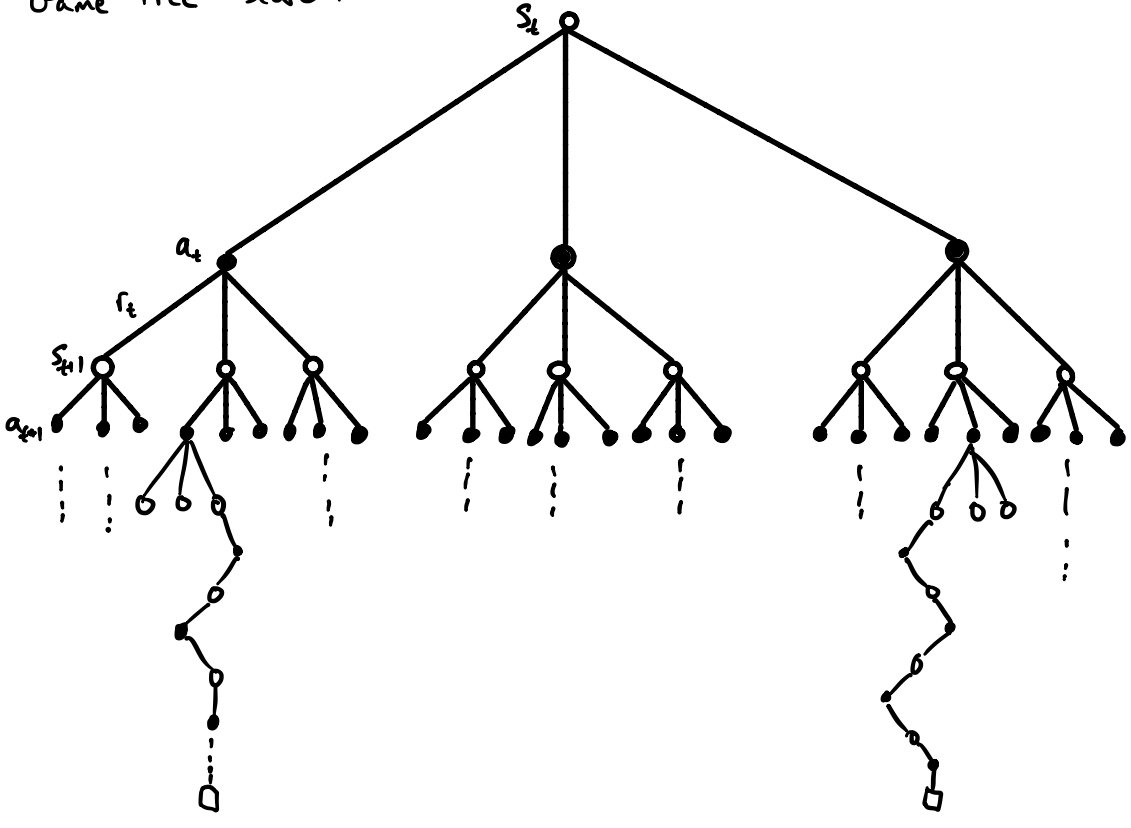
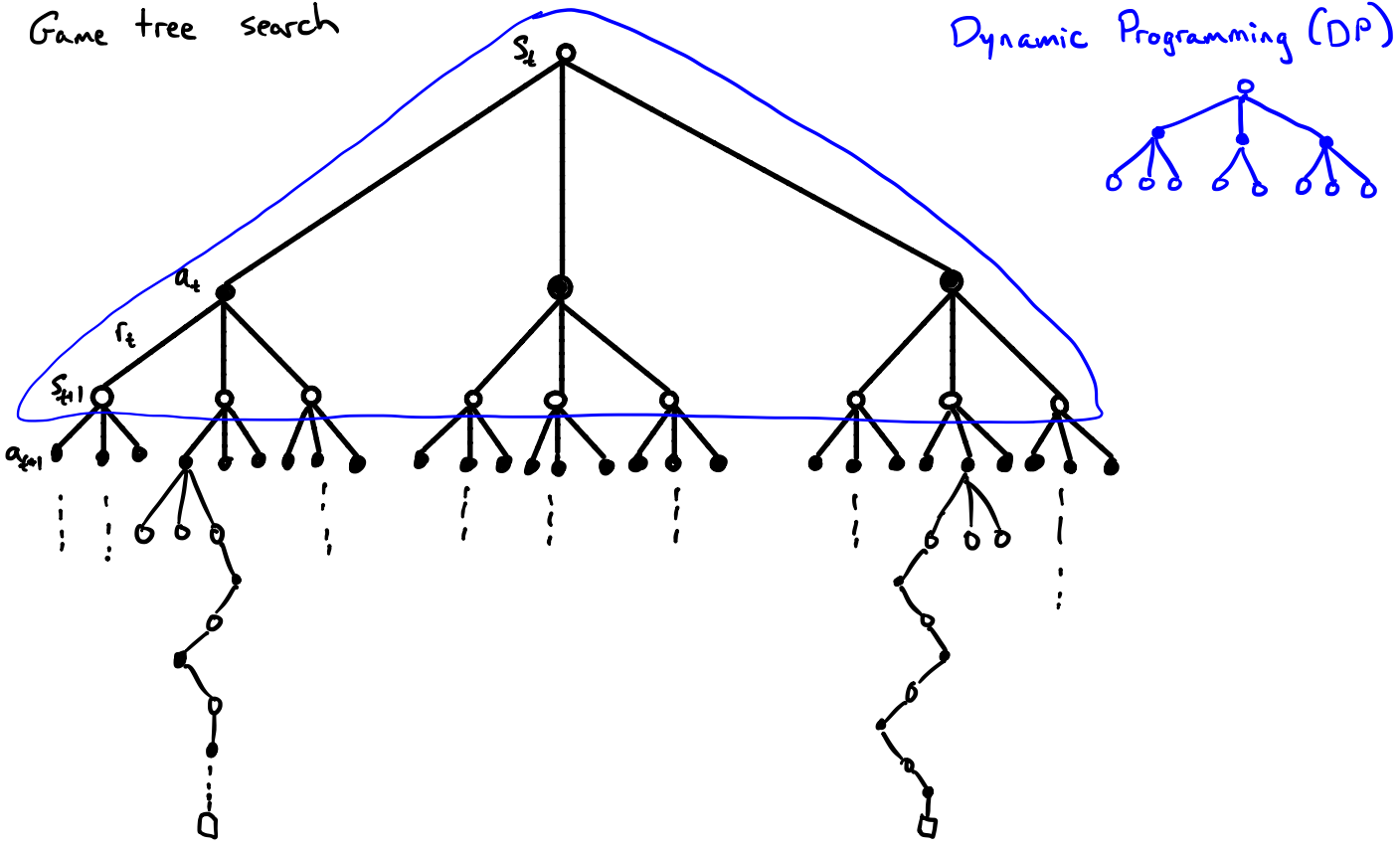
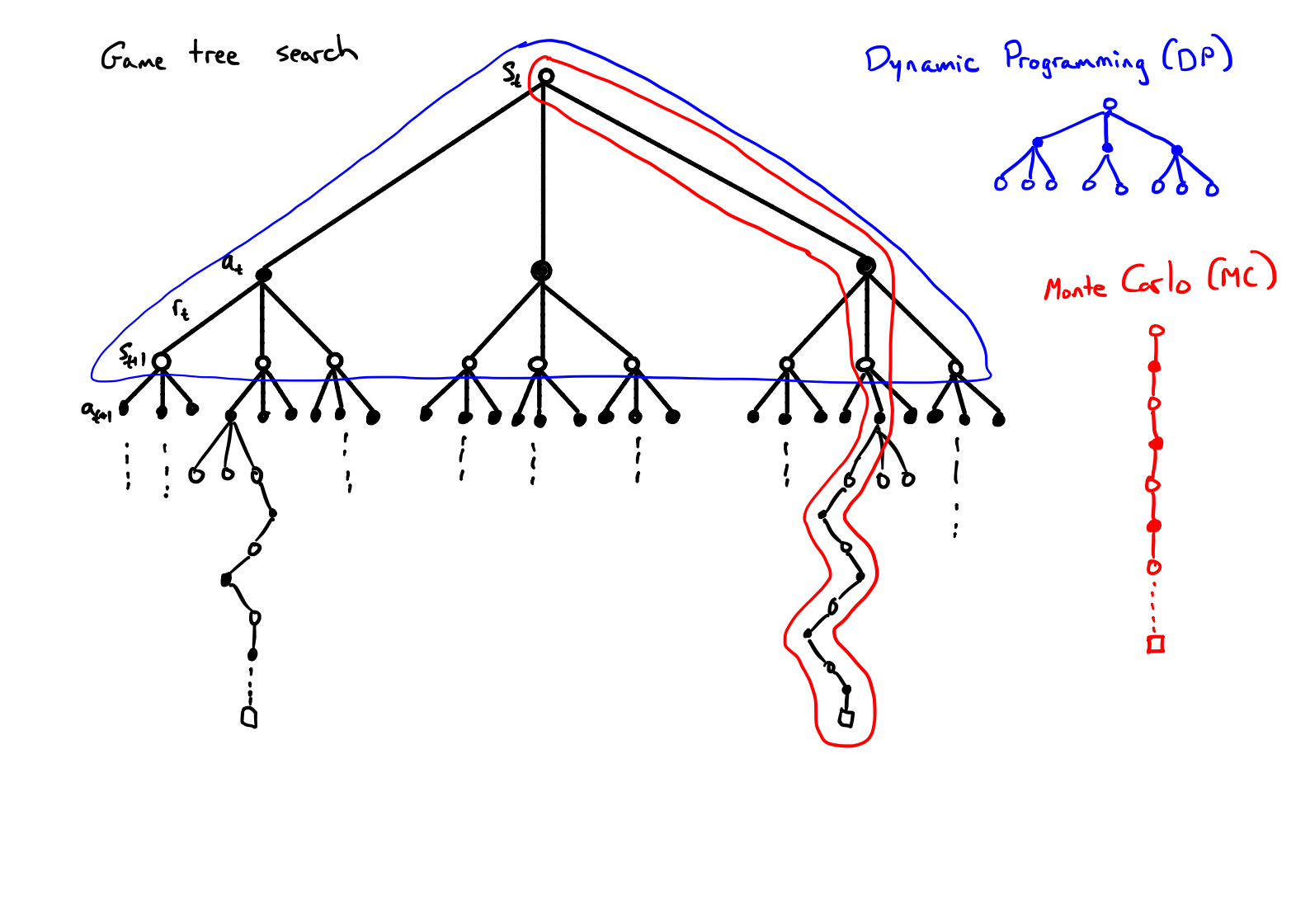
