Efficient Continuous Control with Double Actors and Regularized Critics

Jiafei Lyu[1*], **Xiaoteng Ma**[2*], **Jiangpeng Yan**[2], **Xiu Li**[1†]

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University
[2] Department of Automation, Tsinghua University
{lvjf20, 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 (Brockman et al. 2016) 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

*Equal contribution, † Corresponding author

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.

Algorithms	Double Actors	Double Critics	Value Correction	Regularization	Improvement
DDPG(Lillicrap et al. 2015)	✗	✗	✗	✗	100%
TD3(Fujimoto, Hoof, and Meger 2018)	✗	✓	✓	✗	245%
SAC(Haarnoja et al. 2018a)	✗	✓	✓	✓	191%
DADDPG (this work)	✓	✗	✓	✗	167%
DATD3 (this work)	✓	✓	✓	✗	291%
DARC (this work)	✓	✓	✓	✓	**331%**

$\langle \mathcal{S}, \mathcal{A}, p, r, \gamma \rangle$ where $\mathcal{S}, \mathcal{A}$ denote state space and action space respectively, p denotes transition probability, $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function and $\gamma \in [0, 1)$ is the discount factor. The agent behaves according to a deterministic policy $\pi_\phi : \mathcal{S} \mapsto \mathcal{A}$ parameterized by ϕ. The objective function of reinforcement learning can be written as $J(\phi) = \mathbb{E}_s[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)|s_0, a_0; \pi_\phi(s)]$, which aims at maximizing expected future discounted rewards following policy $\pi_\phi(s)$. We consider continuous control scenario with bounded action space and we further assume a continuous and bounded reward function r.

The policy π_ϕ can be improved by conducting policy gradient ascending in terms of the objective function $J(\phi)$. The Deterministic Policy Gradient (DPG) Theorem (Silver et al. 2014) offers a practical way of calculating the gradient:

$$\nabla_\phi J(\phi) = \mathbb{E}_s[\nabla_\phi \pi_\phi(s) \nabla_a Q_\theta(s, a)|_{a=\pi_\phi(s)}], \qquad (1)$$

where $Q_\theta(s, a)$ is the Q-function with parameter θ that approximates the long-term rewards given state and action. In actor-critic architecture, the critic estimates value function $Q_\theta(s, a)$ to approximate the true parameter θ^{true}, and the actor is updated using Eq. (1). DDPG learns a deterministic policy $\pi_\phi(s)$ to approximate the optimal policy as it is expensive to directly apply the max operator over the continuous action space $\mathcal{A}$. With TD-learning, the critic in DDPG is updated via $\tilde{\theta} \leftarrow \theta + \eta \mathbb{E}_{s,a\sim\rho}[r + \gamma Q_{\theta'}(s', \pi_{\phi'}(s')) - Q_\theta(s, a)] \nabla_\theta Q_\theta(s, a)$, where η is the learning rate, ρ is the sample distribution in the replay buffer, ϕ' and θ' are the parameters of the target actor and critic network respectively. TD3 addresses the overestimation problem in DDPG by employing double critics for value estimation, which is given by $\hat{Q}(s', a') \leftarrow \min_{i=1,2} Q_{\theta_i'}(s', a')$, and one actor for policy improvement. We denote $\mathcal{T}(s')$ as the value estimation function that is utilized to estimate the target value $r + \gamma \mathcal{T}(s')$, and then we have $\mathcal{T}_{\text{DDPG}}(s') = Q_{\theta'}(s', \pi_{\phi'}(s'))$ and $\mathcal{T}_{\text{TD3}}(s') = \min_{i=1,2} Q_{\theta_i'}(s', \pi_{\phi'}(s'))$.

Double Actors for Better Continuous Control

In this section, we demonstrate how double actors work. First, we illustrate that double actors induce a stronger exploration capability in the continuous setting. Then we discuss how to better estimate value function with double actors and also show that double actors can help ease overestimation in DDPG, and underestimation bias in TD3.

