CS 188 Summer 2022 Regular Discussion 6A 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.

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

   The transition function is

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

   The reward function is

$$
\begin{aligned}
R(s, \mathit{Stop}, \mathit{Done}) &= s, s \leq 5 \\
R(s, a, s') &= 0 \text{ otherwise}
\end{aligned}
$$

2. 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

3. 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

4. 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

5. Consider a variant of this problem where we use a discount factor of $\gamma = 0.5$. What are the new optimal values and policy for this formulation? (Hint: You shouldn't need to recalculate all values. Start by thinking about V^* of 3, 4, and 5.)

States	0	2	3	4	5
V^*	1.5	2	3	4	5
π^*	Draw	Stop	Stop	Stop	Stop

2 Something Fishy

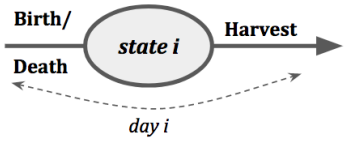
In this problem, we will consider the task of managing a fishery for an infinite number of days. (Fisheries farm fish, continually harvesting and selling them.) Imagine that our fishery has a very large, enclosed pool where we keep our fish.

Harvest (11pm): Before we go home each day at 11pm, we have the option to harvest some (possibly all) of the fish, thus removing those fish from the pool and earning us some profit, x dollars for x fish.

Birth/death (midnight): At midnight each day, some fish are born and some die, so the number of fish in the pool changes. An ecologist has analyzed the ecological dynamics of the fish population. They say that if at midnight there are x fish in the pool, then after midnight there will be exactly $f(x)$ fish in the pool, where f is a function they have provided to us. (We will pretend it is possible to have fractional fish.)

To ensure you properly maximize your profit while managing the fishery, you choose to model it using a Markov decision problem.

For this problem we will define States and Actions as follows:
State: the number of fish in the pool that day (before harvesting)
Action: the number of fish you harvest that day



(a) How will you define the transition and reward functions?

$T(s, a, s') = 1$ if $f(\max(s - a, 0)) = s'$ else 0

$R(s, a) = \min(a, s)$

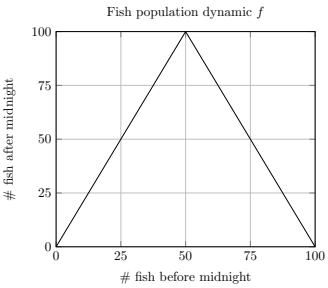
Note that taking the maximum with 0 in T and taking the minimum with s in R were not required for full credit.

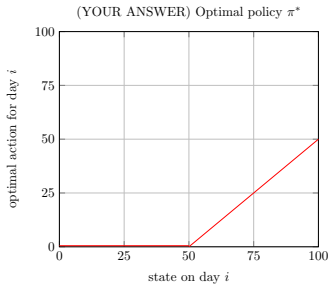
(b) Suppose the discount rate is $\gamma = 0.99$ and f is as below. Graph the optimal policy π^*.

The intuition of the question is that we should always harvest to the point where the birth of fish is highest (*i.e.* 50 in the graph). If γ is infinitely close to 1, then this is definitely true because we can ignore the first few steps before we reach 50 so your utility will be $50T$ where T is the sum of the geometric sequence. No other policies can give you 50 increase in fish population in every night.

Now for the feasibility of the question, we set γ to 0.99 but it is close enough to 1 so it won't affect the optimal policy.

Only answers which depict the piece-wise function that is $\pi^* = 0$ on $s \in [0, 50]$ and $\pi^* = s - 50$ on $s \in [50, 100]$ were accepted.



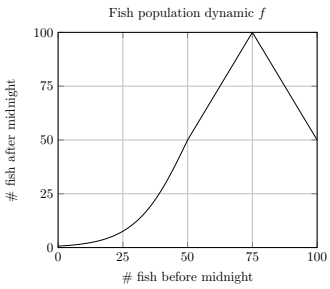


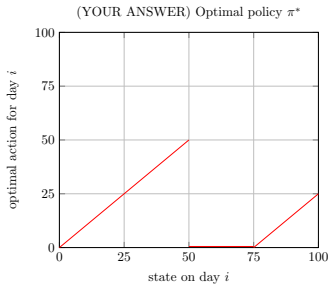
(c) Suppose the discount rate is $\gamma = 0.99$ and f is as below. Graph the optimal policy π^*.

The first part of the curve is below the curve $y = x$, so even if you do not harvest all the fish, the fish will naturally die. As a result, the optimal policy is to directly harvest them all.

The rest two regions need to be obtained by value iteration.

There are three graded components to this answer: the $\pi^* = s$ harvest-all region in $[0, 50]$, the $\pi^* = 0$ grow-to-optimal region in $[50, 75]$, and the $\pi^* = s - 75$ harvest-to-optimal region in $[75, 100]$. The first component was worth most of the points, as the other two components are difficult to come up





with. Additionally, the other two components did not need to border at exactly 75. (Note: This answer was verified by running value iteration on a computer.)