Project2 COSC 6368 Fall 2017

Learning Paths from Feedback Using Q-Learning and SARSA

1. Problem:

In this Project we will use q-learning/SARSA for the PD-World conducting 3 experiments using different parameters and policies, and summarize and interpret the experimental results.

2. Problem Statement:

1) In Experiment 1 you use α=0.3, and run the Q-learning algorithm for 3000 steps with policy PRANDOM; then run PGreedy for 3000 steps.
2) In Experiment 2 you use α=0.3, and run the Q-learning algorithm for 6000 steps with policy PEXPLOIT—however, we use policy PRANDOM for the first 200 steps of the experiment, and then switch to PEXPLOIT for the remainder of the experiment.
3) In Experiment 3 you use α=0.3, and run the SARSA q-learning variation for 6000 steps with policy PEXPLOIT—however, use policy PRANDOM for the first 200 steps of the experiment, and then switch to PEXPLOIT for the remainder of the experiment.
4) Visualizing the Q-Tables thus obtained from the above experiments.

3. Implemented Policies To Run the PD-World:

a) PRANDOM:

If pickup and drop-off is applicable, choose this operator; otherwise, choose an applicable operator randomly.

b) PEXPLOIT:

If pickup and drop-off is applicable, choose this operator; otherwise, apply the applicable operator with the highest q-value (break ties by rolling a dice for operators with the same q-value) with probability 0.85 and choose a different applicable operator randomly with probability 0.15.

c) PGREEDY:

If pickup and drop-off is applicable, choose this operator; otherwise, apply the applicable operator with the highest q-value (break ties by rolling a dice for operators with the same q-value).

4. State Space Used:

1) In this project, we take the reduced Reinforcement Learning state space which is
2) (i, j, block) where (i, j) is the position of the agent in the world and block is a Boolean variable which keeps track of whether the agent is carrying a block or not.
3) Here, we start the experiments by creating a world of total 25 locations, each of which contain 6 different actions with 4 different directions to take.

4) The agent will initially be placed at the 5th location.
5) Out of the 25 locations, 4 locations are pickup locations, i.e., that locations contain some blocks which has to be picked up, 2 locations are drop-off locations.
6) Each pickup location consists of 4 blocks.
7) Each drop-off location can accommodate 8 blocks.
8) The aim of the experiment is to pickup all the blocks from all pickup locations and drop off them at the drop-off locations.
9) If all the blocks at all pickup locations are empty and all the drop-off locations are full with the blocks, we reset the world, i.e., all the pickup locations are filled with 4 blocks each and all drop-off locations are made empty and the agent is placed at the 5th location.
10) We take two Q-Tables to maintain the Q values obtained, i.e., we update one Q-Table when the agent is carrying a block and the other one when the agent is not carrying a block. This is done in order to benefit when the greedy policy is applied, i.e., when the agent is not carrying a block, it is probably looking for a pickup location (that is when it looks for a Q-Table which is updated when agent is not carrying a block), alternately when the agent is carrying a block, it is looking for a drop-off location (that is when it looks for a Q-Table which is updated when agent is carrying a block).
11) We track the number of blocks in the pickup and drop-off locations by maintaining a two hashmaps (one for pickup and the other for drop-off) with location numbers a keys and the number of blocks in that location as values.
12) We keep track of the successor locations that can be reached from a given locations when applied a given action by using a hashmap with "successorsOf_" as a key and the corresponding states in an integer array as the value.

5. Performance measures:

I. The total reward.
II. The total reward after the agent takes a single step.
III. The Q-Tables at different instances like:
    - Display Q-Table after 3000 steps.
    - Display Q-Table after 6000 steps.

6. Experiment 1 Analysis:

I. In this experiment, we have to apply **PRANDOM and PGREEDY** each for 3000 steps in the same order. We use QLearning algorithm to calculate the QValues in this experiment.
II. We apply the **PRANDOM** first in order to populate the initial QValues which are required to apply the PGREEDY policy. If not, all the QValues in the QTable will be initially zeros (0.0).
III. Then, we display the QTables after the 3000 steps which is shown in Fig 1.
IV. Now, we apply PGREEDY policy for 3000 steps. The QTables after the total 6000 steps is displayed, which is shown in Fig 2.
V. While the agent takes each step, we record the total reward value up to that step and present that in a graph shown in Fig 3.

VI. Fig 4 shows the comparison between 2 different runs after 3000 and 6000 steps respectively.

QTable1 after 3000 .

			east	west	north	south	pickup	dropoff
1	1	blockPickedUp	-1.6327	N/A	N/A	-0.5334	0.0000	N/A
1	2	blockPickedUp	-1.4133	-1.2659	N/A	-1.2673	N/A	N/A
1	3	blockPickedUp	-1.1957	-1.6329	N/A	-0.8412	N/A	N/A
1	4	blockPickedUp	-1.4552	-1.4318	N/A	-0.6938	N/A	N/A
1	5	blockPickedUp	N/A	-1.2141	N/A	-0.8837	N/A	N/A
2	1	blockPickedUp	-1.2671	N/A	-1.2662	0.9263	N/A	N/A
2	2	blockPickedUp	-0.8515	-0.5326	-1.6337	-0.5358	N/A	N/A
2	3	blockPickedUp	-0.5861	-1.2670	-1.4147	-0.1645	N/A	N/A
2	4	blockPickedUp	-1.0586	-0.9453	-1.2076	-0.1274	N/A	N/A
2	5	blockPickedUp	N/A	-0.6445	-1.4714	-0.3188	N/A	N/A
3	1	blockPickedUp	-0.5392	N/A	-0.5351	3.8339	N/A	N/A
3	2	blockPickedUp	0.2385	0.9339	-1.2686	0.8747	N/A	N/A
3	3	blockPickedUp	-0.0629	-0.5335	-0.8644	0.8841	0.0000	N/A
3	4	blockPickedUp	0.1368	-0.1613	-0.8014	0.1099	N/A	N/A
3	5	blockPickedUp	N/A	-0.1788	-1.0214	1.0883	N/A	N/A
4	1	blockPickedUp	0.8744	N/A	0.9150	9.3526	0.0000	N/A
4	2	blockPickedUp	1.6212	3.7259	-0.5348	3.4920	N/A	N/A
4	3	blockPickedUp	1.0018	0.8462	-0.0080	0.6850	N/A	N/A
4	4	blockPickedUp	0.9481	0.4730	-0.3675	0.4410	N/A	19.6907
4	5	blockPickedUp	N/A	1.7820	0.0141	0.1734	N/A	N/A
5	1	blockPickedUp	0.0000	N/A	0.0000	N/A	N/A	21.5771
5	2	blockPickedUp	0.6608	9.3452	0.7942	N/A	N/A	N/A
5	3	blockPickedUp	2.5274	3.5232	1.1657	N/A	N/A	N/A
5	4	blockPickedUp	0.1571	0.6572	0.8067	N/A	N/A	N/A
5	5	blockPickedUp	N/A	1.5659	1.7637	N/A	0.0000	N/A

QTable1 after 6000 .

			east	west	north	south	pickup	dropoff
1	1	blockPickedUp	-1.6330	N/A	N/A	-1.6483	0.0000	N/A
1	2	blockPickedUp	-1.4133	-1.4311	N/A	-1.3465	N/A	N/A
1	3	blockPickedUp	-1.1957	-1.6329	N/A	-0.8412	N/A	N/A
1	4	blockPickedUp	-1.4552	-1.4318	N/A	-0.6938	N/A	N/A
1	5	blockPickedUp	N/A	-1.2141	N/A	-0.8837	N/A	N/A
2	1	blockPickedUp	-1.4146	N/A	-1.4099	-1.4055	N/A	N/A
2	2	blockPickedUp	-1.1194	-1.1057	-1.6337	-1.0731	N/A	N/A
2	3	blockPickedUp	-0.9935	-1.2670	-1.4147	-0.2378	N/A	N/A
2	4	blockPickedUp	-1.0586	-1.0835	-1.2076	-0.9982	N/A	N/A
2	5	blockPickedUp	N/A	-0.6445	-1.4714	-0.3188	N/A	N/A
3	1	blockPickedUp	-1.2108	N/A	-1.2710	-1.1936	N/A	N/A
3	2	blockPickedUp	-0.9598	-1.0595	-1.2686	-0.8615	N/A	N/A
3	3	blockPickedUp	-1.1053	-1.1016	-1.0787	3.0685	0.0000	N/A
3	4	blockPickedUp	-0.9647	-1.0551	-1.0354	2.2690	N/A	N/A
3	5	blockPickedUp	N/A	-0.9976	-1.0214	-0.8484	N/A	N/A
4	1	blockPickedUp	-1.1892	N/A	-1.1960	6.7274	0.0000	N/A
4	2	blockPickedUp	1.3563	-0.9980	-0.9880	-0.9557	N/A	N/A
4	3	blockPickedUp	1.1582	-0.8946	-0.9945	-0.8770	N/A	N/A
4	4	blockPickedUp	-0.9221	0.6440	-0.9955	-1.0556	N/A	19.1861
4	5	blockPickedUp	N/A	3.6370	-0.8811	-0.7924	N/A	N/A
5	1	blockPickedUp	-1.0862	N/A	-1.0418	N/A	N/A	21.5027
5	2	blockPickedUp	-0.8519	-0.7875	-0.9741	N/A	N/A	N/A
5	3	blockPickedUp	-0.6760	-0.6407	-0.8222	N/A	N/A	N/A
5	4	blockPickedUp	-1.0887	-1.0421	-0.9733	N/A	N/A	N/A
5	5	blockPickedUp	N/A	-1.1599	0.4435	N/A	0.0000	N/A

