Distillation of RL Policies with Formal Guarantees via Variational Abstraction of Markov Decision Processes

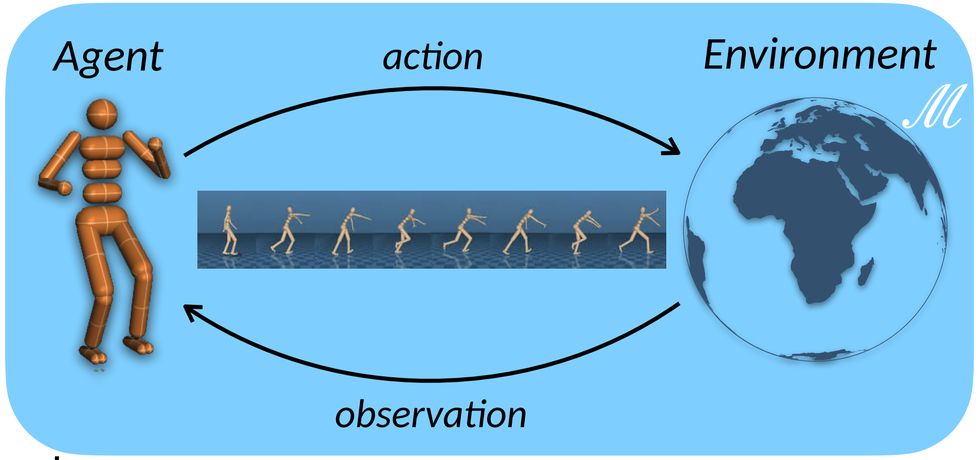
Florent Delgrange, Ann Nowé, Guillermo A. Pérez

AAAI-22

VUB ARTIFICIAL INTELLIGENCE RESEARCH GROUP

Universiteit Antwerpen

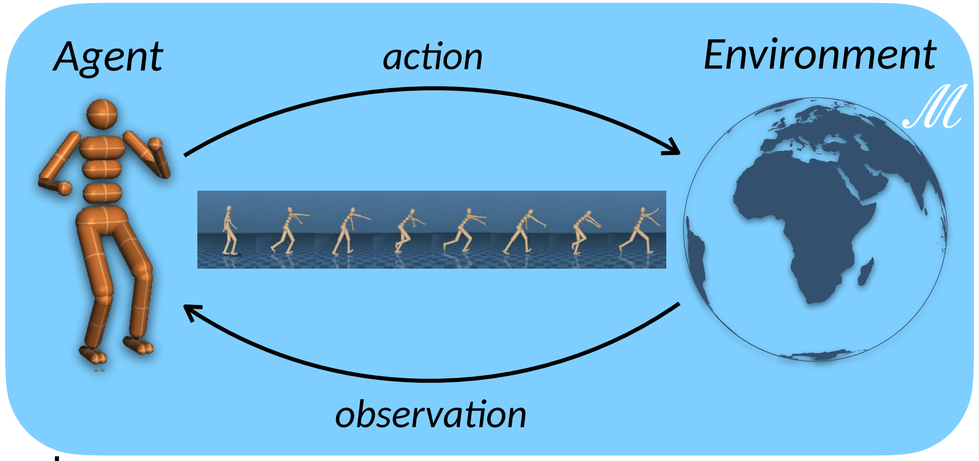
Reinforcement Learning



- Unknown environment
- Continuous state/action spaces

Overview

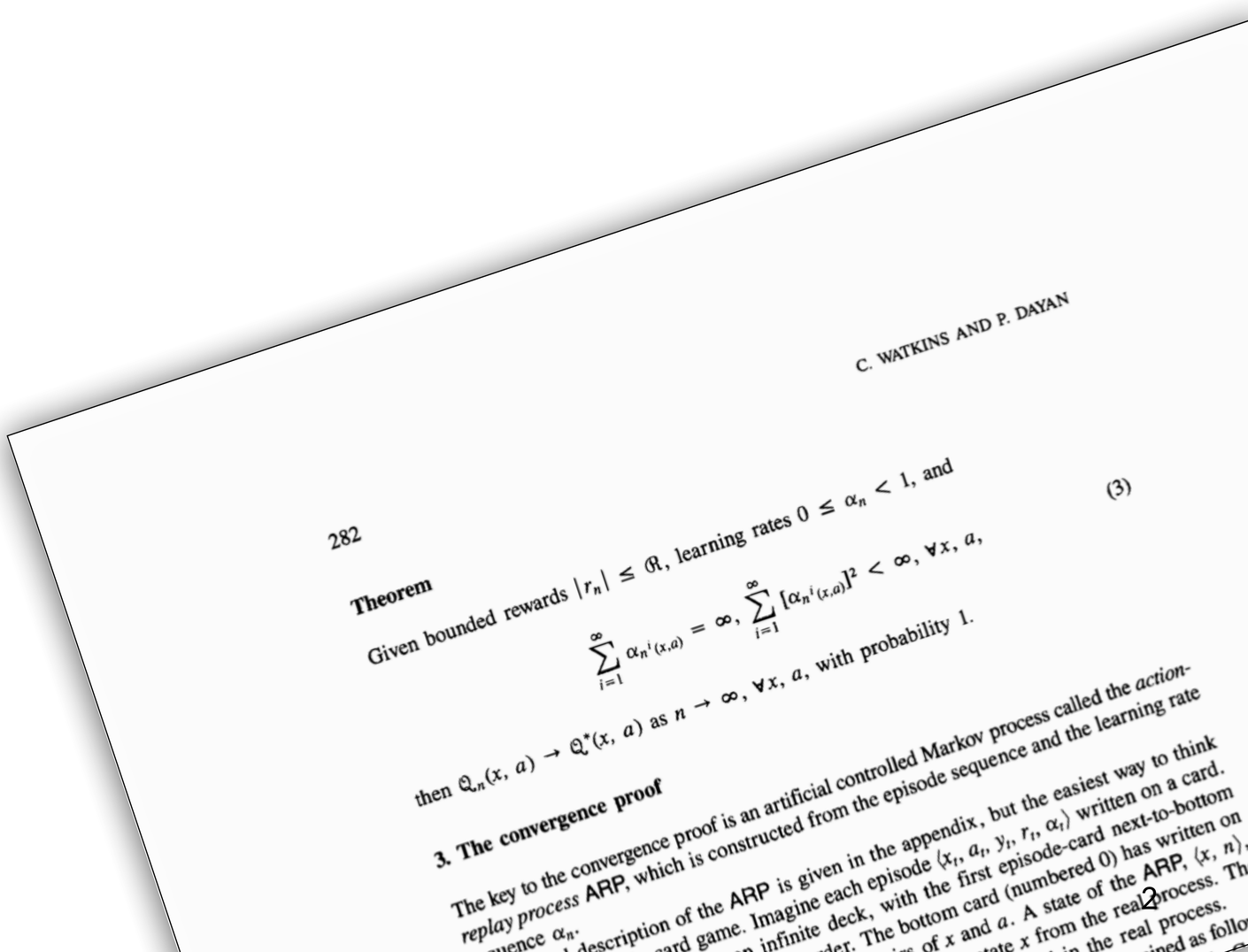
Reinforcement Learning



π

control policy

- Unknown environment
- Continuous state/action spaces


C. WATKINS AND P. DAYAN

282

Theorem

Given bounded rewards $|r_n| \leq \mathcal{R}$, learning rates $0 \leq \alpha_n < 1$, and

$$\sum_{i=1}^{\infty} \alpha_{n^i(x,a)} = \infty, \quad \sum_{i=1}^{\infty} [\alpha_{n^i(x,a)}]^2 < \infty, \ \forall x, a, \tag{3}$$

then $Q_n(x, a) \to Q^*(x, a)$ as $n \to \infty$, $\forall x, a$, with probability 1.

