**046203 – Programming Homework 2**

Dar Arava - 205874951

Yair Nahum - 034462796

**Question 1 – Blackjack**

1. The problem described as a Markov decision process –

Each state consists of two variables where represents the sum of the player’s cards and represents the first dealer’s cards.

The player is initially dealt with two cards therefor the minimal value of is (two s). The maximal sum the player can reach without losing the game is . The dealer’s first card is between and .

Overall the state space is-

In addition, there are three terminal states - therefore - .

The action space is .

The reward for each state is 0 except for the terminal states where –

We do not discount ().

Since the dealer’s policy is given (and independent on the players sum), we can calculate the probability of the final sum of cards for the dealer- marked .

Where B is the value of the card added (generated every time ) and is distributed (independent of ) -

Both and can be initially calculated. The problem’s transition probabilities based on these distributions will be-

All other transitions are with zero probability.

1. As described in section a, initially we shall calculate the probabilities of the total sum reached for the dealer - .

The possible values of are (based on the given policy) –

Where is the case where .

The result after applying the given policy -

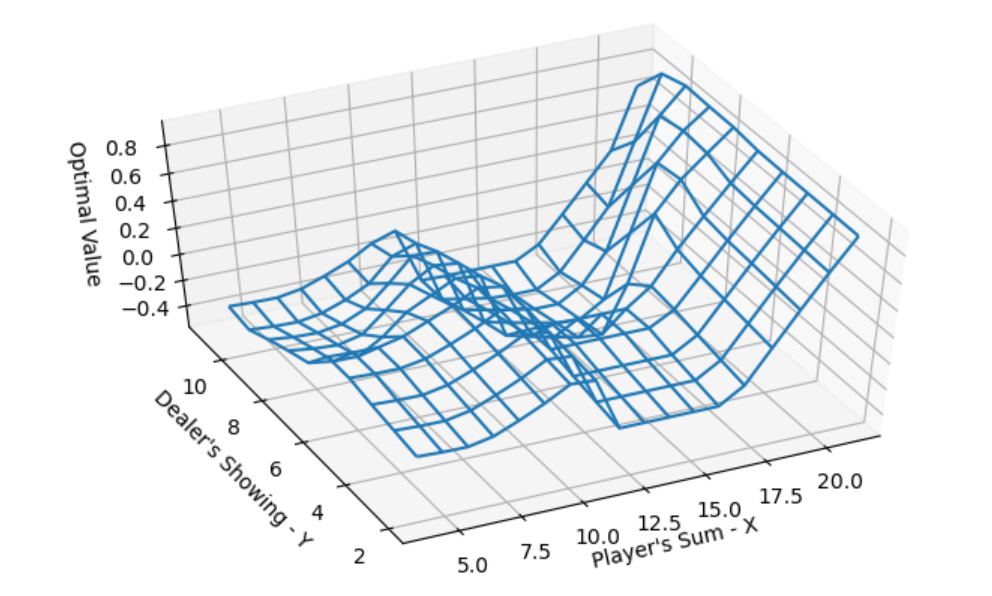
Now we have all the needed possibilities distributions in order to solve the Bellman Equations –

For terminal states –

For non- terminal states -

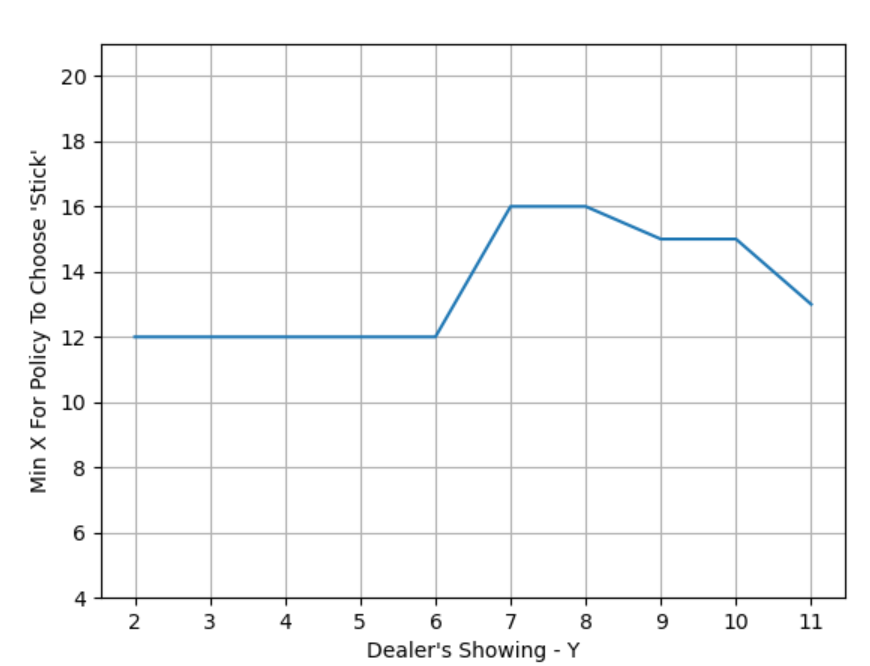
After running the VI algorithm for iterations we converge to an optimal value .

For example, after iterations –



1. After obtaining the optimal value - we must now derive an optimal policy. By using the Greedy operator-

We receive the following result (plot of the minimal value of X for which the policy choses action ‘stick’):



**Q2 – Server CU Rule**

**Part 1 - planning**

1. The state space:  
   }}  
   The state’s defines the remaining jobs that were not served yet.  
   The start state is   
   Therefore, we have   
   As we can select at each stage s whether some job was finished(not in s) or wasn’t (in s)

