On the Reuse Bias in Off-Policy Reinforcement Learning

Chengyang Ying[1] , **Zhongkai Hao**[1] , **Xinning Zhou**[1] , **Hang Su**[1,2] , **Dong Yan**[1] , **Jun Zhu**[*1,2]

[1]Department of Computer Science & Technology, Institute for AI, BNRist Center,
Tsinghua-Bosch Joint ML Center, THBI Lab, Tsinghua University
[2]Pazhou Laboratory (Huangpu), Guangzhou, China
ycy21@mails.tsinghua.edu.cn, dcszj@tsinghua.edu.cn

Abstract

Importance sampling (IS) is a popular technique in off-policy evaluation, which re-weights the return of trajectories in the replay buffer to boost sample efficiency. However, training with IS can be unstable and previous attempts to address this issue mainly focus on analyzing the variance of IS. In this paper, we reveal that the instability is also related to a new notion of Reuse Bias of IS — the bias in off-policy evaluation caused by the reuse of the replay buffer for evaluation and optimization. We theoretically show that the off-policy evaluation and optimization of the current policy with the data from the replay buffer result in an overestimation of the objective, which may cause an erroneous gradient update and degenerate the performance. We further provide a high-probability upper bound of the Reuse Bias and show that controlling one term of the upper bound can control the Reuse Bias by introducing the concept of stability for off-policy algorithms. Based on these analyses, we present a novel yet simple Bias-Regularized Importance Sampling (BIRIS) framework along with practical algorithms, which can alleviate the negative impact of the Reuse Bias, and show that our BIRIS can significantly reduce the Reuse Bias empirically. Moreover, extensive experimental results show that our BIRIS-based methods can significantly improve the sample efficiency on a series of continuous control tasks in MuJoCo.

1 Introduction

Off-policy reinforcement learning algorithms [Mnih *et al.*, 2015; Lillicrap *et al.*, 2016; Haarnoja *et al.*, 2018; Fujimoto *et al.*, 2018] typically have high sample efficiency and are suitable for many real-world applications. The key idea of off-policy algorithms is to reuse historical trajectories in the replay buffer. Though promising, it also causes an issue that these trajectories are not sampled from the current policy. When estimating the expected cumulative return of the current policy with trajectories generated by historical policies, off-policy evaluation, as the core of off-policy algorithms, yields

the *distribution shift* in consequence [Hanna *et al.*, 2019]. To alleviate this issue, the standard method in off-policy evaluation is importance sampling [Kahn and Marshall, 1953; Precup *et al.*, 2000], which is a widely used Monte Carlo technique to evaluate the expected return of the target policy when the training data is generated by a different behavior policy. In practice, it usually requires computing the degree of deviation between the target policy and the behavior policy.

Previous off-policy methods, which are usually combined with importance sampling, have suffered from the problem of unstable training [Li *et al.*, 2015]. To address this problem, previous work mainly focuses on the high variance of importance sampling [Li *et al.*, 2015; Liu *et al.*, 2018; Xie *et al.*, 2019; Fujimoto *et al.*, 2021]. However, we find that reusing trajectories in the replay buffer to optimize and evaluate may introduce a bias to off-policy evaluation (See Fig. 1(1)), which may also lead to unstable training and can not be ignored.

In this work, we present a new analysis beyond variance by introducing the concept of *Reuse Bias*, which formally characterizes the bias in off-policy evaluation caused by the reuse of the replay buffer for evaluation and optimization (See Definition 1). Under some mild assumptions (See Theorem 1, 2), we theoretically show that off-policy evaluation is an overestimation of the true expected cumulative return. To explain how severe the Reuse Bias is, we first prove that the expectation of off-policy evaluation for the optimal policy over the replay buffer may be arbitrarily large when its true return is arbitrarily small (See Theorem 3). Moreover, we experiment on MiniGrid to show that Reuse Bias will cause an erroneous policy evaluation and even gets more significant as the environments become more complex (See Fig. 1(2), and more details are in Sec. 6.2). Consequently, the Reuse Bias can be significantly severe and the inaccurate off-policy evaluation may mislead the direction of the gradient. This becomes particularly problematic for further policy improvement because sub-optimal actions might be reinforced in the next policy update, and further induces unstable training.

To reveal how to optimize the policy for controlling the Reuse Bias, we present a high-probability upper bound of Reuse Bias for any possible policies by extending PAC Bayes techniques [McAllester, 1998; McAllester, 2003] (See Theorem 4). Compared with classical bounds in statistical learning theory, our result holds for any policy hypothesis and is related to the optimized policy for guiding further optimization to con-

---
*Corresponding author.
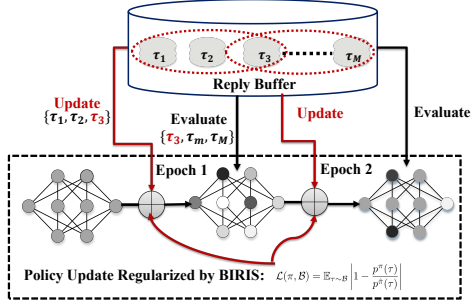
(1) Reuse Bias due to resample τ_3 and our BIRIS
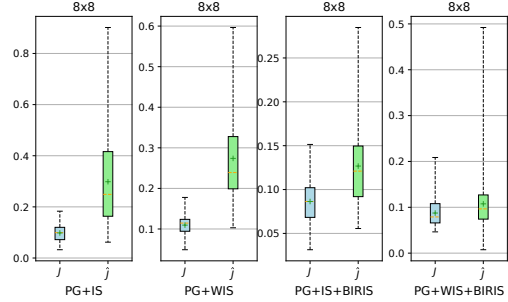
(2) Reuse Bias for PG and PG+BIRIS in MiniGrid

