ANITI

Exploration Methods for Reinforcement Learning Applied to Critical System Testing

Méthodes d'exploration pour l'apprentissage par renforcement appliqué au test de systèmes critiques

Paul-Antoine Le Tolguenec

Director: Emmanuel Rachelson
Co-supervised by: Dennis G. Wilson, Florent Teichteil-Koenigsbuch and Yann Besse

isae Institut Supérieur de l'Aéronautique et de l'Espace SUPAERO

AIRBUS

Critical Software System

Definition

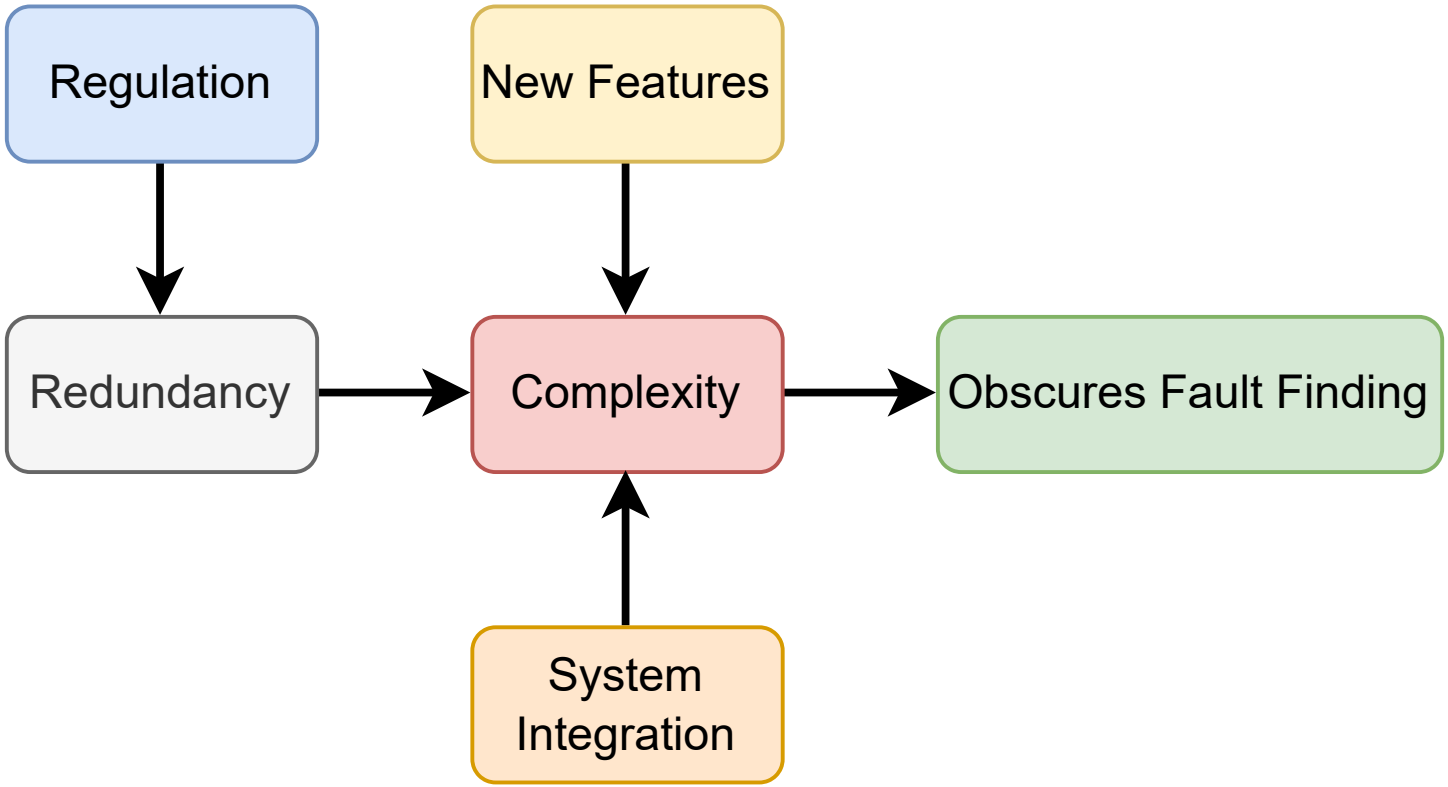
Critical system × failure ⇒ environmental disaster, significant financial loss, or loss of human life.

Ariane 5 (1996)

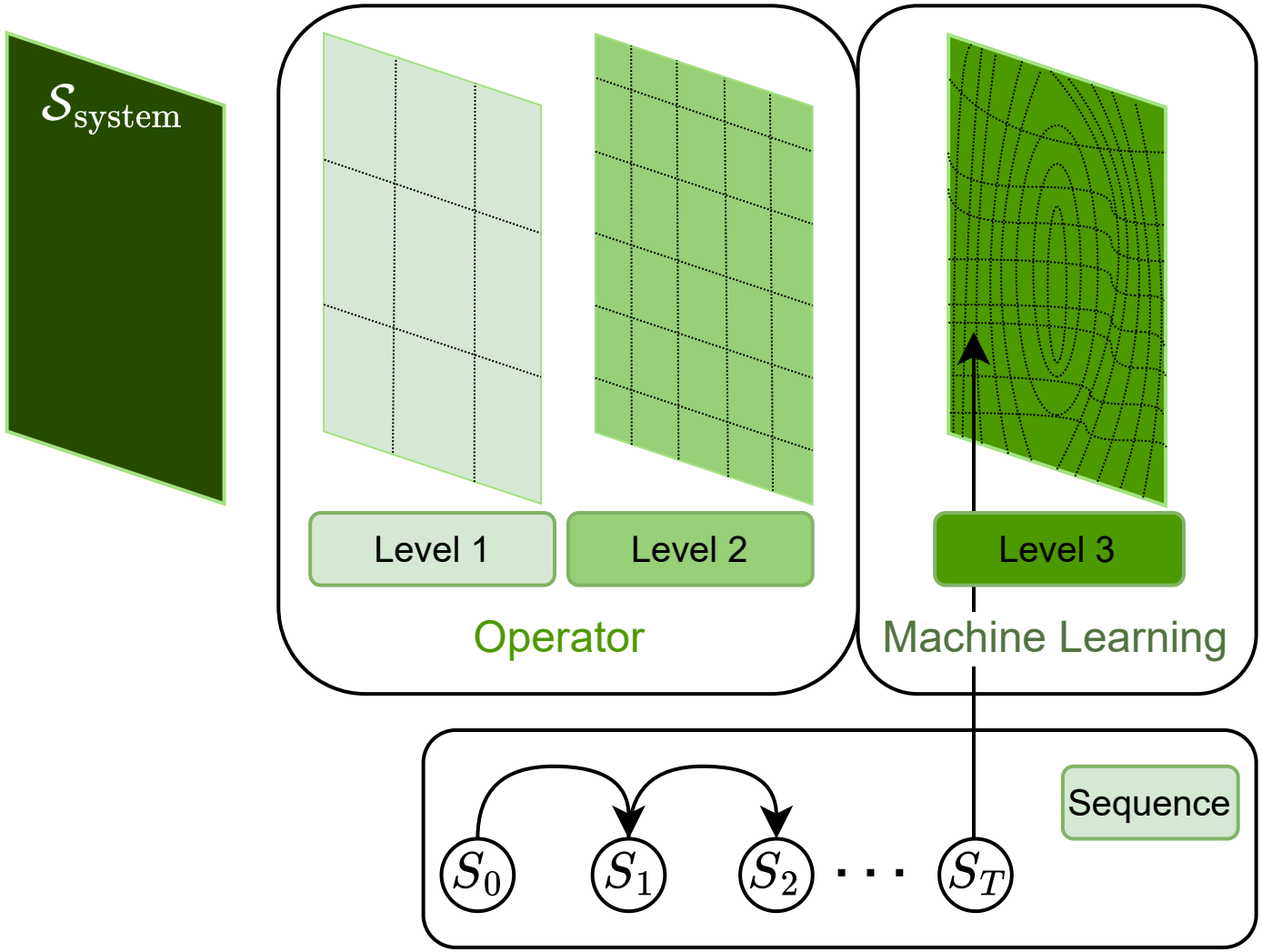
Moorgate Train Crash (1975)

Regulatory constraints for aeronautic systems

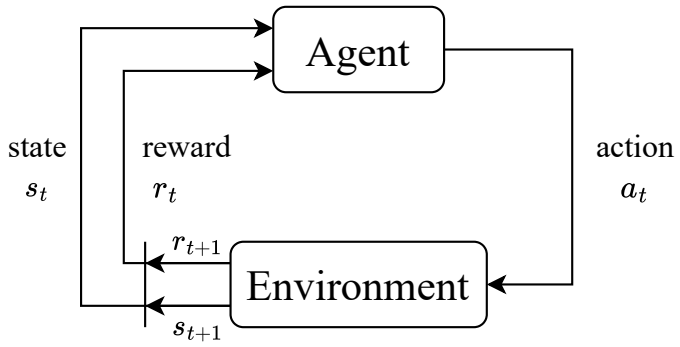
FAR/JAR $\Rightarrow p_{\text{catastrophic failure}} < 10^{-9}$ per flight hour.



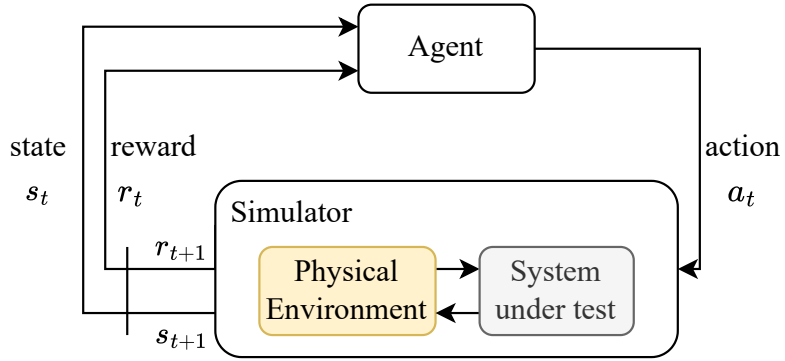
Testing avionic systems



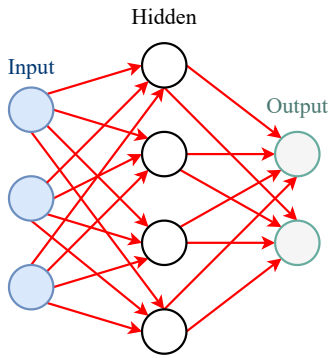
Reinforcement Learning (RL)

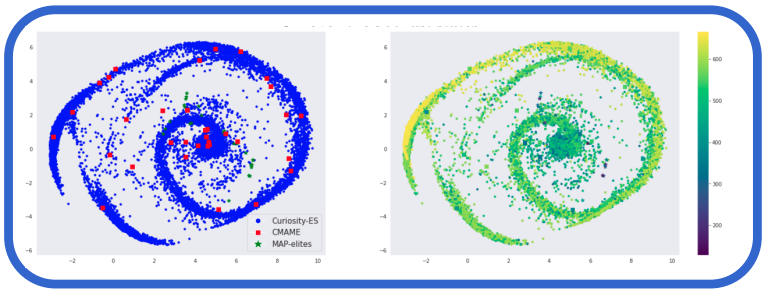


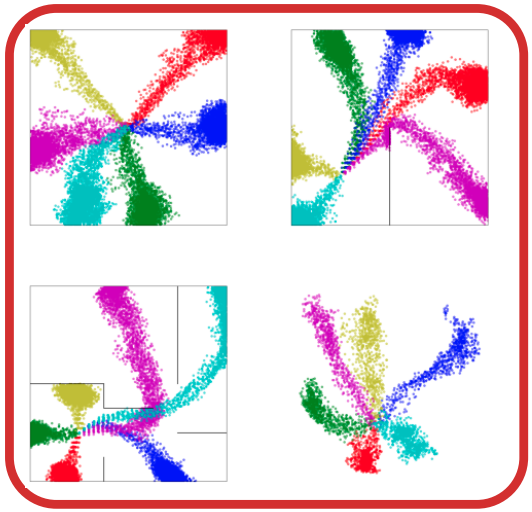
Adaptive Stress Testing (AST)



Policy Representation: $\pi_\theta(s)$





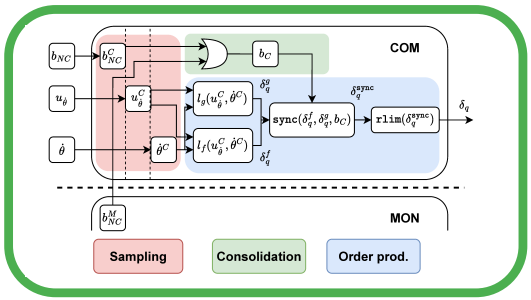


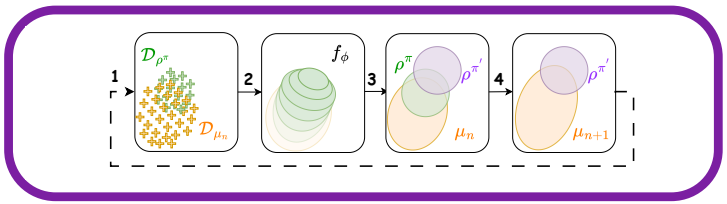
Curiosity-ES

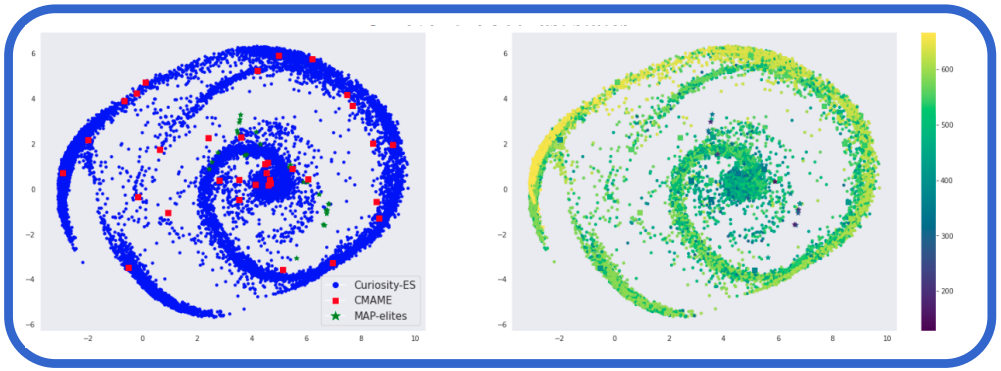
LEADS

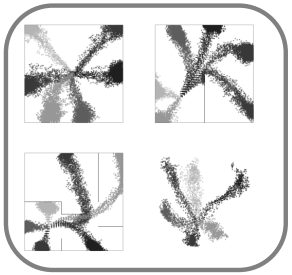
Testing Critical Systems

RAMP







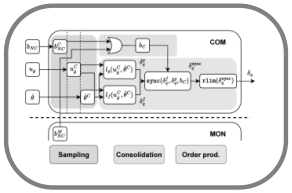


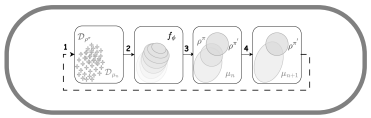
Curiosity-ES

LEADS

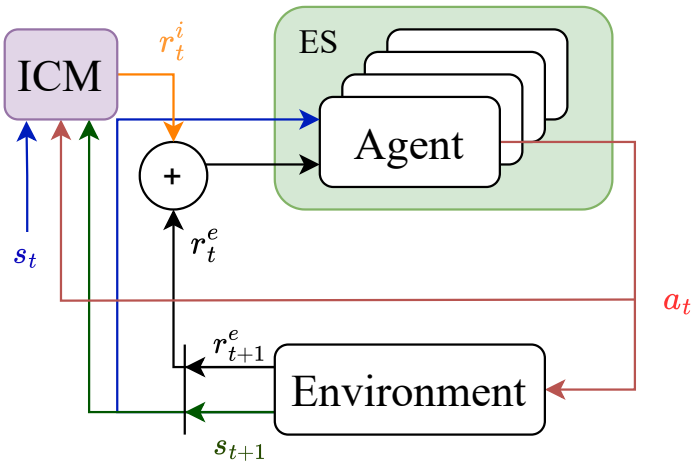
Testing Critical Systems

RAMP

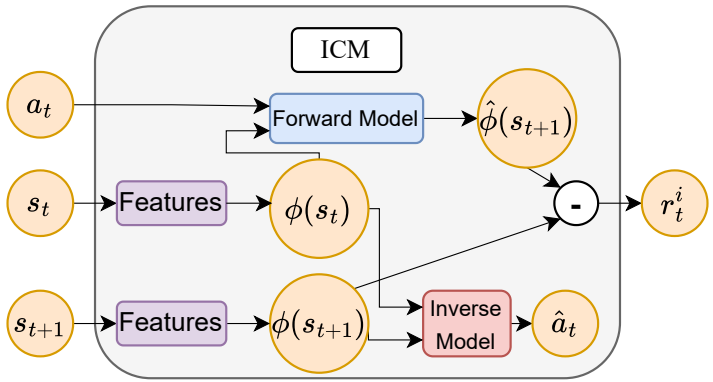




Curiosity-ES



Intrinsic Curiosity Module



Curiosity-ES

Environment Reward

$$f^e = \sum_{t=0}^{T} r_t^e$$

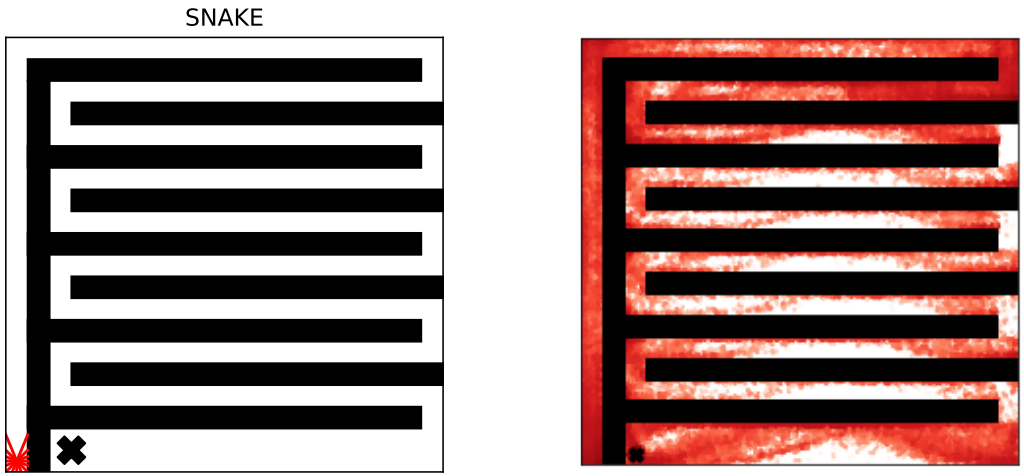
Curiosity Reward

$$f^i = \sum_{t=0}^{T} r_t^i$$

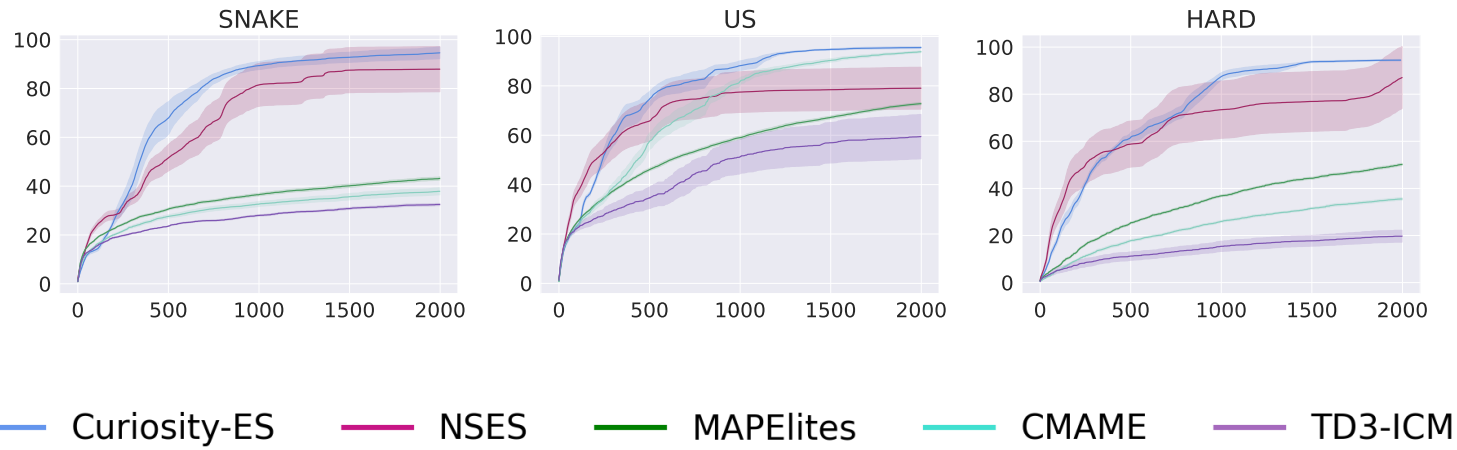
Curiosity-ES Fitness Function

$$\underbrace{f^g}_{\text{global fitness}} = \underbrace{f^e \cdot \alpha}_{\text{extrinsic contribution}} + \underbrace{f^i \cdot (1-\alpha)}_{\text{intrinsic contribution}}$$

