Goal-Conditioned On-Policy Reinforcement Learning

Xudong Gong[1,2] Dawei Feng[1,2] Kele Xu[1,2] Bo Ding[1,2] Huaimin Wang[1,2]

[1]College of Computer, National University of Defense Technology, Changsha, Hunan, China

[2]State Key Laboratory of Complex & Critical Software Environment, Changsha, Hunan, China

Outline

1. Motivation
2. Method
3. Experiments
4. Discussion

Motivation

Multi-Goal Problems

- Controlling robotic arms to grasp objects at any location on a table
- Operating fixed-wing UAVs to navigate towards any specified velocity vector
- ……

Goal-Conditioned Reinforcement Learning

Learns goal-conditioned behaviors that can achieve and generalize across a range of different goals

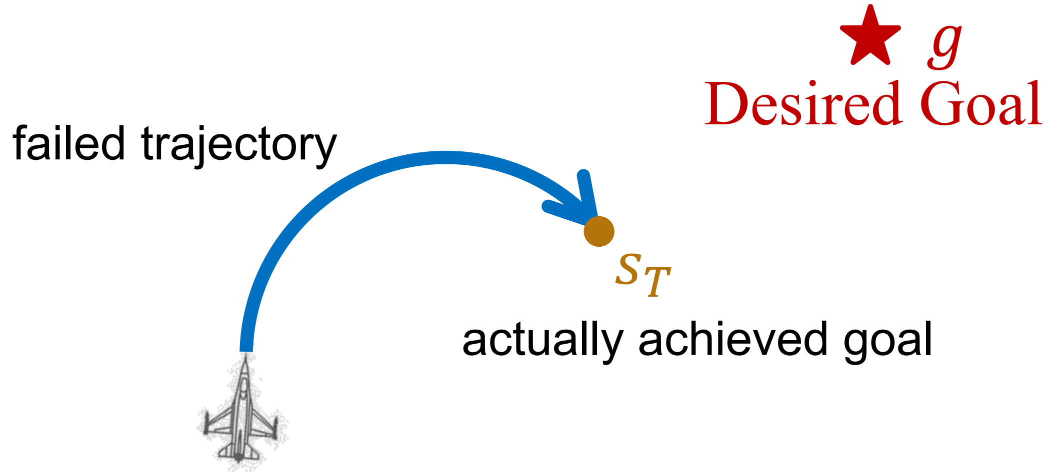
Challenge

- The additional goal space intensifies the complexity of exploration
- The goals used in sampling data and the order of these goals affects GCRL’s training efficiency and effectiveness

Mainstream methods

Based on HER[1] (Hindsight Experience Replay)

1. Relabel failed trajectories[1]



Relabel desired goal and recompute reward.

step	failed traj	relabeled traj
0	$(s_0, a_0, r(s_0, g))$	$(s_0, a_0, r(s_0, s_N))$
1	$(s_1, a_1, r(s_1, g))$	$(s_1, a_1, r(s_1, s_N))$
$\vdots$	$\vdots$	$\vdots$
N	$(s_N, a_N, r(s_N, g))$	$(s_N, a_N, r(s_N, s_N))$

- Failed trajectories contribute almost nothing to policy optimization.
- If the failed but actually achieved goal state s_T is considered as the desired goal, then this failed trajectory becomes a successful one, which can help optimize the policy.
- Besides the last state, any state in the trajectory can be mapped to a potential hindsight goal.

Assigning valuable use to failure experiences, alleviating the exploration challenge!

