When is Mean-Field Reinforcement Learning Tractable and Relevant?

Batuhan Yardim
ETH Zürich
Zürich, Switzerland
yardima@ethz.ch

Artur Goldman
HSE University
Moscow, Russia
agoldman@hse.ru

Niao He
ETH Zürich
Zürich, Switzerland
niao.he@inf.ethz.ch

ABSTRACT

Mean-field reinforcement learning has become a popular theoretical framework for efficiently approximating large-scale multi-agent reinforcement learning (MARL) problems exhibiting symmetry. However, questions remain regarding the applicability of mean-field approximations: in particular, their approximation accuracy of real-world systems and conditions under which they become computationally tractable. We establish explicit finite-agent bounds for how well the MFG solution approximates the true N-player game for two popular mean-field solution concepts. Furthermore, for the first time, we establish explicit lower bounds indicating that MFGs are poor or uninformative at approximating N-player games assuming only Lipschitz dynamics and rewards. Finally, we analyze the computational complexity of solving MFGs with only Lipschitz properties and prove that they are in the class of PPAD-complete problems conjectured to be intractable, similar to general sum N player games. Our theoretical results underscore the limitations of MFGs and complement and justify existing work by proving difficulty in the absence of common theoretical assumptions.

KEYWORDS

Mean-Field Games; Computational Complexity; Approximation

ACM Reference Format:
Batuhan Yardim, Artur Goldman, and Niao He. 2024. When is Mean-Field Reinforcement Learning Tractable and Relevant?. In *Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), Auckland, New Zealand, May 6 – 10, 2024*, IFAAMAS, 9 pages.

1 INTRODUCTION

Multi-agent reinforcement learning (MARL) finds numerous impactful applications in the real world [20, 21, 26, 29, 30, 32]. Despite the urgent need in practice, MARL remains a fundamental challenge, especially in the setting with large numbers of agents due to the so-called "curse of many agents" [31].

Mean-field games (MFG), a theoretical framework first proposed by Lasry and Lions [18] and Huang et al. [16], permits the theoretical study of such large-scale games by introducing mean-field simplification. Under certain assumptions, the mean-field approximation leads to efficient algorithms for the analysis of a particular type of N-agent competitive game where there are symmetries between players and when N is large. Such games appear widely in for instance auctions [17], and cloud resource management [20]. For the mean-field analysis, the game dynamics with N-players must be *symmetric* (i.e., each player must be exposed to the same rules) and *anonymous* (i.e., the effect of each player on the others should be permutation invariant). Under this simplification, works such as [1, 6, 12, 23, 25, 33, 34] and many others have analyzed reinforcement learning (RL) algorithms in the MFG limit $N \to \infty$ to obtain a tractable approximation of many agent games, providing learning guarantees under various structural assumptions.

Being a simplification, MFG formulations should ideally satisfy two desiderata: (1) they should be *relevant*, i.e., they are good approximations of the original MARL problem and (2) they should be *tractable*, i.e., they are at least easier than solving the original MARL problem. In this work, we would like to understand the extent to which MFGs satisfy these two requirements, and we aim to answer two natural questions that remain understudied:

- *When are MFGs good approximations of the finite player games, when are they not?* In particular, are polynomially many agents always sufficient for mean-field approximation to be effective?
- *Is solving MFGs always computationally tractable, or more tractable than directly solving the N-player game?* In particular, can MFGs be solved in polynomial or pseudo-polynomial time?

1.1 Related Work

Mean-field RL has been studied in various mathematical settings. In this work, we focus on two popular formulations in particular: stationary mean-field games (Stat-MFG, see e.g. [1, 12]) and finite-horizon MFG (FH-MFG, see e.g. [23, 25]). In the Stat-MFG setting the objective is to find a stationary policy that is optimal with respect to its induced stationary distribution, while in the FH-MFG setting, a finite-horizon reward is considered with a time-varying policy and population distribution.

Existing results on MFG relevance/approximation. The approximation properties of MFGs have been explored by several works in literature, as summarized in Table 1. Finite-agent approximation bounds have been widely analyzed in the case of stochastic mean-field differential games [3, 4], albeit in the differential setting and without explicit lower bounds. Recent works [1, 6] have established that Stat-MFG Nash equilibria (Stat-MFG-NE) asymptotically approximate the NE of N-player symmetric dynamic games under continuity assumptions. The result by Saldi et al. [28], as the basis of subsequent proofs, shows asymptotic convergence for a large class of MFG variants and only requires continuity of dynamics and rewards as well as minor technical assumptions such as compactness and a form of local Lipschitz continuity. However, such

Proc. of the 23rd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2024), N. Alechina, V. Dignum, M. Dastani, J.S. Sichman (eds.), May 6 – 10, 2024, Auckland, New Zealand. © 2024 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

asymptotic convergence guarantees leave the question unanswered if the MFG models are realistic in real-world games. Many games such as traffic systems, financial markets, etc. naturally exhibit large N, however, if N must be astronomically large for good approximation, the real-world impact of the mean-field analysis will be limited. Recently, [35] provided finite-agent approximation bounds of a special class of stateless MFG, which assumes no state dynamics. We complement existing work on approximation properties of both Stat-MFG and FH-MFG by providing explicit upper and lower bounds for approximation.

Existing results on MFG tractability. The tractability of solving MFGs as a proxy for MARL has been also heavily studied in the RL community under various classes of structural assumptions. Since finding approximate Nash equilibria for normal form games is PPAD-complete, a class believed to be computationally intractable [5, 7], solving the mean-field approximation in many cases can be a tractable alternative. We summarize recent work for computationally (or statistically) solving the two types of MFGs below, with an in-depth comparison also provided in Table 2.

For Stat-MFG, under a contraction assumption RL algorithms such as Q-learning [1, 37], policy mirror ascent [34], policy gradient methods [13], soft Q-learning [6] and fictitious play [33] have been shown to solve Stat-MFG with statistical and computational efficiency. However, all of these guarantees require the game to be heavily regularized as pointed out in [6, 34], inducing a non-vanishing bias on the computed Nash. Moreover, in some works the population evolution is also implicitly required to be contractive under all policies (see e.g. [12, 34]), further restricting the analysis to sufficiently smooth games. While [14] has proposed a method that guarantees convergence to MFG-NE under differentiable dynamics, the algorithm converges only when initialized sufficiently close to the solution. To the best of our knowledge, there are neither RL algorithms that work without regularization nor evidence of difficulty in the absence of such strong assumptions: we complement the line of work by showing that unless dynamics are sufficiently smooth, Stat-MFG is both computationally intractable and a poor approximation.

A separate line of work analyzes the finite horizon problem. In this case, when the dynamics are population-independent and the payoffs are monotone the problem is known to be tractable. Algorithms such as fictitious play [25] and mirror descent [23] have been shown to converge to Nash in corresponding continuous-time equations. Recent work has also focused on the statistical complexity of the finite-horizon problem in very general FH-MFG problems [15], however, the algorithm proposed is in general computationally intractable. In terms of computational tractability and the approximation properties, our work complements these results by demonstrating that (1) when dynamics depend on the population as well an exponential approximation lower bound exists, and (2) in the absence of monotonicity, the FH-MFG is provably as difficult as solving an N-player game.

Finally, we note that there are several other settings and MFG solution concepts have been analyzed. For instance, a certain class of infinite horizon MFG has been shown to be equivalent to concave utility RL, proving finite-time computational guarantees [10].

1.2 Our Contribution

In this work, we formalize and provide answers to the two aforementioned fundamental questions, first focusing on the approximation properties of MFG in Section 3 and later on the computational tractability of MFG in Section 4. Our contributions are summarized as follows.

Firstly, we introduce explicit finite-agent approximation bounds for finite horizon and stationary MFGs (Table 1) in terms of exploitability in the finite agent game. In both cases, we prove explicit upper bounds which quantify how many agents a symmetric game must have to be well-approximated by the MFG, which has been absent in the literature to the best of our knowledge. Our approximation results only require a minimal Lipschitz continuity assumption of the transition kernel and rewards. For FH-MFG, we prove a $\mathcal{O}\left(\frac{(1-L^H)H^2}{(1-L)\sqrt{N}}\right)$ upper bound for the exploitabilty where L is the Lipschitz modulus of the population evolution operator: the upper bound exhibits an exponential dependence on the horizon H. For the Stat-MFG we show that a $\mathcal{O}\left(\frac{(1-\gamma)^{-3}}{\sqrt{N}}\right)$ approximation bound can be established, but only if the population evolution dynamics are non-expansive. Next, for the first time, we establish explicit lower bounds for the approximation proving the shortcomings of the upper bounds are fundamental. For the FH-MFG, we show that unless $N \geq \Omega(2^H)$, an exploitability linear in horizon H is unavoidable when deploying the MFG solution to the N player game: hence in general the MFG equilibrium becomes irrelevant quickly as the problem horizon increases. For Stat-MFG we establish an $\Omega(N^{\log_2 \gamma})$ lower bound when the population dynamics are not restricted to non-expansive population operators, showing that a large discount factor γ also rapidly deteriorates the approximation efficiency. Our lower bounds indicate that in the worst case, the number of agents required for the approximation can grow exponentially in the problem parameters, demonstrating the limitations of the MFG approximation.

