Exercise 10

Kai Schneider

July 6, 2021

Task 1 DQN on the Cart-Pole

a.)

- NNs tend to diverge (due to high correlation states and actions, NOT i.i.d)
  $\rightarrow$ Experience Replay (sampling training data from buffer of past experiences)
- oscillation due to big changes in Q-values
  $\rightarrow$ much lower learning rate than in other DL-scenarios
  $\rightarrow$ reward clipping, normalize $[0, 1]$
- bias introduced at beginning of training and instability due to similar subsequend steps
  $\rightarrow$ 2 NNs. Predictions of a target NN are used to get the Q-values during training. This target network is only synced periodically. The main NN then uses this to backprpagate its weights.