Fig 1: QTable1 instance at after 3000 and 6000 steps for experiment 1

QTable2 after 3000 .

			east	west	north	south	pickup	dropoff
1	1	blockNotPickedUp	-1.8363	N/A	N/A	-1.8188	19.8303	N/A
1	2	blockNotPickedUp	-0.8383	8.1630	N/A	-0.1634	N/A	N/A
1	3	blockNotPickedUp	-1.1210	2.7965	N/A	-0.8255	N/A	N/A
1	4	blockNotPickedUp	-1.5606	0.2611	N/A	-1.2846	N/A	N/A
1	5	blockNotPickedUp	N/A	-0.9936	N/A	-1.6402	N/A	N/A
2	1	blockNotPickedUp	-0.0996	N/A	6.9546	-0.5918	N/A	N/A
2	2	blockNotPickedUp	-0.2975	1.2954	2.3419	-0.8433	N/A	N/A
2	3	blockNotPickedUp	-1.1879	-1.1123	-0.1746	2.8746	N/A	N/A
2	4	blockNotPickedUp	-1.6545	-0.8363	-0.9716	-0.2090	N/A	N/A
2	5	blockNotPickedUp	N/A	-1.3417	-1.7322	-1.1108	N/A	N/A
3	1	blockNotPickedUp	-1.1858	N/A	1.7881	0.3349	N/A	N/A
3	2	blockNotPickedUp	-0.2426	-0.5222	-0.8287	-0.8050	N/A	N/A
3	3	blockNotPickedUp	0.1936	-0.9342	-0.4728	-1.3585	17.6139	N/A
3	4	blockNotPickedUp	-1.1957	3.9762	-1.2330	-0.6775	N/A	N/A
3	5	blockNotPickedUp	N/A	0.4732	-1.6538	-0.1525	N/A	N/A
4	1	blockNotPickedUp	-0.4292	N/A	-0.4684	0.3828	17.6751	N/A
4	2	blockNotPickedUp	0.2132	0.4806	-0.8909	-1.0613	N/A	N/A
4	3	blockNotPickedUp	-0.8691	-0.3163	3.0594	-1.4186	N/A	N/A
4	4	blockNotPickedUp	-0.3514	0.1541	1.1830	-0.1278	N/A	0.0000
4	5	blockNotPickedUp	N/A	-0.9305	-1.1801	0.7607	N/A	N/A
5	1	blockNotPickedUp	-0.9955	N/A	0.1005	N/A	N/A	0.0000
5	2	blockNotPickedUp	-1.3704	0.0338	-0.4157	N/A	N/A	N/A
5	3	blockNotPickedUp	-0.6900	-1.1163	-0.6121	N/A	N/A	N/A
5	4	blockNotPickedUp	3.7925	-1.4218	-0.8864	N/A	N/A	N/A
5	5	blockNotPickedUp	N/A	-0.9465	-0.3632	N/A	19.2334	N/A

QTable2 after 6000 .

			east	west	north	south	pickup	dropoff
1	1	blockNotPickedUp	-1.8363	N/A	N/A	-1.8203	17.2372	N/A
1	2	blockNotPickedUp	-1.7896	-1.7589	N/A	-1.7720	N/A	N/A
1	3	blockNotPickedUp	-1.7947	-1.7752	N/A	-1.7583	N/A	N/A
1	4	blockNotPickedUp	-1.8016	-1.7801	N/A	-1.8134	N/A	N/A
1	5	blockNotPickedUp	N/A	-1.7962	N/A	-1.8147	N/A	N/A
2	1	blockNotPickedUp	-1.7914	N/A	-1.8091	-1.8121	N/A	N/A
2	2	blockNotPickedUp	-1.7634	-1.7769	-1.7963	-1.7813	N/A	N/A
2	3	blockNotPickedUp	-1.7622	-1.7745	-1.7921	-1.7846	N/A	N/A
2	4	blockNotPickedUp	-1.8019	-1.8072	-1.7904	-1.8240	N/A	N/A
2	5	blockNotPickedUp	N/A	-1.7980	-1.8038	-1.8288	N/A	N/A
3	1	blockNotPickedUp	-1.7896	N/A	-1.8114	-1.8183	N/A	N/A
3	2	blockNotPickedUp	-1.7713	-1.7690	-1.7670	-1.7790	N/A	N/A
3	3	blockNotPickedUp	-1.8079	-1.7944	-1.8144	-1.7897	17.2506	N/A
3	4	blockNotPickedUp	-1.8097	-1.8095	-1.8180	-1.8353	N/A	N/A
3	5	blockNotPickedUp	N/A	-1.8309	-1.8114	-1.8161	N/A	N/A
4	1	blockNotPickedUp	-1.8269	N/A	-1.8063	-1.8292	17.3146	N/A
4	2	blockNotPickedUp	-1.8010	-1.7747	-1.7845	-1.7940	N/A	N/A
4	3	blockNotPickedUp	-1.7972	-1.7960	-1.8047	-1.8121	N/A	N/A
4	4	blockNotPickedUp	-1.8160	-1.8304	-1.8259	-1.8110	N/A	0.0000
4	5	blockNotPickedUp	N/A	-1.8186	-1.8267	-1.7996	N/A	N/A
5	1	blockNotPickedUp	-1.8162	N/A	-1.8396	N/A	N/A	0.0000
5	2	blockNotPickedUp	-1.7788	-1.8039	-1.8029	N/A	N/A	N/A
5	3	blockNotPickedUp	-1.7989	-1.7832	-1.7677	N/A	N/A	N/A
5	4	blockNotPickedUp	1.0492	-1.7766	-1.8029	N/A	N/A	N/A
5	5	blockNotPickedUp	N/A	-1.7920	-1.7823	N/A	18.2761	N/A

Fig 2: QTable2 instance at after 3000 and 6000 steps for experiment 1

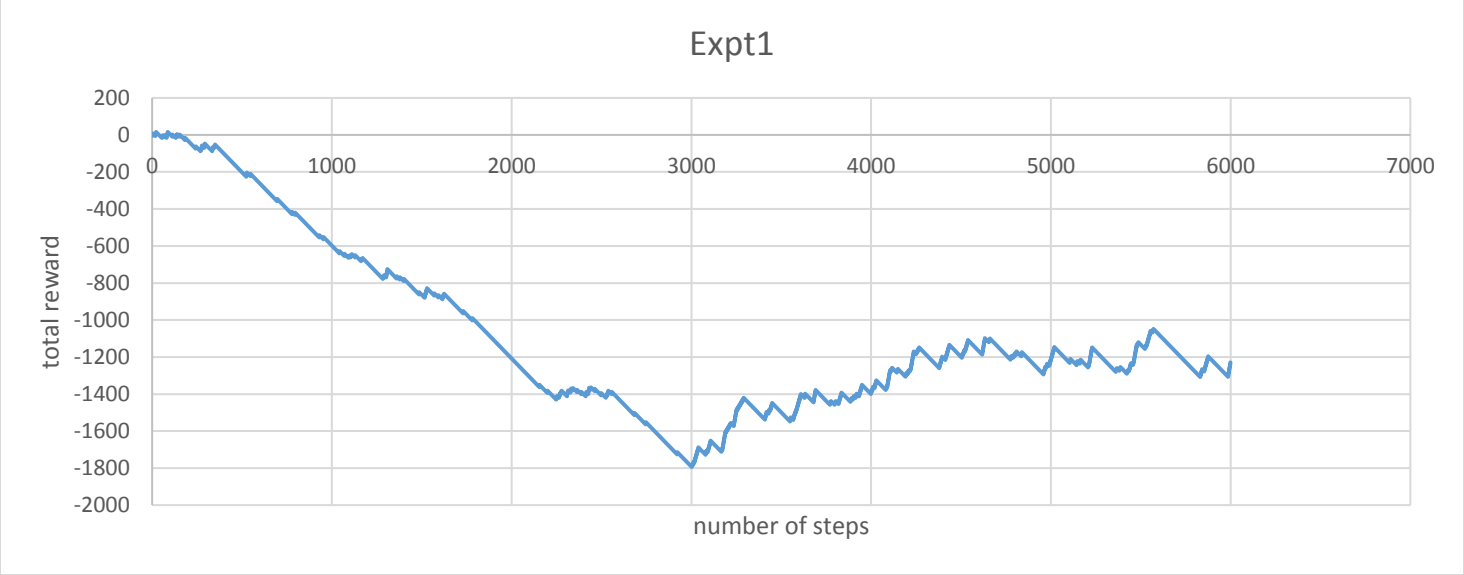

Fig 3: total reward for each step for experiment 1

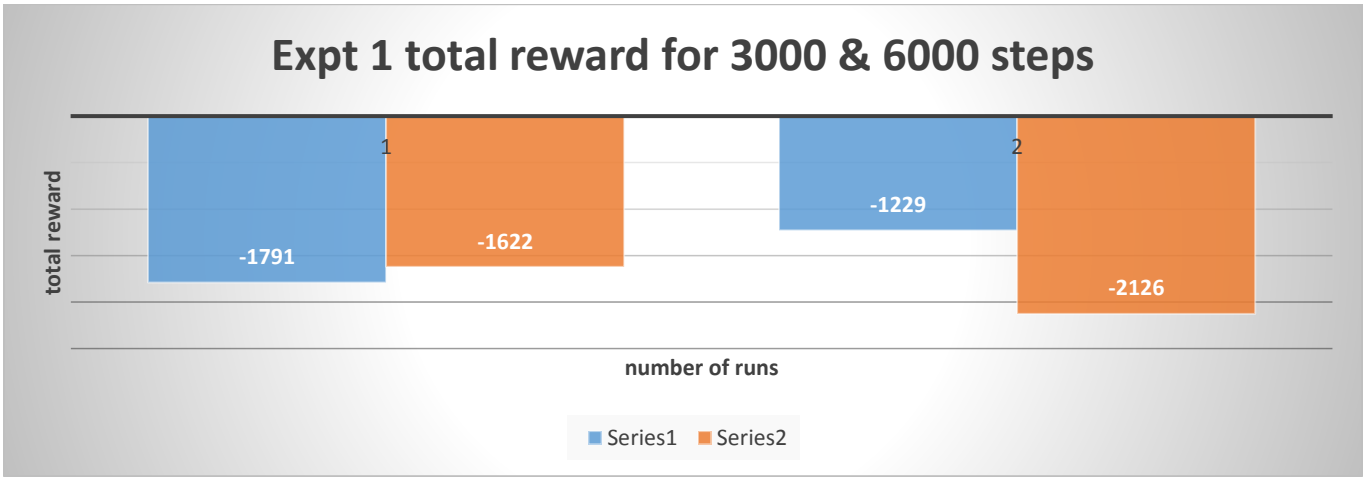

Fig 4: comparison of total reward after 6000 steps for 2 different runs of experiment 1

