CS234: Reinforcement Learning – Problem Session #1

Spring 2023-2024

Problem 1

Suppose we have an infinite-horizon, discounted MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$ with a finite state-action space, $|\mathcal{S} \times \mathcal{A}| < \infty$ and $0 \leq \gamma < 1$. For any two arbitrary sets $\mathcal{X}$ and $\mathcal{Y}$, we denote the class of all functions mapping from $\mathcal{X}$ to $\mathcal{Y}$ as $\{\mathcal{X} \to \mathcal{Y}\} \triangleq \{f \mid f : \mathcal{X} \to \mathcal{Y}\}$. In the questions that follow, let $Q, Q' \in \{\mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$ be any two arbitrary action-value functions and consider any fixed state $s \in \mathcal{S}$. Without loss of generality, you may assume that $Q(s,a) \geq Q'(s,a)$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.

1. Prove that $|\max\limits_{a \in \mathcal{A}} Q(s,a) - \max\limits_{a' \in \mathcal{A}} Q'(s,a')| \leq \max\limits_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|$.

2. Prove that $|\min\limits_{a \in \mathcal{A}} Q(s,a) - \min\limits_{a' \in \mathcal{A}} Q'(s,a')| \leq \max\limits_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|$.

3. Prove that $\left| \frac{1}{|\mathcal{A}|} \sum\limits_{a \in \mathcal{A}} Q(s,a) - \frac{1}{|\mathcal{A}|} \sum\limits_{a' \in \mathcal{A}} Q'(s,a') \right| \leq \max\limits_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|$.

4. Prove that, for any parameter $\omega \in \mathbb{R}$,[1]

$$\left| \frac{1}{\omega} \log \left(\frac{1}{|\mathcal{A}|} \sum_{a \in \mathcal{A}} \exp\left(\omega \cdot Q(s,a)\right) \right) - \frac{1}{\omega} \log \left(\frac{1}{|\mathcal{A}|} \sum_{a' \in \mathcal{A}} \exp\left(\omega \cdot Q'(s,a')\right) \right) \right| \leq \max_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|.$$

Hint: define and introduce $\Delta(a) = Q(s,a) - Q'(s,a)$ for $a \in \mathcal{A}$.

[1] For any $x \in \mathbb{R}$, $\exp(x) = e^x$ and all logarithms are base e.

The remainder of this question focuses on Algorithm 1, which takes as input an operator

$$\bigotimes : \{\mathcal{S} \times \mathcal{A} \to \mathbb{R}\} \to \{\mathcal{S} \to \mathbb{R}\}$$

that adheres to the following property[2]:

$$||\bigotimes Q - \bigotimes Q'||_\infty \leq ||Q - Q'||_\infty, \qquad \forall Q, Q' \in \{\mathcal{S} \times \mathcal{A} \to \mathbb{R}\}. \tag{1}$$

Algorithm 1:

Data: Finite MDP $\mathcal{M}$, Operator $\bigotimes$ satisfying Equation 1
Initialize $V_0(s) = 0, \forall s \in \mathcal{S}$ ▷ Initial value function estimate
Initialize $k = 1$ ▷ Iteration counter
while *not converged* **do**
  for *each state* $s \in \mathcal{S}$ **do**
    $V_k(s) = \bigotimes_{a \in \mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s' \mid s, a) V_{k-1}(s') \right).$
  end
  $k = k + 1$
end
Return V_k

5. For any value function $V \in \{\mathcal{S} \to \mathbb{R}\}$, define the operator $\mathcal{B} : \{\mathcal{S} \to \mathbb{R}\} \to \{\mathcal{S} \to \mathbb{R}\}$ as follows:

$$\mathcal{B}V(s) = \bigotimes_{a \in \mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s' \mid s, a) V(s') \right),$$

where $\bigotimes$ satisfies Equation 1. Prove that $\mathcal{B}$ is a γ-contraction with respect to the L_∞-norm.

[2] As always, $|| \cdot ||_\infty$ denotes the L_∞-norm.

6. Let $\bigotimes_1, \bigotimes_2 : \{\mathcal{S} \times \mathcal{A} \to \mathbb{R}\} \to \{\mathcal{S} \to \mathbb{R}\}$ be two operators satisfying Equation 1. Prove that, for any $0 \leq \lambda \leq 1$,

$$\bigotimes_{\lambda} = \lambda \bigotimes_{1} + (1 - \lambda) \bigotimes_{2}$$

also satisfies Equation 1.

7. For any $0 \leq \varepsilon \leq 1$, define your own operator $\bigotimes_{\varepsilon} : \{\mathcal{S} \times \mathcal{A} \to \mathbb{R}\} \to \{\mathcal{S} \to \mathbb{R}\}$ and prove that running Algorithm 1 with your $\bigotimes_{\varepsilon}$ returns the value function associated with the ε-greedy optimal policy (where the optimal policy maximizes the expected sum of future discounted rewards).