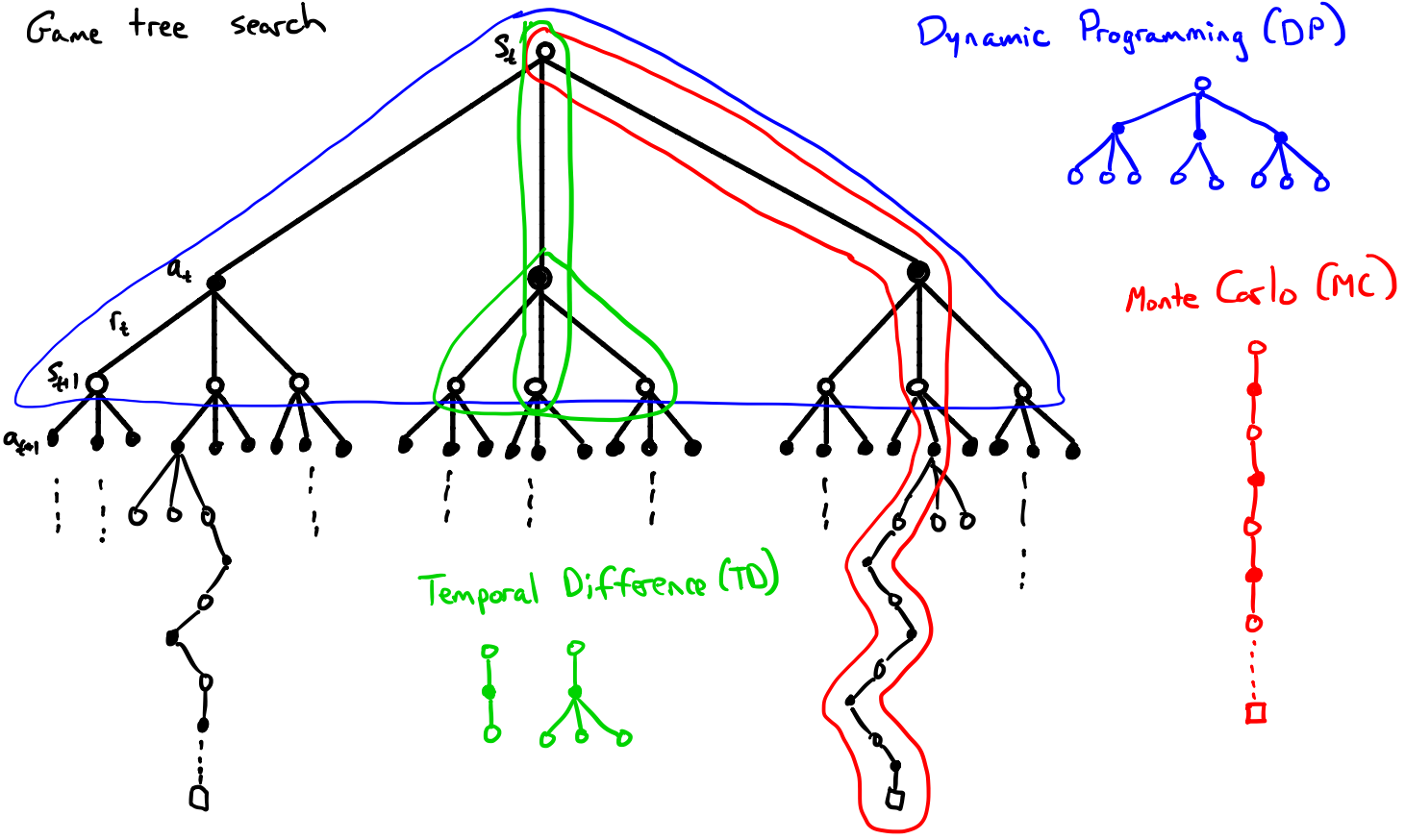
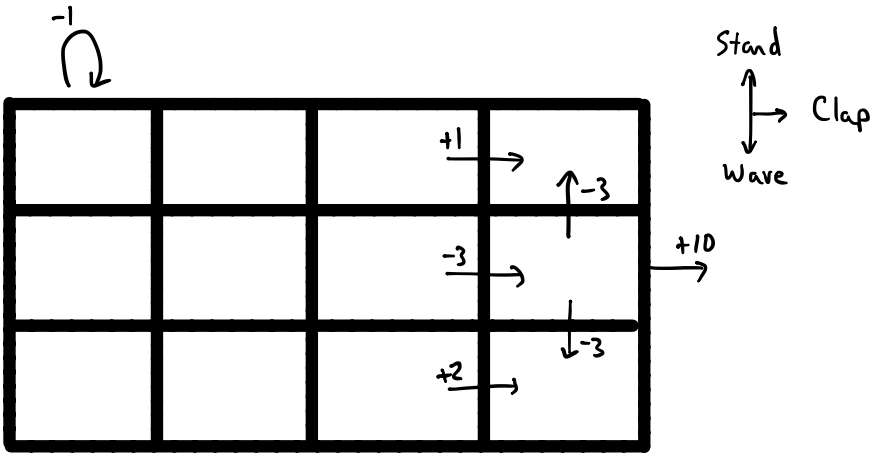
Game tree search

