Quantum Machine Learning, Homework 1

- Author: NCTU 0712238 Yan-Tong Lin
- Lecturer: Dr. Alexey Melnikov

Note for “Projective simulation for artificial intelligence”

- https://arxiv.org/abs/1104.3787 (https://arxiv.org/abs/1104.3787)
  - Projective simulation for artificial intelligence
- https://arxiv.org/pdf/1504.02247 (https://arxiv.org/pdf/1504.02247)
  - Projective simulation with generalization

Model

- Policy
  - $P^{(t)}(a \mid s)$
- How to learn a better policy? By ECM(episodic compositional memory).
- ECM
  - a network of episodes (or clips), which are sequences of remembered percepts and actions.
  - (i) Encounter of percept $s \in S$ which happens with a certain probability $P^{(t)}(s)$. The encounter of percept s triggers the excitation of memory clip $c \in C$ according to a fixed **input-coupler** probability function $I(c \mid s)$.
  - (ii) Random walk through memory/clip space C, which is described by conditional probabilities $p^{(t)}(c' \mid c)$ of calling/exciting clip c' given that c was excited.
  - (iii) Exit of memory through activation of action a, described by a fixed **output-coupler** function $O(a =\mid c)$

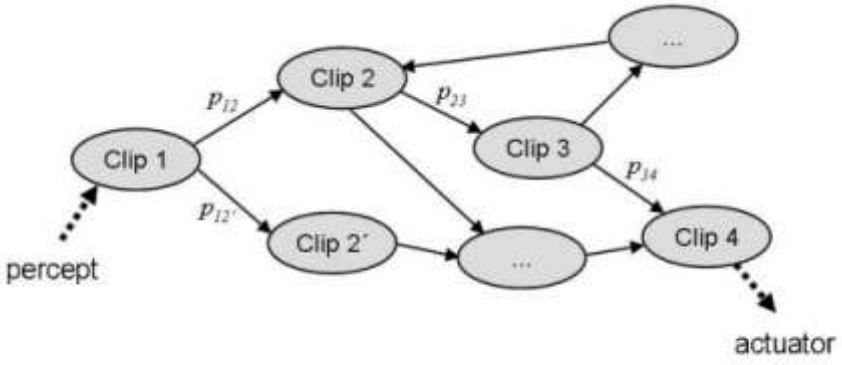


Figure 2 | Model of episodic memory as a network of clips.

- percepts and actions as product of features
  - *Percept space:*
    $s \equiv (s_1, s_2, \ldots, s_N) \in S_1 \times \cdots \times S_N \equiv S$, $s_i = 1, \ldots, |S_i|$. The structure of the percept space S, a cartesian product of sets, reflects the compositional (categorical) structure of percepts (objects). For example, s_1 could label the category of shape, s_2 category of color, s_3 category of size, etc. The maximum number of distinguishable input states is given by the product $|S| = |S_1| \cdots |S_N|$.

  - *Actuator space:*
    $a \equiv (a_1, a_2, \ldots, a_M) \in A_1 \times \cdots \times A_M \equiv A$, $a_j = 1, \ldots, |A_j|$. The structure of the actuator space A reflects the categories (or, in physics terminology, the degrees of freedom) of the agent's actions. For example a_1 could label the state of motion, a_2 the state of a shutter, a_3 the state of a warning signal, etc. All of this depends on the specification of the agent and the environment. The maximum number of different possible actions is given by the product $|A| = |A_1| \cdots |A_M|$.

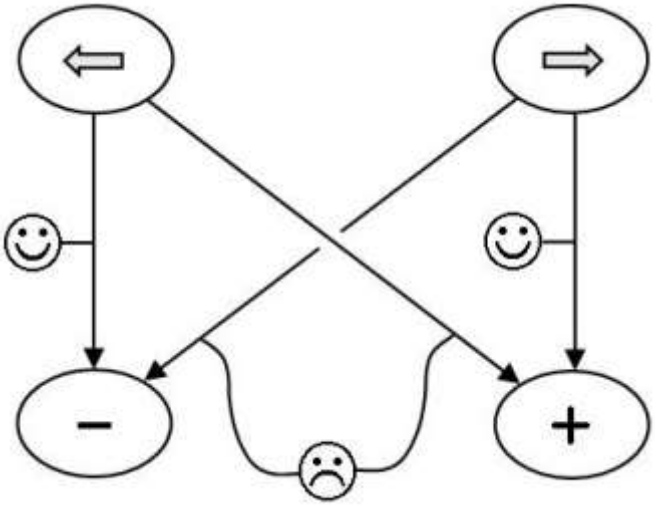
- clips are remembered percect/action time-sequences, in the following we consider clips with length 1 for simplicity
  - *Clip space:*

