Theoretical Framework

The solutions below utilize the standard Bellman operators for Markov Decision Processes (MDPs).

1. **Bellman Expectation Equation (for Policy π):** The value of a state s under policy π is the expected immediate reward plus the discounted value of the next state.

$$V^{\pi}(s) = \sum_{s' \in \mathcal{S}} P(s' \mid s, \pi(s)) \left[R(s, \pi(s), s') + \gamma V^{\pi}(s') \right]$$

2. **Bellman Optimality Equation:** The optimal value $V^*(s)$ is the maximum return achievable from state s.

$$V^*(s) = \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} P(s' \mid s, a) \left[R(s, a, s') + \gamma V^*(s') \right]$$

Problem 1: Policy Iteration (Deterministic)

Parameters: $\mathcal{S} = \{s_1, s_2\}$, $\mathcal{A} = \{a_1, a_2\}$, $\gamma = 0.9$, Convergence $\theta = 0.85$.

1.1 Initialization

Arbitrary policy: $\pi^0(s_1) = a_1, \pi^0(s_2) = a_1$. Value $V_0 = [0, 0]^\top$.

1.2 Policy Evaluation (Iterative)

We iterate $V_{k+1}(s) = R(s, \pi(s), s') + \gamma V_k(s')$.

Iteration 1:

$$V_1(s_1) = r(s_1, a_1, s_1) + 0.9V_0(s_1) = 0 + 0 = 0$$
$$V_1(s_2) = r(s_2, a_1, s_2) + 0.9V_0(s_2) = 1 + 0 = 1$$

$\Delta = 1 > \theta$. Continue.

Iteration 2:

$$V_2(s_1) = 0 + 0.9V_1(s_1) = 0$$
$$V_2(s_2) = 1 + 0.9V_1(s_2) = 1.9$$

$\Delta = 0.9 > \theta$. Continue.

Iteration 3:

$$V_3(s_1) = 0 + 0.9(0) = 0$$
$$V_3(s_2) = 1 + 0.9(1.9) = 2.71$$

$\Delta = 0.81 < \theta$. **Converged.**

$$V^{\pi^0} \approx \begin{bmatrix} 0 \\ 2.71 \end{bmatrix}$$

1.3 Policy Improvement

State s_1:

$$Q(s_1, a_1) = 0 + 0.9(0) = 0$$
$$Q(s_1, a_2) = 1 + 0.9(2.71) = 3.439$$

$3.439 > 0 \implies \pi(s_1) \leftarrow a_2$.

State s_2:

$$Q(s_2, a_1) = 1 + 0.9(2.71) = 3.439$$
$$Q(s_2, a_2) = 0 + 0.9(0) = 0$$

$3.439 > 0 \implies \pi(s_2) \leftarrow a_1$ (Unchanged).

New Policy: $\pi^1 = [a_2, a_1]^\top$.

Problem 2: Value Iteration

Update Rule: $V_{k+1}(s) = \max_a\{R(s, a, s') + \gamma V_k(s')\}$.

Iteration 1 ($V_0 = [0, 0]$):

$$V_1(s_1) = \max\{0, 1\} = 1$$
$$V_1(s_2) = \max\{1, 0\} = 1$$

Iteration 2:

$$V_2(s_1) = \max\{0.9(1), 1 + 0.9(1)\} = 1.9$$
$$V_2(s_2) = \max\{1 + 0.9(1), 0.9(1)\} = 1.9$$

Iteration 3:

$$V_3(s_1) = \max\{0.9(1.9), 1 + 0.9(1.9)\} = \max\{1.71, 2.71\} = 2.71$$
$$V_3(s_2) = \max\{1 + 0.9(1.9), 0.9(1.9)\} = \max\{2.71, 1.71\} = 2.71$$

Policy Extraction:

$$\pi^*(s_1) = a_2, \quad \pi^*(s_2) = a_1$$

Problem 3: Stochastic Policy Iteration

Parameters:

- $A \xrightarrow{a_1} A$ (p=1, r=10); $A \xrightarrow{a_2} A(0.2, -10), B(0.8, -5)$.
- $B \xrightarrow{a_1} B(0.2, 10), A(0.8, 40)$; $B \xrightarrow{a_2} B(0.2, 10), A(0.8, 20)$.

