Defining and Solving RL Environments

Niraj Chetry

March 2021

1 Introduction

This assignment is aimed to define and solve two types of Environments in Reinforcement Learning : **Deterministic** and **Stochastic**. Q-Learning and Temporal Difference : These two processes are used, where a randomly generated Agent searches for his Goal by traveling through different **States** followed by acquiring corresponding **Rewards**. The objective is to **maximize** the **Cumulative Reward** obtained by the random agent. **OpenAI Gym** library is utilized in order to perform the task by using **Google Colaboratory** notebook.

2 Description

Deterministic Environments are the one where there's only one possible action to perform for an agent. While, **Stochastic Environments** incorporate randomness to the process by resulting in multiple possible actions. **Observation Space** is defined by the current state occupied by the agent, without looking into the long term possibilities. Agent performs a specific action **A** in a state **S** and then moves to a different state **S'**. At state **S'** it gets a reward **R**, which is inherent to the state **S'**. Number of time steps required to achieve the goal by an agent can be controlled and the upper bound of which is termed as **Max Time steps**. The process is considered to be completed when either the agent achieves the goal or it exhausts the maximum number of time steps. However, the more future actions are penalized while calculating the cumulative reward and this penalty is called as **Discount Factor**.

Q-Learning : This is a Off Policy and Model Free algorithm which determines the optimal action for an agent by iterative updation of (State, Action) value functions, irrespective of the policy being used. This method is open to both Exploration and Exploitation which gets controlled by the Epsilon Greedy mechanism; where Exploration is dominant during the start of the process and gradually Exploitation becomes dominant once the agent starts learning the optimal policy. The mechanism of Q-Learning is as follows :

Q-learning (off-policy TD control) for estimating $\pi \approx \pi_*$

Algorithm parameters: step size $\alpha \in (0, 1]$, small $\varepsilon > 0$
Initialize $Q(s, a)$, for all $s \in \mathcal{S}^+, a \in \mathcal{A}(s)$, arbitrarily except that $Q(terminal, \cdot) = 0$

Loop for each episode:
  Initialize S
  Loop for each step of episode:
    Choose A from S using policy derived from Q (e.g., ε-greedy)
    Take action A, observe R, S'
    $Q(S, A) \leftarrow Q(S, A) + \alpha\big[R + \gamma \max_a Q(S', a) - Q(S, A)\big]$
    $S \leftarrow S'$
  until S is terminal

Target
Prediction
Immediate Reward
loss

In this problem, the maximum value of **Epsilon** used is **0.9** which can decrease up to **0.1** geometrically at each successive episode.

Temporal Difference : TD methods are On Policy and Model Free in nature, which can learn directly from raw experience without a model of the environment's dynamics. TD methods update estimates based in part on other learned estimates, without waiting for a final outcome. It iteratively updates the State Value functions at each steps. The simplest TD method is where the look ahead step is 1, which is known as TD(0). The mechanism of TD(0) is as follows :

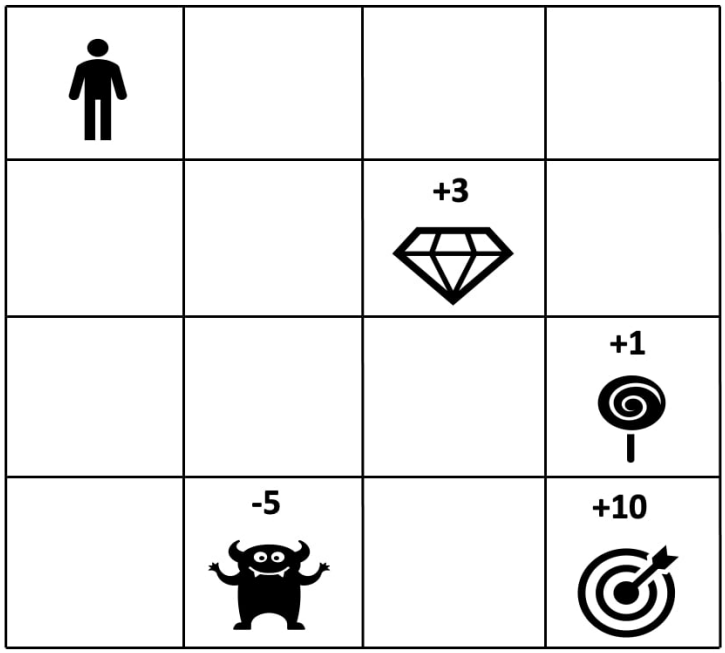
Tabular TD(0) for estimating v_π

Input: the policy π to be evaluated
Algorithm parameter: step size $\alpha \in (0, 1]$
Initialize $V(s)$, for all $s \in \mathcal{S}^+$, arbitrarily except that $V(terminal) = 0$

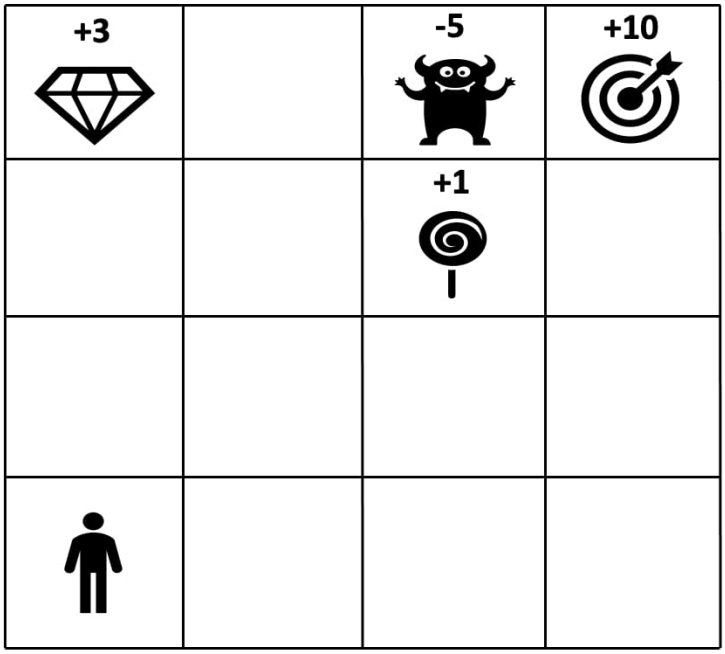
Loop for each episode:
  Initialize S
  Loop for each step of episode:
    $A \leftarrow$ action given by π for S
    Take action A, observe R, S'
    $V(S) \leftarrow V(S) + \alpha\big[R + \gamma V(S') - V(S)\big]$
    $S \leftarrow S'$
  until S is terminal

For both Q-Learning and TD(0), the **Learning Rate** is taken as **0.1** and Total **1000 Episodes** are taken.

3 Problem Statement



Defined above is the 4*4 grid for the **Deterministic Environment** used in this problem. The agent is at state **[0, 0]** and the goal is situated at **[3, 3]**. **Diamond**, **Monster** and **Candy** are placed at intermediate states with the corresponding rewards as shown in figure. Rewards for all other states are 0. Action space for the agent is defined as A : **A = Up, Down, Right, Left**. If the agent is at **[0, 0]** and chooses an action of going **Right**, then the transition probability of arriving at **[0, 1]** is **1** and thus it is an Deterministic Environment.



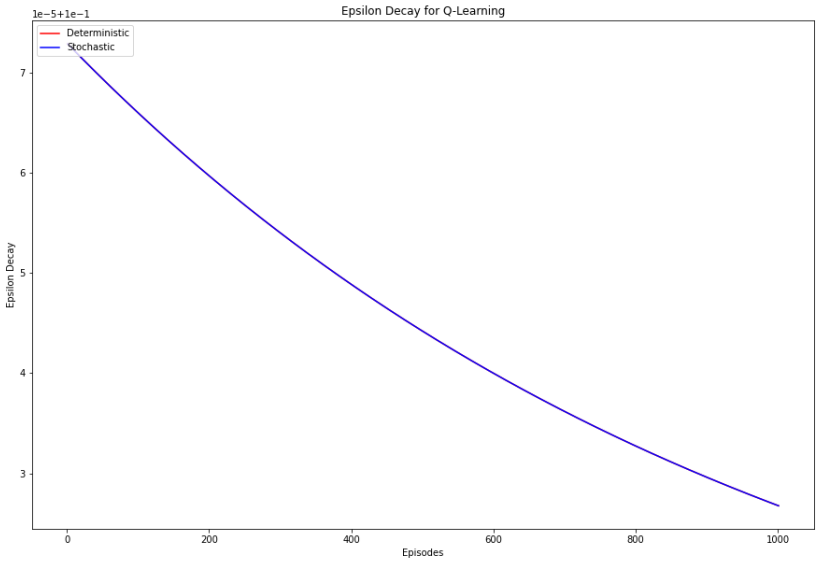
Defined above is the enhanced version of previous environment with agent position at **[3, 0]** and goal position at **[0, 3]**; which is used as the **Stochastic Environment** in this problem. The difference between **Deterministic** and **Stochastic** Environments is due to the randomness present in Stochastic environment, which is not present in the Deterministic environment. Randomness is incorporated such that if the agent situated at **[0, 0]** chooses the action of going **Right**, then it will reach **[0, 1]** with a transition probability of **0.67**; while it will remain in the same position with a probability of **0.33** and this holds true for each of the action it takes - hence it is an Stochastic Environment.

Reward set, R : 0, -5, +1, +3, +10 and State set, S : S1, S2, S3, ..., S16.

Q-Learning and Temporal Difference - TD(0) are tuned with respect to **Maximum Time steps** and **Discount Factor** by taking 3 values of each. Values for Max Time steps are are chosen as **100, 120, 150**; while the values for Discount Factor are **0.1, 0.5, 0.9**. For each of the **9** pair of values, cumulative reward is compared to determine the optimal value of parameters.

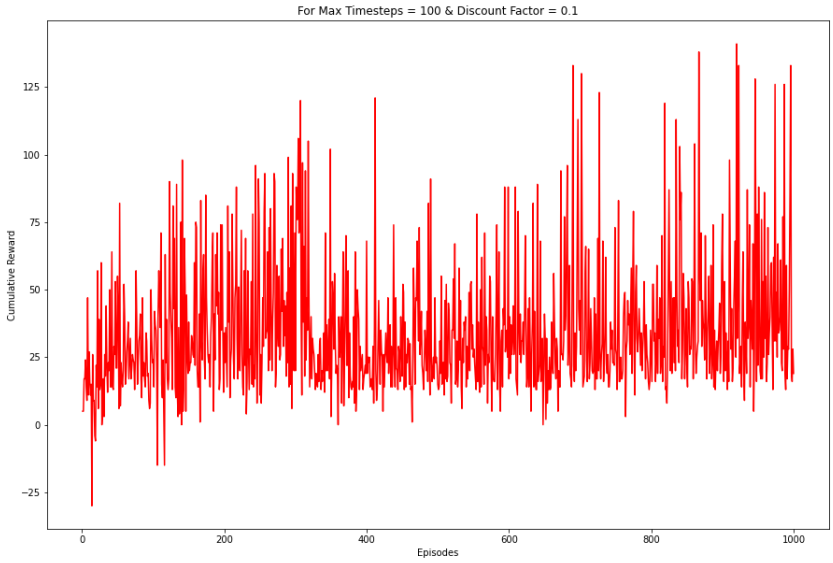
4 Results

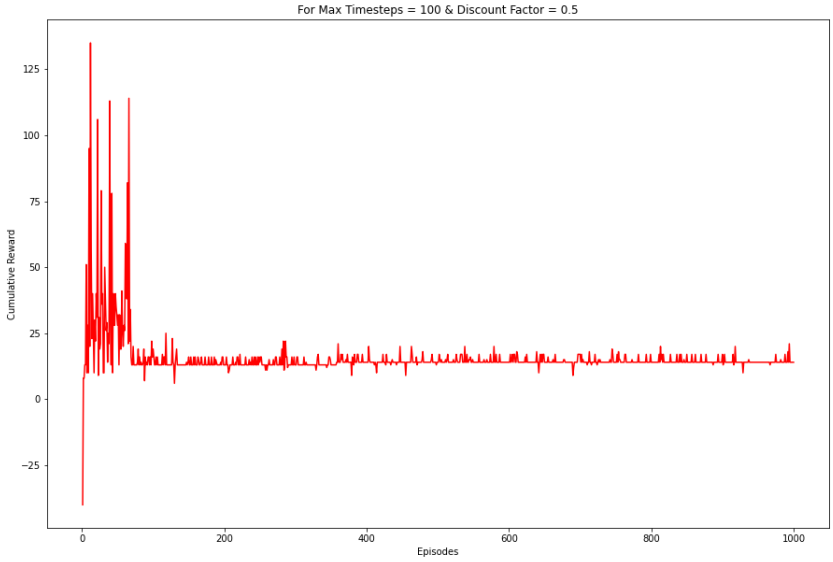
Let's look at the plots obtained from Q-Learning first :

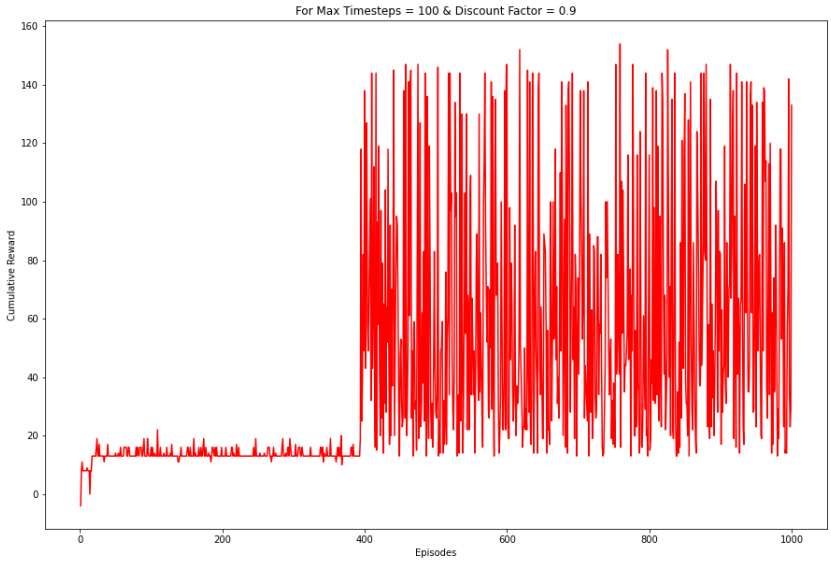


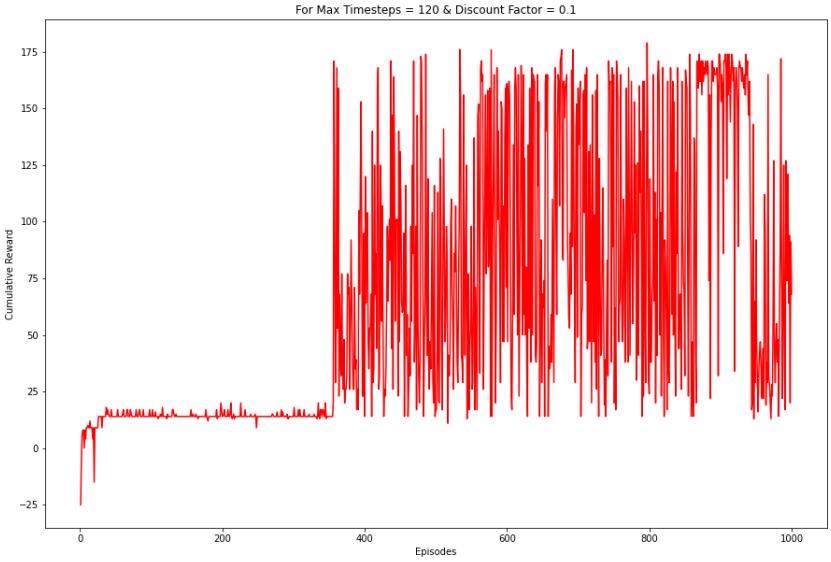
The value of Epsilon is decreasing from the maximum value 0.9 to lower value in the following episodes; i.e. the policy is growing more Greedy.

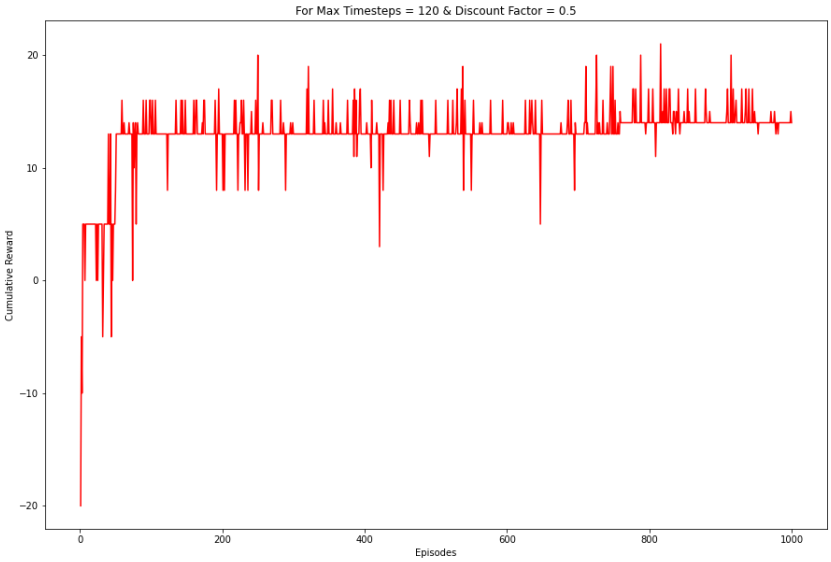
Cumulative Reward Plots :

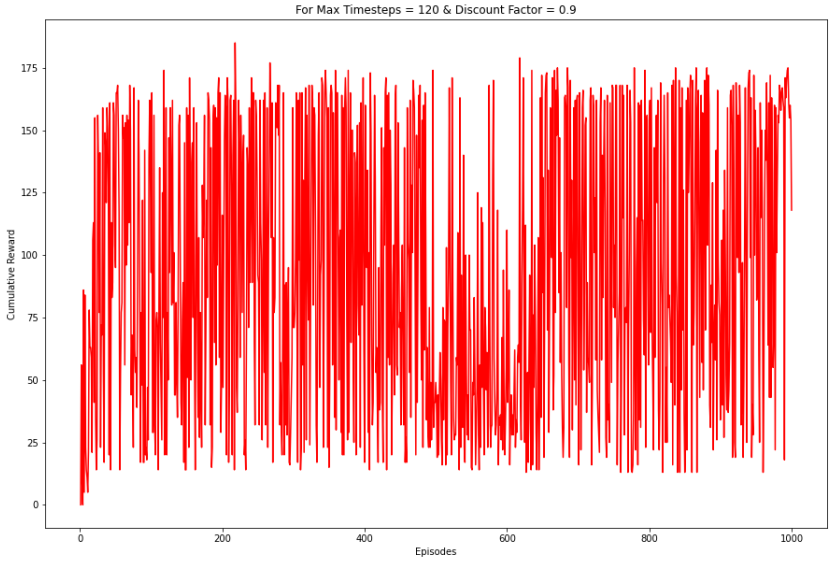

For Max Timesteps = 100 & Discount Factor = 0.1
Cumulative Reward
Episodes
125
100
75
50
25
0
−25
0
200
400
600
800
1000

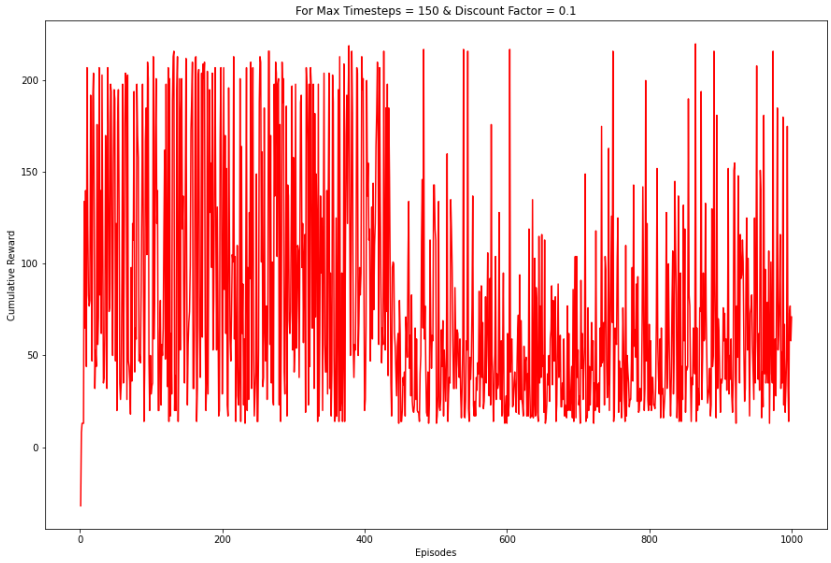

For Max Timesteps = 100 & Discount Factor = 0.5
Cumulative Reward
Episodes
125
100
75
50
25
0
−25
0
200
400
600
800
1000


For Max Timesteps = 100 & Discount Factor = 0.9
Cumulative Reward
Episodes
160
140
120
100
80
60
40
20
0
0
200
400
600
800
1000

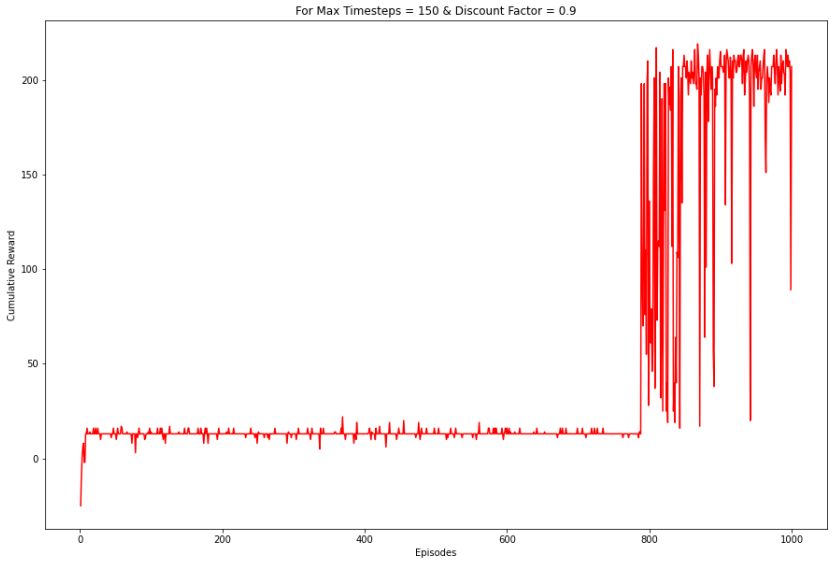

For Max Timesteps = 120 & Discount Factor = 0.1
Cumulative Reward
Episodes
175
150
125
100
75
50
25
0
−25
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Cumulative Reward
Episodes
20
10
0
−10
−20
0
200
400
600
800
1000

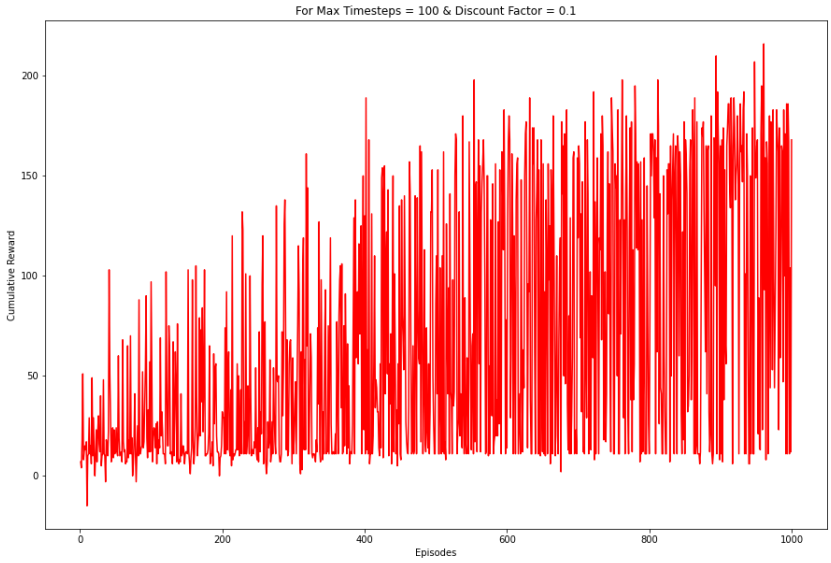

For Max Timesteps = 120 & Discount Factor = 0.9
Cumulative Reward
Episodes
175
150
125
100
75
50
25
0
0
200
400
600
800
1000

