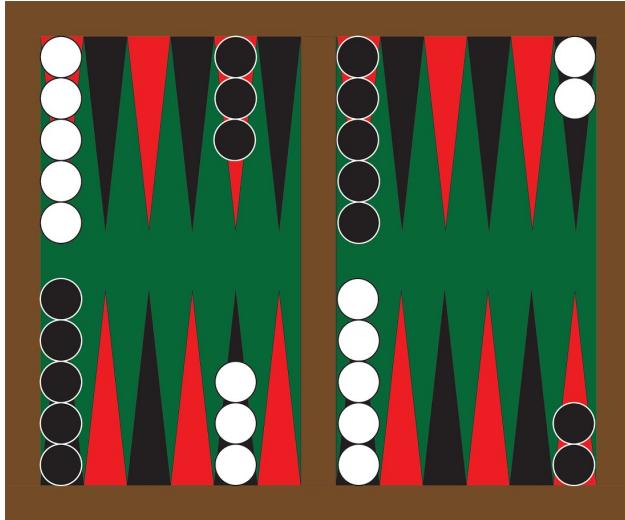


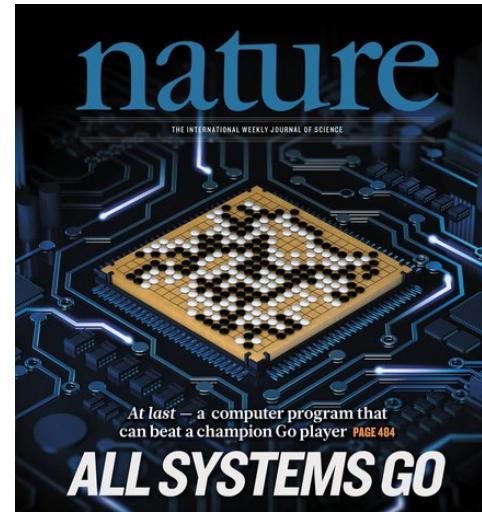
# **Introduction and Basics of Markov Decision Process**

**Sham Kakade and Kianté Brantley**  
**CS 2824: 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:



**Personalization**

# RL in Real World:



Personalization

A screenshot of the msn news homepage. The main headline is "Second US Ebola diagnosis 'deeply concerning' admits CDC chief" from The Guardian. Other visible headlines include "300,000 Evacuated as Strong Cyclone Hits Eastern India" (The New York Times), "Activists: Kurds halt jihad advance in Syria town" (AP Associated Press), and "Teenage Girl Dies In Haunted Hayride Crash" (Sky News). The page features a sidebar with various news items and a "make msn your homepage" button.

online advertising

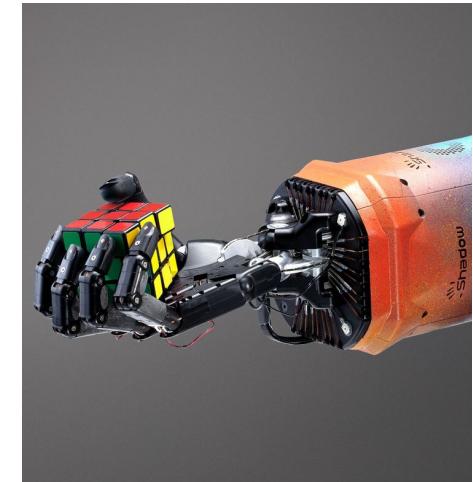
# RL in Real World:



Personalization

A screenshot of the msn news website homepage. The top banner features a photo of three men in blue shirts. Below it, a main article headline reads "Second US Ebola diagnosis 'deeply concerning' admits CDC chief". To the right, there is a sidebar with a purple msn logo and a link to "MAKE MSN YOUR HOMEPAGE". The main content area displays several news stories with small thumbnail images and titles, such as "Protesters stage sit-in at St. Louis gas station", "Teenage Girl Dies In 'Haunted Hayride' Crash", and "McCain urges ground troops to defeat Isis: 'They're winning, and...'".

online advertising



Robotics

# RL in Real World: Coding Assistant

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: Writing Assistant**

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

# RL in Real World: Writing Assistant



Just want to follow up on our lunch plan - I am available tomorrow, if this week's Wed works for you, or next week Mon and Tue; Otherwise, I'll be back on 11/29.

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

# RL in Real World: Writing Assistant



Just want to follow up on our lunch plan - I am available tomorrow, if this week's Wed works for you, or next week Mon and Tue; Otherwise, I'll be back on 11/29.



