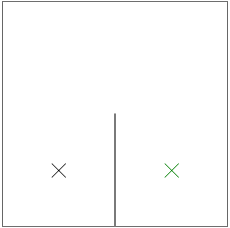
Exploration by Learning Diverse Skills through Successor State Representations

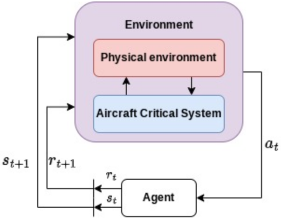
Paul-Antoine Le Tolguenec[1,2,3], Yann Besse[1], Florent Teichteil-Koenigsbuch[1], Dennis Wilson[2,3], Emmanuel Rachelson[2,3]

[1]Airbus, [2]ISAE-Supaero, [3]Université de Toulouse

Exploration in Reinforcement Learning

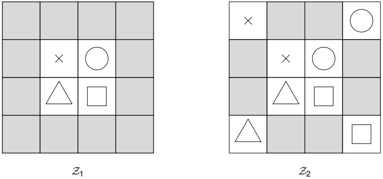


Sparse reward mazes.



Test critical systems using RL.

Diversity using Mutual Information



State distributions showing two sets of four skills on a grid maze. Each skill's visited states are represented by a unique symbol.

Maximizing **mutual information** (MI) $\mathcal{I}(S, Z)$ between state S and skill descriptor Z:

$$\begin{aligned}\mathcal{I}(S, Z) &\triangleq D_{KL}(\mathbb{P}(S, Z)||\mathbb{P}(S)\mathbb{P}(Z)),\\ &= \mathcal{H}(S) - \mathcal{H}(S|Z).\end{aligned}$$

MI is not a perfect measure of exploration:

$$\mathcal{I}_1(S, Z) = \mathcal{H}_1(S) - \mathcal{H}_1(S|Z) = \log 4 \quad \text{and} \quad \mathcal{I}_2(S, Z) = \mathcal{H}_2(S) - \mathcal{H}_2(S|Z) = \log 4$$

How to enforce **exploration**?

Promoting diversity with SSR

The Successor State Representation (SSR) estimators provide a way to estimate conditional probability densities between S and Z: $p(s|s_1, \theta, z)$ represents the **state occupancy measure** of the policy $\pi_\theta(s, z)$, starting from s_1:

$$p(s_2|s_1, \theta, z) = \sum_{t=0}^{\infty} \gamma^t p\left(s_t = s_2 \;\middle|\; \begin{array}{l} s_0 = s_1, \\ a_t \sim \pi_\theta(s_t, z) \end{array}\right).$$

We lower bound MI with SSR estimation ($m(s_1, s_2, z) = p(s_2|s_1, \theta, z)/p(s_2)$):

$$\mathcal{I}(S, Z) \geq \mathop{\mathbb{E}}_{\substack{z \sim p(z) \\ s_1 \sim p(s|z) \\ s_2 \sim p(s|z) \\ a \sim \pi(\cdot|s_1)}} \left[\log\left(\frac{m(s_1, a, s_2, z)}{1 + \sum_{z' \in \mathcal{Z}} m(s_1, a, s_2, z')}\right)\right].$$

Exploring through a novelty measure

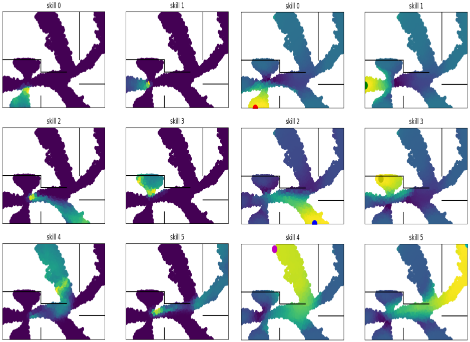
To cope with MI limitations for exploration purposes we use a measure of novelty:

$$u_t(s,z) = \underbrace{\log\left(\frac{m_t(s_0,s,z_i)}{\sum_{k=1}^{t-1}\sum_{z'} m_k(s_0,s,z')}\right)}_{\text{Explore under-visited areas}} + \underbrace{\sum_{z'\neq z}\log\left(\frac{m_t(s^z_{t-1},s,z)}{m_t(s^{z'}_{t-1},s,z')}\right) + \log\left(\frac{m_t(s_0,s,z)}{m_t(s_0,s,z')}\right)}_{\text{Repulsion between skills}}.$$

m_k represents the SSR learned at epoch k of the algorithm. $m_{i\in(1,t-1)}$ are the past SSR learned for the set of skills since the beginning of exploration.

$$\mathcal{G}(\theta) = \mathop{\mathbb{E}}_{\substack{z\sim p(z)\\ s_1\sim p(s|z)\\ a_z\sim\pi_\theta(\cdot|s_1,z)\\ \color{red}{s_2\sim\delta(s|z)}}}\left[\log\left(\frac{m(s_1,a_z,s_2,z)}{1+\sum_{z'\in\mathcal{Z}} m(s_1,a_{z'},s_2,z')}\right)\right].$$

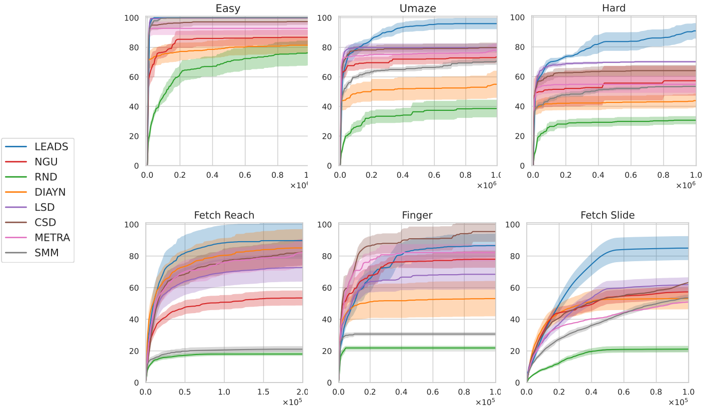
Occupancy measure & Novelty measure estimation



(a) Measure (b) Uncertainty

(a): The SSR $m(s_0, s, z)$. (b): The uncertainty measure $u(s, z)$, per skill, with the maximum state highlighted.

Quantitative evaluation of the coverage



Relative coverage evolution across six tasks. The x-axis represents the number of samples collected since the algorithm began.

Conclusion & Takeaways

- **Challenge:** Addressed exploration in RL by encouraging skill diversity.
- **Key Insight:** Mutual information alone may be insufficient; introduced an novelty-based objective for better coverage.
- **LEADS Algorithm:** Uses SSR estimators to adapt skills to under-explored states while maintaining distinct state distributions.
- **Results:** Achieved superior state coverage in most tested environments over existing methods.
- **Outlook:** The core idea is to explore by moving away from past experiences. This approach could also be achieved with simpler density estimators better suited to the problem.

Thanks for your attention!