Reinforcement Learning & Markov Decision Processes

Lucas Janson and Sham Kakade

CS/Stat 184: Introduction to Reinforcement Learning

Fall 2022

Today

- HW 1 due this thurs
- Today: what is Reinforcement Learning?
  - examples/concepts
  - definition of Markov Decision Processes

Today:

Intro to Markov Decision Processes

Four main themes we will cover in this course:

1. Bandits (horizon $H = 1$)
2. Two models, with horizon $H > 1$:

   - Markov Decision Process: Dynamic Programming & planning

   - Continuous Control

   (technically, this is still an MDP, but with special structure)

3. Learning in “Large” Markov Decision Process
4. Advanced Topics

Supplementary Reading Materials: Reinforcement Learning: Theory & Algorithms

https://rltheorybook.github.io/

This is an advanced RL book.
We will pick **specific subsections,** to further your knowledge.

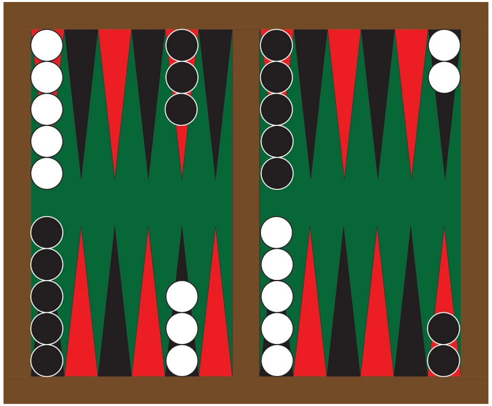
Please let us know if you find any typos or mistakes in the book

Outlines:

1. Introduction: Applications of RL, RL versus Supervised Learning

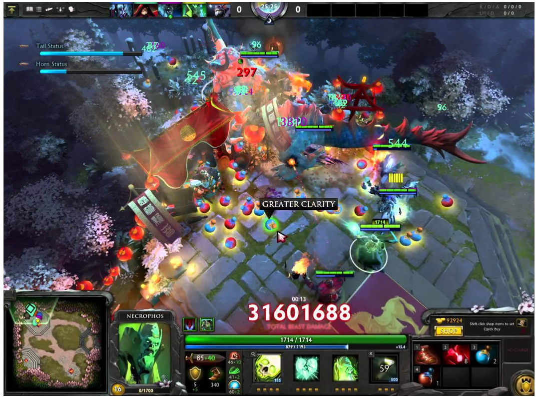
2. Basics of Markov Decision Process (MDP): model, example, V & Q functions

Big Successful Stories of Reinforcement Learning



TD GAMMON [Tesauro 95]

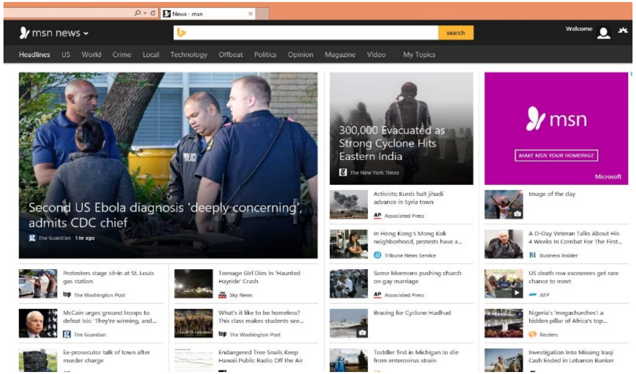
[AlphaZero, Silver et.al, 17]



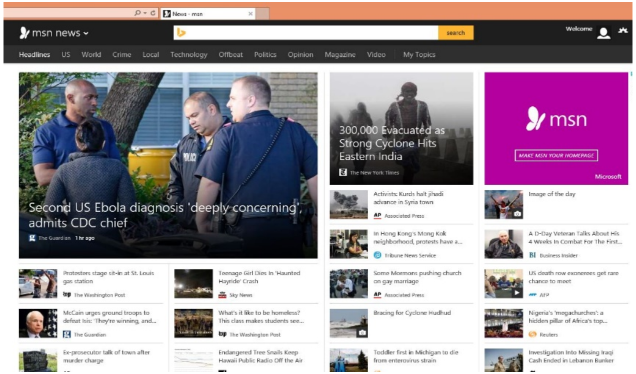
[OpenAI Five, 18]

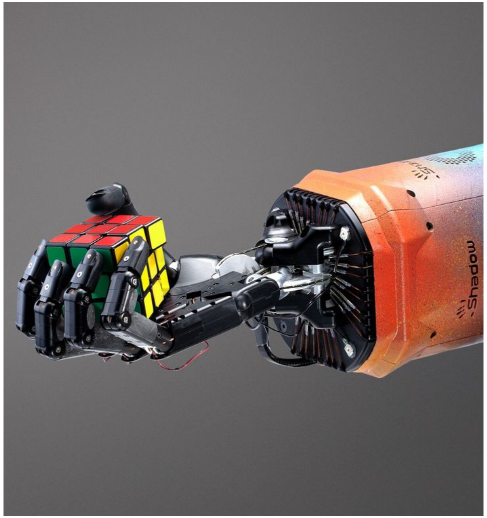
Reinforcement Learning in Real World:

Reinforcement Learning in Real World:



Reinforcement Learning in Real World:





To better understand RL,
let’s summarize “supervised learning”