Finally, from the computational perspective, we establish that both finite-horizon and stationary MFGs can be PPAD-complete problems in general, even when restricted to certain simple subclasses (Table 2). This shows that both MFG problems are in general as hard as finding a Nash equilibrium of N-player general sum games. Furthermore, our results imply that unless PPAD=P there are no polynomial time algorithms for solving FH-MFG and Stat-MFG, a result indicating computational intractability.

2 MEAN-FIELD GAMES: DEFINITIONS, SOLUTION CONCEPTS

Notation. Throughout this work, we assume $\mathcal{S}, \mathcal{A}$ are finite sets. For a finite set $\mathcal{X}$, $\Delta_{\mathcal{X}}$ denotes the set of probability distributions on $\mathcal{X}$. The norm used will not fundamentally matter for our results, we choose to equip $\Delta_{\mathcal{S}}, \Delta_{\mathcal{A}}$ with the norm $\|\cdot\|_1$. We define the set of Markov policies $\Pi := \{\pi : \mathcal{S} \to \Delta_{\mathcal{A}}\}$, $\Pi_H := \{\{\pi_h\}_{h=0}^{H-1} : \pi_h \in \Pi, \forall h\}$ and $\Pi_H^N := \{\{\pi_h^i\}_{h=0,i=0}^{H-1,N} : \pi_h^i \in \Pi, \forall h\}$. For policies $\pi, \pi' \in \Pi$ denote $\|\pi - \pi'\|_1 = \sup_{s\in\mathcal{S}} \|\pi(\cdot|s) - \pi'(\cdot|s)\|_1$. We denote $d(x, y) := \mathbb{1}_{\{x\neq y\}}$ for x, y in $\mathcal{A}$ or $\mathcal{S}$. For $\boldsymbol{\pi} \in \Pi^N$, $\pi' \in \Pi$, we define $(\pi', \boldsymbol{\pi}^{-i}) \in \Pi^N$ as the policy profile where the i-th policy has been replaced by π'. Likewise, for $\boldsymbol{\pi} \in \Pi_H^N, \boldsymbol{\pi'} \in \Pi_H$, we denote by

Work	MFG type	Key Assumptions	Approximation Rate (in Exploitability)
Carmona and Delarue, 2013	Other[a]	Affine drift, Lipschitz derivatives	$O(N^{-1/(d+4)})$ (d dimension of state space)
Saldi et al., 2018	Other[b]	Continuity	$o(1)$ (asymptotic: convergence as $N \to \infty$)
Anahtarci et al., 2022	Stat-MFG	Lipschitz P, R + Regularized + Contractive Γ_P	$o(1)$ (asymptotic: convergence as $N \to \infty$)
Cui and Koeppl, 2021	Stat-MFG	Continuity	$o(1)$ (asymptotic: convergence as $N \to \infty$)
Yardim et al., 2023a	Other[c]	Lipschitz P, R	$O(1/\sqrt{N})$
Theorem 3.2	FH-MFG	Lipschitz P, R	$O\left(\frac{H^2(1-L^H)}{(1-L)\sqrt{N}}\right)$, L Lipschitz modulus of Γ_P
Theorem 3.3	FH-MFG	Lipschitz P, R	$\Omega(H)$ unless $N \geq \Omega(2^H)$
Theorem 3.5	Stat-MFG	Lipschitz P, R + Non-expansive Γ_P	$O((1-\gamma)^{-3}/\sqrt{N})$
Theorem 3.6	Stat-MFG	Lipschitz P, R	$\Omega(N^{-\log_2 \gamma^{-1}}))$

Table 1: Selected approximation results for MFG. Notes: [a] stochastic differential MFG, [b] infinite-horizon discounted setting with non-stationary policies, [c] stateless/static MFG setting.

Work	MFG Type	Key Assumptions	Iteration/Sample Complexity result
Anahtarci et al., 2022	Stat-MFG	Lipschitz P, R + Regularization + Contractive Γ_P	$\widetilde{O}(\varepsilon^{-4\lvert\mathcal{A}\rvert})$ samples, $O(\log \varepsilon^{-1})$ iterations
Geist et al., 2022	Other[a]	Concave potential	$O(\varepsilon^{-2})$ iterations
Perrin et al., 2020	FH-MFG	Monotone R, μ-independent P	$O(\varepsilon^{-1})$ (continuous time analysis)
Pérolat et al., 2022	FH-MFG	Monotone R, μ-independent P	$O(\varepsilon^{-1})$ (continuous time analysis)
Zaman et al., 2023	Stat-MFG	Lipschitz P, R + Regularization + Contractive Γ_P	$O(\varepsilon^{-4})$ samples
Cui and Koeppl, 2021	Stat-MFG	Lipschitz P, R + Regularization	$O(\log \varepsilon^{-1})$ iterations
Yardim et al., 2023a	Other[b]	Monotone and Lipschitz R	$O(\varepsilon^{-2})$ samples (N-player)
Yardim et al., 2023b	Stat-MFG	Lipschitz P, R + Regularization + Contractive Γ_P	$O(\varepsilon^{-2})$ samples (N-player)
Theorem 4.9	Stat-MFG	Lipschitz P, R	PPAD-complete
Theorem 4.12	FH-MFG	Lipschitz P, R + μ-independent P	PPAD-complete
Theorem 4.14	FH-MFG	Linear P, R + μ-independent P	PPAD-complete

Table 2: Selected results for computing MFG-NE from literature. In the assumptions column, contractive Γ_P indicates that for all $\pi \in \Pi$, $\Gamma_P(\cdot, \pi)$ is a contraction, and regularization indicates that a non-vanishing bias is present. Notes: [a] infinite-horizon, population dependence through the discounted state distribution. [b] stateless/static MFG.

$(\boldsymbol{\pi}', \boldsymbol{\pi}^{-i}) \in \Pi_H^N$ the policy profile where the i-th player's policy has been replaced by $\boldsymbol{\pi}'$. For any $N \in \mathbb{N}_{\geq 0}$, $[N] := \{1, \dots, N\}$.

MFGs introduce a dependence on the population distribution over states of the rewards and dynamics. We will strictly consider Lipschitz continuous rewards and dynamics, which is a common assumption in literature [1, 12, 33, 34], formalized below.

Definition 2.1 (Lipschitz dynamics, rewards). For some $L \geq 0$, we define the set of L-Lipschitz reward functions and state transition dynamics as

$$\mathcal{R}_L := \Big\{R : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}} \to [0,1] \ : \ |R(s,a,\mu) - R(s,a,\mu')| \leq L\|\mu - \mu'\|_1, \forall s, a, \mu, \mu'\Big\},$$

$$\mathcal{P}_L := \Big\{P : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}} \to \Delta_{\mathcal{S}} \ : \ \|P(s,a,\mu) - P(s,a,\mu')\|_1 \leq L\|\mu - \mu'\|_1, \forall s, a, \mu, \mu'\Big\}.$$

Moreover, we define the set of Lipschitz rewards and dynamics as $\mathcal{R} := \bigcup_{L \geq 0} \mathcal{R}_L$, $\mathcal{P} := \bigcup_{L \geq 0} \mathcal{P}_L$ respectively.

We note that there are interesting MFGs with non-Lipschitz dynamics and rewards, however, even the existence of Nash is not guaranteed in this case. Lipschitz continuity is a minimal assumption under which solutions to MFG always exist, and as our aim is to prove lower bounds and difficulty we will adopt this assumption. Solving MFG with non-Lipschitz dynamics is more challenging than Lipschitz continuous MFG (the latter being a subset of the former), hence our difficulty results will apply.

