CS 5100: Foundations of Artificial Intelligence (Spring 2022) Robert Platt
Student: Due March 23rd, 2022

Markov Decision Processes and Reinforcement Learning

1 MDP Modeling [3 points]

In a coin game, you repeatedly toss a biased coin (0.75 for head, 0.25 for tail). Each head represent 3 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 ($\gamma = 1$).

(a) What are the states and the actions for this MDP?

State:
Action:

(b) What is the transition function and the reward function for this MDP?
Transition function:
Reward function:

(c) Give an intuitively good policy for this problem (you do not need to calculate the optimal policy).

2 Bellman Equation [4 points]

1. (Modified from Reinforcement Learning: An Introduction 3.14)

   Consider the Bellman equation for deterministic policies and state-only rewards:

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

   We often need to consider stochastic policies as well, which we denote by $\pi(a|s)$ instead of $\pi(s)$. $\pi(a|s)$ specifies the probability of taking action a in state s. When the policy is deterministic, exactly one action a will have probability 1, so we overload notation and refer to that action as $a = \pi(s)$. *Note*: The output types are different; $\pi(a|s)$ outputs a *probability*, whereas $\pi(s)$ outputs an *action*.

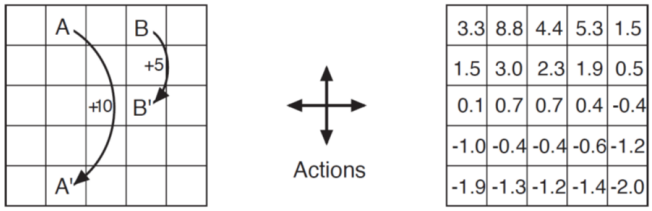
   A more general version of the Bellman equation can be derived for *stochastic* policies and reward functions depending on (s, a, s'):

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

   (a) Explain, in words, what the general version of the Bellman equation means. Additionally, show that it reduces to the simpler version when using deterministic policies $\pi(s)$ and state-only rewards $R(s)$.

   Now, consider the following gridworld MDP:

> **Example 3.5: Gridworld** Figure 3.2 (left) shows a rectangular gridworld representation of a simple finite MDP. The cells of the grid correspond to the states of the environment. At each cell, four actions are possible: `north`, `south`, `east`, and `west`, which deterministically cause the agent to move one cell in the respective direction on the grid. Actions that would take the agent off the grid leave its location unchanged, but also result in a reward of -1. Other actions result in a reward of 0, except those that move the agent out of the special states A and B. From state A, all four actions yield a reward of $+10$ and take the agent to A′. From state B, all actions yield a reward of $+5$ and take the agent to B′.



> **Figure 3.2:** Gridworld example: exceptional reward dynamics (left) and state-value function for the equiprobable random policy (right).

   (b) The general version of the Bellman equation must hold for each state for the value function V^{π} shown in the right figure above. Show numerically that this equation holds for the center state, valued at $+0.7$, with respect to its four neighboring states, valued at $+2.3, +0.4, -0.4, +0.7$. The discount factor is $\gamma = 0.9$.
   Note: The numbers in the figure are accurate only to one decimal place.
   The figure shows the value function for the equiprobable random policy, i.e., $\pi(\cdot|s) = 0.25$ for all 4 actions.

(c) The Bellman equation holds for *all* policies, including optimal policies. Consider V^* and π^* shown in the figure below (middle, right respectively). Similar to the previous part, show numerically that the Bellman equation holds for the center state, valued at +17.8, with respect to its four neighboring states, for the optimal policy π^* shown in the figure (on the right).
Also show numerically that the Bellman *optimality* equation holds for the same center state, valued at +17.8, and verify that the optimal actions at that state are indeed as shown.

Example 3.8: Solving the Gridworld Suppose we solve the Bellman equation for v_* for the simple grid task introduced in Example 3.5 and shown again in Figure 3.5 (left). Recall that state A is followed by a reward of +10 and transition to state A′, while state B is followed by a reward of +5 and transition to state B′. Figure 3.5 (middle) shows the optimal value function, and Figure 3.5 (right) shows the corresponding optimal policies. Where there are multiple arrows in a cell, all of the corresponding actions are optimal.

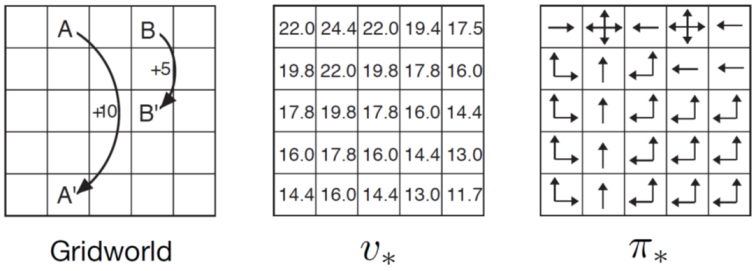


Figure 3.5: Optimal solutions to the gridworld example. ■