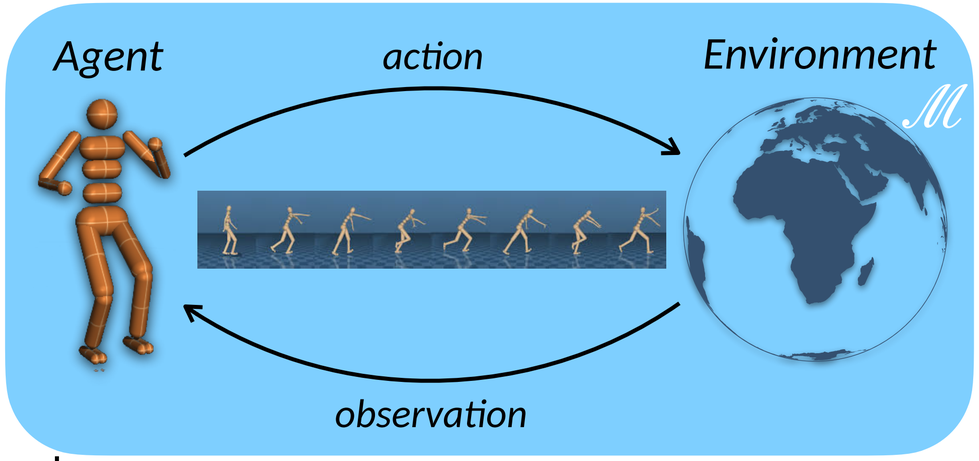
3. The convergence proof

The key to the convergence proof is an artificial controlled Markov process called the *action-replay process* ARP, which is constructed from the episode sequence and the learning rate sequence α_n.

description of the ARP is given in the appendix, but the easiest way to think card game. Imagine each episode $\langle x_t, a_t, y_t, r_t, \alpha_t \rangle$ written on a card. infinite deck, with the first episode-card next-to-bottom der. The bottom card (numbered 0) has written on of x and a. A state of the ARP, $\langle x, n \rangle$, tate x from the real process. Th in the real process. ined as follo

Overview

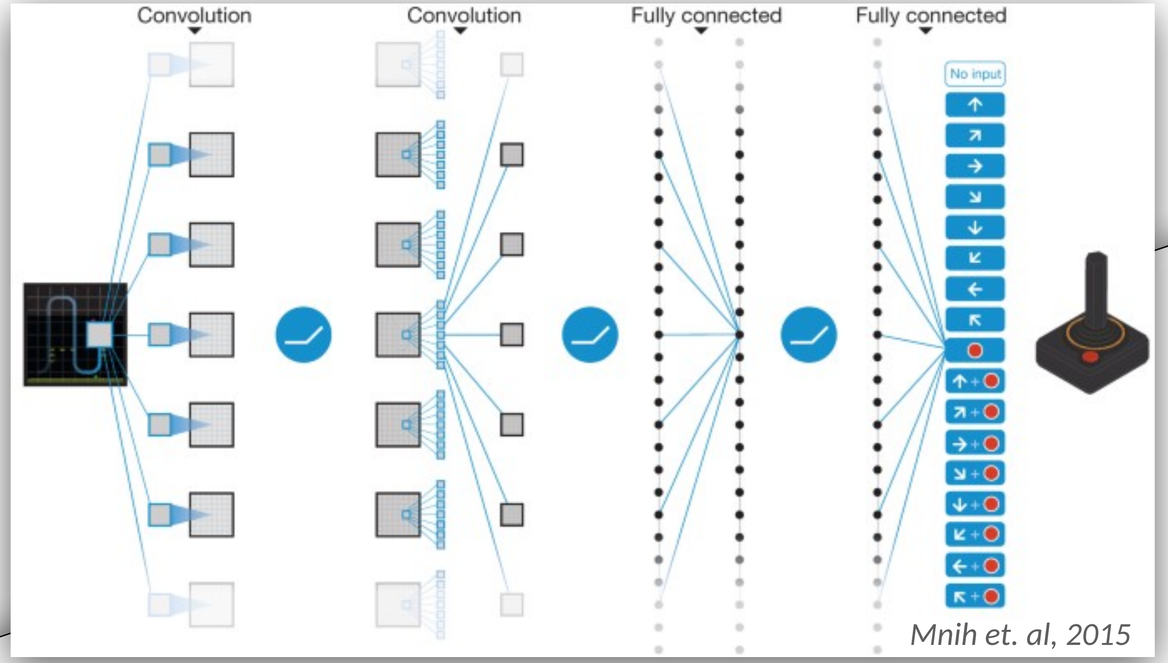
Reinforcement Learning



π

control policy

- Unknown environment
- Continuous state/action spaces



282

Theorem

Given bounded rewards $|r_n| \leq \mathcal{R}$, learning rates $0 \leq \alpha_n < 1$, and

$$\sum_{i=1}^{\infty} \alpha_{n^i(x,a)} = \infty, \quad \sum_{i=1}^{\infty} [\alpha_{n^i(x,a)}]^2 < \infty, \forall x, a,$$

then $Q_n(x, a) \to Q^*(x, a)$ as $n \to \infty$, $\forall x, a$, with probability 1.

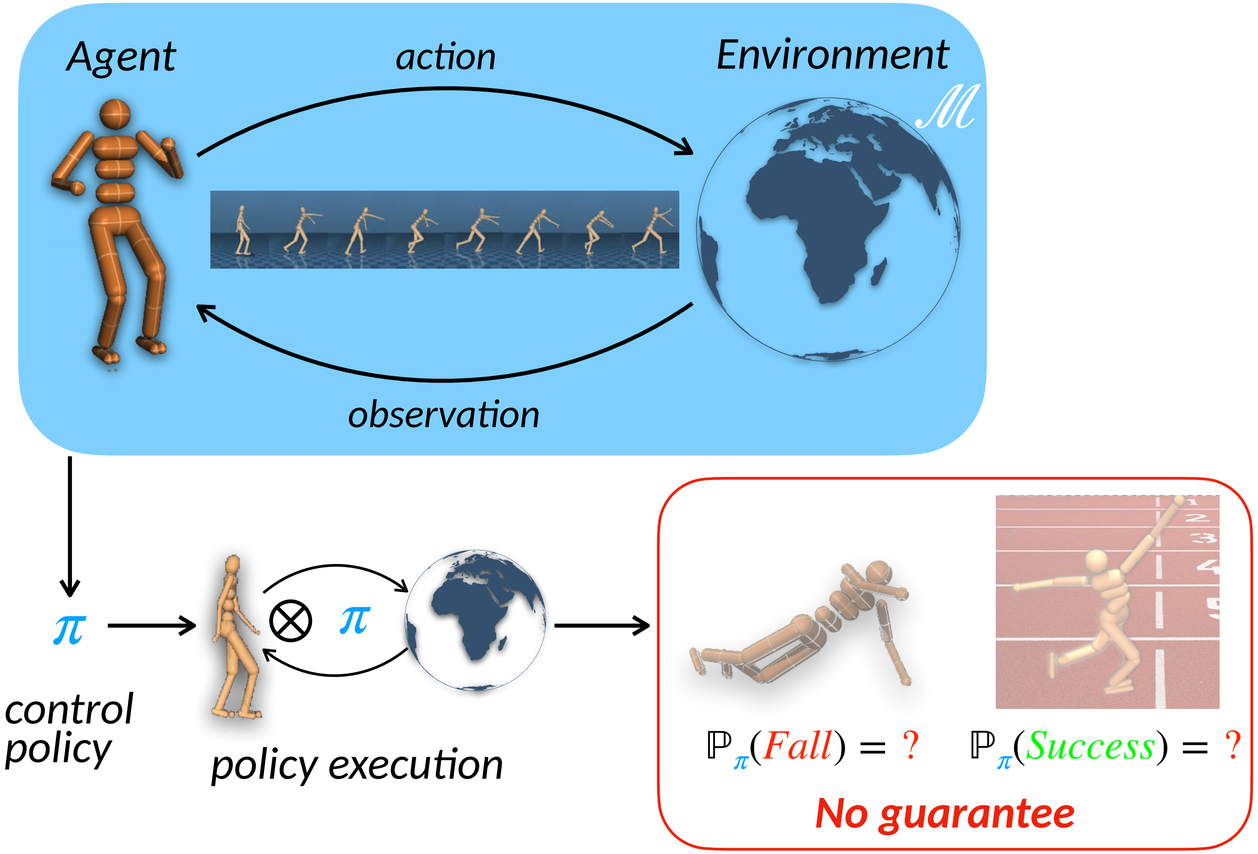
(3)