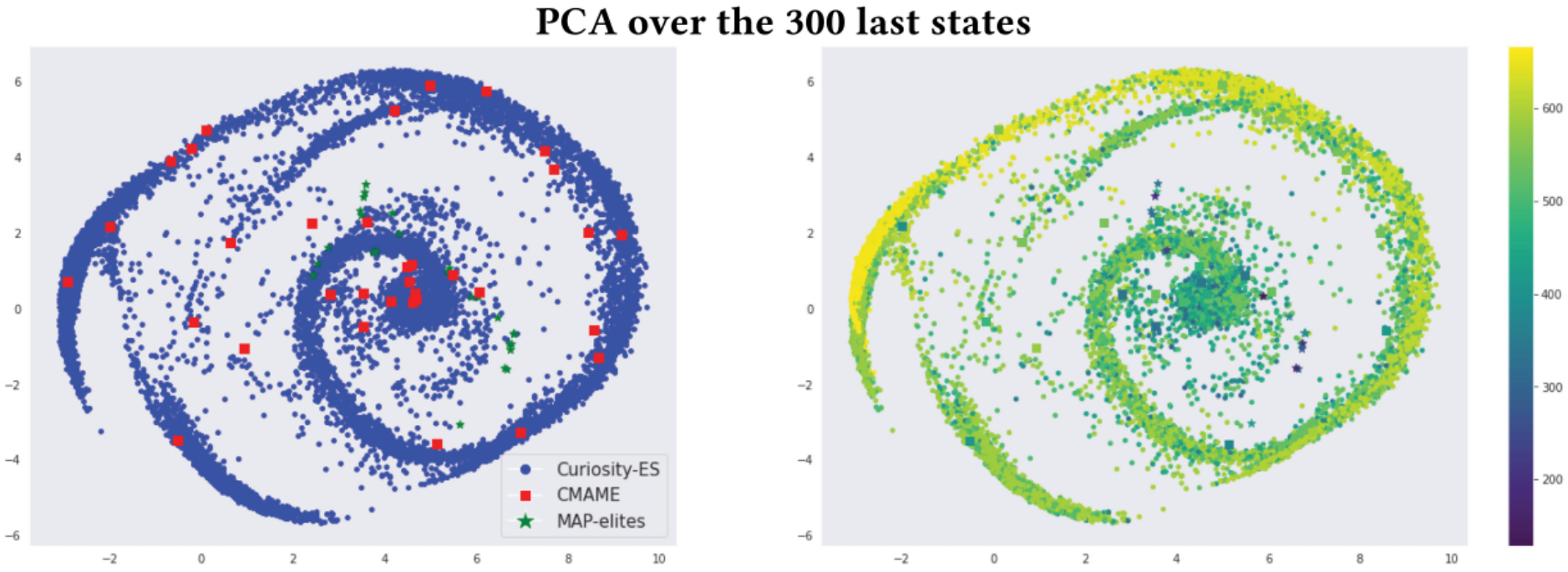
Maze Navigation



Coverage on Maze Navigation

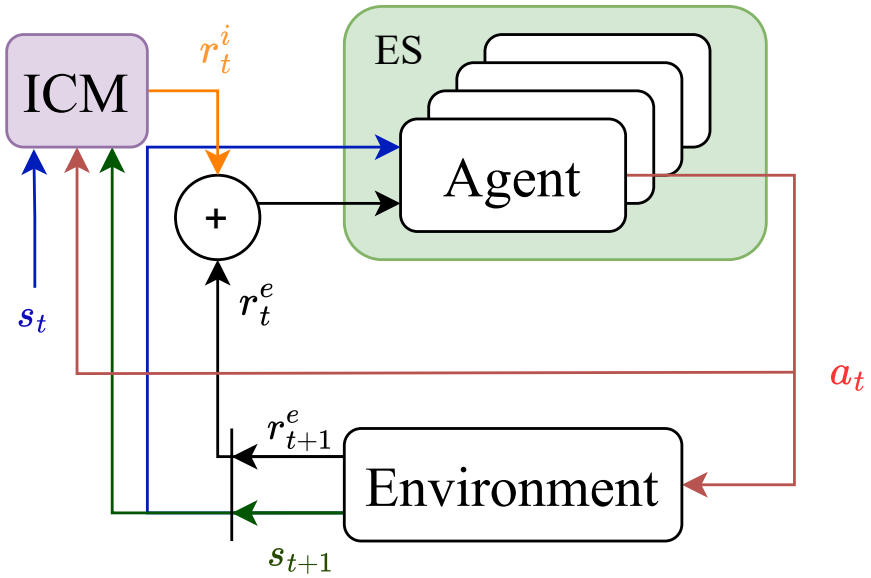


Covering the Policy space



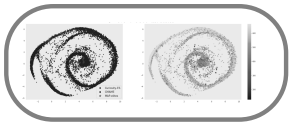
Each policy having found the reward on ”U’s” maze

Main Takeaways

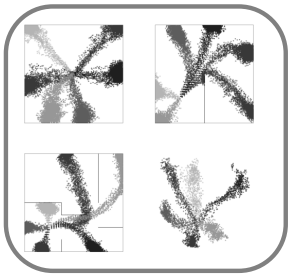


- Curiosity-ES explores the **policy** space
- Curiosity-ES provides efficient exploration of the **state** space

P-A Le Tolguenec et al., "Curiosity Creates Diversity in Policy Search," ACM TELO, 2023.
P-A Le Tolguenec et al., "Summary of 'Curiosity creates Diversity in Policy Search'," GECCO '24 Companion, 2024.



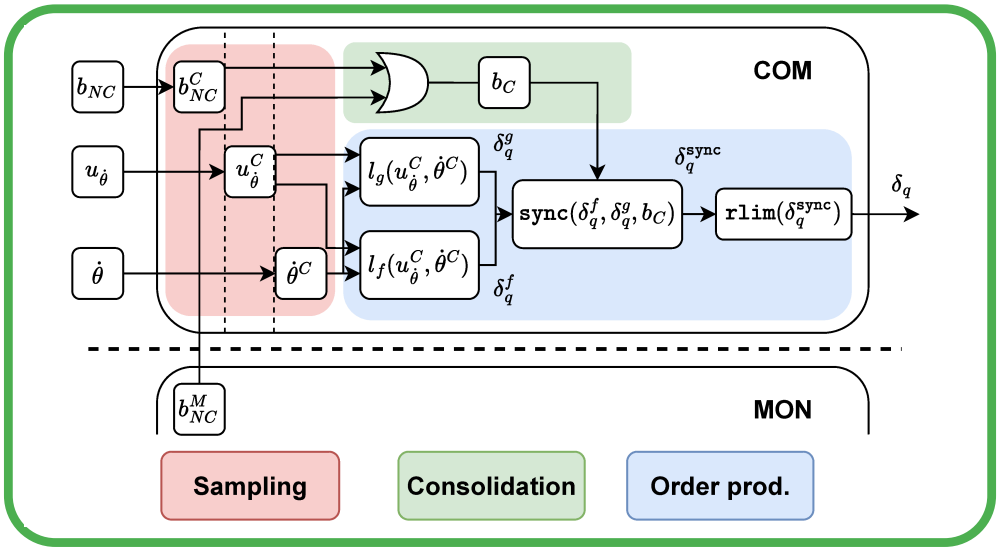
Curiosity-ES

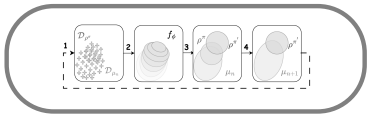


LEADS

Testing Critical Systems

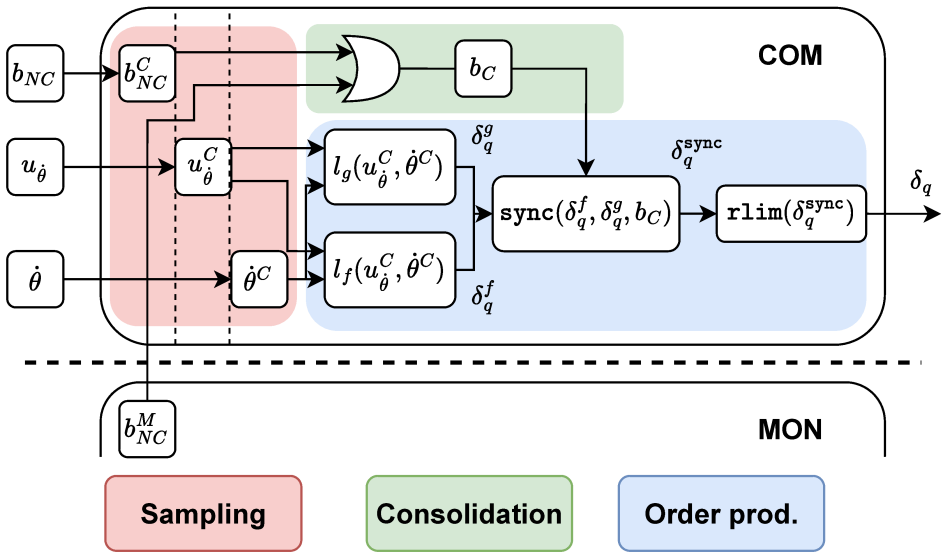
RAMP



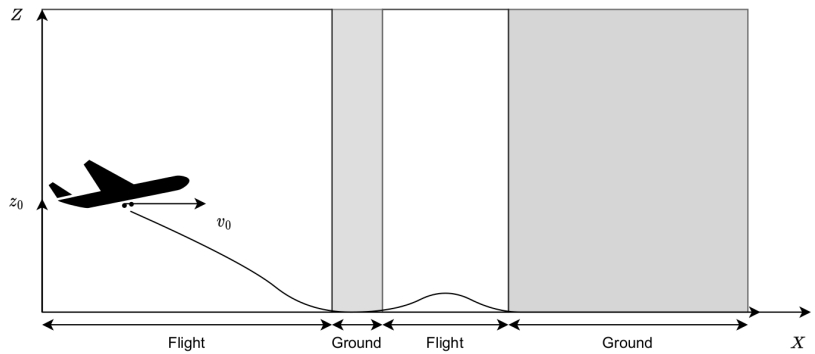


Application & Scenario

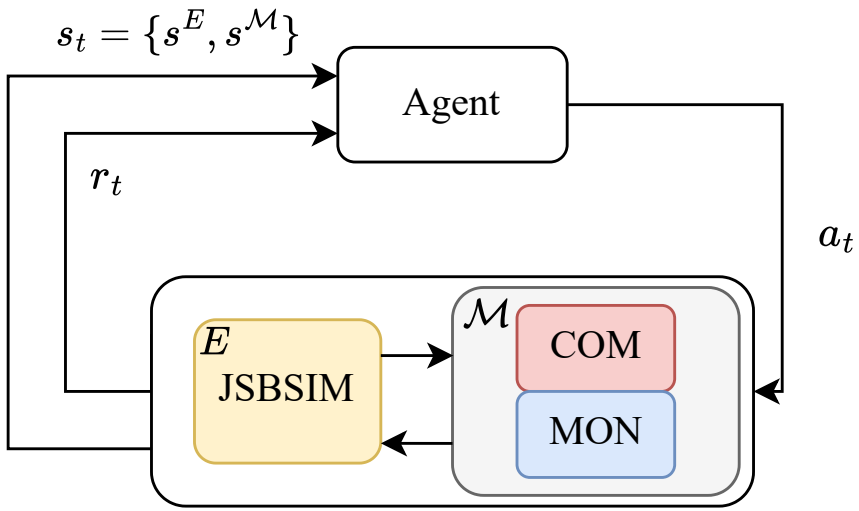
COM/MON System



Landing scenario



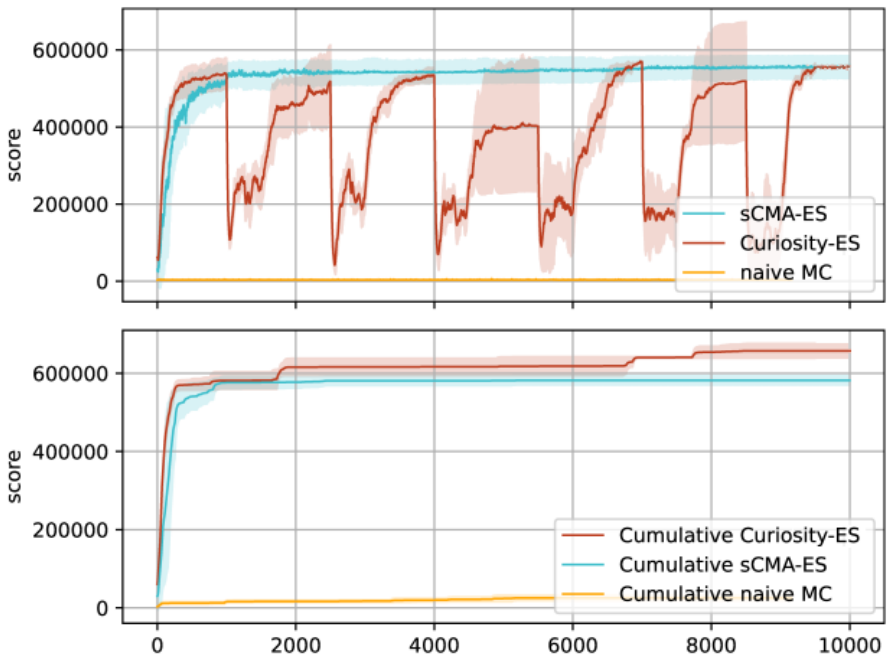
Attacking a COM/MON: a Decision Making Problem



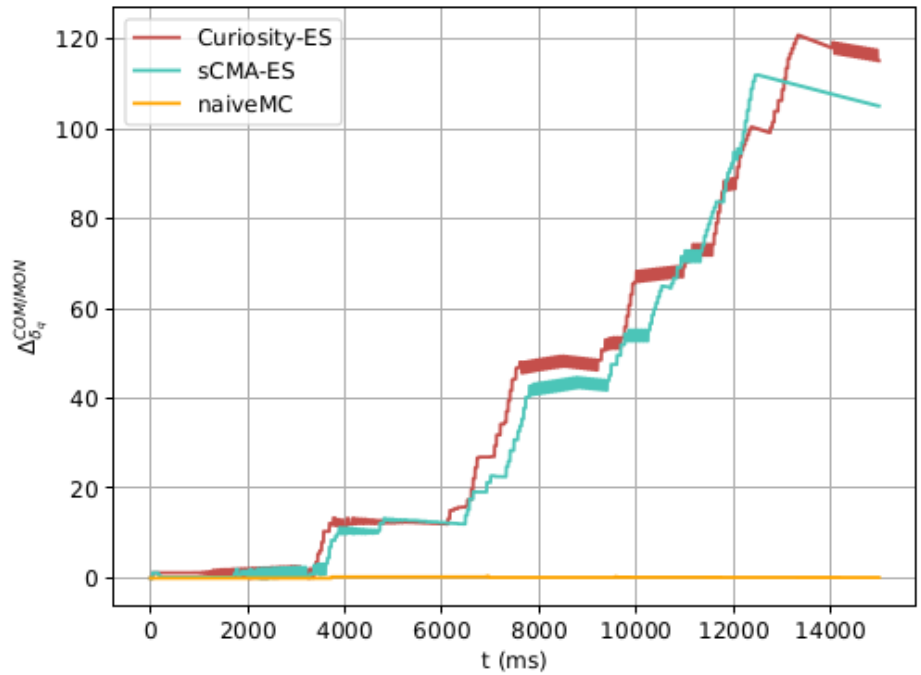
- **Extrinsic Reward**: $r_t^e = |\delta_q^{\text{MON}}(t) - \delta_q^{\text{COM}}(t)|$
- **State Space**: $s_t = \{s^E, s^{\mathscr{M}}\}$
- **Actions**:
  - $u_{\dot{\theta}}$: Angular velocity setpoint
  - $\Delta_t^{\text{COM/MON}}$: Time delay

