3.11

$$E[R_{t+1} \mid S_t = s] = \sum_a \pi(a|s_t) \sum_{s',r} p(s', r|s, a) r$$

3.12

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

3.13

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

3.14

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

$$V_{\text{calc}} = \frac{1}{4} \times 0.9 \times (2.3 + 0.7 - 0.4 + 0.4) = 0.675 \approx 0.7$$

3.15

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$$

Adding a Constant C

~~$G'_t = G_t + C = C$~~

$$G'_t = G_t + \sum_{k=t}^{\infty} \gamma^k C = G_t + \frac{C}{1-\gamma}$$

$$v^*_\pi(s) = E[G^*_t \mid s_t = s] = E\left[G_t + \frac{C}{1-\gamma} \,\middle|\, s_t = s\right] = E[G_t \mid s_t = s] + \frac{C}{1-\gamma}$$

The new $v^*_\pi(s)$ does not affect the relative difference among states.

Q3.16

The sign of the reward has critical influence on the episodic rewards because episodic tasks use negative rewards to accelerate the agent finishing the task. The sign of the agent would impact how the agent moves. Furthermore, if the negative rewards remains negative but the value of it shrinks too much, it will give a wrong signal to the agent that the time of completing the job is not important.

Q3.17

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

$$= E_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a]$$

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

Q3.18

$$V_\pi(s) = E_\pi[q_\pi(S_t, A_t) | S_t = s, A_t = a]$$

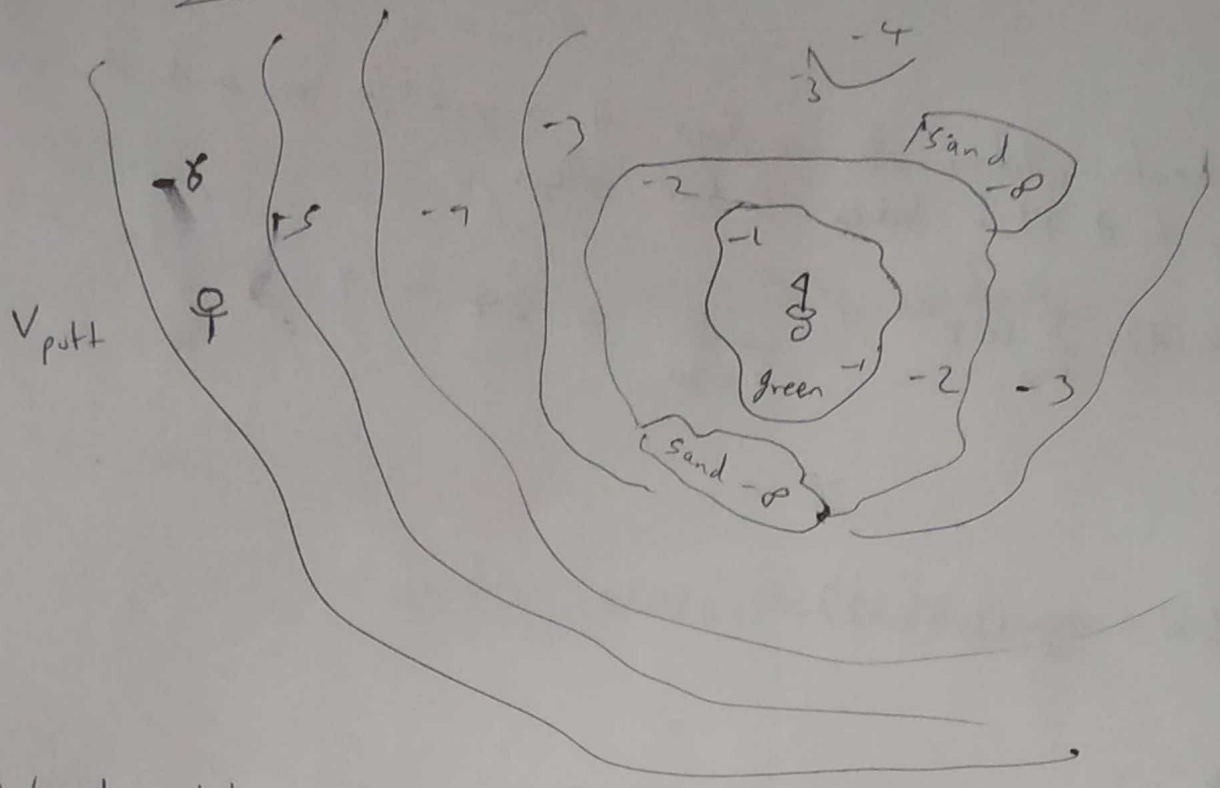
$$= \sum_a \pi(a|s)\, q_\pi(s,a)$$

Q3.19

$$q_\pi(s,a) = E_\pi[R_{t+1} + V_\pi(s') | S_t = s, A_t = a]$$

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

Q 3.20 & 3.21



Optimal state values according to driver when off green, then according to putter on the green

Optimal policy is to use driver when off green & putter when on green.

Q 3.22

$$G_{\pi_{left}} = \sum_{i=0}^{\infty} \gamma^{2i} = \frac{1}{1-\gamma^2} \qquad G_{\pi_{right}} = \sum_{i=0}^{\infty} 2\gamma^{(i+2)} = \frac{2\gamma}{1-\gamma^2}$$

If $\gamma > 0.5$, right is optimal

If $\gamma < 0.5$, left is optimal

If $\gamma = 0.5$, both are optimal

Q.3.24

The best solution after reaching A is quickly go back A after moving to A. That takes 5 time steps.

$$v_*(A) = \sum_{t=0}^{\infty} 10\gamma^{5t} = \frac{10}{1-\gamma^5} \approx 24.419$$

Q 3.25

$$v_*(s) = \max_a (q_\pi(s,a))$$

Q 3.26

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

Q 3.27

$$a_* = \arg \pi_*(a_*|s) = \arg \max_a q_*(s,a)$$

Q 3.28

$$a_* = \arg \pi_*(a_*, s) = \arg \max_{s,r} \sum p'(s,r|s,a)[r + \gamma v_*(s')]$$

3.25

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

$$V_*(s) = E_\pi[G_t \mid S_t = s]$$
$$= \sum [r(s,a) + \gamma \sum_{s'} p(s'|s,a) V_*(s')] \pi_*(s,a)$$

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

$$q_*(s,a) = E_{\pi_*}[G_t \mid S_t = s, A_t = a]$$
$$= E_{\pi_*}[R_{t+1} + \gamma G_{t+1} \mid S_{t+1} = s', A_t = a]$$
$$= r(s,a) + \gamma \sum_{s'} p(s'|s,a) \sum_{a'} q_*(a',s') \pi_*(a'|s')$$