Last Time:

- ☐ Intro to HRI

Lecture 2

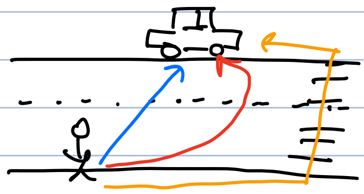
HRI, Fall '24

Andrea Bajcsy

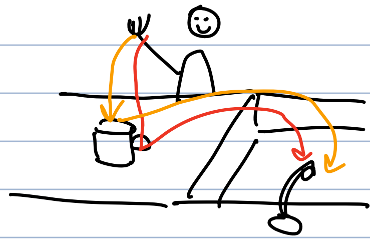
This Time:

- ☐ sequential decision-making
- ☐ MDPs

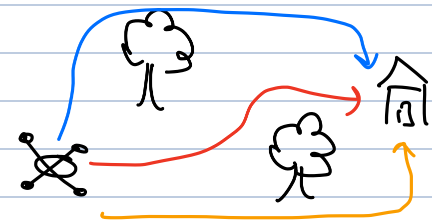
Sequential decision-making is everywhere - play games, making life decisions -- and its present in interaction. In HRI, sequential decision-making will form the "mathematical backbone" of how we model people, robots, and their interaction.



how will this person move to their car?



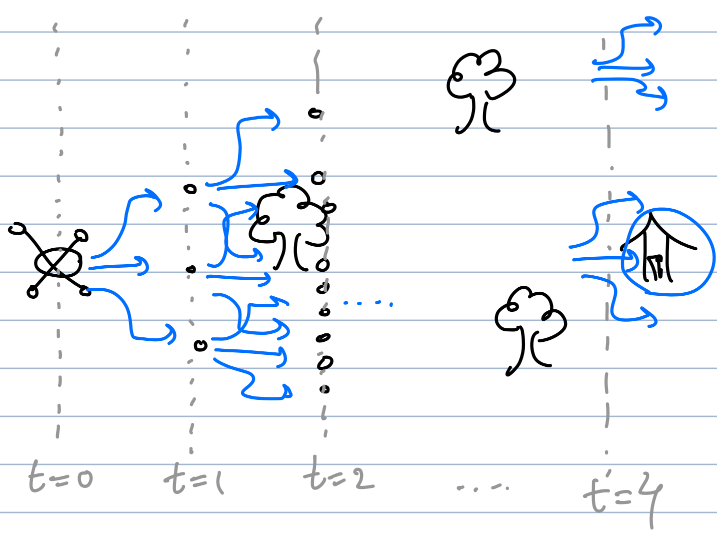
how will this person clean their mug?



how will drone fly home?

Q What makes sequential decision-making hard?

A naïve solution grows exponentially in time horizon.



action space

$|A| = 3$

start from state

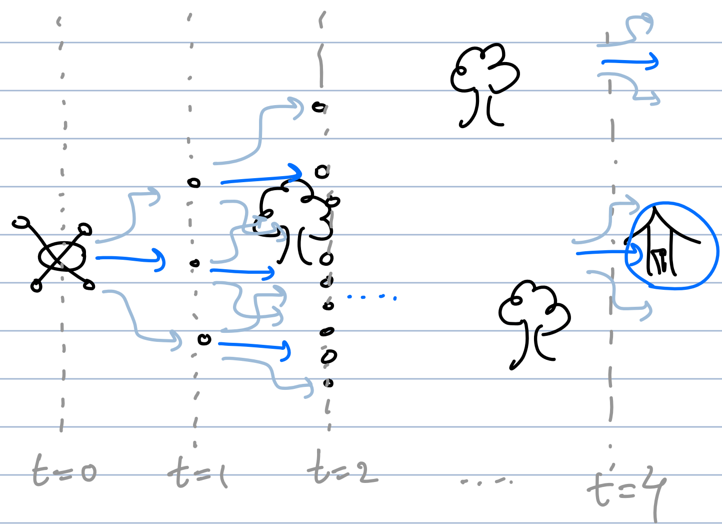
there are:

$|A|^{T+1}$

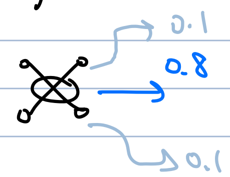
sequences of decisions to choose from.

