Midterm Exam II: Solutions

CS 6300 Artificial Intelligence **Name:**
University of Utah Spring 2016 **UID:**

---

1 Markov Decision Processes (60 points)

1. Formally define the components of a Markov Decision Process (MDP).

   An MDP is a tuple (S, A, T, R, γ)

   S: Set of states

   A: Set of actions

   T: Transition function of the form $T(s, a, s') = Pr(s'|s, a) \rightarrow [0, 1]$

   R: Reward function of the form $R(s, a, s') \rightarrow \mathcal{R}$

   γ: Discount factor $\gamma \in (0, 1]$

2. How is value defined in an MDP?

   Long-term expected reward to be accumulated starting from the current state. For the optimal value function this defined by:

   $$V^*(s) = \max_a \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma V^*(s')\right]$$

   while for an arbitrary fixed policy $\pi(\cdot)$ this is:

   $$V^\pi(s) = \sum_{s'} T(s, \pi(s), s') \left[R(s, a, s') + \gamma V^\pi(s')\right]$$

   Or for the state-action form we have:

   $$Q^*(s, a) = \sum_{s'} T(s, a, s') \left[R(s, a, s') + \gamma \max_{a'} Q^*(s', a')\right]$$

3. What is the objective of a rational agent, given an MDP?

   Find a policy that maximizes value for all states.

4. What is a policy?

   A function that maps states to actions $\pi : S \rightarrow A$

2 Modified MDP (20 points)

Given an arbitrary MDP with reward function $R(s)$ and a given constant $\beta > 0$, consider a modified MDP where everything remains the same, except it has a new reward function $R'(s) = \beta R(s)$. Prove that the modified MDP has the same optimal policy as the original MDP.

Since both the original and modified problems are valid MDPs each problem has an optimal value function. Let us define the optimal state-action value function for the original MDP as $Q_o^*(s,a)$ and for the modified MDP as $Q_m^*(s,a)$.

Consider the case where (s,a) is the penultimate state-action pair in a sequence:

$$\begin{aligned}
Q_m^*(s,a) &= \sum_{s'} T(s,a,s') \left[\beta R(s,a,s') + \gamma \max_{a'} Q_m^*(s',a')\right] \\
&= \sum_{s'} T(s,a,s') \left[\beta R(s,a,s') + \gamma \max_{a'} \sum_{s''} T(s',a',s'')\beta R(s',a',s'')\right] \\
&= \beta \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} \sum_{s''} T(s',a',s'')R(s',a',s'')\right] \\
Q_m^*(s,a) &= \beta Q_o^*(s,a)
\end{aligned}$$

Now consider the case where (s,a) is the third to last state-action pair in a sequence:

$$\begin{aligned}
Q_m^*(s,a) &= \sum_{s'} T(s,a,s') \left[\beta R(s,a,s') + \gamma \max_{a'} Q_m^*(s',a')\right] \\
&= \sum_{s'} T(s,a,s') \left[\beta R(s,a,s') + \gamma \max_{a'} \beta Q_o^*(s',a')\right] \\
&= \beta \sum_{s'} T(s,a,s') \left[R(s,a,s') + \gamma \max_{a'} Q_o^*(s',a')\right] \\
Q_m^*(s,a) &= \beta Q_o^*(s,a)
\end{aligned}$$

Hence by induction $Q_m^*(s,a) = \beta Q_o^*(s,a)$ for any arbitrary sequence length.

Given the optimal state-action value function, the optimal policy $\pi^*(\cdot)$ is defined as $\pi^*(s) = \arg\max_a Q^*(s,a)$ For the modified MDP we thus have:

$$\begin{aligned}
\pi_m^*(s) &= \arg\max_a Q_m^*(s,a) \\
&= \arg\max_a \beta Q_o^*(s,a) \\
&= \arg\max_a Q_o^*(s,a) \\
\pi_m^*(s) &= \pi_o^*(s)
\end{aligned}$$

Q.E.D.