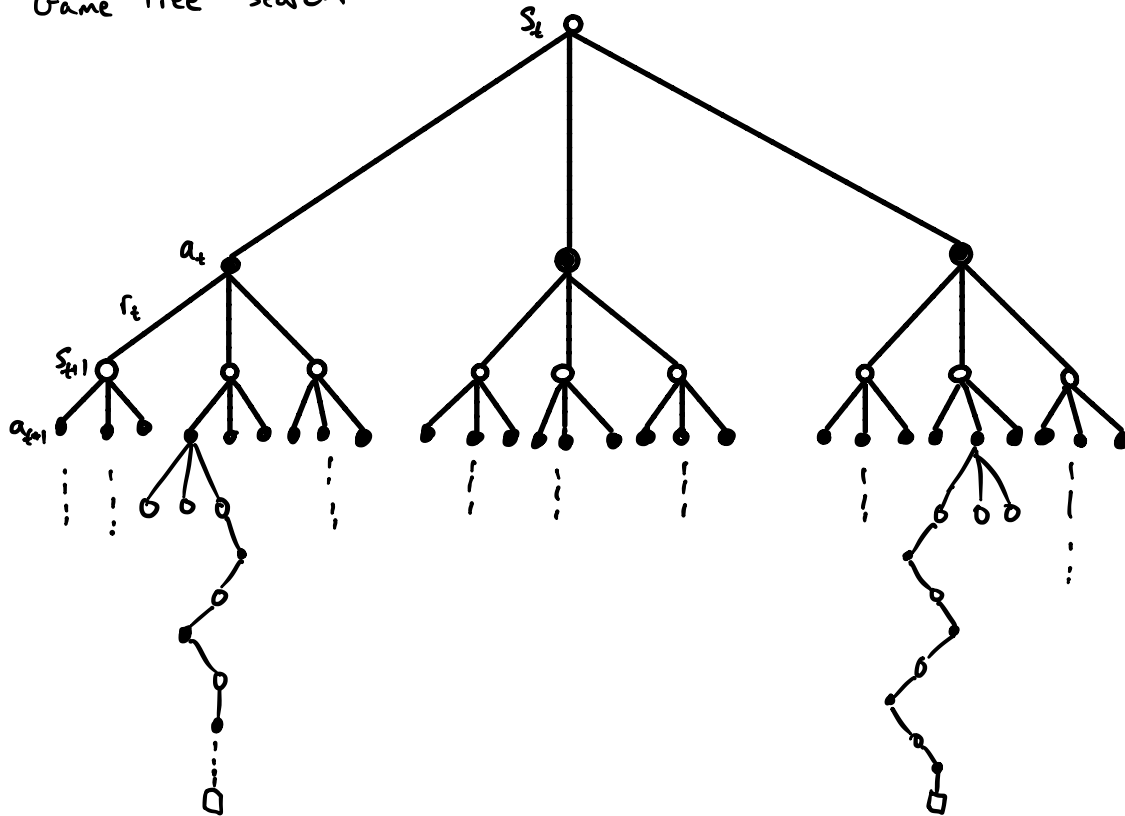


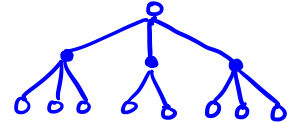
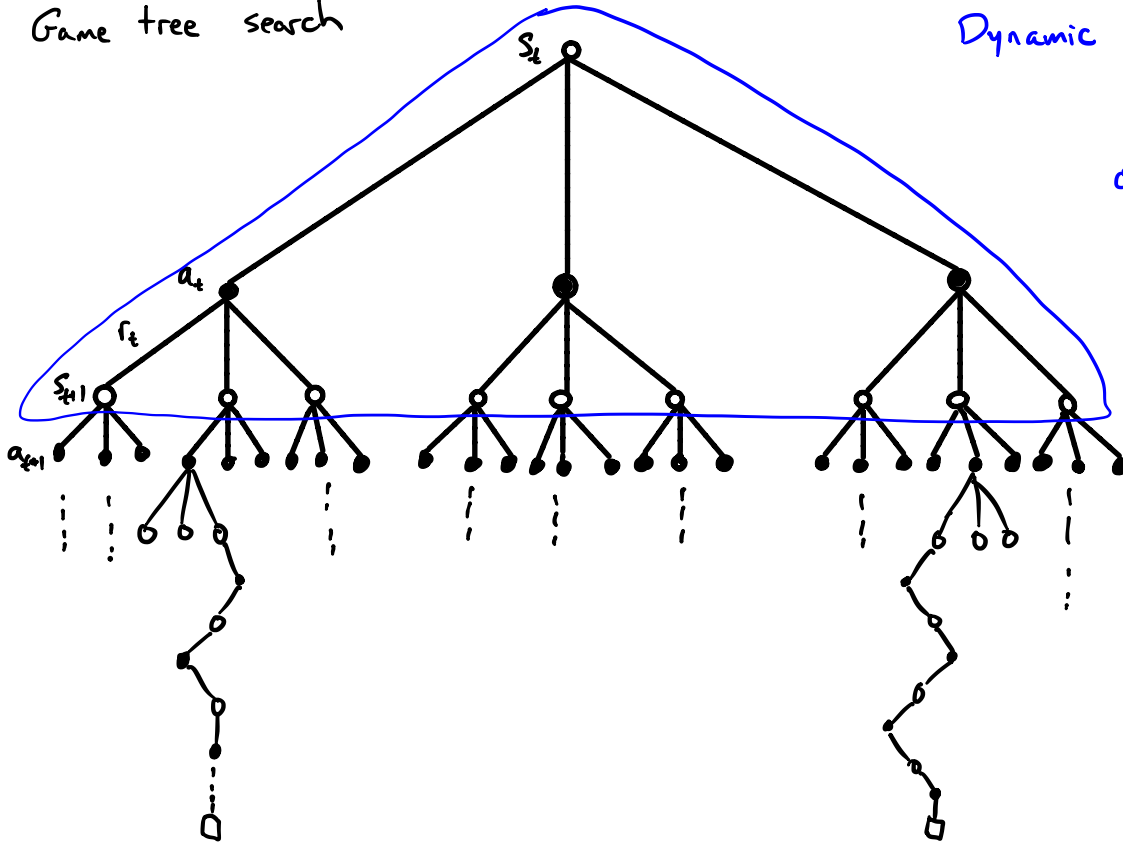


# Game tree search



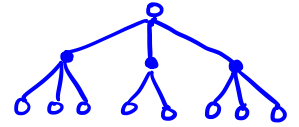
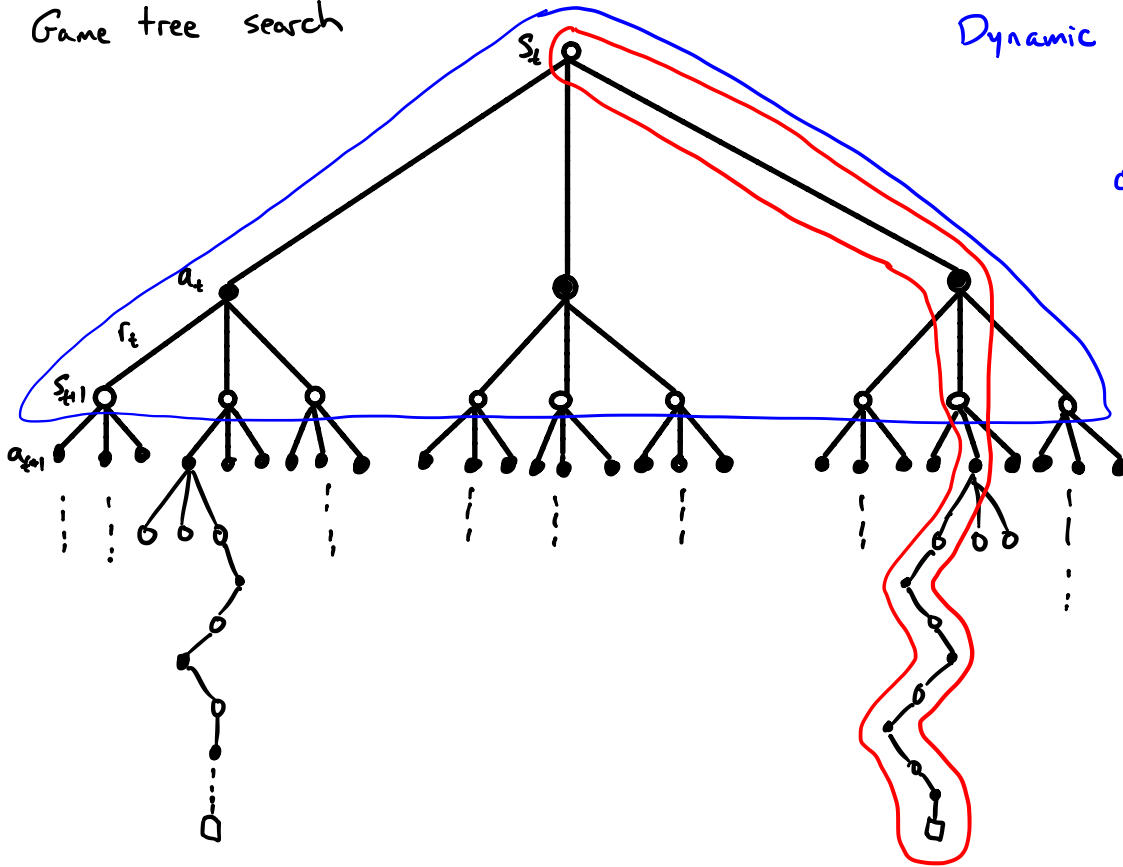
Game tree search

