ダラダラと探さないようにペナルティーを課す

		報酬		
割引率γ	0.7		-1	-1	-1
学習率α	0.5		-1	-1	
			-1	-1	100

episode	1

結果	到着	(エピソード報告)
到着数	1	

Greedy比

ε	0.98

$$\epsilon = 1 - \frac{\text{到着エピソード}}{50}$$

Explore用の確率表

		行動(Action)			
状態	仮	右	上	左	下
1	0	0.50	0.50	0.50	1.00
2	0	0.33	0.33	0.67	1.00
3	0	0.00	0.00	1.00	1.00
4	0	0.33	0.67	0.67	1.00
5	0	0.00	0.33	0.67	1.00
6					
7	0	0.50	1.00	###	###
8	0	0.33	0.67	1.00	1.00
到着	0	1.00	1.00	1.00	1.00

(空) 欄外

step 1 スタート

現Agent位置

行	1
列	1
状態	1 (s_t)

★	
	終

ε	0.98

初期値

状態	現Q値	アクション 右	上	左	下
	1	3.00	欄外	欄外	1.00
	2	4.00	欄外	1.00	5.00
	3	欄外	欄外	9.00	欄外
	4	2.00	6.00	欄外	5.00
	5	欄外	2.00	7.00	1.00
	6	欄外	欄外	欄外	欄外
	7	8.00	2.00	欄外	欄外
	8	8.00	1.00	8.00	欄外
	9(到着)	0	0	0	0

Q値は最初、適当

行動決定

G乱数	0.175	→	Explore
	最大Q	E乱数	Act候補
Exploit 時	3.00		1
Explore時		0.800	4

採用Act	4	(下) →
	(a_t)	

