

Markov Decision Processes

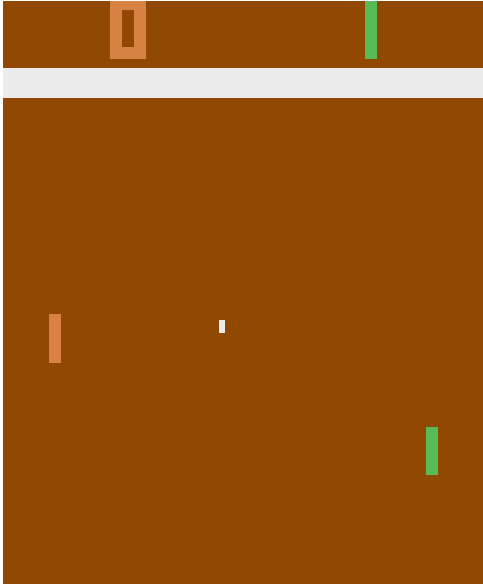
CMPT 729 G100

Jason Peng

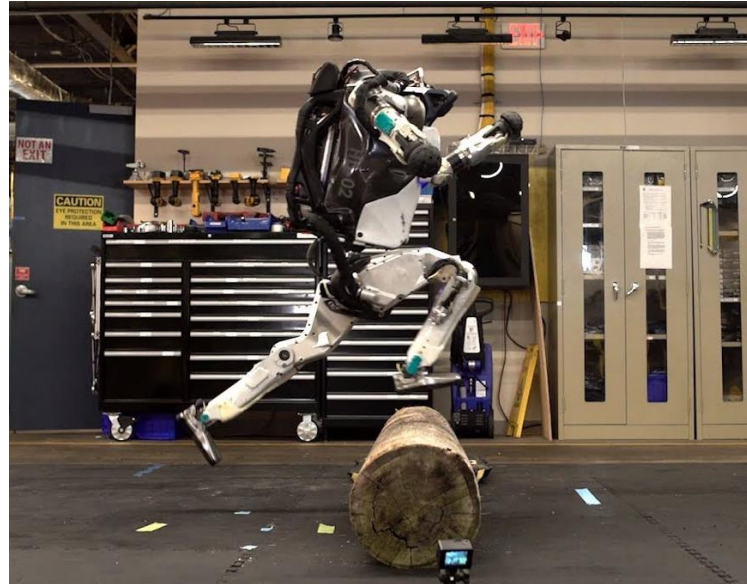
Overview

- Agent-Environment Interface
- Markov Decision Processes
- Partially Observable Markov Decision Processes

Environment Interaction



Pong [Atari]

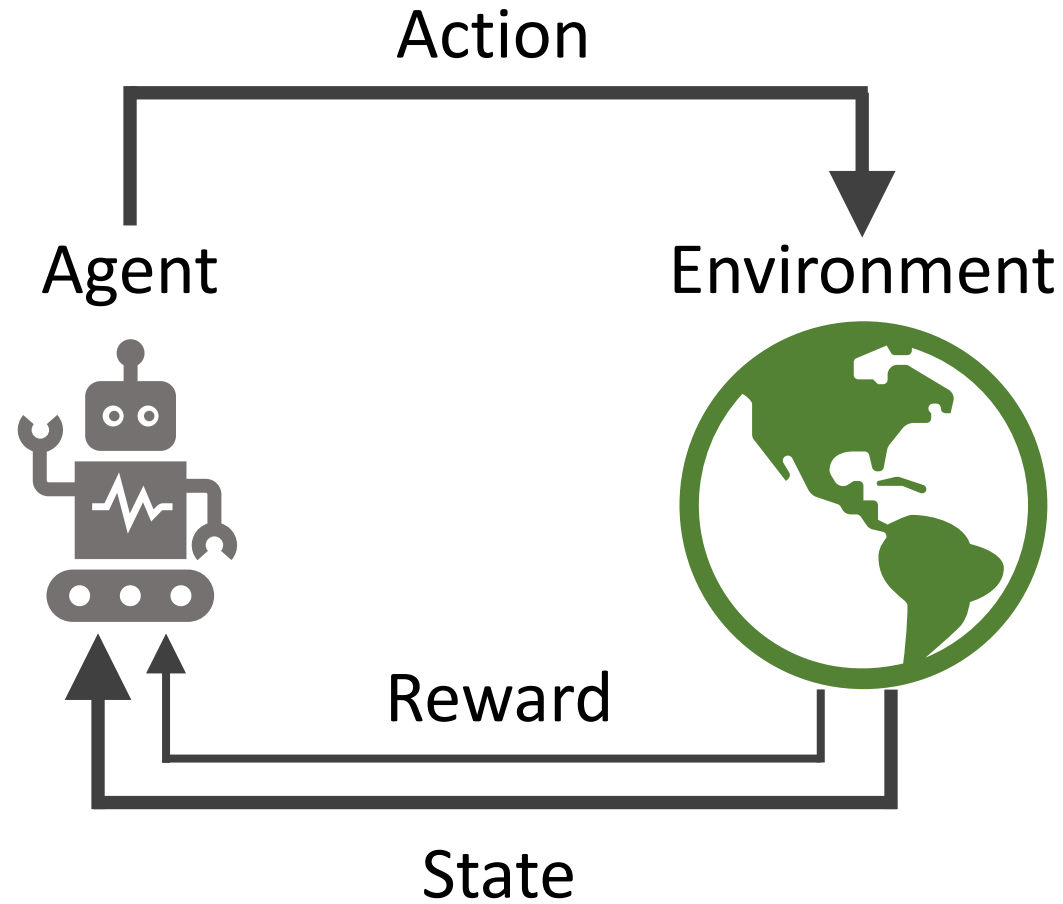


Atlas [Boston Dynamics]



[Smithsonian Magazine]

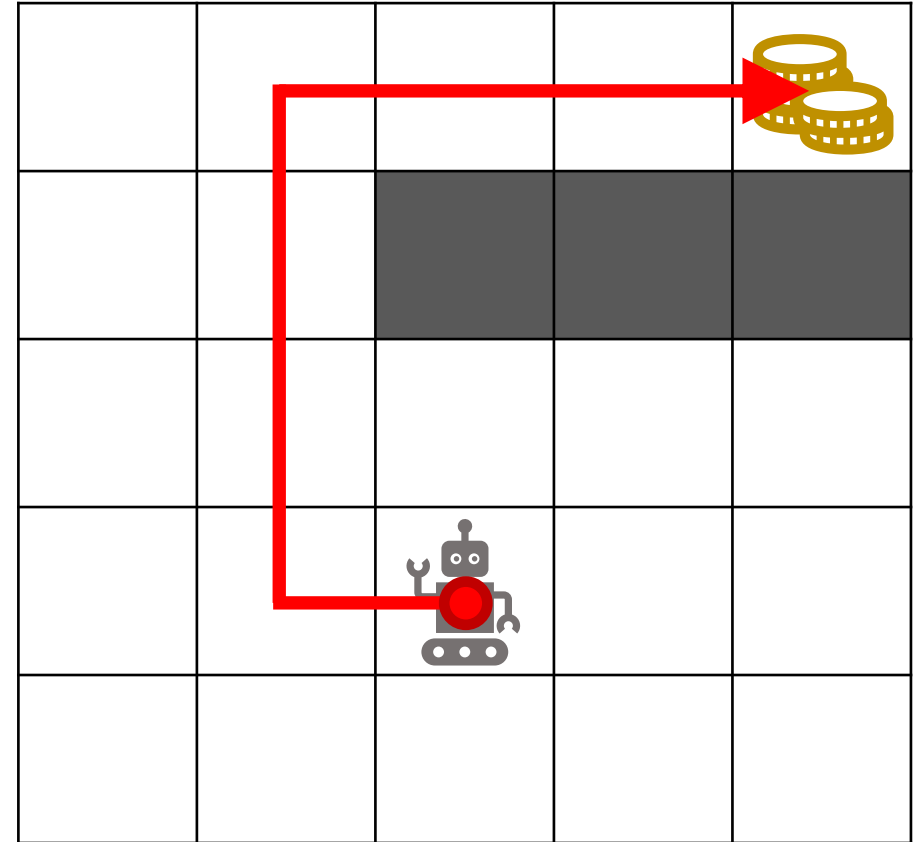
Agent-Environment Interface



Maze

State:

- position



Maze

State:

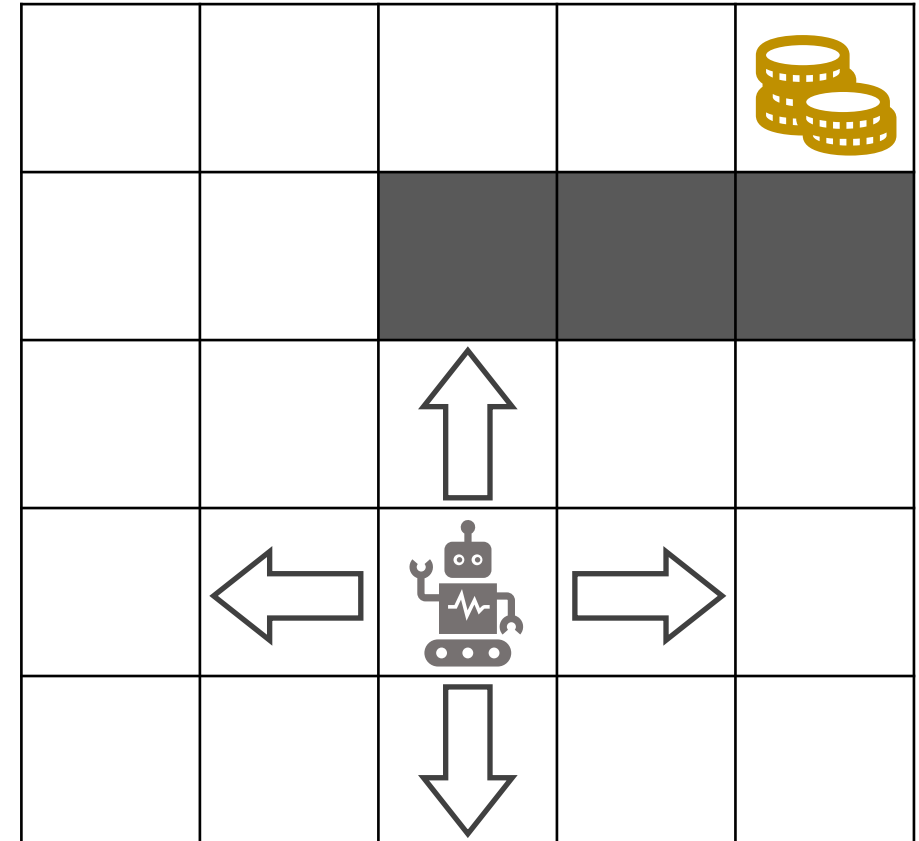
- position

Action:

- up / down / left / right / stop

Reward:

- 1 if goal reached
- 0 otherwise



Pong

State:

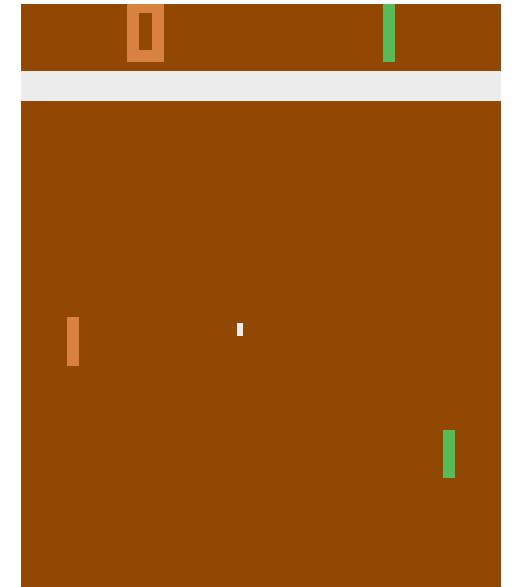
- position + velocity of paddles
- position + velocity of the ball
- scores

Action:

- up / down / stop

Reward:

- 1 if agent scores
- -1 if opponent scores



Pong [Atari]

Humanoid Walking

State:

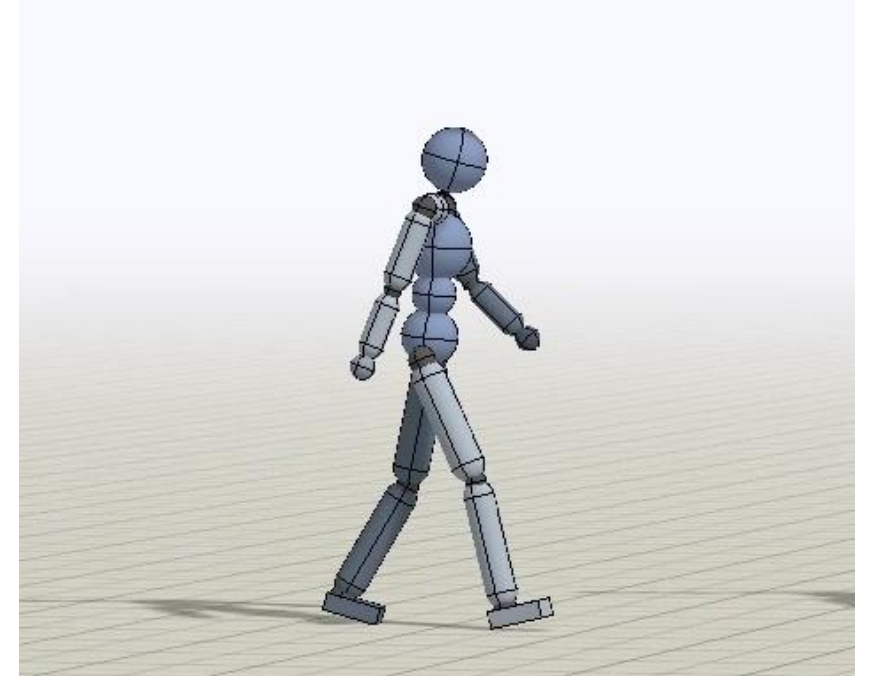
- position + velocity of body parts

Action:

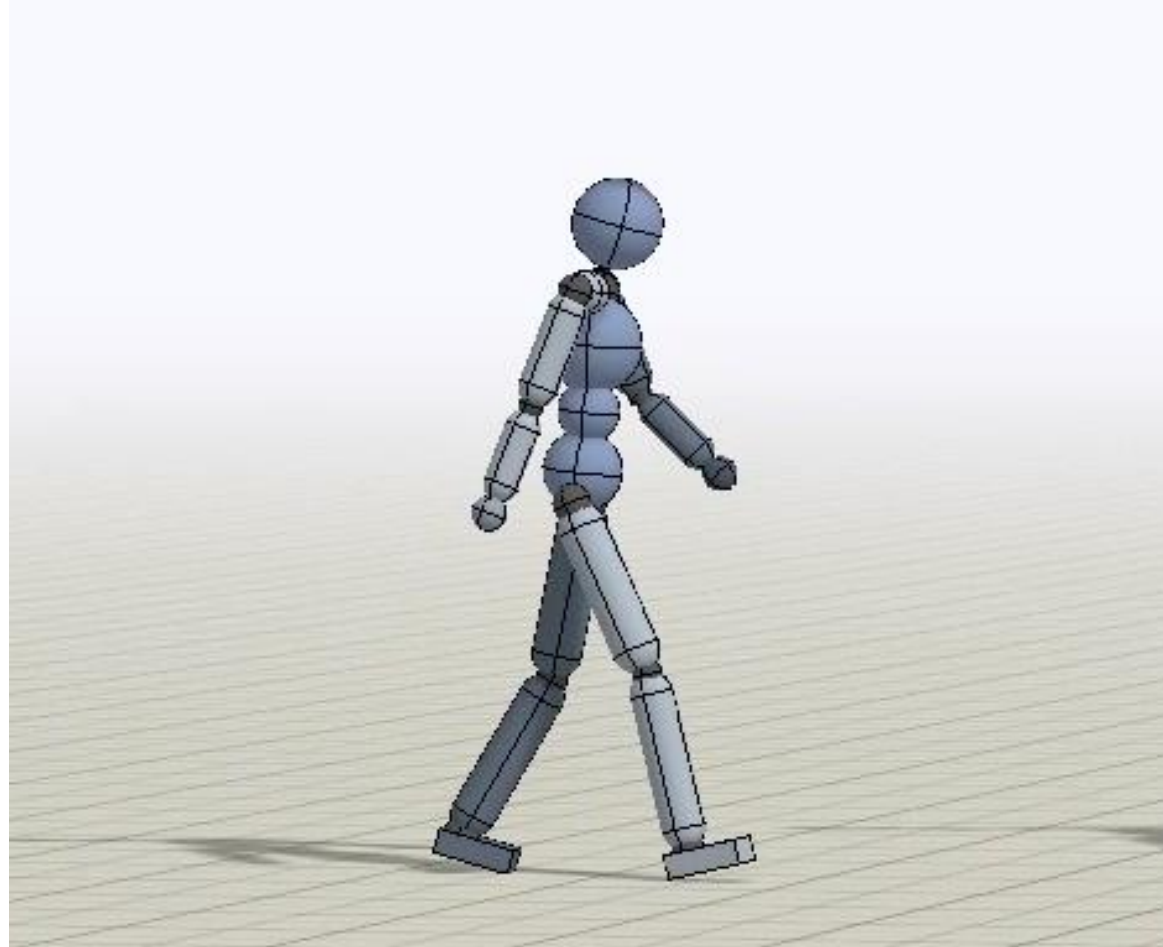
- motor forces

Reward:

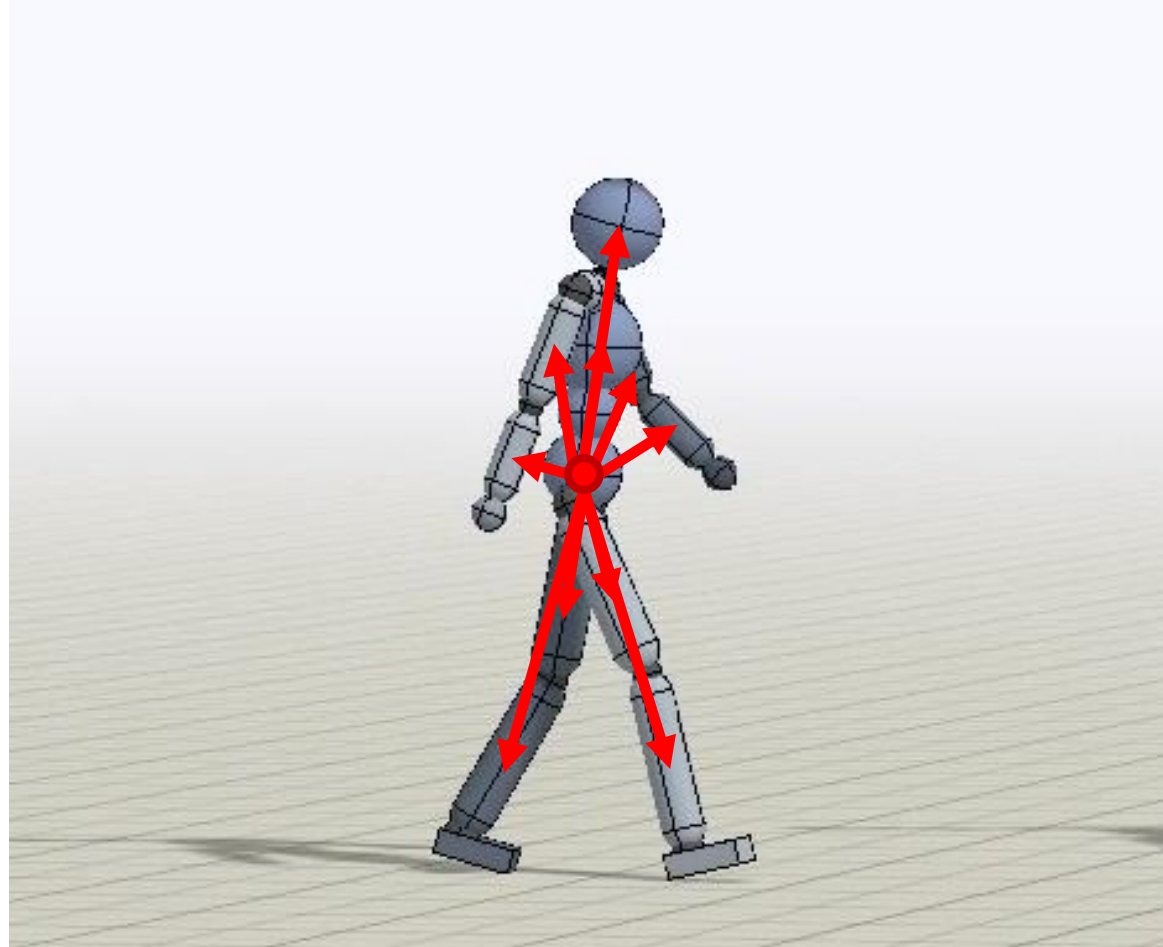
- speed



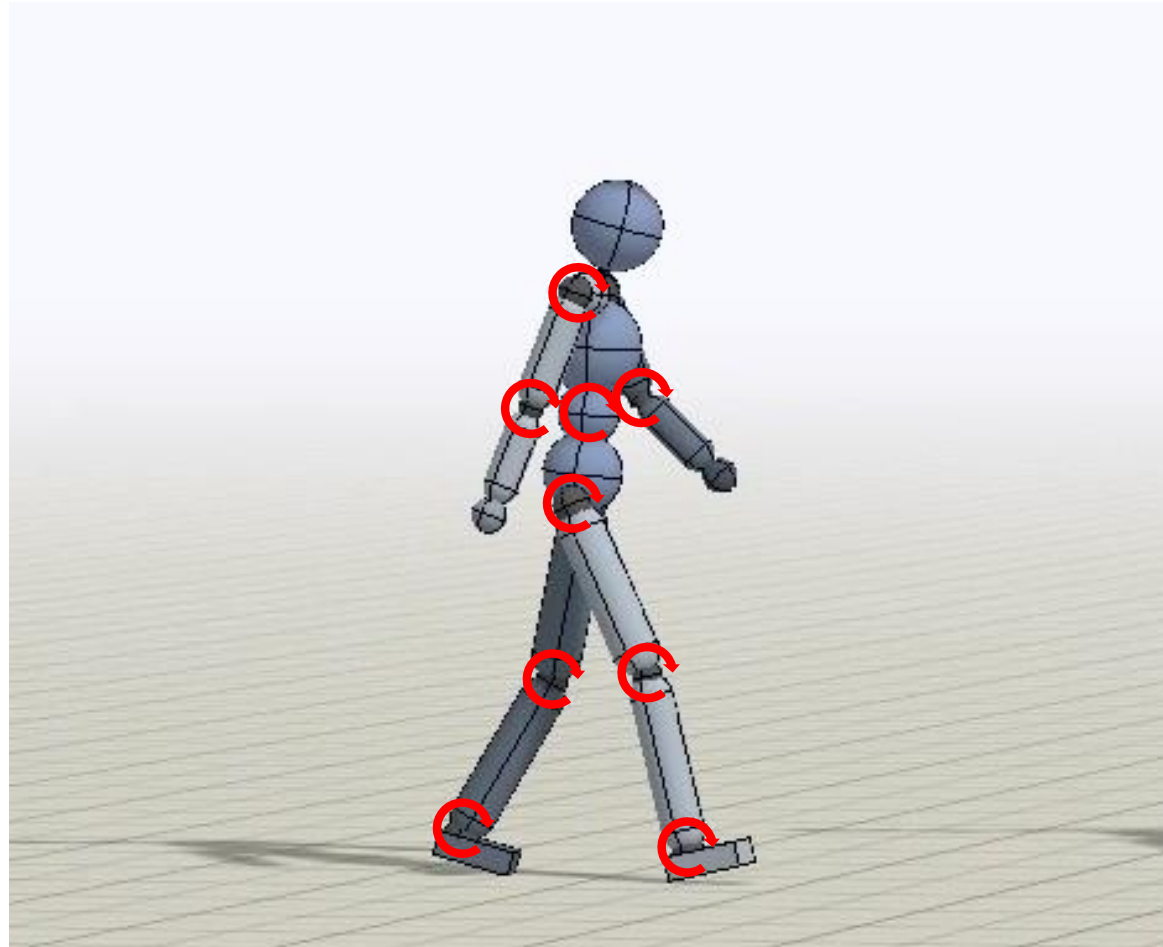
Humanoid Walking



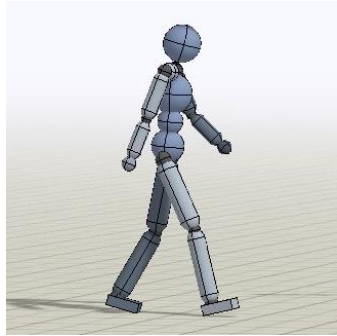
Humanoid Walking (State)



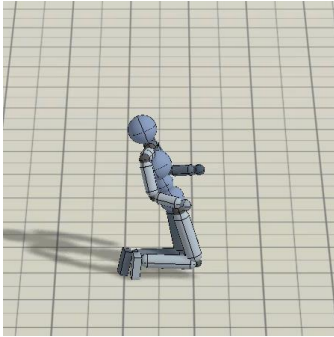
Humanoid Walking (Action)



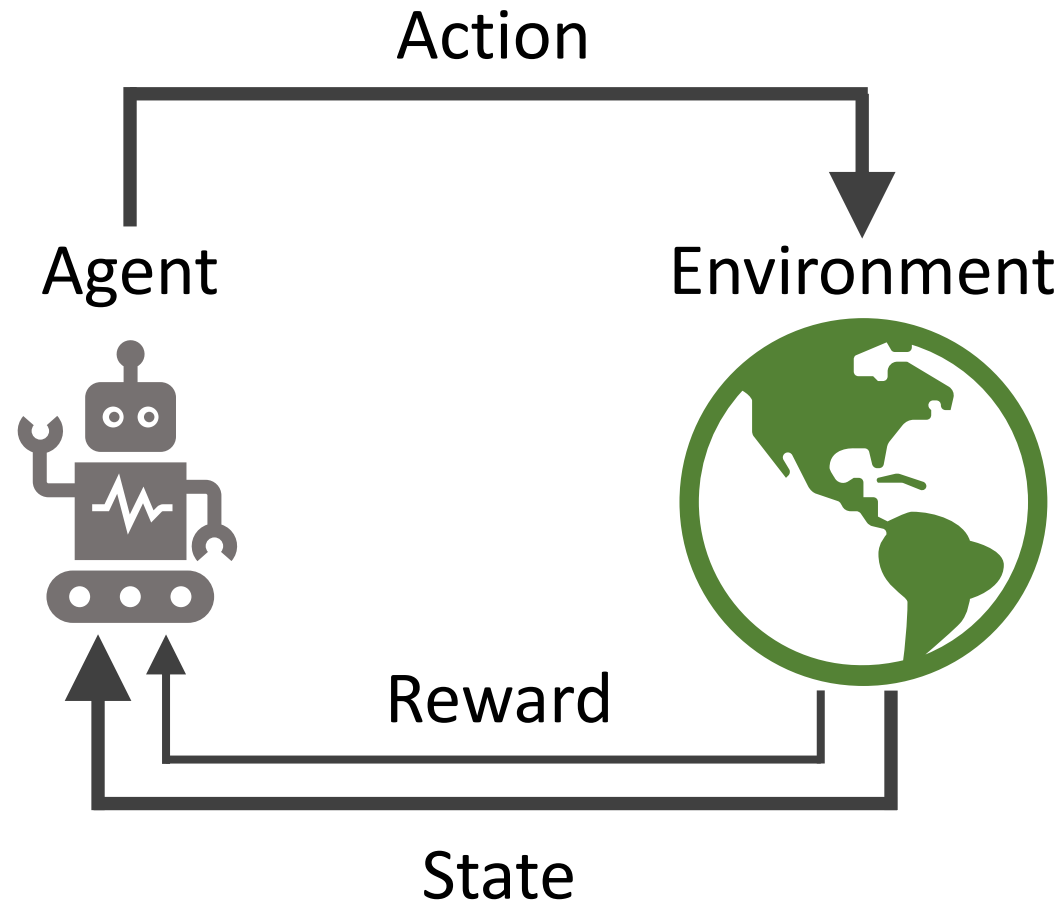
Humanoid Walking (Reward)

$$\text{reward} \left(\text{img} \right) = \text{thumbs up}$$
A blue humanoid robot is shown in a walking pose on a light brown, textured floor. The robot has a blue head, torso, and limbs, with a white joint at the hip. It is walking towards the right. The background is a light blue gradient.

Humanoid Walking (Reward)

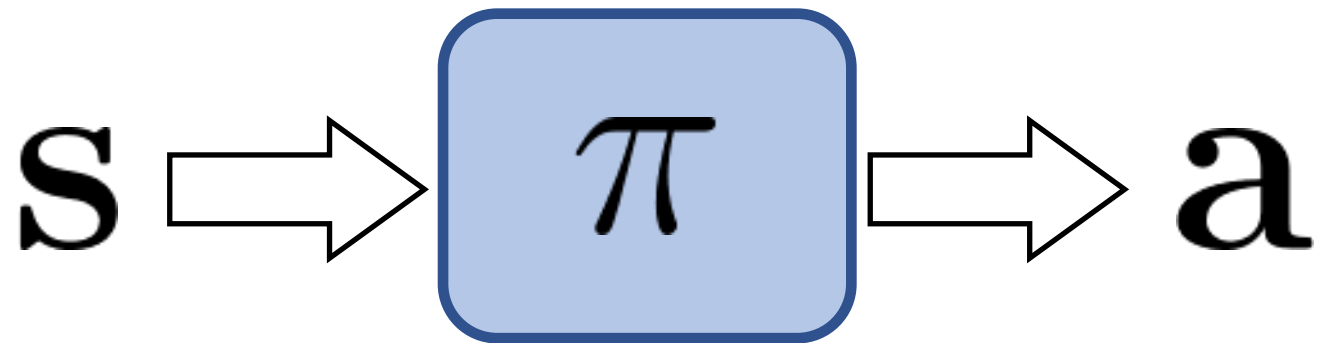
$$\text{reward} \left(\text{img} \right) = \text{thumbs down}$$
A small, blue humanoid robot is shown in a falling or stumbling pose on a light-colored tiled floor. The robot's arms are outstretched, and its legs are bent in a way that suggests it has lost its balance. This image is used to represent a negative reward state in a reinforcement learning context for humanoid walking.

Agent-Environment Interface



Policy

$$\mathbf{a} = \underline{\pi}(\mathbf{s})$$

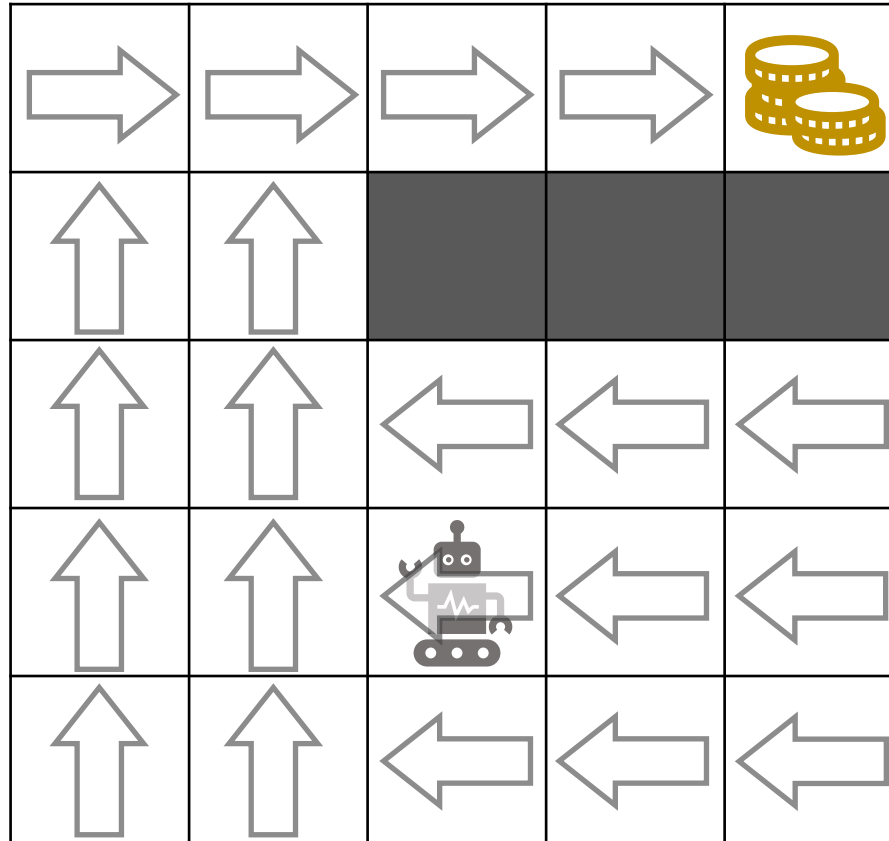


Policy

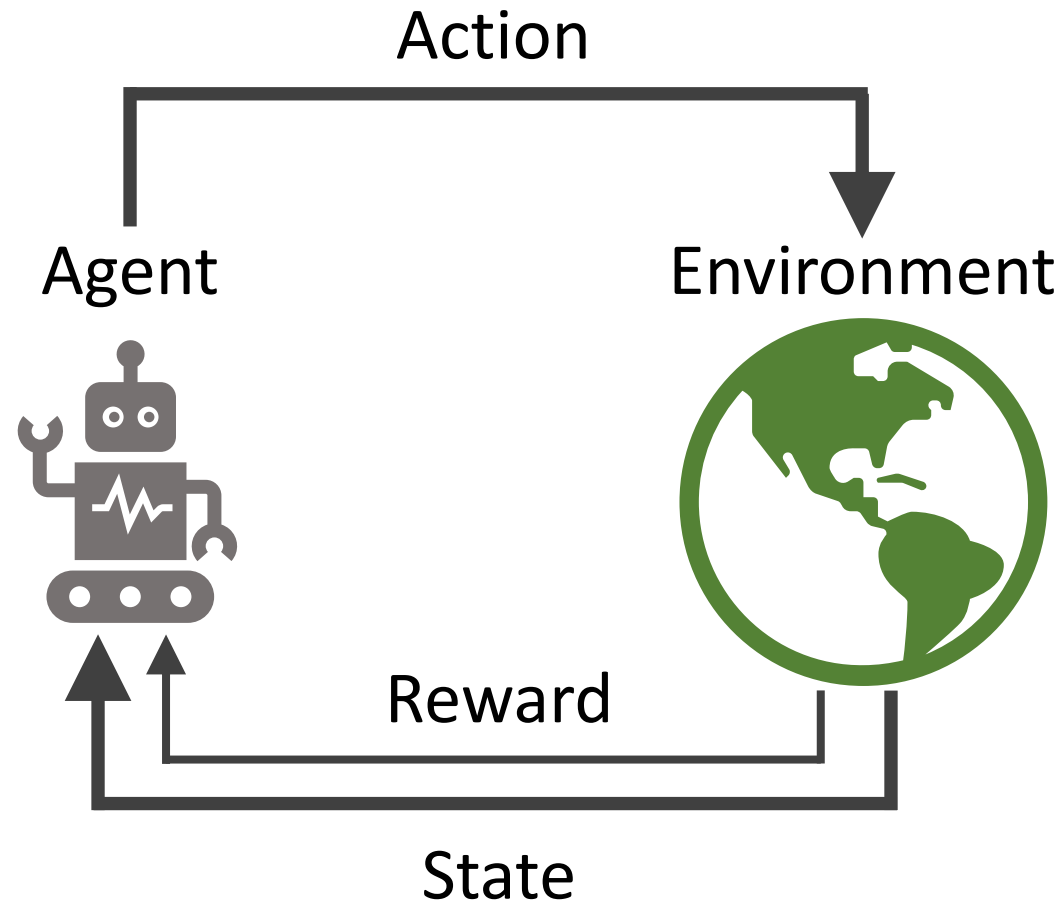
$$\mathbf{a} = \underline{\pi}(\mathbf{s})$$



Maze



Agent-Environment Interface



Markov Decision Process

$\mathbf{s} \in \mathcal{S}$ – state space

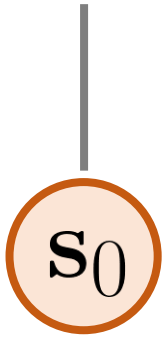
$\mathbf{a} \in \mathcal{A}$ – action space

