Homework 1

ELEN E6885: Introduction to Reinforcement Learning

September 25, 2022

Problem 1 (2-Armed Bandit, 20 Points)

Consider the following 2-armed bandit problem: the first arm has a fixed reward 0.3 and the second arm has a 0-1 reward following a Bernoulli distribution with probability 0.6, i.e., arm 2 yields reward 1 with probability 0.6. Assume we selected arm 1 at $t = 1$, and arm 2 four times at $t = 2, 3, 4, 5$ with reward $0, 1, 0, 0$, respectively. We use the sample-average technique to estimate the action-value, and then use it to guide our choices starting from $t = 6$.

1. [5 pts] Which arm will be played at $t = 6, 7$, respectively, if the greedy method is used to select actions?

2. [10 pts] What is the probability to play arm 2 at $t = 6, 7$, respectively, if the ϵ-greedy method is used to select actions ($\epsilon = 0.1$)?

3. [5 pts] Why could the greedy method perform significantly worse than the ϵ-greedy method in the long run?

Solution. We use Q_t^i to represent the estimated expected payoff for arm i ($i = 1, 2$) after time t.

1. $Q_5^1 = 0.3$ (2 pts) and $Q_5^2 = (0 + 1 + 0 + 0)/4 = 0.25$ (2 pts). Therefore, arm 1 will be played at both $t = 6$ and $t = 7$ if the greedy method is used.(1 pts)

2. At $t = 6$, the probability to play arm 2 using ϵ-greedy method is $\epsilon/2 = 0.05$. (2 pts) For $t = 7$, consider the following three cases:

   a. Arm 1 is played at $t = 6$: Its probability is $1 - \epsilon + \epsilon/2 = 0.95$. We have $Q_6^1 = 0.3$ and $Q_6^2 = 0.25$. Therefore, the probability to play arm 2 at $t = 7$ is $\epsilon/2 = 0.05$. (2 pts)

   b. Arm 2 is played at $t = 6$ and the reward is 0: Its probability is $\epsilon/2 \times 0.4 = 0.02$. We have $Q_6^1 = 0.3$ and $Q_6^2 = (0+1+0+0+0)/5 = 0.2$. Therefore, the probability to play arm 2 at $t = 7$ is $\epsilon/2 = 0.05$. (2 pts)

   c. Arm 2 is played at $t = 6$ and the reward is 1: Its probability is $\epsilon/2 \times 0.6 = 0.03$. We have $Q_6^1 = 0.3$ and $Q_6^2 = (0+1+0+0+1)/5 = 0.4$. Therefore, the probability to play arm 2 at $t = 7$ is $1 - \epsilon + \epsilon/2 = 0.95$. (2 pts)

By combining all above possibilities, the probability to play arm 2 at $t = 7$ using ϵ-greedy method is

$$0.95 \times 0.05 + 0.02 \times 0.05 + 0.03 \times 0.95 = 0.077. (2pts)$$

3. Greedy method may get stuck in the suboptimal action due to the lack of exploration. (5 pts)

Problem 2 (Softmax, 15 Points)

For the softmax action selection, show the following.

1. [5 pts] In the limit as the *temperature* $\tau \to 0$, softmax action selection becomes the same as greedy action selection.

2. [5 pts] In the limit as $\tau \to \infty$, softmax action selection yields equiprobable selection among all actions.

3. [5 pts] In the case of two actions, the softmax operation using the Gibbs distribution becomes the logistic (or sigmoid) function commonly used in artificial neural networks.

Solution.

