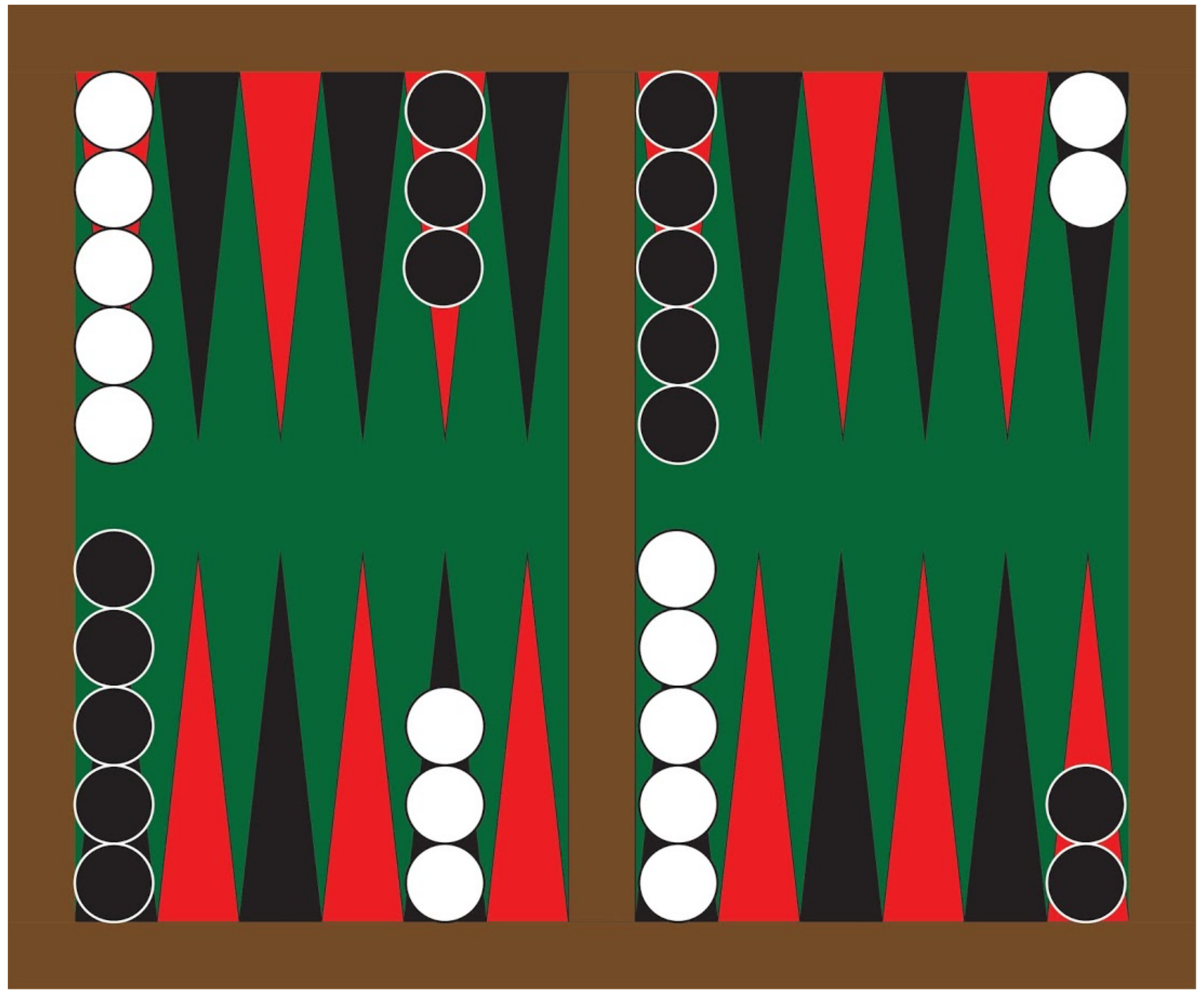
Introduction and Basics of Markov Decision Process

Wen Sun

CS 6789: Foundations of Reinforcement Learning

The very successful stories of ML are based on RL…



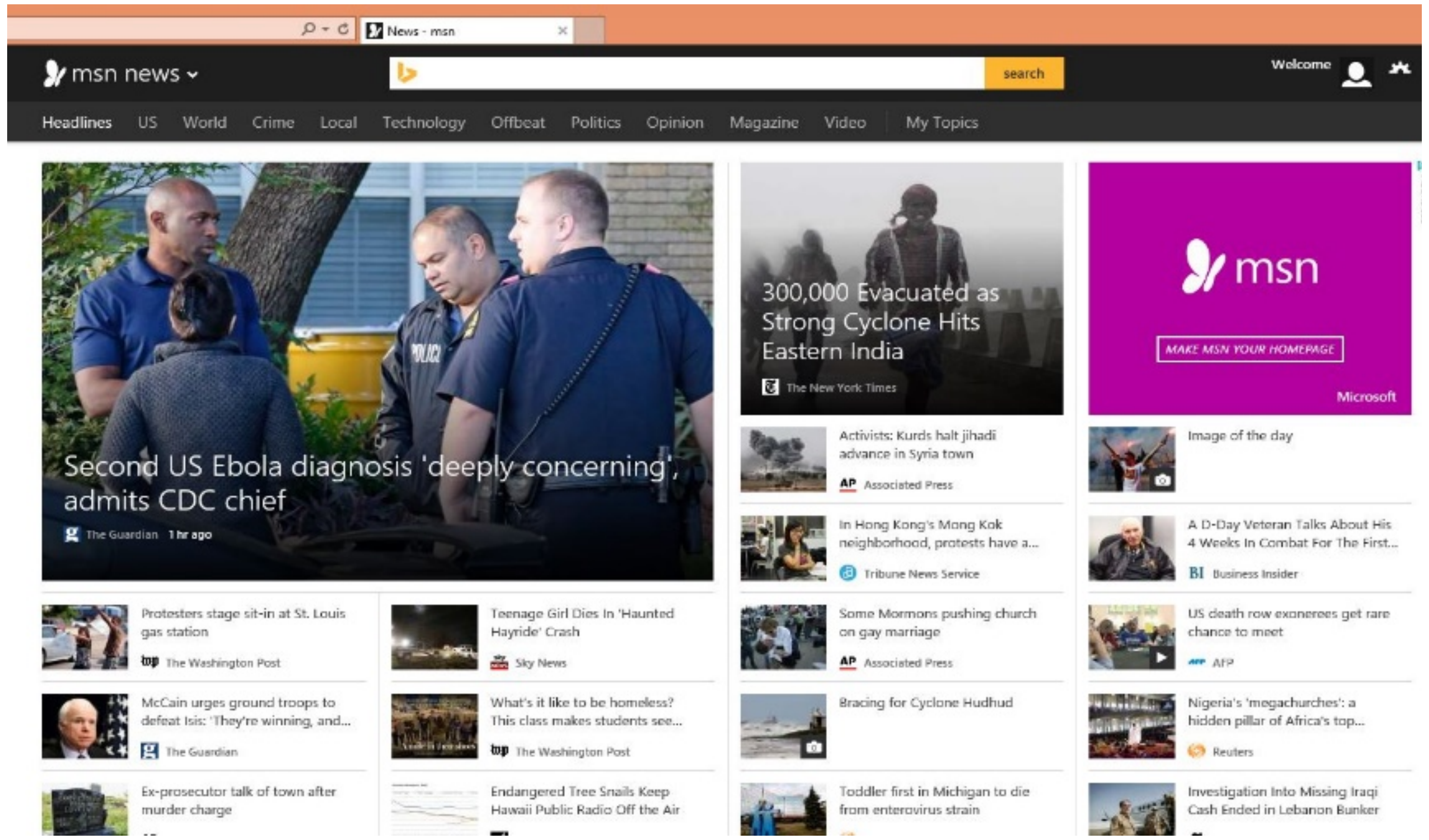
TD GAMMON [Tesauro 95]

[AlphaZero, Silver et.al, 17]

