Iterative Regularized Policy Optimization with Imperfect Demonstrations

Xudong Gong[1,2] Dawei Feng[1,2] Kele Xu[1,2] Yuanzhao Zhai[1,2]

Chengkang Yao[3] Weijia Wang[3] 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

[3]Flight Automatic Control Research Institute, AVIC, Xian, Shanxi, China

Outline

Motivation

Imitation Learning

is a set of learning techniques which use demonstration datasets to directly learn policies.

- advantage: sample efficiency
- disadvantage: depends on the availability of a substantial number of high-quality demonstrations

Challenge

How to learn superior policies from imperfect demonstrations?

Motivation

A method to learn from imperfect demonstrations

fine-tune policy with online KL-regularized Reinforcement Learning

1. learning policy By IL with imperfect $\mathcal{D}_E$:

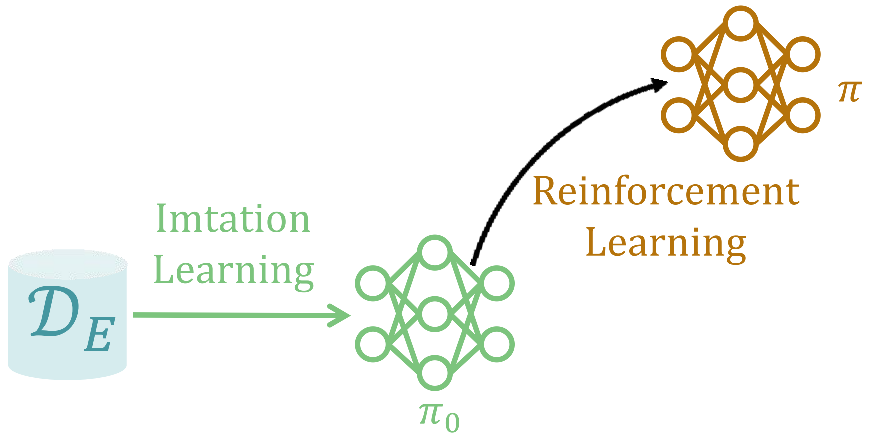
$$\pi_0 \leftarrow \underset{\pi}{\operatorname{argmin}} - \mathbb{E}_{(s,a)\sim\mathcal{D}_E}[\log \pi(a|s)]$$

2. initialize the policy to be learned online with π_0:

$$\pi \leftarrow \pi_0$$

3. as the optimization objective transitions from maximizing the likelihood of expert actions to maximizing cumulative rewards, to stablize the training process, KL-regularized RL is employed to optimize the policy online:

$$\pi \leftarrow \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}\left[\sum_{t=0}^{H-1} \gamma^t\big(r(s_t, a_t)\big)\right] - \lambda D_{KL}(\pi, \pi_0)$$

