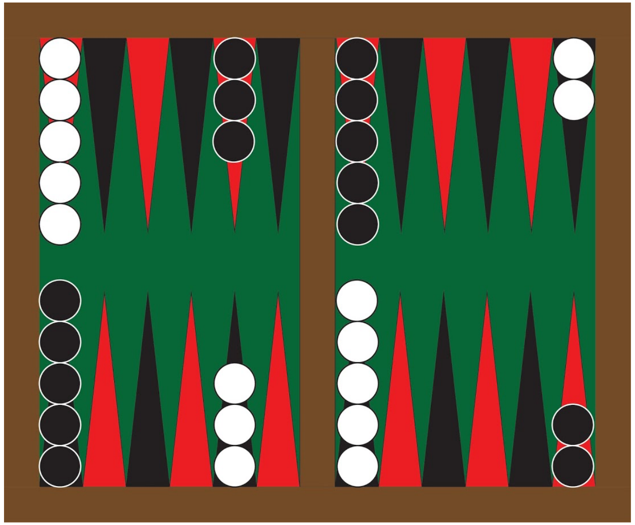
Introduction and Basics of Markov Decision Process

Wen Sun

CS 6789: Foundations of Reinforcement Learning

The very successful stories of ML are based on RL…



TD GAMMON [Tesauro 95]

nature

THE INTERNATIONAL WEEKLY JOURNAL OF SCIENCE

At last — a computer program that can beat a champion Go player PAGE 484

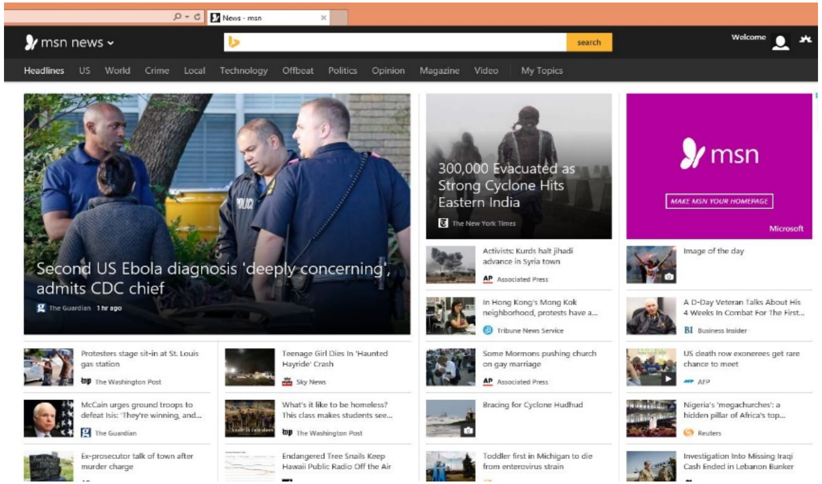
ALL SYSTEMS GO

[AlphaZero, Silver et.al, 17]

[OpenAI Five, 18]

RL in Real World:

RL in Real World:



Second US Ebola diagnosis 'deeply concerning', admits CDC chief
The Guardian 1 hr ago

300,000 Evacuated as Strong Cyclone Hits Eastern India
The New York Times

Activists: Kurds halt jihadi advance in Syria town
AP Associated Press

Image of the day

Protesters stage sit-in at St. Louis gas station
The Washington Post

Teenage Girl Dies In 'Haunted Hayride' Crash
Sky News

In Hong Kong's Mong Kok neighborhood, protests have a...
Tribune News Service

A D-Day Veteran Talks About His 4 Weeks In Combat For The First...
BI Business Insider

McCain urges ground troops to defeat Isis: 'They're winning, and...
The Guardian

What's it like to be homeless? This class makes students see...
The Washington Post

Some Mormons pushing church on gay marriage
AP Associated Press

US death row exonerees get rare chance to meet
AFP

Ex-prosecutor talk of town after murder charge

Endangered Tree Snails Keep Hawaii Public Radio Off the Air

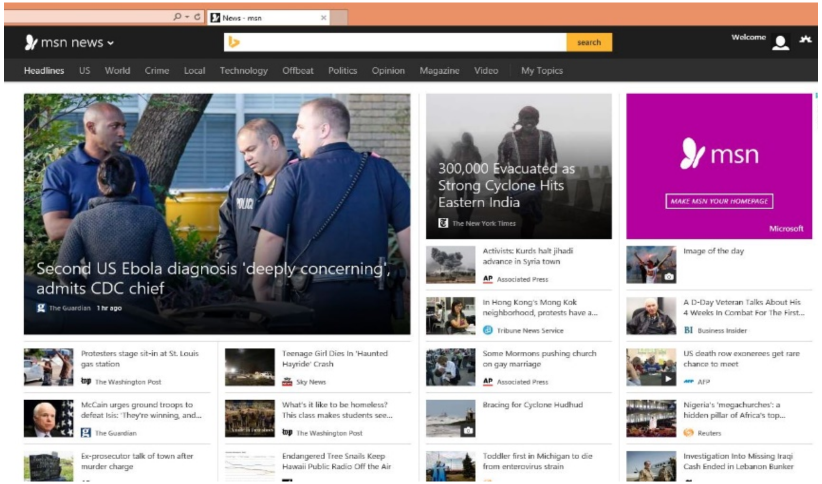
Bracing for Cyclone Hudhud

Nigeria's 'megachurches': a hidden pillar of Africa's top...
Reuters

Toddler first in Michigan to die from enterovirus strain

Investigation Into Missing Iraqi Cash Ended in Lebanon Bunker

RL in Real World:



Second US Ebola diagnosis 'deeply concerning', admits CDC chief
The Guardian 1 hr ago

300,000 Evacuated as Strong Cyclone Hits Eastern India
The New York Times

Activists: Kurds halt jihadi advance in Syria town
AP Associated Press

Image of the day

Protesters stage sit-in at St. Louis gas station
The Washington Post

Teenage Girl Dies In 'Haunted Hayride' Crash
Sky News

In Hong Kong's Mong Kok neighborhood, protests have a...
Tribune News Service

A D-Day Veteran Talks About His 4 Weeks In Combat For The First...
BI Business Insider

McCain urges ground troops to defeat Isis: 'They're winning, and...
The Guardian

What's it like to be homeless? This class makes students see...
The Washington Post

Some Mormons pushing church on gay marriage
AP Associated Press

US death row exonerees get rare chance to meet
AFP

Ex-prosecutor talk of town after murder charge

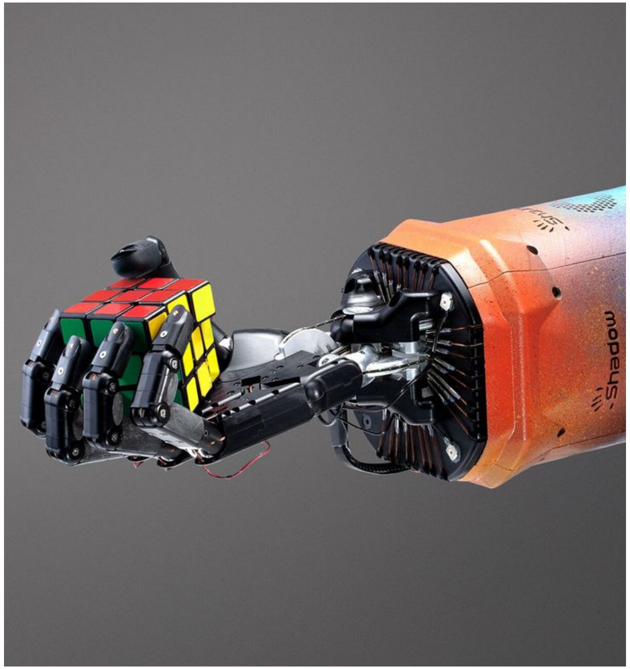
Endangered Tree Snails Keep Hawaii Public Radio Off the Air

Bracing for Cyclone Hudhud

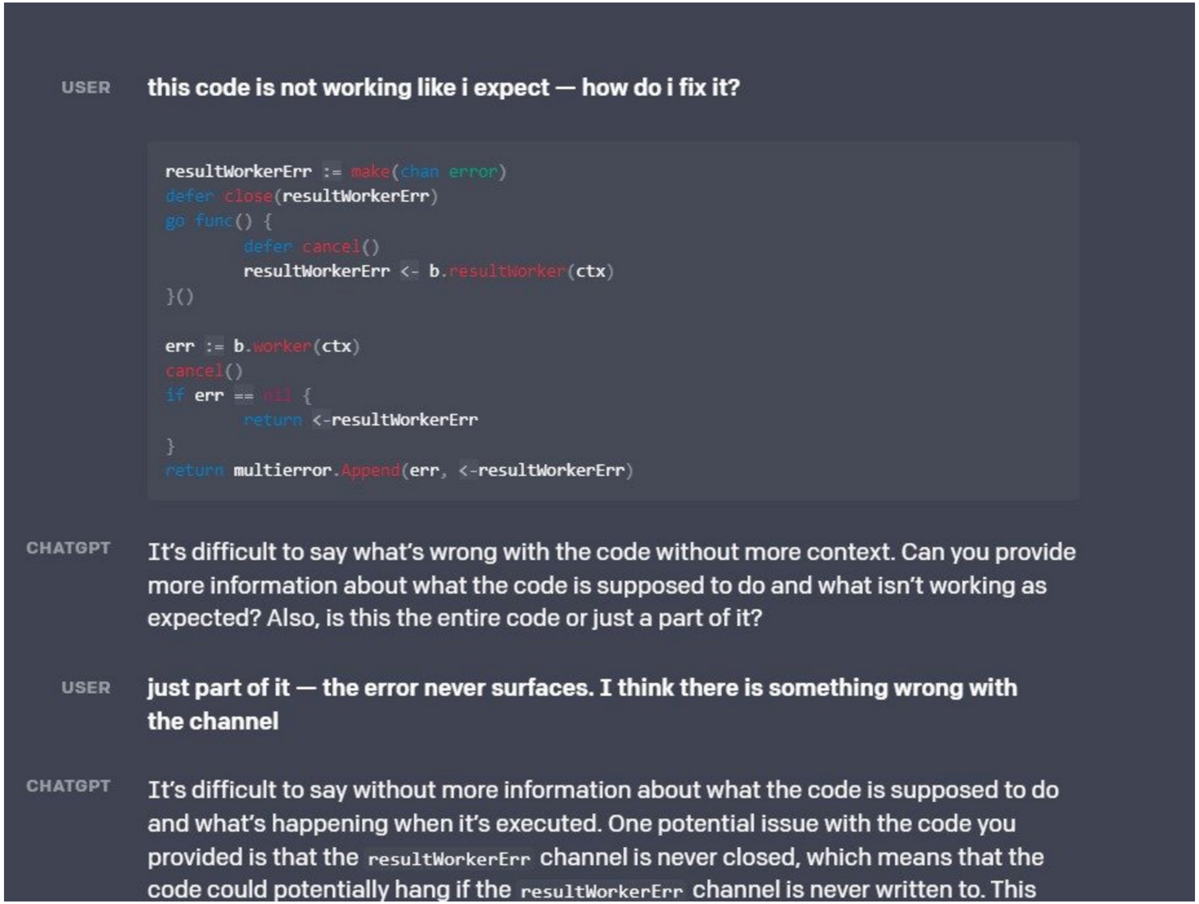
Nigeria's 'megachurches': a hidden pillar of Africa's top...
Reuters

