ERROR CONTROLLED ACTOR-CRITIC METHOD TO REINFORCEMENT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In the reinforcement learning (RL) algorithms which incorporate function approximation methods, the approximation error of value function inevitably cause overestimation phenomenon and have a negative impact on the convergence of the algorithms. To mitigate the negative effects of approximation error, we propose a new actor-critic algorithm called Error Controlled Actor-critic which ensures confining the approximation error in value function. In this paper, we firstly present an analysis of how the approximation error can hinder the optimization process of actor-critic methods. Then, we derive an upper boundary of the approximation error of Q function approximator, and found that the error can be lowered by placing restrictions on the KL-divergence between every two consecutive policies during the training phase of the policy. The results of experiments on a range of continuous control tasks from OpenAI gym suite demonstrate that the proposed actor-critic algorithm apparently reduces the approximation error and significantly outperforms other model-free RL algorithms.

1 INTRODUCTION

In pratice, reinforcement learning (RL) are combined with function approximation methods to adapt to the application scenarios whose state spaces are combinatorial, large, or even continuous. Many function approximation methods RL methods, including the Fourier basis(Konidaris et al., 2011), kernel regression(Xu, 2006; Barreto et al., 2011; Bhat et al., 2012), and neural neworks(Barto et al., 1982; Tesauro, 1992; Boyan et al., 1992; Gullapalli, 1992) have been used to learn value functions. In recent years, many deep reinforcement learning (DRL) methods were implemented by incorporating deep learning into RL methods. Deep Q-learning Network (DQN)(Mnih et al., 2013) reported by Mnih in 2013 is a typical work that uses a deep convolutional neural network (CNN) to represent a suitable action value function estimating future rewards (returns); it successfully learned end-to-end control policies for seven Atari 2600 games directly from large state spaces. Thereafter, deep RL methods, such as Deep Deterministic Policy Gradient (DDPG)(Lillicrap et al., 2016), Proximal Policy Optimization (PPO)(Schulman et al., 2017), Twin Delayed Deep Deterministic policy gradient (TD3)(Fujimoto et al., 2018), and Soft Actor-Critic (SAC)(Haarnoja et al., 2018), started to become mainstream in the field of RL.

Althouth function approximation methods have assisted reinforcement learning (RL) algorithms to perform well in complex problems by providing great representation power; however, they also cause an issue called overestimation phenomenon that jeopardize the optimization process of RL algorithms. Thrun & Schwartz (1993) presented a theoretical analysis of this systematic overestimation phenomenon in Q-learning methods that use function approximation methods. Similar problem persists in the actor-critic methods employed function approximation methods. Thomas (2014) reported that several natural actor-critic algorithms use biased estimates of policy gradient to update parameters when using function approximation to approximate the action value function. Fujimoto et al. (2018) proved that the value estimation in the deterministic policy gradient method also lead to overestimation problem. In brief, the approximation errors of value functions caused the inaccuracy of estimated values, and such inaccuracy induced the overestimation on value function; so that poor performances might be assigned to high reward values. As a result, policies with poor performance might be obtained.

Previous works attempted to find direct strategies to effectively reduce the overestimation.Hasselt (2010) proposed Double Q-learning, in which the samples are divided into two sets to train two independent Q-function estimators. To diminish the overestimation, one Q-function estimator is used to select actions, and the other one is applied to estimate its value. Fujimoto et al. (2018) proposed mechanisms, including clipped double Q-learning and delayed policy updates, to minimize the overestimation.

In contrast to these methods, we focus on actor-critic setting and manage to diminish the approximation error of value function, which is the source of the overestimation, in an indirect but effective way. Before introducing our method, we first prove convergence of actor-critic methods under certain conditions basing on the work reported by Kumar, and conclude that it is difficult to meet the conditions when using function approximators. Then, inspired by the work (Ben-David et al., 2010) on domain adaptation, we use the concepts of domain adaptation to derive an upper boundary of the approximation error of Q function approximator. Further, we find that the least upper bound of this error can be obtained by minimizing the Kullback-Leibler divergence (KL divergence) between the polices before and after each training iteration. This means minimizing the KL divergence when traning policy can indirectly confine the approximation error of value function. Furthermore, this indirect operation can work together with the mentioned direct strategies for reducing overestimation, for example, clipped double Q-learning. Then, a new actor-critic method called Error Controlled Actor-critic (ECAC) is established by adopting an effective operation that minimizes the KL divergence to keep the upper bound as low as possible. In other words, this method ensures the similarity between every two consecutive polices in training process and reduces the optimization difficulty of value function, so that the error in value function approximators can be decreased. Ablation studies were performed to examine the effectiveness of our proposed strategy for decreasing the approximation error, and comparative evaluations were conducted to verify that our method can outperform other mainstream RL algorithms.

The main contributions of this paper are summarized as follow: (1) We present a convergence analysis for actor-critic methods by Banach's Fixed Point Theorem, and find a key condition of convergence, which is hard to meet when function approximators are used to approximate value functions. (2) We analyze the source of approximation error in value function of actor-critic methods, and prove that restricting the KL divergence between every two consecutive policies during training process can decrease the approximation error.

2 PRELIMINARIES

2.1 REINFORCEMENT LEARNING

Reinforcement learning (RL) algorithms are modeled as a mathematical framework called Markov Decision Process (MDP). In each time-step of MDP, an agent generates an action based on current state of its environment, then receives a reward and a new state from the environment. Environmental state and agent's action at time t are denoted as $\boldsymbol{s}_t \in \mathcal{S}$ and $\boldsymbol{a}_t \in \mathcal{A}$, respectively; $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces respectively, which may be either discrete or continuous. Environment is described by a reward function, $r(\boldsymbol{s}_t, \boldsymbol{a}_t)$, and a transition probability distribution, $P(\boldsymbol{s}_{t+1} = \boldsymbol{s}'|\boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a})$. Transition probability distribution specifies the probability that the environment will transition to next state. Initial state distribution is denoted as $P_0(\boldsymbol{s})$.

The goal of RL is to find a policy, π, that maximizes the expected discounted rewards:

$$E_\pi[R_1 + \gamma R_2 + \gamma^2 R_3 + \cdots] = E_\pi[\sum_{t=0}^{\infty} \gamma^k R_{t+1}], \tag{1}$$

where γ denotes a discount rate and $0 \leq \gamma \leq 1$. Different interpretations of this optimization goal result in different routes to the their solutions. RL methods can be roughly divided into three categories: 1) value-based, 2) policy-based, and 3) actor-critic methods. Value-based methods have the advantage of fast convergence while policy-based methods have better convergence guarantees. Actor-critic framework tries to trade off a reduction of learning speed for a better convergence guarantee by incorporating value-based methods into policy-based methods. An actor-critic method consists of two components: actor and critic modules. In n-th iteration, actor module tries to find a better policy, $\pi^{*,n}$, than the current policy, π^n, with the help of value functions (state-value or

action-state-value function or both); and the critic module learns value functions, V^{π^n} or Q^{π^n}, of current policy π^n, which outperforms $V^{\pi^{n-1}}$ or $Q^{\pi^{n-1}}$, by means of $\pi^{*,n-1}$ and learning algorithms, e.g. temporal-difference (TD) learning. Fig 1 illustrates the update process of an actor-critic method from a perspective similar to policy iteration which is a alternating process of value function and policy updates.

