Results:

The following parameters have been varied:

1. Discount factor. **(γ)**
2. Learning Rate. **(α)**

The path chosen is as follows: *(This path has been chosen for ALL results illustrated below)*

Forward → Right → Stop → Left → Forward → Stop

(OR)

F0 → R1 → S3 → L2 → F0 → S3

Discount Factor:

States	F	R	L	S
F0	28.6525	38.7438	28.7792	28.4576
F1	22.5994	20.4063	22.7846	32.3782
F2	29.3042	-1.99553	-1.99863	-1.99732
F3	19.0582	18.9571	32.1115	18.8491
R0	0	0	0	0
R1	34.9921	33.3451	34.956	42.6093
R2	27.3351	-1.99732	-1.99636	-1.99746
R3	20.4635	17.7096	29.4989	18.7091
L0	0	0	0	0
L1	22.8138	19.9556	21.3218	31.2335
L2	41.0718	9.98358	9.9891	9.99017
L3	20.4277	18.3828	31.1132	20.4659
S0	0	0	0	0
S1	22.3179	21.0361	22.9176	31.3149
S2	28.6085	-1.99479	-1.9953	-1.99392
S3	32.2783	31.081	41.5576	30.9583

(Figure 2.4: Gamma = 0.9)

States	F	R	L	S
F0	6.4215	13.2198	8.54023	6.04327
F1	2.35307	2.57579	2.56346	9.60369
F2	6.01301	-1.99718	-1.99553	-1.99718
F3	1.26252	1.11339	8.47032	1.06074
R0	0	0	0	0
R1	14.4673	14.318	14.4558	21.6017
R2	5.87446	-1.99831	-1.99759	-1.99617
R3	1.12093	0.992242	8.34939	1.19741
L0	0	0	0	0
L1	2.57782	2.41608	2.56867	9.6653
L2	17.9752	9.97525	9.98729	9.99066
L3	1.41327	1.20677	8.34243	1.14306
S0	0	0	0	0
S1	2.54241	2.32004	2.62545	9.56156
S2	5.85305	-1.99771	-1.99654	-1.99173
S3	13.2119	12.8805	20.4272	13.1977

(Figure 2.5: Gamma = 0.6)

States	F	R	L	S
F0	-9.31638e-05	1.4503	1.18092	1.59538
F1	-2.097	-2.09644	-2.09952	-0.908223
F2	-1.8426	-1.99597	-1.99291	-1.99505
F3	-2.16965	-2.17698	-0.999972	-2.17609
R0	0	0	0	0
R1	9.88464	9.88756	9.88229	11.0801
R2	-1.8372	-1.99876	-1.99597	-1.99746
R3	-2.17734	-2.17858	-0.997945	-2.17177
L0	0	0	0	0
L1	-2.0963	-2.09834	-2.09767	-0.913804
L2	10.0985	9.99411	9.98358	9.97257
L3	-2.17192	-2.17821	-0.99816	-2.1765
S0	0	0	0	0
S1	-2.09377	-2.09603	-2.09267	-0.907638
S2	-1.84318	-1.99597	-1.99913	-1.99782
S3	9.80875	9.79793	10.9796	9.8143

(Figure 2.6: Gamma = 0.1)

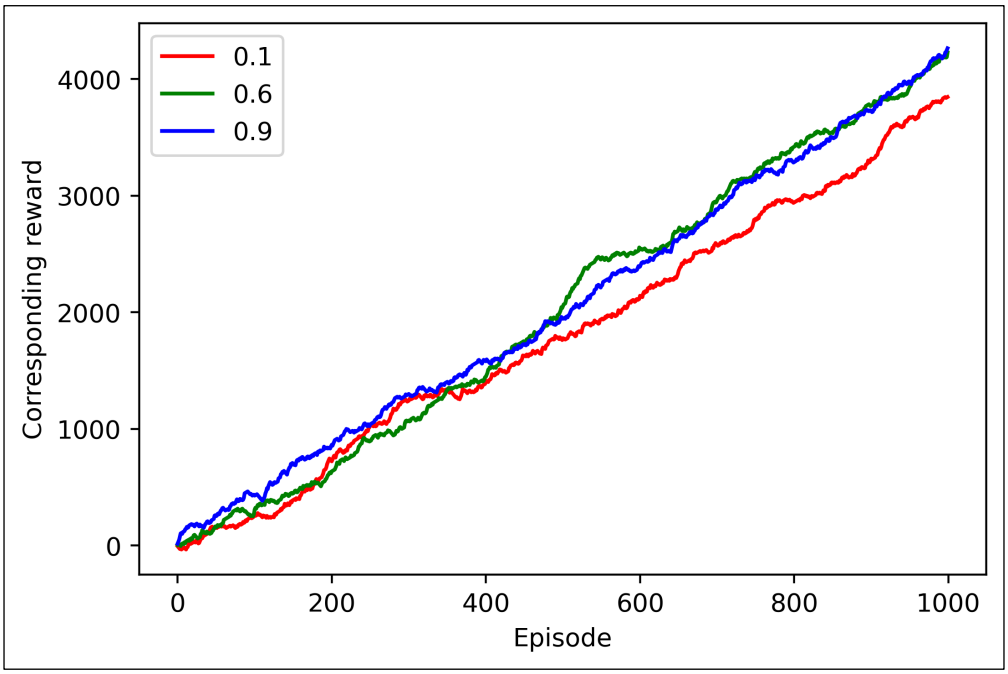

Figure 2.7

(Figure Compares total rewards received for 3 values of γ : 0.1, 0.6, 0.9)