3.1 Iteration 1 ($\pi^0 = [a_2, a_2]^\top$)

Step 1: Evaluation Solving the system:

$$V(A) = -6 + 0.18V(A) + 0.72V(B)$$
$$V(B) = 18 + 0.72V(A) + 0.18V(B)$$

Solving yields $V(A) \approx 52.2$, $V(B) \approx 67.8$.

Step 2: Improvement

- $Q(A, a_1) = 10 + 0.9(52.2) = 56.98 > V(A) \implies$ Switch to a_1.
- $Q(B, a_1) = 0.8(40 + 46.98) + 0.2(10 + 61.02) \approx 83.5 > V(B) \implies$ Switch to a_1.

New Policy: $\pi^1 = [a_1, a_1]^\top$.

3.2 Iteration 2 ($\pi^1 = [a_1, a_1]^\top$)

Step 1: Evaluation $V(A) = 10 + 0.9V(A) \implies V(A) = 100$. $V(B)$ solves to ≈ 129.3.

Step 2: Improvement Check a_2 actions:

- $Q(A, a_2) = 104.88 > 100 \implies$ Switch A to a_2.
- $Q(B, a_2) = 113.22 < 129.3 \implies$ Keep B as a_1.

New Policy: $\pi^2 = [a_2, a_1]^\top$.

Problem 4: Grid World Analysis

4.1 Value Iteration Snapshots

Values are corrected to reflect Step Cost = -1 and Goal Reward = +200. Values generally decrement by 1 per step from the goal.

V_5 (Intermediate)

-5	185	186	187
-5		187	188
-5		188	199
-5			**G**
-5	-5	-5	

V_{13} (Converged)

194	195	196	197
193		197	198
192		198	199
191			**G**
190	189	188	

4.2 Sample Cell Calculation

For Cell 14 (neighbor of G):

$$V(14) = \max_a(R + \gamma V(G)) = -1 + 200 + 0 = 199.$$

The values in the converged table now consistently reflect the Manhattan distance (accounting for walls) from the goal. For example, the bottom-left cell (190) is 10 steps away from G (value $200 - 10 = 190$) via the only valid path around the central wall.

4.3 Optimal Policy π^*

The policy is derived greedily w.r.t V_{13}. Note the bottom-left corner must move Up to escape.

$\rightarrow$	$\rightarrow$	$\rightarrow$	$\downarrow$
$\uparrow$		$\rightarrow$	$\downarrow$
$\uparrow$		$\rightarrow$	$\downarrow$
$\uparrow$			**G**
$\uparrow$	$\leftarrow$	$\leftarrow$	

Problem 5: Bellman Equations Derivation

a. Bellman Expectation Equations (Fixed Policy π)

1. $V^\pi(s)$ via $V^\pi(s')$:

$$V^\pi(s) = \sum_{s'} P(s' \mid s, \pi(s)) \left[R(s, \pi(s), s') + \gamma V^\pi(s')\right]$$

2. $V^\pi(s)$ via $Q^\pi(s, a)$:

$$V^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a \mid s) Q^\pi(s, a)$$

3. $Q^\pi(s, a)$ via $V^\pi(s')$:

$$Q^\pi(s, a) = \sum_{s'} P(s' \mid s, a) \left[R(s, a, s') + \gamma V^\pi(s')\right]$$

4. $Q^\pi(s, a)$ via $Q^\pi(s', a')$:

$$Q^\pi(s, a) = \sum_{s'} P(s' \mid s, a) \left[R(s, a, s') + \gamma \sum_{a'} \pi(a' \mid s') Q^\pi(s', a')\right]$$

b. Bellman Optimality Equations (π^*)

1. $V^*(s)$ via $V^*(s')$:

$$V^*(s) = \max_{a \in \mathcal{A}} \sum_{s'} P(s' \mid s, a) \left[R(s, a, s') + \gamma V^*(s')\right]$$

2. $V^*(s)$ via $Q^*(s, a)$:

$$V^*(s) = \max_{a \in \mathcal{A}} Q^*(s, a)$$

3. $Q^*(s, a)$ via $V^*(s')$:

$$Q^*(s, a) = \sum_{s'} P(s' \mid s, a) \left[R(s, a, s') + \gamma V^*(s')\right]$$

4. $Q^*(s, a)$ via $Q^*(s', a')$:

$$Q^*(s, a) = \sum_{s'} P(s' \mid s, a) \left[R(s, a, s') + \gamma \max_{a'} Q^*(s', a')\right]$$