$p(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ – dynamics function

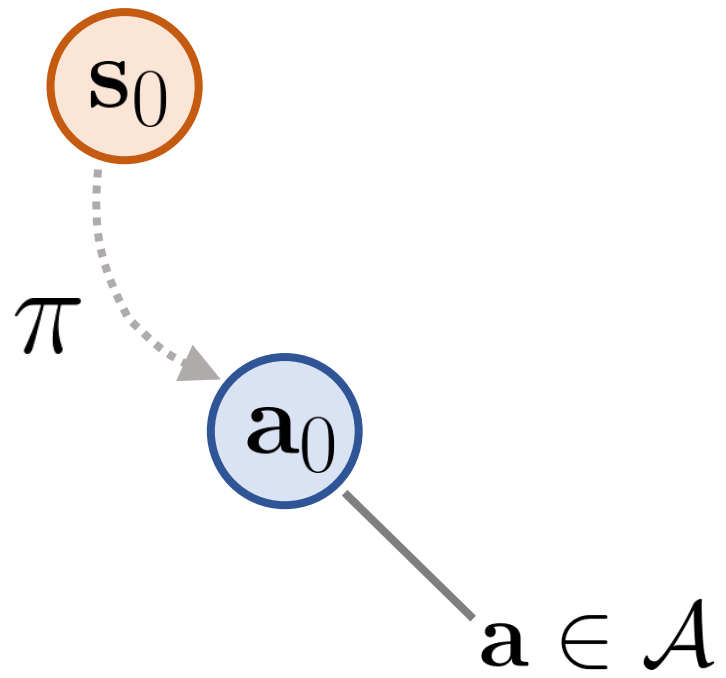
$r(\mathbf{s}, \mathbf{a}, \mathbf{s}')$ – reward function