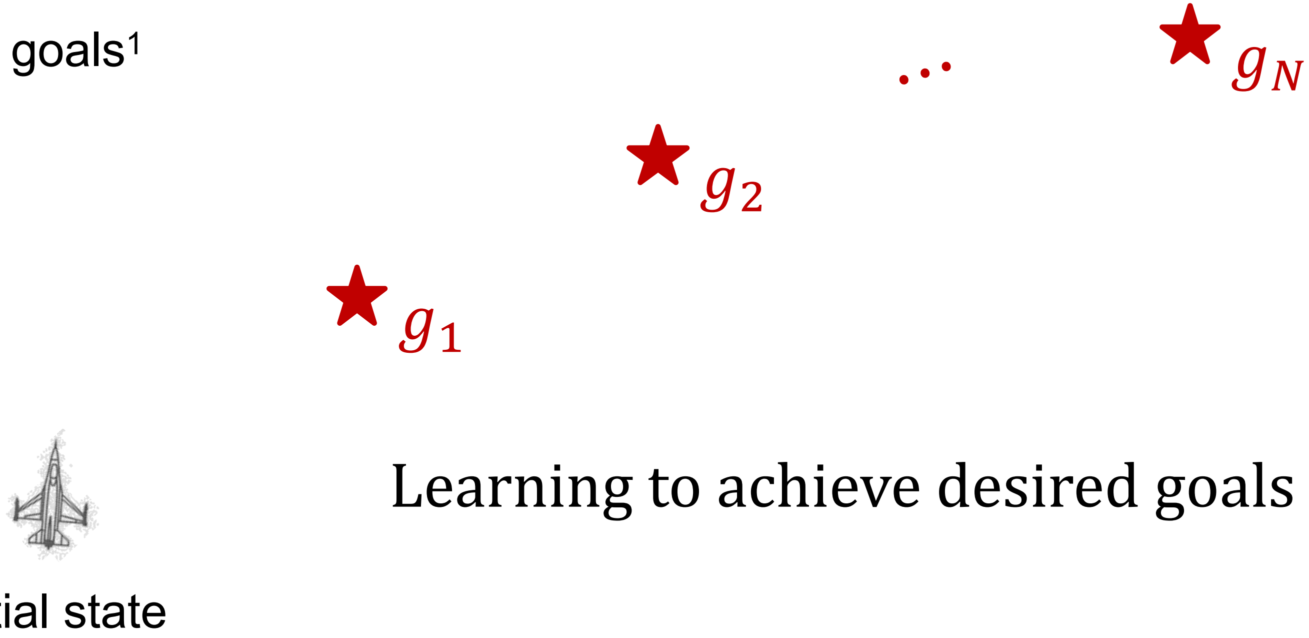
1. Andrychowicz M, Wolski F, Ray A, et al. Hindsight experience replay[J]. Advances in neural information processing systems, 2017, 30.

Motivation

Mainstream methods

Based on HER (Hindsight Experience Replay)

2. Arrange behavioral goals[1]



Learning to achieve desired goals in some orders

- ➢ Evaluate policy’s ability of achieving goals based on the data in the ER, and sample behavioral goals of appropriate difficulty for sampling training data

Further improve training efficiency by arranging behavioral goals!

1. Pitis S, Chan H, Zhao S, et al. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2020: 7750-7761.

Motivation

Mainstream methods

Based on HER (Hindsight Experience Replay)

- Relabel failure experience → alleviate the exploration challenge
- Evaluate policy's ability of achieving goals based on the data in the ER, and select behavioral goals of appropriate difficulty for sampling training data → further improve training efficiency by arranging behavioral goals

Algorithm 1 Unified Framework for Multi-goal Agents

function TRAIN($*args$):
  Alternate between collecting experience using ROLLOUT and optimizing the parameters using OPTIMIZE.

function ROLLOUT (policy π_{explore}, buffer $\mathcal{B}$, $*args$):
  $\boxed{\mathrm{g} \leftarrow \textsc{Select}(*args)}$
  $\mathrm{s}_0 \leftarrow$ initial state
  for t in $0 \ldots T-1$ **do**
    $\mathrm{a}_t, \mathrm{s}_{t+1} \leftarrow$ execute $\pi_{\text{explore}}(\mathrm{s}_t, \mathrm{g})$ in environment
    $\mathrm{r}_t \leftarrow \textsc{Reward}(\mathrm{s}_t, \mathrm{a}_t, \mathrm{s}_{t+1}, \mathrm{g})$
    Store $(\mathrm{s}_t, \mathrm{a}_t, \mathrm{s}_{t+1}, \mathrm{r}_t, \mathrm{g})$ in replay buffer $\mathcal{B}$

function OPTIMIZE (buffer $\mathcal{B}$, algorithm $\mathcal{A}$, parameters θ):
  Sample mini-batch $B = \{(\mathrm{s}, \mathrm{a}, \mathrm{s}', \mathrm{r}, \mathrm{g})_i\}_{i=1}^{N} \sim \mathcal{B}$
  $\boxed{B' \leftarrow \textsc{Relabel}(B, *args)}$
  Optimize θ using $\mathcal{A}$ (e.g., DDPG) and relabeled B'