Toddler first in Michigan to die from enterovirus strain

Investigation Into Missing Iraqi Cash Ended in Lebanon Bunker



RL in Real World:



USER this code is not working like i expect — how do i fix it?

```
resultWorkerErr := make(chan error)
defer close(resultWorkerErr)
go func() {
        defer cancel()
        resultWorkerErr <- b.resultWorker(ctx)
}()

err := b.worker(ctx)
cancel()
if err == nil {
        return <-resultWorkerErr
}
return multierror.Append(err, <-resultWorkerErr)
```

CHATGPT It's difficult to say what's wrong with the code without more context. Can you provide more information about what the code is supposed to do and what isn't working as expected? Also, is this the entire code or just a part of it?

USER just part of it — the error never surfaces. I think there is something wrong with the channel

CHATGPT It's difficult to say without more information about what the code is supposed to do and what's happening when it's executed. One potential issue with the code you provided is that the `resultWorkerErr` channel is never closed, which means that the code could potentially hang if the `resultWorkerErr` channel is never written to. This

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

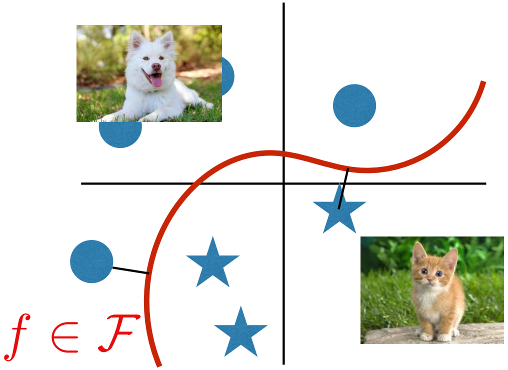
Given i.i.d examples at training:

(

,cat)(

,cat)(

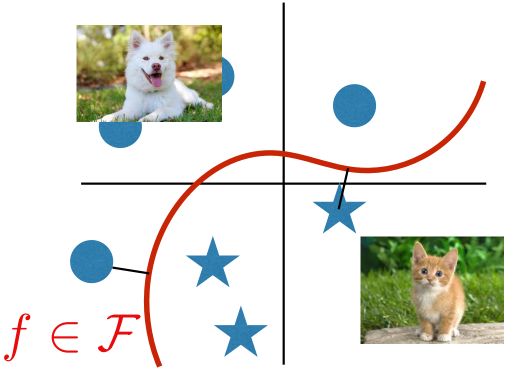
,dog)



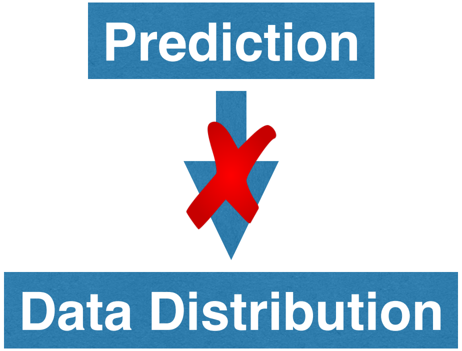
Supervised Learning

Given i.i.d examples at training:

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

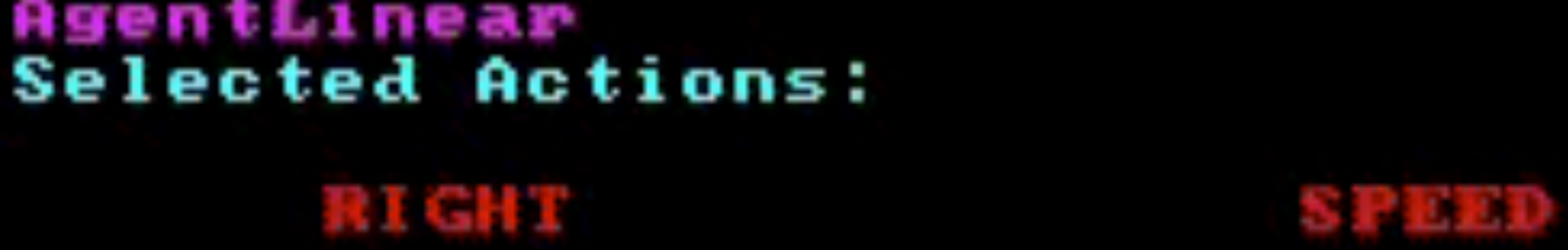


Passive:



Prediction

Data Distribution

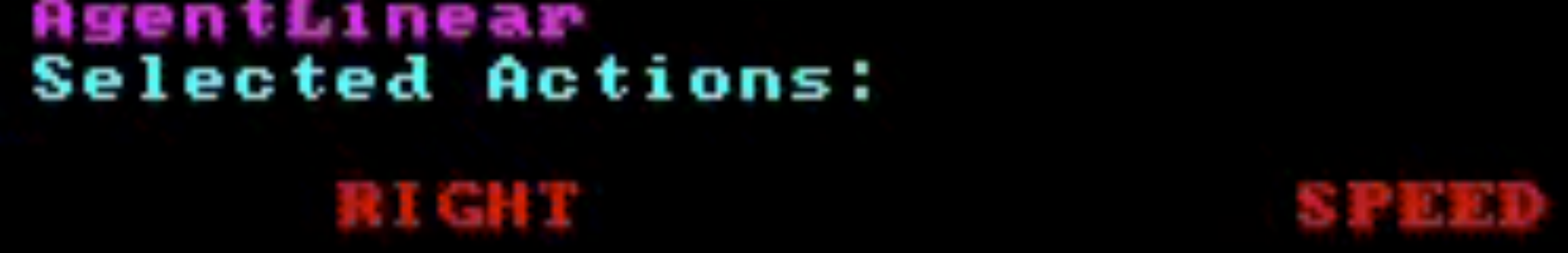


AgentLinear

Selected Actions:

RIGHT SPEED

Active: **Decisions** → **Data Distribution**

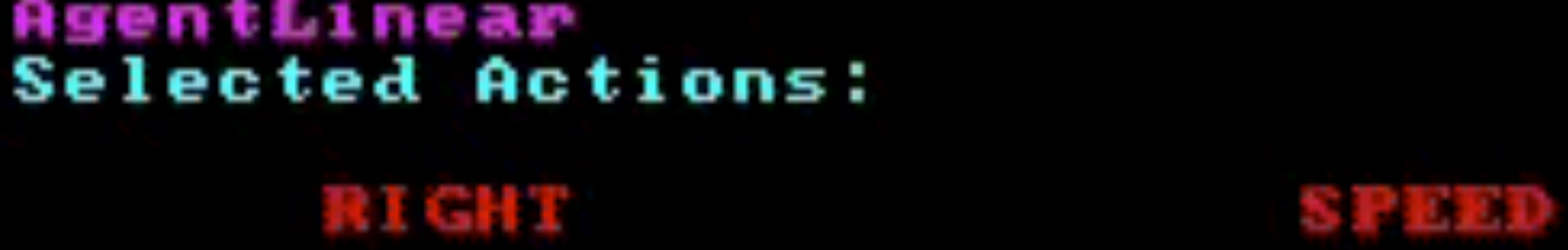


AgentLinear

Selected Actions:

RIGHT SPEED

Active: Decisions → Data Distribution



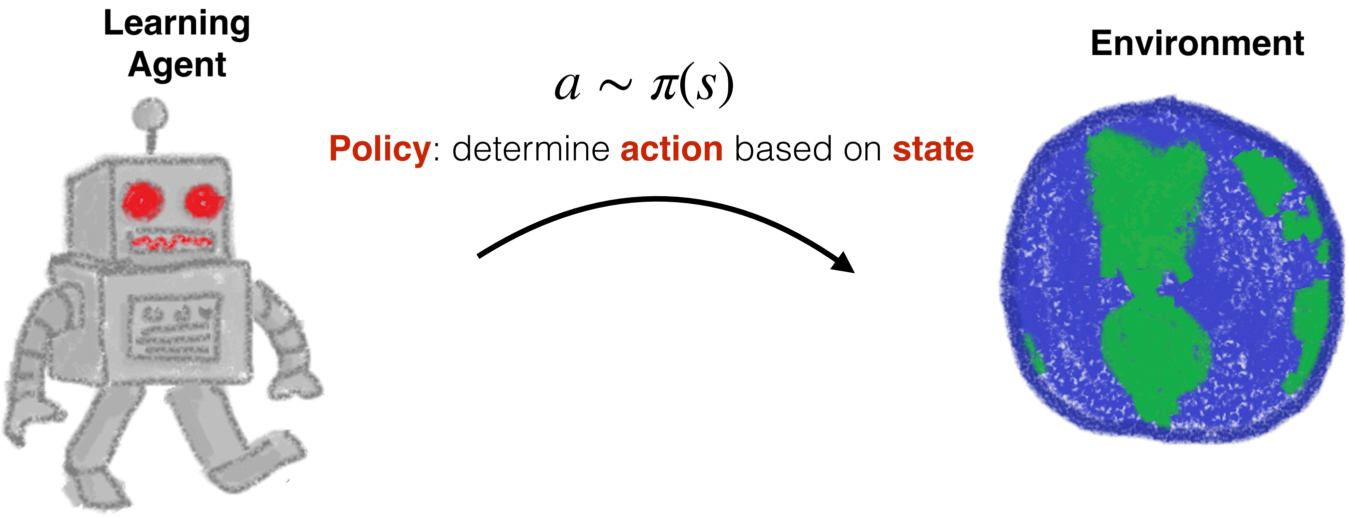
AgentLinear

Selected Actions:

RIGHT SPEED

Active: Decisions → Data Distribution

Markov Decision Process



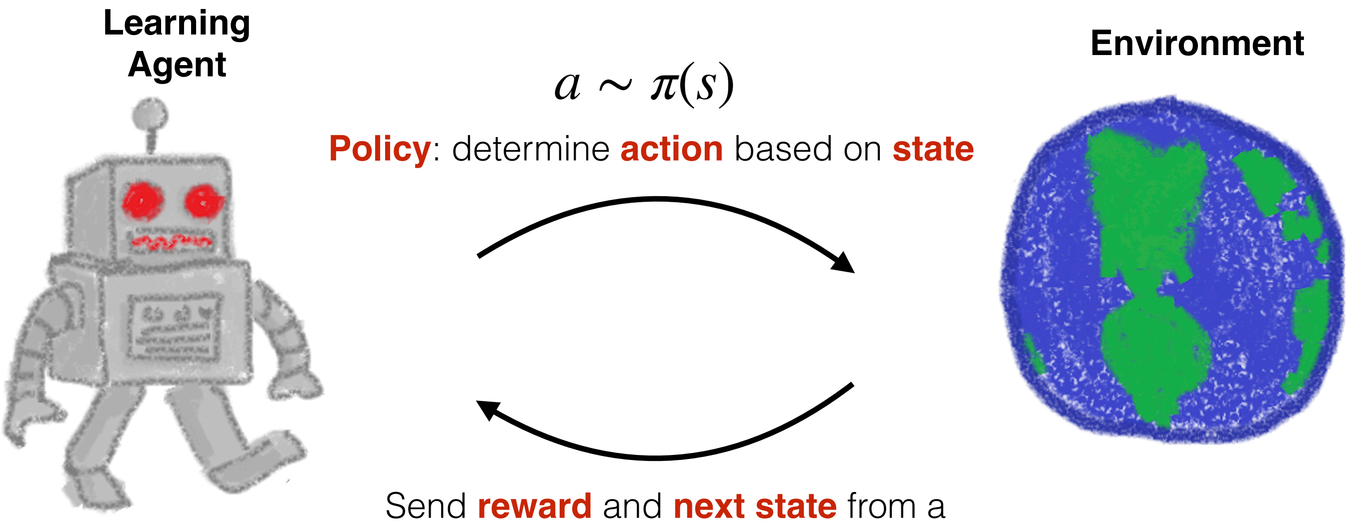
Learning Agent

Environment

$a \sim \pi(s)$

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

Markov Decision Process



Learning Agent

Environment

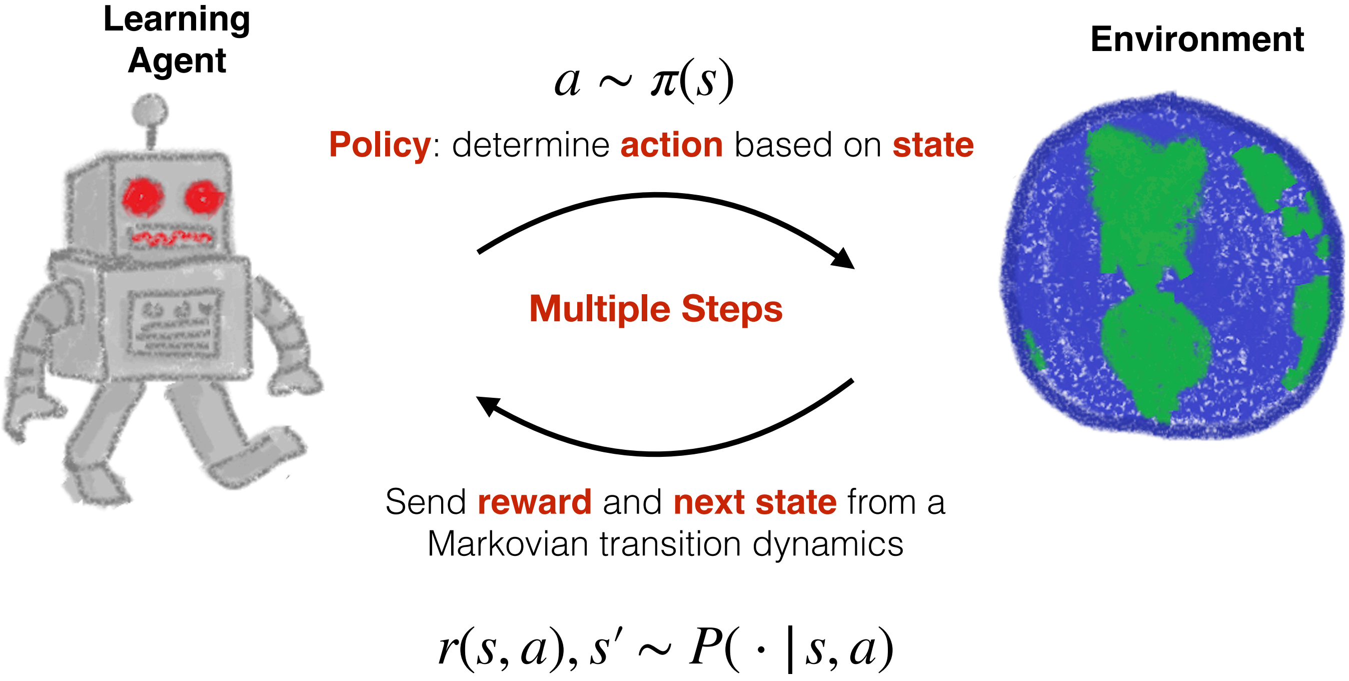
$$a \sim \pi(s)$$

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

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

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

Markov Decision Process



Learning Agent

Environment

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

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

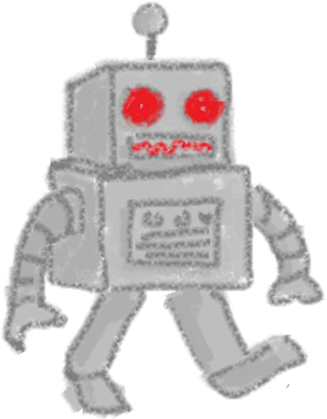
Multiple Steps