MDP

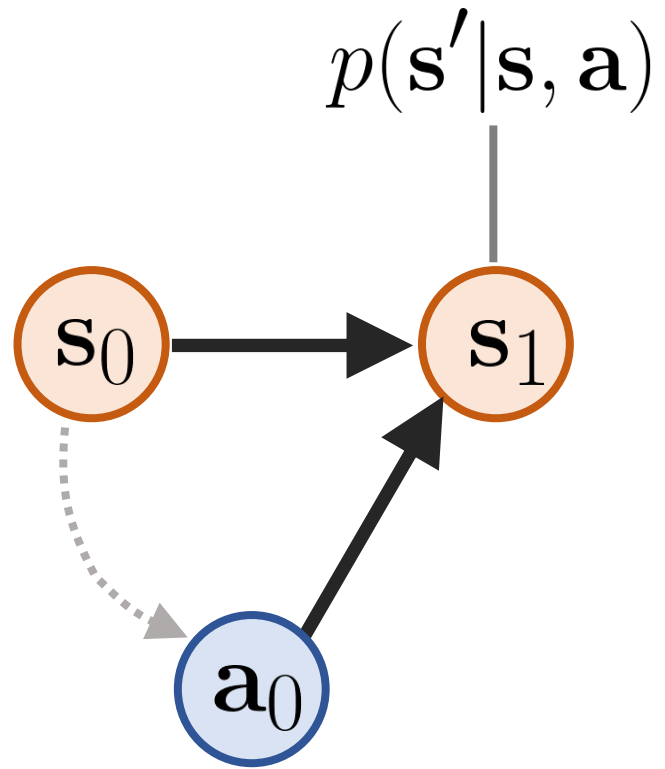
$s \in \mathcal{S}$



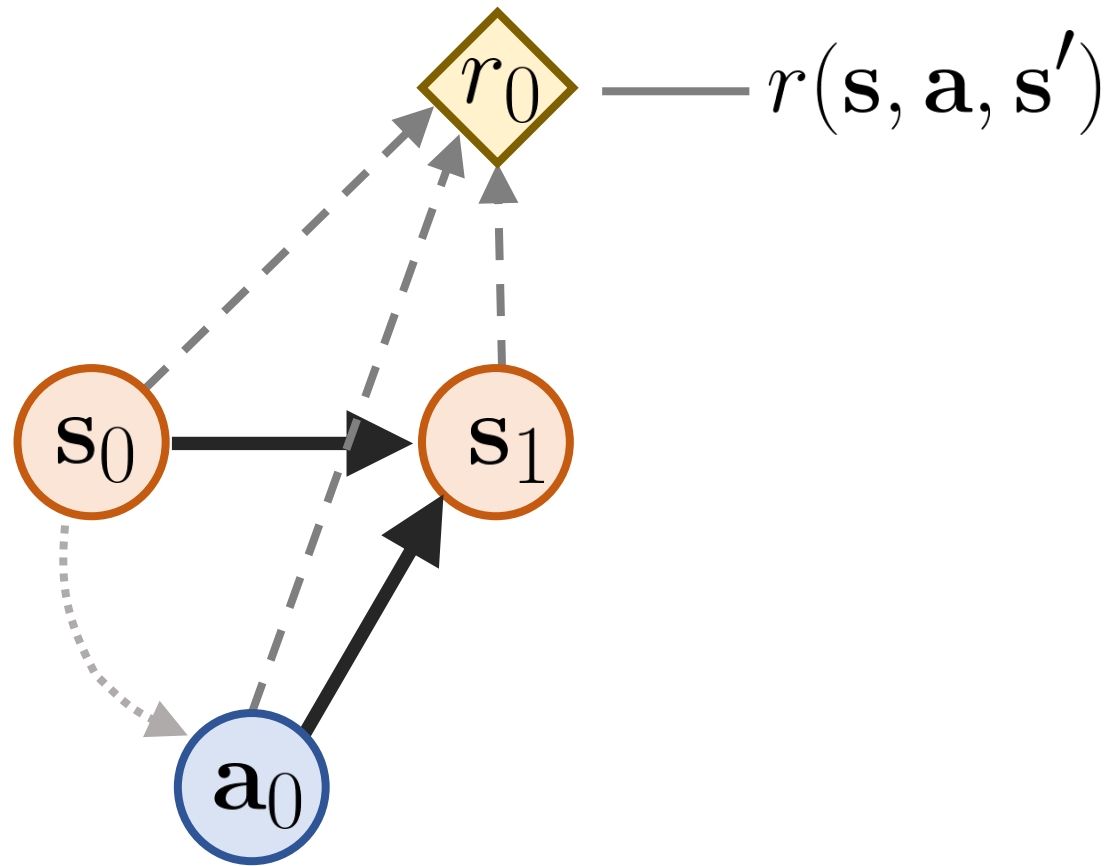
MDP



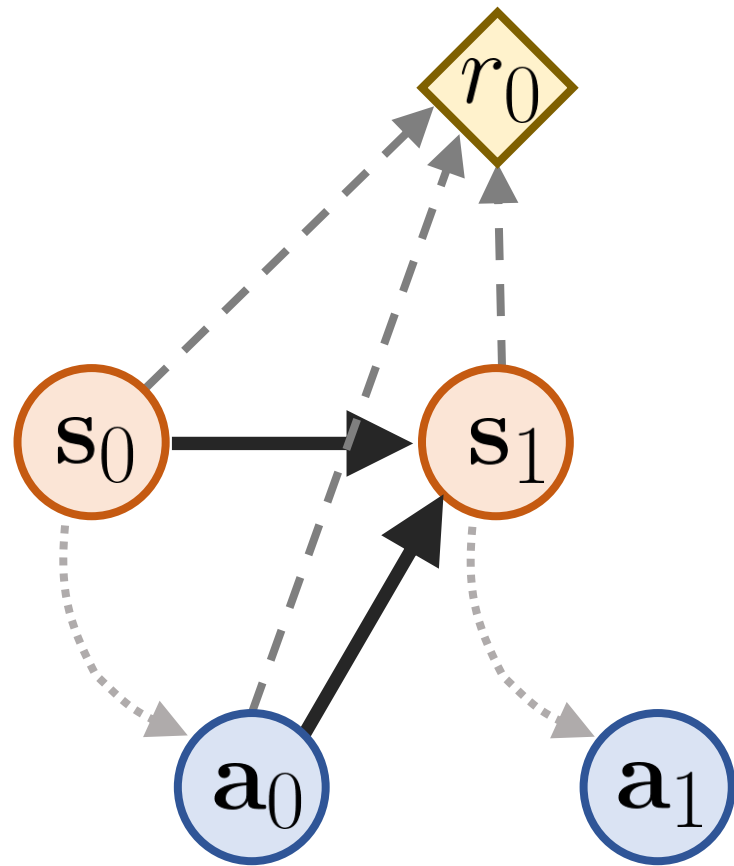
MDP



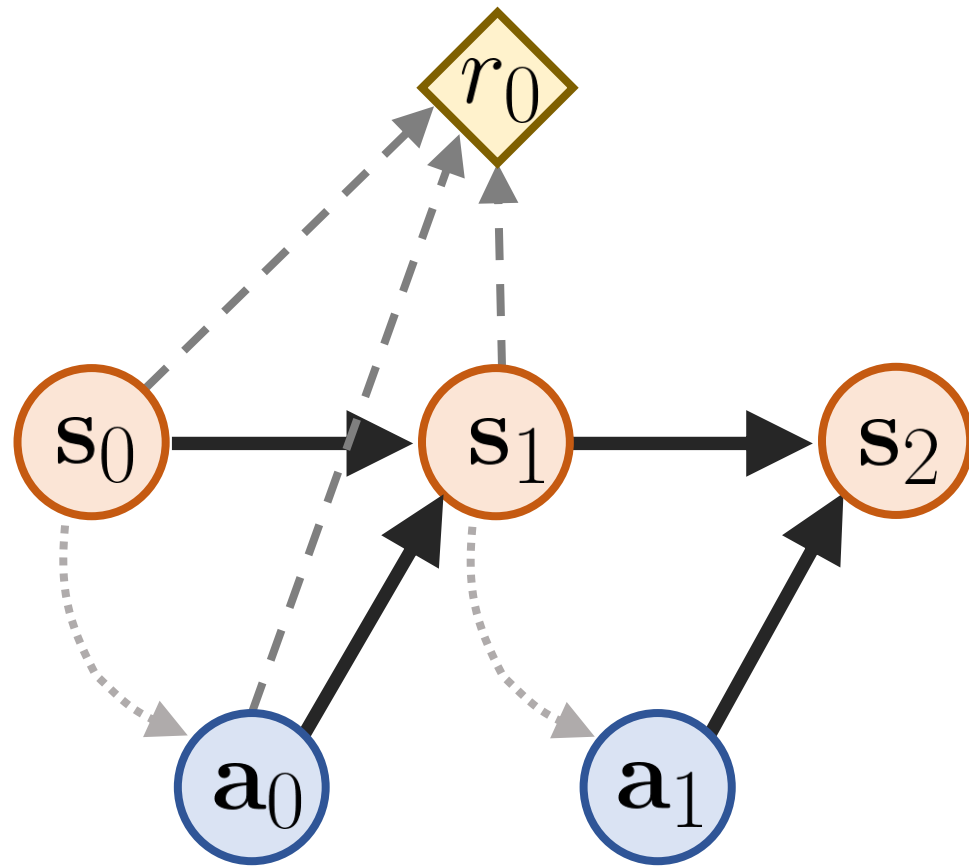
MDP



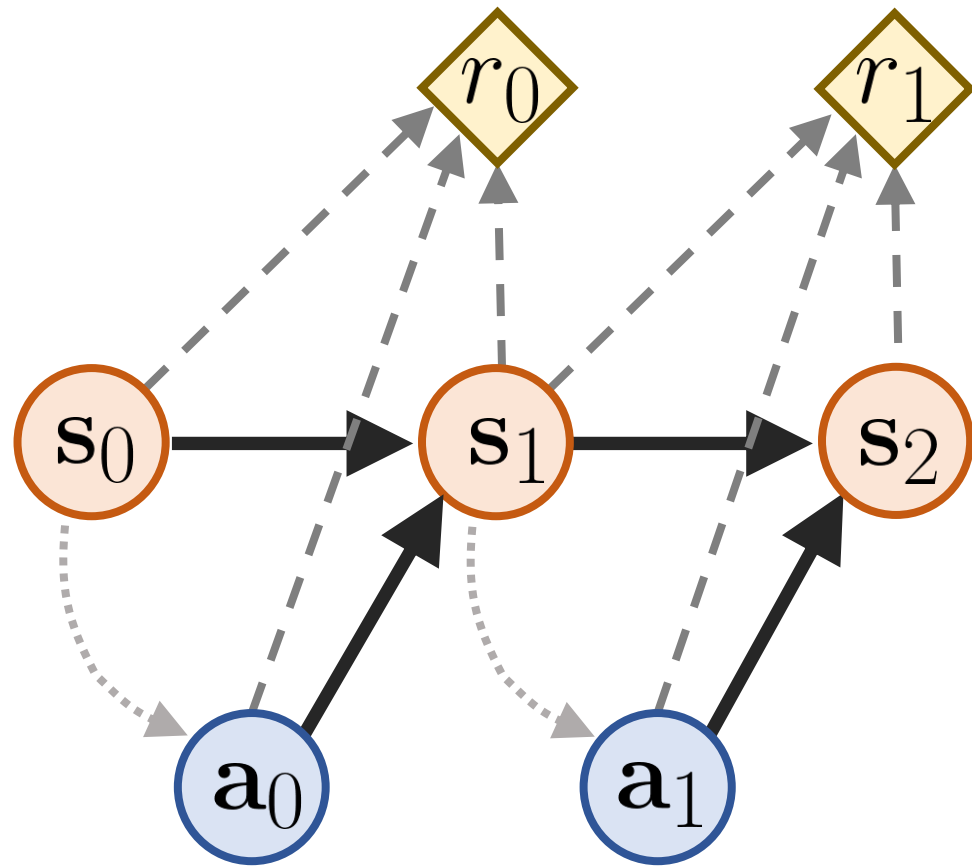
MDP



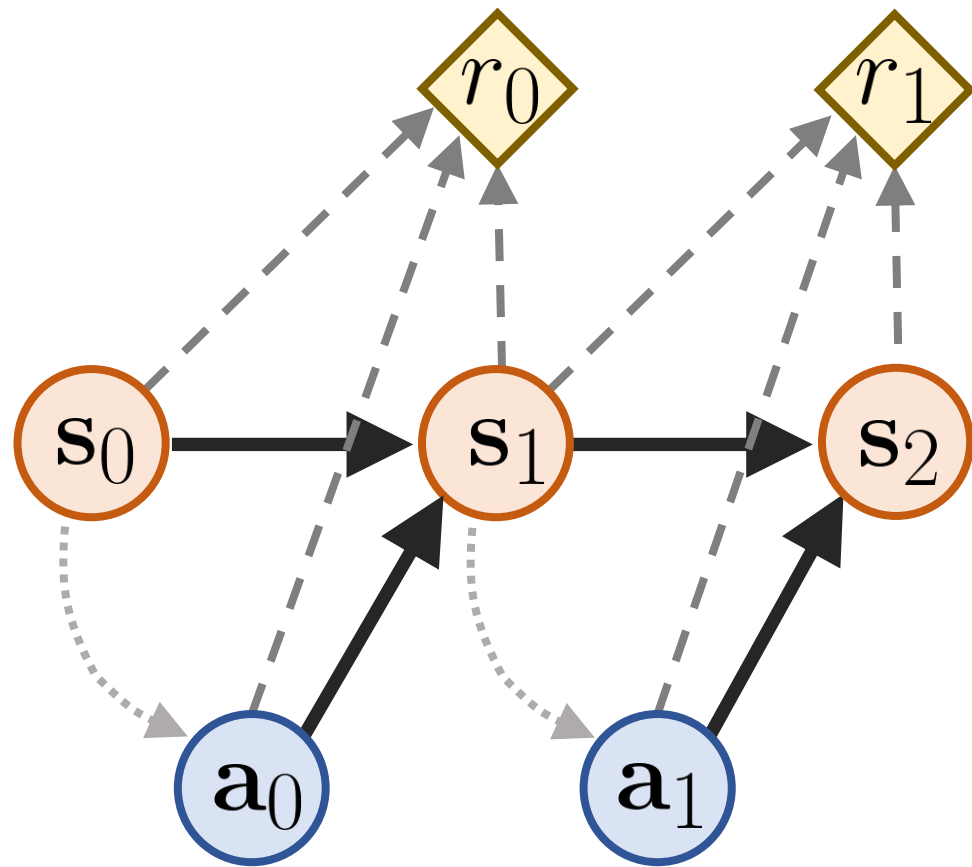
MDP



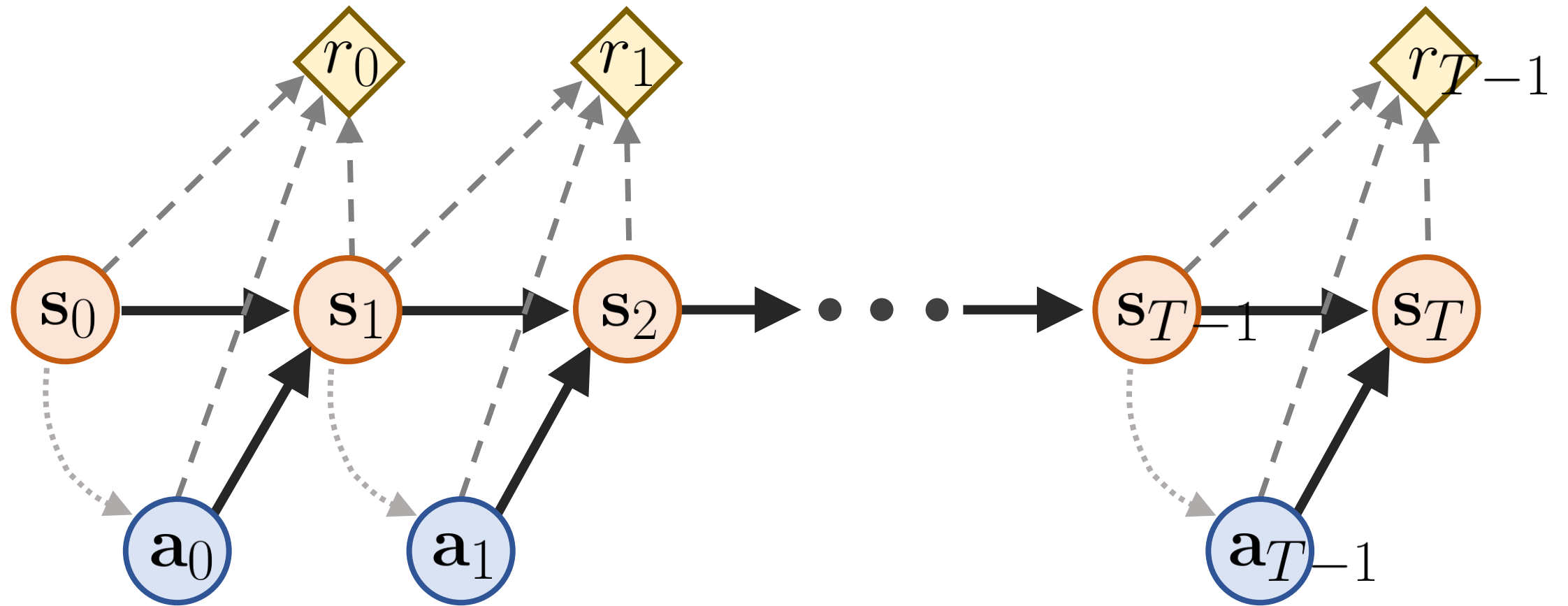
MDP



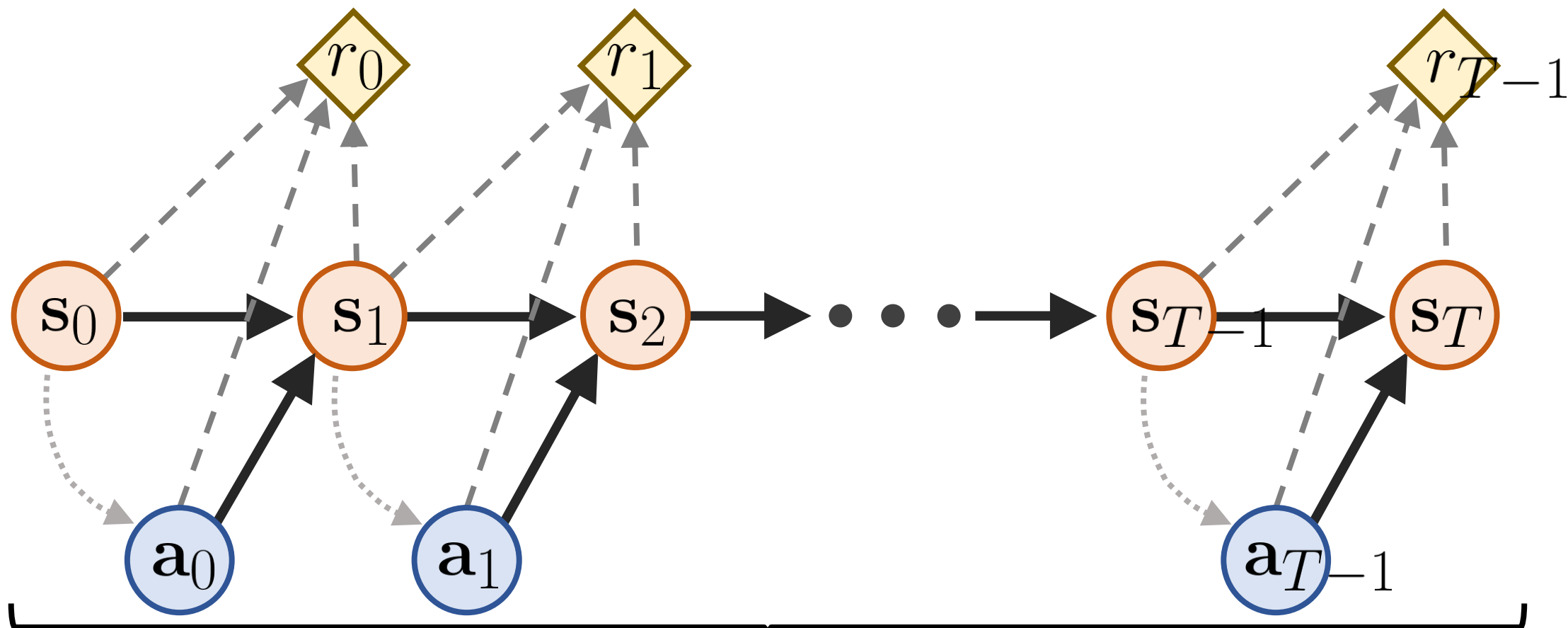
MDP



MDP

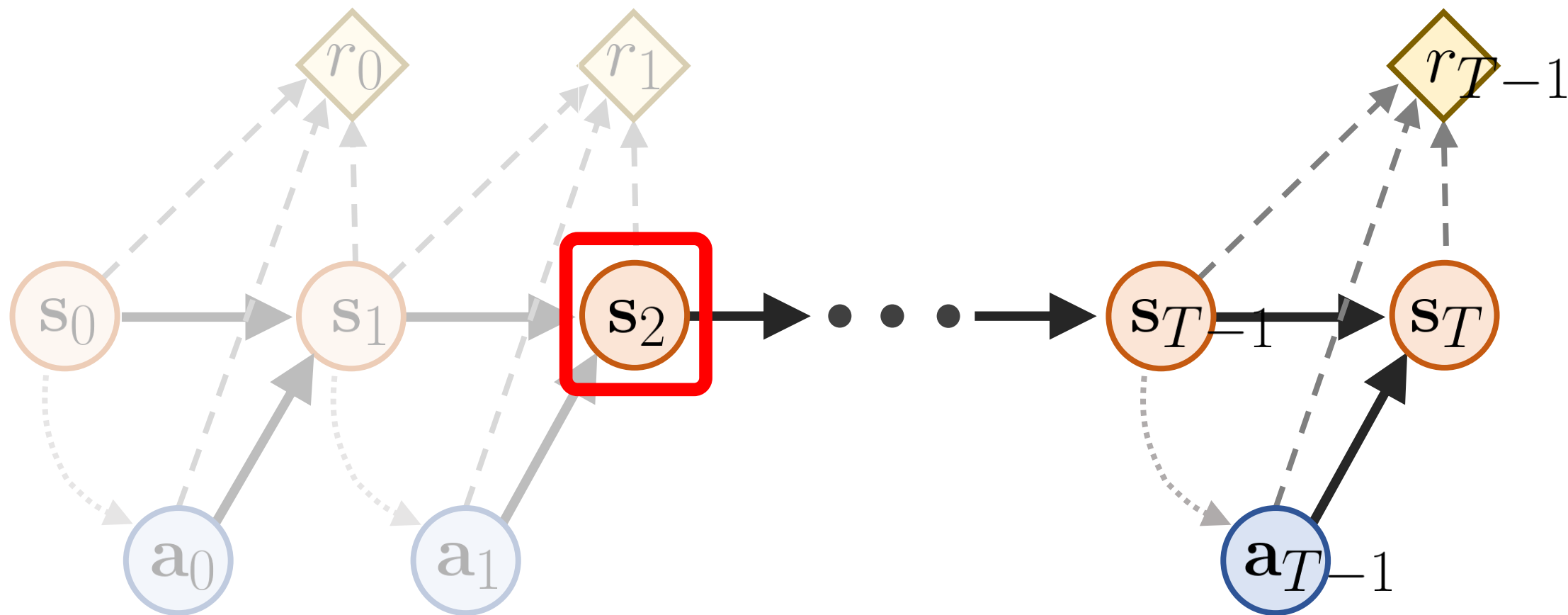


MDP



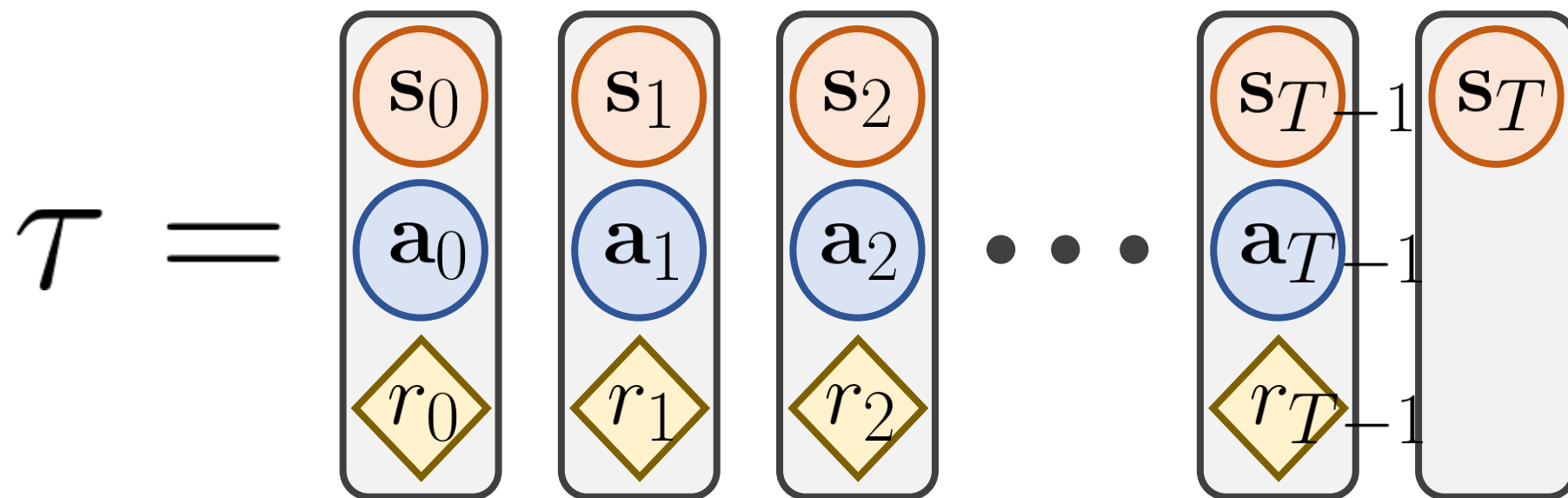
$\max \sum_{t=0}^{T-1} \gamma^t r_t$ — return

MDP

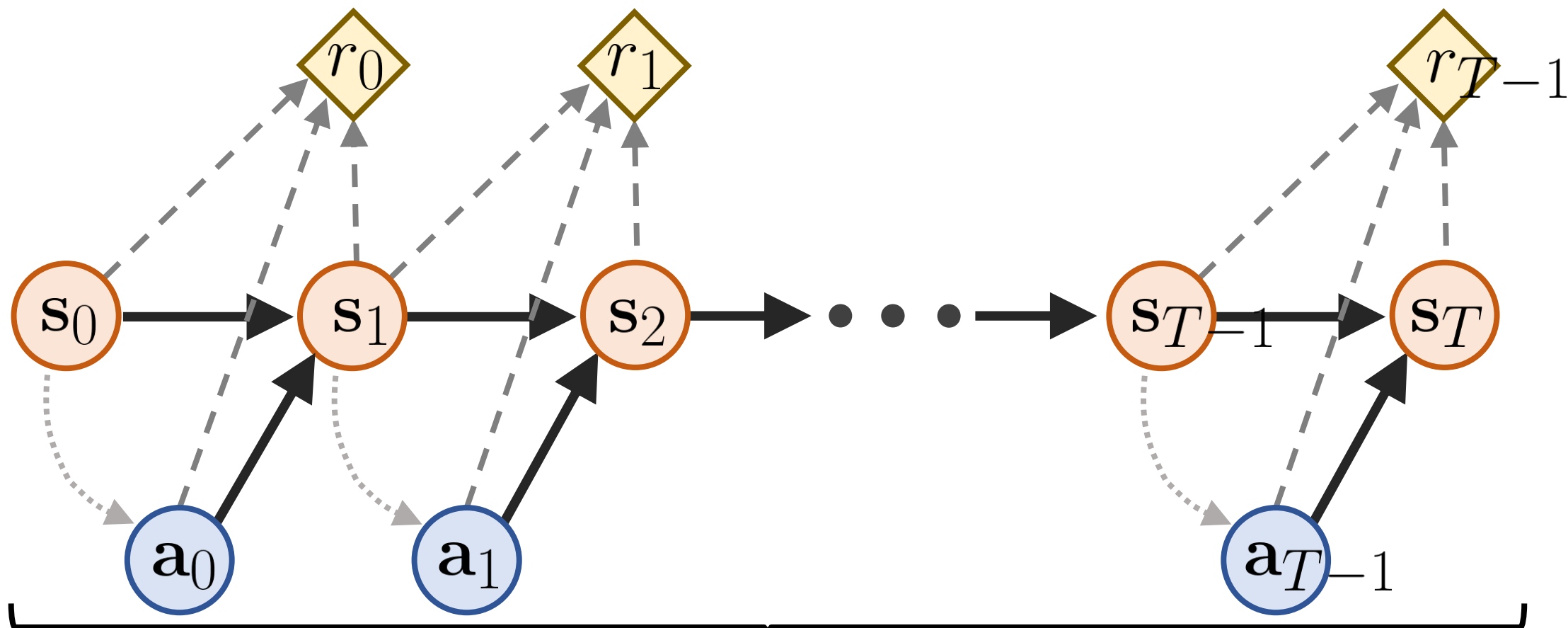


Markov property: $p(s'|s, a)$

Trajectory

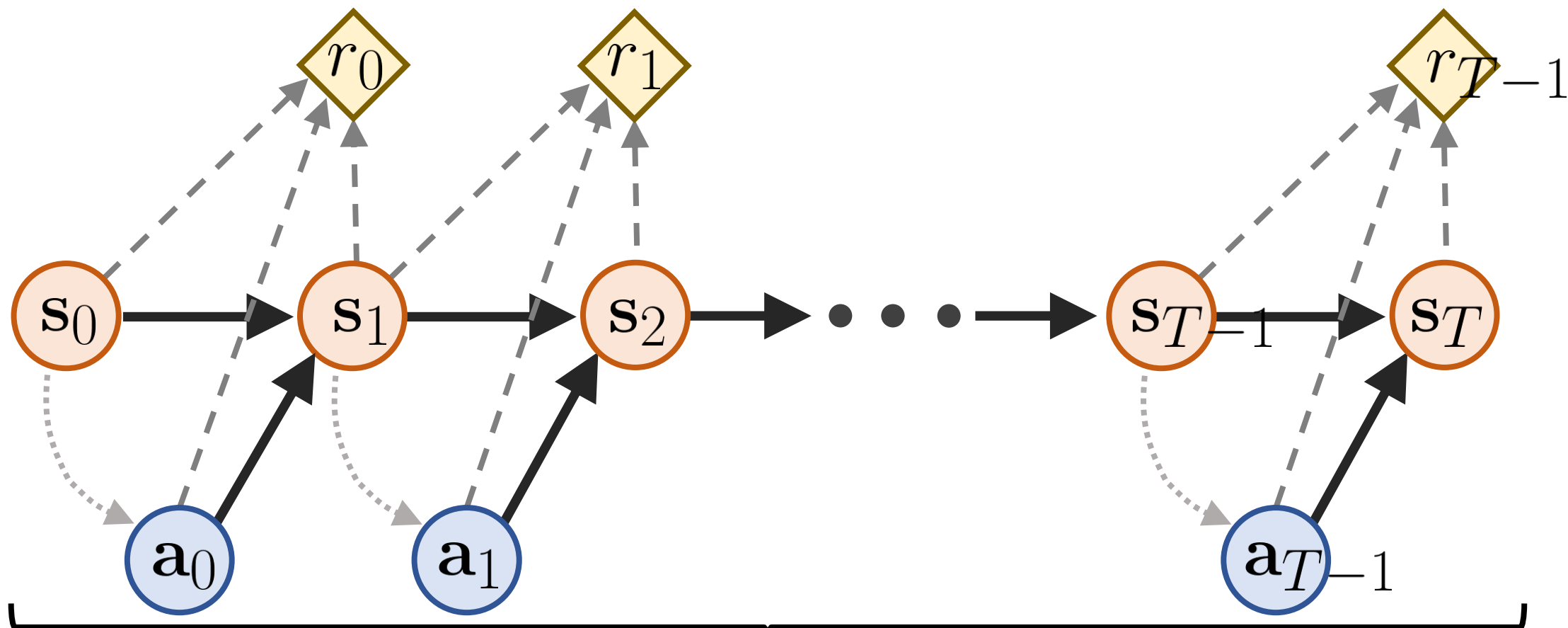


MDP



$$T \in [1, \infty] \quad \max \sum_{t=0}^{T-1} \gamma^t r_t$$

MDP



$$\max \sum_{t=0}^{T-1} \gamma^t r_t \quad \gamma \in [0, 1]$$

Discount Factor

$$\underline{J(\pi)} = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]$$

Discount Factor

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]$$

Discount Factor

$$J(\pi) = \mathbb{E}_{\tau \sim \underline{p(\tau|\pi)}} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]$$

Discount Factor

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \underline{\gamma^t r_t} \right]$$

Discount Factor

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]$$

$$R = r_0 + \gamma r_1 + \gamma^2 r_2 + \dots + \gamma^{T-1} r_{T-1}$$

$\gamma = 1$: maximize rewards at all timesteps equally

Discount Factor

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]$$

$$R = r_0 + r_1 + r_2 + \dots + r_{T-1}$$

$\gamma = 1$: maximize rewards at all timesteps equally

Discount Factor

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]$$

$$R = r_0 + \gamma r_1 + \gamma^2 r_2 + \dots + \gamma^{T-1} r_{T-1}$$

$\gamma = 0$: only maximize reward at first timestep

Discount Factor

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]$$

$$R = r_0$$

$\gamma = 0$: only maximize reward at first timestep

Discount Factor

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]$$

$$R = r_0 + 0.5r_1 + 0.5^2r_2 + \dots + 0.5^{T-1}r_{T-1}$$

$$\gamma = 0.5$$

Discount Factor

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]$$

$$R = r_0 + \gamma r_1 + \gamma^2 r_2 + \dots + \gamma^{T-1} r_{T-1}$$

Large γ : maximize long term rewards (far-sighted)

Small γ : maximize short term rewards (short-sighted / greedy)

