Wasserstein Auto-encoded MDPs

Formal Verification of Efficiently Distilled RL Policies with Many-sided Guarantees

Florent Delgrange, Ann Nowé, Guillermo A. Pérez

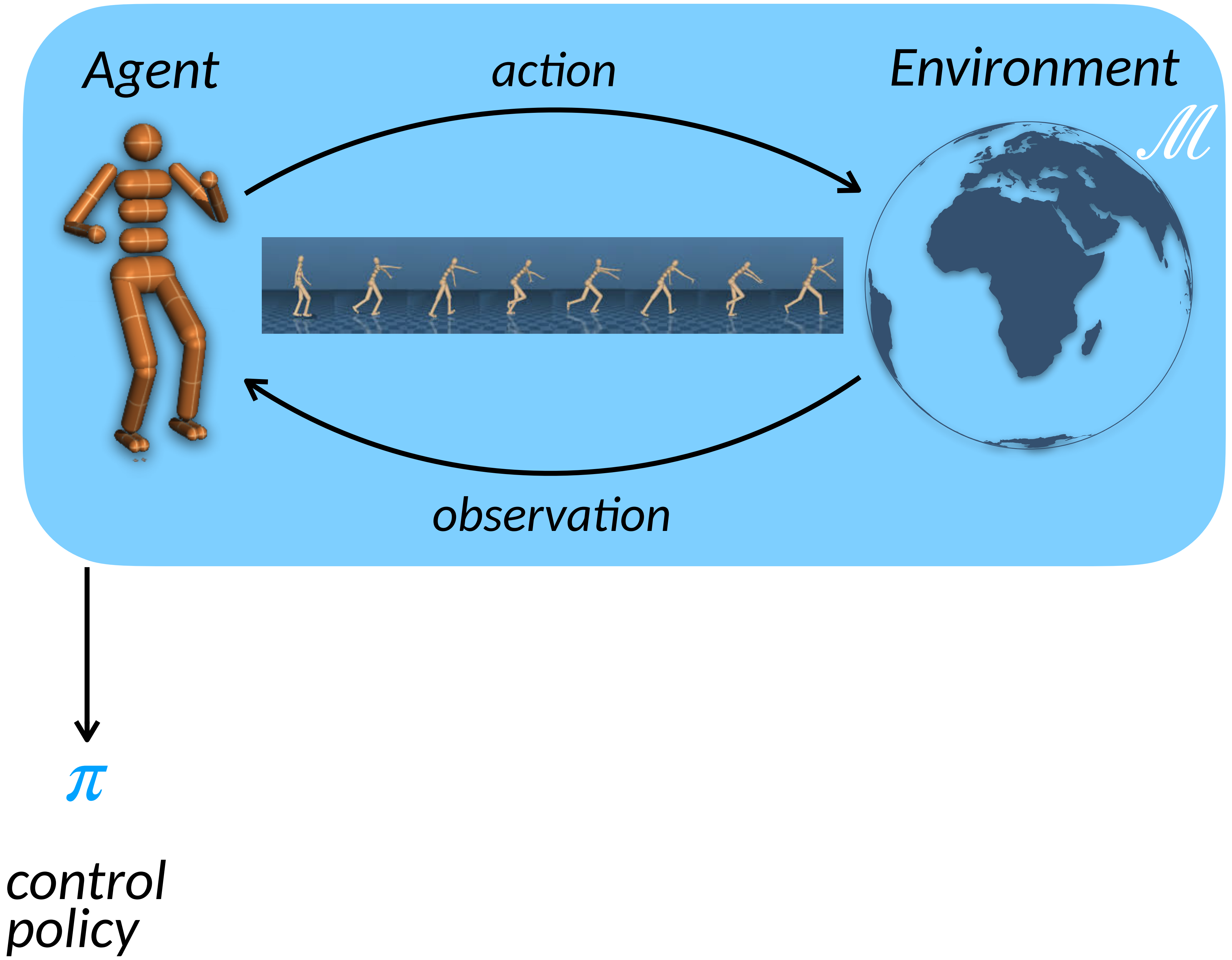
ICLR

VUB ARTIFICIAL INTELLIGENCE RESEARCH GROUP

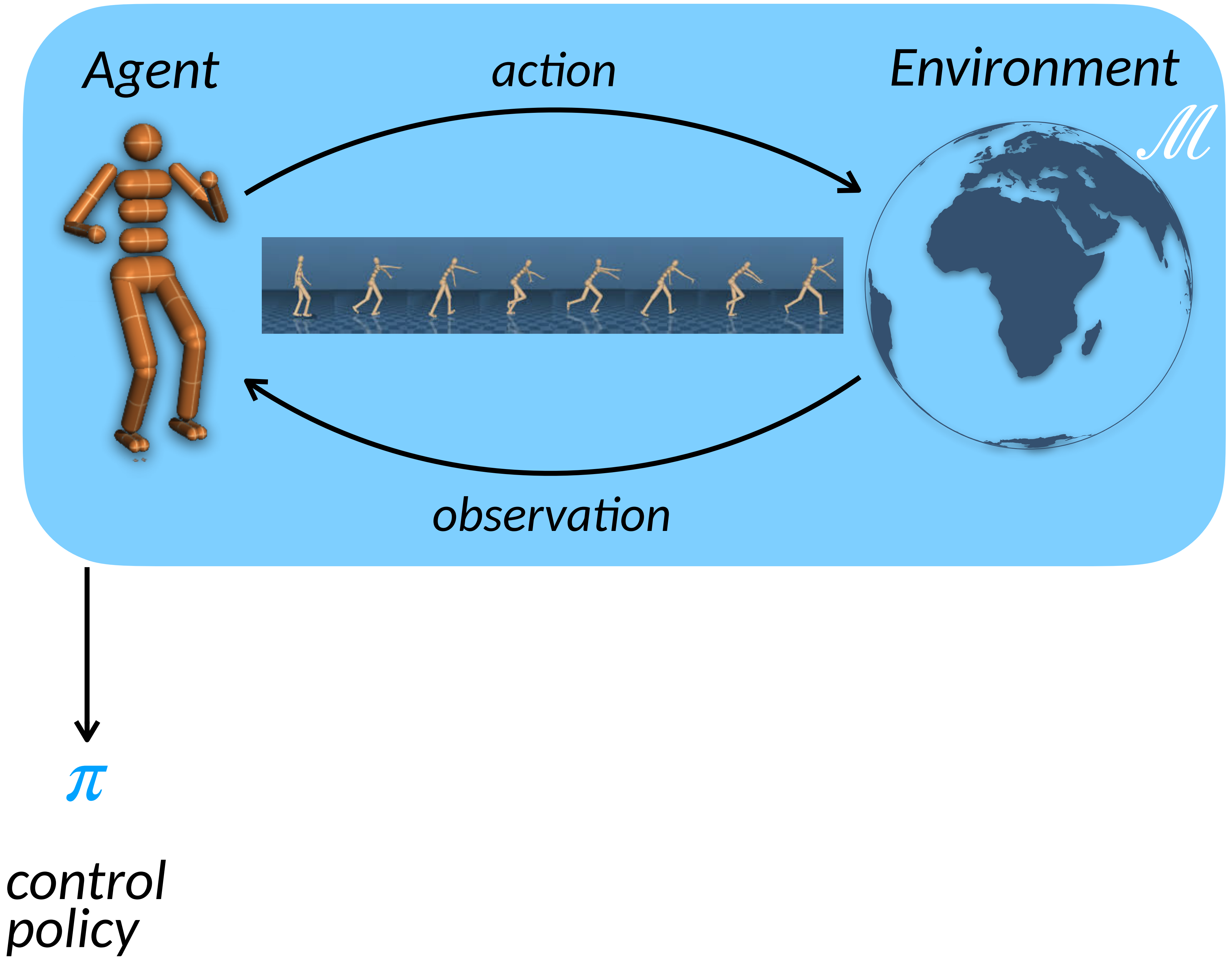
Universiteit Antwerpen

Overview

Reinforcement Learning

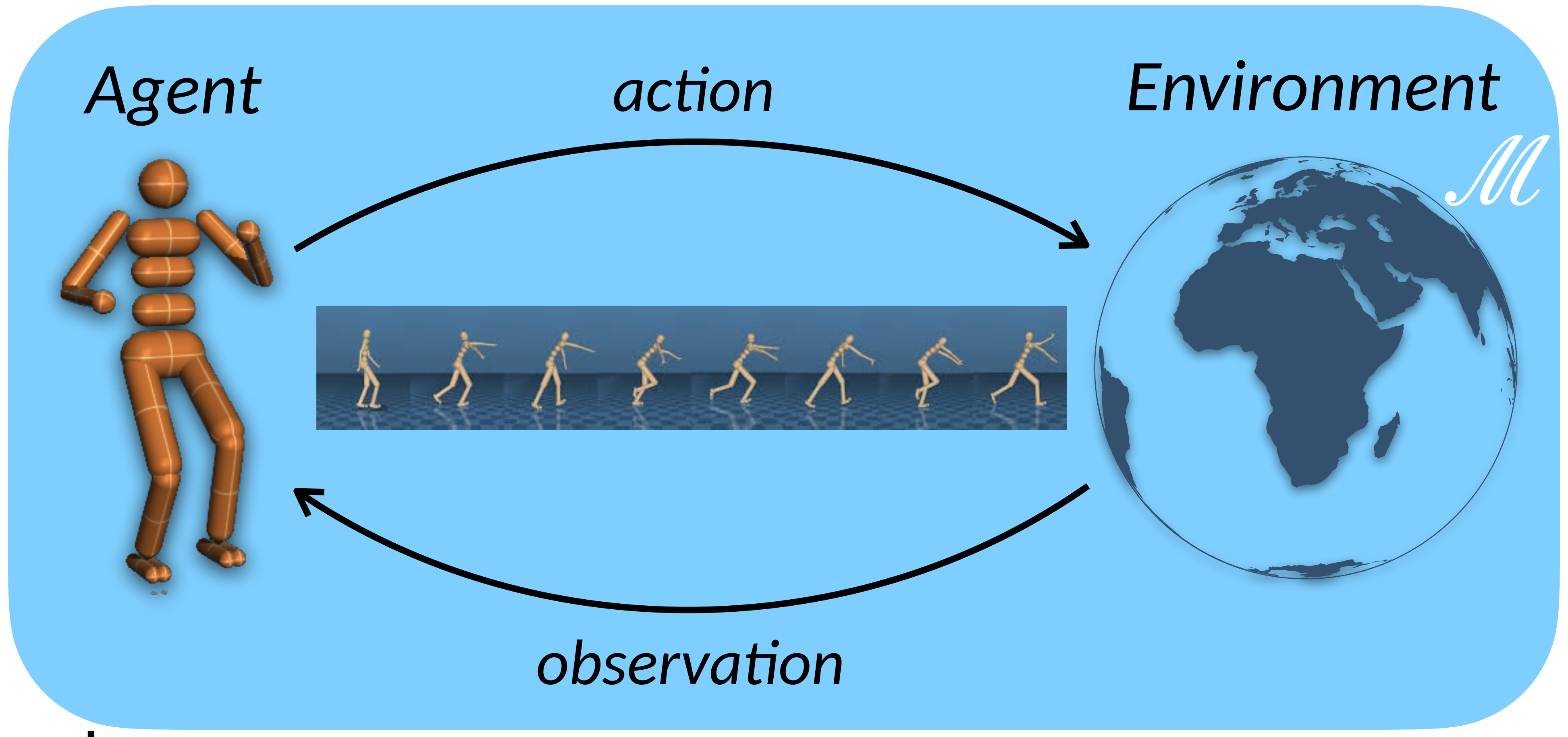


Reinforcement Learning



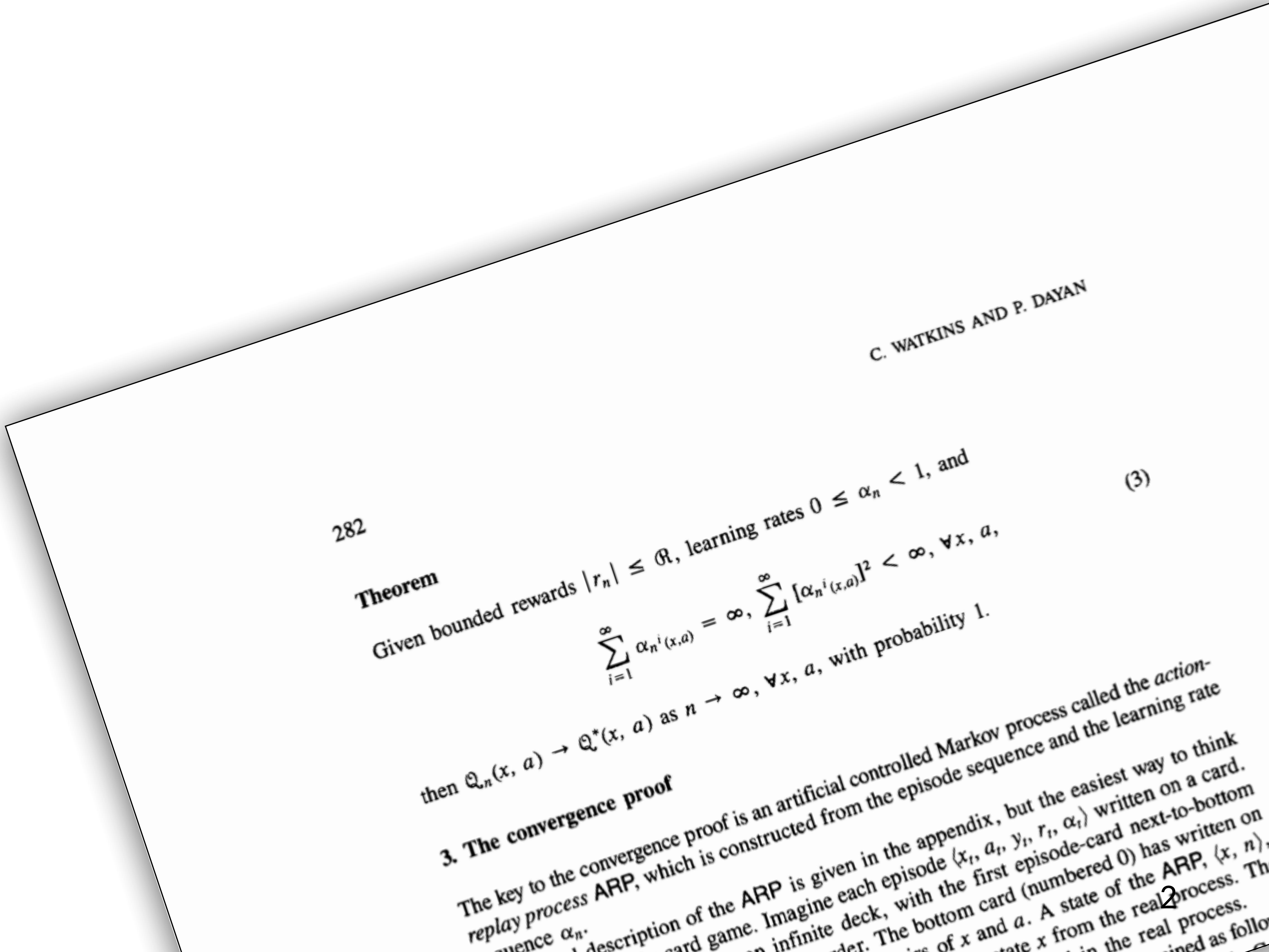
- Unknown environment
- Continuous state/action spaces

