Stanford CME 241 (Winter 2021) - Assignment 4

Assignments:

1. **Manual Value Iteration:** Consider a simple MDP with $\mathcal{S} = \{s_1, s_2, s_3\}, \mathcal{T} = \{s_3\}, \mathcal{A} = \{a_1, a_2\}$. The State Transition Probability function

$$\mathcal{P} : \mathcal{N} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$$

is defined as:

$$\mathcal{P}(s_1, a_1, s_1) = 0.2, \mathcal{P}(s_1, a_1, s_2) = 0.6, \mathcal{P}(s_1, a_1, s_3) = 0.2$$
$$\mathcal{P}(s_1, a_2, s_1) = 0.1, \mathcal{P}(s_1, a_2, s_2) = 0.2, \mathcal{P}(s_1, a_2, s_3) = 0.7$$
$$\mathcal{P}(s_2, a_1, s_1) = 0.3, \mathcal{P}(s_2, a_1, s_2) = 0.3, \mathcal{P}(s_2, a_1, s_3) = 0.4$$
$$\mathcal{P}(s_2, a_2, s_1) = 0.5, \mathcal{P}(s_2, a_2, s_2) = 0.3, \mathcal{P}(s_2, a_2, s_3) = 0.2$$

The Reward Function

$$\mathcal{R} : \mathcal{N} \times \mathcal{A} \to \mathbb{R}$$

is defined as:

$$\mathcal{R}(s_1, a_1) = 8.0, \mathcal{R}(s_1, a_2) = 10.0$$
$$\mathcal{R}(s_2, a_1) = 1.0, \mathcal{R}(s_2, a_2) = -1.0$$

Assume discount factor $\gamma = 1$.

Your task is to determine an Optimal Deterministic Policy *by manually working out* (not with code) simply the first two iterations of Value Iteration algorithm.

- Initialize the Value Function for each state to be it's max (over actions) reward, i.e., we initialize the Value Function to be $v_0(s_1) = 10.0, v_0(s_2) = 1.0, v_0(s_3) = 0.0$. Then manually calculate $q_k(\cdot, \cdot)$ and $v_k(\cdot)$ from $v_{k-1}(\cdot)$ using the Value Iteration update, and then calculate the greedy policy $\pi_k(\cdot)$ from $q_k(\cdot, \cdot)$ for $k = 1$ and $k = 2$ (hence, 2 iterations). $A = \{a_1, a_2\}$

$V_0(S_1) = 10$

$V_0(S_2) = 1$

$V_0(S_3) = 0$

$k=1$ With discount factor of $\lambda = 1$,

$$\begin{cases} q_1(S_1, a_1) = R(S_1, a_1) + 0.2 \cdot 10 + 0.6 \cdot 1 + 0.2 \cdot 0 \\ \quad = 8 + 2 + 0.6 = 10.6 \\ q_1(S_1, a_2) = R(S_1, a_2) + 0.1 \cdot 10 + 0.2 \cdot 1 + 0.7 \cdot 0 \\ \quad = 10 + 1 + 0.2 = 11.2 \end{cases}$$

$$\begin{cases} q_1(S_2, a_1) = R(S_2, a_1) + 0.3 \cdot 10 + 0.3 \cdot 1 + 0.4 \cdot 0 \\ \quad = 1 + 3 + 0.3 = 4.3 \\ q_1(S_2, a_2) = R(S_2, a_2) + 0.5 \cdot 10 + 0.3 \cdot 1 + 0.2 \cdot 0 \\ \quad = -1 + 5 + 0.3 = 4.3 \end{cases}$$

$\Rightarrow \quad V_1(s_1) = 11.2$

$V_1(s_2) = 4.3$

$\Rightarrow \quad \pi_1(s_1) = a_1$

$\pi_1(s_2) = \{a_1, a_2\}$

$k=2$

$q_2(s_1, a_1) = 8 + 0.2 \cdot 11.2 + 0.6 \cdot 4.3 = 12.82$

$q_2(s_1, a_2) = 10 + 0.1 \cdot 11.2 + 0.2 \cdot 4.3 = 11.98$

$q_2(s_2, a_1) = 1 + 0.3 \cdot 11.2 + 0.3 \cdot 4.3 = 5.65$

$q_2(s_2, a_2) = -1 + 0.5 \cdot 11.2 + 0.3 \cdot 4.3 = 5.89$

$\Rightarrow \quad V_2(s_1) = 12.82$

$V_2(s_2) = 5.89$

$\Rightarrow \quad \pi_2(s_1) = a_1$

$\pi_2(s_2) = a_2$

- Now argue that $\pi_k(\cdot)$ for $k > 2$ will be the same as $\pi_2(\cdot)$. Hint: You can make the argument by examining the structure of how you get $q_k(\cdot, \cdot)$ from $v_{k-1}(\cdot)$. With this argument, there is no need to go beyond the two iterations you performed above, and so you can establish $\pi_2(\cdot)$ as an Optimal Deterministic Policy for this MDP.

