**Homework 1**

2018324133 김태우

문제의 해결을 위해 Iterative Policy Evaluation 방법을 사용하였다.

이 문제에서 결정할 수 있는 action은 최대 5개까지 차량을 움직이는 것이다. 이 것을 -5~+5의 정수 (1번째->2번째 : 양수, 2번째->1번째 : 음수)로 정의하면 Policy π는 가능한 각각의 State에서 -5~+5로 가는 함수로 정의할 수 있다.

Value Function v는 (Deterministic policy를 가정하였다. 따라서 Policy에 따른 각각의 a의 확률은 고려하지 않는다.)으로 정의된다.

는 s에서 얻을 수 있는 금액의 합이며, 이는 -(차량의 이동 비용)+(차량의 렌트비)로 정의된다. 이동 비용은 사용한 (Policy π에 의한 action의 절대값) x (한 대의 이동 비용)이며, 차량의 렌트비는 (렌트 비용) x (렌트해준 차량 수) x (그만큼의 차량을 렌트해갈 확률)의 합이다. 렌트비를 계산할 때, 차량이 정류장에 있는 수보다 많이 빌려줄 수는 없음을 고려해야 한다.

는 차량의 대여 확률과 반납 확률을 고려해 보았을 때, s에서 s'으로 이전할 확률이며, 는 s'에서의 Value Function이다.

본 문제의 해결을 위해 Python과 numpy 라이브러리를 사용하였다.

|  |
| --- |
| def poisson(avg) :  ret = np.zeros((21))  for i in range(21) :  ret[i] = (avg \*\* i) \* np.exp(-avg) / np.math.factorial(i)  return ret;  def clip(x) :  if x < 0 :  return 0  elif x > 20 :  return 20  else :  return x |

우선 유용한 함수들을 선언하였다. poisson 함수는 평균값 avg를 갖는 포아송 분포의 0~20의 확률을 가진 vector를 반환하며, clip 함수는 값을 0과 20 사이에 있도록 clipping한다.

|  |
| --- |
| requestof1st = poisson(3)  returnof1st = poisson(3)  requestof2nd = poisson(4)  returnof2nd = poisson(2) |

먼저 첫번째 정류장과 두번째 정류장의 대여와 반납의 확률 분포를 정의하였다. 저 변수들은 21의 크기를 가진 vector로, 인덱스는 그 숫자만큼의 차량이 대여/반납될 확률을 가리킨다.

|  |
| --- |
| transition = np.zeros((21, 21, 21, 21))  transition\_reward = np.zeros((21, 21))  for it\_1 in range(21) :  for it\_2 in range(21) :  transition\_1 = np.zeros(21)  transition\_2 = np.zeros(21)  for reqit in range(21) :  for retit in range(21) :  transition\_1[clip(it\_1 - reqit + retit)] += requestof1st[reqit] \* returnof1st[retit]  transition\_reward[it\_1][it\_2] += RENT\_COST \* \  (reqit if it\_1 - reqit + retit >= 0 else it\_1) \* requestof1st[reqit] \* returnof1st[retit]  for reqit in range(21) :  for retit in range(21) :  transition\_2[clip(it\_2 - reqit + retit)] += requestof2nd[reqit] \* returnof2nd[retit]  transition\_reward[it\_1][it\_2] += RENT\_COST \* \  (reqit if it\_2 - reqit + retit >= 0 else it\_2) \* requestof2nd[reqit] \* returnof2nd[retit]  transition[it\_1][it\_2][:][:] = np.outer(transition\_1, transition\_2) |

다음으로 계산 속도의 향상을 위해 transition matrix를 선언하였다. 앞의 Value Function의 공식을 보면 state s에서 action을 취한 (밤에 차량을 이동만 시킨) 결과를 s\*라 하면, , 는 action과 상관없는 값이므로 미리 만들어서 연산 시간을 절약할 수 있음을 알 수 있다.

