Doubly Robust Off-Policy Actor-Critic Algorithms for Reinforcement Learning

Riashat Islam
McGill University, Mila
School of Computer Science
riashat.islam@mail.mcgill.ca

Raihan Seraj
McGill Univeristy
Electrical and Computer Engineering
raihan.seraj@mail.mcgill.ca

Samin Yeasar Arnob
McGill Univeristy
Electrical and Computer Engineering
samin.arnob@mail.mcgill.ca

Doina Precup
McGill University, Mila
School of Computer Science
dprecup@cs.mcgill.ca

Abstract

We study the problem of off-policy critic evaluation in several variants of value-based off-policy actor-critic algorithms. Off-policy actor-critic algorithms require an off-policy critic evaluation step, to estimate the value of the new policy after every policy gradient update. Despite enormous success of off-policy policy gradients on control tasks, existing general methods suffer from high variance and instability, partly because the policy improvement depends on gradient of the estimated value function. In this work, we present a new way of off-policy policy evaluation in actor-critic, based on the doubly robust estimators. We extend the doubly robust estimator from off-policy policy evaluation (OPE) to actor-critic algorithms that consist of a reward estimator performance model. We find that doubly robust estimation of the critic can significantly improve performance in continuous control tasks. Furthermore, in cases where the reward function is stochastic that can lead to high variance, doubly robust critic estimation can improve performance under corrupted, stochastic reward signals, indicating its usefulness for robust and safe reinforcement learning.

1 Introduction

Policy gradient based methods are widely popular in deep reinforcement learning (RL) for solving continuous control tasks [Schulman et al., 2015]. Several variants of off-policy value gradient based methods have been proposed recently [Haarnoja et al., 2018, Lillicrap et al., 2015] with the goal to solve complex manipulation while being sample efficient due to the ability to re-use off-policy data. Often deep RL policy gradient algorithms rely on using an off-policy estimate of the value function based on which the policy parameters can be directly updated by finding gradients directly through the value function. Existing literature in RL on off-policy evaluation has a long history [Precup, 2000, Jiang and Li, 2016] where the goal is to estimate the value of a policy using data sampled from another behaviour policy. Off-policy methods generally suffer from high variance due to importance sampling corrections, although several approaches have introduced bias by learning a performance model to reduce variance. Additionally, in several control and robotic tasks, the reward function may be corrupted or noisy, e.g rewards from sensors. Stochastic rewards may often make the off-policy learning process even more difficult, especially for learning complex behaviours in robotic applications. The existence of corrupted reward signals can serve as a severe bottleneck towards scaling up RL algorithms for practical applications, with the goal towards robust decision making.

Preprint. Under review.

Several existing work have proposed for conservative policy iteration [Kakade and Langford, 2002] and safe policy improvement [Pirotta et al., 2013], where an important motivation for off-policy evaluation is to guarantee safety before the policy can be deployed in the real world world. While off-policy evaluation (OPE) approaches make use of past data for policy evaluation and have been shown to be beneficial for practical tasks, most of the success with robotic applications have come from policy gradient based methods for continuous control [Schulman et al., 2015, Lillicrap et al., 2015]. Therefore, it is critical to think about safety measures in off-policy actor-critic algorithms, since they have have been shown to be most sample efficient in control tasks. Several off-policy deep RL policy gradient based algorithms rely on off-policy critic evaluation, as in off-policy actor-critic based methods such as the widely used Deep Deterministic Policy Gradient (DDPG) algorithm. However, is it is surprising to see that the gap in literature between off-policy evaluation (OPE) problems and those methods being extended to the control and policy gradient setting.

In this work, we extend the existing doubly robust estimators for off-policy evaluation (OPE) to the control setting, in off-policy actor-critic algorithms. In off-policy value-based policy gradient algorithms such as DDPG and SAC [Lillicrap et al., 2015, Haarnoja et al., 2018], the critic evaluates the performance of a policy and the policy is improved based on the critic estimate. Often existing value-based policy gradient algorithms, however, suffer from high variance and instability particularly in continuous control tasks [Henderson et al., 2018, Islam et al., 2017]. This problem can be further exacerbated in practical sensory-motor driven robotic applications where the reward functions are often noisy and corrupted. As such, existing off-policy policy gradient algorithms would be quite unreliable for use in the real world. We propose doubly robust estimation for critic evaluation towards the goal of reducing variance in the critic estimates, often better stability and safe improvements in performance. Furthermore, we find that using a reward function estimator in the case of noisy rewards as demonstrated in [Romoff et al., 2018] can be quite useful for doubly robust off-policy actor-critic algorithms under stochastic rewards, further justifying that these estimators can possibly play the role of a control variate in policy gradients [Thomas and Brunskill, 2016a].