Figure 1: (a) A high-level illustration of the Reuse Bias and our BIRIS. The Reuse Bias is caused by the fact that off-policy methods optimize and evaluate the policy with the same data in the replay buffer. (b) Experimental results of PG+IS, PG+WIS, PG+BIRIS+IS, and PG+BIRIS+WIS in MiniGrid 8×8 with replay buffer size 30. In each subfigure, J and $\hat{J}$ represent the expected return and the estimated return via the replay buffer of the target policy respectively. We repeat the experiment 50 times and plot the box diagram, where the orange dashed line represents the mean. This figure shows that Reuse Bias is severe to cause an erroneous policy evaluation and our BIRIS can significantly reduce the Reuse Bias. (More details are in Sec. 6.2 and Appendix B.1)

trol the Reuse Bias. Moreover, we present that controlling one term, which only depends on the replay buffer and the policy, of our upper bound can effectively control the Reuse Bias by introducing the concept of stability for off-policy algorithms (See Definition 2, Theorem 5 and 6). Based on these analyses, we propose a novel Bias-Regularized Importance Sampling (BIRIS) framework which can improve the sample efficiency of off-policy algorithms by controlling Reuse Bias. We then experiment on MiniGrid to show that our BIRIS can conspicuously reduce the Reuse Bias (See Fig. 1(2), and more details are in Sec. 6.2). Since our method is orthogonal to existing off-policy algorithms and variance-reduction methods, like weighted importance sampling (WIS) [Mahmood *et al.*, 2014], BIRIS can be easily combined with them. Extensive experimental results show that our BIRIS can significantly reduce the Reuse Bias in MiniGrid compared with PG+IS and PG+WIS. Since practical off-policy algorithms for handling complicated continuous control tasks always consider using actor-critic methods rather than directly using trajectories, we also extend our Reuse Bias to AC-based methods and show that these methods also suffer from the negative impact of the Reuse Bias. Moreover, we extend our BIRIS to this more practical setting and provide two practical algorithms (SAC+BIRIS and TD3+BIRIS). Extensive experiments show that our BIRIS-based methods significantly improve the performance and sample efficiency compared with baselines on a series of continuous control tasks in MuJoCo. In summary, our contributions are:

- To the best of our knowledge, this is the first attempt to 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.

- Moreover, we derive a high-probability bound of the Reuse Bias, which indicates that it cannot be eliminated only by increasing the number of trajectory samples.

Then we introduce the concept of stability and provide an upper bound for the Reuse Bias via stability, which indicates that controlling one term of the high-probability bound can effectively control the Reuse Bias.

- Based on these analyses, we propose BIRIS and show that our BIRIS can conspicuously reduce the Reuse Bias through experiments in MiniGrid. Moreover, we experimentally show that our method can significantly improve the performance and sample efficiency for different off-policy methods in different MuJoCo tasks.

2 Related Work

Off-policy Evaluation. Designing safe, stable, and effective methods are significant for reinforcement learning (RL) [García and Fernández, 2015; Ying *et al.*, 2022], especially for off-policy methods, which hopes to reuse historical trajectories for improving the sample efficiency [Precup *et al.*, 2000; Precup *et al.*, 2001]. It is well known that, when the current policy is independent of the samples, importance sampling is unbiased but has a high variance that is exponentially related to trajectory length [Li *et al.*, 2015], which may cause the result unstable. Consequently, a lot of research focuses on reducing the variance in off-policy evaluation, like weighted importance sampling [Mahmood *et al.*, 2014], doubly robust estimator methods [Jiang and Li, 2016; Thomas and Brunskill, 2016; Su *et al.*, 2020], marginalized importance sampling methods [Liu *et al.*, 2018; Xie *et al.*, 2019; Fujimoto *et al.*, 2021], and so on [Metelli *et al.*, 2018; Hanna *et al.*, 2019; Kumar *et al.*, 2020a; Liu *et al.*, 2021; Zhang *et al.*, 2021; Metelli *et al.*, 2021]. Some researchers also pay attention to providing a high confidence estimation for off-policy evaluation. [Thomas *et al.*, 2015] first applied concentration inequality to provide a lower bound of off-policy evaluation. Follow-up work considered how to provide confidence intervals [Hanna *et al.*, 2017; Shi *et al.*, 2021] as well as a high-probability lower bound for self-normalized importance weighting [Kuzborskij *et al.*, 2021]. Some re-

cent work also provided a high-confidence estimation of the variance [Chandak *et al.*, 2021b] and other parameters of the return distribution [Metelli *et al.*, 2018; Metelli *et al.*, 2021; Chandak *et al.*, 2021a]. However, little work has been done on systematically examining the bias of off-policy evaluation.

Bias in Off-policy and Offline RL. Since bias is significant to analyze and develop algorithms in RL, there are some widely discussed biases in off-policy and offline settings like Q-function overestimation in DQN [Van Hasselt *et al.*, 2016], extrapolation error in offline RL [Fujimoto *et al.*, 2019], and sampling error in TD learning [Pavse *et al.*, 2020]. These bias is mainly caused by out-of-distribution data [Fujimoto *et al.*, 2019] and about Q function estimation in actor-critic methods [Van Hasselt *et al.*, 2016; Fujimoto *et al.*, 2019]. Besides them, in this work, we first discuss the bias resulting from reusing the trajectories in the replay buffer, which can regard as the bias in the policy optimization in actor-critic methods (more details are in Sec. 5.3).

3 Preliminary

We consider the Markov Decision Process (MDP) of $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$ in which $\mathcal{S}$ and $\mathcal{A}$ represent the state and action spaces, respectively. For any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\mathcal{P}(\cdot|s, a)$ is a distribution over $\mathcal{S}$, representing its transition dynamic. Moreover, $\mathcal{R}(s, a)$ is the reward function and $\gamma \in (0, 1)$ is the discount factor. The policy of any agent is modeled as a mapping $\pi : \mathcal{S} \rightarrow \mathcal{D}(\mathcal{A})$, i.e., $\pi(\cdot|s)$ is a distribution over $\mathcal{A}$ given state $s \in \mathcal{S}$.

In RL, the agent interacts with the environment by making its decision at every timestep. At the beginning, we assume that the agent is in the state $s_0 \sim \mu(\cdot)$, where $\mu(\cdot)$ is the initial state distribution. At timestep t, the agent chooses its action $a_t \sim \pi(\cdot|s_t)$, arrives at the next state $s_{t+1} \sim \mathcal{P}(\cdot|s_t, a_t)$, and receive a corresponding reward $r_t = \mathcal{R}(s_t, a_t)$. The performance of policy π is defined as the expectation of the discounted return of the trajectory:

$$J(\pi) = \mathbb{E}_{\tau \sim \pi} \left[R(\tau) \triangleq \sum_{t=0}^{\infty} \gamma^t r_t \right], \tag{1}$$

where $R(\tau)$ represents the discounted return of the trajectory $\tau \triangleq (s_0, a_0, r_0, s_1, a_1, r_1, ...) \in \Gamma$.

