Mountain Car Simulation

Task 1

- **Reward Function:** The agent is rewarded **-1** if it goes on the wrong hill of either side. A reward of **-0.01** is given otherwise to encourage movement and faster learning. A reward of 1 if it gets to the goal.
- **State Space Binning:** The minimum and maximum positions are discretized to 45 bins. 45 bins have been experimentally proven to be better for the agent to learn compared to other bin sizes.
- **Learning Parameters:**
    - **Alpha** is given a value of **0.1**. This is a decent value as a higher alpha will make the agent learns faster on the immediate reward, and a lower value will make it slow to learning.
    - **The Discount Factor** (**Gamma)** is assigned a value of **0.99**. This makes agents prioritize future rewards and not the immediate rewards. Immediate steps will lead to the goal over time.
    - **Epsilon/Decay Rate:** is assigned a value of **0.1** at the start of the simulation. The decay rate is set to **0.94**. Then the epsilon value is decayed gradually in each episode to a minimum of 0.05 at about half of the episodes. The agent has taken a few more random actions and learn from them and then optimized afterwards. The motivation behind choosing this value is that about 50% of the time, the agent tries more random actions, and the remaining 50% is full optimization with only 5% degree of randomness, see Figure 1.
- **Simulation:** This simulation has 10 runs and 20 episodes in each run. Epsilon value and Q-table are reset in every run. This allows each pattern to be observed independently of the other. Agent learns via the episodes in each run.

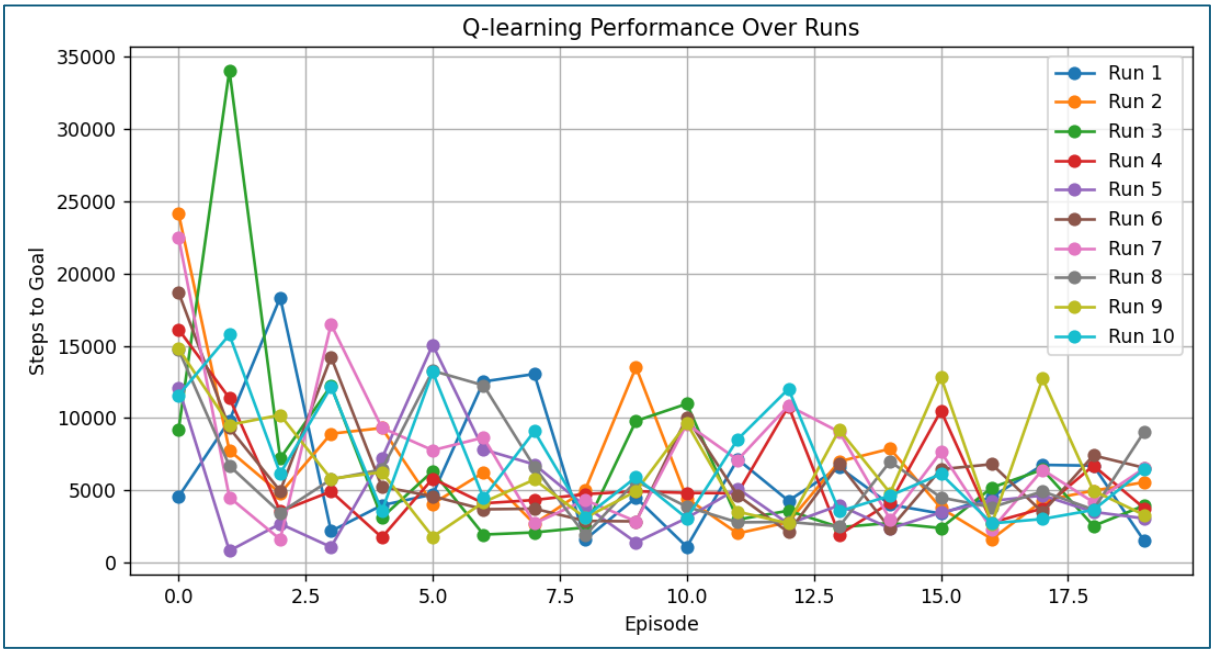

Figure 1. shows plot for 10 runs 10 episodes for Task 1

Observation:

- Convergence can be observed with the graph from the figure above (figure 1). Later runs have lower steps, indicating learning and optimization over time.

