ECE 239AS - Problem Set 1

Swapnil Sayan Saha; UID: 605353215 Date of submission: 04/29/20

1 Exercise 2:

If we are not able to tell which case we face at any step, since the optimal arm alters between a_1 and a_2 for case A and B, the best approach would be to select arms at random for each time-step, assuming each case occurs equally. Thus, the expected maximum reward is as follows:

$$E(a_1) = (0.5 \times 0.1) + (0.5 \times 0.9) = 0.5$$
$$E(a_2) = (0.5 \times 0.2) + (0.5 \times 0.8) = 0.5$$
$$E(\max(r_a)) = P(A) \times E(a_2) + P(B) \times E(a_1) = 0.5 \times 0.5 + 0.5 \times 0.5 = \boxed{0.5}$$

If we are told what case we are facing, then the optimal behavior would be to explore all elements in the action-value pair set $\{V(A, a_1), V(A, a_2), V(B, a_1), V(B, a_2)\}$ to identify the optimal arm to pull for each case at least once, and then pull the arms with the optimal value repeatedly for rest of the rounds (exploitation). This is intuitively similar to Explore-First algorithm, where we consider we consider the two cases as independent bandit problems. The expected maximum reward is: $0.5 \times 0.2 + 0.5 \times 0.9 = \boxed{0.55}$

2 Exercise 5:

1:

$\mathbb{E}[a_1] = 0.5 + 0.5 \times p_1 \times 0.5 + 0.5 \times p_2 \times 0.5 + 0.5 \times (1 - p_2)$ (for two pulls)
$\mathbb{E}[a_2] = 0.5 + 0.5 \times p_1 \times 0.5 + 0.5 \times p_2 \times 0.5 + 0.5 \times (1 - p_1)$(for two pulls)
Hence, $\mathbb{E}[z^2] = \mathbb{E}[a_1] \times (1 - p_1) + \mathbb{E}[a_2] \times (1 - p_2) = \boxed{2 - p1 - p2 - 0.25p_1^2 - 0.25p_2^2 + 0.5p_1p_2}$

2:

The sequence of actions, H performed by L is as follows: Pull a_1 (conversely a_2) repeatedly until 0 reward is reached, thereby switch to a_2 (conversely a_1) and pull repeatedly until 0 reward is reached, repeat. Expected length of H is given by:

$$\mathbb{E}(H_l) = \mathbb{E}(a_1) + \mathbb{E}(a_2) = (1 - p_1)(1 + 2p_1 + 3p_1^2 + ...) + (1 - p2)(1 + 2p_2 + 3p_2^2 + ...)$$
$$\Rightarrow \mathbb{E}(H_l) = \frac{1}{1-p_1} + \frac{1}{1-p_2}$$

The number of 0 rewards in H is 2. Hence, if there are a^T number of H occurring in T pulls, we have:

$$\lim_{T\to\infty} \frac{\mathbb{E}[z^T]}{T} = \lim_{T\to\infty} \frac{\mathbb{E}[2a^T]}{T} = 2\lim_{T\to\infty} \frac{\mathbb{E}[a^T]}{T}$$

As $T \to \infty$, $\frac{\mathbb{E}[a^T]}{T} \to \frac{1}{\mathbb{E}(H_l)}$. Hence,

$$\lim_{T\to\infty} \frac{\mathbb{E}[z^T]}{T} = \lim_{T\to\infty} \frac{\mathbb{E}[2a^T]}{T} = \boxed{\frac{2}{\frac{1}{1-p_1}+\frac{1}{1-p_2}}}$$

3:

Maximum possible reward from T pulls is given by Tp_1. Cumulative reward from T pulls is given by: $T - z^T$. Hence:

$$R^T = Tp_1 - (T - z^T).$$

Thus,

$$\lim_{T\to\infty} \frac{\mathbb{E}[R^T]}{T} = p_1 - 1 + \lim_{T\to\infty} \frac{\mathbb{E}[z^T]}{T} = p_1 - 1 + \frac{2(1-p_1)(1-p_2)}{2-p_1-p_2} = \boxed{\frac{(1-p_1)(p_1-p_2)}{2-p_1-p_2}}$$