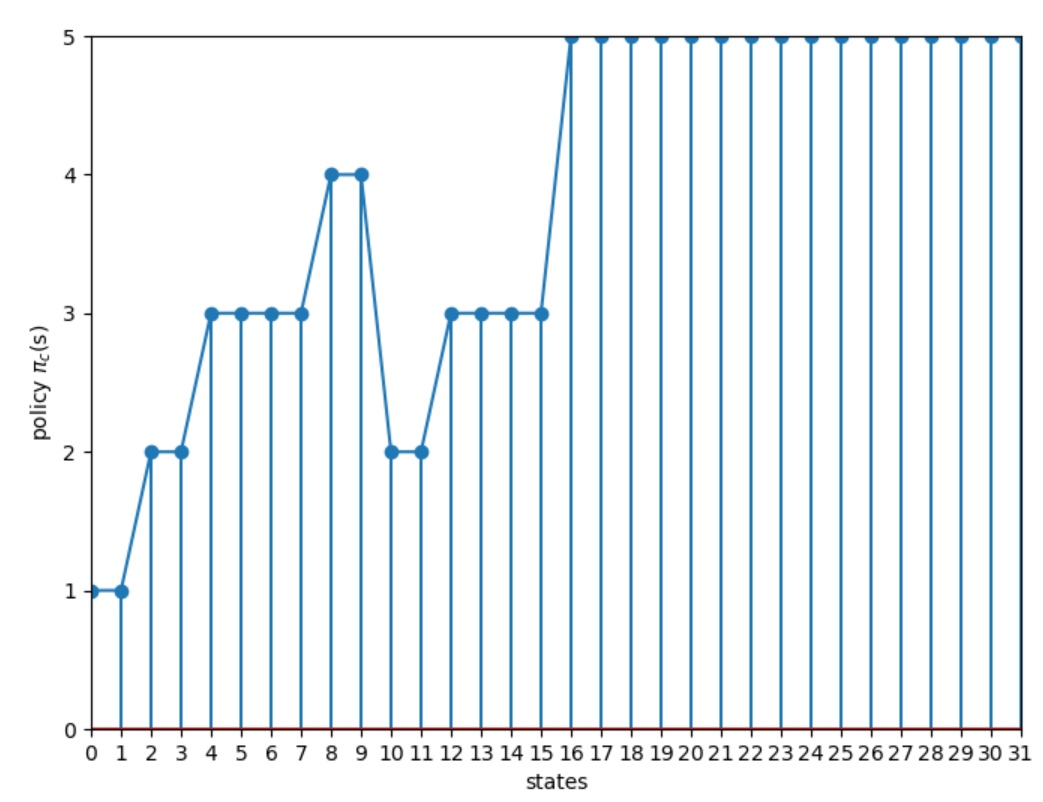
In our case, N=5 so we have 32 different states.

The action space:  
}  
The action index defines the job that the server tries to serve at that time index.  
It cannot be an index that doesn’t belong to the state at that time index.

1. The cost per time step is as follows:  
   We will calculate the value function in python using fixed policy value iteration algorithm.  
   We will use a N bit id per state. In each state, a lit bit defines that the job is unfinished.  
   So in our case the starting state is 31 and the terminal state is 0.

We will use the fact that this is a finite state problem, with a terminal state as absorbing state. Meaning, with probability 1 we will get to the terminal state (with cost 0 defined for it).   
Thus, we can backward recurse in order to compute the value function of the fixed policy.  
The equation for value iteration as we calculated in previous assignment is:

1. Plot of the values of the policy that selects the job with the maximal cost , from the remaining unfinished jobs-



As described in previous section, the representation of state in the plot above is a decimal number matching the binary number where bit if in state job is not yet finished(still in system).

1. For applying the policy iteration algorithm, we need to implement now the policy improvement stage. In our case, as we saw in previous assignment, the bellman equation is as follows (in order to minimize cost):

And the improved policy -

In each iteration in the PI algorithm, we first calculate the Value for policy -

(using FPVI).

We apply the policy improvement above and update our policy to the improved policy found - .

We calculate the Value for policy - (using FPVI).

If we have not improved the value reached by policy and we achieve-

.

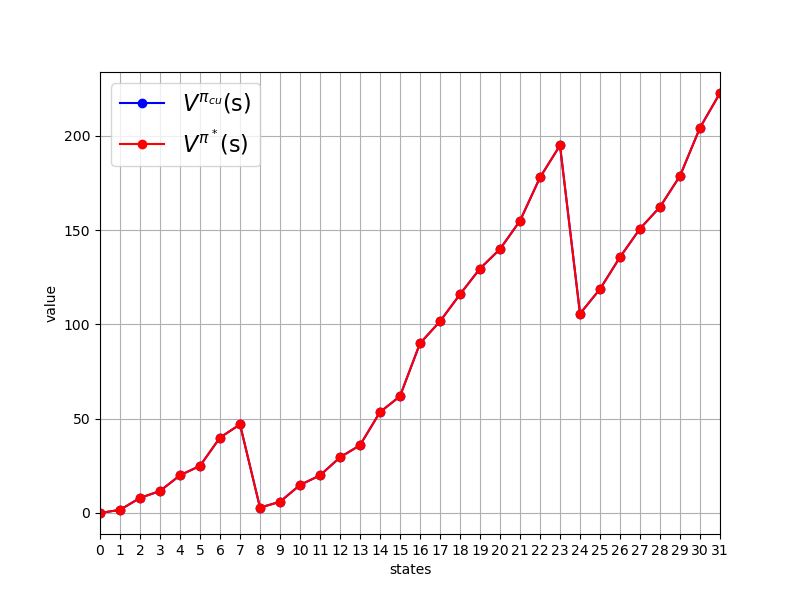
Otherwise, we return the steps for iteration (with policy ).

For initial iteration we set- and got coverage to optimal policy after 2 iterations.  
We actually improved the policy only once (the second time it didn’t improve)

Plot below shows value over iterations of the algorithm (.

A screenshot of a cell phone

Description automatically generated

1. After obtaining ,we get the same value as for the optimal policy. Thus, we can see that the cu rule gives an optimal policy.  
   

