CS 4100/5100: Foundations of Artificial Intelligence (Fall 2015) Robert Platt
Student(s): Oct. 15, 2015

MDP, Value Iteration and Policy Solution

1 MDP Modeling

In a coin game, you repeatedly toss a biased coin (0.6 for head, 0.4 for tail). Each head represent 4 points and tail represents 1 points. You can either Toss or Stop if the total number of points you have tossed is no more than 7. Otherwise, you must Stop. When you Stop, your utility is equal to your total points (up to 7), or zero if you get a total of 8 or higher. When you Toss, you receive no utility. There is no discount ($\gamma = 1$).

1. What are the states and the actions for this MDP?

   State: current points if stop plus a terminal state, that is, 0,1,2,3,4,5,6,7,DONE
   Action: Toss, Stop

2. What is the transition function and the reward function for this MDP?

   Transition function:
   $T(S_i, TOSS, S_{i+4}) = 0.6$ if $i < 4$
   $T(S_i, TOSS, DONE) = 0.6$ if $i \geq 4$
   $T(S_i, TOSS, S_{i+1}) = 0.4$ if $i < 7$
   $T(S_i, TOSS, DONE) = 0.6$ if $i = 7$
   $T(S_i, STOP, DONE) = 1$

   Reward function:
   $R(S_i, TOSS, ANY) = 0$
   $R(S_i, STOP, DONE) = i$
   $R(DONE, STOP, DONE) = 0$

3. What is the optimal policy for this MDP? Please write down the steps to show how you get the optimal policy.

   Optimal policy: Toss for 0,1,2,3; STOP for others.
   You should include the steps of value iteration. The value iteration will converge at iteration 4. Result of iteration 4 is as follow,
   V4: 0: 4.5 from Toss; 1: 5.4 from Toss; 2: 5.9 from Toss; 3: 5.8 from Toss; 4: 4 from Stop; 5: 5 from Stop; 6: 6 from Stop; 7: 7 from Stop

2 MDP Properties and Concepts

We assume there is a MDP which has a finite number of actions and states.

1. What's the condition that can make this MDP guaranteed to converge? Why? HINT: discount factor.

   Discount factor < 1 will guarantee this finite MDP to converge.

2. Do the converged values change based on different initial values? Why?

   No, Since when converge, $V^*(s) = \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^*(s')]$ There will only be one set of values that satisfies this condition, so no matter where we start value iteration, we will always arrive at the same set of values on convergence. That is because the discount factor, which is less than 1 from question 1, will erase the impact of initial values.

3. If the values in value iteration have just converged, is the policy converged as well at that time? Why?

   Yes, policy will converge before values. When values converge, policy is action that yield that value.

4. If the policy in value iteration have just converged, are the values converged as well at that time? Why?

   No, policy will converge before values. When policy converge, there may be still some small changes of values before converge.

5. If two MDPs have the same actions and states, except the discount factor and they both converge, will they have the same policy? If no, please write down a SIMPLE MDP example with the explanation, otherwise write down the reasons.

   No, example as shown in MDP slides deck slide 23 "Discounting example".
   For map a:10,b,c,d,e:1, action is west, east, exit (only available in a, e), transition is deterministic, reward at exit action.

   If $\gamma = 1$, optimal policy for b,c,d is west,west,west
   If $\gamma = 0.1$, optimal policy for b,c,d is west,west,east