COMPSCI 687 Homework 5
(100 Points)

Instructions: Submit your answers as a single typed PDF file in Gradescope. You may use any software to create the PDF, and any reasonable formatting of your answers will suffice. You may collaborate with other students and may use generative AI to complete this assignment without reporting the collaboration or generative AI usage, but you are responsible for the accuracy of your answers.

Part One: Written (60 Points)

1. (40 Points) In this problem you will prove that Dynamic Programming Policy Evaluation converges to v^π. Assume that the MDP is finite, rewards are bounded, and $\gamma \in [0, 1)$. Let $v_0 : \mathcal{S} \to \mathbb{R}$ be arbitrary.

   (a) Write the definition of an operator $T^\pi : \mathbb{R}^{|\mathcal{S}|} \to \mathbb{R}^{|\mathcal{S}|}$ that represents one full, synchronous Dynamic Programming Policy Evaluation update. Write your definition with the left side in the form

   $$\forall s \in \mathcal{S},\, T^\pi v(s) = \cdots$$

   so that $v_{i+1} = T^\pi v_i$.
   For every $v \in \mathbb{R}^{|\mathcal{S}|}$ and $s \in \mathcal{S}$, define

   $$\forall s \in \mathcal{S}, \quad T^\pi v(s) = \sum_{a \in \mathcal{A}} \pi(s, a) \sum_{s' \in \mathcal{S}} p(s, a, s') \left(R(s, a) + \gamma v(s')\right). \tag{1}$$

   (b) Show that v^π is a fixed point of T^π. **Hint:** Compare the equation $T^\pi v^\pi = v^\pi$ with the Bellman equation for v^π. **Note:** The answer can be quite short; a single sentence is enough. This only shows that v^π is a fixed point. It does not show that the updates converge to this fixed point.
   When $T^\pi v = v$, we have that for all $s \in \mathcal{S}$,

   $$v(s) = \sum_{a \in \mathcal{A}} \pi(s, a) \sum_{s' \in \mathcal{S}} p(s, a, s') \left(R(s, a) + \gamma v(s')\right). \tag{2}$$

   That is, v satisfies the Bellman equation, and so v is v^π.

   (c) Prove that T^π is a contraction mapping under the max norm,

   $$\|v - v'\|_\infty = \max_{s \in \mathcal{S}} |v(s) - v'(s)|.$$

   State the contraction rate.
   Let $v, v' \in \mathbb{R}^{|\mathcal{S}|}$. Then

   $$\|T^\pi v - T^\pi v'\|_\infty \overset{(a)}{=} \max_{s \in \mathcal{S}} |T^\pi v(s) - T^\pi v'(s)| \tag{3}$$

   $$\overset{(b)}{=} \gamma \max_{s \in \mathcal{S}} \left| \sum_{a \in \mathcal{A}} \pi(s, a) \sum_{s' \in \mathcal{S}} p(s, a, s') \left(v(s') - v'(s')\right) \right| \tag{4}$$

   $$\overset{(c)}{\leq} \gamma \max_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi(s, a) \sum_{s' \in \mathcal{S}} p(s, a, s') |v(s') - v'(s')| \tag{5}$$

   $$\overset{(d)}{\leq} \gamma \max_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi(s, a) \max_{s' \in \mathcal{S}} |v(s') - v'(s')| \tag{6}$$

   $$\overset{(e)}{=} \gamma \max_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} \pi(s, a) \|v - v'\|_\infty \tag{7}$$

   $$\overset{(f)}{=} \gamma \|v - v'\|_\infty, \tag{8}$$

   where **(a)** follows from the definition of the max norm, **(b)** follows from the definition of T^π, cancellation of the reward terms, and factoring out γ, **(c)** follows from the triangle inequality and the nonnegativity of π and p, **(d)** follows because the average of $|v(s') - v'(s')|$ under $p(s, a, \cdot)$ is at most its maximum over s', **(e)** follows from the definition of the max norm, and **(f)** follows because $\|v - v'\|_\infty$ does not depend on a and $\sum_{a \in \mathcal{A}} \pi(s, a) = 1$. Since $\gamma < 1$, T^π is a contraction mapping with contraction rate γ.

