Abstract—Training a game-playing reinforcement learning agent requires multiple interactions with the environment. Ignorant random exploration may cause a waste of time and resources. It's essential to alleviate such waste. As discussed in this paper, under the settings of the off-policy actor critic algorithms, we demonstrate that the critic can bring more expected discounted rewards than or at least equal to the actor. Thus, the Q value predicted by the critic is a better signal to redistribute the action originally sampled from the policy distribution predicted by the actor. This paper introduces the novel Critic Guided Action Redistribution (CGAR) algorithm and tests it on the OpenAI MuJoCo tasks. The experimental results demonstrate that our method improves the sample efficiency and achieves state-of-the-art performance. Our code can be found at https://github.com/tairanhuang/CGAR.

Index Terms—Reinforcement Learning, Soft actor-critic, MuJoCo tasks

I. INTRODUCTION

In recent years, reinforcement learning has been widely used in games and has made excellent progress in Atari, StarCraft, Dota2, Honor of Kings, and other games [1], [2]. However, training a reinforcement learning model is time-consuming due to the massive interactions between the learning and the environment [3]–[5]. Also, interactions with naive exploration strategies slow down the model’s learning speed and waste resources during model training [6], [7]. These deficiencies restrict the applications of reinforcement learning in games.

Off-policy algorithms store the experience in the buffer and reuse them to reduce interaction costs. Actor critic algorithms use an actor to select actions and use a critic to estimate the value function to reduce the variance of policy gradient and accelerate the convergence. Deep deterministic policy gradient (DDPG) [8] and Soft Actor-Critic (SAC) algorithm [9] combine the advantages of off-policy algorithms and actor critic algorithms. DDPG learns a critic and an actor at the same time. It uses the Bellman function to optimize the critic and uses the critic to optimize the actor. SAC [9] maximizes both expected return and entropy [10], [11] to balance exploration and exploitation. The actor of SAC predicts the action distribution and then samples action in the training stage and uses the mean value of the distribution as the action in the evaluation stage. It further improves the sample efficiency and stability of reinforcement learning.

Following the same purpose, we propose a novel Critic Guided Action Redistribution (CGAR) mechanism and show that SAC with CGAR achieves state-of-the-art performance. In this paper, we first give a theoretical analysis of actor critic algorithms and demonstrate that the critic can bring more expected discounted rewards than or at least equal to the actor in off-policy actor critic algorithms. To utilize such advancement of critic, we redistribute the action distribution predicted by the actor through the Q value predicted by the critic.

In the proposed CGAR model, after getting an action distribution predicted by the actor, we sample $K$ actions from that action distribution. Then, the critic model predicts the Q values for these $K$ actions conditioned on the state. We reset the selection probability of these actions based on their corresponding Q values. Since the critic model is optimized to predict the expected return, with the distribution positively correlated to the Q value, the algorithm tends to select the action with a higher expected return. We deploy our method on SAC and conduct experiments on the OpenAI MuJoCo tasks [12]. The experimental results demonstrate that our approach is effective and achieves state-of-the-art performance. Here we summarize our main contributions:

- We demonstrate that the critic can bring more expected discounted rewards than or at least equal to the actor in the off-policy actor critic algorithm.
- We propose a novel Critic Guided Action Redistribution (CGAR), which uses the Q value predicted by the critic to resample action from the action distribution predicted by the actor.
We apply our method to SAC and achieve state-of-the-art performance on OpenAI MuJoCo tasks.

II. ALGORITHM

A. Motivation

Under the classic setting of the off-policy actor critic learning (environment step), the agent collects data from the environment with the action distribution \(\pi_{\phi_a}(s)\). Then, given the updated data buffer \(D\), the critic is optimized to estimate the future reward. Then, the actor is updated to maximize the estimated expected future reward given the currently learned critic. The procedure forms a dependency circle and loops at each environment step during training.

At a certain environment step, suppose that the critic \(Q_{\phi_c}\) is optimized to \(Q_{\phi_c}^i\) after \(i\) gradient step under the currently collected data \(D\). The loss function to optimize actor is generally written as,

\[
J_{\pi}(\phi_a) = E_{s_t \sim D} E_{a_t \sim \pi_{\phi_a}(a_t|s_t)}(-Q_{\phi_c}^i(s_t, a_t)).
\]

When the loss function is minimized, the action with the highest probability predicted by the actor is the same as the action leading to the highest \(Q_{\phi_c}^i\) predicted by the critic. Then the performance of the actor and the critic is the same. However, most of the time, the actor is less optimized to the optimal distribution of action given the current critic, which leads to poor performance compared with the current critic. This learning procedure is similar to the knowledge distillation or teacher-student method [13]. In our situation, we use the Q value output by the critic to calculate the loss function to train the actor. Hence, critic and actor correspond to teacher and student, respectively. The performance gap between the student and teacher has been demonstrated in many previous works, and the gap exists even though the student network has the same size as the teacher network [14]–[16]. In Section II-B, we also give an empirical demonstration of the motivation. As the performance of the critic is better than or at least equal to the actor in each environment step, we can expect that the critic can bring more expected discounted rewards than or at least equal to the actor during the RL training procedure. This paper proposes a CGAR algorithm to ameliorate the performance gap between the actor and critic.

