S	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)
V_0	0	0	5	0	0	-5
V_1	0	0	5	0	$0{,}8 \times (0{,}9 \times 5) = 3{,}6$	-5
V_2	0	$0{,}8 \times (0{,}9 \times 3{,}6) + 0{,}1 \times (0{,}9 \times -5) = 2{,}14$	5	$0{,}8 \times (0{,}9 \times 3{,}6) = 2{,}59$	$0{,}8 \times (0{,}9 \times 5) + 0{,}1 \times (0{,}9 \times 3{,}6) = 3{,}92$	-5

2) Policy را با توجه به جدول ارزش‌های بالا محاسبه کنید (برای خانه‌هایی که دو یا چند Action با مقدار

2,59	~~3,6~~ 3,92	+5
S	2,14	-5

طبق فرمول داده شده نسبت راهنمایی طبیعی:

$$V_1((2,2)) = 0.8(0 + 0.9(5)) + 0.1(0 + 0.9(0)) + 0.1(0 + 0.9(0))$$
$$= 3.6$$

$$V_2((2,1)) = 0.8(0 + 0.9(3.6)) + 0.1(0 + 0.9(0)) + 0.1(0 + 0.9(0))$$
$$= 2.59$$

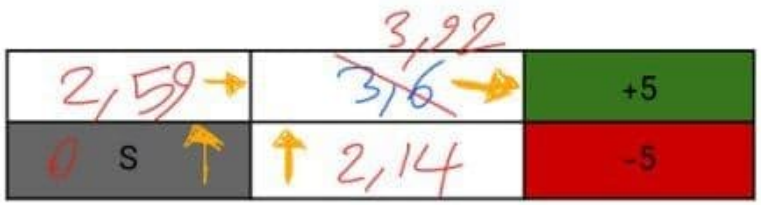
$$V_2((1,2)) = 0.8(0 + 0.9(3.6)) + 0.1(0 + 0.9(-5)) + 0.1(0 + 0.9(0))$$
$$= 2.14$$

به سمت بالا برود از همینجای ماند

$$V_2((2,2)) = 0.8(0 + 0.9(5)) + 0.1(0 + 0.9(3.6)) + 0.1(0 + 0.9(0))$$
$$= 3.92$$

2

S	(1,1)	(1,2)	(1,3)	(2,1)	(2,2)	(2,3)
π*(S)	up ↑	up ↑	-	Right →	Right →	-



2,59 > 2,14 $\xrightarrow{(1,1)}$ so move up

3,92 > -5 $\xrightarrow{(1,2)}$ so move up

3,92 > 9 $\xrightarrow{(2,1)}$ so move right

5 > 2,14 $\xrightarrow{(2,2)}$ so move right

3 با توجه به

2,59	3,92 (~~3,76~~)	+5
S	2,14	-5

و الگوریتم monte carlo :

Summing all the rewards coming after the first visit to (S_t)

(1,1): For episode E_1 : $G = 2{,}14 + (-5) = -2{,}86$

" " E_2 : $G = 2{,}14 + 3{,}92 + 5 = 11{,}06$

" " E_3 : $G = 2{,}59 + 3{,}92 + 5 = 11{,}51$

$$V((1,1)) = \frac{-2{,}86 + 11{,}06 + 11{,}51}{3} = 6{,}57$$

(2,2): For episode E_1 : $G =$ null

" " E_2 : $G = 5$

" " E_3 : $G = 5$

$$V((2,2)) = \frac{5+5}{3} = 3{,}33$$

4 در این قسمت فرض کردم که agent با exploration policy شروع می‌کند، یعنی action ای که انتخاب می‌شود، رندوم خواهد بود.

-1	-1	5
	-1	-5

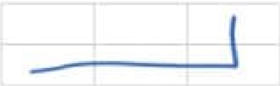
s	V(s)
(1,1)	0
(1,2)	0
⋮	⋮
(2,3)	0

برای بکاپ Value ها داریم:

$$V(S_1)_{obs} = R(S_2) + V(S_2)$$

$$TD_{Error} = V(S_1)_{obs} - V(S_1)_{exp}$$

$$V(S_1) = V(S_1) + \alpha \times TD_{Error}$$



حال دو مسیر زیر را در نظر می‌گیریم :

$V((1,1)) = -0.1$

$V((1,2)) = -0.1$

$V((1,3)) = -5 + 0.1(4.5+5)$

$= -4.05$

بنابراین داریم :

S	V(S)
(1,1)	−0.1
(1,2)	−0.1
(1,3)	−4.05
(2,1)	0
(2,2)	0
(2,3)	0

$v((1,1)) = -0,3$

$v((2,1)) = -0,1$

$v((2,2)) = -0,1$

$(1,1)$	$-0,3$
$(1,2)$	$-0,1$
$(1,3)$	$-4,05$
$(2,1)$	$-0,1$
$(2,2)$	$-0,1$
$(2,3)$	0

بنابراین طبیعی