Improving Intrinsic Exploration by Creating Stationary Objectives

Roger Creus Castanyer

Joshua Romoff

Glen Berseth

Mila

Université de Montréal

UBISOFT MONTRÉAL

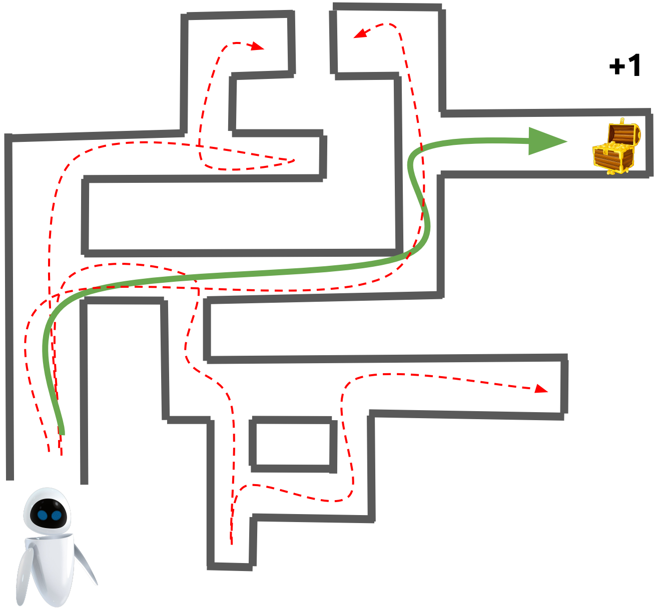
ICLR 2024

Intrinsic Objectives improve Exploration

In hard-exploration problems, exploration is more successful if **directed**, **controlled**, and **efficient**.

Can be achieved by **augmenting the original RL problem**

$$y_t = \mathcal{R}_t + \lambda \mathcal{B}(s_t, a_t) + \gamma \max_{a'} Q_{\theta'}(s_{t+1}, a')$$



Many Intrinsic Objectives for Exploration

Count-based bonuses

Compute the **state-visitation frequencies**

$$\mathcal{B}(s_t, a_t, s_{t+1}|\phi_t) = \frac{\beta}{\sqrt{\mathcal{N}_t(s_{t+1})}}$$

Pseudo-counts [1]

Estimate the state-visitation frequencies

$$\mathcal{B}(s_t, a_t, s_{t+1}|\phi_t) = \psi_t(s_{t+1})^T C_t^{-1} \psi_t(s_{t+1})$$

$$C_t = \sum_{t=0}^{T} \psi_t(s_t)\psi_t(s_t)^T$$

State-Entropy Maximization [2]

Induce a MaxEnt distribution over the **state-visitation distribution**

$$\sum_{t=0}^{T} \mathcal{H}(s_t) = \sum_{t=0}^{T} -\mathbf{E}_{s_t \sim d^{\pi_\theta}(s_t)} [\log d^{\pi_\theta}(s_t)] \leq \sum_{t=0}^{T} \mathbf{E}_{s_t \sim d^{\pi_\theta}(s_t)} [-\log p_{\phi_{t-1}}(s_t)]$$

$$\mathcal{B}(s_t, a_t, s_{t+1}|\phi_t) = -\log p_{\phi_t}(s_{t+1})$$

→ We introduce the parameters ϕ_t to understand the properties of the intrinsic reward distributions

→ If ϕ_t changes over time, then $\mathcal{B}(s_t, a_t, s_{t+1}|\phi_t)$ **is non-stationary**.

→ Non-stationary rewards transform an MDP into a POMDP:
- Require non-Markovian properties (e.g. memory)
- No convergence guarantees

[1] Henaff, Mikael, et al. "Exploration via elliptical episodic bonuses." Advances in Neural Information Processing Systems 35 (2022): 37631-37646.
[2] Berseth, Glen, et al. "SMiRL: Surprise minimizing RL in dynamic environments." *arXiv preprint arXiv:1912.05510* (2019).

SOFE: Stationary Objectives For Exploration

Non-stationary rewards become a deterministic function of the **augmented states**

