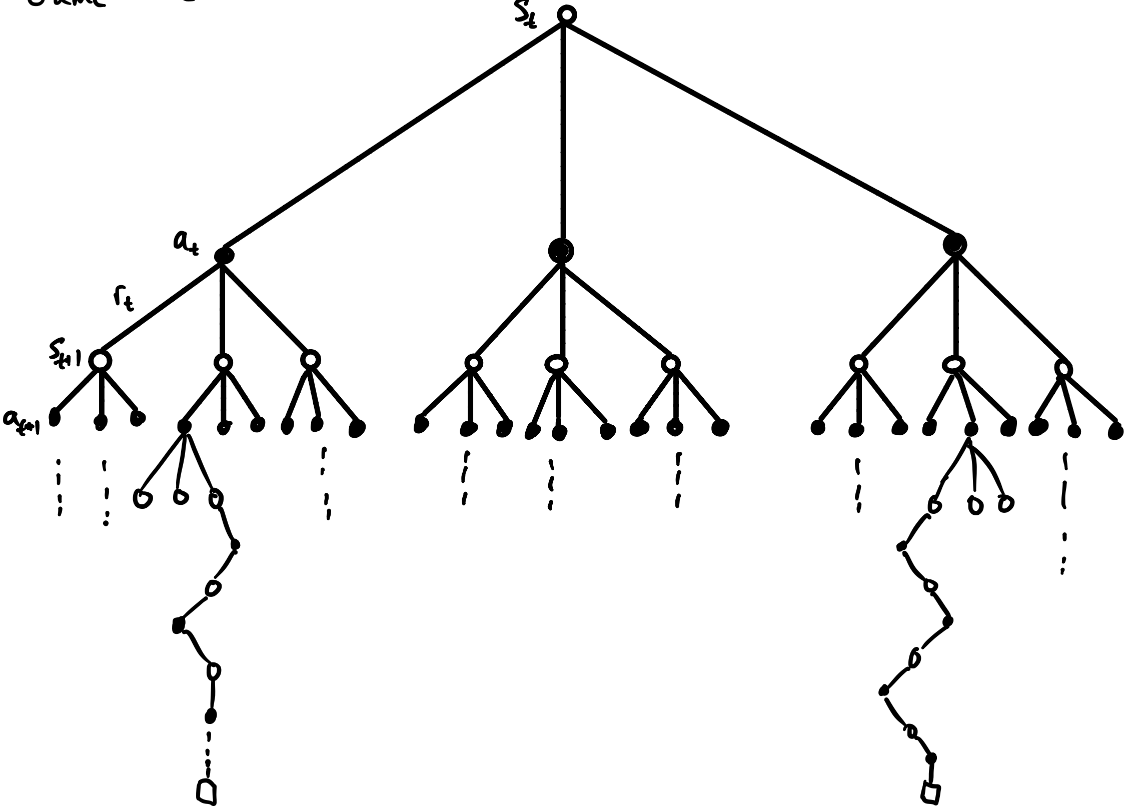
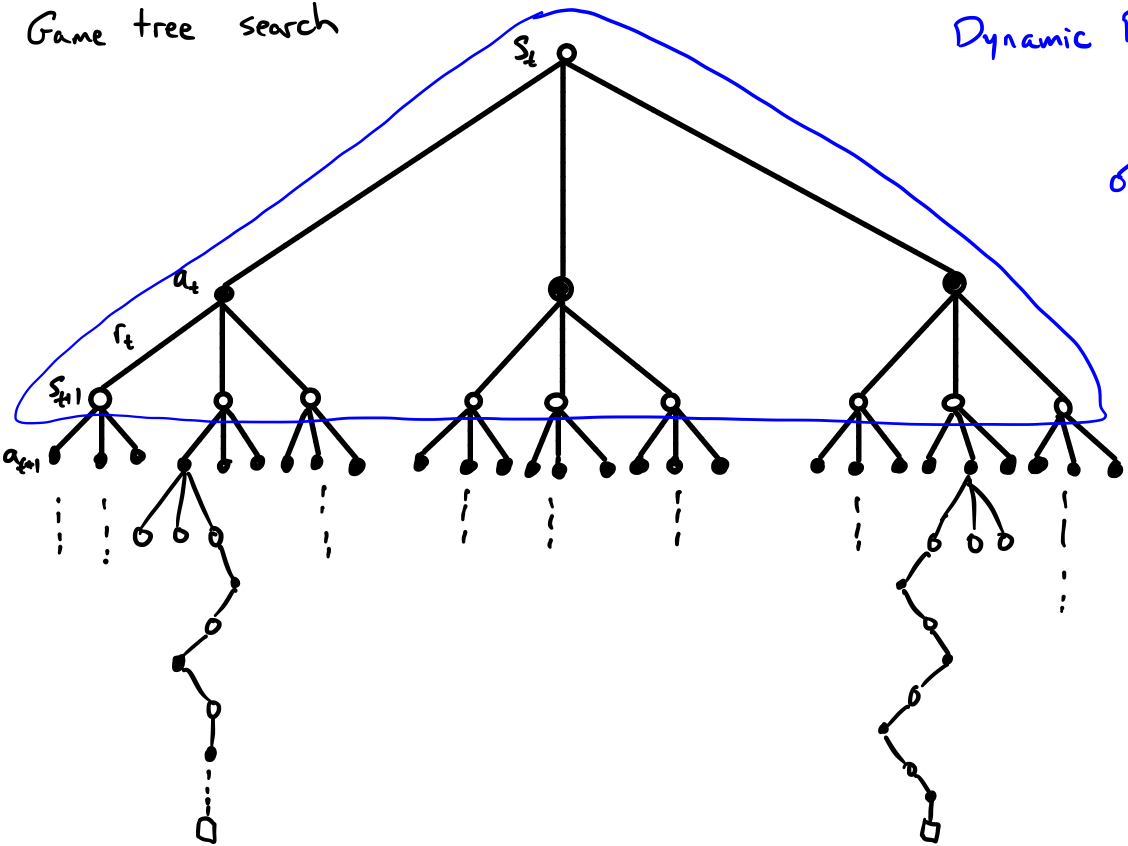
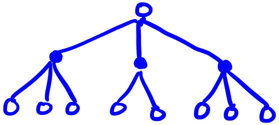
Game tree search