Discount Factor

$$J(\pi) = \mathbb{E}_{\tau \sim p(\tau|\pi)} \left[\sum_{t=0}^{T-1} \gamma^t r_t \right]$$

$$R = r_0 + \gamma r_1 + \gamma^2 r_2 + \dots + \gamma^{T-1} r_{T-1}$$

$\gamma = 1$: maximize rewards at all timesteps equally

$T \neq \infty$

Expectation undefined if $T = \infty$

Discount Factor

- Geometric series with ratio $\gamma < 1$
- $r \in [r_{\min}, r_{\max}]$, $r \geq 0$

$$R = r_0 + \gamma r_1 + \gamma^2 r_2 + \dots + \gamma^{T-1} r_{T-1}$$

Return Bounds

- Geometric series with ratio $\gamma < 1$
- $r \in [r_{\min}, r_{\max}]$, $r \geq 0$

Return Bounds

- Geometric series with ratio $\gamma < 1$
- $r \in [r_{\min}, r_{\max}]$, $r \geq 0$

$$R = \sum_{t=0}^{T-1} \gamma^t r_t$$

Return Bounds

- Geometric series with ratio γ
- $r \in [r_{\min}, r_{\max}]$, $r \geq 0$

$$R = \sum_{t=0}^{T-1} \gamma^t r_t \leq \sum_{t=0}^{T-1} \gamma^t r_{\max}$$

Return Bounds

- Geometric series with ratio γ
- $r \in [r_{\min}, r_{\max}]$, $r \geq 0$

$$\begin{aligned} R &= \sum_{t=0}^{T-1} \gamma^t r_t \leq \sum_{t=0}^{T-1} \gamma^t r_{\max} \\ &\leq \sum_{t=0}^{\infty} \gamma^t r_{\max} \end{aligned}$$

Return Bounds

- Geometric series with ratio γ
- $r \in [r_{\min}, r_{\max}]$, $r \geq 0$

$$\begin{aligned} R &= \sum_{t=0}^{T-1} \gamma^t r_t \leq \sum_{t=0}^{T-1} \gamma^t r_{\max} \\ &\leq \sum_{t=0}^{\infty} \gamma^t r_{\max} = r_{\max} \sum_{t=0}^{\infty} \gamma^t \end{aligned}$$

Return Bounds

- Geometric series with ratio γ
- $r \in [r_{\min}, r_{\max}]$, $r \geq 0$

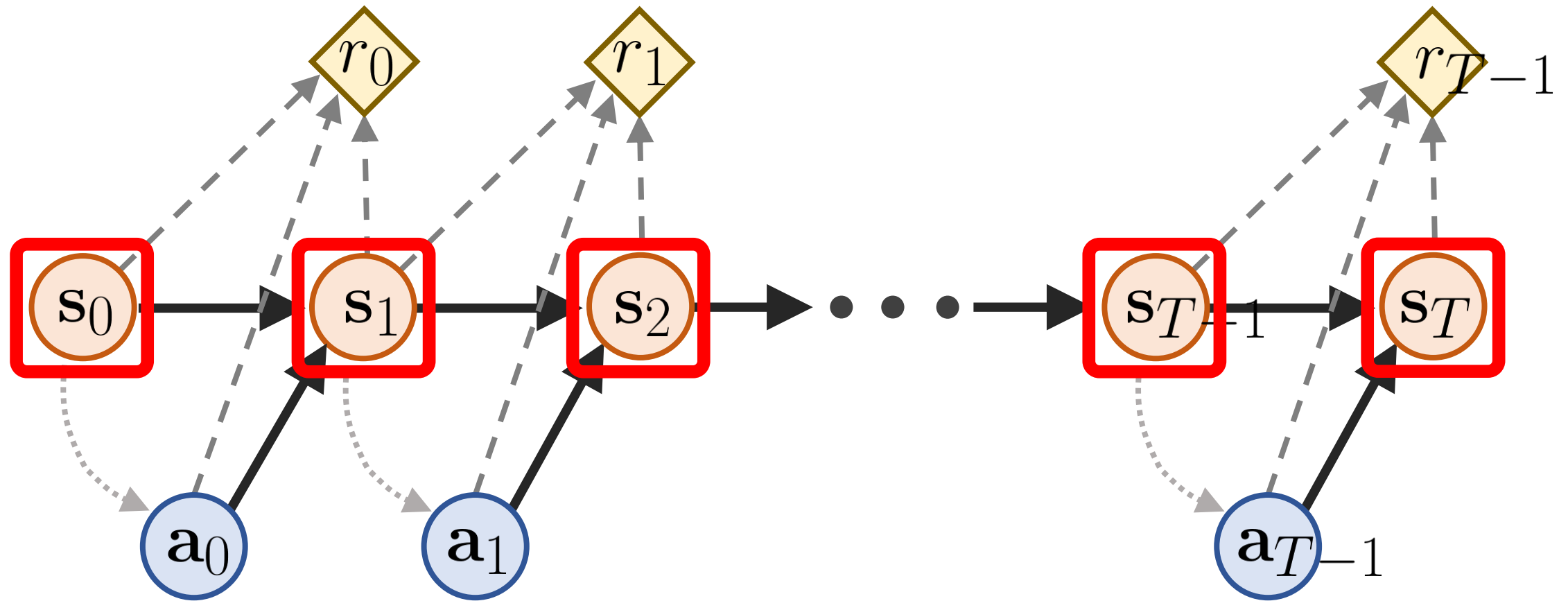
$$\begin{aligned} R &= \sum_{t=0}^{T-1} \gamma^t r_t \leq \sum_{t=0}^{T-1} \gamma^t r_{\max} \\ &\leq \sum_{t=0}^{\infty} \gamma^t r_{\max} = r_{\max} \sum_{t=0}^{\infty} \gamma^t \\ &\leq r_{\max} \frac{1}{1 - \gamma} \end{aligned}$$

Return Bounds

$$R_{\min} = \frac{1}{1 - \gamma} r_{\min} \qquad R_{\max} = \frac{1}{1 - \gamma} r_{\max}$$

$$R_{\min} \leq R \leq R_{\max}$$

MDP



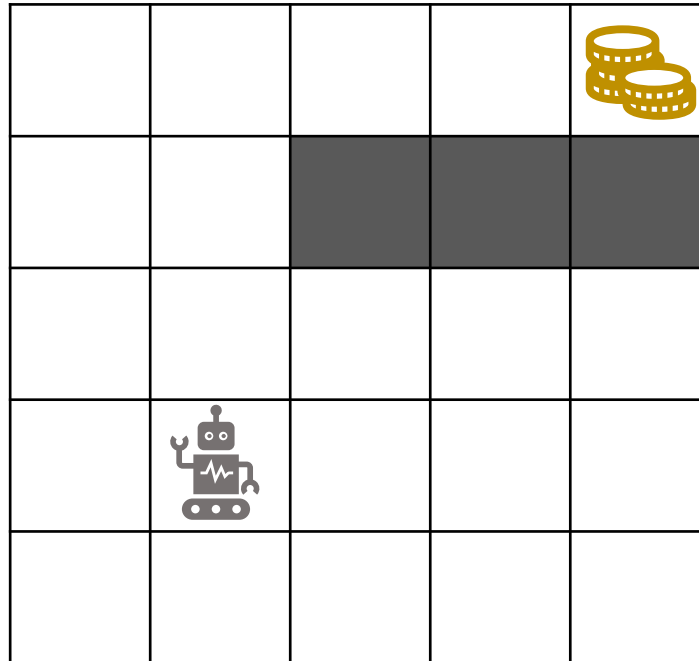
State Spaces

Discrete

$$\mathbf{s} \in \{\mathbf{s}^0, \mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^n\}$$

Continuous

$$\mathbf{s} \in \mathbb{R}^n$$



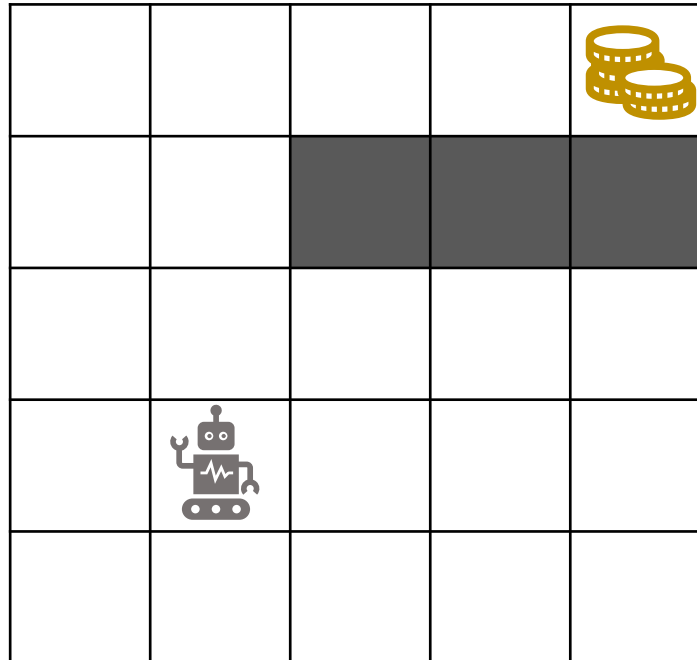
State Spaces

Discrete

$$\mathbf{s} \in \{\mathbf{s}^0, \mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^n\}$$

Continuous

$$\mathbf{s} \in \mathbb{R}^n$$



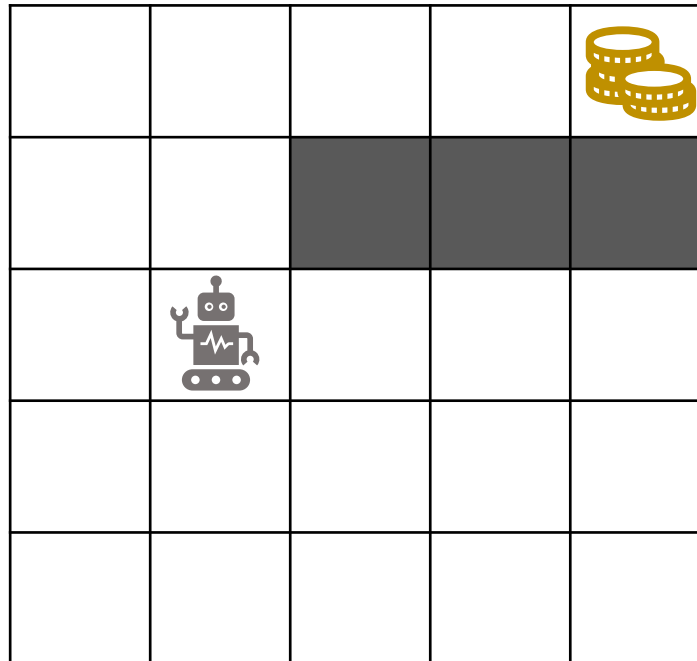
State Spaces

Discrete

$$\mathbf{s} \in \{\mathbf{s}^0, \mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^n\}$$

Continuous

$$\mathbf{s} \in \mathbb{R}^n$$



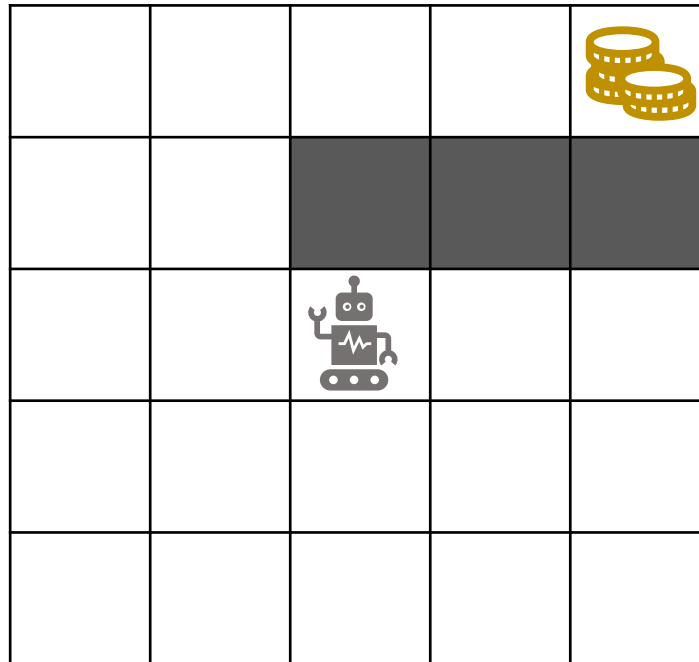
State Spaces

Discrete

$$\mathbf{s} \in \{\mathbf{s}^0, \mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^n\}$$

Continuous

$$\mathbf{s} \in \mathbb{R}^n$$



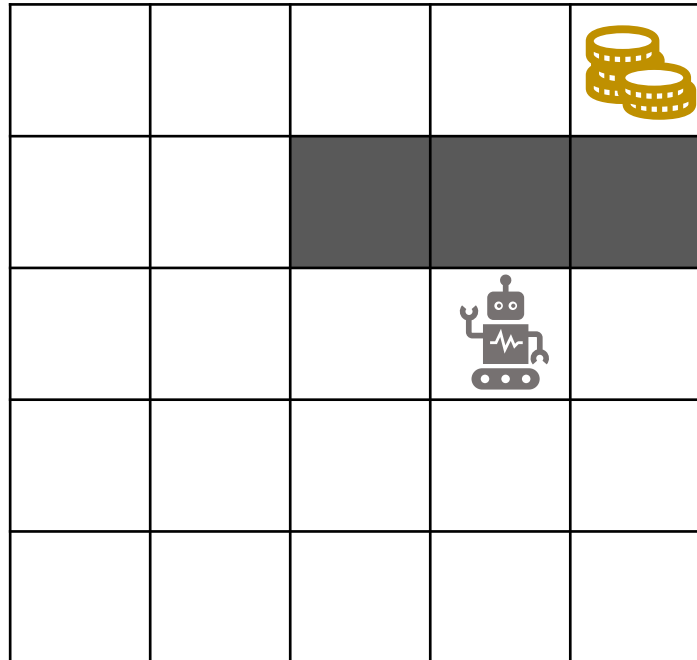
State Spaces

Discrete

$$\mathbf{s} \in \{\mathbf{s}^0, \mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^n\}$$

Continuous

$$\mathbf{s} \in \mathbb{R}^n$$



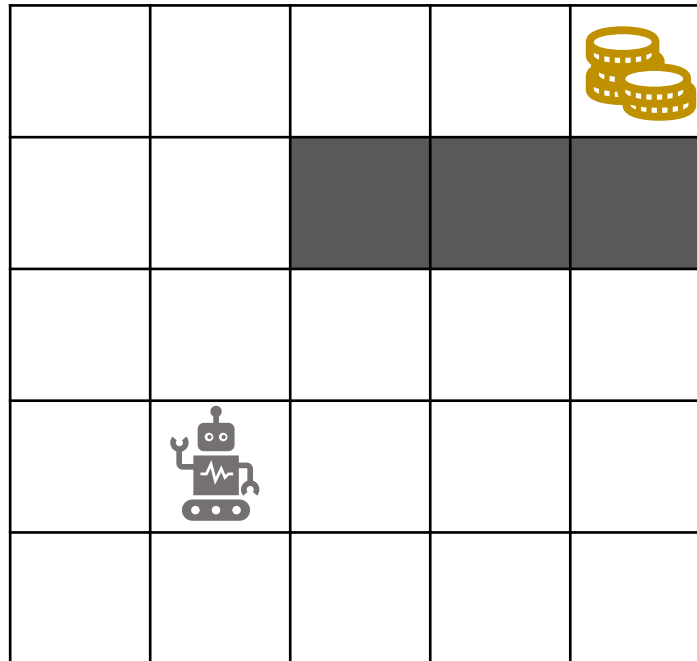
State Spaces

Discrete

$$\mathbf{s} \in \{\mathbf{s}^0, \mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^n\}$$

Continuous

$$\mathbf{s} \in \mathbb{R}^n$$



State Spaces

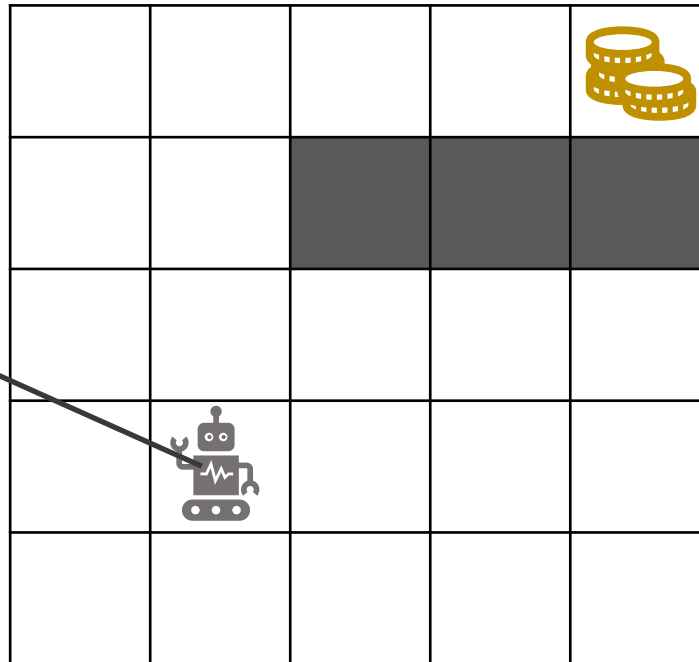
Discrete

$$\mathbf{s} \in \{\mathbf{s}^0, \mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^n\}$$

Continuous

$$\mathbf{s} \in \mathbb{R}^n$$

$$\mathbf{s} = (x, y)$$



State Spaces

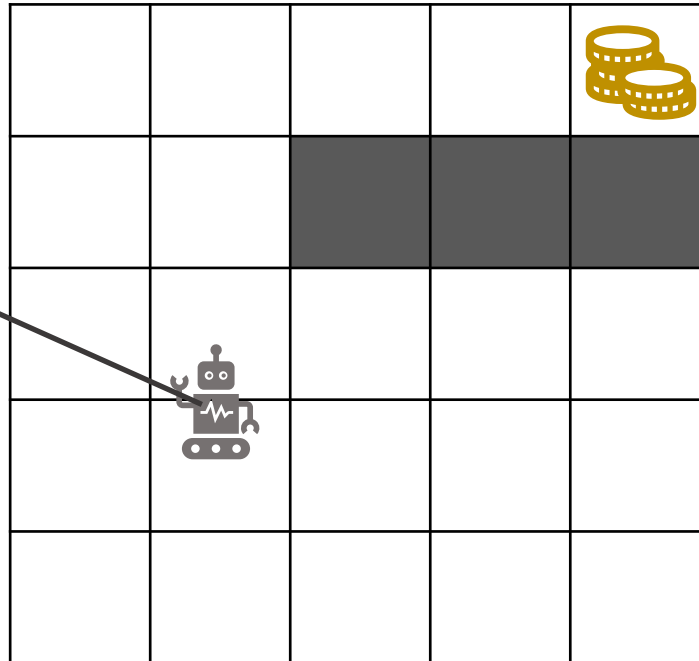
Discrete

$$\mathbf{s} \in \{\mathbf{s}^0, \mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^n\}$$

