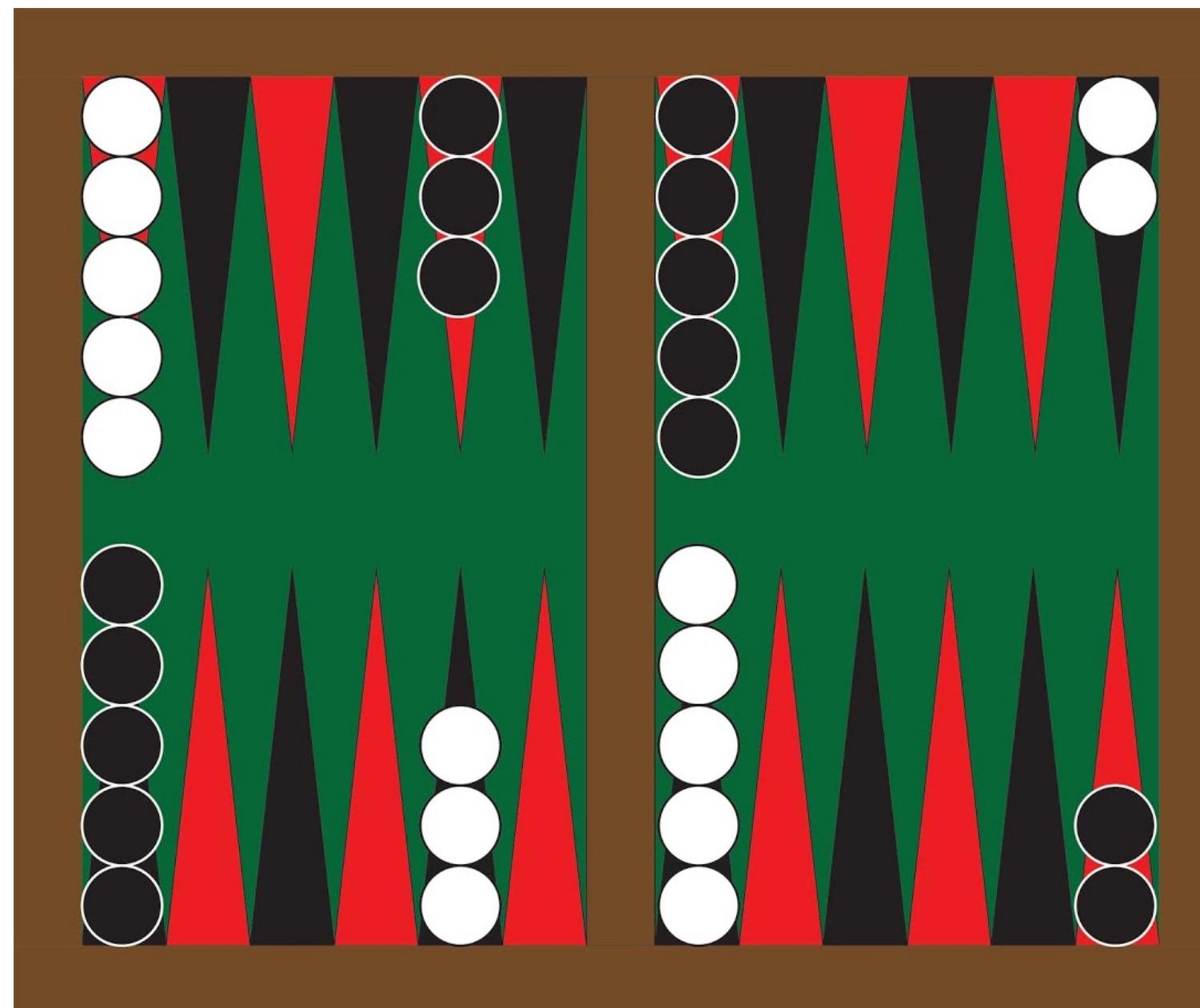


Introduction and Basics of Markov Decision Process

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 top navigation bar includes 'msn news', a search bar, and a 'Welcome' button. Below the navigation is a menu with links like Headlines, US, World, Crime, Local, Technology, Offbeat, Politics, Opinion, Magazine, Video, and My Topics. The main content area features several news stories with images and titles. One prominent story is 'Second US Ebola diagnosis 'deeply concerning'', admitted CDC chief'. Other visible stories include '300,000 Evacuated as Strong Cyclone Hits Eastern India', 'Activists: Kurds halt jihadi advance in Syria town', and 'Teenage Girl Dies In 'Haunted Hayride' Crash'. On the right side, there's a purple sidebar for Microsoft with the text 'MAKE MSN YOUR HOMEPAGE' and 'Image of the day' showing a person cheering. Below the sidebar are more news snippets.

online advertising

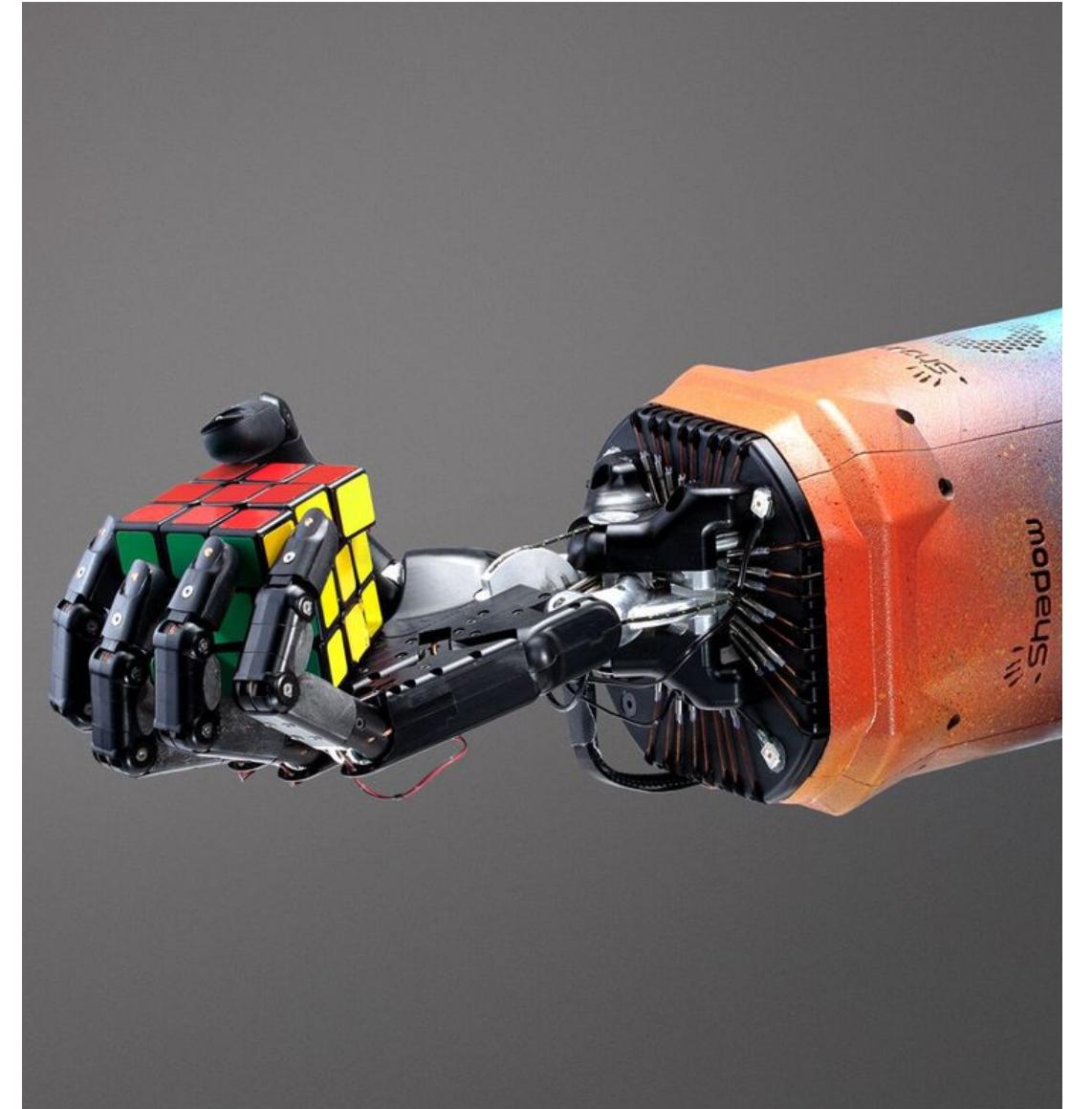
RL in Real World:



Personalization

A screenshot of the MSN News homepage. The top navigation bar includes 'msn news', a search bar, and a 'Welcome' button. Below the navigation is a grid of news articles. One prominent article on the left features a photo of several people and the headline 'Second US Ebola diagnosis 'deeply concerning', admits CDC chief'. Other visible headlines include '300,000 Evacuated as Strong Cyclone Hits Eastern India', 'Activists: Kurds halt jihadi advance in Syria town', and 'Teenage Girl Dies In 'Haunted Hayride' Crash'. The right side of the page features a large Microsoft advertisement for 'msn' with the text 'MAKE MSN YOUR HOMEPAGE' and 'Microsoft'.

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

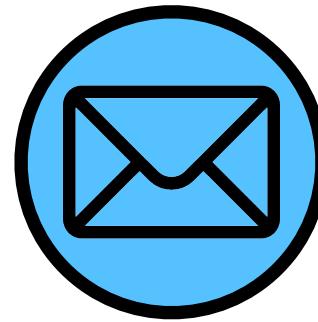
CHATGPT It's difficult to say without more information about what the code is supposed to do and what's happening when it's executed. One potential issue with the code you provided is that the `resultWorkerErr` channel is never closed, which means that the code could potentially hang if the `resultWorkerErr`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

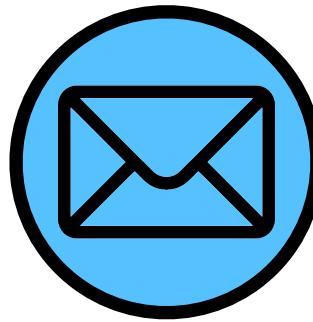
RL in Real World: Writing Assistant



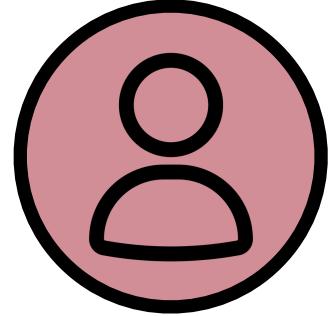
Just want to follow up on our lunch plan - I am available tomorrow, if this week's Wed works for you, or next week Mon and Tue; Otherwise, I'll be back on 11/29.

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

RL in Real World: Writing Assistant



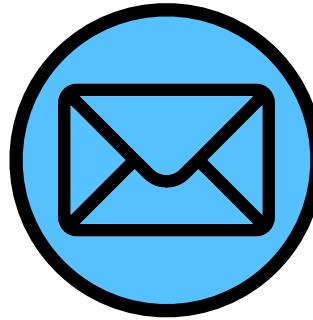
Just want to follow up on our lunch plan - I am available tomorrow, if this week's Wed works for you, or next week Mon and Tue; Otherwise, I'll be back on 11/29.



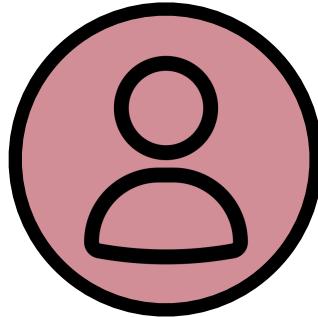
This week is a little rough, next Monday and tuesday are not great for me. Mondays I have lab meetings and research meeting around lunch time. Tuesday I have am teaching a course that is roughly around lunch time. Maybe lets reconnect wehn you are back 11/29.

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.

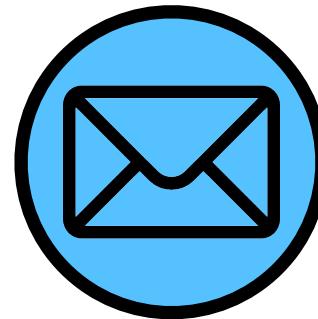


This week is a little rough, next Monday and tuesday are not great for me. Mondays I have lab meetings and research meeting around lunch time. Tuesday I have am teaching a course that is roughly around lunch time. Maybe lets reconnect wehn you are back 11/29.

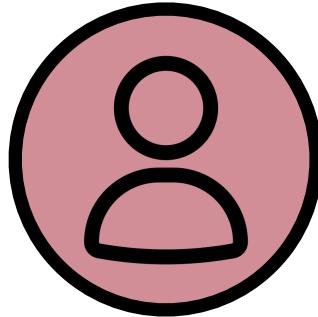
Next Monday and Tuesday are not great for me. On Mondays, I have lab meetings and a research meeting around lunchtime. On Tuesday, I will be teaching a course during roughly the same time. Perhaps we can reconnect when you return on 11/29.

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

RL in Real World: Writing Assistant



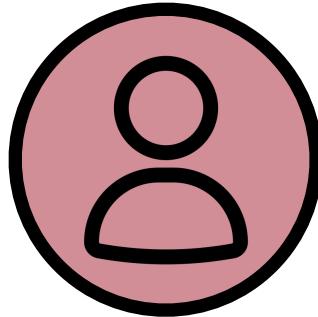
Just want to follow up on our lunch plan - I am available tomorrow, if this week's Wed works for you, or next week Mon and Tue; Otherwise, I'll be back on 11/29.



This week is a little rough, next Monday and Tuesaday 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



