Reinforcement Learning: Problems & Detailed Solutions

2025-09-20

Contents

Notation

States $s \in \mathcal{S}$, actions $a \in \mathcal{A}(s)$, rewards $r \in \mathbb{R}$, discount $\gamma \in [0, 1)$.
State-value $V(s)$, action-value $Q(s, a)$. For episodic tasks we take $V(\text{terminal}) = 0$ and $Q(\text{terminal}, \cdot) = 0$.

1 Bandits — ε-Greedy with Sample-Average Updates [20 pts]

Problem

You are A/B testing two ad creatives. Each impression yields reward 1 (click) or 0 (no click).

- Arm A: Bernoulli($p_A = 0.7$), Arm B: Bernoulli($p_B = 0.5$).
- Use ε-greedy with $\varepsilon = 0.1$ and sample-average action-value estimates. Initialize $Q_1(A) = Q_1(B) = 0$ (break ties uniformly).

(a) Write pseudocode for ε-greedy (sample-average updates).

(b) Given the action–reward sequence over 6 steps: $(A, 1), (B, 1), (A, 0), (A, 1), (B, 0), (A, 1)$, update $Q(A), Q(B)$ after each step and report N and Q.

(c) Using estimates after step 6, compute Pr(choose A at step 7) under ε-greedy.

(d) Compute the cumulative reward over the 6 steps and the (expected) regret relative to always pulling the optimal arm for 6 steps (use $\mu^* = \max\{p_A, p_B\}$).

Detailed Solution

Algorithm 1 ε-greedy k-armed bandit (sample-average updates)

Require: $\varepsilon \in [0, 1]$
Initialize $Q(a) \leftarrow 0$, $N(a) \leftarrow 0$ for all $a \in \{1, \ldots, k\}$
for $t = 1, 2, \ldots$ **do**
  With probability ε: choose A uniformly from arms
  Else: $A \leftarrow \arg\max_a Q(a)$ ▷ break ties uniformly
  Pull A, observe reward R_t
  $N(A) \leftarrow N(A) + 1$
  $Q(A) \leftarrow Q(A) + \dfrac{1}{N(A)}\big(R_t - Q(A)\big)$

(a) Pseudocode (incremental mean).

(b) Step-by-step updates. The incremental mean after the n-th observation of arm a is $Q_n \leftarrow Q_{n-1} + \frac{1}{n}(R_n - Q_{n-1})$.

t	(Action, Reward)	$N(A)$	$N(B)$	$Q(A)$	$Q(B)$
1	(A,1)	1	0	$0 + \frac{1}{1}(1-0) = \mathbf{1.0}$	0.0
2	(B,1)	1	1	1.0	$0 + \frac{1}{1}(1-0) = \mathbf{1.0}$
3	(A,0)	2	1	$1 + \frac{1}{2}(0-1) = \mathbf{0.5}$	1.0
4	(A,1)	3	1	$0.5 + \frac{1}{3}(1-0.5) = \mathbf{0.6667}$	1.0
5	(B,0)	3	2	0.6667	$1 + \frac{1}{2}(0-1) = \mathbf{0.5}$
6	(A,1)	4	2	$0.6667 + \frac{1}{4}(1-0.6667) = \mathbf{0.75}$	0.5

(c) Step 7 selection probability. Greedy action is A (since $0.75 > 0.5$). With two arms,

$$\Pr(A) = (1-\varepsilon) + \varepsilon/2 = 0.9 + 0.05 = \boxed{0.95}.$$

(d) Reward and regret. Observed reward $= 1+1+0+1+0+1 = 4$. Optimal mean $\mu^* = 0.7$; always-optimal expected reward in 6 steps $= 6 \times 0.7 = 4.2$. Regret $= 4.2 - 4 = \boxed{0.2}$.

Reference: Sutton & Barto (2018), Ch. 2.