Policy Search

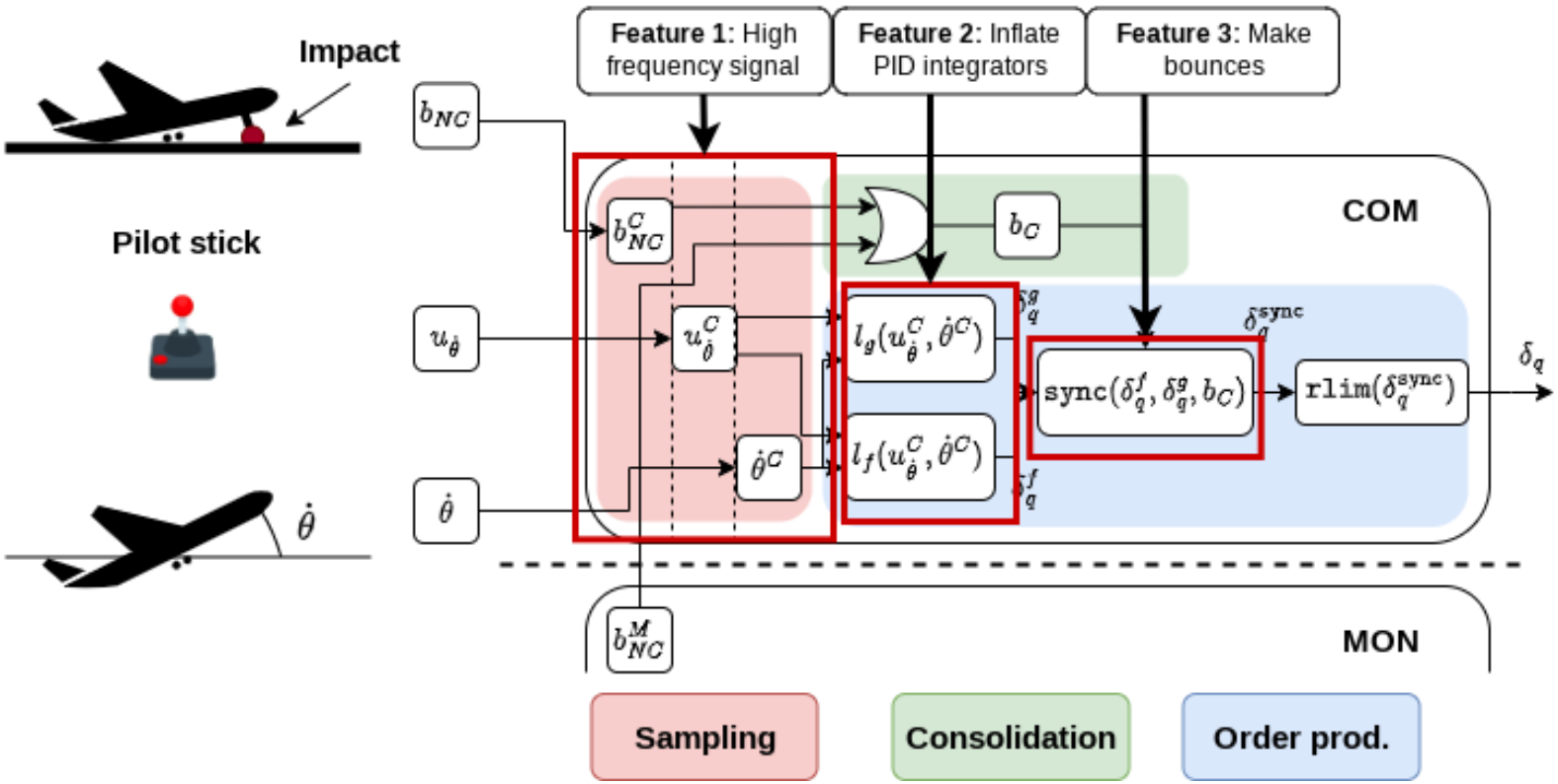
Optimization



Time evolution of $\Delta_{\delta_q}^{\text{COM/MON}}$

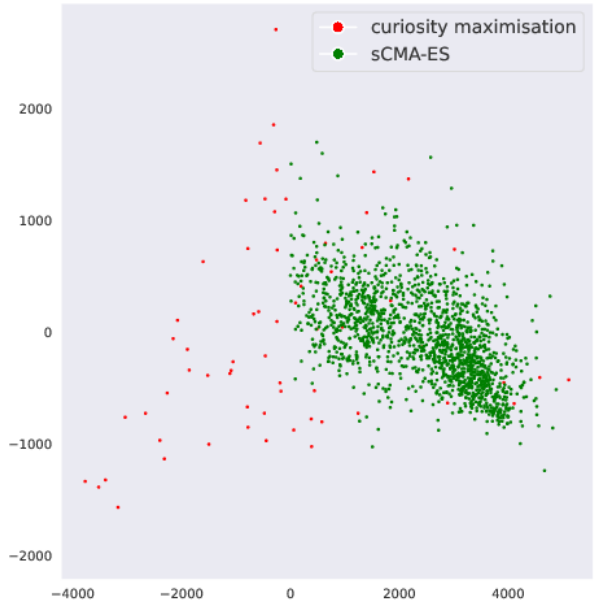


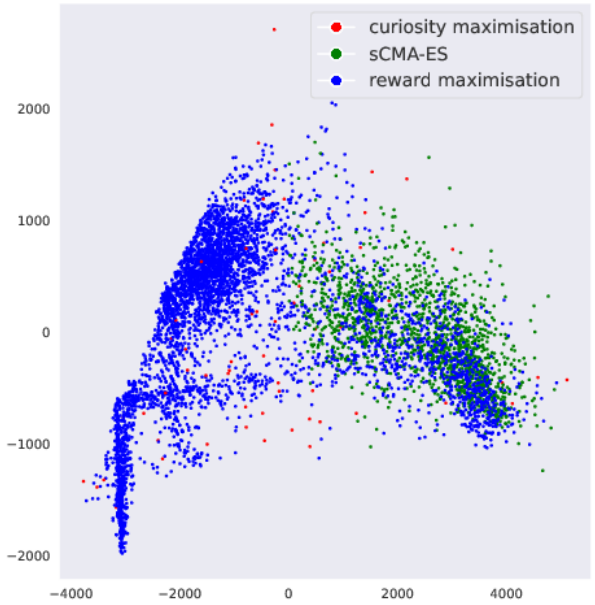
Policy Analysis: The key features

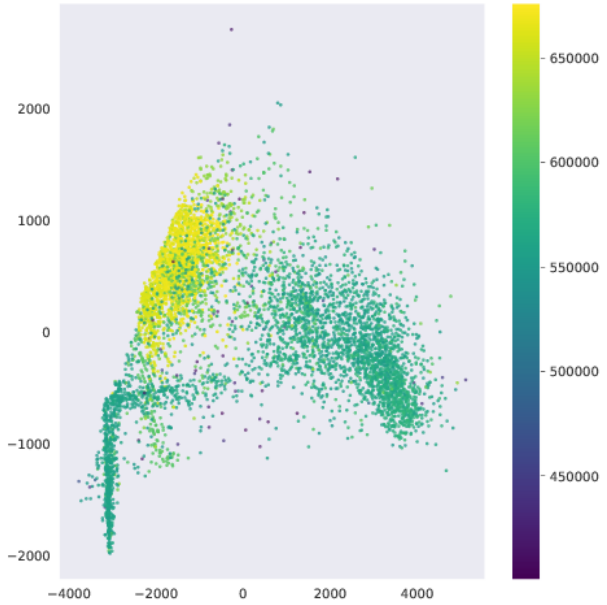


Diversity of failure cases

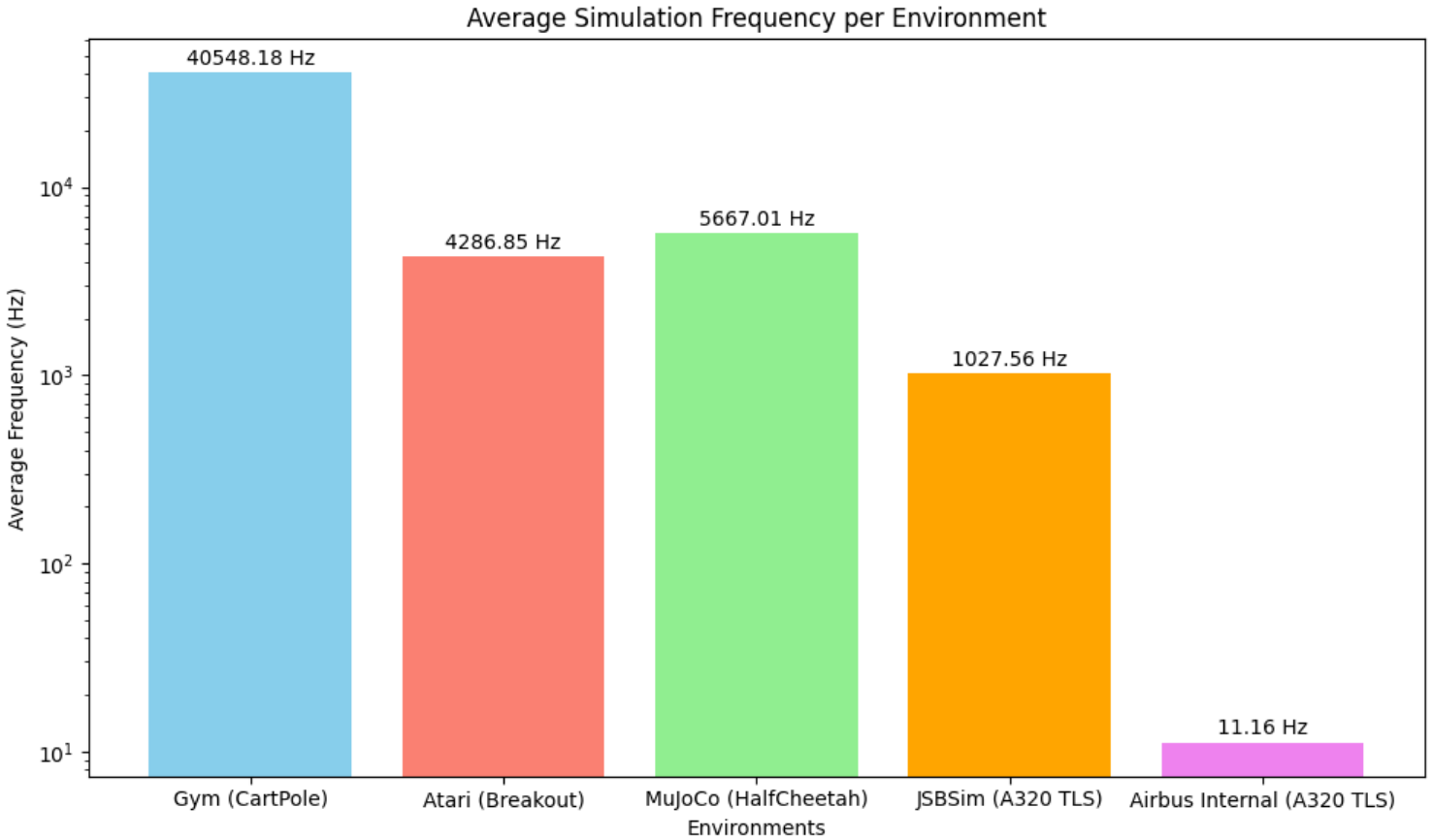
PCA representation of the efficient policies







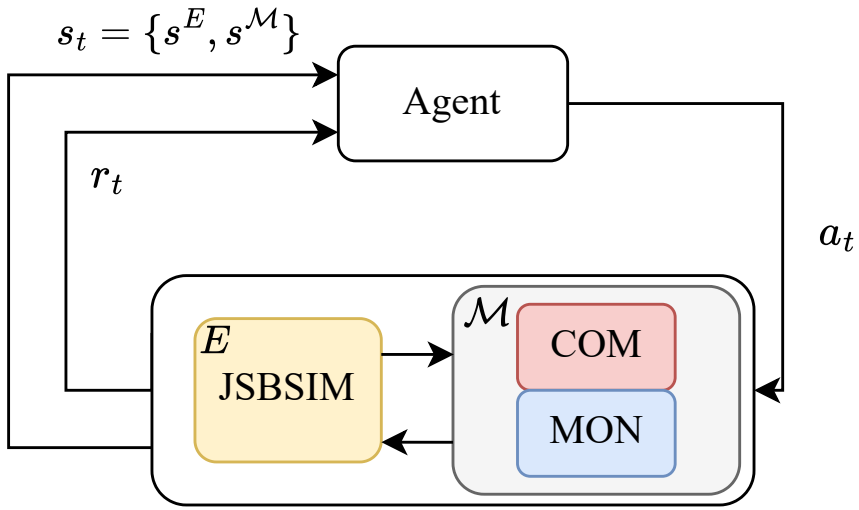
Replicating the study on our internal simulator



Need for sample efficiency

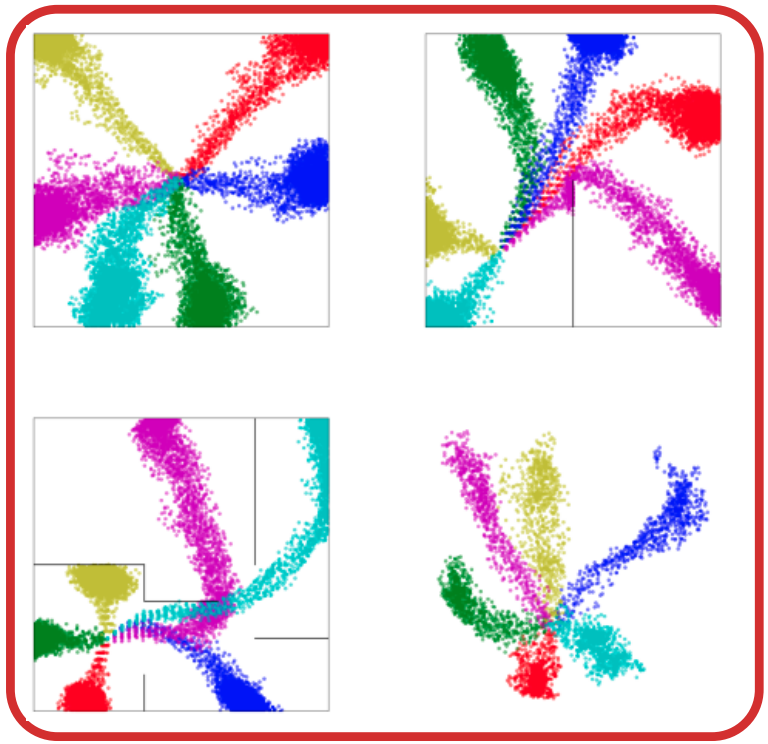
A month and a half to replicate the study

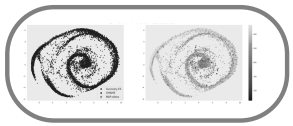
Main Takeaways



- Diverse policies can provide understanding of weaknesses
- Industrial simulators imply the need for sample efficiency

P-A Le Tolguenec et al., "Exploration-Driven Reinforcement Learning for Avionic System Fault Detection (Experience Paper)," ISSTA 2024.
Distinguished Paper Award



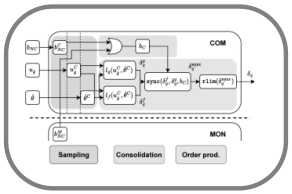


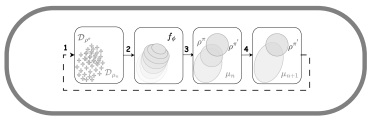
Curiosity-ES

LEADS

Testing Critical Systems

RAMP





Diversity using Mutual Information (MI)

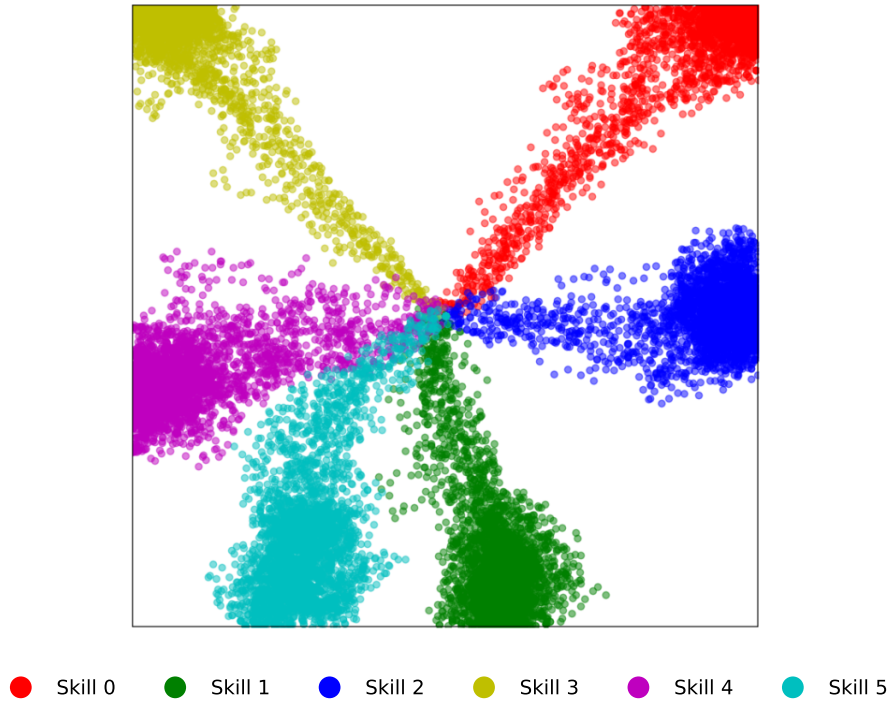
Objective

$$\pi^* = \arg\max_{\pi} \quad I(S, Z)$$

Behavioral Policy

$$\pi : \mathcal{S} \times \mathcal{Z} \to \Delta_{\mathcal{A}}$$

Example of MI Maximization

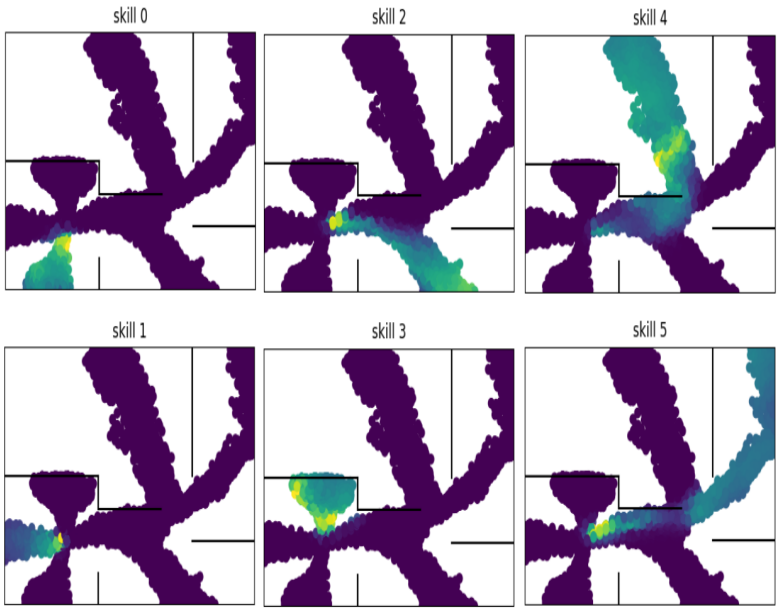


Successor Representation

$$p(s_2|s_1, \pi, z)$$

In practice we estimate: $m^{\pi}_{\phi}(s_1, s_2, z) \approx p(s_2|s_1, \pi, z)/p(s_2)$

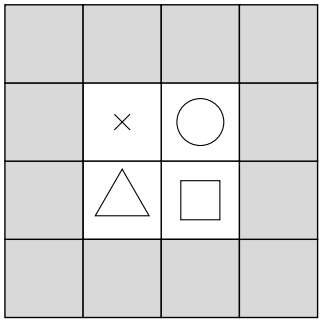
SR Visualization



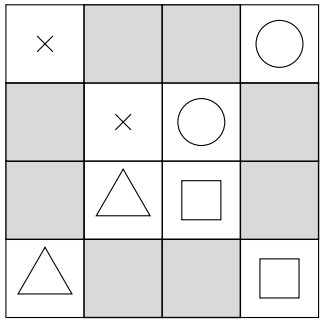
Lower bound on MI

$$I(S, Z) \geq G(\pi, m^{\pi})$$

Limitations of MI Maximization



$\mathcal{Z}_1$



$\mathcal{Z}_2$

How to enforce **exploration**?

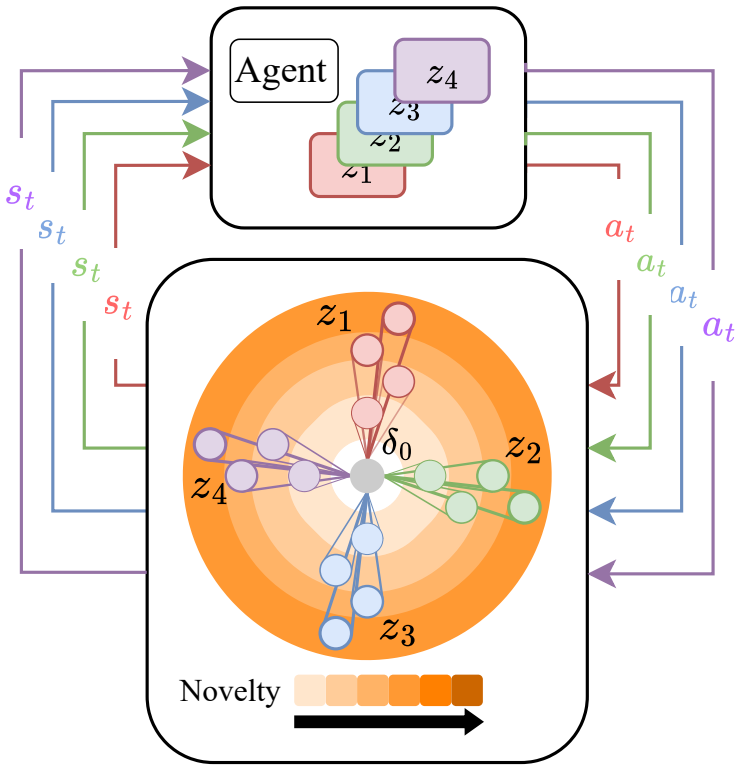
$$I_1(S, Z) = \log 4 \quad \text{and} \quad I_2(S, Z) = \log 4$$

Novelty measure

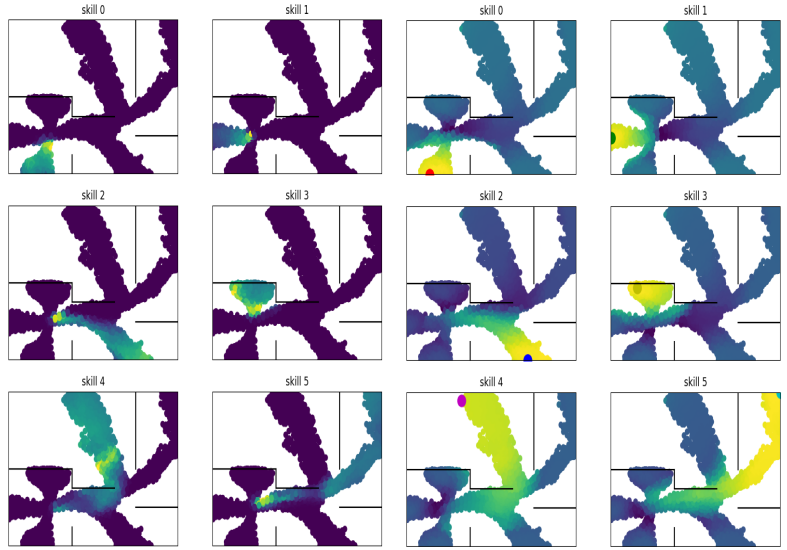
$$u_t(s, z) = \underbrace{\log\left(\frac{m_t(s_0, s, z)}{\sum_{k=1}^{t-1}\sum_{z'} m_k(s_0, s, z')}\right)}_{\text{Explore under-visited areas}} + \underbrace{f(s, z)}_{\text{Repulsion between skills}}$$

Time spent on s by skill z now ■ by all skills previously ■

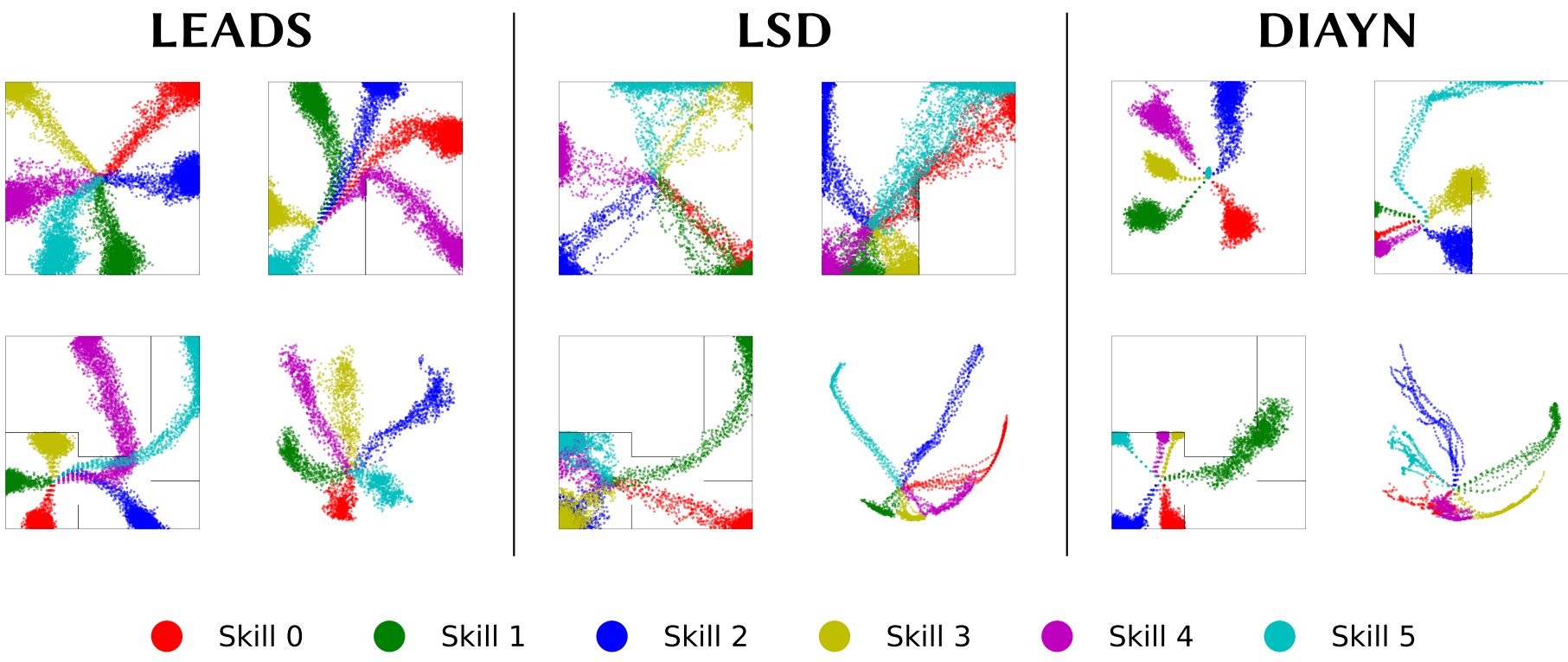
Learning **D**iverse Skills through **S**uccessor Representation



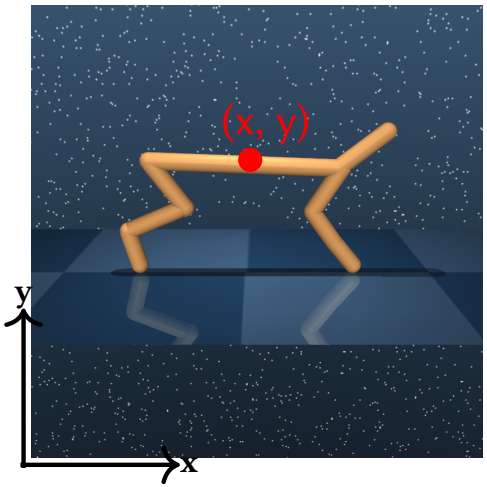
Example of MI Maximization



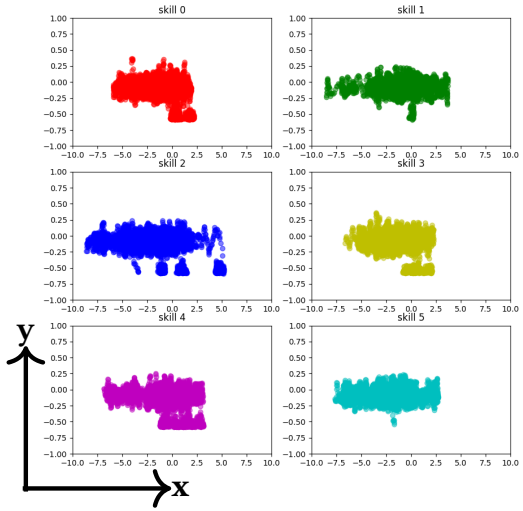
(a) Occupancy (b) Novelty


LEADS
LSD
DIAYN
Skill 0
Skill 1
Skill 2
Skill 3
Skill 4
Skill 5

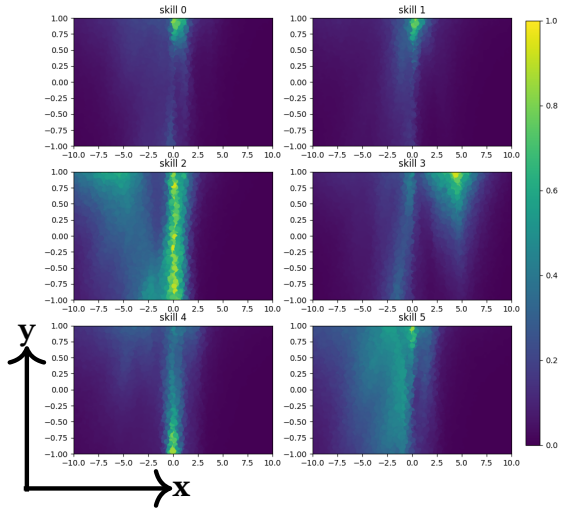
Limitation in high-dimensional state spaces



HalfCheetah ($\dim(\mathcal{S}) = 17$)

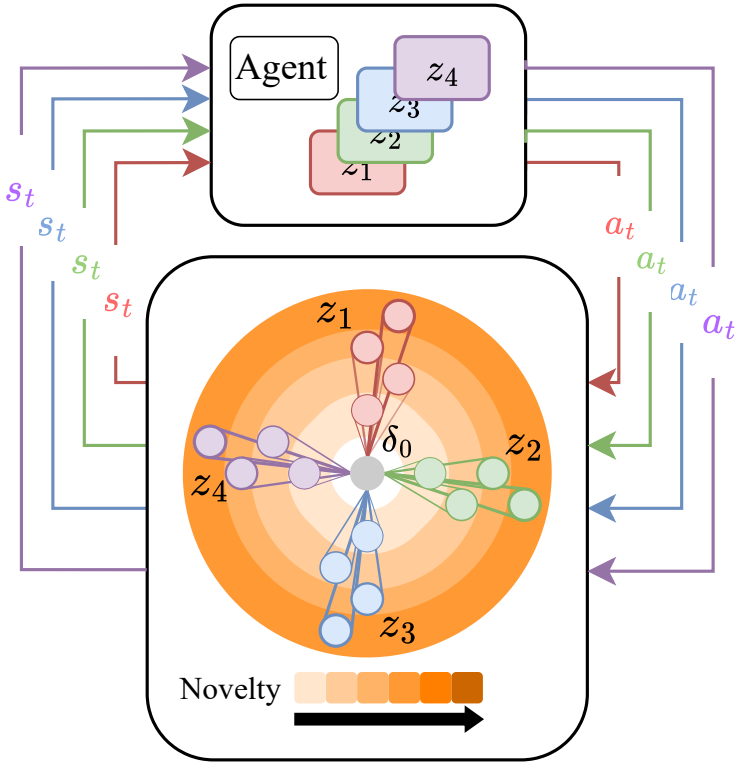


(x,y) torso's coordinates



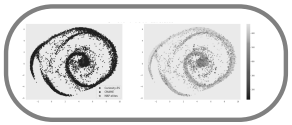
The SR $m(s_0, \pi(s, z), z)$

Main Takeways

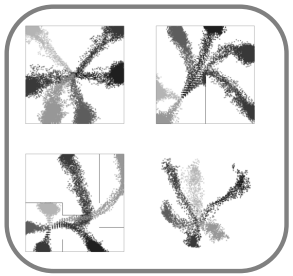


- MI maximization is insufficient for exploration
- LEADS explores efficiently through the novelty measure

Paul-Antoine Le Tolguenec et al., "Exploration by Learning Diverse Skills through Successor Representation," NeurIPS 2024.



