



ICML

International Conference
On Machine Learning



Code

Improving the Continuity of Goal-Achievement Ability via Policy Self-Regularization for Goal-Conditioned Reinforcement Learning


Xudong Gong^{1,2} Sen Yang¹ Dawei Feng^{1,2} Kele Xu^{1,2}
Bo Ding^{1,2} Huaimin Wang^{1,2} Yong Dou¹

¹College of Computer Science and Technology, National University of Defense Technology, Changsha, Hunan, China

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

1. Motivation
2. Margin-Based Policy Self-Regularization
3. Experiments
4. Discussion

Multi-Goal Problems

Markov Decision Process (MDP), $\mathcal{M} := \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma \rangle$  **Goal-Augmented MDP**, $\mathcal{M} := \langle \mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma, \mathcal{G}, p_{dg}, \phi \rangle$

- \mathcal{G} : desired goal space
- p_{dg} : desired goal distribution
- ϕ : mapping function that maps state to a specific goal

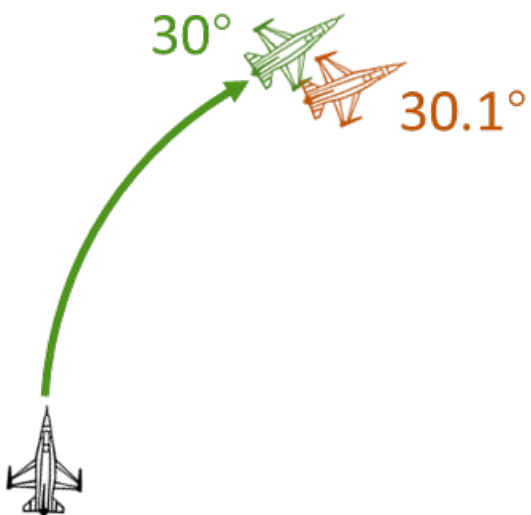
E.g., a UAV must be capable of achieving not only the left-side goal but also the right-side goal

Goal-Conditioned Reinforcement Learning (GCRL)

Learns goal-conditioned policy, $\pi: \mathcal{S} \times \mathcal{G} \rightarrow \Delta(\mathcal{A})$, by maximizing discounted return over desired goal distribution,

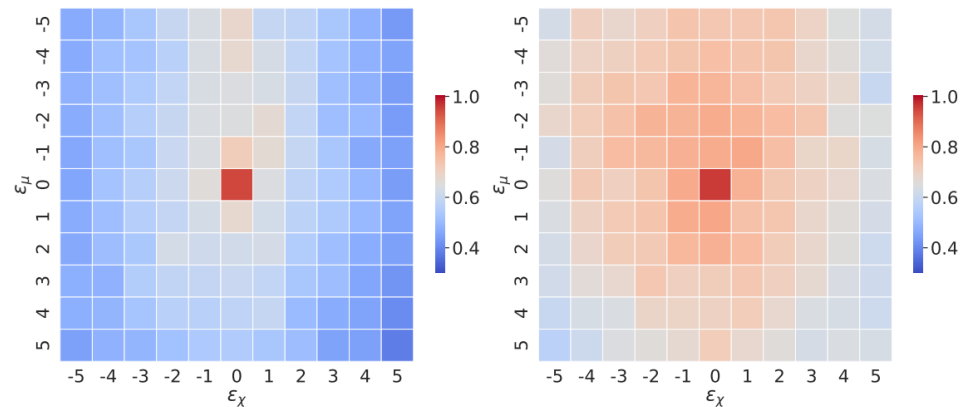
$$\mathbb{E}_{a_t \sim \pi, s_{t+1} \sim \mathcal{T}, g \sim p_{dg}} \left[\sum_t \gamma^t r(s_t, a_t, g) \right]$$

Challenge



Research Question

Suppose the aircraft can turn right 30° . Then, can it turn right 30.1° (error threshold $\delta = 1$)?



