ELEN E6885: Introduction to Reinforcement Learning
Homework #1

Chenye Yang `cy2540@columbia.edu`

October 2, 2019

P1

1.

Ans:

Table 1: The given times in question

time	action	reward
$t = 1$	$A_1 = a_1$	$R_1(a_1) = 0.3$
$t = 2$	$A_2 = a_2$	$R_2(a_2) = 0$
$t = 3$	$A_3 = a_2$	$R_3(a_2) = 1$
$t = 4$	$A_4 = a_2$	$R_4(a_2) = 0$
$t = 5$	$A_5 = a_2$	$R_5(a_2) = 0$

When $t = 6$, the estimated action value $Q_6(a_1) = 0.3$ and $Q_6(a_2) = 1/4 = 0.25$.
With the greedy method being used to select actions,
Because: $Q_6(a_1) = 0.3 > 0.25 = Q_6(a_2)$
Therefore: we choose $A_6 = a_1$, arm 1 will be played, and we get reward $R_6(a_1) = 0.3$.

When $t = 7$, the estimated action value $Q_7(a_1) = (0.3 + 0.3)/2 = 0.3$ and $Q_7(a_2) = 1/4 = 0.25$.
With the greedy method being used to select actions,
Because: $Q_7(a_1) = 0.3 > 0.25 = Q_7(a_2)$
Therefore: we choose $A_7 = a_1$, arm 1 will be played, and we get reward $R_7(a_1) = 0.3$.

2.

Ans:
From **1.**, we know that $Q_6(a_1) = 0.3$ and $Q_6(a_2) = 1/4 = 0.25$
With the ϵ-greedy method being used to select actions ($\epsilon = 0.1$),
When $t = 6$:

$$\begin{aligned} P(A_6 = a_1) &= (1 - \epsilon) + \epsilon \times 0.5 \\ &= 0.9 + 0.05 = 0.95 \\ P(A_6 = a_2) &= \epsilon \times 0.5 = 0.05 \end{aligned} \tag{1}$$

When $t = 7$:
First calculate the estimated action value:

$$\begin{aligned}
Q_7(a_1) &= P(A_6 = a_1) \times \frac{0.3 + 0.3}{2} + P(A_6 = a_2) \times 0.3 \\
&= 0.95 \times 0.3 + 0.05 \times 0.3 = 0.3 \\
Q_7(a_2) &= P(A_6 = a_1) \times \frac{1}{4} + P(A_6 = a_2) \times \frac{1 + 0.6 \times 1 + 0.4 \times 0}{5} \\
&= 0.95 \times 0.25 + 0.05 \times 1.6/5 = 0.2535
\end{aligned} \tag{2}$$

Therefore, $Q_7(a_1) = 0.3 > 0.2535 = Q_7(a_2)$
As a result:

$$\begin{aligned}
P(A_7 = a_1) &= (1 - \epsilon) + \epsilon \times 0.5 = 0.9 + 0.05 = 0.95 \\
P(A_7 = a_2) &= \epsilon \times 0.5 = 0.05
\end{aligned} \tag{3}$$

The probability to play arm 2 at t = 6, 7 respectively is:

$$\begin{aligned}
P(A_6 = a_2) = 0.05 \\
P(A_7 = a_2) = 0.05
\end{aligned}$$

3.

Ans:

Greedy method will only focus on the current optimal actions, however, there may exist other better actions which hasn't been explored. The unexplored or not-well-explored actions may have a greater mean value than the current optimal action. The ϵ-greedy action has the potential to explore every actions thoroughly and find the true, or to say, global optimal action and the focus on the true optimal action. Therefore, in a long run, the greedy method could converge on a sub-optimal action while the ϵ-greedy method could converge on a true optimal action. That is to say, ϵ-greedy method may choose more global optimal actions and get a better average reward, so greedy method performs significantly worse.

