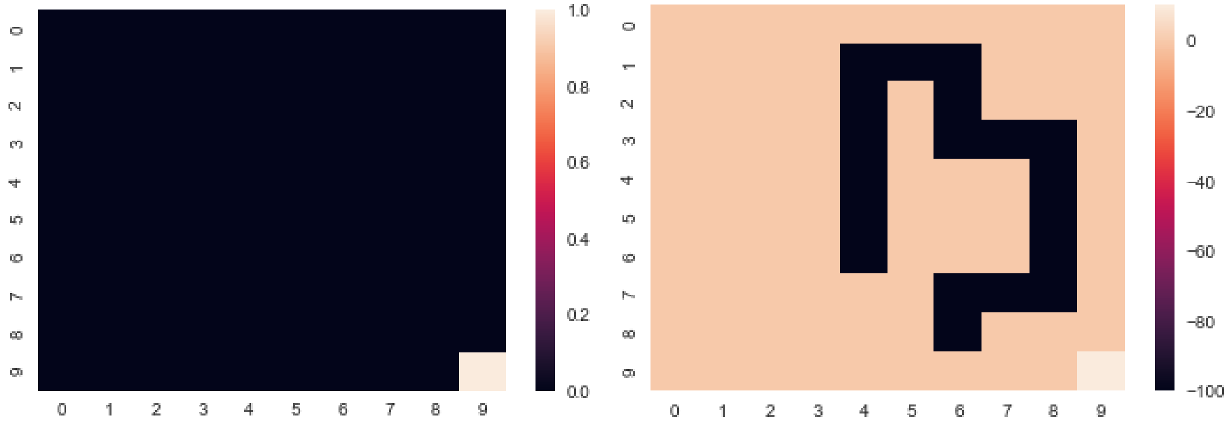
Project 3: Reinforcement learning and Inverse Reinforcement learning

Duan Li 005026839
Di Ma 004945175
Yuan Shao 504880181
Weijie Tang 305029285

Part 1: Reinforcement learning (RL)

Q1

Left plot is heat map for reward function 1 and right plot is heat map for reward function 2.



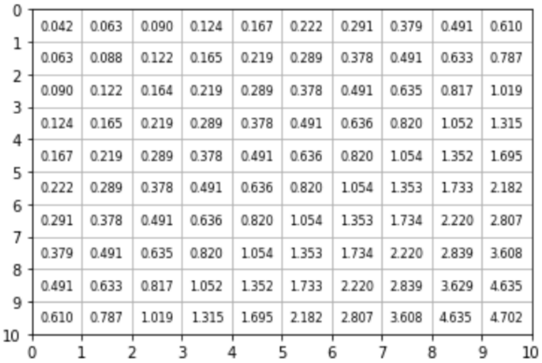
Q2

The optimal state-value function (reward function 1), denoted by V, is defined as below.

$$V(s) \leftarrow \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \mathcal{P}_{ss'}^{a} [\mathcal{R}_{ss'}^{a} + \gamma V(s')]$$

From the equation, it can be observed that V(s) is the maximum value function over all policies. We can use the optimal state-value to determine the deterministic optimal policy.

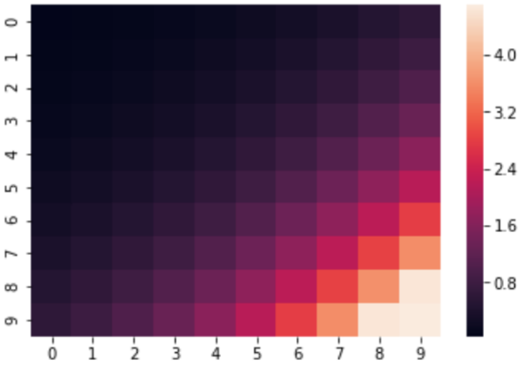
The figure below represents the optimal value of 100 states. From the the figure, we can observe that the state at bottom right corner has the largest optimal value = 4.70154 and the state at top left corner has the smallest optimal value = 0.0417957. Another observation is that the optimal state-value figure is symmetric based on the main diagonal (from the top left corner to the bottom right corner)



