Reinforcement Learning: Homework #5

Due on May 15, 2020 at 11:59am

Professor Ziyu Shao

Junjie He
2019233152

Problem 1

Solution

(a) We have

$$\pi_*(a|s) = \begin{cases} 1 & if\ a = \arg\max_{a \in \mathcal{A}} q_*(s,a) \\ 0 & otherwise \end{cases}$$

$$\begin{aligned} v_*(s) &= \max_a q_*(s,a) \\ &= R_s^{a^*} + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^{a^*} q_*(s', a^*) \end{aligned} \tag{1}$$

We can obtain a^* form π_*. In the problem, we have $P_{ss'}^{a^*} = 1$, since actions are deterministic. For state G,

$$q_*(G,\cdot) = 1 + \gamma \max\{q_*(G,\cdot)\} = 1 + \gamma q_*(G,\cdot) \tag{2}$$

Then we can obtain

$$q_*(G,\cdot) = \frac{1}{1-\gamma}.$$

Since $v_*(s) = \max_a q_*(s,a)$, $v_*(G) = q_*(G,\cdot) = \frac{1}{1-\gamma}$.

$$\begin{aligned} v_*(s_{n-1}) &= R_{s_{n-1}}^{a^*} + \gamma \sum_{s' \in \mathcal{S}} P_{s_{n-1}s'}^{a^*} q_*(s', a^*) = 0 + \gamma q_*(G,\cdot) = \frac{\gamma}{1-\gamma} \\ v_*(s_{n-2}) &= 0 + \gamma q_*(s_{n-1}, a^*) = \frac{\gamma^2}{1-\gamma} \\ &\vdots \\ v_*(s_{n-t}) &= 0 + \gamma q_*(s_{n-t+1}, a^*) = \frac{\gamma^t}{1-\gamma} \\ &\vdots \\ v_*(s_1) &= 0 + \gamma q_*(s_2, a^*) = \frac{\gamma^{n-1}}{1-\gamma} \end{aligned} \tag{3}$$

(b) If $\gamma = 0$, value of γ does not change the ordering of states, so the optimal policy is the same; however, the value of the value function depends on γ. If $\gamma = 0$ then, policy $\pi(s) = a_0, \forall s$ is still an optimal policy; however, this is not the only optimal policy. For example, $\pi(s_1) = a_1$ is also a optimal policy.

(c) No effect on the optimal policy. Adding a constant c to all the rewards only changes the value of each state by a constant v_c for any policy π:

$$\begin{aligned} v_{new}^{\pi}(s_i) &= \sum_{t=0}^{\infty} \gamma^t (r_t + c) \\ &= \sum_{t=0}^{\infty} \gamma^t r_t + \sum_{t=0}^{\infty} \gamma^t c \\ &= v_{old}^{\pi}(s_i) + \frac{c}{1-\gamma} \end{aligned} \tag{4}$$

, since $\sum_{t=0}^{\infty} x^t = \frac{1}{1-x}$, when $x \in [-1, 1]$.

(d)

$$\begin{aligned} v^{\pi}_{new}(s_i) &= \sum_{t=0}^{\infty} \gamma^t a(r_t + c) \\ &= a\sum_{t=0}^{\infty} \gamma^t r_t + a\sum_{t=0}^{\infty} \gamma^t c \\ &= av^{\pi}_{old}(s_i) + \frac{ac}{1-\gamma} \end{aligned} \tag{5}$$

So if $a > 0$ then the optimal policy will not change, and the value of the new optimal policy is a linear mapping of the previous optimal value function $av_*(s_i) + \frac{ac}{1-\gamma}$. If $a = 0$ then all states have reward 0 and any policy is the optimal policy, and the optimal value of all states is 0. If $a < 0$, any policy that never reaches to the state G is the optimal policy with value $\frac{ac}{1-\gamma}$ for all states s_i and $\frac{a(c+1)}{1-\gamma}$ for state G.

Problem 2

Solution

(a)

$$v = \sum_{t=0}^{\infty} \gamma^t r_t = 0 + \sum_{t=1}^{\infty} \gamma^t 1 = \frac{\gamma}{1-\gamma} \tag{6}$$

(b)

$$v = \sum_{t=0}^{\infty} \gamma^t r_t = \frac{\gamma^2}{1-\gamma} + \sum_{t=1}^{\infty} \gamma^t 0 = \frac{\gamma^2}{1-\gamma} \tag{7}$$

Since $\frac{\gamma^2}{1-\gamma} < \frac{\gamma}{1-\gamma}$, optimal action is a_1.

