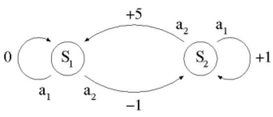
Part 2

Tuesday, 26 January 2021 4:19 pm

Activity: Q-Learning

$$\delta(S_1, a_1) = S_1, \quad r(S_1, a_1) = 0$$
$$\delta(S_1, a_2) = S_2, \quad r(S_1, a_2) = -1$$
$$\delta(S_2, a_1) = S_2, \quad r(S_2, a_1) = +1$$
$$\delta(S_2, a_2) = S_1, \quad r(S_2, a_2) = +5$$

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



ii. Assuming a discount factor of r=0.9, determine:

a. the optimal policy $\pi^* : S \to A$
b. the value function $V^* : S \to R$
c. the "Q" function $Q : S \times A \to R$

$V(S_1, a_1) = 0 + 0.9 \times V(S_1, a_1)$

$= 0$

$V(S_1, a_2) = -1 + 0.9 \times V^*(S_2)$

$\rightarrow V^*(S_2)$

$V(S_2, a_1) = 1 + 0.9 \times V(S_2, a_1)$

$= 10$

$V(S_2, a_2) = 5 + 0.9\,(-1 + 0.9 \times V(S_2, a_2))$

$= 5 - 0.9 + 0.81\, V(S_2, a_2)$

$0.19\, V(S_2, a_2) = 4.1$

$V(S_2, a_2) = 21.57$

$\Rightarrow V^*(S_2) = V(S_2, a_2)$

a): $\pi^* = \begin{cases} S_1 : a_2 \\ S_2 : a_2 \end{cases}$

b) $V^* = \begin{cases} S_1 : -1 + 0.9 \times 21.57 = 18.413 \\ S_2 : 21.57 \end{cases}$

c) Q:

iii) Write the Q values in a matrix (do one and choose the best action)

Q	a_1	a_2
S_1	16.58	18.413
S_2	20.62	21.58

(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.

(Initially, no exploration)

current state	chosen action	new Q value
S_1	a_1	0 + γ*0 = 0
S_1	a_2	-1 + γ*0 = -1
S_2	a_1	1 + γ*0 = +1

Q	a_1	a_2
S_1	0	-1
S_2	1	0

(the Q table)

(If forcing exploration)

current state	chosen action	new Q value
S_2	a_2	5 + γ*0 = 5
S_1	a_1	0 + γ*0 = 0
S_1	a_2	-1 + γ*5 = 3.5
S_2	a_1	1 + γ*5 = 5.5
S_2	a_2	5 + γ*3.5 = 8.15

Q	a_1	a_2
S_1	0	3.5
S_2	5.5	8.15

(the Q table)

(Convergence)

current state	chosen action	new Q value
S_1	a_1	0 + γ*3.5 = 3.15
S_1	a_2	-1 + γ*8.15 = 6.335
S_2	a_1	1 + γ*8.15 = 8.335
S_2	a_2	5 + γ*6.34 = 10.70