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 and 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 $\gamma = 1$.

State	$i=0$	$i=1$	$i=2$	$i=3$
PoorHungry	0	-0.6	-0.48	-0.084
PoorFull	0	3	4.5	5.25
RichHungry	0	4	4.8	4.96
RichFull	0	0	0	0

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

   Looking at the utilities after 3 iterations, the state 'PoorFull' has the maximum utility. Hence, the optimal action is 'Eat'.

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.

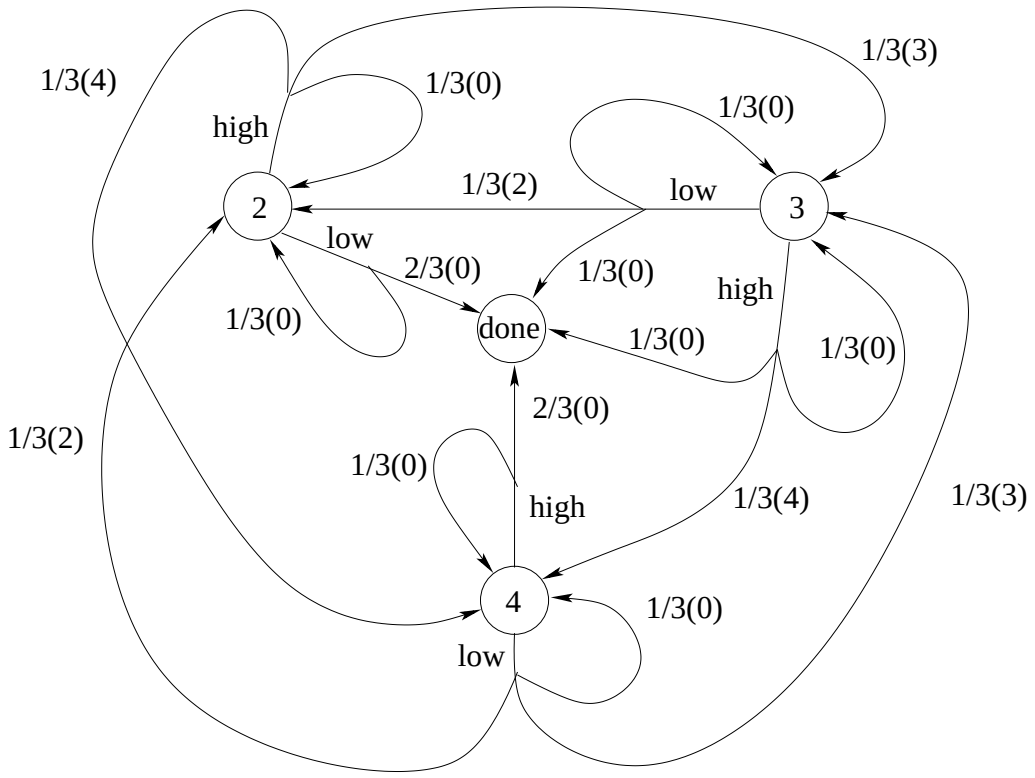
States: 2, 3, 4, done.

Actions: high, low.

Transition rewards and probabilities:

s	a	s'	$R(s,a,s')$	$T(s,a,s')$
2	low	2	0	1/3
2	low	done	0	2/3
2	high	2	0	1/3
2	high	3	3	1/3
a 2	high	4	4	1/3
3	low	2	2	1/3
3	low	3	0	1/3
3	low	done	0	1/3
3	high	done	0	1/3
3	high	3	0	1/3
3	high	4	4	1/3
4	low	2	2	1/3
4	low	3	3	1/3
4	low	4	0	1/3
4	high	done	0	2/3
4	high	4	0	1/3

The MDP is drawn below.



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

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

   For this MDP, the discount factor is $\gamma = 1$. Since the rewards differ, we have to use the general form of the analytical equation for policy evaluation:

$$V^{\pi_i}(s) = \sum_{s'} T(s, \pi_i(s), s') \left[R(s, \pi_i(s), s') + \gamma V^{\pi_i}(s') \right]$$

   For state $s = 2$,

$$\begin{aligned} V^{\pi_0}(2) &= T(2, high, 2)[R(2, high, 2) + V^{\pi_0}(2)] + T(2, high, 3)[R(2, high, 3) + V^{\pi_0}(3)] + \\ &\quad T(2, high, 4)[R(2, high, 4) + V^{\pi_0}(4)] \\ &= \frac{1}{3}[0 + V^{\pi_0}(2)] + \frac{1}{3}[3 + V^{\pi_0}(3)] + \frac{1}{3}[4 + V^{\pi_0}(4)] \\ &= \frac{1}{3}V^{\pi_0}(2) + \frac{1}{3}V^{\pi_0}(3) + \frac{1}{3}V^{\pi_0}(4) + \frac{7}{3} \end{aligned}$$

