REINFORCEMENT LEARNING: THEORY AND PRACTICE

Ch. 3: Finite Markov Decision Processes

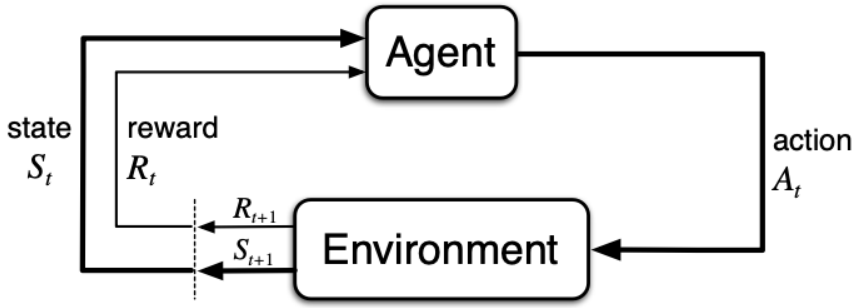
Profs. Amy Zhang and Peter Stone

Questions?

Attendance

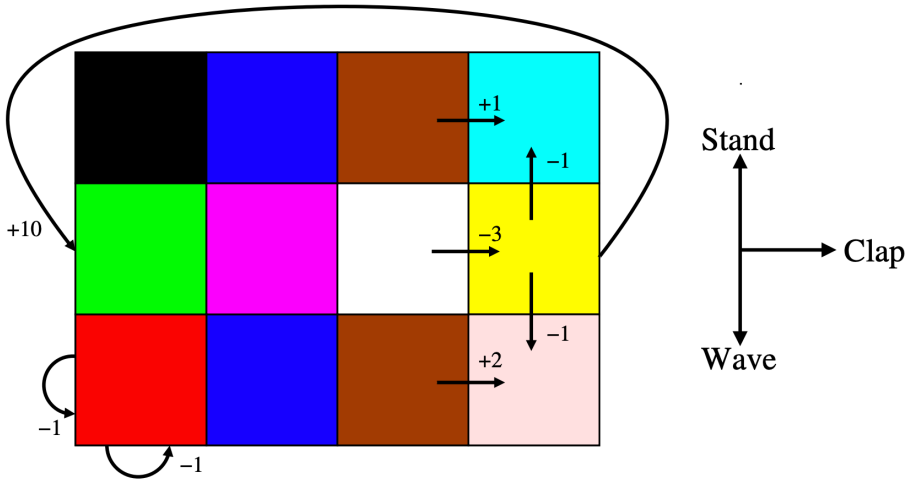
Take the attendance quiz on Canvas!

Answer: Value function

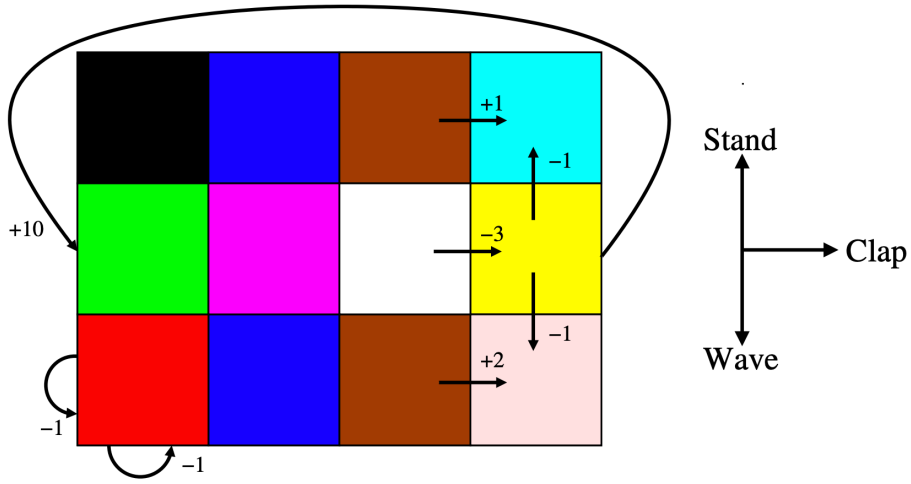


Design Exercise: Define an MDP (S, A, P, R) for one of the following:

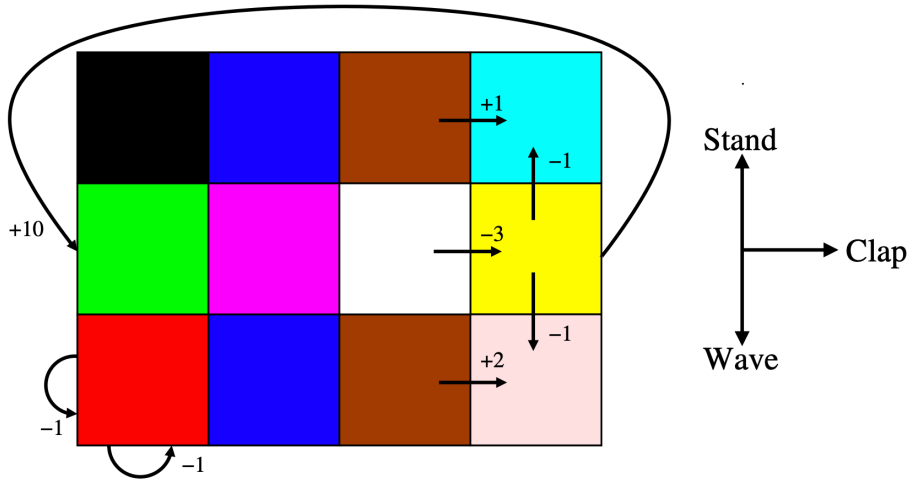
1. Recommendation system
2. Autonomous driving system
3. Advertising
4. A system of your choosing!


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

Is this Markovian?



How would we make it Markovian?



Discussion Groups:

What is the Markov property and when does it come up in real world problems?

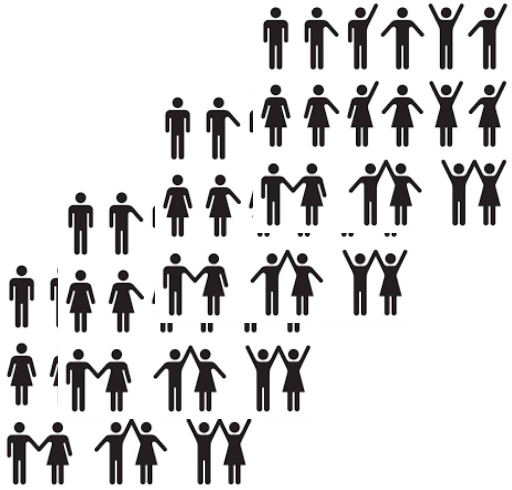
Why is the Markov property desirable?

Discussion Groups:

List some differences across bandits, contextual bandits and MDPs.

Bandits

Contextual Bandits



MDPs

Discussion Groups: Goals vs. Rewards

What are some limitations of the reward formalism?

Discussion Groups: Goals vs. Rewards

What if you have multiple objectives?

Autonomous driving example:

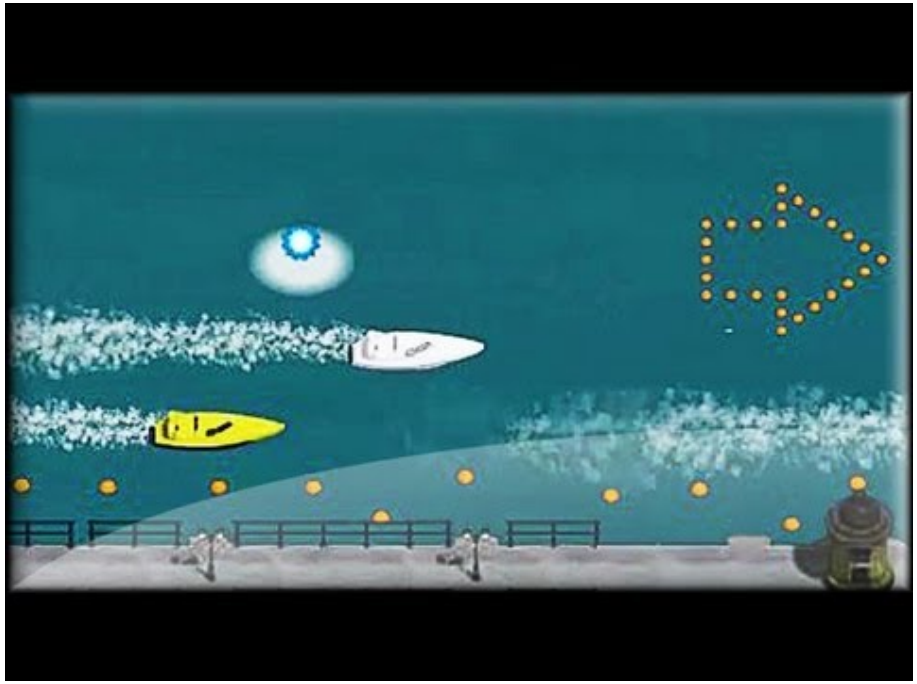
1. Don't hit other cars, people, or objects
2. Obey speed limits
3. Get to your destination quickly
4. Get there safely
5. Get there efficiently

“The reward signal is your way of communicating to the robot *what* you want it to achieve, not *how* you want it achieved.”

"The reward signal is your way of communicating to the robot *what* you want it to achieve, not *how* you want it achieved."

What's the difference, and why?

Coast Runners Game preview



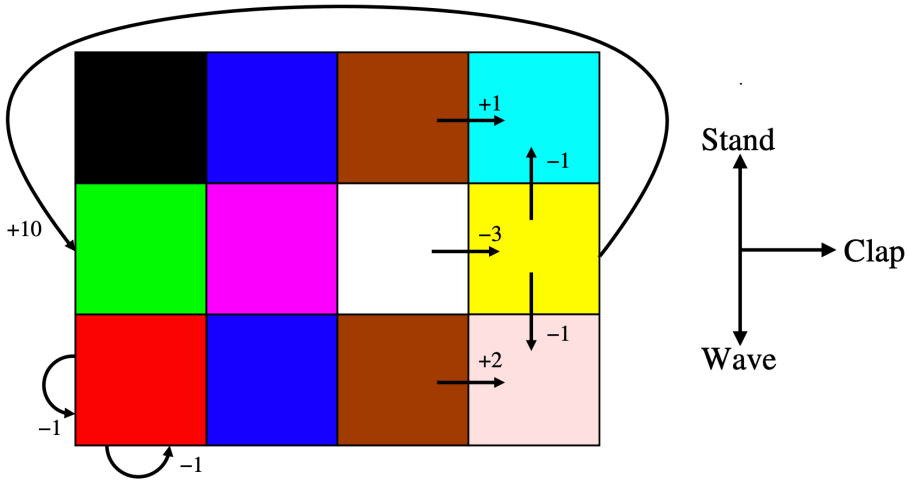
Reward Hacking

Boat gets higher reward for looping through three targets rather than finishing the race!

Discussion Groups:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{3.8}$$

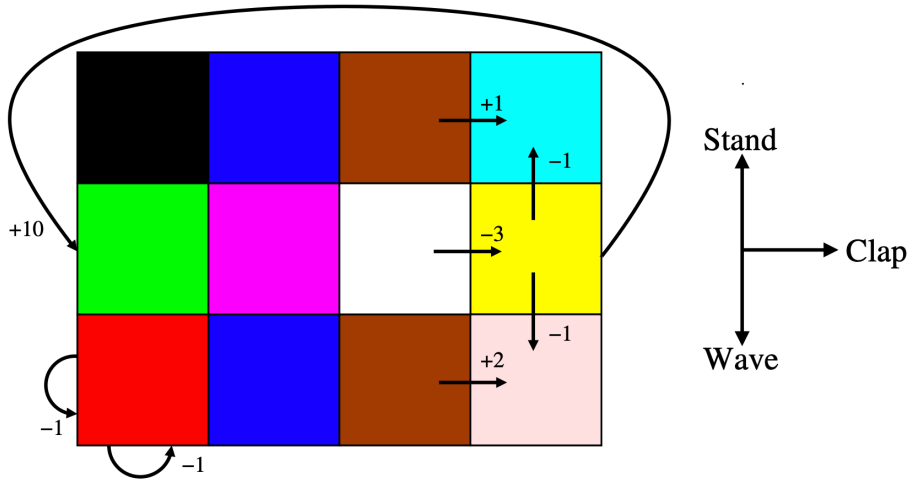
What happens if discount rate is 0?