    $c \equiv (c^{(1)}, c^{(2)}, \ldots, c^{(L)}) \in C$; $c^{(l)} \in \mu(S) \cup \mu(A)$. The index L specifies the length of the clip. A simple example for $L = 2$ is the clip $c = (\mu(s), \mu(a)) \equiv$ (🧐,@), which corresponds to a simple percept-action pair. Clips of length $L = 1$ consist of a single remembered percept or action, respectively. In the subsequent examples, we will mainly consider probabilistic networks of such simple clips.

- emotions are remembered rewards
  - *Emotion space:*
    $e \equiv (e_1, e_2, \ldots, e_K) \in E_1 \times \cdots \times E_K \equiv E$, $e_k = 1, \ldots, |E_k|$. In the simplest case $K = 1$ and $|E_1| = 2$, with a two-valued emotion state $e_1 \equiv e \in \{$☺, ☹$\}$. Emotional states are *tags*, attached to transitions between different clips in the episodic memory. The state of these tags can be changed through feedback (e.g. reward) from the environment. They are internal parameters and should be distinguished from the reward function itself, which is defined externally. Informally speaking, emotional states are *remembered rewards* for previous actions, they have thus a similar status as the clips.

- true reward function $\Lambda : S \times A \to I \in \mathbb{R}$

  The reward function Λ is a mapping from $S \times A$ to $I \subset \mathbb{R}$ (real numbers), where in most subsequent examples we consider the case $I = 0, 1, \ldots, \lambda$. In the simplest case, $\lambda = 1$: If $\Lambda(s, a) = 1$ then the transition $s \to a$ is rewarded; if $\Lambda(s, a) = 0$, it is not rewarded. A rewarded (unrewarded) transition will set certain emotion tags in the episodic memory to ☺ (☹), as discussed previously. We shall also consider situations where the externally defined reward function changes in time, which leads to an adaptation of the flags in the agent's memory.



An Instance of Update Rules

- frequency rule
  - here s, a can be replaced by c_i, c_j to be more general
  - $h^{(1)}(s,a) = 1 \forall s \in S, a \in A$
  - $P^{(n)}(a \mid s) = \frac{h^{(n)}(s,a)}{h^{(n)}(s)}$ where $h^{(n)}(s) = \sum_a h^{(n)}(s,a)$
    - can use softmax to avoid a negative probability
  - let $\lambda^{(n)} = \Lambda(s^{(n)}, a^{(n)})$
  - $h^{(n+1)}(s,a) = h^{(n)}(s,a) + \lambda^{(n)}\delta(s,s^n)\delta(a,a^n) - \gamma(h^{(n)}(s,a) - 1))$
    - the γ part means forgetting part (decay to uniform)
      - not used for non-changing MDPs
    - δ is the delta function (to decide whether the s, a pair is the current pair)
- to model delayed reward: **glow**
  - $h^{(n+1)}(s,a) = h^{(n)}(s,a) + \lambda^{(n)}g^{(n+1)}(s,a) - \gamma(h^{(n)}(s,a) - 1))$
    - $g^{(n+1)}(s,a) = 1$ if s, a was traversed at time step n
    - $g^{(n+1)}(s,a) = (1-\eta)g^{(n)}(s,a)$ otherwise
  - consider delayed reward 100 at time step 3 when a, b is traversed, the edge c, d was traversed earlier in time step 1 with $\eta = 0.1$ then edge c, d by definition will be rewarded with increase in h value by $0.9^2 \times 100$
- choose of meta-params
  - γ is not required for non-chaning environment (e.g. MDP)
  - In theory, we can choose proper η to make the policy to converge to optimal

Reflection time R

- here $\rightarrow$ means activate with prob
- if $R = 1$
  - $\mu(s) \overset{p}{\rightarrow} \mu(a)$

