I think it as a continuing MDP.

1. (a) State space : All the grids in the 4 rooms domin except the wall grids

Action space : {up, down, left, right}

b. total grid : 11×11, wall grids : 17

$|S| = 121 - 17 = 104$ ①

$|A| = 4$ ②

$|S'| = |T(S,A)| \in [1, 3]$ ③

$|r| = |R(S,A,S')| = 1$ ④

From ①,②,③,④, the number of non-zero $p(s',r|s,a)$ falls in the range of $[104 \times 4 \times 1, 104 \times 4 \times 3] = [416, 1248]$

C. Pseudocode : Instead of considering enviroment dynamics, let's consider the stochastic action. In this setting, the environment is deterministic, the transition function $T(S,A)$ is deterministic, but the action is stochastic. For each action in the action space, the agent has 0.8 probability of taking the correct action a, but with 0.1 probability of taking its orthogonal actions n_1, n_2 respectively.

Input : deterministic $T(S,A)$, $R(S,A,s')$

state space S, Action space A

(if $s' = T(S,A)$ is the wall grid, $s' = s$); $R(S,A,s') = 0$ except $s = (10,10)$

Output : full table of $p(s',r|s,a)$.

(see below)

```
For s in State space S :
    For a in Action space A :
①       s'_1 = T(s, a) ; r_1 = R(s, a, s'_1)
        p(r_1, s'_1 | s, a) = 0.8
②       s'_2 = T(s, n_1) ; r_2 = R(s, n_1, s'_2)
        if s'_2 == s'_1 and r_1 == r_2 :
            p(r_1, s'_1 | s, a) = 0.1 + p(r_1, s'_1 | s, a)
        else :
            p(s'_2, r_2 | s, a) = 0.1
③       s'_3 = T(s, n_2) ; r_3 = R(s, n_2, s'_3)
        if s'_3 == s'_1 and r_3 == r_1 ;
            p(r_1, s'_1 | s, a) = p(r_1, s'_1 | s, a) + 0.1
        elif s'_3 == s'_2 and r_3 = r_2 :
            p(r_2, s'_2 | s, a) = p(r_2, s'_2 | s, a) + 0.1
        else :
            p(s'_3, r_3 | s, a) = 0.1
④       for i = 1 : 3 :
            if p(r_i, s'_i | s, a) > 0 :
            store p(r_i, s'_i | s, a) in D
        end for
    end for
end for
Return D
```

2. a.

episodic $\begin{cases} G_t = -\gamma^{T-t-1} \\ G_T = 0 \end{cases}$, where T is the terminal time step.

continuing case $G_t = \sum_{i=1} -\gamma^{k_i-t-1}$, $k \in \{k_1, k_2, k_3 \cdots k_\infty\}$

where k_i is the time step of ith failure after t.

b. We didn't design the reward mechanism well:

The agent doesn't care about how many time steps it takes to complete the task. e.g. Taking 10 steps to get out of the maze has the same return of taking 100 steps to complete it.

3. a.

$G_5 = 0$

$G_4 = R_5 + \gamma \cdot G_5 = 2$

$G_3 = R_4 + \gamma \cdot G_4 = 3 + 0.5 \times 2 = 4$

$G_2 = R_3 + \gamma \cdot G_3 = 6 + 0.5 \times 4 = 8$

$G_1 = R_2 + \gamma \cdot G_2 = 2 + 0.5 \times 8 = 6$

$G_0 = R_1 + \gamma G_1 = -1 + 0.5 \times 6 = 2$

b. $G_0 = 2 + 0.9 \cdot 7 + 0.9^2 \cdot 7 + \cdots + 0.9^n \cdot 7$, where $n \to$ infinity

$= \lim_{n \to \infty} 2 + 7 \cdot \left(0.9 \frac{1-0.9^n}{1-0.9}\right) = 2 + 7 \cdot 9 = 65$

$G_1 = 7 + 0.9 \times 7 + 0.9^2 \cdot 7 + \cdots 0.9^n \cdot 7$, where $n \to \infty$

$= \lim_{n \to \infty} 7 \cdot \left(1 \cdot \frac{1-0.9^{n+1}}{1-0.9}\right) = 70$