0.042	0.063	0.090	0.124	0.167	0.222	0.291	0.379	0.491	0.610
0.063	0.088	0.122	0.165	0.219	0.289	0.378	0.491	0.633	0.787
0.090	0.122	0.164	0.219	0.289	0.378	0.491	0.635	0.817	1.019
0.124	0.165	0.219	0.289	0.378	0.491	0.636	0.820	1.052	1.315
0.167	0.219	0.289	0.378	0.491	0.636	0.820	1.054	1.352	1.695
0.222	0.289	0.378	0.491	0.636	0.820	1.054	1.353	1.733	2.182
0.291	0.378	0.491	0.636	0.820	1.054	1.353	1.734	2.220	2.807
0.379	0.491	0.635	0.820	1.054	1.353	1.734	2.220	2.839	3.608
0.491	0.633	0.817	1.052	1.352	1.733	2.220	2.839	3.629	4.635
0.610	0.787	1.019	1.315	1.695	2.182	2.807	3.608	4.635	4.702

Q3

The heat map of the optimal state values (reward function 1) is shown below.



Q4

From the heat map above, we have the following observations:

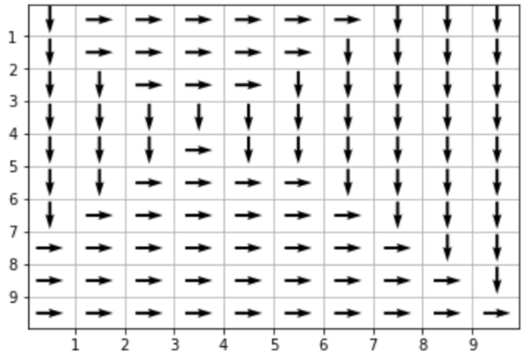
- The state with smallest optimal state value (0~0.8) is black or dark grey, distributed from top left corner to the antidiagonal, covers almost half of the heat map.
- The state with relatively small optimal state value (0.8~1.6) is purple, distributed at the area below the antidiagonal (from the top right to the bottom left corner).
- The state with relatively large optimal state value (1.6~3.6) is red or orange, distributed at the area below the antidiagonal and above the bottom right corner
- The state with largest optimal state value (3.6~4.7)is cream, distributed at the bottom right corner.
- The heap map is almost symmetric based on the main diagonal (from the top left corner to the bottom right corner)

These observations make sense because reward function 1 only reward the state at bottom right corner and all the other states has zero reward. In order to achieve the optimal state value, the agent starting from top left corner will go down or right to approach the bottom right state. Hence, we can conclude that:

- the (initial) state at top left corner has the smallest value
- the (final) state at bottom right corner has largest value
- the state has larger optimal value as it moves away from (initial state) top left corner and moves toward (final goal) bottom right corner.

Q5

The figure below represents the optimal action at 100 states (reward function 1).



The optimal policy of the agent matches our intuition.

As mentioned in Q4, since all the other states except the bottom right state has zero reward, **the intuition** is that the agent from top left corner will go down or right to approach the bottom right state and thus achieve the optimal state value.

The figure above shows that the **optimal policy** is to either (1) go down then go right or (2) go right then go down, which matches the intuition.

Additionally, it is possible for the agent to compute the optimal action to take at each state by observing the optimal values of its neighboring states. The agent should moves towards the neighboring states with higher optimal value. For example, in the figure of Q2, for the state at 2nd row and 1st column, the optimal value of neighbor below (3rd row, 1st col) is 0.090 and the optimal value of right neighbor (2nd row, 2nd col) is 0.088 respectively. Since 0.090 > 0.088, the agent should moves to the state below and takes ↓ action, which matches the figure above.

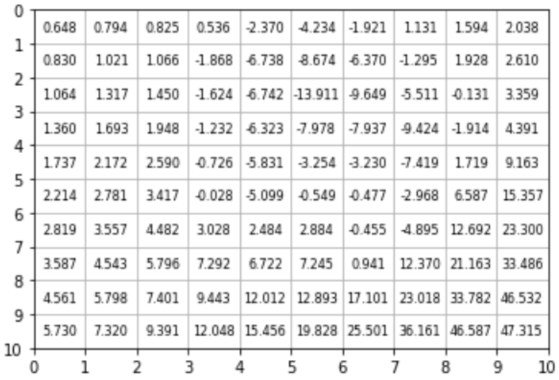
Q6

The optimal state-value function (reward function 2), denoted by V, is defined as below.

$$V(s) \leftarrow \max_{a \in \mathcal{A}} \sum_{s' \in \mathcal{S}} \mathcal{P}^{a}_{ss'}[\mathcal{R}^{a}_{ss'} + \gamma V(s')]:$$

From the equation, it can be observed that V(s) is the maximum value function over all policies. We can use the optimal state-value to determine the deterministic optimal policy.