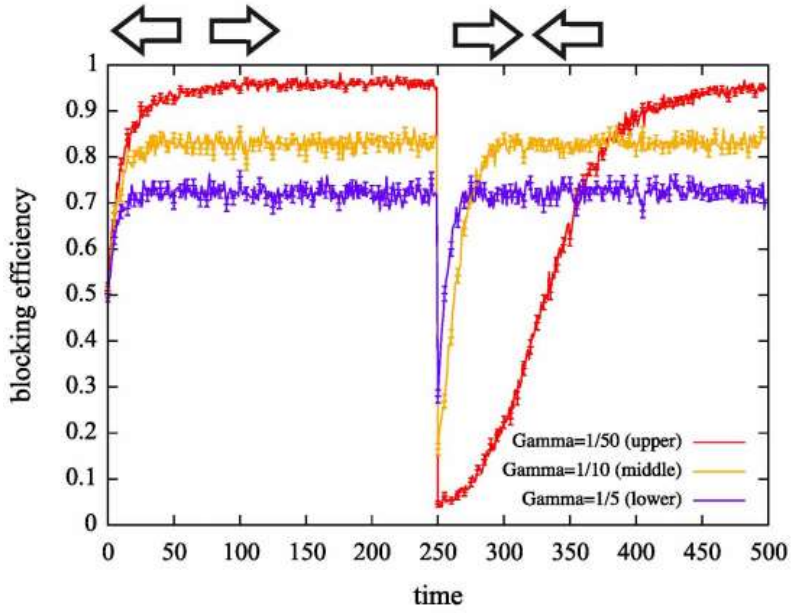
- if $R > 1$
  - same as $\mu(s) \xrightarrow{p} \mu(a)$
  - but if the emotion assigned to $\mu(s), \mu(a)$ is not good enough, repeat til the R_{th} time

Efficiency, the Learning Time / Maximal Efficiency Trade-off

success probability (averaged over different percepts, i.e. symbols shown by the attacker). After the nth round, the blocking efficiency is thus given by

$$r^{(n)} = \sum_{s \in S} P^{(n)}(s) P^{(n)}(a_s^*|s). \qquad (10)$$

In a similar way one can define the *learning time* $\tau(r_{th})$ for a given strategy as the time it takes on average (over an ensemble of identical agents) until the blocking efficiency reaches a certain threshold value r_{th}.



Q1

Go through the material which we considered during the lectures. Think of what is not clear and ask questions.