Basics of Markov Decision Processes

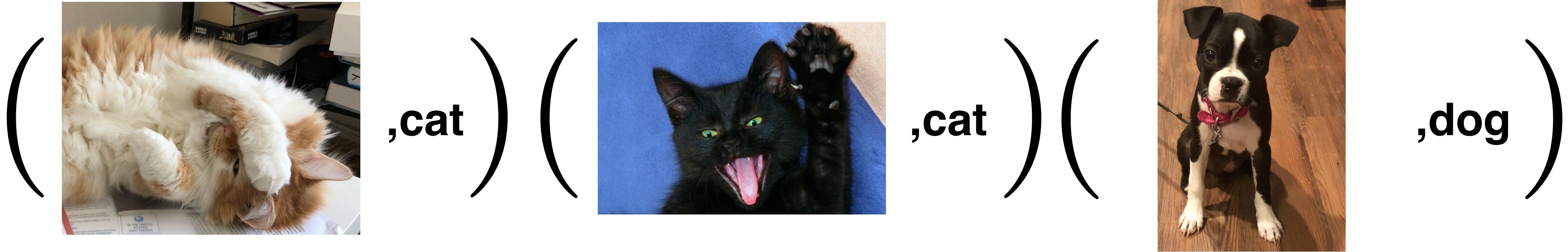
Outline

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

Supervised Learning

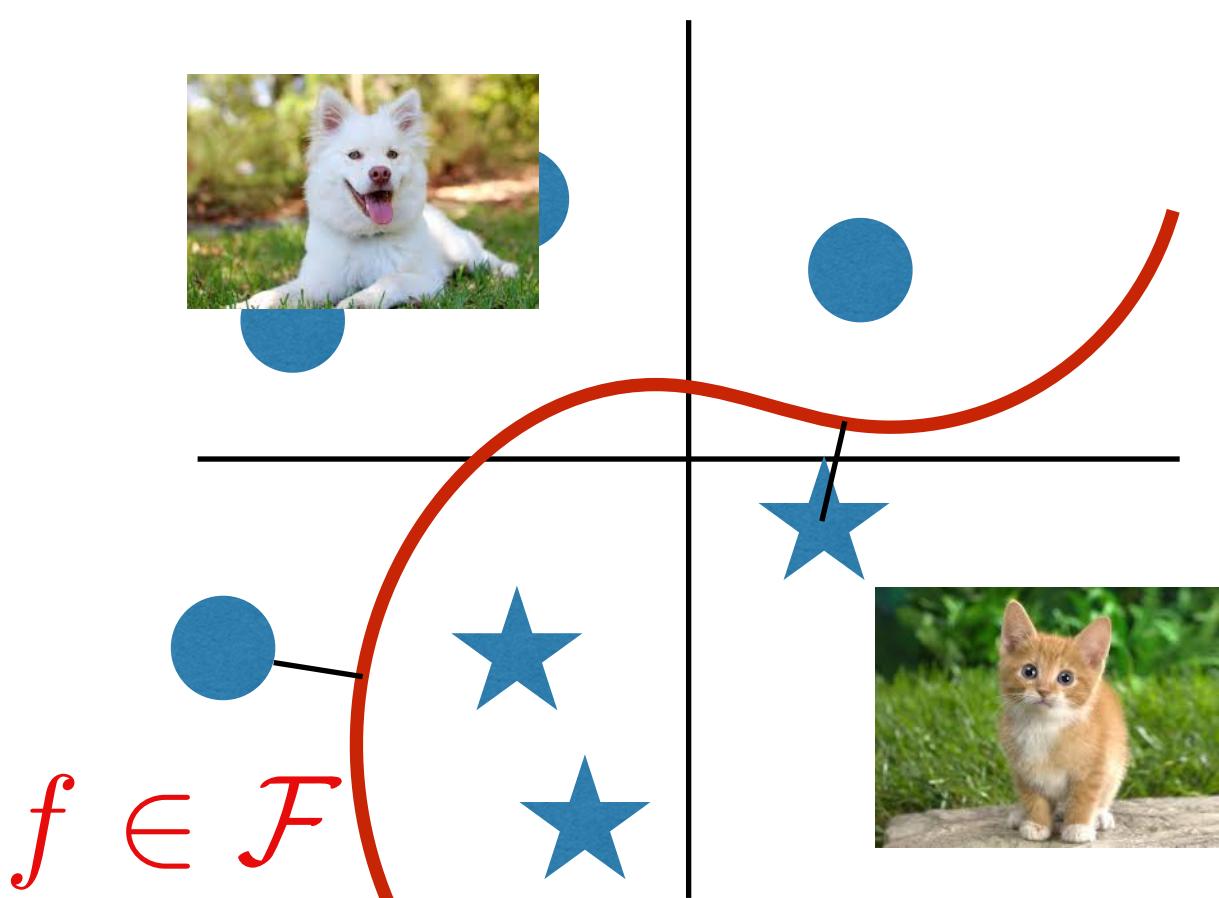
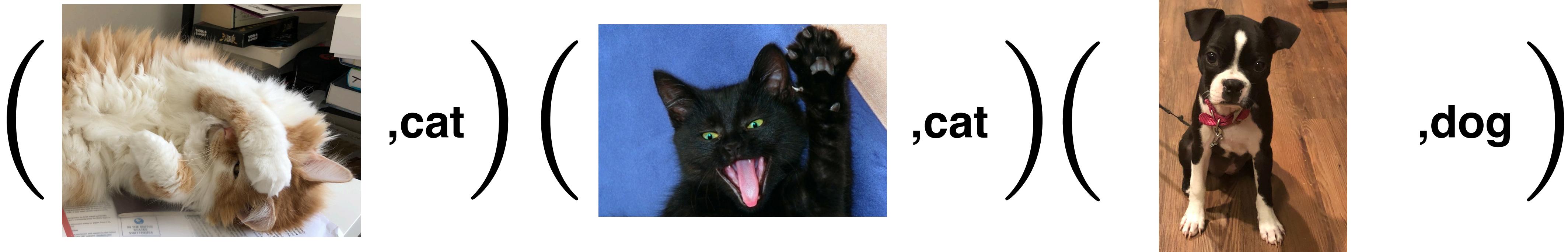
Supervised Learning

Given i.i.d examples at training:



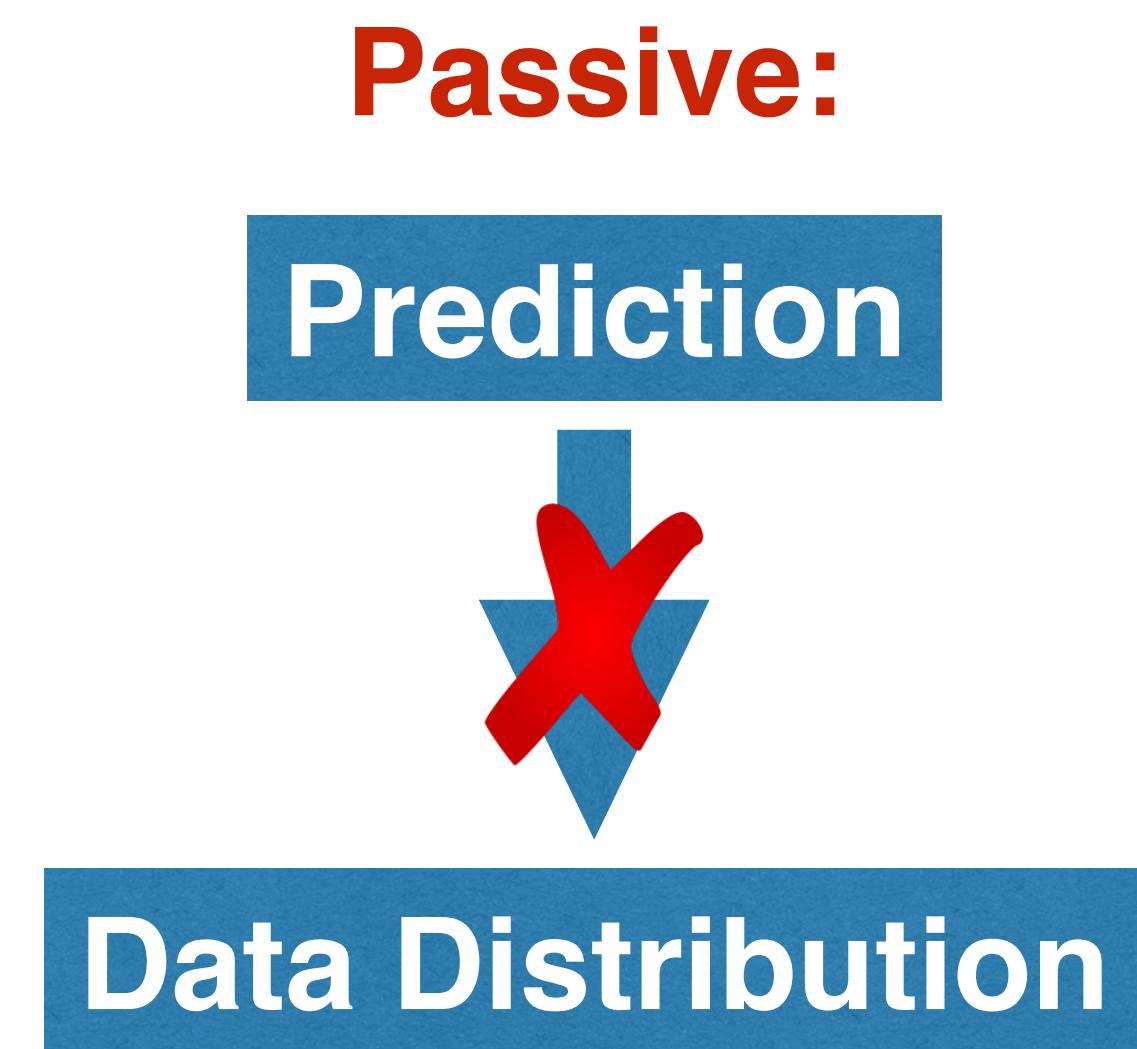
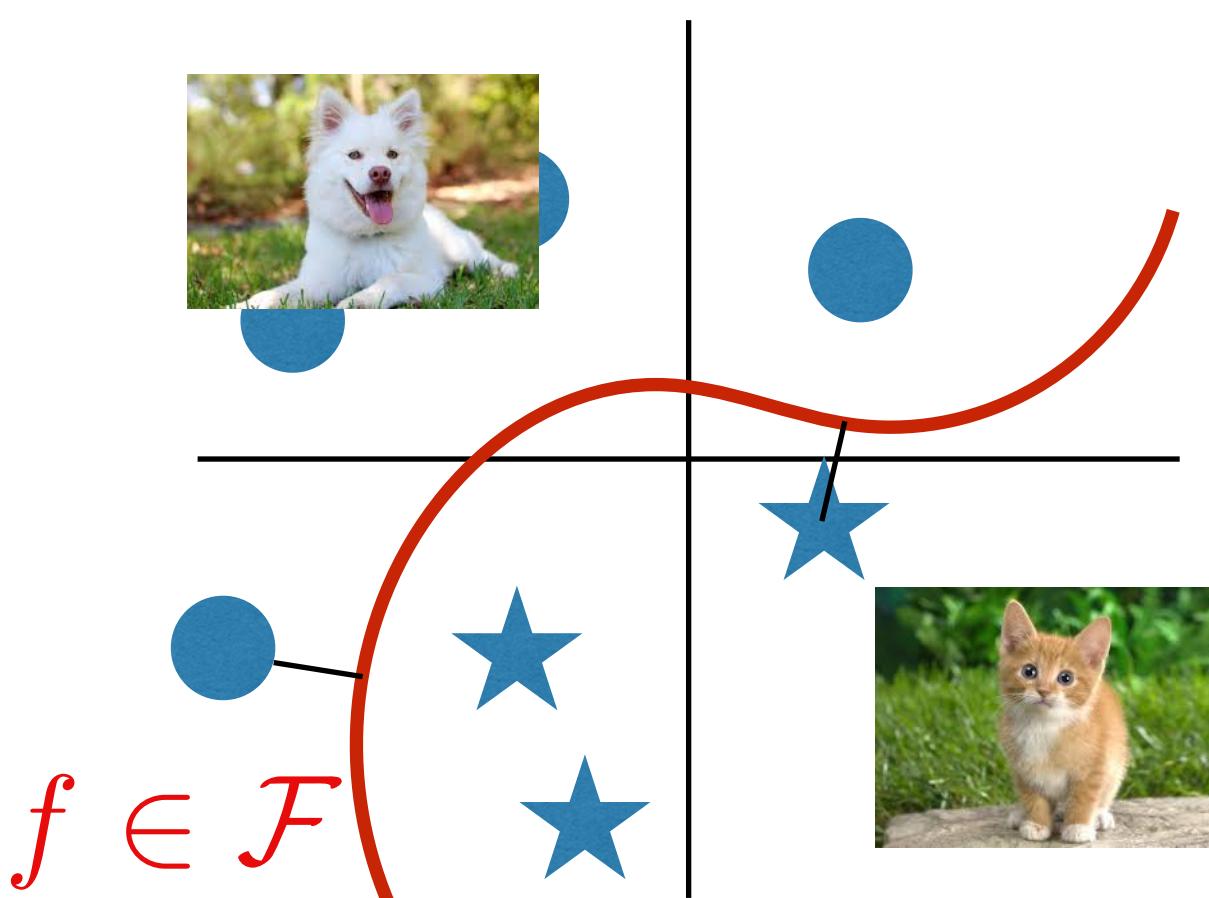
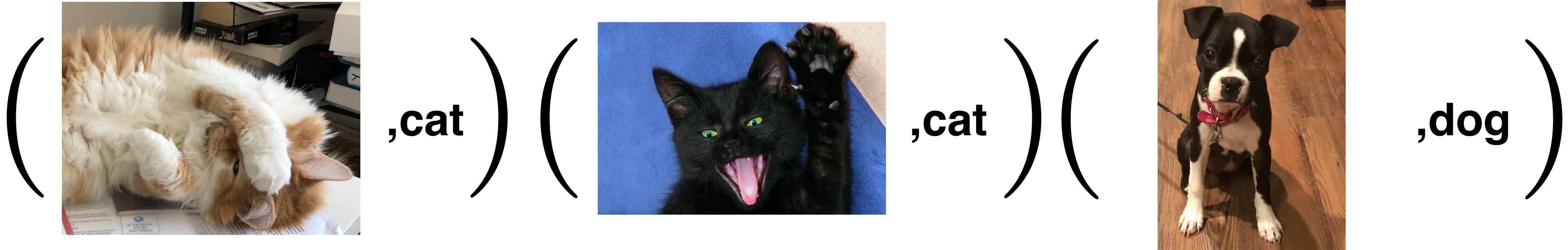
Supervised Learning