Dynamic Programming (DP)



Game tree search

Dynamic Programming (DP)

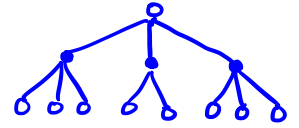
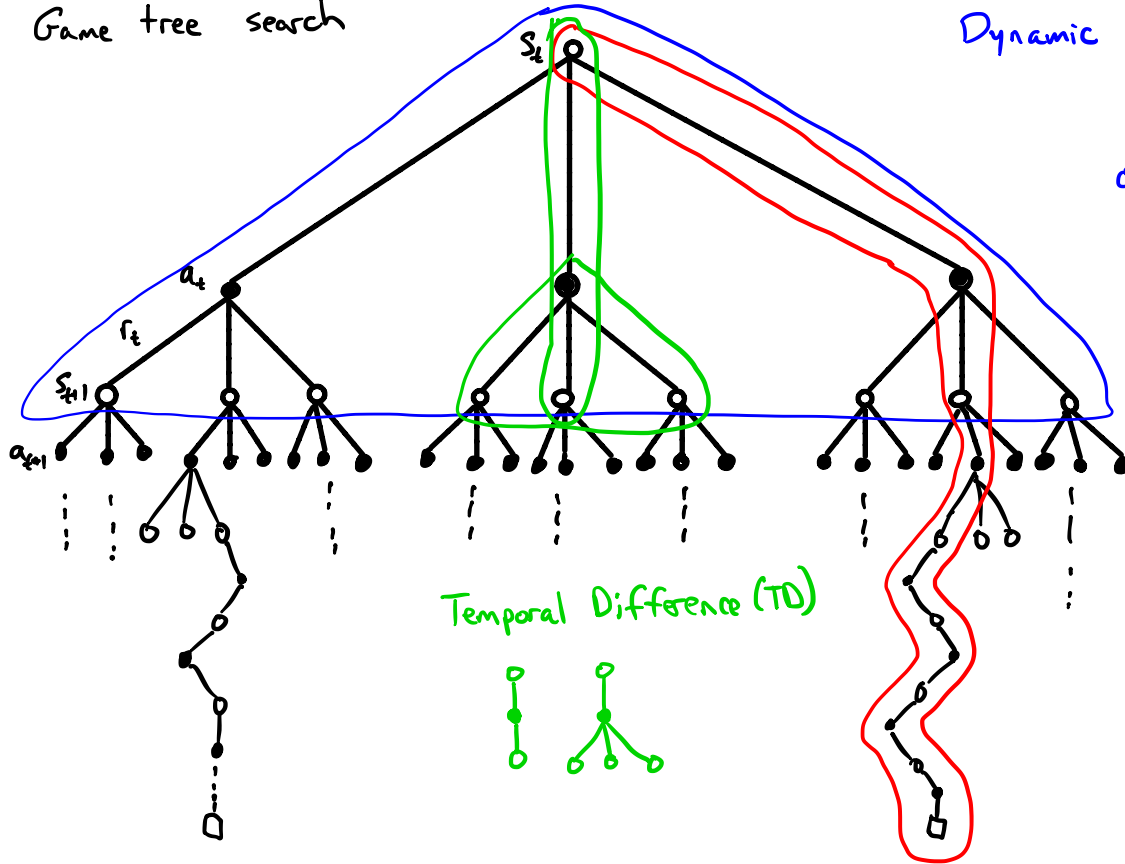


Monte Carlo (MC)



Game tree search

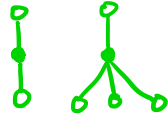
Dynamic Programming (DP)

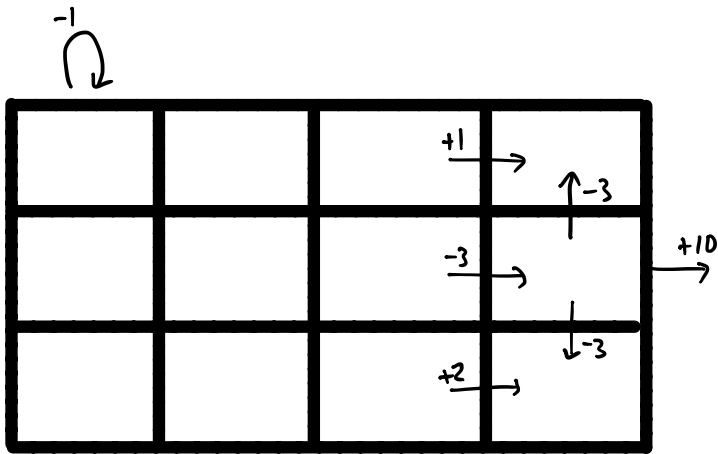


Monte Carlo (MC)



Temporal Difference (TD)





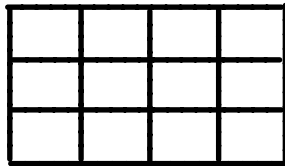
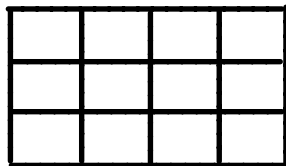
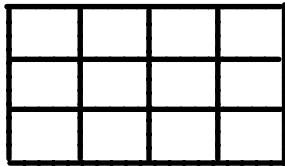
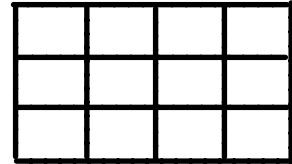
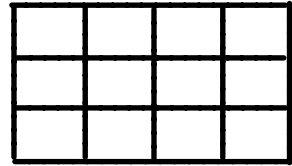
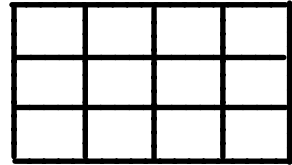
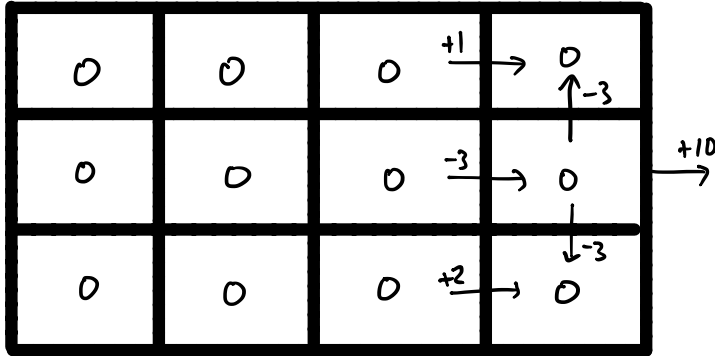
Stand  
 ↑  
 ↓  
 Wave

