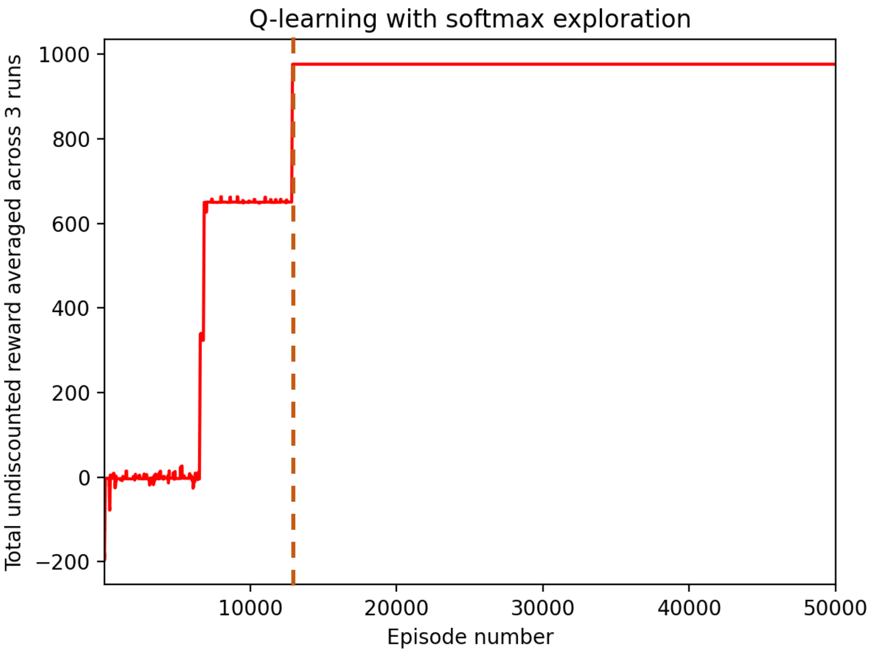
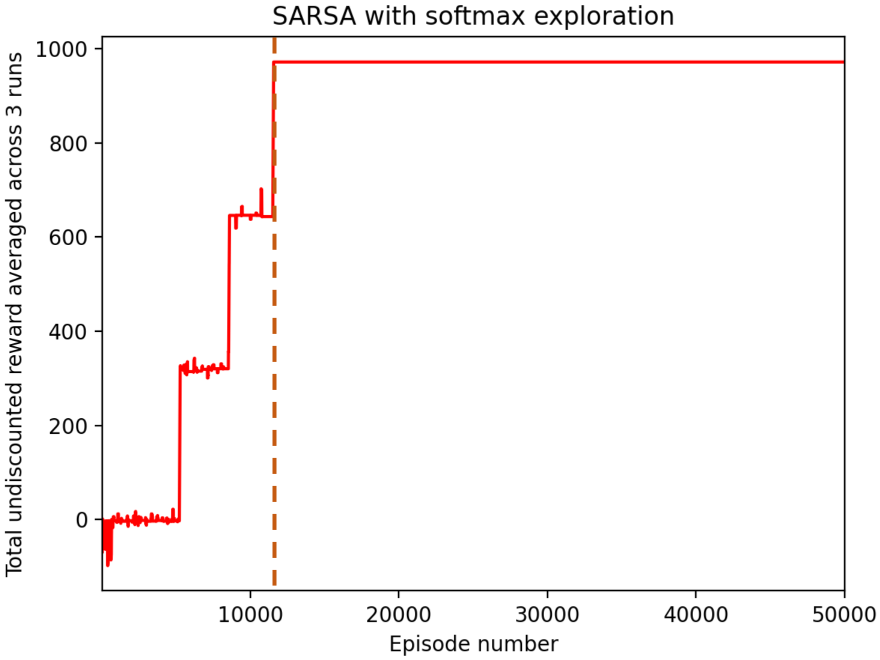
Question 2

