CS234: Reinforcement Learning – Problem Session #6

Spring 2023-2024

Problem 1

Consider an infinite-horizon, discounted MDP $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma \rangle$. As usual, for any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, the value function induced by π is defined as

$$V^\pi(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathcal{R}(s_t, a_t) \mid s_0 = s, \pi\right].$$

1. For an arbitrary $Z \in \mathbb{N}$, consider learning with $Z+1$ distinct discount factors $\gamma_0, \gamma_1, \ldots, \gamma_Z$ where the final discount factor matches that of the MDP $\mathcal{M}$, $\gamma_Z = \gamma$. Letting $[Z] \triangleq \{1, 2, \ldots, Z\}$ denote the index set, we define the following functions for any policy π:

$$V^\pi_{\gamma_z} = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma_z^t \mathcal{R}(s_t, a_t) \mid s_0 = s, \pi\right] \qquad W^\pi_z = V^\pi_{\gamma_z} - V^\pi_{\gamma_{z-1}}, \qquad \forall z \in [Z]$$

where $W_0 = V^\pi_{\gamma_0}$.

Solution: The results of this part were derived by Romoff et al. [2019] who both empirically and theoretically study the benefits of decomposing a single monolithic value function across multiple time-scales through smaller discount factors.

(a) For any $z \in [Z]$; any policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$; and any $s \in \mathcal{S}$, write an expression for $V^\pi_{\gamma_z}(s)$ exclusively in terms of $\{W^\pi_0, W^\pi_1, \ldots, W^\pi_Z\}$.

Solution: From the relationships defined above, we can see that

$$V^\pi_{\gamma_z}(s) = \sum_{i=0}^{z} W^\pi_i(s).$$

(b) Show that W^π_z obeys the following Bellman equation for any $z \in [Z]$ and $s \in \mathcal{S}$:

$$W^\pi_z(s) = \mathbb{E}_{\substack{a \sim \pi(\cdot|s) \\ s' \sim \mathcal{T}(\cdot|s,a)}}\left[(\gamma_z - \gamma_{z-1}) V^\pi_{\gamma_{z-1}}(s') + \gamma_z W^\pi_z(s')\right]$$

Solution: Just by expanding the corresponding Bellman equations for $V^\pi_{\gamma_z}$ and $V^\pi_{\gamma_{z-1}}$, we have

$$\begin{aligned}
W^\pi_z(s) &= V^\pi_{\gamma_z} - V^\pi_{\gamma_{z-1}} \\
&= \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\mathcal{R}(s,a) + \gamma_z \mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[V^\pi_{\gamma_z}(s')\right] - \mathcal{R}(s,a) - \gamma_{z-1}\mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[V^\pi_{\gamma_{z-1}}(s')\right]\right] \\
&= \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\gamma_z \mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[V^\pi_{\gamma_z}(s')\right] - \gamma_{z-1}\mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[V^\pi_{\gamma_{z-1}}(s')\right]\right] \\
&= \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\gamma_z \mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[W^\pi_z(s') + V^\pi_{\gamma_{z-1}}(s')\right] - \gamma_{z-1}\mathbb{E}_{s'\sim\mathcal{T}(\cdot|s,a)}\left[V^\pi_{\gamma_{z-1}}(s')\right]\right] \\
&= \mathbb{E}_{\substack{a \sim \pi(\cdot|s) \\ s' \sim \mathcal{T}(\cdot|s,a)}}\left[(\gamma_z - \gamma_{z-1}) V^\pi_{\gamma_{z-1}}(s') + \gamma_z W_z(s')\right].
\end{aligned}$$

2. Let $\gamma, \beta \in [0, 1)$ be two discount factors such that $\beta \leq \gamma$. Let $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ be an arbitrary policy that induces value functions V_γ^π and V_β^π under the two discount factors, respectively. Similarly, define the Bellman operators

$$\mathcal{B}_\gamma^\pi V(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[\mathcal{R}(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[V(s')\right]\right]$$
$$\mathcal{B}_\beta^\pi V(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[\mathcal{R}(s,a) + \beta \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[V(s')\right]\right].$$

With the reward upper bound $R_{\text{MAX}} = \max\limits_{(s,a) \in \mathcal{S} \times \mathcal{A}} \mathcal{R}(s,a)$, prove that

$$||V_\gamma^\pi - V_\beta^\pi||_\infty \leq \frac{(\gamma - \beta) R_{\text{MAX}}}{(1-\gamma)(1-\beta)}.$$

Solution: This result is given as Theorem 2 of [Petrik and Scherrer, 2008] and highlights the approximation error that can occur by using a smaller discount factor β than that of the true MDP, γ.