(c) For all iterations $v_n(s_2) = 0$, so $q(s_,a_0) = \frac{\gamma^2}{1-\gamma}$. Value iteration keep choosing the sub-optimal action while $q(s_0, a_2) > q(s_0, a_1)$. Value iteration updates are as following,

$$\begin{aligned} q_{n+1}(s_0, a_1) &= 0 + \gamma v_n(s_1) \\ v_{n+1}(s_1) &= 1 + \gamma v_n(s_1) \end{aligned} \tag{8}$$

$$\begin{aligned} q_{n+1}(s_0, a_1) &= 0 + \gamma(1 + \gamma v_n(s_1)) \\ &= \gamma(1 + \gamma + ... + \gamma^{n-1} + \gamma^n v_{n=0}(s_1)) \\ &= \gamma(\frac{1-\gamma^n}{1-\gamma}) \end{aligned} \tag{9}$$

$$\begin{aligned} \gamma(\frac{1-\gamma^{n^*}}{1-\gamma}) &= \frac{\gamma}{1-\gamma} \\ n^* &= \frac{\log(1-\gamma)}{\log(\gamma)} \\ &= \frac{\log(1-\gamma)}{\log(1-1+\gamma)} \\ &\geq \log(1-\gamma)\frac{2+\gamma-1}{2(\gamma-1)} \\ &= -\log(\frac{1}{1-\gamma})\frac{\gamma+1}{-2(1-\gamma)} \\ &\geq \frac{1}{2}\log(\frac{1}{1-\gamma})\frac{1}{1-\gamma} \end{aligned} \tag{10}$$

Where the first inequality follows by $\log(1+x) \leq \frac{2x}{2+x}$ for $x \in (-1, 0]$, and the log is natural logarithm.

Problem 3

Solution

(a) By construction of π, $\tilde{Q}(s, \pi(s)) \geq \tilde{Q}(s, \pi^*(s))$.

$$\begin{aligned} V^*(s) - Q^*(s, \pi(s)) &= V^*(s) - \tilde{Q}(s, \pi(s)) + \tilde{Q}(s, \pi(s)) - Q^*(s, \pi(s)) \\ &\leq V^*(s) - \tilde{Q}(s, \pi^*(s)) + \varepsilon \\ &= Q^*(s, \pi^*(s)) - \tilde{Q}(s, \pi^*(s)) + \varepsilon \\ &\leq 1\varepsilon \end{aligned} \tag{11}$$

(b)

$$\begin{aligned} V^*(s) - V_\pi(s) &= V^*(s) - Q^*(s, \pi(s)) + Q^*(s, \pi(s)) - V_\pi(s) \\ &\leq 2\varepsilon + Q^*(s, \pi(s)) - Q^\pi(s, \pi(s)) \\ &= 2\varepsilon + \gamma \mathrm{E}_{s'}[V^*(s') - V_\pi(s')] \end{aligned} \tag{12}$$

By recursing on this equation and using linearity of expectation we get $V_\pi(s) \geq V^*(s) - \frac{2\varepsilon}{1-\gamma}$.

(c)

$$\begin{aligned} Q^*(s_1, go) &= \frac{2\varepsilon}{1-\gamma} \\ Q^*(s_1, stay) &= \frac{2\varepsilon\gamma}{1-\gamma} \\ V^*(s_1) &= \frac{2\varepsilon}{1-\gamma} \\ V^*(s_2) &= \frac{2\varepsilon\gamma}{1-\gamma} \end{aligned}$$

(d) As observed the difference between two state-value function is 2ε, so one can simply build a state-action value function $\tilde{Q}$ that makes $\pi(s_1) = stay$ the optimal action at s_1.
Let

$$\begin{aligned} \tilde{Q}(s_1, go) =& Q^*(s_1, go) - \varepsilon \\ \tilde{Q}(s_1, stay) =& Q^*(s_1, stay) + \varepsilon \\ V_\pi(s_1) - V^*(s_1) =& \frac{-2\varepsilon}{1-\gamma} \end{aligned}$$

So the bound is tight.

Problem 4

Solution

(a)

$$\begin{aligned} v(s) &= \mathrm{E}[G_t | S_t = s] \\ &= \mathrm{E}[R_{t+1} + \gamma R_{t+2} + ... | S_t = s] \\ &= \mathrm{E}[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + ...) | S_t = s] \\ &= \mathrm{E}[R_{t+1} + \gamma G_{t+1} | S_t = s] \\ &= \mathrm{E}[R_{t+1} + \gamma v(S_{t+1}) | S_t = s] \end{aligned} \tag{13}$$

Then we show $\mathrm{E}[R_{t+1} + \gamma G_{t+1} | S_t = s] = \mathrm{E}[R_{t+1} + \gamma v(S_{t+1}) | S_t = s]$.
By the definition of $v(s)$,

$$v(s) = \mathrm{E}[G_t | S_t = s], v(S_t) = \mathrm{E}[G_t | S_t], v(S_{t+1}) = \mathrm{E}[G_{t+1} | S_{t+1}]$$

