A4 Q2

July 28, 2022 3:26 PM

a) optimal search strategy: visit each grid exactly once.

	a	b	c	d	e
1	24	9	10	11	12
2	23	8	1	2	13
3	22	7	*	3	14
4	21	6	5	4	15
5	20	19	18	17	16

Empty

Suppose the prize appears in each cell with the same possibility, so $p = \frac{1}{24}$

If the prize is at the i^{th} cell, then the discounted reward will be $\gamma^{i-1} = \gamma^i$

Thus, the average discounted award will be:

$$\sum_{i=1}^{24} p \cdot \gamma^i = \frac{1}{24} \sum_{i=1}^{24} \gamma^i$$

$$= \frac{1}{24} \left(\frac{1-\gamma^{25}}{1-\gamma} - \gamma^0 \right)$$

$$= \frac{1}{24} \left(\frac{\gamma - \gamma^{25}}{1-\gamma} \right)$$

$$= \frac{1}{24} \left(\frac{0.95 - 0.95^{25}}{1-0.95} \right)$$

$$\approx 0.5605$$

b)

	a	b	c	d	e
1			■	3	4
2			1	2	5
3	■	■	*	■	■
4					
5			■		

Four room

WLOG, suppose the agent choose the up right room. The prize still appers in the empty cells uniformly randomly: $p = \frac{1}{18}$

Similarly: the average reward will be:

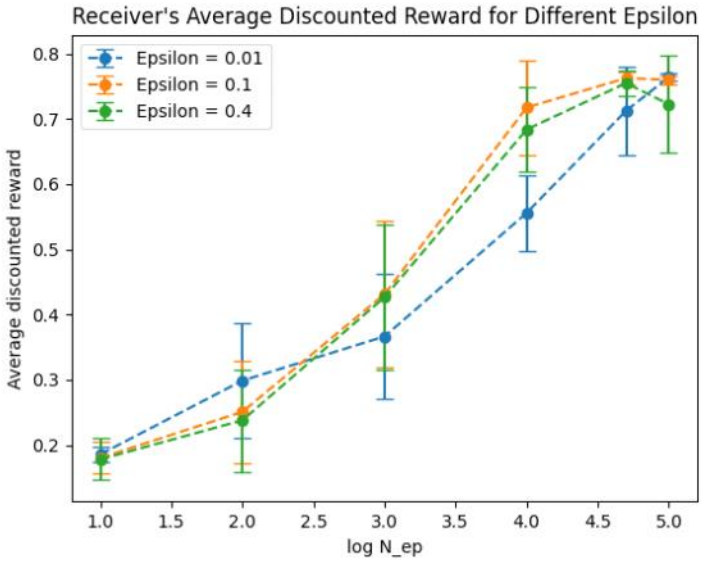
$$p(\gamma + \gamma^2 + \gamma^3 + \gamma^4 + \gamma^5)$$

$$= \frac{1}{18} \left(\sum_{i=1}^{5} \gamma^i \right)$$

$$= \frac{1}{18} \left(\frac{\gamma - \gamma^6}{1-\gamma} \right) = \frac{1}{18} \left(\frac{0.95 - 0.95^6}{1-0.95} \right) \approx 0.2388$$

The agent can randomly choose any room, but the average discounted reward will be the same: 0.2388 (We can also take the average for four rooms, but they have identical results so give the same answer)

C) All log in the following questions are base 10



Messages:

```
[[0 0 0 3 3]
 [0 0 3 3 3]
 [0 0 0 0 0]
 [1 1 1 2 2]
 [1 1 0 2 2]]
```

So:
Msg 0 means the prize is at the top left room
Msg 1 means the prize is at the bottom left room
Msg 2 means the prize is at the bottom right room
Msg 3 means the prize is at the top right room

Actions for message 0

	a	b	c	d	e
0	↓	←		↓	↓
1	→	↑	←	←	←
2			↑		
3	→	→	↑	←	↓
4	↑	↑		↑	←

Receiver's policy for msg 0, top left room

Actions for message 1

	a	b	c	d	e
0	↓	↓		←	↓
1	→	→	↓	←	←
2			↓		
3	→	↓	←	←	←
4	↑	←		↑	↑

Receiver's policy for msg 1, bottom left room

Actions for message 2

	a	b	c	d	e
0	←	↓		↓	←
1	→	→	↓	←	←
2			↓		
3	→	→	→	↓	←
4	→	↑		→	↑

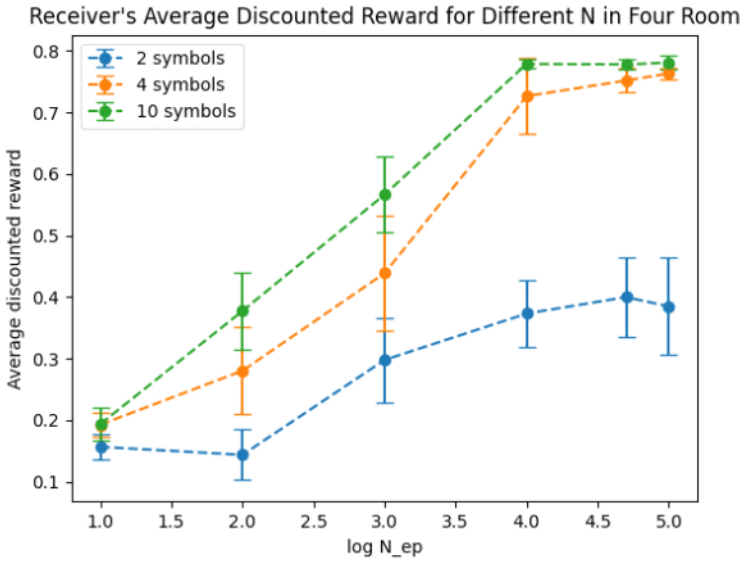
Receiver's policy for msg 2, bottom right room