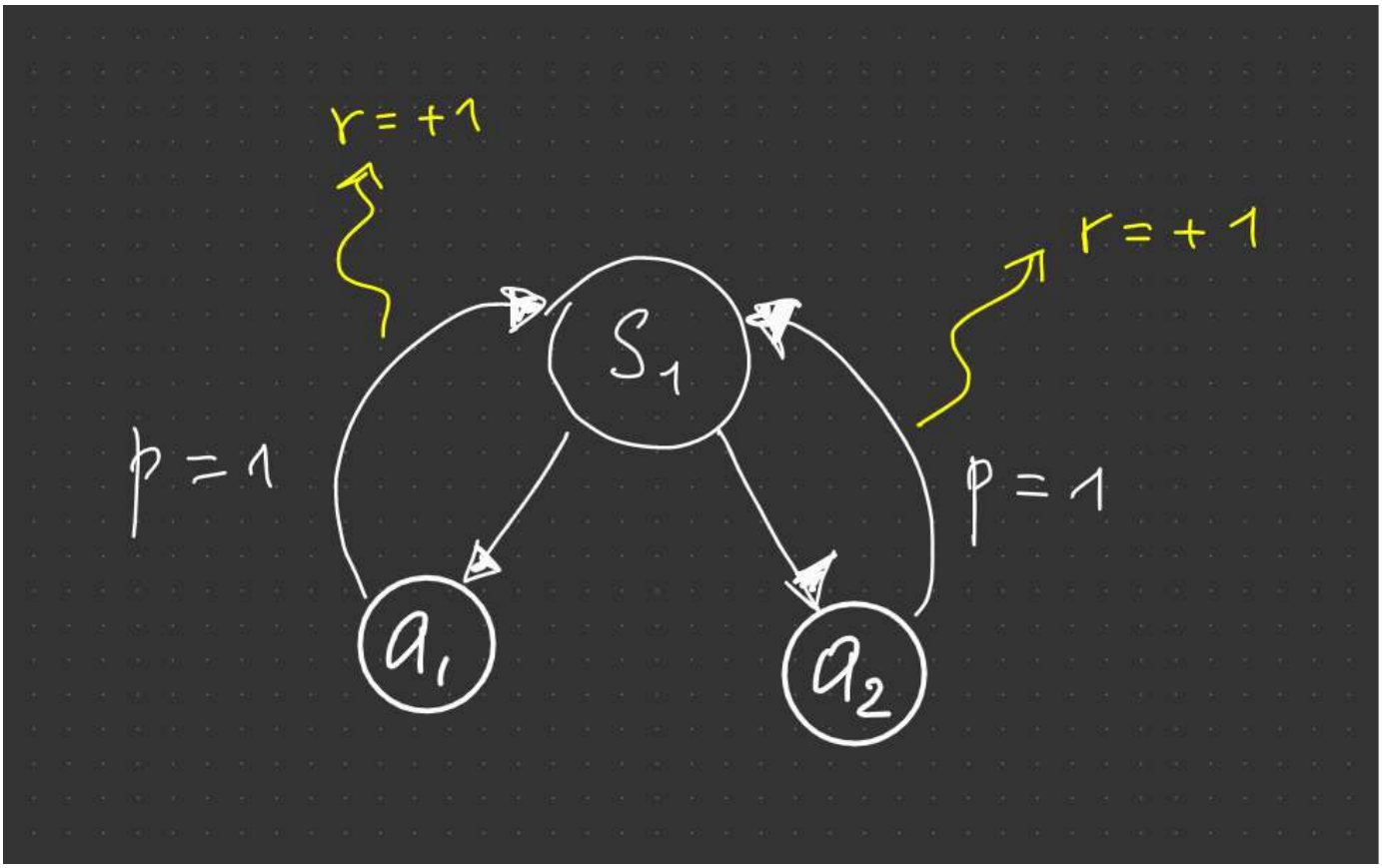
A1

- I go see the original paper for a more rigorous description instead. A note is shown above. If I made any mistake in my note, please let me know. Thanks!
- The equation 5 in the paper "Projective simulation with generalization" is different from the one in the slide by λ^t or λ^{t+1}, is there a reason to define them differently?

- The definition of glow seems confusing at the first glace, I think that adding the time step (i.e. $g^{(n+1)}(s,a)=1$ if s,a was traversed "at time step n") will make it more clear.

Q2 + Q3

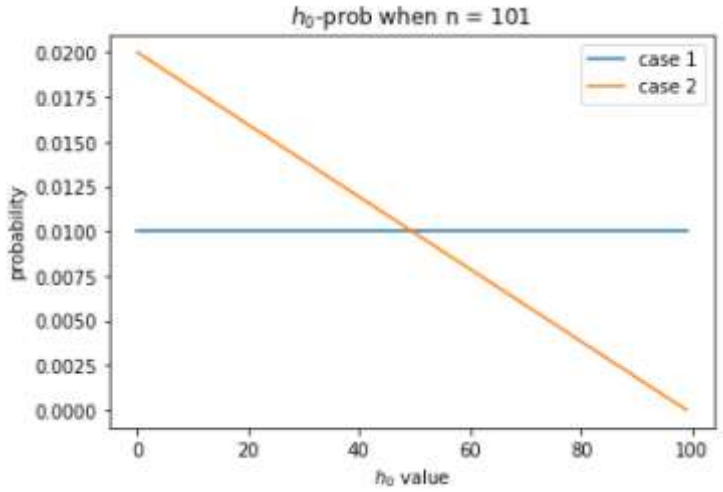
Let us consider the projective simulation (PS) agent.
The agent has a memory construction with only percept clips and action clips. No other clips are created in the PS memory. The learning algorithm is such that it is the simplest (the first rule that we studied): there is no forgetting, and there is no glow.



1. How this will this agent perform in an environment described by the following MDP?
2. How will the h-values change in time?
3. What are the asymptotic h-values at t=infinity?
4. What are the asymptotic probabilities of all the actions?
5. Will all agents behave the same, or some agents will have a preference towards a particular action?

Answer the same questions as above given that the h-values are not initialized as h(0)=1 at time step t=0, but h(0)=1 for action a1 and h(0)=2 for action a2.

