1. MDP

(1) State space (S) is the set of all possible states.

Action space (A) is the set of all possible actions.

Consider the four-room domain from Ex 0,

S is a set of all states in the environment from (0, 0) to (10, 10), except the walls.

A is a set of all actions, which is {left, right, up, down}

(2) It has 102 states in total, except the walls, start and goal.

Most of the state have 4 valid actions.

Others have 3 or 2 when they are against or into the walls.

Each action they take will come with 3 possible movement.

$102 \times 3.5 \times 3 \approx 1070$

Therefore, approximately, the number of non-zero rows is around 1070.

2.

(1) expected return with discounts for episodic case:

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$$

where T is the termial state

Hence, given all rewards zero except for -1 upon failure,

episodic case: $G_{te} = -\gamma^{T-t-1}$

For the countinuing case: $G_{tc} = -\sum_{k=0}^{\infty} \gamma^k$, where time step k is the time a reward received in the future

As we see, the continuing function has a summation at the beginning, which makes the discounted factor become helpful to preventing the return from blowing up.

(2) Because all states are given rewards zero except the goal, the agent never know which state it should go next is better. It doesn't have enough information to learn, which mean we have poor communication with agent. To communicate effectively, we shoul set and update the rewards for each states periodically.

3.

(a) $G_t = R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} \cdots + \gamma^{T-t-1} R_T$

$G_5 = G_T = 0 \quad G_4 = R_5 = 5 \quad G_3 = R_4 + \gamma R_5 = 4 \quad G_2 = R_3 + \gamma R_4 + \gamma^2 R_5 = 8$

$G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \gamma^3 R_5 = 6 \quad G_0 = R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \gamma^4 R_5 = 2$

(b) $G_1 = R_2 + \gamma R_3 + \cdots + \gamma^{n-2} R_n = \sum_{i=0}^{n-2} \gamma^i R_n$

$= \frac{1}{1-\gamma} R_n = \frac{1}{1-0.9} \times 7 = 70$

$G_0 = R_1 + \gamma G_1 = 2 + 0.9 \times 70 = 65$

4. $G_0 = R_1 + \gamma R_2 + \cdots + \gamma^{100} R_{101} \approx R_1 + \frac{1}{1-\gamma} R_2$

To choose UP, $50 - \frac{1}{1-\gamma} > -50 + \frac{1}{1-\gamma}$

$\gamma < 0.98$

Otherwise, choose Down.

5. (a)

$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$

$V_\pi(s) = E_\pi[G_t \mid S_t = s]$

$= E_\pi[\sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \mid S_t = s]$

Adding a constant c to all rewards:

$$V_{\pi c}(s) = E_\pi[\sum_{k=t+1}^{T} \gamma^{k-t-1}(R_k + c) \mid S_t = s]$$

$$= E_\pi[\sum_{k=t+1}^{T} \gamma^{k-t-1} \cdot R_k \mid S_t = s] + E_\pi[\sum_{k=t+1}^{T} \gamma^{k-t-1} \cdot c \mid S_t = s]$$

$$= V_\pi(s) + \frac{c}{1-\gamma} \qquad V_c = \frac{c}{1-\gamma}$$

V_c is a constant. Thus, adding a constant c to all rewards doesn't affect the value.

(b) In episodic task, the equation above no longer exist, because $\gamma = 1$.

Instead, it will become $V_{\pi c}(s) = V_\pi(s) + (T-t-1)c$
where T is the terminal step.

The agent will seek for a longer path (larger T) to get higher expected return. For example, a maze runner task has reward -0.1 at each step and $+10$ at terminal state. If we add 10 to every reward. Then the agent would stay as long as it can to earn more reward, hovering around.

6.

Bellman Equation: $V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r \mid s,a)[r + \gamma V_\pi(s')]$

(a) $$V_\pi(s) = \frac{1}{4} \times 1 \times (0 + 0.9 \times 0.7) + \frac{1}{4} \times 1 \times (0 + 0.9 \times 0.4) + \frac{1}{4} \times 1 \times (0 + 0.9 \times 2.3) + \frac{1}{4} \times 1 \times (0 + 0.9 \times (-0.4))$$

$$= 0.675 \approx 0.7$$

(b) Getting the max V_* by moving up or left.

$$V_*(s) = \max_a \sum_{s',r} p(s',r|s,a)[r + \gamma V_*(s')]$$

$$= 0.5 \times 1 \times (0 + 0.9 \times (19.8)) + 0.5 \times 1 \times (0 + 0.9 \times (19.8))$$

$$= 17.82$$

7. (a) The value function should be $\frac{1}{2}$ as it only got reward +1 on the right with equal probability.

