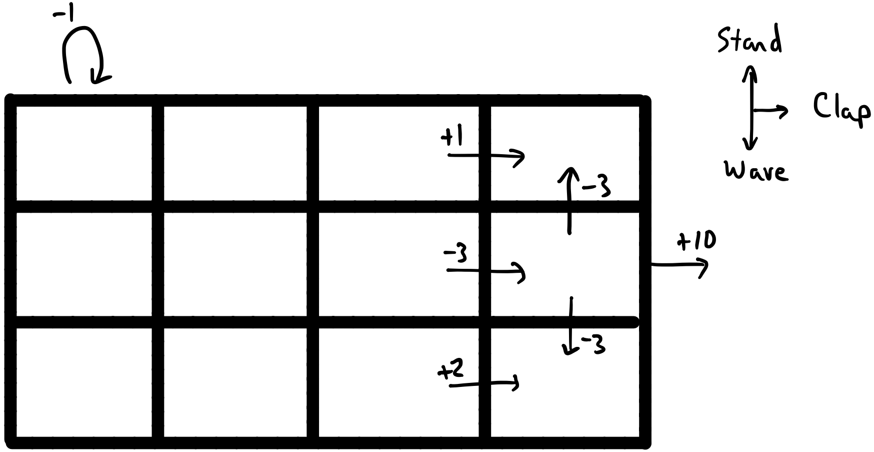
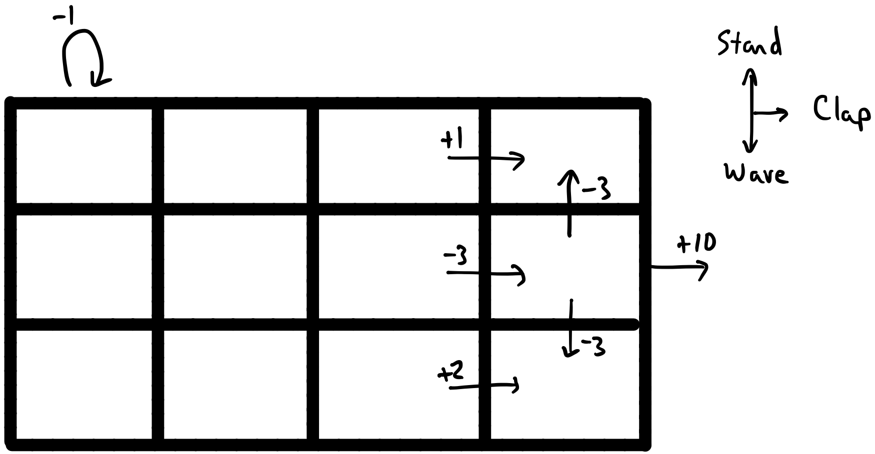
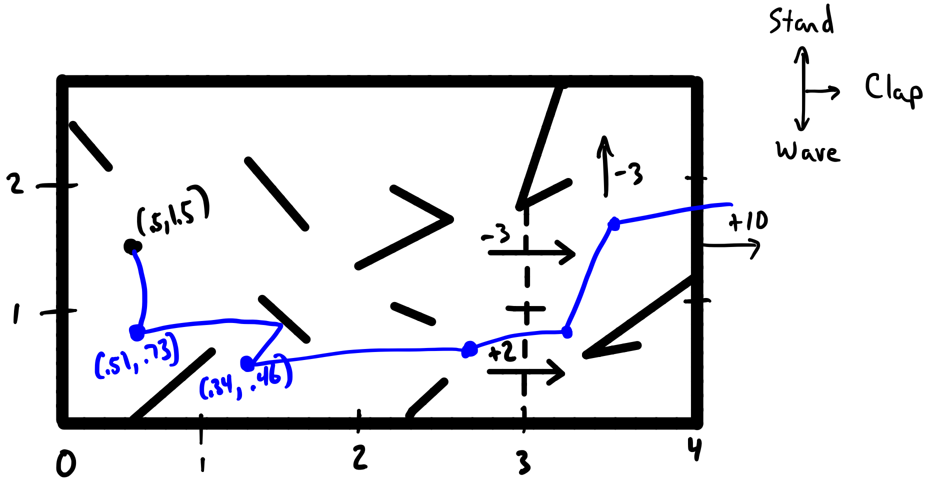
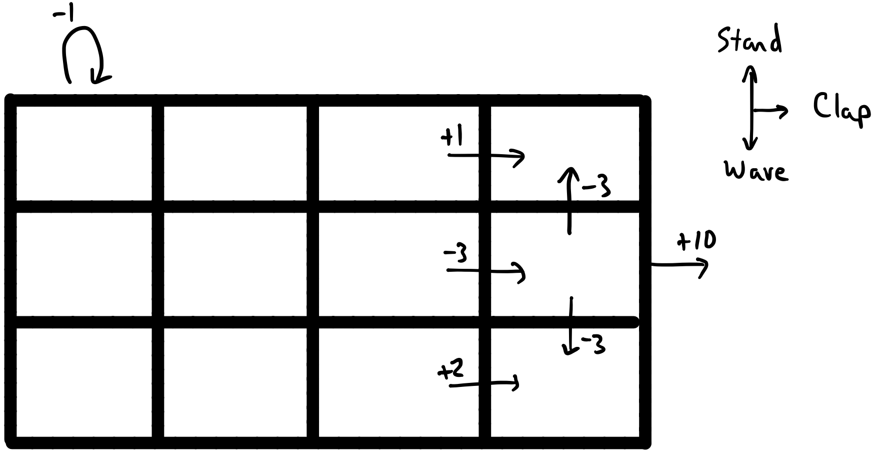
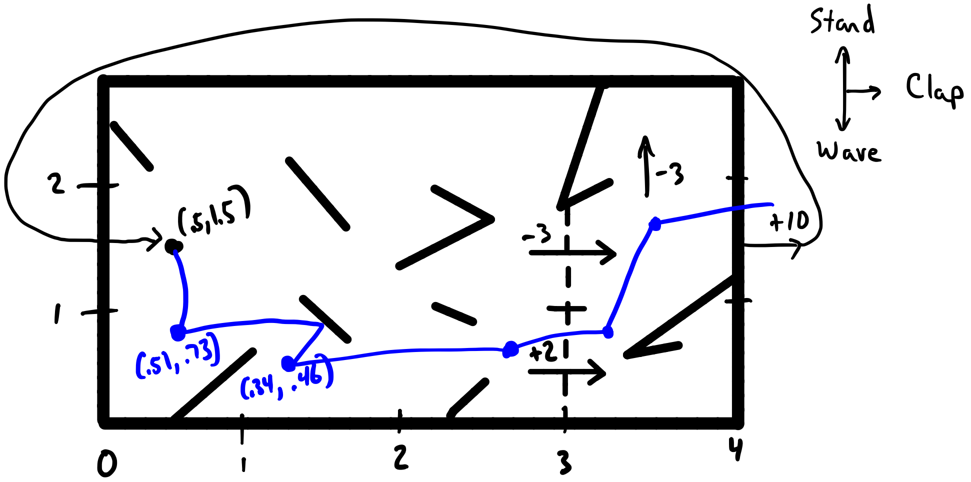

-1
+1
-3
-3
+10
-3
+2
-3
Stand
Clap
Wave

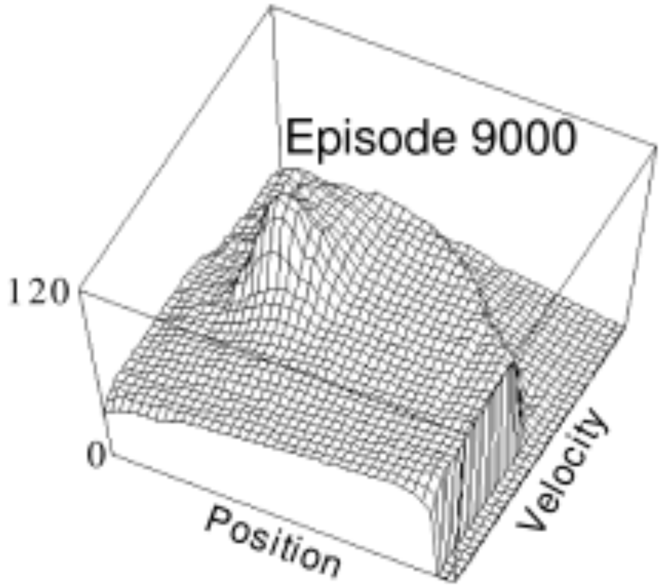

-1
+1
-3
-3
+10
-3
+2
Stand
Clap
Wave


Stand
Clap
Wave
2
1
0
1
2
3
4
(.5, 1.5)
-3
-3
+10
+2
(.51, .73)
(.34, .46)


