INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Stamp / Signature of the Invigilator

EXAMINATION (End Semester)	**SEMESTER (Autumn 2023-2024)**

Roll Number										**Section**		**Name**	

Subject Number	C	S	6	0	0	7	7	**Subject Name**	*REINFORCEMENT LEARNING*

Department / Center of the Student		**Additional sheets**	

Important Instructions and Guidelines for Students

1. You must occupy your seat as per the Examination Schedule/Sitting Plan.
2. Do not keep mobile phones or any similar electronic gadgets with you even in the switched off mode.
3. Loose papers, class notes, books or any such materials must not be in your possession, even if they are irrelevant to the subject you are taking examination.
4. Data book, codes, graph papers, relevant standard tables/charts or any other materials are allowed only when instructed by the paper-setter.
5. Use of instrument box, pencil box and non-programmable calculator is allowed during the examination. However, exchange of these items or any other papers (including question papers) is not permitted.
6. Write on both sides of the answer script and do not tear off any page. **Use last page(s) of the answer script for rough work.** Report to the invigilator if the answer script has torn or distorted page(s).
7. It is your responsibility to ensure that you have signed the Attendance Sheet. Keep your Admit Card/Identity Card on the desk for checking by the invigilator.
8. You may leave the examination hall for wash room or for drinking water for a very short period. Record your absence from the Examination Hall in the register provided. Smoking and the consumption of any kind of beverages are strictly prohibited inside the Examination Hall.
9. Do not leave the Examination Hall without submitting your answer script to the invigilator. **In any case, you are not allowed to take away the answer script with you.** After the completion of the examination, do not leave the seat until the invigilators collect all the answer scripts.
10. During the examination, either inside or outside the Examination Hall, gathering information from any kind of sources or exchanging information with others or any such attempt will be treated as **'unfair means'**. Do not adopt unfair means and do not indulge in unseemly behavior.

Violation of any of the above instructions may lead to severe punishment.

Signature of the Student

To be filled in by the examiner

Question Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Marks Obtained											

Marks obtained (in words)	**Signature of the Examiner**	**Signature of the Scrutineer**

Indian Institute of Technology Kharagpur
Department of Computer Science and Engineering

End-Semester Exam | **Reinforcement Learning (CS60077)** | **Autumn 2023-2024**

Date: 24-Nov-2023 (FN) | **Answer *all* questions.** | **Maximum Marks:** 80

— Write your answers at indicated places inside the question paper. —

— This page is kept blank intentionally. —

— The question paper starts from the next page. —

Q1. [Markov Decision Process and Bellman Equation] **14 marks**

Assume an underlying Markov Decision Process (MDP), $M = (S, A, P, R, \gamma)$, where $s, s' \in S$ are the states of an MDP, $a \in A$ denotes an action in MDP, $r(s,a) = R_s^a$ denotes the reward accumulated when applying action a to state s, $\mathbb{P}[s' \mid s, a] = P_{ss'}^a$ denotes the transition probability to s' from s upon executing action a, and $\gamma \in (0, 1]$ is the discount factor.

(a) The Bellman equations give us the iterative relations that relate the state value functions and the action value functions. The Bellman expectation equations provide the relation between state value function v_π and action value function q_π for any policy π, while the Bellman optimality equations provide the relation between the optimal state value function v_* and the optimal action value function q_*. The two tables below require you to fill out the relations between the quantities marked in the row and the column headings.

(i) Fill up the missing entries from the table for *Bellman Expectation Equations* as given below.

	v_π	q_π
v_π	① $v_\pi(s) = ?$	$v_\pi(s) = \sum_{a \in A} \pi(a\|s) q_\pi(s,a)$
q_π	② $q_\pi(s,a) = ?$	③ $q_\pi(s,a) = ?$

(3)

(ii) Fill up the missing entries from the table for *Bellman Optimality Equations* as given below.

	v_*	q_*
v_*	④ $v_*(s) = ?$	$v_*(s) = \max_{a \in A} q_*(s,a)$
q_*	⑤ $q_*(s,a) = ?$	⑥ $q_*(s,a) = ?$

(3)

Answer:

① $v_\pi(s) = \sum_{a \in A} \pi(a|s) \left[r(s,a) + \gamma \sum_{s' \in S} \mathbb{P}[s' \mid s, a] v_\pi(s') \right]$

② $q_\pi(s,a) = r(s,a) + \gamma \sum_{s' \in S} \mathbb{P}[s' \mid s, a] v_\pi(s')$

③ $q_\pi(s,a) = r(s,a) + \gamma \sum_{s' \in S} \mathbb{P}[s' \mid s, a] \left[\sum_{a' \in A} \pi(a'|s') q_\pi(s', a') \right]$

---

④ $v_*(s) = \max_{a \in A} \left[r(s,a) + \gamma \sum_{s' \in S} \mathbb{P}[s' \mid s, a] v_*(s') \right]$

⑤ $q_*(s,a) = r(s,a) + \gamma \sum_{s' \in S} \mathbb{P}[s' \mid s, a] v_*(s')$

⑥ $q_*(s,a) = r(s,a) + \gamma \sum_{s' \in S} \mathbb{P}[s' \mid s, a] \max_{a' \in A} q_*(s', a')$

(b) Let π be an ε-greedy policy. Let π' be the ε-greedy policy inferred from the action value function q_π (ε-greedy Policy Improvement from π to π'), i.e.,

$$\pi'(a|s) = \begin{cases} 1-\varepsilon+\frac{\varepsilon}{|A|}, & \text{if } a = arg\max_{b\in A} q_\pi(s,b) \\ \frac{\varepsilon}{|A|}, & \text{otherwise} \end{cases}$$

Prove that, $\sum_{a\in A} \pi'(a|s).q_\pi(s,a) \geq v_\pi(s)$, for all $s \in S$,
where $v_\pi(s)$ is state value function for policy π. **(8)**