Operators. We will define the useful population operators $\Gamma_P : \Delta_{\mathcal{S}} \times \Pi \to \Delta_{\mathcal{S}}$, $\Gamma_P^H : \Delta_{\mathcal{S}} \times \Pi \to \Delta_{\mathcal{S}}$, and $\Lambda_P^H : \Delta_{\mathcal{S}} \times \Pi_H \to \Delta_{\mathcal{S}}^H$ as

$$\Gamma_P(\mu, \pi) := \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mu(s)\pi(a|s)P(\cdot|s,a,\mu),$$

$$\Gamma_P^H(\mu, \pi) := \underbrace{\Gamma_P(\dots \Gamma_P(\Gamma_P(\mu, \pi), \pi) \dots), \pi)}_{H \text{ times}},$$

$$\Lambda_P^H(\mu_0, \boldsymbol{\pi}) := \{\underbrace{\Gamma_P(\dots \Gamma_P(\Gamma_P(\mu_0, \pi_0), \pi_1) \dots, \pi_{h-1})}_{h \text{ times}}\}_{h=0}^{H-1}$$

for all $n \in \mathbb{N}_{>0}, \pi \in \Pi, \boldsymbol{\pi} = \{\pi_h\}_{h=0}^{H-1} \in \Pi_H, P \in \mathcal{P}, \mu_0 \in \Delta_{\mathcal{S}}$.

Finally, we will need the following Lipschitz continuity result for the Γ_P operator.

Lemma 2.2. *[34, Lemma 3.2] Let $P \in \mathcal{P}_{K_\mu}$ for $K_\mu > 0$ and*

$$K_s := \sup_{\substack{s,s'\\a,\mu}} \left\|P(s,a,\mu) - P(s',a,\mu)\right\|_1, K_a := \sup_{\substack{a,a'\\s,\mu}} \left\|P(s,a,\mu) - P(s,a',\mu)\right\|_1.$$

Then it holds for all $\mu, \mu' \in \Delta_{\mathcal{S}}, \pi, \pi' \in \Pi$ that:

$$\|\Gamma_P(\mu,\pi) - \Gamma_P(\mu',\pi')\|_1 \le L_{pop,\mu}\|\mu - \mu'\|_1 + \frac{K_a}{2}\|\pi - \pi'\|_1,$$

where $L_{pop,\mu} := (K_\mu + \frac{K_s}{2} + \frac{K_a}{2})$ for all $\pi, \pi' \in \Pi$, $\mu, \mu' \in \Delta_{\mathcal{S}}$.

In particular, in our settings, Lemma 2.2 indicates that Γ_P is always Lipschitz continuous if $P \in \mathcal{P}$, a property which will become significant for approximation analysis.

We will be interested in two classes of MFG solution concepts that lead to different analyses: infinite horizon stationary MFG Nash equilibrium (Stat-MFG-NE) and finite horizon MFG Nash equilibrium (FH-MFG-NE). The first problem widely studied in literature is the stationary MFG equilibrium problem, see for instance [1, 12, 13, 33, 34]. We formalize this solution concept below.

Definition 2.3 (Stat-MFG). A stationary MFG (Stat-MFG) is defined by the tuple $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ for Lipschitz dynamics and rewards $P \in \mathcal{P}, R \in \mathcal{R}$, discount factor $\gamma \in (0,1)$. For any $(\mu, \pi) \in \Delta_{\mathcal{S}} \times \Pi$, we define the γ-discounted infinite horizon expected reward as

$$V^{\gamma}_{P,R}(\mu,\pi) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t,a_t,\mu) \middle| \substack{s_0\sim\mu,\quad a_t\sim\pi(s_t)\\ s_{t+1}\sim P(s_t,a_t,\mu)}\right].$$

A policy-population pair $(\mu^*, \pi^*) \in \Delta_{\mathcal{S}} \times \Pi$ is called a Stat-MFG Nash equilibrium if the two conditions hold:

$$\begin{aligned} &\textit{Stability:} && \mu^* = \Gamma_P(\mu^*, \pi^*),\\ &\textit{Optimality:} && V^{\gamma}_{P,R}(\mu^*,\pi^*) = \max_{\pi\in\Pi} V^{\gamma}_{P,R}(\mu^*,\pi). \end{aligned} \quad \text{(Stat-MFG-NE)}$$

The second MFG concept that we will consider has a finite time horizon, and is also common in literature [15, 19, 24, 25]. In this case, the population distribution is permitted to vary over time, and the objective is to find an optimal non-stationary policy with respect to the population distribution it induces. We formalize this problem and the corresponding solution concept below.

Definition 2.4 (FH-MFG). A finite horizon MFG problem (FH-MFG) is determined by the tuple $(\mathcal{S}, \mathcal{A}, H, P, R, \mu_0)$ where $H \in \mathbb{Z}_{>0}$, $P \in \mathcal{P}, R \in \mathcal{R}, \mu_0 \in \Delta_{\mathcal{S}}$. For $\boldsymbol{\pi} = \{\pi_h\}_{h=0}^{H} \in \Pi_H$, $\boldsymbol{\mu} = \{\mu_h\}_{h=0}^{H-1} \in \Delta_{\mathcal{S}}^{H}$, define the expected reward and exploitability as

$$V^{H}_{P,R}(\boldsymbol{\mu},\boldsymbol{\pi}) := \mathbb{E}\left[\sum_{h=0}^{H-1} R(s_h,a_h,\mu_h) \middle| \substack{s_0\sim\mu_0,\quad a_h\sim\pi_h(s_h)\\ s_{h+1}\sim P(s_h,a_h,\mu_h)}\right],$$

$$\mathcal{E}^{H}_{P,R}(\boldsymbol{\pi}) := \max_{\boldsymbol{\pi}'\in\Pi^H} V^{H}_{P,R}(\Lambda^H_P(\mu_0,\boldsymbol{\pi}),\boldsymbol{\pi}') - V^{H}_{P,R}(\Lambda^H_P(\mu_0,\boldsymbol{\pi}),\boldsymbol{\pi}).$$

Then, the FH-MFG Nash equilibrium is defined as:

$$\begin{aligned} &\textit{Policy } \boldsymbol{\pi}^* = \{\pi^*_h\}_{h=0}^{H-1} \in \Pi_H \text{ such that}\\ &\mathcal{E}^{H}_{P,R}(\{\pi^*_h\}_{h=0}^{H-1}) = 0. \end{aligned} \quad \text{(FH-MFG-NE)}$$

3 APPROXIMATION PROPERTIES OF MFG

As established in literature, the reason the FH-MFG and Stat-MFG problems are studied is the fact that they can approximate the NE of certain symmetric games with N players, establishing the main relevance of the formulations in the real world. Such results are summarized in Table 1.

In this section, we study how efficient this convergence is and also related lower bounds. For these purposes, we first define the corresponding *finite-player* game of each mean-field game problem: to avoid confusion, we call these games *symmetric anonymous dynamic games* (SAG). Afterwards, for each solution concept, we will first establish (1) an upper bound on the approximation error (i.e. the exploitability) due to the mean-field, and (2) a lower bound demonstrating the worst-case rate. We will present the main outlines of proofs, and postpone computation-intensive derivations to the supplementary material of the paper [36].

3.1 Approximation Analysis of FH-MFG

Firstly, we define the finite-player game that is approximately solved by the FH-MFG-NE.

Definition 3.1 (N-FH-SAG). An N-player finite horizon SAG (N-FH-SAG) is determined by the tuple $(N, \mathcal{S}, \mathcal{A}, H, P, R, \mu_0)$ such that $N \in \mathbb{Z}_{>0}, H \in \mathbb{Z}_{>0}$, $P \in \mathcal{P}, R \in \mathcal{R}, \mu_0 \in \Delta_{\mathcal{S}}$. For any $\boldsymbol{\pi} = \{\pi^i_h\}_{h=0,\dots,H-1, i\in[N]} \in \Pi_H^N$, we define the expected mean reward and exploitability of player i as

$$J^{H,N,(i)}_{P,R}(\boldsymbol{\pi}) := \mathbb{E}\left[\sum_{h=0}^{H-1} R(s^i_h,a^i_h,\widehat{\mu}_h) \middle| \substack{\forall j: s^j_0\sim\mu_0,\quad a^j_h\sim\pi^j_h(s^j_h)\\ s^j_{h+1}\sim P(s^j_h,a^j_h,\widehat{\mu}_h), \widehat{\mu}_h := \frac{1}{N}\sum_j \mathbf{e}_{s^j_h}}\right],$$

$$\mathcal{E}^{H,N,(i)}_{P,R}(\boldsymbol{\pi}) := \max_{\boldsymbol{\pi}'\in\Pi^H} J^{H,N,(i)}_{P,R}(\boldsymbol{\pi}',\boldsymbol{\pi}^{-i}) - J^{H,N,(i)}_{P,R}(\boldsymbol{\pi}).$$

Then, the N-FH-SAG Nash equilibrium is defined as:

$$\begin{aligned} &N\textit{-tuple of policies } \{\pi^{(i),*}_h\}_{h=0}^{H-1} \in \Pi_H^N \text{ such that}\\ &\forall i : \mathcal{E}^{H,N,(i)}_{P,R}(\{\pi^*_h\}_{h=0}^{H-1}) = 0. \end{aligned} \quad (N\text{-FH-SAG-NE})$$

If instead $\mathcal{E}^{H,N,(i)}_{P,R}(\boldsymbol{\pi}) \le \delta$ for all i, then $\boldsymbol{\pi}$ is called a δ-N-FH-SAG Nash equilibrium.