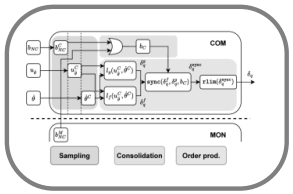
Curiosity-ES

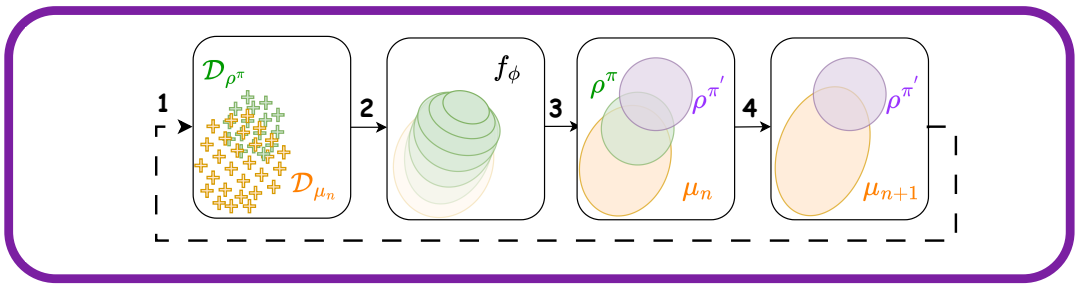


LEADS

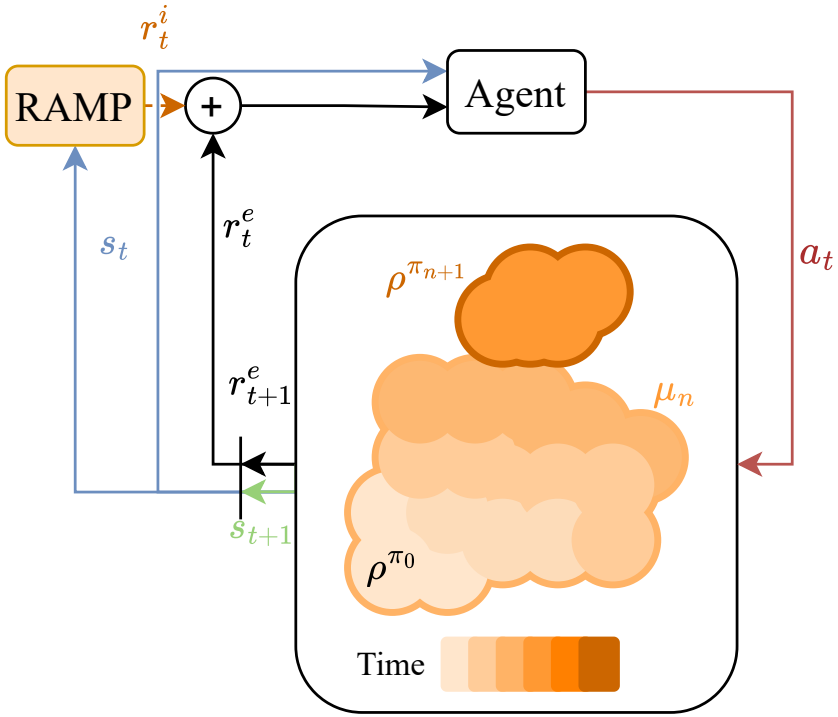
Testing Critical Systems

RAMP





Run Away FroM Your Past (RAMP)



Replay Buffer

$$\mu_{n+1}(s) = \beta \underbrace{\rho_{n+1}(s)}_{\text{The present}} + (1-\beta) \underbrace{\mu_n(s)}_{\text{The past}}$$

Objective

$$\max_{\pi} \quad H_{\mu_{n+1}}[S]$$

Entropy increase rate

$$H_{\mu_{n+1}}[S] - H_{\mu_n}[S] \geq \beta \left(\mathrm{D_{KL}}(\rho_{n+1} \| \mu_{n+1}) + H_{\rho_{n+1}}[S] - H_{\mu_n} \right)$$

Maximizing entropy (Shannon)

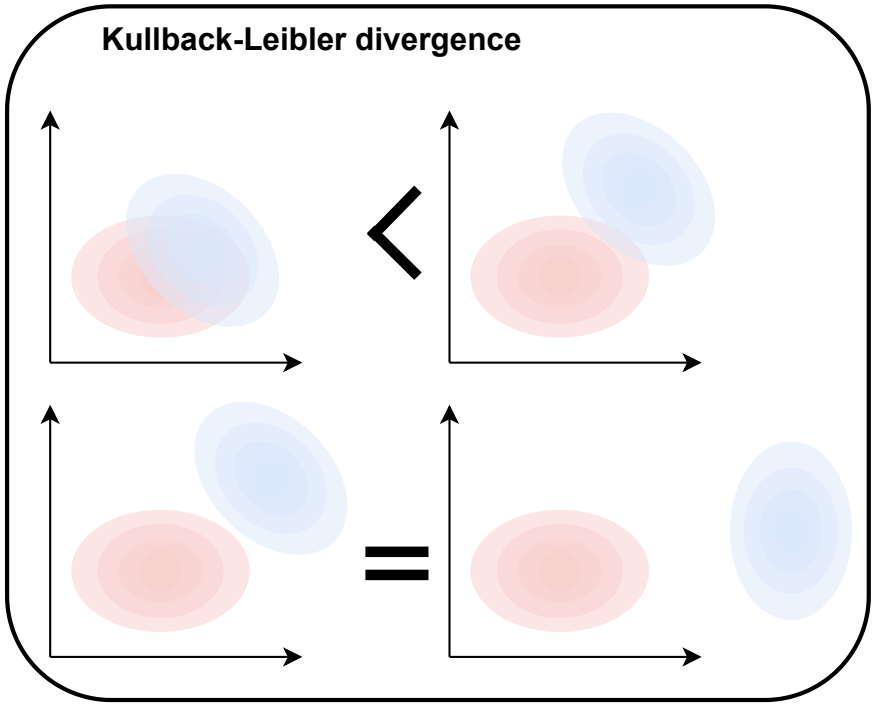
Entropy increase rate

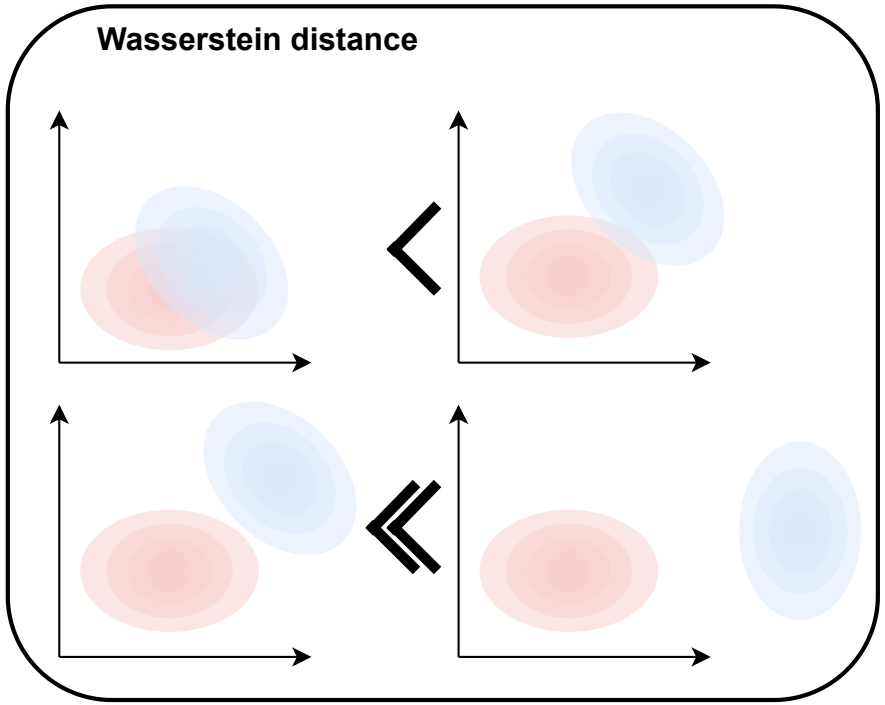
$$H_{\mu_{n+1}}[S] - H_{\mu_n}[S] \geq \beta\Big(\mathrm{D_{KL}}(\rho_{n+1}\|\mu_{n+1}) + H_{\rho_{n+1}}[S] - H_{\mu_n}\Big)$$

Kullback-Leibler divergence

$$\pi_{n+1} = \underset{\pi}{\operatorname{argmax}} \underbrace{\mathrm{D_{KL}}\left(\rho^{\pi}\|\beta\rho^{\pi} + (1-\beta)\mu_n\right)}_{\text{repulsive term}} + \lambda_A \mathop{\mathbb{E}}_{\substack{s\sim\rho^{\pi}\\ a\sim\pi(\cdot|s)}} [-\log(\pi(a|s))]$$

KL divergence vs Wasserstein distance



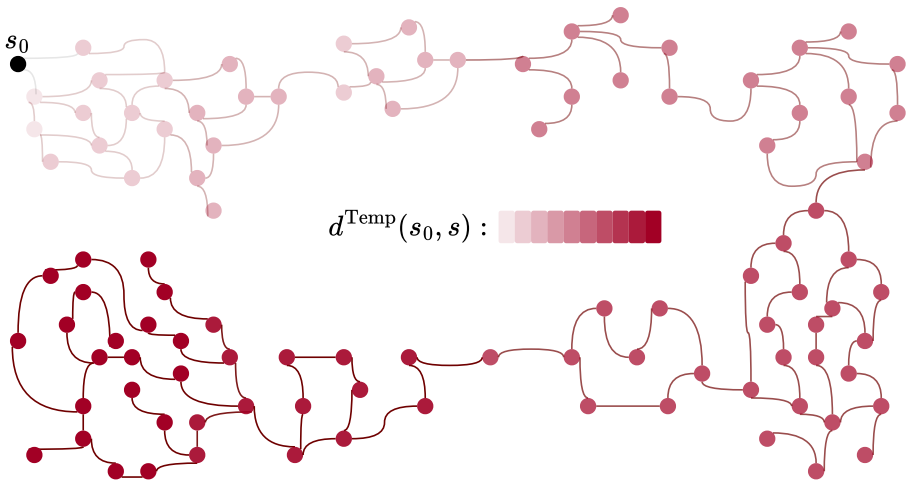


Maximizing entropy (Wasserstein)

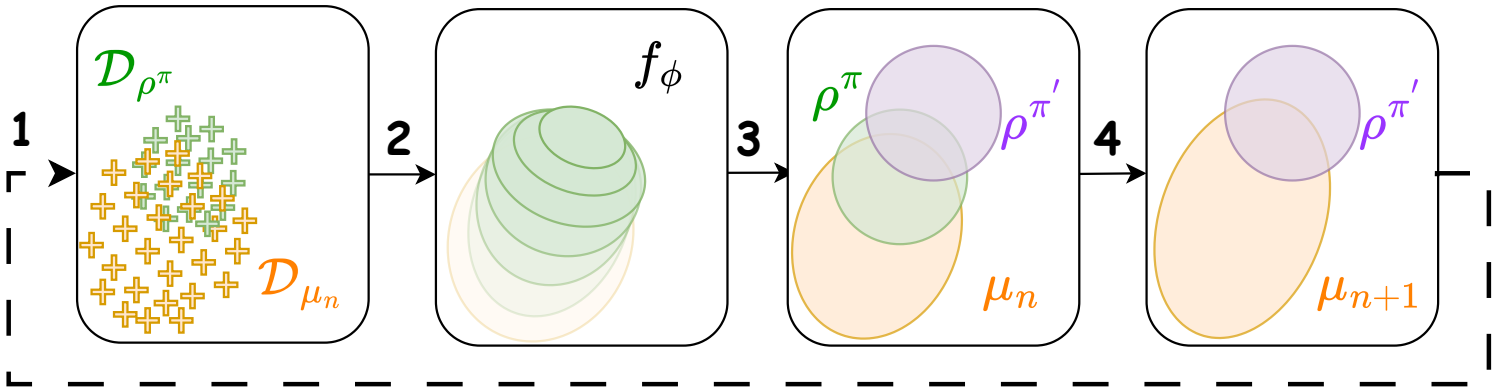
Wasserstein distance

$$\pi_{n+1} = \underset{\pi}{\text{argmax}} \underbrace{\mathscr{W}(\rho^\pi, \beta\rho^\pi + (1-\beta)\mu_n)}_{\text{repulsive term}} + \lambda_A \underset{\substack{s\sim\rho^\pi \\ a\sim\pi(\cdot|s)}}{\mathbb{E}} [-\log(\pi(a|s))]$$

Temporal metric

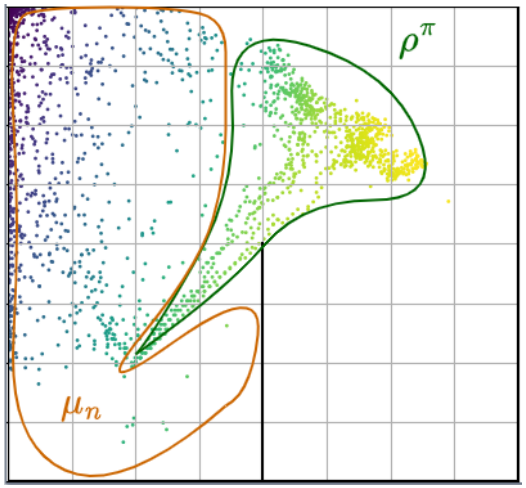


RAMP: a simple idea

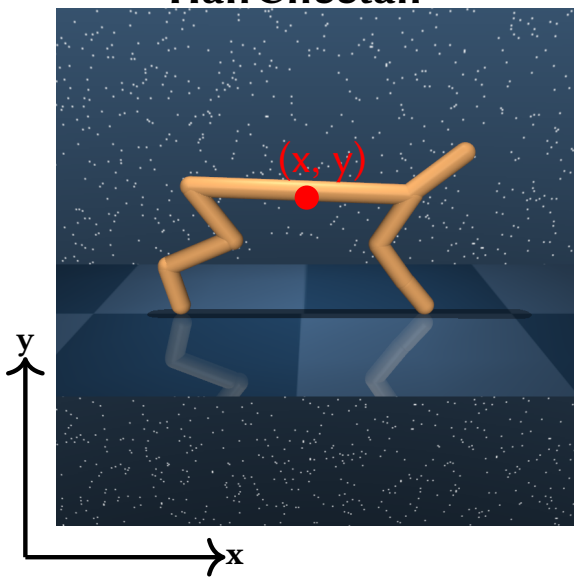


The four steps of the RAMP algorithm.

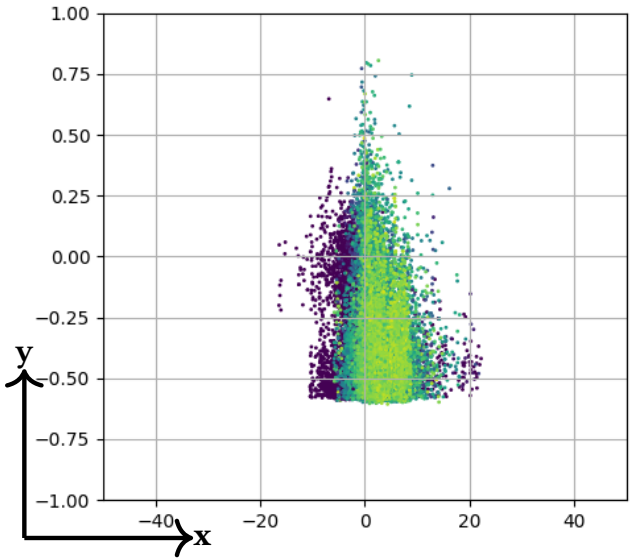
RAMP_{KL}



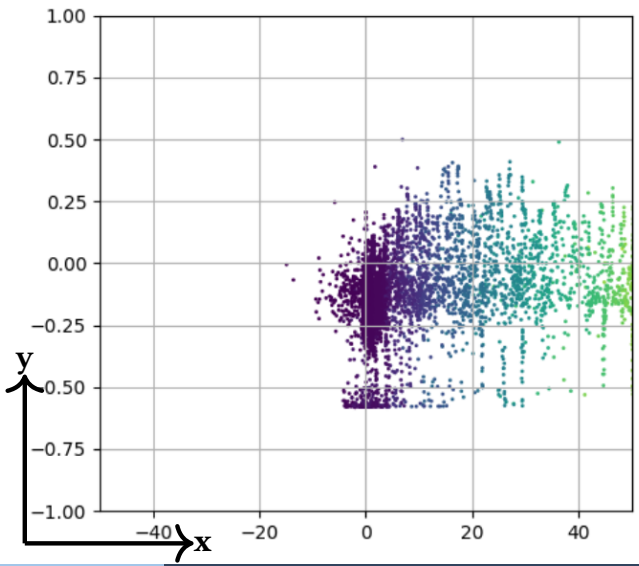
HalfCheetah



RAMP_{KL}



$\text{RAMP}_{\mathscr{W}}$



Exploration using RAMP

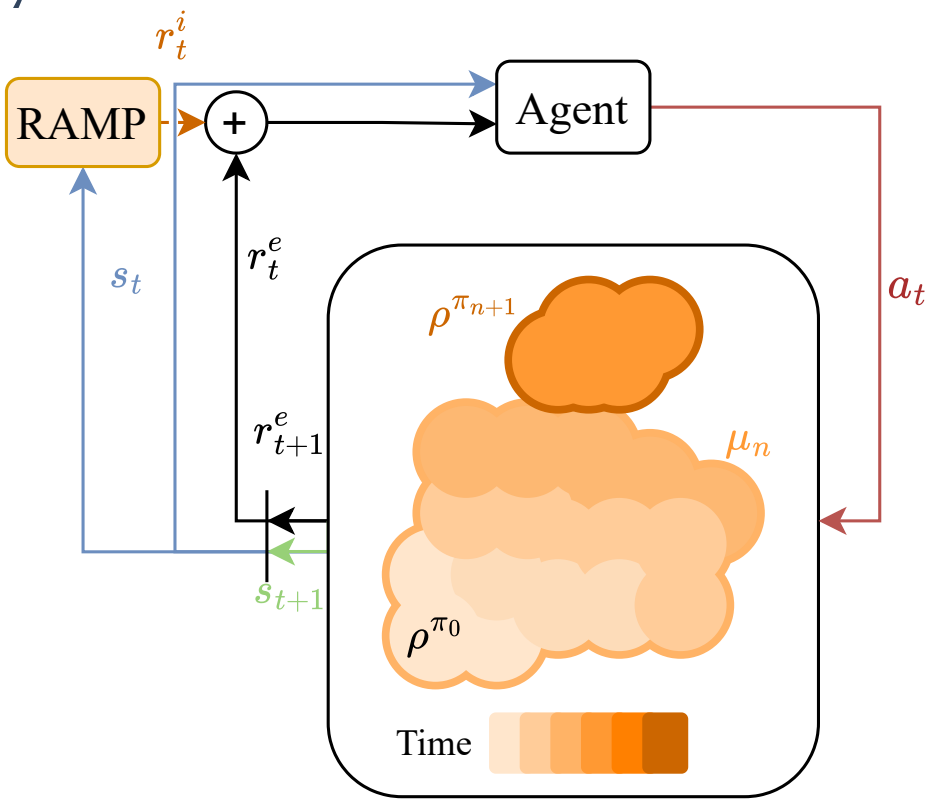
RAMP vs SOTA

APT	DIAYN	LSD	NGU	SMM
AUX	ICM	METRA	RND	SAC

Algorithm	Ant	HalfCheetah	Hopper	Humanoid	Walker2d
APT	7.6 ± 0.9	**93.3 ± 3.3**	55.5 ± 3.7	54.7 ± 2.9	55.3 ± 2.8
METRA	23.5 ± 0.7	73.8 ± 3.0	37.5 ± 3.3	88.3 ± 5.1	36.8 ± 4.1
RAMP$_{\mathscr{W}}$	**78.4 ± 13.5**	40.3 ± 6.7	**74.4 ± 12.1**	**90.5 ± 2.3**	**74.9 ± 11.7**

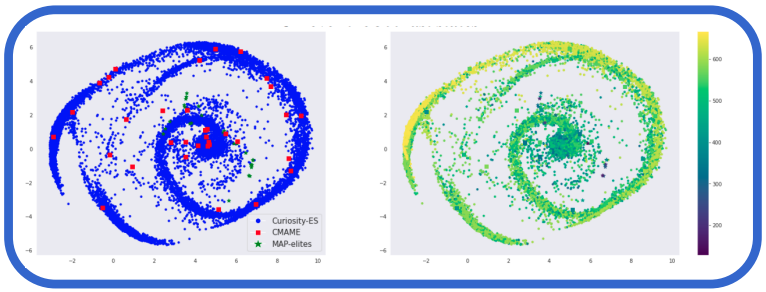
Table: RAMP$_{\mathscr{W}}$ state coverage.

