Q1:

(a) state space: In this continuing case, state space is a set of all positions from (0,0) to (10,10) except the walls.

action space: all actions that can be chosen at each state. In this case all four actions can be chosen at every state, thus action space is {up, down, left, right}

(b) For a (S,a) pair, there are at most 3 non-zero $p(s',r|s,a)$ value, when the agent is not blocked by walls, and the action is "left" or "right". There is at least 1 non-zero $p(s',r|s,a)$ value, when two sides of the agent is blocked by walls and the direction of action is against the wall. And there will be 2 non-zero value if one side is blocked, with action against the wall. There are 104 states which are not walls, so the number of non-zero rows should be at most 1248 (104×4×3) and at least 416 (104×4×1). Since most of the states are unblocked, or one-side blocked, I would say the approximate number is 1000.

(c) pseudo code:

input: all valid states (all states except walls): S

action space: A = {up, down, left, right}

Transition model: $T(s'|s,a)$, return probability of next state s' given current state s and action a.

rules: 1. the probability of taking given action a equals to 0.8, and the agent also has 0.1 chance for both perpendicular move. 2. If there is a wall blocking on the direction of the action which has been chosen according to the probability, the next step S' will be the same as current state S.

reward function: All states have reward of 0, except (10,10) has reward of 1.

$r(s,a,s')=0$ for $s' \neq (10,10)$; $r(s,a,(10,10))=1$

output: prob_table. A dictionary structure that store probability of all four elements pairs $p(s',r|s,a)$.

```
prob_table = {}
for s in S:
    for a in A:
        p(S') = T(s'|s,a)        return a dictionary of the probability of next state equal to s'
        for s' in S':
            reward = r(s,a,s')
            p(s',reward|s,a) = p(S')[s']
            if not p(s',reward|s,a) in prob_table.keys():
                prob_table[p(s',reward|s,a)] = p(S')[s']
            else:
                prob_table[p(s',reward|s,a)] += p(S')[s]

return prob_table
```

Q2:

(a) episodic with discounting:

$$G_T = 0;\ G_{T-1:T} = R_T = -1$$ (T is the terminal state)

$$G_{t:T} = \gamma^{T-t-1} \cdot -1$$
$$= -\gamma^{T-t-1}$$

continuing with discounting:

$$G_t = \sum_j -\gamma^{j-t-1}$$ which j is the step that failure occurs

the main difference is that only one failure occurs in episodic case but multiple failure occurs in continuing case.

(b) No, since there isn't discount for future reward, the return of all state will be the same, which equals to 1, as long as the agent finally escape from the maze. In order to communicate more efficiently, we have to make this task discounted, and then the agent will learn a desired policy by maximizing reward.

Q3:

(a) $G_t = R_{t+1} + \gamma G_{t+1}$

$G_5 = 0$

$G_4 = R_5 + \gamma G_5 = 2$

$G_3 = R_4 + \gamma G_4 = 4$

$G_2 = R_3 + \gamma G_3 = 8$

$G_1 = R_2 + \gamma G_2 = 6$

$G_0 = R_1 + \gamma G_1 = 2$

(b)

$G_1 = R_2 + \gamma R_3 + \gamma^2 R_4 + \cdots \gamma^{n-2} R_n$

$= \sum_{i=2}^{n} \gamma^{n-2} R_n$ $\quad (R_1 = R_2 = \cdots R_n = R = 7)$

when $n \to \infty$:

$G_1 = R \frac{1}{1-\gamma} = 70$

and thus $G_0 = R_1 + \gamma G_1 = 65$

Q4: when $\gamma = 1$, choose "down" action.

when $\gamma \neq 1$, i.e. $0 < \gamma < 1$:

$$G_{up} = R_1 + \sum_{i=2}^{n} \gamma^{n-1} R_n \qquad (R_2 = R_3 = \cdots R_n = -1)$$

$$= 50 - \gamma \left(\frac{1-\gamma^{100}}{1-\gamma} \right)$$

Similarly, $G_{down} = -50 + \gamma \left(\frac{1-\gamma^{100}}{1-\gamma} \right)$

G_{up} will be greater than G_{down} when:

$$G_{up} - G_{down} = 100 - 2\gamma \left(\frac{1-\gamma^{100}}{1-\gamma} \right) > 0 \Rightarrow 50 - \gamma \left(\frac{1-\gamma^{100}}{1-\gamma} \right) > 0$$

otherwise it's better to choose "down" action.

Q5: (a)

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

$$\therefore V_\pi(S) = E_\pi[G_t | S_t = S] = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,|\, S_t = S\right]$$

now since we add a constant c, and reward. will be R_{t+k+1} at each step.

$$V_{\pi c}(S) = E_\pi\left[\sum_{k=0}^{\infty} \gamma^k (R_{t+k+1} + c) \,|\, S_t = S\right]$$

$$= E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} + \sum_{k=0}^{\infty} \gamma^k \cdot c \,|\, S_t = S\right]$$

$$= \underbrace{E_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1}\right]}_{= V_\pi(S)} + \underbrace{E_\pi\left[\sum_{k=0}^{\infty} \gamma^k \cdot c\right]}_{= \sum_{k=0}^{\infty} \gamma^k c}$$

$$= V_\pi(S) + c \cdot \frac{1}{1-\gamma}$$

$$\therefore V_{\pi c} = V_\pi + \frac{c}{1-\gamma} \Rightarrow V_c = \frac{c}{1-\gamma}$$

(b) If we train the model by giving -1 reward everywhere in the maze, then adding a positive constant that greater than 1 will make the agent stay inside the maze forever, to maximize the reward, while if we add a negative constant to the reward the result will remain unchanged, which is the agent will still able to escape from the maze as fast as possible. So there is a possibility that the task is unchanged, but not always.

