

CGAR: Critic Guided Action Redistribution in Reinforcement Learning

Tairan Huang
SCSE
Beihang University
Beijing, China
trhuang@buaa.edu.cn

Xu Li
Cognitive Computing Lab
Baidu Research
Beijing, China
lixu13@baidu.com

Hao Li
ECE
Peking University
Beijing, China
2101212812@pku.edu.cn

Mingming Sun
Cognitive Computing Lab
Baidu Research
Beijing, China
sunmingming01@baidu.com

Ping Li
Cognitive Computing Lab
Baidu Research
Seattle, USA
liping11@baidu.com

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.

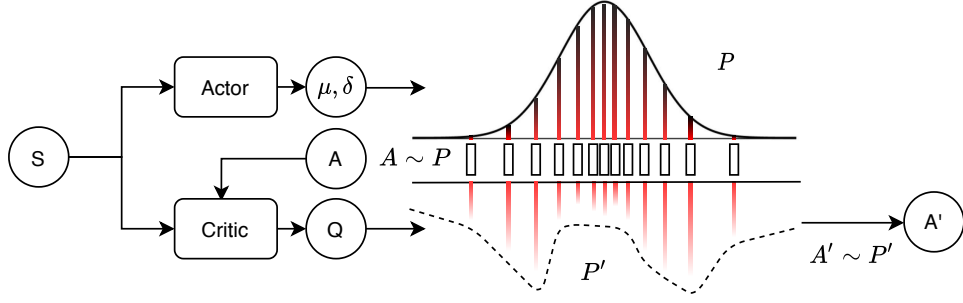


Fig. 1. CGAR: Critic Guided Action Redistribution in Reinforcement Learning

- 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 procedure, at each interaction with the environment (environment step), the agent collects data from the environment under the policy network with the action distribution of $\pi_{\phi_a}(s)$. Then, given the updated data buffer \mathcal{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_{ϕ_c} is optimized to $Q_{\phi_c^i}$ after i gradient step under the currently collected data \mathcal{D} . The loss function to optimize actor is generally written as,

$$J_{\pi}(\phi_a) = \mathbb{E}_{s_t \sim \mathcal{D}} \mathbb{E}_{a_t \sim \pi_{\phi_a}(a_t | s_t)} (-Q_{\phi_c^i}(s_t, a_t)). \quad (1)$$

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 multi-layer 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: $\mathcal{L}_1 = \text{CrossEntropy}(M_1(x), y)$, and the loss function of M_2 is: $\mathcal{L}_2 = \text{CrossEntropy}(M_2(x), M_1(x))$. 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, $\mathcal{L}_1 = \text{MSE}(M_1(x), \bar{y})$, where \bar{y} is the one-hot representation of y , and \mathcal{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.

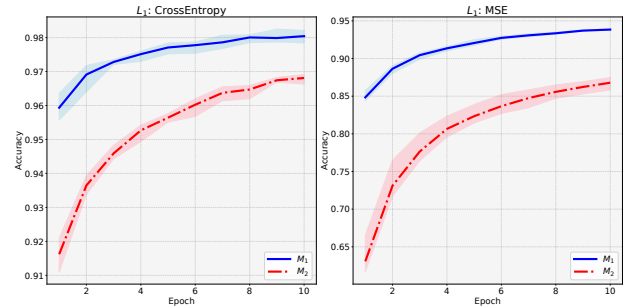


Fig. 2. Comparison results of M_1 and M_2 .

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 π_{ϕ_a} predicts the action distribution \mathcal{P}_t . After that, it samples action a_t from \mathcal{P}_t to interact with the environment.

$$a_t \sim \mathcal{P}_t = \pi_{\phi_a}(s_t). \quad (2)$$

In our algorithm, we first sample K actions $\{a_t^0, a_t^1, \dots, a_t^{K-1}\}$ from \mathcal{P}_t . These actions construct the actions set \mathbb{L}_a^t .

$$\mathbb{L}_a^t := \{a_t^i \sim \mathcal{P}_t | i \in [0, K-1]\}. \quad (3)$$

Then we use the critic to predict the Q value for every action in \mathbb{L}_a^t . The Q value Q_t^i is calculated on every action a_t^i in \mathbb{L}_a^t conditioned with state s_t . These Q values $\{Q_t^0, Q_t^1, \dots, Q_t^{K-1}\}$ construct the Q value set \mathbb{L}_Q^t .

$$\mathbb{L}_Q^t := \{Q_t^i = Q_{\phi_c}(s_t, a_t^i) | i \in [0, K-1]\}. \quad (4)$$

We use the Q value Q_t^i to calculate the probability p_t^i of a_t^i . The probability set results from the Q value set calculated by the Softmax function. In this way, we could get the action probability distribution \mathcal{P}_t' .

