AI 3000 / CS5500 : Reinforcement Learning
Exam № 1

Course Instructor : Easwar Subramanian 18/10/2023, 5.45 PM

Problem 1 : Bellman Equations and Dynamic Programming

Let M_1 and M_2 be two identical MDPs with $|\mathcal{S}| < \infty$ and $|\mathcal{A}| < \infty$ except for reward formulation. That is, $M_1 =< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}_1, \gamma >$ and $M_2 =< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}_2, \gamma >$. Let M_3 be another MDP such that $M_3 =< \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}_1 + \mathcal{R}_2, \gamma >$. Assume the discount factor γ to be less than 1.

(a) For an arbitrary but fixed policy π, suppose we are given action value functions $Q_1^\pi(s,a)$ and $Q_2^\pi(s,a)$, corresponding to MDPs M_1 and M_2, respectively. Explain whether it is possible to combine these action value functions in a simple manner to calculate $Q_3^\pi(s,a)$ corresponding to MDP M_3. (2 Points)

Yes, it is possible to combine the two action value functions of the MDP into a single action value function for the composite MDP since the combination involve only expectation operator and it is linear in nature. (Need to explain in math)

(b) Suppose we are given optimal polices π_1^* and π_2^* corresponding to MDPs M_1 and M_2, respectively. Explain whether it is possible to combine these optimal policies in a simple manner to formulate an optimal policy π_3^* corresponding to MDP M_3. (2 Points)

Combining optimal policies is not straightforward as it involves taking care of the max operator which is a non-linear operator. Hence, optimal policies of the two MDPs cannot be combined in a straightforward fashion. (Again need to explain in math)

(c) Let v denote a value function for MDP M_1 and consider the Bellman optimallity operator given by,

$$\mathcal{L}(v) = \max_{a \in \mathcal{A}} \left[\mathcal{R}_1^a + \gamma \mathcal{P}^a v \right].$$

Prove that the Bellman optimality operator ($\mathcal{L}$) satisfies the monotoniciity property. That is, for any two value functions u and v such that $u \leq v$ (this means, $u(s) \leq v(s)$ for all $s \in \mathcal{S}$), we have $\mathcal{L}(u) \leq \mathcal{L}(v)$. (3 Points)

[**Notation** : For some action $a \in \mathcal{A}$ and $s, s' \in \mathcal{S}$, we have, $\mathcal{P}^a(s) = \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a$ and $\mathcal{R}_1^a(s) = \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^a \mathcal{R}_1(s,a,s')$]

By the defintion of Bellman optimality operator $\mathcal{L}$, for a fixed state $s \in \mathcal{S}$ and for action set $\mathcal{A}$ finite, one can conlude that there exists $a_1, a_2 \in \mathcal{A}$ (with a_1 and a_2 possibly different) such that,

$$L(u(s)) = \left[\mathcal{R}(s, a_1) + \gamma \sum_{s'} P(s'|s, a_1) u(s) \right]$$

and

$$L(v(s)) = \left[\mathcal{R}(s, a_2) + \gamma \sum_{s'} P(s'|s, a_2)v(s)\right]$$

It is then easy to observe (using the defintion of optimality operator) that,

$$L(v(s)) \geq \left[\mathcal{R}(s, a_1) + \gamma \sum_{s'} P(s'|s, a_1)v(s)\right]$$

Now, we have,

$$L(u(s)) - L(v(s)) \leq \left[\gamma \sum_{s'} P(s'|s, a_1)\left(u(s) -)v(s)\right)\right]$$

Since, $u(s) \leq v(s)$, we have,

$$L(u(s)) - L(v(s)) \leq \left[\gamma \sum_{s'} P(s'|s, a_1)\left(u(s) -)v(s)\right)\right] \leq \left[\gamma \sum_{s'} P(s'|s, a_1)\left(v(s) - v(s)\right)\right] = 0$$

Since s was choosen arbitarily, we have the desired result.

(d) For a given MDP, will value and policy iteration converge to the same optimal policy π^* ? Justify (2 Points)

No. For a given MDP, value and policy iteration need not converge to same optimal poliacy π^*

(e) For a given MDP, will value and policy iteration converge to the same optimal value function V_* ? Justify (1 Point)

Yes. For a given MDP, value and policy iteration will converge to same optimal value function V^* since all optimal policies achieve same optimal value function

Problem 2 : Importance Sampling

Consider a finite-state, finite action MDP, such that $|\mathcal{A}| = K$. Futher assume that $\mu(s)$ to be the initial start state distribution (i.e. $s_0 \sim \mu(s)$) 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 Q and P be joint distributions, over the entire trajectory τ induced by the behaviour policy π_b and a target policy π, respectively.

