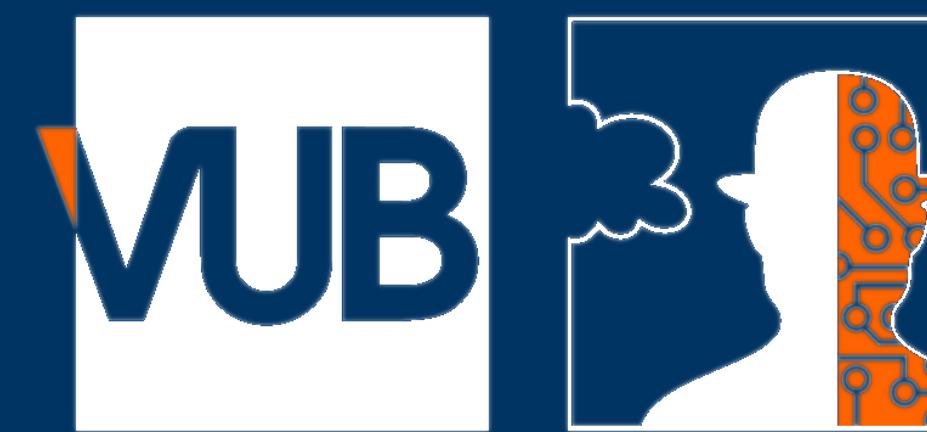


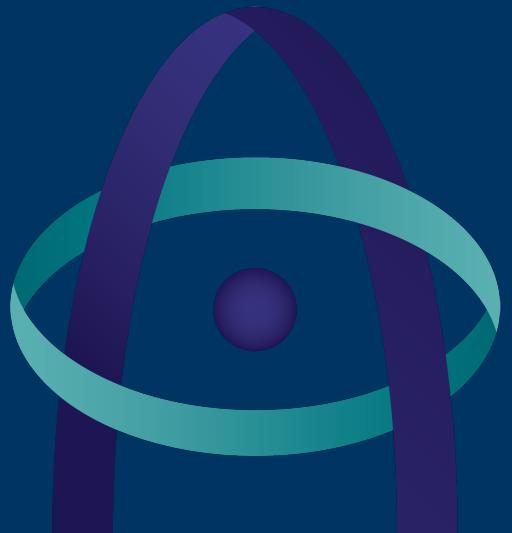
# Distillation of RL Policies with Formal Guarantees via Variational Abstraction of Markov Decision Processes

*Florent Delgrange, Ann Nowé, Guillermo A. Pérez*

December 2021

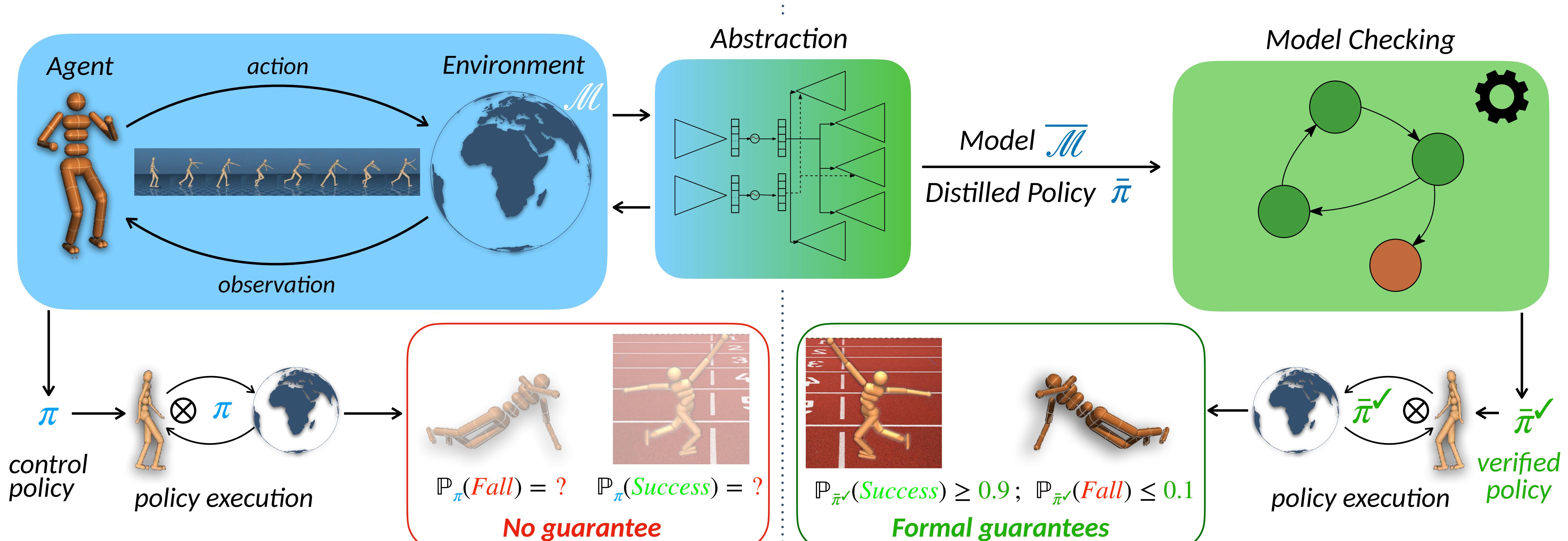


ARTIFICIAL  
INTELLIGENCE  
RESEARCH GROUP



# Overview

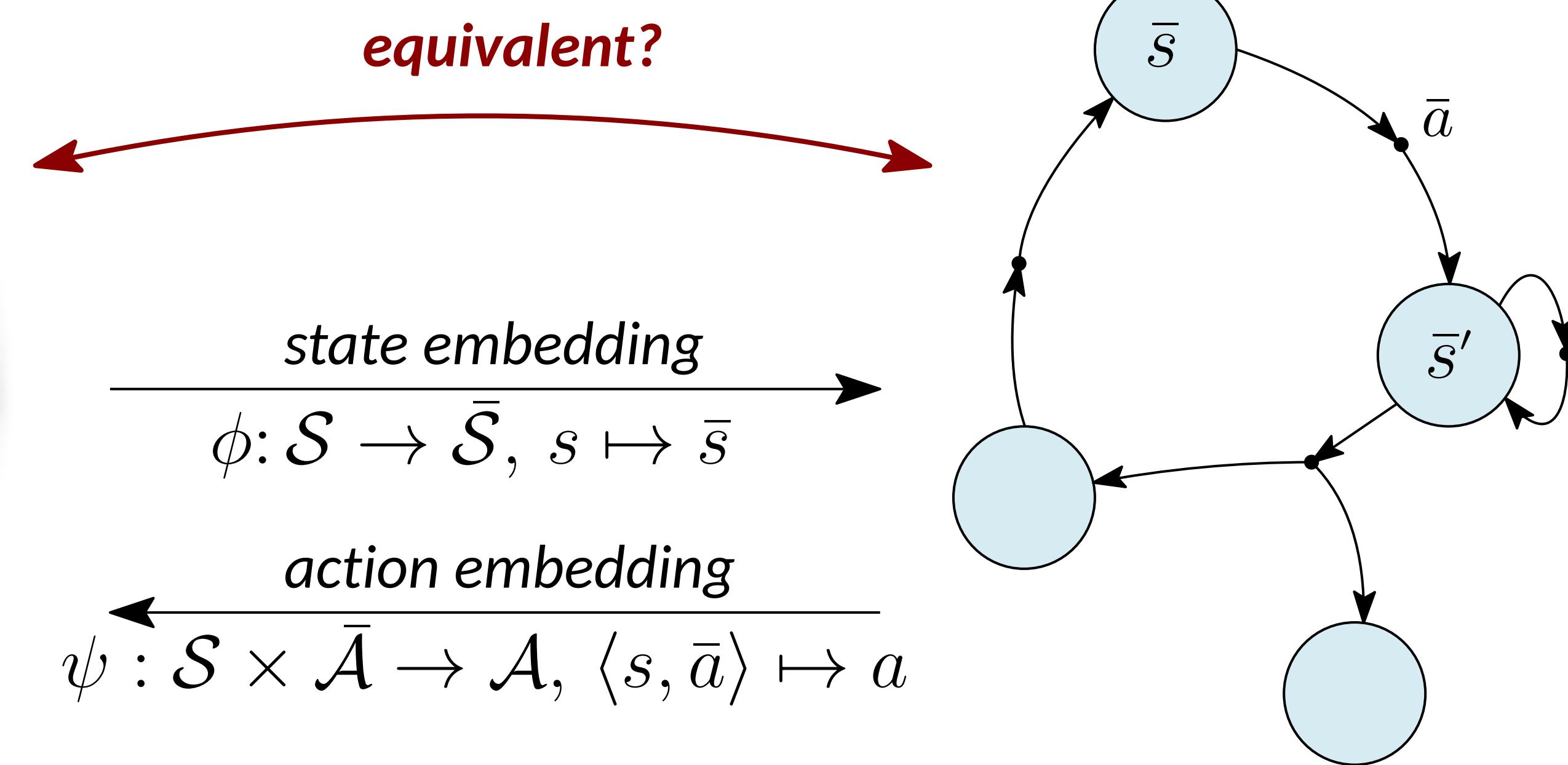
## Reinforcement Learning Policies with Formal Guarantees