$= 3^4 = 81$ possible sequences of decisions w/ just 4 seconds

$\boxed{A2}$ Outcomes of taking actions can sometimes be <u>stochastic</u>.



e.g. maybe the drone is less reliable at taking a sharp left or right than straight

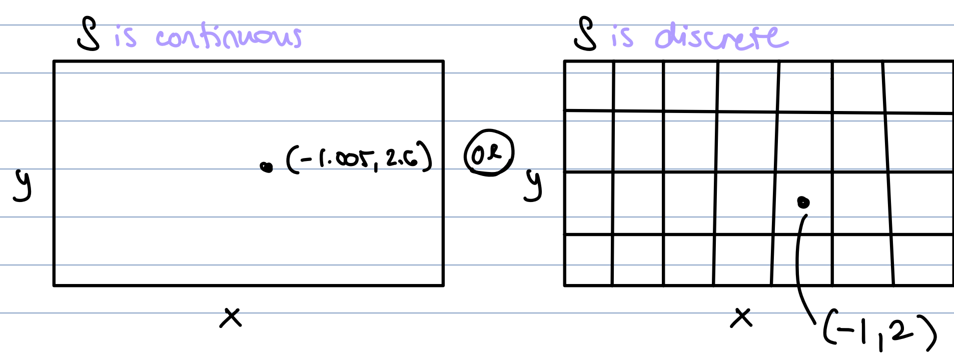


Markov Decision Processes:

this is a mathematical model for sequential decision-making in a fully observable, stochastic environment with a Markovian transition model. lets break down this model into the key "ingredients" and modeling assumptions:

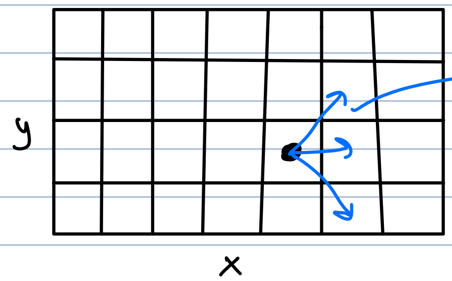
"MDP is a tuple of $< S, A, T, r >$"

- $s \in \underline{\underline{S}}$ = state space.



- $a \in \underline{\underline{A}}$ = <u>action space</u>. This is what our agent can do.

S is discrete



$\equiv A = \{\nearrow, \rightarrow, \searrow\}$

is discrete!

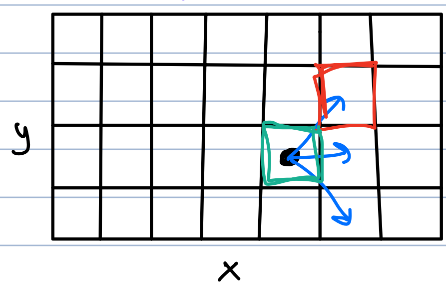
deterministic transition function

- $T : S \times A \rightarrow S$ is the transition function. Is just a map
  $T : S \times A \rightarrow \Delta(S)$

  distribution over next states

  stochastic transition function

from the space of states and actions to a (distribution) over next states.



$$T(s_{t+1} \mid s_t, a_t)$$

$$\underbrace{P(s_{t+1} \mid s_t, a_t)}$$

$$P(s_{t+1} = \square \mid s_t = \square, a_t = \nearrow) = 0.9$$

⊛ In the transition function is precisely where the "Markov" part of Markov Decision-Making Processes comes into play.

Specifically, the Markov Property states that the future is independent of the past, given the present:

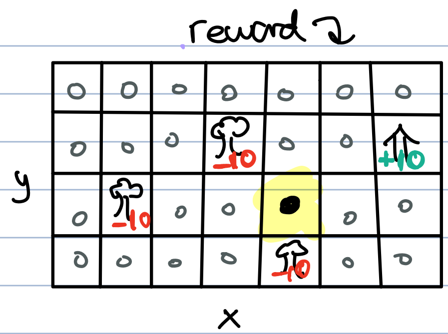
$$P(\text{future} \mid \text{present}, \text{past}) = P(\text{future} \mid \text{present})$$

In MDPs, this translates to the outcome of actions only depending on the current state, and not a history:

$$P(s_{t+1} \mid s_t, a_t, s_{t-1}, a_{t-1}, \dots s_0) = P(s_{t+1} \mid s_t, a_t)$$

by Markov Property!

