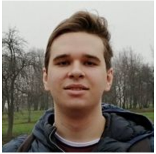
Finite-Sample Convergence Bounds for MF-TRPO

Antonio Ocello[1], Daniil Tiapkin[1], Lorenzo Mancini[1], Mathiéu Laurière[2], Éric Moulines[1,3]

[1]CMAP - École Polytechnique - Institut Polytechnique de Paris, France
[2]NYU Shanghai, China
[3]Mohamed Bin Zayed University of AI, UAE

Introduction to Mean Field Games (MFGs)

- **Definition:**
  - **MFGs:** model strategic interactions among high number of agents.
  - Each individual agent **negligible influence**.
  - Collective behavior: represented through a **mean field term**, summarizing their aggregated effect.
  - **Generalization of the law of large numbers**, allowing for the study of equilibrium dynamics in large-scale multi-agent systems.

- **Applications:** economic modeling (Bassière et al., 2024), finance (Lavigne and Tankov, 2023; Carmona et al., 2013), and energy storage (Alasseur et al., 2020).

Reinforcement Learning (RL) in MFGs

Objective: Use RL techniques to find equilibria in MFGs without explicit knowledge of the system's dynamics.

Setting: **finite state and action spaces**.

Challenges:

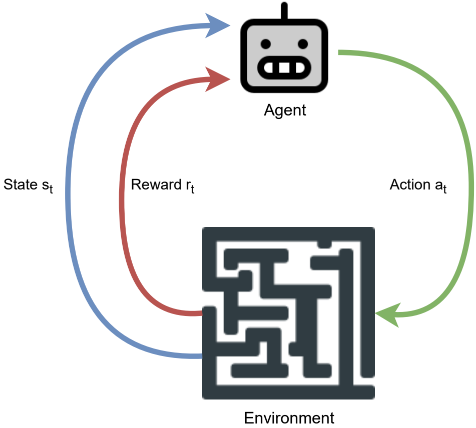
- **Non-linear** nature of the problem, adding significant complexity to the analysis.
- **Ill-conditioned fixed-point solutions**, leading to potential numerical instability.
- Ensuring the convergence of the proposed methods, particularly in high-dimensional settings.

Introduction to Reinforcement Learning (RL)

Definition: RL involves an **agent** learning to **make decisions** by interacting with an environment to maximize cumulative rewards.

Introduction to Reinforcement Learning (RL)

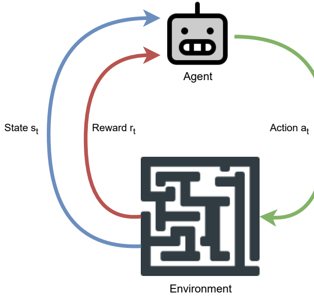
Definition: RL involves an **agent** learning to **make decisions** by interacting with an environment to maximize cumulative rewards.



Introduction to Reinforcement Learning (RL)

Definition: RL involves an **agent** learning to **make decisions** by interacting with an environment to maximize cumulative rewards.

- **Agent:** The learner or decision-maker.
- **Environment:** The external system the agent interacts with.
- **Actions:** The set of choices available to the agent.
- **States:** The situations or contexts in which the agent finds itself.
- **Rewards:** Feedback provided by the environment as a result of the agent's actions.



$$\max_{\pi} \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^t r(a_t, s_t) \middle| s_0 \sim \xi, a_t \sim \pi(\cdot|s_t), s_{t+1} \sim \mathbf{P}(\cdot|a_t, s_t)\right]$$

Multi-Agent Reinforcement Learning (MARL)

MARL: Learning framework where

- **Multiple Agents:**
  - Interact with each other and their environment.
  - Aim to optimize their respective policies.
  - **Must account for the dynamic behavior of other agents**, unlike single-agent RL.
  - Inter-agent interaction renders the **learning environment non-stationary**.
- Each agent $i \in \{1, \ldots, N\}$ in MARL maximizes its own cumulative reward:

$$J_i(\pi_i) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathrm{r}_i(s_t^i, a_t^i, s_t^{-i}) \middle| a_t^i \sim \pi_i(\cdot|s_t^i)\right]$$

  - s_t^i: The state of agent i at time t.
  - a_t^i: The action of agent i at time t.
  - s_t^{-i}: The states of all other agents except i at time t.
  - γ: The discount factor $(0 \leq \gamma \leq 1)$.