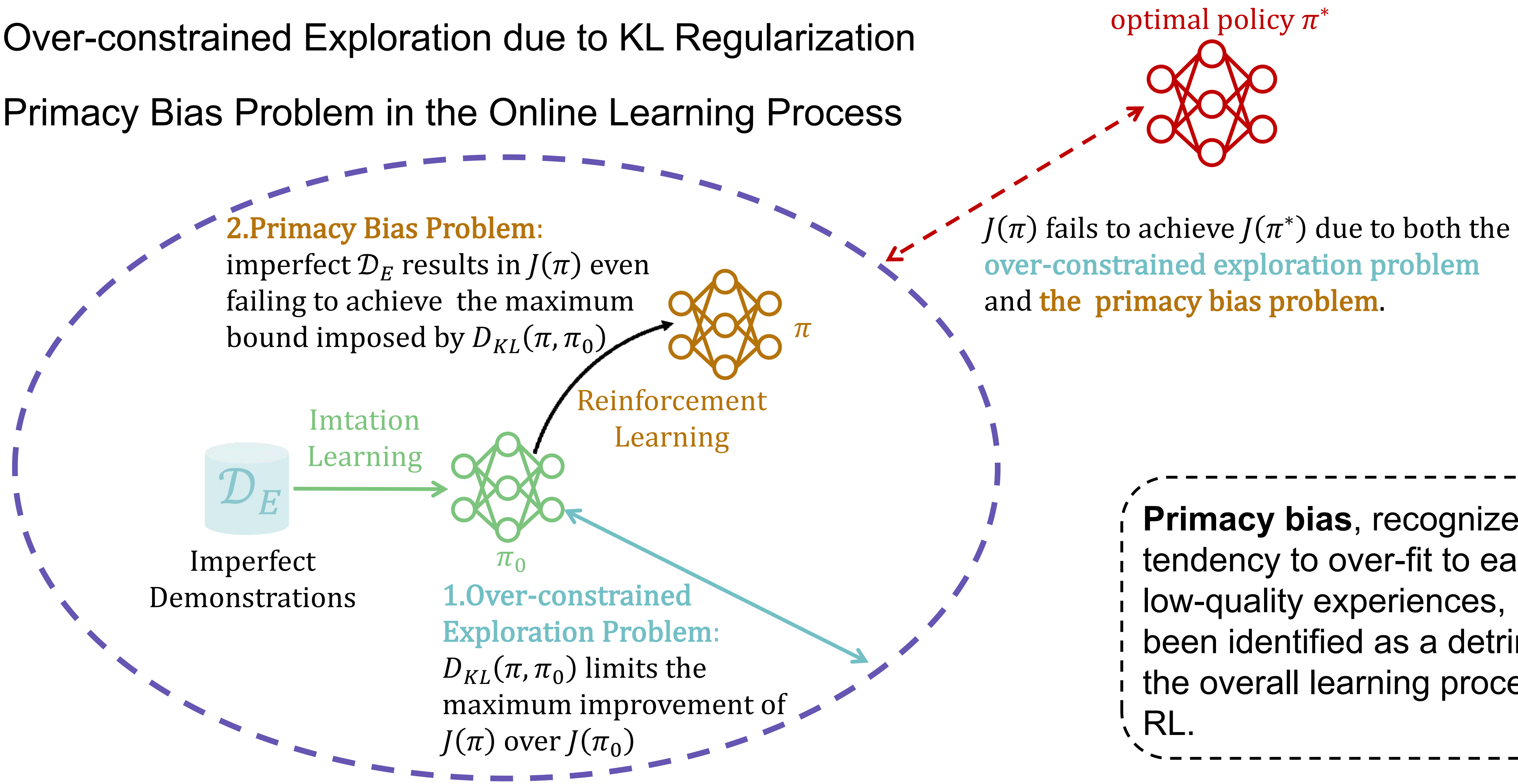


Motivation

Problems

1. Over-constrained Exploration due to KL Regularization

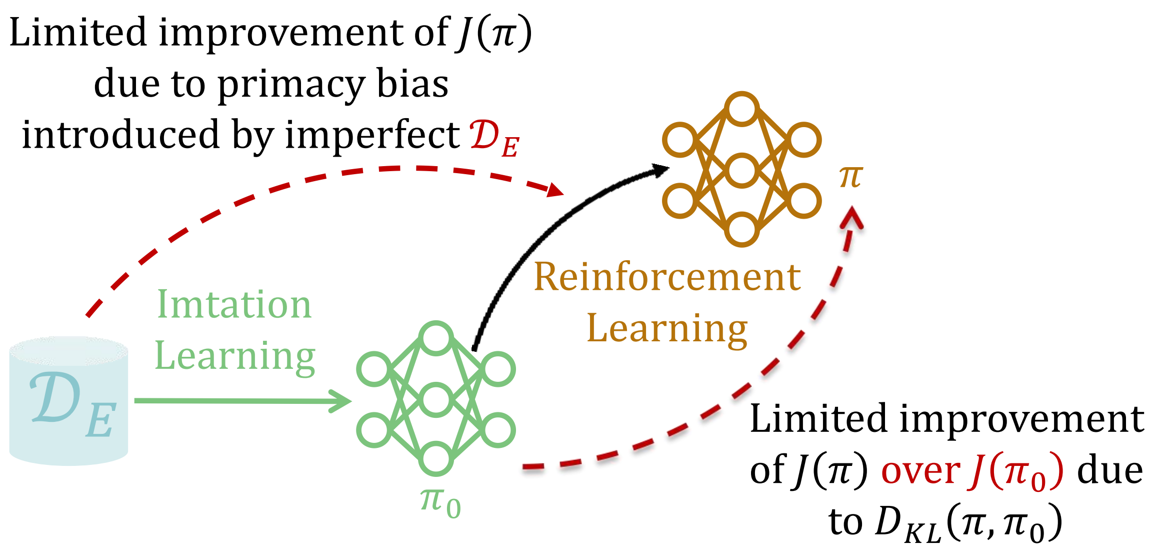
2. Primacy Bias Problem in the Online Learning Process


optimal policy π^*
$J(\pi)$ fails to achieve $J(\pi^*)$ due to both the over-constrained exploration problem and the primacy bias problem.
2.Primacy Bias Problem:
imperfect $\mathcal{D}_E$ results in $J(\pi)$ even failing to achieve the maximum bound imposed by $D_{KL}(\pi, \pi_0)$
π
Reinforcement Learning
Imtation Learning
$\mathcal{D}_E$
Imperfect Demonstrations
π_0
1.Over-constrained Exploration Problem:
$D_{KL}(\pi, \pi_0)$ limits the maximum improvement of $J(\pi)$ over $J(\pi_0)$

Primacy bias, recognized as a tendency to over-fit to early or low-quality experiences, has been identified as a detriment to the overall learning process in RL.