3. The convergence proof

The key to the convergence proof is an artificial controlled Markov process called the *action-replay process* ARP, which is constructed from the episode sequence and the learning rate sequence α_n.

description of the ARP is given in the appendix, but the easiest way to think card game. Imagine each episode $\langle x_t, a_t, y_t, r_t, \alpha_t \rangle$ written on a card. infinite deck, with the first episode-card next-to-bottom der. The bottom card (numbered 0) has written on of x and a. A state of the ARP, $\langle x, n \rangle$, state x from the real process. Th in the real process. ined as follo

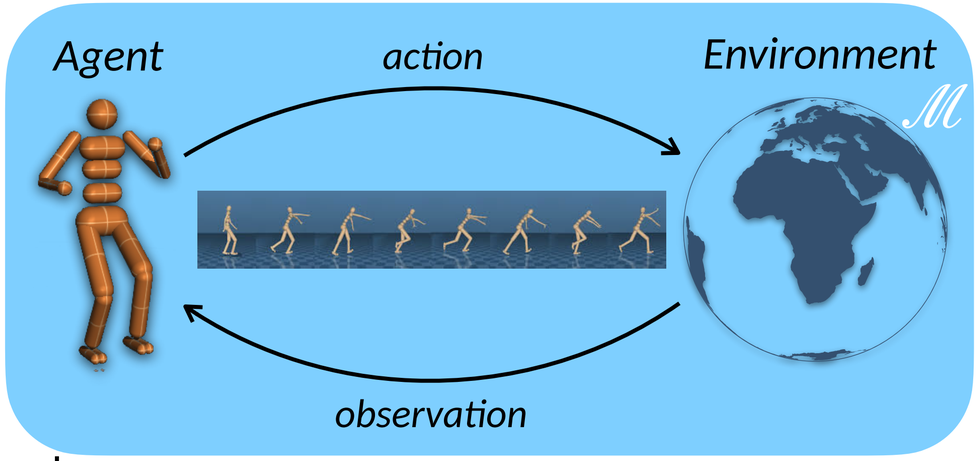
Reinforcement Learning

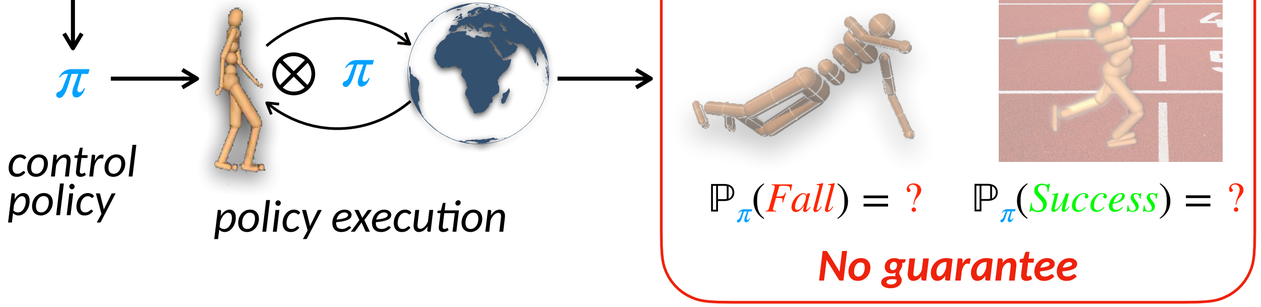


- Unknown environment
- Continuous state/action spaces

Overview

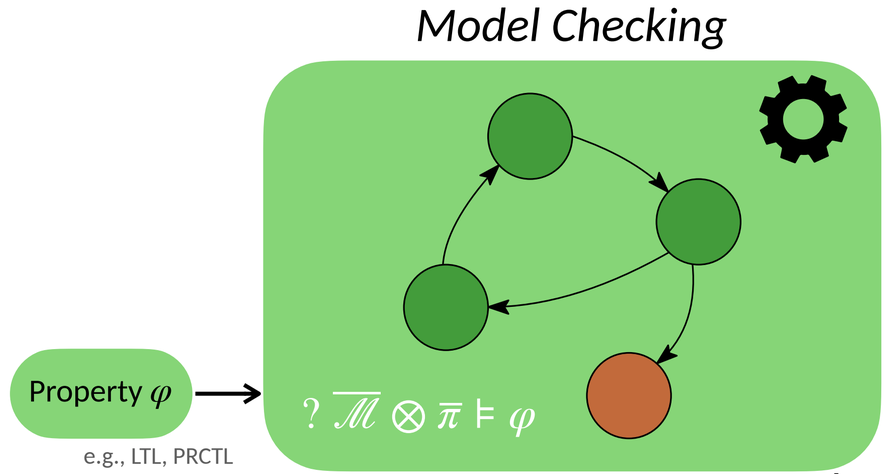
Reinforcement Learning

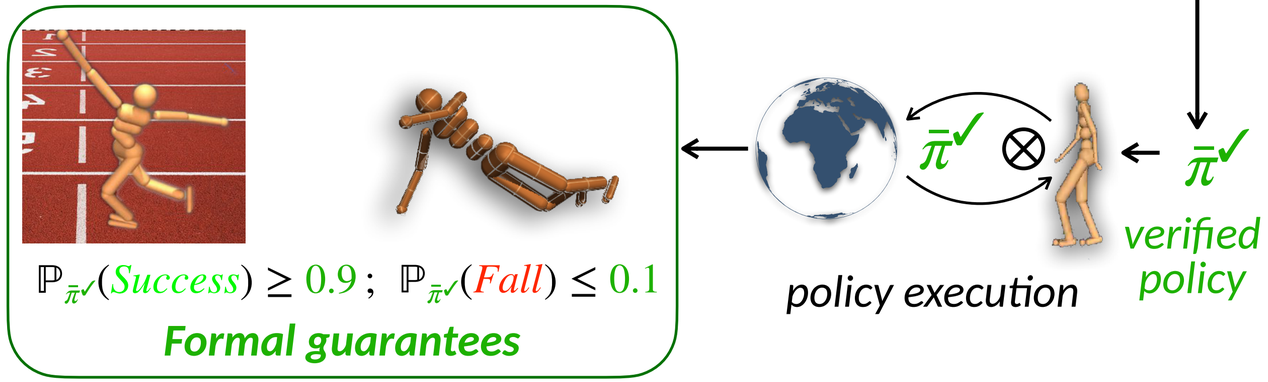