Answer:

It may be noted that,

$$\sum_{a\in A} \pi'(a|s).q_\pi(s,a) = \frac{\varepsilon}{m}\sum_{a\in A} q_\pi(s,a) + (1-\varepsilon)\max_{a\in A} q_\pi(s,a)$$

Now, we make the crucial observation that a max over choices of A is greater than or equal to a weighted average over choices of A. Specifically,

$$\max_{a\in A} q_\pi(s,a) \geq \sum_{a\in A} w_a.q_\pi(s,a),$$

for any choice of weights $w_a \geq 0$, $a \in A$ constrained by $\sum_{a\in A} w_a = 1$. We will make a specific choice of w_a as follows:

$$w_a = \frac{\pi(a|s) - \frac{\varepsilon}{m}}{1-\varepsilon}$$

We note that $w_a \geq 0$ for all $a \in A$ because $\pi(a|s) \geq \frac{\varepsilon}{m}$ (since $\pi(a|s)$ is an ε-greedy policy). We also note that,

$$\sum_{a\in A} w_a = \frac{\sum_{a\in A}\pi(a|s) - \sum_{a\in A}\frac{\varepsilon}{m}}{1-\varepsilon} = \frac{1-\varepsilon}{1-\varepsilon} = 1$$

Having established that,

$$\max_{a\in A} q_\pi(s,a) \geq \sum_{a\in A} \frac{\pi(a|s) - \frac{\varepsilon}{m}}{1-\varepsilon}.q_\pi(s,a)$$

we can go back to the initial equation and state that:

$$\frac{\varepsilon}{m}\sum_{a\in A} q_\pi(s,a) + (1-\varepsilon)\max_{a\in A} q_\pi(s,a) \geq \frac{\varepsilon}{m}\sum_{a\in A} q_\pi(s,a) + (1-\varepsilon)\sum_{a\in A}\frac{\pi(a|s) - \frac{\varepsilon}{m}}{1-\varepsilon}.q_\pi(s,a)$$

Therefore,

$$\begin{aligned} \sum_{a\in A} \pi'(a|s).q_\pi(s,a) &\geq \frac{\varepsilon}{m}\sum_{a\in A} q_\pi(s,a) + (1-\varepsilon)\sum_{a\in A}\frac{\pi(a|s) - \frac{\varepsilon}{m}}{1-\varepsilon}.q_\pi(s,a) \\ &= \frac{\varepsilon}{m}\sum_{a\in A} q_\pi(s,a) + \sum_{a\in A}\pi(a|s).q_\pi(s,a) - \frac{\varepsilon}{m}\sum_{a\in A} q_\pi(s,a) \\ &= \sum_{a\in A}\pi(a|s).q_\pi(s,a) \quad = \quad v_\pi(s) \end{aligned}$$

Hence, $\sum_{a\in A} \pi'(a|s).q_\pi(s,a) \geq v_\pi(s)$, for all $s \in S$. [Proved]

Q2. [Model-free Control: Monte-Carlo vs. SARSA] **10 marks**

In this problem, we deal with the batch Monte-Carlo algorithm and the online Expected-SARSA algorithm. We set the discount factor to $\gamma = 1$ and run 5 episodes (in all episodes first action is always a_1) in a given environment as follows:

- *Episode 1:* $s_1\ a_1\ 0\ s_4\ a_2\ 0\ s_7$
- *Episode 2:* $s_1\ a_1\ 0\ s_4\ a_3\ 0.4\ s_8$
- *Episode 3:* $s_2\ a_1\ 0\ s_4\ a_2\ 2\ s_6$
- *Episode 4:* $s_2\ a_1\ 0\ s_4\ a_3\ 0.4\ s_8$
- *Episode 5:* $s_3\ a_1\ 0.5\ s_4\ a_2\ 0\ s_7$

Here, each episode starts in one of the states s_1, s_2, s_3 and with action a_1. Also, s_6, s_7, s_8 are terminal states. In s_4 there is a choice between actions a_2 and a_3 which are taken with equal probability $\pi(a_2|s_4) = \pi(a_3|s_4) = 0.5$. The rewards are deterministic as mentioned within the episodes with numeric values and only depend on the transition (s,a,s').

(a) Calculate the Q-values in states s_1, s_2, s_3, s_4 using online Expected-SARSA. For a given Q-value $Q(s,a)$, use $\eta = 1$ the FIRST TIME you update this value and $\eta \in [0,0.5]$ for all LATER (subsequent) update steps. Assume that $Q(s,a)$ values in all states are zero intitially. (**Hint:** You can neglect terms of order η^2.) **(5)**

Answer:

After the transition (s,a,r,s'), the update Q-value equation for Expected-SARSA is:

$$Q^{new}(s,a) \leftarrow \begin{cases} (1-\eta)Q^{old}(s,a) + \eta\left(r + \sum_{a'} \pi(s'a')Q^{old}(s',a')\right), & \text{if } s' \text{ is non-teminal} \\ (1-\eta)Q^{old}(s,a) + \eta r, & \text{if } s' \text{ is teminal} \end{cases}$$

Based on the 5 episodes, we get the following updates sequentially:

- *After Episode 1:*

$$\begin{aligned} Q(s_1,a_1) &\leftarrow (1-1)\cdot Q(s_1,a_1) + 1\cdot\left(0 + \frac{Q(s_4,a_2)+Q(s_4,a_3)}{2}\right) = 0 \\ Q(s_4,a_2) &\leftarrow (1-1)\cdot Q(s_4,a_2) + 1\cdot 0 = 0 \end{aligned}$$