The original exploration objective **remains the same**

Identification

Counts

$$\phi_t \leftarrow \mathcal{N}_t$$

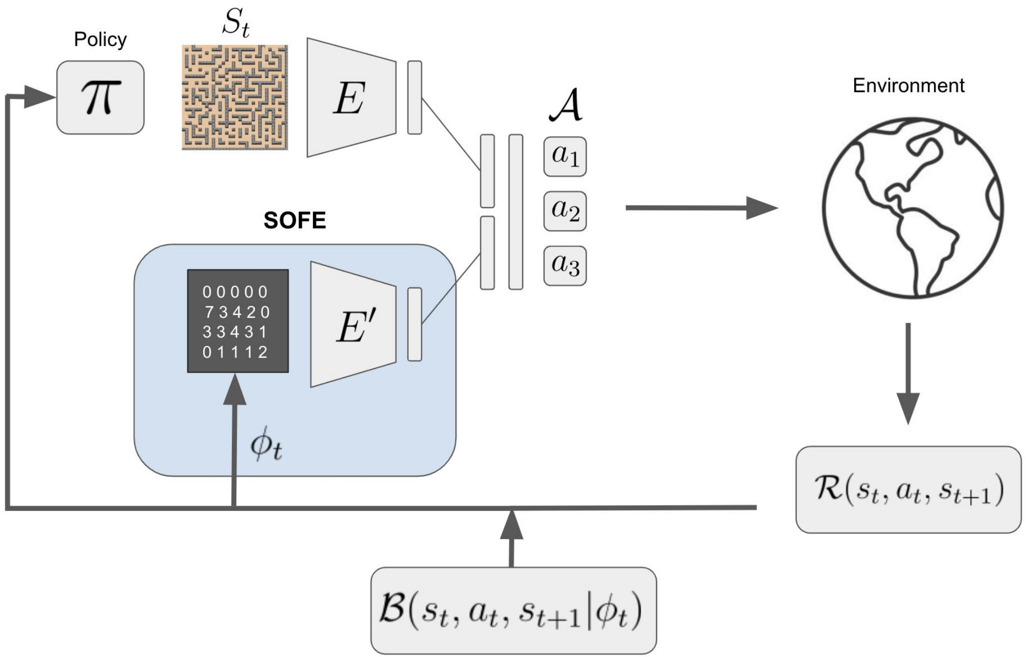
E3B

$$\phi_t \leftarrow \mathcal{C}_t$$

S-Max

$$\phi_t \leftarrow (\mu_t, \sigma_t)$$

Augmentation

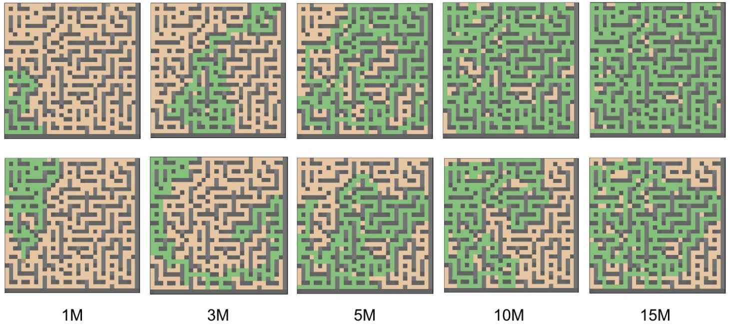


SOFE: Results

SOFE allows the agents **better explore** the state space

SOFE + Counts

Counts

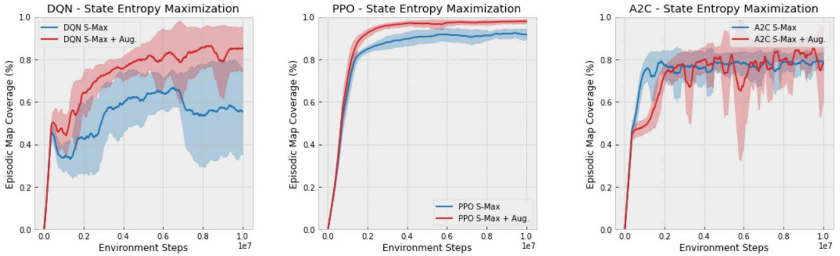


SOFE allows the agents **better solve** sparse-reward tasks

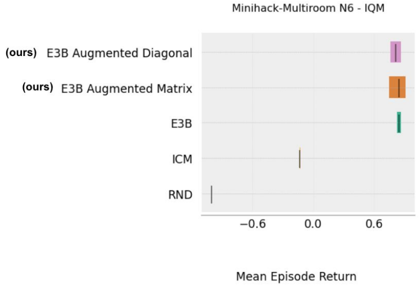
Algorithm	DeepSea 10	DeepSea 14	DeepSea 20	DeepSea 24	DeepSea 30
DeRL-A2C	0.98 ± 0.10	0.65 ± 0.23	0.42 ± 0.16	0.07 ± 0.10	0.09 ± 0.08
DeRL-PPO	0.61 ± 0.20	0.92 ± 0.18	-0.01 ± 0.01	0.63 ± 0.27	-0.01 ± 0.01
DeRL-DQN	$\mathbf{0.98 \pm 0.09}$	$\mathbf{0.95 \pm 0.17}$	0.40 ± 0.08	0.53 ± 0.27	0.10 ± 0.10
SOFE-A2C	0.94 ± 0.19	0.45 ± 0.31	0.11 ± 0.25	0.08 ± 0.14	0.04 ± 0.09
SOFE-PPO	0.77 ± 0.29	0.67 ± 0.33	0.13 ± 0.09	0.07 ± 0.15	0.09 ± 0.23
SOFE-DQN	0.97 ± 0.29	0.78 ± 0.21	$\mathbf{0.70 \pm 0.28}$	$\mathbf{0.65 \pm 0.26}$	$\mathbf{0.42 \pm 0.33}$

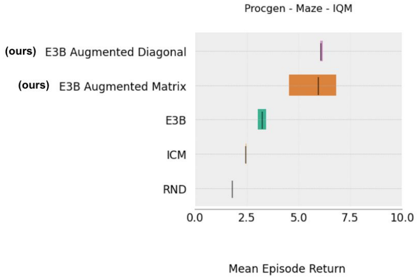
SOFE provides **orthogonal gains** to several exploration objectives

State-Entropy Maximization



Pseudo-counts





SOFE: Conclusion

Make your agent's life easier with **SOFE**