CS 188 Spring 2019 Section Handout 3 Solutions

The preamble is an abbrevation of the lecture notes

Markov Decision Processes

A Markov Decision Process is defined by several properties:

- A set of states S
- A set of actions A.
- A start state.
- Possibly one or more terminal states.
- Possibly a **discount factor** γ.
- A **transition function** $T(s, a, s')$.
- A **reward function** $R(s, a, s')$.

The Bellman Equation

- $V^*(s)$ – the optimal value of s is the expected value of the utility an optimally-behaving agent that starts in s will receive, over the rest of the agent's lifetime.
- $Q^*(s, a)$ - the optimal value of (s, a) is the expected value of the utility an agent receives after starting in s, taking a, and acting optimally henceforth.

Using these two new quantities and the other MDP quantities discussed earlier, the Bellman equation is defined as follows:

$$V^*(s) = \max_a \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V^*(s')]$$

We can also define he equation for the optimal value of a q-state (more commonly known as an optimal **q-value**):

$$Q^*(s, a) = \sum_{s'} T(s, a, s')[R(s, a, s') + \gamma V^*(s')]$$

which allows us to reexpress the Bellman equation as

$$V^*(s) = \max_a Q^*(s, a).$$

Value Iteration

The **time-limited value** for a state s with a time-limit of k timesteps is denoted $V_k(s)$, and represents the maximum expected utility attainable from s given that the Markov decision process under consideration terminates in k timesteps. Equivalently, this is what a depth-k expectimax run on the search tree for a MDP returns.

Value iteration is a **dynamic programming algorithm** that uses an iteratively longer time limit to compute time-limited values until convergence (that is, until the V values are the same for each state as they were in the past iteration: $\forall s, V_{k+1}(s) = V_k(s)$). It operates as follows:

1. $\forall s \in S$, initialize $V_0(s) = 0$. This should be intuitive, since setting a time limit of 0 timesteps means no actions can be taken before termination, and so no rewards can be acquired.

2. Repeat the following update rule until convergence:

$$\forall s \in S,\ \ V_{k+1}(s) \leftarrow \max_a \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V_k(s')]$$

   At iteration k of value iteration, we use the time-limited values for with limit k for each state to generate the time-limited values with limit $(k+1)$. In essence, we use computed solutions to subproblems (all the $V_k(s)$) to iteratively build up solutions to larger subproblems (all the $V_{k+1}(s)$); this is what makes value iteration a dynamic programming algorithm.

Policy Iteration

If all we want is to determine the optimal policy for the MDP value iteration tends to do a lot of overcomputation since the policy as computed by policy extraction generally converges significantly faster than the values themselves. This motivates **policy iteration**, an algorithm that maintains the optimality of value iteration while providing significant performance gains. It operates as follows:

1. Define an *initial policy*. This can be arbitrary, but policy iteration will converge faster the closer the initial policy is to the eventual optimal policy.

2. Repeat the following until convergence:

   - **Policy evaluation:** For a policy π, policy evaluation means computing $V^\pi(s)$ for all states s, where $V^\pi(s)$ is expected utility of starting in state s when following π:

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

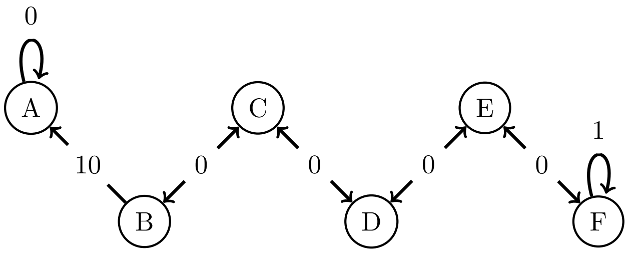
     Define the policy at iteration i of policy iteration as π_i. Since we are fixing a single action for each state, we no longer need the max operator which effectively leaves us with a system of $|S|$ equations generated by the above rule. Each $V^{\pi_i}(s)$ can then be computed by simply solving this system.

   - **Policy improvement:** Policy improvement uses policy extraction on the values of states generated by policy evaluation to generate this new and improved policy:

$$\pi_{i+1}(s) = \underset{a}{\operatorname{argmax}}\, Q^*(s,a) = \underset{a}{\operatorname{argmax}} \sum_{s'} T(s,a,s')[R(s,a,s') + \gamma V^{\pi_i}(s')]$$

     If $\pi_{i+1} = \pi_i$, the algorithm has converged, and we can conclude that $\pi_{i+1} = \pi_i = \pi^*$.

1 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.

2 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

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

The reward function is

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

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

(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?

