NEURAL INFORMATION PROCESSING SYSTEMS

上海纽约大学 NYU SHANGHAI

UCLA

Solving Continuous Mean Field Games: Deep Reinforcement Learning for Non-Stationary Dynamics

Lorenzo Magnino[*]
University of Cambridge

Kai Shao[†]
KTH Royal Institute of Technology

Zida Wu
University of California, Los Angeles

Jiacheng Shen
NYU Center for Data Science

Mathieu Laurière
NYU Shanghai

[*]Work done during period at NYU Shanghai Center for Data Science and the NYU-ECNU Institute of Mathematical Sciences at NYU Shanghai. Contacts: lm2183@cam.ac.uk, kshao@kth.se, zdwu@ucla.edu, shen.patrick.jiacheng@nyu.edu, mathieu.lauriere@nyu.edu.

[†]Work done during period at NYU Shanghai

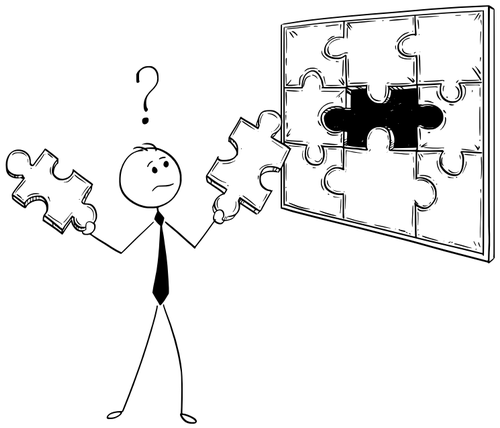
NEURAL INFORMATION PROCESSING SYSTEMS

Challenges!

Extending **RL** frameworks to continuous space, **non-stationary MFGs** presents significant challenges, especially for **learning time-dependent population dynamics** and **solving the resulting fixed-point problems**.

In contrast to MDPs, where the goal is to optimize a single agent's trajectory, **solving MFGs requires learning both an optimal response and a consistent population evolution.**

To the best of our knowledge, **no existing RL algorithms** are capable of learning the solution of non-stationary MFGs with continuous state and action space.

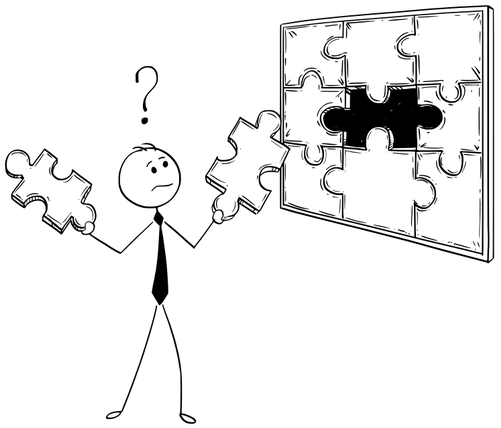


NEURAL INFORMATION PROCESSING SYSTEMS

Related works

Method	Cont. space	General r, P	NE policy	Local. dep.	Non-stat.
DEDA-FP	✓	✓	✓	✓	✓
Zaman et al. [2020]	✓	✗	✓	✗	✓
Perrin et al. [2021]	✓	✓	✗	✓	✗
Laurière et al. [2022a]	✗	✓	✓	✓	✓
Angiuli et al. [2023]	✓	✗	✓	✗	✗

Table 1: Comparison between our approach and related works. Our approach is the first to learn the Nash equilibrium policy and distribution for continuous space non-stationary MFGs with general dynamics and rewards, including possibly local dependence on the mean field.



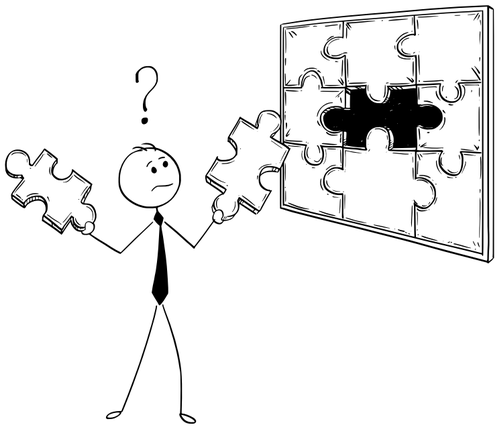
NEURAL INFORMATION PROCESSING SYSTEMS

Major bottleneck

Remaining Challenges. At this point, the algorithm can learn the **equilibrium policy**, which is one part of the solution to the MFGs (see Def. 1). However, the other part, namely the **equilibrium mean field**, is still lacking. Most existing works then approximate the optimal mean field distribution by sampling a large number of trajectories to adequately cover the state space. However, as we will elaborate in Sec. 5, there are several key limitations to this approach:

1. Many mean field games derive their complexity and richness from **local interactions**, where the dynamics or rewards depend on the population density (e.g., congestion, entropy maximization). Without a direct model for the mean field distribution, the density must be estimated indirectly (e.g., via Gaussian convolution), which can alter the nature of the problem.

