Deep SPI

Safe Policy Improvement via World Models

Florent Delgrange, Raphaël Avalos, and Willem Röpke

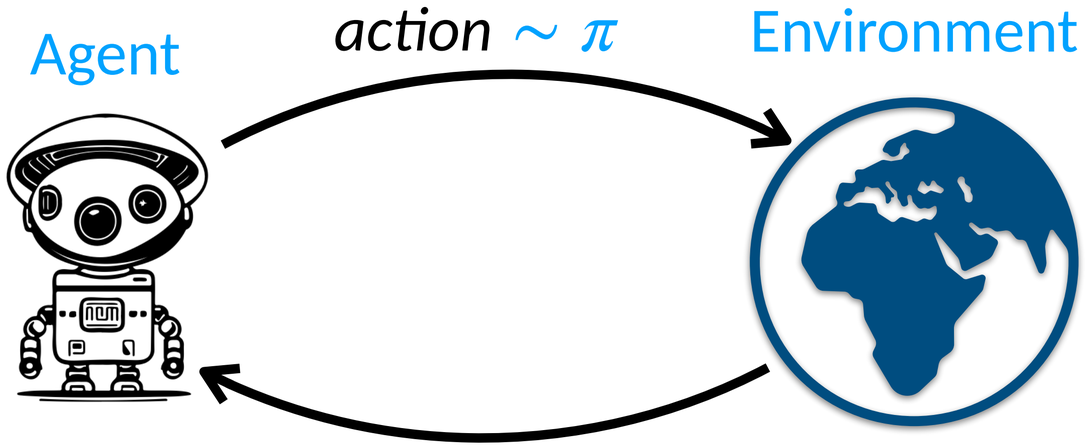
April 2026

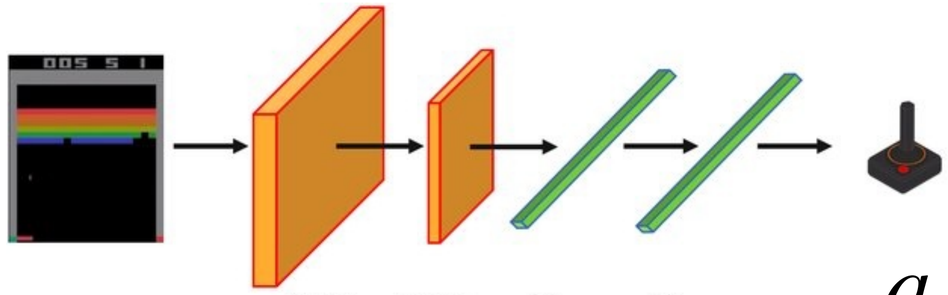
cohere

VUB VRIJE UNIVERSITEIT BRUSSEL

ICLR
International Conference On Learning Representations

Deep Reinforcement Learning



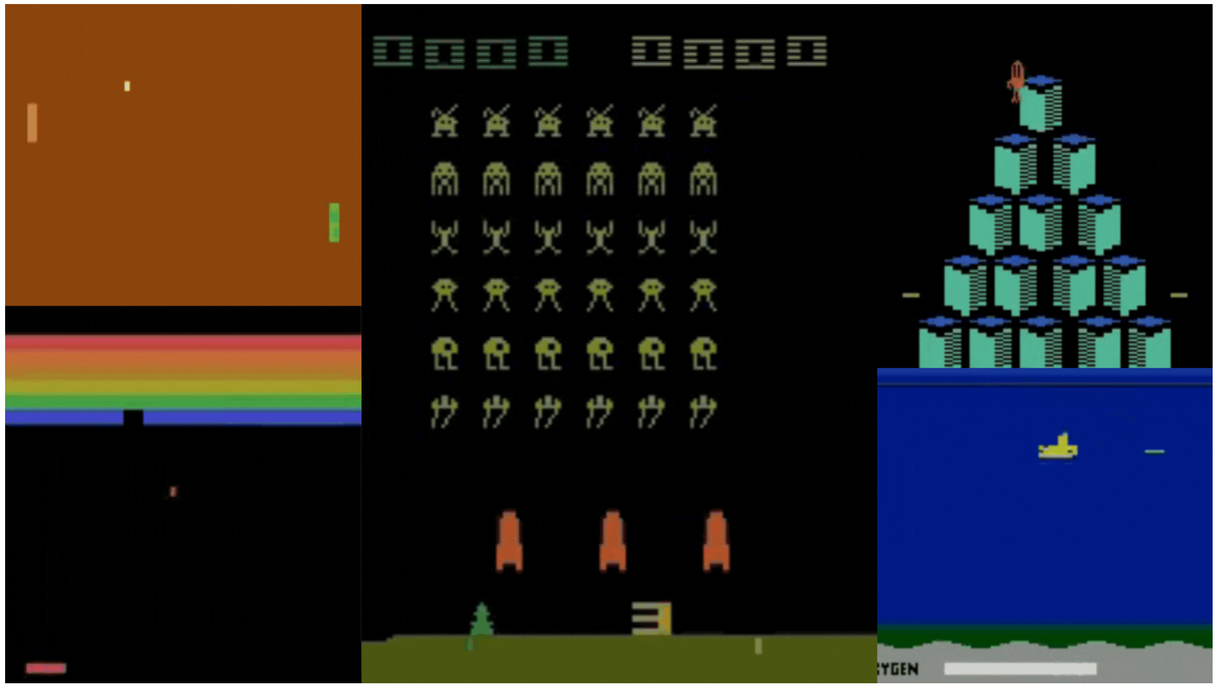


$$\pi := \phi \quad \overline{\pi}$$

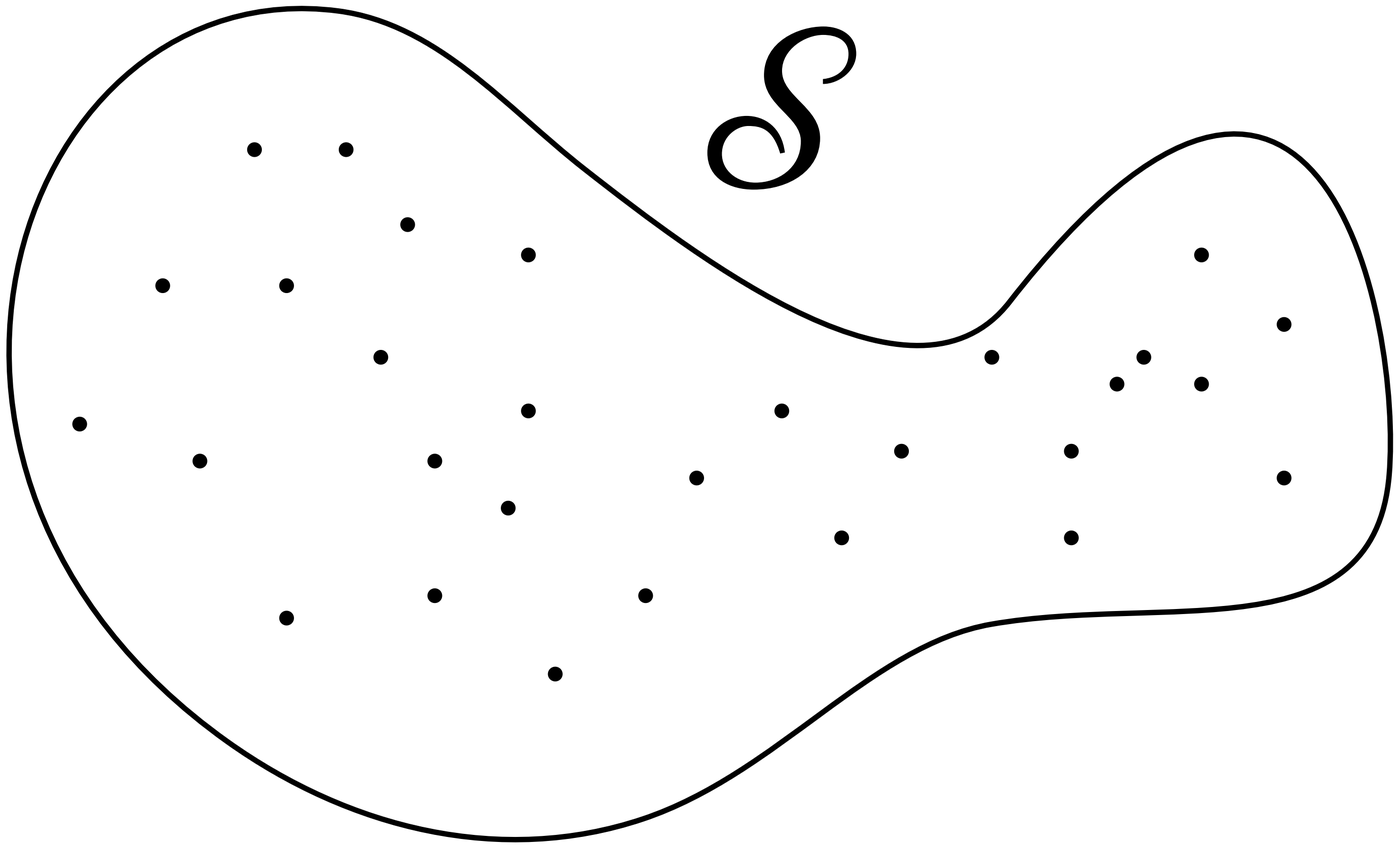
Representation learning

$$\phi : \mathcal{S} \rightarrow \overline{\mathcal{S}}$$