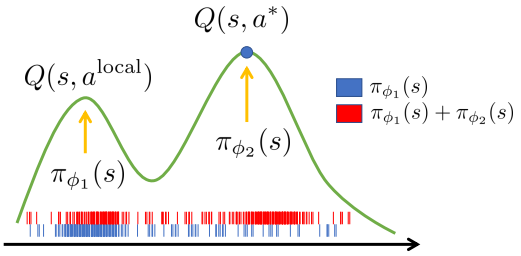


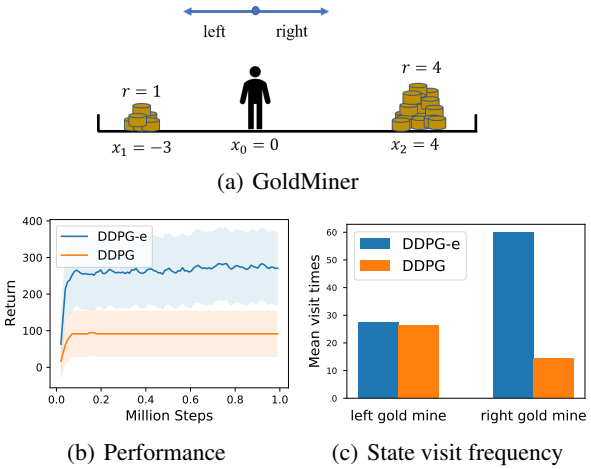
Figure 1: Double actors help escape from local optimum.

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 $\pi_{\phi_1}(s)$ may make the agent stuck in a local optimum $Q(s, a^{\text{local}})$ 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 $\pi_{\phi_1}(s)$ if $Q_\theta(s, \pi_{\phi_1}(s)) \geq Q_\theta(s, \pi_{\phi_2}(s))$ and $\pi_{\phi_2}(s)$ otherwise. In this way, double actors help escape from the local optimum $Q(s, a^{\text{local}})$ 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 $[-1.5, 1.5]$. 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



(a) GoldMiner

(b) Performance

(c) State visit frequency

Figure 2: Exploration ability analysis of double actors on GoldMiner environment.

mines lie) visiting times of each method for every episode. As shown in Fig 2(c), the visiting frequency of DDPG-e to the right gold mine significantly exceed that of DDPG, indicating that the agent would tend to visit the places that could achieve higher rewards with double actors. Single actor, however, would guide the agent to visit the left mine more frequently as it is closer, i.e., it is stuck in a local optimum. DDPG could hardly learn a policy towards the right gold mine, which is shortsighted and pessimistic (see Fig 2).

Moreover, double actors help relieve the *pessimistic underexploration* phenomenon reported in (Ciosek et al. 2019) upon double critics, which is caused by the reliance on pessimistic critics for exploration. This issue can be mitigated naturally with the aid of double actors as only the policy that leads to higher expected return would be executed, which is beneficial for exploration (see Appendix A.2.1 for more details). With double actors, the exploration capability of the agent is decoupled from value estimation and *the agent could benefit from both pessimistic estimation and optimistic exploration*. To conclude, double actors enable the agent to visit more valuable states and enhance the exploration capability of the agent, which makes the application of double actors in continuous control setting appealing.

Better Value Estimation with Double Actors

In actor-critic-style algorithms, a good value estimation lays a foundation for guiding policy learning. In this part, we demonstrate how to get better value estimate with the aid of double actors.

How to use double actors for value estimation correction? We first build double actors upon single Q-network (the critic) parameterized by θ. For each training step, the critic has two paths to choose from: either following $\pi_{\phi'_1}(s)$ or $\pi_{\phi'_2}(s)$, where both paths are estimated to approximate the optimal path for the task. Inspired by double Q-learning, we propose to estimate the value function via

$$\hat{V}(s') \leftarrow \min_{i=1,2} Q_{\theta'}(s', \pi_{\phi'_i}(s')), \tag{2}$$

where $\theta', \phi'_i, i \in \{1, 2\}$ 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 θ'_1, θ'_2 respectively. One naive way to estimate the value function would be taking minimum of Q-networks for each policy $\pi_{\phi'_i}(s), i = 1, 2$, and employing the maximal one for final value estimation, i.e.,

$$\hat{V}(s') \leftarrow \max_{i=1,2} \min_{j=1,2} Q_{\theta'_j}(s', \pi_{\phi'_i}(s')), \tag{3}$$

which we refer to as *Double Actors TD3* (DATD3) algorithm (see Appendix A.3 for more details)[1]. 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 $\mathcal{T}_{\text{DADDPG}}(s') = \min_{i=1,2} Q_{\theta'}(s', \pi_{\phi'_i}(s'))$, $\mathcal{T}_{\text{DATD3}}(s') = \max_{i=1,2} \min_{j=1,2} Q_{\theta'_j}(s', \pi_{\phi'_i}(s'))$. 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 $\mathcal{T}$ as bias($\mathcal{T}$) = $\mathbb{E}[\mathcal{T}(s')] - \mathbb{E}[Q_{\theta^{\text{true}}}(s', \pi_{\phi'}(s'))]$, then we have* $\text{bias}(\mathcal{T}_{\text{DADDPG}}) \leq \text{bias}(\mathcal{T}_{\text{DDPG}})$.

