Problem 1

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

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

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. \( \gamma_Z = \gamma \). Letting \( [Z] \triangleq \{1, 2, \ldots, Z\} \) denote the index set, we define the following functions for any policy \( \pi \):

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

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

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

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

\[
W^\pi_z(s) = V^\pi_{\gamma_z} - V^\pi_{\gamma_{z-1}}
\]

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

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

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

\[
= \mathbb{E}_{a \sim \pi(\cdot|s)} \left[ (\gamma_z - \gamma_{z-1}) V^\pi_{\gamma_{z-1}}(s') + \gamma_z W^\pi_z(s') \right].
\]
2. Let \( \gamma, \beta \in [0, 1) \) be two discount factors such that \( \beta \leq \gamma \). Let \( \pi : \mathcal{S} \rightarrow \Delta(\mathcal{A}) \) be an arbitrary policy that induces value functions \( V^\pi_\gamma \) and \( V^\pi_\beta \) under the two discount factors, respectively. Similarly, define the Bellman operators
\[
B^\pi_\gamma V(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[ \mathcal{R}(s, a) + \gamma \mathbb{E}_{s' \sim T(\cdot|s,a)} [V(s')] \right]
\]
\[
B^\pi_\beta V(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} \left[ \mathcal{R}(s, a) + \beta \mathbb{E}_{s' \sim T(\cdot|s,a)} [V(s')] \right].
\]
With the reward upper bound \( R_{\text{MAX}} = \max_{(s, a) \in \mathcal{S} \times \mathcal{A}} \mathcal{R}(s, a) \), prove that
\[
||V^\pi_\gamma - V^\pi_\beta||_{\infty} \leq \frac{(\gamma - \beta)R_{\text{MAX}}}{(1 - \gamma)(1 - \beta)}.
\]
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 \( \beta \) than that of the true MDP, \( \gamma \).
\[
\begin{align*}
||V^\pi_\gamma - V^\pi_\beta||_{\infty} &= ||B^\pi_\gamma V^\pi_\gamma - B^\pi_\beta V^\pi_\beta||_{\infty} \\
&= ||B^\pi_\gamma V^\pi_\gamma - B^\pi_\beta V^\pi_\gamma + B^\pi_\beta V^\pi_\gamma - B^\pi_\beta V^\pi_\beta||_{\infty} \\
&\leq ||B^\pi_\gamma V^\pi_\gamma - B^\pi_\beta V^\pi_\gamma||_{\infty} + ||B^\pi_\beta V^\pi_\gamma - B^\pi_\beta V^\pi_\beta||_{\infty} \\
&\leq \max_{s \in \mathcal{S}} \left| \mathbb{E}_{a \sim \pi(\cdot|s)} \left[ \mathcal{R}(s,a) + \gamma \mathbb{E}_{s' \sim T(\cdot|s,a)} [V^\pi_\gamma(s')] - \mathbb{E}_{s' \sim T(\cdot|s,a)} [V^\pi_\beta(s')] \right] \right| + \beta ||V^\pi_\gamma - V^\pi_\beta||_{\infty} \\
&\leq \max_{s \in \mathcal{S}} \left| \mathbb{E}_{a \sim \pi(\cdot|s)} \left[ \mathcal{R}(s,a) - \beta \mathbb{E}_{s' \sim T(\cdot|s,a)} [V^\pi_\beta(s')] \right] \right| + \beta ||V^\pi_\gamma - V^\pi_\beta||_{\infty} \\
&= \frac{(\gamma - \beta)R_{\text{MAX}}}{(1 - \gamma)} + \beta ||V^\pi_\gamma - V^\pi_\beta||_{\infty} \\
\implies (1 - \beta)||V^\pi_\gamma - V^\pi_\beta||_{\infty} &\leq \frac{(\gamma - \beta)R_{\text{MAX}}}{(1 - \gamma)} \\
||V^\pi_\gamma - V^\pi_\beta||_{\infty} &\leq \frac{(\gamma - \beta)R_{\text{MAX}}}{(1 - \gamma)(1 - \beta)}
\end{align*}
\]
3. Let \( \alpha, \gamma \in [0, 1) \) be two discount factors such that \( \gamma \leq \alpha \). Consider a new MDP \( \mathcal{M}' = \langle \mathcal{S}, \mathcal{A}, T', \mathcal{R}, \alpha \rangle \) with a different transition function \( T' : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S}) \) defined for \( \lambda \in [0, 1] \) as
\[
T'(s' | s, a) = (1 - \lambda)T(s' | s, a) + \lambda \mathbb{I}(s = s'), \quad \forall (s, a, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}.
\]
In words, the new transition function \( T' \) follows the transitions of the original MDP \( T \) with probability \( (1 - \lambda) \) and takes a self-looping transition with probability \( \lambda \). 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 \( \pi \):
\[
V^\pi_{\mathcal{M}} = (I - \gamma T^\pi)^{-1} R^\pi, \quad V^\pi_{\mathcal{M}'} = (I - \alpha T'^\pi)^{-1} R^\pi,
\]
2
\[ R_\pi(s) = \mathbb{E}_{a \sim \pi(s)} [R(s, a)] \quad T_\pi(s' | s) = \mathbb{E}_{a \sim \pi(s)} [T(s' | s, a)] \quad T'_\pi(s' | s) = \mathbb{E}_{a \sim \pi(s)} [T'(s' | s, a)] \]

