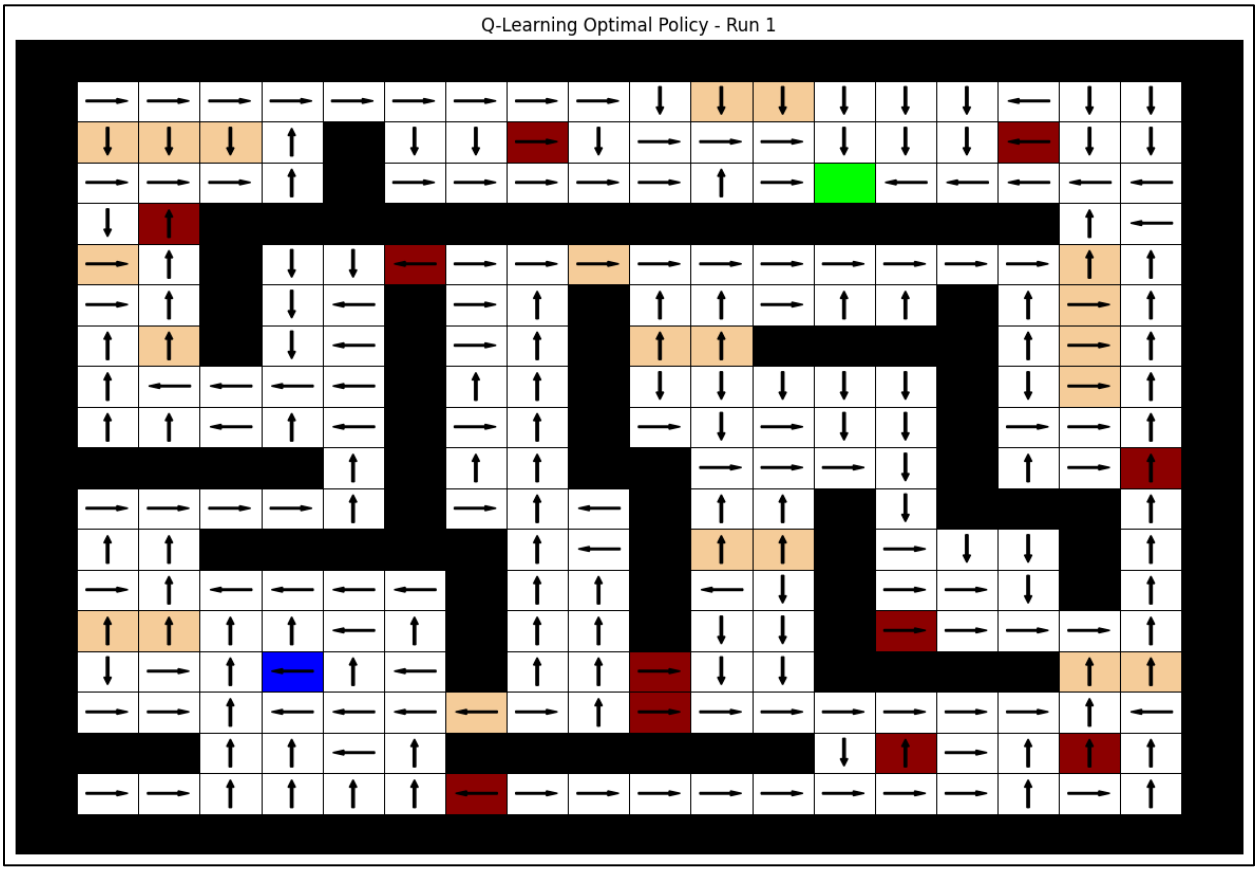
Project 3

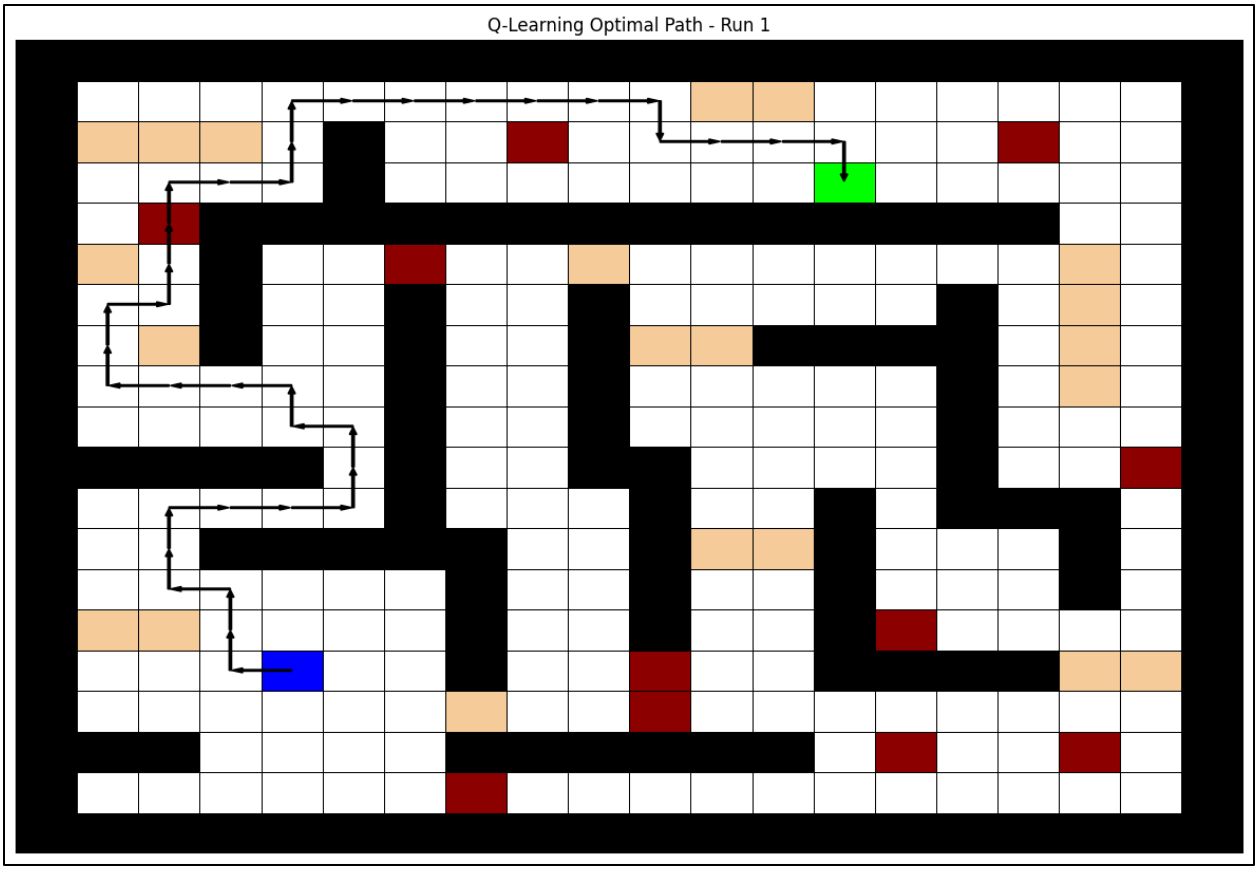
Name: Ameya Padwad
NUID: 002284038