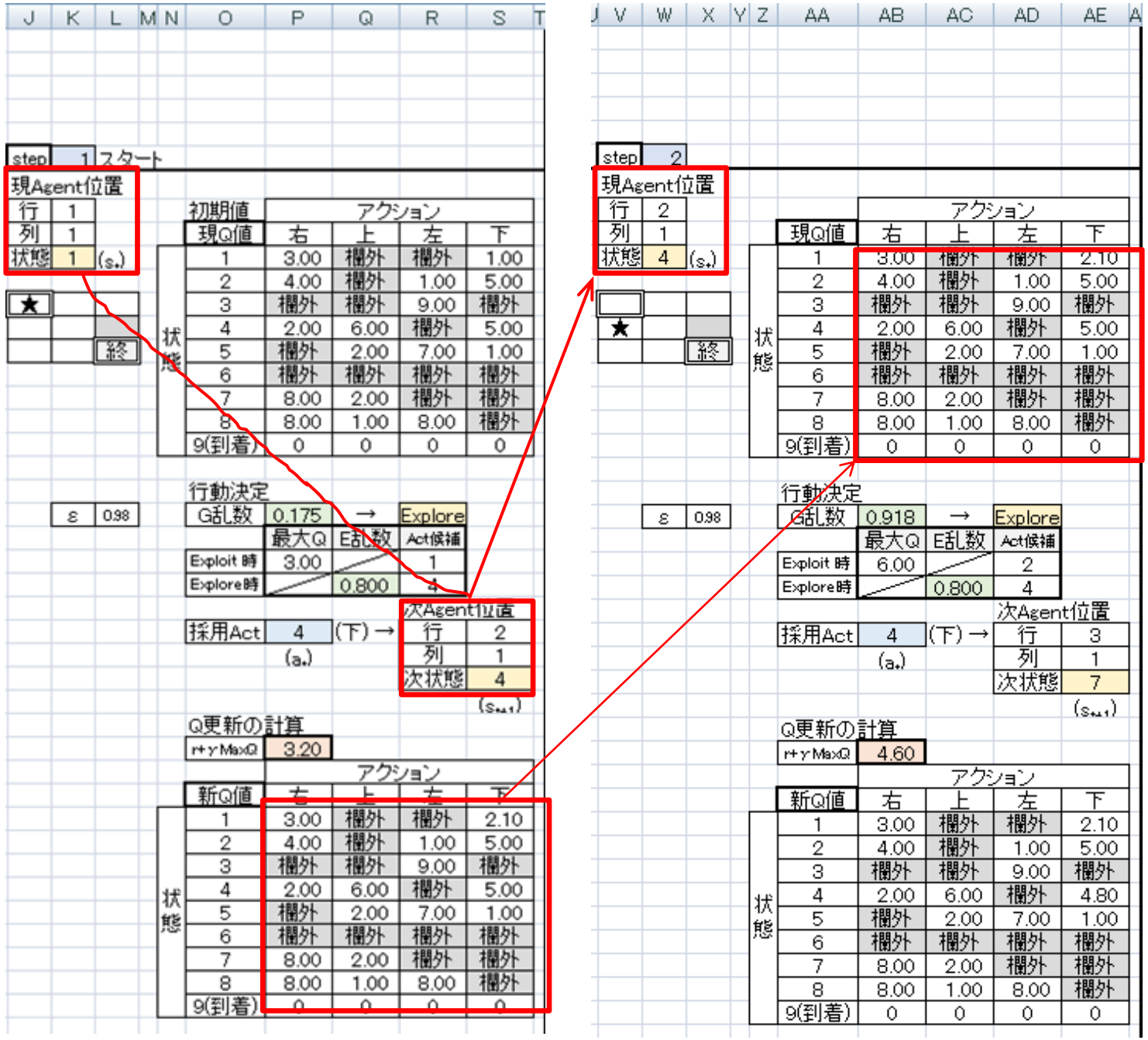
次Agent位置

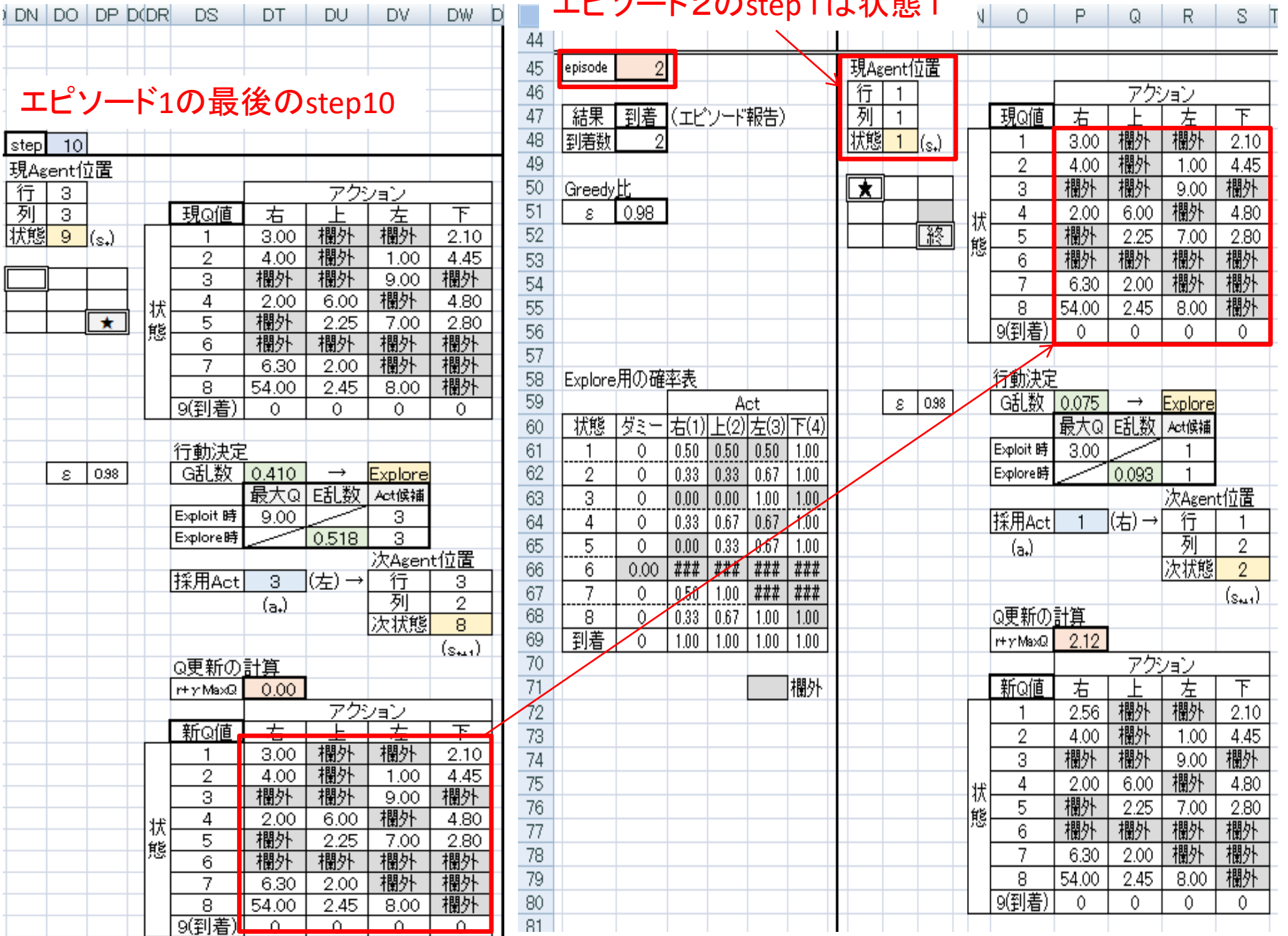
行	2
列	1
次状態	4
	(s_t+1)