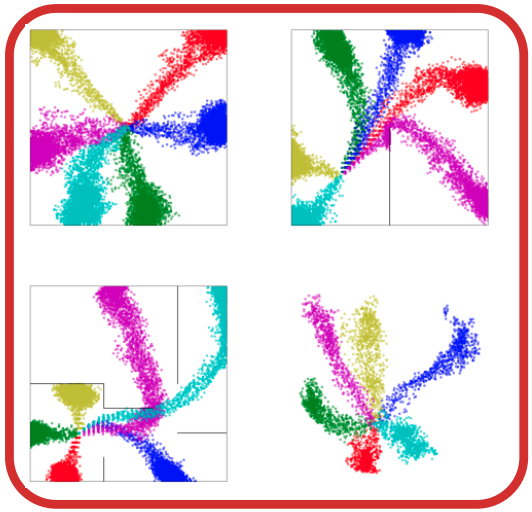
Main Takeaways



- Explore efficiently by seeking novelty.
- Different divergences imply different explorations.

Paul-Antoine Le Tolguenec et al., "Exploration by Running Away from the Past," In preparation for submission.



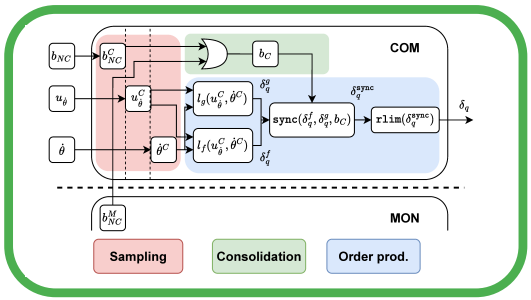


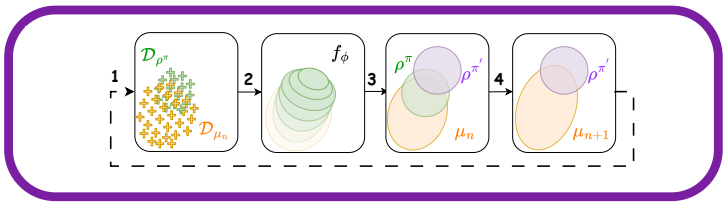
Curiosity-ES

LEADS

Testing Critical Systems

RAMP





New perspectives on Adaptive Stress Testing

A unified objective to support engineers

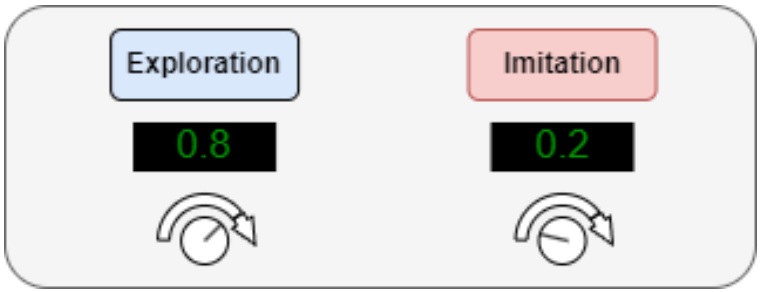
$$\max_{\pi_{\text{agent}}} \mathbb{E}_{s\sim\rho^{\pi_{\text{agent}}}}[r_{\text{attack}}(s)] + \lambda_{\text{exp}} \underbrace{\mathscr{W}_d(\rho^{\pi_{\text{agent}}}, \mu_n)}_{\text{Exploration objective}} - \lambda_{\text{imit}} \underbrace{\mathscr{W}_d(\rho^{\pi_{\text{agent}}}, \rho^{\pi_{\text{operator}}})}_{\text{Imitation objective}}$$

One reward to rule them all

$$r(s) = r_{\text{attack}}(s) + \lambda_{\text{exp}} f_{\text{exp}}(s) - \lambda_{\text{imit}} f_{\text{imit}}(s)$$

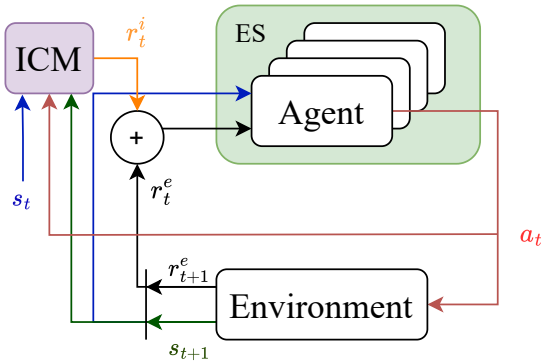
Which d ?

expertise ⇒ metric

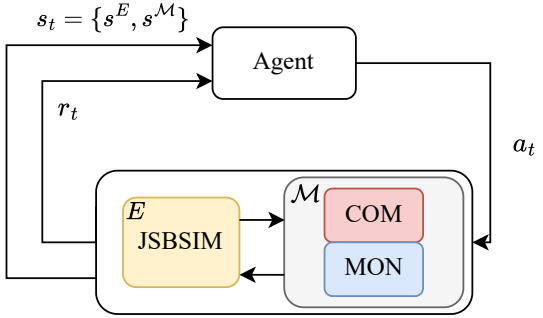


Conclusion

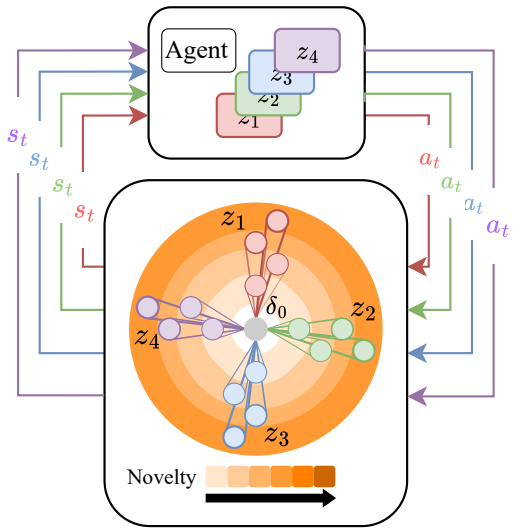
Curiosity-ES



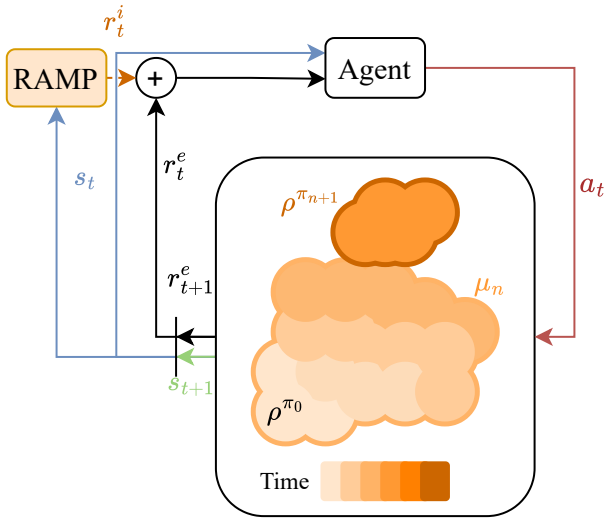
Testing a critical system



LEADS

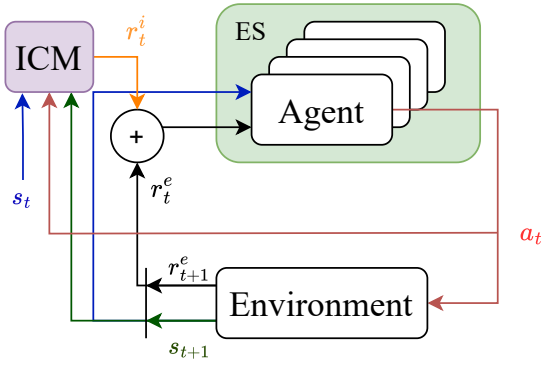


RAMP

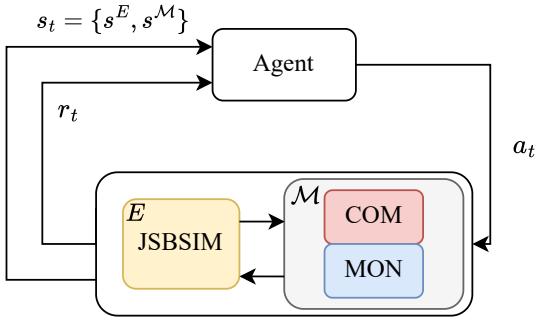


Conclusion

Curiosity-ES

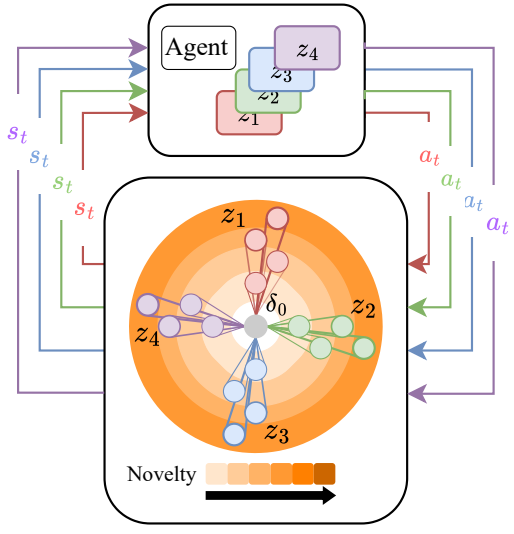


Testing a critical system

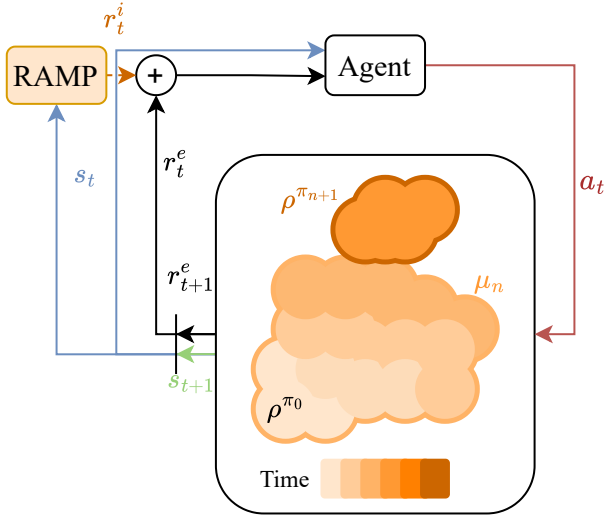


What is next ?

LEADS

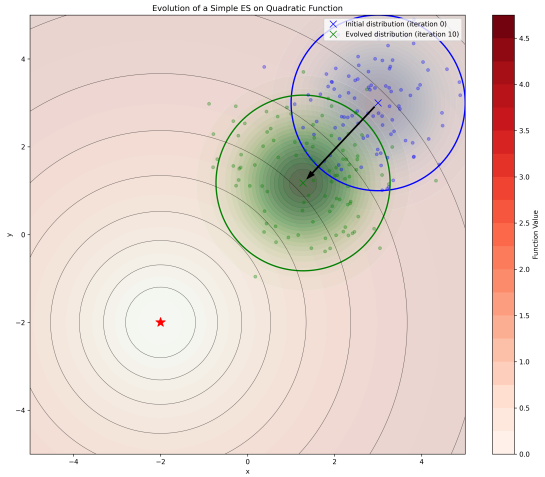


RAMP



Curiosity creates Diversity in Policy Search

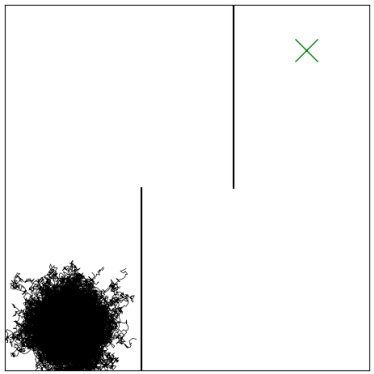
Black-Box optimization



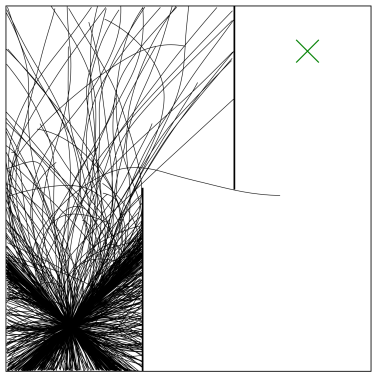
Evolutionary Strategies

$$\theta^{k+1} \leftarrow \theta^k + \nabla_\theta \left(\mathop{\mathbb{E}}_{\theta_i \sim \mathcal{N}(\theta,\Sigma)} [f(\theta_i)] \right)$$

Evolutionary Strategies in Policy Search



White noise action



ES

Curiosity-ES

Environment reward

$$f_\theta^e = \sum_{t=0}^{T} r_t^e$$

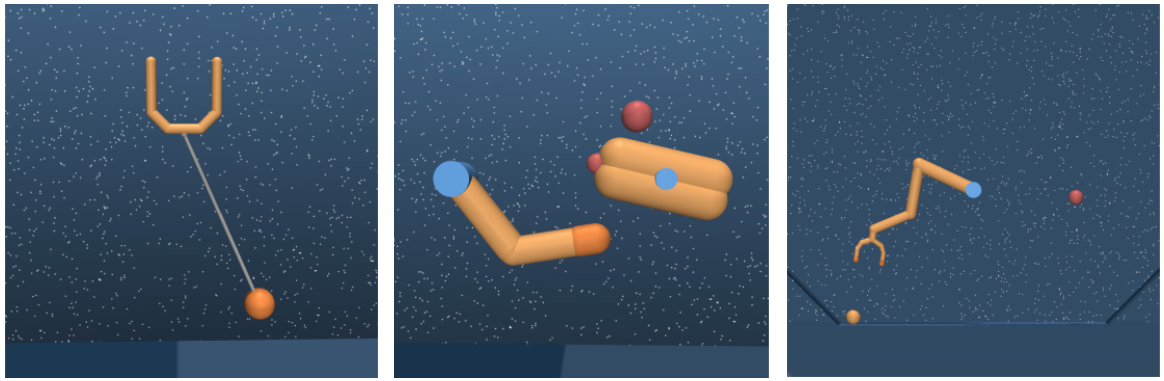
Curiosity reward

$$f_\theta^i = \sum_{t=0}^{T} r_t^i$$

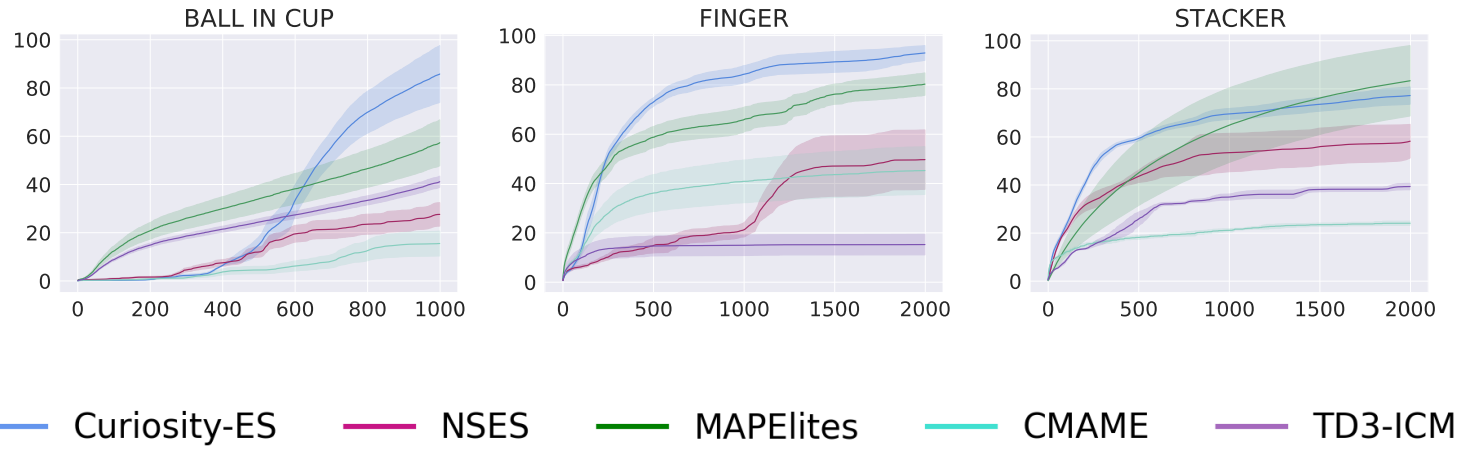
Curiosity-ES Fitness Function

$$\underbrace{f_\theta^g}_{\text{global fitness}} = \underbrace{\frac{f_\theta^e - \mu^e}{\sigma^e}}_{\text{extrinsic contribution}} \varphi + \underbrace{\frac{f_\theta^i - \mu^i}{\sigma^i}}_{\text{intrinsic contribution}} (1 - \varphi)$$

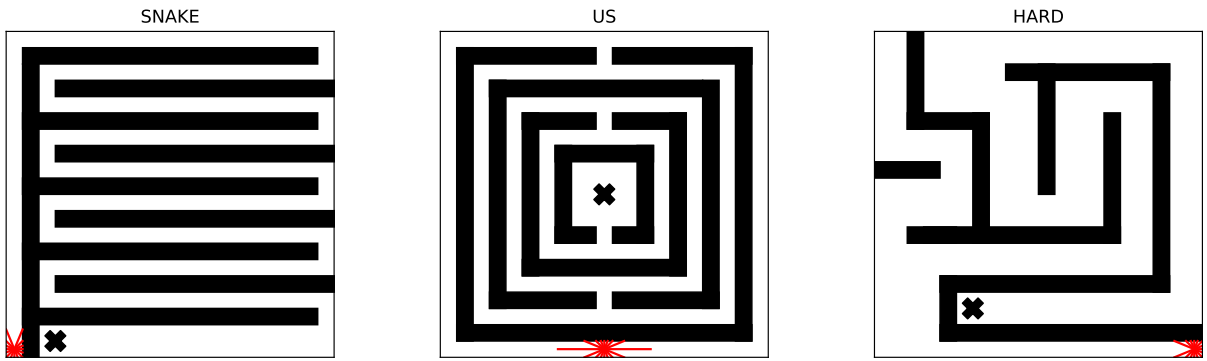
Control Tasks



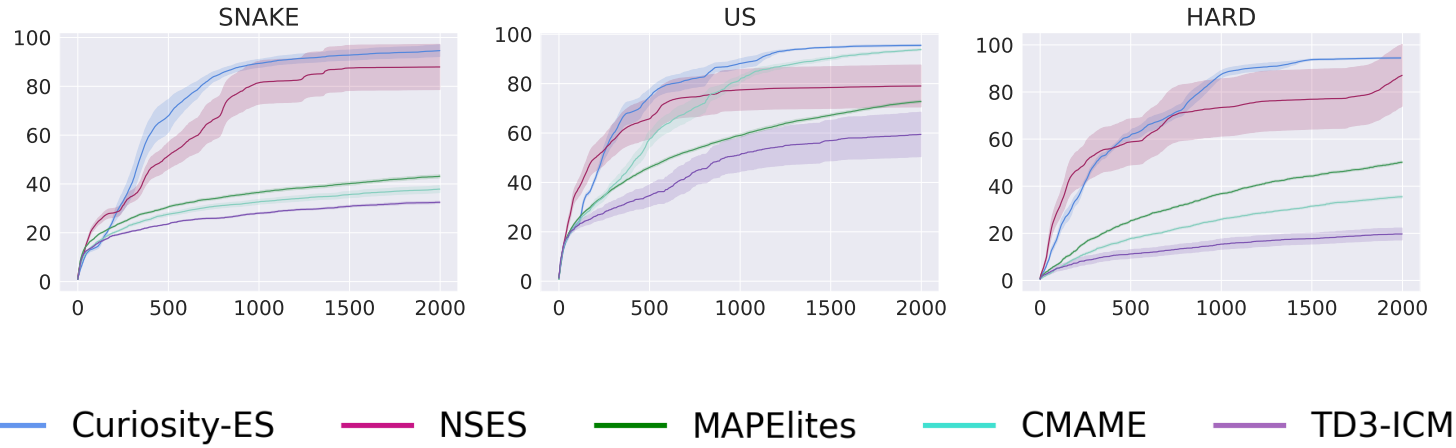
Coverage in Control Tasks



Maze Navigation

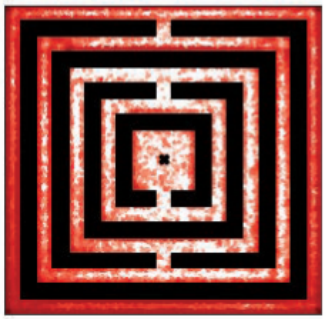


Coverage on Maze Navigation



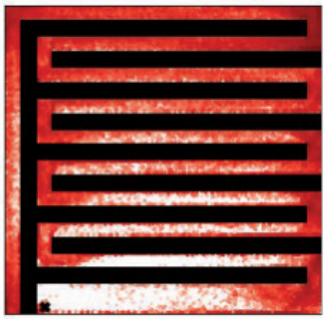
Maze coverage

(a) CMAME on SNAKE



(b) CMAME on US

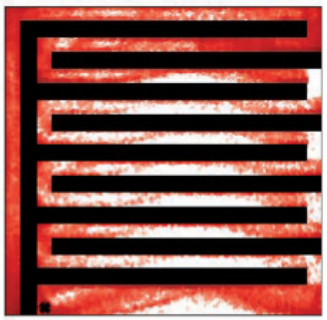
(c) CMAME on HARD



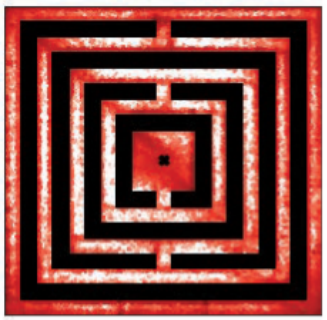
(d) NS-ES on SNAKE

(e) NS-ES on US

(f) NS-ES on HARD



(g) Curiosity-ES on SNAKE



(h) Curiosity-ES on US

(i) Curiosity-ES on HARD

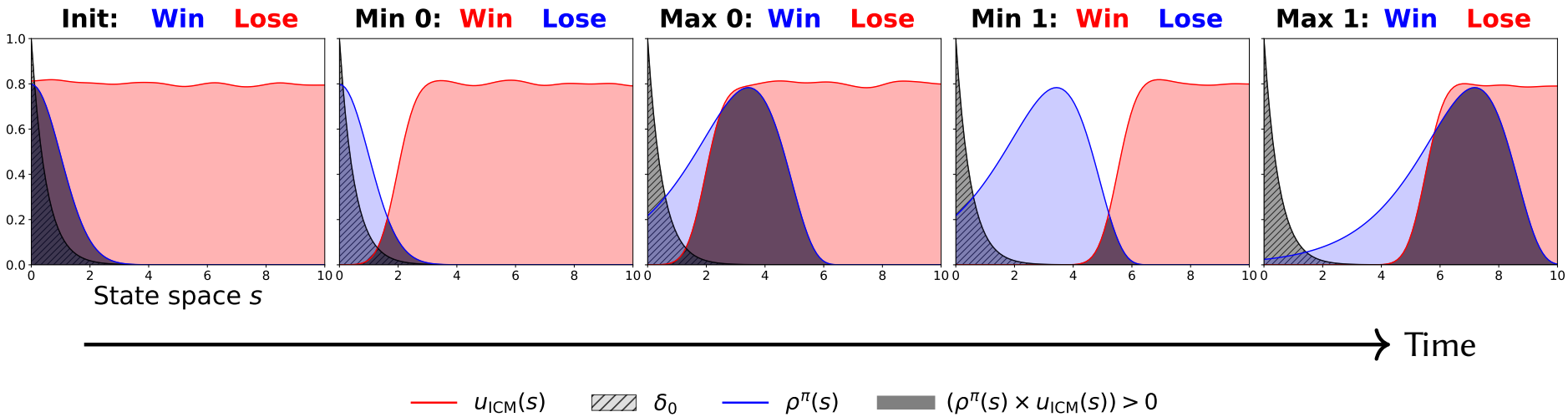
Two-player adversarial game

Uncertainty measure

$$u(s_t, w) = \mathop{\mathbb{E}}_{\substack{a_t \sim U(A) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} [l_{\text{ICM}}(w, s_t, a_t, s_{t+1})]$$

