HW 1: MDPs, Policy Iteration, and Value Iteration

EECE 571N – Sequential Decision-Making (EECE 571N)

Instructor: Cyrus Neary
Due: 2025-09-29 at 23:59 PT

Name: Dominic Klukas **Student Number:** 64348378

Instructions

Submit a single PDF to Canvas. Please include your name and student number at the top of that PDF. For all questions, show your work and clearly justify your steps and thinking. Please feel free to include any code as an attachment at the end of the PDF. State any assumptions. Unless otherwise specified, you may collaborate conceptually but must write up your own solutions independently.

Grading

Points for each part are indicated. The total number of achievable points is 100. Partial credit is available for incorrect answers with clear reasoning.

Problem 1

$V^\pi(s)$ represents the expected reward that a policy π will earn in a Markov Decision Process if it starts in state s. $V^*(s)$ represents the expected reward that an optimal policy will earn in a Markov Decision Process if it starts in state s. Bellman's equation for $V^\pi(s)$ is given by

$$V^\pi(s) = \sum_{a \in A} \pi(a|s) \left(R(s,a) + \gamma \cdot \sum_{s' \in S} T(s'|s,a) V^\pi(s') \right).$$

Bellman's optimality equation for $V^*(s)$ is given by

$$V^*(s) = \max_{a \in A} \left(R(s,a) + \gamma \cdot \sum_{s' \in S} T(s'|s,a) V^*(s') \right).$$

Problem 2

Now, suppose V and W are any two functions of s. Let π be any policy. First, we compute:

$$\begin{aligned}
(T_\pi V - T_\pi W)(s) &= \sum_{a \in A} \pi(a|s) \left(R(s,a) + \gamma \sum_{s' \in S} T(s'|s,a) V(s') \right) \\
&\quad - \sum_{a \in A} \pi(a|s) \left(R(s,a) + \gamma \sum_{s' \in S} T(s'|s,a) W(s') \right) \\
&= \sum_{a \in A} \pi(a|s) \left(\gamma \sum_{s' \in S} T(s'|s,a) (V(s') - W(s')) \right).
\end{aligned}$$

From the definition of the $\|\cdot\|_\infty$ norm as $\sup_{s\in S}|f(s)|$, and applying the traingle inequality, we have

$$\begin{aligned}\|T_\pi V - T_\pi W\|_\infty &= \sup_{s\in S}|T_\pi V(s) - T_\pi W(s)| \\ &\le \sup_{s\in S}\sum_{a\in A}\gamma\cdot\pi(a|s)\sum_{s'\in S}T(s'|s,a)\left|V(s') - W(s')\right| \text{ by triangle inequality} \\ &\le \gamma\cdot\sup_{s\in S}\sum_{a\in A}\pi(a|s)\sum_{s'\in S}T(s'|s,a)\sup_{s''\in S}|V(s'') - W(s'')|.\end{aligned}$$

But then, $\sup_{s''\in S}|V(s'') - W(s'')| = \|V - W\|_\infty$ which is just a constant, and $\sum_{s'\in S}T(s'|s,a) = 1$ since it is a probability distribution. Likewise, $\sum_{a\in A}\pi(a|s) = 1$. Putting these together:

$$\begin{aligned}\|T_\pi V - T_\pi W\|_\infty &\le \gamma\cdot\sup_{s\in S}\sum_{a\in A}\pi(a|s)\sum_{s'\in S}T(s'|s,a)\|V - W\|_\infty \\ \|T_\pi V - T_\pi W\|_\infty &\le \gamma\cdot\|V - W\|_\infty\cdot\sup_{s\in S}\sum_{a\in A}\pi(a|s)\cdot 1 \\ \|T_\pi V - T_\pi W\|_\infty &\le \gamma\cdot\|V - W\|_\infty\cdot\sup_{s\in S}(1) = \gamma\cdot\|V - W\|_\infty,\end{aligned}$$

as desired.