Recap: Supervised Learning

Recap: Supervised Learning

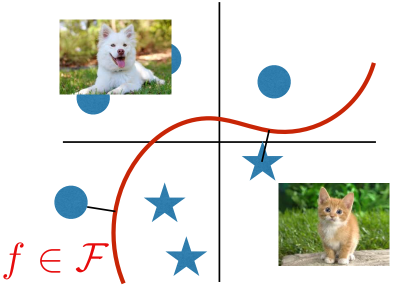
Given i.i.d examples at training:

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

Recap: Supervised Learning

Given i.i.d examples at training:

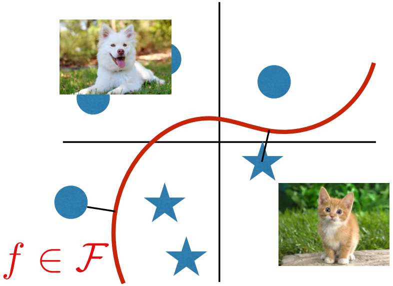
( ,cat) ( ,cat) ( ,dog)



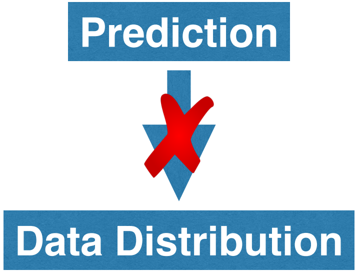
Recap: Supervised Learning

Given i.i.d examples at training:

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



Passive:



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

Summary so far:

1. In RL, we often start from zero data

Summary so far:

1. **In RL, we often start from zero data**

2. In RL, **decisions/predictions have consequences:**
Future data is determined by our past historical decisions/predictions

Summary so far:

1. **In RL, we often start from zero data**

2. In RL, **decisions/predictions have consequences:**
Future data is determined by our past historical decisions/predictions

3. To solve the task, we often need to make a **long sequence of decisions**

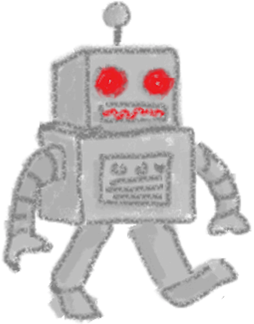
Outlines:

1. Introduction: Applications of RL, RL versus Supervised Learning

2. Basics of Markov Decision Process (MDP): model, example, V & Q functions

The Mathematical framework:
Markov Decision Process

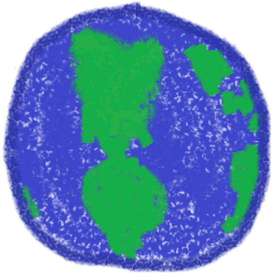
Learning Agent



$$\pi(s) \rightarrow a$$

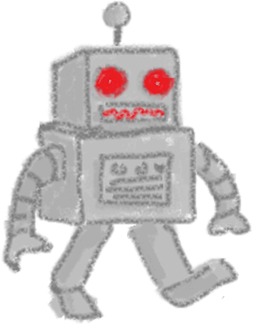
Policy: determine **action** based on **state**

Environment



The Mathematical framework:
Markov Decision Process

Learning Agent



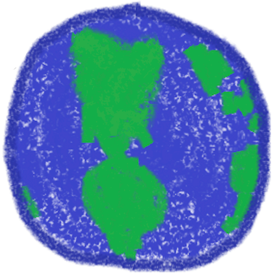
$$\pi(s) \rightarrow a$$

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

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

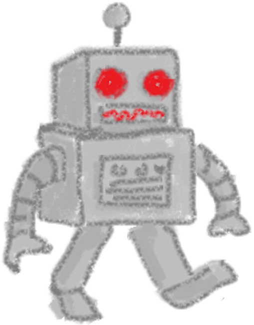
$$r(s, a), s' \sim P(\cdot \,|\, s, a)$$