The above definition corresponds to a real-world problem as the function $J^{H,N,(i)}_{P,R}$ expresses the expected total payoff of each player: hence a δ-N-MFG-NE is a Nash equilibrium of a concrete N-player game in the traditional game theoretical sense. Also, note that now in the definition transition probabilities and rewards depend on $\widehat{\mu}_h$ which is the $\mathcal{F}(\{s^i_h\}_i) = \mathcal{F}_h$-measurable random vector of the empirical state distribution at time h of all agents.

Firstly, we provide a positive result well-known in literature: the N-FH-SAG is approximately solved by the FH-MFG-NE policy. Unlike some past works, we establish an explicit rate of convergence in terms of N and problem parameters.

Theorem 3.2 (Approximation of N-FH-SAG). *Let $(\mathcal{S}, \mathcal{A}, H, P, R, \mu_0)$ be a FH-MFG with $P \in \mathcal{P}, R \in \mathcal{R}$ and with a FH-MFG-NE $\boldsymbol{\pi}^* \in \Pi_H$, and for any $N \in \mathbb{N}_{>0}$ let $\boldsymbol{\pi}^*_N := (\underbrace{\boldsymbol{\pi}^*, \dots, \boldsymbol{\pi}^*}_{N \text{ times}}) \in \Pi_H^N$. Let $L > 0$ be the Lipschitz constant of Γ_P in μ, and let $\mathcal{G}_N := (N, \mathcal{S}, \mathcal{A}, H, P, R, \mu_0)$ be the corresponding N-player game. Then:*

(1) *If* $L = 1$, *then for all* $i \in [N]$, $\mathcal{E}^{H,N,(i)}_{P,R}(\boldsymbol{\pi}^*_N) \le O(\frac{H^3}{\sqrt{N}})$, *that is,* $\boldsymbol{\pi}^*_N$ *is a* $O(\frac{H^3}{\sqrt{N}})$*-NE of* $\mathcal{G}_N$.

(2) *If* $L \neq 1$, *then for all* $i \in [N]$, $\mathcal{E}^{H,N,(i)}_{P,R}(\boldsymbol{\pi}^*_N) \le O\left(\frac{H^2(1-L^H)}{(1-L)\sqrt{N}}\right)$, *that is,* $\boldsymbol{\pi}^*_N$ *is a* $O\left(\frac{H^2(1-L^H)}{(1-L)\sqrt{N}}\right)$*-NE of* $\mathcal{G}_N$.

Proof. *(sketch)* Certain aspects of our proof will mirror the techniques introduced by [28], although we establish an explicit bound. We first bound the expected empirical population deviation given by $\mathbb{E}[\|\widehat{\mu}_h - \mu^{\boldsymbol{\pi}}_h\|_1] = O\left(\frac{L^h}{\sqrt{N}}\right)$ with an inductive concentration argument: at each step $h+1$, given past states $\widehat{\mu}_h$, the empirical distribution $\widehat{\mu}_h$ is a sum of N independent identically distributed sub-Gaussian random variables. Next, by utilizing the Lipschitz property of rewards and bounding deviation from the theoretical rewards the result follows in two computational steps: (1) we show that $\left|J^{H,N,(1)}_{P,R}(\boldsymbol{\pi},\dots,\boldsymbol{\pi}) - V^H_{P,R}(\Lambda^H_P(\mu_0,\boldsymbol{\pi}),\boldsymbol{\pi})\right| \le O(1/\sqrt{N})$, and similarly (2) we show that for any policy sequence $\boldsymbol{\pi}' \in \Pi_h$, we have $\left|J^{H,N,(1)}_{P,R}(\boldsymbol{\pi}',\boldsymbol{\pi},\dots,\boldsymbol{\pi}) - V^H_{P,R}(\Lambda^H_P(\mu_0,\boldsymbol{\pi}),\boldsymbol{\pi}')\right| \le O(1/\sqrt{N})$. The result follows by definition of exploitability, with explicit constants shown in the appendix [36]. □

Γ_P in Theorem 3.2 is always L-Lipschitz in μ for some L by Lemma 2.2. When $L > 1$, the upper bound $O\left((1+L^H)H^2/\sqrt{N}\right)$ has an exponential dependence on the Lipschitz constant of the operator Γ_P. However, for games with longer horizons, the upper bound might require an unrealistic amount of agents N to guarantee a good approximation due to the exponential dependency. Next, we establish a worst-case result demonstrating that this is not avoidable without additional assumptions.

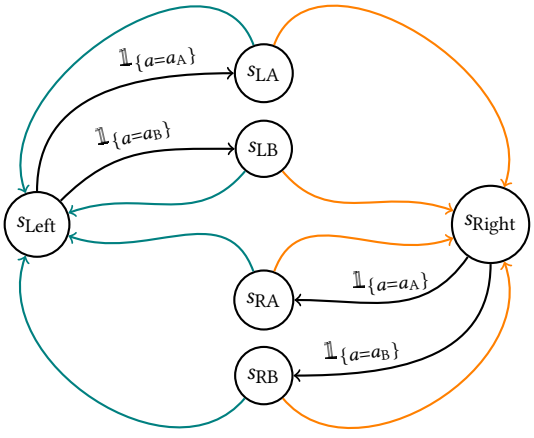

Figure 1: Visualization of the counterexample. All orange edges have probability $\omega_\varepsilon(\mu(s_{\mathbf{RA}}) + \mu(s_{\mathbf{RB}}))$**, green edges have probability** $\omega_\varepsilon(\mu(s_{\mathbf{LA}}) + \mu(s_{\mathbf{LB}}))$ **independent of action taken. Edges with probability** 0 **are not drawn.**

Theorem 3.3 (Approximation lower bound for N-FH-SAG). *There exists* $\mathcal{S}, \mathcal{A}$ *and* $P \in \mathcal{P}_8, R \in \mathcal{R}_2, \mu_0 \in \Delta_{\mathcal{S}}$ *such that the following hold:*

(1) *For each* $H > 0$, *the FH-MFG defined by* $(\mathcal{S}, \mathcal{A}, H, P, R, \mu_0)$ *has a* unique *solution* $\boldsymbol{\pi}^*_H$ *(up to modifications on zero-probability sets),*

(2) *For any* $H, h > 0$, *in the* N*-FH-SAG it holds that* $\mathbb{E}_H[\|\widehat{\mu}_h - \Lambda^H_P(\mu_0, \boldsymbol{\pi}^*_H)_h\|_1] \ge \Omega\left(\min\{1, \frac{2^H}{\sqrt{N}}\}\right)$.

(3) *For any* $H, N > 0$ *either* $N \ge \Omega(2^H)$, *or for each player* $i \in [N]$ *it holds that* $\mathcal{E}^{H,N,(i)}_{P,R}(\boldsymbol{\pi}^*_H,\dots,\boldsymbol{\pi}^*_H) \ge \Omega(H)$.

