# Diversity Actor-Critic: Sample-Aware Entropy Regularization for Sample-Efficient Exploration

Seungyul Han and Youngchul Sung

Dept. of Electrical Engineering

KAIST

ICML 2021

Jun. 20, 2021

# Exploration in Reinforcement Learning (RL)

- Exploration to visit diverse samples is one of most important issues in RL community.
- Exploration can allow policy to converge on better points without falling into local optima.
- Random noise (Gaussian policy, parameter noise)
- Intrinsic reward (Counting, prediction error)
- Diversity gain (Maximum entropy RL, mutual information gain)

➔ We focus on the maximum entropy framework since it is widely used in RL and its optimal convergence is guaranteed.

# Maximum Entropy (MaxEnt) RL

- Information entropy $\mathcal{H}(p) = \mathbb{E}_{x\sim p}[-\log p(x)]$: Amount of uncertainty(information).
- MaxEnt RL adds the sum of policy entropy $\mathcal{H}(\pi)$ to the return objective of standard RL.

$$J(\pi) = \mathbb{E}_{\tau_0\sim\pi}\left[\sum_{t=0}^{T-1} r_t + \beta\mathcal{H}(\pi)\right], \quad (1)$$

$\tau_t$: A sample trajectory $(s_t, a_t, s_{t+1}, a_{t+1}\cdots)$, $\beta \in (0,\infty)$: Entropy weighting factor.

- MaxEnt RL framework can lead to wider exploration compared to standard RL.

# Soft Actor-Critic (SAC)

- Haarnoja et al., (2018) extends MaxEnt RL to the infinite-horizon MDP:

$$J_{SAC}(\pi) = \mathbb{E}_{\tau_0 \sim \pi}\left[\sum_{t=0}^{\infty} \gamma^t (r_t + \beta \mathcal{H}(\pi))\right], \tag{2}$$

- Soft policy iteration (SPI) theoretically guarantees the optimal convergence.
- Soft actor-critic (SAC) is a practical actor-critic algorithm for SPI.
- SAC has a good performance compared to standard RL algorithms.
- However, $\mathcal{H}(\pi)$ does not capture the previous sample distribution in off-policy RL.

# Contributions

We proposed,

- Sample-aware entropy regularization that uses previous action distribution for better exploration,
- Diverse policy iteration: Prove the optimal convergence of sample-aware entropy regularization,
- Diversity actor-critic (DAC): Practical implementation of sample-aware entropy framework,
- Adaptation scheme: Adaptive weighting factor in the mixture distribution.

# Sample-Aware Entropy Regularization

- $q$: the distribution of previous action samples stored in the replay buffer $\mathcal{D}$.
- We draw current samples from policy $\pi$ and store them in the replay buffer.
- The updated sample action distribution will be a mixture of $\pi$ and $q$:

$$q_{mix}^{\pi,\alpha} := \alpha\pi + (1-\alpha)q. \tag{3}$$

$\alpha \in [0, 1]$: Weighting factor of the mixture distribution.

- We regularizes the entropy of the mixture distribution $\mathcal{H}(q_{mix}^{\pi,\alpha})$:

$$J(\pi) = \mathbb{E}_{\tau_0\sim\pi}\left[\sum_{t=0}^{T-1}\gamma^t(r_t + \beta\mathcal{H}(q_{mix}^{\pi,\alpha}))\right]. \tag{4}$$

➔ Previously sampled actions will be given low probabilities to make $q_{mix}^{\pi,\alpha}$ uniform.

# A Toy Example

- Consider 1-step MDP with $s_0$ is the unique initial state.
- There is $N_a$ discrete actions ($\mathcal{A} = \{A_1, \cdots, A_{N_a}\}$), and $s_1$ is the terminal state, and $r$ is a deterministic reward function.
- There are $N_a$ state-action pairs in total.
- We assume there are already $N_a - 1$ samples in the buffer $\mathcal{D} = \{(s_0, A_1, r(s_0, A_1)), \cdots, (s_0, A_{N_a-1}, r(s_0, A_{N_a-1}))\}$.
- To estimate $Q$-function for all possible pairs, the policy should sample the last action $A_{N_a}$.

# A Toy Example

- Simple policy entropy maximization requires $N_a$ samples on average to visit $A_{N_a}$.
- $q$ is defined as $q(a_0|s_0) = \frac{1}{N_a-1}$ for $a_0 \in \{A_1, \cdots, A_{N_a-1}\}$ and $q(A_{N_a}|s_0) = 0$.
- If we set $\alpha = \frac{1}{N_a}$ in $q_{mix}^{\pi,\alpha} = \alpha\pi + (1-\alpha)q$, $\pi(A_{N_a}|s_0) = 1$ maximizes $\mathcal{H}(q_{mix}^{\pi,\alpha})$.
- Thus, we only need one sample to visit the action $A_{N_a}$.

➔ The proposed sample-aware entropy framework leads sample-efficient exploration!