The next paragraph contains the visualization of the QTable. (White squares indicate normal states with default application actions viz. East (E), West (W), North (N), and South (S). Green squares indicate pickup/drop-off locations with default application actions and additional pickup (Pickup) and drop-off (Drop-off). The QValues are indicated with the action prepended. For pickup/drop-off locations the values displayed inside the green square. The black line indicate that the invalid direction from that cell. “N/A” indicate “Not Applicable”.

If we consider the Fig 5 and Fig 6, we observe the following:

a. The states which are around the pickup state or drop-off state, the QValue for that action which results into the pickup or drop-off will be the highest among the other action’s QValues in the same location.
   This observation states that, when the agent gets some experience with the world, it will be able to recognize the pickup and drop-off states when it is adjacent to it based on those QValues.
b. From the drop-off position’s QValues, we can infer that both the drop-off locations are equally explored by the agent.

c. From the pickup position's QValues, we can infer that all the pickup locations are equally explored by the agent.
d. We can see that after applying PGREEDY policy makes the agent reach the terminal state more times, i.e., agent does more number of pickups and drop-offs when compared to that of made by it during PRANDOM.

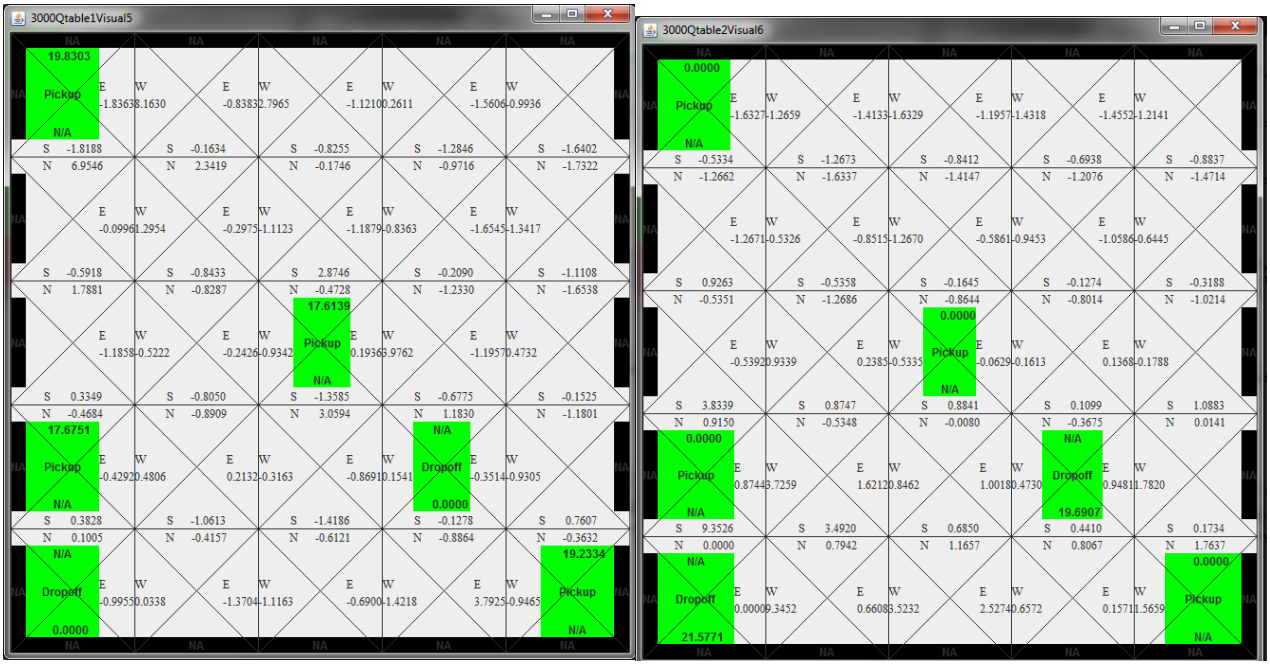

Fig 5: QTable1 & QTable2 after 3000 steps for experiment 1

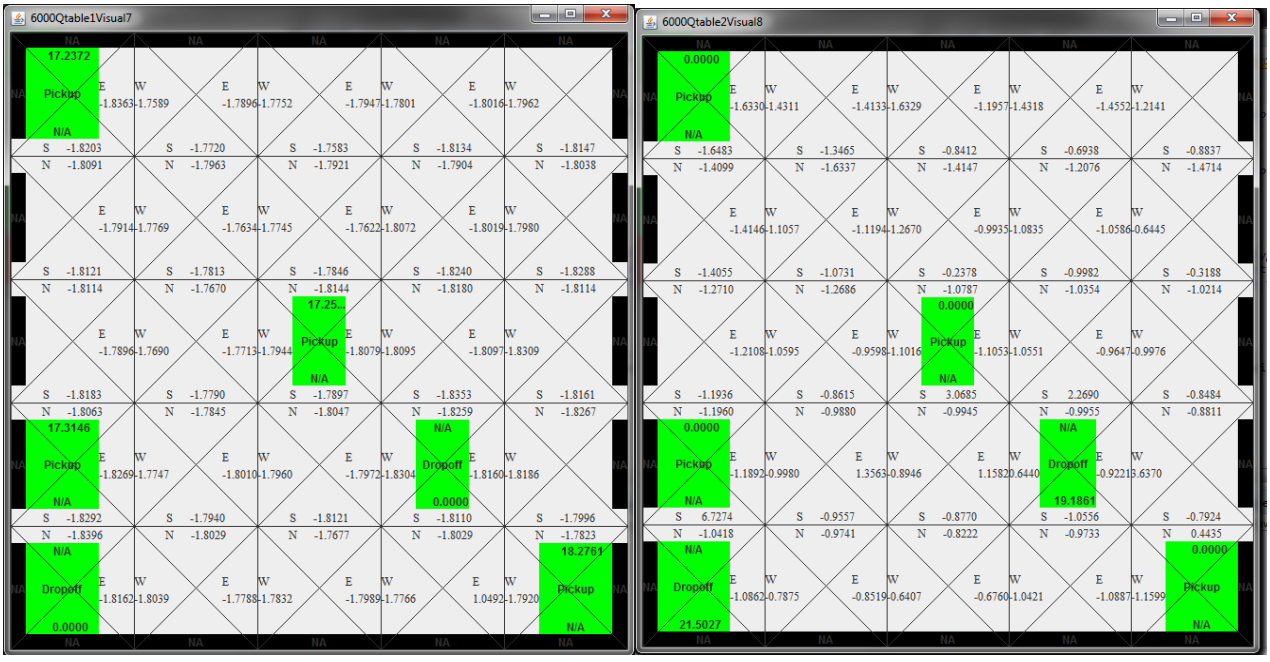

Fig 6: QTable1 & QTable2 after 6000 steps for experiment 1

In conclusion, Experiment 1 exposes random behavior with respect to the agent. The agent got lucky sometimes, reaping huge benefits in the former stages itself. In other words,

there was no intension the agent could display while learning attractive paths. The rate at which the agent was learning (alpha) helped it a lot in this experiment.

7. Experiment 2 Analysis:

a. In this experiment, we have to apply **PRANDOM** for first 200 steps and **PEXPLOIT** for each of the remaining steps. We use QLearning algorithm to calculate the QValues in this experiment.
b. We apply the **PRANDOM** first in order to populate the initial QValues which are required to apply the PEXPLOIT policy. If not, all the QValues in the QTable will be initially zeros (0.0).
c. Then, we display the QTables after the 3000 steps which is shown in Fig 9.
d. Now, we apply PEXPLOIT policy for 3000 steps. The QTables after the total 6000 steps is displayed, which is shown in Fig 10.
e. While the agent takes each step, we record the total reward value up to that step and present that in a graph shown in Fig 11.
f. Fig 12 shows the comparison between 2 different runs after 3000 and 6000 steps respectively.