$$a \sim \overline{\pi}(\cdot \mid \phi(s))$$



What is a "Good Representation?"



$$\phi : \mathcal{S} \to \overline{\mathcal{S}}$$

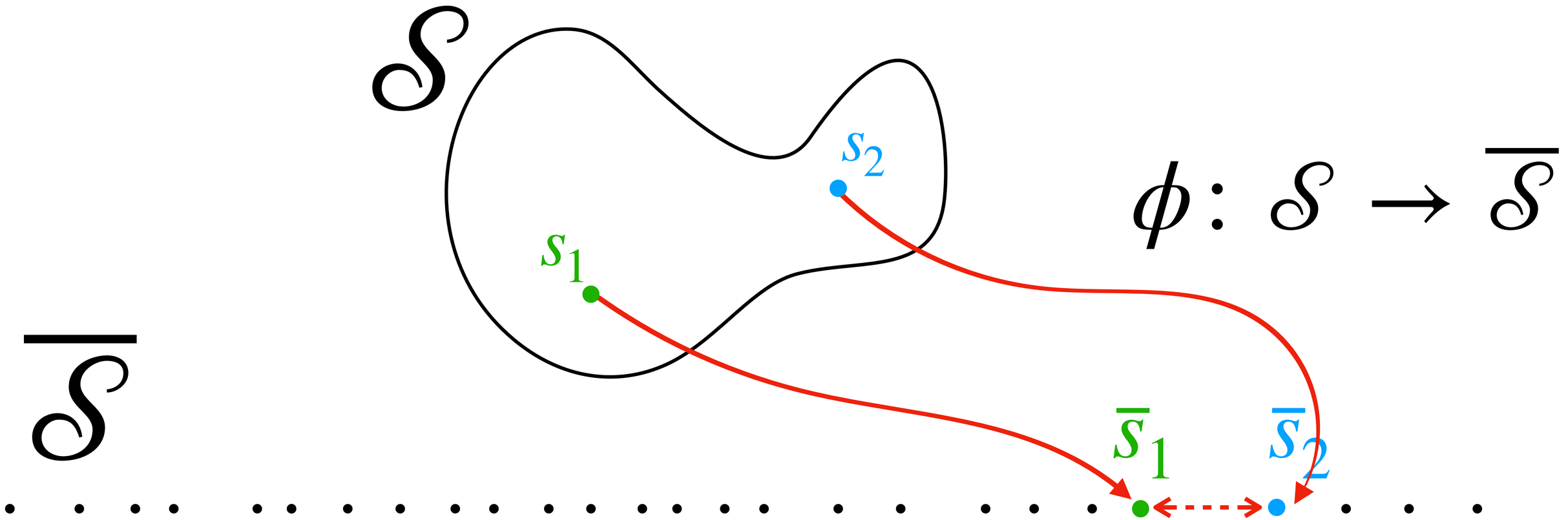
$$a \sim \overline{\pi}(\cdot \mid \phi(s))$$

Our goal is to maximize V^{π}

$\Longrightarrow \phi$ should tell us something about V^{π}

$$V^{\pi}(s) := \mathbb{E}_{\pi}\left[\sum_{t=0}^{\infty} \gamma^{t} \cdot R(s_t, a_t) \mid s_0 = s,\, a_t \sim \pi\left(\cdot \mid s_t \right) \right]$$

What is a "Good Representation?"