Problem 1

Q-Learning

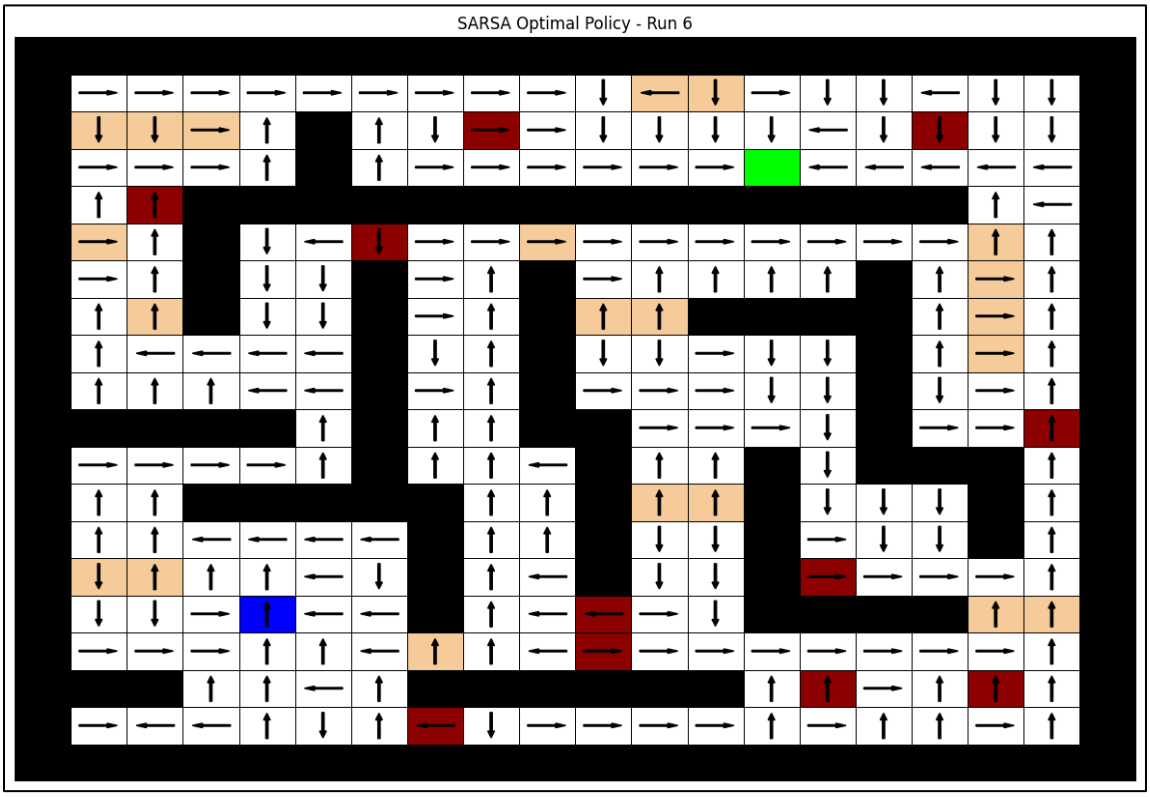
No of times path from start to goal found = 10/10