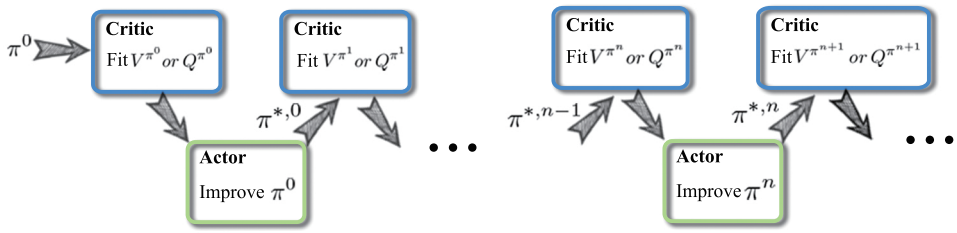


Figure 1: The alternating process of value function and policy updates.

2.2 Domain Adaptation

Domain adaptation is the task which aims at adapting a well performing model from a source domain to a different target domain. Domain is defined as a specific pair consisting of a distribution, $\mathcal{D}$, on an input space, $\mathcal{X}$, and a labeling function, $f : \mathcal{X} \to \mathbb{R}$. In domain adaption, source and target domains are denoted as $< \mathcal{D}_S, f_S >$ and $< \mathcal{D}_T, f_T >$, respectively. A function, $h : \mathcal{X} \to \mathbb{R}$, is defined as a hypothesis. And the difference between a hypothesis, $h(x)$, and a labeling function, $f(x)$, on a source distribution is denoted as follow:

$$\epsilon_S(h, f) = \underset{x \sim \mathcal{D}_S}{E} [|h(x) - f(x)|] \tag{2}$$

3 Problem Definition–Hidden Dangers in Actor-critic

In this section, we investigate the continuous state and action space, which are the most complex setting, to define an underlying problem in actor-critic methods. First, we prove convergence of the actor-critic methods that utilize action value function to improve policies. Then, we find a hidden danger in the optimization process of actor-critic methods when use the generalizing approximation methods.

3.1 Actor-critic Convergence Theorem

We now consider a simple actor-critic setting, as described below. There are a Q function to be estimated and a policy to be learned. At each iteration of optimization, the actor module attempts to find a better policy than the current one with the help of the Q function. We define this attempting operations by:

$$\pi^{*,n} = f_{Actor}(Q^{\pi^n}(\boldsymbol{s}, \boldsymbol{a})), \tag{3}$$

where the mapping, f_{Actor}, satisfies the following condition:

$$\forall \pi', \underset{\substack{\boldsymbol{s} \sim \rho_\pi, \\ \boldsymbol{a} \sim \pi'(\cdot|\boldsymbol{s})}}{E} [Q^{\pi^n}(\boldsymbol{s}, \boldsymbol{a})] \leq \underset{\substack{\boldsymbol{s} \sim \rho_\pi, \\ \boldsymbol{a} \sim \pi^{*,n}(\cdot|\boldsymbol{s})}}{E} [Q^{\pi^n}(\boldsymbol{s}, \boldsymbol{a})], \tag{4}$$

where ρ_π denotes state marginals of the trajectory distribution induced by the policy $\pi(\boldsymbol{a}|\boldsymbol{s})$.

The critic module tries to estimate the Q function by using the mentioned bootstrapping operation, for example, the following operation in a recursive form:

$$Q^{\pi^{n+1}}(\boldsymbol{s}, \boldsymbol{a}) = \underset{\substack{\boldsymbol{s}' \sim p(\cdot|\boldsymbol{s}, \boldsymbol{a}), \\ \boldsymbol{a}' \sim \pi^{*,n}(\cdot|\boldsymbol{s}')}}{E} [R_{t+1} + \gamma Q^{\pi^n}(\boldsymbol{s}', \boldsymbol{a}')|\boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}] \tag{5}$$

where $p(\cdot|s, a)$ denotes the transition probability distribution.

We provide the convergence theorem for this actor-critic setting, and its proof is provided in Appendix A.

Theorem 1 (Actor-critic Convergence Theorem)**.** *Given the following two conditions:*

(1) Actor operation maps an actor value function, Q^{π^n}, into a policy, $\pi^{*,n}$, that satisfies:

$$\forall \pi', \mathop{E}_{\substack{\boldsymbol{s}\sim\rho_\pi,\\ \boldsymbol{a}\sim\pi'(\cdot|\boldsymbol{s})}} [Q^{\pi^n}(\boldsymbol{s},\boldsymbol{a})] \leq \mathop{E}_{\substack{\boldsymbol{s}\sim\rho_\pi,\\ \boldsymbol{a}\sim\pi^{*,n}(\cdot|\boldsymbol{s})}} [Q^{\pi^n}(\boldsymbol{s},\boldsymbol{a})]. \tag{6}$$

(2) Critic operation updates actor value function via Bellman equation (or bootstrapping operation).

Then repeated application of actor-critic operation from any $\pi \in \Pi$ converges to a policy π^*.

3.2 Hidden Dangers in Actor-critic

Although actor-critic methods are theoretically guaranteed to converge to an optimal policy, when using function approximation, it is practically difficult to strictly meet the first condition of Theorem 1, i.e. it is practically hard to find a better policy, $\pi^{*,n}(\cdot|\boldsymbol{s})$, by using a noised Q function. In addition, because of the error of value function approximator, there is an overestimation on the value function as presented in the previous works (Thrun & Schwartz, 1993; Thomas, 2014; Fujimoto et al., 2018). Further, the overestimated Q values result in the bias of gradient estimation, which will also affect the system's convergence.

4 Error Controlled Actor-critic

To solve the problem described in Section 3—the approximation error of value function can lead actor-critic algorithms to fail to converge, we propose a new off-policy actor-critic algorithm called error controlled actor-critic (ECAC) to reduce the impact of the approximation error. We derive an upper boundary of the error of Q function approximator, and find that the error can be reduced by limiting the KL divergence between every two consecutive policies. Although this operation is conducted when training the policy, it can indirectly reduce the optimization difficulty of Q function. Moreover, this indirect operation can work together with the strategies for diminishing overestimation phenomenon. We incorporate clipped double-Q learning into the critic module.

