Efficient Continuous Control with Double Actors and Regularized Critics

Jiafei Lyu1*, Xiaoteng Ma2*, Jiangpeng Yan2, Xiu Li1†

1 Tsinghua Shenzhen International Graduate School, Tsinghua University
2 Department of Automation, Tsinghua University
{ljf20, ma-xt17, yanjp17}@mails.tsinghua.edu.cn, li.xiu@sz.tsinghua.edu.cn

Abstract
How to obtain good value estimation is a critical problem in Reinforcement Learning (RL). Current value estimation methods in continuous control, such as DDPG and TD3, suffer from unnecessary over- or under-estimation. In this paper, we explore the potential of double actors, which has been neglected for a long time, for better value estimation in the continuous setting. First, we interestingly find that double actors improve the exploration ability of the agent. Next, we uncover the bias alleviation property of double actors in handling overestimation with single critic, and underestimation with double critics respectively. Finally, to mitigate the potentially pessimistic value estimate in double critics, we propose to regularize the critics under double actors architecture. Together, we present Double Actors Regularized Critics (DARC) algorithm. Extensive experiments on challenging continuous control benchmarks, MuJoCo and PyBullet, show that DARC significantly outperforms current baselines with higher average return and better sample efficiency.

Introduction
Actor-Critic methods (Prokhorov and Wunsch 1997; Konda and Tsitsiklis 2000) are among the most popular methods in Reinforcement Learning (RL) (Sutton and Barto 2018) which involve value approximation (Baird 1995; Gordon 1995) and policy gradients (Williams 1992; Weng 2018). Built upon actor-critic framework, Deep Deterministic Policy Gradient (DDPG) (Lillicrap et al. 2015) is a typical and widely-used RL algorithm for continuous control. It has been revealed that DDPG results in severe overestimation bias (Fujimoto, Hoof, and Meger 2018) using single critic for function approximation (Thrun and Schwartz 1993) since the actor network is trained to execute action with the highest value estimate.

To tackle the overestimation issue in DDPG, Fujimoto et al. (Fujimoto, Hoof, and Meger 2018) borrow ideas from the Double Q-learning algorithm (Hasselt 2010; Hasselt, Guez, and Silver 2016) and propose Twin Delayed Deep Deterministic Policy Gradient (TD3), which utilizes the minimum value from double critic networks for value estimation. With clipped double Q-learning, TD3 alleviates the overestimation bias problem and significantly improves the performance of DDPG. However, it turns out that TD3 may lead to large underestimation bias (Ciosek et al. 2019), which negatively affects its performance.

While many previous works have focused on enhancing double critics (Pan, Cai, and Huang 2020; Kuznetsov et al. 2020) for better value estimation, the role and advantages of double actors have long been overlooked. We uncover an essential property of double actors, i.e., they enhance the exploration ability of the agent. Double actors offer double paths for policy optimization instead of making the agent confined by a single policy, significantly reducing the agent’s probability of being trapped locally.

We also explore the advantages of double actors for value estimation correction and how they benefit continuous control. We first develop Double Actors DDPG (DADDPG), showing that double actors can remove overestimation bias when built upon a single critic. We experimentally find out that DADDPG significantly outperforms DDPG, which sheds light on the potential of double actors. Similarly, we demonstrate that double actors lessen underestimation bias of double critics method, and develop Double Actors TD3 (DATD3) algorithm. Finally, we propose a soft combination of value estimates over double actors to control the bias flexibly.

To alleviate the potential over pessimistic value estimates of double independent critics, we propose critic regularization, which restricts critics from differing too much. Together, we present our Double Actors Regularized Critics (DARC) algorithm. For illustrating the effectiveness of DARC, a thorough component comparison of relevant algorithms is given in Table 1, reporting the average performance improvement on MuJoCo (Todorov, Erez, and Tassa 2012) environments.

We perform extensive experiments on two challenging continuous control benchmarks, MuJoCo and PyBullet (Ellenberger 2018), where we compare our DARC algorithm against the current common baselines, including TD3 and Soft Actor-Critic (SAC) (Haarnoja et al. 2018a,b). The results show that DARC significantly outperforms them with much higher sample efficiency.

