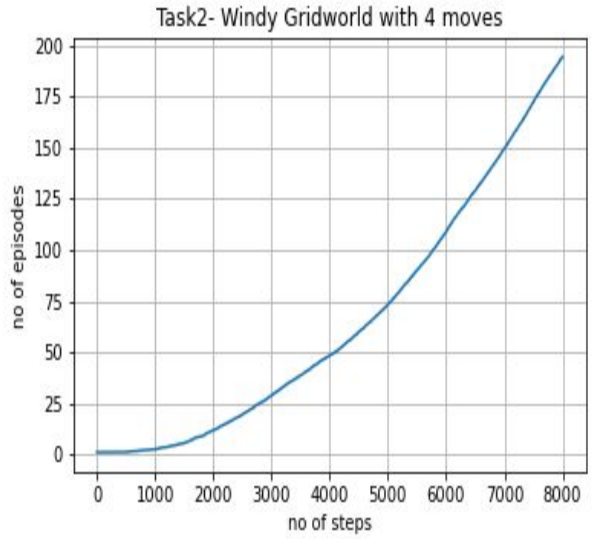
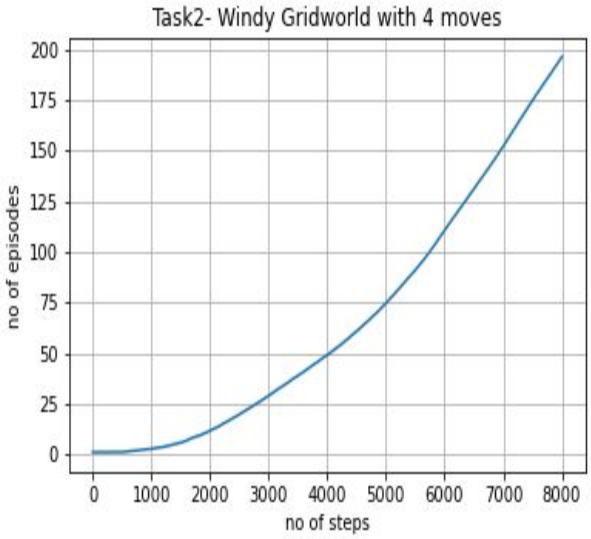
Task 2



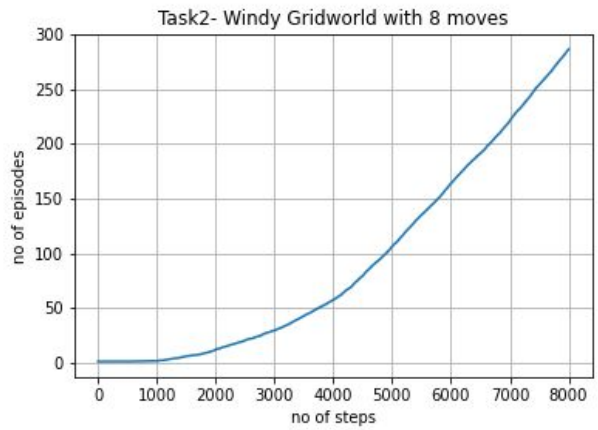


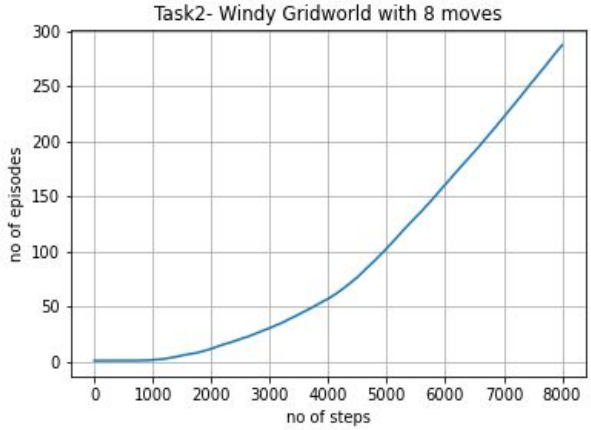
As sarsa is on policy algorithm and we don't have the optimal policy at the start so to find the optimal policy was solely dependent on the exploration done in the policy step. Thus initially it has to search more for initial episodes and as the no of episodes increases the policy gets better and better and we start getting lesser and lesser no of steps in every episode.
As clearly we can see in the graph that initially very much time is taken for every episode and later the slope somewhat becomes constant representing that the optimal behavior is nearly achieved.
We are not annihilating epsilon and that also leads to a lesser no. of steps because even after achieving the optimal policy it continues to explore.
The curve is somewhat edgy because of the randomness in taking the policy and if we have taken average over a very large no of random seeds it should become more curved.