Environment



The Mathematical framework:
Markov Decision Process

Learning Agent

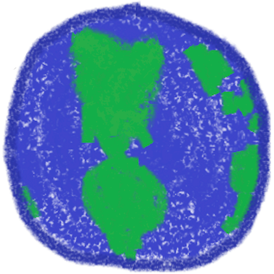


$$\pi(s) \rightarrow a$$

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

Multiple Steps

Environment

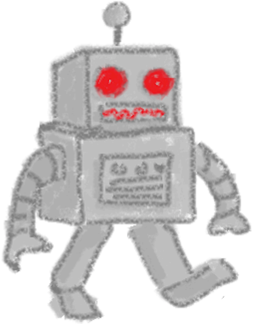


Send **reward** and **next state** from a Markovian transition dynamics

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

The Mathematical framework:
Markov Decision Process

Learning Agent



Environment

$$\pi(s) \rightarrow a$$

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

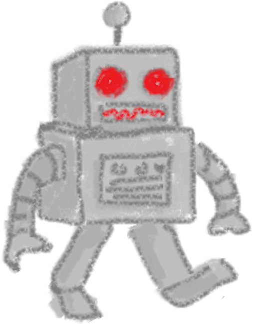
Multiple Steps

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

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

The Mathematical framework:
Markov Decision Process

Learning Agent



$$\pi(s) \rightarrow a$$

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

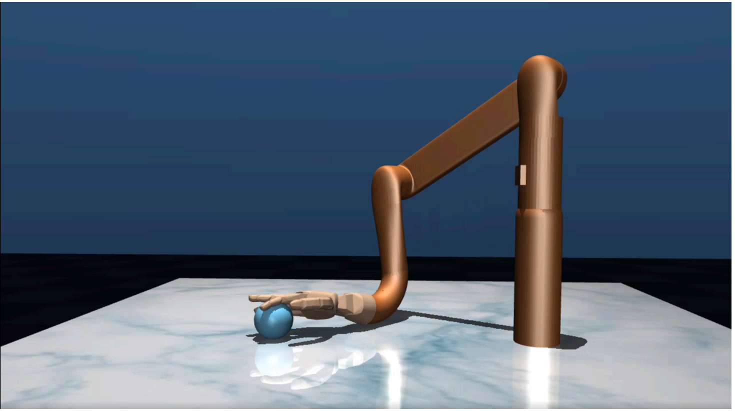
Multiple Steps

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

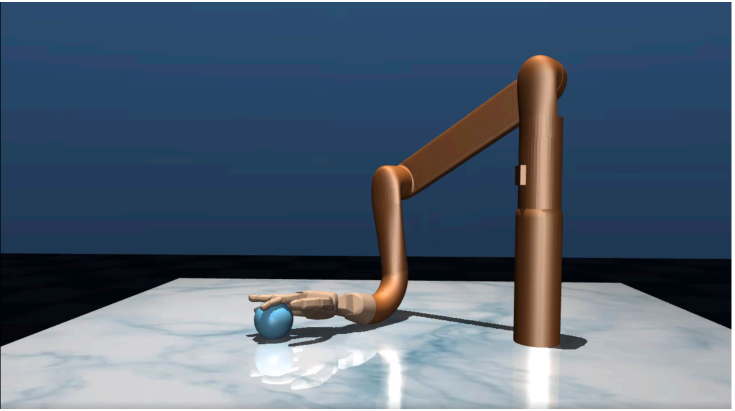
$$r(s,a), s' \sim P(\cdot \mid s,a)$$

Environment

Example:
robot hand needs to pick the ball and hold it in a goal (x,y,z) position

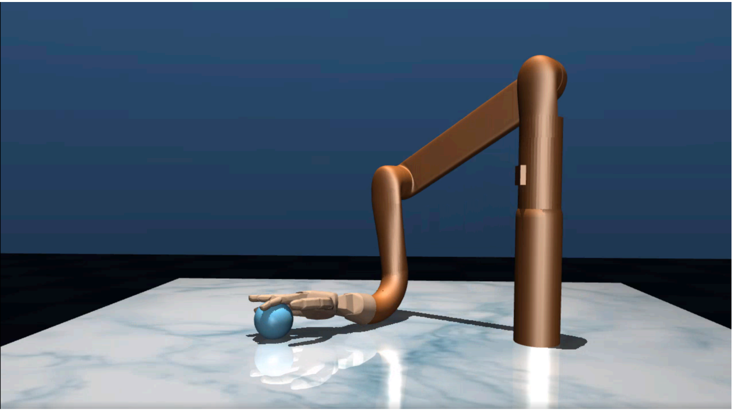


Example:
robot hand needs to pick the ball and hold it in a goal (x,y,z) position



State s: robot configuration (e.g., joint angles) and the ball's position

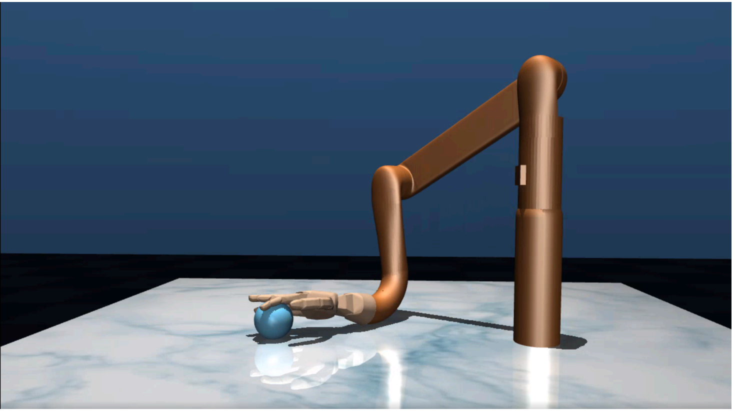
Example:
robot hand needs to pick the ball and hold it in a goal (x,y,z) position



State s: robot configuration (e.g., joint angles) and the ball's position

Action a: Torque on joints in arm & fingers

Example:
robot hand needs to pick the ball and hold it in a goal (x,y,z) position

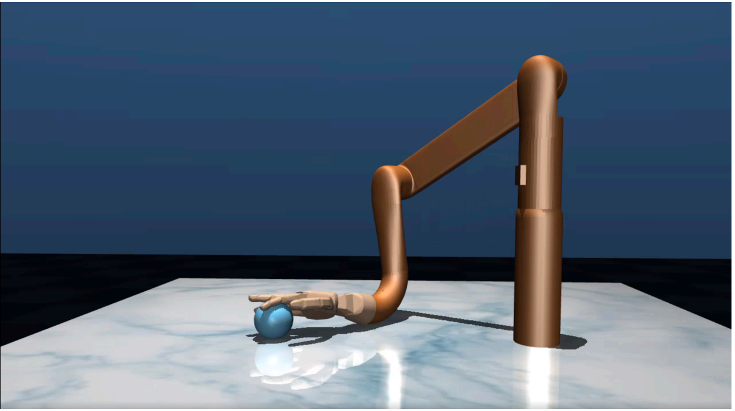


State s: robot configuration (e.g., joint angles) and the ball's position

Action a: Torque on joints in arm & fingers

Transition $s' \sim P(\cdot \mid s, a)$: physics + some noise

Example:
robot hand needs to pick the ball and hold it in a goal (x,y,z) position