In this section, we present the details of a version of ECAC for continous control tasks, and, more importantly, explain the rationale for confining the KL divergence.

4.1 Critic module–learning action value functions

The learning task of critic module is to approximate Q functions. In the critic module of ECAC, two Q functions are approximated by two neural networks with weight $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, respectively. As noted previously, we adopt clipped double-Q strategy to directly reduce overestimation. Furthermore, we adopt experience replay mechanism (Lin, 1992)—agent's experiences at each timestep, $(\boldsymbol{s}_t, \boldsymbol{a}_t, r_{t+1}, \boldsymbol{s}_{t+1})$, are stored in a replay buffer, $\mathcal{D}$; and training samples are uniformly drawn from this buffer. The two Q networks, $Q_{\boldsymbol{\phi}_1}$ and $Q_{\boldsymbol{\phi}_2}$, are trained by using temporal-difference learning. Notice that the bootstrapping operation in this setting—uses function approximation—means to minimize the following TD error of Q funtion:

$$\delta_t = Q(\boldsymbol{s}_t, \boldsymbol{a}_t) - (R_{t+1} + \gamma Q(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1})) \tag{7}$$

Moreover, clipped double-Q learning uses the smaller of the two Q values to form the TD error. The two Q networks are respectively trained by minimizing the following two loss functions:

$$L(\boldsymbol{\phi}_1) = \mathop{\mathrm{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a},r,\boldsymbol{s}')\sim\mathcal{D}\\ s'\sim p(\cdot|\boldsymbol{s},\boldsymbol{a}),\\ \boldsymbol{a}'\sim\pi(\cdot|\boldsymbol{s}')}} [Q_{\boldsymbol{\phi}_1}(\boldsymbol{s},\boldsymbol{a}) - (r + \gamma \min_{i=1,2} Q_{\boldsymbol{\phi}_i}(\boldsymbol{s}',\boldsymbol{a}'))]^2 \tag{8}$$

$$L(\boldsymbol{\phi}_2) = \mathop{\mathrm{E}}_{\substack{(\boldsymbol{s},\boldsymbol{a},r,\boldsymbol{s}')\sim\mathcal{D}\\ s'\sim p(\cdot|\boldsymbol{s},\boldsymbol{a}),\\ \boldsymbol{a}'\sim\pi(\cdot|\boldsymbol{s}')}} [Q_{\boldsymbol{\phi}_2}(\boldsymbol{s},\boldsymbol{a}) - (r + \gamma \min_{i=1,2} Q_{\boldsymbol{\phi}_i}(\boldsymbol{s}',\boldsymbol{a}'))]^2 \tag{9}$$

where $\mathcal{D}$ denotes replay buffer, $p(\cdot|\boldsymbol{s},\boldsymbol{a})$ denotes the transition probability distribution, and π denotes the target policy.

4.2 THE UPPER BOUND OF THE APPROXIMATION ERROR OF Q FUNCTION

For convenience, we analyze the setting with only one Q function. Here, the concept of domain adaptation from transfer learning is used to describe the task of critic module and to derive the upper bound of the approximation error of Q function.

At the $(n+1)$-th iteration, the learning task of critic module is a subtask that can be views adapting the learned Q-network, Q_{ϕ^n}, to the next one, $Q_{\phi^{n+1}}$. The distribution on $\boldsymbol{s}$ and $\boldsymbol{a}$ is denoted as $\mathcal{D}^{n+1}$, and the target for training the Q-network, $y_{\pi^{n+1}}(\boldsymbol{s},\boldsymbol{a})$, is defined by:

$$y_{\pi^{n+1}}(\boldsymbol{s},\boldsymbol{a}) = R_{t+1} + \gamma Q_{\phi^n}(\boldsymbol{s}',\boldsymbol{a}'), \boldsymbol{s}' \sim p(\cdot|\boldsymbol{s},\boldsymbol{a}), \boldsymbol{a}' \sim \pi^{*,n}(\cdot|\boldsymbol{s}), \tag{10}$$

where $p(\cdot|\boldsymbol{s},\boldsymbol{a})$ denotes the transition probability distribution, and $\pi^{*,n}$ denotes the policy updated in actor module by using Q_{ϕ^n}.

Furthermore, the difference between the Q-network, $Q_\phi(\boldsymbol{s},\boldsymbol{a})$, and its target at the n-th iteration, y_{π^n}, on a *source* distribution is given by

$$\epsilon_S(Q_\phi) = \underset{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}^n}{E}[|Q_\phi(\boldsymbol{s},\boldsymbol{a}) - y_{\pi^n}(\boldsymbol{s},\boldsymbol{a})|] = \underset{\substack{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}^n,\\ \boldsymbol{s}'\sim p(\cdot|\boldsymbol{s},\boldsymbol{a}),\\ \boldsymbol{a}'\sim\pi^{*,n}}}{E}[|Q_\phi(\boldsymbol{s},\boldsymbol{a}) - [R_{t+1} + \gamma Q_{\phi^{n-1}}(\boldsymbol{s}',\boldsymbol{a}')]|]. \tag{11}$$

The difference between the Q-network, $Q_\phi(\boldsymbol{s},\boldsymbol{a})$, and its target at $(n+1)$-th iteration, $y_{\pi^{n+1}}$, on a *target* distribution is given by

$$\epsilon_T(Q_\phi) = \underset{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}^{n+1}}{E}[|Q_\phi(\boldsymbol{s},\boldsymbol{a}) - y_{\pi^{n+1}}(\boldsymbol{s},\boldsymbol{a})|] = \underset{\substack{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}^{n+1},\\ \boldsymbol{s}'\sim p(\cdot|\boldsymbol{s},\boldsymbol{a}),\\ \boldsymbol{a}'\sim\pi^{*,n+1}}}{E}[|Q_\phi(\boldsymbol{s},\boldsymbol{a}) - [R_{t+1} + \gamma Q_{\phi^n}(\boldsymbol{s}',\boldsymbol{a}')]|]. \tag{12}$$

The difference between the Q-network, $Q(\boldsymbol{s},\boldsymbol{a})$, and its target at $(n+1)$-th iteration, $y_{\pi^{n+1}}$, on a *source* distribution is given by

$$\epsilon_S(Q_\phi, y_{\pi^{n+1}}) = \underset{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}^n}{E}[|Q_\phi(\boldsymbol{s},\boldsymbol{a}) - y_{\pi^{n+1}}(\boldsymbol{s},\boldsymbol{a})|] = \underset{\substack{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}^n,\\ \boldsymbol{s}'\sim p(\cdot|\boldsymbol{s},\boldsymbol{a}),\\ \boldsymbol{a}'\sim\pi^{*,n+1}}}{E}[|Q_\phi(\boldsymbol{s},\boldsymbol{a}) - [R_{t+1} + \gamma Q_{\phi^n}(\boldsymbol{s}',\boldsymbol{a}')]|] \tag{13}$$