Nash Equilibrium in MARL Systems

Definition: A **Nash equilibrium** in a MARL system is a strategy profile $(\pi_1^*, \ldots, \pi_N^*)$ and a space configuration $(s_1^*, \ldots, s_N^*)$ where no agent has an incentive to unilaterally deviate:

$$J_i(s_i^*, \pi_i^*, s_{-i}^*) \geq J_i(s_i^*, \pi_i, s_{-i}^*),$$

for any π_i and $i \in \{1, \ldots, N\}$.

Nash Equilibrium in MARL Systems

Definition: A **Nash equilibrium** in a MARL system is a strategy profile $(\pi_1^*, \ldots, \pi_N^*)$ and a space configuration $(s_1^*, \ldots, s_N^*)$ where no agent has an incentive to unilaterally deviate:

$$J_i(s_i^*, \pi_i^*, s_{-i}^*) \geq J_i(s_i^*, \pi_i, s_{-i}^*),$$

for any π_i and $i \in \{1, \ldots, N\}$.

Problem: Exponential Complexity: Finding a Nash equilibrium in an N-player game is computationally hard as the strategy space growing exponentially.

From MARL to MFG

Motivation:

- Simplify the complexity of interacting agents.
- Address the non-stationarity of the learning procedure to enhance stability and convergence.

From MARL to MFG

Motivation:

- Simplify the complexity of interacting agents.
- Address the non-stationarity of the learning procedure to enhance stability and convergence.

Assumptions:

From MARL to MFG

Motivation:

- Simplify the complexity of interacting agents.
- Address the non-stationarity of the learning procedure to enhance stability and convergence.

Assumptions:

- **Anonymity:** Each agent interacts with the population as a whole rather than individual agents.
  - The influence of a single agent becomes negligible as the number of agents $N \to \infty$.

From MARL to MFG

Motivation:

- Simplify the complexity of interacting agents.
- Address the non-stationarity of the learning procedure to enhance stability and convergence.

Assumptions:

- **Anonymity:** Each agent interacts with the population as a whole rather than individual agents.
  - The influence of a single agent becomes negligible as the number of agents $N \to \infty$.
- **Homogeneity:** All agents have similar objectives, dynamics, and rewards.
  - Agents are considered *exchangeable*, leading to the assumption that their policies can be symmetric.

From MARL to MFG

Motivation:

- Simplify the complexity of interacting agents.
- Address the non-stationarity of the learning procedure to enhance stability and convergence.

Assumptions:

- **Anonymity:** Each agent interacts with the population as a whole rather than individual agents.
  - The influence of a single agent becomes negligible as the number of agents $N \to \infty$.
- **Homogeneity:** All agents have similar objectives, dynamics, and rewards.
  - Agents are considered *exchangeable*, leading to the assumption that their policies can be symmetric.

$$J_i(\pi_i) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \mathrm{r}_i(s_t^i, a_t^i, s_t^{-i})\right]$$

From MARL to MFG

Motivation:

- Simplify the complexity of interacting agents.
- Address the non-stationarity of the learning procedure to enhance stability and convergence.

Assumptions:

- **Anonymity:** Each agent interacts with the population as a whole rather than individual agents.
  - The influence of a single agent becomes negligible as the number of agents $N \to \infty$.
- **Homogeneity:** All agents have similar objectives, dynamics, and rewards.
  - Agents are considered *exchangeable*, leading to the assumption that their policies can be symmetric.

$$J\left(\pi, \mu^{(N)} = \frac{1}{N}\sum_{j=1}^{N} \delta_{s_t^j}, \xi\right) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \mathbf{r}\left(s_t^i, a_t^i, \frac{1}{N}\sum_{j=1}^{N}\delta_{s_t^j}\right) \middle| \begin{array}{l} s_0^i \sim \xi,\ a_t^i \sim \pi(\cdot|s_t^i), \\ s_{t+1}^i \sim \mathbf{P}(\cdot|s_t^i, a_t^i, \mu) \end{array}\right]$$
$$= \xi\left(\mathbb{I} - \mathbf{P}^{\pi}_{\mu^{(N)}}\right)^{-1} \mathbf{r}^{\pi}_{\mu^{(N)}}$$