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 DADDPG.

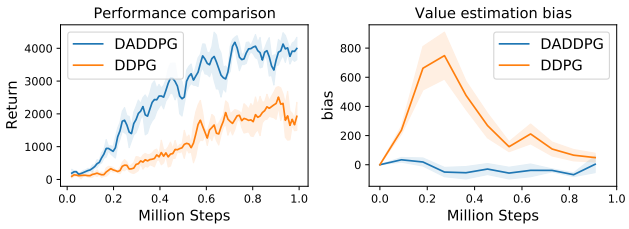
Similarly, we present the relationship of the value estimation of DATD3 (double-actor-double-critic structure) and TD3 in Theorem 2, whose proof is in Appendix B.2.

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

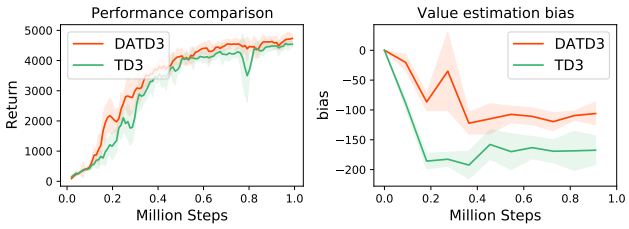
Theorem 1 and Theorem 2 theoretically ensure the bias alleviation property of double actors, i.e., double actors architecture helps mitigate overestimation bias if built upon single critic, and underestimation issues if built upon double critics.

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

[1]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., $\hat{V}(s') \leftarrow \min_{i=1,2} \max_{j=1,2} Q_{\theta'_j}(s', \pi_{\phi'_i}(s'))$. Such update scheme would induce large overestimation bias by taking maximum like DDPG.



(a) Single critic network



(b) Double critic networks

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 tiemstep. 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:

$$\hat{V}(s';\nu) = (1-\nu)\max_{i=1,2}\min_{j=1,2} Q_{\theta'_j}(s', \pi_{\phi'_i}(s')) + \nu \min_{i=1,2}\min_{j=1,2} Q_{\theta'_j}(s', \pi_{\phi'_i}(s')), \tag{4}$$

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*

$$\|\min_{i=1,2}\min_{j=1,2} Q_{\theta'_j}(s, \pi_{\phi'_i}(s)) - \max_{a\in\mathcal{A}} Q(s,a)\|_\infty \le \Delta_1,$$
$$\|\max_{i=1,2}\min_{j=1,2} Q_{\theta'_j}(s, \pi_{\phi'_i}(s)) - \max_{a\in\mathcal{A}} Q(s,a)\|_\infty \le \Delta_2.$$

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:

$$\|V_t(s) - V^*(s)\|_\infty \le \gamma^t \|V_0(s) - V^*(s)\|_\infty + \frac{\nu\Delta_1}{1-\gamma} + \frac{(1-\nu)\Delta_2}{1-\gamma}.$$

This theorem guarantees the rationality of employing the value estimation induced by the double actors. Δ_1 and Δ_2 measure the deviance of minimal and maximal estimates over double actors against optimal value functions respectively. The upper bound would be $O(\frac{1}{1-\gamma})$ if Δ_1 and Δ_2 can be controlled in a valid scale. The hyperparameter ν is important in compromising between overestimation and underestimation. If one uses large ν, 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, we assume that $Q_{\theta_i}(s,a) = Q^{\text{true}}(s,a) + U_i(a)$ with independent noise $U_i(a)$ such that $\forall a \in \mathcal{A}, \mathbb{E}_U[U_i(a)] = 0, \forall i = 1, 2$. Then for TD3, we have

$$\begin{aligned}\mathbb{E}_U\left[\min_{i=1,2} Q_{\theta_i}(s,a)\right] &= \mathbb{E}_U\left[\min_{i=1,2}(Q^{\text{true}}(s,a) + U_i(a))\right] \\ &\le \mathbb{E}_U\left[Q^{\text{true}}(s,a)\right] = Q^{\text{true}}(s,a)\end{aligned} \tag{5}$$