Proof. *(sketch)* We provide the basic idea of the proof and leave the cumbersome computations to the appendix. The proof is constructive: we construct an explicit FH-MFG where the statements hold, depicted in Figure 1. The FH-MFG will have 6 states and two actions defined as sets $\mathcal{S} = \{s_{\text{Left}}, s_{\text{Right}}, s_{\text{LA}}, s_{\text{LB}}, s_{\text{RA}}, s_{\text{RB}}\}$ and $\mathcal{A} = \{a_A, a_B\}$. We define the initial state distribution with $\boldsymbol{\mu}_0(s_{\text{Left}}) = \boldsymbol{\mu}_0(s_{\text{Right}}) = 1/2$. The colored state transition probabilities are given by the function:

$$\omega_\epsilon(x) = \begin{cases} 1, & x > 1/2 + \epsilon \\ 0, & x < 1/2 - \epsilon \\ \frac{1}{2} + \frac{x - 1/2}{2\epsilon}, & x \in [1/2 - \epsilon, 1/2 + \epsilon] \end{cases}.$$

The uniform policy over all actions $\boldsymbol{\pi}^*$ at all states will be the unique FH-MFG-NE for all H, and the mean-field population distribution for all even h will be $\mu^*_h(s_{\text{Left}}) = \mu^*_h(s_{\text{Right}}) = 1/2$. However, for finite N, using an anti-concentration bound on the binomial, we can show that with probability at least $1/10$, $\|\mu^*_0 - \widehat{\mu}_h\|_1 \ge 1/\sqrt{N}$. Using the fact that ω_ϵ is $(2\epsilon)^{-1}$-expansive in the interval $[1/2 - \epsilon, 1/2 + \epsilon]$, we can then show that the empirical population distribution exponentially diverges from the mean-field, that is $\mathbb{E}[\|\mu^*_{2h} - \widehat{\mu}_{2h}\|_1] \ge \Omega(5^h/\sqrt{N})$ until time $K := \log_5 \sqrt{N}$. Moreover, with a series of concentration bounds, it can be shown that within an expected number of $O(\log N)$ steps, all agents will converge to either s_{Left} or s_{Right} during even rounds. Only the colored transitions are defined to have non-zero rewards, whose definition (provided in the supplementary) guarantees that the exploitability suffered scales linearly with H after N agents concentrate on the same state in even steps. □

This result shows that without further assumptions, the FH-MFG solution might suffer from exponential exploitability in H in the N-player game. In such cases, to avoid the concrete N-player game from deviating from the mean-field behavior too fast, either H must be small or P must be sufficiently smooth in μ. We note that the typical assumption in the finite-horizon setting that $P \in \mathcal{P}_0$ (see e.g. [10, 25]) avoids this lower bound since in this case $\Gamma_P(\cdot, \pi)$ is simply multiplication by a stochastic matrix which is always non-expansive ($L = 1$). We also note at the expense of simplicity a stronger counter-example inducing exploitability $\Omega(H)$ unless $N \ge \Omega((L-\epsilon)^H)$ for all $\epsilon > 0$ can be constructed, where $P \in \mathcal{P}_L$.

A remark. The proof of Theorem 3.3 in fact suggests that for finite N and large horizon H, there exists a time-homogenous policy $\overline{\pi}^\star \in \Pi$ different than the FH-MFG solution such that for $\overline{\boldsymbol{\pi}}^\star_H := \{\overline{\pi}^\star\}_{h=0}^{H-1} \in \Pi_H$, the time-averaged exploitability of $\overline{\boldsymbol{\pi}}^\star_H$ is small: $\forall i \in [N] : H^{-1}\mathcal{E}^{H,N,(i)}_{P,R}(\overline{\boldsymbol{\pi}}^\star_H,\dots,\overline{\boldsymbol{\pi}}^\star_H) \le O(H^{-1}\log_2 N)$.

3.2 Approximation Analysis of Stat-MFG

Similarly, we introduce the N-player game corresponding to the Stat-MFG solution concept.

Definition 3.4 (N-Stat-SAG). An N-player stationary SAG (N-Stat-SAG) problem is defined by the tuple $(N, \mathcal{S}, \mathcal{A}, P, R, \gamma)$ for Lipschitz dynamics and rewards $P \in \mathcal{P}, R \in \mathcal{R}$, discount factor $\gamma \in (0,1)$. For any $(\mu, \boldsymbol{\pi}) \in \Delta_{\mathcal{S}} \times \Pi^N$, the N-player γ-discounted infinite horizon expected reward is defined as:

$$J_{P,R}^{\gamma,N,(i)}(\mu, \boldsymbol{\pi}) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t^i, a_t^i, \widehat{\mu}_t) \middle| \begin{matrix} a_t^j \sim \pi^j(s_t^j), \widehat{\mu}_t := \frac{\sum_j \mathbf{e}_{s_t^j}}{N} \\ s_0^j \sim \mu, s_{t+1}^i \sim P(s_t^i, a_t^i, \widehat{\mu}_t) \end{matrix}\right].$$

A policy profile-population pair $(\mu^*, \boldsymbol{\pi}^*) \in \Delta_{\mathcal{S}} \times \Pi^N$ is called an N-Stat-SAG Nash equilibrium if:

$$J_{P,R}^{\gamma,N,(i)}(\mu^*, \boldsymbol{\pi}^*) = \max_{\pi \in \Pi} J_{P,R}^{\gamma,N,(i)}(\mu^*, (\pi, \boldsymbol{\pi}^{*,-i})). \quad (N\text{-Stat-SAG-NE})$$

If instead $J_{P,R}^{\gamma,N,(i)}(\mu^*, \boldsymbol{\pi}^*) \geq \max_{\pi \in \Pi} J_{P,R}^{\gamma,N,(i)}(\mu^*, (\pi, \boldsymbol{\pi}^{*,-i})) - \delta$, then we call μ^*, π^* a δ-N-Stat-SAG Nash equilibrium.

Theorem 3.5 (Approximation of N-Stat-SAG). *Let $(\mathcal{S}, \mathcal{A}, H, P, R, \gamma)$ be a Stat-MFG and $(\mu^*, \pi^*) \in \Delta_{\mathcal{S}} \times \Pi$ be a corresponding Stat-MFG-NE. Furthermore, assume that $\Gamma_P(\cdot, \pi)$ is non-expansive in the ℓ_1 norm for any π, that is, $\|\Gamma_P(\mu, \pi) - \Gamma_P(\mu', \pi)\|_1 \leq \|\mu - \mu'\|_1$. Then, $(\mu^*, \boldsymbol{\pi}^*) \in \Delta_{\mathcal{S}} \times \Pi^N$ is a $O\left(\frac{1}{\sqrt{N}}\right)$ Nash equilibrium for the N-player game where $\boldsymbol{\pi}^*_N := (\pi^*, \ldots, \pi^*)$, that is, for all i,*

$$J_{P,R}^{\gamma,N,(i)}(\mu^*, \boldsymbol{\pi}^*_N) \geq \max_{\pi \in \Pi} J_{P,R}^{\gamma,N,(i)}(\mu^*, (\pi, \boldsymbol{\pi}^{*,-i}_N)) - O\left(\frac{(1-\gamma)^{-3}}{\sqrt{N}}\right).$$

Proof. *(sketch)* Let (μ^*, π^*) be a Stat-MFG-NE. The proof method is very similar to the FH-MFG case: we first bound the expected deviation from the stable distribution μ^* given by $\mathbb{E}[\|\widehat{\mu} - \mu^*\|_1]$. The truncated expected rewards can be controlled using similar arguments to the FH-MFG case, and an application of the dominated convergence theorem yields the exploitability for the infinite horizon discounted setting. □

We also establish an approximation lower bound for the N-Stat-SAG. In this case, the question is if the non-expansive Γ_P assumption is necessary for the optimal $O(1/\sqrt{N})$ rate. The below results affirm this: in for Stat-MFG-NE with expansive Γ_P, we suffer from an exploitability of $\omega(1/\sqrt{N})$ in the N-agent case.