2. In the **evaluation** or rollout phase, estimating the mean field and its density requires sampling many trajectories at each step. This becomes computationally expensive, especially in state spaces of dimension $d \geq 2$, and can significantly slow down execution.



NEURAL INFORMATION PROCESSING SYSTEMS

Proposed Solution: DEDA-FP!!!

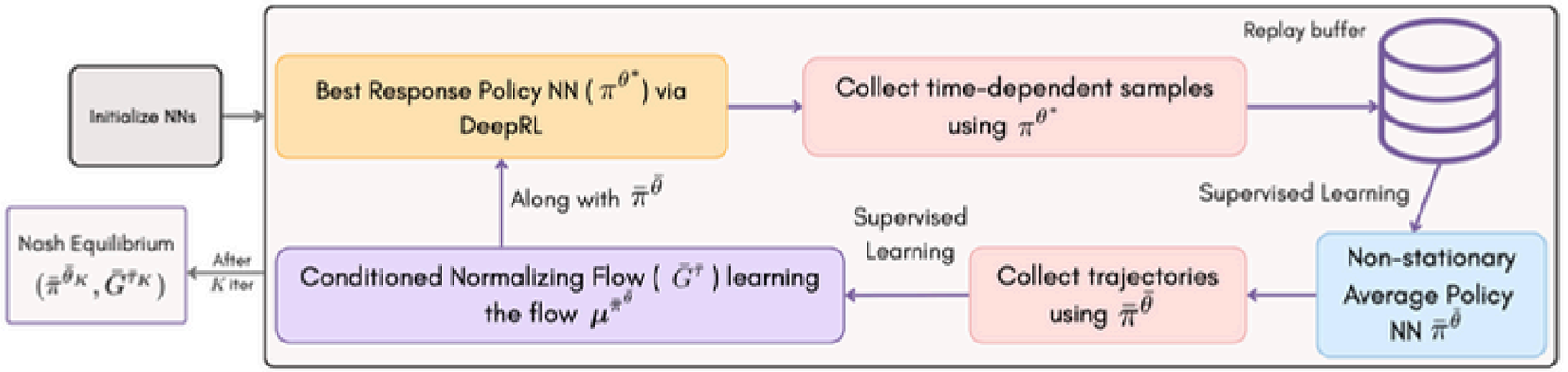
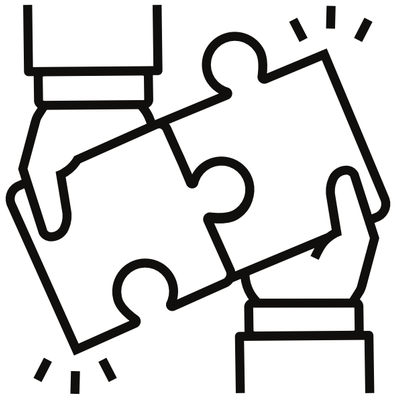


Figure 1: Overview of our **DEDA-FP** model. Our framework uses three main steps, built upon the Fictitious Play algorithm, to fully solve the MFG problem (details in Section 3): (1) computation of the best response using Deep RL algorithms; (2) learning a policy neural network to approximate the average policy over past policies; and (3) learning a **Time-Conditioned Normalizing Flow** to approximate the average distribution over past mean-field flows.



NEURAL INFORMATION PROCESSING SYSTEMS

Proposed Solution: DEDA-FP!!!