- 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

# Abstraction Quality Criterion



Continuous-spaces MDP

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

Discrete latent MDP

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

- **Policies** prescribe which action to choose at each step:  $\pi: \mathcal{S} \rightarrow \mathcal{D}(\mathcal{A}), a_t \sim \pi(\cdot | s_t)$ 
  - A latent policy  $\bar{\pi}: \bar{\mathcal{S}} \rightarrow \mathcal{D}(\bar{\mathcal{A}})$  can be executed in  $\mathcal{M}$  via  $\bar{a} \sim \bar{\pi}(\cdot | \phi(s)), a = \psi(s, \bar{a})$

# Bisimulation

## Bisimulation Relation

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

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

**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  $\mathcal{L}$ , 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  $s_1, s_2 \in \mathcal{S}, a \in \mathcal{A}, T/B$

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

**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  $\mathcal{L}$ , 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  $\epsilon$  are  $\neq$

# Bisimulation distance

Continuous-spaces MDP

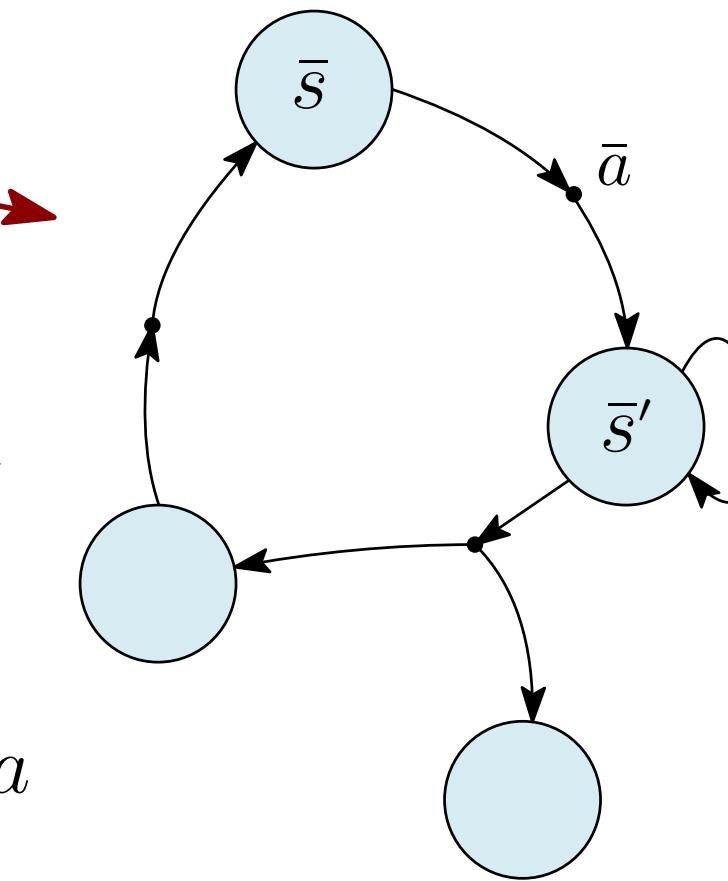


$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbf{P}, \ell \rangle$$

distance?

$$\begin{array}{c} \xrightarrow{\text{state embedding}} \\ \phi: \mathcal{S} \rightarrow \bar{\mathcal{S}}, s \mapsto \bar{s} \\ \xleftarrow{\text{action embedding}} \\ \psi: \mathcal{S} \times \bar{\mathcal{A}} \rightarrow \mathcal{A}, \langle s, \bar{a} \rangle \mapsto a \end{array}$$

Discrete latent MDP



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

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

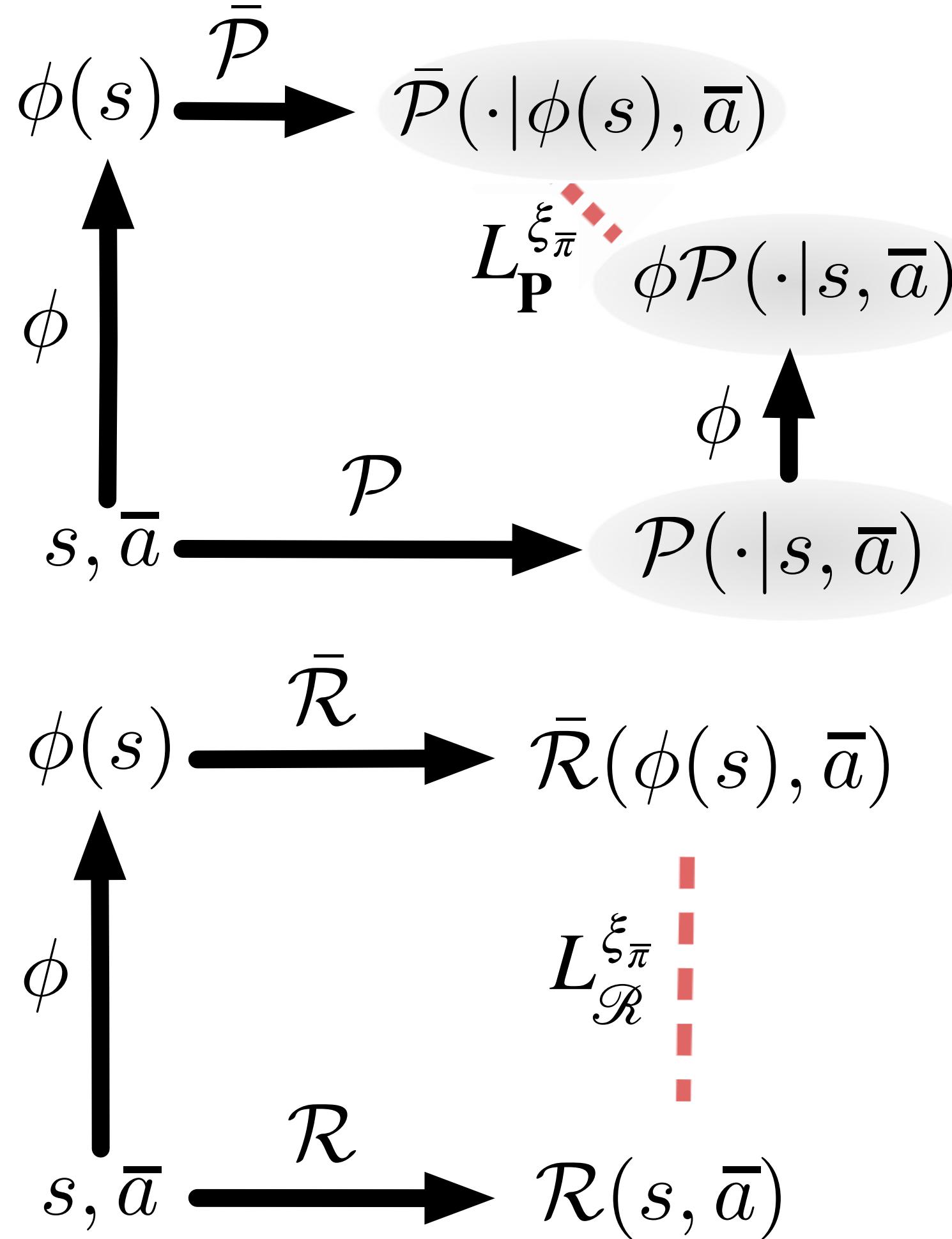
→ **Bisimulation distance:** largest behavioral difference (Desharnais et. al, 2004)

