REINFORCEMENT LEARNING: THEORY AND PRACTICE

Ch. 10: On-policy Control with Approximation

Profs. Amy Zhang and Peter Stone

Previously

Chapter 9 On-policy Prediction with Approximation

Focus on value estimation with various function approximation methods

Semi-gradient Control

Gradient-descent update for action-value prediction

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \Big[U_t - \hat{q}(S_t, A_t, \mathbf{w}_t) \Big] \nabla \hat{q}(S_t, A_t, \mathbf{w}_t)$$

Semi-gradient Control

Gradient-descent update for action-value prediction

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \Big[U_t - \hat{q}(S_t, A_t, \mathbf{w}_t) \Big] \nabla \hat{q}(S_t, A_t, \mathbf{w}_t)$$

One-step Sarsa:

$$\mathbf{w}_{t+1} \doteq \mathbf{w}_t + \alpha \Big[R_{t+1} + \gamma \hat{q}(S_{t+1}, A_{t+1}, \mathbf{w}_t) - \hat{q}(S_t, A_t, \mathbf{w}_t) \Big] \nabla \hat{q}(S_t, A_t, \mathbf{w}_t)$$

N-step semi-gradient Sarsa

N-step return in function approximation form:

N-step semi-gradient Sarsa

N-step return in function approximation form:

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n \hat{q}(S_{t+n}, A_{t+n}, \mathbf{w}_{t+n-1})$$

N-step update equation:

N-step semi-gradient Sarsa

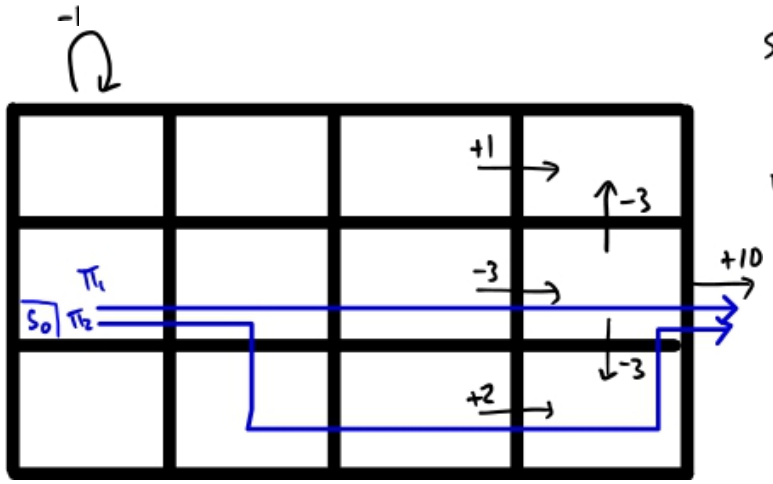
N-step return in function approximation form:

$$G_{t:t+n} \doteq R_{t+1} + \gamma R_{t+2} + \cdots + \gamma^{n-1} R_{t+n} + \gamma^n \hat{q}(S_{t+n}, A_{t+n}, \mathbf{w}_{t+n-1})$$

N-step update equation:

$$\mathbf{w}_{t+n} \doteq \mathbf{w}_{t+n-1} + \alpha \left[G_{t:t+n} - \hat{q}(S_t, A_t, \mathbf{w}_{t+n-1}) \right] \nabla \hat{q}(S_t, A_t, \mathbf{w}_{t+n-1})$$