$$\begin{aligned} \mathcal{P}_t' &= \text{Softmax}(\mathbb{L}_Q^t) \\ &= \left\{ \frac{e^{Q_t^i}}{\sum_{j=0}^{K-1} e^{Q_t^j}} | i \in [0, K-1] \right\}. \end{aligned} \quad (5)$$

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.

$$a_t \sim \mathcal{P}_t'. \quad (6)$$

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 π_{ϕ_a} , a Q value function Q_{ϕ_c} , and a temperature coefficient α , with parameters ϕ_a , $\phi_c = (\phi_c^1, \phi_c^2)$, and α separately. The loss functions of SAC are defined below, which are introduced in the SAC paper.

$$\begin{aligned} J_Q(\phi_c) &= \mathbb{E}_{s_t, a_t \sim \mathcal{D}} \frac{1}{2} (Q_{\phi_c}(s_t, a_t) \\ &\quad - (r_t + \gamma \mathbb{E}_{s_{t+1} \sim p}(V_{\bar{\phi}_c}(s_{t+1})))^2, \end{aligned} \quad (7)$$

$$\begin{aligned} V_{\bar{\phi}_c}(s_t) &= \mathbb{E}_{a_t \sim \pi_{\phi_a}(a_t | s_t)} Q_{\bar{\phi}_c}(s_t, a_t) \\ &\quad - \alpha \log \pi_{\phi_a}(a_t | s_t), \end{aligned} \quad (8)$$

$$\begin{aligned} J_\pi(\phi_a) &= \mathbb{E}_{s_t \sim \mathcal{D}} \mathbb{E}_{a_t \sim \pi_{\phi_a}(a_t | s_t)} \\ &\quad (\alpha \log \pi_{\phi_a}(a_t | s_t) - Q_{\phi_c}(s_t, a_t)). \end{aligned} \quad (9)$$

$$J_\alpha(\alpha) = \mathbb{E}_{a_t \sim \pi_{\phi_a}(a_t | s_t)} - \alpha \log \pi_{\phi_a}(a_t | s_t) - \alpha \mathcal{H}, \quad (10)$$

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  $\bar{\phi}_c = \phi_c$ ,  $\mathcal{D} = \emptyset$ 
5: for  $k = 1, \dots, N_{\text{init}}$  do
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:  $\mathcal{D} \leftarrow \mathcal{D} \cup \{s_t, a_t, s_{t+1}, r_{t+1}\}$ 
9: end for
10: for  $k = 1, \dots, N_{\text{train}}$  do
11:   Predict action distribution:  $\mathcal{P}_t = \pi(s_t)$ 
12:   Sample action:  $a_t \sim \mathcal{P}_t$ 
13:   Sample  $K$  actions from  $\mathcal{P}_t$ :
14:    $\mathbb{L}_a^t = \{a_t^i \sim \mathcal{P}_t | i \in [0, K-1]\}$ 
15:   Calculate the Q value set:
16:    $\mathbb{L}_Q^t = \{Q_t^i = Q_{\phi_c}(s_t, a_t^i) | i \in [0, K-1]\}$ 
17:   Get the new distribution:  $\mathcal{P}_t' = \text{Softmax}(\mathbb{L}_Q^t)$ 
18:   Sample action:  $a_t \sim \mathcal{P}_t'$ 
19:   Execute action:  $s_{t+1}, r_{t+1}, \text{done} \sim \text{ENV}(a_t)$ 
20:   Collect data:  $\mathcal{D} \leftarrow \mathcal{D} \cup \{s_t, a_t, s_{t+1}, r_{t+1}\}$ 
21:   Train critic:  $\phi_c \leftarrow \phi_c - \lambda_Q \nabla_{\phi_c} J_Q(\phi_c)$ 
22:   Train actor:  $\phi_a \leftarrow \phi_a - \lambda_\pi \nabla_{\phi_a} J_\pi(\phi_a)$ 
23:   Train alpha:  $\alpha \leftarrow \alpha - \lambda_\alpha \nabla_\alpha J_\alpha(\alpha)$ 
24:   Update the target critic:  $\bar{\phi}_c \leftarrow \tau \phi_c + (1 - \tau) \bar{\phi}_c$ 
25: 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 every 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 overall 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 overall seeds during each evaluation and calculated their mean

TABLE I
MEAN VALUE OF THE AVERAGE RETURN FOR THE WHOLE TRAINING PROCESS

Task	BCC	CR	RE	WW	WS	FS
CGAR-SAC	843\pm73	593\pm51	821\pm81	761\pm71	857\pm32	793\pm47
SAC	841 \pm 89	574 \pm 59	817 \pm 82	746 \pm 110	846 \pm 29	760 \pm 49

over the entire training process. In Table I, we use the initials of the first letter of every environment as the table title, and the full names can be found in Fig. 3. The value of each item in the table represents the mean value of the average return, and the subscript denotes the standard deviation. From Table I, we can see that the mean value of the average return of our method during the training process is higher than SAC in every task, and the standard deviation is lower than SAC in most tasks. We can conclude that CGAR improves the sample efficiency of SAC.

