1- montecarlo would be better in the case our prior state values are not much more reliable because of some changes. So montecarlo would understant the change more rapidly and avoid bias.

2- a) Because In Q_lerng we are estimaty Q^*, the optimal action_value and we execute a's $\arg\max_a Q(s',a)$ for next iteration, Independent of the Poliy beg followed.

2-b) No_ Because Q-leary is still off-Policy and for behaivior Policy could select any Policy (ϵ-greedy) while SARSA is on-Policy and greed Selection cause it not have exploration so weight updates will be differet

3) a) No I do not think so.

No I don't think other differet value of α would have better result. either larger and smaller α have some disadvantages, for example smaller α will take long time for convergence so although it may reach a closer approximation of optimal value but it takes too long time. on the other hand, large α will cause noisy convergence and some distance from the

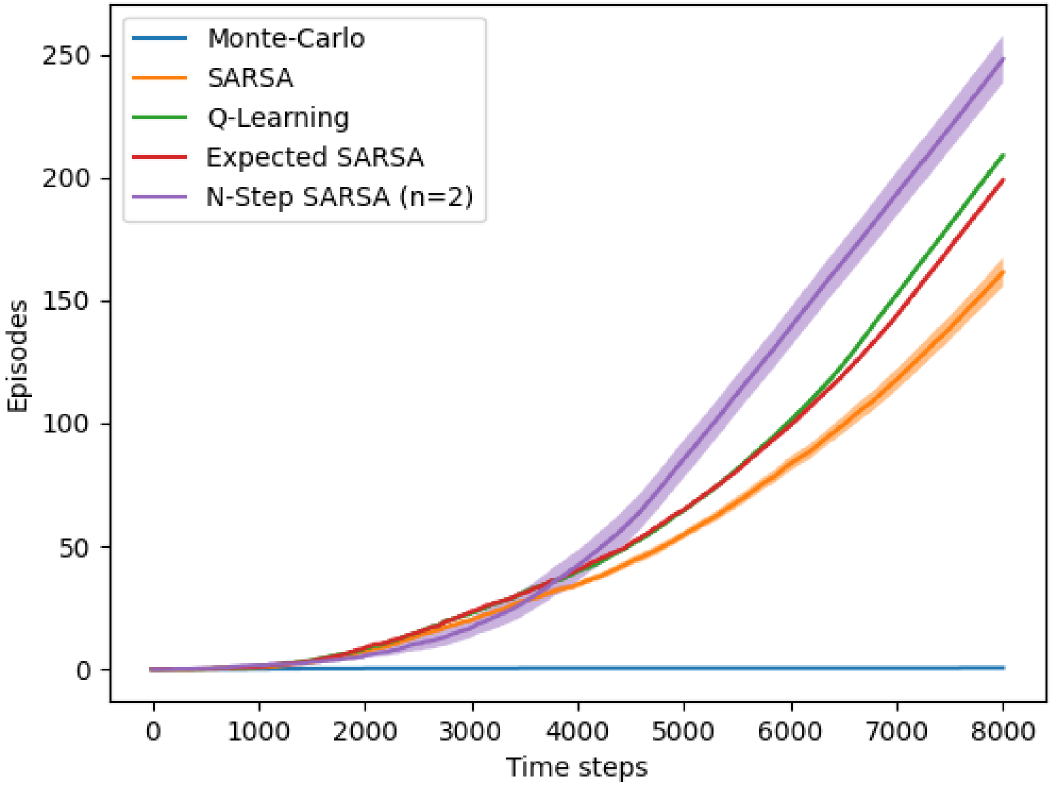
3_b) the larger α the more dependancy to next value function so in this case when α is larg and the initial values are seleted less than optimal value at initial steps our estimation will more under optimal value (under shooting) and then will be corrected. ~~of~~

3-b) Because we are using n-step we want to avoid finish early for random walk. If we have small problem, the expected value of number of states of travely will be smaller than n, making the performance less fair for larger n. So for very small problem smaller n would be better