Observations:

- If **γ**~0, the agent will be completely myopic and will only ***learn about actions that produce an immediate reward***. If **γ**~1, the agent ***learns actions that produce rewards that complement future transitions***.
- As observed in figure 2.4, for **γ** value of 0.9 the rewards are: ***High*** (Numerically) and ***Well distributed.***
- As observed in figure 2.6, for **γ** value of 0.1 the rewards are: ***Low*** (Numerically) and ***extremely biased.***
- The given path: **F0 → R1 → S3 → L2 → F0 → S3**
- Let's take the first transition **F0 → R1.** In figure 2.4 the max Q value for **F0** is 38.7 in column **R.** Which confirms with the path, since the agent is transitioning from **F0 → R1.**
- However, in figure 2.6 the max Q-value for **F0** is 1.59 in column **S**. Which is **wrong**.
- This is inaccuracy in max Q-values confirms the concept behind the discount factor. In figure 3, the agent with a **γ** value of 0.1 only learnt actions with immediate rewards. ***Thus, it failed to learn the eventual actions for future state transitions along the path.***

Learning Rate And Epsilon:

- Epsilon is a parameter used to implement the epsilon-greedy policy of q-learning.
- Simply put for **ε = 1**, the agent will ***only explore***. For **ε = 0**, the agent will ***only exploit***.
- The policy is essentially a probability distribution of the actions the agent has a propensity to take at any time state.
- Below figure 2.8 is the State transition table. Both the x and y axes are states. The numbers represent the Q-values (Added across 5000 episodes) used by the agent to transition from a ***row to a column***. Rows represent initial state, the columns represent final states.

row	F0	F1	F2	F3	R0	R1	R2	R3	L0	L1	L2	L3	S0	S1	S2	S3
F0	0	14134.7	0	0	0	15380.6	0	0	0	13919.7	0	0	0	13371.1	0	0
F1	0	0	0	3328.39	0	0	0	3468.11	0	0	0	2754.78	0	0	0	3968.41
F2	3011.05	0	0	0	−113.895	0	0	0	−119.895	0	0	0	−119.895	0	0	0
F3	0	0	2523.64	0	0	0	2965.47	0	0	0	4968.51	0	0	0	2908.37	0
R0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
R1	0	0	0	3615.95	0	0	0	3693.83	0	0	0	3698.13	0	0	0	4291.33
R2	3132.39	0	0	0	−113.895	0	0	0	−129.895	0	0	0	−129.895	0	0	0
R3	0	0	3064.42	0	0	0	2904.49	0	0	0	3581.83	0	0	0	2498.43	0
L0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
L1	0	0	0	3610.06	0	0	0	2531.43	0	0	0	2996.81	0	0	0	4317.45
L2	4584.42	0	0	0	619.474	0	0	0	699.474	0	0	0	669.474	0	0	0
L3	0	0	2465.84	0	0	0	2994.84	0	0	0	3131.24	0	0	0	3313.38	0
S0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
S1	0	0	0	3054.67	0	0	0	2758.42	0	0	0	2877.21	0	0	0	4112.44
S2	3731.76	0	0	0	−125.895	0	0	0	−125.895	0	0	0	−133.895	0	0	0
S3	0	0	3547.27	0	0	0	3502.6	0	0	0	5001.73	0	0	0	4495.65	0

Figure 2.8

(State Transition Table for the path: F0 → R1 → S3 → L2 → F0 → S3)

Observations:

- The difference in Q-values shown in the Q-table for **ε** values ranging from 0.05 to 0.95 was ***minuscule***.
- The reason is that, the difference in Q-values for exploring and exploiting ***further accentuates, when there are a plethora of actions possible.***
- Imagine we have 100 actions possible. For an **ε** value of 0.1, the distribution would be extremely biased towards one action and less towards the rest 99. ***This situation would accurately illustrate the sheer difference in outcomes in an agent trying to explore and exploit. Because of the sheer number of actions possible, the agent will take a very long time figuring out the path if it exploits.***
- ***In leman's terms, the agent has an extremely high propensity to get biased and not recover from it, if there are a plethora of actions.***
- However, in our environment, the difference is barely noticed due to there only being 4 possible actions. Thus, even if **ε = 0.05** (Propensity to exploit). The action probability distributions would look like [0.05 0.05 0.85 0.05]. Which is a very limited number of actions to accentuate the difference in outcomes of exploring and exploiting.
- Lastly, figure 2 illustrates the state transition table to give us a better idea of the transitions. The numbers represent Q-values.
-