Chapter 4

Exercise 4.1:

$$q_\pi(11, down) = -1 + v_\pi(terminal)$$
$$q_\pi(14, down) = -1 + v_\pi(11) = -1 - 14 = -15$$

Exercise 4.2:

With the original transitions not changed:

$$\begin{aligned}
v_\pi(15) = {} & 0.25 * [-1 + v_\pi(13)] + 0.25 * [-1 + v_\pi(12)] + 0.25 * [-1 + v_\pi(14)] \\
& + 0.25 * [-1 + v_\pi(15)] \\
0.75 * v_\pi(15) = {} & -1 + 0.25 * [v_\pi(13) + v_\pi(14) + v_\pi(12)] \\
0.75 * v_\pi(15) = {} & -1 + 0.25 * [-20 - 14 - 22] \\
0.75 * v_\pi(15) = {} & -15 \\
v_\pi(15) = {} & -20
\end{aligned}$$

We can see that $v_\pi(15) = v_\pi(13)$ which is expected because State 15 is essentially just a copy of State 13. Therefore, even if the dynamics of the game are changed such that going down from 13 leads to 15, $v_\pi(15) = v_\pi(13) = -20$ will still hold.

Exercise 4.3:

$$q_\pi(s,a) = E_\pi[G_t | S_t = s,\ A_t = a]$$
$$q_\pi(s,a) = E_\pi[R_{t+1} + \gamma * G_{t+1} | S_t = s,\ A_t = a]$$
$$q_\pi(s,a) = E_\pi[R_{t+1} | S_t = s,\ A_t = a] + \gamma * E_\pi[G_{t+1} | S_t = s,\ A_t = a]$$
$$q_\pi(s,a) = E_\pi[R_{t+1} | S_t = s,\ A_t = a] + \gamma \sum_{s^1,\, r} \sum_{a^1} \pi(a^1 | s^1) * p(s^1, r | s, a) * E_\pi[G_{t+1} | S_{t+1} = s^1,\ A_{t+1} = a^1]$$
$$q_\pi(s,a) = \sum_{s^1,\, r} p(s^1, r | s, a) * [r + \gamma * \sum_{a^1} \pi(a^1 | s^1) * q_\pi(s^1, a^1)]$$
$$q_{k+1}(s,a) = \sum_{s^1,\, r} p(s^1, r | s, a) * [r + \gamma * \sum_{a^1} \pi(a^1 | s^1) * q_k(s^1, a^1)]$$

Exercise 4.4:

Instead of checking $old\ action \neq \pi(s)$, we should check $old\ action\ \notin\ \{a \mid q_\pi(s,a) = max_a\ q_\pi(s,a)\}$ (the set of all actions which are optimal).

Exercise 4.5:

1. Initialization

$$q_\pi(s,a) \in R\ and\ \pi(s) \in A(s)\ arbitrarily\ for\ all\ s \in S\ and\ a \in A$$

```
2. Policy Evaluation
   Loop:
         $\triangle \leftarrow 0$
         Loop for all $s \in S\ and\ a \in A$
               $q \leftarrow q_\pi(s,a)$
               $q_\pi(s,a) \leftarrow \sum_{s^1,\, r} p(s^1, r|s,a) * [r + \gamma * \sum_{a^1} \pi(a^1|s^1) * q_\pi(s^1, a^1)]$
               $\triangle = max(\triangle\ ,\ |q - q_\pi(s,a)|)$
   Until $\triangle < \theta$ (a small number)

3. Policy Improvement
   $policy\ stable \leftarrow True$
   Loop for all $s \in S\ and\ a \in A$
         $old\ action = \pi(s)$
         $\pi(s) = arg\ max_a\ q_\pi(s,a)$
         If $old\ action\ \notin \{a \mid q_\pi(s,a) = max_a\ q_\pi(s,a)\}$
               $policy\ stable \leftarrow False$
   If $policy\ stable = True$ then stop and return $q_* \approx q_\pi\ and\ \pi_* \approx \pi$;
   else go to 2
```

<u>Exercise 4.6:</u>

The below changes are for and $\varepsilon - greedy\ policy$

Changes in Step 3:

$$old\ action\ \ =\ \ arg\ max_a\ \pi(a|s)$$

$$a^1\ = arg\ max_a \sum_{s^1,\, r} p(s^1, r|s,a) * v_\pi(s^1)\}$$

$$\pi(a^1|s) = 1 - \varepsilon + \varepsilon/|A(s)|$$

$\pi(a|s) = \varepsilon/|A(s)|\, for\ all\ a \neq a^1$

Changes in Step 2:

$$v_\pi(s) \leftarrow \sum_{s^1,\, r,\, a} p(s^1, r|s, a) * \pi(a|s) * [r + \gamma * v_\pi(s^1)]$$

Changes in Step 1:

Define some ε

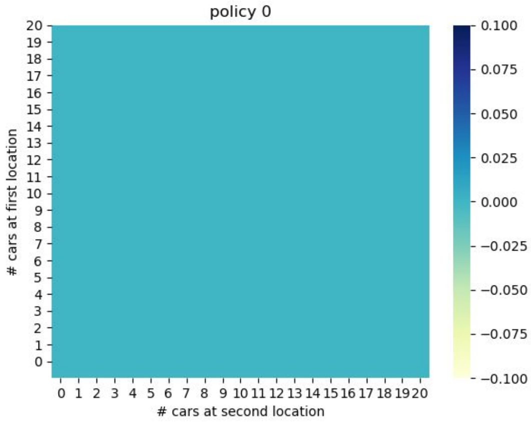
$\pi(s) = arbitrary\ \varepsilon - greedy\ policy$

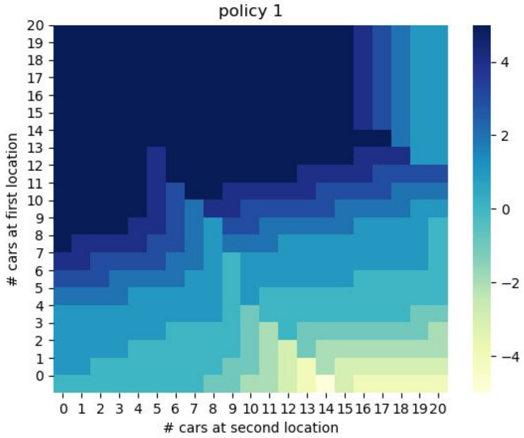
Exercise 4.7 (Programming):

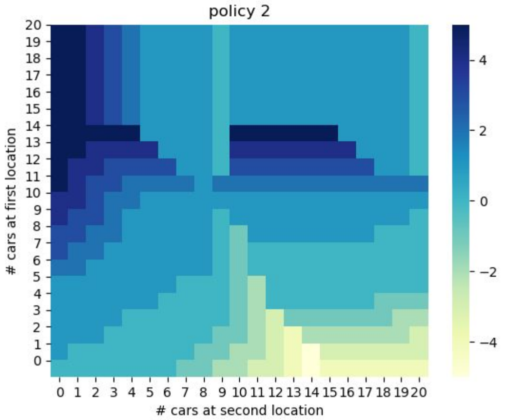
I have also reproduced the example output.

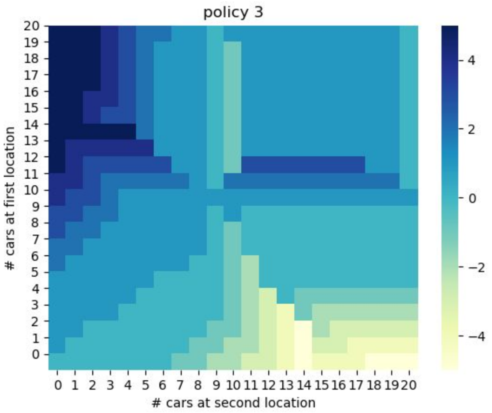
Code and Results: Check Github

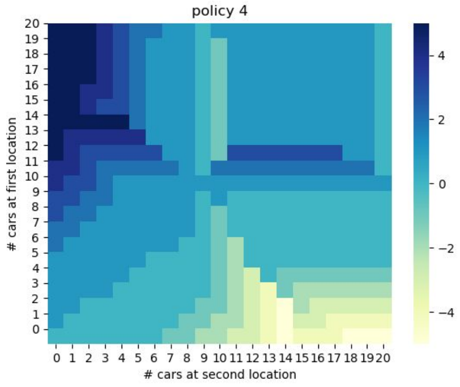
Results:

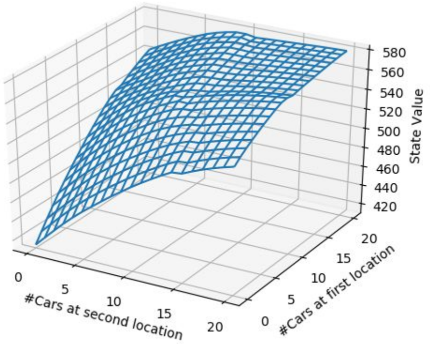










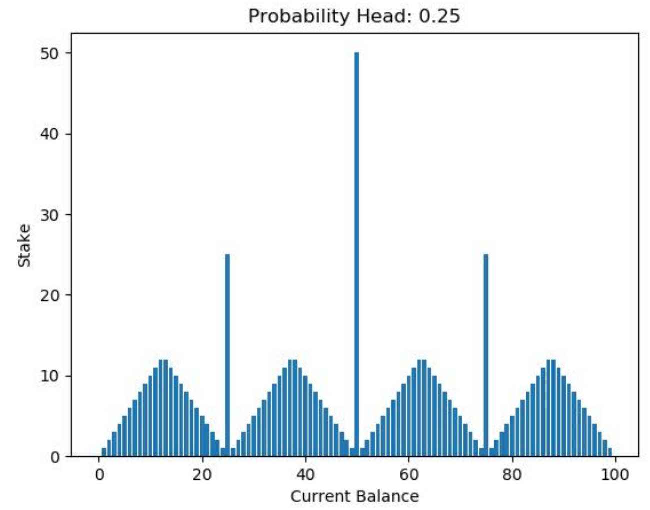


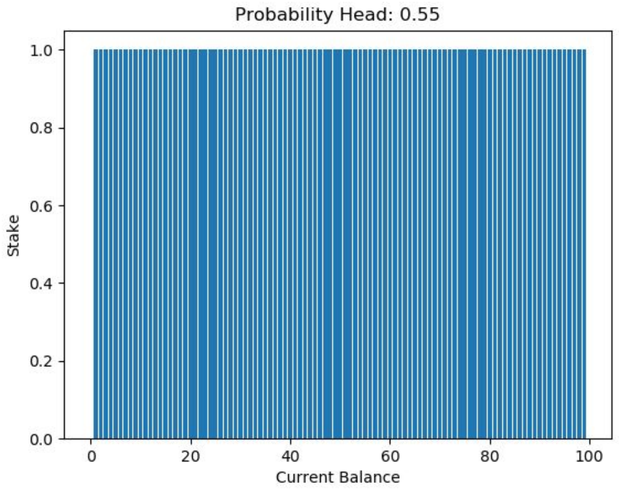
Exercise 4.8:

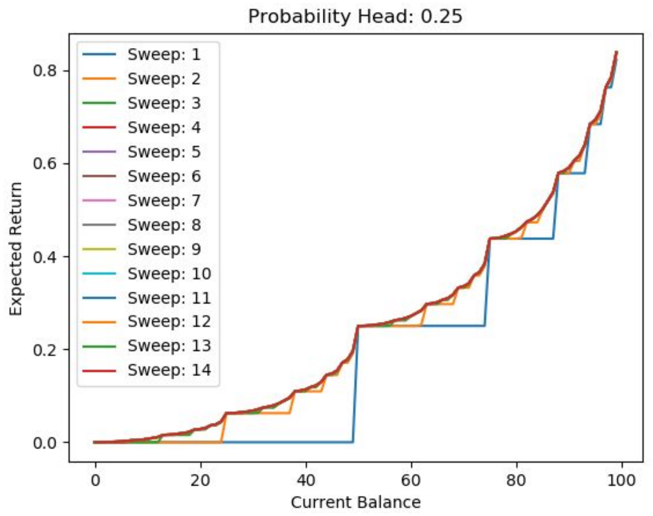
At Capital=50, the policy finds it optimal to bet it all and reach 100 with a probability of p_h . Similarly at Capital=75, the policy finds it optimal to bet maximum (25) and reach 100 with probability p_h . At Capital=51, the policy does not bet maximum in an attempt to reach 100 because it is better to bet small and try to reach the higher milestone of 75. It bets only 1 because it is safe to do so. So the optimal policy leads not only to more chances of winning but also if the gambler loses a bet, the policy ensures that he lands in a safe spot.

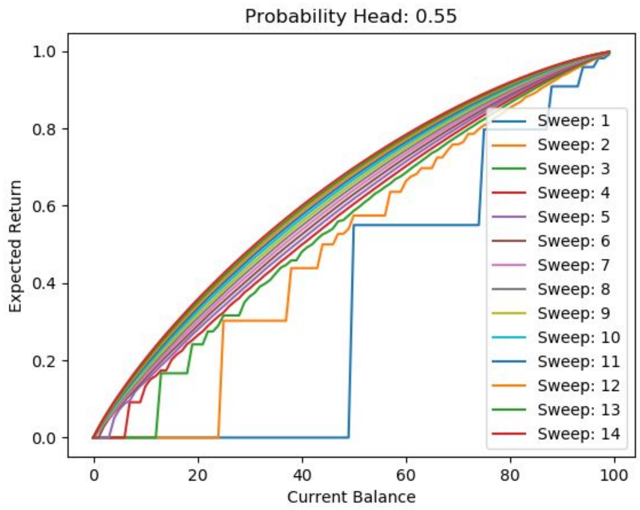
Exercise 4.9 (Programming):

I have also reproduced the example output.
Code and Results: Check Github
Results:
When p_h=0.25, the result is similar to the example which is expected.
When p_h=0.55, the resultant policy is to always bet 1. In fact for any p_h>0.5, I get the same policy.









Exercise 4.10:

$$q_{k+1}(s,a) = \sum_{s^1,\, r} p(s^1, r|s,a) * [r + \gamma * max_{a^1}\ q_k(s^1, a^1)]$$