(d) Use the Banach fixed-point theorem to prove that the sequence $v_{i+1} = T^\pi v_i$ converges to v^π. State why the theorem applies. **Note:** This part might be brief; it should bring together the results that you have shown above.

The space $\mathbb{R}^{|\mathcal{S}|}$ is non-empty and complete under the max norm. By part (c), T^π is a contraction mapping on this space. The Banach fixed-point theorem therefore implies that, for any initial v_0, the sequence $v_{i+1} = T^\pi v_i$ converges to the unique fixed point of T^π. By part (b), this fixed point is v^π. Therefore, the sequence converges to v^π.

2. (20 Points) Construct a finite episodic MDP and policy for which a one-episode every-visit Monte Carlo estimate of $v^\pi(s)$ is biased. Assume that the episode starts in s, and let $t_1 < \cdots < t_N$ be all of the times that s is visited in that episode. The estimate is

$$\widehat{v}_{\text{EV}}(s) = \frac{1}{N}\sum_{i=1}^{N} G_{t_i}.$$

Specify all components of your MDP and the policy. Compute $v^\pi(s)$ and $\mathbf{E}[\widehat{v}_{\text{EV}}(s)]$, and show that they are not equal. This is an average over all visits within one episode, not an average over visits collected from multiple episodes. Your example must have positive probability of visiting s more than once.

Let $\mathcal{S} = \{s, s_\infty\}$, $\mathcal{A} = \{a\}$, $d_0(s) = 1$, and $\gamma = 1$. The episode ends when s_∞ is reached. Define

$$p(s,a,s) = \frac{1}{2}, \qquad p(s,a,s_\infty) = \frac{1}{2}, \qquad p(s_\infty,a,s_\infty) = 1, \tag{9}$$

$$R(s,a) = 1, \qquad R(s_\infty,a) = 0, \tag{10}$$

with all unspecified transition probabilities equal to zero and all rewards deterministic given the current state and action. Let $\pi(s,a) = \pi(s_\infty,a) = 1$.

As in similar examples in the course, the expected return from s can be written as the geometric series

$$v^\pi(s) = 1 + \frac{1}{2} + \frac{1}{4} + \cdots = \frac{1}{1-\frac{1}{2}} = 2. \tag{11}$$

Let N be the number of times state s occurs in an episode. Conditioned on $N = n$, the returns from the visits to s are $n, n-1, \ldots, 1$. The formula for the sum of the first n positive integers is

$$\sum_{i=1}^{n} i = 1 + 2 + \cdots + n = \frac{n(n+1)}{2}. \tag{12}$$

Therefore, conditioned on $N = n$, the every-visit estimate is

$$\widehat{v}_{\text{EV}}(s) = \frac{1}{n}\left(n + (n-1) + \cdots + 1\right) \tag{13}$$

$$= \frac{1}{n}\sum_{i=1}^{n} i \tag{14}$$

$$= \frac{1}{n}\frac{n(n+1)}{2} \tag{15}$$

$$= \frac{n+1}{2}. \tag{16}$$

The probability of exactly n visits is

$$\Pr(N = n) = \left(\frac{1}{2}\right)^n, \qquad n = 1, 2, \ldots. \tag{17}$$

For $|x| < 1$, we will use the power-series identities

$$\sum_{n=1}^{\infty} x^n = \frac{x}{1-x} \qquad \text{and} \qquad \sum_{n=1}^{\infty} n x^n = \frac{x}{(1-x)^2}. \tag{18}$$

