Shaun Mathew

Running Programs:

Type python <pia.py or via.py> [p1] [p2] [r\_up] [r\_down] [r\_left] [r\_right] [discount\_factor] [theta].

The default values are those given in the assignment and the program can simply be run by typing python pia.py or python via.py

**Policy Iteration using p1 = 0.7, p2 = 0.2, all rewards = -1, discount faction=0.95 and accuracy=0.001**

Number of iterations of Evaluation Improvement Cycle: 3

Average Iteration Time: ﻿0.004286050796508789

﻿Total Time Taken: 0.014050006866455078

﻿Optimal Policy: ['LEFT', 'LEFT', 'LEFT', 'UP', 'LEFT', 'LEFT', 'DOWN', 'UP', 'UP', 'DOWN', 'DOWN', 'UP', 'RIGHT', 'RIGHT']

**Value Iteration using p1 = 0.7, p2 = 0.2, all rewards = -1, discount faction=0.95 and accuracy=0.001**

﻿Num Iterations of Policy Valuation: 10

Average Iteration Time: 0.00034618377685546875

Total Time Taken: 0.003813028335571289

﻿Optimal Policy: ['LEFT', 'LEFT', 'DOWN', 'UP', 'UP', 'DOWN', 'DOWN', 'UP', 'RIGHT', 'RIGHT', 'DOWN', 'RIGHT', 'RIGHT', 'RIGHT']

The average time per iteration differs by a factor of 12 but this is not as meaningful because policy iteration waits until convergence of the value function before improving. The total time is a better indicator in this regard. From total time taken metric, we see that value iteration converged much faster, 4 times as fast. This is to be expected since we don’t need to wait for value function convergence before performing policy improvement on each iteration.

**Policy Iteration using p1 = 1.0, p2 = 0.0, all rewards = -1, discount faction=0.95 and accuracy=0.001 i.e. Deterministic Environment**

Num Iterations of Evaluation Improvement Cycle: 194

Average Iteration Time: 0.0004715268144902495

Total Time Taken: 0.09159207344055176

Optimal Policy: ['LEFT', 'LEFT', 'LEFT', 'UP', 'UP', 'DOWN', 'DOWN', 'UP', 'LEFT', 'DOWN', 'DOWN', 'RIGHT', 'RIGHT', 'RIGHT']

**Value Iteration using p1 = 1.0, p2 = 0.0, all rewards = -1, discount faction=0.95 and accuracy=0.001 i.e. Deterministic Environment**

Num Iterations of Policy Valuation: 4

Average Iteration Time: 0.0005649328231811523

Total Time Taken: 0.0028679370880126953

Optimal Policy: ['LEFT', 'LEFT', 'LEFT', 'UP', 'UP', 'UP', 'DOWN', 'UP', 'DOWN', 'DOWN', 'DOWN', 'UP', 'RIGHT', 'RIGHT']

In a deterministic environment, the speed up is even more evident. Policy iteration converges in 0.092 seconds, taking about 194 iterations; whereas, value iteration finds the optimal policy in 0.003 seconds. Note that both these policies are optimal. The reason for the difference is when faced with 2 actions that have identical values, an optimizing action is randomly chosen.