CS747 Assignment 4

Gurparkash Singh 160050112

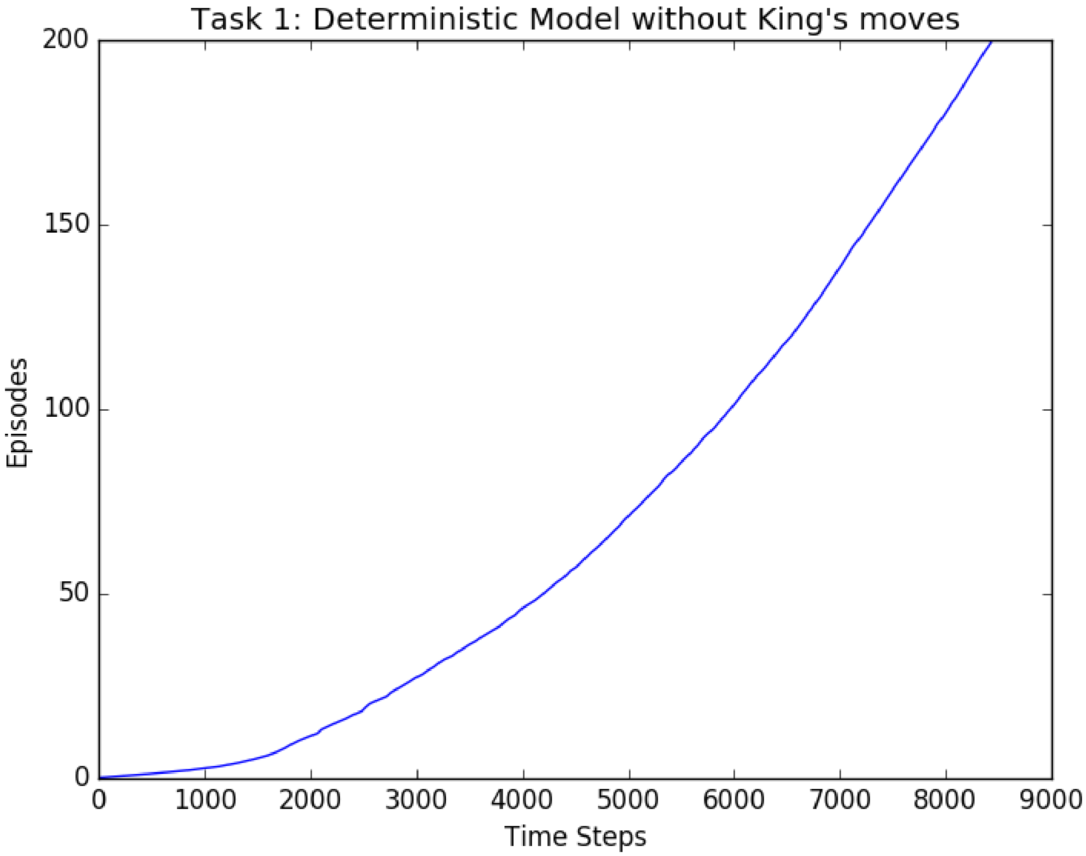
November 8, 2019

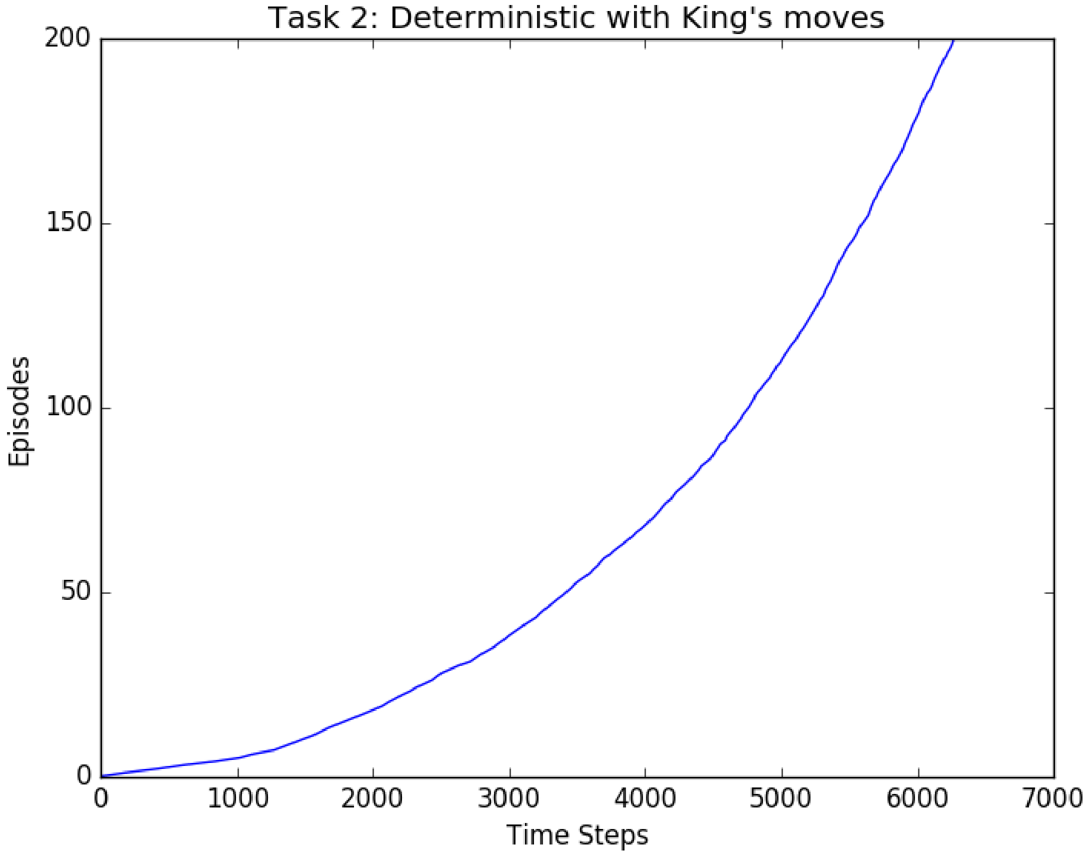
Movements Definition & Boundary Conditions