# Ratio Function and Diverse Policy Iteration

- $q$ estimation requires discretization/counting/dimension reduction $\cdots$.
- We aim to maximize $J(\pi)$ by using the ratio function $R^{\pi,\alpha}$ without using explicit $q$.

$$R^{\pi,\alpha} = \frac{\alpha\pi}{\alpha\pi + (1-\alpha)q}: \text{ the ratio of } \alpha\pi \text{ to } q_{mix}^{\pi,\alpha}, \tag{5}$$

- Diverse policy iteration: the optimal convergence proof in terms of $R^{\pi_{old},\alpha}$.

**Theorem 1** (Diverse Policy Iteration)**.** By repeating iteration of the diverse policy evaluation and the diverse policy improvement, any initial policy converges to the optimal policy $\pi^*$ s.t. $Q^{\pi^*}(s_t, a_t) \geq Q^{\pi'}(s_t, a_t)$, $\forall\ \pi' \in \Pi$, $\forall\ (s_t, a_t) \in \mathcal{S} \times \mathcal{A}$. Also, such $\pi^*$ achieves maximum $J$, i.e., $J_{\pi^*}(\pi^*) \geq J_\pi(\pi)$ for any $\pi \in \Pi$.

**Theorem 2.** Suppose that the policy is parameterized with parameter $\theta$. Then, for parameterized policy $\pi_\theta$, the two objective functions $J_{\pi_{\theta_{old}}}(\pi_\theta(\cdot|s_t))$ and $\tilde{J}_{\pi_{\theta_{old}}}(\pi_\theta(\cdot|s_t))$ have the same gradient direction for $\theta$ at $\theta = \theta_{old}$ for all $s_t \in \mathcal{S}$, where $\theta_{old}$ is the parameter of the given current policy $\pi_{old}$.

# Diversity Actor-Critic

- Diversity actor-critic (DAC): practical implementation of sample-aware entropy regularized RL.
- $R^{\pi_{old},\alpha}$ can be estimated by $R_\eta^\alpha$ based on density ratio estimation [Sugiyama et al., 2012].
- All objective/loss functions in DAC can be represented in terms of $R_\eta^\alpha$:

$$\hat{J}_\pi(\theta) = \mathbb{E}_{s_t\sim\mathcal{D},\ a_t\sim\pi_\theta}[Q_\phi(s_t,a_t) + \alpha\log R_\eta^\alpha(s_t,a_t) - \alpha\log\pi_\theta(a_t|s_t)], \tag{6}$$

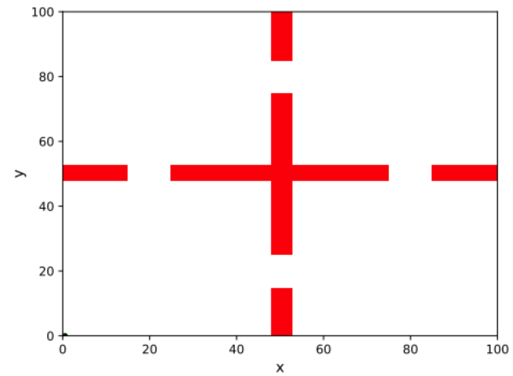
$$\hat{J}_{R^\alpha}(\eta) = \mathbb{E}_{s_t\sim\mathcal{D}}[\alpha\mathbb{E}_{a_t\sim\pi_\theta}[\log R_\eta^\alpha(s_t,a_t)] + (1-\alpha)\mathbb{E}_{a_t\sim\mathcal{D}}[\log(1-R_\eta^\alpha(s_t,a_t))]], \tag{7}$$

$$\hat{L}_Q(\phi) = \mathbb{E}_{(s_t,\ a_t)\sim\mathcal{D}}\left[\frac{1}{2}(Q_\phi(s_t,a_t) - \hat{Q}(s_t,a_t))^2\right], \tag{8}$$

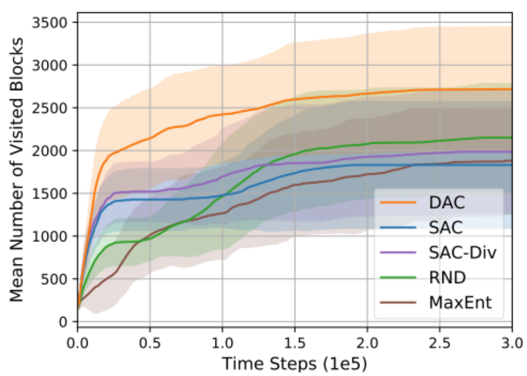
$$\hat{L}_V(\psi) = \mathbb{E}_{s_t\sim\mathcal{D}}\left[\frac{1}{2}(V_\psi(s_t) - \hat{V}(s_t))^2\right], \tag{9}$$