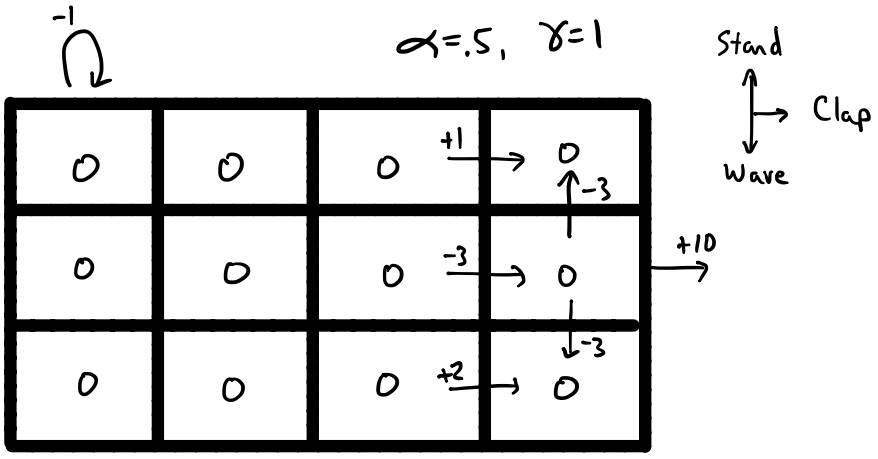


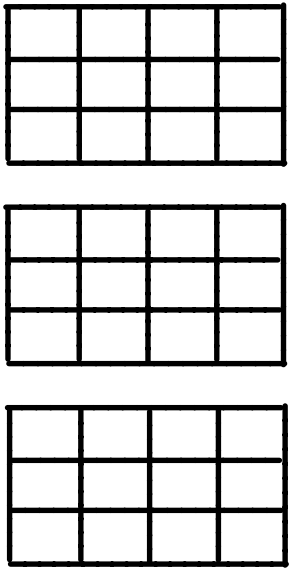

Game tree search
s_t
a_t
r_t
s_{t+1}
a_{t+1}
Dynamic Programming (DP)


Game tree search
s_t
a_t
r_t
s_{t+1}
a_{t+1}
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)

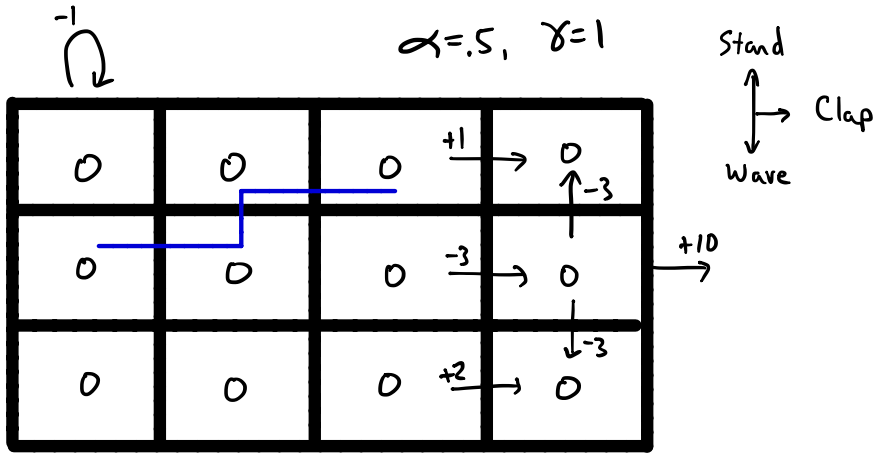

-1
+1
-3
-3
+10
-3
+2
Stand
Clap
Wave

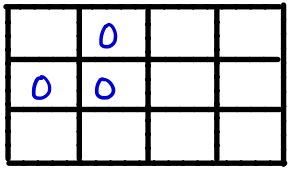


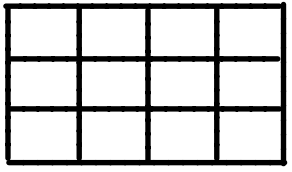


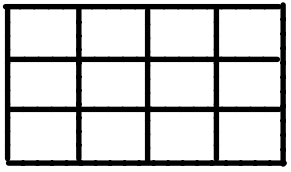
$$V(s) \leftarrow V(s) + \alpha[R + \gamma V(s') - V(s)]$$

"target"



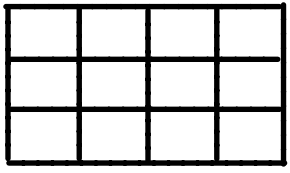


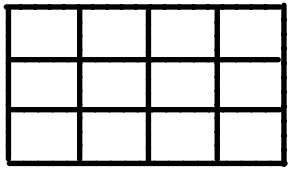


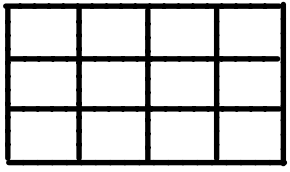


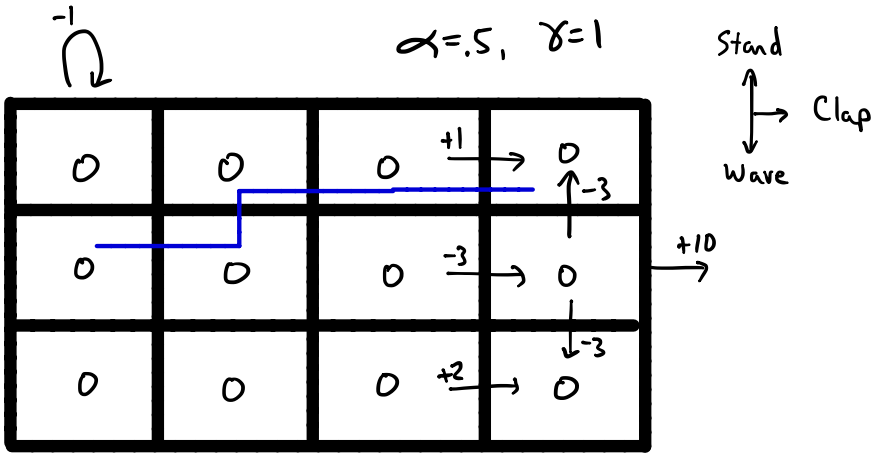
$$V(s) \leftarrow V(s) + \alpha[R + \gamma V(s') - V(s)]$$

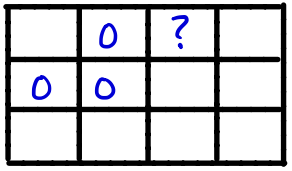
"target"

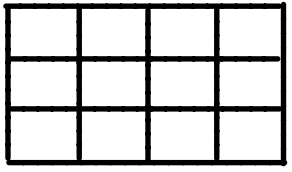


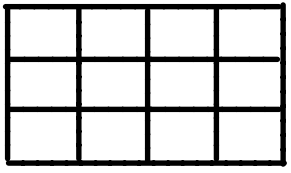






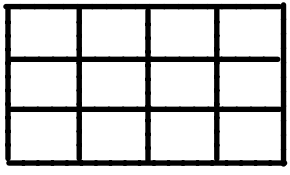


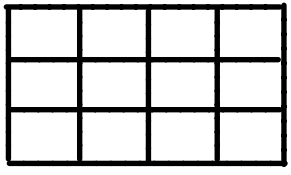


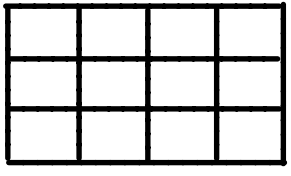


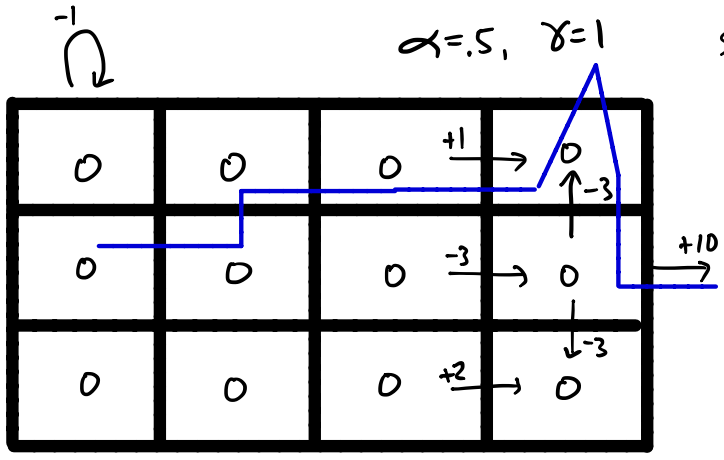
$$V(s) \leftarrow V(s) + \alpha[R + \gamma V(s') - V(s)]$$