Motivation

Core Ideas

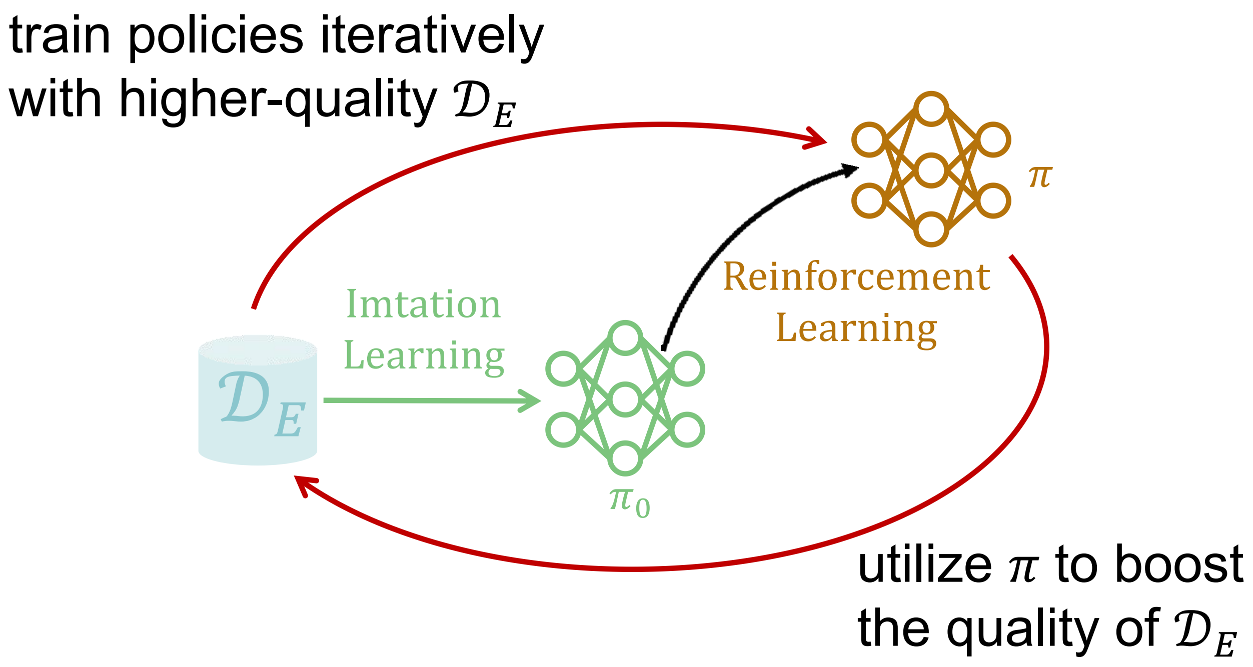


Limited improvement of $J(\pi)$ due to primacy bias introduced by imperfect $\mathcal{D}_E$
$\mathcal{D}_E$
Imtation Learning
π_0
Reinforcement Learning
π
Limited improvement of $J(\pi)$ over $J(\pi_0)$ due to $D_{KL}(\pi, \pi_0)$

Both the two problems are rooted in the imperfect demonstrations $\mathcal{D}_E$.

Motivation

Core Ideas


train policies iteratively with higher-quality $\mathcal{D}_E$
π
Reinforcement Learning
Imtation Learning
$\mathcal{D}_E$
π_0
utilize π to boost the quality of $\mathcal{D}_E$

Improve the quantity and quality of $\mathcal{D}_E$: utilize the policy learned online as the demonstrator for the subsequent training

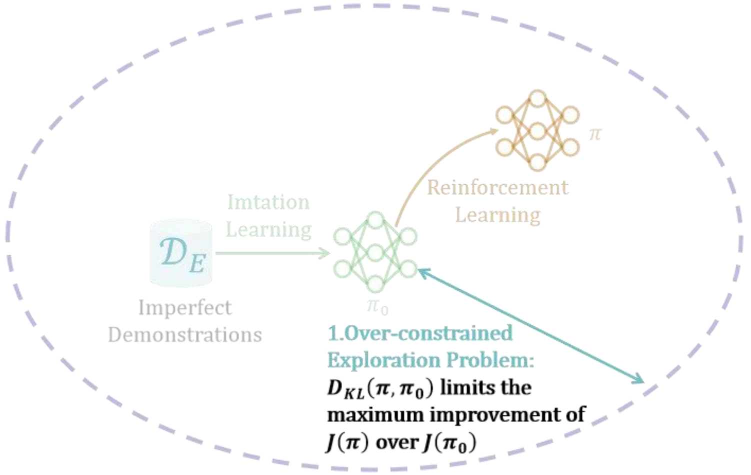
- **Demonstration boosting**: utilizes demonstrations generated by the online-learned policy to consistently enhance demonstration quality
- **Iterative training**: trains policies with IL and RL iteratively with higher-quality $\mathcal{D}_E$

Outline

The Over-Constrained Exploration Problem

Theorem 4.2. Given a finite-horizon MDP $\langle S, A, T, r, \gamma, H \rangle$, in which the reward function depends only on states $r : S \rightarrow R$, and a demonstrator π_E, for any policy π, the difference in the average of expected return with π_E is bounded by the KL divergence between π and π_E:

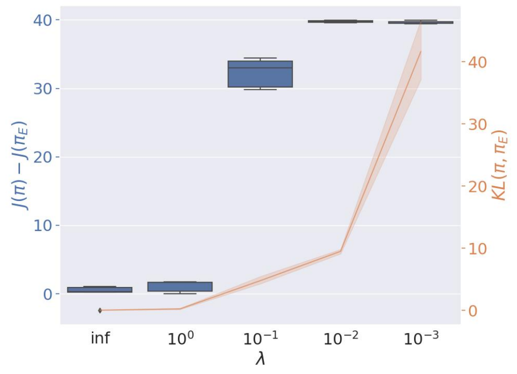
$$|J(\pi) - J(\pi_E)| \leq \max_s |r(s)| \sqrt{2HD_{KL}^{s\sim d_\pi}(\pi(\cdot|s), \pi_E(\cdot|s))}$$


$\mathcal{D}_E$
Imperfect Demonstrations
Imtation Learning
π_0
Reinforcement Learning
π
1.Over-constrained Exploration Problem:
$D_{KL}(\pi, \pi_0)$ limits the maximum improvement of $J(\pi)$ over $J(\pi_0)$

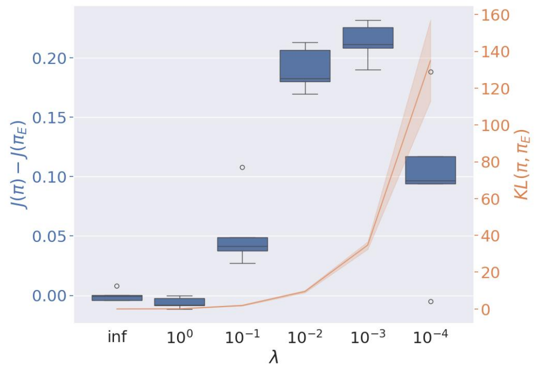
Theorem 4.2 implies that the KL divergence constraint between policies bounds the upper limit of policy returns.

Experimental verification:

Stronger regularization λ results in smaller KL values, and these smaller KL values are highly correlated with lower relative returns $|J(\pi) - J(\pi_E)|$.



Reach



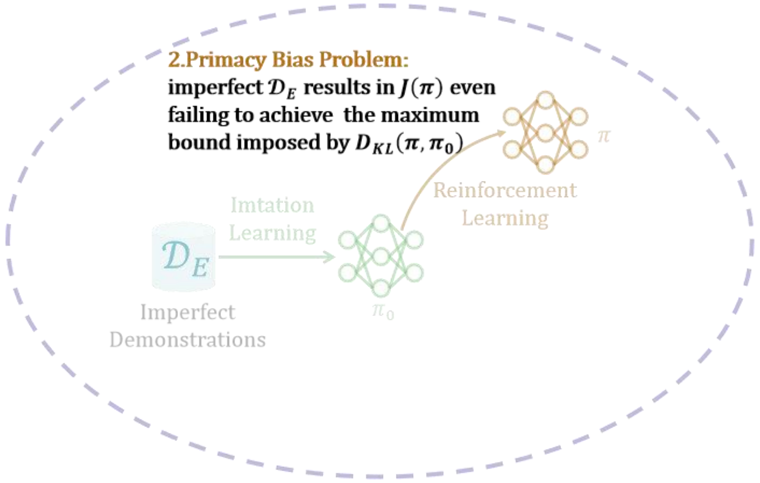
Fixed-wing attitude control

Method

The Primacy Bias Problem

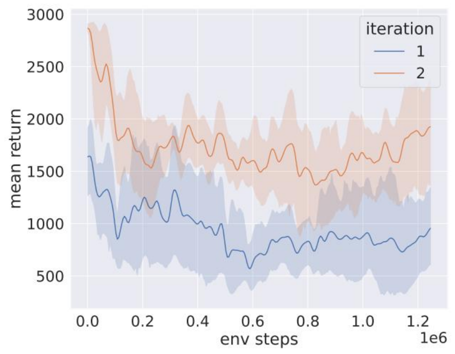
Primacy Bias originally refers to a tendency to overfit early interactions with the environment preventing the agent from improving its behavior on subsequent experiences[1].

However, we find that during the learning of offline IL to online RL, although only IL directly utilizes imperfect demonstrations, these imperfect demonstrations can still indirectly affect the learning of online RL in the form of a primacy bias.

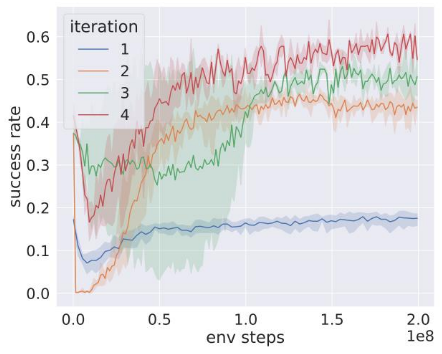

2.Primacy Bias Problem:
imperfect $\mathcal{D}_E$ results in $J(\pi)$ even failing to achieve the maximum bound imposed by $D_{KL}(\pi, \pi_0)$
π
Reinforcement Learning
Imtation Learning
$\mathcal{D}_E$
π_0
Imperfect Demonstrations

experimental verification：

Lower-quality demonstrations correspond to inferior offline and online learning policies. This observation implies that lower-quality demonstrations induce the primacy bias problem during online learning.