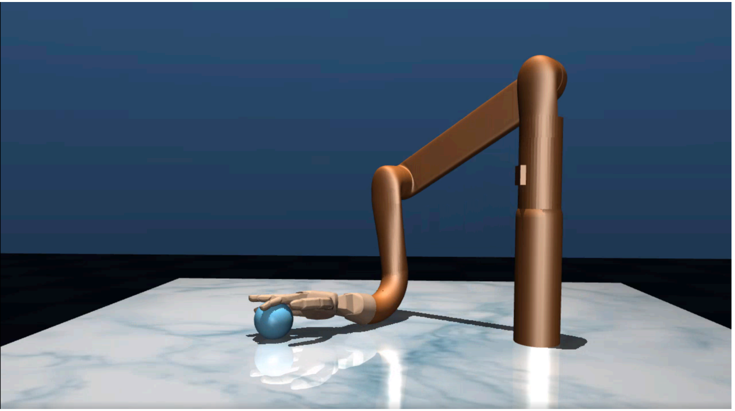
State s: robot configuration (e.g., joint angles) and the ball's position

Action a: Torque on joints in arm & fingers

Transition $s' \sim P(\cdot \,|\, s, a)$: physics + some noise

policy $\pi(s)$: a function mapping from robot state to action (i.e., torque)

Example:
robot hand needs to pick the ball and hold it in a goal (x,y,z) position



State s: robot configuration (e.g., joint angles) and the ball's position

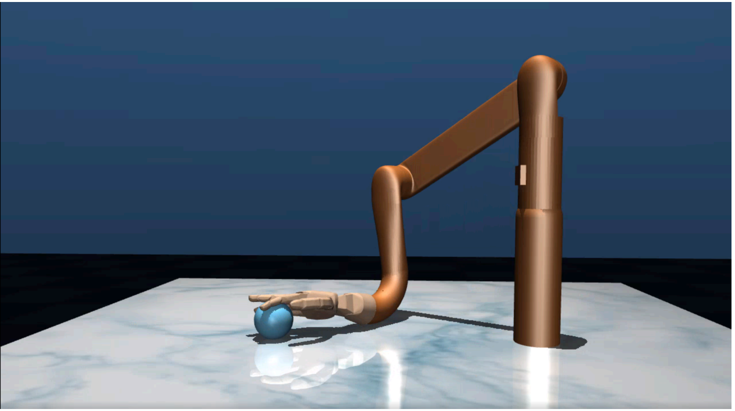
Action a: Torque on joints in arm & fingers

Transition $s' \sim P(\cdot \mid s, a)$: physics + some noise

policy $\pi(s)$: a function mapping from robot state to action (i.e., torque)

Cost $c(s, a)$: torque magnitude + dist to goal

Example:
robot hand needs to pick the ball and hold it in a goal (x,y,z) position



State s: robot configuration (e.g., joint angles) and the ball's position

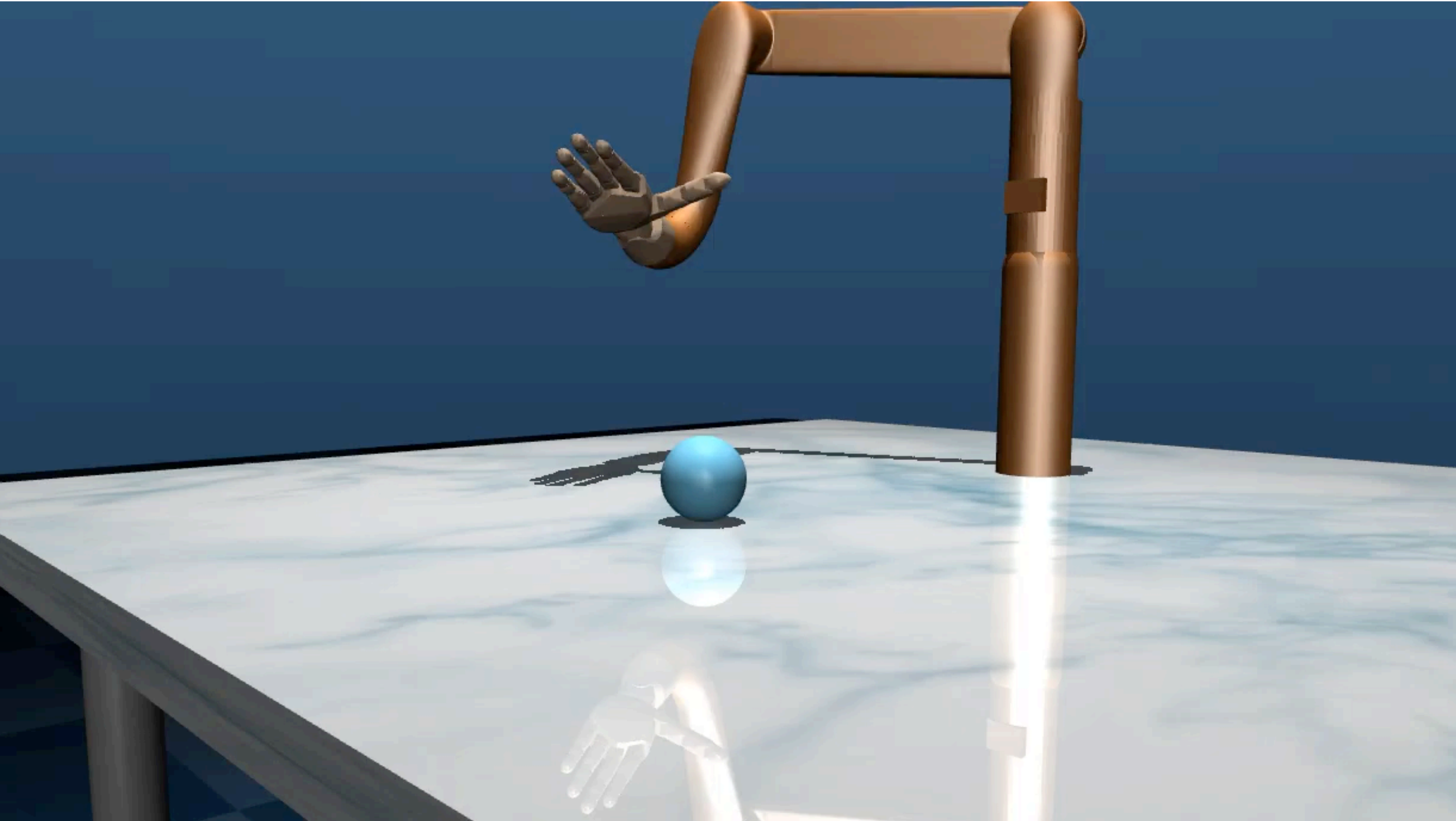
Action a: Torque on joints in arm & fingers

