ECE 239AS: Problem Set

Exercises 1-6 (minimum 1 problem)

2. In the first case, we have no way to select which arm to pull. Thus, we can maximize the expected reward by playing a random action at each step. Hence,

$$A_1 = 0.5(0.1 + 0.9) = 0.5$$

$$A_2 = 0.5(0.2 + 0.8) = 0.5$$

The value estimates of both actions are the same. Hence, we can choose an action randomly and still get the maximum expected reward.

In the second case, we are told that which case we are facing. Thus, we can pose this as a 2-arm bandit problem and learn to play the optimal arm to maximize the expectation of success. This means taking action 2 in case A and taking action 1 in case B. Hence,

$$\mathrm{E}(success) = 0.5(0.2) + 0.5(0.9) = 0.55$$

4. **a)** Each reward is independent from each other and identically distributed. This means

$$Var(x^T) = Var\left(\sum_{i=0}^{T-1} r_i\right) = \sum_{i=0}^{T-1} Var(r_i)$$

Since the rewards follow a Bernoulli distribution, we have $P(r_i = 1) = \frac{p_1 + \ldots + p_n}{n}$ and $P(r_i = 0) = 1 - \frac{p_1 + \ldots + p_n}{n}$ if we sample randomly. Hence, we get

$$Var(x^T) = \frac{T(p_1 + \ \ldots + p_n)}{n}\left(1 - \frac{p_1 + \ \ldots + p_n}{n}\right)$$

b) Since we want to maximize the expected reward, we would make a greedy algorithm L_* that picks the arm i which maximizes c where $c = \max(p_i, 1 - p_i)$.

5. **a)** Both arms are equally likely in the first pull. For the first 2 pulls, we have

$$\mathrm{E}(a_1 \ in \ first \ 2 \ pulls) = \frac{1}{2}\left(1 + \frac{p_1}{2} + \frac{p_2}{2} + 1 - p_2\right)$$
$$\mathrm{E}(a_2 \ in \ first \ 2 \ pulls) = \frac{1}{2}\left(1 + \frac{p_2}{2} + \frac{p_1}{2} + 1 - p_1\right)$$

Hence, we get

$$\mathrm{E}(0\ rewards) = \frac{1}{2}\left(1 + \frac{p_1}{2} + \frac{p_2}{2} + 1 - p_2\right)(1 - p_1) + \frac{1}{2}(1 + \frac{p_2}{2} + \frac{p_1}{2} + 1 - p_1)(1 - p_2)$$

b) We have

$$\mathrm{E}(a_1) = 1 - p_1 + 2p_1(1 - p_1) + 3p_1p_2(1 - p_1) + \cdots + \left(T(1 - p_1)\prod_{i=1}^{T-1} p_i\right) = \frac{1}{1-p_1}$$

Then we also have $\mathrm{E}(a_2) = \frac{1}{1-p_2}$. Hence, by the law of large numbers we get

$$\lim_{T\to\infty} \frac{\mathrm{E}(\mathrm{z}^{\mathrm{T}})}{T} = \frac{1}{\frac{1}{1-p_1} + \frac{1}{1-p_2}} = \frac{2(1-p_1)(1-p_2)}{1-p_1+1-p_2}$$

c) After T pulls, the total reward is $T - z^T$. Then, $R^T = Tp_1 - (T - z^T) = Tp_1 + z^T - T$. Hence,

$$\lim_{T\to\infty} \frac{\mathrm{E}(\mathrm{R}^{\mathrm{T}})}{T} = p_1 - 1 + \frac{2(1-p_1)(1-p_2)}{1-p_1+1-p_2} = \frac{(1-p_1)(1-p_2)}{1-p_1+1-p_2}$$

Exercises 7-10 (minimum 1 problem)

7. Each transition has reward -1, including the transition to the terminal state. $q_\pi(11, down)$ is a transition to the terminal state. Hence,

$$q_\pi(11, down) = -1$$

$q_\pi(7, down)$ is a transition to state 11. At state 11, we have 4 actions that are equiprobable. Using the table for policy evaluation, we know the state-value function at state 11. Hence,

$$q_\pi(7, down) = -1 + v_\pi(11) = -1 - 14 = -15$$

8. We have $v_\pi(15) = -1 + 0.25\left(v_\pi(12) + v_\pi(13) + v_\pi(14) + v_\pi(15)\right)$. Using the policy evaluation table, we get

$$v_\pi(15) = -1 + 0.25(-22 - 20 - 14 + v_\pi(15))$$

$$\therefore v_\pi(15) = -20$$

If the dynamics of state 13 changes, then $v_\pi(15)$ remains the same since every state satisfies the Bellman equation for v_π.

