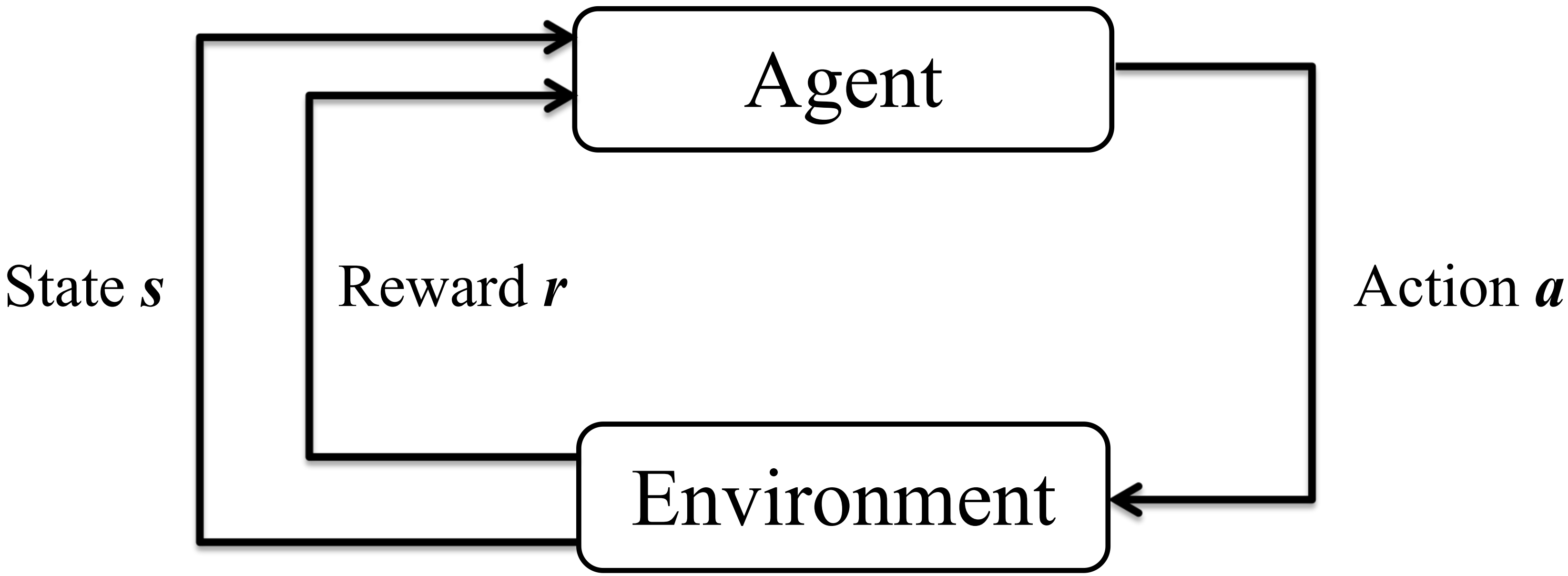
On the Reuse Bias in Off-Policy Reinforcement Learning

Chengyang Ying, Zhongkai Hao, Xinning Zhou,
Hang Su, Dong Yan, Jun Zhu
Tsinghua University
Contact: ycy21@mails.tsinghua.edu.cn

Reinforcement Learning

- Reinforcement learning formulates the sequence decision problem as a Markov desision process. At each time step, the agent perceives the current state, chooses its action by its policy, interates with the environment, obtains a reward, and arrives at the next state.



Objective: $$\max J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[R(\tau) \triangleq \sum_{t=0}^{\infty} \gamma^t r_t\right]$$