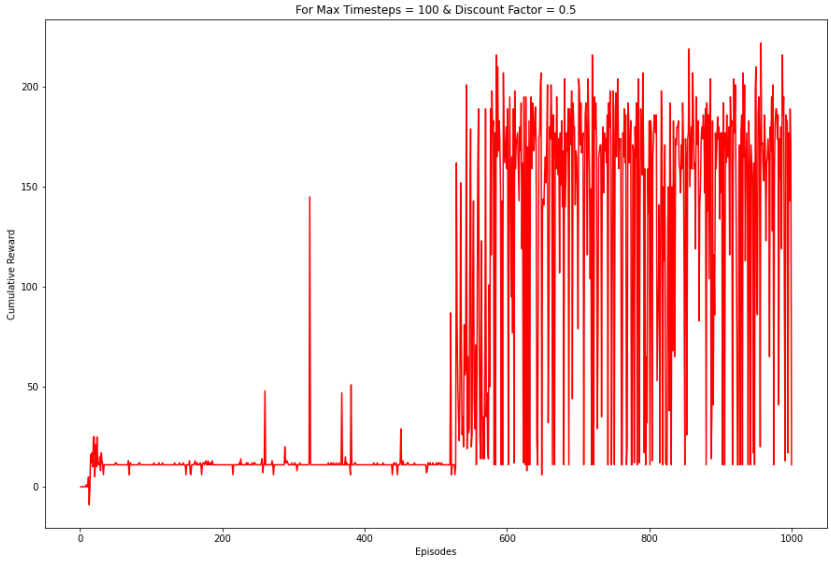

For Max Timesteps = 150 & Discount Factor = 0.1
Cumulative Reward
Episodes
200
150
100
50
0
0
200
400
600
800
1000

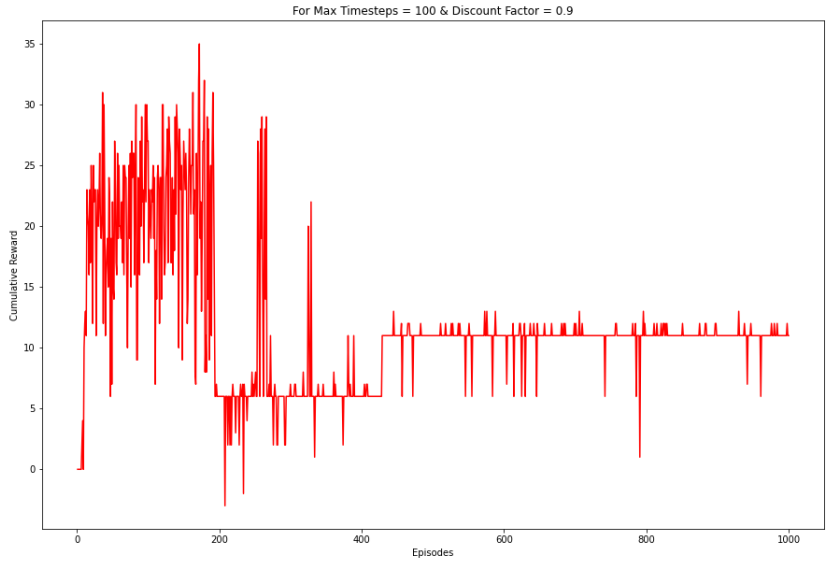

For Max Timesteps = 150 & Discount Factor = 0.5
Cumulative Reward
Episodes
25
20
15
10
5
0
−5
−10
0
200
400
600
800
1000

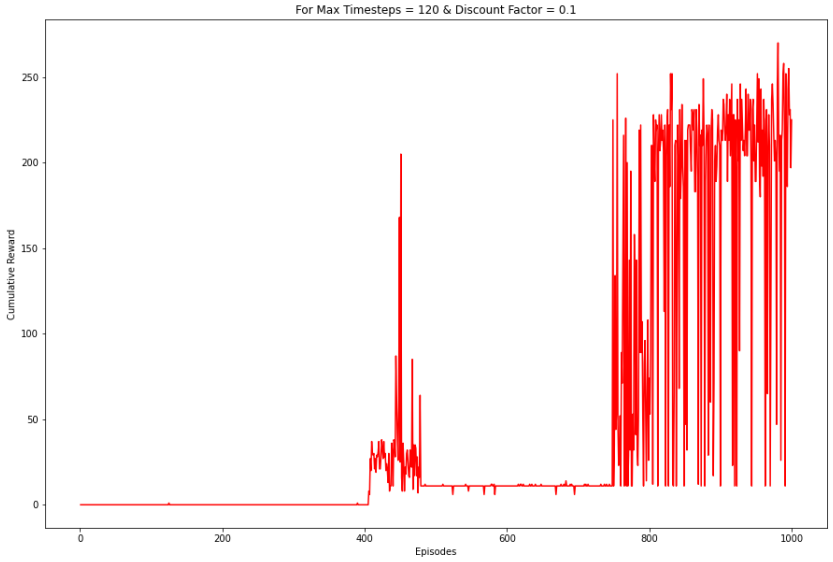


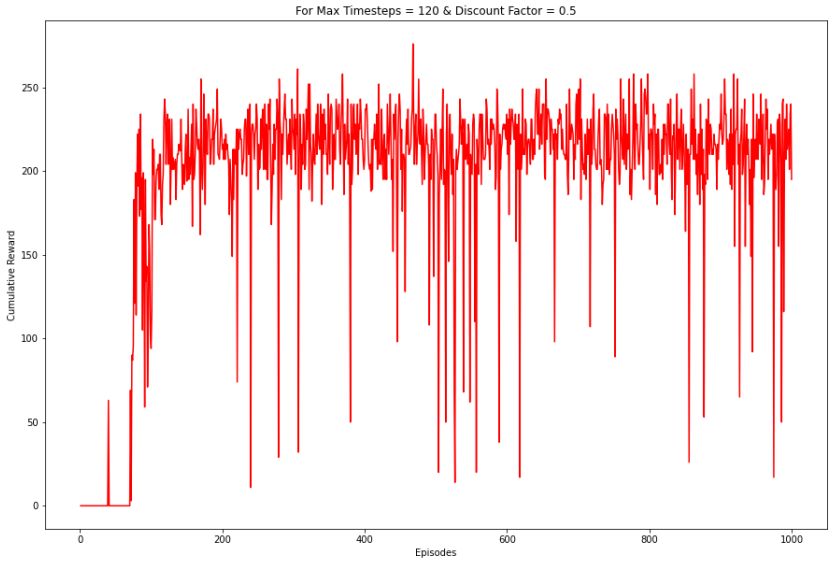
Plots above are the **Cumulative Rewards** for each of the Episodes for the **Deterministic Environment** across different pair of (**Max Time steps, Discount Factor**). Most of the plots have a increasing trend; however, equal number of other plots have fluctuating behaviour. It is observed that Cumulative Reward is maximum (**R = 200**) in case of **Maximum Time steps = 150** and **Discount Factor = 0.9**

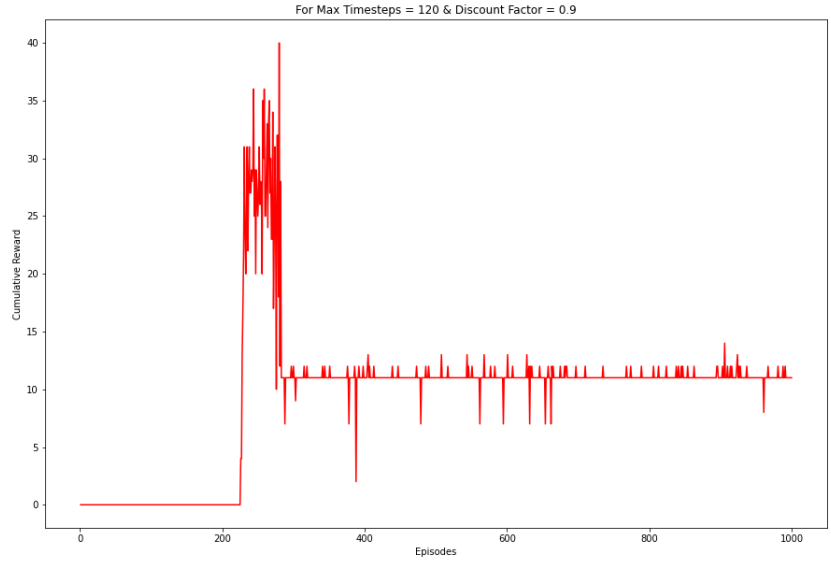

For Max Timesteps = 100 & Discount Factor = 0.1
Cumulative Reward
Episodes
0
50
100
150
200
0
200
400
600
800
1000

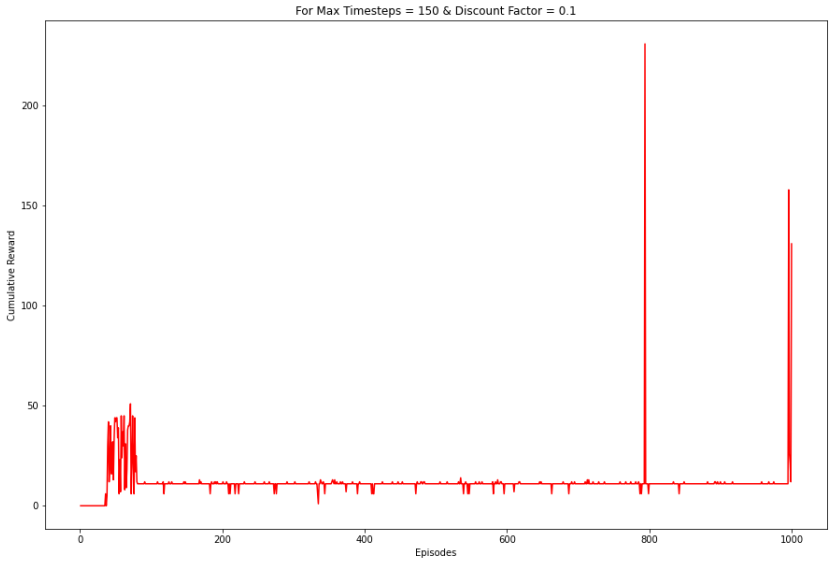

For Max Timesteps = 100 & Discount Factor = 0.5
Cumulative Reward
Episodes
0
50
100
150
200
0
200
400
600
800
1000

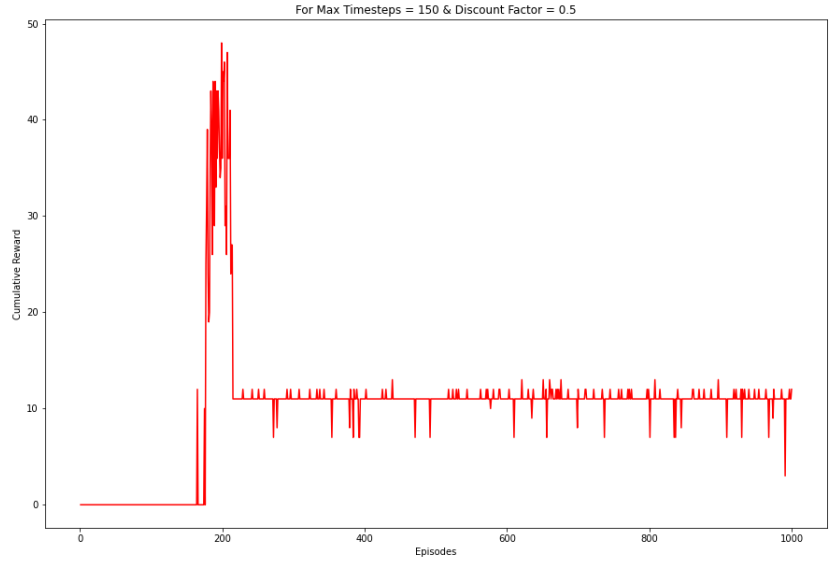

For Max Timesteps = 100 & Discount Factor = 0.9
Cumulative Reward
Episodes
35
30
25
20
15
10
5
0
0
200
400
600
800
1000

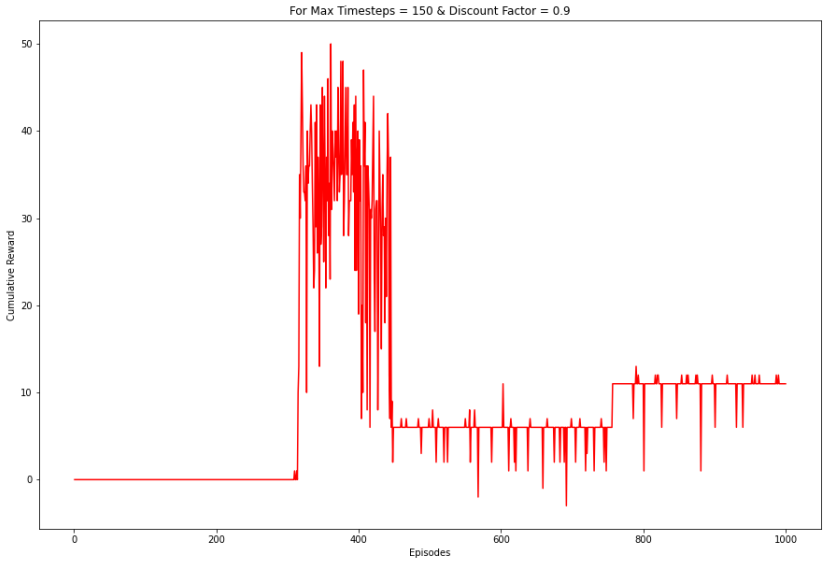

For Max Timesteps = 120 & Discount Factor = 0.1
Cumulative Reward
Episodes
250
200
150
100
50
0
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Cumulative Reward
Episodes
0
50
100
150
200
250
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.9
Cumulative Reward
Episodes
0
5
10
15
20
25
30
35
40
0
200
400
600
800
1000

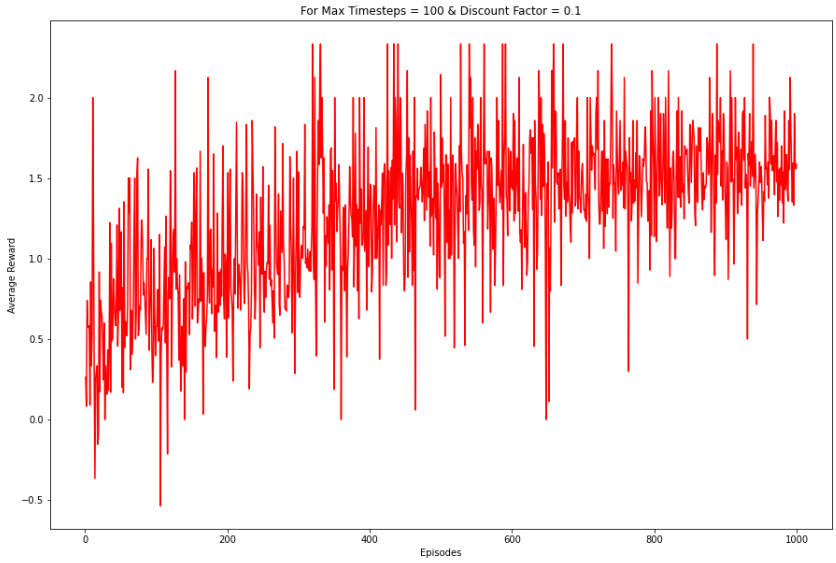

For Max Timesteps = 150 & Discount Factor = 0.1
Cumulative Reward
Episodes
0
50
100
150
200
0
200
400
600
800
1000

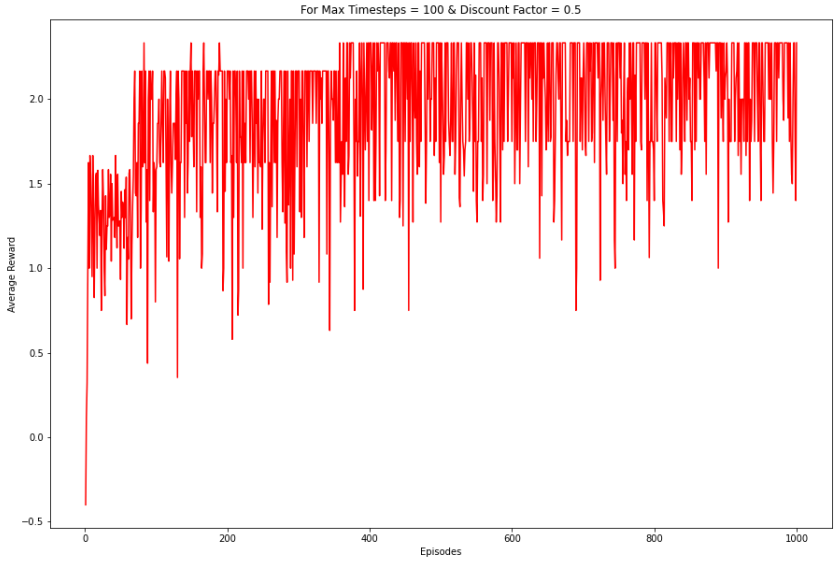

For Max Timesteps = 150 & Discount Factor = 0.5
Cumulative Reward
Episodes
0
10
20
30
40
50
0
200
400
600
800
1000

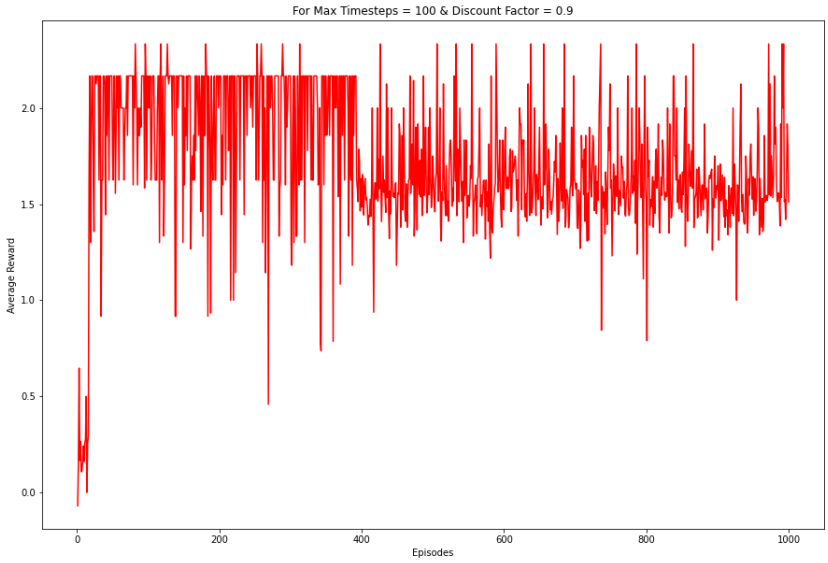


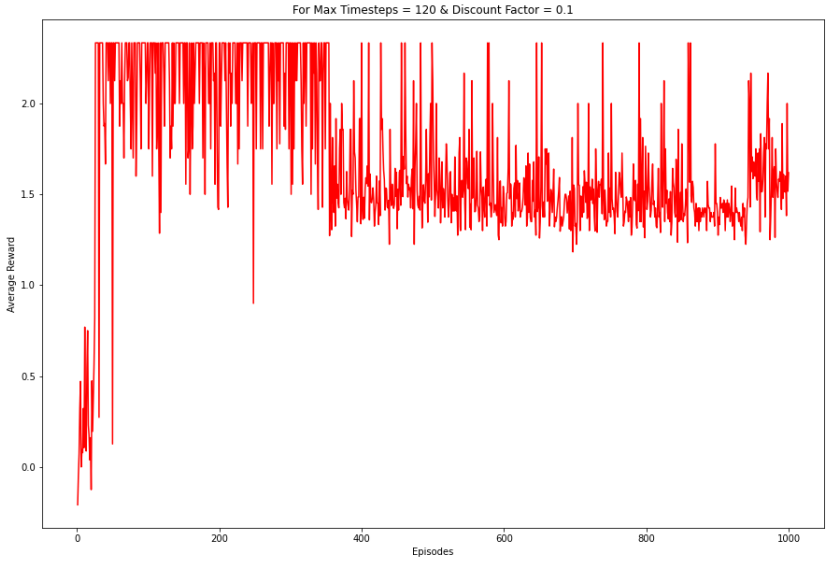
Plots above are the **Cumulative Rewards** for each of the Episodes for the **Stochastic Environment** across different pair of (**Max Time steps, Discount Factor**). Most of the plots have a increasing trend; however, equal number of other plots have fluctuating behaviour. It is observed that Cumulative Reward is maximum (**R = 250**) in case of **Maximum Time steps = 120** and **Discount Factor = 0.5**

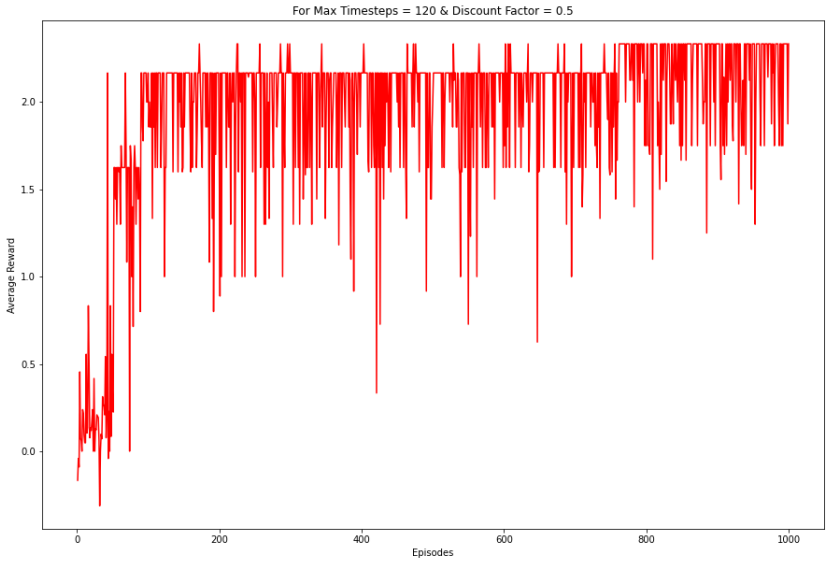
Average Reward Plots :

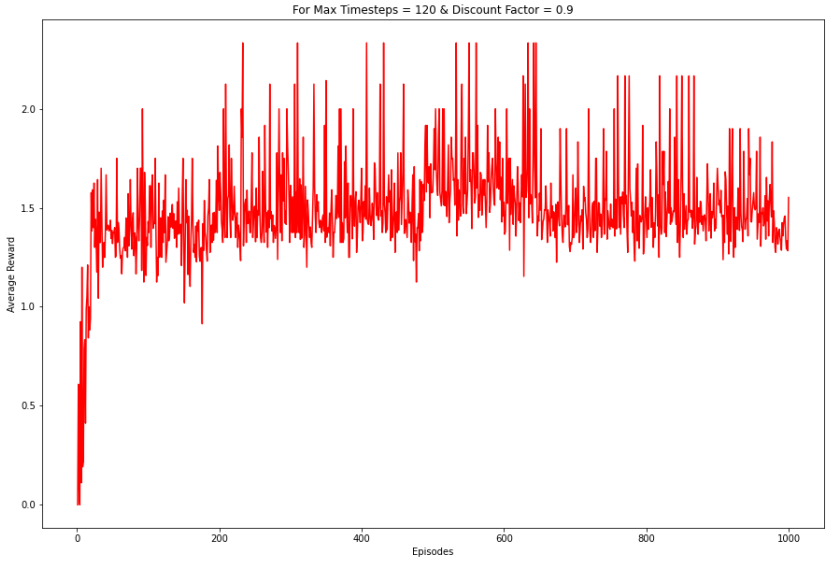

For Max Timesteps = 100 & Discount Factor = 0.1
Average Reward
Episodes
2.0
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000

