Problem 1: Tabular Q-Learning

System Dynamics:

- States $\mathcal{S} = \{A, B\}$. Actions $\mathcal{A} = \{a^1, a^2\}$.
- Parameters: $\alpha = 0.5$, $\gamma = 0.9$.
- Initial Q-Values: $Q(A, a^2) = 0.5$, $Q(B, a^1) = -0.1$, others 0.
- Policy: Greedy ($\epsilon = 0$).

The Q-Learning update rule is:

$$Q_{k+1}(S_t, A_t) \leftarrow Q_k(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_{a'} Q_k(S_{t+1}, a') - Q_k(S_t, A_t) \right]$$

Step-by-Step Execution

Step 1: State A

Current Q-Table: $Q(A, a^2) = 0.5$, $Q(B, a^1) = -0.1$. 1. **Select Action:** $\pi(A) = \text{argmax}(0, 0.5) = a^2$. 2. **Observe:** Transition $A \xrightarrow{a^2} B$, Reward $r = 1$. 3. **Update:**

$$\begin{aligned} \text{Target} &= r + \gamma \max_{a'} Q(B, a') \\ &= 1 + 0.9 \times \max(-0.1, 0) = 1 + 0 = 1. \\ Q_{new}(A, a^2) &= 0.5 + 0.5[1 - 0.5] \\ &= 0.5 + 0.25 = \mathbf{0.75}. \end{aligned}$$

Step 2: State B

Current Q-Table: $Q(A, a^2) \to 0.75$. 1. **Select Action:** $\pi(B) = \text{argmax}(-0.1, 0) = a^2$. 2. **Observe:** Transition $B \xrightarrow{a^2} A$, Reward $r = -1$. 3. **Update:**

$$\begin{aligned} \text{Target} &= -1 + 0.9 \max(Q(A, a^1), Q(A, a^2)) \\ &= -1 + 0.9 \max(0, 0.75) \\ &= -1 + 0.675 = -0.325. \\ Q_{new}(B, a^2) &= 0 + 0.5[-0.325 - 0] \\ &= \mathbf{-0.1625}. \end{aligned}$$

Step 3: State A

Current Q-Table: $Q(B, a^2) \to -0.1625$. 1. **Select Action:** $\pi(A) = \text{argmax}(0, 0.75) = a^2$. 2. **Observe:** Transition $A \xrightarrow{a^2} B$, Reward $r = 1$. 3. **Update:**

$$\begin{aligned} \text{Target} &= 1 + 0.9 \max(Q(B, a^1), Q(B, a^2)) \\ &= 1 + 0.9 \max(-0.1, -0.1625) \\ &= 1 + 0.9(-0.1) = 0.91. \\ Q_{new}(A, a^2) &= 0.75 + 0.5[0.91 - 0.75] \\ &= 0.75 + 0.08 = \mathbf{0.83}. \end{aligned}$$

Step 4: State B

Current Q-Table: $Q(A, a^2) \to 0.83$. 1. **Select Action:** $\pi(B) = \text{argmax}(-0.1, -0.1625) = a^1$. 2. **Observe:** Transition $B \xrightarrow{a^1} A$, Reward $r = 0$. 3. **Update:**

$$\begin{aligned} \text{Target} &= 0 + 0.9 \max(Q(A, a^1), Q(A, a^2)) \\ &= 0.9(0.83) = 0.747. \\ Q_{new}(B, a^1) &= -0.1 + 0.5[0.747 - (-0.1)] \\ &= -0.1 + 0.4235 = \mathbf{0.3235}. \end{aligned}$$

Step 5: State A

Current Q-Table: $Q(B, a^1) \to 0.3235$. 1. **Select Action:** $\pi(A) = \text{argmax}(0, 0.83) = a^2$. 2. **Observe:** Transition $A \xrightarrow{a^2} B$, Reward $r = 1$. 3. **Update:**

$$\begin{aligned} \text{Target} &= 1 + 0.9 \max(Q(B, a^1), Q(B, a^2)) \\ &= 1 + 0.9(0.3235) = 1.29115. \\ Q_{new}(A, a^2) &= 0.83 + 0.5[1.29115 - 0.83] \\ &= 0.83 + 0.230575 = \mathbf{1.0606}. \end{aligned}$$

Final Q-Values and Policy

State	$Q(S, a^1)$	$Q(S, a^2)$	$\pi^*(S)$
A	0	**1.0606**	a^2
B	**0.3235**	-0.1625	a^1

Problem 2: Grid World Q-Learning Analysis

a. Detailed Update Calculations

Setup:

- $Q(s, a)$ initialized to 0. Step reward $r = -1$. Goal reward $R_G \approx 100$.

