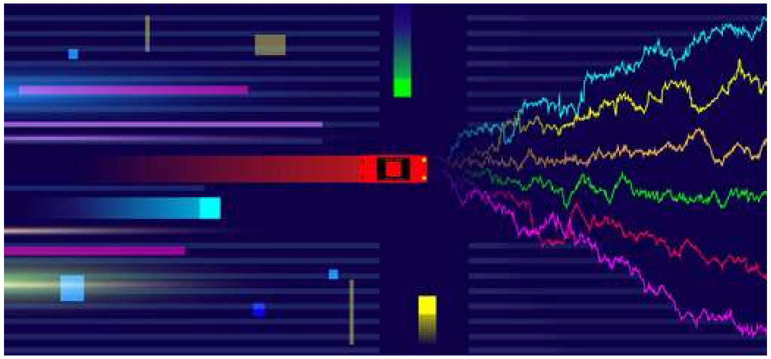
RL for Operations
Day 1: MDP Basics, VI+PI, Deep RL

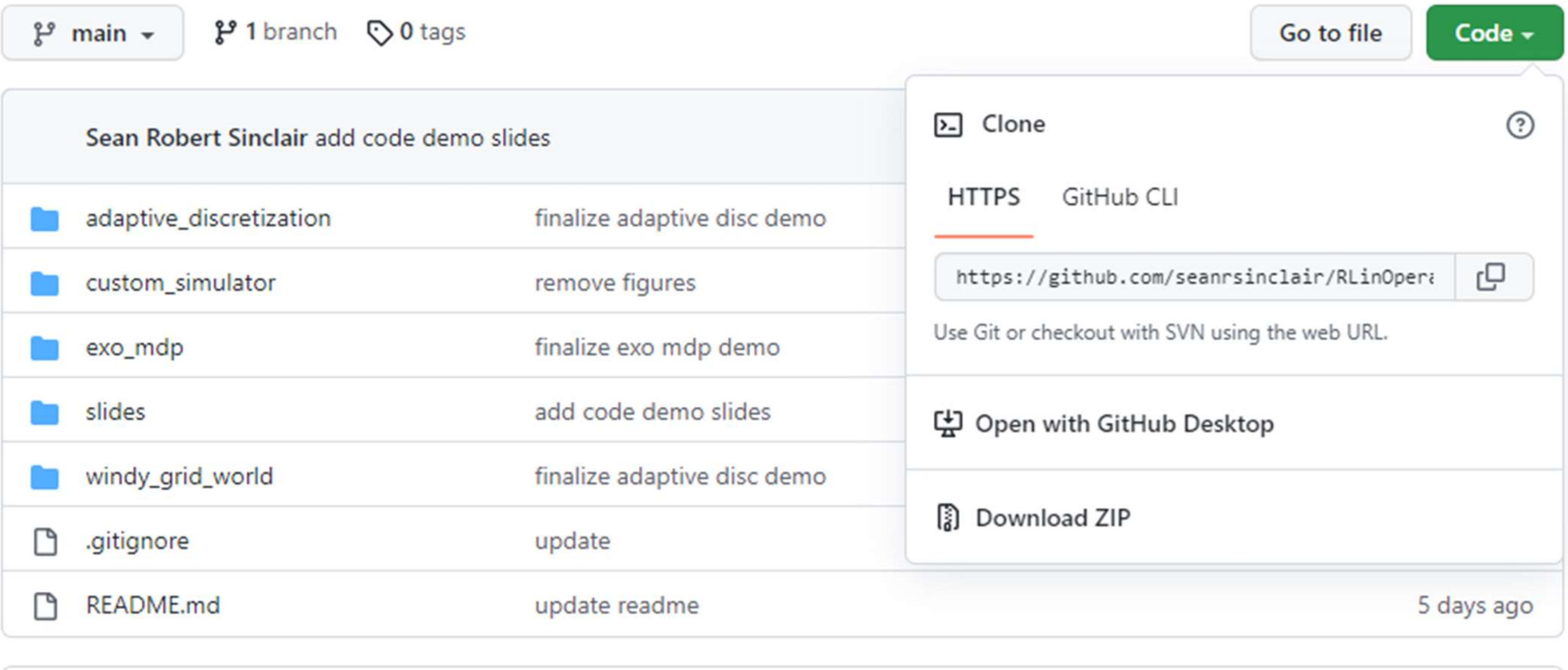
Sean Sinclair, Sid Banerjee, Christina Yu
Cornell University



