Learning Discrete-Time Major-Minor Mean Field Games

Kai Cui[*1], **Gökçe Dayanıklı**[*2], **Mathieu Laurière**[3], **Matthieu Geist**[4], **Olivier Pietquin**[5], **Heinz Koeppl**[1]

[1]Technische Universität Darmstadt,
[2]University of Illinois at Urbana-Champaign,
[3]NYU Shanghai,
[4]Google DeepMind,
[5]Cohere
kai.cui@tu-darmstadt.de, gokced@illinois.edu, mathieu.lauriere@nyu.edu,
mfgeist@google.com, olivier@cohere.com, heinz.koeppl@tu-darmstadt.de

Abstract

Recent techniques based on Mean Field Games (MFGs) allow the scalable analysis of multi-player games with many similar, rational agents. However, standard MFGs remain limited to homogeneous players that weakly influence each other, and cannot model major players that strongly influence other players, severely limiting the class of problems that can be handled. We propose a novel discrete time version of major-minor MFGs (M3FGs), along with a learning algorithm based on fictitious play and partitioning the probability simplex. Importantly, M3FGs generalize MFGs with common noise and can handle not only random exogeneous environment states but also major players. A key challenge is that the mean field is stochastic and not deterministic as in standard MFGs. Our theoretical investigation verifies both the M3FG model and its algorithmic solution, showing firstly the well-posedness of the M3FG model starting from a finite game of interest, and secondly convergence and approximation guarantees of the fictitious play algorithm. Then, we empirically verify the obtained theoretical results, ablating some of the theoretical assumptions made, and show successful equilibrium learning in three example problems. Overall, we establish a learning framework for a novel and broad class of tractable games.

Introduction

While reinforcement learning (RL) has achieved tremendous recent success (Mnih et al. 2015; Sutton and Barto 2018), multi-agent RL (MARL) as its game-theoretic counterpart remains difficult due to its many challenges (Zhang, Yang, and Başar 2021). In particular, the scalability challenge is hard to overcome due to the notorious complexity of non-cooperative stochastic games (Daskalakis, Goldberg, and Papadimitriou 2009; Yang and Wang 2021). Here, the recent introduction of mean field games (MFGs, (Lasry and Lions 2007; Huang, Malhamé, and Caines 2006; Saldi, Başar, and Raginsky 2018)) has contributed a mathematically rigorous and tractable approach to handling large-scale games, finding application in a variety of domains such as finance (Carmona 2020) and engineering (Djehiche, Tcheukam, and Tembine 2017). The general idea is to summarize many similar agents (players) and their interaction through their state distribution – the mean field (MF). Owing to the amenable complexity of MFGs, many recent efforts have formulated equilibrium learning algorithms for MFGs (Laurière et al. 2022), including approaches based on regularization (Cui and Koeppl 2021; Laurière et al. 2022; Guo, Xu, and Zariphopoulou 2022), optimization (Guo, Hu, and Zhang 2023; Guo et al. 2023), fictitious play (Perrin et al. 2020; Geist et al. 2022) and online mirror descent (Pérolat et al. 2022; Laurière et al. 2022; Yardim et al. 2023). For less-familiar readers, we refer to the survey of Laurière et al. (2022).

So far however, most MFG learning frameworks remain unable to handle common noise (Carmona, Delarue, and Lacker 2016), or more generally major players. Contrary to minor players, a major player directly affects all minor players and is affected by the MF of minor players, whereas common noise also affects all minor players, but is exogeneous and can be understood as a static major player without actions (Huang and Wang 2016). Notably, Perrin et al. (2020) formulate an algorithm handling common noise using a continuous learning Lyapunov argument (Harris 1998; Hofbauer and Sandholm 2002), assuming however that the common noise is known, while Cui, Fabian, and Koeppl (2023) consider a cooperative setting. Common noise and major players remain important in practice, as a system seldom consists only of many similar minor players. For example, strategic players on the market do not exist in a vacuum but must contend for instance with idiosyncratic shocks (Carmona 2020) or government regulators (Aurell et al. 2022), while many cars on a road network (Cabannes et al. 2022) may be subject to traffic accidents or traffic lights. In continuous-time, such systems are known as MFGs with major and minor players (Carmona, Delarue et al. 2018), and have been considered, e.g., by Huang (2010); Nguyen and Huang (2012); Bensoussan, Chau, and Yam (2016) for LQG systems, by (Nourian and Caines 2013; Sen and Caines 2016) in non-linear and partially observed settings, and more recently by Carmona and Zhu (2016); Carmona and Wang (2017); Lasry and Lions (2018); Cardaliaguet, Cirant,

[*]These authors contributed equally.
Copyright © 2024, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Urheberrechtlich geschützt / In copyright https://rightsstatements.org/page/InC/1.0/

and Porretta (2020). Major agents also generalize common noise, an important problem in MFG literature (Carmona, Delarue, and Lacker 2016; Perrin et al. 2020; Motte and Pham 2022). For an additional overview, we also point to Carmona, Delarue et al. (2018). In contrast to prior work, we focus on a computational learning framework that is in discrete time. Additionally, even existing discrete-time MFG frameworks with only common noise such as by Perrin et al. (2020) have to the best of our knowledge not yet rigorously connected MFGs with the finite games of practical interest. We note that another setting with major players has already been explored: Stackelberg MFGs. (Elie, Mastrolia, and Possamaï 2019; Carmona and Wang 2021; Carmona, Dayanıklı, and Laurière 2022) consider a Stackelberg equilibrium instead of a Nash equilibrium, wherein a 'major' principal player chooses their policy first and has priority (like a government or regulator); see (Guo, Hu, and Zhang 2022; Vasal and Berry 2022) for discrete time versions of the problem. Though the Stackelberg setting is of importance, it is distinct from computing Nash equilibria where major and minor players are "on the same level": in the Stackelberg setting, minor players only respond with a Nash equilibrium between themselves *after* the principal's policy choice. Furthermore, we are not aware of any propagation of chaos results even in discrete-time Stackelberg MFGs, for which our result also applies. The field of Stackelberg MFGs remains part of continued active research, to which our M3FG setting may also contribute, and vice versa.

By the preceding motivation, we propose the first general discrete-time Major-Minor MFG (M3FG) learning framework. We begin with providing a theoretical foundation of the proposed M3FG model, showing that equilibria in finite games with many players can be approximately learned in the M3FG instead. The proof is based upon showing propagation of chaos i.e., convergence of the empirical MF, which – in contrast to its counterpart in MFGs without common noise – converges only in distribution. We then move on to provide a learning algorithm based on fictitious play to solve M3FGs, with convergence results and approximation guarantees for its tractable and practical, tabular variant. Empirically, our learned policies do not assume that common noise is known a priori. Due to the resulting stochastic MF, for tractable dynamic programming we allow conditioning of player actions and policies also on the MF instead of just the player's own state. Finally, we verify the M3FG framework on three problems, empirically supporting theoretical claims, even when the assumptions are not entirely fulfilled.

Major-Minor Mean Field Games

In this section, we begin by giving a description of considered problems and their corresponding mean field system. Proofs and additional details can be found in the full preprint version (Cui et al. 2023).

Notation: Equip finite sets $\mathcal{S}$ with the discrete metric, products with the product sup *metric, and probability measures $\mathcal{P}(\mathcal{S})$ on $\mathcal{S}$ with the L_1 norm. Let $[N] \coloneqq \{1, \dots, N\}$.*

Finite Player Game

We consider a game with N minor players and one major player. Let $\mathcal{X}$ and $\mathcal{U}$ be finite state and action spaces for minor players, respectively. Let $\mathcal{X}^0$ and $\mathcal{U}^0$ be finite state and action spaces for the major player, respectively. Let $T \in \mathbb{N}$ be a finite time horizon and let $\mathcal{T} \coloneqq \{0, 1, \dots, T-1\}$. We denote the state and the action of minor player $i \in [N]$ at time $t \in \mathcal{T}$ by $x_t^{i,N}$ and $u_t^{i,N}$, respectively. Similarly, we denote by $x_t^{0,N}$ and $u_t^{0,N}$ the state and the action of the major player at time t. Let μ_0 and μ_0^0 be initial probability distributions on $\mathcal{X}$ and $\mathcal{X}^0$, respectively. Define the empirical MF $\mu_t^N \coloneqq \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{x_t^{i,N}}$, where $\mathbf{1}_x$ is the indicator function equal to 1 for the argument x and 0 otherwise. The MF can be viewed as a histogram with $|\mathcal{X}|$ many bins.