Reinforcement Learning



π

control policy


282 C. WATKINS AND P. DAYAN

Theorem

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

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

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

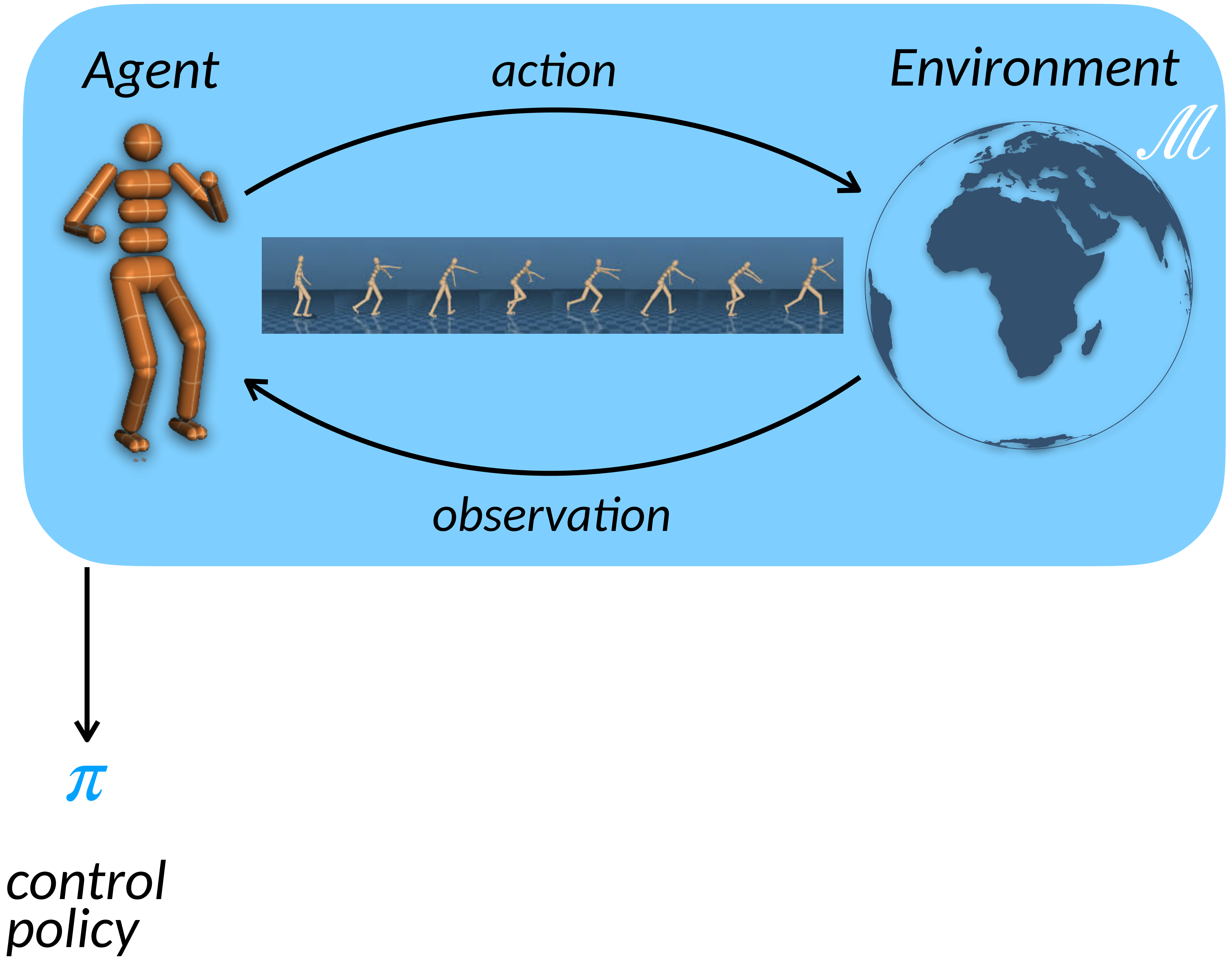
3. The convergence proof

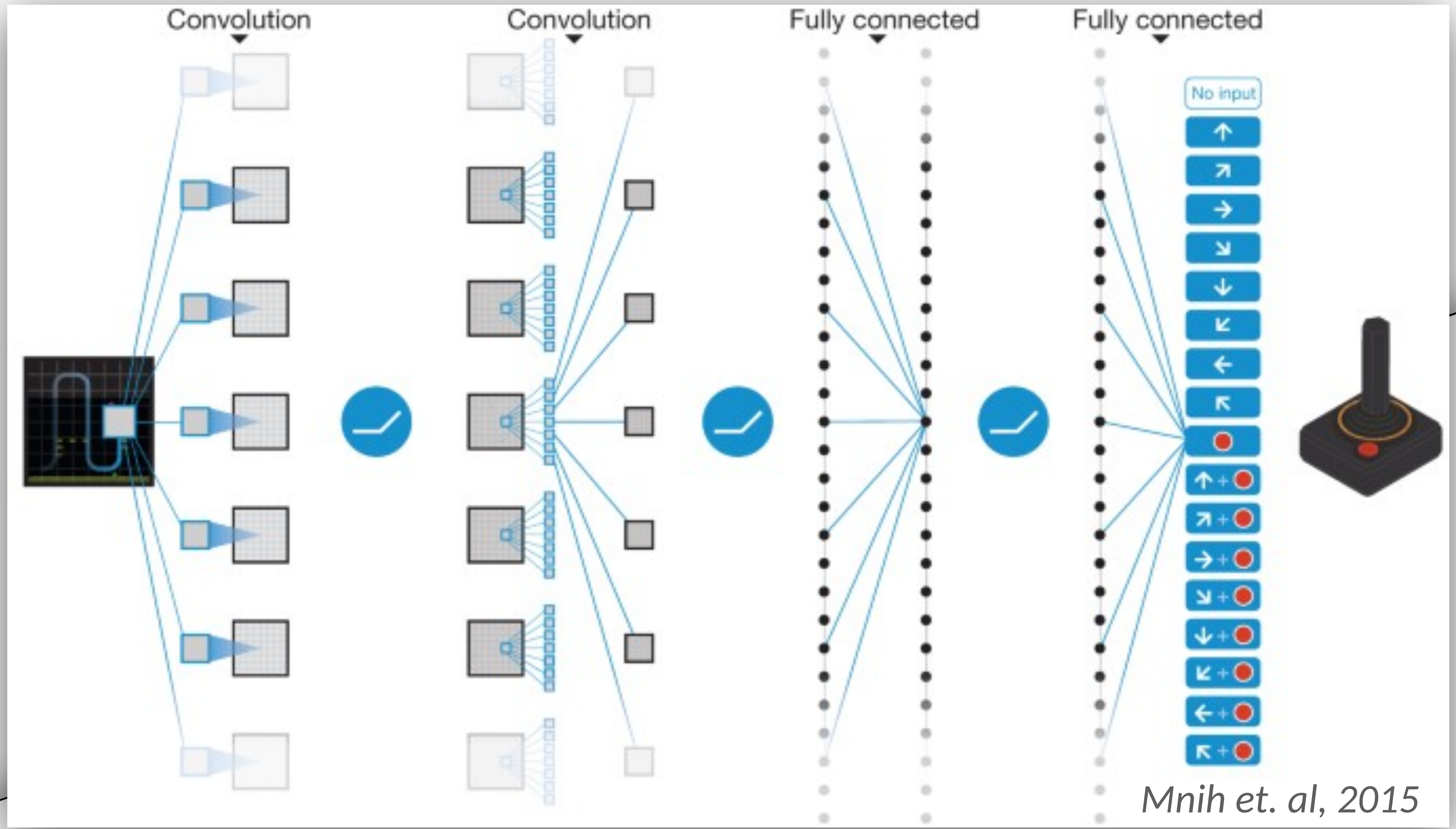
The key to the convergence proof is an artificial controlled Markov process called the *action-replay process* ARP, which is constructed from the episode sequence and the learning rate sequence α_n.

- Unknown environment
- Continuous state/action spaces

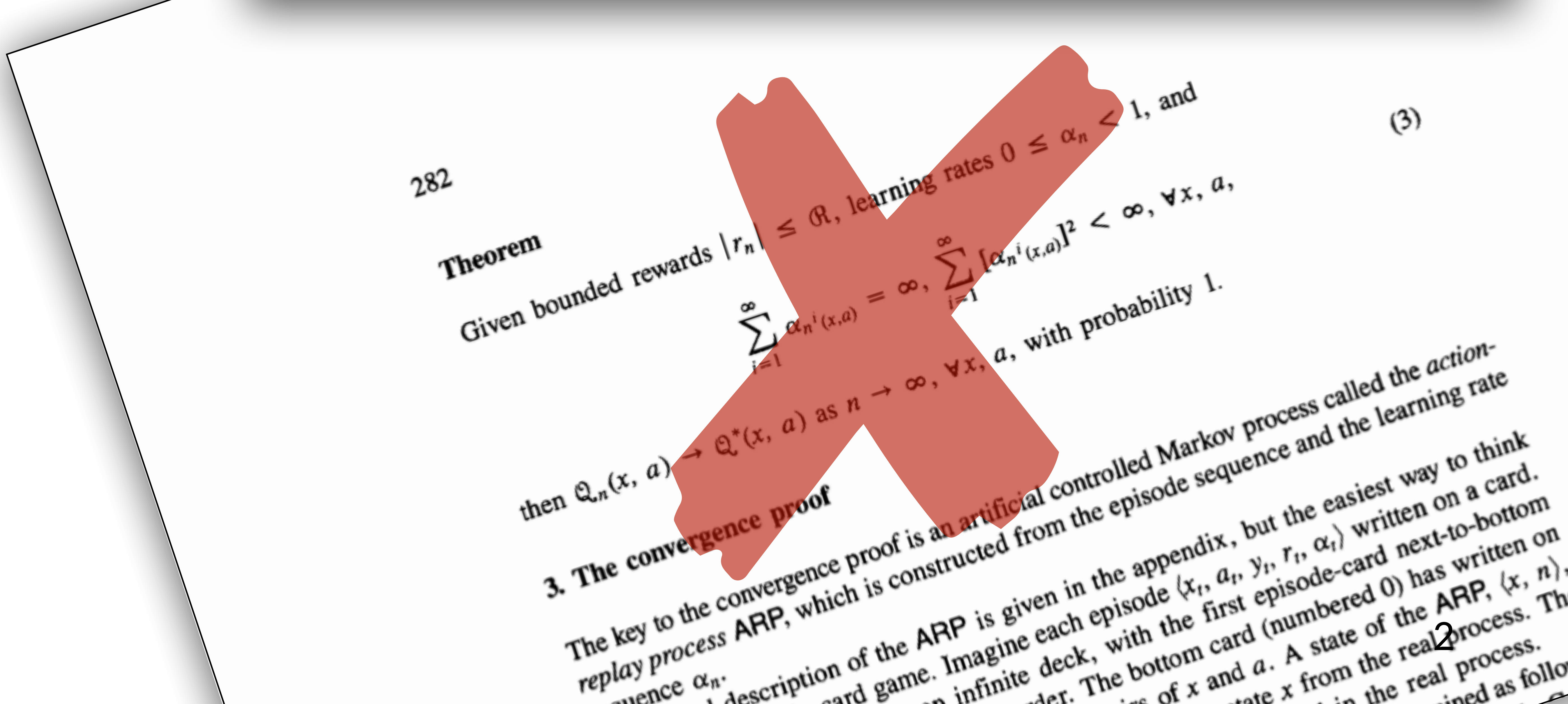
Overview

Reinforcement Learning





Mnih et. al, 2015


282

Theorem

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

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

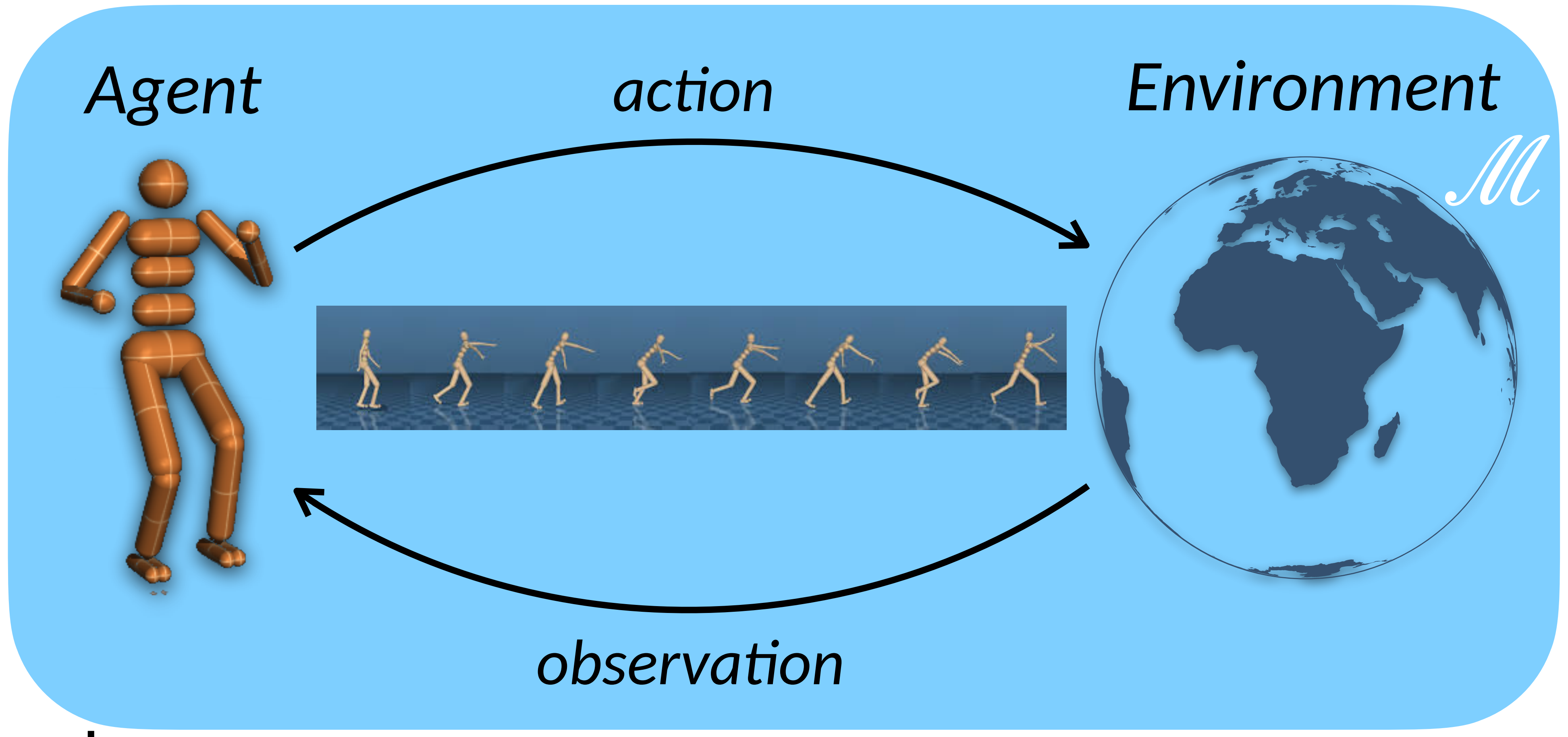
then $\mathcal{Q}_n(x, a) \to \mathcal{Q}^*(x, a)$ as $n \to \infty$, $\forall x, a$, with probability 1.

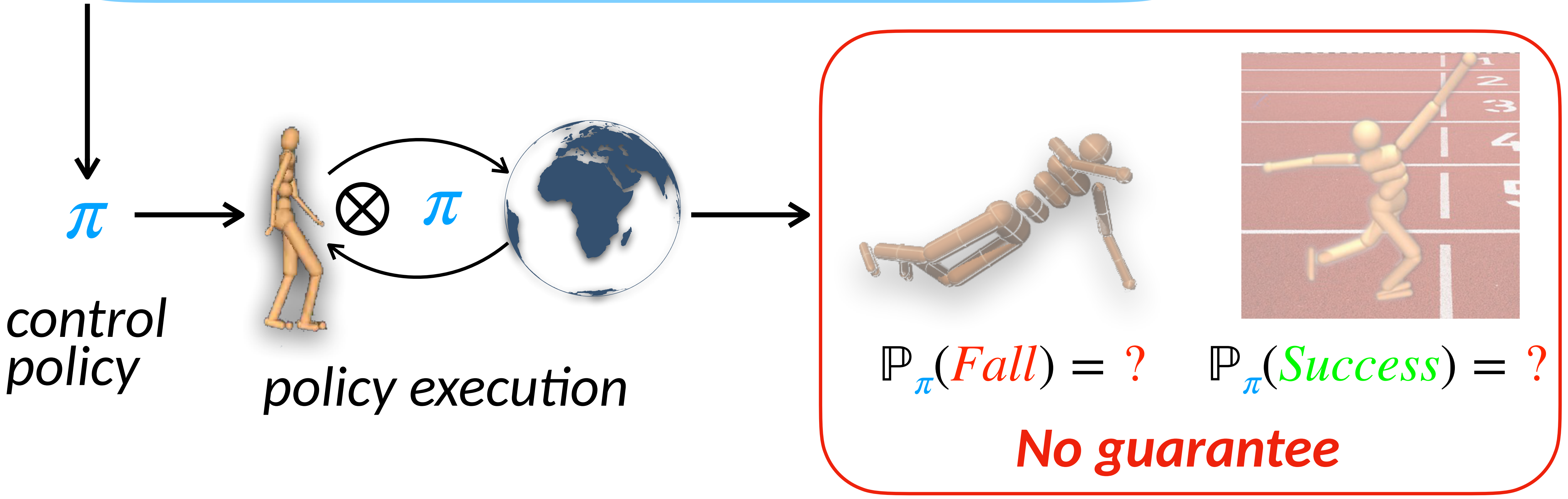
3. The convergence proof

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.