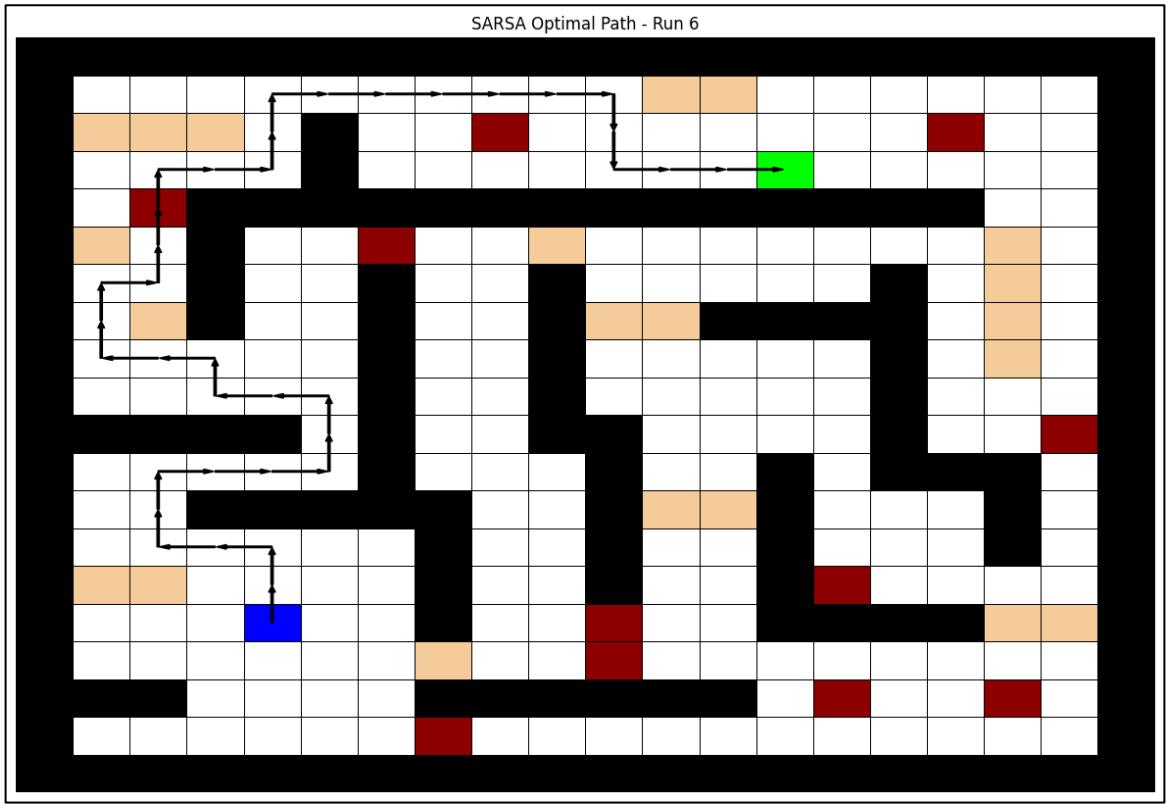



Q-Learning Optimal Path - Run 1

SARSA

