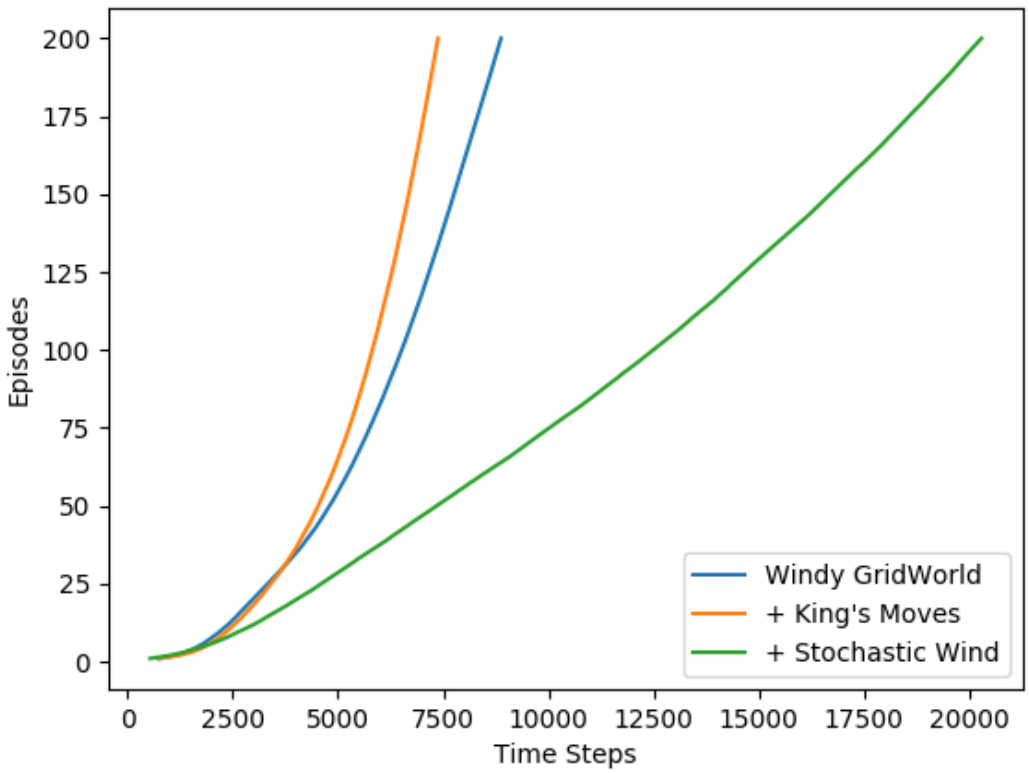
CS 747: Programming Assignment 4 REPORT

Name: Vamsi Krishna Reddy Satti **Roll Number:** 160050064

Notes about my experiments

- I have set the learning rate (called `alpha` in code) as 0.5 and exploration rate (called `epsilon` in code) as 0.1. These can be set by using the flags `--alpha` and `--epsilon` while running `main.py`.
- My experiment was run for 200 episodes in each case and statistics averaged across 150 random seeds.
- In my code, a negative value of wind means that wind is in opposite direction. Rest all the behaviour is what is mentioned in Sutton and Barto (2018).
- Also, for handling corner cases, the effects of wind and action are added and then the bounds are applied to keep the agent within the grid.

Observations



Note that the slope of the graph gives us episodes per unit time. Hence a higher slope implies the agent is performing better. Also note that the limit of 200 episodes doesn't mean convergence has happened. I make note that it is sufficient to make the following conclusions from just those episodes.

Task 1

Initially when the agent is exploring for the environment for the first time, so it takes longer time to cover an episode. As SARSA is applied and agent learns the Q value, it performs well and cover more episodes per unit time. The agent is expected to learn the optimal policy, which directs the actions such that it follows the shortest path (as given in the book). It was indeed the case and the final Q value the agent learnt when used for inference from starting state took 15 time steps to reach goal. The convergence to this value is also seen when I use an exploration rate given by $\epsilon = \frac{1}{t}$. (by convergence, I mean that the inverse of slope converges to this value).

Task2

This case is very similar to the previous case, and obviously since we give the agent more ability and choice of actions to take (*i.e.* the action space here is a strict superset of the that in task1), we expect that the agent takes lesser steps per episode on convergence. And, that indeed is the case as seen from the plot. (The slope is higher than that in task 1). Also, initially it takes more time per episode compared to task 1 since actions space is larger, it requires more exploration to learn the right actions and also has more chance to take a wrong move initially on an unexplored state. Again, on inference over the Q value learnt finally, the agent took 7 time steps to reach the goal from start state which is indeed the shortest path.

Task3

Due to stochasticity in wind, it is expected that the agent must learn to take actions (in theory) that in expectation minimize the steps taken to reach the goal from that state. Thus, it just can't simply learn the shortest path from source to goal (by shortest, I mean shortest with the wind added). Thus, the agent requires a lot more exploration to find the optimal policy in expectation as observed in reality from the plots.