When is Mean-Field Reinforcement Learning Tractable and Relevant?

Anonymous Authors[1]

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.

1. Introduction

Multi-agent reinforcement learning (MARL) finds numerous impactful applications in the real world (Shavandi & Khedmati, 2022; Wiering, 2000; Samvelyan et al., 2019; Rashedi et al., 2016; Matignon et al., 2007; Mao et al., 2022). 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" (Wang et al., 2020).

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the Workshop on Foundations of Reinforcement Learning and Control at the International Conference on Machine Learning (ICML). Do not distribute.

Mean-field games (MFG), a theoretical framework first proposed by Lasry & Lions (2007) and Huang et al. (2006), 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 (Iyer et al., 2014), and cloud resource management (Mao et al., 2022). 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 (Perrin et al., 2020; Anahtarci et al., 2022; Guo et al., 2019; Pérolat et al., 2022; Xie et al., 2021) 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. (Anahtarci et al., 2022; Guo et al., 2019)) and finite-horizon

MFG (FH-MFG, see e.g. (Perrin et al., 2020; Pérolat et al., 2022)). 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 (Carmona & Delarue, 2013; Carmona et al., 2018), albeit in the differential setting and without explicit lower bounds. Recent works (Anahtarci et al., 2022; Cui & Koeppl, 2021) 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. (2018), 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 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, (Yardim et al., 2023b) 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 (Daskalakis et al., 2009; Chen et al., 2009), 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 (Zaman et al., 2023; Anahtarci et al., 2022), policy mirror ascent (Yardim et al., 2023a), policy gradient methods (Guo et al., 2022a), soft Q-learning (Cui & Koeppl, 2021) and fictitious play (Xie et al., 2021) 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 (Cui & Koeppl, 2021; Yardim et al., 2023a), 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. (Guo et al., 2019; Yardim et al., 2023a)), further restricting the analysis to sufficiently smooth games. While (Guo et al., 2022b) 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 (Perrin et al., 2020) and mirror descent (Pérolat et al., 2022) 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 (Huang et al., 2023), 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.

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 $(\boldsymbol{\pi}', \boldsymbol{\pi}^{-i}) \in \Pi_H^N$ 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 (Guo et al., 2019; Anahtarci et al., 2022; Yardim et al., 2023a; Xie et al., 2021), 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 $\mathcal{R} := \bigcup_{L\geq 0} \mathcal{R}_L, \quad \mathcal{P} := \bigcup_{L \geq 0} \mathcal{P}_L$ 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

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

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

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

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 (Lemma 3.2 of (Yardim et al., 2023a))**.** *Let $P \in \mathcal{P}_{K_\mu}$ for $K_\mu > 0$ and*

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

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

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

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

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

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

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

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 (Anahtarci et al., 2022; Yardim et al., 2023a; Guo et al., 2019; 2022a; Xie et al., 2021). 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

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

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

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

The second MFG concept that we will consider has a finite time horizon, and is also common in literature (Perolat et al., 2015; Perrin et al., 2020; Laurière et al., 2022; Huang et al., 2023). 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

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

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

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

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.

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 $\boldsymbol{\pi} = \{\pi_h^i\}_{h=0,\dots,H-1, i\in[N]} \in \Pi_H^N$, we define the expected mean reward and exploitability of player i as

$$J_{P,R}^{H,N,(i)}(\boldsymbol{\pi}) := \mathbb{E}\left[\sum_{h=0}^{H-1} R(s_h^i, a_h^i, \widehat{\mu}_h) \middle| \substack{\forall j: s_0^j \sim \mu_0, a_h^j \sim \pi_h^j(s_h^j) \\ \widehat{\mu}_h := \frac{1}{N}\sum_j \mathbf{e}_{s_h^j} \\ s_{h+1}^j \sim P(s_h^j, a_h^j \widehat{\mu}_h)}\right],$$
$$\mathcal{E}_{P,R}^{H,N,(i)}(\boldsymbol{\pi}) := \max_{\boldsymbol{\pi}' \in \Pi^H} J_{P,R}^{H,N,(i)}(\boldsymbol{\pi}', \boldsymbol{\pi}^{-i}) - J_{P,R}^{H,N,(i)}(\boldsymbol{\pi}).$$

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

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

If instead $\mathcal{E}_{P,R}^{H,N,(i)}(\boldsymbol{\pi}) \leq \delta$ 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 $J_{P,R}^{H,N,(i)}$ 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 $\mathcal{F}(\{s_h^i\}_i) = \mathcal{F}_h$-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 $i \in [N]$, $\mathcal{E}_{P,R}^{H,N,(i)}(\boldsymbol{\pi}_N^*) \leq \mathcal{O}(\frac{H^3}{\sqrt{N}})$, that is, $\boldsymbol{\pi}_N^*$ is a $\mathcal{O}(\frac{H^3}{\sqrt{N}})$-NE of $\mathcal{G}_N$.*
2. *If $L \neq 1$, then for all $i \in [N]$, $\mathcal{E}_{P,R}^{H,N,(i)}(\boldsymbol{\pi}_N^*) \leq \mathcal{O}\left(\frac{H^2(1-L^H)}{(1-L)\sqrt{N}}\right)$, that is, $\boldsymbol{\pi}_N^*$ is a $\mathcal{O}\left(\frac{H^2(1-L^H)}{(1-L)\sqrt{N}}\right)$-NE of $\mathcal{G}_N$.*

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 $\mathcal{O}\left((1 + L^H)H^2/\sqrt{N}\right)$ 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.

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 $\mathbb{E}_H[\|\widehat{\mu}_h - \Lambda_P^H(\mu_0, \boldsymbol{\pi}_H^*)_h\|_1] \geq \Omega\left(\min\{1, \frac{2^H}{\sqrt{N}}\}\right)$.*
3. *For any $H, N > 0$ either $N \geq \Omega(2^H)$, or for each player $i \in [N]$ it holds that $\mathcal{E}_{P,R}^{H,N,(i)}(\boldsymbol{\pi}_H^*, \dots, \boldsymbol{\pi}_H^*) \geq \Omega(H)$.*

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. (Perrin et al., 2020; Geist et al., 2022)) 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 $N \geq \Omega((L-\epsilon)^H)$ 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: $\forall i \in [N]$: $H^{-1}\mathcal{E}_{P,R}^{H,N,(i)}(\overline{\boldsymbol{\pi}}^\star_H, \dots, \overline{\boldsymbol{\pi}}^\star_H) \leq \mathcal{O}(H^{-1}\log_2 N)$.

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:

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

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

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

If $J_{P,R}^{\gamma,N,(i)}(\mu^*, \boldsymbol{\pi}^*) \geq \max_{\pi \in \Pi} J_{P,R}^{\gamma,N,(i)}(\mu^*, (\pi, \boldsymbol{\pi}^{*,-i})) - \delta$, 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 $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$ Nash equilibrium for the N-player game where $\boldsymbol{\pi}^*_N := (\pi^*, \dots, \pi^*)$, that is, for all i,*

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

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 $\mathcal{O}(1/\sqrt{N})$ 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 $\boldsymbol{\pi}^*_N := (\pi^*, \dots, \pi^*) \in \Pi^N$, it holds that $J_{P,R}^{\gamma,N,(i)}(\boldsymbol{\pi}^*_N) \leq \max_\pi J_{P,R}^{\gamma,N,(i)}(\pi, \boldsymbol{\pi}_N^{*,-i}) - \Omega(N^{-\log_2 \gamma^{-1}})$.*

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 $\mathcal{O}(1/\sqrt{N})$ 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 (1994), containing the complete problem of finding N-player Nash equilibrium in general sum normal form games and finding the fixed point of continuous maps (Daskalakis et al., 2009; Chen et al., 2009). The class PPAD is conjectured to contain difficult problems with no polynomial time algorithms (Beame et al., 1995; Goldberg, 2011), 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 (Daskalakis et al., 2023). Once again, several computation-intensive aspects of our proofs will be postponed to the supplementary material.

Due to a technicality, 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. 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 (Daskalakis & Papadimitriou, 2011) for a similar idea.

4.1. The Complexity Class PPAD

The PPAD class is defined by the complete problem END-OF-THE-LINE (Daskalakis et al., 2009), whose formal definition we defer to the appendix as it is not used in 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 (Rubinstein, 2015))**.** A generalized circuit $\mathcal{C} = (\mathcal{V}, \mathcal{G})$ 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 $v \in V$ is the output node of the gate. The collection $\mathcal{G}$ satisfies the property that if $G_1(\theta|v_1, v_2|v), G_2(\theta'|v_1', v_2'|v') \in \mathcal{G}$ are distinct, 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 (Rubinstein, 2015))**.** Given a generalized circuit $\mathcal{C} = (\mathcal{V}, \mathcal{G})$, 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 $\mathcal{C}$,
find an ε-satisfying assignment of $\mathcal{C}$.

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

Theorem 4.5. *(Rubinstein, 2015) 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 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

$$\max_{\sigma \in \Delta_{K_1}} U_A(\sigma, \sigma_2) - U_A(\sigma_1, \sigma_2) \leq \varepsilon,$$

$$\max_{\sigma \in \Delta_{K_2}} U_B(\sigma_1, \sigma) - U_B(\sigma_1, \sigma_2) \leq \varepsilon,$$

where $U_M(\sigma_1, \sigma_1) := \sum_{i \in [K_1]} \sum_{j \in [K_2]} M_{i,j} \sigma_1(i) \sigma_2(j)$ for any matrix $M \in [0,1]^{K_1,K_2}$.

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. *(Chen et al., 2009)* 2-NASH *is* PPAD-*complete.*

4.2. Complexity of Stat-MFG

Next, we provide our hardness 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, without additional assumptions (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 approximate 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.*

Consequently, 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 (Cui & Koeppl, 2021) 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 (Xie et al., 2021; Anahtarci et al., 2022; Yardim et al., 2023a) 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 $\mathcal{E}_{P,R}^H(\{\pi_h\}_{h=0}^{H-1}) \leq \varepsilon/|\mathcal{S}|$.

Our main result for FH-MFG 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.*

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 (Perrin et al., 2020; Pérolat et al., 2022) might also be necessary, as in the absence of any structural assumptions the problems are provably hard.

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. This follows from a reduction to 2-NASH. 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 $\mathcal{E}_{P,R}^H(\{\pi_h\}_{h=0}^{H-1}) \leq \varepsilon$.

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

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.

References

Anahtarci, B., Kariksiz, C. D., and Saldi, N. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, pp. 1–29, 2022.

Beame, P., Cook, S., Edmonds, J., Impagliazzo, R., and Pitassi, T. The relative complexity of np search problems. In *Proceedings of the twenty-seventh annual ACM symposium on Theory of computing*, pp. 303–314, Las Vegas, Nevada, USA, 1995.

Carmona, R. and Delarue, F. Probabilistic analysis of mean-field games. *SIAM Journal on Control and Optimization*, 51(4):2705–2734, 2013.

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

Chen, X., Deng, X., and Teng, S.-H. Settling the complexity of computing two-player nash equilibria. *Journal of the ACM (JACM)*, 56(3):1–57, 2009.

Cui, K. and Koeppl, H. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 1909–1917. PMLR, 2021.

Daskalakis, C. and Papadimitriou, C. Continuous local search. In *Proceedings of the twenty-second annual ACM-SIAM symposium on Discrete Algorithms*, pp. 790–804. SIAM, 2011.

Daskalakis, C., Goldberg, P. W., and Papadimitriou, C. H. The complexity of computing a nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009.

Daskalakis, C., Golowich, N., and Zhang, K. The complexity of markov equilibrium in stochastic games. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 4180–4234. PMLR, 2023.

Geist, M., Pérolat, J., Laurière, M., Elie, R., Perrin, S., Bachem, O., Munos, R., and Pietquin, O. Concave utility reinforcement learning: The mean-field game viewpoint. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '22, pp. 489–497, Richland, SC, 2022. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450392136.

Goldberg, P. W. A survey of ppad-completeness for computing nash equilibria. *arXiv preprint arXiv:1103.2709*, 2011.

Guo, X., Hu, A., Xu, R., and Zhang, J. Learning mean-field games. *Advances in Neural Information Processing Systems*, 32, 2019.

Guo, X., Hu, A., Xu, R., and Zhang, J. A general framework for learning mean-field games. *Mathematics of Operations Research*, 2022a.

Guo, X., Hu, A., and Zhang, J. Mf-omo: An optimization formulation of mean-field games. *arXiv preprint arXiv:2206.09608*, 2022b.

Huang, J., Yardim, B., and He, N. On the statistical efficiency of mean field reinforcement learning with general function approximation. *arXiv preprint arXiv:2305.11283*, 2023.

Huang, M., Malhamé, R. P., and Caines, P. E. Large population stochastic dynamic games: closed-loop mckean-vlasov systems and the nash certainty equivalence principle. *Communications in Information & Systems*, 6(3): 221–252, 2006.

Iyer, K., Johari, R., and Sundararajan, M. Mean field equilibria of dynamic auctions with learning. *Management Science*, 60(12):2949–2970, 2014.

Kaas, R. and Buhrman, J. M. Mean, median and mode in binomial distributions. *Statistica Neerlandica*, 34(1): 13–18, 1980.

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

Laurière, M., Perrin, S., Girgin, S., Muller, P., Jain, A., Cabannes, T., Piliouras, G., P'erolat, J., Elie, R., Pietquin, O., and Geist, M. Scalable deep reinforcement learning algorithms for mean field games. In *International Conference on Machine Learning*, 2022.

Mao, W., Qiu, H., Wang, C., Franke, H., Kalbarczyk, Z., Iyer, R., and Basar, T. A mean-field game approach to cloud resource management with function approximation. In *Advances in Neural Information Processing Systems*, 2022.

Matignon, L., Laurent, G. J., and Le Fort-Piat, N. 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*, pp. 64–69. IEEE, 2007.

McDiarmid, C. et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

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

Perolat, J., Scherrer, B., Piot, B., and Pietquin, O. Approximate dynamic programming for two-player zero-sum markov games. In *International Conference on Machine Learning*, pp. 1321–1329. PMLR, 2015.