- Unknown environment
- Continuous state/action spaces

Reinforcement Learning

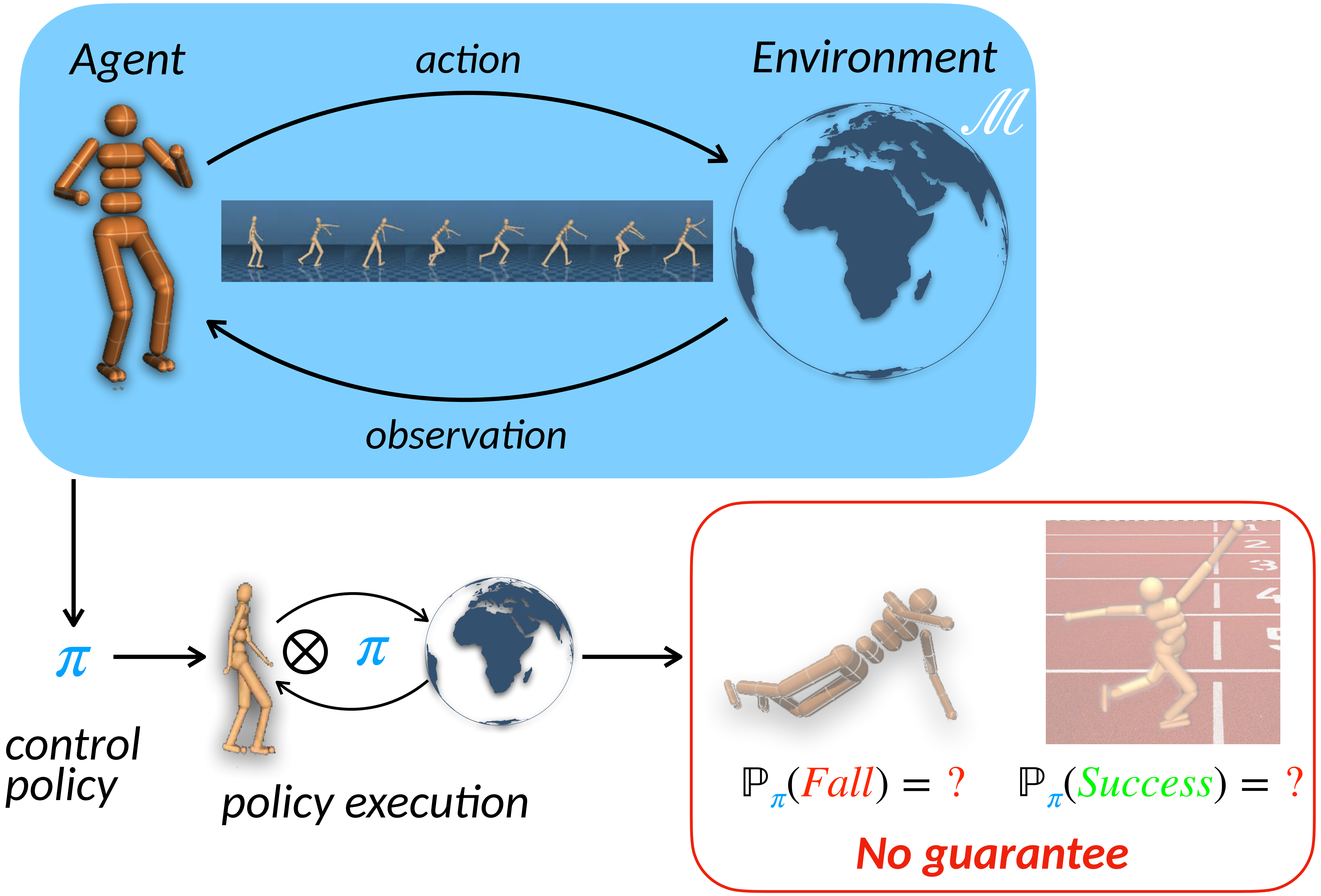




- Unknown environment
- Continuous state/action spaces

Overview

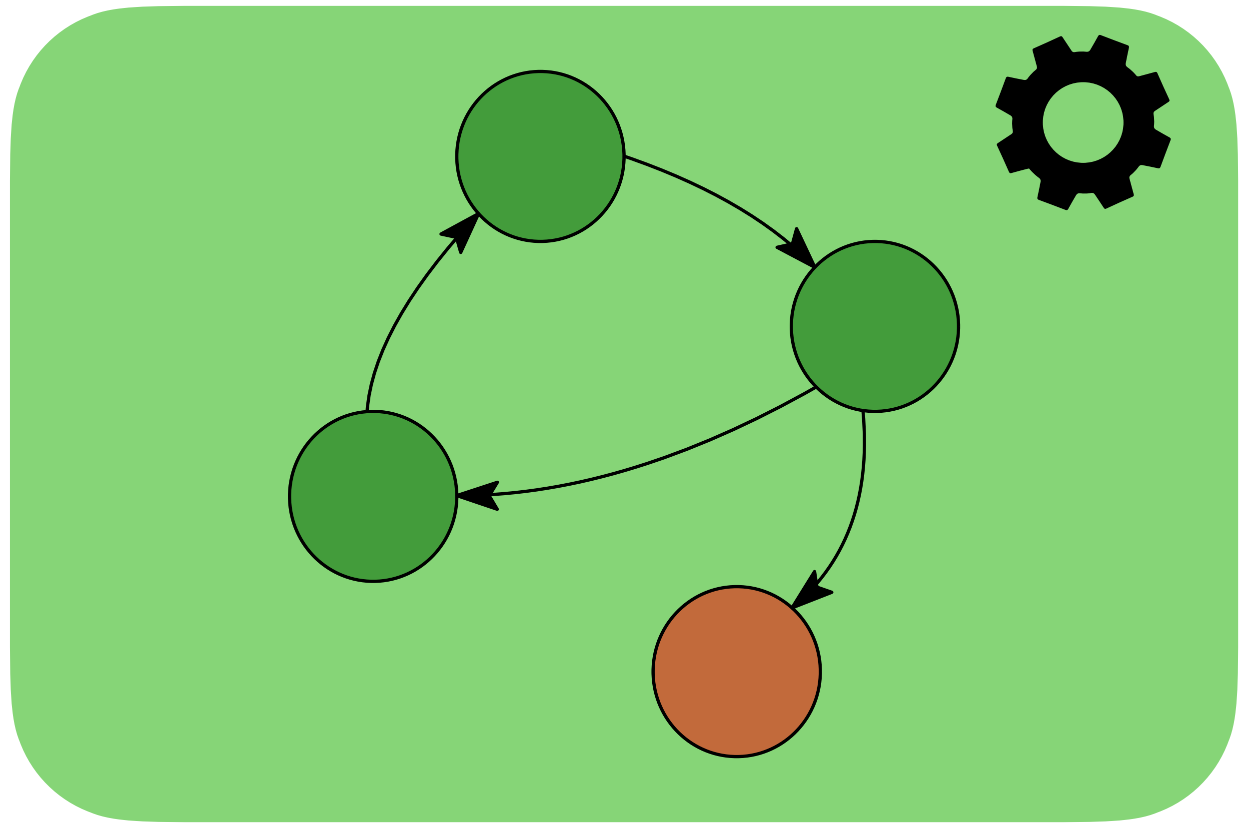
Reinforcement Learning



- Unknown environment
- Continuous state/action spaces

Formal Guarantees

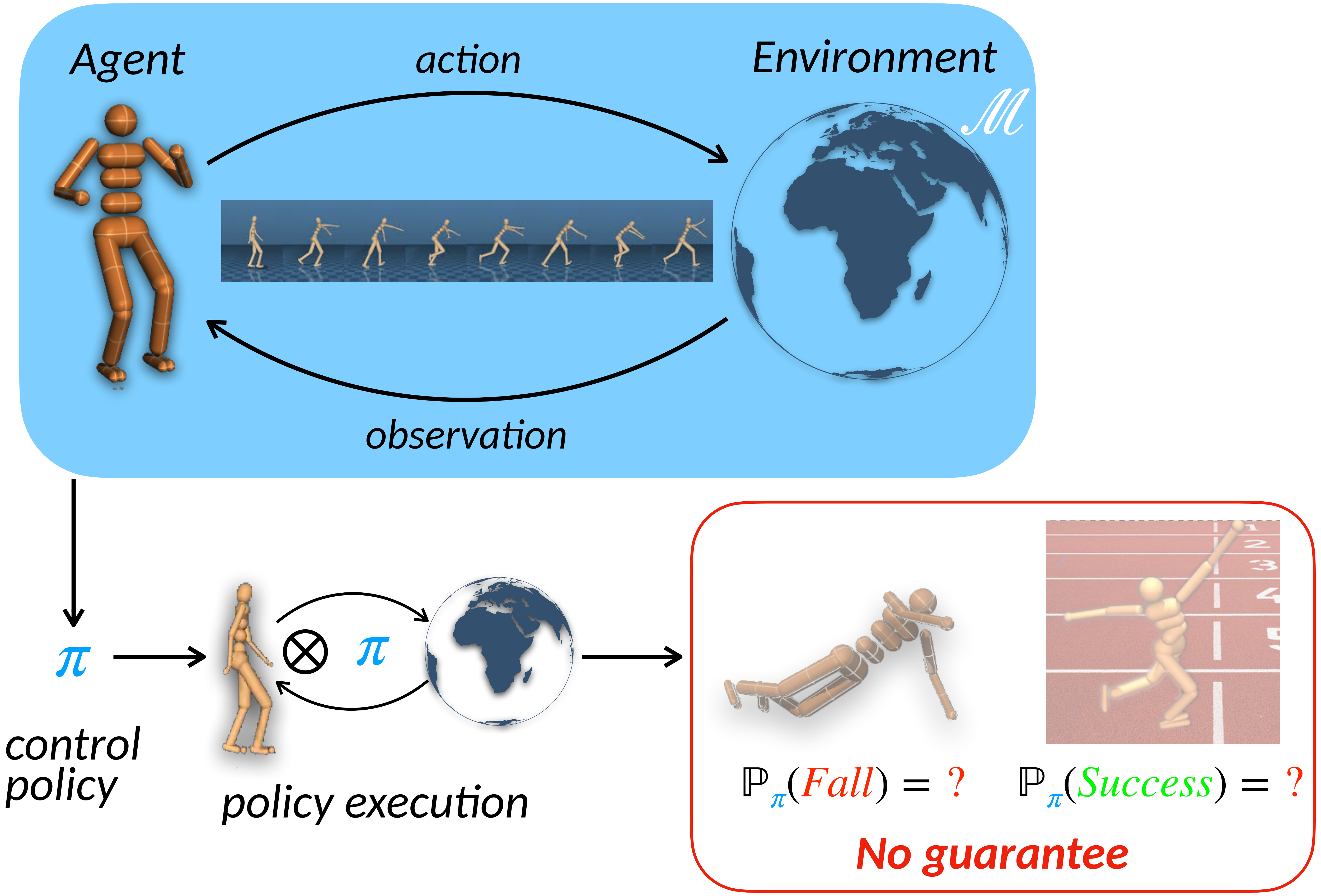
Model Checking



- Full knowledge of the model of the interaction

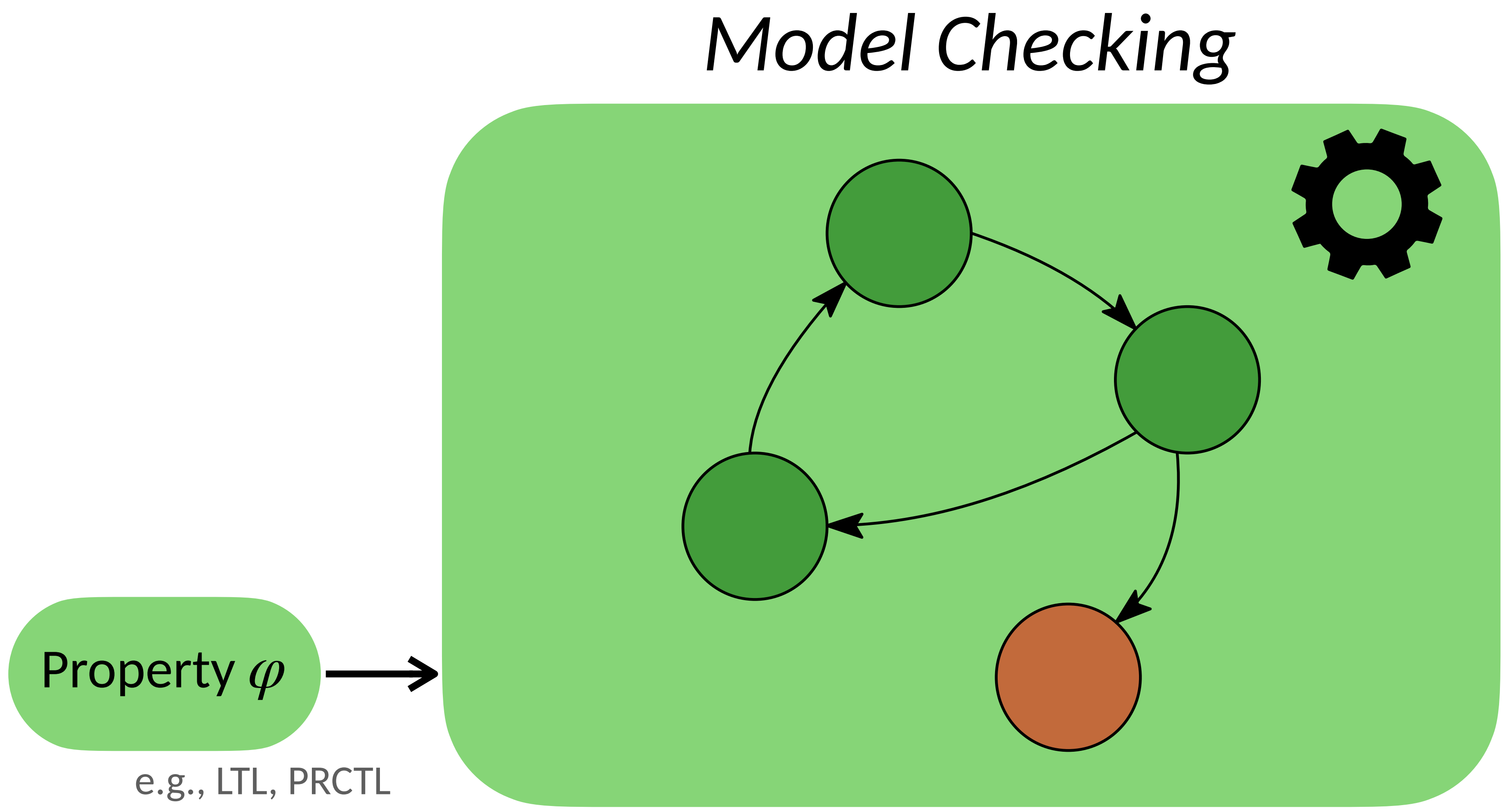
Overview

Reinforcement Learning



- Unknown environment
- Continuous state/action spaces

Formal Guarantees



- Full knowledge of the model of the interaction

Overview

Reinforcement Learning

Formal Guarantees

- Unknown environment
- Continuous state/action spaces

- Full knowledge of the model of the interaction

Overview

Reinforcement Learning

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

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

Bisimulation distance

Continuous-spaces MDP

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

distance?