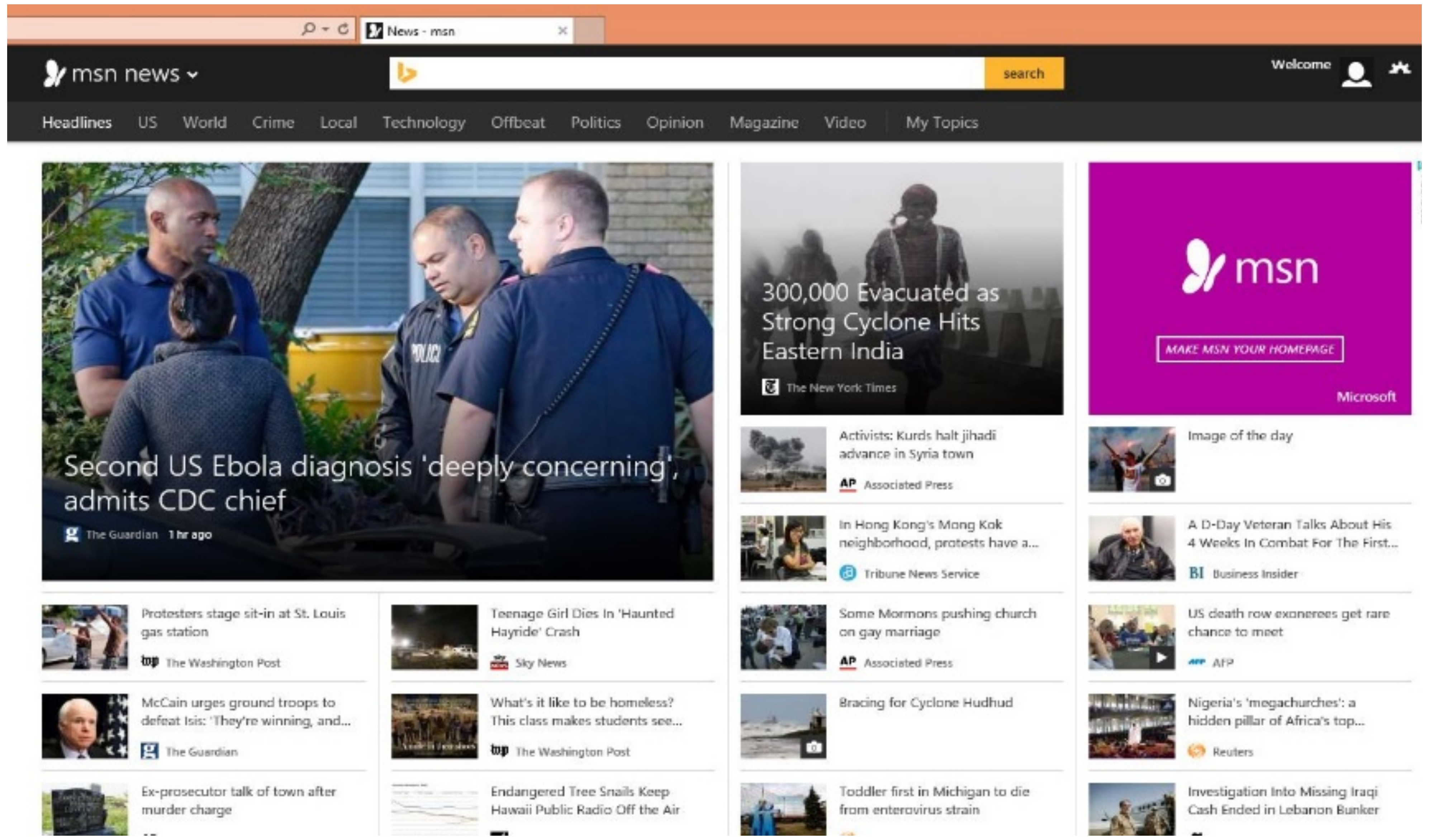
[OpenAI Five, 18]

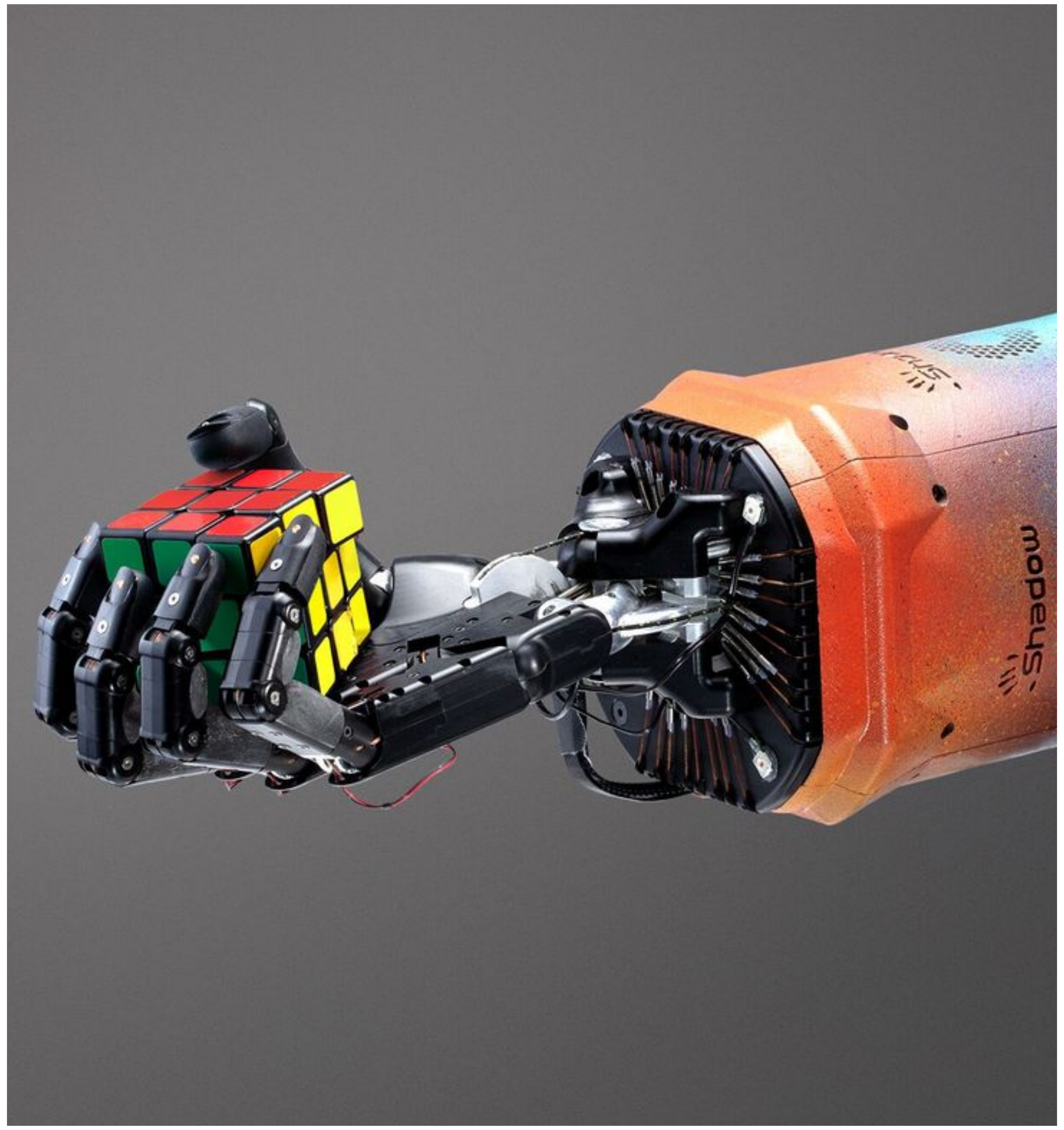
RL in Real World:

RL in Real World:

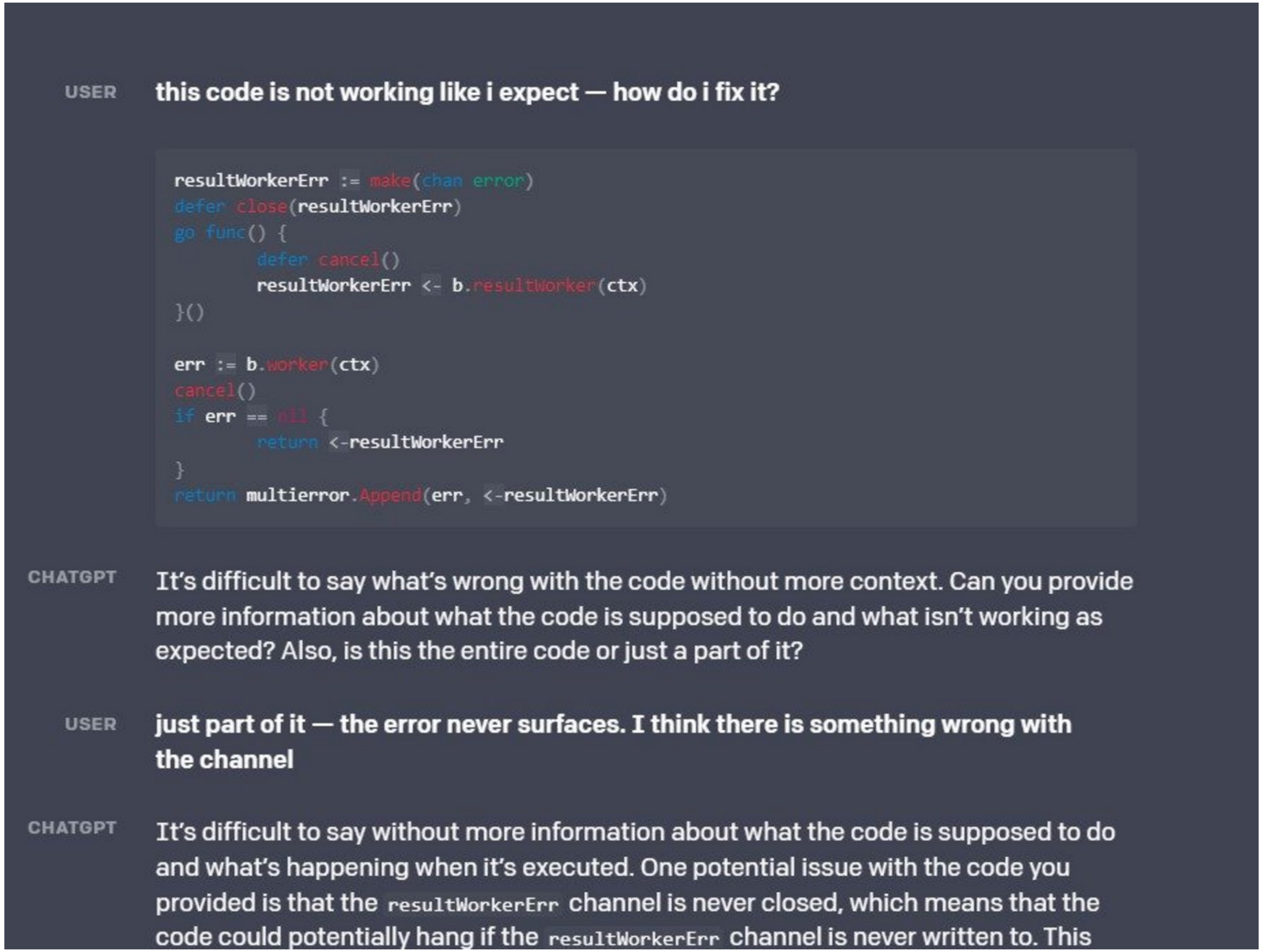


RL in Real World:





RL in Real World:



Training Language models using RL, e.g., chatGPT

RL in Real World:

Genearting creative images that would never appeared in real world

This course mostly focuses on RL Theory

When and Why RL works!
(Convergence, sample / computation complexity, etc)

Four main themes we will cover in this course:

1. Fundamentals (MDPs and Optimal planning)
2. Exploration (sample complexity)
3. Policy Gradient (global convergence)
4. Learning from human feedback

Logistics

Four (HW0-HW3) assignments (total 55%), Course Project (40%), Reading (5%)

(HW0 10%, HW1-3 15% each)

HW0 is out today and due in one week

Prerequisites (HW0)

Deep understanding of Machine Learning, Optimization, Statistics

ML: sample complexity analysis for supervised learning (PAC)

Opt: Convex (linear) optimization, e.g., gradient decent for convex functions

Stats: basics of concentration (e.g., Hoeffding’s), tricks such as union bound

Prerequisites (HW0)

Deep understanding of Machine Learning, Optimization, Statistics

ML: sample complexity analysis for supervised learning (PAC)

Opt: Convex (linear) optimization, e.g., gradient decent for convex functions

Stats: basics of concentration (e.g., Hoeffding’s), tricks such as union bound

Undergrad & MEng students: I need to see your HW0 performance

Course projects (40%)

- Team work: size 3
- Midterm report (5%), Final presentation (15%), and Final report (20%)

- Basics: **survey** of a set of similar RL theory papers. Reproduce analysis and provide a coherent story
- Advanced: **identify** extensions of existing RL papers, **formulate** theory questions, and **provide** proofs

Course Notes:
Reinforcement Learning Theory & Algorithms

- Book website: https://rltheorybook.github.io/

- Many lectures will correspond to chapters in Version 3.
- Reading assignment (5%) is from this book and additional papers

- Please let us know if you find typos/errors in the book!
  We appreciate it!

Outline

1. Definition of infinite horizon discounted MDPs

2. Bellman Optimality

3. State-action distribution

Supervised Learning

Supervised Learning

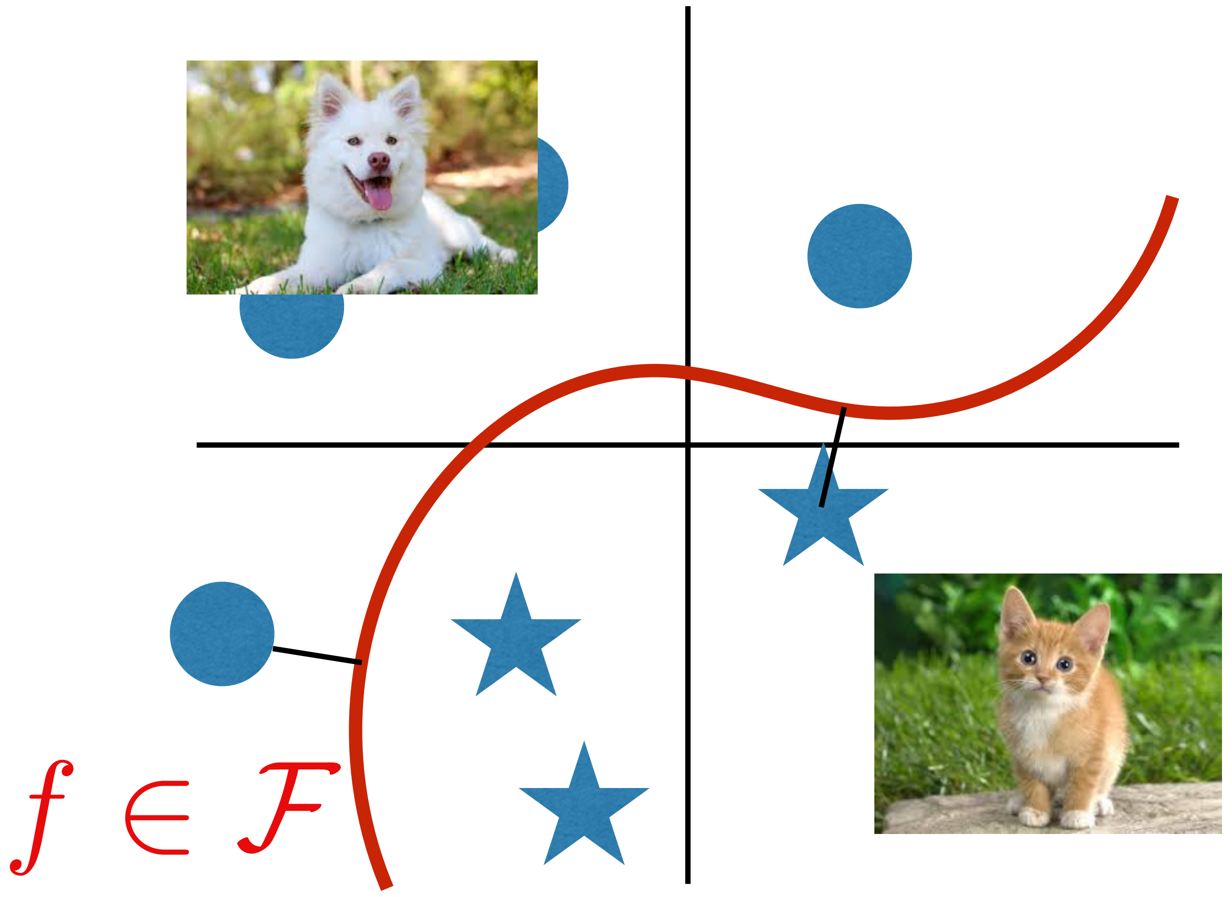
Given i.i.d examples at training:

( ,cat) ( ,cat) ( ,dog)

Supervised Learning

Given i.i.d examples at training:

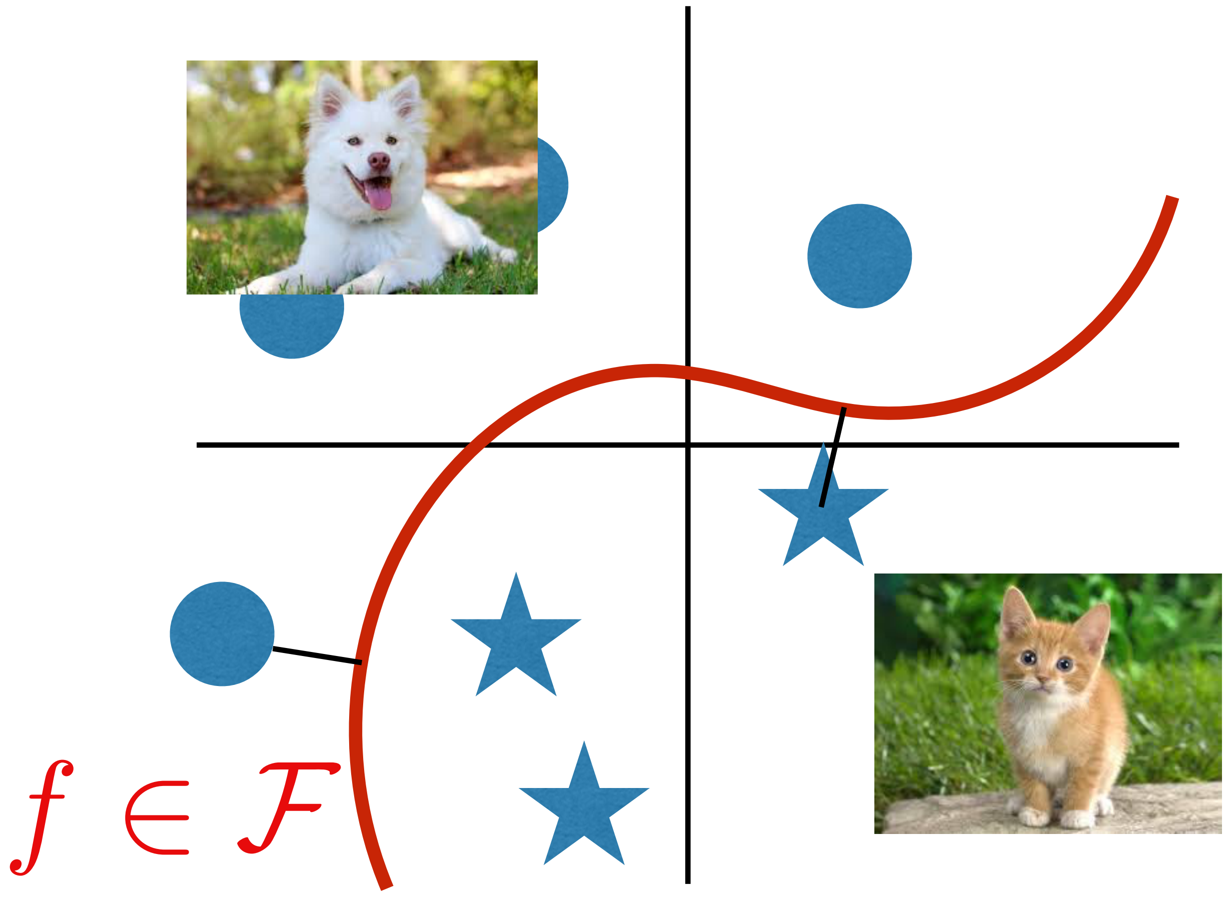
( ,cat) ( ,cat) ( ,dog)



Supervised Learning

Given i.i.d examples at training:

( ,cat) ( ,cat) ( ,dog)



Passive:

Prediction

Data Distribution

AgentLinear

Selected Actions:

RIGHT SPEED

Active: Decisions → Data Distribution

AgentLinear

Selected Actions:

RIGHT SPEED

Active: Decisions → Data Distribution

AgentLinear

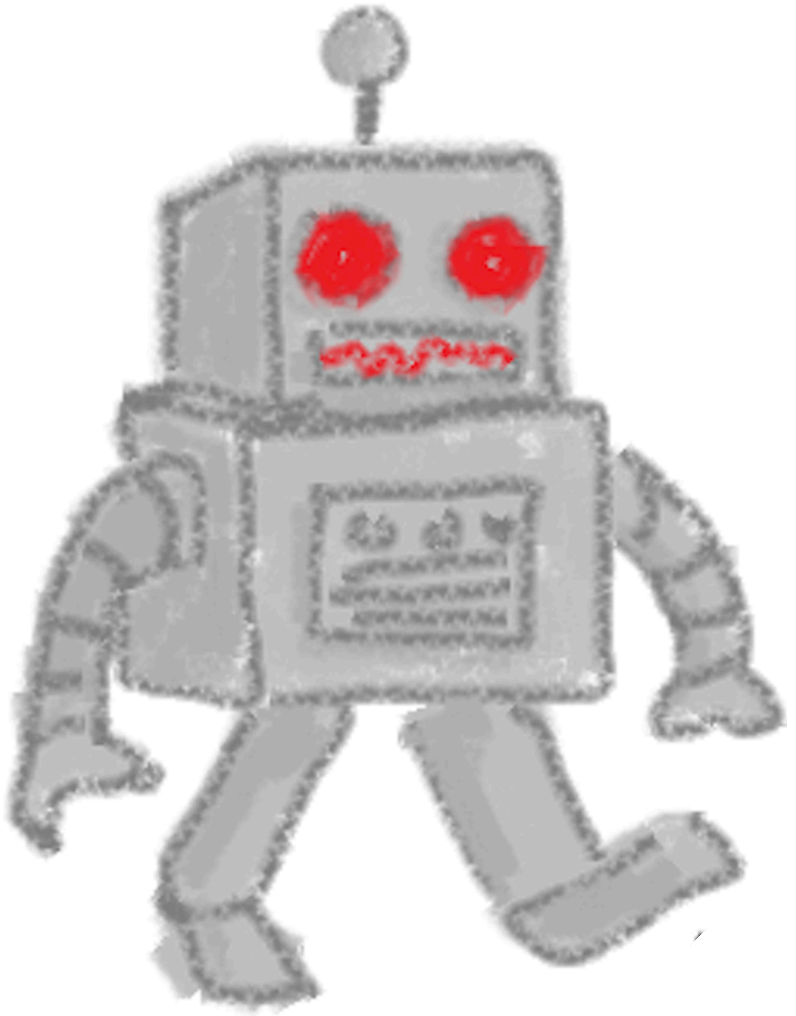
Selected Actions:

RIGHT SPEED

Active: Decisions → Data Distribution

Markov Decision Process

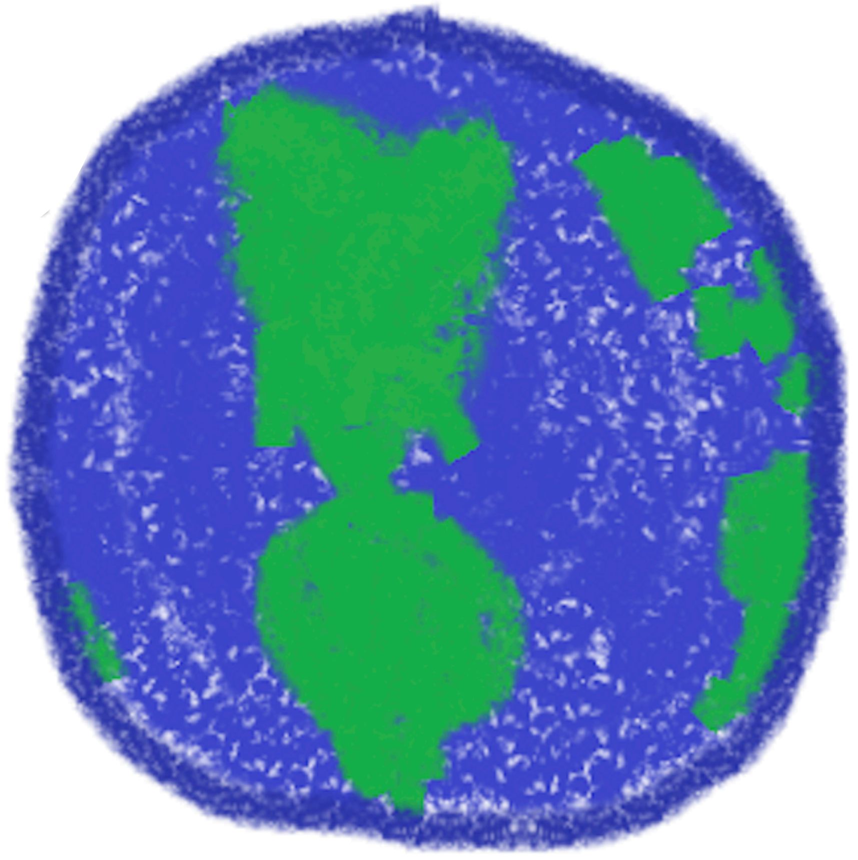
Learning Agent



$$a \sim \pi(s)$$

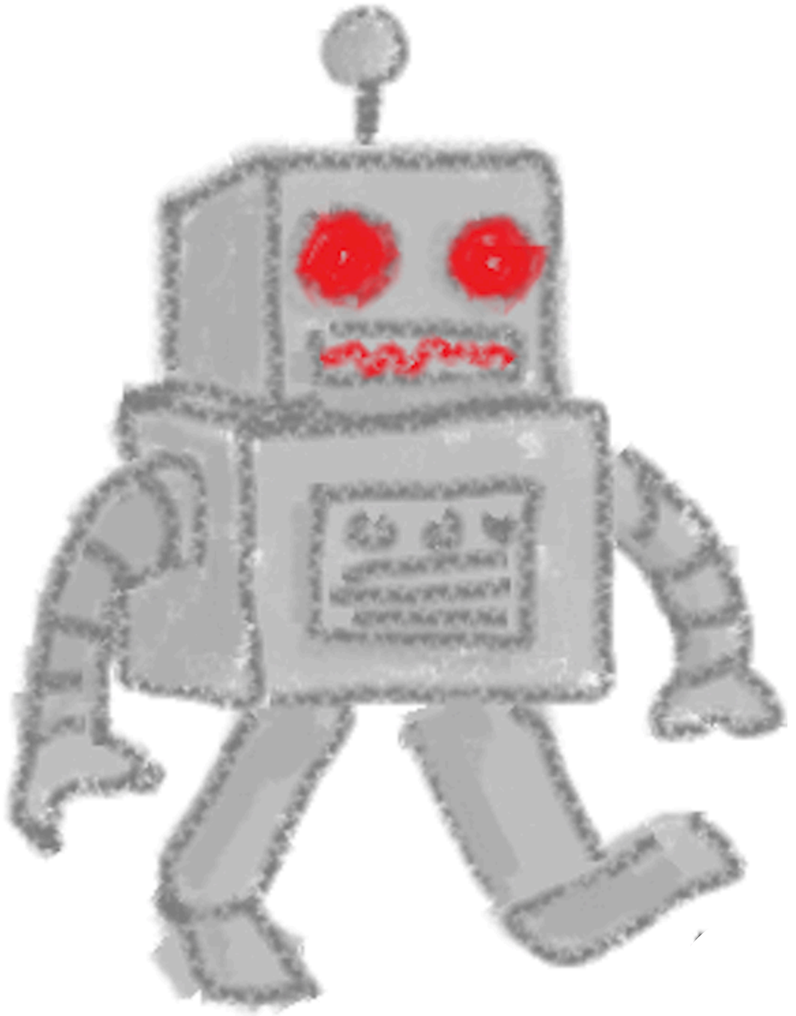
Policy: determine **action** based on **state**