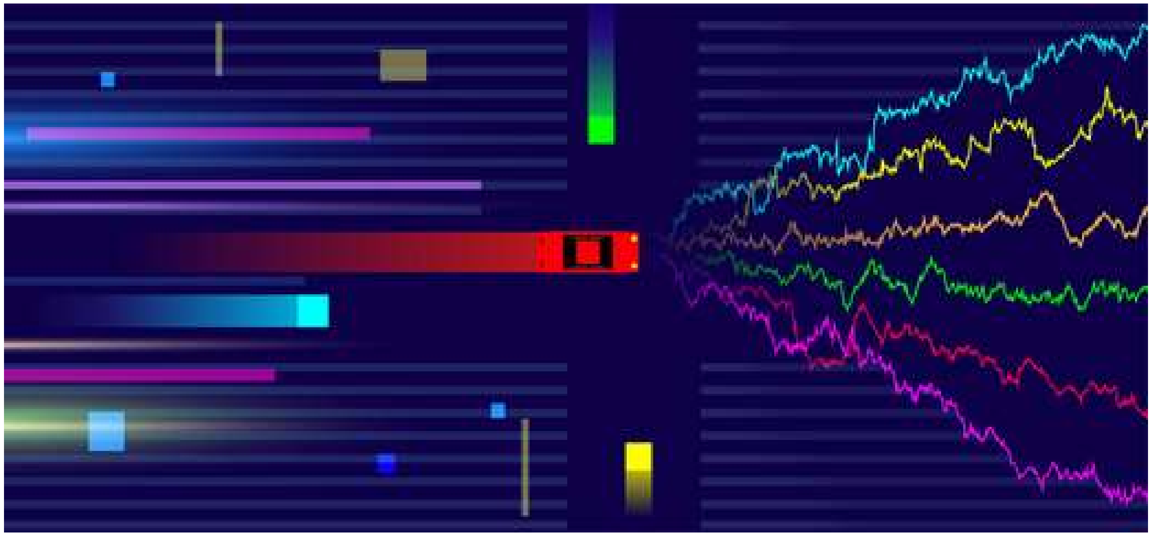
RL for Operations
Day 1: MDP Basics, VI+PI, Deep RL

Sean Sinclair, Sid Banerjee, Christina Yu
Cornell University

https://github.com/seanrsinclair/RLinOperations



https://github.com/seanrsinclair/RLinOperations



Data Driven Decision Making at Simons

- First bootcamp August 22nd to August 26th
- Live streamed on youtube + zoom
- https://simons.berkeley.edu/workshops/datadriven-2022-bc

Plan for Today

MDP Basics

- Basic framework for Markov Decision Processes
- Tabular RL Algorithms with policy iteration + value iteration
- DeepRL algorithms (and their “tabular” counterparts)

Simulation Packages

- OpenAI Framework for simulation design
- Existing packages and code-bases for RL algorithm development

Simulation Implementation

- Develop simulator for problem using OpenAI Gym API

Tabular RL Algorithms

- Implement basic tabular RL algorithms to understand key algorithmic design aspects of *value estimates + value iteration*, *policy iteration*

Plan for Today

MDP Basics

- Basic framework for Markov Decision Processes
- Tabular RL Algorithms with policy iteration + value iteration
- DeepRL algorithms (and their "tabular" counterparts)

Simulation Packages

- OpenAI Framework for simulation design
- Existing packages and code-bases for RL algorithm development

Simulation Implementation

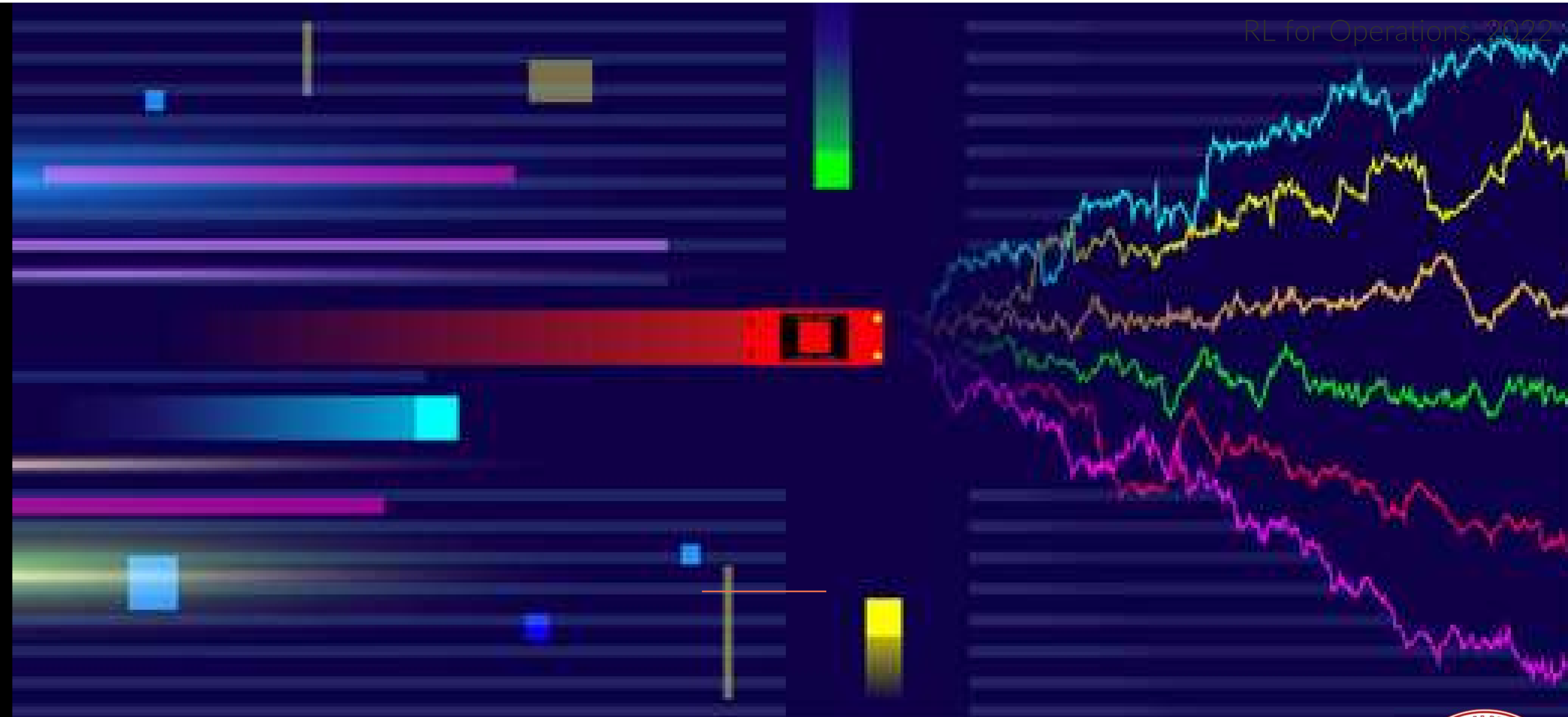
- Develop simulator for problem using OpenAI Gym API