Algo. 3 Density-Enhanced Deep Average Fictitious Play (**DEDA-FP**)

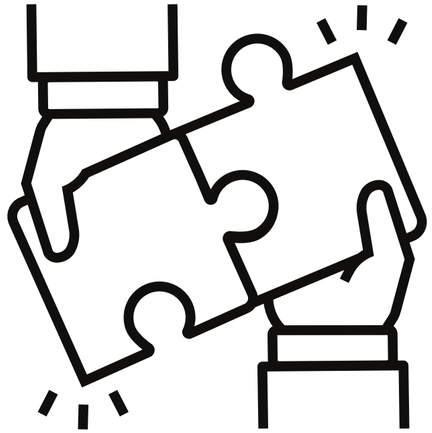
1: **Input:** μ_0: initial distribution; N_{sa}: number of state-action pairs to collect at every iteration, N: population size in population simulation; K: number of iterations.
2: **Initialize:** $(\theta_0^* = \bar{\theta}_0, \bar{\tau}_0)$ at random; empty replay buffer $\mathcal{M}_{SL}$ for supervised learning of average policy; using $\boldsymbol{\pi}_0^* := \boldsymbol{\pi}^{\theta_0^*}$, sample N_{sa} triples $(0, s, a)$ and store them in $\mathcal{M}_{SL}$.
3: **for** iteration $k = 1$ to K **do**
4: Find the best response $\boldsymbol{\pi}_k^* := \boldsymbol{\pi}^{\theta_k^*}$ vs the $N-1$ agents using $\bar{\boldsymbol{\pi}}_{k-1} := \bar{\boldsymbol{\pi}}^{\bar{\theta}_{k-1}}$ using **Deep RL**:

$$\boldsymbol{\pi}_k^* = \arg\max_{\boldsymbol{\pi}} J_{\mu_0}^N(\boldsymbol{\pi}, \bar{\boldsymbol{G}}_{k-1})$$

5: Collect N_{sa} time-state-action samples of the form (t, s, a) using $\boldsymbol{\pi}_k^*$ and store in $\mathcal{M}_{SL}$.
6: Train the **NN policy** $\bar{\boldsymbol{\pi}}_k := \bar{\boldsymbol{\pi}}^{\bar{\theta}_k}$ using supervised learning to minimize the categorical loss:

$$\mathcal{L}_{\text{NLL}}(\bar{\theta}) = \mathbb{E}_{(t,s,a)\sim\mathcal{M}_{SL}}\left[-\log \bar{\boldsymbol{\pi}}^{\bar{\theta}}(a|t,s)\right]$$

7: Train a **Conditional Normalizing Flow** $\bar{\boldsymbol{G}}_k := \bar{\boldsymbol{G}}^{\bar{\tau}_k}$ for the time-dependent mean field $\boldsymbol{\mu}^{\bar{\boldsymbol{\pi}}_k}$ using trajectories generated by $\bar{\boldsymbol{\pi}}_k$ and initialization $\bar{\boldsymbol{G}}_{k-1}$.
8: **end for**
9: **return** $\bar{\boldsymbol{\pi}}_K, \bar{\boldsymbol{G}}_K$



NEURAL INFORMATION PROCESSING SYSTEMS

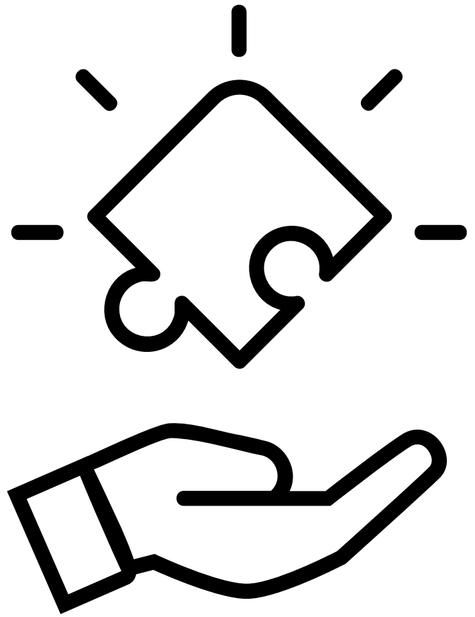
Contribution of the paper

Contributions. This paper introduces Density-Enhanced Deep-Average Fictitious Play (DEDA-FP) (see Figure 1), a novel deep reinforcement learning (DRL) algorithm for **non-stationary** Mean Field Games (MFGs) with **continuous** state and action spaces. Our approach extends **Fictitious Play (FP)**, a classical game-theoretic learning scheme that iteratively updates each agent's policy to optimally respond to the evolving population behavior.

To address the challenge of averaging neural policies, we use **DRL** (Soft Actor-Critic and Proximal Policy Optimization) to compute approximate best responses and **supervised learning** to represent the averaged policy across FP iterations. This hybrid strategy ensures scalability and accurate policy approximation.