As for this case, from **1.**, we can infer that the greedy method will always choose arm 1, because the estimated sample average value of arm 1 is greater than that of arm 2, from the reward of first 5 actions. Even though the actual expect value of arm 2 (0.6) is greater than that of arm 1 (0.3). However, from **2.**, with the ϵ-greedy method, we have the possible to explore arm 2 and the estimated action value $Q_n(a_2)$ of arm 2 will increase with time. Finally it will surpass the $Q_n(a_1)$ and we can choose the true optimal action (a_2) with probability 0.95. The ϵ-greedy method will get average reward $\mathbb{E}R_n = 0.6 \times 0.95 + 0.3 \times 0.05 = 0.585$ greater than greedy method $\mathbb{E}R_n = 0.3$, and perform better in a long run.

P2

1.

Ans:
Softmax action selection means choosing action a at t-th play with possibility

$$P(A_t = a) = \frac{e^{Q_t(a)/\tau}}{\sum_{i=1}^{n} e^{Q_t(i)/\tau}}$$

When $\tau \to 0$, $Q_t(i)/\tau \to +\infty$. Considering the figure of $f(x) = e^x$, which will significantly increase when the independent variable is great and increases slightly, thus $e^{Q_t(a)/\tau}$ will increasing to $+\infty$ sharply.

In this situation, a bigger $Q_t(i)/$ will lead to a much bigger $Q_t(i)/\tau$, or to say, $Q_t(i)/\tau$ will move to $+\infty$ faster. Therefore, the action with the biggest $Q_t(a)$ is most possible to be chosen. And when $\tau \to 0$, that action will almost always be chosen, which is the same as greedy action selection.

2.

Ans:

$$\begin{aligned}\lim_{\tau\to+\infty} P(A_t = a) &= \lim_{\tau\to+\infty} \frac{e^{Q_t(a)/\tau}}{\sum_{i=1}^{n} e^{Q_t(i)/\tau}} \\ &= \frac{e^0}{\sum_{i=1}^{n} e^0} = \frac{1}{n}\end{aligned} \tag{4}$$

Therefore, when $\tau \to +\infty$, the probability of selecting every action is equal. Softmax action selection yields equiprobable selection among all actions.

3.

Ans:
Sigmoid function reflects an independent real variable to interval $(0, 1)$, commonly having form as $S(x) = 1/(1 + e^{-x}), x \in \mathbb{R}, y \in (0, 1)$.
Let the two actions be a_1 and a_2.

$$\begin{aligned}P(A = a_1) &= \frac{e^{Q_t(a_1)/\tau}}{e^{Q_t(a_1)/\tau} + e^{Q_t(a_2)/\tau}} \\ &= \frac{1}{1 + e^{Q_t(a_2)/\tau - Q_t(a_1)/\tau}} \\ &= \frac{1}{1 + e^{[Q_t(a_2) - Q_t(a_1)]/\tau}}\end{aligned} \tag{5}$$

Because $e^{[Q_t(a_2)-Q_t(a_1)]/\tau} \in (0, +\infty)$, thus $P(A = a_1) \in (0, 1)$. Also:

$$\begin{aligned}\tau \to +\infty, \quad P(A = a_1) \to 1/2 \qquad \tau \to -\infty, \quad P(A = a_1) \to 1/2 \\ \tau \to 0^+, \quad P(A = a_1) \to 0 \qquad \tau \to 0^-, \quad P(A = a_1) \to 1\end{aligned} \tag{6}$$

Following is a figure of P to τ when $Q_t(a_2) - Q_t(a_1) = 1$:

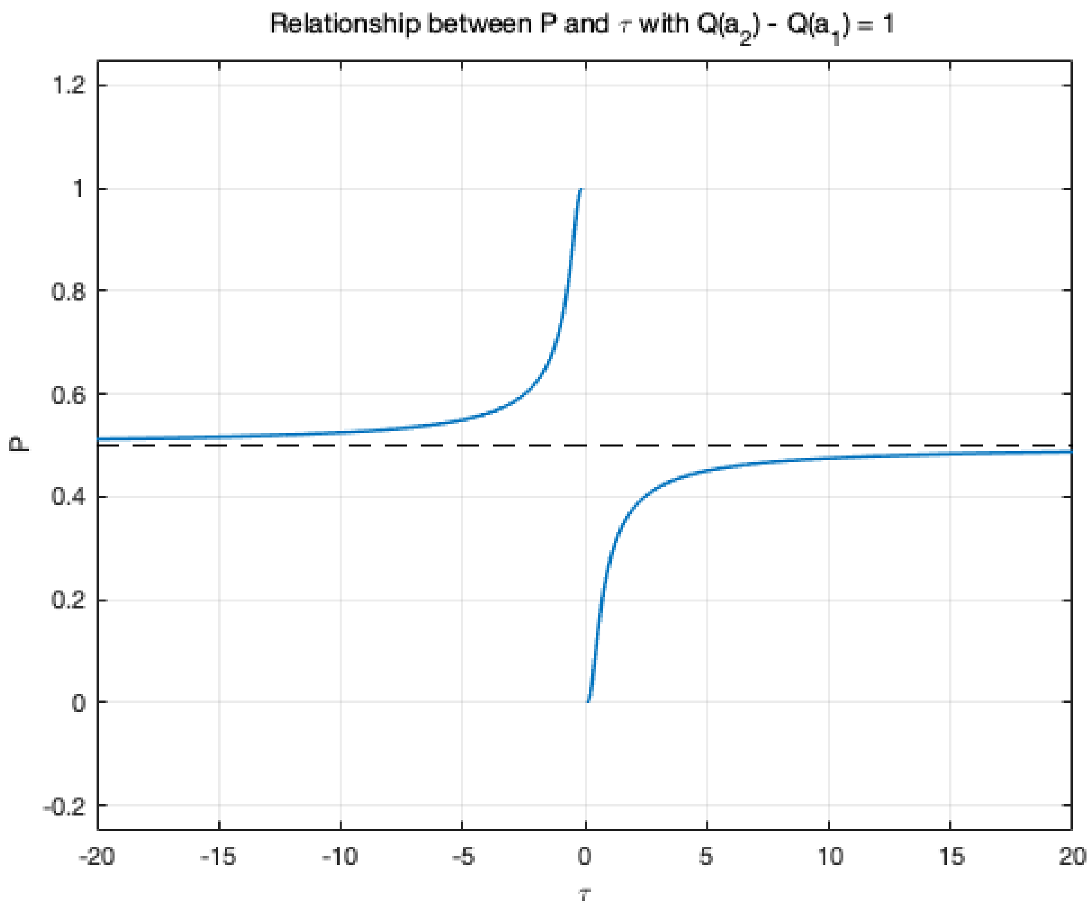

Figure 1: Relationship between P and τ

Things are exactly same with $P(A = a_2)$.
Therefore, in the case of two actions, the softmax operation using the Gibbs distribution becomes the sigmoid function.

P3

Ans:
From the question, we know the cumulative sum of the weights $C_n = W_1+W_2+...+W_n$. Obviously $C_{n+1} = C_n + W_{n+1}$ is true for $n \geq 1$.
For $n \geq 2$:

$$\begin{aligned} V_n &= \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k} \\ &= \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n-1} W_k} \frac{\sum_{k=1}^{n} W_k}{\sum_{k=1}^{n} W_k} \\ &= \frac{\sum_{k=1}^{n-1} W_k G_k}{\sum_{k=1}^{n} W_k} \frac{C_n}{C_{n-1}} \\ &= \frac{\sum_{k=1}^{n} W_k G_k - W_n G_n}{\sum_{k=1}^{n} W_k} \frac{C_n}{C_{n-1}} \\ &= (V_{n+1} - \frac{W_n G_n}{C_n}) \frac{C_n}{C_{n-1}} \\ &= V_{n+1} \frac{C_n}{C_{n-1}} - \frac{W_n G_n}{C_{n-1}} \end{aligned} \tag{7}$$

$\Rightarrow$

$$\begin{aligned} V_{n+1} &= \frac{C_{n-1}}{C_n} V_n + \frac{W_n G_n}{C_n} \\ &= V_n - \frac{W_n}{C_n} V_n + \frac{W_n}{C_n} G_n \\ &= V_n + \frac{W_n}{C_n} (G_n - V_n) \end{aligned} \tag{8}$$

Also as definition:

$$V_2 = \frac{W_1 G_1}{W_1} = G_1 \tag{9}$$

For $n = 1$, from equation 8:

$$V_1 + \frac{W_1}{C_1}(G_1 - V_1) = G_1 = V_2 \tag{10}$$

Therefore, equation 8 is true for $n \geq 1$.
Therefore, the update rule for $V_{n+1}, n \geq 1$ is as stated in problem.