Preliminaries
Reinforcement learning studies sequential decision making problems and it can be formulated by a Markov Decision Process (MDP). The MDP is defined as a 5-tuple
Table 1: Algorithmic component comparison and average performance improvement compared to DDPG baseline. The improvement refers to the averaged relative improvement on the mean final scores of MuJoCo environments with respect to the DDPG baseline over 5 runs.

![Figure 1: Double actors help escape from local optimum.](image)

Enhanced Exploration with Double Actors

Intuitively, double actors allow the agent to evaluate different policies instead of being restricted by single policy path. Single actor \( \pi_{s_0}(s) \) may make the agent stuck in a local optimum \( Q(s, a_{local}) \) rather than the global optimum \( Q(s, a^*) \) due to lack of exploration as demonstrated in Fig 1. Double actors could enhance the exploration ability of the agent by following the policy that results in higher return, i.e., the agent would follow \( \pi_{s_0}(s) \) if \( Q_{\theta}(s, \pi_{s_0}(s)) \geq Q_{\theta}(s, \pi_{s_2}(s)) \) and \( \pi_{s_2}(s) \) otherwise. In this way, double actors help escape from the local optimum \( Q(s, a_{local}) \) and reach the global optimum \( Q(s, a^*) \).

To illustrate the exploration effect of double actors, we design a 1-dimensional, continuous state and action toy environment GoldMiner in Fig 2(a) (see Appendix A.2 for detailed environmental setup). There are two gold mines centering at position \( x_1 = -3, x_2 = 4 \) with neighboring region length to be 1. The miner can receive a reward of +4 and +1 if he digs in the right and left gold mine respectively, and a reward of 0 elsewhere. The miner always starts at position \( x_0 = 0 \) and could move left or right to dig for gold with actions ranging from \([-1, 1, 5]\). The boundaries for the left and right sides are \(-4 \) and 5, and the episode length is 200 steps. We build double actors upon DDPG, which we refer to as DDPG-e (DDPG-exploration). The second actor in DDPG-e is merely used for exploration. We run DDPG and DDPG-e on GoldMiner for 40 independent runs. It can be found that DDPG-e significantly outperforms DDPG as shown in Fig 2(b) where the shaded region denotes one-third a standard deviation for better visibility.

To better understand the effectiveness of double actors, we collect the average high-reward state (where the gold
As shown in Fig 2(c), the visiting frequency of DDPG-e (the critic) parameterized by $\phi'_j$, $i \in \{1, 2\}$ are the parameters of the target networks, which leads to Double Actors DDPG (DADDPG) algorithm (see Appendix A.1 for detailed algorithm and its comparison to TD3). We then build double actors upon double critics networks parameterized by $\theta'_1, \theta'_2$ respectively. One naive way to estimate the value function would be taking minimum of Q-networks for each policy $\pi'_i(s)$, $i = 1, 2$, and employing the maximal one for final value estimation, i.e.,

$$\hat{V}(s') \leftarrow \max_{i=1, 2} \min_{j=1, 2} Q_{\theta'_i}(s', \pi'_{\phi'_i}(s')),$$

where we refer to as Double Actors TD3 (DATD3) algorithm (see Appendix A.3 for more details). Eq. (2) and Eq. (3) provide us with a novel way of estimating value function upon single and double critics. It is worth noting that our method is different from the double Q-learning algorithm as we adopt double actors for value estimation correction instead of constructing double target values for the individual update of actor-critic pairs.

**What benefits can value estimation with double actors bring?** We demonstrate that double actors help mitigate the severe overestimation problem in DDPG and the underestimation bias in TD3. By the definition, we have $T_{\text{DADDPG}}(s') = \min_{i=1, 2} Q_{\theta'_i}(s', \pi'_{\phi'_i}(s'))$, $T_{\text{DATD3}}(s') = \max_{i=1, 2} \min_{j=1, 2} Q_{\theta'_i}(s', \pi'_{\phi'_i}(s'))$. For DADDPG (double-actor-single-critic structure), we show in Theorem 1 that double actors effectively alleviate the overestimation bias problem in DDPG (see the proof in Appendix B.1).

**Theorem 1.** Denote the value estimation bias deviating the true value induced by $T$ as $\text{bias}(T) = \mathbb{E}[T(s')] - \mathbb{E}[Q_{\text{true}}(s') \pi'(s')]$, then we have $\text{bias}(T_{\text{DADDPG}}) \leq \text{bias}(T_{\text{DDPG}})$.