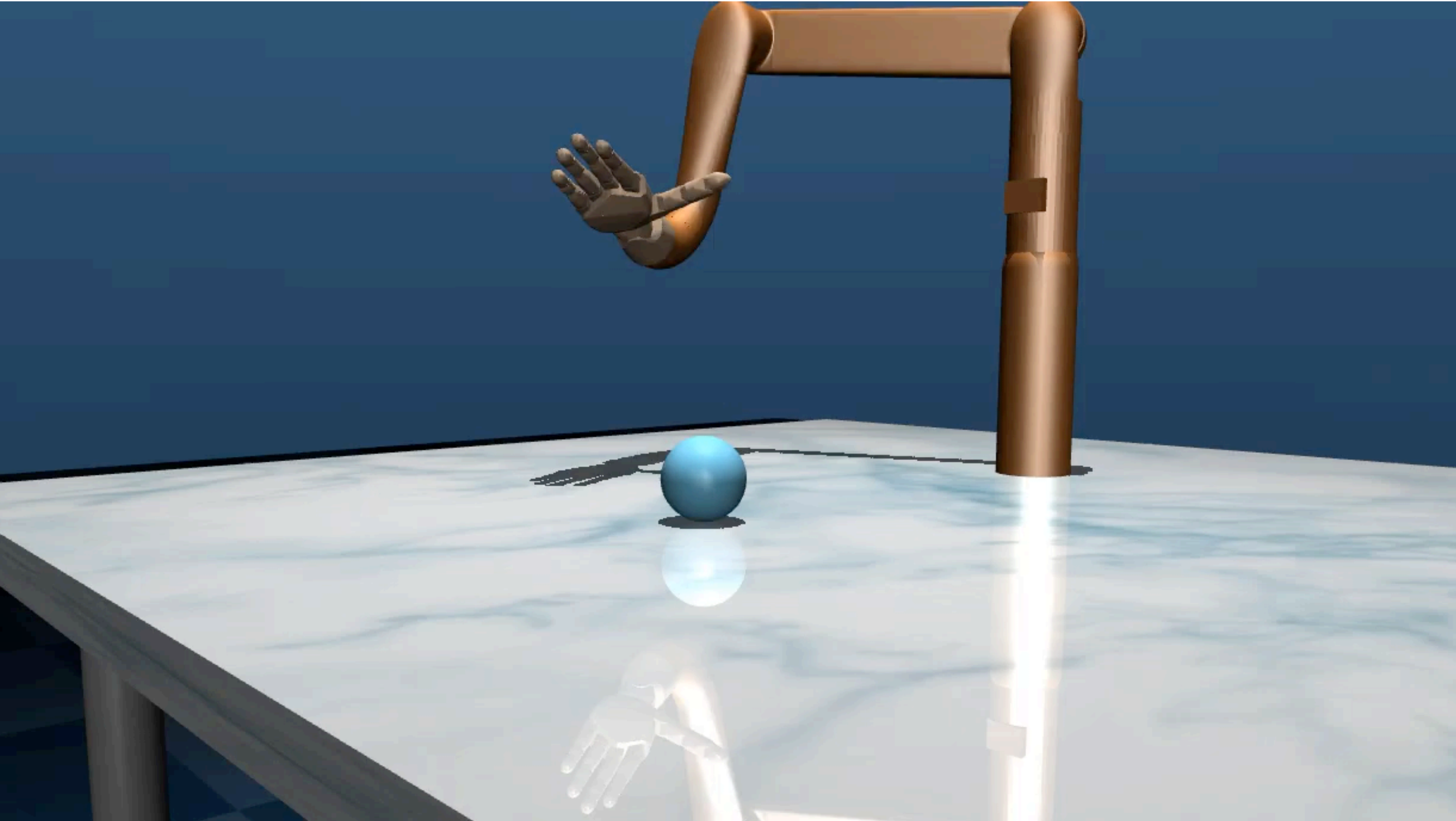
Transition $s' \sim P(\cdot \mid s, a)$: physics + some noise

policy $\pi(s)$: a function mapping from robot state to action (i.e., torque)