Problem 3

The statement of the Banach fixed point theorem, essentially verbatim from Wikipedia with small modifications to fit our notations, is as follows:

Let (X, d) be a non-empty complete metric space with a γ-contraction mapping $T : X \to X$. Then T admits a unique fixed-point x^* in X (i.e. $T(x^*) = x^*$). Furthermore, x^* can be found as follows: start with an arbitrary element $x_0 \in X$ and define a sequence $(x_n)_{n\in\mathbb{N}}$ by $x_n = T(x_{n-1})$ for $n \ge 1$. Then, $\lim_{n\to\infty} x_n = x^*$. Furthermore, the speed of convergence is bounded in the sense that $d(x^*, x_n) \le \frac{\gamma^n}{1-\gamma}d(x_1, x_0)$.

Problem 4

In order to apply the Banach fixed point theorem on an operator T, we have to make sure (X, d) which it is acting on is a metric space, and with that, a complete metric space. In our case, X is the set of functions $V : S \to \mathbb{R}$, and d is $\|\cdot\|_\infty$. Since S is finite, $V : S \to \mathbb{R}$ is clealry isomorphic to $\mathbb{R}^{|S|}$. But then:

Proof. Let $\{x_n\}_{n\in\mathbb{N}} \in \mathbb{R}^d$ be a Cauchy sequence with respect to the $\|\cdot\|_\infty$ norm. We prove that this sequence is also a pointwise Cauchy sequence. Fix some index i. Let $\varepsilon > 0$. Then, there exists some N such that for all $n, m > N$, we have $\|x_n - x_m\|_\infty < \varepsilon$. By the definition of the ∞-norm, it follows that for each index $1 \le j \le N$, $|(x_n)_j - (x_m)_j| < \varepsilon$. In particular, $\{(x_n)_i\}_{n\in\mathbb{N}}$ is a Cauchy sequence. We proved in Math 320 that Cauchy sequences converge in $\mathbb{R}$: the proof works by showing that first, the sequences are bounded. Next, every bounded sequence has a convergent subsequence (this result uses compactness), and finally, that the whole sequence

converges to the same limit as this subsequence. This pointwise convergence, $\{(x_n)_i\}_{n\in\mathbb{N}} \to x_i \in \mathbb{R}$, implies $x_n = ((x_n)_1, \ldots, (x_n)_d) \to (x_1, \ldots, x_d) \in \mathbb{R}^d$, as desired. □

Furthermore, $V(s) = 0$ is a valid function in our space, so X is non-empty.

Finally T_π is indeed a γ-contraction mapping, as we proved in Question 2. Therefore, the requirements of the Banach fixed point theorem are met and we can use the theorem: a unique fixed point exists. That is, there exists a function $V : S \to \mathbb{R}$ such that

$$T_\pi V(s) = \sum_{a\in A} \pi(a|s) \left(R(s,a) + \gamma \sum_{s'\in S} T(s'|s,a)V(s') \right)$$

(V is a fixed point of T_π), and V is the limit of any sequence given by $V_{n+1} = T_\pi V_n$ initialized by any bounded function $V_0 : S \to \mathbb{R}$.

However, the equation satisfied by this fixed point is precisely the equation satisfied by the value function V^π for policy π, so we conclude that the value function is the function that the contraction mapping converges to, regardless of the initial function V_0 in the sequence.

Problem 5

We define each component of the MDP in turn.

- S

  We define the state by the Robot's location on the gridworld. Every location is possible except for the impassable wall. We enumerate the states by their co-ordinates, where n is the row from the top and m is the column from the left. Thus,

  $$S = \{(n,m)\}_{\substack{1\le n\le 5 \\ 1\le m\le 4}} \setminus \{(2,3)\}.$$

- A

  At each state, we have the set of actions $\{\text{left}, \text{right}, \text{up}, \text{down}\}$. To make these easier to work with, define $A = \{(-1,0),(1,0),(0,1),(0,-1)\}$. Note: actions that result in moving into the wall are allowed, as specified in the question.