By Adam's Law, we have

$$\mathrm{E}[\mathrm{E}[Y|X]] = \mathrm{E}[Y].$$

Adam's Law with extra conditioning,

$$\hat{\mathrm{E}}(\cdot) = \mathrm{E}(\cdot|Z).$$

$$\hat{\mathrm{E}}[\hat{\mathrm{E}}(Y|X)] = \hat{\mathrm{E}}(Y)$$

$$\mathrm{E}[\mathrm{E}(Y|X,Z)|Z] = \mathrm{E}[Y|Z]$$

$$\mathrm{E}[\mathrm{E}(G_{t+1}|S_{t+1},S_t)|S_t] = \mathrm{E}[E[G_{t+1}|S_{t+1}]|S_t] \quad \text{By Markov property} \tag{14}$$

$$\begin{aligned} \mathrm{E}[\mathrm{E}(G_{t+1}|S_{t+1},S_t)|S_t] &= \mathrm{E}[G_{t+1}|S_t] \\ &= \mathrm{E}[v(S_{t+1}|S_t)] \quad \text{By Adam's Law} \end{aligned} \tag{15}$$

thus

$$\mathrm{E}[G_{t+1}|S_t] = \mathrm{E}[v(S_{t+1})|S_t]\mathrm{E}[G_{t+1}|S_t=s] = \mathrm{E}[v(S_{t+1})|S_t=s] \tag{16}$$

$$\begin{aligned} v(s) = \mathrm{E}[G_t|S_t=s] = \mathrm{E}[R_{t+1}+\gamma G_{t+1}|S_t=s] &= \mathrm{E}[R_{t+1}|S_t=s] + \gamma \mathrm{E}[G_{t+1}|S_t=s] \\ = \mathrm{E}[R_{t+1}|S_t=s] + \gamma \mathrm{E}[v(S_{t+1})|S_t=s] \\ = E[R_{t+1}+\gamma v(S_{t+1})|S_t=s] \end{aligned} \tag{17}$$

(b)

$$v_\pi(s) = \mathrm{E}[R_{t+1}+\gamma v_\pi(S_{t+1})|S_t=s]$$

Equivalently,

$$v_\pi(S_t) = \mathrm{E}[R_{t+1}+\gamma v_\pi(S_{t+1})|S_t]$$

$$\begin{aligned} v_\pi(s) &= \mathrm{E}[G_t|S_t=s] \\ &= \mathrm{E}[R_{t+1}+\gamma R_{t+2}+...|S_t=s] \\ &= \mathrm{E}[R_{t+1}+\gamma(R_{t+2}+\gamma R_{t+3}+...)|S_t=s] \\ &= \mathrm{E}[R_{t+1}+\gamma G_{t+1}|S_t=s] \\ &= \mathrm{E}[R_{t+1}+\gamma v_\pi(S_{t+1})|S_t=s] \end{aligned} \tag{18}$$

Then we show $\mathrm{E}[R_{t+1}+\gamma G_{t+1}|S_t=s] = \mathrm{E}[R_{t+1}+\gamma v_\pi(S_{t+1})|S_t=s]$.
By the definition of $v_\pi(s)$,

$$v_\pi(s) = \mathrm{E}[G_t|S_t=s], v_\pi(S_t) = \mathrm{E}[G_t|S_t], v_\pi(S_{t+1}) = \mathrm{E}[G_{t+1}|S_{t+1}]$$

By Adam's Law, we have

$$\mathrm{E}[\mathrm{E}[Y|X]] = \mathrm{E}[Y].$$

Adam's Law with extra conditioning,

$$\hat{\mathrm{E}}(\cdot) = \mathrm{E}(\cdot|Z).$$

$$\hat{\mathrm{E}}[\hat{\mathrm{E}}(Y|X)] = \hat{\mathrm{E}}(Y)$$

$$\mathrm{E}[\mathrm{E}(Y|X,Z)|Z] = \mathrm{E}[Y|Z]$$

$$\mathrm{E}[\mathrm{E}(G_{t+1}|S_{t+1},S_t)|S_t] = \mathrm{E}[E[G_{t+1}|S_{t+1}]|S_t] \quad \text{By Markov property} \tag{19}$$

$$\begin{aligned} \mathrm{E}[\mathrm{E}(G_{t+1}|S_{t+1},S_t)|S_t] &= \mathrm{E}[G_{t+1}|S_t] \\ &= \mathrm{E}[v_\pi(S_{t+1}|S_t)] \quad \text{By Adam's Law} \end{aligned} \tag{20}$$

thus

$$\mathrm{E}[G_{t+1}|S_t] = \mathrm{E}[v_\pi(S_{t+1})|S_t]\mathrm{E}[G_{t+1}|S_t=s] = \mathrm{E}[v_\pi(S_{t+1})|S_t=s] \tag{21}$$

$$\begin{aligned} v_\pi(s) &= \mathrm{E}[G_t|S_t = s] = \mathrm{E}[R_{t+1} + \gamma G_{t+1}|S_t = s] = \mathrm{E}[R_{t+1}|S_t = s] + \gamma \mathrm{E}[G_{t+1}|S_t = s] \\ &= \mathrm{E}[R_{t+1}|S_t = s] + \gamma \mathrm{E}[v_\pi(S_{t+1})|S_t = s] \\ &= E[R_{t+1} + \gamma v_\pi(S_{t+1})|S_t = s] \end{aligned} \tag{22}$$

$$q_\pi(s,a) = \mathrm{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a]$$

$$\begin{aligned} q_\pi(s,a) &= \mathrm{E}_\pi[G_t|S_t = s, A_t = a] \\ q_\pi(S_t, A_t) &= \mathrm{E}_\pi[G_t|S_t, A_t] \\ q_\pi(S_{t+1}, A_{t+1}) &= \mathrm{E}_\pi[G_{t+1}|S_t, A_{t+1}] \end{aligned} \tag{23}$$

