Artificial Intelligence *COMP 5/600, Fall 2020*

HW07: Solving MDP

Remember that only PDF submissions are accepted.

1. Consider two finite MDPs, M_1 and M_2, having the same state set, S, the same action set, A, and respective optimal action-value functions Q_1^* and Q_2^*. (For simplicity, assume all actions are possible in all states.) Suppose that the following is true for some function $f: S \rightarrow$ Reals:

   $$Q_2^*(s, a) = Q_1^*(s, a) - f(s)$$

   for all $s \in S$ and $a \in A$. Show mathematically that M_1 and M_2 have the same optimal policies.

   $Q_1(s,a) = E[R| s, a, 1]$

   $1^*(s) = \max a\ Q_1(s,a) = \max a\ \{r_{t+1} + \gamma\ V^{\pi}(s_{t+1}) \mid s_t = s, a_t = a\} = \max a\ \Sigma\ P^{a}_{11'}[R^{a}_{11'} + \gamma\ V^{\pi}(1')]$

   $Q_2(s,a) = E[R| s, a, 2] - f(s)$

   $2^*(s) = \max a\ Q_2(s,a) = \max a\ \{r_{t+1} + \gamma\ V^{\pi}(s_{t+1}) \mid s_t = s, a_t = a\} = \max a\ \Sigma\ P^{a}_{22'}[R^{a}_{22'} + \gamma\ V^{\pi}(2')]$

   Thus 1*(s) = 2*(s), so M1 and M2 have the same optimal policies.

2. **True or False and justify your answers (aka, you need to give justifications)**:

   (a) **T** **F** Suppose you are given some arbitrary MDP M with finite state set, S, and you are also given some arbitrary function, f, that maps S to the real numbers. Then there exists a policy π for M such that $V^{\pi} = f$.

   Consider an MDP M where all the rewards are zero. This function f cannot be equal to V^{π} because f cannot be zero. V^{π} must be equal to 0 if the rewards are zero. This is a contradiction. Thus this is false.

   (b) **T** **F** If a policy π is greedy with respect to its own value function, V^{π}, then it is an optimal policy.

   A greedy policy chooses the best action among all possible actions. If we derive the V^{π} from π by using the policy improvement theorem, the policy is guaranteed to get a policy that is at least as good as policy π. If π is optimal then the derived policy will be equivalent to policy π. Thus, if a policy is greedy with respects to its own value function then it is an optimal policy.