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 $Q(s,a) \geq Q'(s,a)$, $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$.
Solution: The first three parts of this question are proven simultaneously and in more generality via Theorem 8 of Littman and Szepesvári [1996].

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

   Solution: We can start by simply ignoring the absolute value signs on the left-hand side. Let $a^\star = \arg\max\limits_{a \in \mathcal{A}} Q(s,a)$. Then,

   $$\begin{aligned}\max_{a \in \mathcal{A}} Q(s,a) - \max_{a' \in \mathcal{A}} Q'(s,a') &= Q(s,a^\star) - \max_{a' \in \mathcal{A}} Q'(s,a') \\ &\leq Q(s,a^\star) - Q'(s,a^\star) \\ &\leq \max_{a \in \mathcal{A}} \left(Q(s,a) - Q'(s,a)\right) \\ &\leq \max_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|.\end{aligned}$$

   Now, take absolute values on both sides of the inequality (the left-hand side is already non-negative) to get
   $$|\max_{a \in \mathcal{A}} Q(s,a) - \max_{a' \in \mathcal{A}} Q'(s,a')| \leq \max_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|.$$

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

   Solution: We can start by simply ignoring the absolute value signs on the left-hand side. Let $a^\star = \arg\min\limits_{a' \in \mathcal{A}} Q'(s,a')$. Then,

   $$\begin{aligned}\min_{a \in \mathcal{A}} Q(s,a) - \min_{a' \in \mathcal{A}} Q'(s,a') &= \min_{a \in \mathcal{A}} Q(s,a) - Q'(s,a^\star) \\ &\leq Q(s,a^\star) - Q'(s,a^\star) \\ &\leq \max_{a \in \mathcal{A}} \left(Q(s,a) - Q'(s,a)\right) \\ &\leq \max_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|.\end{aligned}$$

   Now, take absolute values on both sides of the inequality (the left-hand side is already non-negative) to get
   $$|\min_{a \in \mathcal{A}} Q(s,a) - \min_{a' \in \mathcal{A}} Q'(s,a')| \leq \max_{a \in \mathcal{A}} |Q(s,a) - Q'(s,a)|.$$

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

   Solution: We can start by simply ignoring the absolute value signs on the left-hand side.

$$\begin{aligned}
\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} Q(s,a) - \frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} Q'(s,a') &= \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \left(Q(s,a) - Q'(s,a)\right) \\
&\leq \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} |Q(s,a) - Q'(s,a)| \\
&\leq \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \max_{a'\in\mathcal{A}} |Q(s,a') - Q'(s,a')| \\
&= \frac{1}{|\mathcal{A}|} \cdot |\mathcal{A}| \cdot \max_{a'\in\mathcal{A}} |Q(s,a') - Q'(s,a')| \\
&= \max_{a\in\mathcal{A}} |Q(s,a) - Q'(s,a)|.
\end{aligned}$$

   Now, take absolute values on both sides of the inequality (the left-hand side is already non-negative) to get

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

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 $a \in \mathcal{A}$.

   Solution: This is the so-called mellowmax operator introduced by Asadi and Littman [2017] which, unlike the Boltzmann softmax operator (see Lemma C.3 of Littman [1996]), obeys the stated property. Let $\Delta(a) = Q(s,a) - Q'(s,a)$

$$\begin{aligned}
\left|\frac{1}{\omega}\log\left(\frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \exp\left(\omega \cdot Q(s,a)\right)\right) - \frac{1}{\omega}\log\left(\frac{1}{|\mathcal{A}|}\sum_{a'\in\mathcal{A}} \exp\left(\omega \cdot Q'(s,a')\right)\right)\right| &= \left|\frac{1}{\omega}\log\left(\frac{\sum\limits_{a\in\mathcal{A}} \exp\left(\omega \cdot Q(s,a)\right)}{\sum\limits_{a'\in\mathcal{A}} \exp\left(\omega \cdot Q'(s,a')\right)}\right)\right| \\
&= \left|\frac{1}{\omega}\log\left(\frac{\sum\limits_{a\in\mathcal{A}} \exp\left(\omega \cdot \left(Q'(s,a) + \Delta(a)\right)\right)}{\sum\limits_{a'\in\mathcal{A}} \exp\left(\omega \cdot Q'(s,a')\right)}\right)\right| \\
&\leq \left|\frac{1}{\omega}\log\left(\frac{\sum\limits_{a\in\mathcal{A}} \exp\left(\omega \cdot \left(Q'(s,a) + \max\limits_{a'\in\mathcal{A}} \Delta(a')\right)\right)}{\sum\limits_{a'\in\mathcal{A}} \exp\left(\omega \cdot Q'(s,a')\right)}\right)\right| \\
&= \left|\frac{1}{\omega}\log\left(\exp\left(\omega \cdot \max_{a'\in\mathcal{A}} \Delta(a')\right)\frac{\sum\limits_{a\in\mathcal{A}} \exp\left(\omega \cdot Q'(s,a)\right)}{\sum\limits_{a'\in\mathcal{A}} \exp\left(\omega \cdot Q'(s,a')\right)}\right)\right| \\
&= \left|\frac{1}{\omega}\log\left(\exp\left(\omega \cdot \max_{a'\in\mathcal{A}} \Delta(a')\right)\right)\right| \\
&= |\max_{a\in\mathcal{A}} \Delta(a)| \\
&\leq \max_{a\in\mathcal{A}} |Q(s,a) - Q'(s,a)|.
\end{aligned}$$

[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]:

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

Solution: Equation 1 is known as the *non-expansion* property and all operators $\bigotimes$ which obey this property are known as *non-expansion operators*. Technically, the following convergence results also rely on $\bigotimes$ obeying the following conservative property, which all the above operators also satisfy but we didn't have you prove:

$$\min_{a \in \mathcal{A}} Q(s,a) \leq \bigotimes Q(s) \leq \max_{a \in \mathcal{A}} Q(s,a).$$

Algorithm 1: Solution: Generalized Value Iteration (GVI) [Littman and Szepesvári, 1996]

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.