- *After Episode 2:*

$$\begin{aligned} Q(s_1,a_1) &\leftarrow (1-\eta)\cdot Q(s_1,a_1) + \eta\cdot\left(0 + \frac{Q(s_4,a_2)+Q(s_4,a_3)}{2}\right) = 0 \\ Q(s_4,a_3) &\leftarrow (1-1)\cdot Q(s_4,a_3) + 1\cdot 0.4 = 0.4 \end{aligned}$$

- *After Episode 3:*

$$\begin{aligned} Q(s_2,a_1) &\leftarrow (1-1)\cdot Q(s_2,a_1) + 1\cdot\left(0 + \frac{Q(s_4,a_2)+Q(s_4,a_3)}{2}\right) = 0.2 \\ Q(s_4,a_2) &\leftarrow (1-\eta)\cdot Q(s_4,a_2) + \eta\cdot 2 = 2\eta \end{aligned}$$

- *After Episode 4:*

$$\begin{aligned} Q(s_2,a_1) &\leftarrow (1-\eta)\cdot Q(s_2,a_1) + \eta\cdot\left(0 + \frac{Q(s_4,a_2)+Q(s_4,a_3)}{2}\right) = 0.2+\eta^2 \approx 0.2 \\ Q(s_4,a_3) &\leftarrow (1-\eta)\cdot Q(s_4,a_3) + \eta\cdot 0.4 = 0.4 \end{aligned}$$

- *After Episode 5:*

$$\begin{aligned} Q(s_3,a_1) &\leftarrow (1-1)\cdot Q(s_3,a_1) + 1\cdot\left(0.5 + \frac{Q(s_4,a_2)+Q(s_4,a_3)}{2}\right) = 0.7+\eta \\ Q(s_4,a_2) &\leftarrow (1-\eta)\cdot Q(s_4,a_2) + \eta\cdot 0 = 2\eta - 2\eta^2 \approx 2\eta \end{aligned}$$

As a result:

$$Q(s_1,a_1) = 0;\ Q(s_2,a_1) = 0.2;\ Q(s_3,a_1) = 0.7+\eta;\ Q(s_4,a_2) = 2\eta;\ Q(s_4,a_3) = 0.4$$

(b) Calculate the Q-values in states s_1, s_2, s_3, s_4 using batch Monte-Carlo control (i.e., average total returns from each starting state). **(3)**

Answer:

$$
\begin{aligned}
Q(s_1, a_1) &= \frac{(0+0)+(0+0.4)}{2} = 0.2 \\
Q(s_2, a_1) &= \frac{(0+2)+(0+0.4)}{2} = 1.2 \\
Q(s_3, a_1) &= \frac{(0.5+0)}{1} = 0.5 \\
Q(s_4, a_2) &= \frac{0+2+0}{3} = 0.67 \\
Q(s_4, a_3) &= \frac{0.4+0.4}{2} = 0.4
\end{aligned}
$$

(c) You can choose the initial state for episode 6. Which initial state looks best in Part (a), i.e. for online Expected SARSA? Which initial state looks best in Part (b), i.e. for batch Monte-Carlo? **(2)**

Answer:

With online Expected-SARSA, s_3 looks best.
But, with batch Monte-Carlo, s_2 looks best.

Q3. [TD(λ) and Eligibility Traces] **16 marks**

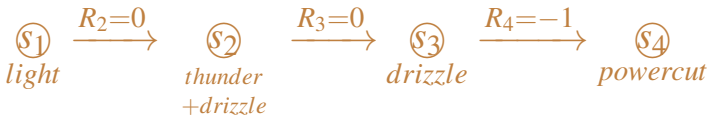
Sitting at hall on a rainy evening, you hear an episode of experience as follows: At the first step you saw a lightning. At the second step you hear a thunder with a drizzle of rain. At the third step you saw only a drizzle of rain. Then you had a powercut, worth -1 reward, and the episode terminates on the fourth step. All other rewards were zero. The experiment is undiscounted (i.e., $\gamma = 1$).

We may represent the state s that you witnessed by a vector of three binary features, $light(s) \in \{0,1\}$, $thunder(s) \in \{0,1\}$ and $drizzle(s) \in \{0,1\}$. So, the sequence of feature vectors corresponding to the four steps of this episode can be expressed as, $[1,0,0]^\intercal$, $[0,1,1]^\intercal$, $[0,0,1]^\intercal$ and $[0,0,0]^\intercal$.

(a) Approximate the state-value function by a linear combination of these features with two parameters: $\phi(s) = \big[l \times light(s) + t \times thunder(s) + d \times drizzle(s)\big]$. If $l = -3$, $t = -2$ and $d = 1$, then write down the sequence of approximate values corresponding to this episode. **(2)**

Answer:

The activity of the presented episode is given as follows:



Given, $l = -3$, $t = -2$ and $d = 1$. Let the feature vectors corresponding to each episode are given as, $\phi(s_1) = [1,0,0]^\intercal$, $\phi(s_2) = [0,1,1]^\intercal$, $\phi(s_3) = [0,0,1]^\intercal$, $\phi(s_4) = [0,0,0]^\intercal$.
So, the sequence of approximate values corresponding to each step in this episode is denoted as,

$$\begin{aligned} V(s_1) &= [l,t,d].\phi(s_1) = [-3,-2,1].[1,0,0]^\intercal = -3 \\ V(s_2) &= [l,t,d].\phi(s_2) = [-3,-2,1].[0,1,1]^\intercal = -1 \\ V(s_3) &= [l,t,d].\phi(s_3) = [-3,-2,1].[0,0,1]^\intercal = 1 \\ V(s_4) &= [l,t,d].\phi(s_4) = [-3,-2,1].[0,0,0]^\intercal = 0 \quad \text{(for terminal state)} \end{aligned}$$

It may also be noted that, $R_2 = 0$, $R_3 = 0$, but $R_4 = -1$.