(a) GC-PPO (baseline)

(b) MSR-GC-PPO (ours)

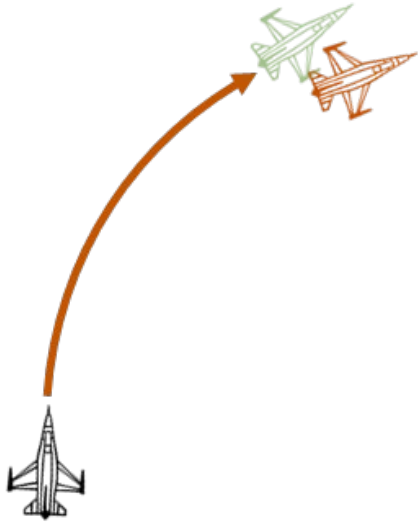
Experimental Validation

Left: for general GCRL algorithms, the success rate for goals in the vicinity of an achievable goal typically falls below 60%

Right: our method significantly improves the success rate to well above 60%

1. Motivation
- 2. Margin-Based Policy Self-Regularization**
3. Experiments
4. Discussion

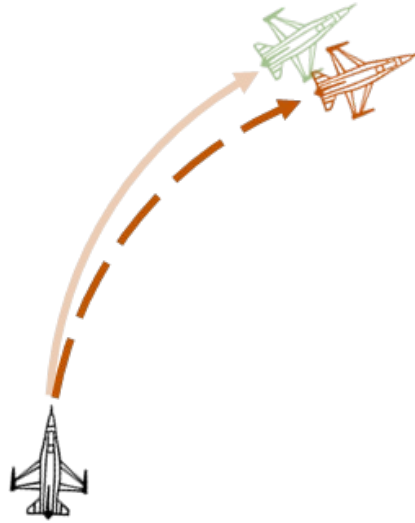
Insights



(a) Directly reusing the policy for turning right 30° can achieve 30.1° and obtain a guaranteed return.



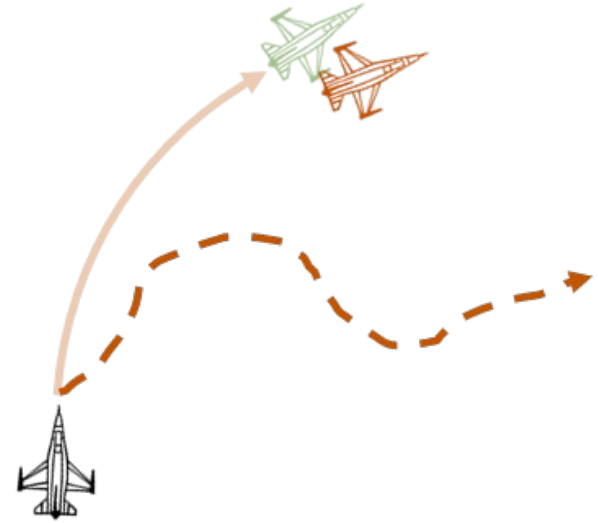
Theory 3.1, Corollary 3.2



(b) It is necessary to change the reused policy in order to obtain higher returns.



Theory 3.3



(c) If the policy differs too much, it may lead to failure.



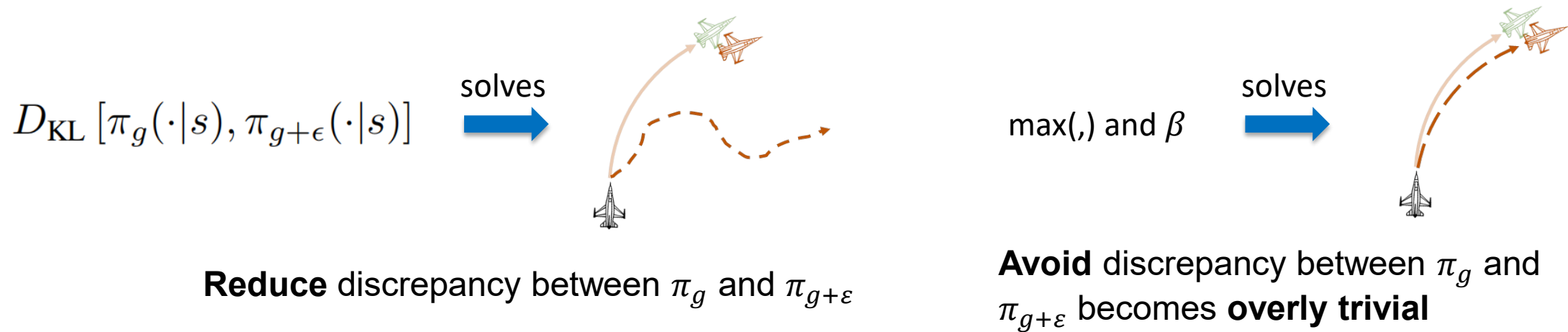
Theory 3.4, Corollary 3.5, Corollary 3.6