- Unknown environment
- Continuous state/action spaces

Formal Guarantees

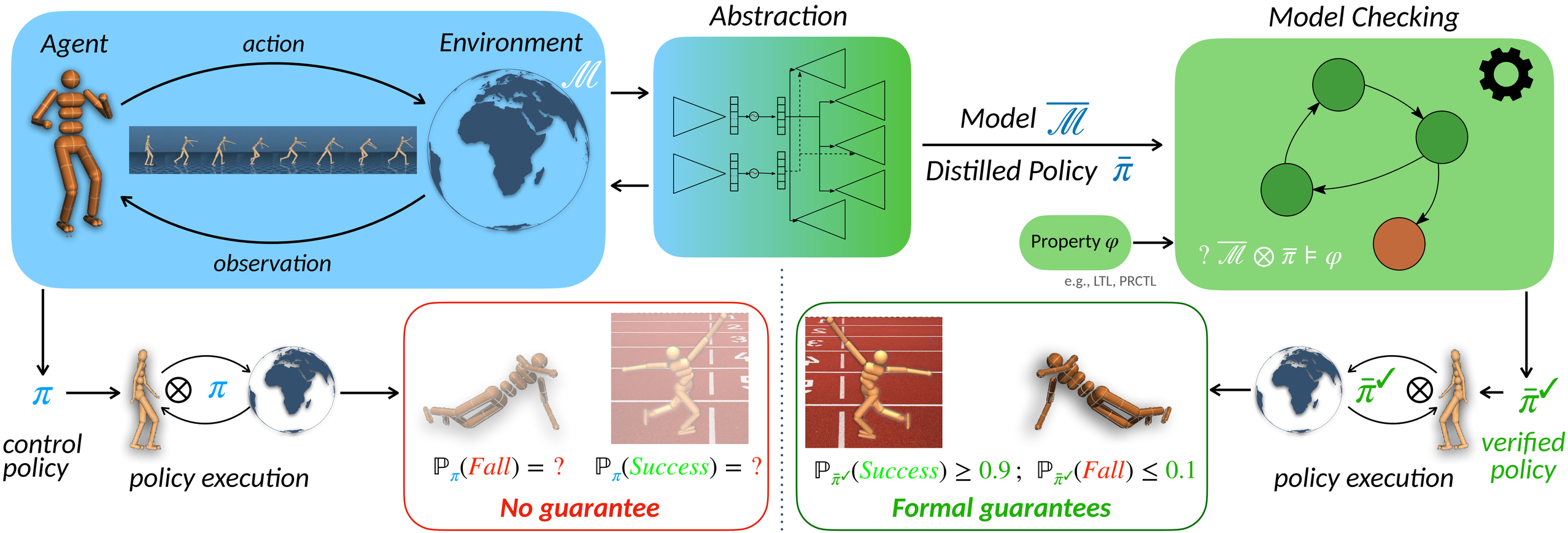




- Full knowledge of the model of the interaction
- Exhaustive exploration of the model
- Sensitive to the state space explosion problem

Overview

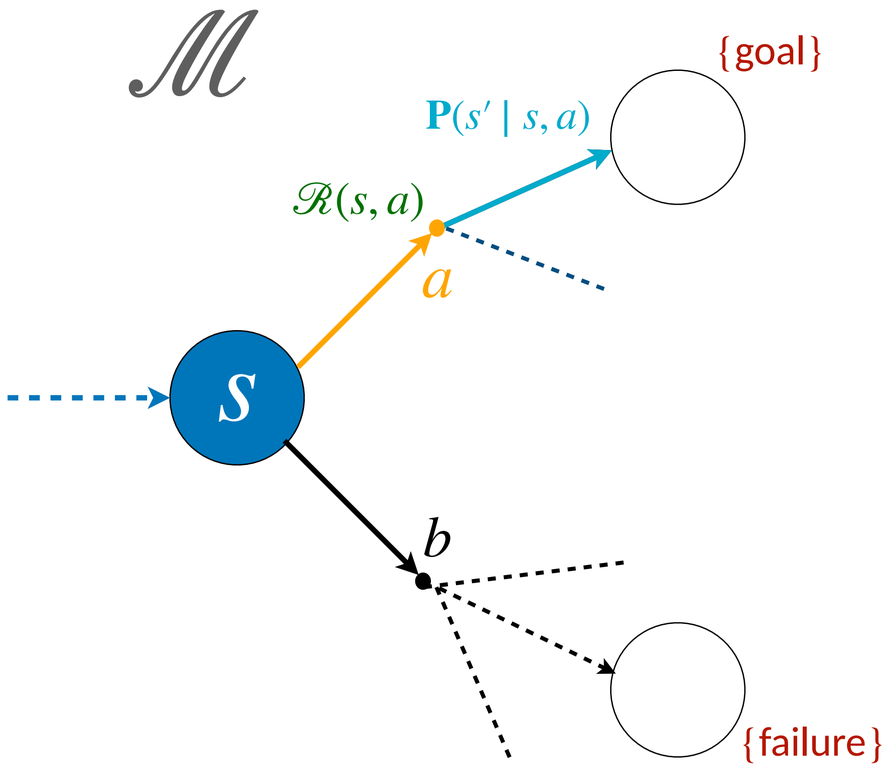
Reinforcement Learning Policies with Formal Guarantees



- Unknown environment
- Continuous state/action spaces

- Full knowledge of the model of the interaction
- Exhaustive exploration of the model
- Sensitive to the state space explosion problem

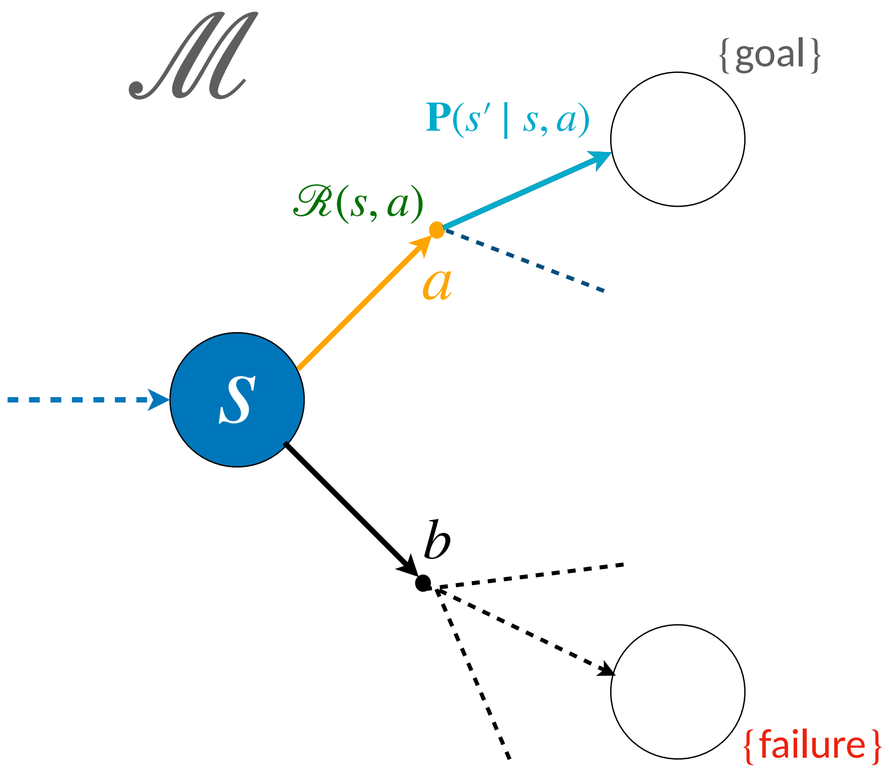
Markov Decision Processes



- State space $\mathscr{S}$
- Action space $\mathscr{A}$
- Reward function $\mathscr{R} : \mathscr{S} \times \mathscr{A} \to \mathbb{R}$
- Probability transition function $\mathbf{P}(s' \mid s, a)$
- Atomic propositions $\mathbf{AP}$ and labeling function $\mathscr{l} : \mathscr{S} \to 2^{\mathbf{AP}}$