Hopper



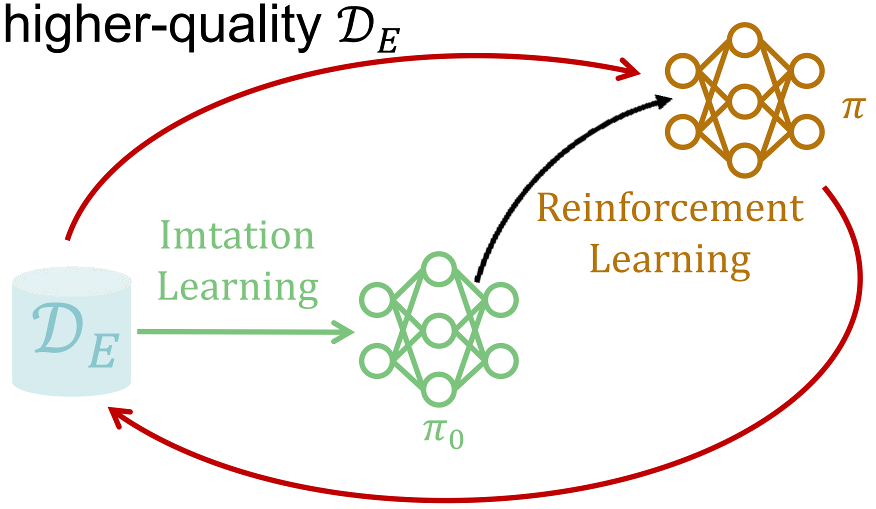
Fixed-wing attitude control

1. Nikishin E, Schwarzer M, D’Oro P, et al. The primacy bias in deep reinforcement learning[C]//International conference on machine learning. PMLR, 2022: 16828-16847.

Method

The Proposed IRPO Method

Iterative training:
train policies iteratively
with higher-quality $\mathcal{D}_E$



Demonstration boosting:
utilize π to boost the quality of $\mathcal{D}_E$

Algorithm 1 Iterative Regularized Policy Optimization (IRPO) method

```
Require: demonstrations $\mathcal{D}_E$, number of iteration: $K$
Ensure: $\pi^*_{1...K}$
1: $\mathcal{D}_0 \leftarrow \mathcal{D}_E$
2: for all $k \in 1 \dots K$ do
3:     train $\pi^0_k$ by Eq. 6 on $\mathcal{D}_{k-1}$
4:     $\pi_k \leftarrow \pi^0_k$
5:     optimize $\pi_k$ by Eq.7 with the imitation policy as $\pi^0_k$
       to get the $\pi^*_k$
6:     sample with $\pi^*_k$ to get $\mathcal{D}'_k$
7:     update demonstrations $\mathcal{D}_k \leftarrow \mathcal{D}_{k-1} \uparrow_f \mathcal{D}'_k$
8: end for
```

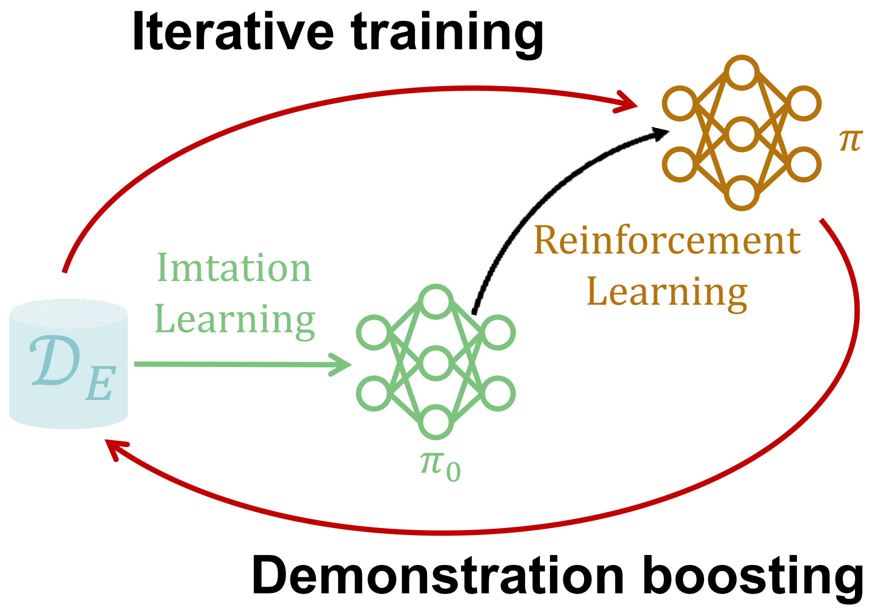
Method

Demonstration boosting

For demonstration $\mathcal{D}$ and a trajectory τ, the demonstration update operator $\uparrow_f$ is defined:

$$\mathcal{D} \uparrow_f \tau = \begin{cases} \mathcal{D} \setminus \tau' \cup \{\tau\}, & \text{if} \exists \tau', \text{s.t.} f(\tau') < f(\tau) \\ \mathcal{D}, & \text{else} \end{cases}$$

After obtaining π_k^* from the k^{th} iteration, we roll out demonstrations $\mathcal{D}_k^{'}$ with π_k^* and update demonstrations as $\mathcal{D}_k \leftarrow \mathcal{D}_{k-1} \uparrow_f \mathcal{D}_k^{'}$.