Q更新の計算

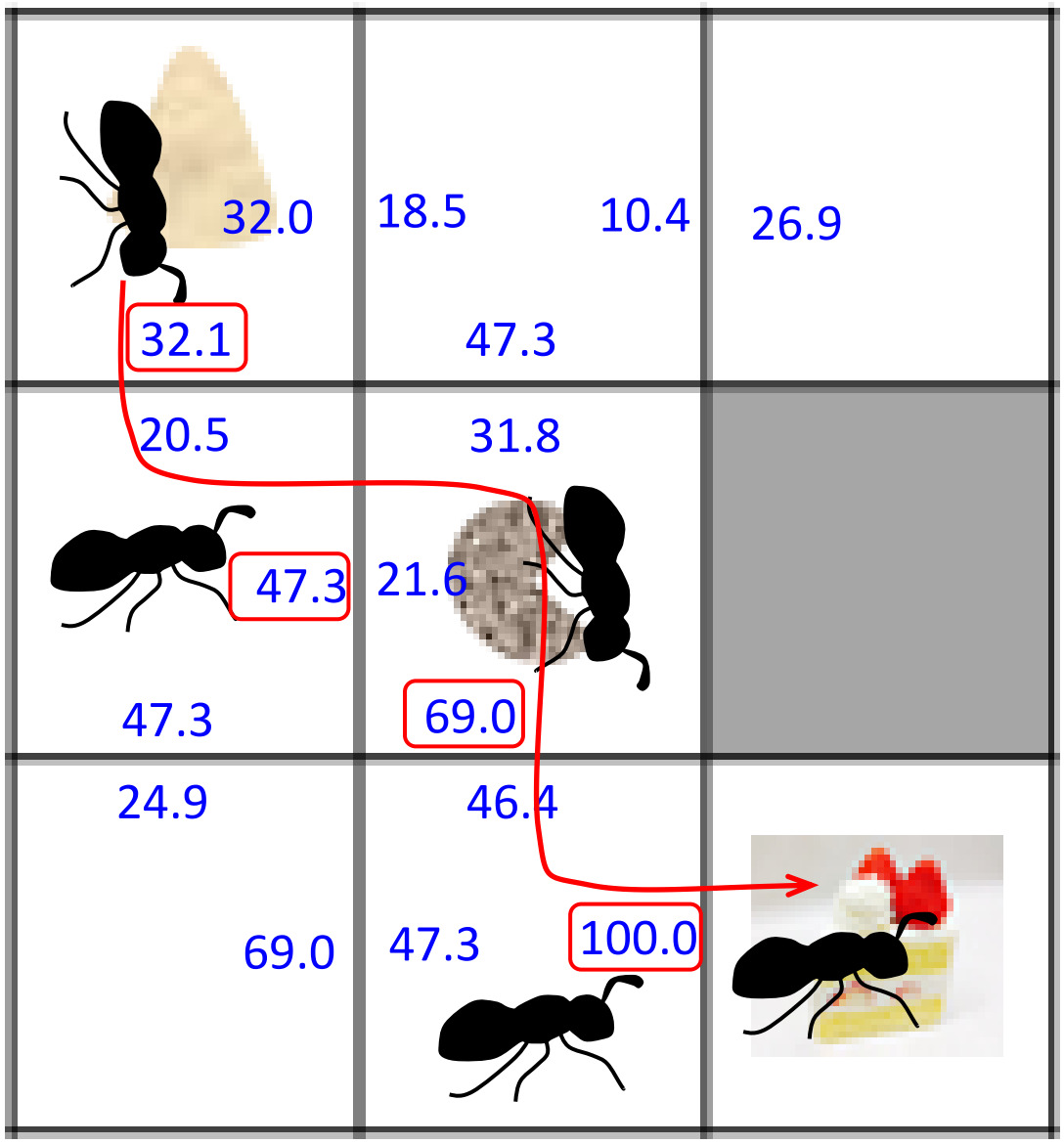
r+γMaxQ	3.20

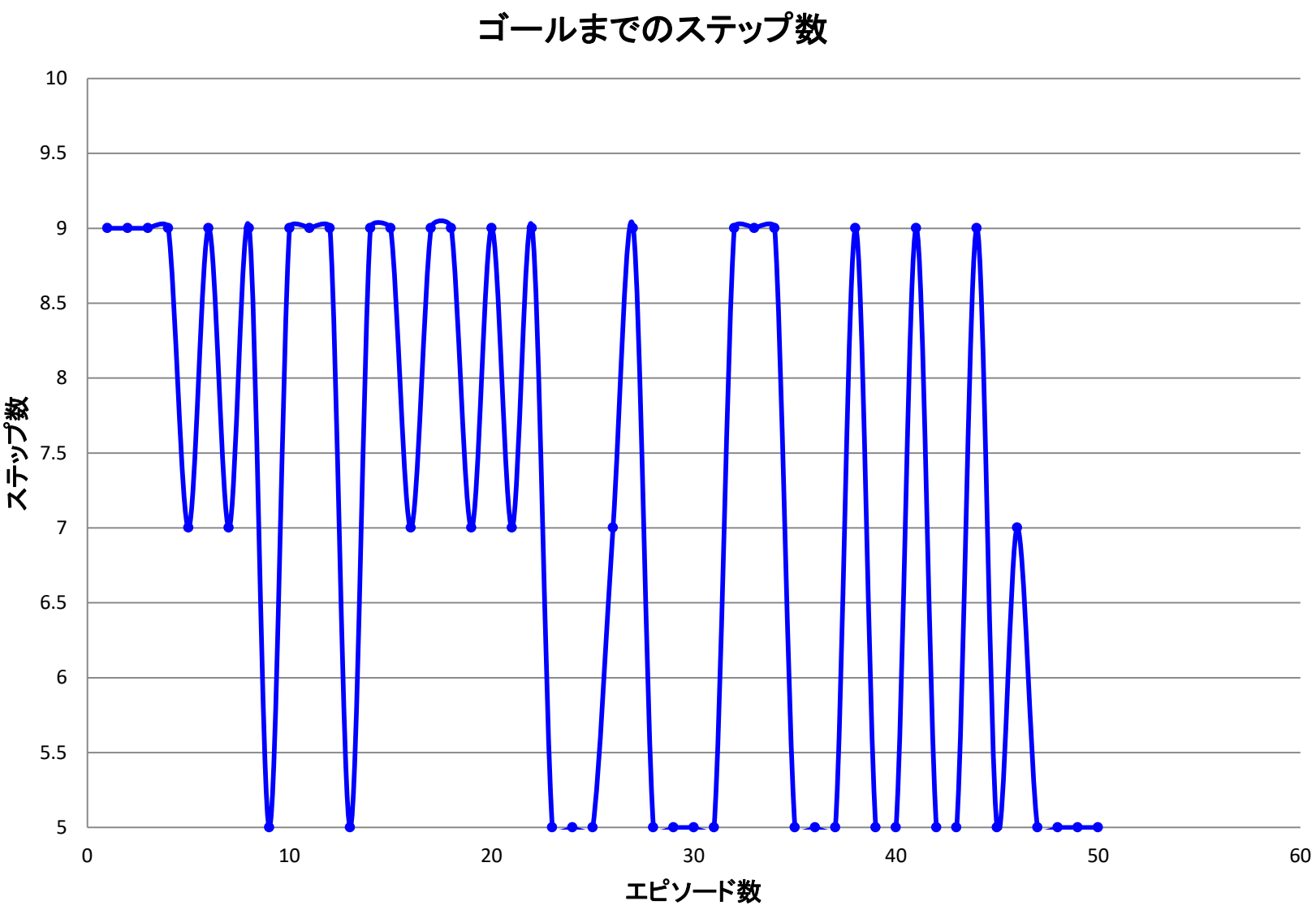
状態	新Q値	アクション 右	上	左	下
	1	3.00	欄外	欄外	2.10
	2	4.00	欄外	1.00	5.00
	3	欄外	欄外	9.00	欄外
	4	2.00	6.00	欄外	5.00
	5	欄外	2.00	7.00	1.00
	6	欄外	欄外	欄外	欄外
	7	8.00	2.00	欄外	欄外
	8	8.00	1.00	8.00	欄外
	9(到着)	0	0	0	0