Episodic tasks

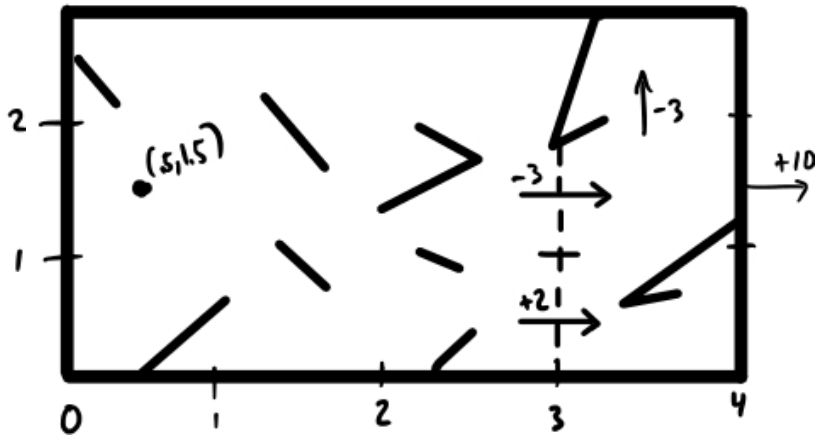


Stand
Clap
Wave

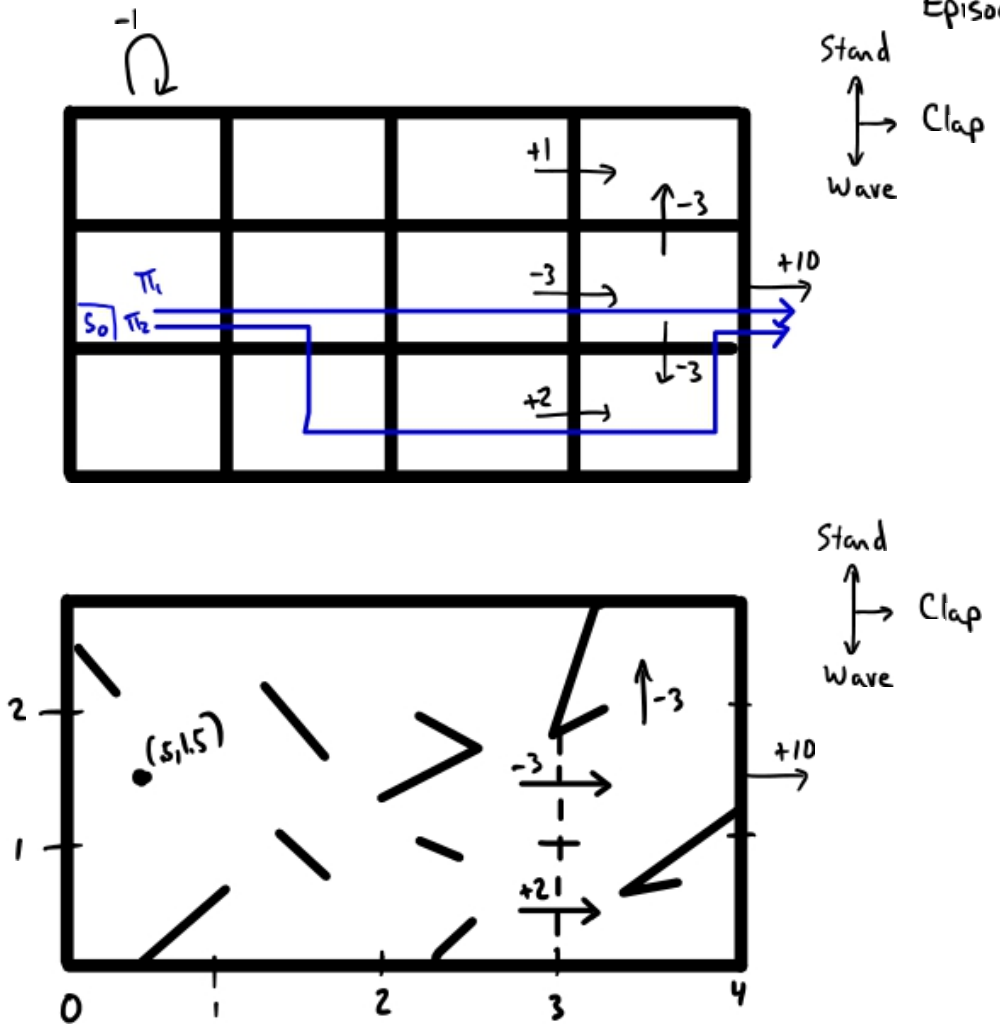
Discounting : γ

$V_{\pi_1}(s_0) =$

$V_{\pi_2}(s_0) =$



Peter's slide



Episodic tasks

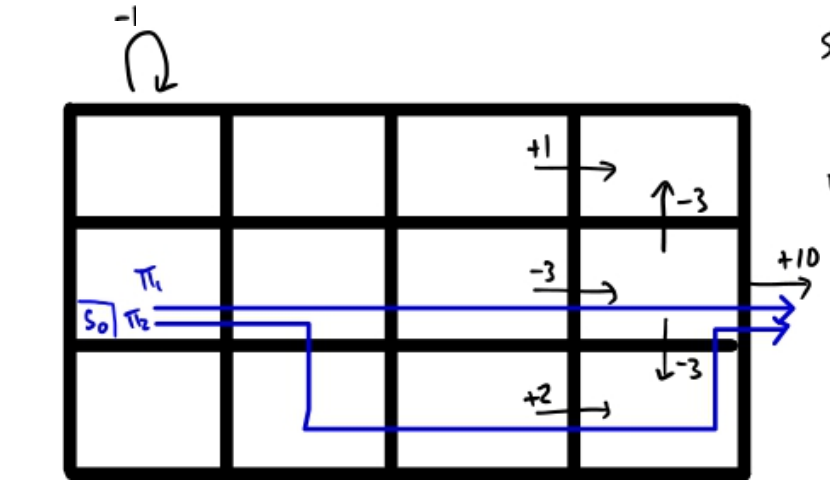
Discounting : γ

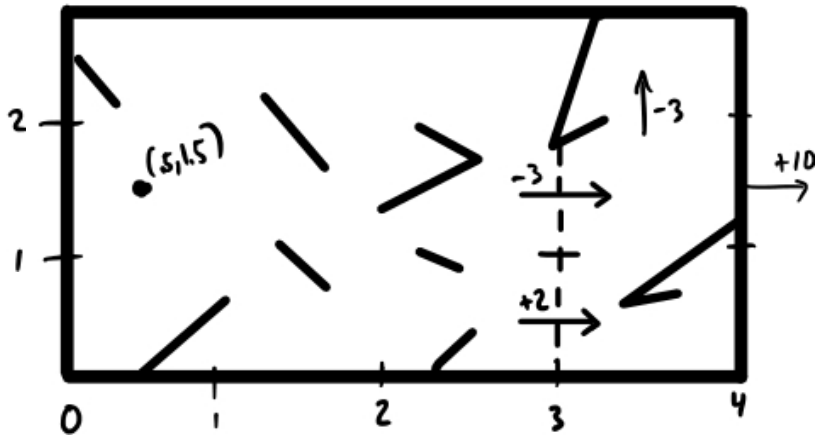
$V_{\pi_1}(s_0) = 0 + 0 - 3\gamma^2 + 10\gamma^3$

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

Which policy is better?

Peter's slide





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?

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

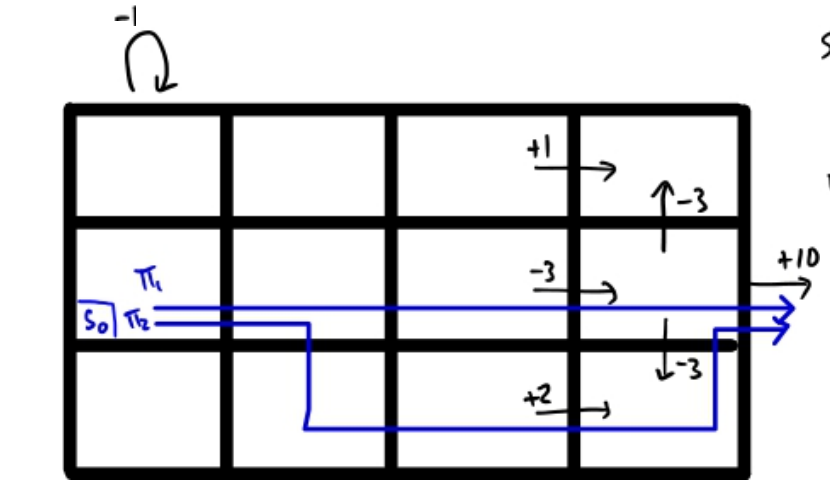
$\gamma = .5$: $V_{\pi_1}(s_0) = 1.175$ $V_{\pi_2}(s_0) = .5625$

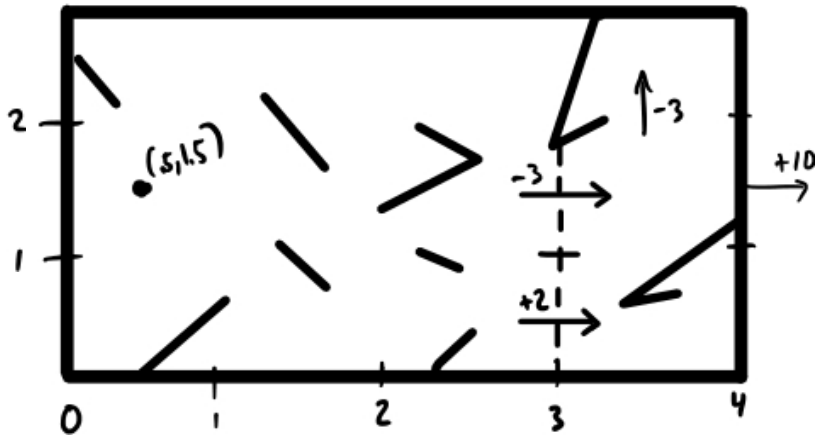
Two meaning of γ:

1)

2)

Peter's slide





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?

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

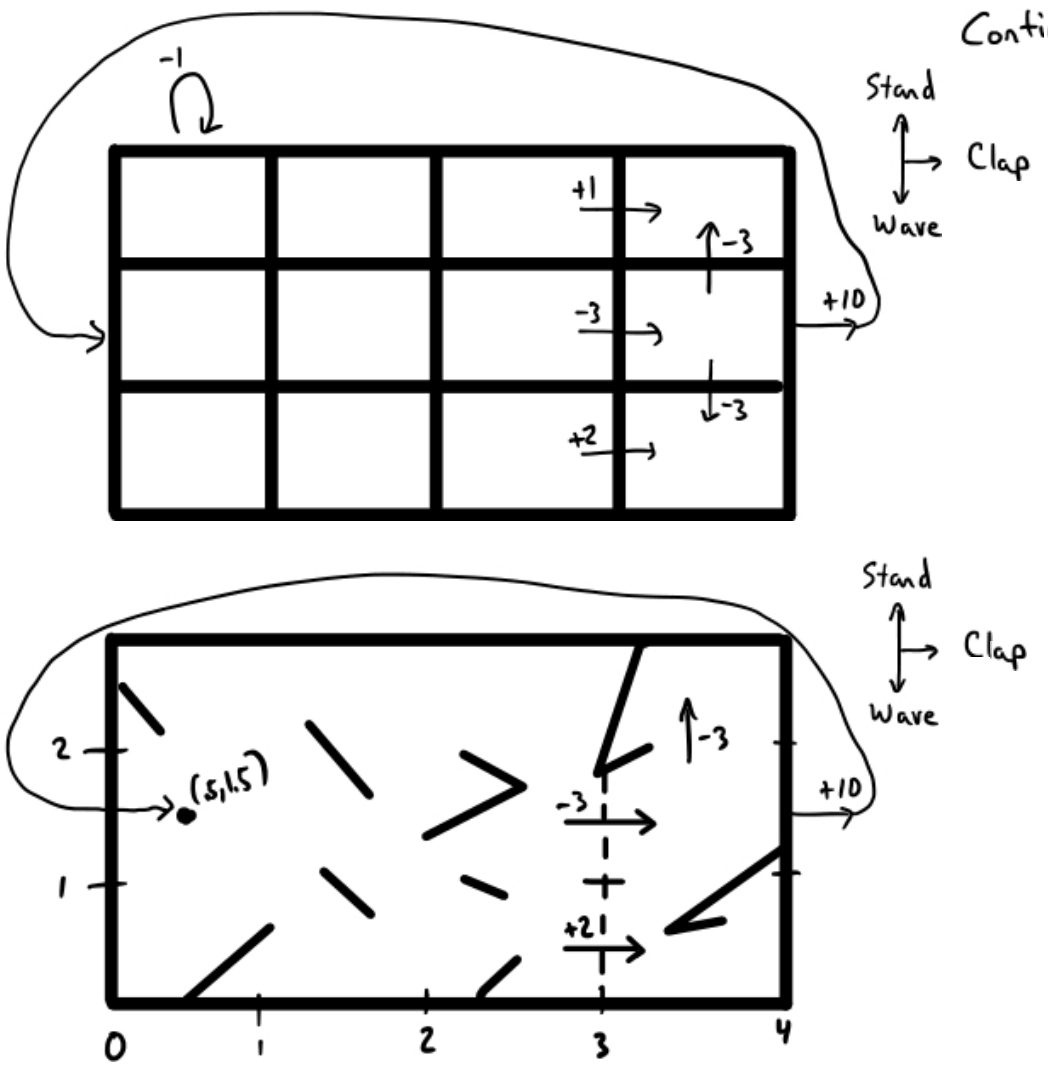
$\gamma = .5$: $V_{\pi_1}(s_0) = 1.175$ $V_{\pi_2}(s_0) = .5625$