We aim to merge the gap between off-policy evaluation (OPE) estimators with guarantees towards unbiasedness and low variance, and estimators used in off-policy actor-critic based methods. Our goal is to achieve low variance regression based critic evaluation based on minimizing the mean squared error between next state critic target and current critic estimates, while keeping the critic estimator unbiased. We propose to use doubly robust estimators for off-policy actor-critic based methods, to show the significance of being at the intersection of model-based and model-free actor-critic algorithms. There are primarily two classes of off-policy value evaluation, where either a model-based approach is taken to fit an MDP model and evaluate the policy based on the learned model as in the *Direct Method (DM)*, or approaches based on *Importance Sampling (IS)* estimators which are completely model-free, unbiased and independent of the state space, but can suffer from uncontrolled variance especially in long horizon tasks. While several existing actor-critic algorithms either use model-based estimates [Janner et al., 2019] or use IS corrections and truncations [Wang et al., 2017], we propose a novel approach towards extending doubly robust estimators, based on a combination of direct model-based approach and model-free IS based estimators in the off-policy actor-critic setting for deep RL. We achieve this by proposing a reward function estimator, to estimate the reward function of the MDP, based on which we can estimate the DR estimator relying on the predicted MDP rewards. Such an approach can be particularly useful in settings where the reward function is often noisy [Romoff et al., 2018] as often found in control and robotics applications. Our key contributions are as follows:

- We extend the doubly robust estimators, previously proposed for contextual bandits [Dudík et al., 2015] and off-policy evaluation (OPE) in RL [Jiang and Li, 2016] to the off-policy actor-critic setting in RL, for low variance critic evaluation in policy gradients
- We present a novel formulation for the model-based part of the reward estimator, based on using a function approximator for estimating the MDP rewards. We then derive the doubly robust (DR) estimator based on the predicted rewards and use it for minimizing the mean squared error in critic evaluation.
- Our proposed doubly robust (DR) estimators for actor-critic algorithms can also be interpreted as a novel formulation as an action-dependent control variate, while keeping the policy gradient estimator unbiased but loweing variance of the gradient estimates
- We find that extending the DR estimator for off-policy actor-critic can be significantly useful for reducing variance of the critic evaluation, since in most off-policy value gradient

based approaches, the estimate of the critic plays a key role in performance. Our proposed extension is evaluated on a wide range of benchmark continuous control tasks, for both growing batch and fixed batch off-policy deep RL settings

- We find that for stochastic or corrupted reward signals, doubly robust estimators can be particularly useful, since existing methods suffer from high variance in presence of noisy reward signals. Doubly robust estimators in actor-critic can significantly reduce the variance in critic evaluation under stochastic rewards, leading to robust and safe algorithms for practical applications. We evaluate our proposed algorithm on noisy reward versions of existing control benchmark tasks, and find that using DR estimation can significantly reduce variance and improve performance.

2 Preliminaries

In policy gradient methods, the aim is to learn a parameterized policy $\pi_\theta(a|s)$ to maximize the discounted sum of cumulative returns along the sampled trajectories, given by $J(\pi_\theta) = \mathbb{E}\pi_\theta[\sum_{i=t+1}^{\infty} \gamma^i r(s_i, a_i)]$. Based on the policy gradient theorem Sutton et al. [2000], we can improve the policy parameters θ using the policy gradient, which can be computed with Monte-Carlo estimation $\nabla_\theta J(\theta) = \mathbb{E}\pi_\theta[\nabla_\theta \log \pi_\theta(a|s) Q^{\pi_\theta}(s, a)]$, where, the $\mathbb{E}$ above uses samples under the current policy π. Often the policy gradient estimator can suffer from high variance, and hence an advantage function $A^{\pi_\theta}(s, a) = Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s)$ or a state dependent baseline, or a combination of both is used in practice to reduce baseline $\nabla_\theta J(\theta) = \mathbb{E}\pi_\theta[\nabla_\theta \log \pi_\theta(a|s)(Q^{\pi_\theta}(s, a) - V^{\pi_\theta}(s))]$. Alternatively, often the gradient is also estimated with an advantage function critic and using a state-action dependent baseline, such that $A^w(s, a) - Q(s, a)$ where the critic uses a function approximator parameterized by w and a separately learned baseline is used.

