Reinforcement Learning—HW 7 Solution

March 30, 2020

1. Your TD algorithm generates the following episode using policy π when interacting with its environment. This is the first episode that has been generated.

Timestep	**Reward**	**State**	**Action**
0		S1	A1
1	16	S1	A2
2	12	S1	A1
3	24	S1	A1
4	16	T	

Assume the discount factor, γ, is $\frac{1}{2}$, the step size, α, is 0.1, and that all q_π values are currently 0.

What are the estimates of: $q_\pi(S1, A1)$ and $q_\pi(S1, A2)$ assuming the use of 2-step Sarsa? Show your work.

Solution: Let's look at the rewards:

After the second timestep,

$$\begin{aligned}
q_\pi(S_0, A_0) &= q_\pi(S_0, A_0) + \alpha[R_1 + \gamma R_2 + \gamma^2 q_\pi(S_2, A_2) - q_\pi(S_0, A_0)] \\
q_\pi(S1, A1) &= q_\pi(S1, A1) + \alpha[R_1 + \gamma R_2 + \gamma^2 q_\pi(S1, A1) - q_\pi(S1, A1)] \\
&= 0 + .1[16 + .5 \times 12 + .5^2 \times 0 - 0] \\
&= 0 + .1[16 + 6] \\
&= 2.2
\end{aligned}$$

After the third timestep:

$$\begin{aligned}
q_\pi(S_1, A_1) &= q_\pi(S_1, A_1) + \alpha[R_2 + \gamma R_3 + \gamma^2 q_\pi(S_3, A_3) - q_\pi(S_1, A_1)] \\
q_\pi(S1, A2) &= q_\pi(S1, A2) + \alpha[R_2 + \gamma R_3 + \gamma^2 q_\pi(S1, A1) - q_\pi(S1, A2)] \\
&= 0 + .1[12 + .5 \times 24 + .5^2 \times 2.2 - 0] \\
&= 0 + .1[12 + 12 + 0.55] \\
&= 2.455
\end{aligned}$$

After the fourth timestep:

$$\begin{aligned}
q_\pi(S_2, A_2) &= q_\pi(S_2, A_2) + \alpha[R_3 + \gamma R_4 + \gamma^2 q_\pi(S_4, A_4) - q_\pi(S_2, A_2)] \\
q_\pi(S1, A1) &= q_\pi(S1, A1) + \alpha[R_3 + \gamma R_4 + \gamma^2 q_\pi(T, \odot) - q_\pi(S1, A1)] \\
&= 2.2 + .1[24 + .5 \times 16 + .5^2 \times 0 - 2.2] \\
&= 2.2 + .1[24 + 8 + 0 - 2.2] \\
&= 2.2 + 2.98 \\
&= 5.18
\end{aligned}$$

After the fifth timestep:

$$\begin{aligned}
q_\pi(S_3, A_3) &= q_\pi(S_3, A_3) + \alpha[R_4 + \gamma q_\pi(S_4, A_4) - q_\pi(S_3, A_3)] \\
q_\pi(S1, A1) &= q_\pi(S1, A1) + \alpha[16 + \gamma q_\pi(T, \odot) - q_\pi(S1, A1)] \\
&= 5.18 + .1[16 + .5 \times 0 - 5.18] \\
&= 5.18 + .1[16 - 5.18] \\
&= 5.18 + 1.082 \\
&= 6.262
\end{aligned}$$

So, our estimates are: $q_\pi(S1, A1) = 6.262, q_\pi(S1, A2) = 2.455$.