If we consider the Fig 7 and Fig 8, we observe the following:

a. The states which are around the pickup state or drop-off state, the QValue for that action which results into the pickup or drop-off will be the highest among the other action's QValues in the same location.
   This observation states that, when the agent gets some experience with the world, it will be able to recognize the pickup and drop-off states when it is adjacent to it based on those QValues.
b. From the drop-off position's QValues, we can infer that all the drop-off locations are equally explored.
c. From the pickup position's QValues, we can infer that all the pickup locations are equally explored by the agent.
d. We can see that after applying PEXPLOIT policy, makes the agent reach the terminal state more times, i.e., agent does more number of pickups and drop-offs when compared to that of made by it during PRANDOM + PGREEDY.

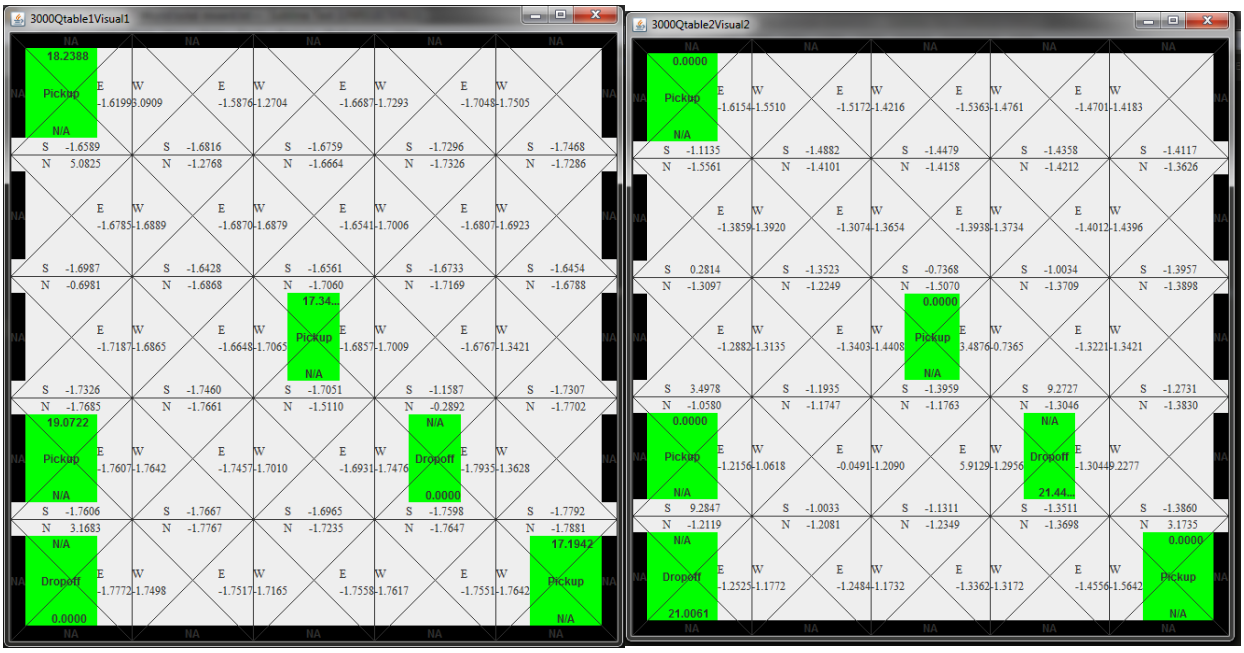

Fig 7: QTable1 & QTable2 after 3000 steps for experiment 2

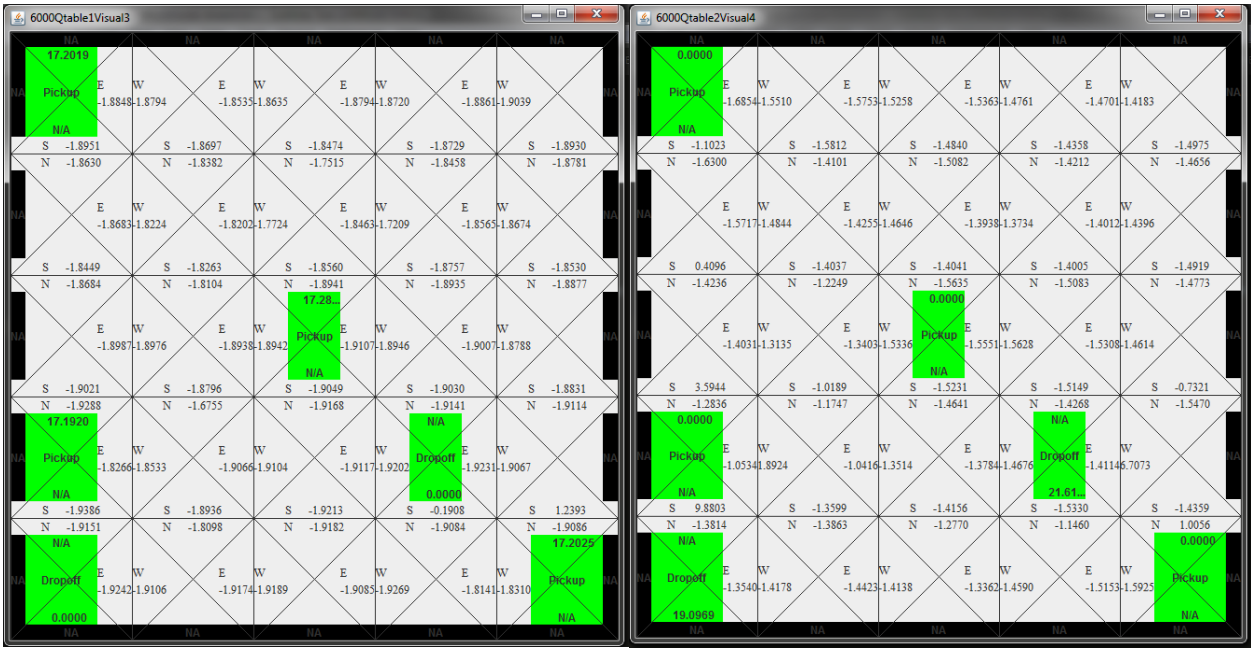

Fig 8: QTable1 & QTable2 after 6000 steps for experiment 2

QTable1 after 3000 .

			east	west	north	south	pickup	dropoff
1	1	blockPickedUp	-1.6154	N/A	N/A	-1.1135	0.0000	N/A
1	2	blockPickedUp	-1.5172	-1.5510	N/A	-1.4882	N/A	N/A
1	3	blockPickedUp	-1.5363	-1.4216	N/A	-1.4479	N/A	N/A
1	4	blockPickedUp	-1.4701	-1.4761	N/A	-1.4358	N/A	N/A
1	5	blockPickedUp	N/A	-1.4183	N/A	-1.4117	N/A	N/A
2	1	blockPickedUp	-1.3859	N/A	-1.5561	0.2814	N/A	N/A
2	2	blockPickedUp	-1.3074	-1.3920	-1.4101	-1.3523	N/A	N/A
2	3	blockPickedUp	-1.3938	-1.3654	-1.4158	-0.7368	N/A	N/A
2	4	blockPickedUp	-1.4012	-1.3734	-1.4212	-1.0034	N/A	N/A
2	5	blockPickedUp	N/A	-1.4396	-1.3626	-1.3957	N/A	N/A
3	1	blockPickedUp	-1.2882	N/A	-1.3097	3.4978	N/A	N/A
3	2	blockPickedUp	-1.3403	-1.3135	-1.2249	-1.1935	N/A	N/A
3	3	blockPickedUp	3.4876	-1.4408	-1.5070	-1.3959	0.0000	N/A
3	4	blockPickedUp	-1.3221	-0.7365	-1.3709	9.2727	N/A	N/A
3	5	blockPickedUp	N/A	-1.3421	-1.3898	-1.2731	N/A	N/A
4	1	blockPickedUp	-1.2156	N/A	-1.0580	9.2847	0.0000	N/A
4	2	blockPickedUp	-0.0491	-1.0618	-1.1747	-1.0033	N/A	N/A
4	3	blockPickedUp	5.9129	-1.2090	-1.1763	-1.1311	N/A	N/A
4	4	blockPickedUp	-1.3044	-1.2956	-1.3046	-1.3511	N/A	21.4459
4	5	blockPickedUp	N/A	9.2277	-1.3830	-1.3860	N/A	N/A
5	1	blockPickedUp	-1.2525	N/A	-1.2119	N/A	N/A	21.0061
5	2	blockPickedUp	-1.2484	-1.1772	-1.2081	N/A	N/A	N/A
5	3	blockPickedUp	-1.3362	-1.1732	-1.2349	N/A	N/A	N/A
5	4	blockPickedUp	-1.4556	-1.3172	-1.3698	N/A	N/A	N/A
5	5	blockPickedUp	N/A	-1.5642	3.1735	N/A	0.0000	N/A

QTable2 after 3000 .

