ECE 289A - An Introduction to Reinforcement Learning HW#2

Ahmed H. Mahmoud

October, 26th 2017

Q.1

The Bellman equation for q_* for the recycling robot is:

$$q_*(h,w) = p(h|h,w)[r(h,w,h) + \gamma max_{a'} q_*(h,a')] = r_w + \gamma max_{a'} q_*(h,a') \tag{1}$$

$$\begin{aligned} q_*(h,s) &= p(h|h,s)[r(h,s,h) + \gamma max_{a'} q_*(h,a')] + p(l|h,s)[r(h,s,l) + \gamma max_{a'} q_*(l,a')] \\ &= \alpha[r_s + \gamma max_{a'} q_*(h,a')] + (1-\alpha)[r_s + \gamma max_{a'} q_*(l,a')] \end{aligned} \tag{2}$$

$$q_*(l,w) = p(l|l,w)[r(l,w,l) + \gamma max_{a'} q_*(l,a')] = r_w + \gamma max_{a'} q_*(l,a') \tag{3}$$

$$\begin{aligned} q_*(l,s) &= p(l|l,s)[r(l,s,l) + \gamma max_{a'} q_*(l,a')] + p(h|l,s)[r(l,s,h) + \gamma max_{a'} q_*(h,a')] \\ &= \beta[r_s + \gamma max_{a'} q_*(l,a')] + (1-\beta)[-3 + \gamma max_{a'} q_*(h,a')] \end{aligned} \tag{4}$$

$$q_*(l,re) = p(h|l,re)[r(l,re,h) + \gamma max_{a'} q_*(l,a')] = \gamma max_{a'} q_*(l,a') \tag{5}$$

where l and h are the low and high state, the actions *search*, *wait* and *recharge* are given the abbreviated as s, w and re. Also, r_s, r_w are the expected number of cans to be collected while searching and waiting respectively i.e., the reward. We omitted all terms with zero probability.

Q.2

The following shows the optimal value function for Gridworld problem given in Example 3.12 by solving the Bellman equation for v_*.

21.9775	24.4194	21.9775	19.4194	17.4775
19.7797	21.9775	19.7797	17.8018	16.0216
17.8018	19.7797	17.8018	16.0216	14.4194
16.0216	17.8018	16.0216	14.4194	12.9775
14.4194	16.0216	14.4194	12.9775	11.6797

Q.3

For this question, we started first by generating the original figures to the problem as descried in Example 4.2 to make sure the code works correctly before adding any additional conditions. Figure 1 shows the solution; the policy as it improves until we reach the optimal policy along with the optimal value function. These results are identical to the one given in the book (Figure 4.2 in the book).