Continuous

$$\mathbf{s} \in \mathbb{R}^n$$

$$\mathbf{s} = (x, y)$$



State Spaces

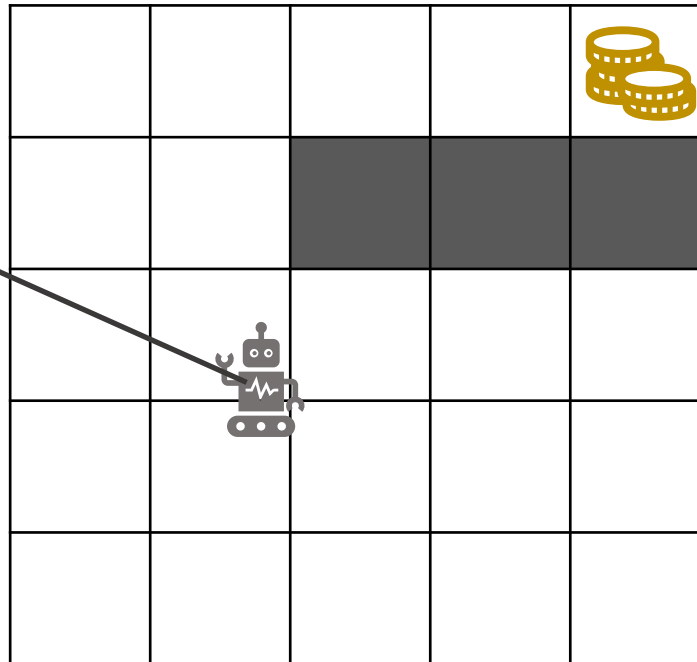
Discrete

$$\mathbf{s} \in \{\mathbf{s}^0, \mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^n\}$$

Continuous

$$\mathbf{s} \in \mathbb{R}^n$$

$$\mathbf{s} = (x, y)$$



State Spaces

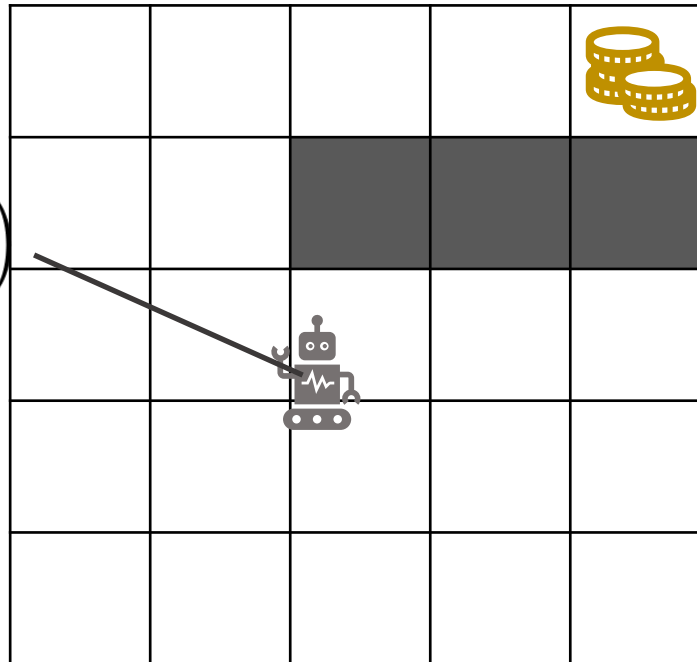
Discrete

$$\mathbf{s} \in \{\mathbf{s}^0, \mathbf{s}^1, \mathbf{s}^2, \dots, \mathbf{s}^n\}$$

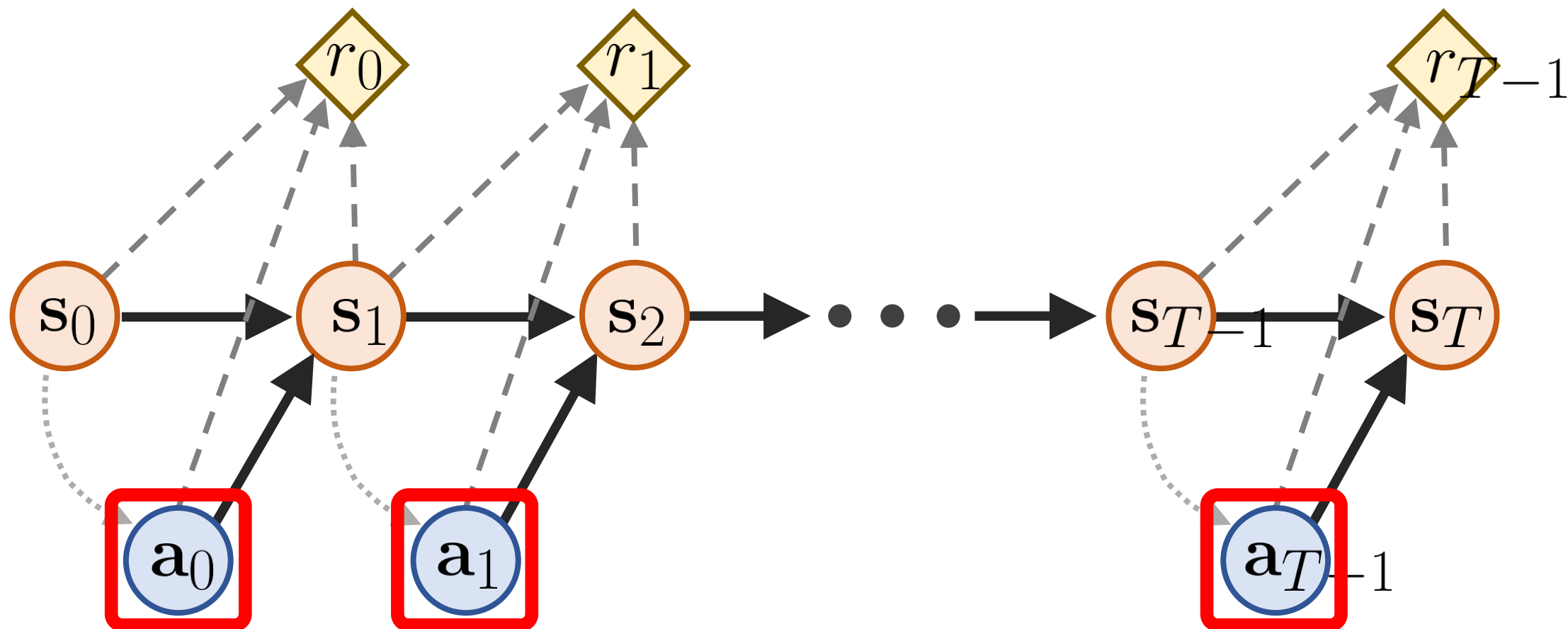
Continuous

$$\mathbf{s} \in \mathbb{R}^n$$

$$\mathbf{s} = (x, y)$$



MDP



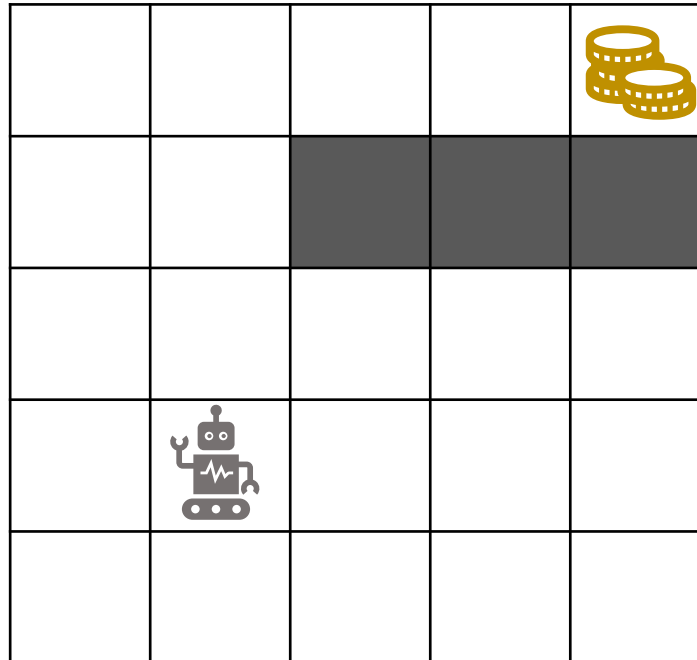
Action Spaces

Discrete

$$\mathbf{a} \in \{\mathbf{a}^0, \mathbf{a}^1, \mathbf{a}^2, \dots, \mathbf{a}^m\}$$

Continuous

$$\mathbf{a} \in \mathbb{R}^m$$



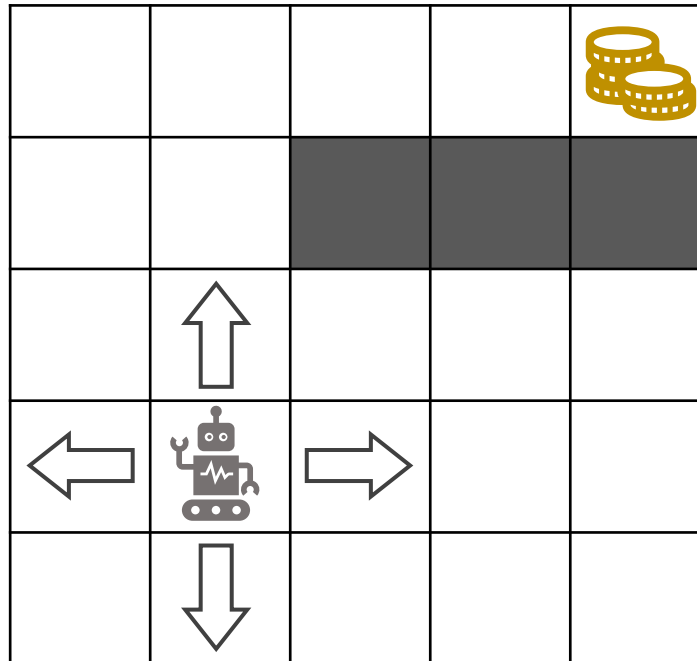
Action Spaces

Discrete

$$\mathbf{a} \in \{\mathbf{a}^0, \mathbf{a}^1, \mathbf{a}^2, \dots, \mathbf{a}^m\}$$

Continuous

$$\mathbf{a} \in \mathbb{R}^m$$



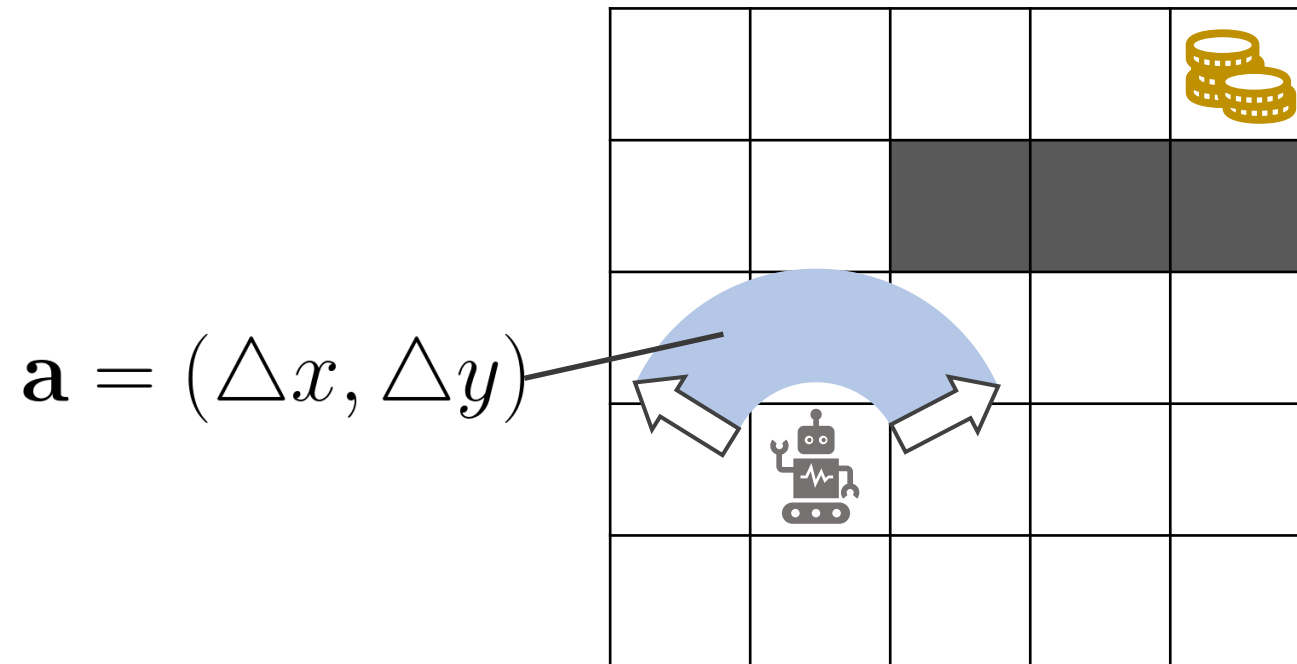
Action Spaces

Discrete

$$\mathbf{a} \in \{\mathbf{a}^0, \mathbf{a}^1, \mathbf{a}^2, \dots, \mathbf{a}^m\}$$

Continuous

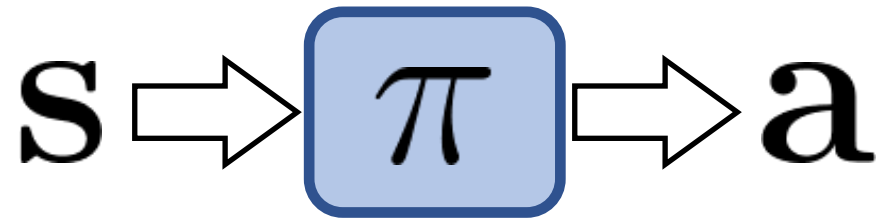
$$\mathbf{a} \in \mathbb{R}^m$$



Policies

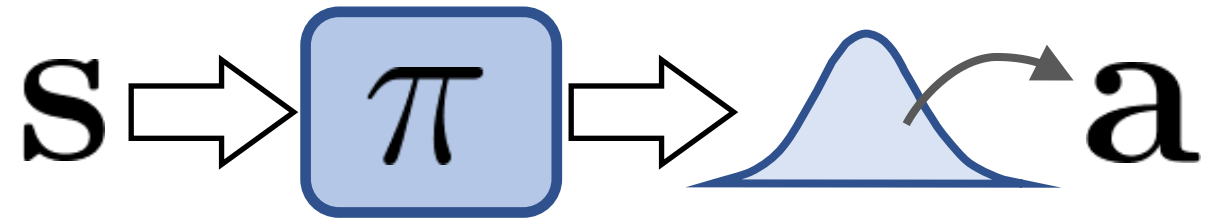
Deterministic

$$\mathbf{a} = \pi(\mathbf{s})$$



Stochastic

$$\pi(\mathbf{a}|\mathbf{s})$$



Deterministic Policy

$$\pi^{\text{det}}(\mathbf{s}) = \mathbf{a}^*$$

$$\pi(\mathbf{a}|\mathbf{s}) = \begin{cases} 1 & \text{if } \mathbf{a} = \pi^{\text{det}}(\mathbf{s}) \\ 0 & \text{otherwise} \end{cases}$$

Deterministic Policy

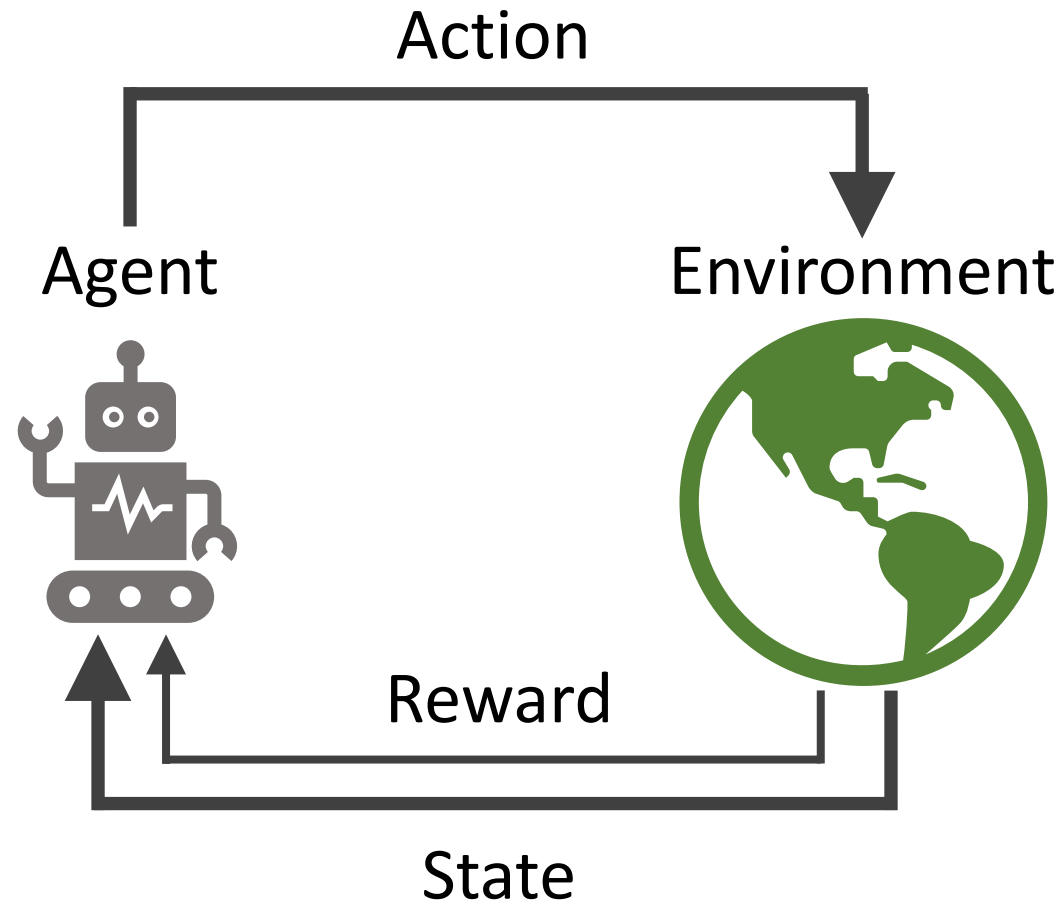
$$\pi(\mathbf{a}|\mathbf{s})$$

Deterministic Policy

$$\mathbf{a}^* = \arg \max_{\mathbf{a}} \pi(\mathbf{a}|\mathbf{s})$$

$$\pi^{\text{det}}(\mathbf{s}) = \mathbf{a}^*$$

Agent-Environment Interface



Partially Observable Markov Decision Process

$\mathbf{s} \in \mathcal{S}$ – state space

$\mathbf{o} \in \mathcal{O}$ – observation space

$\mathbf{a} \in \mathcal{A}$ – action space

$p(\mathbf{o}|\mathbf{s})$ – observation function

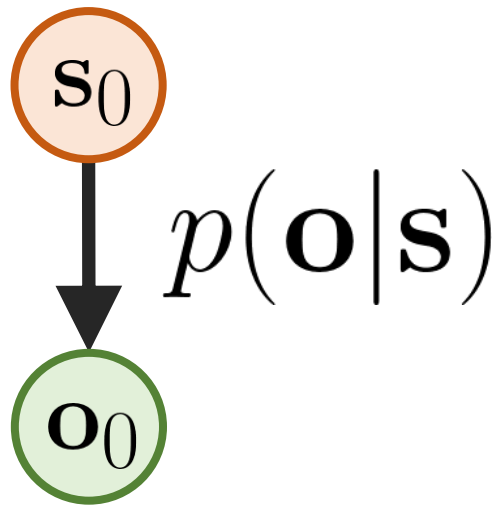
$p(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ – dynamics function

