Problem 1: Discrete LSPI (Scalar Basis)

Setup:

- State Space: $\mathcal{S} = \{-1, 1, 2\}$
- Action Space: $\mathcal{A} = \{-1, 0, 1\}$
- Transitions D: $\{(s_0 = 1, a_0 = 1, r_1 = 1, s_1 = 2), (2, 0, -1, 1), (1, -1, 0, -1)\}$
- Basis Function: $\phi(s, a) = a^2 s + as + a$
- Initial Weight: $\omega^0 = 1$
- Discount: $\gamma = 0.9$

1.1 Derivation of Policy π^1

The policy is greedy with respect to $Q(s, a) = \phi(s, a)\omega^0$.

$$\pi^1(s) = \underset{a \in \mathcal{A}}{\text{argmax}} \, (a^2 s + as + a) \cdot 1$$

State $s = 2$:

$$\pi^1(2) = \underset{a}{\text{argmax}} \, (2a^2 + 3a) = \text{argmax}\{-1, 0, 5\} = 1$$

State $s = 1$:

$$\pi^1(1) = \underset{a}{\text{argmax}} \, (a^2 + 2a) = \text{argmax}\{-1, 0, 3\} = 1$$

State $s = -1$:

$$\pi^1(-1) = \underset{a}{\text{argmax}} \, (-a^2) = \text{argmax}\{-1, 0, -1\} = 0$$

Resulting Policy:

$$\pi^1(s) = \begin{cases} 0 & \text{if } s = -1 \\ 1 & \text{if } s = 1 \\ 1 & \text{if } s = 2 \end{cases}$$

1.2 Policy Evaluation (Calculation of ω^1)

We compute the scalar A and b using the LSTDQ statistics:

$$A = \frac{1}{L} \sum_{i=0}^{L-1} \phi(s_i, a_i) \left(\phi(s_i, a_i) - \gamma \phi(s_{i+1}, \pi^1(s_{i+1})) \right)$$

$$b = \frac{1}{L} \sum_{i=0}^{L-1} \phi(s_i, a_i) r_{i+1}$$

Substituting the samples from D:

$$A = \frac{1}{3} \Big[3(3 - 0.9(5)) + 0(0 - 0.9(3)) + (-1)(-1 - 0.9(0)) \Big]$$

$$= \frac{1}{3} \Big[3(-1.5) + 0 + 1 \Big] = \frac{1}{3}[-3.5] = -1.1667$$

$$b = \frac{1}{3}\Big[3(1) + 0(-1) + (-1)(0)\Big] = \frac{1}{3}[3] = 1$$

Solving for weights:

$$\omega^1 = A^{-1}b = \frac{1}{-1.1667} = -0.8571$$

1.3 Derivation of Policy π^2

Update policy greedy w.r.t $Q(s,a) = \phi(s,a)(-0.8571)$.

$$\pi^2(s) = \underset{a\in\mathcal{A}}{\text{argmax}}\,(a^2 s + as + a)(-0.8571)$$

State $s = 2$:

$$\pi^2(2) = \text{argmax}\{0.8571, 0, -4.2855\} = -1$$

State $s = 1$:

$$\pi^2(1) = \text{argmax}\{0.8571, 0, -2.5713\} = -1$$

State $s = -1$:

$$\pi^2(-1) = \text{argmax}\{0.8571, 0, 0.8571\} \to -1 \quad \text{(Tie-break)}$$

Resulting Policy:

$$\pi^2(s) = -1 \quad \forall s \in \mathcal{S}$$

1.4 Policy Evaluation (Calculation of ω^2)

Next actions are all -1.

$$A = \frac{1}{3}\Big[3(3 - 0.9(-1)) + 0(0 - 0.9(-1)) + (-1)(-1 - 0.9(-1))\Big]$$
$$= \frac{1}{3}\Big[3(3.9) + 0 + (-1)(-0.1)\Big] = \frac{1}{3}[11.7 + 0.1] = 3.9333$$

$$b = \frac{1}{3}\Big[3(1) + 0(-1) + (-1)(0)\Big] = 1$$

Solving for weights:

$$\omega^2 = A^{-1}b = \frac{1}{3.9333} = 0.2542$$

1.5 Derivation of Policy π^3

Update policy greedy w.r.t $Q(s,a) = \phi(s,a)(0.2542)$.

$$\pi^3(2) = \text{argmax}\{-0.2542, 0, 1.271\} = 1$$
$$\pi^3(1) = \text{argmax}\{-0.2542, 0, 0.7626\} = 1$$
$$\pi^3(-1) = \text{argmax}\{-0.2542, 0, -0.2542\} = 0$$

Conclusion: $\pi^3 = \pi^1$. The policy has converged.