Tabular RL Algorithms

- Implement basic tabular RL algorithms to understand key algorithmic design aspects of *value estimates + value iteration*, *policy iteration*

MDP Basics

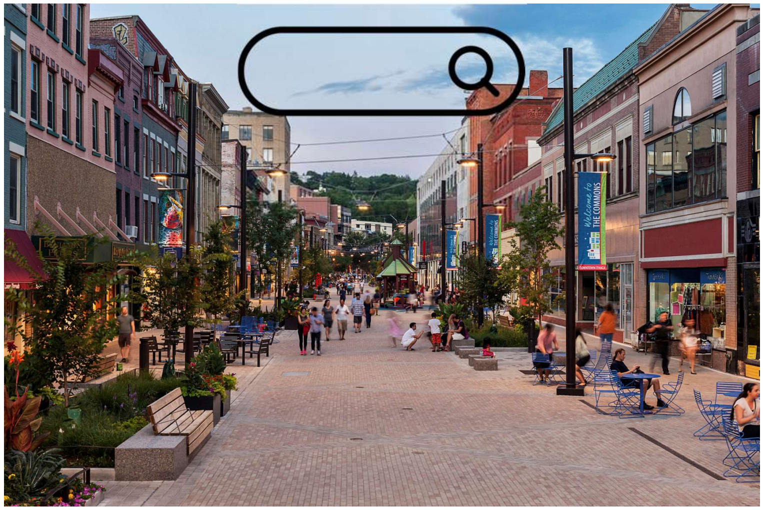
Sid Banerjee
Cornell University



Background

A Story

Typical question: “which decision is better”



City of Ithaca homepage photo

A/B Test

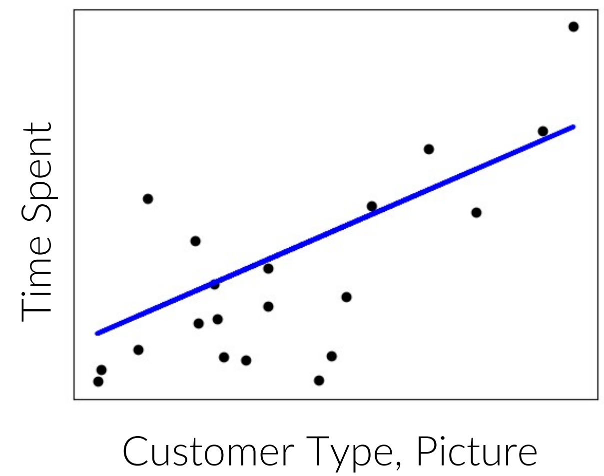
Take users, divide randomly, observe which has longer visit times

Supervised Learning

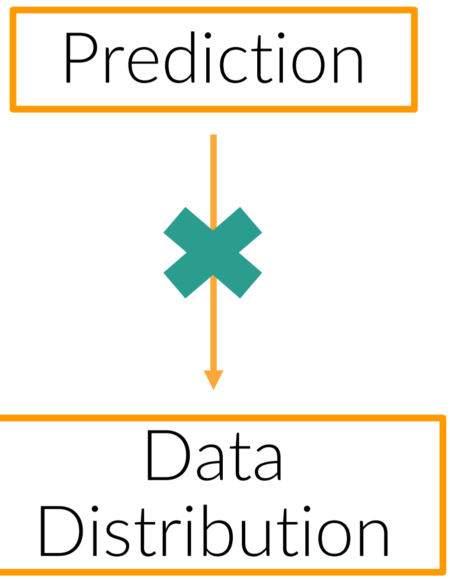
( ,  , 14 mins)