It follows that

$$\mathbf{E}[\widehat{v}_{\mathrm{EV}}(s)] \overset{(a)}{=} \sum_{n=1}^{\infty} \Pr(N=n)\frac{n+1}{2} \tag{19}$$

$$\overset{(b)}{=} \frac{1}{2}\left(\sum_{n=1}^{\infty} n\left(\frac{1}{2}\right)^n + \sum_{n=1}^{\infty}\left(\frac{1}{2}\right)^n\right) \tag{20}$$

$$\overset{(c)}{=} \frac{1}{2}\left(\frac{\frac{1}{2}}{\left(1-\frac{1}{2}\right)^2} + \frac{\frac{1}{2}}{1-\frac{1}{2}}\right) \tag{21}$$

$$\overset{(d)}{=} \frac{3}{2}, \tag{22}$$

where **(a)** follows from the law of total expectation and the expression above for the estimate conditioned on $N=n$, **(b)** follows by substituting $\Pr(N=n) = \left(\frac{1}{2}\right)^n$ and splitting the sum, **(c)** follows from the two power-series identities, and **(d)** follows by simplifying. Thus,

$$\mathbf{E}[\widehat{v}_{\mathrm{EV}}(s)] = \frac{3}{2} \neq 2 = v^{\pi}(s), \tag{23}$$

and the bias is $\frac{3}{2} - 2 = -\frac{1}{2}$.

Part Two: Programming (40 Points)

3. (40 Points) Implement Value Iteration for the state-value function in any programming language. Your program should read an MDP from a file and write the resulting optimal value function and the optimal action or actions for each state to two output files. Use the MDP file format from Homework 4, repeated below. All numbers are separated by spaces, and there are no headers or comments.

   MDP file: States are numbered $0, \ldots, n-1$ and actions are numbered $0, \ldots, m-1$. If the MDP has a terminal absorbing state, s_∞, it is included in these n states. The file contains:

   - Line 1: n and m, the numbers of states and actions.
   - Line 2: γ.
   - Line 3: $d_0(0), \ldots, d_0(n-1)$, written as n numbers separated by spaces.
   - The remaining nm lines: one line for each state-action pair, ordered first by state and then by action, starting with $(0,0), (0,1), \ldots, (0,m-1), (1,0), \ldots$. Each line contains $n+3$ numbers:

   $$s \quad a \quad R(s,a) \quad p(s,a,0) \quad p(s,a,1) \quad \cdots \quad p(s,a,n-1).$$

   Value output file: Write n lines, with one number per line: $\widehat{v}^*(0), \ldots, \widehat{v}^*(n-1)$, in that order. Write at least eight digits after the decimal point.

   Optimal-action output file: Write n lines. Line $s+1$ contains the numbers of all optimal actions for state s, in increasing order and separated by spaces. If only one action is optimal, the line contains one number. If multiple actions are optimal, output all of them.

   Use double precision, initialize every state value to zero, and use the synchronous Value Iteration update from class: each complete update uses only values from the previous iteration. Stop when the maximum absolute change over all states in one complete update is at most 10^{-10}, and output the newly computed vector. Compute

   $$\widehat{q}(s,a) = \sum_{s' \in \mathcal{S}} p(s,a,s') \left(R(s,a) + \gamma \widehat{v}^*(s') \right) \quad \text{and} \quad M_s = \max_{a \in \mathcal{A}} \widehat{q}(s,a).$$

   For the supplied files, output every action a satisfying $M_s - \widehat{q}(s,a) \leq 10^{-8}$.

   First, test your program using `687_gridworld_MDP.txt`. The files `687_Gridworld_opt_values.txt` and `687_Gridworld_opt_actions.txt` contain the correct outputs. **There may be slight numerical differences** due to rounding and stopping tolerances; your values should agree with the reference to within 2×10^{-6} at every state, and your optimal-action output should match the reference. The notes' gridworld locations $1, \ldots, 23$ correspond to file states $0, \ldots, 22$, and the terminal absorbing state is file state 23. Actions $0, 1, 2, 3$ are AU, AD, AL, AR, respectively.

   Then run your program using `test_MDP.txt`. This MDP has seven states and five actions. Some states have one optimal action and some have multiple optimal actions. In your PDF answer, report your estimates of $v^*(0), \ldots, v^*(6)$ and all optimal actions for each state. Give at least six digits after the decimal point for the values. You do not need to submit your source code or output files, or report the gridworld values or actions.

s	$\widehat{v}^*(s)$	optimal actions
0	1.458000000000	0, 1
1	1.620000000000	2
2	1.620000000000	0, 1
3	1.800000000000	2
4	1.800000000000	0, 3
5	2.000000000000	0, 1
6	0.000000000000	0, 1, 2, 3, 4