This theorem uncovers the advantages of value estimation correction with double actors as it holds without any special requirement or assumption on double actors. The theorem indicates that DADDPG naturally eases the overestimation bias in DDPG. TD3 also alleviates the overestimation issue while it leverages double critic networks for value correction, which differs from that of DADDGP.

Similarly, we present the relationship of the value estimation of DATD3 (double-actor-double-critic structure) and TD3 in Theorem 2, whose proof is in Appendix B.2.

**Theorem 2.** The bias of DATD3 is larger than that of TD3, i.e., $\text{bias}(T_{\text{DATD3}}) \geq \text{bias}(T_{\text{TD3}})$.

Theorem 1 and Theorem 2 theoretically ensure the bias alleviation property of double actors, i.e., double actors architecture helps mitigate overestimation bias if built upon single critic, and underestimation issues if built upon double critics.

To illustrate the bias alleviation effect with double actors, we conduct experiments in a typical MuJoCo (Todorov, Erez, and Tassa 2012) environment, Walker2d-v2. The value

---

One may wonder why we adopt value estimation using Eq. (3) instead of taking maximal value of two policies first and employing smaller estimation for target value update, i.e., $\hat{V}(s') \leftarrow \min_{i=1, 2} \max_{j=1, 2} Q_{\theta'_i}(s', \pi'_{\phi'_i}(s'))$. Such update scheme would induce larger overestimation bias by taking maximum like DDPG.
Appendix C.1. which reveals the effectiveness and advantages of utilizing double actors in the following (see Appendix D for proof).

![Performance comparison](image1)

Figure 3: Comparison of performance and value estimation bias on Walker2d-v2. Double actors help (a) relieve the overestimation bias in DDPG; (b) mitigate the underestimation bias in TD3.

estimates are calculated by averaging over 1000 states sampled from the replay buffer each timestep. True values are estimated by rolling out the current policy using the sampled states as the initial states and averaging the discounted long-term rewards. The experimental setting is identical as in Section Experiments and the result is presented in Fig 3. Fig 3(a) shows that DADDPG reduces the overestimation bias in DDPG and significantly outperforms DDPG in both sample efficiency and final performance. As shown in Fig 3(b), DATD3 outperforms TD3 and preserves larger bias than TD3, which reveals the effectiveness and advantages of utilizing double actors to correct value estimation. The estimation bias comparison on broader environments can be found in Appendix C.1.

**Soft target value.** To control the underestimation bias more flexible, we propose to use a convex combination of value estimate over two actors, which is given by:

\[
\hat{V}(s'; \nu) = (1 - \nu) \min_{i=1,2} \min_{j=1,2} Q_{\theta_j}^{\pi_i}(s', \pi_{\phi_i}(s')) + \nu \min_{i=1,2} \min_{j=1,2} Q_{\theta_j}^{\pi_i}(s', \pi_{\phi_i}(s')) ,
\]

where \( \nu \in \mathbb{R} \) and \( \nu \in [0, 1) \). Eq. (3) is a special case of Eq. (4). \( \hat{V}(s'; \nu) \) would lean towards the maximal value of two evaluation paths with \( \nu \to 0 \) and vice versa if \( \nu \to 1 \). If we set \( \nu = 0 \), then slight overestimation may be introduced as shown in Fig 3(b).

We further give estimation error analysis induced by double actors in the following (see Appendix D for proof).

**Theorem 3** (Upper Error Bound). Assume that

\[
\begin{align*}
\| \min_{i=1,2} \min_{j=1,2} Q_{\theta_j}^{\pi_i}(s, \pi_{\phi_i}(s)) - \max_{a \in \mathcal{A}} Q(s, a) \|_\infty & \leq \Delta_1 ,
\| \max_{i=1,2} \min_{j=1,2} Q_{\theta_j}^{\pi_i}(s, \pi_{\phi_i}(s)) - \max_{a \in \mathcal{A}} Q(s, a) \|_\infty & \leq \Delta_2 ,
\end{align*}
\]

Then for any iteration \( t \), the difference between the optimal value function \( V^*(s) \) and the value function \( V(s) \) induced by the double actors satisfies:

\[
\| V_t(s) - V^*(s) \|_\infty \leq \gamma^t \| V_0(s) - V^*(s) \|_\infty + \frac{\nu_t \Delta_1}{1 - \gamma} + \frac{(1 - \nu_t) \Delta_2}{1 - \gamma}.
\]

