NEUROPHYSICS 2015 – EXERCISE 4

Richard Hahnloser

Solutions of Joachim Ott and Benjamin Ellenberger

1. REINFCORCEMENT LEARNING (A LA SUTTON AND BARTO)

a) Try to run the Matlab script Sarsa.m by inserting the missing line implementing temporal difference learning of the action-value function Q.

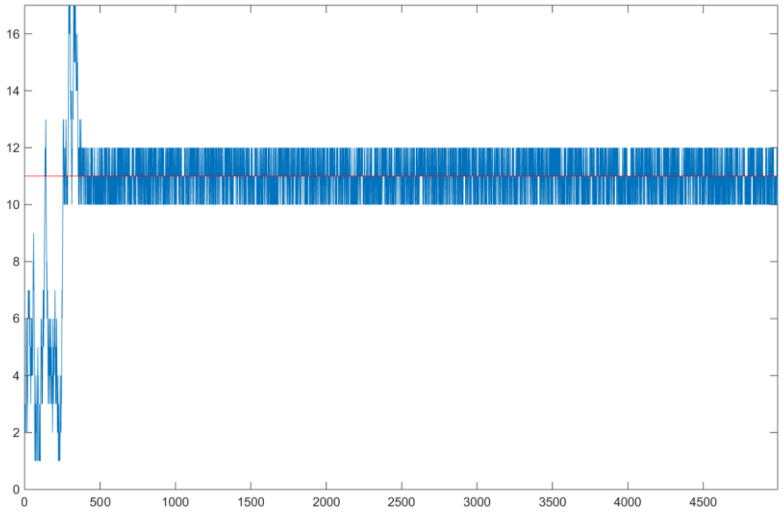
Our Solution:
Q(s_old,h_old)=Q(s_old,h_old)+alpha*(rs(t+rl)+g*Q(s,h)-Q(s_old,h_old));

b) Describe the chosen policy and its behavior during learning.

The chosen policy changes the action value function with learning rate alpha times a weighted sum of the current total reward, the expected future reward and the actual reward gained from the last step. Thereby it chooses the next state according to the highest future value and gets punished if the last decision was not as great as expected. As a behavior we can see that it initially explores the function space and finally finds out where the reward is highest.

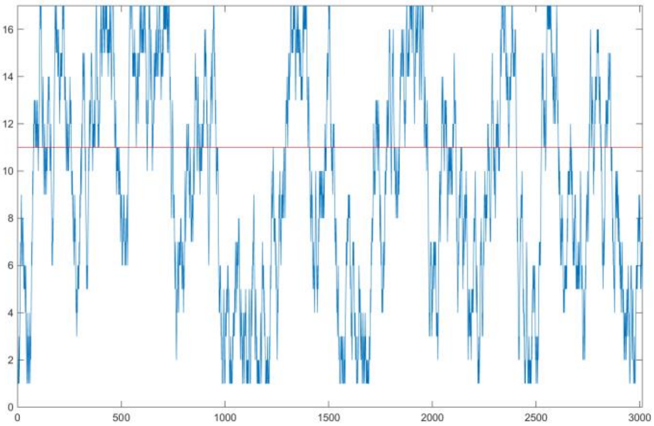
c) There is also an off-policy TD control learning rule (one-step Q-learning), in which case Q converges to the optimum regardless of the policy. Try to find its definition in the literature and implement it. Show that it converges even for a policy in which every action is equally likely. Is convergence of off-policy TD learning slow or fast? (hint: explore convergence for various reward latencies).

Q(s_old,h_old)=Q(s_old,h_old)+alpha*(rs(t+rl));
Works:



Q(s_old,h_old)=Q(s_old,h_old)+alpha*(-Q(s_old,h_old));

Does not Work:



0
2
4
6
8
10
12
14
16
0
500
1000
1500
2000
2500
3000