Pérolat, J., Perrin, S., Elie, R., Laurière, M., Piliouras, G., Geist, M., Tuyls, K., and Pietquin, O. Scaling mean field games by online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 1028–1037, 2022.

Perrin, S., Pérolat, J., Laurière, M., Geist, M., Elie, R., and Pietquin, O. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in Neural Information Processing Systems*, 33:13199–13213, 2020.

Rashedi, N., Tajeddini, M. A., and Kebriaei, H. Markov game approach for multi-agent competitive bidding strategies in electricity market. *IET Generation, Transmission & Distribution*, 10(15):3756–3763, 2016.

Rubinstein, A. Inapproximability of nash equilibrium. In *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*, pp. 409–418, 2015.

Saldi, N., Basar, T., and Raginsky, M. Markov–nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

Samvelyan, M., Rashid, T., Schroeder de Witt, C., Farquhar, G., Nardelli, N., Rudner, T. G. J., Hung, C.-M., Torr, P. H. S., Foerster, J., and Whiteson, S. The starcraft multi-agent challenge. In *Proc. of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2019)*, AAMAS '19, pp. 2186–2188, Richland, SC, 2019. International Foundation for Autonomous Agents and Multiagent Systems.

Shavandi, A. and Khedmati, M. A multi-agent deep reinforcement learning framework for algorithmic trading in financial markets. *Expert Systems with Applications*, 208: 118124, 2022.

Wang, L., Yang, Z., and Wang, Z. Breaking the curse of many agents: Provable mean embedding q-iteration for mean-field reinforcement learning. In *International conference on machine learning*, pp. 10092–10103. PMLR, 2020.

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

Xie, Q., Yang, Z., Wang, Z., and Minca, A. Learning while playing in mean-field games: Convergence and optimality. In *International Conference on Machine Learning*, pp. 11436–11447. PMLR, 2021.

Yardim, B., Cayci, S., Geist, M., and He, N. Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*, pp. 39722–39754. PMLR, 2023a.

Yardim, B., Cayci, S., and He, N. Stateless mean-field games: A framework for independent learning with large populations. In *Sixteenth European Workshop on Reinforcement Learning*, 2023b.

Zaman, M. A. U., Koppel, A., Bhatt, S., and Basar, T. Oracle-free reinforcement learning in mean-field games along a single sample path. In *International Conference on Artificial Intelligence and Statistics*, pp. 10178–10206. PMLR, 2023.

A. MFG Approximation Results

A.1. Preliminaries

To establish explicit upper bounds on the approximation rate, we will use standard concentration tools.

Definition A.1 (Sub-Gaussian)**.** Random variable ξ is called sub-Gaussian with variance proxy σ^2 if $\forall \lambda \in \mathbb{R}$: $\mathbb{E}\left[e^{\lambda(\xi - \mathbb{E}[\xi])}\right] \leq e^{\frac{\lambda^2\sigma^2}{2}}$. In this case, we write $\xi \in SG(\sigma^2)$.

It is easy to show that if $\xi \in SG(\sigma^2)$, then $\alpha\xi \in SG(\alpha^2\sigma^2)$ for any constant $\alpha \in \mathbb{R}$. Furthermore, if $\xi_1, \ldots, \xi_n$ are independent random variables with $\xi_i \in SG(\sigma_i^2)$, then $\sum_i \xi_i \in SG(\sum_i \sigma_i^2)$. Finally, if ξ is almost surely bounded in $[a, b]$, then $\xi_i \in SG(^{(b-a)^2}/_4)$. We also state the well-known Hoeffding concentration bound and a corollary, Lemma A.3.

Lemma A.2 (Hoeffding inequality (McDiarmid et al., 1989))**.** *Let $\xi \in SG(\sigma^2)$. Then for any $t > 0$ it holds that* $\mathbb{P}\left(|\xi - \mathbb{E}[\xi]| \geq t\right) \leq 2e^{-\frac{t^2}{2\sigma^2}}$.

Lemma A.3. *Let $\xi \in SG(\sigma^2)$. Then*

$$\mathbb{E}\left[|\xi - \mathbb{E}[\xi]|\right] \leq \sqrt{2\pi\sigma^2}, \quad \mathbb{E}\left[(\xi - \mathbb{E}[\xi])^2\right] \leq 4\sigma^2$$

Proof.

$$\begin{aligned}\mathbb{E}\left[|\xi - \mathbb{E}[\xi]|\right] &= \int_0^\infty \mathbb{P}(|\xi - \mathbb{E}[\xi]| \geq t)dt \\ &\overset{(I)}{\leq} 2\int_0^\infty e^{-\frac{t^2}{2\sigma^2}}dt = \sqrt{2\pi\sigma^2}\end{aligned}$$

Inequality (I) is true due to Lemma A.2. Likewise,

$$\begin{aligned}\mathbb{E}\left[(\xi - \mathbb{E}[\xi])^2\right] &= \int_0^\infty \mathbb{P}((\xi - \mathbb{E}[\xi])^2 \geq t)dt \\ &= \int_0^\infty \mathbb{P}(|\xi - \mathbb{E}[\xi]| \geq \sqrt{h})dt \\ &\overset{(II)}{\leq} 2\int_0^\infty e^{-\frac{h}{2\sigma^2}}dt = 4\sigma^2\end{aligned}$$

□

Establishing lower bounds for the mean-field approximation of the N-player game will be more challenging as it will require different tools. To establish lower bounds, we will need to use the following anti-concentration result for the binomial distribution.

Lemma A.4 (Anti-concentration for binomial)**.** *Let $N \in \mathbb{N}_{>0}$ and $X \sim \mathrm{Binom}(N, p)$ be drawn from a binomial distribution for some $p \in [^1/_2, 1]$. Then,* $\mathbb{P}\left[X \geq \frac{N}{2} + \frac{\sqrt{N}}{2}\right] \geq \frac{1}{20}$.

Proof. For $k_0 := \left\lceil \frac{N}{2} + \frac{\sqrt{N}}{2} \right\rceil$, we will lower bound $\sum_{k=k_0}^N \binom{N}{k} p^k (1-p)^{N-k}$ when N is large enough. If $k_0 < \lceil Np \rceil$, then the probability in the statement above is bounded below trivially by $^1/_2$ since $\lfloor Np \rfloor$ lower bounds the median of the binomial (Kaas & Buhrman, 1980). Otherwise, if $k_0 \geq \lceil Np \rceil$, then the function $\bar{p} \to \bar{p}^k(1-\bar{p})^{N-k}$ is increasing in $\bar{p}$ in the interval $[0, p]$. As $^1/_2 \in [0, p]$, it is then sufficient to assume $p = {}^1/_2$, and to upper bound $\mathbb{P}\left[\frac{N}{2} - \frac{\sqrt{N}}{2} < X < \frac{N}{2} + \frac{\sqrt{N}}{2}\right]$ by $^9/_{10}$ as the binomial probability mass is symmetric around $\frac{N}{2}$ when $p = {}^1/_2$.

First assuming N is even, we obtain by monotonicity $\binom{N}{k} \leq \binom{N}{N/2}$. Using the Stirling bound $\sqrt{2\pi}k^{k+\frac{1}{2}}\mathrm{e}^{-k} \leq k! \leq \mathrm{e}k^{k+\frac{1}{2}}\mathrm{e}^{-k}$, we further upper bound $\binom{N}{N/2} \leq \frac{e}{\pi}\frac{2^N}{\sqrt{N}}$, resulting in the bound $\mathbb{P}\left[\frac{N}{2} - \frac{\sqrt{N}}{2} < X < \frac{N}{2} + \frac{\sqrt{N}}{2}\right] \leq 2^{-N}\sqrt{N}\binom{N}{N/2} \leq \frac{e}{\pi} \leq {}^9/_{10}$, since there are at most $\sqrt{N}$ binomial coefficients being summed. Finally, assume $N = 2m+1$ is odd, then by the binomial formula $\binom{2m+1}{m+1} = \binom{2m}{m+1} + \binom{2m}{m} \leq 2\binom{2m}{m} \leq \frac{2e}{\pi}\frac{2^{2m}}{\sqrt{2m}}$. Hence we have the bound on the sum

$\mathbb{P}\left[\frac{N}{2} - \frac{\sqrt{N}}{2} < X < \frac{N}{2} + \frac{\sqrt{N}}{2}\right] \leq \frac{e\sqrt{N}}{\pi}\frac{1}{\sqrt{N-1}}$. It is easy to verify that for $N \geq 16$, $\frac{e\sqrt{N}}{\pi\sqrt{N-1}} \leq 9/10$, and the case when $N < 16$ and N is odd follows by manual computation. $\square$

Finally, we prove slightly more general upper bounds than presented in the main text that approximates the exploitability of an *approximate* MFG-NE in a finite population setting. Hence we define the following notions approximate FH-MFG and Stat-MFG.

Definition A.5 (δ-FH-MFG-NE)**.** Let $(\mathcal{S}, \mathcal{A}, H, P, R, \mu_0)$ be a FH-MFG. Then, a δ-FH-MFG Nash equilibrium is defined as:

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

Definition A.6 (δ-Stat-MFG-NE)**.** Let $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$ be a Stat-MFG. A policy-population pair $(\mu^*_\delta, \pi^*_\delta) \in \Delta_\mathcal{S} \times \Pi$ is called a δ-Stat-MFG Nash equilibrium if the two conditions hold:

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

A.2. Upper Bound for FH-MFG: Extended Proof of Theorem 3.2

Throughout this section we work with fixed $P \in \mathcal{P}_{K_\mu}$ and $R \in \mathcal{R}_{L_\mu}$. For any $\mathcal{X}$ valued random variable x denote $\mathcal{L}(x)(\cdot) \in \Delta_\mathcal{X}$ as the distribution of x. We start by introducing some notation.

For given R and P define the following constants:

$$\begin{aligned} L_s &:= \sup_{s,s',a,\mu} \left| R(s,a,\mu) - R(s',a,\mu) \right|, \\ L_a &:= \sup_{s,a,a',\mu} \left| R(s,a,\mu) - R(s,a',\mu) \right|, \\ K_s &:= \sup_{s,s',a,\mu} \left\| P(\cdot|s,a,\mu) - P(\cdot|s',a,\mu) \right\|, \\ K_a &:= \sup_{s,a,a',\mu} \left\| P(\cdot|s,a,\mu) - P(\cdot|s,a',\mu) \right\|. \end{aligned}$$

R and P are bounded due to Definition 2.1, thus all constants K_a, K_s, L_a, L_s are finite and well-defined, and it always holds that $K_s, K_a \leq 2$ and $L_s, L_a \leq 1$. With the above definition of constants, the more general Lipschitz condition holds: $\forall\, s, s' \in \mathcal{S}, a, a' \in \mathcal{A}, \mu, \mu' \in \Delta_\mathcal{S}$

$$\begin{aligned} \|P(\cdot|s,a,\mu) - P(\cdot|s',a',\mu')\|_1 \leq& K_\mu \|\mu - \mu'\|_1 + K_s d(s,s') \\ &+ K_a d(a,a'), \\ |R(s,a,\mu) - R(s',a',\mu')| \leq& L_\mu \|\mu - \mu'\|_1 + L_s d(s,s') \\ &+ L_a d(a,a'). \end{aligned}$$

We also introduce the shorthand notation for any $s \in \mathcal{S}, u \in \Delta_\mathcal{A}, \mu \in \Delta_\mathcal{S}$:

$$\begin{aligned} \overline{P}(\cdot|s,u,\mu) &:= \sum_{a\in\mathcal{A}} u(a) P(\cdot|s,a,\mu), \\ \overline{R}(s,u,\mu) &:= \sum_{a\in\mathcal{A}} u(a) R(s,a,\mu). \end{aligned}$$

By (**?**)Lemma C.1]yardim2023policy, it holds that

$$\begin{aligned} \|\overline{P}(\cdot|s,u,\mu) - \overline{P}(\cdot|s',u',\mu')\|_1 \leq& K_\mu \|\mu - \mu'\|_1 + K_s d(s,s') \\ &+ \frac{K_a}{2}\|u - u'\|_1, \\ |\overline{R}(s,u,\mu) - \overline{R}(s',u',\mu')| \leq& L_\mu \|\mu - \mu'\|_1 + L_s d(s,s') \\ &+ \frac{L_a}{2}\|u - u'\|_1. \end{aligned} \tag{1}$$

We will define a new operator for tracking the evolution of the population distribution over finite time horizons for a time-varying policy $\forall \boldsymbol{\pi} = \{\pi_h\}_{h=0}^{H-1} \in \Pi_H$:

$$\begin{aligned}\boldsymbol{\Gamma}_P^h(\mu, \boldsymbol{\pi}) &:= \underbrace{\Gamma_P(\dots \Gamma_P(\Gamma_P(\mu, \pi_0), \pi_1) \dots, \pi_{h-1})}_{h \text{ times}} \\ &= \mu_h^{\boldsymbol{\pi}} = \Lambda_P^H(\mu_0, \boldsymbol{\pi})_h,\end{aligned}$$

so $\boldsymbol{\Gamma}_P^0(\mu, \boldsymbol{\pi}) = \mu_0$. By repeated applications of Lemma 2.2, we obtain the Lipschitz condition:

$$\begin{aligned}&\|\boldsymbol{\Gamma}_P^n(\mu, \{\pi_i\}_{i=0}^{n-1}) - \boldsymbol{\Gamma}_P^n(\mu', \{\pi'_i\}_{i=0}^{n-1})\|_1 \\ &\quad\le L_{pop,\mu} \|\boldsymbol{\Gamma}_P^{n-1}(\mu, \{\pi_i\}_{i=0}^{n-2}) - \boldsymbol{\Gamma}_P^{n-1}(\mu', \{\pi'_i\}_{i=0}^{n-2})\|_1 \\ &\quad+ \frac{K_a}{2} \|\pi_{n-1} - \pi'_{n-1}\|_1 \\ &\quad\le L_{pop,\mu}^n \|\mu - \mu'\|_1 + \frac{K_a}{2} \sum_{i=0}^{n-1} L_{pop,\mu}^{n-1-i} \|\pi_i - \pi'_i\|_1, \end{aligned} \tag{2}$$

where $L_{pop,\mu} = (K_\mu + \frac{K_s}{2} + \frac{K_a}{2})$.