By Adam's Law, we have

$$\mathrm{E}[\mathrm{E}[Y|X]] = \mathrm{E}[Y].$$

Adam's Law with extra conditioning,

$$\hat{\mathrm{E}}(\cdot) = \mathrm{E}(\cdot|Z).$$

$$\hat{\mathrm{E}}[\hat{\mathrm{E}}(Y|X)] = \hat{\mathrm{E}}(Y)$$

$$\mathrm{E}[\mathrm{E}(Y|X,Z)|Z] = \mathrm{E}[Y|Z]$$

Let $Y = G_{t+1}$,$Z = (S_t, A_t), X = (S_{t+1}, A_{t+1})$ then we have

$$\begin{aligned} \mathrm{E}[\mathrm{E}(G_{t+1}|S_{t+1}, S_t, A_{t+1}, A_t)|S_t, A_t] &= \mathrm{E}[E[G_{t+1}|S_{t+1}, A_{t+1}]|S_t, A_t] \quad \text{By Markov property} \\ &= \mathrm{E}[q_\pi(S_{t+1}, A_{t+1})|S_t, A_t] \end{aligned} \tag{24}$$

$$\begin{aligned} \mathrm{E}[\mathrm{E}(G_{t+1}|S_{t+1}, S_t, A_{t+1}, A_t)|S_t, A_t] &= \mathrm{E}[G_{t+1}|S_t, A_t] \\ &= \mathrm{E}[q_\pi(S_{t+1}, A_{t+1})|S_t, A_t] \quad \text{By Adam's Law} \end{aligned} \tag{25}$$

thus

$$\mathrm{E}_\pi[G_{t+1}|S_t = s, A_t = a] = \mathrm{E}_\pi[q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a] \tag{26}$$

Then we have

$$\begin{aligned} q_\pi(s,a) &= \mathrm{E}_\pi[G_t|S_t = s, A_t = a] = \mathrm{E}_\pi[R_{t+1} + \gamma G_{t+1}|S_t = s, A_t = a] \\ &= \mathrm{E}_\pi[R_{t+1}|S_t = s, A_t = a] + \gamma \mathrm{E}_\pi[G_{t+1}|S_t = s, A_t = a] \\ &= \mathrm{E}_\pi[R_{t+1}|S_t = s, A_t = a] + \gamma \mathrm{E}_\pi[q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a] \\ &= \mathrm{E}_\pi[R_{t+1} + \gamma q_\pi(S_{t+1}, A_{t+1})|S_t = s, A_t = a] \end{aligned} \tag{27}$$

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) q_\pi(s,a)$$

Proof

$$\begin{aligned} v_\pi(s) = \mathrm{E}_\pi[G_t|S_t = s] &= \sum_{a \in \mathcal{A}} \mathrm{E}_\pi[G_t|S_t = s, A_t = a] P(A_t = a|S_t = s) \quad \text{(LOTE)} \\ &= \sum_{a \in \mathcal{A}} q_\pi(s,a) \pi(a|s) \end{aligned} \tag{28}$$

End proof

$$q_\pi(s,a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v_\pi(s')$$

Problem 4 continued on next page. . .

$$\begin{aligned}
\mathrm{E}[q_\pi(S_{t+1},A_{t+1})|S_{t+1}=s',S_t=s,A_t=a] &= \mathrm{E}[q_\pi(S_{t+1},A_{t+1})|S_{t+1}] \quad \text{(By Markov property)} \\
&= \sum_{a\in\mathcal{A}} \mathrm{E}[q_\pi(S_{t+1},A_{t+1})|S_{t+1}=s',A_{t+1}=a]P(A_{t+1}=a|S_{t+1}=s') \\
&= \sum_{a\in\mathcal{A}} q_\pi(s',a)\pi(a|s') \\
&= v_\pi(s')
\end{aligned} \tag{29}$$