"target"





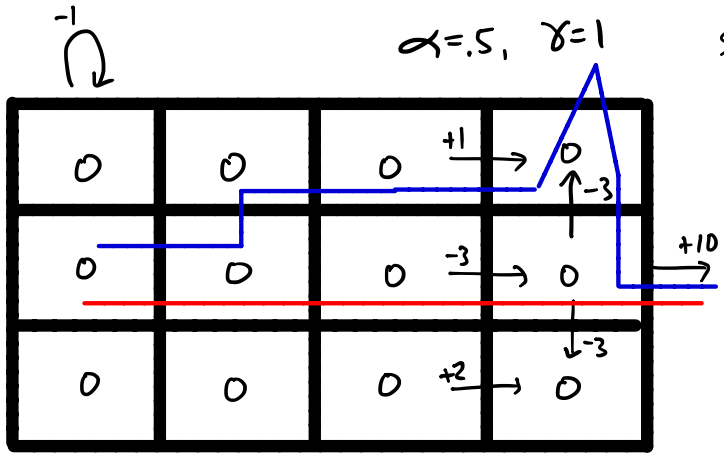




	0	.5	?
0	0		?

$$V(s) \leftarrow V(s) + \alpha[R + \gamma V(s') - V(s)]$$

"target"

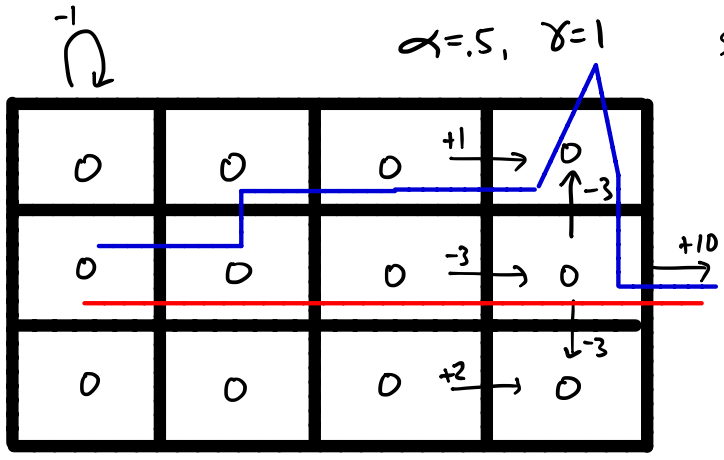


	0	.5	-.25
0	0		5

0	0	?	?

$$V(s) \leftarrow V(s) + \alpha[R + \gamma V(s') - V(s)]$$

"target"

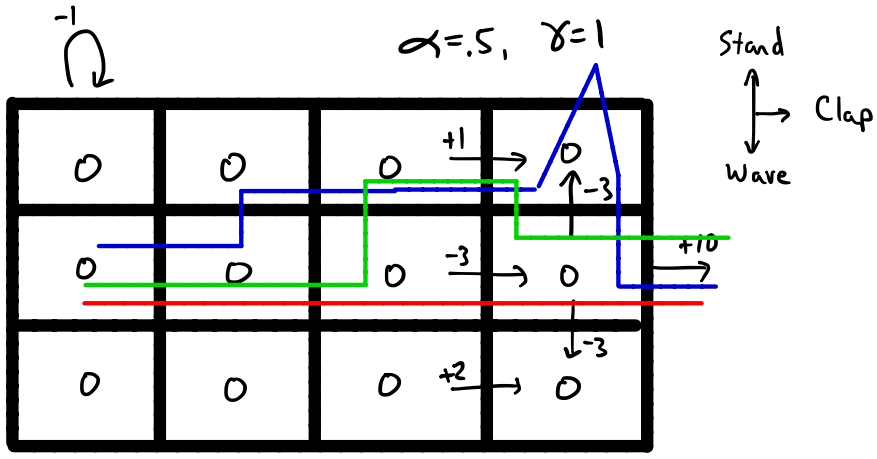


	0	.5	-.25
0	0		5

0	0	1	7.5

$$V(s) \leftarrow V(s) + \alpha[R + \gamma V(s') - V(s)]$$

"target"



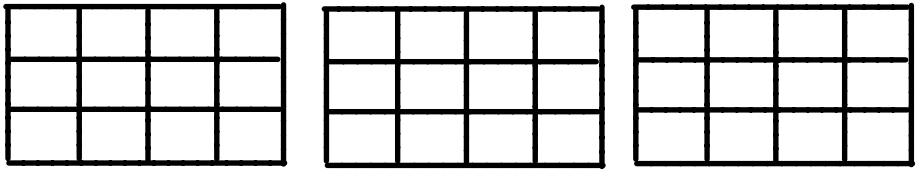
	0	.5	-.25
0	0		5

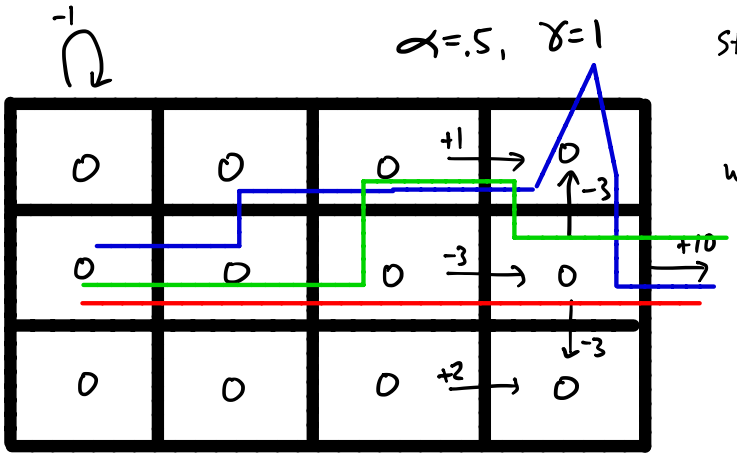
0	0	.5	-.25
0	0	1	7.5
0	0	0	0

		?	?
0	.5	.75	8.75

$$V(s) \leftarrow V(s) + \alpha[R + \gamma V(s') - V(s)]$$

"target"





TD

	0	.5	-.25
0	0		5

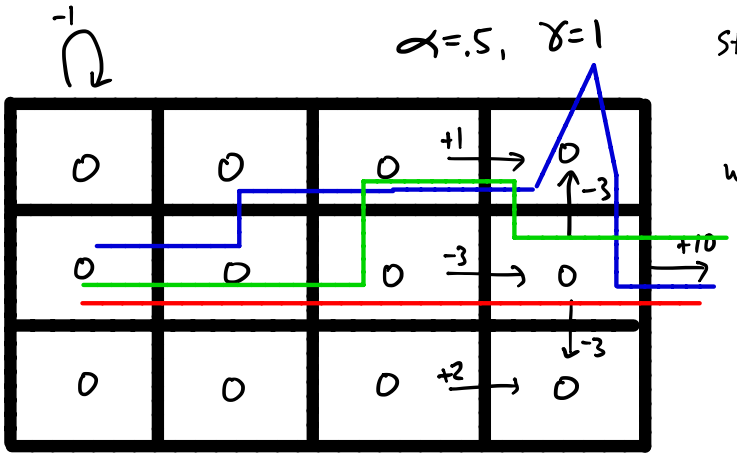
0	0	.5	-.25
0	0	1	7.5
0	0	0	0

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

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

MC (first visit)

	?	?	?
?	?		?