Nash Equilibrium and Exploitability in MFG

Definition: A Mean-Field Nash Equilibrium (MFNE) is a couple $(\pi_\star, \mu_\star)$ where:

- Each agent chooses a strategy that maximizes their own utility, given the average effect of all other agents, *i.e.*,

$$J(\pi_\star, \mu_\star, \mu_\star) = \max_{\pi} \; J(\pi, \mu_\star, \mu_\star) \; .$$

- The mean-field profile $\mu_\star$ is stable for the optimal strategy $\pi_\star$ at a macroscopic level, *i.e.*,

$$\mu_\star = \mu_\star \, \mathbf{P}_{\mu_\star}^{\pi_\star} \; .$$

Exploitability: measures of improvement of an agent by deviating unilaterally from π, given the mean-field parameter as the stationary distribution $\lambda_{\pi,\mu}$.

$$\phi(\pi, \mu) := \max_{\pi'} J\left(\pi', \lambda_{\pi,\mu}, \lambda_{\pi,\mu}\right) - J\left(\pi, \lambda_{\pi,\mu}, \lambda_{\pi,\mu}\right).$$

Definition: $(\pi_\star, \mu_\star)$ is an ε-MFNE, if its exploitability is bounded by ε, *i.e.*,

$$\phi(\pi, \mu) \leq \varepsilon.$$

Trust Region Policy Optimization (TRPO)

Key Insight:

- Trust Region Policy Optimization (TRPO) is a state-of-the-art reinforcement learning algorithm that strikes a balance between stability and exploration.

Advantages:

- Prevents drastic policy updates, ensuring stable learning.
- Leverages policy improvement guarantees, making it robust to policy changes.

Our Goal:

- Adapt TRPO to the mean field setting.
- Analyze how much data (sample complexity) is needed to ensure convergence to the Nash equilibrium.

TRPO: Adaptive Trust Region Planning

Overview:

- TRPO: trust region planning algorithm with an adaptive proximity term.
- Despite the non-convexity we still have convergence guarantees: $\mathcal{O}(1/k)$

TRPO: Adaptive Trust Region Planning

Overview:

- TRPO: trust region planning algorithm with an adaptive proximity term.
- Despite the non-convexity we still have convergence guarantees: $\mathcal{O}(1/k)$

Update Rule: TRPO iterates, for a fixed μ,

$$\pi_{k+1} \in \arg\max_{\pi} \left\langle \nabla J_{\mu}^{\pi_k}, \pi - \pi_k \right\rangle - \eta(k+2) \left(\mathbb{I} - \gamma \mathbf{P}_{\mu}^{\pi_k} \right)^{-1} D_{\mathrm{KL}}(\pi || \pi_k).$$

TRPO(μ, K) Algorithm

1: **Initialize:** π_0 uniform policy
2: **for** $k \in [K]$ **do**
3: $\quad J_{\mu}^{\pi_k} \leftarrow (\mathbb{I} - \gamma \mathbf{P}_{\mu}^{\pi_k})^{-1} \mathbf{r}_{\eta,\mu}^{\pi_k}$ `Value function`
4: $\quad$ **for** $s \in S$ **do**
5: $\quad\quad$ **for** $a \in A$ **do**
6: $\quad\quad\quad q_{\mu}^{\pi_k}(s,a) \leftarrow \mathbf{r}_{\eta,\mu}^{\pi_k}(s,a) + \gamma \sum_{s'} \mathbf{P}(s'|s,a,\mu) J_{\mu}^{\pi_k}(s')$ `Action-value function`
7: $\quad\quad$ **end for**
8: $\quad\quad \pi_{k+1}(a|s) \leftarrow \dfrac{\pi_k(a|s) \exp\left(\frac{1}{\eta(k+2)} \left(q_{\mu}^{\pi_k}(s,a) + \lambda \log \pi_k(a|s) \right) \right)}{\sum_{a' \in A} \pi_k(a'|s) \exp\left(\frac{1}{\eta(k+2)} \left(q_{\mu}^{\pi_k}(s,a') + \lambda \log \pi_k(a'|s) \right) \right)}$ `Policy Update`
9: $\quad$ **end for**
10: **end for**