따라서, 정류장 1에 n1, 정류장 2에 n2대가 있는 state를 (n1, n2)로 표기할 때, transition(o1, o2, t1, t2)는 s\* = (o1, o2)에서 s' = (t1, t2)로 전이할 확률을 미리 계산해놓은 matrix이며, transition\_reward(o1, o2)는 s\* = (o1, o2)에서 얻을 것으로 예상되는 렌트비의 평균이다.

|  |
| --- |
| #Firstly Set policy as zeros  policy = np.zeros((21, 21), dtype=np.int32)  reward = np.zeros((21, 21))  # 1000 Iterations  for iter in range(1, 1001) :  new\_reward = np.zeros((21, 21))  # Calculate next policy greedy  for it\_1 in range(21) :  for it\_2 in range(21) :  for policy\_it in range ( max(it\_1 - 20, -it\_2, -5), min(it\_1, 20 - it\_2, 5) + 1 ) :  cur\_reward = MOVING\_COST \* abs(policy\_it) + \  transition\_reward[it\_1 - policy\_it][it\_2 + policy\_it] +\  GAMMA \* (np.sum(np.multiply(transition[it\_1 - policy\_it][it\_2 + policy\_it], reward)))  if new\_reward[it\_1][it\_2] < cur\_reward :  new\_reward[it\_1][it\_2] = cur\_reward  policy[it\_1][it\_2] = policy\_it  reward = new\_reward |

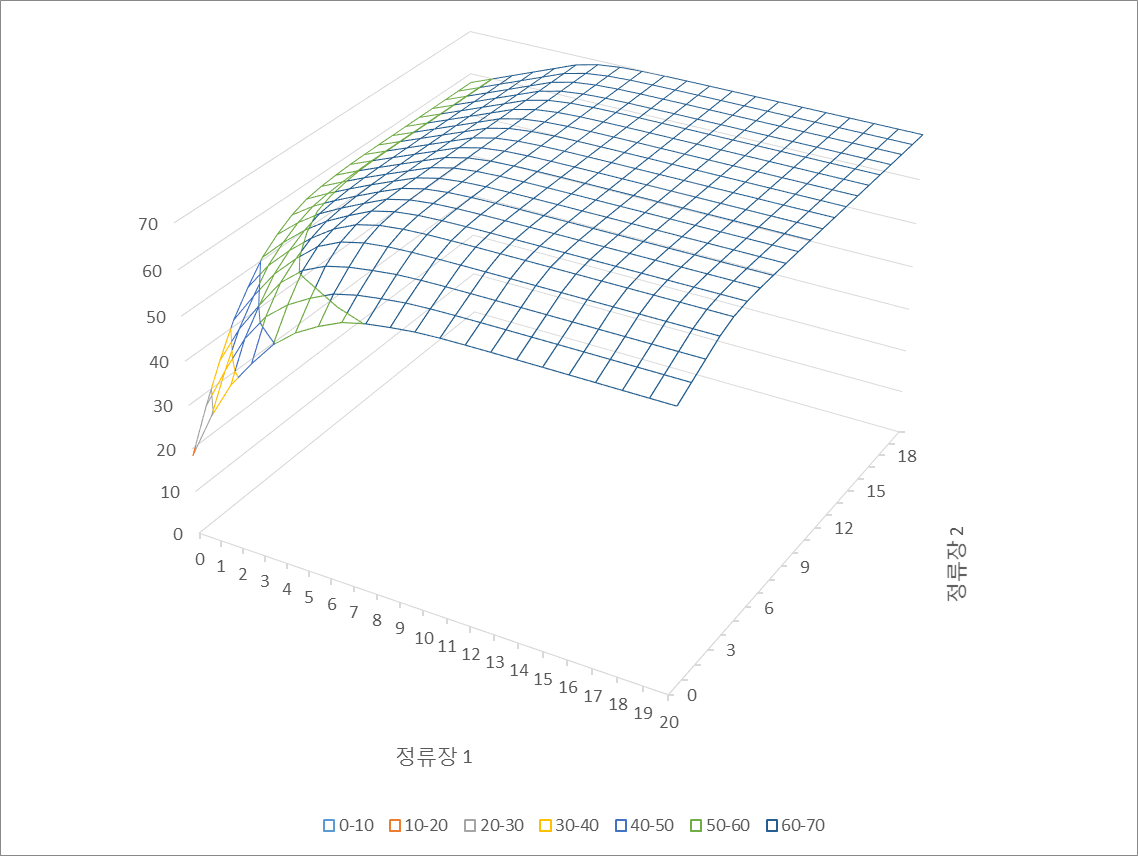
먼저 initial policy는 차량을 전혀 움직이지 않는 uniform zero로 설정하였다. 그리고 매 iteration마다 각state별로 greedy하게 가장 나은 policy를 찾는다.