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

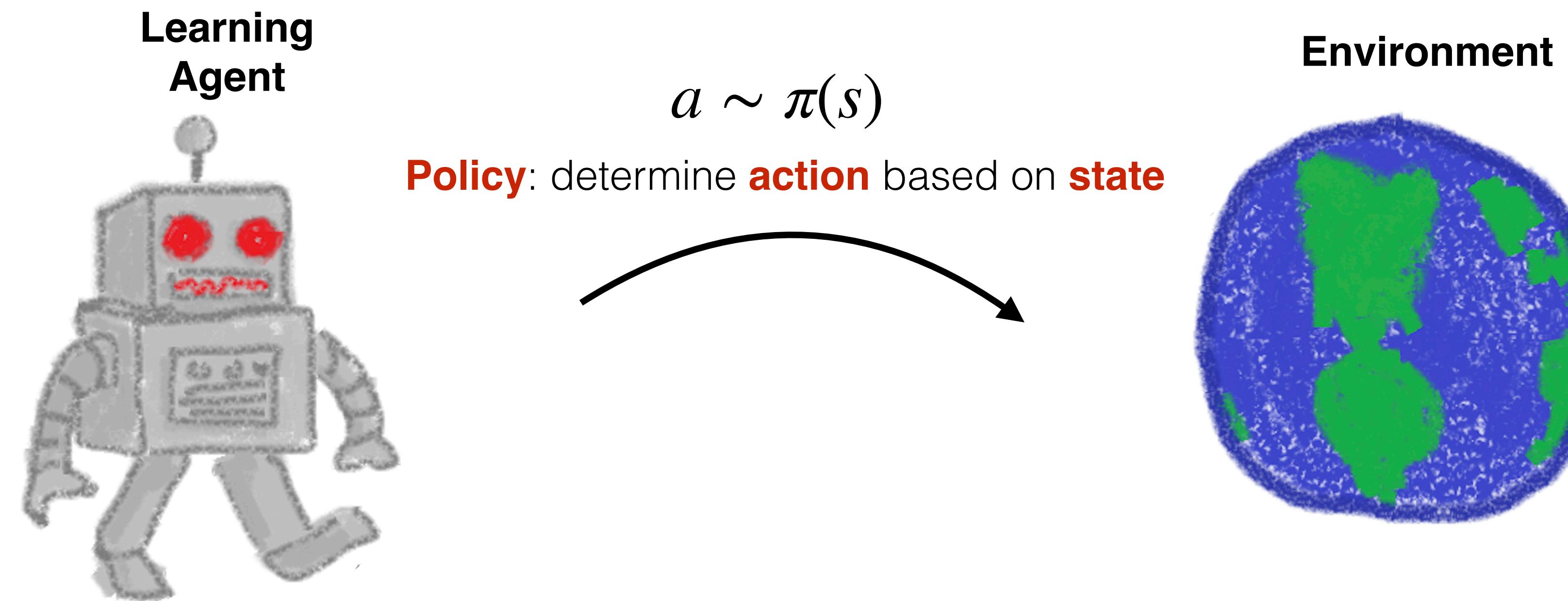
Agent Linear Selected Actions:

RIGHT

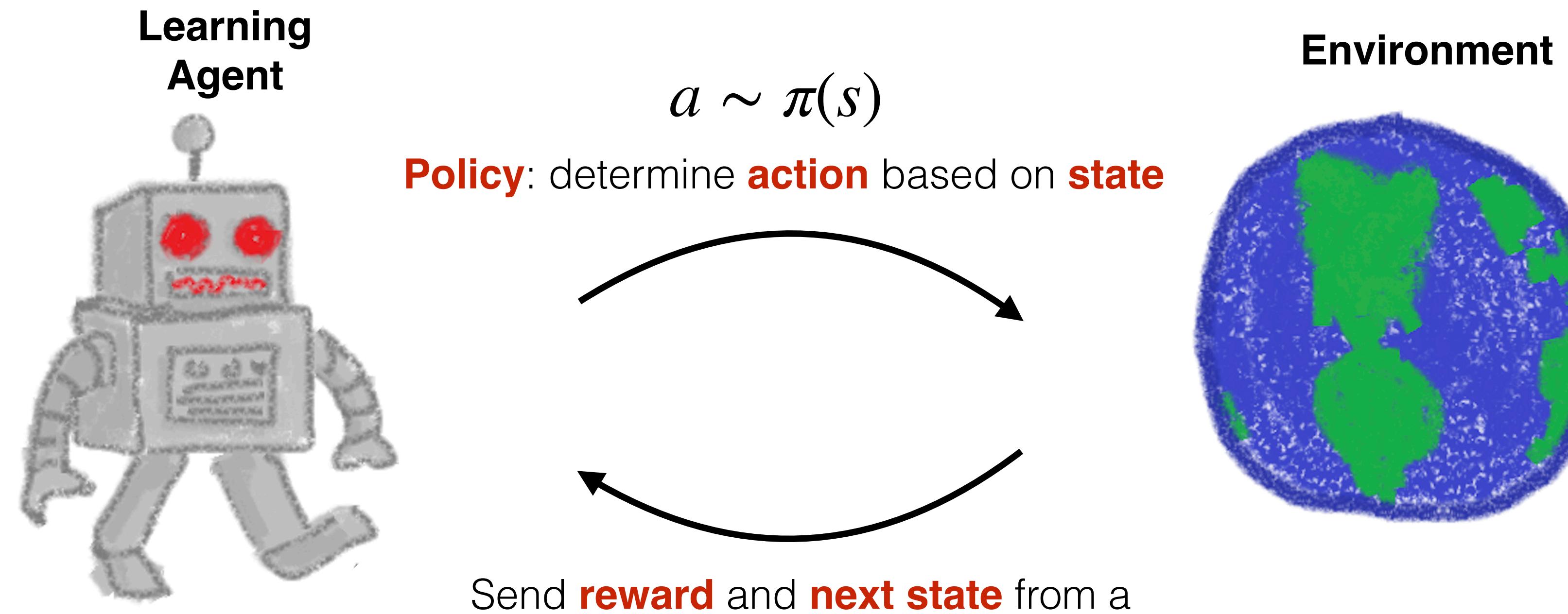
SPEED

Active: Decisions → Data Distribution

Markov Decision Process

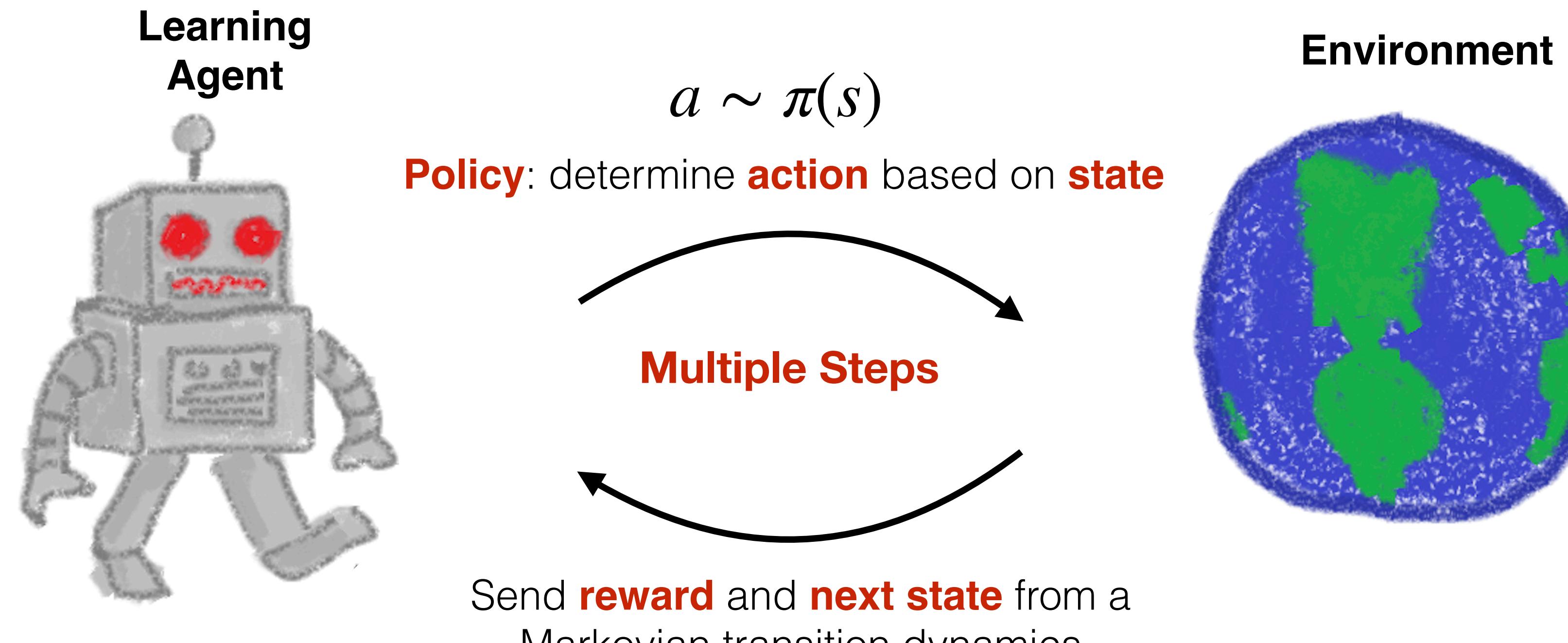


Markov Decision Process



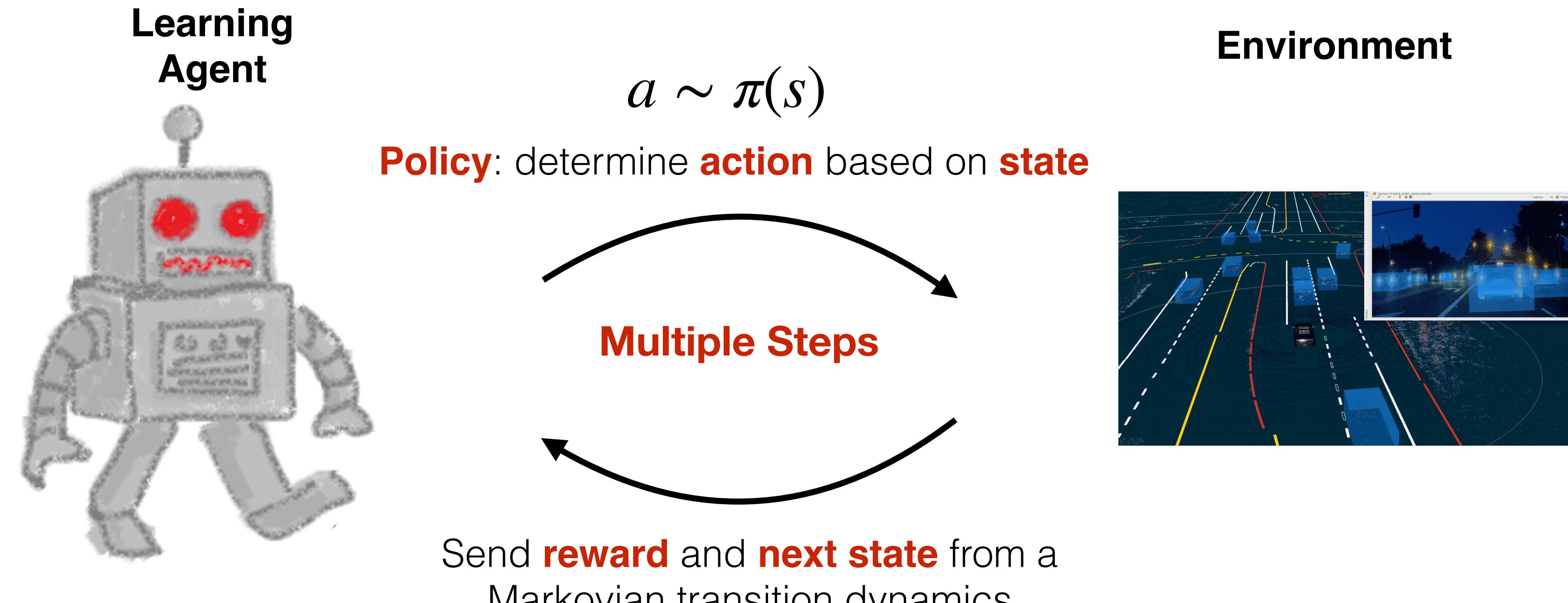
$$r(s, a), s' \sim P(\cdot | s, a)$$

Markov Decision Process



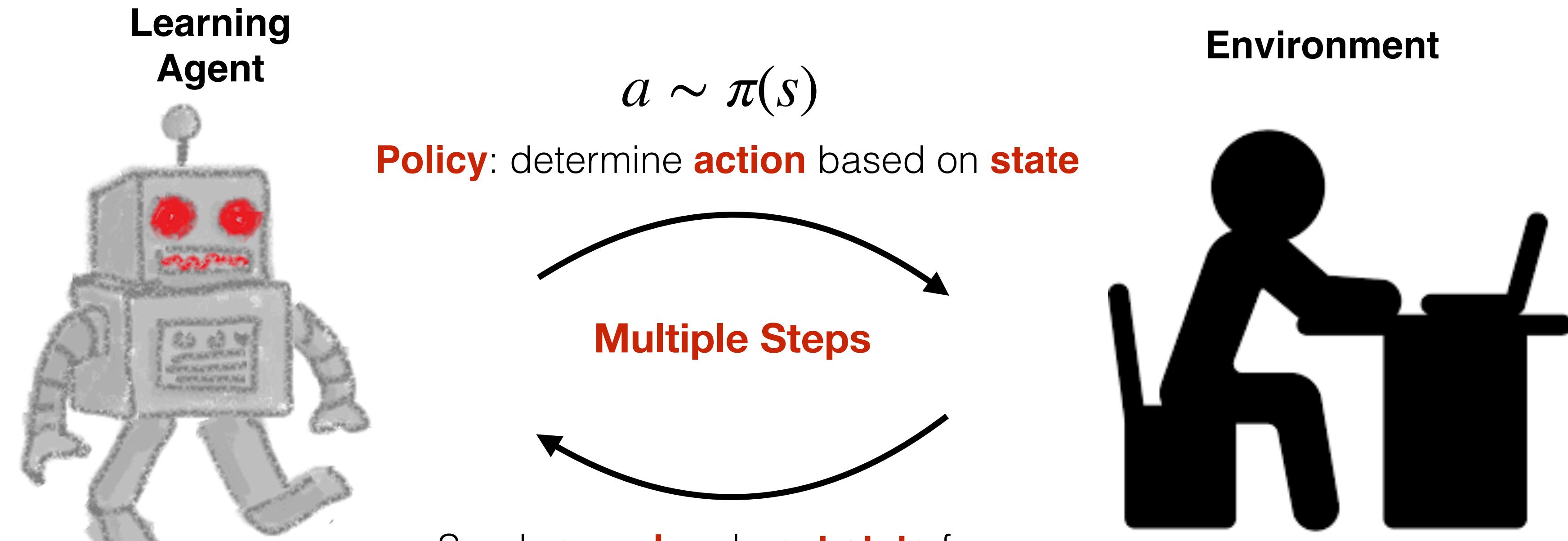
$$r(s, a), s' \sim P(\cdot | s, a)$$

Markov Decision Process



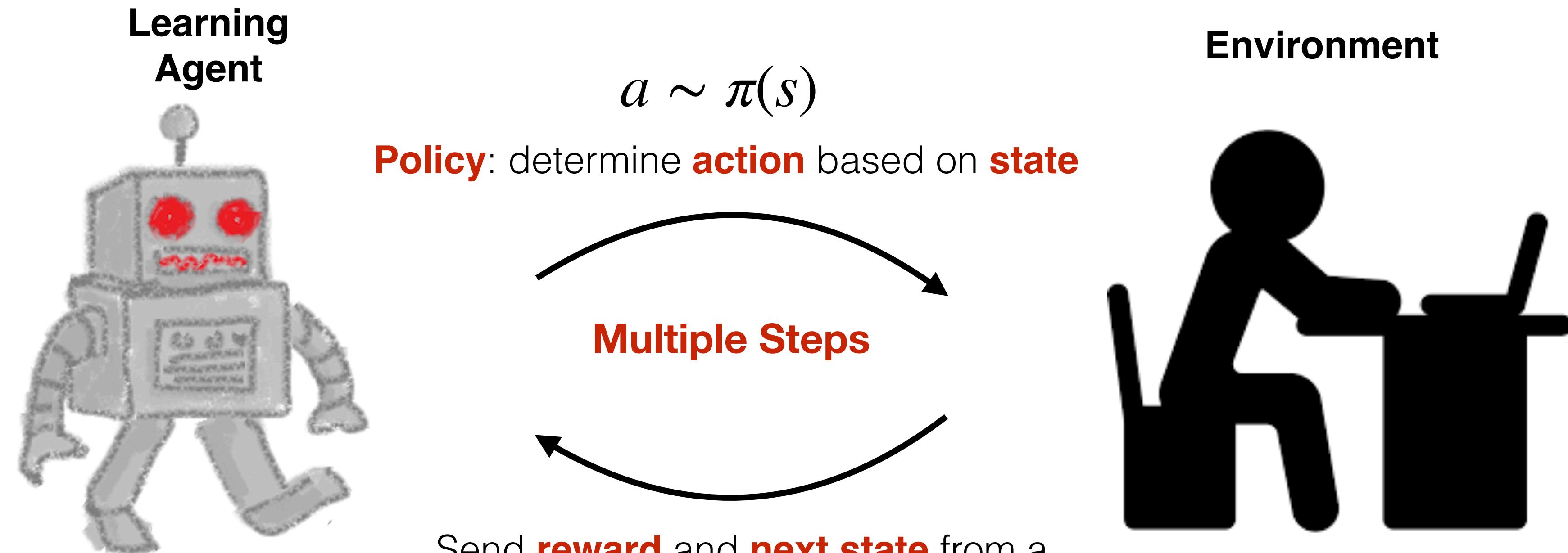
$$r(s, a), s' \sim P(\cdot | s, a)$$

Markov Decision Process



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

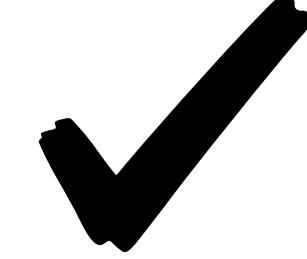
Markov Decision Process



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

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

	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	✓	✓	✓		

	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

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

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

Infinite horizon Discounted MDP

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

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

$$\text{Policy } \pi : S \mapsto \Delta(A)$$

Infinite horizon Discounted MDP

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

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

$$\text{Policy } \pi : S \mapsto \Delta(A)$$

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

Infinite horizon Discounted MDP

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

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

$$\text{Policy } \pi : S \mapsto \Delta(A)$$

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

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

Bellman Equation:

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

Bellman Equation:

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

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

Bellman Equation:

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

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

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

Bellman Equation:

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

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