Environment



Markov Decision Process

Learning Agent



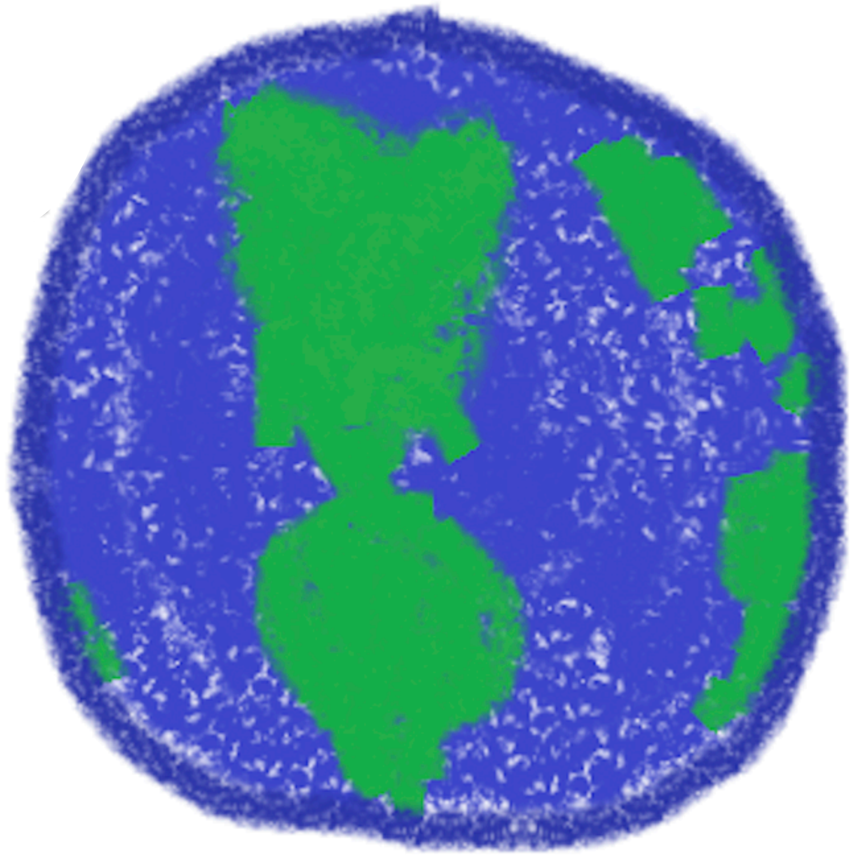
$$a \sim \pi(s)$$

Policy: determine **action** based on **state**

Send **reward** and **next state** from a Markovian transition dynamics

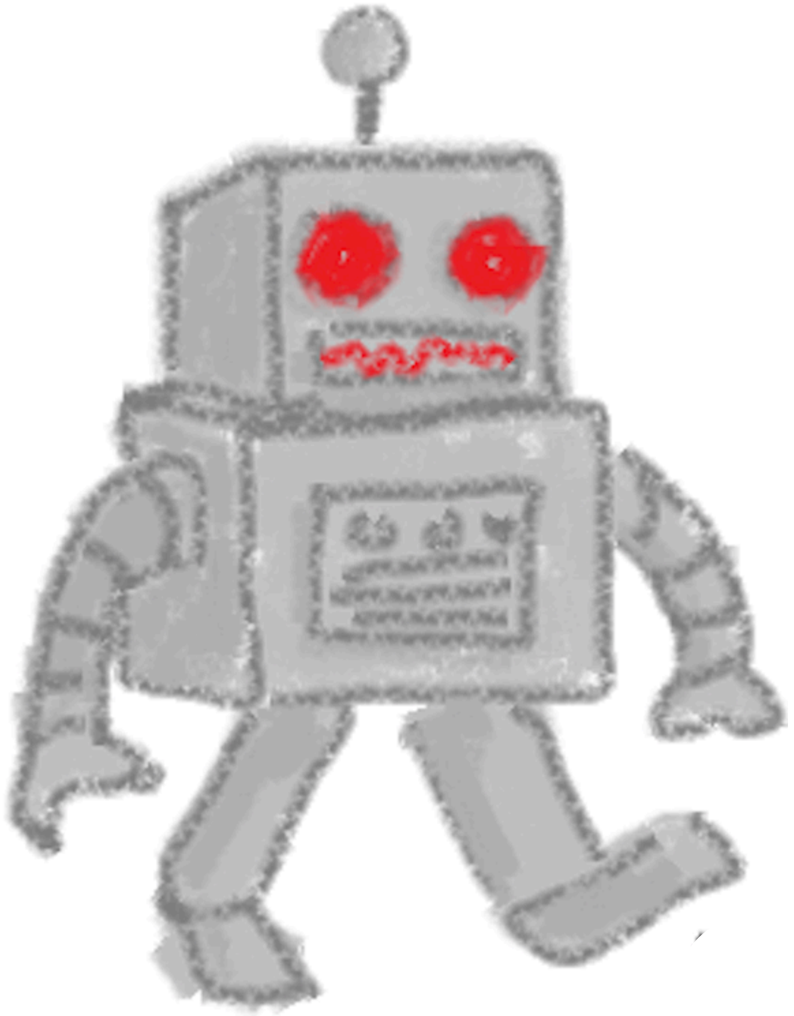
$$r(s, a), s' \sim P(\cdot \,|\, s, a)$$

Environment



Markov Decision Process

Learning Agent



$$a \sim \pi(s)$$

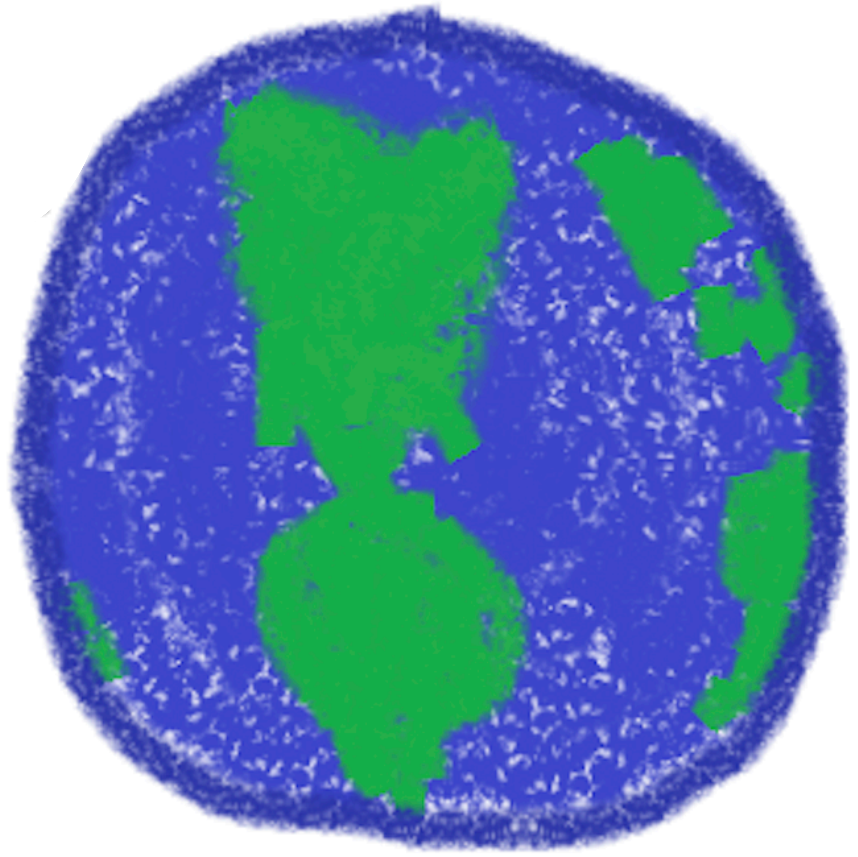
Policy: determine **action** based on **state**

Multiple Steps

Send **reward** and **next state** from a Markovian transition dynamics

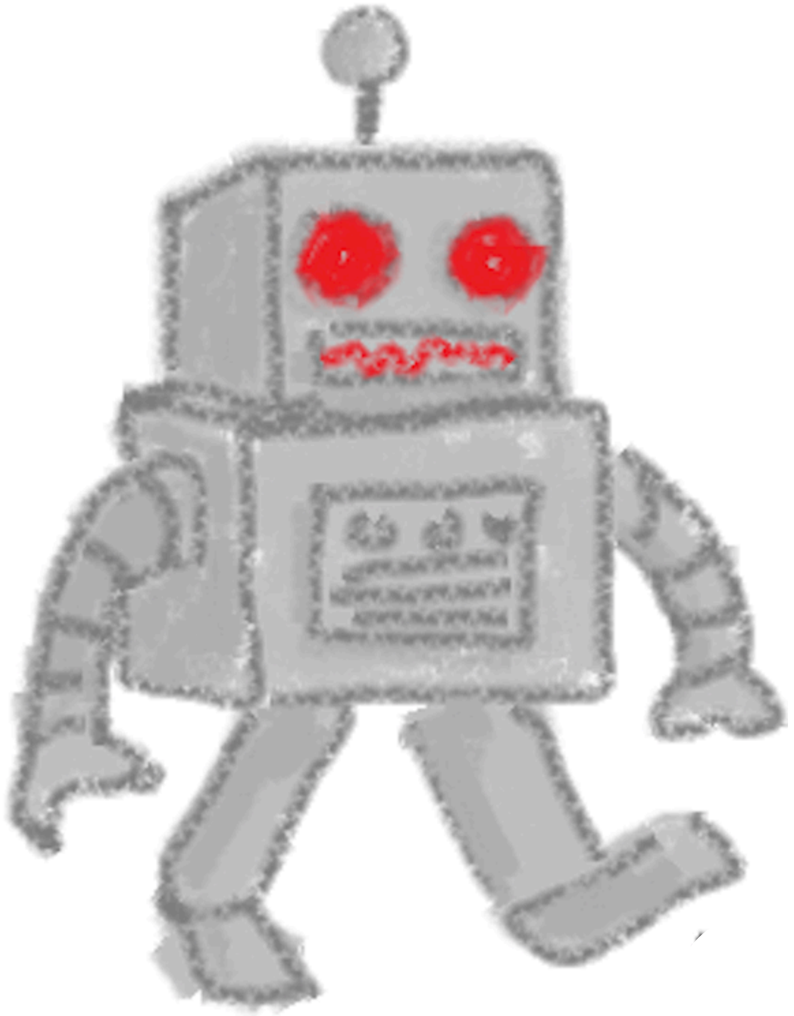
$$r(s,a), s' \sim P(\cdot \mid s,a)$$

Environment



Markov Decision Process

Learning Agent



$$a \sim \pi(s)$$

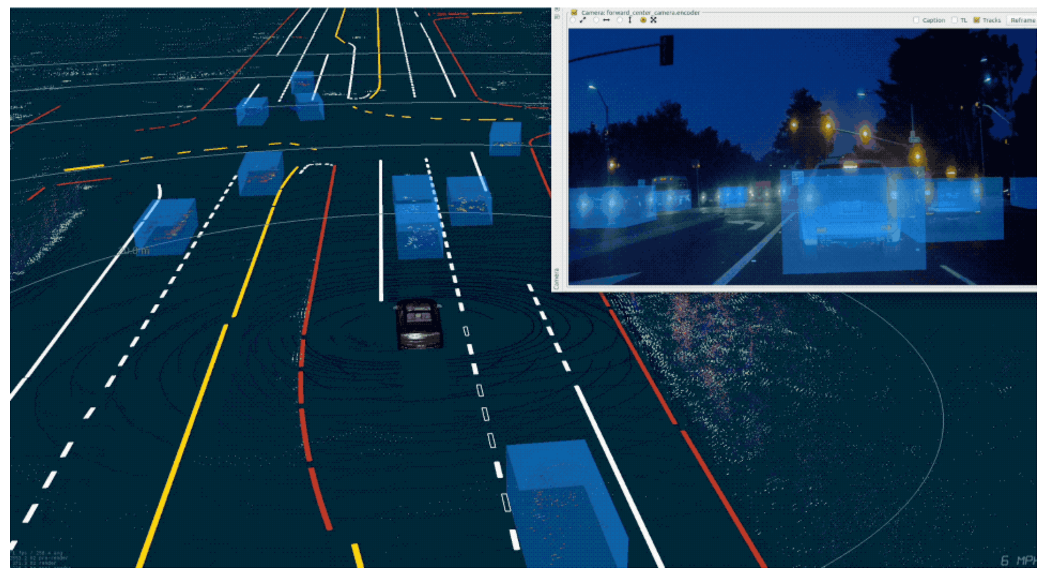
Policy: determine **action** based on **state**