The proof will proceed in three steps:

- **Step 1.** Bounding the expected deviation of the empirical population distribution from the mean-field distribution $\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h^{\boldsymbol{\pi}}\|_1\right]$ for any given policy $\boldsymbol{\pi}$.
- **Step 2.** Bounding difference of N agent value function $J_{P,R}^{H,N,(i)}$ and the infinite player value function $V_{P,R}^H$.
- **Step 3.** Bounding the exploitability of an agent when each of N agents are playing the FH-MFG-NE policy.

Step 1: Empirical distribution bound. Due to its relevance for a general connection between the FH-MFG and the N-player game, we state this result in the form of an explicit bound.

Lemma A.7. *Suppose for the N-FH-MFG $(N, \mathcal{S}, \mathcal{A}, N, P, R, \gamma)$, agents $i = 1, \dots, N$ follow policies $\boldsymbol{\pi}^i = \{\pi_h^i\}_h$. Let $\overline{\boldsymbol{\pi}} = \{\overline{\pi}_h\}_h \in \Pi^H$ be arbitrary and $\boldsymbol{\mu}^{\overline{\boldsymbol{\pi}}} := \{\mu_h^{\overline{\boldsymbol{\pi}}}\}_{h=0}^{H-1} = \Lambda_P^H(\mu_0, \overline{\boldsymbol{\pi}})$. Then for all $h \in \{0, \dots, H-1\}$, it holds that:*

$$\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h^{\overline{\boldsymbol{\pi}}}\|_1\right] \le \frac{1 - L_{pop,\mu}^{h+1}}{1 - L_{pop,\mu}} |\mathcal{S}| \sqrt{\frac{\pi}{2N}} + \frac{K_a}{2N} \sum_{i=0}^{h-1} L_{pop,\mu}^{h-i-1} \Delta_{\pi_i},$$

where $\Delta_h := \frac{1}{N} \sum_i \|\overline{\pi}_h - \pi_h^i\|_1$

Proof. The proof will proceed inductively over h. First, for time $h = 0$, we have

$$\mathbb{E}\left[\|\widehat{\mu}_0 - \mu_0\|_1\right] = \sum_{s \in \mathcal{S}} \mathbb{E}\left[\left|\frac{1}{N} \sum_{i=1}^N (\mathbb{1}_{\{s_0^i = s\}} - \mu_0(s))\right|\right] \le |\mathcal{S}| \sqrt{\frac{\pi}{2N}},$$

where the last line is due to Lemma A.3 and the fact that $\mathbb{1}_{\{s_0^i = s\}}$ are bounded (hence subgaussian) random variables, and that in the finite state space we have $\mathbb{E}\left[\mathbb{1}_{\{s_0^i = s\}}\right] = \mu_0(s)$.

Next, denoting the σ-algebra induced by the random variables $(\{s_h^i\})_{i, h' \le h}$ as $\mathcal{F}_h$, we have that:

$$\begin{aligned}&\mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mu_{h+1}^{\overline{\boldsymbol{\pi}}}\|_1 \,|\mathcal{F}_h\right] \\ &\le \underbrace{\mathbb{E}\left[\|\mathbb{E}\left[\widehat{\mu}_{h+1} \,|\mathcal{F}_h\right] - \Gamma_P(\widehat{\mu}_h, \overline{\pi}_h)\|_1 \,|\mathcal{F}_h\right]}_{(\square)} \\ &\quad + \underbrace{\mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mathbb{E}\left[\widehat{\mu}_{h+1} \,|\mathcal{F}_h\right]\|_1 \,|\mathcal{F}_h\right]}_{(\triangle)} + \underbrace{\mathbb{E}\left[\|\Gamma_P(\widehat{\mu}_h, \overline{\pi}_h) - \mu_{h+1}^{\overline{\boldsymbol{\pi}}}\|_1 \,|\mathcal{F}_h\right]}_{(\heartsuit)}\end{aligned} \tag{3}$$

We upper bound the three terms separately. For $(\triangle)$, it holds that

$$
\begin{aligned}
(\triangle) =& \mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right]\|_1\,|\mathcal{F}_h\right] \\
=& \sum_{s\in\mathcal{S}} \mathbb{E}\left[|\widehat{\mu}_{h+1}(s) - \mathbb{E}\left[\widehat{\mu}_{h+1}(s)\,|\mathcal{F}_h\right]|\,|\mathcal{F}_h\right] \leq |\mathcal{S}|\sqrt{\frac{\pi}{2N}},
\end{aligned}
$$

since each $\widehat{\mu}_{h+1}(s)$ is an average of independent subgaussian random variables given $\mathcal{F}_h$. Specifically, each indicator is bounded $\mathbb{1}_{\{s^i_{h+1}=s\}} \in [0,1]$ a.s. and therefore is sub-Gaussian with $\mathbb{1}_{\{s^i_{h+1}=s\}} \in SG(1/4)$. Thus we get $\widehat{\mu}_{h+1}(s) \in SG(1/(4N))$ and apply bound on expected value discussed in Appendix A.1.

Next, for $(\square) = \|\mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right] - \Gamma_P(\widehat{\mu}_h, \overline{\pi}_h)\|_1$, we note that

$$
\mathbb{E}\left[\widehat{\mu}_{h+1}(s)\,|\mathcal{F}_h\right] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\{s^i_{h+1}=s\}}\,|\mathcal{F}_h\right] = \frac{1}{N}\sum_{i=1}^N \overline{P}(s|s^i_h, \pi^i_h(s^i_h), \widehat{\mu}_h),
$$

therefore

$$
\begin{aligned}
(\square) &= \left\| \frac{1}{N}\sum_{i=1}^N \overline{P}(\cdot|s^i_h, \pi^i_h(\cdot|s^i_h), \widehat{\mu}_h) - \sum_{s'} \widehat{\mu}_h(s')\overline{P}(\cdot|s', \pi_h(\cdot|s'), \widehat{\mu}_h) \right\|_1 \\
&= \left\| \frac{1}{N}\sum_{i=1}^N \left(\overline{P}(\cdot|s^i_h, \pi^i_h(\cdot|s^i_h), \widehat{\mu}_h) - \overline{P}(\cdot|s^i_h, \pi_h(\cdot|s^i_h), \widehat{\mu}_h)\right) \right\|_1 \\
&\leq \frac{1}{N}\sum_{i=1}^N \|\overline{P}(\cdot|s^i_h, \pi^i_h(\cdot|s^i_h), \widehat{\mu}_h) - \overline{P}(\cdot|s^i_h, \pi_h(\cdot|s^i_h), \widehat{\mu}_h)\|_1 \\
&\overset{(I)}{\leq} \frac{K_a}{2N}\sum_{i=1}^N \|\pi^i_h(\cdot|s^i_h) - \pi_h(\cdot|s^i_h)\|_1 \leq \frac{K_a}{2}\Delta_h,
\end{aligned}
$$

where (I) follows from the Lipschitz property (1). Finally, the last term $(\heartsuit)$ can be bounded using:

$$
(\heartsuit) = \mathbb{E}\left[\|\Gamma_P(\widehat{\mu}_h, \overline{\pi}_h) - \Gamma_P(\mu^{\overline{\boldsymbol{\pi}}}_h, \overline{\pi}_h)\|_1\,|\mathcal{F}_h\right] \leq L_{pop,\mu}\|\widehat{\mu}_h - \mu^{\overline{\boldsymbol{\pi}}}_h\|_1.
$$

To conclude, merging the bounds on the three terms in Inequality (3) and taking the expectations we obtain:

$$
\mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mu^{\overline{\boldsymbol{\pi}}}_{h+1}\|_1\right] \leq L_{pop,\mu}\mathbb{E}\left[\|\widehat{\mu}_h - \mu^{\overline{\boldsymbol{\pi}}}_h\|_1\right] + |\mathcal{S}|\sqrt{\frac{\pi}{2N}} + \frac{K_a\Delta_h}{2}.
$$

Induction on h yields the statement of the lemma.

□

Step 2: Bounding difference of N agent value function. Next, we bound the difference between the N-player expected reward function $J^{H,N,(1)}_{P,R}$ and the infinite player expected reward function $V^H_{P,R}$. For ease of reading, expectations, probabilities, and laws of random variables will be denoted $\mathbb{E}_\infty, \mathbb{P}_\infty, \mathcal{L}_\infty$ respectively over the infinite player finite horizon game and $\mathbb{E}_N, \mathbb{P}_N, \mathcal{L}_N$ respectively over the N-player game. We use the regular notation $\mathbb{E}[\cdot], \mathbb{P}[\cdot], \mathcal{L}(\cdot)$ without subscripts if the underlying randomness is clearly defined. We state the main result of this step in the following lemma.

Lemma A.8. *Suppose N-FH-MFG agents follow the same sequence of policies $\boldsymbol{\pi} = \{\pi_h\}_{h=0}^{H-1}$. Then*

$$
\left|J^{H,N,(1)}_{P,R}(\boldsymbol{\pi}, \ldots, \boldsymbol{\pi}) - V^H_{P,R}(\Lambda^H_P(\mu_0, \boldsymbol{\pi}), \boldsymbol{\pi})\right| \leq (L_\mu + \frac{L_s}{2})|\mathcal{S}|\sqrt{\frac{\pi}{2N}}\sum_{h=0}^{H-1}\frac{1 - L^{h+1}_{pop,\mu}}{1 - L_{pop,\mu}}.
$$

Proof. Due to symmetry in the N agent game, any permutation $\sigma : [N] \to [N]$ of agents does not change their distribution, that is $\mathcal{L}_N(s_h^1, \dots, s_h^N) = \mathcal{L}_N(s_h^{\sigma(1)}, \dots, s_h^{\sigma(N)})$. We can then conclude that:

$$\begin{aligned}\mathbb{E}_N\left[R(s_h^1, a_h^1, \widehat{\mu}_h)\right] &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}_N\left[R(s_h^i, a_h^i, \widehat{\mu}_h)\right] \\ &= \mathbb{E}_N\left[\sum_{s\in\mathcal{S}} \widehat{\mu}_h(s)\overline{R}(s, \pi_h(s), \widehat{\mu}_h).\right]\end{aligned}$$

Therefore, we by definition:

$$J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}, \dots, \boldsymbol{\pi}) = \mathbb{E}_N\left[\sum_{h=0}^{H-1}\sum_{s\in\mathcal{S}} \widehat{\mu}_h(s)\overline{R}(s, \pi_h(s), \widehat{\mu}_h)\right].$$

Next, in the FH-MFG, under the population distribution $\{\mu_h\}_{h=0}^{H-1} = \Lambda_P^H(\mu_0, \boldsymbol{\pi})$ we have that for all $h \in 0, \dots, H-1$,

$$\begin{aligned}\mathbb{P}_\infty(s_0 = \cdot) &= \mu_0, \\ \mathbb{P}_\infty(s_{h+1} = \cdot) &= \sum_{s\in\mathcal{S}} \mathbb{P}_\infty(s_h = s)\,\mathbb{P}_\infty(s_h = \cdot | s_h = s) \\ &= \Gamma_P(\mathbb{P}_\infty(s_h = \cdot), \pi_h),\end{aligned}$$

so by induction $\mathbb{P}_\infty(s_h = \cdot) = \mu_h$. Then we can conclude that

$$\begin{aligned}V_{P,R}^H(\Lambda_P^H(\mu_0, \boldsymbol{\pi}), \boldsymbol{\pi}) &= \mathbb{E}_\infty\left[\sum_{h=0}^{H-1} R(s_h, \pi_h(s_h), \mu_h)\right] \\ &= \sum_{h=0}^{H-1}\sum_{s\in\mathcal{S}} \mu_h(s) R(s, \pi_h(s), \mu_h).\end{aligned}$$

Merging the two equalities for J, V, we have the bound:

$$\begin{aligned}&|J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}, \dots, \boldsymbol{\pi}) - V_{P,R}^H(\Lambda_P^H(\mu_0, \boldsymbol{\pi}), \boldsymbol{\pi})| \\ =&\left|\mathbb{E}_N\left[\sum_{h=0}^{H-1}\sum_{s\in\mathcal{S}} \widehat{\mu}_h(s)\overline{R}(s, \pi_h(s), \widehat{\mu}_h)\right] - \sum_{h=0}^{H-1}\sum_{s\in\mathcal{S}} \mu_h(s)R(s, \pi_h(s), \mu_h)\right| \\ \leq&\mathbb{E}_N\left[\sum_{h=0}^{H-1}\left|\sum_{s\in\mathcal{S}}\left(\widehat{\mu}_h(s)\overline{R}(s, \pi_h(s), \widehat{\mu}_h) - \mu_h(s)R(s, \pi_h(s), \mu_h)\right)\right|\right] \\ \leq&\mathbb{E}_N\left[\sum_{h=0}^{H-1}\left(\frac{L_s}{2}\|\mu_h - \widehat{\mu}_h\|_1 + L_\mu\|\mu_h - \widehat{\mu}_h\|_1\right)\right].\end{aligned}$$

The statement of the lemma follows by an application of Lemma A.7. □

Step 3: Bounding difference in policy deviation. Finally, to conclude the proof of the main theorem of this section, we will prove that the improvement in expectation due to single-sided policy changes are at most of order $\mathcal{O}\left(\frac{1}{\sqrt{N}}\right)$.

