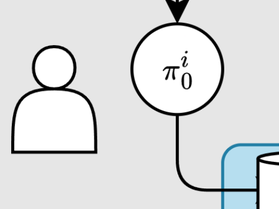
DiffFP: Learning Behaviors from Scratch via Diffusion-based Fictitious Play

Akash Karthikeyan Yash Vardhan Pant
Department of Electrical and Computer Engineering, University of Waterloo

Overview

How to train agents in multi-agent environments with unseen opponents?



Expressive, Multimodal Policy Representation + Game-theoretic Algorithm

We tackle the problem of learning policies in dynamic, continuous state-action games. We propose DiffFP, which:

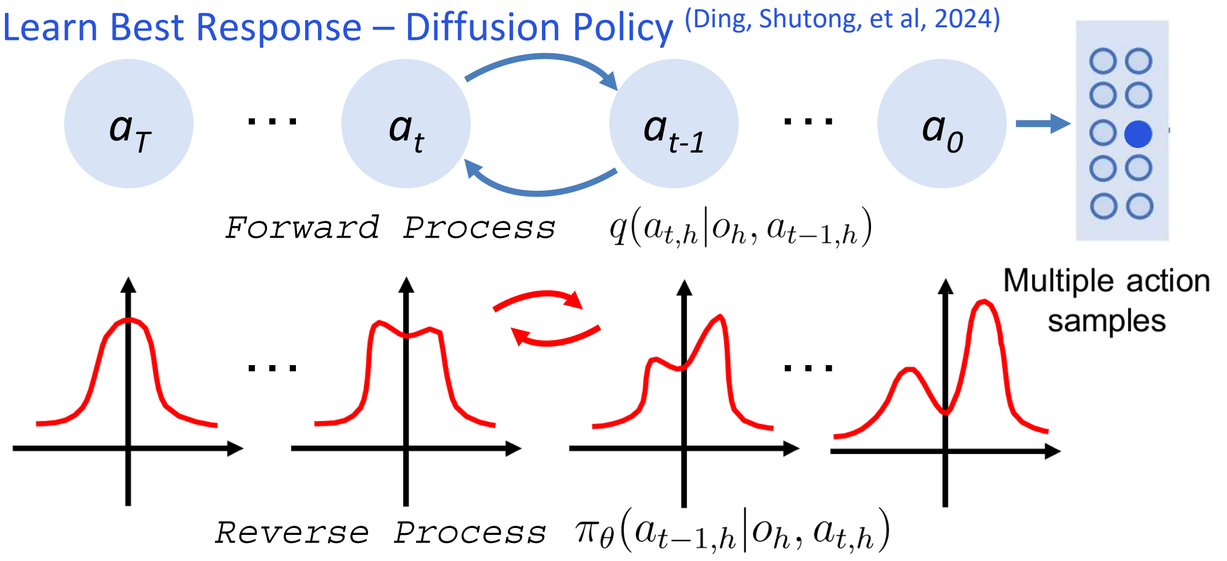
A. Learns from scratch via iterative generalized weakened fictitious play.
B. Compute approximate best response via diffusion policy.
C. Enables learning of sample-efficient and multimodal policies.

Method

DiffFP builds upon the Fictitious Play framework with two core components: **(1)** The outer loop of Fictitious Play, where agents maintain and update empirical average strategies. **(2)** A generative best-response policy implemented as a conditional diffusion model, trained using policy gradients to approximate optimal behavior against the average opponent.

Fictitious Play (Leslie and Collins, 2006 ; Chen, Jingxiao, et al, 2024)

Learn Best Response – Diffusion Policy (Ding, Shutong, et al, 2024)