Markov Decision Processes

Events, Policies, and Value Functions



- **Trajectories** $\langle s_{0:\infty}, a_{0:\infty} \rangle \in Traj(\mathscr{M})$ s.t. $\mathbf{P}\left(s_{t+1} \mid s_t, a_t\right) > 0, \forall t \in \mathbb{N}$
- **Events** $\varphi \subseteq Traj(\mathscr{M})$
  - *Example:* $\square \neg \text{failure} = \left\{ s_{0:\infty}, a_{0:\infty} \mid \forall i, \text{failure} \notin \ell\left(s_i\right) \right\}$
- **Policies** prescribe which action to choose at each step
  - ➡ $\pi \colon \mathcal{S} \to \mathscr{D}(\mathcal{A}), a_t \sim \pi(\cdot \mid s_t)$
  - ➡ Induce a **probability measure** $\mathbb{P}_\pi^{\mathscr{M}}$ over the events of $\mathscr{M}$
  - ➡ **Goal of RL:** $\arg\max_\pi \ \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \mathscr{R}(s_t, a_t) \right]$ for $\gamma \in]0,1[$
- **Value functions:**
  - $V_\pi(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \mathscr{R}(s_t, a_t) \mid s_0 = s \right]$
  - $V_\pi^\varphi(s)$, where $\lim_{\gamma \to 0} V_\pi^\varphi(s) = \mathbb{P}_\pi^{\mathscr{M}_s}\left(\varphi\right)$

Abstraction Quality Criterion

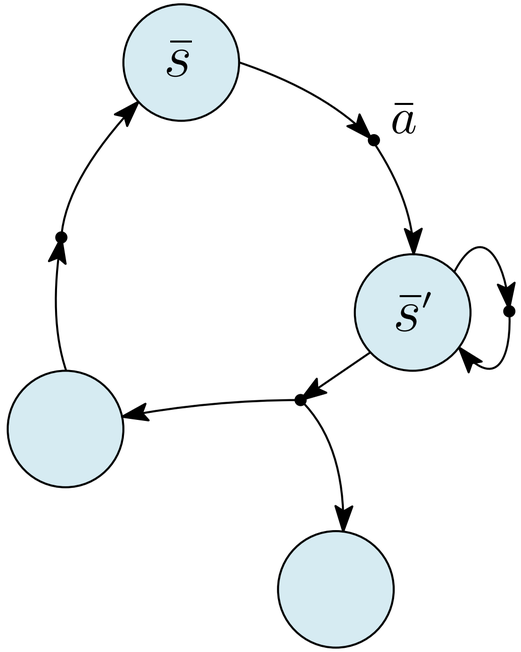
equivalent?

state embedding

$$\phi\colon \mathcal{S} \to \bar{\mathcal{S}},\ s \mapsto \bar{s}$$

action embedding

$$\psi : \mathcal{S} \times \bar{\mathcal{A}} \to \mathcal{A},\ \langle s, \bar{a} \rangle \mapsto a$$



Continuous-spaces MDP

$$\mathscr{M} = \langle \mathcal{S}, \mathcal{A}, \mathscr{R}, \mathbf{P}, \ell \rangle$$

Discrete latent MDP

$$\overline{\mathscr{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathscr{R}}, \overline{\mathbf{P}}, \ell \rangle$$

➡ A latent policy $\bar{\pi}\colon\ \overline{\mathcal{S}} \to \mathscr{D}(\overline{\mathcal{A}})$ can be executed in $\mathscr{M}$ via $\bar{a} \sim \bar{\pi}(\,\cdot \mid \phi(s)), a = \psi(s, \bar{a})$

Bisimulation

Bisimulation Relation

$B \in \mathcal{S}^2$ is a ***stochastic bisimulation*** iff for all $s_1, s_2 \in \mathcal{S}, a \in \mathcal{A}, T/B$

$$\ell(s_1) = \ell(s_2) \qquad \mathscr{R}(s_1, a) = \mathscr{R}(s_2, a) \qquad \text{and} \qquad \mathbf{P}(T \mid s_1, a) = \mathbf{P}(T \mid s_2, a)$$

Largest: $\sim$

(Larsen and Skou 1989; Givan, Dean, and Greig 2003)

- Behavioral equivalence between states
- Compare two MDPs: take the disjoint union of their state space: $\mathcal{S} \uplus \overline{\mathcal{S}}$

➡ Trajectory, value, and optimal policy equivalence

➡ For a given formal logic $\mathscr{L}$, two bisimilar models satisfy the same set of properties, i.e.,

➡ **They *behave the same***

Bisimulation

Bisimulation Relation

$B \in \mathcal{S}^2$ is a ~~**stochastic bisimulation iff**~~ for all $s_1, s_2 \in \mathcal{S}, a \in \mathcal{A}, T/B$

$$\ell(s_1) = \ell(s_2) \qquad \mathscr{R}(s_1, a) \neq \mathscr{R}(s_2, a) + \epsilon \text{ and } \qquad \mathbf{P}(T \mid s_1, a) \neq \mathbf{P}(T \mid s_2, a) + \epsilon$$

Largest: $\sim$

(Larsen and Skou 1989; Givan, Dean, and Greig 2003)

- Behavioral equivalence between states
- Compare two MDPs: take the disjoint union of their state space: $\mathcal{S} \uplus \overline{\mathcal{S}}$

➡ Trajectory, value, and optimal policy equivalence

➡ For a given formal logic $\mathscr{L}$, two bisimilar models satisfy the same set of properties, i.e.,

➡ **They *behave the same***

◉ *All or nothing:* two states *nearly identical* with slight numerical difference ϵ are $\neq$

Bisimulation distance

Continuous-spaces MDP

distance?

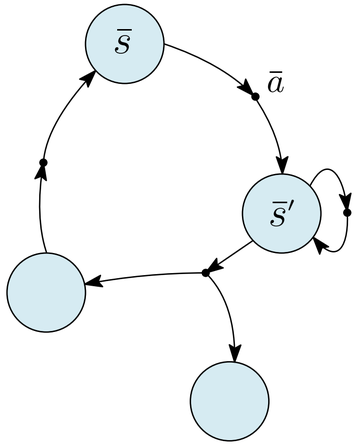
state embedding

$\phi: \mathcal{S} \to \bar{\mathcal{S}},\ s \mapsto \bar{s}$

action embedding

$\psi : \mathcal{S} \times \bar{\mathcal{A}} \to \mathcal{A},\ \langle s, \bar{a} \rangle \mapsto a$

Discrete latent MDP



$\mathscr{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbf{P}, \ell \rangle$

$\overline{\mathscr{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathcal{R}}, \overline{\mathbf{P}}, \ell \rangle$

- For policy π, $\gamma \in [0,1[$, and formal logic $\mathscr{L}$:

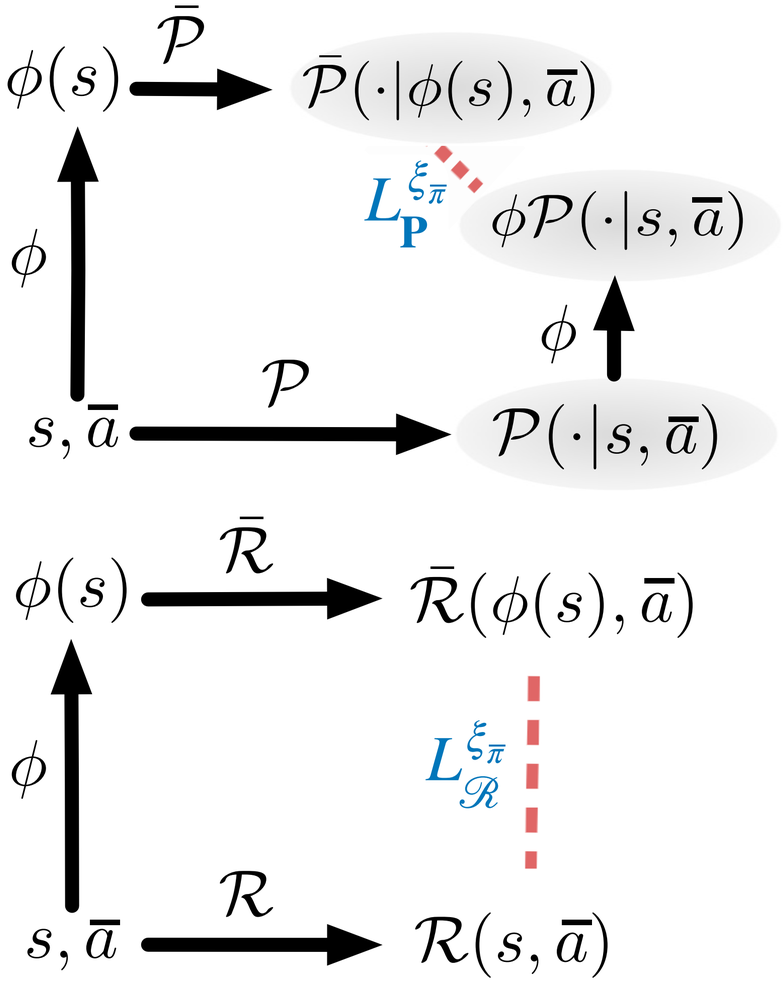
  ➡ ***Bisimulation distance:*** *largest behavioral difference* (*Desharnais et. al, 2004*)

$$\tilde{d}_\pi(s_1, s_2) = \sup_{f \in \mathcal{F}_\gamma^{\mathscr{L}}(\pi)} \left| f(s_1) - f(s_2) \right| \qquad \forall s_1, s_2 \in \mathcal{S}$$

where $\mathcal{F}_\gamma^{\mathscr{L}}(\pi)$ *is a logical family of functional expressions defining the semantics of* $\mathscr{L}$

➡ ***Kernel is bisimilarity:*** $\tilde{d}_\pi(s_1, s_2) = 0 \iff s_1 \sim s_2$

Checking the bisimulation distance via Local Losses Bounds



Gelada et. al, 2019

- Latent policy $\bar{\pi}$, stationary distribution $\xi_{\bar{\pi}}$

$$L_{\mathbf{P}}^{\xi_{\bar{\pi}}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} W_{d_{\bar{S}}}\left(\phi\mathbf{P}\left(\cdot \mid s,\bar{a}\right), \bar{\mathbf{P}}\left(\cdot \mid \phi(s),\bar{a}\right)\right)$$

$$L_{\mathscr{R}}^{\xi_{\bar{\pi}}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} \left|\mathscr{R}\left(s,\bar{a}\right) - \bar{\mathscr{R}}\left(\phi(s),\bar{a}\right)\right|$$

- *Expected bisimulation distance*

$$(1-\gamma)\,\mathbb{E}_{s\sim\xi_{\bar{\pi}}}\left|V_{\bar{\pi}}(s) - V_{\bar{\pi}}\left(\phi(s)\right)\right| \leq \mathbb{E}_{s\sim\xi_{\bar{\pi}}}\tilde{d}_{\bar{\pi}}^{\mathscr{R}}\left(s,\phi(s)\right) \leq L_{\mathscr{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}\frac{K_{\mathscr{R}}^{\bar{\pi}}}{1-\gamma K_{\mathbf{P}}^{\bar{\pi}}}$$

$$\mathbb{E}_{s\sim\xi_{\bar{\pi}}}\left|V_{\bar{\pi}}^{\varphi}(s) - V_{\bar{\pi}}^{\varphi}\left(\phi(s)\right)\right| \leq \mathbb{E}_{s\sim\xi_{\bar{\pi}}}\tilde{d}_{\bar{\pi}}^{\mathscr{L}}\left(s,\phi(s)\right) \leq \frac{\gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}$$

- **PAC scheme from samples**: let *trace* $\langle s_{0:T}, \bar{a}_{0:T-1}, r_{0:T-1}\rangle \sim \xi_{\bar{\pi}}$, $\epsilon, \delta \in]0,1[$ and $T \geq \left\lceil \frac{-\log(\delta/4)}{2\epsilon^2} \right\rceil$:

$$\hat{L}_{\mathscr{R}}^{\xi_{\bar{\pi}}} = \frac{1}{T}\sum_{t=0}^{T-1}\left|r_t - \bar{\mathscr{R}}\left(\phi(s_t),\bar{a}_t\right)\right| \quad \text{and} \quad \hat{L}_{\mathbf{P}}^{\xi_{\bar{\pi}}} = \frac{1}{T}\sum_{t=0}^{T-1}\left[1 - \bar{\mathbf{P}}\left(\phi(s_{t+1}) \mid \phi(s_t),\bar{a}_t\right)\right]$$

Then, $\left|L_{\mathscr{R}}^{\xi_{\bar{\pi}}} - \hat{L}_{\mathscr{R}}^{\xi_{\bar{\pi}}}\right| \leq \epsilon$ and $\left|\dot{L}_{\mathbf{P}}^{\xi_{\bar{\pi}}} - \hat{L}_{\mathbf{P}}^{\xi_{\bar{\pi}}}\right| \leq \epsilon$ **with probability** $1-\delta$

Variational Markov Decision Process

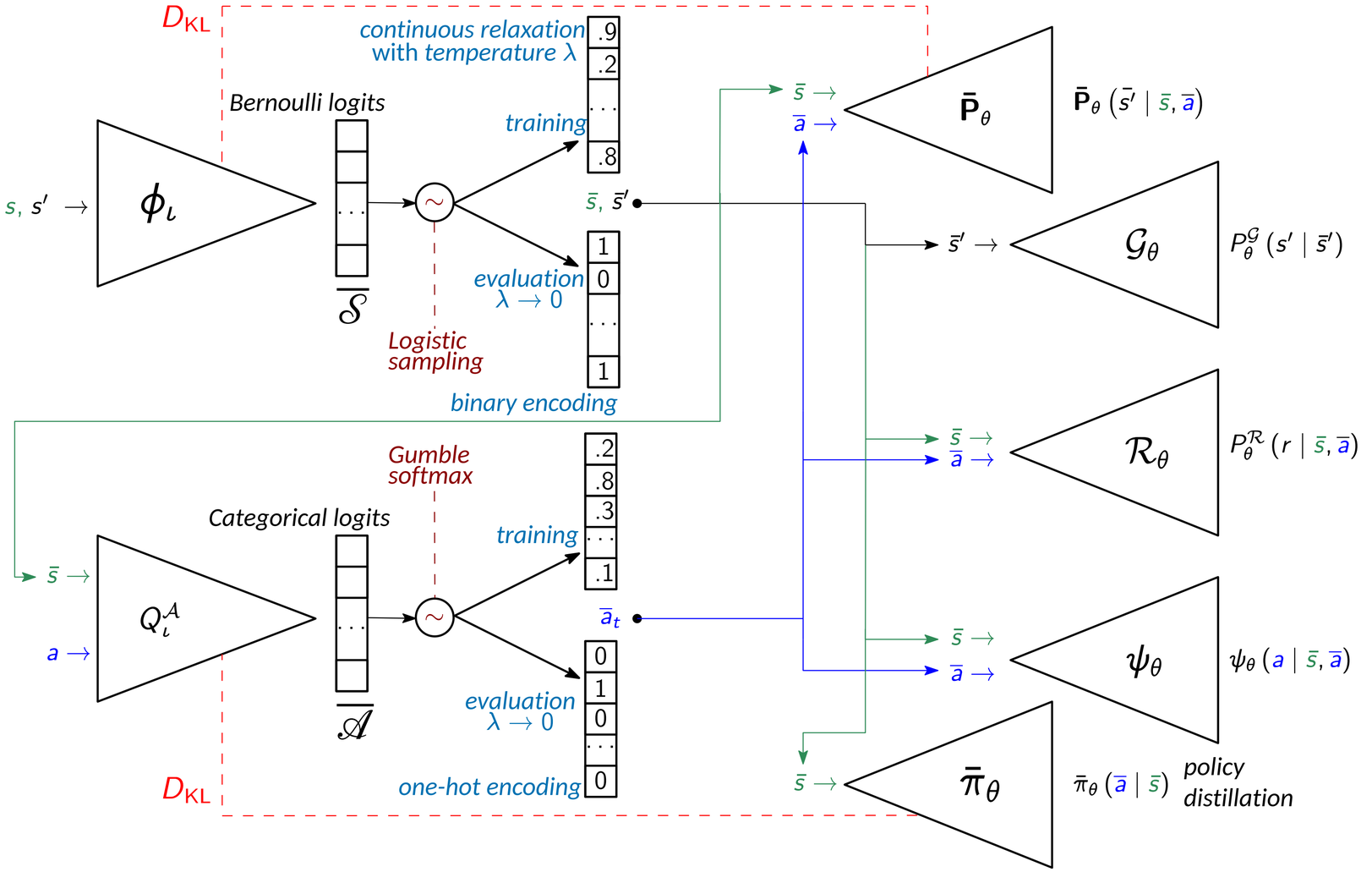
$$\max_{\iota,\theta} \; ELBO\left(\overline{\mathscr{M}}_\theta, \phi_\iota, \psi_\theta\right) = -\min_{\iota,\theta}\left\{\mathbf{D}_{\iota,\theta} + \mathbf{R}_{\iota,\theta}\right\}$$