(b) Write down the sequence of λ-returns G_t^λ $(1 \le t \le 3)$ corresponding to this episode, for $\lambda = \frac{1}{2}$ and $l = -3$, $t = -2$, $d = 1$. Clearly show the detailed evaluations. **(6)**

Answer:

The n-step return and λ-return G_t^λ are given as,

$$\begin{aligned} G_t^{(n)} &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n V(s_{t+n}) \\ G_t^\lambda &= (1-\lambda) \sum_{n=1}^{\infty} \lambda^{n-1} . G_t^{(n)} \end{aligned}$$

where, $G_t^{(\infty)} = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-1} R_T$, with multiplication factor being λ^{T-t-1}

Here, the calculation for $G_1^{(\lambda)}$ goes as follows. In our case, $T = 4$ and hence $R_T = R_4 = -1$.

$$\begin{aligned}
G_1^{(1)} &= R_2 + \gamma V(s_2) = 0 + 1 \times (-1) = -1 \\
G_1^{(2)} &= R_2 + \gamma R_3 + \gamma^2 V(s_3) = 0 + 1 \times 0 + 1 \times 1 = 1 \\
G_1^{(\infty)} &= R_2 + \gamma R_3 + \gamma^2 R_4 = 0 + 1 \times 0 + 1 \times (-1) = -1 \\
\therefore\ G_1^{\lambda} &= (1 - \frac{1}{2}) \times \Big[(\frac{1}{2})^0 G_1^{(1)} + (\frac{1}{2})^1 G_1^{(2)} \Big] + (\frac{1}{2})^{4-1-1} G_1^{(\infty)} = \frac{1}{2} \Big[1 \times (-1) + \frac{1}{2} \times 1 + \frac{1}{2} \times (-1) \Big] = -\frac{1}{2}
\end{aligned}$$

Similarly, the calculation for $G_2^{(\lambda)}$ goes as follows.

$$\begin{aligned}
G_2^{(1)} &= R_3 + \gamma V(s_3) = 0 + 1 \times 1 = 1 \\
G_2^{(\infty)} &= R_3 + \gamma R_4 = 0 + 1 \times (-1) = -1 \\
\therefore\ G_2^{\lambda} &= (1 - \frac{1}{2}) \times \Big[(\frac{1}{2})^0 G_1^{(1)} \Big] + (\frac{1}{2})^{4-2-1} G_2^{(\infty)} = \frac{1}{2} \Big[1 \times 1 + 1 \times (-1) \Big] = 0
\end{aligned}$$

And, the calculation for $G_3^{(\lambda)}$ goes as follows.

$$\begin{aligned}
G_3^{(\infty)} &= R_4 = -1 \\
\therefore\ G_3^{\lambda} &= (\frac{1}{2})^{4-3-1} G_3^{(\infty)} = (\frac{1}{2})^0 \times (-1) = -1
\end{aligned}$$

(c) Using the forward-view TD(λ) algorithm and your linear function approximator, what are the sequence of updates to weight d corresponding to the drizzle? What is the total update to weight d? Use $\lambda = \frac{1}{2}$, $\gamma = 1$, $\alpha = \frac{1}{2}$ and start with $l = -3$, $t = -2$, $d = 1$. **(3)**

Answer:

The sequence of updates to weight d is given as,

$$\begin{aligned}
\Delta d_1 &= \alpha \Big(G_1^{\lambda} - V(s_1) \Big) drizzle(s_1) = \frac{1}{2} \times \Big(-\frac{1}{2} - (-3) \Big) \times 0 = 0 \\
\Delta d_2 &= \alpha \Big(G_2^{\lambda} - V(s_2) \Big) drizzle(s_2) = \frac{1}{2} \times \Big(0 - (-1) \Big) \times 1 = \frac{1}{2} \\
\Delta d_3 &= \alpha \Big(G_3^{\lambda} - V(s_3) \Big) drizzle(s_3) = \frac{1}{2} \times \Big(-1 - 1 \Big) \times 1 = -1
\end{aligned}$$

The total update to weight d is, $\sum \Delta d = (\Delta d_1 + \Delta d_2 + \Delta d_3) = 0 + \frac{1}{2} + (-1) = -\frac{1}{2}$.

(d) Using linear value function approximation, write down the sequence of TD(λ) accumulating eligibility trace e_t corresponding to the drizzle, using $\lambda = \frac{1}{2},\ \gamma = 1$. **(2)**

Answer:

The equation for eligibility trace is given as, $e_t = \gamma\alpha e_{t-1} + drizzle(s_t)$.
So, the sequence of eligibility traces e_t corresponding to $drizzle(s_t)$ are,

$$\begin{aligned} e_1 &= 1 \times \frac{1}{2} \times 0 + drizzle(s_1) = 0 \\ e_2 &= 1 \times \frac{1}{2} \times 0 + drizzle(s_2) = 1 \\ e_2 &= 1 \times \frac{1}{2} \times 1 + drizzle(s_3) = \frac{3}{2} \end{aligned}$$

(e) Using the backward-view TD(λ) algorithm and your linear function approximator, what are the sequence of updates to weight d? (Use offline updates, i.e., do not actually change your weights, just accumulate your updates). What is the total update to weight d? Use $\lambda = \frac{1}{2},\ \gamma = 1,\ \alpha = \frac{1}{2}$ and start with $l = -3,\ t = -2,\ d = 1$. **(3)**

Answer:

The sequence of updates to weight b is given as,

$$\begin{aligned} \Delta d_1 &= \alpha\delta_1 e_1 = \alpha\Big[R_2 + \gamma V(s_2) - V(s_1)\Big]e_1 = \frac{1}{2} \times \Big[0 + 1 \times (-1) - (-3)\Big] \times 0 = 0 \\ \Delta d_2 &= \alpha\delta_2 e_2 = \alpha\Big[R_3 + \gamma V(s_3) - V(s_2)\Big]e_2 = \frac{1}{2} \times \Big[0 + 1 \times 1 - (-1)\Big] \times 1 = 1 \\ \Delta d_3 &= \alpha\delta_3 e_3 = \alpha\Big[R_4 + \gamma V(s_4) - V(s_3)\Big]e_3 = \frac{1}{2} \times \Big[-1 + 1 \times 0 - 1\Big] \times \frac{3}{2} = -\frac{3}{2} \end{aligned}$$

The total update to weight b is, $\sum \Delta d = (\Delta d_1 + \Delta d_2 + \Delta d_3) = 0 + 1 + -\frac{3}{2} = -\frac{1}{2}$.

