Assignment 1: Optimizing Finite-Horizon Expected Total Reward

Steve Hill

April 2015

1 Lunch in Corvallis Problem

The MDP selected for evaluation of the Value Iteration algorithm was eating lunch in Corvallis. The idea is this, you want to eat lunch on Monroe, but you have h-minutes to eat lunch. Different restaurants have different rewards for being more delicious and having a better beer selection. The goal is to accumulate as much delicious food and beer as possible before you run out of lunch time.

1.1 States

The states in this problem represent city blocks. Some of these blocks contain restaurants, which the agent will earn a reward for consuming a delicious food or drink.

1.2 Actions

There are two actions, do nothing, or move forward. Moving forward involves moving from one city block to the "next" possible block. Unfortunately, sometimes there is a lot of traffic on Monroe, and the move action will occasionally result in no movement.

1.2.1 Move action

The move action has a 80% probability of moving to the next tile, and a 20% probability of being blocked by traffic and remaining in the current tile. Moving is deterministic only when moving from a restaurant, as it makes optimality more clear.

1.2.2 Nothing action

The nothing action always keeps the agent where he is. Hopefully this is somewhere with delicious food or drinks.

1.3 Rewards

The reward for each restaurants increases incrementally as you move down Monroe. For some reason, the restaurants with the best food and drink are the farthest away. Regular city block provide no reward. More specifically the reward values are:

$$[1.0, 0.0, 0.0, 2.0, 0.0, 0.0, 3.0, 0.0, 0.0, 4.0, 0.0, 0.0, 5.0, 0.0, 0.0, 6.0, 0.0, 0.0, 7.0, 0.0]$$

2 Results

The planner was run on h=10 and h=20

2.1 Optimal Policy

There are two clear cases, that the agent is at a restaurant, or he is not. If the agent is not at a restaurant it's obvious the best choice is to continue forward. However, if the agent is at a restaurant, the optimal policy is to move to continue on ONLY if the amount of time remaining after the expected arrival is greater than the time it takes to travel to the next restaurant.

Given our move action probability of 80% for non-restaurant tiles, we have an expected time of travel on non-restaurant tiles of 2.5 steps. Keeping this in mind, our reward is equal to our reward at the first restaurant, plus however many remaining tiles we have in the next restaurant tiles. The following is a formula for calculating the reward of a given restaurant and a horizon h. Let n represent the nth restaurant and R_n represent the reward of that restaurant.

$R_n \in Restaurant, R_n + 2.5 * 0 + R_n + 1 * (h - 2.5)$

It's clear that it takes an additional 3 time steps to make moving on worth it, which is pretty intuitive by looking at the setup of the problem.

2.2 Value function

The tables representing the value functions are in the results. Each table is Hxn, where H is the horizon and n is the number of states. Each column i represents i steps remaining. Each row represents a state. There is a header row and column for both feature. The first is h = 6, which represents the threshold for which it is suitable to try and get to the second restaurant. The second is of size 10. The extended problem shows the pattern continuing onto larger horizons. These are represented in tables 3.1 and 3.3.

2.3 Optimal policy

The table are organized exactly as described in the previous section, however it is important to note that 0 time left means there are no more actions that can be taken. Thus, the None is appropriate for this column (cannot take an action!). You can find the optimal policies in 3.2 and 3.4. These figures are particularly interesting because you can see a clear diagonal of selecting the first action which is completely dependent on how much time is remaining.

3 Data tables and Part 3 of Homework

3.1 Monroe Restaurant Value function H = 6

	5	4	3	2	1	0
0	6.02	5.00	4.00	3.00	2.00	1.00
1	7.00	5.02	3.07	1.28	0.00	0.00
2	9.50	7.50	5.50	3.52	1.60	0.00
3	12.00	10.00	8.00	6.00	4.00	2.00
4	10.51	7.53	4.61	1.92	0.00	0.00
5	14.25	11.25	8.26	5.28	2.40	0.00
6	18.00	15.00	12.00	9.00	6.00	3.00
7	14.01	10.04	6.14	2.56	0.00	0.00
8	19.00	15.00	11.01	7.04	3.20	0.00
9	24.00	20.00	16.00	12.00	8.00	4.00
10	17.51	12.54	7.68	3.20	0.00	0.00
11	23.75	18.75	13.76	8.80	4.00	0.00
12	30.00	25.00	20.00	15.00	10.00	5.00
13	21.01	15.05	9.22	3.84	0.00	0.00
14	28.50	22.50	16.51	10.56	4.80	0.00
15	36.00	30.00	24.00	18.00	12.00	6.00
16	24.51	17.56	10.75	4.48	0.00	0.00
17	33.25	26.25	19.26	12.32	5.60	0.00
18	42.00	35.00	28.00	21.00	14.00	7.00
19	0.00	0.00	0.00	0.00	0.00	0.00