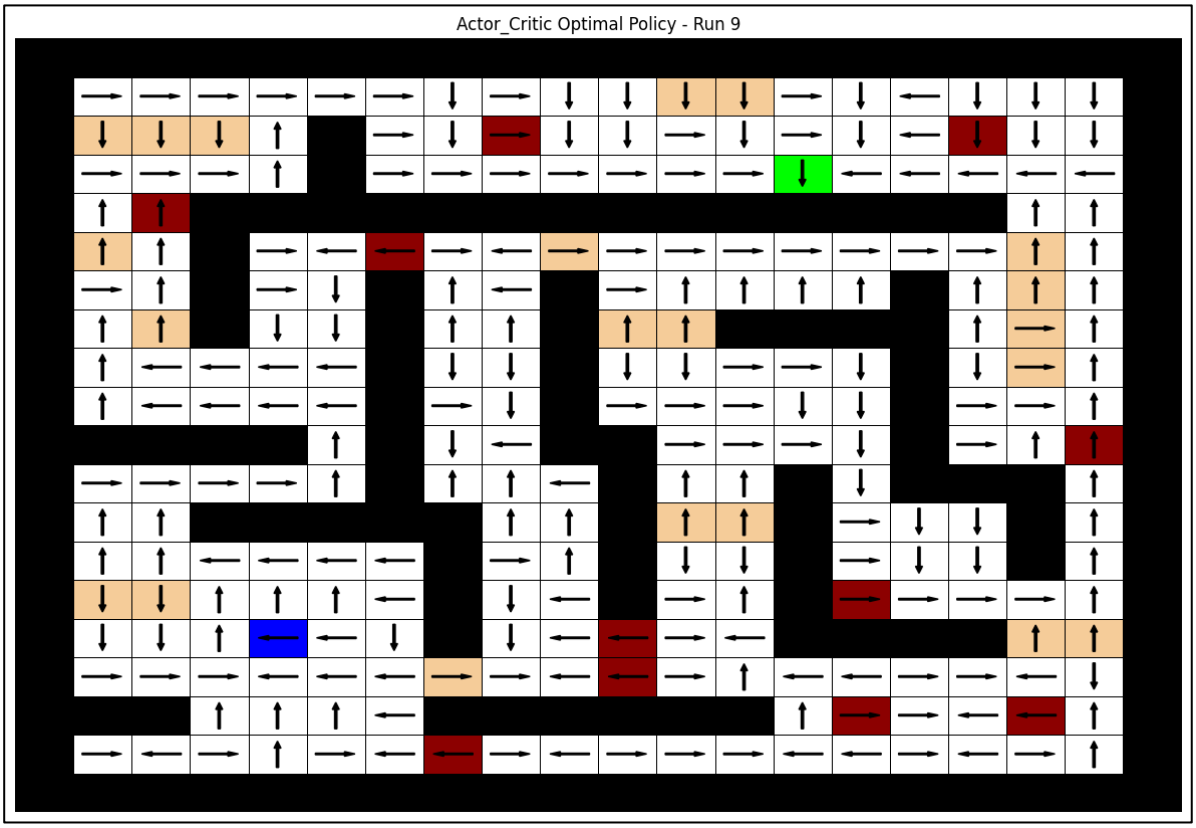
No of times path from start to goal found = 9/10