We first derive the following inequality:

$$\begin{aligned}\epsilon_T(Q_\phi) =& \epsilon_T(Q_\phi) + \epsilon_S(Q_\phi) - \epsilon_S(Q_\phi) + \epsilon_S(Q_\phi, y_{\pi^{n+1}}) - \epsilon_S(Q_\phi, y_{\pi^{n+1}})\\ \le& \epsilon_S(Q_\phi) + |\epsilon_S(Q_\phi, y_{\pi^{n+1}}) - \epsilon_S(Q_\phi)| + |\epsilon_T(Q_\phi) - \epsilon_S(Q_\phi, y_{\pi^{n+1}})|\\ \le& \epsilon_S(Q_\phi) + \underset{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}^n}{E}[|y_{\pi^{n+1}}(\boldsymbol{s},\boldsymbol{a}) - y_{\pi^n}(\boldsymbol{s},\boldsymbol{a})|] + |\epsilon_T(Q_\phi) - \epsilon_S(Q_\phi, y_{\pi^{n+1}})|,\end{aligned} \tag{14}$$

where the third term in the third line, $|\epsilon_T(Q_\phi) - \epsilon_S(Q_\phi, y_{\pi^{n+1}})|$, is transformed further as:

$$\begin{aligned}&|\epsilon_T(Q_\phi) - \epsilon_S(Q_\phi, y_{\pi^{n+1}})|\\ =&\left|\underset{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}^{n+1}}{E}[|Q_\phi(\boldsymbol{s},\boldsymbol{a}) - y_{\pi^{n+1}}(\boldsymbol{s},\boldsymbol{a})|] - \underset{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}^n}{E}[|Q_\phi(\boldsymbol{s},\boldsymbol{a}) - y_{\pi^{n+1}}(\boldsymbol{s},\boldsymbol{a})|]\right|\\ \le&\gamma \underset{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}}{E}\left|\underset{\substack{\boldsymbol{s}'\sim p(\cdot|\boldsymbol{s},\boldsymbol{a}),\\ \boldsymbol{a}'\sim\pi^{*,n+1}(\cdot|\boldsymbol{s}')}}{E}[Q_{\phi^n}(\boldsymbol{s}',\boldsymbol{a}')] - \underset{\substack{\boldsymbol{s}'\sim p(\cdot|\boldsymbol{s},\boldsymbol{a}),\\ \boldsymbol{a}'\sim\pi^{*,n}(\cdot|s')}}{E}[Q_{\phi^n}(\boldsymbol{s}',\boldsymbol{a}')]\right|\end{aligned} \tag{15}$$

Recall that $\mathcal{D}$ is actually the replay buffer; and the number of samples in $\mathcal{D}^{n+1}$ is only a little more than in $\mathcal{D}^n$, we ignore the difference between $\mathcal{D}$ and $\mathcal{D}^{n+1}$. Finally, the upper bound of error in Q-network is determined by

$$\begin{aligned}\epsilon_T(Q_\phi) \le& \epsilon_S(Q_\phi) + \underset{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}^n}{E}[|y_{\pi^{n+1}}(\boldsymbol{s},\boldsymbol{a}) - y_{\pi^n}(\boldsymbol{s},\boldsymbol{a})|]\\ &+ \gamma \underset{\boldsymbol{s},\boldsymbol{a}\sim\mathcal{D}}{E}\left|\underset{\substack{\boldsymbol{s}'\sim p(\cdot|\boldsymbol{s},\boldsymbol{a}),\\ \boldsymbol{a}'\sim\pi^{*,n+1}(\cdot|\boldsymbol{s}')}}{E}[Q_{\phi^n}(\boldsymbol{s}',\boldsymbol{a}')] - \underset{\substack{\boldsymbol{s}'\sim p(\cdot|\boldsymbol{s},\boldsymbol{a}),\\ \boldsymbol{a}'\sim\pi^{*,n}(\cdot|s')}}{E}[Q_{\phi^n}(\boldsymbol{s}',\boldsymbol{a}')]\right|.\end{aligned} \tag{16}$$

It is noticeable that the third term in the upper bound will be smaller if the more similar the two consecutive policies, $\pi^{*,n}$ and $\pi^{*,n+1}$ are. Hence, we can conclude that confining the KL divergence between every two consecutive policies can help limit the approximation error during the optimization process of actor-critic. This conclusion is used to design the learning method of the policy.

4.3 ACTOR MODULE–LEARNING A POLICY

The learning task of the actor module is to learn a parameterized stochastic policy, $\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})$.In order to lower the upper bound of approximation error of Q function (or to reduce the optimization difficulty of the Q functions), the goal of the policy is converted from maximizing expect discounted return two parts: maximizing the estimated Q values—the minimun of the two Q approximators—and, concurrently, minimizing the KL divergence between two successive policies. The optimization objective is specified by

$$\max_{\boldsymbol{\theta}} \mathop{E}_{\boldsymbol{s}\sim\mathcal{D}}[\min_{i=1,2} Q_{\boldsymbol{\phi}_i}(\boldsymbol{s}, \widetilde{\boldsymbol{a}}_{\boldsymbol{\theta}}(\boldsymbol{s})) - D_{KL}(\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s}), \pi_{\boldsymbol{\theta}_{old}}(\cdot|\boldsymbol{s}))], \tag{17}$$

where $\mathcal{D}$ is the replay buffer; $\boldsymbol{\theta}$ is the parameters of the policy network; $\boldsymbol{\theta}_{old}$ is the parameters of the policy updated in the last iteration; $\widetilde{\boldsymbol{a}}_{\boldsymbol{\theta}}(\boldsymbol{s})$ is the samples drawn from the target stochastic policy$\pi_{\boldsymbol{\theta}}(\boldsymbol{a}|\boldsymbol{s})$. Note that, in order to back-propagate the error through this sampling operation, we use a squashed Gaussian policy and the reparameterization trick, i.e. samples are obtained according to

$$\widetilde{\boldsymbol{a}}_{\boldsymbol{\theta}}(\boldsymbol{s}) = \mu_{\boldsymbol{\theta}}(\boldsymbol{s}) + \sigma_{\boldsymbol{\theta}}(\boldsymbol{s}) \odot \boldsymbol{\xi}, \quad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}), \tag{18}$$

where $\mu_{\boldsymbol{\theta}}(\boldsymbol{s})$ and $\sigma_{\boldsymbol{\theta}}(\boldsymbol{s})$ are the output of the policy network, and denotes the mean and covariance matrix's diagonal elements, respectively.