3.2 Monroe Restaurant Optimal Policy H = 6

	5	4	3	2	1	0
0	0	1	1	1	1	None
1	0	0	0	0	0	None
2	0	0	0	0	0	None
3	1	1	1	1	1	None
4	0	0	0	0	0	None
5	0	0	0	0	0	None
6	1	1	1	1	1	None
7	0	0	0	0	0	None
8	0	0	0	0	0	None
9	1	1	1	1	1	None
10	0	0	0	0	0	None
11	0	0	0	0	0	None
12	1	1	1	1	1	None
13	0	0	0	0	0	None
14	0	0	0	0	0	None
15	1	1	1	1	1	None
16	0	0	0	0	0	None
17	0	0	0	0	0	None
18	1	1	1	1	1	None
19	0	0	0	0	0	None

3.3 Monroe Restaurant Value function H = 10

	9	8	7	6	5	4	3	2	1	0
0	14.00	12.00	10.00	8.00	6.02	5.00	4.00	3.00	2.00	1.00
1	15.00	13.00	11.00	9.00	7.00	5.02	3.07	1.28	0.00	0.00
2	17.90	15.50	13.50	11.50	9.50	7.50	5.50	3.52	1.60	0.00
3	21.50	18.50	16.00	14.00	12.00	10.00	8.00	6.00	4.00	2.00
4	22.50	19.50	16.50	13.50	10.51	7.53	4.61	1.92	0.00	0.00
5	26.25	23.25	20.25	17.25	14.25	11.25	8.26	5.28	2.40	0.00
6	30.00	27.00	24.00	21.00	18.00	15.00	12.00	9.00	6.00	3.00
7	30.00	26.00	22.00	18.00	14.01	10.04	6.14	2.56	0.00	0.00
8	35.00	31.00	27.00	23.00	19.00	15.00	11.01	7.04	3.20	0.00
9	40.00	36.00	32.00	28.00	24.00	20.00	16.00	12.00	8.00	4.00
10	37.50	32.50	27.50	22.50	17.51	12.54	7.68	3.20	0.00	0.00
11	43.75	38.75	33.75	28.75	23.75	18.75	13.76	8.80	4.00	0.00
12	50.00	45.00	40.00	35.00	30.00	25.00	20.00	15.00	10.00	5.00
13	45.00	39.00	33.00	27.00	21.01	15.05	9.22	3.84	0.00	0.00
14	52.50	46.50	40.50	34.50	28.50	22.50	16.51	10.56	4.80	0.00
15	60.00	54.00	48.00	42.00	36.00	30.00	24.00	18.00	12.00	6.00
16	52.50	45.50	38.50	31.50	24.51	17.56	10.75	4.48	0.00	0.00
17	61.25	54.25	47.25	40.25	33.25	26.25	19.26	12.32	5.60	0.00
18	70.00	63.00	56.00	49.00	42.00	35.00	28.00	21.00	14.00	7.00
19	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

3.4 Monroe Restaurant Optimal Policy H = 10

	9	8	7	6	5	4	3	2	1	0
0	0	0	0	0	0	1	1	1	1	None
1	0	0	0	0	0	0	0	0	0	None
2	0	0	0	0	0	0	0	0	0	None
3	0	0	1	1	1	1	1	1	1	None
4	0	0	0	0	0	0	0	0	0	None
5	0	0	0	0	0	0	0	0	0	None
6	1	1	1	1	1	1	1	1	1	None
7	0	0	0	0	0	0	0	0	0	None
8	0	0	0	0	0	0	0	0	0	None
9	1	1	1	1	1	1	1	1	1	None
10	0	0	0	0	0	0	0	0	0	None
11	0	0	0	0	0	0	0	0	0	None
12	1	1	1	1	1	1	1	1	1	None
13	0	0	0	0	0	0	0	0	0	None
14	0	0	0	0	0	0	0	0	0	None
15	1	1	1	1	1	1	1	1	1	None
16	0	0	0	0	0	0	0	0	0	None
17	0	0	0	0	0	0	0	0	0	None
18	1	1	1	1	1	1	1	1	1	None
19	0	0	0	0	0	0	0	0	0	None

