CS 188 Fall 2018 Introduction to Artificial Intelligence Written HW 4 Sol.

Self-assessment due: Monday 10/1/2018 at 11:59pm (submit via Gradescope)

For the self assessment, **fill in the self assessment boxes in your original submission** (you can download a PDF copy of your submission from Gradescope). For each subpart where your original answer was correct, write "correct." Otherwise, write and explain the correct answer.

Q1. MDPs: Dice Bonanza

A casino is considering adding a new game to their collection, but need to analyze it before releasing it on their floor. They have hired you to execute the analysis. On each round of the game, the player has the option of rolling a fair 6-sided die. That is, the die lands on values 1 through 6 with equal probability. Each roll costs 1 dollar, and the player **must** roll the very first round. Each time the player rolls the die, the player has two possible actions:

1. *Stop*: Stop playing by collecting the dollar value that the die lands on, or
2. *Roll*: Roll again, paying another 1 dollar.

Having taken CS 188, you decide to model this problem using an infinite horizon Markov Decision Process (MDP). The player initially starts in state *Start*, where the player only has one possible action: *Roll*. State s_i denotes the state where the die lands on i. Once a player decides to *Stop*, the game is over, transitioning the player to the *End* state.

(a) In solving this problem, you consider using policy iteration. Your initial policy π is in the table below. Evaluate the policy at each state, with $\gamma = 1$.

State	s_1	s_2	s_3	s_4	s_5	s_6
$\pi(s)$	*Roll*	*Roll*	*Stop*	*Stop*	*Stop*	*Stop*
$V^{\pi}(s)$	3	3	3	4	5	6

We have that $s_i = i$ for $i \in \{3, 4, 5, 6\}$, since the player will be awarded no further rewards according to the policy. From the Bellman equations, we have that $V(s_1) = -1 + \frac{1}{6}(V(s_1) + V(s_2) + 3 + 4 + 5 + 6)$ and that $V(s_2) = -1 + \frac{1}{6}(V(s_1) + V(s_2) + 3 + 4 + 5 + 6)$. Solving this linear system yields $V(s_1) = V(s_2) = 3$.

(b) Having determined the values, perform a policy update to find the new policy π'. The table below shows the old policy π and has filled in parts of the updated policy π' for you. If both *Roll* and *Stop* are viable new actions for a state, write down both *Roll*/*Stop*. In this part as well, we have $\gamma = 1$.

State	s_1	s_2	s_3	s_4	s_5	s_6
$\pi(s)$	*Roll*	*Roll*	*Stop*	*Stop*	*Stop*	*Stop*
$\pi'(s)$	*Roll*	*Roll*	*Roll*/*Stop*	*Stop*	*Stop*	*Stop*

For each s_i in part (a), we compare the values obtained via Rolling and Stopping. The value of Rolling for each state s_i is $-1 + \frac{1}{6}(3 + 3 + 3 + 4 + 5 + 6) = 3$. The value of Stopping for each state s_i is i. At each state s_i, we take the action that yields the largest value; so, for s_1 and s_2, we Roll, and for s_4 and s_5, we stop. For s_3, we Roll/Stop, since the values from Rolling and Stopping are equal.

(c) Is $\pi(s)$ from part (a) optimal? Explain why or why not.

Yes, the old policy is optimal. Looking at part (b), there is a tie between 2 equally good policies that policy iteration considers employing. One of these policies is the same as the old policy. This means that both new policies are as equally good as the old policy, and policy iteration has converged. Since policy iteration converges to the optimal policy, we can be sure that $\pi(s)$ from part (a) is optimal.

(d) Suppose that we were now working with some $\gamma \in [0, 1)$ and wanted to run **value iteration**. Select the **one** statement that would hold true at convergence, or write the correct answer next to Other if none of the options are correct.

○ $V^*(s_i) = \max\left\{-1 + \frac{i}{6} \ , \ \sum_j \gamma V^*(s_j)\right\}$

○ $V^*(s_i) = \max\left\{i \ , \ \frac{1}{6} \cdot \left[-1 + \sum_j \gamma V^*(s_j)\right]\right\}$

○ $V^*(s_i) = \max\left\{-\frac{1}{6} + i \ , \ \sum_j \gamma V^*(s_j)\right\}$

○ $V^*(s_i) = \max\left\{i \ , \ -\frac{1}{6} + \sum_j \gamma V^*(s_j)\right\}$

○ $V^*(s_i) = \frac{1}{6} \cdot \sum_j \max\{i \ , \ -1 + \gamma V^*(s_j)\}$

○ $V^*(s_i) = \frac{1}{6} \cdot \sum_j \max\left\{-1 + i \ , \ \sum_k V^*(s_j)\right\}$