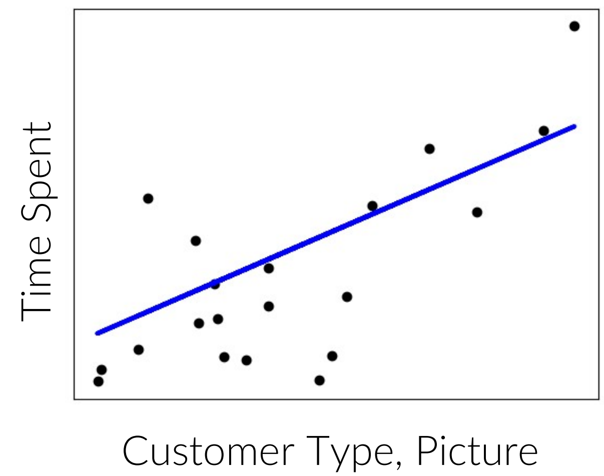
( ,  , 8 mins)

( ,  , 24 mins)



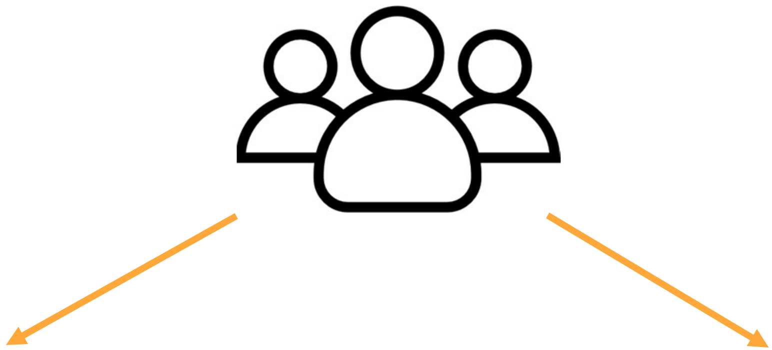
Supervised Learning

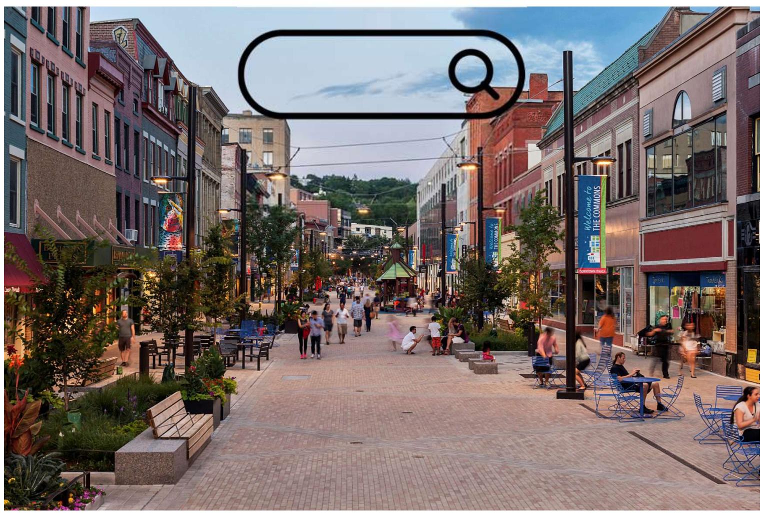




Theory and practice relies on prediction not affecting data distribution

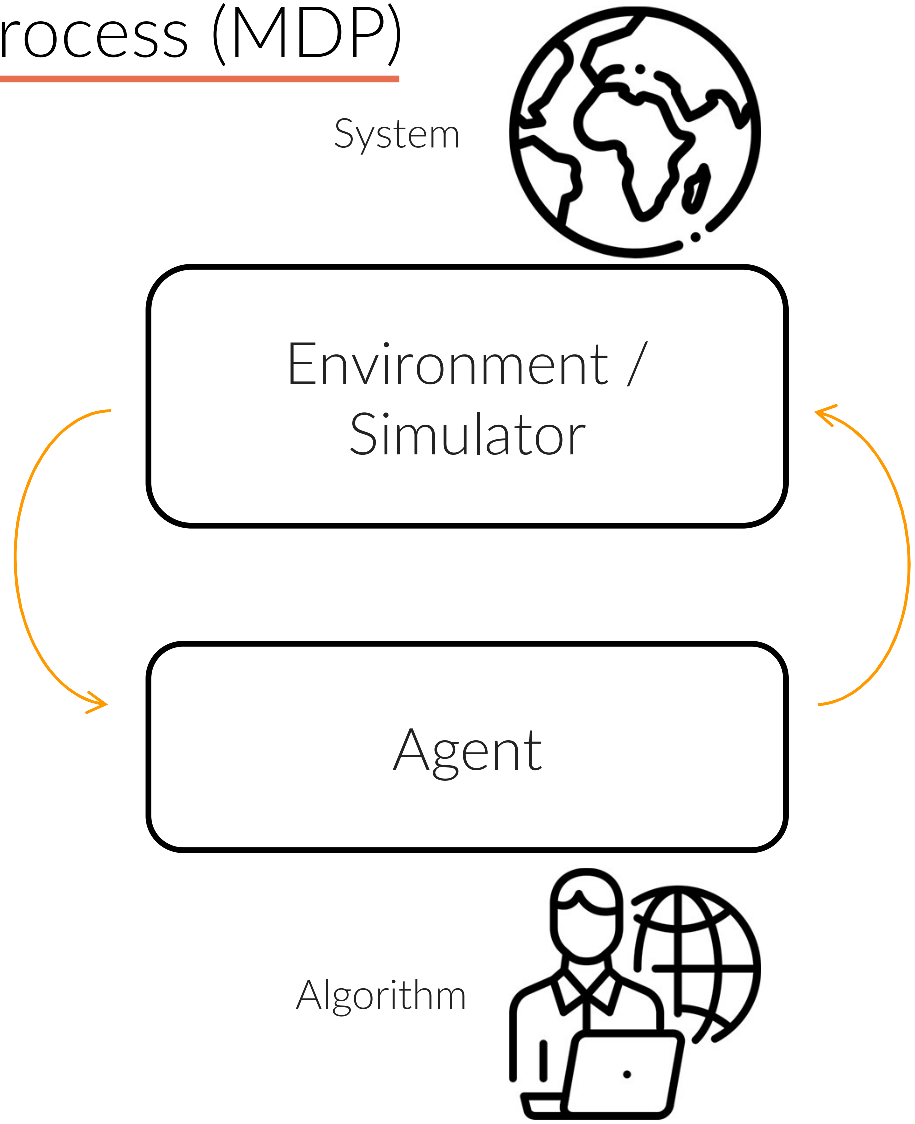
Bandit Algorithms





Adaptively partition users based on observed feedback thus far

Markov Decision Process (MDP)



Markov Decision Process (MDP)

$$\pi(s) \rightarrow \Delta(\mathcal{A})$$

Policy: Determine action based on state

Markov Decision Process (MDP)

Environment: Determine reward and new state

$$r(s,a), s' \sim T(\cdot \mid s, a)$$

$$\pi(s) \to \Delta(\mathcal{A})$$

Policy: Determine action based on state