# Experiments: Pure Exploration

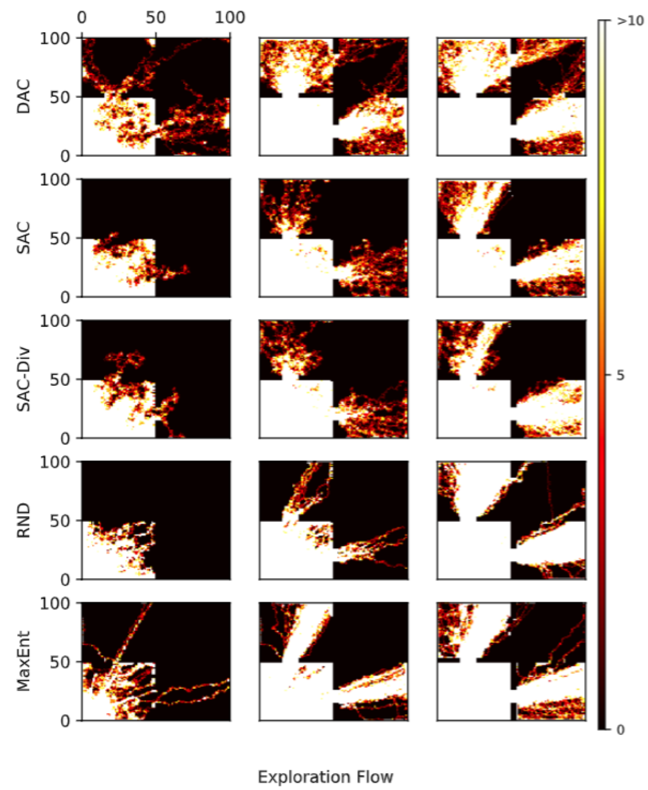
- DAC has better sample-efficiency for exploration than other exploration methods.




(a) Continuous 4-room maze




(b) Number of state visitation

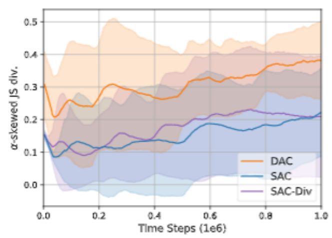



(c) State visit histogram at 5k (left) 50k (middle) 300k (right) steps

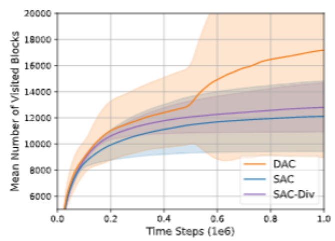
**Figure 1:** Pure exploration comparison

# Experiments: Sparse Rewarded Tasks

- Evaluation on SparseMujoco (Reward: 1 if the agent exceeds the threshold).
- Compared DAC with SAC baselines.
- DAC chooses more diverse action and visit more states, and it yields better performance.

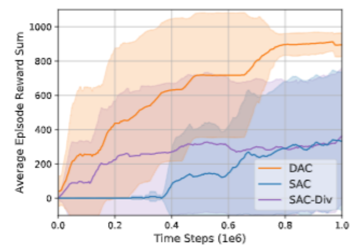



(a) Divergence $D^{\alpha}_{JS}$

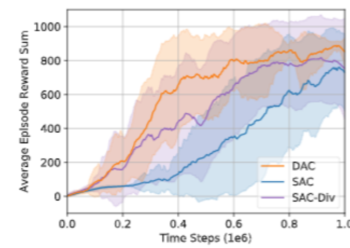



(b) Number of state visitation

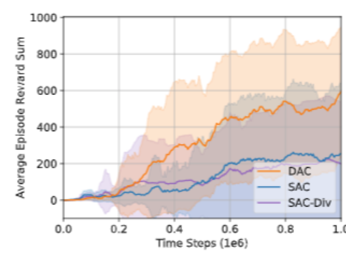
(a) $D^{\alpha}_{JS}(\pi||q)$ with $\alpha = 0.5$ and
(b) the corresponding mean number of state visitation



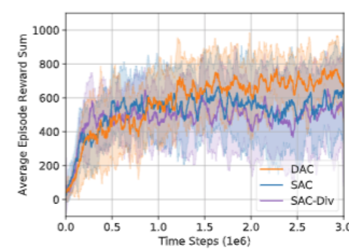

SparseHalfCheetah-v1




SparseHopper-v1




SparseWalker2d-v1




SparseAnt-v1

Performance comparison on Sparse Mujoco tasks

**Figure 2:** (a) $D^{\alpha}_{JS}$ with $\alpha = 0.5$ (b) state visitation (left) and the performance comparison (right) on Sparse-Mujoco tasks

# Conclusion

We proposed,

1. Sample-aware entropy regularization that considers the previous distribution for better exploration.

2. Diverse policy iteration to guarantee the convergence.

3. Diversity actor-critic to implement sample-aware entropy regularized RL.

4. DAC shows better performance compared to SAC baselines and recent RL algorithms.

Thank You!!