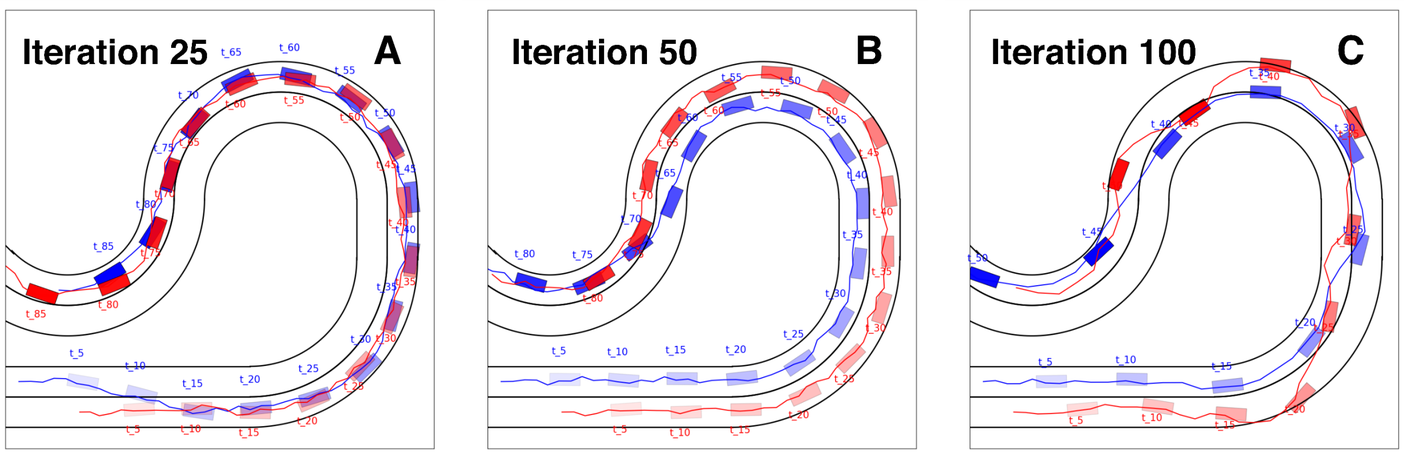


$$\text{DDPM Loss } \mathbb{E}_{\substack{t\sim\mathcal{U}\{1,\dots,T\}\\ a_0\sim\mathcal{D},\ \epsilon\sim\mathcal{N}(0,I)}}\left[\left\|\epsilon-\epsilon_\theta\left(\sqrt{\bar{\alpha}_t}\,a_{0,h}+\sqrt{1-\bar{\alpha}_t}\,\epsilon,\ o_h,\ t\right)\right\|^2\right]$$

$$\text{Action Gradient } \tilde{a}_h^{n+1}\leftarrow a_h^n+\eta\nabla_a\left(\min_{j=1,2}Q_{\psi_j}(o_h,a)\right)\Big|_{a=a_h^n}$$

Simulation Studies

We evaluate DiffFP on various continuous action space environments for convergence, efficiency and robustness to unseen opponents.



Training Progression in Racing. DiffFP learns to race and perform overtakes and wins more frequently on head-to-head trials.

Exploitability: Measures how much a player can improve their payoff by unilaterally deviating from their current strategy.

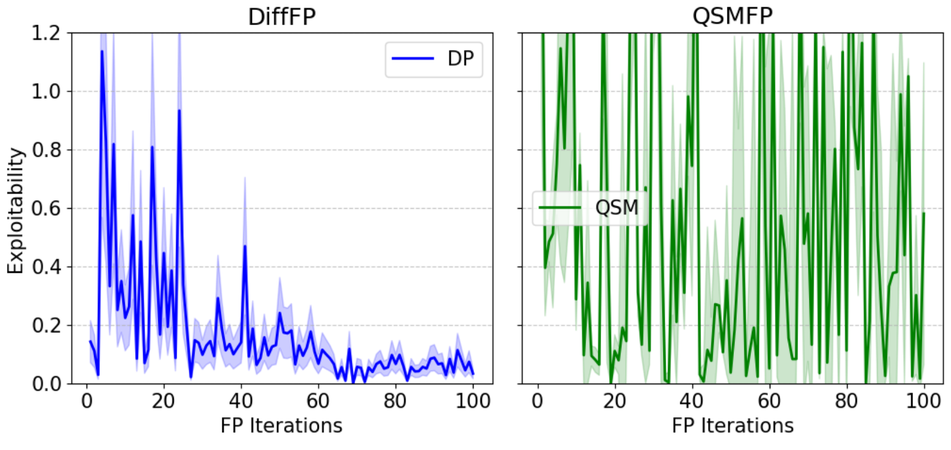
$\sum_{i\in\mathcal{I}}$ (Cumulative rewards of BR policy *i* − Cumulative rewards of average policy *i*)