This week is a little rough, next Monday and tuesday are not great for me. Mondays I have lab meetings and research meeting around lunch time. Tuesday I have am teaching a course that is roughly around lunch time. Maybe lets reconnect wehn you are back 11/29.

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

# RL in Real World: Writing Assistant



Just want to follow up on our lunch plan - I am available tomorrow, if this week's Wed works for you, or next week Mon and Tue; Otherwise, I'll be back on 11/29.



This week is a little rough, next Monday and tuesday are not great for me. Mondays I have lab meetings and research meeting around lunch time. Tuesday I have am teaching a course that is roughly around lunch time. Maybe lets reconnect wehn you are back 11/29.

Next Monday and Tuesday are not great for me. On Mondays, I have lab meetings and a research meeting around lunchtime. On Tuesday, I will be teaching a course during roughly the same time. Perhaps we can reconnect when you return on 11/29.

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

# RL in Real World: Writing Assistant



Just want to follow up on our lunch plan - I am available tomorrow, if this week's Wed works for you, or next week Mon and Tue; Otherwise, I'll be back on 11/29.



This week is a little rough, next Monday and tuesday are not great for me. Mondays I have lab meetings and research meeting around lunch time. Tuesday I have am teaching a course that is roughly around lunch time. Maybe lets reconnect wehn you are back 11/29.

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

Next Monday and Tuesday are not great for me. On Mondays, I have lab meetings and a research meeting around lunchtime. On Tuesday, I will be teaching a course during roughly the same time. Perhaps we can reconnect when you return on 11/29.



Monday and Tuesday are **typically** not great for me. **On Mondays,** I have **lab meetings and**, research meetings. **On Tuesday, I will be teaching** and teach a course **during roughly the same time around lunchtime.** Perhaps we can reconnect when you return on 11/29.

# RL in Real World:

Genearting creative images that would never appeared in real world



# Logistics

# Course staff introductions

- **Instructors:** : Kianté Brantley and Sham Kakade
- **TFs:** Lukas Fesser, Jaeyeon Kim, and Alex Meterez
- **We will post Homework 0 today!**
  - We will make minor updates on the HW and post it on Ed.
  - This should be a review;  
**you should be familiar with the material** to take the course.

# Course Overview

All policies are stated on the course website:

<https://harvard-cs2824-s26.github.io/>

- We want u to obtain fundamental knowledge of RL.
- **Grades: Participation; Reading; HW0 +HW1-HW3; Project**
- **Readings:** Readings will be assigned. It is important you do these and turn them in on time. They help with learning the material.
- **HWs:** HW is designed to target to many of the concepts in the class.
- **Project:** 3 people per project. It must be theoretical (fine to also have an empirical component).
- **Bonus (5%):**

# Enrollment/Auditing

- Priority will be given to PhD students + having appropriate pre-requisites.
  - You needed to have filled out the form linked to on website for consideration.
  - You also need to add yourself to the petition via the registrar enrollment.
- You are welcome to audit/sit in on the course, though please give seats to the enrolled students (in case it is tight).
- Please hit “enroll” if you have been accepted in the course  
(so we have an accurate count to let more people in)
- Please drop if you know you will not take the course  
(so we can let others in)
  - Please see HW0.

# Other Points

- **Attendance:** it is expected to attend and do the readings.
- Communication: please use Ed to contact us
- Late policy (basically): you have 96 cumulative hours of late time.
  - *Please use this to plan for unforeseen circumstances.*

# Course Overview

- Fundamentals:
  - Sample Complexity
  - Tabular exploration (“UCB-VI”)
- Generalization:
  - RL in “large” (of inf dim) state spaces.
  - Upper bounds: What conditions lets us have guaranteed success.  
(e.g. Bellman rank)
  - Lower bounds: Why are getting such conditions so difficult in RL?  
(say in comparison to SL)
- (Direct) Policy Optimization:
  - Policy gradient methods are what work in practice. (why?)
  - theory/practice of them
- Other topics: RLHF/LLMs, imitation learning.