TD

	0	.5	-.25
0	0		5

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

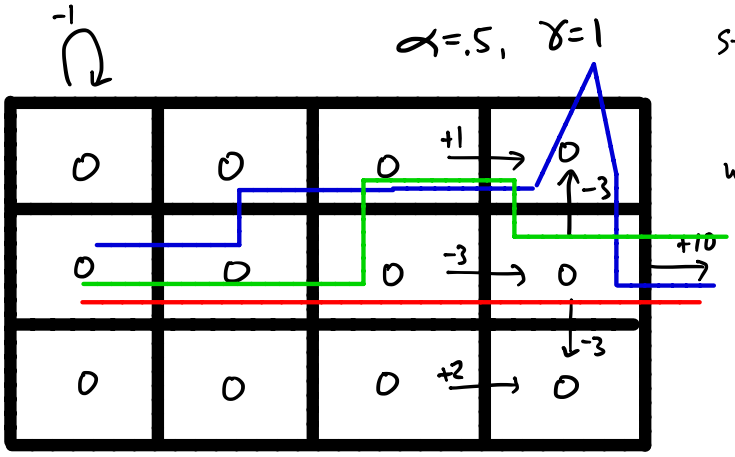
0	0	.625	3.625
0	.5	.75	8.75
0	0	0	0

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

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

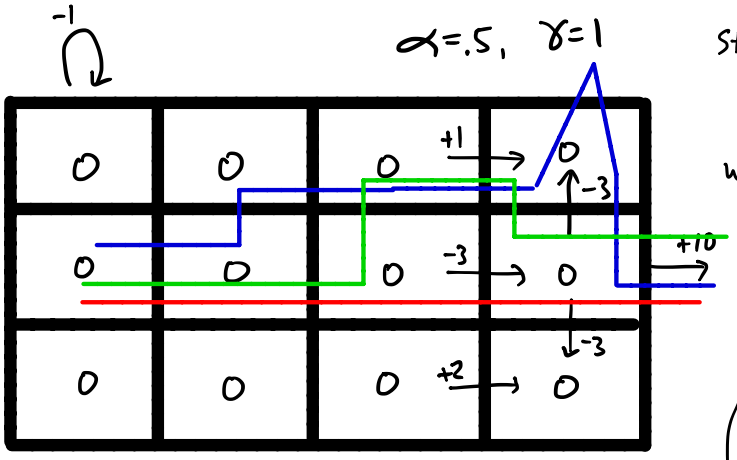
$$V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$$

"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



Stand
Clap
Wave

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

$$V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$$

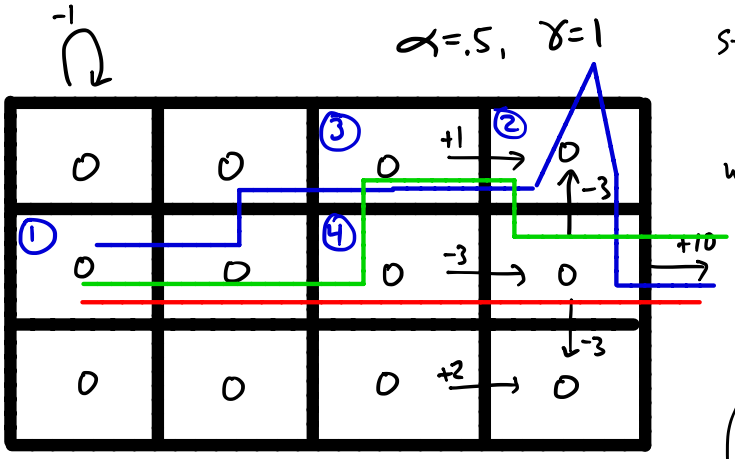
"target"

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



Stand
Clap
Wave

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

$$V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$$

"target"

MC (first visit)

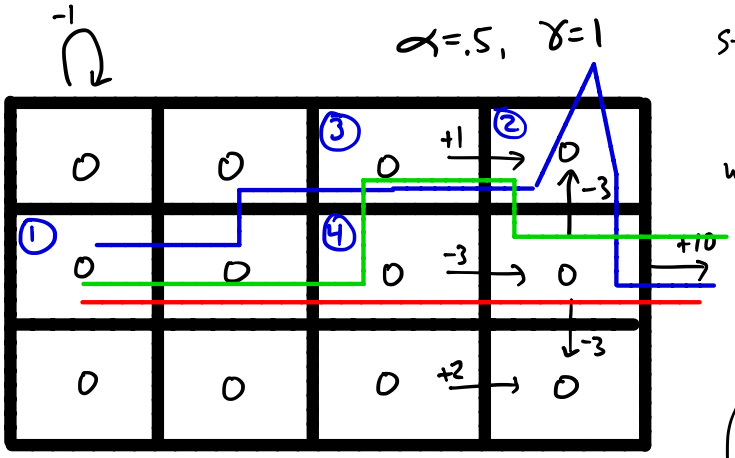
	5	5	4.5
5	5		5

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	?	?



Stand
Clap
Wave

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

$$V(S) \leftarrow V(S) + \alpha[R + \gamma V(S') - V(S)]$$

"target"

MC (first visit)

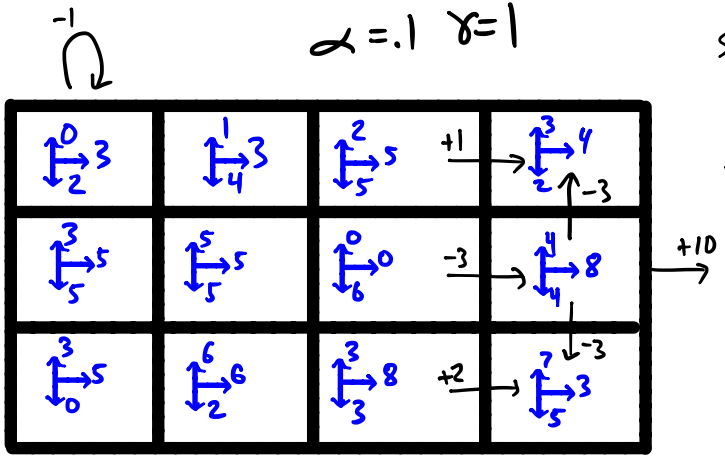
	5	5	4.5
5	5		5

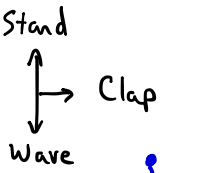
0	5	5	4.5
6	6	3.5	7.5
0	0	0	0

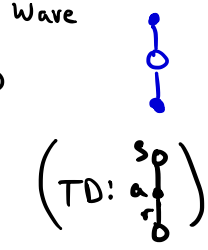
		8	7.25
8.5	8.5	7.25	8.75

Convergence of Batch methods

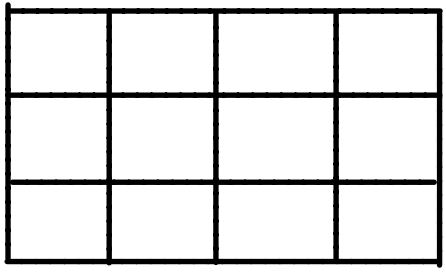
	Batch MC	Batch TD
1	avg(7,11,10)=9.33	
2		9.5
3		10.5
4	avg(10,5,7)=8.75	avg(11,7)=9