$V(L) = V(R) = 0$, since these are terminal states

Verify: $V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a)[r+\gamma V_\pi(s')]$

$= \frac{1}{2} + 0 = \frac{1}{2}$

Hence, $V(s) = \frac{1}{2}$ is consistent with Bellman equation.

(b) Guess: $\begin{cases} V(A) = \frac{1}{6} & V(D) = \frac{2}{3} \\ V(B) = \frac{1}{3} & V(E) = \frac{5}{6} \\ V(C) = \frac{1}{2} \\ V(L) = V(R) = 0 \end{cases}$

Verify: $\begin{cases} V(A) = \frac{1}{2}\times 0 + \frac{1}{2}V(B) = \frac{1}{6} \\ V(B) = \frac{1}{2}V(A) + \frac{1}{2}V(C) = \frac{1}{3} \\ V(C) = \frac{1}{2}V(B) + \frac{1}{2}V(D) = \frac{1}{2} \\ V(D) = \frac{1}{2}V(C) + \frac{1}{2}V(E) = \frac{2}{3} \\ V(E) = \frac{1}{2}V(D) + \frac{1}{2} = \frac{5}{6} \end{cases}$

(c) Assuming there are states n, the value function of k th state is:

$$V(k) = \frac{k-1}{n-1}$$

8. (a) Bellman equation: $V_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s',r|a,s)[r+\gamma V_\pi(s')]$

Expand this for the 2 states, H for high, L for low,

$$V_\pi(S_H) = \pi(\text{search}|S_H)\cdot[\alpha(r_{search}+\gamma V_\pi(S_H)) + (1-\alpha)(r_{search}+\gamma V_\pi(S_L))]$$
$$+ \pi(\text{wait}|S_H)(r_{wait}+\gamma V_\pi(S_H))$$

$$V_\pi(S_L) = \pi(\text{search}|S_L)\cdot[\beta(r_{search}+\gamma V_\pi(S_L)) + (1-\beta)(-3+\gamma V_\pi(S_H))]$$
$$+ \pi(\text{wait}|S_L)\cdot(r_{wait}+\gamma V_\pi(S_L)) + \pi(\text{recharge}|S_L)\cdot \gamma V_\pi(S_H)$$

(b) $\begin{cases} V_\pi(S_H) = 1\times[0.7\times(10+0.9V_\pi(S_H)) + 0.3\times(10+0.9V_\pi(S_L))] \\ V_\pi(S_L) = 0.5\times(3+0.9V_\pi(S_L)) + 0.5\times 0.9V_\pi(S_H) \end{cases}$

Solved $\Rightarrow \begin{cases} V_\pi(S_H) = 72.012 \\ V_\pi(S_L) = 61.646 \end{cases}$ Checked. It satisfies the Bellman equation.

(c) Rewrite the formulation, using θ:

$$\begin{cases} V_\pi(S_H) = 1 \times [0.7 \times (10 + 0.9 V_\pi(S_H)) + 0.3 \times (10 + 0.9 V_\pi(S_L))] \\ V_\pi(S_L) = \theta \cdot (3 + 0.9 V_\pi(S_L)) + (1-\theta) \times 0.9 V_\pi(S_H) \end{cases}$$

Solved

$$\begin{cases} V_\pi(S_H) = 27.03 + 0.73 V_\pi(S_L) \\ V_\pi(S_L) = \dfrac{900 - 789\theta}{12.7 + 57.6\theta} \\ \quad = \dfrac{1073}{12.7 + 57.6\theta} - 13.7 \end{cases}$$

Hence, $\theta = 0$ will maximize the value function.

$$V_\pi(S_H) = 78.71 \qquad V_\pi(S_L) = 70.79$$

9. (a) Equation can be given as $V_\pi(s) = \sum_a \pi(a|s) q_\pi(s,a)$

(b) Equation can be given as $q_\pi(s,a) = \sum_{s',r} p(s',r|a,s) [r + \gamma V_\pi(s')]$

(c) $q_\pi(s,a) = \sum_{s',r} p(s',r|a,s) [r + \gamma \sum_{a'} \pi(a'|s') q_\pi(s',a')]$