hw4_mdps_q8_policy_iteration_cycle

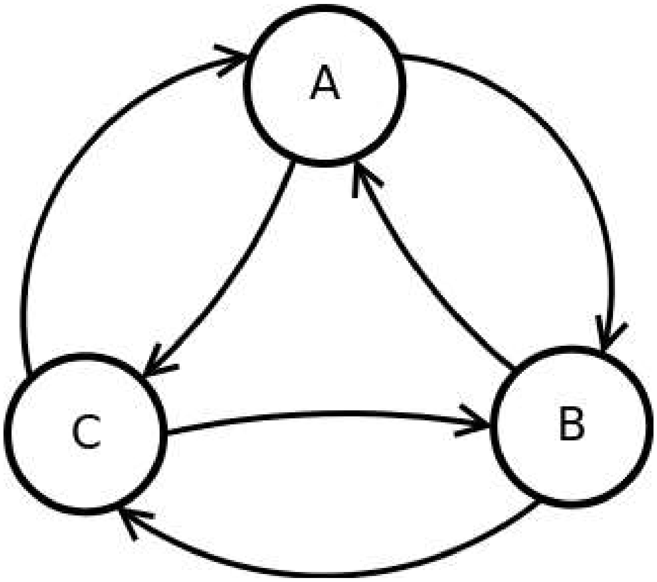
Question 8: Policy Iteration: Cycle

16/16 points (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 the following transition diagram, transition function and reward function for an MDP.

Discount Factor, γ = 0.5



s	a	s'	T(s,a,s')	R(s,a,s')
A	Clockwise	B	1.0	-2.0
A	Counterclockwise	C	1.0	1.0
B	Clockwise	C	1.0	2.0
B	Counterclockwise	A	1.0	-1.0
C	Clockwise	A	1.0	0.0
C	Counterclockwise	A	0.4	-2.0
C	Counterclockwise	B	0.6	-2.0

Suppose we are doing policy evaluation, by following the policy given by the left-hand side table below. Our current estimates (at the end of some iteration of policy evaluation) of the value of states when following the current policy is given in the right-hand side table.

A	B	C
Counterclockwise	Counterclockwise	Counterclockwise

$V_k^\pi(A)$	$V_k^\pi(B)$	$V_k^\pi(C)$
0.000	-0.500	-2.100

Part 1: What is $V_{k+1}^\pi(B)$?

-1 ✓

Correct: Your answer evaluated to -1.000, which is close enough to the correct answer, -1.000.

Suppose that policy evaluation converges to the following value function, V_∞^π.

$V_\infty^\pi(A)$	$V_\infty^\pi(B)$	$V_\infty^\pi(C)$
-0.182	-1.091	-2.364

Now let's execute policy improvement.
Part 2: What is Q_∞^π(B, clockwise)?

0.818 ✓

Correct: Your answer evaluated to 0.818, which is close enough to the correct answer, 0.818.

Part 3: What is Q_∞^π(B, counterclockwise)?

-1.091 ✓

Correct: Your answer evaluated to -1.091, which is close enough to the correct answer, -1.091.

Part 4: What is the updated action for state B? Enter clockwise or counterclockwise.

clockwise ✓

Submit

✓ Correct (16/16 points)