Reinforcement Learning—HW 7

March 23, 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.