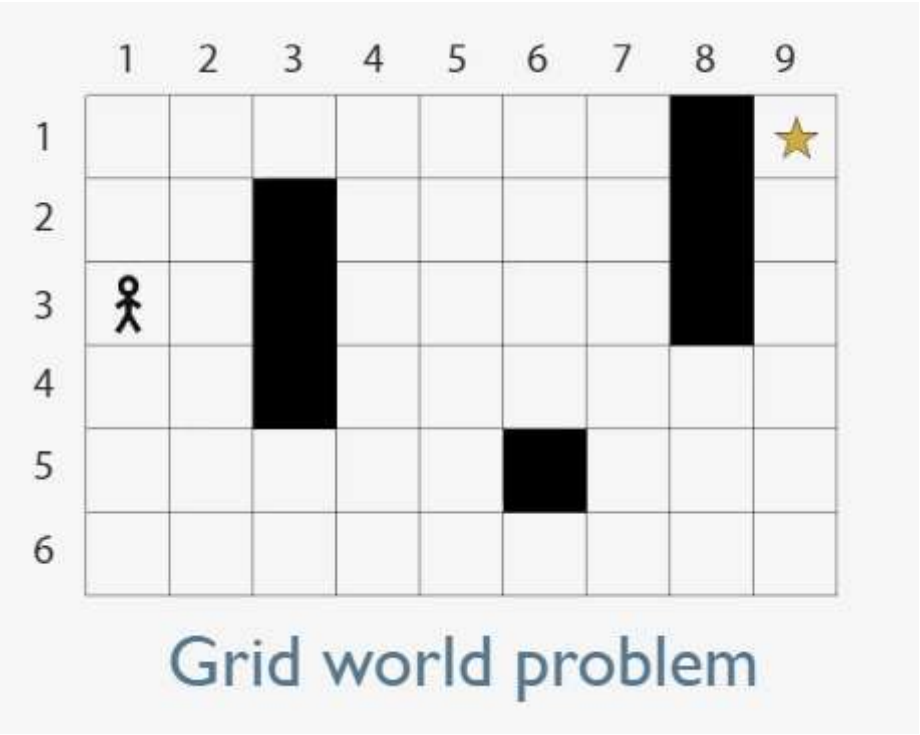
A2 + A3, a DP solution using jyputer notebook



1. According to the model description and I will discuss how its h value changes as time goes to infinity
2. The dynamic programming transition shows it
   - $P(a,b) = \frac{a-1}{a+b-1}P(a-1,b) + \frac{b-1}{a+b-1}P(a,b-1)$
3. What are the asymptotic h-values at t=infinity?
   - the figure clearly shows the two cases
   - case 1
     - any h value pairs are equally possible
     - $[h_0] = \frac{1}{2}n + 1$
   - case 2
     - A h value pair with lower $h(0)$ is more possible
     - $[h_0] = \frac{1}{3}n + 1$
4. What are the asymptotic probabilities of all the actions?
   - case 1
     - $\frac{1}{2}, \frac{1}{2}$
   - case 2
     - $\frac{1}{3}, \frac{2}{3}$
5. Will all agents behave the same, or some agents will have a preference towards a particular action?
   - Agents will have their preferences, with uniform prob in case 1, and with higher prob to prefer action 2 in case 2