Q4. [Policy Gradient Methods] **12 marks**

(a) Let us consider the linear MDP as shown below. Here the (shaded) states, s_1 and s_7, are

s_1	s_2	s_3	s_4	s_5	s_6	s_7
$+1$	0	-1	0	-1	0	$+10$

terminal states. The rewards presented below are received when you enter a particular state. There are two actions, `Left` and `Right`. The action, `Left`, transitions from state s_i to s_{i-1} with 0.5 probability and stays in state s_i with 0.5 probability. Similarly, the action, `Right`, transitions from state s_i to s_{i+1} with 0.5 probability and stays in state s_i with 0.5 probability. Let $\gamma = 1$. We want to apply Monte-Carlo policy gradient algorithm, REINFORCE, to learn a policy in this linear MDP setting. Let our feature representation be a one-hot encoding using the state, action pair. More concretely, let us denote $a_1 = \texttt{Left}$ and $a_2 = \texttt{Right}$. Then, assuming the vector is 0-indexed, our feature representation is, $\phi(s_i, a_j)_k = \begin{cases} 1, & \text{if } 7(j-1)+(i-1) = k \\ 0, & \text{otherwise} \end{cases}$. Let us use a softmax policy parameterized by θ: $\pi_\theta(s,a) = \frac{\exp\left(\phi(s,a)^\intercal \theta\right)}{\sum_a \exp\left(\phi(s,a)^\intercal \theta\right)}$ and run REINFORCE algorithm. Assume θ is initialized to be all zeros. We execute one rollout of the policy π_θ to obtain the following episode: $(s_4, a_1, -1, s_3, a_2, 0, s_4, a_2, -1, s_5, a_2, 0, s_6, a_1, 0, s_6, a_2, +10)$. Run REINFORCE to update θ *three times* using the provided episode. For simplicity, let $\alpha = 1$. **(6)**

Answer:

The score function is: $\nabla_\theta \log \pi_\theta(s,a) = \phi(s,a) - \mathbb{E}_{\pi_\theta}\left[\phi(s,\cdot)\right]$

So, the update using REINFORCE algorithm will be:

$$\theta \leftarrow \theta + \alpha \nabla_\theta \log \pi_\theta(s_t, a_t) G_t = \theta + \alpha \left[\phi(s,a) - \sum_b \pi_\theta(s,b) \cdot \phi(s,b)\right] G_t$$

Iteratively,

- After the first update:

$$\begin{aligned} \theta &= [0,0,0,0,0,0,0,0,0,0,0,0,0,0,0] + 1 \cdot \Big[[0,0,0,1,0,0,0,0,0,0,0,0,0,0,0] - \\ &\quad \Big(0.5 \cdot [0,0,0,1,0,0,0,0,0,0,0,0,0,0,0] + 0.5 \cdot [0,0,0,0,0,0,0,0,0,0,0,1,0,0,0]\Big)\Big] \cdot 8 \\ &= [0,0,0,4,0,0,0,0,0,0,-4,0,0,0] \end{aligned}$$

- After the second update:

$$\begin{aligned} \theta &= [0,0,0,4,0,0,0,0,0,0,-4,0,0,0] + 1 \cdot \Big[[0,0,0,0,0,0,0,0,0,1,0,0,0,0,0] - \\ &\quad \Big(0.5 \cdot [0,0,1,0,0,0,0,0,0,0,0,0,0,0,0] + 0.5 \cdot [0,0,0,0,0,0,0,0,0,0,1,0,0,0,0]\Big)\Big] \cdot 9 \\ &= [0,0,-4.5,4,0,0,0,0,0,4.5,-4,0,0,0] \end{aligned}$$

- After the third update:

$$\begin{aligned} \theta &= [0,0,-4.5,4,0,0,0,0,0,4.5,-4,0,0,0] + 1 \cdot \Big[[0,0,0,0,0,0,0,0,0,0,1,0,0,0] - \\ &\quad \Big(0.5 \cdot [0,0,0,1,0,0,0,0,0,0,0,0,0,0,0] + 0.5 \cdot [0,0,0,0,0,0,0,0,0,0,0,1,0,0,0]\Big)\Big] \cdot 9 \\ &= [0,0,-4.5,-0.5,0,0,0,0,0,4.5,0.5,0,0,0] \end{aligned}$$

Note that, instead of updating θ in place, we use the original θ used to collect the data in the computation of π_θ.

(b) Suppose you have a Gaussian policy, π_θ, that samples actions a from a normal distribution with mean $\mu = \phi(s)^\intercal\theta$ and variance σ^2.

(**Hint:** Recall that, the Gaussian PDF is as follows: $f(x;\,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$.)