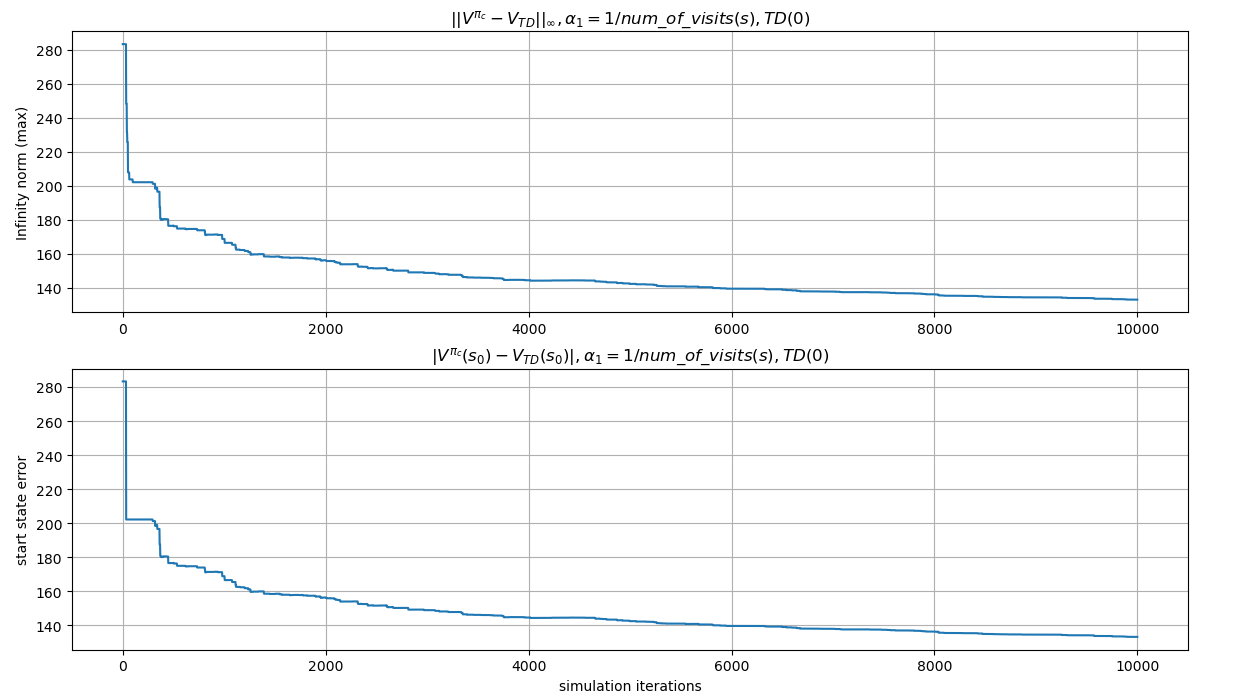
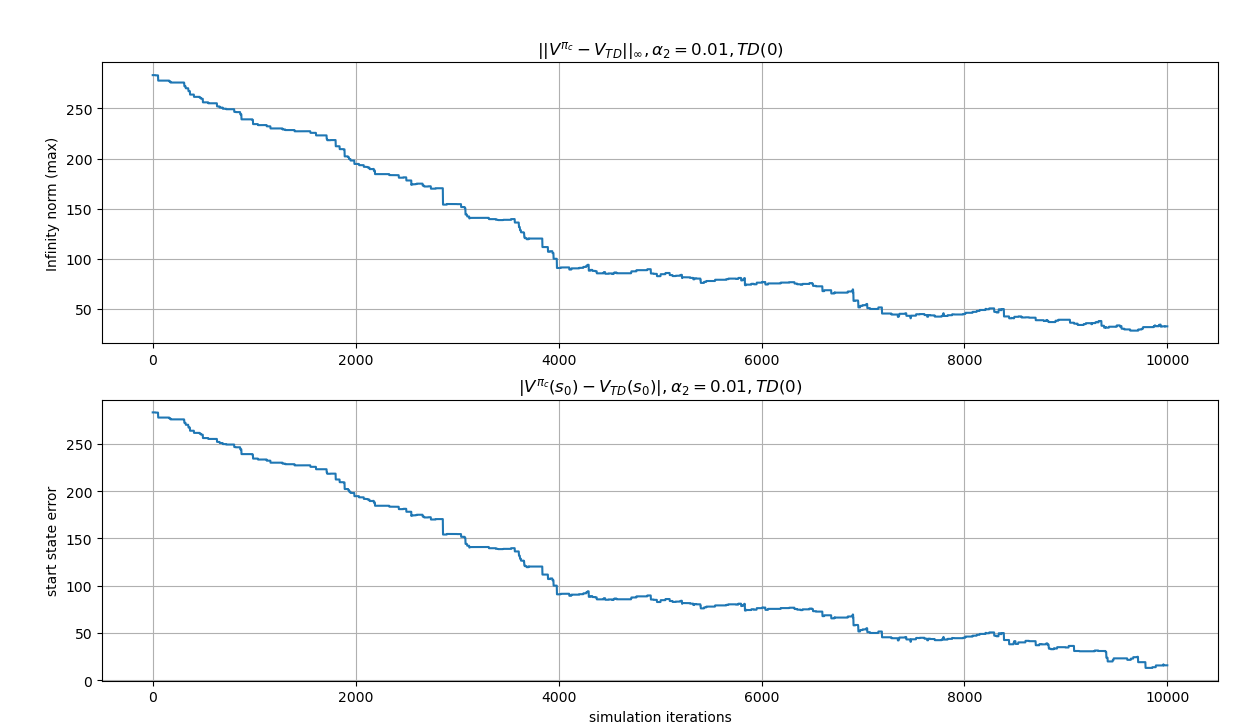
Plot of vs. :  
A screenshot of a cell phone

