SARSA and Q-Learning

Implement SARSA and Q-learning algorithms with a linear epsilon-decay behaviour policy. The epsilon value should decrease linearly (by a specified decay parameter) every episode but maintain a minimum value of 0.01.

Apply to the FrozenLake and Taxi Gym environments with the following parameters:

- Learning rate = 0.1
- Discount factor = 0.95
- 5000 episodes
- Initial epsilon = 0.4
- Epsilon decay value = 0.001

Test the policies every 200 episodes by running 1000 trajectories (using the policy at that point in time) and averaging the cumulative returns. Compare the performance of the SARSA and Q-learning algorithms by plotting these results.

Evaluate (estimate) the optimal policy for each environment using DP (choose either policy or value iteration and set the discount factor to 0.95). How do the SARSA and Q-learning results compare to these?