Week 5: Reinforcement Learning (Week 5 Lecture)

Tutorial 10: Reinforcement Learning

10.1 (Activity 9.2: Q-Learning - Open learning)

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:

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

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

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

(a) the optimal policy $\pi^* : S \rightarrow A$
(b) the value function $V^* : S \rightarrow R$
(c) the Q function $Q : S \times A \rightarrow R$

(iii)Write the Q values in a table.

Q	a_1	a_2
S_1		
S_2		

(iv) Trace through the first few steps of the Q-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.