Multiple Steps

Send **reward** and **next state** from a Markovian transition dynamics

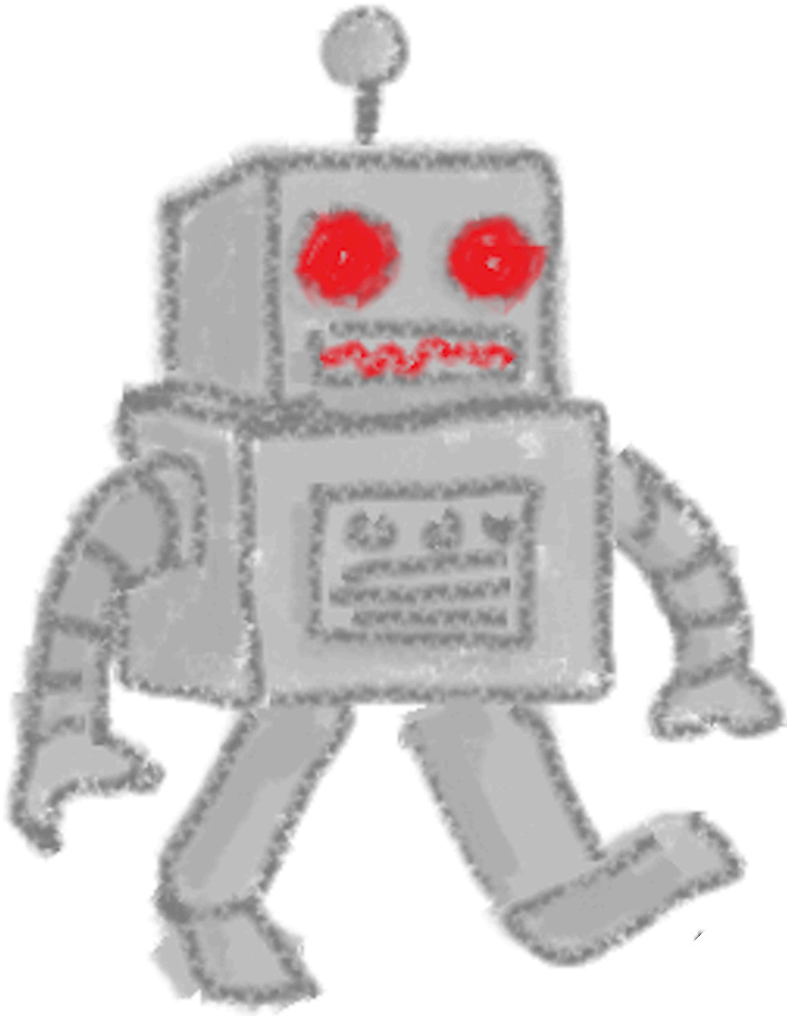
$$r(s,a), s' \sim P(\cdot \mid s,a)$$

Environment



Markov Decision Process

Learning Agent



Environment

$$a \sim \pi(s)$$

Policy: determine **action** based on **state**

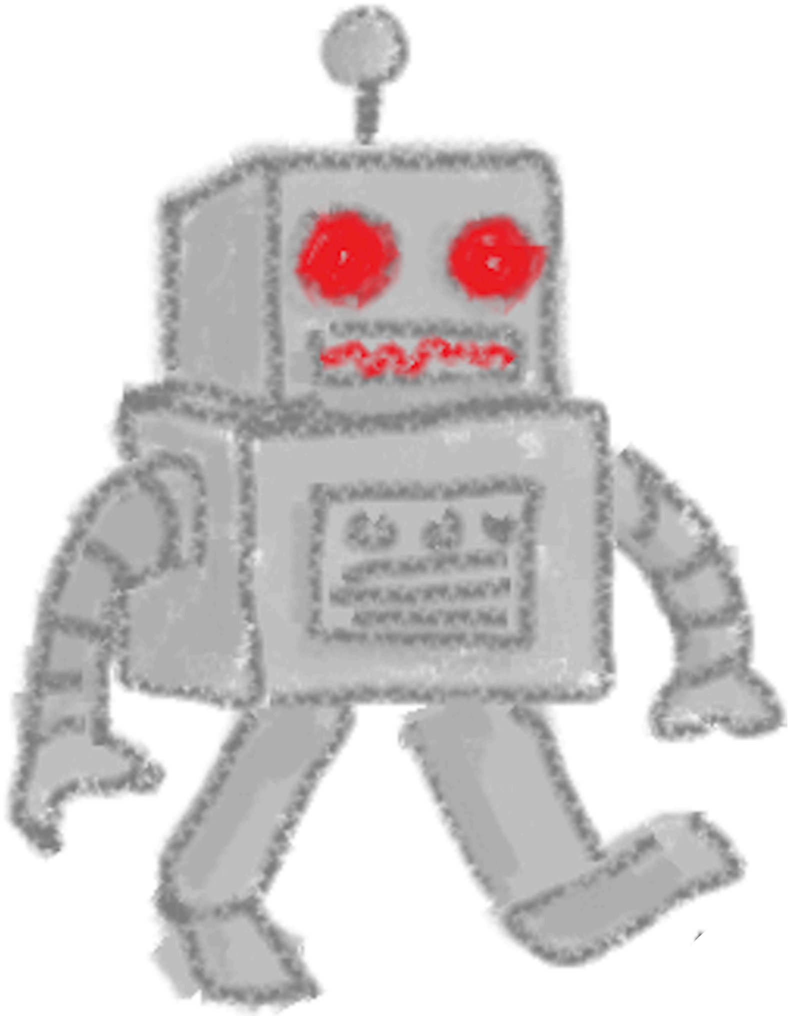
Multiple Steps

Send **reward** and **next state** from a Markovian transition dynamics

$$r(s,a), s' \sim P(\cdot \mid s,a)$$

Markov Decision Process

Learning Agent



$$a \sim \pi(s)$$

Policy: determine **action** based on **state**

Multiple Steps

Send **reward** and **next state** from a Markovian transition dynamics

Environment

$$r(s, a), s' \sim P(\cdot \,|\, s, a)$$

$$s_0 \sim \mu_0, a_0 \sim \pi(s_0), r_0, s_1 \sim P(s_0, a_0), a_1 \sim \pi(s_1), r_1 \ldots$$

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning					
Reinforcement Learning					

Table content based on slides from Emma Brunskill

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning	✔				
Reinforcement Learning	✔				

Table content based on slides from Emma Brunskill

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning	✔	✔			
Reinforcement Learning	✔	✔			

Table content based on slides from Emma Brunskill

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning	✓	✓			
Reinforcement Learning	✓	✓	✓		

Table content based on slides from Emma Brunskill

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning	✓	✓			
Reinforcement Learning	✓	✓	✓	✓	

Table content based on slides from Emma Brunskill

	Learn from Experience	Generalize	Interactive	Exploration	Credit assignment
Supervised Learning	✓	✓			
Reinforcement Learning	✓	✓	✓	✓	✓

Table content based on slides from Emma Brunskill

Infinite horizon Discounted MDP

$$\mathcal{M} = \{S, A, P, r, \mu_0, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Infinite horizon Discounted MDP

$$\mathcal{M} = \{S, A, P, r, \mu_0, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto \Delta(A)$

Infinite horizon Discounted MDP

$$\mathscr{M} = \{S, A, P, r, \mu_0, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto \Delta(A)$

Value function $V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$

Infinite horizon Discounted MDP

$$\mathscr{M} = \{S, A, P, r, \mu_0, \gamma\}$$

$$P : S \times A \mapsto \Delta(S), \quad r : S \times A \to [0,1], \quad \gamma \in [0,1)$$

Policy $\pi : S \mapsto \Delta(A)$

Value function $V^{\pi}(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \middle| s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$

Q function $Q^{\pi}(s, a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \middle| (s_0, a_0) = (s, a), a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$

Bellman Equation:

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\,\cdot \mid s_h, a_h)\right]$$

Bellman Equation:

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$$

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(s)}\left[r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} V^\pi(s')\right]$$

Bellman Equation:

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$$

$$V^\pi(s) = \mathbb{E}_{a \sim \pi(s)}\left[r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\pi(s')\right]$$

$$Q^\pi(s, a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, (s_0, a_0) = (s, a), a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \mid s_h, a_h)\right]$$

Bellman Equation:

$$V^\pi(s) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, s_0 = s, a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \,|\, s_h, a_h)\right]$$

$$V^\pi(s) = \mathbb{E}_{a\sim\pi(s)}\left[r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} V^\pi(s')\right]$$

$$Q^\pi(s,a) = \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^h r(s_h, a_h) \,\middle|\, (s_0, a_0) = (s,a), a_h \sim \pi(s_h), s_{h+1} \sim P(\cdot \,|\, s_h, a_h)\right]$$

$$Q^\pi(s,a) = r(s,a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} V^\pi(s')$$

Outline

1. Definition of infinite horizon discounted MDPs

2. Bellman Optimality

3. State-action distribution

Optimal Policy

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

$$\pi^\star : S \mapsto A, \text{ s.t., } V^{\pi^\star}(s) \geq V^{\pi}(s), \forall s, \pi$$