Off-policy Evaluation

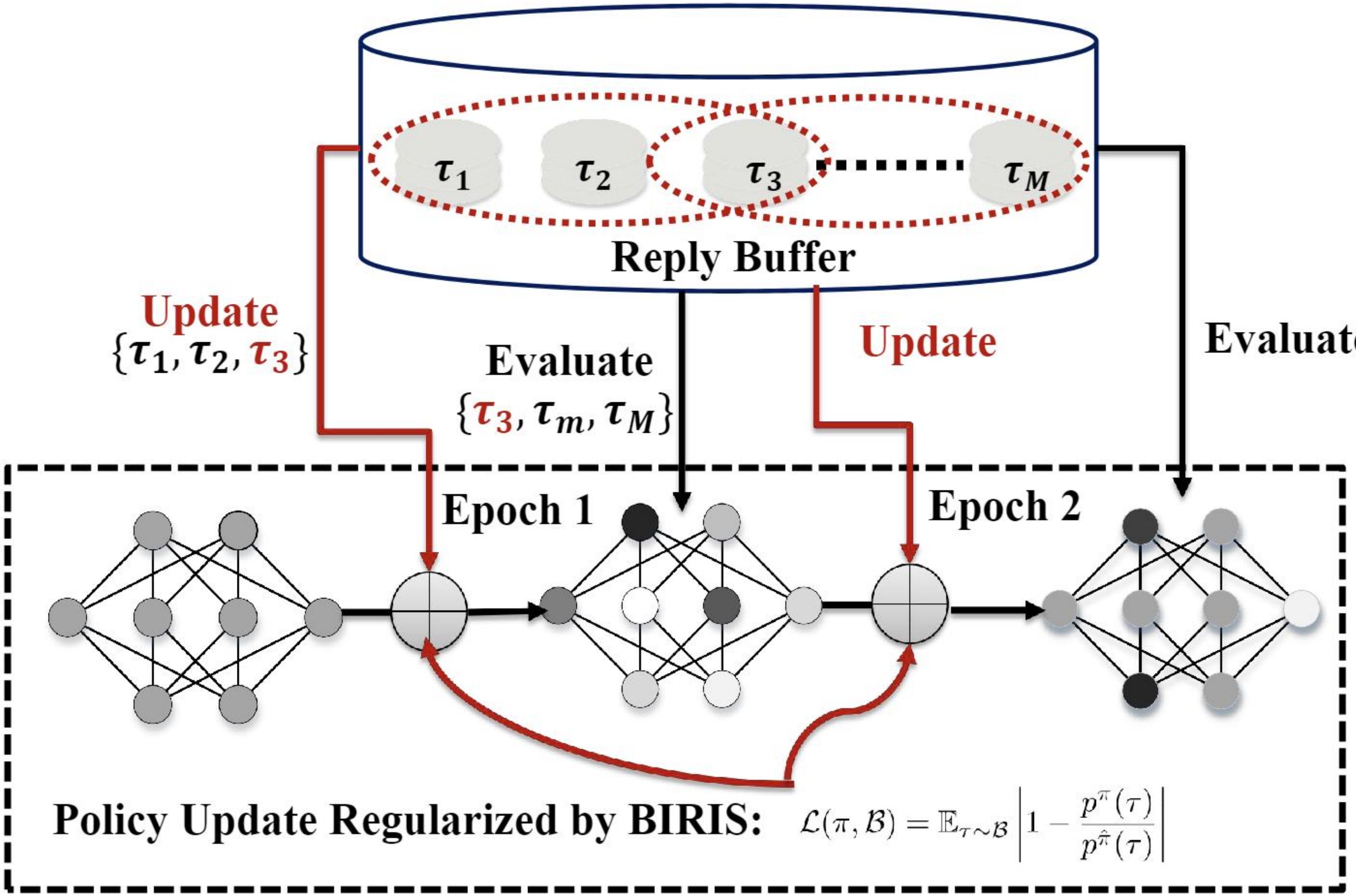
- In off-policy RL, the agent will store hitorical data into the Replay Buffer and reuse them to improve the sample efficiency.

- Due to the *distribution shift* between the target policy and the behavior policy, off-policy evaluation (OPE) utilizes importance sampling to provide an unbiased estimation.

$$\underbrace{\frac{1}{m}\sum_{i=1}^{m}\frac{p^{\pi}(\tau_i)}{p^{\hat{\pi}_i}(\tau_i)}R(\tau_i)}_{\hat{J}_{\hat{\Pi},\mathcal{B}}(\pi)} \approx \underbrace{\mathbb{E}_{\tau\sim\pi}\left[R(\tau) \triangleq \sum_{t=0}^{\infty}\gamma^t r_t\right]}_{J(\pi)}$$

- Unfortunately, reusing trajectories in the replay buffer to optimize and evaluate the policy may introduce a bias, which is systematically examined in this work.