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

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

→ **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



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

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

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

- Expected bisimulation distance

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

$$\mathbb{E}_{s \sim \xi_{\bar{\pi}}} \tilde{d}_{\bar{\pi}}^{\ell}(s, \phi(s)) \leq \frac{\gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1 - \gamma}$$

- 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

$$T \geq \left\lceil \frac{-\log(\delta/4)}{2\epsilon^2} \right\rceil :$$

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

Gelada et. al, 2019

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

# Variational Markov Decision Process

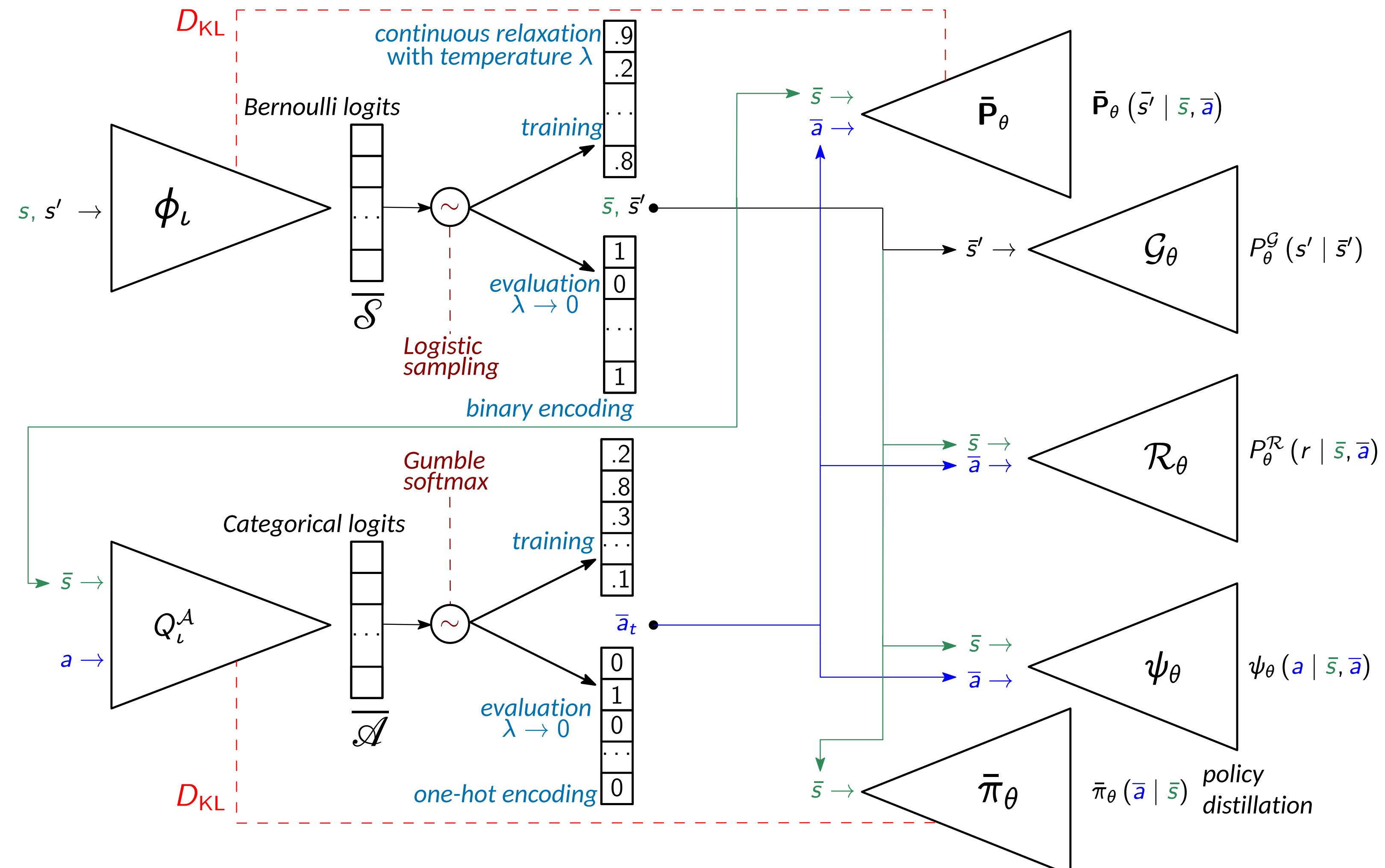
$$\max_{\iota, \theta} ELBO(\bar{\mathcal{M}}_\theta, \phi_\iota, \psi_\theta) = - \min_{\iota, \theta} \{ \mathbf{D}_{\iota, \theta} + \mathbf{R}_{\iota, \theta} \}$$