2 Bandits — UCB1 (Upper Confidence Bound) [15 pts]

Problem

Three-armed Bernoulli bandit with unknown means $\mu_A = 0.50,\ \mu_B = 0.60,\ \mu_C = 0.40$. Use UCB1 $\mathrm{UCB}_t(a) = \hat{Q}_t(a) + \sqrt{\frac{2\ln t}{N_t(a)}}$, initializing by pulling each arm once (order A, B, C) and observing $R_1(A)$=1, $R_2(B)$=0, $R_3(C)$=1.

(a) Write pseudocode for UCB1.

(b) Compute choices at $t = 4, 5, 6, 7$ given realized rewards: R_4=1, R_5=0, R_6=0, R_7=1 (use $\ln 4$=1.3863, $\ln 5$=1.6094, $\ln 6$=1.7918, $\ln 7$=1.9459).

(c) Give cumulative reward by $t = 7$ and expected regret vs. always pulling B.

Detailed Solution

Algorithm 2 UCB1

Pull each arm once; set $N(a) = 1$, $Q(a) =$ observed mean
for $t = k+1, k+2, \ldots$ **do**
  for each arm a **do**
    $\mathrm{UCB}(a) \leftarrow Q(a) + \sqrt{\frac{2\ln t}{N(a)}}$
  Play $A_t = \arg\max_a \mathrm{UCB}(a)$; observe R_t
  $N(A_t) \leftarrow N(A_t) + 1;\ Q(A_t) \leftarrow Q(A_t) + \dfrac{1}{N(A_t)}(R_t - Q(A_t))$

(a) Pseudocode.

(b) Numeric UCBs and updates. After inits: $N = (1, 1, 1)$, $Q = (1.0, 0.0, 1.0)$.

- $t = 4$: $\sqrt{2\ln 4} \approx 1.665$.

  $\mathrm{UCB}(A) = 1 + 1.665 = 2.665, \quad \mathrm{UCB}(B) = 0 + 1.665 = 1.665, \quad \mathrm{UCB}(C) = 1 + 1.665 = 2.665.$

  Tie $\Rightarrow$ choose A. Reward 1. Update $N(A) = 2$, $Q(A) = \frac{1+1}{2} = 1.0$.
- $t = 5$: $\sqrt{2\ln 5} \approx 1.793$, $\sqrt{(2\ln 5)/2} \approx 1.269$.

  $\mathrm{UCB}(A) = 1 + 1.269 = 2.269, \quad \mathrm{UCB}(B) = 0 + 1.793 = 1.793, \quad \mathrm{UCB}(C) = 1 + 1.793 = 2.793.$

  Choose C. Reward 0. Update $N(C) = 2$, $Q(C) = \frac{1+0}{2} = 0.5$.
- $t = 6$: $\sqrt{2\ln 6} \approx 1.893$, $\sqrt{(2\ln 6)/2} \approx 1.338$.

  $\mathrm{UCB}(A) = 1 + 1.338 = 2.338, \quad \mathrm{UCB}(B) = 0 + 1.893 = 1.893, \quad \mathrm{UCB}(C) = 0.5 + 1.338 = 1.838.$

  Choose A. Reward 0. Update $N(A) = 3$, $Q(A) = \frac{1+1+0}{3} = 0.667$.
- $t = 7$: $\sqrt{2\ln 7} \approx 1.973$, $\sqrt{(2\ln 7)/3} \approx 1.139$, $\sqrt{(2\ln 7)/2} \approx 1.395$.

  $\mathrm{UCB}(A) = 0.667 + 1.139 = 1.806, \quad \mathrm{UCB}(B) = 0 + 1.973 = 1.973, \quad \mathrm{UCB}(C) = 0.5 + 1.395 = 1.895.$

  Choose B. Reward 1.

(c) Totals. Chosen arms: A, C, A, B with rewards $1, 0, 0, 1$. Including the first three pulls (rewards $1, 0, 1$), total reward $= 4$. Regret vs. always B: $7 \cdot 0.6 - 4 = \boxed{0.2}$.