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 $Q_{\theta_1}(s,a), Q_{\theta_2}(s,a)$ 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:

$$\begin{aligned}&\min_{\theta_i} \mathbb{E}_{s,a\sim\rho}[(Q_{\theta_i}(s,a) - y)^2], \\ &\text{s.t.} \quad \mathbb{E}_{s,a\sim\rho}\left|Q_{\theta_1}(s,a) - Q_{\theta_2}(s,a)\right| \le \delta,\end{aligned} \tag{6}$$

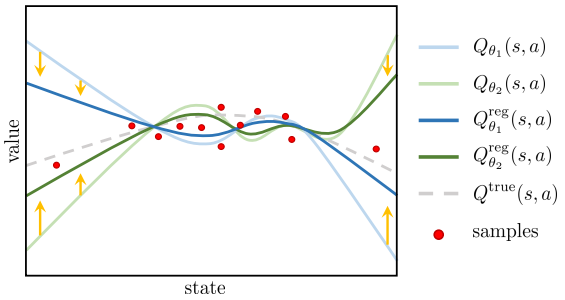


Figure 4: Illustration of pessimistic value estimation from double critics and how Regularized Critics (RC) help mitigate the phenomenon. Red dots represent transition samples. The yellow arrows show the effect of RC.

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,

$$Q_{\theta_1}(s,a) \to Q^{\text{reg}}_{\theta_1}(s,a), Q_{\theta_2}(s,a) \to Q^{\text{reg}}_{\theta_2}(s,a).$$

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^{true} less than the light ones).

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

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

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)

Initialize critic networks $Q_{\theta_1}, Q_{\theta_2}$ and actor networks $\pi_{\phi_1}, \pi_{\phi_2}$ with random parameters
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 $\mathcal{B} = \{\}$
for t = 1 to T **do**
  Select action a with $\max_i \max_j Q_{\theta_i}(s, \pi_{\phi_j}(s))$ added $\epsilon \sim \mathcal{N}(0, \sigma)$
  Execute action a and observe reward r, new state s' and done flag d
  Store transitions in the replay buffer, i.e., $\mathcal{B} \leftarrow \mathcal{B} \bigcup \{(s, a, r, s', d)\}$
  for $i = 1, 2$ **do**
    Sample N transitions $\{(s_j, a_j, r_j, s'_j, d_j)\}_{j=1}^N \sim \mathcal{B}$
    Get actions: $a' \leftarrow \pi_{\phi'_1}(s') + \epsilon$, $a'' \leftarrow \pi_{\phi'_2}(s') + \epsilon$, $\epsilon \sim \text{clip}(\mathcal{N}(0, \bar{\sigma}), -c, c)$
    Calculate $\hat{V}(s')$ with Eq. (4)
    $y_t \leftarrow r + \gamma(1-d)\hat{V}(s')$
    Update critic θ_i with Eq. (7)
    Update actor ϕ_i with Eq. (8)
    Update target networks: $\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i$, $\phi'_i \leftarrow \tau\phi_i + (1-\tau)\phi'_i$
  end for
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 ν is mainly selected from $[0, 0.5]$ with 0.05 as interval by using grid search.

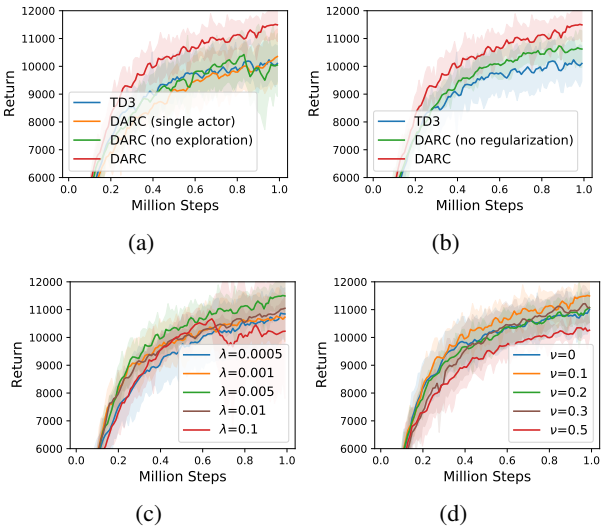


Figure 5: Ablation study on HalfCheetah-v2 (5 runs, mean $\pm$ standard deviation). (a) Components; (b) Regularization; (c) Regularization parameter λ; (d) Weighting coefficient ν.