Then we have

$$\begin{aligned}
\mathrm{E}[q_\pi(S_{t+1},A_{t+1})|S_t=s,A_t=a] &= \sum_{s'\in\mathcal{S}} \mathrm{E}[q_\pi(S_{t+1},A_{t+1})|S_{t+1}=s',S_t=s,A_t=a]P(S_{t+1}=s'|S_t=s,A_t=a) \quad \text{(LOTE)} \\
&= \sum_{s'\in\mathcal{S}} v_\pi(s')P^a_{ss'}
\end{aligned} \tag{30}$$

$$\begin{aligned}
q_\pi(s,a) &= \mathrm{E}_\pi[R_{t+1}+\gamma q_\pi(S_{t+1},A_{t+1})|S_t=s,A_t=a] \\
&= \mathrm{E}_\pi[R_{t+1}] + \gamma \mathrm{E}_\pi[q_\pi(S_{t+1},A_{t+1})|S_t=s,A_t=a] \\
&= R^a_s + \gamma \sum_{s'\in\mathcal{S}} P^a_{ss'} v_\pi(s')
\end{aligned} \tag{31}$$

$$\begin{aligned}
v_\pi(s) &= \sum_{a\in\mathcal{A}} \pi(a|s) q_\pi(s,a) \\
&= \sum_{a\in\mathcal{A}} \pi(a|s)(R^a_s + \gamma \sum_{s'\in\mathcal{S}} P^a_{ss'} v_\pi(s'))
\end{aligned} \tag{32}$$

$$\begin{aligned}
q_\pi(s,a) &= R^a_s + \gamma \sum_{s'\in S} P^a_{ss'} v_\pi(s') \\
&= R^a_s + \gamma \sum_{s'\in S} P^a_{ss'} (\sum_{a'\in\mathcal{A}} \pi(a'|s') q_\pi(s',a'))
\end{aligned} \tag{33}$$

(c)

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

In part (b), we have $q_\pi(s,a) = R^a_s + \gamma \sum_{s'\in\mathcal{S}} P^a_{ss'} v_\pi(s')$.

$$\begin{aligned}
q_*(s,a) &= \max_\pi q_\pi(s,a) = R^a_s + \gamma \sum_{s'\in\mathcal{S}} P^a_{ss'} \max_\pi v_\pi(s') \\
&= R^a_s + \gamma \sum_{s'\in\mathcal{S}} P^a_{ss'} v_*(s')
\end{aligned} \tag{34}$$

$$\mathrm{E}[R_{t+1}|S_t=s,A_t=a] = R^a_s$$

$$\begin{aligned}
\mathrm{E}[v_*(S_{t+1})|S_t=s,A_t=a] &= \sum_{s'\in\mathcal{S}} \mathrm{E}[v_*(S_{t+1})|S_{t+1}=s',S_t=s,A_t=a]P(S_{t+1}=s'|S_t=s,A_t=a) \\
&= \sum_{s'\in\mathcal{S}} v_*(s')P^a_{ss'}
\end{aligned} \tag{35}$$

$$q_*(s,a) = \mathrm{E}[R_{t+1}+\gamma v_*(S_{t+1})|S_t=s,A_t=a] \tag{36}$$

$$\begin{aligned}
v_*(s) &= \max_a q_*(s,a) \\
&= \max_a \mathrm{E}[R_{t+1}+\gamma v_*(S_{t+1})|S_t=s,A_t=a]
\end{aligned} \tag{37}$$

Then we also have

$$v_*(s) = \max_a (R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v_*(s'))$$

$$\begin{aligned} q_*(s,a) &= R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v_*(s') \\ &= R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a \max_{a'} q_*(s',a') \end{aligned} \tag{38}$$

$$\mathrm{E}[\max_{a'} q_*(S_{t+1},a')|S_t = s, A_t = a] = \sum_{s' \in \mathcal{S}} \mathrm{E}[\max_{a'} q_*(S_{t+1},a')|S_{t+1}, S_t = s, A_t = a] P(s_{t+1} = s'|S_t = s, A_t = a) \quad \text{(LOTE)}$$

$$= \sum_{s' \in \mathcal{S}} \max_{a'} q_*(s',a') P_{ss'}^a \tag{39}$$

Then

$$q_*(s,a) = \mathrm{E}[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1},a')|S_t = s, A_t = a]$$

Problem 5

Solution

(a)

$$v(s) = R_s + \gamma \sum_{s' \in \mathcal{S}} P_{ss'} v(s')$$

The $\gamma = 1$, then

$$\begin{aligned} v(pass) &= R_{pass} + \gamma \sum_{s' \in \mathcal{S}} P_{pass,s'} v(s') = 10 + 1 \cdot 0 = 10 \\ v(c_3) &= R_{c_3} + \gamma \sum_{s' \in \mathcal{S}} P_{c_3 s'} v(s') = -2 + 0.4 \cdot 0.8 + 0.6 \cdot 10 = 4.32 \\ v(c_2) &= R_{c_2} + \gamma \sum_{s' \in \mathcal{S}} P_{c_2 s'} v(s') = -2 + 0.2 \cdot 0 + 0.8 \cdot 4.32 = 1.456 \\ v(c_1) &= R_{c_1} + \gamma \sum_{s' \in \mathcal{S}} P_{c_1 s'} v(s') = -2 + 0.5 \cdot -22.543 + 0.5 \cdot 1.456 = -12.543 \\ v(pub) &= R_{pub} + \gamma \sum_{s' \in \mathcal{S}} P_{pub,s'} v(s') = 1 + 0.2 \cdot -12.543 + 0.4 \cdot 1.456 + 0.4 \cdot 4.32 = 0.802 \\ v(facebook) &= R_{facebook} + \gamma \sum_{s' \in \mathcal{S}} P_{facebook,s'} v(s') = -1 + 0.9 \cdot v(facebook) + 0.1 \cdot -12.543 = -22.543 \\ v(sleep) &= 0 + \gamma 0 = 0 \end{aligned} \tag{40}$$