**결과**

Iteration 1 - Policy

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 0 | -1 | -1 | -2 | -2 | -2 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -3 | -3 |
| 1 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | -2 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 3 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 3 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

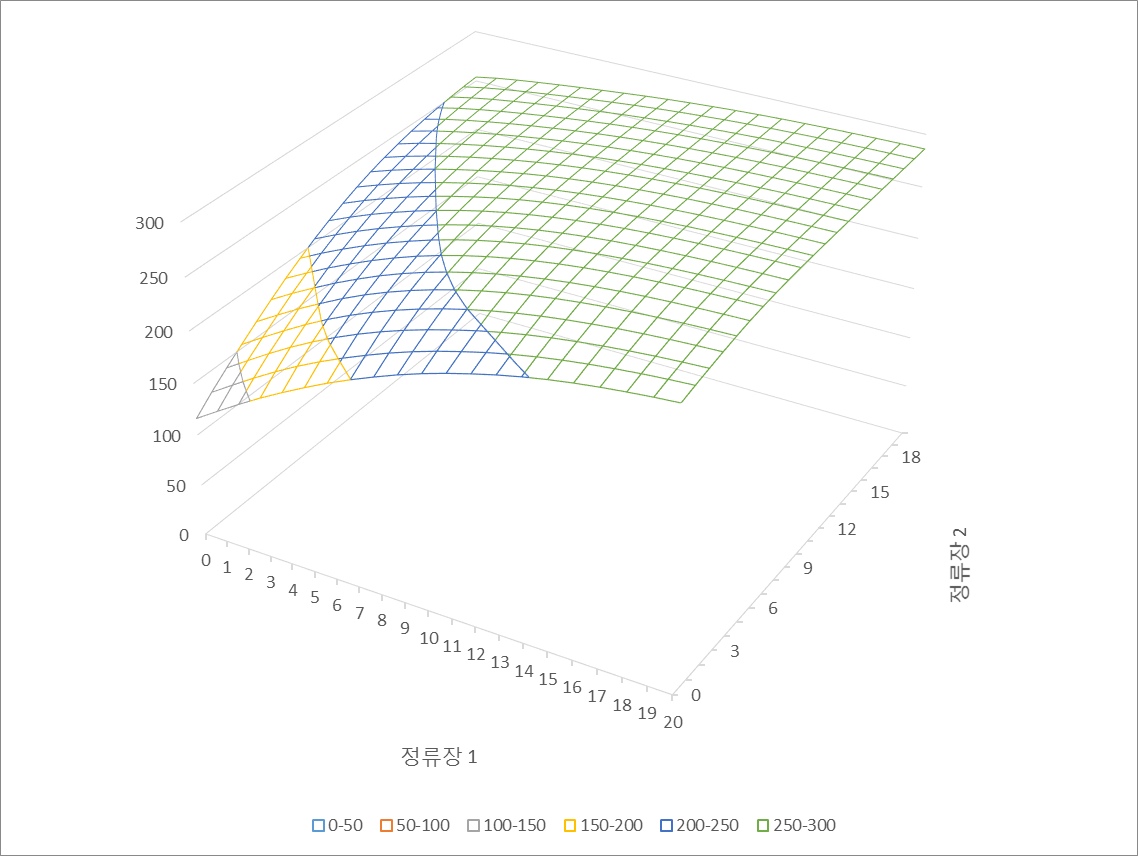
Iteration 1 - Value



Iteration 5 - Policy

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -2 | -2 | -2 | -2 | -3 | -3 | -3 | -3 | -3 | -3 | -4 | -4 | -4 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -3 | -3 | -3 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -1 | -1 | -1 | -2 | -2 | -2 | -2 | -2 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -1 | -1 | -1 |
| 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 3 | 3 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 4 | 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 4 | 3 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 3 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