# Basics of Markov Decision Processes

# **Outline**

1. Definition of infinite horizon discounted MDPs
2. Bellman Optimality
3. State-action distribution

# **Supervised Learning**

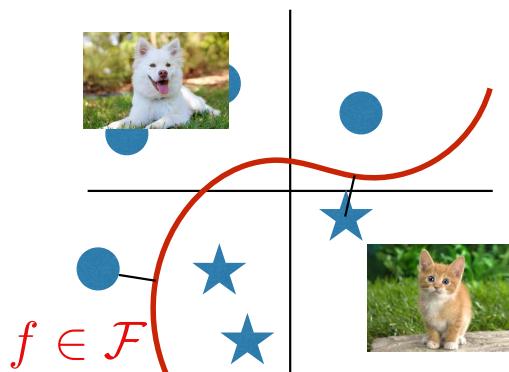
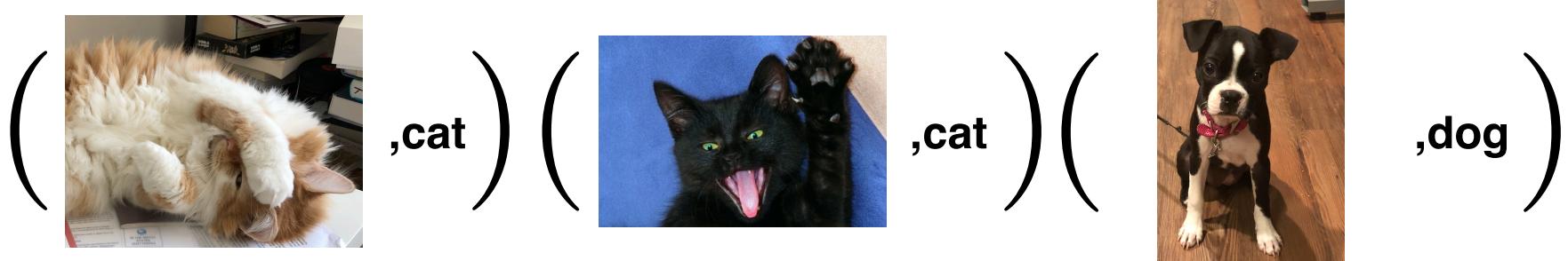
# Supervised Learning

Given i.i.d examples at training:



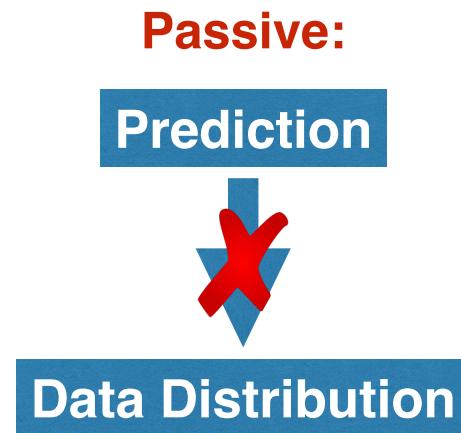
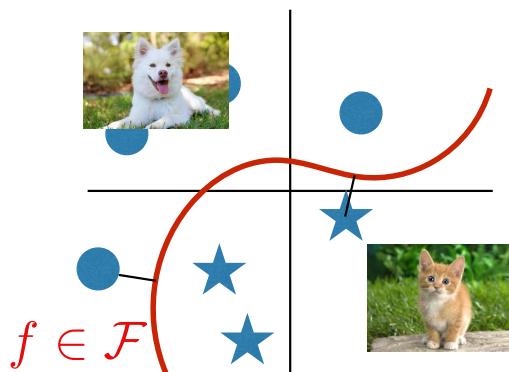
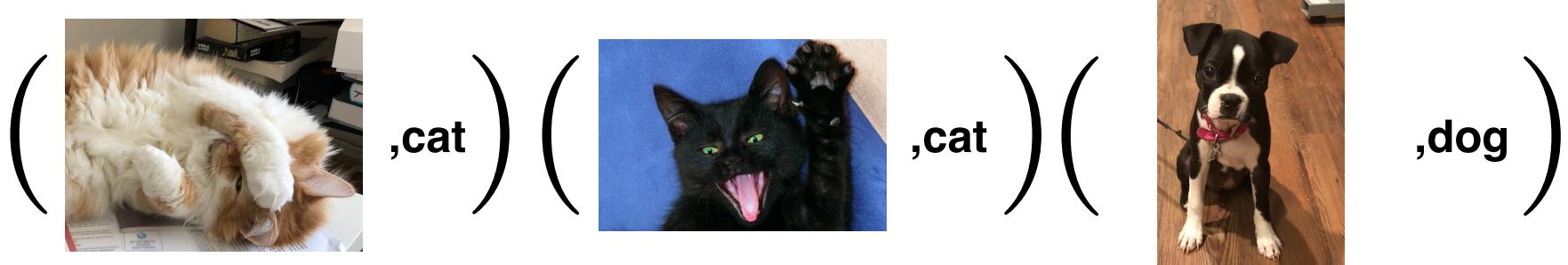
# Supervised Learning

