AAI Assignment 4

Name: 吉辰卿 SID: 12332452

1. Solution

proof: To prove this problem, we should reveal that the optimal policy remains unchanged for this modified MDP.

Firstly, we denote the state-value function for a policy π in the original MDP is $V_\pi(s)$ and in the modified MDP is $V'_\pi(s)$.

so: $V_\pi(s) = E_\pi[G_t \mid \pi, S_t = s] = E_\pi[R_{t+1} + \gamma R_{t+2} + \cdots \mid \pi, S_t = s]$ --①

Then, because for the modified MDP, the new reward function will become: $R'(s) = \alpha R(s) + \beta$ $(\alpha > 0)$

Therefore, the state-value function in the modified MDP under the policy π will become $V'_\pi(s)$:

$$V'_\pi(s) = E_\pi[G'_\pi \mid \pi, S_t = s] = E_\pi[\cancel{\alpha R_{t+1} + \beta}\ R'_{t+1} + \gamma R'_{t+2} + \cdots \mid \pi, S_t = s]$$

$$= E_\pi[\alpha R_{t+1} + \beta + \gamma \alpha R_{t+2} + \beta + \cdots \mid \pi, S_t = s] \text{ --②}$$

let's observe the structure of ① and ② and we can find β is only considered as a constant term and the α is only considered as a coefficient of the expected reward. so, we can get:

$V_\pi(s) = E_\pi[G_\pi \mid \pi, S_t = s]$ and $V'_\pi(s) = E_\pi[G'_\pi \mid \pi, S_t = s] = E_\pi[\alpha G_\pi + \beta \mid \pi, S_t = s]$

So: G'_π in the modified MDP is only a linear combination of the G_π in the original MDP. That is: $G'_\pi = \alpha G_\pi + \beta$

Since the relationship between G'_π and G_π is linear, ~~which~~ and $\alpha > 0$, which is linear positive correlation. Therefore, the modified reward function $R'(s)$ will not affect the formulation of the optimal policy.

So, the modified MDP has the same optimal policy as the original MDP.

2. solution

(1) when MDP stops after two steps:

$$G_t = R_{t+1} + \gamma R_{t+2}$$

so: the state-value function is: $v_\pi(s) = E[G_t \mid S_t = s, \pi]$

$$= E[R_{t+1} + \gamma v_\pi(s') \mid S_t = s, \pi] \quad \text{--- (1)}$$

In this problem, we denote H as the state high, L as the state low, W as the action wait, S as the action search, R as the action recharge.

And for stochastic $\pi(a|s)$, (1) will become:

$$v_\pi(s) = \sum_{a \in A} \pi(a|s) \sum_{s' \in S} p(s, a, s') \left(R(s, a, s') + \gamma v_\pi(s') \right).$$

And: $\pi(S|H)=1$, $\pi(S|L)=0.4$ and $\pi(R|L)=0.6$, and others are 0.

So: $v_\pi(H) = \pi(S|H)\times[P(H,S,H)(R(H,S,H)+\gamma v'_\pi(H)) + P(H,S,L)(R(H,S,L)+\gamma v'_\pi(L))]$

$+ \pi(W|H)\times[P(H,W,H)(R(H,W,H)+\gamma v'_\pi(H)) + P(H,W,L)(R(H,W,L)+\gamma v'_\pi(L))]$

$+ \pi(R|H)\times[P(H,R,H)(R(H,R,H)+\gamma v'_\pi(H)) + P(H,R,L)(R(H,R,L)+\gamma v'_\pi(L))]$.

$v_\pi(L) = \pi(S|L)\times[P(L,S,H)(R(L,S,H)+\gamma v'_\pi(H)) + P(L,S,L)(R(L,S,L)+\gamma v'_\pi(L))]$

$+ \pi(R|L)\times[P(L,R,H)(R(L,R,H)+\gamma v'_\pi(H)) + P(L,R,L)(R(L,R,L)+\gamma v'_\pi(L))]$

$+ \pi(W|L)\times[P(L,W,H)(R(L,W,H)+\gamma v'_\pi(H)) + P(L,W,L)(R(L,W,L)+\gamma v'_\pi(L))]$