Reference: Sutton & Barto (2018), §2.7; Auer et al. (2002).

3 MDP (Inventory) — Laddu Shop [25 pts]

Problem (clarified statement)

Demand $D \in \{1, 2, 3\}$ with $P(1) = 0.2,\ P(2) = 0.5,\ P(3) = 0.3$). The morning state $s \in \{0, 1, 2\}$ is the number of leftover batches from the previous day. Upon choosing action $a \in \{1, 2\}$ (batches prepared now), the available inventory for the day is

$$I = s + a.$$

During the day, stochastic demand D realizes. Sales equal $\min(D, I)$; unused inventory $L = \max(0, I - D)$ is carried to the next morning but capped at 2 to keep the state space closed:

$$s' = \min\{2,\ L\}.$$

Per-day profit is

$$r = 2500 \cdot \min(D, I) \ - \ 500 \cdot a \ - \ 100 \cdot L.$$

(a) Write down the transition probabilities $P(s'|s, a)$ for all $s \in \{0, 1, 2\}$ and $a \in \{1, 2\}$.

(b) Compute the expected one-step reward $R(s, a) = \mathbb{E}[r \mid s, a]$ for all (s, a).

Detailed Solution

(a) Transitions. For each (s, a), set $I = s + a$ and enumerate demand outcomes:

- $a = 1$: $I = 1, 2, 3$ for $s = 0, 1, 2$ respectively, giving

$$P^{(a=1)} = \begin{bmatrix} 1 & 0 & 0 \\ 0.8 & 0.2 & 0 \\ 0.3 & 0.5 & 0.2 \end{bmatrix} \quad \text{(rows } s\text{, columns } s'\text{)}.$$

- $a = 2$: $I = 2, 3, 4$ for $s = 0, 1, 2$:

$$P^{(a=2)} = \begin{bmatrix} 0.8 & 0.2 & 0 \\ 0.3 & 0.5 & 0.2 \\ 0 & 0.3 & 0.7 \end{bmatrix}.$$

The entries follow from $L = \max(0, I - D)$ mapped to $s' = \min\{2, L\}$ (e.g., s=2, a=2, I=4: if $D = 1, 2, 3$ we get $L = 3, 2, 1$ hence $s' = \{2, 2, 1\}$ with probabilities $0.2, 0.5, 0.3$).

(b) Expected rewards. For each (s, a):

$$R(s, a) = \sum_{d \in \{1,2,3\}} p(d)\big(2500 \min(d, I) - 500a - 100 \max(0, I - d)\big).$$

Examples (all amounts in INR):

$$R(0,2):\ I = 2 \ \Rightarrow \ \begin{cases} d = 1: & 2500(1) - 500(2) - 100(1) = 1400, \\ d = 2: & 2500(2) - 1000 - 0 = 4000, \\ d = 3: & 2500(2) - 1000 - 0 = 4000, \end{cases}$$
$$\Rightarrow \ R = 0.2 \cdot 1400 + 0.5 \cdot 4000 + 0.3 \cdot 4000 = \boxed{3480}.$$

$$R(1,1):\ I = 2 \ \Rightarrow \ \begin{cases} d = 1: & 2500(1) - 500(1) - 100(1) = 1900, \\ d = 2: & 5000 - 500 - 0 = 4500, \\ d = 3: & 5000 - 500 - 0 = 4500, \end{cases} \quad \Rightarrow R = \boxed{3980}.$$

$$R(2,1):\ I = 3 \ \Rightarrow \ \begin{cases} d = 1: & 2500 - 500 - 200 = 1800, \\ d = 2: & 5000 - 500 - 100 = 4400, \\ d = 3: & 7500 - 500 - 0 = 7000, \end{cases} \Rightarrow R = 0.2 \cdot 1800 + 0.5 \cdot 4400 + 0.3 \cdot 7000 = \boxed{4660}.$$