$$\mathbf{D}_{\iota,\theta} = - \mathop{\mathbb{E}}_{\substack{s,a,r,s'\sim\xi_\pi \\ \bar{s},\bar{s}'\sim\phi_\iota(\cdot|s,s') \\ \bar{a}\sim Q_\iota^{\mathcal{A}}(\cdot|\bar{s},a)}} \left[\log P_\theta^{\mathcal{G}}(s' \mid \bar{s}') + \log \psi_\theta(a \mid \bar{s}, \bar{a}) + \log P_\theta^{\mathcal{R}}(r \mid \bar{s}, \bar{a})\right]$$

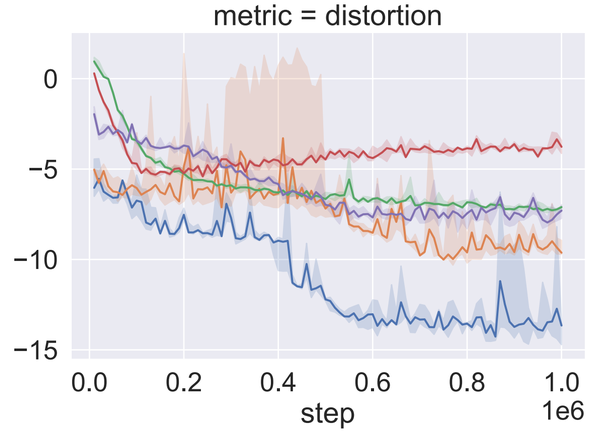
$$\mathbf{R}_{\iota,\theta} = \mathop{\mathbb{E}}_{\substack{s,a,s'\sim\xi_\pi \\ \bar{s}\sim\phi_\iota(\cdot|s) \\ \bar{a}\sim Q_\iota^{\mathcal{A}}(\cdot|\bar{s},a)}} \left[D_{\mathrm{KL}}\left(\phi_\iota(\cdot \mid s') \,\|\, \bar{\mathbf{P}}_\theta(\cdot \mid \bar{s}, \bar{a})\right) + D_{\mathrm{KL}}\left(Q_\iota^{\mathcal{A}}(\cdot \mid \bar{s}, a) \,\|\, \bar{\pi}_\theta(\cdot \mid \bar{s})\right)\right]$$

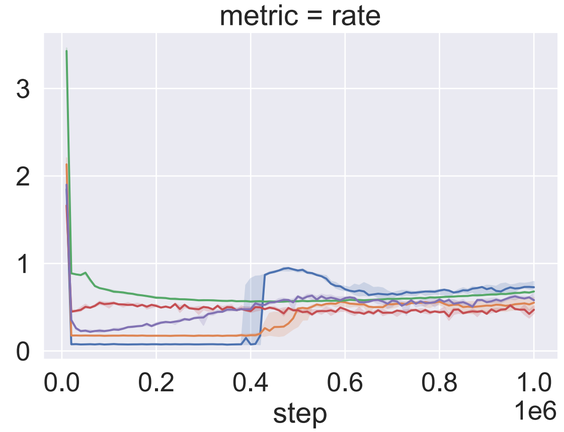
Discrete latent model:

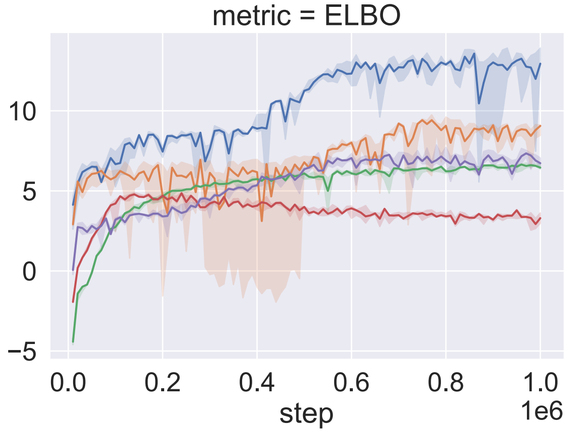
- $\langle \overline{\mathscr{M}}_\theta, \phi_\iota, \psi_{\iota,\theta} \rangle$
- $\overline{\mathscr{M}} = \langle \overline{\mathcal{S}}, \overline{\mathcal{A}}, \overline{\mathscr{R}}_\theta, \overline{\mathbf{P}}_\theta, \ell \rangle$