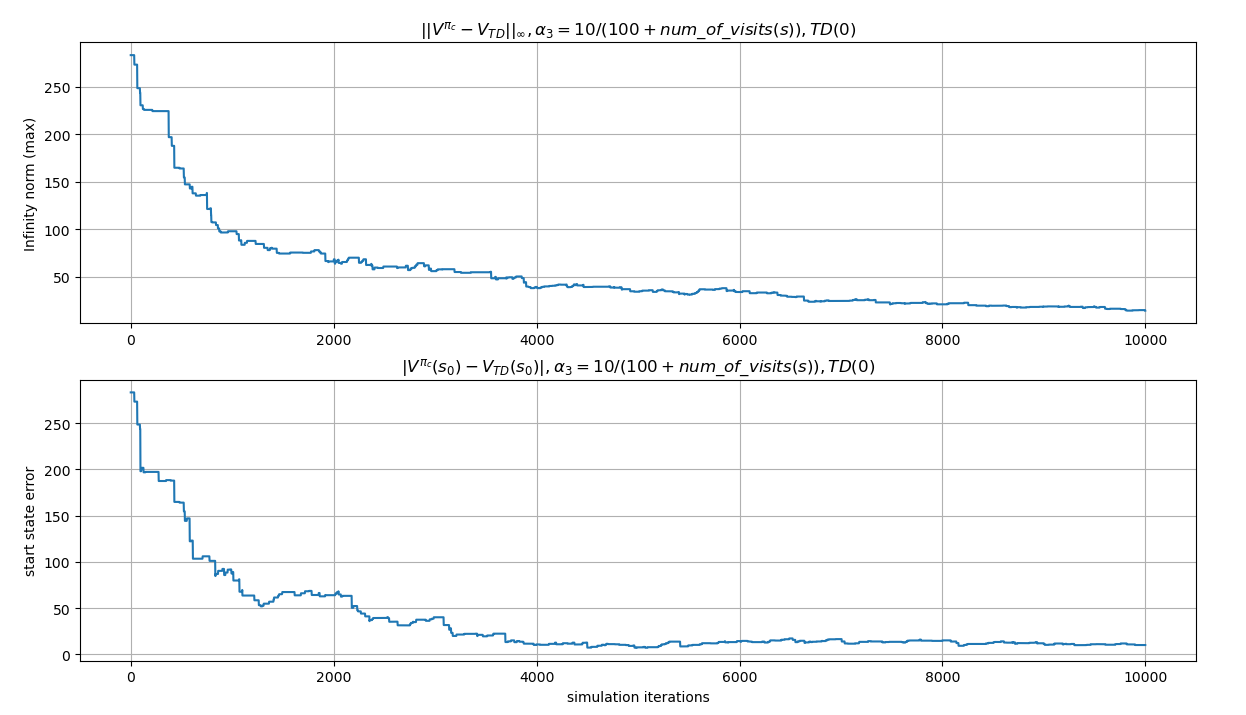
Description automatically generated

For all states as expected of optimal policy.

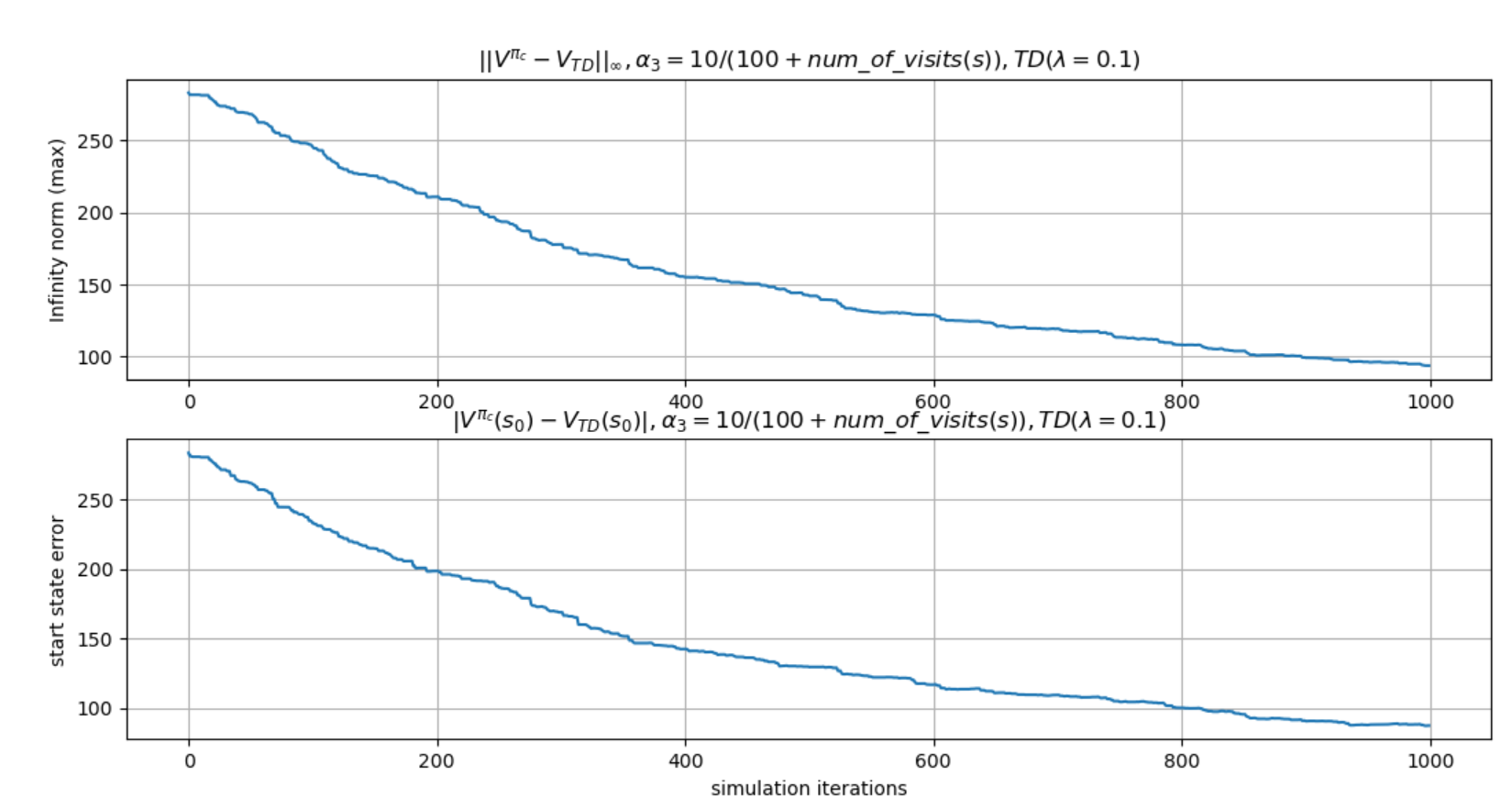
1. The simulator function implemented (assuming ) –

