COMPSCI 687 Homework 4
(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)

Note: For this part, assume that we're considering only finite MDPs with bounded rewards and $\gamma < 1$. You may use results proved in the course notes, but should reference any such results that you use (e.g., "By Property 3 from the course notes...").

1. (8 Points) Derive the Bellman equation for q^π. Begin with the definition of $q^\pi(s, a)$ as an expected discounted return, and end with a recursive expression for $q^\pi(s, a)$ that uses only $R, p, \gamma, \pi, \mathcal{S}, \mathcal{A}$, and q^π. Your final answer should not contain v^π, expectations, or random variables. Show your work and explain the steps.

2. (8 Points) Before looking up the answer, write what you think the Bellman optimality equation for q^* is, using the notation of this course. You may guess or derive it; a derivation is not required. Then look up the equation in the first edition of Sutton and Barto's *Reinforcement Learning: An Introduction* at `http://incompleteideas.net/book/first/ebook/node35.html`. Write the correct equation in our notation. You may use AI to help translate the notation. State whether your original expression was correct. If it was not, explain the difference. Include both your original expression and the correct equation, even if they agree.

   Notation: Their Q^* is our q^*, and their $\mathcal{P}^a_{ss'}$ is our $p(s, a, s')$. Their $\mathcal{R}^a_{ss'}$ is the expected reward conditioned on the next state being s', so our reward function is $R(s, a) = \sum_{s' \in \mathcal{S}} p(s, a, s')\mathcal{R}^a_{ss'}$.

3. (14 Points) Let π be a policy such that q^π satisfies the Bellman optimality equation from Question 2. Prove that $q^\pi = q^*$.

4. (6 Points) Give an example of an MDP and policy for which the Dynamic Programming Policy Evaluation algorithm presented in class, initialized with $v_0(s) = 0$ for all $s \in \mathcal{S}$, does not reach v^π after any finite number of iterations, but does converge to v^π in the limit. Specify all components of your MDP and the policy. Prove that the algorithm does not converge to v^π in a finite number of iterations, and also prove that it converges to v^π in the limit as the number of iterations goes to infinity. **Hint:** To make the proofs simple, try to make the MDP have as few states and actions as possible.

5. (6 Points) Once you've completed Question 1, look up the Bellman equation for q^π (e.g., ask an AI tool, or check in Sutton and Barto's book). Ensuring that you have the correct form of the Bellman equation for q^π, use it to create a dynamic programming policy evaluation update for estimating q^π. Use q_i to denote the current estimate and q_{i+1} to denote the next estimate. You do not need to prove convergence.

6. (18 Points) Prove the following stochastic policy improvement theorem. Let π and π' be policies, and suppose that for every $s \in \mathcal{S}$,

$$\sum_{a \in \mathcal{A}} \pi'(s, a) q^\pi(s, a) \geq v^\pi(s). \tag{1}$$

   Prove that $v^{\pi'}(s) \geq v^\pi(s)$ for every $s \in \mathcal{S}$. Your proof must allow π' to be stochastic.

Part Two: Programming (40 Points)

7. (40 Points) Implement Dynamic Programming Policy Evaluation for the state-value function in any programming language. Your program should read an MDP and a policy from two files, evaluate that policy, and write the resulting value function to an output file. Use the following file formats. 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:

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

   Policy file: The file has n lines and m numbers per line. Line $s+1$ contains $\pi(s,0), \ldots, \pi(s,m-1)$. These are the action probabilities themselves, not parameters to transform using a softmax.

   Output file: Write n lines, with one number per line: $\widehat{v}(0), \ldots, \widehat{v}(n-1)$, in that order. Here $\widehat{v}$ is the estimate of v^π computed using DP policy evaluation. Write at least eight digits after the decimal point.

   Use double precision, initialize every state value to zero, and use the synchronous DP policy evaluation 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.

   First, test your program using `687_gridworld_MDP.txt` and `687_gridworld_policy.txt`, which describe 687-Gridworld and the uniform random policy. The file `687_Gridworld_values.txt` contains the correct output. **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. 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` and `test_policy.txt`. This MDP has seven states and ten actions; every state can transition to every state with positive probability under every action. The supplied policy is stochastic and is not uniform. In your PDF answer, report your estimates of $v^\pi(0), \ldots, v^\pi(6)$, giving at least six digits after the decimal point. You do not need to submit your source code or output files, or report the gridworld values.