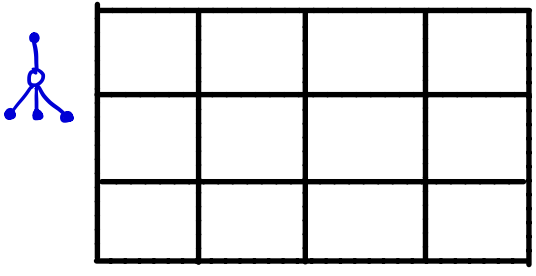


SARSA

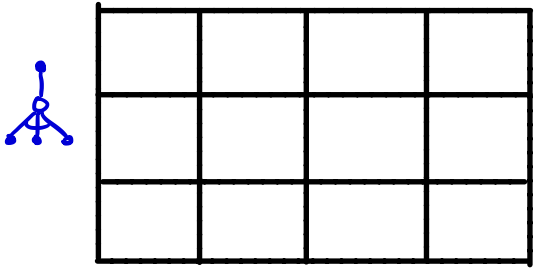


Policy: ϵ-greedy, $\epsilon = .75$; Ties: →, ↓

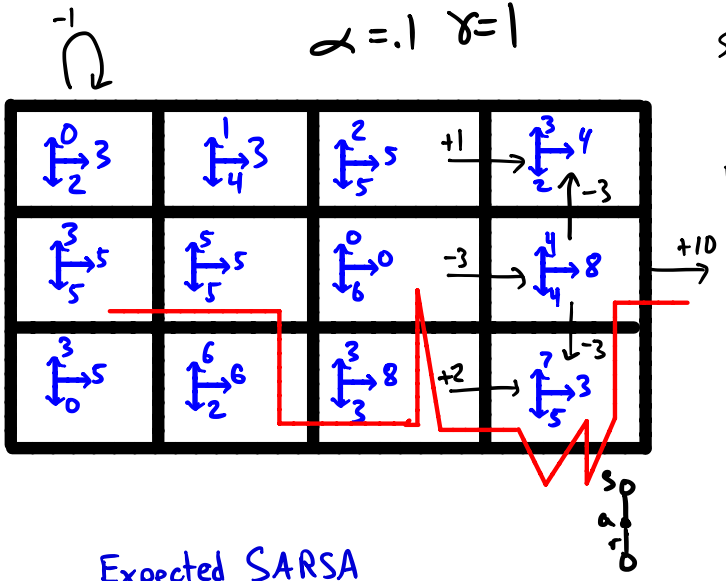
Expected SARSA

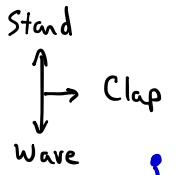


Q-learning

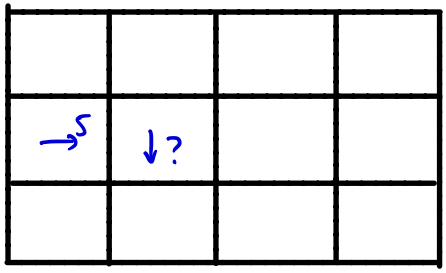


$$Q(S,A) \leftarrow Q(S,A) + \alpha \left[\text{Target} - Q(S,A) \right]$$



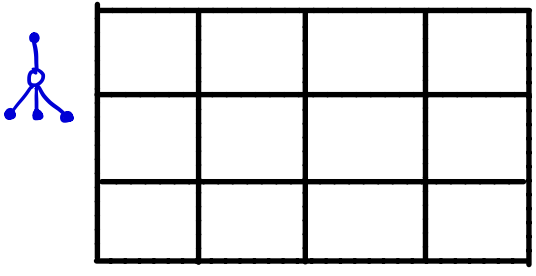


SARSA

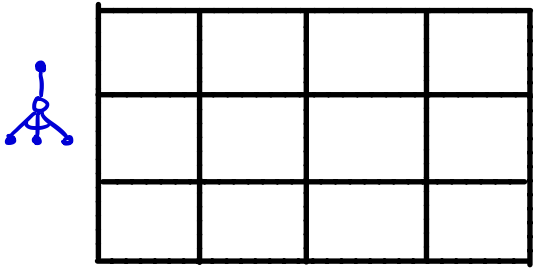


Policy: ϵ-greedy, $\epsilon = .75$; Ties: →, ↓

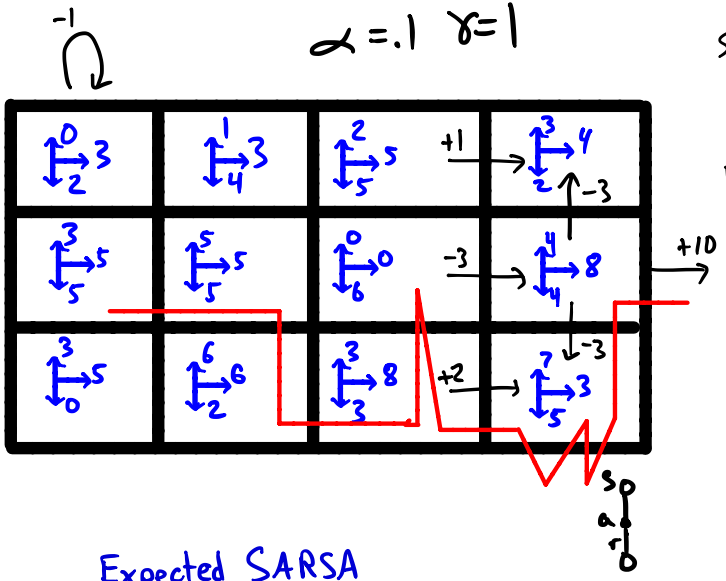
Expected SARSA

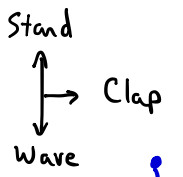


Q-learning

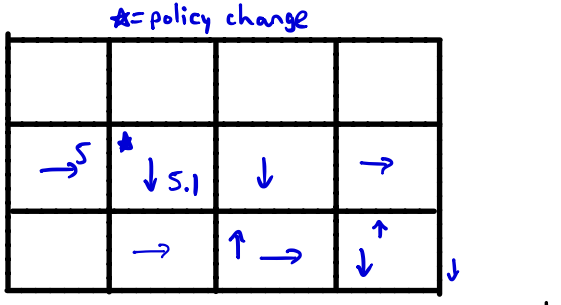


$$Q(S,A) \leftarrow Q(S,A) + \alpha\,[\text{Target} - Q(S,A)]$$



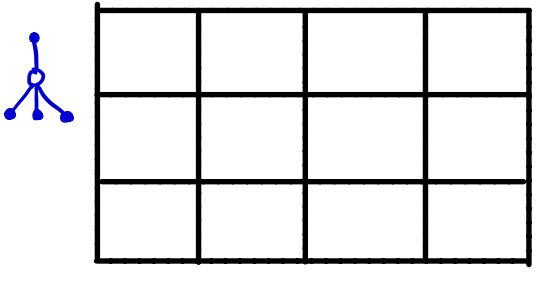


SARSA

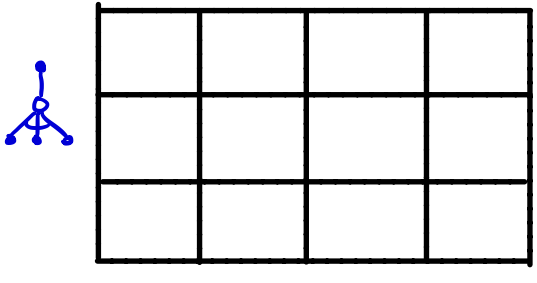


Policy: ϵ-greedy, $\epsilon = .75$; Ties: →, ↓

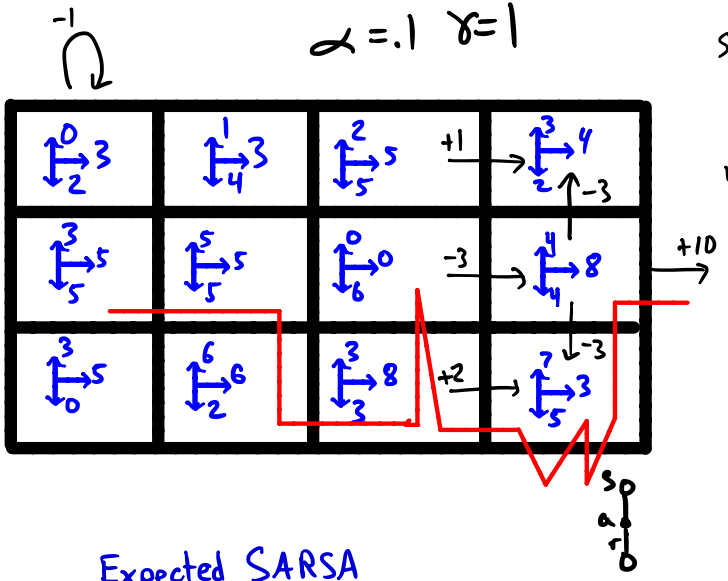
Expected SARSA