Problem 2: Discrete LSPI (Vector Basis)

Setup:

- Same Data D.
- Basis Function: $\phi(s,a) = \begin{bmatrix} as+a \\ a^2 s \end{bmatrix}$.
- Initial Weights: $\boldsymbol{\omega}^0 = [1,1]^\top$.

2.1 Derivation of Policy π^1

$Q(s,a) = \phi(s,a)^\top \boldsymbol{\omega}^0 = (as+a) + a^2 s$. This functional form is identical to Problem 1.

$$\pi^1(s) = \{1, 1, 0\} \quad \text{for } s = \{2, 1, -1\}.$$

2.2 Policy Evaluation (Calculation of ω^1)

We compute the matrix $\mathbf{A}$ and vector $\mathbf{b}$.

$$\mathbf{A} = \frac{1}{L}\sum_{i=0}^{L-1} \phi_i(\phi_i - \gamma\phi_i')^\top, \quad \mathbf{b} = \frac{1}{L}\sum_{i=0}^{L-1} \phi_i r_{i+1}$$

Term 1: $s=1, a=1 \to s'=2, \pi(2)=1$.

$$\phi = \begin{bmatrix} 2 \\ 1 \end{bmatrix}, \quad \phi' = \begin{bmatrix} 3 \\ 2 \end{bmatrix}. \quad \text{Diff} = \begin{bmatrix} -0.7 \\ -0.8 \end{bmatrix}.$$

$$\phi(\text{Diff})^\top = \begin{bmatrix} -1.4 & -1.6 \\ -0.7 & -0.8 \end{bmatrix}.$$

Term 2: $s=2, a=0 \to s'=1, \pi(1)=1$.

$$\phi = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad \text{Contribution is } \mathbf{0}.$$

Term 3: $s=1, a=-1 \to s'=-1, \pi(-1)=0$.

$$\phi = \begin{bmatrix} -2 \\ 1 \end{bmatrix}, \quad \phi' = \begin{bmatrix} 0 \\ 0 \end{bmatrix}. \quad \text{Diff} = \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

$$\phi(\text{Diff})^\top = \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix}.$$

Aggregation:

$$\mathbf{A} = \frac{1}{3}\left(\begin{bmatrix} -1.4 & -1.6 \\ -0.7 & -0.8 \end{bmatrix} + \begin{bmatrix} 4 & -2 \\ -2 & 1 \end{bmatrix} \right) = \begin{bmatrix} 0.867 & -1.2 \\ -0.9 & 0.067 \end{bmatrix}.$$

$$\mathbf{b} = \frac{1}{3}\left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}(1) + \mathbf{0} + \begin{bmatrix} -2 \\ 1 \end{bmatrix}(0) \right) = \begin{bmatrix} 0.67 \\ 0.33 \end{bmatrix}.$$

Solving $\mathbf{A}\boldsymbol{\omega}^1 = \mathbf{b}$:

$$\boldsymbol{\omega}^1 = \mathbf{A}^{-1}\mathbf{b} \approx \begin{bmatrix} -0.43 \\ -0.87 \end{bmatrix}.$$

2.3 Derivation of Policy π^2

Maximize $Q(s,a) = -0.43(as+a) - 0.87(a^2 s)$. For $s=1$, $Q(1,0)=0$ and $Q(1,-1) \approx 0$. The tie is broken with -1.

Resulting Policy: $\pi^2 = \{0,-1,-1\}$.

2.4 Policy Evaluation (Calculation of ω^2)

Next actions: $\pi^2(2)=0, \pi^2(1)=-1, \pi^2(-1)=-1$.