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

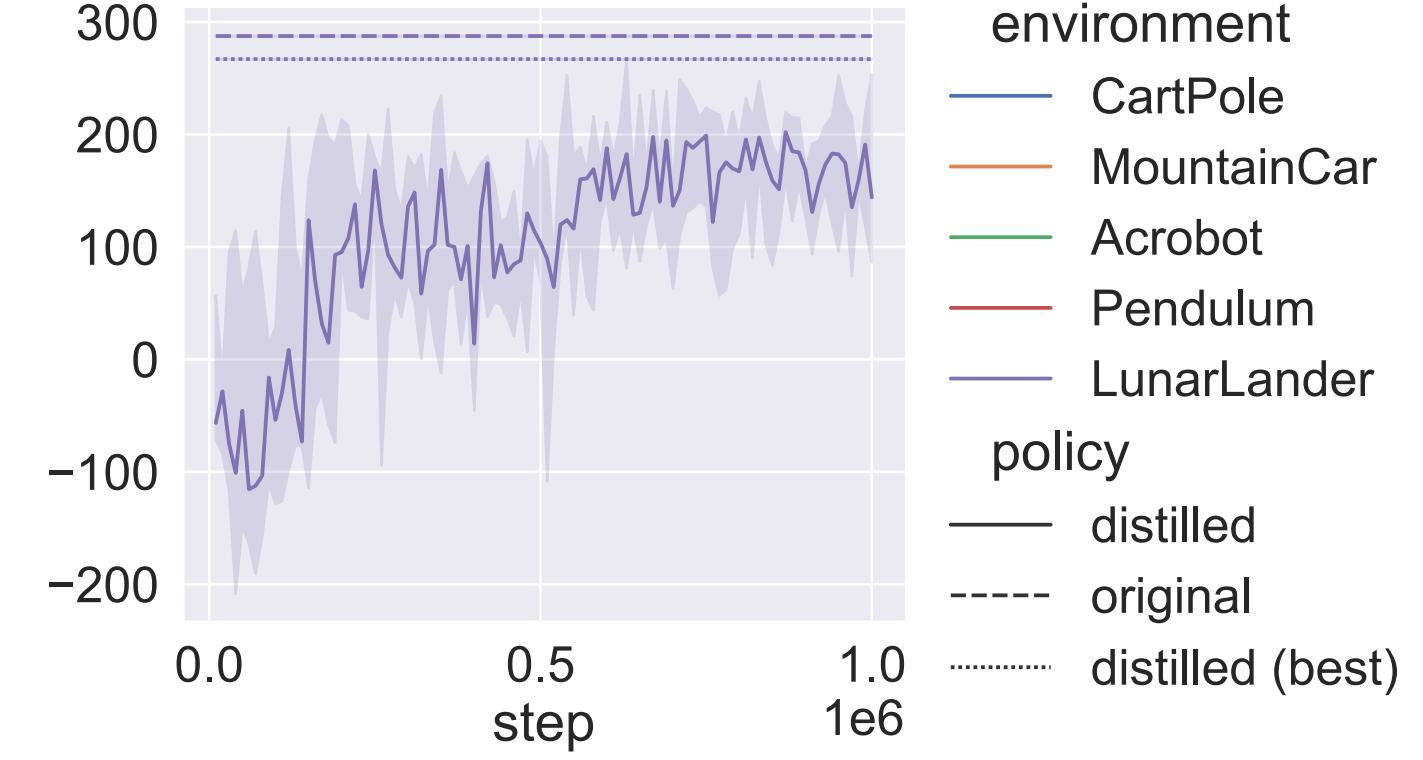
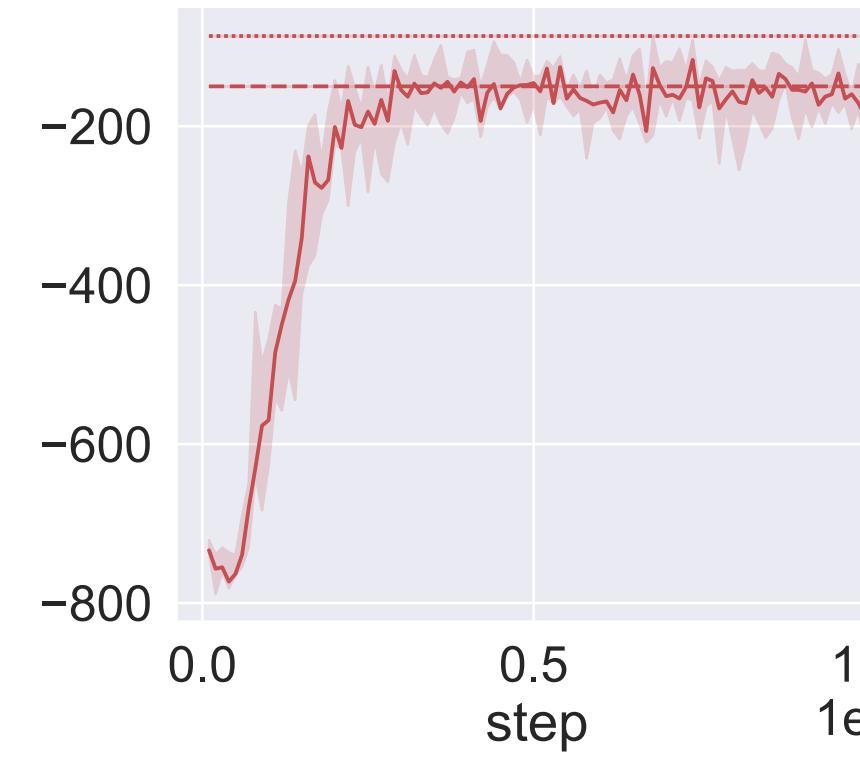
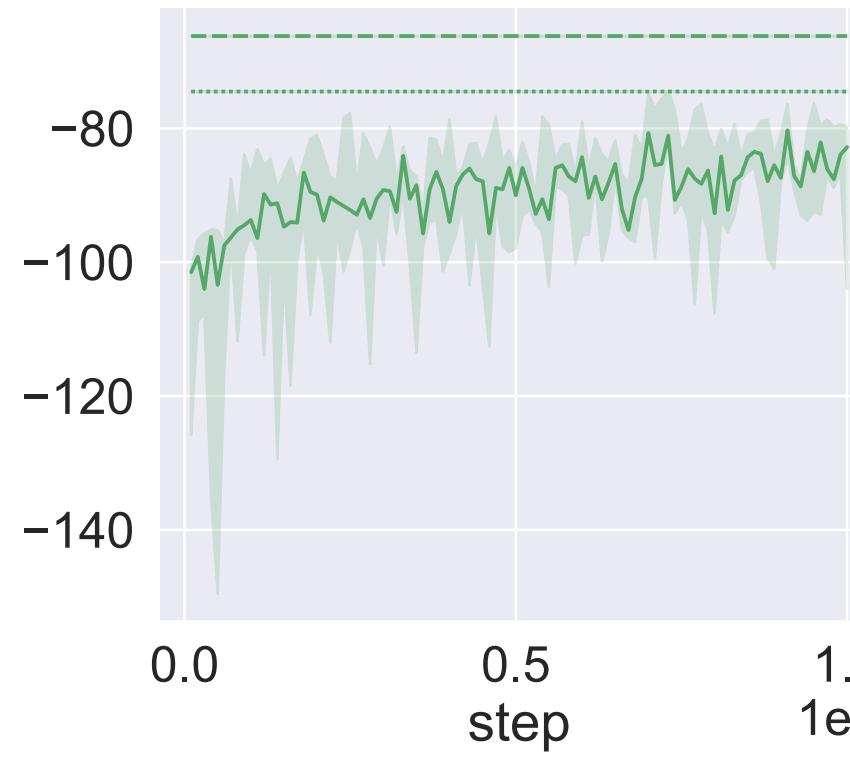
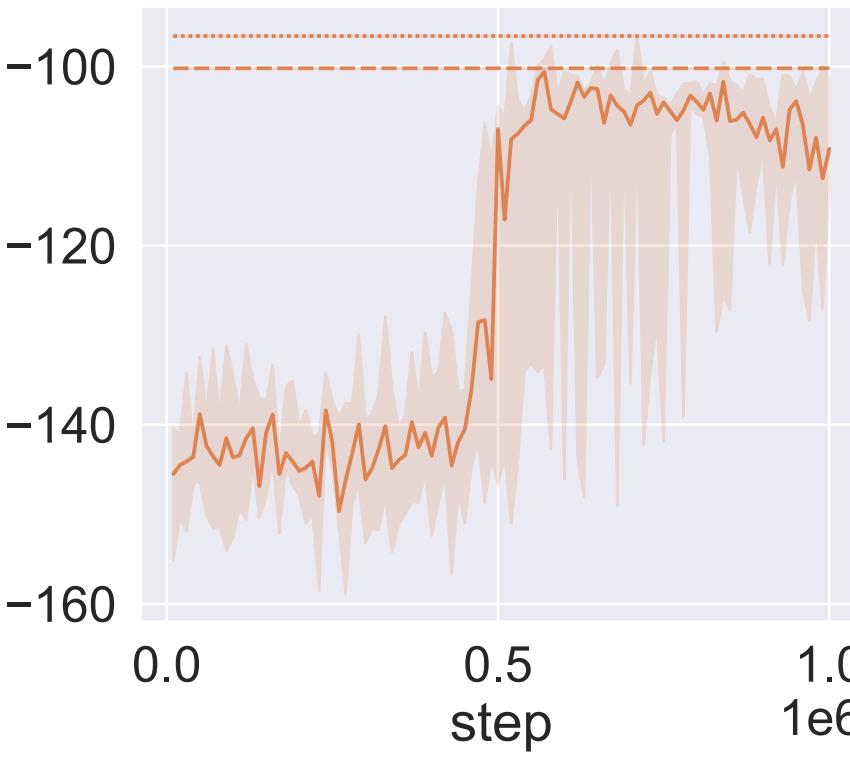
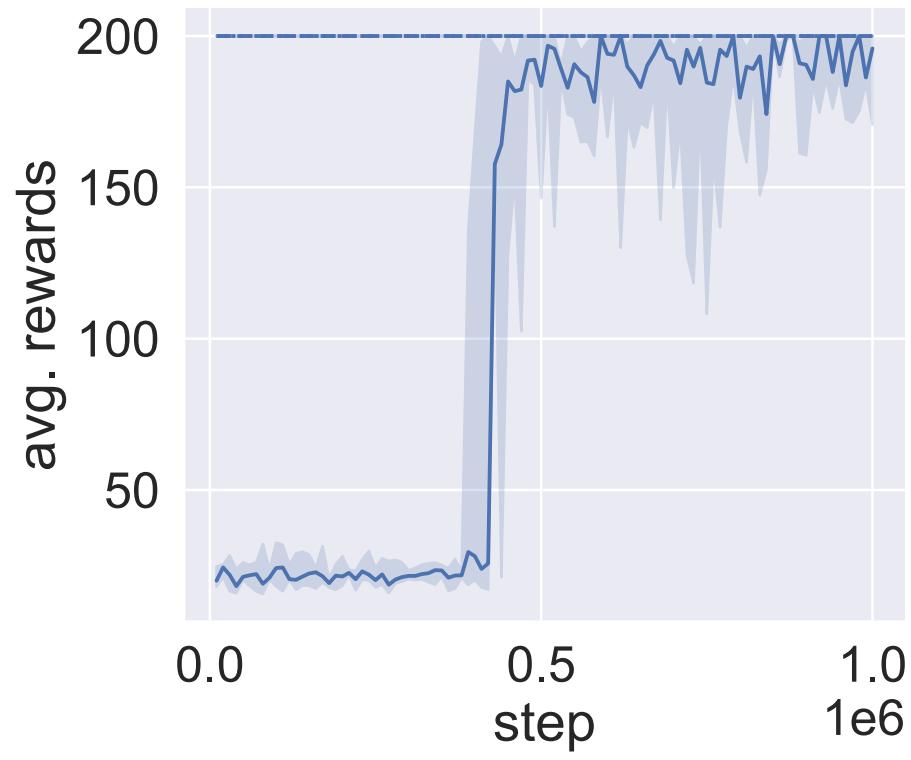