Task 2

- Learning parameters from task 1 are unchanged.
- **State Space Binning:** The gradient at any point in the graph is a function of the position at that same point. To discretise the gradient, a bin size of 45 is used with the boundaries of 1 and -1 inclusive. The value of the bins' boundaries is chosen based on the gradient values ranging from -1 and 1, a cosine function.

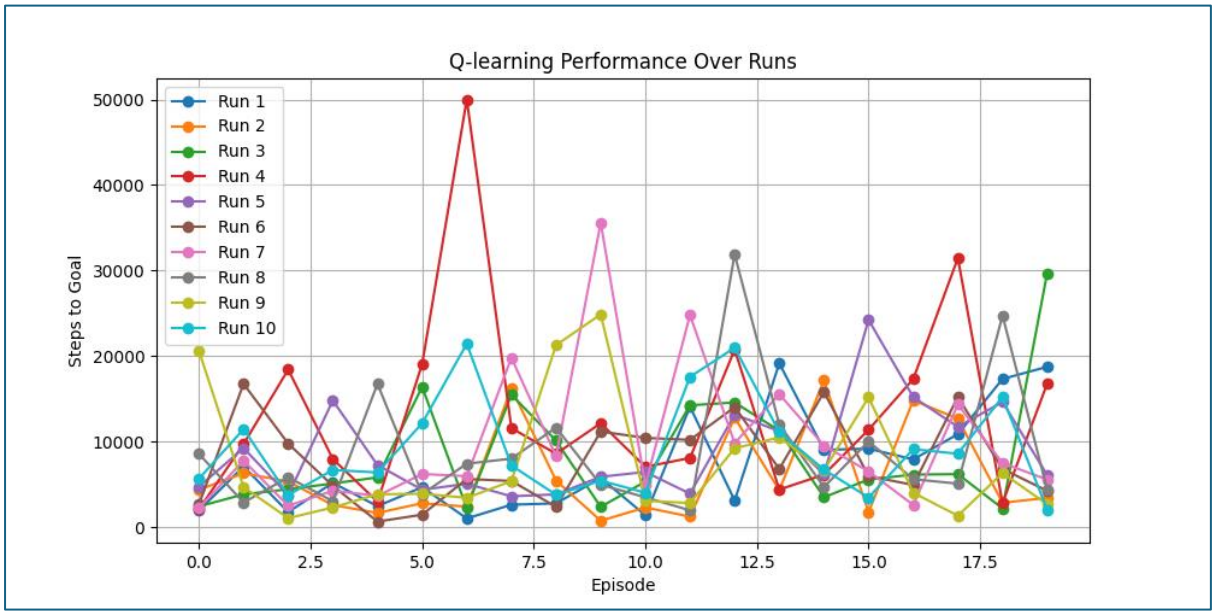

Figure 2. shows 10 runs, 20 episodes for the gradient state space for Task 2

Observation

- The gradient states space is better at the beginning with lower steps per episode as compared to task 1, with slightly higher steps because the position is not known any lower.
- It doesn't have a converging pattern at the span of the episodes recorded as compared to task 1. This is because the position is not known.