Iterative training

Theorem 4.3. Define π_k as the policy obtained the k^{th} iteration by optimizing the objective:

$$\pi_k \leftarrow \arg\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{H-1} \gamma^t \left(r(s_t, a_t) - \lambda \log\left(\frac{\pi(a_t|s_t)}{\pi_{k-1}(a_t|s_t)}\right)\right)\right],$$

π_0 is the pre-trained policy. Let $v_{max}^{\lambda} := \frac{r_{max} + \lambda \ln |A|}{1-\gamma}$, ϵ_j is the approximation error of value function at j^{th} iteration, then:

$$|J(\pi_k) - J(\pi^*)| \leq \frac{4}{1-\gamma}\left\|\frac{1}{k+1}\sum_{j=0}^{k}\epsilon_j\right\| + \frac{8}{1-\gamma}\frac{v_{max}^{\lambda}}{k+1}, k \in N.$$

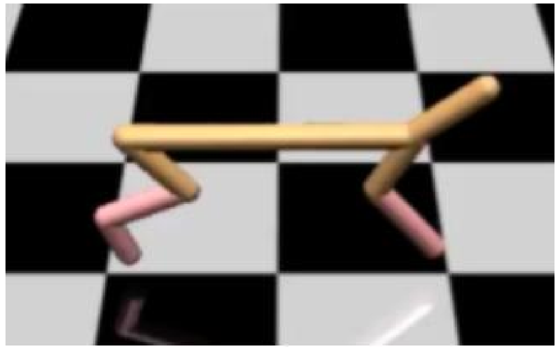
Theorem 4.3 implies that as iterations proceed, $J(\pi)$ approaches $J(\pi^*)$ in an expected sense.

Outline

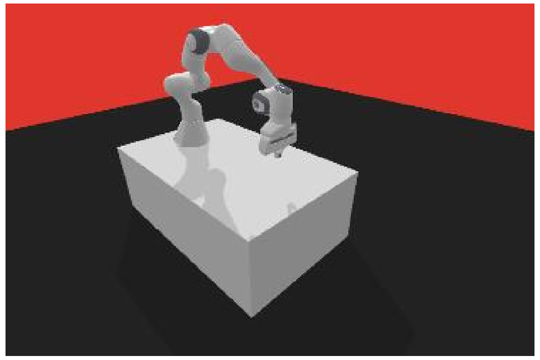
Experiments

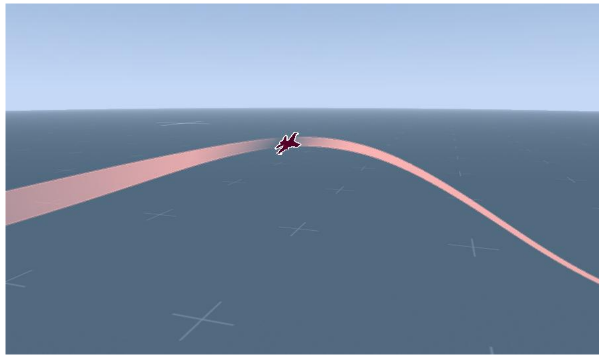
Settings

➢ Tasks


Halfcheetah

Hopper


Reach

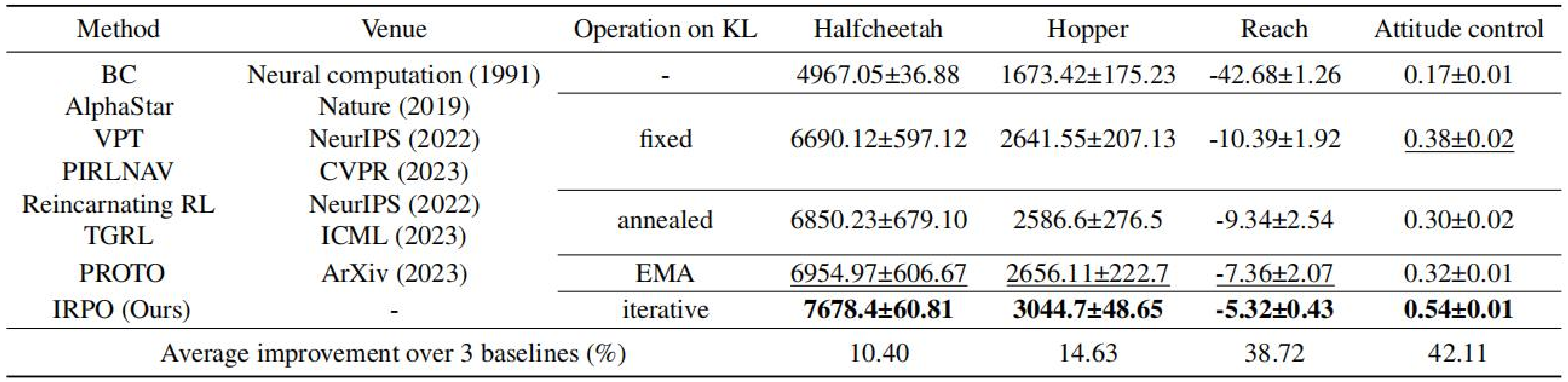

Fixed-wing attitude control

➢ Baselines

Operation on KL	Method	Venue
fixed	AlphaStar	Nature (2019)
	VPT	NeurIPS (2022)
	PIRLNAV	CVPR (2023)
annealed	Reincarnating RL	NeurIPS (2022)
	TGRL	ICML (2023)
EMA	PROTO	ArXiv (2023)