The full table:

s	a	I	$R(s, a)$
0	1	1	2000
0	2	2	3480
1	1	2	3980
1	2	3	4160
2	1	3	4660
2	2	4	4060

4 DP — Value Iteration (Mini-MDP) [20 pts]

Problem (clarified statement)

Discount $\gamma = 0.9$. States $\{s_1, s_2, s_3\}$ where s_3 is *terminal* (no actions, zero future value).
Actions and deterministic dynamics (explicit):

- At s_1:
  - a_1: *Immediate reward* $+2$, then transition deterministically to s_2.
  - a_2: *Immediate reward* $+1$, then transition deterministically to s_1 (self-loop).
- At s_2:
  - b_1: *Immediate reward* $+5$, then transition deterministically to s_3 (terminal).
  - b_2: *Immediate reward* 0; with probability 0.5 go to s_1, with probability 0.5 stay in s_2.

Task. Run *three full sweeps* of value iteration starting from $V_0 \equiv 0$. After each sweep k, report all state action-values $Q_k(s,a) = r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_{k-1}(s')$ and then $V_k(s) = \max_a Q_k(s,a)$. Finally determine the optimal policy π^* and optimal values V^*.

Detailed Solution

The Bellman optimality update is $V_k(s) = \max_a \left[r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_{k-1}(s') \right]$.

Sweep 1 (using $V_0 \equiv 0$).

$$Q_1(s_1,a_1) = 2 + 0.9\, V_0(s_2) = 2, \qquad Q_1(s_1,a_2) = 1 + 0.9\, V_0(s_1) = 1,$$
$$Q_1(s_2,b_1) = 5 + 0.9\, V_0(s_3) = 5, \qquad Q_1(s_2,b_2) = 0 + 0.9\,[0.5 V_0(s_1) + 0.5 V_0(s_2)] = 0.$$

Thus $V_1(s_1) = \max\{2,1\} = \mathbf{2}$, $V_1(s_2) = \max\{5,0\} = \mathbf{5}$, $V_1(s_3) = 0$.

Sweep 2 (use V_1).

$$\begin{aligned} Q_2(s_1,a_1) &= 2 + 0.9\, V_1(s_2) = 2 + 0.9 \cdot 5 = 6.5, \\ Q_2(s_1,a_2) &= 1 + 0.9\, V_1(s_1) = 1 + 0.9 \cdot 2 = 2.8, \\ Q_2(s_2,b_1) &= 5 + 0.9\, V_1(s_3) = 5, \\ Q_2(s_2,b_2) &= 0 + 0.9\,[0.5 \cdot V_1(s_1) + 0.5 \cdot V_1(s_2)] = 0.9 \cdot 0.5(2+5) = 3.15. \end{aligned}$$

Therefore $V_2(s_1) = \max\{6.5, 2.8\} = \mathbf{6.5}$, $V_2(s_2) = \max\{5, 3.15\} = \mathbf{5}$, $V_2(s_3) = 0$.

Sweep 3 (use V_2).

$$\begin{aligned} Q_3(s_1,a_1) &= 2 + 0.9\, V_2(s_2) = 6.5, \\ Q_3(s_1,a_2) &= 1 + 0.9\, V_2(s_1) = 1 + 0.9 \cdot 6.5 = 6.85, \\ Q_3(s_2,b_1) &= 5 + 0.9\, V_2(s_3) = 5, \\ Q_3(s_2,b_2) &= 0 + 0.9\,[0.5 \cdot V_2(s_1) + 0.5 \cdot V_2(s_2)] = 0.9 \cdot 0.5(6.5+5) = 5.175. \end{aligned}$$

Hence $V_3(s_1) = \max\{6.85, 6.5\} = \mathbf{6.85}$, $V_3(s_2) = \max\{5, 5.175\} = \mathbf{5.175}$, $V_3(s_3) = 0$.