Two meaning of γ:

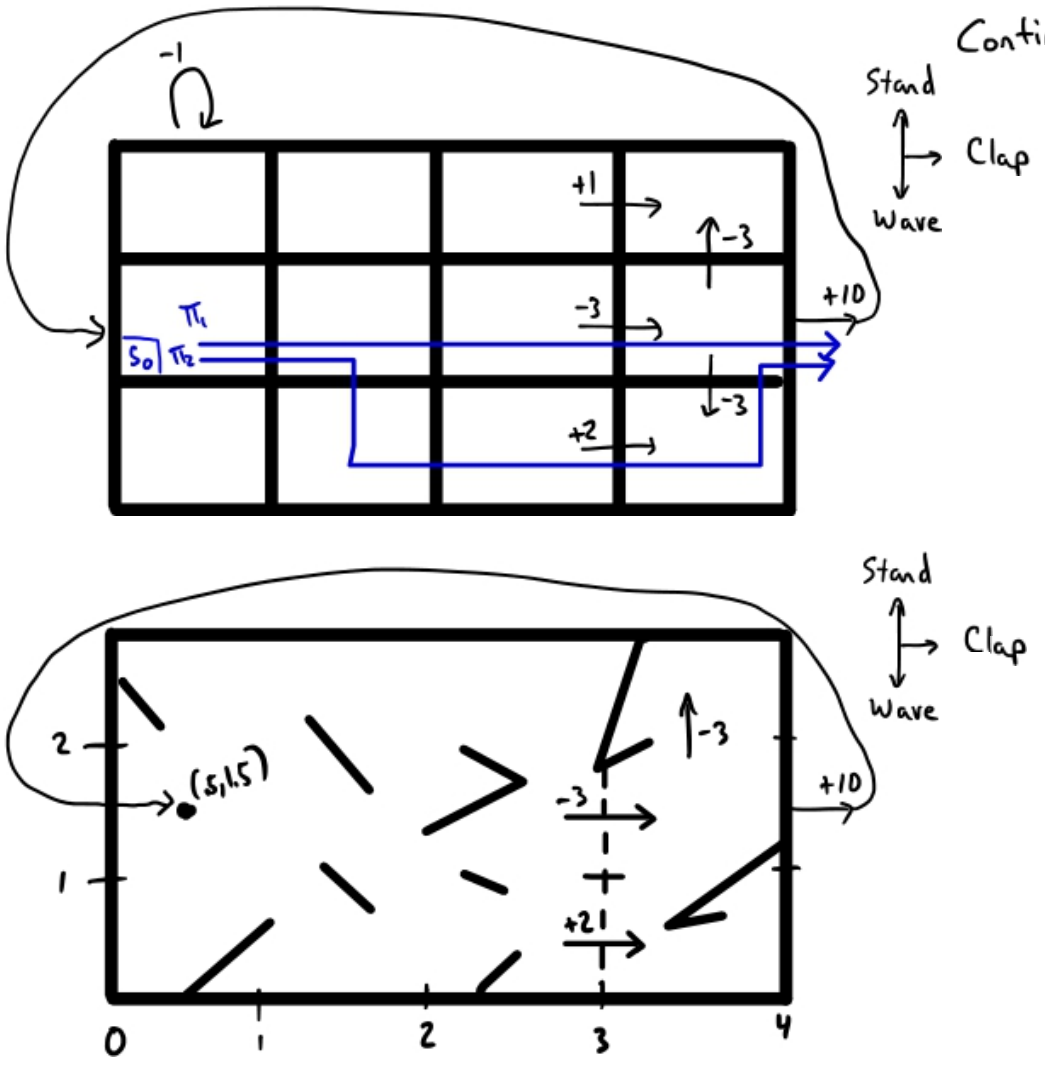
1) interest / inflation

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

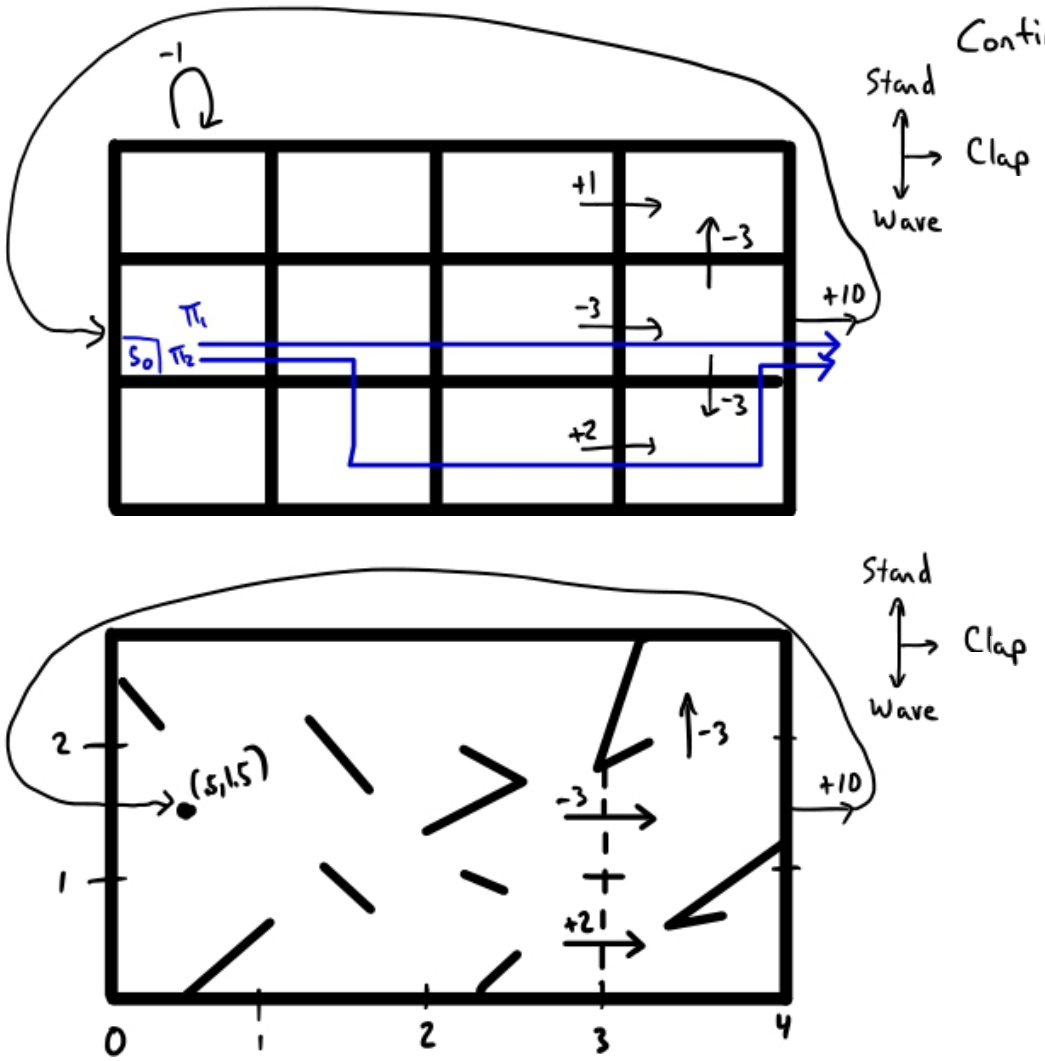
Peter's slide



Peter's slide



Peter's slide



Continuing tasks

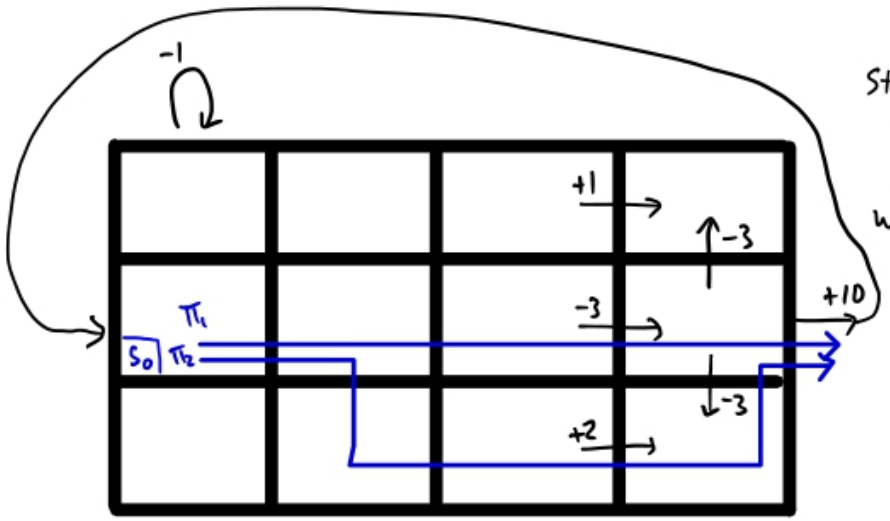
Discounting : γ

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

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

which policy is better?