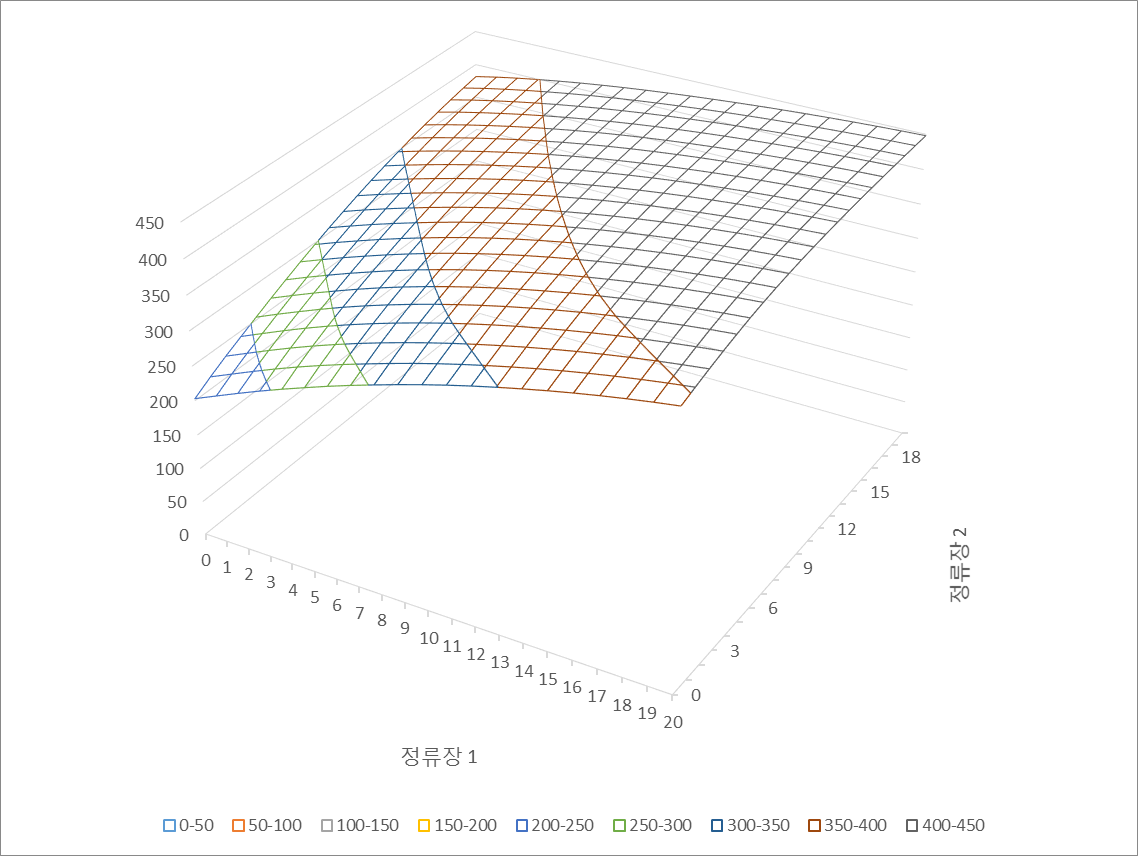
Iteration 5 - Value



Iteration 10 - Policy

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -2 | -2 | -2 | -3 | -3 | -3 | -3 | -3 | -3 | -3 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -1 | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -2 | -2 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -2 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 3 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 3 | 3 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 4 | 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 3 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 3 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 3 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 3 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 3 | 3 | 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 4 | 3 | 3 | 2 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