→ Clap

-1

$\alpha = 0.5, \gamma = 1$

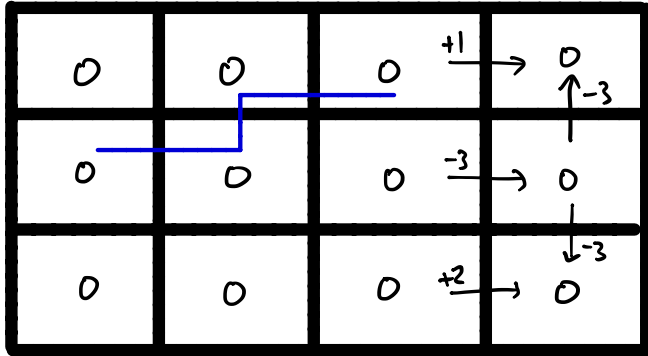
Stand  
↕  
Wave  
Clap



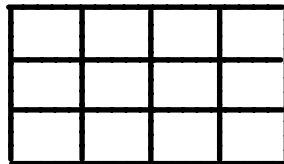
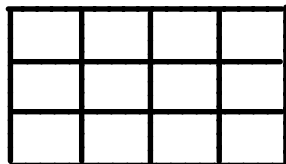
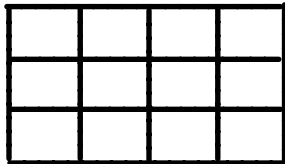
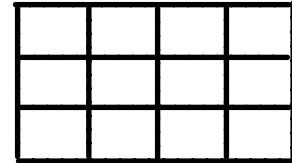
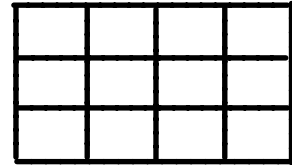
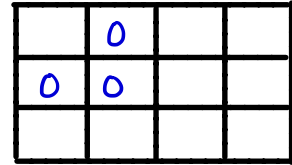
$$v(s) \leftarrow v(s) + \alpha [ \underbrace{R + \gamma v(s')}_{\text{"target"}} - v(s) ]$$

-1

$\alpha = 0.5, \gamma = 1$



Stand  
 $\updownarrow$  Clap  
 Wave

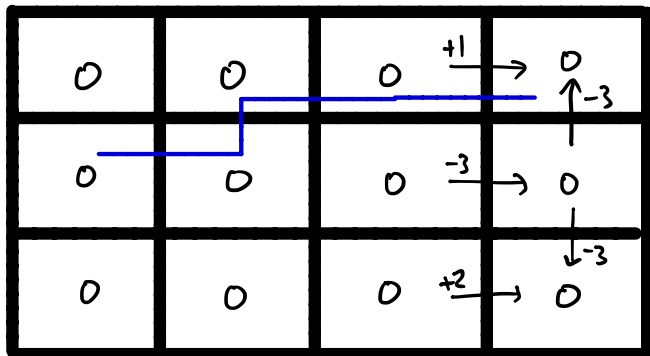


$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s')}_{\text{"target"}} - v(s)]$$



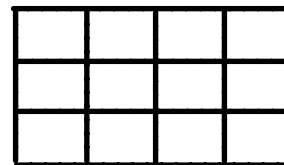
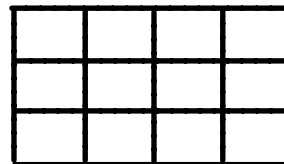
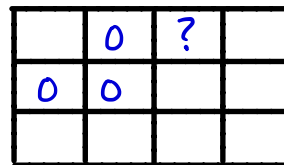
-1

$\alpha = 0.5, \gamma = 1$

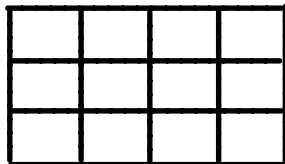
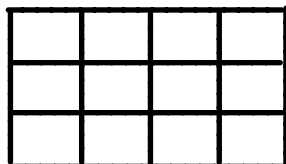
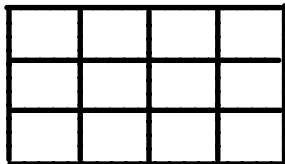


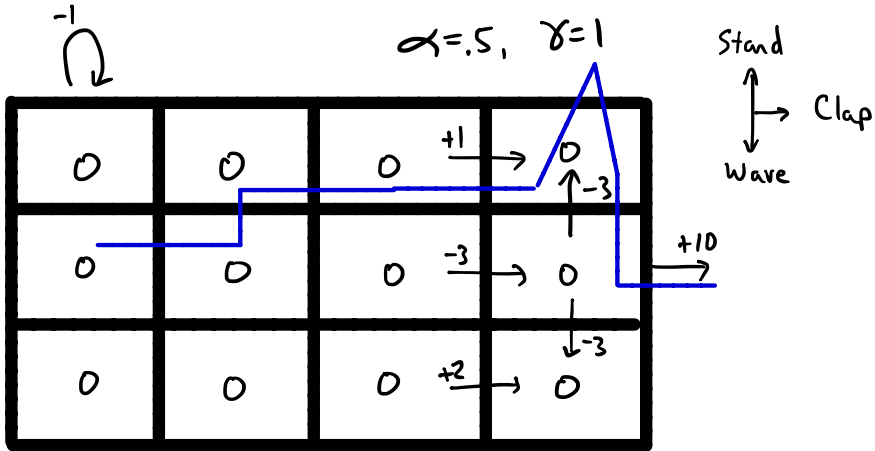
Stand  
 ↑  
 ↓  
 Wave

Clap



$$v(s) \leftarrow v(s) + \alpha [ \underbrace{R + \gamma v(s')}_{\text{"target"}} - v(s) ]$$



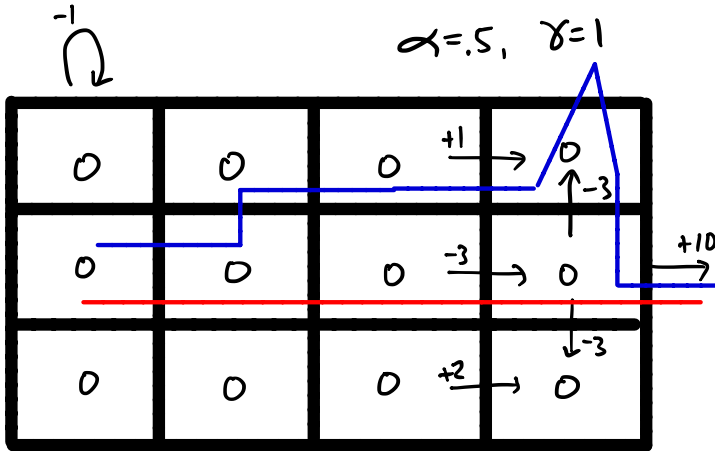


	0	.5	?
0	0		?






$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s')}_{\text{"target"}} - v(s)]$$



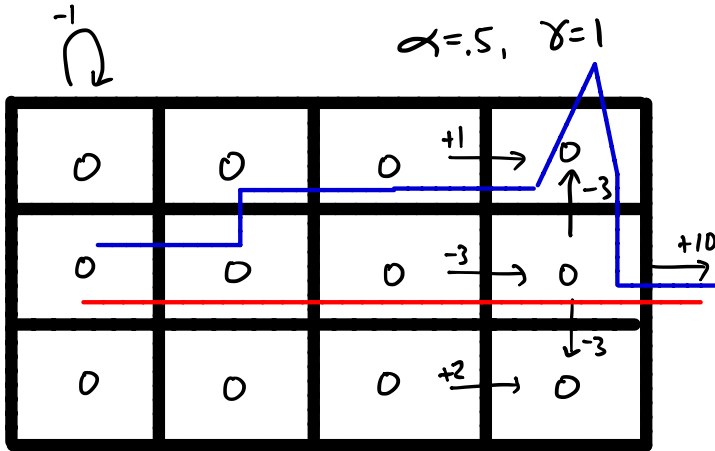
	0	.5	-.25
0	0		5

0	0	?	?