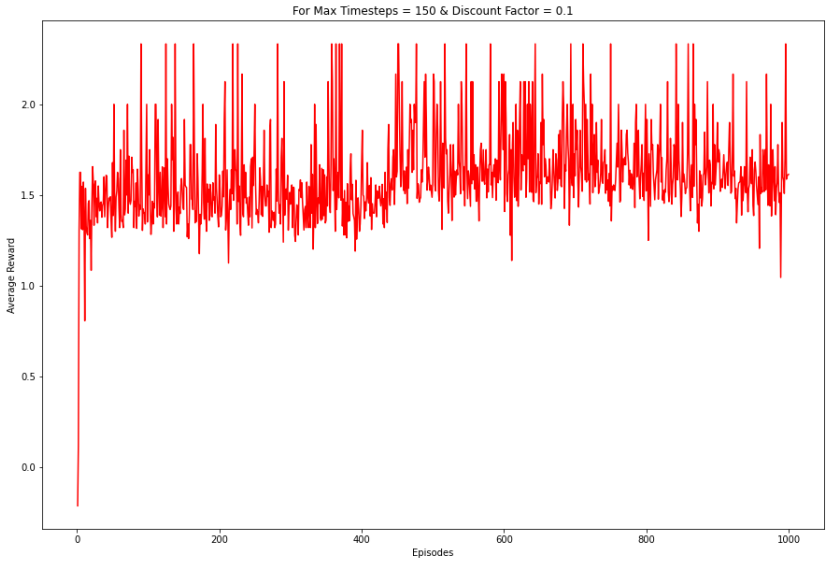

For Max Timesteps = 100 & Discount Factor = 0.5
Average Reward
Episodes
2.0
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000

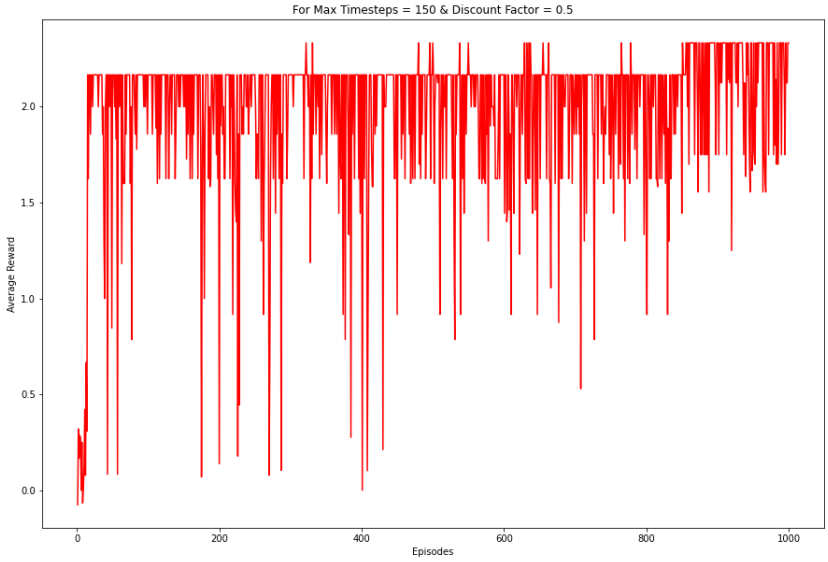

For Max Timesteps = 100 & Discount Factor = 0.9
Average Reward
Episodes
0.0
0.5
1.0
1.5
2.0
0
200
400
600
800
1000

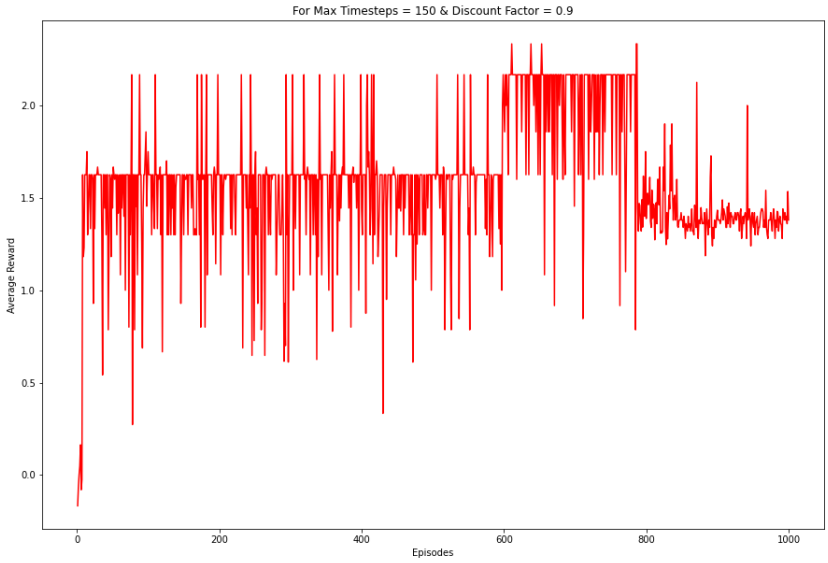

For Max Timesteps = 120 & Discount Factor = 0.1
Average Reward
Episodes
0.0
0.5
1.0
1.5
2.0
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Average Reward
Episodes
0.0
0.5
1.0
1.5
2.0
0
200
400
600
800
1000

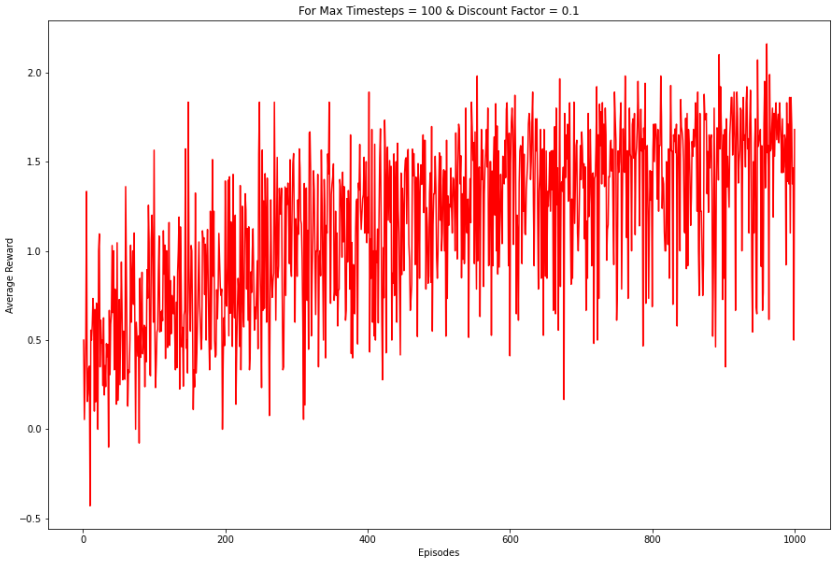

For Max Timesteps = 120 & Discount Factor = 0.9
Average Reward
Episodes
0.0
0.5
1.0
1.5
2.0
0
200
400
600
800
1000

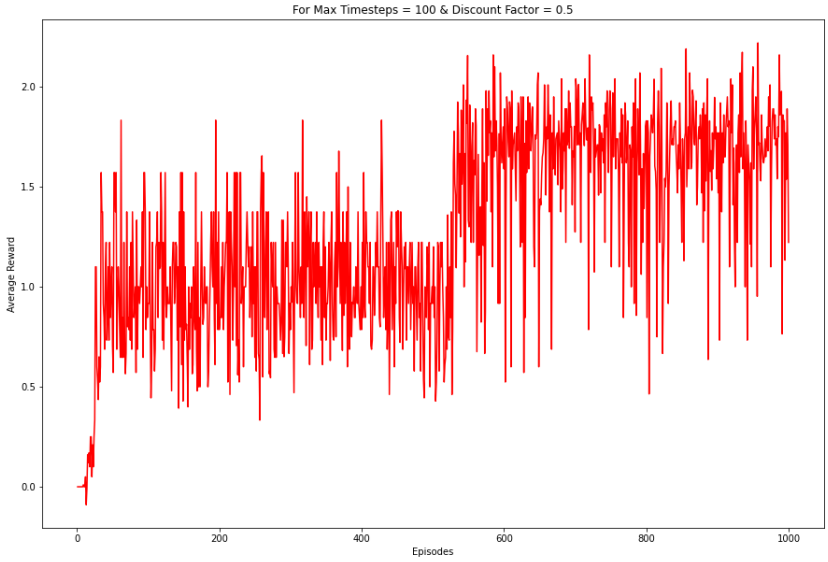

For Max Timesteps = 150 & Discount Factor = 0.1
Average Reward
Episodes
0.0
0.5
1.0
1.5
2.0
0
200
400
600
800
1000

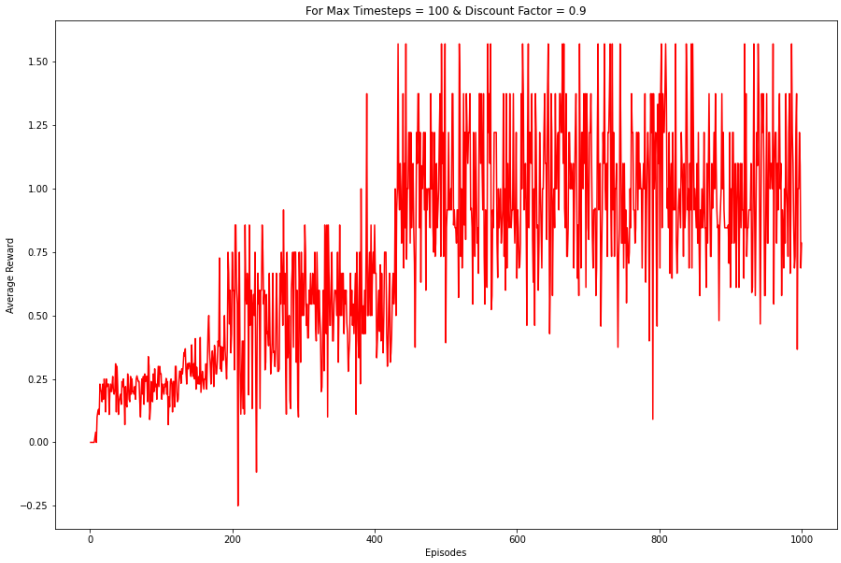

For Max Timesteps = 150 & Discount Factor = 0.5
Average Reward
Episodes
0.0
0.5
1.0
1.5
2.0
0
200
400
600
800
1000

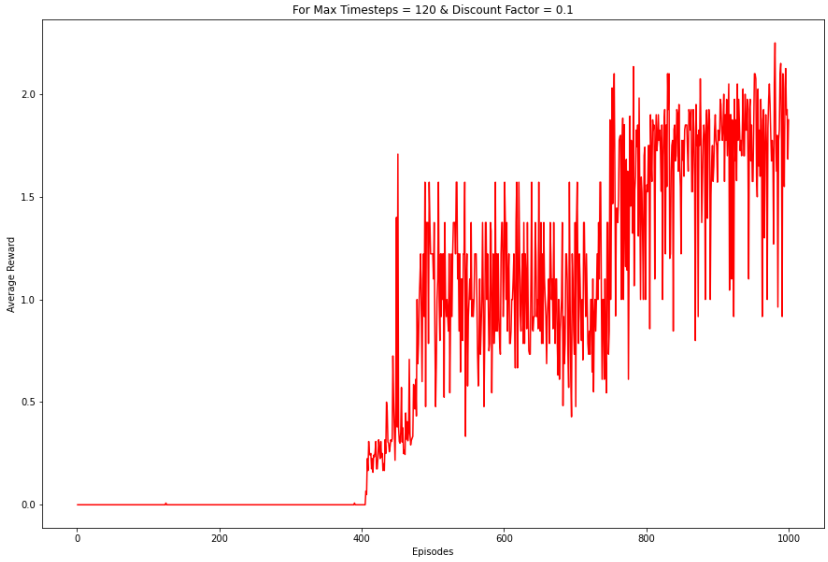


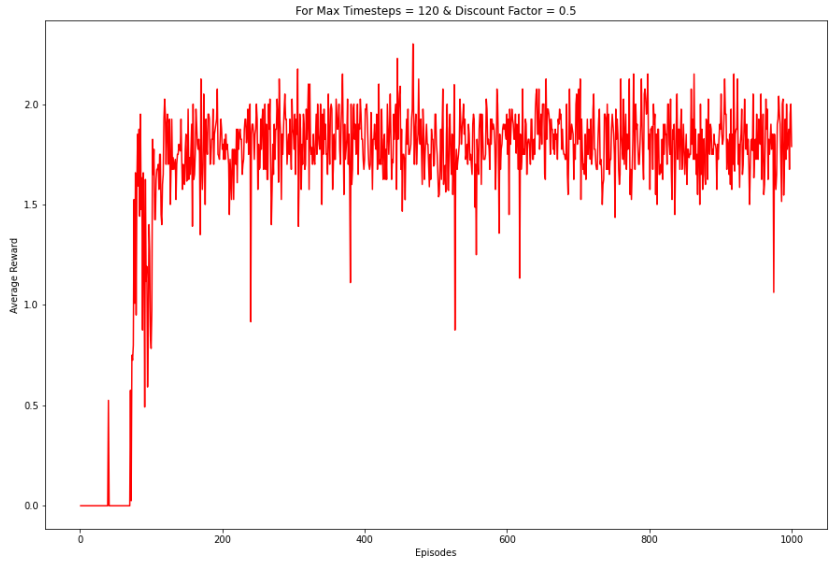
Plots above are the **Average Rewards** for each of the Episodes for the **Deterministic Environment** across different pair of (**Max Time steps, Discount Factor**). All of the plots are fluctuating across the value **2**

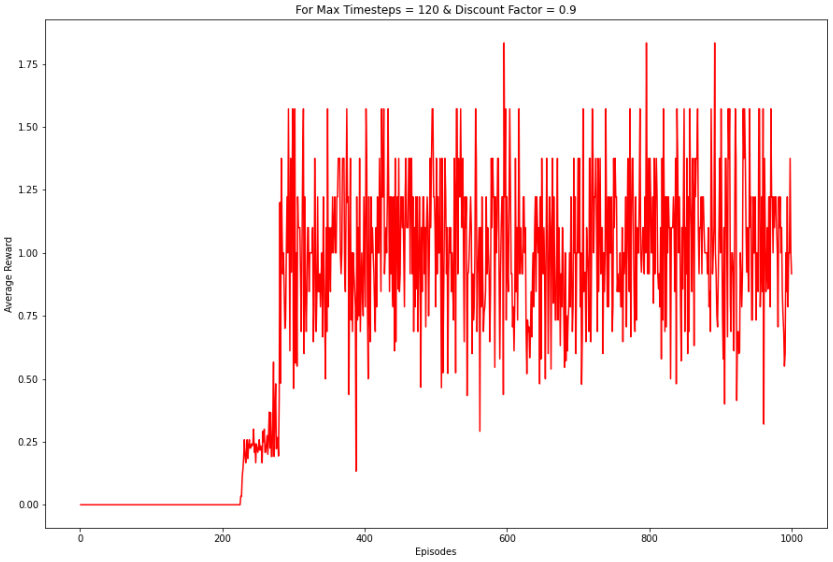

For Max Timesteps = 100 & Discount Factor = 0.1
Average Reward
Episodes
2.0
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000

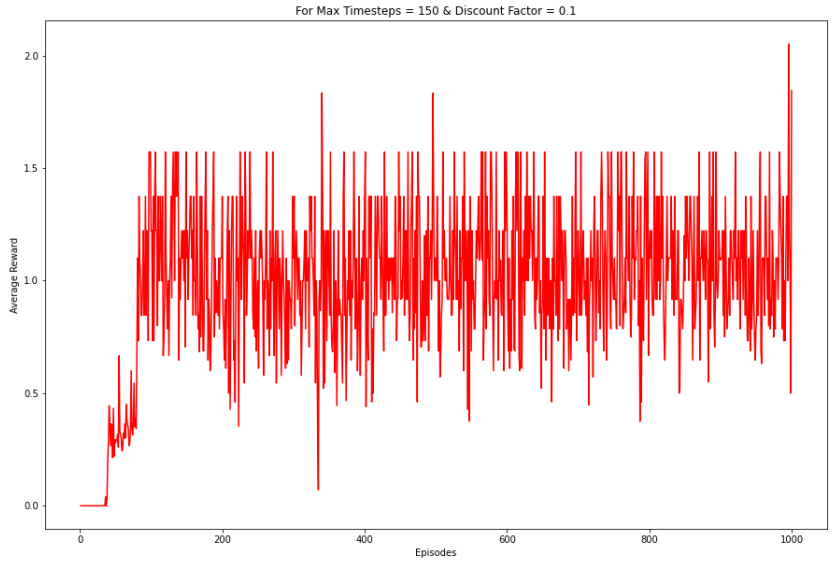

For Max Timesteps = 100 & Discount Factor = 0.5
Average Reward
Episodes
2.0
1.5
1.0
0.5
0.0
0
200
400
600
800
1000

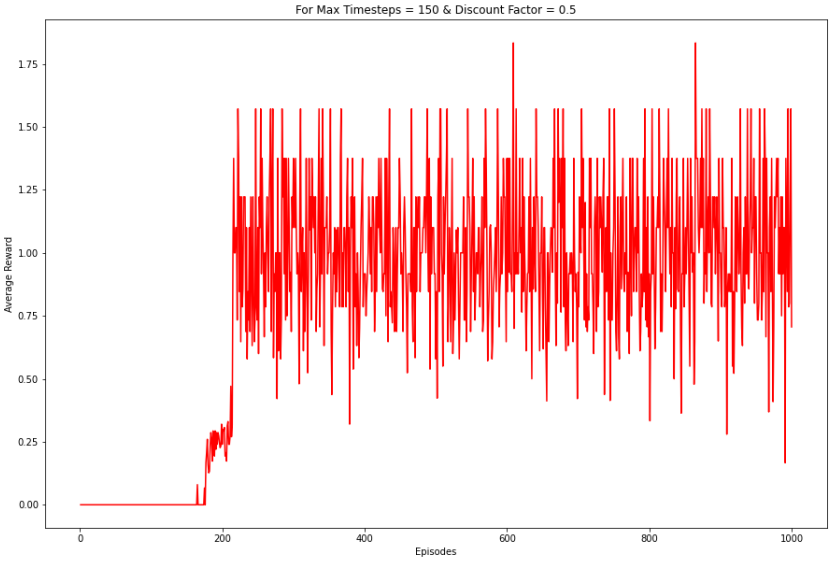

For Max Timesteps = 100 & Discount Factor = 0.9
Average Reward
Episodes
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
0
200
400
600
800
1000

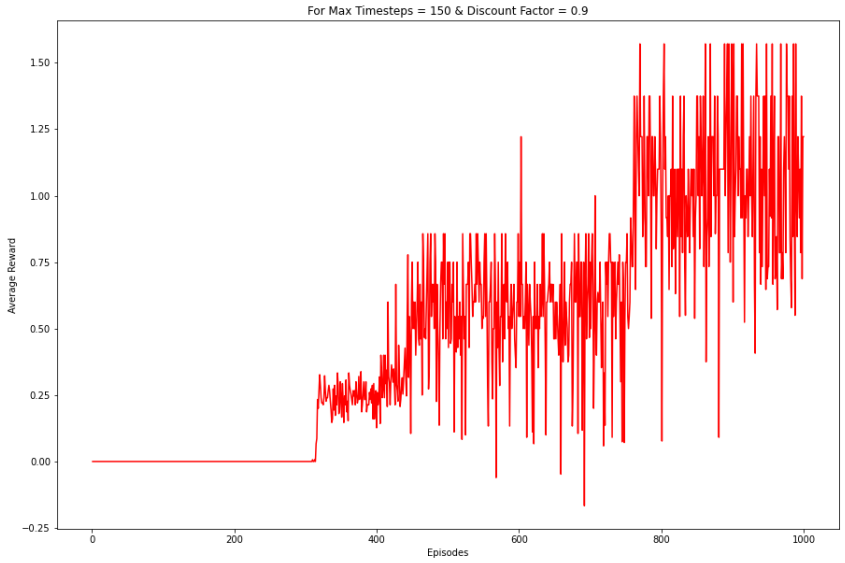

For Max Timesteps = 120 & Discount Factor = 0.1
Average Reward
Episodes
2.0
1.5
1.0
0.5
0.0
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Average Reward
Episodes
0.0
0.5
1.0
1.5
2.0
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.9
Average Reward
Episodes
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0
200
400
600
800
1000

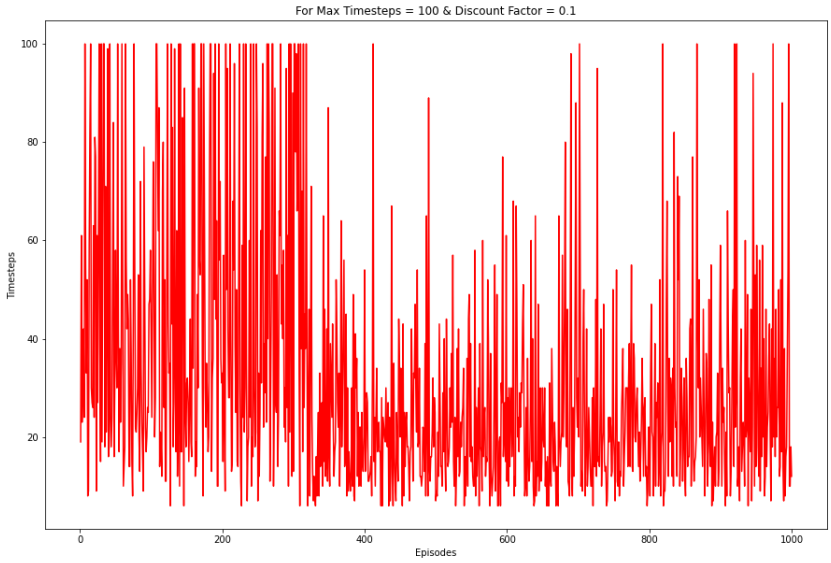

For Max Timesteps = 150 & Discount Factor = 0.1
Average Reward
Episodes
0.0
0.5
1.0
1.5
2.0
0
200
400
600
800
1000

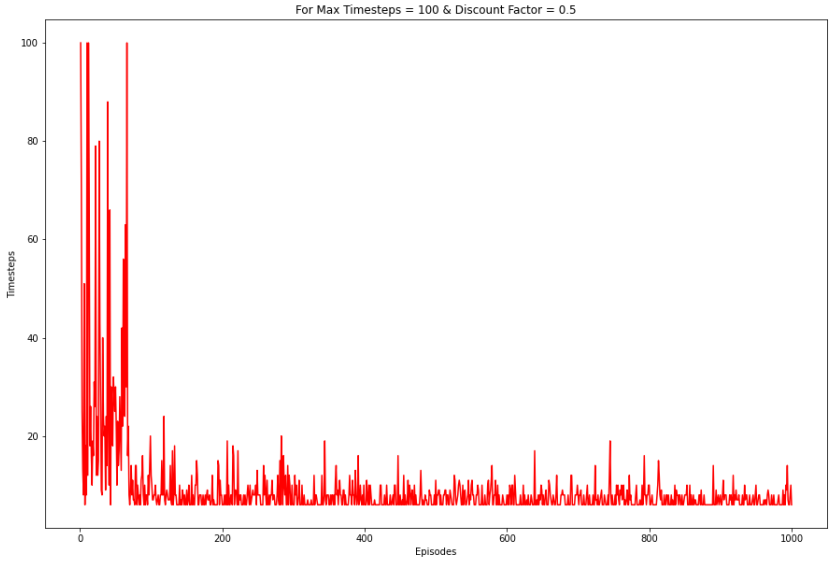

For Max Timesteps = 150 & Discount Factor = 0.5
Average Reward
Episodes
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
0
200
400
600
800
1000

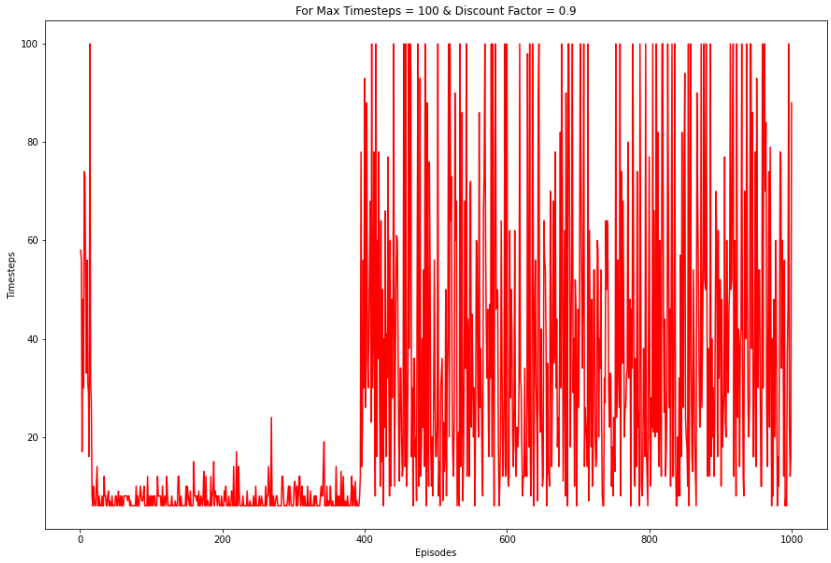


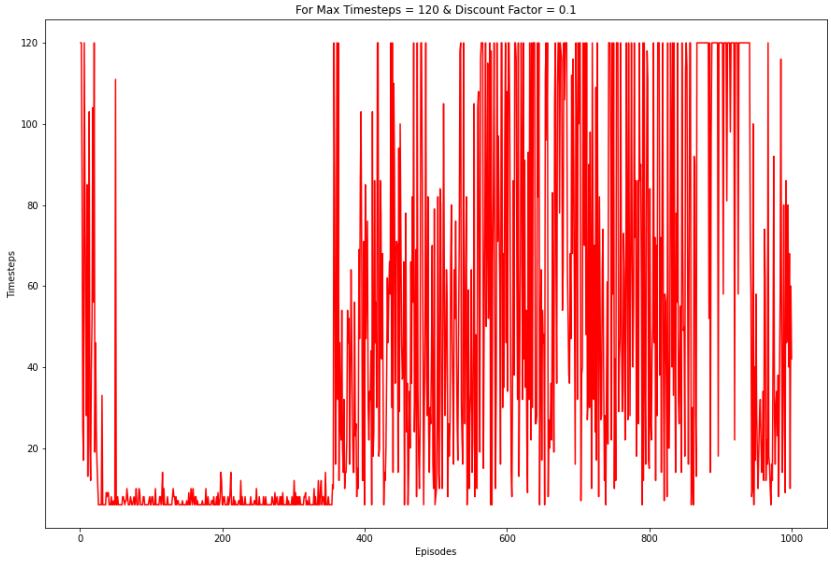
Plots above are the **Average Rewards** for each of the Episodes for the **Stochastic Environment** across different pair of (**Max Time steps, Discount Factor**). Some of the plots are converging towards the value **2**, while some others are fluctuating across **1.5**. The best value is obtained for **Maximum Time steps = 120** and **Discount Factor = 0.5**.

