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

Lorenzo Magnino Kai Shao Zida Wu Jiacheng Shen Mathieu Laurière

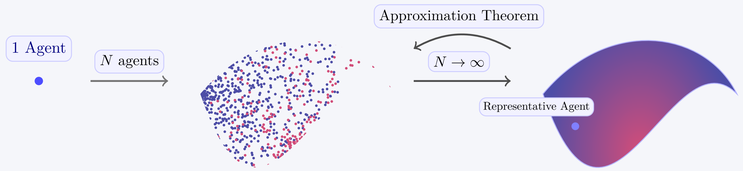
上海纽约大学 NYU SHANGHAI UCLA

NEURAL INFORMATION PROCESSING SYSTEMS

Why This Matters!

- BRIDGE the GAP: practical implementation in multi-agent settings.
- Time-conditioned Normalizing Flow: solving INTRACTABILITY issues
- Take into account Local Interactions (e.g *congestion*)
- Population Inference at SCALE!
- Full Access to the Nash Equilibrium

Mean Field Game Model



Dynamics of the population:

$$\mu_{t+1}^{\pi}(x') = \int_{\mathcal{X}\times\mathcal{A}} \mu_t^{\pi}(x)\pi_t(a|x)P(x'|x,a,\mu_t^{\pi})dxda$$

> A mean-field Nash equilibrium (MFNE) is a pair $(\boldsymbol{\mu}^*, \boldsymbol{\pi}^*) = (\mu_t, \pi_t)_{t\geq 0}$ that satisfies:
>
> - $\boldsymbol{\pi}^*$ maximizes $\boldsymbol{\pi} \mapsto J_{\boldsymbol{\mu}^*}(\boldsymbol{\pi})$;
> - For every $t \geq 0$, μ_t^* is the distribution of x_t.

It can be expressed by the **Exploitability** of a policy $\boldsymbol{\pi}$ is defined as:

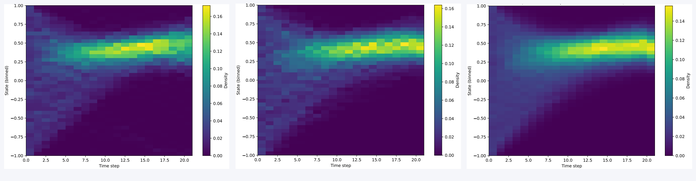
$$\mathbb{E}(\boldsymbol{\pi}) = \max_{\boldsymbol{\pi}'} J_{\boldsymbol{\mu}^{\boldsymbol{\pi}}}(\boldsymbol{\pi}') - J_{\boldsymbol{\mu}^{\boldsymbol{\pi}}}(\boldsymbol{\pi}).$$

Then, $(\boldsymbol{\pi}, \boldsymbol{\mu}^{\boldsymbol{\pi}})$ is Nash equilibrium if and only if $\mathbb{E}(\boldsymbol{\pi}) = 0$.

LQ dynamics

Dynamics: $x_{t+1} = Ax_t + Ba_t + \bar{A}\bar{\boldsymbol{\mu}}_t + \epsilon_t$, where $\bar{\boldsymbol{\mu}}_t = \int_{\mathcal{X}} x\boldsymbol{\mu}_t(x)\mathrm{d}x$.

Reward: $r(x, a, \mu) = -c_X|x - x_{\text{target}}|^2 - c_A|a|^2 - c_M|x - \bar{\boldsymbol{\mu}}|^2$



DEDA-FP

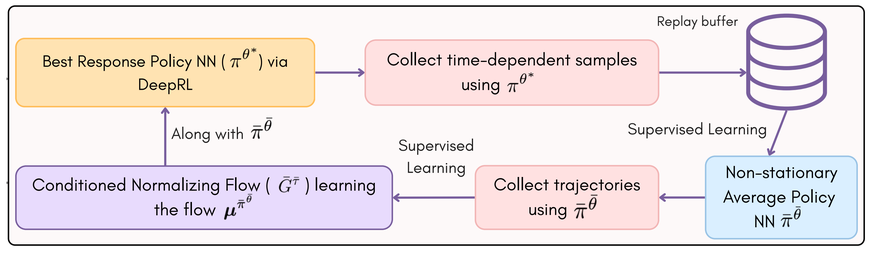
Challenges of Previous Approaches:

- Inaccessible Local Interactions: The density $\mu_t(x)$ must be estimated (e.g with convolution)
- Computationally Intractable: In the rollout phase, estimating the mean field and its density requires sampling many trajectories!

