Centralized Cooperative Exploration Policy for Continuous Control Tasks

Extended Abstract

Chao Li*
Institute of Automation, Chinese Academy of Sciences, China
lichao2021@ia.ac.cn

Chen Gong∗
Institute of Automation, Chinese Academy of Sciences, China
gongchen2020@ia.ac.cn

Xinwen Hou†
Institute of Automation, Chinese Academy of Sciences, Beijing, China
xinwen.hou@ia.ac.cn

Qiang He
University of Tubingen, Germany
qianghe97@gmail.com

Yu Liu
Institute of Automation, Chinese Academy of Sciences, Beijing, China
yu.liu@ia.ac.cn

ABSTRACT

Despite recent works making great progress in continuous control tasks, exploration in these tasks has remained insufficiently investigated. This paper proposes CCEP (Centralized Cooperative Exploration Policy), which utilizes estimation biases of value functions to contribute to the exploration capacity. CCEP keeps two value functions initialized with different parameters, and generates diverse policies with multiple exploration styles from a pair of value functions. In addition, a centralized policy framework ensures that CCEP achieves message delivery between multiple policies, further contributing to exploring the environment cooperatively. Extensive experimental results demonstrate that CCEP achieves higher exploration capacity. Empirical analysis shows diverse exploration styles in the learned policies by CCEP, reaping benefits in more exploration regions. Besides, the exploration capabilities of CCEP have been demonstrated to outperform current state-of-the-art methods on multiple continuous control tasks.

KEYWORDS

Cooperative Exploration; Continuous Control Tasks

ACM Reference Format:

1 INTRODUCTION

In DRL settings, an agent aims to learn an optimal policy to maximize its expected cumulative rewards through trial and error. It is essential that during the training phase, the agent should be encouraged to explore the environments and gather sufficient reward signals for well-training. Therefore, exploration has obsessed with a critical problem: submitting solutions too quickly without sufficient exploration, leading to getting stuck at local minima or even complete failure. Whereas existing exploration methods [1, 3, 6, 7, 11, 16] remain a problematic drawback – lacking diversity to explore. However, in massive situations, diverse styles of exploration are necessary. For instance, in chess games, players should perform different styles of policies to keep competitive.

Our insights originate from a non-trivial phenomenon during the critic update process: the different critic functions may have great differences even if they approximate the same target due to the function approximation error which accumulates to the estimation bias. Although many proposed methods are dedicated to reducing estimation bias [4, 5, 8, 9, 14, 15], this knotty problem is impossible to completely avoid. We raise the question of whether we can “transform an enemy into a friend” – utilizing estimation bias to

Figure 1: The workflow of CCEP Algorithm. The agent \( \pi \) interacts with the environment with diverse style cooperatively. A centralized policy with four different styles is learned from the multi-styled critics.

![Image of the workflow of CCEP Algorithm]

These two authors contributed equally.

†The corresponding author.


2454
Table 1: Max Average Return over 10 trials of 1 million time steps. The mean values have been listed. The maximum value for each task is bolded.

<table>
<thead>
<tr>
<th>Env</th>
<th>Ours</th>
<th>OAC</th>
<th>SAC</th>
<th>TD3</th>
<th>ICM</th>
<th>RND</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCheck</td>
<td>11945</td>
<td>9921</td>
<td>11129</td>
<td>9758</td>
<td>10085</td>
<td>10629</td>
</tr>
<tr>
<td>Hopper</td>
<td>3636</td>
<td>3364</td>
<td>3357</td>
<td>3479</td>
<td>3504</td>
<td>3419</td>
</tr>
<tr>
<td>Walker2d</td>
<td>4706</td>
<td>4458</td>
<td>4349</td>
<td>4229</td>
<td>4255</td>
<td>4197</td>
</tr>
<tr>
<td>Ant</td>
<td>5630</td>
<td>4519</td>
<td>5084</td>
<td>5142</td>
<td>5166</td>
<td>4990</td>
</tr>
<tr>
<td>Pusher</td>
<td>-21</td>
<td>-25</td>
<td>-20</td>
<td>-25</td>
<td>-23</td>
<td>-23</td>
</tr>
<tr>
<td>Humanoid</td>
<td>5325</td>
<td>5747</td>
<td>5523</td>
<td>5356</td>
<td>5374</td>
<td>5490</td>
</tr>
</tbody>
</table>

Figure 2: Measuring the exploration region. The points represent region explored by each method in 10 episodes. All the states get dimension reduction by the same t-SNE transformation for better visualization.

2 OUR METHOD

CCEP start by maintaining double randomly initialized value functions $Q_1$ and $Q_2$, with parameters $\theta_1$ and $\theta_2$ respectively and update the value function with TD3 [4]. But the two randomly initialized value functions potentially have different value estimations for a given state-action pair due to the accumulated function approximation error. This difference leads to the result that the two critics may give two different suggestions for the best action choice. These different criteria for a given state-action pair may lead to a different style of action choice. It is reasonable the estimation is radical if we choose the maximum value of the two to estimate and the estimation is conservative if we choose the minimum value of the two. Additionally, rather of constantly providing conservative or radical estimates for the current batch of state-action pairs, we would like to take into account random conservative or radical estimates. Thus, we involve four critics during the update of policy networks. There exists controversy among these critics, and the controversy can influence the performance of the policy learned. With four critics, we train a centralized cooperative policy to encourage multi-styled explorations through diverse value estimations. The target is to train multiple policies, with each policy targeting an individual value function. We express the policy function as $\pi(s, z)$, with state $s$ and latent variable $z$ as input. The latent variable $z$, which is a one-hot label in our method, identifies different policies. For a given latent variable $z$, the policy targets $z$-th value functions. With different latent variable $z$, the policy shows diverse styles due to the multi-styled targets. We make an experiment showing that there exists different exploration preferences for these policies. In the exploration procedure, we randomly sample latent variable $z$ and make decisions by policy $\pi(s, z)$. This approach enables diverse styles to be applied at each time step. The workflow of CCEP Algorithm is shown in Figure 1. Broadly speaking, our exploration policy has the following characteristics: Multi-styled, Centralized, and Cooperative.

3 EXPERIMENTS

To evaluate our method, we test our algorithm on the suit of MuJoCo [12] continuous control tasks. We show the max average return over 10 trials of 1 million time steps in Table 1. Further, We compare the exploration capabilities of CCEP with that of baselines [1, 2, 4, 7, 11] (Figure 2). For a fair comparison, these methods are trained in Ant-v3 with the same seed at half of the training process. In order to get reliable results, the states explored are gathered in 10 episodes with different seeds. We apply the same t-SNE [13] transformation to the states explored by all of the algorithms for better visualization. While there are great differences between the states explored by different algorithms, the result shows that our algorithm (red) explores a wider range of states involving that other algorithms has explored.
REFERENCES