Optimal policy and values. Notice b_1 at s_2 yields 5 while b_2 yields at most 5.175 at this sweep; but the true optimal $V^*(s_2)$ is obtained by comparing *fixed points*. For s_2 the action b_1 terminates with $+5$, so $V^*(s_2) \geq 5$. For s_1, the self-loop a_2 has fixed point

$$V(s_1) = 1 + 0.9\, V(s_1) \;\Rightarrow\; V^*(s_1) = \frac{1}{1-0.9} = \boxed{10}.$$

Thus $Q^*(s_1, a_1) = 2 + 0.9 \cdot V^*(s_2) = 2 + 0.9 \cdot 5 = 6.5 < 10$, so $\pi^*(s_1) = a_2$. At s_2, b_1 is optimal because b_2's value cannot exceed $0.9 \max\{V(s_1), V(s_2)\} < 0.9 \cdot 10 = 9$ without extra rewards; with zero immediate reward and mixing back to s_1/s_2, its best achievable fixed point under $\gamma = 0.9$ is still dominated by b_1's sure $+5$ followed by termination in this construction. Hence

$$\boxed{\pi^*(s_1) = a_2, \quad \pi^*(s_2) = b_1, \quad V^*(s_1) = 10,\ V^*(s_2) = 5,\ V^*(s_3) = 0.}$$

Reference: Sutton & Barto (2018), Ch. 4; Puterman (1994), Ch. 2.

5 DP — Policy Iteration (Mini-MDP) [20 pts]

Problem

Same state/action setup as above but with stochastic self-loop at s_1 and recycling at s_2:

- s_1: a (to s_2 with $+2$), b (reward 0; stay s_1 w.p. 0.8, go s_3 w.p. 0.2).
- s_2: c (to s_3 with $+4$), d (reward $+1$; to s_1 w.p. 0.5, stay s_2 w.p. 0.5).

Discount $\gamma = 0.9$. Start with policy $\pi_0(s_1)$=a, $\pi_0(s_2)$=c. Perform exact policy evaluation and policy improvement until convergence. Report π^* and V^*.

Detailed Solution

Evaluate π_0.

$$V_0(s_2) = 4 + 0.9 \cdot V_0(s_3) = 4,$$
$$V_0(s_1) = 2 + 0.9\, V_0(s_2) = 2 + 0.9 \cdot 4 = \boxed{5.6}.$$

Improve $\pi_0 \to \pi_1$.

$$\begin{aligned}
Q(s_1, a) &= 2 + 0.9\, V_0(s_2) = 5.6, \\
Q(s_1, b) &= 0 + 0.9\, [0.8 V_0(s_1) + 0.2 V_0(s_3)] = 0.72 \cdot 5.6 = 4.032 \Rightarrow \text{keep } a. \\
Q(s_2, c) &= 4, \\
Q(s_2, d) &= 1 + 0.9\, [0.5 V_0(s_1) + 0.5 V_0(s_2)] \\
&= 1 + 0.9 \cdot 0.5(5.6 + 4) = \boxed{5.32} \Rightarrow \text{switch to } d.
\end{aligned}$$

Thus $\pi_1(s_1) = a,\ \pi_1(s_2) = d$.

Evaluate π_1. Solve

$$V_1(s_1) = 2 + 0.9\, V_1(s_2),$$
$$V_1(s_2) = 1 + 0.9\, [0.5 V_1(s_1) + 0.5 V_1(s_2)].$$

From the second: $0.55V_1(s_2) = 1 + 0.45V_1(s_1) \Rightarrow V_1(s_2) = \frac{1+0.45V_1(s_1)}{0.55}$. Plug into the first:

$$\begin{aligned} V_1(s_1) &= 2 + 0.9 \cdot \frac{1 + 0.45V_1(s_1)}{0.55} = 2 + 1.63636(1 + 0.45V_1(s_1)) \\ &= 3.63636 + 0.73636V_1(s_1) \;\Rightarrow\; (1 - 0.73636)V_1(s_1) = 3.63636 \\ &\Rightarrow\; V_1(s_1) = \boxed{13.793}, \quad V_1(s_2) = \frac{1 + 0.45 \cdot 13.793}{0.55} = \boxed{13.103}. \end{aligned}$$