3 Exercise 6:

$$Q_{n+1} = Q_n + \frac{\alpha}{\overline{o}_n}(R_n - Q_n)$$
$$\Rightarrow \overline{o}_n Q_{n+1} = \alpha R_n - (\overline{o}_n - \alpha)Q_n$$
$$\Rightarrow \overline{o}_n Q_{n+1} = \alpha R_n - (\overline{o}_{n-1} + \cancel{\alpha} - \alpha\overline{o}_{n-1} - \cancel{\alpha})Q_n = \alpha R_n - (1-\alpha)\overline{o}_{n-1}Q_n$$
$$\Rightarrow \overline{o}_n Q_{n+1} = \alpha R_n - (1-\alpha)(\alpha R_{n-1} - (\overline{o}_{n-1} - \alpha)Q_{n-1})$$
$$\Rightarrow \overline{o}_n Q_{n+1} = \alpha R_n - (1-\alpha)(\alpha R_{n-1} - (1-\alpha)\overline{o}_{n-2}Q_{n-1}) = \alpha R_n - (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2\overline{o}_{n-2}Q_{n-1}$$

Expanding the sequence until Q_1, we have:

$$\Rightarrow \overline{o}_n Q_{n+1} = \alpha R_n - (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2\alpha R_{n-2} + ... + (1-\alpha)^{n-1}\alpha R_1 + (-1)^n(1-\alpha)^n\overline{o}_0 Q_1$$

But, $\overline{o}_0 = 0$, hence:

$$\Rightarrow \overline{o}_n Q_{n+1} = \alpha R_n - (1-\alpha)\alpha R_{n-1} + (1-\alpha)^2\alpha R_{n-2} + ... + (1-\alpha)^{n-1}\alpha R_1$$
$$\Rightarrow \overline{o}_n Q_{n+1} = \sum_{i=1}^{n} \alpha(1-\alpha)^{i-1} R_i \text{ } (\textbf{proved})$$

4 Exercise 7:

$$q_\pi(s,a) = \mathbb{E}_\pi\{G_t | s_t = s, a_t = a\} = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} | s_t = s, a_t = a]$$

Using Bellman backup (or backup diagram) in Markov Decision Process (MDP), we can transform the above equation as follows:

$$q_\pi(s,a) = \sum_{s'\in S} P^a_{s',s}[R^a_{s',s} + \gamma V^\pi(s')] = \sum_{s'\in S} R^a_{s',s}P^a_{s',s} + \gamma \sum_{s'\in S} P^a_{s',s} v^\pi(s')$$

Moving down 1 step from state 11 transitions the agent into terminal state T, where, $v^\pi(T) = 0$, $P^{\text{down}}_{T,11} = 1, R^{\text{down}}_{T,11} = -1$. Hence:

$$q_\pi(11,\text{ down}) = -1 \times 1 + \gamma \times 1 \times 0 = \boxed{\text{-1}}$$

Moving down 1 step from state 7 transitions the agent into state 11 where, $v^\pi(11) = -14$, $P^{\text{down}}_{11,7} = 1, R^{\text{down}}_{11,7} = -1$. Furthermore, we observe that $\gamma = 1$. Hence:

$$q_\pi(7,\text{ down}) = -1 \times 1 + 1 \times 1 \times -14 = \boxed{\text{-15}}.$$

5 Exercise 8:

First part:

$$v^\pi(s) = \mathbb{E}_\pi(s)\{R_t | s_t = s\} = \textstyle\sum_a \pi(s,a) q_\pi(s,a)$$
$$\Rightarrow v^\pi(s) = \textstyle\sum_a \pi(s,a)[\sum_{s' \in S} R^a_{s',s} P^a_{s',s} + \gamma \sum_{s' \in S} P^a_{s',s} v^\pi(s')]$$

Since the policy is equiprobable at each state, $\pi(s,a) = 0.25$. For the new state 15:

$$v^\pi(15) = 0.25 \times (R^{\text{left}}_{12,15} \times P^{\text{left}}_{12,15} + R^{\text{up}}_{13,15} \times P^{\text{up}}_{13,15} + R^{\text{right}}_{14,15} \times P^{\text{right}}_{14,15} + R^{\text{down}}_{15,15} \times P^{\text{down}}_{15,15}) + 0.25 \times \gamma \times$$
$$(v^\pi(12) \times P^{\text{left}}_{12,15} + v^\pi(13) \times P^{\text{up}}_{13,15} + v^\pi(14) \times P^{\text{right}}_{14,15} + v^\pi(15) \times P^{\text{down}}_{15,15})$$