Overview



τ_1
τ_2
τ_3
τ_M
Reply Buffer
Update
$\{\tau_1, \tau_2, \tau_3\}$
Evaluate
$\{\tau_3, \tau_m, \tau_M\}$
Update
Evaluate
Epoch 1
Epoch 2
Policy Update Regularized by BIRIS: $\mathcal{L}(\pi, \mathcal{B}) = \mathbb{E}_{\tau\sim\mathcal{B}}\left|1 - \frac{p^{\pi}(\tau)}{p^{\hat{\pi}}(\tau)}\right|$

Reuse Bias

- *Reuse Bias*: The bias of OPE caused by reusing the replay buffer.

Definition 1 (Reuse Bias). For any off-policy algorithm $\mathcal{O}$, initialized policy π_0 and replay buffer $\mathcal{B} \sim \hat{\Pi}$, we define the Reuse Error of $\mathcal{O}$ on π_0 and $\mathcal{B}$ as

$$\epsilon_{\mathrm{RE}}(\mathcal{O}, \pi_0, \mathcal{B}) \triangleq \hat{J}_{\hat{\Pi},\mathcal{B}}(\mathcal{O}(\pi_0, \mathcal{B})) - J(\mathcal{O}(\pi_0, \mathcal{B})).$$

Moreover, we define its expectation as the Reuse Bias:

$$\epsilon_{\mathrm{RB}}(\mathcal{O}, \pi_0) \triangleq \mathbb{E}_{\mathcal{B}}[\epsilon_{\mathrm{RE}}(\mathcal{O}, \pi_0, \mathcal{B})].$$

- When the trained policy is independent of the replay buffer, OPE is unbiased, i.e.,

$$\epsilon_{\mathrm{RB}}(\mathcal{O}, \pi_0) = \mathbb{E}_{\mathcal{B}}\left[\epsilon_{\mathrm{RE}}(\mathcal{O}, \pi_0, \mathcal{B})\right] = 0$$

- However, when we use historical data to update our policy (as most off-policy algorithms do), the Reuse Bias is no longer 0.

Overestimation

- We first analyze the *overestimation* for OPE under two practical situations: trained policies own the highest estimated return (Theorem 1), or are trained by one-step PG (Theorem 2).

Theorem 1 (Overestimation for Off-Policy Evaluation). *Assume that $\mathcal{O}^*(\pi_0, \mathcal{B})$ is the optimal policy of $\mathcal{H}$ over the replay buffer $\mathcal{B}$, i.e.,*

$$\mathcal{O}^*(\pi_0, \mathcal{B}) = \arg\max_{\pi \in \mathcal{H}} \hat{J}_{\hat{\Pi}, \mathcal{B}}(\pi) = \arg\max_{\pi \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \left[\frac{p^{\pi}(\tau_i)}{p^{\hat{\pi}_i}(\tau_i)} R(\tau_i) \right].$$

We can show that $\hat{J}_{\hat{\Pi}, \mathcal{B}}(\mathcal{O}^*(\pi_0, \mathcal{B}))$ is an overestimation of $J(\mathcal{O}^*(\pi_0, \mathcal{B}))$, i.e., $\epsilon_{\mathrm{RB}}(\mathcal{O}^*, \pi_0) = \mathbb{E}_{\mathcal{B} \sim \hat{\Pi}} [\epsilon_{\mathrm{RE}}(\mathcal{O}^*, \pi_0, \mathcal{B})] \geq 0$. If the equality holds, for $\forall \mathcal{B}, \mathcal{B}' \sim \hat{\Pi}$, we have $\mathcal{O}^*(\pi_0, \mathcal{B}) = \arg\max_{\pi \in \mathcal{H}} \hat{J}_{\hat{\Pi}, \mathcal{B}'}(\pi)$.