Discussion Groups:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{3.8}$$

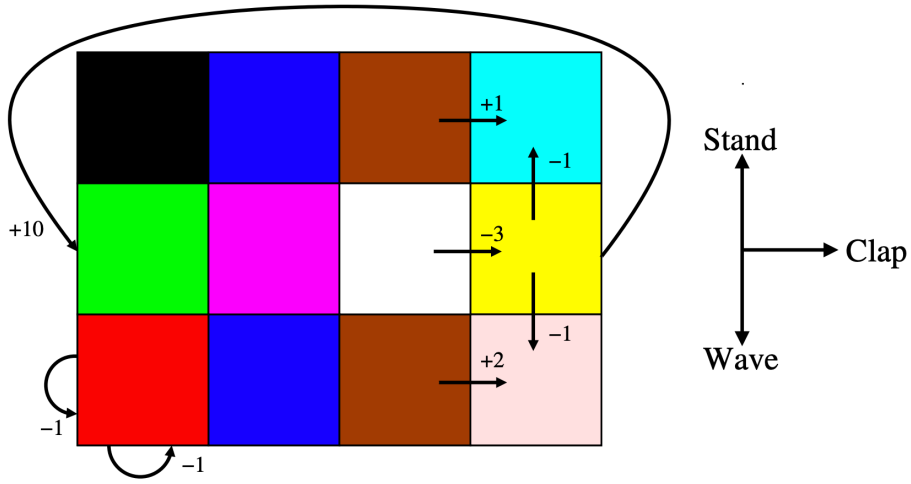
What happens if discount rate is 1?



Discussion Groups:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{3.8}$$

What happens if discount rate is 1?

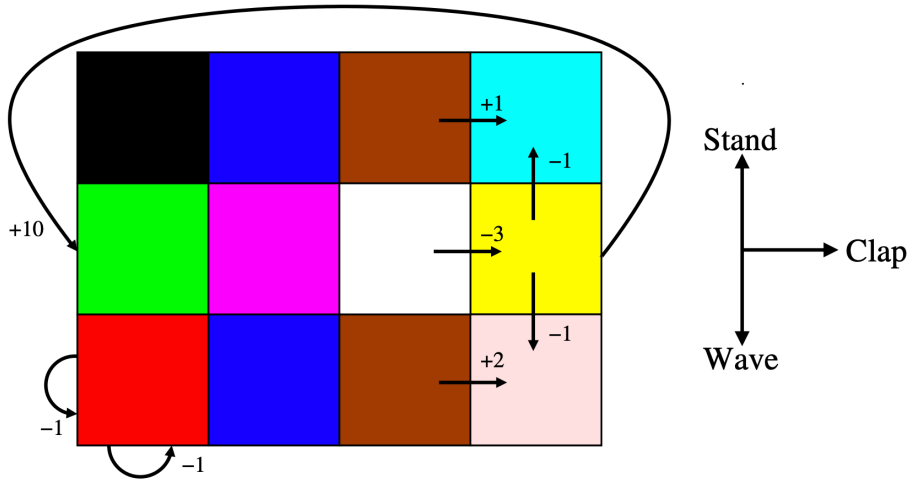


What if it's episodic?

Discussion Groups:

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \tag{3.8}$$

What happens if discount rate is 1?



What if it's episodic?

What if it's continuous?

Discussion Groups: Episodic vs. Continuous

Come up with some examples of applications that are naturally episodic vs continuous.