Theorem 3.6 (Lower bound for N-Stat-SAG). *For any $N \in \mathbb{N}_{>0}, \gamma \in (1/\sqrt{2}, 1)$ there exists $\mathcal{S}, \mathcal{A}$ with $|\mathcal{S}| = 6, |\mathcal{A}| = 2$ and $P \in \mathcal{P}_7, R \in \mathcal{R}_3$ such that:*

1. *The Stat-MFG $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ has a* unique *NE μ^*, π^*,*
2. *For any N and $\boldsymbol{\pi}^*_N := (\pi^*, \ldots, \pi^*) \in \Pi^N$, it holds that $J_{P,R}^{\gamma,N,(i)}(\boldsymbol{\pi}^*_N) \leq \max_\pi J_{P,R}^{\gamma,N,(i)}(\pi, \boldsymbol{\pi}^{*,-i}_N) - \Omega(N^{-\log_2 \gamma^{-1}})$.*

Proof. *(sketch)* The counter-example will be similar to the case in the FH-MFG, with minor modifications to make the Stat-MFG-NE unique. Intuitively, due to the same anti-concentration bound as before for $T = \log_2 \sqrt{N}$, at times $t = 0, 2, 4, \ldots, T-1$ the population deviation from μ^* can be lower bounded by $\mathbb{E}[\|\widehat{\mu}_t - \mu^*\|_1] \geq \Omega(\frac{2^t}{\sqrt{N}})$. By the design of reward functions, this yields an exploitability of

$$\Omega\left(\frac{1 + 2\gamma^2 + \ldots + (2\gamma^2)^{T-1}}{\sqrt{N}}\right) = \Omega\left(N^{-\log_2 \gamma^{-1}}\right).$$

The proof is postponed to the supplementary material. □

The result above shows that unless the relevant Γ_P operator is contracting in some potential, in general, the exploitability of the Stat-MFG-NE in the N-player game might be very large unless the effective horizon $(1-\gamma)^{-1}$ is small. Hence, in these cases, the mean-field Nash equilibrium might be uninformative regarding the true NE of the N player game. In the case of Stat-MFG, our lower bound is even stronger in the sense that the exploitability no longer decreases with $O(1/\sqrt{N})$ for large γ. For a sufficiently long effective horizon $(1-\gamma)^{-1}$ and large enough Lipschitz constant L, the rate in terms of N can be arbitrarily slow. Furthermore, if we take the ergodic limit $\gamma \to 1$, we will observe a non-vanishing exploitability $\Omega(1)$ for *all* finite N.

4 COMPUTATIONAL TRACTABILITY OF MFG

The next fundamental question for mean-field reinforcement learning will be whether it is always computationally easier than finding an equilibrium of a N-player general sum normal form game. We focus on the computational aspect of solving mean-field games in this section, and not statistical uncertainty: we assume we have full knowledge of the MFG dynamics. We will show that unless additional assumptions are introduced (as typically done in the form of contractivity or monotonicity), solving MFG can in general be as hard as finding N-player general sum Nash.

We will prove that the problems are PPAD-complete, where PPAD is a class of computational problems studied in the seminal work by Papadimitriou [22], containing the complete problem of finding N-player Nash equilibrium in general sum normal form games and finding the fixed point of continuous maps [5, 7]. The class PPAD is conjectured to contain difficult problems with no polynomial time algorithms [2, 11], hence our results can be seen as a proof of difficulty. Our results are significant since they imply that the MFG problems studied in literature are in the same complexity class as general-sum N-player normal form games or N-player Markov games [8]. Once again, several computation-intensive aspects of our proofs will be postponed to the appendix [36].

Due to a technical detail, we will prove the complexity results for a subset of possible reward and transition probability functions. We formalize this subset of possible rewards and dynamics as "simple" rewards/dynamics and also linear rewards, defined below.

Definition 4.1 (Simple/Linear Dynamics and Rewards). $R \in \mathcal{R}$ and $P \in \mathcal{P}$ are said to be *simple* if for any $s, s' \in \mathcal{S}, a \in \mathcal{A}$, $P(s'|s, a, \mu)$ and $R(s, a, \mu)$ are functions of μ that are expressible as finite combinations of arithmetic operations $+, -, \times, \div$ and functions $\max\{\cdot,\cdot\}, \min\{\cdot,\cdot\}$ of coordinates of μ. They are called *linear* if $P(s'|s, a, \mu)$ and $R(s, a, \mu)$ are linear functions of μ for all s, a, s'. The set of simple rewards and dynamics are denoted by $\mathcal{R}^{\text{Sim}}$ and $\mathcal{P}^{\text{Sim}}$ respectively, and the set of linear rewards and transitions are denoted $\mathcal{R}^{\text{Lin}}, \mathcal{P}^{\text{Lin}}$ respectively.

A note on simple functions. We define simple functions as above as in general there is no known efficient encoding of a Lipschitz continuous function as a sequence of bits. This is significant since a Turing machine accepts a finite sequence of bits as input. To solve this issue, we prove a slightly stronger hardness result that even games where $P(s'|s,a,\mu), R(s,a,\mu)$ are Lipschitz functions with strong structure are PPAD-complete. Since we are proving hardness, other larger classes of P, R including $\mathcal{P}^{\text{Sim}}, \mathcal{R}^{\text{Sim}}$ will have similar intractability. See also arithmetic circuits with max, min gates [9] for a similar idea.

4.1 The Complexity Class PPAD

The PPAD class is defined by the complete problem End-of-The-Line [7], whose formal definition we defer to the appendix [36] as it is not used in our proofs.

Definition 4.2 (PPAD, PPAD-hard, PPAD-complete). The class PPAD is defined as all search problems that can be reduced to End-of-The-Line in polynomial time. If End-of-The-Line can be reduced to a search problem $\mathcal{S}$ in polynomial time, then $\mathcal{S}$ is called PPAD-hard. A search problem $\mathcal{S}$ is called PPAD-complete if it is both a member of PPAD and it is PPAD-hard.

While End-of-the-Line defines the problem class PPAD, it is hard to construct direct reductions to it. We will instead use two problems that are known to be PPAD-complete (and hence can be equivalently used to define PPAD): solving generalized circuits and finding a NE for an N-player general sum game.

Definition 4.3 (Generalized Circuits [8, 27]). A generalized circuit $C = (\mathcal{V}, \mathcal{G})$ is a finite set of nodes $\mathcal{V}$ and gates $\mathcal{G}$. Each gate $G \in \mathcal{G}$ is characterized by the tuple $G(\theta|v_1, v_2|v)$ where $G \in \{G_{\leftarrow}, G_{\times,+}, \mathcal{G}_{<}\}$, $\theta \in \mathbb{R}^{\star}$ is a parameter (possibly of length 0), $v_1, v_2 \in V \cup \{\perp\}$ are the input nodes (with $\perp$ indicating an empty input) and $v \in V$ it the output node of the gate. The collection of gates $\mathcal{G}$ satisfies the property that if $G_1(\theta|v_1, v_2|v), G_2(\theta'|v_1', v_2'|v') \in G$ are distinct gates, then $v \neq v'$.

Such circuits define a set of constraints on values assigned to each gate, and finding such an assignment will be the associated computational problem for such a circuit desription. We formally define the ε-GCircuit problem to this end. ε-GCircuit is a standard complete problem for the class PPAD, and we will work with it for our reductions. We will use the shorthand notation $x = y \pm \varepsilon$ to indicate that $x \in [y - \varepsilon, y + \varepsilon]$ for $x, y \in \mathbb{R}$.

Definition 4.4 (ε-GCircuit [27]). Given a generalized circuit $C = (\mathcal{V}, \mathcal{G})$, a function $p : V \to [0, 1]$ is called an ε-satisfying assignment if:

- For every gate $G \in \mathcal{G}$ of the form $G_{\leftarrow}(\zeta||v)$ for $\zeta \in 0, 1$, it holds that $p(v) = \zeta \pm \varepsilon$,
- For every gate $G \in \mathcal{G}$ of the form $G_{\times,+}(\alpha, \beta|v_1, v_2|v)$ for $\alpha, \beta \in [-1, 1]$, it holds that

$$p(v) \in [\max\{\min\{0, \alpha p(v_1) + \beta p(v_2)\}\}] \pm \varepsilon,$$

- For every gate $G \in \mathcal{G}$ of the form $G_{<}(|v_1, v_1|v)$ it holds that

$$p(v) = \begin{cases} 1 \pm \varepsilon, & p(v_1) \leq p(v_2) - \varepsilon, \\ 0 \pm \varepsilon, & p(v_1) \geq p(v_2) + \varepsilon. \end{cases}$$

The ε-GCircuit problem is defined as follows:

Given generalized circuit C, find an ε-satisfying assignment of C.

ε-GCircuit is one of the prototypical hard instances of PPAD problems as the result below suggests.

Theorem 4.5. *[27] There exists $\varepsilon > 0$ such that ε-GCircuit is PPAD-complete.*

In other words, ε-GCircuit is representative of the most difficult problem in PPAD which suggests intractability. The ε-GCircuit computational problem will be used in our proofs by reducing an arbitrary generalized circuit into solving a particular MFG.

We will also use the general sum 2-player Nash computation problem, which is the standard problem of finding an approximate Nash equilibrium of a general sum bimatrix game.