We can consider several classes of policies. In this presentation, we focus on Markovian feedback policies in the following sense: the policy $\pi^{i,N}$ for minor player i is a function of her own state, the major player's state and the MF; the policy $\pi^{0,N}$ for the major player is a function of her own state and the MF. We denote respectively by Π and Π_0 the sets of such minor and major player policies.

For a given tuple of policies $(\underline{\pi}^N, \pi^{0,N}) = ((\pi^{1,N}, \dots, \pi^{N,N}), \pi^{0,N}) \in \Pi^N \times \Pi_0$, the game begins with states $x_0^{0,N} \sim \mu_0^0$, $x_0^{i,N} \sim \mu_0$ and subsequently, for $t = 0, 1, \dots, T-2$, let

$$u_t^{i,N} \sim \pi_t^{i,N}(x_t^{i,N}, x_t^{0,N}, \mu_t^N), \quad i \in [N] \tag{1a}$$

$$u_t^{0,N} \sim \pi_t^{0,N}(x_t^{0,N}, \mu_t^N), \tag{1b}$$

$$x_{t+1}^{i,N} \sim P(x_t^{i,N}, u_t^{i,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N), \quad i \in [N] \tag{1c}$$

$$x_{t+1}^{0,N} \sim P^0(x_t^{0,N}, u_t^{0,N}, \mu_t^N). \tag{1d}$$

where $P\colon \mathcal{X} \times \mathcal{U} \times \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ and $P^0\colon \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{X})$ are transition kernels.

In contrast to classic MFGs such as studied e.g, in (Saldi, Başar, and Raginsky 2018), the minor players' dynamics depend also on the major player's state. An important consequence is that the minor players' dynamics are influenced by a form of common noise. This explains why we decide to consider policies that depend on the MF μ_t^N. Furthermore, this form of common noise is not simply an exogenous source of randomness because it is influenced by the major player's choice of policy. This makes the problem more challenging than MFGs with common noise.

Next, we define the minor and major total rewards

$$J_N^i(\underline{\pi}^N, \pi^{0,N}) = \mathbb{E}\left[\sum_{t\in\mathcal{T}} r(x_t^{i,N}, u_t^{i,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N)\right],$$

$$J_N^0(\underline{\pi}^N, \pi^{0,N}) = \mathbb{E}\left[\sum_{t\in\mathcal{T}} r^0(x_t^{0,N}, u_t^{0,N}, \mu_t^N)\right],$$

for some reward functions $r\colon \mathcal{X}\times\mathcal{U}\times\mathcal{X}^0\times\mathcal{U}^0\times\mathcal{P}(\mathcal{X}) \to \mathbb{R}$ and $r^0\colon \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$.

In this work, we focus on the non-cooperative scenario where players try to maximize their own objectives while anticipating the behavior of other players. This is formalized by the solution concept of (approximate) Nash equilibria.

Definition 1. *Let $\varepsilon \geq 0$. An approximate ε-Nash equilibrium is a tuple $(\underline{\pi}^N, \pi^{0,N}) = ((\pi^{1,N}, \dots, \pi^{N,N}), \pi^{0,N}) \in \Pi^N \times \Pi_0$ of policies, such that $J_N^0(\underline{\pi}^N, \pi^{0,N}) \geq \sup_{\tilde{\pi}^0} J_N^0((\pi^{1,N}, \dots, \pi^{N,N}), \tilde{\pi}^0) - \varepsilon$ and $J_N^i(\underline{\pi}^N, \pi^{0,N}) \geq \sup_{\tilde{\pi}^i \in \Pi} J_N^i((\pi^1, \dots, \pi^{i-1}, \tilde{\pi}^i, \pi^{i+1}, \dots, \pi^N), \pi^{0,N}) - \varepsilon$ for all $i \in [N]$. A Nash equilibrium is an approximate 0-Nash equilibrium.*

Remark 1. *We can also consider time-dependent dynamics or rewards, multiple major players, and infinite-horizon discounted objectives. Some results we prove below can be extended to such settings (e.g., propagation of chaos, equilibrium approximation, and fictitious play; see also generalized infinite-horizon experiments in Cui et al. (2023, Appendix I)). Similarly, we can extend the model to multiple minor agent populations with small changes, see e.g. Pérolat et al. (2022). Another possibility is to simply include types of players into their state (Mondal et al. 2022).*

Mean Field Game

When the number of minor players N is large, we can approximate the game by an MFG, which corresponds formally to the limit $N \to \infty$. In an MFG, the empirical MF is replaced by a random limiting MF. Unlike standard MFGs, the limiting MF does not evolve in a deterministic way due to the influence of the major player. Fixing major and minor player policies π^0, π for all players, except for a single minor player deviating to $\hat{\pi}$, when $N \to \infty$, we obtain (intuitively by a law of large numbers argument) the major and deviating minor player M3FG dynamics $x_0^0 \sim \mu_0^0$, $x_0 \sim \mu_0$,

$$u_t \sim \hat{\pi}_t(x_t, x_t^0, \mu_t), \tag{2a}$$

$$u_t^0 \sim \pi_t^0(x_t^0, \mu_t), \tag{2b}$$

$$x_{t+1} \sim P(x_t, u_t, x_t^0, u_t^0, \mu_t), \tag{2c}$$

$$x_{t+1}^0 \sim P^0(x_t^0, u_t^0, \mu_t), \tag{2d}$$

$$\mu_{t+1} = T_t^\pi(x_t^0, u_t^0, \mu_t) \tag{2e}$$

with the deterministic transition operator $T_t^\pi(x^0, u^0, \mu) \coloneqq \iint P(x, u, x^0, u^0, \mu)\, \pi_t(\mathrm{d}u \mid x, x^0, \mu)\, \mu(\mathrm{d}x)$ as the conditional "expectation" of the next MF given the current major state x^0, action u^0, and random MF μ. The policy π is shared by all minor players except one who is deviating and using $\hat{\pi}$. This means that we look for symmetric Nash equilibria where all exchangeable minor players use the same policy, as usual in MFG literature. Still, a mean field equilibrium suffices as an approximate Nash equilibrium in the finite game, which is not to say that there cannot be other heterogeneous policy tuples in the finite game that are Nash.

M3FGs now consist of *two* Markov decision process (MDP) optimality conditions, one for all minor players and one for the major player. An equilibrium is then optimal in each MDP simultaneously. More precisely, from the point of view of a minor player, the goal is to optimize over $\hat{\pi}$ while (π, π^0) are fixed. This yields the minor player MDP with state $(x_t, x_t^0, \mu_t) \in \mathcal{X} \times \mathcal{X}^0 \times \mathcal{P}(\mathcal{X})$, and action $u_t \in \mathcal{U}$, and with the objective

$$J(\hat{\pi}, \pi, \pi^0) = \mathbb{E}\left[\sum_{t \in \mathcal{T}} r(x_t, u_t, x_t^0, u_t^0, \mu_t)\right]. \tag{3}$$

Note that, although μ_{t+1} is given by a deterministic function of (x_t^0, u_t^0, μ_t), from the point of view of a minor player, the evolution of $(\mu_t)_t$ is stochastic since it depends on the sequence $(x_t^0, u_t^0)_t$, which is random. By definition of a Nash equilibrium, only a *single* minor player can deviate arbitrarily to $\hat{\pi}$, and by symmetry it does not matter which "representative" minor player deviates. Therefore there is only one MDP optimality condition for all minor players. We also stress that since $N \to \infty$, the representative player is insignificant and her deviation does not affect the mean field.

On a similar note, from the major player's point of view, we obtain the major player MDP with $(\mathcal{X}^0 \times \mathcal{P}(\mathcal{X}))$-valued states (x_t^0, μ_t) and $\mathcal{U}^0$-valued actions u_t^0 of the major player, using the same dynamics, forgetting about the (insignificant for the major player) deviating minor player, and optimizing instead for π^0, the corresponding major objective

$$J^0(\pi, \pi^0) = \mathbb{E}\left[\sum_{t \in \mathcal{T}} r^0(x_t^0, u_t^0, \mu_t)\right]. \tag{4}$$

Mean field equilibrium. The Nash equilibrium in the finite game hence corresponds to a major-minor mean field equilibrium, as a fixed point of both MDPs *at once*. In other words, major and minor policies π^0, π that are optimal against themselves in the major and minor player MDPs.