Given i.i.d examples at training:



# Supervised Learning

Given i.i.d examples at training:



# Agent Linear Selected Actions:

RIGHT

SPEED

Active: Decisions → Data Distribution

# Agent Linear Selected Actions:

RIGHT

SPEED

Active: Decisions → Data Distribution

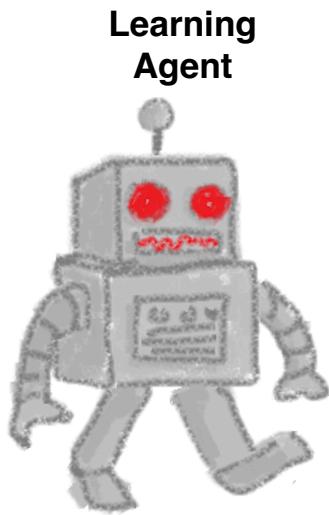
# Agent Linear Selected Actions:

RIGHT

SPEED

Active: Decisions → Data Distribution

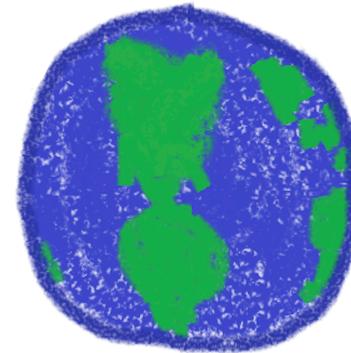
# Markov Decision Process



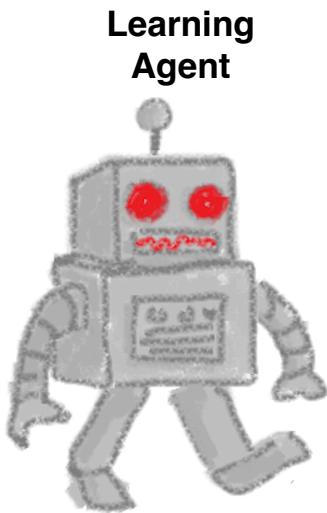
$$a \sim \pi(s)$$

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

Environment

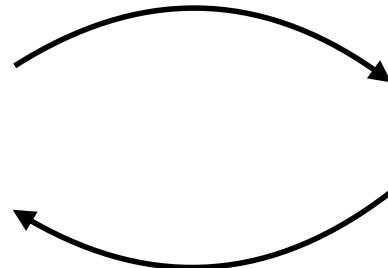


# Markov Decision Process

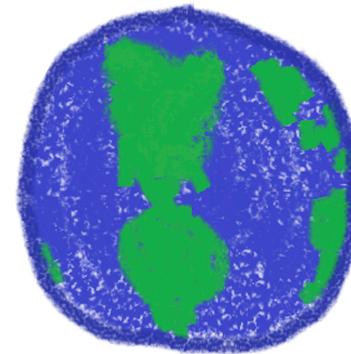


$$a \sim \pi(s)$$

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



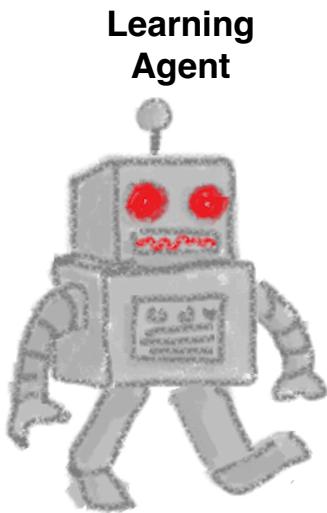
Environment



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

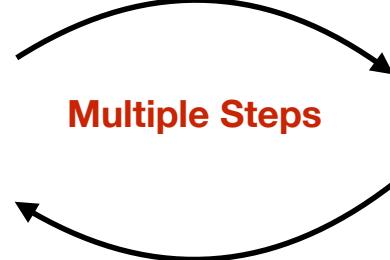
$$r(s, a), s' \sim P(\cdot | s, a)$$

# Markov Decision Process

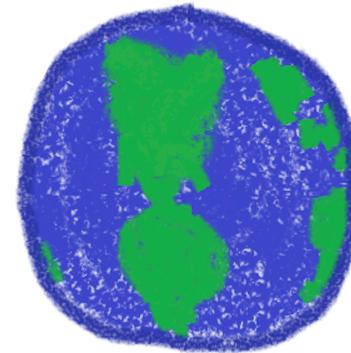


$$a \sim \pi(s)$$

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



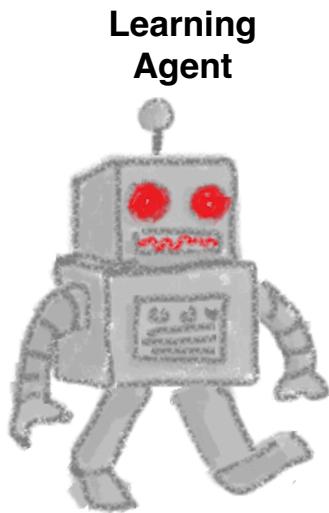
Environment



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

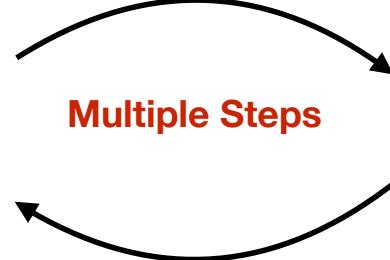
$$r(s, a), s' \sim P(\cdot | s, a)$$

# Markov Decision Process



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

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



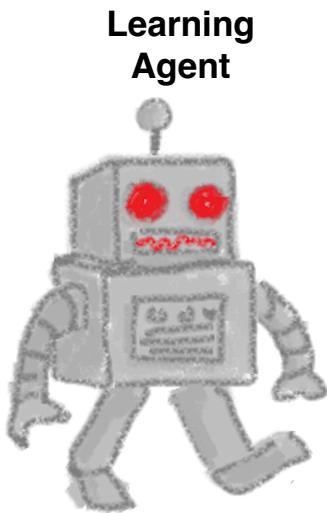
Send **reward** and **next state** from a  
Markovian transition dynamics

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

Environment

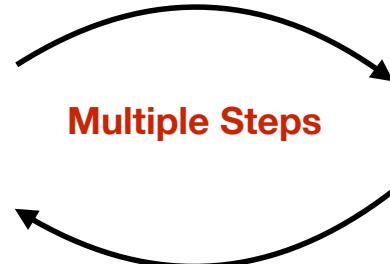


# Markov Decision Process



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

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



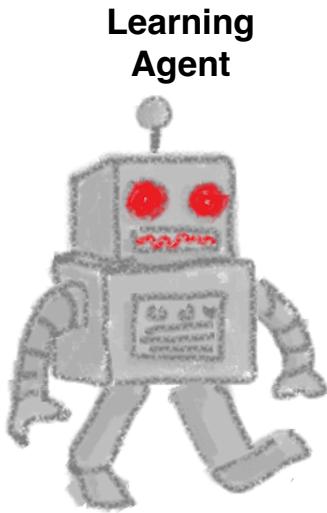
Send **reward** and **next state** from a  
Markovian transition dynamics

Environment



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

# Markov Decision Process



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

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



Environment



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

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

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

