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 $R_{\text{MAX}} \triangleq \max_{s,a} \mathcal{R}(s,a)$. 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 $a = \arg\max_{a' \in \mathcal{A}} \widehat{Q}^\star(s,a')$
  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**
    $\widehat{\mathcal{R}}(s,a) = \frac{r(s,a)}{N(s,a)}$
    $\widehat{\mathcal{T}}(s,a,s') = \frac{N(s,a,s')}{N(s,a)}$
    $\widehat{Q}^\star = \texttt{value\_iteration}(\mathcal{S}, \mathcal{A}, \widehat{\mathcal{R}}, \widehat{\mathcal{T}}, \gamma)$
  end
end

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

   Solution: Algorithm 1 is a model-based reinforcement-learning algorithm known as R-MAX [Brafman and Tennenholtz, 2002]. The way to see this is by noting how a R-MAX agent is gradually building a model of the MDP (that is, the transition function and reward function) in order to plan and recover an optimal policy.

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.

   Solution: As time passes, state-action pairs the agent has not yet tried become more likely since the reward attributed to these state-action pairs is set to be maximal (that is, R_{MAX}) until the state-action

pair has been tried M times. This general principle for exploration is known as *optimism in the face of uncertainty*.

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. $M = 100$.

Solution: The cited result is known as the simulation lemma [Kearns and Singh, 2002]. The simulation lemma tells us that when we can recover an accurate approximation to the true reward function and transition function of our MDP, the return obtained by policies in the approximate MDP will be close to those of the original MDP. In the R-MAX algorithm, the parameter M controls how long an agent waits before deciding it has observed enough information about each state-action pair in order to accurately estimate the associated reward and next-state transition distribution, based on maximum-likelihood estimates. In a stochastic MDP, $M = 1$ will lose out on valuable information in each state-action pair, leading to poor model estimates and poor policy performance. In contrast, $M = 100$ is likely to provide enough information for accurate model estimation; the simulation lemma then tells us that this accurate model will translate into accurate reasoning about $\mathcal{M}$ through $\widehat{\mathcal{M}}$.

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 $s \in \mathcal{S}$

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

Solution:

$$\begin{aligned}
V^{\pi^{\star}_{\mathcal{M}}}_{\mathcal{M}} - V^{\rho^{\star}_{\widehat{\mathcal{M}}}}_{\mathcal{M}} &= (V^{\pi^{\star}_{\mathcal{M}}}_{\mathcal{M}} - V^{\rho^{\star}_{\mathcal{M}}}_{\mathcal{M}}) + (V^{\rho^{\star}_{\mathcal{M}}}_{\mathcal{M}} - V^{\rho^{\star}_{\mathcal{M}}}_{\widehat{\mathcal{M}}}) - (V^{\rho^{\star}_{\widehat{\mathcal{M}}}}_{\mathcal{M}} - V^{\rho^{\star}_{\widehat{\mathcal{M}}}}_{\widehat{\mathcal{M}}}) - (V^{\rho^{\star}_{\widehat{\mathcal{M}}}}_{\widehat{\mathcal{M}}} - V^{\rho^{\star}_{\mathcal{M}}}_{\widehat{\mathcal{M}}}) \\
&\leq (V^{\pi^{\star}_{\mathcal{M}}}_{\mathcal{M}} - V^{\rho^{\star}_{\mathcal{M}}}_{\mathcal{M}}) + (V^{\rho^{\star}_{\mathcal{M}}}_{\mathcal{M}} - V^{\rho^{\star}_{\mathcal{M}}}_{\widehat{\mathcal{M}}}) - (V^{\rho^{\star}_{\widehat{\mathcal{M}}}}_{\mathcal{M}} - V^{\rho^{\star}_{\widehat{\mathcal{M}}}}_{\widehat{\mathcal{M}}}) \\
&\leq |V^{\pi^{\star}_{\mathcal{M}}}_{\mathcal{M}} - V^{\rho^{\star}_{\mathcal{M}}}_{\mathcal{M}}| + |V^{\rho^{\star}_{\mathcal{M}}}_{\mathcal{M}} - V^{\rho^{\star}_{\mathcal{M}}}_{\widehat{\mathcal{M}}}| + |V^{\rho^{\star}_{\widehat{\mathcal{M}}}}_{\mathcal{M}} - V^{\rho^{\star}_{\widehat{\mathcal{M}}}}_{\widehat{\mathcal{M}}}| \\
&\leq |V^{\pi^{\star}_{\mathcal{M}}}_{\mathcal{M}} - V^{\rho^{\star}_{\mathcal{M}}}_{\mathcal{M}}| + 2 \max_{\rho \in \overline{\Pi}} |V^{\rho}_{\mathcal{M}} - V^{\rho}_{\widehat{\mathcal{M}}}|
\end{aligned}$$

The claim follows since, by definition of $\pi^{\star}_{\mathcal{M}}$, $V^{\pi^{\star}_{\mathcal{M}}}_{\mathcal{M}} - V^{\rho^{\star}_{\widehat{\mathcal{M}}}}_{\mathcal{M}} \geq 0$. This claim was originally proven for a particular choice of $\overline{\Pi}$ based on the discount factor in Jiang et al. [2015] and was generalized to arbitrary restricted policy classes in Arumugam et al. [2018].

References

Dilip Arumugam, David Abel, Kavosh Asadi, Nakul Gopalan, Christopher Grimm, Jun Ki Lee, Lucas Lehnert, and Michael L Littman. Mitigating planner overfitting in model-based reinforcement learning. *arXiv preprint arXiv:1812.01129*, 2018.

Ronen I Brafman and Moshe Tennenholtz. R-MAX–a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3:213–231, 2002.

Nan Jiang, Alex Kulesza, Satinder Singh, and Richard Lewis. The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 International Conference on Autonomous Agents and Multiagent Systems*, pages 1181–1189. Citeseer, 2015.

Michael Kearns and Satinder Singh. Near-optimal reinforcement learning in polynomial time. *Machine learning*, 49(2-3):209–232, 2002.