Since the steps are deterministic and $s' = a$ for all the 4 possible steps, $P^a_{s',s} = 1$ for the above steps. Furthermore $R^a_{s',s} = -1$ and $\gamma = 1$. Substituting $v(12), v(13)$ and $v(14)$ from Figure 2, we have:

$$v^\pi(15) = -1 + 0.25 \times (-22 + -20 - 14 + v^\pi(15)) = -15 + 0.25 v^\pi(15)$$
$$\Rightarrow v^\pi(15) = \boxed{\text{-20.0}}$$

Second part:

Since dynamics of state 13 has changed, we need to recalculate the value of $v^\pi(13)$ to find the value of $v^\pi(15)$ via a system of linear equations.

$$v^\pi(13) = 0.25 \times (R^{\text{left}}_{12,13} \times P^{\text{left}}_{12,13} + R^{\text{up}}_{9,13} \times P^{\text{up}}_{9,13} + R^{\text{right}}_{14,13} \times P^{\text{right}}_{14,13} + R^{\text{down}}_{15,13} \times P^{\text{down}}_{15,13}) + 0.25 \times \gamma \times$$
$$(v^\pi(12) \times P^{\text{left}}_{12,13} + v^\pi(9) \times P^{\text{up}}_{9,13} + v^\pi(14) \times P^{\text{right}}_{9,13} + v^\pi(15) \times P^{\text{down}}_{15,13})$$

Substituting $P^a_{s',s} = 1$, $R^a_{s',s} = -1$ and $\gamma = 1$ and the values of V from Figure 2:

$$\Rightarrow v^\pi(13) = -1 + 0.25 \times (-56 + v^\pi(15)) = -15 + 0.25 v^\pi(15) \tag{1}$$

Again, for state 15:

$$v^\pi(15) = 0.25 \times (R^{\text{left}}_{12,15} \times P^{\text{left}}_{12,15} + R^{\text{up}}_{13,15} \times P^{\text{up}}_{13,15} + R^{\text{right}}_{14,15} \times P^{\text{right}}_{14,15} + R^{\text{down}}_{15,15} \times P^{\text{down}}_{15,15}) + 0.25 \times \gamma \times$$
$$(v^\pi(12) \times P^{\text{left}}_{12,15} + v^\pi(13) \times P^{\text{up}}_{13,15} + v^\pi(14) \times P^{\text{right}}_{14,15} + v^\pi(15) \times P^{\text{down}}_{15,15})$$
$$\Rightarrow v^\pi(15) = -1 + 0.25 \times (-36 + v^\pi(13) + v^\pi(15)) = -10 + 0.25 v^\pi(13) + 0.25 v^\pi(15)$$

$$\Rightarrow 0.75 v^\pi(15) = -10 + 0.25 v^\pi(13) \tag{2}$$

Solving the system of equations, we have:

$$v^\pi(13) = -20.0, v^\pi(15) = \boxed{\text{-20.0}}$$

P.S. Since discount factor is 1, $v^\pi(15)$ is unchanged despite dynamics change of state 13, however, if $1 < \gamma < 0$, one would observe a decrease in the value of $v^\pi(15)$.

6 Exercise 11:

The Q value for state 34, action 7 is updated. The update expression is given by:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[R_t + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t) \tag{3}$$

Expression for new value:

$$\Rightarrow Q(34, 7) \leftarrow Q(34, 7) + \alpha[3 + \gamma \max_a Q(65, a) - Q(34, 7)]$$

7 Exercise 12:

The agent may not find the most optimal action / policy (action with highest Q-value does not guarantee than an action is optimal) as it does not explore those actions with a lower Q-value. To force the agent to explore we could:

- Use an algorithm that consists of a random exploration part (e.g. ϵ-greedy or UCB) apart from exploitation part.
- Initialize Q-values high enough such that exploration is encouraged in the beginning. When Q-values are high, the agent tends to sees all the possible states as near-optimal in the beginning and hence tends to explore.