	<b>Learn from Experience</b>	<b>Generalize</b>	<b>Interactive</b>	<b>Exploration</b>	<b>Credit assignment</b>
<b>Supervised Learning</b>					
<b>Reinforcement Learning</b>					

	<b>Learn from Experience</b>	<b>Generalize</b>	<b>Interactive</b>	<b>Exploration</b>	<b>Credit assignment</b>
<b>Supervised Learning</b>	✓				
<b>Reinforcement Learning</b>	✓				

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

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

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

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

# Infinite horizon Discounted MDP

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

state space  
action space  
transition dynamics

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

# Infinite horizon Discounted MDP

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

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

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

$$s_0 \sim \mu_0$$

$$a_0 \sim \pi(s_0)$$

$$s_1 \sim P(\cdot | s_0, a_0)$$

$$r(s_0, a_0)$$

⋮

$$\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

# Infinite horizon Discounted MDP

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

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

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

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



# Infinite horizon Discounted MDP

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

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

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

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

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



# 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:

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

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

$$\begin{aligned} V^\pi(s) &= \sum r(s_0, a_0) + \gamma^1 r(s_1, a_1) + \gamma^2 r(s_2, a_2) + \dots \\ &= \mathbb{E}_{\substack{a \sim \pi(s)}} [r(s_0, a_0) + \gamma^1 r(s_1, a_1) + \gamma^2 r(s_2, a_2) + \dots] \\ &= \mathbb{E}_{\substack{a \sim \pi(s)}} [r(s_0, a_0) + \gamma \mathbb{E}_{\substack{a \sim \pi(s_1)}} [r(s_1, a_1) + \gamma^2 r(s_2, a_2) + \dots]] \\ &\quad \text{V}^\pi(s, \cdot) \end{aligned}$$