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

(a) Give a value of \( \lambda \) such that, for any policy \( \pi \),

\[ V^\pi_{M'} = \frac{1 - \gamma}{1 - \alpha} \cdot V^\pi_M. \]

Solution: We can write the transition matrix in the new MDP \( M' \) induced by any policy \( \pi \) as

\[ T'_{\pi'}(s' | s) = (1 - \lambda)T^\pi + \lambda I, \]

where \( I \) is the \(|S| \times |S|\) identity matrix. So, substituting in directly, we have

\[ V^\pi_{M'} = (1 - \alpha T^\pi)^{-1} R^\pi = (I - \alpha ((1 - \lambda)T^\pi + \lambda I))^{-1} R^\pi = ((1 - \alpha \lambda)I - \alpha(1 - \lambda)T^\pi)^{-1} R^\pi = \left(1 - \frac{\alpha(1 - \lambda)}{1 - \alpha \lambda} T^\pi\right)^{-1} R^\pi = \frac{1}{1 - \alpha \lambda} \left(1 - \frac{\alpha(1 - \lambda)}{1 - \alpha \lambda} T^\pi\right)^{-1} R^\pi. \]

We can compute the required value of \( \lambda \) as

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

which means

\[ \frac{1}{1 - \alpha \lambda} = \frac{1}{1 - \frac{\alpha}{\alpha - \gamma}} = \frac{1 - \gamma}{1 - \gamma - \alpha + \gamma} = \frac{1 - \gamma}{1 - \alpha}. \]

Substituting back in to the earlier equation yields

\[ V^\pi_{M'} = \frac{1}{1 - \alpha \lambda} \left( I - \frac{\alpha(1 - \lambda)}{1 - \alpha \lambda} T^\pi\right)^{-1} R^\pi = \frac{1 - \gamma}{1 - \alpha} (I - \gamma T^\pi)^{-1} R^\pi = \frac{1 - \gamma}{1 - \alpha} \cdot V^\pi_M. \]

(b) If \( \pi^* \) is the optimal policy of MDP \( M \), prove that \( \pi^* \) is also optimal in \( M' \).

Solution: By definition of the optimal policy, we know that \( \pi^* \) obeys the following inequality for any other policy \( \pi \):

\[ V^\pi_{M'}(s) \geq V^\pi_M(s), \quad \forall s \in S. \]

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

\[ \frac{1 - \gamma}{1 - \alpha} \cdot V^\pi_M(s) \geq \frac{1 - \gamma}{1 - \alpha} \cdot V^\pi_M(s), \quad \forall s \in S. \]
Applying this previous part, we see that for any other policy \( \pi \),

\[
V_{M'}^\pi(s) \geq V_{M'}^\pi(s), \quad \forall s \in S.
\]

Thus, by definition, \( \pi^* \) is also the optimal policy in MDP \( M' \). 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