Definition 2. *A Major-Minor Mean Field Nash Equilibrium (M3FNE) is a tuple $(\pi, \pi^0) \in \Pi \times \Pi_0$ of policies, such that $\pi \in \arg\max J(\cdot, \pi, \pi^0)$ and $\pi^0 \in \arg\max J^0(\pi, \cdot)$.*

We slightly weaken the concept of optimality to *approximate* optimality, since the solution of a limiting MFG provides approximate Nash equilibria for the finite game, which are still achieved by solving for approximate M3FNE.

Definition 3. *An approximate ε-M3FNE is a tuple $(\pi, \pi^0) \in \Pi \times \Pi_0$ of policies, such that $J(\pi, \pi, \pi^0) \geq \sup J(\cdot, \pi, \pi^0) - \varepsilon$ and $J^0(\pi, \pi^0) \geq \sup J^0(\pi, \cdot) - \varepsilon$.*

The minimal such ε for minor and major agents are also referred to as the minor and major exploitabilities $\mathcal{E}(\pi, \pi^0)$ and $\mathcal{E}^0(\pi, \pi^0)$ of (π, π^0). Accordingly, an exploitability of 0 means that (π, π^0) is an exact M3FNE.

Theoretical Analysis

The M3FG is a theoretically rigorous formulation for large corresponding finite games. Note in particular that the MF will be stochastic due to the randomness of major players and their states, and therefore standard results based on determinism of MFs will no longer hold. We provide such a theoretical foundation of M3FG by propagation of chaos.

Continuity assumptions. We provide theoretical guarantees to prove that the M3FNE is an approximate Nash equilibrium in the finite game, despite having a non-deterministic MF in the limiting case, contrary to most of the existing literature (Huang et al. 2006; Guo et al. 2019). For this, we need some common Lipschitz continuity assumptions (Gu et al. 2021; Pásztor, Krause, and Bogunovic 2023).

Assumption 1. *The kernels P, P^0 are L_P, L_{P^0}-Lipschitz.*

Assumption 2. *The rewards r, r^0 are L_r, L_{r^0}-Lipschitz.*

Assumption 3. *The classes of major and minor policies Π^0, Π are equi-Lipschitz, i.e. there are $L_{\Pi^0}, L_\Pi > 0$ s.t. for all t, $\pi^0 \in \Pi^0$, $\pi \in \Pi$, we have that $\pi_t^0 \colon \mathcal{X}^0 \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{U})$ and $\pi_t \colon \mathcal{X} \times \mathcal{X}^0 \times \mathcal{P}(\mathcal{X}) \to \mathcal{P}(\mathcal{U})$ are L_{Π^0}, L_Π-Lipschitz.*

Here, we always consider Lipschitz continuity for all arguments using the sup metric for products, and the L_1 distance for probability measures, see e.g., Cui et al. (2023, Appendix B). We note that the Lipschitz assumption for policies – while standard – is technical. Empirically, the only piecewise Lipschitz policies obtained in Section for tractability nonetheless remain close to the following approximations in the finite system. A theoretical investigation of guarantees for piecewise Lipschitz policies is left for future work.

Propagation of chaos. We achieve propagation of chaos "in distribution" for major and minor players to the M3FG at rate $\mathcal{O}(1/\sqrt{N})$, which is shown inductively in Cui et al. (2023, Appendix F). Here, propagation of chaos refers to the conditional independence of minor agents, and thus convergence in the limit to the deterministic mean field (Chaintron and Diez 2022). In contrast to MFGs with deterministic MFs, a stronger mode of convergence such as the one considered by Saldi, Başar, and Raginsky (2018) fails by stochasticity of the MF.

Theorem 1. *Consider Assumptions 1 and 3, and any equi-Lipschitz family of functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}\times\mathcal{U}\times\mathcal{X}^0\times\mathcal{U}^0\times\mathcal{P}(\mathcal{X})}$ with shared Lipschitz constant $L_\mathcal{F}$. Then, the random variable $(x_t^{1,N}, u_t^{1,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N)$ in system (1) under $((\hat{\pi}, \pi, \pi, \ldots), \pi^0)$ converges weakly, uniformly over $f \in \mathcal{F}$ and $(\hat{\pi}, \pi, \pi^0) \in \Pi\times\Pi\times\Pi^0$, to $(x_t, u_t, x_t^0, u_t^0, \mu_t)$ in system (2) under $(\hat{\pi}, \pi, \pi^0)$,*

$$\forall t \in \mathcal{T},\ \sup_{\hat{\pi},\pi,\pi^0} \sup_{f\in\mathcal{F}} \Big| \mathbb{E}\left[f(x_t^{1,N}, u_t^{1,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N)\right] - \mathbb{E}\left[f(x_t, u_t, x_t^0, u_t^0, \mu_t)\right] \Big| = \mathcal{O}(1/\sqrt{N}). \quad (5)$$

Corollary 1. *Similarly, consider Assumptions 1 and 3, and any family of equi-Lipschitz functions $\mathcal{F}^0 \subseteq \mathbb{R}^{\mathcal{X}^0\times\mathcal{U}^0\times\mathcal{P}(\mathcal{X})}$ with shared Lipschitz constant $L_{\mathcal{F}^0}$. Then the random variable $(x_t^{0,N}, u_t^{0,N}, \mu_t^N)$ in system (1) under $((\hat{\pi}, \pi, \pi, \ldots), \pi^0)$ converges weakly, uniformly over $f \in \mathcal{F}^0$, to (x_t^0, u_t^0, μ_t) in system (2) under $(\hat{\pi}, \pi, \pi^0)$,*

$$\forall t \in \mathcal{T},\ \sup_{\hat{\pi},\pi,\pi^0} \sup_{f\in\mathcal{F}^0} \Big| \mathbb{E}\left[f(x_t^{0,N}, u_t^{0,N}, \mu_t^N)\right] - \mathbb{E}\left[f(x_t^0, u_t^0, \mu_t)\right] \Big| = \mathcal{O}(1/\sqrt{N}). \quad (6)$$

Approximate Nash equilibrium. From propagation of chaos, the approximate Nash property of M3FNE follows, suggesting that a solution of M3FGs provides a good game-theoretic solution of interest to practical N-player games, see Cui et al. (2023, Appendix G) for the proof based on propagation of chaos.

Corollary 2. *Consider Assumptions 1, 2, 3, and a M3FNE $(\pi, \pi^0) \in \Pi \times \Pi^0$. Then, the policies $((\pi, \ldots, \pi), \pi^0)$ constitute an $\mathcal{O}(1/\sqrt{N})$-Nash equilibrium in the finite game.*

Finally, existence of a M3FNE is a difficult question under policies that depend on the stochastic MF. While assuming reactive policies unconditioned on the MF could help, choosing such policies makes the design of our algorithm based on dynamic programming difficult, as policies computed via dynamic programming need to depend on the entire M3FG system state. In contrast, in usual deterministic MFGs it is sufficient to remove policy dependence on the MF, which is deterministic. For practical purposes, learning equilibria and then checking the exploitability by Theorem 3 may suffice.

Fictitious Play

To find M3FNE and solve the fixed-point problem, we formulate a fictitious play (FP) algorithm and provide a theoretical analysis. Following the exact algorithm, as empirical contribution we provide and analyze an approximate, numerically tractable algorithm that does not assume knowledge of common noise, contrary to Perrin et al. (2020), and extend it to the setup with major and minor players. Since the space of MFs is continuous and does not allow general exact computation of value functions, we project MFs onto a finite partition with guarantees for policy evaluation.