$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s')}_{\text{"target"}} - v(s)]$$



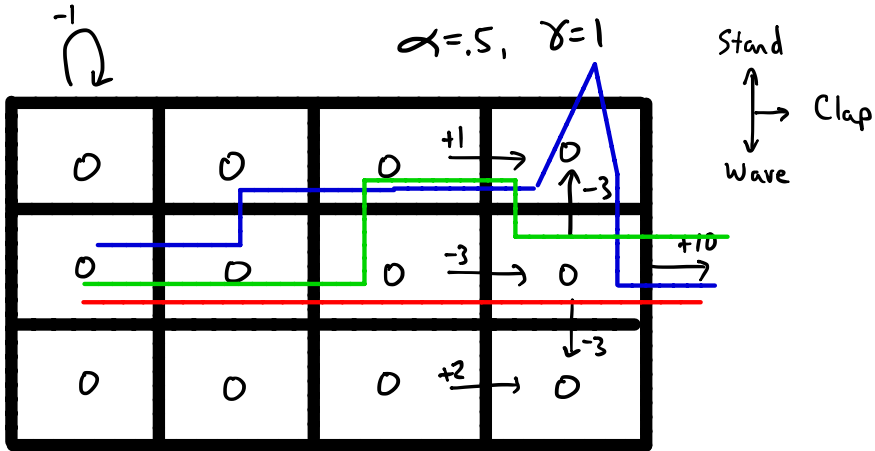
	0	.5	-.25
0	0		5

0	0	1	7.5





$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s')}_{\text{"target"}} - v(s)]$$



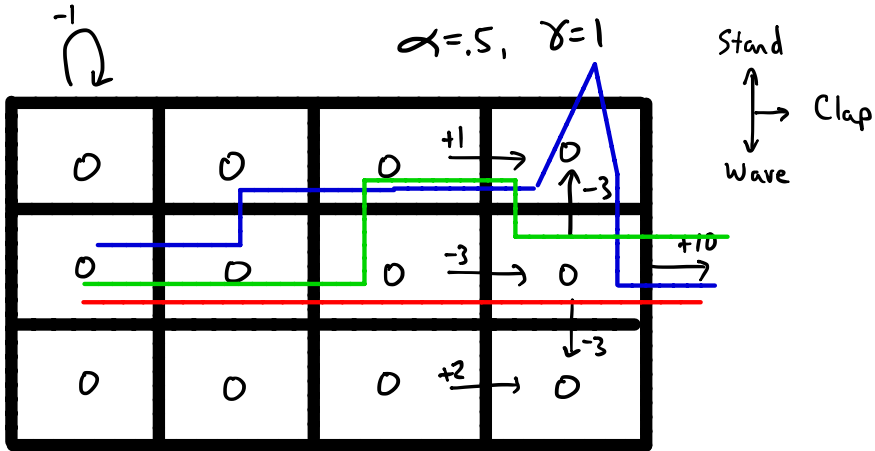
	0	.5	-.25
0	0		5

0	0	.5	-.25
0	0	1	7.5
0	0	0	0

		?	?
0	.5	.75	8.75




$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s')}_{\text{"target"}} - v(s)]$$



TD

	0	.5	-.25
0	0		5

0	0	.5	-.25
0	0	1	7.5
0	0	0	0

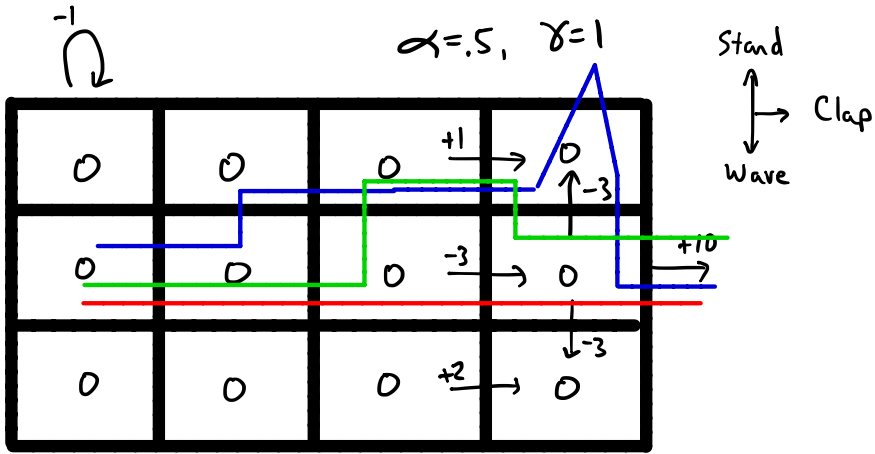
0	0	.625	3.625
0	.5	.75	8.75
0	0	0	0

MC (first visit)

	?	?	?
?	?		?



$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s') - v(s)}_{\text{"target"}}]$$



TD

	0	.5	-.25
0	0		5

0	0	.5	-.25
0	0	1	7.5
0	0	0	0

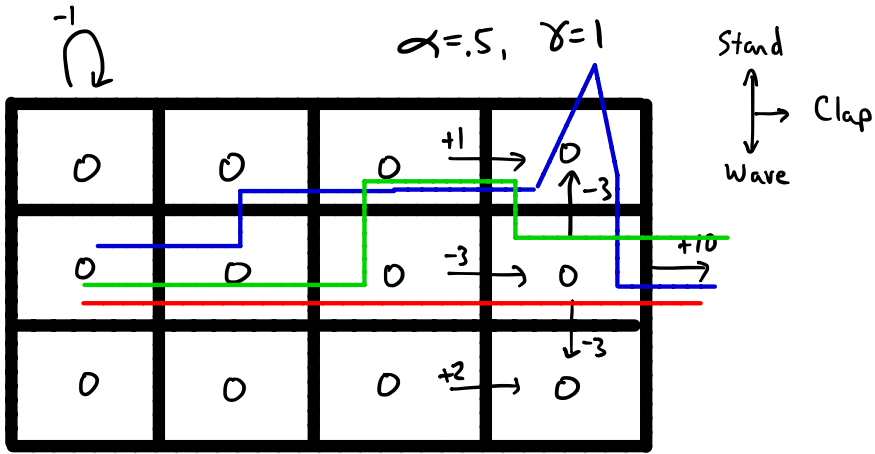
0	0	.625	3.625
0	.5	.75	8.75
0	0	0	0

MC (first visit)

	5	5	4.5
5	5		5

?	?	?	7.5


$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s') - v(s)}_{\text{"target"}}]$$



TD

	0	.5	-.25
0	0		5

0	0	.5	-.25
0	0	1	7.5
0	0	0	0

0	0	.625	3.625
0	.5	.75	8.75
0	0	0	0

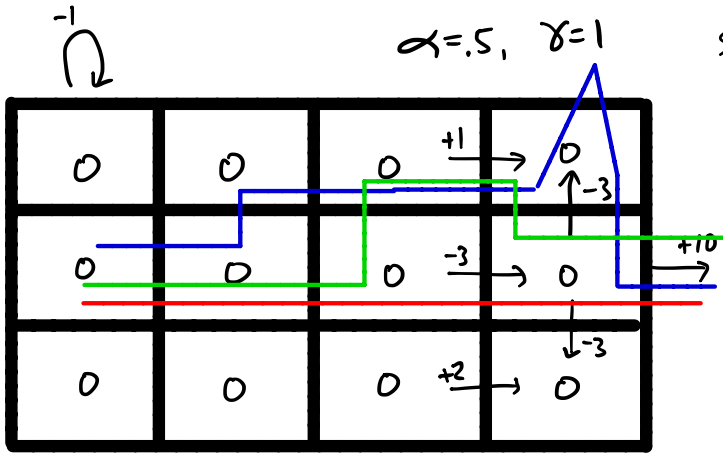
MC (first visit)

	5	5	4.5
5	5		5

0	5	5	4.5
6	6	3.5	7.5
0	0	0	0


$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s')}_{\text{"target"}} - v(s)]$$





Stand  
↕  
Wave  
→ Clap

TD

	0	.5	-.25
0	0		5

0	0	.5	-.25
0	0	1	7.5
0	0	0	0

0	0	.625	3.625
0	.5	.75	8.75
0	0	0	0

TD(0):