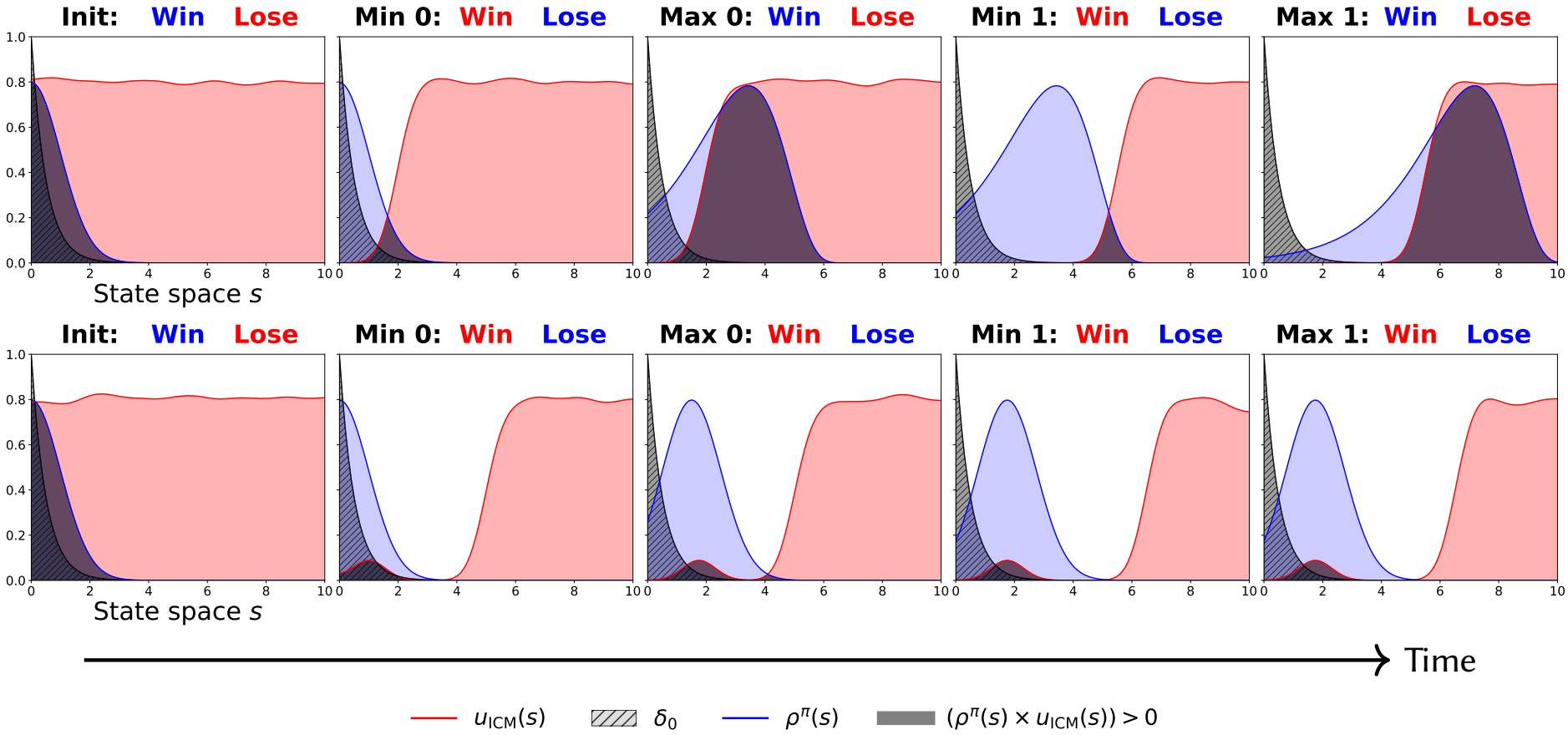
Max-Min

$$\max_{\theta} \min_{w \text{ s.t. } \langle \rho^{\pi_\theta}, u_w \rangle \geq \varepsilon_c} \langle \rho^{\pi_\theta}, u_w \rangle - \varepsilon_p$$



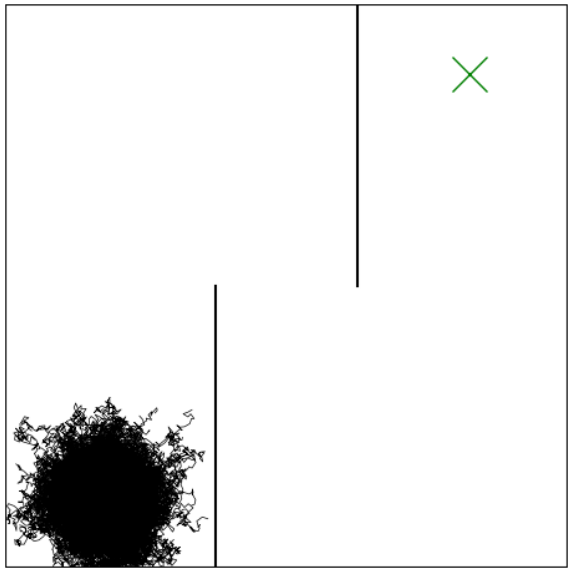
Evolution of payoffs for both players as a function of time and in the optimal case.

Two-player adversarial game

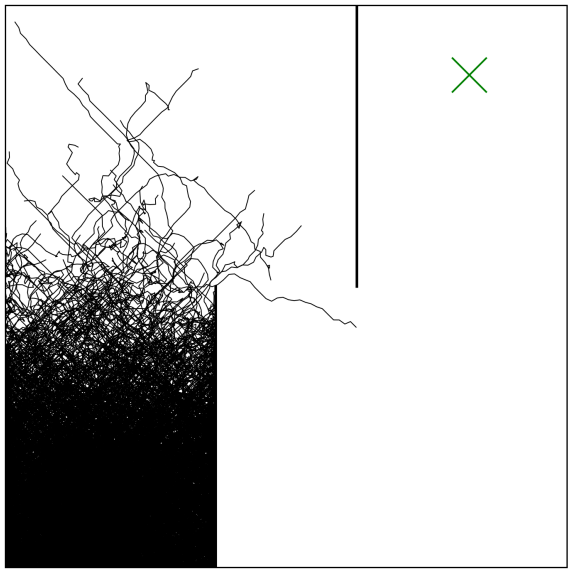


Evolution of payoffs for both players as a function of time and in the optimal case (top) and worst case (bottom).

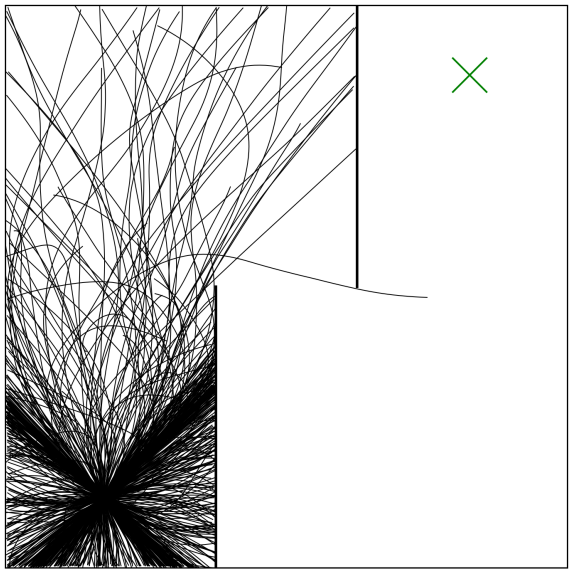
Exploration



White noise exploration:
$a_t \sim \mathcal{N}(0, 1\mathrm{I})$.



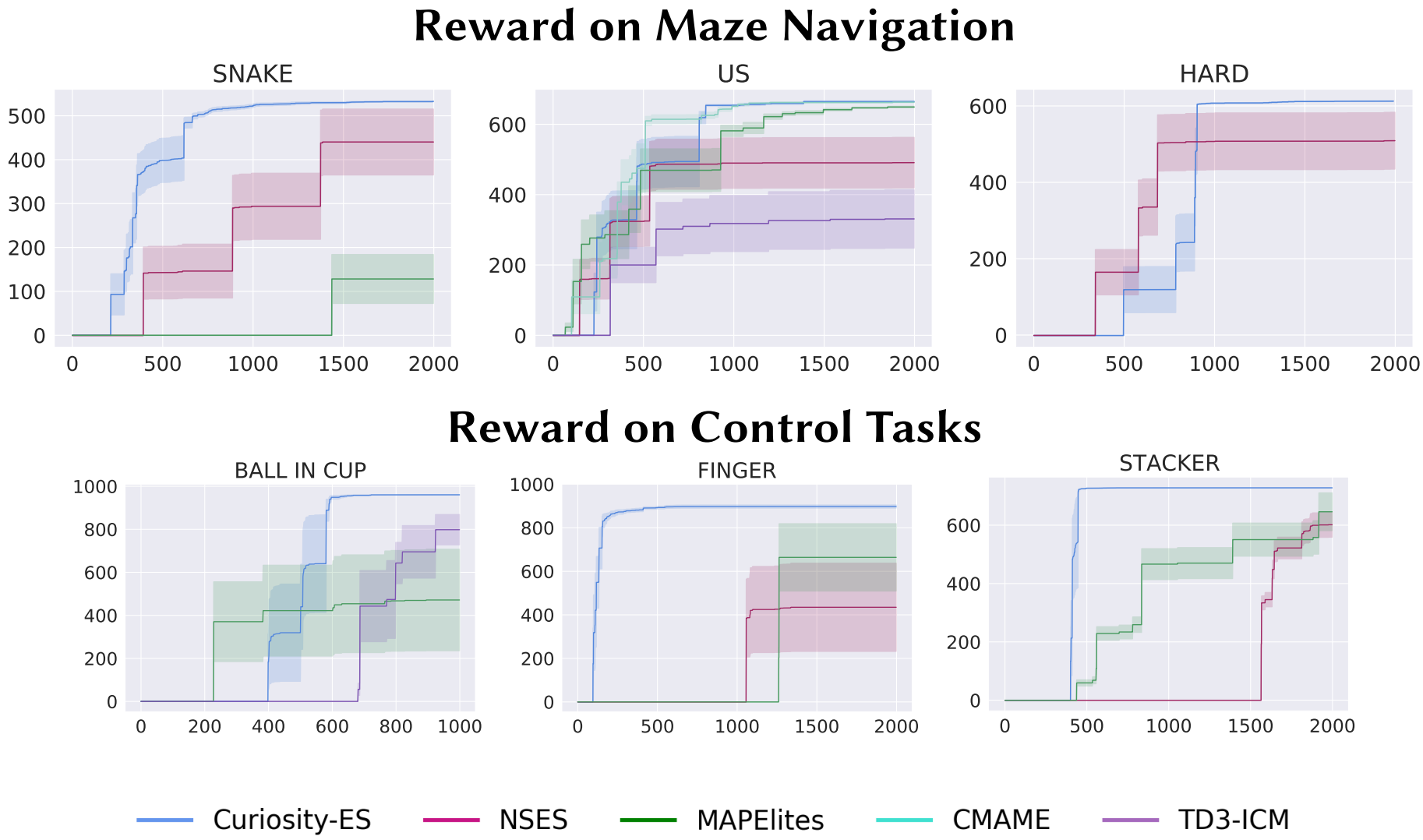
OU noise exploration:
$\beta = 0.1$.



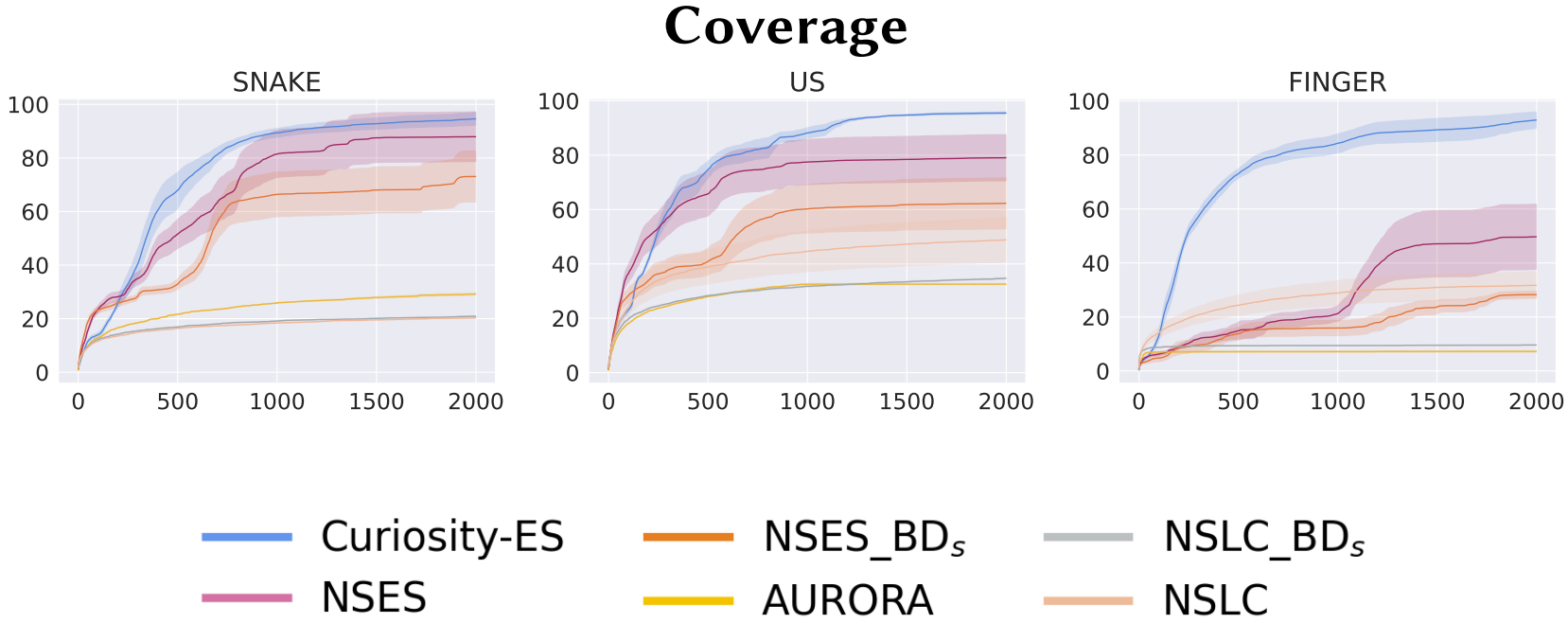
ES exploration:
$\theta \sim \mathcal{N}(0, 0.1\mathrm{I})$, with
$\theta \in \mathbb{R}^{322}$.

10^3 trajectories of a 2D agent in a continuous maze exploring through white noise (left), OU noise (center), and ES (right).

Rewards


Reward on Maze Navigation
SNAKE
US
HARD
Reward on Control Tasks
BALL IN CUP
FINGER
STACKER
Curiosity-ES
NSES
MAPElites
CMAME
TD3-ICM

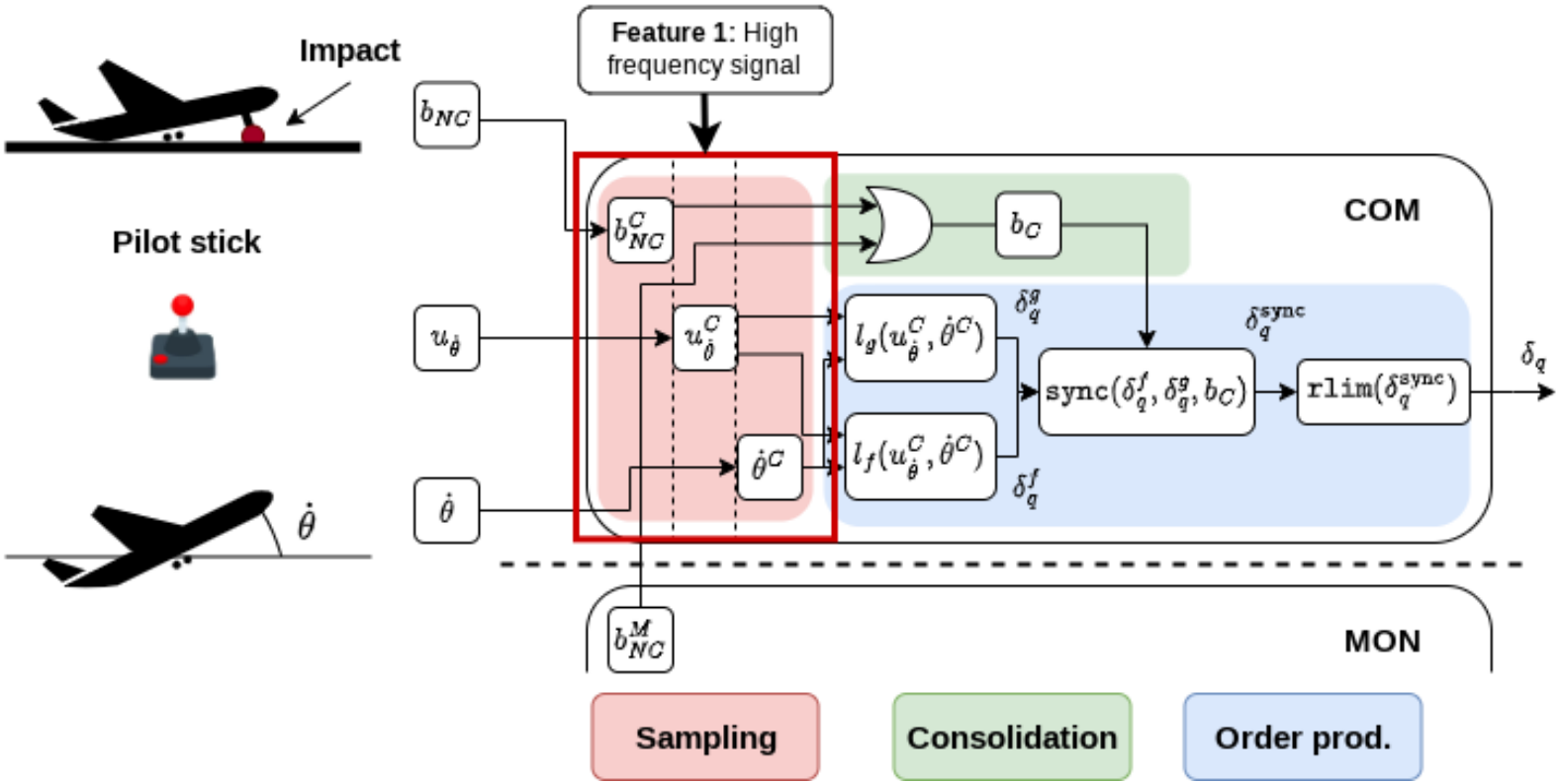
Coverage for various behavior descriptors



The percentage of terminal states reached throughout policy search. For SNAKE and US, coverage measures final robot position, and for Finger, it measures the final finger joint and hinge positions.

