Assignment 5: Reinforcement Learning

AIGS/CSED515 Machine Learning

Instructor: Jungseul Ok

jungseul@postech.ac.kr

Due: 20:00pm Dec 23, 2020

Remarks

- Group study and open discussion via LMS board are encouraged, however, assignment that your hand-in must be **of your own work**, and **hand-written** unless you're asked a coding task.
- Submit a scanned copy of your answer on LMS online in **a single PDF file**, to which you also print and add your code if apply, i.e., no zip file, just a single PDF containing everything.
- Delayed submission may get some penalty in score: 5% off for delay of $0 \sim 4$ hours; 20% off for delay of $4 \sim 24$ hours; and delay longer than 24 hours will not be accepted.

1. (Life MDP) Consider an MDP, which may give a short lesson on how we live or how we face this final exam, with four states $\{-1, 0, 1, 2\}$, at each of which two actions (+: try, −: give-up) are available, and the reward and state transition have no randomness. Figure 1 summaries the reward function and state transition. We want to find optimal deterministic policy $\pi_* : \mathcal{S} \to \mathcal{A}$ maximizing the cumulated reward with discount factor γ on infinite horizon, i.e., $\pi_*(s) \in \arg\max_\pi v_\pi(s) \forall s \in \mathcal{S}$ where $v_\pi(s) := \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R_{t+1} | S_0 = s\right]$.

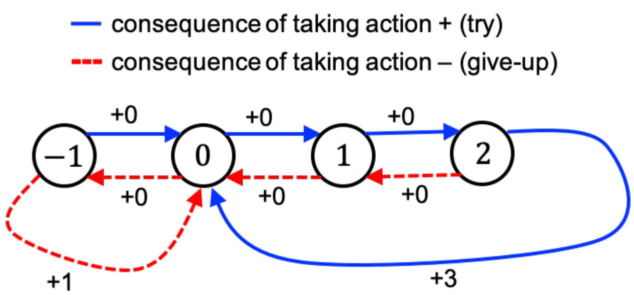

Figure 1: The reward $r(s, a)$ of taking action a at state s is non-zero only for $(s, a) \in \{(-1, -), (2, +)\}$, where $r(-1, -) = 1$ and $r(2, +) = 3$. The next state when taking action a at state s is the state which the corresponding arrow head is pointing at.

(a) [2 pt] Consider π that selects (+: try) at every state. Then, its value function at state 2 is computed as follows:

$$v_\pi(1) = 0 + 3\gamma + 0 + 0 + 3\gamma^4 + 0 + 0 + 3\gamma^7 + 0 + \ldots = \frac{3\gamma}{1 - \gamma^3} . \tag{1}$$

Bellman equation can be written as follows:

$$v_\pi(s) = \sum_a \pi(a \mid s) \sum_{r,s'} p(r, s' \mid s, a)\big(r + \gamma v_\pi(s')\big) . \tag{2}$$

Using $v_\pi(1)$ in (1) and Bellman equation in (2), compute $v_\pi(0)$ and $v_\pi(-1)$.

(b) [1 pt] Is the optimal action $\pi_*(-1)$ at state (-1) always (-: give-up) for any discount factor $\gamma \in [0, 1)$?

(c) [1 pt] When discount factor γ is 0.0001, which action is the optimal action $\pi_*(0)$ at state 0, (+: try) or (-: give-up)?

(d) [1 pt] When discount factor γ is 0.9999, which action is the optimal action $\pi_*(0)$ at state 0, (+: try) or (-: give-up)?

(e) [1 pt] Is the optimal action constant for all discount factor?

2. [Policy Iteration] Consider an MDP of finite state space $\mathcal{S}$ and action space $\mathcal{A}$ on infinite horizon with discount factor $\gamma \in [0, 1)$. Assume bounded reward, i.e., $|R_t| < \infty$. From the sequence of questions asking properties of Bellman operator and value function, you will show the convergence of policy iteration to the optimal policy described in the following:

   - **Initialization**: pick a deterministic policy π_0
   - **Loop**: for $n = 0, 1, ...$
     - **Evaluation**: Obtain $V_n = V^{\pi_n}$, i.e., solve $V_n = \mathcal{B}^{\pi_n} V_n$
     - **Improvement**: Update π_{n+1} s.t. $\forall s \in \mathcal{S}$,

       $$\pi_{n+1}(s) = \arg\max_{a \in \mathcal{A}(s)} r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V(s')$$

