Assignment 5 Report for Part A
Chen Bai(chenb24; 1560405)

1. 3-disks, no noise, one goal, living R = 0, discount = 1

1a: 4 times.

1b: 8 times.

1c:
The policy cannot lead the start state to the final goal, and most of states are doing invalid actions and stay in the place.

This is a bad policy.

There is no violation to a selected direction because of the lack of noise, so the T is 1 for all cases. Every state gives a 100 because there is no discount (gamma = 1.0) and there is no living cost(R = 0), which means every q value of a state is the same(1 * [0 + 1.0 * 100] = 100). Therefore, the state will just stick to the first action in the loop while extracting the policy because other values in the Q_Values_Dict are all 100, which will not make any difference.

2. 3-disks, noise = 20%, one goal, living R = 0, discount = 1

2a: 8 times

2b:
This policy makes much more sense than the one in question 1. This policy provides the optimal solution and does no invalid action. So, it can be a good policy.

The only difference in this setup is the noise. T * Vk will yield a q value that points to the direction from which the goal state comes and dominates other q value pointing to other directions because Vk from that goal state direction direction should be larger.

2c: 56 times (total)

2d: It does not change, because after convergence, Vk will not make changes that are large enough to make difference on Vk+1, and so does the q value, which means the policy will stay where they are.

3. 3-disks, noise = 20%, two goal, living R = 0, discount = 0.5

3a: The start state shows 0.82, and the policy leads to R = 10.0 goal state.

3b: The start state shows 36.9, and the policy leads to R = 100.0 goal state.

4.

4a: 6 times(4 times not arrive goal)

4b: 6 times

4c: 3, 1, 2, 1

4d:
the 9-nodes triangle at upper half of the graph;
the 3-nodes triangle at the right-bottom of the 9-node triangle at left-bottom of the graph;
the 3-nodes triangle at the left-bottom of the 9-node triangle at right-bottom of the graph.

5.

5a: No. When the iteration is large enough, like near the convergence, the last several delta between Vk and Vk+1 will not make any large difference on the domination relation between the directions that are in policy and off policy.

5b:The revisisting is important. Because the relationship among states will be ambiguous unless there is a enough number of revisiting or iteration to “fix” the imprecise expectation observed at the beginning of the exploration. And a good policy should be built based on the relationship. So, the revisiting is important.