8 Exercise 18:

1:

For first-visit Monte-Carlo:

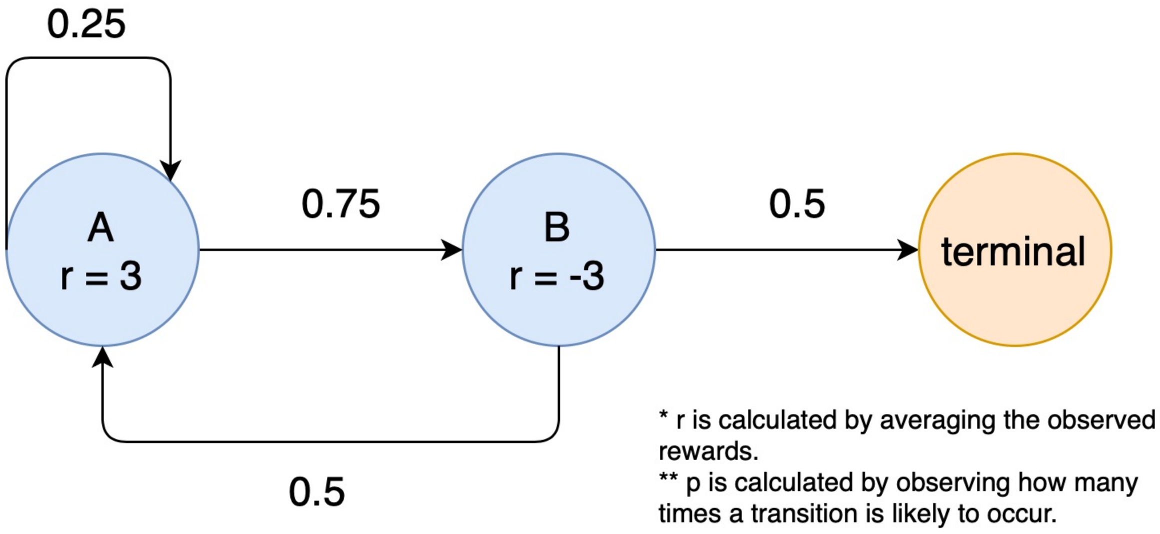
$$V(A) = \frac{(3+2-4+4-3)+(3-3)}{1+1} = \boxed{1}$$
$$V(B) = \frac{(-4+4-3)+(-2+3-3)}{1+1} = \boxed{\text{-2.5}}$$

2:

For every-visit Monte-Carlo:

$$V(A) = \frac{(3+2-4+4-3)+(2-4+4-3)+(4-3)+(3-3)}{1+1+1+1} = \boxed{0.5}$$
$$V(B) = \frac{(4+4-3)+(-3)+(-2+3-3)+(-3)}{1+1+1+1} = \boxed{\text{-2.75}}$$

3:



* r is calculated by averaging the observed rewards.
** p is calculated by observing how many times a transition is likely to occur.

Figure 1: MRP for Exercise 18.

4:

$$V = R + \gamma PV \tag{4}$$

where, $V = V(s)$ = state-value function, $R = r(s)$ = immediate reward, γ = discount factor, $P = P_{s',s}$ = state transition matrix.

5:

Assuming $\gamma = 1$:

$$V(A) = 3 + 1 \times (0.25V(A) + 0.75V(B))$$

$$\Rightarrow 0.75V(A) - 0.75V(B) = 3 \tag{5}$$

Again,

$$V(B) = -3 + 1 \times (0.5V(A))$$

$$\Rightarrow 0.5V(A) - V(B) = 3 \tag{6}$$

Solving the system of linear equations, $V(A) = \boxed{2}$, $V(B) = \boxed{\text{-2}}$

6:

TD(0) gives the same solution as solution to Bellman equation for MRP. $V(A) = \boxed{2}$ $V(B) = \boxed{\text{-2}}$

7:

TD(1) gives the same answer as every-visit Monte-Carlo. $V(A) = \boxed{0.5}$

$V(B) = \boxed{-2.75}$

9 Exercise 23:

1:

- Uniformly offset tilings across all dimensions of the state-space causes different states to be generalized in distinct ways, generating diagonal artifacts and significant variation in generalization of the states. This is not preferable as feature vectors generated by tile coding should generalize for all states homogeneously. Asymmetrically offset tilings allows the diagonal artifacts to be avoided, resulting in a feature space that is more spherical, homogenous and well centered on the training states void of any significant asymmetries.
- Asymmetrically offset tilings cover more area than uniformly offset tilings around the training state, i.e. has a greater receptive field, using same number of tiles, which is preferable. A larger receptive field helps capture more salient features/information of the training set using small number of tiles (small number of tiles reduces memory and computational resource usage), improving function approximation performance as tile coding uses binary feature vectors, making value function approximation trivial to compute.