$v'_\pi(H) = \pi(S|H)\times[P(H,S,H)\times R(H,S,H) + P(H,S,L)\times R(H,S,L)]$

$+ \pi(W|H)\times[P(H,W,H)\times R(H,W,H) + P(H,W,L)\times R(H,W,L)]$

$+ \pi(R|H)\times[P(H,R,H)\times R(H,R,H) + P(H,R,L)\times R(H,R,L)]$

$v'_\pi(L) = \pi(S|L)\times[P(L,S,H)\times R(L,S,H) + P(L,S,L)\times R(L,S,L)]$

$+ \pi(R|L)\times[P(L,R,H)\times R(L,R,H) + P(L,R,L)\times R(L,R,L)]$

$+ \pi(W|L)\times[P(L,W,H)\times R(L,W,H) + P(L,W,L)\times R(L,W,L)]$

Therefore:

$$v'_\pi(H) = 1\times[\alpha\times r_{search} + (1-\alpha)\times r_{search}]$$

$$= 0.5\times 3 + 0.5\times 3 = 3$$

$$v'_\pi(L) = 0.4\times[(1-\beta)\times(-3) + \beta\times r_{search}] + 0.6\times[1\times 0]$$

$$= 0.4\times[0.7\times(-3) + 0.3\times 3] = -0.48$$

$$v_\pi(H) = 1\times[\alpha(r_{search} + \gamma v'_\pi(H)) + (1-\alpha)(r_{search} + \gamma v'_\pi(L))]$$

$$= 0.5\times(3+0.8\times 3) + 0.5\times[3+0.8\times(-0.48)] = 4.008$$

$$v_\pi(L) = 0.4\times[(1-\beta)(-3+\gamma v'_\pi(H)) + \beta(r_{search} + \gamma v'_\pi(L))] + 0.6\times[1\times(0+\gamma v'_\pi(H))]$$

$$= 0.4\times[0.7\times(-3+0.8\times 3) + 0.3\times(3-0.8\times 0.48)] + 0.6\times 0.8\times 3$$

$$= 1.58592$$

So, the state-value function is: $V_\pi(H) = 4.008$, $V_\pi(L) = 1.58592$.

(2) The expression of the action-value function in the case the MDP stops after a single step is:

$$q_\pi(s,a) = E[G_t \mid \pi, s, a] = E[R_{t+1} \mid \pi, s, a]$$

So, we use the same notation as (1).

$$q_\pi(H,W) = P(H,W,H) \times R(H,W,H) + P(H,W,L) \times R(H,W,L)$$

$$q_\pi(H,S) = P(H,S,H) \times R(H,S,H) + P(H,S,L) \times R(H,S,L)$$

$$q_\pi(H,R) = P(H,R,H) \times R(H,R,H) + P(H,R,L) \times R(H,R,L)$$

$$q_\pi(L,W) = P(L,W,H) \times R(L,W,H) + P(L,W,L) \times R(L,W,L)$$

$$q_\pi(L,S) = P(L,S,H) \times R(L,S,H) + P(L,S,L) \times R(L,S,L)$$

$$q_\pi(L,R) = P(L,R,H) \times R(L,R,H) + P(L,R,L) \times R(L,R,L)$$

Therefore:

$$q_\pi(H,W) = 1 \times r_{wait} = 1 \times 0 = 0$$

$$q_\pi(H,S) = \alpha \times r_{search} + (1-\alpha) \times r_{search} = 0.5 \times 3 + 0.5 \times 3 = 3$$

$$q_\pi(L,W) = 1 \times r_{wait} = 1 \times 0 = 0$$

$$q_\pi(L,R) = 1 \times 0 = 0$$

$$q_\pi(L,S) = (1-\beta) \times (-3) + \beta \times r_{search} = -3 \times 0.7 + 0.3 \times 3 = -1.2$$

Therefore, the action-value function for each value pair is: $q_\pi(H,W) = 0$, $q_\pi(H,S) = 3$, $q_\pi(L,W) = 0$, $q_\pi(L,S) = -1.2$ and $q_\pi(L,R) = 0$.

3 solution

(1) For a policy that always takes the action "produce" in the "Good" state, the conditional probability distribution over actions given states can be expressed by: (assume "G" is good state, "B" is broken state, "P" is produce action, "I" is inactive action and "R" is repair action).

$$\pi(P|G) = 1 \quad \pi(I|G) = 0 \quad \pi(R|B) = 1 \quad \pi(R|G) = 0$$

so: $v_\pi(S) = E[G_t | S_t = s, \pi] = E[R_{t+1} + \gamma V_\pi(s') | \pi, S_t = s]$

Therefore: $v_\pi(G) = \pi(P|G) \times [P(G,P,G) \times (R(G,P,G) + \gamma V_\pi'(G))$
$+ P(G,P,B) \times (R(G,P,B) + \gamma V_\pi'(B))]$
$+ \pi(I|G) \times [P(G,I,G) \times (R(G,I,G) + \gamma V_\pi'(G)) + P(G,I,B)$
$\times (R(G,I,B) + \gamma V_\pi'(B))]$

