Artificial Intelligence

Exercises week 4 - Reinforcement learning

COMP9414

Question 1: Value functions

Consider a world with two states $S = \{S_1, S_2\}$ and two actions $A = \{a_1, a_2\}$, where the transitions δ and reward r for each state and action are as follows:

$$\begin{array}{ll} \delta(S_1, a_1) = S_1 & r(S_1, a_1) = 0 \\ \delta(S_1, a_2) = S_2 & r(S_1, a_2) = -1 \\ \delta(S_2, a_1) = S_2 & r(S_2, a_1) = +1 \\ \delta(S_2, a_2) = S_1 & r(S_2, a_2) = +5 \end{array}$$

i. Draw a picture of this world, using circles for the states and arrows forthe transitions.

ii. Assuming a discount factor of $\gamma = 0.9$, determine:

   (a) the optimal policy $\pi^* : S \to A$

   (b) the state-value function $V^* : S \to R$

   (c) the action-value function $Q^* : S \times A \to R$

iii. Write the Q-values in a table (a.k.a. Q-table) as follows:

Q	a_1	a_2
S_1		
S_2		

iv. Trace through the first few steps of the action-value function learning algorithm, with all Q-values initially set to zero. Explain why it is necessary to force exploration through probabilistic choice of actions in order to ensure convergence to the true Q-values.

Question 2: Temporal-difference learning

Consider the same world as the previous question. Assume the use of temporal-difference learning with the following parameters: learning rate $\alpha = 0.3$, discount factor $\gamma = 0.9$, and ϵ-greedy action selection method with $\epsilon = 0.1$. After a few steps of iterating, the learning agent performs action a_1 from state S_1 with the Q-table containing the following values:

Q	a_1	a_2
S_1	0.15	3.55
S_2	5.72	9.18

i. How would look the Q-table after one iteration of the off-policy method Q-learning?

ii. How would look the Q-table after one iteration of the on-policy method Sarsa? Assume a random value $rnd = 0.01$.

iii. Explain how differently would the on-policy Sarsa method converge to the optimal value function in comparison to off-policy Q-learning.

Question 3: Returns

Consider a robot learning a task with a discount factor $\gamma = 0.5$ and receiving the following reward sequence: $R_1 = -1$, $R_2 = 2$, $R_3 = 6$, $R_4 = 3$, and $R_5 = 2$, and then 0 all the time. What are $G_0, G1, ..., G_5$? Hint: Work backwards.