Problem 1 and 2 are finished. Problem 3 is doable but I ran out of time.

In problem 1, I used a random choice function to do policy evaluation.

In problem 2, I used a policy evaluation function such that it has noise probability to go to the intended direction.

In problem 3, just need to choose the direction that returns maximum rewards.

I spent 10 hours in total.