$r(\mathbf{s}, \mathbf{a}, \mathbf{s}')$ – reward function

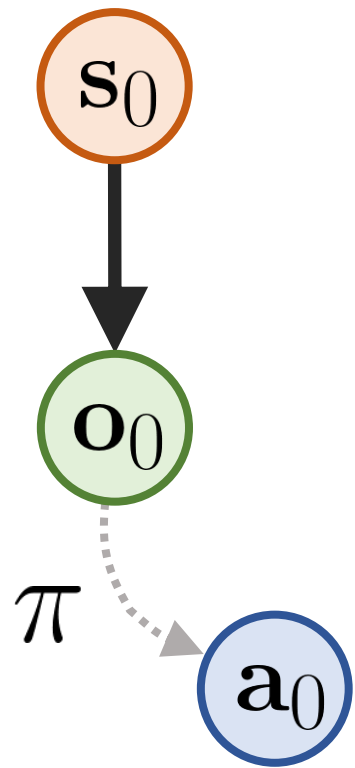
POMDP



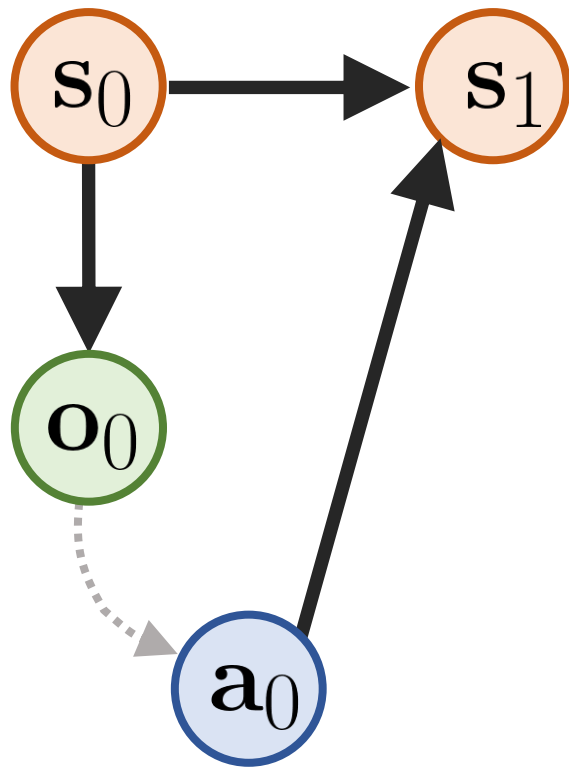
POMDP



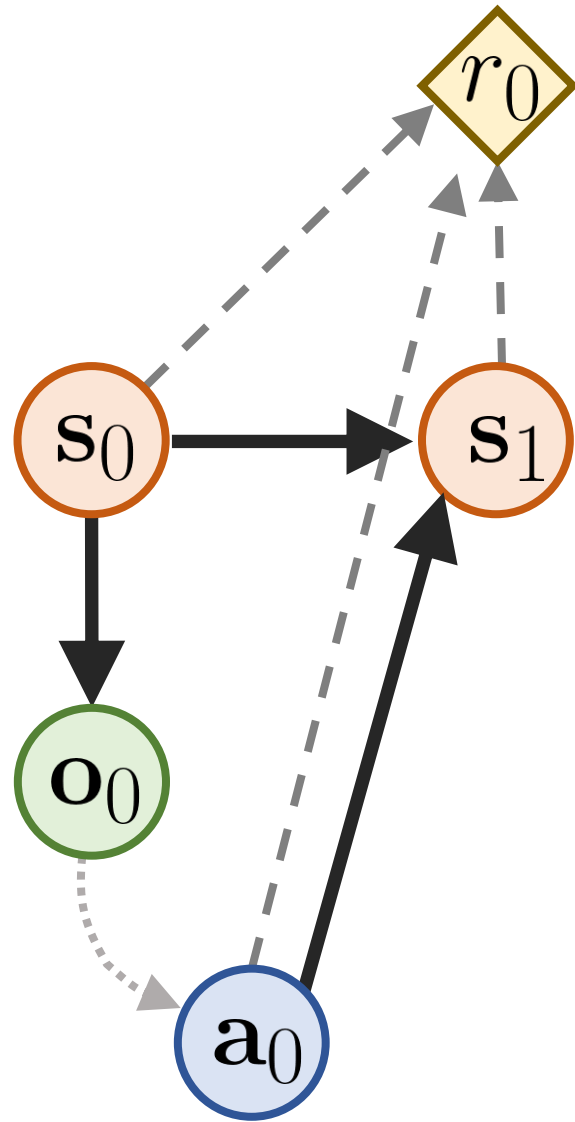
POMDP



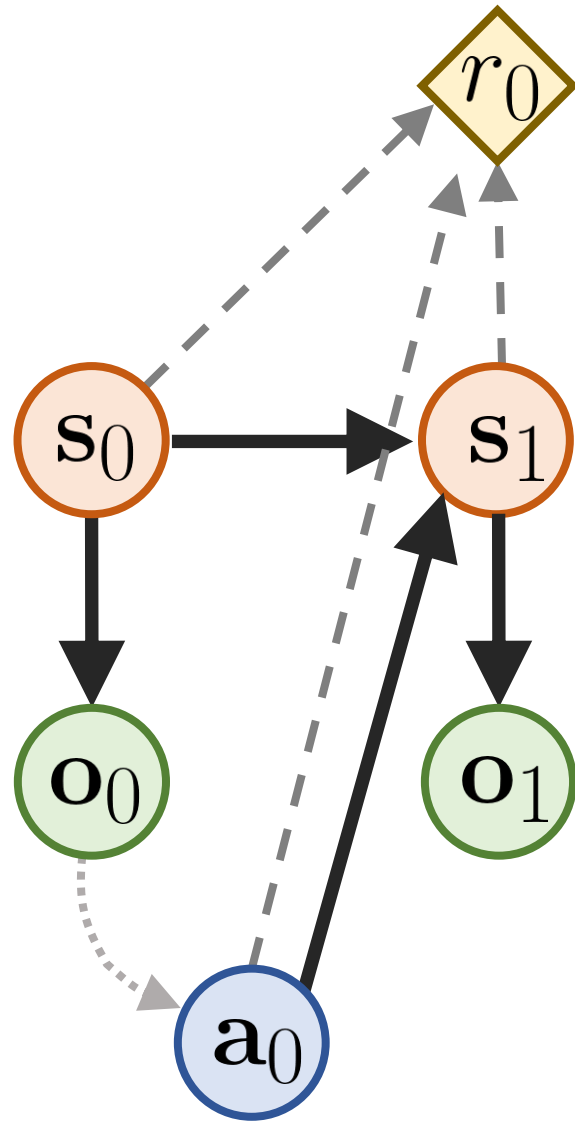
POMDP



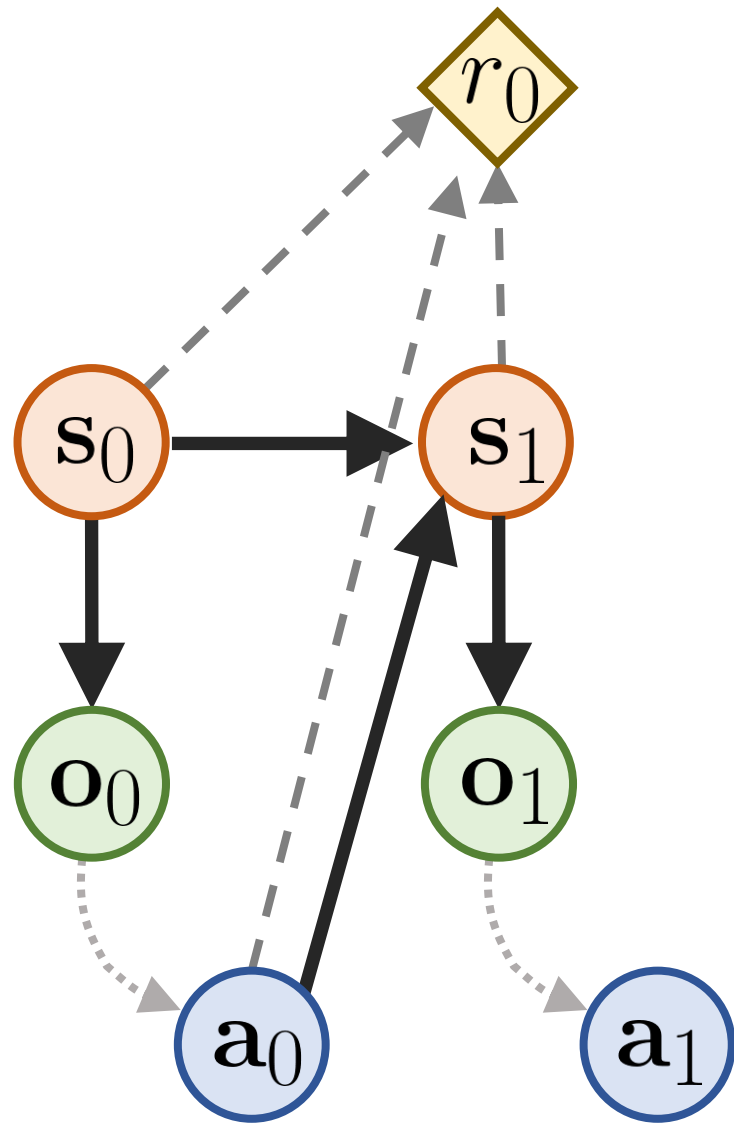
POMDP



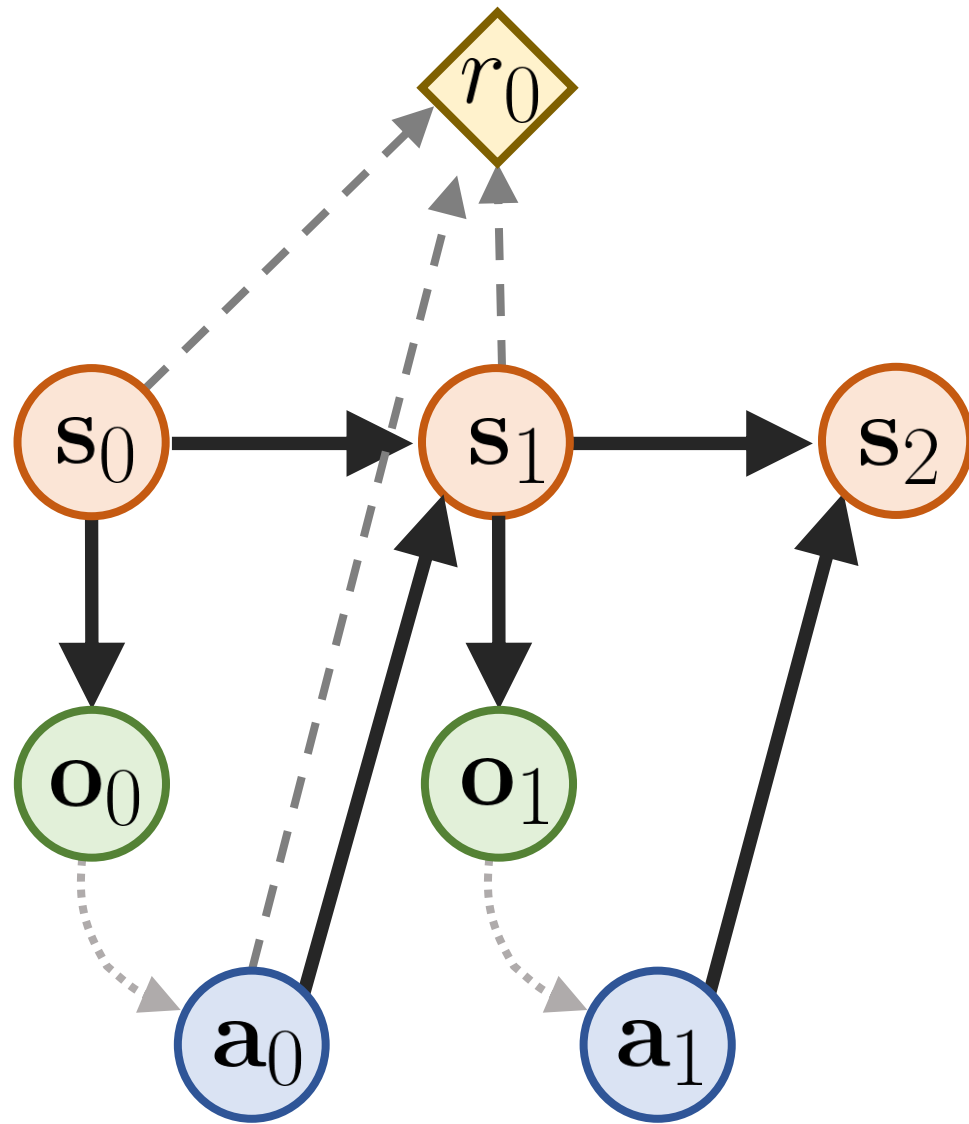
POMDP



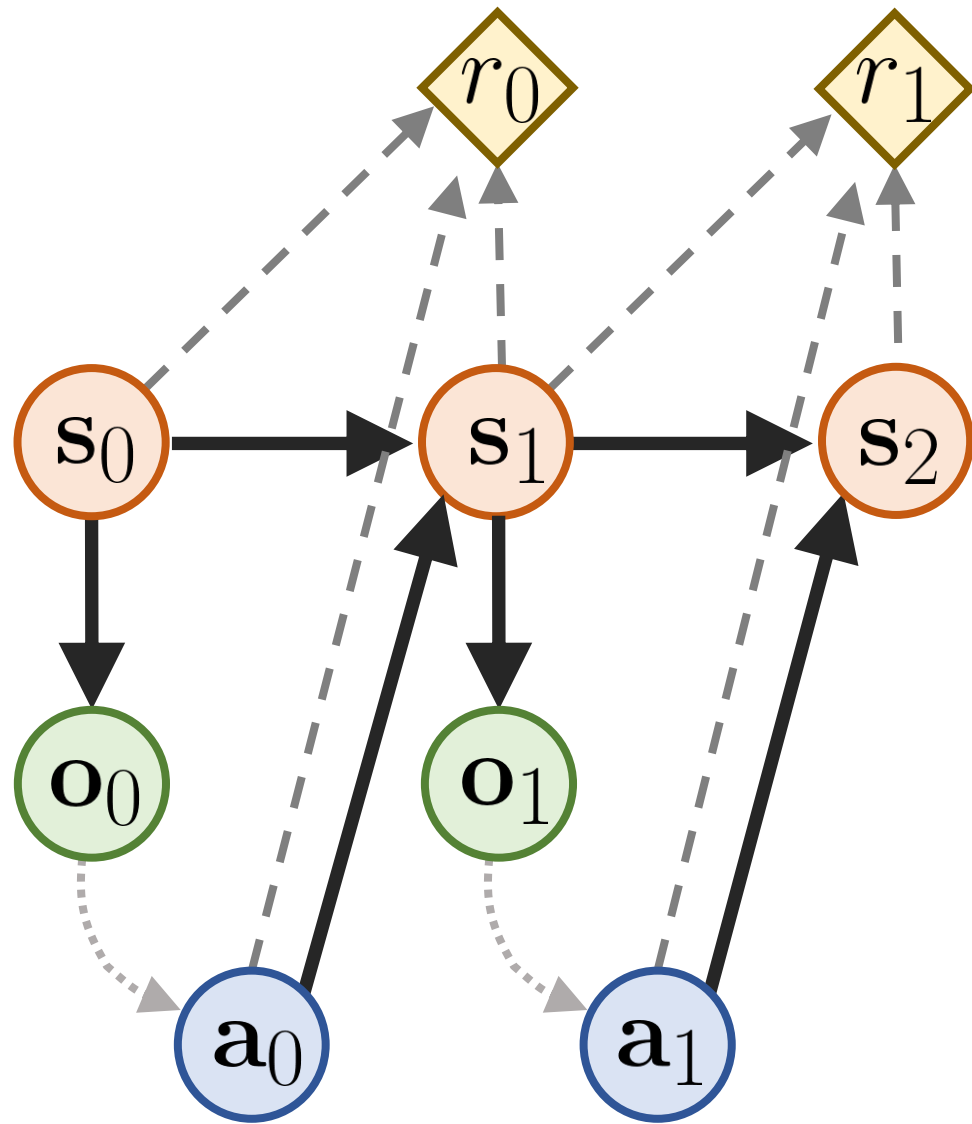
POMDP



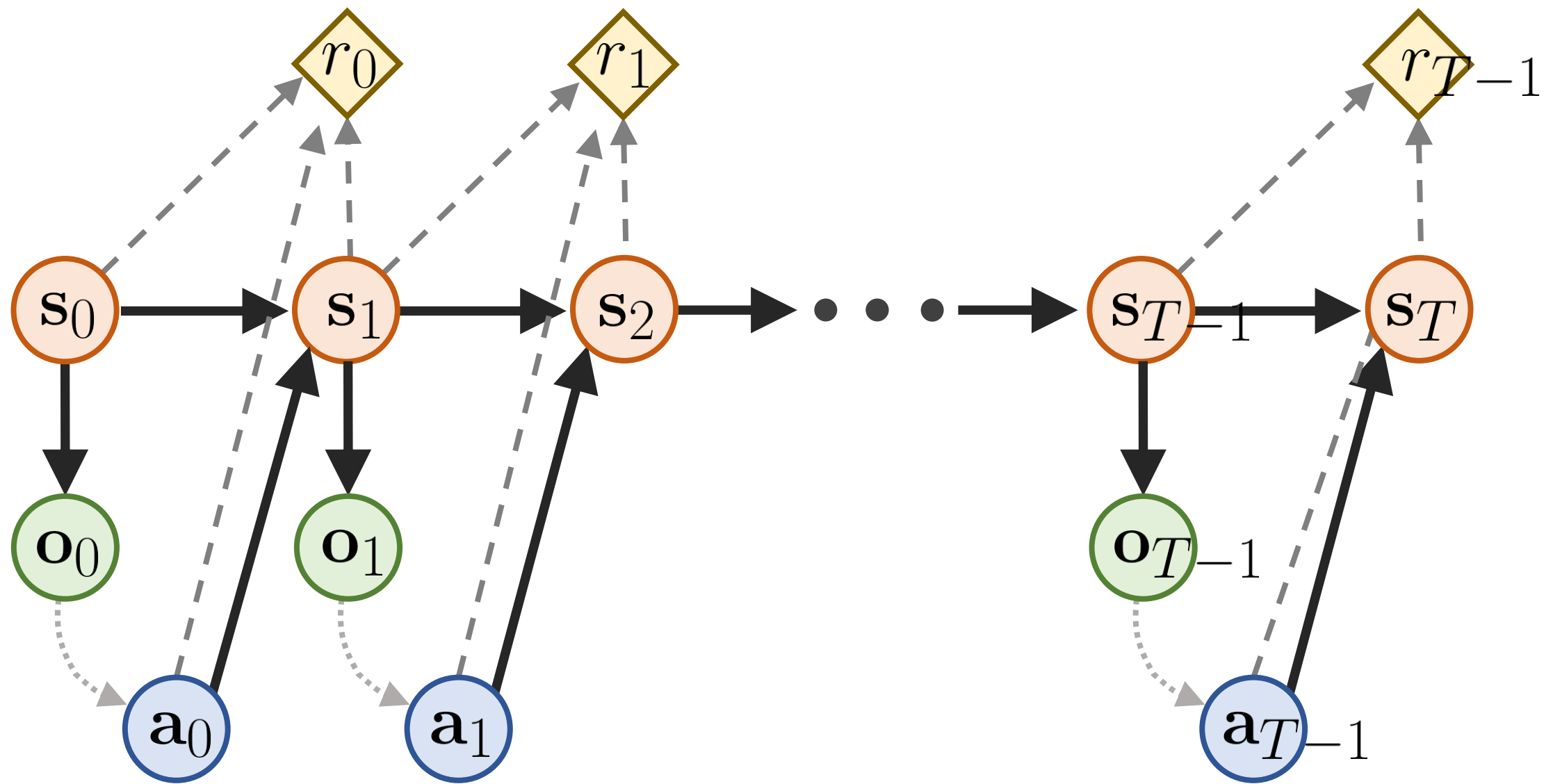
POMDP



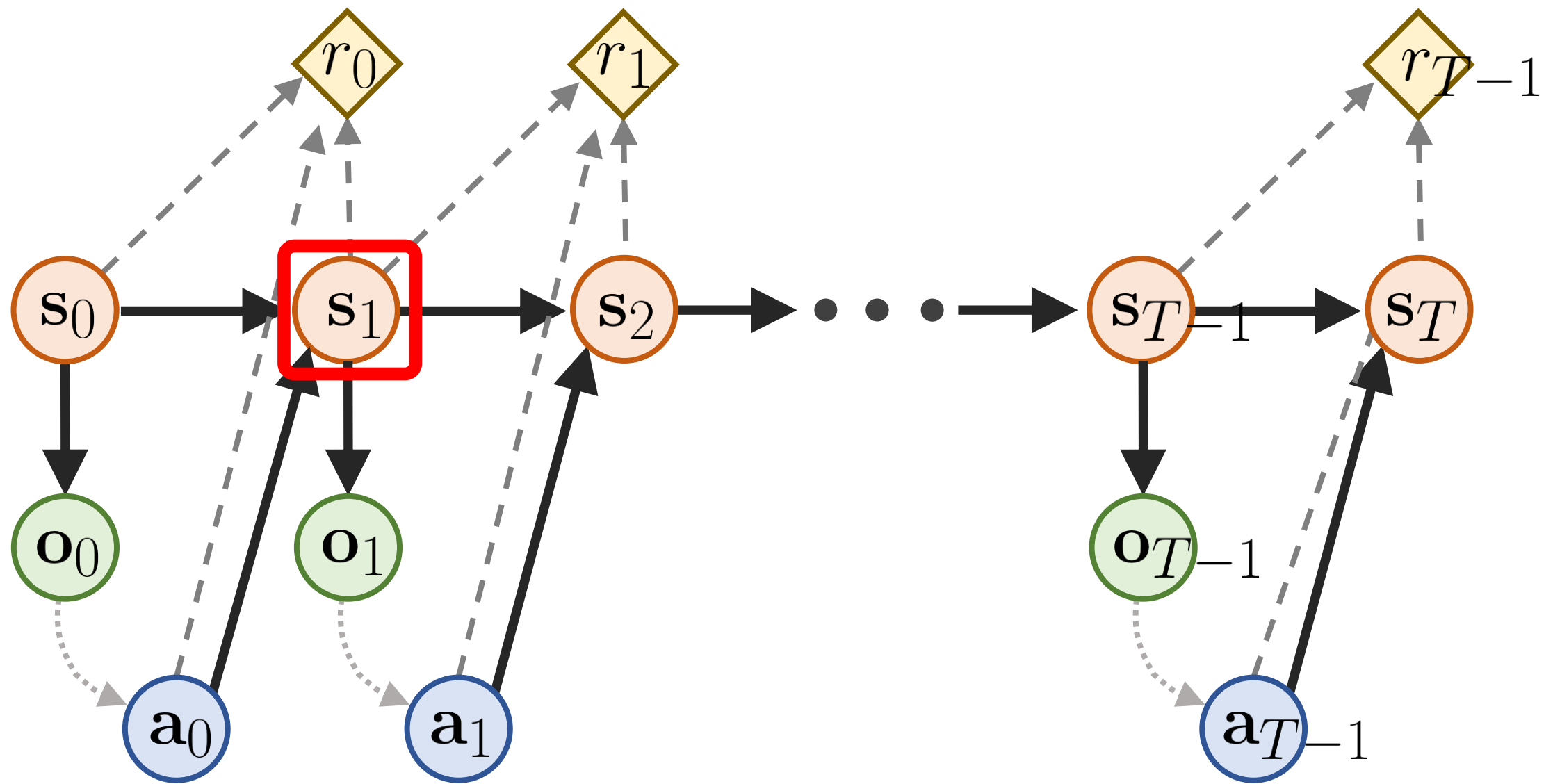
POMDP



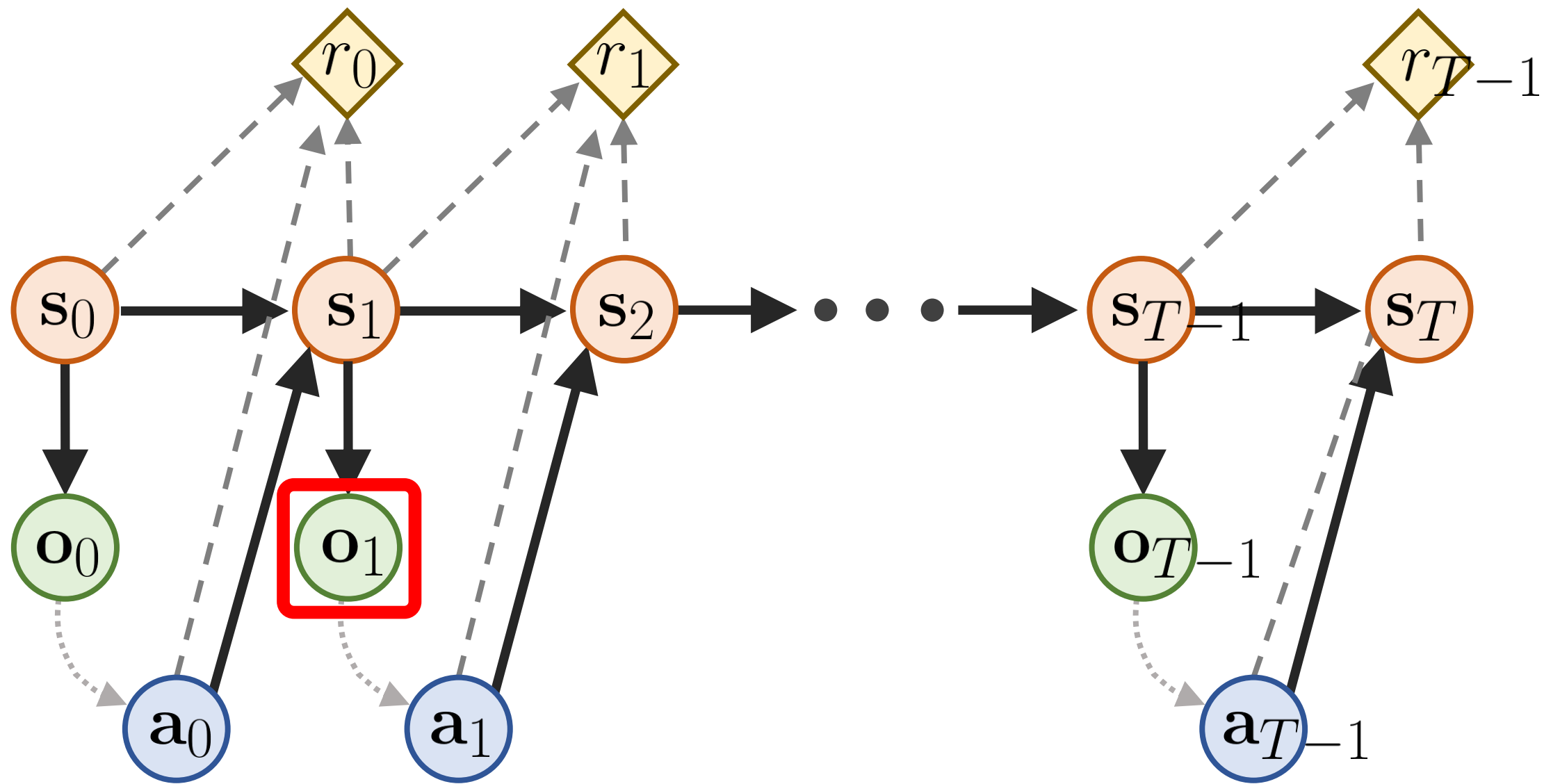