Theorem 2 (Overestimation for One-Step PG). *Given a parameterized policy π_θ which is independent with the replay buffer $\mathcal{B}$ and is differentiable to the parameter θ, we consider the one-step policy gradient*

$$\theta' = \theta + \alpha \nabla_\theta \hat{J}_{\hat{\Pi}, \mathcal{B}}(\pi_\theta),$$

*where α is the learning rate. If $\nabla_\theta \hat{J}_{\hat{\Pi}, \mathcal{B}}(\pi_\theta)$, as the function of $\mathcal{B}$, is **not** constant, and $\alpha > 0$ is sufficiently small, then the Reuse Bias is strictly larger than 0, i.e.,*

$$\mathbb{E}_{\mathcal{B} \sim \hat{\Pi}} \hat{J}_{\hat{\Pi}, \mathcal{B}}(\pi_{\theta'}) > \mathbb{E}_{\mathcal{B} \sim \hat{\Pi}} J(\pi_{\theta'}).$$

High-probability Bound

IJCAI/2023 MACAO

- Also, we provide a high-probability upper bound for Reuse Error. Compared with previous results, our results hold for any hypothesis sets, and is related to the optimized policy.

Theorem 4 (High-Probability Bound for Reuse Error). *Assume that, for any trajectory τ, we can bound its return as $0 \leq R(\tau) \leq 1$. Then, for **any** off-policy algorithm $\mathcal{O}$ and initialized policy $\pi_0 \in \mathcal{H}$, with a probability of at least $1-\delta$ over the choice of an i.i.d. training set $\mathcal{B} = \{\tau_i\}_{i=1}^m$ sampled by the same original policy $\hat{\pi}$, the following inequality holds:*

$$|\epsilon_{\mathrm{RE}}(\mathcal{O}, \pi_0, \mathcal{B})| \leq \sqrt{\frac{m\epsilon_1 + \log\left(\frac{m^2}{\delta}\right)}{m-1}} + \epsilon_2,$$

where ϵ_1 and ϵ_2 are defined as:

$$\epsilon_1 = \mathrm{KL}[p^{\mathcal{O}(\pi_0,\mathcal{B})}(\cdot)||p^{\hat{\pi}}(\cdot)],$$
$$\epsilon_2 = \frac{1}{m}\sum_{i=1}^{m}\left|1 - \frac{p^{\mathcal{O}(\pi_0,\mathcal{B})}(\tau_i)}{p^{\hat{\pi}}(\tau_i)}\right| = \mathbb{E}_{\tau\sim\mathcal{B}}\left|1 - \frac{p^{\mathcal{O}(\pi_0,\mathcal{B})}(\tau)}{p^{\hat{\pi}}(\tau)}\right|.$$

can be directly calculated by the replay buffer

Stability Analyses

- Moreover, we establish the concept of the stability for off-policy algorithms and further show that we can control the Reuse Bias in off-policy stochastic policy gradient just by controlling ε_2.

Definition 2 (Stability for Off-Policy Algorithm)**.** A randomized off-policy algorithm $\mathcal{O}$ is β-uniformly stable if for all Replay Buffer $\mathcal{B}, \mathcal{B}'$, such that $\mathcal{B}, \mathcal{B}'$ differ in at most one trajectory, we have

$$\forall \tau, \pi_0, \quad \mathbb{E}_{\mathcal{O}}\left[p^{\mathcal{O}(\pi_0, \mathcal{B})}(\tau) - p^{\mathcal{O}(\pi_0, \mathcal{B}')}(\tau)\right] \leq \beta. \tag{4}$$

Theorem 5 (Bound for the Reuse Error of Stable Algorithm)**.** *Suppose a randomized off-policy algorithm $\mathcal{O}$ is β-uniformly stable, then we can prove that*

$$\forall \pi_0, \quad |\mathbb{E}_{\mathcal{B}\sim\hat{\pi}}\mathbb{E}_{\mathcal{O}}\left[\epsilon_{\mathrm{RE}}(\mathcal{O}, \pi_0, \mathcal{B})\right]| \leq \beta. \tag{5}$$