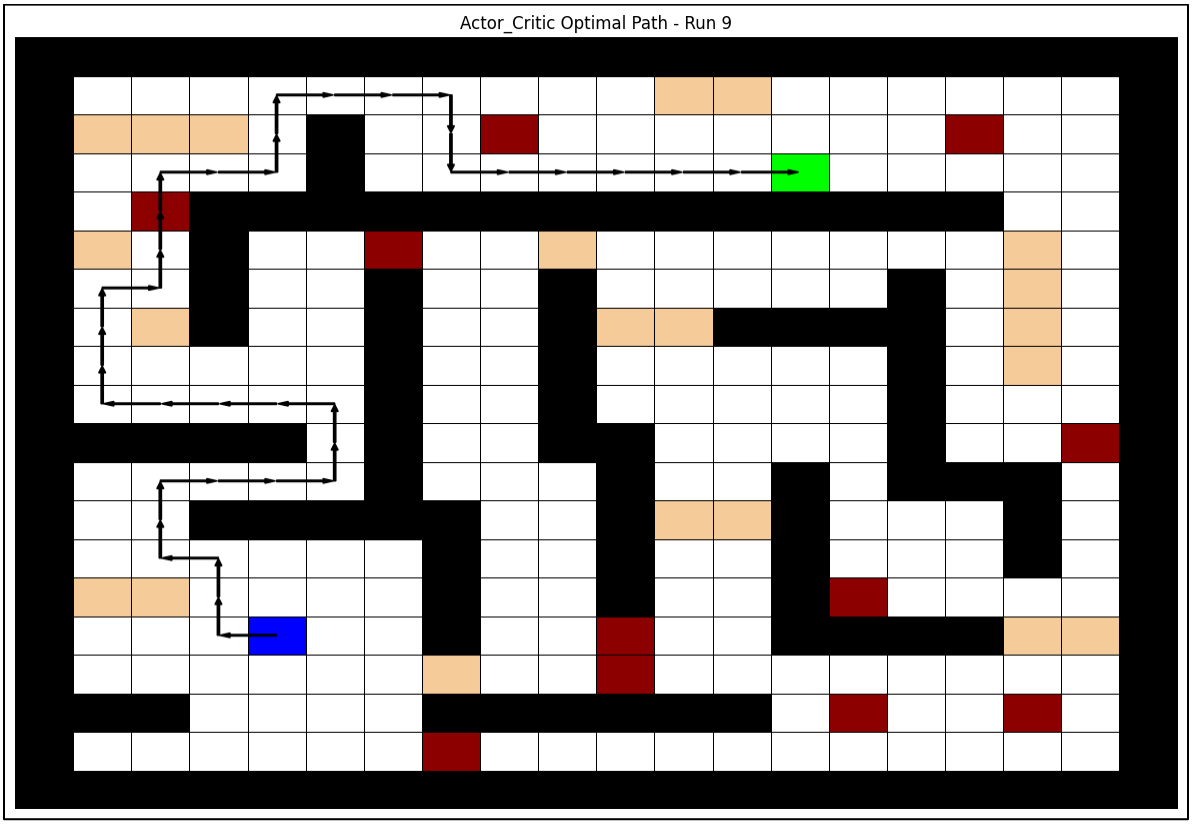




Actor Critic

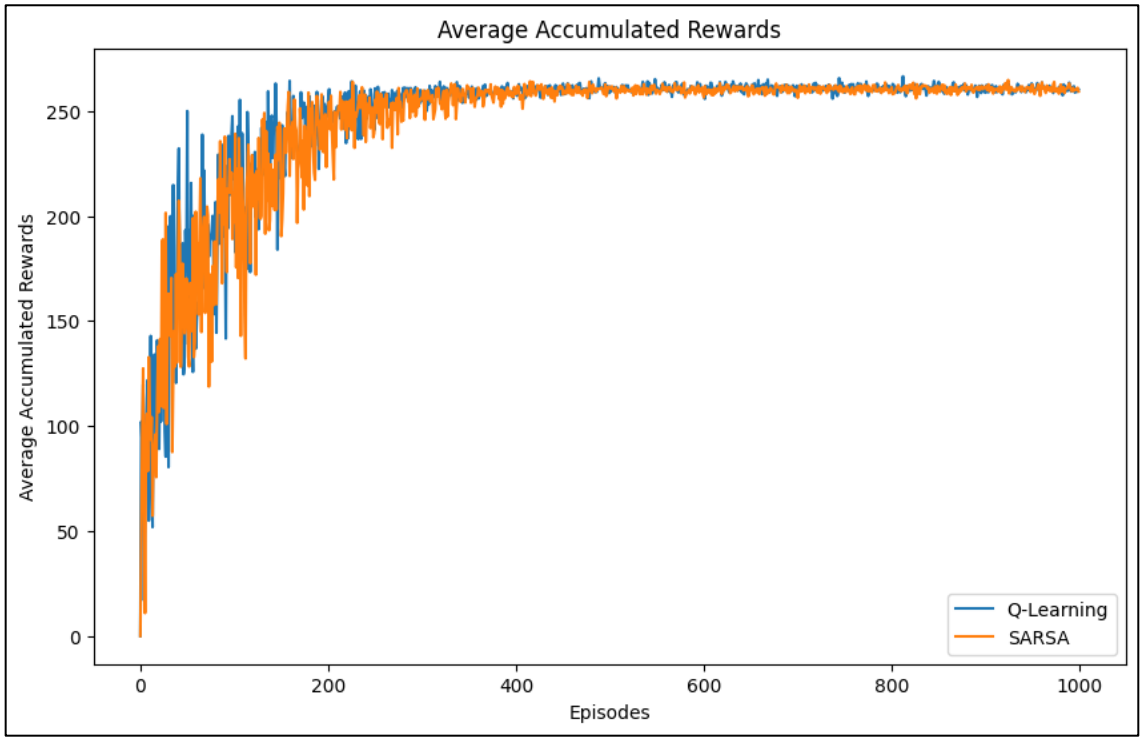
No of times path from start to goal found = 5/10