# Bellman Equation:

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

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

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

# 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

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

[Puterman 94 chapter 6, also see theorem 1.7 in the RL monograph]

# Optimal Policy

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

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

[Puterman 94 chapter 6, also see theorem 1.7 in the RL monograph]

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

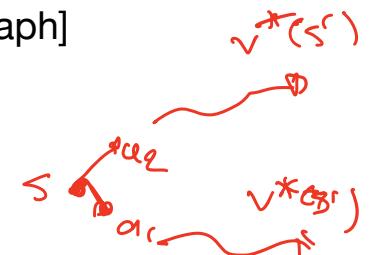
# Optimal Policy

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

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

[Puterman 94 chapter 6, also see theorem 1.7 in the RL monograph]

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



## Theorem 1: Bellman Optimality

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

# Proof of Bellman Optimality

**Theorem 1:** Bellman Optimality

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

# Proof of Bellman Optimality

**Theorem 1:** Bellman Optimality

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

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

$$Q^\star(s, a)$$

# Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

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

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

Bellman Equation

# Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

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

$$\begin{aligned} \underline{V^\star(s)} &= r(s, \underline{\pi^\star(s)}) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s') \\ &\leq \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^\star(s') \right] = r(s, \underline{\hat{\pi}(s)}) + \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}(s))} \boxed{V^\star(s')} \end{aligned}$$

*f(s,a) \leq \max\_a f(a)*

*$Q^\star(s,a)$*

# Proof of Bellman Optimality

**Theorem 1:** Bellman Optimality

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

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

$$\begin{aligned} V^\star(s) &= \underbrace{r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s')}_{\text{Bellman Optimality}} \\ &\leq \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^\star(s') \right] = r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}(s))} V^\star(s') \\ &= r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}(s))} \left[ \underbrace{r(s', \pi^\star(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \pi^\star(s'))} V^\star(s'')}_{\text{Bellman Optimality}} \right] \end{aligned}$$

# Proof of Bellman Optimality

**Theorem 1:** Bellman Optimality

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

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

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

# Proof of Bellman Optimality

**Theorem 1: Bellman Optimality**

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

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

$$\begin{aligned} V^\star(s) &= r(s, \pi^\star(s)) + \gamma \mathbb{E}_{s' \sim P(s, \pi^\star(s))} V^\star(s') \\ &\leq \max_a \left[ r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^\star(s') \right] = r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}(s))} V^\star(s') \\ &= r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}(s))} \left[ r(s', \pi^\star(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \pi^\star(s'))} V^\star(s'') \right] \\ &\leq r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}(s))} \left[ r(s', \hat{\pi}(s')) + \gamma \mathbb{E}_{s'' \sim P(s', \hat{\pi}(s'))} V^\star(s'') \right] \\ &\leq r(s, \hat{\pi}(s)) + \gamma \mathbb{E}_{s' \sim P(s, \hat{\pi}(s))} \left[ \underbrace{r(s', \hat{\pi}(s'))}_{\text{blue underline}} + \gamma \mathbb{E}_{s'' \sim P(s', \hat{\pi}(s'))} \left[ \underbrace{r(s'', \hat{\pi}(s''))}_{\text{blue underline}} + \gamma \mathbb{E}_{s''' \sim P(s'', \hat{\pi}(s''))} \underbrace{V^\star(s''')}_{\text{blue box}} \right] \right] \end{aligned}$$

# Proof of Bellman Optimality

**Theorem 1: Bellman Optimality**

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

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

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

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

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

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

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

$$\leq \mathbb{E} [r(s, \hat{\pi}(s)) + \gamma r(s', \hat{\pi}(s')) + \dots] = V^{\hat{\pi}}(s)$$

$$\textcircled{2} \quad V^\star(s) \geq V^{\hat{\pi}}(s)$$

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

$$\textcircled{3} \quad V^\star(s) \leq V^{\hat{\pi}}(s)$$

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

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

$\hat{\pi}$

# Proof of Bellman Optimality

## Theorem 2:

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

# Proof of Bellman Optimality

## Theorem 2:

For any  $V : S \rightarrow \mathbb{R}$ , 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  $V(s) = V^*(s), \forall s$