- Games (chess, video games, soccer)
- Robots (when battery runs down)
- Autonomous driving
- Recommendation systems

Episodic

- Games (chess, video games, soccer)
- Robots (when battery runs down)
- Autonomous driving
- Recommendation systems

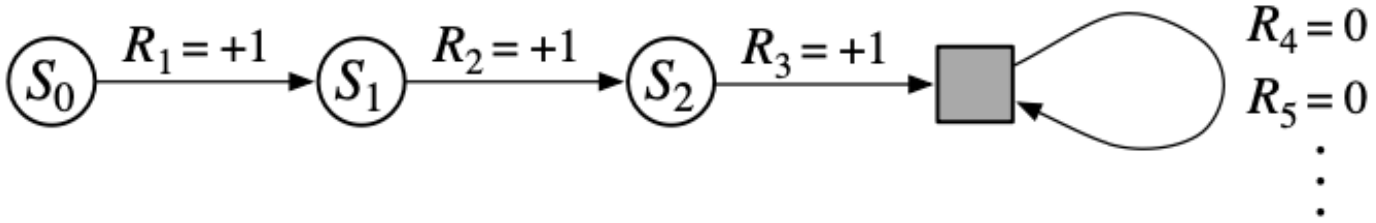
Episodic

- Robots
- Autonomous driving
- Ads
- Recommendation systems

Continuous

Unifying Episodic vs. Continuous

Unifying Episodic vs. Continuous



Transferring Episodic into Continuous

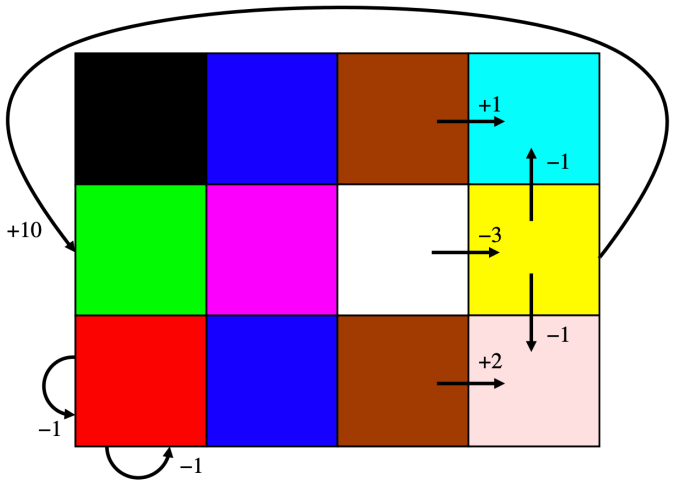
Discussion Groups: Policies and value functions

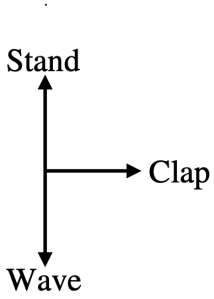
Pros and cons of learning a state value function vs. state-action value function

Compute Monte Carlo updates

Random policy: Agent selects all three actions with equal probability in all states.

