Homework II

Deadline: 2022-11-24

1. (5 pts) In a finite state MDP $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, suppose every reward function $r(s, a, s')$ is changed by an affine transformation to $a \cdot r(s, a, s') + b$, where $a > 0$. Show that the optimal policies remain unchanged.

2. (10 pts) Recall the definition of the advantage function in Lecture 2:

$$g(\pi', \pi) = \mathcal{T}_{\pi'} v_\pi - v_\pi,$$

   where $\mathcal{T}_{\pi'}$ is the Bellman operator associated with the policy π'. Show that π^* is the optimal policy if and only if for any π there holds $g(\pi, \pi^*) \leq 0$ (elementwise).

3. (10 pts) Reproduce the figure on pg. 21 of Lecture 4 for the test of RM on root finding.

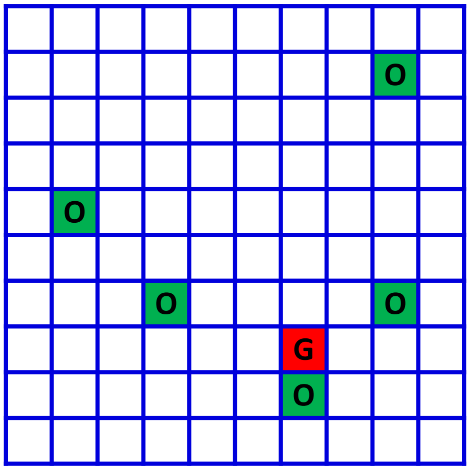
4. (5 pts) Present and prove a general policy improvement lemma that has been used in the proof of Theorem 1 (about the improvement of ϵ-greedy policy) in Lecture 4.

5. (10 pts) Test TD(0) for policy evaluation (pg. 5 in Lecture 5) by applying the algorithm on the following single episode **repeatedly**:

$$A,\ a_0,\ r = 0,\ B,\ a_1,\ r = 0,\ A,\ a_0,\ r = 0,\ B,\ a_2,\ r = 1,\ T,$$

   where T is the terminal state (that is, $v(T) = 0$). Try different values for the discount factor γ and 1) report the state values $v(A)$ and $v(B)$ that the algorithm converges to; 2) try to find some pattern from your results and guess what kind of solution the algorithm converges to. [Assume $v(A) = v(B) = 0$ for the initial guess and you can use a small constant stepsize in the algorithm or use the reciprocal of the visit times as the stepsize]

6. (10 pts) Implement MC Learning with ϵ-Greedy Exploration (pg. 23 of Lecture 4) and Off-policy MC Learning (pg. 30 of Lecture 4) and test them on the 10 gridworld problem shown in Figure 1.



								O	
	O								
			O					O	
						G			
						O			

Figure 1: Hit obstacle grid:-10; reach goal state (from other states): 10. Goal state is the terminal state, that is, if the agent leaves the goal state no matter what action it takes, it will return to the goal state with reward 0. The other settings are the same as the one in Homework I.