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

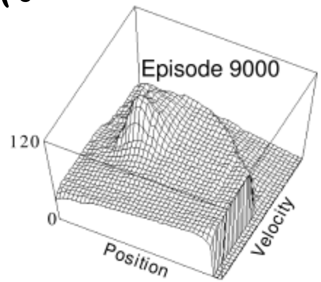

Stand
Clap
Wave
-3
-3
+10
+2
(.5, 1.5)
(.51, .73)
(.34, .46)
0
1
2
3
4
1
2

Continuous state: V(s)

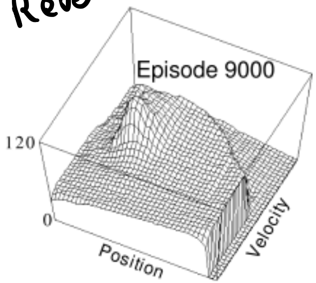


Continuous state : $Q(s,a)$

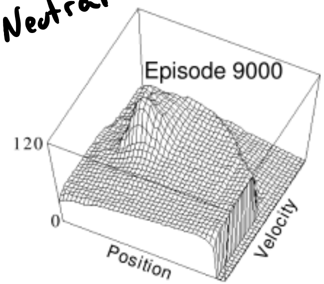
Forward



Reverse



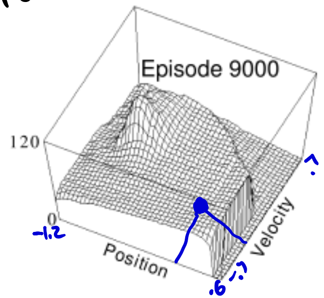
Neutral



Continuous state: $Q(s,a)$

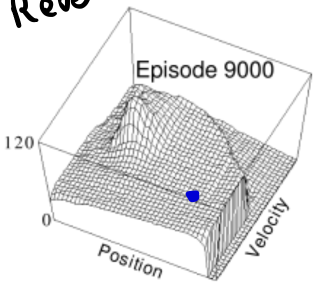
pos $= .3$, vel $= -.3$

Forward



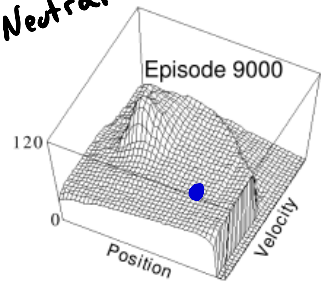
$Q(s, \text{forward}) = 60$

Reverse



$Q(s, \text{reverse}) = 75$

Neutral

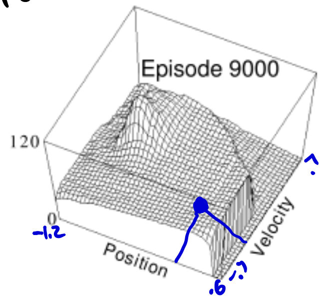


$Q(s, \text{neutral}) = 68$

Continuous state: $Q(s,a)$; <u>Discrete</u> actions

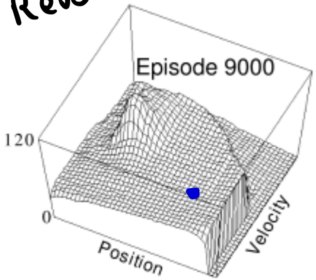
$pos = .3, \; vel = -.3$

Forward



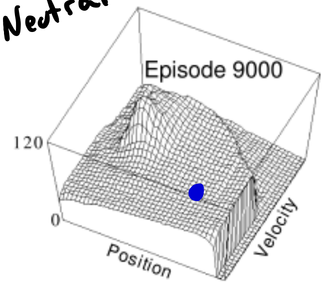
$Q(s, forward) = 60 \leftarrow$

Reverse



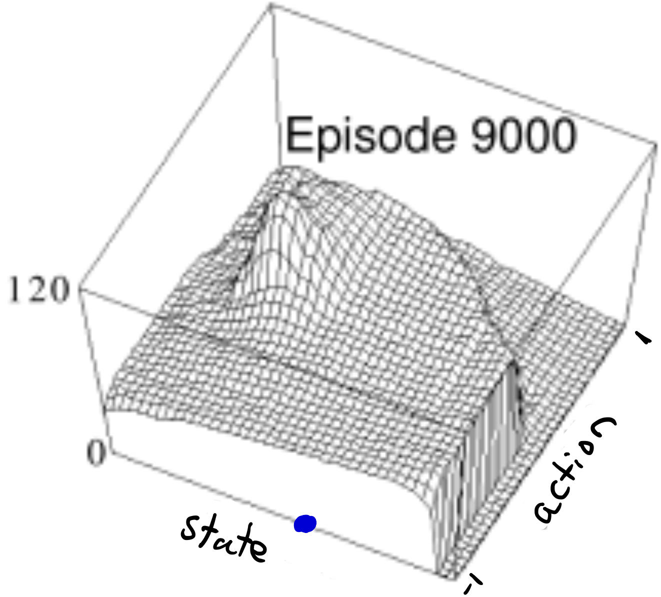
$Q(s, reverse) = 75$

Neutral

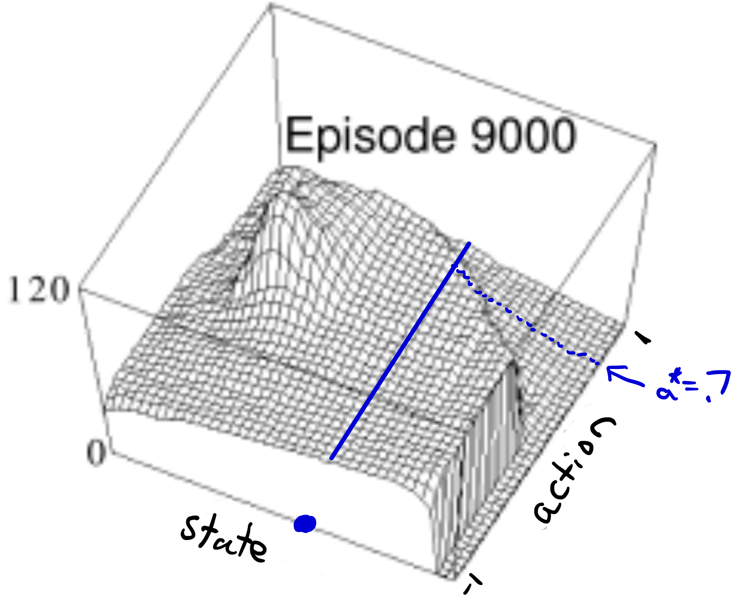


$Q(s, neutral) = 68$

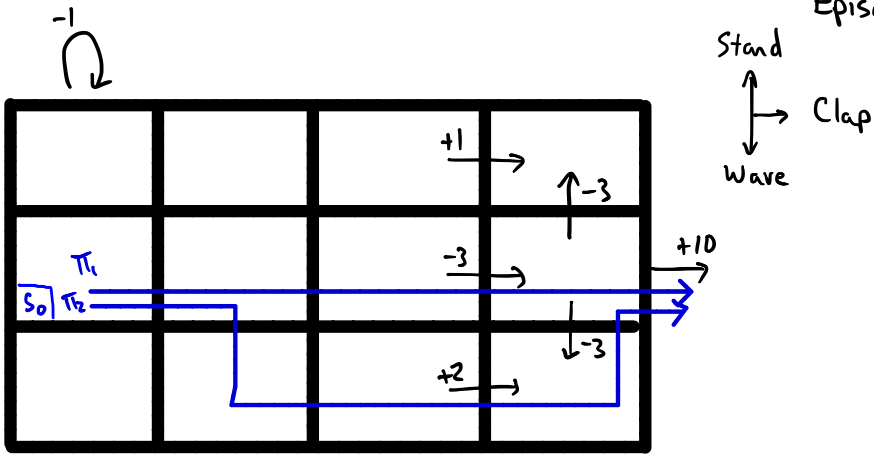
Continuous state, continuous action : $Q(s,a)$