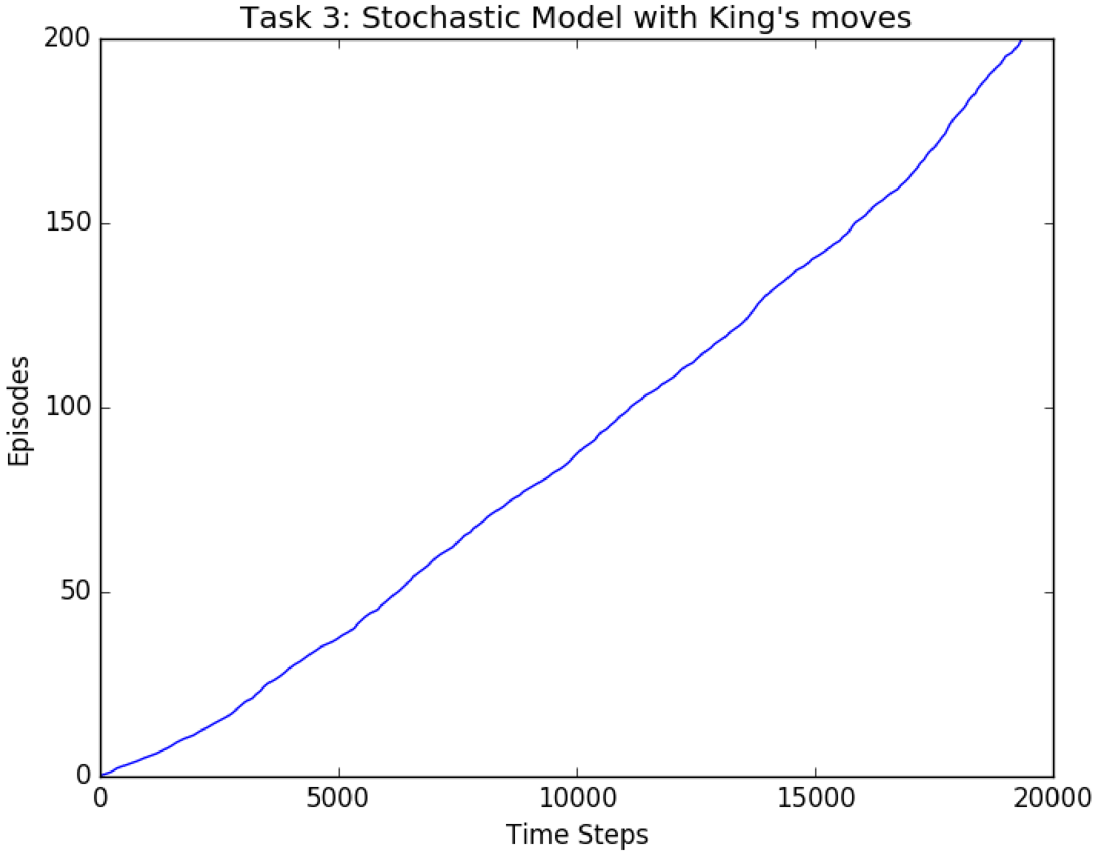
To take a step corresponding to an action, first I am calculating the "effect" of the action on the agent. For example, if the action is to move towards North-West (or Upper-Left), then the effect is (1,-1). Then, I add the effect of the wind. For example, if the wind in the current column is 1 in the North (or Upward) direction, the the effect of the wind is (1,0), thereby making the combined effect = (2,-1). I then add this to the current position. For example, if the current position is (r,c), then the resulting position will become (r+2,c-1). If the new position overflows the grid, I clamp the values between 0 and n-1. where n is the number of rows or columns. For example, if the updated position is (n+1,-1), the position will be clamped to (n-1,0). This is the final updated position, or the next state.

