**Dear all students,**

**Follow the instructions carefully, please:**

* Kindly you can find the following 3 questions you have whole **120 minutes** to answer the questions and **10 minutes of upload** time (not extendable more).
* For the questions that you need to answer in the text, add your answer in this world file and upload it into Moodle. For the questions that you need to write code, upload your python file into Moodle. Rename both files to your English name or student ID.
* You can use Slides and class code examples.
* No need to mention any communication during the exam is acceptable.
  + No open chat apps in background phones only in the pocket all the time, and using the internet is completely prohibited, (you must disconnect your personal laptop or computer from the network during the exam).

**Good Luck.**

--------------------------------------------------------------------------------------------------

English Name: Nick Shen

Chinese name: 沈峻宇

Student ID: 61075029h

--------------------------------------------------------------------------------------------------

**Questions:**

1. What do you think is the main advantage of Thompson over Upper Confidence bound (UCB) algorithms? explain two with reasons.

Answer:

1. It can get good choices quickly. If we repeat it many times the sampling the distribution and getting Qvalue we will form our desired (true) estimation.
2. continuing enough iterations our probability estimations will be close to real machines winning probability, but UCB just select the greediest action to maximize the upper confidence bound.
3. As we know Modified Policy Iteration can achieve better performance than policy Iteration and value iteration. However, what we can do to find optimum k automatically for Modified Policy Iteration? Propose two applicable ideas one considering unlimited and one limited computational power.

Answer:

1. If we have unlimited computational power, we can set k to a large number. Like 10000. Then we can get a best value to improve the policy.
2. If we just have limited computational power, the k can set a tiny nember. Like 5. Then It can quickey do a few Evaluation and Improve the policy.
3. Implement the MC Epsilon Greedy algorithm with only one of the UCB **or** Thompson sampling algorithms and compare results with the MC Epsilon Greedy algorithm (just print optimal policy for both algorithms after running to see if results).

**Note:** You already extended the given class example “Monte-Carlo -Exploring Starts” to the “MC Epsilon Greedy algorithm” as homework. Now you can use the same codes to answer this question.

Answer:

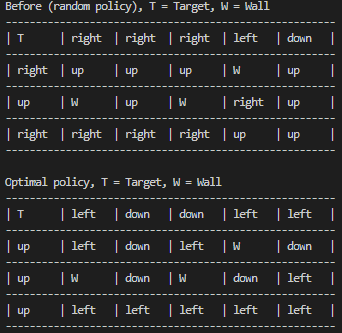
        self.Max\_iteration = 100

        self.gamma = 0.9

        self.Horizon = 15 #Max episode\_length

        self.epsilon = 0.2

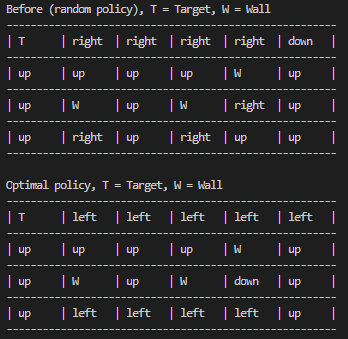
1. Epsilon Greedy algorithm:





Not Good.

1. Add UCB in MC Epsilon Greedy



Perfact!

1. Optimal policy in Epsilon Greedy algorithm.