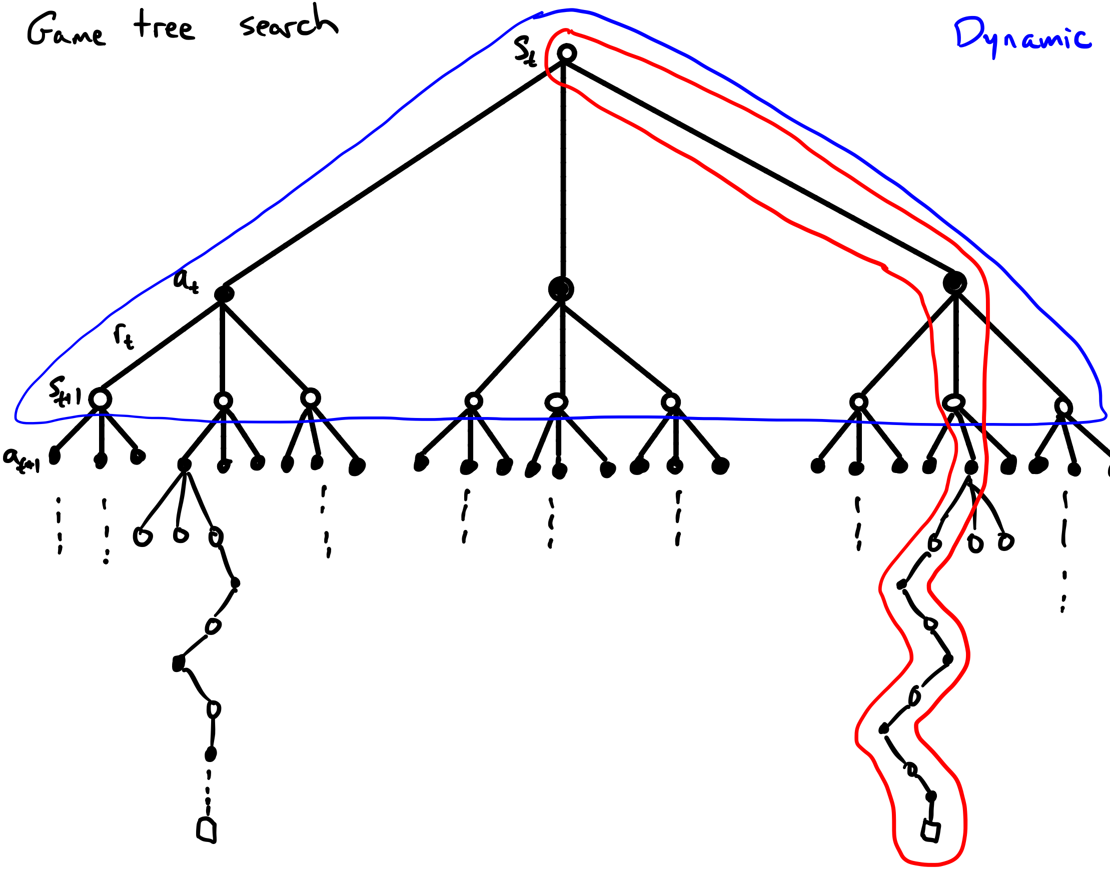
Game tree search

Dynamic Programming (DP)

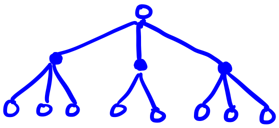




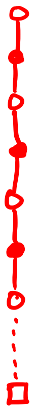
Game tree search

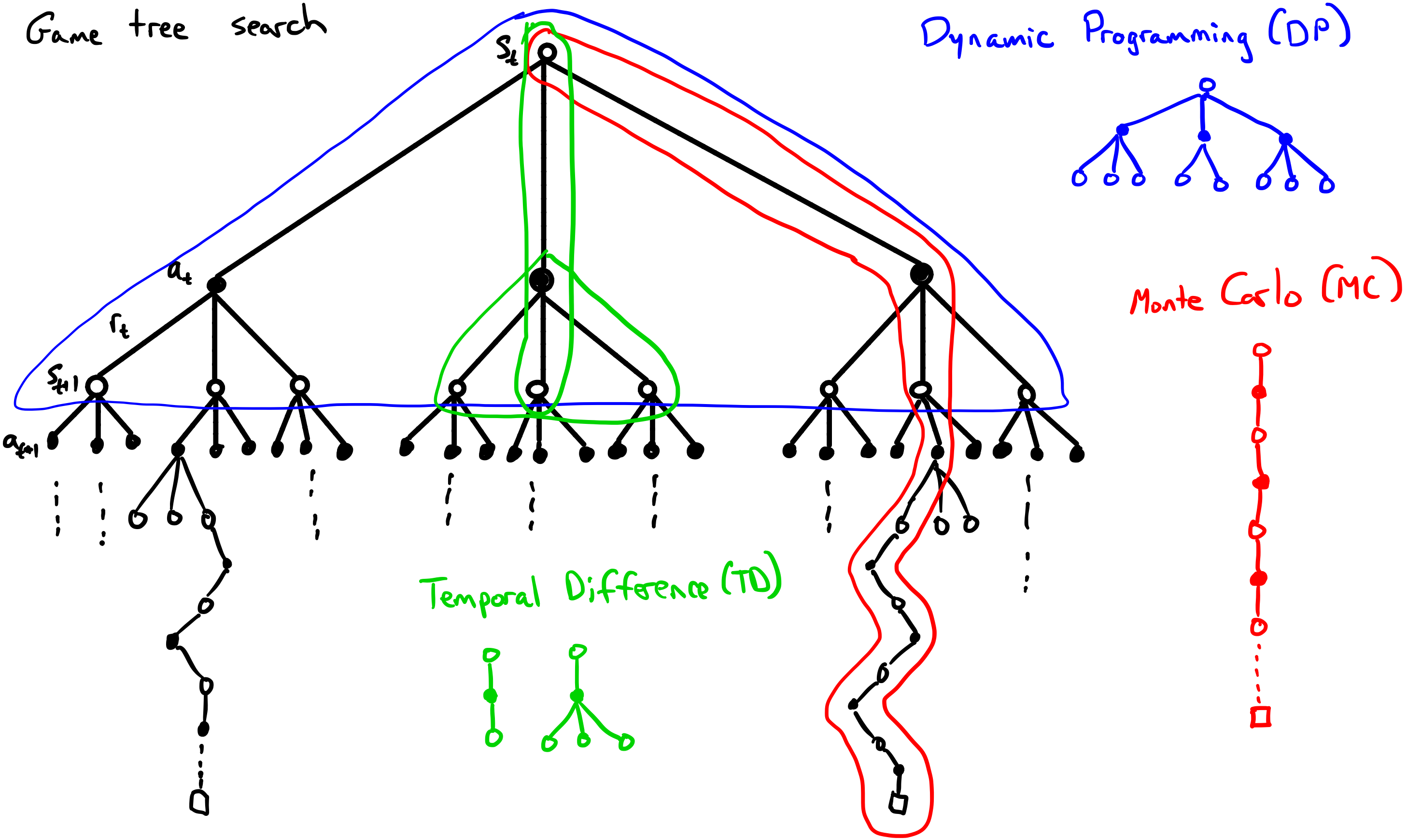


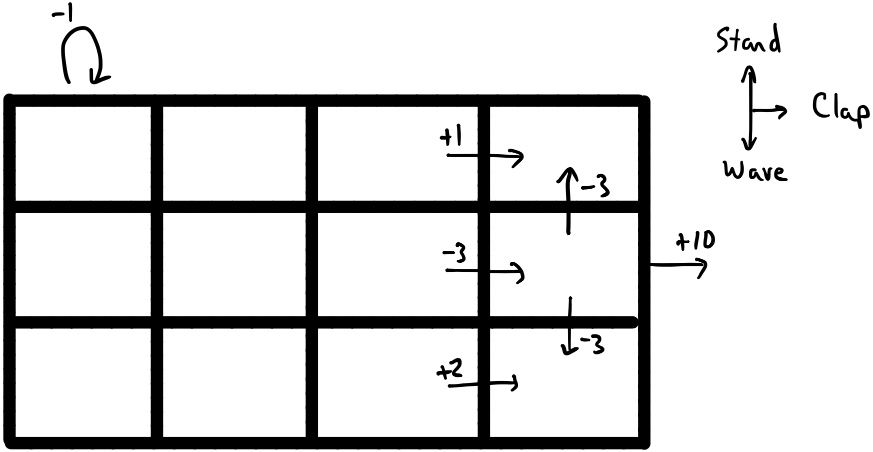
Dynamic Programming (DP)



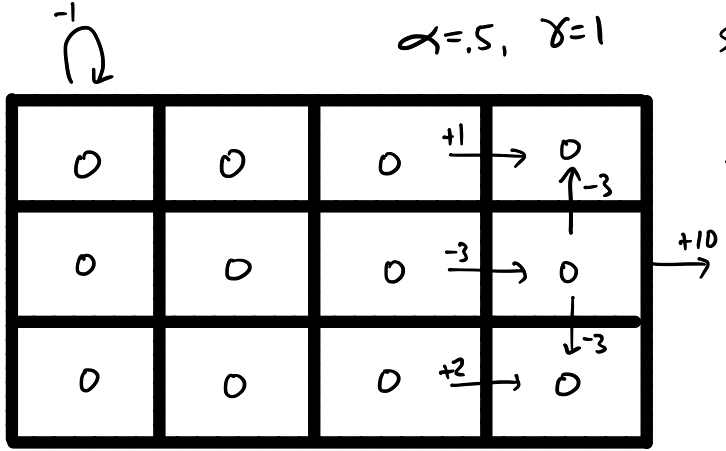
Monte Carlo (MC)