			east	west	north	south	pickup	dropoff
1	1	blockNotPickedUp	-1.6199	N/A	N/A	-1.6589	18.2388	N/A
1	2	blockNotPickedUp	-1.5876	3.0909	N/A	-1.6816	N/A	N/A
1	3	blockNotPickedUp	-1.6687	-1.2704	N/A	-1.6759	N/A	N/A
1	4	blockNotPickedUp	-1.7048	-1.7293	N/A	-1.7296	N/A	N/A
1	5	blockNotPickedUp	N/A	-1.7505	N/A	-1.7468	N/A	N/A
2	1	blockNotPickedUp	-1.6785	N/A	5.0825	-1.6987	N/A	N/A
2	2	blockNotPickedUp	-1.6870	-1.6889	-1.2768	-1.6428	N/A	N/A
2	3	blockNotPickedUp	-1.6541	-1.6879	-1.6664	-1.6561	N/A	N/A
2	4	blockNotPickedUp	-1.6807	-1.7006	-1.7326	-1.6733	N/A	N/A
2	5	blockNotPickedUp	N/A	-1.6923	-1.7286	-1.6454	N/A	N/A
3	1	blockNotPickedUp	-1.7187	N/A	-0.6981	-1.7326	N/A	N/A
3	2	blockNotPickedUp	-1.6648	-1.6865	-1.6868	-1.7460	N/A	N/A
3	3	blockNotPickedUp	-1.6857	-1.7065	-1.7060	-1.7051	17.3412	N/A
3	4	blockNotPickedUp	-1.6767	-1.7009	-1.7169	-1.1587	N/A	N/A
3	5	blockNotPickedUp	N/A	-1.3421	-1.6788	-1.7307	N/A	N/A
4	1	blockNotPickedUp	-1.7607	N/A	-1.7685	-1.7606	19.0722	N/A
4	2	blockNotPickedUp	-1.7457	-1.7642	-1.7661	-1.7667	N/A	N/A
4	3	blockNotPickedUp	-1.6931	-1.7010	-1.5110	-1.6965	N/A	N/A
4	4	blockNotPickedUp	-1.7935	-1.7476	-0.2892	-1.7598	N/A	0.0000
4	5	blockNotPickedUp	N/A	-1.3628	-1.7702	-1.7792	N/A	N/A
5	1	blockNotPickedUp	-1.7772	N/A	3.1683	N/A	N/A	0.0000
5	2	blockNotPickedUp	-1.7517	-1.7498	-1.7767	N/A	N/A	N/A
5	3	blockNotPickedUp	-1.7558	-1.7165	-1.7235	N/A	N/A	N/A
5	4	blockNotPickedUp	-1.7551	-1.7617	-1.7647	N/A	N/A	N/A
5	5	blockNotPickedUp	N/A	-1.7642	-1.7881	N/A	17.1942	N/A

Fig 9: QTable1 instance at after 3000 and 6000 steps for experiment 2

QTable1 after 6000 .

			east	west	north	south	pickup	dropoff
1	1	blockPickedUp	-1.6854	N/A	N/A	-1.1023	0.0000	N/A
1	2	blockPickedUp	-1.5753	-1.5510	N/A	-1.5812	N/A	N/A
1	3	blockPickedUp	-1.5363	-1.5258	N/A	-1.4840	N/A	N/A
1	4	blockPickedUp	-1.4701	-1.4761	N/A	-1.4358	N/A	N/A
1	5	blockPickedUp	N/A	-1.4183	N/A	-1.4975	N/A	N/A
2	1	blockPickedUp	-1.5717	N/A	-1.6300	0.4096	N/A	N/A
2	2	blockPickedUp	-1.4255	-1.4844	-1.4101	-1.4037	N/A	N/A
2	3	blockPickedUp	-1.3938	-1.4646	-1.5082	-1.4041	N/A	N/A
2	4	blockPickedUp	-1.4012	-1.3734	-1.4212	-1.4005	N/A	N/A
2	5	blockPickedUp	N/A	-1.4396	-1.4656	-1.4919	N/A	N/A
3	1	blockPickedUp	-1.4031	N/A	-1.4236	3.5944	N/A	N/A
3	2	blockPickedUp	-1.3403	-1.3135	-1.2249	-1.0189	N/A	N/A
3	3	blockPickedUp	-1.5551	-1.5336	-1.5635	-1.5231	0.0000	N/A
3	4	blockPickedUp	-1.5308	-1.5628	-1.5083	-1.5149	N/A	N/A
3	5	blockPickedUp	N/A	-1.4614	-1.4773	-0.7321	N/A	N/A
4	1	blockPickedUp	-1.0534	N/A	-1.2836	9.8803	0.0000	N/A
4	2	blockPickedUp	-1.0416	1.8924	-1.1747	-1.3599	N/A	N/A
4	3	blockPickedUp	-1.3784	-1.3514	-1.4641	-1.4156	N/A	N/A
4	4	blockPickedUp	-1.4114	-1.4676	-1.4268	-1.5330	N/A	21.6147
4	5	blockPickedUp	N/A	6.7073	-1.5470	-1.4359	N/A	N/A
5	1	blockPickedUp	-1.3540	N/A	-1.3814	N/A	N/A	19.0969
5	2	blockPickedUp	-1.4423	-1.4178	-1.3863	N/A	N/A	N/A
5	3	blockPickedUp	-1.3362	-1.4138	-1.2770	N/A	N/A	N/A
5	4	blockPickedUp	-1.5153	-1.4590	-1.1460	N/A	N/A	N/A
5	5	blockPickedUp	N/A	-1.5925	1.0056	N/A	0.0000	N/A

QTable2 after 6000 .

			east	west	north	south	pickup	dropoff
1	1	blockNotPickedUp	-1.8848	N/A	N/A	-1.8951	17.2019	N/A
1	2	blockNotPickedUp	-1.8535	-1.8794	N/A	-1.8697	N/A	N/A
1	3	blockNotPickedUp	-1.8794	-1.8635	N/A	-1.8474	N/A	N/A
1	4	blockNotPickedUp	-1.8861	-1.8720	N/A	-1.8729	N/A	N/A
1	5	blockNotPickedUp	N/A	-1.9039	N/A	-1.8930	N/A	N/A
2	1	blockNotPickedUp	-1.8683	N/A	-1.8630	-1.8449	N/A	N/A
2	2	blockNotPickedUp	-1.8202	-1.8224	-1.8382	-1.8263	N/A	N/A
2	3	blockNotPickedUp	-1.8463	-1.7724	-1.7515	-1.8560	N/A	N/A
2	4	blockNotPickedUp	-1.8565	-1.7209	-1.8458	-1.8757	N/A	N/A
2	5	blockNotPickedUp	N/A	-1.8674	-1.8781	-1.8530	N/A	N/A
3	1	blockNotPickedUp	-1.8987	N/A	-1.8684	-1.9021	N/A	N/A
3	2	blockNotPickedUp	-1.8938	-1.8976	-1.8104	-1.8796	N/A	N/A
3	3	blockNotPickedUp	-1.9107	-1.8942	-1.8941	-1.9049	17.2831	N/A
3	4	blockNotPickedUp	-1.9007	-1.8946	-1.8935	-1.9030	N/A	N/A
3	5	blockNotPickedUp	N/A	-1.8788	-1.8877	-1.8831	N/A	N/A
4	1	blockNotPickedUp	-1.8266	N/A	-1.9288	-1.9386	17.1920	N/A
4	2	blockNotPickedUp	-1.9066	-1.8533	-1.6755	-1.8936	N/A	N/A
4	3	blockNotPickedUp	-1.9117	-1.9104	-1.9168	-1.9213	N/A	N/A
4	4	blockNotPickedUp	-1.9231	-1.9202	-1.9141	-0.1908	N/A	0.0000
4	5	blockNotPickedUp	N/A	-1.9067	-1.9114	1.2393	N/A	N/A
5	1	blockNotPickedUp	-1.9242	N/A	-1.9151	N/A	N/A	0.0000
5	2	blockNotPickedUp	-1.9174	-1.9106	-1.8098	N/A	N/A	N/A
5	3	blockNotPickedUp	-1.9085	-1.9189	-1.9182	N/A	N/A	N/A
5	4	blockNotPickedUp	-1.8141	-1.9269	-1.9084	N/A	N/A	N/A
5	5	blockNotPickedUp	N/A	-1.8310	-1.9086	N/A	17.2025	N/A