Definition 4.6 (2-Nash). Given $\varepsilon > 0$, $K_1, K_2 \in \mathbb{N}_{>0}$, payoff matrices $A, B \in [0, 1]^{K_1, K_2}$, find an approximate Nash equilibrium $(\sigma_1, \sigma_2) \in \Delta_{K_1} \times \Delta_{K_2}$ such that

$$\max_{\sigma \in \Delta_{K_1}} \sum_{i \in [K_1]} \sum_{j \in [K_2]} A_{i,j} \sigma(i) \sigma_2(j) - \sum_{i \in [K_1]} \sum_{j \in [K_2]} A_{i,j} \sigma_1(i) \sigma_2(j) \leq \varepsilon$$

$$\max_{\sigma \in \Delta_{K_2}} \sum_{i \in [K_2]} \sum_{a \in [K_2]} B_{i,j} \sigma_1(i) \sigma(j) - \sum_{i \in [K_1]} \sum_{j \in [K_2]} B_{i,j} \sigma_1(i) \sigma_2(j) \leq \varepsilon$$

The following is the well-known result that even the 2-Nash general sum problem is PPAD-complete. In fact, any N-player general sum normal form game is PPAD-complete.

Theorem 4.7. *[5] 2-Nash is PPAD-complete.*

4.2 Complexity of Stat-MFG

Next, we provide our difficulty results for the Stat-MFG problem. Notably, for Stat-MFG, the stability subproblem of finding a stable distribution for a fixed policy π itself is PPAD-hard. Even without considering the optimality conditions, finding a stable distribution in general for a fixed policy is intractable, unless additional assumptions are introduced (e.g. Γ_P is contractive or non-expansive). We define the computational problem below and state the results.

Definition 4.8 (ε-StatDist). Given finite state-action sets $\mathcal{S}, \mathcal{A}$, simple dynamics $P \in \mathcal{P}^{\text{Sim}}$ and policy π, find $\mu^* \in \Delta_{\mathcal{S}}$ such that $\|\Gamma_P(\mu^*, \pi) - \mu^*\|_\infty \leq \frac{\varepsilon}{|\mathcal{S}|}$.

The computational problem as described above is to find an approximate fixed point of $\Gamma_P(\cdot, \pi)$ which corresponds to an approximately stable distribution of policy π. We show that ε-StatDist is PPAD-complete for some fixed constant ε.

Theorem 4.9 (ε-StatDist is PPAD-complete). *For some $\varepsilon > 0$, the problem ε-StatDist is PPAD-complete.*

Proof. *(sketch)* The reduction from ε-StatDist to a fixed point problem (or the Sperner problem [7]) is straightforward, showing ε-StatDist is in PPAD. The main challenge of the proof is showing ε-StatDist is simultaneously PPAD-hard. This is achieved by showing any ε-GCircuit problem can be reduced to a ε-StatDist for some ε'. For simplicity, we reduce ε-GCircuit to finding the stable distribution of a transition kernel $P(s'|s, \mu)$. Given a generalized circuit $C = (\mathcal{V}, \mathcal{G})$, we construct a Stat-MFG that has one

base state s_{base}, one additional state s_v for each $v \in \mathcal{V}$ that is the output of a gate. Let $\theta := \frac{1}{8V}, B := \frac{1}{4}$. Also define the function $u_\alpha(x) := \max\{0, \min\{\alpha, x\}\}$ for any $\alpha \in [0,1]$. We present the construction and defer the analysis to the appendix: any gate of the form $G_{\leftarrow}(\zeta||v)$, we will add one state s_v such that $P(s_{\text{base}}|s_v, \mu) = 1$, $P(s_v|s_{\text{base}}, \mu) = \frac{\zeta\theta}{\max\{B, \mu(s_{\text{base}})\}}$. For any weighted addition gate $G_{\times,+}(\alpha, \beta|v_1, v_2|v)$, we add a state s_v such that $P(s_{\text{base}}|s_v, \mu) = 1$ and $P(s_v|s_{\text{base}}, \mu) = \frac{u_\theta(\alpha\mu(v_1)+\beta\mu(v_2))}{\max\{B, \mu(s_{\text{base}})\}}$. Finally, for each comparison gate $G_<(|v_1, v_1|v)$, also add a state s_v and define the transition probabilities:

$$P(s_v|s_{\text{base}}, \mu) = \frac{\theta p_{\varepsilon/8}(\theta^{-1}\mu(s_1), \theta^{-1}\mu(s_2))}{\max\{B,\ \mu(s_{\text{base}})\}},$$
$$P(s_v|s_v, \mu) = 0, \quad P(s_{\text{base}}|s_v, \mu) = 1,$$

where $p_\varepsilon(x, y) := u_1\left(\frac{1}{2} + \varepsilon^{-1}(x - y)\right)$. Once all gates are added, the construction is completed by defining $P(s_{\text{base}}|s_{\text{base}}, \mu) = 1 - \sum_{s' \in \mathcal{S}} P(s'|s_{\text{base}}, \mu)$. Simple computation verifies that for any *exact* stationary distribution μ^* of the above P, an exact assignment the the generalized circuit can be read by the map $v \to u_1(\frac{\mu^*(s_v)}{\theta})$. □

As a corollary, there is no polynomial time algorithm for ε-StatDist unless PPAD=P, which is conjectured to be not the case.

Corollary 4.10. *There exists a $\varepsilon > 0$ such that there exists no polynomial time algorithm for ε-StatDist, unless P = PPAD.*

Most notably, these results show that the stable distribution oracle of [6] might be intractable to compute in general, and the shared assumption that $\Gamma_P(\cdot, \pi)$ is contractive in some norm found in many works [1, 33, 34] might not be trivial to remove without sacrificing tractability.

4.3 Complexity of FH-MFG

We will show that finding an ε solution to the finite horizon problem is also PPAD-complete, in particular even if we restrict our attention to the case when $H = 2$ and the transition probabilities P do not depend on μ. We formalize the structured computational FH-MFG problem.

Definition 4.11 ((ε, H)-FH-Nash). Given simple reward function $R \in \mathcal{R}^{\text{Sim}}$, transition matrix $P(s'|s, a)$, and initial distribution $\mu_0 \in \Delta_{\mathcal{S}}$, find a time dependent policy $\{\pi_h\}_{h=0}^{H-1}$ such that $\mathcal{E}^H_{P,R}(\{\pi_h\}_{h=0}^{H-1}) \leq \varepsilon/|\mathcal{S}|$.

Our result in the case of the finite horizon MFG problem is that even in the case of $H = 2$, the problem is PPAD-complete.

Theorem 4.12 ($(\varepsilon, 2)$-FH-Nash is PPAD-complete). *There exists an $\varepsilon > 0$ such that the problem $(\varepsilon, 2)$-FH-Nash is PPAD-complete.*

Proof. *(sketch)* Once again, showing $(\varepsilon, 2)$-FH-Nash is in PPAD is simple: it follows from the fact that a FH-MFG-NE is a fixed point of an easy-to-compute function (see e.g. [15]). To show that $(\varepsilon, 2)$-FH-Nash is also PPAD-hard, for an arbitrary generalized circuit $C = (\mathcal{V}, \mathcal{G})$ we construct a FH-MFG whose δ-NE will be δ'-satisfying assignments for C for some δ'. □

Corollary 4.13. *There exists a $\varepsilon > 0$ such that there exists no polynomial time algorithm for $(\varepsilon, 2)$-FH-Nash, unless P= PPAD.*

These results for the FH-MFG show that the (weak) monotonicity assumption present in works such as [23, 25] might also be necessary, as in the absence of any structural assumptions the problems are provably difficult.

Finally, we also show that even if $R(s, a, \mu)$ is a linear function of μ for all s, a (that is, $R \in \mathcal{R}^{\text{Lin}}$), the intractability holds, although not for fixed ε. We define the linear computational problem below.

Definition 4.14 (H-FH-Linear). Given $\varepsilon > 0$, linear reward function $R \in \mathcal{R}^{\text{Lin}}$, transition matrix $P(s'|s, a)$, find a time dependent policy $\{\pi_h\}_{h=0}^{H-1}$ such that $\mathcal{E}^H_{P,R}(\{\pi_h\}_{h=0}^{H-1}) \leq \varepsilon$.