This theorem guarantees the rationality of employing the value estimation induced by the double actors. \( \Delta_1 \) and \( \Delta_2 \) measure the deviance of minimal and maximal estimates over double actors against optimal value functions respectively. The upper bound would be \( O(\frac{1}{\gamma^t}) \) if \( \Delta_1 \) and \( \Delta_2 \) can be controlled in a valid scale. The hyperparameter \( \nu \) is important in compromising between overestimation and underestimation. If one uses large \( \nu \), then the upper bound would be dominant by severe underestimation bias, i.e., \( \Delta_1 > \Delta_2 \), which is harmful to the performance of the agent.

**Beyond Double Critics: Double Actors with Regularized Critics**

In this section, we first illustrate that the independence in double critics leads to pessimistic underestimation and value estimate uncertainty. We then propose to regularize critics for reduced uncertainty in value estimate. Furthermore, we build double actors upon regularized critics and present Double Actors Regularized Critics (DARC) algorithm.

**Pessimistic Estimation in Double Critics**

Despite the success in addressing the severe overestimation bias in single critic, the double critics in TD3 introduce pessimism. Though underestimation is much better than that of overestimation, TD3 is still not satisfying. We dig deeply into the root of underestimation bias in TD3. Given \( s, a \), we assume that \( Q_{\theta_i}(s, a) = Q^{true}_i(s, a) + U_i(a) \) with independent noise \( U_i(a) \) such that \( \forall a \in \mathcal{A}, \mathbb{E}[U_i(a)] = 0, \forall i = 1, 2 \). Then for TD3, we have

\[
\mathbb{E}_U \left[ \min_{i=1,2} Q_{\theta_i}(s, a) \right] = \mathbb{E}_U \left[ \min_{i=1,2} (Q^{true}_i(s, a) + U_i(a)) \right] 
\leq \mathbb{E}_U \left[ Q^{true}_i(s, a) \right] = Q^{true}_i(s, a).
\]

We note that the independence in critics is responsible for the underestimation bias of TD3. Eq. (5) illustrates that double critics would intrinsically induce underestimation because of the negative bias from the minimum operation. Furthermore, there exists some uncertainty when double critics estimate value functions at the same state. As shown in Fig 4 (light blue line and light green line), double critics \( Q_{\theta_i}(s, a), Q_{\theta_j}(s, a) \) may have large disagreement in value estimation at the same state. If we always take minimum over double critics, the resulting value estimate would deviate from the true value (the grey line) largely.

**Critic regularization.** We then propose to constrain the value estimates of the critics to mitigate the pessimistic value estimation, which leads to solving the optimization problem:

\[
\begin{align*}
& \min_{\theta_i} \mathbb{E}_{s, a \sim \rho} (Q_{\theta_i}(s, a) - y)^2 , \\
& \text{s.t.} \quad \mathbb{E}_{s, a \sim \rho} |Q_{\theta_i}(s, a) - Q_{\theta_j}(s, a)| \leq \delta ,
\end{align*}
\]

7658
Figure 4: Illustration of pessimistic value estimation from double critics and how Regularized Critics (RC) help mitigate the phenomenon. Red dots represent transition samples. The yellow arrows show the effect of RC.

where \( i \in \{1, 2\} \) and \( y = r + \gamma \hat{V}(s'; \nu) \) is the target value. The regularization term pushes the double critics to be close to each other while simultaneously approximate the target value. Specifically, in Fig 4,

\[
Q_{\theta_1}(s, a) \rightarrow Q_{\theta_1}^\text{reg}(s, a), \quad Q_{\theta_2}(s, a) \rightarrow Q_{\theta_2}^\text{reg}(s, a).
\]

With Regularized Critics (RC), the approximation error \( U_i(a) \) has less possibility of inducing a large negative bias. The pessimistic underestimation phenomenon can therefore be mitigated as also demonstrated in Fig 4 (the dark blue and green lines deviate true value \( Q^\text{true} \) less than the light ones).