MF-TRPO

Tabular TRPO for MFG Algorithm

Initialize: Initial distribution $\mu_0 = \mathcal{U}(\mathcal{S})$, initial policy $\pi_{0,0} = \mathcal{U}(\mathcal{A})$.
for $p \in [P]$ **do**
  Initialize: Initial policy $\pi_{p+1,0} = \pi_{p,K}$.
  for $k \in [K]$ **do**
    $J_{\mu_p}^{\pi_{p+1,k}} \leftarrow \left(\mathbb{I} - \gamma \mathbf{P}_{\mu_p}^{\pi_{p+1,k}}\right)^{-1} \mathbf{r}_{\eta,\mu_p}^{\pi_{p+1,k}}$ Value function
    for $s \in S$ **do**
      for $a \in A$ **do**
        $q_{\mu_p}^{\pi_{p+1,k}}(s,a) \leftarrow \mathbf{r}_{\eta,\mu_p}^{\pi_{p+1,k}}(s,a) + \gamma \sum_{s'} \mathbf{P}(s'|s,a,\mu_p) J_{\mu_p}^{\pi_{p+1,k}}(s')$ Action-value function
      end for
      $$\pi_{p+1,k+1}(a|s) \leftarrow \frac{\pi_{p+1,k}(a|s) \exp\left(\frac{1}{\eta(k+2)}\left(q_{\mu_p}^{\pi_{p+1,k}}(s,a) + \lambda \log \pi_{p+1,k}(a|s)\right)\right)}{\sum_{a' \in A} \pi_{p+1,k}(a'|s) \exp\left(\frac{1}{\eta(k+2)}\left(q_{\mu_p}^{\pi_{p+1,k}}(s,a') + \lambda \log \pi_{p+1,k}(a'|s)\right)\right)}$$ Policy Update
    end for
  end for
  $\mu_{p+1} \leftarrow \mu_{p-1} + \beta_p \left(\mu_{p-1}\left(\mathbf{P}_{\mu_{p-1}}^{\pi_{p+1,K}}\right)^M - \mu_{p-1}\right)$ Update population
end for

Bound for the exact algorithm

Convergence bound of `Tabular TRPO for MFG`

Let $\{\mu_p\}_{p\geq 0}$ and $\{\pi_{p,k}\}_{p,k\geq 0}$ be the sequence generated by `Tabular TRPO for MFG`. Then, under some assumptions (which implies the uniqueness of the MFNE $(\pi_\star, \mu_\star)$). for some $C, \tau > 0$, we have that

$$\max_{\pi} J(\pi, \mu_p, \mu_p) - J(\pi_{p,K}, \mu_p, \mu_p) \leq \frac{C \log K}{K}\,, \qquad \text{for } p \in [P],$$

$$\left\|\mu_P - \mu_\star\right\|^2 \leq \exp\left(-\frac{\tau}{2}\sum_{j=1}^{P} \beta_j\right) \left\|\mu_0 - \mu_\star\right\|^2 + \frac{C \log K}{K}\,.$$

Bound for the exact algorithm

Convergence bound of `Tabular TRPO for MFG`

Let $\{\mu_p\}_{p\geq 0}$ and $\{\pi_{p,k}\}_{p,k\geq 0}$ be the sequence generated by `Tabular TRPO for MFG`. Then, under some assumptions (which implies the uniqueness of the MFNE $(\pi_\star, \mu_\star)$). for some $C, \tau > 0$, we have that

$$\max_\pi J(\pi, \mu_p, \mu_p) - J(\pi_{p,K}, \mu_p, \mu_p) \leq \frac{C \log K}{K}, \quad \text{for } p \in [P],$$

$$\|\mu_P - \mu_\star\|^2 \leq \exp\left(-\frac{\tau}{2}\sum_{j=1}^{P}\beta_j\right)\|\mu_0 - \mu_\star\|^2 + \frac{C \log K}{K}.$$

Moreover, $(\pi_{P+1,K}, \mu_P)$ is ε_P-MFNE, with

$$\varepsilon_P = C\exp\left(-\frac{\tau}{4}\sum_{j=1}^{P}\beta_j\right) + C\sqrt{\frac{\log(K)}{K}}.$$

