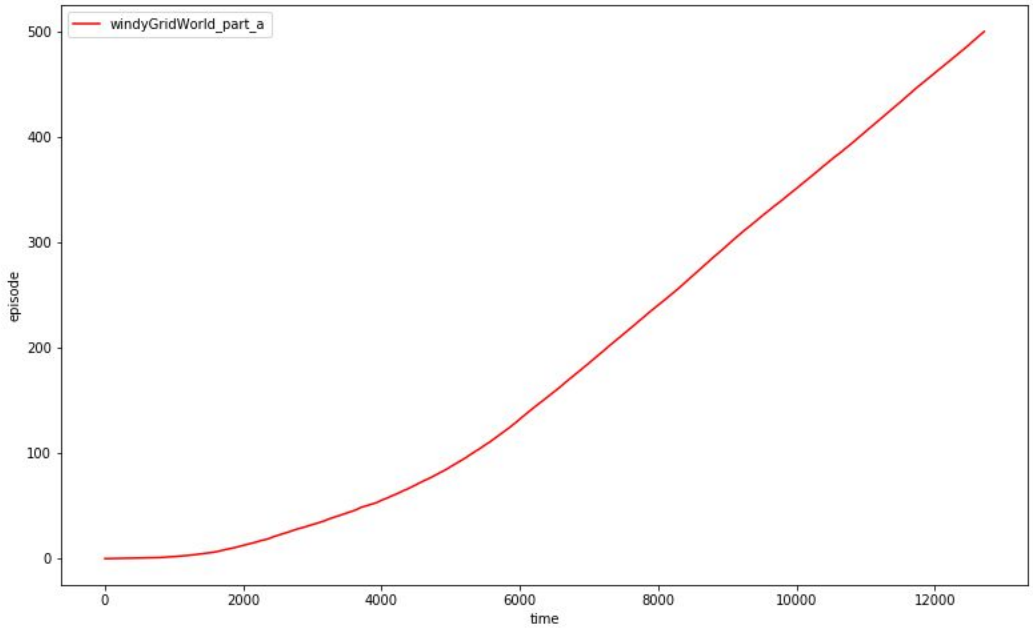
CS 747: Assignment-4

Sudhir Kumar Suman
16D070027

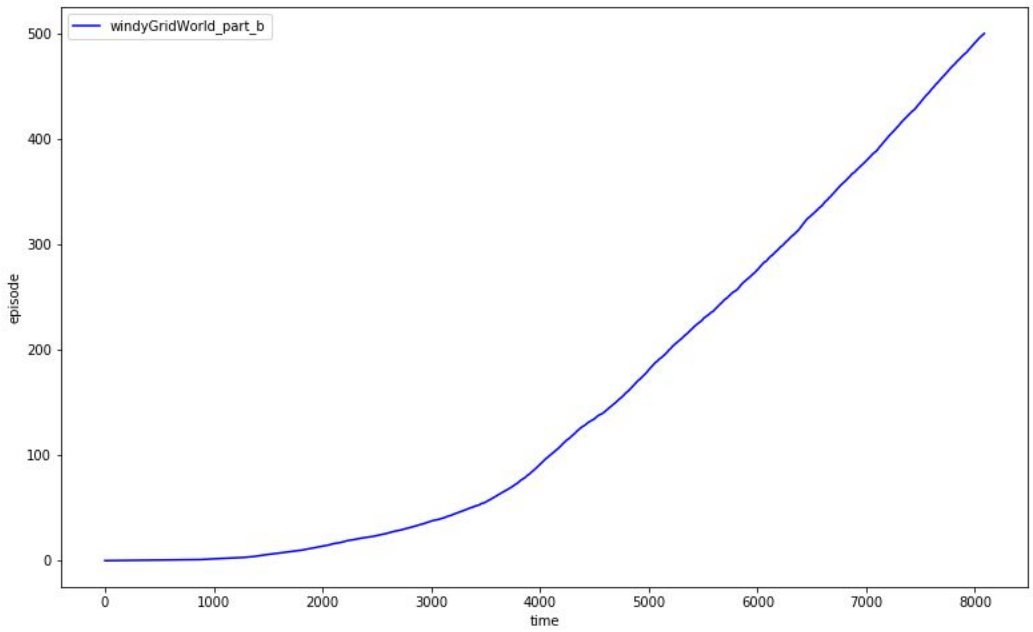
I have used epsilon = 0.05, alpha = 0.5 and decay factor as 1 in all three cases