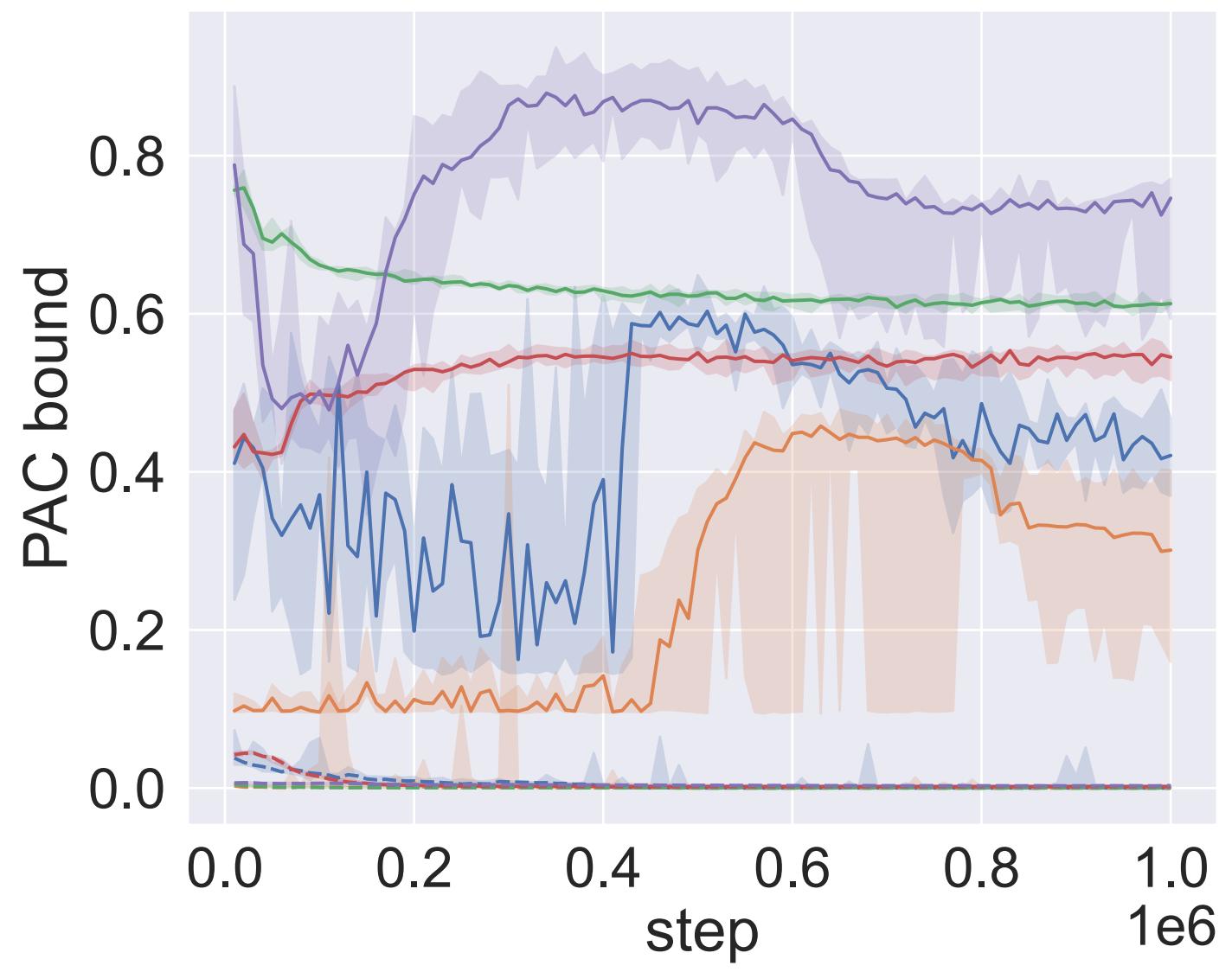
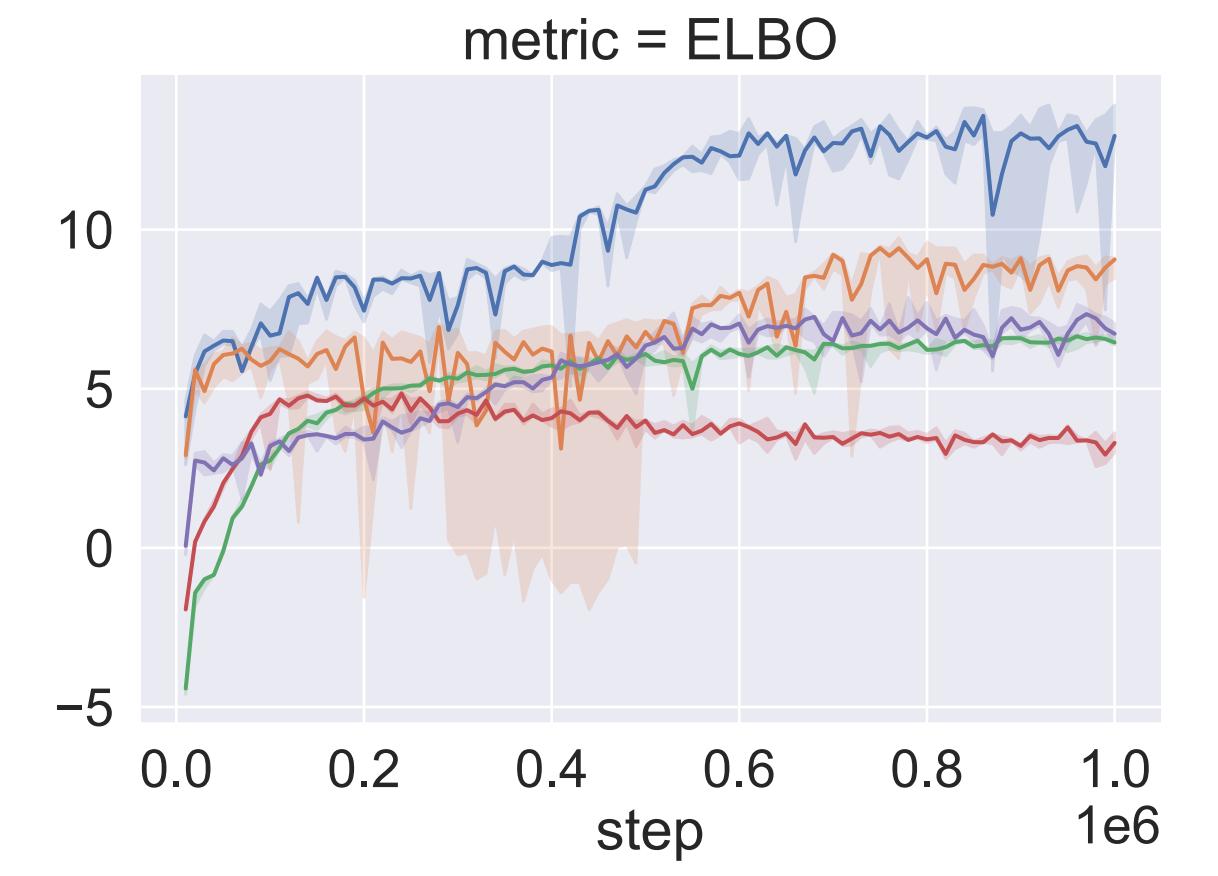
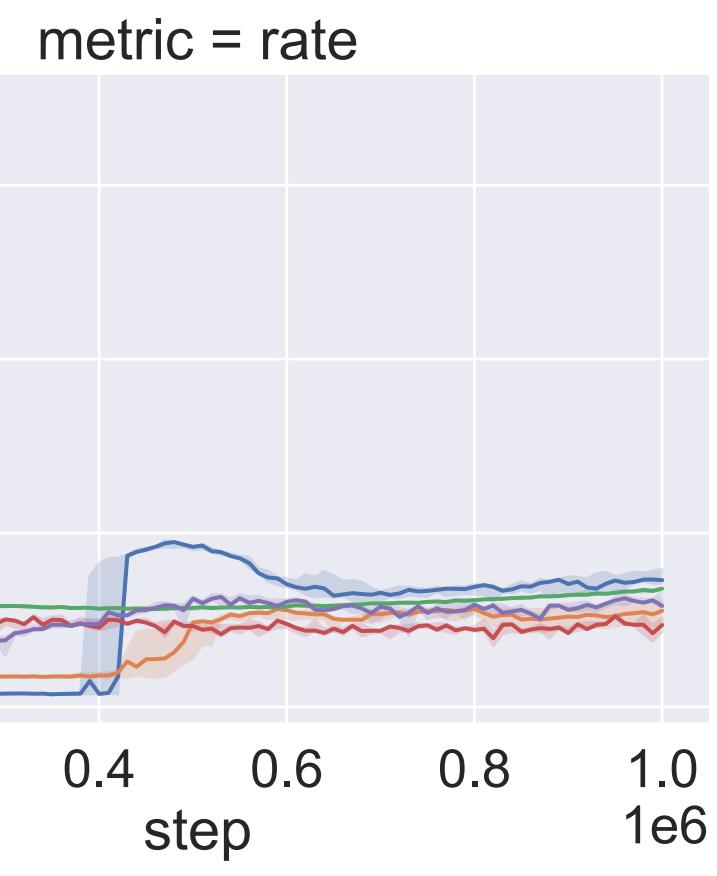
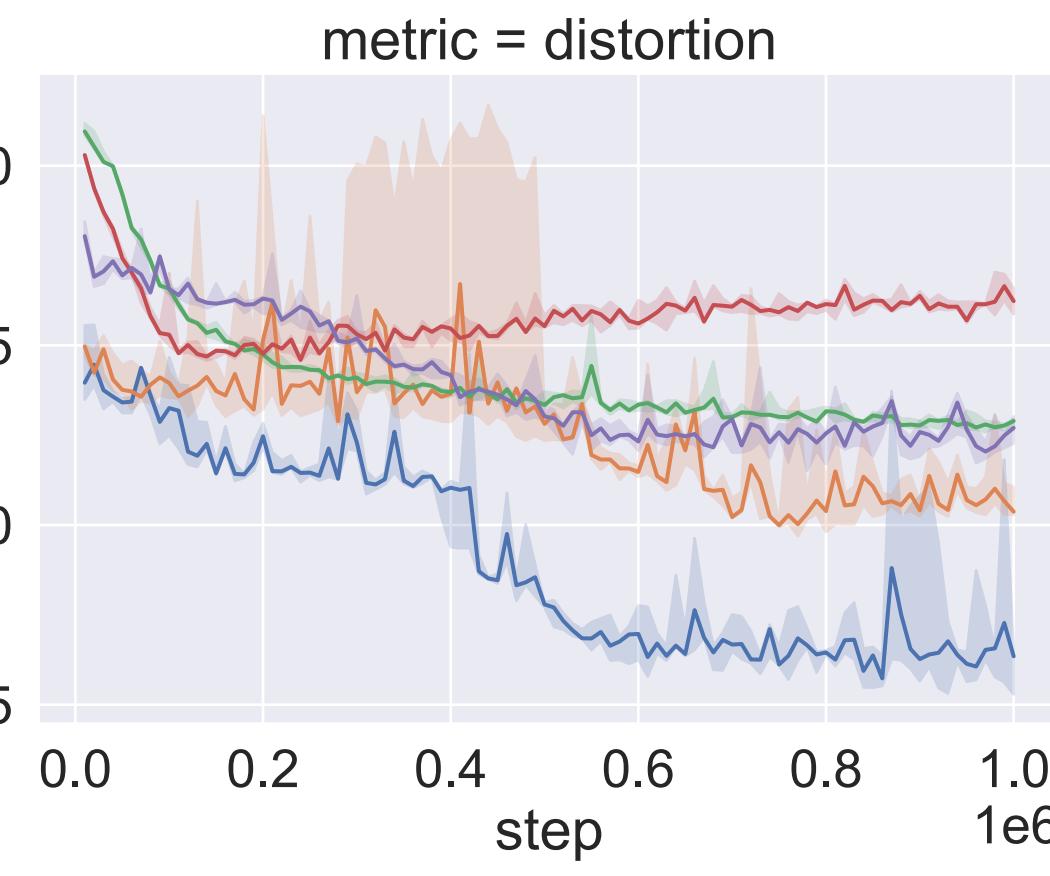
$$\mathbf{R}_{\iota, \theta} = \mathbb{E}_{\substack{s, a, s' \sim \xi_\pi \\ \bar{s} \sim \phi_\iota(\cdot | s) \\ \bar{a} \sim Q_\iota^A(\cdot | \bar{s}, a)}} [D_{KL}(\phi_\iota(\cdot | s') || \bar{\mathbf{P}}_\theta(\cdot | \bar{s}, \bar{a})) + D_{KL}(Q_\iota^A(\cdot | \bar{s}, a) || \bar{\pi}_\theta(\cdot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

