AI 3000 / CS 5500 : Reinforcement Learning

Assignment № 2

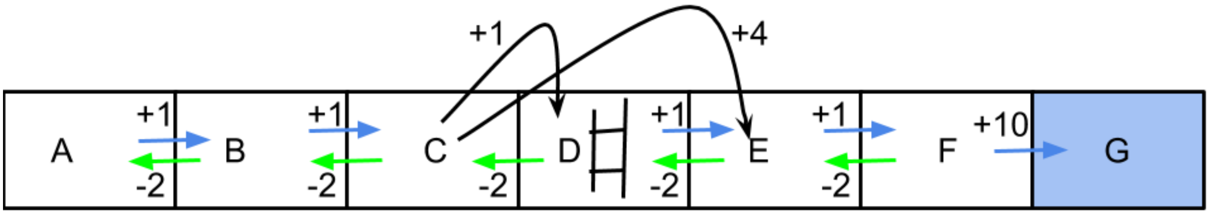
Due Date : 18/10/2021

Teaching Assistants : Chaitanya Devaguptapu and Deepayan Das

Easswar Subramanian, IIT Hyderabad 06/10/2021

Problem 1 : Model Free Prediction and Control

Consider the MDP shown below with states $\{A, B, C, D, E, F, G\}$. Normally, an agent can either move *left* or *right* in each state. However, in state C, the agent has the choice to either move *left* or *jump* forward as the state D of the MDP has an hurdle. There is no *right* action from state C. The *jump* action from state C will place the agent either in square D or in square E with probability 0.5 each. The rewards for each action at each state s is depicted in the figure below alongside the arrow. The terminal state is G and has a reward of zero. Assume a discount factor of $\gamma = 1$.



Consider the following samples of Markov chain trajectories with rewards to answer the questions below

- $A \xrightarrow{+1} B \xrightarrow{+1} C \xrightarrow{-2} B \xrightarrow{+1} C \xrightarrow{+1} D \xrightarrow{+1} E \xrightarrow{+1} F \xrightarrow{+10} G$
- $A \xrightarrow{+1} B \xrightarrow{+1} C \xrightarrow{+1} D \xrightarrow{+1} E \xrightarrow{+1} F \xrightarrow{+10} G$
- $A \xrightarrow{+1} B \xrightarrow{+1} C \xrightarrow{+4} E \xrightarrow{+1} F \xrightarrow{+10} G$
- $A \xrightarrow{+1} B \xrightarrow{+1} C \xrightarrow{+4} E \xrightarrow{-2} D \xrightarrow{+1} E \xrightarrow{+1} F \xrightarrow{+10} G$
- $A \xrightarrow{+1} B \xrightarrow{+1} C \xrightarrow{+4} E \xrightarrow{-2} D \xrightarrow{+1} E \xrightarrow{+1} F \xrightarrow{-2} E \xrightarrow{+1} F \xrightarrow{+10} G$

(a) Evaluate $V(s)$ using first visit Monte-Carlo method for all states s of the MDP. (2 Points)

(i) $V(A) = (14 + 15 + 17 + 16 + 15)/5 = 77/5 = 15.4$

(ii) $V(B) = (13 + 14 + 16 + 15 + 14)/5 = 72/5 = 14.4$

(iii) $V(C) = (12 + 13 + 15 + 14 + 13)/5 = 67/5 = 13.4$

(iv) $V(D) = (12 + 12 + 12 + 11)/4 = 47/4 = 11.75$

(v) $V(E) = (11 + 11 + 11 + 10 + 9)/5 = 47/4 = 10.4$

(vi) $V(F) = (10 + 10 + 10 + 10 + 9)/5 = 9.8$ and $V(G) = 0$

(b) Which states are likely to have different value estimates if evaluated using every visit MC as compared to first visit MC ? Why ? (1 Point)

States $\{B, C, E, F\}$ are likely to have different value estimates when evaluated using every visit MC as these are visited more than once in a single rollout.

(c) Now consider a policy π_f that always move forward (using actions *right* or *jump*). Compute **true** values of $V^{\pi_f}(s)$ for all states of the MDP. (2 Points)

For the forward policy π_f, we would have $V(G) = 0, V(F) = 10, V(E) = 11, V(D) = 12, V(C) = (13 + 15)/2 = 14, V(B) = 15, V(A) = 16$

(d) Consider trajectories 2, 3 and 4 from the above list of rollouts. Compute $V^{\pi_f}(s)$ for all states of the MDP using maximum likelihood estimation (certainity equivalence estimate) (2 Points) [**Hint** : A MLE based value estimation is computed from sample trajectories. For example, to compute $V(B)$ we need to compute $V(C)$ and one need to calculate state transition probabilities to go from state C to D and E respectively using samples. Use the transition probabilities obtained to compute $V(C)$.]