- one-step
- tabular
- model free

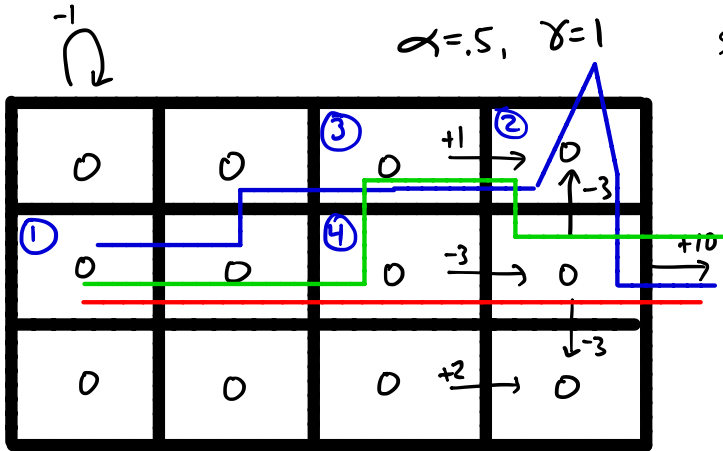
MC (first visit)

	5	5	4.5
5	5		5

0	5	5	4.5
6	6	3.5	7.5
0	0	0	0

		8	7.25
8.5	8.5	7.25	8.75

$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s')}_{\text{"target"}} - v(s)]$$



TD(0):

- one-step
- tabular
- model free

	0	.5	-.25
0	0		5

0	0	.5	-.25
0	0	1	7.5
0	0	0	0

0	0	.625	3.625
0	.5	.75	8.75
0	0	0	0

MC (first visit)

	5	5	4.5
5	5		5

0	5	5	4.5
6	6	3.5	7.5
0	0	0	0

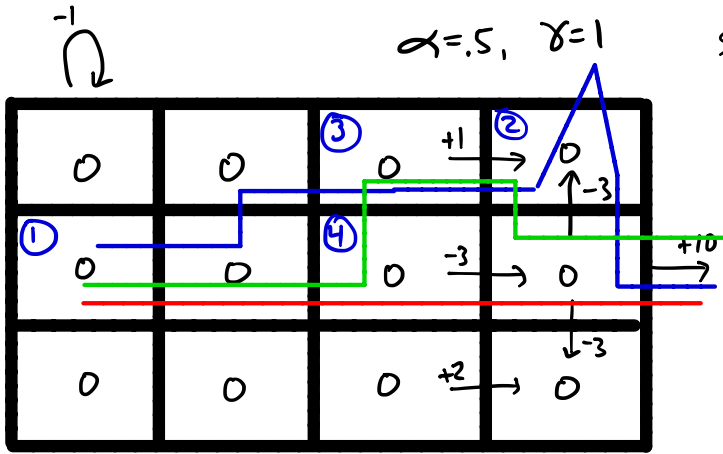
		8	7.25
8.5	8.5	7.25	8.75

$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s')}_{\text{"target"}} - v(s)]$$

Batch MC      Batch TD

- |   |   |   |
|---|---|---|
| ① | ? | ? |
| ② |   | ? |
| ③ |   | ? |
| ④ | ? | ? |

Convergence of Batch methods



Stand  
↕  
Wave  
→ Clap

TD

	0	.5	-.25
0	0		5

0	0	.5	-.25
0	0	1	7.5
0	0	0	0

0	0	.625	3.625
0	.5	.75	8.75
0	0	0	0

TD(0):  
- one-step  
- tabular  
- model free

MC (first visit)

	5	5	4.5
5	5		5

0	5	5	4.5
6	6	3.5	7.5
0	0	0	0

		8	7.25
8.5	8.5	7.25	8.75

$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s')}_{\text{"target"}} - v(s)]$$

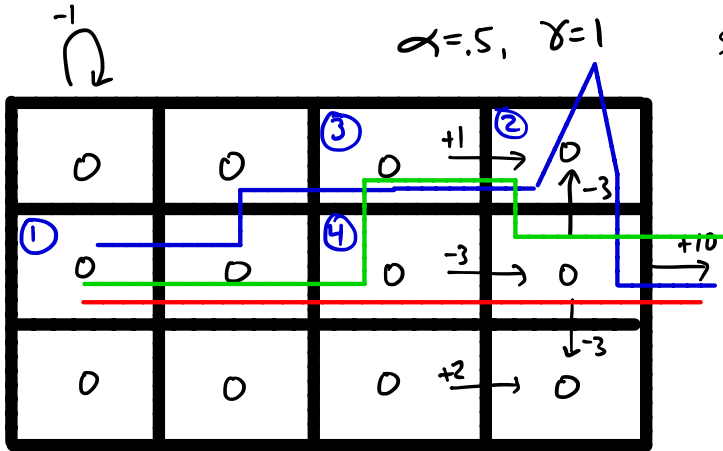
Batch MC

Batch TD

- ①  $\text{avg}(7, 11, 10) = 9.33$
- ②
- ③
- ④ ?

- 9.5
- 10.5
- ?

Convergence of Batch methods



Stand  
↕  
Wave

Clap

TD

	0	.5	-.25
0	0		5

0	0	.5	-.25
0	0	1	7.5
0	0	0	0

0	0	.625	3.625
0	.5	.75	8.75
0	0	0	0

TD(0):

- one-step
- tabular
- model free

MC (first visit)

	5	5	4.5
5	5		5

0	5	5	4.5
6	6	3.5	7.5
0	0	0	0

		8	7.25
8.5	8.5	7.25	8.75

$$v(s) \leftarrow v(s) + \alpha [R + \underbrace{\gamma v(s')}_{\text{"target"}} - v(s)]$$

Batch MC      Batch TD)

①  $avg(7, 11, 10) = 9.33$

②

9.5

③

10.5

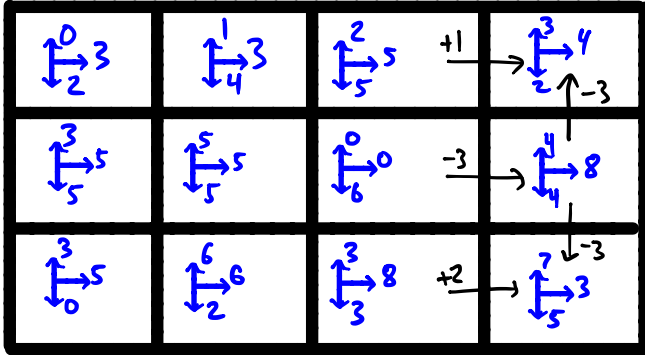
④

$avg(11, 7) = 9$        $avg(10.5, 7) = 8.75$

Convergence  
of Batch methods

-1

$\alpha = .1 \quad \gamma = 1$



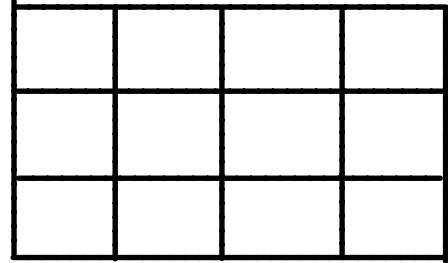
Stand  
↑  
Wave  
↓

Clap



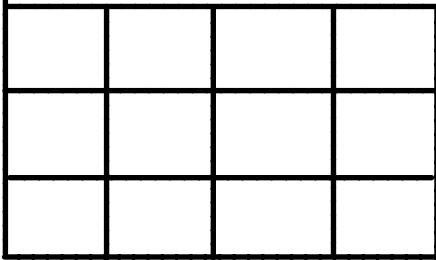
(TD:  $\begin{matrix} s \\ a \\ r \\ o$ )

SARSA

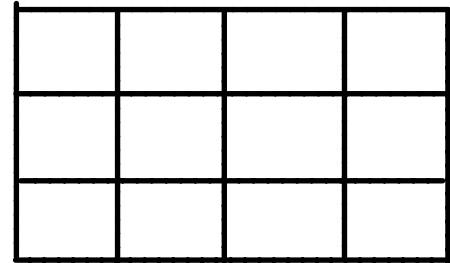


Policy:  $\epsilon$ -greedy,  $\epsilon = .75$ ; Ties:  $\rightarrow, \downarrow$