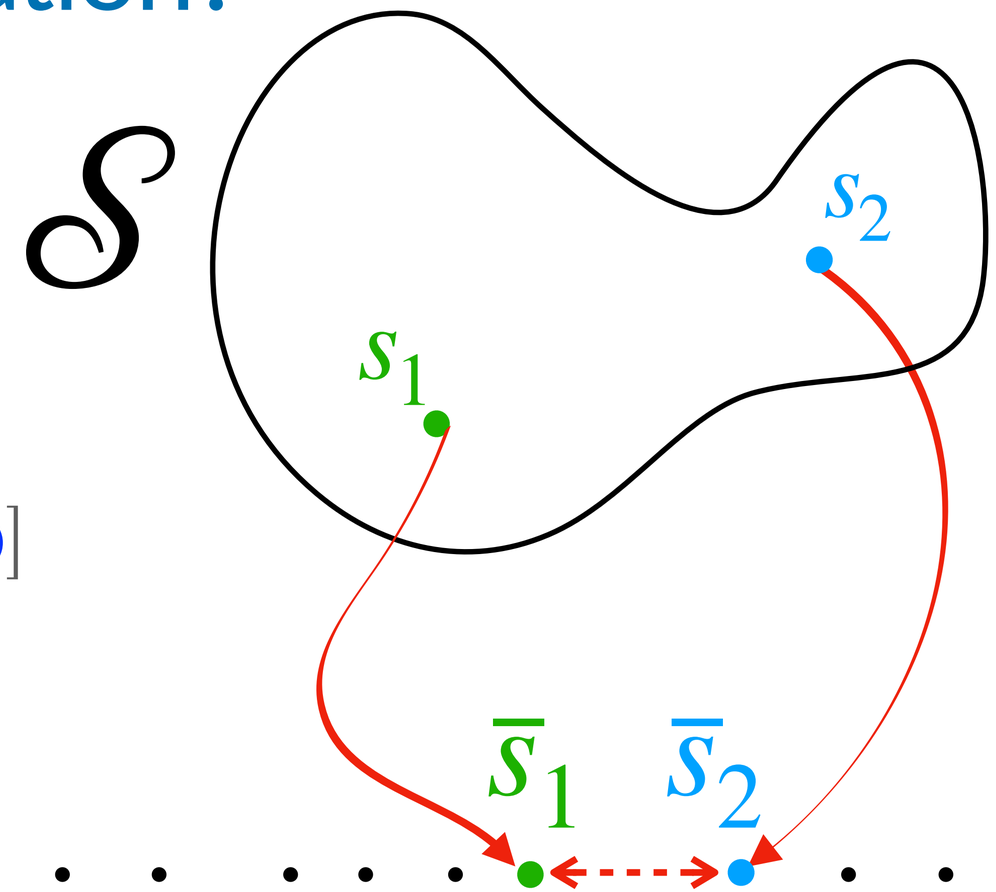


$$\left| V^{\pi}(s_1) - V^{\pi}(s_2) \right| \leq K \cdot \left\| \overline{s}_1 - \overline{s}_2 \right\|$$

What is a "Good Representation?"

$$\text{Loss} = \mathbb{E}_{s,a,r,s' \sim \pi}\left[L_{\text{RL}}(s,a,r,s') + L_{\text{aux. task}}(s,a,r,s')\right]$$

on-policy data

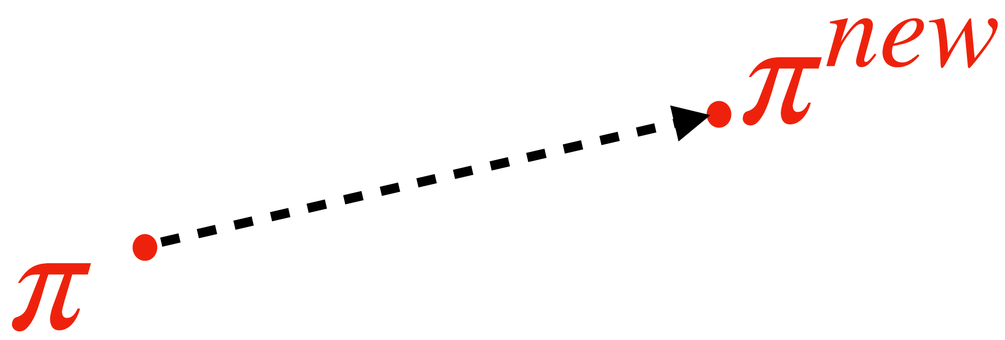


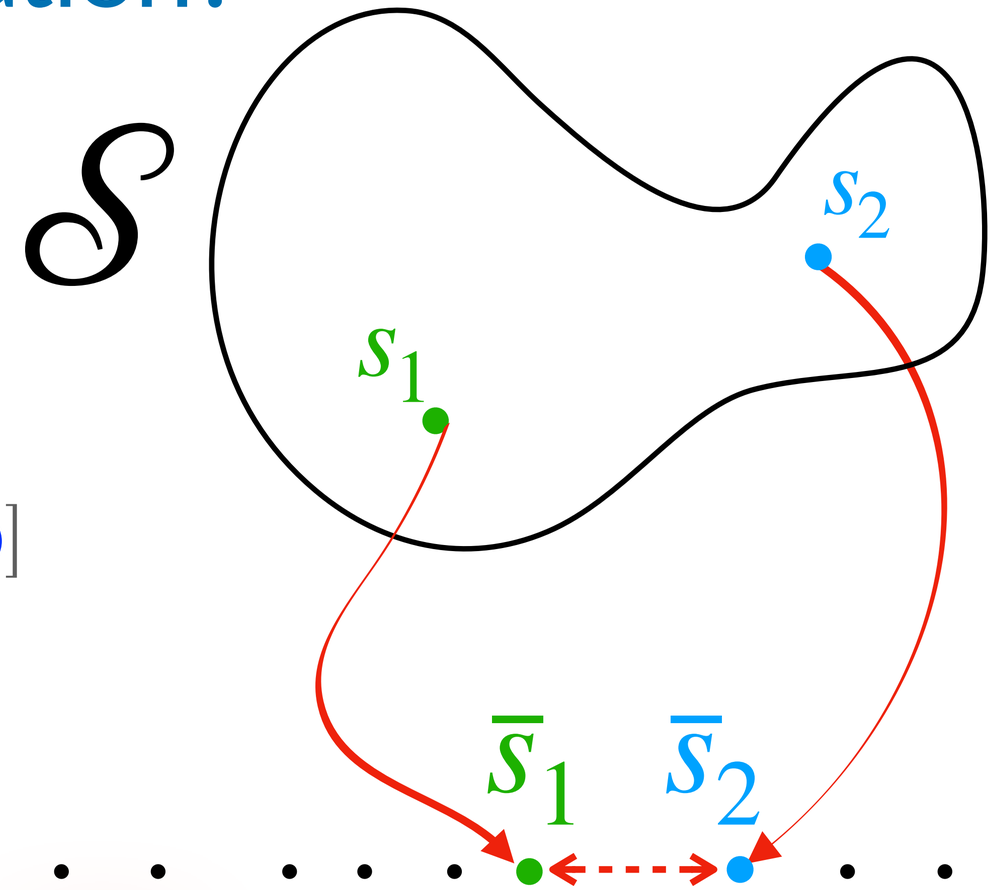
$\overline{\mathcal{S}}$

$$\left| V^{\pi}(s_1) - V^{\pi}(s_2) \right| \leq K \cdot \left\| \bar{s}_1 - \bar{s}_2 \right\|$$