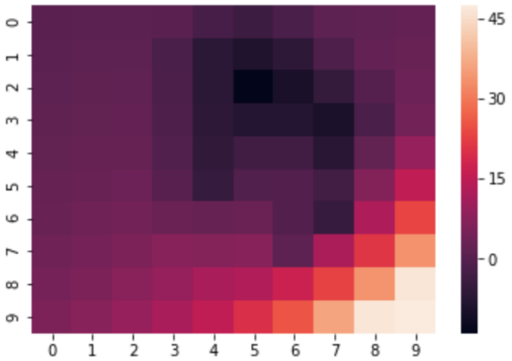
The figure below represents the optimal value of 100 states. From the the figure, we can observe that the state at bottom right corner has the largest optimal value = 47.315 and the state at top left corner has the smallest optimal value = 0.648. Another observation is that



0.648	0.794	0.825	0.536	-2.370	-4.234	-1.921	1.131	1.594	2.038
0.830	1.021	1.066	-1.868	-6.738	-8.674	-6.370	-1.295	1.928	2.610
1.064	1.317	1.450	-1.624	-6.742	-13.911	-9.649	-5.511	-0.131	3.359
1.360	1.693	1.948	-1.232	-6.323	-7.978	-7.937	-9.424	-1.914	4.391
1.737	2.172	2.590	-0.726	-5.831	-3.254	-3.230	-7.419	1.719	9.163
2.214	2.781	3.417	-0.028	-5.099	-0.549	-0.477	-2.968	6.587	15.357
2.819	3.557	4.482	3.028	2.484	2.884	-0.455	-4.895	12.692	23.300
3.587	4.543	5.796	7.292	6.722	7.245	0.941	12.370	21.163	33.486
4.561	5.798	7.401	9.443	12.012	12.893	17.101	23.018	33.782	46.532
5.730	7.320	9.391	12.048	15.456	19.828	25.501	36.161	46.587	47.315

Q7

The heat map of the optimal state values (reward function 2) is shown below.



Q8

From the heat map above, we have the following observations:

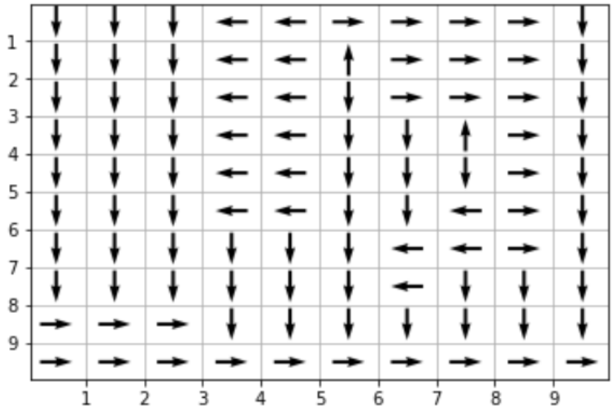
- The state with smallest optimal state value (-14~-5) is black or dark grey, distributed at the thumb-up area around the antidiagonal (from the top right to the bottom left corner). (Check the reward function 2 head map in the Q1 we can observe the black thumb-up.)
- The state with relatively small optimal state value (-1~12) is purple, distributed from the top left corner to the antidiagonal excluding the thumb-up in the middle, covers almost half of the heat map
- The state with relatively large optimal state value (15~36) is red or orange, distributed at the area below the antidiagonal and above the bottom right corner
- The state with largest optimal state value (45~47)is cream, distributed at the bottom right corner.
- The heap map is almost symmetric based on the main diagonal (from the top left corner to the bottom right corner)

These observations make sense because reward function 2 has negative reward at the thumb-up area around the antidiagonal and the only positive reward is at the bottom right state. In order to achieve the optimal state value, the agent starting from top left corner will avoid moving toward state with negative reward but move down or right to approach the bottom right state. Hence, we can conclude that

- the (initial) state at top left corner has the value around zero
- the state with negative reward has the smallest value (negative value)
- the (final) state at bottom right corner has the largest value
- the state has less optimal value as it moves toward negative reward (thumb-up area enclosed by negative reward states)
- the state has larger optimal value as it moves away from (initial state) top left corner, moves away from state with negative reward, and moves toward (final goal) bottom right corner.

Q9

The figure below represents the optimal action at 100 states (reward function 2).



The optimal policy of the agent matches our intuition.

