S²AC: Energy-Based Reinforcement Learning With Stein Soft Actor Critic

Safa Messaoud, Billel Mokeddem*, Zhenghai Xue*, Linsey Pang, Bo An, Haipeng Chen and Sanjay Chawla

NANYANG TECHNOLOGICAL UNIVERSITY SINGAPORE · salesforce · WILLIAM & MARY CHARTERED 1693 · QCRI معهد قطر لبحوث الحوسبة Qatar Computing Research Institute · HAMAD BIN KHALIFA UNIVERSITY جامعة حمد بن خليفة

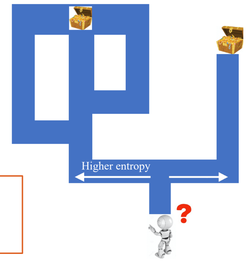
Motivation: A more optimal solution to MaxEntr RL

MaxEntr RL searches for the policy that maximizes the expected **future entropy** jointly with the expected **future reward**

$$\pi^* = \underset{\pi\in\Pi}{\operatorname{argmax}} \sum_t \gamma^t \mathbb{E}_{(s_t,a_t)\sim\rho_\pi}\Big[r(s_t,a_t) + \alpha\mathcal{H}(\pi(\cdot|s_t))\Big]$$

Why **MaxEntr RL**?

- Provably robust
- Better exploration
- Training stability
- Sample efficiency



Open Challenge
How can we estimate the entropy of arbitrarily complex policies?

Landmark MaxEntr RL Algorithms

- Soft Q-learning (**SQL**) [Haarnoja *et al.*, ICML 2017]
  - ✓ Implicit entropy computation for arbitrary policies
  - × Not scalable (uses importance sampling)
- Soft Actor Critic (**SAC**) [Haarnoja *et al.*, ICML 2018]
  - ✓ Tractable entropy estimate (policy=Gaussian)
  - × Limited to unimodal policies
- Soft Actor Critic with Normalizing Flows (**SAC-NF**) [Mazoure *et al.*, 2020]
  - ✓ Tractable entropy estimate (policy=Normalizing Flow)
  - × Hard to train (collapses to local minima)

Our Approach: Stein Soft Actor Critic (**S²AC**)

- Policy = Stein Variational Gradient Descent (SVGD) Sampler [Liu *et al., 2017*]
- ✓ **Closed form expression of the entropy of arbitrary complex distribution**
- ✓ Computationally efficient (only vector dot products and first order derivatives)
- ✓ Parameter efficient (same num. parameters as **SAC**)

Background

Change of Variable Formula (CVF):

$$F: Z \to X \text{ is invertible} \implies p_X(x) = p_Z(z)\left|\det\frac{\partial F(z)}{\partial z}\right|^{-1}$$

S²AC: Stein Soft Actor Critic

Optimal policy: $\pi^*(a|s) = \dfrac{\exp(Q_\phi(s,a)/\alpha)}{Z}$ ← Untractable

Critic: Learn the **Q-values** by optimizing the Bellman loss

$$\phi^* = \arg\min_\phi \mathbb{E}_{(s_t,a_t)\sim\rho_{\pi_\theta}}\left[(Q_\phi(s_t,a_t) - \hat{y})^2\right]$$

$$\hat{y} = r_t(s_t,a_t) + \gamma\mathbb{E}_{(s_{t+1},a_{t+1})\sim\rho_{\pi_\theta}}\left[Q_{\bar{\phi}}(s_{t+1},a_{t+1}) + \alpha\mathcal{H}(\pi_\theta(\cdot|s_{t+1}))\right]$$

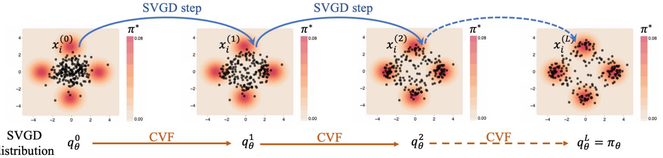
Actor: Learn a **policy (SVGD sampler)** via variational inference

$$\theta^* = \underset{\theta}{\operatorname{argmin}} D_{KL}\left(\pi_\theta(s|a)\middle\|\frac{\exp(Q_\phi(s,a)/\alpha)}{Z}\right)$$