Expected SARSA



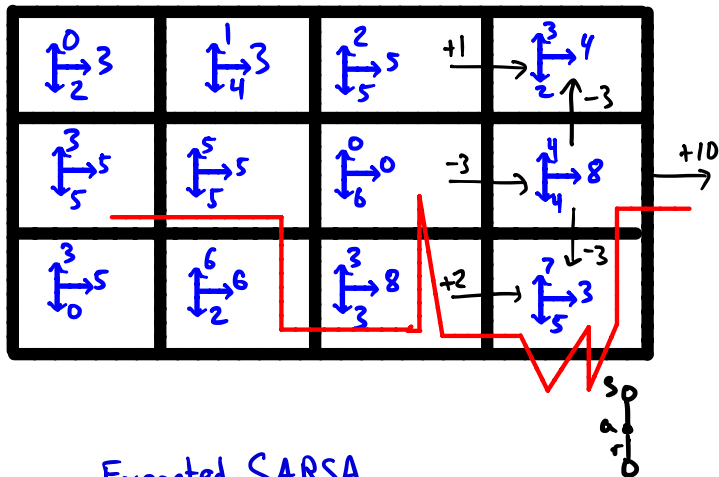
Q-learning



$$Q(s,A) \leftarrow Q(s,A) + \alpha [Target - Q(s,A)]$$

-1

$\alpha = .1 \quad \gamma = 1$



Stand

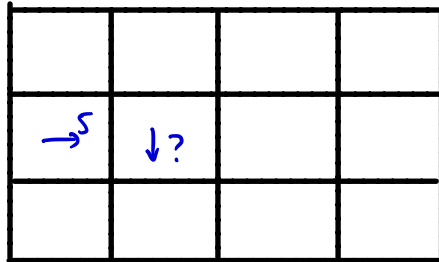


Wave

Clap

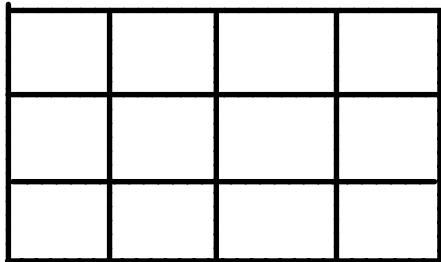


SARSA

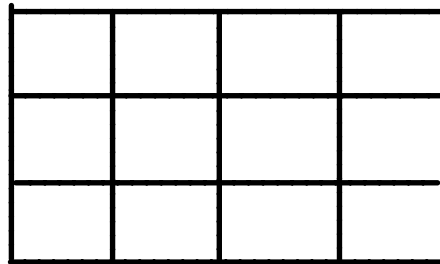


Policy:  $\epsilon$ -greedy,  $\epsilon = .75$ ; Ties:  $\rightarrow, \downarrow$