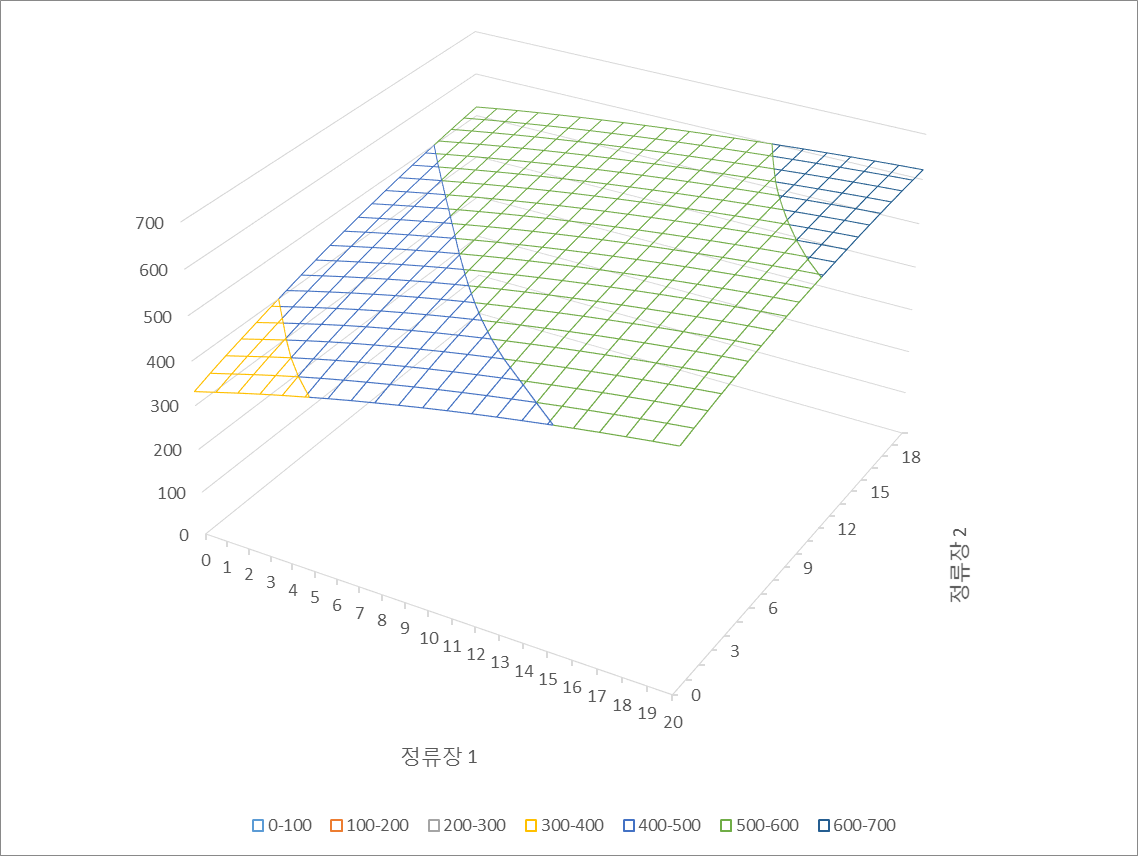
Iteration 10 - Value



Iteration 100 - Policy

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -3 | -3 | -3 | -3 | -3 | -3 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -1 | -1 | -1 | -2 | -2 | -2 | -2 | -2 | -2 | -2 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 | -1 | -1 | -1 | -1 | -1 | -2 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 | -1 |
| 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | -1 |
| 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 3 | 3 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 3 | 3 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 4 | 4 | 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 3 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 4 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 3 | 2 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 3 | 3 | 2 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 4 | 4 | 3 | 2 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 3 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| 5 | 5 | 5 | 4 | 4 | 4 | 3 | 3 | 3 | 3 | 3 | 3 | 2 | 2 | 2 | 2 | 1 | 1 | 1 | 0 | 0 |

Iteration 100 - Value



100번째에서 충분히 수렴하였지만, 과제의 예시와는 약간 다른 값을 나타내고 있었다. reward를 구할 때, 대여 가능한 차량 수를 (갖고 있던 차량 수) + (반납받은 차량 수)로 잡았는데 여기서 약간의 오차가 있지 않았을까 하는 생각이 든다. 그래도 꽤 유사한 수준까지는 수렴했음을 확인할 수 있었다.