Continuous state, continuous action : $Q(s,a)$



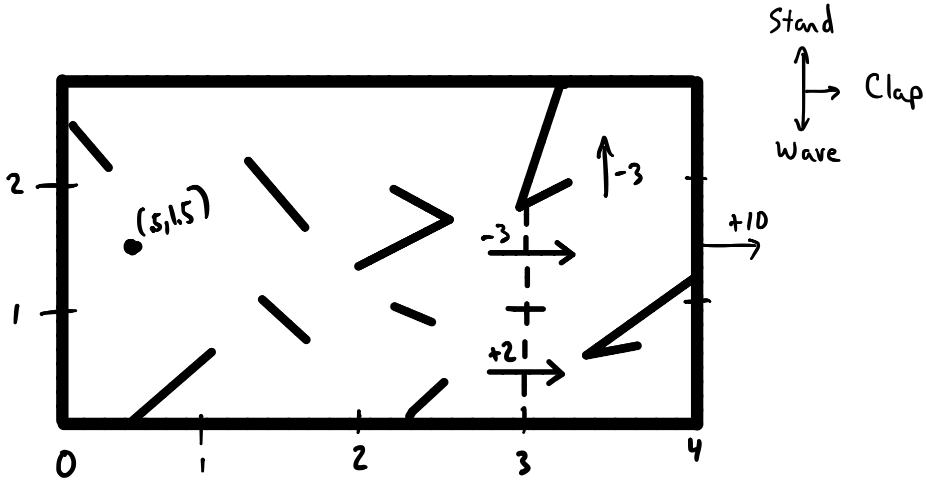
Episodic tasks



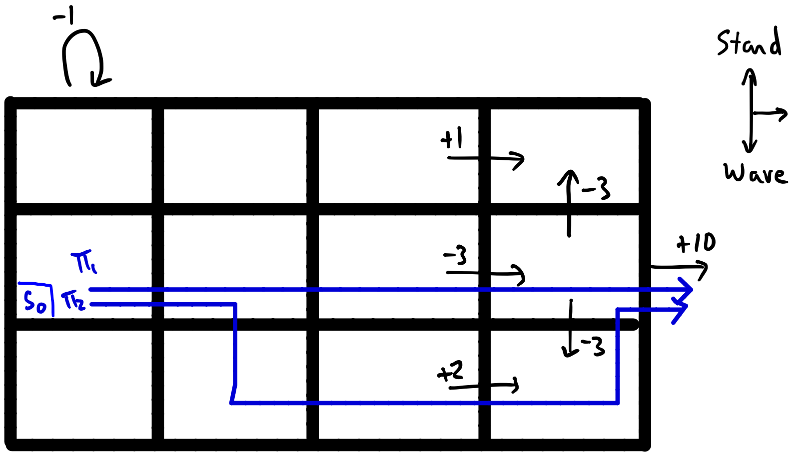
Discounting : γ

$V_{\pi_1}(s_0) =$

$V_{\pi_2}(s_0) =$



Episodic tasks

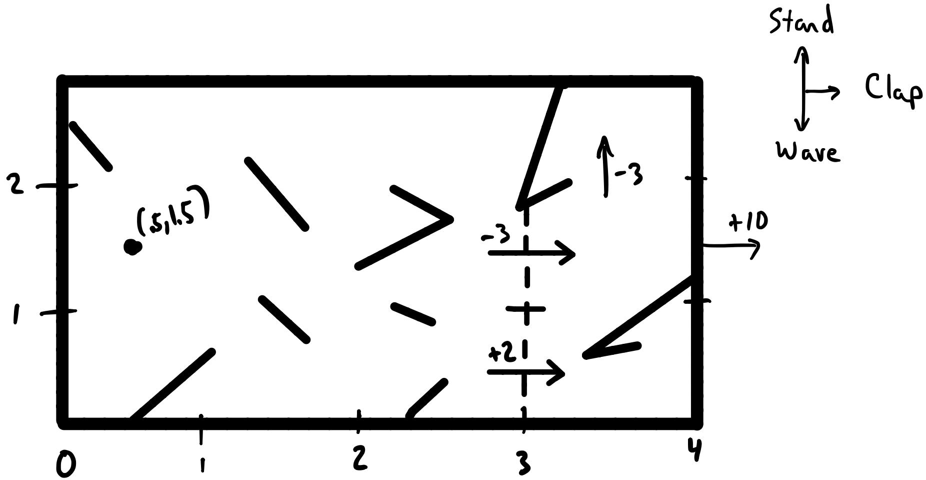


Discounting : γ

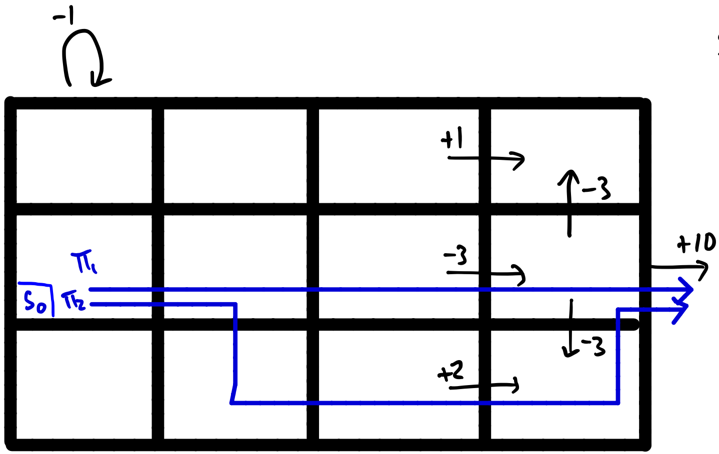
$$V_{\pi_1}(s_0) = 0 + 0 - 3\gamma^2 + 10\gamma^3$$

$$V_{\pi_2}(s_0) = 0 + 0 + 0 + 2\gamma^3 + 0 + 10\gamma^5$$

Which policy is better?



Episodic tasks



Stand
Clap
Wave

Discounting : γ

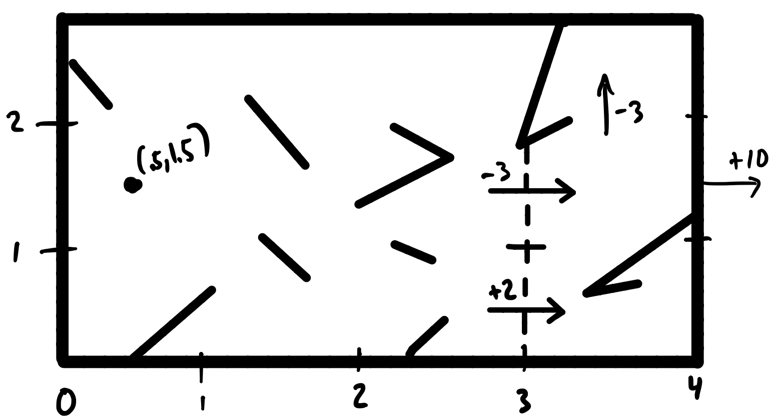
$v_{\pi_1}(s_0) = 0 + 0 - 3\gamma^2 + 10\gamma^3$

$v_{\pi_2}(s_0) = 0 + 0 + 0 + 2\gamma^3 + 0 + 10\gamma^5$

Which policy is better?

$\gamma = 1$: $v_{\pi_1}(s_0) = 7$ $\quad v_{\pi_2}(s_0) = 12$

$\gamma = .5$: $v_{\pi_1}(s_0) = 1.175$ $\quad v_{\pi_2}(s_0) = .5625$



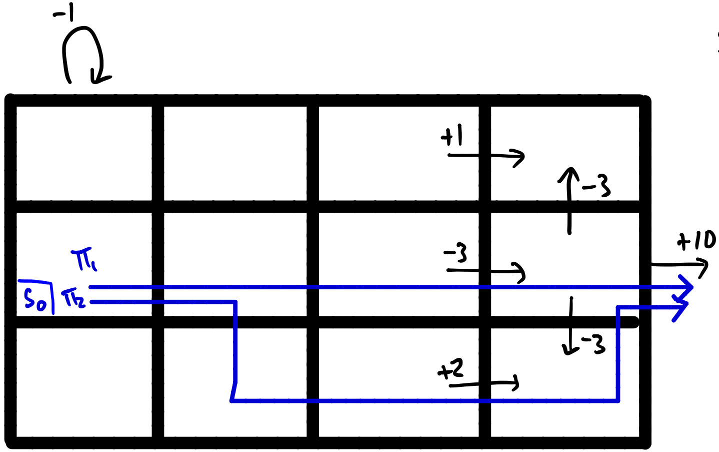
Stand
Clap
Wave

Two meaning of γ:

1)

2)

Episodic tasks



Stand
Clap
Wave

Discounting : γ

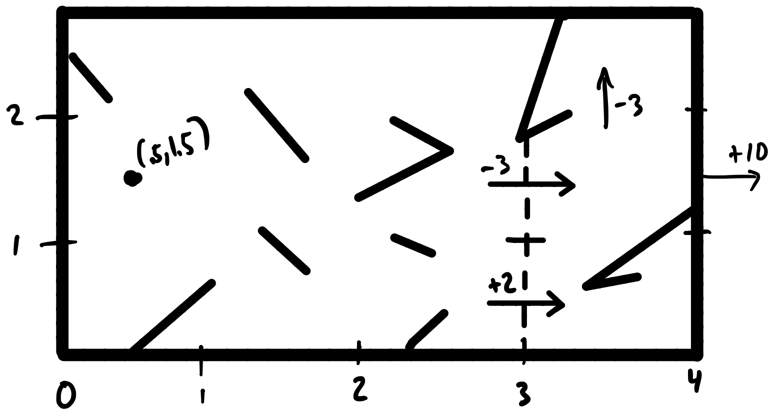
$v_{\pi_1}(s_0) = 0 + 0 - 3\gamma^2 + 10\gamma^3$

$v_{\pi_2}(s_0) = 0 + 0 + 0 + 2\gamma^3 + 0 + 10\gamma^5$

Which policy is better?