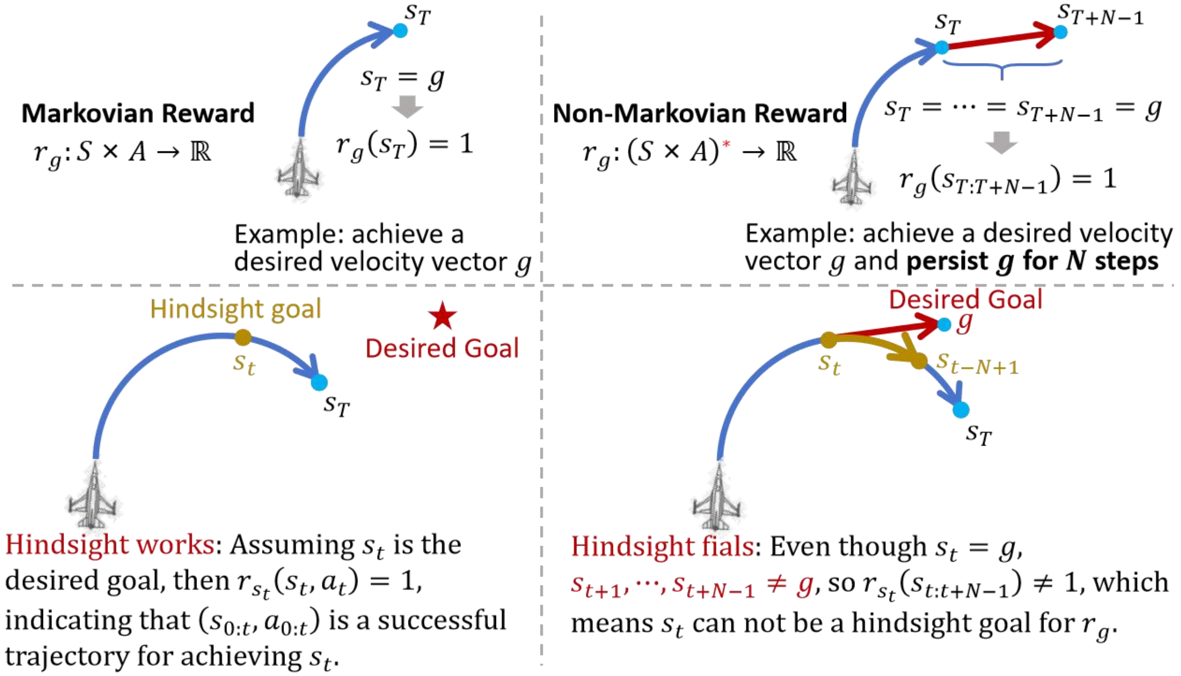
A general GCRL training framework[1]

1. Pitis S, Chan H, Zhao S, et al. Maximum entropy gain exploration for long horizon multi-goal reinforcement learning[C]//International Conference on Machine Learning. PMLR, 2020: 7750-7761.

Challenge of HER-based methods

- HER-based methods make an implicit assumption that the rewards **depends only on the current state** (Markovian Reward, MR). Under this assumption, any single state can potentially become a hindsight goal.

- However, when the computation of rewards **depends on multi-step states** (Non-Markovian Reward, NMR), a failed trajectory may not contain any state sequence that satisfies the reward.



Can a GCRL framework be proposed that does not rely on HER and can simultaneously address both MR and NMR problems?

Outline

1. Motivation
2. Method
3. Experiments
4. Discussion

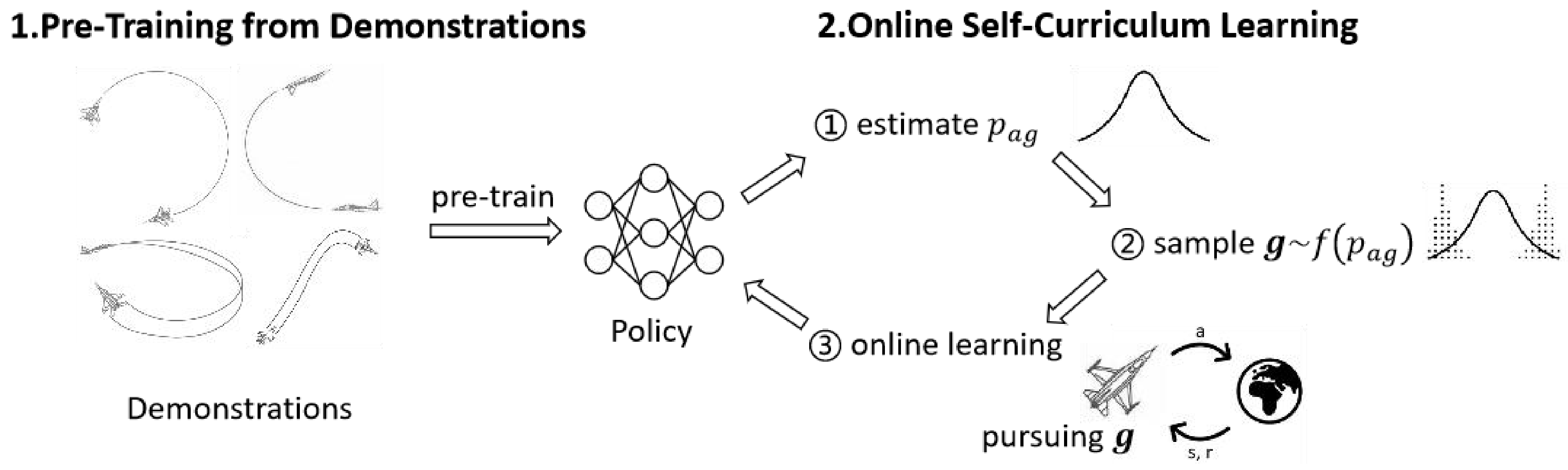
Goal-conditioned on-policy reinforcement learning (GCPO)

