CS 6300 **HW04: Utility and Value Iteration** **Ryan Dalby** **Due February 24, 2022**

1 Value Iteration

You decide to go to Las Vegas for spring break, to take in some shows and play a little blackjack. Casino hotels typically offer very cheap buffets, and so you have two possible actions: Eat Buffet or Play Blackjack. You start out Poor and Hungry, and would like to leave the casino Rich and Full. If you Play while you are Full you are more likely to become Rich, but if you are Poor you may have a hard time becoming Full on your budget. We can model your decision making process as the following MDP:

State Space	{PoorHungry, PoorFull, RichHungry, RichFull}
Actions	{Eat, Play}
Initial State	PoorHungry
Terminal State	RichFull

s	a	s'	$T(s,a,s')$
PoorHungry	Play	PoorHungry	0.8
PoorHungry	Play	RichHungry	0.2
PoorHungry	Eat	PoorHungry	0.8
PoorHungry	Eat	PoorFull	0.2
PoorFull	Play	PoorFull	0.5
PoorFull	Play	RichFull	0.5
RichHungry	Eat	RichHungry	0.2
RichHungry	Eat	RichFull	0.8

Transition Model

s'	$R(s')$
PoorHungry	-1
PoorFull	1
RichHungry	0
RichFull	5

Rewards

1. Perform 3 iterations of Value Iteration. Fill out tables of both the Q-values and the Values. Assume $\gamma = 1$.

   Note that all Values or Q-values whose state or state-value parameters are not specified in the table are 0. Also note the following notation shorthand:

State Space	{PoorHungry, PoorFull, RichHungry, RichFull} = {PH, PF, RH, RF}
Actions	{Eat, Play} = {E, P}
Initial State	PoorHungry = PH
Terminal State	RichFull = RF

i	$Q_i(PH, P)$	$Q_i(PH, E)$	$Q_i(PF, P)$	$Q_i(RH, E)$
0	0	0	0	0
1	-0.8	-0.6	3.0	4.0
2	-0.48	-0.48	4.5	4.8
3	-0.224	-0.084	5.25	4.96

i	$V_i(PH)$	$V_i(PF)$	$V_i(RH)$
0	0	0	0
1	-0.6	3.0	4.0
2	-0.48	4.5	4.8
3	-0.084	5.25	4.96

Using (assuming $\gamma = 1$):

$$\begin{aligned} Q^*_{i+1}(s,a) &= \sum_{s'} T(s,a,s')[R(s,a,s') + V^*_i(s')] \\ V^*_{i+1}(s) &= \max_{a_i} Q^*_{i+1}(s,a) \end{aligned}$$

$$\begin{aligned} Q_1(PH,P) &= 0.8[-1+0] + 0.2[0+0] = -0.8 \\ Q_1(PH,E) &= 0.8[-1+0] + 0.2[1+0] = -0.6 \\ Q_1(PF,P) &= 0.5[1+0] + 0.5[5+0] = 3.0 \\ Q_1(RH,E) &= 0.2[0+0] + 0.8[5+0] = 4.0 \end{aligned}$$

$$\begin{aligned} V_1(PH) &= -0.6 \\ V_1(PF) &= 3.0 \\ V_1(RH) &= 4.0 \end{aligned}$$

$$\begin{aligned} Q_2(PH,P) &= 0.8[-1+-0.6] + 0.2[0+4.0] = -0.48 \\ Q_2(PH,E) &= 0.8[-1+-0.6] + 0.2[1+3.0] = -0.48 \\ Q_2(PF,P) &= 0.5[1+3.0] + 0.5[5+0] = 4.5 \\ Q_2(RH,E) &= 0.2[0+4.0] + 0.8[5+0] = 4.8 \end{aligned}$$

$$
\begin{aligned}
V_2(PH) &= -0.48 \\
V_2(PF) &= 4.5 \\
V_2(RH) &= 4.8
\end{aligned}
$$

$$
\begin{aligned}
Q_3(PH,P) &= 0.8[-1+-0.48]+0.2[0+4.8]=-0.224 \\
Q_3(PH,E) &= 0.8[-1+-0.48]+0.2[1+4.5]=-0.084 \\
Q_3(PF,P) &= 0.5[1+4.5]+0.5[5+0]=5.25 \\
Q_3(RH,E) &= 0.2[0+4.8]+0.8[5+0]=4.96
\end{aligned}
$$

$$
\begin{aligned}
V_3(PH) &= -0.084 \\
V_3(PF) &= 5.25 \\
V_3(RH) &= 4.96
\end{aligned}
$$

2. Assuming that we are acting for three time steps, what is the optimal action to take from the starting state? Justify your answer.

   Using (assuming $\gamma = 1$):

$$
\pi_i^*(PH) = \text{arg}\max_a \sum_{s'} T(PH,a,s')[R(PH,a,s') + V_i^*(s')]
$$

   The optimal policy from the start state given three time steps is Eat (E).