○ $V^*(s_i) = \sum_j \max\left\{-1 + i \ , \ \frac{1}{6} \cdot \gamma V^*(s_j))\right\}$

○ $V^*(s_i) = \sum_j \max\left\{\frac{i}{6} \ , \ -1 + \gamma V^*(s_j)\right\}$

● $V^*(s_i) = \max\left\{i \ , \ -1 + \frac{\gamma}{6} \sum_j V^*(s_j)\right\}$

○ $V^*(s_i) = \sum_j \max\left\{i \ , \ -\frac{1}{6} + \gamma V^*(s_j)\right\}$

○ $V^*(s_i) = \sum_j \max\left\{\frac{-i}{6} \ , \ -1 + \gamma V^*(s_j)\right\}$

○ Other _____

At convergence,

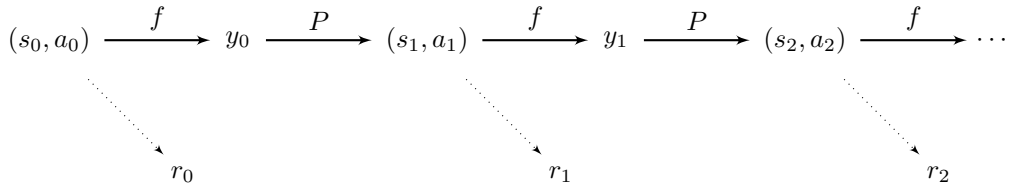
$$\begin{aligned} V^*(s_i) &= \max_a Q^*(s_i, a) \\ &= \max\{Q^*(s_i, \text{stop}), \quad Q^*(s_i, \text{roll})\} \\ &= \max\left\{R(s_i, \text{stop}), \quad R(s_i, \text{roll}) + \gamma \sum_j T(s_i, \text{roll}, s_j) V^*(s_j)\right\} \\ &= \max\left\{i, \quad -1 + \frac{\gamma}{6} \sum_j V^*(s_j)\right\} \end{aligned}$$

Q2. Bellman Equations for the Post-Decision State

Consider an infinite-horizon, discounted MDP (S, A, T, R, γ). Suppose that the transition probabilities and the reward function have the following form:

$$T(s, a, s') = P(s'|f(s, a)), \qquad R(s, a, s') = R(s, a)$$

Here, f is some deterministic function mapping $S \times A \to Y$, where Y is a set of states called *post-decision states*. We will use the letter y to denote an element of Y, i.e., a post-decision state. In words, the state transitions consist of two steps: a deterministic step that depends on the action, and a stochastic step that does not depend on the action. The sequence of states (s_t), actions (a_t), post-decision-states (y_t), and rewards (r_t) is illustrated below.



You have learned about $V^\pi(s)$, which is the expected discounted sum of rewards, starting from state s, when acting according to policy π.

$$V^\pi(s_0) = E\left[R(s_0, a_0) + \gamma R(s_1, a_1) + \gamma^2 R(s_2, a_2) + \dots\right] \qquad \text{given } a_t = \pi(s_t) \text{ for } t = 0, 1, 2, \dots$$

$V^*(s)$ is the value function of the optimal policy, $V^*(s) = \max_\pi V^\pi(s)$.

This question will explore the concept of computing value functions on the post-decision-states y. [1]

$$W^\pi(y_0) = E\left[R(s_1, a_1) + \gamma R(s_2, a_2) + \gamma^2 R(s_3, a_3) + \dots\right]$$

We define $W^*(y) = \max_\pi W^\pi(y)$.

[1] In some applications, it is easier to learn an approximate W function than V or Q. For example, to use reinforcement learning to play Tetris, a natural approach is to learn the value of the block pile *after* you've placed your block, rather than the value of the pair (current block, block pile). TD-Gammon, a computer program developed in the early 90s, was trained by reinforcement learning to play backgammon as well as the top human experts. TD-Gammon learned an approximate W function.