As mentioned in Q8, since the reward function has negative reward of -100 at thumb-up area (check the reward function 2 head map in Q1) and has positive reward of 10 at the bottom right state, **the intuition** is that the agent from top left corner will try to move far away from the negative reward, move toward the positive reward, and thus achieve the optimal state value.

From the figure above, we can observe that **the optimal policy**:

- is unique: go down then go right
- the state at thumb-up area has the action pointing to the opposite direction of the negative reward.
  - actions of state from (1st row, 4th col) to (6th row, 5th col) are all ← to avoid agent moving toward negative rewards -100 from (2nd row, 5th col) to (7th row, 5th col).
  - actions of state enclosed by negative reward from (3rd row, 6th col) to (8th row, 6th col) are all ↓ to escape the surrounding of negative reward and to move toward the (final state) bottom right corner.

All those observations of optimal policy matches the intuition.

Part 2: Inverse Reinforcement learning (IRL)

Q10

To simplify the expression, first we **denote** following matrices, assuming we have 4 actions and a1 is the action given by the optimal policy.

$$\mathbf{V}_0 = \begin{bmatrix} 0 & 0 & \dots & 0 \end{bmatrix} \text{dimension: } (1, |\mathrm{S}|)$$

$$\mathbf{V}_1 = \begin{bmatrix} 1 & 1 & \dots & 1 \end{bmatrix} \text{dimension: } (1, |\mathrm{S}|)$$

$$\mathbf{t} = \begin{bmatrix} t_1 & t_2 & \dots & t_{|S|} \end{bmatrix}^{\mathrm{T}} \text{dimension: } (|\mathrm{S}|,)$$

$$\mathbf{M} = \begin{bmatrix} R_{\max} & \dots & R_{\max} \end{bmatrix}^{\mathrm{T}} \text{dimension: } (|\mathrm{S}|,)$$

$$\mathbf{PP_a} = \begin{bmatrix} (\mathbf{P}_{a1}(1) - \mathbf{P}_a(1))(\mathbf{I} - \gamma \mathbf{P}_{a1})^{-1} \\ (\mathbf{P}_{a1}(2) - \mathbf{P}_a(2))(\mathbf{I} - \gamma \mathbf{P}_{a1})^{-1} \\ \vdots \\ (\mathbf{P}_{a1}(|S|) - \mathbf{P}_a(|S|))(\mathbf{I} - \gamma \mathbf{P}_{a1})^{-1} \end{bmatrix} \text{dimension: } (|\mathrm{S}|, |\mathrm{S}|)$$

$$\mathbf{D}_1 = \begin{bmatrix} \mathbf{-PP_{a2}} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{-PP_{a3}} & \mathbf{I} & \mathbf{0} & \mathbf{0} \\ \mathbf{-PP_{a4}} & \mathbf{I} & \mathbf{0} & \mathbf{0} \end{bmatrix} \text{dimension: } (3|\mathrm{S}|, 4|\mathrm{S}|)$$

$$\mathbf{D}_2 = \begin{bmatrix} \mathbf{-PP_{a2}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{-PP_{a3}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{-PP_{a4}} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \text{dimension: } (3|\mathrm{S}|, 4|\mathrm{S}|)$$

$$\mathbf{D}_3 = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{-I} & \mathbf{0} \end{bmatrix} \text{dimension: } (|\mathrm{S}|, 4|\mathrm{S}|)$$

$$\mathbf{D}_4 = \begin{bmatrix} \mathbf{-I} & \mathbf{0} & \mathbf{-I} & \mathbf{0} \end{bmatrix} \text{dimension: } (|\mathrm{S}|, 4|\mathrm{S}|)$$

$$\mathbf{D}_5 = \begin{bmatrix} \mathbf{I} & \mathbf{0} & \mathbf{0} & \mathbf{-I} \end{bmatrix} \text{dimension: } (|\mathrm{S}|, 4|\mathrm{S}|)$$

$$\mathbf{D}_6 = \begin{bmatrix} \mathbf{-I} & \mathbf{0} & \mathbf{0} & \mathbf{-I} \end{bmatrix} \text{dimension: } (|\mathrm{S}|, 4|\mathrm{S}|)$$

Answer: Then we can express **c**,**x**,**D**:

$$\mathbf{x} = \begin{bmatrix} \mathbf{R} \\ \mathbf{t} \\ \mathbf{u} \\ \mathbf{M} \end{bmatrix} \text{dimension: } (4|\mathrm{S}|,)$$

$$\mathbf{c} = \begin{bmatrix} \mathbf{V}_0 & \mathbf{V}_1 & -\lambda \mathbf{V}_1 & \mathbf{V}_0 \end{bmatrix}^{\mathrm{T}} \text{dimension: } (4|\mathrm{S}|,)$$