↖ $k=2$

$A = \{a_1, a_2\}$

For $k>2$, need to examine the magnitude between $q_k(s_1, a_1)$, $q_k(s_1, a_2)$ and between $q_k(s_2, a_1)$, $q_k(s_2, a_2)$

Given that $R(s_1, a_1) = 8$ and $R(s_1, a_2) = 10$,

and that $P(s_1, a_1, s_1) = 0.2$ $\quad P(s_1, a_2, s_1) = 0.1$

$P(s_1, a_1, s_2) = 0.6$ $\quad P(s_1, a_2, s_2) = 0.2$

$\boxed{\text{WTS : At } s_1,\ \pi(s_1) = a_1}$

$$q_k(s_1, a_1) - q_k(s_1, a_2)$$

$$= (8 - 10) + (0.2 - 0.1) V_{k-1}(s_1) + (0.6 - 0.2) V_{k-1}(s_2)$$

$$= -2 + 0.1 V_{k-1}(s_1) + 0.4 V_{k-1}(s_2)$$

—— ①

want > 0

Also given that $R(s_2, a_1) = 1$ and $R(s_2, a_2) = -1$,

and that $P(s_2, a_1, s_1) = 0.3$ $\quad P(s_2, a_2, s_1) = 0.5$

$P(s_2, a_1, s_2) = 0.3$ $\quad P(s_2, a_2, s_2) = 0.3$

$\boxed{\text{WTS : At } s_2,\ \pi(s_2) = a_2}$

$$q_k(s_2, a_2) - q_k(s_2, a_1)$$

$$= (-1 - 1) + (0.5 - 0.3) V_{k-1}(s_1) + (0.3 - 0.3) V_{k-1}(s_2)$$

$$= -2 + 0.2 V_{k-1}(s_1)$$

—— ②

want > 0

Combining ①, ②

② > 0 when $V_{k-1}(s_1) > 10$

then $\Rightarrow$ ① > 0 when $V_{k-1}(s_2) > 2.5$

s.t. $\pi(s_1) = a_1$

$\pi(s_2) = a_2$

$\because \quad V_1(s_1) = 11.2 \quad > 10$

$\quad\quad V_1(s_2) = 4.3 \quad > 2.5$

when $k = 1$

$\therefore$ By def of Value Iteration, we could establish $\pi_2(\cdot)$ as the Optimal Deterministic Policy for this MDP.

There's no need to go beyond $K = 2$.

Hence, it holds that $\forall k > 2$

$$\pi^*(s_1) = a_1$$

$$\pi^*(s_2) = a_2$$

#

3. **Job-Hopping and Wages-Utility-Maximization:** You are a worker who starts every day either employed or unemployed. If you start your day employed, you work on your job for the day (one of n jobs, as elaborated later) and you get to earn the wage of the job for the day. However, at the end of the day, you could lose your job with probability $\alpha \in [0,1]$, in which case you start the next day unemployed. If at the end of the day, you do not lose your job (with probability $1-\alpha$), then you will start the next day with the same job (and hence, the same daily wage). On the other hand, if you start your day unemployed, then you will be randomly offered one of n jobs with daily wages $w_1, w_2, \ldots w_n \in \mathbb{R}^+$ with respective job-offer probabilities $p_1, p_2, \ldots p_n \in [0,1]$ (with $\sum_{i=1}^{n} p_i = 1$). You can choose to either accept or decline the offered job. If you accept the job-offer, your day progresses exactly like the *employed-day* described above (earning the day's job wage and possibly (with probability α) losing the job at the end of the day). However, if you decline the job-offer, you spend the day unemployed, receive the unemployment wage $w_0 \in \mathbb{R}^+$ for the day, and start the next day unemployed. The problem is to identify the optimal choice of accepting or rejecting any of the job-offers the worker receives, in a manner that maximizes the infinite-horizon *Expected Discounted-Sum of Wages Utility.* Assume the daily discount factor for wages (employed or unemployed) is $\gamma \in [0,1)$. Assume Wages Utility function to be $U(w) = \log(w)$ for any wage amount $w \in \mathbb{R}^+$. So you are looking to maximize

$$\mathbb{E}[\sum_{u=t}^{\infty} \gamma^{u-t} \cdot \log(w_{i_u})]$$

at the start of a given day t (w_{i_u} is the wage earned on day u, $0 \leq i_u \leq n$ for all $u \geq t$).

- Express with clear mathematical notation the state space, action space, transition function, reward function, and write the Bellman Optimality Equation customized for this MDP.

∘ $S = \{ s \mid s = \underbrace{1, 2, \ldots, n}_{\text{job offered}}, \underbrace{n+1}_{\text{state of unemployment}} \}$

∘ $A = \{ \underbrace{a}_{\text{accept}}, \underbrace{r}_{\text{reject}} \}$

a new job

∘ $P(s,a,s') = \mathbb{P}[S_{t+1} = s' \mid S_t = s, A = a]$

$P(n+1, a, n+1) = 0$

i.e. cannot accept an offer but still stay unemployed

∘ $R(s,a)$

$R(n+1, r) = \log(W_0)$ where $W_0 \in \mathbb{R}^+$ is the unemployment wage

Bellman $V(s) = \max \{ R(s,a) + \gamma \sum_{s' \in S} P(s,a,s') V(s') \}$

$\gamma \in [0,1)$