EECS 498 HW1

Casper Guo

September 2023

1 Bandits

1. $(1 - 0.5) + \frac{\epsilon}{2} = 0.75$
2. The random action selection could have possibly occured on any of the time steps. It definitely occured on time step t_3 as $Q_3(2) > 0$. It definitely occured on time step t_4 as well for the same reason. Random action selection also happened at time step t_5 because $Q_5(4) > Q_5(3)$.
3. The $\epsilon = 0.01$ method will perform the best in the long run in terms of both cumulative reward and probability of selecting the best action. Both ϵ-greedy methods are guarenteed to eventually identify the local actions but then the $\epsilon = 0.1$ will select it 91% of the time whereas the $\epsilon = 0.01$ method will select it 99.1% of the time. Thus the later method will eventually achieve higher average reward.

2 MDP

1. The reward setting is inappropriate. There is no clear incentive to quickly escape from the maze as there is no penalty for lingering in the maze indefinitely (0 reward). An improvement might be to add a -1 reward for each time step not at the goal state.
2. $$G_t = R_{t+1} + \gamma R_{t+2} \dots$$
$$G_1 = R_2 + \gamma R_3 \dots = 7 \times \frac{1}{1-\gamma} = 70$$
$$G_0 = R_1 + \gamma G_1 = 2 + 0.9 \times G_1 = 66.8$$
3. $$v_\pi(s) = \sum_{a \in \mathcal{A}(s)} \pi(a|s) q_\pi(s, a)$$
4. $$q_\pi(s, a) = \sum_{s' \in \mathcal{S}^+} \sum_{a \in \mathcal{A}(s)} \sum_{r \in \mathcal{R}} p(s', r|s, a)(r + \gamma v_\pi(s'))$$

5.
$$G_{\text{center}} = \frac{\gamma(G_N + G_E + G_S + G_W)}{4} = \frac{0.9 \times 3}{4} = 0.675 \approx 0.7$$

6. We have $G_t = R_{t+1} + \gamma R_{t+2} \ldots$. If we add constant c to all the rewards, resulting in the sequence $R_{t+1} + c$, $\gamma(R_{t+2} + c)$ and so on. Then the total change to the value of G_t is $\frac{c}{1-\gamma}$.

   Since $v_\pi(s) = \mathbb{E}_\pi[G_t | S_t = s]$ and $\mathbb{E}[x + c] = \mathbb{E}[x] + c$ where c is a constant. All the state values will change by $v_c = \frac{c}{1-\gamma}$.

7. Since both policies are just the same two timesteps being repeated indefinitely. We only need to compare the rewards received in the first two timesteps.

   When $\gamma = 0$, π_{right} is better. When $\gamma = 0.9$, see $0.9 \times 1 + 0.9^2 \times 2 < 0.9 \times 0 + 0.9^2 \times 2$, so π_{right} is better. When $\gamma = 0.5$, $0.5 \times 1 = 0.5^2 \times 2$, so the two policies are equivalent in terms of rewards received.

3 DP

1. Check $q_\pi(s, \text{old-action}) = q_\pi(s, \text{new-action})$. If true, do nothing. Otherwise, set policy-stable to false.

2. Almost analogous. In policy evaluation, also loop over all $a \in \mathcal{A}$ to update q estimates.

3.
$$q_{k+1}(s, a) = \sum_{s', r} p(s', r | s, a)[r + \gamma v_{s'}(k)]$$

4 Monte Carlo

1. Maintain a count of the number of returns seen previously n and the previous mean mu. Then when a new return G is added, update μ to $\frac{n\mu + g}{n+1}$ and increment n.

2. First-visit: 10. Every-visit: $\frac{\sum_{i=1}^{10} i}{10} = 5.5$.