We can solve the Bellman equation, then we obtain results, iterative method or solve directly.

$$\begin{aligned} v &= R + \gamma P v \\ v &= (I - \gamma P)^{-1} R \end{aligned}$$

(b) Let policy π is uniform random and discount factor $\gamma = 1$.

$$v_\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s)[R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v_\pi(s')]$$

$$\begin{aligned} v_1 = v_\pi(s_1) &= \pi(study|s_1)(R_{s_1}^{study} + 1 \cdot 1 \cdot v_\pi(s_2)) + \pi(fb|s_1)(R_{s_1}^{fb} + 1 \cdot 1 \cdot v_\pi(s_4)) \\ &= 0.5(-2 + v_2) + 0.5(-1 + v_4) \end{aligned} \tag{41}$$

$$\begin{aligned} v_2 = v_\pi(s_2) &= 0.5(-2 + v_3) + 0.5(0 + 0) \\ v_3 = v_\pi(s_3) &= 0.5(1 + 0.2v_1 + 0.4v_2 + 0.4v_3) + 0.5(10 + 0) \\ v_4 = v_\pi(s_4) &= 0.5(0 + v_1) + 0.5(-1 + v_4) \end{aligned} \tag{42}$$

So we have

$$v_1 = -1.3, v_2 = 2.7, v_3 = 7.4, v_4 = -2.3.$$

$$q_\pi(s, a) = R_s^a + \gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v_\pi(s')$$

$$\begin{aligned} q_\pi(s_1, study) &= -2 + 1 \cdot v_2 = 0.7 \\ q_\pi(s_1, fb) &= -1 + 1 \cdot v_4 = -3.3 \\ q_\pi(s_2, sleep) &= 0 + 0 = 0 \\ q_\pi(s_2, study) &= -2 + 1 \cdot v_3 = 5.4 \\ q_\pi(s_3, study) &= 10 + 0 = 10 \\ q_\pi(s_3, pub) &= 1 + 0.2v_1 + 0.4v_2 + 0.4v_3 = 4.78 \\ q_\pi(s_4, fb) &= -1 + 1 \cdot v_4 = -3.3 \\ q_\pi(s_4, quit) &= 0 + 1 \cdot v_1 = -1.3 \end{aligned} \tag{43}$$

(c) We obtain $q_*(s, a)$ first, then obtain $v_*(s)$.

$$q_\pi(s, a) = R_s^a + +\gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v_\pi(s')$$

$$q_*(s, a) = \max_\pi q_\pi(s, a)$$

Since $v_\pi(sleep) = 0, \forall \pi$, then we have

$$q_\pi(s_2, sleep) = R_{s_2}^{sleep} + 1 \cdot v_\pi(sleep) = 0 + 0, \forall \pi.$$

Then $q_*(s_2, sleep) = 0$.

$$q_\pi(s_3, study) = R_{s_3}^{sleep} + 1 \cdot v_\pi(sleep) = 10 + 0 = 10, \forall \pi \tag{44}$$

Then $q_*(s_3, study) = 10$. We also have

$$q_*(s, a) = R_s^a + +\gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a v_*(s') = R_s^a + +\gamma \sum_{s' \in \mathcal{S}} P_{ss'}^a \max_{a'} q_*(s', a').$$

$$\begin{aligned} q_*(s_1, study) &= R_{s_1}^{study} + \gamma P_{s_1 s_2}^{study} \max_{a'} q_*(s_2, a') \\ &= R_{s_1}^{study} + \gamma P_{s_1 s_2}^{study} \max\{q_*(s_2, sleep), q_*(s_2, study)\} \\ &= -2 + \max\{0, q_*(s_2, study)\} \end{aligned} \tag{45}$$

$$\begin{aligned} q_*(s_1, facebook) &= R_{s_1}^{facebook} + \gamma P_{s_1 s_4}^{facebook} \max_{a'} q_*(s_4, a') \\ &= R_{s_1}^{facebook} + \gamma P_{s_1 s_4}^{facebook} \max\{q_*(s_4, quit), q_*(s_4, facebook)\} \\ &= -1 + \max\{q_*(s_4, quit), q_*(s_4, facebook)\} \end{aligned} \tag{46}$$

$$\begin{aligned} q_*(s_2, sleep) &= 0 \\ q_*(s_2, study) &= -2 + \max\{10, q_*(s_3, pub)\} q_*(s_3, study) \quad = 10 \end{aligned}$$

