CS 188
Summer 2024

Regular Discussion 9 Solutions

1 MDPs: Micro-Blackjack

In micro-blackjack, you repeatedly draw a card (with replacement) that is equally likely to be a 2, 3, or 4. You can either Draw or Stop if the total score of the cards you have drawn is less than 6. If your total score is 6 or higher, the game ends, and you receive a utility of 0. When you Stop, your utility is equal to your total score (up to 5), and the game ends. When you Draw, you receive no utility. There is no discount ($\gamma = 1$). Let's formulate this problem as an MDP with the following states: $0, 2, 3, 4, 5$ and a *Done* state, for when the game ends.

(a) What is the transition function and the reward function for this MDP? The transition function is

$$\begin{aligned}
T(s, Stop, Done) &= 1 \\
T(0, Draw, s') &= 1/3 \text{ for } s' \in \{2, 3, 4\} \\
T(2, Draw, s') &= 1/3 \text{ for } s' \in \{4, 5, Done\} \\
T(3, Draw, s') &= \begin{matrix} 1/3 \text{ if } s' = 5 \\ 2/3 \text{ if } s' = Done \end{matrix} \\
T(4, Draw, Done) &= 1 \\
T(5, Draw, Done) &= 1 \\
T(s, a, s') &= 0 \text{ otherwise}
\end{aligned}$$

The reward function is

$$R(s, Stop, Done) = s, s \leq 5$$

$$R(s, a, s') = 0 \text{ otherwise}$$

(b) Fill in the following table of value iteration values for the first 4 iterations.

States	0	2	3	4	5
V_0	0	0	0	0	0
V_1	0	2	3	4	5
V_2	3	3	3	4	5
V_3	10/3	3	3	4	5
V_4	10/3	3	3	4	5

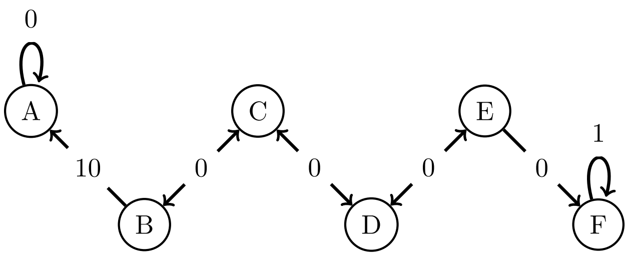
(c) You should have noticed that value iteration converged above. What is the optimal policy for the MDP?

States	0	2	3	4	5
π^*	Draw	Draw	Stop	Stop	Stop

(d) Perform one iteration of policy iteration for one step of this MDP, starting from the fixed policy below:

States	0	2	3	4	5
π_i	Draw	Stop	Draw	Stop	Draw
V^{π_i}	2	2	0	4	0
π_{i+1}	Draw	Stop	Stop	Stop	Stop

2 MDP



Consider the MDP above, with states represented as nodes and transitions as edges between nodes. The rewards for the transitions are indicated by the numbers on the edges. For example, going from state B to state A gives a reward of 10, but going from state A to itself gives a reward of 0. Some transitions are not allowed, such as from state A to state B. Transitions are deterministic (if there is an edge between two states, the agent can choose to go from one to the other and will reach the other state with probability 1).

(a) For this part only, suppose that the max horizon length is 15. Write down the optimal action at each step if the discount factor is $\gamma = 1$.

A: Go to A
B: Go to C
C: Go to D
D: Go to E
E: Go to F
F: Go to F

(b) Now suppose that the horizon is infinite. For each state, does the optimal action depend on γ? If so, for each state, write an equation that would let you determine the value for γ at which the optimal action changes.

A: Only staying at A is a possible action. For the other states, let n be the number of steps to B, and m be the number of steps to F. Then, the value of going left is $10\gamma^n$ and the value of going right is $\sum_{k=m}^{\infty} \gamma^k = \frac{1}{1-\gamma} - \frac{1-\gamma^m}{1-\gamma}$ because of the geometric series. Now we find the value of γ at which these are equal.

$$10\gamma^n = \frac{1}{1-\gamma} - \frac{1-\gamma^m}{1-\gamma} = \frac{\gamma^m}{1-\gamma}$$
$$10 - 10\gamma = \gamma^{m-n}$$
$$\gamma^{m-n} + 10\gamma - 10 = 0$$

The roots of the above polynomial are the points at which the optimal action changes.