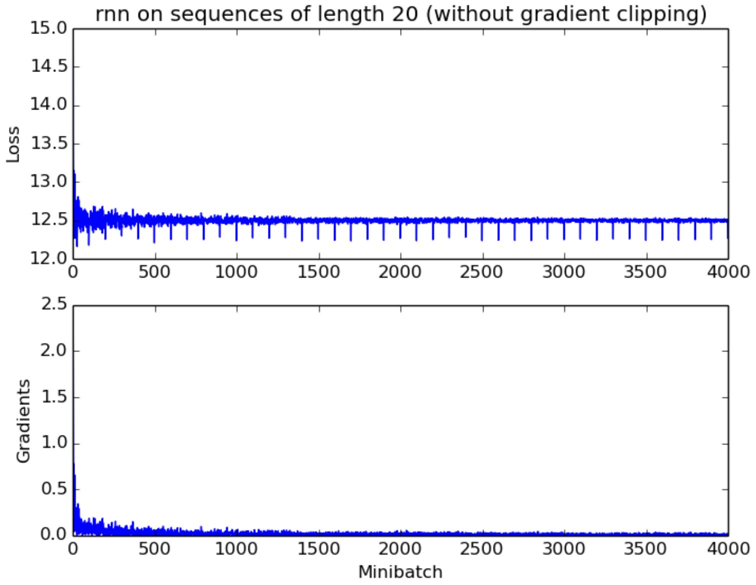
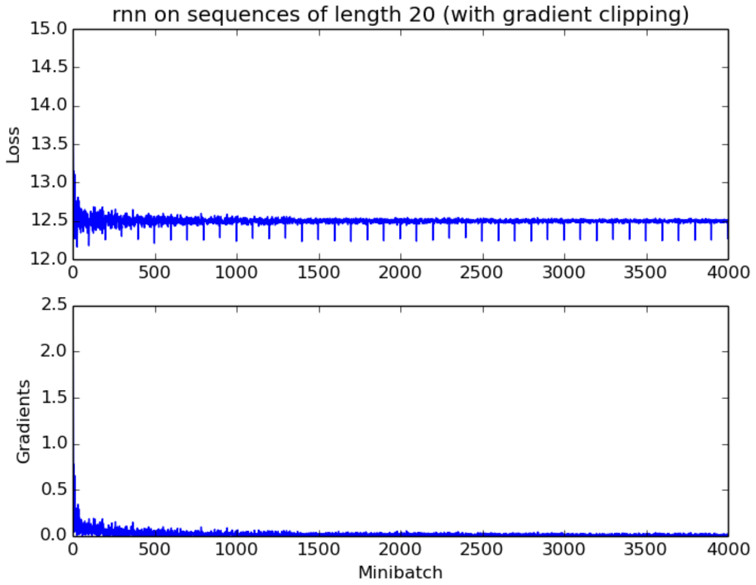
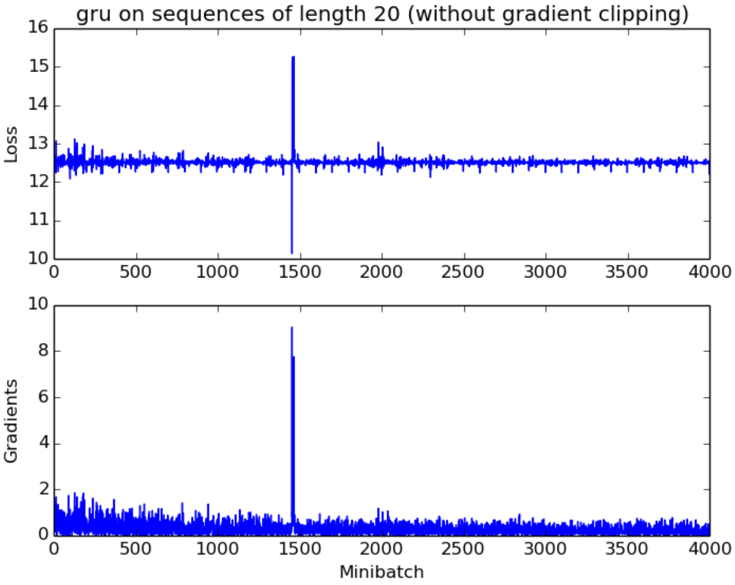
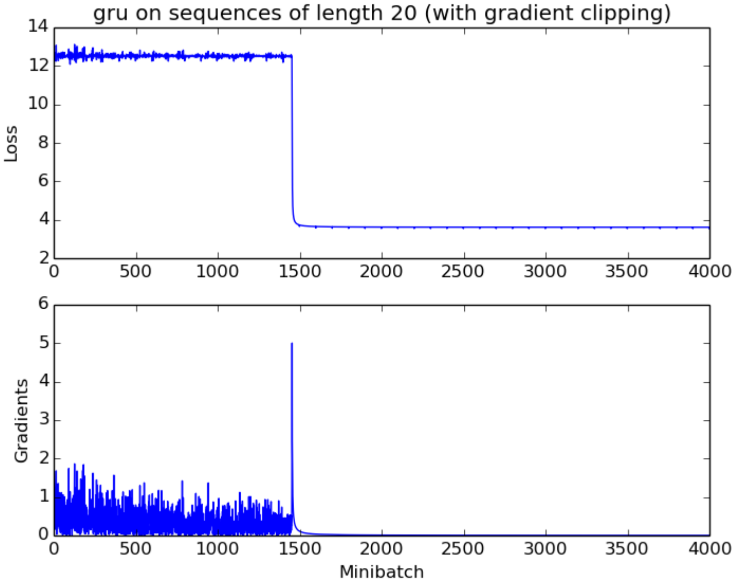
Question 3.d)


rnn on sequences of length 20 (without gradient clipping)
Loss
15.0
14.5
14.0
13.5
13.0
12.5
12.0
0
500
1000
1500
2000
2500
3000
3500
4000
Gradients
2.5
2.0
1.5
1.0
0.5
0.0
0
500
1000
1500
2000
2500
3000
3500
4000
Minibatch


rnn on sequences of length 20 (with gradient clipping)
Loss
15.0
14.5
14.0
13.5
13.0
12.5
12.0
0
500
1000
1500
2000
2500
3000
3500
4000
Gradients
2.5
2.0
1.5
1.0
0.5
0.0
0
500
1000
1500
2000
2500
3000
3500
4000
Minibatch



gru on sequences of length 20 (without gradient clipping)



gru on sequences of length 20 (with gradient clipping)

Question 3.e)i.

It seems like the RNN model experience vanishing gradients, since the gradients become very small and the model get stuck in some local minima. On the other hand, the GRU model seems to experience exploding gradient, since at some point it reaches a loss significantly better than before, but a large gradient pushes it to a worse local minimia region. Clipping does help, as we can see from the last pair of graphs: after it reaches this better loss, the clipped gradient is not large enough to push it to the worse local minima, and then we stay in the same region, eventually converging to a significantly better loss.

Question 3.e)ii.

GRU does better since it is able to achieve a much lower loss. This doesn't necessarily guarantees a better generalization error, but the fit is definitely better, we achieve a much lower error and the model seems to allow for better accuracy.