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

Fill me in

Your name here.

Due February XX, 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 $\gamma = 0.5$. 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

The value function for state $1 = -1 + 0.5 * 0 = -1$. The value function for state $2 = 0$. The value function for state $3 = 0$.

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 choose to move back to state 1 when in state 2.

Solution

Denote value function on state 1 by v_1, and value function on state 2 by v_2. Then we have the following Bellman equation:

$$v_1 = -1 + 0.5 * v_2$$
$$v_2 = -2 + 0.5 * v_1,$$

solve these equations and we get $v_1 = -\frac{8}{3}$, $v_2 = -\frac{10}{3}$.

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

By the definition of value function, we have

$$V^\pi(s) = \sum_a Q^\pi(s, a)\pi(a|s),$$

therefore since $V^\pi(s) \geq Q^\pi(s,a)$ for all (s,a), $\pi(a|s) \neq 0$ must implies $Q^\pi(s,a) = V^\pi(s)$, therefore $V^\pi(s) = \text{argmax}_a Q(s,a)$ holds for all $s \in \mathcal{S}$. By Bellman equation, π is the optimal policy.

Question 2.b

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

Solution

No it's not. For example, consider a contextual bandit with only one state s and action space $\mathcal{A} = \{a_1, \ldots, a_n\}$, and $r(s, a_i) = c$ for all $a_i \in \mathcal{A}$. In this case, any policy is optimal

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

No it cannot, because the $Q^*(s,a)$ is only changed by a constant $\frac{c}{1-\gamma}$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and this does not change the optimal policy.

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 differ.

Solution

No it doesn't. For example, consider an MDP with two states s_1, s_2, two actions a_1 and a_2. s_2 is the terminating state. $\gamma = 1/2$. The agent always starts at s_1. If she chooses a_1, then she will stay in s_1, otherwise, she will be transferred to s_2 and the MDP terminates. $r(s_1, a_1) = x_1$ and $r(s_2, a_2) = x_2$. In this case, if the agent always chooses a_1 under s_1 (denoted by policy π_1), her value function will be $2x_1$, and if she always chooses a_2 (denoted by policy π_2) her value function will be x_2. Set $x_1 = \frac{1}{4}$, $x_2 = 1$, then her optimal policy would be always choosing a_2. However, if we further select $c = 1$, then value function for π_1 becomes $2(1 + \frac{1}{4}) = 2.5$ and the value function for π_2 becomes $1 + 1 = 2$, therefore her optimal policy should always be choosing a_1, which is different.

Question 3. Bellman residue

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^π is equivalent to solving the following Bellman equation:

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

For an arbitrary function $V : \mathcal{S} \to \mathbb{R}$, define the Bellman residual to be $(\mathbb{B}V - V)$, and its magnitude by $\|(\mathbb{B}V - V)\|_\infty$. 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):

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

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

Solution

We only need to prove $\|V - V^\pi\| \le \frac{\|V-\mathbb{B}^\pi V\|_\infty}{1-\gamma}$. First note that $V^\pi = \mathbb{B}^\pi V^\pi$, we have

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

and the conclusion follows. Here in the second inequality, we use the following inequality:

$$\begin{aligned}\|\mathbb{B}^\pi V - \mathbb{B}^\pi V^\pi\|_\infty &= \gamma\| \sum_a p(s'|s,a)\big(V(s') - V^\pi(s')\big)\|_\infty \\ &\le \gamma\|V - V^\pi\|_\infty.\end{aligned}$$

Similarly, utilizing $|\max_a f_1 - \max_a f_2| \leq \max_a |f_1 - f_2|$, we have

$$\begin{aligned}\|\mathbb{B}V - \mathbb{B}V^*\|_\infty &= \|\max_a\{r(s,a) + \gamma \sum_{s'} p(s'|s,a)V(s')\} - \max_a\{r(s,a) + \gamma \sum_{s'} p(s'|s,a)V^*(s')\}\|_\infty \\ &\leq \gamma \|\max_a\{\sum_{s'} p(s'|s,a)(V(s') - V^*(s'))\}\|_\infty \\ &\leq \gamma \|V - V^*\|_\infty,\end{aligned}$$

we have

$$\|V - V^*\|_\infty \leq \|V - \mathbb{B}V\|_\infty + \|\mathbb{B}V - \mathbb{B}V^*\|_\infty \leq \|V - \mathbb{B}V\|_\infty + \gamma\|V - V^*\|_\infty,$$

solving this inequality and we conclude the proof.

Question 3.2

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

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

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

Note that in this case, $\mathbb{B}^\pi V = \mathbb{B}V$. Therefore

$$\|V^* - V^\pi\|_\infty \leq \|V^* - V\|_\infty + \|V - V^\pi\|_\infty \leq \frac{\|V - \mathbb{B}V\|_\infty}{1-\gamma} + \frac{\|V - \mathbb{B}^\pi V\|_\infty}{1-\gamma} \leq \frac{2\epsilon}{1-\gamma}.$$

Question 4. Coding