*Discrete latent* model:

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



# Evaluation



environment

- CartPole
- MountainCar
- Acrobot
- Pendulum
- LunarLander

policy

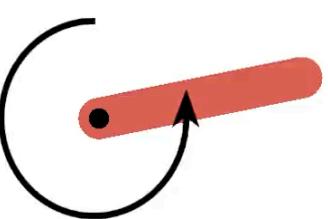
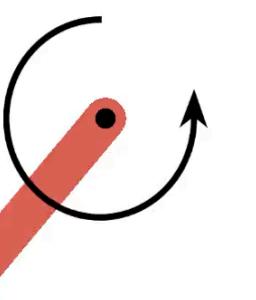
- distilled
- original
- distilled (best)

# Evaluation

 $\mathcal{M}_\pi$  $\mathcal{M}_{\bar{\pi}_\theta}$ 

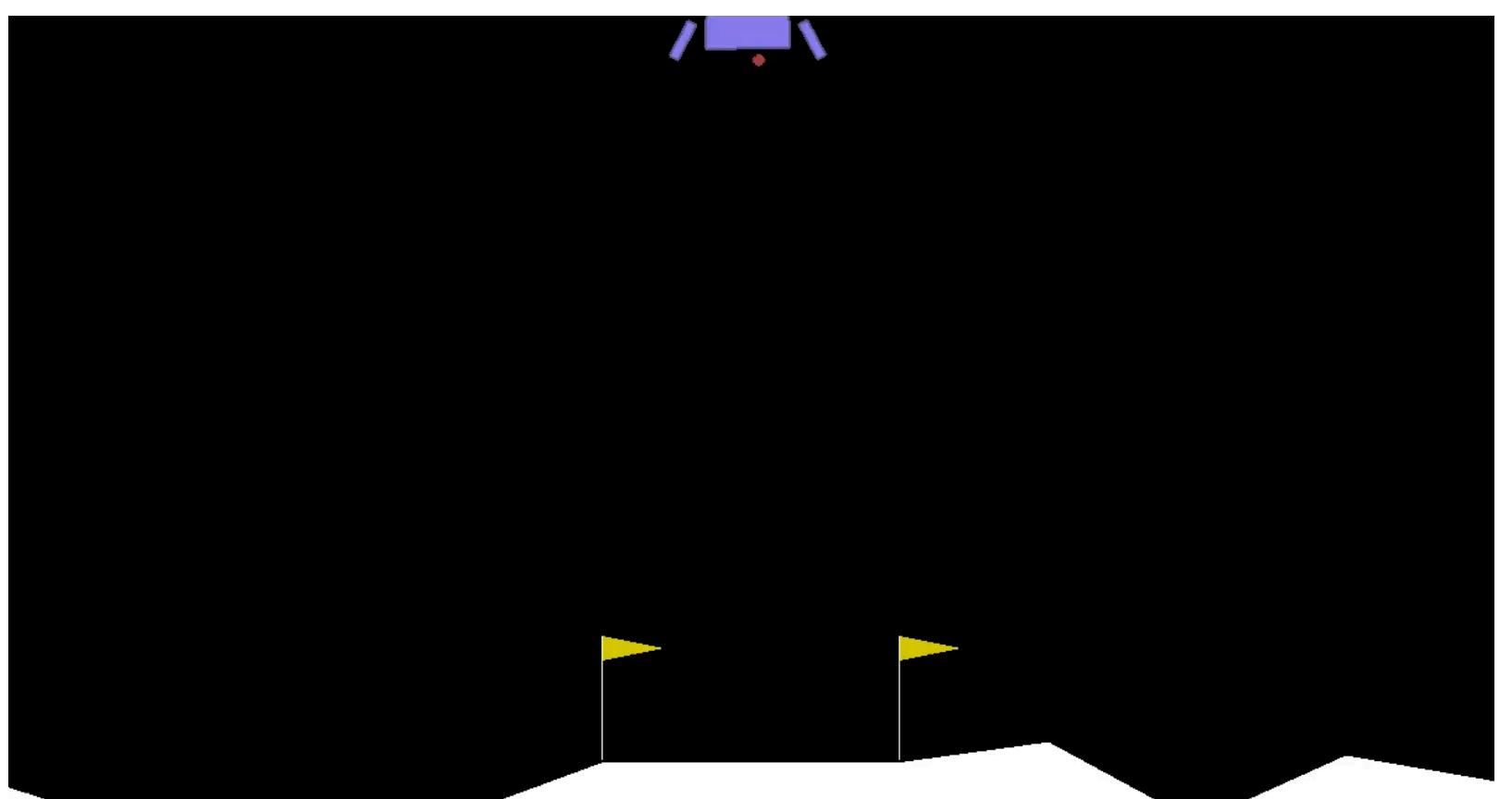
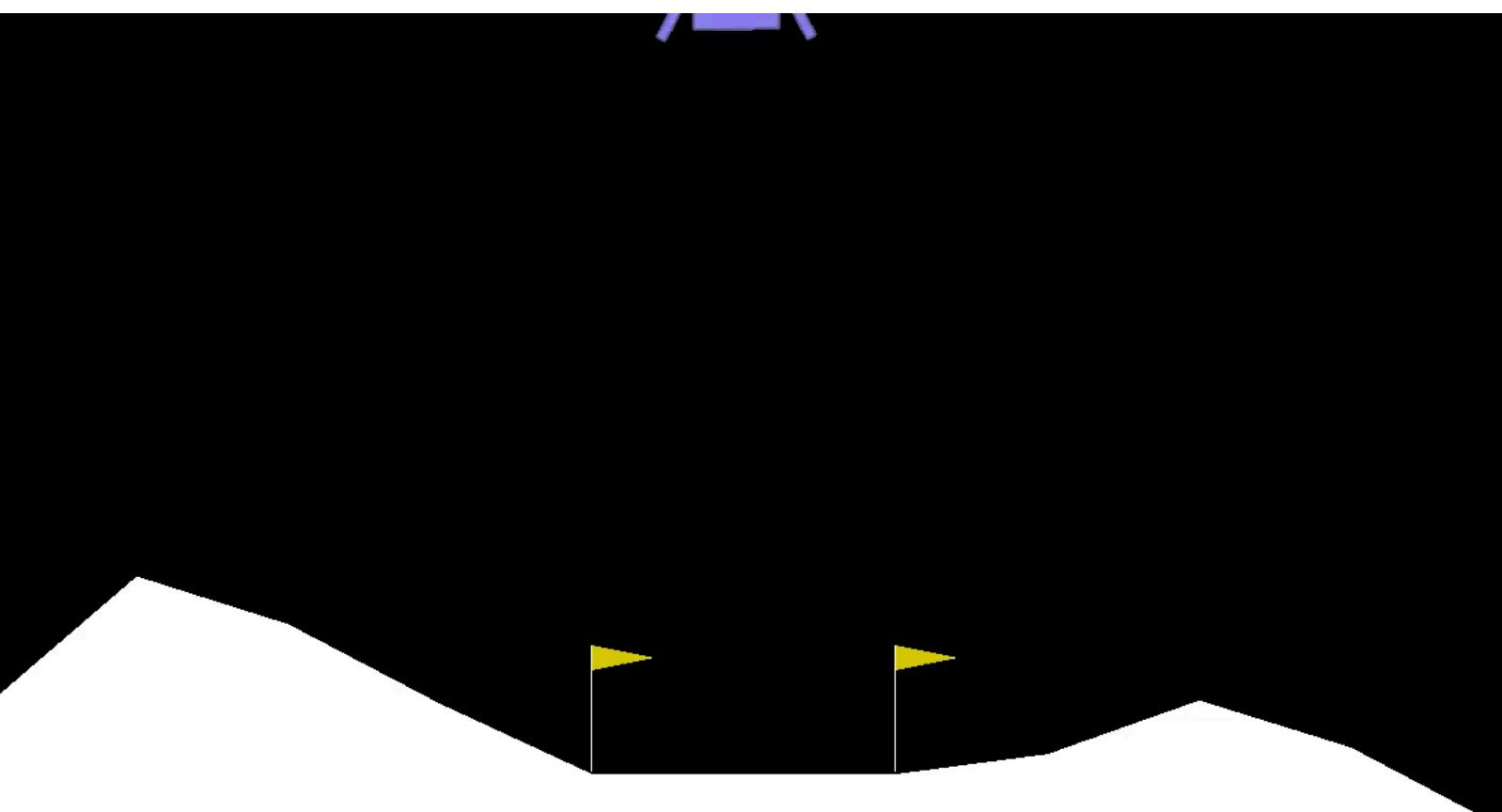
## Pendulum