Expected SARSA



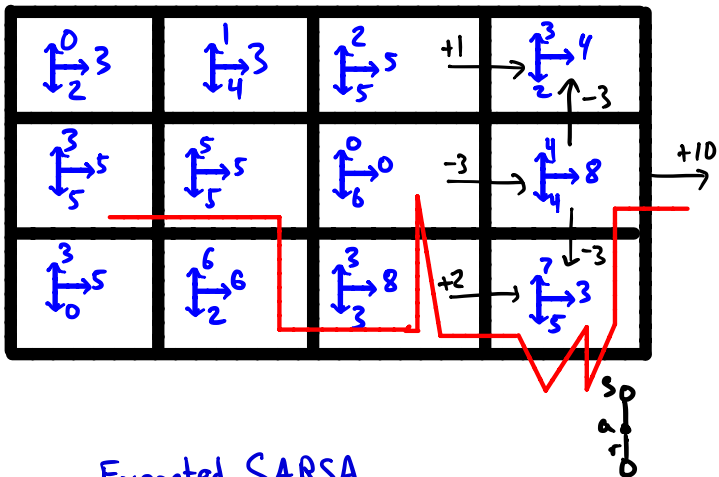
Q-learning



$$Q(S,A) \leftarrow Q(S,A) + \alpha [Target - Q(S,A)]$$

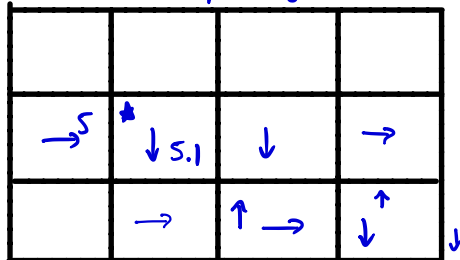
-1

$\alpha = .1 \quad \gamma = 1$



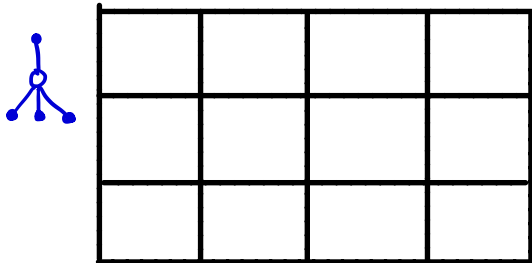
SARSA

\* = policy change

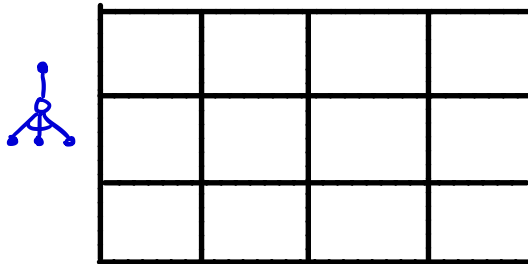


Policy:  $\epsilon$ -greedy,  $\epsilon = .75$ ; Ties: →, ↓

Expected SARSA



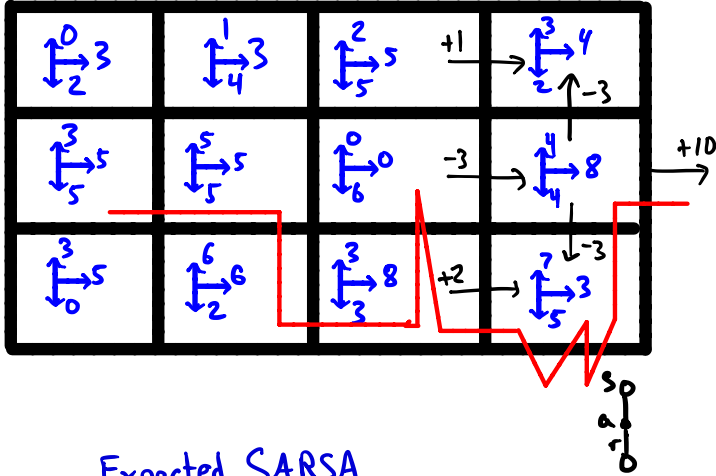
Q-learning



$$Q(S,A) \leftarrow Q(S,A) + \alpha [Target - Q(S,A)]$$

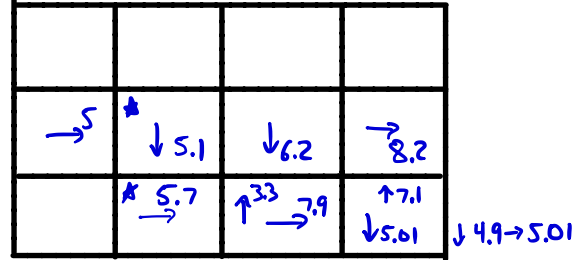
-1

$\alpha = .1 \quad \gamma = 1$



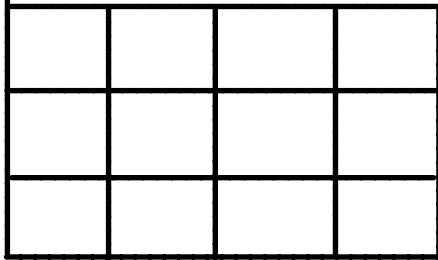
## SARSA

\* = policy change

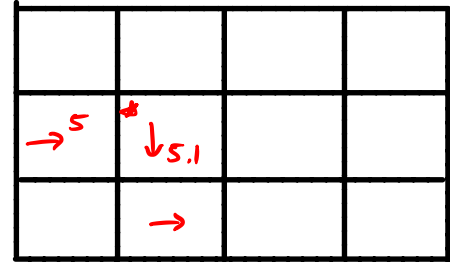


Policy:  $\epsilon$ -greedy,  $\epsilon = .75$ ; Ties: →, ↓

## Expected SARSA



## Q-learning

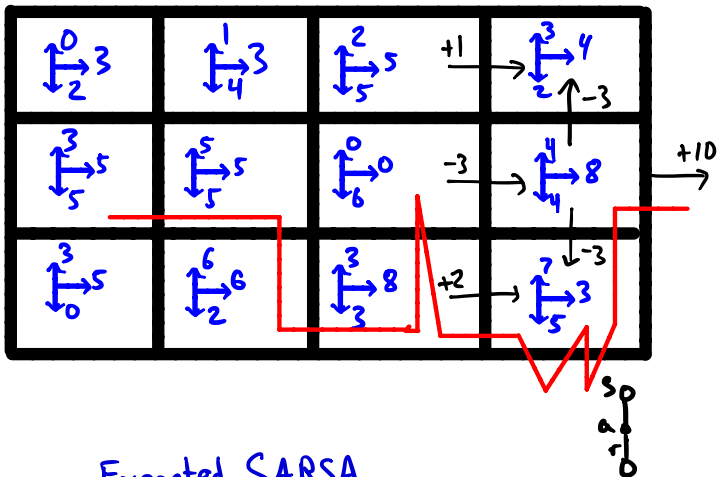


$$Q(s,A) \leftarrow Q(s,A) + \alpha [Target - Q(s,A)]$$



-1

$\alpha = .1 \quad \gamma = 1$



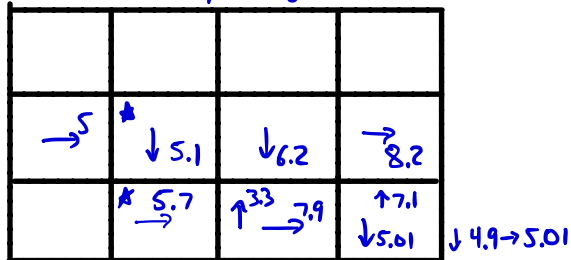
Stand  
↑  
Wave  
↓

Clap



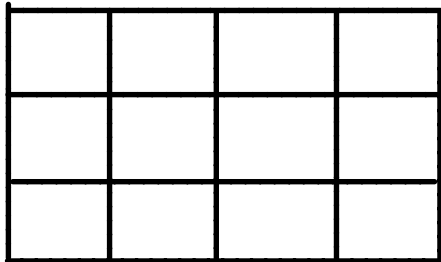
### SARSA

\* = policy change

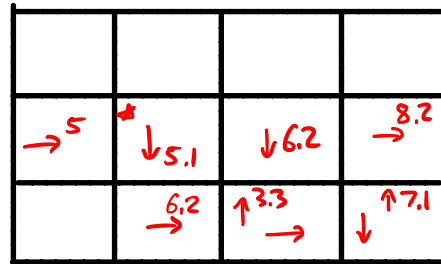


Policy:  $\epsilon$ -greedy,  $\epsilon = .75$ ; Ties: →, ↓

### Expected SARSA



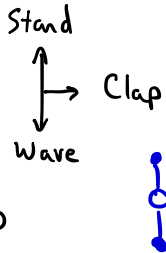
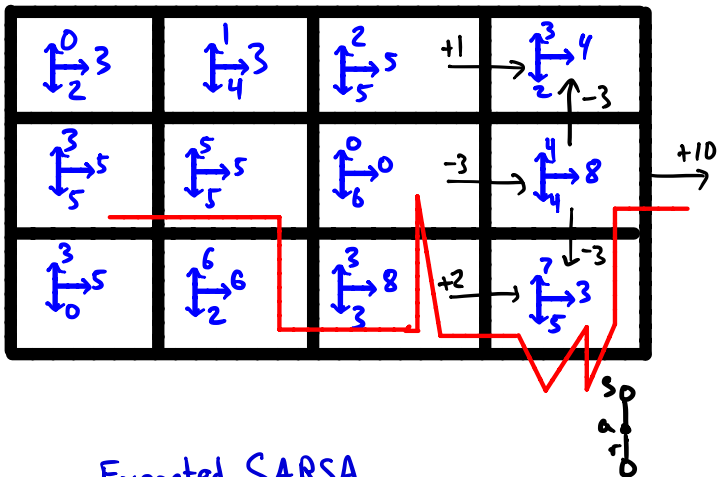
### Q-learning



$$Q(S,A) \leftarrow Q(S,A) + \alpha [Target - Q(S,A)]$$

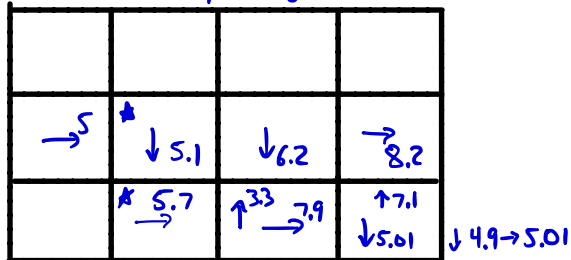
-1

$\alpha = .1 \quad \gamma = 1$



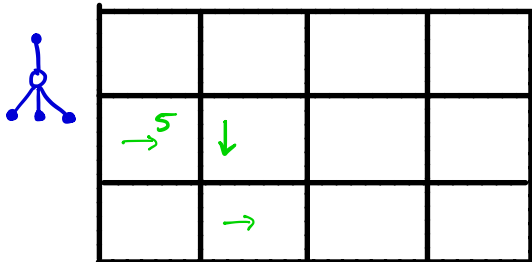
### SARSA

\* = policy change

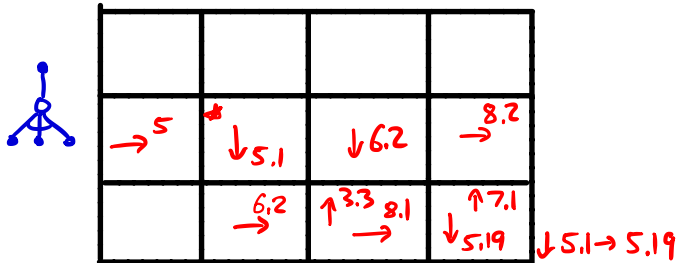


Policy:  $\epsilon$ -greedy,  $\epsilon = .75$ ; Ties: →, ↓

### Expected SARSA



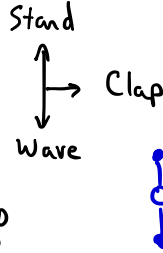
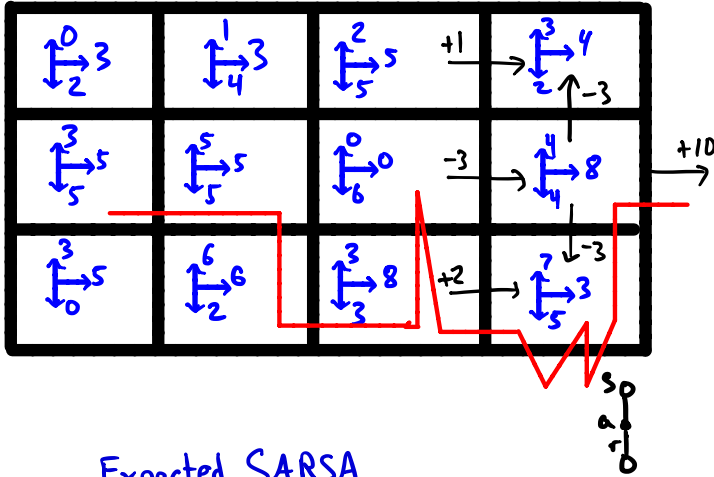
### Q-learning