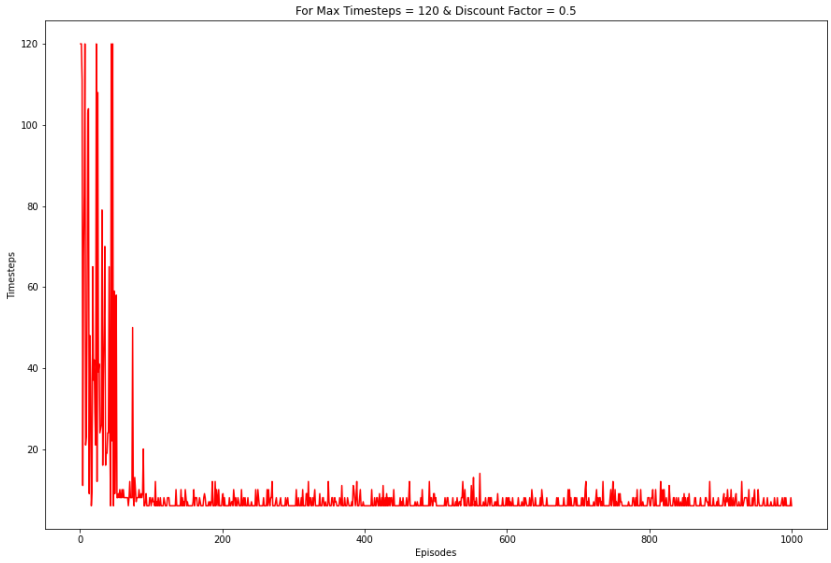
Timesteps per Episode Plots :

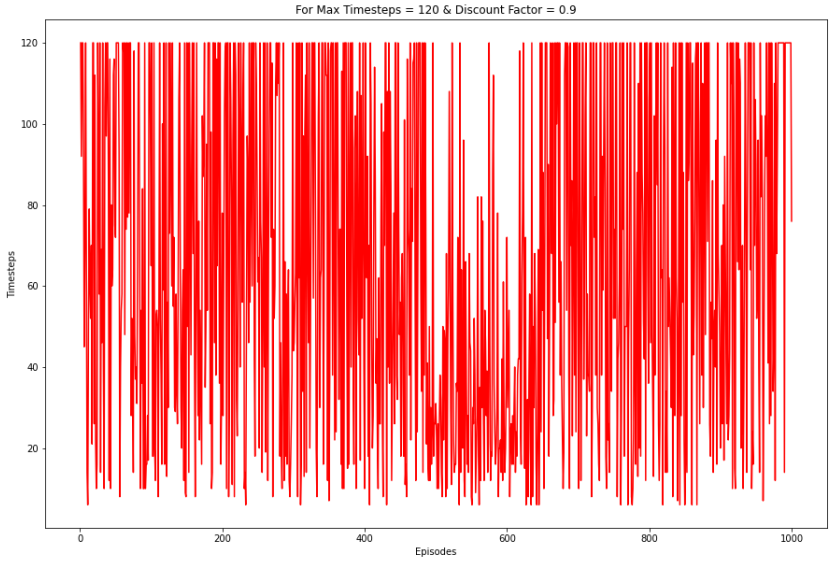

For Max Timesteps = 100 & Discount Factor = 0.1
Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

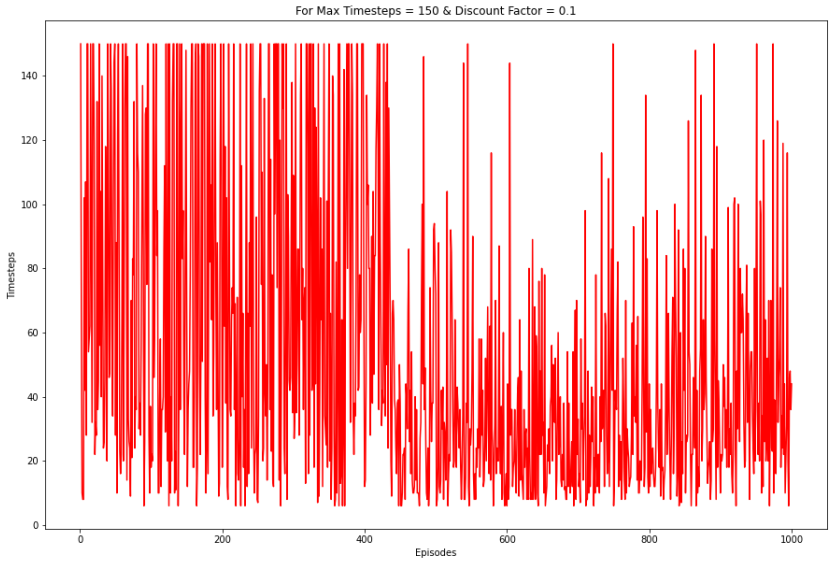

For Max Timesteps = 100 & Discount Factor = 0.5
Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

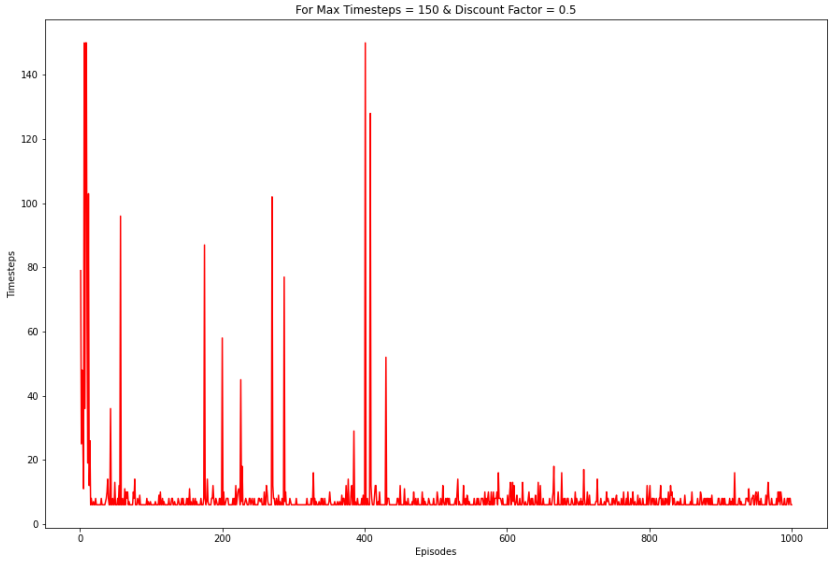

For Max Timesteps = 100 & Discount Factor = 0.9
Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

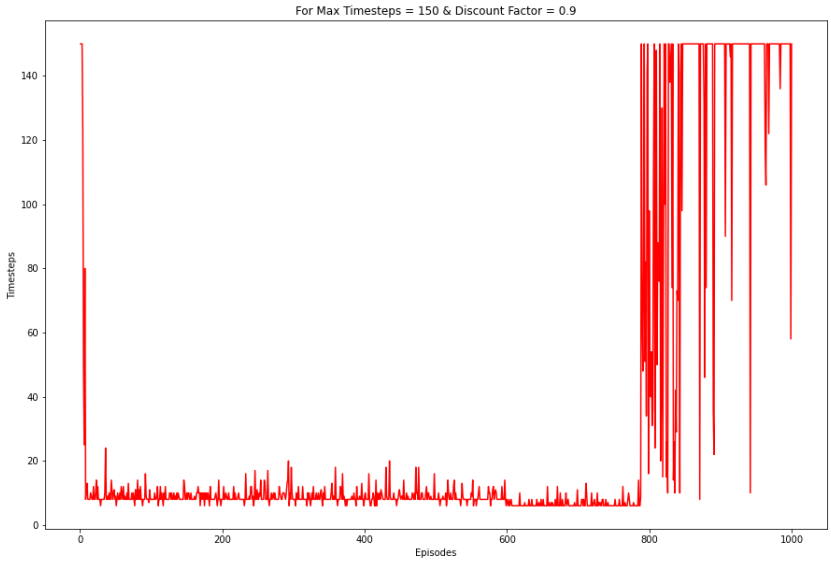

For Max Timesteps = 120 & Discount Factor = 0.1
Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000

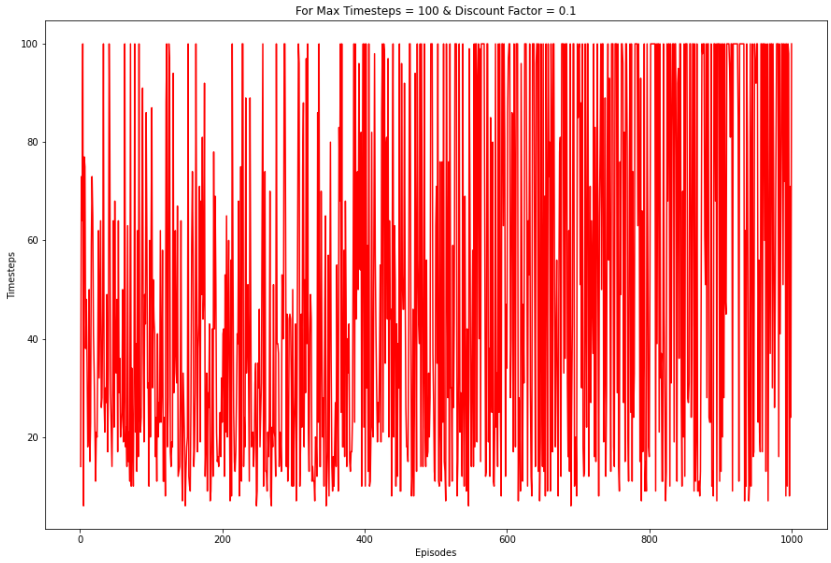

For Max Timesteps = 120 & Discount Factor = 0.9
Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000

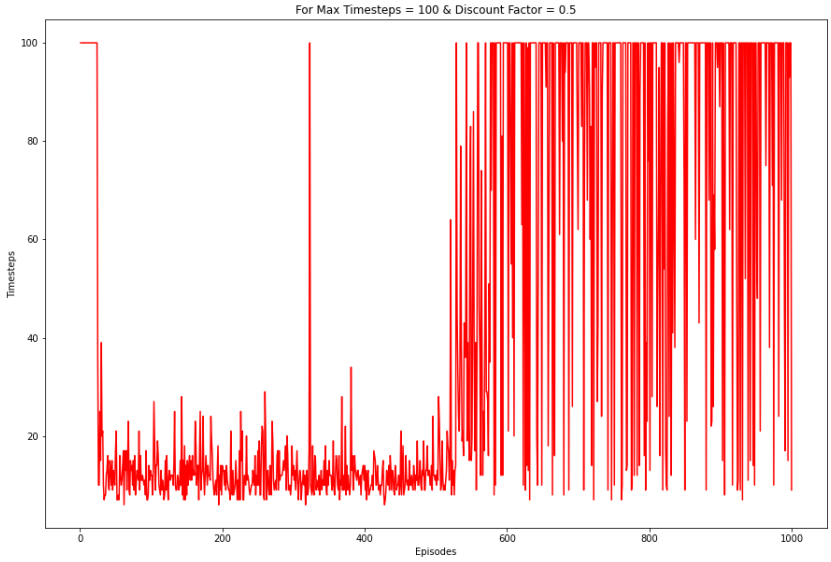

For Max Timesteps = 150 & Discount Factor = 0.1
Timesteps
Episodes
0
20
40
60
80
100
120
140
0
200
400
600
800
1000

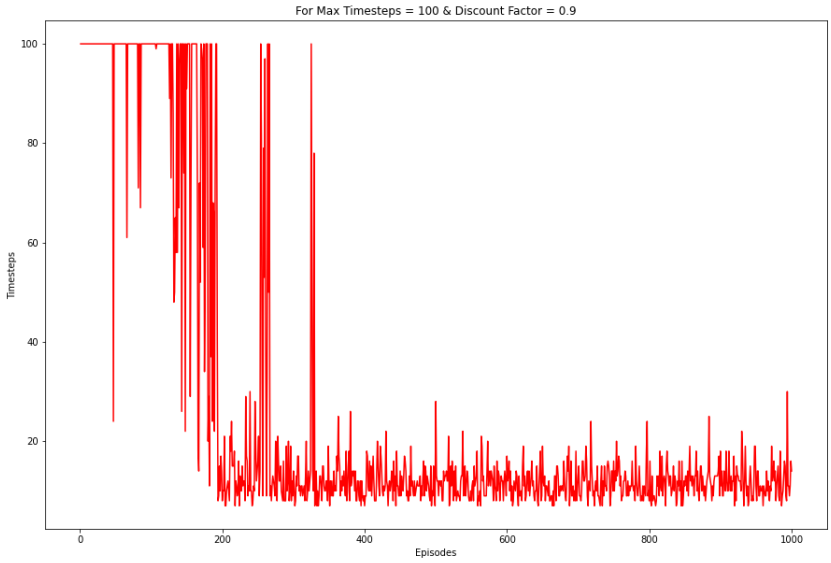

For Max Timesteps = 150 & Discount Factor = 0.5
Timesteps
Episodes
0
20
40
60
80
100
120
140
0
200
400
600
800
1000

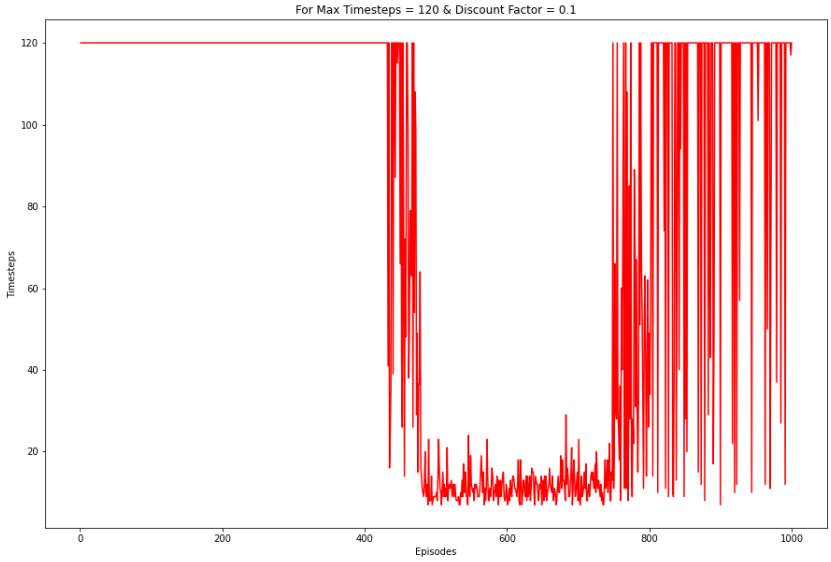


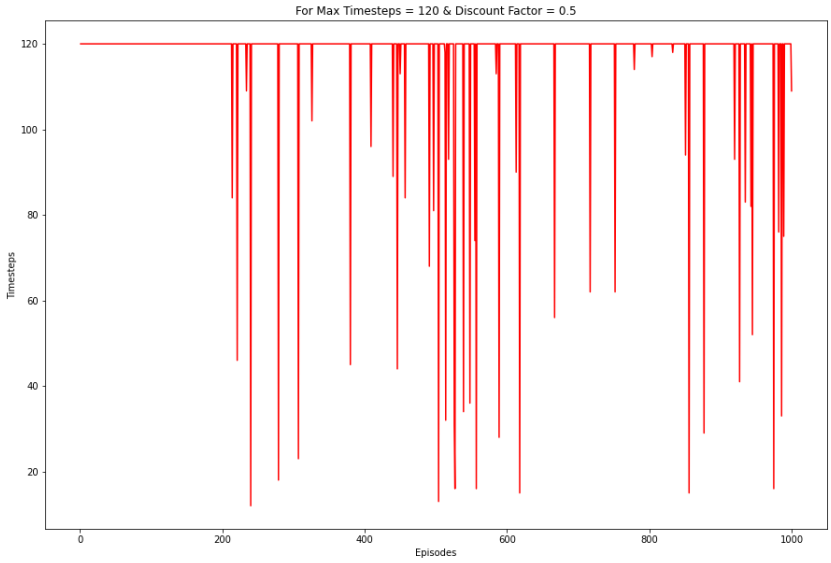
Plots above are the **Timesteps Used** for each of the Episodes for the **Deterministic Environment** across different pair of (**Max Time steps, Discount Factor**).

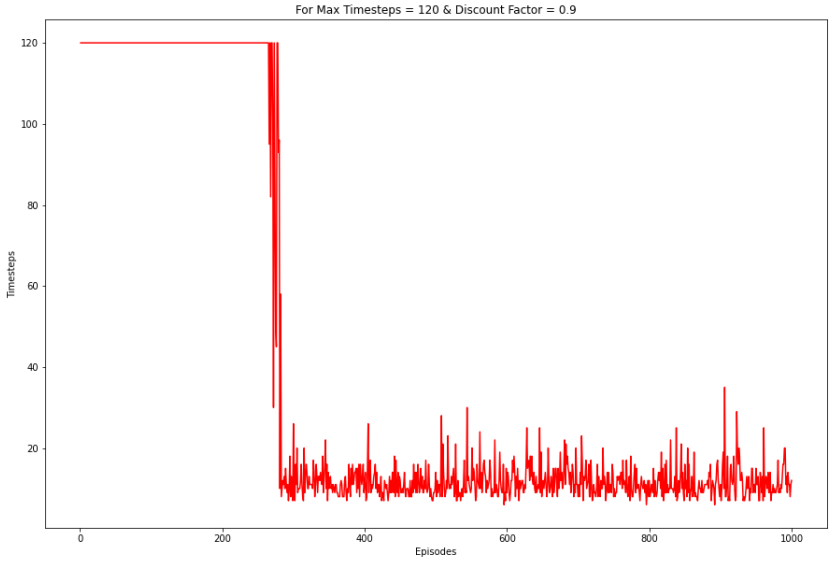

For Max Timesteps = 100 & Discount Factor = 0.1
Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

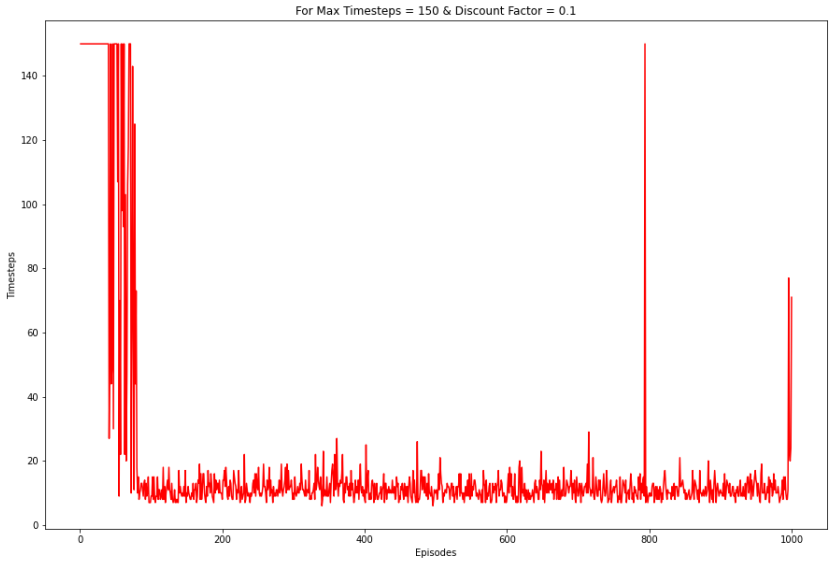

For Max Timesteps = 100 & Discount Factor = 0.5
Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

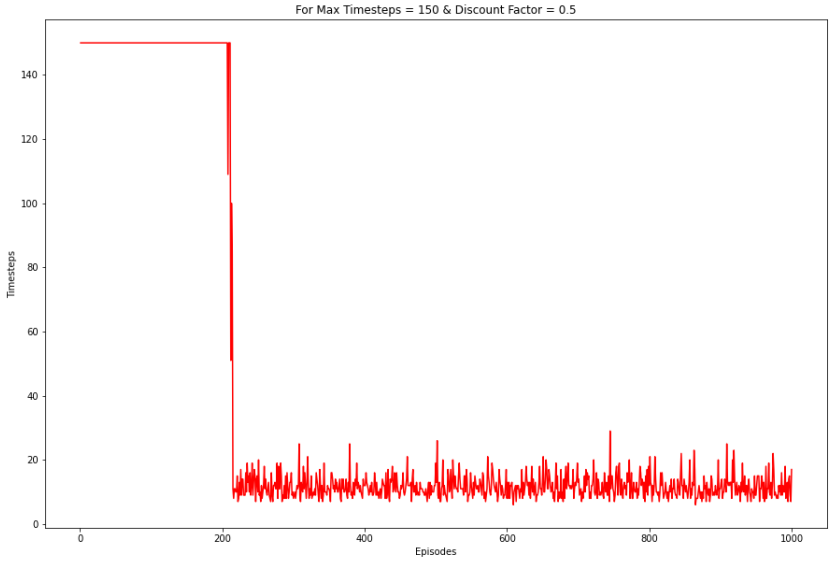

For Max Timesteps = 100 & Discount Factor = 0.9
Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

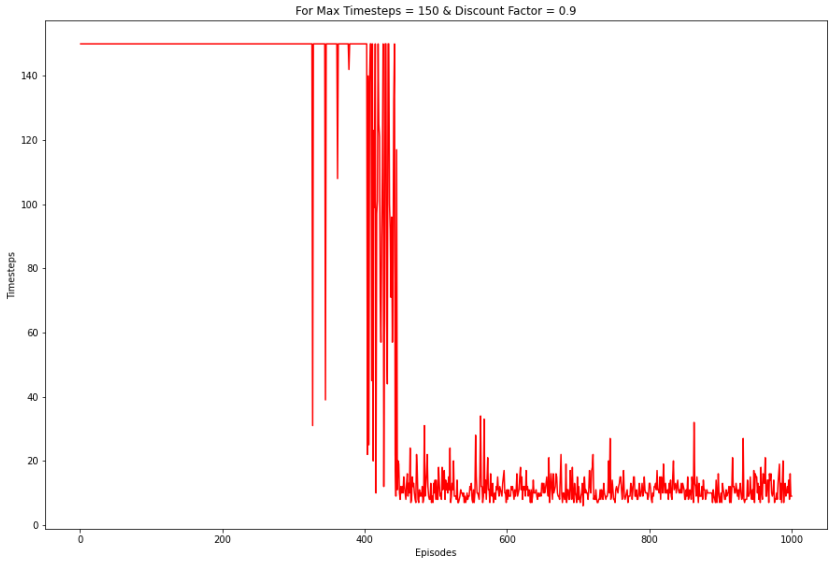

For Max Timesteps = 120 & Discount Factor = 0.1
Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.9
Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000