Fictitious Play for M3FNE

In order to learn an M3FNE, we first propose an exact analytic algorithm based on FP (Perrin et al. 2020) and provide a theoretical analysis of convergence. For this part, we will assume that the major player's action does not affect the minor players' transition kernel. To simplify the presentation and the analysis, we will use conditioning with respect to the sources of randomness that affect the MF, i.e., the minors' distribution. For every $t \geq 0$, let the major and minor players' actions be determined not by the mean field μ_t, but instead by the history of major states and actions, $u_t^0 \sim \pi_t^0(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)$, $x_{0:t-1}^0 := (x_0^0, x_1^0, \ldots, x_{t-1}^0)$, $u_{0:t-1}^0 := (u_0^0, u_1^0, \ldots, u_{t-1}^0)$. By induction, we can in fact view μ_t as a deterministic function of $(x_{0:t-1}^0, u_{0:t-1}^0)$ given the minor players' policy π, since we simply have $\mu_{t+1} = T_t^\pi(x_t^0, u_t^0, \mu_t)$ recursively and deterministically. This means that for fixed policies such as a given Nash equilibrium, any policies dependent on μ_t can instead be rewritten as functions of $(x_{0:t-1}^0, u_{0:t-1}^0)$. Therefore, instead of seeing policies as functions of μ_t, we will see them as functions of the major player randomness $(x_{0:t-1}^0, u_{0:t-1}^0)$ and we will write (slightly abusing notation) $\pi_t(x_t, x_{0:t}^0, u_{0:t-1}^0)$ and $\pi_t^0(x_t^0, x_{0:t-1}^0, u_{0:t-1}^0)$ respectively for the minor players' and the major player's policies. The results we prove below go beyond existing results by (i) analyzing also the major exploitability similarly to the minor exploitability, and (ii) expanding analysis of minor exploitability under presence of major players. To this end, we formulate Assumption 4.(c) and 4.(d), which provide the conditions for convergence in the presence of major players. See Cui et al. (2023, Appendix A) for more detail.

We first start by introducing the discrete time FP before analyzing it in continuous time. Here, time refers to the algorithm's current iteration and not to the time of the M3FG

system, which remains discrete throughout the whole paper. At any given step j of FP, we have:

$$\mu^{\bar{\pi}^j}_{t|x^0_{0:t-1},u^0_{0:t-1}} = \frac{j-1}{j}\mu^{\bar{\pi}^{j-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}} + \frac{1}{j}\mu^{\pi^{BR,j}}_{t|x^0_{0:t-1},u^0_{0:t-1}} \tag{7}$$

where we use the notation $\mu^{\pi}_{t|x^0_{0:t-1},u^0_{0:t-1}}$ for the minor state distribution at time t induced by the minor agent policy π and conditioned on the past sequence $(x^0_{0:t-1}, u^0_{0:t-1})$. Here, $\mu^{\pi^{BR,j}}_{t|x^0_{0:t-1},u^0_{0:t-1}}$ is the conditional distribution induced by the best response (BR) policy $\pi^{BR,j}$ against $\bar{\pi}^{j-1}$ and $\bar{\pi}^{0,j-1}$, i.e., $\pi^{BR,j} := \arg\max_\pi J(\pi, \bar{\pi}^{j-1}, \bar{\pi}^{0,j-1})$. The policy generating this average distribution is

$$\bar{\pi}^j_t(u|x, x^0_{0:t-1}, u^0_{0:t-1}) = \frac{\sum_{i=0}^j \mu^{\pi^{BR,i}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x)\pi^{BR,i}_t(u|x, x^0_{0:t-1}, u^0_{0:t-1})}{\sum_{i=0}^j \mu^{\pi^{BR,i}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x)}. \tag{8}$$

Meanwhile, the major player state distribution is

$$\mu^{\bar{\pi}^{0,j}}_t = \frac{j-1}{j}\mu^{\bar{\pi}^{0,j-1}}_t + \frac{1}{j}\mu^{\pi^{0,BR,j}}_t$$

where $\bar{\pi}^{0,j}_t$ analogous to (8), but in contrast to minor agents using joint distributions $\mu^{\pi^{0,BR,i}}_t(x^0_t, x^0_{0:t-1}, u^0_{0:t-1})$ and $\pi^{0,BR,i}_t(u^0_t \mid x^0_t, x^0_{0:t-1}, u^0_{0:t-1})$.

For the convergence analysis, we study the continuous time version of above discrete time FP, as Perrin et al. (2020). In the continuous time FP algorithm, we denote the time of the algorithm (its "iterations") with τ and we first initialize the algorithm for $\tau < 1$ with arbitrary policies for the minor players, $\bar{\pi}^{\tau<1} = \{\bar{\pi}^{\tau<1}_t\}_{t\in\mathcal{T}}$, and major player, $\bar{\pi}^{0,\tau<1} = \{\bar{\pi}^{0,\tau<1}_t\}_{t\in\mathcal{T}}$. For all $\tau \geq 1$, $t \in \mathcal{T}$ and $x^0_{0:t-1}, u^0_{0:t-1}$, define the FP process

$$\begin{aligned}\bar{\mu}^\tau_{t|x^0_{0:t-1},u^0_{0:t-1}} &= \frac{1}{\tau}\int_0^\tau \mu^{\pi^{BR,s}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}} ds \\ \bar{\mu}^{0,\tau}_t &= \frac{1}{\tau}\int_0^\tau \mu^{\pi^{0,BR,s}_{0:t-1}}_t ds\end{aligned} \tag{9}$$

where $\mu^{\pi^{BR,\tau}_{0:t-1}}_{t|x^0_{0:t-1},u^0_{0:t-1}}$ and $\mu^{\pi^{0,BR,\tau}_{0:t-1}}_t$ are conditional and joint distributions respectively, induced by the BR policies $\pi^{BR,\tau}$ and $\pi^{0,BR,\tau}$ up to time $t-1$ against $\mu^{\bar{\pi}^\tau}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x)$ and $\mu^{\bar{\pi}^{0,\tau}}_t(x^0_t, x^0_{0:t-1}, u^0_{0:t-1})$. In other words, $\pi^{BR,\tau} := \arg\min_\pi J(\pi, \bar{\pi}^\tau, \bar{\pi}^{0,\tau})$ and $\pi^{0,BR,\tau} := \arg\min_{\pi^0} J^0(\bar{\pi}^\tau, \pi^0)$.

Note that the distributions induced by the averaged policies $\{\bar{\pi}^\tau_t\}_{t\in\mathcal{T}}$ and $\{\bar{\pi}^{0,\tau}_t\}_{t\in\mathcal{T}}$ for $\tau \geq 1$ are given as

$$\begin{aligned}&\bar{\pi}^\tau_t(u|x, x^0_{0:t-1}, u^0_{0:t-1})\int_{s=0}^\tau \mu^{\pi^{BR,s}}_{t|x^0_{0:t-1},u^0_{0:t-1}}(x)ds \\ &= \int_{s=0}^\tau \mu^{\pi^{BR,s}}_{t,|x^0_{0:t-1},u^0_{0:t-1}}(x)\pi^{BR,s}_t(u|x, x^0_{0:t-1}, u^0_{0:t-1})ds, \\ &\bar{\pi}^{0,\tau}_t(u^0|x^0, x^0_{0:t-1}, u^0_{0:t-1})\int_{s=0}^\tau \mu^{\pi^{0,BR,s}}_t(x^0, x^0_{0:t-1}, u^0_{0:t-1})ds \\ &= \int_{s=0}^\tau \mu^{\pi^{0,BR,s}}_t(x^0, x^0_{0:t-1}, u^0_{0:t-1}) \\ &\qquad \cdot \pi^{0,BR,s}_t(u^0|x^0, x^0_{0:t-1}, u^0_{0:t-1})ds,\end{aligned} \tag{10}$$

for all $t \in \mathcal{T}$ and $x^0_{0:t-1}, u^0_{0:t-1}$. For $s < 1$, $\pi^{BR,s}$ and $\pi^{0,BR,s}$ are chosen arbitrarily. The proof and the differential form of equations (9) and (10) can be found in Cui et al. (2023, Appendix A).

As a result, below we give a convergence analysis together with assumptions for continuous time FP, converging in both minor and major exploitability $\mathcal{E}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) = \max_{\pi'} J(\pi', \bar{\pi}^\tau, \bar{\pi}^{0,\tau}) - J(\bar{\pi}^\tau, \bar{\pi}^\tau, \bar{\pi}^{0,\tau})$, $\mathcal{E}^0(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) = \max_{\pi^{0\prime}} J^0(\bar{\pi}^\tau, \pi^{0\prime}) - J^0(\bar{\pi}^\tau, \bar{\pi}^{0,\tau})$, summarized as the total exploitability $\mathcal{E}_{\text{tot}}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) = \mathcal{E}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) + \mathcal{E}^0(\bar{\pi}^\tau, \bar{\pi}^{0,\tau})$.