   This is because:

$$
\begin{aligned}
\pi_3^*(PH) &= \text{arg}\max_{a\in\{P,E\}} \begin{cases} 0.8[-1+0.084]+0.2[0+4.96]=0.1248(P) \\ 0.8[-1+0.084]+0.2[0+5.25]=0.1828(E) \end{cases} \\
\pi_3^*(PH) &= E
\end{aligned}
$$

2 Policy Iteration (30pts)

You didn't do so well playing blackjack, so you decide to play the card game of high-low. High-low is played with an infinite deck whose only cards are 2, 3, and 4 in equal proportion. You start with one of the cards showing, and say either *high* or *low*. Then a new card is flipped, and you compare the value of the new card to that of the old card.

- If you are right, you get the value of the new card.
- If the new card has the same value, you don't get any points.
- If you are wrong, the game is done.

If you are not done, the new card then becomes the reference card for drawing the next new card. You accumulate points as above until you are wrong and the game ends.

1. Formulate high-low as an MDP, by listing the states, actions, transition rewards, and transition probabilities.

State Space	{2, 3, 4, F}
Actions	{High, Low} = {H, L}
Initial State	$s_0 \in \{2, 3, 4\}$
Terminal State	F

s	a	s'	$T(s,a,s')$	$R(s,a,s')$
2	H	F	$1/3$	0
2	H	3	$1/3$	3
2	H	4	$1/3$	4
2	L	F	1	0
3	H	F	$2/3$	0
3	H	4	$1/3$	4
3	L	2	$1/3$	2
3	L	F	$2/3$	0
4	H	F	1	0
4	L	2	$1/3$	2
4	L	3	$1/3$	3
4	L	F	$1/3$	0

Transition Model and Rewards

2. You will be doing one iteration of policy iteration. Assume the initial policy $\pi_0(s) = high$.

   Note: Since not specified, assuming $\gamma = 1$.

(a) Perform policy evaluation to solve for the utility values $V^{\pi_0}(s)$ for the appropriate states s. Please solve these equations analytically.

Using (assuming $\gamma = 1$):

$$V^{\pi_0}(s) \;=\; \sum_{s'} T(s,\pi_0(s),s')[R(s,\pi_0(s),s') + V^{\pi_0}(s')] = \sum_{s'} T(s,H,s')[R(s,H,s') + V^{\pi_0}(s')]$$

We get equations (Note since F is a terminal state $V^{\pi_0}(F) = 0$):

$$\begin{aligned}
V^{\pi_0}(2) &= \frac{1}{3}[0 + V^{\pi_0}(F)] + \frac{1}{3}[3 + V^{\pi_0}(3)] + \frac{1}{3}[4 + V^{\pi_0}(4)] \\
V^{\pi_0}(3) &= \frac{2}{3}[0 + V^{\pi_0}(F)] + \frac{1}{3}[4 + V^{\pi_0}(4)] \\
V^{\pi_0}(4) &= 1[0 + V^{\pi_0}(F)] \\
V^{\pi_0}(F) &= 0
\end{aligned}$$

Thus algebraically solving for each value gives:

$$\begin{aligned}
V^{\pi_0}(2) &= \frac{25}{9} \\
V^{\pi_0}(3) &= \frac{4}{3} \\
V^{\pi_0}(4) &= 0 \\
V^{\pi_0}(F) &= 0
\end{aligned}$$

(b) Perform policy improvement to find the next policy $\pi_1(s)$.

Using (assuming $\gamma = 1$):

$$\pi_1(s) \;=\; \text{arg}\max_a \sum_{s'} T(s,a,s')[R(s,a,s') + V^{\pi_0}(s')]$$

$$\begin{aligned}
\pi_1(2) &= \text{arg}\max_{a\in\{H,L\}} \begin{cases} \frac{1}{3}[0+0] + \frac{1}{3}[3+\frac{4}{3}] + \frac{1}{3}[4+0] = \frac{25}{9} \approx 2.78(H) \\ 1[0+0] = 0(L) \end{cases} \\
\pi_1(3) &= \text{arg}\max_{a\in\{H,L\}} \begin{cases} \frac{2}{3}[0+0] + \frac{1}{3}[4+0] = \frac{4}{3} \approx 1.33(H) \\ \frac{1}{3}[2+\frac{25}{9}] + \frac{2}{3}[0+0] = \frac{43}{27} \approx 1.59(L) \end{cases} \\
\pi_1(4) &= \text{arg}\max_{a\in\{H,L\}} \begin{cases} 1[0+0] = 0(H) \\ \frac{1}{3}[2+\frac{25}{9}] + \frac{1}{3}[3+\frac{4}{3}] + \frac{1}{3}[0+0] = \frac{82}{27} \approx 3.04(L) \end{cases}
\end{aligned}$$

$$\begin{aligned}\pi_1(2) &= H\\ \pi_1(3) &= L\\ \pi_1(4) &= L\end{aligned}$$