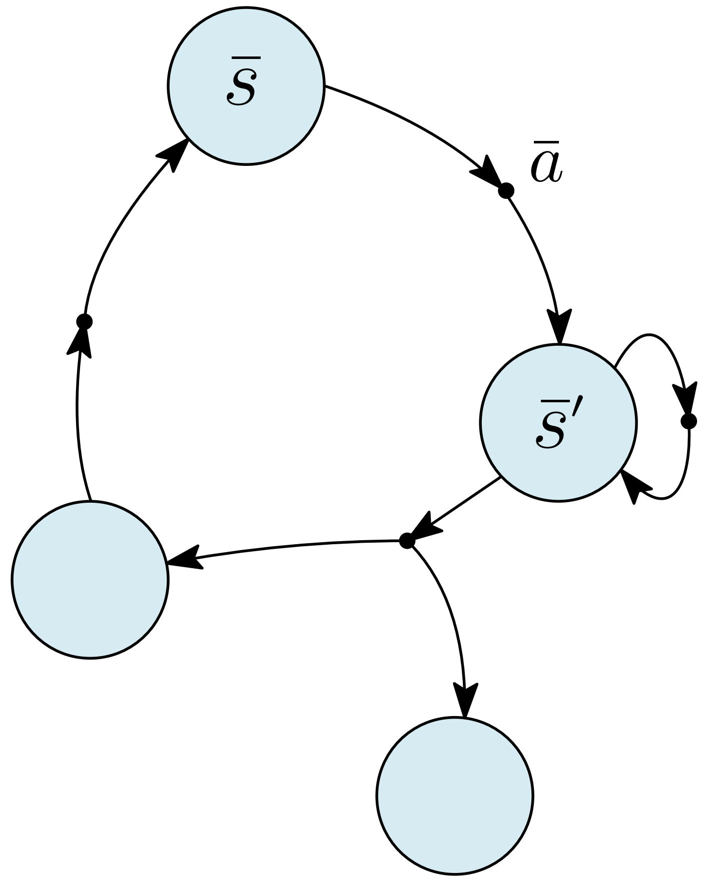
state embedding

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

action embedding

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

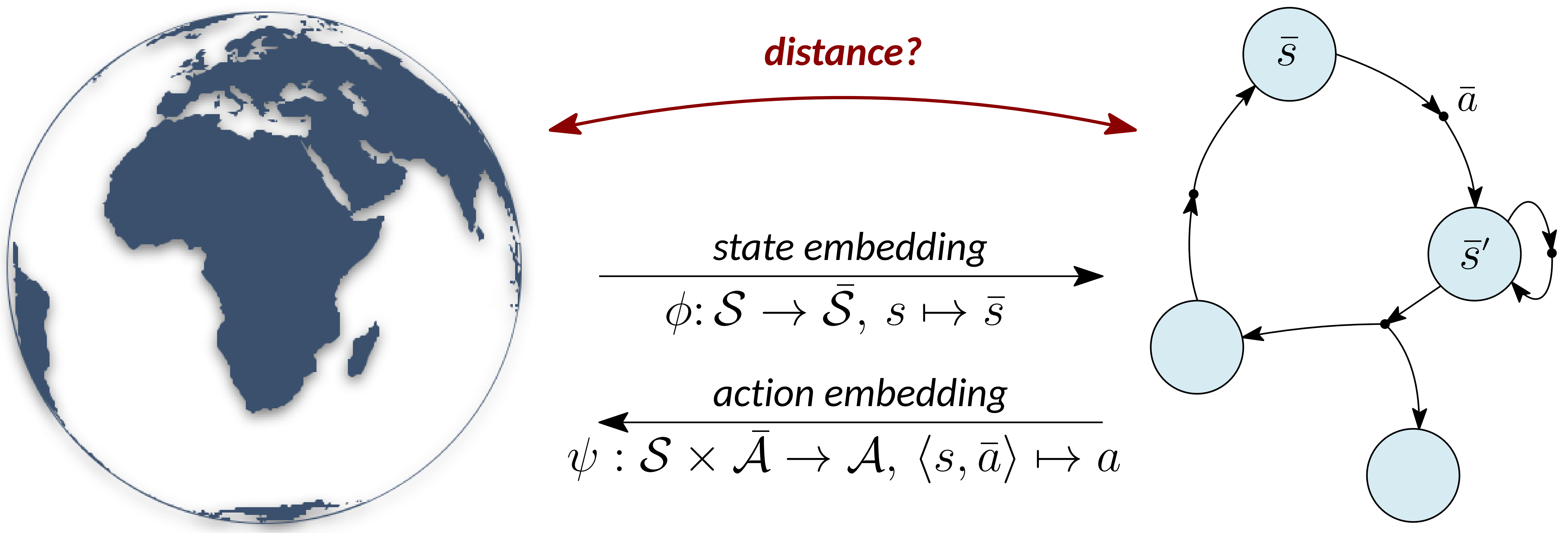
Discrete latent MDP



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

Bisimulation distance

Continuous-spaces MDP　　　　*Discrete latent* MDP



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

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

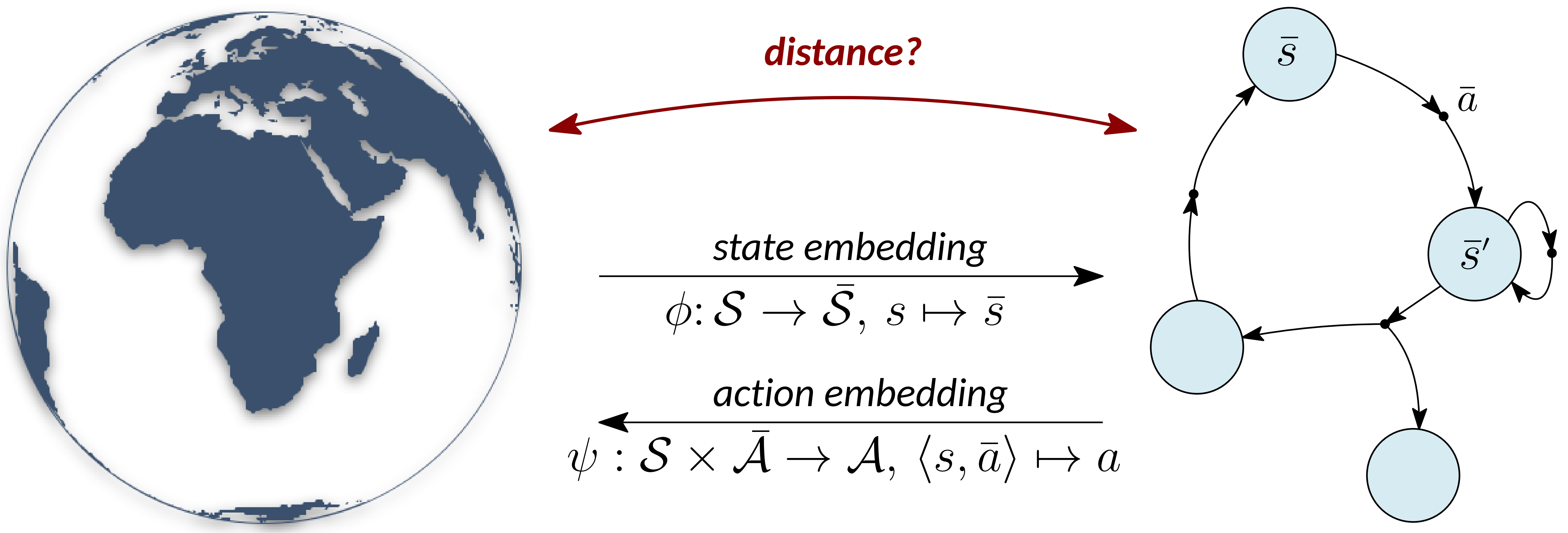
➡ ***Bisimulation distance:*** *largest behavioral difference* *(Desharnais et. al, 2004)*

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

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

Bisimulation distance

Continuous-spaces MDP | *Discrete latent* MDP



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

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

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

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

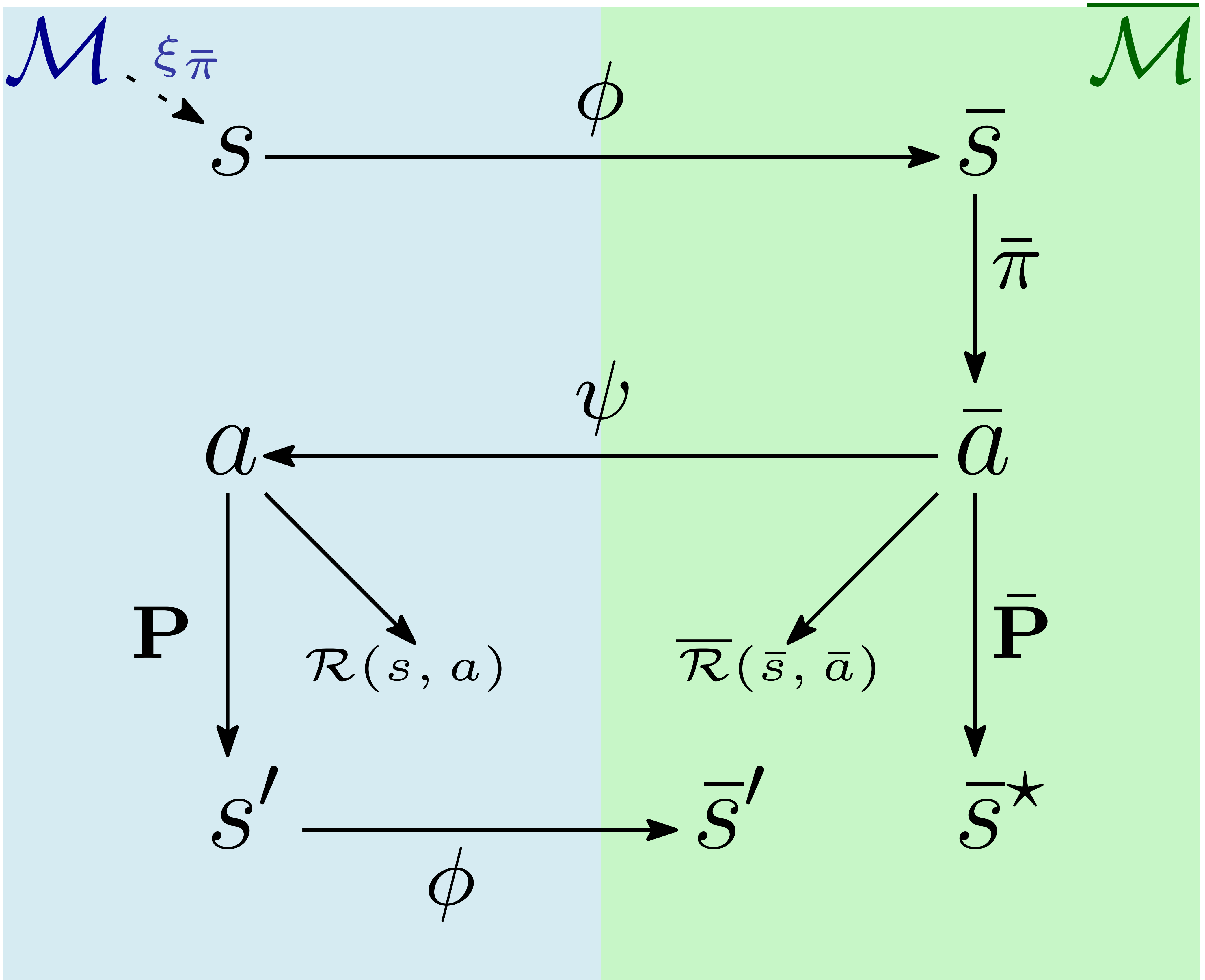
$$\tilde{d}_\pi(s_1, s_2) = \sup_{V \in \mathcal{F}_\gamma^{\mathscr{L}}(\pi)} \left| V_\pi(s_1) - V_\pi(s_2) \right| \quad \forall s_1, s_2 \in \mathcal{S}$$

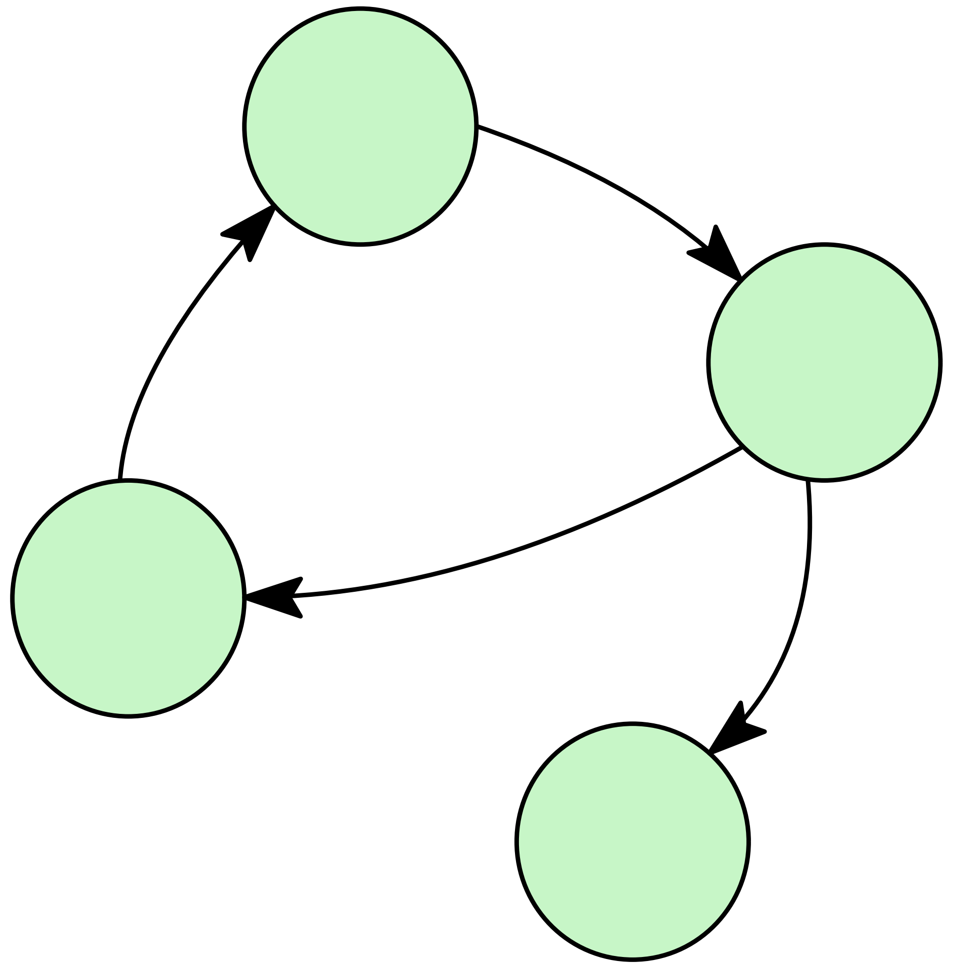
where $\mathcal{F}_\gamma^{\mathscr{L}}(\pi)$ *is a logical family of functional expressions defining the semantics of* $\mathscr{L}$

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

Latent Flow

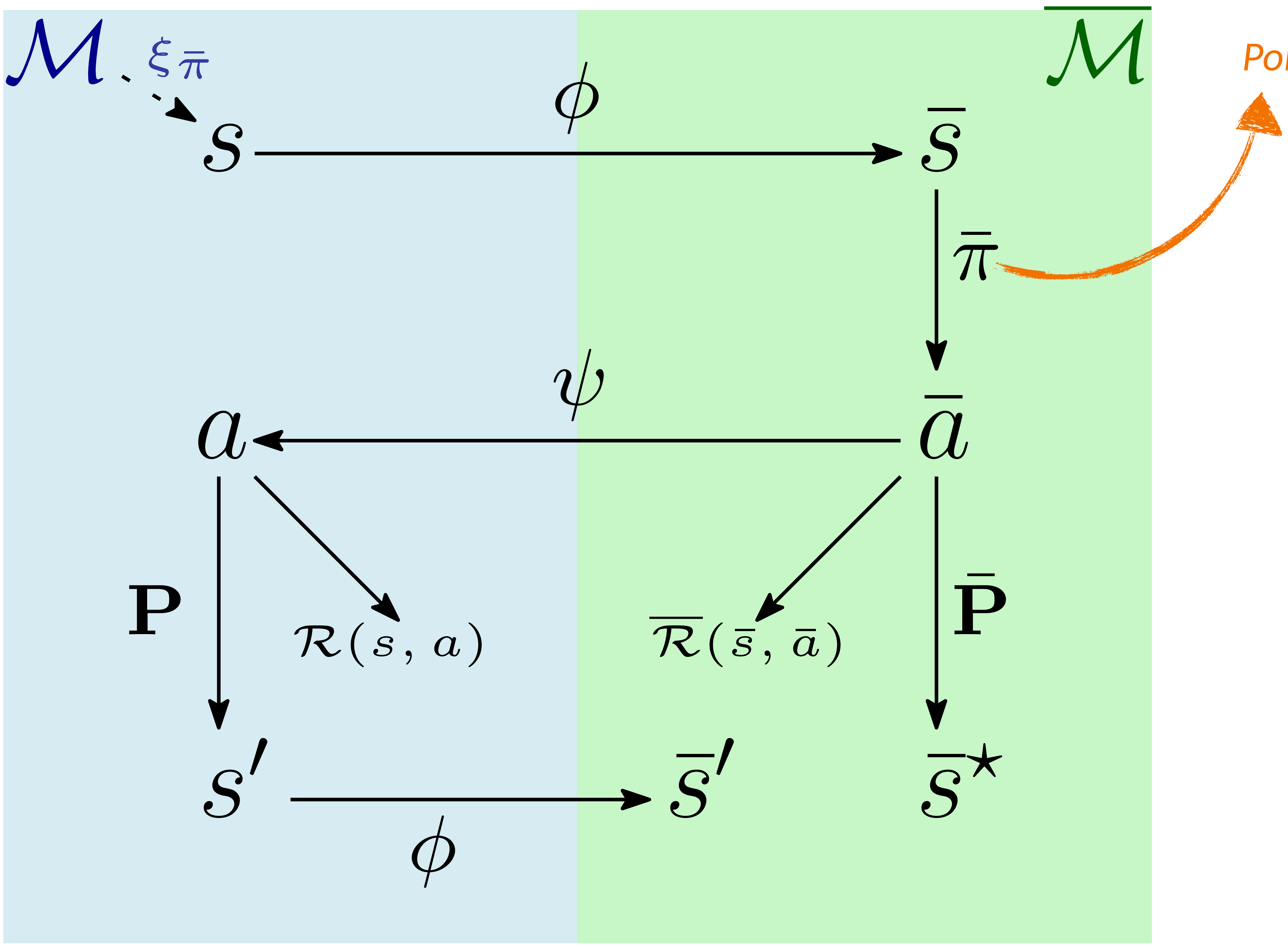
Execution of a latent policy $\bar{\pi}$ in the original model