$$q_*(s_3, pub) = 1 + 0.2\max\{q_*(s_1, study), q_*(s_1, facebook)\} + 0.4\max\{q_*(s_2, study), 0\} + 0.4\max\{q_*(s_3, pub), 10\} \tag{47}$$

$$q_*(s_4, facebook) = -1 + \max\{q_*(s_4, facebook), q_*(s_4, quit)\} \tag{48}$$

$$q_*(s_4, quit) = 0 + \max\{q_*(s_1, facebook), q_*(s_1, study)\} \tag{49}$$

$$\begin{aligned}
q_*(s_2, study) &= -2 + \max\{10, q_*(s_3, pub)\} \\
q_*(s_1, study) &= -2 + \max\{0, q_*(s_2, study)\} = -2 + q_*(s_2, study) \\
q_*(s_4, facebook) &= -1 + \max\{q_*(s_4, facebook), q_*(s_4, quit)\} = -1 + q_*(s_4, quit) \\
q_*(s_1, facebook) &= -1 + \max\{q_*(s_4, facebook), q_*(s_4, quit)\} = -1 + q_*(s_4, quit) \\
q_*(s_4, quit) &= \max\{q_*(s_1, study), q_*(s_1, facebook)\} = \max\{q_*(s_1, study), -1 + q_*(s_4, quit)\} = q_*(s_1, study) \\
q_*(s_3, pub) &= 1 + 0.2\max\{q_*(s_1, study), q_*(s_1, facebook)\} + 0.4\max\{q_*(s_2, study), 0\} + 0.4\max\{q_*(s_3, pub), 10\} \\
&= 0.6 + 0.6q_*(s_2, study) + 0.4\max\{q_*(s_3, pub), 10\}
\end{aligned} \tag{50}$$

If $\max\{q_*(s_3, pub), 10\} = q_*(s_3, pub)$, then $q_*(s_3, pub) \geq 10$.

$$\begin{aligned}
q_*(s_3, pub) &= 0.6 + 0.6q_*(s_2, study) + 0.4q_*(s_3, pub) \\
q_*(s_3, pub) &= 1 + q_*(s_2, study)
\end{aligned}$$

$$q_*(s_2, study) = -2 + \max\{10, q_*(s_3, pub)\} = -2 + q_*(s_3, pub)$$

Then

$$q_*(s_3, pub) = 1 - 2 + q_*(s_3, pub)$$

, which means $q_*(s_3, pub) < 10$. Then $q_*(s_3, pub) = 0.6 + 0.6q_*(s_2, study) + 4 = 4.6 + 0.6q_*(s_2, study)$.

$$q_*(s_2, study) = -2 + \max\{10, q_*(s_3, pub)\} = -2 + 10 = 8$$

Then we have

$$\begin{aligned}
q_*(s_1, study) &= -2 + q_*(s_2, study) = -2 + 8 = 6 \\
q_*(s_3, pub) &= 4.6 + 0.6q_*(s_2, study) = 4.6 + 0.6 * 8 = 9.4 \\
q_*(s_4, quit) &= q_*(s_1, study) = 6 \\
q_*(s_4, facebook) &= -1 + q_*(s_4, quit) = -1 + 6 = 5 \\
q_*(s_1, facebook) &= -1 + q_*(s_4, quit) = -1 + 6 = 5
\end{aligned} \tag{51}$$

Since

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

$$\begin{aligned}
v_*(s_1) &= \max_a q_*(s_1, a) = \max\{q_*(s_1, study), q_*(s_1, facebook)\} = \max\{6, 5\} = 6 \\
v_*(s_2) &= \max_a q_*(s_2, a) = \max\{q_*(s_2, study), q_*(s_2, sleep)\} = \max\{8, 0\} = 8 \\
v_*(s_3) &= \max_a q_*(s_3, a) = \max\{q_*(s_3, study), q_*(s_3, pub)\} = \max\{10, 9.4\} = 10 \\
v_*(s_4) &= \max_a q_*(s_4, a) = \max\{q_*(s_4, quit), q_*(s_4, facebook)\} = \max\{6, 5\} = 6 \\
v_*(sleep) &= 0
\end{aligned} \tag{52}$$

Problem 6

Solution

(a)

$$v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s)[R_s^a + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a v_\pi(s')]$$

By description in Reinforcement Learning:introduction, $\gamma = 0.9$. For example, the value function in first row is

$$\begin{aligned}
&3.3 = 1/4(-1 + 0.9 * 1 * 3.3) + 1/4(0 + 0.9 * 1 * 8.8) + 1/4(0 + 0.9 * 1 * 1.5) + 1/4(-1 + 0.9 * 1 * 3.3) \\
&8.8 = 1 * (10 + 0.9 * 1 * -1.3) \\
&4.4 = 1/4(-1 + 0.9 * 1 * 4.4) + 1/4(0 + 0.9 * 1 * 5.3) + 1/4(0 + 0.9 * 1 * 2.3) + 1/4(0 + 0.9 * 1 * 8.8) \\
&5.3 = 1 * (5 + 0.9 * 1 * 0.4) \\
&1.5 = 1/4(-1 + 0.9 * 1 * 1.5) + 1/4(-1 + 0.9 * 1 * 1.5) + 1/4(0 + 0.9 * 1 * 0.5) + 1/4(0 + 0.9 * 1 * 5.3)
\end{aligned} \tag{53}$$