2.1 Off-Policy Actor-Critic Algorithms

Instead of on-policy gradient estimators, which can be sample-inefficient in practice, due to their inability to re-use data from past experiences, often an off-policy gradient estimate is preferred, based on the deterministic policy gradient (DPG) theorem Silver et al. [2014a]. For continuous control tasks, the DDPG algorithm Lillicrap et al. [2015] is often used due to their ease ability to learn from experience replay buffer. The off-policy policy gradient estimator is given by $\nabla_\theta J(\theta) = \mathbb{E}\mu[\nabla_\theta Q^w(s, \pi_\theta(s)]$, where Q^w is a critic estimate and π_θ is a deterministic policy which outputs continuous actions, to allow directly finding the gradient of the action-value function. Due to the instability of off-policy gradient methods with function approximators, we often require careful fine-tuning of this algorithm Henderson et al. [2018] as the gradient estimate is directly related to the estimate of the critic. Since the DPG Silver et al. [2014a] algorithm can avoid importance sampling (IS) corrections, typically required in off-policy learning Precup [2000], the critic can be evaluated with a regression based objective without any high variance IS corrections being required $L(w) = \mathbb{E}_\mu[(r(s, a) + \gamma Q(s', \pi_\theta(s')) - Q(s, \pi_\theta(s)))^2]$, where the $\mathbb{E}_\mu$ is under samples from the experience replay buffer, and the off-policy critic evaluation is a regular one-step temporal difference (TD) based update without requiring off-policy corrections, even though we are using old samples from the replay buffer. This is due to the DPG theorem Silver et al. [2014b] which avoids an integral over the action space, avoiding the need for IS corrections. However, the DDPG algorithm can often be unstable to use in practice Henderson et al. [2018], and a state-action dependent or state dependent baseline can be used in the gradient estimate.

2.2 Doubly Robust Off-Policy Evaluation

In off-policy evaluation, given a fixed batch or historical data generated by some behaviour or unknown policies $\beta(a, s)$, often the goal is to produce an estimate of the value function $V^\pi(s)$ such that the estimator has low mean squared error (MSE) between the true value function V^{π_e} and the estimated $V^\pi(s)$. Doubly robust estimation is an idea extended from statistics to produce regression estimates, lowering the MSE, in the case of missing or incomplete data. The idea of DR estimators extended from statistics were initially proposed for the contextual bandits setting where often the assumption was that the estimated reward function is given $\hat{R}(s, a)$, to define the DR estimator for contextual bandits $V_{DR} = \hat{V}(s) + \rho\big[r - \hat{R}(s, a)\big]$, where $\hat{R}(s, a)$ is the estimated reward, ρ being the IS corrections for the mismatch in the action distributions. From there, DR estimators were extended

to the off-policy evaluation in RL setting Jiang and Li [2016],Thomas and Brunskill [2016b] to reduce the variance of off-policy evaluation, while keeping the regression based estimators unbiased. Jiang and Li [2016] argued that instead of using importance sampling corrections, which is unbiased, but can have high variance, it is better to use DR estimators in off-policy evaluation tasks. The key step in DR estimator is to use the following unbiased estimator

$$V_{DR}(s) = \hat{V}(s) + \rho[r(s,a) + \gamma V_{DR}(s') - \hat{Q}(s,a)] \tag{1}$$

where we replace V^{π} with V^{π}_{DR} to denote a DR estimation of the off-policy evaluation. A key requirement in DR estimators is to use an approximation to the MDP model since the $\hat{V}$ requires the rewards from an approximation of the MDP. In other words, $\hat{R}$ used to compute $\hat{V}$ is the model's prediction of the reward. Given the samples from past data and an approximate model of the MDP, the goal of DR estimators is to produce a low variance regression mean equated error estimate $MSE(V_{DR}, V^{\pi})$.

3 Related Work

Off policy evaluation in Markov decision processes is the task of evaluating a expected return of one policy with data generated by a different *behavior policy*. In Hanna et al. [2019], the authors propose a regression importance sampling method where the behavior policy is estimated using the same set of data which are used for calculating the importance sampling estimate. Such an estimate of the behavior policy leads to a lower mean squared error for off policy evaluation compared with the true behavior policy. Methods related to regression importance sampling had been studied for Monte Carlo methods in Henmi et al. [2007], Delyon et al. [2016]. Doubly Robust estimators for off policy value evaluation had been studied in Jiang and Li [2016], where the authors used Doubly Robust estimates for Bandits as a control variate for variance reduction. An extension of this work was proposed in Thomas and Brunskill [2016b], which uses doubly robust estimator and proposes a way to mix between model based estimates and importance sampling based estimates to predict the performance of a policy with historical data where the data was generated using a behavior policy. Since Doubly Robust estimators, require a reward predictor, we have adapted the reward estimator method in Romoff et al. [2018]. Other data driven approaches in reward estimation has been discussed in Fu et al. [2018], Hadfield-Menell et al. [2017], Sermanet et al. [2016]. A more extensive view of the Doubly Robust estimator has been proposed by Farajtabar et al. [2018] where the authors present the formulation for learning DR model in RL and the model parameters are learned by minimizing the variance of the DR estimators.