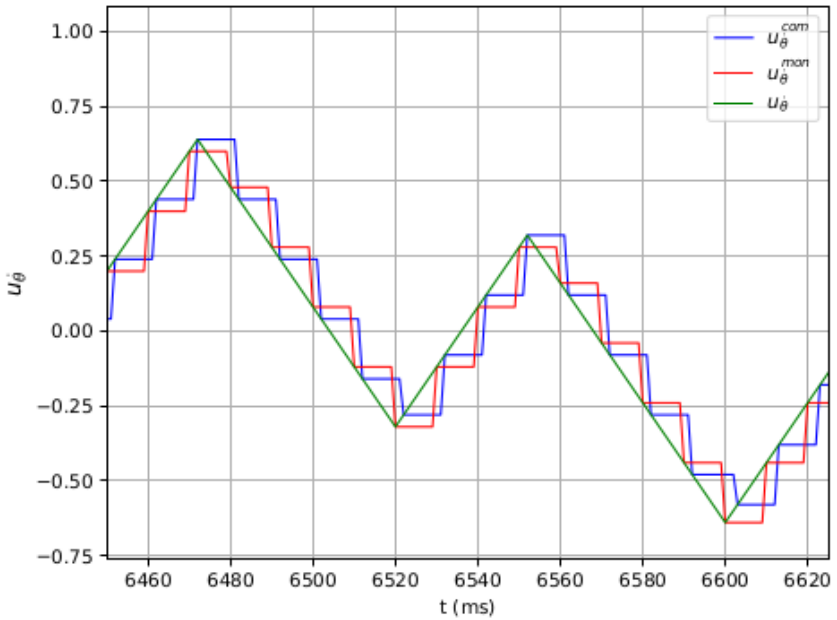
Exploration-Driven Reinforcement Learning for Avionic System Fault Detection

System weakness (1): High Frequency Signals

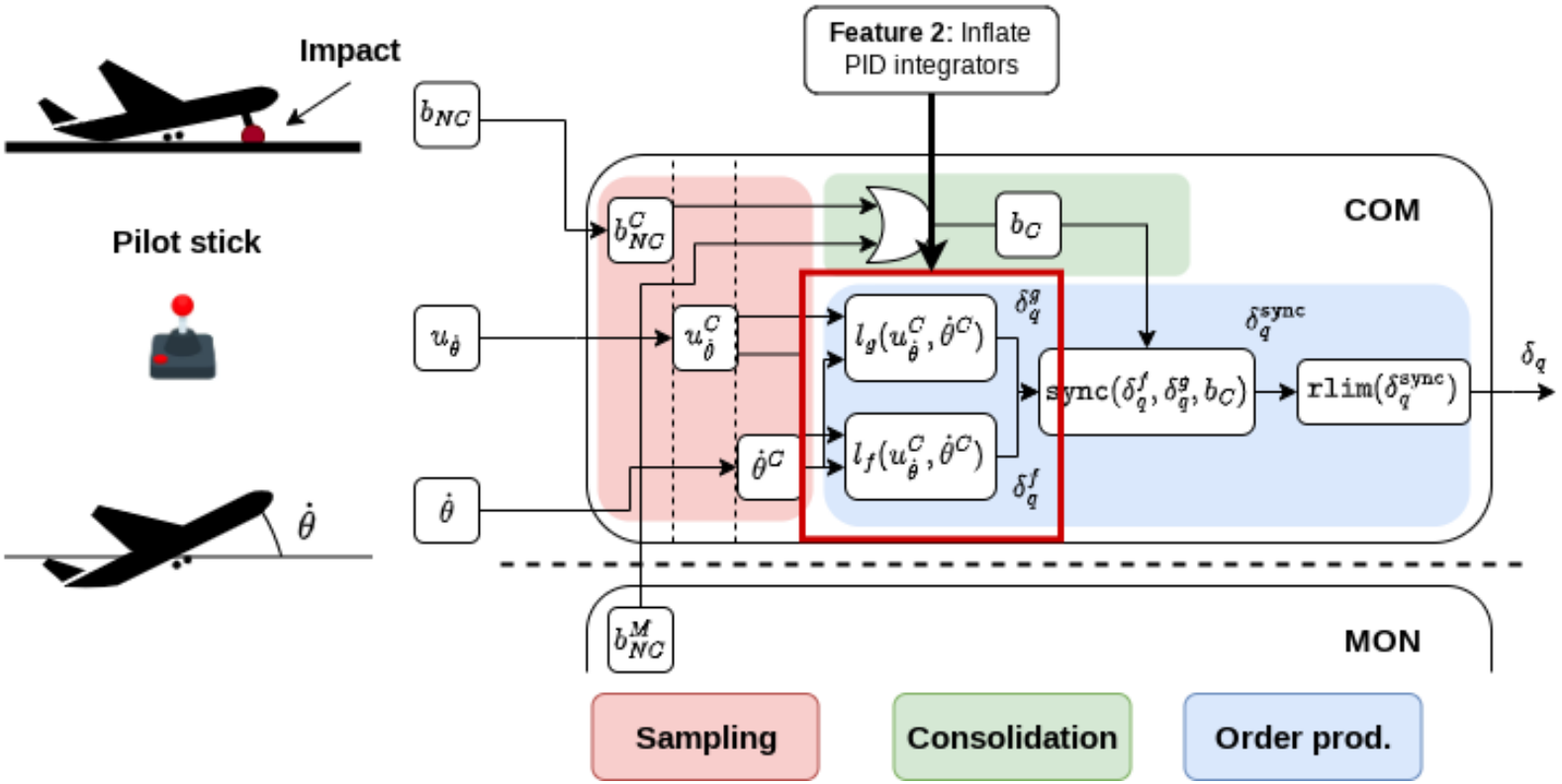


High Frequency Signals

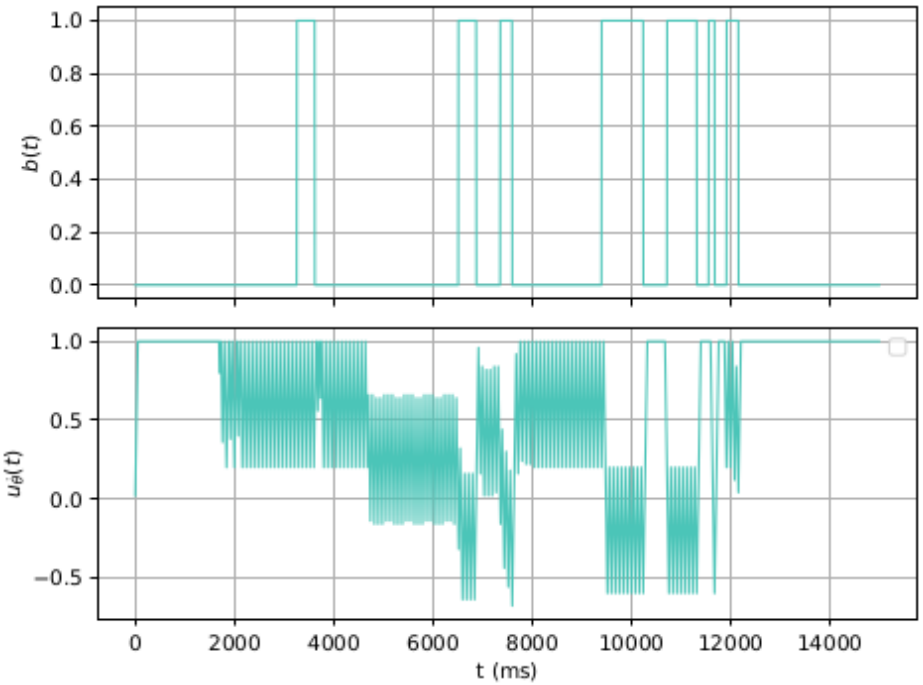
- $u_{\dot{\theta}}$ sampled every 10 ms
- COM sample $u_{\dot{\theta}}(t)$
- MON sample $u_{\dot{\theta}}(t + \Delta_t^{\text{COM/MON}})$
- High dynamic leads to a large gap

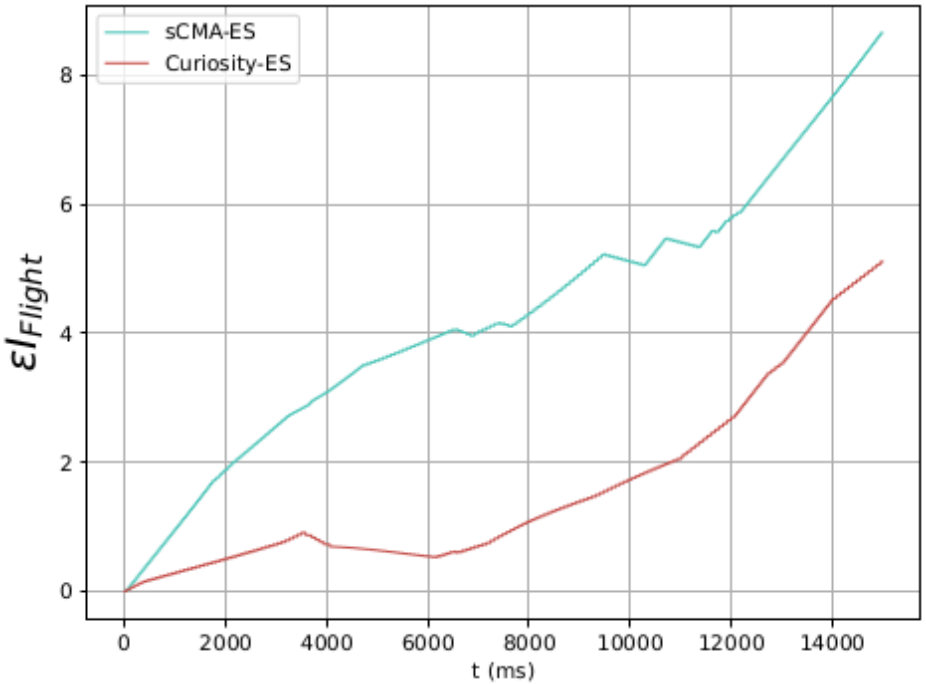


System weakness (2): Integrator Inflation

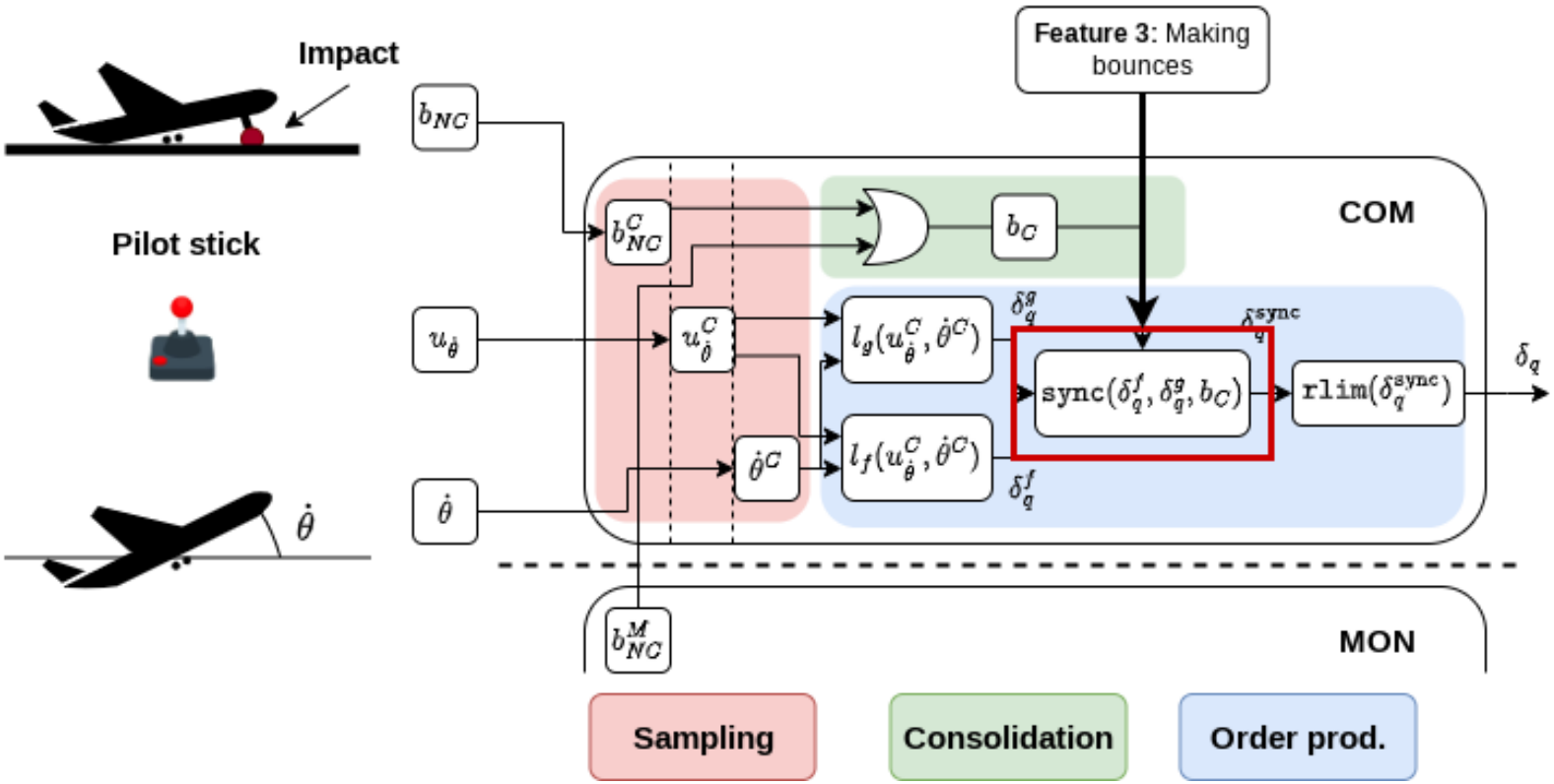


Inflate the PID integrator

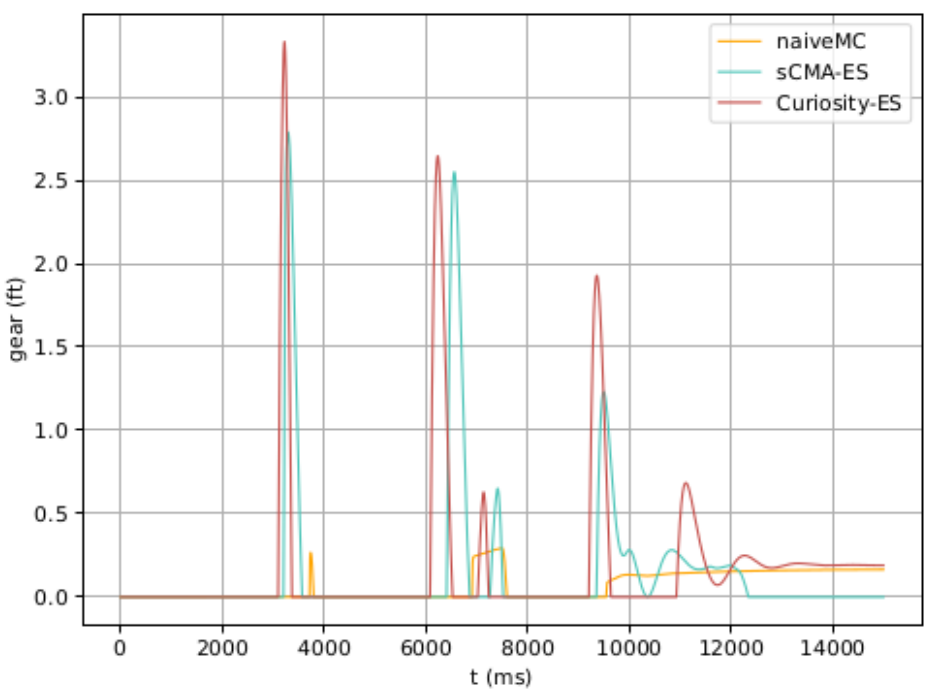


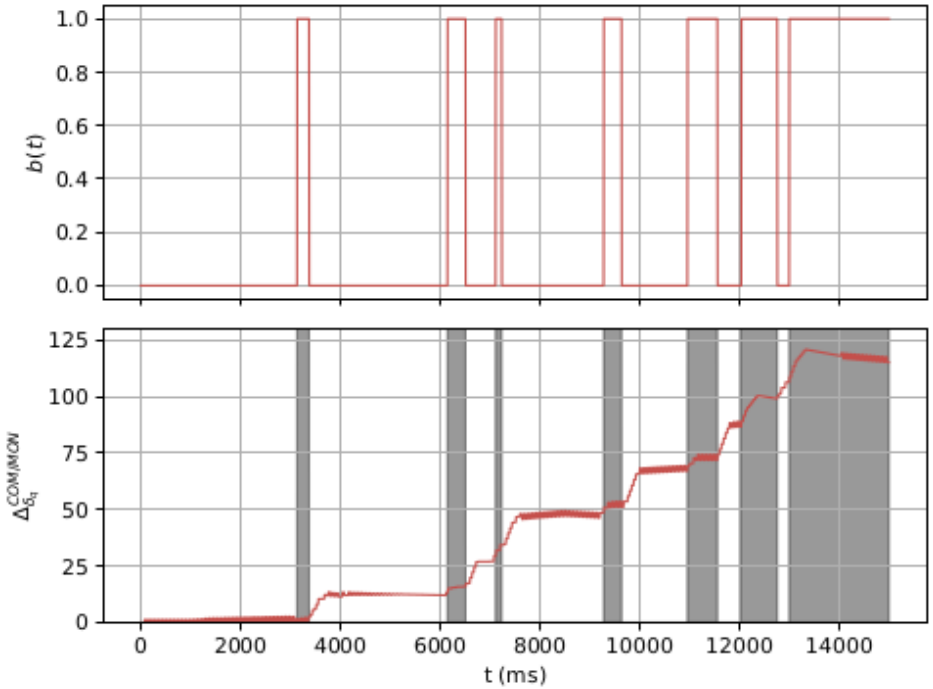


System weakness (3): Making Bounces

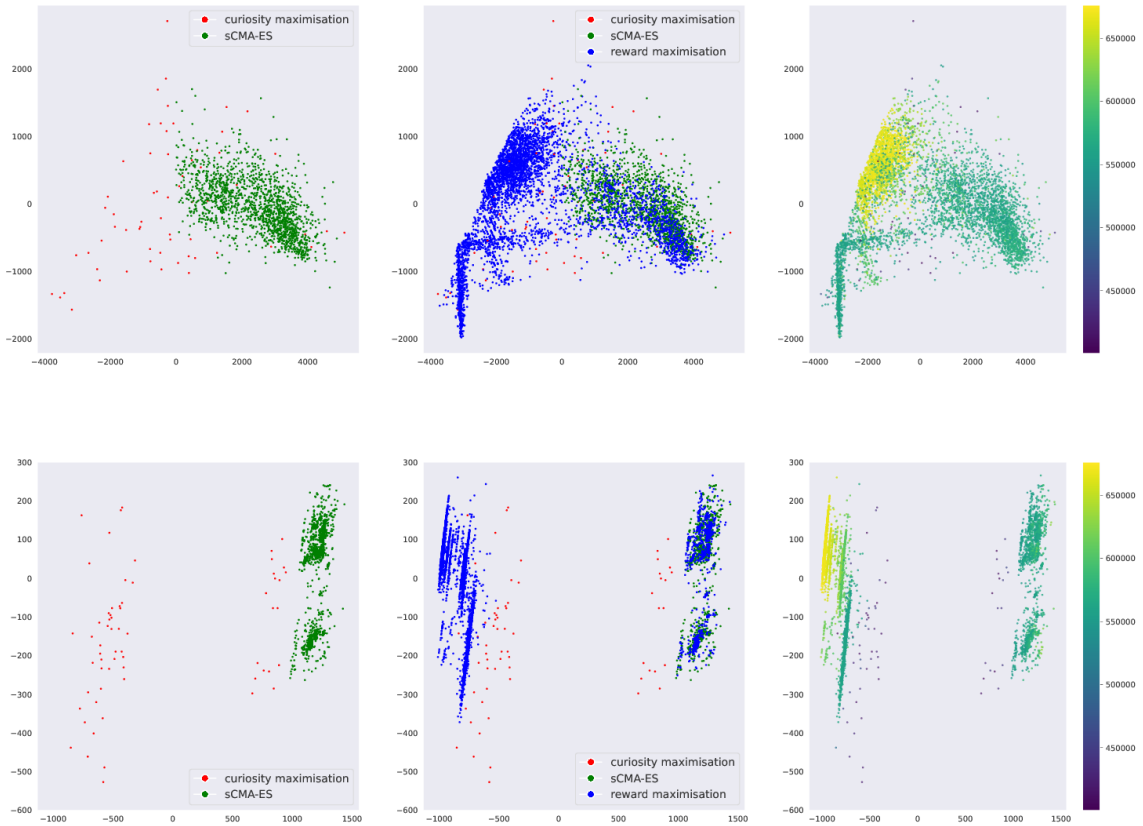


Making landing bounces





Diversity



(top) PCA on frequency signal of actions. (bottom) PCA on $n = 20$ most rewarding states. Relative importance of eigenvalues in the PCA projection: (top) $\lambda_1^{\%} = 44.51\%$, $\lambda_2^{\%} = 22.134\%$, $\lambda_3^{\%} = 3.54\%$. (Bottom) $\lambda_1^{\%} = 49.52\%$, $\lambda_2^{\%} = 32.77\%$, $\lambda_3^{\%} = 12.69\%$.