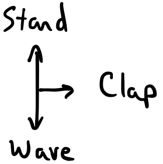



Game tree search
S_t
a_t
r_t
S_{t+1}
a_{t+1}
Dynamic Programming (DP)
Monte Carlo (MC)
Temporal Difference (TD)


-1
+1
-3
-3
+10
-3
+2
Stand
Clap
Wave

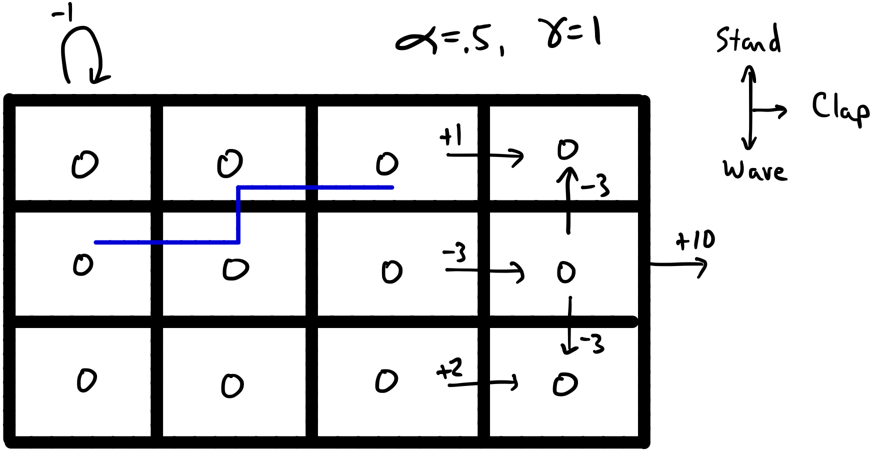
$\alpha = .5, \quad \gamma = 1$





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

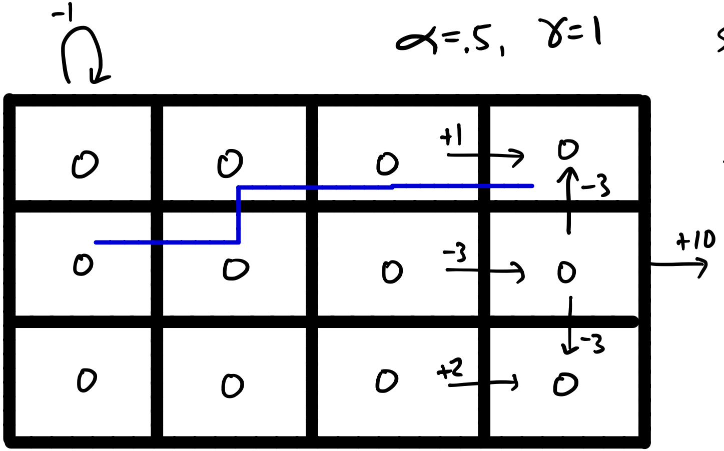
$\alpha = .5, \quad \gamma = 1$

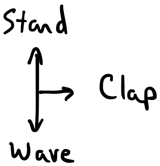


	0		
0	0		

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

$\alpha = .5, \quad \gamma = 1$

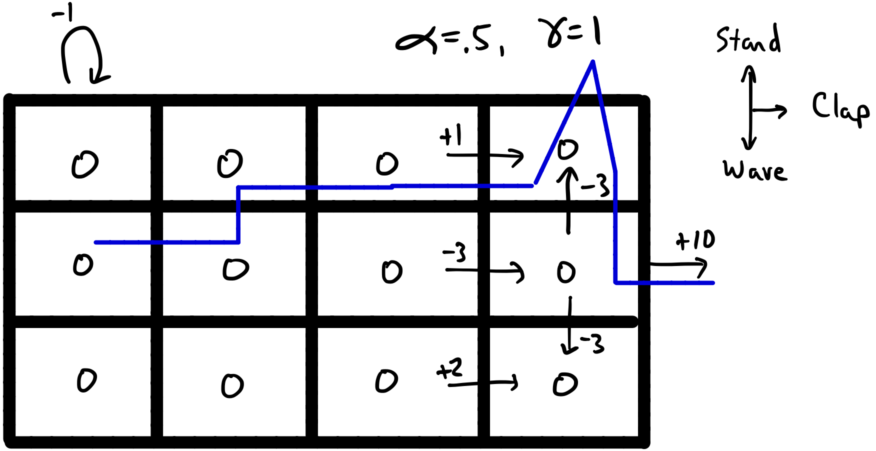




	0	?	
0	0		

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

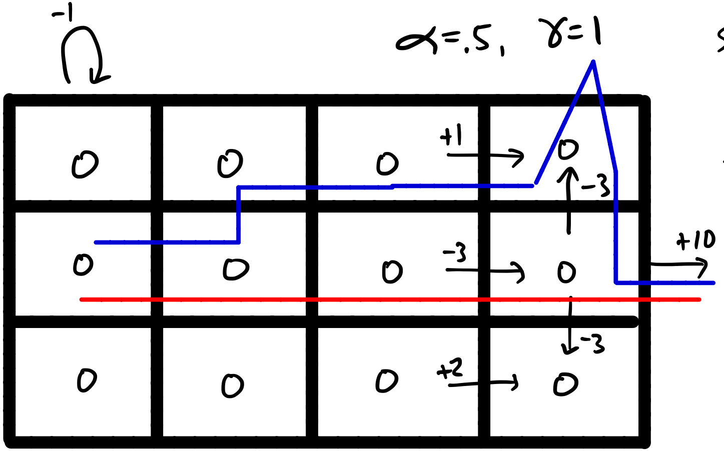
$\alpha = .5, \quad \gamma = 1$

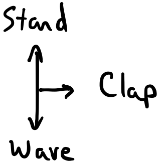


	0	.5	?
0	0		?

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

$\alpha = .5, \quad \gamma = 1$

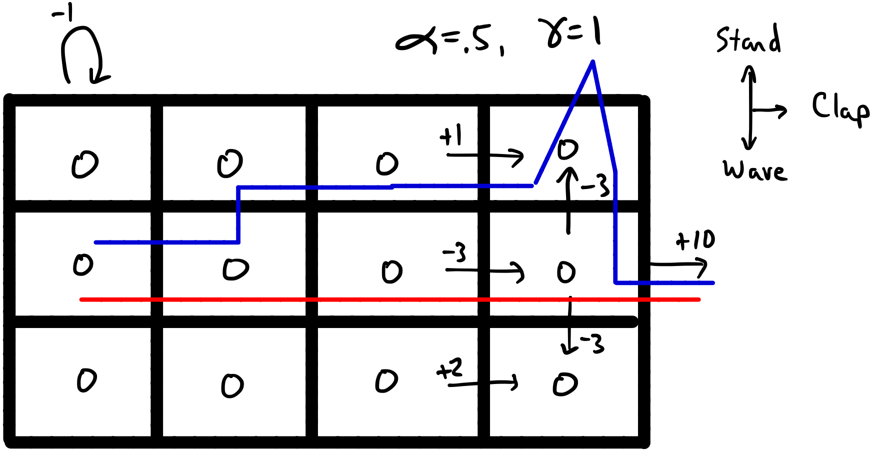




	0	.5	-.25
0	0		5

0	0	?	?