(i) Prove that, $\nabla_\theta \log \pi_\theta(s,a) = \frac{\left(a-\phi(s)^\intercal\theta\right)\phi(s)}{\sigma^2}$. (show your derivation in details) **(3)**

Answer:

$$\begin{aligned}
\nabla_\theta \log \pi_\theta(s,a) &= \frac{1}{\pi_\theta(s,a)} \nabla_\theta \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{a-\phi(s)^\intercal\theta}{\sigma}\right)^2} \\
&= \frac{1}{\pi_\theta(s,a)} \pi_\theta(s,a) \nabla_\theta -\frac{1}{2}\left(\frac{a-\phi(s)^\intercal\theta}{\sigma}\right)^2 \\
&= -\frac{1}{2\sigma^2} \cdot 2\left(a-\phi(s)^\intercal\theta\right)\left(-\phi(s)\right) \quad = \quad \frac{\left(a-\phi(s)^\intercal\theta\right)\phi(s)}{\sigma^2}
\end{aligned}$$

Alternatively, writing the log density

$$\log \pi_\theta(s,a) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{1}{2}\left(\frac{\phi(s)^\intercal\theta - a}{\sigma}\right)^2 = -\frac{1}{2}\left(\log(2\pi) + 2\log\sigma + \left(\frac{\phi(s)^\intercal\theta - a}{\sigma}\right)^2\right)$$

and differentiating with respect to θ,

$$\begin{aligned}
\nabla_\theta \log \pi_\theta(s,a) &= -\frac{1}{2} \nabla_\theta \left(\frac{\phi(s)^\intercal\theta - a}{\sigma}\right)^2 \\
&= -\frac{1}{2} \cdot 2\left(\left(\frac{\phi(s)^\intercal\theta - a}{\sigma}\right)\frac{\phi(s)}{\sigma}\right) \quad = \quad \frac{\left(a-\phi(s)^\intercal\theta\right)\phi(s)}{\sigma^2}
\end{aligned}$$

(ii) Prove that, $\nabla_\sigma \log \pi_\theta(s,a) = \frac{\left(a-\phi(s)^\intercal\right)^2}{\sigma^3} - \frac{1}{\sigma}$. (show your derivation in details) **(3)**

Answer:

$$\begin{aligned}
\nabla_\sigma \log \pi_\theta(s,a) &= \frac{1}{\pi_\theta(s,a)} \nabla_\sigma \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{a-\phi(s)^\intercal\theta}{\sigma}\right)^2} \\
&= \frac{1}{\pi_\theta(s,a)} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \nabla_\sigma e^{-\frac{1}{2}\left(\frac{a-\phi(s)^\intercal\theta}{\sigma}\right)^2} + e^{-\frac{1}{2}\left(\frac{a-\phi(s)^\intercal\theta}{\sigma}\right)^2} \nabla_\sigma \frac{1}{\sqrt{2\pi\sigma^2}}\right] \\
&= \nabla_\sigma \frac{-1}{2\sigma^2}\left(a-\phi(s)^\intercal\theta\right)^2 + \frac{1}{\pi_\theta(s,a)} e^{-\frac{1}{2}\left(\frac{a-\phi(s)^\intercal\theta}{\sigma}\right)^2} \frac{1}{\sqrt{2\pi}} \nabla_\sigma \frac{1}{\sigma} \\
&= \frac{1}{\sigma^3}\left(a-\phi(s)^\intercal\right)^2 + \frac{1}{\pi_\theta(s,a)} e^{-\frac{1}{2}\left(\frac{a-\phi(s)^\intercal\theta}{\sigma}\right)^2} \frac{1}{\sqrt{2\pi}} \cdot \frac{-1}{\sigma^2} \\
&= \frac{\left(a-\phi(s)^\intercal\right)^2}{\sigma^3} - \frac{1}{\sigma}
\end{aligned}$$

Alternatively, directly differentiating the log density with respect to σ,

$$\begin{aligned}
\frac{\partial}{\partial\sigma} \log \pi_\theta(s,a) &= -\frac{1}{2}\left(2\frac{\partial \log\sigma}{\partial\sigma} + \frac{\partial}{\partial\sigma}\left(\frac{\phi(s)^\intercal\theta - a}{\sigma}\right)^2\right) \\
&= -\frac{1}{2}\left(\frac{2}{\sigma} - \frac{2\left(\phi(s)^\intercal\theta - a\right)^2}{\sigma^3}\right) \quad = \quad \frac{\left(a-\phi(s)^\intercal\right)^2}{\sigma^3} - \frac{1}{\sigma}
\end{aligned}$$

Q5. [Multi-arm Bandits] **12 marks**

(a) In the *2-armed bandit* problem, one has to choose one of 2 actions. Assume action a_1 yields a reward of $r = 1$ with probability $p = 0.25$ and 0 otherwise. If you take action a_2, you will receive a reward of $r = 0.4$ with probability $p = 0.75$ and 0 otherwise. The *2-armed bandit* game is played several times and Q-values are updated using the update rule, $\Delta Q(s,a) = \eta\left[r_t - Q(s,a)\right]$.

(i) Assume that you initialize all Q-values at zero. You first try both actions: in trial 1, you choose a_1 and get $r = 1$; in trial 2, you choose a_2 and get $r = 0.4$. Update your Q-values ($\eta = 0.2$). **(3)**

Answer:

In the beginning, $Q(a_1, t = 0) = Q(a_2, t = 0) = 0$ (we dropped the state index s since there is only a single state). After choosing action a_1 and receiving a reward of $r = 1$, its Q-value is updated to:

$$Q(a_1, t = 1) = Q(a_1, t = 0) + \Delta Q(a_1) = 0 + \eta\left(r - Q(a_1, t = 0)\right) = 0 + 0.2 \times 1 = 0.2$$

After choosing action a_2 and receiving a reward of $r = 0.4$, its Q-value is updated to:

$$Q(a_2, t = 2) = Q(a_2, t = 0) + \Delta Q(a_2) = 0 + \eta\left(r - Q(a_2, t = 0)\right) = 0 + 0.2 \times 0.4 = 0.08$$

Continuing with a greedy method implies that in the next round, action a_1 will be chosen.