Moreover, the KL divergence between two distributions, for example $p(x)$ and $q(x)$, can be thought of the difference between the cross entropy and entropy, which is specified by

$$D_{KL}(p||q) = H(p, q) - H(p), \tag{19}$$

where $H(p, q)$ denotes the cross entropy between $p(x)$ and $q(x)$, and $H(p)$ denotes the entropy of $p(x)$. In practice, we find it is more effective to minimize the cross entropy between two successive policies and to maximize the entropy of current policy, separately, than to minimize the KL divergence directly. Hence, we expand the original objective Eq. 17 into the following one:

$$\max_{\boldsymbol{\theta}} \mathop{E}_{\boldsymbol{s}\sim\mathcal{D}}[\min_{i=1,2} Q_{\boldsymbol{\phi}_i}(\boldsymbol{s}, \widetilde{\boldsymbol{a}}_{\boldsymbol{\theta}}(\boldsymbol{s})) - \alpha H(\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s}), \pi_{\boldsymbol{\theta}_{old}}(\cdot|\boldsymbol{s})) + \beta H(\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s}))], \tag{20}$$

where α and β denote the coefficients of the cross entropy and the entropy, respectively.

Moreover, we adopt a mechanism to automatically adjust α and β. To do this, α is adjusted by making the current cross entropy close to a target value. This mechanism is specified by the following optimization objective:

$$\min_{\alpha} \mathop{E}_{\boldsymbol{s}\sim\mathcal{D}}[\log \alpha \cdot ((\delta_{KL} + \delta_{entropy}) - H(\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s}), \pi_{\boldsymbol{\theta}_{old}}(\cdot|\boldsymbol{s})))], \tag{21}$$

where δ_{KL} denotes the target KL value and $\delta_{entropy}$ denotes the target entropy. Note that $\log(\cdot)$ is use to ensure that α is greater than 0. β is adjusted in the same way:

$$\min_{\beta} \mathop{E}_{\boldsymbol{s}\sim\mathcal{D}}[\log \beta \cdot (H(\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s})) - \delta_{entropy})]. \tag{22}$$

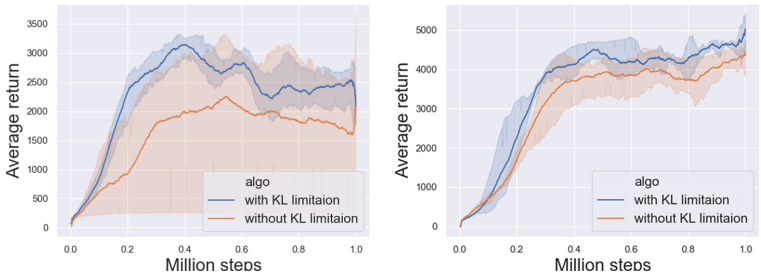
The overall training process is summarized in Appendix B, and code can be found on our GitHub `https://github.com/SingerGao/ECAC`.

5 EXPERIMENTS

The experiments aim to evaluate the effectiveness of the proposed strategy to lower the approximation error and to verify that the proposed method can outperform other mainstream RL algorithms. The experiments for ablation study and comparative evaluation are conducted on a few challenging continuous control tasks from the OpenAI Gym(Brockman et al., 2016) environments, which includes Mujoco(Todorov et al., 2012) and Pybullet (Coumans & Bai, 2016–2019) versions. Implementation details and hyperparameters of ECAC are presented in Appendix C.

5.1 ABLATION STUDY

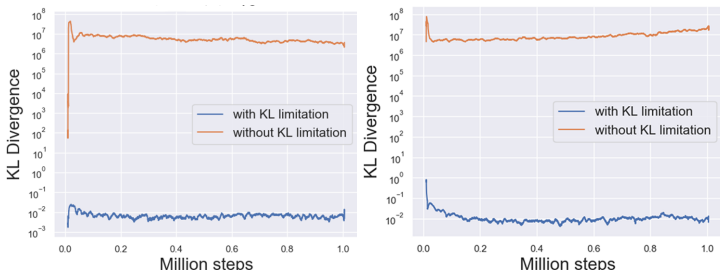
Ablation studies are performed to verify the contribution of the operation of KL limitation. We compared the performance of ECAC with the method removing KL limitation from ECAC.



(a) Hopper-v3 (b) Walker2d-v3

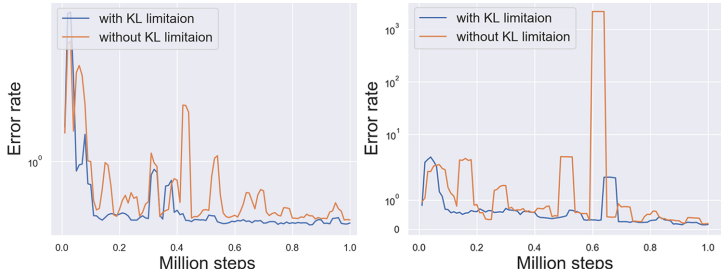
Figure 2: Performance comparison of the method with KL limitation and the one without KL limitation on the Hopper-v3 and Walker2d-v3 benchmark. The method with KL limitation performs better. Curves are smoothed uniformly for visual clarity.

Figure 2 compares five different instances for both methods with and without KL limition using different random seeds; and each instance performs five evaluation episodes every $1,000$ environment steps. The solid curves corresponds to the mean and the shaded region to the minimum and maximum returns over the five runs. The experimental result shows that our method with KL limitation performs better than the one without KL limitation. Figure 3 demonstrates that by using ECAC the KL divergence remains comfortably low during all the training process.



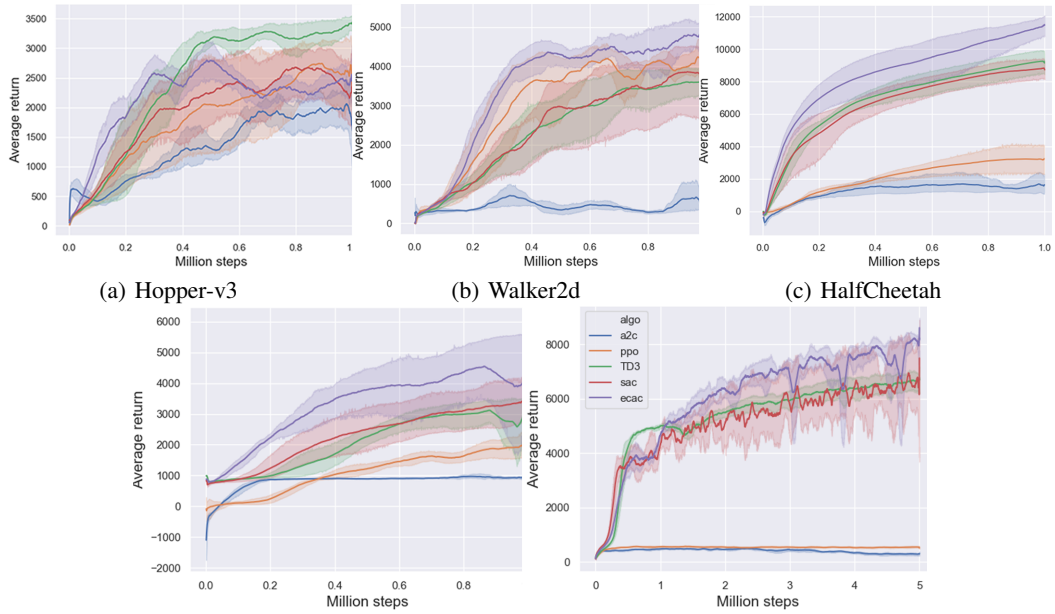
(a) Hopper-v3 (b) Walker2d-v3

Figure 3: The KL divergence comparison of the method with KL limitation and the one without KL limitation on the Hopper-v3 and Walker2d-v3 benchmark.