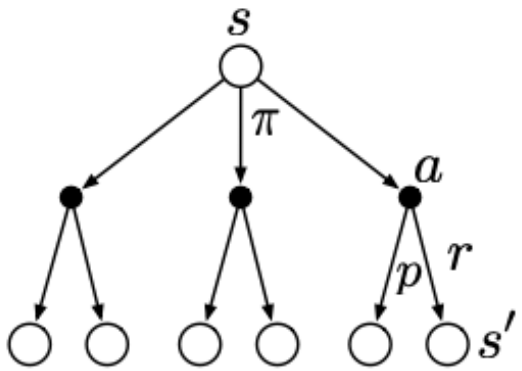
Discount factor = 0.5



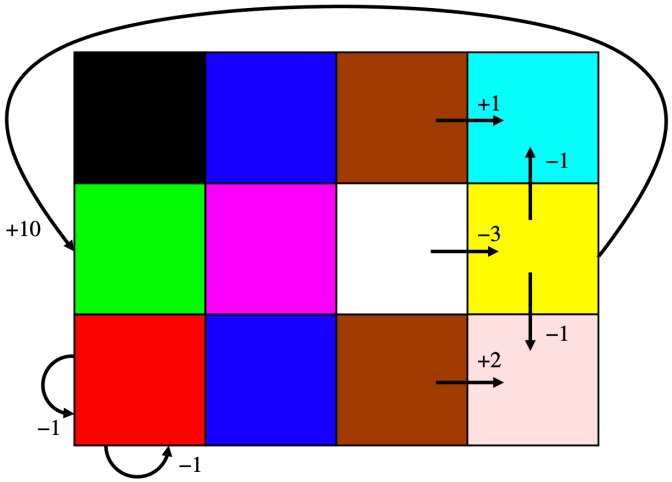


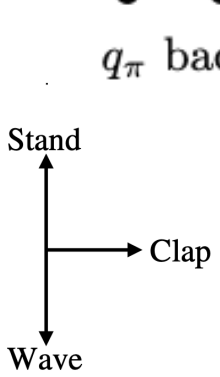
1	2	3	4
5	6	7	8
9	10	11	12

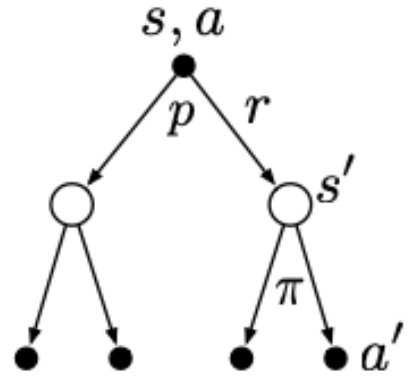
Discussion Groups: Backup Diagram



Backup diagram for v_π







q_π backup diagram

Draw out 2 steps of the backup diagram for a random policy

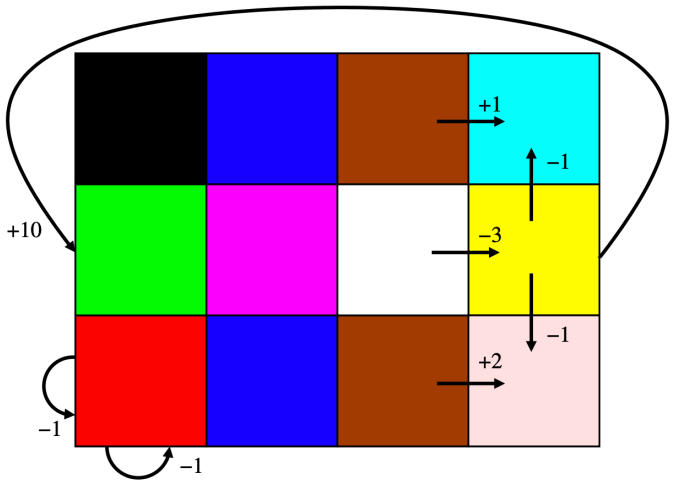
Random policy: Agent selects all three actions with equal probability in all states.

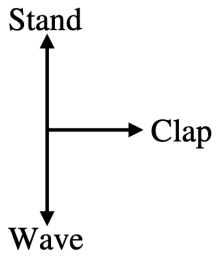
Compute Monte Carlo updates

Random policy: Agent selects all three actions with equal probability in all states.

Discount factor = 0.5

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right], \text{ for all } s \in \mathcal{S},$$





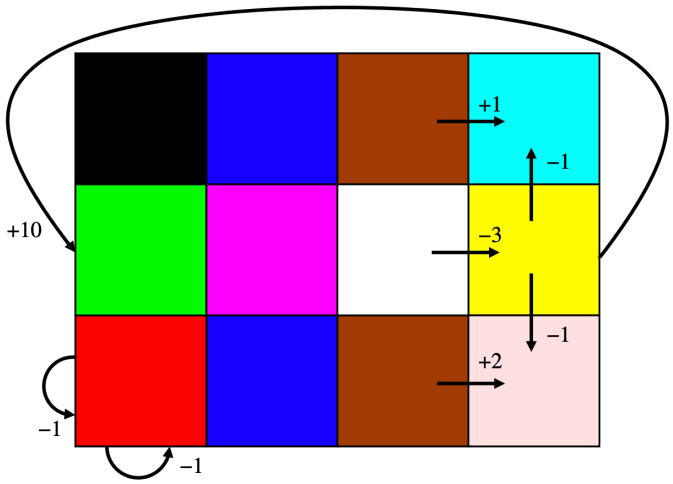
1	2	3	4
5	6	7	8
9	10	11	12

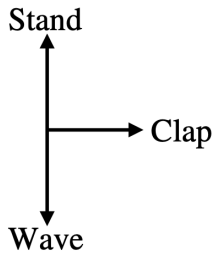
Compute Monte Carlo updates

Random policy: Agent selects all three actions with equal probability in all states.