Aggregation: *Note: For the third term,* $\pi(-1) = -1 \implies a=-1, s=-1$. *Thus* $\phi' = [as+a, a^2 s]^\top = [0,-1]^\top$.

$$\mathbf{A} = \frac{1}{3}\left[\begin{bmatrix}2\\1\end{bmatrix}\left(\begin{bmatrix}2\\1\end{bmatrix} - 0.9\mathbf{0}\right)^\top + \mathbf{0} + \begin{bmatrix}-2\\1\end{bmatrix}\left(\begin{bmatrix}-2\\1\end{bmatrix} - 0.9\begin{bmatrix}0\\-1\end{bmatrix}\right)^\top\right]$$
$$= \begin{bmatrix}2.67 & -0.6\\ 0 & 0.967\end{bmatrix}.$$
$$\mathbf{b} = \begin{bmatrix}0.67\\0.33\end{bmatrix}.$$

Solving $\mathbf{A}\boldsymbol{\omega}^2 = \mathbf{b}$:

$$\boldsymbol{\omega}^2 \approx \begin{bmatrix}0.33\\0.34\end{bmatrix}.$$

2.5 Derivation of Policy π^3

Maximize $Q(s,a) = 0.33(as+a) + 0.34(a^2 s)$. Since weights are positive, the logic mirrors π^1.

Resulting Policy: $\pi^3 = \{1,1,0\}$. This matches π^1.

Problem 3: Continuous State LSPI

Setup:

- Continuous States $\mathcal{S} \in [-2,2]$. Continuous Actions $\mathcal{A} \in [-1.5,2]$.
- Data: $D = \{(2,1,1,1),(1,-1,2,-1)\}$.
- Basis: $\phi(s,a) = [s, sa]^\top$. $\boldsymbol{\omega}^0 = [0.5,1]^\top$.

3.1 Derivation of Policy π^1

Maximize $Q(s,a) = 0.5s + 1(sa) = s(0.5+a)$ with respect to $a \in [-1.5,2]$.

- If $s>0$: Maximize $(0.5+a) \implies a=2$.
- If $s<0$: Minimize $(0.5+a)$ to make product positive $\implies a=-1.5$.

For Data points: $\pi^1(1)=2$, $\pi^1(-1)=-1.5$.

3.2 Policy Evaluation (Calculation of ω^1)

Aggregation:

$$\mathbf{A} = \frac{1}{2}\begin{bmatrix} 4.1 & -1.95 \\ 0.3 & 2.75 \end{bmatrix} = \begin{bmatrix} 2.05 & -0.975 \\ 0.15 & 1.375 \end{bmatrix}.$$
$$\mathbf{b} = \frac{1}{2}\Big([2,2]^\top(1) + [1,-1]^\top(2)\Big) = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

Solving $\mathbf{A}\boldsymbol{\omega}^1 = \mathbf{b}$ yields $\boldsymbol{\omega}^1 \approx [0.93, -0.10]^\top$.

3.3 Derivation of Policy π^2

Maximize $Q(s,a) = 0.93s - 0.10sa = s(0.93 - 0.10a)$.

- If $s > 0$: Maximize $(0.93 - 0.10a) \implies$ Min $a \implies a = -1.5$.
- If $s < 0$: Minimize $(0.93 - 0.10a) \implies$ Max $a \implies a = 2$.

For Data points: $\pi^2(1) = -1.5$, $\pi^2(-1) = 2$.

3.4 Policy Evaluation (Calculation of ω^2)

Aggregation: *Note: For the second term,* $\pi(-1) = 2 \implies sa = (-1)(2) = -2$. *Thus* $\boldsymbol{\phi}' = [-1,-2]^\top$.

$$\mathbf{A} = \frac{1}{2}\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix} (\begin{bmatrix} 2 \\ 2 \end{bmatrix} - 0.9 \begin{bmatrix} 1 \\ -1.5 \end{bmatrix})^\top + \begin{bmatrix} 1 \\ -1 \end{bmatrix} (\begin{bmatrix} 1 \\ -1 \end{bmatrix} - 0.9 \begin{bmatrix} -1 \\ -2 \end{bmatrix})^\top \right)$$
$$= \begin{bmatrix} 2.05 & 3.75 \\ 0.15 & 2.95 \end{bmatrix}.$$
$$\mathbf{b} = \begin{bmatrix} 2 \\ 0 \end{bmatrix}.$$

Solving for weights:

$$\boldsymbol{\omega}^2 \approx \begin{bmatrix} 1.076 \\ -0.055 \end{bmatrix}.$$

3.5 Interpretation of Final Policy π^3

Maximize $Q(s,a) = s(1.076 - 0.055a)$. The logic remains the same as π^2.

$$\pi^3(s) = \begin{cases} -1.5 & \text{if } s > 0 \\ 2 & \text{if } s < 0 \end{cases}$$