$\gamma = 1$: $v_{\pi_1}(s_0) = 7$ $\quad v_{\pi_2}(s_0) = 12$

$\gamma = .5$: $v_{\pi_1}(s_0) = 1.175$ $\quad v_{\pi_2}(s_0) = .5625$



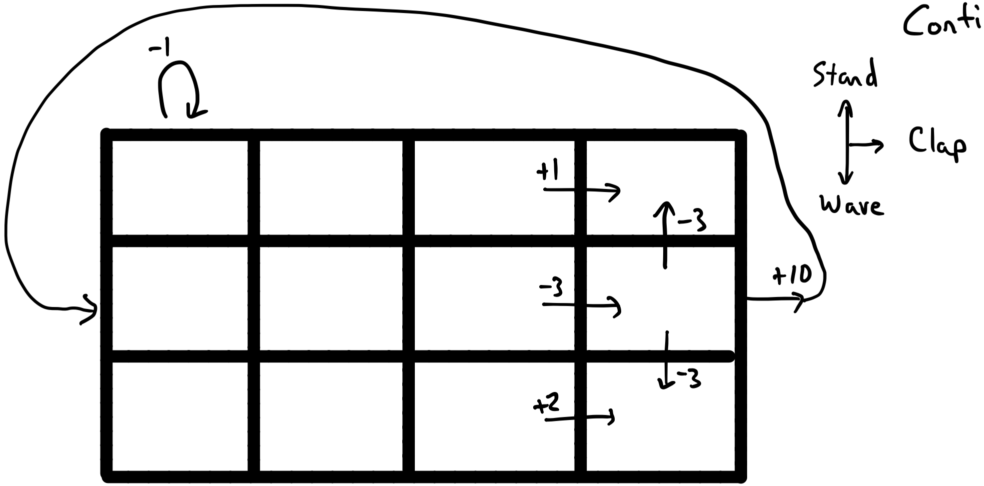
Stand
Clap
Wave

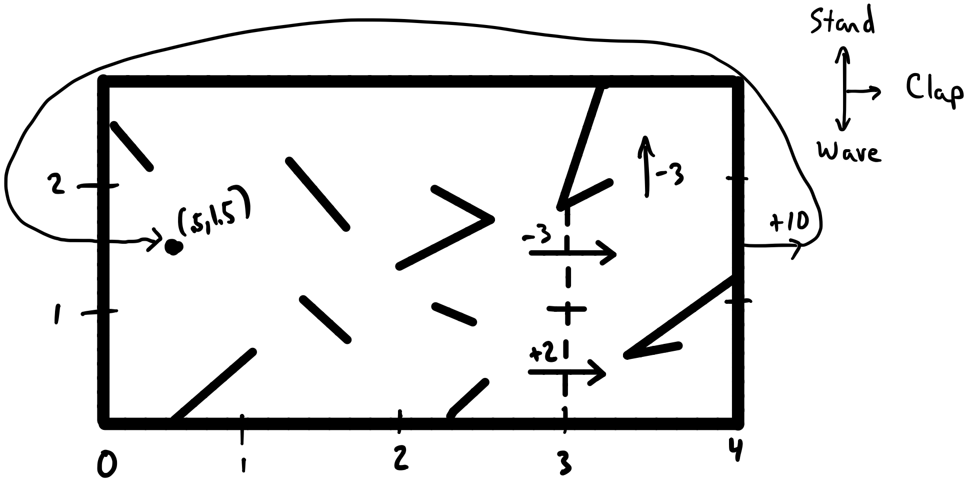
Two meaning of γ:

1) interest / inflation

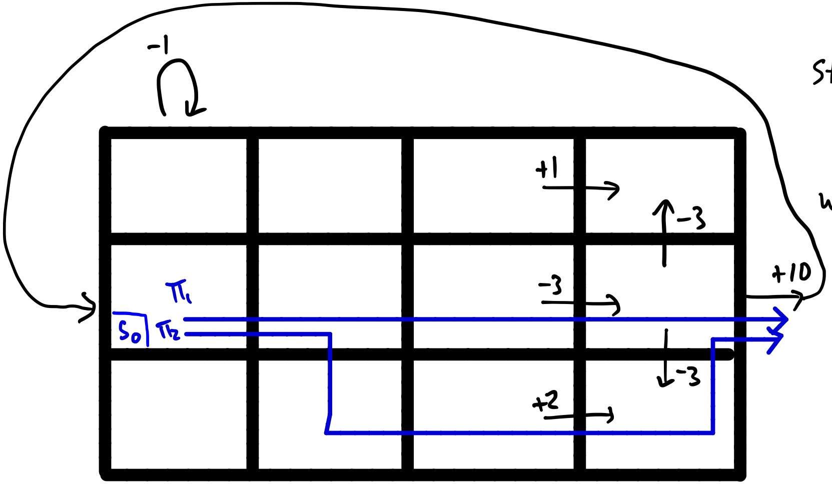
2) probability of episode ending $(1-\gamma)$