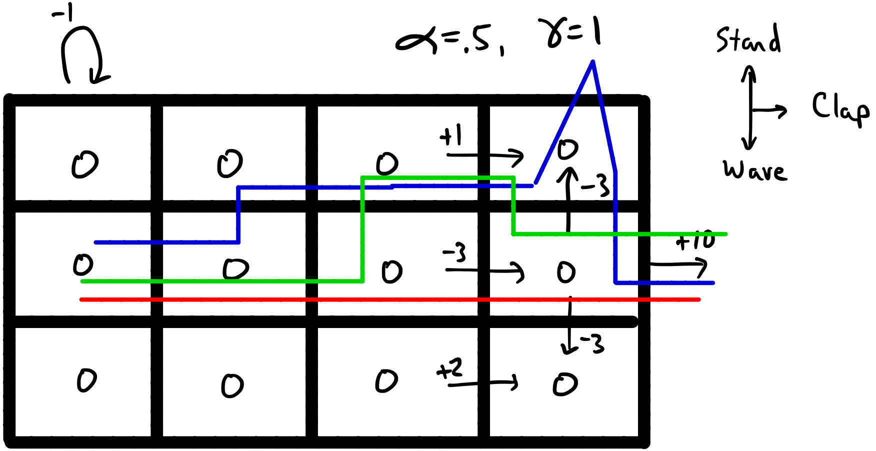
$$v(s) \leftarrow v(s) + \alpha\underbrace{[R + \gamma v(s') - v(s)]}_{\text{“target”}}$$



	0	.5	-.25
0	0		5

0	0	1	7.5

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

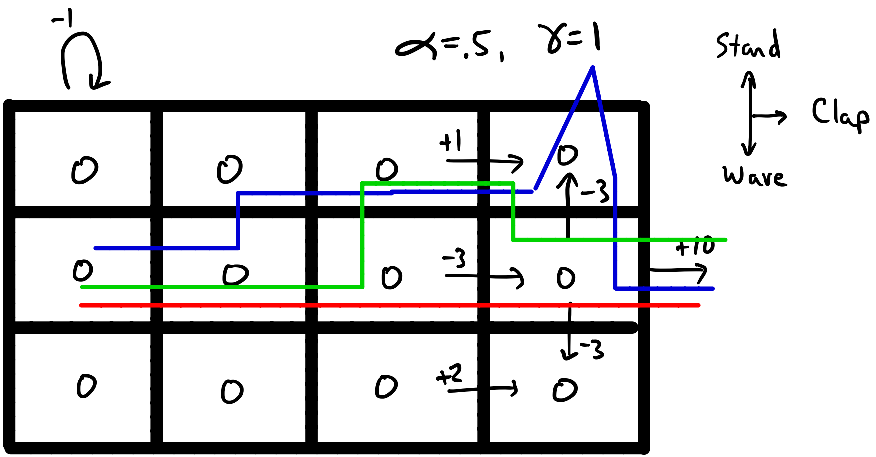


	0	.5	-.25
0	0		5

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

		?	?
0	.5	.75	8.75

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



TD

	0	.5	-.25
0	0		5

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

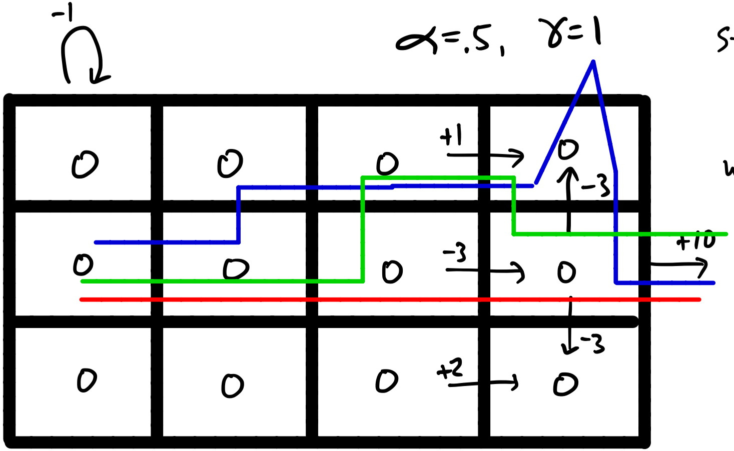
0	0	.625	3.625
0	.5	.75	8.75
0	0	0	0

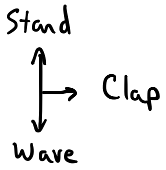
$$v(s) \leftarrow v(s) + \alpha\underbrace{[R + \gamma v(s') - v(s)]}_{\text{"target"}}$$

MC (first visit)

	?	?	?
?	?		?

$\alpha = .5, \ \gamma = 1$





TD

	0	.5	-.25
0	0		5

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

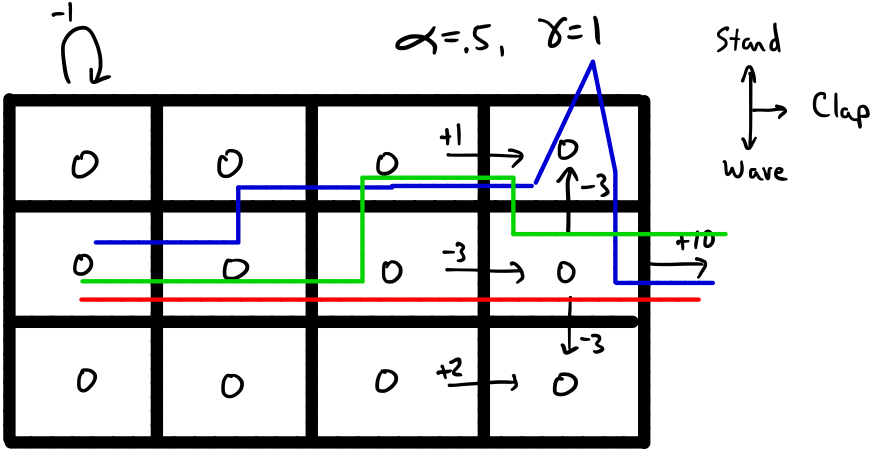
0	0	.625	3.625
0	.5	.75	8.75
0	0	0	0

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

MC (first visit)

	5	5	4.5
5	5		5

?	?	?	7.5



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

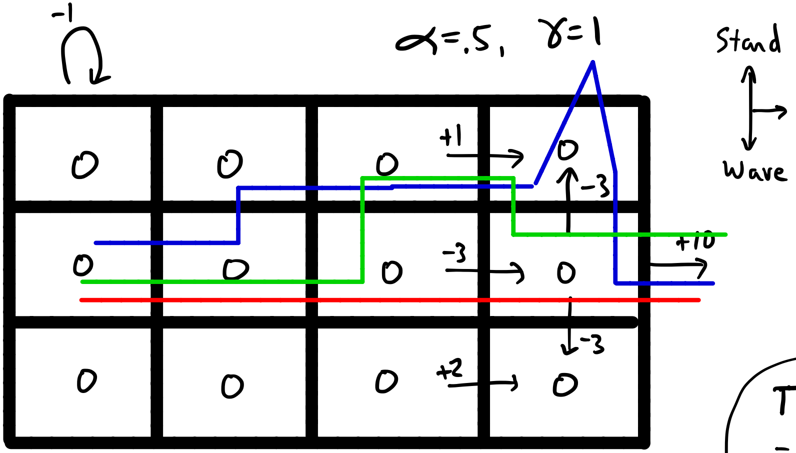
$$v(s) \leftarrow v(s) + \alpha \underbrace{[R + \gamma v(s') - v(s)]}_{\text{"target"}}$$

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

$\alpha = .5, \ \gamma = 1$



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

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

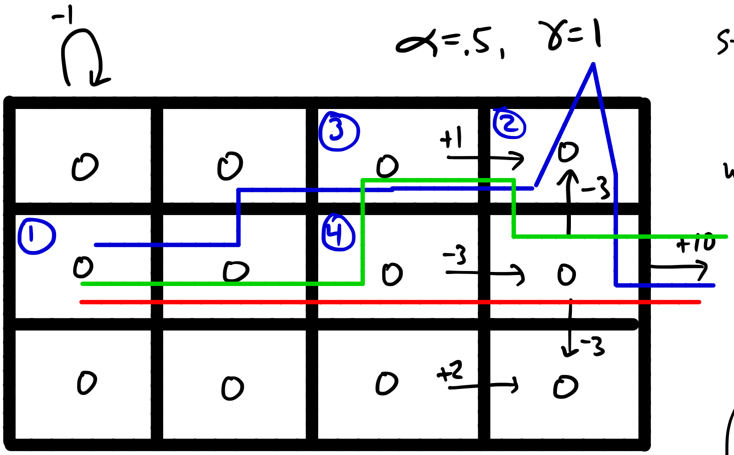
MC (first visit)