Markov Decision Process (MDP)

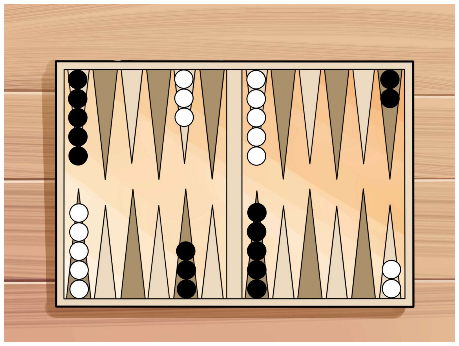
Environment: Determine reward and new state

Policy: Determine action based on state

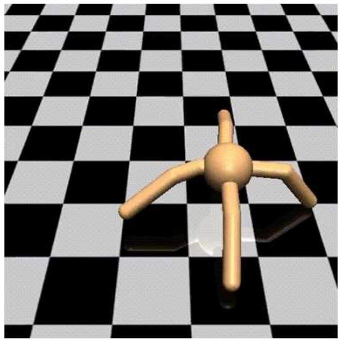
MDP vs Supervised Learning

	Learn from Experience	Generalize	Interactive	Exploration	Credit Assignment
Supervised Learning	✓	✓	✗	✗	✗
Reinforcement Learning	✓	✓	✓	✓	✓

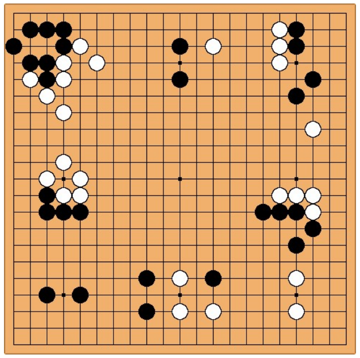
Success of RL



Backgammon



MuJoCo Simulator



AlphaGo Zero

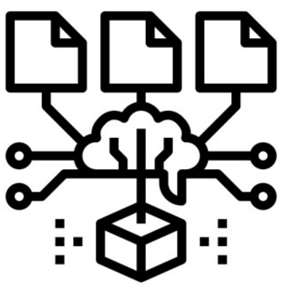
Focused on game playing + robotics

[Silver2017, Tesauro1995]

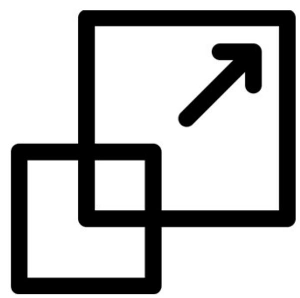
This Workshop

This workshop focuses on RL for Operations

We care about:

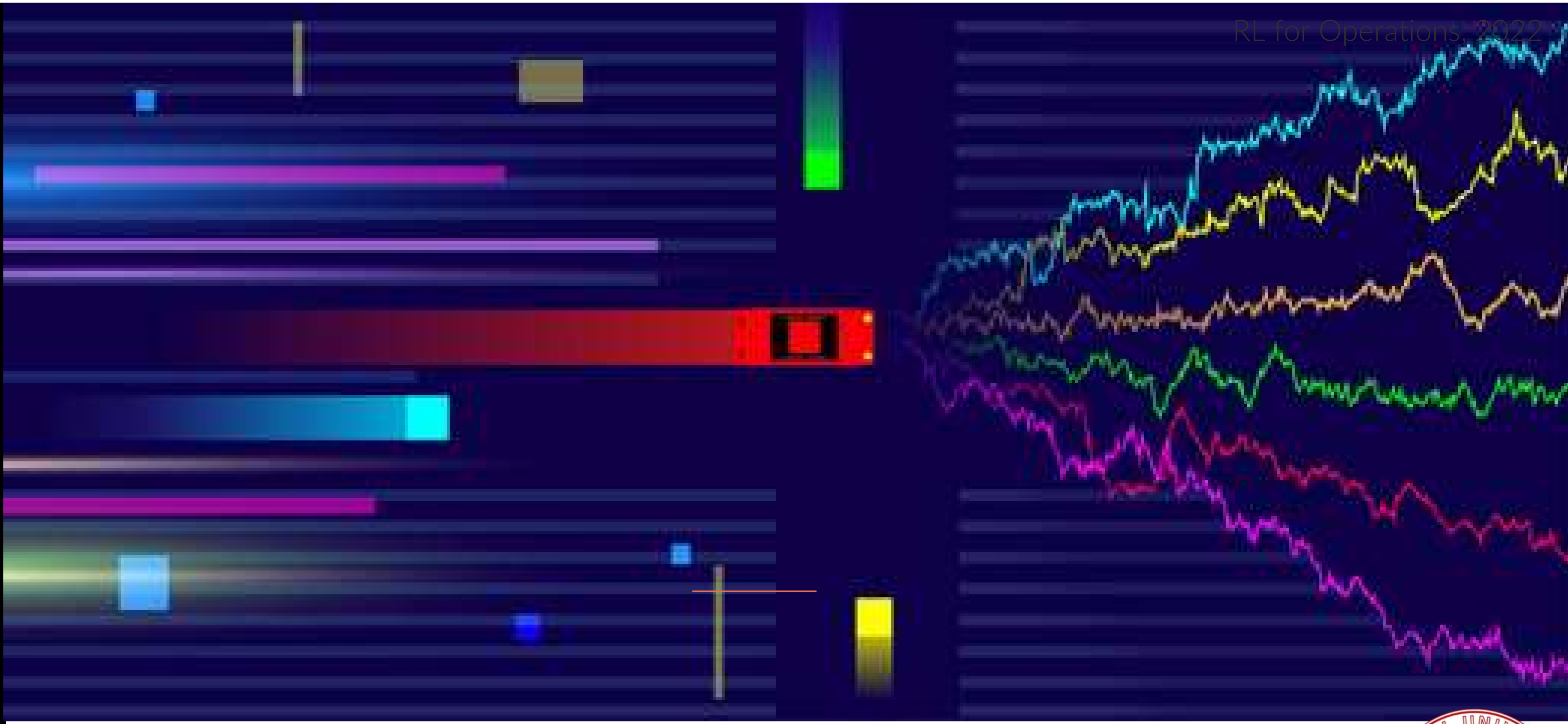


OR Models



Computation + Scale

Impact



Formulating an MDP

3 ‘Flavors’ of MDPs

- Finite horizon
- Infinite horizon (discounted)
- Infinite horizon (average cost)

Finite Horizon

defined by: $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, r, T, s_0, H\}$

$\mathcal{S}$	State space
$\mathcal{A}$	Action space
$r_h : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$	Rewards
$T_h : \mathcal{S} \times \mathcal{A} \rightarrow \Delta(\mathcal{S})$	Transitions
H	Time horizon

Finite Horizon

Infinite Horizon (Discounted)

defined by: $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, r, T, s_0, \gamma\}$

$\mathcal{S}$	State space
$\mathcal{A}$	Action space
$r : \mathcal{S} \times \mathcal{A} \to [0, 1]$	Reward
$T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$	Transitions
$\gamma \in [0, 1)$	Discount

Infinite Horizon (Discounted)

Infinite Horizon (Average Cost)

defined by: $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, r, T, s_0\}$

$\mathcal{S}$	State space
$\mathcal{A}$	Action space
$r : \mathcal{S} \times \mathcal{A} \to [0, 1]$	Reward
$T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$	Transitions

Which flavor for you?