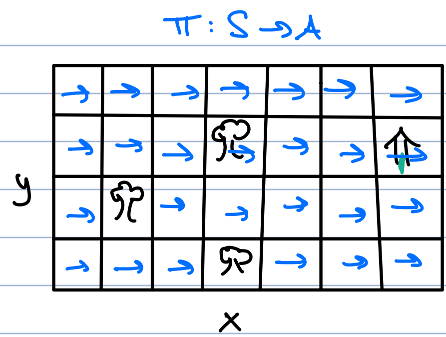
• $r : S \to R$ is the <u>reward function</u>. Its called "instantaneous reward" b/c its only thinking about where you are <u>RIGHT NOW</u>!



we have all these components of what it means to make decisions, but we need a way for our agent to <u>solve</u> or <u>know</u> what the "best" decision is for any state they may be at.

The object we seek to solve for is a <u>policy</u>.

$\pi : S \to A$ is a mapping from states to actions

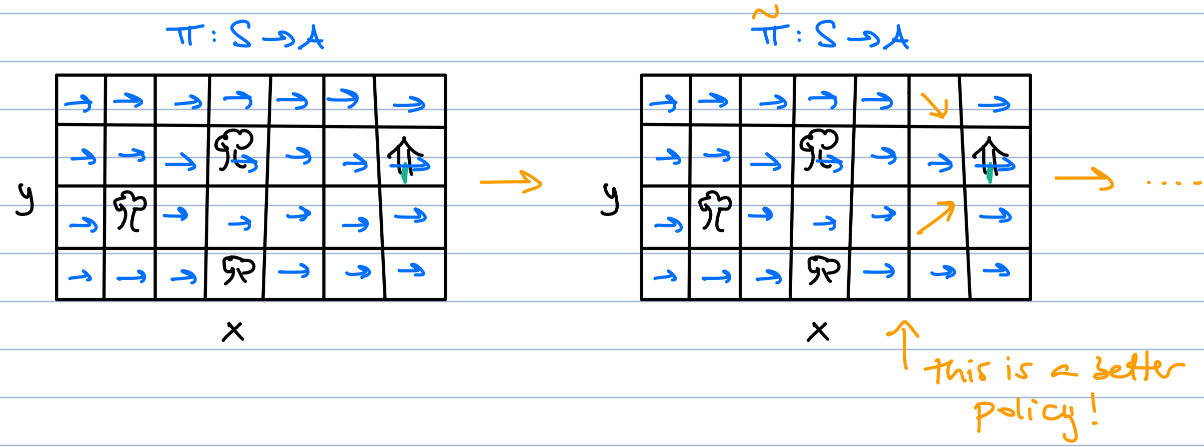


Evaluate the quality of a policy π by the <u>expected cumulative reward</u> induced by the policy.

$$R(s_0, s_1, s_2 \ldots) = r(s_0) + r(s_1) + r(s_2) + \ldots$$

$$= \sum_{t=0} r(s_t)$$

An <u>optimal policy</u> $\pi^*: S \to A$ yields the highest expected reward for all states.



- γ = <u>discount factor</u>, $\gamma \in [0,1]$

$$R(s_0, s_1, s_2 \ldots) = \gamma^0 r(s_0) + \gamma^1 r(s_1) + \gamma^2 r(s_2) + \ldots .$$
$$= \sum_{t=0}^{\infty} \gamma^t r(s_t)$$

Discount factor describes the preference of an agent for <u>current</u> rewards over future ones when $\gamma < 1$. When $\gamma = 1$, then our agent wants the max reward over all time steps.

⊛ discounting appears to be a good model of <u>human</u> + <u>animal</u> preferences over time.

Ultimately, we are searching for a policy $\pi^*: S \to A$ that maximizes the sum of discounted rewards

$$\pi^* := \arg\max_{\pi} \mathbb{E}_{s_{t+1} \sim P(s_{t+1} | s_t, \pi)} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t) \right]$$

<u>MORE INFO:</u>

"Artificial Intelligence: A Modern Approach" by Russel & Norvig.