Exercises 11-15 (minimum 1 problem)

11. Using the Q-learning update rule, the values that get updated are S_t , A_t , $and\ R_{t+1}$. The final expression is

$$Q(34,7) \leftarrow Q(34,7) + \alpha \left[3 + \gamma \max_a Q(64,a) - Q(34,7)\right]$$

12. If the agent always chooses the action that maximizes the Q-value, then the agent can become stuck in a local minimum and choose non-optimal policies. To force exploration, the agent can (1) use an ϵ-greedy policy and (2) using large optimistic initial values so exploration is encouraged.

Exercises 16-20 (minimum 1 problem)

16. The transitions are presented below as a table.

Step	1	2	3	4	5	6	7	9	9	10
State	s_1	s_1	s_1	s_1	s_1	s_1	s_1	s_1	s_1	T
Reward	1	1	1	1	1	1	1	1	1	-

First visit MC = $\frac{\sum_{i=1}^{T} 1}{1} = 10$

Every visit MC = $\frac{\sum_{i=1}^{T} i}{10} = 5.5$

Exercises 21-30 (minimum 2 problem)

24. First, we need to find the TD(0) estimates to find the true value of the states. They are $V_\pi(s_1) = 4$, $V_\pi(s_2) = 0$ and $V_\pi(s_3) = 5$. Under policy π_{111}, we have $P(s_1) = 0.25$, $P(s_2) = 0.5$ and $P(s_3) = 0.25$. If we apply TD(1), then the converged θ_c is the θ for which the following expression is minimized.

$$\theta_c = \min_\theta 0.25(\theta - 4)^2 + 0.5\theta^2 + 0.25(2\theta - 5)^2$$

In linear approximation, we want to find the θ that minimizes the sum of the error squared. Thus, we take the derivative and set it equal to 0. Hence,

$$0.5(\theta - 4) + \theta + 2\theta - 5 = 0$$

$$=> 3.5\theta = 7$$

$$=> \theta = \theta_c = 2$$

25. a) In linear generalization scheme, we want to find the w^{opt} that minimizes the squared error. With Monte Carlo policy evaluation, we have

$$w^{opt} = \min_{w} 0.5(2w_1 + w_2 - 4)^2 + 0.25(w_1 + w_2 - 6)^2 + 0.25(-w_1 + w_2 + 5)^2$$

Where $w = (w_1, w_2) \in \mathbb{R}^2$

By taking $\frac{dw}{dw_1}$ and $\frac{dw}{dw_2}$ and setting them equal to 0, we can find the optimal weights. Thus we have,

$$\frac{dw}{dw_1} = 2(2w_1 + w_2 - 4) + 0.5(w_1 + w_2 - 6) - 0.5(-w_1 + w_2 + 5) = 0$$

$$=> 5w_1 + 2w_2 = 13.5$$

$$\frac{dw}{dw_2} = (2w_1 + w_2 - 4) + 0.5(w_1 + w_2 - 6) + 0.5(-w_1 + w_2 + 5) = 0$$

$$=> 2w_1 + 2w_2 = 4.5$$

Now we have $5w_1 + 2w_2 = 13.5$ and $2w_1 + 2w_2 = 4.5$. Solving this system of linear equations, we get $w_1^{opt} = 3$ and $w_2^{opt} = -0.75$.

b) If the agent uses linear TD(0), then the values of the weights will still converge to some $w_1^{TD(0)}$ and $w_2^{TD(0)}$. However, in general it can be the case that $w_1^{TD(0)} \neq w_1^{opt}$ and $w_2^{TD(0)} \neq w_2^{opt}$. It would be reasonable to assume that the difference between the optimal weights $\begin{pmatrix} w_1^{opt} \\ w_2^{opt} \end{pmatrix}$ and $\begin{pmatrix} w_1^{TD(0)} \\ w_2^{TD(0)} \end{pmatrix}$ is finite and bounded.