Lemma A.9. *Suppose* $\boldsymbol{\pi} = \{\pi_h\}_{h=0}^{H-1} \in \Pi^H$ *and* $\boldsymbol{\pi}' = \{\pi'_h\}_{h=0}^{H-1} \in \Pi^H$ *arbitrary policies, and* $\boldsymbol{\mu}^{\boldsymbol{\pi}} := \Lambda_P^H(\mu_0, \boldsymbol{\pi})$ *is the population distribution induced by* $\boldsymbol{\pi}$. *Then*

$$\begin{aligned}&\left|J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}', \boldsymbol{\pi}, \dots, \boldsymbol{\pi}) - V_{P,R}^H(\Lambda_P^H(\mu_0, \boldsymbol{\pi}), \boldsymbol{\pi}')\right| \\ &\qquad\leq \sum_{h=0}^{H-1}\left(\frac{L_\mu}{2}\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h^{\boldsymbol{\pi}}\|_1\right] + K_\mu \sum_{h'=0}^{h-1}\mathbb{E}\left[\|\widehat{\mu}_{h'} - \mu_{h'}^{\boldsymbol{\pi}}\|_1\right]\right).\end{aligned}$$

Proof. Define the random variables $\{s_h^i, a_h^i\}_{i,h}, \{\widehat{\mu}_h\}_h$ as in the definition of N-FH-SAG (Definition 3.1). In addition, define the random variables $\{s_h, a_h\}_h$ evolving according to the FH-MFG with population $\boldsymbol{\mu}^{\boldsymbol{\pi}} := \{\mu_h^{\boldsymbol{\pi}}\}_h := \Lambda_P^H(\mu_0, \boldsymbol{\pi})$ and representative policy $\boldsymbol{\pi}'$, independent from the random variables $\{s_h^i, a_h^i\}_{i,h}$. Hence $s_0 \sim \mu_0, a_h \sim \pi'(\cdot|s_h), s_{h+1} \sim P(\cdot|s_h, a_h, \mu_h^{\boldsymbol{\pi}})$. Define also for simplicity

$$E_N := \left| J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}', \boldsymbol{\pi}, \dots, \boldsymbol{\pi}) - V_{P,R}^H(\Lambda_P^H(\mu_0 \boldsymbol{\pi}), \boldsymbol{\pi}') \right|.$$

With these definitions, we have

$$\begin{aligned} E_N &= \left| \mathbb{E}\left[\sum_{h=0}^{H-1} R(s_h, a_h, \mu_h^{\boldsymbol{\pi}}) - \sum_{h=0}^{H-1} R(s_h^1, a_h^1, \widehat{\mu}_h) \right] \right| \\ &\leq \sum_{h=0}^{H-1} \left| \mathbb{E}\left[R(s_h, a_h, \mu_h^{\boldsymbol{\pi}}) - R(s_h^1, a_h^1, \widehat{\mu}_h) \right] \right|. \end{aligned} \tag{4}$$

Furthermore, for any $h \in \{0, \dots, H-1\}$,

$$\begin{aligned} &|\, \mathbb{E}\left[R(s_h, a_h, \mu_h^{\boldsymbol{\pi}}) - R(s_h^1, a_h^1, \widehat{\mu}_h) \right] | \\ &\leq \left| \mathbb{E}\left[R(s_h, a_h, \mu_h^{\boldsymbol{\pi}}) - R(s_h^1, a_h^1, \mu_h^{\boldsymbol{\pi}}) \right] \right| \\ &\quad + \left| \mathbb{E}\left[R(s_h^1, a_h^1, \mu_h^{\boldsymbol{\pi}}) - R(s_h^1, a_h^1, \widehat{\mu}_h) \right] \right| \\ &\leq \left| \mathbb{E}\left[R(s_h, \pi_h'(s_h), \mu_h^{\boldsymbol{\pi}}) - R(s_h^1, \pi_h'(s_h^1), \mu_h^{\boldsymbol{\pi}}) \right] \right| \\ &\quad + L_\mu\, \mathbb{E}\left[\|\mu_h^{\boldsymbol{\pi}} - \widehat{\mu}_h\|_1 \right] \\ &\leq \frac{1}{2} \| \mathbb{P}[s_h = \cdot] - \mathbb{P}[s_h^1 = \cdot] \|_1 + L_\mu\, \mathbb{E}\left[\|\mu_h^{\boldsymbol{\pi}} - \widehat{\mu}_h\|_1 \right], \end{aligned}$$

where the last line follows since R is bounded in $[0, 1]$. Replacing this in Equation (4),

$$E_N \leq \frac{1}{2} \sum_h \| \mathbb{P}[s_h = \cdot] - \mathbb{P}[s_h^1 = \cdot] \|_1 + L_\mu \sum_h \mathbb{E}\left[\|\mu_h^{\boldsymbol{\pi}} - \widehat{\mu}_h\|_1 \right]. \tag{5}$$

The first sum above we upper bound in the rest of the proof inductively.

Firstly, by definitions of N-FH-SAG and FH-MFG, both s_0^1 and s_0 have distribution μ_0, hence $\| \mathbb{P}[s_0 = \cdot] - \mathbb{P}[s_0^1 = \cdot] \|_1 = 0$. Assume that $h \geq 1$. We note that P takes values in $\Delta_{\mathcal{S}}$ and the random vector $\widehat{\mu}_h$ takes values in the discrete set $\{\frac{1}{N} u : u \in \{0, \dots, N\}^{\mathcal{S}}, \sum_s u(s) = N\} \subset \Delta_{\mathcal{S}}$, hence we have the bounds:

$$\begin{aligned} &\| \mathbb{P}[s_{h+1} = \cdot] - \mathbb{P}[s_{h+1}^1 = \cdot] \|_1 \\ &\leq \left\| \sum_{s,\mu} P(s, \pi_h'(s), \mu)\, \mathbb{P}[s_h^1 = s, \widehat{\mu}_h = \mu] - \sum_s P(s, \pi_h'(s), \mu_h^{\boldsymbol{\pi}})\, \mathbb{P}[s_h = s] \right\|_1 \\ &\leq \left\| \sum_s P(s, \pi_h'(s), \mu_h^{\boldsymbol{\pi}})\, \mathbb{P}[s_h^1 = s] - \sum_s P(s, \pi_h'(s), \mu_h^{\boldsymbol{\pi}})\, \mathbb{P}[s_h = s] \right\|_1 \\ &\quad + \left\| \sum_{s,\mu} (P(s, \pi_h'(s), \mu) - P(s, \pi_h'(s), \mu_h^{\boldsymbol{\pi}}))\, \mathbb{P}[s_h^1 = s, \widehat{\mu}_h = \mu] \right\|_1 \\ &\leq \left\| \mathbb{P}[s_h^1 = \cdot] - \mathbb{P}[s_h = \cdot] \right\|_1 + \sum_{s,\mu} K_\mu \|\mu - \mu_h^{\boldsymbol{\pi}}\|_1\, \mathbb{P}[s_h^1 = s, \widehat{\mu}_h = \mu] \\ &\leq \left\| \mathbb{P}[s_h^1 = \cdot] - \mathbb{P}[s_h = \cdot] \right\|_1 + K_\mu\, \mathbb{E}\left[\|\widehat{\mu}_h^{\boldsymbol{\pi}} - \mu_h^{\boldsymbol{\pi}}\|_1 \right] \end{aligned}$$

where the last two lines follow from the fact that P is K_μ Lipschitz in μ and stochastic matrices are non-expansive in the total-variation norm over probability distributions. By induction, we conclude that for all $h \geq 0$, it holds that:

$$\| \mathbb{P}[s_h = \cdot] - \mathbb{P}[s_h^1 = \cdot] \|_1 \leq K_\mu \sum_{h'=0}^{h} \mathbb{E}\left[\|\widehat{\mu}_{h'}^{\boldsymbol{\pi}} - \mu_{h'}^{\boldsymbol{\pi}}\|_1 \right].$$

Placing this result into Equation (5), we obtain the statement of the lemma.

□

Since $\mathbb{E}\left[\|\widehat{\mu}_{h'} - \mu^{\boldsymbol{\pi}}_{h'}\|_1\right]$ above in the theorem is of the order of $\mathcal{O}\left({}^1\!/\!\sqrt{N}\right)$ by the result in step 1, the result above allows us to bound exploitability in the N-FH-SAG.

Conclusion and Statement of Result. Finally, we can merge the results up until this stage to upper bound the exploitability. By definition of the FH-MFG-NE, we have:

$$\delta \geq \max_{\boldsymbol{\pi}' \in \Pi^H} V^H_{P,R}(\Lambda^H_P(\mu_0, \boldsymbol{\pi}_\delta), \boldsymbol{\pi}') - V^H_{P,R}(\Lambda^H_P(\mu_0, \boldsymbol{\pi}_\delta), \boldsymbol{\pi}_\delta)$$

The upper bounds on the deviation between $V^H_{P,R}$ and $J^{H,N,(1)}_{P,R}$ from the previous steps directly yields the statement of the theorem. We state it below for completeness.

Theorem A.10. *It holds that*

$$\mathcal{E}^{H,N,(1)}_{P,R}(\boldsymbol{\pi}_\delta, \dots, \boldsymbol{\pi}_\delta) \leq 2\delta + \frac{C_1}{\sqrt{N}} + \frac{C_2}{N} = O\left(\delta + \frac{1}{\sqrt{N}}\right)$$

where $\boldsymbol{\pi}_\delta$ is a δ-FH-MFG Nash equilibrium and

$$C_1 = |\mathcal{S}|\sqrt{\frac{\pi}{2}}\left((2L_\mu + \frac{L_s}{2})\sum_{h=0}^{H-1}\frac{1 - L^{h+1}_{pop,\mu}}{1 - L_{pop,\mu}} + K_\mu \sum_{h=0}^{H-1}\sum_{i=0}^{h-1}\frac{1 - L^{i+1}_{pop,\mu}}{1 - L_{pop,\mu}}\right)$$

$$C_2 = L_\mu K_a \sum_{h=0}^{H-1}\frac{1 - L^{h}_{pop,\mu}}{1 - L_{pop,\mu}} + K_a K_\mu \sum_{h=0}^{H-1}\sum_{i=0}^{h-1}\frac{1 - L^{i}_{pop,\mu}}{1 - L_{pop,\mu}},$$

where we use shorthand notation $\frac{1-L^k_{pop,\mu}}{1-L_{pop,\mu}} := k - 1$ *when* $L_{pop,\mu} = 1$.

A note on constants. Note that constants C_1, C_2 in Theorem A.10 depend on horizon with $\frac{H^2}{1-L_{pop,\mu}}$ if $L_{pop,\mu} < 1$, with H^3 if $L_{pop,\mu} = 1$ and with $H^2\frac{1-L^{H+1}_{pop,\mu}}{1-L_{pop,\mu}}$ if $L_{pop,\mu} > 1$.

A.3. Lower Bound for FH-MFG: Extended Proof of Theorem 3.3

The proof will be by construction: we will explicitly define an FH-MFG where the optimal policy for the N-agent game diverges quickly from the FH-MFG-NE policy.

Preliminaries. We first define a few utility functions. Define $\mathbf{g} : \Delta_2 \to B^2_{\infty,+} := \{\mathbf{x} \in \mathbb{R}^2 : \|\mathbf{x}\|_\infty = 1,\ x_1, x_2 \geq 0\}$ and $\mathbf{h} : \Delta_2 \to [0,1]^2$ as follows:

$$\mathbf{g}(x_1, x_2) := \begin{pmatrix}\mathbf{g}_1(x_1,x_2)\\ \mathbf{g}_2(x_1,x_2)\end{pmatrix} := \begin{pmatrix}\frac{x_1}{\max\{x_1,x_2\}}\\ \frac{x_2}{\max\{x_1,x_2\}}\end{pmatrix},$$

$$\mathbf{h}(x_1, x_2) := \begin{pmatrix}\mathbf{h}_1(x_1,x_2)\\ \mathbf{h}_2(x_1,x_2)\end{pmatrix} := \begin{pmatrix}\max\{4x_2, 1\}\\ \max\{4x_1, 1\}\end{pmatrix}.$$

Furthermore, for any $\epsilon > 0$ we define $\omega_\epsilon : [0,1] \to [0,1]$ as:

$$\omega_\epsilon(x) = \begin{cases} 1, & x > {}^1\!/\!{}_2 + \epsilon \\ 0, & x < {}^1\!/\!{}_2 - \epsilon \\ \frac{1}{2} + \frac{x - {}^1\!/\!{}_2}{2\epsilon}, & x \in [{}^1\!/\!{}_2 - \epsilon, {}^1\!/\!{}_2 + \epsilon] \end{cases}.$$

$\epsilon \in (0, {}^1\!/\!{}_2)$ will be specified later.

It is straightforward to verify that $\mathbf{g}$ has an inverse in its domain given by

$$\mathbf{g}^{-1}(x_1, x_2) = \left(\frac{x_1}{x_1 + x_2}, \frac{x_2}{x_1 + x_2}\right), \forall (x_1, x_2) \in B^2_{\infty,+}.$$

Furthermore, it holds for $\mathbf{x} = (x_1, x_2) \in B^2_{\infty,+}, \mathbf{y} = (y_1, y_2) \in B^2_{\infty,+}$

$$\begin{aligned}
&\|\mathbf{g}^{-1}(\mathbf{x}) - \mathbf{g}^{-1}(\mathbf{y})\|_1 \\
&= \left|\frac{x_1}{x_1 + x_2} - \frac{y_1}{y_1 + y_2}\right| + \left|\frac{x_2}{x_1 + x_2} - \frac{y_2}{y_1 + y_2}\right| \\
&= \left|\frac{x_1(y_2 - x_2) + x_2(x_1 - y_1)}{(x_1 + x_2)(y_1 + y_2)}\right| + \left|\frac{x_2(y_1 - x_1) + x_1(x_2 - y_2)}{(x_1 + x_2)(y_1 + y_2)}\right| \\
&\leq 2\|\mathbf{x} - \mathbf{y}\|_1,
\end{aligned}$$

and likewise for $\mathbf{u}, \mathbf{v} \in \Delta_2$, letting $u_+ := \max\{u_1, u_2\}, v_+ := \max\{v_1, v_2\}$,

$$\begin{aligned}
\|\mathbf{g}(\mathbf{u}) - \mathbf{g}(\mathbf{v})\|_1 &= \left|\frac{u_1}{u_+} - \frac{v_1}{v_+}\right| + \left|\frac{u_2}{u_+} - \frac{v_2}{v_+}\right| \\
&= \left|\frac{u_1 v_+ - v_1 u_+}{u_+ v_+}\right| + \left|\frac{u_2 v_+ - u_+ v_2}{u_+ v_+}\right| \leq 2\|\mathbf{u} - \mathbf{v}\|_1.
\end{aligned}$$

This follows from considering cases and observation that $u_+ \geq 1/2$, $v_+ \geq 1/2$. Then for all $\mathbf{u}, \mathbf{v} \in \Delta_2$, $\mathbf{g}, \mathbf{h}$ have the bi-Lipschitz and Lipschitz properties:

$$\frac{1}{2}\|\mathbf{u} - \mathbf{v}\|_1 \leq \|\mathbf{g}(\mathbf{u}) - \mathbf{g}(\mathbf{v})\|_1 \leq 2\|\mathbf{u} - \mathbf{v}\|_1, \tag{6}$$

$$\|\mathbf{h}(\mathbf{u}) - \mathbf{h}(\mathbf{v})\|_1 \leq 4\|\mathbf{u} - \mathbf{v}\|_1. \tag{7}$$

Likewise, ω_ϵ, being piecewise linear, also satisfies the Lipschitz condition: $|\omega_\epsilon(x) - \omega_\epsilon(y)| \leq \frac{1}{2\epsilon}|x - y|, \quad \forall x, y \in [0, 1]$.