Fig 10: QTable2 instance at after 3000 and 6000 steps for experiment 2

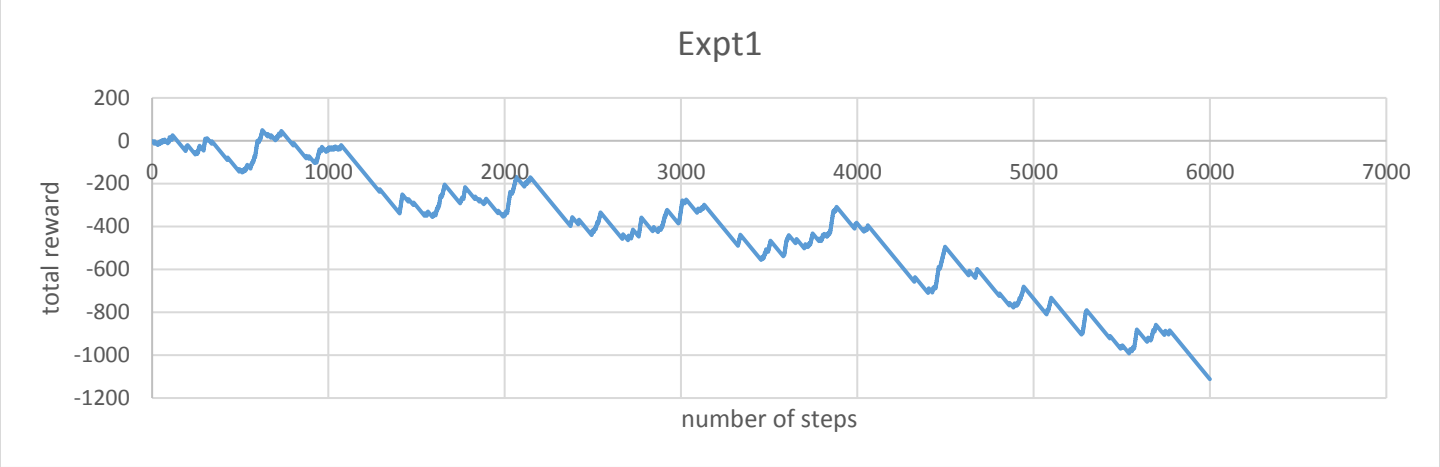

Fig 11: total reward for each step for experiment 2

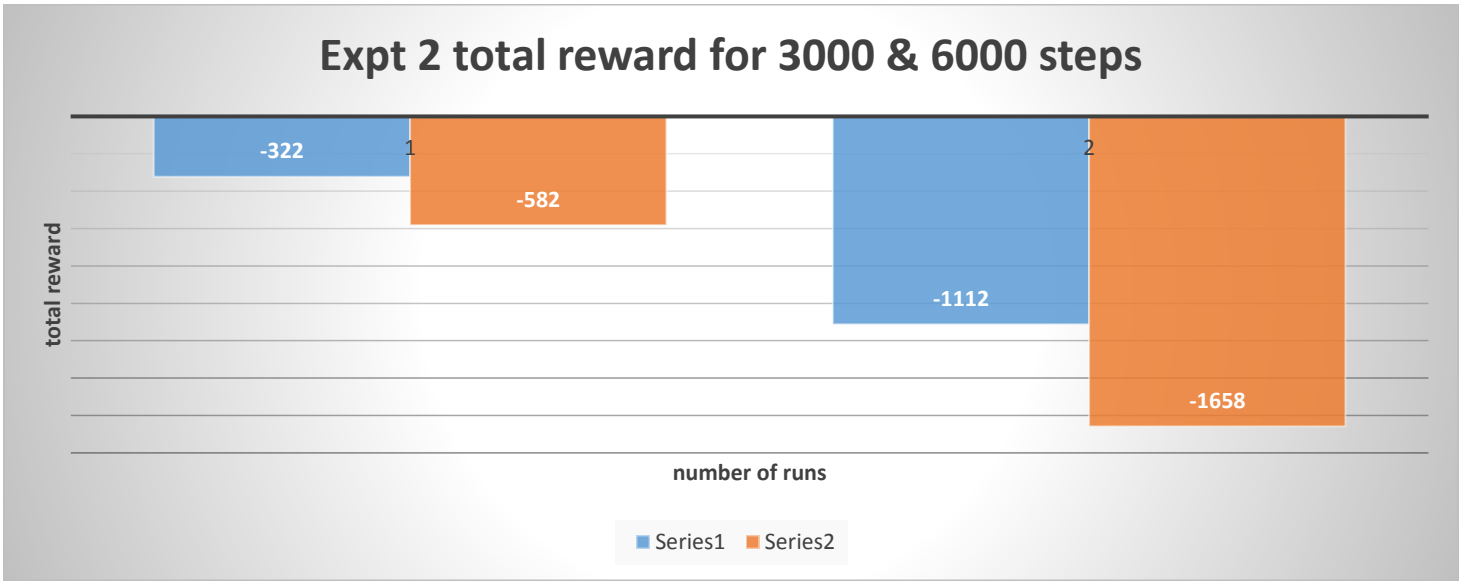

Fig 12: comparison of total reward after 3000 & 6000 steps for 2 different runs of experiment 2

8. Experiment 3 analysis:

a. In this experiment, we have to apply **PRANDOM** for first 200 steps and **PEXPLOIT** for each of the remaining steps. We use SARSA algorithm to calculate the QValues in this experiment.
b. We apply the **PRANDOM** first in order to populate the initial QValues which are required to apply the PEXPLOIT policy. If not, all the QValues in the QTable will be initially zeros (0.0).
c. Then, we display the QTables after the 3000 steps which is shown in Fig 13.
d. Now, we apply PEXPLOIT policy for 3000 steps. The QTables after the total 6000 steps is displayed, which is shown in Fig 14.
e. While the agent takes each step, we record the total reward value up to that step and present that in a graph shown in Fig 17.

f. Fig 19 shows the comparison between 2 different runs after 3000 and 6000 steps respectively.

QTable1 after 3000 .

			east	west	north	south	pickup	dropoff
1	1	blockPickedUp	-1.7793	N/A	N/A	-1.6607	0.0000	N/A
1	2	blockPickedUp	-1.6757	-1.6743	N/A	-1.7037	N/A	N/A
1	3	blockPickedUp	-1.6537	-1.5994	N/A	-1.6164	N/A	N/A
1	4	blockPickedUp	-1.6477	-1.6340	N/A	-1.6029	N/A	N/A
1	5	blockPickedUp	N/A	-1.6429	N/A	-1.6475	N/A	N/A
2	1	blockPickedUp	-1.6998	N/A	-1.6890	-1.1562	N/A	N/A
2	2	blockPickedUp	-1.6272	-1.6211	-1.6259	-1.5556	N/A	N/A
2	3	blockPickedUp	-1.5826	-1.6168	-1.5785	-1.6000	N/A	N/A
2	4	blockPickedUp	-1.6172	-1.5609	-1.5664	-1.6236	N/A	N/A
2	5	blockPickedUp	N/A	-1.6432	-1.6370	-1.5921	N/A	N/A
3	1	blockPickedUp	-1.5918	N/A	-1.5196	-1.5175	N/A	N/A
3	2	blockPickedUp	-1.5041	-1.5164	-1.5275	-1.5017	N/A	N/A
3	3	blockPickedUp	-1.6238	-1.5773	-1.6191	-1.5855	0.0000	N/A
3	4	blockPickedUp	-1.6215	-1.6082	-1.5777	-1.5793	N/A	N/A
3	5	blockPickedUp	N/A	-1.5515	-1.6121	-1.6606	N/A	N/A
4	1	blockPickedUp	-1.5384	N/A	-1.5974	-1.5375	0.0000	N/A
4	2	blockPickedUp	-1.5653	-1.4765	-1.5405	-1.5644	N/A	N/A
4	3	blockPickedUp	-1.5336	-1.4492	-1.5780	-1.5159	N/A	N/A
4	4	blockPickedUp	-1.5136	-1.5719	-1.5632	-1.5227	N/A	22.1616
4	5	blockPickedUp	N/A	1.6567	-1.6205	-1.6274	N/A	N/A
5	1	blockPickedUp	-1.5767	N/A	-1.5749	N/A	N/A	19.1101
5	2	blockPickedUp	-1.5846	-1.5462	-1.5485	N/A	N/A	N/A
5	3	blockPickedUp	-0.2533	-1.4543	-1.4813	N/A	N/A	N/A
5	4	blockPickedUp	-1.4389	-1.4702	7.7173	N/A	N/A	N/A
5	5	blockPickedUp	N/A	-0.9071	-1.6656	N/A	0.0000	N/A

QTable2 after 3000 .

			east	west	north	south	pickup	dropoff
1	1	blockNotPickedUp	-1.6242	N/A	N/A	-1.5526	17.2287	N/A
1	2	blockNotPickedUp	-1.5464	-1.5575	N/A	-1.3099	N/A	N/A
1	3	blockNotPickedUp	-1.6139	-1.5210	N/A	-1.5928	N/A	N/A
1	4	blockNotPickedUp	-1.6671	-1.6790	N/A	-1.6691	N/A	N/A
1	5	blockNotPickedUp	N/A	-1.7074	N/A	-1.6881	N/A	N/A
2	1	blockNotPickedUp	-1.5543	N/A	-1.5575	-1.5960	N/A	N/A
2	2	blockNotPickedUp	-1.5455	-1.3756	-1.5986	-1.5691	N/A	N/A
2	3	blockNotPickedUp	-1.5910	-1.5887	-1.5450	1.1911	N/A	N/A
2	4	blockNotPickedUp	-1.6735	-1.6222	-1.6793	-1.6236	N/A	N/A
2	5	blockNotPickedUp	N/A	-1.6856	-1.6842	-1.6528	N/A	N/A
3	1	blockNotPickedUp	-1.6971	N/A	0.7655	-1.7145	N/A	N/A
3	2	blockNotPickedUp	1.3035	-1.6602	-1.6023	-1.6006	N/A	N/A
3	3	blockNotPickedUp	-1.6363	-1.6514	-1.6154	-1.6576	19.0130	N/A
3	4	blockNotPickedUp	-1.6797	-1.6397	-1.6578	-1.6489	N/A	N/A
3	5	blockNotPickedUp	N/A	-1.6452	-1.6558	-1.6025	N/A	N/A
4	1	blockNotPickedUp	-1.4920	N/A	-1.1054	-1.5691	17.2183	N/A
4	2	blockNotPickedUp	-1.5908	-1.5256	-1.6475	-1.6021	N/A	N/A
4	3	blockNotPickedUp	-1.6457	-1.6567	-1.6060	-1.6040	N/A	N/A
4	4	blockNotPickedUp	0.8620	-1.6787	-1.6546	-1.6752	N/A	0.0000
4	5	blockNotPickedUp	N/A	-1.4600	-1.6281	0.2902	N/A	N/A
5	1	blockNotPickedUp	-1.6830	N/A	-0.7924	N/A	N/A	0.0000
5	2	blockNotPickedUp	-1.6108	-1.6450	-1.3537	N/A	N/A	N/A
5	3	blockNotPickedUp	-1.6008	-1.6437	-1.5846	N/A	N/A	N/A
5	4	blockNotPickedUp	-1.6678	-1.6270	-1.6626	N/A	N/A	N/A
5	5	blockNotPickedUp	N/A	-1.6458	-0.6498	N/A	17.2036	N/A

Fig 13: QTable1 instance at after 3000 and 6000 steps for experiment 3

QTable1 after 6000 .