2:

Since one of the two state dimensions is more likely to have an effect on the value function than other, we would prefer generalization primarily across (perpendicular) that dimension rather than along it. as the feature vectors biased towards that dimension would more likely generate predicted values closer to true state values. This can be achieved using stripe tilings (e.g. rectangular stripes), which are elongated along one dimension. For example, consider the case shown in Figure 2. Since the tilings are denser and thinner across the dimension more likely to affect value functions (x_1), the extracted feature vectors will be primarily biased towards that dimension, while generalization would occur along (parallel to) the other dimension (x_2).

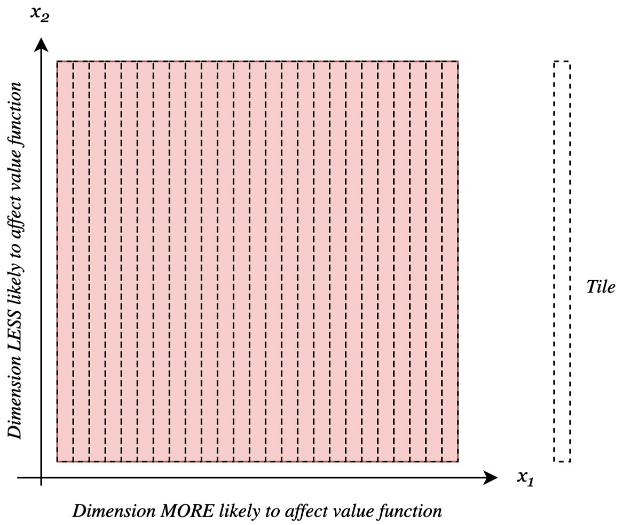

x_2
Dimension LESS likely to affect value function
Tile
x_1
Dimension MORE likely to affect value function

Figure 2: Rectangular stripes, promoting generalization perpendicular to the dimension more likely to affect value function.

10 Exercise 25:

1:

The loss function J that needs to be minimized to find final values of w_1 and w_2 (w_1^f and w_2^f) is:

$$J(w_1^f, w_2^f) = \mathbb{E}_\pi(V^\pi(s) - \widehat{V}_w(s))$$
$$\Rightarrow J(w_1^f, w_2^f) = \textstyle\sum_{s \in S} P^\pi_{s',s}(V^\pi(s) - \widehat{V}_w(s))$$
$$\Rightarrow J(w_1^f, w_2^f) = \textstyle\sum_{s \in S} P^\pi_{s',s}(V^\pi(s) - (w_1 f(s) + w_2))$$

Substituting values of $f(s)$, $V^\pi(s)$ and $P^\pi_{s',s}$, we have:

$$\Rightarrow J(w_1^f, w_2^f) = 0.5 \times (4 - 2w_1 - w_2)^2 + 0.25 \times (6 - w_1 - w_2)^2 + 0.25 \times (-5 + w_1 - w_2)^2$$

Taking partial derivatives of J with respect to w_1 and w_2, we have:

$$\nabla_{w_1} J(w_1^f, w_2^f) = (4 - 2w_1 - w_2) \times (-2) + 0.5 \times (6 - w_1 - w_2) \times (-1) + 0.5 \times (-5 + w_1 - w_2)$$
$$\Rightarrow \nabla_{w_1} J(w_1^f, w_2^f) = -13.5 + 5w_1 + 2w_2$$
$$\nabla_{w_2} J(w_1^f, w_2^f) = (4 - 2w_1 - w_2) \times (-1) + 0.5 \times (6 - w_1 - w_2) \times (-1) + 0.5 \times (-5 + w_1 - w_2) \times (-1)$$
$$\Rightarrow \nabla_{w_1} J(w_1^f, w_2^f) = -4.5 + 2w_1 + 2w_2$$

Setting the value of the system of linear equations to 0, we have: $w_1^f = \boxed{3}$, $w_2^f = \boxed{\text{-0.75}}$

2:

Linear TD(0) may not converge to the exact values of w_1^f and w_2^f obtained above, but the values will be within a certain distance of w_1^f and w_2^f.