P4

1.

Ans:

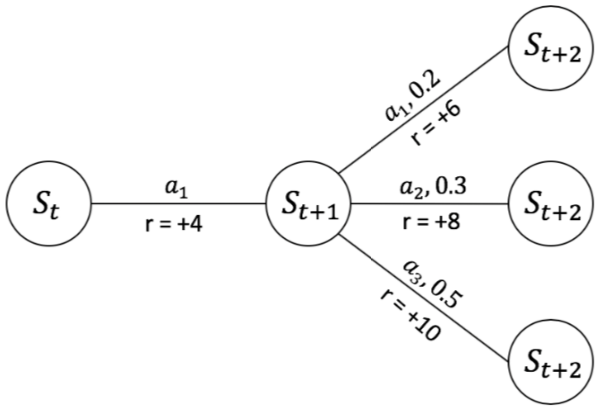

Figure 2: MDP with deterministic transitions

$$
\begin{aligned}
v_\pi(S_{t+1}) &= \mathbb{E}_\pi[R_{t+2} + \gamma v_\pi(S_{t+2}) | S = S_{t+1}] \\
&= 0.2 \times 6 + 0.3 \times 8 + 0.5 \times 10 \\
&= 8.6 \\
v_\pi(S_t) &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(S_{t+1}) | S = S_t] \\
&= 4 + 8.6 = 12.6
\end{aligned} \tag{11}
$$

2.

Ans:
For the $s = S_{t+1}$ with relation to $s = S_{t+2}$ in the top-right of Figure 3:

$$
\begin{aligned}
v_\pi(S_{t+1}) &= \mathbb{E}_\pi[R_{t+2} + \gamma v_\pi(S_{t+2}) | S = S_{t+1}] \\
&= 0.2 \times 6 + 0.3 \times 8 + 0.5 \times 10 \\
&= 8.6
\end{aligned} \tag{12}
$$

For the $s = S_t$:

$$
\begin{aligned}
v_\pi(S_t) &= \sum_{a \in A} \pi(a|s)(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')) \\
&= 0.5 \times 4 + 0.5 \times [4 + 1 \times (0.4 \times 8.6 + 0.6 \times 0)] \\
&= 5.72
\end{aligned} \tag{13}
$$

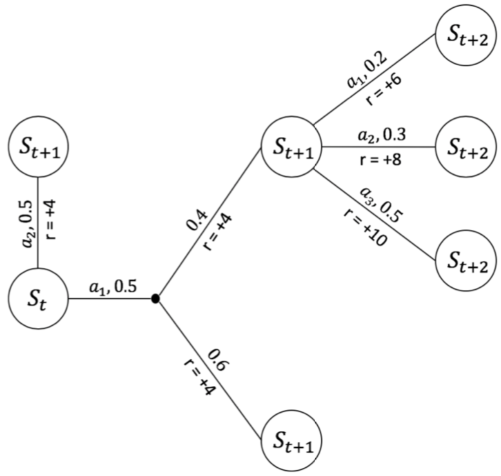

Figure 3: MDP with stochastic transitions

3.