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

# Proof of Bellman Optimality

## Theorem 2:

For any  $V : S \rightarrow \mathbb{R}$ , 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  $V(s) = V^*(s), \forall s$

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

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

$$\left| \max_a f(a) - \max_a g(a) \right| \leq \max_a |f(a) - g(a)|$$

# Proof of Bellman Optimality

## Theorem 2:

For any  $V : S \rightarrow \mathbb{R}$ , 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  $V(s) = V^*(s), \forall s$

$$\begin{aligned} |V(s) - V^*(s)| &= \left| \max_a (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V(s')) - \max_a (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s')) \right| \\ &\leq \max_a \left| (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V(s')) - (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s')) \right| \\ &\leq \max_a \gamma \mathbb{E}_{s' \sim P(s, a)} |V(s') - V^*(s')| \end{aligned}$$

$$\left| \mathbb{E}_{x \sim P} f(x) - \mathbb{E}_{x \sim P} g(x) \right| \leq \mathbb{E}_{x \sim P} |f(x) - g(x)|$$

# Proof of Bellman Optimality

## Theorem 2:

For any  $V : S \rightarrow \mathbb{R}$ , 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  $V(s) = V^*(s), \forall s$

$$\begin{aligned} \underbrace{|V(s) - V^*(s)|}_{\text{ }} &= \left| \max_a (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V(s')) - \max_a (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s')) \right| \\ &\leq \max_a \left| (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V(s')) - (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s')) \right| \\ &\leq \max_a \gamma \mathbb{E}_{s' \sim P(s, a)} \boxed{\left| V(s') - V^*(s') \right|} \\ &\leq \max_a \gamma \mathbb{E}_{s' \sim P(s, a)} \left( \max_{a'} \gamma \mathbb{E}_{s'' \sim P(s', a')} \left| V(s'') - V^*(s'') \right| \right) \end{aligned}$$

# Proof of Bellman Optimality

## Theorem 2:

For any  $V : S \rightarrow \mathbb{R}$ , 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  $V(s) = V^*(s), \forall s$

$$\begin{aligned} |V(s) - V^*(s)| &= \left| \max_a (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V(s')) - \max_a (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s')) \right| \\ &\leq \max_a \left| (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V(s')) - (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^*(s')) \right| \\ &\leq \max_a \gamma \mathbb{E}_{s' \sim P(s, a)} |V(s') - V^*(s')| \\ &\stackrel{\textcircled{*}}{\leq} \max_a \gamma \mathbb{E}_{s' \sim P(s, a)} \left( \underbrace{\max_{a'} \gamma \mathbb{E}_{s'' \sim P(s', a')} |V(s'') - V^*(s'')|}_{\gamma \in [0, 1]} \right) \\ &\leq \max_{a_1, a_2, \dots, a_{k-1}} \gamma^k \mathbb{E}_{s_k} |V(s_k) - V^*(s_k)| \quad \gamma \in [0, 1] \xrightarrow{k \rightarrow \infty} \gamma^k \rightarrow 0 \end{aligned}$$

# Outline

✓ 1. Definition of infinite horizon discounted MDPs

✓ 2. Bellman Optimality

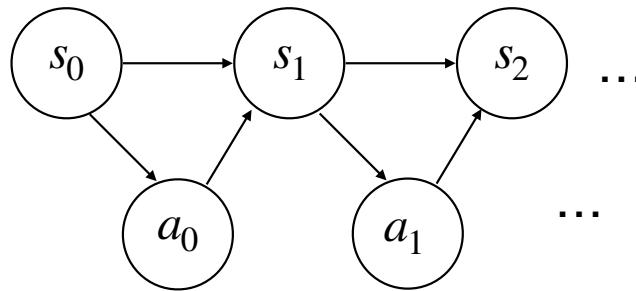
3. State-action distribution

# Trajectory distribution and state-action distribution

Q: Assume we start at  $s_0$ , following  $\pi$  to the step  $h$ , what is the probability of generating a trajectory  $\tau = \{s_0, a_0, s_1, a_1, \dots, s_h, a_h\}$ ?

