CS 6300 **HW4: Value Iteration** **Due February 21, 2017**

Please use the LaTeX template to produce your writeups. See the Homework Assignments page on the class website for details. Hand in at: `https://webhandin.eng.utah.edu/index.php`.

1 Value Iteration

At the AI casino, there are two things to do: Eat Buffet and Play AI 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:

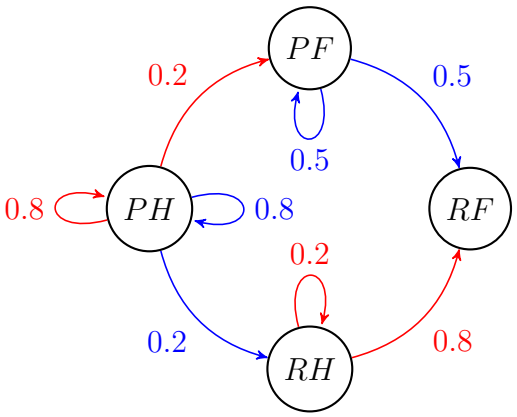
State Space	{PoorHungry, PoorFull, RichHungry, RichFull}
Actions	{Eat, Play}
Initial State	PoorHungry
Terminal State	RichFull

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



Where red denotes the action to Eat and blue denotes Play.

1. Complete the table for the first 3 iterations of Value Iteration. Assume $\gamma = 1$.

State	$i = 0$	$i = 1$	$i = 2$	$i = 3$
PoorHungry	0.000	-0.600	-0.480	-0.084
PoorFull	0.000	3.000	4.500	5.250
RichHungry	0.000	4.000	4.800	4.960
RichFull	0.000	0.000	0.000	0.000

The equation used is the *Bellman Update* Equation which is defined as

$$Q^*_{i+1}(s,a) = \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_a Q^*_i(s') \right]$$

The values of *PoorHungry* (PH) is calculated for two iterations to show the steps.

$$Q_1(PH) = 0.8(-1+0) + \max_a \begin{cases} 0.2(1+0) = -0.60 \\ 0.2(0+0) = -0.80 \end{cases}$$
$$Q_2(PH) = 0.8(-1-0.6) + \max_a \begin{cases} 0.2(1+3) = -0.48 \\ 0.2(0+4) = -0.48 \end{cases}$$

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

   After 3 iterations, the Q-values have converged enough to where we can make a reasonable decision on the action that should be taken. From the above table, we can see that the best action to be chosen is to *Eat*, as the expected return value of *PoorFull* is 5.250 compared to the value of 4.960.

   This is expected as we can get "more" by taking PF, as there is a reward on reaching that state, even if looping back, compared to $R(RH) = 0$.

   As the number of iterations approaches infinity, the system converges such that $Q_\infty(PH) = 3.0$, $Q_\infty(PF) = 6.0$, and $Q_\infty(RH) = 5.0$.