			east	west	north	south	pickup	dropoff
1	1	blockPickedUp	-1.8978	N/A	N/A	-1.8850	0.0000	N/A
1	2	blockPickedUp	-1.8606	-1.8528	N/A	-1.8680	N/A	N/A
1	3	blockPickedUp	-1.8312	-1.8174	N/A	-1.8116	N/A	N/A
1	4	blockPickedUp	-1.8099	-1.8165	N/A	-1.8140	N/A	N/A
1	5	blockPickedUp	N/A	-1.8158	N/A	-1.8054	N/A	N/A
2	1	blockPickedUp	-1.8488	N/A	-1.8499	-1.7833	N/A	N/A
2	2	blockPickedUp	-1.7966	-1.7857	-1.8622	-1.7628	N/A	N/A
2	3	blockPickedUp	-1.7768	-1.7728	-1.7604	-1.4452	N/A	N/A
2	4	blockPickedUp	-1.7996	-1.7798	-1.7884	-1.7720	N/A	N/A
2	5	blockPickedUp	N/A	-1.7864	-1.8138	-1.8076	N/A	N/A
3	1	blockPickedUp	-1.8029	N/A	-1.7645	-1.7664	N/A	N/A
3	2	blockPickedUp	-1.7906	-1.8233	-1.7972	-1.7832	N/A	N/A
3	3	blockPickedUp	-1.7887	-1.7913	-1.7907	0.6822	0.0000	N/A
3	4	blockPickedUp	-1.7899	-1.7804	-1.8125	-1.7834	N/A	N/A
3	5	blockPickedUp	N/A	-1.8005	-1.8066	-1.7848	N/A	N/A
4	1	blockPickedUp	-1.7906	N/A	-1.7793	-1.7621	0.0000	N/A
4	2	blockPickedUp	-1.7721	-1.7714	-1.7973	-1.7637	N/A	N/A
4	3	blockPickedUp	7.4930	-1.7717	-1.7682	-1.7872	N/A	N/A
4	4	blockPickedUp	-1.8119	-1.8014	-1.8067	-1.8169	N/A	21.8072
4	5	blockPickedUp	N/A	-1.8079	-1.8161	-1.8099	N/A	N/A
5	1	blockPickedUp	-1.7646	N/A	-1.7570	N/A	N/A	19.0305
5	2	blockPickedUp	-1.6772	-1.6856	-1.7282	N/A	N/A	N/A
5	3	blockPickedUp	-1.7123	-1.7406	-1.7104	N/A	N/A	N/A
5	4	blockPickedUp	-1.7844	-1.7526	-1.7629	N/A	N/A	N/A
5	5	blockPickedUp	N/A	-1.8018	-1.8253	N/A	0.0000	N/A

QTable2 after 6000 .

			east	west	north	south	pickup	dropoff
1	1	blockNotPickedUp	-1.8901	N/A	N/A	-1.9044	17.2337	N/A
1	2	blockNotPickedUp	-1.8814	-1.8991	N/A	-1.8898	N/A	N/A
1	3	blockNotPickedUp	-1.8876	-1.8854	N/A	-1.8882	N/A	N/A
1	4	blockNotPickedUp	-1.8743	-1.8681	N/A	-1.8733	N/A	N/A
1	5	blockNotPickedUp	N/A	-1.8873	N/A	-1.8835	N/A	N/A
2	1	blockNotPickedUp	-1.8971	N/A	-1.8937	-1.8853	N/A	N/A
2	2	blockNotPickedUp	-1.8868	-1.8782	-1.8883	-1.8859	N/A	N/A
2	3	blockNotPickedUp	-1.8637	-1.8768	-1.8896	-1.8785	N/A	N/A
2	4	blockNotPickedUp	-1.8699	-1.8574	-1.8607	-1.8521	N/A	N/A
2	5	blockNotPickedUp	N/A	-1.8687	-1.8737	-1.8857	N/A	N/A
3	1	blockNotPickedUp	-1.8792	N/A	-1.8827	-1.8670	N/A	N/A
3	2	blockNotPickedUp	-1.8911	-1.8756	-1.8837	-1.8852	N/A	N/A
3	3	blockNotPickedUp	-1.8666	-1.8775	-1.8835	-1.8834	17.2149	N/A
3	4	blockNotPickedUp	-1.8884	-1.8822	-1.8677	-1.8855	N/A	N/A
3	5	blockNotPickedUp	N/A	-1.8756	-1.8733	-1.8775	N/A	N/A
4	1	blockNotPickedUp	-1.8920	N/A	-1.8903	-1.9040	17.2320	N/A
4	2	blockNotPickedUp	-1.9028	-1.8824	-1.8759	-1.8755	N/A	N/A
4	3	blockNotPickedUp	-1.8885	-1.8956	-1.8706	-1.8887	N/A	N/A
4	4	blockNotPickedUp	-1.8945	-1.9021	-1.8817	-1.9034	N/A	0.0000
4	5	blockNotPickedUp	N/A	-1.8962	-1.9052	0.9697	N/A	N/A
5	1	blockNotPickedUp	-1.9017	N/A	-1.9014	N/A	N/A	0.0000
5	2	blockNotPickedUp	-1.8822	-1.8852	-1.8908	N/A	N/A	N/A
5	3	blockNotPickedUp	-1.8773	-1.8930	-1.8799	N/A	N/A	N/A
5	4	blockNotPickedUp	-1.8882	-1.8978	-1.8886	N/A	N/A	N/A
5	5	blockNotPickedUp	N/A	-1.8965	-1.8704	N/A	18.2510	N/A

Fig 14: QTable2 instance at after 3000 and 6000 steps for experiment 3

If we consider the Fig 15 and Fig 16, we observe the following:

a. The states which are around the pickup state or drop-off state, the QValue for that action which results into the pickup or drop-off will be the highest among the other action's QValues in the same location.

   This observation states that, when the agent gets some experience with the world, it will be able to recognize the pickup and drop-off states when it is adjacent to it based on those QValues.

b. From the drop-off position's QValues, we can infer that the drop-off locations (4, 1) is explored more than that of the drop-off location at (3, 3) up to .

c. From the (2, 2), (3, 1), (4, 4) pickup position's QValues, we can infer that these pickup locations are equally explored by the agent. The QValue of pickup location (0, 0) is higher than the other pickup locations. So, (0, 0) is visited more than the others upto 3000 steps. (2, 2), (3, 1) pickup location's QValues after 6000 steps indicate that there is not much increase which implies that they are not visited regularly. But there is fair

increase in the QValues of the (4, 4) pickup location, which infer that they are more regularly visited than others. There is decrease in the QValues of the (0, 0).