In our implementation, we resort to penalty function methods by regularizing the original objective with deviation of critics to avoid the complex and expensive nonlinear programming costs. The critics are therefore trained by minimizing:

\[
\frac{1}{N} \sum_{s} \left\{ (Q_{\theta_i}(s, a) - y)^2 + \lambda \left[ Q_{\theta_1}(s, a) - Q_{\theta_2}(s, a) \right] \right\} \tag{7}
\]

where \( i \in \{1, 2\} \). The policy parameter \( \phi_i, i \in \{1, 2\} \) can then be optimized via policy gradient:

\[
\frac{1}{N} \sum_{s} \nabla_a Q_{\theta_i}(s, a)_{|a=\pi_{\phi_i}(s)} \nabla_{\phi_i} \pi_{\phi_i}(s) \tag{8}
\]

**Full Algorithm**

In summary, we present Double Actors Regularized Critics (DARC) algorithm in Algorithm 1, which has three key components: (1) action that incurs higher return is executed to enjoy better exploration capability; (2) double actors are used for value correction to balance under- and overestimation bias, e.g. via Eq. (4); (3) critics are regularized for reduced uncertainty in value estimate. We also adopt a cross update scheme (see graphical illustration in Appendix A.5) where only one actor-critic pair is updated each timestep and meanwhile the other pair is only used for value correction. Such scheme naturally leads to the delayed update of the target network and contributes to policy smoothing.

The estimation bias of DARC is less conservative than that of DATD3. We present detailed bias comparison of DARC with DATD3 and TD3 in Appendix C.2. DARC is more optimistic and efficient with soft value estimate and critic regularization.

**Algorithm 1: Double Actors Regularized Critics (DARC)**

1. Initialize critic networks \( Q_{\theta_1}, Q_{\theta_2} \) and actor networks \( \pi_{\phi_1}, \pi_{\phi_2} \) with random parameters
2. Initialize target networks \( \theta_1' \leftarrow \theta_1, \theta_2' \leftarrow \theta_2, \phi_1' \leftarrow \phi_1, \phi_2' \leftarrow \phi_2 \) and replay buffer \( B = \{ \} \)
3. for \( t = 1 \) to \( T \) do
4. Select action \( a \) with \( \max_i \max_j Q_{\theta_i}(s, \pi_{\phi_i}(s)) \) added \( \epsilon \sim \mathcal{N}(0, \sigma) \)
5. Execute action \( a \) and observe reward \( r \), new state \( s' \) and done flag \( d \)
6. Store transitions in the replay buffer, i.e., \( B \leftarrow B \cup \{(s, a, r, s', d)\} \)
7. for \( i = 1, 2 \) do
8. Sample \( N \) transitions \( \{(s_j, a_j, r_j, s_j', d_j)\}_{j=1}^{N} \sim B \)
9. Get actions: \( a' \leftarrow \pi_{\phi'_i}(s') + \epsilon, a'' \leftarrow \pi_{\phi'_2}(s') + \epsilon, \epsilon \sim \text{clip}(\mathcal{N}(0, \sigma), -c, c) \)
10. Calculate \( \hat{V}(s') \) with Eq. (4)
11. \( y_t \leftarrow r + \gamma(1 - d) \hat{V}(s') \)
12. Update critic \( \theta_i \) with Eq. (7)
13. Update actor \( \phi_i \) with Eq. (8)
14. Update target networks: \( \theta_i' \leftarrow \tau \theta_i + (1 - \tau) \theta_i', \phi_i' \leftarrow \tau \phi_i + (1 - \tau) \phi_i' \)
15. end for
16. end for

DARC benefits from double actors while it is different from A3C (Mnih et al. 2016) because: (1) A3C adopts multiple processes while DARC only adopts one; (2) There is no synchronization of actors in DARC; (3) The actors in A3C are merely used for exploration while DARC also uses them for value correction.

**Experiments**

In this section, we first conduct a detailed ablation study on DARC to investigate what contributes most to the performance improvement. We then extensively evaluate our method on two continuous control benchmarks, where we compare with common baseline methods including TD3 (Fujimoto, Hoof, and Meger 2018) and SAC (Haarnoja et al. 2018a). Moreover, we extensively compare DARC with other value estimation correction methods to further illustrate the effectiveness of DARC.