Peter's slide



Stand

Clap

Wave

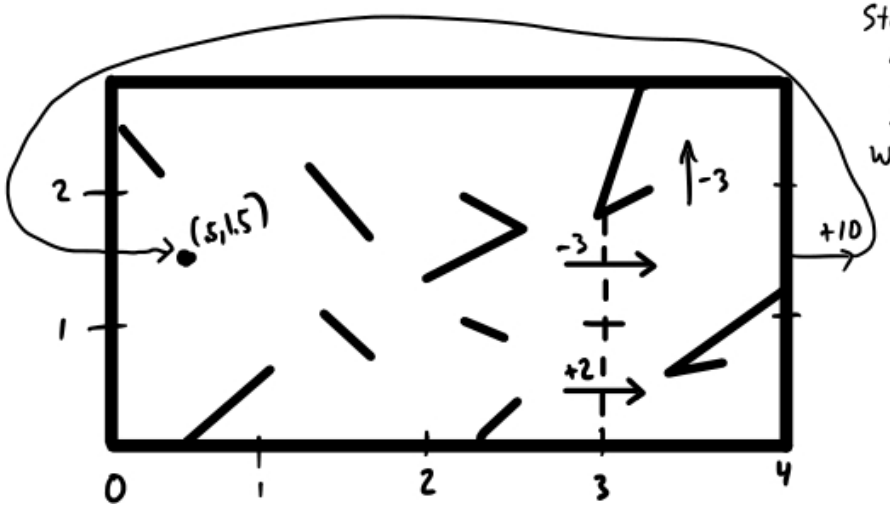
Continuing tasks

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?



Stand

Clap

Wave

Discrete state (tabular):
Depends on γ

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

Peter's slide

Average Reward

For the continuing problem setting (alternative to episodic and discounted settings)

$$
\begin{aligned}
r(\pi) &\doteq \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}[R_t \mid A_{0:t-1} \sim \pi] \\
&= \lim_{t\to\infty} \mathbb{E}[R_t \mid A_{0:t-1} \sim \pi] \,, \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a|s) \sum_{s',r} p(s', r \,|\, s, a) r
\end{aligned}
$$

Average Reward

For the continuing problem setting (alternative to episodic and discounted settings)

$$
\begin{aligned}
r(\pi) &\doteq \lim_{h\to\infty} \frac{1}{h} \sum_{t=1}^{h} \mathbb{E}[R_t \mid A_{0:t-1} \sim \pi] \\
&= \lim_{t\to\infty} \mathbb{E}[R_t \mid A_{0:t-1} \sim \pi] \,, \\
&= \sum_s \mu_\pi(s) \sum_a \pi(a|s) \sum_{s',r} p(s', r \,|\, s, a) r
\end{aligned}
$$

Steady state assumption: $\sum_s \mu_\pi(s) \sum_a \pi(a|s) p(s' \,|\, s, a) = \mu_\pi(s').$

Reading Responses

[Linus Zhang]
"In the average-reward setting, returns are defined in terms of differences between rewards and the average reward" - isn't it strange to have the return depend on the current policy? Doesn't this mean that a state's new value is somehow dependent on how the algorithm has gotten to its current value? Perhaps I am understanding this wrong. The book also talks a lot about "ordering"; does this describe a generalization of the meaning of the value function to represent something specific to a run of the algorithm, instead of an objective value that can be compared across various runs and algorithms?

Reading Responses

[Linus Zhang]
"In the average-reward setting, returns are defined in terms of differences between rewards and the average reward" - isn't it strange to have the return depend on the current policy? Doesn't this mean that a state's new value is somehow dependent on how the algorithm has gotten to its current value? Perhaps I am understanding this wrong. The book also talks a lot about "ordering"; does this describe a generalization of the meaning of the value function to represent something specific to a run of the algorithm, instead of an objective value that can be compared across various runs and algorithms?
Return always depends on a policy!
The ordering refers to the ordering of what policy is better than another according to the expected return.

Differential Return

In the average-reward setting, returns are defined in terms of differences between rewards and the average reward:

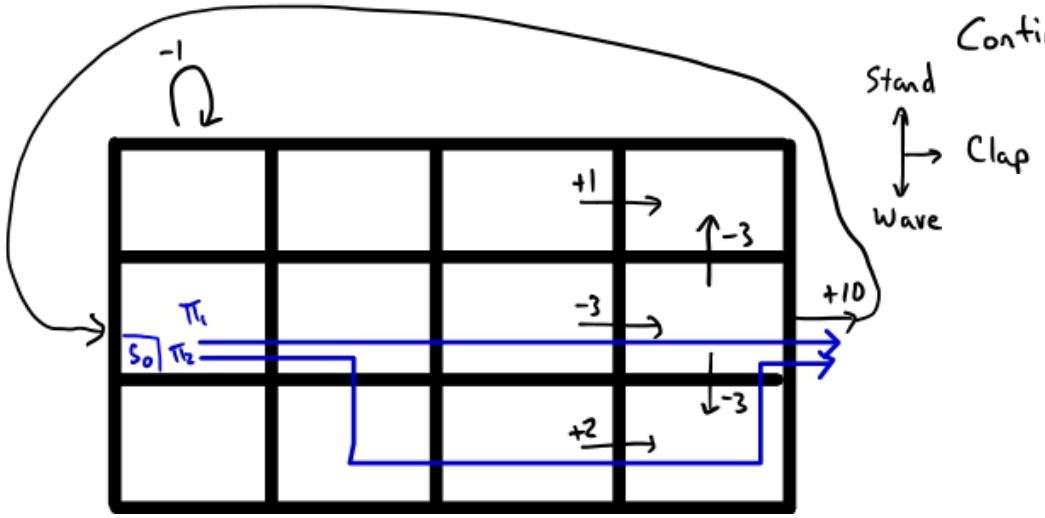
$$G_t \doteq R_{t+1} - r(\pi) + R_{t+2} - r(\pi) + R_{t+3} - r(\pi) + \cdots .$$

Differential form of TD errors:

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \hat{v}(S_{t+1}, \mathbf{w}_t) - \hat{v}(S_t, \mathbf{w}_t),$$

and

$$\delta_t \doteq R_{t+1} - \bar{R}_t + \hat{q}(S_{t+1}, A_{t+1}, \mathbf{w}_t) - \hat{q}(S_t, A_t, \mathbf{w}_t)$$