As the trajectory distribution $\tau \sim \pi$ is unknown, in practice, we estimate $J(\pi)$ by Monte Carlo sampling. In order to improve sample efficiency, off-policy algorithms utilize trajectories sampled from other policies via importance sampling. We denote that there are m trajectories in the replay buffer $\mathcal{B}$ and the trajectory τ_i is sampled from policy $\hat{\pi}_i$, i.e. $\mathcal{B} = \{\tau_i\}_{i=1}^{m} \in \Gamma^m, \tau_i \sim \hat{\pi}_i$.

For notational clarity, we denote $\hat{\Pi} = (\hat{\pi}_1, \hat{\pi}_2, ..., \hat{\pi}_m)$ and $\mathcal{B} \sim \hat{\Pi} = \tau_1 \sim \hat{\pi}_1, \tau_2 \sim \hat{\pi}_2, ..., \tau_m \sim \hat{\pi}_m$. In off-policy evaluation, to estimate $J(\pi)$ of a different policy π, we need to modify the weight of the return for each trajectory in $\mathcal{B}$ via importance sampling [Kahn and Marshall, 1953; Precup *et al.*, 2000]. The resulting importance sampling estimate of the return for π is:

$$\hat{J}_{\hat{\Pi}, \mathcal{B}}(\pi) = \frac{1}{m} \sum_{i=1}^{m} \frac{p^\pi(\tau_i)}{p^{\hat{\pi}_i}(\tau_i)} R(\tau_i), \tag{2}$$

where $p^\pi(\tau)$ represents the probability of generating the trajectory τ with policy π, i.e.,

$$p^\pi(\tau) = \mu(s_0) \prod_{i=0}^{\infty} \left[\pi(a_i|s_i) \mathcal{P}(s_{i+1}|s_i, a_i) \right]. \tag{3}$$

4 Reuse Bias

In this section, we first introduce a novel concept of *Reuse Bias* to measure the bias in current off-policy evaluation methods. Then, we show that the reuse of the replay buffer causes overestimation and provide a high probability upper bound of the Reuse Bias with some insights into this result.

4.1 Reuse Bias

Assuming that the hypothesis set of policies is $\mathcal{H}$, for any off-policy algorithm, we formalize it as a mapping $\mathcal{O} : \mathcal{H} \times \Gamma^m \rightarrow \mathcal{H}$ that takes an initialized policy $\pi_0 \in \mathcal{H}$ and a replay buffer $\mathcal{B} \in \Gamma^m$ as inputs, and outputs a policy $\mathcal{O}(\pi_0, \mathcal{B}) \in \mathcal{H}$. In off-policy evaluation, we use the importance sampling estimate $\hat{J}_{\hat{\Pi}, \mathcal{B}}(\mathcal{O}(\pi_0, \mathcal{B}))$ defined in Eq. (2) as an approximation of the expected return $J(\mathcal{O}(\pi_0, \mathcal{B}))$. We define their difference as the Reuse Error of an off-policy algorithm $\mathcal{O}$ on π_0 as well as $\mathcal{B}$, and the expectation of the Reuse Error as the Reuse Bias:

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

$$\epsilon_{\text{RE}}(\mathcal{O}, \pi_0, \mathcal{B}) \triangleq \hat{J}_{\hat{\Pi}, \mathcal{B}}(\mathcal{O}(\pi_0, \mathcal{B})) - J(\mathcal{O}(\pi_0, \mathcal{B})). \tag{4}$$

Moreover, we define is expectation as the Reuse Bias:

$$\epsilon_{\text{RB}}(\mathcal{O}, \pi_0) \triangleq \mathbb{E}_{\mathcal{B}}[\epsilon_{\text{RE}}(\mathcal{O}, \pi_0, \mathcal{B})]. \tag{5}$$

It is well-known that importance sampling is an unbiased estimation when the estimated distribution is independent of the samples. In other words, the off-policy evaluation is unbiased, i.e., $\epsilon_{\text{RB}}(\mathcal{O}, \pi_0) = \mathbb{E}_{\mathcal{B}}[\epsilon_{\text{RE}}(\mathcal{O}, \pi_0, \mathcal{B})] = 0$, when $\mathcal{O}(\pi_0, \mathcal{B})$ is independent of $\mathcal{B}$. However, off-policy algorithms [Haarnoja *et al.*, 2018; Fujimoto *et al.*, 2018] utilize trajectories in the replay buffer $\mathcal{B}$ to optimize the policy. In this case, the policy $\mathcal{O}(\pi_0, \mathcal{B})$ does depend on $\mathcal{B}$ which makes the off-policy evaluation no longer unbiased. And we will provide a first systematic investigation of the Reuse Bias, as detailed below.

We first assume that the off-policy algorithm $\mathcal{O}^*$ satisfies the condition that $\mathcal{O}^*(\pi_0, \mathcal{B})$ has the highest estimated performance with $\mathcal{B}$ and consider the Reuse Bias under this assumption. This assumption is mild and practical since off-policy algorithms always hope to maximize the expected performance of the policy w.r.t. the replay buffer in practice. Under this assumption, the estimation $\hat{J}_{\hat{\Pi}, \mathcal{B}}(\mathcal{O}^*(\pi_0, \mathcal{B}))$ is an overestimation of $J(\mathcal{O}^*(\pi_0, \mathcal{B}))$ as summarized below:

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

$$\mathcal{O}^*(\pi_0, \mathcal{B}) = \arg\max_{\pi \in \mathcal{H}} \hat{J}_{\hat{\Pi}, \mathcal{B}}(\pi)$$

$$= \arg\max_{\pi \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} \left[\frac{p^\pi(\tau_i)}{p^{\hat{\pi}_i}(\tau_i)} R(\tau_i) \right]. \tag{6}$$

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., $\epsilon_{\mathrm{RB}}(\mathcal{O}^*,\pi_0) = \mathbb{E}_{\mathcal{B}\sim\hat{\Pi}}\left[\epsilon_{\mathrm{RE}}(\mathcal{O}^*,\pi_0,\mathcal{B})\right]$ ≥ 0. If the equality holds, then for any $\mathcal{B},\mathcal{B}' \sim \hat{\Pi}$, we have

$$\mathcal{O}^*(\pi_0,\mathcal{B}) = \arg\max_{\pi\in\mathcal{H}} \hat{J}_{\hat{\Pi},\mathcal{B}'}(\pi). \tag{7}$$

However, in practice, we always optimize our policy through several steps of gradient descent rather than directly getting the optimal policy $\mathcal{O}^*(\pi_0,\mathcal{B})$ over the replay buffer $\mathcal{B}$. To analyze the Reuse Bias in more practical cases, we consider the one-step policy gradient optimization and prove that the estimation $\hat{J}_{\hat{\Pi},\mathcal{B}}$ is strictly an overestimation of J in this case only under some mild assumptions as:

Theorem 2 (Overestimation for One-Step PG, Proof in Appendix A.2). *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*

$$\theta' = \theta + \alpha\nabla_\theta\hat{J}_{\hat{\Pi},\mathcal{B}}(\pi_\theta), \tag{8}$$

*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.,*

$$\mathbb{E}_{\mathcal{B}\sim\hat{\Pi}}\hat{J}_{\hat{\Pi},\mathcal{B}}(\pi_{\theta'}) > \mathbb{E}_{\mathcal{B}\sim\hat{\Pi}}J(\pi_{\theta'}). \tag{9}$$

Furthermore, to explain how severe the Reuse Bias will be, we provide Theorem 3 to show that the estimated return of $\mathcal{O}^*(\pi_0,\mathcal{B})$ might be arbitrarily large when its true return is arbitrarily small, which indicates that the Reuse Bias will result in the degeneration of the policy optimization and the final performance.

Theorem 3 (Proof in Appendix A.3). *For any fixed replay buffer size n, assume that Π is the set of random policies, and the reward of any state-action pair (s,a) is not negative, i.e., $\mathcal{R}(s,a) \geq 0$, then there exists an environment satisfying that $\mathbb{E}_{\mathcal{B}\sim\hat{\Pi}}[J(\mathcal{O}^*(\pi_0,\mathcal{B}))]$ is arbitrarily small when $\mathbb{E}_{\mathcal{B}\sim\hat{\Pi}}\left[\hat{J}_{\hat{\Pi},\mathcal{B}}(\mathcal{O}^*(\pi_0,\mathcal{B}))\right]$ is arbitrarily large. In other words, for $\forall n, M \geq 1, \epsilon > 0, \exists$ an environment satisfying that*

$$J(\pi^*) = 1, |\mathcal{B}| = n, \mathbb{E}_{\mathcal{B}\sim\hat{\Pi}}\left[J(\mathcal{O}^*(\pi_0,\mathcal{B}))\right] \leq \epsilon,$$
$$\mathbb{E}_{\mathcal{B}\sim\hat{\Pi}}\left[\hat{J}_{\hat{\Pi},\mathcal{B}}(\mathcal{O}^*(\pi_0,\mathcal{B}))\right] \geq M. \tag{10}$$

here π^* is the optimal policy of this environment.

In Theorem 3, we point out that the Reuse Bias may cause an erroneous off-policy policy evaluation, which might be sufficiently severe, and affect further policy improvement, which is also shown in experiments of MiniGrid (See Sec. 6.2).

4.2 High Probability Bound for Reuse Error

Since the Reuse Bias may mislead the optimization direction and further affect the performance, which cannot be ignored, it is important to analyze the connection of the replay buffer $\mathcal{B}$ and the Reuse Error to better control it. Without losing generality, we mainly consider the case that all trajectories are sampled from the same original policy $\hat{\pi}$.

When the hypothesis set has bounded statistical complexity, for example, $\mathcal{H}$ is finite or its VC-dimension is finite, we can

naturally provide a high-probability upper bound of the Reuse Error via using some classical results of statistical learning theory [Shalev-Shwartz and Ben-David, 2014] (We provide the result when $\mathcal{H}$ is finite in Appendix A.4 as an example). However, these results are based on the statistic complexity of the hypothesis set and $\max_{\pi,\tau}\left[\frac{p^\pi(\tau)}{p^{\hat{\pi}}(\tau)}\right]$. In practice, the statistic complexity of the hypothesis set is extremely huge while the sample number m is relatively small, and $\max_{\pi,\tau}\left[\frac{p^\pi(\tau)}{p^{\hat{\pi}}(\tau)}\right]$ is the term related on the environment as well as the hypothesis set, which can not be optimized in the training stage.

Here, we provide another high-probability upper bound for the Reuse Error as below. Compared with prior results, our bound holds for all $\mathcal{H}$ even with unbounded statistical complexity, and is related with the optimized policy $\mathcal{O}(\pi_0,\mathcal{B})$, which guides us to optimize $\mathcal{O}(\pi_0,\mathcal{B})$ by controlling Reuse Error.

Theorem 4 (High-Probability Bound for Reuse Error, Proof in Appendix A.5). *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:*

$$|\epsilon_{\mathrm{RE}}(\mathcal{O},\pi_0,\mathcal{B})| \leq \sqrt{\frac{m\epsilon_1 + \log\left(\frac{m^2}{\delta}\right)}{m-1}} + \epsilon_2, \tag{11}$$

where ϵ_1 and ϵ_2 are defined as:

$$\epsilon_1 = \mathrm{KL}[p^{\mathcal{O}(\pi_0,\mathcal{B})}(\cdot)||p^{\hat{\pi}}(\cdot)],$$
$$\epsilon_2 = \frac{1}{m}\sum_{i=1}^m\left|1 - \frac{p^{\mathcal{O}(\pi_0,\mathcal{B})}(\tau_i)}{p^{\hat{\pi}}(\tau_i)}\right| = \mathbb{E}_{\tau\sim\mathcal{B}}\left|1 - \frac{p^{\mathcal{O}(\pi_0,\mathcal{B})}(\tau)}{p^{\hat{\pi}}(\tau)}\right|. \tag{12}$$

In Theorem 4, ϵ_1, as the KL divergence of the trajectory distribution under $\mathcal{O}(\pi_0,\mathcal{B})$ and $\hat{\pi}$, indicates their similarity over the trajectory space Γ. Moreover, ϵ_2 only focuses on the similarity in the replay buffer $\mathcal{B}$ of these two policies. Then we will introduce some insights into this theorem.

First, when the replay buffer size m tends to infinity, our high-probability bound will converge to $\sqrt{\epsilon_1} + \epsilon_2$ rather than 0. This is reasonable because the policy $\mathcal{O}(\pi_0,\mathcal{B})$ depends on $\mathcal{B}$ and the hypothesis space in our results may be arbitrary large since the statistical complexity of practical policies might be very huge or even difficult to analyze. Also, for better explaining that the Reuse Bias may still exist even when the sample number m tends to infinity, we conduct a concrete example satisfying that $|\epsilon_{\mathrm{RE}}(\mathcal{O},\hat{\pi},\mathcal{B})| = 1$ holds for any finite m and any replay buffer $\mathcal{B}$ in Appendix C.