Crowd Modeling: Four Rooms Environment

- **Environment:** two-dimensional grid divided into four interconnected rooms.
- Agents move through narrow passageways between rooms.
- The reward function **discourages overcrowding**:

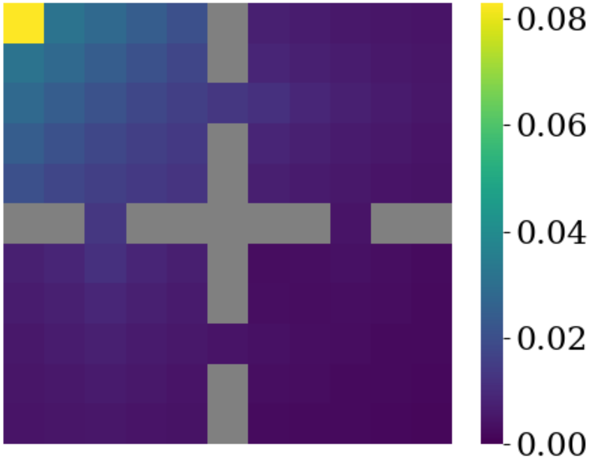
$$\mathrm{r}(s, a, \mu) = -K \log(\mu(s)) + \Gamma(a),$$

with

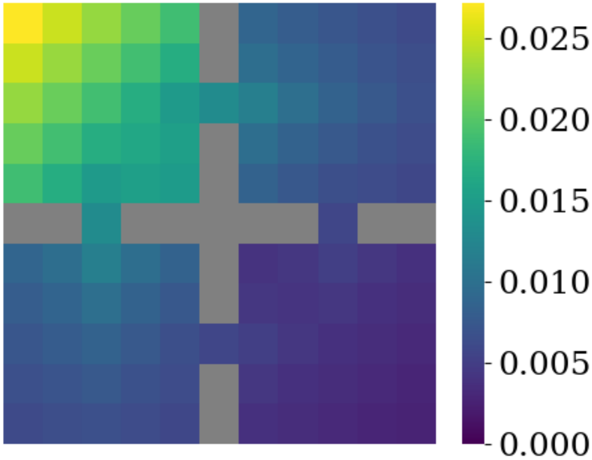
$$\Gamma(a) = \begin{cases} 0.2 & \text{if } a = 0 \quad \text{(Stay)} \\ -0.2 & \text{if } a \in \{\text{Left}, \text{Right}, \text{Up}, \text{Down}\} \quad \text{(Move)} \end{cases}$$