$$\begin{aligned}
||V_\gamma^\pi - V_\beta^\pi||_\infty &= ||\mathcal{B}_\gamma^\pi V_\gamma^\pi - \mathcal{B}_\beta^\pi V_\beta^\pi||_\infty \\
&= ||\mathcal{B}_\gamma^\pi V_\gamma^\pi - \mathcal{B}_\beta^\pi V_\gamma^\pi + \mathcal{B}_\beta^\pi V_\gamma^\pi - \mathcal{B}_\beta^\pi V_\beta^\pi||_\infty \\
&\leq ||\mathcal{B}_\gamma^\pi V_\gamma^\pi - \mathcal{B}_\beta^\pi V_\gamma^\pi||_\infty + ||\mathcal{B}_\beta^\pi V_\gamma^\pi - \mathcal{B}_\beta^\pi V_\beta^\pi||_\infty \\
&\leq ||\mathcal{B}_\gamma^\pi V_\gamma^\pi - \mathcal{B}_\beta^\pi V_\gamma^\pi||_\infty + \beta ||V_\gamma^\pi - V_\beta^\pi||_\infty \\
&= \max_{s \in \mathcal{S}} |\mathbb{E}_{a \sim \pi(\cdot|s)} \left[\mathcal{R}(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[V_\gamma^\pi(s')\right] - \mathcal{R}(s,a) - \beta \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[V_\gamma^\pi(s')\right]\right] | + \beta ||V \\
&= \max_{s \in \mathcal{S}} |\mathbb{E}_{a \sim \pi(\cdot|s)} \left[\gamma \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[V_\gamma^\pi(s')\right] - \beta \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[V_\gamma^\pi(s')\right]\right] | + \beta ||V_\gamma^\pi - V_\beta^\pi||_\infty \\
&= \max_{s \in \mathcal{S}} |\mathbb{E}_{a \sim \pi(\cdot|s)} \left[(\gamma - \beta) \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[V_\gamma^\pi(s')\right]\right] | + \beta ||V_\gamma^\pi - V_\beta^\pi||_\infty \\
&\leq \max_{s \in \mathcal{S}} |\mathbb{E}_{a \sim \pi(\cdot|s)} \left[(\gamma - \beta) \mathbb{E}_{s' \sim \mathcal{T}(\cdot|s,a)} \left[\frac{R_{\text{MAX}}}{(1-\gamma)}\right]\right] | + \beta ||V_\gamma^\pi - V_\beta^\pi||_\infty \\
&= \frac{(\gamma - \beta) R_{\text{MAX}}}{(1-\gamma)} + \beta ||V_\gamma^\pi - V_\beta^\pi||_\infty \\
\implies (1-\beta) ||V_\gamma^\pi - V_\beta^\pi||_\infty &\leq \frac{(\gamma - \beta) R_{\text{MAX}}}{(1-\gamma)} \\
||V_\gamma^\pi - V_\beta^\pi||_\infty &\leq \frac{(\gamma - \beta) R_{\text{MAX}}}{(1-\gamma)(1-\beta)}
\end{aligned}$$

3. Let $\alpha, \gamma \in [0, 1)$ be two discount factors such that $\gamma \leq \alpha$. Consider a new MDP $\mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, \mathcal{T}', \mathcal{R}, \alpha \rangle$ with a different transition function $\mathcal{T}' : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ defined for $\lambda \in [0, 1]$ as

$$\mathcal{T}'(s' \mid s, a) = (1-\lambda)\mathcal{T}(s' \mid s, a) + \lambda \mathbb{1}(s = s'), \qquad \forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}.$$

In words, the new transition function $\mathcal{T}'$ follows the transitions of the original MDP $\mathcal{T}$ with probability $(1-\lambda)$ and takes a self-looping transition with probability λ. We will use subscripts to distinguish between value functions of $\mathcal{M}$ versus those of $\mathcal{M}'$.

Assuming that both $\mathcal{M}$ and $\mathcal{M}'$ are tabular, recall the matrix form of the Bellman equations for any policy π:

$$V_\mathcal{M}^\pi = (I - \gamma \mathcal{T}^\pi)^{-1} \mathcal{R}^\pi \qquad V_{\mathcal{M}'}^\pi = (I - \alpha \mathcal{T}'^\pi)^{-1} \mathcal{R}^\pi,$$

where

$$\mathcal{R}^\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\mathcal{R}(s,a)\right] \qquad \mathcal{T}^\pi(s' \mid s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\mathcal{T}(s' \mid s,a)\right] \qquad \mathcal{T}'^\pi(s' \mid s) = \mathbb{E}_{a\sim\pi(\cdot|s)}\left[\mathcal{T}'(s' \mid s,a)\right]$$

Solution: The results of this question are proven as part of Theorem 1 in [Jiang et al., 2015].

