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* → Reals:

*Q*∗2(*s, a*) = *Q* (∗1*s, a*) − *f* (*s*)

for all *s* ∈ *S* and *a* ∈ *A*. Show mathematically that *M*1 and *M*2 have the same optimal policies.

Q1(s,a) = E[R| s, a, 1]

1\*(s) = max a Q1(s,a) = max a {rt+1 + γ *V π*(st+1) | st = s, at = a} = max a Σ Pa11’[Ra11’+ γ *V π*(1’)

Q2(s,a) = E[R| s, a, 2] – f(s)

2\*(s) = max a Q2(s,a) = max a {rt+1 + γ *V π*(st+1) | st = s, at = a} = max a Σ Pa22’[Ra22’+ γ *V π*(2’)

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

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

* 1. **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.

1