Langevin Soft Actor-Critic: Efficient Exploration through Uncertainty-Driven Critic Learning

Haque Ishfaq * [1,2] ✉, Guangyuan Wang * [1,2], Sami Nur Islam [1,2], Doina Precup [1,2]

[1]Mila – Quebec AI Institute [2]McGill University

Motivations

- Explorations techniques are crucial for an agent to be able to solve novel complex problems.
- Existing actor-critic algorithms, which are popular for continuous control tasks, suffer from poor sample efficiency due to lack of principled exploration mechanism within them.
- Langevin Monte Carlo based Thompson sampling is a powerful and efficient approach for performing principled exploration in RL.

Challenges:

- **Multidimensional continuous action spaces:** naively selecting exact greedy actions based on Q posterior approximations is computationally intractable.
- **Value approximation errors:** Overestimation bias of Q-function; instability of LMC in DNN.

Our Contributions

- A novel way to perform Thompson Sampling in actor-critic algorithm through distributional critic learning and **adaptive Langevin Monte Carlo**.
- Enabling sampling from multimodal Q-posteriors using **parallel tempering** approach.
- Synthetic data generation using **diffusion Q action gradient** method.

Langevin Monte Carlo for Reinforcement Learning

- Langevin Monte Carlo update:

$$w_{k+1} = w_k - \eta_k \nabla L(w_k) + \sqrt{2\eta_k \beta^{-1}}\epsilon_k,$$

  - It approximately samples from $\pi_k \propto \exp(-\beta L(w))$.
- It is computationally efficient due to
  - it only needs to sample ϵ_k from isotropic Gaussian $\mathcal{N}(0, I)$.
  - it only needs to perform noisy gradient descent updates.

Preliminary

- Denote the entropy augmented cumulative return from s_t, by $G_t = \sum_{i=t}^{\infty} \gamma^i [r_i - \alpha \log \pi(a_i \mid s_i)]$.
- The soft Q-value of policy π is defined as $Q^\pi(s_t, a_t) \coloneqq r_t + \gamma \mathbb{E}[G_{t+1}]$.
- Define soft state-action return, a random variable, by $Z^\pi(s_t, a_t) \coloneqq r_t + \gamma G_{t+1}$.
- Observe that $Q^\pi(s, a) = \mathbb{E}[Z^\pi(s, a)]$.

Distributional Critic

- Instead of the expected state-action return $Q^\pi(s, a)$, we aim to model the distribution of the random variable $Z^\pi(s, a)$.
- We define, value distribution function, $\mathcal{Z}^\pi(Z^\pi(s, a) \mid s, a) : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(Z^\pi(s, a))$ as a mapping from (s, a) to a distribution over the soft state-action return $Z^\pi(s, a)$.
- We define the distributional Bellman operator in the maximum entropy framework as

$$\mathcal{T}^\pi Z^\pi(s, a) \stackrel{D}{\coloneqq} r + \gamma(Z^\pi(s', a') - \alpha \log \pi(a' \mid s')). \tag{1}$$

- We model the value distribution function $\mathcal{Z}_\psi(\cdot \mid s, a)$ as Gaussian distribution $\mathcal{Z}_\psi(\cdot \mid s, a) = \mathcal{N}(Q_\psi(s, a), \sigma_\psi(s, a)^2)$.

Algorithm: Langevin Soft Actor-Critic (LSAC)

Distributional Critic Learning with Adaptive Langevin Monte Carlo:

- Distributional Critic Loss Function:

$$L_{\mathcal{Z}}(\psi) \coloneqq \mathbb{E}_{(s,a)\sim B} D_{\mathrm{KL}}(\mathcal{T}^{\pi_\phi} \mathcal{Z}_{\bar{\psi}}(s, a) \| \mathcal{Z}_\psi(s, a)), \tag{2}$$

- Under some mild assumptions, the posterior over Q_ψ is of the form $\exp(-L_{\mathcal{Z}}(\psi))/Z$, where Z is the partition function.
- Approximate sampling from the posterior using adaptive LMC:

$$\boxed{\psi_{k+1} \leftarrow \psi_k - \eta(\nabla_\psi L_{\mathcal{Z}}(\psi_k) + a\zeta_{\psi_k}) + \sqrt{2\eta\beta^{-1}}\epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, I_d).} \tag{3}$$

  with adaptive preconditioner ζ_k is defined as $\zeta_{\psi_k} \coloneqq m_k \oslash \sqrt{v_k + \lambda \mathbf{1}}$ where,

$$m_k = \alpha_1 m_{k-1} + (1 - \alpha_1)\nabla L_{\mathcal{Z}}(\psi_k) \quad \text{and} \quad v_k = \alpha_2 v_{k-1} + (1 - \alpha_2)\nabla L_{\mathcal{Z}}(\psi_k) \odot \nabla L_{\mathcal{Z}}(\psi_k).$$