4 Approach

In this work, we extend the existing value-based off-policy policy gradient algorithms such as DDPG [Lillicrap et al., 2015] and Soft Actor-Critic (SAC) [Haarnoja et al., 2018] with a doubly robust (DR) estimation of the critic Q^{DR}_{ϕ}. We propose to use DR based critic estimation in the critic evaluation step, to reduce the variance of the critic estimate in value-based policy gradient algorithms. Since value-based gradient algorithms relies on directly finding the gradient of the action-value function, we hypothesize that reducing the variance of critic estimation can significantly improve the performance and lead to better stability in these algorithms.

4.1 Policy Gradient with Doubly Robust Estimator

In this section, we derive the doubly robust estimator for policy gradient algorithms. The key idea of this estimator comes from observing equation 1 of doubly robust estiimation in OPE which provides an unbiased but low variance estimation of the value function. This depends on the estimated reward function $\hat{R}(s,a)$, where accurate estimation of the MDP rewards can reduce the variance of the value function. In the next section, we will discuss our approach to estimate the MDP rewards $\hat{R}(s,a)$ for a practical algorithm. By using the unbiased DR estimator as a control variate [Thomas and Brunskill, 2016a], for the off-policy actor-critic algorithm, we can achieve a low variance unbiased estimator of

the critic, which helps to improve the stability of existing off-policy policy gradient algorithms. The policy gradient update is given by

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim d_\beta}\Big[\nabla_\theta Q_\phi^{DR}(s, \pi_\theta(s)\Big] \tag{2}$$

where the critic and policies are separately parameterized with ϕ and θ, and we update the policy parameters with stochastic gradient optimization. Considering algorithms such as DDPG, here we denote the policy improvement phase with deterministic policies $\pi_\theta(s)$, which has been extended to stochastic Gaussian policies with a reparameterization trick in algorithms such as Soft Actor-Critic (SAC) [Haarnoja et al., 2018]. The key step in our algorithm is that, we replace the critic as in DDPG with a doubly robust estimation of the critic denoted by $Q_\phi^{DR}(s, \pi_\theta(s))$, such that the critic now mininizes the mean squared regression loss with the following TD error :

$$Q_\phi^{DR}(s,a) = \hat{Q}(s, \pi_\theta(s)) + \Big[r(s,a) + \gamma Q_\phi^{DR}(s', \pi_\theta(s')) - \hat{V}(s)\Big] \tag{3}$$

where $\hat{Q}(s, \pi_\theta(s))$ is following the DR estimate for action-value functions $\hat{Q}$ instead of the value function $\hat{V}$. Equation 3 shows that the critic update for the policy gradient now requires a separate estimation action-value $\hat{Q}$ and value function $\hat{V}$ as well, based on the predicted MDP rewards Jiang and Li [2016], ie, model-based estimation of the reward function, for the doubly robust estimation of the critic. In the next section, we discuss our approach for approximation of the rewards $\hat{R}(s,a)$.

4.2 Reward Function Approximator

Since the doubly robust estimation of the critic requires an approximate MDP model, we estimate the true rewards of the mDP $R(s,a)$ with an approximate reward $\hat{R}(s,a)$ by using a separately parameterized function approximator with parameters ψ_R, denoted by $\hat{R} = f_{\psi_R}(s,a)$. The MDP rewards are estimated based on samples from the replay buffer, and we use a similar approach as in [Romoff et al., 2018] where we train the reward function estimator based on the following regression loss

MSE

$$L(\psi_R) = \mathbb{E}_{s,a,r\sim Buffer}[(\hat{R}(s,a) - R(s,a))^2] \tag{4}$$

We then use this reward for further estimating the approximated action-value function $\hat{Q}(s,a)$ and value function $\hat{V}(s)$ that are required for the DR estimation. Note that, to estimate this, which is a form of the advantage function or a control variate, we typically use a separate function approximator for the control variate estimate. We use another separately parameterized network with parameterss ψ_{QV} which outputs both the approximated $\hat{Q}$ and $\hat{V}$ and trained with the samples from the replay buffer

Temporal Difference

$$\mathcal{L}_{\psi_{\mathcal{QV}}} = \mathbb{E}_{s,a,r,s'\sim Buffer}[(\hat{R}(s,a) + \gamma\hat{Q}(s', \pi_\theta(s')) - \hat{Q}(s,a))^2] \tag{5}$$

where we use the approximated reward $\hat{R}(s,a)$ in the TD error for miniziming the loss $\mathcal{L}(\psi_{QV})$. Based on minimizing the losses in equation 4 and 5, we therefore get an approximation of $\hat{R}$, $\hat{Q}$ and $\hat{V}$ that are required for the doubly robust critic estimation in equation 3.