For Max Timesteps = 150 & Discount Factor = 0.1
Timesteps
Episodes
0
20
40
60
80
100
120
140
0
200
400
600
800
1000


For Max Timesteps = 150 & Discount Factor = 0.5
Timesteps
Episodes
0
20
40
60
80
100
120
140
0
200
400
600
800
1000

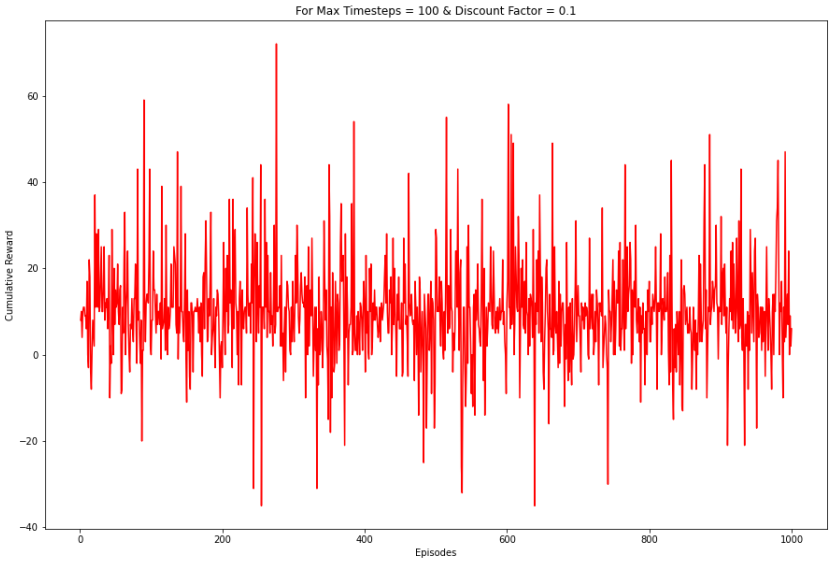


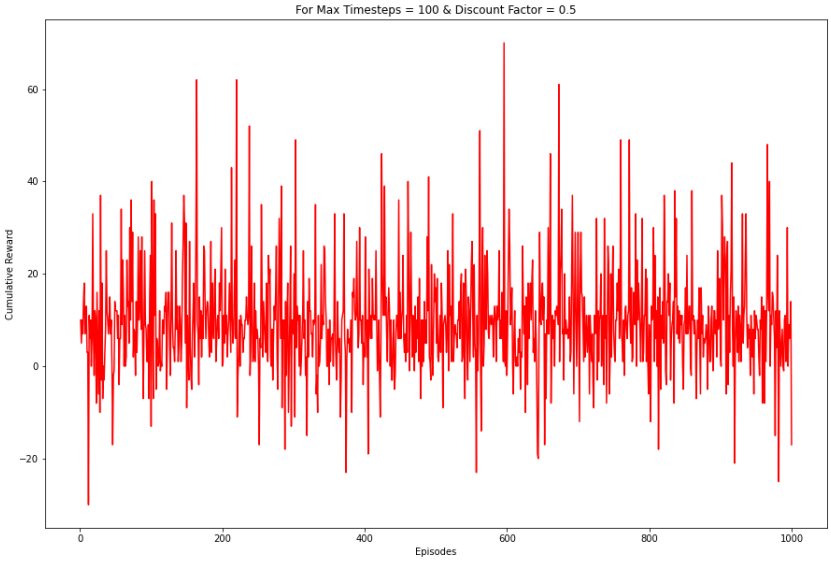
Plots above are the **Timesteps Used** for each of the Episodes for the **Stochastic Environment** across different pair of (**Max Time steps, Discount Factor**).

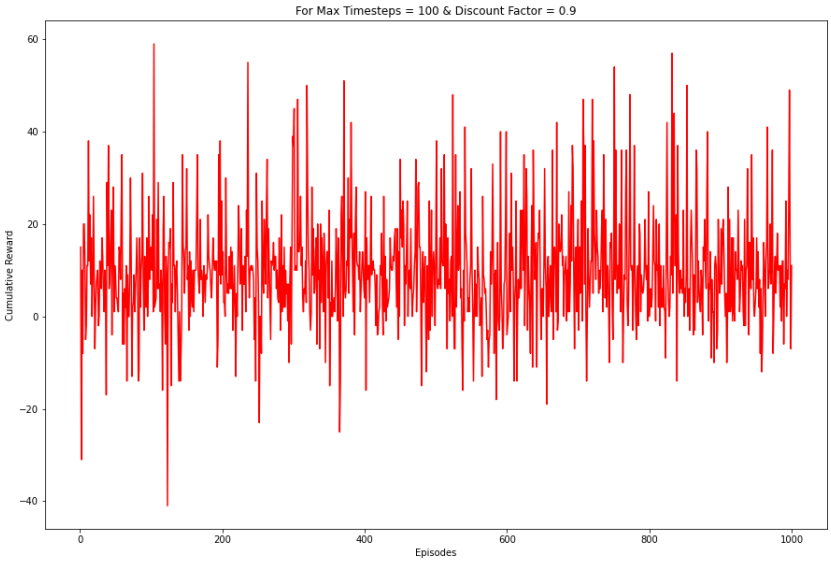
Most of episodes are utilizing maximum timesteps without reaching the goal, i.e. it's better to increase the value more so that the agent can reach the goal and hence maximize it's reward.

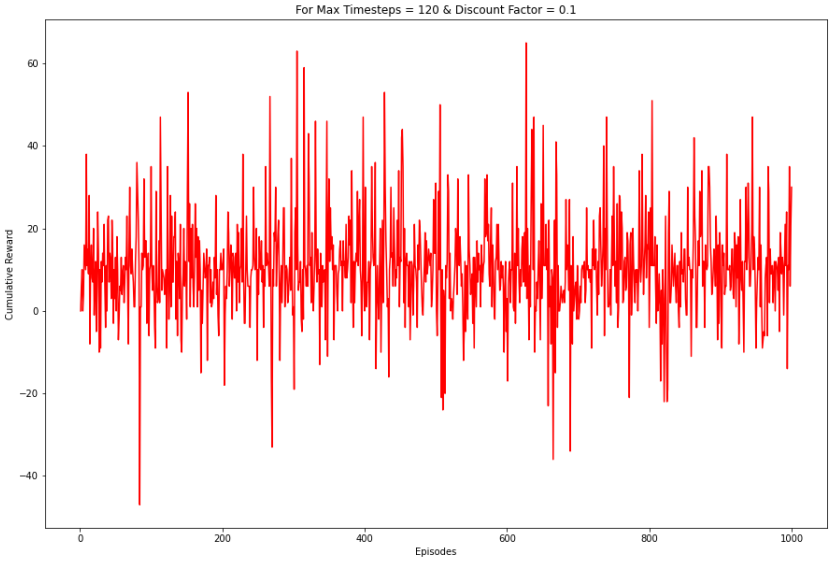
Now, Let's look at the plots obtained from TD(0) :

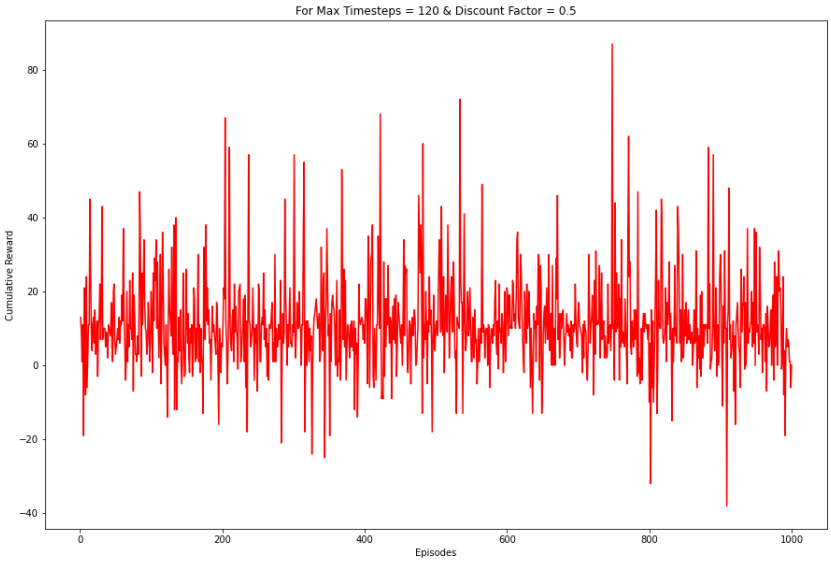
Cumulative Reward Plots :

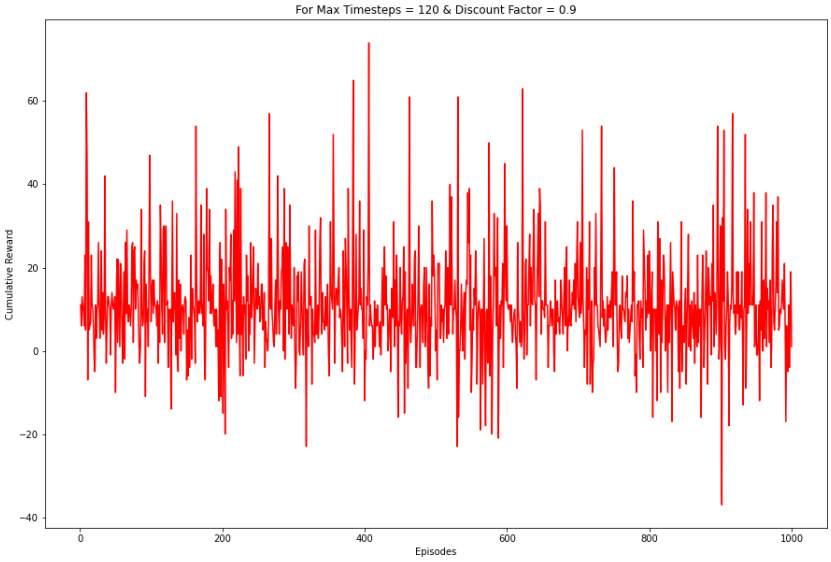

For Max Timesteps = 100 & Discount Factor = 0.1
Cumulative Reward
Episodes
60
40
20
0
−20
−40
0
200
400
600
800
1000

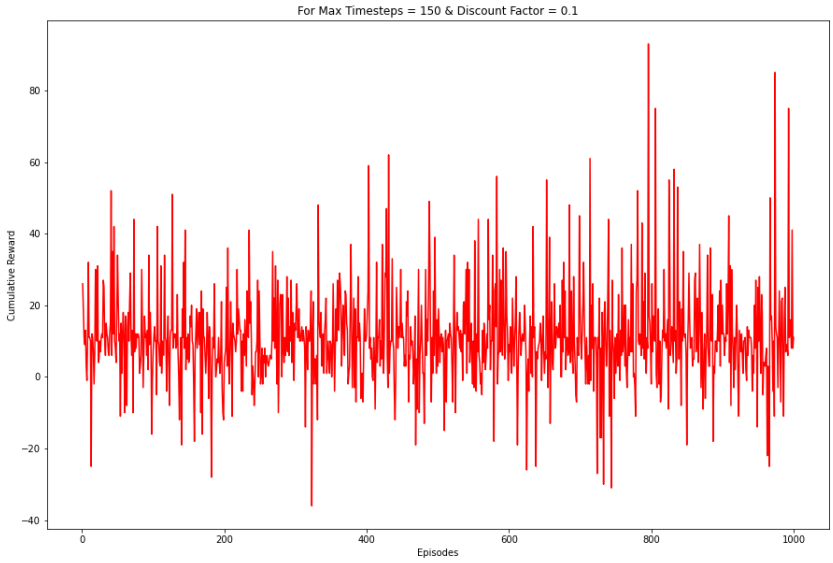

For Max Timesteps = 100 & Discount Factor = 0.5
Cumulative Reward
Episodes
60
40
20
0
−20
0
200
400
600
800
1000

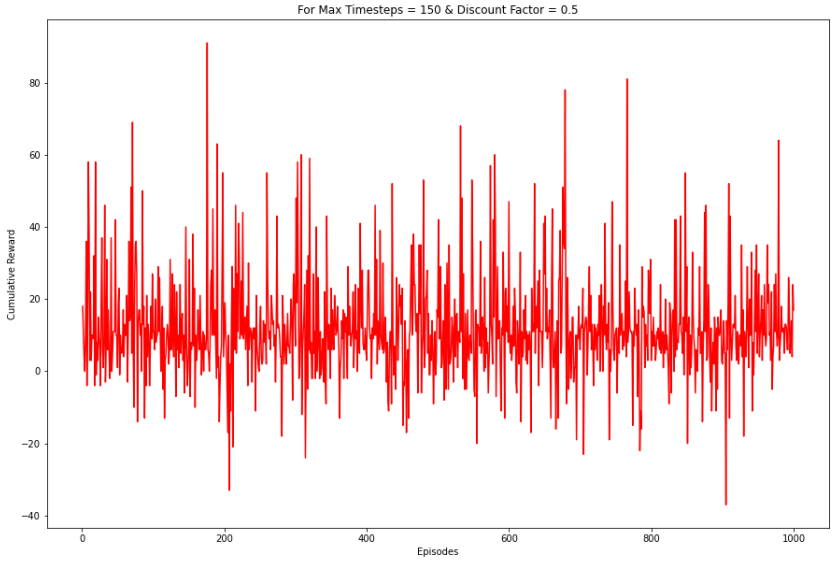

For Max Timesteps = 100 & Discount Factor = 0.9
Cumulative Reward
Episodes
60
40
20
0
−20
−40
0
200
400
600
800
1000

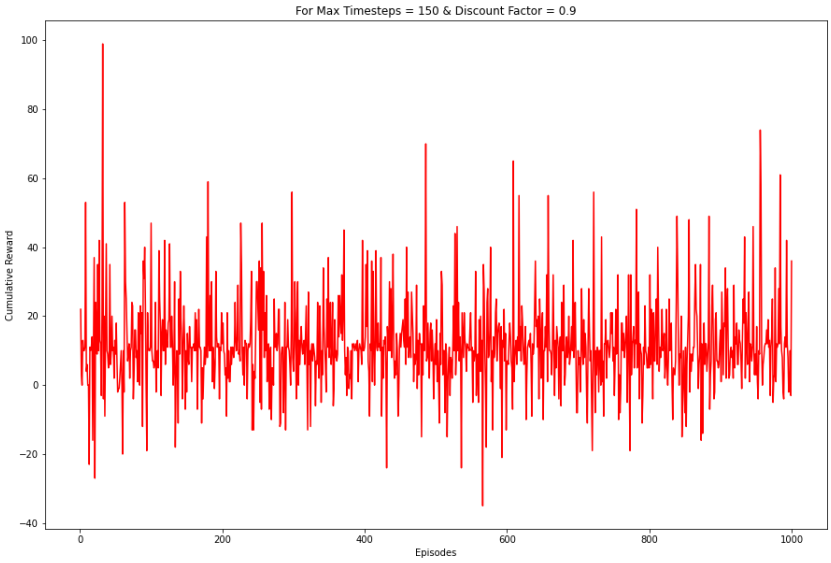

For Max Timesteps = 120 & Discount Factor = 0.1
Cumulative Reward
Episodes
60
40
20
0
−20
−40
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Cumulative Reward
Episodes
80
60
40
20
0
−20
−40
0
200
400
600
800
1000

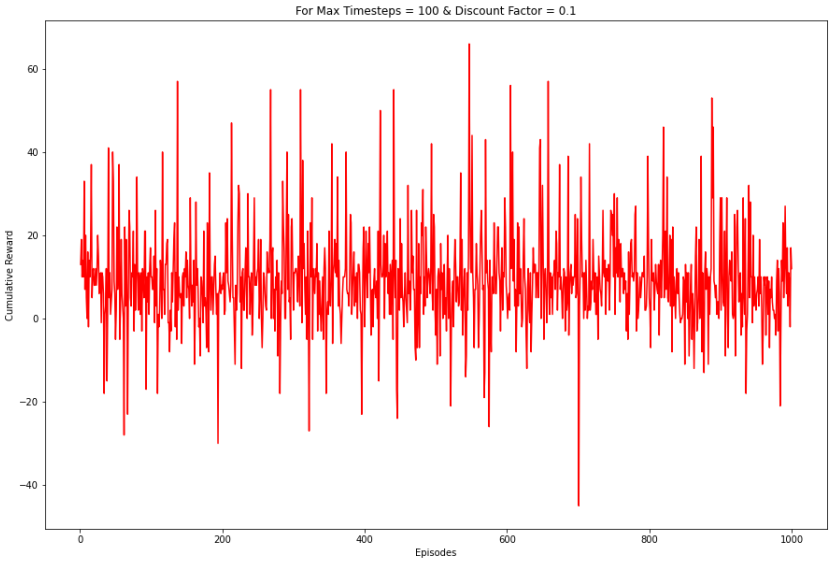

For Max Timesteps = 120 & Discount Factor = 0.9
Cumulative Reward
Episodes
60
40
20
0
−20
−40
0
200
400
600
800
1000

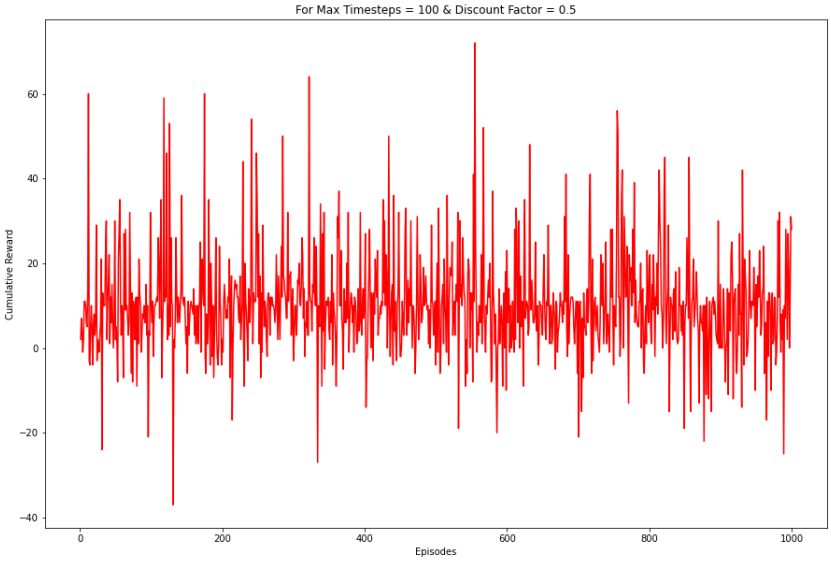

For Max Timesteps = 150 & Discount Factor = 0.1
Cumulative Reward
Episodes
80
60
40
20
0
−20
−40
0
200
400
600
800
1000

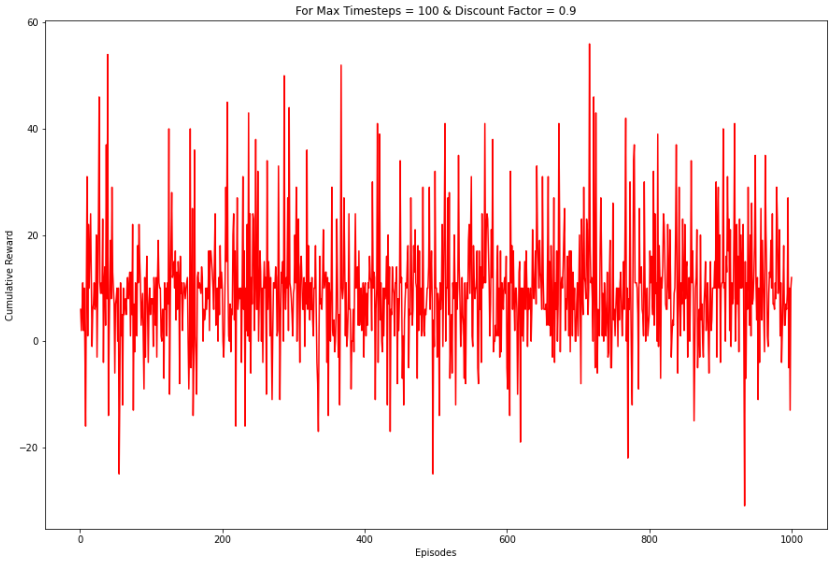

For Max Timesteps = 150 & Discount Factor = 0.5
Cumulative Reward
Episodes
80
60
40
20
0
−20
−40
0
200
400
600
800
1000

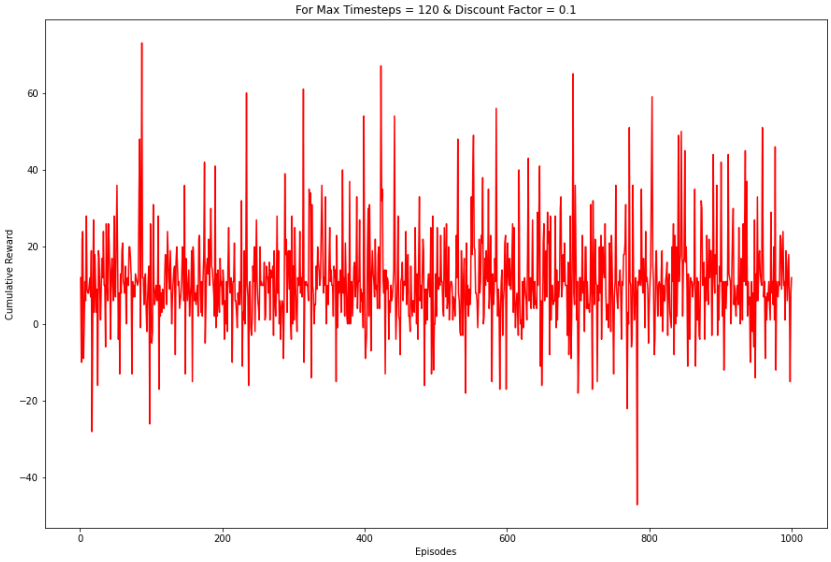


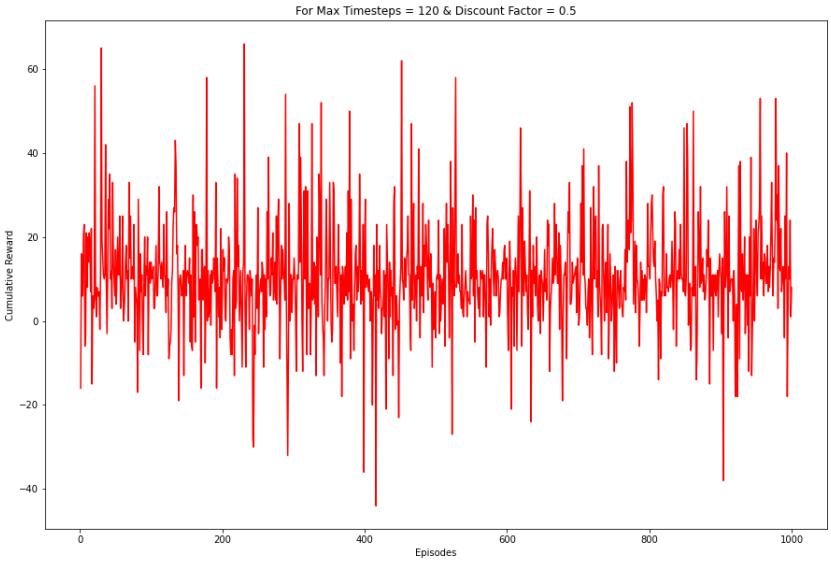
Plots above are the **Cumulative Rewards** for each of the Episodes for the **Deterministic Environment** across different pair of (**Max Time steps, Discount Factor**).