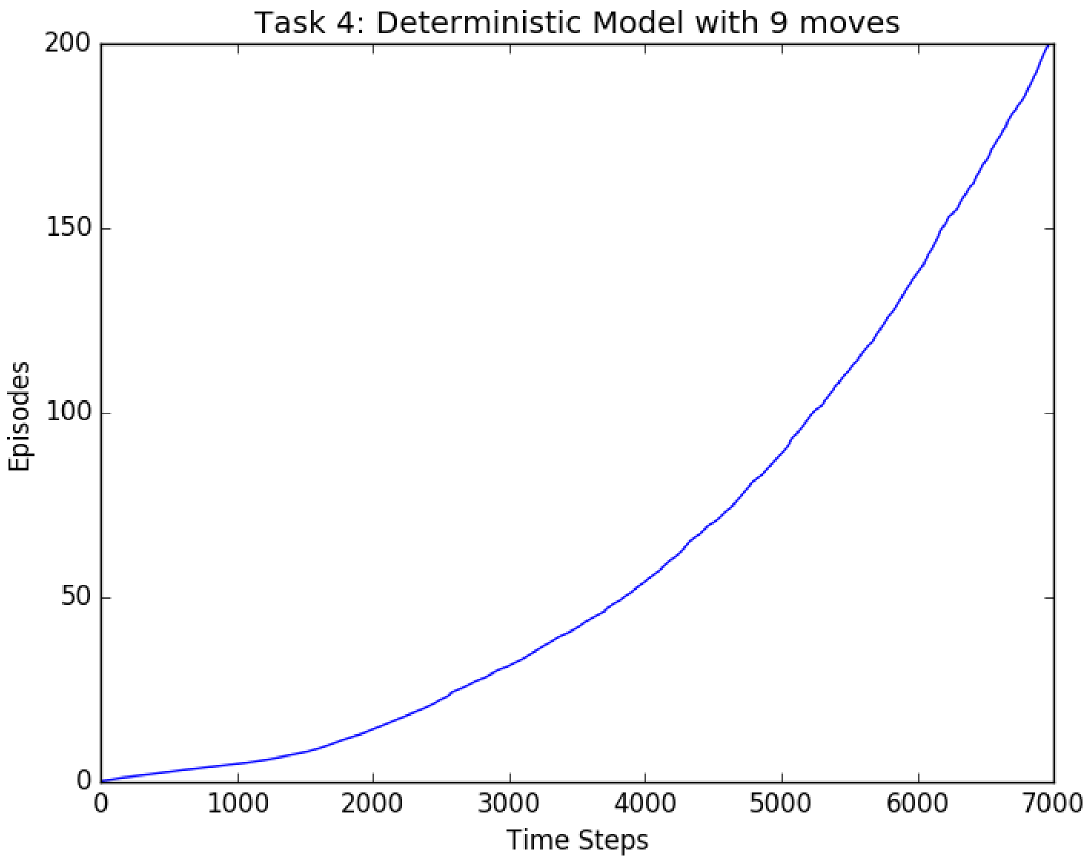
Plots

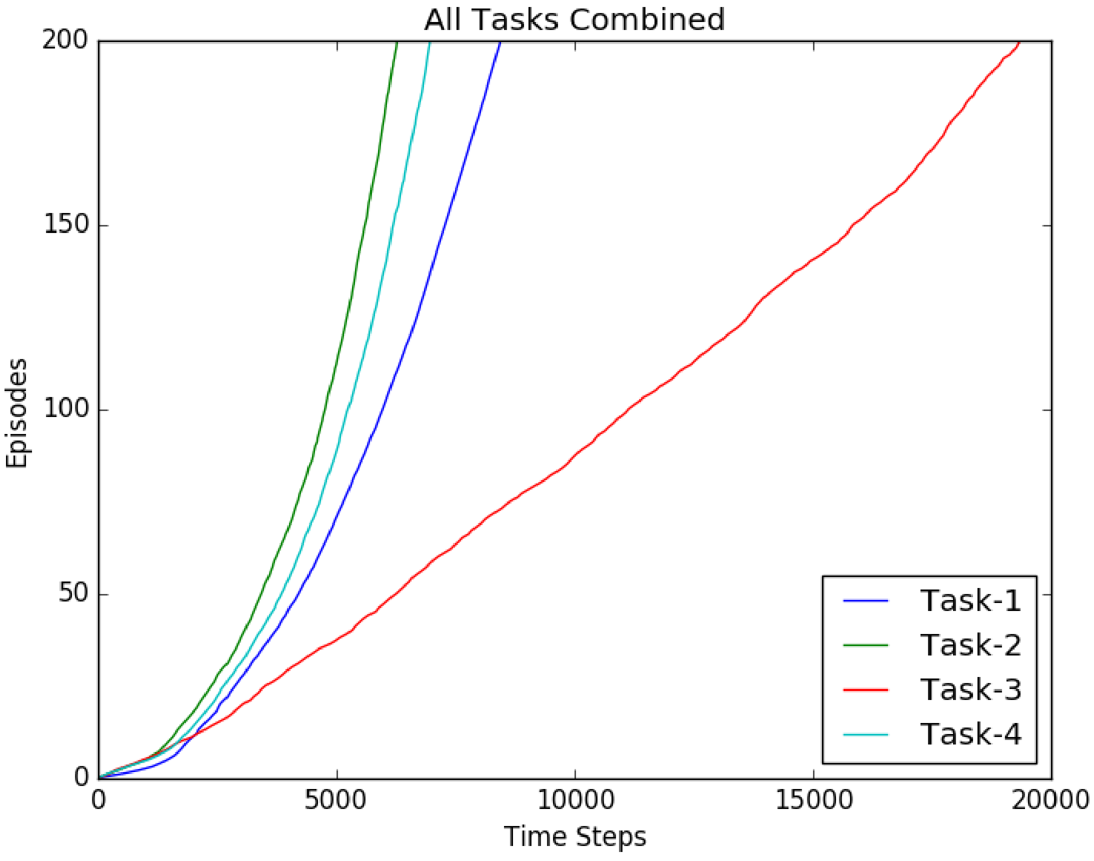
I have made plots for a total of 4 tasks. Task 1 is for a deterministic environment with 4 moves for the agent. Task 2 is also deterministic but it allows a total of 8 tasks. Task 3 adds stochasticity to Task 3. Task 4 is my own experiment which also allows the agent to take no action and let the wind carry it. The environment is deterministic for this task.




Task 2: Deterministic with King's moves
Episodes
Time Steps
0
50
100
150
200
0
1000
2000
3000
4000
5000
6000
7000


Task 3: Stochastic Model with King's moves
Episodes
Time Steps
0
50
100
150
200
0
5000
10000
15000
20000


Task 4: Deterministic Model with 9 moves
Episodes
Time Steps
0
50
100
150
200
0
1000
2000
3000
4000
5000
6000
7000


All Tasks Combined
Episodes
Time Steps
0
50
100
150
200
0
5000
10000
15000
20000
Task-1
Task-2
Task-3
Task-4

Observations & Explanations

We see that for all the tasks, the plots get increasingly steeper, i.e. the slope keeps on increasing. This shows that the number of timesteps per episode keeps on increasing as we are training. This is expected since as the agent learns, the estimated Q moves closer to the actual values, therefore, the agent starts taking optimal actions to minimize the path length.

Furthermore, we can observe from the plot for Task 3, that the number of steps per episode is the highest. This is because even though the agent learns the optimal action successfully, it is not guaranteed to be able to reach the desired state because of stochasticity. This makes it take more steps per episode than the optimal case.

Finally, if we compare Tasks 1 and 2, we see that the number of steps per episode is higher for Task 1 as compared to Task 2. This is because in Task 2, the agent has a higher number of actions. For example, if an agent has to take 1 step in both left and upward directions, it would have took him 2 time steps, whereas it would take only 1 time step in Task 2 (assuming no wind in that column). This logic can similarly be extended for the case when wind isn't 0.

However, it might be slightly surprising to see that despite having another action (to stay at the same place), task 4 has a higher number of time steps per episode. This is possibly because the new action doesn't provide the agent with more power. Still, in a separate environment, having the possibility of not taking an action can in fact decrease the number of timesteps. Consider the case the start state is exactly 1 grid below the goal state and the wind in this column is 1. If the agent has the option of not taking an action, then it will reach the state in 1 time step. However, if that action is not available, it will have to take a longer route.

I tried tuning the values of parameters alpha and epsilon by varying them from 0 to 1. The value of gamma was simply kept to 1. Closer to the extremes (0 and 1), the algorithm didn't perform very well whereas closer to the given values in the book, most values gave similar results. So I simply used the values given in the book, i.e., alpha = 0.5 and epsilon = 0.2.