4.3 Algorithm

Our full algorithm is as given below. Since our entire algorithm requires actor and critic updates, additional estimates of$\hat{R}$, $\hat{Q}$ and $\hat{V}$, we are effectively introducing a three time-scale algorithm. Therefore, in our algorithm, to ensure convergence, we use the large learning rate $\alpha_{\hat{Q}}$ for the approximated $\hat{Q}$, followed by a larger learning rate for the critic estimate compared to the actor as in actor-critic algorithms. In our algorithm, as detailed below, we therefore require the MDP rewards $\hat{R}(s,a)$ to be predicted first, based on which we can train the function approximator with a TD error based on $\hat{R}(s,a)$ to compute $\hat{Q}(s,a)$ and $\hat{V}(s)$. We then use the estimates from the model for off-policy DR evaluation of the critic Q_ϕ^{DR} such as to get unbiased but low variance estimates of the critic. Following this, we can then use *any* off-policy value-based policy gradient algorithm including DDPG, SAC or TD3.

Requirement of Independence : Following [Jiang and Li, 2016], note that, we use a different set of samples from the replay buffer to estimate $\hat{R}(s,a)$, $\hat{Q}(s,a)$ and $\hat{V}(s)$ than the samples used for actor and critic updates. Although this is not a strict requirement that the samples used for estimating $\pi_\theta(a,s)$ and $\hat{R}(s,a)$ be independent

of each other, we find that using separate random samples for the model-based estimates and the model-free updates typically works better in practice.

Algorithm 1: Off-Policy Actor-Critic Algorithms with Doubly Robust Critic Estimator

Require: A policy network $\pi_\theta(a, s)$, critic network $Q_\phi(s, a)$, and networks $\hat{R}_{\psi_R}$ and $\hat{Q}_{\psi_{QV}}(s, a)$
Require: The number of episodes, E and update interval, N.
for $e = 1$ to E **do**
Take action a_t,get reward r_t and observe next state s_{t+1}
Store tuple $(s_t, a_t, r_{t+1}, s_{t+1})$ as trajectory rollouts or in replay buffer $\mathcal{B}$
Estimate the MDP rewards $\hat{R}$, by minimizing loss 4
Estimate approximate $\hat{Q}$ and $\hat{V}$, minimize loss 5
Estimate critic $Q_\phi^{DR}(s, a)$ using $\hat{Q}$ and $\hat{V}$ for DR estimation, by minimizing loss 3
Update policy parameters θ following *any* value-based policy gradient method such as DDPG, SAC, TD3 according to $\nabla_\theta \tilde{J}(\theta) = \left[\nabla_\theta Q_\phi^{DR}(s, \pi_\theta(s))\right]$
end for

4.4 Stochastic and Noisy Rewards

We further consider the case of using stochastic and noisy rewards, where conventional algorithms often fail due to high variance estimates of the gradient. We consider the setting where in addition to the MDP rewards $R(s, a)$, we add a Gaussian noise $\mathcal{N}(\mu, \sigma)$ to the true rewards, to make a corrupted version of the rewards. This is similar to the Corruped Reward MDP (CRMDP) [Everitt et al., 2017], similar to the stochastic reward control experiments considered in [Romoff et al., 2018]. This is similar to many practical robotic tasks where the reward function may often be noisy due to noise in the sensory data. We further examine the significance of using the DR estimator in cases with noisy rewards, given by $\tilde{r}(s, a) = r(s, a) + \mathcal{N}(\mu, \sigma)$ and examine the significance of DR estimator to reduce the variance of noisy critic estimates. For our experiments, we add noisy rewards to all the benchmark control Mujoco tasks, and examine how standard off-policy algorithms such as DDPG and SAC perform in presence of stochastic rewards. All our experiments, as discussed below are for noisy rewards with $\sigma = 0.5$ and $\sigma = 1.0$.

5 Experimental Results

In all our experiments, we compare with existing value-based off-policy gradient algorithms including DDPG, SAC and TD3, and highlight the significance of reducing variance of the critic estimate with DR estimator, in terms of performance improvement. We evaluate the performance of different actor critic algorithms with a doubly robust critic estimator on several continuous control Mujoco tasks Todorov et al. [2012]. Experiments are evaluated on the Half-Cheetah-v1, Walker-2d-v1, Hopper-v1, environments, and evaluated an average over 3 runs with random seeds.