Defining the FH-MFG. We take a particular FH-MFG with 6 states, 2 actions. Define the state-actions sets:

$$\mathcal{S} = \{s_{\text{Left}}, s_{\text{Right}}, s_{\text{LA}}, s_{\text{LB}}, s_{\text{RA}}, s_{\text{RB}}\}, \quad \mathcal{A} = \{a_{\text{A}}, a_{\text{B}}\}.$$

Intuitively, the "main" states of the game are $s_{\text{Left}}, s_{\text{Right}}$ and the 4 states $s_{\text{LA}}, s_{\text{LB}}, s_{\text{RA}}, s_{\text{RB}}$ are dummy states that keep track of which actions were taken by which percentage of players used to introduce a dependency of the rewards on the distribution of agents over actions as well as states. Define the initial probabilities μ_0 by:

$$\begin{aligned}
\boldsymbol{\mu}_0(s_{\text{Left}}) = \boldsymbol{\mu}_0(s_{\text{Right}}) = 1/2, \\
\boldsymbol{\mu}_0(s_{\text{LA}}) = \boldsymbol{\mu}_0(s_{\text{RA}}) = \boldsymbol{\mu}_0(s_{\text{RA}}) = \boldsymbol{\mu}_0(s_{\text{RB}}) = 0.
\end{aligned}$$

When at the states $s_{\text{Left}}, s_{\text{Right}}$, the transition probabilities are defined for all $\mu \in \Delta_{\mathcal{S}}$ by:

$$\begin{aligned}
P(s_{\text{LA}}|s_{\text{Left}}, a_{\text{A}}, \mu) = 1, \quad P(s_{\text{LB}}|s_{\text{Left}}, a_{\text{B}}, \mu) = 1, \\
P(s_{\text{RA}}|s_{\text{Right}}, a_{\text{A}}, \mu) = 1, \quad P(s_{\text{RB}}|s_{\text{Right}}, a_{\text{B}}, \mu) = 1.
\end{aligned}$$

That is, the agent transitions to one of $\{s_{\text{LA}}, s_{\text{RA}}, s_{\text{RB}}, s_{\text{LB}}\}$ to remember its last action and left-right state. When at states $\{s_{\text{LA}}, s_{\text{RA}}, s_{\text{RB}}, s_{\text{LB}}\}$, the transition probabilities are:

$$\begin{aligned}
&\text{If } s \in \{s_{\text{LA}}, s_{\text{LB}}, s_{\text{RA}}, s_{\text{RB}}\}: \\
&P(s'|s, a, \mu) = \begin{cases} \omega_\epsilon(\mu(s_{\text{LA}}) + \mu(s_{\text{LB}})), \text{ if } s' = s_{\text{Left}} \\ \omega_\epsilon(\mu(s_{\text{RA}}) + \mu(s_{\text{RB}})), \text{ if } s' = s_{\text{Right}} \end{cases}, \forall \mu, a.
\end{aligned}$$

The other non-defined transition probabilities are of course 0.

Finally, let $\alpha, \beta > 0$ such that $\alpha + \beta < 1$ (to be also defined later). The reward functions are defined for all $\mu \in \Delta_{\mathcal{S}}$ as follows:

$$\begin{aligned}
R(s_{\text{Left}}, a_{\text{A}}, \mu) =& R(s_{\text{Left}}, a_{\text{B}}, \mu) = 0,\\
R(s_{\text{Right}}, a_{\text{A}}, \mu) =& R(s_{\text{Right}}, a_{\text{B}}, \mu) = 0,\\
\begin{pmatrix} R(s_{\text{LA}}, a_{\text{A}}, \mu) \\ R(s_{\text{LB}}, a_{\text{A}}, \mu) \end{pmatrix} =& (1-\alpha-\beta)\mathbf{g}\big(\mu(s_{\text{LA}}) + \mu(s_{\text{LB}}), \mu(s_{\text{RA}}) + \mu(s_{\text{RB}})\big)\\
&+ \alpha \mathbf{h}(\mu(s_{\text{LA}}), \mu(s_{\text{LB}}))\\
\begin{pmatrix} R(s_{\text{LA}}, a_{\text{B}}, \mu) \\ R(s_{\text{LB}}, a_{\text{B}}, \mu) \end{pmatrix} =& (1-\alpha-\beta)\mathbf{g}\big(\mu(s_{\text{LA}}) + \mu(s_{\text{LB}}), \mu(s_{\text{RA}}) + \mu(s_{\text{RB}})\big)\\
&+ \alpha \mathbf{h}(\mu(s_{\text{LA}}), \mu(s_{\text{LB}})) + \beta \mathbf{1}\\
\begin{pmatrix} R(s_{\text{RA}}, a_{\text{A}}, \mu) \\ R(s_{\text{RB}}, a_{\text{A}}, \mu) \end{pmatrix} =& (1-\alpha-\beta)\mathbf{g}\big(\mu(s_{\text{RA}}) + \mu(s_{\text{RB}}), \mu(s_{\text{LA}}) + \mu(s_{\text{LB}})\big)\\
&+ \alpha \mathbf{h}(\mu(s_{\text{RA}}), \mu(s_{\text{RB}}))\\
\begin{pmatrix} R(s_{\text{RA}}, a_{\text{B}}, \mu) \\ R(s_{\text{RB}}, a_{\text{B}}, \mu) \end{pmatrix} =& (1-\alpha-\beta)\mathbf{g}\big(\mu(s_{\text{RA}}) + \mu(s_{\text{RB}}), \mu(s_{\text{LA}}) + \mu(s_{\text{LB}})\big)\\
&+ \alpha \mathbf{h}(\mu(s_{\text{RA}}), \mu(s_{\text{RB}})) + \beta \mathbf{1}
\end{aligned}$$

Note that only at odd steps do the agents get a reward, and at this step, it does not matter which action the agent plays, only the state among $\{s_{\text{LA}}, s_{\text{LA}}, s_{\text{RA}}, s_{\text{RB}}\}$ and the population distribution. The parameters ϵ, α, β of the above FH-MFG are "free" parameters to be specified later. We visualize the FH-MFG in Figure 1.

A minor remark. The arguments of $\mathbf{g}$ above will be with probability one in the set Δ_2 at odd-numbered time steps, but to formally satisfy the Lipschitz condition $R \in \mathcal{R}_2$ one can for instance replace $\mathbf{g}\big(\mu(s_{\text{RA}}) + \mu(s_{\text{RB}}), \mu(s_{\text{LA}}) + \mu(s_{\text{LB}})\big)$ with $\mathbf{g}\big(\mu(s_{\text{RA}}) + \mu(s_{\text{RB}}) + \mu(s_{\text{Left}}), \mu(s_{\text{LA}}) + \mu(s_{\text{LB}}) + \mu(s_{\text{Right}})\big)$ in the definitions, which will not impact the analysis since at odd timesteps $\mu(s_{\text{Right}}) = \mu(s_{\text{Left}}) = 0$ for both the FH-MFG and N-FH-SAG.

Note that with these definitions, $P \in \mathcal{P}_{1/2\epsilon}, R \in \mathcal{R}_2$ since only $\forall\ s, s' \in \mathcal{S}, a, a' \in \mathcal{A}, \mu, \mu' \in \Delta_{\mathcal{S}}$, we have by the definitions:

$$\|P(\cdot|s,a,\mu) - P(\cdot|s',a',\mu')\|_1 \leq 2d(s,s') + 2d(a,a') + \frac{1}{2\epsilon}\|\mu - \mu'\|_1, \tag{8}$$

$$|R(s,a,\mu) - R(s',a',\mu')| \leq d(s,s') + d(a,a') + 2\|\mu - \mu'\|_1, \tag{9}$$

for any $\alpha, \beta > 0$ with $\alpha + \beta < 1$ and $\alpha < \frac{1}{4}$, using the Lipschitz conditions in (6), (7).

Step 1: Solution of the FH-MFG. Next, we solve the infinite player FH-MFG and show that the policy $\boldsymbol{\pi}^*_H := \{\pi^*_h\}_{h=0}^{H-1}$ given by:

$$\pi^*_h(a|s) := \begin{cases} 1, \text{if } h \text{ odd and } a = a_{\text{B}} \\ \frac{1}{2}, \text{if } h \text{ even} \\ 0, \text{if } h \text{ odd and } a = a_{\text{B}} \end{cases}$$

It is easy to verify in this case that, if $\boldsymbol{\mu}^* := \{\mu^*_h\}_h$ is induced by $\boldsymbol{\pi}^*$:

$$\begin{aligned}
\mu^*_h(s_{\text{LA}}) = \mu^*_h(s_{\text{LB}}) = \mu^*_h(s_{\text{RA}}) = \mu^*_h(s_{\text{RB}}) = {}^1\!/\!_4, \text{ if } h \text{ odd},\\
\mu^*_h(s_{\text{Left}}) = \mu^*_h(s_{\text{Right}}) = {}^1\!/\!_2, \text{ if } h \text{ even}.
\end{aligned}$$

In this case, the induced rewards in odd steps are state-independent (it is the same for all states $s_{\text{RA}}, s_{\text{RB}}, s_{\text{LA}}, s_{\text{LB}}$), therefore the policy $\boldsymbol{\pi}^*$ is the optimal best response to the population and a FH-MFG.

In fact, $\boldsymbol{\pi}^*$ is unique up to modifications in zero-probability sets (e.g., modifying $\pi^*_h(s_{\text{Left}})$ for odd h, for which $\mathbb{P}[s_h = s_{\text{Left}}] = 0$). To see this, for *any* policy $\boldsymbol{\pi} \in \Pi_H$, it holds that

$$\begin{aligned}
\mu^{\boldsymbol{\pi}}_h(s_{\text{Left}}) = \mu^{\boldsymbol{\pi}}_h(s_{\text{Right}}) = {}^1\!/\!_2, \text{ if } h \text{ even},\\
\mu^{\boldsymbol{\pi}}_h(s_{\text{LA}}) + \mu^{\boldsymbol{\pi}}_h(s_{\text{LB}}) = \mu^{\boldsymbol{\pi}}_h(s_{\text{RA}}) + \mu^{\boldsymbol{\pi}}_h(s_{\text{RB}}) = {}^1\!/\!_2, \text{ if } h \text{ odd},
\end{aligned}$$

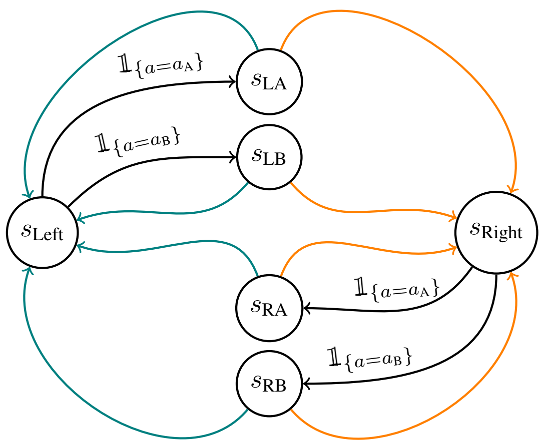

Figure 1. Visualization of the counterexample. All orange edges have probability $\omega_\varepsilon(\mu(s_{\text{RA}}) + \mu(s_{\text{RB}}))$, green edges have probability $\omega_\varepsilon(\mu(s_{\text{LA}}) + \mu(s_{\text{LB}}))$ independent of action taken. Edges with probability 0 are not drawn.

as the action of the agent does not affect transition probabilities between $s_{\text{Left}}, s_{\text{Right}}$ in even rounds. Moreover, as odd stages, the action rewards terms only depend on the state apart from the positive additional term $\beta\mathbf{1}$, so the only optimal action will be a_{B}. Finally, for $\alpha > 0$, the actions $a_{\text{A}}, a_{\text{B}}$ must be played with equal probability as otherwise the term $\alpha\mathbf{h}(\mu(s_{\text{RA}}), \mu(s_{\text{RB}}))$ will lead to the action with lower probability assigned by being optimal.

Step 2: Population divergence in N-FH-MFG. We will analyze the empirical population distribution deviation from $\boldsymbol{\mu}^*$, namely, we will lower bound $\mathbb{E}[\|\mu_h^* - \widehat{\mu}_h\|_1]$. The results in this step will be valid for *any* policy profile $(\boldsymbol{\pi}^1, \ldots, \boldsymbol{\pi}^N) \in \Pi$: we emphasize that at even h, $\widehat{\mu}_h$ is independent of agent policies in the N player game. In this step, we also fix $1/2\varepsilon = 8$.

