CS234: Reinforcement Learning – Problem Session #1

Winter 2021-2022

Problem 1

Suppose we have a MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$ and we know that the maximal reward we can observe in $\mathcal{M}$ is given by $R_{\text{MAX}} \triangleq \max\limits_{s,a} \mathcal{R}(s,a)$. The following questions focus on Algorithm 1 which assumes access to a sub-routine for running Value Iteration (`value_iteration`).

Algorithm 1:

Data: MDP $\mathcal{M}$, Threshold parameter $M \in \mathbb{N}$, Reward upper bound $R_{\text{MAX}} \in \mathbb{R}$
Initialize $N(s,a) = 0, \forall s,a \in \mathcal{S} \times \mathcal{A}$ ▷ Counter for state-action pair (s,a)
Initialize $N(s,a,s') = 0, \forall s,a,s' \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ ▷ Counter for transition (s,a,s')
Initialize $r(s,a) = 0, \forall s,a \in \mathcal{S} \times \mathcal{A}$ ▷ Total reward observed for state-action pair (s,a)
Initialize approximate reward function $\widehat{\mathcal{R}}(s,a) = R_{\text{MAX}}, \forall s,a \in \mathcal{S} \times \mathcal{A}$
Initialize approximate transition function $\widehat{\mathcal{T}}(s,a,s') = \mathbb{1}_{s=s'}, \forall s,a,s' \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$
Initialize approximate action-value function $\widehat{Q}^\star(s,a) = \frac{R_{\text{MAX}}}{(1-\gamma)}, \forall s,a \in \mathcal{S} \times \mathcal{A}$
for $t = 1,2,3,\ldots$ **do**
    Observe state s
    Take action $a = \arg\max\limits_{a' \in \mathcal{A}} \widehat{Q}^\star(s,a')$
    Observe reward r and next state s'
    $r(s,a) = r(s,a) + r$
    $N(s,a) = N(s,a) + 1$
    $N(s,a,s') = N(s,a,s') + 1$
    if $N(s,a) = M$ **then**
        $\widehat{\mathcal{R}}(s,a) = \frac{r(s,a)}{N(s,a)}$
        $\widehat{\mathcal{T}}(s,a,s') = \frac{N(s,a,s')}{N(s,a)}$
        $\widehat{Q}^\star =$ `value_iteration`$(\mathcal{S}, \mathcal{A}, \widehat{\mathcal{R}}, \widehat{\mathcal{T}}, \gamma)$
    end
end

1. Is Algorithm 1 a model-free or model-based reinforcement-learning algorithm? Provide a brief explanation of your answer.

2. Consider all of the unvisited state-action pairs in each timestep. Is the agent more likely or less likely to visit these state-action pairs as time passes? In other words, do you expect the total number of unvisited state-action pairs to increase or decrease as time passes? Provide a brief justification.

3. Consider the MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$ and the MDP $\widehat{\mathcal{M}} = \langle \mathcal{S}, \mathcal{A}, \widehat{\mathcal{R}}, \widehat{\mathcal{T}}, \gamma \rangle$. We will use subscripts to distinguish between arbitrary value functions $V_{\mathcal{M}}$ and $V_{\widehat{\mathcal{M}}}$ of MDPs $\mathcal{M}$ and $\widehat{\mathcal{M}}$, respectively. For simplicity, we will assume that $0 \leq V_{\mathcal{M}}(s) \leq 1$ and $0 \leq V_{\widehat{\mathcal{M}}}(s) \leq 1, \forall s \in \mathcal{S}$. If $\exists$ two constants $\varepsilon_1, \varepsilon_2 \geq 0$ such that

$$\max_{s,a \in \mathcal{S} \times \mathcal{A}} |\mathcal{R}(s,a) - \widehat{\mathcal{R}}(s,a)| \leq \varepsilon_1 \qquad \max_{s,a \in \mathcal{S} \times \mathcal{A}} \sum_{s' \in \mathcal{S}} |\mathcal{T}(s'|s,a) - \widehat{\mathcal{T}}(s'|s,a)| \leq \varepsilon_2,$$

then we know that for any policy $\pi : \mathcal{S} \to \mathcal{A}$, $||V^{\pi}_{\mathcal{M}} - V^{\pi}_{\widehat{\mathcal{M}}}||_{\infty} \leq \frac{\varepsilon_1 + \gamma\varepsilon_2}{(1-\gamma)}$. Discuss the importance of this result in the context of Algorithm 1. In particular, contrast running Algorithm 1 on $\mathcal{M}$ with $M = 1$ vs. $M = 100$.

4. Now, instead of assuming that we may freely represent any policy, let's account for the approximation error that we incur when we can only represent a subset of all policies. Let $\Pi = \{\pi \mid \pi : \mathcal{S} \to \mathcal{A}\}$ denote the set of all possible stationary policies and define $\overline{\Pi} \subseteq \Pi$ as some restricted subset of policies. Take $\mathcal{M}$ and $\widehat{\mathcal{M}}$ as defined in the previous part and let $\pi^{\star}_{\mathcal{M}}$ and $\pi^{\star}_{\widehat{\mathcal{M}}}$ denote the optimal policies for $\mathcal{M}$ and $\widehat{\mathcal{M}}$, respectively. Similarly, let $\rho^{\star}_{\mathcal{M}}$ and $\rho^{\star}_{\widehat{\mathcal{M}}}$ denote the optimal policies **in $\overline{\Pi}$** for $\mathcal{M}$ and $\widehat{\mathcal{M}}$, respectively. Show that for any state $s \in \mathcal{S}$

$$|V^{\pi^{\star}_{\mathcal{M}}}_{\mathcal{M}}(s) - V^{\rho^{\star}_{\widehat{\mathcal{M}}}}_{\mathcal{M}}(s)| \leq |V^{\pi^{\star}_{\mathcal{M}}}_{\mathcal{M}}(s) - V^{\rho^{\star}_{\mathcal{M}}}_{\mathcal{M}}(s)| + 2 \max_{\rho \in \overline{\Pi}} |V^{\rho}_{\mathcal{M}}(s) - V^{\rho}_{\widehat{\mathcal{M}}}(s)|.$$