(ii) In trials 3 to 5, you play greedy and always choose the action which looks best (i.e., has the highest Q-value). Which action has the higher Q-value after trial 5? (Assume that the actual reward is $r = 0$ in trials 3-5.). **(3)**

Answer:

In trial 3 you take action a_1. If the return is 0,

$$\begin{aligned} Q(a_1, t = 3) &= Q(a_1, t = 2) + \eta\left(r - Q(a_1, t = 2)\right) \\ &= (1 - \eta)Q(a_1, t = 2) + \eta r = 0.8 \times 0.2 = 0.16 \end{aligned}$$

Thus, in trial 4 we take again action a_1. If the return is again 0,

$$Q(a_1, t = 4) = (1 - \eta)Q(a_1, t = 3) + \eta r = 0.8 \times 0.16 = 0.128$$

In trial 5 we take again action a_1. If the return is again 0,

$$Q(a_1, t = 5) = (1 - \eta)Q(a1, t = 4) + \eta r = 0.8 \times 0.128 = 0.1024$$

Thus, with a greedy policy, also in trial 6 action a_1 will be taken. If by chance some of the returns were 1 instead of 0, $Q(a_1, t = 5)$ would be even higher, while $Q(a_2, t = 5) = Q(a_2, t = 2) = 0.08$ because action a_2 was never taken.

(iii) Calculate the expected reward for both actions. Which one is the best? **(2)**

Answer:

For action a_1, the expected reward per round is given by,

$$\mathbb{E}[r1] = p \cdot 1 + (1-p) \cdot 0 = 0.25$$

For action a_2, the expected reward per round is evaluated to

$$\mathbb{E}[r2] = 0.75 \times 0.4 + 0.25 \times 0 = 0.3$$

The second action yields a higher reward on average.

(b) After 12 iterations of the UCB 1 algorithm applied on a 4-arm bandit problem, we have $n_1 = 3,\ n_2 = 4,\ n_3 = 3,\ n_4 = 2$ and $Q_{12}(1) = 0.55,\ Q_{12}(2) = 0.63,\ Q_{12}(3) = 0.61,\ Q_{12}(4) = 0.40$. Which arm should be played next? Show the calculations to justify your answer. **(4)**

Answer:

The next action, A_{13}, will be the action with the maximum upper confidence bound among the four arms. Calculating these values, we have

$$\begin{aligned}
Q_{12}(1) + \sqrt{\frac{2\ln 12}{n_1}} &= 0.55 + \sqrt{\frac{2\ln 12}{3}} = 1.837 \\
Q_{12}(2) + \sqrt{\frac{2\ln 12}{n_2}} &= 0.63 + \sqrt{\frac{2\ln 12}{4}} = 1.745 \\
Q_{12}(3) + \sqrt{\frac{2\ln 12}{n_3}} &= 0.61 + \sqrt{\frac{2\ln 12}{3}} = 1.897 \\
Q_{12}(4) + \sqrt{\frac{2\ln 12}{n_4}} &= 0.40 + \sqrt{\frac{2\ln 12}{2}} = 1.976
\end{aligned}$$

Clearly, arm 4 has the highest upper confidence bound and hence will be selected by the UCB 1 algorithm.

Q6. [Approximation Guarantees over MDPs] **16 marks**

(a) For an MDP $\langle S, A, P, R, \gamma \rangle$, let $V_0 : S \to \mathbb{R}$ be an initial guess of the optimal value function V^*. Let this guess be progressively updated using Value Iteration: i.e., by setting $V_{t+1} \leftarrow T^*(V_t)$ for $t = 0, 1, 2, \ldots$. Recall that T^* is the Bellman optimality operator.

In this question, we examine the design of a stopping condition for Value Iteration. As usual, let $||\cdot||_\infty$ denote the max norm. We would like that our computed solution, V_u for some $u \in \{1, 2, \ldots\}$, be within ε of V^* for some given tolerance $\varepsilon > 0$. In other words, we would like to stop after u applications of T^*, so long as we can guarantee $||V_u - V^*||_\infty \leq \varepsilon$. Naturally, we cannot use V^* itself in our stopping rule, since it is not known! *Prove that* it suffices to stop when $||V_u - V_{u-1}||_\infty \leq \frac{\varepsilon(1-\gamma)}{\gamma}$ and thereafter return V_u as the answer. **(5)**

You are likely to find two results handy: (1) that T^* is a contraction mapping with contraction factor γ, and (2) the triangle inequality: for $X : S \to \mathbb{R}$, $Y : S \to \mathbb{R}$, $||X + Y||_\infty \leq ||X||_\infty + ||Y||_\infty$.

Answer:

Let $\varepsilon' = \frac{\varepsilon(1-\gamma)}{\gamma}$. We are given $||V_u - V_{u-1}||_\infty \leq \varepsilon'$; by successive application of the result that T^* is a contraction mapping with contraction factor γ, we get

$$\begin{aligned} ||V_u - V_{u-1}||_\infty &\leq \varepsilon' \\ ||T^*(V_u) - T^*(V_{u-1})||_\infty &\leq \varepsilon'\gamma \\ ||(T^*)^2(V_u) - (T^*)^2(V_{u-1})||_\infty &\leq \varepsilon'\gamma^2 \\ &\vdots \\ ||(T^*)^k(V_u) - (T^*)^k(V_{u-1})||_\infty &\leq \varepsilon'\gamma^k \end{aligned}$$

for all $k \geq 0$. By using the triangle inequality, we obtain

$$||(T^*)^k(V_u) - V_u||_\infty \leq \sum_{j=1}^{k} ||(T^*)^j(V_u) - (T^*)^j(V_{u-1})||_\infty \leq \varepsilon'\left(\gamma + \gamma^2 + \gamma^3 + \cdots + \gamma^k\right)$$

for all $k \geq 0$. Taking the limit as, $k \to \infty$ yields

$$||V^* - V_u||_\infty \leq \frac{\varepsilon'\gamma}{1-\gamma} = \varepsilon,$$

thereby guaranteeing the stopping condition. [Proved]