Assumption 4. *(a) The transition kernels are in the form of $P(x_{t+1} \mid x_t, u_t, x^0_t, u^0_t)$ and $P^0(x^0_{t+1} \mid x^0_t, u^0_t)$ for minor players and major player, respectively.*

(b) The reward of minor and major players are separable, i.e. for some reward functions $\tilde{r}, \bar{r}, \tilde{r}^0, \hat{r}^0, \check{r}^0$, we have

$$r(x, u, x^0, u^0, \mu) = \tilde{r}(x, x^0, u) + \bar{r}(x, x^0, \mu),$$
$$r^0(x^0, u^0, \mu) = \tilde{r}^0(x^0, u^0) + \bar{r}^0(x^0, \mu).$$

(c) The game is monotone; i.e., satisfies Lasry-Lions monotonicity condition: For minor players, we have $\forall x^0 \in \mathcal{X}^0$, $\forall \mu, \mu'$: $\sum_{x\in\mathcal{X}}(\mu(x) - \mu'(x))(\bar{r}(x, x^0, \mu) - \bar{r}(x, x^0, \mu')) \leq 0$. Meanwhile, for major players, we have $\frac{d}{d\tau}\mu^{\bar{\pi}^{0,\tau}_{0:t}}_{t+1}(x^0_{t+1}, x^0_{0:t}, u^0_{0:t}) \cdot \left\langle \nabla_\mu \bar{r}^0(x^0_{t+1}, \mu^{\bar{\pi}^\tau_{0:t}}_{t+1|x^0_{0:t},u^0_{0:t}}), \frac{d}{d\tau}\mu^{\bar{\pi}^\tau_{0:t}}_{t+1|x^0_{0:t},u^0_{0:t}} \right\rangle \leq 0$.

(d) We have $\tilde{\mathcal{E}}(\bar{\pi}^\tau, \pi^{0,BR,\tau}, \bar{\pi}^{0,\tau}) \leq \mathcal{E}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau})$, where we define $\tilde{\mathcal{E}}(\bar{\pi}^\tau, \pi^{0,BR,\tau}, \bar{\pi}^{0,\tau}) = J(\pi^{BR,\tau}, \bar{\pi}^\tau, \pi^{0,BR,\tau}) - J(\bar{\pi}^\tau, \bar{\pi}^\tau, \pi^{0,BR,\tau})$ with any BR policy given as $\pi^{BR,\tau} = \arg\max_\pi J(\pi, \bar{\pi}^\tau, \bar{\pi}^{0,\tau})$.

Remark 2. *Assumption 4.(c) is fulfilled for major players if $\bar{r}^0(x^0, \mu) = \bar{r}^0(x^0)$. Assumption 4.(d) is satisfied for instance if $r(x, u, x^0, \mu) = r(x, u, \mu)$ and $P(x_{t+1} \mid x_t, u_t, x^0_t, u^0_t) = P(x_{t+1} \mid x_t, u_t)$. Then, we trivially have $\tilde{\mathcal{E}}(\bar{\pi}^\tau, \pi^{0,BR,\tau}, \bar{\pi}^{0,\tau}) = \mathcal{E}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau})$ by obtaining a minor player MFG independent of the major player.*

Theorem 2. *Under Assumption 4, the total exploitability is a strong Lyapunov function such that $\frac{d}{d\tau}\mathcal{E}_{\text{tot}}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) \leq -\frac{1}{\tau}\mathcal{E}_{\text{tot}}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau})$; i.e., we have $\mathcal{E}_{\text{tot}}(\bar{\pi}^\tau, \bar{\pi}^{0,\tau}) = \mathcal{O}(1/\tau)$ in the continuous time FP algorithm.*

The proof of Theorem 2 can be found in Cui et al. (2023, Appendix A) and is based on a monotonic decrease of exploitability, at the same rate as standard FP in MFGs (Perrin et al. 2020).

In numerical experiments, for applicability and computational tractability (due to the exponential complexity of the histories in the horizon), we condition policies on the random MF and major state instead of the histories, averaging policies uniformly instead of for each possible major state-action sequence. Further, numerically we partition and represent the (naturally continuous) MFs as described in the following, to obtain tabular Algorithm 1. Experimentally, in Section we nonetheless find that the algorithm optimizes exploitability, even if Assumption 4 is not fully satisfied. The dependence of policy actions on the MF and major state has the additional advantage of allowing standard dynamic programming for major and minor MDPs, as their full MDP states include both the MF and major state.

Projected Mean Field

Observe that for given current MF and major state-actions, we obtain deterministic transitions from one MF to the next. Therefore, by partitioning we can obtain deterministic transitions in-between parts of a partition of $\mathcal{P}(\mathcal{X})$, and a Bellman equation over *finite* spaces.

Definition 4. *A δ-partition $\mathcal{M} = \{\mathcal{P}_i\}_{i\in[|\mathcal{M}|]}$ is a partition of $\mathcal{P}(\mathcal{X})$, with $\|\mu - \nu\| < \delta$ for any $i \in [|\mathcal{M}|]$, $\mu, \nu \in \mathcal{P}_i$.*

Since $\mathcal{P}(\mathcal{X})$ is compact, a finite δ-partition of $\mathcal{P}(\mathcal{X})$ exists for any $\delta > 0$. We will henceforth assume for any $\delta > 0$ some δ-partition $\mathcal{M}$ of $\mathcal{P}(\mathcal{X})$ with $M = M(\delta)$ parts.

Discretized finite MDPs. To each part $\mathcal{P}_i$, we associate an arbitrary element $\hat{\mu}^{(i)} \in \mathcal{P}_i$ and write $\mathrm{proj}_\delta \mu$ for the δ-partition projection of MFs $\mu \in \mathcal{P}(\mathcal{X})$, i.e. whenever $\mu \in \mathcal{P}_i$ we project to the representative $\mathrm{proj}_\delta \mu = \hat{\mu}^{(i)} \in \mathcal{P}_i$.

As a result, we obtain discretized, *finite* MDP versions of the major and minor player MDPs, where the continuous MF state is replaced by finitely many states in $\hat{\mathcal{P}}(\mathcal{X}) \coloneqq \{\hat{\mu}^{(1)}, \dots, \hat{\mu}^{(M)}\}$, evolving by discretized MF evolutions in (2), i.e. $\hat{\mu}_{t+1} = \mathrm{proj}_\delta T_t^\pi(x^0, u^0, \hat{\mu}_t)$ for any $x^0, u^0, \hat{\mu}_t$.

We can solve the discretized MDPs in a tabular manner: To compute best responses under policies (π, π^0), observe that the true action-value function Q^0_{π,π^0} of the (not discretized) major player MDP follows the Bellman equation

$$Q^0_{\pi,\pi^0}(t, x^0, u^0, \mu) = r^0(x^0, u^0, \mu) + \sum_{x^{0\prime}} P^0(x^{0\prime} \mid x^0, u^0, \mu) \cdot \max_{u^{0\prime}} Q^0_{\pi,\pi^0}(t+1, x^{0\prime}, u^{0\prime}, T_t^\pi(x^0, u^0, \mu)).$$

The tabular approximate action-value function $\hat{Q}^0_{\pi,\pi^0}$ for the major player follows instead the Bellman equation of the discretized major player MDP (letting the domain of $\hat{Q}_{\pi,\pi^0}$ be the entirety of $\mathcal{P}(\mathcal{X})$ as constants over each part $\mathcal{P}_i$),

$$\begin{aligned}\hat{Q}^0_{\pi,\pi^0}(t, x^0, u^0, \mu) &= \hat{Q}^0_{\pi,\pi^0}(t, x^0, u^0, \mathrm{proj}_\delta \mu) \\ &= r^0(x^0, u^0, \mathrm{proj}_\delta \mu) + \sum_{x^{0\prime}} P^0(x^{0\prime} \mid x^0, u^0, \mathrm{proj}_\delta \mu) \\ &\quad \cdot \max_{u^{0\prime}} \hat{Q}^0_{\pi,\pi^0}(t+1, x^{0\prime}, u^{0\prime}, T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \mu))\end{aligned}$$

with terminal condition zero, and the minor action-values analogously. The above can nonetheless provide a good approximation that can be computed in *tabular* form, see Cui et al. (2023, Appendix C) and empirical support in Section .