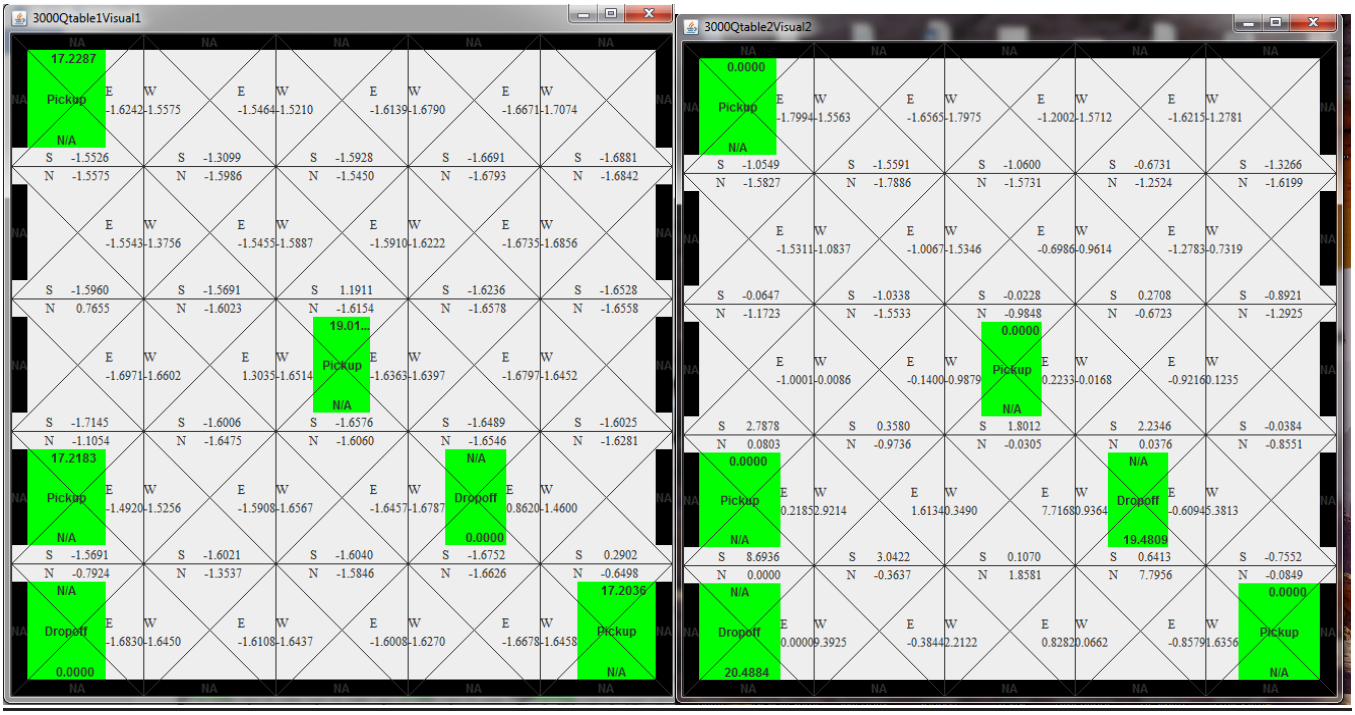

Fig 15: QTable1 after 3000 steps for experiment 3

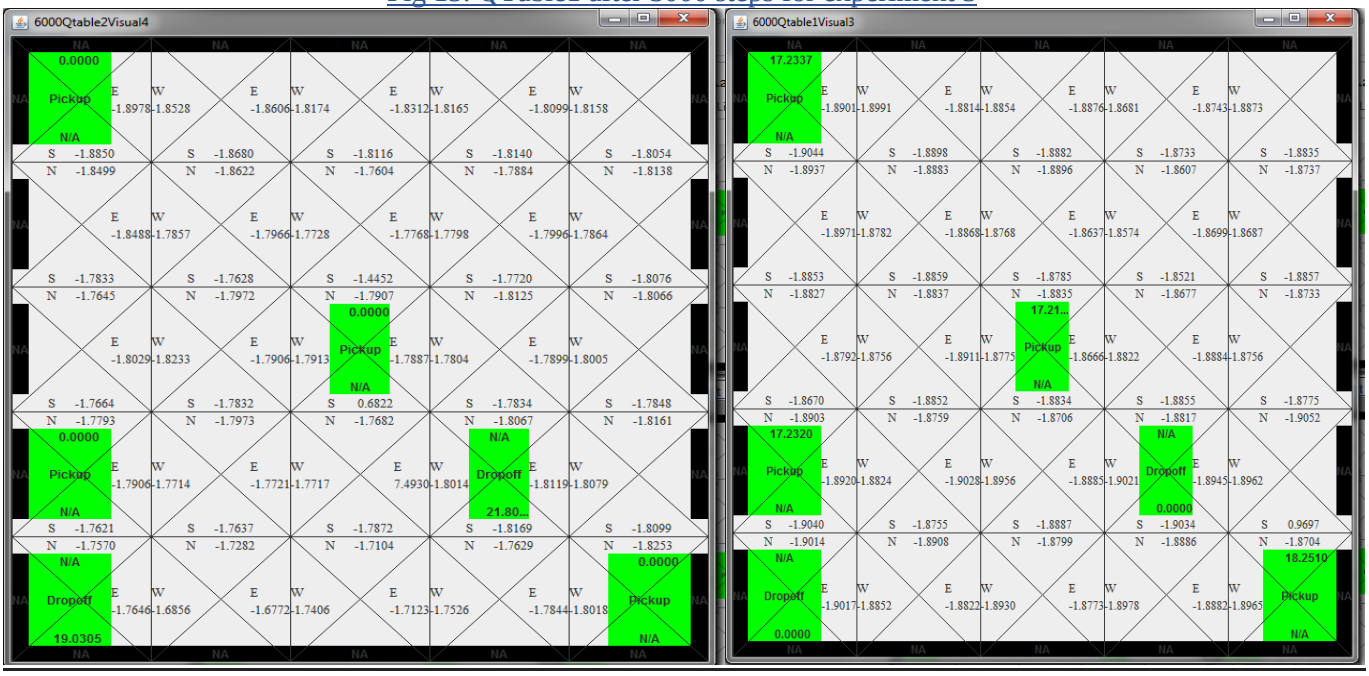

Fig 16: QTable1 & QTable2 after 6000 steps for experiment 3

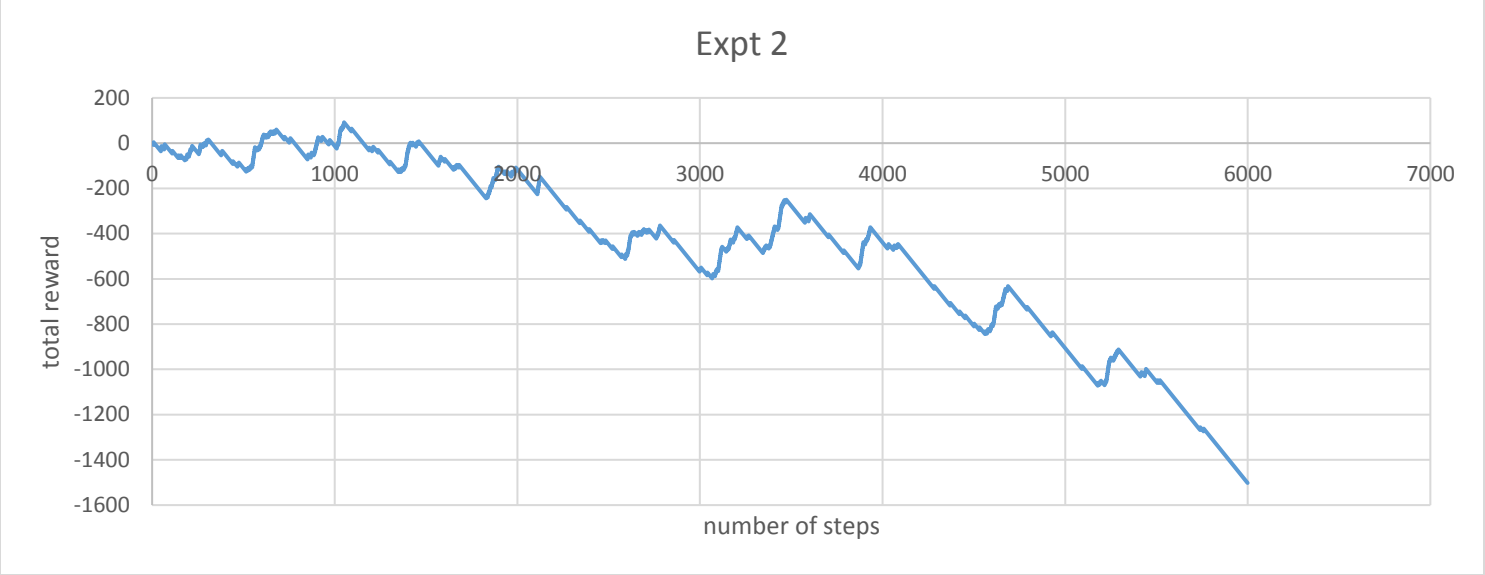

Fig 17: total reward for each step for experiment3

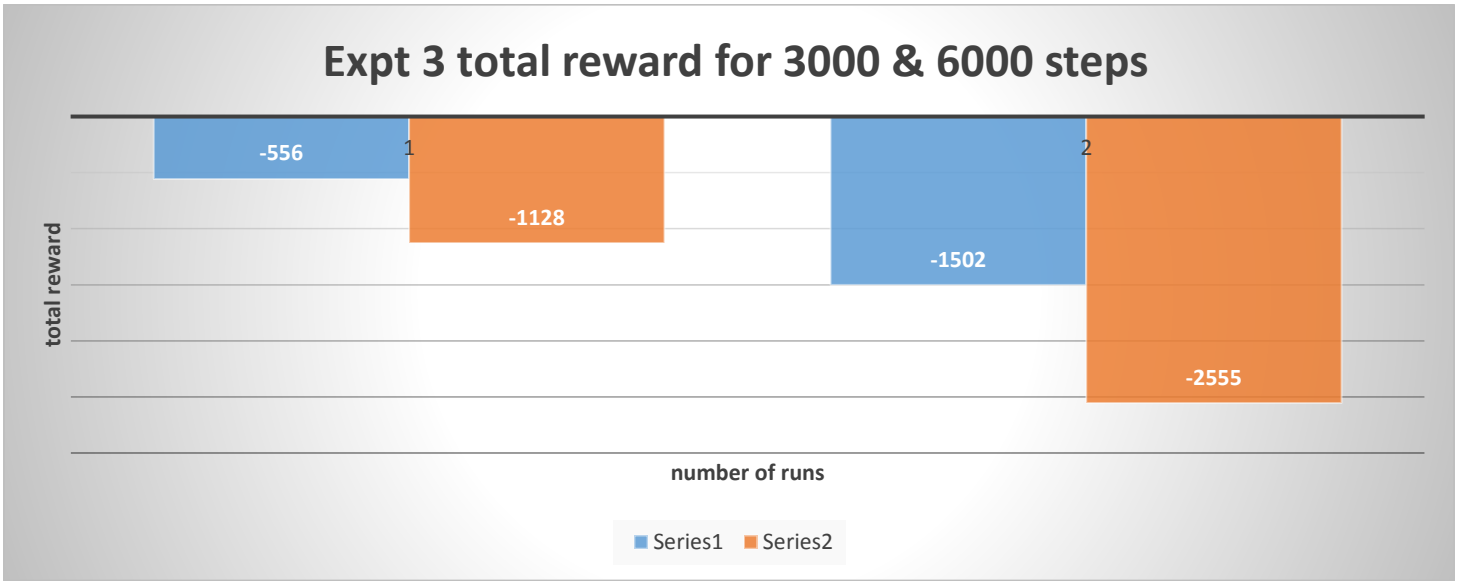

Fig 18: comparison of total reward after 3000 & 6000 steps for 2 different runs of experiment 3

9. Comparison between QLearning and SARSA Techniques

The number of times the agent reaches a termination state is more in SARSA when compared to QLearning.

The QValues of the world in the two Algorithms are almost comparable. So, the agent has visited the pickup and drop-off states almost in an equal manner when both algorithms are applied.

But when we compare the final reward value of the two algorithms, it is high for the QLearning algorithm, given this world.

10. Conclusion:

The best results are obtained for the Experiment two as the PEXPLOIT ran for more number of steps than in other experiments.