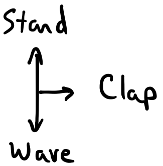
	5	5	4.5
5	5		5

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

$\alpha = .5, \ \gamma = 1$





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

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

MC (first visit)

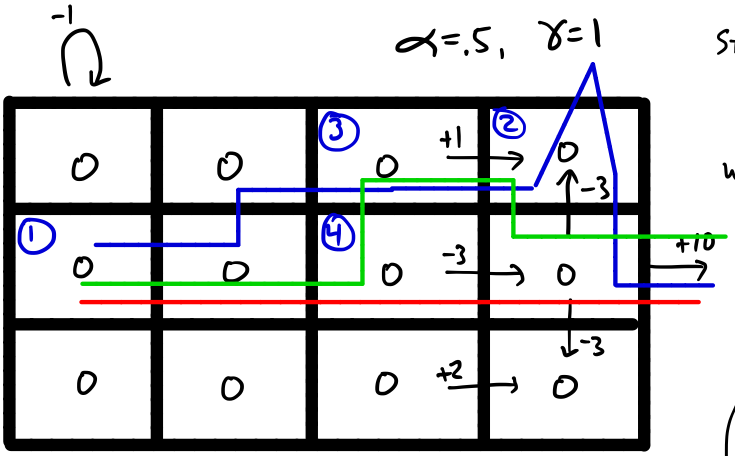
	5	5	4.5
5	5		5

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

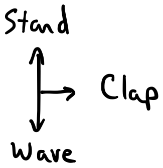
		8	7.25
8.5	8.5	7.25	8.75

Convergence of Batch methods

	Batch MC	Batch TD
①	?	
②		?
③		?
④	?	?



$\alpha = .5, \ \gamma = 1$



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

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

MC (first visit)

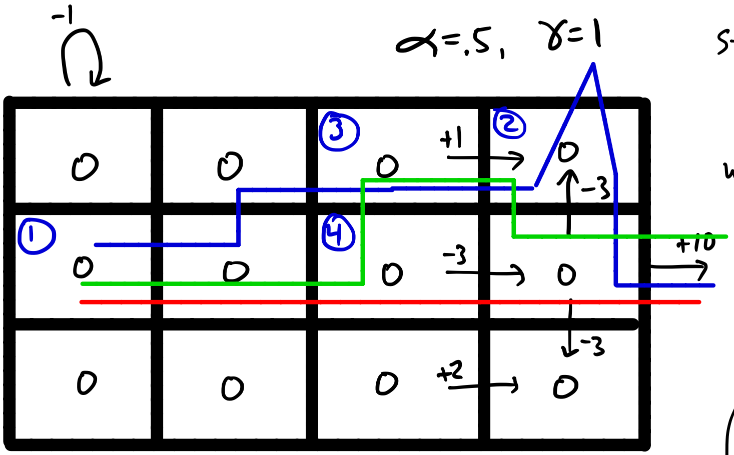
	5	5	4.5
5	5		5

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

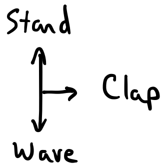
		8	7.25
8.5	8.5	7.25	8.75

Convergence of Batch methods

	Batch MC	Batch TD
①	avg(7,11,10)=9.33	
②		9.5
③		10.5
④	?	?



$\alpha = .5, \; \gamma = 1$



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

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

MC (first visit)

	5	5	4.5
5	5		5

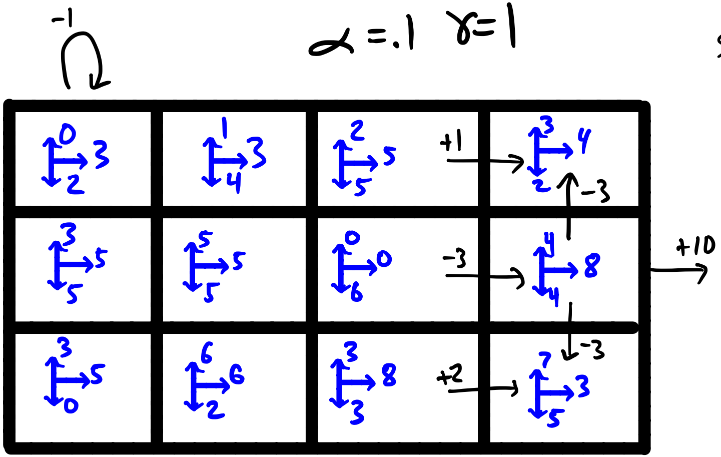
0	5	5	4.5
6	6	3.5	7.5
0	0	0	0

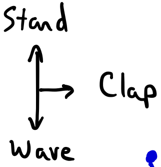
		8	7.25
8.5	8.5	7.25	8.75

Convergence of Batch methods

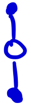
	Batch MC	Batch TD
①	avg(7,11,10)=9.33	
②		9.5
③		10.5
④	avg(11,7)=9	avg(10.5,7)=8.75

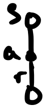
$\alpha = .1 \quad \gamma = 1$





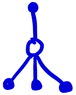
SARSA



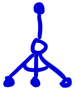
(TD: )

Policy: ε-greedy, ε = .75; Ties: →, ↓

Expected SARSA

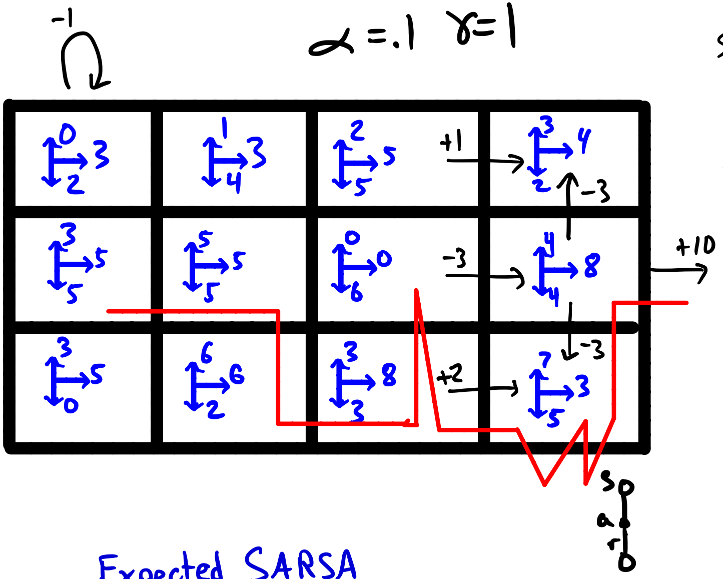


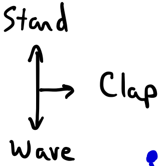
Q-learning



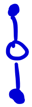
$$Q(S,A) \leftarrow Q(S,A) + \alpha\left[\text{Target} - Q(S,A)\right]$$

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





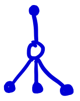
SARSA



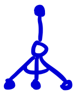
→ 5	↓ ?		