Task 3

- In this task, the action space has been changed to [-1.9, 1.8]. this enables the agents to decelerate downhill to the garage. The episode is increased to 120 due to an increase in the complexity of the task. A bin size of 30 was experimentally better for this task.
- **State Space Binning:** The gradient at any point in the graph is a function of the position at that same point. To discretise the gradient, a bin size of 30 of continuous variables is used with the boundaries of 1 and -1 inclusive. The hill slope on the garage side is a replica of the hill. The variable ranges from 0 1, -1 to 1.
- **This vs Task2:** Task3 has convergence while Task2 does not. But due to the increased complexity, it has more steps compared to task 2.
- **This vs Task 1:** Due to the increased complexity, Task 1 has fewer steps compared to Task 3.

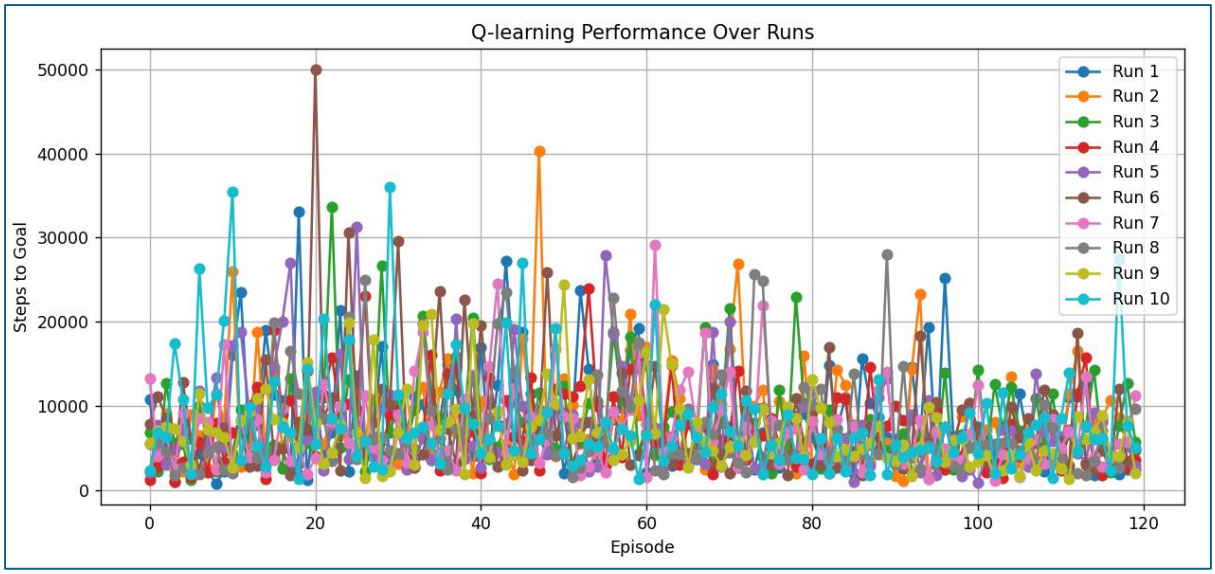

Figure 3 shows 10 runs, 120 episodes for Task 3

Task4

- The gradient space (-1, 1) was discretized into 4 bins. The Q table is made of 60 states, indicating 30 for green bins and 30 for red bins.
- The gradient flows in both directions (0, -1, 0, 1). When a gradient value falls in a state and the green bin, the discretised state will be between 0 and 29. If the gradient falls in a state in the red bin, the value is a number between 30 and 60.
- **Bin Choice:** the state space (-1.5 and 1.5) was segmented into two. The left-hand side (-1.5 – 0.0) indicates the green bin, and the right, the red bin (0.0 – 1.5). The agent was less penalized for being in the red zone to move in the direction of the goal.

