Continual Learning of Control Primitives: Skill Discovery via Reset-Games

Kelvin Xu*[1] Siddharth Verma*[1] Chelsea Finn[2] Sergey Levine[1]

[1]UC Berkeley [2]Stanford University *equal contribution

Motivation: What Problems Will Agents Face In The Real World?

- (1) When we learn in the real world, agents must contend with non-episodic learning dynamics (no reset)
- (2) Complex temporally extended behavior can be exceedingly hard to to acquire with simple exploration
- On the surface, these two problems seem unconnected, but the non-episodic learning problem can be mitigated by learning skills to reset the system
- What if those same skills also help accelerate downstream learning?
- We study this question and propose Learning Skills from Resets (**LSR**) as a means of tackling these two challenges

The Reset-Free Problem Statement

- An RL problem is defined on a Markov decision process (MDP) $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}_s, r, \gamma, \mathcal{P}_0)$, $\mathcal{S}$ is a set of continuous states and $\mathcal{A}$ is a set of continuous actions, $\mathcal{P}_s : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the transition probability density, $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function, γ is the discount factor and $\mathcal{P}_0$ is the initial state distribution
- The γ-discounted return $R(\tau)$ of a trajectory $\tau = (s_0, a, \ldots s_{T-1}, a_{T-1})$ is $\sum_{t=0}^{T-1} \gamma^t r(s_t, a_t)$
- In **episodic**, finite horizon tasks of length T, learn a policy $\pi_\theta : \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ that maximizes the objective:

$$J(\pi_\theta) = \mathbb{E}_{\pi_\theta, \mathcal{P}_0, \mathcal{P}_s}[R(\tau)] \qquad (1)$$

- The **reset-free** problem assumes you want to optimize Eq. 1 that you cannot sample and reset from $\mathcal{P}_0$

Learning Skills via Mutual Information Maximization

- Skills can be represented by conditioning the policy on a latent skill z, which is sampled from $\mathcal{P}(z)$ and held constant during execution
- To learn which z should correspond to a which behavior, can use the mutual information (MI) between skills and the states they visit:

$$\mathcal{I}(s; z) = \mathcal{H}(s) - \mathcal{H}(s|z) \qquad (2)$$

- Maximizing $\mathcal{I}(s; z)$ entails maximizing the state entropy (high state coverage) while minimizing conditional state entropy (high predictability for each z):

$$\mathcal{I}(s; z) - \mathcal{I}(a; s, z) = \mathbb{E}_\pi\left[\log \frac{p(z|s)}{p(z)} - \log \frac{\pi(a|s, z)}{\pi(a)}\right]$$
$$\geq \mathbb{E}_\pi[\log q_\omega(z|s) - \log p(z) - \log \pi(a|s, z)] = \mathcal{G}(\theta, \omega),$$

where we replace $p(z|s)$ with an approximate learned discriminator $q_\omega(z|s)$ with parameters ω to obtain a variational lower bound. We can maximize $\mathcal{G}(\theta, \omega)$ with RL using the pseudo-reward:

$$r_{\text{skill}}(\pi, q_\omega) = \log q_\omega(z|s) - \log p(z) - \log \pi(a|s, z). \qquad (3)$$

The Reset Game

- Aim to have a forward policy $\pi_\theta(a|s)$ and a set of reset skills $\pi_\phi^{\text{reset}}(a|s, z)$
- We propose a general sum formulation of this problem which we call the "Reset Game"

$$\boxed{\max_{\pi_\theta} \mathcal{J}^{\text{forward}}(\pi_\theta, \pi_\phi^{\text{reset}}), \quad \max_{\pi_\phi^{\text{reset}}} \mathcal{J}^{\text{reset}}(\pi_\theta, \pi_\phi^{\text{reset}})} \qquad (4)$$