We can also see that at about 5000 steps it takes a sudden change in the slope that is most probably due to the fact that the policy has finally found the optimal policy if not the Q_values.

One more thing I observed was that if the reward for reaching the final stage is set very high then the policy misbehaves and that is mainly due to the fact that if a wrong policy is picked initially some actions in some states ultimately gets very high q_values and it takes more time to find the optimal policy.
I took a reward for reaching equal to 1000 and after about 4800 steps it gets stuck in a loop and took about 1000-4000 steps just to complete one episode even more than to complete the very first episode which usually took around 500-600 steps.

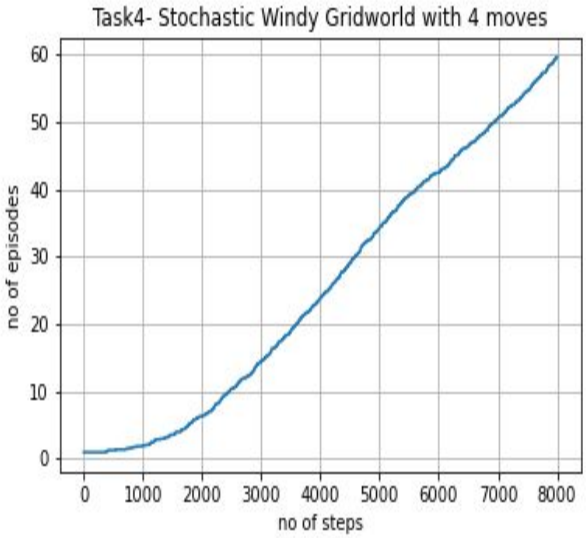
Task3

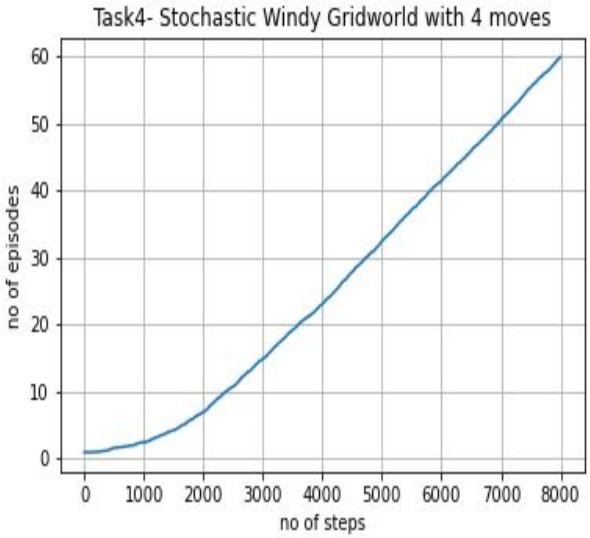




Here also we are using the sarsa method and therefore we have to take more steps initially for every episode due to the algorithm's ignorance of optimal policy. So it keeps to explore and become better. It has also got the change in the slope earlier than the previous task and also completes many more episodes than him as well. This is because of the availability of extra steps and that can be used with wind direction favoring us. I saw some episodes completing in just 9 actions for this task!

