Reinforcement Learning

Ex-06

1)

(a) What justifies classifying Q-learning as an off-policy method?

Q-learning is termed off-policy because it updates its value function based on the next state *s*' and the greedy choice of action *a*, independent of the policy being followed. This means that the policy used to generate the behavior does not influence the updating of the value estimates. Consequently, Q-learning is recognized as an off-policy method of control.

(b) If we use a greedy policy for choosing actions, does Q-learning become identical to SARSA in terms of its algorithmic structure and the way it selects actions and updates values?

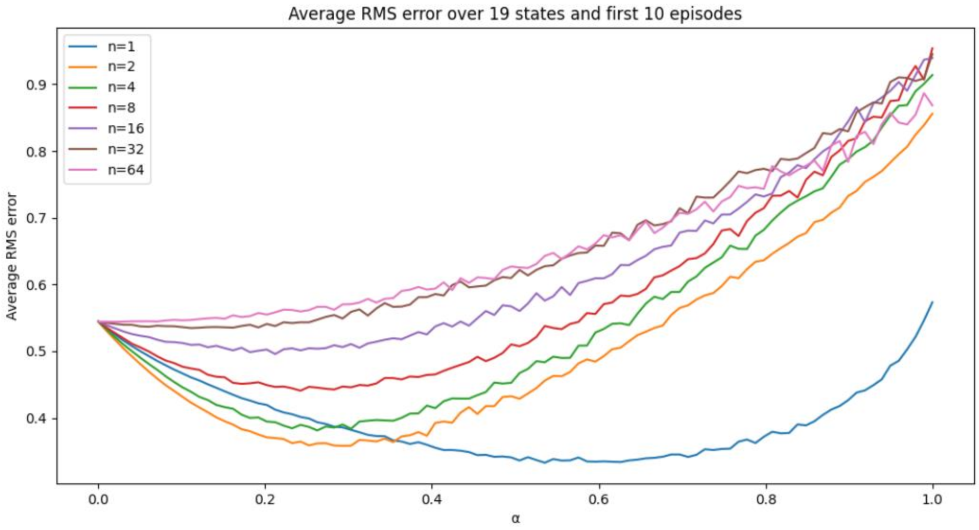
When a greedy policy is applied to action selection, the SARSA algorithm aligns with the Q-learning algorithm. As a result, both algorithms would perform action selection and value updates identically.

2)

To assess the impact of varying 'n' on the efficacy of the Temporal Difference (TD) method, it's necessary to experiment with different 'n' values. The initial setup involves an environment with five states, which presents challenges in performance evaluation.

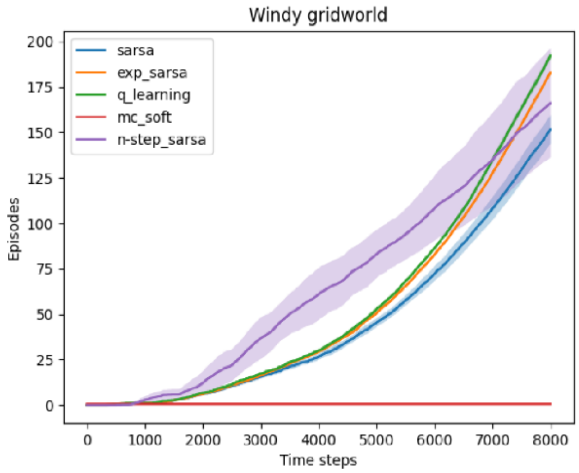
For a smaller walk problem, the expected value of number of states for travelling is smaller than n. Then the performance is more accurate than a larger n. Therefore, a smaller walk task does affect the results regarding which value of n yields the best performance, the smaller n should be better.