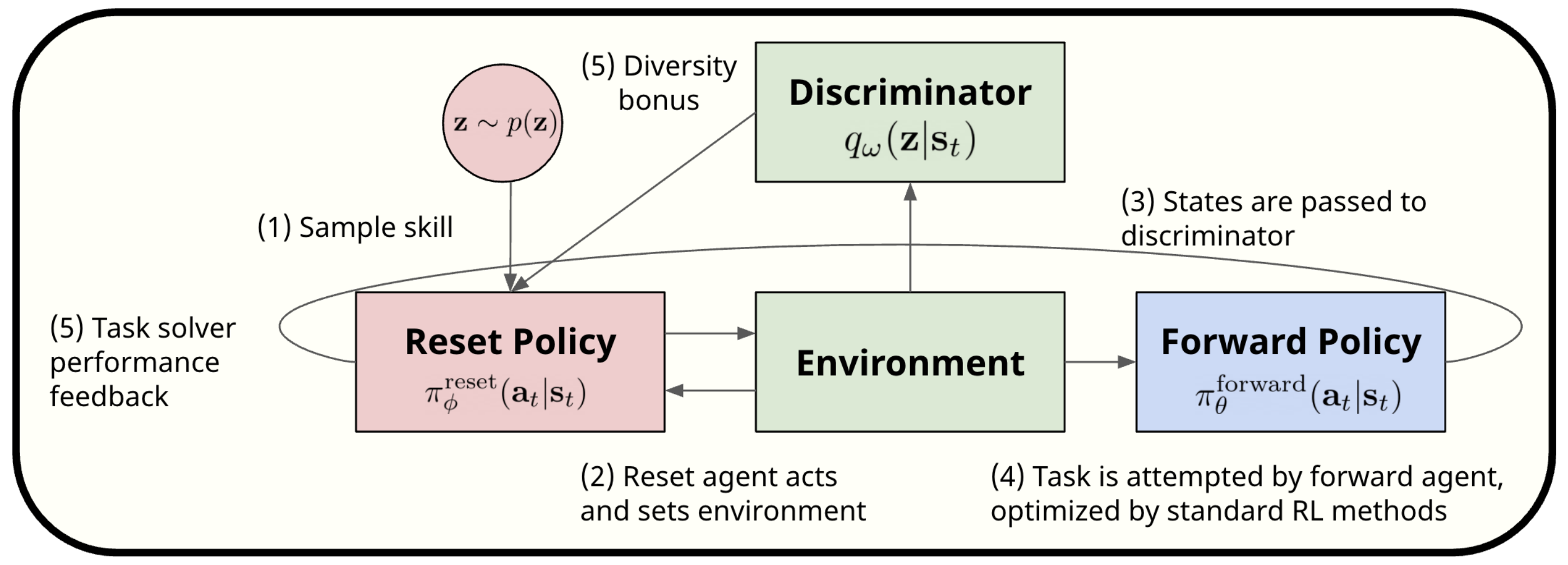
Learning Skills from Resets (LSR)

- Instantiate a practical version of this algorithm using an MI based reward balanced by task reward using λ

$$\max_\phi \underbrace{\mathbb{E}_{s_t', a' \sim \pi_\phi^{\text{reset}}}\left[\sum_{t=0}^{T_{reset}-1} \gamma^t r_{skill}(a', s_t') - \lambda\, \mathbb{E}_{\pi_{\theta^*}}\left[\sum_{t=0}^{T-1} \gamma^t r(a, s_t)\right]\right]}_{\mathcal{J}^{\text{reset}}(\pi_\theta, \pi_\phi^{\text{reset}})} \qquad (5)$$

$$\text{such that } \theta^* = \arg\max_\theta \underbrace{\mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{T-1} \gamma^t r(a, s_t)\right]}_{\mathcal{J}^{\text{forward}}(\pi_\theta, \pi_\phi^{\text{reset}})}. \qquad (6)$$

Learning Skills from Resets (LSR)



$z \sim p(z)$
(5) Diversity bonus
Discriminator
$q_\omega(\mathbf{z}|\mathbf{s}_t)$
(1) Sample skill
(3) States are passed to discriminator
(5) Task solver performance feedback
Reset Policy
$\pi_\phi^{\text{reset}}(\mathbf{a}_t|\mathbf{s}_t)$
Environment
Forward Policy
$\pi_\theta^{\text{forward}}(\mathbf{a}_t|\mathbf{s}_t)$
(2) Reset agent acts and sets environment
(4) Task is attempted by forward agent, optimized by standard RL methods