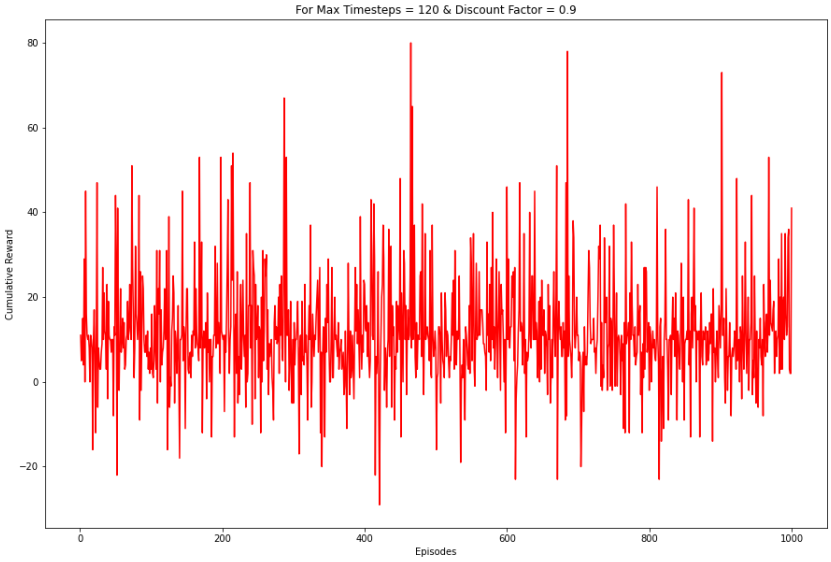

For Max Timesteps = 100 & Discount Factor = 0.1
Cumulative Reward
Episodes
60
40
20
0
−20
−40
0
200
400
600
800
1000

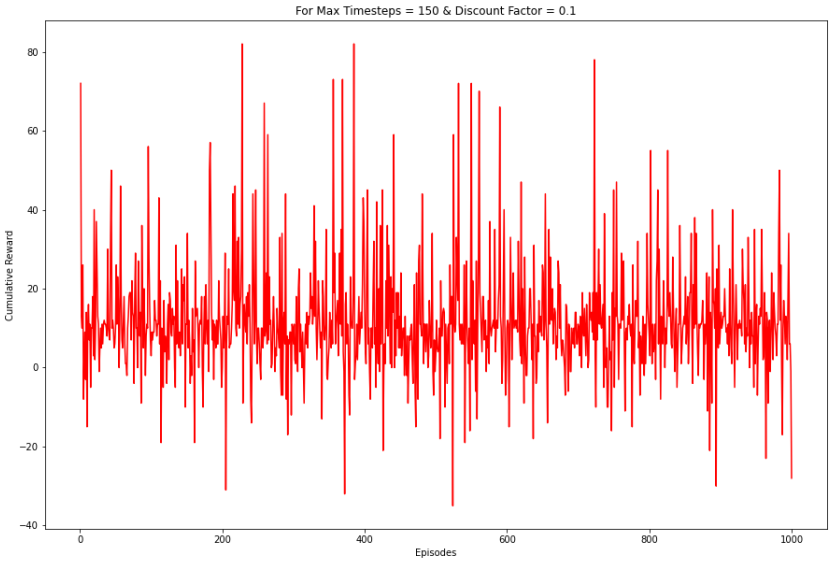

For Max Timesteps = 100 & Discount Factor = 0.5
Cumulative Reward
Episodes
60
40
20
0
−20
−40
0
200
400
600
800
1000

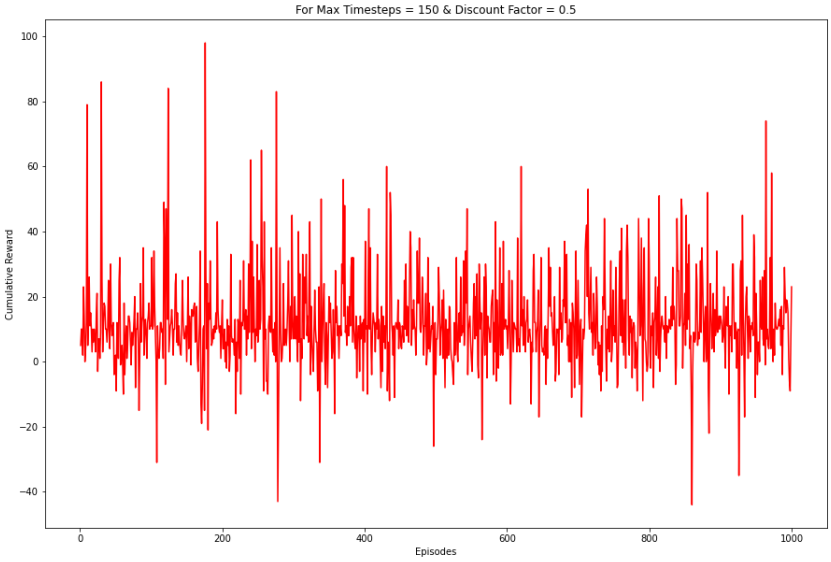

For Max Timesteps = 100 & Discount Factor = 0.9
Cumulative Reward
Episodes
60
40
20
0
−20
0
200
400
600
800
1000

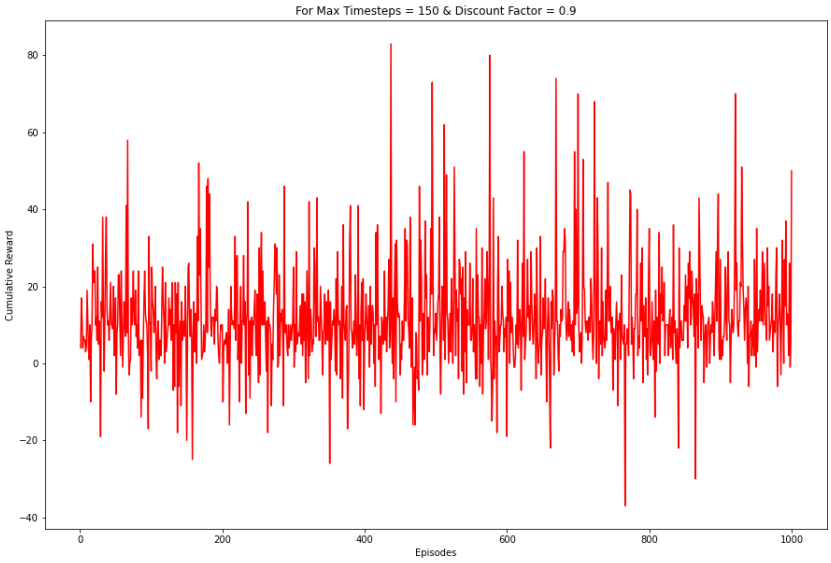

For Max Timesteps = 120 & Discount Factor = 0.1
Cumulative Reward
Episodes
60
40
20
0
−20
−40
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Cumulative Reward
Episodes
60
40
20
0
−20
−40
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.9
Cumulative Reward
Episodes
80
60
40
20
0
−20
0
200
400
600
800
1000


For Max Timesteps = 150 & Discount Factor = 0.1
Cumulative Reward
Episodes
80
60
40
20
0
−20
−40
0
200
400
600
800
1000

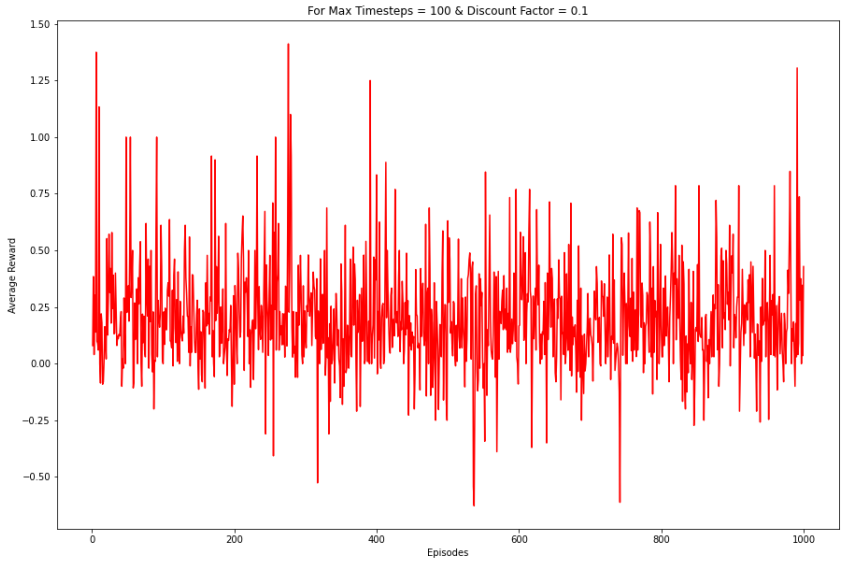

For Max Timesteps = 150 & Discount Factor = 0.5
Cumulative Reward
Episodes
100
80
60
40
20
0
−20
−40
0
200
400
600
800
1000

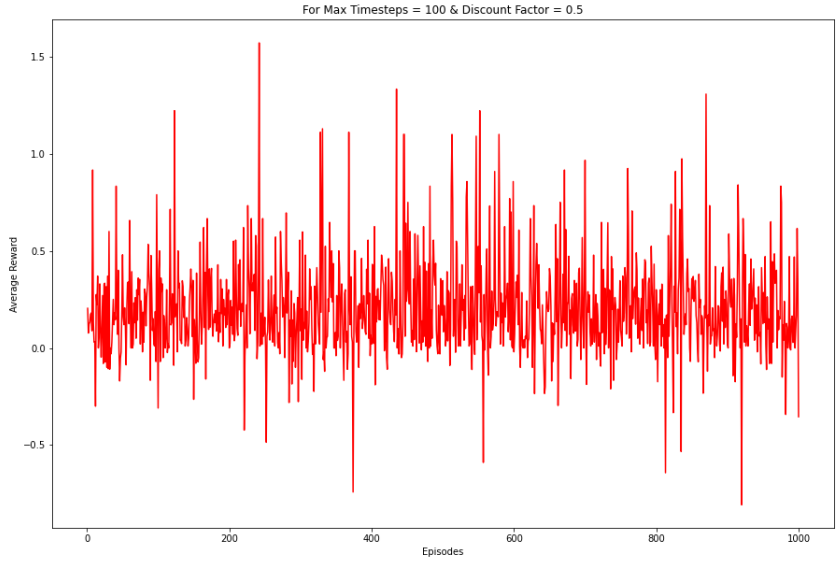


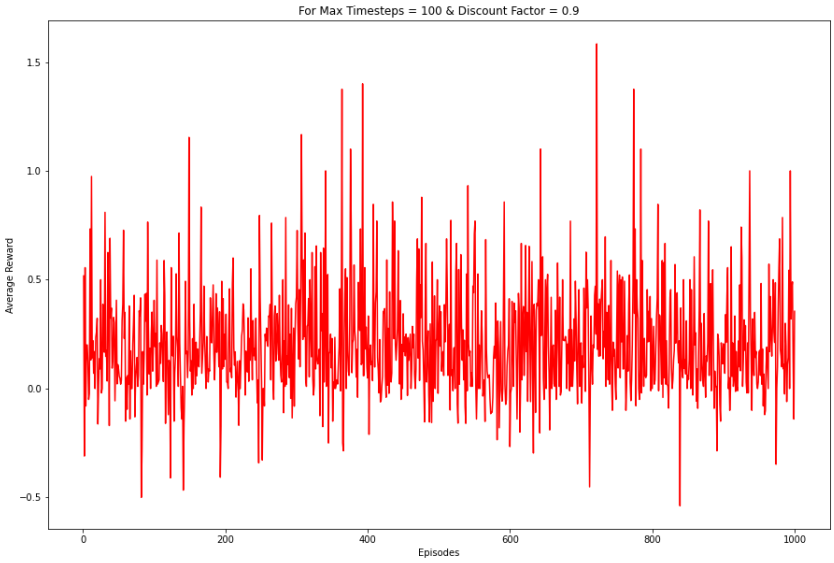
Plots above are the **Cumulative Rewards** for each of the Episodes for the **Stochastic Environment** across different pair of (**Max Time steps, Discount Factor**).

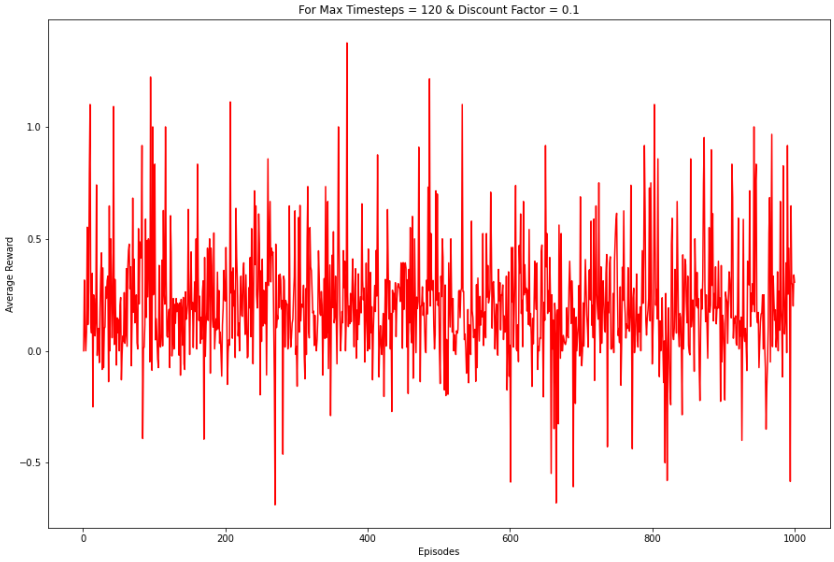
It is observed that, the Cumulative Rewards are fluctuating around a point for both types of Environments; which is around 10 in this case.

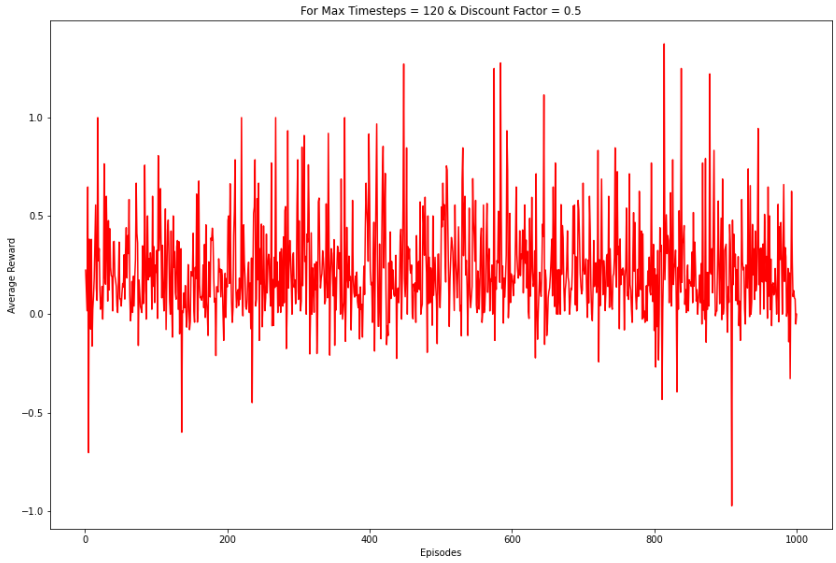
Average Reward Plots :

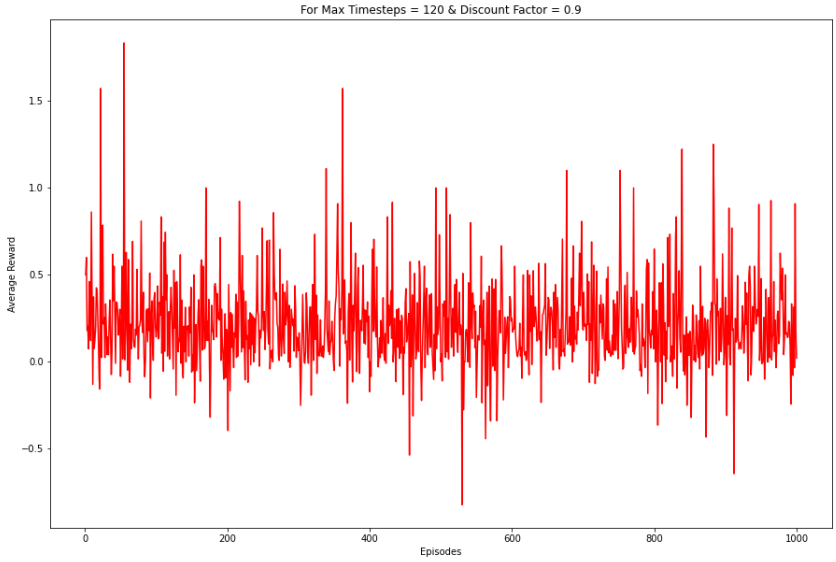

For Max Timesteps = 100 & Discount Factor = 0.1
Average Reward
Episodes
1.50
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
200
400
600
800
1000

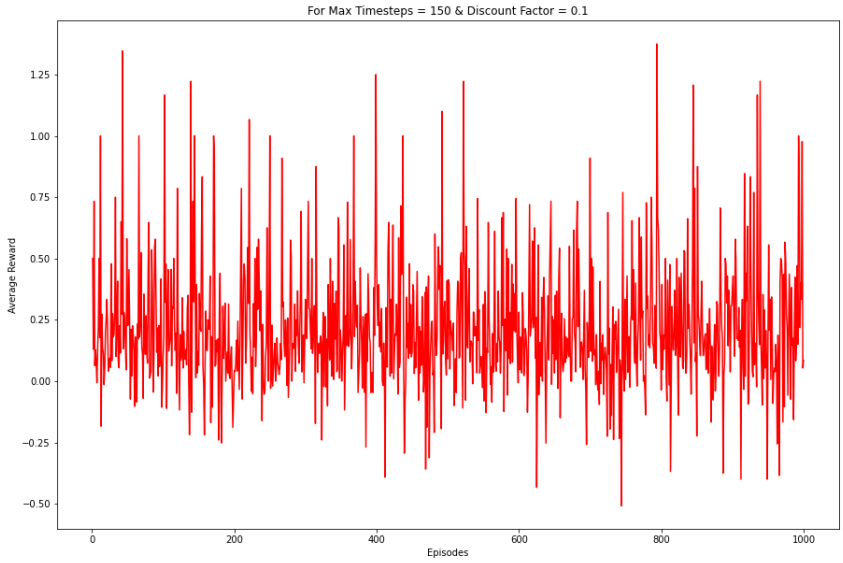

For Max Timesteps = 100 & Discount Factor = 0.5
Average Reward
Episodes
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000

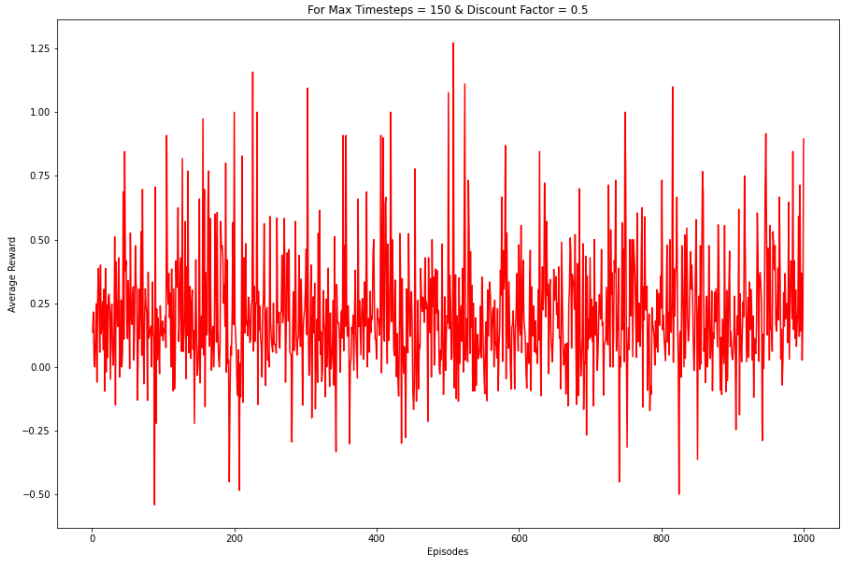

For Max Timesteps = 100 & Discount Factor = 0.9
Average Reward
Episodes
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000

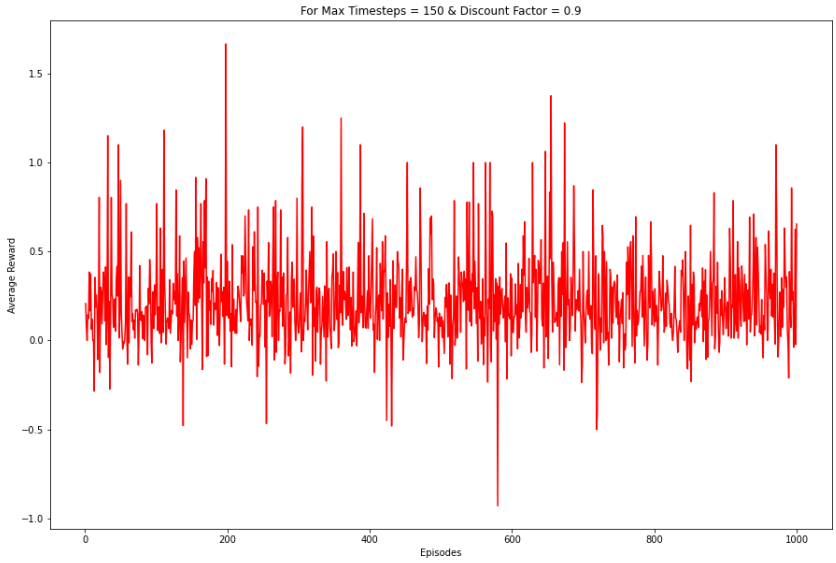

For Max Timesteps = 120 & Discount Factor = 0.1
Average Reward
Episodes
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Average Reward
Episodes
1.0
0.5
0.0
−0.5
−1.0
0
200
400
600
800
1000

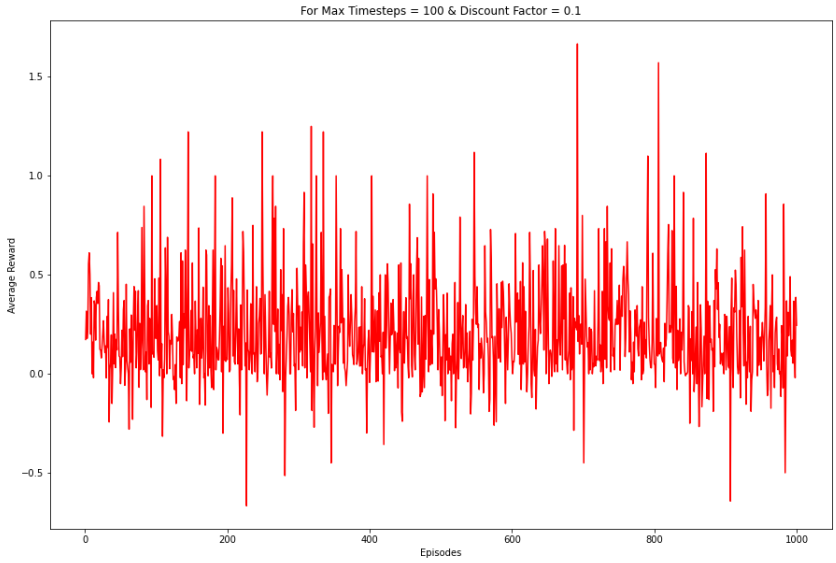

For Max Timesteps = 120 & Discount Factor = 0.9
Average Reward
Episodes
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000