Theorem 6 (Details and Proof are in Appendix)**.** *We assume that the policy π_θ is parameterized with θ, and $|\nabla_\theta \log p^{\pi_\theta}(\tau)| \leq L_1$ holds for any θ, τ, and $p^{\pi_\theta}(\tau)$ is L_2-Lipsticz to θ for any τ. If we constrain the policy by $\mathcal{L}(\pi, \mathcal{B}) \leq M$, then off-policy stochastic policy gradient algorithm (detailed in Appendix A.7) is β-uniformly stable where β is positively correlated with M, L_1 and L_2.*

BIRIS

$$\mathcal{O}_{\text{BIRIS}}(\pi_0, \mathcal{B}) = \arg\min_{\pi \in \mathcal{H}} \mathcal{L}_{\text{BIRIS}}(\pi, \mathcal{B}),$$

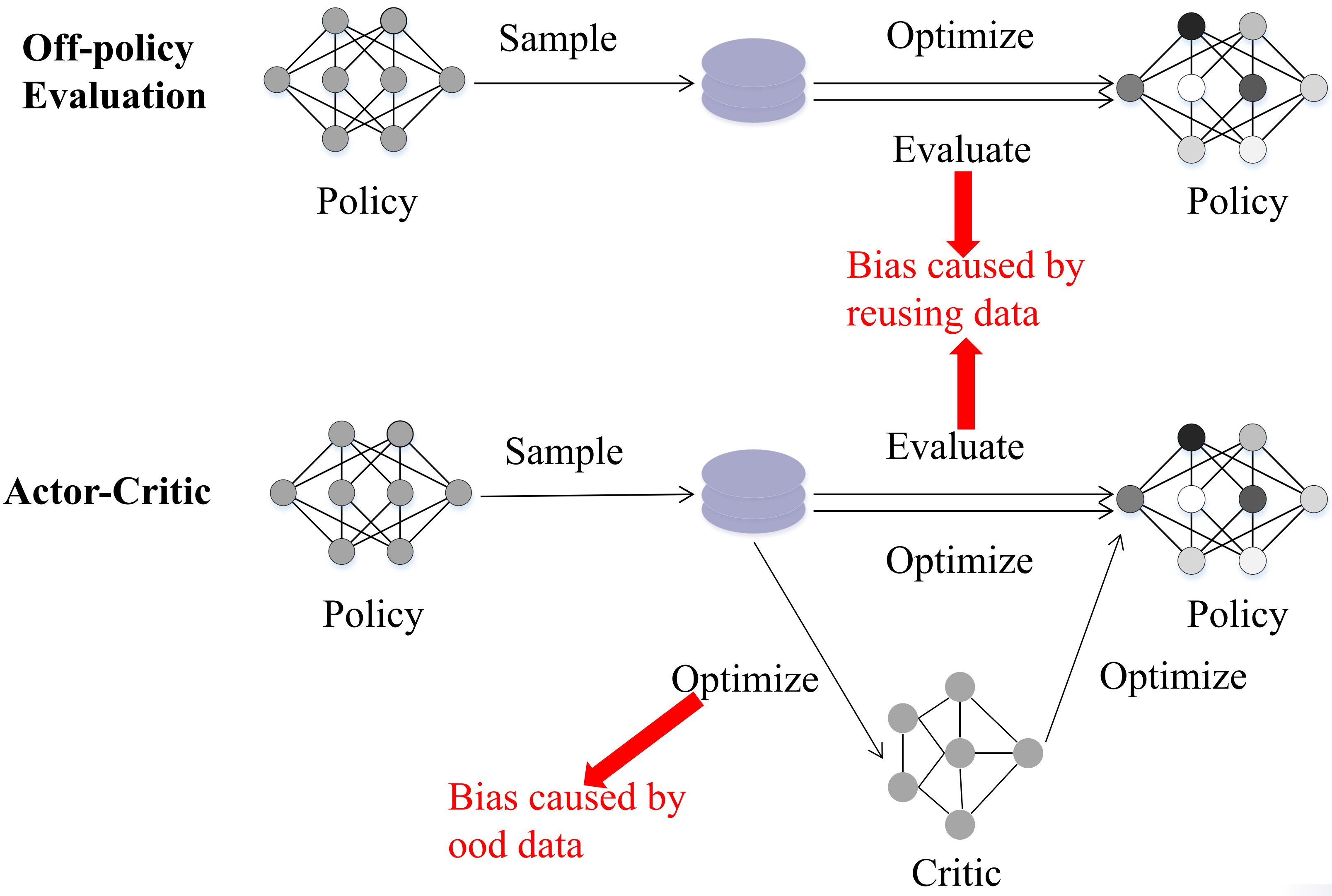
$$\text{where } \mathcal{L}_{\text{BIRIS}}(\pi, \mathcal{B}) = \mathcal{L}_{\text{RL}}(\pi, \mathcal{B}) + \alpha \mathcal{L}(\pi, \mathcal{B}).$$