4. $$q(\text{start}, \text{up}) = 50 - \gamma - \gamma^2 - \gamma^3 \cdots - \gamma^{100}$$
$$= 50 - \gamma \cdot \frac{1-\gamma^{100}}{1-\gamma} \qquad (\gamma \neq 1)$$

$$q(\text{start}, \text{down}) = -50 + \gamma + \gamma^2 + \cdots + \gamma^{100}$$
$$= -50 + \gamma \cdot \frac{1-\gamma^{100}}{1-\gamma} \qquad (\gamma \neq 1)$$

For $q(\text{start}, \text{up}) > q(\text{start}, \text{down})$ and $0 \leq \gamma < 1$,

$100 > 2 \cdot \gamma \frac{1-\gamma^{100}}{1-\gamma}$ and $0 \leq \gamma < 1$ $\rightarrow$ up action better

$100 < 2 \cdot \gamma \frac{1-\gamma^{100}}{1-\gamma}$ and $0 \leq \gamma < 1$ $\rightarrow$ down action better.

$100 = 2 \cdot \gamma \frac{1-\gamma^{100}}{1-\gamma}$ and $0 \leq \gamma < 1$ $\rightarrow$ actions are equal

For $\gamma = 1$, $q(s, \text{up}) = 50 - 100 = -50$
$q(s, \text{down}) = -50 + 100 = 50$ $\longrightarrow$ down action better.

5. a. $G_t + V_c = (R_{t+1} + c) + \gamma \cdot (R_{t+2} + c) + \cdots + \gamma^n (R_{t+n+1} + c)$,
where $n \rightarrow$ infinity, and $\gamma \neq 1$ for continuing case.

$$V_c = c + \gamma c + \gamma^2 c + \cdots + \gamma^n c, \quad \text{where } n \rightarrow \infty$$
$$= \lim_{n \to \infty} c \cdot \frac{1-\gamma^{n+1}}{1-\gamma} = c \cdot \frac{1}{1-\gamma} = \frac{c}{1-\gamma}$$

b.) It would affect the episodic case. e.g. The agent try to get outs of the maze, it get -1 reward at each step ~~unit~~ until geting out. If we add 2 to the reward, it would stay in the maze to get more return instead of finding a way to get out.

6. Bellman equation

$$V(s) = \sum_a \pi(a|s) \sum_{s',r} p(r,s'|s,a)\,(r + \gamma \cdot V(s'))$$

For the state we want to calculate:

~~0.7~~ ≠ $V(s) = \frac{1}{4} \cdot 1 \cdot (0 + 0.9 \cdot 0.4) +$ (right action)

$\frac{1}{4} \cdot 1 \cdot (0 + 0.9 \cdot (-0.4)) +$ (down)

$\frac{1}{4} \cdot 1 \cdot (0 + 0.9 \cdot 0.7) +$ (left)

$\frac{1}{4} \cdot 1 \cdot (0 + 0.9 \cdot 2.3) +$ (up)

$= \frac{1}{4} \cdot (0.36 - 0.36 + 0.63 + 2.07)$

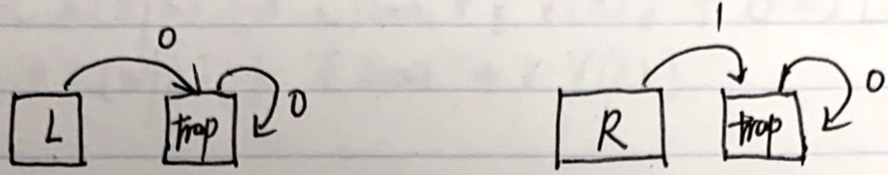
$= 0.675 \approx 0.7$

b.) For the center state has two optimal actions: up, left. We could assign random probability of the two actions only requiring the sum is equal to 1. I will show two case selecting the up action only and select the left action only.

$V(s) = q(s, up) = 1 \cdot 1 \cdot (0 + 0.9 \times 19.8) = 17.82$

$V(s) = q(s, down) = 1 \cdot 1 \cdot (0 + 0.9 \times 19.8) = 17.82$

7. a. Guess $V(L) = 0$, $V(A) = \frac{1}{2}$, $V(R) = 1$



verify: $V(A) = \frac{1}{2} \cdot 1 \cdot [0 + 1 \cdot V(L)] + \frac{1}{2} \cdot 1 \cdot [0 + 1 \cdot V(R)]$

$= \frac{1}{2}$.