Ans:

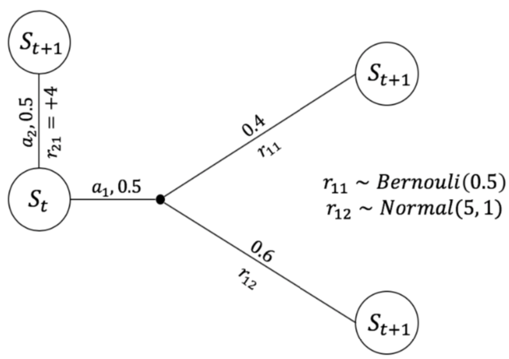

Figure 4: MDP with stochastic rewards

The reward of action a_1 at state S_t is:

$$\begin{aligned} R_{S_t}^{a_1} &= \mathbb{E}[R_{t+1}|S_t = s] \\ &= 0.4 \times \mathbb{E}(r_{11}) + 0.6 \times \mathbb{E}(r_{12}) \\ &= 0.4 \times 0.5 + 0.6 \times 5 = 3.2 \end{aligned} \tag{14}$$

The state value for S_t:

$$\begin{aligned} v_\pi(S_t) &= \sum_{a \in A} \pi(a|s)(R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_\pi(s')) \\ &= 0.5 \times (4 + 0) + \underline{0.5 \times [3.2 + 1 \times (0.4 \times 0 + 0.6 \times 0)}] \\ &= 3.6 \end{aligned} \tag{15}$$

This is a misunderstanding, should be:
0.5x[1x0.4x(0.5+0)+1x0.6x(5+0)]

P5

1.

Ans:
The action-value function is the expected return starting from state s, taking action a, and following policy π:

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

The optimal action-value function is the maximum action-value function over all policies.

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

An optimal policy can be found by maximising over $q_*(s,a)$:

$$\pi_*(a|s) = \begin{cases} 1 & \text{if } a = \arg\max_{a\in A} q_*(s,a) \\ 0 & \text{else} \end{cases}$$

Let $q_\pi^{'}(s,a)$ and $q_*^{'}(s,a)$ be the new action-value function and new optimal action-value function, we have:

$$\begin{aligned} q_\pi^{'}(s,a) &= \mathbb{E}_\pi[\sum_{k=0}^{\infty}\gamma^k(R_{t+k+1}+\alpha)|S_t = s, A_t = a] \\ &= \mathbb{E}_\pi[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1} + \sum_{k=0}^{\infty}\gamma^k\alpha|S_t = s, A_t = a] \\ &= q_\pi(s,a) + \sum_{k=0}^{\infty}\gamma^k\alpha \\ &= q_\pi(s,a) + \frac{\alpha}{1-\gamma} \\ q_*^{'}(s,a) &= \max_\pi q_\pi^{'}(s,a) \\ &= \max_\pi[q_\pi(s,a) + \frac{\alpha}{1-\gamma}] \\ &= \max_\pi q_\pi(s,a) + \frac{\alpha}{1-\gamma} \end{aligned}$$

The new optimal policy $\pi_*^{'}(a|s)$ is:

$$\begin{aligned} \pi_*^{'}(a|s) &= \begin{cases} 1 & \text{if } a = \arg\max_{a\in A} q_*^{'}(s,a) \\ 0 & \text{else} \end{cases} \\ &= \begin{cases} 1 & \text{if } a = \arg\max_{a\in A} q_*(s,a) \\ 0 & \text{else} \end{cases} = \pi_*(a|s) \end{aligned}$$

Therefore, the modified MDP in **1.** has the same optimal policy as the original MDP.

2.

Ans:
The definitions are the same as **1.**
Let $q_{\pi}^{'}(s,a)$ and $q_{*}^{'}(s,a)$ be the new action-value function and new optimal action-value function, we have:

$$\begin{aligned}
q_{\pi}^{'}(s,a) &= \mathbb{E}_{\pi}[\sum_{k=0}^{\infty}\gamma^k(\beta \times R_{t+k+1})|S_t = s, A_t = a] \\
&= \beta\mathbb{E}_{\pi}[\sum_{k=0}^{\infty}\gamma^k R_{t+k+1}|S_t = s, A_t = a] \\
&= \beta q_{\pi}(s,a) \\
q_{*}^{'}(s,a) &= \max_{\pi} q_{\pi}^{'}(s,a) \\
&= \beta \max_{\pi} q_{\pi}(s,a)
\end{aligned}$$

Because $\beta > 0$, the new optimal policy $\pi_{*}^{'}(a|s)$ is:

$$\begin{aligned}
\pi_{*}^{'}(a|s) &= \begin{cases} 1 & \text{if } a = \arg\max_{a \in A} q_{*}^{'}(s,a) \\ 0 & \text{else} \end{cases} \\
&= \begin{cases} 1 & \text{if } a = \arg\max_{a \in A} q_{*}(s,a) \\ 0 & \text{else} \end{cases} = \pi_{*}(a|s)
\end{aligned}$$

Therefore, the modified MDP in **2.** has the same optimal policy as the original MDP.