For the forward policy π_f, we would have $V(G) = 0, V(F) = 10, V(E) = 11, V(D) = 12, V(C) = (13 * 0.33 + 15 * 0.66) = 14.333, V(B) = 15.333, V(A) = 16.333$

Correction! This question should not include trajectory 4 as it cannot be generated by forward policy. Hence, using the only trajectories 2 and 3, the answer should be same as part (c) (Pls. check)

(e) Suppose, using policy π_f, we collect infinitely many trajectories of the above MDP. If we compute the value function V^{π_f} using Monte Carlo and TD(0) evaluations, would the two methods converge to the same value function ? Justify your answer. (2 Points)

(f) Fill in the blank cells of the table below with the Q-values that result from applying the Q-learning update for the 4 transitions specified by the episode below. You may leave Q-values that are unaffected by the current update blank. Use learning rate $\alpha = 0.5$. Assume all Q-values are initialized to 0. (2 Points)

If we have infinitely many trajectories from a policy π, then the value function estimates using the MC method will converge to the true value function due to law of large numbers. For the TD(0), the MDP needs to be Markovian which in our example holds true. If the choice of the learning rate of the TD(0) algorithm, obeys the Robbins-Monroe condition, generally TD(0) methods converge to the true value function.

[Note : If proper reasoning is not given, then marks are reduced accordingly.]

s	a	r	s	a	r	s	a	r	s	a	r	s
C	jump	4	E	right	1	F	left	-2	E	right	+1	F

	Q(C, left)	Q(C, jump)	Q(E, left)	Q(E, right)	Q(F, left)	Q(F, right)
Initial	0	0	0	0	0	0
Transition 1						
Transition 2						
Transition 3						
Transition 4						

Q-Evaluations are provided in the table. A state-action is only updated when a transition is made from it. Q(C; left), Q(E; left), and Q(F; right) state-actions are never experienced and so these values are never updated. The Q-learning update rule is given by,

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

Using the above update rule, the four updates are given by,

$$\begin{aligned} 2 &= 0 + 0.5(4 + 0 - 0) \\ 0.5 &= 0 + 0.5(1 + 0 - 0) \\ -0.75 &= 0 + 0.5(-2 + 0.5 - 0) \\ 0.75 &= 0.5 + 0.5(1 + 0 - 0.5) \end{aligned}$$

On transition 2, the Qs for F are still both 0, so the update increases the value by the reward +1 times the learning rate. On transition 3, the reward of -2 and $Q(E;right) = 0:5$ are included in the update. On transition 4, $Q(F;left)$ is now -0:75 but $Q(F;right)$ is still 0 so the next update to $Q(E;right)$ uses 0 in the max over the next state's action

	Q(C, left)	Q(C, jump)	Q(E, left)	Q(E, right)	Q(F, left)	Q(F, right)
Initial	0	0	0	0	0	0
Transition 1		2				
Transition 2				0.5		
Transition 3					- 0.75	
Transition 4				0.75		

(g) After running the Q-learning algorithm using the four transitions given above, construct a greedy policy using the current values of the Q-table in states C, E and F. (1 Point)

The greedy policy in states C, E and F is given by,

$$\pi(s) = \left\{ \begin{array}{ll} \text{jump}, & \text{for } s = C \\ \text{right}, & \text{for } s = E \\ \text{right}, & \text{for } s = F \end{array} \right\}$$

Problem 2 : On Learning Rates

In any TD based algorithm, the update rule is of the following form

$$V(s) \leftarrow V(s) + \alpha_t[r + \gamma V(s') - V(s)]$$

where α_t is the learning rate at the t-th time step. In here, the time step t refers to the t-th time we are updating the value of the state s. Among other conditions, the learning rate α_t has to obey the Robbins-Monroe condition given by,

$$\begin{aligned}\sum_{t=0}^{\infty} \alpha_t &= \infty \\ \sum_{t=0}^{\infty} \alpha_t^2 &< \infty\end{aligned}$$

for convergence to true $V(s)$. Other conditions being same, reason out if the following values for α_t would result in convergence. (5 Points)

