$$Q(s,a) = E[G_t \mid S_t = s, A_t = a]$$

$$E[G_t \mid s, a] \Rightarrow$$

then given $P(s', r \mid s, a)$ $\boxed{E[R_{t+1} + \gamma G_{t+1} \mid s, a] \ (1)}$

$$Q(s,a) = \sum_{s', r} P(s', r \mid s, a) \cdot G_t \quad (2)^*$$

here $Q(s,a)$ is defined as sum of possible returns, weighted by it's pobabilities.

if we substitude (1) inte (2) will get

$$Q(s,a) = \sum_{s', r} P(s', r \mid s, a) \cdot (r + \gamma \cdot E[G_t \mid s'])$$

so here next reward R_{t+1} is r and expected return of next step depend on next state $E[G_t \mid s']$

$E[G_t \mid s]$ is value function so

$$Q(s,a) = \sum_{s', r} P(s', r \mid s, a) \cdot (r + \gamma \cdot V(s'))$$

so Q funtion is defined as sum of reward from geting from state s to s' plus discount factor multiplied on Value funktion of the next state.

Now we could link Q function at current state to Value function of next state.

Now Value function:

V(s) is sum of all possible returns weighted by ~~por~~ probabilities that agent take action (policy) $(4)^*$

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

$$V_\pi(s) = \sum_a \pi(a \mid s) \cdot \cancel{E} \ G_{t+1}$$

$$G_t = R_{t+1} + \gamma \cdot G_{t+1} \quad (3)^*$$

from (3) it is clear that G_t is return = reward after taking some action a plus future rewards.

Then Value function is expected return at state s.

Now it is more clear definition (4) and formula looks:

$$V(s) = \sum_a P(a|s) \cdot E[R_{t+1} + \gamma \cdot G_{t+1} | a, s]$$

the return expected on state is sum of returns based on state and action $E[R_{t+1} + \gamma G_{t+1} | A_t = a, S_t = s]$ weighted by probabilities of taking this action by policy.

Where $E[R_{t+1} + \gamma G_{t+1} | A_t = a, S_t = s] = Q(s, a)$

so

$$V(s) = \sum_a P(a|s) \cdot Q(s, a)$$

Now we also have link between action value and Q function

So we have

$$Q(s, a) = \sum_{s', r} P(s', r | s, a) \cdot [r + \gamma \cdot V(s')] \quad (5)^*$$

$$V(s) = \sum_a P(a|s) \cdot Q(s, a) \quad (6)^*$$

Now to get relation between current and next value function we could substitude (5) into (6)

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

Assignment 3

(1)

(1.1) $P(a_1) = P(a_2) = 0.5$

define value functions for terminal states as

$V(s_1) = -1$ $V(s_5) = 1$, these might be also zeros, but $V(s_5) \geq$ reward r, $V(s_1) \leq$ reward

For non terminal states $\in \{s_2, s_3, s_4\}$

$$V(s) = \sum_a P(a|s) \cdot \sum_{s',r} P(s',r|s,a) \cdot [r + \gamma \cdot V(s')]$$

since $P(a|s) = \frac{1}{2}$ for any action and $P(s',r|s,a) = 1$ deterministic environment

$$V(s) = \frac{1}{2} \cdot [(r_{s-1} + \gamma \cdot V_{(s-1)}) + (r_{s+1} + \gamma \cdot V(s+1))]$$

For our case:

- $V(s_2) = \frac{1}{2} \cdot [(0 + \frac{1}{2} \cdot V(s_1)) + (0 + \frac{1}{2} \cdot V(s_3))] =$

$$= \frac{1}{4} \cdot (-1) + \frac{1}{4} V(s_3) = \frac{1}{4}(V(s_3) - 1)$$

- $V(s_3) = \frac{1}{4}(V(s_2) + V(s_4))$

- $V(s_4) = \frac{1}{4}(V(s_3) + 1)$

Terminal states doesn't have any possible actions so we just assign constants

$V(s_1) = -1$ $V(s_5) = 1$

Additional solve gpi might use this equations to gradually update state values.