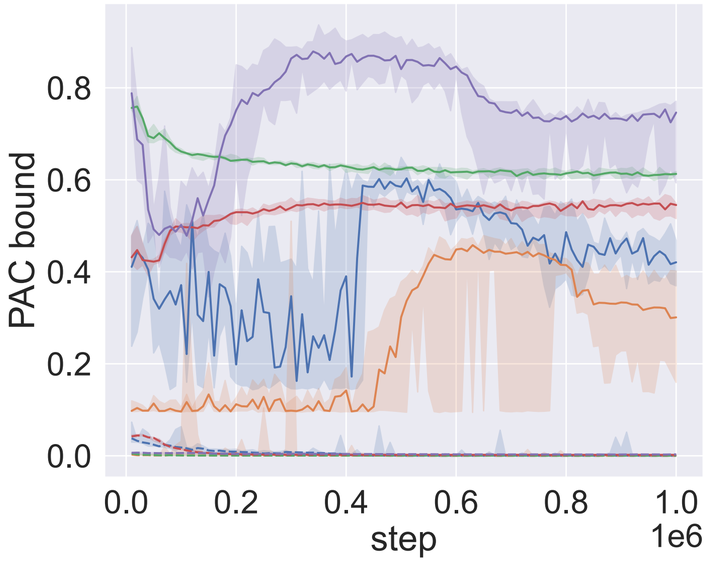


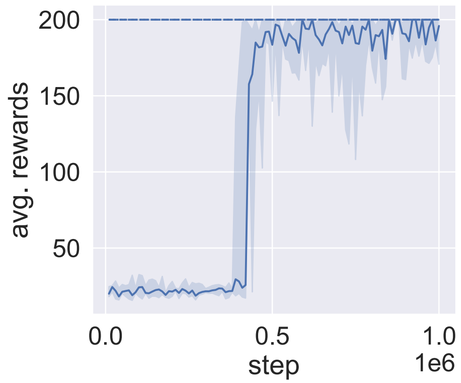
Evaluation

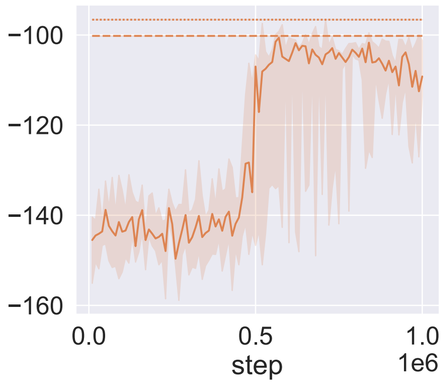

metric = distortion
0
−5
−10
−15
0.0
0.2
0.4
0.6
0.8
1.0
1e6
step

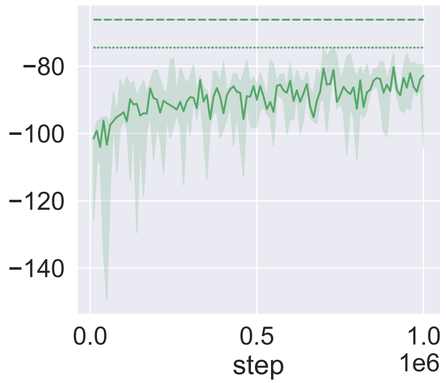

metric = rate
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
step

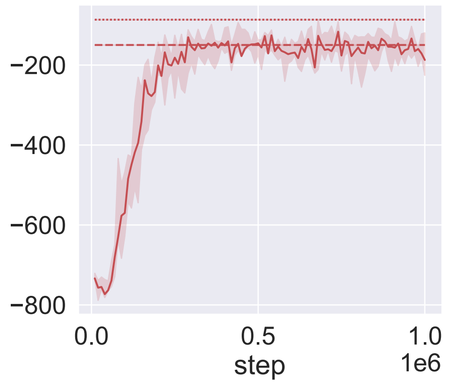

metric = ELBO
10
5
0
−5
0.0
0.2
0.4
0.6
0.8
1.0
1e6
step

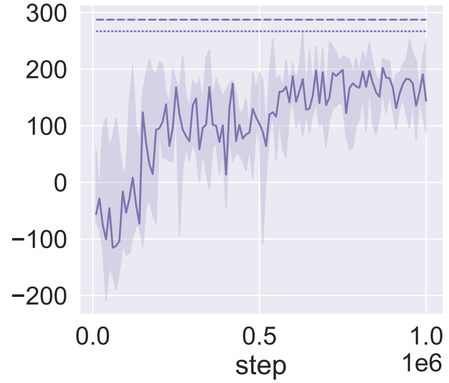

PAC bound
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1e6
step

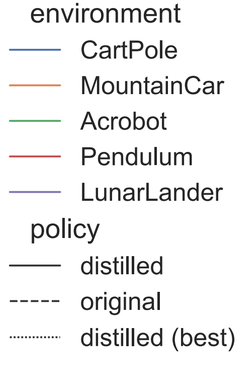

avg. rewards
200
150
100
50
0.0
0.5
1.0
1e6
step


−100
−120
−140
−160
0.0
0.5
1.0
1e6
step


−80
−100
−120
−140
0.0
0.5
1.0
1e6
step


−200
−400
−600
−800
0.0
0.5
1.0
1e6
step


300
200
100
0
−100
−200
0.0
0.5
1.0
1e6
step


environment
CartPole
MountainCar
Acrobot
Pendulum
LunarLander
policy
distilled
original
distilled (best)

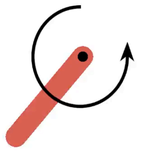
Evaluation

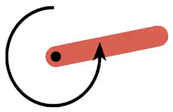
$\mathscr{M}_{\pi}$ ← *original*

$\mathscr{M}_{\overline{\pi}_\theta}$ ← *distilled*

Pendulum

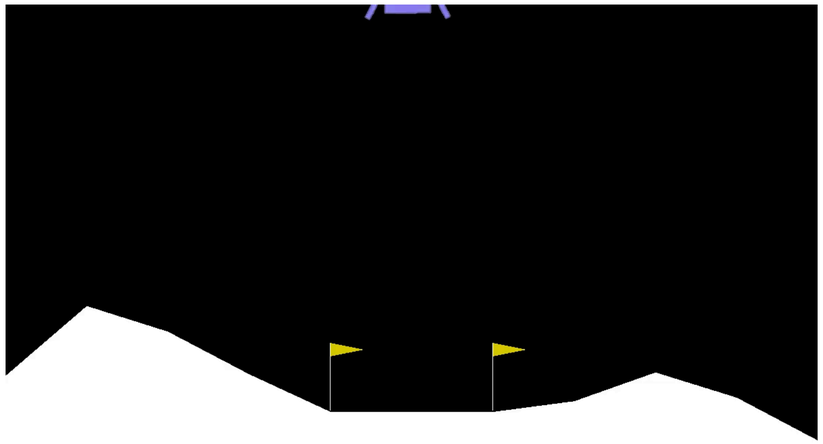
- $\mathcal{S} = [-1,1]^2 \times [-8,8] \subseteq \mathbb{R}^3$
- $\mathscr{A} = [-2,2] \subseteq \mathbb{R}$
- $\left|\overline{\mathcal{S}}\right| = 2^{13}$
- $\left|\overline{\mathscr{A}}\right| = 3$

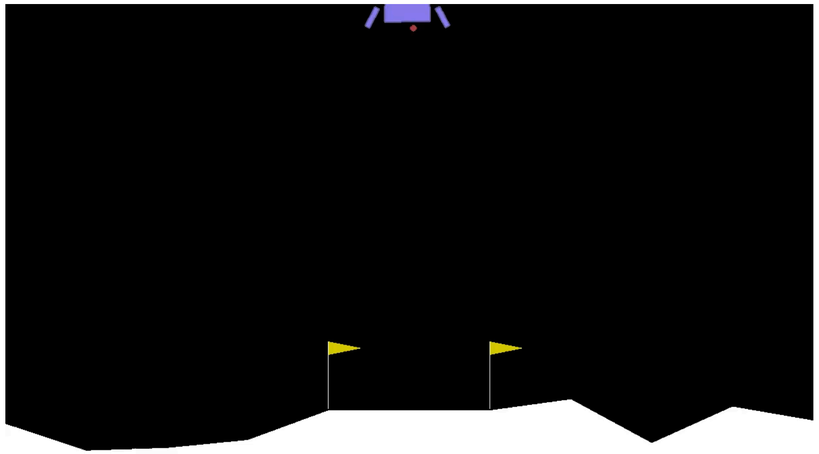




LunarLander

- $\mathcal{S} \subseteq \mathbb{R}^8$
- $\mathscr{A} = [-1,1]^2 \subseteq \mathbb{R}^2$
- $\left|\overline{\mathcal{S}}\right| = 2^{16}$
- $\left|\overline{\mathscr{A}}\right| = 5$





Conclusion

- ***VAE-MDPs***, a framework for learning **discrete latent models** of **unknown continuous-spaces** environment with **bisimulation guarantees**
  - Can be learned by executing an RL policy in the environment
  - Yields a **distilled version of the RL policy**
  - **New local losses bounds** for (i) bisimulation guarantees (ii) discrete setting (iii) action embedding function
  - **PAC schemes**, derived from the execution of the distilled policy
- Our tool can be used to **highlight the lack of robustness of input policies** when the distillation fails
- *Future work:*
  - Safe RL via formal methods
  - Real-world case study
  - Multi agent systems