[Puterman 94 chapter 6, also see theorem 1.4 in the RL monograph]

Optimal Policy

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

$$\pi^\star : S \mapsto A, \text{ s.t., } V^{\pi^\star}(s) \geq V^\pi(s), \forall s, \pi$$

[Puterman 94 chapter 6, also see theorem 1.4 in the RL monograph]

We denote $V^\star := V^{\pi^\star}, Q^\star := Q^{\pi^\star}$

Optimal Policy

For infinite horizon discounted MDP, there exists a deterministic stationary policy
$\pi^\star : S \mapsto A$, s.t., $V^{\pi^\star}(s) \geq V^\pi(s), \forall s, \pi$
[Puterman 94 chapter 6, also see theorem 1.4 in the RL monograph]

We denote $V^\star := V^{\pi^\star}, Q^\star := Q^{\pi^\star}$

Theorem 1: Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] = r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} V^\star(s')$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] = r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} V^\star(s')$$

$$= r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} \left[r(s', \pi^\star(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \pi^\star(s'))} V^\star(s'') \right]$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] = r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} V^\star(s')$$

$$= r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} \left[r(s', \pi^\star(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \pi^\star(s'))} V^\star(s'') \right]$$

$$\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \widehat{\pi}(s'))} V^\star(s'') \right]$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] = r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} V^\star(s')$$

$$= r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} \left[r(s', \pi^\star(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \pi^\star(s'))} V^\star(s'') \right]$$

$$\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \widehat{\pi}(s'))} V^\star(s'') \right]$$

$$\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \widehat{\pi}(s'))} \left[r(s'', \widehat{\pi}(s'')) + \gamma \mathbb{E}_{s''' \sim P(s'', \widehat{\pi}(s''))} V^\star(s''') \right] \right]$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we will prove $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) = r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

$$\leq \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s') \right] = r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} V^\star(s')$$

$$= r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} \left[r(s', \pi^\star(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \pi^\star(s'))} V^\star(s'') \right]$$

$$\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \widehat{\pi}(s'))} V^\star(s'') \right]$$

$$\leq r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} \left[r(s', \widehat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \widehat{\pi}(s'))} \left[r(s'', \widehat{\pi}(s'')) + \gamma \mathbb{E}_{s''' \sim P(s'', \widehat{\pi}(s''))} V^\star(s''') \right] \right]$$

$$\leq \mathbb{E} \left[r(s, \widehat{\pi}(s)) + \gamma r(s', \widehat{\pi}(s')) + \ldots \right] = V^{\widehat{\pi}}(s)$$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we just proved $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

Denote $\widehat{\pi}(s) := \arg\max_a Q^\star(s,a)$, we just proved $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

This implies that $\arg\max_a Q^\star(s,a)$ is an optimal policy

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$|V(s) - V^\star(s)| = \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$| V(s) - V^\star(s) | = \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

$$\leq \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$|V(s) - V^\star(s)| = \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

$$\leq \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

$$\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left| V(s') - V^\star(s') \right|$$

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$|V(s) - V^\star(s)| = \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

$$\leq \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

$$\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left| V(s') - V^\star(s') \right|$$

$$\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left(\max_{a'} \gamma \mathbb{E}_{s'' \sim P(s',a')} \left| V(s'') - V^\star(s'') \right| \right)$$

Proof of Bellman Optimality

Theorem 2:

For any $V : S \to \mathbb{R}$, if $V(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V(s') \right]$ for all s,

then $V(s) = V^\star(s), \forall s$

$$|V(s) - V^\star(s)| = \left| \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - \max_a (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

$$\leq \max_a \left| (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V(s')) - (r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^\star(s')) \right|$$

$$\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left| V(s') - V^\star(s') \right|$$

$$\leq \max_a \gamma \mathbb{E}_{s' \sim P(s,a)} \left(\max_{a'} \gamma \mathbb{E}_{s'' \sim P(s',a')} \left| V(s'') - V^\star(s'') \right| \right)$$

$$\leq \max_{a_1, a_2, \ldots a_{k-1}} \gamma^k \mathbb{E}_{s_k} |V(s_k) - V^\star(s_k)|$$

Outline

1. Definition of infinite horizon discounted MDPs

2. Bellman Optimality

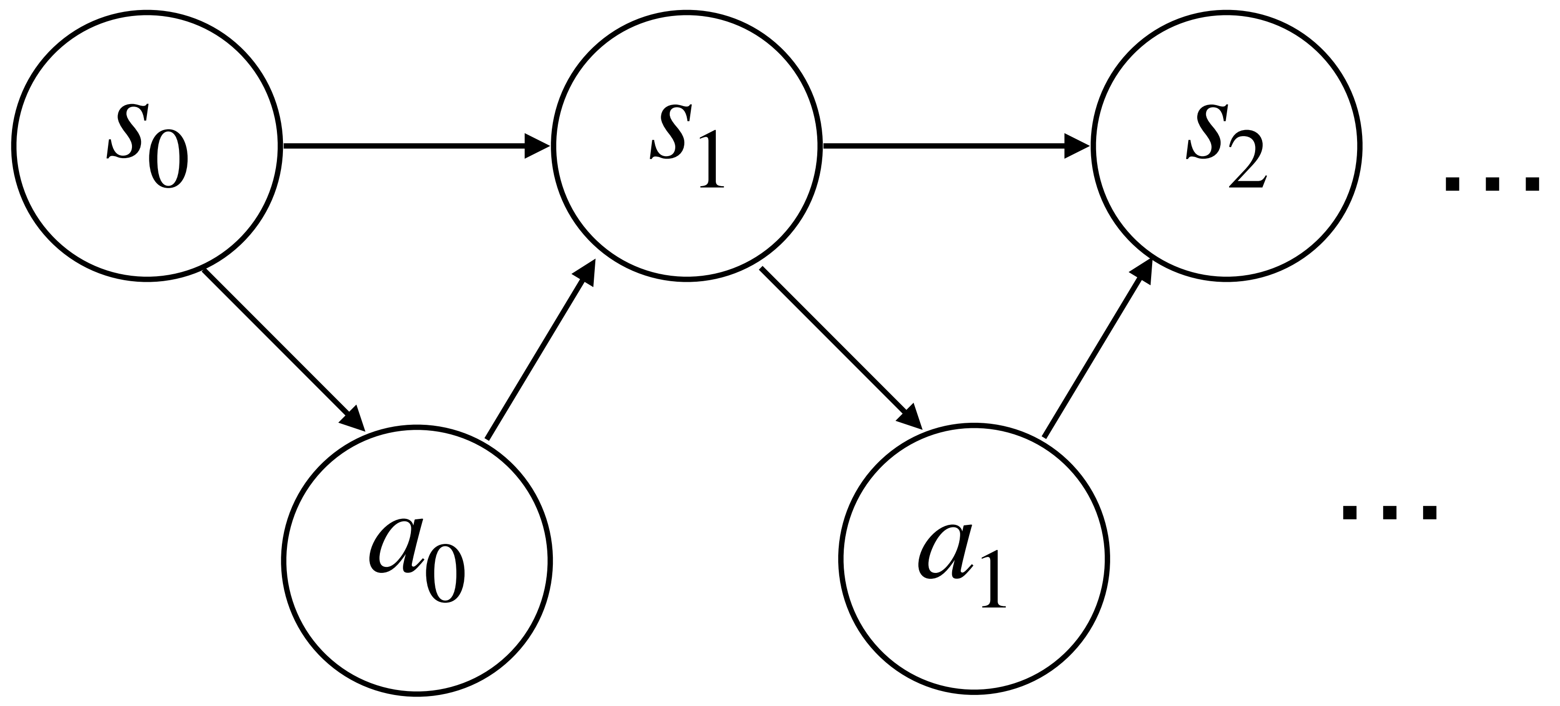
3. State-action distribution

Trajectory distribution and state-action distribution

Q: what is the probability of π generating trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_h, a_h\}$?

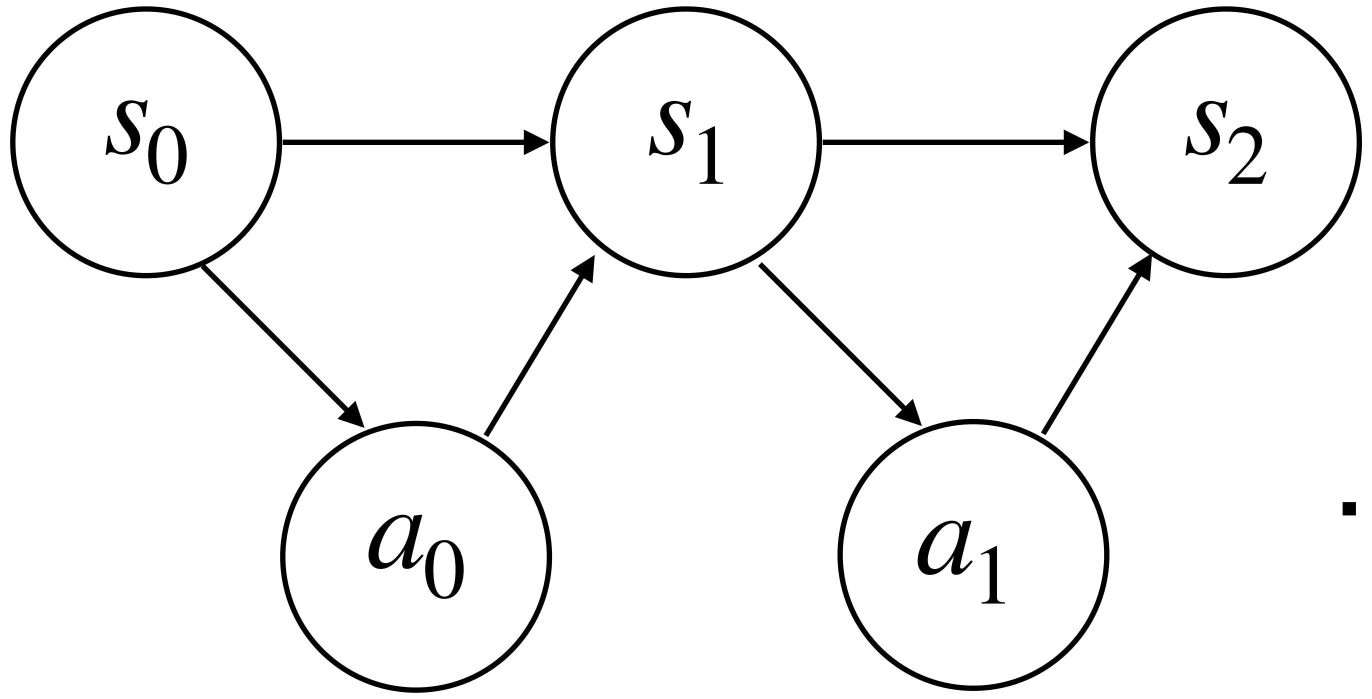
Trajectory distribution and state-action distribution

Q: what is the probability of π generating trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_h, a_h\}$?