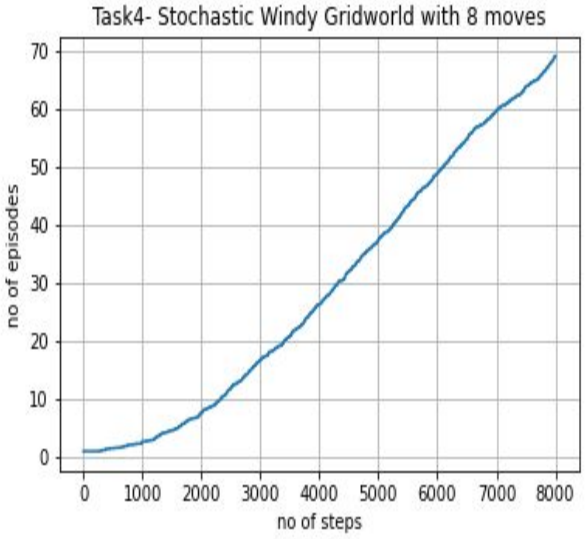
Task4

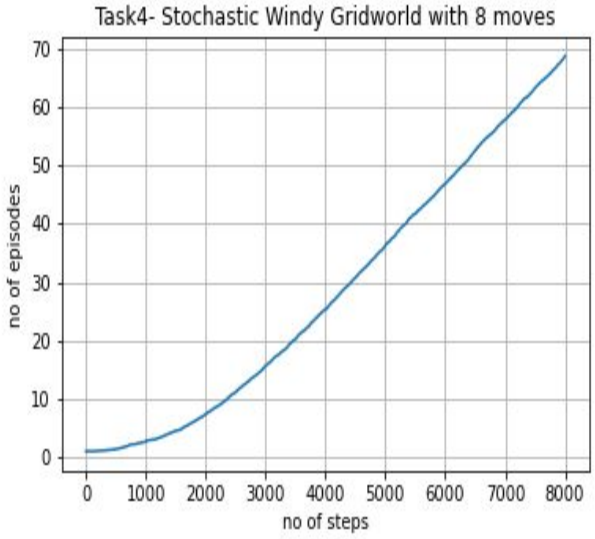




Here due to very much randomness in the environment, the policy can't get an optimal policy very soon and there is not such an optimal policy as well. So it goes with the best possible policy but that all depends on the environment and it can throw the agent to a very different stage and

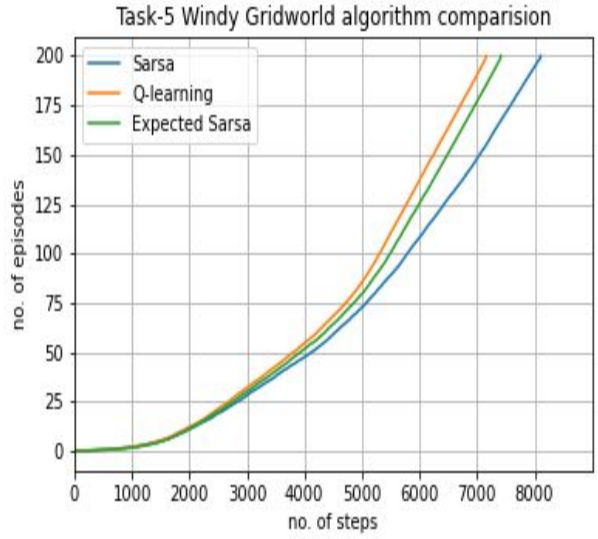
the agent has to again work to get to the target. So it takes much more time to complete an episode and thus the total no of episodes is very less. Increasing the no of random seed smoothens the curve here much more because there is too much disturbance and at less random seeds it doesn't average out much so we have to take more no. of seeds to get a better average plot.

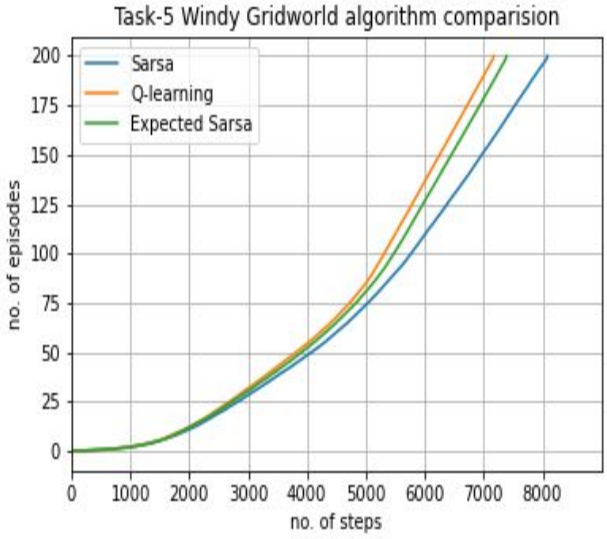




For 8 moves the explanation is the same as before but due to the availability of more moves, it takes lesser steps to complete an episode. So here no. of episodes are very less compared to the task3 but has more episodes than 4 moves.

Task5





Initially, every method performs almost similar due to a lack of knowledge of optimal policy but as the Q-learning takes actions according to the maximum q_values and next is expected sarsa as it also takes the actions according to maximum q_values but also include the effect of other actions so it is less good than Q-learning but better than sarsa which is completely on policy.