Figure: An outline of our approach for learning the reset policy and forward policy $\pi_\phi^{\text{reset}}, \pi_\theta$

Experimental Domains

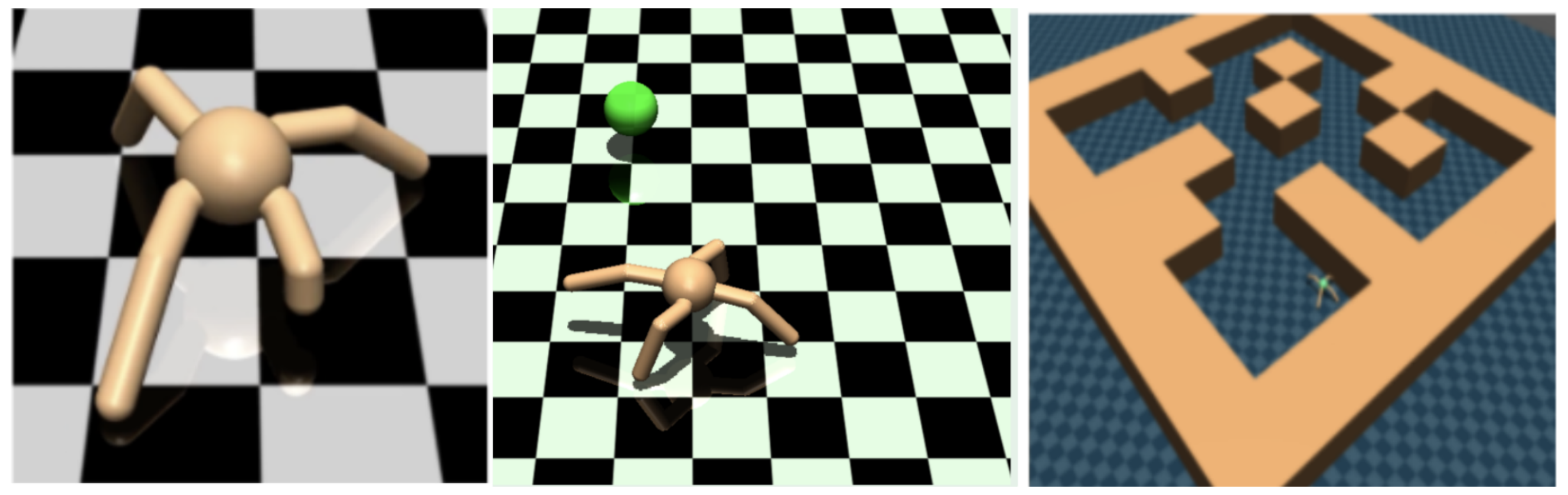


Figure: Diagram of the hierarchical locomotion tasks considered in this work. The agent must first acquire locomotion skills in a reset game (left), where the task policy must learn to walk to the origin. The skills are subsequently used as the action space for a hierarchical policy, which must learn navigation tasks (second and third image).