(a) Hopper-v3 (b) Walker2d-v3

Figure 4: The approximation error comparison of the method with KL limitation and the one without KL limitation on the Hopper-v3 and Walker2d-v3 benchmark.



(a) Hopper-v3 (b) Walker2d (c) HalfCheetah

(d) Ant-v3 (e) Humanoid-v3

Figure 5: The results of comparative evaluation on Mujoco version of the OpenAI gym continuous control tasks. Curves are smoothed uniformly for visual clarity.

Furthermore, to verify that confining the KL divergence can decrease the approximation error in Q function, we measured the normalized approximation error in 100 random states every $10,000$ environment steps. The normalized approximation error is specified by

$$e_Q = \frac{|Q_{approx} - Q_{ture}|}{|Q_{ture}|} \tag{23}$$

where Q_{approx} is the approximate Q value given by the current Q network, and Q_{ture} is the true discounted return. The true value is estimated using the average discounted return over 100 episodes following the current policy, starting from states sampled from the replay buffer. Figure 4 shows that the method with KL limitation have lower error in the Q function.

The results of all the ablation studies indicates that the approximation error of Q functioncan be decreased and the performance of the RL algorithm can be improved by placing restrictions on KL divergence between every two consecutive policies.

5.2 Comparative Evaluation

Comparative evaluation are conducted to verify that our method can outperform other traditional RL methods including A2C, PPO, TD3, and SAC. Five individual runs of each algorithm with different random seeds are done; and each run performs five evaluation episodes every $1,000$ environment steps. Figure 5 demonstrates the results of comparative evaluation on the Mujoco version of continuous control tasks. ECAC outperforms all other algorithms on the tasks except Hopper-v3 and are only next to TD3 on Hopper-v3. The results of comparative evaluation on the Pybullet version of continuous control tasks are presented in Appendix D.

6 Conclusion

This paper presented a model-free actor-critic method based on a finding that the approximation error in value function of RL methods can be decreased by placing restrictions on KL-divergence between every two consecutive policies. Our method increases the similarity between every two consecutive polices in the training process and therefore reduces the optimization difficulty of value function. In the ablation studies, we compare the approximation error in Q function, KL divergence, and

performance of the methods with and without KL limitation. The experimental results show that the proposed method can decrease the approximation error and improved the performance. Moreover, the results of comparative evaluation demonstrate that ECAC outperforms other model-free deep RL algorithm including A2C, PPO, TD3, and SAC.

REFERENCES

Andre Barreto, Doina Precup, and Joelle Pineau. Reinforcement learning using kernel-based stochastic factorization. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 24*, pp. 720–728. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4217-reinforcement-learning-using-kernel-based-stochastic-factorization.pdf.

Andrew G Barto, Charles W Anderson, and Richard S Sutton. Synthesis of nonlinear control surfaces by a layered associative search network. *Biological Cybernetics*, 43(3):175–185, 1982.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010. doi: 10.1007/s10994-009-5152-4. URL https://doi.org/10.1007/s10994-009-5152-4.

Nikhil Bhat, Vivek Farias, and Ciamac C Moallemi. Non-parametric approximate dynamic programming via the kernel method. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25*, pp. 386–394. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4547-non-parametric-approximate-dynamic-programming-via-the-kernel-method.pdf.

Justin A. Boyan, Justin A. Boyan, Justin A. Boyan, and Justin A. Boyan. Modular neural networks for learning context-dependent game strategies. Technical report, Masters thesis, Computer Speech and Language Processing, 1992.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. http://pybullet.org, 2016–2019.

Scott Fujimoto, Herke van Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1587–1596, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR. URL http://proceedings.mlr.press/v80/fujimoto18a.html.

Vijaykumar Gullapalli. *Reinforcement Learning and Its Application to Control*. PhD thesis, USA, 1992.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Jennifer G. Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1856–1865. PMLR, 2018. URL http://proceedings.mlr.press/v80/haarnoja18b.html.

Hado V. Hasselt. Double q-learning. In J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (eds.), *Advances in Neural Information Processing Systems 23*, pp. 2613–2621. Curran Associates, Inc., 2010. URL http://papers.nips.cc/paper/3964-double-q-learning.pdf.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv: Learning*, 2014.

George Dimitri Konidaris, Sarah Osentoski, and Philip S. Thomas. Value function approximation in reinforcement learning using the fourier basis. In Wolfram Burgard and Dan Roth (eds.), *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011*. AAAI Press, 2011. URL http://www.aaai.org/ocs/index.php/AAAI/AAAI11/paper/view/3569.

Vaibhav Kumar. Mathematical analysis of reinforcement learning bellman optimality equation. [EB/OL]. URL https://towardsdatascience.com/mathematical-analysis-of-reinforcement-learning-bellman-equation-ac9f0954e19f.

Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1509.02971.

Long-Ji Lin. *Reinforcement Learning for Robots Using Neural Networks*. PhD thesis, USA, 1992.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. Playing atari with deep reinforcement learning. *CoRR*, abs/1312.5602, 2013. URL http://arxiv.org/abs/1312.5602.

Arthur L. Samuel. Some studies in machine learning using the game of checkers. *IBM J. Res. Dev.*, 3 (3):210–229, 1959. doi: 10.1147/rd.33.0210. URL https://doi.org/10.1147/rd.33.0210.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017. URL http://arxiv.org/abs/1707.06347.

David Silver, Aja Huang, Chris J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Vedavyas Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy P. Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529(7587):484–489, 2016. doi: 10.1038/nature16961. URL https://doi.org/10.1038/nature16961.

Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998. ISBN 0262193981. URL http://www.cs.ualberta.ca/%7Esutton/book/ebook/the-book.html.

