1. (1) $V_\pi(s = high) = 0.8 \times 4 + 0.2 \times 0 = 3.2$

$V_\pi(s = low) = 0.25 \times 4 + 0.2 \times 0 + 0.5 \times 0 = 1$

(2) $q(low, search) = r_{search} + \gamma \cdot \sum_{s'} P(s' | low, search) \cdot V_\pi(s')$

$= 4 + 0.9 \times (0.5 \times 3.2 + 0.5 \times 1)$

$= 4 + 0.9 \times (2.1)$

$= 5.89$

2. (1) Markov Decision Processes (MDP)

$M = \langle S, A, P, R \rangle$

S: set of all possible states

A: set of all possible actions

$P: S \times A \times S \rightarrow [0, 1]$, transition probability function

R: immediate reward function.

If both S and A are finite, then M is a finite MDP

MDP formally describe an environment for RL where the environment is fully observable.

(2) Define state value function $V_\pi : S \rightarrow \mathbb{R}$ of policy π as

$V_\pi(s) = E[G_t | \pi, S_t = s]$

Define action value function $q_\pi : S \times A \rightarrow \mathbb{R}$ of policy π as

$q_\pi(s, a) = E[G_t | \pi, S_t = s, A_t = a]$

(3) We can prove that the modified MDP has the same optimal action values as the original MDP by showing that the Bellman equation for the modified MDP is a scaled version of the

equation for the modified MDP is a scaled version of the Bellman equation for the original MDP

The Bellman equation for the orignal MDP is given by

$$q_\pi(s,a) = R(s) + \gamma \sum_{s'} P(s'|s,a) * V_\pi(s')$$

If we substitude $RO(s) = \beta R(s)$ into the Bellman equation, we get:

$$q_\pi(s,a) = RO(s) + \gamma \sum_{s'} P(s'|s,a) \times V_\pi(s')$$

$$= \beta R(s) + \gamma \sum_{s'} P(s'|s,a) \times V_\pi(s')$$

This is simply a scaled version of the original Bellman equation, where the reward term is multiplied by a constant factor β. Therefore, the optimal action values $q^*(s,a)$ will be the same in both MDPs.

Therefore, we have proved that the modified MDP has the same optimal policy as the original MDP.

3. (1) $V^\pi(g) = 0.9 \times (1 + \gamma \cdot V^\pi(g)) + 0.1 \times (0 + \gamma \cdot V^\pi(b))$

$V^\pi(b) = P_{b\to g} \times (0 + \gamma \cdot V^\pi(g)) + (1 - P_{b\to g}) \times (-10 + \gamma \cdot V^\pi(b))$

Known that $V^\pi(g) = 7.29$ and $V^\pi(b) = 5.39$

Then we can get $\gamma = 0.8$ $P_{b\to g} = 0.39$

(2) $V^*(g) = \max(0.9 \times (1 + \gamma * V^*(g)), 0.1 \times (0 + \gamma \times V^*(b)))$

$V^*(b) = \max(P_{b\to g} \times (0 + \gamma * V^*(g)), (1 - P_{b\to g}) * (-10 + \gamma * V^*(b)))$

Known the results in Q1, we can know that

$V^*(g) = \max(0.9 \times (1 + 0.8 \cdot V^*(g)), 0.1 \times (0 + 0.8 \cdot V^*(g)))$

$V^*(b) = \max(0.39 \times (0 + 0.8 \cdot V^*(g)), 0.61 \times (-10 + 0.8 \cdot V^*(b)))$