Continuing tasks

Stand

Clap

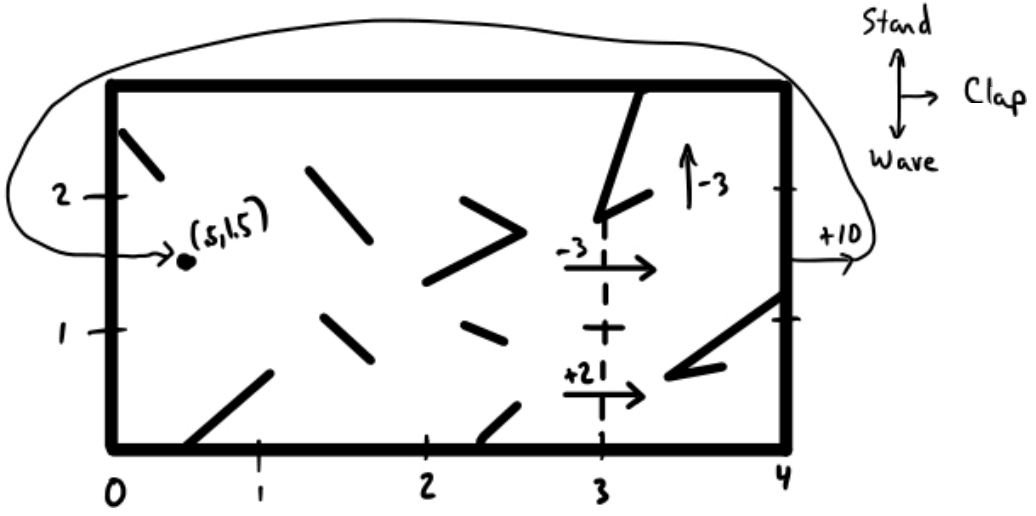
Wave

Average reward RL

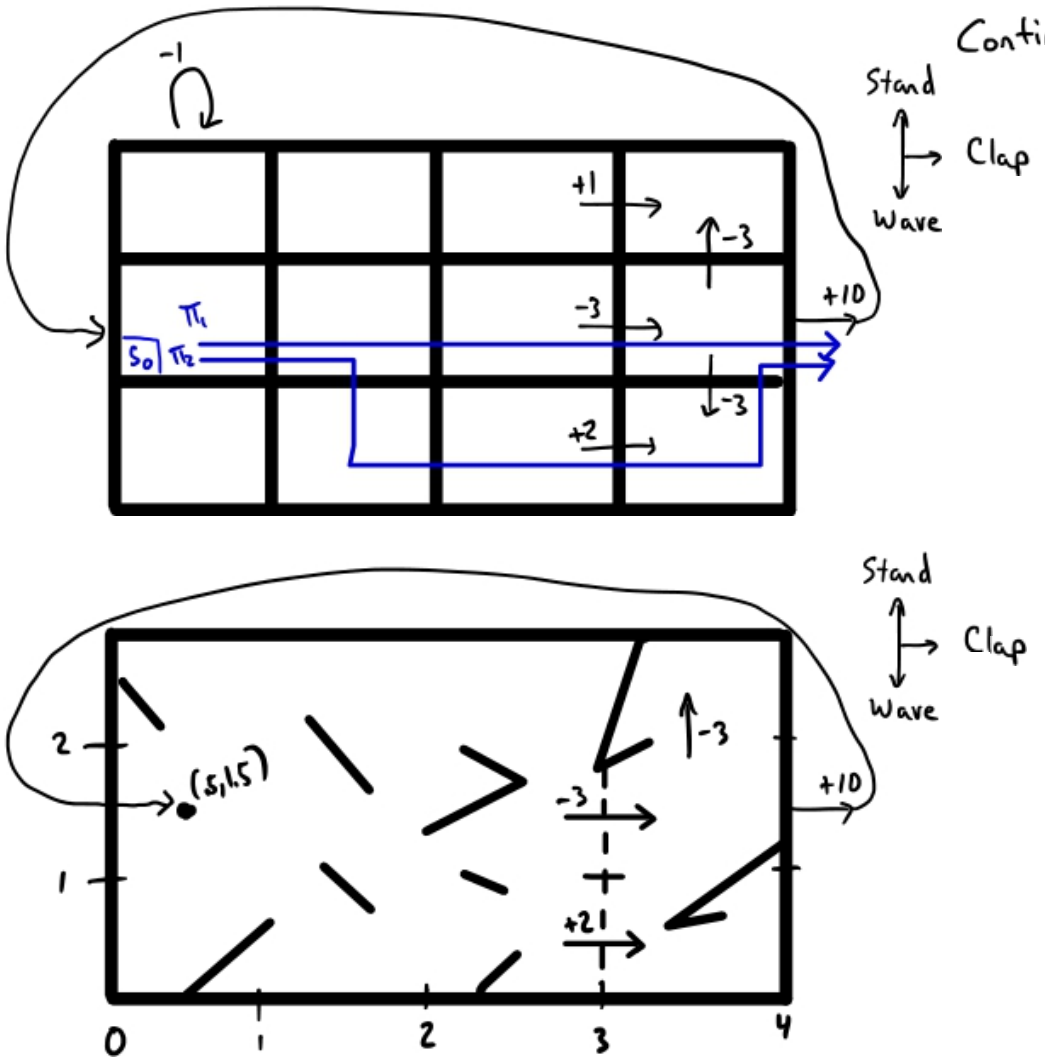
$r(\pi_1) =$

$r(\pi_2) =$

which policy is better?



Peter's slide



Peter's slide

Reading Responses

[Abhi Sridhar]
Chapter 10 introduces the concept of average reward as a new problem setting for continuing tasks. How does this setting differ from the episodic and discounted settings, and why is it relevant for function approximation?

[Garvit Mohata]
Can't we use differential returns in the earlier part where we were not using function approximation? Would we still run the risk of getting not defined values (+inf, -inf)? Then if that is true, what makes the function approximation case not have such undefined values? Does this have to do with the sequence in differential return that is being generated?

Reading Responses

[Abhi Sridhar]
Chapter 10 introduces the concept of average reward as a new problem setting for continuing tasks. How does this setting differ from the episodic and discounted settings, and why is it relevant for function approximation?
We find that discount factor does not affect the ordering of policies in the continuous state setting which requires function approximation.

[Garvit Mohata]
Can't we use differential returns in the earlier part where we were not using function approximation? Would we still run the risk of getting not defined values (+inf, -inf)? Then if that is true, what makes the function approximation case not have such undefined values? Does this have to do with the sequence in differential return that is being generated?
We can! Not defined values not an issue with *average* returns.

Reading Responses

[Krystal An]
Why is it necessary to transition from a discounting formulation to an average-reward formulation in the context of function approximation for continuing tasks?

[Emin Arslan]
I don't understand what is different between the tabular and function approximation methods that makes discounting unnecessary in the latter case and not so in the former. I understand the symmetry argument and how the discount rate would have no effect on the ordering of policies, but I don't fully understand why this only applies to the approximation methods.

Reading Responses

We find that discounting and average reward formulations are the same in continuing tasks with continuous state spaces. Average reward means we don't have to consider discount factor anymore.
Also see Discounted Reinforcement Learning is Not an Optimization Problem by Naik, Shariff, Yasui, and Sutton on resources page.

Problems with discounting in continuous settings

Discounting is equivalent to average reward

Discounting algorithms with function approximation do not optimize discounted value over the on-policy distribution, and thus are not guaranteed to optimize average reward.

With function approximation we have lost the policy improvement theorem!

Reading Responses

[Linus Zhang]
"Once we introduce function approximation we can no longer guarantee improvement for any setting." - this means that we cannot prove theoretical bounds for any of the algorithms involving function approximation? Are there special classes of MDPs or certain function approximators that we can prove bounds for?

[Ahmet Aydin]
In Section 10.4, it is mentioned that the policy improvement theorem is lost with function approximation. Could you elaborate on this further? Is this due to the approximation and working in a smaller space?

Reading Responses

[Linus Zhang]
"Once we introduce function approximation we can no longer guarantee improvement for any setting." - this means that we cannot prove theoretical bounds for any of the algorithms involving function approximation? Are there special classes of MDPs or certain function approximators that we can prove bounds for?
We will talk about policy gradient methods and the policy gradient theorem in Chapter 13.

[Ahmet Aydin]
In Section 10.4, it is mentioned that the policy improvement theorem is lost with function approximation. Could you elaborate on this further? Is this due to the approximation and working in a smaller space?
With function approximation, value predictions for other states get affected when we update the value for a target state, so we lose the policy improvement theorem.

n-step Differential Semi-gradient Sarsa

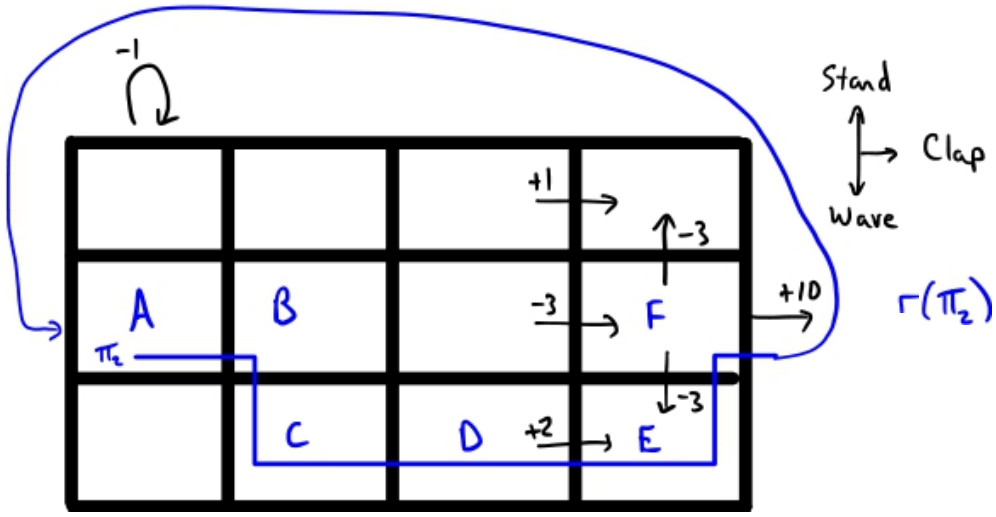
Differential form of n-step return with function approximation:

$$G_{t:t+n} \doteq R_{t+1} - \bar{R}_{t+n-1} + \cdots + R_{t+n} - \bar{R}_{t+n-1} + \hat{q}(S_{t+n}, A_{t+n}, \mathbf{w}_{t+n-1}),$$

N-step TD error:

$$\delta_t \doteq G_{t:t+n} - \hat{q}(S_t, A_t, \mathbf{w})$$

Differential value function



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

Differential semi-gradient SARSA (R-learning)

$V(A) =$

$V(B) =$?