Gerald Tesauro. Practical issues in temporal difference learning. *Mach. Learn.*, 8:257–277, 1992. doi: 10.1007/BF00992697. URL https://doi.org/10.1007/BF00992697.

Philip Thomas. Bias in natural actor-critic algorithms. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, volume 32 of *JMLR Workshop and Conference Proceedings*, pp. 441–448. JMLR.org, 2014. URL http://proceedings.mlr.press/v32/thomas14.html.

Sebastian Thrun and Anton Schwartz. Issues in using function approximation for reinforcement learning. In *In Proceedings of the Fourth Connectionist Models Summer School*. Erlbaum, 1993.

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

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

Xin Xu. A sparse kernel-based least-squares temporal difference algorithm for reinforcement learning. In Licheng Jiao, Lipo Wang, Xinbo Gao, Jing Liu, and Feng Wu (eds.), *Advances in Natural Computation, Second International Conference, ICNC 2006, Xi'an, China, September 24-28, 2006. Proceedings, Part I*, volume 4221 of *Lecture Notes in Computer Science*, pp. 47–56. Springer, 2006. doi: 10.1007/11881070_8. URL https://doi.org/10.1007/11881070_8.

APPENDICES

A PROOF OF CONVERGENCE OF ACTOR-CRITIC METHODS

We prove the convergence of actor-critic methods by using Banach's fixed point theorem.

Definition 1 (Contraction)**.** *Let X be a metric space equipped with a distance d. $f : X \to X$ is considered as a contraction mapping, if there is a constant, $0 \leq \lambda < 1$, thus:*

$$d(f(x_1), f(x_2)) \leq \lambda d(f(x_1), f(x_2)), \forall x_1, x_2 \in X. \tag{24}$$

Lemma 1 (Banach's Fixed Point Theorem)**.** *Let f be a contraction on a complete metric space X. Then f has a unique fixed point $x^* \in X$.*

Proof of Theorem 1. For convenience, the mentioned two actor-critic operations, i.e. equation 3 and 5, is combined into the following recursive expression:

$$Q^{\pi^{n+1}} = f_{AC}(Q^{\pi^n}(\boldsymbol{s}, \boldsymbol{a})). \tag{25}$$

Let (X, d) denotes the metric space of action value function. The set $X = \{Q(\boldsymbol{s}, \boldsymbol{a})\}$ is the hypothesis space of value functions; d is a metric measuring the distance between two action value functions, which is defined as follow:

$$d(Q_1(\boldsymbol{s}, \boldsymbol{a}), Q_2(\boldsymbol{s}, \boldsymbol{a})) = \max_{\boldsymbol{s}, \boldsymbol{a}} [|Q_1(\boldsymbol{s}, \boldsymbol{a}) - Q_2(\boldsymbol{s}, \boldsymbol{a})|]. \tag{26}$$

Let Q_1 and Q_2 be two action value functions, then we have:

$$\begin{aligned}
&||f_{AC}(Q_1(\boldsymbol{s}, \boldsymbol{a})) - f_{AC}(Q_2(\boldsymbol{s}, \boldsymbol{a}))|| \\
\leq &|| \mathop{E}_{\substack{\boldsymbol{s}' \sim p(\cdot|\boldsymbol{s}, \boldsymbol{a}), \\ \boldsymbol{a}'_1 \sim \pi_1^{*,n}(\cdot|\boldsymbol{s}')}} [R_{t+1} + \gamma Q_1(\boldsymbol{s}', \boldsymbol{a}'_1) | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}] \\
&- \mathop{E}_{\substack{\boldsymbol{s}' \sim p(\cdot|\boldsymbol{s}, \boldsymbol{a}), \\ \boldsymbol{a}'_2 \sim \pi_2^{*,n}(\cdot|\boldsymbol{s}')}} [R_{t+1} + \gamma Q_2(\boldsymbol{s}', \boldsymbol{a}'_2) | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}]|| \\
\leq &|| \mathop{E}_{\substack{\boldsymbol{s}' \sim p(\cdot|\boldsymbol{s}, \boldsymbol{a}), \\ \boldsymbol{a}' \sim \pi_1^{*,n}(\cdot|\boldsymbol{s}')}} [R_{t+1} + \gamma Q_1(\boldsymbol{s}', \boldsymbol{a}') - [R_{t+1} + \gamma Q_2(\boldsymbol{s}', \boldsymbol{a}')] | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}]|| \\
\leq &\gamma || \mathop{E}_{\substack{\boldsymbol{s}' \sim p(\cdot|\boldsymbol{s}, \boldsymbol{a}), \\ \boldsymbol{a}' \sim \pi_1^{*,n}(\cdot|\boldsymbol{s}')}} [Q_1(\boldsymbol{s}', \boldsymbol{a}') - Q_2(\boldsymbol{s}', \boldsymbol{a}') | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}]|| \\
\leq &\gamma || \max_{\boldsymbol{s}', \boldsymbol{a}'} (Q_1(\boldsymbol{s}', \boldsymbol{a}') - Q_2(\boldsymbol{s}', \boldsymbol{a}'))|| \\
\leq &\gamma \max_{\boldsymbol{s}, \boldsymbol{a}} |Q_1(\boldsymbol{s}, \boldsymbol{a}) - Q_2(\boldsymbol{s}, \boldsymbol{a})| \\
= &\gamma ||Q_1(\boldsymbol{s}, \boldsymbol{a}) - Q_2(\boldsymbol{s}, \boldsymbol{a})||.
\end{aligned} \tag{27}$$

In Definition 1, we determine that f_{AC} is a contraction mapping for the complete metric space (X, d); in addition, since the results of operations of actor-critic always stay in hypothesis space of value functions, the space, f_{AC}, is complete. Then the condition of lemma 1 holds. According to the Banach's fixed point theorem, the repeated application of f_{AC} will converge to a unique optimal actor value function; and an optimal policy can be derived by using the Banach's fixed point theorem. □

B PSEUDO CODE OF ECAC

Algorithm 1 Error controlled Actor-critic.

Require: initial policy patameters, $\boldsymbol{\theta}$; Q function parameters, $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$; discount rate, γ; the coefficient of KL term, β; empty replay buffer $\mathcal{D}$; the number of episodes, M; the maximum number of steps in each episode, T; minibatch size, N.