Trajectory distribution and state-action distribution

Q: what is the probability of π generating trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_h, a_h\}$?

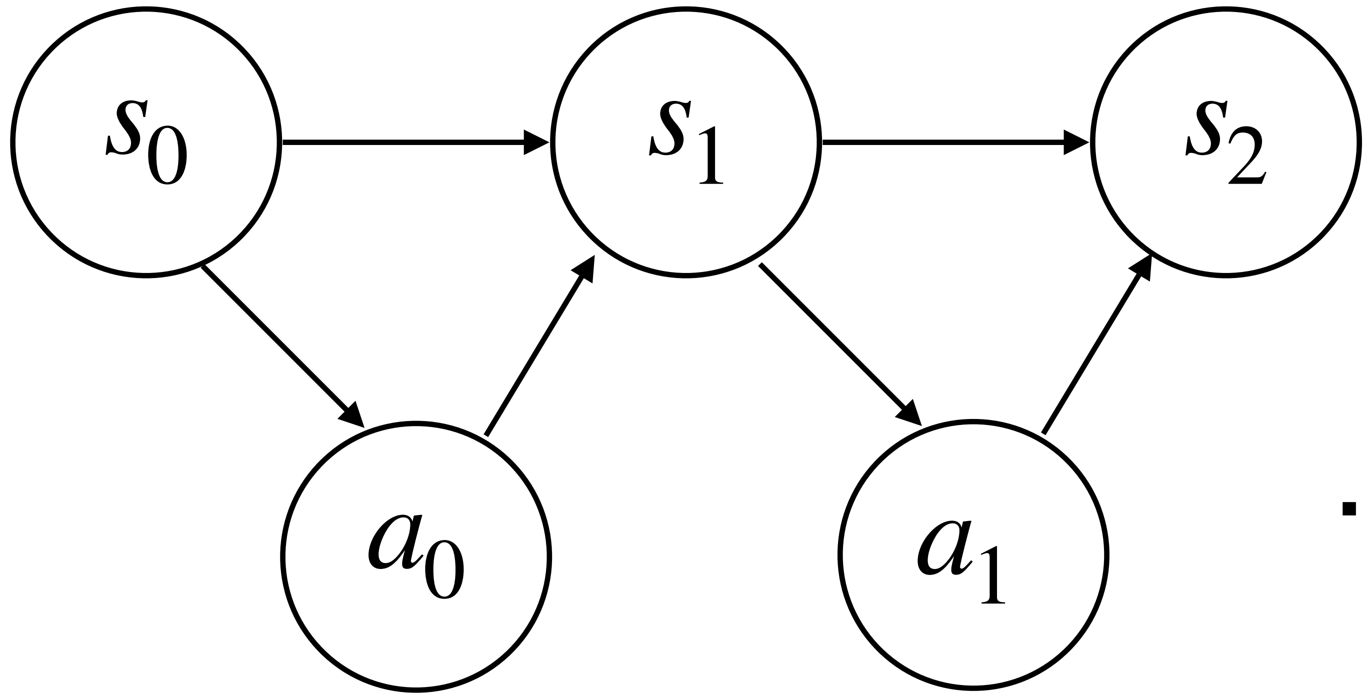


… $\mathbb{P}^\pi(s_0, a_0, \ldots, s_h, a_h)$

… $= \mu(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)P(s_2 \mid s_1, a_1)\ldots P(s_h \mid s_{h-1}, a_{h-1})\pi(a_h \mid s_h)$

Trajectory distribution and state-action distribution

Q: what is the probability of π generating trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_h, a_h\}$?



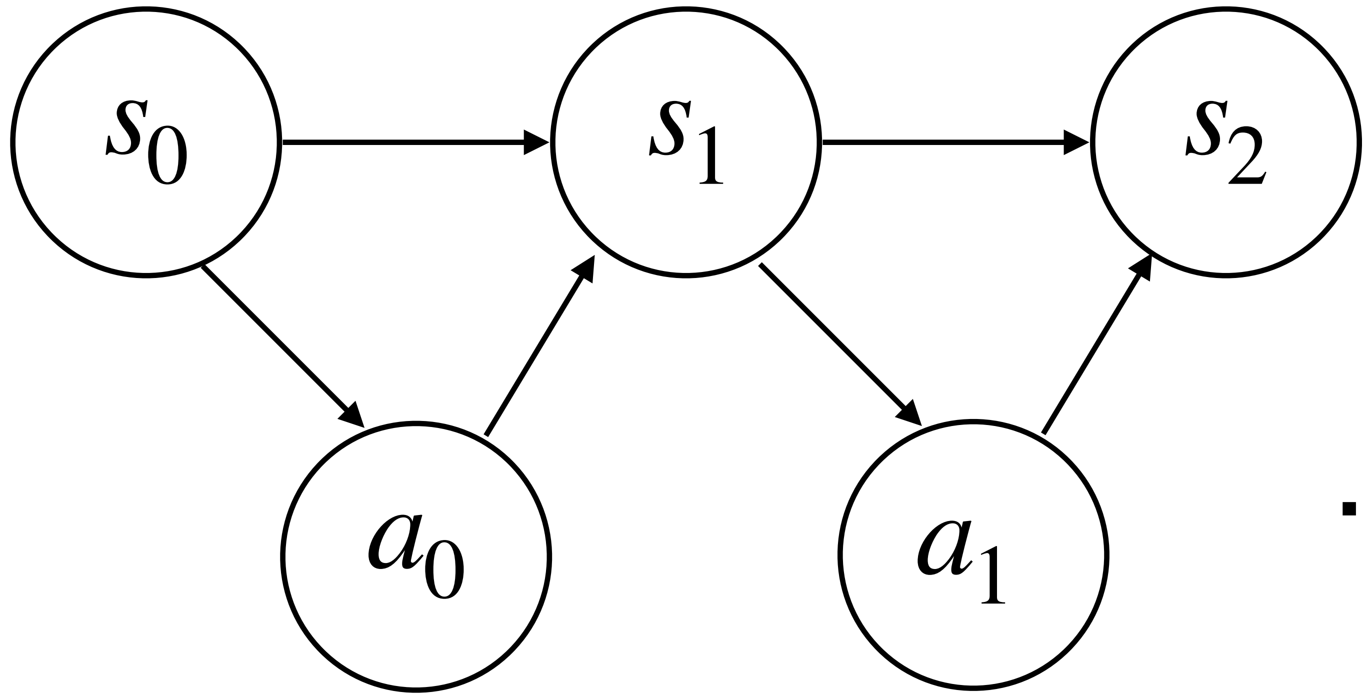
… $\mathbb{P}^\pi(s_0, a_0, \ldots, s_h, a_h)$

… $= \mu(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)P(s_2 \mid s_1, a_1)\ldots P(s_h \mid s_{h-1}, a_{h-1})\pi(a_h \mid s_h)$

Q: what's the probability of π visiting state (s,a) at time step h?

Trajectory distribution and state-action distribution

Q: what is the probability of π generating trajectory $\tau = \{s_0, a_0, s_1, a_1, \ldots, s_h, a_h\}$?



$$\ldots \quad \mathbb{P}^\pi(s_0, a_0, \ldots, s_h, a_h)$$

$$\ldots \quad = \mu(s_0)\pi(a_0 \mid s_0)P(s_1 \mid s_0, a_0)\pi(a_1 \mid s_1)P(s_2 \mid s_1, a_1)\ldots P(s_h \mid s_{h-1}, a_{h-1})\pi(a_h \mid s_h)$$

Q: what's the probability of π visiting state $(s$,a) at time step h?

$$\mathbb{P}_h^\pi(s, a) = \sum_{s_0, a_0, s_1, a_1, \ldots, s_{h-1}, a_{h-1}} \mathbb{P}^\pi(s_0, a_0, \ldots, s_{h-1}, a_{h-1}, s_h = s, a_h = a)$$

State action occupancy measure

$\mathbb{P}_h^\pi(s, a)$: probability of π visiting (s, a) at time step $h \in \mathbb{N}$

$$d^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a)$$

State action occupancy measure

$\mathbb{P}_h^\pi(s, a)$: probability of π visiting (s, a) at time step $h \in \mathbb{N}$

$$d^\pi(s, a) = (1 - \gamma) \sum_{h=0}^{\infty} \gamma^h \mathbb{P}_h^\pi(s, a)$$

$$\mathbb{E}_{s_0 \sim \mu} V^\pi(s_0) = \frac{1}{1 - \gamma} \sum_{s,a} d^\pi(s, a) r(s, a)$$

Summary for today

Key definitions: MDPs, Value / Q functions, State-action distribution

Key property: Bellman optimality (the two theorems and their proofs)