Infinite Horizon

- Transition, rewards, policy, not allowed to depend on timestep
  - Optimal policy is stationary
- "Less" importance on future rewards
  - Initial/terminal conditions 'wash out'

$$\lim_{T\to\infty} \frac{1}{T}(r_1 + r_2 + \ldots + r_T)$$

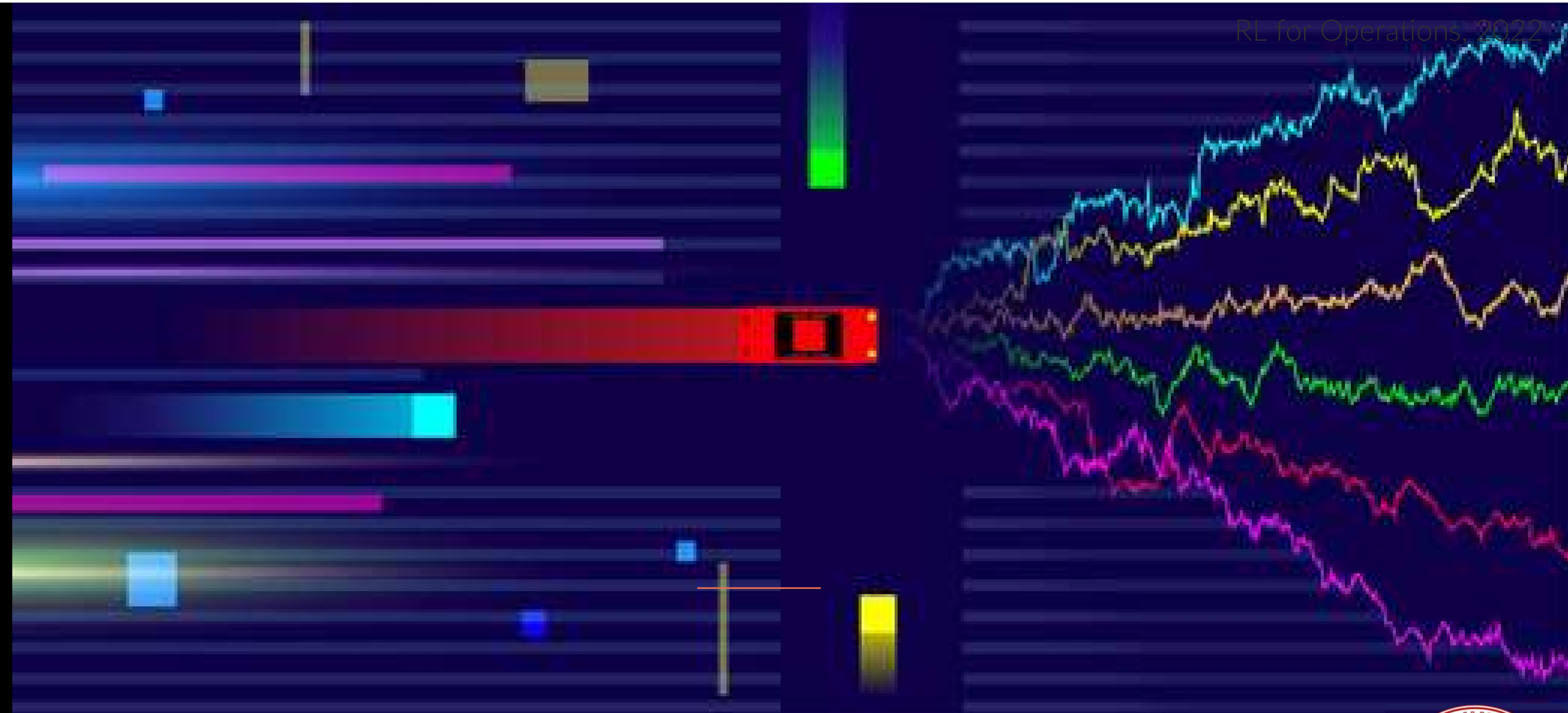
Finite Horizon

- Transition, rewards, policy *allowed* to depend on timestep
  - Optimal policy is time-dependent
- "More" importance on future rewards
  - Initial/terminal conditions matter

$$r_1 + r_2 + r_3 + \ldots + r_H$$

$$r_1 + \gamma r_2 + \gamma^3 r_3 + \gamma^4 r_4 \ldots$$

Comments



Solving an MDP

Markov Decision Process (MDP)

Environment: Determine reward and new state

$$r(s,a), s' \sim T(\cdot \mid s,a)$$

$$\pi(s) \to \Delta(\mathcal{A})$$

Policy: Determine action based on state

State-Action Frequencies

Suppose you want to measure performance of given policy $\pi(s) \rightarrow \Delta(\mathcal{A})$

State-Action Frequencies

Suppose you want to measure performance of given policy $\pi(s) \to \Delta(\mathcal{A})$

State-Action Frequencies

What can we say about $\nu^{\pi}(s, a)$

State-Action Frequencies

What can we say about $\nu^\pi(s,a)$

The state-action frequency LP

Putting things together, we have the following LP

$$\max \quad \sum_{s,a} \nu(s,a) r(s,a)$$

subject to

$$\sum_{a} \nu(s,a) = \mathbb{I}_{[s_0 = s]} + \gamma \sum_{s',a'} \nu(s',a') T(s|s',a') \qquad \forall\, s \in \mathcal{S}$$

$$\nu(s,a) \geq 0 \qquad \forall\, s \in \mathcal{S}, a \in \mathcal{A}$$

Some duality magic!