Moreover, the high-probability upper bound depends on the KL divergence of these two policies as well as on their probability over $\mathcal{B}$. Consequently, controlling their divergence, especially over $\mathcal{B}$, guides us on how to optimize the policy for controlling the Reuse Bias.

5 Methodology

In this section, introducing the concept of stability for off-policy algorithms, we further analyze the connection between

controlling ϵ_2 in Theorem 4 and controlling the Reuse Bias. Moreover, we propose a general framework for controlling Reuse Bias with practical algorithms by extending actor-critic methods for complicated continuous control tasks.

5.1 Theoretical Analysis on Controlling Reuse Bias

Inaccurate off-policy evaluation caused by Reuse Bias may lead the policy optimization to the wrong direction and further affect both the performance and sample utilization. Therefore, it is necessary to develop methods to control the Reuse Bias.

Based on Theorem 4, we can control ϵ_1 and ϵ_2 to control the Reuse Bias. However, directly using the trajectories in the replay buffer to estimate ϵ_1, which is the KL divergence of $p^{\mathcal{O}(\pi_0,\mathcal{B})}$ and $p^{\hat{\pi}}$, may also be biased because $\mathcal{O}(\pi_0,\mathcal{B})$ is dependent on the replay buffer. Thus we only consider constraining ϵ_2 and use $\mathcal{L}(\pi,\mathcal{B})$ to denote it, i.e.,

$$\mathcal{L}(\pi,\mathcal{B}) \triangleq \mathbb{E}_{\tau \sim \mathcal{B}} \left| 1 - \frac{p^\pi(\tau)}{p^{\hat{\pi}}(\tau)} \right|. \tag{13}$$

Moreover, we hope to analyze the Reuse Bias of policies with bounded $\mathcal{L}(\pi,\mathcal{B})$. Borrowing ideas of stability [Hardt *et al.*, 2016], we similarly define the concept of stability for off-policy algorithms and further provide an upper bound of stable off-policy algorithms as shown below

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*

$$\forall \tau, \pi_0, \quad \mathbb{E}_{\mathcal{O}} \left[p^{\mathcal{O}(\pi_0,\mathcal{B})}(\tau) - p^{\mathcal{O}(\pi_0,\mathcal{B}')}(\tau) \right] \le \beta. \tag{14}$$

Theorem 5 (Bound for the Reuse Error of Stable Algorithm, Proof in Appendix A.6). *Suppose a randomized off-policy algorithm $\mathcal{O}$ is β-uniformly stable, then we can prove that*

$$\forall \pi_0, \quad |\mathbb{E}_{\mathcal{B} \sim \hat{\pi}} \mathbb{E}_{\mathcal{O}} \left[\epsilon_{\mathrm{RE}}(\mathcal{O},\pi_0,\mathcal{B}) \right]| \le \beta. \tag{15}$$

Theorem 5 controls Reuse Bias for stable off-policy algorithms. Furthermore, we provide Theorem 6 to point out that, under some mild assumptions, Reuse Bias of off-policy stochastic policy gradient can be controlled just by $\mathcal{L}(\pi,\mathcal{B})$.

Theorem 6 (Details and Proof are in Appendix A.7). *We assume that the policy π_θ is parameterized with θ, and $|\nabla_\theta \log p^{\pi_\theta}(\tau)| \le L_1$ 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 $\mathcal{L}(\pi,\mathcal{B}) \le M$, 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.*

5.2 Bias-Regularized Importance Sampling Framework

Based on above analyses, for any given off-policy algorithm $\mathcal{O}$, we focus on controlling $\mathcal{L}(\mathcal{O}(\pi_0,\mathcal{B}),\mathcal{B})$. To simplify the notation, we denote $\pi \triangleq \mathcal{O}(\pi_0,\mathcal{B})$ as the policy trained by the off-policy algorithm. For any trajectory τ, we have

$$\frac{p^\pi(\tau)}{p^{\hat{\pi}}(\tau)} = \frac{\mu(s_0)}{\mu(s_0)} \prod_i \left[\frac{\pi(a_i|s_i)\mathcal{P}(s_{i+1}|s_i,a_i)}{\hat{\pi}(a_i|s_i)\mathcal{P}(s_{i+1}|s_i,a_i)} \right] = \prod_i \frac{\pi(a_i|s_i)}{\hat{\pi}(a_i|s_i)}. \tag{16}$$

Thus, we can derive $\mathcal{L}(\pi,\mathcal{B})$ as

$$\mathcal{L}(\pi,\mathcal{B}) = \mathbb{E}_{\tau \sim \mathcal{B}} \left| 1 - \frac{p^\pi(\tau)}{p^{\hat{\pi}}(\tau)} \right| = \mathbb{E}_{\tau \sim \mathcal{B}} \left| 1 - \prod_i \frac{\pi(a_i|s_i)}{\hat{\pi}(a_i|s_i)} \right|. \tag{17}$$

Consequently, we propose a novel off-policy reinforcement learning framework named Bias-Regularized Importance Sampling (BIRIS) for controlling the Reuse Bias as below

$$\mathcal{O}_{\mathrm{BIRIS}}(\pi_0,\mathcal{B}) = \arg\min_{\pi \in \mathcal{H}} \mathcal{L}_{\mathrm{BIRIS}}(\pi,\mathcal{B}),$$
$$\text{where } \mathcal{L}_{\mathrm{BIRIS}}(\pi,\mathcal{B}) = \mathcal{L}_{\mathrm{RL}}(\pi,\mathcal{B}) + \alpha\mathcal{L}(\pi,\mathcal{B}), \tag{18}$$

here $\mathcal{L}_{\mathrm{RL}}$ denotes the nominal loss function of the basic off-policy algorithm, and $\alpha \ge 0$ is a hyperparameter to control the trade-off between standard RL loss and regularized loss.

5.3 Connection with Actor-Critic Methods

However, since directly using the return of trajectories to optimize the policy suffers from instability, many off-policy practical methods for complicated environments consider using TD-based actor-critic methods, like Soft Actor Critic (SAC) [Haarnoja *et al.*, 2018] and Twin Delayed Deep Deterministic Policy Gradient (TD3) [Fujimoto *et al.*, 2018]. These methods always alternately optimize the critic and the actor as below

- Optimizing the Q function via TD-learning, i.e., minimizing $\sum_{(s,a) \in \mathcal{B}}(Q_\theta(s,a) - Q_{\mathrm{target}}(s,a))^2$
- Optimize the policy via the Q function, i.e., maximizing $\sum_{s \in \mathcal{B}} Q_\theta(s,\pi_\phi(s))$

Current work mainly considers the error in Q function optimization caused by out-of-distribution data. For example, Extrapolation Error [Fujimoto *et al.*, 2019] is because erroneously estimating target Q values of unseen state-action pairs.