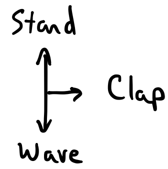
Continuing tasks





Continuing tasks

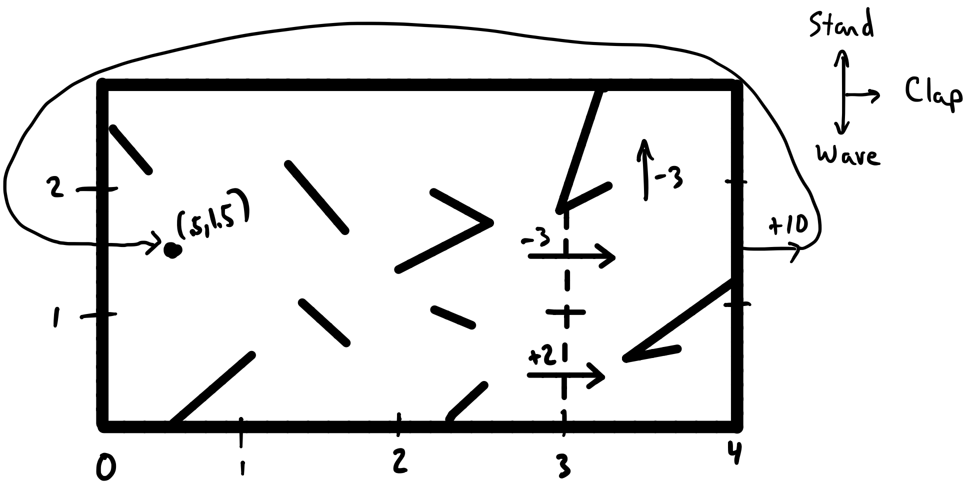


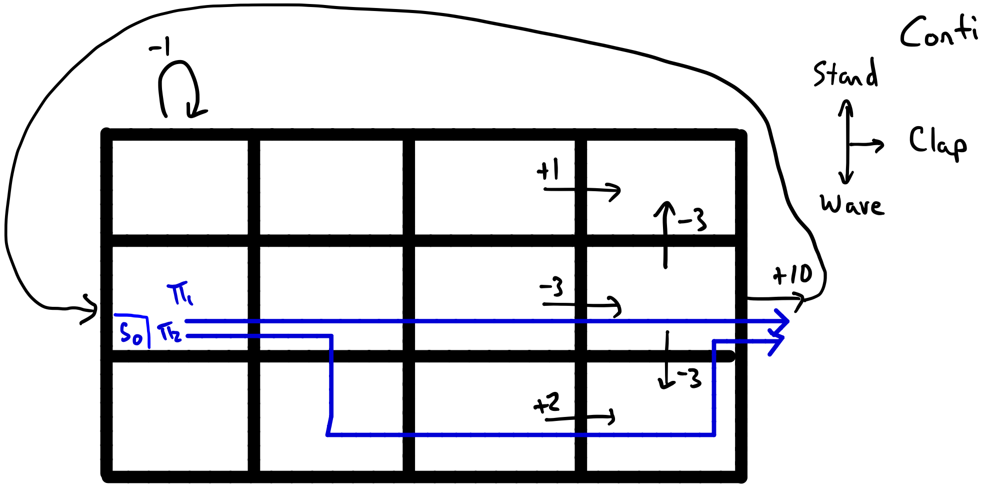


Discounting : γ

$V_{\pi_1}(s_0) =$

$V_{\pi_2}(s_0) =$





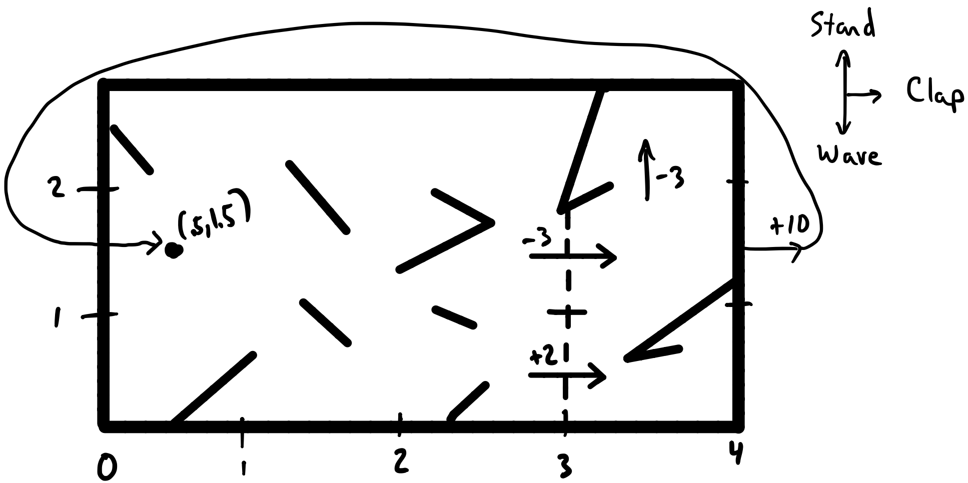
Continuing tasks

Discounting : γ

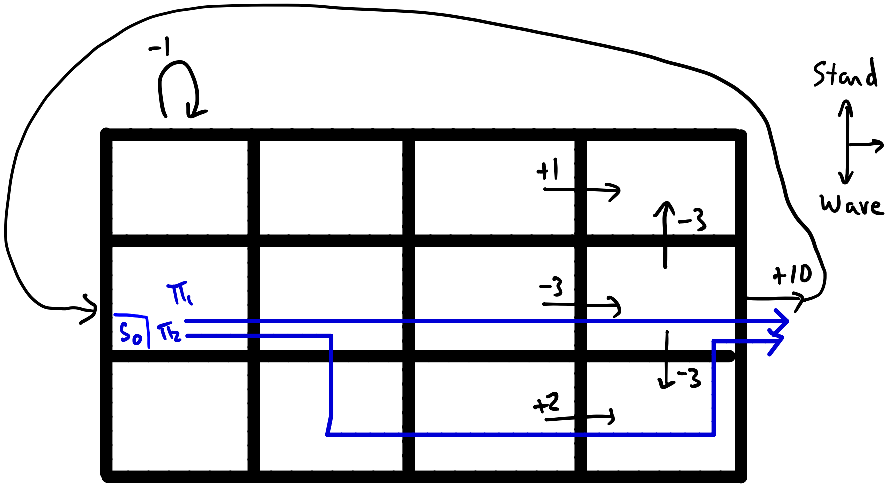
$$V_{\pi_1}(s_0) = 0 + 0 - 3\gamma^2 + 10\gamma^3 + 0 + 0 - 3\gamma^6 + 10\gamma^7 + 0 + 0 - 3\gamma^{10} + 10\gamma^{11} + \cdots$$

$$V_{\pi_2}(s_0) = 0 + 0 + 0 + 2\gamma^3 + 0 + 10\gamma^5 + 0 + 0 + 0 + 2\gamma^9 + 0 + 10\gamma^{11} + \cdots$$

which policy is better?



Continuing tasks



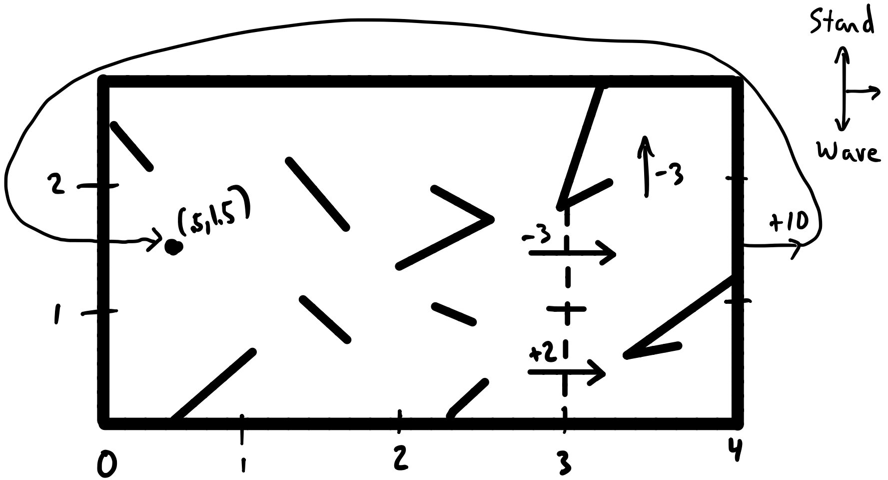
Discounting : γ

$$V_{\pi_1}(s_0) = 0 + 0 - 3\gamma^2 + 10\gamma^3 + 0 + 0 - 3\gamma^6 + 10\gamma^7 + 0 + 0 - 3\gamma^{10} + 10\gamma^{11} + \dots$$

$$V_{\pi_2}(s_0) = 0 + 0 + 0 + 2\gamma^3 + 0 + 10\gamma^5 + 0 + 0 + 0 + 2\gamma^9 + 0 + 10\gamma^{11} + \dots$$

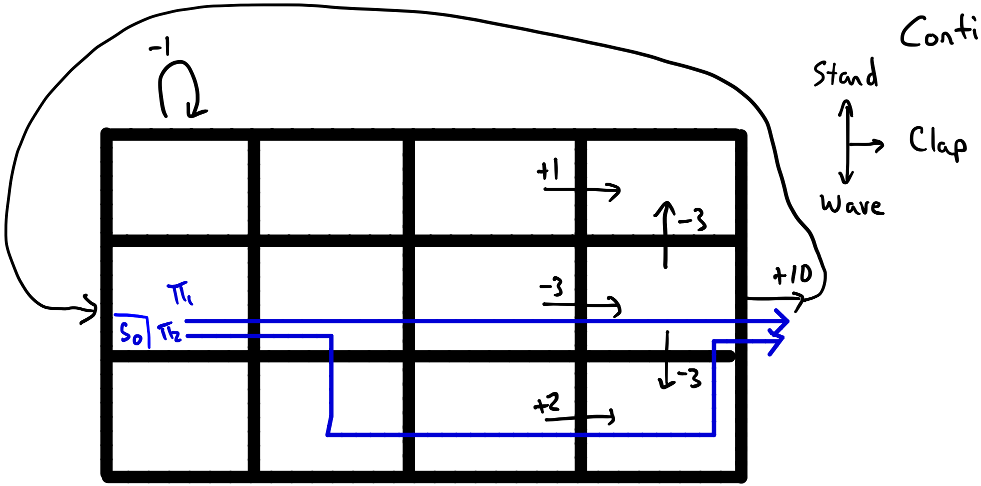
which policy is better?

Discrete state (tabular): Depends on γ



Continuous state (function approx.): Might not depend on γ!