$V(C) = 0$

$V(D) =$

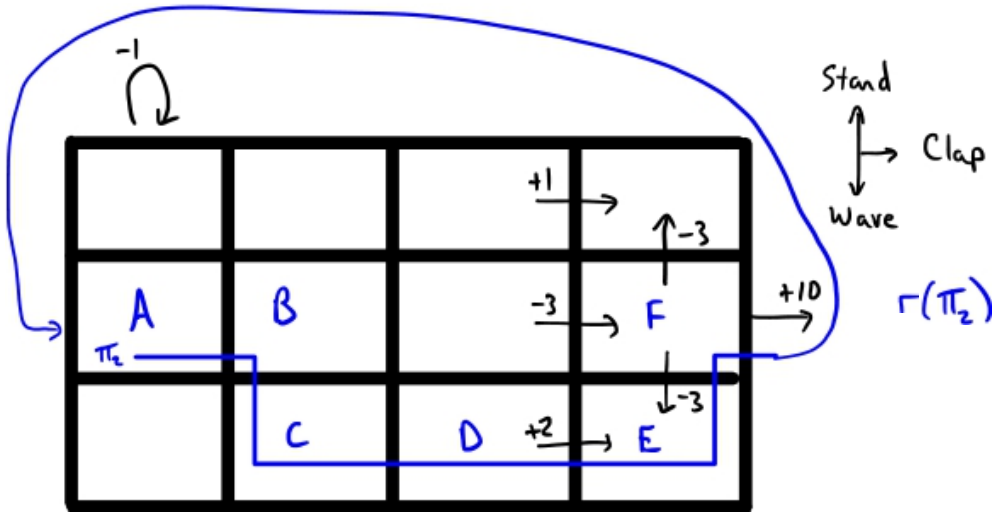
$V(E) =$

$V(F) =$

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

Peter's slide

Differential value function



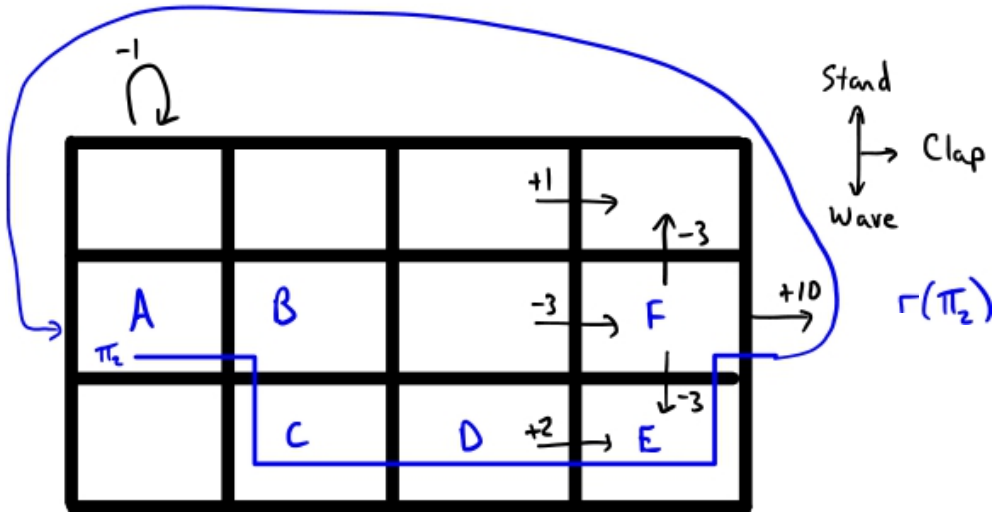
$r(\pi_2) = 2$

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

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

Peter's slide

Differential value function



$r(\pi_2) = 2$

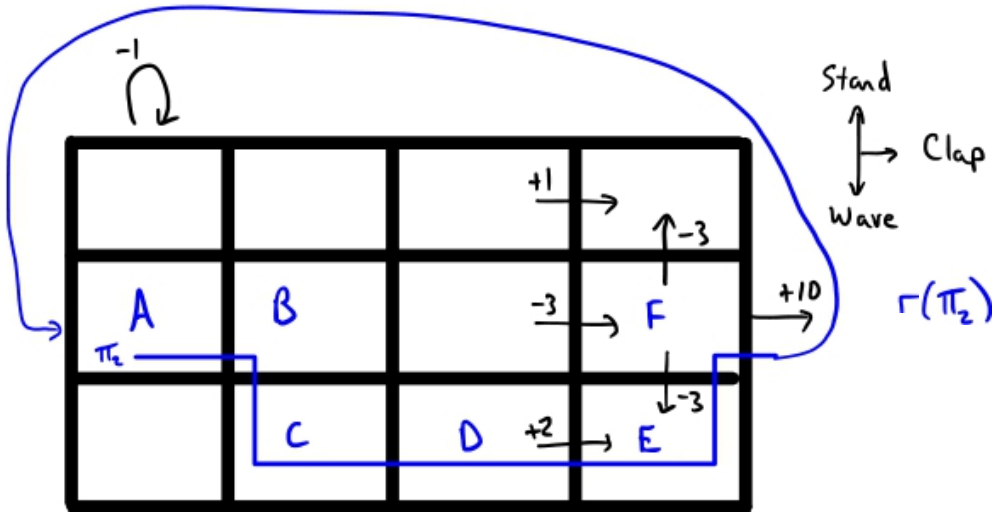
$V(A) = -4$
$V(B) = -2$
$V(C) = 0$
$V(D) = 2$
$V(E) = 2$
$V(F) = 4$

Can this be V?

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

Peter's slide

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?

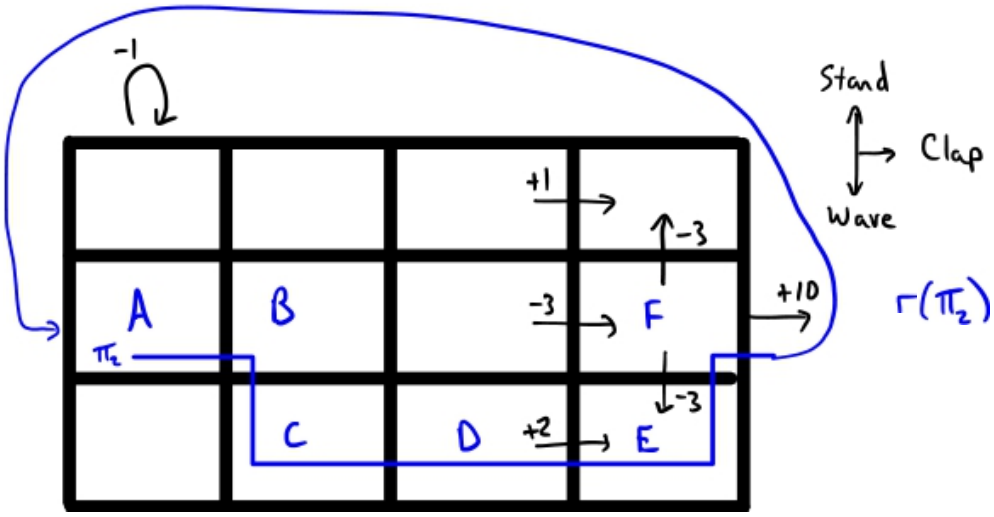
$v(A) + v(B) + \dots v(F) = 2$

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

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

Peter's slide

Differential value function



$r(\pi_2) = 2$

What's the steady state distribution of π_2?

$V(A) = -4 - 1/3 = -15/3$
$V(B) = -2 - 1/3 = -7/3$
$V(C) = 0 - 1/3 = -1/3$
$V(D) = 2 - 1/3 = 5/3$
$V(E) = 2 - 1/3 = 5/3$
$V(F) = 4 - 1/3 = 11/3$

Can this be V?