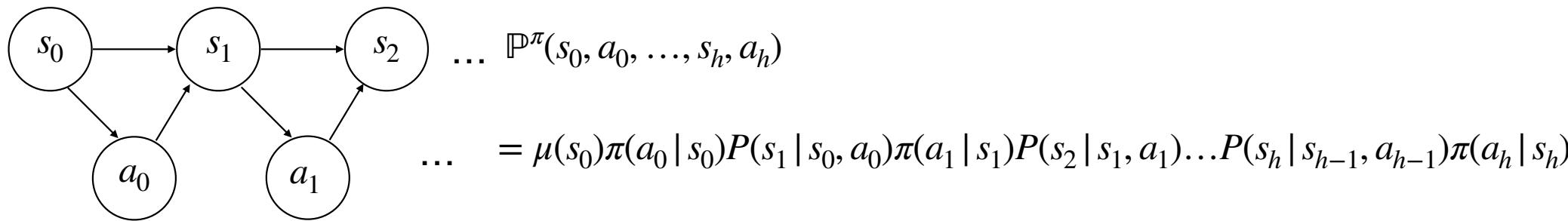
# Trajectory distribution and state-action distribution

Q: Assume we start at  $s_0$ , following  $\pi$  to the step  $h$ , what is the probability of generating a trajectory  $\tau = \{s_0, a_0, s_1, a_1, \dots, s_h, a_h\}$ ?



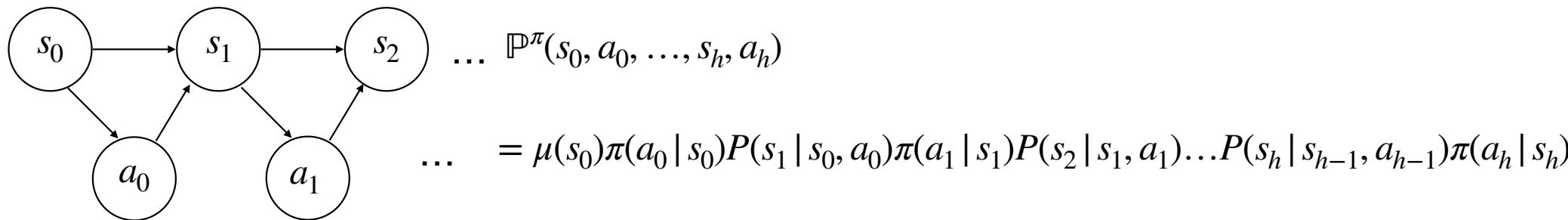
# Trajectory distribution and state-action distribution

Q: Assume we start at  $s_0$ , following  $\pi$  to the step  $h$ , what is the probability of generating a trajectory  $\tau = \{s_0, a_0, s_1, a_1, \dots, s_h, a_h\}$ ?



# Trajectory distribution and state-action distribution

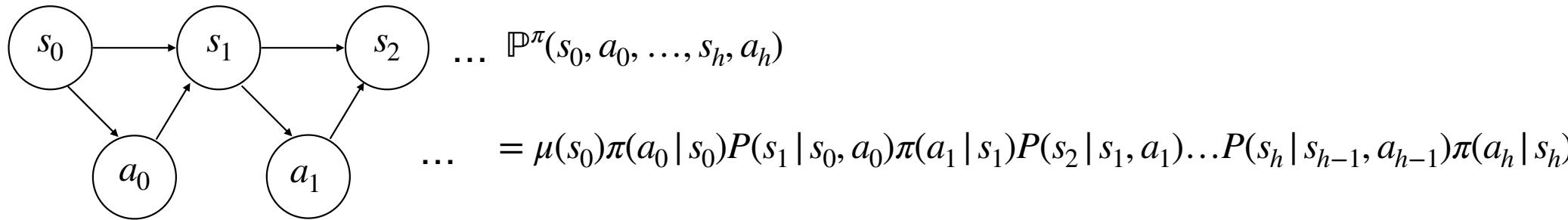
Q: Assume we start at  $s_0$ , following  $\pi$  to the step h, what is the probability of generating a trajectory  $\tau = \{s_0, a_0, s_1, a_1, \dots, s_h, a_h\}$ ?



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

# Trajectory distribution and state-action distribution

Q: Assume we start at  $s_0$ , following  $\pi$  to the step  $h$ , what is the probability of generating a trajectory  $\tau = \{s_0, a_0, s_1, a_1, \dots, s_h, a_h\}$ ?



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

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

## Average State-Action occupancy measure

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

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

# Average State-Action occupancy measure

$\mathbb{P}_h^\pi(s, a)$ : probability of  $\pi$  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)