B. Empirical Demonstration of the Motivation

We conduct experiments under supervised learning settings to empirically demonstrate the above motivation. In detail, We use MNIST [17] as a dataset and maintain two identical multilayer perceptron models, \(M_1\) and \(M_2\). We let \(M_1\) fit the dataset and let \(M_2\) fit \(M_1\). Given input \(x\) and its label \(y\), the loss function of \(M_1\) is: \(L_1 = \text{CrossEntropy}(M_1(x), y)\), and the loss function of \(M_2\) is: \(L_2 = \text{CrossEntropy}(M_2(x), M_1(x))\). Under this setting, \(M_1\) is the signal provided to \(M_2\), and \(y\) is the signal provided to \(M_1\). Besides, \(M_1\) and \(M_2\) are updated one after another iteratively. This setting is quite similar to the off-policy actor critic algorithm in Section II-A, where the reward is provided to calculate the critic’s target, and the Q is the signal provided to the actor. To simulate the loss function of critic, we further design another loss function, \(L_1 = \text{MSE}(M_1(x), \bar{y})\), where \(\bar{y}\) is the one-hot representation of \(y\), and \(L_2\) remains unchanged. We test the evaluation accuracy of \(M_1\) and \(M_2\) every epoch over five seeds and report the results in Fig. 2. We can see that under both loss functions, \(M_2\) learns more slowly than \(M_1\), demonstrating our motivation’s correctness.

C. Critic Guided Action Redistribution

We propose our Critic Guided Action Redistribution (CGAR) based on the above motivation and empirical demonstration. In actor critic algorithms, given state \(s_t\) at environment step \(t\), the actor \(\pi_{\phi_a}\) predicts the action distribution \(P_t\). After that, it samples action \(a_t\) from \(P_t\) to interact with the environment.

\[
a_t \sim P_t = \pi_{\phi_a}(s_t).
\]
In our algorithm, we first sample $K$ actions $\{a_1, a_2, \ldots, a_K\}$ from $P_t$. These actions construct the actions set $\mathbb{L}^a$. Then we use the critic to predict the Q value for each action in $\mathbb{L}^a$. The Q value $Q_i$ is calculated on each action $a_i$ in $\mathbb{L}^a$ conditioned with state $s_t$. These values $\{Q_0, Q_1, \ldots, Q_K\}$ construct the Q value set $\mathbb{L}^Q$.

$$
\mathbb{L}^a := \{a_i \sim P_t | i \in [0, K - 1]\}.
$$

(3)

We use the Softmax function to make actions with large Q values more likely to be sampled while maintaining a certain degree of exploration. Finally, we select the action $a_t$ from the new distribution. Fig. 1 is our model’s diagram.

$$
a_t \sim P_t'.
$$

(6)

### D. Implementation

We apply CGAR to SAC. SAC is an off-policy algorithm based on the Maximum Entropy Principle. Its optimization goal is simultaneously maximizing both the expected return and the entropy. It learns a policy $\pi_{\phi_a}$, a Q value function $Q_{\phi_c}$, and a temperature coefficient $\alpha$, with parameters $\phi_c$, $\phi_a = (\phi_{a1}, \phi_{a2})$, and $\alpha$ separately. The loss functions of SAC are defined below, which are introduced in the SAC paper.

$$
J_Q(\phi_c) = \mathbb{E}_{s_t, a_t, r_t} \left[ \frac{1}{2} (Q_{\phi_c}(s_t, a_t) - r_t + \gamma \mathbb{E}_{s_{t+1} \sim p} (V_{\phi_c}(s_{t+1}))^2 \right],
$$

(7)

$$
V_{\phi_c}(s_t) = \mathbb{E}_{a_t, s_t, r_t} \pi_{\phi_a} \pi_{\phi_c}(a_t, s_t) Q_{\phi_c}(s_t, a_t) - \alpha \log \pi_{\phi_a}(a_t | s_t),
$$

(8)

$$
J_\pi(\phi_a) = \mathbb{E}_{s_t, a_t \sim P_t} \pi_{\phi_a}(a_t | s_t) \mathbb{E}_{a_t, s_t, r_t} \pi_{\phi_c} \pi_{\phi_a}(a_t, s_t) \left( -\alpha \log \pi_{\phi_c}(a_t | s_t) - Q_{\phi_c}(s_t, a_t) \right).
$$

(9)

$$
J_\alpha(\alpha) = \mathbb{E}_{a_t \sim \pi_{\phi_a} | s_t} - \alpha \log \pi_{\phi_a}(a_t | s_t) - \alpha H.
$$

(10)

Note that different from Eq. (1), there’s an entropy term in Eq. (9), which doesn’t affect the demonstration in Section II-A. Our complete algorithm is shown in Algorithm 1. Our unique operations are marked with red. And the deleted operations in SAC are marked with blue and strikethrough.