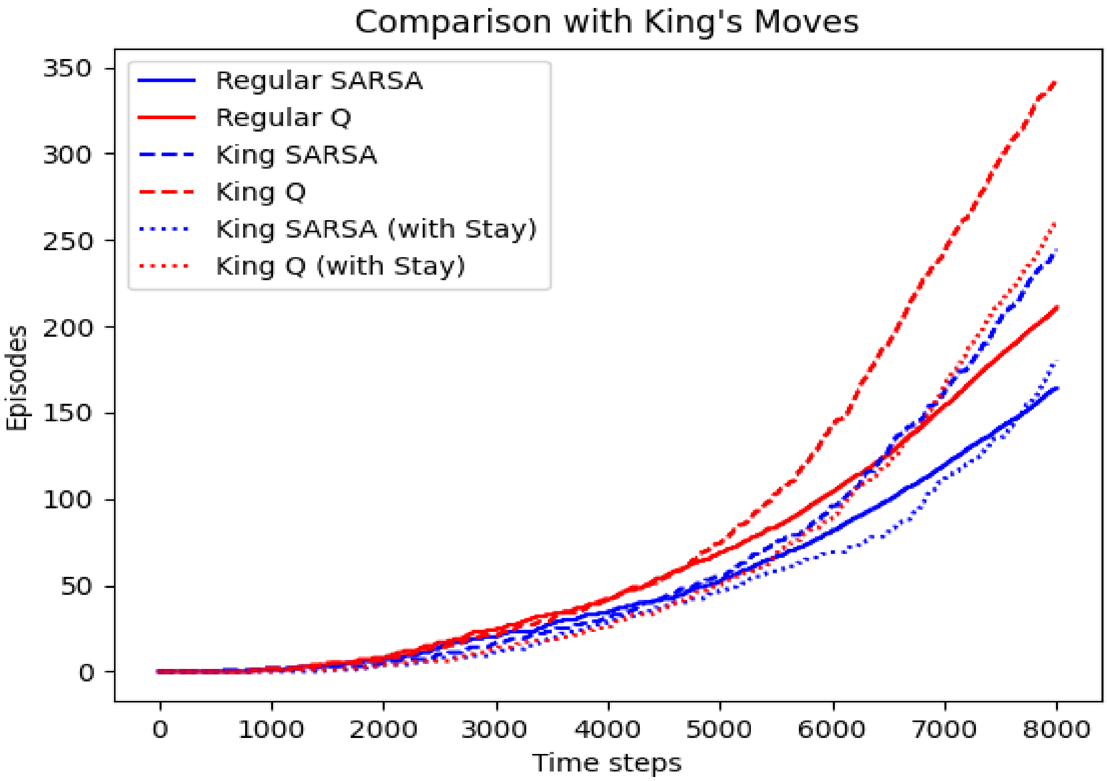
The algorithm related to Q4 and Q5 are in Algorithms file. Right now, 4a is uncommented in the main function at the bottom of the file.

Q4)

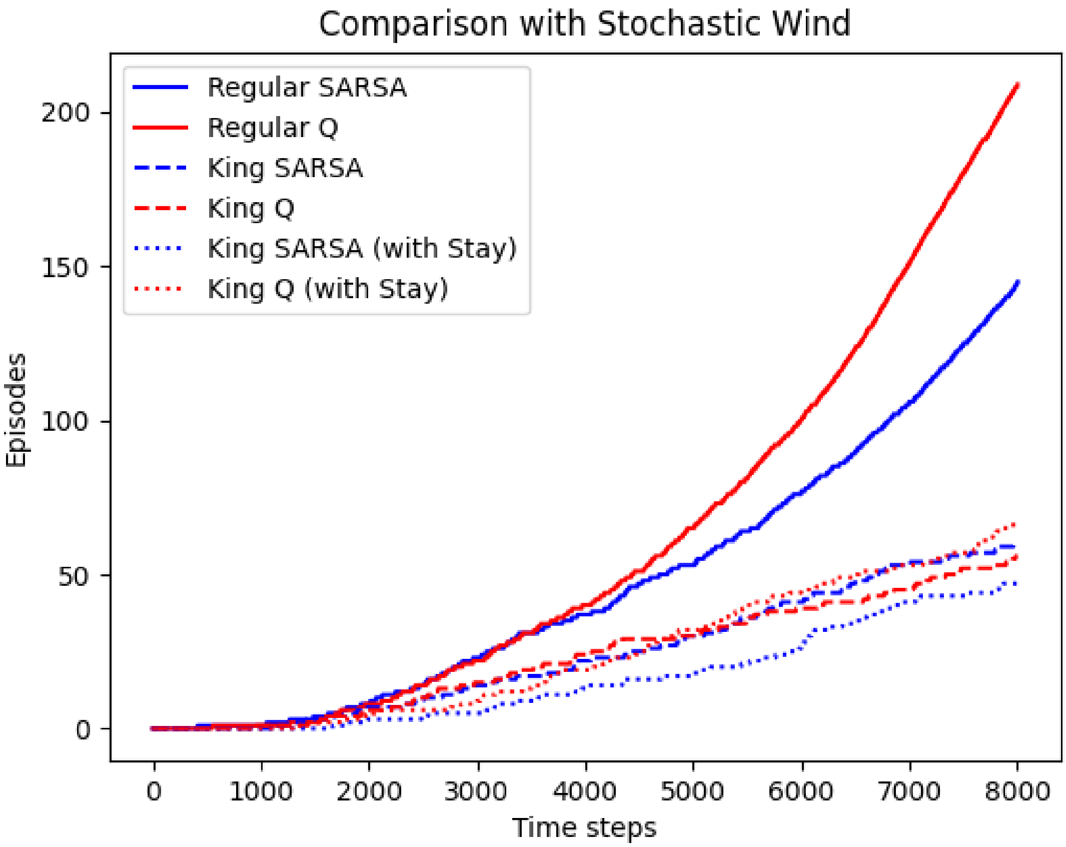
4-b-a) I used n = 2 for N-Step SARSA because it gave me better performance