P6

1.

Ans:
From the statement in question, let the set of action be $A = \{a_1, a_2\}$, a_1 means 'draw' while a_2 means 'stop', and set of state be $S = \{S_0, S_2, S_3, S_4, S_5, S_D\}$, $S_0, ...S_5$ means score is 0,...,5, S_D means end of game. Thus we can draw the state transition figure as follow:

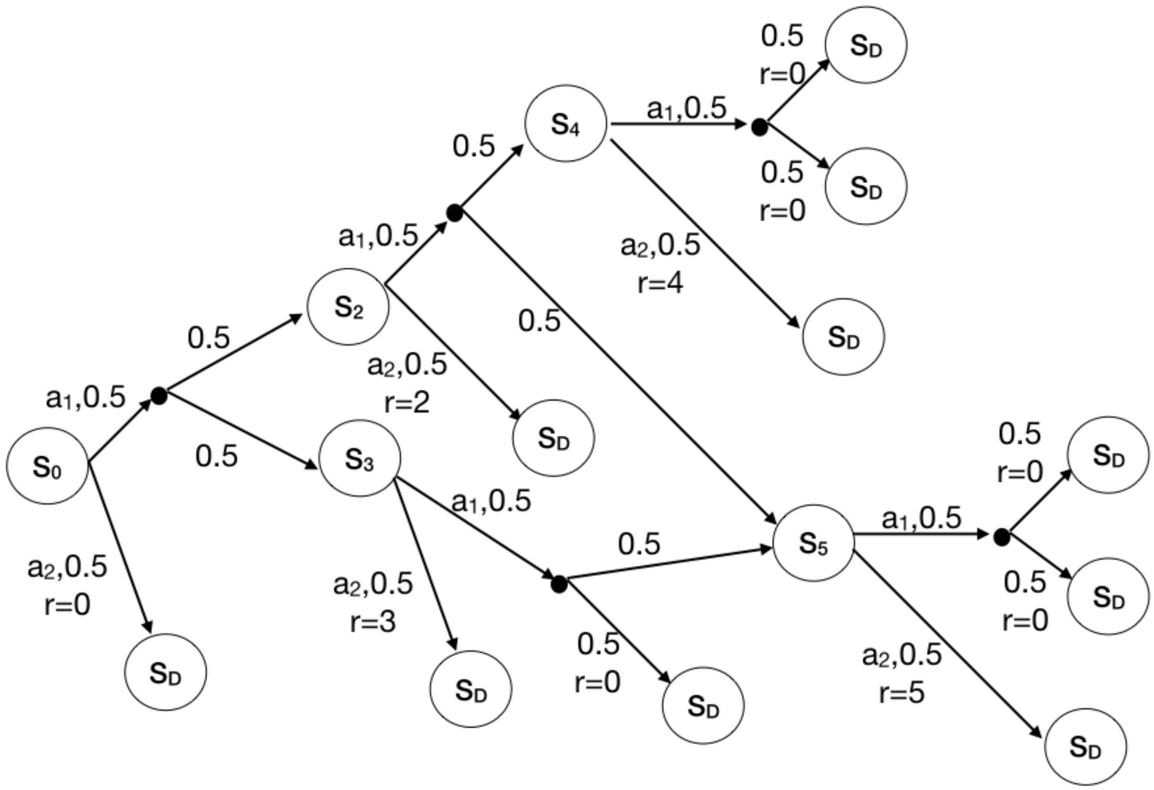

Figure 5: State Transition Figure

The state transition function:

$$
\begin{array}{lll}
P_{S_0S_2} = 0.25 & P_{S_0S_3} = 0.25 & P_{S_0S_D} = 0.5 \\
P_{S_2S_4} = 0.25 & P_{S_2S_5} = 0.25 & P_{S_2S_D} = 0.5 \\
P_{S_3S_5} = 0.25 & P_{S_3S_D} = 0.75 & \\
P_{S_4S_D} = 1 & & \\
P_{S_5S_D} = 1 & &
\end{array}
$$