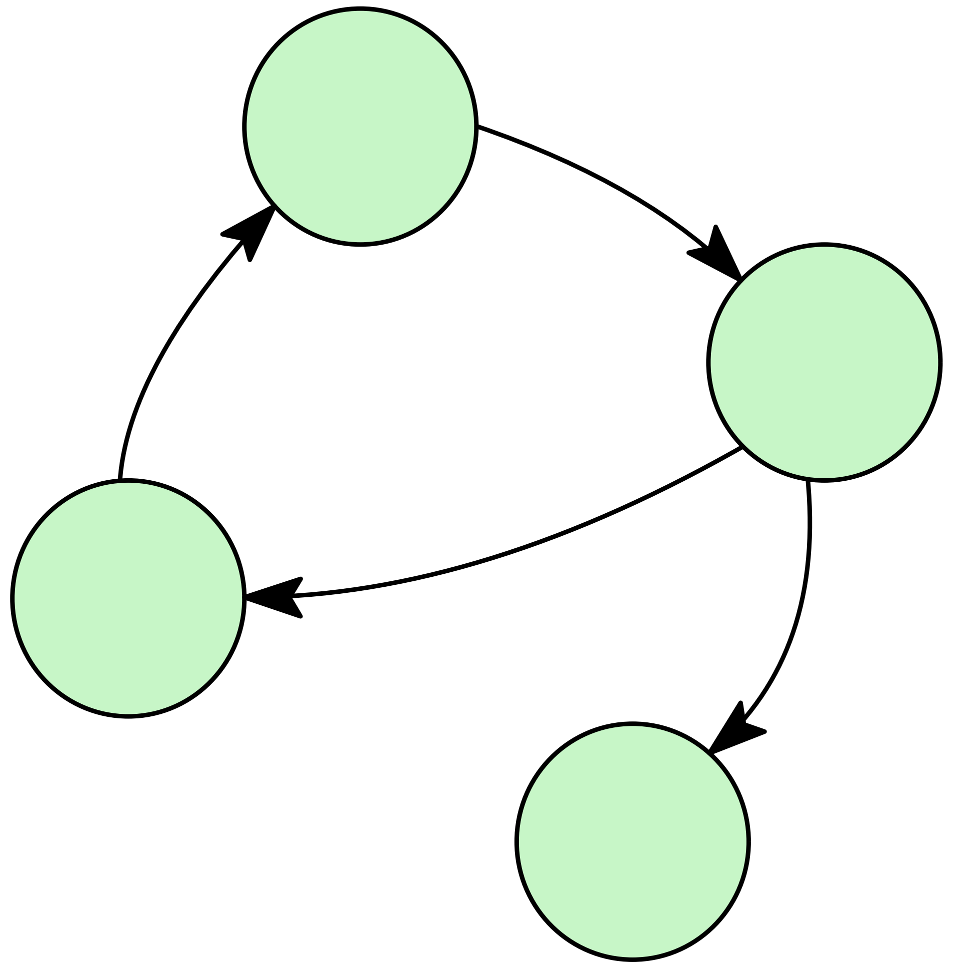




Latent Flow

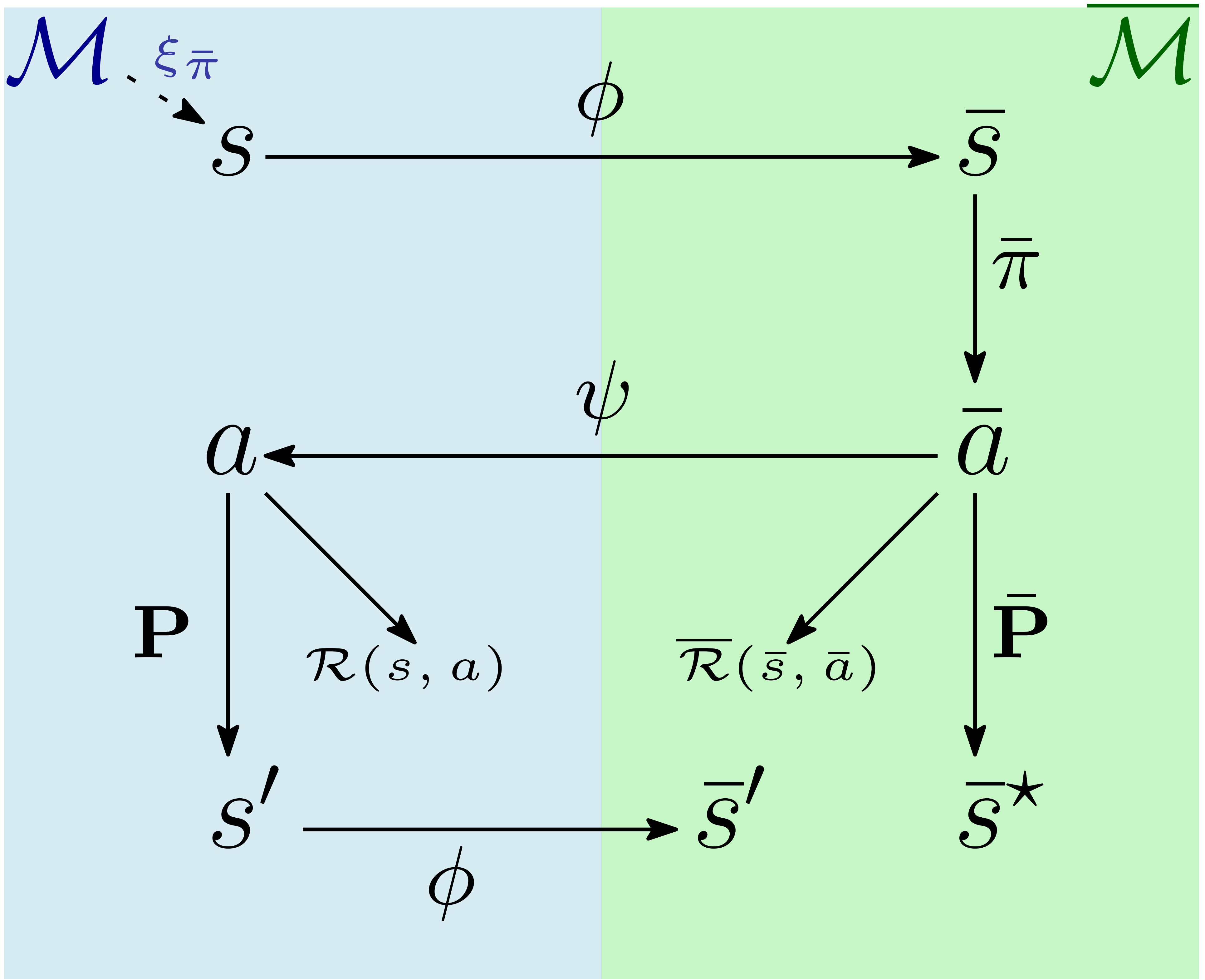
Execution of a latent policy $\bar{\pi}$ in the original model

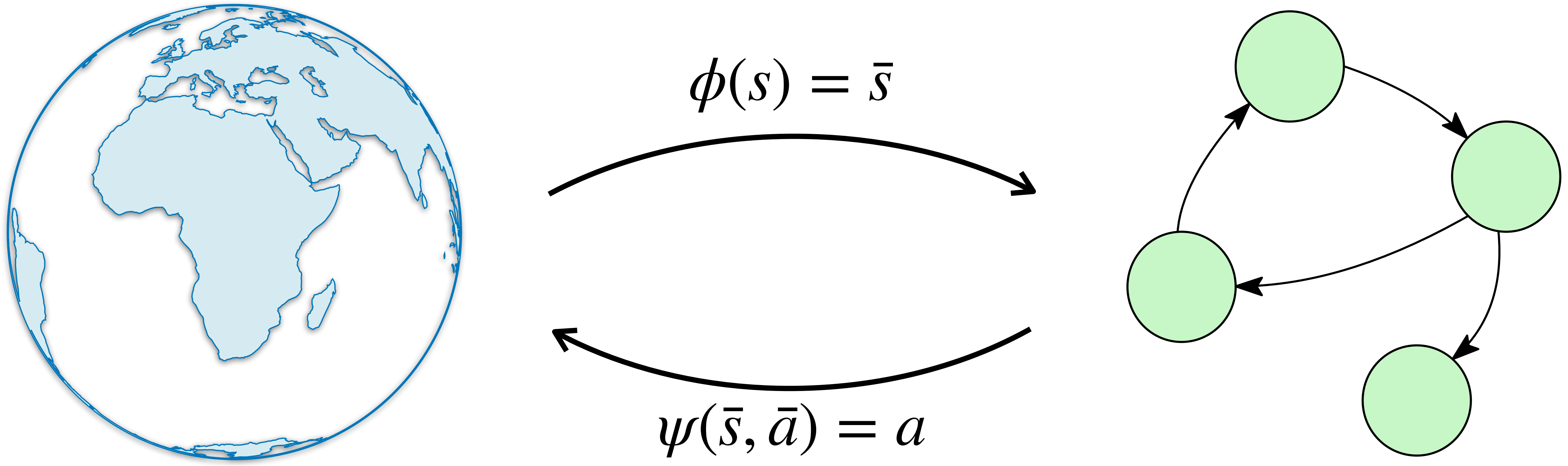




Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**

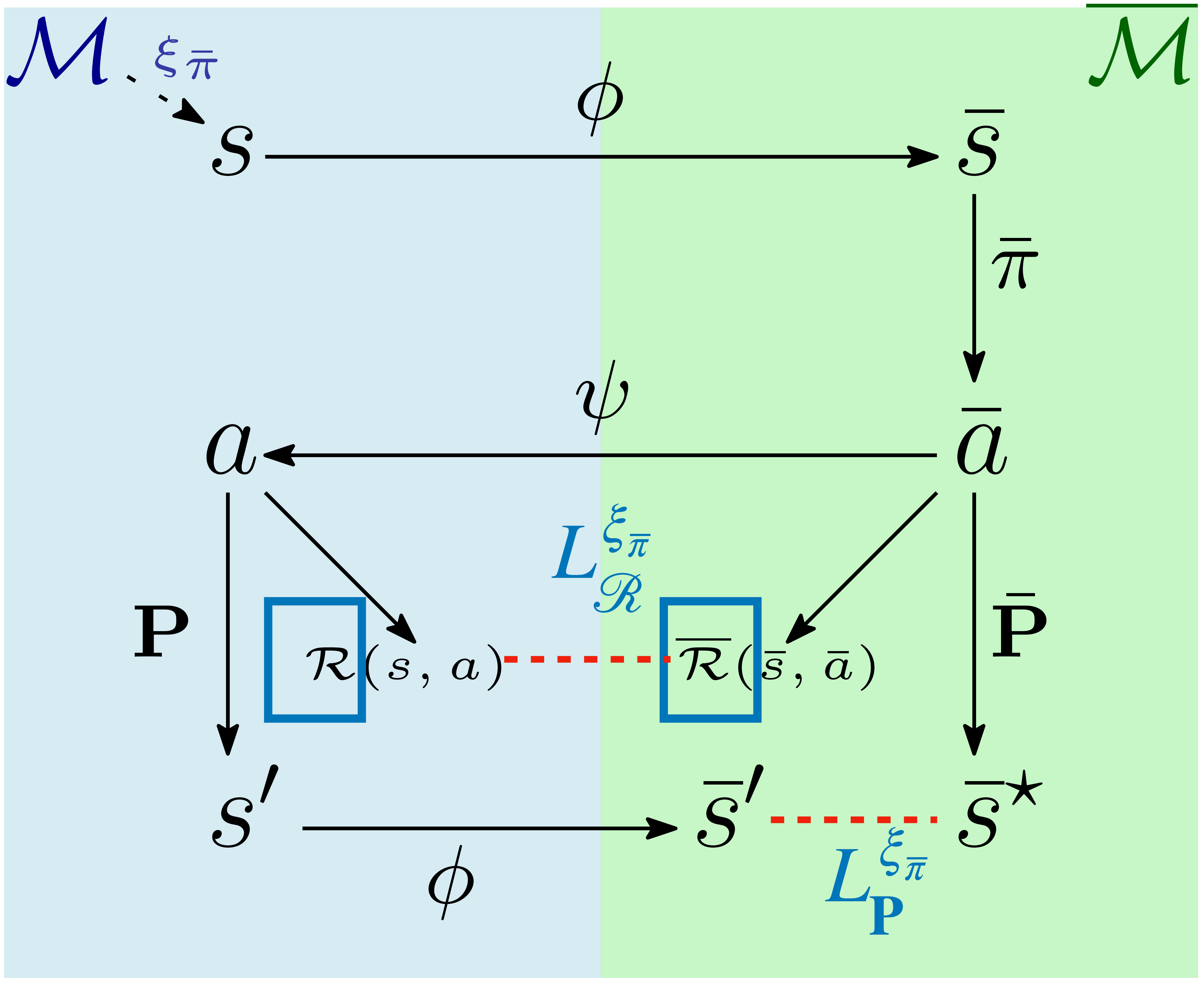
- Latent policy $\bar{\pi}$, stationary distribution $\xi_{\bar{\pi}}$





Latent Flow

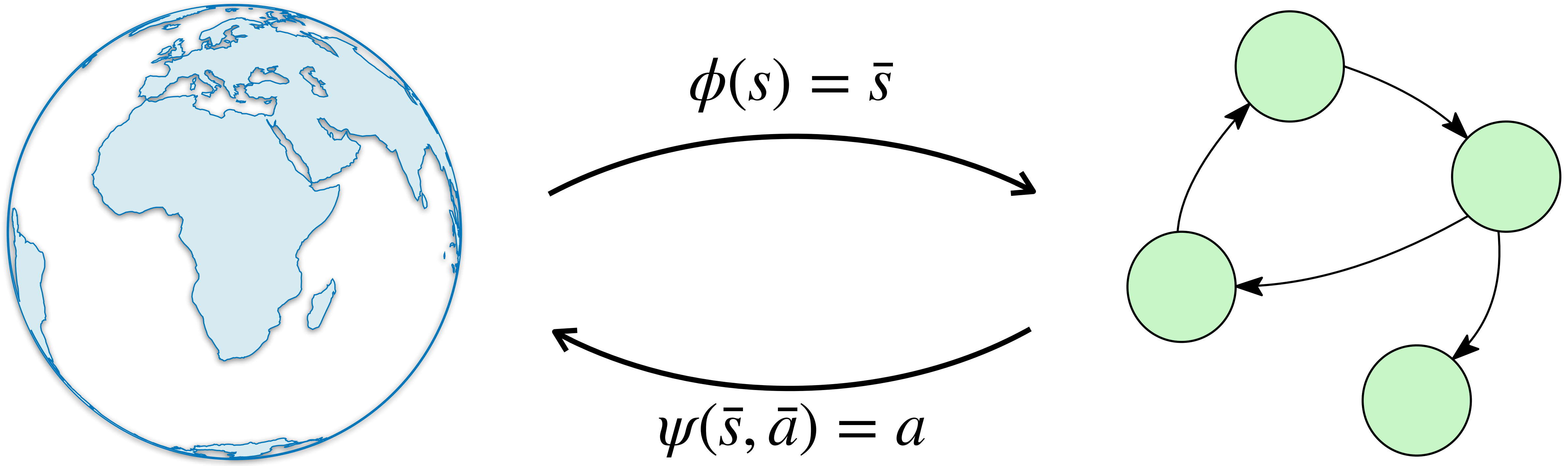
Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**



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

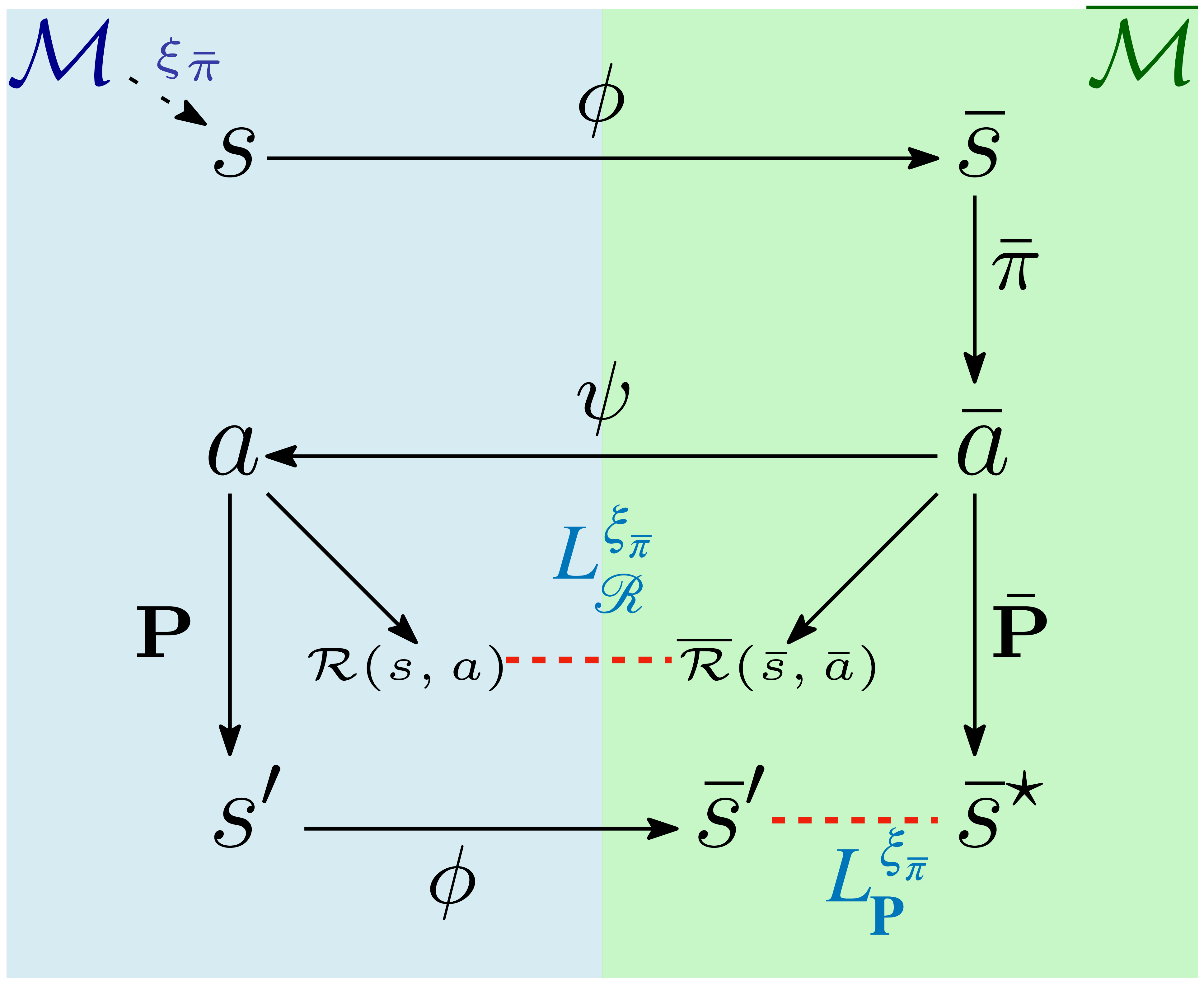
$$L_{\mathbf{P}}^{\xi_{\bar{\pi}}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} W_{d_{\bar{\mathcal{S}}}}\left(\phi\mathbf{P}\left(\cdot \mid s, \bar{a}\right), \overline{\mathbf{P}}\left(\cdot \mid \phi(s), \bar{a}\right)\right)$$