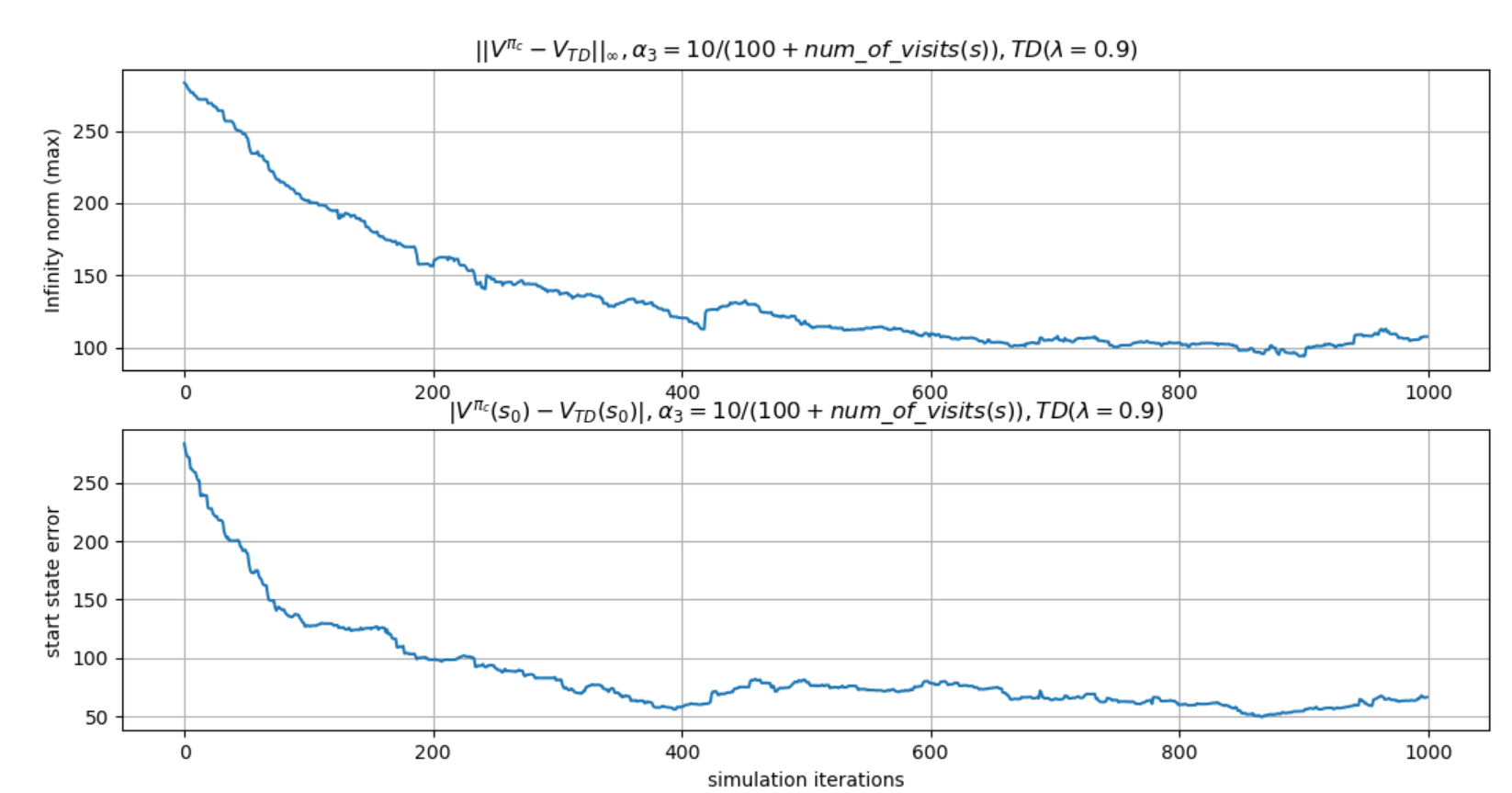
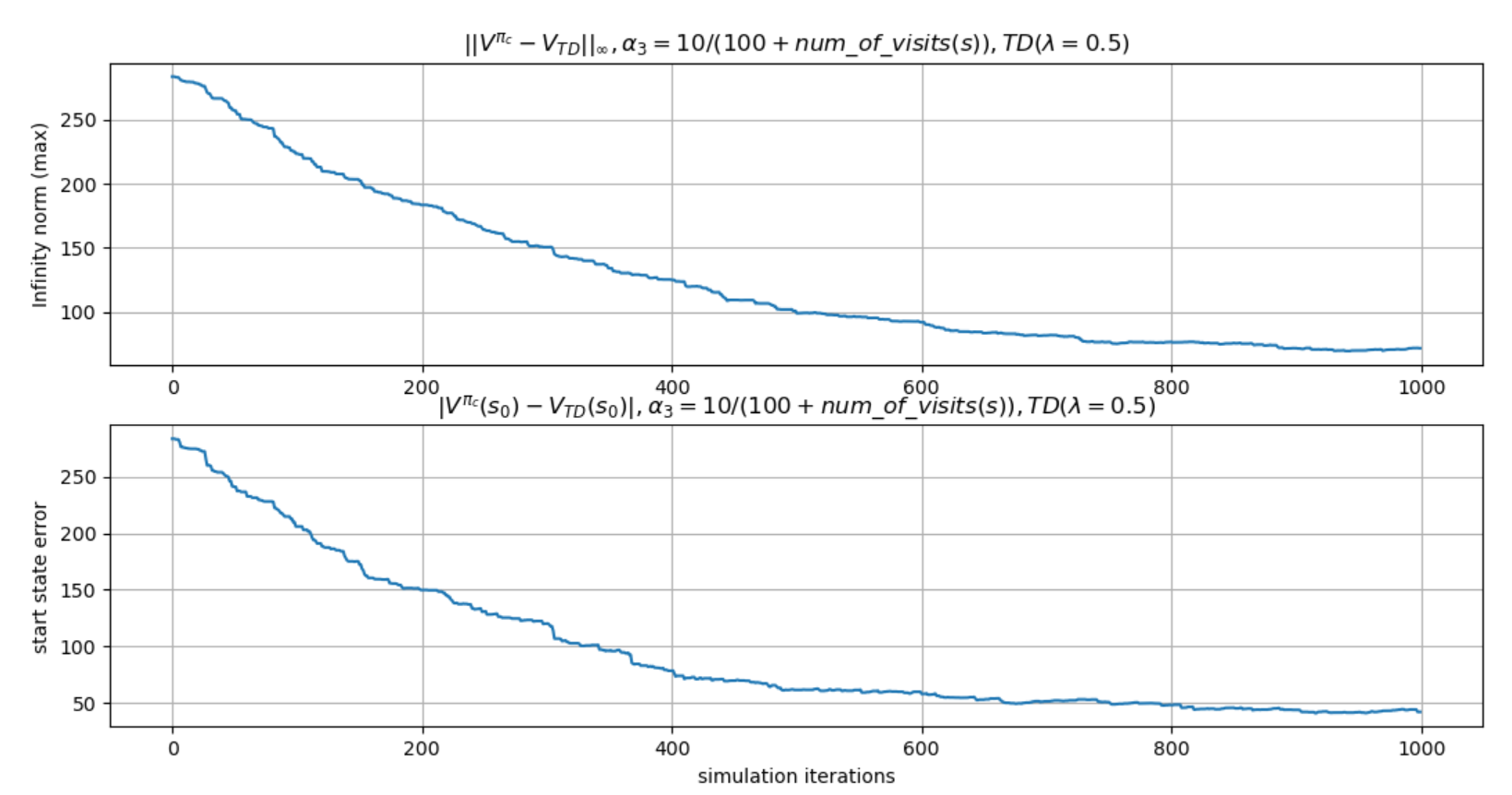
**Part 2 - learning**

