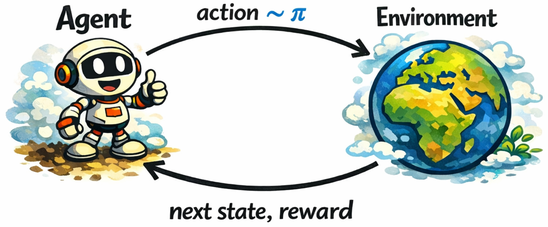
Deep SPI: Safe Policy Improvement via World Models

Florent Delgrange, Raphaël Avalos, Willem Röpke

VUB VRIJE UNIVERSITEIT BRUSSEL · fwo · cohere

REINFORCEMENT LEARNING



Goal: learn a policy π that optimizes the **value** function

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

REPRESENTATION LEARNING

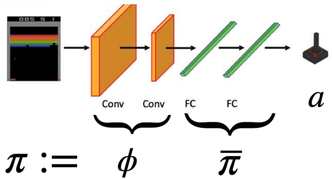
Goal: learn an encoder

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

to a **tractable latent space** from which values/policies can be approximated

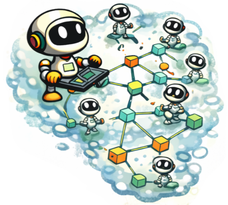
→ Learn a **latent policy:**

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

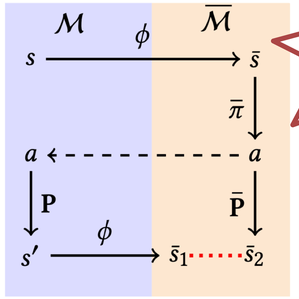


$$\pi := \overline{\pi} \circ \phi$$

WORLD MODELS



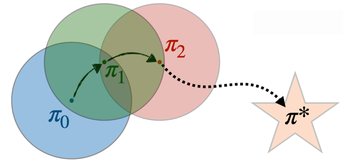
Latent predictive losses



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

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

DEEP, SAFE POLICY IMPROVEMENT!



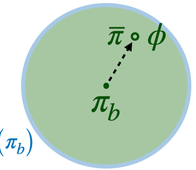
$$\mathcal{N}^C(\pi_b) = \left\{ \pi' \in \Pi \;\middle|\; \forall s \in \mathcal{S}, a \in \mathcal{A},\quad 2 - C \le \frac{\pi'(a \mid s)}{\pi_b(a \mid s)} \le C, \right\}$$

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

$$A^{\pi}(s,a) = Q^{\pi}(s,a) - V^{\pi}(s)$$

➡ $V^{\pi_n} \xrightarrow[n \to \infty]{} V^{*}$ *convergence guarantees*

RESTRICT POLICY + REPRESENTATION UPDATES!



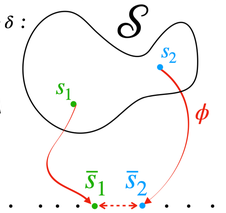
Latent planning → improvement

$$V^{\overline{\pi} \circ \phi}(\mathcal{M}) - V^{\overline{\pi}_b \circ \phi}(\mathcal{M}) \ge V^{\overline{\pi}}(\overline{\mathcal{M}}) - V^{\overline{\pi}_b}(\overline{\mathcal{M}}) - \zeta$$

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

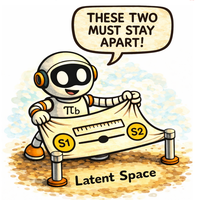
$$\forall \epsilon > 0 \text{ and } \delta \propto 1/\epsilon \cdot \left(L_R^{\pi_b} + \gamma K_V \cdot L_P^{\pi_b}\right),$$

with proba ≥ 1 - δ :



$$\left| V^{\pi^{new}}(s_1) - V^{\pi^{new}}(s_2) \right| \le K_V \cdot \left\| \bar{s}_1 - \bar{s}_2 \right\| + \epsilon$$

where $\pi^{new} := \overline{\pi} \circ \phi$



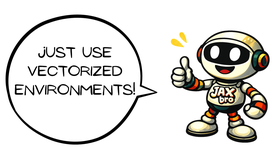
DEEP-SPI – THE ALGORITHM

How to enforce the neighborhood constraint in practice? Modify PPO to incorporate the world model losses into the clipping objective:

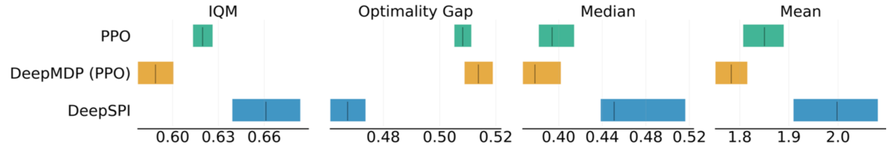
$$\pi_{n+1} := \arg\sup_{\pi' \in \Pi} \mathbb{E}_{s \sim \xi_{\pi_n}}\, \mathbb{E}_{a \sim \pi_n}\left[\min\left\{ \frac{\pi'(a \mid s)}{\pi_n(a \mid s)} \cdot U^{\pi_n}(s,a),\quad \operatorname{clip}\left(\frac{\pi'(a \mid s)}{\pi_n(a \mid s)}, 1 \pm \epsilon \right) \cdot U^{\pi_n}(s,a) \right\} \right]$$

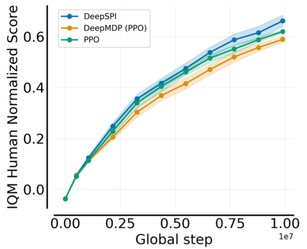
take $\epsilon = C - 1$

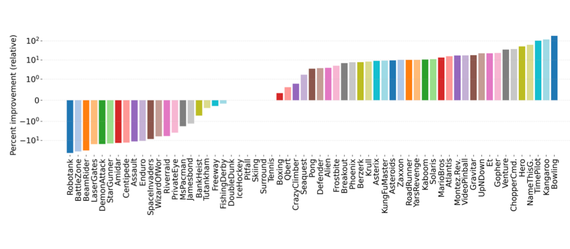
$$U^{\pi_n}(s,a) := A^{\pi_n}(s,a) - \alpha_R \cdot L_R - \alpha_P \cdot L_P$$



EVALUATION ON STOCHASTIC ATARI-57







in/florent-delgrange · @f_delgrange · https://delgrange.me/