Carter Roberts 11/15/2025
Loyola University New Orleans COSC-A405-001: Artificial Intelligence

Deterministic Reinforcement Learning Assignment: Comparing Value Iteration, Policy Iteration, and Q-Learning

In this assignment, I implemented Value Iteration, Policy Iteration, and Q-Learning all on the same state/reward/action set-environment and compared their outputs all at multiple gamma values, those being 0.9, 0.5, and 0.1. Below is the virtualized state maze, with rewards included:

3			+20
2		+10	
1		-10	-20
0			
	0	1	2

The following represents the outputs.

VALUE ITERATION (gamma = 0.9)

Iterations to Convergence: **5**

Final Policy

(0,0): U | (1, 0): L | (2, 0): L | (0, 1): U | (0, 2): U | (2, 2): U | (0, 3): R | (1, 3): R

Final V[]

(0, 0): 13.122 | (1, 0): 11.809800000000001 | (2, 0): 10.628820000000001 | (0, 1): 14.58 | (0, 2): 16.2 | (2, 2): 20.0 | (0, 3): 18.0 | (1, 3): 20.0

VALUE ITERATION (gamma = 0.5)

Iterations to Convergence: **4**

Final Policy

(0,0): U | (1, 0): L | (2, 0): L | (0, 1): U | (0, 2): U | (2, 2): U | (0, 3): R | (1, 3): R

Final V[]

(0, 0): 2.5 | (1, 0): 1.25 | (2, 0): 0.625 | (0, 1): 5.0 | (0, 2): 10.0 | (2, 2): 20.0 | (0, 3): 10.0 | (1, 3): 20.0

VALUE ITERATION (gamma = 0.1)

Iterations to Convergence: **4**

Final Policy

(0,0): U | (1, 0): L | (2, 0): L | (0, 1): U | (0, 2): R | (2, 2): U | (0, 3): R | (1, 3): R

Final V[]
(0, 0): 0.1 | (1, 0): 0.0100000000000000002 | (2, 0): 0.0100000000000000002 | (0, 1): 1.0 | (0, 2): 10.0 | (2, 2): 20.0 | (0, 3): 2.0 | (1, 3): 20.0

POLICY ITERATION (gamma = 0.9)
Iterations to Convergence for V: **16**
Iterations to Convergence for Policy: **4**
Final Policy
(0,0): U | (1, 0): L | (2, 0): L | (0, 1): U | (0, 2): U | (2, 2): U | (0, 3): R | (1, 3): R
Final V[]
(0, 0): 13.122 | (1, 0): 11.809800000000001 | (2, 0): 10.628820000000001 | (0, 1): 14.58 | (0, 2): 16.2 | (2, 2): 20.0 | (0, 3): 18.0 | (1, 3): 20.0

POLICY ITERATION (gamma = 0.5)
Iterations to Convergence for V: **9**
Iterations to Convergence for Policy: **3**
Final Policy
(0,0): U | (1, 0): L | (2, 0): L | (0, 1): U | (0, 2): R | (2, 2): U | (0, 3): R | (1, 3): R
Final V[]
(0, 0): 2.5 | (1, 0): 1.25 | (2, 0): 0.625 | (0, 1): 5.0 | (0, 2): 10.0 | (2, 2): 20.0 | (0, 3): 10.0 | (1, 3): 20.0

POLICY ITERATION (gamma = 0.1)
Iterations to Convergence for V: **11**
Iterations to Convergence for Policy: **3**
Final Policy
(0,0): U | (1, 0): L | (2, 0): L | (0, 1): U | (0, 2): R | (2, 2): U | (0, 3): R | (1, 3): R
Final V[]
(0, 0): 0.1 | (1, 0): 0.0100000000000000002 | (2, 0): 0.0100000000000000002 | (0, 1): 1.0 | (0, 2): 10.0 | (2, 2): 20.0 | (0, 3): 2.0 | (1, 3): 20.0

Q LEARNING (gamma = 0.9)
Iterations to Convergence: **6**
Final Policy
(0,0): U | (1, 0): L | (2, 0): L | (0, 1): U | (0, 2): U | (2, 2): U | (0, 3): R | (1, 3): R

Final Q[] (coord, coord): [Lq, Rq, Uq, Dq]
(0, 0): [13.122, 10.62882…1, 11.8098…1, 11.8098…1]
(1, 0): [-10.0, 9.565938…1, 11.8098…1, 10.62882…1]
(2, 0): [-20.0, 9.565938…1, 10.62882…1, 9.565938…1]
(0, 1): [14.58, -10.0, 13.122, 11.8098…1]
(0, 2): [16.2, 10.0, 14.58, 13.122]
(2, 2): [20.0, 18.0, 10.0, -20.0]
(0, 3): [16.2, 18.0, 16.2, 14.58]
(1, 3): [18.0, 20.0, 16.2, 10.0]

Q LEARNING (gamma = 0.5)
Iterations to Convergence: **5**
Final Policy
(0,0): U | (1, 0): L | (2, 0): L | (0, 1): U | (0, 2): R | (2, 2): U | (0, 3): R | (1, 3): R
Final Q[] (coord, coord): [Lq, Rq, Uq, Dq]
(0, 0): [2.5, 0.625, 1.25, 1.25]
(1, 0): [-10.0, 0.3125, 0.625, 1.25]
(2, 0): [-20.0, 0.3125, 0.3125, 0.625]
(0, 1): [5.0, -10.0, 1.25, 2.5]
(0, 2): [5.0, 10.0, 2.5, 5.0]
(2, 2): [20.0, 10.0, -20.0, 10.0]
(0, 3): [5.0, 10.0, 5.0, 5.0]
(1, 3): [10.0, 20.0, 10.0, 5.0]

Q LEARNING (gamma = 0.1)
Iterations to Convergence: **5**
Final Policy
(0,0): U | (1, 0): L | (2, 0): L | (0, 1): U | (0, 2): R | (2, 2): U | (0, 3): R | (1, 3): R
Final Q[] (coord, coord): [Lq, Rq, Uq, Dq]
(0, 0): [0.001…2, 0.01…2, 0.01…2, 0.1]
(1, 0): [0.0001…3, 0.01…2, 0.001…2, -10.0]
(2, 0): [0.001…3, 0.001…2, 0.0001…3, -20.0]
(0, 1): [-10.0, 0.1, 0.01…2, 1.0]
(0, 2): [10.0, 1.0, 0.1, 0.2]
(2, 2): [2.0, 10.0, -20.0, 20.0]

(0, 3): [2.0, 0.2, 1.0, 0.2]
(1, 3): [20.0, 0.2, 10.0, 2.0]

QUESTIONS & ANSWERS ABOUT RESULTS

- Which algorithm converges fastest?

The Value Iteration converges fastest.

- How does the discount gamma factor effect:
    - Value function magnitude?

The gamma function massively decides the degree to which values decay along a path to a reward. 0.9 and 0.1 are worlds apart in value magnitude.

    - Preference for long-term / short-term rewards?

Lower gamma factors can make a state, such as (0, 2), which is along a longer path to +20 but also directly adjacent to +10, decide to pick the adjacent +10 due to an extremely decaying gamma factor like 0.1 which would make the +20 worth less by the time it would reach that terminal state.

    - Policy behavior?

The policy is mostly the same regardless of gamma, except for (0, 2) for reasons outlined just previously.