   For state $s = 3$,

$$
\begin{aligned}
V^{\pi_0}(3) &= T(3, high, done)[R(3, high, done) + V^{\pi_0}(done)] + \\
&\quad T(3, high, 3)[R(3, high, 3) + V^{\pi_0}(3)] + \\
&\quad T(3, high, 4)[R(3, high, 4) + V^{\pi_0}(4)] \\
&= \frac{1}{3}[0 + 0] + \frac{1}{3}[0 + V^{\pi_0}(3)] + \frac{1}{3}[4 + V^{\pi_0}(4)] \\
&= \frac{1}{3}V^{\pi_0}(3) + \frac{1}{3}V^{\pi_0}(4) + \frac{4}{3}
\end{aligned}
$$

For state $s = 4$,

$$
\begin{aligned}
V^{\pi_0}(4) &= T(4, high, done)[R(4, high, done) + V^{\pi_0}(done)] + \\
&\quad T(4, high, 4)[R(4, high, 4) + V^{\pi_0}(4)] \\
&= \frac{2}{3}[0 + 0] + \frac{1}{3}[0 + V^{\pi_0}(4)] \\
&= \frac{1}{3}V^{\pi_0}(4)
\end{aligned}
$$

The last equation immediately gives $V^{\pi_0}(4) = 0$. For state $s = 3$,

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

Finally, for state $s = 2$,

$$
\begin{aligned}
V^{\pi_0}(2)[1 - \frac{1}{3}] &= \frac{1}{3} \cdot 2 + \frac{1}{3} \cdot 0 + \frac{7}{3} \\
V^{\pi_0}(2) &= \frac{9}{2}
\end{aligned}
$$

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

The general analytical equation is also used for policy improvement:

$$\pi_{i+1}(s) = \arg\max_a \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma V^{\pi_i}(s')\right]$$

In all states, we have already calculated the utility values for the action $a = high$, so the new calculations are for the actions $a = low$ ($\gamma = 1$ has been substituted).

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

For state $s = 2$,

$$\begin{aligned}
\sum_{s'} T(2,low,s') \left[R(2,low,s') + V^{\pi_0}(s')\right] &= T(2,low,2)[R(2,low,2) + V^{\pi_0}(2)] + \\
&\quad T(2,low,done)[R(2,low,done) + V^{\pi_0}(done)] \\
&= \frac{1}{3}[0 + \frac{9}{2}] + \frac{2}{3}[0+0] \\
&= \frac{3}{2}
\end{aligned}$$

Hence

$$\begin{aligned}
\pi_1(2) &= \arg\max_a \begin{cases} a = high : \frac{9}{2} \\ a = low : \frac{3}{2} \end{cases} \\
&= high
\end{aligned}$$

The policy does not change. For state $s = 3$, the low calculation is:

$$\begin{aligned}
\sum_{s'} T(3,low,s') \left[R(3,low,s') + V^{\pi_0}(s')\right] &= T(3,low,2)[R(3,low,2) + V^{\pi_0}(2)] + \\
&\quad T(3,low,3)[R(3,low,3) + V^{\pi_0}(3)] + \\
&\quad T(3,low,done)[R(3,low,done) + V^{\pi_0}(done)] \\
&= \frac{1}{3}[2 + \frac{9}{2}] + \frac{1}{3}[0+2] + \frac{1}{3}[0+0] \\
&= \frac{17}{6}
\end{aligned}$$

Hence

$$\begin{aligned}\pi_1(3) &= \arg\max_a \begin{cases} a = high : 2 \\ a = low : \frac{17}{6} \end{cases} \\ &= low\end{aligned}$$

The new policy for $s = 3$ changes to low. For state $s = 4$, the low calculation is:

$$\begin{aligned}\sum_{s'} T(4, low, s') \left[R(4, low, s') + V^{\pi_0}(s')\right] &= T(4, low, 2)[R(4, low, 2) + V^{\pi_0}(2)] + \\ &\quad T(4, low, 3)[R(4, low, 3) + V^{\pi_0}(3)] + \\ &\quad T(4, low, 4)[R(4, low, 4) + V^{\pi_0}(4)] \\ &= \frac{1}{3}[2 + \frac{9}{2}] + \frac{1}{3}[3 + 2] + \frac{1}{3}[0 + 0] \\ &= \frac{23}{6}\end{aligned}$$

Hence

$$\begin{aligned}\pi_1(4) &= \arg\max_a \begin{cases} a = high : 0 \\ a = low : \frac{23}{6} \end{cases} \\ &= low\end{aligned}$$

The new policy for $s = 4$ also changes to low.