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



Latent Flow

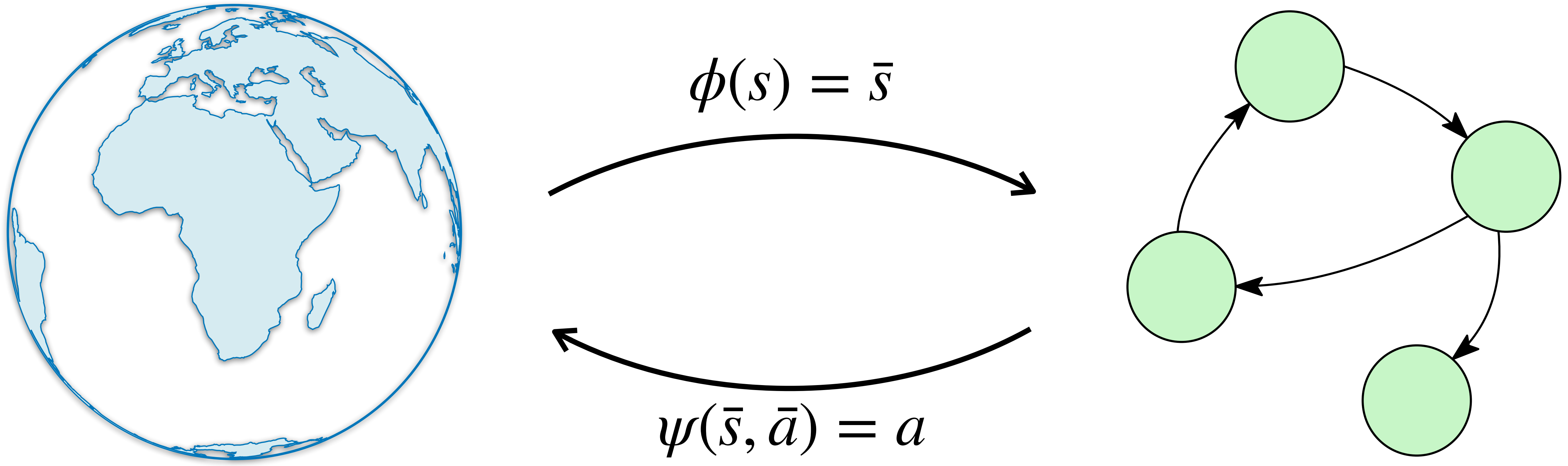
Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**



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

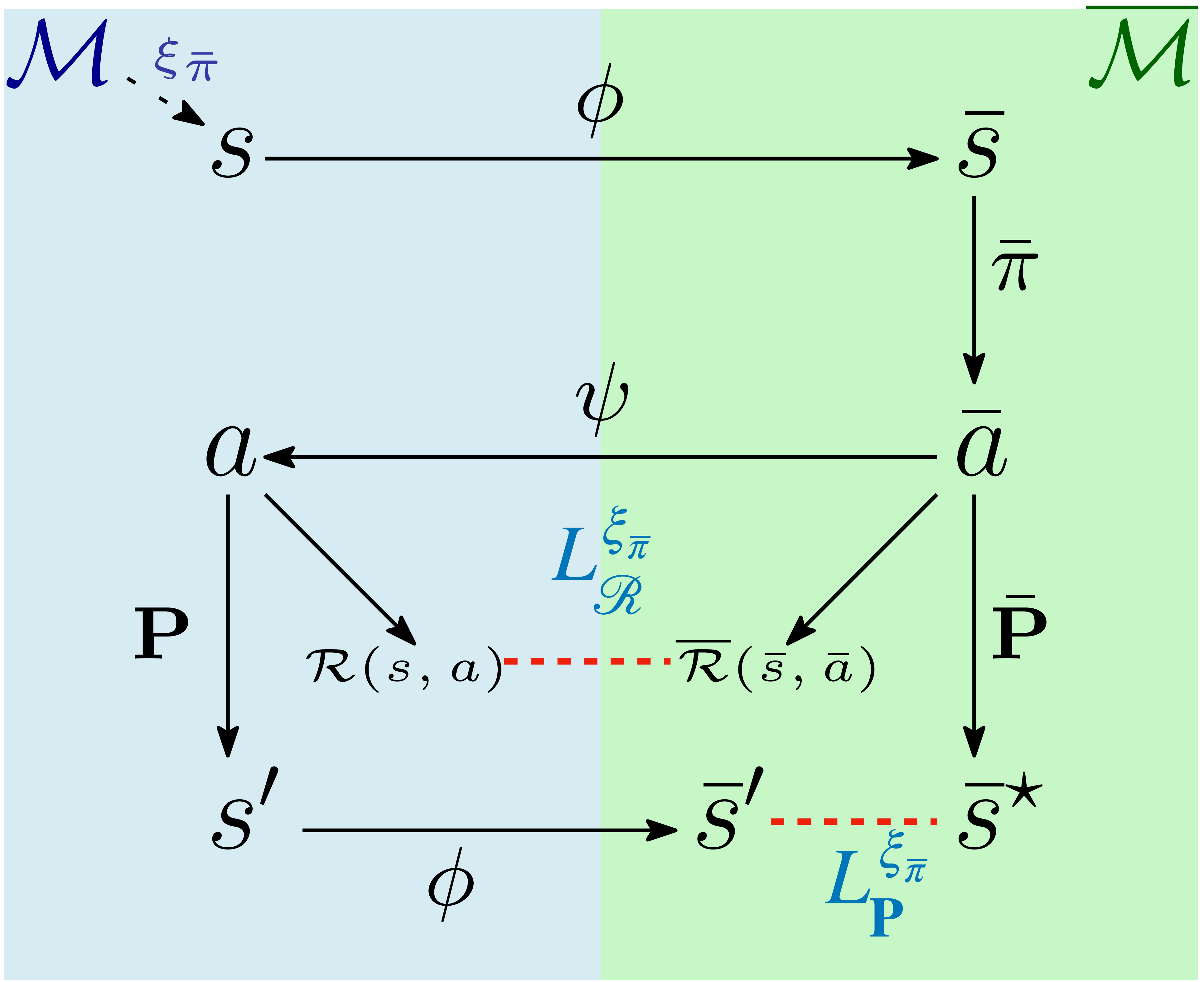
$$L^{\xi_{\bar{\pi}}}_{\mathbf{P}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} W_{d_{\bar{\mathcal{S}}}}\left(\phi\mathbf{P}\left(\cdot \mid s, \bar{a}\right), \overline{\mathbf{P}}\left(\cdot \mid \phi(s), \bar{a}\right)\right)$$

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



Latent Flow

Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**

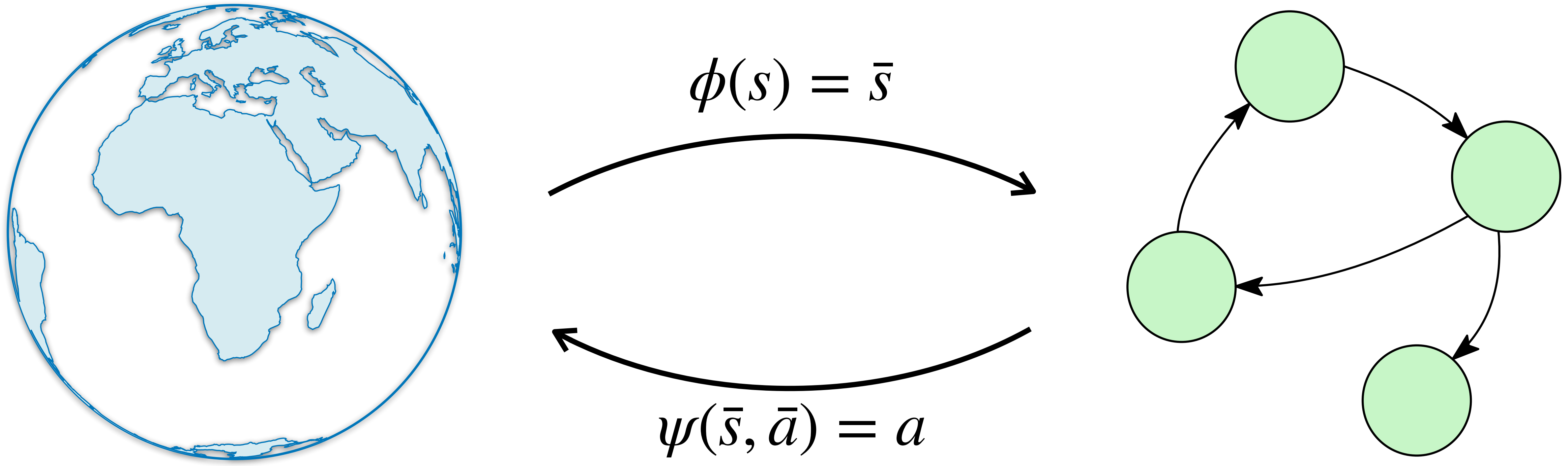


- Latent policy $\bar{\pi}$, stationary distribution $\xi_{\bar{\pi}}$

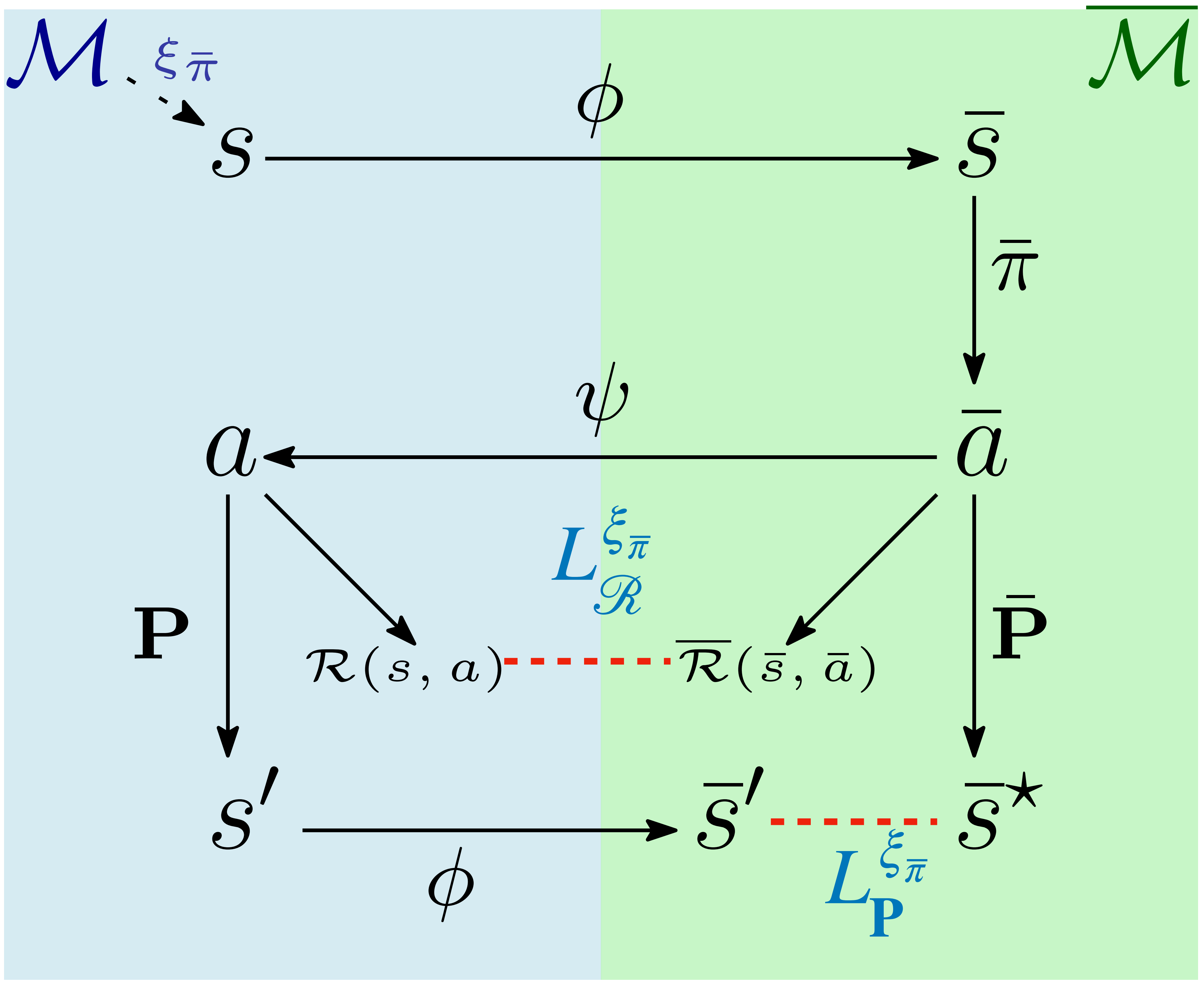
$$L_{\mathbf{P}}^{\xi_{\bar{\pi}}} = \mathbb{E}_{s,\bar{a}\sim\xi_{\bar{\pi}}} W_{d_{\bar{\mathcal{S}}}}\left(\phi\mathbf{P}\left(\cdot \mid s, \bar{a}\right), \bar{\mathbf{P}}\left(\cdot \mid \phi(s), \bar{a}\right)\right)$$

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

- ***Abstraction quality:*** $\mathbb{E}_{s\sim\xi_{\bar{\pi}}} \tilde{d}_{\bar{\pi}}\left(s, \phi(s)\right) \leq \dfrac{L_{\mathcal{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}$



Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**



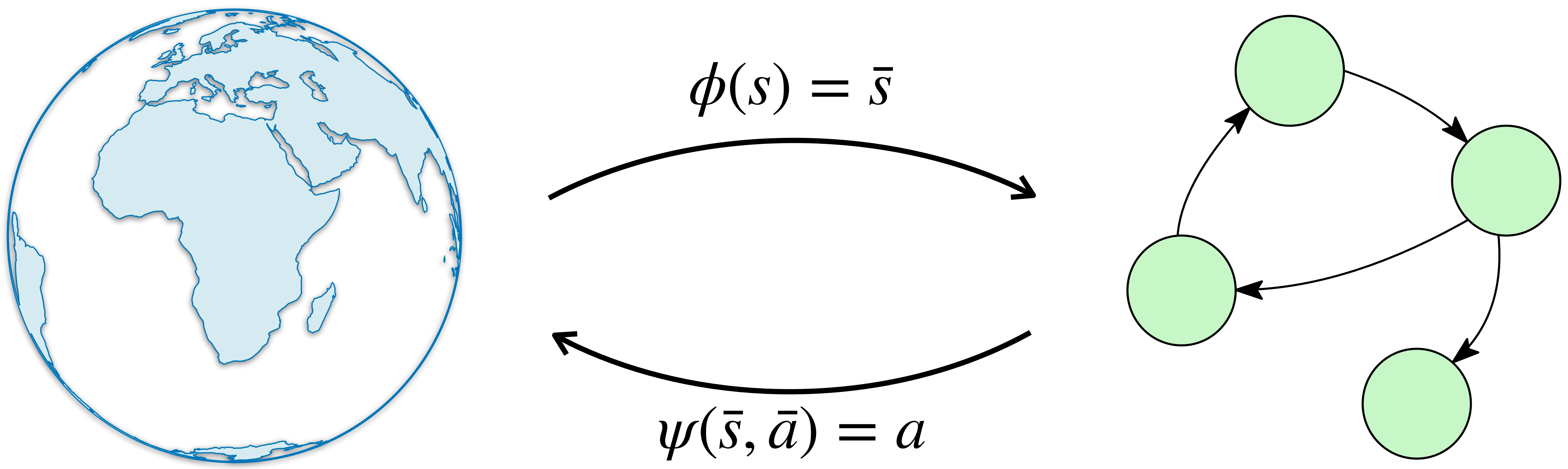
- Latent policy $\bar{\pi}$, stationary distribution $\xi_{\bar{\pi}}$

