Homework II

Deadline: 2024-12-01

1. (40 pts) Reproduce the figures for MC Learning with ϵ-Greedy Exploration, Off-policy MC Learning , and TD Learning (SARSA and Q-learning) on the 10 gridworld problem shown in Figure 1.

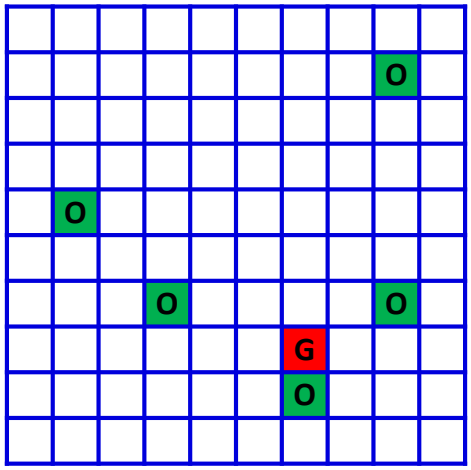

Figure 1: Hit obstacle grid:-10; reach goal state (from other states): 10. Goal state is the terminal state, that is, if the agent leaves the goal state no matter what action it takes, it will return to the goal state with reward 0. The other settings are the same as the one in Homework I.

2. Consider the MDP presented in Figure 2.
   - (5 pts)Compute $Q*(s_0, a_0)$.
   - (20 pts) Implement Q-learning and double Q-learning for this MDP and plot Q-values at (s_0, a_0). What is your observation?

3. (5 pts) We have seen the definition of Bellman error for Bellman operator in Lecture 6. The Bellman error for Bellman optimality operator can be similarly defined (under infinity norm),

$$\mathrm{BE}(V) := \|\mathcal{T}V - V\|_\infty.$$

   Show that

$$\|V - V^*\|_\infty \leq \frac{1}{1-\gamma}\mathrm{BE}(V).$$

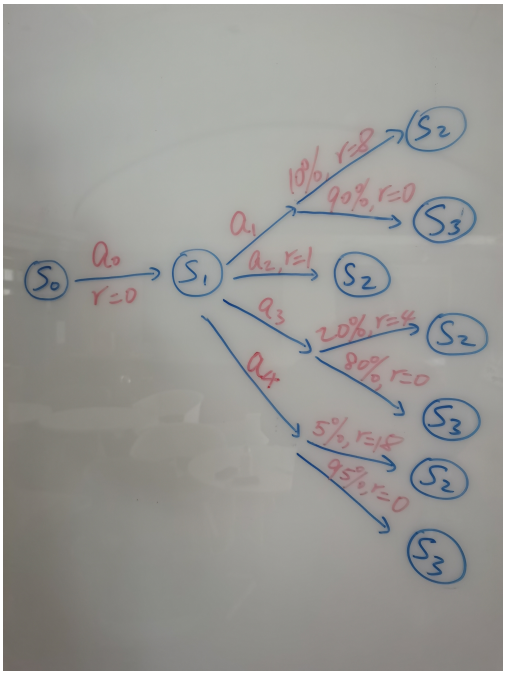

Figure 2: An MDP where s_2, s_3 are terminal states, the percentage means transition probability.

4. (5 pts) Show the following alternative expression for performance difference lemma:

$$V^{\pi_1}(\mu) - V^{\pi_2}(\mu) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d_\mu^{\pi_1}}\left[\sum_a (\pi_1(a|s) - \pi_2(a|s))Q^{\pi_2}(s,a)\right].$$

5. Given an initial distribution, recall the definition of visitation measure given in Lecture 7:

$$d_\mu^\pi(s) = \mathbb{E}_{s_0\sim\mu}[d_{s_0}^\pi(s)] = \mathbb{E}_{s_0\sim\mu}[(1-\gamma)\sum_{t=0}^{\infty}\gamma^t P(s_t = s|s_0,\pi)].$$

Assume $\tilde{\mu} = \min_s \mu(s) > 0$.

- (5 pts) Show that $d_\mu^\pi(s) \geq (1-\gamma)\tilde{\mu}$.
- (10 pts) Consider the Policy Iteration (PI) method presented in Lecture 2, and let π_k be the output policy of PI in the k-th iteration. Show that

$$V^*(\mu) - V^{\pi_k}(\mu) \leq (1-(1-\gamma)\tilde{\mu})^k \left(V^*(\mu) - V^{\pi_0}(\mu)\right).$$

  Comparing this result with the one we get in Lecture 2 based on contraction of the Bellman optimality operator, which one is better?