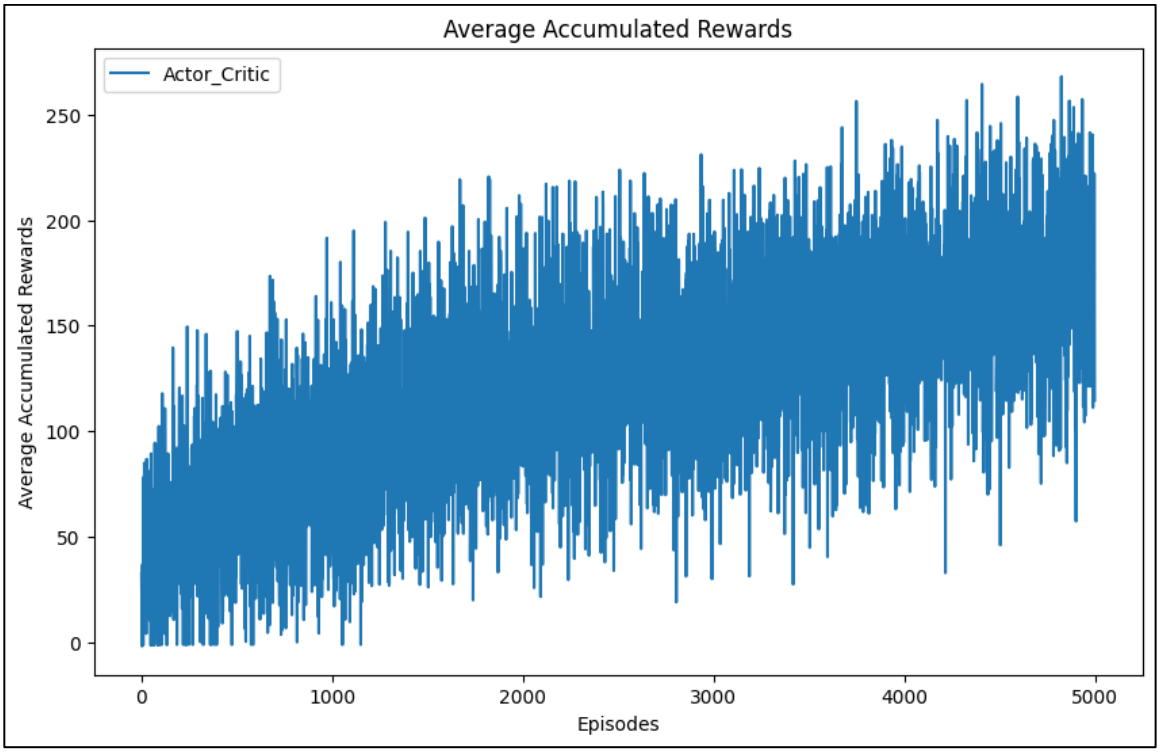




For the actor critic method, when the base parameters were used, optimal path was found 0 times, and the optimal policy was sub-optimal. The above results were obtained using the following parameters:

$p = 0.025$, $\gamma = 0.96$, $\alpha = 0.01$, $\epsilon = 0.1$, $\beta = 0.05$, max_no_of_episodes = 5000 and max_length_of_episode = 200.





The Q-Learning and SARSA algorithms converge rapidly and exhibit stable performance, both reaching high accumulated rewards (around 270) within 1000 episodes. This indicates effective learning and policy optimization under the given settings.

The Actor-Critic plot, however, runs for 5000 episodes, and displays a noisier but upward-trending reward curve. Initially, rewards are low and fluctuate widely, but over time, the algorithm steadily improves, achieving comparable reward levels to Q-Learning and SARSA by the end. This extended and noisy convergence is likely due to the dual learning structure of Actor-Critic — where both actor (policy) and critic (value function) must learn in tandem, making it more sensitive to hyperparameter tuning and variance.