Q6. (a)

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

$\therefore$ $V(s) = 0.25\times(1\times(0+0.9\times2.3)) + 0.25\times(1\times(0+0.9\times0.4)) + 0.25(1\times(0+0.9\times(-0.4)))$
$+ 0.25\times(1\times(0+0.9\times0.7))$
$= 0.675 \approx 0.7$.

(b) $V(s) = 0.5\times(1\times(0+0.9\times19.8) + 0.5\times1\times(0+0.9\times19.8))$
$= 17.82 \approx 17.8$.

Q7. (a) since all actions are chose with equal probability, I guess the state value is $\frac{1}{2}(1+0)=\frac{1}{2}$

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

$= \frac{1}{2}\times(1\times(1+1\times0)) + \frac{1}{2}\times(1\times(0+1\times0))$
$= \frac{1}{2}$

(b) derive Bellman equation for all states.

$$\begin{cases} V(A) = \frac{1}{2}\times(1\times(0+1\times0)) + \frac{1}{2}\times(1\times(0+1\times V(B))) \\ V(B) = \frac{1}{2}\times(1\times(0+1\times V(A))) + \frac{1}{2}\times(1\times(0+1\times V(C))) \\ \vdots \\ V(D) = \frac{1}{2}\times(1\times(0+1\times V(C))) + \frac{1}{2}\times(1\times(0+1\times V(E))) \\ V(E) = \frac{1}{2}\times(1\times(0+1\times1)) + \frac{1}{2}\times(1\times(0+1\times V(D))) \end{cases} \Rightarrow \begin{cases} V(A) = \frac{1}{2}V(B) \\ V(B) = \frac{1}{2}V(A) + \frac{1}{2}V(C) \\ V(C) = \frac{1}{2}V(B) + \frac{1}{2}V(D) \\ V(D) = \frac{1}{2}V(C) + \frac{1}{2}V(E) \\ V(E) = \frac{1}{2} + \frac{1}{2}V(C) \end{cases}$$

$$\Rightarrow \begin{cases} V(A) = \frac{1}{6} \\ V(B) = \frac{2}{6} \\ V(C) = \frac{3}{6} \\ V(D) = \frac{4}{6} \\ V(E) = \frac{5}{6} \end{cases}$$

(c) If there are n states, the value of the kth state from the left is:

$V(S_k) = \frac{k-1}{n-1}$ when $k \in (2,3,\cdots n-1)$

Q8:

(a)

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

$$= \pi(search|high)[\alpha\cdot(r_{search}+\gamma V(high))+(1-\alpha)(r_{search}+\gamma V(low))] +$$

$$\pi(wait|high)[1\cdot(r_{wait}+\gamma\cdot V(high))]$$

Similarly. $V(low) = \pi(search|low)\cdot[(1-\beta)(-3+\gamma V(high))+\beta(r_{search}+\gamma V(low))] +$

$$\pi(wait|low)\cdot[1\cdot(r_{wait}+\gamma\cdot V(low))] +$$

$$\pi(recharge|low)\cdot[1\cdot(r_{wait}+\gamma\cdot V(high))]$$

(b) $\pi(search|high)=1$, $\pi(wait|low)=0.5$, $\pi(recharge|low)=0.5$, $\alpha=0.8$, $\beta=0.6$, $\gamma=0.9$

$V_H - V(high)$, $V_L - V(low)$:

$$\begin{cases} V_H = 0.8\times(10+0.9\times V_H)+0.2\times(10+0.9V_L) \\ V_L = 0.5\times1\times(3+0.9\times V_L)+0.5\times1\times(0.9+V_H) \end{cases} \Rightarrow \begin{cases} 0.55V_L = 1.5+0.45V_H \\ 0.28V_H = 10+0.18V_L \end{cases}$$

solve the equations, and obtain: $\begin{cases} V_H = 79.04 \\ V_L = 67.39 \end{cases}$

Verification: $0.8\times(10+0.9\times79.04)+0.2\times(10+0.9\times67.39) = 64.90+14.13 = 79.03 \approx V_H$

$0.5\times1\times(3+0.9\times67.39)+0.5\times1\times(0.9\times79.04) = 31.83+35.57 = 67.40 \approx V_L.$

(c). substitute θ into V_L:

$V_L = \theta(3+0.9V_L)+(1-\theta)\times0.9V_H$, substitute into equations:

$$\Rightarrow \begin{cases} (1-0.9\theta)V_L = 3\theta+0.9(1-\theta)V_H \\ 0.28V_H = 10+0.18V_L \end{cases}$$

Solve the equation, we get:

$$V_L = \frac{9-8.16\theta}{0.118-0.09\theta}$$

$$= 90.67 - \frac{1.6986}{0.118-0.09\theta}$$

$\therefore$ when $\theta=0$, V_L maximized and equal to 76.28, and then $V_H = 84.75$

Q9.

(a)

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

(b)

$$q_\pi(s,a) = E[R_{t+1} + \gamma V_\pi(S_{t+1}) \mid S_t = s, A_t = a]$$
$$= \sum_{s',r} p(s',r|s,a)\,[r + \gamma V_\pi(s')]$$

(c)

- substitute $V_\pi(s') = \sum_{a'} \pi(a'|s')\, q(s',a')$

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