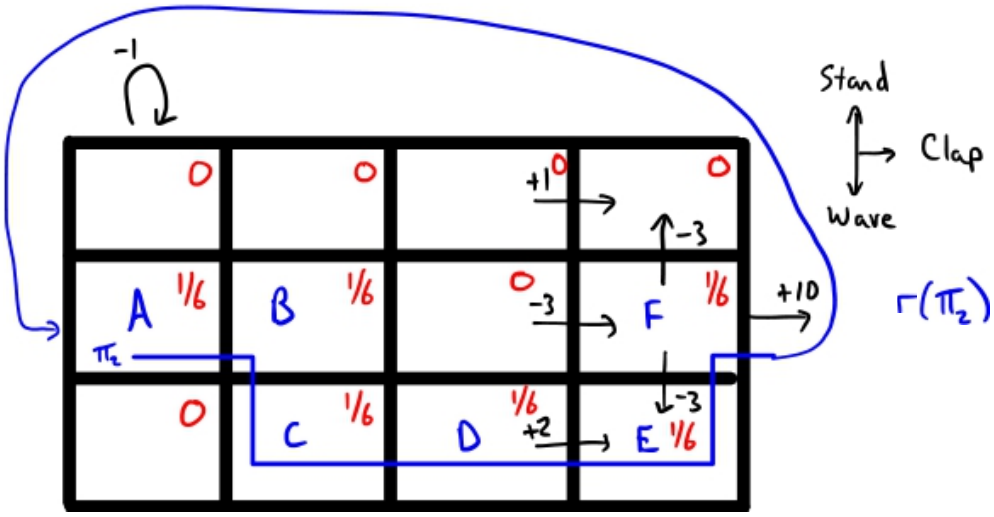
$V(A) + V(B) + \dots V(F) = 2$

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

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

Peter's slide

Differential value function



$r(\pi_2) = 2$

What's the steady state distribution of π_2?

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

$V(A) = -4 - 1/3 = -15/3$
$V(B) = -2 - 1/3 = -7/3$
$V(C) = 0 - 1/3 = -1/3$
$V(D) = 2 - 1/3 = 5/3$
$V(E) = 2 - 1/3 = 5/3$
$V(F) = 4 - 1/3 = 11/3$

Can this be V?

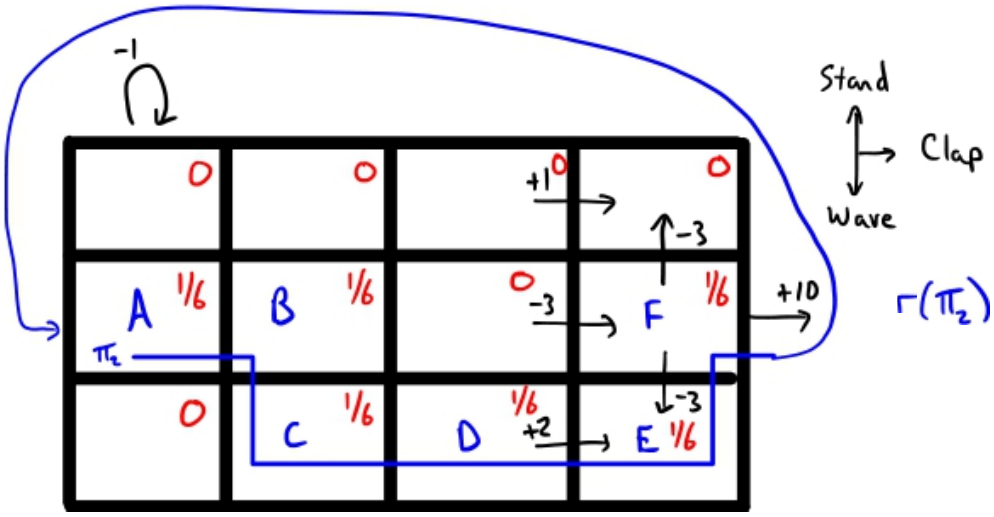
$V(A) + V(B) + \dots V(F) = 2$

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

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

Peter's slide

Differential value function



$r(\pi_2) = 2$

$V(A) = -4 - 1/3 = -15/3$
$V(B) = -2 - 1/3 = -7/3$
$V(C) = 0 - 1/3 = -1/3$
$V(D) = 2 - 1/3 = 5/3$
$V(E) = 2 - 1/3 = 5/3$
$V(F) = 4 - 1/3 = 11/3$

Can this be V?

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

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

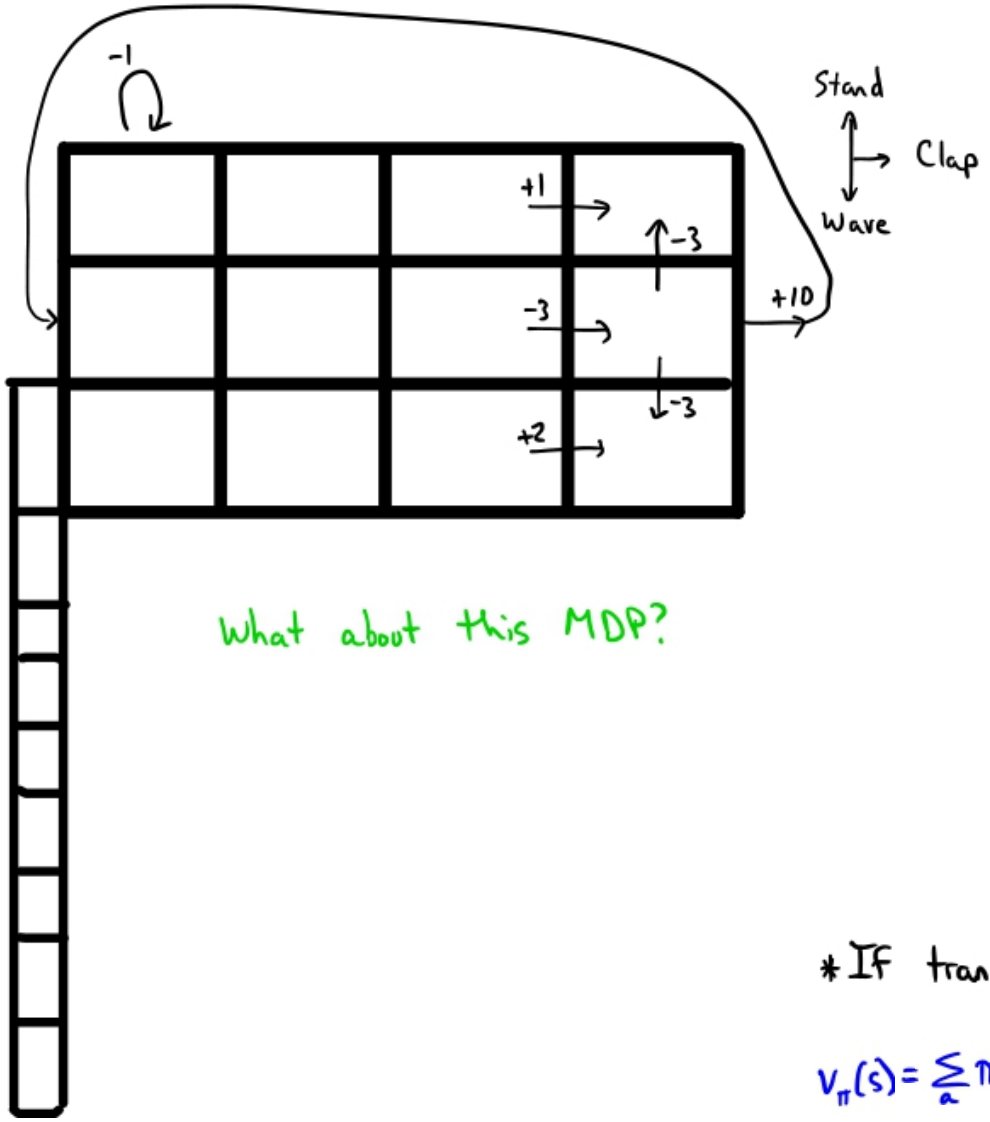
What's the steady state distribution of π_2?

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

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

Peter's slide

Differential value function



What about this MDP?

What's the steady state distribution of π_e?