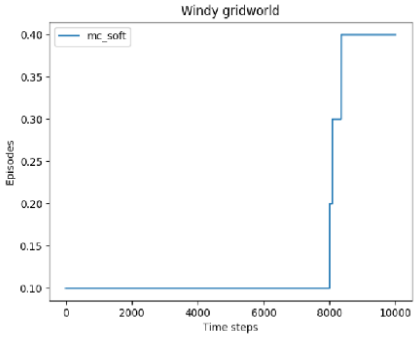
Changing the left-side outcome from 0 to -1 won't make any difference in the best value of n. n only depends on the environment, as long as the state and transition unchanged, the best value of n is unchanged.



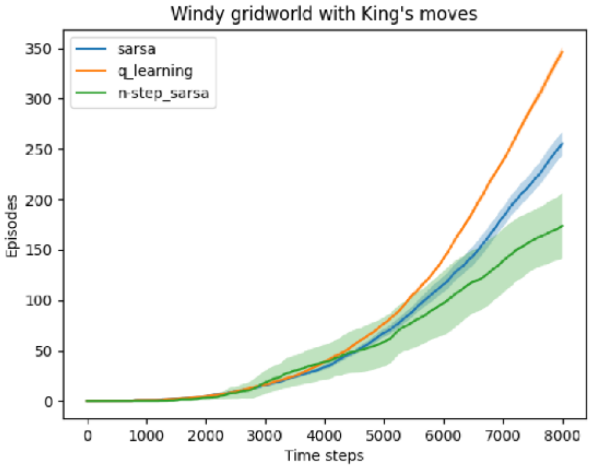
3)

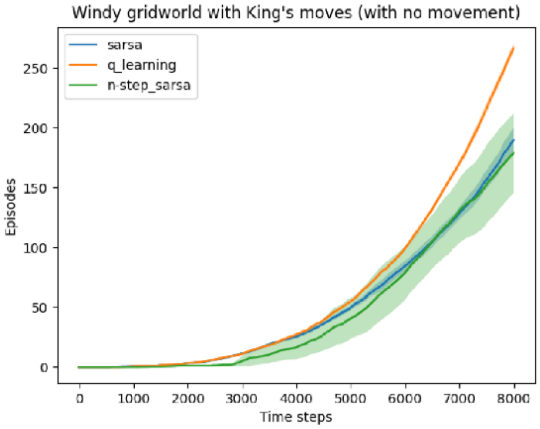
a),b)





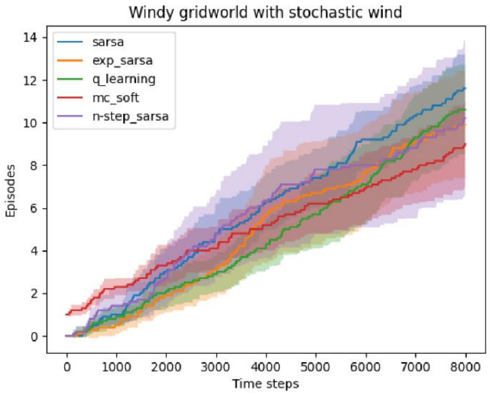
c)