1. We’ve run 10,000 iterations. In each iteration we’ve run an episode that start in a random state and finishes at the terminal state.  
   The first step size- is actually a moving average, as we saw in the lectures. It basically does averaging on all occurrences in which we simulate over some stage. We can see the error remains high as alpha gets low too rapidly, thus the update of the new value is relatively small and the convergence is slow due to it.  
   plot:  
     
     
     
   The second step size- remains constant to all iterations. Thus, the convergence relatively keeps the same high rate to convergence in the long term. At start it’s slow, but in the long term it’s much higher than the other steps tested (get smaller as time goes by). The problem is, it’s not really the average between estimations as in the first step size and may be noisier.  
   Plot:  
   

The third step size- starts with ~0.1 alpha and the pace in which the step size getting smaller is less sharp than the first step size. Meaning, it’s some kind of a merge between the previous steps - . We keep high rate of convergence for a longer period of time, but also does averaging between value estimations in the long term.  
Plot:  


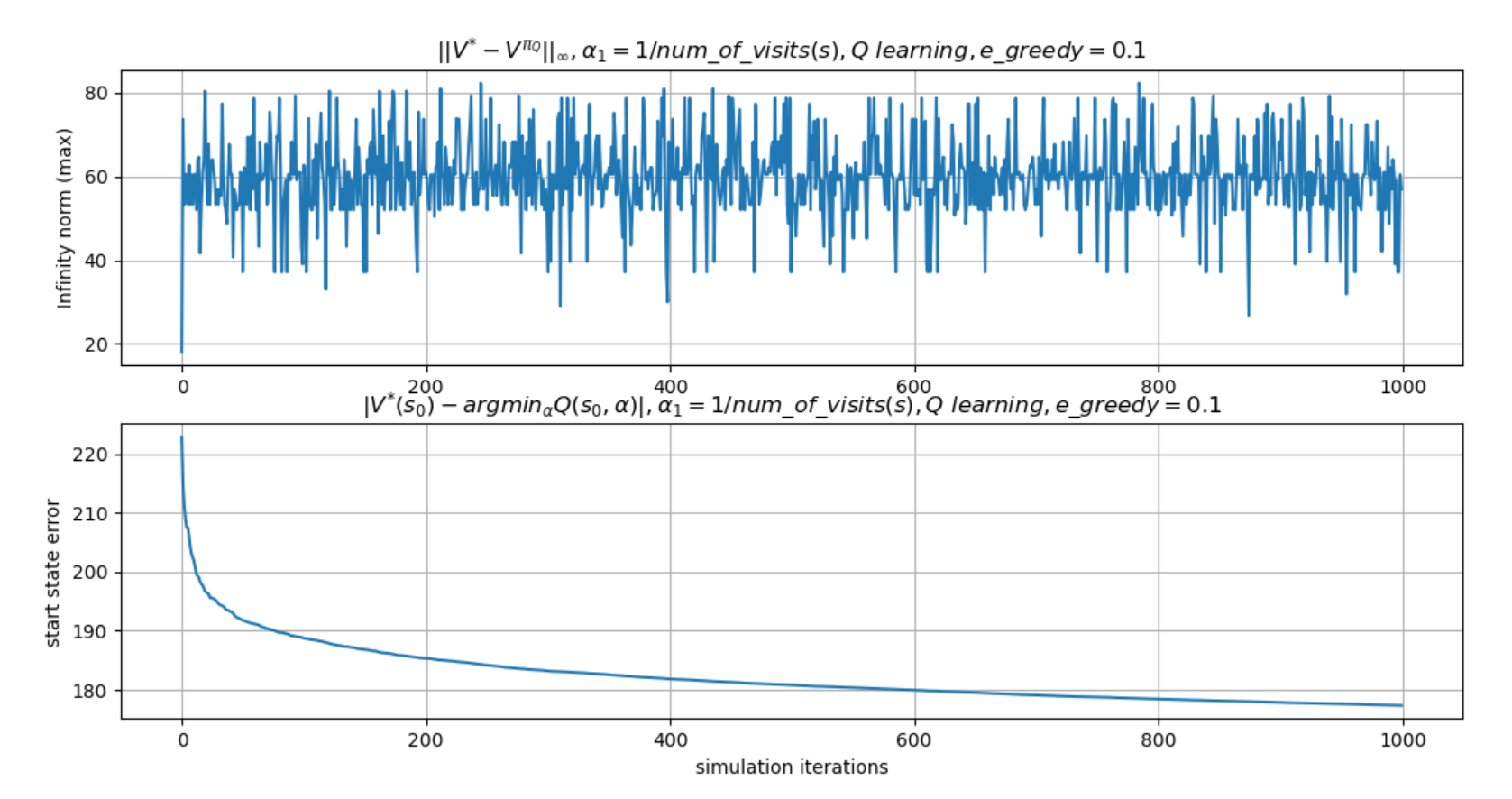
1. We’ve tested the third step size using different values for lambda (0.1,0.5,0.9). Plots of the average results( 20 runs) for :