Policy: ε-greedy, ε = .75; Ties: →, ↓

Expected SARSA

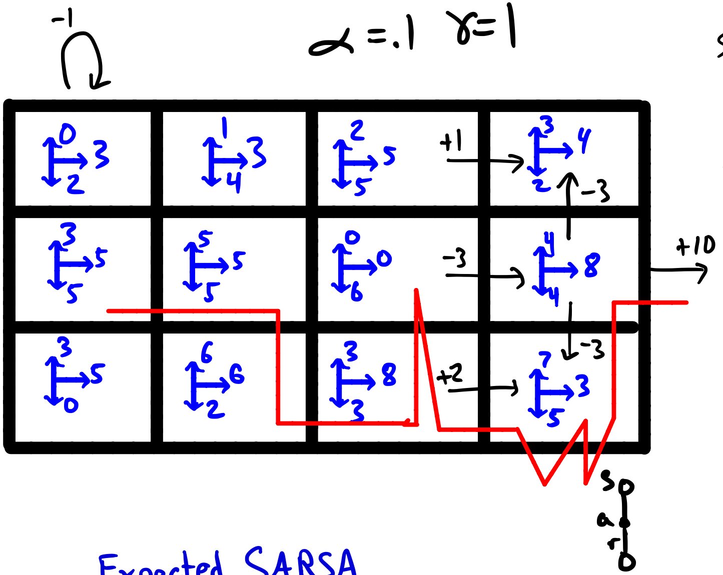


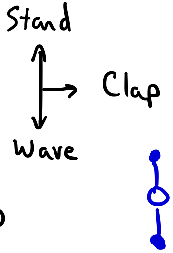
Q-learning



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

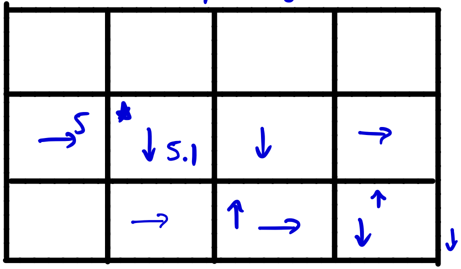
$\alpha = .1 \quad \gamma = 1$





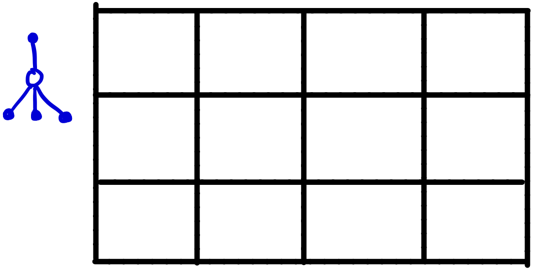
SARSA

★= policy change

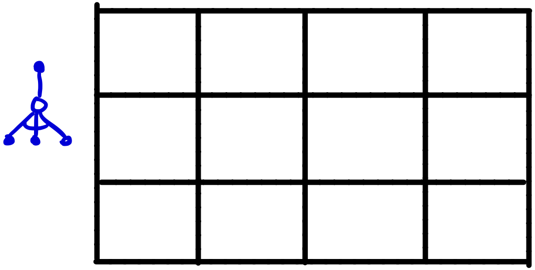


Policy: ε-greedy, ε=.75; Ties: →, ↓

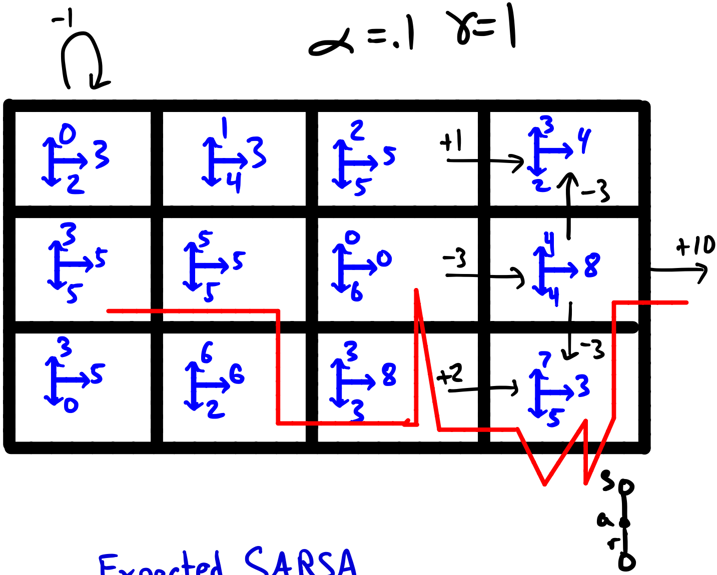
Expected SARSA

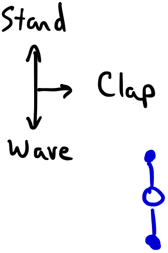


Q-learning



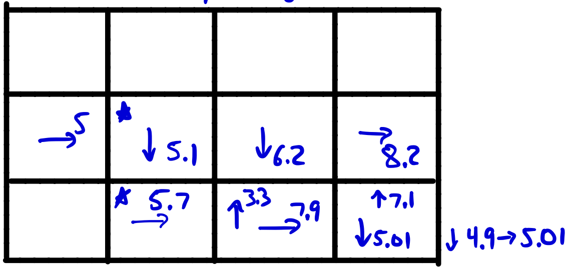
$$Q(S,A) \leftarrow Q(S,A) + \alpha\left[\text{Target} - Q(S,A)\right]$$





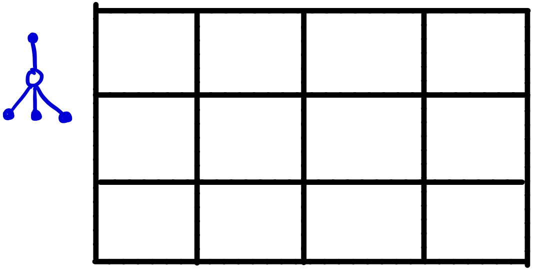
SARSA

★= policy change

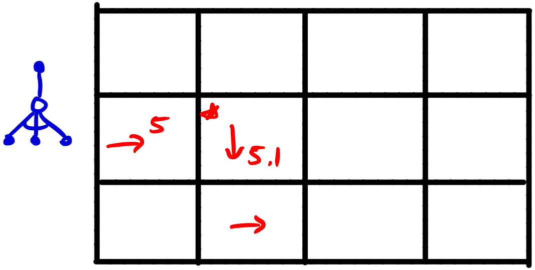


Policy: ε-greedy, ε=.75; Ties: →, ↓

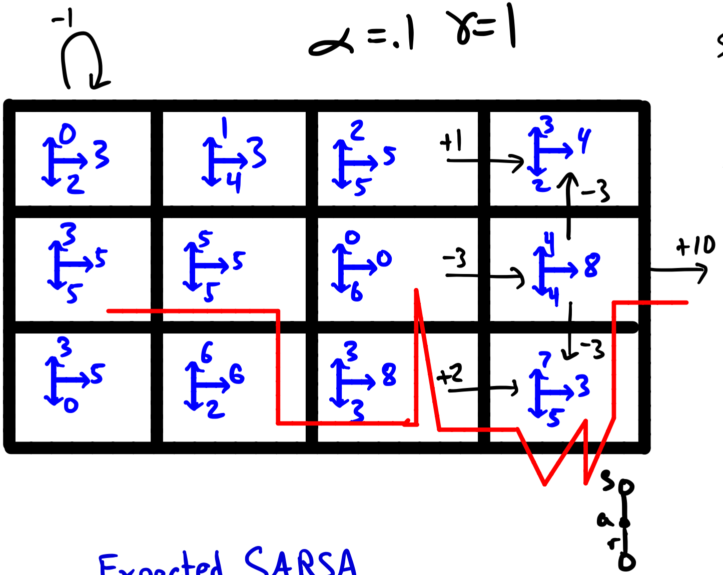
Expected SARSA

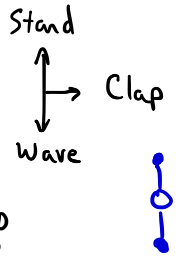


Q-learning



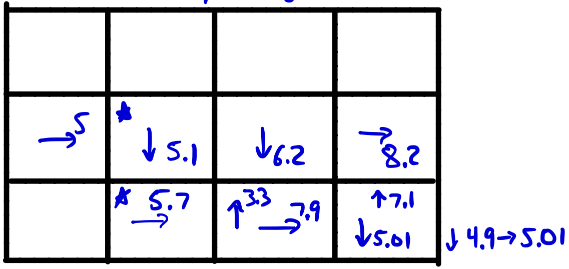
$$Q(S,A) \leftarrow Q(S,A) + \alpha\left[\text{Target} - Q(S,A)\right]$$