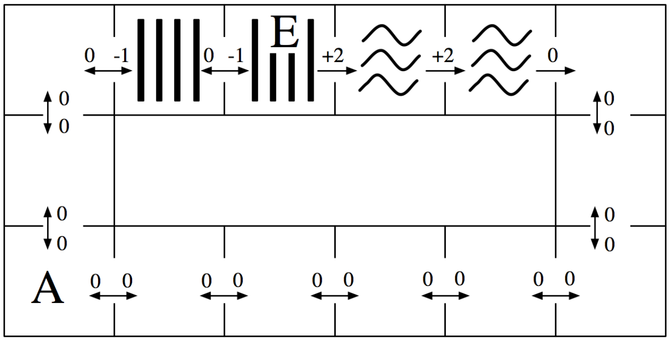
States	0	2	3	4	5
π^*	Draw	Draw	Stop	Stop	Stop

(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

Q3. MDPs: Grid-World Water Park

Consider the MDP drawn below. The state space consists of all squares in a grid-world water park. There is a single waterslide that is composed of two ladder squares and two slide squares (marked with vertical bars and squiggly lines respectively). An agent in this water park can move from any square to any neighboring square, unless the current square is a slide in which case it must move forward one square along the slide. The actions are denoted by arrows between squares on the map and all deterministically move the agent in the given direction. The agent cannot stand still: it must move on each time step. Rewards are also shown below: the agent feels great pleasure as it slides down the water slide (+2), a certain amount of discomfort as it climbs the rungs of the ladder (-1), and receives rewards of 0 otherwise. The time horizon is infinite; this MDP goes on forever.



(a) How many (deterministic) policies π are possible for this MDP?

2^{11}

(b) Fill in the blank cells of this table with values that are correct for the corresponding function, discount, and state. *Hint: You should not need to do substantial calculation here.*

	γ	$s = A$	$s = E$
$V_3^*(s)$	1.0	0	4
$V_{10}^*(s)$	1.0	2	4
$V_{10}^*(s)$	0.1	0	2.2
$Q_1^*(s, \text{west})$	1.0	——	0
$Q_{10}^*(s, \text{west})$	1.0	——	3
$V^*(s)$	1.0	∞	∞
$V^*(s)$	0.1	0	2.2

$V_{10}^*(A), \gamma = 1$: In 10 time steps with no discounting, the rewards don't decay, so the optimal strategy is to climb the two stairs (-1 reward each), and then slide down the two slide squares (+2 rewards each). You only have time to do this once. Summing this up, we get $-1 - 1 + 2 + 2 = 2$.

$V_{10}^*(E), \gamma = 1$: No discounting, so optimal strategy is sliding down the slide. That's all you have time for. Sum of rewards $= 2 + 2 = 4$.

$V_{10}^*(A), \gamma = 0.1$. The discount rate is 0.1, meaning that rewards 1 step further into the future are discounted by a factor of 0.1. Let's assume from A, we went for the slide. Then, we would have to take the actions $A \rightarrow B, B \rightarrow C, C \rightarrow D, D \rightarrow E, E \rightarrow F, F \rightarrow G$. We get the first -1 reward from $C \rightarrow D$, discounted by γ^2 since it is two actions in the future. $D \rightarrow E$ is discounted by γ^3, $E \rightarrow F$ by γ^4, and $F \rightarrow G$ by γ^5. Since γ is low, the positive rewards you get from the slide have less of an effect as the larger negative rewards you get from climbing up. Hence, the sum of rewards of taking the slide path would be negative; the optimal value is 0.

$V_{10}^*(E), \gamma = 0.1$. Now, you don't have to do the work of climbing up the stairs, and you just take the slide down. Sum of rewards would be 2 (for $E \rightarrow F$) + 0.2 (for $F \rightarrow G$, discounted by 0.1) = 2.2.

$Q_{10}^*(E, west), \gamma = 1$. Remember that a Q-state (s,a) is when you start from state s and are committed to taking a. Hence, from E, you take the action West and land in D, using up one time step and getting an immediate reward of 0. From D, the optimal strategy is to climb back up the higher flight of stairs and then slide down the slide. Hence, the rewards would be $-1(D \rightarrow E) + 2(E \rightarrow F) + 2(F \rightarrow G) = 3$.

$V^*(s), \gamma = 1$. Infinite game with no discount? Have fun sliding down the slide to your content from anywhere.

$V^*(s), \gamma = 0.1$. Same reasoning apply to both A and E from $V_{10}^*(s)$. With discounting, the stairs are more costly to climb than the reward you get from sliding down the water slide. Hence, at A, you wouldn't want to head to the slide. From E, since you are already at the top of the slide, you should just slide down.