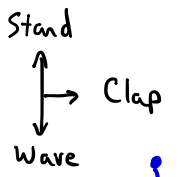


Q-learning

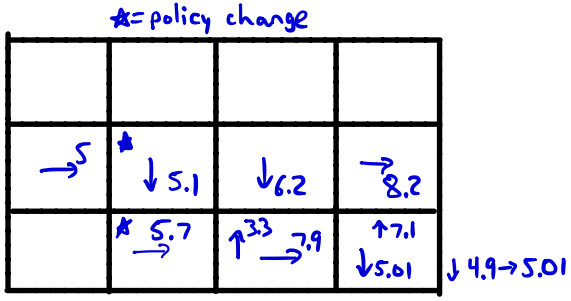


$$Q(S,A) \leftarrow Q(S,A) + \alpha\,[\text{Target} - Q(S,A)]$$



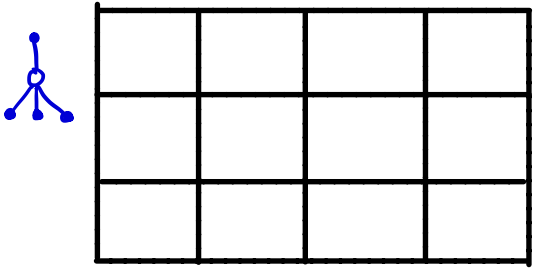


SARSA



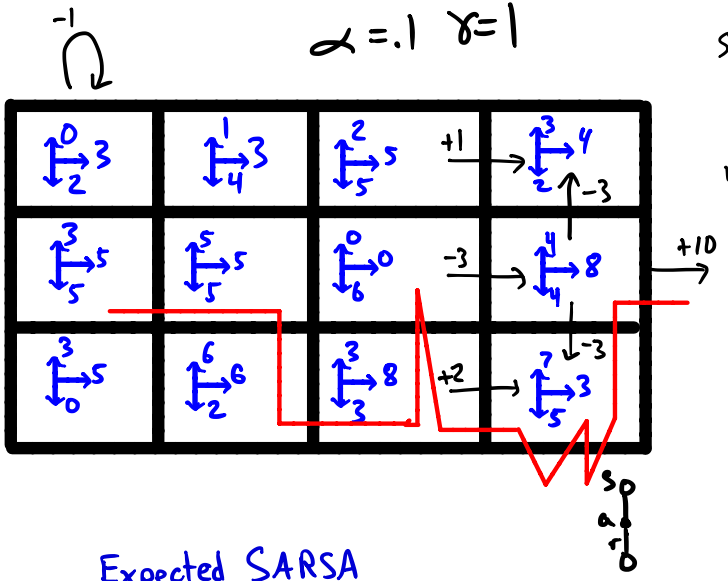
Policy: ϵ-greedy, $\epsilon = .75$; Ties: →, ↓

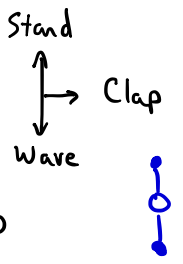
Expected SARSA



Q-learning

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





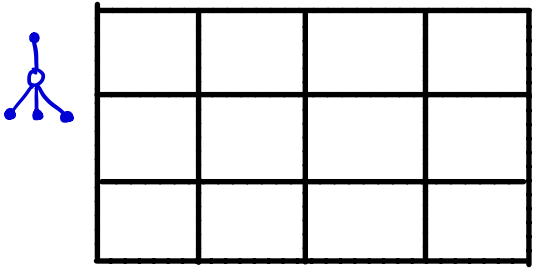
SARSA

★= policy change

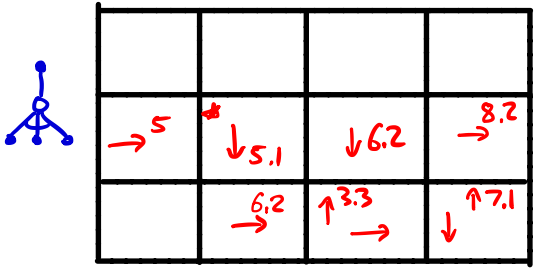
→5	★↓5.1	↓6.2	→8.2
	★→5.7	↑3.3 →7.9	↑7.1 ↓5.01 ↓4.9→5.01

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

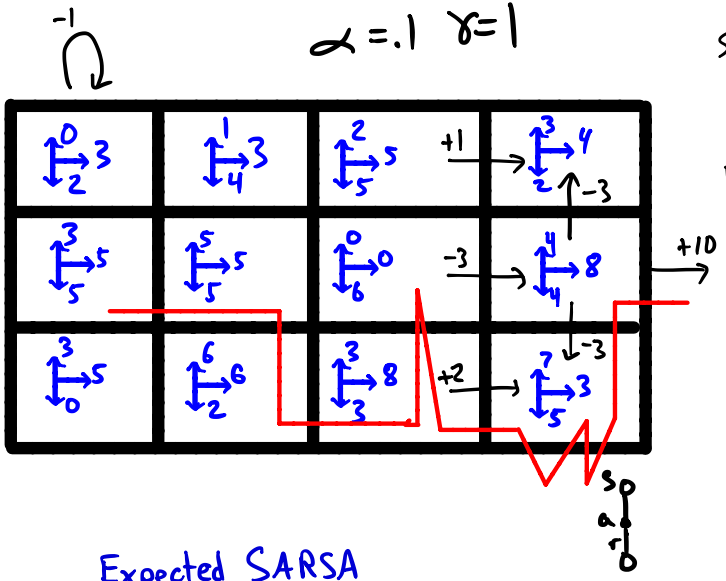
Expected SARSA

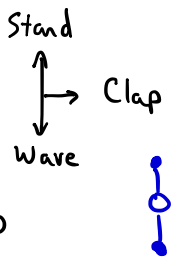


Q-learning



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





SARSA

★= policy change

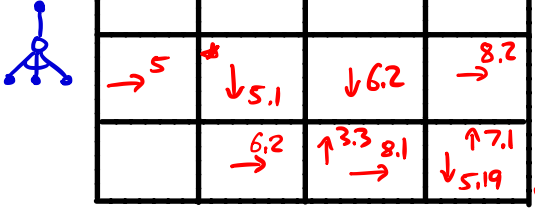
→ 5	★ ↓ 5.1	↓ 6.2	→ 8.2
	★ → 5.7	↑ 3.3 → 7.9	↑ 7.1 ↓ 5.01