$= 1 \times [0.8 \times (2 + \gamma V_\pi'(G)) + 0.2 \times \gamma V_\pi'(B)]$

$= 1.6 + 0.8\gamma V_\pi'(G) + 0.2\gamma V_\pi'(B)$

So: $V_\pi(G) = 1.6 + 0.8\gamma V_\pi'(G) + 0.2\gamma V_\pi'(B)$ ($V_\pi'(G)$ is similar to $V_\pi(G)$ and it's a recursive process).

$V_\pi(B) = \pi(R|B) \times [P(B,R,G) \times (R(B,R,G) + \gamma V_\pi'(G)) + P(B,R,B) \times$
$(R(B,R,B) + \gamma V_\pi'(B))]$

$= 1 \times [0.8 \times \gamma V_\pi'(G) + 0.2 \times (-5 + \gamma V_\pi'(B))]$

$= -1 + 0.8\gamma V_\pi'(G) + 0.2\gamma V_\pi'(B)$

So: $V_\pi(B) = -1 + 0.8\gamma V_\pi'(G) + 0.2\gamma V_\pi'(B)$ ($V_\pi'(B)$ is similar to $V_\pi(B)$ and it's a recursive process)

(2) For the optimal value function in "Good" state and "Broken" state denoted by $V^*(G)$ and $V^*(B)$.

$$V^*(G) = \max_{a \in A} \sum_{s' \in S} P(G,a,s')(R(G,a,s') + \gamma V^*(s'))$$

$$V^*(B) = \max_{a \in A} \sum_{s' \in S} P(B,a,s')(R(B,a,s') + \gamma V^*(s')).$$

Therefore, $V^*(G) = \max_{a \in A} \sum_{s' \in S} P(G,a,s')(R(G,a,s') + \gamma V^*(s'))$

$$= \max\{[P(G,P,G)\cdot(R(G,P,G) + \gamma V'_{\pi}(G)) + P(G,P,B)(R(G,P,B) + \gamma V'_{\pi}(B))],$$

$$P(G,I,G)(R(G,I,G) + \gamma V'_{\pi}(G))\}$$

$$= \max\{[0.8\times(2+\gamma V'_{\pi}(G)) + 0.2\cdot\gamma V'_{\pi}(B)],\ \gamma V'_{\pi}(G)\}$$

~~So:~~

$$= \max\{[1.6 + 0.8\gamma V'_{\pi}(G) + 0.2\gamma V'_{\pi}(B)],\ \gamma V'_{\pi}(G)\}$$

$$= \max\{[1.6 + 0.8\gamma V'^{*}_{\pi}(G) + 0.2\gamma V'^{*}_{\pi}(B)],\ \gamma V'^{*}_{\pi}(G)\}$$

$$= \max\{(1.6 + 0.8\gamma V'^{*}(G) + 0.2\gamma V'^{*}(B)),\ \gamma V'^{*}(G)\}$$

So: $V^{*}(G) = \max\{1.6 + 0.8\gamma V'^{*}(G) + 0.2\gamma V'^{*}(B),\ \gamma V'^{*}(G)\}$

Then: $V^{*}(B) = \max\limits_{a\in A}\sum\limits_{s'\in S} P(B,a,s')(R(B,a,s') + \gamma V^{*}(s'))$

$$= P(B,R,G)(R(B,R,G) + \gamma V'_{\pi}(G)) + P(B,R,B)(R(B,R,B) + \gamma V'_{\pi}(B))$$

$$= 0.8\cdot\gamma V'_{\pi}(G) + 0.2\times(-5 + \gamma V'_{\pi}(B))$$

$$= -1 + 0.8\gamma V'^{*}(G) + 0.2\gamma V'^{*}(B)$$

So: $V^{*}(B) = -1 + 0.8\gamma V'^{*}(G) + 0.2\gamma V'^{*}(B)$

Therefore:

$$\begin{cases} V^{*}(G) = \max\{(1.6 + 0.8\gamma V'^{*}(G) + 0.2\gamma V'^{*}(B)),\ \gamma V'^{*}(G)\} \\ V^{*}(B) = -1 + 0.8\gamma V'^{*}(G) + 0.2\gamma V'^{*}(B) \end{cases}$$

(And $V'^{*}(G)$, $V'^{*}(B)$ is the optimal value function by the next state related to "G" and "B").