### Algorithm 1 CGAR applied to SAC

1. **Input:** ENV
2. **Output:** $\phi_c$, $\phi_a$, and $\alpha$
3. Initialize parameters $\phi_c$, $\phi_a$, and $\alpha$
4. Set $\phi_c = \phi_c$, $D = \emptyset$
5. for $k = 1, \ldots, N_{\text{init}}$
   6. Sample random action: $a_t \sim \pi_{\text{random}}(s_t)$
   7. Execute action: $s_{t+1}, r_{t+1}, \text{done} \sim \text{ENV}(a_t)$
   8. Collect data: $D \leftarrow D \cup \{s_t, a_t, s_{t+1}, r_{t+1}\}$
5. end for
10. for $k = 1, \ldots, N_{\text{train}}$
11. Predict action distribution: $P_t = \pi(s_t)$
12. Sample action: $a_t \sim P_t$
13. Sample $K$ actions from $P_t$: $\mathbb{L}_t = \{a_t \sim P_t | i \in [0, K - 1]\}$
14. Calculate the Q value set: $\mathbb{L}_Q = \{Q_i \sim Q(s_t) | i \in [0, K - 1]\}$
15. Get the new distribution: $P_t' = \text{Softmax}(L_Q)$
16. Sample action: $a_t \sim P_t'$
17. Execute action: $s_{t+1}, r_{t+1}, \text{done} \sim \text{ENV}(a_t)$
18. Collect data: $D \leftarrow D \cup \{s_t, a_t, s_{t+1}, r_{t+1}\}$
19. Train critic: $\phi_c \leftarrow \phi_c - \alpha \nabla \phi_c J_Q(\phi_c)$
20. Train actor: $\phi_a \leftarrow \phi_a - \lambda \nabla \phi_a J_\pi(\phi_a)$
21. Train alpha: $\alpha = \alpha - \lambda_{\alpha} \nabla_\alpha J_\alpha(\alpha)$
22. Update the target critic: $\phi_c \leftarrow \tau \phi_c + (1 - \tau) \bar{\phi}_c$
23. end for

### III. Experiment

#### A. Implementation Details

We implement CGAR on the SAC algorithm noted as CGAR-SAC. The implementation of the SAC comes from [18]. We evaluate our method on OpenAI MuJoCo tasks, including tasks such as standing, walking, and running. The state of the agent consists of parameters such as positions and velocities. Action is a real-valued vector that represents the control of the agent’s joints. The purpose of model learning is to maximize the expected discounted rewards. We test the agent performance every 10,000 environment steps. We compute the mean episode returns an agent obtains over ten episodes for each evaluation. All results are over five different seeds, and we keep the minimum, maximum, and mean values over these seeds.

#### B. Comparison between CGAR-SAC with SAC

We compare the performance of CGAR-SAC with SAC. We counted the mean, maximum, and minimum values of the average return obtained over all seeds during each evaluation and plotted them in Fig. 3. The curve represents the average return, and the shading represents the range between the maximum and minimum. From Fig. 3 we can see that in most tasks, CGAR-SAC converges faster and achieves better final performance than SAC, especially Cheetah Run, Walker Walk, and Finger Spin. And CGAR-SAC is not weaker than SAC in other tasks. We also counted the average return obtained over all seeds during each evaluation and calculated their mean.
of SAC.

We can conclude that CGAR improves the sample efficiency and the standard deviation is lower than SAC in most tasks. We can see that the mean value of the average return of our method over the entire training process is higher than SAC in every task, and the standard deviation is lower than SAC in most tasks. We can therefore conclude that CGAR improves the sample efficiency of SAC.

Fig. 3. Comparison results between CGAR-SAC and SAC.

### IV. CONCLUSION

This paper proposes a novel action redistribution algorithm, Critic Guided Action Redistribution for game playing. We demonstrate that the critic can bring more expected discounted rewards than or at least equal to the actor in the off-policy actor critic algorithm. Based on the demonstration, we use the Q value predicted by the critic to redistribute the actions probability distribution generated by the actor. Then we sample actions from the new distribution to interact with the environment. We implement our algorithm on SAC and test it on the OpenAI MuJoCo tasks. The experimental results demonstrate that our method improves the sample efficiency and achieves state-of-the-art performance. Future research can be done by applying CGAR to other games or analyzing the distribution of Q value on the multimodal distribution.

### REFERENCES


[16] X. Deng and Z. Zhang, “Can students outperform teachers in knowledge distillation based model compression?”