Sid's maxim: When life gives you an LP, take its dual!

$$\max \quad \sum_{s,a} \nu(s,a) r(s,a) \quad \text{subject to}$$

$$\sum_{a} \nu(s,a) = \mathbb{I}_{[s_0 = s]} + \gamma \sum_{s',a'} \nu(s',a') T(s|s',a') \qquad \forall\, s \in \mathcal{S}$$

$$\nu(s,a) \geq 0 \qquad \forall\, s \in \mathcal{S}, a \in \mathcal{A}$$

The Dual LP

$$\max \quad \sum_{s} \mathbb{I}[s_0 = s] V(s)$$

subject to

$$V(s) \leq r(s,a) + \gamma \sum_{s'} T(s'|s,a) V(s') \qquad \forall\, s \in \mathcal{S}, a \in \mathcal{A}$$

The 'Bellman' LP

$$\max \quad \sum_{s} \mathbb{I}[s_0 = s] V(s)$$

subject to

$$V(s) \leq \min_{a \in \mathcal{A}} \left\{ r(s,a) + \gamma \mathbb{E}_{S' \sim T(\cdot|s,a)} [V(S')] \right\} \qquad \forall\, s \in \mathcal{S}$$

Value Function

The Value Function is expected return for policy $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(S_h, A_h) \mid S_0 = s, A_h \sim \pi(S_h), S_{h+1} \sim T(\cdot \mid S_h, A_h)\right]$$

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(S_h, A_h) \mid (S_0, A_0) = (s,a), A_h \sim \pi(S_h), S_{h+1} \sim T(\cdot \mid S_h, A_h)\right]$$

Starting State

Actions by policy

Next state by environment

Expectation over randomness in policy and transitions

Bellman Equation

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(S_h, A_h) \mid S_0 = s, A_h \sim \pi(S_h), S_{h+1} \sim T(\cdot \mid S_h, A_h)\right]$$

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(S_h, A_h) \mid (S_0, A_0) = (s,a), A_h \sim \pi(S_h), S_{h+1} \sim T(\cdot \mid S_h, A_h)\right]$$

In other words, the Bellman Equations encode that:

$$V^\pi(s) = \mathbb{E}_{A\sim\pi(s)}[r(s,A) + \gamma \mathbb{E}_{S'\sim T(\cdot|s,A)}[V^\pi(S')]]$$

$$Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{S'\sim T(\cdot|s,a)}[V^\pi(S')]$$

Optimal Policy

For an infinite horizon discounted MDP, there exists a deterministic stationary policy:

$$\pi^* : \mathcal{S} \rightarrow \mathcal{A}, \text{ s. t. } V^{\pi^*}(s) \geq V^{\pi}(s) \quad \forall s, \pi$$

Optimal Policy

For an infinite horizon discounted MDP, there exists a deterministic stationary policy:

$$\pi^* : \mathcal{S} \to \mathcal{A}, \text{ s. t. } V^{\pi^*}(s) \geq V^\pi(s) \;\; \forall s, \pi$$

Denote $V^* = V^{\pi^*}, Q^* = Q^{\pi^*}$

Our goal is to find this policy, either looking at:

- Sample complexity (statistics)
- Optimization complexity

Bellman Optimality

The optimal policy satisfies Bellman Optimality equation:

$$V^*(s) = \max_{a \in \mathcal{A}} r(s,a) + \gamma \mathbb{E}_{S' \sim T(\cdot|s,a)}[V^*(S')]$$

Q-greedy policy: $\pi^*(s) = \operatorname{argmax}_a Q^*(s,a)$

$$V^\pi(s) = \mathbb{E}_{A \sim \pi(s)}[r(s,A) + \gamma \mathbb{E}_{S' \sim T(\cdot|s,A)}[V^\pi(S')]]$$

Fixed Point Uniqueness

If $V(s) = \max_{a \in \mathcal{A}} r(s,a) + \gamma \mathbb{E}_{S' \sim T(\cdot|s,a)}[V(S')]$

then $V(s) = V^*(s) \forall s$

What about the other MDP formulations

What about the other MDP formulations

Are all formulations equal?

References

[Puterman1994] Martin Puterman. “Markov Decision Processes: Discrete Stochastic Dynamic Programming”. *John Wiley + Sons*, 1994.
[Sutton2018] Richard Sutton. “Reinforcement Learning: An Introduction.” *MIT Press*, 2018.
[Agarwal2021] Alekh Agarwal, Nan Jiang, Sham M. Kakade, Wen Sun. “Reinforcement Learning: Theory and Algorithms”. 2021.
[Slivkins2019] Alexsandrs Slivkins. “Introduction to Multi-Armed Bandits.” *Foundations and Trends in ML*, 2019.
[Powell2021] Warren Powell. “Reinforcement Learning and Stochastic Optimization.” 2021.
[Meyn2021] Sean Meyn. “Control Systems and Reinforcement Learning”. *Cambridge University Press*, 2021.

Course Slides

Cornell CS6789: Foundations of Reinforcement Learning
https://wensun.github.io/CS6789_fall_2021.html
Stanford CS 234: Reinforcement Learning
https://web.stanford.edu/class/cs234/
UCL COMPM050: Course on RL
https://www.davidsilver.uk/teaching/