Algorithm 1: Discrete-time, projected fictitious play

1: Input: δ-partition $\{\mathcal{P}_i\}_{i=1\dots,M}$.
2: Initialize initial policies $\bar{\pi}_{(0)}$, $\bar{\pi}^0_{(0)}$.
3: **for** iteration $n = 0, 1, 2, \dots$ **do**
4: Compute discretized BR (as in Definition 5)

$$\pi_{(n+1)} \in \arg\max \hat{Q}_{\bar{\pi}_{(n)}, \bar{\pi}^0_{(n)}},$$
$$\pi^0_{(n+1)} \in \arg\max \hat{Q}^0_{\bar{\pi}_{(n)}, \bar{\pi}^0_{(n)}}.$$

5: Compute next average policies

$$\bar{\pi}_{(n+1)} \coloneqq \frac{n}{n+1}\bar{\pi}_{(n)} + \frac{1}{n+1}\pi_{(n+1)},$$
$$\bar{\pi}^0_{(n+1)} \coloneqq \frac{n}{n+1}\bar{\pi}^0_{(n)} + \frac{1}{n+1}\pi^0_{(n+1)}.$$

6: **end for**

Discretized equilibria. Building upon the preceding approximations, we define an approximate equilibrium as a fixed point of the discretized system.

Definition 5. *A δ-partition M3FNE is a tuple $(\pi, \pi^0) \in \hat{\Pi} \times \hat{\Pi}^0$ with $\pi \in \arg\max \hat{Q}_{\pi,\pi^0}$ and $\pi^0 \in \arg\max \hat{Q}^0_{\pi,\pi^0}$ where policies in $\hat{\Pi}, \hat{\Pi}^0$ are instead defined as blockwise constant over each part $\mathcal{P}_i$ of the δ-partition.*

Here, we understand $\hat{\pi} \in \arg\max \hat{Q}_{\pi,\pi^0}$ by the defining equation $\sum_{u \in \arg\max_{u'} \hat{Q}_{\pi,\pi^0}(t,x,u',x^0,\hat{\mu})} \hat{\pi}_t(x, x^0, \hat{\mu}, u) = 1$ for all $(t, x, x^0, \hat{\mu}) \in \mathcal{T} \times \mathcal{X} \times \mathcal{X}^0 \times \hat{\mathcal{P}}(\mathcal{X})$, and similarly for major players, noting that $\hat{\pi}$ optimizes the preceding discretized finite MDP (Hernández-Lerma and Lasserre 2012).

We note that while the discretized solutions only piecewise fulfill Assumption 3 by not being Lipschitz, in Section we empirically find that the approximation of finite games and exploitability can nonetheless be accurate.

Approximation guarantees. We evaluate solutions by tabular evaluation in the discretized MDP, for which we are able to obtain theoretical guarantees for evaluating the true exploitability via the approximate tabular exploitability. Under a δ-partition, define the major approximate objective

$$\hat{J}^0(\pi, \pi^0) \coloneqq \sum_{x^0} \mu^0_0(x^0) \hat{V}^{0,\pi^0}_{\pi,\pi^0}(0, x^0, \mu_0)$$

and approximate exploitability

$$\hat{\mathcal{E}}^0(\pi, \pi^0) \coloneqq \sum_{x^0} \mu^0_0(x^0) \cdot \Big(\max_{\hat{\pi}^{0\prime} \in \hat{\Pi}} \hat{V}^{0,\hat{\pi}^{0\prime}}_{\pi,\pi^0}(0, x^0, \mu_0)) - \hat{V}^{0,\pi^0}_{\pi,\pi^0}(0, x^0, \mu_0) \Big),$$

with approximate values $\hat{V}^{0,\hat{\pi}^0}_{\pi,\pi^0}$ of major deviation under (π, π^0) to $\hat{\pi}^0$, following the "discretized" Bellman equation

$$\begin{aligned}\hat{V}^{0,\hat{\pi}^0}_{\pi,\pi^0}(t, x^0, \mu) = \sum_{u^{0\prime}} \hat{\pi}^0_t(u^{0\prime} \mid x^{0\prime}, \mathrm{proj}_\delta \mu) \Big[r^0(x^0, u^0, \mathrm{proj}_\delta \mu) + \sum_{x^{0\prime}} P^0(x^{0\prime} \mid x^0, u^0, \mathrm{proj}_\delta \mu) \\ \hat{V}^{0,\hat{\pi}^0}_{\pi,\pi^0}(t+1, x^{0\prime}, u^{0\prime}, T_t^\pi(x^0, u^0, \mathrm{proj}_\delta \mu)) \Big],\end{aligned}$$

and similarly for the minor player. Note that only for the major player, π^0 is irrelevant (replaced by $\hat{\pi}^0$). In other words, we approximate values and exploitability via the discretized finite MDPs, which has the advantage of enabling dynamic programming (backwards induction, value iteration).

By analyzing the value functions under continuity, we show in Cui et al. (2023, Appendix H) that these approximations are generally close to the true objectives and exploitabilities respectively, as the discretization becomes sufficiently fine.

Theorem 3. *Under Assumptions 1, 2, 3, as $\delta \to 0$, approximate minor and major values tend to the exact values, and approximate exploitabilities tend to the exact exploitabilities, at rate $\mathcal{O}(\delta)$ uniformly over $(\pi, \pi^0) \in \Pi \times \Pi^0$.*

Experiments

We evaluate FP by comparing against fixed-point iteration (FPI), which iterates discretized best response policies. For reproducibility, note that the algorithms used are deterministic, and details can be found in Cui et al. (2023, Appendix I). For code, see https://github.com/tudkcui/M3FG-learning.

Problems

For the evaluation, we use the following problem instances for exemplary, practically applicable M3FG scenarios.

SIS epidemics control. The SIS problem is an epidemics control scenario, where each individualistic minor player may decide whether to take costly preventative actions against becoming infected at a rate proportional to the proportion of infected. The major player (e.g. government) is responsible for the well-being of minor players, and can encourage preventative actions, while its state models random low- and high-infectivity seasons. The finite time horizon can be considered the time until a cure is found. The original problem without major players has been used as a benchmark for MFG learning (Cui and Koeppl 2021; Laurière et al. 2022).

Buffet problem. In the Buffet problem, we consider the following scenario: At a conference with multiple buffet locations, players desire to be at locations that are filled with food and uncrowded. However, each location depletes faster with increasing number of players. The major player (caterer) must keep buffets full and equally filled. The Buffet problem fulfills most assumptions (except Assm. 4.(d)) and shows accordingly stable FP learning.

Advertisement duopoly model. Lastly, in the advertisement model, a regulator sets the price of advertisement. Depending on the regulator's state and price of advertisement, two companies exogeneously decide on advertisement efficiencies of their subscription service. Minor players are consumers and choose whether to change to subscriptions for the better-funded product, while the regulator avoids formation of a monopoly. Duopoly advertisement competition in a static MFG was modeled in (Carmona and Dayanıklı 2021).

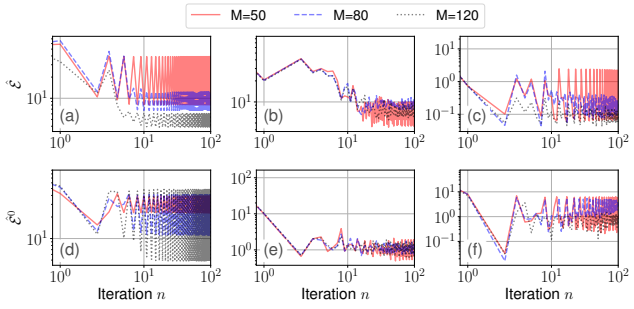

Figure 1: The approximate exploitability oscillates over iterations of FPI. (a, d): SIS, (b, e): Buffet, (c, f): Advertisement. (a-c): Minor exploitability, (d-f): major exploitability.

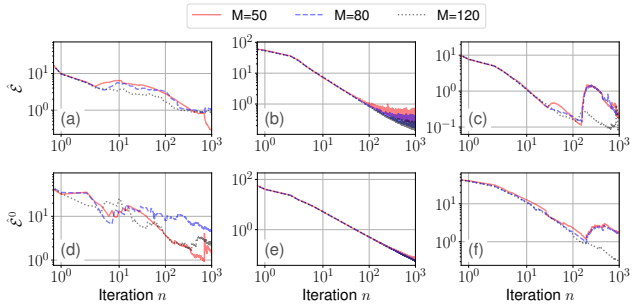

Figure 2: The approximate exploitability is optimized via FP. (a, d): SIS, (b, e): Buffet, (c, f): Advertisement. (a-c): Minor exploitability, (d-f): major exploitability.