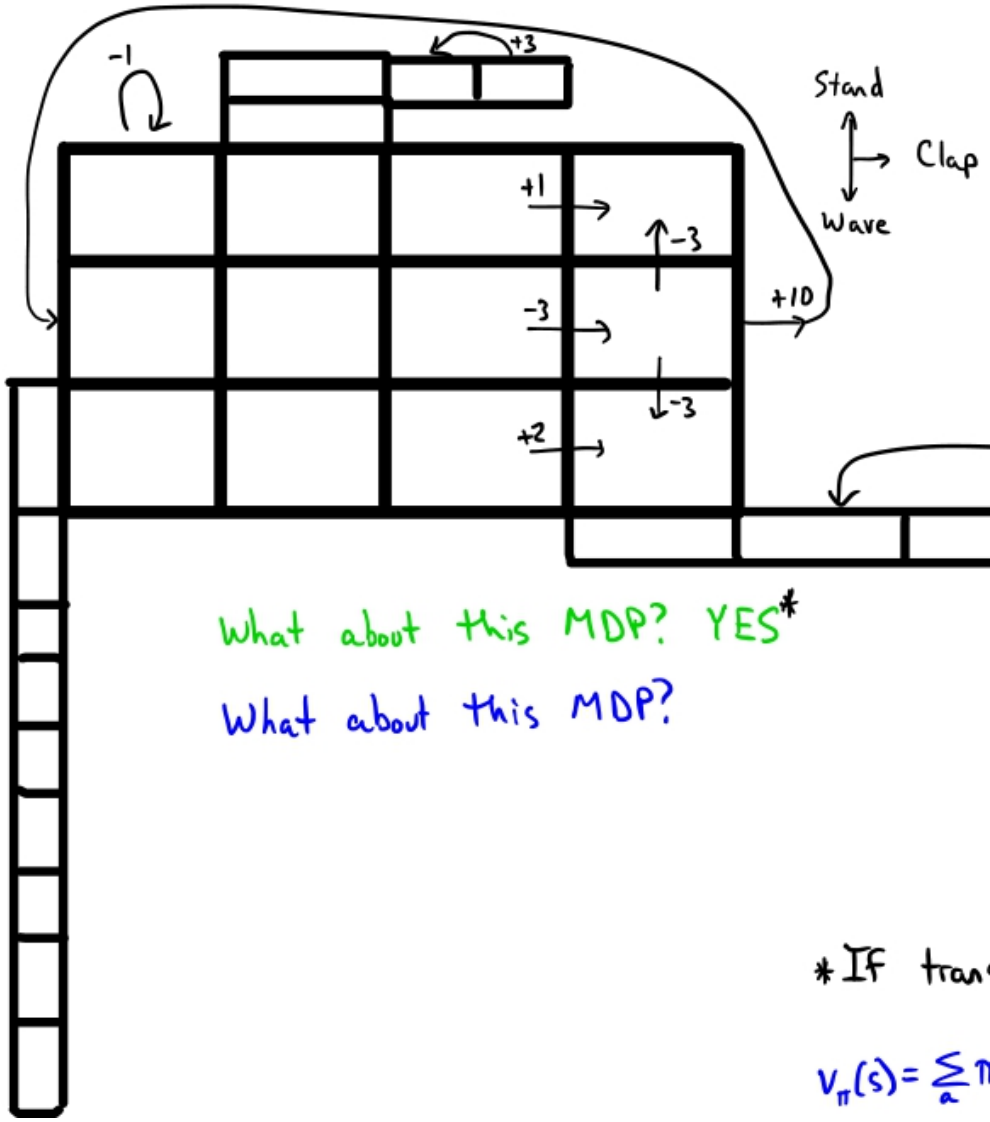
Is this MDP ergodic? YES*

* If transitions are at least slightly stochastic(↕)

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

Peter's slide

Differential value function



What's the steady state distribution of π_e?

Is this MDP ergodic? YES

What about this MDP? YES*

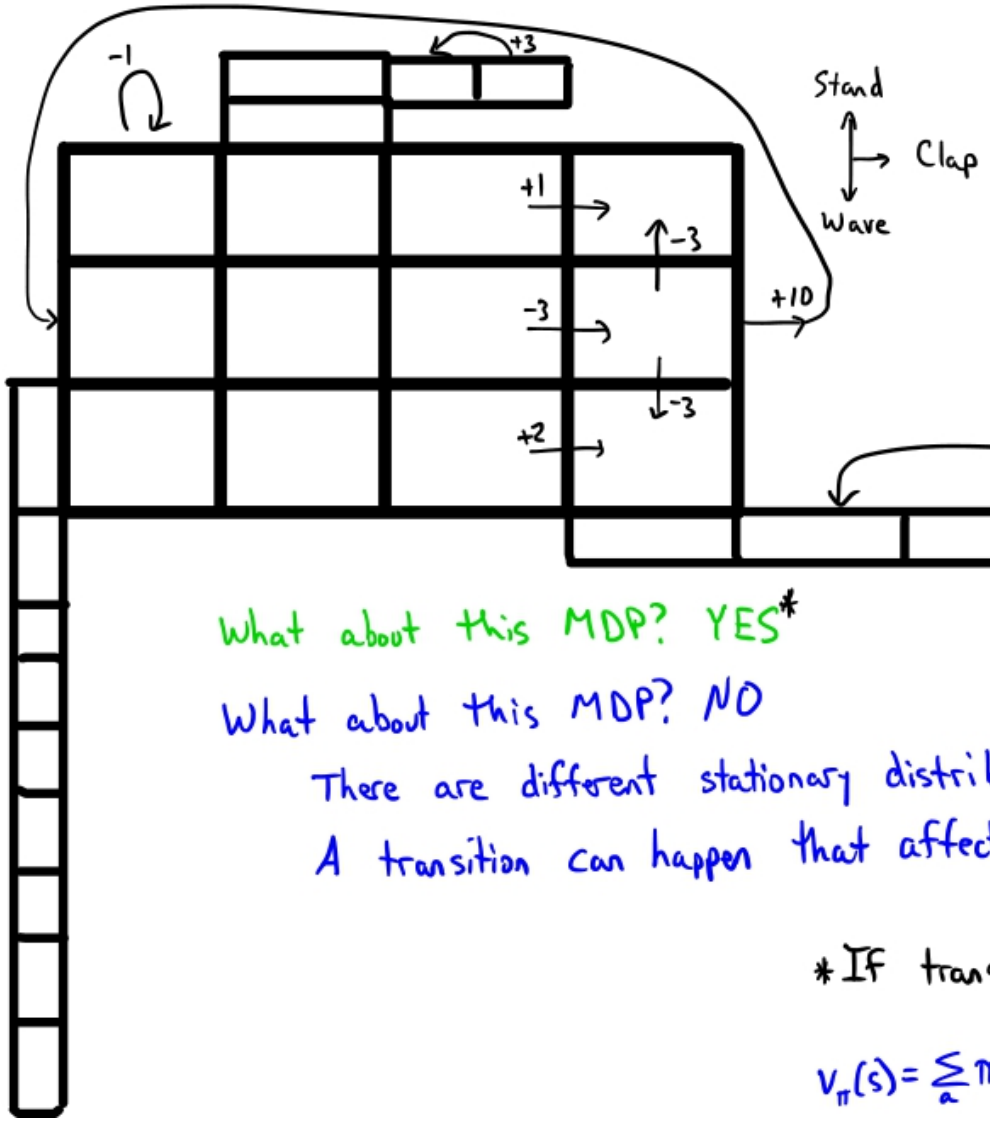
What about this MDP?

* If transitions are at least slightly stochastic(↕)

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

Peter's slide

Differential value function



What's the steady state distribution of π_e?

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(↕→)

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

Peter's slide

Reading Responses

[Gizem Toplu-Tutay]
How does the concept of ergodicity ensure the existence of the limits in the equations defining the average reward setting, and why is it important in the study of dynamic programming and reinforcement learning?

[Shikhar Gupta]
Is ergodicity less strict than the Markov condition we introduced in previous chapters? How is is the same or different, and why do we not use the same conditions for both?

Reading Responses

[Gizem Toplu-Tutay]
How does the concept of ergodicity ensure the existence of the limits in the equations defining the average reward setting, and why is it important in the study of dynamic programming and reinforcement learning?
Ergodicity ensures there is a steady state distribution that does not depend on initial state. Makes average reward easier to compute.

[Shikhar Gupta]
Is ergodicity less strict than the Markov condition we introduced in previous chapters? How is is the same or different, and why do we not use the same conditions for both?
Different assumptions. Can have ergodicity without Markov assumption, and vice versa.

Final Logistics

Next lecture:
Chapter 11 (through 11.4): Off-policy Methods with Approximation
Reading assignments due **2PM Monday**

Office hours:

Mon: Michael 1-2PM GDC Basement TA Station #5
Tues: Caroline 11:15-12:15PM
Wed: Amy 2-3PM EER 6.878
Thurs: Haoran 11-12PM; Siddhant 5-6PM
Fri: Shuozhe 4-5PM

Final Logistics

Final project proposal due at **11:59pm on Thursday, 3/7**

Complete Homework for Chapters 10+11 on edx by **Friday 11:59 PM CST**

Programming Assignment 3 is Ch 9 + 10 and due in 2 weeks

Complete Programming Assignment 2 on edx by **Sunday at 11:59 PM CST**