Improve π_1.

$$\begin{aligned} &Q(s_1, a) = 2 + 0.9 \cdot 13.103 = 13.793, \\ &Q(s_1, b) = 0.9 \cdot (0.8 \cdot 13.793) = 9.933 \Rightarrow \text{keep } a. \\ &Q(s_2, c) = 4, \\ &Q(s_2, d) = 1 + 0.9 \cdot 0.5(13.793 + 13.103) = 13.103 \Rightarrow \text{keep } d. \end{aligned}$$

Policy stable $\Rightarrow \boxed{\pi^*(s_1) = a,\ \pi^*(s_2) = d}$ with values above.

6 Monte Carlo — First-Visit Prediction [20 pts]

Problem

States $\{A, B, C\}$ and terminal T, discount $\gamma = 0.9$. Three episodes observed:

$$\begin{aligned} &\text{E1: } A \xrightarrow{+2} B \xrightarrow{-1} C \xrightarrow{+3} T, \\ &\text{E2: } B \xrightarrow{+0} A \xrightarrow{+2} B \xrightarrow{+2} T, \\ &\text{E3: } C \xrightarrow{+1} C \xrightarrow{-2} A \xrightarrow{+4} T. \end{aligned}$$

Use first-visit MC to estimate $V(A), V(B), V(C)$ (explicitly show each return).

Detailed Solution

For a first visit at time t, $G_t = \sum_{k \geq 1} \gamma^{k-1} R_{t+k}$.

- E1: $G(A) = 2 + 0.9(-1) + 0.9^2 \cdot 3 = \boxed{3.53}$; $G(B) = -1 + 0.9 \cdot 3 = \boxed{1.7}$; $G(C) = \boxed{3.0}$.
- E2: $G(B) = 0 + 0.9 \cdot 2 + 0.9^2 \cdot 2 = \boxed{3.42}$; $G(A) = 2 + 0.9 \cdot 2 = \boxed{3.8}$; (the later first-visit for B) $G(B) = \boxed{2.0}$.
- E3: $G(C) = 1 + 0.9(-2) + 0.9^2 \cdot 4 = \boxed{2.44}$; $G(A) = \boxed{4.0}$.

Averages: $V(A) = \frac{3.53+3.8+4}{3} = \boxed{3.777}$, $V(B) = \frac{1.7+3.42}{2} = \boxed{2.56}$, $V(C) = \frac{3+2.44}{2} = \boxed{2.72}$.

7 Monte Carlo — On-Policy MC Control (ε-soft) [20 pts]

Problem

States $\{A, B, C\}$, actions $\{L, R\}$, terminal T, $\gamma = 0.9$. Three episodes:

$$\begin{aligned} &\text{E1: } (A, L) \xrightarrow{+2} B,\ (B, R) \xrightarrow{+1} C,\ (C, L) \xrightarrow{+3} T, \\ &\text{E2: } (B, L) \xrightarrow{0} A,\ (A, R) \xrightarrow{+2} B,\ (B, R) \xrightarrow{+2} T, \\ &\text{E3: } (C, R) \xrightarrow{+1} C,\ (C, L) \xrightarrow{-2} A,\ (A, L) \xrightarrow{+4} T. \end{aligned}$$

(1) Compute first-visit $Q(s,a)$ sample averages. (2) With $\varepsilon = 0.1$ and $|\mathcal{A}| = 2$, perform ε-greedy improvement.