We will analyze $\widehat{\mu}_h$ at all even steps $h = 2m$ where $m \in \mathbb{N}_{\geq 0}$. Define the sequence of random variables for all $m \in \mathbb{N}_{\geq 0}$ as $X_m := \widehat{\mu}_{2m}(s_{\text{Left}})$. Define $\mathcal{G} := \{\frac{k}{N} : k = 0, \ldots, N\}$. Note that for all even $h = 2m$, it holds almost surely that $\widehat{\mu}_h(s_{\text{Left}}), \widehat{\mu}_h(s_{\text{Right}}) \in \mathcal{G}$. By the definition of the MFG, it holds for any $m \geq 0, k \in [N]$ that

$$\mathbb{P}[NX_0 = k] = \binom{N}{k} 2^{-N},$$
$$\mathbb{P}[NX_{m+1} = k | X_m] = \binom{N}{k} (\omega_\varepsilon(X_m))^k (1 - \omega_\varepsilon(X_m))^k,$$

that is, given X_m, NX_{m+1} is binomially distributed with $NX_{m+1} \sim \text{Binom}(N, \omega_\epsilon(X_m))$ without any dependence on the actions played by agents. Therefore

$$\mathbb{E}\left[X_{m+1} | X_m\right] = \omega_\epsilon(X_m), \quad \mathbb{V}\text{ar}[X_{m+1} | X_m] \leq \frac{1}{4N}.$$

We define the following set $\mathcal{G}_* := \{0, 1\} \subset \mathcal{G}$. By the definition of the mechanics, if $x \in \mathcal{G}_*, m \in \mathbb{N}_{\geq 0}$, it holds for all $m' > m$ that $\mathbb{P}[X_{m'} = X_m | X_m = x] = 1$, that is once the Markovian random process X_m hits $\mathcal{G}_*$, it will remain in $\mathcal{G}_*$. Furthermore, for $K := \lfloor \log_5 \sqrt{N} \rfloor$, and for $k = 0, \ldots, K$ define the level sets:

$$\mathcal{G}_{-1} := \mathcal{G}, \quad \mathcal{G}_k := \left\{ x \in \mathcal{G} : \left| x - \frac{1}{2} \right| \geq \frac{5^k}{2\sqrt{N}} \right\}.$$

For all $k \geq K$, define $\mathcal{G}_k := \mathcal{G}_*$.

Firstly, we have that

$$\begin{aligned}\mathbb{P}[X_0 \in \mathcal{G}_0] =& \mathbb{P}\left[\left|\frac{1}{N}\sum_i \mathbb{1}_{\{s_0^i = s_{\text{Left}}\}} - \frac{1}{2}\right| \geq \frac{1}{2\sqrt{N}}\right] \\ =& \mathbb{P}\left[\left|\sum_i \mathbb{1}_{\{s_0^i = s_{\text{Left}}\}} - \frac{N}{2}\right| \geq \frac{\sqrt{N}}{2}\right] \geq \frac{1}{10},\end{aligned}$$

where in the last line we applied the anti-concentration result of Lemma A.4 on the sum of independent Bernoulli random variables $\mathbb{1}_{\{s_0^i = s_{\text{Left}}\}}$ for $i \in [N]$.

Next, assume that for some $m \in 1, \ldots, K-1$ we have $p \in \mathcal{G}_m$. If $\omega_\epsilon(p) \in \{0,1\}$, it holds trivially that $\mathbb{P}[X_{m+1} \in \mathcal{G}_{m+1}|X_m = p] = 1$. Otherwise, if $\omega_\epsilon(p) \in (0,1)$,

$$\begin{aligned}&\mathbb{P}[X_{m+1} \in \mathcal{G}_{m+1}|X_m = p] \\ =& \mathbb{P}\left[|X_{m+1} - \frac{1}{2}| \geq \frac{5^{m+1}}{2\sqrt{N}} \,\middle|\, X_m = p\right] \\ \geq& \mathbb{P}\left[|\omega_\epsilon(p) - \frac{1}{2}| - |X_{m+1} - \omega_\epsilon(p)| \geq \frac{5^{m+1}}{2\sqrt{N}} \middle| X_m = p\right].\end{aligned}$$

Since in this case $|\omega_\epsilon(X_m) - \frac{1}{2}| = |\omega_\epsilon(X_m) - \omega_\epsilon(\frac{1}{2})| \geq {}^{1}/_{2\epsilon}|X_m - \omega_\epsilon(\frac{1}{2})|$, we have

$$\begin{aligned}&\mathbb{P}[X_{m+1} \in \mathcal{G}_{m+1}|X_m = p] \\ \geq& \mathbb{P}\left[|\omega_\epsilon(p) - \frac{1}{2}| - |X_{m+1} - \omega_\epsilon(p)| \geq \frac{5^{m+1}}{2\sqrt{N}} \middle| X_m = p\right] \\ =& \mathbb{P}\left[\left|X_{m+1} - \omega_\epsilon(p)\right| \leq \left|\omega_\epsilon(p) - \frac{1}{2}\right| - \frac{5^{m+1}}{2\sqrt{N}} \middle| X_m = p\right] \\ \geq& \mathbb{P}\left[\left|X_{m+1} - \omega_\epsilon(p)\right| \leq 8\frac{5^m}{2\sqrt{N}} - \frac{5^{m+1}}{2\sqrt{N}} \middle| X_m = p\right] \\ =& \mathbb{P}\left[\left|X_{m+1} - \omega_\epsilon(p)\right| \leq 3\frac{5^m}{2\sqrt{N}} \middle| X_m = p\right] \\ \geq& 1 - 2\exp\left\{-\frac{9}{50}25^{m+1}\right\}\end{aligned}$$

where in the last line we invoked the Hoeffding concentration bound (Lemma A.2).

Using the above result inductively for $m \in 0, \ldots, K$ it holds that

$$\begin{aligned}\mathbb{P}[X_m \in \mathcal{G}_m|X_0 \in \mathcal{G}_0] \geq& \prod_{m'=1}^{m} \mathbb{P}[X_{m'} \in \mathcal{G}_{m'}|X_{m'-1} \in \mathcal{G}_{m'-1}] \\ \geq& \prod_{m'=1}^{m}\left(1 - 2\exp\left\{-\frac{9}{50}25^{m'}\right\}\right) \\ \geq& \left(1 - 2\sum_{m'=0}^{\infty}\exp\left\{-\frac{9}{50}25^{m'+1}\right\}\right) \\ \geq& \left(1 - 2\sum_{m'=0}^{\infty}\exp\left\{-\frac{9}{2}m' - \frac{9}{2}\right\}\right) \\ \geq& \left(1 - \frac{2e^{-9/2}}{1 - e^{-9/2}}\right) \geq \frac{9}{10}.\end{aligned}$$

Since for $k > K$, $\mathbb{P}[X_{k+1} \in \mathcal{G}_*|X_k \in \mathcal{G}_*] = 1$ and $\mathbb{P}[X_0 \in \mathcal{G}_0] \geq {}^{1}/_{10}$, it also holds that

$$\mathbb{P}[X_m \in \mathcal{G}_m, \forall m \geq 0] \geq \frac{9}{100}.$$

Finally, we use the above lower bound on the probability to lower bound the expectation:

$$
\begin{aligned}
\mathbb{E}\left[\|\widehat{\mu}_{2m}-\mu_{2m}\|_1\right] &\geq \mathbb{P}[X_m \in \mathcal{G}_m]\,\mathbb{E}\left[\|\widehat{\mu}_{2m}-\mu_{2m}\|_1 \,|X_m \in \mathcal{G}_m\right] \\
&\geq \mathbb{P}[X_m \in \mathcal{G}_m]\,\mathbb{E}\left[2|X_m - 1/2| \;\; |X_m \in \mathcal{G}_m\right] \\
&\geq \frac{9}{100}\,\min\left\{\frac{5^m}{\sqrt{N}}, 1\right\}.
\end{aligned}
$$

For odd $h = 2m+1$, we also have the inequality

$$
\begin{aligned}
\mathbb{E}\left[\|\widehat{\mu}_{2m+1}-\mu_{2m+1}\|_1\right] &\geq \mathbb{E}\left[\|\widehat{\mu}_{2m}-\mu_{2m}\|_1\right] \\
&\geq \frac{9}{100}\,\min\left\{\frac{5^m}{\sqrt{N}}, 1\right\}.
\end{aligned}
$$

which completes the first statement of the theorem (as $5^{H/2} = \Omega(2^H)$).

Step 3: Hitting time for $\mathcal{G}_*$. We will show that the empirical distribution of agent states almost always concentrates on one of $s_{\text{Left}}, s_{\text{Right}}$ during the even rounds in the N-player game, and bound the expected waiting time for this to happen. The distributions of agents over states $s_{\text{Left}}, s_{\text{Right}}$ in the even rounds are policy independent (they are not affected by which actions are played): hence the results from Step 2 still hold for the population distribution and the expected time computed in this step will be valid for any policy.

For simplicity, we define the FH-MFG for the non-terminating infinite horizon chain, and we will compute value functions up to horizon H. Define the (random) hitting time τ as follows:

$$
\tau := \inf\{m \geq 0 : \widehat{\mu}_{2m}(s_{\text{Left}}) \in \mathcal{G}_*\} = \inf\{m \geq 0 : X_m \in \mathcal{G}_*\}.
$$

Note that for any $p \in \mathcal{G}$, it holds that $\mathbb{P}[X_{m+1} \in \mathcal{G}_* | X_m = p] = \widehat{\mu}_{2m}(s_{\text{Left}})^N + \widehat{\mu}_{2m}(s_{\text{Right}})^N = p^N + (1-p)^N \geq 2^{-N}$. Therefore for all m it holds that $\mathbb{P}[\widehat{\mu}_{2m} \notin \mathcal{G}_*] \leq \left(1 - 2^{-N}\right)^{m-1}$. By the Borel-Cantelli lemma, we can conclude that $\tau < \infty$ almost surely, and in particular $T_\tau := \mathbb{E}[\tau | X_0 = x] < \infty$ for any $x \in \mathcal{G}$.

Next, we compute the expected value T_τ. Define the following two quantities:

$$
\begin{aligned}
T_{-1} &:= \sup_{x \in \mathcal{G}_{-1}} \{\mathbb{E}[\tau | X_0 = x]\} \\
T_0 &:= \sup_{x \in \mathcal{G}_0} \{\mathbb{E}[\tau | X_0 = x]\}.
\end{aligned}
$$

First, we compute an upper bound for T_0. Define the event:

$$
E_0 := \bigcap_{m' \in [K]} \{X_{m'} \in \mathcal{G}_{m'}\}.
$$

Then, T_0 is upper bounded by:

$$
\begin{aligned}
T_0 &= \sup_{x \in \mathcal{G}_0} \mathbb{E}[\tau | X_0 = x] \\
&= \sup_{x \in \mathcal{G}_0} \mathbb{E}[\tau | E_0, X_0 = x]\,\mathbb{P}[E_0 | X_0 = x] \\
&\quad + \mathbb{E}[\tau | E_0^c, X_0 = x]\,\mathbb{P}[E_0^c | X_0 = x] \\
&\leq \sup_{x \in \mathcal{G}_0} \mathbb{E}[\tau | E_0, X_0 = x]\,\mathbb{P}[E_0 | X_0 = x] \\
&\quad + \mathbb{E}[\tau | E_0^c, X_0 = x]\,\mathbb{P}[E_0^c | X_0 = x] \\
&\leq K\frac{9}{10} + (K + T_{-1})\,\frac{1}{10} = K + \frac{T_{-1}}{10}
\end{aligned}
$$

where in the last step we used the lower bound on $\mathbb{P}[E_0]$ from Step 2. Similarly for T_{-1}, from the one-sided anti-concentration bound (Lemma A.4) it holds that:

$$\begin{aligned}
T_{-1} \leq & \sup_{x \in \mathcal{G}_{-1}} \mathbb{E}[\tau | X_0 = x] \\
\leq & \mathbb{E}[\tau | x \in \mathcal{G}_0, X_0 = x]\,\mathbb{P}[x \in \mathcal{G}_0 | X_0 = x] \\
& + \mathbb{E}[\tau | x \notin \mathcal{G}_0, X_0 = x]\,\mathbb{P}[x \notin \mathcal{G}_0 | X_0 = x] \\
\leq & \frac{1}{20}(T_0 + 1) + \frac{19}{20}(T_{-1} + 1),
\end{aligned}$$

the last line following since $T_{-1} > T_0$ by definition. Solving the two inequalities, we obtain

$$T_\tau \leq T_{-1} \leq \frac{200}{9} + \frac{10K}{9} \leq 23 + \frac{5}{9}\log_5 N.$$

Step 4: Ergodic optimal response to N-players. Next, we formulate a policy $\boldsymbol{\pi}^{\text{br}} = \{\pi_h^{\text{br}}\}_{h=0}^{H-1} \in \Pi^H$ that is ergodically optimal for the N-player game and can exploit a population that deploys the unique FH-MFG-NE. For all h, the optimal policy will be defined by:

$$\pi_h^{\text{br}}(a|s) = \begin{cases} 1, \text{ if } s = s_{\text{Left}}, a = a_{\text{A}} \\ 1, \text{ if } s = s_{\text{Right}}, a = a_{\text{B}} \\ 1, \text{ if } s \notin \{s_{\text{Left}}, s_{\text{Right}}\}, a = a_{\text{B}} \\ 0, \text{ otherwise} \end{cases}$$

Intuitively, π_h^{br} becomes optimal once all the agents are concentrated in the same states during the even rounds, which happens very quickly as shown in Step 3. Assume that agents $i = 2, \ldots N$ deploy the unique FH-MFG-NE $\boldsymbol{\pi}^i = \boldsymbol{\pi}^*$, and for agent $i = 1$, $\boldsymbol{\pi}^1 = \boldsymbol{\pi}^{\text{br}}$. We decompose the three components of the rewards for the first agent, as defined in the construction of the MFG (Step 1):

$$\begin{aligned}
& J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}^{\text{br}}, \boldsymbol{\pi}^*, \ldots, \boldsymbol{\pi}^*) \\
& = \mathbb{E}\left[\sum_{\substack{h \text{ odd} \\ 0 \leq h \leq H}} (1 - \alpha - \beta) R_h^{1,\mathbf{g}} + \alpha R_h^{1,\mathbf{h}} + \beta \mathbb{1}_{\{a_h^1 = a_B\}}\right] \\
& \geq (1 - \alpha - \beta)\mathbb{E}\left[\sum_{\text{odd } h=0}^{H-1} R_h^{1,\mathbf{g}}\right] + \beta \left\lfloor \frac{H}{2} \right\rfloor
\end{aligned}$$

as by definition clearly $\mathbb{E}\left[\mathbb{1}_{\{a_h^1 = a_B\}}\right] = 1$ for all odd h and $R_h^{\mathbf{h}} \geq 0$ almost surely.

