ECE433/COS435 Introduction to RL
Assignment 3: Value Iteration and Policy Gradient
Spring 2024

Fill me in

Your name here.

Due February 26, 2024

Collaborators

Fill me in

Please fill in the names and NetIDs of your collaborators in this section.

Instructions

Writeups should be typesetted in Latex and submitted as PDFs. You can work with whatever tool you like for the code, but **please submit the asked-for snippet and answer in the solutions box as part of your writeup. We will only be grading your write-up.** Make sure still also to attach your notebook/code with your submission.

Question 1. Finite-state Value Function

An agent is navigating a very simple environment structured as a straight path with three states labeled 1, 2, and 3, where state 3 is a terminal state. At each step, the agent can choose to move to the next state or stay in the current state. Assume the discount factor = 0.5. The actions available in each state are:

- In state 1: the agent can either “move” to state 2 or “stay” in state 1.
- In state 2: the agent can “move” to state 3, ”stay” in state 2, or “move back” to state 1.

The rewards are as follows:

- Moving from state 1 to state 2 gives a reward of -1.

- Moving from state 2 to state 3 gives a reward of 0.
- Moving back from state 2 to state 1 gives a reward of -2.
- Staying in state 1 or state 2 gives a reward of -1.

Question 1.a

Calculate the value function for each state when the agent always decides to ”move” to the next state when possible.

Solution

Your solution here...

Question 1.b

Calculate the value function at each state when the agent always chooses to move to state 2 when in state 1, and always chooses to move back to state 1 when in state 2.

Solution

Your solution here...

Question 2. Properties of value functions

In this question, we consider an infinite horizon MDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, r, p, \gamma)$. We aim to derive some interesting properties for its value function.

Question 2.a

Show that a policy π is optimal if and only if its corresponding value functions $V^\pi : \mathcal{S} \to \mathbb{R}$ and $Q^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ satisfies $V^\pi(s) \geq Q^\pi(s, a)$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.

Solution

Your solution here...

Question 2.b

Is the optimal policy of an MDP unique? Please answer by proof or a counter-example.

Solution

Your solution here...

Question 2.d

Suppose that $\mathcal{M}$ has no terminating state. The agent will work forever. Now someone decides to add a small reward bonus c to all transitions in the MDP, which results in a new reward $r'(s,a) = r(s,a) + c$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Note that r is the original reward function. Could this change the optimal value function of $\mathcal{M}$?

Solution

Your solution here...

Question 2.e

If $\mathcal{M}$ is allowed to have terminating states, does the answer in Question 2.d still hold? If not, provide an example MDP where your answers to Question 2.d and this one differs.

Solution

Your solution here...

Question 3. Bellman residue [Optional]

This problem will not be graded, but we encourage those who are interested in the theoretical aspect of reinforcement learning to try it out. In the lecture, we introduce the (optimal) Bellman operator for an infinite horizon MDP with discount factor γ and transition p:

$$(\mathbb{B}V)(s) = \max_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \sum_{s' \in \mathcal{S}} p(s'|s,a) V(s') \right\},$$

and the Bellman operator with respect to a certain policy π:

$$(\mathbb{B}^{\pi}V)(s) = \sum_{a \in \mathcal{A}} r(s,a)\pi(a|s) + \gamma \sum_{s' \in \mathcal{S}, a \in \mathcal{A}} p(s'|s,a)\pi(a|s)V(s').$$

We denote the optimal policy by π^* and the optimal value function by V^*. As we know from the lecture, learning V^* is equivalent to solving the following Bellman equation:

$$V(s) - \mathbb{B}V(s) = 0, \forall s \in \mathcal{S}.$$

For an arbitrary function $V : \mathcal{S} \to \mathbb{R}$, define the Bellman residual to be $(\mathbb{B}V - V)$, and its magnitude by $\|(\mathbb{B}V - V)\|_{\infty}$. Recall that for a vector $x = (x_i)_{i \in [d]}$, $\|\cdot\|_{\infty}$ is defined by $\max_{i \in [d]} |x_i|$. As we will see through the course, this Bellman residual is an important component of several important RL algorithms such as the Deep Q-network.

Question 3.1

Prove the following statements for an arbitrary $V : \mathcal{S} \to \mathbb{R}$ (not necessarily a value function for any policy):

$$\begin{aligned} \|V - V^\pi\|_\infty &\le \frac{\|V - \mathbb{B}^\pi V\|_\infty}{1-\gamma}, \\ \|V - V^*\|_\infty &\le \frac{\|V - \mathbb{B} V\|_\infty}{1-\gamma}. \end{aligned}$$

(Hint: use Bellman equation to expand V^π, then apply triangle inequality.)

Solution

Your solution here...

Question 3.2

Now let's assume that π is the greedy policy extracted from V, and assume $\|V - \mathbb{B}V\|_\infty \le \epsilon$. Prove the following inequality by utilizing the results in Question 3.1:

$$V^* - V^\pi \le \frac{2\epsilon}{1-\gamma}.$$

This shows that as long as the Bellman residue of V is small, then the policy learned from V will be not too bad.

Solution

Your solution here...

Question 4. Coding

Policy Improvement

Paste the **entire cell** implementing value iteration below.

Solution

```
# YOUR CODE HERE!
```

Value Iteration

Paste the **entire cell** implementing value iteration below.

Solution

```
# YOUR CODE HERE!
```

Policy Gradient

Please paste the **entire cell** implementing value iteration below. Also, plot the curve of average cumulative reward v.s. training episodes with `matplotlib.pyplot`, and insert it below.

Solution

```
# YOUR CODE AND FIGURE HERE!
```