# hw5_rl_q5_model_free_rl_cycle

## Question 5: Model-Free RL: Cycle

0 points possible (ungraded)

*We recommend you work out the solutions to the following questions on a sheet of scratch paper, and then enter your results into the answer boxes.*

Consider an MDP with 3 states, A, B and C; and 2 actions Clockwise and Counterclockwise. We do not know the transition function or the reward function for the MDP, but instead, we are given with samples of what an agent actually experiences when it interacts with the environment (although, we do know that we do not remain in the same state after taking an action). In this problem, instead of first estimating the transition and reward functions, we will directly estimate the Q function using Q-learning.

Assume, the discount factor, $\gamma$ is 0.5 and the step size for Q-learning, $\alpha$ is 0.5.

Our current Q function, $Q(s, a)$, is as follows.

| | A | B | C |
|---|---|---|---|
| Clockwise | 0.0 | 0.0 | 0.0 |
| Counterclockwise | 0.0 | -5.813 | -1.0 |

The agent encounters the following samples.

| s | a | s' | r |
|---|---|---|---|
| A | Clockwise | B | 0.0 |
| B | Clockwise | C | 0.0 |

Process the samples given above. Below fill in the Q-values after both samples have been accounted for.

| | A | B | C |
|---|---|---|---|
| Clockwise | 0<br>Answer: $0.000$ | 0<br>Answer: $0.000$ | 0<br>Answer: $0.000$ |
| Countercl ockwise | 0<br>Answer: $0.000$ | -5.813<br>Answer: $-5.812$ | -1.0<br>Answer: $-1.000$ |

**For this problem, you may press "Check" as many times as you want without resetting the problem, so that you don't have to reset the problem for trivial math mistakes.**

For each $s, a, s', r$ transition sample, you update the Q value function as follows:
$Q(s,a) = (1-\alpha) * Q(s,a) + \alpha * (R(s,a,s') + \gamma * max_{a'} Q(s',a'))$

Note that because there are only two samples, at least four of the values stay the same.

Submit