$$\mathcal{L}(\pi, \mathcal{B}) = \mathbb{E}_{\tau \sim \mathcal{B}} \left| 1 - \frac{p^{\pi}(\tau)}{p^{\hat{\pi}}(\tau)} \right| = \mathbb{E}_{\tau \sim \mathcal{B}} \left| 1 - \prod_i \frac{\pi(a_i|s_i)}{\hat{\pi}(a_i|s_i)} \right|$$

a surrogate when trajectories are so long that their probabilities are difficult to calculate and numerically unstable

$$\mathcal{L}_{\text{BR}}(\pi, \mathcal{B}) \triangleq \mathbb{E}_{(s,a) \in \mathcal{B}} \left| \frac{\pi(a|s)}{\hat{\pi}(a|s)} - 1 \right|$$

Reuse Bias in Actor-Critic



Experimental Results

We present empirical results to answer the questions:

- How severe is Reuse Bias in the practical experiments and can our BIRIS effectively reduce Reuse Bias?

  Gridworld: calculate and compare the Reuse Bias of PG+IS, PG+WIS, PG+IS+BIRIS, and PG+WIS+BIRIS

- What is the empirical performance of our BIRIS for actor-critic methods in complicated continuous control tasks?

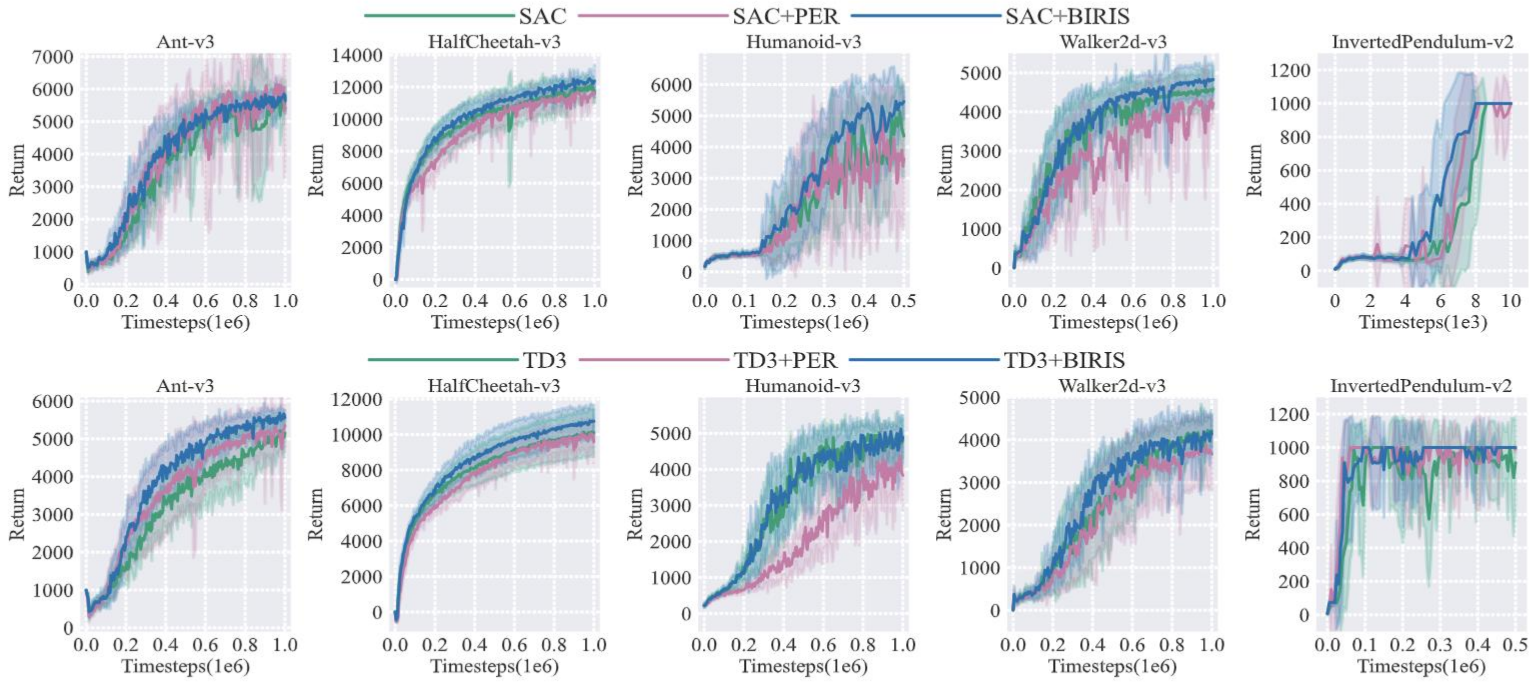
  MuJoCo: evaluate BIRIS compared with SAC and TD3, with uniform sampling and prioritized experience replay (PER)