Insight: refer to the two successful designs of HER-based methods

HER-based methods		**GCPO**
Relabel failure experience ←	alleviate the exploration challenge	→ Pre-training from demonstrations to provide a behavioral prior for the policy
(1) Evaluate policy's goal-achieving ability **with the help of ER** (2) select appropriate behavioral goals (3) optimize policy with off-policy RL ←	further improve training efficiency by arranging behavioral goals	→ (1) evaluate policy periodically (2) estimate policy's goal achieving ability with **Off-Policy Evaluation (OPE)** method (3) select appropriate behavioral goals (4) optimize polify with KL-Regularized on-policy RL

Method-GCPO

General framework



1.Pre-training from demonstrations

pre-training provides the policy with an initial ability to achieve some of the desired goals, enhancing informative rewards during online learning.

2.Online self-curriculum learning

design an online self-curriculum learning mechanism that autonomously constructs a curriculum, generating behavioral goals that are incrementally more difficult than those the policy is currently capable of achieving.

A practical implementation

1.Pre-training from demonstrations

utilize Behavioral Cloning (BC) to pre-train policies

$$\mathcal{L}(\theta) = -\mathbb{E}_{(s,a)\sim\mathcal{D}_E}[\log \pi_\theta(a|s)]$$

2.Online self-curriculum learning

2.1 estimating the current policy's goal-achieving ability, p_{ag}

employ GMM to estimate p_{ag} with historical evaluation data

2.2 sampling progressively challenging behavioral goals

utilize inverse probability weighting, $[f_{MEGA}(p_{ag}, p_{dg})](g) = \frac{\frac{1}{p_{ag}(g)}}{\sum_{p'} \frac{1}{p_{ag}(g')}}$

2.3 Conducting online RL learning with behavioral goals

optimize policy with KL-regularized RL, $J_{kl}(\pi_\theta) = \mathbb{E}[\sum_t \gamma^t (r - \lambda log(\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_0}(a_t|s_t)}))]$

Algorithm 1 Goal-Conditioned Policy Optimization (GCPO) framework

Require: demonstrations $\mathcal{D}_E$, distribution of desired goal p_{dg}, goal weight discount factor κ, online evaluation budget N, probability transform function f

Ensure: $\pi_\theta(\cdot|s, g)$

Initialize goal-conditioned policy $\pi_\theta(\cdot|s, g)$, goal buffer B_g which stores tuples of achieved goals in evaluation and their corresponding weight (g, w_g)

pre-train $\pi_\theta(\cdot|s, g)$ by Eq. 1 ▷ pre-train policy

while Not converge **do**

  for all (g, w_g) in B_g **do**

    $w_g \leftarrow \kappa \cdot w_g$ ▷ decay weight of historically achieved goals

  end for

  sample N goals, $g_1, g_2, \ldots, g_N$ uniformly from p_{dg} ▷ online policy evaluation