4-c) Having more actions did give us better performance, as we could continue to move straight right even through weak wind by using the down-right action. Adding the stay action reduced performance, as I believe this would be a trap action to take, which means the agent wastes time learning that it is a bad action.

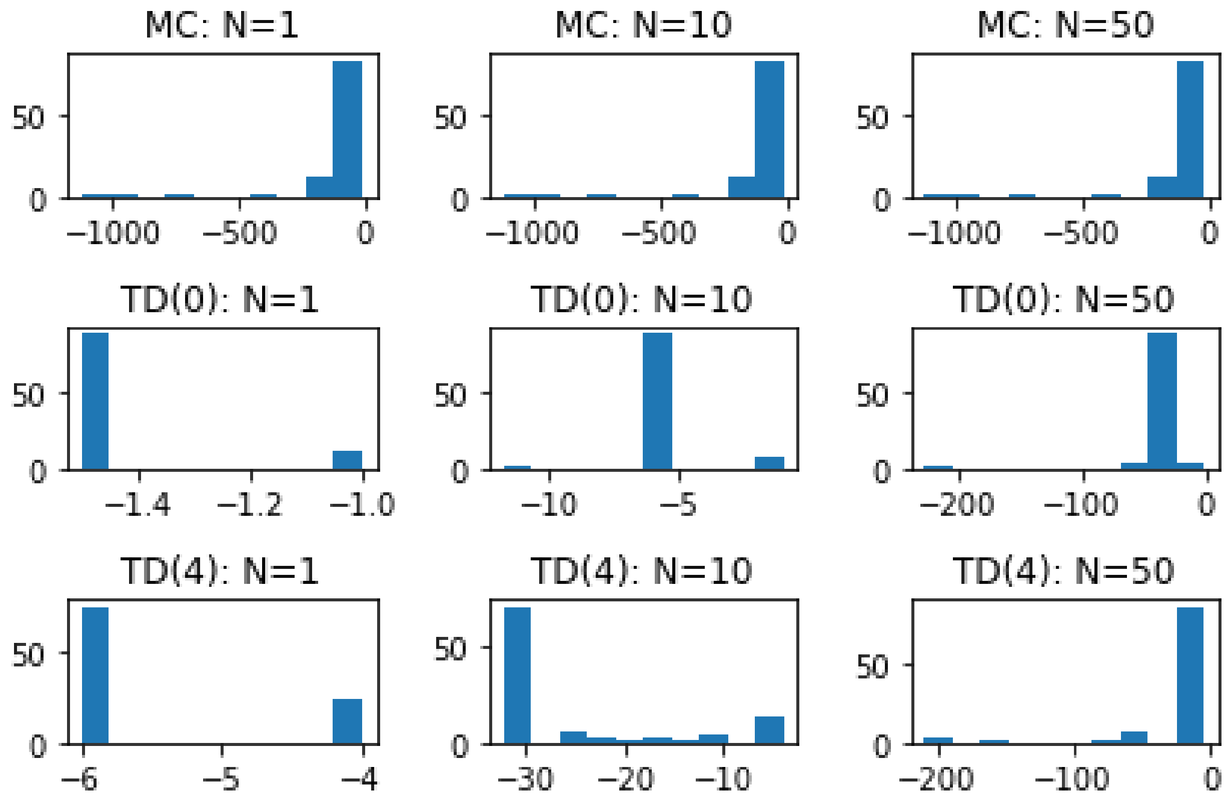


4-d) Stochastic wind causes the problem to be harder to solve


Comparison with Stochastic Wind
Regular SARSA
Regular Q
King SARSA
King Q
King SARSA (with Stay)
King Q (with Stay)
Episodes
0
50
100
150
200
0
1000
2000
3000
4000
5000
6000
7000
8000
Time steps

Q 5)

5-a)



5-b)

There seems to be high variance in the Monte-Carlo episodes, but no bias, as the plots do not change with more training episodes. On the other hand, TD methods have incredibly small variance, but high bias, as they continue to change with more training episodes