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

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

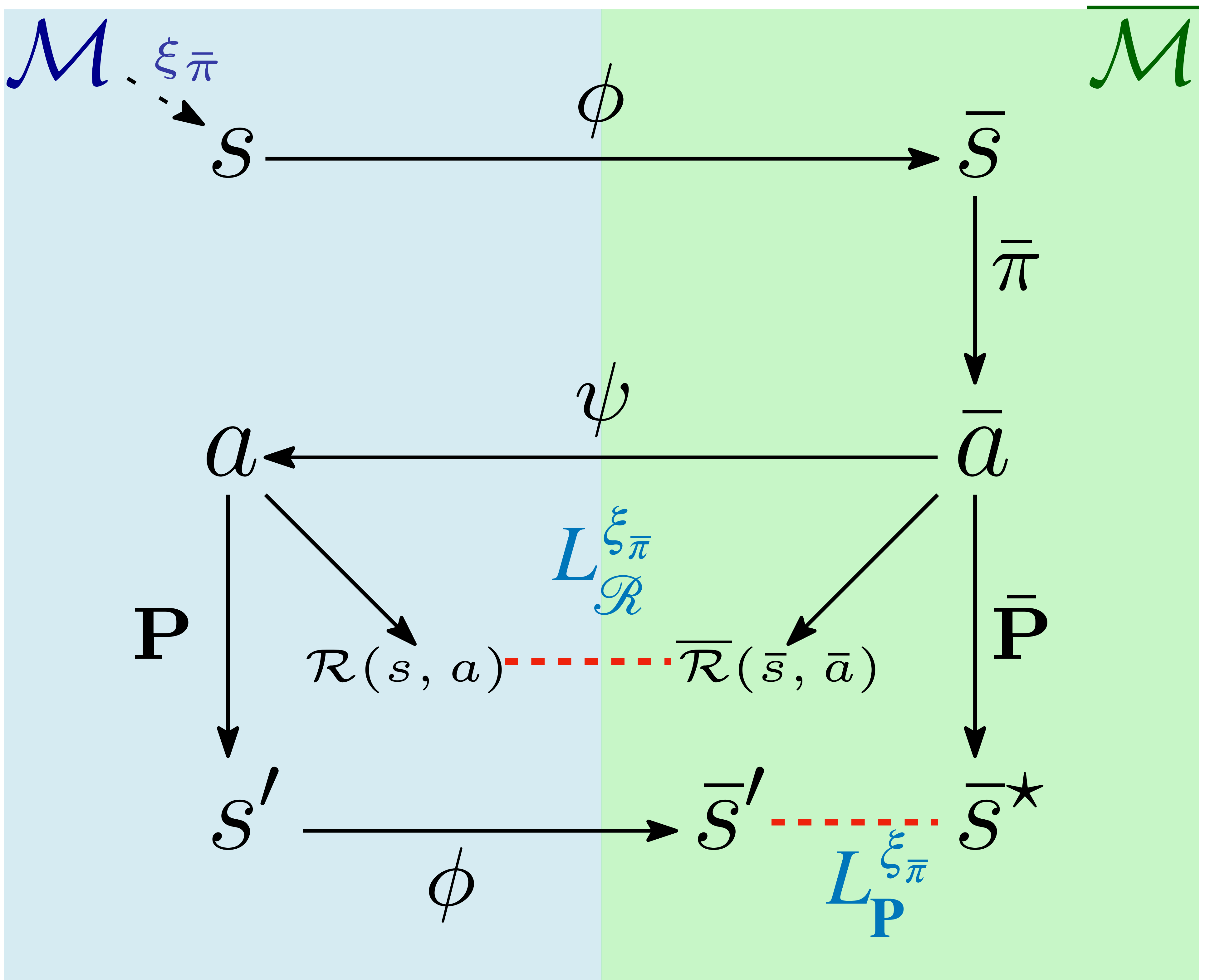
- ***Abstraction quality:*** $\mathbb{E}_{s\sim\xi_{\bar{\pi}}} \tilde{d}_{\bar{\pi}}\left(s, \phi(s)\right) \leq \dfrac{L_{\mathcal{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}$
- ***Representation quality:*** for all $s_1, s_2 \in \mathcal{S}$ such that $\phi(s_1) = \phi(s_2)$

$$\tilde{d}_{\bar{\pi}}\left(s_1, s_2\right) \leq \left(\frac{L_{\mathcal{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}\right) \cdot \left(\xi_{\bar{\pi}}^{-1}\left(s_1\right) + \xi_{\bar{\pi}}^{-1}\left(s_2\right)\right)$$



Latent Flow

Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**



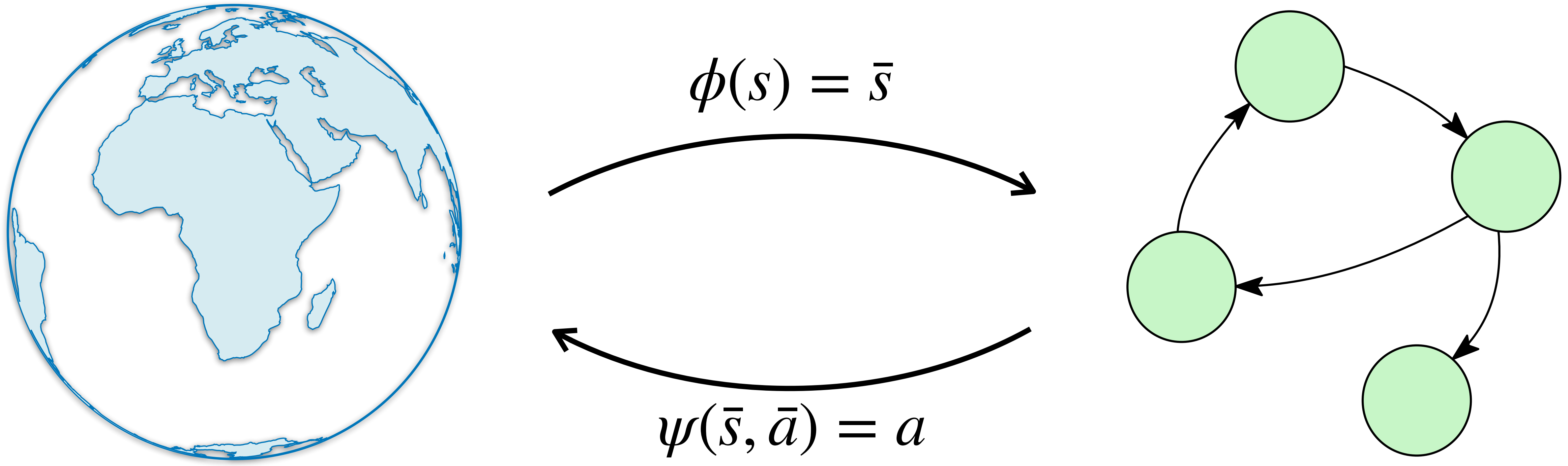
- Latent policy $\bar{\pi}$, stationary distribution $\xi_{\bar{\pi}}$

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

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

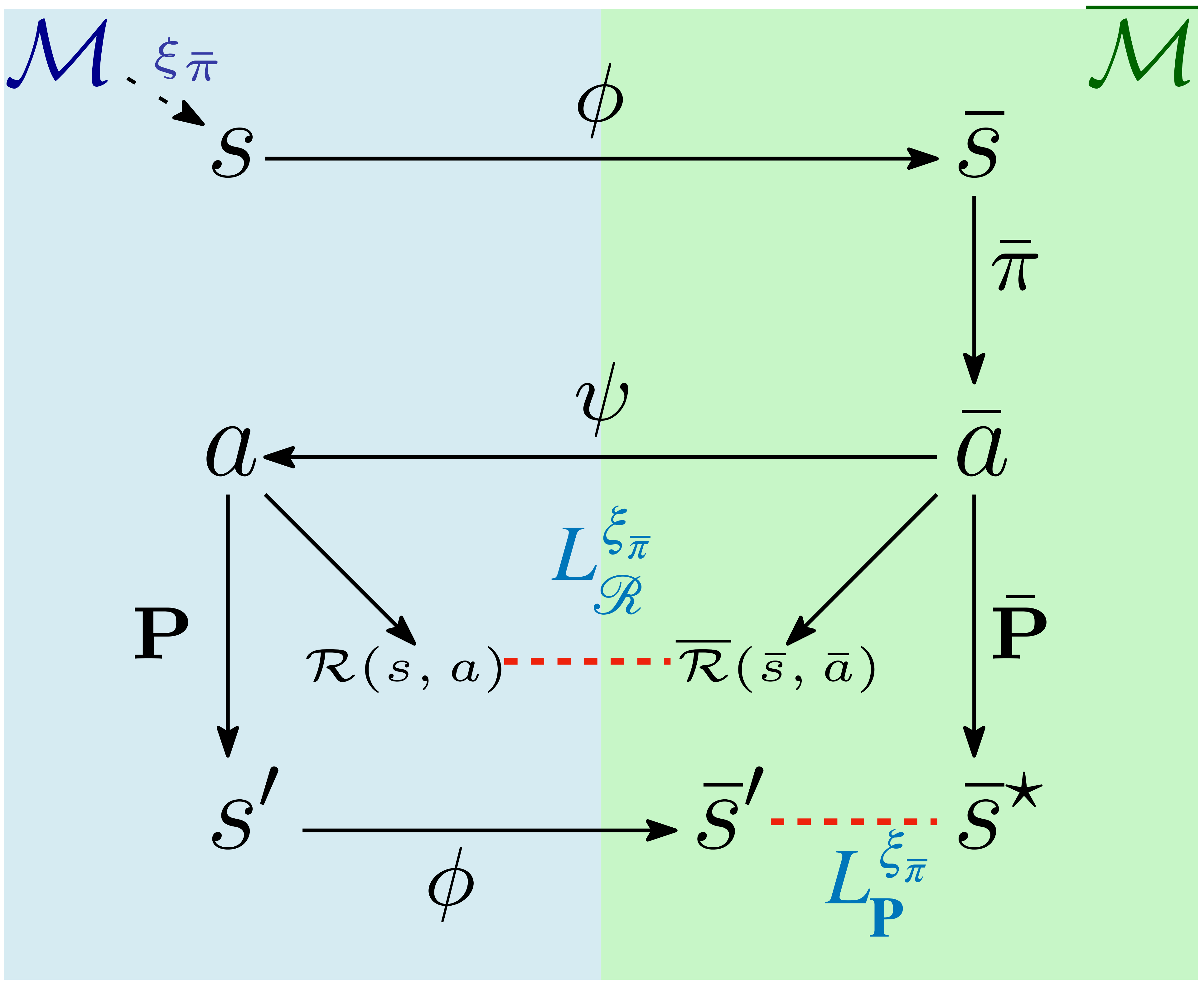
- ***Abstraction quality:*** $\mathbb{E}_{s\sim\xi_{\bar{\pi}}} \tilde{d}_{\bar{\pi}}\left(s, \phi(s)\right) \leq \dfrac{L_{\mathcal{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}$
- ***Representation quality:*** for all $s_1, s_2 \in \mathcal{S}$ such that $\phi(s_1) = \phi(s_2)$

$$\tilde{d}_{\bar{\pi}}\left(s_1, s_2\right) \leq \left(\frac{L_{\mathcal{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}\right) \cdot \left(\xi_{\bar{\pi}}^{-1}\left(s_1\right) + \xi_{\bar{\pi}}^{-1}\left(s_2\right)\right)$$



Latent Flow

Execution of a latent policy $\bar{\pi}$ in the original model: **Local Losses**



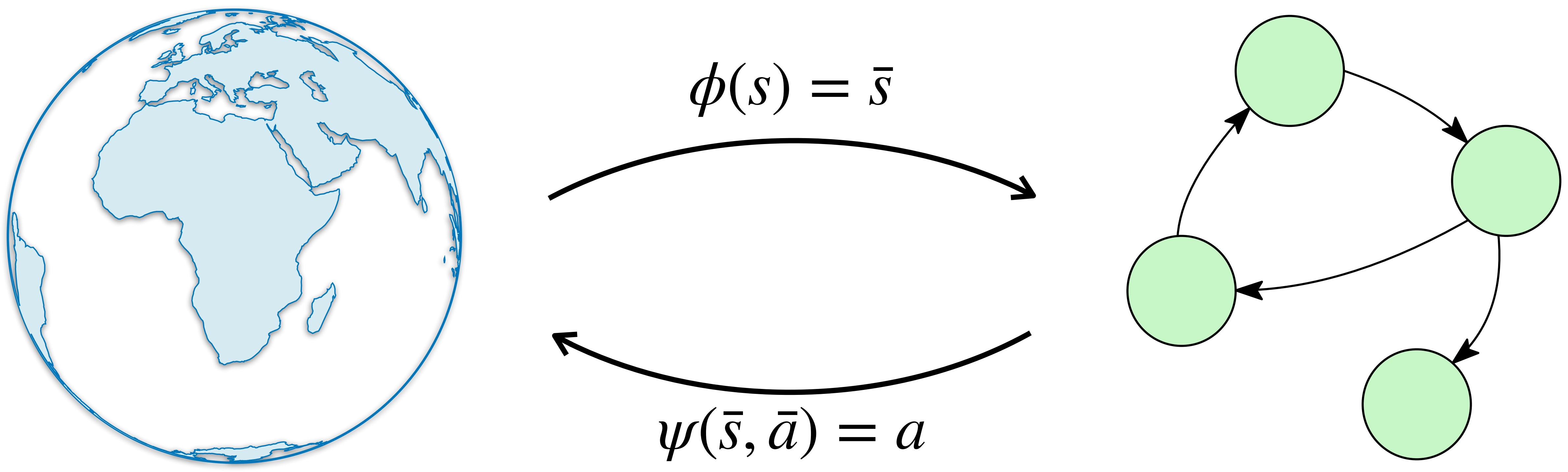
- Latent policy $\bar{\pi}$, stationary distribution $\xi_{\bar{\pi}}$

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

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

- ***Abstraction quality:*** $\mathbb{E}_{s\sim\xi_{\bar{\pi}}} \tilde{d}_{\bar{\pi}}\left(s, \phi(s)\right) \leq \dfrac{L_{\mathcal{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}$
- ***Representation quality:*** for all $s_1, s_2 \in \mathcal{S}$ such that $\phi(s_1) = \phi(s_2)$