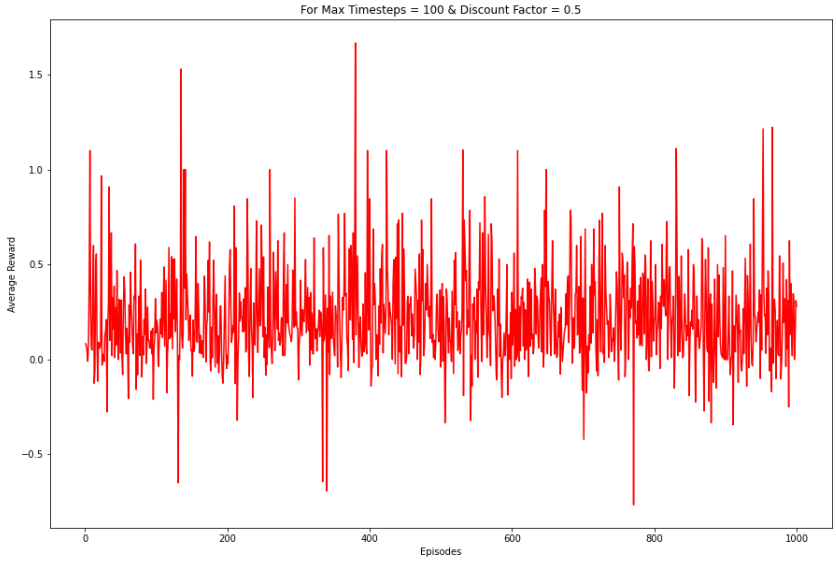

For Max Timesteps = 150 & Discount Factor = 0.1
Average Reward
Episodes
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
200
400
600
800
1000

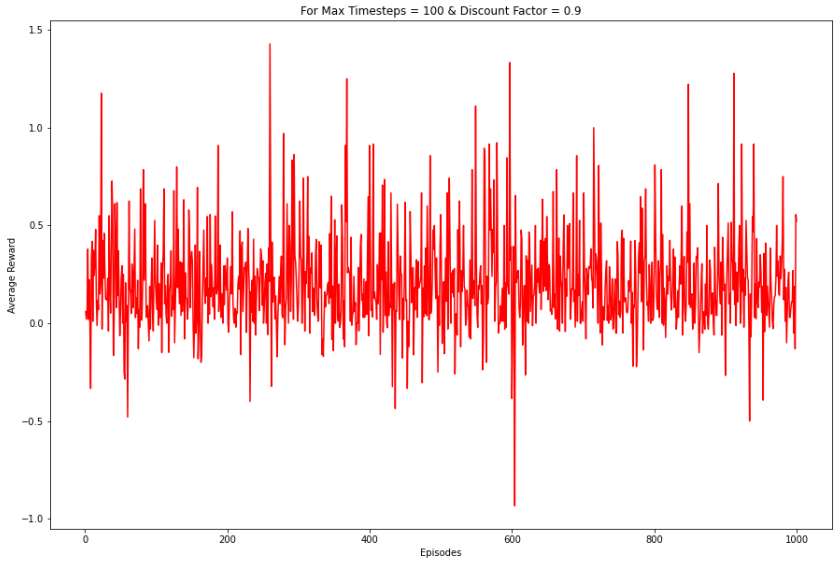

For Max Timesteps = 150 & Discount Factor = 0.5
Average Reward
Episodes
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
0
200
400
600
800
1000

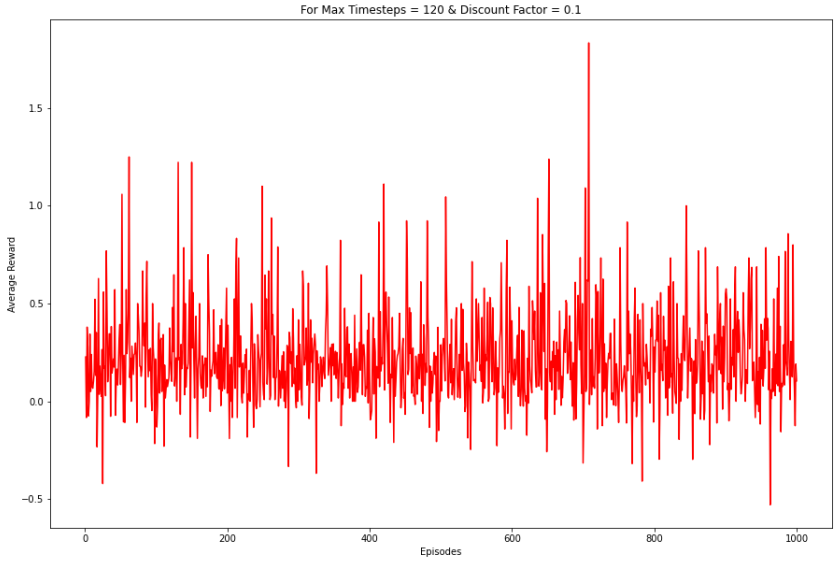


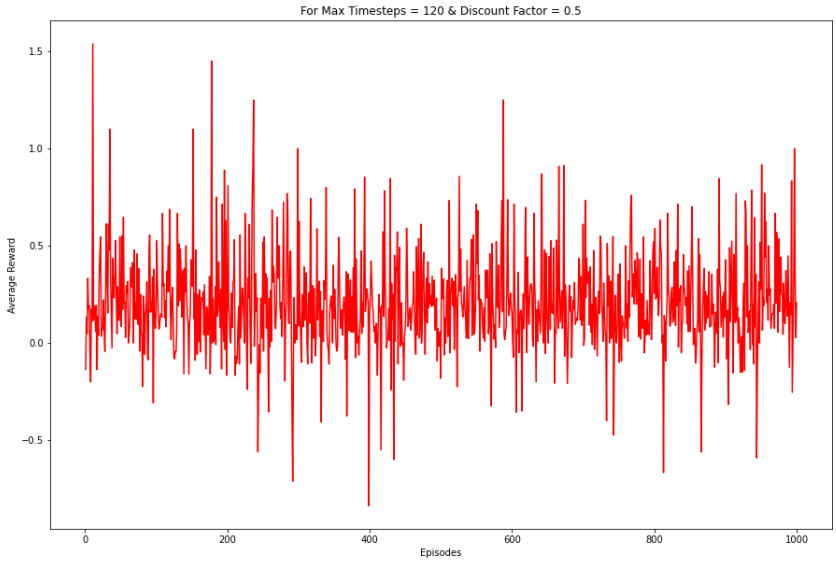
Plots above are the **Average Rewards** for each of the Episodes for the **Deterministic Environment** across different pair of (**Max Time steps, Discount Factor**).

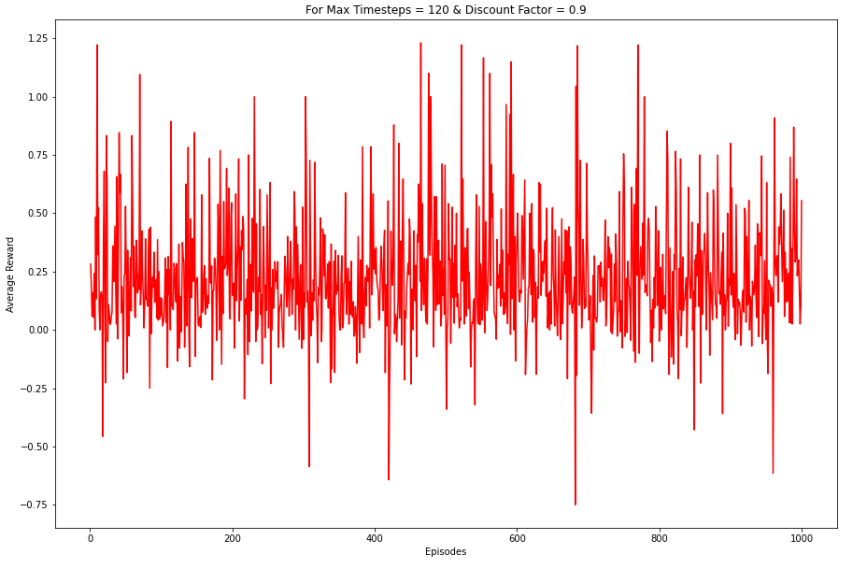

For Max Timesteps = 100 & Discount Factor = 0.1
Average Reward
Episodes
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000

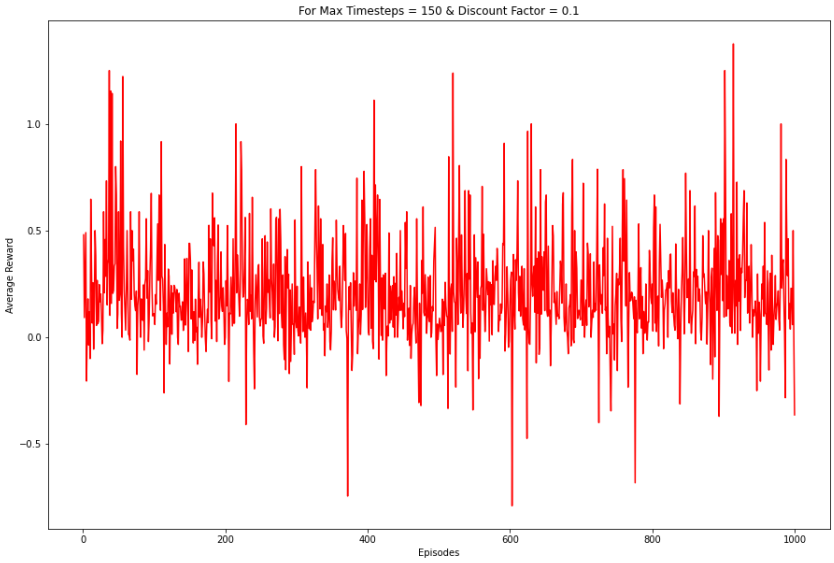

For Max Timesteps = 100 & Discount Factor = 0.5
Average Reward
Episodes
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000

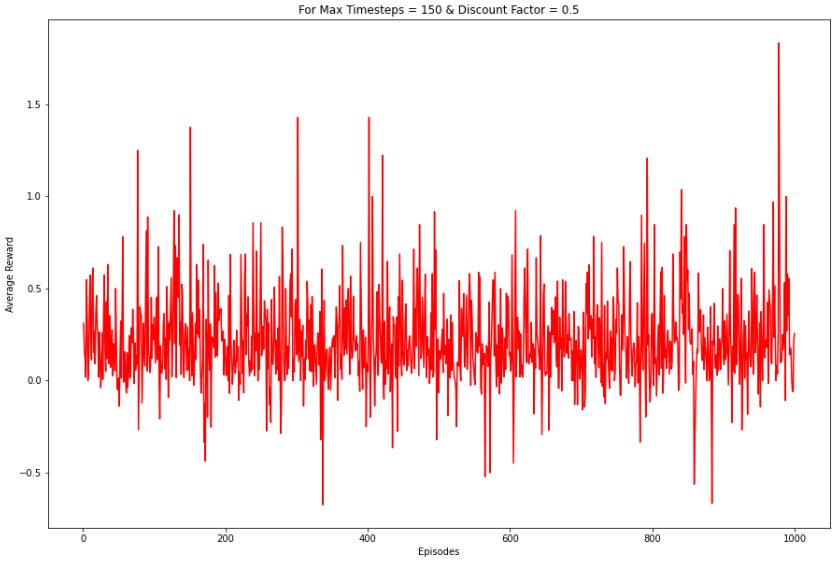

For Max Timesteps = 100 & Discount Factor = 0.9
Average Reward
Episodes
1.5
1.0
0.5
0.0
−0.5
−1.0
0
200
400
600
800
1000

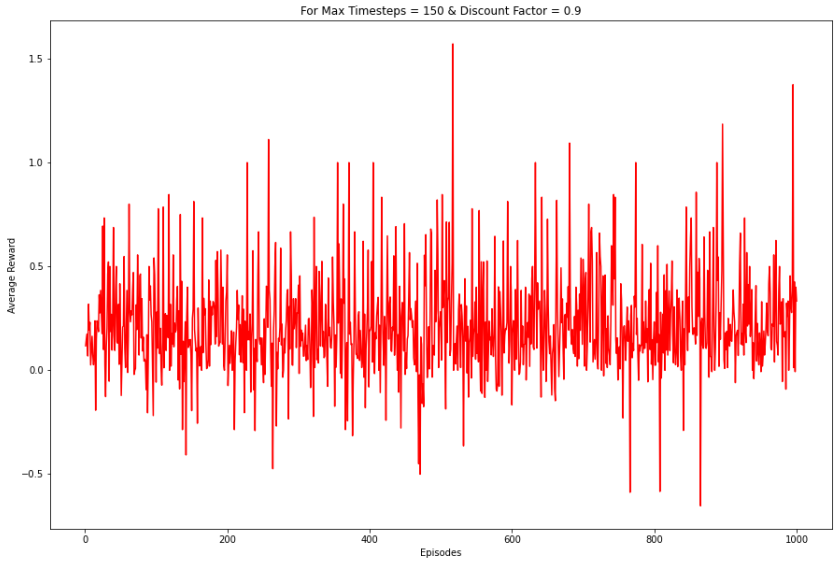

For Max Timesteps = 120 & Discount Factor = 0.1
Average Reward
Episodes
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Average Reward
Episodes
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.9
Average Reward
Episodes
1.25
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
0
200
400
600
800
1000


For Max Timesteps = 150 & Discount Factor = 0.1
Average Reward
Episodes
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000

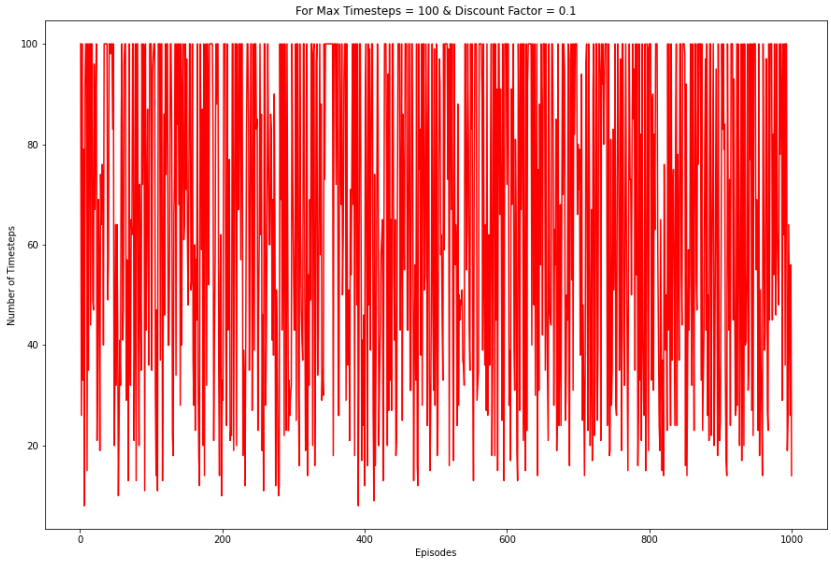

For Max Timesteps = 150 & Discount Factor = 0.5
Average Reward
Episodes
1.5
1.0
0.5
0.0
−0.5
0
200
400
600
800
1000

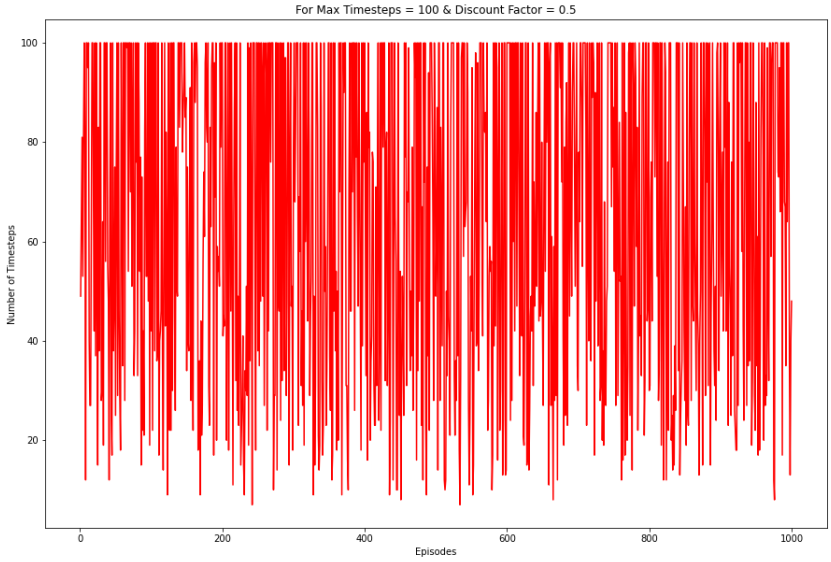


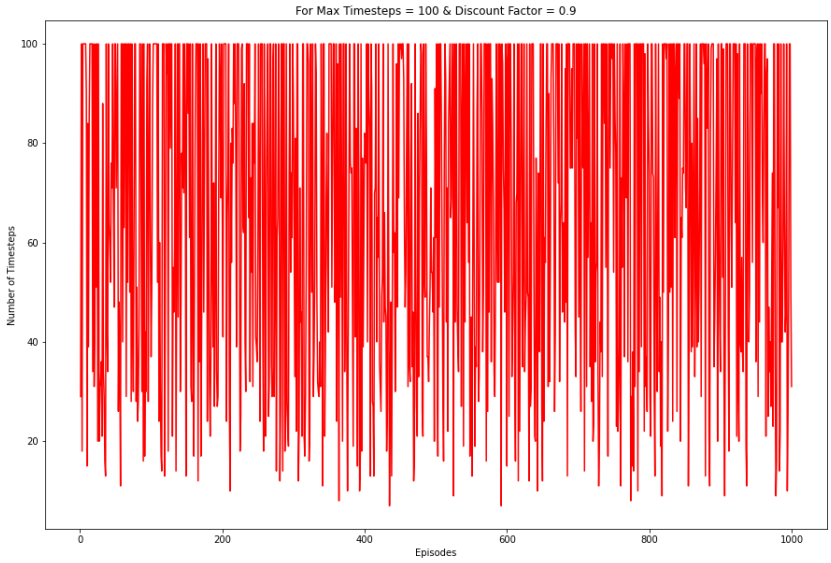
Plots above are the **Average Rewards** for each of the Episodes for the **Stochastic Environment** across different pair of (**Max Time steps, Discount Factor**).

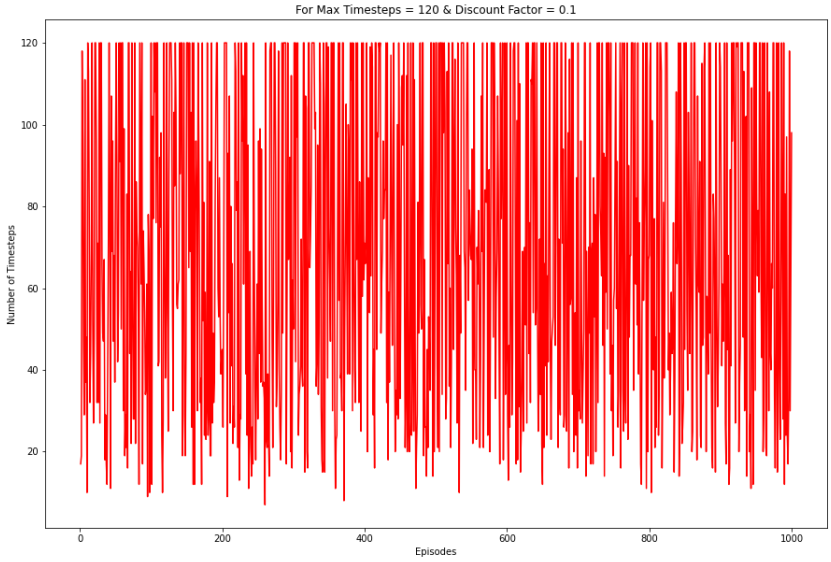
It is observed that, the Average Rewards are fluctuating around a point for both types of Environments; which is around 0.2 in this case.

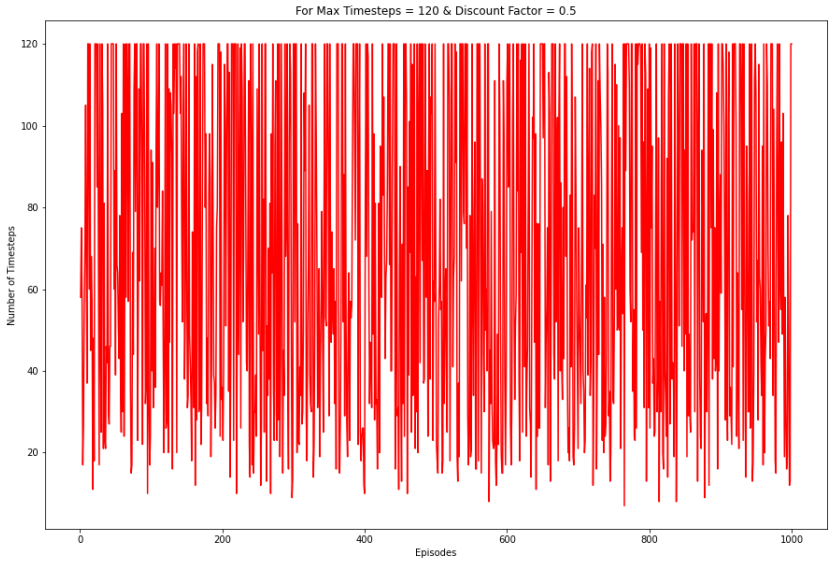
Timesteps per Episode Plots :

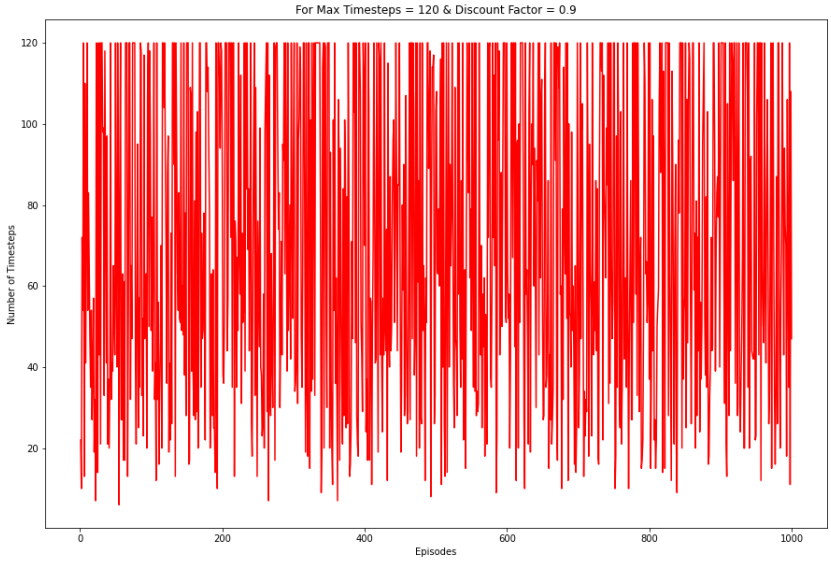

For Max Timesteps = 100 & Discount Factor = 0.1
Number of Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

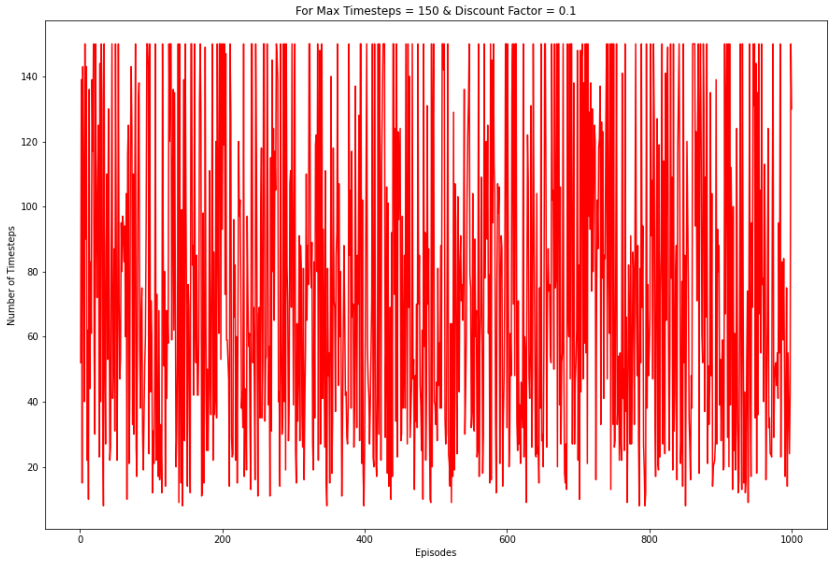

For Max Timesteps = 100 & Discount Factor = 0.5
Number of Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