step 1 スタート
現Agent位置
行 1
列 1
状態 1 (s.)
終
ε 0.98
初期値
アクション
現Q値 右 上 左 下
状態
1 3.00 欄外 欄外 1.00
2 4.00 欄外 1.00 5.00
3 欄外 欄外 9.00 欄外
4 2.00 6.00 欄外 5.00
5 欄外 2.00 7.00 1.00
6 欄外 欄外 欄外 欄外
7 8.00 2.00 欄外 欄外
8 8.00 1.00 8.00 欄外
9(到着) 0 0 0 0
行動決定
G乱数 0.175 → Explore
最大Q E乱数 Act候補
Exploit 時 3.00 1
Explore時 0.800 4
採用Act 4 (下) →
(a.)
次Agent位置
行 2
列 1
次状態 4
(s.+1)
Q更新の計算
r+γMaxQ 3.20
新Q値 右 上 左 下
状態
1 3.00 欄外 欄外 2.10
2 4.00 欄外 1.00 5.00
3 欄外 欄外 9.00 欄外
4 2.00 6.00 欄外 5.00
5 欄外 2.00 7.00 1.00
6 欄外 欄外 欄外 欄外
7 8.00 2.00 欄外 欄外
8 8.00 1.00 8.00 欄外
9(到着) 0 0 0 0
step 2
現Agent位置
行 2
列 1
状態 4 (s.)
終
ε 0.98
アクション
現Q値 右 上 左 下
状態
1 3.00 欄外 欄外 2.10
2 4.00 欄外 1.00 5.00
3 欄外 欄外 9.00 欄外
4 2.00 6.00 欄外 5.00
5 欄外 2.00 7.00 1.00
6 欄外 欄外 欄外 欄外
7 8.00 2.00 欄外 欄外
8 8.00 1.00 8.00 欄外
9(到着) 0 0 0 0
行動決定
G乱数 0.918 → Explore
最大Q E乱数 Act候補
Exploit 時 6.00 2
Explore時 0.800 4
採用Act 4 (下) →
(a.)
次Agent位置
行 3
列 1
次状態 7
(s.+1)
Q更新の計算
r+γMaxQ 4.60
アクション
新Q値 右 上 左 下
状態
1 3.00 欄外 欄外 2.10
2 4.00 欄外 1.00 5.00
3 欄外 欄外 9.00 欄外
4 2.00 6.00 欄外 4.80
5 欄外 2.00 7.00 1.00
6 欄外 欄外 欄外 欄外
7 8.00 2.00 欄外 欄外
8 8.00 1.00 8.00 欄外
9(到着) 0 0 0 0