(1.2)

again

$V(s_1) = -1 \quad V(s_5) = 1$

For non-terminal state

When agent always go to the green terminal state, without cost of living is optimal policy.

$$V(s_2) = 1 \cdot (\gamma V(s_3) + 0) = \gamma V(s_3) = \frac{V(s_3)}{2}$$

$$V(s_3) = 1 \cdot (\gamma V(s_4) + 0) = \gamma \cdot V(s_4) = \frac{V(s_4)}{2}$$

$$V(s_4) = 1 \cdot (\gamma \cdot 1 + 0) = \gamma 1 = \frac{1}{2}$$

(2)

(2.1) $V(s_1) = -1 \quad V(s_5) = 1$

$$V(s_i) = \frac{1}{2} \cdot \left(2 \cdot c + \gamma \cdot V(s_{i-1}) + \gamma \cdot V(s_{i+1})\right)$$

random agent

$$V(s_2) = \frac{1}{2} \cdot \left[-0.1 + \gamma \cdot V(s_3) + \left((-0.1) + \gamma \cdot V(s_1)\right)\right] =$$

$$= \frac{1}{2} \cdot \left(-0.7 + \frac{1}{2} \cdot V(s_3)\right)$$

$$V(s_3) = \frac{1}{2} \cdot \left[-0.1 + \gamma \cdot V(s_2) + \left(-0.1 + \gamma \cdot V(s_4)\right)\right] =$$

$$= \frac{1}{2} \cdot \left(-0.2 + \frac{1}{2} \cdot V(s_2) + \frac{1}{2} \cdot V(s_3)\right)$$

$$V(s_4) = \frac{1}{2} \cdot \left(-0.2 + \frac{1}{2} \cdot V(s_3) + \frac{1}{2} \cdot V(s_5)\right) =$$

$$= \frac{1}{4}\left(0.8 + \frac{1}{2} \cdot V(s_3)\right)$$

(2.2)

$V(S_1) = -1 \qquad V(S_5) = 1$

$$V(S_2) = \frac{1}{2} \cdot \left(-3 + \frac{1}{2} \cdot V(S_3)\right)$$
$$V(S_3) = \frac{1}{2} \cdot \left(-2 + \frac{1}{2} \cdot V(S_2) + \frac{1}{2} V(S_4)\right)$$
$$V(S_4) = \frac{1}{2} \cdot \left(-1 + \frac{1}{2} \cdot V(S_3)\right)$$

(2.3) Agent goal of policy is to maximise return so as in 2.2 example it's not always necessary to go for positive termination.

c - cost of living

Initial state is S_2

so

$$r_1 + c \cdot 1 < r_5 + 3 \cdot c$$
$$-1 + c < 1 + 3c$$
$$c < -1$$

If $c < -1$ it make sense for agent to go left and exit rather then going to S_5

(3)

(3.1) to get Q function equation substitude (6) into (5)

$$Q(s,a) = \sum_{s',r} p(s',r \mid s,a) \cdot \left[r + \gamma \cdot \sum_{\hat{a}} P(\hat{a} \mid s') \cdot Q(s',\hat{a}\right.$$

since we have deterministic environme

$P(s',r \mid s,a) = 1$ for any state, action

so $\quad Q(s,a) = r + \gamma \cdot \sum_{\hat{a}} P(\hat{a} \mid s) \cdot Q(s',\hat{a})$

For S_1 ~~Actio~~

$Q(S_1,\leftarrow) = r = -1$

$Q(S_1,\rightarrow) = 0 + \gamma \cdot \left(\frac{1}{2}\left(Q(S_2,\leftarrow) + Q(S_2,\rightarrow)\right)\right)$

$Q(S_2,\leftarrow) = 0 + \gamma \cdot \left(\frac{1}{2}\left(Q(S_1,\leftarrow) + Q(S_1,\rightarrow)\right)\right)$

