

Problem 1: Deterministic Policy Iteration

Setup:

- States: $\mathcal{S} = \{A, B\}$
- Actions: $\mathcal{A} = \{a^1, a^2\}$
- Transitions (P) and Rewards (R) derived from problem statement:
 - a^1 : $A \rightarrow B$ (0.8), $A \rightarrow A$ (0.2). $B \rightarrow B$ (1).
 - a^2 : $A \rightarrow A$ (1). $B \rightarrow B$ (0.9), $B \rightarrow A$ (0.1).
- Discount factor: $\gamma = 0.9$.

a. Transition Matrices and Reward Vectors

For action a^1 :

$$M(a^1) = \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix}, \quad R_{ss'}^{a^1} = \begin{bmatrix} -1.5 & 0.5 \\ 0 & 0.5 \end{bmatrix}$$

Expected immediate reward vector $R_s^{a^1}$:

$$R_s^{a^1} = (M(a^1) \odot R_{ss'}^{a^1})\mathbf{1} = \begin{bmatrix} 0.2(-1.5) + 0.8(0.5) \\ 0(0) + 1(0.5) \end{bmatrix} = \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix}$$

For action a^2 :

$$M(a^2) = \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix}, \quad R_{ss'}^{a^2} = \begin{bmatrix} -1 & 0 \\ -1 & 1 \end{bmatrix}$$

Expected immediate reward vector $R_s^{a^2}$:

$$R_s^{a^2} = (M(a^2) \odot R_{ss'}^{a^2})\mathbf{1} = \begin{bmatrix} 1(-1) + 0(0) \\ 0.1(-1) + 0.9(1) \end{bmatrix} = \begin{bmatrix} -1 \\ 0.8 \end{bmatrix}$$

b. Policy Iteration

Iteration 1: $\pi^0 = [a^2, a^1]^\top$

1. Policy Evaluation:

$$P(\pi^0) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad R_s^{\pi^0} = \begin{bmatrix} -1 \\ 0.5 \end{bmatrix}$$

Solving $V^{\pi^0} = (I - \gamma P(\pi^0))^{-1} R_s^{\pi^0}$:

$$V^{\pi^0} = \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} -1 \\ 0.5 \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix} \begin{bmatrix} -1 \\ 0.5 \end{bmatrix} = \begin{bmatrix} -10 \\ 5 \end{bmatrix}$$

2. Policy Improvement:

$$Q(A, a^1) = 0.1 + 0.9(0.2(-10) + 0.8(5)) = 0.1 + 0.9(2) = 1.9$$

$$Q(A, a^2) = -1 + 0.9(1(-10) + 0(5)) = -1 - 9 = -10$$

$$\implies \pi^1(A) = a^1 \quad (1.9 > -10)$$

$$Q(B, a^1) = 0.5 + 0.9(0(-10) + 1(5)) = 0.5 + 4.5 = 5$$

$$Q(B, a^2) = 0.8 + 0.9(0.1(-10) + 0.9(5)) = 0.8 + 0.9(3.5) = 3.95$$

$$\implies \pi^1(B) = a^1 \quad (5 > 3.95)$$

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

Iteration 2: $\pi^1 = [a^1, a^1]^\top$

1. Policy Evaluation:

$$P(\pi^1) = \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix}, \quad R_s^{\pi^1} = \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix}$$

Solving linear system:

$$\begin{aligned} V(B) &= 0.5 + 0.9(1)V(B) \implies 0.1V(B) = 0.5 \implies V(B) = 5 \\ V(A) &= 0.1 + 0.9(0.2V(A) + 0.8(5)) = 0.1 + 0.18V(A) + 3.6 \\ 0.82V(A) &= 3.7 \implies V(A) \approx 4.512 \end{aligned}$$

$$V^{\pi^1} \approx [4.512, 5]^\top.$$

2. Policy Improvement:

$$\begin{aligned} Q(A, a^1) &= 4.512 \quad (\text{Current}) \\ Q(A, a^2) &= -1 + 0.9(4.512) = 3.06 \\ &\implies \pi^2(A) = a^1 \\ Q(B, a^1) &= 5 \quad (\text{Current}) \\ Q(B, a^2) &= 0.8 + 0.9(0.1(4.512) + 0.9(5)) = 0.8 + 0.9(4.9512) = 5.256 \\ &\implies \pi^2(B) = a^2 \quad (5.256 > 5) \end{aligned}$$

New Policy: $\pi^2 = [a^1, a^2]^\top$.