... where $\pi := \bar{\pi} \circ \phi$

What is a "Good Representation?"





$$\text{Loss} = \mathbb{E}_{s,a,r,s' \sim \pi}\left[L_{\text{RL}}(s,a,r,s') + L_{\text{aux. task}}(s,a,r,s')\right]$$

$$\bar{\pi}^{new}, \phi^{new} \leftarrow \text{update}\left(\left(\bar{\pi}, \phi\right), \nabla \text{Loss}\right)$$

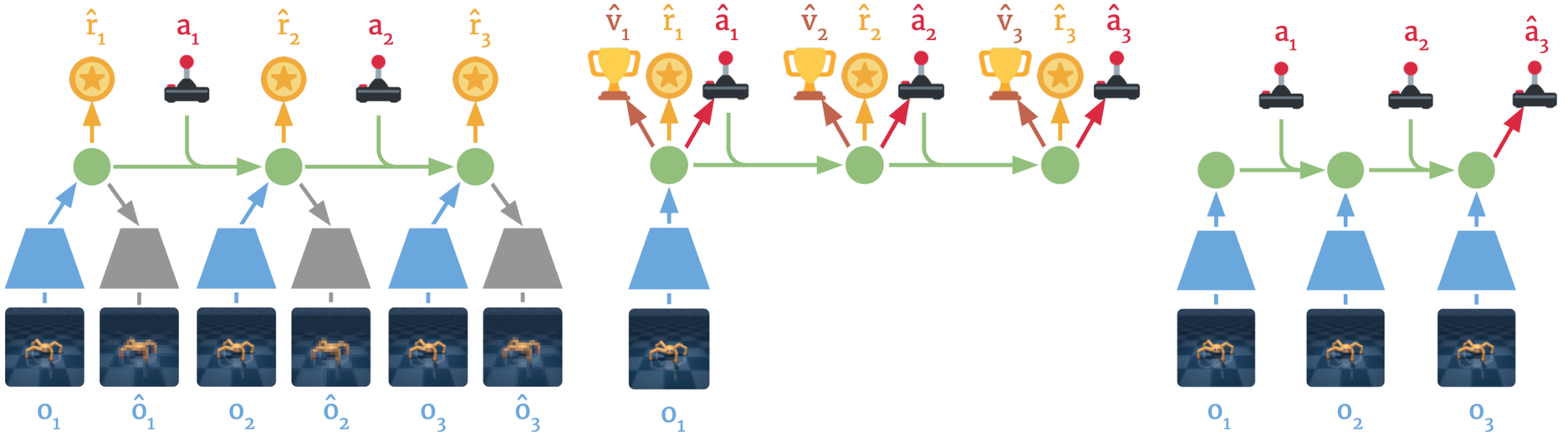
$\overline{\mathcal{S}}$

$$\left| V^{\pi^{new}}(s_1) - V^{\pi^{new}}(s_2) \right| \overset{?}{\leq} K \cdot \left\| \bar{s}_1 - \bar{s}_2 \right\|$$

What about $\pi^{new} := \bar{\pi}^{new} \circ \phi^{new}$?

Model-Based, Deep Reinforcement Learning

World Models



(a) Learn dynamics from experience (b) Learn behavior in imagination (c) Act in the environment

From **Dreamer-v[X]** *by D. Hafner et al,* **v1***: ICLR 2020,* **v2***: ICLR 2021 &* **v3***: Nature 2025,* **v4***: arXiv 2025*

- Sample efficiency
- ***Auxiliary task for representation learning***

- *Deep MDPs — Gelada et. al, ICML 2019*
- *Learning Invariant Representations for Reinforcement Learning without Reconstruction — Zhang et al., ICLR 2020*
- *Data-Efficient Reinforcement Learning with Self-Predictive Representations — Schwarzer el. al, ICLR 2021*

How to leverage world models to…

- learn **theoretically** & **practically useful** representations that ensure **policy improvement**?
- ensure **model calibration** — i.e., reliable planning and learning within the world model to support **policy improvement**?

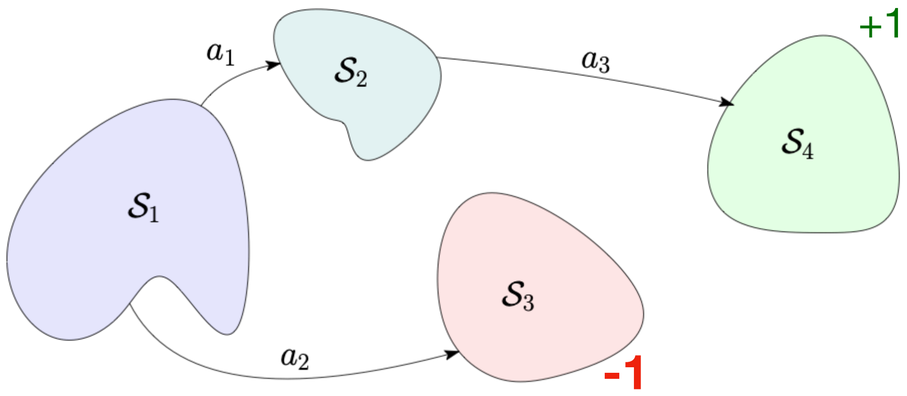
What is a "World Model"?

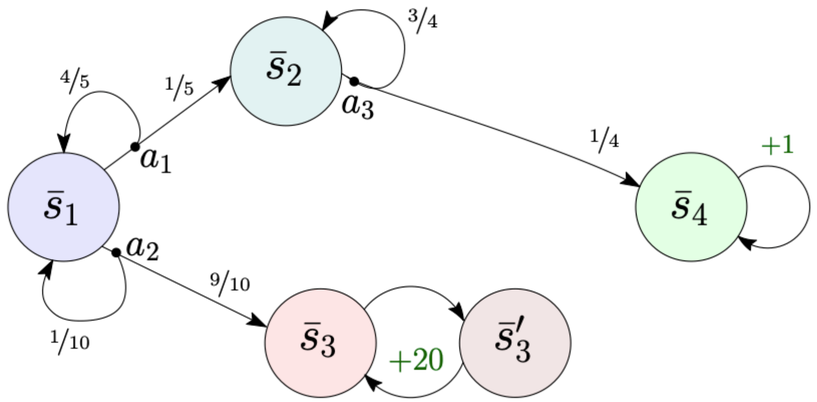
- World models are learned by minimizing reward/transition losses