Experiments

Main results

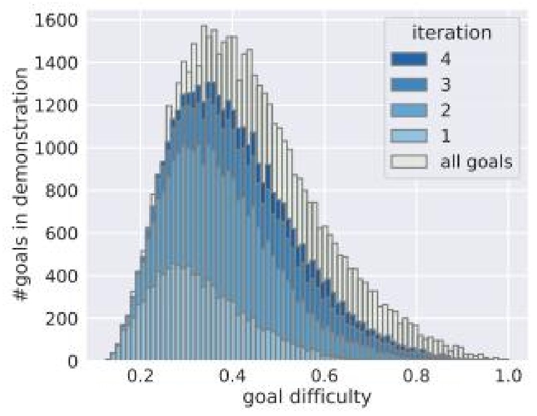


Method	Venue	Operation on KL	Halfcheetah	Hopper	Reach	Attitude control
BC	Neural computation (1991)	-	4967.05±36.88	1673.42±175.23	-42.68±1.26	0.17±0.01
AlphaStar	Nature (2019)	fixed	6690.12±597.12	2641.55±207.13	-10.39±1.92	0.38±0.02
VPT	NeurIPS (2022)					
PIRLNAV	CVPR (2023)					
Reincarnating RL	NeurIPS (2022)	annealed	6850.23±679.10	2586.6±276.5	-9.34±2.54	0.30±0.02
TGRL	ICML (2023)					
PROTO	ArXiv (2023)	EMA	6954.97±606.67	2656.11±222.7	-7.36±2.07	0.32±0.01
IRPO (Ours)	-	iterative	**7678.4±60.81**	**3044.7±48.65**	**-5.32±0.43**	**0.54±0.01**
Average improvement over 3 baselines (%)			10.40	14.63	38.72	42.11

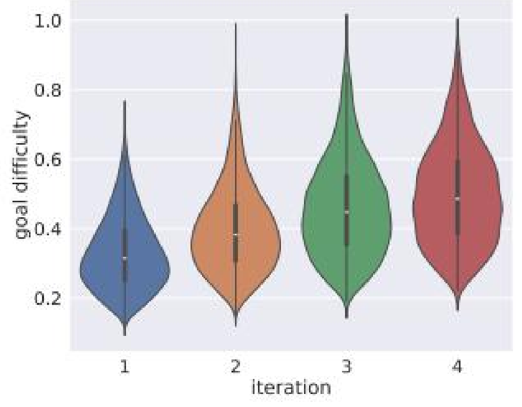
- The table indicates that IRPO outperforms all three baseline algorithms across all four tasks. This suggests that IRPO exhibits effectiveness in learning superior policies from imperfect emonstrations and exhibits applicability across a diverse range of tasks.

Experiments

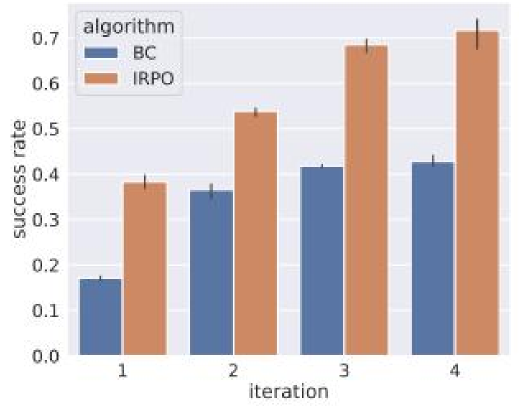
IRPO Improves Policy’s Expected Return Progressively



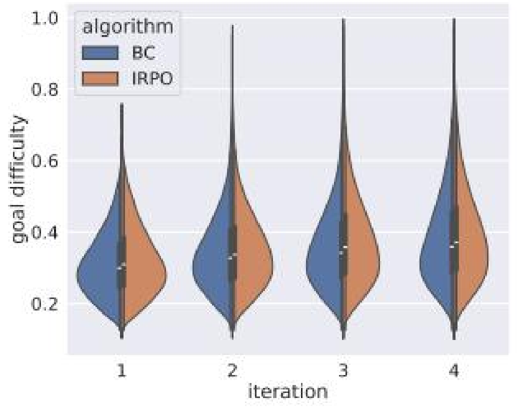
(a) histogram of goals on goal difficulty



(b) distribution of goal difficulty of new-added trajectories



(c) number of goals finished by BC and IRPO



(d) distribution of difficulty of goals finished by BC and IRPO

- Figure (a) shows that **demonstrations** cover **more** goals as IRPO iterates.
- Figure (b) shows that **demonstrations** cover **more difficult** goals as IRPO iterates.
- Figure (c) shows that **policy** achieve **more** goals as IRPO iterates.
- Figure (d) shows that **policy** achieve **more difficult** goals as IRPO iterates.