Parallel Tempering and Multimodal Q Posteriors:

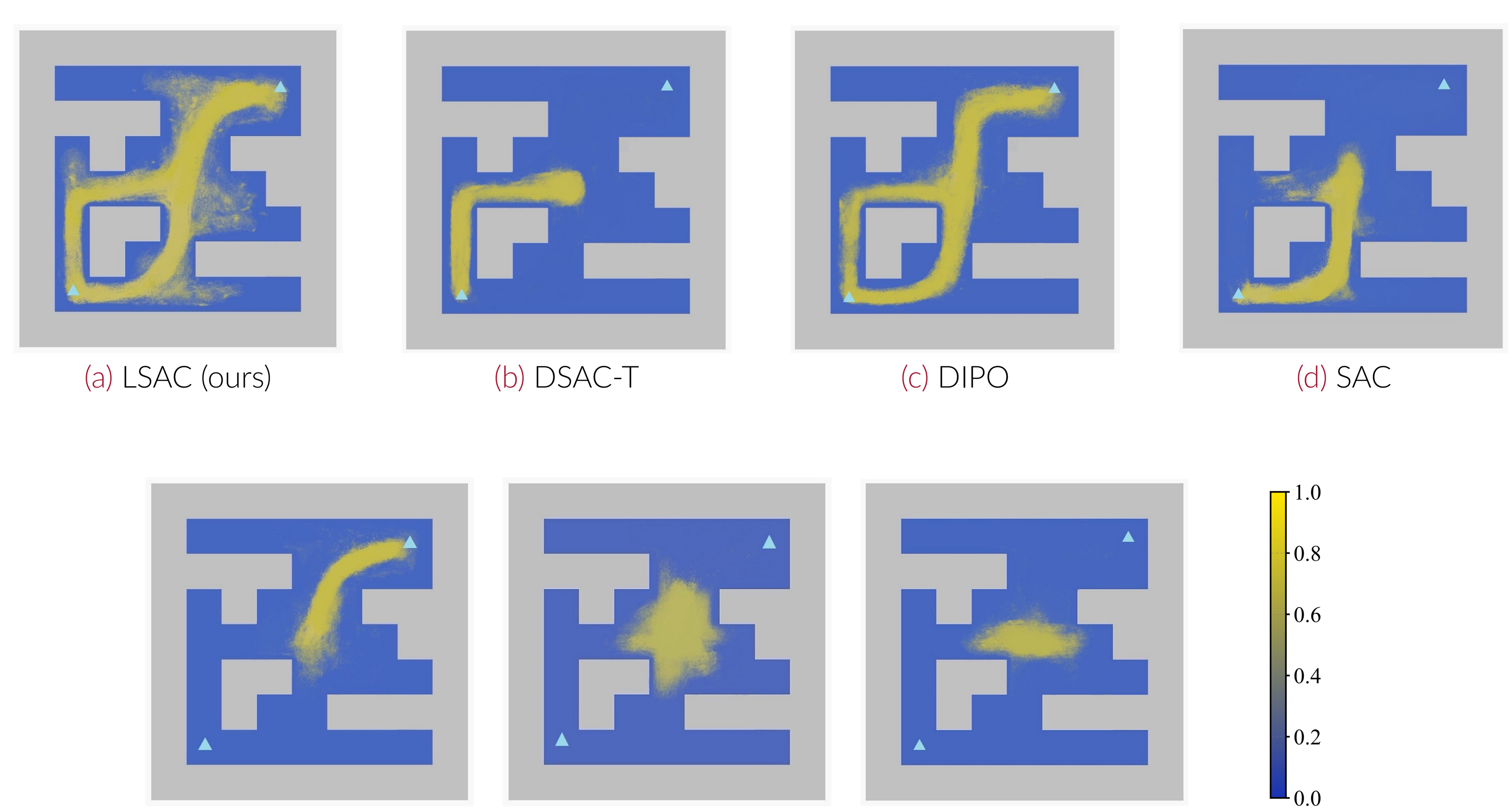
- Performing naive LMC to approximately sample from multimodal Q posterior can converge very slowly due to its slow mixing rate.
- We use a simplified version of parallel tempering with all replicas having same temperature for efficient exploration in the parameter space.
- By running multiple LMC chains $\Psi_Q = \{\psi^{(i)}\}_{i=1}^n$, we can sample Q-functions for critics from distinct modes of the multimodal posterior while ensuring faster convergence and mixing time.