(b) The *future state distribution* gives the probability of a state s appearing anywhere in a trajectory τ when a policy π is followed. It is denoted by $T^{\pi}(s) = \sum_{t=0}^{\infty} \mathbb{P}\big[S_t = s \mid \pi\big]$. Similarly,the *discounted future state distribution* provides the probability of a state s appearing anywhere in a trajectory but discounted by when is the state visited. It is denoted by, $d^{\pi}(s)$ and is defined as, $d^{\pi}(s) = (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}\big[S_t = s \mid \pi\big]$ where γ is the discount factor of an infinite horizon MDP. $\mathbb{P}\big[S_t = s \mid \pi\big]$ denotes the probability of the state s to appear at timestep t.

We can use the discounted future state distribution to rewrite the objective function of a RL problem in the infinite horizon discounted reward setting. The objective i.e., the expectation of the discounted sums over trajectories can be rewritten in terms of expectations over states and actions. For any policy π and any reward function $r : S \times A \to \mathbb{R}$, the relation can be written as,

$$\mathop{\mathbb{E}}_{\tau \sim p(\tau)}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\Big] = \frac{1}{1-\gamma}\sum_{s \in S} d^{\pi}(s) \sum_{a \in A} \pi(a|s) r(s,a) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s \sim d^{\pi}(\cdot) \\ a \sim \pi(\cdot|s)}} \big[r(s,a)\big]$$

where, $s \sim d^{\pi}(\cdot)$ is a shorthand to denote the fact that states are drawn according to the discounted future state distribution and similarly $a \sim \pi(\cdot|s)$ is a shorthand to denote the fact that actions are distributed according to π. Your task is to *prove this above relation*. **(7)**

(**Hint:** Start proving from the opposite direction!)

Answer:

$$\begin{aligned}
\frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s \sim d^{\pi}(\cdot) \\ a \sim \pi(\cdot|s)}} \big[r(s,a)\big] &= \frac{1}{1-\gamma}\sum_{s \in S}\sum_{a \in A} r(s,a) d^{\pi}(s) \pi(a|s) && \text{(Definition of Expectation)} \\
&= \frac{1}{1-\gamma}\sum_{s \in S} d^{\pi}(s) \sum_{a \in A} r(s,a)\pi(a|s) \\
&= \frac{1}{1-\gamma}\sum_{s \in S} (1-\gamma)\sum_{t=0}^{\infty} \gamma^t \mathbb{P}\big[S_t = s \mid \pi\big] \sum_{a \in A} r(s,a)\pi(a|s) && \text{(Definition of } d^{\pi}(s)\text{)} \\
&= \sum_{s \in S}\sum_{t=0}^{\infty} \gamma^t \mathbb{P}\big[S_t = s \mid \pi\big] \sum_{a \in A} r(s,a)\pi(a|s) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s \in S} \mathbb{P}\big[S_t = s \mid \pi\big] \sum_{a \in A} r(s,a)\pi(a|s) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s \in S}\sum_{a \in A} r(s,a)\mathbb{P}\big[S_t = s \mid \pi\big]\pi(a|s) \\
&= \sum_{t=0}^{\infty} \gamma^t \sum_{s \in S}\sum_{a \in A} r(s,a)\mathbb{P}\big[s, a \mid \pi\big] \\
&= \mathop{\mathbb{E}}_{\tau \sim p(\tau)}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\Big] && \text{(Linearity of Expectation over summation)}
\end{aligned}$$

Therefore, in reverse,

$$\mathop{\mathbb{E}}_{\tau \sim p(\tau)}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\Big] = \frac{1}{1-\gamma}\sum_{s \in S} d^{\pi}(s) \sum_{a \in A} \pi(a|s) r(s,a) = \frac{1}{1-\gamma} \mathop{\mathbb{E}}_{\substack{s \sim d^{\pi}(\cdot) \\ a \sim \pi(\cdot|s)}} \big[r(s,a)\big]$$

[Proved]

(c) Suppose an MDP $\langle S, A, R, P, \gamma \rangle$ is defined such that $R(s,a) \geq 0$ for all state action pairs $(s,a) \in S \times A$. Furthermore, suppose that for every state $s \in S$, there is some action $a_s \in A$ such that $\mathbb{P}(s' = s \mid s, a_s) \geq p$, where $0 \leq p \leq 1$ is some constant probability. Consider performing value iteration on this MDP. Let $V_t(s)$ be the value of state s after t iterations. We initialize to $V_0(s) = 0$ for all states $s \in S$. *Prove that* for all states $s \in S$ and $t \geq 0$, $\quad V_{t+1}(s) \geq p\gamma V_t(s)$. **(4)**

Answer:

Consider an arbitrary state $s \in S$ and an arbitrary iteration $t \geq 0$. From the value iteration algorithm, we have:

$$\begin{aligned} V_{t+1}(s) &= \max_{a \in A} R(s,a) + \gamma \sum_{s' \in S} \mathbb{P}(s' \mid s, a) V_t(s') \\ &\geq R(s,a_s) + \gamma \sum_{s' \in S} \mathbb{P}(s' \mid s, a_s) V_t(s') \\ &\geq R(s,a_s) + \gamma p V_t(s) \\ &\geq \gamma p V_t(s) \end{aligned}$$

Where the 3rd line follows by the fact that $\forall s \; : \; V_t(s) \geq 0$ because $R(s,a) \geq 0$ and initialization is 0. Last line uses $R(s,a) \geq 0$. This completes the proof.

— The question paper ends here. —