Cost $c(s, a)$: torque magnitude + dist to goal

$$\pi^\star = \arg\min_{\pi} \mathbb{E}\left[c(s_0, a_0) + \gamma c(s_1, a_1) + \gamma^2 c(s_2, a_2) + \gamma^3 c(s_3, a_3) + \ldots \ldots \middle| s_0, \pi\right]$$





MDPs, more formally:

MDPs, more formally:

- An MDP: $\mathscr{M} = \{S, A, P, r, \gamma\}$

MDPs, more formally:

- An MDP: $\mathscr{M} = \{S, A, P, r, \gamma\}$
  - S a set of states

MDPs, more formally:

- An MDP: $\mathscr{M} = \{S, A, P, r, \gamma\}$
  - S a set of states
  - A a set of actions

MDPs, more formally:

- An MDP: $\mathscr{M} = \{S, A, P, r, \gamma\}$
  - S a set of states
  - A a set of actions
  - $P : S \times A \mapsto \Delta(S)$ specifies the dynamics model,
    i.e. $P(s' \mid s, a)$ is the probability of transitioning to s' form states s under action a

MDPs, more formally:

- An MDP: $\mathscr{M} = \{S, A, P, r, \gamma\}$
  - S a set of states
  - A a set of actions
  - $P : S \times A \mapsto \Delta(S)$ specifies the dynamics model, i.e. $P(s' \mid s, a)$ is the probability of transitioning to s' form states s under action a
  - $r : S \times A \to [0,1]$
    - let's assume this is deterministic
    - (sometimes we use a cost $c : S \times A \to [0,1]$)

• easy to extend MDPs to have stochastic reward functions.

• in bandits, r was stochastic.

MDPs, more formally:

$s_0 \}$

- An MDP: $\mathcal{M} = \{S, A, P, r, \gamma\}$
  - S a set of states
  - A a set of actions
  - $P : S \times A \mapsto \Delta(S)$ specifies the dynamics model, i.e. $P(s' \mid s, a)$ is the probability of transitioning to s' form states s under action a
  - $r : S \times A \rightarrow [0,1]$
    - let's assume this is deterministic
    - (sometimes we use a cost $c : S \times A \rightarrow [0,1]$)
  - A discount factor $\gamma \in [0,1)$

- sometimes we often specify a starting state s_0

The Objective

The Objective

- A "stationary" policy $\pi : S \mapsto A$ ⟵ suppose this is deterministic.
  - "stationary" means not history dependent
  - we could also consider π to be random and a function of the history

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

The Objective

$\{(s_0, a_0, r_0), (s_1, a_1, r_1), \ldots\}$

- A "stationary" policy $\pi : S \mapsto A$
  - "stationary" means not history dependent
  - we could also consider π to be random and a function of the history
- Sampling a trajectory: from a given policy π starting at state s_0:
  - For $t = 0,1,2,\ldots\infty$
    - Take action $a_t = \pi(s_t)$
    - Observe reward $r_t = r(s_t, a_t)$
    - Transition to (and observe) s_{t+1} where $s_{t+1} \sim P(\cdot \mid s_t, a_t)$

$s_0, a_0 = \pi(s_0), r(s_0, a_0),$

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

$s_1, a_1 = \pi(s_1), r(s_1, a_1)$

The Objective

- A "stationary" policy $\pi : S \mapsto A$
  - "stationary" means not history dependent
  - we could also consider π to be random and a function of the history
- Sampling a trajectory: from a given policy π starting at state s_0:
  - For $t = 0,1,2,\ldots\infty$
    - Take action $a_t = \pi(s_t)$
    - Observe reward $r_t = r(s_t, a_t)$
    - Transition to (and observe) s_{t+1} where $s_{t+1} \sim P(\cdot \mid s_t, a_t)$
- Objective: given state starting state s,
  find a policy π that maximizes our expected, discounted future reward:

$$\max_{\pi} \mathbb{E}\left[r(s_0, a_0) + \gamma r(s_1, a_1) + \gamma^2 r(s_2, a_2) + \ldots\ldots \,\Big|\, s_0 = s, \pi\right]$$

"effective horizon" $\approx \frac{1}{1-\gamma}$

H-finite horizon

alternative objective

$$\max_{\pi} E\left[r(s_0, a_0) + r(s_1, a_1) + \ldots r(s_H, a_H) \,\Big|\, s_0 = s, \pi\right]$$

Question:

Assume we have $|S|$ many states, and $|A|$ many actions, how many different polices there are?

stationary

Question:

Assume we have $|S|$ many states, and $|A|$ many actions, how many different polices there are?

(Hint: a policy is a mapping from s to a, we have A many choices per state s)

possible det. stat. policies

is $|A|^{|S|}$

Infinite horizon Discounted Setting

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

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

Policy $\pi : S \mapsto A$

Infinite horizon Discounted Setting

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

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

Policy $\pi : S \mapsto A$

Quantities that allow us to reason policy's long-term effect:

Infinite horizon Discounted Setting

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

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

Policy $\pi : S \mapsto A$

depends on π

Quantities that allow us to reason policy's long-term effect:

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

Infinite horizon Discounted Setting

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

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

Policy $\pi : S \mapsto A$

Quantities that allow us to reason policy's long-term effect:

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

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

Understanding Value function and Q functions

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

start s_0, take $a_0 = \pi(s_0)$, $r(s_0, a_0)$, $s_1 \sim P(\cdot | s_0, a_0)$

s_1, take $a_1 = \pi(s_1)$, $r(s_1, a_1)$, $s_2 \sim P(\cdot | s_1, a_1)$

Understanding Value function and Q functions

suppose $\exists \tilde{a}$,

$Q^{\pi}(s, \tilde{a}) > V^{\pi}(s)$

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

possible $a \neq \pi(s)$

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

start s_0, take $a_0 = a$, $r(s_0, a_0)$, $s_1 \sim P(\cdot \mid s_0, a_0)$

s_1, take $a_1 = \pi(s_1)$, $r(s_1, a_1)$, $s_2 \sim P(\cdot \mid s_1, a_1)$

Bellman Consistency Equation for V-function:

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

for stationary deterministic π.