(1) $\alpha_t = \frac{1}{t}$

(2) $\alpha_t = \frac{1}{t^2}$

(3) $\alpha_t = \frac{1}{t^{\frac{2}{3}}}$

(4) $\alpha_t = \frac{1}{t^{\frac{1}{2}}}$

Generalize the above result for $\alpha_t = \frac{1}{t^p}$ for any positive real number p (i.e. $p \in \mathbb{R}^+$)

The series $\sum_{i=1}^{\infty} \frac{1}{t}$ is harmonic series and it does not converge. In fact, one can rewrite the series in the following way (by re-grouping terms)

$$\begin{aligned}\sum_{i=1}^{\infty} \frac{1}{t} &= 1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots \\ &> 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{4} + \frac{1}{4} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} + \frac{1}{8} \cdots \\ &= 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \cdots = \infty\end{aligned}$$

A generalization of the Harmonic series is the p-series (Hyperharmonic series) defined as $\sum_{i=1}^{\infty} \frac{1}{t^p}$ for any +ve real number p. The p-series converges for all $p > 1$ (overharmonic series) and diverges for all $p \leq 1$. So, one can now use the above property to get the following results.

α_t	$\sum \alpha_t$	$\sum \alpha_t^2$	Algo converges
$\frac{1}{t^2}$	$< \infty$	$< \infty$	No
$\frac{1}{t}$	∞	$< \infty$	Yes
$\frac{1}{t^{\frac{2}{3}}}$	∞	$< \infty$	Yes
$\frac{1}{t^{0.5}}$	∞	∞	No

Problem 3 : Q-Learning

Consider a single state state MDP with two actions. That is, $\mathcal{S} = \{s\}$ and $\mathcal{A} = \{a_1, a_2\}$. Assume the discount factor of the MDP γ and horizon length to be 1. The expected reward for both actions is a constant c. That is,

$$\mathbb{E}(r|a_1) = c \text{ and } \mathbb{E}(r|a_2 = c)$$

.

(a) What are true values of $Q(s, a_1)$, $Q(s, a_2)$ and $V(s)$? (1 Point)

$Q(s, a_1) - Q(s, a_2) = V(s) = c$

(b) Consider two collections of n prior samples of reward r obtained by choosing action a_1 and a_2 from state s respectively. Denote $\hat{Q}(s, a_1)$ and $\hat{Q}(s, a_2)$ to be the final sample estimates of action value function $Q(s, a_1)$ and $Q(s, a_2)$, repectively. Let $\hat{\pi}$ be a greedy policy obtained with respect to the estimated $Q(\hat{s}, a_i)$. That is,

$$\hat{\pi}(s) = \arg\max_a \hat{Q}(s, a)$$

Prove that the estimated value of the policy $\hat{\pi}$, denoted by $\hat{V}^{\hat{\pi}}$, is biased. (4 Points) The unbiasaed sample estimates for $Q(s, a_i) = \frac{1}{n}\sum_{ij=1}^{n} r_j$ for $i = \{1, 2\}$ $\hat{\pi}$ is the greedy policy with respect to $\hat{Q}$ Then

$$\begin{aligned} \hat{V}^{\hat{\pi}}(s) &= \mathbb{E}(\max(\hat{Q}(s, a_1), \hat{Q}(s, a_2)) \\ &\geq \max(\mathbb{E}(\hat{Q}(s, a_1), \hat{Q}(s, a_2)) \\ &= \max(c, c) = V^{\pi}(s) \end{aligned} \tag{1}$$

The second inequivality is due to Jensen's inequality.

(c) Let us now consider that first action a_1 always gives a reward of c whereas the second action a_2 gives a reward $c + \mathcal{N}(-0.2, 1)$ (Normal distribution with mean -0.2 and unit variance). Which is the better action to take in expectation and would the TD control algorithms Q-learning or SARSA always favor the action that is best in expectation ? Explain. (3 Points)

Action a_1 is a better action in expectation. But Q-learning and SARSA control can choose action a_2 because they use sample estimates to decide the best action

Problem 4 : Importance Sampling

Consider a single state, single time-step MDP with finite action space, such that $|\mathcal{A}| = K$. Assume discount factor $\gamma = 1$. Let $\mathcal{R}^a(r)$ denote the unknown distribution of reward r, bounded in the range $[0, 1]$, for taking action $a \in \mathcal{A}$. Suppose we have collected a dataset consisting of action-reward pairs $\{(a, r)\}$ by sampling $a \sim \pi_b$, where π_b is a stochastic behaviour policy and $r \sim \mathcal{R}^a$. Using this dateset, we now wish to estimate $V^{\pi} = \mathbb{E}_{\pi}[r|a \sim \pi]$ for some target policy π. We assume that π is fully supported on π_b.

