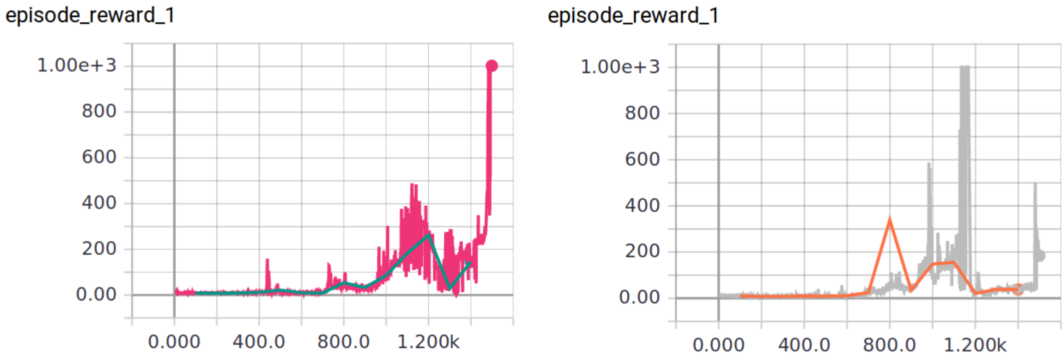
Deep Learning Lab Course
Exercise 4

Julian Stock
4509044

Antoine Schmidt
4613278

Exercise 2.1

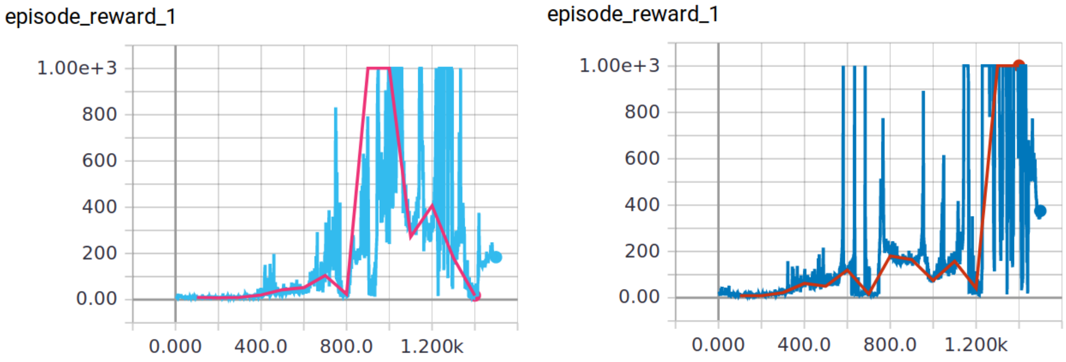
e-greedy



Left q-learning with e-greedy and on the right side dq-learning with e-greedy. For training we used an epsilon value of 0.1. Q-Learning reaches in this setting after 1500 episodes a full score of 1002 +-0 while dq only reaches 176 +-3.58. We only evaluated for 1000 steps, the score of 1002 is caused by the given while loop implementation. We trained all settings with a replay buffer of size 1e4. This might be the reason for the poor performance of dq in this setup.

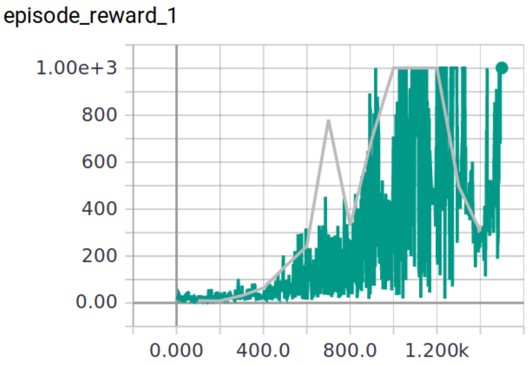
Besides double-q learning we also implemented e-anneal and boltzmann exploration.

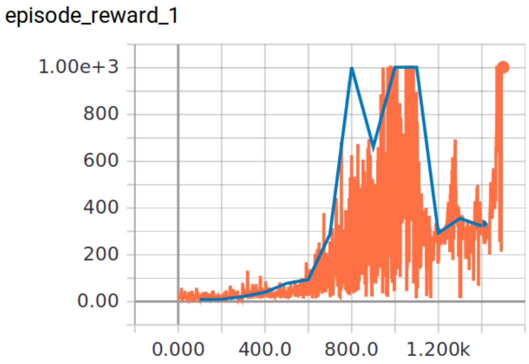
boltzmann



Boltzmann with dq - depicted on the right side - had, after training (1500 episodes) a better result (392.13 +-14.6) than boltzmann with q (left) (212.6 +-10.2)

e-anneal



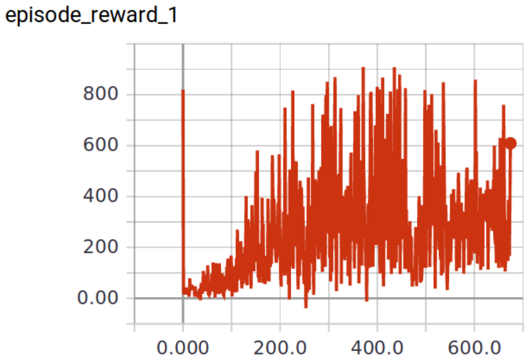


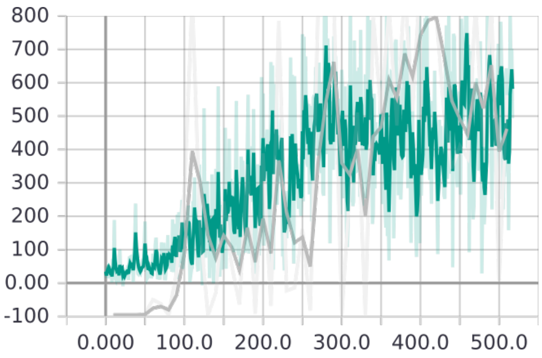
With e-anneal and a decay of 0.999 both settings had perfect solutions after 1500 episodes. Still double-q learning had a better performance because the pole always converges to a fixpoint and thus getting always 100 percent of the possible score, while the q solution survives 1000 steps but slowly strides to the left till out of frame. It can be observed that the worst reward results get slowly lifted because we behave more and more greedy over time.

All solutions balance to a 50/50 action distribution.

We also tried out our RL-solution on the mountaincar problem using dq-learning and e-annealing, with the same settings as in cartpole we achieved a test score of -122.5 +-15.9 after 1000 episodes.

Exercise 2.2





Left shows the training curve with a history of 0 (only one frame), the right is trained on history of 1 (two frames). Both where trained with e-anneal and dq-learning. For better exploration we used a modified random action selection being 12% straight, 6% left/right, 12% accelerate, 1% brake. We additionally implemented an early termination for the training episode in case the racing car gets lost in the green. Additionally using the skip-frame parameter with a value of 3 we achieved a score of 364.8 +-349.5 for the history of one frame and a score of 616.4 +-154.1 for two frames after 1000 episodes (unfortunately we don't have the full graphs anymore). The main issues being for the history of 1 frame getting stuck and standing still till end of episode while the 2 frame history is often too fast to take the narrow curves. Training for more episodes should improve the results.