Consider Action!

$$P = \begin{bmatrix} 0 & 0.25 & 0.25 & 0 & 0 & 0.5 \\ 0 & 0 & 0 & 0.25 & 0.25 & 0.5 \\ 0 & 0 & 0 & 0 & 0.25 & 0.75 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

The reward function $R_s^a = \mathbb{E}[R_{t+1}|S_t = s, A_t = a]$:

$$\begin{aligned} R_{S_0}^{a_1} &= 0 & R_{S_0}^{a_2} &= 0 \\ R_{S_2}^{a_1} &= 0 & R_{S_2}^{a_2} &= 2 \\ R_{S_3}^{a_1} &= 0 & R_{S_3}^{a_2} &= 3 \\ R_{S_4}^{a_1} &= 0 & R_{S_4}^{a_2} &= 4 \\ R_{S_5}^{a_1} &= 0 & R_{S_5}^{a_2} &= 5 \end{aligned}$$

2.

Ans:
For S_4:

$$\begin{aligned} q_*(S_4, a_1) &= 0 + 1 \times 1 \times 0 = 0 \\ q_*(S_4, a_2) &= 4 + 1 \times 1 \times 0 = 4 \\ v_*(S_4) &= \max_{a \in A} q_*(S_4, a) = 4 \end{aligned}$$

For S_5:

$$\begin{aligned} q_*(S_5, a_1) &= 0 + 1 \times 1 \times 0 = 0 \\ q_*(S_5, a_2) &= 5 + 1 \times 1 \times 0 = 5 \\ v_*(S_5) &= \max_{a \in A} q_*(S_5, a) = 5 \end{aligned}$$

For S_3:

$$\begin{aligned} q_*(S_3, a_1) &= 0 + 1 \times \underline{0.25} \times v_*(S_5) = 1.25 \\ q_*(S_3, a_2) &= 3 + 0 = 3 \\ v_*(S_3) &= \max_{a \in A} q_*(S_3, a) = 3 \end{aligned}$$

For S_2:

$$\begin{aligned} q_*(S_2, a_1) &= 0 + 1 \times [\underline{0.25} \times v_*(S_4) + \underline{0.25} \times v_*(S_5)] = 2.25 \\ q_*(S_2, a_2) &= 2 + 0 = 2 \\ v_*(S_2) &= \max_{a \in A} q_*(S_2, a) = 2.25 \end{aligned}$$

Already take the action a1, the probability to S4 or S5 must be 0.5, instead of 0.25

For S_0:

$$\begin{aligned} q_*(S_0, a_1) &= 0 + 1 \times [\underline{0.25} \times v_*(S_2) + \underline{0.25} \times v_*(S_3)] = 1.3125 \\ q_*(S_0, a_2) &= 0 + 0 = 0 \\ v_*(S_0) &= \max_{a \in A} q_*(S_0, a) = 1.3125 \end{aligned}$$

3.

Ans:
The optimal policy for this MDP:

$$\pi_*(a|S_0) = \begin{cases} 1 & a = a_1 \\ 0 & a = a_2 \end{cases}$$

$$\pi_*(a|S_2) = \begin{cases} 1 & a = a_1 \\ 0 & a = a_2 \end{cases}$$

$$\pi_*(a|S_3) = \begin{cases} 1 & a = a_2 \\ 0 & a = a_1 \end{cases}$$

$$\pi_*(a|S_4) = \begin{cases} 1 & a = a_2 \\ 0 & a = a_1 \end{cases}$$

$$\pi_*(a|S_5) = \begin{cases} 1 & a = a_2 \\ 0 & a = a_1 \end{cases}$$