We adopt two widely-used continuous control benchmarks, OpenAI Gym (Brockman et al. 2016) simulated by MuJoCo (Todorov, Erez, and Tassa 2012) and Box2d (Catto 2011), and PyBullet Gym simulated by PyBullet (Ellenberger 2018). We compare our method against DDPG, TD3, and SAC. We use the fine-tuned version of DDPG proposed in TD3 and temperature auto-tuned SAC (Haarnoja et al. 2018b). The baselines are conducted by open-sourced implementations (Fujimoto 2018; Tianhong 2019). Each algorithm is repeated with 5 independent seeds and evaluated for 10 times every 5000 timesteps. DARC shares the identical network configuration with TD3. The regularization coefficient is set to be 0.005 by default and the value estimation weight \( \nu \) is mainly selected from \([0, 0.5]\) with 0.05 as interval by using grid search.
We conduct the ablation study and parameter study on one typical MuJoCo environment, HalfCheetah-v2, which is adequate to show the influence of different parts and parameters.

**Components.** We show in Fig 5(a) that double actors are vital for DARC, where DARC with a single actor significantly underperforms DARC. We further exclude the exploration effect by executing actions following the first actor in DARC (see Fig 5(a)), which results in a bad performance. Moreover, a decrease in the performance on HalfCheetah-v2 occurs without either double actors or regularization on critics as demonstrated in Fig 5(a) and Fig 5(b). We find that value correction with double actors contributes most to the performance improvement upon TD3, e.g., DARC without regularization outperforms DARC with a single actor. Critic regularization can only be powerful if the value estimation is good enough.

**The regularization parameter** $\lambda$. $\lambda$ balances the influence of the difference in two critic networks. Large $\lambda$ may cause instability in value estimation and impede the agent from learning a good policy. While small $\lambda$ may induce a slow and conservative update, which weakens the effect and benefits of critic regularization. Luckily, there does exist an intermediate value ($\lambda = 0.005$) that could achieve a trade-off as shown in Fig 5(c). Since DARC is not sensitive to $\lambda$ as long as it is not too large, we set $\lambda = 0.005$ as default in the rest experiments.

**Comparison with Other Value Correction Methods**

We additionally compare DARC with other recent value correction methods, SD3 (Pan, Cai, and Huang 2020) and TADD (Wu et al. 2020), where SD3 leverages softmax operator on value function for a softer estimation and TADD leverages triple critics by weighting over them for better estimation. We also compare DARC with Double Actors TD3 (DATD3). We conduct numerous experiments in identical environments in the above section and report the final mean score over 5 independent runs in Table 2, showing that DARC significantly outperforms these value correction methods in all tasks.

**Related Work**

Actors-Critic methods (Konda and Tsitsiklis 2000; Prokhorov and Wunsch 1997; Konda and Borkar 1999) are widely used in Reinforcement Learning (RL). The quality of the learned critic network is vital for a good performance in an RL agent when applying function approximation (Barth-Maron et al. 2018b), e.g., we can get an unbiased estimate of policy gradient if we enforce the critic to meet the compatibility conditions (Silver et al. 2014).

How to estimate the value function in a good way remains an ongoing problem in RL, and has been widely investigated in deep Q-network (DQN) (Hasselt, Guez, and Silver 2016; Sabry and Khalifa 2019) in discrete regime control. Lan et al. (Lan et al. 2020) propose to take the minimum Q-value under the ensemble scheme to control the estimation bias in DQN, while Anschel et al. (Anschel, Baram, and Shimkin 2017) leverage the average value of an ensemble of Q-networks for variance reduction. Apart from Q-ensemble methods, many effective methods involving estimation weighting (Zhang, Pan, and Kochenderfer 2017), softmax operator (Song, Parr, and Carin 2019) are also explored.

In continuous control domain, DDPG suffers from large overestimation bias. The improvement upon DDPG includes distributional (Barth-Maron et al. 2018a; Bellemare, Dabney, and Munos 2017), model-based (Feinberg et al. 2018), prioritized experience replay (Horgan et al. 2018) method, etc. TD3 tackles the issue by using double critics for value correction, while it may suffer from severe underestimation.
Table 2: Numerical performance comparison on final score (3M steps for Humanoid and 1M steps for the rest) between DARC and other value estimation correction methods. The best results are in bold.