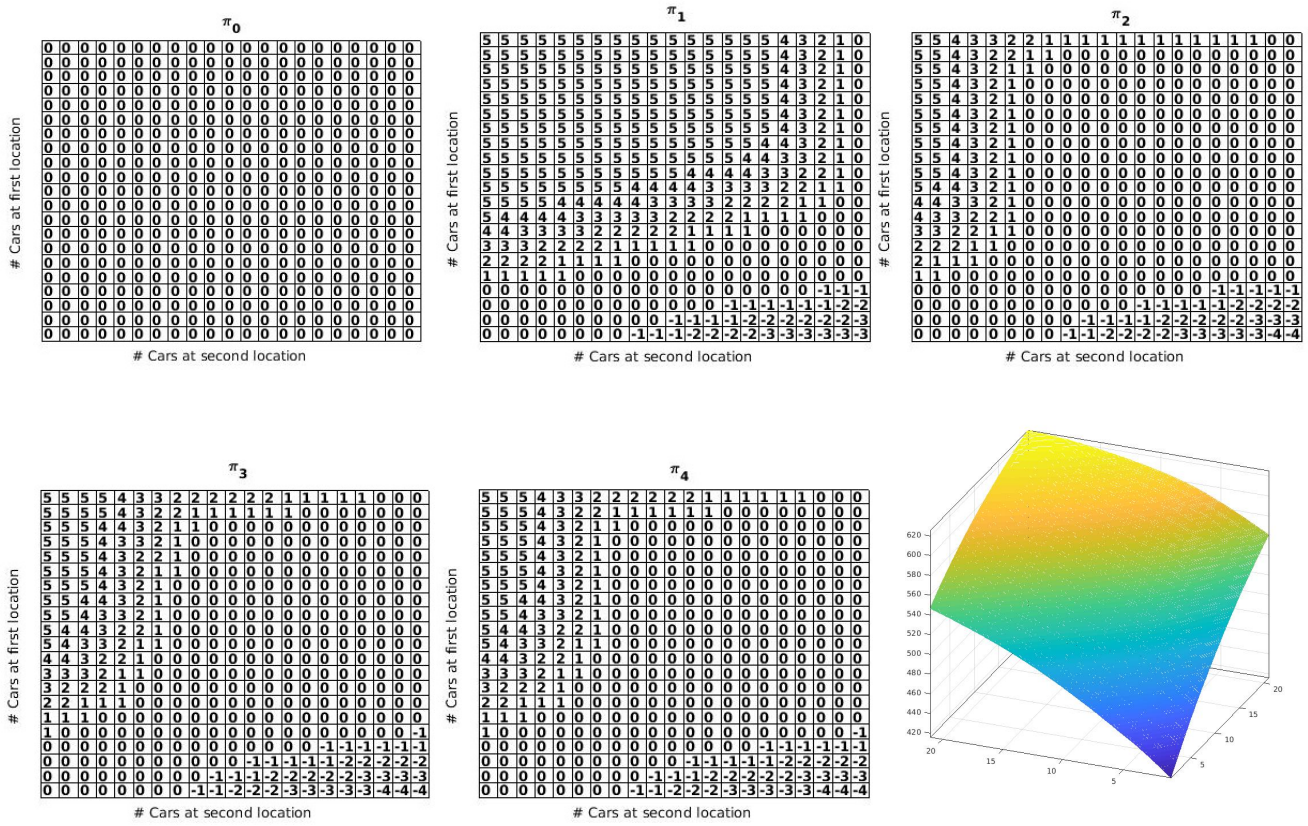

Figure 1: Regenerated original Jack's car rental as described in Example 4.2 without any additional conditions.

Next, we added the new conditions by incurring no cost for the first car moved from the first location to second location. Also, we decreased the reward by 4 whenever there are more than 10 cars in any location (strictly more than). Figure 2 shows the improved policy and the optimal value function.

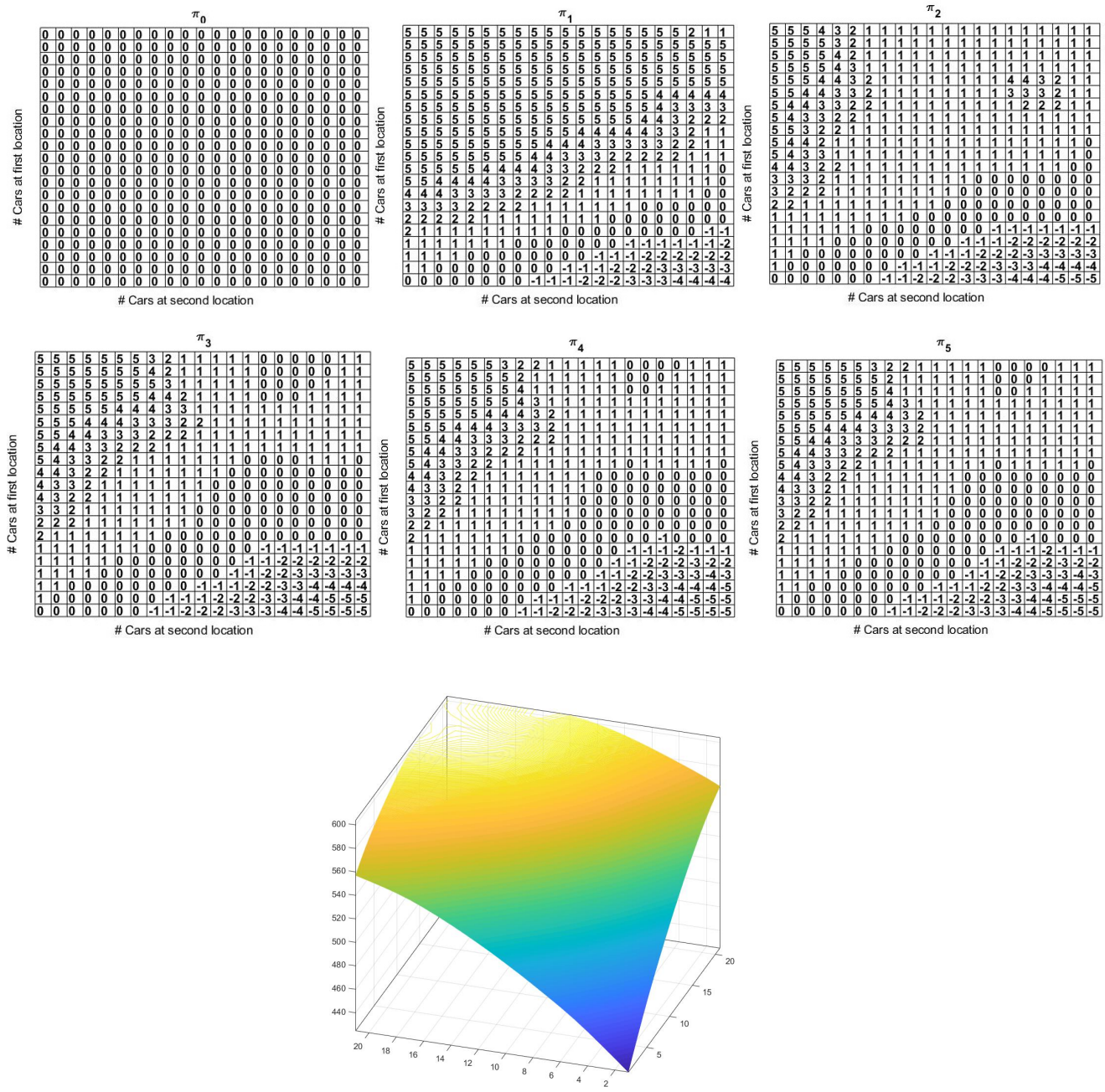

Figure 2: Solution for Jack's car rental after adding the new conditions as described in Exercise 4.4; the policy as it improves till we reach the optimal policy along with the optimal value function.