Ensure: optimal policy parameters $\boldsymbol{\theta}^*$.

for $episode = 1, \mathrm{M}$ **do**
  Reset environment.
  for $t = 1, \mathrm{T}$ **do**
    Obeserve state $\boldsymbol{s}$ and select action $\boldsymbol{a} \sim \pi(\cdot|\boldsymbol{s})$.
    Execute $\boldsymbol{a}$ in the enviroment.
    Observe next state $\boldsymbol{s}'$ and reward r.
    Store transition $(\boldsymbol{s}, \boldsymbol{a}, r, \boldsymbol{s}')$ in replay buffer $\mathcal{D}$.
    Randomly sample a minibatch of N transitions, $B = \{(\boldsymbol{s}, \boldsymbol{a}, r, \boldsymbol{s}')\}$ from $\mathcal{D}$.
    Compute targets for the Q functions:

$$y(r, \boldsymbol{s}') = r + \gamma \min_{i=1,2} Q_{\boldsymbol{\phi}_i}(\boldsymbol{s}', \widetilde{\boldsymbol{a}}'),\ \widetilde{\boldsymbol{a}}' \sim \pi_\theta(\cdot|\boldsymbol{s}').$$

    Update Q functions by one step of gradient descent using

$$\nabla_{\boldsymbol{\phi}} \frac{1}{N} \sum_{(\boldsymbol{s},\boldsymbol{a},r,\boldsymbol{s}')\in B} (Q_{\boldsymbol{\phi}_i}(\boldsymbol{s}, \boldsymbol{a}) - y(r, \boldsymbol{s}'))^2,\ i = 1, 2.$$

    Backup old policy, $\boldsymbol{\theta}_{old} \leftarrow \boldsymbol{\theta}$.
    Update α and β by one step of gradient ascent using

$$\nabla_\alpha \frac{1}{N} \sum_{s\in B} [\log \alpha \cdot ((\delta_{KL} + \delta_{entropy}) - H(\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s}), \pi_{\boldsymbol{\theta}_{old}}(\cdot|\boldsymbol{s})))],$$

$$\nabla_\beta \frac{1}{N} \sum_{s\in B} [\log \beta \cdot (H(\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s})) - \delta_{entropy})],$$

    where δ_{KL} and $\delta_{entropy}$ denote target KL divergence and target entropy, respectively.
    Update policy by one step of gradient ascent using

$$\nabla_{\boldsymbol{\theta}} \frac{1}{N} \sum_{s\in B} (Q_{\boldsymbol{\phi}_1}(\boldsymbol{s}, \widetilde{\boldsymbol{a}}_{\boldsymbol{\theta}}(\boldsymbol{s})) + \alpha H(\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s})) - \beta D_{KL}(\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s}), \pi_{\boldsymbol{\theta}_{old}}(\cdot|\boldsymbol{s}))),$$

    where $\widetilde{\boldsymbol{a}}_{\boldsymbol{\theta}}(s)$ is a sample from $\pi_{\boldsymbol{\theta}}(\cdot|\boldsymbol{s})$, which is differentiable with respect to θ via the reparameterization trick.
  end for
end for

Table 1: ECAC Hyperparameters

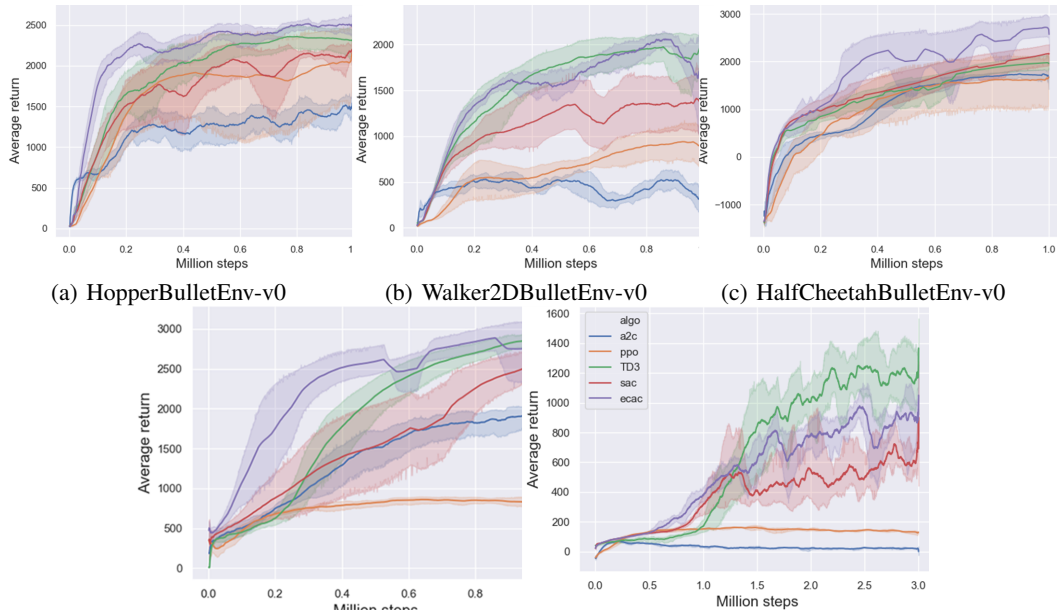
Parameter	Value
learning rate	10^{-3}
discount(γ)	0.99
replay buffer size	$5 \cdot 10^5$
batch size	128
target KL	$5 \cdot 10^{-3}$
target entropy	$-dim(\mathcal{A})/2.0$ (e.g. , -3 for Walker2d-v3)
target smoothing coefficient (τ)	$5 \cdot 10^{-3}$
number of hidden layers (all networks)	2
number of hidden units per layer	256
nonlinearity	ReLU

Table 2: Reward Scale Parameter

Environment	Reward Scale
Hopper-v3	5
Walker2d-v3	5
HalfCheetah-v3	5
Ant-v3	5
Humanoid-v3	20
HopperBulletEnv-v0	5
Walker2DBulletEnv-v0	5
HalfCheetahBulletEnv-v0	5
AntBulletEnv-v0	5
HumanoidBulletEnv-v0	20

C IMPLEMENTATION DETAILS AND HYPERPARAMETERS OF ECAC

For the implementation of ECAC, a two layer feedforward neural network of 256 hidden units, with rectified linear units (ReLU) between each layer are used to build the two Q functions an the policy. The parameters of the neural networks and the two coefficients (i.e α and β) are optimized by using Adam(Kingma & Ba, 2014). The hyperparameters of ECAC are listed in Table 1. Moreover, we adopt the target policy smoothing regularization technique in ECAC, which is proposed in Fujimoto et al. (2018). We also adopt reward scale trick which is presented in Haarnoja et al. (2018); and the reward scale parameter is listed in Table 2.



(a) HopperBulletEnv-v0 (b) Walker2DBulletEnv-v0 (c) HalfCheetahBulletEnv-v0

(d) AntBulletEnv-v0 (e) HumanoidBulletEnv-v0

Figure 6: The results of comparative evaluation on Pybullet version of the OpenAI gym continuous control tasks. Curves are smoothed uniformly for visual clarity.

D Additional Results of Comparative Evaluation

Figure 6 shows that ECAC outperforms all other algorithms on the tasks except HumanoidBulletEnv-v0 and are only next to TD3 on HumanoidBulletEnv-v0.