Continuing tasks

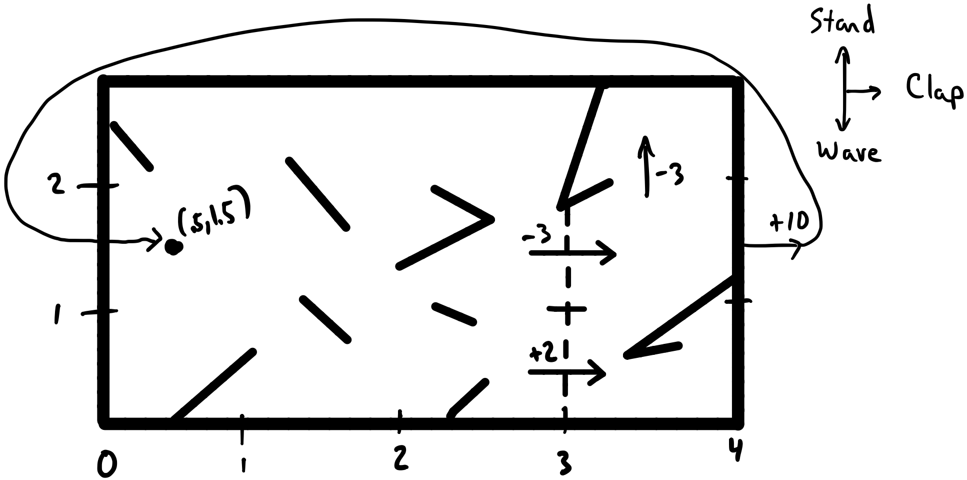


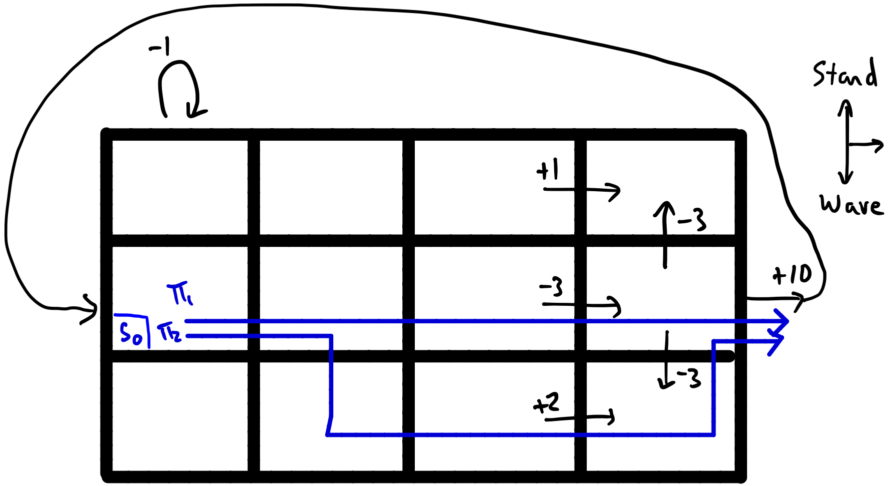
Average reward RL

$r(\pi_1) =$

$r(\pi_2) =$

which policy is better?





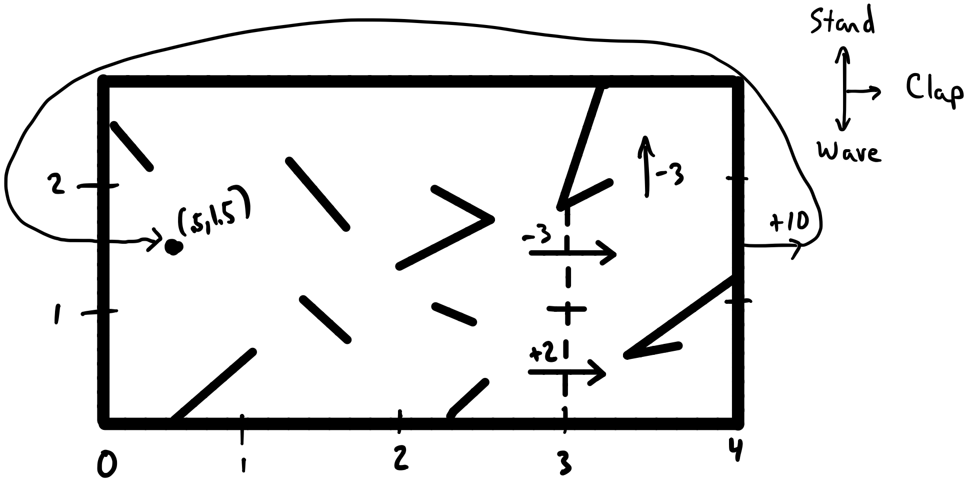
Continuing tasks

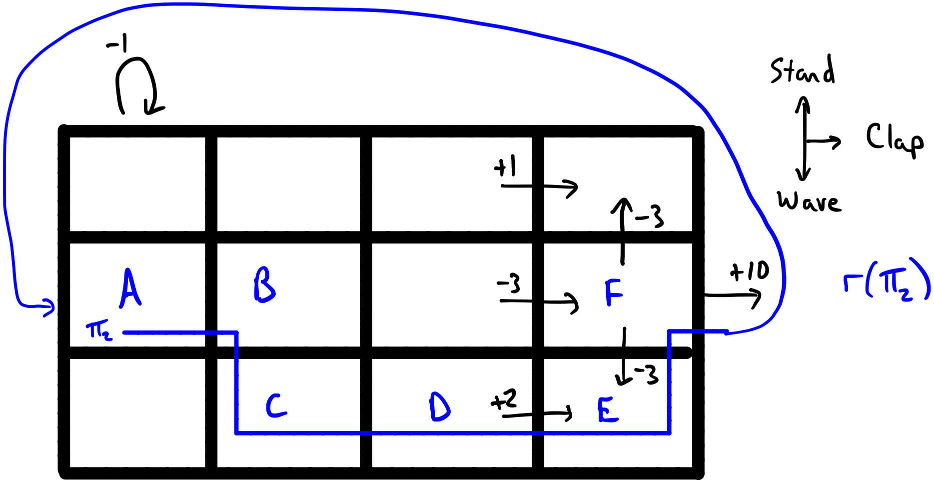
Average reward RL

$r(\pi_1) = 7/4$

$r(\pi_2) = 12/6 = 2$

which policy is better?





Differential value function

$r(\pi_2) = 2$ ← estimated by algorithm in the book: β

Differential semi-gradient SARSA (R-learning)

$V(A) =$

$V(B) =$?

$V(C) = 0$

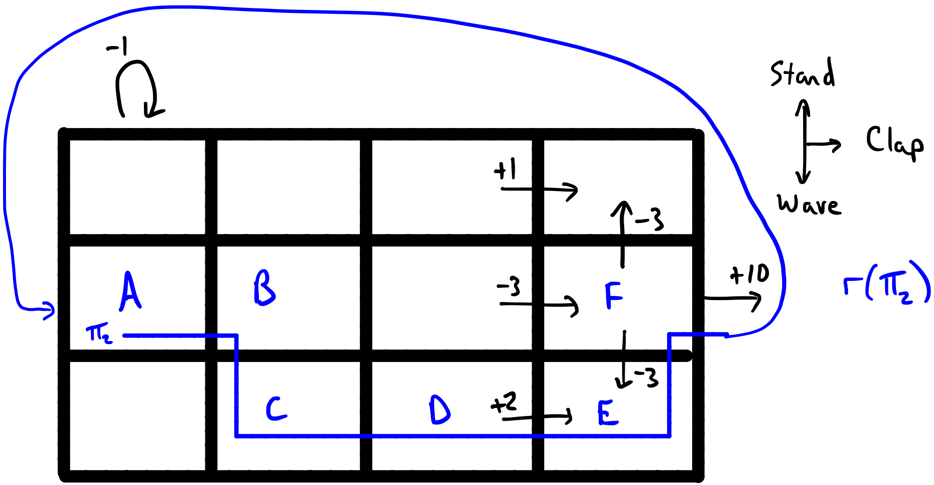
$V(D) =$

$V(E) =$

$V(F) =$

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(s',r|s,a)\left[r - r(\pi) + V_\pi(s')\right]$$

Differential value function



$r(\pi_2) = 2$

$V(A) = ?$

$V(B) = -2$

$V(C) = 0$

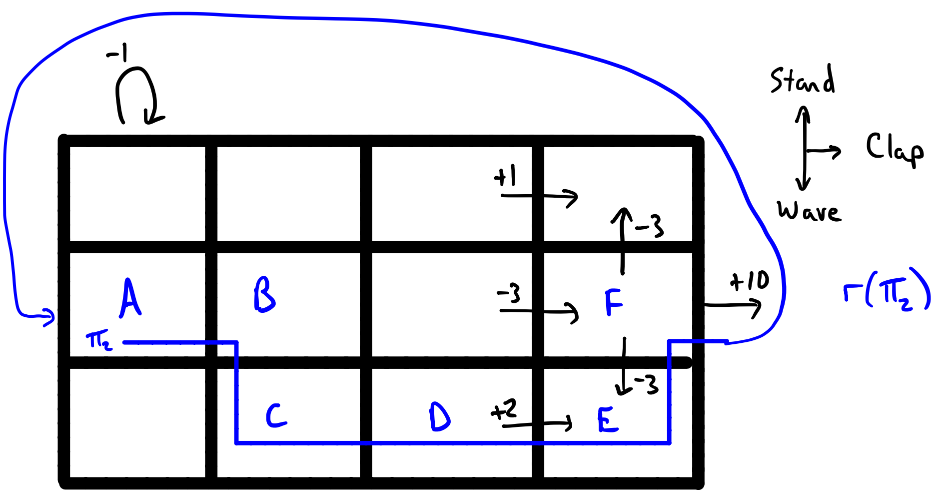
$V(D) = ?$

$V(E) = ?$

$V(F) = ?$

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(s',r|s,a)\left[r - r(\pi) + V_\pi(s')\right]$$