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}_1 \\ \mathbf{D}_2 \\ \vdots \\ \mathbf{D}_6 \end{bmatrix} \text{dimension: } (10|\mathrm{S}|, 4|\mathrm{S}|)$$

Explanation: for this project, |S| = 100, and **D**1 is for

$$[(\mathbf{P}_{a_1}(i) - \mathbf{P}_a(i))(\mathbf{I} - \gamma \mathbf{P}_{a_1})^{-1}\mathbf{R}] \geq t_i, \quad \forall a \in \mathcal{A} \setminus a_1, \forall i$$

D2 is for

$$(\mathbf{P}_{a_1} - \mathbf{P}_a)(\mathbf{I} - \gamma \mathbf{P}_{a_1})^{-1}\mathbf{R} \succeq 0, \quad \forall a \in \mathcal{A} \setminus a_1$$

D3 and **D**4 are for

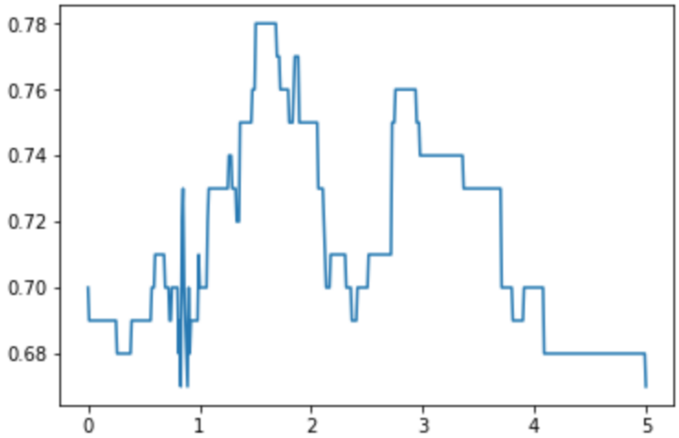
$$-\mathbf{u} \preceq \mathbf{R} \preceq \mathbf{u}$$

D5 and **D**6 are for

$$|\mathbf{R}_i| \leq R_{max}, \quad i = 1, 2, \cdots, |\mathcal{S}|$$

Q11

The plot of λ against accuracy (using optimal policy from Q5) is shown below:



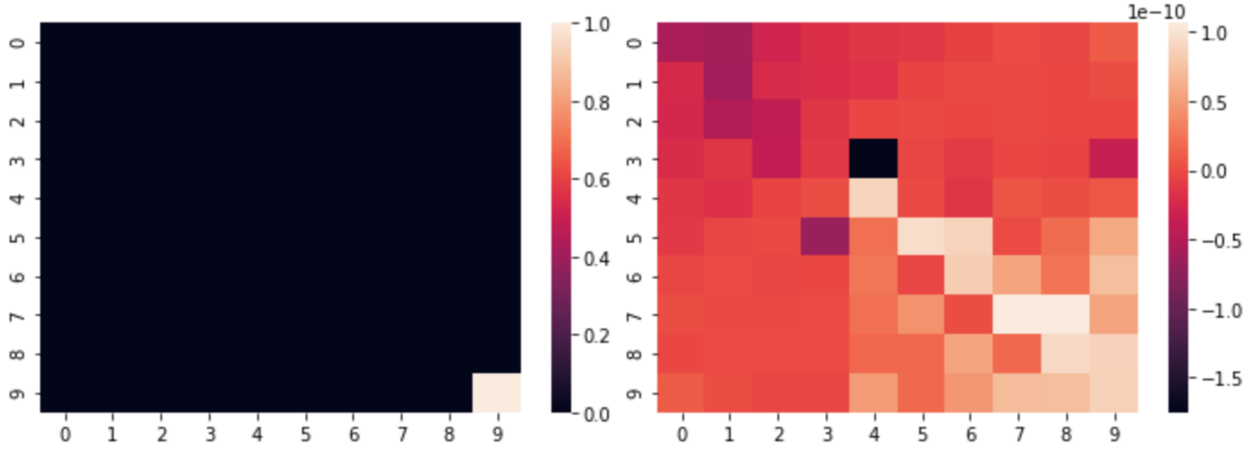
Q12

From the plot above, we can get the best λ with the best accuracy of **0.78**:

$$\lambda_{\max}^{(1)} = 1.503006012024048$$

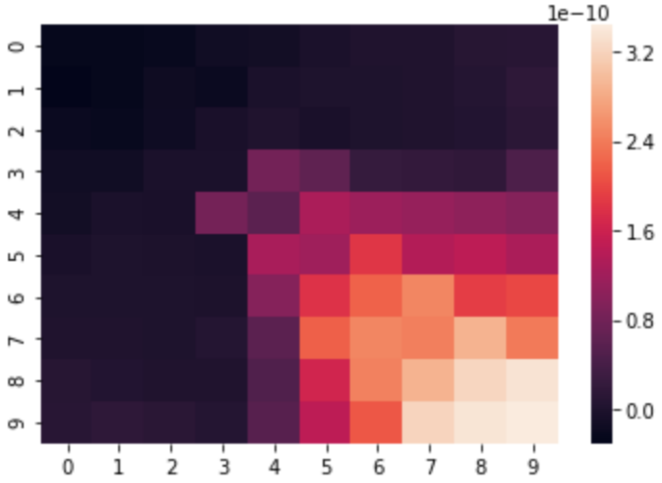
Q13

Left plot is the heatmap of the ground truth reward (reward function 1), and right plot is the heatmap of the extracted reward corresponding to Q12.



Q14

The heatmap of the optimal state values corresponding to Q12 is shown below.



Q15

After comparing the heatmaps of Q3 and Q14, we can observe that:

Similarities:

- For both the heatmaps, the optimal state values increase from the top left corner to the bottom right corner.
- Both heatmaps are almost symmetric based on the main diagonal (from the top left corner to the bottom right corner).

Differences:

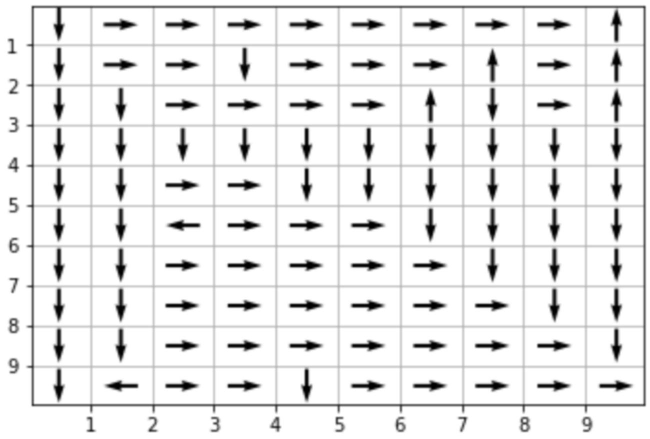
- The smallest and biggest state values of the heatmap in Q14 are lower than those of the one in Q3.
- The heatmap in Q14 is not so symmetric compared with the one in Q3.
- The heatmap in Q14 has more states with relatively large optimal state value (red or orange) than the one in Q3.
- The optimal state values at the top right and the bottom left corner of the heatmap in Q14 are much smaller than those in the heatmap of Q3.

Explanation:

- Since the goal is still to start from top left corner and get to bottom right corner, the optimal state values increase along the main diagonal.
- Because the optimal policy from Q5 may not be unique, so the heatmap in Q14 is not so symmetric.
- When the agent gets to either the top right corner or the bottom left corner, it is still relatively far away from its goal, the bottom right corner, thus in the heatmap of Q14, the optimal state values in those areas are small.

Q16

The figure below represents the optimal action at 100 states (extracted reward function in Q13).



Q17

After comparing the optimal policies from Q5 and Q16, we can observe that:

Similarities:

- Both optimal policies matches the intuition that the agent should either (1) go down then go right or (2) go right then go down.

Differences:

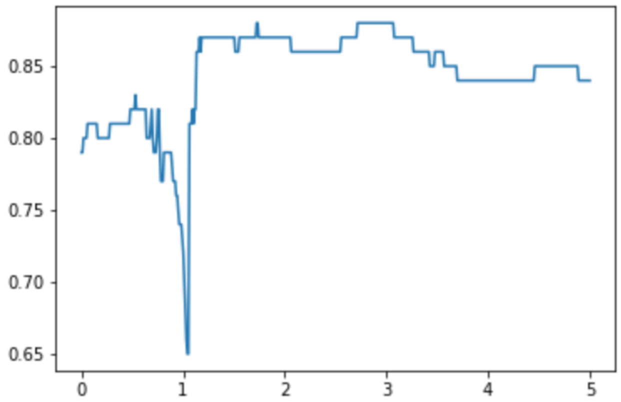
- For some states, the agent can either go right or go down, which causes some differences in the optimal policies.
- In the optimal policy from Q16, there are several states indicating that the agent should go up or go left.
- In the optimal policy, there are several states which do not lead the agent to the (final) bottom right corner, such as the state at the top right corner.