Solution: Take any two value functions $V_1, V_2 \in \{\mathcal{S} \to \mathbb{R}\}$. Then,

$$\begin{aligned}
||\mathcal{B}V_1 - \mathcal{B}V_2||_\infty &= \max_{s \in \mathcal{S}} |\mathcal{B}V_1(s) - \mathcal{B}V_2(s)| \\
&= \max_{s \in \mathcal{S}} \left| \bigotimes_{a \in \mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s' \mid s,a) V_1(s') \right) - \bigotimes_{a \in \mathcal{A}} \left(\mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s' \mid s,a) V_2(s') \right) \right| \\
&\leq \max_{s,a \in \mathcal{S} \times \mathcal{A}} \left| \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s' \mid s,a) V_1(s') - \mathcal{R}(s,a) + \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s' \mid s,a) V_2(s') \right| \\
&= \max_{s,a \in \mathcal{S} \times \mathcal{A}} \left| \gamma \sum_{s' \in \mathcal{S}} \mathcal{T}(s' \mid s,a) \left[V_1(s') - V_2(s') \right] \right| \\
&\leq \max_{s,a \in \mathcal{S} \times \mathcal{A}} \gamma \left| \max_{s' \in \mathcal{S}} \left[V_1(s') - V_2(s') \right] \right| \\
&\leq \gamma \max_{s \in \mathcal{S}} |V_1(s) - V_2(s)| = \gamma ||V_1 - V_2||_\infty.
\end{aligned}$$

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

Therefore, we have shown that the generalized Bellman operator is a γ-contraction with respect to 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.

Solution: Take any $Q, Q' \in \{\mathcal{S} \times \mathcal{A} \to \mathbb{R}\}$. Then,

$$\begin{aligned}
||\bigotimes_\lambda Q - \bigotimes_\lambda Q'||_\infty &= \max_{s\in\mathcal{S}} \left| \bigotimes_\lambda Q(s) - \bigotimes_\lambda Q'(s) \right| \\
&= \max_{s\in\mathcal{S}} \left| \lambda \bigotimes_1 Q(s) + (1-\lambda) \bigotimes_2 Q(s) - \lambda \bigotimes_1 Q'(s) - (1-\lambda) \bigotimes_2 Q'(s) \right| \\
&= \max_{s\in\mathcal{S}} \left| \lambda \left(\bigotimes_1 Q(s) - \bigotimes_1 Q'(s) \right) + (1-\lambda) \left(\bigotimes_2 Q(s) - \bigotimes_2 Q'(s) \right) \right| \\
&\leq \max_{s\in\mathcal{S}} \left[\lambda \left| \bigotimes_1 Q(s) - \bigotimes_1 Q'(s) \right| + (1-\lambda) \left| \bigotimes_2 Q(s) - \bigotimes_2 Q'(s) \right| \right] \\
&\leq \lambda \max_{s\in\mathcal{S}} \left| \bigotimes_1 Q(s) - \bigotimes_1 Q'(s) \right| + (1-\lambda) \max_{s\in\mathcal{S}} \left| \bigotimes_2 Q(s) - \bigotimes_2 Q'(s) \right| \\
&= \lambda ||\bigotimes_1 Q - \bigotimes_1 Q'||_\infty + (1-\lambda) ||\bigotimes_2 Q - \bigotimes_2 Q'||_\infty \\
&\leq \lambda ||Q - Q'||_\infty + (1-\lambda) ||Q - Q'||_\infty = ||Q - Q'||_\infty .
\end{aligned}$$

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).

Solution: Define the non-expansion operators

$$\bigotimes_1 Q(s) = \frac{1}{|\mathcal{A}|} \sum_{a\in\mathcal{A}} Q(s,a) \qquad \bigotimes_2 Q(s) = \max_{a\in\mathcal{A}} Q(s,a).$$

A policy acting uniformly at random achieves the average Q-value over all actions at each state. Thus, $\bigotimes_1$ is the non-expansion operator associated with this uniform random policy whereas $\bigotimes_2$ corresponds to the usual definition of optimal policy that maximizes the Q-value at each state. Therefore, the ε-greedy optimal policy is formed by taking the convex combination:

$$\bigotimes_\varepsilon Q = \varepsilon \bigotimes_1 Q + (1-\varepsilon) \bigotimes_2 Q.$$

By parts (1) and (3) above, we know that $\bigotimes_1, \bigotimes_2$ are both non-expansion operators. Thus, by the previous part (6), we immediately have that $\bigotimes_\varepsilon$ is also a non-expansion operator implying that it is compatible with GVI. By part (5), we have that any non-expansion operator is a γ-contraction on value functions with respect to the L_∞-norm. Therefore, by the Banach Fixed-Point Theorem, we are guaranteed the existence of and the convergence of GVI to a unique fixed point.

References

Kavosh Asadi and Michael L. Littman. An alternative softmax operator for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 243–252, 2017.

Michael L. Littman. *Algorithms for Sequential Decision-Making*. PhD thesis, Brown University, 1996.

Michael L. Littman and Csaba Szepesvári. A generalized reinforcement-learning model: convergence and applications. In *Proceedings of the Thirteenth International Conference on International Conference on Machine Learning*, pages 310–318, 1996.