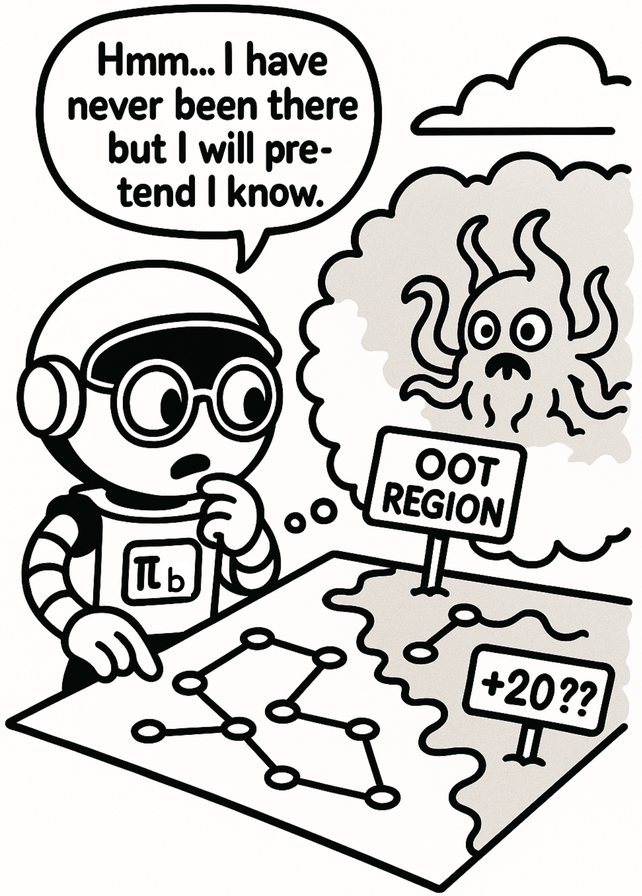
$$L_R := \mathbb{E}_{\eta \sim \mathcal{B}} f_R\left(\phi, \overline{R};\ \eta\right) \qquad L_P := \mathbb{E}_{\eta \sim \mathcal{B}} f_P\left(\phi, \overline{P};\ \eta\right)$$

- η : experiences
- $\mathcal{B}$: (mini-)**batch** of experiences or **replay buffer**
- f_R assigns a "cost" relative to the error between R and $\overline{R}$ w.r.t. experiences η
- f_P assigns a "cost" relative to the error between P and $\overline{P}$ w.r.t. experiences η
- π_b : **baseline/reference policy** used to insert experiences in $\mathcal{B}$

The **out-of-trajectory** (OOT) problem







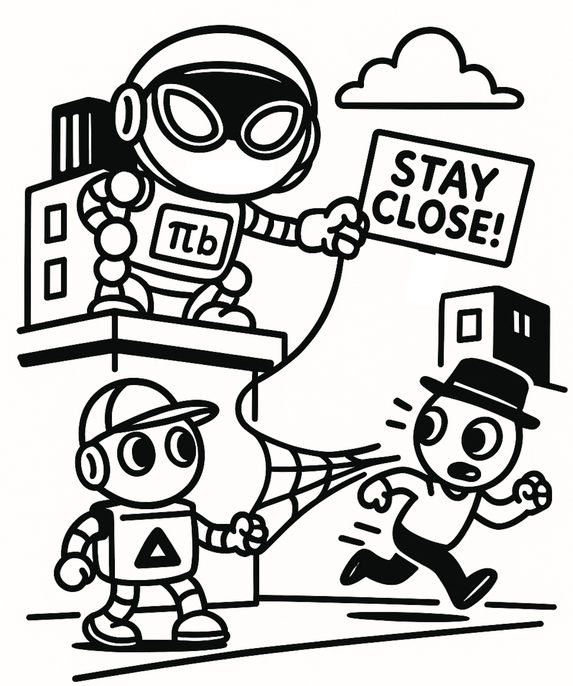
- Train the world model by collecting trajectories from π_b , with $\pi_b(a_2 \mid s) \leq \epsilon \quad \forall s \in \mathcal{S}_1$

➡ L_R, L_P low even if the region $\mathcal{S}_3$ is under-explored, leading to a large error in $\overline{\mathcal{M}}$ (+20)

Your Friendly Neighborhood Policy

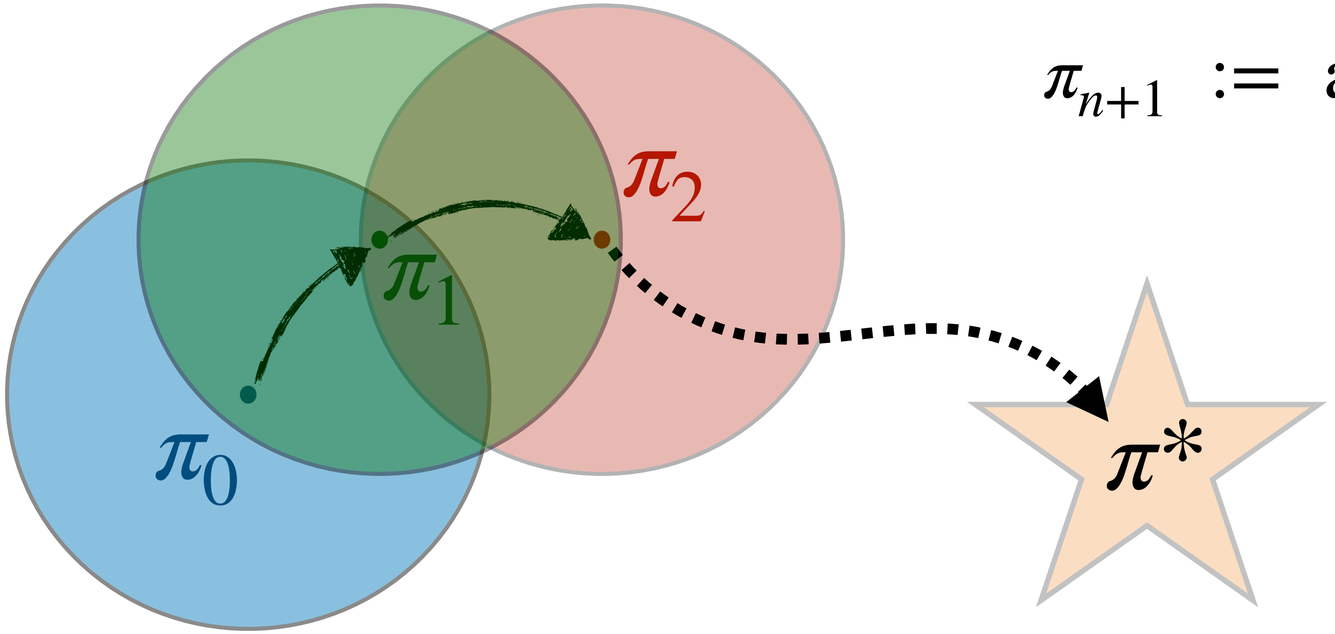
$$\mathcal{N}^C(\pi_b) = \left\{ \pi' \in \Pi \;\middle|\; \begin{array}{l} \forall s \in \mathcal{S}, a \in \mathcal{A}, \quad 2 - C \leq \dfrac{\pi'(a \mid s)}{\pi_b(a \mid s)} \leq C, \\ \\ \operatorname{supp}(\pi_b(\cdot \mid s)) = \operatorname{supp}(\pi'(\cdot \mid s)) \end{array} \right\}$$