1. Let the actions be $a_1, a_2, ...a_n$ and the action with the maximum Q-value (i.e., the greedy action) be a_m. Then, we have $Q(a_m) > Q(a_i)$ for all $i \neq m$. The probability of choosing action a_i using the softmax action selection is

$$\begin{aligned} p_i &= \frac{\exp(Q(a_i)/\tau)}{\sum_{j=1}^{n} \exp(Q(a_j)/\tau)} \\ &= \frac{\exp\left(\frac{Q(a_i)-Q(a_m)}{\tau}\right)}{1 + \sum_{j \neq m} \exp\left(\frac{Q(a_j)-Q(a_m)}{\tau}\right)} (3pts) \end{aligned} \tag{1}$$

Since $Q(a_m) > Q(a_j)$ for all $j \neq m$, it implies that for $j \neq m$,

$$\lim_{\tau \to 0} \exp\left(\frac{Q(a_j) - Q(a_m)}{\tau}\right) = 0. (1pts)$$

Therefore, (1) reduces to

$$p_i = \begin{cases} 1, & \text{if } i = m, \\ 0, & \text{if } i \neq m. (1pts) \end{cases}$$

This is exactly the greedy action selection policy.

2. When $\tau \to \infty$, we have for all j,

$$\lim_{\tau \to \infty} \exp\left(\frac{Q(a_j) - Q(a_m)}{\tau}\right) = 1. (3pts)$$

Substituting this fact into (1), we have $p_i = 1/n$ for all $i = 1, 2, \cdots, n$. (2 pts) Therefore, all actions are selected with equal probability.

3. When there are only two actions a_1 and a_2, the probability of choosing action a_1 is

$$\begin{aligned} p_1 &= \frac{\exp\left(Q(a_1)/\tau\right)}{\exp\left(Q(a_1)/\tau\right) + \exp\left(Q(a_2)/\tau\right)} \\ &= \frac{1}{1 + \exp\left(\frac{-(Q(a_1) - Q(a_2))}{\tau}\right)}. (3pts) \end{aligned}$$

This is the form of the logistic/sigmoid function, i.e., $f(x) = \frac{1}{1+e^{-x}}$. (2 pts)

Problem 3 (**Incremental Implementation, 15 Points**)

Suppose we have a sequence of returns $G_1, G_2, \cdots, G_{n-1}$, all starting in the same state and each with a corresponding random weight W_i, $i = 1, 2, \cdots, n-1$. We wish to form the estimate

$$V_n = \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k}, \; n \geq 2,$$

and keep it up-to-date as we obtain an additional return G_n. In addition to keeping track of V_n, we must maintain for each state the cumulative sum C_n of the weights given to the first n returns. Show that the update rule for $V_{n+1}, n \geq 1$ is

$$V_{n+1} = V_n + \frac{W_n}{C_n}\Big[G_n - V_n\Big],$$

and

$$C_{n+1} = C_n + W_{n+1},$$

where $C_0 = 0$ (and V_1 is arbitrary and thus need not be specified).

Solution. By definition, C_n is the cumulative sum of weights, i.e., $C_n = \sum_{k=1}^{n} W_k$. Thus, we have

$$C_{n+1} = C_n + W_{n+1}.$$

And for $n \geq 1$,

$$\begin{aligned} V_{n+1} &= \frac{\sum_{k=1}^{n} W_k G_k}{\sum_{k=1}^{n} W_k} \\ &= \frac{\sum_{k=1}^{n-1} W_k G_k + W_n G_n}{C_n} (3pts) \\ &= \frac{C_{n-1} V_n + W_n G_n}{C_n} (3pts) \\ &= \frac{(C_n - W_n) V_n + W_n G_n}{C_n} (3pts) \\ &= \frac{C_n V_n + W_n (G_n - V_n)}{C_n} (3pts) \\ &= V_n + \frac{W_n}{C_n} (G_n - V_n). (3pts) \end{aligned}$$

Problem 4 (**MDP: State-Value Function, 15 Points**)

Compute the state value for S_t in all the MDPs in Fig. 1 – 3. The decimal number above lines refers to the probability of choosing the corresponding action. The value r refers to reward, which can be deterministic or stochastic. Assume $\gamma = 1$ for all questions and all terminal states (i.e., no successors in the graph) always have zero values.