We have that:

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

$s' \sim P(\cdot \mid s, \pi(s))$

Proof: Bellman Consistency for V-function:

Proof: Bellman Consistency for V-function:

- By definition:

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

$s_1 \sim P(\cdot \mid s, \pi(s))$

$$= r(s, \pi(s)) + \gamma \mathbb{E}\left[r(s_1, a_1) + \gamma r(s_2, a_2) + \ldots \,\middle|\, s_0 = s, \pi\right]$$

Proof: Bellman Consistency for V-function:

$\mathbb{E}_{x\sim D}[f(x)] = \sum_{x\in Domain} D(x)f(x)$

- By definition:

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

$$= r(s,\pi(s)) + \gamma\mathbb{E}\left[r(s_1,a_1) + \gamma r(s_2,a_2) + \ldots \,\middle|\, s_0 = s, \pi\right]$$

* by Markov

- By the "tower property" and due to that $s_1 = s'$ with probability $P(s' \mid s, \pi(s))$,

$$= r(s,\pi(s)) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,\pi(s))}\left[\mathbb{E}\left[r(s_1,a_1) + \gamma r(s_2,a_2) + \ldots \,\middle|\, s_0 = s, s_1 = s', \pi\right]\right]$$

$$= r(s,\pi(s)) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,\pi(s))}\left[\mathbb{E}\left[r(s_1,a_1) + \gamma r(s_2,a_2) + \ldots \,\middle|\, s_1 = s', \pi\right]\right]$$

by def.

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

$\sum_{s'\in S} P(s'|s,a)V^\pi(s')$

Bellman Consistency Equation for Q-function:

Bellman Consistency Equation for Q-function:

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

$$V^\pi(s) = Q^\pi(s, \pi(s))$$

Bellman Consistency Equation for Q-function:

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

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

P_{sa}

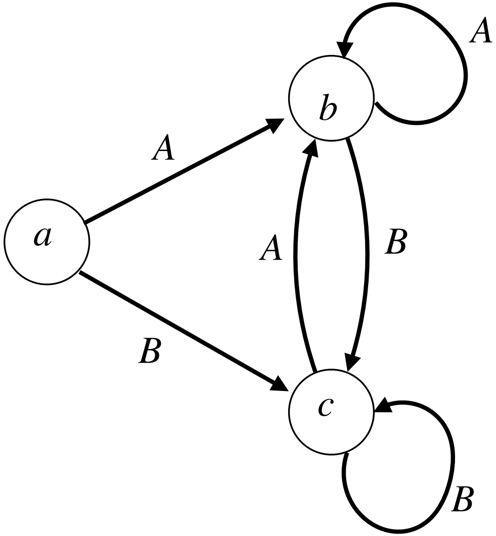
$P(s'|s,a)$

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

$f(x)$, $f(\cdot)$

Example of Optimal Policy $\pi^\star$

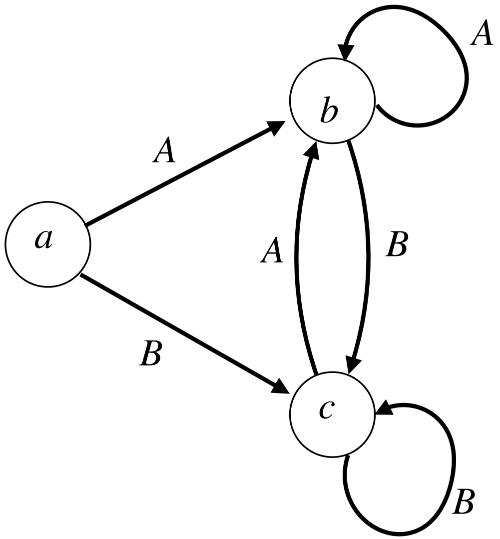
Consider the following **deterministic** MDP w/ 3 states & 2 actions



Reward: $r(b, A) = 1$, & 0 everywhere else

Example of Optimal Policy $\pi^\star$

Consider the following **deterministic** MDP w/ 3 states & 2 actions

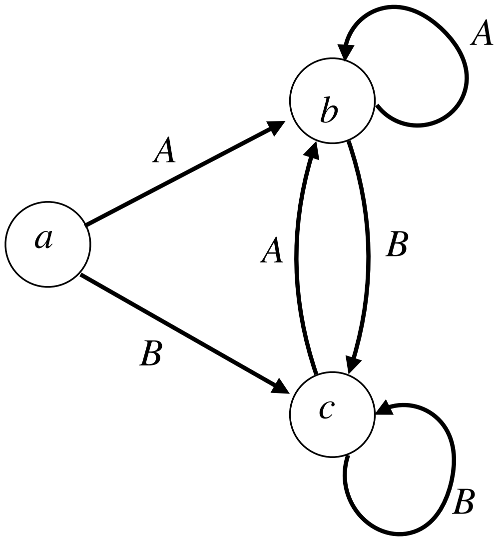


Let's say $\gamma \in (0,1)$
What's the optimal policy?