(a) Suppose the dataset consists of a single sample (a, r). Estimate V^π using importance sampling (IS). Is the obtained IS estimate of V^π is unbiased ? Explain. (2 Points)

The unbiased IS estimate of V^π is given by $\rho \, r$ where $\rho = \frac{\pi(a|s)}{\pi_b(a|s)}$. One can argue that the estimate is unbiased in the following way.

$$V^\pi(s) = \mathbb{E}_{a\sim\pi}(r) = \mathbb{E}_{a\sim\pi_b}\left(\frac{\pi(a|s)}{\pi_b(a|s)} r\right)$$

The entity $\rho \, r$ is sample estimate of the expecation in RHS

(b) Compute

$$\mathbb{E}_{a\sim\pi_b}\left[\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)}\right]$$

(1 Point)

$$\mathbb{E}_{a\sim\pi_b}\left[\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)}\right] = \sum_{a\in\mathcal{A}}\left[\frac{\pi(a|\cdot)}{\pi_b(a|\cdot)}\pi_b(a|\cdot)\right] = 1$$

(c) For the case that π_b is a uniformly random policy $a \sim U$ (all K actions are equiprobable) and π a deterministic policy, provide an expression for importance sampling ratio. (1 Point)

$$\rho = \frac{1_{a=\pi(s)}}{1/K}$$

(d) For this sub-question, consider a special case when the reward r obtained for action a is a deterministic function, i.e $r = \mathcal{R}(a)$. For a uniform behaviour policy π_b and a deterministic target policy π, calculate the variance of V^π obtained using MC and importance sampling (IS). (5 Points)

$$\begin{aligned}
V[\rho \, r|a\sim U] &= r^2 V[\rho|a\sim U] \\
&= r^2\left(\mathbb{E}(\rho^2|a\sim U) - \mathbb{E}(\rho|a\sim U)^2\right) \\
&= r^2\left(\mathbb{E}(\rho^2|a\sim U) - 1\right) \\
&= r^2\left(\mathbb{E}(\left[\frac{1_{a=\pi(s)}}{1/K}\right]^2|a\sim U) - 1\right) \\
&= r^2(K-1)
\end{aligned}$$

(e) Derive an upper bound for the variance of the IS estimate of V^π for the general case of the reward distribution being bounded in the range $[0, 1]$. (3 Points)

$$V[\rho \, r|a\sim U] \leq \mathbb{E}(\rho^2 r^2|a\sim U) \leq \mathbb{E}(\rho^2 r^2|a\sim U) = K$$

(f) We now consider the case of multi-state (i.e $|\mathcal{S}| > 1$), multi-step MDP. We futher assume that $P(s_0)$ to be the initial start state distribution (i.e. $s_0 \sim P(s_0)$) where s_0 is the start state of the MDP. Let τ denote a trajectory (state-action sequence) given by, $(s_0, a_0, s_1, a_1, \cdots, s_t, a_t, \cdots)$ with actions $a_{0:\infty} \sim \pi_b$. Let $\mathbf{P}$ and $\mathbf{Q}$ be joint distributions, over the entire trajectory τ induced by the behaviour policy π_b and target policy π, respectively. Provide a compact expression for the importance sampling weight $\frac{P(\tau)}{Q(\tau)}$. (3 Points)

Let $\tau \sim \pi_\theta$ denote the state-action sequence given by $s_0, a_0, s_1, a_1, \cdots, s_t, a_t, \cdots$. Then, $P(\tau; \theta)$ be the probability of finding a trajectory τ with policy π

$$P(\tau; \pi) = P(s_0) \prod_{t=0}^{\infty} \pi(a_t|s_t) P(s_{t+1}|s_t, a_t)$$

$$\frac{\mathbf{P}(\tau|\pi)}{\mathbf{Q}(\tau|\pi_b)} = \frac{\mu(s_0) \prod_{t=0}^{\infty} P(s_{t+1}|s_t, a_t)\pi(a_t|s_t)}{\mu(s_0) \prod_{t=0}^{\infty} P(s_{t+1}|s_t, a_t)\pi_b(a_t|s_t)} = \prod_{t=0}^{\infty} \frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}$$

The point is that the dynamics and start state distribution gets cancelled as they don't depend on policy.

Correction ! The correct answer should be reciprocal of the what is stated above, because of the way the question is framed; because, $\mathbf{P}$ is over π_b and $\mathbf{Q}$ is over π. Also, introduced bold-face notation for trajectory probablities to cover up for the usage of iniital state distribution $P(s_0)$

ALL THE BEST