(a) Give a value of λ such that, for any policy π,

$$V^\pi_{\mathcal{M}'} = \frac{1-\gamma}{1-\alpha} \cdot V^\pi_{\mathcal{M}}.$$

Solution: We can write the transition matrix in the new MDP $\mathcal{M}'$ induced by any policy π as

$$\mathcal{T}'^\pi = (1-\lambda)\mathcal{T}^\pi + \lambda I,$$

where I is the $|\mathcal{S}| \times |\mathcal{S}|$ identity matrix. So, substituting in directly, we have

$$\begin{aligned}
V^\pi_{\mathcal{M}'} &= \left(I - \alpha\mathcal{T}'^\pi\right)^{-1}\mathcal{R}^\pi \\
&= \left(I - \alpha\left((1-\lambda)\mathcal{T}^\pi + \lambda I\right)\right)^{-1}\mathcal{R}^\pi \\
&= \left((1-\alpha\lambda)I - \alpha(1-\lambda)\mathcal{T}^\pi\right)^{-1}\mathcal{R}^\pi \\
&= \left((1-\alpha\lambda)\left(I - \frac{\alpha(1-\lambda)}{1-\alpha\lambda}\mathcal{T}^\pi\right)\right)^{-1}\mathcal{R}^\pi \\
&= \frac{1}{1-\alpha\lambda}\left(I - \frac{\alpha(1-\lambda)}{1-\alpha\lambda}\mathcal{T}^\pi\right)^{-1}\mathcal{R}^\pi.
\end{aligned}$$

We can compute the required value of λ as

$$\frac{\alpha(1-\lambda)}{1-\alpha\lambda} = \gamma \implies \lambda = \frac{\alpha-\gamma}{\alpha(1-\gamma)},$$

which means

$$\frac{1}{1-\alpha\lambda} = \frac{1}{1-\frac{\alpha-\gamma}{(1-\gamma)}} = \frac{1-\gamma}{1-\gamma-\alpha+\gamma} = \frac{1-\gamma}{1-\alpha}.$$

Substituting back in to the earlier equation yields

$$\begin{aligned}
V^\pi_{\mathcal{M}'} &= \frac{1}{1-\alpha\lambda}\left(I - \frac{\alpha(1-\lambda)}{1-\alpha\lambda}\mathcal{T}^\pi\right)^{-1}\mathcal{R}^\pi \\
&= \frac{1-\gamma}{1-\alpha}\left(I - \gamma\mathcal{T}^\pi\right)^{-1}\mathcal{R}^\pi \\
&= \frac{1-\gamma}{1-\alpha} \cdot V^\pi_{\mathcal{M}}.
\end{aligned}$$

(b) If $\pi^\star$ is the optimal policy of MDP $\mathcal{M}$, prove that $\pi^\star$ is also optimal in $\mathcal{M}'$.

Solution: By definition of the optimal policy, we know that $\pi^\star$ obeys the following inequality for any other policy π:

$$V^{\pi^\star}_{\mathcal{M}}(s) \geq V^\pi_{\mathcal{M}}(s), \qquad \forall s \in \mathcal{S}.$$

Since $\frac{1-\gamma}{1-\alpha} > 0$, we can scale both sides to get

$$\frac{1-\gamma}{1-\alpha} \cdot V^{\pi^\star}_{\mathcal{M}}(s) \geq \frac{1-\gamma}{1-\alpha} \cdot V^\pi_{\mathcal{M}}(s), \qquad \forall s \in \mathcal{S}.$$

Applying this previous part, we see that for any other policy π,

$$V_{\mathcal{M}'}^{\pi^\star}(s) \geq V_{\mathcal{M}'}^{\pi}(s), \qquad \forall s \in \mathcal{S}.$$

Thus, by definition, $\pi^\star$ is also the optimal policy in MDP $\mathcal{M}'$. This result illustrates that, for any MDP with a particular discount factor, there exists a transition function for another MDP with a larger discount factor such that the two MDPs have the same optimal policy.

References

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.

Marek Petrik and Bruno Scherrer. Biasing approximate dynamic programming with a lower discount factor. *Advances in Neural Information Processing Systems*, 21, 2008.

Joshua Romoff, Peter Henderson, Ahmed Touati, Emma Brunskill, Joelle Pineau, and Yann Ollivier. Separating value functions across time-scales. In *International Conference on Machine Learning*, pages 5468–5477. PMLR, 2019.