Q4

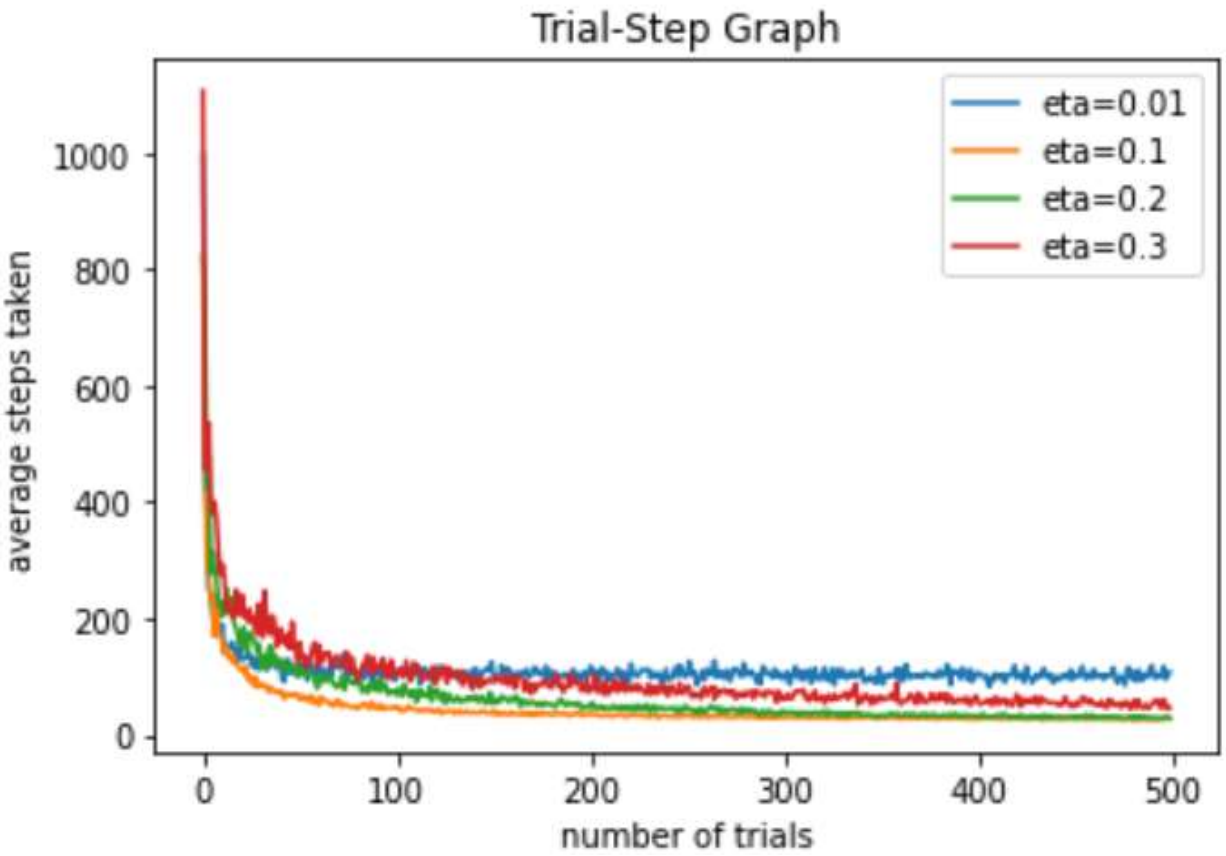
Consider the grid-world problem.



Program PS agent in this environment (the same maze as in the presentation slides containing 54 positions). Use glow in this environment and no forgetting. Which meta parameter do you find to be optimal?
Use any programming language you prefer.

A4

- The optimal can be defined on ourself. (confirmed with Dr. Alexey Melnikov)
- In this case, my observation is that agents with different ηs in certain range all converge to η as the number of trials grow.
- For the observation, I would define the optimal as "the η that converges to the optimal policy with high probability with the least number of trials"
- In practice, I plot a $\mathrm{number\ of\ trials}$ - $\mathrm{average\ steps}$ graph to show which value among $\{0.01, 0.1, 0.2, 0.3\}$ is optimal



    - note: the population is set to 50
  - By the graph, it seems that $\eta = 0.1, 0.2$ are good
  - My observation is that agents with larger η converge to smaller value, but takes more trials.
- Also, an example of the policy (from its h-values) of an $\eta = 0.2$ agent after 100 trials. Please note that the policy is optimal in the agent's mind does not means that it acted optimally during the training process.

```
In [252]: env = GridWorld()
          agent = PSAgent(54, 4, gamma=0.0, eta=.2)
          print(np.array(env.nogo, dtype=int))
          sim_grid(agent, env, T=100)
          pi = show_policy(agent.h, env.nogo)
```

```
[[0 0 0 0 0 0 0 1 0]
 [0 0 1 0 0 0 0 1 0]
 [0 0 1 0 0 0 0 1 0]
 [0 0 1 0 0 0 0 0 0]
 [0 0 0 0 0 1 0 0 0]
 [0 0 0 0 0 0 0 0 0]]
→ → → → ↓ ↓ ↓ x ↑
→ ↓ x ↓ → ↓ ↓ x ↑
→ ↓ x ↓ → ↓ ↓ x ↑
↓ ↓ x → → → → → ↑
→ → → → ↑ x ↑ → ↑
↑ → ↑ ↑ ↑ → → → ↑
```

Notes

- Sicne I was at a conference (TAAI2020) from 12/3-5, I handed in the homework after the deadline. However, I completed all Q1-3 without any reference to the the lecture video on 12/4. And I consulted with the lecturer to make sure the "optimal" in Q4 is defined by

ourself and gave the “definition of my own” and came up with “how to program it” independently, before seeing the video.

- For the answer to question 4, one more thing I can to is to focus on the region $[0.1, 0.3]$ with more population and see which value actually yields optimal result (14 steps) with high probability. But since the computational cost is high for my computer and the core of the method is catched, I will omit this part.
- To see my answer with code, please refer to the other file that is generated from jupyter notebook.