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

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

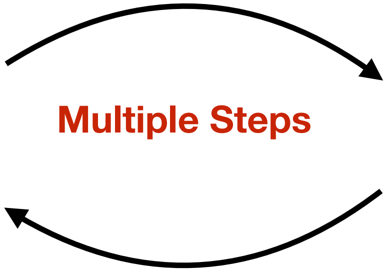
Markov Decision Process

Learning Agent



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

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

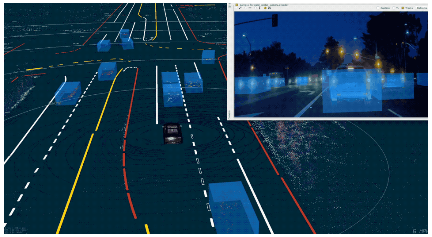


Multiple Steps

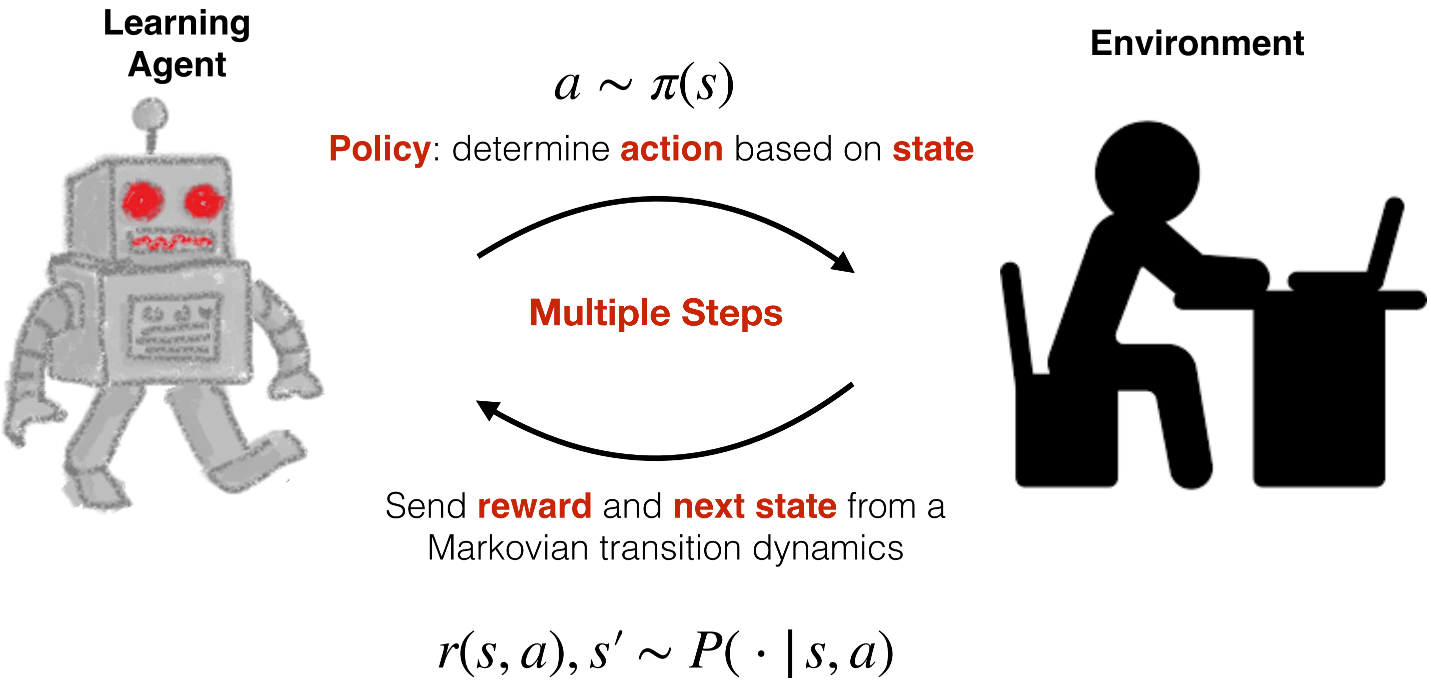
Send **reward** and **next state** from a Markovian transition dynamics

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

Environment



Markov Decision Process



Learning Agent

Environment

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

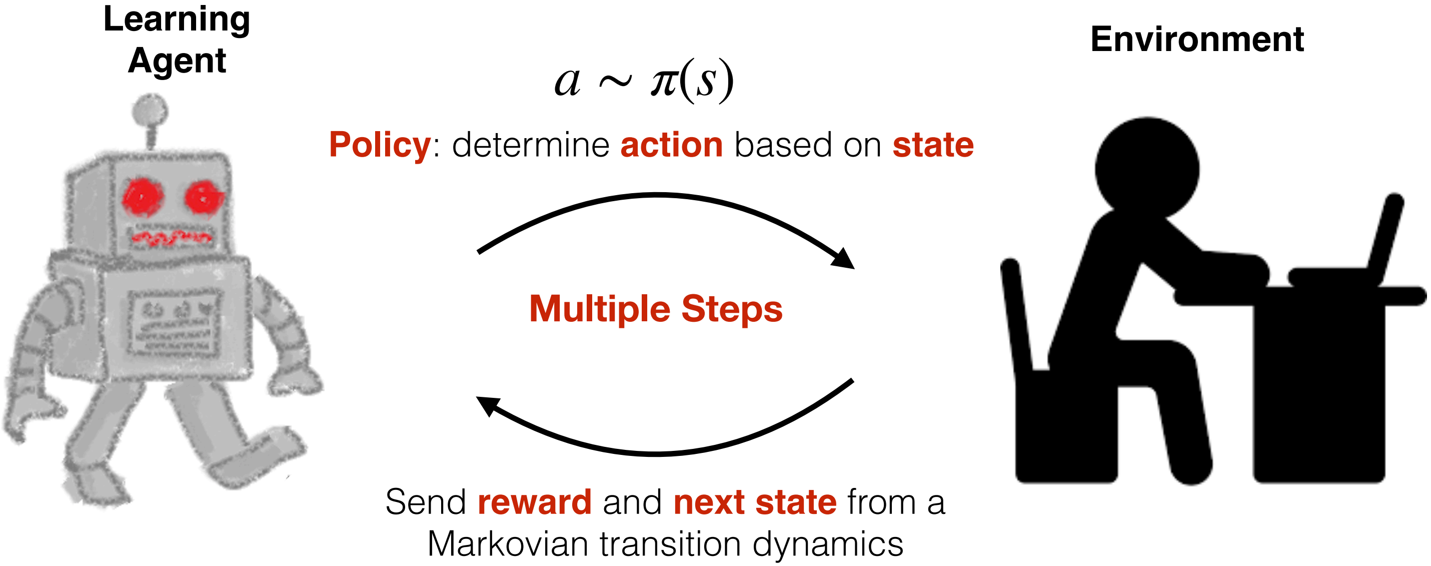
Policy: determine **action** based on **state**