Improved Reset Free Learning

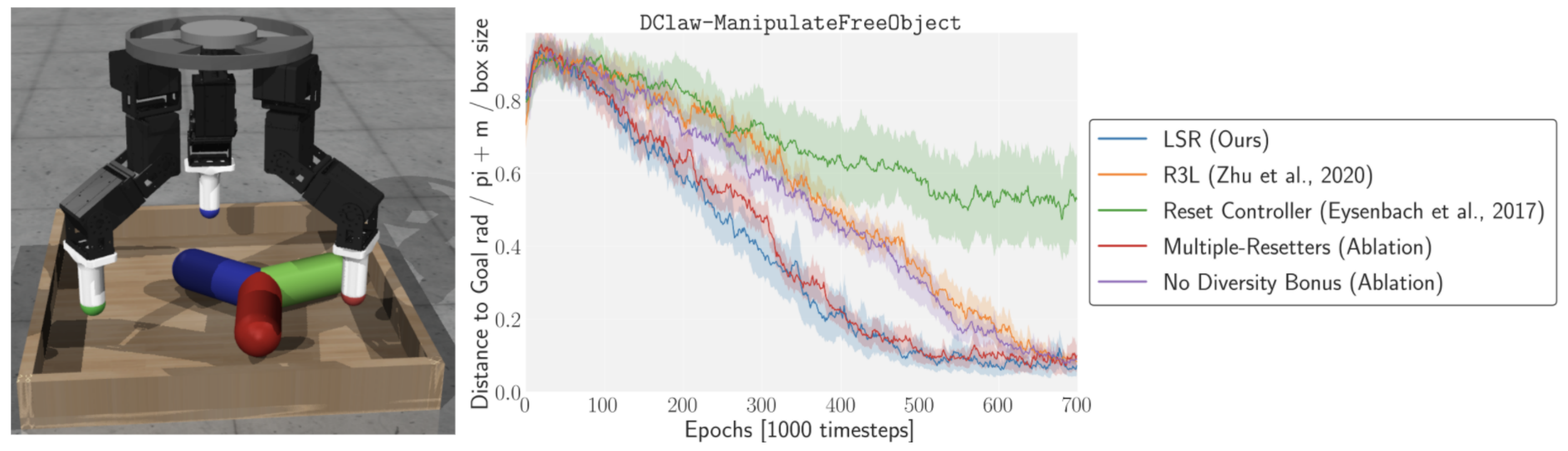


Figure: Reset-free learning comparison (lower is better). Our method (blue) outperforms prior methods (orange, green). Our ablations show that the most important aspect of our approach is having multiple resetters. This indicates that reset state convergence is crucial to good performance, in contrast to prior methods [Eysenbach et al., 2017] that learn a explicit reset policy.

Downstream Hiearchical Learning

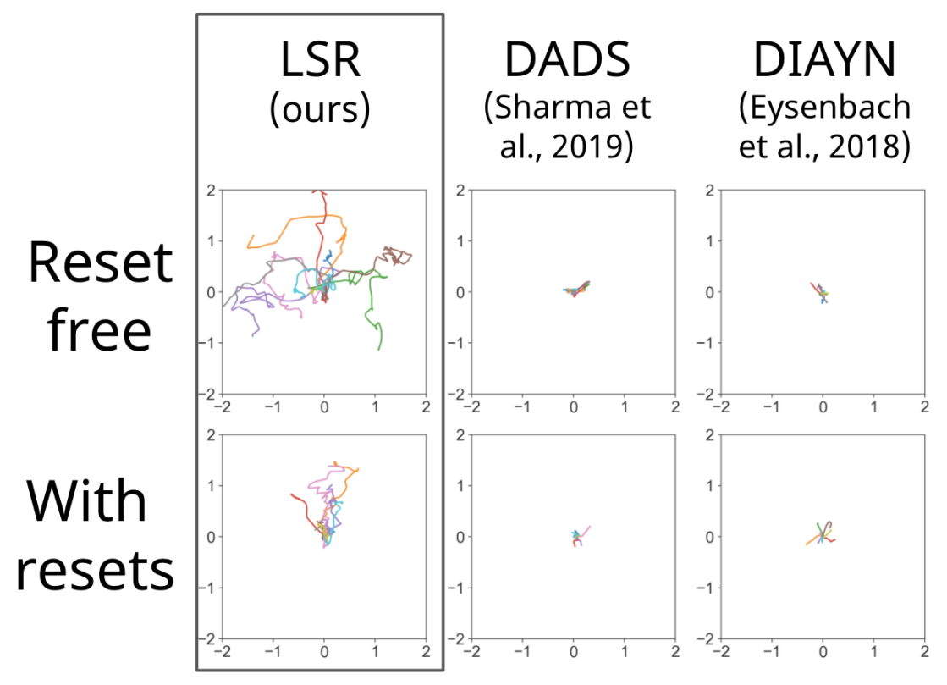


Figure: We visualize the (x, y) trajectories of learned skills (left). In contrast to prior approach Eysenbach et al. [2018], Sharma et al. [2019].