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

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

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 1.7 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 1.7 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 1.7 in the RL monograph]

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

Theorem 1: Bellman Optimality

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

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Proof of Bellman Optimality

Theorem 1: Bellman Optimality

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

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

Proof of Bellman Optimality

Theorem 2:

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

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

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

Proof of Bellman Optimality

Theorem 2:

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

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

Proof of Bellman Optimality

Theorem 2:

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

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

Proof of Bellman Optimality

Theorem 2:

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

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

Proof of Bellman Optimality

Theorem 2:

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

$$\begin{aligned} |V(s) - V^\star(s)| &= \left| \max_a (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V(s')) - \max_a (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^\star(s')) \right| \\ &\leq \max_a \left| (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V(s')) - (r(s, a) + \gamma \mathbb{E}_{s' \sim P(s, a)} V^\star(s')) \right| \\ &\leq \max_a \gamma \mathbb{E}_{s' \sim P(s, a)} |V(s') - V^\star(s')| \\ &\leq \max_a \gamma \mathbb{E}_{s' \sim P(s, a)} \left(\max_{a'} \gamma \mathbb{E}_{s'' \sim P(s', a')} |V(s'') - V^\star(s'')| \right) \\ &\leq \max_{a_1, a_2, \dots, a_{k-1}} \gamma^k \mathbb{E}_{s_k} |V(s_k) - V^\star(s_k)| \end{aligned}$$

Outline

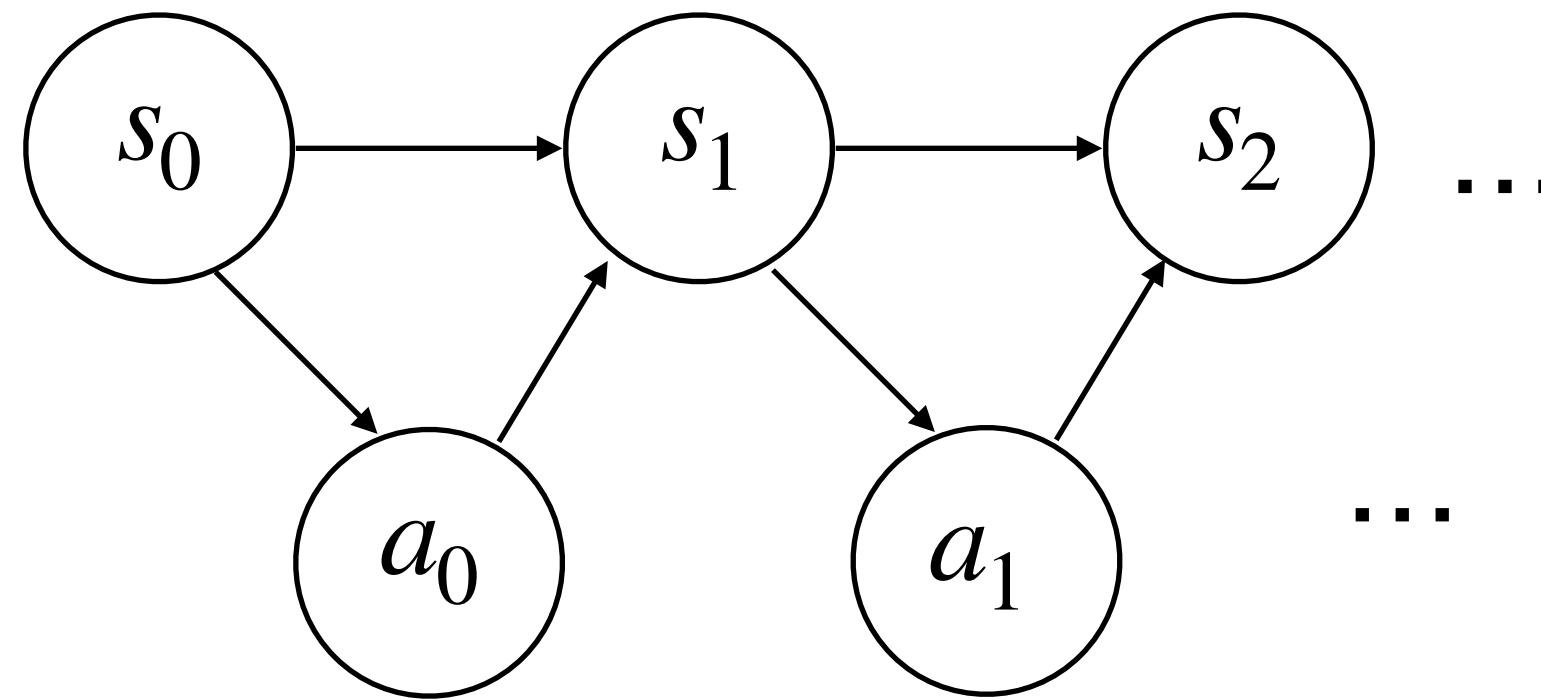
- ✓ 1. Definition of infinite horizon discounted MDPs
- ✓ 2. Bellman Optimality
- 3. State-action distribution