Differential value function



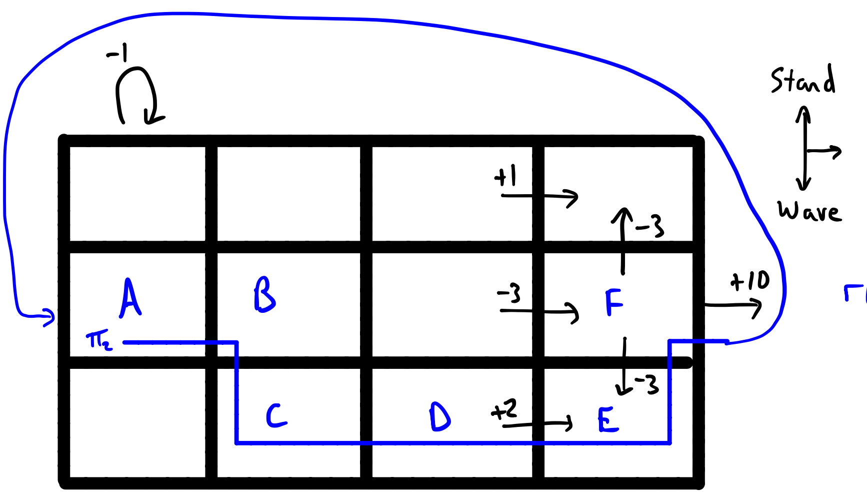
$r(\pi_2) = 2$

$V(A) = -4$
$V(B) = -2$
$V(C) = 0$
$V(D) = 2$
$V(E) = 2$
$V(F) = 4$

Can this be V?

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(s', r | s, a)\left[r - r(\pi) + V_\pi(s')\right]$$

Differential value function



$r(\pi_2) = 2$

$V(A) = -4$
$V(B) = -2$
$V(C) = 0$
$V(D) = 2$
$V(E) = 2$
$V(F) = 4$

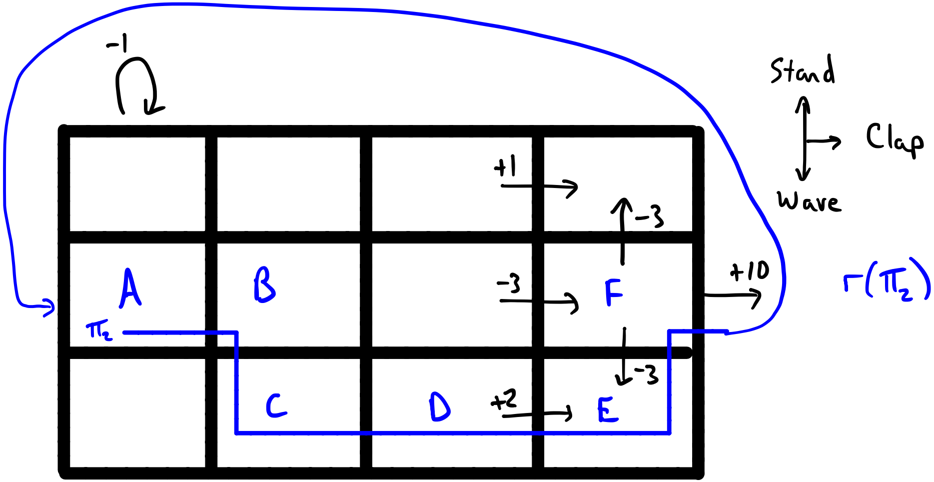
Can this be V?

$V(A) + V(B) + \dots V(F) = 2$

But avg. value of a cycle must be 0....

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(s', r | s, a) \left[r - r(\pi) + V_\pi(s') \right]$$

Differential value function



$r(\pi_2) = 2$

What's the steady state distribution of π_2?

$V(A) = -4 - 1/3 = -13/3$
$V(B) = -2 - 1/3 = -7/3$
$V(C) = 0 - 1/3 = -1/3$
$V(D) = 2 - 1/3 = 5/3$
$V(E) = 2 - 1/3 = 5/3$
$V(F) = 4 - 1/3 = 11/3$

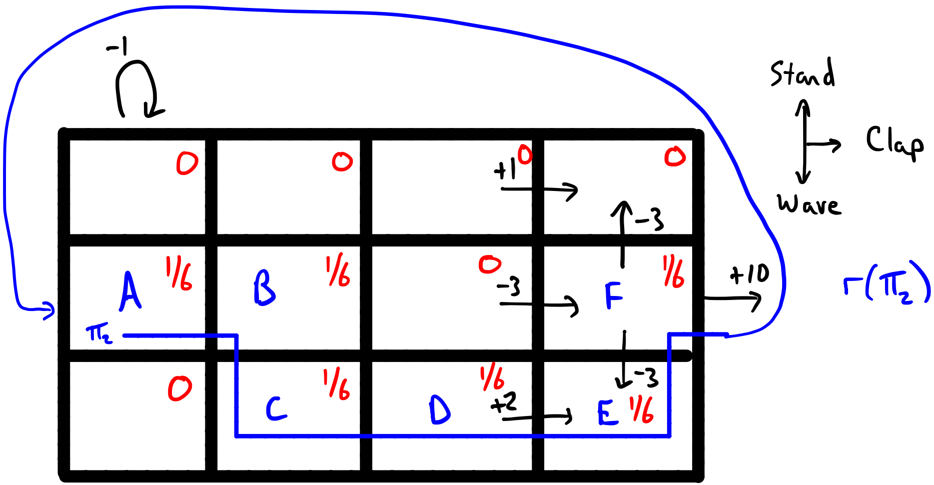
Can this be V?

$V(A) + V(B) + \dots V(F) = 2$

But avg. value of a cycle must be 0....

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(s', r | s, a)\left[r - r(\pi) + V_\pi(s')\right]$$

Differential value function



$r(\pi_2) = 2$

$V(A) = -4 - 1/3 = -13/3$
$V(B) = -2 - 1/3 = -7/3$
$V(C) = 0 - 1/3 = -1/3$
$V(D) = 2 - 1/3 = 5/3$
$V(E) = 2 - 1/3 = 5/3$
$V(F) = 4 - 1/3 = 11/3$

Can this be V?

$V(A) + V(B) + \dots V(F) = 2$

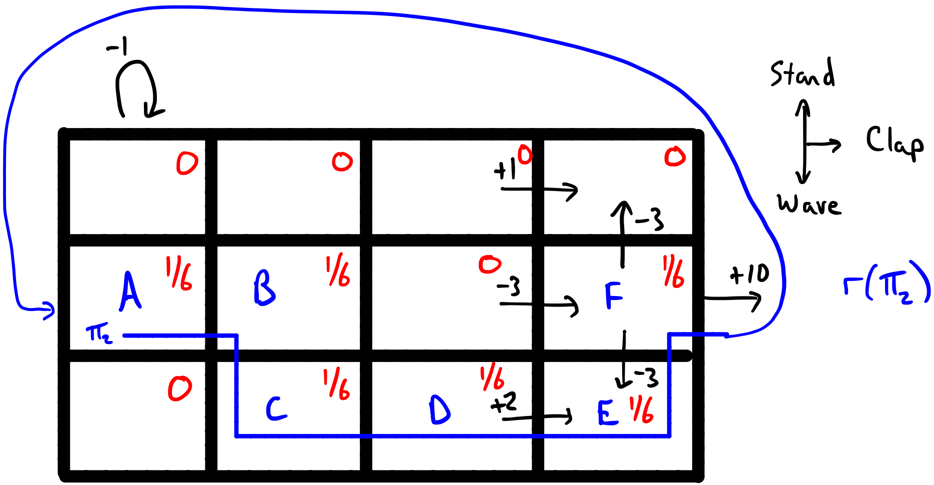
But avg. value of a cycle must be 0....

What's the steady state distribution of π_2?

Is this MDP ergodic?
(what does ergodic mean?)

$$V_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(s', r | s, a)\left[r - r(\pi) + V_\pi(s')\right]$$

Differential value function



$r(\pi_2) = 2$

What's the steady state distribution of π_2?

Is this MDP ergodic?
(i.e. does every policy have a steady state distribution independent of S_0?)