1.2) Using $\alpha = 0.5, \gamma = 0.99,$

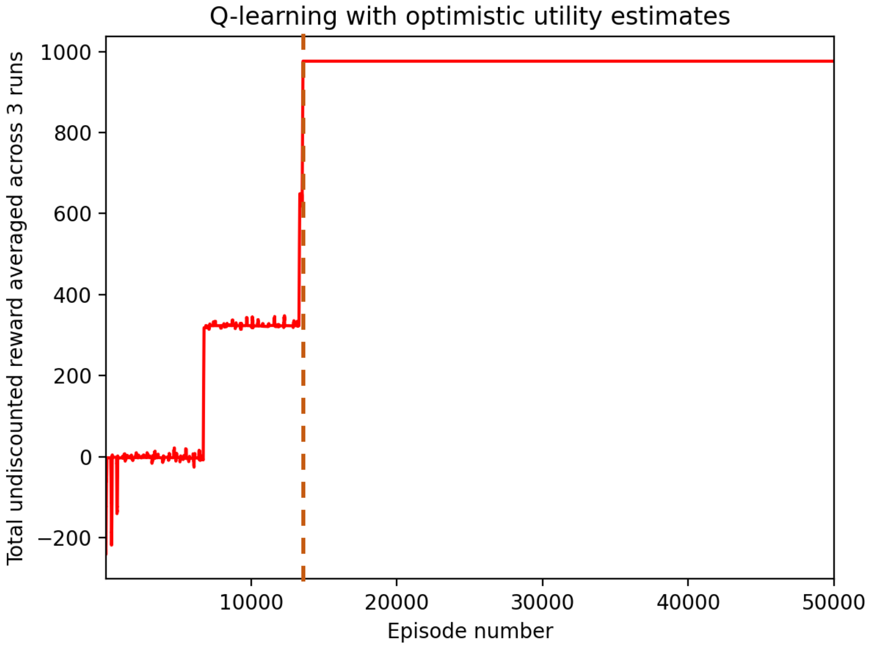
a. Q-learning + softmax exploration with $T = 0.1$



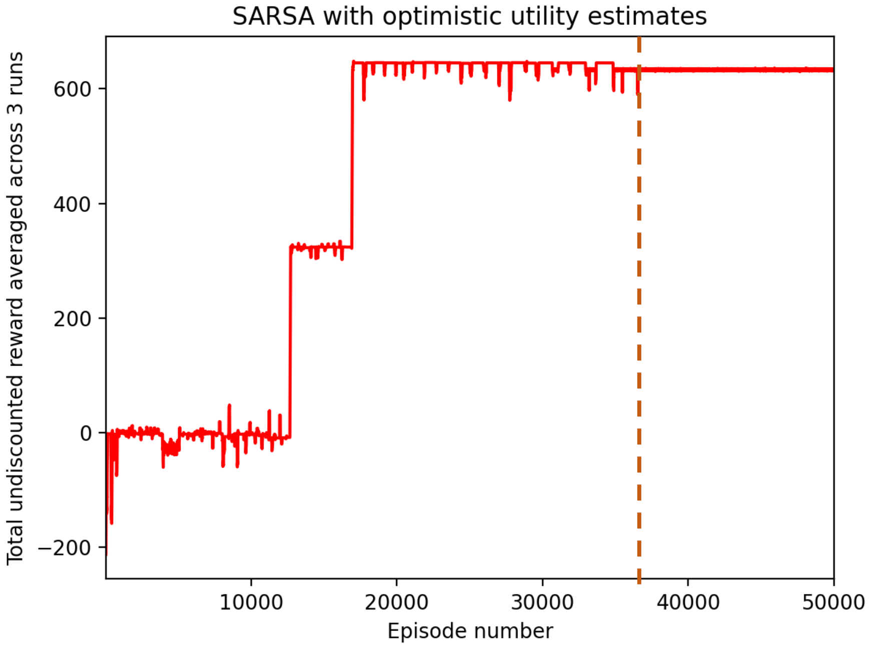
b. SARSA + softmax exploration with $T = 0.1$



c. Q-learning + optimistic utility exploration with N_e = 2, R^+ = 999



d. SARSA + optimistic utility exploration with N_e = 2, R^+ = 999



2) If I were to explore the Wumpus world, I would choose SARSA with softmax exploration.

Based on the above graphs, SARSA with softmax exploration stabilized the fastest after 12,000 episodes, followed by Q-learning with softmax exploration (13,000 episodes), Q-learning with optimistic utility estimates (14,000 episodes), and SARSA with optimistic utility estimates (38,000 episodes). Note that all except SARSA with optimistic utility estimates have comparable rewards of 1,000, while SARSA with optimistic utility estimates have a lower reward of 600.