We analyze the terms $R_h^{1,\mathbf{g}}$ when the first agent follows $\boldsymbol{\pi}^{\text{br}}$. By the definition of the dynamics and $\boldsymbol{\pi}^{\text{br}}$, it holds that

$$R_h^{1,\mathbf{g}} = g_1(\widehat{\mu}_{h-1}(s_{h-1}^1), \widehat{\mu}_{h-1}(\bar{s}_{h-1}^1))$$

where $\bar{s}_{h-1}^1 := s_{\text{Left}}$ if $s_{h-1}^1 = s_{\text{Right}}$ and $\bar{s}_{h-1}^1 := s_{\text{Right}}$ if $s_{h-1}^1 = s_{\text{Left}}$. As $\mathbb{P}[s_{h-1}^1 = \cdot, \ldots, s_{h-1}^N = \cdot]$ at even step $h - 1$ is permutation invariant, it holds that $\mathbb{P}[s_{h-1}^1 = \cdot | \widehat{\mu}_{h-1} = \mu] = \mu(\cdot)$ for any $\mu \in \mathcal{G}$. Therefore,

$$\begin{aligned}
\mathbb{E}[R_h^{1,\mathbf{g}}] = & \sum_{\substack{\mu \in \mathcal{G} \\ s \in \{s_{\text{Left}}, s_{\text{Right}}\}}} \mathbb{P}[\widehat{\mu}_{h-1} = \mu]\,\mathbb{P}[s_{h-1}^1 = s | \widehat{\mu}_{h-1} = \mu] \\
& \mathbb{E}[R_h^{1,\mathbf{g}} | s_{h-1}^1 = s, \widehat{\mu}_{h-1} = \mu] \\
= & \sum_{\substack{\mu \in \mathcal{G} \\ s \in \{s_{\text{Left}}, s_{\text{Right}}\}}} \mathbb{P}[\widehat{\mu}_{h-1} = \mu] \mu(s) g_1(\mu(s), \mu(\bar{s})) \geq 1/2,
\end{aligned}$$

as for any μ, if s is such that $\mu(s) \geq \mu(\bar{s})$ then $g_1(\mu(s), \mu(\bar{s})) = 1$. Furthermore, by the definition of the hitting time τ, for any odd $h \geq 1$, $\mathbb{E}\,[R_h^{\mathbf{g}}\,|2\tau < h] = \mathbb{E}\,[R_h^{\mathbf{g}}\,|\widehat{\mu}_{h-1}(s_{\text{Left}}) \in \mathcal{G}_*] = 1$, as after time 2τ the action a_{A} will be optimal with reward $R_h^{\mathbf{g}} = 1$ almost surely, as $\boldsymbol{\pi}^{br}$ chooses action a_{A} at even steps.

Finally, using the lower bound of $1/2$ for $R_h^{\mathbf{g}}$ when $h < 2\tau$ and that $R_h^{\mathbf{g}} = 1$ when $h > 2\tau$, we obtain:

$$
\begin{aligned}
\mathbb{E}\left[\sum_{\substack{h \text{ odd} \\ 0\leq h\leq H}} R_h^{\mathbf{g}}\right] =& \mathbb{E}\left[\sum_{\substack{h \text{ odd} \\ 0\leq h\leq \min\{2\tau, H\}}} R_h^{1,\mathbf{g}} + \sum_{\substack{h \text{ odd} \\ \min\{2\tau,H\}+1\leq h<H}} R_h^{1,\mathbf{g}}\right] \\
\geq& \mathbb{E}\left[\frac{1}{2}\min\left\{\tau, \left\lfloor\frac{H}{2}\right\rfloor\right\} + \left(\left\lfloor\frac{H}{2}\right\rfloor - \min\left\{\tau, \left\lfloor\frac{H}{2}\right\rfloor\right\}\right)\right] \\
\geq& \left\lfloor\frac{H}{2}\right\rfloor - \frac{1}{2}\mathbb{E}\left[\min\left\{\tau, \left\lfloor\frac{H}{2}\right\rfloor\right\}\right] \\
\geq& \left\lfloor\frac{H}{2}\right\rfloor - \frac{1}{2}\mathbb{E}\left[\tau\right] = \left\lfloor\frac{H}{2}\right\rfloor - \frac{T_\tau}{2}
\end{aligned}
$$

Merging the inequalities above, we obtain

$$
J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}^{\text{br}}, \boldsymbol{\pi}^*, \ldots, \boldsymbol{\pi}^*) \geq (1-\alpha-\beta)\left(\left\lfloor\frac{H}{2}\right\rfloor - \frac{T_\tau}{2}\right) + \beta\left\lfloor\frac{H}{2}\right\rfloor.
$$

Step 5: Bounding exploitability. Finally, we will upper bound also the expected reward of the FH-MFG-NE policy $\boldsymbol{\pi}^*$ and hence lower bound the exploitability. Our conclusion will be that $\boldsymbol{\pi}^*$ suffers from a non-vanishing exploitability for large H, as $\boldsymbol{\pi}^{\text{br}}$ becomes the best response policy after $H \gtrsim \log N$. In this step, we assume the probability space induced by all N agents following FH-MFG-NE policy $\boldsymbol{\pi}^{\text{br}}$.

We have the definition

$$
\begin{aligned}
J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}^*, \boldsymbol{\pi}^*, \ldots, \boldsymbol{\pi}^*) =& \mathbb{E}\left[\sum_{h=0}^{H-1} R(s_h^1, a_h^1, \widehat{\mu}_h)\right] \\
\leq& (1-\alpha-\beta)\,\mathbb{E}\left[\sum_{\text{odd } h=0}^{H-1} R_h^{1,\mathbf{g}}\right] + (\alpha+\beta)\left\lfloor\frac{H}{2}\right\rfloor
\end{aligned}
$$

This time, when h odd and $h > 2\tau$, it holds that $\mathbb{E}[R_h^{\mathbf{g}}|h > 2\tau] = 1/2$ since $\boldsymbol{\pi}^*$ takes actions $a_{\text{A}}, a_{\text{B}}$ with equal probability in even steps, yielding $R_h^{\mathbf{g}} = 1$ and $R_h^{\mathbf{g}} = 0$ respectively almost surely. As before,

$$
\begin{aligned}
\mathbb{E}\left[\sum_{\substack{h \text{ odd} \\ 0\leq h\leq H}} R_h^{\mathbf{g}}\right] =& \mathbb{E}\left[\sum_{\substack{h \text{ odd} \\ 0\leq h\leq \min\{2\tau, H\}}} R_h^{1,\mathbf{g}} + \sum_{\substack{h \text{ odd} \\ \min\{2\tau,H\}+1\leq h<H}} R_h^{1,\mathbf{g}}\right] \\
\leq& \mathbb{E}\left[\min\left\{\tau, \left\lfloor\frac{H}{2}\right\rfloor\right\} + \frac{1}{2}\left(\left\lfloor\frac{H}{2}\right\rfloor - \min\left\{\tau, \left\lfloor\frac{H}{2}\right\rfloor\right\}\right)\right] \\
=& \frac{1}{2}\mathbb{E}\left[\left\lfloor\frac{H}{2}\right\rfloor + \min\left\{\tau, \left\lfloor\frac{H}{2}\right\rfloor\right\}\right] \\
\leq& \frac{1}{2}\left\lfloor\frac{H}{2}\right\rfloor + \frac{1}{2}\,\mathbb{E}[\tau] = \frac{1}{2}\left\lfloor\frac{H}{2}\right\rfloor + \frac{1}{2}T_\tau.
\end{aligned}
$$

The statement of the theorem then follows by lower bounding the exploitability as follows:

$$
\begin{aligned}
&\mathcal{E}_{P,R}^{H,N,(1)}(\boldsymbol{\pi}^*, \boldsymbol{\pi}^*, \dots, \boldsymbol{\pi}^*) \\
&\qquad = \max_{\boldsymbol{\pi}} J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}, \boldsymbol{\pi}^*, \dots, \boldsymbol{\pi}^*) - J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}^*, \boldsymbol{\pi}^*, \dots, \boldsymbol{\pi}^*) \\
&\qquad \geq J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}^{\text{br}}, \boldsymbol{\pi}^*, \dots, \boldsymbol{\pi}^*) - J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}^*, \boldsymbol{\pi}^*, \dots, \boldsymbol{\pi}^*) \\
&\qquad \geq (1-\alpha-\beta)\left(\left\lfloor \frac{H}{2} \right\rfloor - \frac{T_\tau}{2} - \frac{1}{2}\left\lfloor \frac{H}{2} \right\rfloor - \frac{T_\tau}{2}\right) - \alpha \left\lfloor \frac{H}{2} \right\rfloor \\
&\qquad \geq (1-\alpha-\beta)\left(\frac{H}{4} - 24 - \frac{5}{9}\log_5 N\right) - \alpha \left\lfloor \frac{H}{2} \right\rfloor
\end{aligned}
$$

The above inequality implies that if $H \geq \log_2 N$, then

$$
\begin{aligned}
&\mathcal{E}_{P,R}^{H,N,(1)}(\boldsymbol{\pi}^*, \boldsymbol{\pi}^*, \dots, \boldsymbol{\pi}^*) \\
&\qquad \geq (1-\alpha-\beta)\left(\frac{1}{4} - \frac{5}{9\log_2 5}\right) H - \alpha \frac{H}{2} - 24,
\end{aligned}
$$

which implies $\mathcal{E}_{P,R}^{H,N,(1)}(\boldsymbol{\pi}^*, \boldsymbol{\pi}^*, \dots, \boldsymbol{\pi}^*) \geq \Omega(H)$ by choosing α, β small constants as $\frac{1}{4} - \frac{5}{9\log_2 5} > 0$.

A.4. Upper Bound for Stat-MFG: Extended Proof of Theorem 3.5

Let μ^*, π^* be a δ-Stat-MFG-NE. As before, the proof will proceed in three steps:

- **Step 1.** Bounding the expected deviation of the empirical population distribution from the mean-field distribution $\mathbb{E}\left[\|\widehat{\mu}_h - \mu^*\|_1\right]$ for any given policy $\boldsymbol{\pi}$.

- **Step 2.** Bounding difference of N agent value function $J_{P,R}^{\gamma,N,(i)}$ and the infinite player value function $V_{P,R}^{\gamma}$ in the stationary mean-field game setting.

- **Step 3.** Bounding the exploitability of an agent when each of N agents are playing the Stat-MFG-NE policy.

Step 1: Empirical distribution bound. We first analyze the deviation of the empirical population distribution $\widehat{\mu}_t$ over time from the stable distribution μ^*. For this, we state the following lemma and prove it using techniques similar to Corollary D.4 of (Yardim et al., 2023a).

Lemma A.11. *Assume that the conditions of Theorem 3.5 hold, and that $(\mu^*, \pi^*) \in \Delta_{\mathcal{S}}$ is a Stat-MFG-NE. Furthermore, assume that the N agents follow policies $\{\pi^i\}_{i=1}^N$ in the N-Stat-MFG, define $\Delta_{\overline{\pi}} := \frac{1}{N}\sum_i \|\overline{\pi} - \pi^i\|_1$. Then, or any $t \geq 0$, we have*

$$
\mathbb{E}\left[\|\mu^* - \widehat{\mu}_t\|_1\right] \leq \frac{tK_a\Delta_\pi}{2} + \frac{2(t+1)\sqrt{|\mathcal{S}|}}{\sqrt{N}}.
$$

Proof. $\mathcal{F}_t$ as the σ-algebra generated by the states of agents $\{s_t^i\}$ at time t. For $\widehat{\mu_0}$, we have by definitions that

$$
\mathbb{E}\left[\widehat{\mu_0}\right] = \mathbb{E}\left[\frac{1}{N}\sum_i \mathbf{e}_{s_t^i}\right] = \mu^*
$$

$$
\mathbb{E}\left[\|\widehat{\mu_0} - \mu^*\|_2^2\right] = \mathbb{E}\left[\frac{1}{N^2}\sum_i \left\|\left(\mathbf{e}_{s_t^i} - \mu^*\right)\right\|_2^2\right] \leq \frac{4}{N}
$$

where the last line follows by independence. The two above imply $\mathbb{E}\left[\|\widehat{\mu_0} - \mu^*\|_1\right] \leq \frac{2\sqrt{|\mathcal{S}|}}{\sqrt{N}}$.