We also train a time-dependent Conditional Normalizing Flow (CNF) to model the non-stationary evolution of the **population distribution**, enabling **sampling** from the equilibrium mean field and **density estimation**. This model accurately captures MFGs with local dependence on population density, unlike empirical distributions, and **improves sampling time efficiency tenfold** compared to our benchmarks.

We validate our method with three experiments of increasing complexity and provide an error propagation analysis (Theorem 1). Our contributions address key challenges in applying RL to MFGs, including **time-dependence**, **continuous spaces**, and **local density effects**, representing a significant step toward scalable, model-free solutions for real-world multi-agent systems.



NEURAL INFORMATION PROCESSING SYSTEMS

Results: 4-rooms exploration

$$r(x, v, \mu) = -c_A ||v||_2^2 - c_M \log(\mu(x) + \epsilon)$$

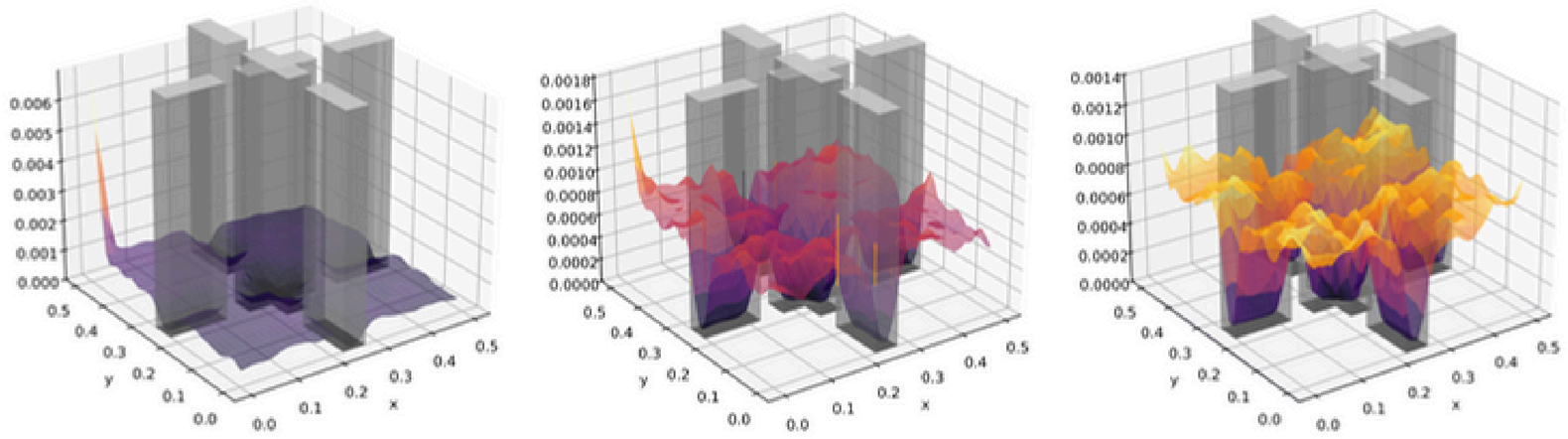


Figure 5: **4-rooms Exploration - NE flow.** The three plots represent the dynamics of the Nash Equilibrium mean field flow $\bar{G}_K$ at time $t = 6, 15, 20$, obtained by **DEDA-FP**. It can be seen how the population is spreading across the 4 different rooms.