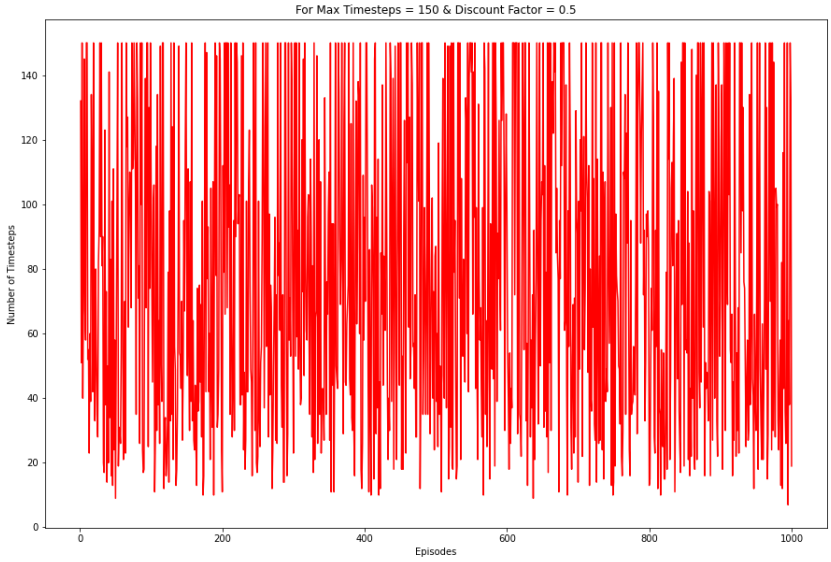

For Max Timesteps = 100 & Discount Factor = 0.9
Number of Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

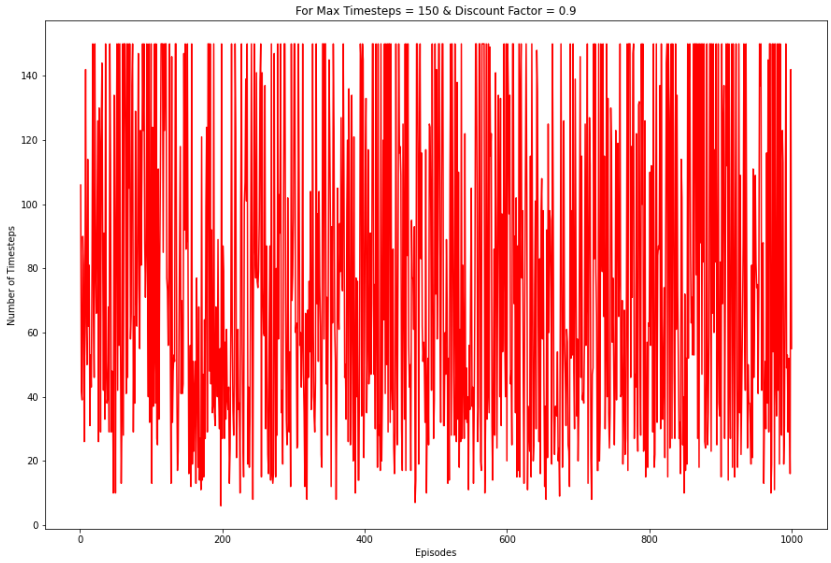

For Max Timesteps = 120 & Discount Factor = 0.1
Number of Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Number of Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000

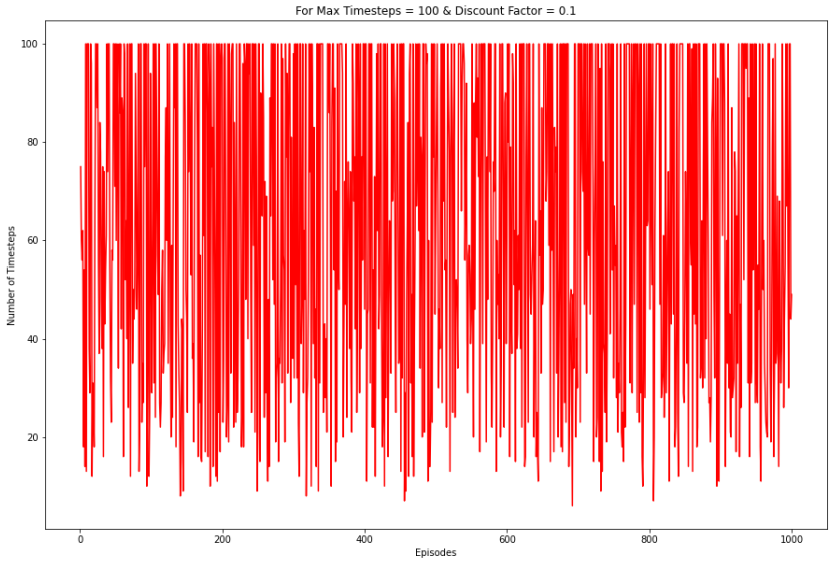

For Max Timesteps = 120 & Discount Factor = 0.9
Number of Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000

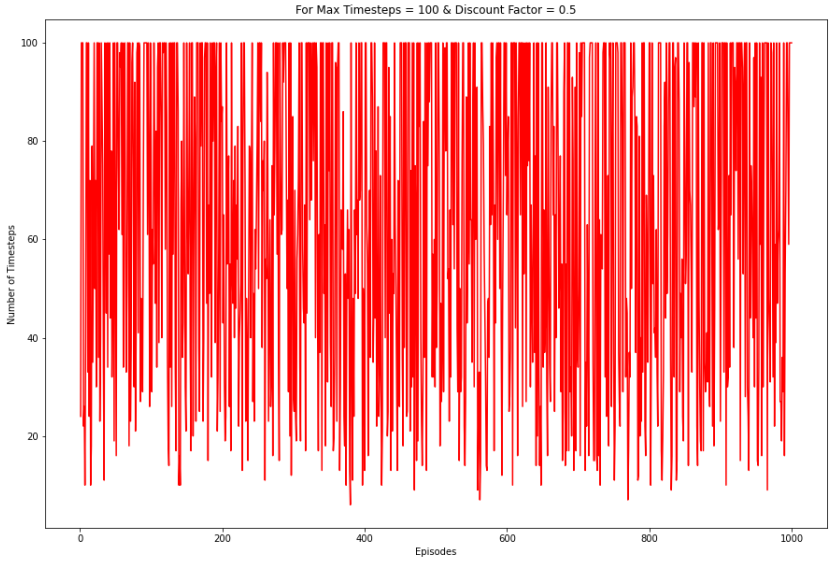

For Max Timesteps = 150 & Discount Factor = 0.1
Number of Timesteps
Episodes
0
200
400
600
800
1000
20
40
60
80
100
120
140

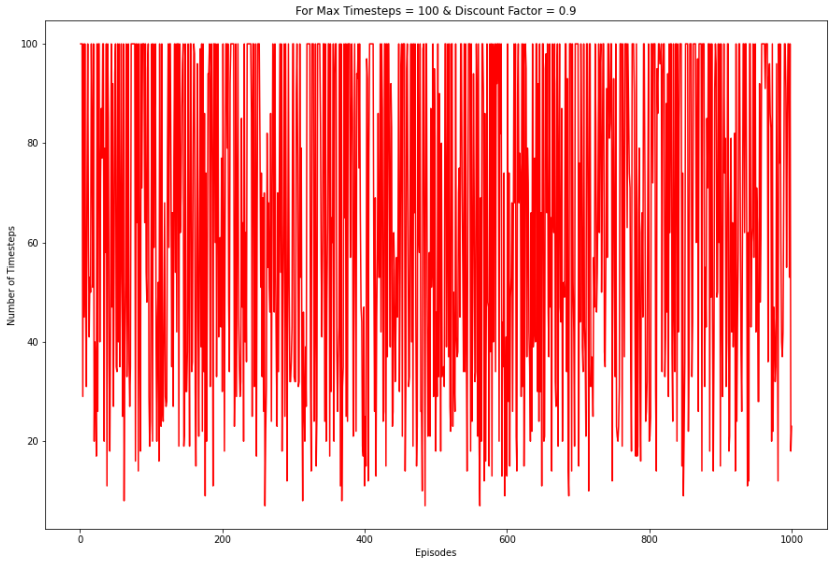

For Max Timesteps = 150 & Discount Factor = 0.5
Number of Timesteps
Episodes
0
200
400
600
800
1000
0
20
40
60
80
100
120
140

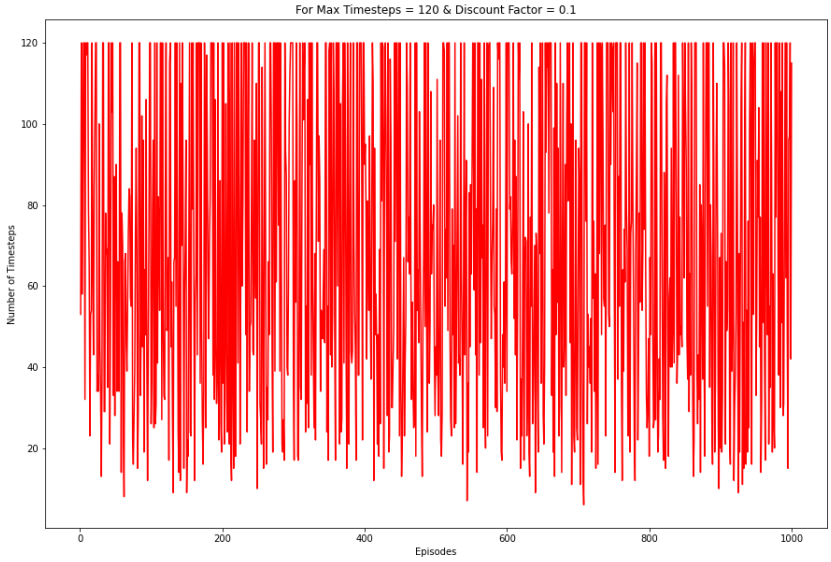


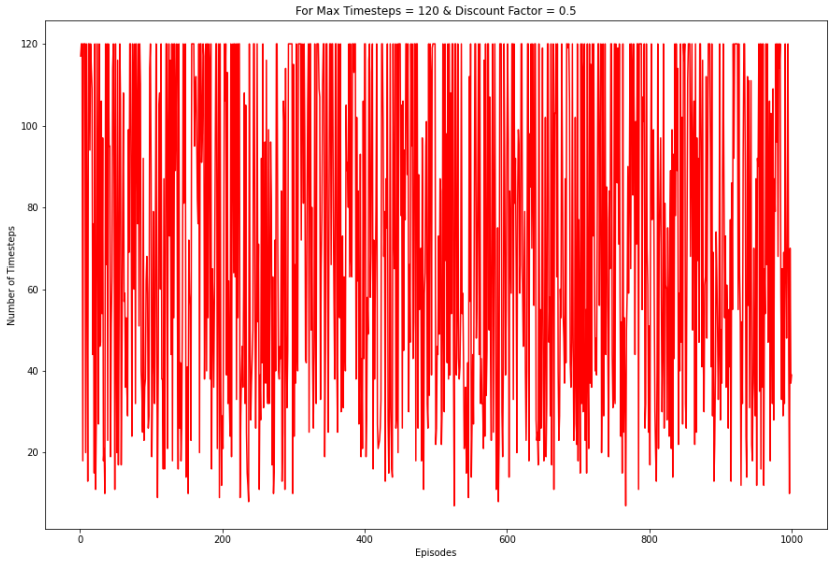
Plots above are the **Timesteps Used** for each of the Episodes for the **Deterministic Environment** across different pair of (**Max Time steps, Discount Factor**).

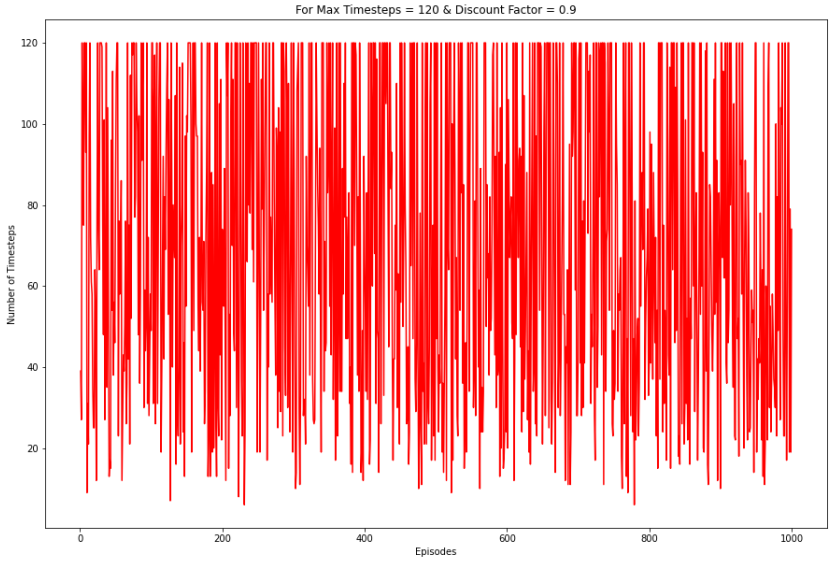

For Max Timesteps = 100 & Discount Factor = 0.1
Number of Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

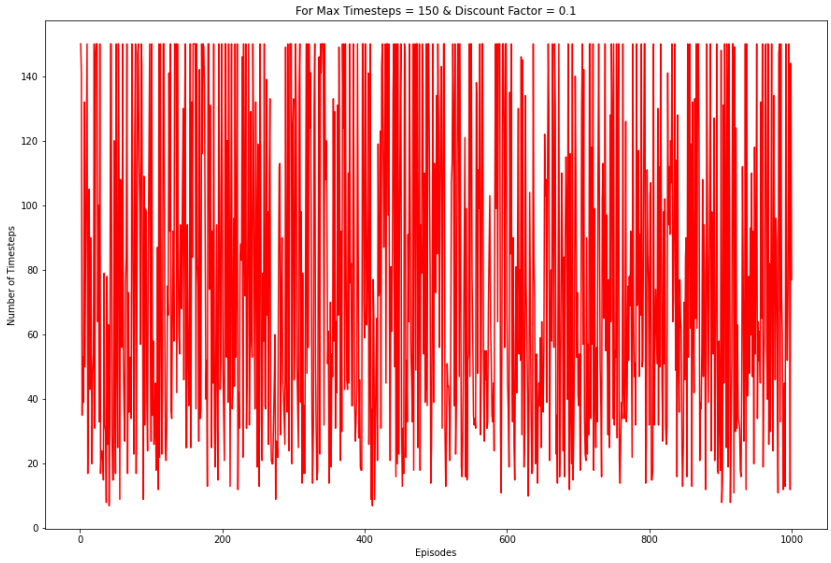

For Max Timesteps = 100 & Discount Factor = 0.5
Number of Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

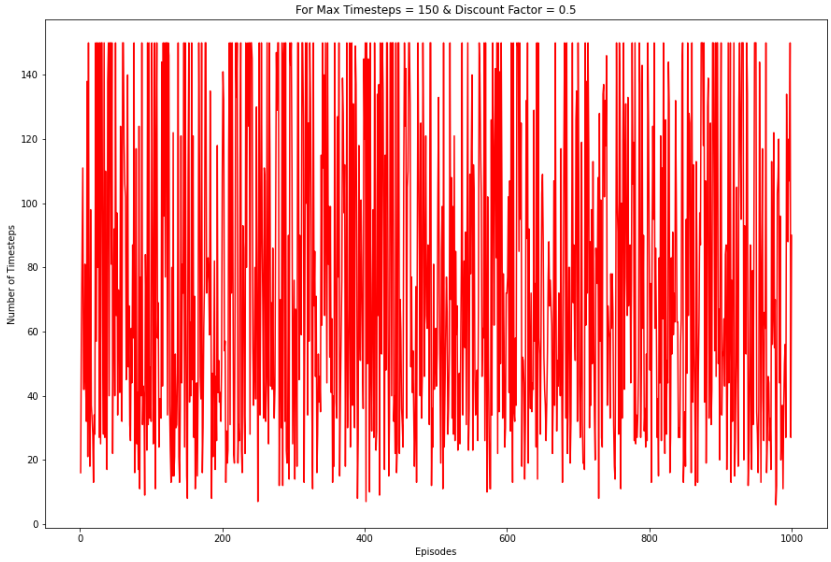

For Max Timesteps = 100 & Discount Factor = 0.9
Number of Timesteps
Episodes
100
80
60
40
20
0
200
400
600
800
1000

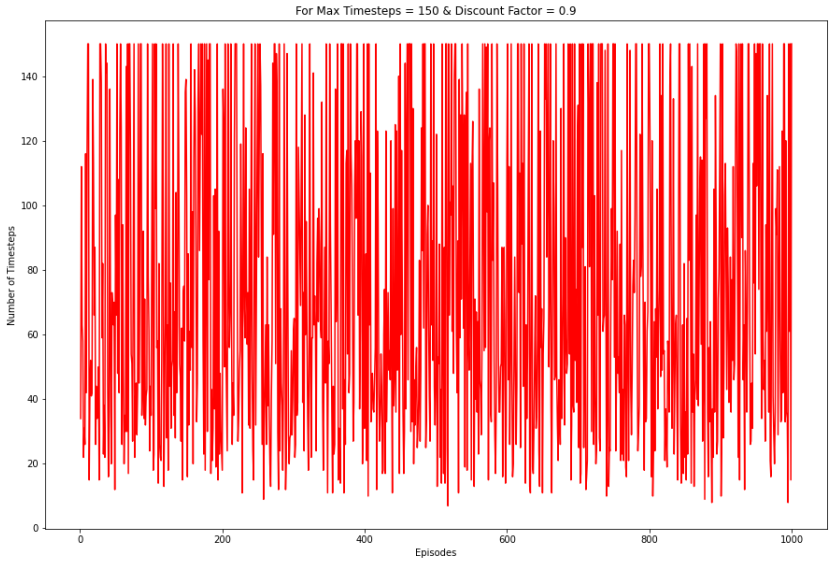

For Max Timesteps = 120 & Discount Factor = 0.1
Number of Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.5
Number of Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000


For Max Timesteps = 120 & Discount Factor = 0.9
Number of Timesteps
Episodes
120
100
80
60
40
20
0
200
400
600
800
1000


For Max Timesteps = 150 & Discount Factor = 0.1
Number of Timesteps
Episodes
0
20
40
60
80
100
120
140
0
200
400
600
800
1000

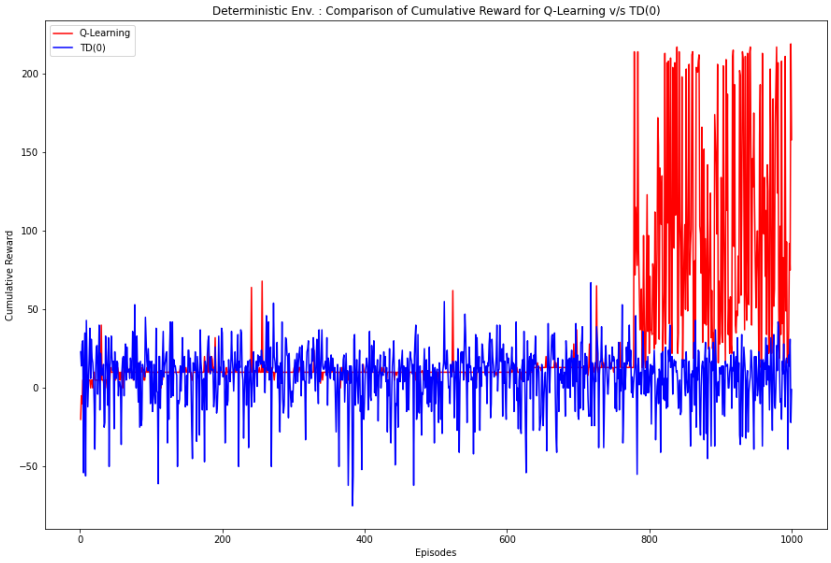

For Max Timesteps = 150 & Discount Factor = 0.5
Number of Timesteps
Episodes
0
20
40
60
80
100
120
140
0
200
400
600
800
1000

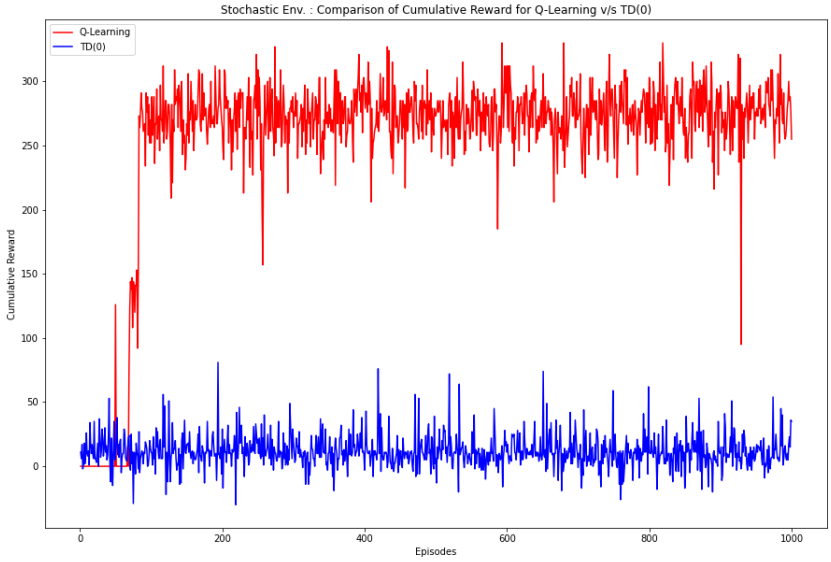


Plots above are the **Timesteps Used** for each of the Episodes for the **Stochastic Environment** across different pair of (**Max Time steps, Discount Factor**).

Most of episodes are utilizing maximum timesteps without reaching the goal, i.e. it's better to increase the value more so that the agent can reach the goal and hence maximize it's reward.

Now, let's compare the Rewards graph for the two algorithms applied :

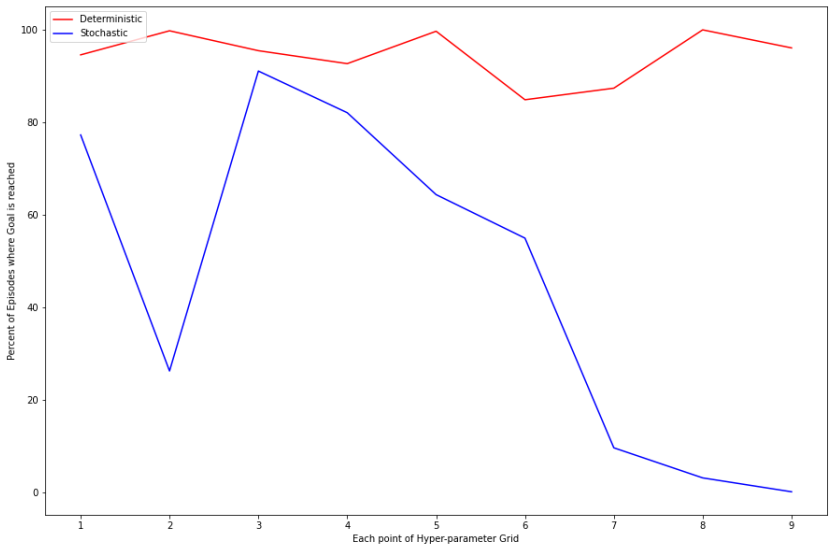




It is observed that Q-Learning tends to obtain higher Cumulative Rewards than TD(0), because it is greedy in nature i.e. prone to higher reward values.

Due to the straightforward logic behind Deterministic Environments, they

are capable of reaching the Goal more than Stochastic Environments as shown below :



It is because, our Stochastic Environment is defined such that the agent retains its last position 33 out of 100 times and thus involves lesser attempt to reach its Goal.

5 Discussion

It is observed that Deterministic Environments are more favorable in terms of reaching the Goal due to the less randomness in their behaviour. However, both types of environments reach nearly the same Cumulative Rewards for either of the two algorithms applied. Q-Learning tends to provide higher value of Cumulative Rewards than TD(0), due to its Greedy behaviour.

6 Safety

First safety concern is that the agent shouldn't go outside the environment grid. Therefore, the co-ordinates of the agent are clipped so that it remains within the environment. Another potential threat in case of Stochastic Environment is getting the same Reward successively when it retains the same position. For example : If the agent is in the state of the Diamond, then it already obtained

the corresponding Reward. For the next time step, if it retains the same position, then it will obtain the same Reward again! Therefore, the environment is adjusted such that the Reward is considered only for the first time and no Reward for the successive retention of the same State.