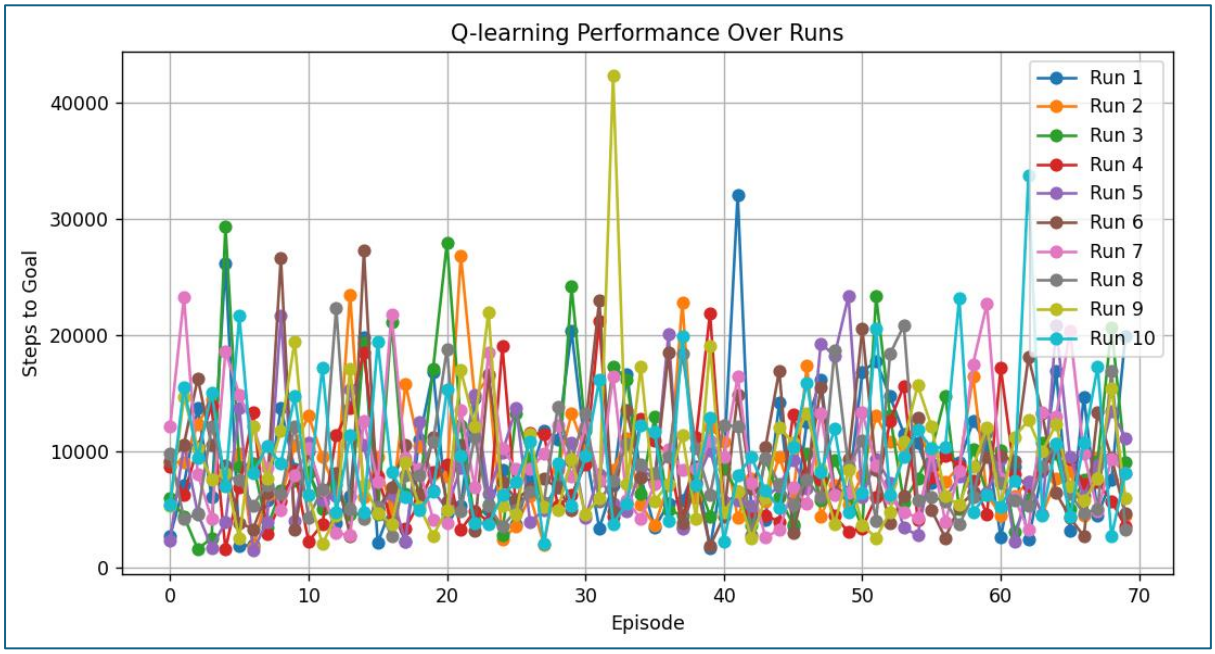

Figure 4 show 10 runs 70 episodes for Task 4.

Observation

- **Task 4 vs Task 3:** Task 4 has a reduced number of steps (about 10, 000) compared to Task 3. And converges slightly better at the 70th episode to 20,000 steps. This is a lesser step if compared to the 70th step in task 3. Using 2 bins and reward adjustment is the reason for better performance
- **Task 4 vs Task 2:** Task 4 slightly converges while Task 2 does not. It has slightly fewer steps (about 5,000) compared to task 2.
- **Task 4 vs Task 1:** Task1 show lesser steps count. Task 1 has less complexity than Task 4 and one of the reasons for this result.