NEURAL INFORMATION PROCESSING SYSTEMS

Results: 4-rooms exploration. N-agent simulation

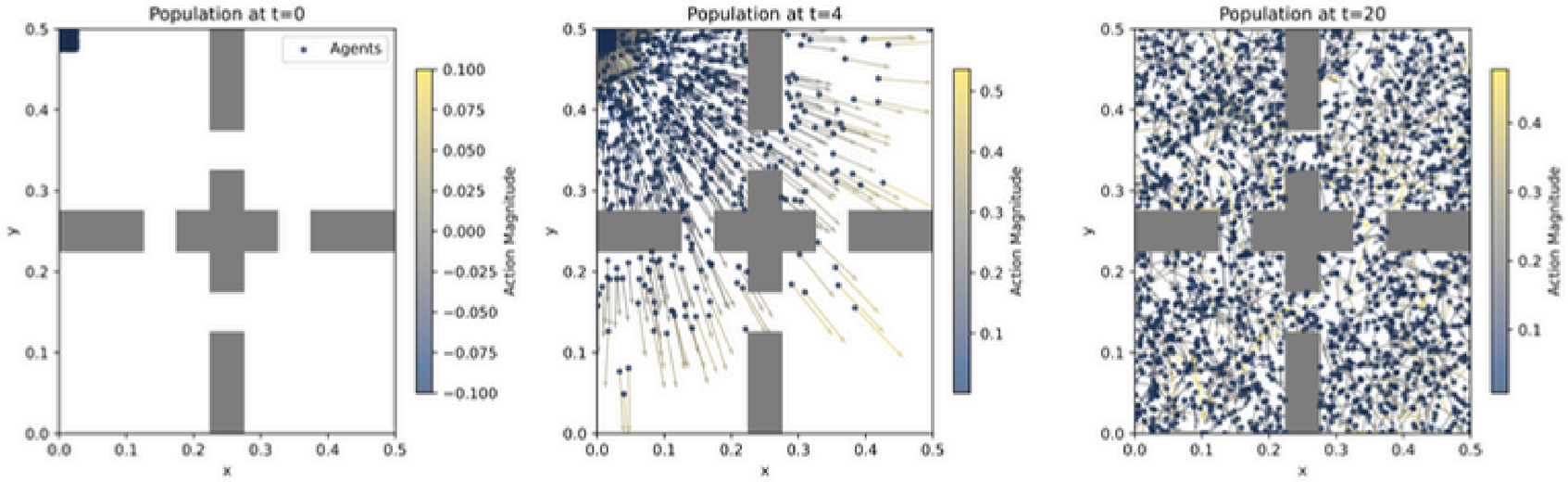


Figure 4: **4-rooms Exploration.** Visualization of a large, finite population of 2000 agents and their velocity vectors during exploration in the 4-rooms environment. The agents' behavior is governed by the mean-field Nash equilibrium policy learned by the DEDA-FP solver, Algo. 3. This shows how well the mean-field approximation captures the behavior of a large-population system.

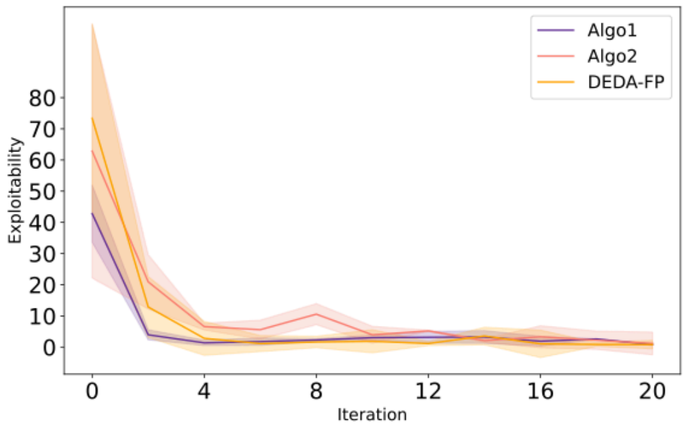
NEURAL INFORMATION PROCESSING SYSTEMS

Strenghts of DEDA-FP

DEDA-FP learns through a **time-conditional normalizing flow** the population distribution at each time step.

- **Direct access** to the **local dependence** in the reward function, without the need to compute any approximation (e.g., convolution).
- **Scalable Sampling**: generates trajectories over 10 times faster. This advantage is critical during rollout when applying the mean field policy in a finite agent context.

...and all this without compromising performance.



Algorithm	Time (s)
Algo1	16.76±1.54
Algo2	15.14±1.19
DEDA-FP	1.52±0.23

NEURAL INFORMATION PROCESSING SYSTEMS

Limitations and future work

- Still **lacking** a **complete theoretical understanding** of the proposed algorithm, particularly due to the complexity of analyzing deep neural networks training

Left for future work:

- Extensions beyond standard MFGs, such as **multiple populations** and **graphon games**, or MFGs with **common noise** and real-world applications.
- Our present evaluation relies on **approximate exploitability**, which, while a state-of-the-art technique for assessing Nash equilibria, provides an evaluation that is inherently dependent on the environment approximation. We will **investigate** this aspect **further**.

NEURAL INFORMATION PROCESSING SYSTEMS

If you have additional **ideas** regarding the potential **applications** of our method or **extensions** related to the **multi-agent dynamics** in **continuous time**, please don't hesitate to **REACH OUT**!

Lorenzo Magnino: lm2183@cam.ac.uk

Mathieu Lauriere: mathieu.lauriere@nyu.edu