ECSE 6965

Assignment 5

Ishtiak Mahmud
RIN 662066901

October 2024

1. The average simulation (discounted) return is going to be similar to the value of one of your MDP's states. Which state is that? What statistical property (i.e., theorem) that we saw in class ensures that the average reward will converge to that state's value?

Ans: During all the simulation, we were always starting from state 1 and calculating our average reward over a large number of simulation. By the *Law of large number* we know that, if we do simulations for a large number of times, the sample average matches closely with the actual average. However, by definition $V_\pi(1)$ is the expected reward you get when you start from state 1. That is why the average simulation (discounted) return is going to be similar to the value function of state 1 or $V_\pi(1)$.

2. Why does policy $\pi_{0.2}$ result in lower returns?

Ans: Policy $\pi_{0.2}$ results in lower returns because it introduces more randomness: it follows the optimal action with probability 0.8 and a random action with probability 0.2. This extra randomness increases the chance of making suboptimal moves, especially dangerous in the cliff environment where stepping into a cliff cell results in a large penalty of -100 and resets the agent to the start. The agent then spends additional steps (each costing -1) to return toward the goal, further reducing the discounted return. Thus, the increased exploration in $\pi_{0.2}$ leads to lower expected returns due to higher risks outweighing any potential benefits.

3. Suppose you use the matrix form of the Bellman equation in order to calculate the state values. What would happen to the calculation if we set the discount factor to 1, i.e., $\gamma = 1$?

Ans: Setting the discount factor $\gamma = 1$ in the Bellman equation leads to computational issues. The matrix form is:

$$V = R + \gamma P_\pi V \implies (I - \gamma P_\pi)V = R.$$

With $\gamma = 1$, this becomes:

$$(I - P_\pi)V = R.$$

The matrix $I - P_\pi$ may be singular (non-invertible) if P_π has eigenvalue 1, which is common in MDPs with cycles. This means we can't uniquely solve for V. Also, with $\gamma = 1$, the total return might diverge to infinity if there are cycles with positive rewards, making V undefined for some states. Therefore, setting $\gamma = 1$ prevents us from obtaining meaningful value estimates using standard methods.