Figure 1 shows that using the DR estimator, we can obtain improvements in the performance over standard baselines. In case of SAC and TD3 algorithms, for Hopper-v1, we obtain an equivalent performance to that of the baselines. For experiments with the HalfCheetah-v1, environemnts, we observe that the baselines for SAC is higher than our DR estimator. These results show the case where the per step rewards are exactly observed by the agent (with out any noise) and that the agent directly use these rewards in the reward estimator. We also observe that for most of the mujoco environments, the variance of the DR estimator is lower compared to the baselines.

Figure 2 and 3 shows the performance comparison using DR estimator where reward prediction is done in the presence of a Gaussian noise added on top of true rewards. Since the reward estimator that we use performs best if the rewards are noisy we observe that the DR estimator thus obtained significantly outperforms in the majority of the mujoco tasks. We also run our DR estimator with different noise levels added to the rewards. In Figure 2 shows the performance plots where the variance of the added Gaussian noise is 0.5 and Figure 3 shows the performance plots where the variance of the added Gaussian noise is 1.0. We also observe that performance with DR estimator achieves lower variance compared to the baselines.

6 Discussion and Conclusion

We proposed an extension of doubly robust estimators from off-policy evaluation (OPE) methods to the off-policy policy gradient and actor-critic algorithms. This, we believe, is an important step towards extending the literature on OPE to the control setting. A large number of previous works on OPE are motivated for low variance but

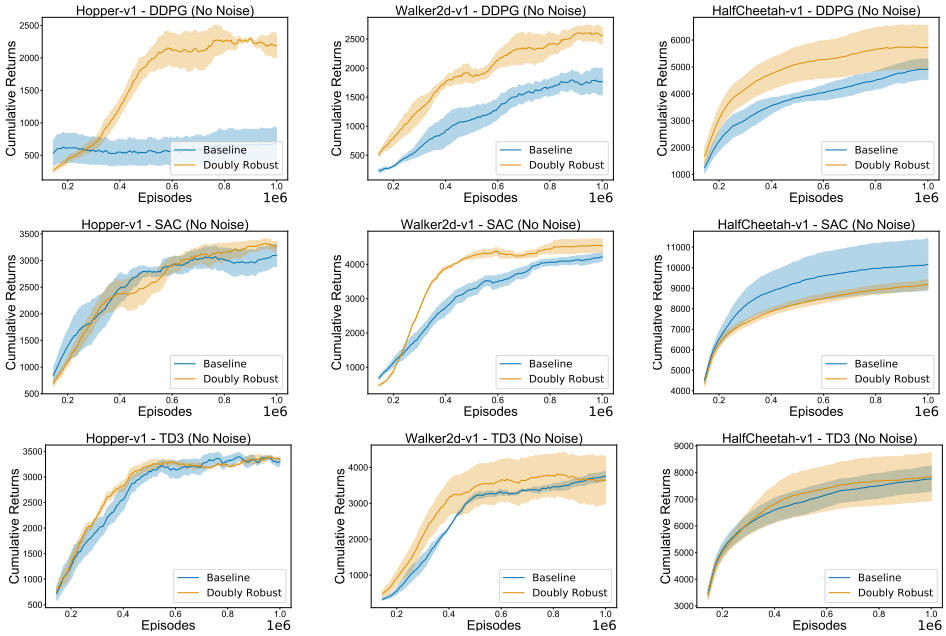

Figure 1: Performance comparison of DDPG,SAC and TD3 with DR estimator using rewards with no noise