Explanation:

- Since the main goal remains unchanged, the optimal policies have the aforementioned similarities.
- The optimal policy from Q16 is based on the extracted reward function in Q13. And from the heatmap of the extracted reward in Q13, we can see that the values are not guaranteed to increase if the agent go right or go down. That is the reason of the last two differences. For example, the extracted reward decrease from (3rd row, 10th col) to the top right corner, thus causing the optimal policy in Q16 going up in those states, rather than going down.

Q18

The plot of λ against accuracy (using optimal policy from Q9) is shown below:



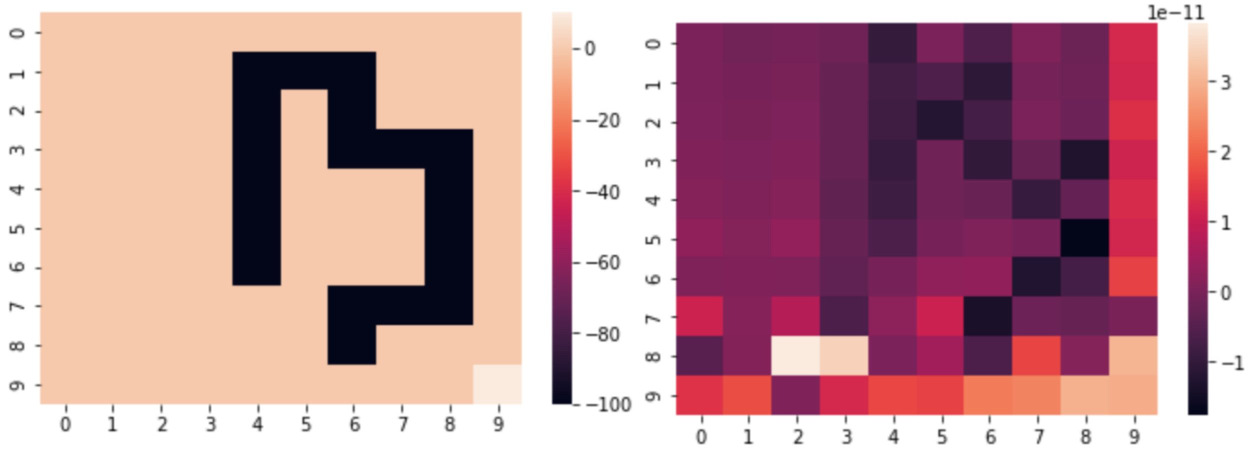
Q19

From the plot above, we can get the best λ with the best accuracy of **0.88**:

$$\lambda_{\max}^{(2)} = 1.723446893787575$$

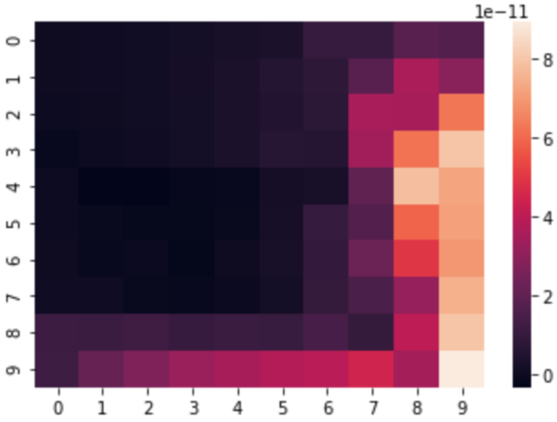
Q20

Left plot is the heatmap of the ground truth reward (reward function 2), and right plot is the heatmap of the extracted reward corresponding to Q19.



Q21

The heatmap of the optimal state values corresponding to Q20 is shown below.



Q22

After comparing the heatmaps of Q7 and Q21, we can observe that:

Similarities:

- For both the heatmaps, the bottom right corner has the largest optimal state value.
- Both heatmaps are somewhat symmetric based on the main diagonal (from the top left corner to the bottom right corner).