Problem 2

Q-Learning

	Run 1	**Run 2**	**Run 3**	**Run 4**	**Run 5**	**Run 6**	**Run 7**	**Run 8**	**Run 9**	**Run 10**
S1	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S2	a2	a2	a3	a2	a2	a2	a2	a2	a2	a2
S3	a2	a2	a3	a2	a2	a2	a2	a2	a2	a2
S4	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S5	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S6	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S7	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S8	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S9	a4	a3	a3	a3	a4	a3	a3	a1	a1	a3
S10	a2	a2	a3	a2	a2	a2	a3	a2	a2	a2
S11	a2	a2	a2	a2	a2	a2	a4	a2	a2	a2
S12	a2	a3	a2	a2	a2	a2	a2	a2	a2	a2
S13	a4	a1	a4	a4	a4	a4	a4	a4	a4	a4
S14	a2	a2	a1	a2	a2	a2	a2	a2	a2	a2
S15	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S16	a2	a2	a2	a2	a2	a2	a1	a2	a2	a2

SARSA

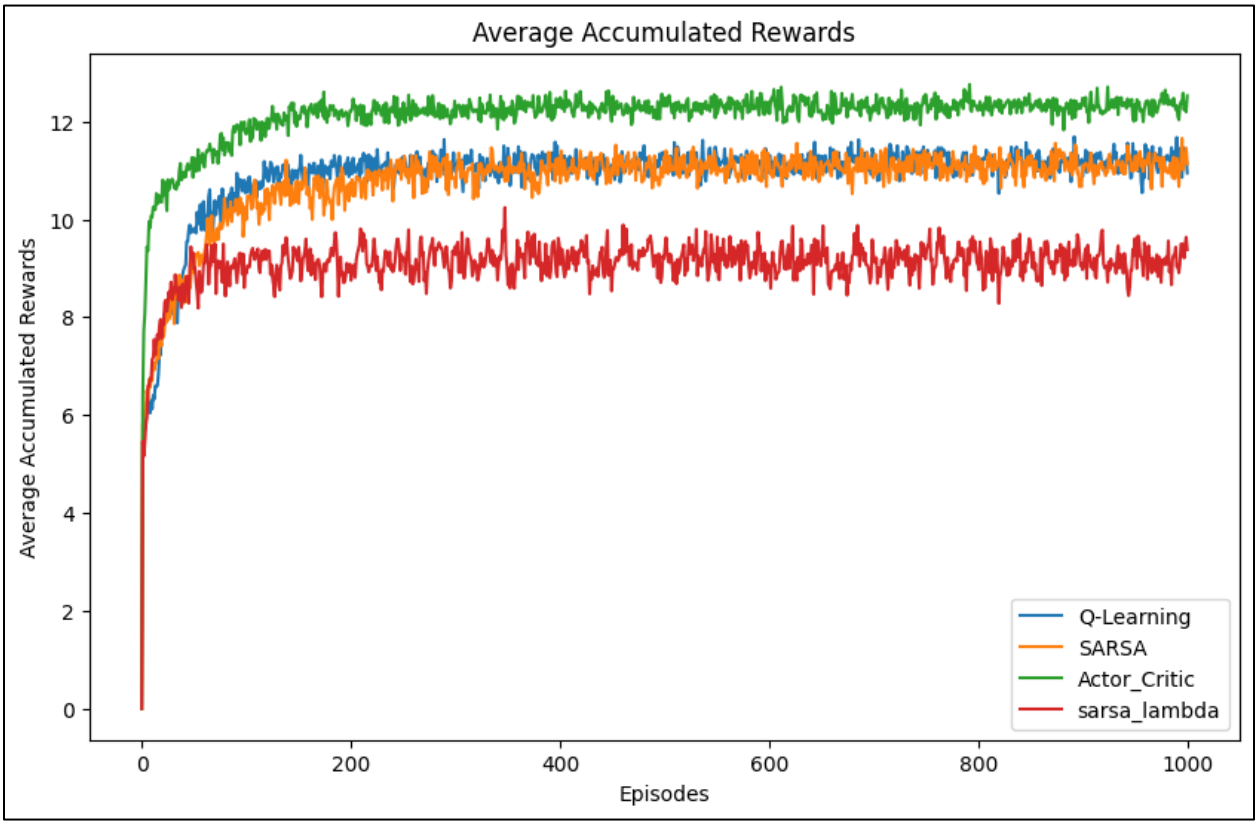
	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
S1	a2	a2	a2	a2	a3	a2	a2	a2	a2	a3
S2	a2	a2	a2	a2	a3	a4	a2	a2	a2	a2
S3	a2	a2	a3	a3	a2	a2	a3	a2	a3	a2
S4	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S5	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S6	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S7	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S8	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S9	a3	a4	a3	a1	a4	a3	a3	a1	a4	a3
S10	a2	a4	a2	a4	a1	a2	a2	a2	a2	a4
S11	a2	a2	a2	a2	a2	a2	a4	a2	a2	a2
S12	a2	a2	a2	a3	a2	a2	a2	a2	a2	a2
S13	a4	a4	a4	a4	a4	a4	a1	a4	a4	a4
S14	a2	a2	a2	a2	a2	a2	a4	a2	a2	a2
S15	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S16	a2	a1	a2	a2	a2	a2	a2	a2	a2	a2

