Policy Iteration Exercise*

Consider the following game. On each round, you have the option of rolling a fair 6-sided die. The first roll is required and comes for free. In all subsequent steps, you have two possible actions:

1. Stop: Stop playing by collecting the dollar value that the die lands on, or
2. Roll: Roll again, paying 1 dollar

You initially start in state Start, where you only have one possible action: Roll. State s_i denotes the state where the die lands on i. Once you decide to Stop, the game is over, transitioning the game to the End state.

In this problem we consider using policy iteration to find the optimal strategy.

Go to next page ➜

*adapted from UCB Sp16 midterm

Q1: The initial policy π is given in the table below. Write down (*but don't solve*) a set of 6 Bellman equations (one for each state) that would help you find the values for this policy. Assume a discount factor of γ = 1. The first equation is given for reference: $V^{\pi}(s_1) = -1 + \frac{1}{6} (V^{\pi}(s_1) + V^{\pi}(s_2) + V^{\pi}(s_3) + V^{\pi}(s_4) + V^{\pi}(s_5) + V^{\pi}(s_6))$

State	s_1	s_2	s_3	s_4	s_5	s_6
π(s)	Roll	Roll	Stop	Stop	Stop	Stop

Q2: We have provided the values for π in the table below. Now perform a policy update to find the new policy π'. The table below shows the old policy π and has filled in parts of the updated policy π' for you.

State	s1	s2	s3	s4	s5	s6
π(s)	Roll	Roll	Stop	Stop	Stop	Stop
$V^{\pi}(s)$	3	3	3	4	5	6
π'(s)	Roll					Stop

Q3: Is π from Q1 optimal? Explain why or why not.

Q1: The initial policy π is given in the table below. Write down (*but don't solve*) a set of 6 Bellman equations (one for each state) that would help you find the values for this policy. Assume a discount factor of γ = 1. The first equation is given for reference: $V^{\pi}(s_1) = -1 + \frac{1}{6} (V^{\pi}(s_1) + V^{\pi}(s_2) + V^{\pi}(s_3) + V^{\pi}(s_4) + V^{\pi}(s_5) + V^{\pi}(s_6))$

State	s_1	s_2	s_3	s_4	s_5	s_6
π(s)	Roll	Roll	Stop	Stop	Stop	Stop

Solution:

$V^{\pi}(s_1) = -1 + \frac{1}{6} (V^{\pi}(s_1) + V^{\pi}(s_2) + V^{\pi}(s_3) + V^{\pi}(s_4) + V^{\pi}(s_5) + V^{\pi}(s_6))$
$V^{\pi}(s_2) = -1 + \frac{1}{6} (V^{\pi}(s_1) + V^{\pi}(s_2) + V^{\pi}(s_3) + V^{\pi}(s_4) + V^{\pi}(s_5) + V^{\pi}(s_6))$
$V^{\pi}(s_3) = 3$
$V^{\pi}(s_4) = 4$
$V^{\pi}(s_4) = 5$
$V^{\pi}(s_4) = 6$

To solve this system (not required for this problem): plug in the values for the last four equations into the first two equations. Solving the resulting equations, we would then obtain the values 3, 3, 3, 4, 5, 6 for the 6 states respectively.

Q2: We have provided the values for π in the table below. Now perform a policy update to find the new policy π'. The table below shows the old policy π and has filled in parts of the updated policy π' for you.

State	s1	s2	s3	s4	s5	s6
π(s)	Roll	Roll	Stop	Stop	Stop	Stop
$V^{\pi}(s)$	3	3	3	4	5	6
π'(s)	Roll	Roll	Roll/Stop	Stop	Stop	Stop

Solution:

For each s_i, we compare the values obtained via Rolling and Stopping. The value of Rolling for each state s_i is $-1 + \frac{1}{6}(3 + 3 + 3 + 4 + 5 + 6) = 3$. The value of Stopping for each state s_i is i. At each state s_i, we take the action that yields the largest value; so, for s_1 and s_2, we Roll, and for s_4 and s_5, we stop. For s_3, we Roll/Stop, since the values from Rolling and Stopping are equal.

Q3: Is π from Q1 optimal? Explain why or why not.

Solution:

Yes, the old policy is optimal. Looking at Q2, there is a tie between 2 equally good policies that policy iteration considers employing. One of these policies is the same as the old policy. This means that both new policies are as equally good as the old policy, and policy iteration has converged. Since policy iteration converges to the optimal policy, we can be sure that π from Q1 is optimal.