Trajectory distribution and state-action distribution

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

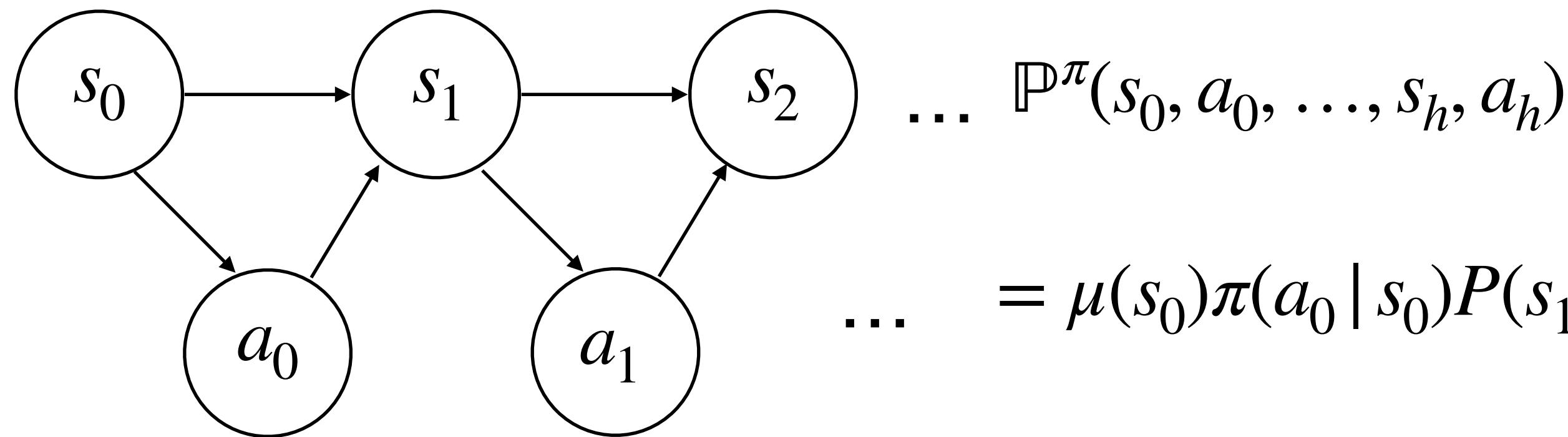
Trajectory distribution and state-action distribution

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

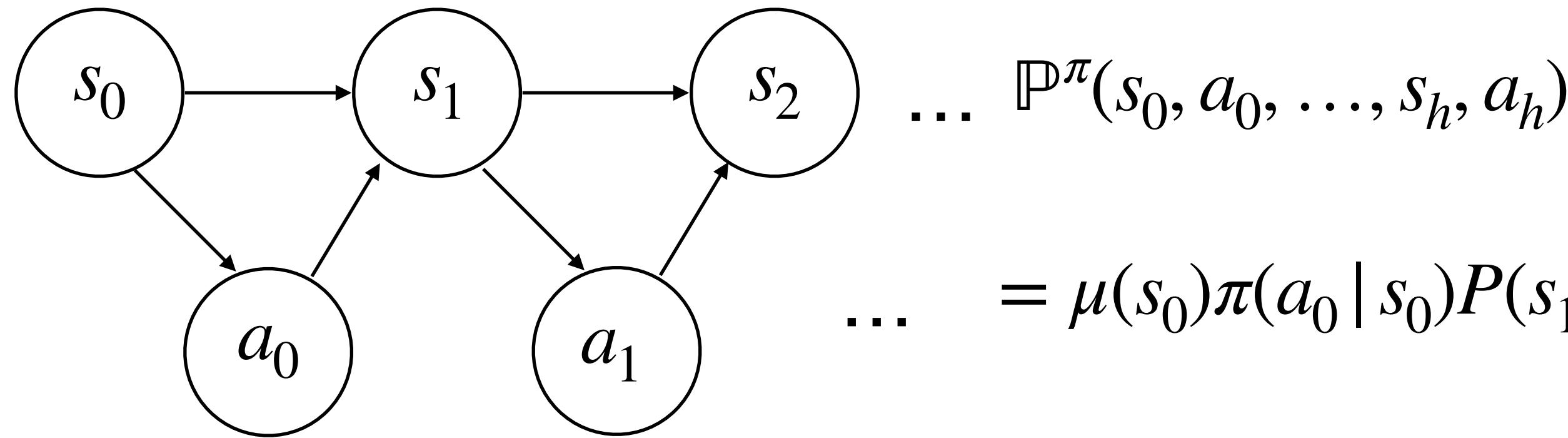


$$\dots \mathbb{P}^\pi(s_0, a_0, \dots, s_h, a_h)$$

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

Trajectory distribution and state-action distribution

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



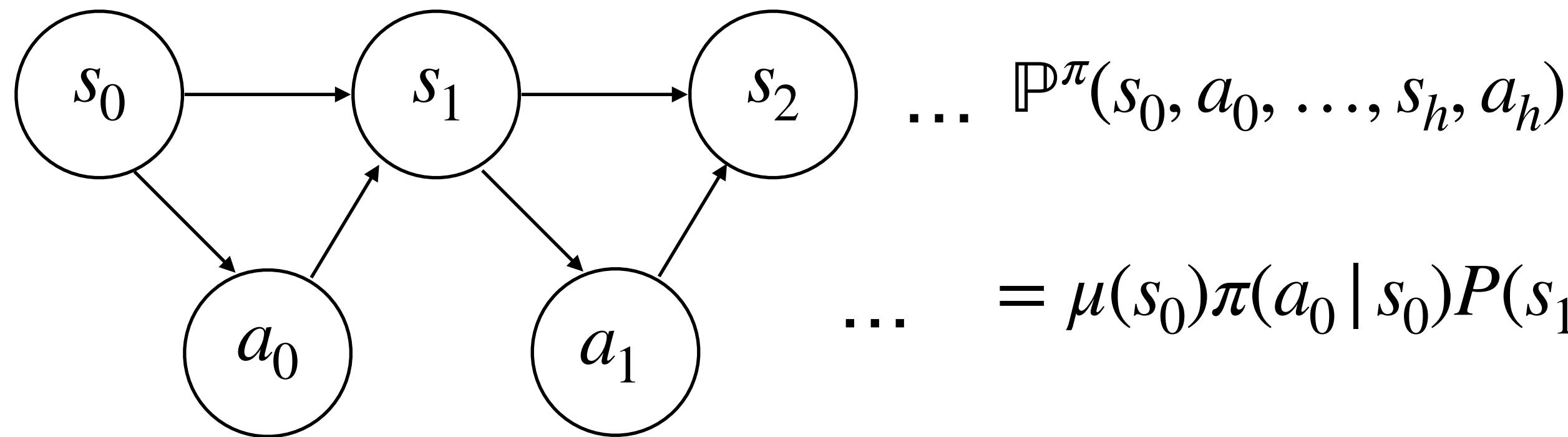
$$\dots \mathbb{P}^{\pi}(s_0, a_0, \dots, s_h, a_h)$$

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

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

Trajectory distribution and state-action distribution

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



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

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

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

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

Average State-Action occupancy measure

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

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

Average 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)