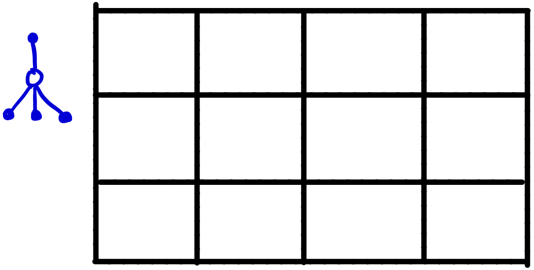
SARSA

★= policy change

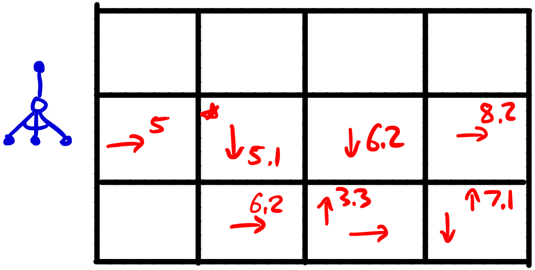


Policy: ε-greedy, ε=.75; Ties: →, ↓

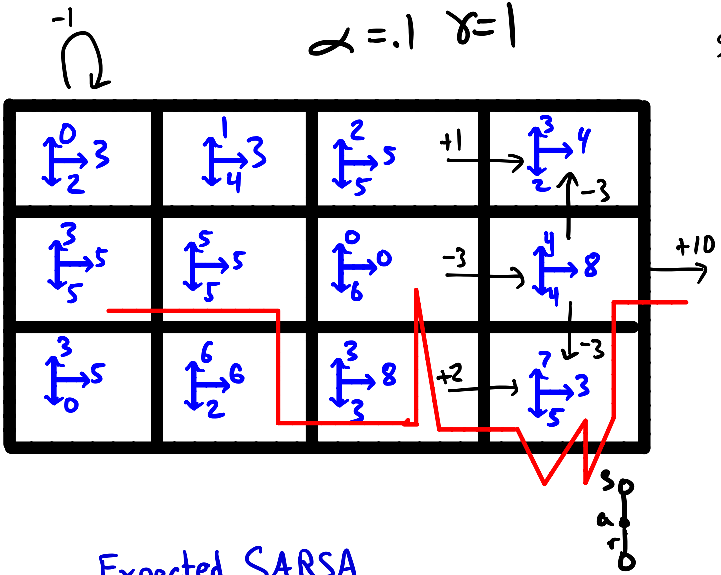
Expected SARSA

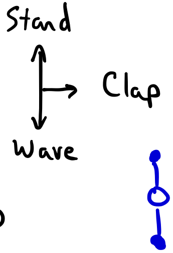


Q-learning



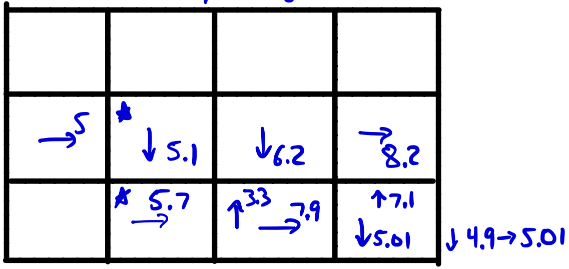
$$Q(S,A) \leftarrow Q(S,A) + \alpha\left[\text{Target} - Q(S,A)\right]$$