POMDP



POMDP



POMDP



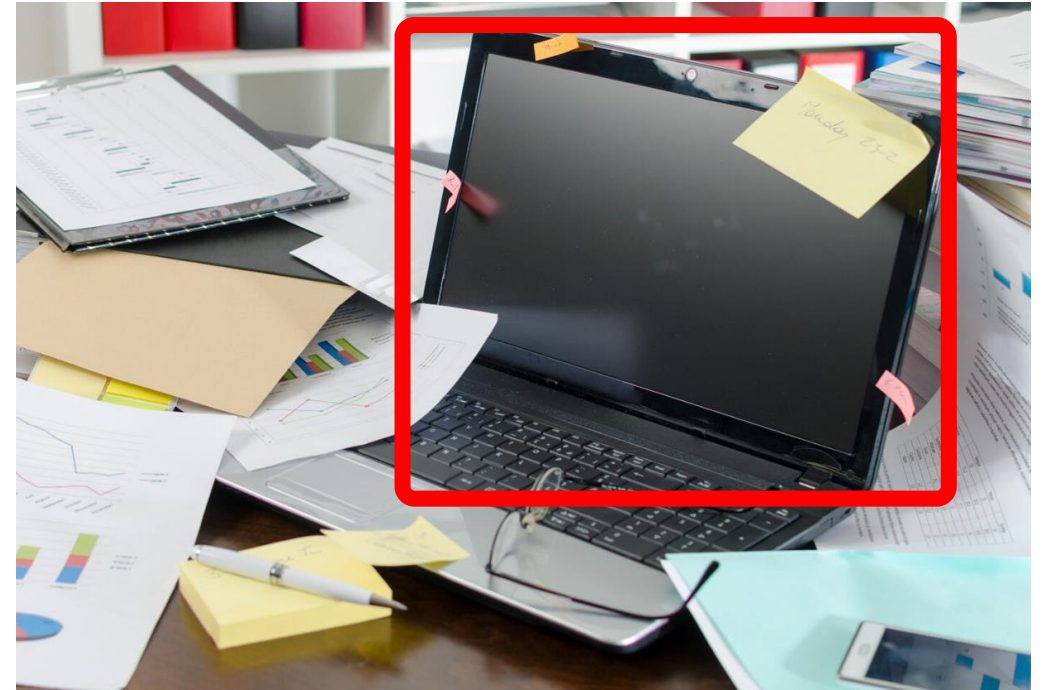
Observation Function

State:

- Location of every item

Observation

- image of the desk



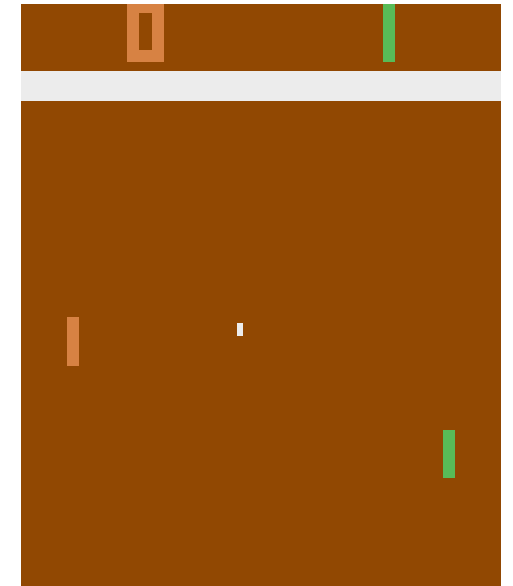
Pong

State:

- position + velocity of paddles
- position + velocity of the ball
- scores

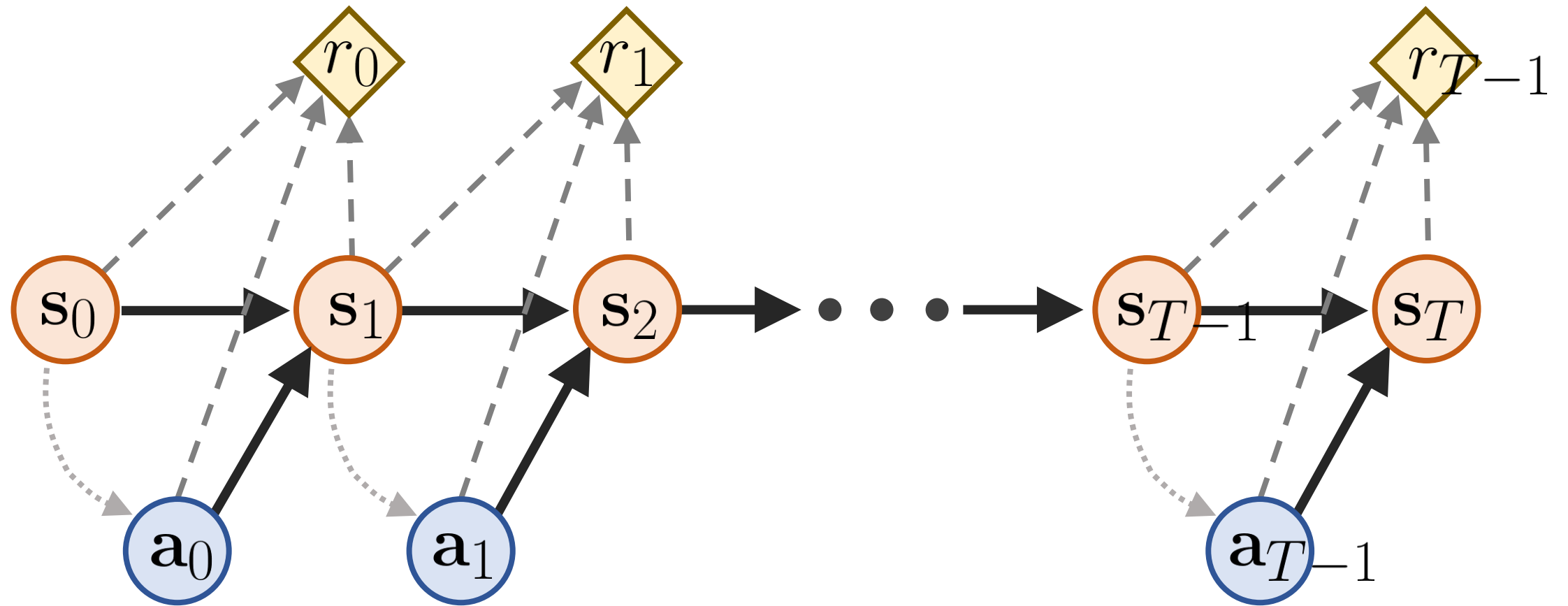
Observation

- image of the game screen



Pong [Atari]

MDP



Summary

- Agent-Environment Interface
- Markov Decision Processes
- Partially Observable Markov Decision Processes