We use the same hyperparameters in DARC as the default setting for TD3 on all tasks except Humanoid-v2 where all these methods fail with default hyperparameters. We run for 3×10^6 timesteps for Humanoid-v2 and 1×10^6 timesteps for the rest of the tasks for better illustration. Details for hyperparameters are listed in Appendix E.

Ablation Study

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 λ. λ balances the influence of the difference in two critic networks. Large λ may cause instability in value estimation and impede the agent from learning a good policy. While small λ 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 λ as long as it is not too large, we set $\lambda = 0.005$ as default in the rest experiments.

The weighting coefficient ν. The weighting coefficient ν directly influences the value estimation of DARC. Large ν would yield underestimation issues and small ν may induce overestimation bias. We show in Fig 5(d) that there exists a suitable ν that could offer the best trade-off.

Extensive Experiments

The overall performance comparison is presented in Fig 6 where the solid line represents the averaged return and the shaded region denotes standard deviation. We use the smoothing strategy with sliding window 3 that is suggested in OpenAI baselines (Dhariwal et al. 2017) for better demonstration. As demonstrated in Fig 6, DARC significantly outperforms TD3 with much higher sample efficiency, e.g., DARC consumes 50% fewer interaction times to reach the highest return than TD3 in HalfCheetah-v2 task with around 30% additional training steps. DARC learns much faster than other methods.

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

Actor-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

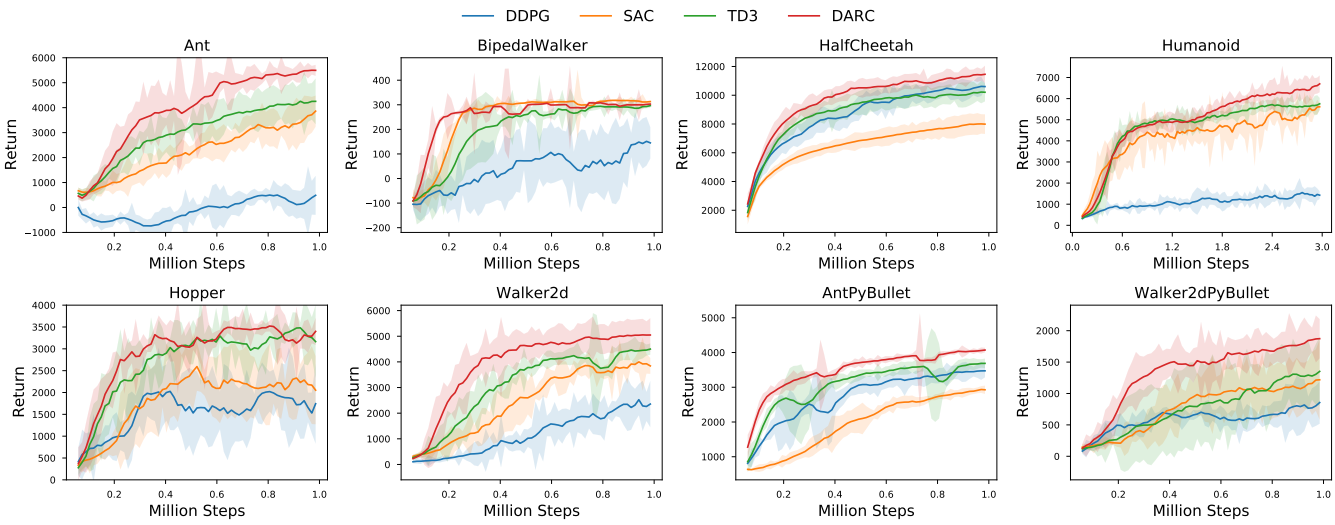


Figure 6: Performance comparison in OpenAI Gym and PyBullet Gym environments.

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.

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

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, Pereverzev, and Rosasco 2007), computer vision (Girosi, 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 Ilin 2020; D'Oro and Jaśkowski 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

Anschel, O.; Baram, N.; and Shimkin, N. 2017. Averaged-dqn: Variance reduction and stabilization for deep reinforcement learning. In *International Conference on Machine Learning*, 176–185. PMLR.

