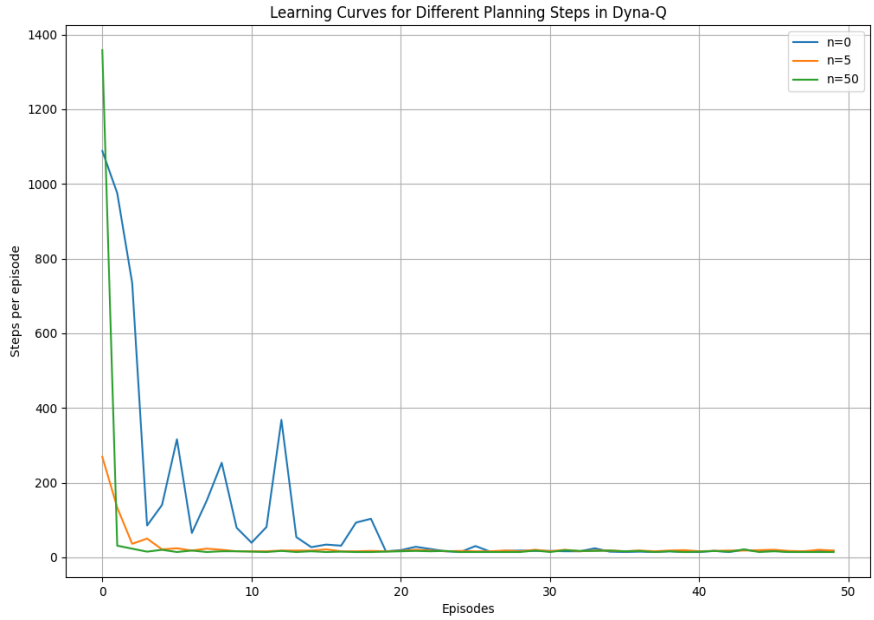
Sumeet Shanbhag

Student ID : 641020714

Reinforcement Learning Programming Assignment 5

In this programming assignment, a 6-by-9 grid maze was constructed with obstacles, a designated start, and a goal state. The agent's actions included moving up, down, left, or right, each associated with a Q-value indicating the expected utility of performing the action in a given state. Three variations of the Dyna-Q algorithm are tested, differing in the number of planning steps per real step: *n*=0, *n*=5, and *n*=50, over 50 episodes to assess learning efficiency. For each state-action pair, the model predicted the next state and reward, allowing the algorithm to 'plan' by updating Q-values based on these predictions. The experiment maintained a consistent seed for random number generation to ensure comparability.



The learning curves plotted in Figure 1 represent the mean steps per episode for each planning step condition. All conditions exhibited learning progression, with initial episodes showing the highest steps per episode, reflecting the exploration phase. As episodes progressed, the learning curves for *n*=5 and *n*=50 steeply declined, *n*=50 achieving the quickest reduction in steps, indicative of efficient policy learning. The *n*=0 condition, representing standard Q-learning, showed a more gradual decline, underscoring the importance of planning in learning acceleration.

The improvement with increased planning steps illustrates the balance between exploration and exploitation. Planning steps allow the agent to benefit from indirect learning without additional interactions with the environment, leading to a more informed policy and reduced steps to reach the goal. However, the marginal benefit seemed to plateau, suggesting an optimal range for the number of planning steps.