3.5 MDP1.txt for Part 3 Value function H = 10

	9	8	7	6	5	4	3	2	1	0
0	3.03	3.00	3.00	2.03	2.00	2.00	1.03	1.00	1.00	0.00
1	3.00	2.89	2.02	2.00	1.89	1.02	1.00	0.89	0.00	0.00
2	2.90	2.80	2.01	1.90	1.80	1.01	0.90	0.80	0.01	0.00
3	2.90	2.89	2.02	1.90	1.89	1.02	0.90	0.89	0.00	0.00
4	4.00	3.03	3.00	3.00	2.03	2.00	2.00	1.03	1.00	1.00
5	2.89	2.65	2.00	1.89	1.65	1.00	0.89	0.66	0.00	0.00
6	2.98	2.88	2.65	1.98	1.88	1.65	0.98	0.88	0.66	0.00
7	2.90	2.85	2.01	1.90	1.85	1.01	0.90	0.85	0.00	0.00
8	3.00	3.00	2.03	2.00	2.00	1.03	1.00	1.00	0.03	0.00
9	3.02	2.90	2.89	2.02	1.90	1.89	1.02	0.90	0.89	0.00

3.6 MDP1.txt for Part 3 Optimal Policy H = 10

	9	8	7	6	5	4	3	2	1	0
0	3	3	3	3	3	3	3	3	3	None
1	1	3	3	1	3	3	1	3	0	None
2	1	2	2	1	2	2	1	2	2	None
3	0	0	0	0	0	0	0	0	0	None
4	0	0	0	0	0	0	0	0	0	None
5	2	0	2	2	0	2	2	0	0	None
6	1	1	1	1	1	1	1	1	1	None
7	2	2	2	2	2	2	2	2	0	None
8	1	1	1	1	1	1	1	1	3	None
9	3	0	3	3	0	3	3	0	3	None

3.7 MDP2.txt for Part 3 Value function H = 10

	9	8	7	6	5	4	3	2	1	0
0	4.06	4.00	3.06	3.00	2.06	2.00	1.06	1.00	0.06	0.00
1	4.00	3.99	3.00	2.99	2.00	1.99	0.99	0.99	0.00	0.00
2	5.31	4.62	4.31	3.62	3.31	2.62	2.30	1.60	1.25	0.49
3	3.99	3.04	2.99	2.04	1.99	1.04	0.99	0.07	0.00	0.00
4	4.57	4.00	3.57	3.00	2.57	2.00	1.57	1.00	0.57	0.00
5	5.00	4.00	4.00	3.00	3.00	2.00	2.00	1.00	1.00	0.00
6	5.00	5.00	4.00	4.00	3.00	3.00	2.00	2.00	1.00	1.00
7	3.96	3.08	2.96	2.08	1.96	1.08	0.97	0.08	0.08	0.00
8	4.00	3.57	3.00	2.57	2.00	1.57	1.00	0.57	0.01	0.00
9	5.00	4.00	4.00	3.00	3.00	2.00	2.00	1.00	1.00	0.00

3.8 MDP2.txt for Part 3 Optimal policy H = 10

	9	8	7	6	5	4	3	2	1	0
0	3	0	3	0	3	0	3	0	3	None
1	0	0	0	0	0	0	0	0	0	None
2	0	0	0	0	0	0	0	0	0	None
3	1	1	1	1	1	1	1	2	1	None
4	2	1	2	1	2	1	2	1	2	None
5	2	0	2	0	2	0	2	0	2	None
6	2	2	2	2	2	2	2	2	0	None
7	1	1	1	1	1	1	1	1	3	None
8	0	1	0	1	0	1	0	1	3	None
9	2	2	2	2	2	2	2	2	2	None