  for all g in $g_1, g_2, \ldots, g_N$ **do**

    if π_θ finishes g successfully **then**

      add $(g, 1.0)$ to B_g

    end if

  end for

  estimate p_{ag} with GMM on B_g ▷ estimate p_{ag}

  $\mathcal{D} \leftarrow \varnothing$ ▷ roll-out samples

  while Not collect enough online samples **do**

    sample a goal g from Eq. 2

    sample a trajectory τ by π_θ on g

    append τ to $\mathcal{D}$

  end while

  update π_θ by Eq. 3 on $\mathcal{D}$ ▷ update policy

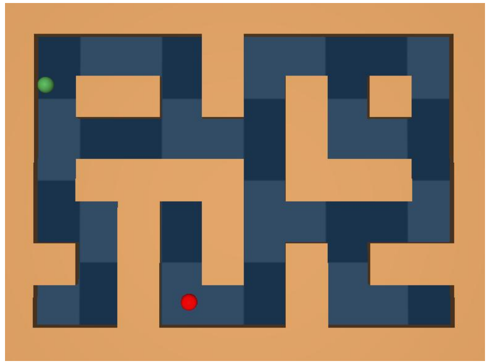
end while

Outline

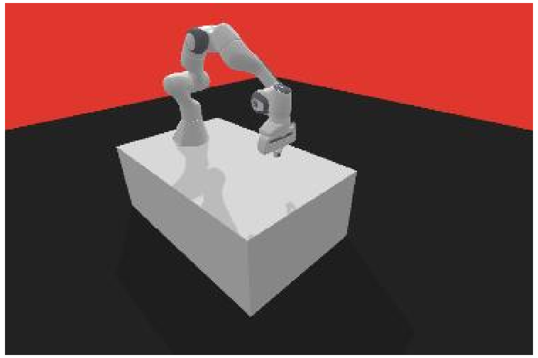
Experiments

Settings

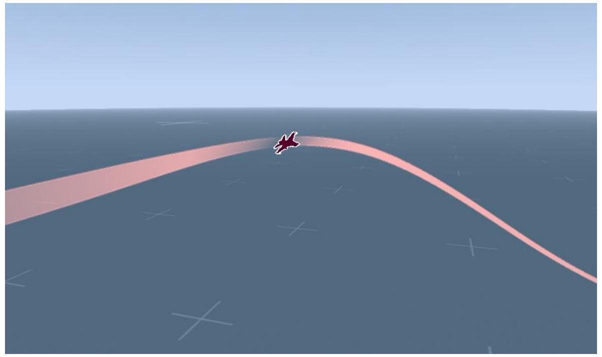
- Tasks



PointMaze



Reach



Fixed-wing velocity vector control

Experiments

Main Results

On PointMaze and Reach

Table 7: Comparison between GCPO and baselines on Reach and PointMaze. The mean and variance of % success rates are presented over 5 random seeds. Optimal values are highlighted in bold, and sub-optimal values are underlined.

Task	Reward	SAC + HER + MEGA	BC	GCPO
Reach	MR	**100.0±0.0**	<u>70.63±2.99</u>	**100.0±0.0**
	NMR	0.72±1.34	<u>10.52±11.70</u>	**80.26±17.01**
PointMaze	MR	**100.0±0.0**	75.96±5.34	<u>93.33±3.06</u>
	NMR	4.17±0.93	<u>22.8±3.71</u>	**47.50±8.06**

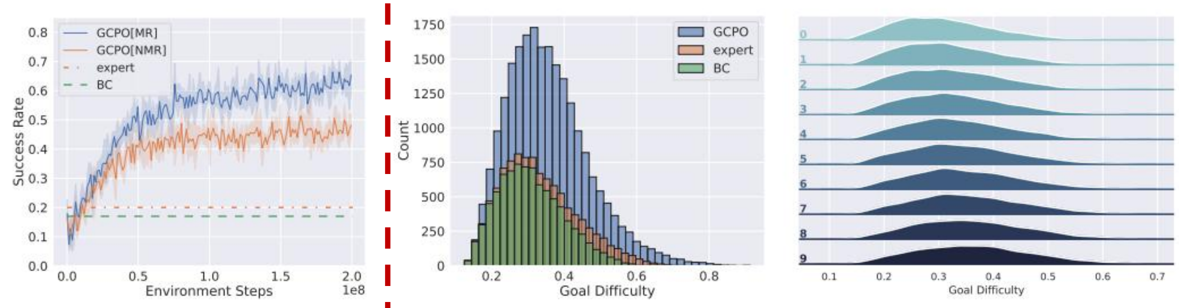
- under MR settings, GCPO exhibits similar performance to HER-based methods.
- under NMR settings, GCPO shows significantly superior performance than HER-based methods.

Main Results

On fixed-wing velocity vector control

Table 2: Comparison between GCPO and baselines on NMR. The mean and variance of % success rates are presented over 5 random seeds. Optimal values are highlighted in bold, and sub-optimal values are underlined.