Detailed Solution

First-visit returns (explicit):

$$(A,L): 2 + 0.9 \cdot 1 + 0.9^2 \cdot 3 = \boxed{5.33}, \quad (A,R): 2 + 0.9 \cdot 2 = \boxed{3.80},$$
$$(B,L): 0 + 0.9 \cdot 2 + 0.9^2 \cdot 2 = \boxed{3.42}, \quad (B,R): 1 + 0.9 \cdot 3 = \boxed{3.70}, \quad \boxed{2.00},$$
$$(C,L): \boxed{3.00}, \ \boxed{1.60}, \qquad (C,R): 1 + 0.9(-2) + 0.9^2 \cdot 4 = \boxed{2.44}.$$

Averages: $(A,L) = \boxed{4.665}$, $(A,R) = \boxed{3.80}$, $(B,L) = \boxed{3.42}$, $(B,R) = \boxed{2.85}$, $(C,L) = \boxed{2.30}$, $(C,R) = \boxed{2.44}$. Greedy: A:L, B:L, C:R. With $\varepsilon = 0.1$ and 2 actions: $\pi'(\text{greedy}|s) = \boxed{0.95}$, $\pi'(\text{other}|s) = \boxed{0.05}$.

8 TD — SARSA (On-Policy Control) [20 pts]

Problem

Episodic MDP with $\gamma = 0.9$, $\alpha = 0.5$, actions $\{U, R\}$. Trajectory: $(s_0, U) \xrightarrow{0} s_1$, $(s_1, U) \xrightarrow{2} s_2$, $(s_2, U) \xrightarrow{-1} T$. Initialize $Q \equiv 0$. Perform the three SARSA updates *showing each target explicitly* and give ε-greedy action probabilities ($\varepsilon = 0.1$).

Detailed Solution

Update rule: $Q \leftarrow Q + \alpha\big(R + \gamma Q(S', A') - Q\big)$.

$$t{=}0: Q(s_0,U) \leftarrow 0 + 0.5\big(0 + 0.9 \cdot Q(s_1,U) - 0\big) = 0,$$
$$t{=}1: Q(s_1,U) \leftarrow 0 + 0.5\big(2 + 0.9 \cdot Q(s_2,U) - 0\big) = 1.0,$$
$$t{=}2: Q(s_2,U) \leftarrow 0 + 0.5\big(-1 + 0 - 0\big) = -0.5.$$

ε-greedy with 2 actions: at s_1 greedy $U \Rightarrow (0.95, 0.05)$; at s_2 greedy $R \Rightarrow (0.95, 0.05)$; at s_0 tie (split or random tiebreak).

9 TD — Q-Learning (Off-Policy Control) [20 pts]

Problem

Same MDP, $\gamma = 0.9$, $\alpha = 0.5$. Episode 1 identical to SARSA; Episode 2: $(s_0, U) \xrightarrow{0} s_1$, $(s_1, R) \xrightarrow{0} s_1$, $(s_1, U) \xrightarrow{2} s_2$, $(s_2, U) \xrightarrow{-1} T$. Show each target.

Detailed Solution

Q-learning: $Q \leftarrow Q + \alpha\big(R + \gamma \max_a Q(S', a) - Q\big)$.

- After Ep. 1: $Q(s_0,U) = 0$, $Q(s_1,U) = 1.0$, $Q(s_2,U) = -0.5$.
- Ep. 2, $t = 0$: target $0 + 0.9 \max\{Q(s_1,U){=}1.0, Q(s_1,R){=}0\} = 0.9 \Rightarrow Q(s_0,U) = 0.45$.
- $t = 1$: target $0 + 0.9 \max\{1.0, 0\} = 0.9 \Rightarrow Q(s_1,R) = 0.45$.

- $t = 2$: target $2 + 0.9 \max\{Q(s_2, U) = -0.5, Q(s_2, R) = 0\} = 2 \Rightarrow Q(s_1, U) = 1.5$.
- $t = 3$: target $-1 + 0 = -1 \Rightarrow Q(s_2, U) = -0.75$.

Greedy: U at s_0, s_1; R at s_2.