Differences:

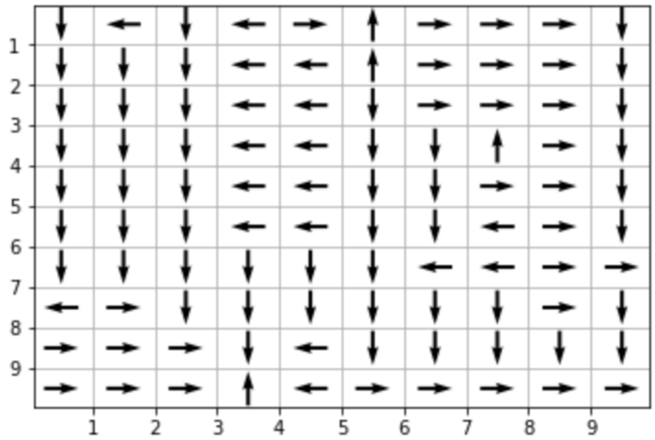
- The smallest and biggest state values of the heatmap in Q21 are lower than those of the one in Q7.
- The heatmap in Q21 has more states with small optimal state value (black or grey) than the one in Q7, especially in the top left corner.
- From the heatmap of Q21, we can no longer observe the thumb-up area.

Explanation:

- Since the goal is still to start from top left corner and get to bottom right corner meanwhile avoid the thumb-up area, in both heatmaps the (final) state at the bottom right corner has the largest optimal state value, and the states near the (initial) top left corner and the thumb-up area have small optimal state values.
- As for the area between the top left corner and the thumb-up area, if the agent goes into that area, it is still far away from its goal, because it needs to avoid the thumb-up area. As a result, the optimal state values of that area in the heatmap of Q21 are also small.

Q23

The figure below represents the optimal action at 100 states (extracted reward function in Q20).



Q24

After comparing the optimal policies from Q9 and Q23, we can observe that:

Similarities:

- Both optimal policies try to guide the agent move from the top left corner to the bottom right corner while avoiding the thumb-up area.

Differences:

- For several states in the optimal policy of Q23, the agent can either go right or go down, such as (8th row, 9th col). Both choices lead the agent to the bottom right corner.

- In the optimal policy of Q23, some states cannot get the agent to the bottom right corner, such as the (8th row, 1st col) and the (9th row, 4th col).

Explanation:

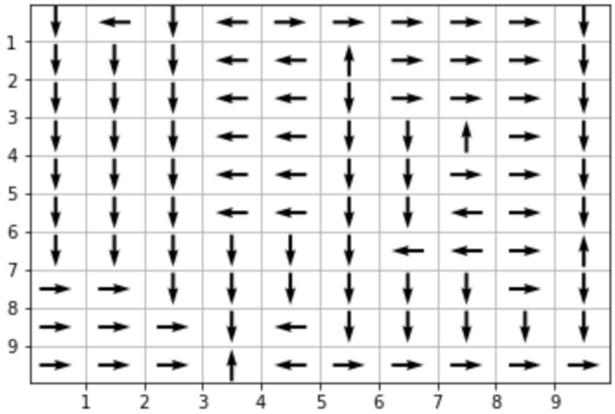
- Since the optimal policy in Q23 is based on the extracted reward function in Q20, and from that heatmap we can see that there are certain areas that form local minimum areas. As a result, if an agent goes into those area, it cannot get out and go to the bottom right corner according to the optimal policy in Q23.
- Due to the same reason, an agent may keep changing states indefinitely, such as the area of (9th row, 4th col) and (10th row, 4th col).

Q25

From the figure in Q23, we observe the optimal policy has following two major discrepancies:

1. For some states on the edges, the agent would not get to the bottom right state according to the optimal policy of Q23, including (8th row, 1st col), (1st row, 6th col), and (7th row, 10th col).
2. The optimal policy in Q23 may lead the agent into a dilemma (local minima). To be specific, if an agent is at (9th row, 4th col) or (10th row, 4th col), then it will keep changing between the two states.

To fix the first discrepancy, we slightly modify the value iteration algorithm by adding constraint in the computation step, so that in the optimal policy, the agent would not get off the board until getting the bottom right corner. After this modification, we get the new optimal policy as shown below and the maximum accuracy increases a little bit to **0.89**.



To fix the second discrepancy, we propose that at any time step t, the agent has a small probability, such as 0.01, to choose a random action, rather than the best action according to the optimal policy. With this small probability, the agent will be able to get out of local minima, and get a better balance between the exploration and the exploitation.