Iteration 3: $\pi^2 = [a^1, a^2]^\top$

1. Policy Evaluation:

$$P(\pi^2) = \begin{bmatrix} 0.2 & 0.8 \\ 0.1 & 0.9 \end{bmatrix}, \quad R_s^{\pi^2} = \begin{bmatrix} 0.1 \\ 0.8 \end{bmatrix}$$

Solving linear system:

$$\begin{aligned} 0.82V(A) - 0.72V(B) &= 0.1 \\ -0.09V(A) + 0.19V(B) &= 0.8 \end{aligned}$$

Solving yields $V^{\pi^2} \approx [6.538, 7.307]^\top$.

2. Policy Improvement:

$$\begin{aligned} Q(A, a^1) &= 6.538 \quad (\text{Current}) \\ Q(A, a^2) &= -1 + 0.9(6.538) = 4.88 \\ Q(B, a^1) &= 0.5 + 0.9(7.307) = 7.076 \\ Q(B, a^2) &= 7.307 \quad (\text{Current}) \end{aligned}$$

Policy stable. Optimal Policy $\pi^* = [a^1, a^2]^\top$.

Problem 2: Value Iteration

Setup: Same parameters as Problem 1. Update rule: $V_{k+1}(s) = \max_a \{R_s^a + \gamma P(a)V_k\}$.

Iterations

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

$$\begin{aligned} V_1(A) &= \max\{0.1, -1\} = 0.1 \\ V_1(B) &= \max\{0.5, 0.8\} = 0.8 \end{aligned}$$

Iteration 2:

$$\begin{aligned} V_2(A) &= \max\{0.1 + 0.9(0.2(0.1) + 0.8(0.8)), -1 + 0.9(0.1)\} = 0.694 \\ V_2(B) &= \max\{0.5 + 0.9(0.8), 0.8 + 0.9(0.1(0.1) + 0.9(0.8))\} = 1.457 \end{aligned}$$

Iteration 3: Using truncated inputs from text (0.69, 1.45):

$$\begin{aligned} V_3(A) &= \max\{0.1 + 0.9(0.2(0.69) + 0.8(1.45)), -0.379\} = 1.268 \\ V_3(B) &= \max\{0.5 + 0.9(1.45), 0.8 + 0.9(0.1(0.69) + 0.9(1.45))\} = 2.036 \end{aligned}$$

Iteration 4:

$$V_4(A) = 1.794, \quad V_4(B) = 2.563$$

Iteration 5:

$$V_5(A) = 2.268, \quad V_5(B) = 3.037$$

Convergence check: $\Delta \approx 0.47 < 0.5$. Converged.

Policy Extraction

Using V_5 :

$$\begin{aligned} \pi^*(A) &= \operatorname{argmax}(2.694, 1.041) = a^1 \\ \pi^*(B) &= \operatorname{argmax}(3.233, 3.464) = a^2 \end{aligned}$$

Matches Problem 1 result.

Problem 3: Monte Carlo Policy Evaluation

Note: The reward structure and transitions in the provided rollouts differ from Problem 1. We analyze the rollouts as ground truth for this specific problem instance.

Method: First-visit Monte Carlo.

Rollout Analysis (Initial Policy π^0)

- **Ep 1** (A, a^1): Returns 0. $G = 0$.
- **Ep 2** (A, a^2): $A \xrightarrow{a^2, 4} B \xrightarrow{a^2, -1} A \dots$
 $G = 4 + 0.9(-1) + 0 = 3.1$.

- **Ep 3** (B, a^1) : $B \xrightarrow{a^1, 5} B \xrightarrow{a^2, -1} A \dots$
 $G = 5 + 0.9(-1) + 0 = 4.1.$
- **Ep 4** (B, a^2) : $B \xrightarrow{a^2, -1} A \dots$
 $G = -1 + 0 = -1.$

Policy Update 1

Estimated Q-Values:

$$Q(A, a^1) \approx 0, \quad Q(A, a^2) \approx 3.1$$

$$Q(B, a^1) \approx 4.1, \quad Q(B, a^2) \approx -1$$

New Policy π^1 : $\pi(A) = a^2, \pi(B) = a^1.$

Evaluation of π^1 (New Rollouts)

We observe new traces generated by π^1 . Based on the provided finite-horizon problem context, the returns are calculated as:

- **Start** (A, a^1) : $G \approx 14.57.$
- **Start** (A, a^2) : $G \approx 19.47.$
- **Start** (B, a^1) : $G \approx 16.19.$
- **Start** (B, a^2) : $G \approx 13.57.$

Policy Update 2 (π^2)

Comparing Q-values:

- $\pi'(A) = \operatorname{argmax}(14.57, 19.47) = a^2.$
- $\pi'(B) = \operatorname{argmax}(16.19, 13.57) = a^1.$

New Policy: $\pi^2 = [a^2, a^1]^\top$. Since $\pi^2 = \pi^1$, the policy has converged.