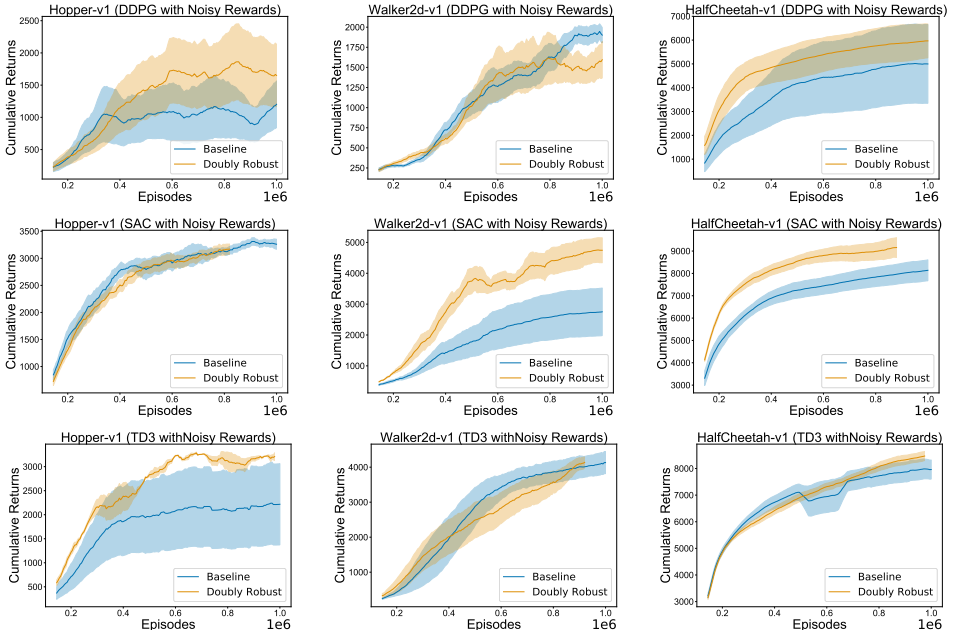

Figure 2: Performance comparison of DDPG, SAC and TD3 with DR estimator using noisy rewards,with $\sigma = 0.5$

unbiased estimators of the value function. We highlight that such OPE estimators can also be extended to the actor-critic setting, where the critic evaluates the policy based on past data.

Our proposed algorithm uses doubly robust estimators, for providing unbiased and low variance critic evaluation. This is particularly useful since majority of popular off-policy methods for control tasks rely on direct value based policy gradient estimates where having useful estimates of the value function plays an important role. We find that since the DR estimator plays the role of control variates to reduce variance of the critic estimate, it has a significant effect in terms of improving performance and lowering variance of existing popular off-policy

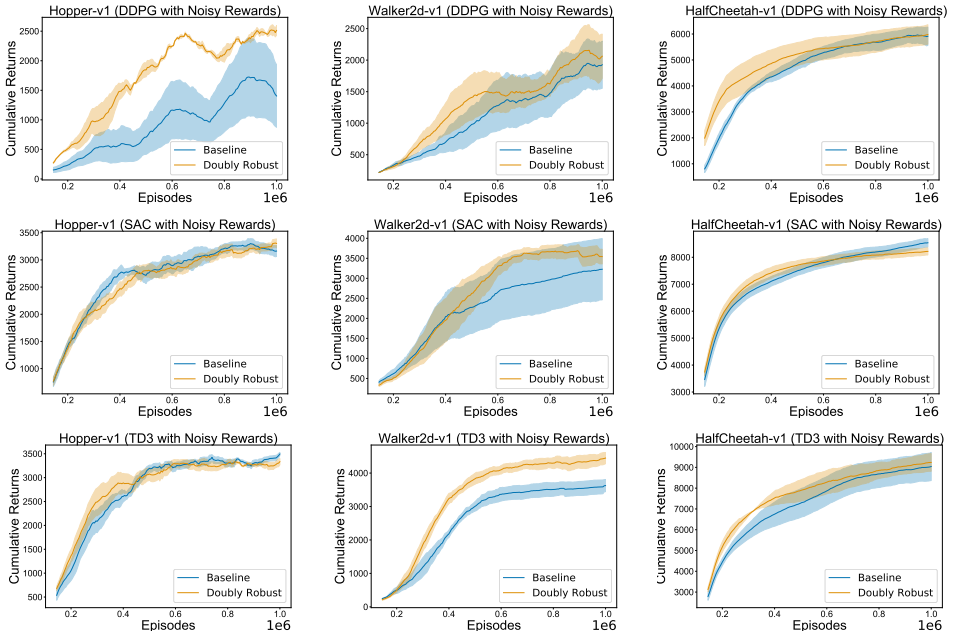

Figure 3: Performance comparison of DDPG, SAC and TD3 with DR estimator using noisy rewards, with $\sigma = 1$

gradient algorithms. We achieve DR estimation for the model-free setting by using a separate reward function estimator to predict the MDP rewards, since DR estimators use a combination of model-fre and model-based approaches. Our approach of estimating the rewards with a separate function approximator plays a further important role in settings where the reward function is stochastic and corrupted. We find that existing policy gradient algorithms can perform poorly in stochastic reward environments, due to high variance in the critic estimates. In such cases, DR estimators can be quite useful for further reducing the variance. Our algorithm plays an important step towards robust and safe RL methods, which is a crucial step for extending current advances of deep RL algorithms for real world applications.