$v(A) = -4 - 1/3 = -13/3$
$v(B) = -2 - 1/3 = -7/3$
$v(C) = 0 - 1/3 = -1/3$
$v(D) = 2 - 1/3 = 5/3$
$v(E) = 2 - 1/3 = 5/3$
$v(F) = 4 - 1/3 = 11/3$

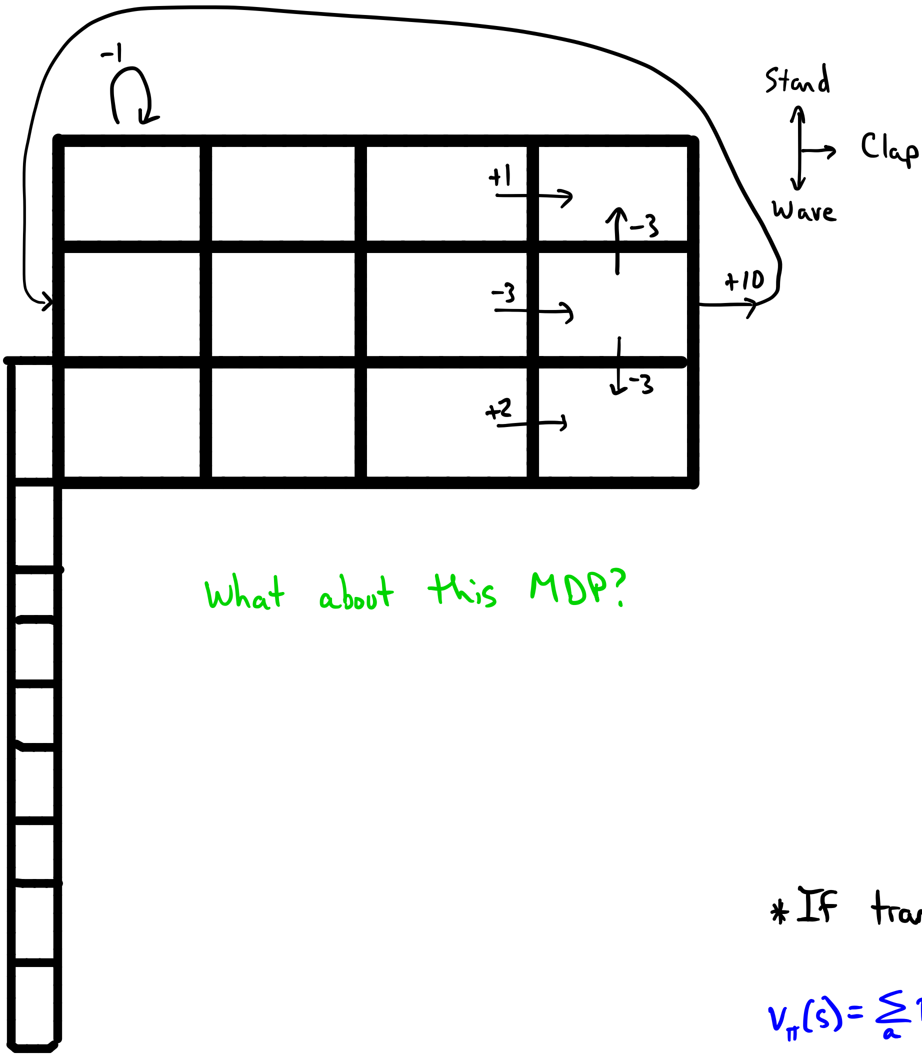
Can this be v?

$v(A) + v(B) + \dots v(F) = 2$

But avg. value of a cycle must be 0....

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(s', r | s, a) \left[r - r(\pi) + v_\pi(s') \right]$$

Differential value function



What about this MDP?

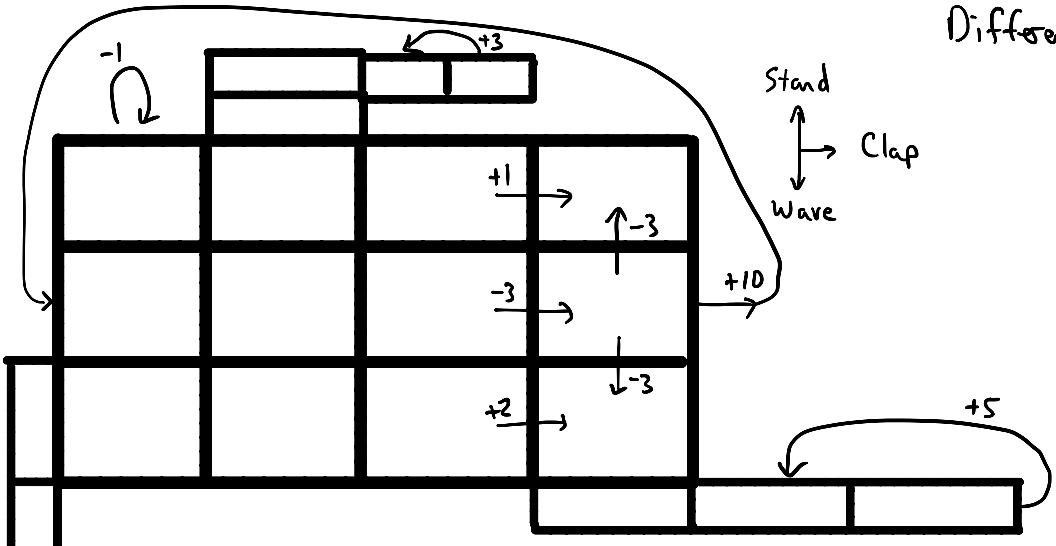
What's the steady state distribution of π_2?

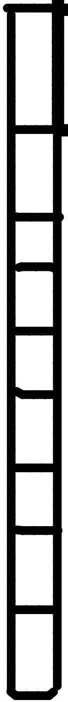
Is this MDP ergodic? YES*

*If transitions are at least slightly stochastic ($\updownarrow\rightarrow$)

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(s',r|s,a)\left[r - r(\pi) + v_\pi(s')\right]$$

Differential value function





What's the steady state distribution of π_2?

Is this MDP ergodic? YES

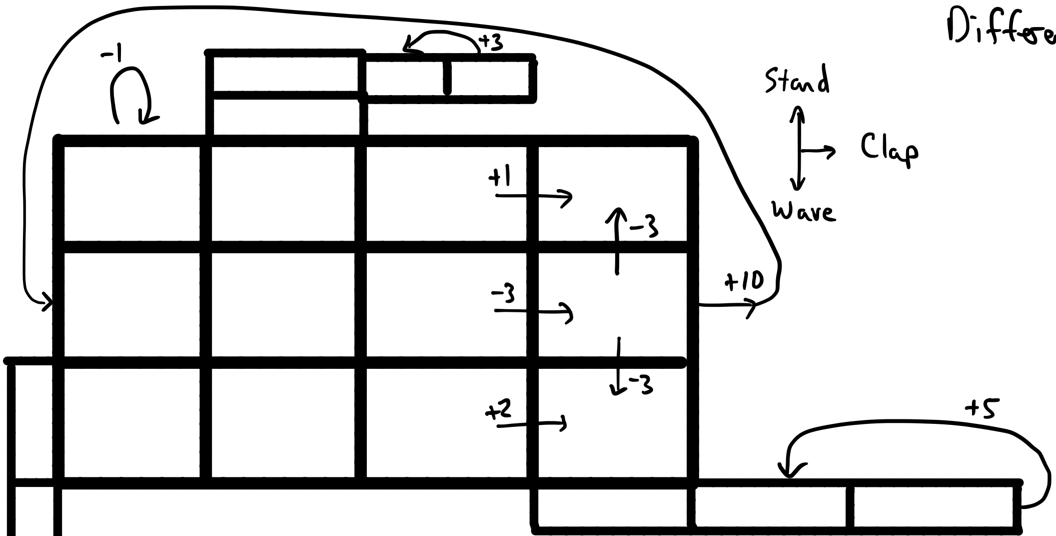
What about this MDP? YES*

What about this MDP?

* If transitions are at least slightly stochastic (↕→)

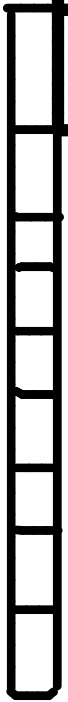
$$v_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(s', r | s, a) \left[r - r(\pi) + v_\pi(s') \right]$$

Differential value function



What's the steady state distribution of π_2?

Is this MDP ergodic? YES



What about this MDP? YES*

What about this MDP? NO

There are different stationary distributions for the same policy

A transition can happen that affects where the agent can get (eventually)

* If transitions are at least slightly stochastic()

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{r,s'} p(s',r|s,a)\left[r - r(\pi) + v_\pi(s')\right]$$