エピソード1の最後のstep10
エピソード2のstep1は状態1
step 10
現Agent位置
行 3
列 3
状態 9 (s_t)
ε 0.98
行動決定
G乱数 0.410 → Explore
最大Q E乱数 Act候補
Exploit 時 9.00 3
Explore時 0.518 3
採用Act 3 (左) →
(a_t)
次Agent位置
行 3
列 2
次状態 8
(s_t+1)
Q更新の計算
r+γMaxQ 0.00
アクション
新Q値 右 上 左 下
1 3.00 棚外 棚外 2.10
2 4.00 棚外 1.00 4.45
3 棚外 棚外 9.00 棚外
4 2.00 6.00 棚外 4.80
5 棚外 2.25 7.00 2.80
6 棚外 棚外 棚外 棚外
7 6.30 2.00 棚外 棚外
8 54.00 2.45 8.00 棚外
9(到着) 0 0 0 0
状態
episode 2
結果 到着 (エピソード報告)
到着数 2
Greedy比
ε 0.98
Explore用の確率表
Act
状態 ダミー 右(1) 上(2) 左(3) 下(4)
1 0 0.50 0.50 0.50 1.00
2 0 0.33 0.33 0.67 1.00
3 0 0.00 0.00 1.00 1.00
4 0 0.33 0.67 0.67 1.00
5 0 0.00 0.33 0.67 1.00
6 0.00 ### ### ### ###
7 0 0.50 1.00 ### ###
8 0 0.33 0.67 1.00 1.00
到着 0 1.00 1.00 1.00 1.00
棚外
現Agent位置
行 1
列 1
状態 1 (s_t)
終
現Q値 右 上 左 下
1 3.00 棚外 棚外 2.10
2 4.00 棚外 1.00 4.45
3 棚外 棚外 9.00 棚外
4 2.00 6.00 棚外 4.80
5 棚外 2.25 7.00 2.80
6 棚外 棚外 棚外 棚外
7 6.30 2.00 棚外 棚外
8 54.00 2.45 8.00 棚外
9(到着) 0 0 0 0
ε 0.98
行動決定
G乱数 0.075 → Explore
最大Q E乱数 Act候補
Exploit 時 3.00 1
Explore時 0.093 1
採用Act 1 (右) →
(a_t)
次Agent位置
行 1
列 2
次状態 2
(s_t+1)
Q更新の計算
r+γMaxQ 2.12
新Q値 右 上 左 下
1 2.56 棚外 棚外 2.10
2 4.00 棚外 1.00 4.45
3 棚外 棚外 9.00 棚外
4 2.00 6.00 棚外 4.80
5 棚外 2.25 7.00 2.80
6 棚外 棚外 棚外 棚外
7 6.30 2.00 棚外 棚外
8 54.00 2.45 8.00 棚外
9(到着) 0 0 0 0

数字はQ値


32.0
18.5
10.4
26.9
32.1
47.3
20.5
31.8
47.3
21.6
47.3
69.0
24.9
46.4
69.0
47.3
100.0


ゴールまでのステップ数
10
9.5
9
8.5
8
7.5
7
6.5
6
5.5
5
ステップ数
0
10
20
30
40
50
60
エピソード数