$\forall\, 1 < C < 2$



Theorem:

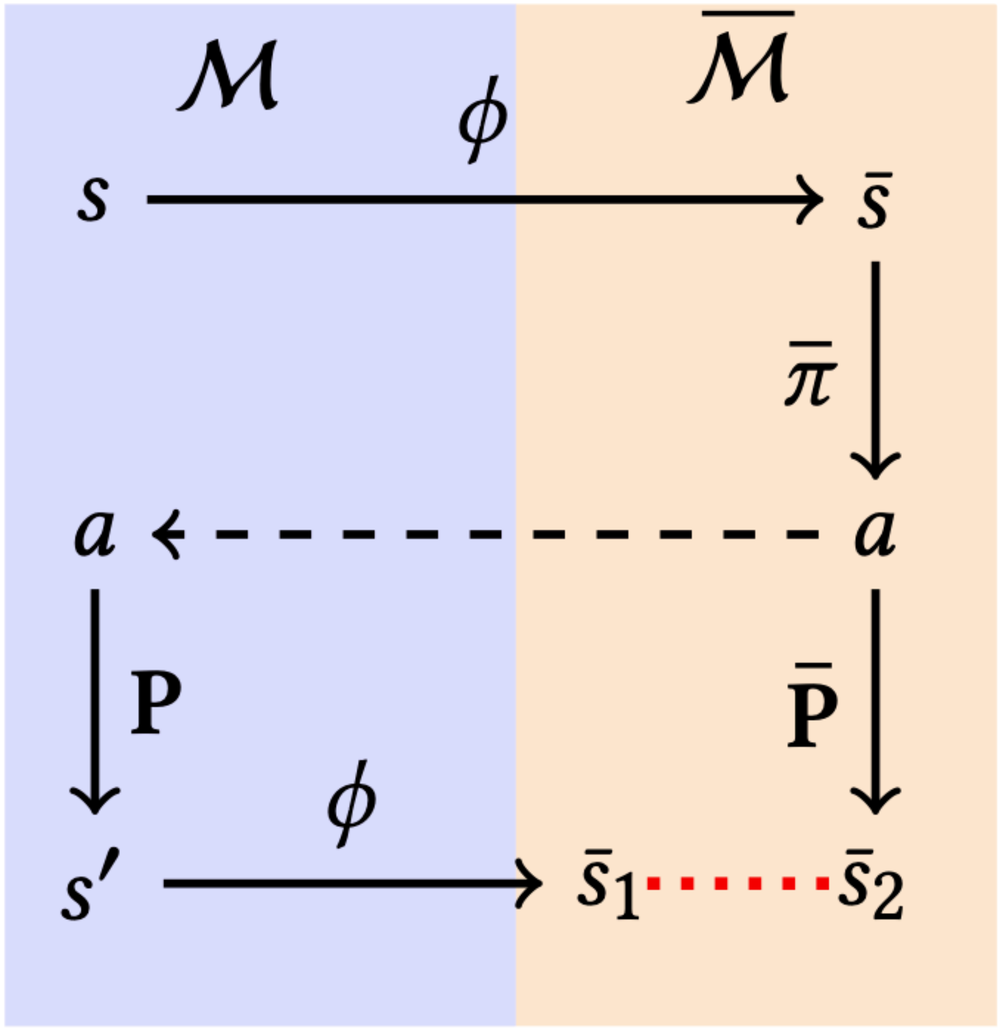
$$\pi_{n+1} := \arg \sup_{\pi' \in \mathcal{N}^C(\pi_n)} \mathbb{E}_{s \sim \mu_{\pi_n}} \mathbb{E}_{a \sim \pi'(\cdot|s)} \left[A^{\pi_n}(s, a) \right]$$



$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ is the ***advantage***;
Then:

- monotonic improvement
- $V^{\pi_n} \xrightarrow[n \to \infty]{} V^*$

Learning a **sound** world model



$$L_R := \mathbb{E}_{s,a,r\sim\pi} \left| r - \overline{R}(\phi(s), a) \right|$$

$$L_P := \mathbb{E}_{s,a\sim\pi} W\Big(\phi_\sharp P(\cdot \mid s, a),\ \overline{P}(\cdot \mid \phi(s), a)\Big)$$

Next state prediction:

$$L_P \le \mathbb{E}_{s,a,s'\sim\pi}\ \mathbb{E}_{\bar{s}'\sim\overline{P}(\cdot \mid \phi(s), a)} \left\| \phi(s') - \bar{s}' \right\|$$

Deep, Safe Policy Improvement

For model-based planning

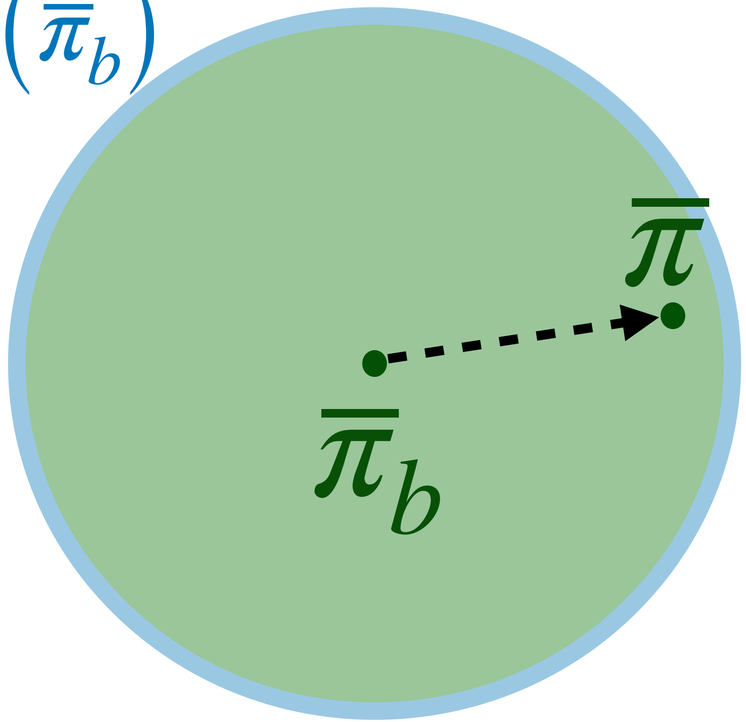
$\mathscr{M}$: real environment

$\overline{\mathscr{M}}$: world model

Theorem: $V^{\bar{\pi}\circ\phi}(\mathscr{M}) - V^{\bar{\pi}_b\circ\phi}(\mathscr{M}) \geq V^{\bar{\pi}}(\overline{\mathscr{M}}) - V^{\bar{\pi}_b}(\overline{\mathscr{M}}) - \zeta$

