XCS234 Assigment 2

Ammar Husain

August 2022

1 Q1

1.1 a

Infinite-horizon MDP $\mathcal{M}$

Trajectory $\tau = (s_0, a_0, s_1, a_1, ...)$

The probability of sampling τ given a policy π and a certain MDP $\mathcal{M}$ is

$$\rho^{\pi}(\tau) = \Pi_{t=0}^{\infty} P^{\mathcal{M}}(s_{t+1}|s_t, a_t)\pi(a_t|s_t)$$

That is the probability of sampling an action a_t given a state s_t using policy π multiplied by the probability of reaching state s_{t+1} from state s_t and taking action a_t in an MDP $\mathcal{M}$

2 Q2

2.1 a

The maximum sum of rewards that can be achieved in a single trajectory in the test environment is 6.2. This can be achieved by the following:

$\mathrm{s}_0 = 0, a_0 = 2, R_0 = 0.0$
$\mathrm{s}_1 = 2, a_0 = 3, R_0 = 3.0$
$\mathrm{s}_2 = 3, a_0 = 2, R_0 = 0.0$
$\mathrm{s}_3 = 2, a_0 = 3, R_0 = 3.0$
$\mathrm{s}_4 = 3, a_0 = 0, R_0 = 0.2$
$\mathrm{s}_5 = 0$

No other trajectory can achieve such a high reward because going from state 2 to state 3 really amplifies the reward 10X more than any other transition. So even though going back to state 2 yields zero reward it pays off enormously. The last step from 3 to 0 helps add a 0.2 reward toward the end which is the only positive reward for originating states that are not state 2.

3 Q3

3.1 b

Assuming that our Q function is an unbiased estimator of the optimal Q^* function we can easily see that it will still overestimate the real target value. Consider a single state s where there are many actions a whose true values $Q^*(s,a)$ all zero but the estimated values $Q(s,a)$ are uncertain and thus distributed some above and some below zero. The maximum of the true values $Q^*(s,a)$ is zero, but the maximum of the estimates is positive thereby yielding a positive bias. This causes the estimator to overestimate the real target:

$$E[max_a Q(s,a)] \geq max_a Q^*(s,a)$$