SVGD Update Rule (step):

$$a_i^{l+1} = a_i^l + \epsilon\mathbb{E}_{a_j^l\sim q_\theta^l}\Big[\underbrace{k(a_j^l,a_j^l)\nabla_{a_j^l}\log\pi^*(a_j^l|s)}_{\text{Weighted average of particles gradients}} + \underbrace{\nabla_{a_j^l}k(a_i^l,a_j^l)}_{\text{Deterministic repulsion}}\Big]$$

l: SVGD iteration
$k(\cdot,\cdot)$: kernel



A Closed Form Estimate of The SVGD Induced Distribution:

The closed-form estimate of $\log q^L(a^L|s)$ for the SVGD based sampler with an RBF kernel is:

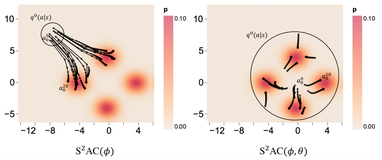
$$\log q^L(a^L|s) \approx \log q^0(a^0|s) + \frac{\epsilon}{m\sigma^2}\sum_{l=0}^{L-1}\sum_{j=1,a^l\neq a_j^l}^{m} k(a_j^l,a^l)\Big((a^l-a_j^l)^\top\nabla_{a_j^l}Q(s,a_j^l) + \frac{\alpha}{\sigma^2}\|a^l-a_j^l\|^2 - d\alpha\Big)$$

neighborhood · Curvature · dispersion

Parametrized Initial Distribution $q_\theta^0 = \mathcal{N}(\mu_\theta, \sigma_\theta)$

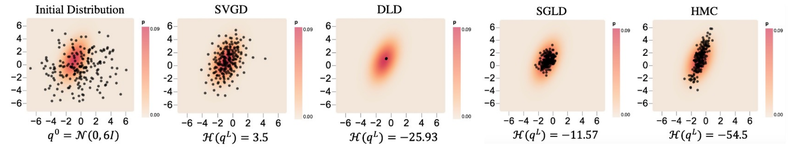
- Faster convergence
- Robustness to non-smooth deepnet landscapes by spatially constraining the actions:

$$(\mu - 3\sigma_\theta \le a_\theta \le \mu + 3\sigma_\theta)$$



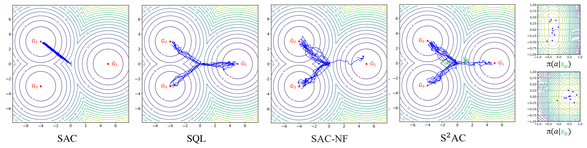
Results

Entropy Evaluation ($\mathcal{H}(\pi^*) = 3.41$)

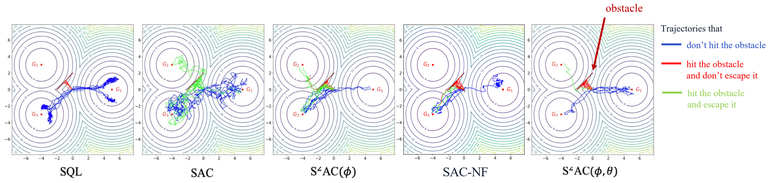


Multigoal Environment

- Multi-Modality: S²AC maximizes the expected future entropy and recovers all three modes

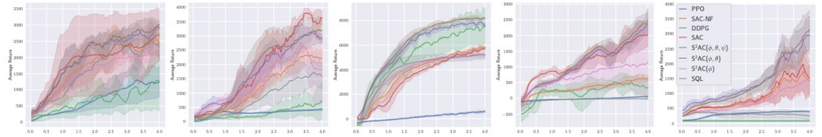


- Robustness: S²AC escapes obstacles at test time more frequently than other baselines



MuJoCo Environment

- Performance: S²AC beats the baselines on 4 out of 5 environments
- Run-time: Amortized S²AC has comparable run-time to SAC wo trading-off performance



(a) Hopper-v2 (b) Walker2d-v2 (b) HalfCheetah-v2 (c) Ant-v2 (d) Humnoid-v2