Solution.

1. We have $v(S_{t+2}) = 0$. By Bellman expectation equation,

$$v(S_{t+1}) = \sum_a \pi(a|S_{t+1}) \left(r + \gamma v(S_{t+2})\right) = 0.2 \times 6 + 0.3 \times 8 + 0.5 \times 10 = 8.6, (3pts)$$

$$v(S_t) = r + \gamma v(S_{t+1}) = 4 + 8.6 = 12.6. (2pts)$$

2. We have $v(S_{t+2}) = 0$, and $v(S_{t+1}) = 0$ for S_{t+1} on the top of S_t and the bottom-right of S_t. By Bellman expectation equation,

$$\begin{aligned} v(S_{t+1}) &= \sum_a \pi(a|S_{t+1}) \left(r + \gamma v(S_{t+2})\right) \\ &= 0.2 \times 6 + 0.3 \times 8 + 0.5 \times 10 = 8.6, (2pts) \end{aligned}$$

for S_{t+1} on the top-right of S_t, and

$$\begin{aligned} v(S_t) &= \sum_a \pi(a|S_t) \sum_{S_{t+1}} \sum_r p(S_{t+1}, r|S_t, a) \left(r + \gamma v(S_{t+1})\right) \\ &= 0.5 \times (4 + 0) + 0.5 \times [0.4 \times (4 + 8.6) + 0.6 \times (4 + 0)] \\ &= 2 + 0.5(0.4 \times 12.6 + 2.4) \\ &= 5.72. (3pts) \end{aligned}$$

3. We have $v(S_{t+1}) = 0$(1 pts). By Bellman expectation equation,

$$\begin{aligned} v(S_t) &= \sum_a \pi(a|S_t) \sum_{S_{t+1}} \sum_r p(S_{t+1}, r|S_t, a)\,(r + \gamma v(S_{t+1})) \\ &= 0.5 \times (4 + 0) + 0.5 \times (0.4\mu_{r_{11}} + 0.6\mu_{r_{12}}) \\ &= 2 + 0.5 \times (0.4 \times 0.5 + 0.6 \times 5) \\ &= 3.6(4pts) \end{aligned}$$

Problem 5 (**MDP: Optimal Policy, 10 Points**)

Given an arbitrary MDP with reward function $\mathcal{R}_s^a$ and constants α and $\beta > 0$, prove that the following modified MDPs have the same optimal policy as the original MDP.

1. [5 pts] Everything remains the same as the original MDP, except it has a new reward function $\alpha + \mathcal{R}_s^a$. Assume that there is no terminal state and discount factor $\gamma < 1$.

2. [5 pts] Everything remains the same as the original MDP, except it has a new reward function $\beta \cdot \mathcal{R}_s^a$.

Solution. Let $Q_\pi(s, a)$ and $Q'_\pi(s, a)$ denote the action-value functions of the original MDP and the modified MDP following the same policy π, respectively. We will prove that $Q'_\pi(s, a) = Q_\pi(s, a) + \frac{\alpha}{(1-\gamma)}$ for 1) and $Q'_\pi(s, a) = \beta Q_\pi(s, a)$ for 2).

1. For the original MDP,

$$\begin{aligned} Q_\pi(s, a) &= \mathsf{E}_\pi[G_t|S_t = s, A_t = a] \\ &= \mathsf{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...|S_t = s, A_t = a].(1pts) \end{aligned}$$

   Similarly for the modified MDP,

$$\begin{aligned} Q'_\pi(s, a) &= \mathsf{E}_\pi[G'_t|S_t = s, A_t = a] \\ &= \mathsf{E}_\pi[R'_{t+1} + \gamma R'_{t+2} + \gamma^2 R'_{t+3} + ...|S_t = s, A_t = a] \\ &= \mathsf{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ... + \alpha + \gamma\alpha + \gamma^2\alpha + ...|S_t = s, A_t = a].(2pts) \end{aligned}$$