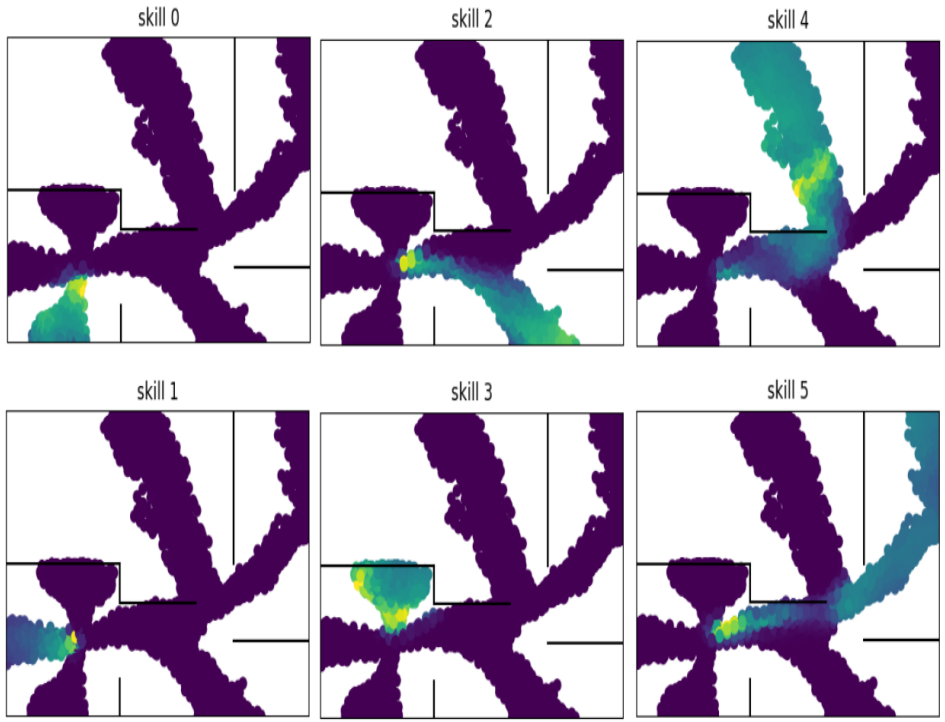
Learning **D**iverse Skills through **S**uccessor Representation

Successor Representation

$$p(s_2|s_1,\pi,z) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t \mathbb{P}\left(s_{t+1}=s_2 \;\middle|\; \begin{array}{l} s_0 = s_1, \\ a_{t+1}\sim\pi(s_{t+1},z)\end{array}\right)$$

In practice we estimate: $m_\phi^\pi(s_1,s_2,z) \approx p(s_2|s_1,\pi,z)/p(s_2)$

SR Visualization



Learning **D**iverse Skills through **S**uccessor Representation

Lower bound on MI

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

LEADS Objective

Objective

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

Novelty

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

LEADS quantitative coverage

Method	Easy (%)	Umaze (%)	Hard (%)	Fetch Reach (%)	Finger (%)	Fetch Slide (%)
RND	76.6 ± 7.3	39.6 ± 5.4	30.8 ± 4.3	17.6 ± 2.7	21.3 ± 1.6	21.6 ± 3.2
DIAYN	81.4 ± 8.6	55.0 ± 8.2	43.8 ± 4.5	85.6 ± 8.7	53.8 ± 12.3	52.3 ± 8.5
SMM	100.0 ± 0.0	70.6 ± 3.7	53.4 ± 0.5	22.3 ± 5.5	31.2 ± 1.4	53.3 ± 3.4
NGU	86.8 ± 8.4	73.4 ± 6.0	57.2 ± 8.3	53.4 ± 4.5	76.4 ± 5.6	57.8 ± 4.5
LSD	99.8 ± 0.4	79.8 ± 4.5	70.0 ± 0.6	71.9 ± 5.2	69.8 ± 8.8	61.5 ± 5.2
CSD	97.4 ± 3.4	79.8 ± 5.4	64.0 ± 6.2	83.2 ± 0.5	96.2 ± 9.1	63.0 ± 2.8
METRA	92.8 ± 4.2	78.0 ± 5.3	54.8 ± 9.5	82.5 ± 1.5	83.4 ± 7.5	50.7 ± 2.2
LEADS	100.0 ± 0.0	$\mathbf{96.0 \pm 4.3}$	$\mathbf{90.8 \pm 8.6}$	89.7 ± 8.8	87.4 ± 4.6	$\mathbf{85.4 \pm 7.2}$

Table: Final coverage percentages for each method on each environments. Bold indicates when a single method is statistically superior to all other methods ($p < 0.05$).

Exploration by **R**unning **A**way Fro**M** the **P**ast

RAMP Definitions

Occupancy measure

$$\rho^{\pi}(s) = \mathop{\mathbb{E}}_{\substack{s_1 \sim \delta_0 \\ a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\frac{1}{T} \sum_{t=1}^{T} \mathbb{1}_s(s_t) \right]$$

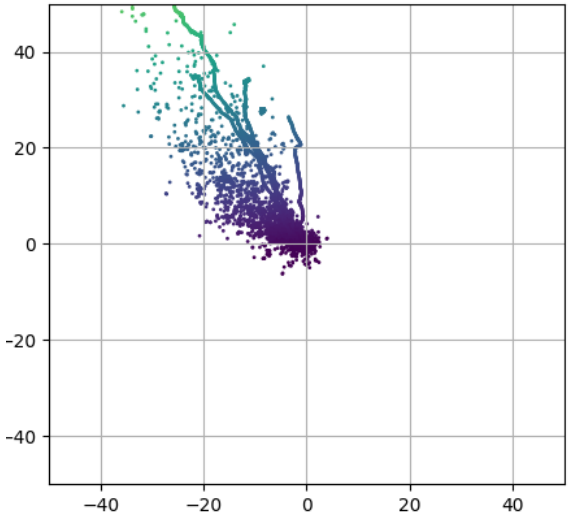
Replay buffer definition

$$\mu_n(s) = \beta \sum_{k=1}^{n} (1-\beta)^{n-k} \rho_k(s)$$

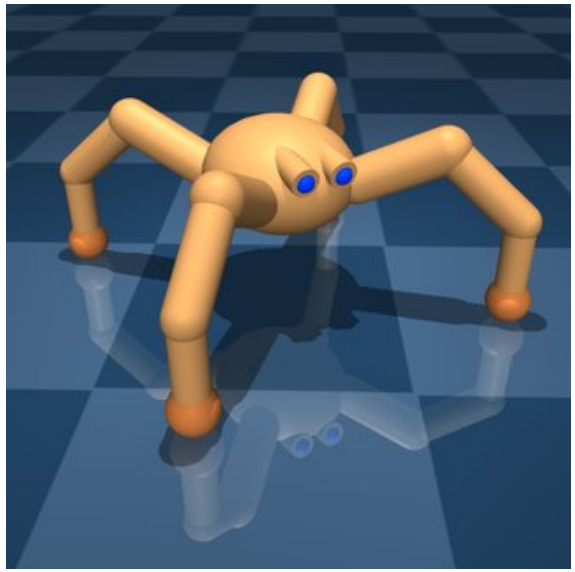
Recursive definition

$$\mu_{n+1}(s) = \beta \underbrace{\rho_{n+1}(s)}_{\text{The present}} + (1-\beta) \underbrace{\mu_n(s)}_{\text{The past}}$$

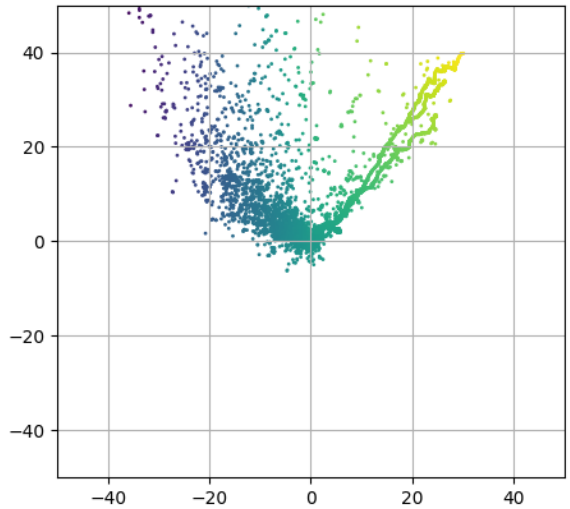
$T = 6 \times 10^5$



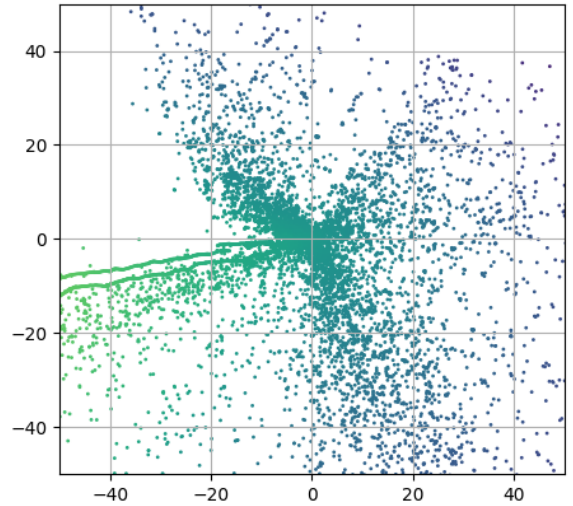
Ant



$T = 8 \times 10^5$



$T = 2 \times 10^6$



Exploration using RAMP

Algorithm	Ant	HalfCheetah	Hopper	Humanoid	Walker2d
APT	7.68 ± 0.9	**93.39 ± 3.39**	55.52 ± 3.75	54.73 ± 2.97	55.37 ± 2.83
AUX	4.53 ± 0.03	18.87 ± 0.21	5.65 ± 0.16	58.2 ± 0.24	11.65 ± 0.5
DIAYN	11.76 ± 0.61	58.18 ± 5.65	15.03 ± 8.89	70.68 ± 4.11	14.84 ± 1.34
ICM	3.26 ± 0.13	28.9 ± 1.55	41.63 ± 0.56	58.96 ± 0.87	33.81 ± 1.95
LSD	7.01 ± 2.15	30.43 ± 2.79	18.38 ± 5.27	69.89 ± 2.25	17.26 ± 2.32
METRA	23.46 ± 0.74	73.82 ± 3.0	37.5 ± 3.33	88.35 ± 5.05	36.88 ± 4.18
NGU	2.79 ± 0.07	25.53 ± 0.42	19.55 ± 0.62	44.02 ± 4.3	27.26 ± 2.01
RND	4.57 ± 0.07	19.1 ± 0.14	6.95 ± 0.51	57.91 ± 0.2	13.19 ± 0.31
SAC	4.4 ± 0.05	18.42 ± 0.24	5.83 ± 0.15	57.94 ± 0.27	13.11 ± 0.92
SMM	10.61 ± 1.28	58.91 ± 5.2	43.41 ± 14.93	32.64 ± 3.09	44.21 ± 13.19
RAMP$_{\mathscr{W}}$	**78.35 ± 13.45**	40.33 ± 6.69	**74.43 ± 12.14**	**90.5 ± 2.29**	**74.89 ± 11.71**

Table: Final relative mean coverage for the robot locomotion tasks. Bold indicates the highest mean per environment.

Exploring with extrinsic reward

Algorithm	Ant	HalfCheetah	Hopper	Humanoid	Walker2d
APT	1042 ± 69	10316 ± 138	3036 ± 442	3330 ± 739	2205 ± 313
AUX	5434 ± 263	11127 ± 243	2249 ± 465	3477 ± 706	4767 ± 91
ICM	4450 ± 402	11161 ± 163	3675 ± 126	3880 ± 491	5513 ± 191
NGU	975 ± 12	2976 ± 584	1360 ± 60	415 ± 93	1689 ± 96
RND	4427 ± 158	10901 ± 108	3084 ± 381	5103 ± 34	5723 ± 56
SAC	4972 ± 95	12197 ± 79	**3875 ± 40**	5163 ± 70	5650 ± 108
RAMP$_{KL}$	4768 ± 381	**13826 ± 361**	3636 ± 206	**5358 ± 49**	**5939 ± 524**
RAMP$_{\mathcal{W}}$	**7100 ± 47**	12997 ± 987	1036 ± 67	5342 ± 74	5933 ± 174

Table: Cumulative episodic return for the robot locomotion tasks. Bold indicates the highest mean per environment.

Kullback-Leibler Divergence

$$\mathrm{D}_{\mathrm{KL}}\left(\rho^{\pi}(s)\middle\|\beta\rho^{\pi}(s)+(1-\beta)\mu_n(s)\right)=\mathop{\mathbb{E}}_{s\sim\rho^{\pi}}\left[\underbrace{\log\left(\frac{\rho^{\pi}(s)}{(1-\beta)\mu_n(s)+\beta\rho^{\pi}(s)}\right)}_{r_{\mathrm{kl}}}\right]$$

Reward for KL

The reward r_{KL} is derived from a simple binary classification problem with the following labeling:

$$\begin{cases} s^{+}\sim\rho^{\pi} & \iff L=1 \\ s^{-}\sim\beta\rho^{\pi}+(1-\beta)\mu_n & \iff L=0 \end{cases} \tag{2}$$

$$r_{\mathrm{KL}}=\max_{f}\mathop{\mathbb{E}}_{\substack{s^{+}\sim\rho^{\pi}\\ s^{-}\sim\beta\rho^{\pi}+(1-\beta)\mu_n}}\left[\log\left(\sigma(f(s^{+}))\right)+\log\left(1-\sigma(f(s^{-}))\right)\right] \tag{3}$$

RAMP: a max-max problem

Uncertainty measure

$$\max_{\theta} \min_{w} \langle \rho^{\pi_\theta}, u_w \rangle$$

Novelty measure RAMP

$$\max_{\theta} \max_{\phi} \langle \rho^{\pi_\theta}, f_\phi \rangle$$

Wasserstein/Kantorovich Distance

$$\mathscr{W}(\rho^\pi, \beta\rho^\pi + (1-\beta)\mu_n) = \max_{\substack{f \\ \text{s.t.} \\ \|f\|_{\text{Lip}} \leq 1}} \mathop{\mathbb{E}}_{\substack{s^+ \sim \rho^\pi \\ s^- \sim (1-\beta)\mu_n + \beta\rho^\pi}} \left[f(s^+) - f(s^-) \right]$$

Reward for Wasserstein Distance

The reward $r_{\mathscr{W}}$ is derived from a constrained optimization problem:

$$\begin{aligned} r_{\mathscr{W}} = \min_f \max_\lambda & - \mathop{\mathbb{E}}_{s^+ \sim \rho^\pi} \left[f(s^+) \right] + \mathop{\mathbb{E}}_{s^- \sim \beta\rho^\pi + (1-\beta)\mu_n} \left[f(s^-) \right] \\ & + \lambda \cdot \mathop{\mathbb{E}}_{\substack{s \sim (1+\beta)\rho^\pi + (1-\beta)\mu_n \\ a \sim \pi(\cdot|s) \\ s' \sim P(\cdot|s,a)}} \left(\max \left(|f_\phi(s) - f_\phi(s')| - 1, -\epsilon \right) \right) \end{aligned} \tag{4}$$

Wasserstein Distance & Optimal Transport

Wasserstein Distance

$$W_p(\rho_1, \rho_2) = \left(\inf_{\gamma \in \Gamma(\rho_1, \rho_2)} \int_{\mathcal{X} \times \mathcal{X}} d(x, y)^p \, d\gamma(x, y) \right)^{1/p}$$

Valid Transport Plan

$$\int_{\mathcal{Y}} \gamma(x, y) \, dy = \rho_1(x) \quad \text{and} \quad \int_{\mathcal{X}} \gamma(x, y) \, dx = \rho_2(y)$$

	ρ_2					
ρ_1	γ_{11}	γ_{12}	γ_{13}	$\cdots$	γ_{1n}	$\rho_1(x_1)$
	γ_{21}	γ_{22}	γ_{23}	$\cdots$	γ_{2n}	$\rho_1(x_2)$
	γ_{31}	γ_{32}	γ_{33}	$\cdots$	γ_{3n}	$\rho_1(x_3)$
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
	γ_{m1}	γ_{m2}	γ_{m3}	$\cdots$	γ_{mn}	$\rho_1(x_m)$
	$\rho_2(y_1)$	$\rho_2(y_2)$	$\rho_2(y_3)$	$\cdots$	$\rho_2(y_n)$	

Theorem (reward improvement for KL Divergence)

Given policy π, let ε_1 be the approximation error of $\hat{r}_{D_{KL}}$, i.e. $\|\hat{r}_{D_{KL}} - r^{\pi}_{D_{KL}}\|_\infty \leq \varepsilon_1$.

Let π' be another policy and $\varepsilon_0 \in \mathbb{R}$ such that $\|\frac{\rho^{\pi'}}{\rho^{\pi}} - 1\|_\infty \geq \varepsilon_0$ ($\rho^{\pi'}$ is close to ρ^{π}).

Finally, let ε_2 measure how much π' improves on π for $\hat{r}_{D_{KL}}$: $\langle \rho^{\pi'}, \hat{r}_{D_{KL}} \rangle - \langle \rho^{\pi}, \hat{r}_{D_{KL}} \rangle = \varepsilon_2$.
If $\varepsilon_2 \geq 2\varepsilon_1 - \log(1-\varepsilon_0)$, then $D_{KL}\left(\rho^{\pi'} \| \rho^{\pi'}\beta + (1-\beta)\mu_n\right) \geq D_{KL}\left(\rho^{\pi} \| \rho^{\pi}\beta + (1-\beta)\mu_n\right)$.

Theorem (reward improvement for Wasserstein distance)

Given policy π, let ε_1 be the approximation error of $\hat{r}_{\mathscr{W}}$, i.e. $\|\hat{r}_{\mathscr{W}} - r^{\pi}_{\mathscr{W}}\|_\infty \leq \varepsilon_1$.
Let π' be another policy and ε_2 measure how much π' improves on π for $\hat{r}_{\mathscr{W}}$:
$\langle \rho^{\pi'}, \hat{r}_{\mathscr{W}} \rangle - \langle \rho^{\pi}, \hat{r}_{\mathscr{W}} \rangle = \varepsilon_2$. If $\varepsilon_2 \geq 2\varepsilon_1(1+\beta)$, then
$\mathscr{W}(\rho^{\pi'}, \beta\rho^{\pi'} + \mu_n(\beta - 1)) > \mathscr{W}(\rho^{\pi}, \beta\rho^{\pi} + \mu_n(\beta - 1))$.

Future work

Credit assignment

Answer the question: Why does my policy work?

Inverse RL

$$Q^{z_r}(s,a) = \phi(s,a)^T \underbrace{\mathbb{E}_{s\sim\mathcal{D}_{\text{training}}}[\psi(s)r(s)]}_{z_r} \qquad \min_r D(\rho^{\pi_{z_r}}, \rho^{\text{accident}})$$

What rewards should be maximized to achieve this incident?

Definitions

Occupancy Measure

$$\mu^\pi(s) = (1-\gamma) \mathop{\mathbb{E}}_{\substack{s_0 \sim \delta_0 \\ a_t \sim \pi(\cdot|s_t) \\ s_{t+1} \sim P(\cdot|s_t, a_t)}} \left[\sum_{t=0}^{\infty} \gamma^t \mathbb{1}_s(s_{t+1}) \right]$$

Tabular Representation

$$\mu^\pi = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t (P^\pi)^{t+1} \delta_0$$

Stationary Distribution

$$\rho^\pi = \lim_{t \to +\infty} (P^\pi)^t \nu$$

Successor State Framework

Model Construction

$$M^\pi = (1-\gamma)\sum_{t\geq 0}^{\infty}\gamma^t (P^\pi)^{t+1}$$

Matrix Form

$$M^\pi = \begin{bmatrix} \mu^\pi(s_0 \mid s_0) & \cdots & \mu^\pi(s_n \mid s_0) \\ \vdots & \ddots & \vdots \\ \mu^\pi(s_0 \mid s_n) & \cdots & \mu^\pi(s_n \mid s_n) \end{bmatrix}$$

Forward-Bellman Operator

$$M^\pi = \mathcal{T}^\pi(M^\pi) = (1-\gamma)P^\pi + \gamma P^\pi M^\pi$$

Probability Interpretation

- $(1-\gamma)P^\pi$: reach s_2 directly
- $\gamma P^\pi M^\pi$: reach s_2 via next state

Continuous State Spaces

Successor State Definition

$$m^\pi(s_a|s_d, a_d) =$$
$$(1-\gamma)\sum_{t=0}^{\infty}\gamma^t\mathbb{P}\left(s_{t+1}=s_a \;\middle|\; \begin{array}{l} s_0 = s_d, a_0 = a_d \\ a_{t+1}\sim\pi(s_{t+1}) \end{array}\right)$$

Linear Parameterization

$$m^\pi_\phi(s_2|s_1,a_1) \approx f_\phi(s_1,a_1,s_2)\,\rho(s_2)$$

Operator Definition

$$\begin{aligned} m^\pi(\cdot|s,a) &= \mathcal{T}^\pi(m^\pi)(\cdot|s,a) \\ &= (1-\gamma)\mathbb{P}(\cdot|s,a) + \gamma \mathop{\mathbb{E}}_{\substack{s'\sim\mathbb{P}(\cdot|s,a)\\ a'\sim\pi(\cdot|s')}} [m^\pi(\cdot|s',a')] \end{aligned}$$

Room for improvement

- Measure on $\mathcal{S}$
- Norm used for $(m - \mathcal{T}^\pi m)$
- The way we bring m on $\mathcal{T}^\pi m$
- The operator we use

Estimation and Optimization

Squared Norm

$$\| m^{\pi_1} - m^{\pi_2} \|_\rho^2 = \mathop{\mathbb{E}}_{\substack{s_1 \sim \rho \\ s_2 \sim \rho}} \left[\left(f_1(s_1, s_2) - f_2(s_1, s_2) \right)^2 \right] \quad (5)$$

Loss Function

$$\begin{aligned} \mathscr{L}(\phi) &= \frac{1}{2} \| m_\phi - \mathscr{T}^\pi m_\phi \|_\rho^2 \\ &= \frac{1}{2} \| m_\phi \|_\rho^2 - \langle m_\phi, \mathscr{T}^\pi m_\phi \rangle + \frac{1}{2} \| \mathscr{T} m_\phi \|_\rho^2 \end{aligned}$$

Dot Product

$$\langle m^{\pi_1}, m^{\pi_2} \rangle_\rho = \int_{\mathcal{S}} \int_{\mathcal{S}} f_1(s_1, s_2) \cdot f_2(s_1, s_2) \rho(\mathrm{d}s_1) \rho(\mathrm{d}s_2) \quad (6)$$

Gradient

$$\begin{aligned} \nabla_\phi \mathscr{L}(\phi) &= \frac{1}{2} \nabla_\phi \| m_\phi \|_\rho^2 - \langle \nabla_\phi m_\phi, \mathscr{T}^\pi m_\phi \rangle \\ &= \int_{\mathcal{S}} \int_{\mathcal{S}} \nabla_\phi (f_\phi(s_1, s_2)) f_\phi(s_1, s_2) \rho(\mathrm{d}s_1) \rho(\mathrm{d}s_2) \end{aligned}$$

Stochastic Gradient Descent

Loss Function

$$L(\theta) = \mathbb{E}_{X \sim P}[\ell(X, \theta)]$$

Empirical Risk

$$\hat{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, \theta)$$

Stochastic Differential Equation (SDE) of SGD

$$d\theta_t = - \underbrace{\nabla \hat{L}_n(\theta_t)}_{\text{drift term}} dt + \underbrace{\left(\nabla \hat{L}_n(\theta_t) - \frac{1}{B} \sum_{i=1}^{B} \nabla \ell(x_i, \theta_t) \right)}_{\text{noise term}} dt$$

L2 Loss Minimization and Gaussian Assumption

Model Assumption

$$y = \underbrace{f_\theta(x)}_{\text{prediction}} + \underbrace{\varepsilon}_{\text{error}}$$

where $\varepsilon \sim \mathcal{N}(0, \sigma^2)$

Conditional Distribution

$$p(\hat{y}|x, \theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y - f_\theta(x))^2}{2\sigma^2}}$$

Maximum Likelihood Estimation

$$\theta^* = \arg\max_\theta \sum_{i=1}^n \log p(y_i|x_i, \theta)$$

Equivalence to L2 Minimization

$$\theta^* = \arg\min_\theta \sum_{i=1}^n (y_i - f_\theta(x_i))^2$$