$$Q(s,A) \leftarrow Q(s,A) + \alpha [\text{Target} - Q(s,A)]$$

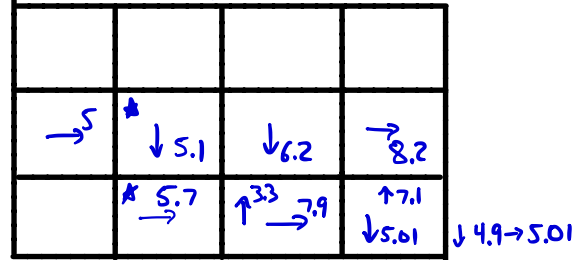
$\alpha = 1$   $\gamma = 1$

$\alpha = 1$   $\gamma = 1$



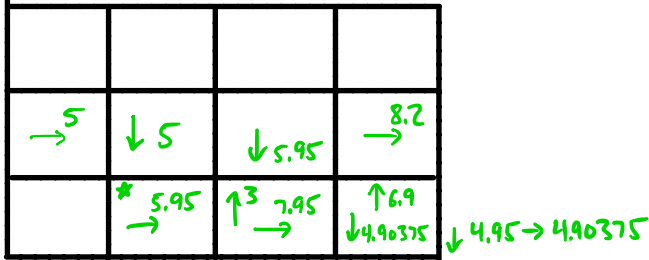
### SARSA

\* = policy change

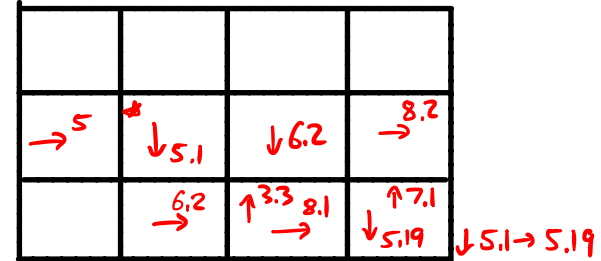


Policy:  $\epsilon$ -greedy,  $\epsilon = .75$ ; Ties:  $\rightarrow, \downarrow$

### Expected SARSA



### Q-learning



$$Q(S,A) \leftarrow Q(S,A) + \alpha [\text{Target} - Q(S,A)]$$

-1

$\alpha = .1 \quad \gamma = 1$

$\begin{matrix} 0 \\ \rightleftarrows \\ 2 \end{matrix} \begin{matrix} 3 \\ \end{matrix}$	$\begin{matrix} 1 \\ \rightleftarrows \\ 4 \end{matrix} \begin{matrix} 3 \\ \end{matrix}$	$\begin{matrix} 2 \\ \rightleftarrows \\ 5 \end{matrix} \begin{matrix} 5 \\ \end{matrix}$	+1 → $\begin{matrix} 3 \\ \rightleftarrows \\ 2 \end{matrix} \begin{matrix} 4 \\ \end{matrix}$
$\begin{matrix} 3 \\ \rightleftarrows \\ 5 \end{matrix} \begin{matrix} 5 \\ \end{matrix}$	$\begin{matrix} 5 \\ \rightleftarrows \\ 5 \end{matrix} \begin{matrix} 5 \\ \end{matrix}$	$\begin{matrix} 0 \\ \rightleftarrows \\ 6 \end{matrix} \begin{matrix} 0 \\ \end{matrix}$	-3 → $\begin{matrix} 4 \\ \rightleftarrows \\ 4 \end{matrix} \begin{matrix} 8 \\ \end{matrix}$
$\begin{matrix} 3 \\ \rightleftarrows \\ 0 \end{matrix} \begin{matrix} 5 \\ \end{matrix}$	$\begin{matrix} 6 \\ \rightleftarrows \\ 2 \end{matrix} \begin{matrix} 6 \\ \end{matrix}$	$\begin{matrix} 3 \\ \rightleftarrows \\ 3 \end{matrix} \begin{matrix} 8 \\ \end{matrix}$	+2 → $\begin{matrix} 7 \\ \rightleftarrows \\ 5 \end{matrix} \begin{matrix} 3 \\ \end{matrix}$

+10 →

Stand  
↕  
Wave

Clap



SARSA ← on policy

\* = policy change

→ 5	* ↓ 5.1	↓ 6.2	→ 8.2
	* → 5.7	↑ 3.3 → 7.9	↓ 7.1
			↓ 5.01

↓ 4.9 → 5.01

Policy:  $\epsilon$ -greedy,  $\epsilon = .75$ ; Ties: →, ↓

Expected SARSA ← Can be off policy



→ 5	↓ 5	↓ 5.95	→ 8.2
	* → 5.95	↑ 3 → 7.95	↑ 6.9
			↓ 4.90375

↓ 4.95 → 4.90375

Q-learning ← off policy



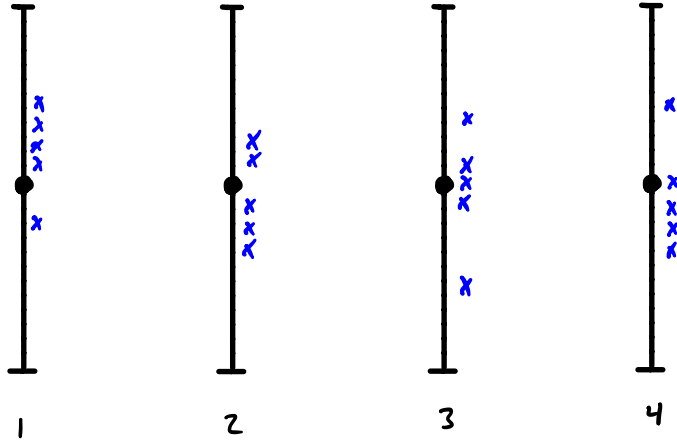
→ 5	* ↓ 5.1	↓ 6.2	→ 8.2
	→ 6.2	↑ 3.3 → 8.1	↑ 7.1
			↓ 5.19

↓ 5.1 → 5.19

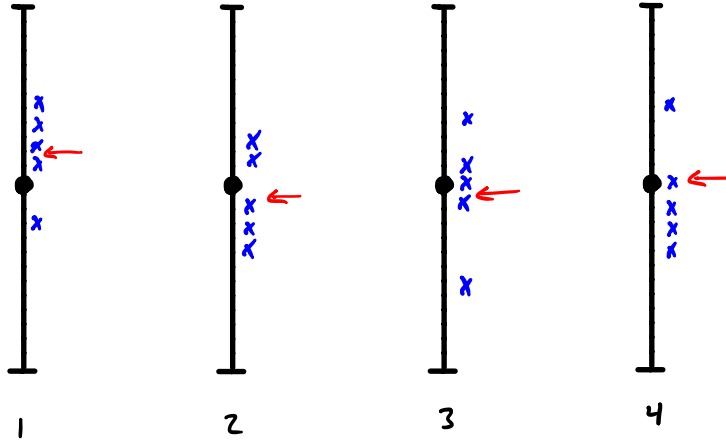
- How do learned policies differ?
- Conditions for convergence?
- Why Expected SARSA not as known?

$$Q(S,A) \leftarrow Q(S,A) + \alpha [Target - Q(S,A)]$$

Double Q learning - addresses maximization bias, regression to the mean  
(illustrated in bandit setting)

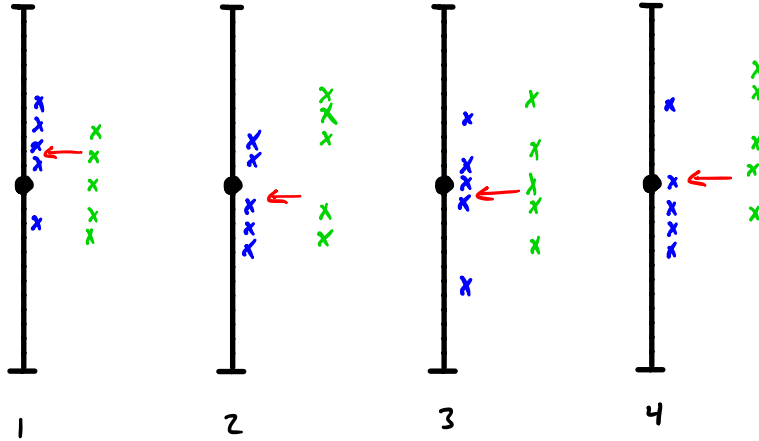


Double Q learning - addresses maximization bias, regression to the mean  
(illustrated in bandit setting)



● true means  
← sample means

Double Q learning - addresses maximization bias, regression to the mean  
(illustrated in bandit setting)



● true means  
← sample means  
x new, independent samples

## Ch 6 summary

Prediction:  $TD(0)$  = one-step, tabular, model-free TD

Control: SARSA  
Q-learning } expected SARSA

The core:

bootstrapping

Also: Double Q

After states