Actions for message 3

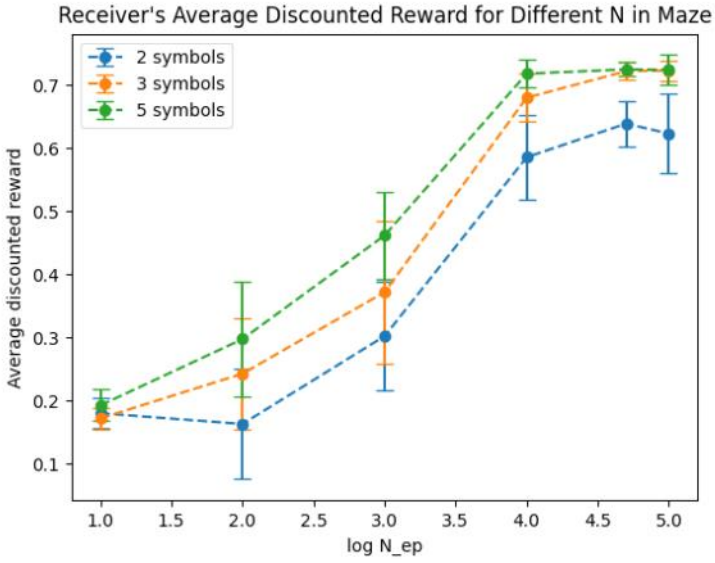
	a	b	c	d	e
0	↓	↓		↓	←
1	→	→	→	→	↑
2			↑		
3	→	→	↑	←	←
4	→	↑		↑	↑

Receiver's policy for msg 3, top right room

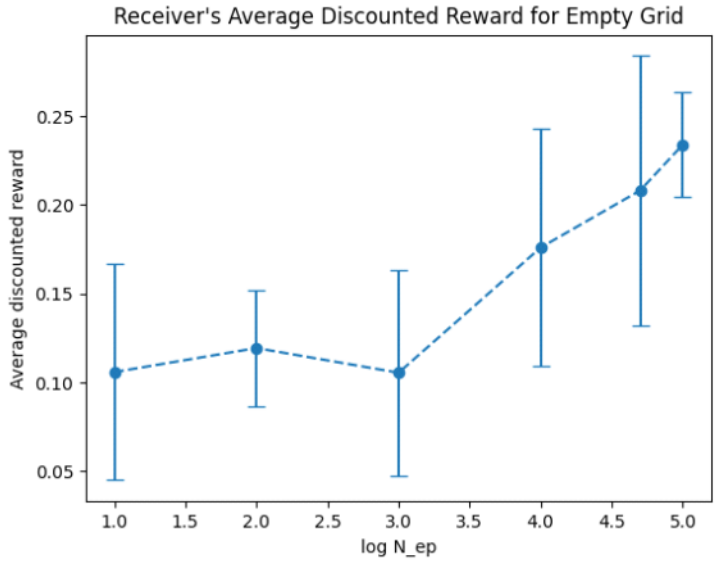
D)


Receiver's Average Discounted Reward for Different N in Four Room
2 symbols
4 symbols
10 symbols
Average discounted reward
log N_ep

E)


Receiver's Average Discounted Reward for Different N in Maze
2 symbols
3 symbols
5 symbols
Average discounted reward
log N_ep

F)



G)

Assuming that the avg discounted reward is approximately linear to log N_ep according to the graph we found in f)

By interpolating:

reward = 0.106 at N=10 log Nep = 1

reward = 0.234 at N = 10 0000 log Nep = 5

Slope $k = \frac{0.234 - 0.106}{5 - 1} \approx 0.032$

from a): the optimal award we found is 0.5605

so: $\log N_{opt} = 5 + \frac{0.5605 - 0.234}{0.032} \approx 15.203$

$\Rightarrow N_{opt} \approx 10^{15.203} \approx 1.6 * 10^{15}$

Although from the graph in f), it looks like the reward grows slowly for small Nep, and faster for large Nep, which makes it like linear from $\log Nep = 3$ to $\log Nep = 5$. If we only use that part for interpolation,

$k = \frac{0.234 - 0.108}{5 - 3} \approx 0.063$, $\log N_{opt} = (5 + \frac{0.5605 - 0.234}{0.063}) \approx 10.18$, $N_{opt} \approx 10^{10}$.

From the result for c.d.e, the curve all goes like a "sigmoid" shape, meaning growing slow at the beginning and the end. So this might be an overestimate.

So I probably need to run some iterations between 10^{10} and 10^{15} to get close enough to the optimal