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

Environment

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

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

Multiple Steps

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

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

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

	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

state

Actions

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

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

$$P(\cdot | s, a)$$

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

$\pi(\cdot | s)$

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

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

$\gamma \cdot r(s', a')$

$r(s, a) \leftarrow$ $a \sim \pi(\cdot|s)$ $a' \sim \pi(\cdot|s')$ $\gamma^2 r(s'', a'') + \gamma^3 r(s''', a''')$

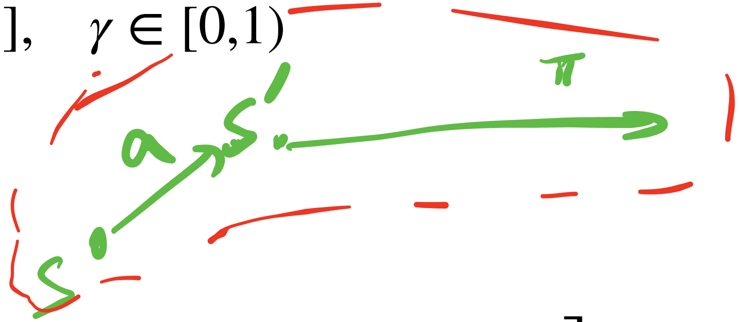
$s^0 \rightarrow s' \rightarrow s'' \sim P(\cdot|s', a')$

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

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

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

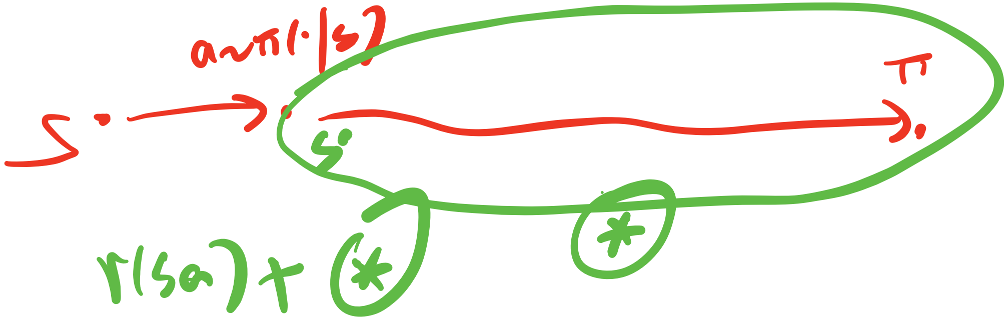
Bellman Equation:

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

Bellman Equation:

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

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



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:

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

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

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

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

$$s \xrightarrow{a} s' \longrightarrow \pi$$

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

$\pi^\star : S \mapsto A$, 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]

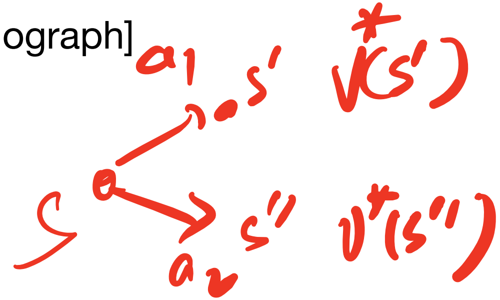
We denote $V^\star := V^{\pi^\star}, Q^\star := Q^{\pi^\star}$

Optimal Policy

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

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

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



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

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

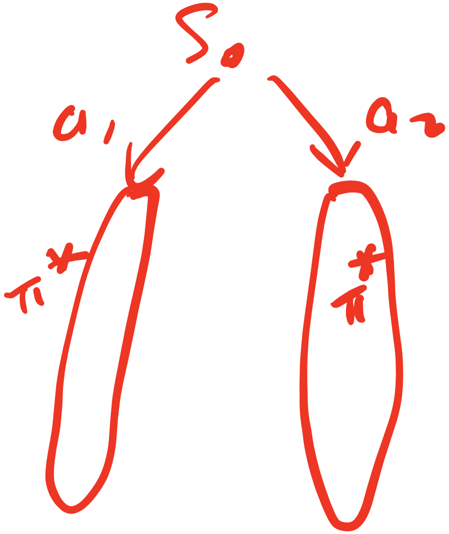
$$V^\star(s) = \max_a \left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} V^\star(s') \right]$$

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 $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

$$V^\star(s) \underset{\substack{\uparrow \\ \text{Bell-Eqn}}}{=} r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')$$

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 $V^{\widehat{\pi}}(s) = V^{\star}(s), \forall s$

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

$$\leq \underline{\max_{a}} \underbrace{\left[r(s,a) + \gamma \mathbb{E}_{s' \sim P(s,a)} V^{\star}(s') \right]}_{\text{Bell-Eqn} \quad Q^{\star}(s,a)} = r(s, \widehat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \widehat{\pi}(s))} V^{\star}(s')$$

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 $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

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

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

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

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 $V^{\widehat{\pi}}(s) = V^\star(s), \forall s$

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

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

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

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

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

$\forall s$:

$V^\star(s) \le V^{\hat{\pi}}(s)$

$V^\star(s) \ge V^{\hat{\pi}}(s)$

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

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

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

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

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

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

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