However, similar to our analyses above, in the policy optimization stage, reusing the state-action pairs in the replay buffer will also cause the Reuse Error, and we can also implement our BIRIS to reduce the negative impact. Moreover, for complicated environments, trajectories are so long that their probabilities are difficult to calculate and numerically unstable. If we assume $\left| \frac{\pi(a|s)}{\hat{\pi}(a|s)} \right| \le \epsilon$ for all $(s,a) \in \mathcal{B}$ and the length of the trajectory is T, we can show that

$$\left| \prod_{i=1}^T \frac{\pi(a_i|s_i)}{\hat{\pi}(a_i|s_i)} - 1 \right| \le \max(1-(1-\epsilon)^T, (1+\epsilon)^T - 1). \tag{19}$$

Consequently, for handling complicated environments, we use the state-level ratio $\mathcal{L}_{\mathrm{BR}}$ as a surrogate

$$\mathcal{L}_{\mathrm{BR}}(\pi,\mathcal{B}) \triangleq \mathbb{E}_{(s,a) \in \mathcal{B}} \left| \frac{\pi(a|s)}{\hat{\pi}(a|s)} - 1 \right|. \tag{20}$$

Our $\mathcal{L}_{\mathrm{BR}}$ has some similar points with pessimistic offline reinforcement learning methods [Kumar *et al.*, 2020b; Liu *et al.*, 2020; Fujimoto and Gu, 2021] and is a supplement to the effectiveness of pessimistic methods based on our theoretical analyses on the view of the Reuse Bias. Furthermore, for handling complicated continuous control tasks, we concretize our framework to specific algorithms, like SAC and TD3, of which the core is to calculate $\frac{\pi(a|s)}{\hat{\pi}(a|s)}$ for any state-action pair $(s,a) \in \mathcal{B}$ (More details are in Appendix D).

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**

Table 1: Relative Reuse Bias of policies trained by PG+IS, PG+WIS, PG+IS+BIRIS, and PG+WIS+BIRIS in different MiniGrid environments.

6 Experiments

We now present empirical results to answer the questions:

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

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

6.1 Experiment Setup

We experiment in two different domains to answer the questions above. Before that, we first provide a short description of environments, including algorithms as well as metrics.

Gridworld. For the first question, we experiment in Mini-Grid[1], which includes different shapes of grids with discrete state space and action space, and is simple for optimizing and evaluating our policy. We will calculate and compare the Reuse Bias of PG+IS, PG+WIS [Mahmood *et al.*, 2014], PG+IS+BIRIS, and PG+WIS+BIRIS.

Simulated Robotics. To evaluate BIRIS in complicated environments, we choose several continuous control tasks from MuJoCo and use two popular off-policy algorithms: SAC [Haarnoja *et al.*, 2018] and TD3 [Fujimoto *et al.*, 2018], as baselines, with uniform sampling and prioritization experience replay (PER) [Schaul *et al.*, 2016]. Our implementation is based on Tianshou [Weng *et al.*, 2022].

6.2 Results for Reuse Bias

We first present our experimental results in MiniGrid to answer the first question, i.e., how severe the Reuse Bias is for practical off-policy evaluation and whether our BIRIS can significantly control the Reuse Bias. For different grids (5 × 5, 5 × 5-random, 6 × 6, 6 × 6-random, 8 × 8, 16 × 16), we initialize our policy parameterized with a simple three-layer convolutional neural network. Moreover, we choose 30, 40, and 50 as the replay buffer size and sample trajectories to fill the replay buffer, to test its impact. Then we use PG [Sutton *et al.*, 2000] and PG+BIRIS, respectively, to optimize the policy by maximizing the estimated return via replay buffer by IS and weighted importance sampling (WIS) [Mahmood *et al.*, 2014]. In the test stage, we compute the return of our policy estimated

---

[1]https://github.com/mit-acl/gym-minigrid

by the replay buffer as well as the true expected return of our policy. To mitigate the impact of randomness caused by the replay buffer and sampling, we repeat our experiments for each task and each algorithm with 50 different random seeds.

We report part of our results (8×8 with the size of replay buffer 30) in Fig. 1(2) in the form of box diagrams (All figures are in Appendix B.1). In each subgraph, the distribution of the true return J of the optimized policy is shown on the left (filled in with blue) while the distribution of the estimated return $\hat{J}$ is on the right (filled in with green). In each box diagram, the horizontal lines at the top and bottom indicate the largest and smallest value respectively, while the horizontal lines at the top and bottom of the box indicate the first quartile and third quartile respectively. Moreover, the orange dotted line and the green plus sign represent the median and the expectation separately. In Table 1, we also report the Relative Reuse Bias, i.e., $\epsilon_{RB}(\mathcal{O}, \pi_0, \mathcal{B})/J(\mathcal{O}(\pi_0, \mathcal{B}))$ of all algorithms.

As shown in Fig. 1(2) and Table 1, the estimated return is obviously higher than the true return in all environments for PG+IS, which matches our theoretical analyses, i.e., using the same replay buffer to optimize and evaluate the policy will lead to overestimation. Furthermore, as shown in Table 1, the Relative Reuse Bias shown in this experiment is greater when the grid is larger and more complicated, which means that the Reuse Bias will be more serious in more complicated environments. It is probably because the policy is more likely to converge to local optimization in complicated environments.

As shown in Table 1, when the replay buffer size increases, the relative Reuse Bias of PG+IS will reduce to some degree but is still large even if the size is 50. Moreover, the relative Reuse Bias of PG+WIS is lower than it of PG+IS but is still relatively large, which shows that variance-reduced methods like WIS can somehow reduce the impact of the Reuse Bias. Finally, in almost all tasks and all replay buffer sizes, the Relative Reuse Bias of PG+IS+BIRIS and PG+WIS+BIRIS are conspicuously smaller than PG+IS and PG+WIS, which means that our BIRIS can significantly reduce the Reuse Bias.