- T

  For T, we define the set $N(s) = \{s+(1,0), s+(0,1), s+(-1,0), s+(0,-1)\} \cap S$. In other words, this is the set of valid "neighbors" in the set S that the robot can move to from s.

  $$T(s'|s,a) = \begin{cases} 1 & \text{if } s' = s \text{ and } s \in \{(5,4),(5,1)\} \\ 0 & \text{if } s' \neq s \text{ and } s \in \{(5,4),(5,1)\} \\ (1-p) & \text{if } s' = s+a \text{ and } s' \in N(s), s \notin \{(5,4),(5,1)\} \\ (1-p) & \text{if } s' = s \text{ and } s+a \notin N(s), s \notin \{(5,4),(5,1)\} \\ p/|N(s)| & \text{if } s' \in N(s), s \notin \{(5,4),(5,1)\} \\ 0 & \text{else} \end{cases}.$$

  The first two lines describe the case where s is in one of the sink states. The third line describes the case when s is not in a sink state, and the robot wants to move to a valid neighbor. The

fourth line describes the case when s is not in a sink state and the robot is trying to move into a wall. The fifth line describes the case when s is not in a sink state and the robot is slipping. The last case is a catch all case.

- γ

  Since this is a finite horizon MDP we set $\gamma = 1$.

- R

  We can express the reward function for the problem in the form $R(s, a)$ with the following expression:

$$R(s, a) = \sum_{s' \in S} T(s'|s, a) R(s, a, s').$$

  Then, we define $R(s, a, s')$:

$$R(s, a, s') = \begin{cases} 1 & \text{if } s' = (1, 5) \\ -1 & \text{if } s' = (4, 5) \\ 0 & \text{if } s \in \{(1, 4), (4, 4)\} \\ -0.1 & \text{else} \end{cases}.$$

- μ

  This is the initial state function. We put the robot at its starting position, so that we have:
  $\mu(s) = \begin{cases} 1 & \text{if } s = (1, 4) \\ 0 & \text{else} \end{cases}$

Problem 6

Credit: I wrote all of the code myself, except for the code to generate the figures. See page 5.