Ego	Adv	Ego Wins	Adv Wins	Draws
Games where DiffFP is Ego				
DiffFP	SACFP	79	10	11
DiffFP	TD3FP	75	11	14
DiffFP	QSMFP	75	12	13
Games where DiffFP is Adversary				
SACFP	DiffFP	62 (17↓)	16 (6↑)	22 (11↑)
TD3FP	DiffFP	50 (25↓)	18 (7↑)	32 (18↑)
QSMFP	DiffFP	63 (12↓)	27 (15↑)	10 (3↓)

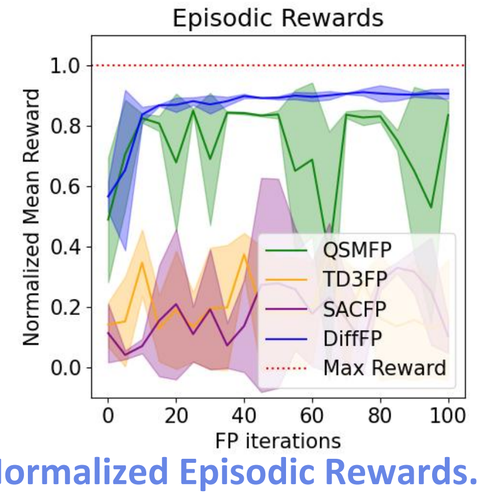
Head-to-Head MPE. Colored numbers indicate relative gains (↑) / losses (↓) when ego and adversary roles are swapped

Attacking	Gap (m)	Successes	Crashes	Mean Reward
QSMFP	8.26	3 / 20	4 / 20	0.80
DiffFP	**17.70**	**12 / 20**	**1 / 20**	**0.92**

Head-to-Head Trials Racing. Gap denotes the mean relative distance gained by the attacker. Successes refer to the number of episodes where the attacker successfully closed the initial gap or overtook the target.



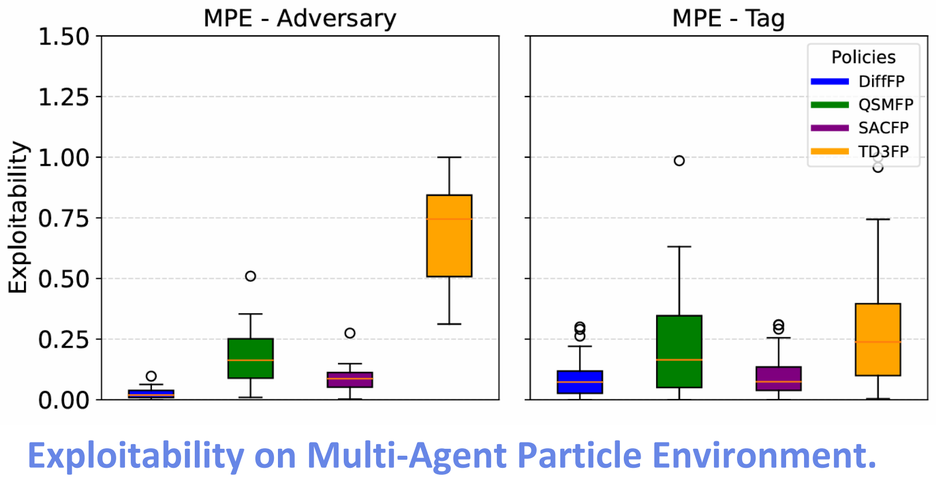
Exploitability on the Racing Task.



Normalized Episodic Rewards.

Opponent	Crashes	Mean Reward
QSMFP	11 / 20	0.556
DiffFP	**4 / 20**	**0.756**

Robustness to Unseen Opponents. Evaluation of crash rates (lower is better) and mean rewards (higher is better) when the agent is tested against up to five previously unseen adversaries.



Exploitability on Multi-Agent Particle Environment.

Future Work

- **Population Learning:** Learn a population of strategies within a single conditional network, using the compute resources of self-play.
- **Scaling:** Extend to complex games with higher-dimensional actions and benchmark on systems with more unstable dynamics.
- **Efficiency:** Faster convergence and sampling in diffusion-based policies for adaptation in agile systems.