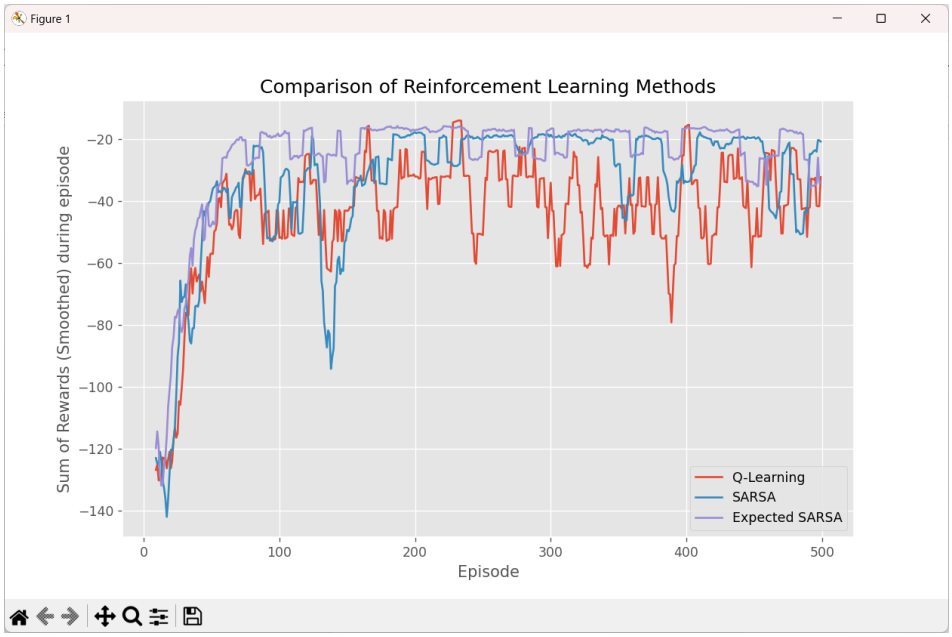
Assignment1: SARSA, Q-learning, Expected-SARSA

20211145 이하은

Cliff-Walking 문제를 SARSA, Q-Learning, Expected-SARSA 알고리즘으로 돌렸을 때 결과물:



위와 같은 결과가 나오는 이유 분석:

Q-Learning은 보상이 가장 큰 action을 선택하는 방식으로 학습되며, 주로 Cliff의 가장자리 근처에서 탐험을 하게 됩니다. 이때 엡실론-그리디(epsilon-greedy) 정책을 사용해 무작위로 행동을 선택하기 때문에 Cliff에 떨어져 큰 벌점(-100)을 받을 가능성이 존재합니다. 이로 인해 Q-Learning은 세 알고리즘 중 가장 낮은 보상의 합을 기록합니다.

SARSA는 on-policy 알고리즘으로, 에이전트가 보수적인 탐험을 하여 Cliff 근처에서 위험을 피하는 경향이 있습니다. 그 결과, 절벽에 떨어질 확률이 낮아 Q-Learning보다 더 높은 보상을 기록합니다.

Expected SARSA는 행동을 선택할 때, 가장 큰 보상을 주는 Action을 선택하는 것이 아니라, 모든 action의 기대값을 바탕으로 Q값을 업데이트합니다. 이로 인해 SARSA보다 적극적인 탐험을 하면서도 Q-Learning보다는 보수적으로 행동하여 탐험과 안전 사이에서 적절한 균형을 유지합니다. 따라서, 가장 높은 보상을 기록하는 경향이 있습니다.