`OUR SOLUTION: for iteration` $k = 1$ `to` K:

1. Best response (DeepRL):
$$\boldsymbol{\pi_k}^* = \arg\max_{\boldsymbol{\pi}} J_{\mu_0}^N(\boldsymbol{\pi}, \bar{\boldsymbol{G}}_{k-1})$$
2. Learn the Average Policy (Behavioral Cloning):
$$\mathcal{L}_{\text{BC}}(\bar{\theta}) = \mathbb{E}_{(t,s,a)\sim\mathcal{M}_{SL}}\left[-\log \bar{\boldsymbol{\pi}}^{\bar{\theta}}(a|t,s)\right]$$
3. Learn the Average Mean Field (Time-Conditioned Normalizing Flow):
$$\mathcal{L}_{\text{NLL}}(\phi) = -\mathbb{E}_{\tau=(t,x)\sim\bar{\boldsymbol{\pi}}_k}\left[\log p_0(f_\phi^{-1}(\mathbf{x}, t)) + \log\left|\det\left(\frac{\partial f_\phi^{-1}(\mathbf{x}, t)}{\partial \mathbf{x}}\right)\right|\right]$$

The **equilibrium** is the last-iteration `(average policy, average flow)`



Inference at SCALE

> $> 10\times$ efficiency advantage with respect to all previous works that do not learn a mean field model. This is CRITICAL, during LEARNING and ROLL-OUT in the finite agent case.

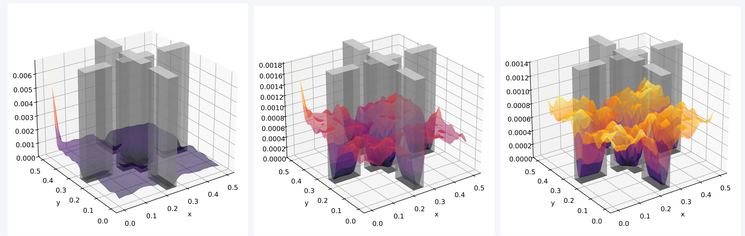
Algorithm	Time (s)
Algo1	16.76±1.54
Algo2	15.14±1.19
DEDA-FP	1.52±0.33

Propagation of Errors

Main Theorem:

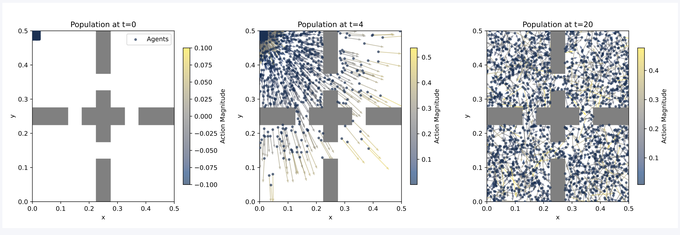
$$e_k^{\text{true}} < C_0\boldsymbol{e}_{\boldsymbol{cnf}}^{\boldsymbol{0}} + \frac{1}{k}\sum_{i=1}^{k-1}\left[(i+1)\boldsymbol{\epsilon}_{\boldsymbol{br}}^{\boldsymbol{i+1}} + C_1(\boldsymbol{\epsilon}_{\boldsymbol{sl}}^{\boldsymbol{i+1}} + \boldsymbol{\epsilon}_{\boldsymbol{cnf}}^{\boldsymbol{i+1}}) + \frac{C_2}{i}\right]$$

Case Study: Exploration of 4-rooms

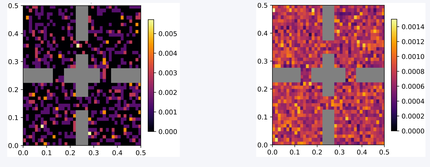


The reward function is defined by the entropy maximization:

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



DEDA-FP yields a superior mean-field distribution representation during training (given a fixed-time budget for all algorithms)



Let's set up a follow-up!

My Latest Research at Univ. of Cambridge

- Multi-Agent Robotics
- Collective Intelligence