↓ 4.9 → 5.01

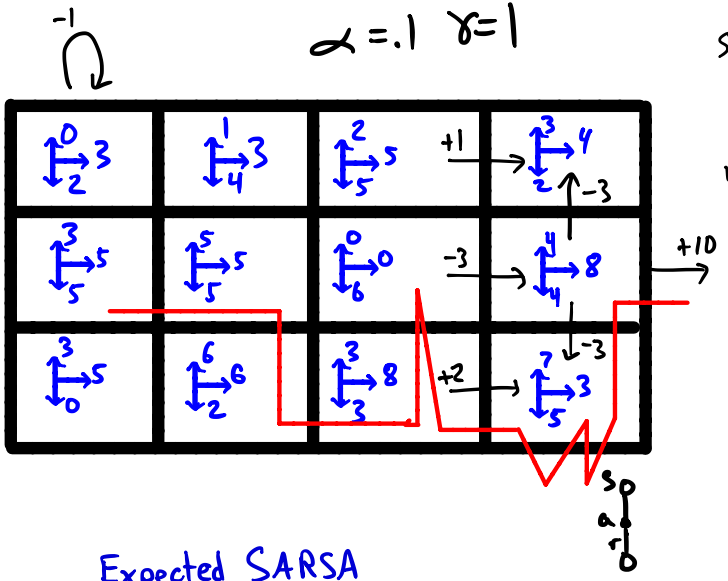
Policy: ϵ-greedy, $\epsilon = .75$; Ties: →, ↓

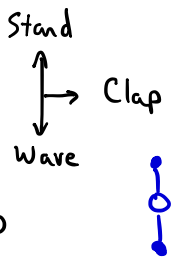
Expected SARSA

Q-learning



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





SARSA

★= policy change

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

↓4.9→5.01

Policy: ϵ-greedy, ϵ=.75; Ties: →, ↓

Expected SARSA

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

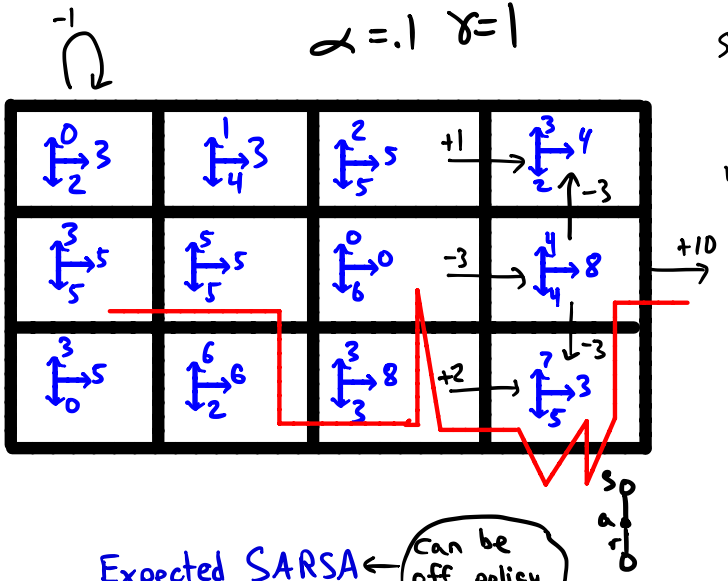
↓4.95→4.90375

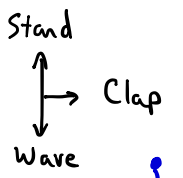
Q-learning

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

↓5.1→5.19

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





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: ϵ-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

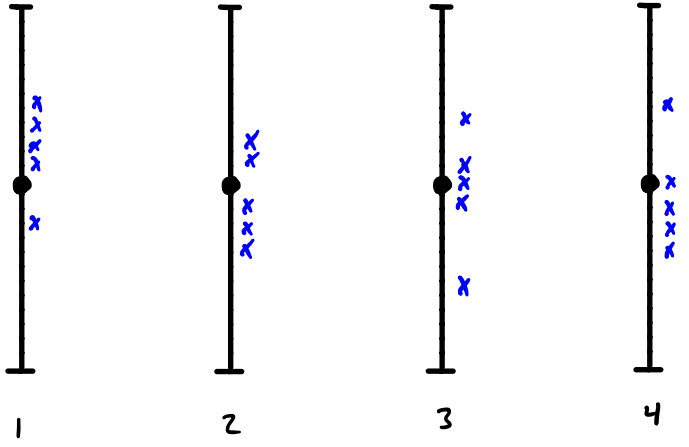
Q-learning ← off policy

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

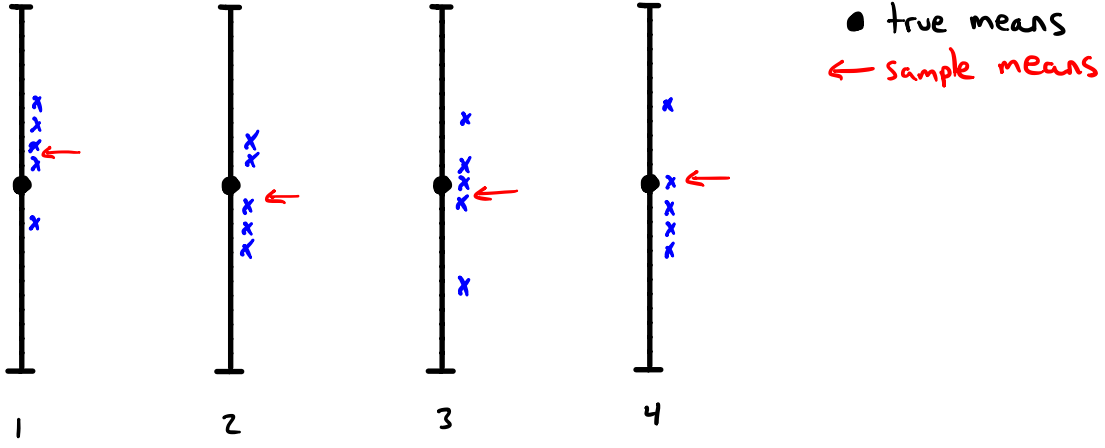
- How do learned policies differ?
- Conditions for convergence?
- Why Expected SARSA not as known?

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

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



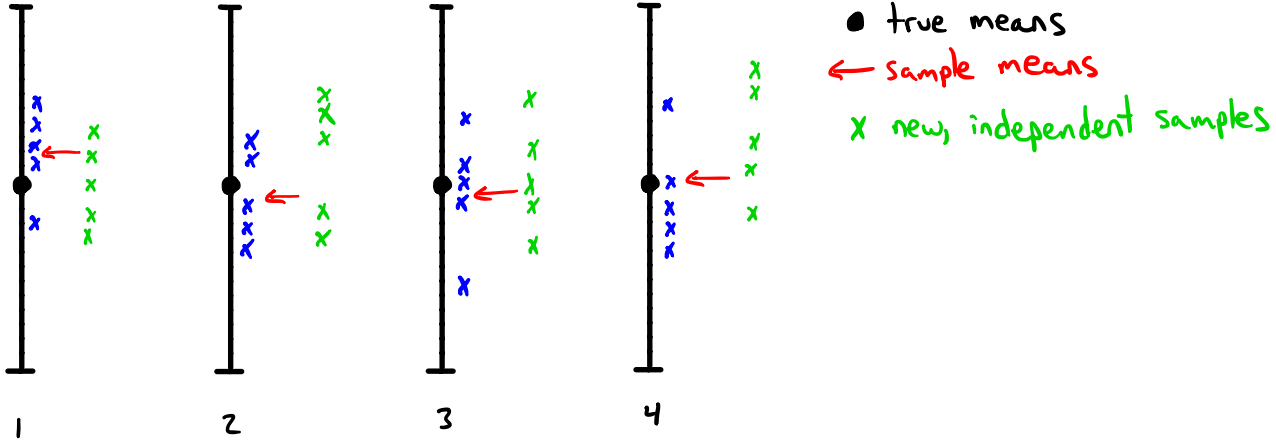
Double Q learning – addresses maximization bias, regression to the mean (illustrated in bandit setting)



● true means

← sample means

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: bootstrapping

Also: Double Q

After states