Method

Introduce an auxiliary loss to regularize the policy divergence between g and $g + \varepsilon$,

$$L(\pi) = L_{\text{RL}}(\pi) + \lambda L_{\text{MSR}}(\pi)$$

$$L_{\text{MSR}}(\pi) = \max\left\{\mathbb{E}_{\substack{s \sim d_{\pi_g} \\ \epsilon \sim (-\epsilon', \epsilon')}} D_{\text{KL}}[\pi_g(\cdot|s), \pi_{g+\epsilon}(\cdot|s)] - \beta, 0\right\}$$

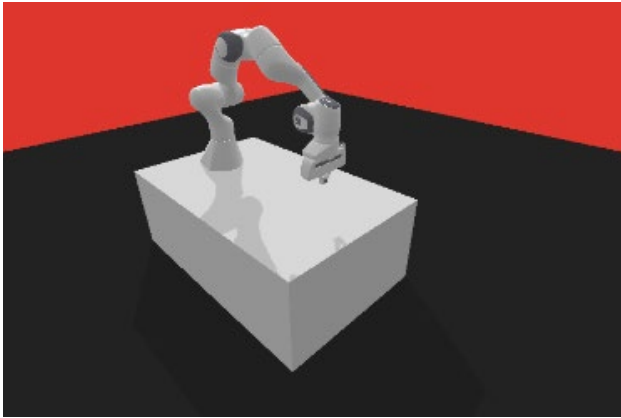


Theoretical analysis: refer to Section 3 and Appendix A of our manuscript.

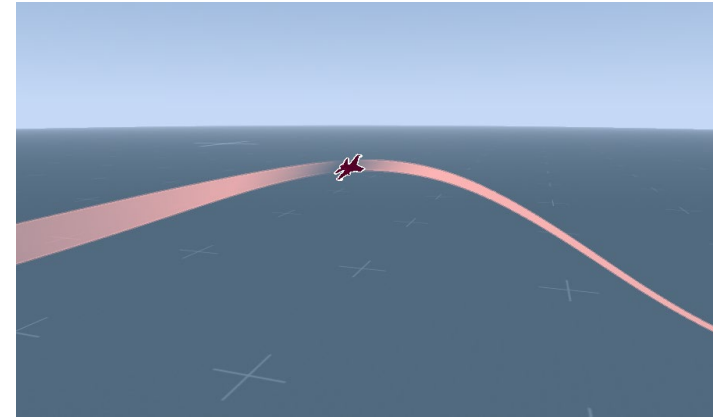
1. Motivation
2. Margin-Based Policy Self-Regularization
- 3. Experiments**
4. Discussion

Settings

- 3 tasks



Panda **Reach** and **Push**



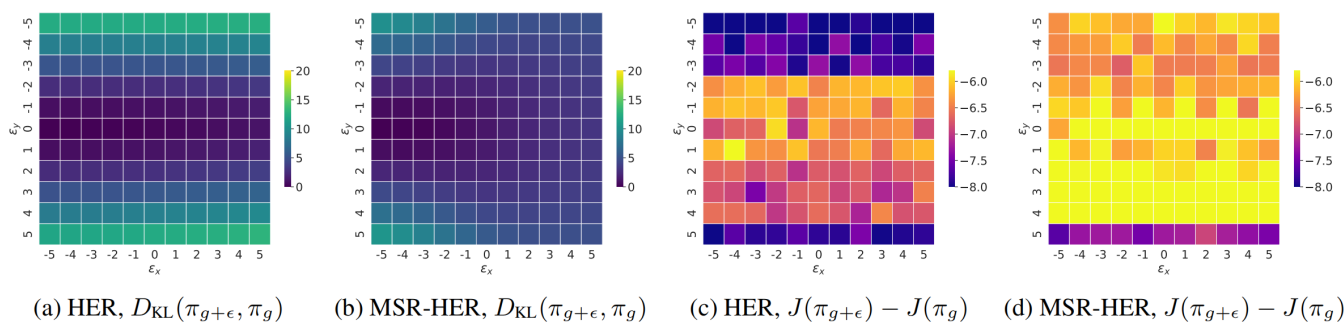
Fixed-wing UAV **velocity vector control**