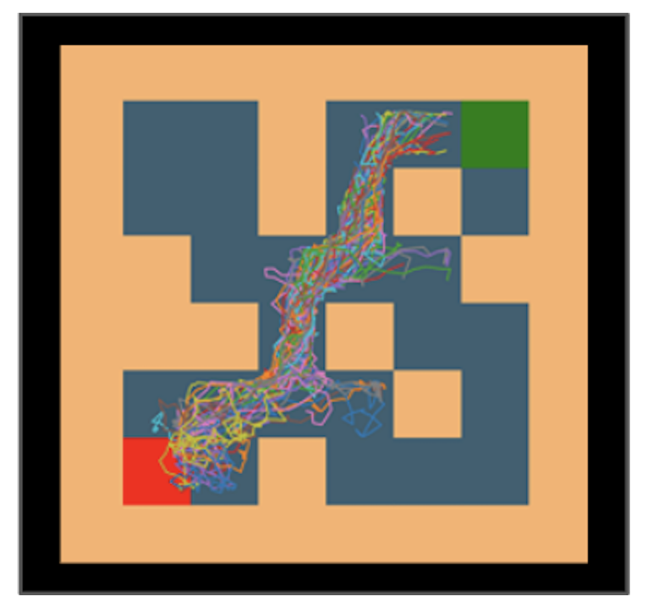


Figure: A visualization of the paths taken by a hierarchical policy using our learned reset skills.

