1) Average steps 2512
2) Value Iteration converges much more quickly than Policy Iteration

Value Iteration Steps:

Varied with gamma, 440 for 0.96 , 1666 for 0.99 (Initial Gamma Value = 0.1 Increase by 0.05, till >=1, complete data in notebook)

Varied threshold and saw the differences in optimal policies.

theta	diff vals
0.1	10
1.00E-02	9
1.00E-03	8
1.00E-06	7
1.00E-10	9
1.00E-15	4
1.00E-20	2
1.00E-25	2

Time taken 12.055 (for theta = 1e-20), 4.963 (for theta = 1e-5)

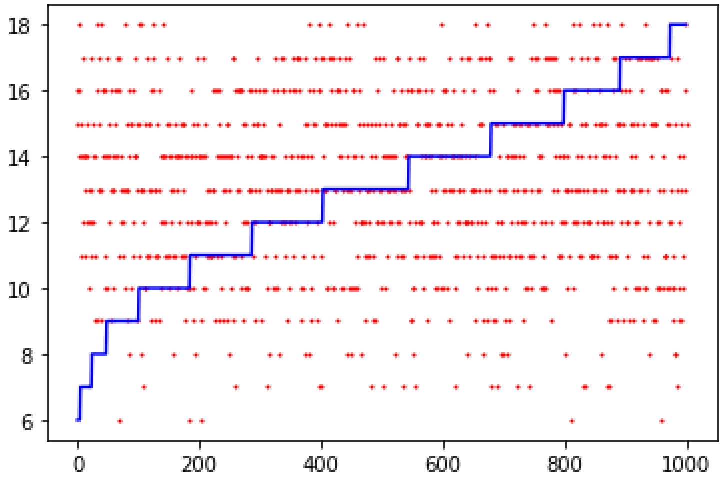
Policy Iteration Steps:15 (Very little variation wrt gamma)

Time: 96.61

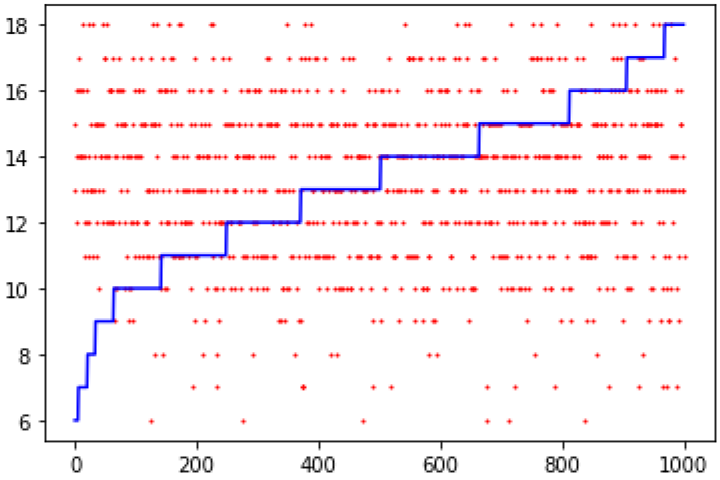
Policy wasn’t exactly the same for theta = 1e-20, 50 states had different actions out of the 500 states.

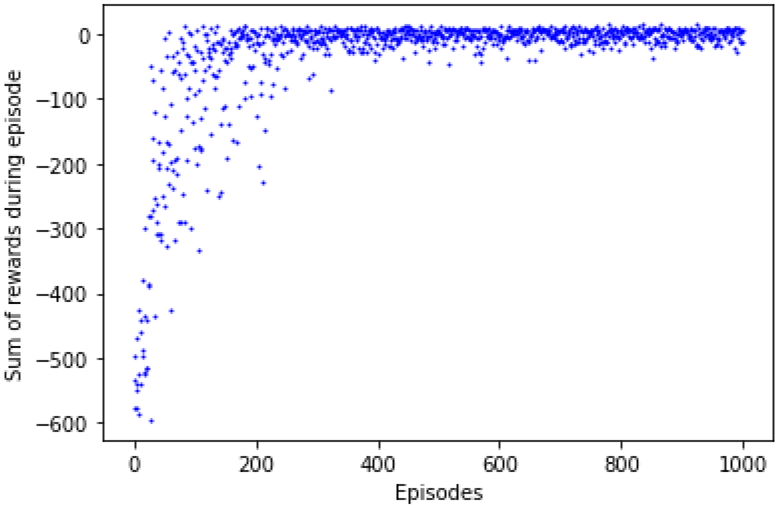
For theta=1e-5, they were exactly the same.

Value Iteration: steps vs episodes, Avg 13.061



Policy Iteration: steps vs episodes, Avg 13.261

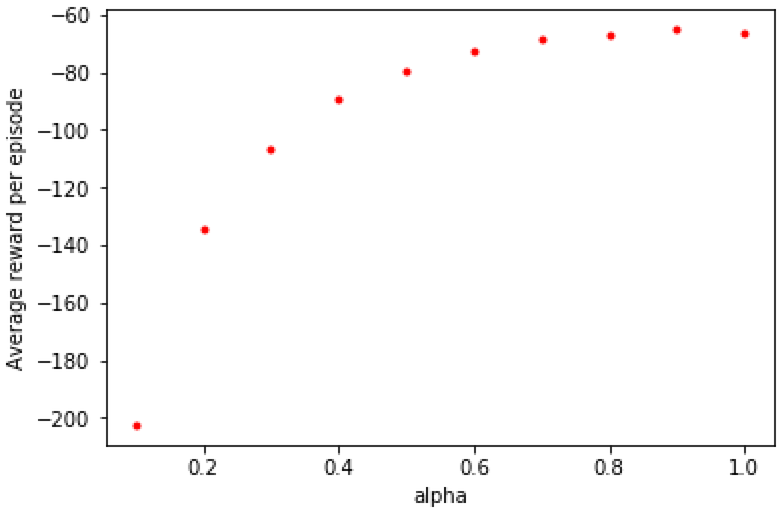




3)

Idea of convergence, as sum of rewards during episodes start converging.

Alpha graph(using 500 episodes):



4) Refer Notebook

References:
For Plotting Q4: https://github.com/SamKirkiles/

Others:
http://gym.openai.com/docs/
https://github.com/rharish101
https://github.com/ceteke/RL