For future work, it would be interesting to see other ways the reward function approximator can be used to predict the rewards $\hat{R}$. Since predicting the rewards can be considered as a supervised learning problem, it would be interesting to see the effect of over-fitting in the reward function approximation. Furthermore, we would require theoretical analysis of the bias variance trade-off of the DR estimator in the actor update, to fully understand the potential of DR estimators in the actor-critic setting, similar to previous studies of DR and variants of DR in terms of bias-variance typically done in OPE methods.

References

John Schulman, Sergey Levine, Pieter Abbeel, Michael I. Jordan, and Philipp Moritz. Trust region policy optimization. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pages 1889–1897, 2015. URL `http://jmlr.org/proceedings/papers/v37/schulman15.html`.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 1856–1865, 2018. URL `http://proceedings.mlr.press/v80/haarnoja18b.html`.

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. *arXiv preprint arXiv:1509.02971*, 2015.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 652–661, 2016. URL http://jmlr.org/proceedings/papers/v48/jiang16.html.

Sham M. Kakade and John Langford. Approximately optimal approximate reinforcement learning. In *Machine Learning, Proceedings of the Nineteenth International Conference (ICML 2002), University of New South Wales, Sydney, Australia, July 8-12, 2002*, pages 267–274, 2002.

Matteo Pirotta, Marcello Restelli, Alessio Pecorino, and Daniele Calandriello. Safe policy iteration. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, pages 307–315, 2013. URL http://proceedings.mlr.press/v28/pirotta13.html.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3207–3214, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16669.

Riashat Islam, Peter Henderson, Maziar Gomrokchi, and Doina Precup. Reproducibility of benchmarked deep reinforcement learning tasks for continuous control. *CoRR*, abs/1708.04133, 2017. URL http://arxiv.org/abs/1708.04133.

Joshua Romoff, Peter Henderson, Alexandre Piché, Vincent François-Lavet, and Joelle Pineau. Reward estimation for variance reduction in deep reinforcement learning. In *2nd Annual Conference on Robot Learning, CoRL 2018, Zürich, Switzerland, 29-31 October 2018, Proceedings*, pages 674–699, 2018. URL http://proceedings.mlr.press/v87/romoff18a.html.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2139–2148, 2016a. URL http://proceedings.mlr.press/v48/thomasa16.html.

Michael Janner, Justin Fu, Marvin Zhang, and Sergey Levine. When to trust your model: Model-based policy optimization. *CoRR*, abs/1906.08253, 2019. URL http://arxiv.org/abs/1906.08253.

Ziyu Wang, Victor Bapst, Nicolas Heess, Volodymyr Mnih, Remi Munos, Koray Kavukcuoglu, and Nando de Freitas. Sample efficient actor-critic with experience replay. *International Conference on Learning Representations*, 2017.

Miroslav Dudík, Dumitru Erhan, John Langford, and Lihong Li. Doubly robust policy evaluation and optimization. *CoRR*, abs/1503.02834, 2015. URL http://arxiv.org/abs/1503.02834.

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*, 2000.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, 2014a.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *International Conference on Machine Learning*, 2014b.

Philip S. Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016*, pages 2139–2148, 2016b. URL http://jmlr.org/proceedings/papers/v48/thomasa16.html.

Josiah Hanna, Scott Niekum, and Peter Stone. Importance sampling policy evaluation with an estimated behavior policy. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pages 2605–2613, 2019. URL http://proceedings.mlr.press/v97/hanna19a.html.

Masayuki Henmi, Ryo Yoshida, and Shinto Eguchi. Importance sampling via the estimated sampler. *Biometrika*, 94(4):985–991, 2007.

Bernard Delyon, François Portier, et al. Integral approximation by kernel smoothing. *Bernoulli*, 22(4):2177–2208, 2016.

Justin Fu, Avi Singh, Dibya Ghosh, Larry Yang, and Sergey Levine. Variational inverse control with events: A general framework for data-driven reward definition. In *Advances in Neural Information Processing Systems*, pages 8538–8547, 2018.

Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart J Russell, and Anca Dragan. Inverse reward design. In *Advances in neural information processing systems*, pages 6765–6774, 2017.

Pierre Sermanet, Kelvin Xu, and Sergey Levine. Unsupervised perceptual rewards for imitation learning. *arXiv preprint arXiv:1612.06699*, 2016.

Mehrdad Farajtabar, Yinlam Chow, and Mohammad Ghavamzadeh. More robust doubly robust off-policy evaluation. *arXiv preprint arXiv:1802.03493*, 2018.

Tom Everitt, Victoria Krakovna, Laurent Orseau, and Shane Legg. Reinforcement learning with a corrupted reward channel. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4705–4713, 2017. doi: 10.24963/ijcai.2017/656. URL `https://doi.org/10.24963/ijcai.2017/656`.

Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 5026–5033. IEEE, 2012.