Diffusion Q Action Gradient:

- We use diffusion synthesized state-action samples regularized with Q action gradients.
- This ensures that the synthetic actions are not only diverse but also accurately reflect regions of high Q value

$$a \leftarrow a + \gamma \nabla_a Q_{\psi^{(i)}}(s, a).$$

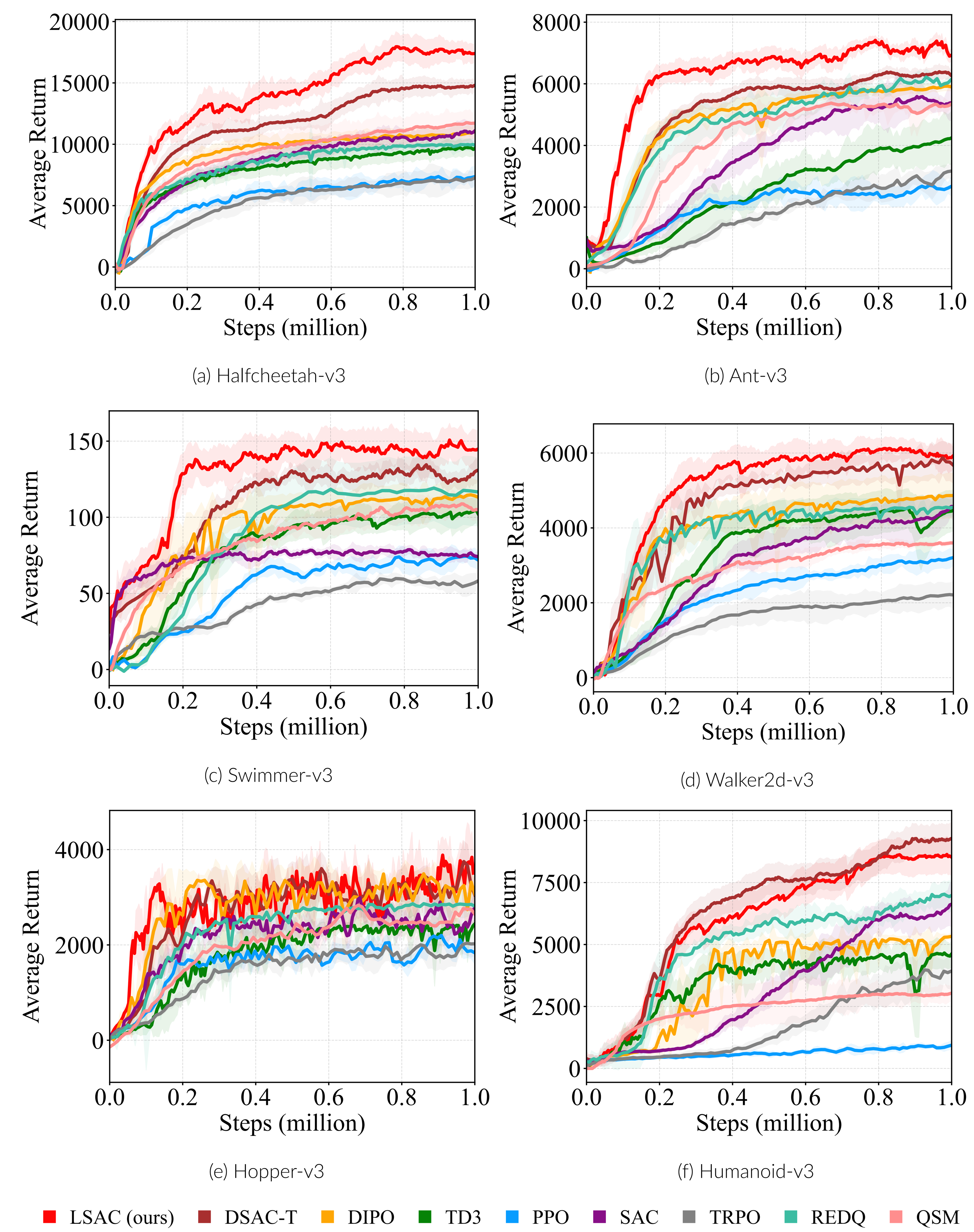
LSAC Encourages Exploration



(a) LSAC (ours) (b) DSAC-T (c) DIPO (d) SAC
(e) TD3 (f) PPO (g) TRPO

Figure 1. Exploration density map in the maze environment. The two goals are located in the upper-right and lower-left corners, as shown by the triangle markers. The starting position is at the center of the maze map.

Experiment: MuJoCo Continuous Control Tasks



(a) Halfcheetah-v3 (b) Ant-v3
(c) Swimmer-v3 (d) Walker2d-v3
(e) Hopper-v3 (f) Humanoid-v3

For more details check the paper!

References:

- Ishfaq, Haque, et al. "Provable and Practical: Efficient Exploration in Reinforcement Learning via Langevin Monte Carlo." ICLR 2024.
- Ishfaq, Haque, et al. "More Efficient Randomized Exploration for Reinforcement Learning via Approximate Sampling." Reinforcement Learning Conference 2024.

✉ haque.ishfaq@mail.mcgill.ca