$\mathscr{N}^{1/\gamma}\left(\bar{\pi}_b\right)$

where $\zeta \propto L_R^{\pi_b}/\gamma + K_V \cdot L_P^{\pi_b}$



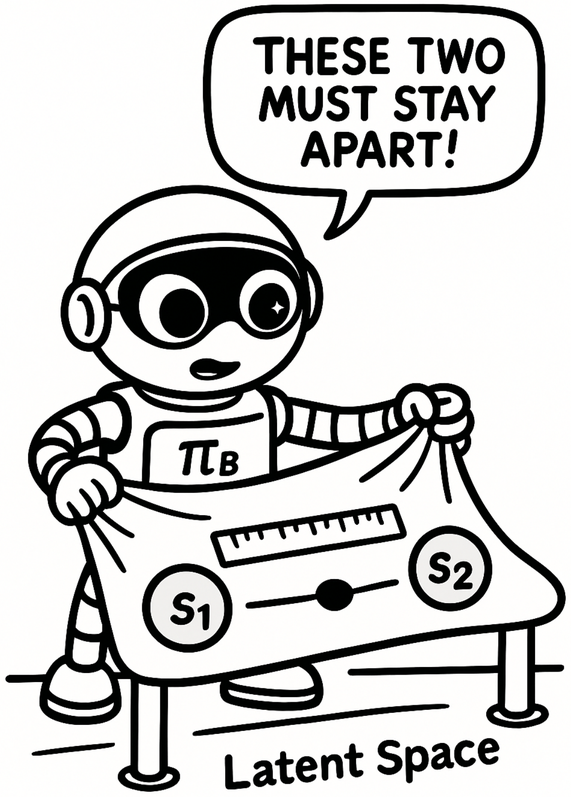
Deep, Safe Policy Improvement for **representation learning**

Theorem: for all policies $\bar{\pi}$ in the neighborhood of π_b,

$$\varepsilon > 0 \text{ and } \delta \propto 1/\epsilon \cdot \left(L_R^{\pi_b} + \gamma K_V \cdot L_P^{\pi_b} \right),$$

we have with probability $1 - \delta$ that:

$$\left| V^{\bar{\pi}}(s_1) - V^{\bar{\pi}}(s_2) \right| \leq K_V \cdot \bar{d}\left(\phi(s_1), \phi(s_2) \right) + \varepsilon$$



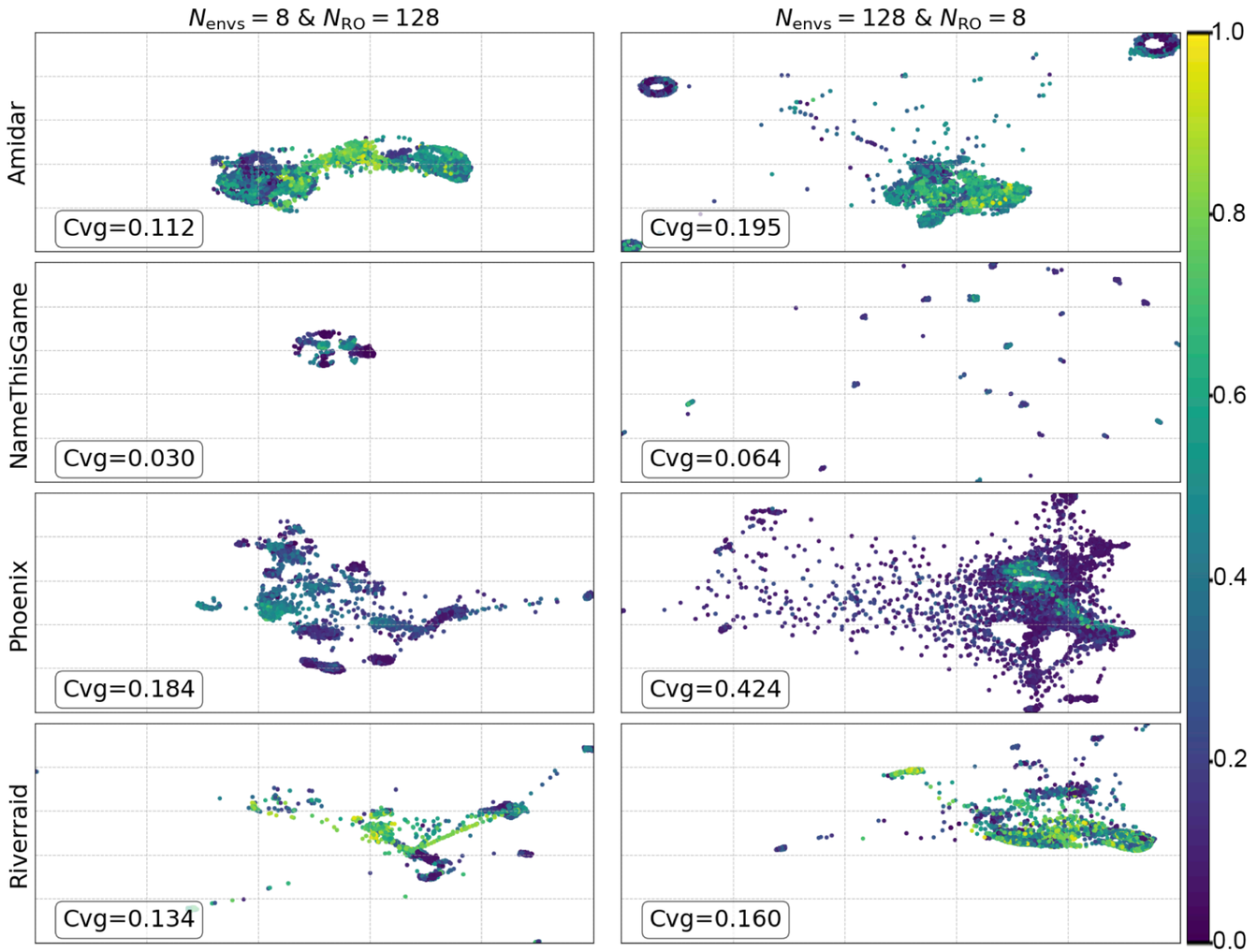
Deep SPI

PPO comes into play!