  States 5, 6, 7 (Path to Goal):

$$\begin{aligned} Q(5, U) &\leftarrow 0 + 0.5[-1 + 0.9(0) - 0] = -\mathbf{0.5}. \\ Q(6, U) &\leftarrow 0 + 0.5[-1 + 0.9(0) - 0] = -\mathbf{0.5}. \\ Q(7, U) &\leftarrow 0 + 0.5[-1 + 0.9(0) - 0] = -\mathbf{0.5} \quad \text{(Hit Wall)}. \end{aligned}$$

  State 9 (Transition to Goal): Action D leads to state 11 (Goal).

$$\begin{aligned} Q(9, D) &\leftarrow 0 + 0.5[-1 + 100 + 0.9(0) - 0] \\ &= 0.5[99] = \mathbf{49.5}. \end{aligned}$$

  State 10 (Correction from 9): Agent moves L to 9, utilizing the new $Q(9, D)$.

$$\begin{aligned} Q(10, L) &\leftarrow 0 + 0.5[-1 + 0.9(49.5) - 0] \\ &= 0.5[-1 + 44.55] = 0.5[43.55] = \mathbf{21.775}. \end{aligned}$$

b. Resulting Policy Trajectory

7 (D)	8 (D)	9 (D)	10 (L)
6 (R)		**11 (G)**	12 (L)
5 (R)		13 (U)	14 (U)
4 (U)			
3 (U)	2 (U)	1 (U)	

Problem 3: Actor-Critic (One-Step TD)

Setup:

- Critic $V(s)$: $\alpha = 0.5$. Actor $H(s,a)$: $\beta = 0.1$.
- Softmax Policy: $\pi(a|s) = \frac{e^{H(s,a)}}{\sum_b e^{H(s,b)}}$.
- Update: $H(s,a) \leftarrow H(s,a) + \beta\delta[1 - \pi(a|s)]$ (for taken action).

Episode Trace

Step 1: $A \xrightarrow{a^1} A$**,** $r = 10$

$\pi(a^1|A) = 0.5$.

$$\delta = 10 + 0.9(0) - 0 = 10.$$
$$V(A) \leftarrow 0.5(10) = \mathbf{5}.$$
$$H(A, a^1) \leftarrow 0 + 0.1(10)(0.5) = \mathbf{0.5}.$$

Step 2: $A \xrightarrow{a^2} B$**,** $r = -5$

$\pi(a^1|A) \approx 0.6225, \pi(a^2|A) \approx 0.3775$.

$$\delta = -5 + 0.9(0) - 5 = -10.$$
$$V(A) \leftarrow 5 + 0.5(-10) = \mathbf{0}.$$
$$H(A, a^2) \leftarrow 0 + 0.1(-10)(1 - 0.3775) = -\mathbf{0.6225}.$$

Step 3: $B \xrightarrow{a^1} A$**,** $r = 40$

$\pi(a^1|B) = 0.5$.

$$\delta = 40 + 0.9(0) - 0 = 40.$$
$$V(B) \leftarrow 0.5(40) = \mathbf{20}.$$
$$H(B, a^1) \leftarrow 0 + 0.1(40)(0.5) = \mathbf{2}.$$

Step 4: $A \xrightarrow{a^2} A$, $r = -5$

$\pi(a^2|A) \approx 0.2456$.

$$\begin{aligned}
\delta &= -5 + 0.9(0) - 0 = -5. \\
V(A) &\leftarrow 0 + 0.5(-5) = \mathbf{-2.5}. \\
H(A, a^2) &\leftarrow -0.6225 + 0.1(-5)(0.7544) = \mathbf{-0.9997}.
\end{aligned}$$

Step 5: $A \xrightarrow{a^2} A$, $r = 20$

$\pi(a^2|A) \approx 0.1824$.

$$\begin{aligned}
\delta &= 20 + 0.9(-2.5) - (-2.5) = 20.25. \\
V(A) &\leftarrow -2.5 + 0.5(20.25) = \mathbf{7.625}. \\
H(A, a^2) &\leftarrow -0.9997 + 0.1(20.25)(0.8176) = \mathbf{0.6559}.
\end{aligned}$$

Step 6: $A \xrightarrow{a^1} A$, $r = 10$

$\pi(a^1|A) \approx 0.4611$.

$$\begin{aligned}
\delta &= 10 + 0.9(7.625) - 7.625 = 9.2375. \\
V(A) &\leftarrow 7.625 + 0.5(9.2375) = \mathbf{12.244}. \\
H(A, a^1) &\leftarrow 0.5 + 0.1(9.2375)(0.5389) = \mathbf{0.9978}.
\end{aligned}$$

Final Convergence Check

$$\pi(a^1|A) = \frac{e^{0.9978}}{e^{0.9978} + e^{0.6559}} \approx \mathbf{0.585}$$

$$\pi(a^1|B) = \frac{e^2}{e^2 + e^0} \approx \mathbf{0.881}$$

Greedy Policy: $\pi^*(A) = a^1, \pi^*(B) = a^1$.