- The above results suggest that IRPO can improve the quantity and quality of demonstrations, and converge the policy to a higher performance with each iteration.

Experiments

Ablation Studies

① Does iteration with rolling out data perform better than iteration with remaining policy?

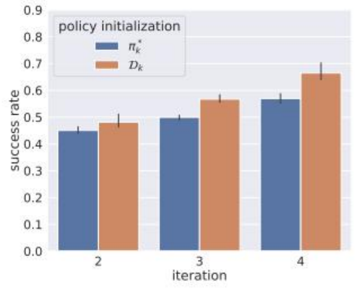
Data is more valuable than policy for the next training iteration.

② How many demonstrations should be rolled out for the next iteration?

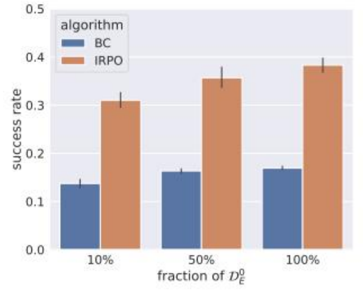
Roll out as many demonstrations as possible.

③ Which part of goal space should be focused on rolling out demonstrations for multi-goal problems?

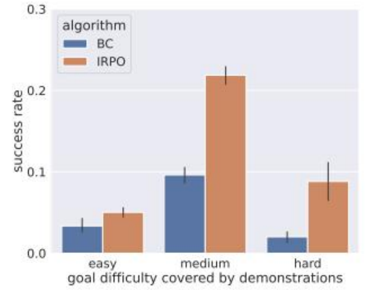
Roll out demonstrations in areas with dense goal distribution.



(a) Success rate of policies with different remains for the next iteration.



(b) Success rate of policies with different quantity of demonstrations.



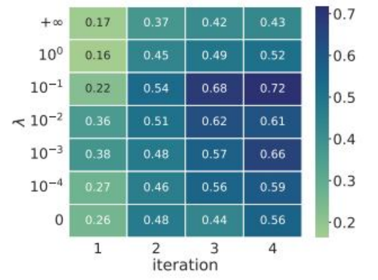
(c) Success rate of policies with demonstrations from different part of goal distribution.

Experiments

Ablation Studies

④ How to set the strength of KL regularization for different iterations?

Relax the KL regularization when demonstration quality improves.



(d) Success rate of policies with different strength of KL regularizations in different iterations.

⑤ How does the indicator function influence the performance?

$f(\tau)$	Demonstration (quantity ↑)	Demonstration (trajectory length ↓)	Demonstration (trajectory smooth ↓)	Policy (success rate ↑)	Policy (trajectory length ↓)	Policy (trajectory smooth ↓)
no $f(\tau)$	24924	193.61±128.29	7.23±5.53	0.43±0.01	285.32±131.95	17.12±16.61
$f(\tau) = -smooth(\tau)$	24924	132.07±60.37	**5.76±6.35**	0.42±0.05	**221.27±122.67**	**12.74±10.58**
$f(\tau) = -length(\tau)$	24924	**124.64±53.07**	7.42±6.38	**0.54±0.01**	224.88±120.59	40.94±36.61
reference: demonstrations and policy in 1st iteration	10184	281.83±149.48	2.11±2.21	0.38±0.02	223.14±131.06	11.01±11.74

IRPO demonstrates robustness to the setting of indicator functions.

⑥ How well does IRPO perform on easily accessible imperfect demonstrations, such as human play data?

IRPO performs well on human play data.

Data Source	Iteration	Demonstration (quantity ↑)	Demonstration (trajectory length ↓)	Policy (success rate ↑)
Human Play	1	613	143.71±23.91	0.29±0.02
Human Play	2	21014	133.96±51.57	0.36±0.03
PID	1	10184	281.83±149.48	0.38±0.02
PID	2	24924	124.64±53.07	0.54±0.01

Outline

Discussion

Our Contributions

- We provide an analysis of the **over-constrained exploration problem** and **primacy bias problem** that arise during offline IL to online RL with imperfect demonstrations.
- We propose IRPO, a framework that consists of **iterative training** and **demonstration boosting**. These components work synergistically to address both the over-constrained exploration and primacy bias issues simultaneously.
- We **systematically assess** the efficacy of IRPO across diverse and complex tasks. Our results demonstrate that IRPO exhibits a consistent ability to enhance policy and demonstration quality over successive iterations.

Limitations

- the design of the indicator function in the demonstration boosting mechanism relies on our understanding of the task.
- Our experimental validation has been limited to on-policy RL. It is still unclear whether IRPO is applicable to off-policy RL settings.

Thanks for watching!

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

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