We can solve the Bellman Expectation equation.

$$v_\pi = R^\pi - \gamma P^\pi v_\pi$$

$$v_\pi = (I - \gamma P^\pi)^{-1} R^\pi$$

(b)

$$q_*(s,a) = R_s^a + +\gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a v_*(s') = R_s^a + +\gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a \max_{a'} q_*(s', a')$$

$$v_*(s) = \max_a (R_s^a + \gamma \sum_{s'\in\mathcal{S}} P_{ss'}^a v_*(s'))$$

Let location (i,j) in gridworld is state $s_{(i-1)*5+j}$, then we should have 25 states.

$$\begin{aligned}
v_*(s_1) &= \max_a (R_{s_1}^a + \gamma \sum_{s'\in\mathcal{S}} P_{s_1 s'}^a v_*(s')) \\
&= \max\{-1 + 0.9 * v_*(s_1), 0 + 0.9 * v_*(s_2), 0 + 0.9 * v_*(s_6), -1 + 0.9 * v_*(s_1)\} \\
v_*(s_2) &= \max\{10 + 0.9 * v_*(s_{22}), 10 + 0.9 * v_*(s_{22}), 10 + 0.9 * v_*(s_{22}), 10 + 0.9 * v_*(s_{22})\} \\
v_*(s_3) &= \max\{-1 + 0.9 * v_*(s_3), 0 + 0.9 * v_*(s_4), 0 + 0.9 * v_*(s_8), 0 + 0.9 * v_*(s_2)\} \\
&\vdots \\
v_*(s_t) &= \max\{0.9 * v_*(s_{t-5}), 0.9 * v_*(s_{t+1}), 0.9 * v_*(s_{t+5}), 0.9 * v_*(s_{t-1})\} \\
&\vdots \\
v_*(s_{25}) &= \max\{0.9 * v_*(s_{20}), -1 + 0.9 * v_*(s_{25}), -1 + 0.9 * v_*(s_{25}), 0.9 * v_*(s_{24})\}
\end{aligned} \tag{54}$$

For example,

$$\begin{aligned}
v_*(s_1) &= \max\{-1 + 0.9 * v_*(s_1), 0 + 0.9 * v_*(s_2), 0 + 0.9 * v_*(s_6), -1 + 0.9 * v_*(s_1)\} \\
&= \max\{-1 + 0.9 * 22, 0.9 * 24.4, 0.9 * 19.8, -1 + 0.9 * 22\} = 22
\end{aligned} \tag{55}$$

And we also have

$$\begin{aligned} q_*(s_1, up) &= R_{s_1}^{left} + \gamma \sum_{s' \in \mathcal{S}} P_{s_1 s'}^{left} v_*(s') = -1 + 0.9 * v_*(s_1) \\ q_*(s_1, right) &= R_{s_1}^{right} + \gamma \sum_{s' \in \mathcal{S}} P_{s_1 s'}^{right} v_*(s') = 0.9 * v_*(s_2) \\ q_*(s_1, down) &= R_{s_1}^{down} + \gamma \sum_{s' \in \mathcal{S}} P_{s_1 s'}^{down} v_*(s') = 0.9 * v_*(s_6) \\ q_*(s_1, left) &= R_{s_1}^{left} + \gamma \sum_{s' \in \mathcal{S}} P_{s_1 s'}^{left} v_*(s') = -1 + 0.9 * v_*(s_1) \end{aligned} \tag{56}$$

Then

$$\begin{aligned} q_*(s_1, up) = -1 + 0.9 * v_*(s_1) = 18.8 \\ q_*(s_1, right) = 0.9 * v_*(s_2) = 22 \\ q_*(s_1, down) = 0.9 * v_*(s_6) = 17.82 \\ q_*(s_1, left) = -1 + 0.9 * v_*(s_1) = 18.8 \end{aligned} \tag{57}$$

Since

$$a^* = \arg\max_{a'} q_*(s, a')$$

, then a^* in s_1 is $\arg\max_{a'}\{18.8, 22, 17.82, 18.8\} = right$. optimal policy is

$$\pi(a|s_1) = \begin{cases} 1 & \text{if } a = right \\ 0 & \text{otherwise} \end{cases}$$

. I find that if I want to obtain $v_*(s)$, I have to get $q_*(s, a)$, i.e. I have to solve optimal Bellman equation recursively.
To obtain optimal value function, we can use iterative solution methods, such as Value Iteration, Policy Iteration, Q-learning and Sarsa.

Problem 7

Solution

Problem 8

Solution

Problem 9

Solution