Part a:



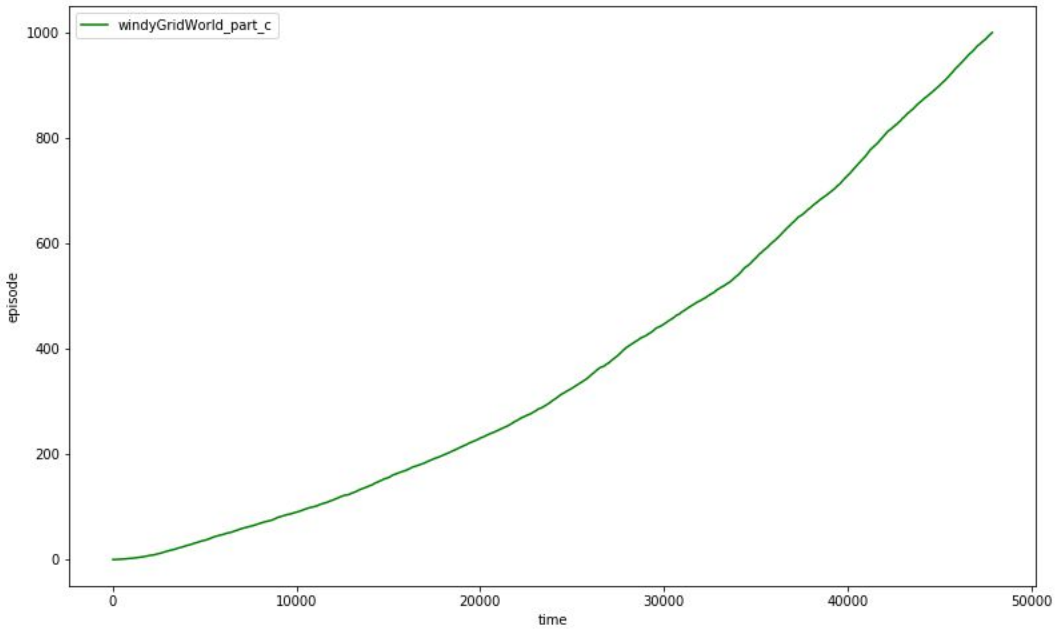
Calculated Time per episode of 17.67 in the last 10 episodes.
At the beginning of the learning process, the agent doesn't know the estimate of optimal policy and therefore it takes more time per episode, as the number of episodes increases than the agent estimates the policy better and takes lesser time and due to this time per episode decreases. After a few episodes, the slope gets almost constant as the agent gets the estimates of optimal policy. Due to fewer moves available, it takes more time per episode than taken by the agent in part b with eight moves.

Part b:



At the beginning of the learning process, the agent doesn't know the estimate of optimal policy and therefore it takes more time per episode, as the number of episodes increases than the agent estimates the policy better and takes lesser time and due to this time per episode decreases.
Due to more move, it takes lesser time than part a. The slope gets almost constant after a few episodes as the agent have an approx estimate of optimal policy.
Calculated Time per episode of 9.5 on last 10 episode

Part c:



At the beginning of the learning process, the agent doesn't know the estimate of optimal policy and therefore it takes more time per episode, as the number of episodes increases than the agent estimates the policy better and takes lesser time and due to this time per episode decreases.
Due to stochasticity in wind nature, the agent has to estimate the different best policy every time and therefore it takes more time per episode and takes longer time to converge as compared to part b.
The time per episode obtained in this case is 29.36 calculated on the last 10 episodes.

In epsilon greedy when more than one action is optimal than I choose randomly from it in order to provide more randomness.

Some Observation:

In all the three cases I tried to train the agent with a lower value of epsilon and on decreasing the value of epsilon the time per step was also decreasing and the slope of the graph was converging to constant value more quickly as compared to larger epsilon. The best time per episode of 15 and 7 for part a and part b was obtained with epsilon 0 and the time per episode gets converged to this value around 70-80 episodes.