     - Stop if $\pi_n = \pi_{n+1}$
   - Return π_{n+1}

   (a) [2 pt] Prove that the Bellman operator $\mathcal{B}^\pi$ for stationary policy π is contraction with factor γ, i.e., for any $V, V' \in \mathbb{R}^{|\mathcal{S}| \times 1}$

   $$\|\mathcal{B}^\pi V' - \mathcal{B}^\pi V\|_\infty \leq \gamma \|V' - V\|_\infty ,$$

   where $\|V\|_\infty := \max_s V(s)$ and

   $$(\mathcal{B}^\pi V)(s) := \sum_{a \in \mathcal{A}(s)} \pi(a \mid s) \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V(s') \right) .$$

   (b) [2 pt] Using (a), show the uniqueness and existence of value function V^π for any stationary policy π.

   (c) [2 pt] Prove the monotonicity of values in policy iteration, i.e., for the successive value functions V_n and V_{n+1} in the policy iteration,

   $$V_{n+1} \geq V_n , \quad \text{i.e.,} V_{n+1}(s) \geq V_n(s) \quad \forall s \in \mathcal{S} .$$

   (d) [2 pt] Prove that the optimal Bellman operator $\mathcal{B}^*$ is contraction with factor γ, i.e., for any $V, V' \in \mathbb{R}^{|\mathcal{S}| \times 1}, \|\mathcal{B}^* V' - \mathcal{B}^* V\|_\infty \leq \gamma \|V' - V\|_\infty$, where $(\mathcal{B}^* V)(s) := \max_{a \in \mathcal{A}(s)} \left(r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s' \mid s, a) V(s') \right)$.

   (e) [2 pt] Show the uniqueness and existence of value function V^*.

   (f) [2 pt] Show that V_n converges to V^* and thus π_n converges to optimal policy, where V^* is the optimal value function, i.e., $V^*(s) = \max_{\pi \in \text{stationary}} V^\pi(s)$ for all $s \in \mathcal{S}$.

3. [Q-learning] Consider an MDP with state space $\mathcal{S} = \{1, 2, 3\}$ and action space $\mathcal{A} = \{-, +\}$, where state 3 is only terminal state. We evaluate policy π using discounted value $v_\pi(s) := \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma R_{t+1} \mid S_0 = s \right]$ where $\gamma = 0.5$ is discount factor and T is terminating time, i.e., $S_T = 3$. For this MDP, we will perform Q-learning in Algorithm 1.

Algorithm 1 Q-learning

Initialize $Q(s, a) = 0$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$
repeat (for each episode)
    Initialize $S = 1$
    repeat (for each step of episode)
        Choose A from S using some behavior policy
        Take action A, observe R, S'
        $Q(S, A) \leftarrow Q(S, A) + 0.5[R + \max_{a \in \mathcal{A}} Q(S', a) - Q(S, A)]$
        $S \leftarrow S'$
    until S is terminal, i.e., $S = 3$
until Q converges

(a) [2 pt] Suppose that in the first episode, we observe the sequence of state transitions and rewards in Table 1. Compute $Q(\cdot, \cdot)$ after the first episode.

S_0	A_0	R_1	S_1	A_1	R_2	S_2
1	$+$	**-1**	2	$+$	**1**	3

Table 1: The first episode

	$-$	$+$
1	0	
2	0	
3	0	0

Table 2: $Q(\cdot, \cdot)$ after the first episode in Table 1

(b) [3 pt] Suppose that in the second episode, we observe the sequence of state transitions and rewards in Table 3. Compute $Q(\cdot, \cdot)$ after the second episode.

S_0	A_0	R_1	S_1	A_1	R_2	S_2	A_2	R_3	S_3
1	$-$	**-1**	1	$+$	**-1**	2	$+$	**1**	3

Table 3: The second episode

	$-$	$+$
1		
2	0	
3	0	0

Table 4: $Q(\cdot, \cdot)$ after the second episode in Table 3

(c) [2 pt] Suppose that after few hundreds of episodes, we have the convergence of $Q(\cdot, \cdot)$ in Table 5. What are the optimal actions at state 1 and state 2?

	$-$	$+$
1	-1	0
2	-1	1
3	0	0

Table 5: $Q(\cdot, \cdot)$ after convergence

(d) [0 pt] This link provides a simple implementation of Q-learning for this specific MDP. Have a fun with it. (e.g., does it converge to the optimal Q-function in Table 5 if a trajectory with full of transition $(S, A, R, S') = (1, -, -1, 1)$ is sampled? When does Q-learning converge to the optimal Q-function?)