Demonstration			SAC + HER + MEGA	BC	GCPO w/o pre-training	GCPO w/o self-curriculum	GCPO
notation	#traj	traj length					
$\mathcal{D}_E^0$	10264	281.83±149.48	8.32±1.86	17.08±0.57	0.04±0.03	31.28±8.97	**45.87±3.09**
$\mathcal{D}_E^1$	27021	119.64±47.55		36.54±1.97		43.49±3.85	**49.12±1.67**
$\mathcal{D}_E^2$	34952	115.76±45.65		41.79±0.44		51.28±2.07	**57.45±2.49**
$\mathcal{D}_E^3$	39835	116.56±47.62		42.77±1.35		53.51±3.18	**59.90±1.78**



(a) Success rate on MR and NMR (b) Histogram of achieved goals (c) Distribution of goals from self-curriculum during learning

Figure 3: Main results of GCPO. 'expert' refers to the demonstrator that generates demonstrations. 'BC' refers to the pre-trained policy. Results are derived from experiments across 5 random seeds. For sub-figure (a), expert and BC are both evaluated in the NMR setting. For sub-figure (c), the vertical axis represents the training progress, where $0, 1, \cdots, 9$ correspond to $10\%, 20\%, \cdots, 100\%$ of the training progress, respectively.

- **GCPO is applicable to both MR and NMR problems**
  - GCPO outperforms all baselines on NMR
  - The learning progression of GCPO for both NMR and MR shows that GCPO is effective in solving both types of problems

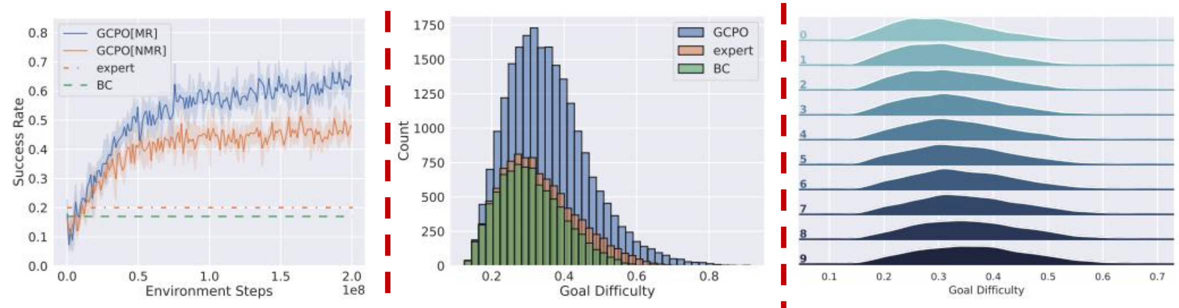
Experiments

Main Results

On fixed-wing velocity vector control

Table 2: Comparison between GCPO and baselines on NMR. The mean and variance of % success rates are presented over 5 random seeds. Optimal values are highlighted in bold, and sub-optimal values are underlined.

Demonstration notation	#traj	traj length	SAC + HER + MEGA	BC	GCPO w/o pre-training	GCPO w/o self-curriculum	GCPO
$\mathcal{D}_E^0$	10264	281.83±149.48	8.32±1.86	17.08±0.57	0.04±0.03	31.28±8.97	**45.87±3.09**
$\mathcal{D}_E^1$	27021	119.64±47.55		36.54±1.97		43.49±3.85	**49.12±1.67**
$\mathcal{D}_E^2$	34952	115.76±45.65		41.79±0.44		51.28±2.07	**57.45±2.49**
$\mathcal{D}_E^3$	39835	116.56±47.62		42.77±1.35		53.51±3.18	**59.90±1.78**



(a) Success rate on MR and NMR (b) Histogram of achieved goals (c) Distribution of goals from self-curriculum during learning

Figure 3: Main results of GCPO. 'expert' refers to the demonstrator that generates demonstrations. 'BC' refers to the pre-trained policy. Results are derived from experiments across 5 random seeds. For sub-figure (a), expert and BC are both evaluated in the NMR setting. For sub-figure (c), the vertical axis represents the training progress, where $0, 1, \cdots, 9$ correspond to $10\%, 20\%, \cdots, 100\%$ of the training progress, respectively.