<table>
<thead>
<tr>
<th>Environment</th>
<th>TD3</th>
<th>TADD</th>
<th>SD3</th>
<th>DATD3 (ours)</th>
<th>DARC (ours)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ant</td>
<td>4164.10</td>
<td>4593.01</td>
<td>4541.71</td>
<td>5180.29</td>
<td><strong>5642.33±188.82</strong></td>
</tr>
<tr>
<td>BipedalWalker</td>
<td>294.08</td>
<td>303.42</td>
<td>299.69</td>
<td>305.09</td>
<td><strong>311.25±2.66</strong></td>
</tr>
<tr>
<td>HalfCheetah</td>
<td>10237.62</td>
<td>10099.80</td>
<td>10934.72</td>
<td>10623.96</td>
<td><strong>11600.74±499.11</strong></td>
</tr>
<tr>
<td>Hopper</td>
<td>3145.20</td>
<td>3142.16</td>
<td>3286.24</td>
<td>2822.94</td>
<td><strong>3577.93±133.97</strong></td>
</tr>
<tr>
<td>Humanoid</td>
<td>5992.28</td>
<td>6182.54</td>
<td>5809.18</td>
<td>5960.03</td>
<td><strong>6737.63±743.95</strong></td>
</tr>
<tr>
<td>Walker2d</td>
<td>4605.25</td>
<td>4834.59</td>
<td>4622.89</td>
<td>4694.75</td>
<td><strong>5159.84±687.84</strong></td>
</tr>
<tr>
<td>AntPyBullet</td>
<td>3683.49</td>
<td>3216.75</td>
<td>3762.93</td>
<td>3949.02</td>
<td><strong>4100.01±19.24</strong></td>
</tr>
<tr>
<td>Walker2dPybullet</td>
<td>1385.01</td>
<td>1150.07</td>
<td>1497.06</td>
<td>1777.24</td>
<td><strong>1902.46±217.25</strong></td>
</tr>
</tbody>
</table>

Figure 6: Performance comparison in OpenAI Gym and PyBullet Gym environments.

problem. There are many efforts in utilizing TD3 for distributional training (Ma et al. 2020), Hierarchical RL (HRL) (Nachum et al. 2018), and so on, while value estimation correction methods for performance improvement are rarely investigated. Wu et al. (Wu et al. 2020) adopt triple critics and correct the value estimation by weighting over these critics. There are also some prior works (Kuznetsov et al. 2020; Roy, Bakshi, and Maharaj 2020) that adopt critic ensemble for bias alleviation. Also, two parallel actor-critic architecture are explored to learn better options (Zhang and Whiteson 2019). Despite these advances, few of them investigate the role and benefits of double actors in value correction, which is the focus of our work. Also, training multiple critic or actor networks can be expensive, while DARC is efficient.

Finally, our method is related to the regularization method, which has been broadly used outside RL, for instance, machine learning (Bauer, Pereverzey, and Rosasco 2007), computer vision (Giroi, Jones, and Poggio 1995; Wan et al. 2013; Xu et al. 2021), etc. Inside RL, regularization strategy is widely used in offline RL (Lange, Gabel, and Riedmiller 2012; Wu, Tucker, and Nachum 2019), model-based RL (Boney, Kannala, and Ilan 2020; D’Oro and Jaskowski 2020), and maximum entropy RL (Haarnoja et al. 2018a; Zhao, Sun, and Tresp 2019). We, however, propose to regularize critics to ensure that the value estimation from them would not deviate far from each other, which reduces value estimate uncertainty.

Conclusion

In this paper, we explore and illustrate the benefits of double actors in continuous control tasks, which has long been ignored. We show the preeminent exploration property and the bias alleviation property of double actors on both single critic and double critics. We further propose to regularize critics to mitigate large difference in value estimation from two independent critics. Putting together, we present Double Actors Regularized Critics (DARC) algorithm which extensively and significantly outperforms baseline methods as well as other value estimation correction methods on standard continuous control benchmarks.

For future work, it will be interesting to extend DARC from double-actor-double-critic architecture into multi-actor-multi-critic structure. Critic ensemble can be utilized to measure the uncertainty of value estimate. Meanwhile, multiple actors show strength in better exploration and have more advantages to tackle multimodal distribution.
Acknowledgements

This research was partly supported by the National Natural Science Foundation of China (Grant No. 41876098), the National Key Research and Development Program of China (Grant No. 2020AAA0108303), and Shenzhen Science and Technology Project (Grant No. JCYJ20200109143041798).

References


