Homework 2: Gridworld

劉安得

106061151

1. Implementation
   - figure_2_1()

     Implement bellman equation

     • Bellman Equation for $v_\pi(s)$

     $$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + \gamma v_\pi(s')].$$

```
for i in range(WORLD_SIZE):
    for j in range(WORLD_SIZE):
        for action in ACTIONS:
            next_state, reward = step([i, j], action)
            new_value[i, j] += ACTION_PROB * (reward +
DISCOUNT*value[next_state[0], next_state[1]])
```

   - figure_2_2()

     Implement optimal value function.

     Record $q_\pi(s, a)$ by `values.append(x)` first, and choose the optimal action value by

     `np.max(values)`

```
for i in range(WORLD_SIZE):
    for j in range(WORLD_SIZE):
        values = []
        for action in ACTIONS:
            next_state, reward = step([i, j], action)
            values.append((reward + DISCOUNT*value[next_state[0],
next_state[1]]))
        new_value[i, j] += np.max(values)
```

2. Experiments and Analysis

	1	2	3	4	5
1	3.31	8.79	4.43	5.32	1.49
2	1.52	2.99	2.25	1.91	0.55
3	0.05	0.74	0.67	0.36	-0.4
4	-0.97	-0.44	-0.35	-0.59	-1.18
5	-1.86	-1.35	-1.23	-1.42	-1.98

↑ equiprobable random policy

	1	2	3	4	5
1	21.98	24.42	21.98	19.42	17.48
2	19.78	21.98	19.78	17.8	16.02
3	17.8	19.78	17.8	16.02	14.42
4	16.02	17.8	16.02	14.42	12.98
5	14.42	16.02	14.42	12.98	11.68

↑ optimal policy

The state values of optimal policy are more reasonable because it chooses actions according to the highest action value, unlike equiprobable random policy chooses actions no matter what action value it got. Choosing the action which gives the best result is more reasonable.

- Gamma = 0

	1	2	3	4	5
1	-0.5	10.0	-0.25	5.0	-0.5
2	-0.25	0.0	0.0	0.0	-0.25
3	-0.25	0.0	0.0	0.0	-0.25
4	-0.25	0.0	0.0	0.0	-0.25
5	-0.5	-0.25	-0.25	-0.25	-0.5

	1	2	3	4	5
1	0.0	10.0	0.0	5.0	0.0
2	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0

- Gamma = 0.1

	1	2	3	4	5
1	-0.27	9.97	0.13	5.0	-0.4
2	-0.26	0.24	0.01	0.12	-0.27
3	-0.27	-0.0	-0.0	-0.0	-0.27
4	-0.28	-0.01	-0.01	-0.01	-0.28
5	-0.54	-0.28	-0.27	-0.28	-0.54

	1	2	3	4	5
1	1.0	10.0	1.0	5.01	0.5
2	0.1	1.0	0.1	0.5	0.05
3	0.01	0.1	0.01	0.05	0.01
4	0.0	0.01	0.0	0.01	0.0
5	0.0	0.0	0.0	0.0	0.0

- Gamma = 0.5

	1	2	3	4	5
1	0.96	9.76	1.89	5.01	0.13
2	-0.01	1.29	0.49	0.66	-0.23
3	-0.34	0.11	0.07	0.03	-0.38
4	-0.47	-0.11	-0.08	-0.13	-0.48
5	-0.83	-0.48	-0.43	-0.48	-0.83

	1	2	3	4	5
1	5.16	10.32	5.16	5.71	2.86
2	2.58	5.16	2.58	2.86	1.43
3	1.29	2.58	1.29	1.43	0.71
4	0.65	1.29	0.65	0.71	0.36
5	0.32	0.65	0.32	0.36	0.18

- Gamma = 0.9

	1	2	3	4	5
1	3.31	8.79	4.43	5.32	1.49
2	1.52	2.99	2.25	1.91	0.55
3	0.05	0.74	0.67	0.36	-0.4
4	-0.97	-0.44	-0.35	-0.59	-1.18
5	-1.86	-1.35	-1.23	-1.42	-1.98

	1	2	3	4	5
1	21.98	24.42	21.98	19.42	17.48
2	19.78	21.98	19.78	17.8	16.02
3	17.8	19.78	17.8	16.02	14.42
4	16.02	17.8	16.02	14.42	12.98
5	14.42	16.02	14.42	12.98	11.68

Because environment is very predictable, we should make policy as farsighted as possible. Therefore, choosing gamma close to 1 is preferable.