Baird, L. 1995. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, 30–37. Elsevier.

Barth-Maron, G.; Hoffman, M. W.; Budden, D.; Dabney, W.; Horgan, D.; Tb, D.; Muldal, A.; Heess, N.; and Lillicrap, T. 2018a. Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*.

Barth-Maron, G.; Hoffman, M. W.; Budden, D.; Dabney, W.; Horgan, D.; TB, D.; Muldal, A.; Heess, N.; and Lillicrap, T. 2018b. Distributional Policy Gradients. In *International Conference on Learning Representations*.

Bauer, F.; Pereverzev, S.; and Rosasco, L. 2007. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1): 52–72.

Bellemare, M. G.; Dabney, W.; and Munos, R. 2017. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, 449–458. PMLR.

Boney, R.; Kannala, J.; and Ilin, A. 2020. Regularizing model-based planning with energy-based models. In *Conference on Robot Learning*, 182–191. PMLR.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Catto, E. 2011. Box2d: A 2d physics engine for games. *URL: http://www. box2d. org*.

Ciosek, K.; Vuong, Q.; Loftin, R.; and Hofmann, K. 2019. Better Exploration with Optimistic Actor-Critic. In *Advances in Neural Information Processing Systems*, 1785–1796.

Dhariwal, P.; Hesse, C.; Klimov, O.; Nichol, A.; Plappert, M.; Radford, A.; Schulman, J.; Sidor, S.; Wu, Y.; and Zhokhov, P. 2017. OpenAI Baselines. https://github.com/openai/baselines.

D'Oro, P.; and Jaśkowski, W. 2020. How to Learn a Useful Critic? Model-based Action-Gradient-Estimator Policy Optimization. In *Advances in Neural Information Processing Systems*, volume 33.

Ellenberger, B. 2018. PyBullet Gymperium. https://github.com/benelot/pybullet-gym.

Feinberg, V.; Wan, A.; Stoica, I.; Jordan, M. I.; Gonzalez, J. E.; and Levine, S. 2018. Model-based value estimation for efficient model-free reinforcement learning. *arXiv preprint arXiv:1803.00101*.

Fujimoto, S. 2018. Open-source implementation for TD3. https://github.com/sfujim/TD3.

Fujimoto, S.; Hoof, H.; and Meger, D. 2018. Addressing Function Approximation Error in Actor-Critic Methods. In *International Conference on Machine Learning*, 1587–1596.

Girosi, F.; Jones, M.; and Poggio, T. 1995. Regularization theory and neural networks architectures. *Neural computation*, 7(2): 219–269.

Gordon, G. J. 1995. Stable function approximation in dynamic programming. In *Machine Learning Proceedings 1995*, 261–268. Elsevier.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018a. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 1861–1870. PMLR.

Haarnoja, T.; Zhou, A.; Hartikainen, K.; Tucker, G.; Ha, S.; Tan, J.; Kumar, V.; Zhu, H.; Gupta, A.; Abbeel, P.; et al. 2018b. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*.

Hasselt, H. 2010. Double Q-learning. *Advances in Neural Information Processing Systems*, 23: 2613–2621.

Hasselt, H. v.; Guez, A.; and Silver, D. 2016. Deep reinforcement learning with double Q-Learning. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2094–2100.

Horgan, D.; Quan, J.; Budden, D.; Barth-Maron, G.; Hessel, M.; Van Hasselt, H.; and Silver, D. 2018. Distributed prioritized experience replay. *arXiv preprint arXiv:1803.00933*.

Konda, V. R.; and Borkar, V. S. 1999. Actor-critic-type learning algorithms for markov decision processes. *SIAM Journal on Control and Optimization*, 38(1): 94–123.

Konda, V. R.; and Tsitsiklis, J. N. 2000. Actor-critic algorithms. In *Advances in neural information processing systems*, 1008–1014. Citeseer.

Kuznetsov, A.; Shvechikov, P.; Grishin, A.; and Vetrov, D. 2020. Controlling overestimation bias with truncated mixture of continuous distributional quantile critics. In *International Conference on Machine Learning*, 5556–5566. PMLR.

Lan, Q.; Pan, Y.; Fyshe, A.; and White, M. 2020. Maxmin Q-learning: Controlling the estimation bias of Q-learning. In *International Conference on Learning Representations*.