- **Pre-training is crucial for the success of GCPO**
  - without pre-training, GCPO struggles to learn meaningful skills
  - even with a pre-trained policy that initially exhibits inferior performance compared to the demonstrator, GCPO's online self-curriculum learning facilitates significant improvement in the policy's performance, surpassing that of the demonstrator

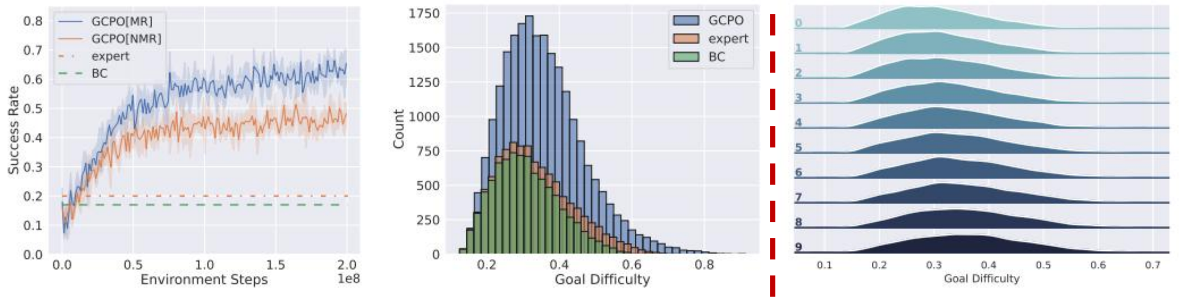
Experiments

Main Results

On fixed-wing velocity vector control

Table 2: Comparison between GCPO and baselines on NMR. The mean and variance of % success rates are presented over 5 random seeds. Optimal values are highlighted in bold, and sub-optimal values are underlined.

Demonstration			SAC + HER + MEGA	BC	GCPO w/o pre-training	GCPO w/o self-curriculum	GCPO
notation	#traj	traj length					
$\mathcal{D}_E^0$	10264	281.83±149.48	8.32±1.86	17.08±0.57	0.04±0.03	31.28±8.97	**45.87±3.09**
$\mathcal{D}_E^1$	27021	119.64±47.55		36.54±1.97		43.49±3.85	**49.12±1.67**
$\mathcal{D}_E^2$	34952	115.76±45.65		41.79±0.44		51.28±2.07	**57.45±2.49**
$\mathcal{D}_E^3$	39835	116.56±47.62		42.77±1.35		53.51±3.18	**59.90±1.78**



(a) Success rate on MR and NMR (b) Histogram of achieved goals (c) Distribution of goals from self-curriculum during learning

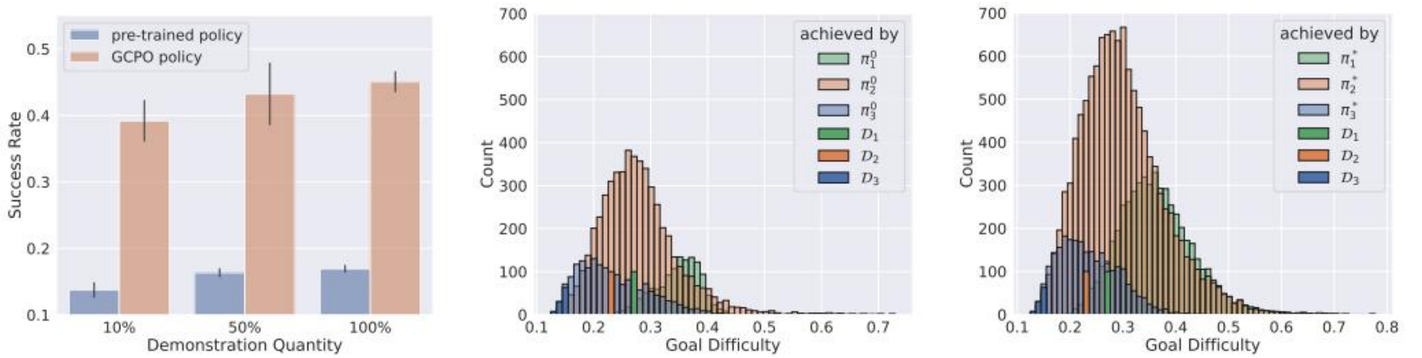
Figure 3: Main results of GCPO. 'expert' refers to the demonstrator that generates demonstrations. 'BC' refers to the pre-trained policy. Results are derived from experiments across 5 random seeds. For sub-figure (a), expert and BC are both evaluated in the NMR setting. For sub-figure (c), the vertical axis represents the training progress, where $0, 1, \cdots, 9$ correspond to $10\%, 20\%, \cdots, 100\%$ of the training progress, respectively.