- Modify **PPO** to incorporate L_P, L_R ***into the clipped objective***, *which defines the expected trust region…*

- … but **PPO** is **on-policy**
  → tight state coverage
  → **not enough data diversity**
  → **difficult to learn a model**

- **Use JAX** (vectorized environment)
  → from 8 envs / 128 rollouts
  to 128 envs / 8 rollouts
  → **good state coverage**



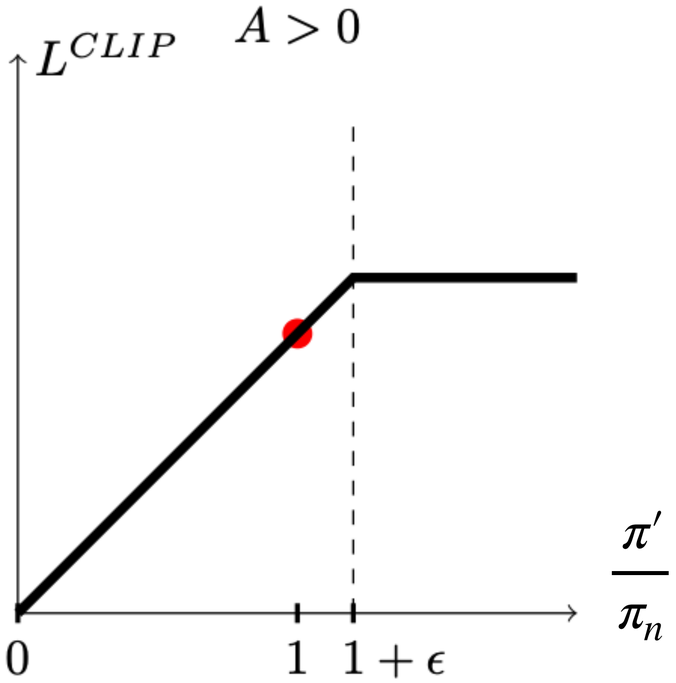
Mayor et al.: **The Impact of On-Policy Parallelized Data Collection on Deep Reinforcement Learning Networks**, ICML 2025

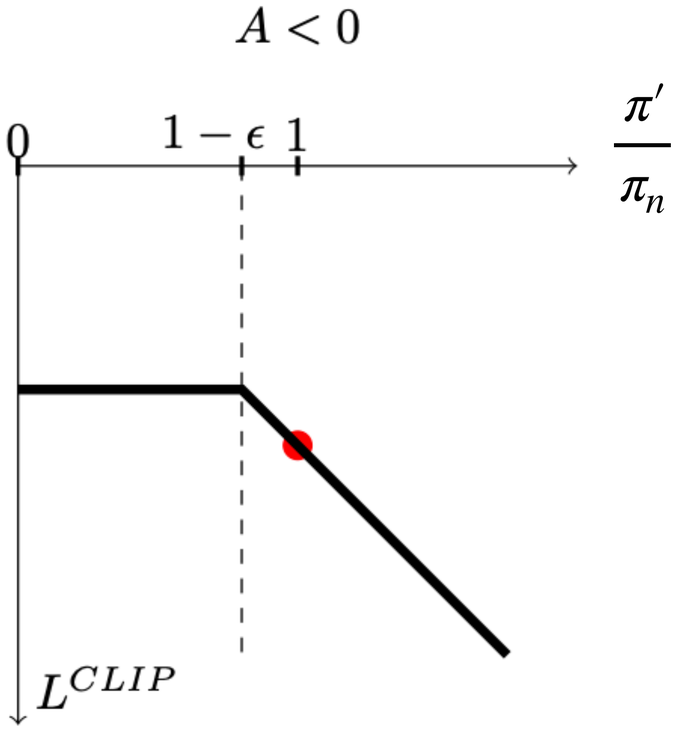
Deep SPI

How to enforce neighborhood constraints? PPO comes into play!

$$\pi_{n+1} := \arg\sup_{\pi' \in \Pi} \mathbb{E}_{s \sim \xi_{\pi_n}} \mathbb{E}_{a \sim \pi_n} \left[\min \left\{ \frac{\pi'(a \mid s)}{\pi_n(a \mid s)} \cdot A^{\pi_n}(s,a), \quad \text{clip}\left(\frac{\pi'(a \mid s)}{\pi_n(a \mid s)}, 1 \pm \epsilon \right) \cdot A^{\pi_n}(s,a) \right\} \right]$$

Schulman et al., 2017





Deep SPI

$$\pi_{n+1} := \arg\sup_{\pi' \in \Pi} \mathbb{E}_{s \sim \xi_{\pi_n}} \mathbb{E}_{a \sim \pi_n} \left[\min \left\{ \frac{\pi'(a \mid s)}{\pi_n(a \mid s)} \cdot A^{\pi_n}(s,a), \quad \operatorname{clip}\left(\frac{\pi'(a \mid s)}{\pi_n(a \mid s)}, 1 \pm \epsilon \right) \cdot A^{\pi_n}(s,a) \right\} \right]$$

$$L_R := \mathbb{E}_{s,a \sim \mathcal{B}} \left| R(s,a) - \overline{R}\left(\phi(s), a\right) \right|$$

$$L_P := \mathbb{E}_{s,a \sim \mathcal{B}} W\left(\phi_\sharp P(\cdot \mid s,a), \overline{P}(\cdot \mid \phi(s), a) \right)$$

- Let's just add L_R, L_P to the loss?
  - ➡ ⚠️ No, we can't!
  - ➡ L_R, L_P update ϕ and might push $\pi_{new} := \left(\overline{\pi}_{new} \circ \phi\right)$ outside $\mathcal{N}^C$
  - ➡ Solution: incorporate L_R, L_P as auxiliary objectives **to the advantage**!

Deep SPI

$$\pi_{n+1} := \arg\sup_{\pi' \in \Pi} \mathbb{E}_{s \sim \xi_{\pi_n}} \mathbb{E}_{a \sim \pi_n} \left[\min \left\{ \frac{\pi'(a \mid s)}{\pi_n(a \mid s)} \cdot U^{\pi_n}(s,a), \quad \mathrm{clip}\left(\frac{\pi'(a \mid s)}{\pi_n(a \mid s)}, 1 \pm \epsilon \right) \cdot U^{\pi_n}(s,a) \right\} \right]$$

$$U^{\pi_n}(s,a,s') := A^{\pi_n}(s,a) - \alpha_R \cdot \ell_R(s,a) - \alpha_P \cdot \ell_P(s,a,s')$$

$$\ell_R(s,a) := \left| R(s,a) - \overline{R}\left(\phi(s), a\right) \right| \qquad \ell_P(s,a,s') := \mathbb{E}_{\bar{s}' \sim \overline{P}(\,\cdot \mid \phi(s), a)} \, \bar{d}\left(\phi(s'), \bar{s}'\right)$$

- ℓ_R, ℓ_P allow retrieving L_R, L_P in expectation!
- **restrict both the policy and the representation updates**!

Deep SPI (57 stochastic environments! IQM, Median, Mean: higher is better; Optimality Gap: lower is better)