   Now, as there is no terminal state and $\gamma < 1$,

$$\begin{aligned} Q'_\pi(s, a) &= \mathsf{E}_\pi[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + ...|S_t = s, A_t = a] + \frac{\alpha}{1-\gamma} \\ &= Q_\pi(s, a) + \frac{\alpha}{1-\gamma}.(2pts) \end{aligned}$$

2. Similar to 1, we see that

$$\begin{aligned} Q'_\pi(s,a) &= \mathsf{E}_\pi[G'_t|S_t=s,A_t=a] \\ &= \mathsf{E}_\pi[R'_{t+1}+\gamma R'_{t+2}+\gamma^2 R'_{t+3}+...|S_t=s,A_t=a](1pts) \\ &= \mathsf{E}_\pi[\beta R_{t+1}+\gamma\beta R_{t+2}+\gamma^2\beta R_{t+3}+...|S_t=s,A_t=a](2pts) \\ &= \beta\,\mathsf{E}_\pi[R_{t+1}+\gamma R_{t+2}+\gamma^2 R_{t+3}+...|S_t=s,A_t=a](2pts) \\ &= \beta Q_\pi(s,a). \end{aligned}$$

Now for 1), we have

$$Q'_*(s,a)=\max_\pi Q'_\pi(s,a)=\max_\pi Q_\pi(s,a)+\frac{\alpha}{1-\gamma}=Q_*(s,a)+\frac{\alpha}{1-\gamma},$$

and for 2), we have

$$Q'_*(s,a)=\max_\pi Q'_\pi(s,a)=\beta\max_\pi Q_\pi(s,a)=\beta Q_*(s,a).$$

Therefore, the modified MDP has the same (deterministic) optimal policy as the original MDP.

Problem 6 (**MDP: Simple Card Game, 25 Points**)

In a card game, you repeatedly draw a card (with replacement) that is equally likely to be a 2 or 3. You can either *Draw* or *Stop* if the total score of the cards you have drawn is less than 6. Otherwise, you must Stop. When you Stop, your reward is equal to your total score (up to 5), or zero if you get a total of 6 or higher. When you Draw, you receive no reward. Assume there is no discount ($\gamma = 1$). We formulate this problem as an MDP with the following states: $0, 2, 3, 4, 5$, and a *Done* state for when the game ends.

1. [10 pts] What is the state transition function and the reward function for this MDP?
2. [12 pts] What is the optimal state-value function and optimal action-value function for this MDP? (Hint: Solve Bellman optimality equation starting from states 4 and 5.)
3. [3 pts] What is the optimal policy for this MDP?

Solution.

1. Transition function: (1 pts each)

   $P^{Stop}_{s,Done}=1,\ s\in\{0,2,3,4,5\}$;
   $P^{Draw}_{0,s'}=0.5,\ s'\in\{2,3\}$;
   $P^{Draw}_{2,s'}=0.5,\ s'\in\{4,5\}$;
   $P^{Draw}_{3,s'}=0.5,\ s'\in\{5,Done\}$;
   $P^{Draw}_{s,Done}=1,\ s\in\{4,5\}$.

   Reward function: (2.5 pts each)

   $R^{Stop}_s=s,\ s\in\{0,2,3,4,5\}$;
   $R^a_s=0$, otherwise.

2. Optimal state-value function: (1 pts each)

   $Q_*(5, Draw) = 0,\ Q_*(5, Stop) = 5;$
   $Q_*(4, Draw) = 0,\ Q_*(4, Stop) = 4;$
   $Q_*(3, Draw) = 2.5,\ Q_*(3, Stop) = 3;$
   $Q_*(2, Draw) = 4.5,\ Q_*(2, Stop) = 2;$
   $Q_*(0, Draw) = 3.75,\ Q_*(0, Stop) = 0.$

   Optimal action-value function: (0.4 pts each)