Task5

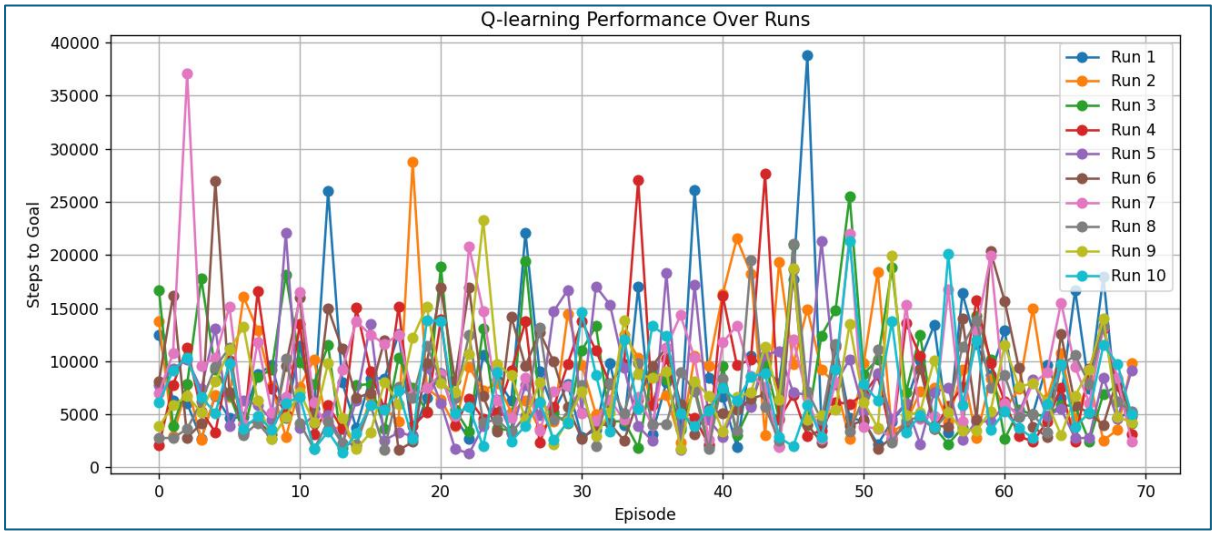

Figure 5 show 10 runs 70 episodes for task 5

- **Task 6 vs Task 5:** Task 5 has improved performance over Task 4. It has fewer steps than task 4 (about 10,000 fewer steps). Using many more bins and more rewards is the cause of this positive effect
- **Task 6 vs Task 4:** Task 5 has improved performance over Task 3 with fewer steps per episode.
- **Task 6 vs Task 3:** Task1 show lesser steps count. Task 1 has less complexity than Task 4, and one of the reasons for this result. The number of steps per episode that stays below 25, 000 in task 5 is higher than that of task 2.
- **Task 6 vs Task 2** The Number of steps in Task 1 is fewer than this task. Again, task 1 has less complexity.
- **Task 6 vs Task 2** The Number of steps in Task 1 is fewer than this task. Again, task 1 has less complexity.

Task 6: In task 6 I added an overlap bin towards to the red bin closer to the goal but it did not have significant better impact compared to other previous