6.3 Results for Simulated Robotics

We now report results to evaluate BIRIS-based methods in more complicated environments against actor-critic methods. We compare the performance of our methods against common off-policy algorithms like SAC and TD3 in MuJoCo envi-

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**

Table 2: Cumulative reward (mean ± one std) of the best policy trained by SAC, SAC+PER, SAC+BIRIS, TD3, TD3+PER, TD3+BIRIS in different MuJoCo games. In each column, we **bold** the best performance over all algorithms.
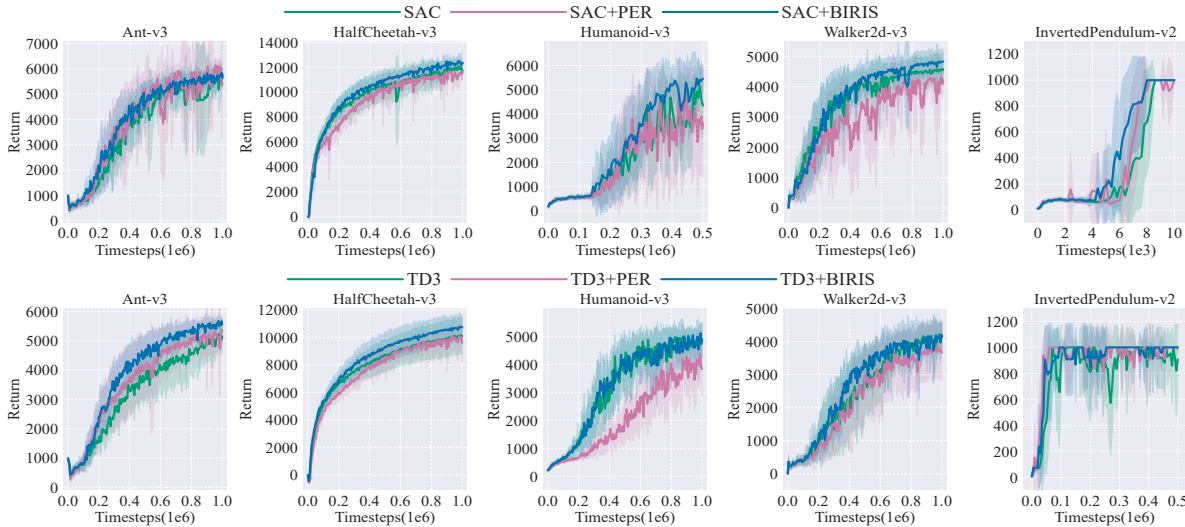

Figure 2: Cumulative reward curves for SAC, SAC+PER, SAC+BIRIS, TD3, TD3+PER, and TD3+BIRIS. The x-axes indicate the number of steps interacting with the environment, and the y-axes indicate the performance of the agent, including average rewards with std.

ronments, including Ant, Halfcheetah, Humanoid, Walker2d, and InvertedPendulum. To mitigate the effects of randomness caused by stochastic environments and policies, we train 10 policies with different random seeds for each algorithm in each task and plot the mean with standard deviations of these 10 policies as a function of timestep in the training stage. Moreover, we use the solid line and the part with a lighter color to represent the average reward and standard deviations of 10 strategies, respectively. We also report the optimal mean and standard deviations of the cumulative return of 10 policies trained by each algorithm in each environment in Table 2. As shown in this table, PER does not outperform uniform sampling, which is also reported in [Novati and Koumoutsakos, 2019] for DDPG [Lillicrap *et al.*, 2016] in MuJoCo.

In Fig. 2 and Table 2, we compare SAC, SAC+PER, and SAC+BIRIS. As we can see, SAC+BIRIS learns a better policy compared with SAC, i.e., SAC+BIRIS can effectively improve the performance and the sample efficiency compared with SAC. Overall, the performance of SAC+BIRIS increases more smoothly in the training stage, especially in HalfCheetah and Humanoid. This is consistent with our theoretical results, in that BIRIS can better estimate off-policy performance and provide a more precise optimization direction.

The comparison of TD3, TD3+PER, and TD3+BIRIS is reported in Fig. 2 and Table 2. Similarly, for all tasks,

TD3+BIRIS performs better than TD3, i.e., TD3+BIRIS can improve the performance as well as the sample efficiency compared with TD3, especially in Ant and HalfCheetah. We also notice that TD3+BIRIS can conspicuously reduce the standard deviations of the policies compared with TD3 in Ant and HalfCheetah. This is also because BIRIS can better reduce the Reuse Bias and provide a more stable optimization direction. Moreover, we also do some ablation studies in Appendix B.2.

7 Conclusion

In this paper, we first show that the bias in off-policy evaluation is problematic and introduce the concept of Reuse Bias to describe it. We theoretically prove the overestimation bias in off-policy evaluation, which is because of reusing historical trajectories in the same replay buffer. Moreover, we provide a high-probability upper bound of Reuse Bias as well as an expectation upper bound of Reuse Bias for stable off-policy algorithms. Based on these analyses, we provide a framework of Bias-Regularized Importance Sampling (BIRIS) with practical algorithms for controlling Reuse Bias. Experimental results demonstrate that BIRIS can significantly reduce Reuse Bias and mitigating Reuse Bias via BIRIS can significantly improve the sample efficiency for off-policy methods in MuJoCo. As our modifications are independent of existing variance reduction methods, they can be easily integrated.

Ethical Statement

Off-policy methods are important to improve sample efficiency and apply RL in larger scenarios. This paper studies the Reuse Bias of off-policy evaluation and is beneficial for further studying more efficient and stable off-policy methods. There are no serious ethical issues as this is basic research.

Acknowledgments

References

[Chandak *et al.*, 2021a] Yash Chandak, Scott Niekum, Bruno da Silva, Erik Learned-Miller, Emma Brunskill, and Philip S Thomas. Universal off-policy evaluation. *Advances in Neural Information Processing Systems*, 34:27475–27490, 2021.

[Chandak *et al.*, 2021b] Yash Chandak, Shiv Shankar, and Philip S Thomas. High-confidence off-policy (or counterfactual) variance estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6939–6947, 2021.

[Fujimoto and Gu, 2021] Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *Advances in neural information processing systems*, 34:20132–20145, 2021.

[Fujimoto *et al.*, 2018] Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International Conference on Machine Learning*, pages 1587–1596. PMLR, 2018.

[Fujimoto *et al.*, 2019] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, pages 2052–2062. PMLR, 2019.

[Fujimoto *et al.*, 2021] Scott Fujimoto, David Meger, and Doina Precup. A deep reinforcement learning approach to marginalized importance sampling with the successor representation. In *International Conference on Machine Learning*, pages 3518–3529. PMLR, 2021.

[Garcıa and Fernández, 2015] Javier Garcıa and Fernando Fernández. A comprehensive survey on safe reinforcement learning. *Journal of Machine Learning Research*, 16(1):1437–1480, 2015.

