NEUROPHYSICS 2015 – EXERCISE 4

Richard Hahnloser

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