(a) 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.

(b) For the case when the horizon length of the MDP is 1, the behaviour policy π_b being uniform (all K actions are equiprobable) and target policy π being deterministic, provide an expression for importance sampling weight. (1 Point)

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

(c) Does the importance sampling technique rely on the transition function and the Markovian assumption in the MDP ? (2 Points)

No it does not. See answer to part (a). There is no P there.

(d) Provide an expression for expected sum of discounted rewards where the expectation is calculated under the target policy π (samples collected from behaviour policy π_b) (2 Points)

For a horizon length H, we have,

$$\mathbb{E}_{\tau\sim\pi}\left[\sum_{t=1}^{H}\gamma^{t-1}r_t\right] = \mathbb{E}_{\tau\sim\pi_b}\left[\frac{\mathbf{P}(\tau|\pi)}{\mathbf{Q}(\tau|\pi_b)}\sum_{t=1}^{H}\gamma^{t-1}r_t\right] = \mathbb{E}_{\tau\sim\pi_b}\left[\prod_{t=0}^{H}\frac{\pi(a_t|s_t)}{\pi_b(a_t|s_t)}\sum_{t=1}^{H}\gamma^{t-1}r_t\right]$$

(e) For a given state $s \in \mathcal{S}$, compute

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

(2 Points)

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

Problem 3 : On Value Functions

Consider the following grid world problem shown in Figure 1 below. The grid has a terminal state (labelled as G) with payoff +1. The agent could start from state A or state B. As usual, the agent has four possible actions, namely, $\mathcal{A} = \{\text{Left}, \text{Right}, \text{Up}, \text{Down}\}$ and actions that take the agent off the grid leave the state unchanged. Further there is a wall between the two middle states of row two and row three (see the dots) and hence movements between those four squares are not possible. For example, there can be no movement from State A to the upwards square. Actions that take the agent to any intermediate state does not fetch any reward. The environment is determinsitc and assume that the discount factor $\gamma = 1$.

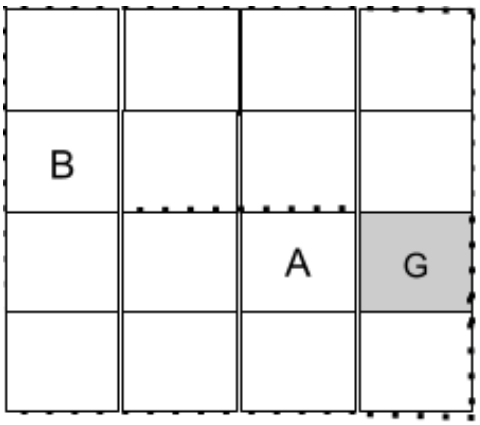

Figure 1: Grid World

(a) What is the optimal value of state A, $V_*(A)$? (1 Point)

1

(b) What is the optimal value of state B, $V_*(B)$? (2 Points)

1

The reason the answers are the same for both (b) and (a) is that there is no penalty for existing. With a discount factor of 1, reaching the goal at any future step is just as valuable as reaching it on the next step. An optimal policy will definitely find the goal state, so the optimal value of any state is always 1.

(c) If we run the value iteration algorithm to find $V_*(\cdot)$, by initializing value of all states to 0, that is, $V_0(s) = 0$, for all $s \in \mathcal{S}$, at what iteration k, will $V_k(B)$ first be non-zero ? (2 Points)

4. The value function at iteration k is equivalent to the maximum reward possible within k steps of the state in question which is B. Since the food pellet is exactly 4 steps away from Pacman in state B, $V_4(B) = 1$ and $V_{k<4}(B) = 0$.

(d) How does the optimal Q-value function at state A, for actions L and R compare ? (2 Points)

$Q^*(A, L) = Q^*(A, R)$. Once again, since $\gamma = 1$, the optimal value of every state is the same, since the optimal policy will eventually find the goal

(e) In the current formulation of the MDP, is the optimal policy found guaranteed to produce the shortest path from the start state of the agent ? If not, how could one modify the MDP formulation to make the optimal policy produce the shortest path from the start state of the MDP. (3 Points)

No. The Q-values for going Left and Right from state A are equal so there is no preference given to the shortest path to the goal state. Adding a negative living reward (example: -1 for every time step) will help differentiate between two paths of different lengths. Setting $\gamma < 1$ will make rewards seen in the future worth less than those seen right now, incentivizing the agent to arrive at the goal as early as possible.

ALL THE BEST