[Haarnoja *et al.*, 2018] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. PMLR, 2018.

[Hanna *et al.*, 2017] Josiah P Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[Hanna *et al.*, 2019] Josiah Hanna, Scott Niekum, and Peter Stone. Importance sampling policy evaluation with an estimated behavior policy. In *International Conference on Machine Learning*, pages 2605–2613. PMLR, 2019.

[Hardt *et al.*, 2016] Moritz Hardt, Ben Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning*, pages 1225–1234. PMLR, 2016.

[Jiang and Li, 2016] Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 652–661. PMLR, 2016.

[Kahn and Marshall, 1953] Herman Kahn and Andy W Marshall. Methods of reducing sample size in monte carlo computations. *Journal of the Operations Research Society of America*, 1(5):263–278, 1953.

[Kumar *et al.*, 2020a] Aviral Kumar, Abhishek Gupta, and Sergey Levine. Discor: Corrective feedback in reinforcement learning via distribution correction. *Advances in Neural Information Processing Systems*, 33:18560–18572, 2020.

[Kumar *et al.*, 2020b] Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

[Kuzborskij *et al.*, 2021] Ilja Kuzborskij, Claire Vernade, Andras Gyorgy, and Csaba Szepesvári. Confident off-policy evaluation and selection through self-normalized importance weighting. In *International Conference on Artificial Intelligence and Statistics*, pages 640–648. PMLR, 2021.

[Li *et al.*, 2015] Lihong Li, Rémi Munos, and Csaba Szepesvári. Toward minimax off-policy value estimation. In *Artificial Intelligence and Statistics*, pages 608–616. PMLR, 2015.

[Lillicrap *et al.*, 2016] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2016.

[Liu *et al.*, 2018] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: infinite-horizon off-policy estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 5361–5371, 2018.

[Liu *et al.*, 2020] Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch off-policy reinforcement learning without great exploration. *Advances in neural information processing systems*, 33:1264–1274, 2020.

[Liu *et al.*, 2021] Xu-Hui Liu, Zhenghai Xue, Jingcheng Pang, Shengyi Jiang, Feng Xu, and Yang Yu. Regret minimization experience replay in off-policy reinforcement learning. *Advances in Neural Information Processing Systems*, 34:17604–17615, 2021.

[Mahmood *et al.*, 2014] A Rupam Mahmood, Hado P van Hasselt, and Richard S Sutton. Weighted importance sampling for off-policy learning with linear function approximation. *Advances in Neural Information Processing Systems*, 27:3014–3022, 2014.

[McAllester, 1998] David A. McAllester. Some pac-bayesian theorems. In Peter L. Bartlett and Yishay Mansour, editors, *Proceedings of the Eleventh Annual Conference on Computational Learning Theory, COLT 1998, Madison, Wisconsin, USA, July 24-26, 1998*, pages 230–234. ACM, 1998.

[McAllester, 2003] David A. McAllester. Simplified pac-bayesian margin bounds. In Bernhard Schölkopf and Manfred K. Warmuth, editors, *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel 2003, Washington, DC, USA, August 24-27, 2003, Proceedings*, volume 2777 of *Lecture Notes in Computer Science*, pages 203–215. Springer, 2003.

[Metelli *et al.*, 2018] Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. *Advances in Neural Information Processing Systems*, 31, 2018.

[Metelli *et al.*, 2021] Alberto Maria Metelli, Alessio Russo, and Marcello Restelli. Subgaussian and differentiable importance sampling for off-policy evaluation and learning. *Advances in Neural Information Processing Systems*, 34:8119–8132, 2021.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin A. Riedmiller, Andreas Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[Novati and Koumoutsakos, 2019] Guido Novati and Petros Koumoutsakos. Remember and forget for experience replay. In *International Conference on Machine Learning*, pages 4851–4860. PMLR, 2019.

[Pavse *et al.*, 2020] Brahma Pavse, Ishan Durugkar, Josiah Hanna, and Peter Stone. Reducing sampling error in batch temporal difference learning. In *International Conference on Machine Learning*, pages 7543–7552. PMLR, 2020.

[Precup *et al.*, 2000] Doina Precup, Richard S Sutton, and Satinder P Singh. Eligibility traces for off-policy policy evaluation. In *ICML*, 2000.

[Precup *et al.*, 2001] Doina Precup, Richard S Sutton, and Sanjoy Dasgupta. Off-policy temporal-difference learning with function approximation. In *ICML*, pages 417–424, 2001.

[Schaul *et al.*, 2016] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. Prioritized experience replay. In *ICLR (Poster)*, 2016.

[Shalev-Shwartz and Ben-David, 2014] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, USA, 2014.

[Shi *et al.*, 2021] Chengchun Shi, Runzhe Wan, Victor Chernozhukov, and Rui Song. Deeply-debiased off-policy interval estimation. In *International Conference on Machine Learning*, pages 9580–9591. PMLR, 2021.

[Su *et al.*, 2020] Yi Su, Maria Dimakopoulou, Akshay Krishnamurthy, and Miroslav Dudík. Doubly robust off-policy evaluation with shrinkage. In *International Conference on Machine Learning*, pages 9167–9176. PMLR, 2020.

[Sutton *et al.*, 2000] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems (NeurIPS)*, pages 1057–1063, 2000.

[Thomas and Brunskill, 2016] Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.

[Thomas *et al.*, 2015] Philip Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, 2015.

[Van Hasselt *et al.*, 2016] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

[Weng *et al.*, 2022] Jiayi Weng, Huayu Chen, Dong Yan, Kaichao You, Alexis Duburcq, Minghao Zhang, Yi Su, Hang Su, and Jun Zhu. Tianshou: A highly modularized deep reinforcement learning library. *Journal of Machine Learning Research*, 23(267):1–6, 2022.

[Xie *et al.*, 2019] T Xie, Y Ma, and YX Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. *Advances in neural information processing systems*, 2019.

[Ying *et al.*, 2022] ChengYang Ying, Xinning Zhou, Hang Su, Dong Yan, Ning Chen, and Jun Zhu. Towards safe reinforcement learning via constraining conditional value-at-risk. In Lud De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3673–3680. International Joint Conferences on Artificial Intelligence Organization, 7 2022. Main Track.

[Zhang *et al.*, 2021] Junyu Zhang, Chengzhuo Ni, Csaba Szepesvari, Mengdi Wang, et al. On the convergence and sample efficiency of variance-reduced policy gradient method. *Advances in Neural Information Processing Systems*, 34:2228–2240, 2021.