$V(L) = 1 \cdot 1 \cdot [0 + 1 \cdot V(\text{trap})]$ *

$= 0$

$V(R) = 1 \cdot 1 \cdot [1 + 1 \cdot V(\text{trap})]$

$= 1$

b. Guess $V(L) = 0$, $V(A) = \frac{1}{6}$, $V(B) = \frac{2}{6}$, $V(C) = \frac{3}{6}$, $V(D) = \frac{4}{6}$
$V(E) = \frac{5}{6}$, $V(R) = 1$

Not need to verify $V(L)$, $V(R)$ again.

verify:

$V(A) = \frac{1}{2} \cdot 1 \cdot [0 \cdot 1 \cdot V(L)] + \frac{1}{2} \cdot 1 \cdot [0 \cdot 1 \cdot V(B)] = \frac{1}{6}$

$V(B) = \frac{1}{2} \cdot V(A) + \frac{1}{2} \cdot V(C) = \frac{2}{6}$

same for C, D, E

$V(n) = \frac{1}{2} \cdot V(n-1) + \frac{1}{2} V(n+1)$,

c. With arbitrary n state including L, R, the state value for the i th left state ($i = 0 \dots n-1$) is $\frac{i}{n-1}$

8. (a)

$$V(h) = \pi(s|h)\cdot\left(\alpha\cdot[r_{search} + \gamma\cdot V(h)] + (1-\alpha)[r_{search} + \gamma\cdot V(L)]\right)$$
$$+\pi(W|h)\cdot 1\cdot[r_{wait} + \gamma\cdot V(h)]$$

$$V(L) = \pi(s|L)\left(\beta\cdot[r_{search} + \gamma V(low)] + (1-\beta)[-3 + \gamma\cdot V(L)]\right)$$
$$+\pi(W|L)\left(1\cdot[r_{wait} + \gamma\cdot V(L)]\right)$$
$$+\pi(Recharge|L)\left(1\cdot[0 + \gamma\cdot V(h)]\right)$$

(b).

$$\begin{cases} V(h) = 1\cdot[0.8\cdot(10 + 0.9\cdot V(h)) + 0.2(10 + 0.9V(L))] \\ V(L) = 0.5\cdot 1\cdot[3 + 0.9V(L)] + 0.5\cdot 1\cdot(0 + 0.9V(h)) \end{cases}$$

$\downarrow$

$$\begin{cases} V(h) = 10 + 0.18V(L) + 0.72V(h) \\ V(L) = 1.5 + 0.45V(h) + 0.45V(L) \end{cases} \longrightarrow \begin{cases} V(h) = 79.0 \\ V(L) = 67.3 \end{cases}$$

Verify:

$$\begin{cases} 79 = 0.8(10 + 0.9\times 79) + 0.2(10 + 0.9\times 67.3) \\ 79 = 0.8\times 81.1 + 0.2\times 70.6 \quad (\checkmark \text{ correct}) \end{cases}$$

$$67.3 = 0.5\times(3 + 0.9\times 67.3) + 0.5\cdot 0.9\cdot 79 \quad (\checkmark \text{ correct})$$

(c). Form b, we get $0.28V(h) = 10 + 0.18V(L)$.

$\therefore V(h) = \dfrac{10 + 0.18V(L)}{0.28}$ ①

$V(L) = \theta \cdot (3 + 0.9V(L)) + (1-\theta)[0 + 0.9V(h)]$ ②

put ① into ②:

$\underline{V(L) = 90.31 - \dfrac{60316.6}{4222 - 32222\theta}}$

From ①, we could find $V(h)$, $V(L)$ are positive related.

When $\theta = 0$, $V(h)$ and $V(L)$ get their maximum value.

$$\begin{cases} V(h) = 84.6 \\ V(L) = 76.0 \end{cases}$$

9. (a) $V_\pi(s) = \sum_a \pi(a|s) \cdot q(s,a)$

(b) $q_\pi(s,a) = \sum_{s',r} p(s',r \mid s,a) \cdot [r + \gamma V_\pi(s')]$

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