$Q(S_2,\rightarrow) = 0 + \gamma \cdot \frac{1}{2}\left(Q(S_3,\leftarrow) + Q(S_3,\rightarrow)\right)$

$Q(S_3,\leftarrow) = 0 + \gamma \cdot \frac{1}{2}\left(Q(S_2,\leftarrow) + Q(S_2,\rightarrow)\right)$

$Q(S_3,\rightarrow) = 0 + \gamma \cdot \frac{1}{2}\left(Q(S_4,\leftarrow) + Q(S_4,\rightarrow)\right)$

$Q(S_4,\leftarrow) = 0 + \gamma \cdot \frac{1}{2}\left(Q(S_3,\leftarrow) + Q(S_3,\rightarrow)\right)$

$Q(S_4,\rightarrow) =$
$= 1$

Then

- $Q(S_1,\leftarrow) = -1 \qquad Q(S_1,\rightarrow) = \frac{1}{4} Q(S_2,\leftarrow) + \frac{1}{4} Q(S_2,\rightarrow)$
- $Q(S_2,\leftarrow) = \frac{1}{4} \cdot Q(S_1,\leftarrow) + \frac{1}{4} \cdot Q(S_1,\rightarrow)$

  $Q(S_2,\rightarrow) = \frac{1}{4} \cdot Q(S_3,\leftarrow) + \frac{1}{4} \cdot Q_1(S_3,\rightarrow)$
- $Q(S_3,\leftarrow) = \frac{1}{4} \cdot Q(S_2,\leftarrow) + \frac{1}{4} \cdot Q(S_2,\rightarrow)$

  $Q(S_3,\rightarrow) = \frac{1}{4} \cdot Q(S_4,\leftarrow) + \frac{1}{4} \cdot Q(S_4,\rightarrow) = \frac{1}{4} Q(S_4,\leftarrow) + \frac{1}{4}$
- $Q(S_4,\leftarrow) = \frac{1}{4} \cdot Q(S_3,\leftarrow) + \frac{1}{4} \cdot Q(S_3,\rightarrow)$

  $Q(S_4,\rightarrow) = 1$

3.2 When agent always move right

$P(\rightarrow|S) = 1$ for any state

so $Q(s,a) = r + \gamma \cdot Q(s',\rightarrow)$

so

- $Q(s_2, \leftarrow) = -1$

$Q(s_2, \rightarrow) = 0 + \gamma \cdot Q(s_3, \rightarrow) = \frac{1}{2} \cdot Q(s_3, \rightarrow)$

- $Q(s_3, \leftarrow) = 0 + \gamma \cdot Q(s_2, \rightarrow) = \frac{1}{2} \cdot Q(s_2, \rightarrow)$

$Q(s_3, \rightarrow) = 0 + \gamma \cdot Q(s_4, \rightarrow) = \frac{1}{2} \cdot Q(s_4, \rightarrow) = \frac{1}{4}$

- $Q(s_4, \leftarrow) = 0 + \gamma \cdot Q(s_3, \rightarrow)$

- $Q(s_4, \rightarrow) = 1$

3.3

When we calculated Q values for a random policy we get good exploration of environment

When

$$V(s) = \sum_a \pi(a|s) \cdot q(s,a) = \\ = \frac{1}{2} \cdot q(s,a)$$

optimal policy means taking always maximal $q(s,a)$ so

$V(s) = \max_a q(s,a)$, so value function should change

~~The Q function doesn't change because it doesn't depend on the policy.~~

When we change to the policy which always select highest Q function the expected value function of this state is maximal from all possible, so it at least keep V function the same (if random policy already does best performance) or increas V funtion for each state.

Q function also use V function inside even though action is already taken and policy doesn't affect on it,

with the optimal policy we lead to increase future rewards, which are used in Q calculation.

So Q function will also increase which is also reflected in previous assignments.

So getting to optimal policy lead to better performance. This is what GPI is doing