10 TD — n-Step TD Prediction (Forward View, n=3) [20 pts]

Problem

Evaluate π with $\gamma = 0.9$, $\alpha = 0.5$, $n = 3$ on episode $A \to^{+1} B \to^{+2} B \to^{-1} C \to^{+2} T$. Initialize $V(A) = V(B) = V(C) = 0$. Compute updates for $\tau = 0, 1, 2, 3$ and show each $G_\tau^{(3)}$.

Detailed Solution

$$
\begin{aligned}
G_0^{(3)} &= 1 + 0.9 \cdot 2 + 0.9^2(-1) + 0.9^3 V(C) = \boxed{1.99} \Rightarrow V(A) = 0.995, \\
G_1^{(3)} &= 2 + 0.9(-1) + 0.9^2 \cdot 2 = \boxed{2.72} \Rightarrow V(B) = 1.36, \\
G_2^{(3)} &= -1 + 0.9 \cdot 2 = \boxed{0.8} \Rightarrow V(B) = 1.08, \\
G_3^{(3)} &= 2 = \boxed{2.0} \Rightarrow V(C) = 1.0.
\end{aligned}
$$

11 TD — n-Step SARSA (On-Policy, n=2) [20 pts]

Problem

Discount $\gamma = 0.9$, $\alpha = 0.5$, two states s_0, s_1, terminal T, actions $\{U, R\}$. Dynamics: s_0: $U : +1 \to s_1$, $R : 0 \to s_0$; s_1: $U : +2 \to s_1$, $R : 0 \to T$. Episode: $(s_0, U) \to^{+1} s_1$, $(s_1, U) \to^{+2} s_1$, $(s_1, R) \to^{0} T$. Compute n=2 SARSA updates for $\tau = 0, 1, 2$ showing each target.

Detailed Solution

$$
\begin{aligned}
G_0^{(2)} &= 1 + 0.9 \cdot 2 + 0.9^2 Q(s_1, R) = \boxed{2.8} \Rightarrow Q(s_0, U) = 1.4, \\
G_1^{(2)} &= 2 + 0.9 \cdot 0 = \boxed{2.0} \Rightarrow Q(s_1, U) = 1.0, \\
G_2^{(2)} &= 0 = \boxed{0} \Rightarrow Q(s_1, R) = 0.
\end{aligned}
$$

With $\varepsilon = 0.1$: greedy U at both s_0, s_1 (0.95/0.05).

12 Bayesian Bandit — Beta–Bernoulli & Thompson Sampling [20 pts]

Problem

Independent Beta priors on click rates. Observations: A has s_A=5, f_A=3; B has s_B=3, f_B=1. Priors: (a) uninformative Beta(1, 1), (b) regularized Beta(2, 2).

(a) Derive posteriors and posterior means for each arm.

(b) Under Bayes-greedy (pick larger posterior mean), which arm is chosen and what is the posterior-predictive success probability?

(c) One Thompson step under (a): suppose samples $\tilde{\theta}_A$=0.58, $\tilde{\theta}_B$=0.72. Which arm and expected reward?

Detailed Solution

For Beta–Bernoulli, $\theta \mid \text{data} \sim \text{Beta}(\alpha_0 + s, \beta_0 + f)$ with mean $\frac{\alpha_0+s}{\alpha_0+\beta_0+s+f}$.

- (a) Beta$(1,1)$: $A \sim \text{Beta}(6,4)$ mean 0.60; $B \sim \text{Beta}(4,2)$ mean 0.6667.
- (b) Beta$(2,2)$: $A \sim \text{Beta}(7,5)$ mean $7/12 \approx 0.5833$; $B \sim \text{Beta}(5,3)$ mean 0.625.

Bayes-greedy picks B in both cases (predictive success equals the posterior mean).
Thompson step (samples given): pick B; expected instantaneous reward $= \tilde{\theta}_B = \boxed{0.72}$.

References: Sutton & Barto (2018), Ch. 2,4–7; Puterman (1994), Ch. 2.