Numerical Results

In the following, we provide a numerical evaluation via exploitability as the primary metric of interest, since it describes the quality of achieved equilibria. Additional experiments and parameter details are shown in Cui et al. (2023, Appendix I), including more qualitative results, the effect of alternative initializations, and analogous results for infinite-horizon discounted objectives. Beyond supporting the theoretical results, we also ablate both convergence assumptions for the algorithm and the Lipschitz policy assumption for propagation of chaos in the finite player system.

Exploitability convergence. As observed in Figure 1, naive FPI usually fails to converge and runs into limit cycles, motivating FP. In Figure 2, we see that the proposed FP algorithm optimizes both approximate major and minor exploitabilities $\hat{\mathcal{E}}$, $\hat{\mathcal{E}}^0$ over its iterations. Especially for Buffet, which fulfills most of Assumption 4, learning is smooth and exploitability descends monotonically as in Theorem 2, while exploitability is nevertheless optimized in the other problems. Overall, the proposed FP algorithm improves achieved exploitabilities significantly over FPI.

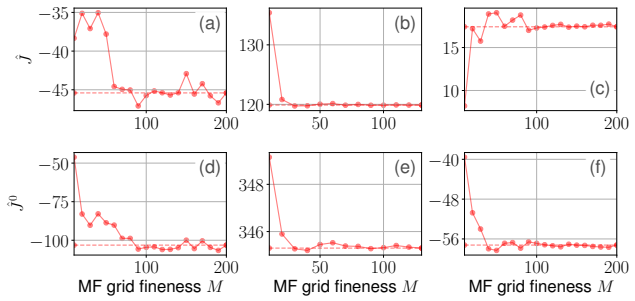

Figure 3: The final objectives of FP under discretization (dashed: right-most entry) are stable with high discretization. (a, d): SIS, (b, e): Buffet, (c, f): Advertisement. (a-c): Minor exploitability, (d-f): major exploitability.

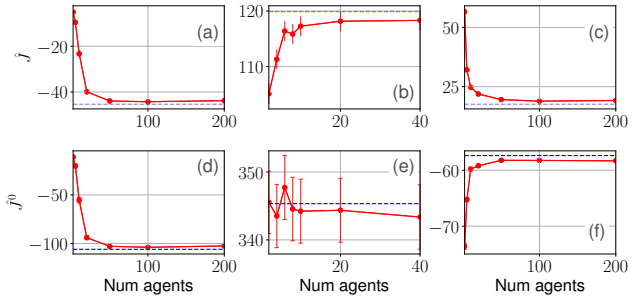

Figure 4: The mean N-player objective (red) over 1000 (or 5000 for Buffet) episodes, with 95% confidence interval, against MF predictions $\hat{J}$, $\hat{J}^0$ for FP and $M = 120$ (blue, dashed). (a, d): SIS, (b, e): Buffet, (c, f): Advertisement.

Stability over discretization. Comparing approximation results empirically over discretization bins M per dimension, i.e. using δ-partitions with $\delta \approx \frac{2}{M}$, in Figure 3 we observe that the FP-learned policies quickly stabilize as the discretization becomes sufficiently fine. The result supports not only the discretization approximation in Theorem 3, but also shows insensitivity of our FP algorithm to the fineness of the grid, as long as it is sufficiently fine to approximate the problem well. Hence, in the following we will use $M = 120$.

Finite-player convergence. In Figure 4, the convergence of episodic returns by propagation of chaos is depicted as the number of players $N \to \infty$. The limiting performance as the number of players grows, quickly approaches the performance of the projected MF prediction, up to a small, negligible error from discretization and finite players. The result supports propagation of chaos in Theorem 1 by convergence of the empirical objective to the limiting objective, despite the non-Lipschitz projected MF policies. In Cui et al. (2023, Appendix I), similar results are shown for a Lipschitz policy.

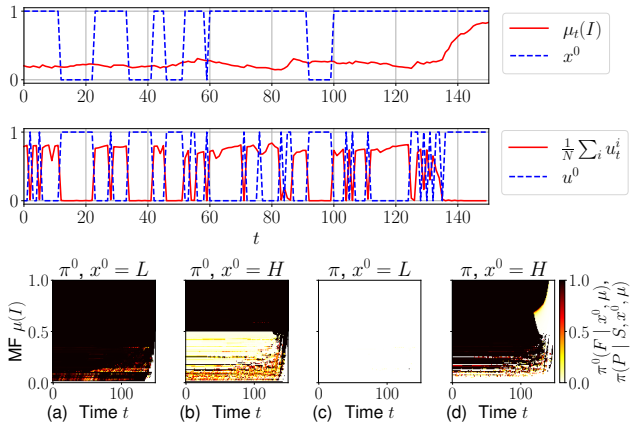

Figure 5: The FP-learned M3FNE in SIS, for $M = 120$. Top: example trajectory (for visualization, $L = \bar{P} = \bar{F} = 0, H = P = F = 1$, see Cui et al. (2023, Appendix I) for the details); bottom (a-d): policy heatmaps.

Qualitative analysis. Lastly, we visualize the qualitative behavior obtained and find plausible equilibrium behavior, e.g., for the SIS problem. As seen in Figure 5, the equilibrium behavior plausibly reaches an equilibrium of infected players, where the cost of actions equilibrates. The number of infected increases over time due to the finite horizon, discounting costs of infection beyond the horizon. Furthermore, minor players take precautions only down to some infection threshold, at which point the expected cost of not taking precautions is higher. The major player prevents infections in the low-infectivity regime ($x^0 = L$), while in the high-infectivity regime ($x^0 = H$) the high infection probability for minor players already encourages preventative actions.

Conclusion and Discussion

We have developed a new model and algorithm for a novel, broad class of tractable games. The framework allows scalable analysis of a large number of players with theoretical guarantees. The proposed methods have been empirically supported through a variety of experiments. Still, for problems with multiple Nash equilibria, the FP algorithm finds only some equilibrium. Future work could address finding all or specific, e.g., socially-optimal equilibria. One could also try to relax theoretical assumptions. Lastly, since scalability of the discretization method remains an issue for larger minor state spaces, one may consider deep RL methods.

Acknowledgments

This work has been co-funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center. The authors acknowledge the Lichtenberg high performance computing cluster of the TU Darmstadt for providing computational facilities for the calculations of this research. M. Lauriere is affiliated with the Shanghai Frontiers Science Center of Artificial Intelligence and Deep Learning, and with the NYU-ECNU Institute of Mathematical Sciences at NYU Shanghai, NYU Shanghai, 567 West Yangsi Road, Shanghai, 200126, People's Republic of China. O. Pietquin was affiliated with Google DeepMind during the preparation of this work.

References

Aurell, A.; Carmona, R.; Dayanıklı, G.; and Laurière, M. 2022. Optimal incentives to mitigate epidemics: a Stackelberg mean field game approach. *SIAM J. Control Optim.*, 60(2): S294–S322.

Bensoussan, A.; Chau, M. H.; and Yam, S. C. 2016. Mean field games with a dominating player. *Appl. Math. Optim.*, 74: 91–128.

Cabannes, T.; Laurière, M.; Pérolat, J.; Marinier, R.; Girgin, S.; Perrin, S.; Pietquin, O.; Bayen, A. M.; Goubault, E.; and Elie, R. 2022. Solving N-Player Dynamic Routing Games with Congestion: A Mean-Field Approach. In *Proc. AAMAS*, volume 21, 1557–1559.

Cardaliaguet, P.; Cirant, M.; and Porretta, A. 2020. Remarks on Nash equilibria in mean field game models with a major player. *Proc. Am. Math. Soc.*, 148(10): 4241–4255.

Carmona, R. 2020. Applications of Mean Field Games in Financial Engineering and Economic Theory. arXiv:2012.05237.

Carmona, R.; and Dayanıklı, G. 2021. Mean field game model for an advertising competition in a duopoly. *International Game Theory Review*, 23(04): 2150024.