(a) Write W^* in terms of V^*.
$W^*(y) =$

- [x] $\sum_{s'} P(s' \mid y)V^*(s')$
- [ ] $\sum_{s'} P(s' \mid y)[V^*(s') + \max_a R(s', a)]$
- [ ] $\sum_{s'} P(s' \mid y)[V^*(s') + \gamma \max_a R(s', a)]$
- [ ] $\sum_{s'} P(s' \mid y)[\gamma V^*(s') + \max_a R(s', a)]$
- [ ] None of the above

Consider the expected rewards under the optimal policy.

$$\begin{aligned} W^*(y_0) &= E\left[R(s_1, a_1) + \gamma R(s_2, a_2) + \gamma^2 R(s_3, a_3) + \dots \mid y_0\right] \\ &= \sum_{s_1} P(s_1 \mid y_0) E\left[R(s_1, a_1) + \gamma R(s_2, a_2) + \gamma^2 R(s_3, a_3) + \dots \mid s_1\right] \\ &= \sum_{s_1} P(s_1 \mid y_0) V^*(s_1) \end{aligned}$$

V^* is time-independent, so we can replace y_0 by y and replace s_1 by s', giving

$$W^*(y) = \sum_{s'} P(s' \mid y) V^*(s')$$

(b) Write V^* in terms of W^*.
$V^*(s) =$

- [ ] $\max_a[W^*(f(s, a))]$
- [ ] $\max_a[R(s, a) + W^*(f(s, a))]$
- [x] $\max_a[R(s, a) + \gamma W^*(f(s, a))]$
- [ ] $\max_a[\gamma R(s, a) + W^*(f(s, a))]$
- [ ] None of the above

$$\begin{aligned} V^*(s_0) &= \max_{a_0} Q(s_0, a_0) \\ &= \max_{a_0} E\left[R(s_0, a_0) + \gamma R(s_1, a_1) + \gamma^2 R(s_2, a_2) + \dots \mid s_0, a_0\right] \\ &= \max_{a_0} \left(E\left[R(s_0, a_0) \mid s_0, a_0\right] + E\left[\gamma R(s_1, a_1) + \gamma^2 R(s_2, a_2) + \dots \mid s_0, a_0\right]\right) \\ &= \max_{a_0} \left(R(s_0, a_0) + E\left[\gamma R(s_1, a_1) + \gamma^2 R(s_2, a_2) + \dots \mid f(x_0, a_0)\right]\right) \\ &= \max_{a_0} \left(R(s_0, a_0) + \gamma W^*(f(s_0, a_0))\right) \end{aligned}$$

Renaming variables, we get

$$V^*(s) = \max_a \left(R(s, a) + \gamma W^*(f(s, a))\right)$$

(c) Recall that the optimal value function V^* satisfies the Bellman equation:

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

which can also be used as an update equation to compute V^*.

Provide the equivalent of the Bellman equation for W^*.

$W^*(y) =$ $\sum_{s'} P(s'|y) \max_a \left(R(s',a) + \gamma W^*(f(s',a))\right)$

The answer follows from combining parts (a) and (b)

(d) Fill in the blanks to give a policy iteration algorithm, which is guaranteed return the optimal policy π^*.

- Initialize policy $\pi^{(1)}$ arbitrarily.
- For $i = 1, 2, 3, \ldots$
  - Compute $W^{\pi^{(i)}}(y)$ for all $y \in Y$.
  - Compute a new policy $\pi^{(i+1)}$, where $\pi^{(i+1)}(s) = \arg\max_a$ (1) for all $s \in S$.
  - If (2) for all $s \in S$, **return** $\pi^{(i)}$.

Fill in your answers for blanks (1) and (2) below.

(1)
- ○ $W^{\pi^{(i)}}(f(s,a))$
- ○ $R(s,a) + W^{\pi^{(i)}}(f(s,a))$
- ● $R(s,a) + \gamma W^{\pi^{(i)}}(f(s,a))$
- ○ $\gamma R(s,a) + W^{\pi^{(i)}}(f(s,a))$
- ○ None of the above

(2) $\pi^{(i)}(s) = \pi^{(i+1)}(s)$

Policy iteration performs the following update:

$$\pi^{(i+1)}(s) = \arg\max_a Q^{\pi^{(i)}}(s,a)$$

Next we express Q^π in terms of W^π (similarly to part b):

$$\begin{aligned} Q^\pi(s_0,a_0) &= E\left[R(s_0,a_0) + \gamma R(s_1,a_1) + \gamma^2 R(s_2,a_2) + \ldots \;\middle|\; s_0, a_0\right] \\ &= R(s_0,a_0) + \gamma E\left[R(s_1,a_1) + \gamma R(s_2,a_2) + \ldots \;\middle|\; f(s_0,a_0)\right] \\ &= R(s_0,a_0) + \gamma W^\pi(f(s_0,a_0)) \end{aligned}$$