- **Online self-curriculum facilitates the mastery of challenging goals**
  - the application of self-curriculum within GCPO leads to an average 8.2% increase in policy performance compared to its absence
  - online self curriculum mechanism systematically introduces more difficult goals into the learning progression as the policy gains proficiency

Experiments

Ablation Studies

1. Ablation on Quantity of Demonstrations
2. Ablation on Goal Distribution of Demonstrations



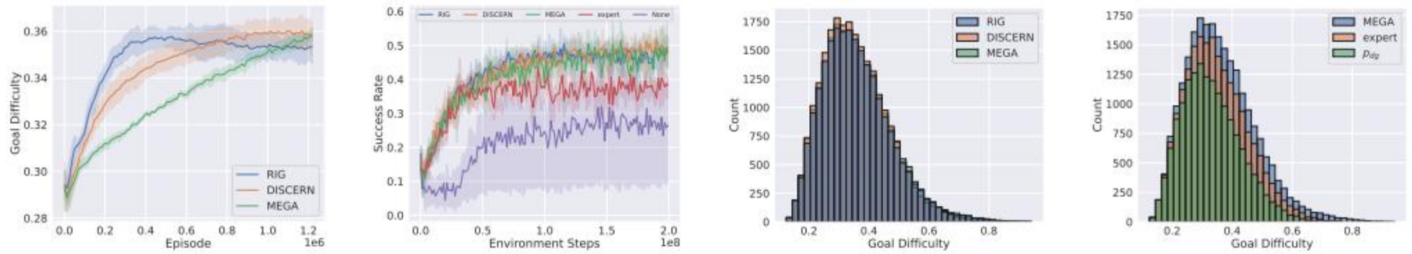
(a) Success rate of GCPO with different demonstration quantity (b) Histogram of achieved goals of the pre-trained policy (c) Histogram of achieved goals of GCPO policy

Figure 4: The influence of demonstration quantity and the distribution of goals covered by demonstrations on GCPO. $\mathcal{D}_1, \mathcal{D}_2, \mathcal{D}_3$ represent sets of demonstrations that are difficult, medium, and easy, respectively. The pre-trained policies obtained from $\mathcal{D}_1$, $\mathcal{D}_2$, and $\mathcal{D}_3$ are denoted as π_1^0, π_2^0, and π_3^0, respectively. The corresponding GCPO policies are denoted as π_1^*, π_2^*, and π_3^*, respectively. Results are derived from experiments across 5 random seeds.

- An increase in the quantity of demonstrations can enhance the performance of GCPO, yet the marginal gains diminish as the quantity of demonstrations grows
- when preparing demonstrations for GCPO, it is preferable to sample goals and generate demonstrations as closely as possible to the desired goal distribution p_{dg}

Ablation Studies

3. Comparison on Different Self-Curriculum Methods



(a) Difficulty of goals sampled by self-curriculum methods

(b) Success rate of self-curriculum and non-curriculum methods

(c) Histogram of achieved goals of different self-curriculum methods

(d) Histogram of achieved goals of MEGA and non-curriculum methods

Figure 5: Analysis of the influence of different self-curriculum methods on the learning progression of GCPO, as well as a comparison between self-curriculum and non-curriculum methods. ‘expert’ and ‘None’ are two non-curriculum methods, where ‘expert’ refers to sampling goals from those that the demonstrator can achieve, and ‘None’ signifies directly sampling from p_{dg}. Results are derived from experiments across 5 random seeds.

- Fig 5(a) suggests that different self-curriculum methods exhibit distinctly different learning progressions
- Figs 5(b) and 5(c) show that there is no significant difference in performance between different self-curriculum methods, whether in the learning progression or in the final policy
- Figs 5(b) and 5(d) show that self-curriculum methods outperform non-curriculum methods in both the learning progression and the final policy performance

Outline

1. Motivation
2. Method
3. Experiments
4. Discussion

Discussion

Our Contributions

- We propose an on-policy goal-conditioned reinforcement learning framework, GCPO, designed to address the limitations of existing methods in solving non-Markovian reward (NMR) problems.
- We demonstrate the effectiveness of GCPO in handling both Markovian reward (MR) and NMR problems through experimental evaluation.

Limitations

- In the implementation of the two components within GCPO, we employ relatively simple methods, such as behavioral cloning and Gaussian mixture model. Whether the use of alternative methods could lead to more efficient learning and better-performing policies is yet to be further validated
- Under the sparse reward setting, the successful training of GCPO relies on the pre-trained policy possessing a certain level of goal-achieving capability. Otherwise, if the policy achieves nothing, it becomes ineffective in establishing a self-curriculum.
- The specific implementation of GCPO has not explicitly incorporated components that are specifically designed to handle NMR problems. It is not clear whether integrating the most advanced methods for handling NMR problems within GCPO would lead to a more effective resolution

Thanks for watching!

- Code is available at https://github.com/GongXudong/GCPO
- Happy to answer any questions by email:

gongxudong_cs@aliyun.com davyfeng.c@qq.com