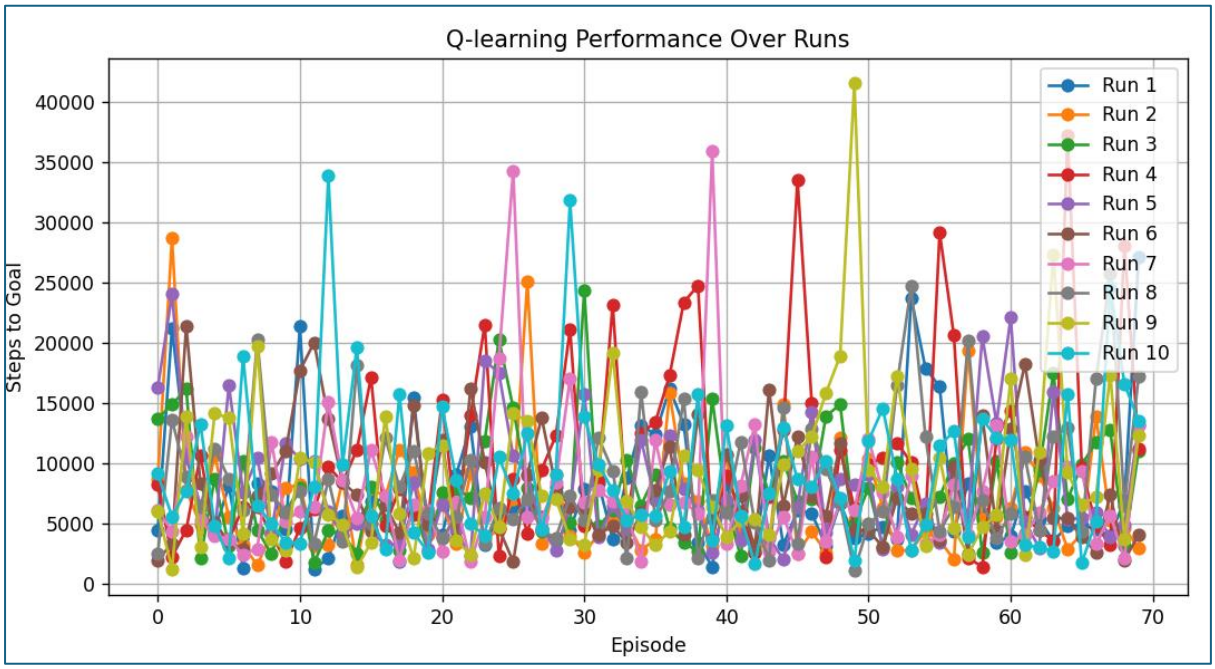

Figure 5 show 10 runs 70 episodes for task 6

Bonus Question: In this task, I attempted to use PID control to reduce overshooting, especially near the goal. Although I did not achieve any success record, overshooting and convergence was set in. Improvement will be made on this area.

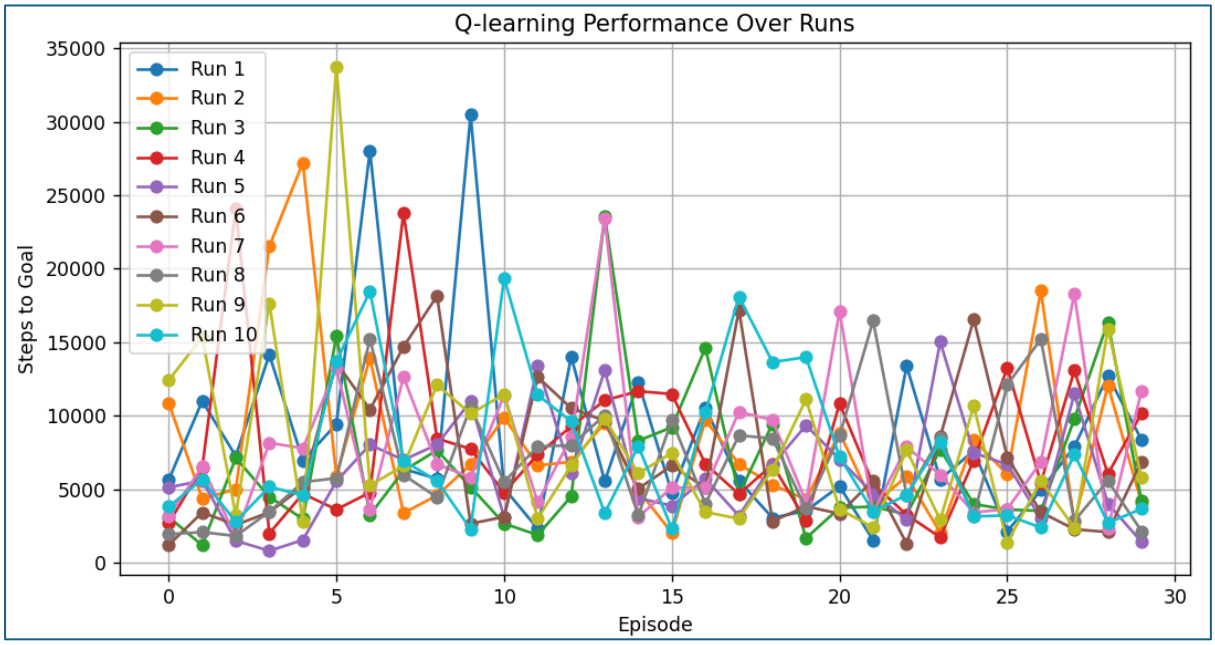

Figure 7 showing 10 runs 30 episodes for bonus task.