Theorem 4.15 (2-FH-Linear is PPAD-complete). *The problem 2-FH-Linear is PPAD-complete.*

Proof. *(sketch)* In this case, we provide a reduction from 2-Nash. For a given 2-Nash instance $K_1, K_2 \in \mathbb{N}_{>0}$ with payoff matrices $A, B \in [0,1]^{K_1, K_2}$, we construct an FH-MFG with one initial state for each player and one additional state for each strategy of each of the players, resulting in a FH-MFG with $K_1 + K_2 + 2$ states, $\mathcal{S} := \{s^1_{\text{base}}, s^2_{\text{base}}, s^1_1, \ldots, s^1_{K_1}, s^2_1, \ldots, s^2_{K_2}\}$. We set $\mu_0(s^1_{\text{base}}) = \mu_0(s^2_{\text{base}}) = 1/2$. The action set will consist of $\max\{K_1, K_2\}$ actions. In the first round, an agent starting from s^1_{base} will be transitioned to one of states $s^1_1, \ldots, s^1_{K_1}$ depending on the action picked receiving zero reward, and likewise and agent starting from s^2_{base} will transition to one of states $s^2_1, \ldots, s^2_{K_2}$. In the second round, the agent will receive a population-dependent reward regardless of the action player, which is equal to the expected utility of an action (a linear function). We postpone the cumbersome details relating to error analysis and dealing with the case $K_1 \neq K_2$ to the appendix. □

We emphasize that for 2-FH-Linear the accuracy ε is also an input of the problem: hence the existence of a pseudo-polynomial time algorithm is not ruled out.

5 DISCUSSION AND CONCLUSION

We provided novel results on when mean-field RL is relevant for real-world applications and when it is tractable from a computational perspective. Our results differ from existing work by provably characterizing cases where MFGs might have practical shortcomings. From the approximation perspective, we show clear conditions and lower bounds on when the MFGs efficiently approximate real-world games. Computationally, we show that even simple MFGs can be as hard as solving N-player general sum games.

We emphasize that our results do not discard MFGs, but rather identify potential bottlenecks (and conditions to overcome these) when using mean-field RL to compute a good approximate NE.

ACKNOWLEDGMENTS

This project is supported by Swiss National Science Foundation (SNSF) under the framework of NCCR Automation and SNSF Starting Grant. A.Goldman was supported by the ETH Student Summer Research Fellowship.

REFERENCES

[1] Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. 2022. Q-learning in regularized mean-field games. *Dynamic Games and Applications* (2022), 1–29.

[2] Paul Beame, Stephen Cook, Jeff Edmonds, Russell Impagliazzo, and Toniann Pitassi. 1995. The relative complexity of NP search problems. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing*. Las Vegas, Nevada, USA, 303–314.

[3] René Carmona and François Delarue. 2013. Probabilistic analysis of mean-field games. *SIAM Journal on Control and Optimization* 51, 4 (2013), 2705–2734.

[4] René Carmona, François Delarue, et al. 2018. *Probabilistic theory of mean field games with applications I-II.* Springer.

[5] Xi Chen, Xiaotie Deng, and Shang-Hua Teng. 2009. Settling the complexity of computing two-player Nash equilibria. *Journal of the ACM (JACM)* 56, 3 (2009), 1–57.

[6] Kai Cui and Heinz Koeppl. 2021. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 1909–1917.

[7] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. 2009. The complexity of computing a Nash equilibrium. *Commun. ACM* 52, 2 (2009), 89–97.

[8] Constantinos Daskalakis, Noah Golowich, and Kaiqing Zhang. 2023. The complexity of markov equilibrium in stochastic games. In *The Thirty Sixth Annual Conference on Learning Theory*. PMLR, 4180–4234.

[9] Constantinos Daskalakis and Christos Papadimitriou. 2011. Continuous local search. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*. SIAM, 790–804.

[10] Matthieu Geist, Julien Pérolat, Mathieu Laurière, Romuald Elie, Sarah Perrin, Oliver Bachem, Rémi Munos, and Olivier Pietquin. 2022. Concave Utility Reinforcement Learning: The Mean-field Game Viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems* (Virtual Event, New Zealand) *(AAMAS '22)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 489–497.

[11] Paul W Goldberg. 2011. A survey of PPAD-completeness for computing Nash equilibria. *arXiv preprint arXiv:1103.2709* (2011).

[12] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. 2019. Learning mean-field games. *Advances in Neural Information Processing Systems* 32 (2019).

[13] Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. 2022. A general framework for learning mean-field games. *Mathematics of Operations Research* (2022).

[14] Xin Guo, Anran Hu, and Junzi Zhang. 2022. MF-OMO: An optimization formulation of mean-field games. *arXiv preprint arXiv:2206.09608* (2022).

[15] Jiawei Huang, Batuhan Yardim, and Niao He. 2023. On the Statistical Efficiency of Mean Field Reinforcement Learning with General Function Approximation. *arXiv preprint arXiv:2305.11283* (2023).

[16] Minyi Huang, Roland P Malhamé, and Peter E Caines. 2006. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle. *Communications in Information & Systems* 6, 3 (2006), 221–252.

[17] Krishnamurthy Iyer, Ramesh Johari, and Mukund Sundararajan. 2014. Mean field equilibria of dynamic auctions with learning. *Management Science* 60, 12 (2014), 2949–2970.

[18] Jean-Michel Lasry and Pierre-Louis Lions. 2007. Mean field games. *Japanese journal of mathematics* 2, 1 (2007), 229–260.

[19] Mathieu Laurière, Sarah Perrin, Sertan Girgin, Paul Muller, Ayush Jain, Théophile Cabannes, Georgios Piliouras, Julien P'erolat, Romuald Elie, Olivier Pietquin, and Matthieu Geist. 2022. Scalable Deep Reinforcement Learning Algorithms for Mean Field Games. In *International Conference on Machine Learning*.

[20] Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, Ravi Iyer, and Tamer Basar. 2022. A Mean-Field Game Approach to Cloud Resource Management with Function Approximation. In *Advances in Neural Information Processing Systems*.

[21] Laëtitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. 2007. Hysteretic q-learning: an algorithm for decentralized reinforcement learning in cooperative multi-agent teams. In *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 64–69.

[22] Christos H Papadimitriou. 1994. On the complexity of the parity argument and other inefficient proofs of existence. *Journal of Computer and system Sciences* 48, 3 (1994), 498–532.

[23] Julien Pérolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. 2022. Scaling Mean Field Games by Online Mirror Descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*. 1028–1037.

[24] Julien Perolat, Bruno Scherrer, Bilal Piot, and Olivier Pietquin. 2015. Approximate dynamic programming for two-player zero-sum markov games. In *International Conference on Machine Learning*. PMLR, 1321–1329.

[25] Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. 2020. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems* 33 (2020), 13199–13213.

[26] Navid Rashedi, Mohammad Amin Tajeddini, and Hamed Kebriaei. 2016. Markov game approach for multi-agent competitive bidding strategies in electricity market. *IET Generation, Transmission & Distribution* 10, 15 (2016), 3756–3763.

[27] Aviad Rubinstein. 2015. Inapproximability of Nash equilibrium. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. 409–418.

[28] Naci Saldi, Tamer Basar, and Maxim Raginsky. 2018. Markov–Nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization* 56, 6 (2018), 4256–4287.

[29] Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. 2019. The StarCraft Multi-Agent Challenge. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)* (Montreal QC, Canada) *(AAMAS '19)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 2186–2188.

[30] Ali Shavandi and Majid Khedmati. 2022. A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets. *Expert Systems with Applications* 208 (2022), 118124.

[31] Lingxiao Wang, Zhuoran Yang, and Zhaoran Wang. 2020. Breaking the curse of many agents: Provable mean embedding Q-iteration for mean-field reinforcement learning. In *International conference on machine learning*. PMLR, 10092–10103.

[32] Marco A. Wiering. 2000. Multi-agent reinforcement learning for traffic light control. In *Machine Learning: Proceedings of the Seventeenth International Conference (ICML'2000)*. 1151–1158.

[33] Qiaomin Xie, Zhuoran Yang, Zhaoran Wang, and Andreea Minca. 2021. Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*. PMLR, 11436–11447.

[34] Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. 2023. Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*. PMLR, 39722–39754.

[35] Batuhan Yardim, Semih Cayci, and Niao He. 2023. Stateless Mean-Field Games: A Framework for Independent Learning with Large Populations. In *Sixteenth European Workshop on Reinforcement Learning*.

[36] Batuhan Yardim, Artur Goldman, and Niao He. 2024. When is Mean-Field Reinforcement Learning Tractable and Relevant? https://arxiv.org/abs/2402.05757. arXiv:2402.05757

[37] Muhammad Aneeq Uz Zaman, Alec Koppel, Sujay Bhatt, and Tamer Basar. 2023. Oracle-free reinforcement learning in mean-field games along a single sample path. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 10178–10206.