Reward: $r(b, A) = 1$, & 0 everywhere else

Example of Optimal Policy $\pi^\star$

Consider the following **deterministic** MDP w/ 3 states & 2 actions



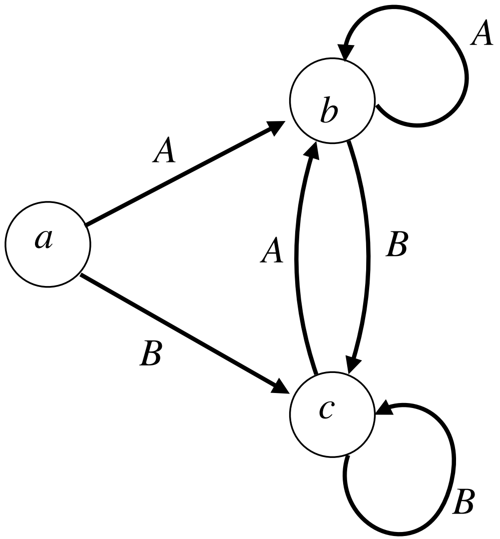
Let's say $\gamma \in (0,1)$
What's the optimal policy?

$$\pi^\star(s) = A, \forall s$$

Reward: $r(b, A) = 1$, & 0 everywhere else

Example of Optimal Policy $\pi^\star$

Consider the following **deterministic** MDP w/ 3 states & 2 actions



Reward: $r(b, A) = 1$, & 0 everywhere else

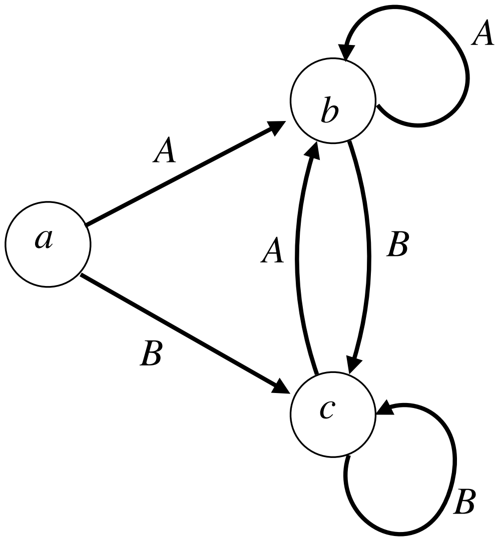
Let's say $\gamma \in (0,1)$
What's the optimal policy?

$$\pi^\star(s) = A, \forall s$$

$$V^\star(a) = \frac{\gamma}{1-\gamma}, V^\star(b) = \frac{1}{1-\gamma}, V^\star(c) = \frac{\gamma}{1-\gamma}$$

Example of Optimal Policy $\pi^\star$

Consider the following **deterministic** MDP w/ 3 states & 2 actions



Reward: $r(b, A) = 1$, & 0 everywhere else

Let's say $\gamma \in (0,1)$

What's the optimal policy?

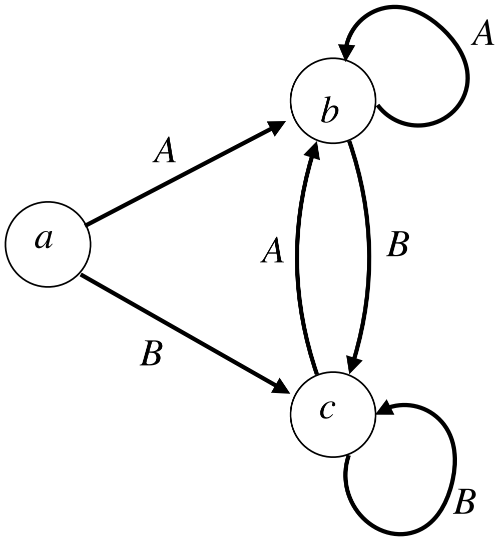
$$\pi^\star(s) = A, \forall s$$

$$V^\star(a) = \frac{\gamma}{1-\gamma}, V^\star(b) = \frac{1}{1-\gamma}, V^\star(c) = \frac{\gamma}{1-\gamma}$$

What about policy $\pi(s) = B, \forall s$

Example of Optimal Policy $\pi^\star$

Consider the following **deterministic** MDP w/ 3 states & 2 actions



Reward: $r(b, A) = 1$, & 0 everywhere else

Let's say $\gamma \in (0,1)$

What's the optimal policy?

$$\pi^\star(s) = A, \forall s$$

$$V^\star(a) = \frac{\gamma}{1-\gamma}, V^\star(b) = \frac{1}{1-\gamma}, V^\star(c) = \frac{\gamma}{1-\gamma}$$

What about policy $\pi(s) = B, \forall s$

$$V^\pi(a) = 0, V^\pi(b) = 0, V^\pi(c) = 0$$

Summary:

- **RL is different from Supervised Learning:**
  - Our actions have consequences
  - Need to make sequence of decisions to complete the task

- **Discounted infinite horizon MDP:**
  - State, action, policy, transition, reward (or cost), discount factor
  - **V function and Q function**
  - Key concept: **Bellman consistency equations**

Summary:

- **RL is different from Supervised Learning:**
  - Our actions have consequences
  - Need to make sequence of decisions to complete the task

- **Discounted infinite horizon MDP:**
  - State, action, policy, transition, reward (or cost), discount factor
  - **V function and Q function**
  - Key concept: **Bellman consistency equations**

1-minute feedback form: https://bit.ly/3RHtlxy