Lange, S.; Gabel, T.; and Riedmiller, M. 2012. Batch reinforcement learning. In *Reinforcement learning*, 45–73. Springer.

Lillicrap, T. P.; Hunt, J. J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; and Wierstra, D. 2015. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*.

Ma, X.; Xia, L.; Zhou, Z.; Yang, J.; and Zhao, Q. 2020. DSAC: Distributional Soft Actor Critic for Risk-Sensitive Reinforcement Learning. In *Reinforcement Learning for Real Life Workshop at ICML 2019*.

Mnih, V.; Badia, A. P.; Mirza, M.; Graves, A.; Lillicrap, T.; Harley, T.; Silver, D.; and Kavukcuoglu, K. 2016. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*, 1928–1937. PMLR.

Nachum, O.; Gu, S.; Lee, H.; and Levine, S. 2018. Data-efficient hierarchical reinforcement learning. In *Advances in Neural Information Processing Systems*.

Pan, L.; Cai, Q.; and Huang, L. 2020. Softmax Deep Double Deterministic Policy Gradients. *Advances in Neural Information Processing Systems*, 33.

Prokhorov, D. V.; and Wunsch, D. C. 1997. Adaptive critic designs. *IEEE transactions on Neural Networks*, 8(5): 997–1007.

Roy, S.; Bakshi, S.; and Maharaj, T. 2020. OPAC: Opportunistic Actor-Critic. *arXiv preprint arXiv:2012.06555*.

Sabry, M.; and Khalifa, A. 2019. On the Reduction of Variance and Overestimation of Deep Q-Learning. *arXiv preprint arXiv:1910.05983*.

Silver, D.; Lever, G.; Heess, N.; Degris, T.; Wierstra, D.; and Riedmiller, M. 2014. Deterministic Policy Gradient Algorithms. In *International Conference on Machine Learning*, 387–395.

Song, Z.; Parr, R.; and Carin, L. 2019. Revisiting the softmax bellman operator: New benefits and new perspective. In *International Conference on Machine Learning*, 5916–5925. PMLR.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement learning: An introduction*. MIT press.

Thrun, S.; and Schwartz, A. 1993. Issues in using function approximation for reinforcement learning. In *Proceedings of the Fourth Connectionist Models Summer School*, 255–263. Hillsdale, NJ.

Tianhong, D. 2019. Open-source implementation for SAC. https://github.com/TianhongDai/reinforcement-learning-algorithms.

Todorov, E.; Erez, T.; and Tassa, Y. 2012. Mujoco: A physics engine for model-based control. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026–5033. IEEE.

Wan, L.; Zeiler, M.; Zhang, S.; Le Cun, Y.; and Fergus, R. 2013. Regularization of neural networks using dropconnect. In *International conference on machine learning*, 1058–1066. PMLR.

Weng, L. 2018. Policy Gradient Algorithms. lilianweng.github.io/lil-log. https://lilianweng.github.io/lil-log/2018/04/08/policy-gradient-algorithms.html.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4): 229–256.

Wu, D.; Dong, X.; Shen, J.; and Hoi, S. C. 2020. Reducing estimation bias via triplet-average deep deterministic policy gradient. *IEEE Transactions on Neural Networks and Learning Systems*, 31(11): 4933–4945.

Wu, Y.; Tucker, G.; and Nachum, O. 2019. Behavior regularized offline reinforcement learning. *arXiv preprint arXiv:1911.11361*.

Xu, Z.; Yan, J.; Luo, J.; Wells, W.; Li, X.; and Jagadeesan, J. 2021. Unimodal Cyclic Regularization For Training Multimodal Image Registration Networks. In *IEEE 18th International Symposium on Biomedical Imaging*, 1660–1664. IEEE.

Zhang, S.; and Whiteson, S. 2019. DAC: The Double Actor-Critic Architecture for Learning Options. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 2010–2020.

Zhang, Z.; Pan, Z.; and Kochenderfer, M. J. 2017. Weighted Double Q-learning. In *International Joint Conferences on Artificial Intelligence*, 3455–3461.

Zhao, R.; Sun, X.; and Tresp, V. 2019. Maximum entropy-regularized multi-goal reinforcement learning. In *International Conference on Machine Learning*, 7553–7562. PMLR.