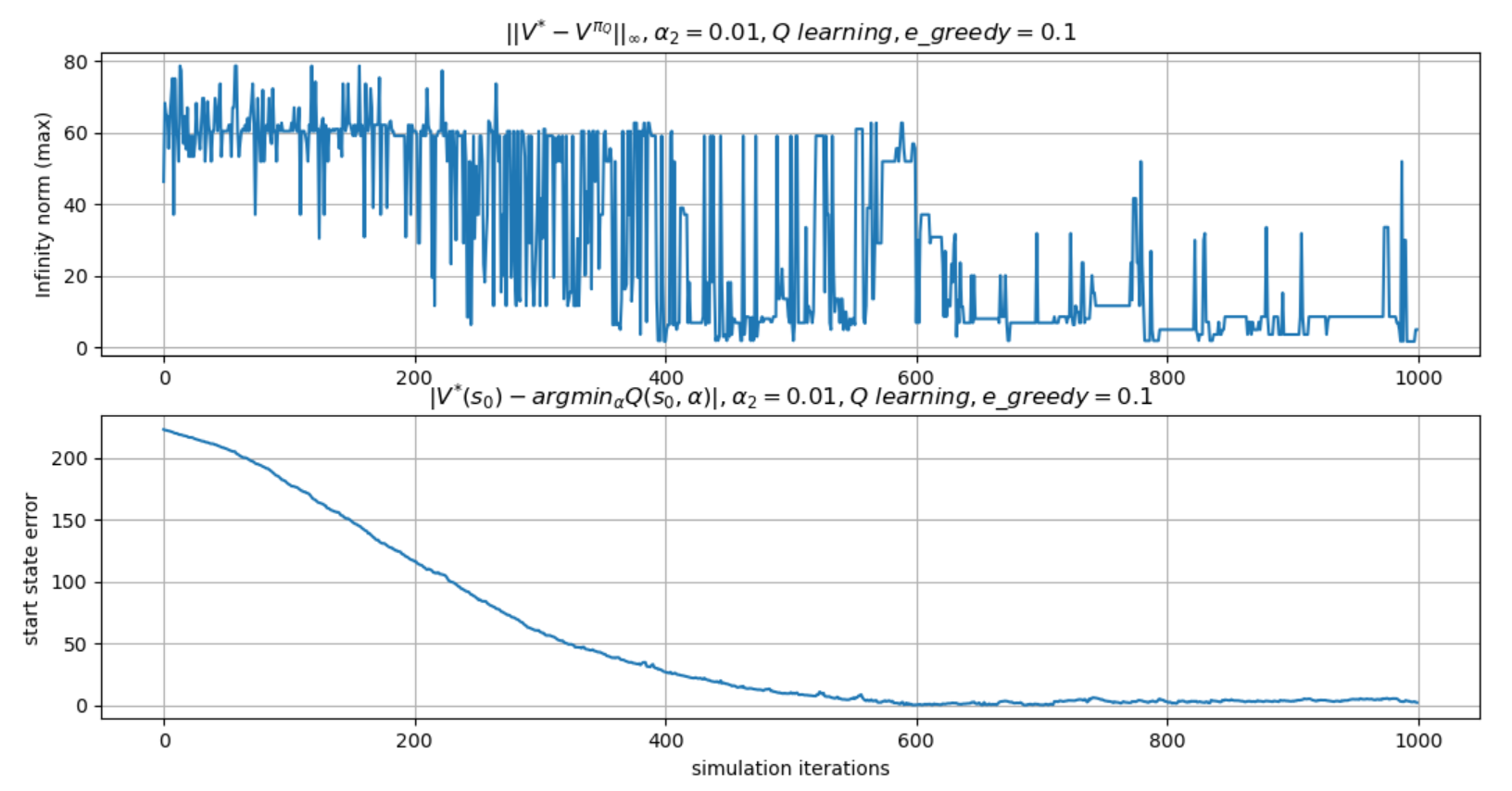


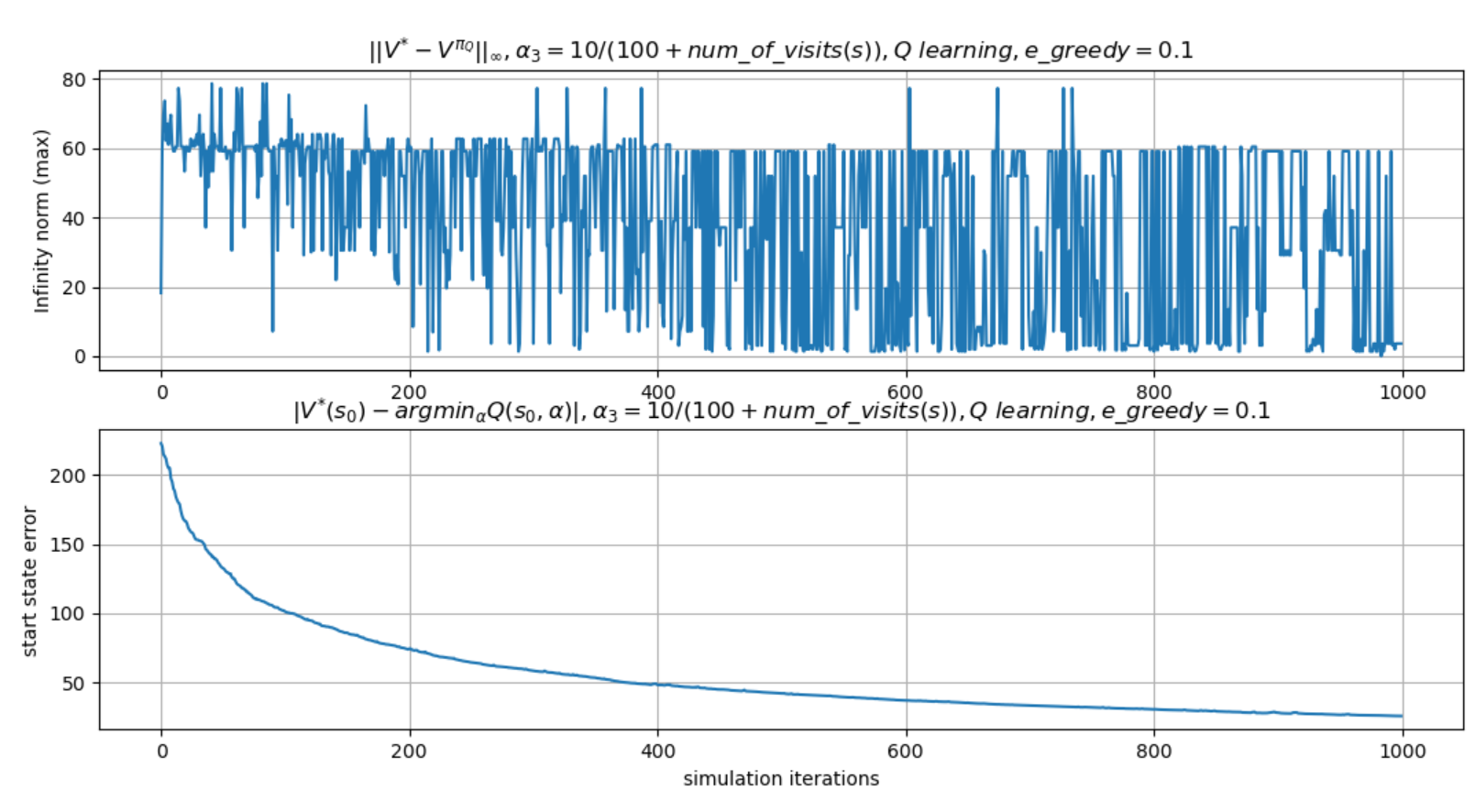


1. Results of the Q-Learning for over 100,000 iterations-

Displayed in plots are 1000 sampled points (sample for every 100th iteration).

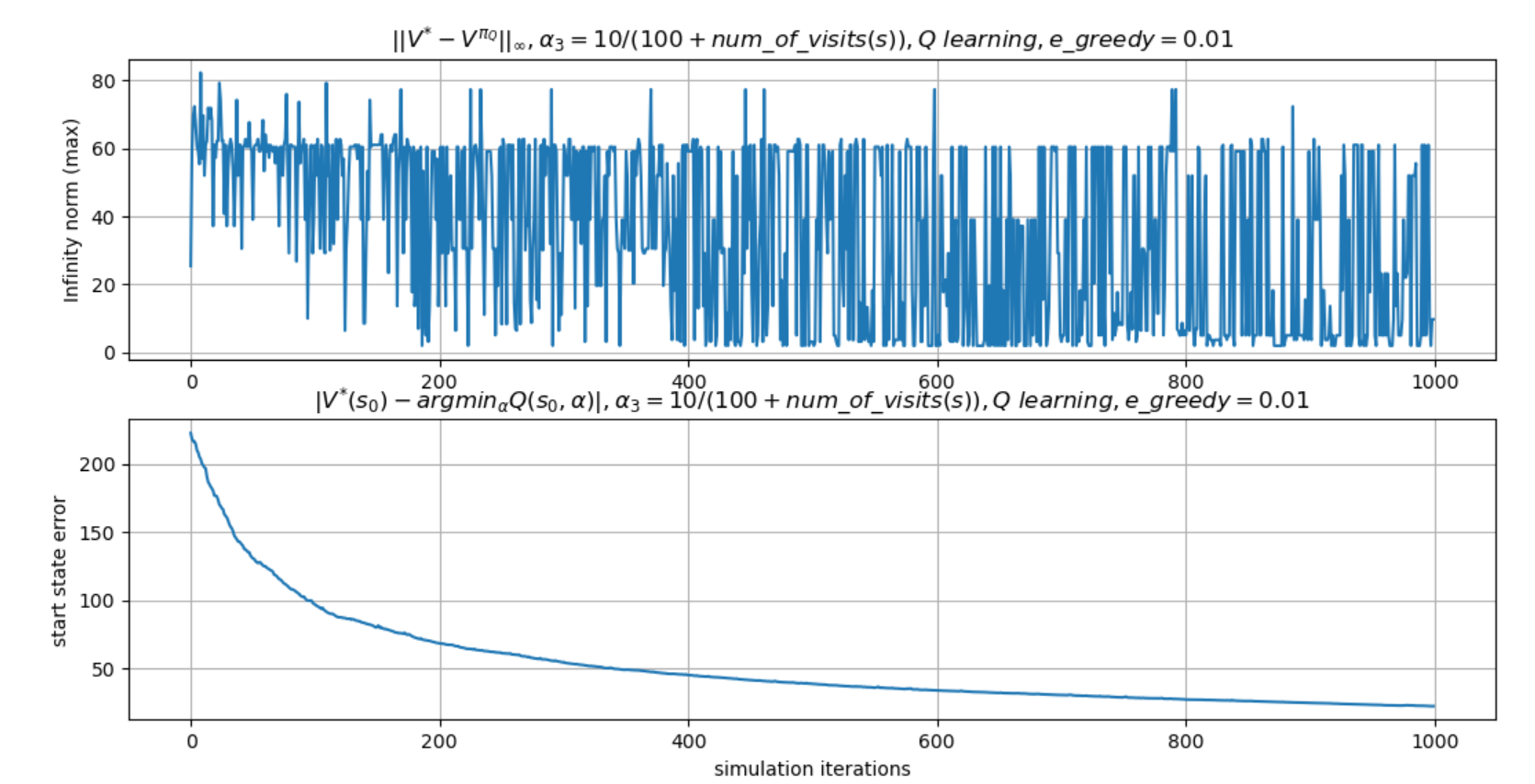






1. Results of the Q-Learning for and step size over 100,000 iterations-

Displayed in plots are 1000 sampled points (sample for every 100th iteration).



Changing the value of affects the ratio between exploration and exploitation in the learning process. For smaller the probability of exploring some of the states is smaller, however we reach similar performance for both cases. This could be explained by the relatively small action space for each state and the large number of iterations meaning we reach exploration of most states in both cases.