- 3 baselines + 3 MSR algorithms

Baselines	MSR algorithms
GC-PPO	MSR-GC-PPO
GC-SAC	MSR-GC-SAC
HER	MSR-HER

Main Results

- MSR effectively reduces the discrepancy between policies for adjacent desired goals
- MSR also effectively diminishes the return gap between policies for adjacent desired goals
- MSR enhances the cumulative rewards of the policy



A micro view

(a) Reach			
Algorithms	$D_{KL}(\pi_{g+\epsilon}, \pi_g)$	$J(\pi_{g+\epsilon}) - J(\pi_g)$	$J(\pi)$
GC-SAC	0.51±1.28	-0.89±0.68	-12.03±0.63
MSR-GC-SAC	0.25±0.03	0.00±0.00	-0.85±0.02
HER	0.71±0.25	-0.01±0.01	-1.16±0.10
MSR-HER	0.31±0.04	-0.01±0.01	-0.96±0.08
GC-PPO	0.01±0.03	-5.29±2.11	-24.89±1.62
MSR-GC-PPO	2.85±0.72	-0.01±0.02	-1.03±0.15
(b) Push			
Algorithms	$D_{KL}(\pi_{g+\epsilon}, \pi_g)$	$J(\pi_{g+\epsilon}) - J(\pi_g)$	$J(\pi)$
GC-SAC	9.69±6.81	-7.58±3.05	-25.50±8.00
MSR-GC-SAC	4.27±1.42	-6.48±4.14	-23.80±6.76
HER	5.62±0.57	-7.05±2.28	-18.00±2.24
MSR-HER	3.80±0.31	-7.02±1.33	-16.31±1.92
GC-PPO	0.80±0.42	-8.21±1.82	-24.76±2.99
MSR-GC-PPO	0.61±0.17	-7.08±1.78	-23.08±5.06
(c) VVC			
Algorithms	$D_{KL}(\pi_{g+\epsilon}, \pi_g)$	$J(\pi_{g+\epsilon}) - J(\pi_g)$	$J(\pi)$
GC-SAC	0.37±0.12	-22.00±5.06	-138.20±14.16
MSR-GC-SAC	0.33±0.09	-20.94±7.62	-132.09±8.06
HER	0.65±0.22	-7.91±3.22	-69.21±12.87
MSR-HER	0.58±0.16	-7.52±3.15	-64.73±13.55
GC-PPO	0.08±0.08	-44.50±12.09	-169.04±25.57
MSR-GC-PPO	0.19±0.20	-28.89±14.75	-146.64±28.05

An overall view

More results: refer to Section 4 and Appendix D our manuscript.

1. Motivation
2. Margin-Based Policy Self-Regularization
3. Experiments
- 4. Discussion**

Our Contributions

- We assess multiple prevalent GCRL algorithms across a range of tasks, substantiating the prevalent challenge of discontinuity in the goal-achievement capabilities of GCRL algorithms.
- We conduct a theoretical analysis to elucidate the reasons for the discontinuity in goal-achievement capabilities of GCRL algorithms. And design a margin-based policy self-regularization method.
- We conduct systematic evaluations on two robotic arm control tasks and a fixed-wing aircraft control task.

Future Work

- Exploring the scalability of our method to even more complex tasks and environments, as well as investigating its applicability to more GCRL algorithms, remain important directions.
- Exploring methods that address the discontinuity in goal-achievement capabilities in the worst-case scenarios.

Thanks for watching!

- Code is available at:
 - <https://github.com/GongXudong/fly-craft-examples>



Code

- Happy to answer any questions by email:

`gongxudong_cs@aliyun.com`

`davyfeng.c@qq.com`