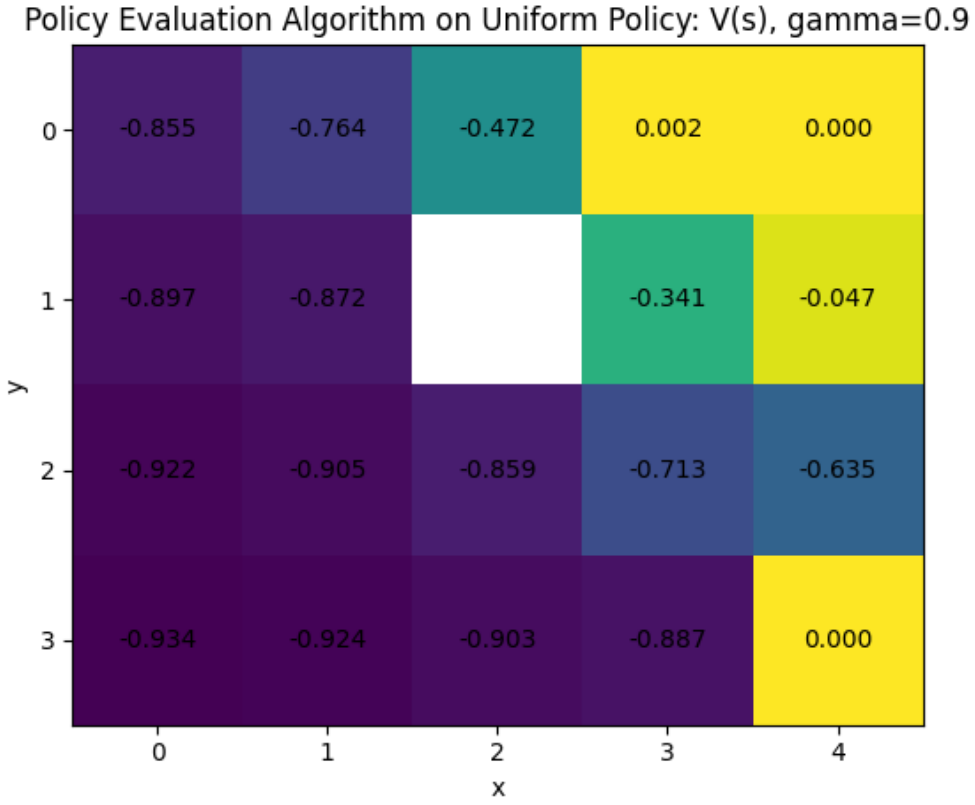

Figure 1: Heatmap for the computed value function V^{π} from the Policy Evaluation Iteration algorithm. π is the policy that assigns a uniform probability distribution over the set of actions at each state. Since the time horizon is infinite, we chose $\gamma = 0.9$.

Problem 7

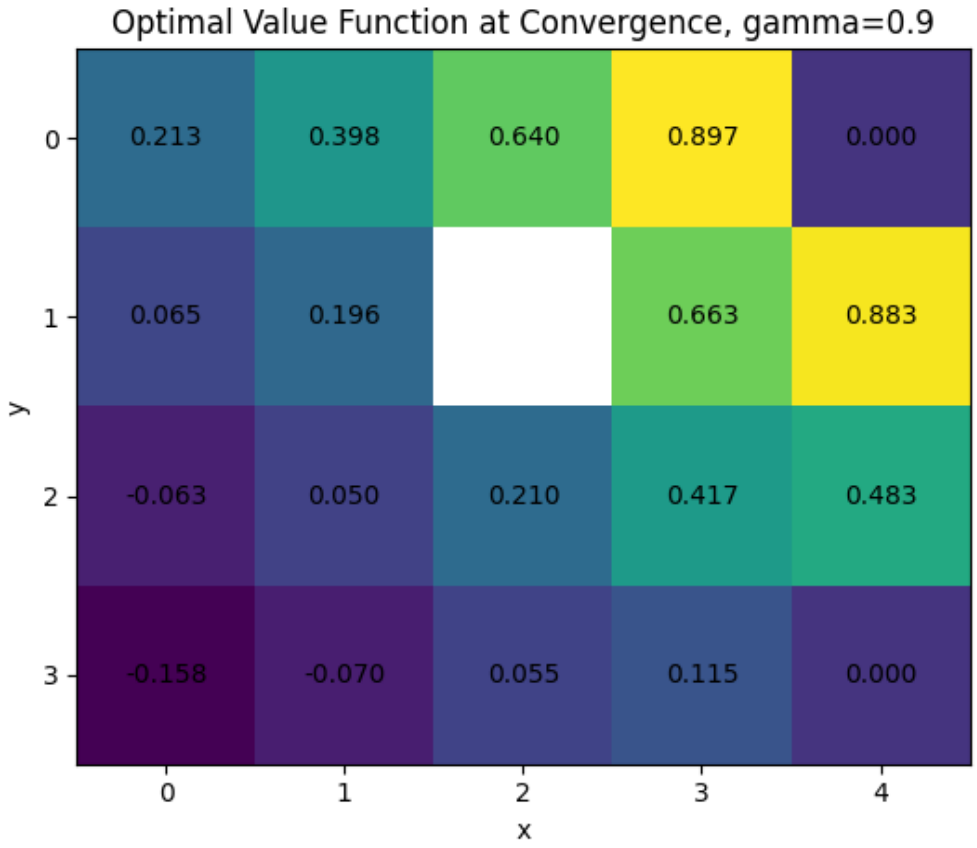

Figure 2: Heatmap for the the converged optimal value function V^* from the policy iteration algorithm, where the initial policy is π from Problem 6 and the initial $V(s) = 0$. Since the time horizon is technically infinite, we chose $\gamma = 0.9$.