Gridworld

Size of Replay Buffer	Method	5×5	5×5-random	6×6	6×6-random	8×8	16×16
30	PG+IS	0.57	0.26	0.86	0.55	2.04	19.04
	PG+WIS	0.36	0.24	0.72	0.43	1.50	5.75
	PG+IS+BIRIS	0.19	**0.08**	**0.11**	**0.25**	0.47	0.43
	PG+WIS+BIRIS	**0.16**	0.12	0.17	0.28	**0.23**	**0.05**
40	PG+IS	0.38	0.24	0.67	0.29	1.99	80.62
	PG+WIS	**0.20**	0.21	0.49	0.32	1.20	4.75
	PG+IS+BIRIS	0.23	0.18	0.29	0.25	0.39	**0.44**
	PG+WIS+BIRIS	0.21	**0.14**	**0.25**	**0.24**	**0.26**	0.51
50	PG+IS	0.44	0.21	0.60	0.42	2.01	13.40
	PG+WIS	0.26	0.22	0.51	0.31	1.22	4.65
	PG+IS+BIRIS	0.23	0.20	**0.17**	**0.11**	**0.23**	0.26
	PG+WIS+BIRIS	**0.14**	**0.18**	0.25	0.16	0.24	**0.21**

MuJoCo

Method	Ant	HalfCheetah	Humanoid	Walker2d	InvertedPendulum
SAC	5797.9±492.1	12096.6±597.7	5145.4±567.4	4581.1±541.4	**1000.0±0.0**
SAC+PER	**6133.5±269.0**	11695.1±603.2	4860.8±1117.1	4320.5±392.5	**1000.0±0.0**
SAC+BIRIS	5843.8±159.9	**12516.5±613.3**	**5466.1±493.9**	**4836.3±405.6**	**1000.0±0.0**
TD3	5215.7±488.2	10147.6±1291.6	5012.9±211.1	**4223.0±350.5**	**1000.0±0.0**
TD3+PER	5351.1±530.1	10091.4±830.3	4365.5±608.3	3879.6±557.2	**1000.0±0.0**
TD3+BIRIS	**5675.1±132.6**	**10774.2±907.0**	**5117.9±181.6**	4189.4±485.9	**1000.0±0.0**



Conclusion

- We first systematically discuss the bias of off-policy evaluation due to reusing the replay buffer. We show that the off-policy evaluation via importance sampling is an overestimation when optimized by the same replay buffer, which is recognized as the *Reuse Bias* in this paper.

- We derive a high-probability bound of the Reuse Bias holds for any hypothesis sets. Also, we introduce the concept of stability and provide an upper bound for the Reuse Bias via stability.

- We propose BIRIS to control Reuse Biase. BIRIS can conspicuously reduce the Reuse Bias in experiments of MiniGrid. Moreover, experiments show that BIRIS can improve the performance and sample efficiency for different off-policy methods in MuJoCo tasks.

Thanks