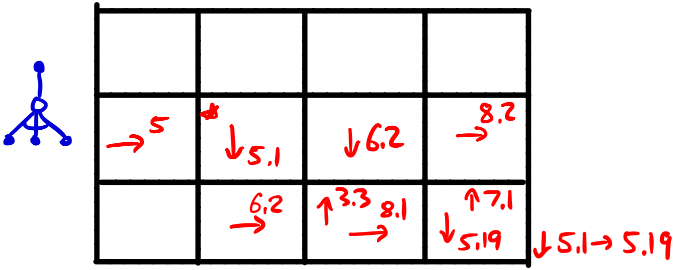
SARSA

★= policy change

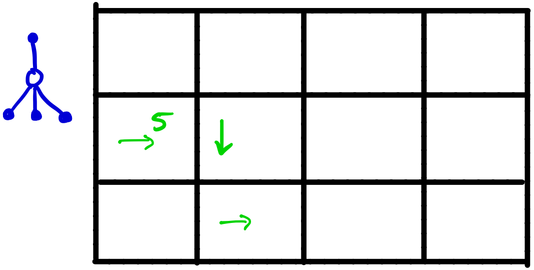


Policy: ε-greedy, ε=.75; Ties: →, ↓

Q-leasning

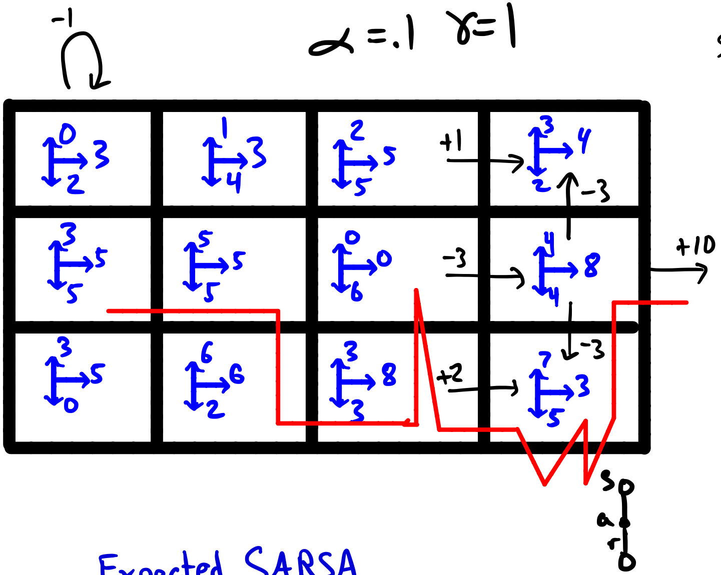


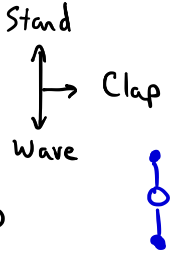
Expected SARSA



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

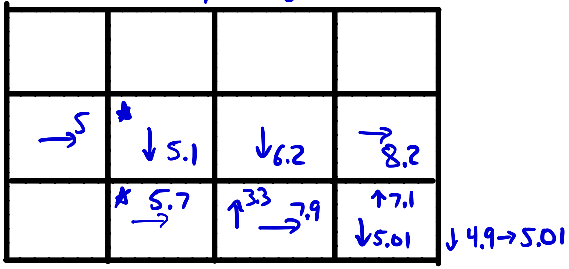
$\alpha = .1 \quad \gamma = 1$





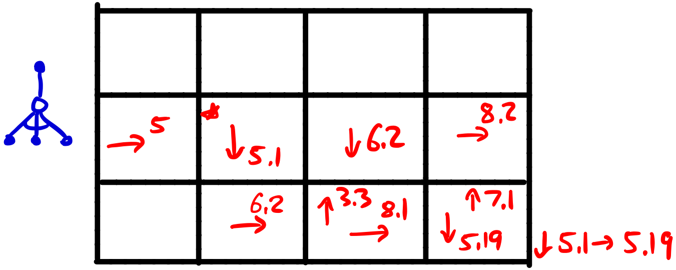
SARSA

★= policy change

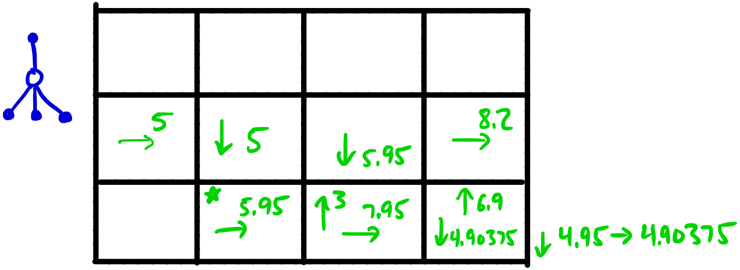


Policy: ϵ-greedy, ϵ=.75; Ties: →, ↓

Q-learning

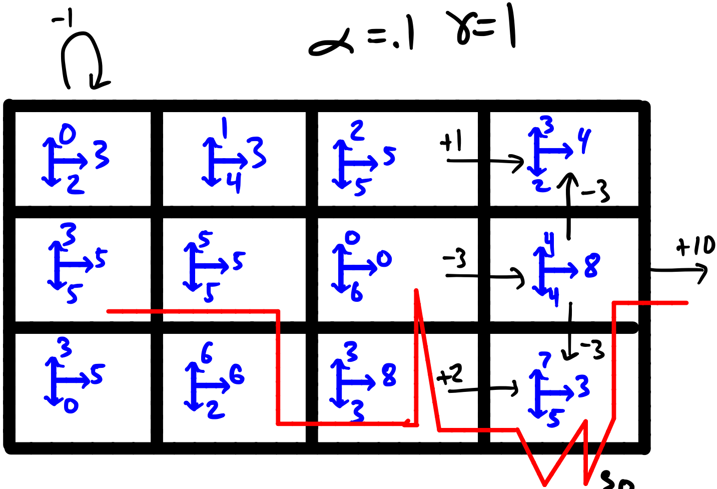


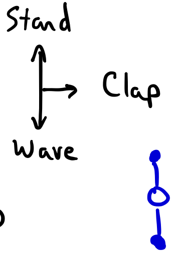
Expected SARSA

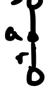


$$Q(S,A) \leftarrow Q(S,A) + \alpha \left[\text{Target} - Q(S,A) \right]$$

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

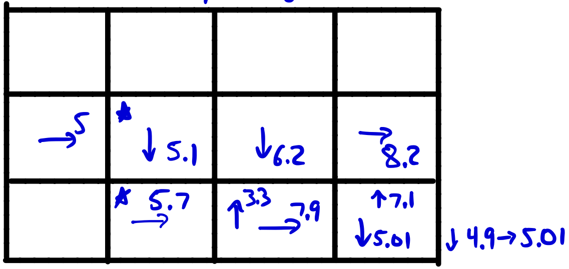






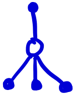
SARSA ← on policy

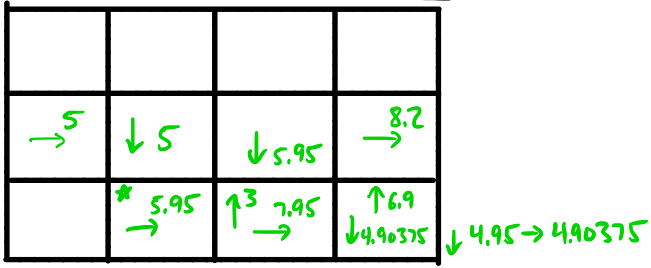
★= policy change



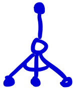
Policy: ε-greedy, ε=.75; Ties: →, ↓

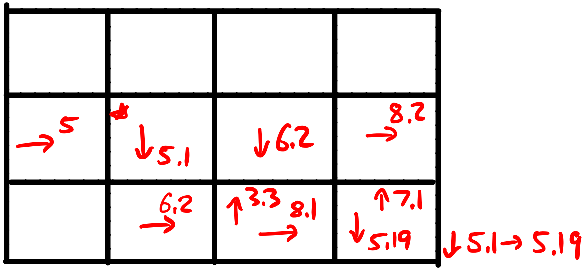
Expected SARSA ← Can be off policy





Q-learning ← off policy

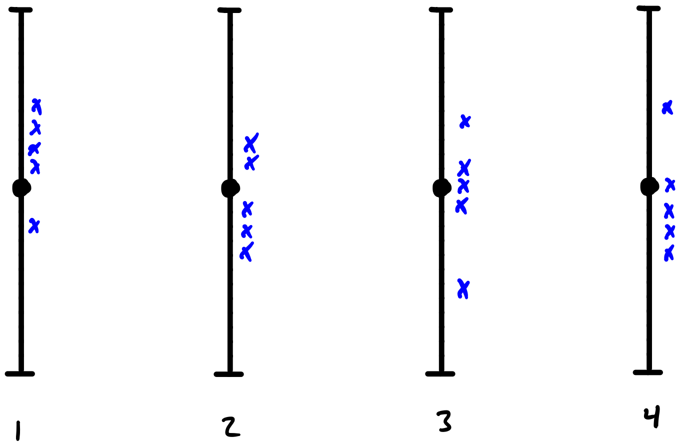




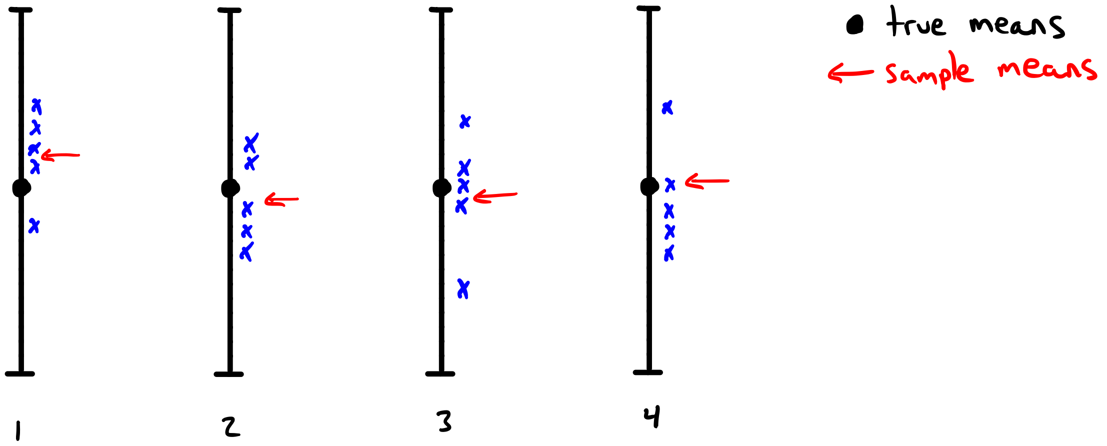
- How do learned policies differ?
- Conditions for convergence?
- Why Expected SARSA not as known?

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

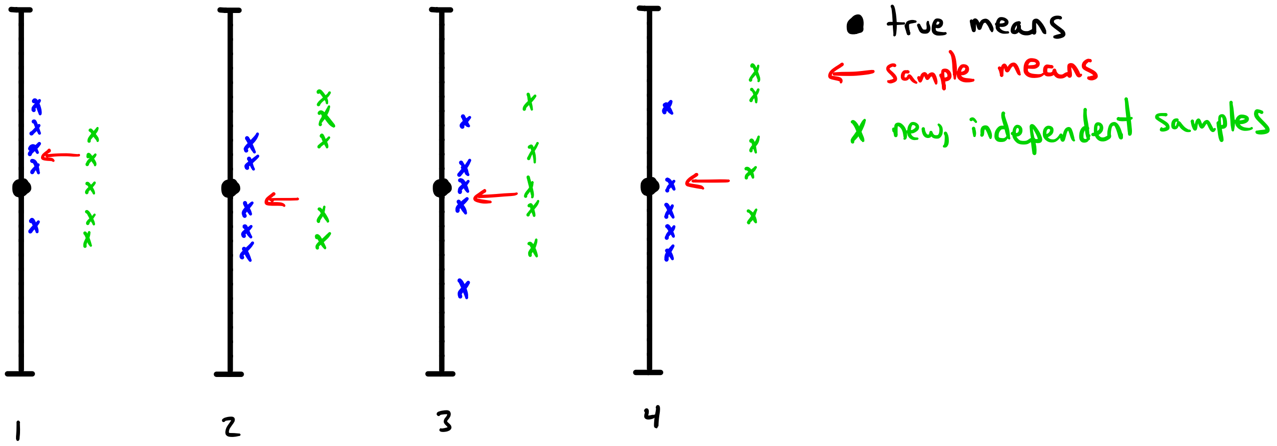
Double Q learning – addresses maximization bias, regression to the mean (illustrated in bandit setting)



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



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



Ch 6 summary

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

Control: SARSA
Q-learning } expected SARSA

The core: <u>bootstrapping</u>

Also: Double Q

After states