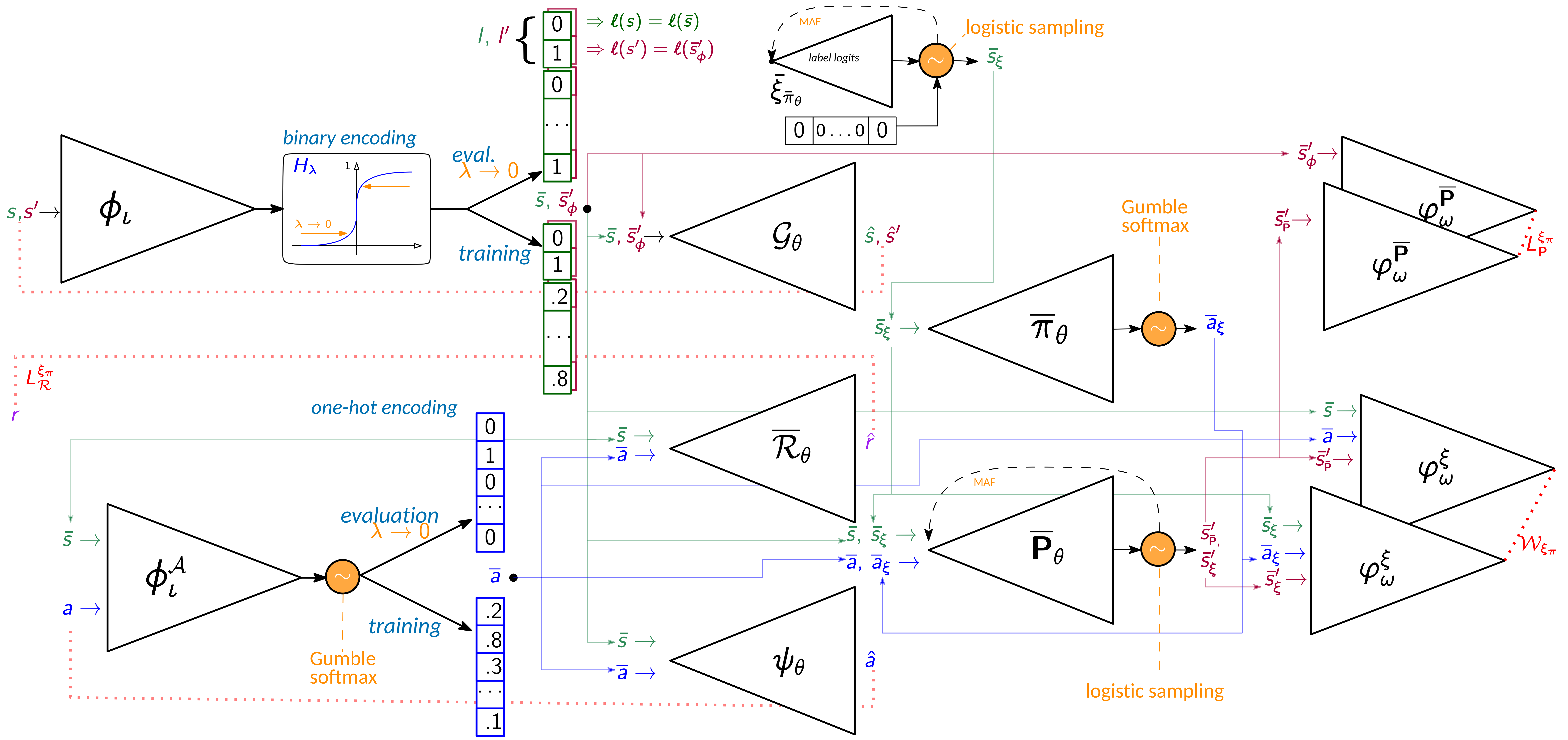
$$\tilde{d}_{\bar{\pi}}\left(s_1, s_2\right) \leq \left(\frac{L_{\mathcal{R}}^{\xi_{\bar{\pi}}} + \gamma L_{\mathbf{P}}^{\xi_{\bar{\pi}}}}{1-\gamma}\right) \cdot \left(\xi_{\bar{\pi}}^{-1}\left(s_1\right) + \xi_{\bar{\pi}}^{-1}\left(s_2\right)\right)$$



$$\min_{\iota,\theta} \mathbb{E}_{s,a,s' \sim \xi_\pi} \mathbb{E}_{\bar{s},\bar{a},\bar{s}' \sim \phi_\iota(\,\cdot \mid s,a,s')} \left\| \langle s, a, s' \rangle - \langle \mathcal{G}_\theta(\bar{s}), \psi_\theta(\bar{s}, \bar{a}), \mathcal{G}_\theta(\bar{s}') \rangle \right\| + L_{\mathcal{R}}^{\xi\pi} + \beta \left(\mathcal{W}_{\xi_\pi} + L_{\mathbf{P}}^{\xi_\pi} \right)$$

Wasserstein Auto-encoded Markov Decision Process

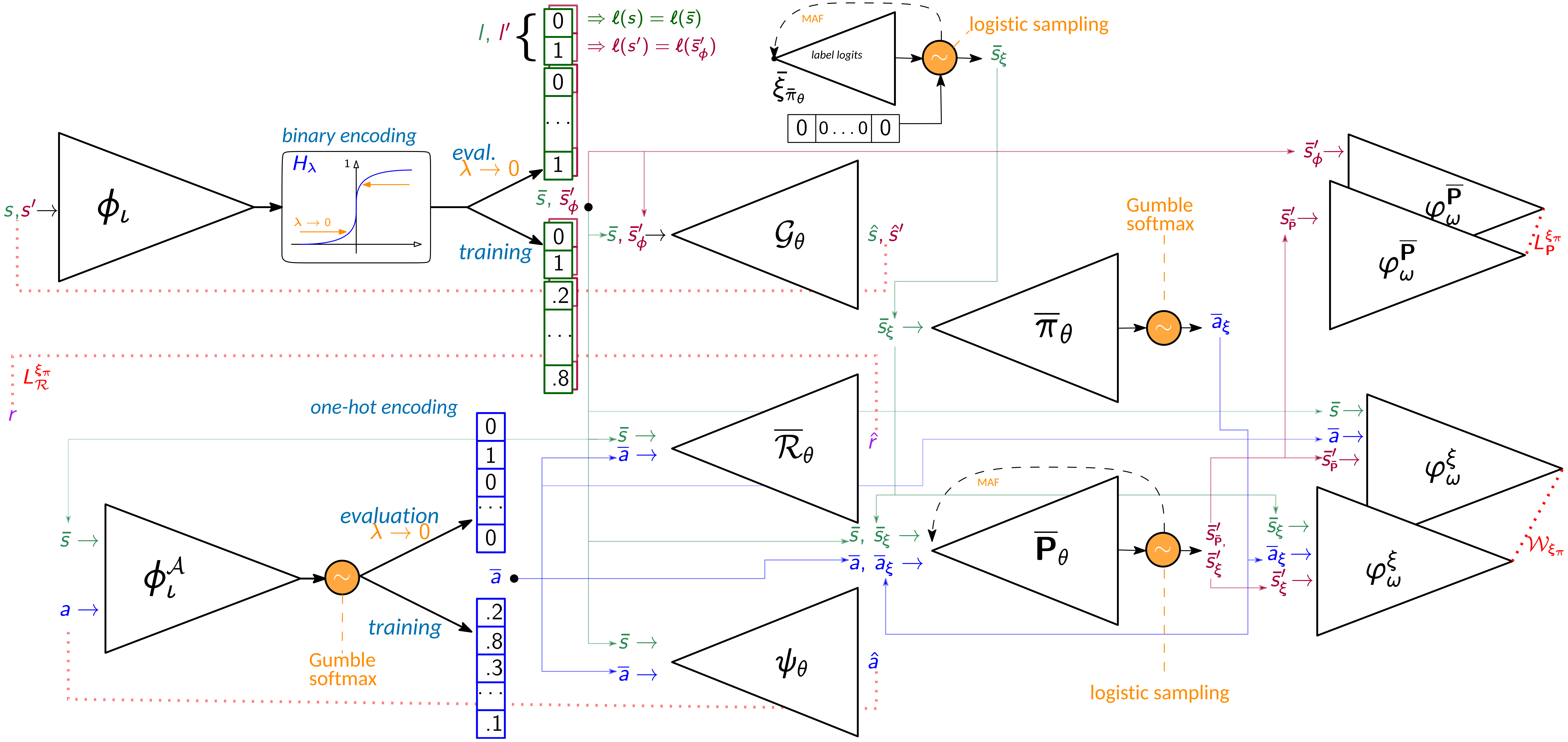
$$\min_{\iota,\theta} \mathbb{E}_{s,a,s'\sim\xi_\pi}\mathbb{E}_{\bar{s},\bar{a},\bar{s}'\sim\phi_\iota(\cdot\mid s,a,s')} \left\|\langle s,a,s'\rangle - \langle \mathcal{G}_\theta(\bar{s}),\psi_\theta(\bar{s},\bar{a}),\mathcal{G}_\theta(\bar{s}')\rangle\right\| + L_{\mathcal{R}}^{\xi_\pi} + \beta\left(\mathcal{W}_{\xi_\pi} + L_{\mathbf{P}}^{\xi_\pi}\right)$$



- $\mathcal{W}_{\xi_\pi} = \max_{\omega:\ \|\varphi_\omega^\xi\|\leq 1} \mathbb{E}_{s,a\sim\xi_\pi}\mathbb{E}_{\bar{a}\sim\phi_\iota^{\mathcal{A}}(\cdot\mid\phi_\iota(s),a)}\mathbb{E}_{\bar{s}'\sim\overline{\mathbf{P}}_\theta(\cdot\mid\bar{s},\bar{a})}\ \varphi_\omega^\xi\left(\phi_\iota(s),\bar{a},\bar{s}'\right) - \mathbb{E}_{\bar{s},\bar{a},\bar{s}'\sim\xi_{\bar{\pi}}}\ \varphi_\omega^\xi(\bar{s},\bar{a},\bar{s}')$
- $L_{\mathbf{P}}^{\xi_\pi} = \max_{\omega:\ \|\varphi_\omega^{\mathbf{P}}\|\leq 1} \mathbb{E}_{s,a,s'\sim\xi_\pi}\mathbb{E}_{\bar{s},\bar{a},\bar{s}'\sim\phi_\iota(\cdot\mid s,a,s')}\left[\varphi_\omega^{\mathbf{P}}(s,a,\bar{s},\bar{a},\bar{s}') - \mathbb{E}_{\bar{s}'_{\mathbf{P}}\sim\overline{\mathbf{P}}_\theta(\cdot\mid\bar{s},\bar{a})}\ \varphi_\omega^{\mathbf{P}}\left(s,a,\bar{s},\bar{a},\bar{s}'_{\mathbf{P}}\right)\right]$

Wasserstein Auto-encoded Markov Decision Process

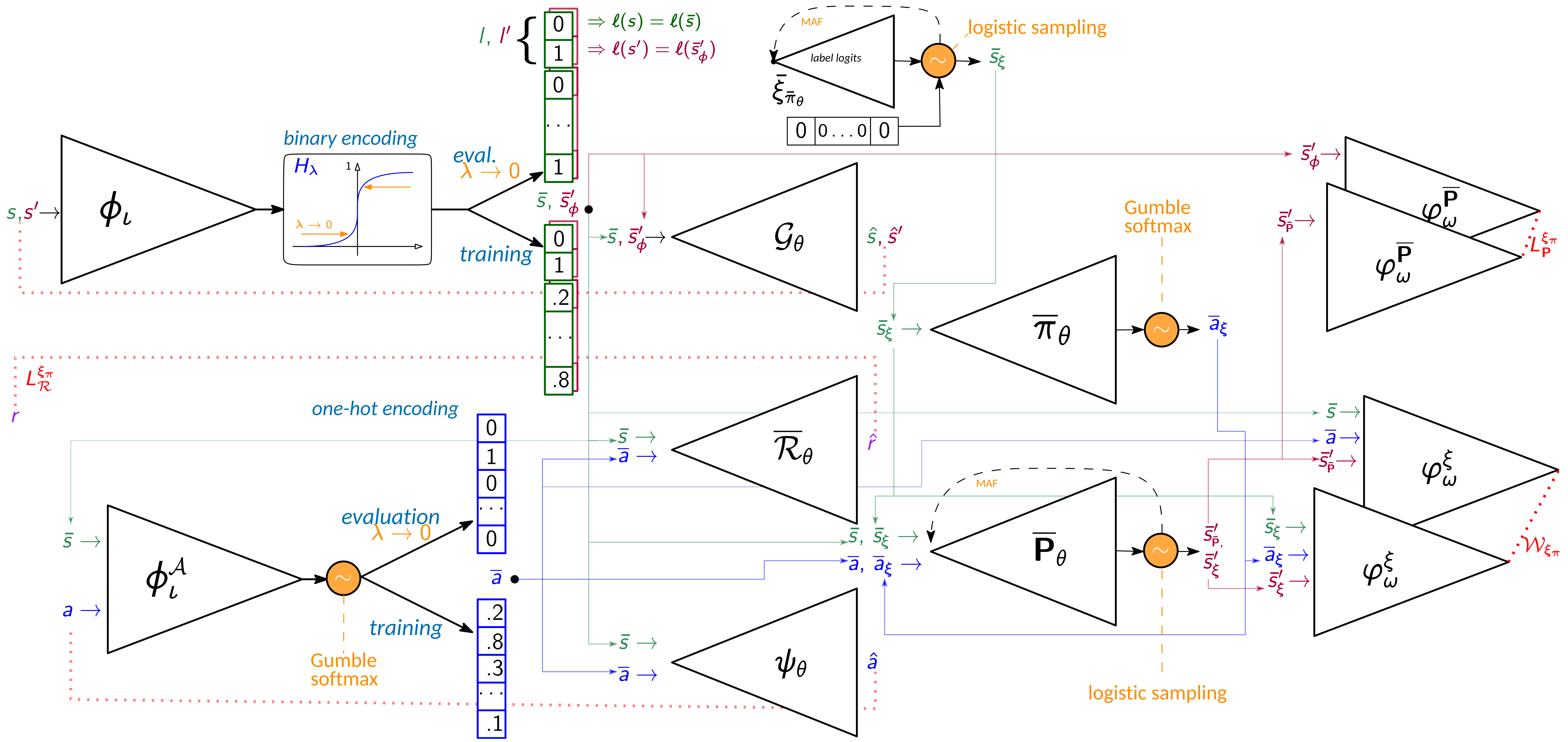
$$\min_{\iota,\theta} \mathbb{E}_{s,a,s'\sim\xi_\pi}\mathbb{E}_{\bar s,\bar a,\bar s'\sim\phi_\iota(\cdot\mid s,a,s')} \|\langle s,a,s'\rangle - \langle \mathcal{G}_\theta(\bar s), \psi_\theta(\bar s,\bar a), \mathcal{G}_\theta(\bar s')\rangle\| + L^{\xi_\pi}_{\mathcal{R}} + \beta\left(\mathcal{W}_{\xi_\pi} + L^{\xi_\pi}_{\mathbf{P}}\right)$$



- $\mathcal{W}_{\xi_\pi} = \max_{\omega:\ \|\varphi^\xi_\omega\|\le 1} \mathbb{E}_{s,a\sim\xi_\pi}\mathbb{E}_{\bar a\sim\phi^{\mathcal{A}}_\iota(\cdot\mid\phi_\iota(s),a)}\mathbb{E}_{\bar s'\sim\bar{\mathbf{P}}_\theta(\cdot\mid\bar s,\bar a)}\ \varphi^\xi_\omega\left(\phi_\iota(s),\bar a,\bar s'\right) - \mathbb{E}_{\bar s,\bar a,\bar s'\sim\xi_{\bar\pi}}\ \varphi^\xi_\omega(\bar s,\bar a,\bar s')$
- $L^{\xi_\pi}_{\mathbf{P}} = \max_{\omega:\ \|\varphi^{\mathbf{P}}_\omega\|\le 1} \mathbb{E}_{s,a,s'\sim\xi_\pi}\mathbb{E}_{\bar s,\bar a,\bar s'\sim\phi_\iota(\cdot\mid s,a,s')}\left[\varphi^{\mathbf{P}}_\omega(s,a,\bar s,\bar a,\bar s') - \mathbb{E}_{\bar s'_{\mathbf{P}}\sim\bar{\mathbf{P}}_\theta(\cdot\mid\bar s,\bar a)}\ \varphi^{\mathbf{P}}_\omega\left(s,a,\bar s,\bar a,\bar s'_{\mathbf{P}}\right)\right]$

Wasserstein Auto-encoded Markov Decision Process

$$\min_{\iota,\theta} \mathbb{E}_{s,a,s'\sim\xi_\pi}\mathbb{E}_{\bar{s},\bar{a},\bar{s}'\sim\phi_\iota(\cdot\mid s,a,s')} \|\langle s,a,s'\rangle - \langle \mathcal{G}_\theta(\bar{s}), \psi_\theta(\bar{s},\bar{a}), \mathcal{G}_\theta(\bar{s}')\rangle\| + L^{\xi_\pi}_{\mathcal{R}} + \beta\left(\mathcal{W}_{\xi_\pi} + L^{\xi_\pi}_{\mathbf{P}}\right)$$



- $\mathcal{W}_{\xi_\pi} = \max_{\omega:\, \|\varphi^{\xi}_\omega\|\le 1} \mathbb{E}_{s,a\sim\xi_\pi}\mathbb{E}_{\bar a\sim\phi^{\mathcal{A}}_\iota(\cdot\mid\phi_\iota(s),a)}\mathbb{E}_{\bar s'\sim\bar{\mathbf{P}}_\theta(\cdot\mid\bar s,\bar a)}\, \varphi^{\xi}_\omega\left(\phi_\iota(s),\bar a,\bar s'\right) - \mathbb{E}_{\bar s,\bar a,\bar s'\sim\xi_{\bar\pi}}\varphi^{\xi}_\omega(\bar s,\bar a,\bar s')$
- $L^{\xi_\pi}_{\mathbf{P}} = \max_{\omega:\, \|\varphi^{\mathbf{P}}_\omega\|\le 1} \mathbb{E}_{s,a,s'\sim\xi_\pi}\mathbb{E}_{\bar s,\bar a,\bar s'\sim\phi_\iota(\cdot\mid s,a,s')}\left[\varphi^{\mathbf{P}}_\omega(s,a,\bar s,\bar a,\bar s') - \mathbb{E}_{\bar s'_{\mathbf{P}}\sim\bar{\mathbf{P}}_\theta(\cdot\mid\bar s,\bar a)}\,\varphi^{\mathbf{P}}_\omega\left(s,a,\bar s,\bar a,\bar s'_{\mathbf{P}}\right)\right]$

Evaluation

Distillation: performance of $\bar{\pi}$

WAE-MDPs distill policies up to 10 times faster than VAE-MDPs

Distillation of RL Policies with Formal Guarantees via Variational Abstraction of Markov Decision Processes.
Florent Delgrange, Ann Nowé, Guillermo A. Pérez (2022). AAAI 2022.

- *Faster*
- *Better performance*
- *Learning guarantees*
- *Similar or even better model quality*