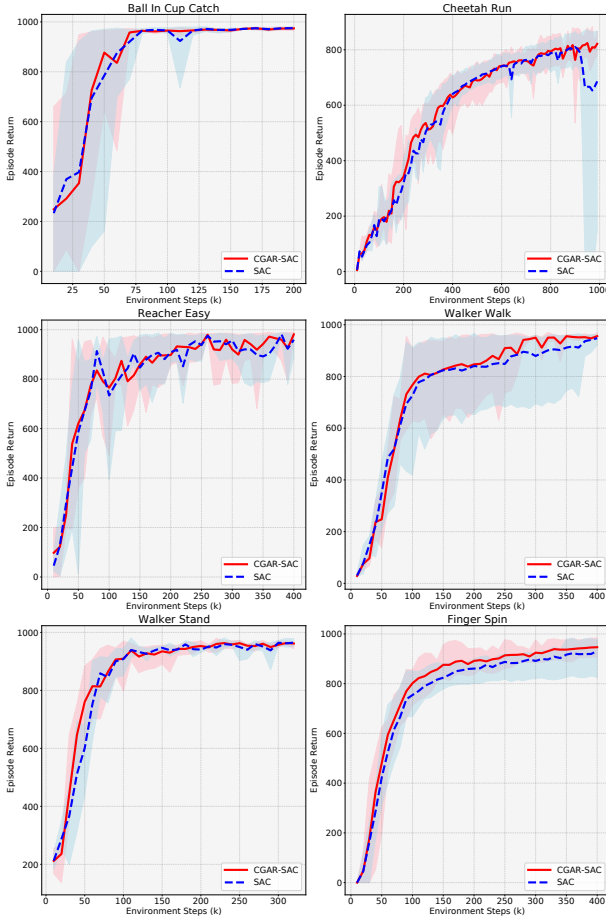


Fig. 3. Comparison results between CGAR-SAC and SAC.

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

- [1] D. Ye, G. Chen, W. Zhang, S. Chen, B. Yuan, B. Liu, J. Chen, Z. Liu, F. Qiu, H. Yu, Y. Yin, B. Shi, L. Wang, T. Shi, Q. Fu, W. Yang, L. Huang, and W. Liu, "Towards playing full MOBA games with deep reinforcement learning," in *NeurIPS 2020, December 6-12, 2020, virtual*.
- [2] N. Brown, A. Bakhtin, A. Lerer, and Q. Gong, "Combining deep reinforcement learning and search for imperfect-information games," in *NeurIPS 2020, December 6-12, 2020, virtual*.
- [3] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," 2011.
- [4] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, "Human-level control through deep reinforcement learning," *nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [5] T. Malloy, C. R. Sims, T. Klinger, M. Liu, M. Riemer, and G. Tesauro, "Capacity-limited decentralized actor-critic for multi-agent games," in *COG 2021, Copenhagen, Denmark, August 17-20, 2021*. IEEE, 2021, pp. 1–8.
- [6] R. Liang, Y. Zhu, Z. Tang, M. Yang, and X. Zhu, "Proximal policy optimization with elo-based opponent selection and combination with enhanced rolling horizon evolution algorithm," in *COG 2021, Copenhagen, Denmark, August 17-20, 2021*. IEEE, 2021, pp. 1–4.
- [7] J. Büttner and S. von Mammen, "Training a reinforcement learning agent based on XCS in a competitive snake environment," in *COG 2021, Copenhagen, Denmark, August 17-20, 2021*. IEEE, 2021, pp. 1–5.
- [8] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," in *ICLR*, 2016.
- [9] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *ICML*, 2018.
- [10] T. Haarnoja, H. Tang, P. Abbeel, and S. Levine, "Reinforcement learning with deep energy-based policies," in *ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. PMLR, 2017, pp. 1352–1361.
- [11] B. D. Ziebart, A. L. Maas, J. A. Bagnell, and A. K. Dey, "Maximum entropy inverse reinforcement learning," in *AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*. AAAI Press, pp. 1433–1438.
- [12] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. P. Lillicrap, and M. A. Riedmiller, "Deepmind control suite," *CoRR*, vol. abs/1801.00690, 2018.
- [13] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NIPS Deep Learning and Representation Learning Workshop*, 2015.
- [14] S. I. Mirzadeh, M. Farajtabar, A. Li, N. Levine, A. Matsukawa, and H. Ghasemzadeh, "Improved knowledge distillation via teacher assistant," in *AAAI 2020*, vol. 34, no. 04, 2020, pp. 5191–5198.
- [15] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *ICCV 2019*.
- [16] X. Deng and Z. Zhang, "Can students outperform teachers in knowledge distillation based model compression?"
- [17] Y. LeCun and C. Cortes, "MNIST handwritten digit database," 2010.
- [18] D. Yarats and I. Kostrikov, "Soft actor-critic (sac) implementation in pytorch," https://github.com/denisyarats/pytorch_sac.