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



## LunarLander

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



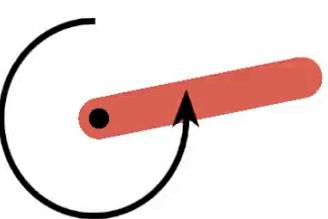
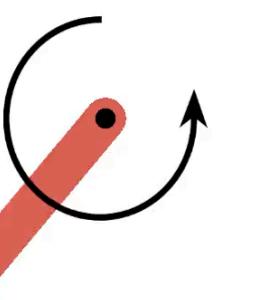
# Evaluation

$$\mathcal{M}_\pi$$

$$\mathcal{M}_{\bar{\pi}_\theta}$$

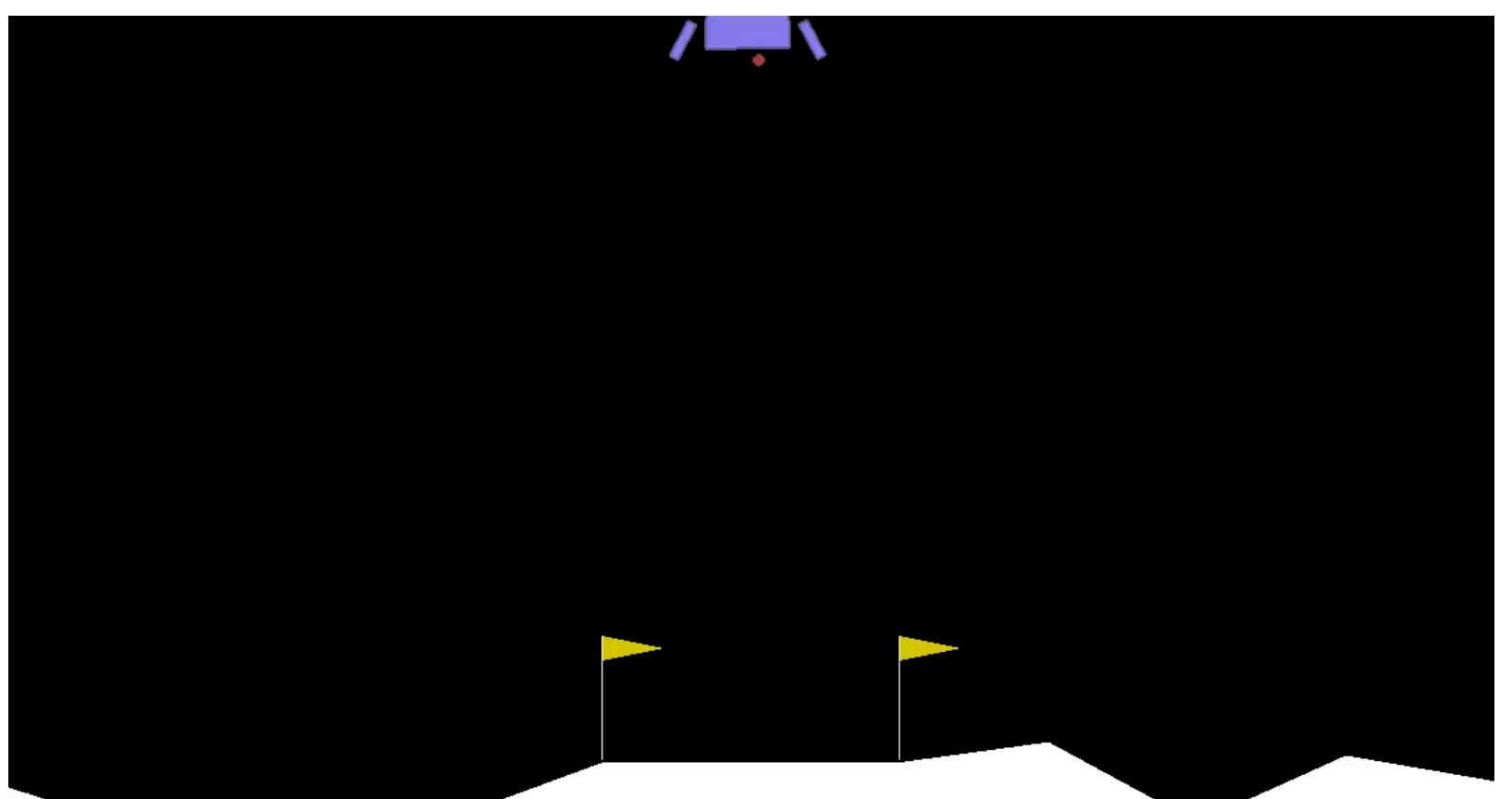
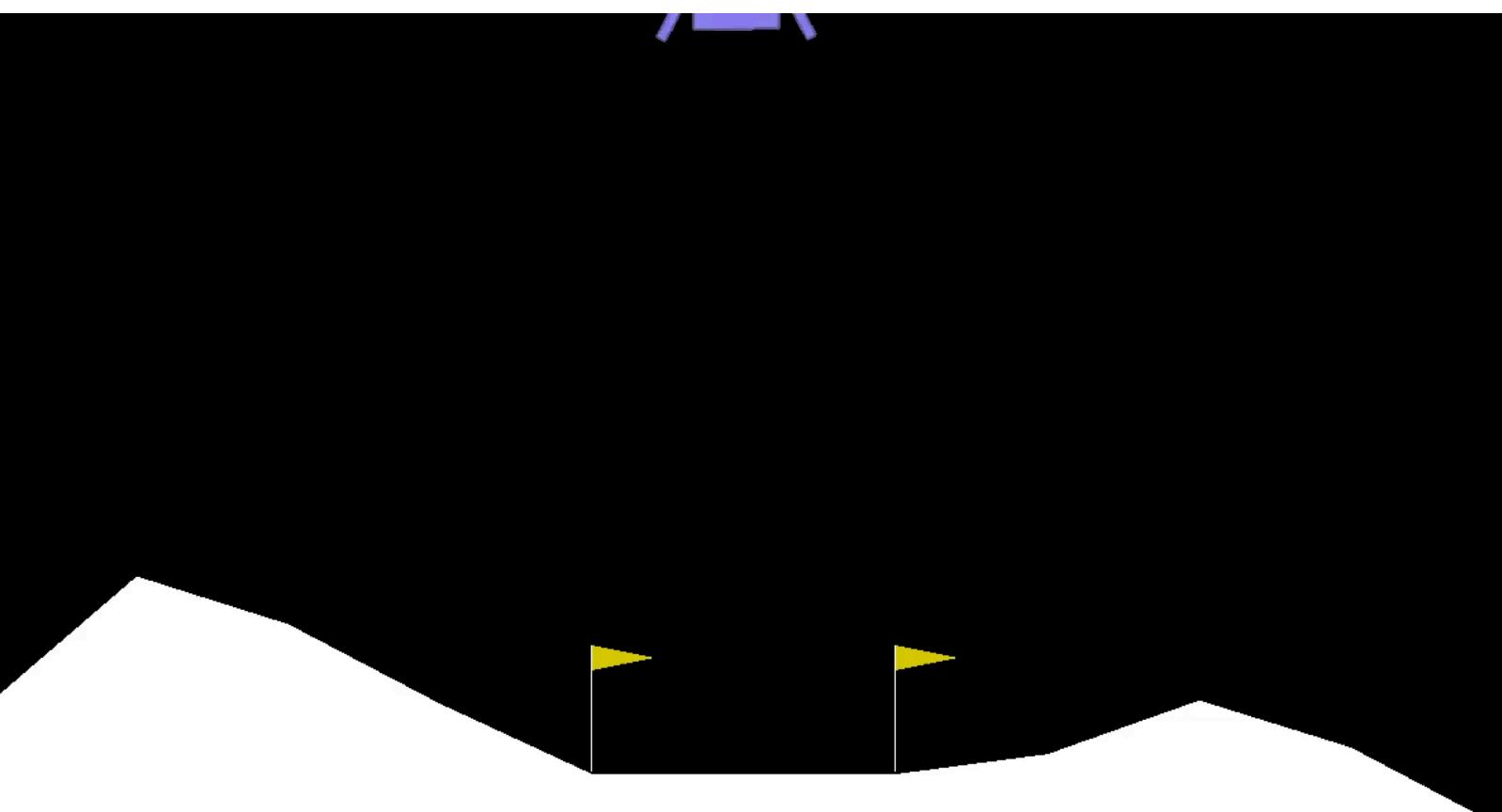
## Pendulum

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



## LunarLander

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



# 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