Crowd Modeling: Four Rooms Environment

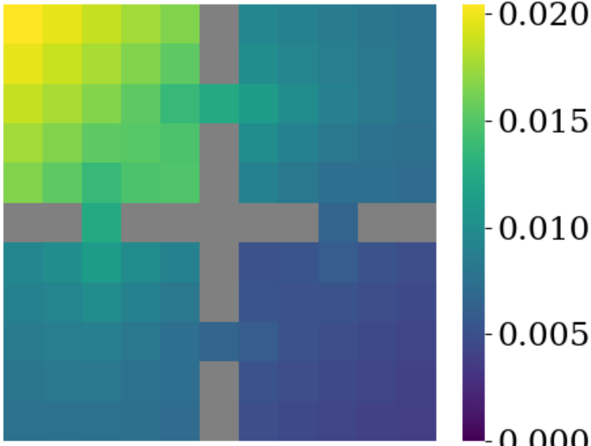
Crowd Modeling: Four Rooms Environment



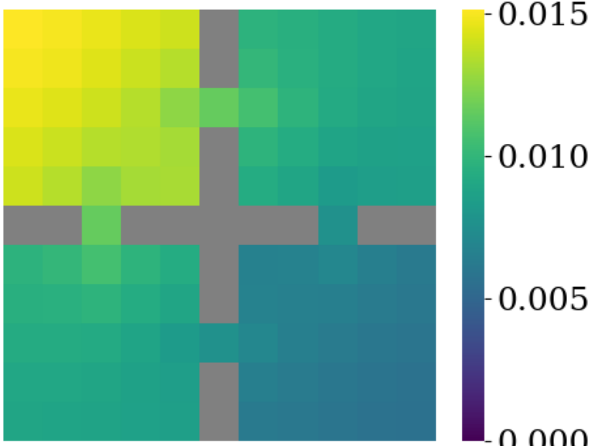
Crowd Modeling: Four Rooms Environment



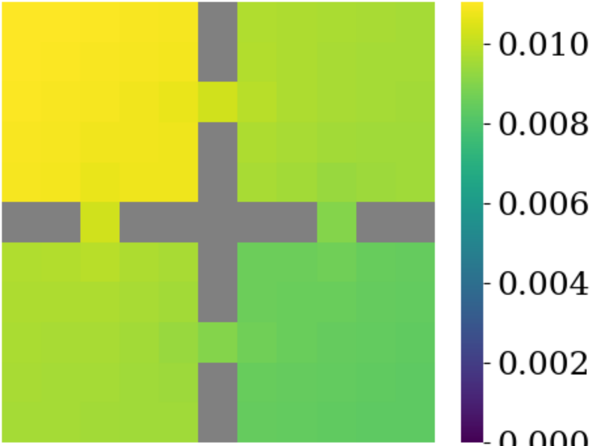
Crowd Modeling: Four Rooms Environment



Crowd Modeling: Four Rooms Environment



Crowd Modeling: Four Rooms Environment



Performance Evaluation: Exploitability Metric

Measure of the performance of the learned policy:

- **Exploitability:** it quantifies the deviation from a Nash equilibrium by measuring the best possible improvement for any agent:

$$\phi(\pi) = \max_{\pi'} J(\pi', \mu^\pi) - J(\pi, \mu^\pi).$$

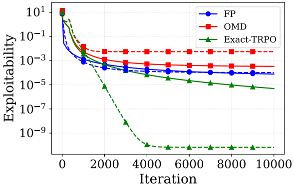
- **Quality:** Evaluates how well a given policy performs under a fixed population distribution:

$$T(\pi, \mu) = \max_{\pi'} J(\pi', \mu) - J(\pi, \mu).$$

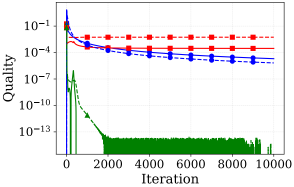
- **Mean-field distribution convergence:** increments in the mean-field distribution parameter between consecutive iterations.

We benchmark our approach against **Fictitious Play** (Perrin et al., 2020) and **Online Mirror Descent** (Pérolat et al., 2022).

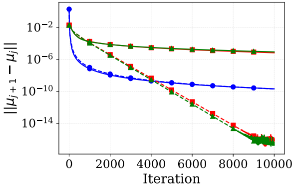
Performance Evaluation: Exploitability Metric



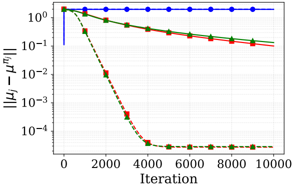
(a) Exploitability



(b) Quality



(c) Mean-field Deltas



(d) Mu Pi Deltas

Questions

Menù del giorno

- Introduction to MFGs and RL
- From MARL to MFGs
- Problem Setting and MF-TRPO
- Algorithm and "Results"
- Visualizations

Future perspectives

- The non-stationary case
- Mean field control
- Continuous-time version of the algorithm
- Robust version of the algorithm

References I

Clémence Alasseur, Imen Ben Taher, and Anis Matoussi. An extended mean field game for storage in smart grids. *Journal of Optimization Theory and Applications*, 184:644–670, 2020.

Alicia Bassière, Roxana Dumitrescu, and Peter Tankov. A mean-field game model of electricity market dynamics. In *Quantitative Energy Finance: Recent Trends and Developments*, pages 181–219. Springer, 2024.

René Carmona, Jean-Pierre Fouque, and Li-Hsien Sun. Mean field games and systemic risk. *arXiv preprint arXiv:1308.2172*, 2013.

Pierre Lavigne and Peter Tankov. Decarbonization of financial markets: a mean-field game approach. *arXiv preprint arXiv:2301.09163*, 2023.

Julien Pérolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling mean field games by online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pages 1028–1037, 2022.

Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in neural information processing systems*, 33:13199–13213, 2020.

Thank you for the attention

Funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.