Discount factor = 0.5

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t \mid S_t = s] = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \,\middle|\, S_t = s\right], \text{ for all } s \in \mathcal{S},$$



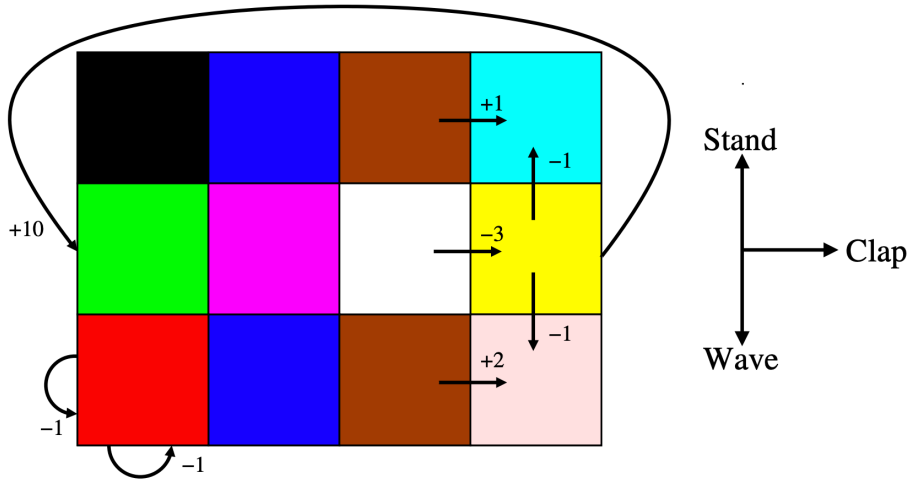


-0.33	-0.33	0	-0.67
0	0	-1	2.67
-0.33	-0.33	0.33	-0.67

Discussion Groups: Optimal Policies and Value Functions

What's the optimal policy?

How did you find it?



Reading Response Questions

(Arnav Jain) For tasks with large amounts of states or actions, what are some ways to abstract them and make them solvable using RL?

(Jackson Paull) When modeling and programming a reinforcement learning problem in practice it seems there are several areas of potential fault if the program is not behaving as expected (i.e. learning), those being issues with the description of the state space (potentially leaving out some important signal from the environment), issues with the reward schema (not properly incentivizing what you want), or even choice of algorithm. Of these, which is most common as an issue? Which would affect an RL agent the most?

Reading Response Questions

(Emin Arslan) The reward is always considered to be a part of the environment outside the agent. What if the reward was partially under the control of the agent? This would allow the agent to change its goals. This sounds like something we do as humans. It could also give the agent the ability to shape its reward function in a way such that it complements the part of its reward function that it can't change.

(Jiaheng Hu) In a normal MDP formulation, we assume no knowledge of the reward function. What can we do to speed up the learning if we do have some knowledge about how the reward is computed?

A Summary

1. MDPs
2. Goals and Rewards
3. Episodic vs. Continuous Environments
4. Discount factor
5. Backup diagrams
6. Policies and value functions
7. Optimal policy and value functions

Final Logistics

Next lecture:
Chapter 4: Dynamic Programming
Reading assignments due **2PM Monday**

Final Logistics

Next lecture:
Chapter 4: Dynamic Programming
Reading assignments due **2PM Monday**

Office hours start **this week:**

Mon: Michael 1-2PM GDC Basement TA Station #5
Tues: Peter 11-11:50AM GDC 3.508; 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

Coding assignment for Chapter 2 on edX **due Feb 4th 11:59PM CST**

Chapter 2 Homework on edX was **due Sunday 11:59PM CST**

Chapter 3+4 Homework on edX **due Friday 11:59PM CST**