SARSA - Lambda

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
S1	a3	a4	a4	a3	a3	a2	a2	a2	a2	a2
S2	a3	a2	a2	a2	a1	a2	a2	a4	a2	a2
S3	a2	a2	a2	a3	a2	a2	a3	a2	a3	a2
S4	a3	a2	a2	a2	a2	a2	a3	a2	a2	a2
S5	a2	a1	a2	a2	a1	a2	a2	a2	a2	a2
S6	a2	a2	a2	a2	a2	a2	a2	a1	a2	a4
S7	a2	a2	a2	a2	a2	a2	a2	a2	a2	a4
S8	a1	a1	a1	a4	a2	a1	a1	a3	a1	a1
S9	a3	a1	a4	a3	a1	a1	a4	a4	a4	a4
S10	a2	a2	a4	a4	a2	a4	a2	a2	a2	a3
S11	a3	a4	a2	a2	a2	a2	a2	a2	a2	a2
S12	a3	a2	a2	a2	a4	a2	a4	a2	a2	a2
S13	a4	a4	a4	a4	a4	a4	a4	a4	a4	a4
S14	a4	a2	a2	a2	a2	a2	a4	a2	a4	a2
S15	a2	a1	a1	a2	a2	a1	a2	a2	a2	a2
S16	a2	a1	a4	a2	a2	a2	a2	a2	a2	a2

Actor Critic

	Run 1	Run 2	Run 3	Run 4	Run 5	Run 6	Run 7	Run 8	Run 9	Run 10
S1	a3	a2	a2	a2	a3	a2	a2	a2	a2	a2
S2	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S3	a2	a2	a2	a2	a2	a2	a3	a2	a2	a2
S4	a2	a2	a2	a2	a2	a2	a2	a3	a2	a2
S5	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S6	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S7	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S8	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S9	a1	a1	a4	a3	a4	a4	a1	a3	a4	a4
S10	a2	a3	a3	a2	a3	a2	a2	a2	a2	a2
S11	a3	a2	a3	a2	a2	a3	a2	a3	a2	a2
S12	a2	a2	a3	a4	a2	a2	a2	a2	a2	a2
S13	a4	a4	a1	a4	a4	a4	a1	a4	a4	a4
S14	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2
S15	a2	a2	a2	a2	a2	a2	a2	a1	a2	a2
S16	a2	a2	a2	a2	a2	a2	a2	a2	a2	a2



The Actor-Critic algorithm clearly outperforms the rest, converging quickly and stabilizing at the highest average reward (~12.5). Q-Learning and SARSA follow closely with similar trends, leveling off around 11.5 and 11.2 respectively. SARSA(λ), however, lags noticeably behind, struggling to exceed an average reward of 9 and showing higher variance throughout.

Actor-Critic combines the benefits of both value-based and policy-based methods, enabling faster and more stable convergence, especially in noisy environments like this one with Bernoulli transition noise. Q-Learning performs slightly better than SARSA since it is off-policy and tends to be more aggressive in exploiting optimal actions. SARSA, being on-policy, is more conservative and thus slightly slower in reaching high rewards. SARSA(λ), despite theoretically offering faster learning through eligibility traces, may be underperforming here due to suboptimal tuning of λ or interference caused by the stochasticity in transitions. Its high variance suggests it is more sensitive to noisy updates, leading to instability in learning.