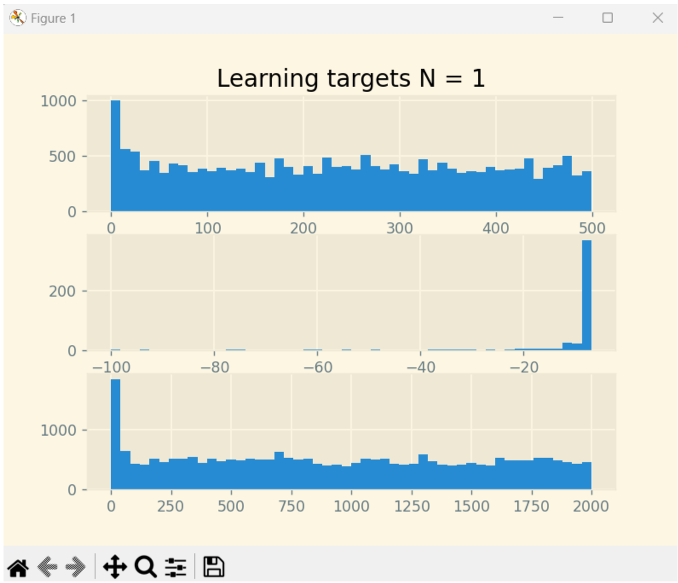


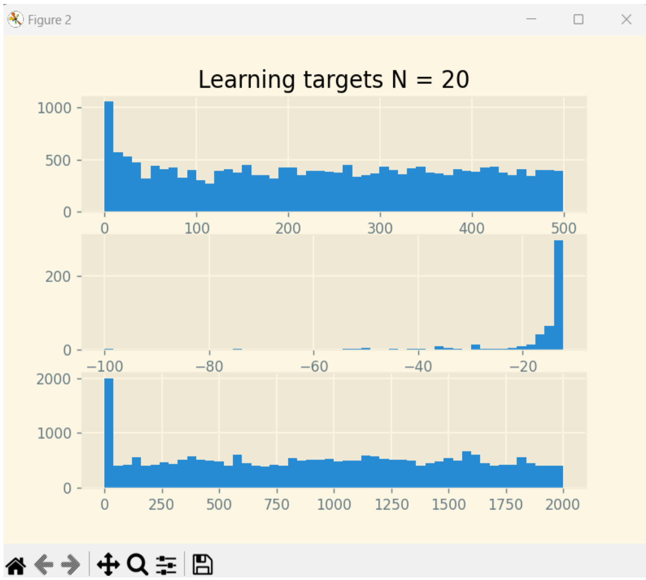
Based on the graphical data, Q-learning outperforms Expected Sarsa, which in turn exceeds N-step Sarsa, with Sarsa following and First visit Monte-Carlo ranking last.

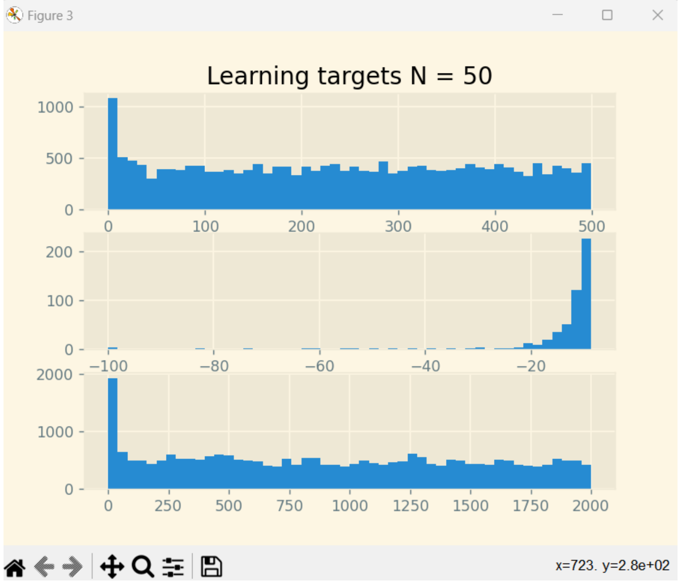
d) Every algorithm exhibits its poorest performance within the stochastic environment



4) a)







b) It is observable that as the number of learning targets N increases, the distribution of outcomes becomes more spread out, suggesting a variance shift. The histograms with lower N values show tighter clustering of results, indicating lower variance but potentially higher bias. This pattern suggests a bias-variance trade-off, where increasing N may reduce bias at the cost of higher variance.

The impact of training volume also appears evident. With more training (as seen in the progression of the x-axis), the histograms for higher N values might demonstrate a convergence trend, hinting at a possible reduction in variance over time. This implies that the amount of training could be a mitigating factor for variance, with extensive training potentially offsetting the increased variance introduced by larger N.

c)
Monte-Carlo methods remain unaffected by the training set as they do not rely on bootstrapping from state values. However, this is not the case for other algorithms, which will be influenced.