Carmona, R.; Dayanıklı, G.; and Laurière, M. 2022. Mean field models to regulate carbon emissions in electricity production. *Dynamic Games and Applications*, 12(3): 897–928.

Carmona, R.; Delarue, F.; and Lacker, D. 2016. Mean field games with common noise. *Ann. Probab.*, 44(6): 3740–3803.

Carmona, R.; Delarue, F.; et al. 2018. *Probabilistic theory of mean field games with applications I-II*. Springer.

Carmona, R.; and Wang, P. 2017. An alternative approach to mean field game with major and minor players, and applications to herders impacts. *Appl. Math. Optim.*, 76: 5–27.

Carmona, R.; and Wang, P. 2021. Finite-state contract theory with a principal and a field of agents. *Management Science*, 67(8): 4725–4741.

Carmona, R. A.; and Zhu, X. 2016. A probabilistic approach to mean field games with major and minor players. *Ann. Appl. Probab.*, 26(3): 1535–1580.

Chaintron, L.-P.; and Diez, A. 2022. Propagation of chaos: A review of models, methods and applications. I. Models and methods. *Kinetic and Related Models*, 15(6): 895–1015.

Cui, K.; Dayanıklı, G.; Laurière, M.; Geist, M.; Pietquin, O.; and Koeppl, H. 2023. Learning Discrete-Time Major-Minor Mean Field Games. arXiv:2312.10787.

Cui, K.; Fabian, C.; and Koeppl, H. 2023. Multi-Agent Reinforcement Learning via Mean Field Control: Common Noise, Major Agents and Approximation Properties. arXiv:2303.10665.

Cui, K.; and Koeppl, H. 2021. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *Proc. AISTATS*, 1909–1917.

Daskalakis, C.; Goldberg, P. W.; and Papadimitriou, C. H. 2009. The complexity of computing a Nash equilibrium. *SIAM J. Comput.*, 39(1): 195–259.

Djehiche, B.; Tcheukam, A.; and Tembine, H. 2017. Mean-Field-Type Games in Engineering. *AIMS Electron. Electr. Eng.*, 1(1): 18–73.

Elie, R.; Mastrolia, T.; and Possamaï, D. 2019. A tale of a principal and many, many agents. *Mathematics of Operations Research*, 44(2): 440–467.

Geist, M.; Pérolat, J.; Laurière, M.; Elie, R.; Perrin, S.; Bachem, O.; Munos, R.; and Pietquin, O. 2022. Concave Utility Reinforcement Learning: The Mean-field Game Viewpoint. In *Proc. AAMAS*, 489–497.

Gu, H.; Guo, X.; Wei, X.; and Xu, R. 2021. Mean-field controls with Q-learning for cooperative MARL: convergence and complexity analysis. *SIAM J. Math. Data Sci.*, 3(4): 1168–1196.

Guo, X.; Hu, A.; Xu, R.; and Zhang, J. 2019. Learning mean-field games. In *Proc. NeurIPS*, 4966–4976.

Guo, X.; Hu, A.; and Zhang, J. 2022. Optimization frameworks and sensitivity analysis of Stackelberg mean-field games. arXiv:2210.04110.

Guo, X.; Hu, A.; and Zhang, J. 2023. MF-OMO: An Optimization Formulation of Mean-Field Games. arXiv:2206.09608.

Guo, X.; Li, L.; Nabi, S.; Salhab, R.; and Zhang, J. 2023. MESOB: Balancing Equilibria & Social Optimality. arXiv:2307.07911.

Guo, X.; Xu, R.; and Zariphopoulou, T. 2022. Entropy regularization for mean field games with learning. *Math. Oper. Res.*

Harris, C. 1998. On the rate of convergence of continuous-time fictitious play. *Games Econ. Behav.*, 22(2): 238–259.

Hernández-Lerma, O.; and Lasserre, J. B. 2012. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media.

Hofbauer, J.; and Sandholm, W. H. 2002. On the global convergence of stochastic fictitious play. *Econometrica*, 70(6): 2265–2294.

Huang, J.; and Wang, S. 2016. Dynamic optimization of large-population systems with partial information. *J. Optim. Theory Appl.*, 168: 231–245.

Huang, M. 2010. Large-population LQG games involving a major player: the Nash certainty equivalence principle. *SIAM J. Control Optim.*, 48(5): 3318–3353.

Huang, M.; Caines, P. E.; Malhamé, R. P.; et al. 2006. Distributed multi-agent decision-making with partial observations: asymptotic Nash equilibria. In *Proc. 17th Internat. Symp. MTNS*, 2725–2730.

Huang, M.; Malhamé, R. P.; and Caines, P. E. 2006. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Commun. Inf. Syst.*, 6(3): 221–252.

Lasry, J.-M.; and Lions, P.-L. 2007. Mean field games. *Japanese J. Math.*, 2(1): 229–260.

Lasry, J.-M.; and Lions, P.-L. 2018. Mean-field games with a major player. *Comptes Rendus Mathematique*, 356(8): 886–890.

Laurière, M.; Perrin, S.; Girgin, S.; Muller, P.; Jain, A.; Cabannes, T.; Piliouras, G.; Pérolat, J.; Elie, R.; Pietquin, O.; et al. 2022. Scalable deep reinforcement learning algorithms for mean field games. In *Proc. ICML*, 12078–12095. PMLR.

Laurière, M.; Perrin, S.; Geist, M.; and Pietquin, O. 2022. Learning Mean Field Games: A Survey. arXiv:2205.12944.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529–533.

Mondal, W. U.; Agarwal, M.; Aggarwal, V.; and Ukkusuri, S. V. 2022. On the Approximation of Cooperative Heterogeneous Multi-Agent Reinforcement Learning (MARL) using Mean Field Control (MFC). *JMLR*, 23(129): 1–46.

Motte, M.; and Pham, H. 2022. Mean-field Markov decision processes with common noise and open-loop controls. *Ann. Appl. Probab.*, 32(2): 1421–1458.

Nguyen, S. L.; and Huang, M. 2012. Linear-quadratic-Gaussian mixed games with continuum-parametrized minor players. *SIAM J. Control Optim.*, 50(5): 2907–2937.

Nourian, M.; and Caines, P. E. 2013. ϵ-Nash mean field game theory for nonlinear stochastic dynamical systems with major and minor agents. *SIAM J. Control Optim.*, 51(4): 3302–3331.

Pásztor, B.; Krause, A.; and Bogunovic, I. 2023. Efficient Model-Based Multi-Agent Mean-Field Reinforcement Learning. *Trans. Mach. Learn. Res.*

Pérolat, J.; Perrin, S.; Elie, R.; Laurière, M.; Piliouras, G.; Geist, M.; Tuyls, K.; and Pietquin, O. 2022. Scaling Mean Field Games by Online Mirror Descent. In *Proc. AAMAS*, 1028–1037.

Perrin, S.; Perolat, J.; Laurière, M.; Geist, M.; Elie, R.; and Pietquin, O. 2020. Fictitious play for mean field games: continuous time analysis and applications. In *Proc. NeurIPS*, 13199–13213.

Saldi, N.; Başar, T.; and Raginsky, M. 2018. Markov–Nash equilibria in mean-field games with discounted cost. *SIAM J. Control Optim.*, 56(6): 4256–4287.

Sen, N.; and Caines, P. E. 2016. Mean field game theory with a partially observed major agent. *SIAM J. Control Optim.*, 54(6): 3174–3224.

Sutton, R. S.; and Barto, A. G. 2018. *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT press.

Vasal, D.; and Berry, R. 2022. Master Equation for Discrete-Time Stackelberg Mean Field Games with a Single Leader. In *Proc. CDC*, 5529–5535. IEEE.

Yang, Y.; and Wang, J. 2021. An Overview of Multi-Agent Reinforcement Learning from Game Theoretical Perspective. arXiv:2011.00583.

Yardim, B.; Cayci, S.; Geist, M.; and He, N. 2023. Policy mirror ascent for efficient and independent learning in mean field games. In *Proc. ICML*, 39722–39754. PMLR.

Zhang, K.; Yang, Z.; and Başar, T. 2021. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. In Vamvoudakis, K. G.; Wan, Y.; Lewis, F. L.; and Cansever, D., eds., *Handbook of Reinforcement Learning and Control*, 321–384. Cham: Springer International Publishing.