   $V_*(5) = 5,\ V_*(4) = 4,\ V_*(3) = 3,\ V_*(2) = 4.5,\ V_*(5) = 3.75.$

3. Optimal policy:

$$\pi_*(a|s) = \begin{cases} 1, & \text{if } a = \text{Draw}, s \in \{0, 2\} \text{ or } a = \text{Stop}, s \in \{3, 4, 5\}, (1.5pts) \\ 0, & \text{otherwise.} (1.5pts) \end{cases}$$

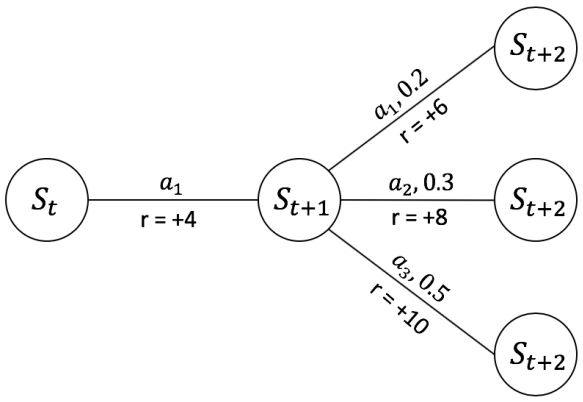

Figure 1: MDP with deterministic transitions

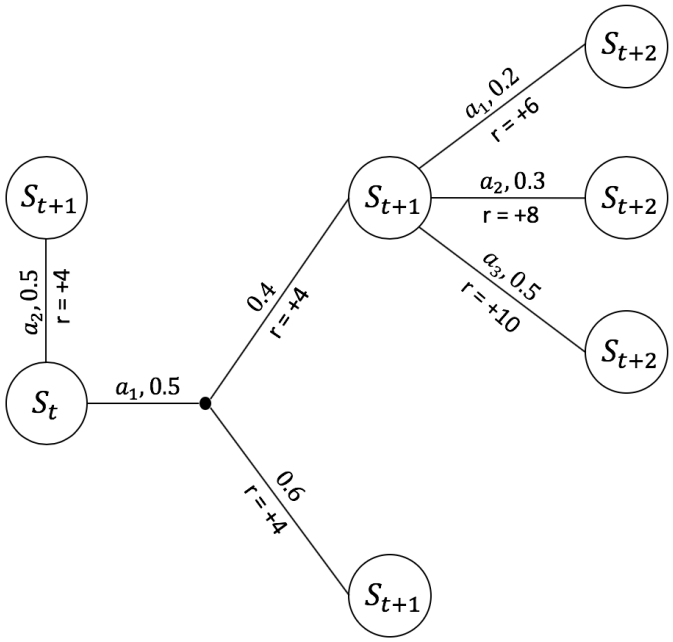

Figure 2: MDP with stochastic transitions

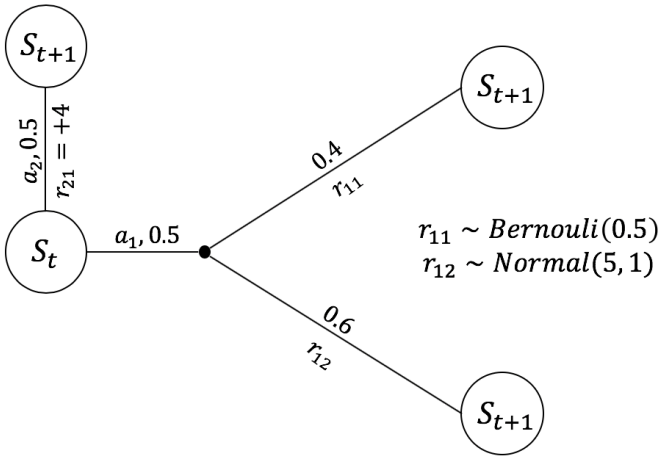

Figure 3: MDP with stochastic rewards