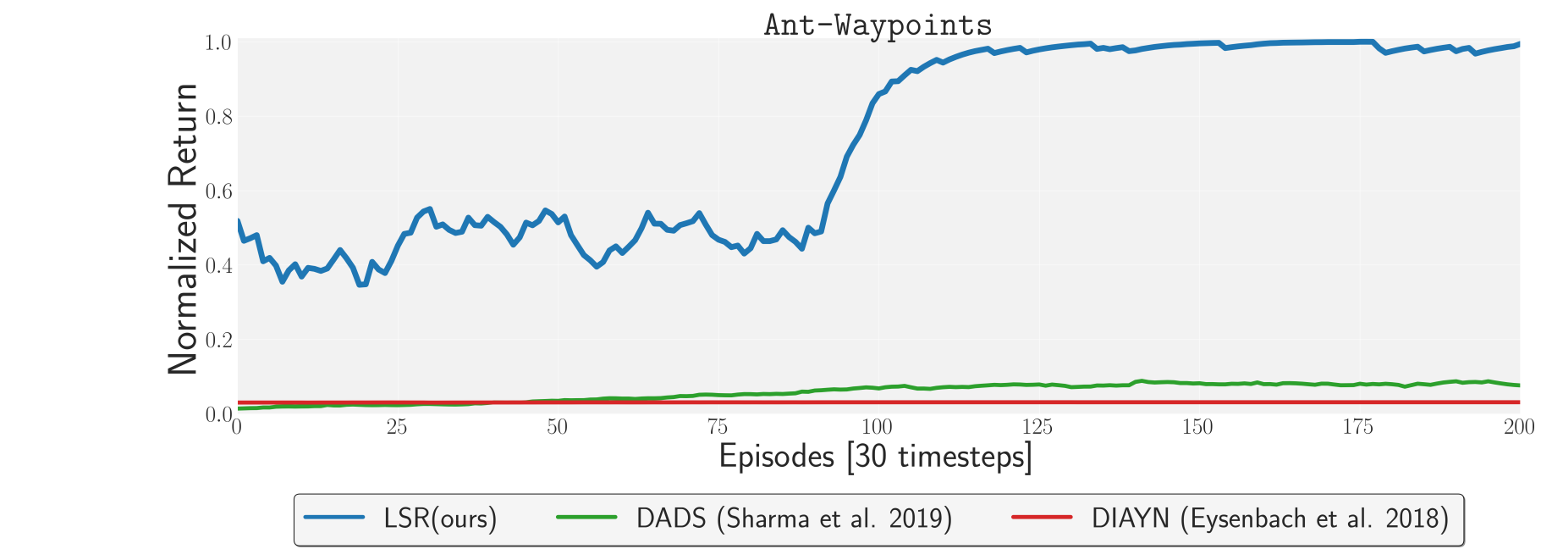


Figure: Performance on the Ant-Waypoints environment compared to DADS and DIAYN. DADS and DIAYN makes small progress on this simpler domain (orange), but our method (blue) is still able to successfully navigate between the waypoints more quickly.

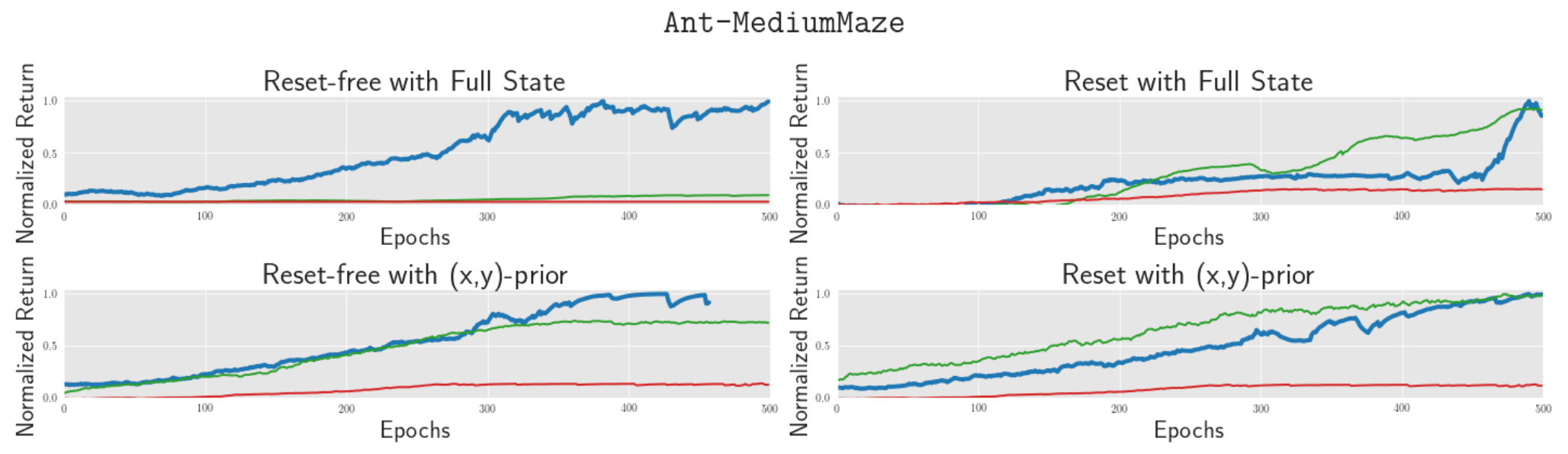


Figure: A evaluation of the skills learned when varying the state representation ((x, y)-prior or full state) and initial state distribution (reset-free or with resets). We normalize the return of plot by the best performing algorithm. We find that with the (x, y)-prior is necessary to allow prior methods to perform comparably to LSR.

Conclusions

- Across a range of tasks, learning a set of diverse reset skills can help enable non-episodic learning and aid a downstream hierarchical learner.

References

B. Eysenbach, S. Gu, J. Ibarz, and S. Levine. Leave no trace: Learning to reset for safe and autonomous reinforcement learning. *arXiv preprint arXiv:1711.06782*, 2017.

B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

A. Sharma, S. Gu, S. Levine, V. Kumar, and K. Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

{kelvinxu,vsiddharth,svlevine}@berkeley.edu, cbfinn@cs.stanford.edu

https://github.com/siddharthverma314/clcp-neurips-2020