Next, we inductively calculate:

$$
\begin{aligned}
\mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{s'\in\mathcal{S}}\sum_{i=1}^{N}\mathbb{1}(s^i_{t+1}=s')\mathbf{e}_{s'}\middle|\mathcal{F}_t\right] \\
&= \sum_{s'\in\mathcal{S}}\mathbf{e}_{s'}\sum_{i=1}^{N}\frac{1}{N}\overline{P}(s'|s^i_t,\pi^i(s^i_t),\widehat{\mu}_t), \qquad (10)\\
\mathbb{E}[\|\widehat{\mu}_{t+1}-\mathbb{E}[\widehat{\mu}_{t+1}|\mathcal{F}_t]\|_2^2|\mathcal{F}_t] & \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}[\|\mathbf{e}_{s^i_{t+1}}-\mathbb{E}[\mathbf{e}_{s^i_{t+1}}|\mathcal{F}_t]\|_2^2|\mathcal{F}_t] \leq \frac{4}{N}. \qquad (11)
\end{aligned}
$$

We bound the ℓ_1 distance to the stable distribution as

$$
\begin{aligned}
&\mathbb{E}\left[\|\widehat{\mu}_{t+1}-\mu^*\|_1|\mathcal{F}_t\right] \\
&\quad\leq \underbrace{\mathbb{E}\left[\|\,\mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right]|\mathcal{F}_t\right]-\mu^*\|_1}_{(\square)} + \underbrace{\mathbb{E}\left[\|\,\mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right]-\widehat{\mu}_{t+1}\|_1\mathcal{F}_t\right]}_{(\triangle)}.
\end{aligned}
$$

The two terms can be bounded separately using Inequalities (10) and (11).

$$
\begin{aligned}
(\triangle) \leq& \sqrt{|\mathcal{S}|}\,\mathbb{E}\left[\|\,\mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right]-\widehat{\mu}_{t+1}\|_2\mathcal{F}_t\right] \\
\leq& \sqrt{|\mathcal{S}|}\sqrt{\mathbb{E}\left[\|\,\mathbb{E}\left[\widehat{\mu}_{t+1}|\mathcal{F}_t\right]-\widehat{\mu}_{t+1}\|_2^2\mathcal{F}_t\right]} \leq \frac{2\sqrt{|\mathcal{S}|}}{\sqrt{N}} \\
(\square) =& \left\|\sum_{s'\in\mathcal{S}}\mathbf{e}_{s'}\sum_{i=1}^{N}\frac{1}{N}\overline{P}(s'|s^i_t,\pi^i(s^i_t),\widehat{\mu}_t)-\mu^*\right\|_1 \\
=& \left\|\sum_{s'\in\mathcal{S}}\mathbf{e}_{s'}\sum_{i=1}^{N}\frac{1}{N}\overline{P}(s'|s^i_t,\pi^i(s^i_t),\widehat{\mu}_t)-\Gamma_{pop}(\pi^*,\mu^*)\right\|_1 \\
\leq& \left\|\sum_{i=1}^{N}\frac{1}{N}\overline{P}(\cdot|s^i_t,\pi^i(s^i_t),\widehat{\mu}_t)-\sum_{i=1}^{N}\frac{1}{N}\overline{P}(\cdot|s^i_t,\pi^*(s^i_t),\widehat{\mu}_t)\right\|_1 \\
&+ \left\|\sum_{s'\in\mathcal{S}}\widehat{\mu}_t(s')\overline{P}(s'|s^i_t,\pi^i(s^i_t),\widehat{\mu}_t)-\Gamma_{pop}(\pi^*,\mu^*)\right\|_1 \\
\leq& \frac{K_a}{2N}\sum_i\|\pi^*-\pi^i\|_1+\|\Gamma_{pop}(\pi^*,\widehat{\mu}_t)-\Gamma_{pop}(\pi^*,\mu^*)\|_1 \\
\leq& \frac{K_a\Delta_\pi}{2}+\|\mu^*-\widehat{\mu}_t\|_1
\end{aligned}
$$

Hence, by the law of total expectation, we can conclude

$$
\mathbb{E}\left[\|\mu^*-\widehat{\mu}_{t+1}\|_1\right] \leq \mathbb{E}\left[\|\mu^*-\widehat{\mu}_t\|_1\right]+\frac{K_a\Delta_\pi}{2}+\frac{2\sqrt{|\mathcal{S}|}}{\sqrt{N}}
$$

or inductively,

$$
\mathbb{E}\left[\|\mu^*-\widehat{\mu}_t\|_1\right] \leq \frac{tK_a\Delta_\pi}{2}+\frac{2(t+1)\sqrt{|\mathcal{S}|}}{\sqrt{N}}.
$$

$\square$

Step 2: Bounding difference in value functions. Next, we bound the differences in the infinite-horizon

Lemma A.12. *Suppose N-Stat-MFG agents follow the same sequence of policy π^*. Then for all i,*

$$\begin{aligned}&|J^{\gamma,N,(i)}_{P,R}(\pi^*,\dots,\pi^*) - V^{\gamma}_{P,R}(\mu^*,\pi^*)|\\ &\leq \frac{\gamma}{1-\gamma}\left(L_\mu + \frac{L_s}{2}\right)\frac{2\sqrt{|\mathcal{S}|}}{\sqrt{N}}\end{aligned}$$

Proof. For ease of reading, in this proof expectations, probabilities, and laws of random variables will be denoted $\mathbb{E}_\infty, \mathbb{P}_\infty, \mathcal{L}_\infty$ respectively over the infinite player finite horizon game and $\mathbb{E}_N, \mathbb{P}_N, \mathcal{L}_N$ respectively over the N-player game. Due to symmetry in the N agent game, any permutation $\sigma : [N] \to [N]$ of agents does not change their distribution, that is $\mathcal{L}_N(s_t^1,\dots,s_t^N) = \mathcal{L}_N(s_t^{\sigma(1)},\dots,s_t^{\sigma(N)})$. We can then conclude that:

$$\begin{aligned}\mathbb{E}_N\left[R(s_t^1,a_t^1,\widehat{\mu}_h)\right] &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}_N\left[R(s_t^i,a_t^i,\widehat{\mu}_t)\right]\\ &= \mathbb{E}_N\left[\sum_{s\in\mathcal{S}}\widehat{\mu}_t(s)\overline{R}(s,\pi_t(s),\widehat{\mu}_t).\right]\end{aligned}$$

Therefore, we by definition:

$$J^{\gamma,N,(1)}_{P,R}(\boldsymbol{\pi},\dots,\boldsymbol{\pi}) = \mathbb{E}_N\left[\sum_{t=0}^\infty\sum_{s\in\mathcal{S}}\widehat{\mu}_t(s)\overline{R}(s,\pi^*(s),\widehat{\mu}_t)\right].$$

Next, in the Stat-MFG, we have that for all $t \geq 0$,

$$\begin{aligned}\mathbb{P}_\infty(s_t = \cdot) &= \mu^*,\\ \mathbb{P}_\infty(s_{t+1} = \cdot) &= \sum_{s\in\mathcal{S}}\mathbb{P}_\infty(s_t = s)\,\mathbb{P}_\infty(s_t = \cdot|s_t = s)\\ &= \Gamma_P(\mathbb{P}_\infty(s_t = s),\pi^*) = \mu^*,\end{aligned}$$

so by induction $\mathbb{P}_\infty(s_t = \cdot) = \mu^*$. Then we can conclude that

$$\begin{aligned}V^\gamma_{P,R}(\mu^*,\pi^*) &= \mathbb{E}_\infty\left[\sum_{t=0}^\infty \gamma^t R(s_t,\pi^*(s_t),\mu_t)\right]\\ &= \sum_{t=0}^\infty\gamma^t\sum_{s\in\mathcal{S}}\mu^*(s)R(s,\pi^*(s),\mu^*),\end{aligned}$$

by a simple application of the dominated convergence theorem. We next bound the differences in truncated expect reward until some time $T > 0$:

$$\begin{aligned}&\Bigg|\mathbb{E}_N\left[\sum_{t=0}^T\gamma^t\sum_{s\in\mathcal{S}}\widehat{\mu}_t(s)\overline{R}(s,\pi^*(s),\widehat{\mu}_t)\right]\\ &\quad-\sum_{t=0}^T\gamma^t\sum_{s\in\mathcal{S}}\mu_t(s)R(s,\pi^*(s),\mu_t)\Bigg|\\ &\leq\mathbb{E}_N\left[\sum_{t=0}^T\gamma^t\left|\sum_{s\in\mathcal{S}}\left(\widehat{\mu}_t(s)\overline{R}(s,\pi^*(s),\widehat{\mu}_t) - \mu^*(s)R(s,\pi^*(s),\mu^*)\right)\right|\right]\\ &\leq\mathbb{E}_N\left[\sum_{t=0}^T\gamma^t\left(\frac{L_s}{2}\|\mu^*-\widehat{\mu}_t\|_1 + L_\mu\|\mu^*-\widehat{\mu}_t\|_1\right)\right]\\ &\leq\sum_{t=0}^T\gamma^t\left(L_\mu+\frac{L_s}{2}\right)\mathbb{E}_N\left[\|\mu^*-\widehat{\mu}_t\|_1\right]\\ &\leq\frac{1}{(1-\gamma)^2}\left(L_\mu+\frac{L_s}{2}\right)\frac{2\sqrt{|\mathcal{S}|}}{\sqrt{N}}\end{aligned}$$

Taking $T \to \infty$ and applying once again the dominated convergence theorem the result is obtained. $\square$

Step 3: Bounding difference in policy deviation. Finally, to conclude the proof of the main theorem of this section, we will prove that the improvement in expectation due to single-sided policy changes are at most of order $\mathcal{O}\left(\delta + \frac{1}{\sqrt{N}}\right)$.

Lemma A.13. *Suppose we have two policy sequences $\pi^*, \pi \in \Pi$ and $\mu^* \in \Delta_{\mathcal{S}}$ such that $\Gamma_P(\mu^*, \pi^*) = \mu^*$ and $\Gamma_P(\cdot, \pi^*)$ is non-expansive. Then,*

$$\begin{aligned}
&\left| J_{P,R}^{\gamma,N,(1)}(\pi', \pi^*, \dots, \pi^*) - V_{P,R}^{\gamma}(\mu^*, \pi') \right| \\
&\leq \sum_{t=0}^{\infty} \gamma^t \left(L_\mu \mathbb{E}\left[\|\widehat{\mu}_t - \mu_t^{\boldsymbol{\pi}}\|_1\right] + K_\mu \sum_{t'=0}^{t-1} \mathbb{E}\left[\|\widehat{\mu}_{t'} - \mu_{t'}^{\boldsymbol{\pi}}\|_1\right] \right) \\
&\leq \left(\frac{K_a}{2N} + \frac{2\sqrt{|\mathcal{S}|}}{\sqrt{N}} \right) \frac{L_\mu/2 + K_\mu}{(1-\gamma)^3}
\end{aligned}$$

Proof. For the truncated game T, it still holds by the derivation in the FH-MFG that:

$$\begin{aligned}
&\left|\mathbb{E}_N\left[R(s_t^1, a_t^1, \widehat{\mu}_t)\right] - \mathbb{E}_\infty\left[R(s_t, a_t, \mu_t^{\boldsymbol{\pi}})\right]\right| \\
&\quad \leq \frac{L_\mu}{2} \mathbb{E}_N\left[\|\mu_t^{\boldsymbol{\pi}} - \widehat{\mu}_t\|_1\right] + K_\mu \sum_{t'=0}^{t-1} \mathbb{E}_N\left[\|\mu_{t'}^{\boldsymbol{\pi}} - \widehat{\mu}_{t'}\|_1\right].
\end{aligned}$$

We take the limit $T \to \infty$ and apply the dominated convergence theorem to obtain the state bound, also noting that $1/2 \cdot \sum_t (t+1)(t+2)\gamma^t \leq \frac{1}{(1-\gamma)^3}$. $\square$

Conclusion and Statement of the Result. Finally, if μ^*, π^* is a δ-Stat-MFG-NE, by definition we have that: By definition of the Stat-MFG-NE, we have:

$$\delta \geq \mathcal{E}_{P,R}^{H}(\boldsymbol{\pi}_\delta) = \max_{\pi' \in \Pi} V_{P,R}^{\gamma}(\mu^*, \pi') - V_{P,R}^{\gamma}(\mu^*, \pi^*)$$

Then using the two bounds from Steps 2,3 and the fact that π^* δ-optimal with respect to μ^*:

$$\begin{aligned}
&\max_{\pi' \in \Pi} J_{P,R}^{H,N,(1)}(\pi', \pi^*, \dots, \pi^*) - J_{P,R}^{H,N,(1)}(\pi^*, \pi^*, \dots, \pi^*) \\
&\leq 2\delta + \left(\frac{K_a}{2N} + \frac{2\sqrt{|\mathcal{S}|}}{\sqrt{N}} \right) \frac{L_\mu/2 + K_\mu}{(1-\gamma)^3} + \frac{L_\mu + L_s/2}{(1-\gamma)^2} \left(\frac{2\sqrt{|\mathcal{S}|}}{\sqrt{N}} \right)
\end{aligned}$$

A.5. Lower Bound for Stat-MFG: Extended Proof of Theorem 3.6

Similar to the finite horizon case, we define constructively the counter-example: the idea and the nature of the counter-example remain the same. However, minor details of the construction are modified, as it will not hold immediately that all agents are on states $\{s_{\text{Left}}, s_{\text{Right}}\}$ on even times t, and that the Stat-MFG-NE is unique as before.

Defining the Stat-MFG. We use the same definitions for $\mathcal{S}, \mathcal{A}, \mathbf{g}, \mathbf{h}, \omega_\epsilon$ as in the FH-MFG case. Define the convenience functions Q_L, Q_R as

$$\begin{aligned}
Q_L(\mu) &:= \frac{\mu(s_{\text{LA}}) + \mu(s_{\text{LB}})}{\max\{\mu(s_{\text{LA}}) + \mu(s_{\text{LB}}) + \mu(s_{\text{RA}}) + \mu(s_{\text{RB}}), 4/9\}}, \\
Q_R(\mu) &:= \frac{\mu(s_{\text{RA}}) + \mu(s_{\text{RB}})}{\max\{\mu(s_{\text{LA}}) + \mu(s_{\text{LB}}) + \mu(s_{\text{RA}}) + \mu(s_{\text{RB}}), 4/9\}}.
\end{aligned}$$

We define the transition probabilities:

$$\text{If } s \in \{s_{\text{LA}}, s_{\text{LB}}, s_{\text{RA}}, s_{\text{RB}}\}, \forall \mu, a:$$
$$P(s'|s,a,\mu) = \begin{cases} \omega_\epsilon(Q_L(\mu)), \text{ if } s' = s_{\text{Right}}, s \in \{s_{\text{LA}}, s_{\text{LB}}\} \\ \omega_\epsilon(Q_R(\mu)), \text{ if } s' = s_{\text{Left}}, s \in \{s_{\text{LA}}, s_{\text{LB}}\} \\ \omega_\epsilon(Q_L(\mu)), \text{ if } s' = s_{\text{Right}}, s \in \{s_{\text{RA}}, s_{\text{RB}}\} \\ \omega_\epsilon(Q_R(\mu)), \text{ if } s' = s_{\text{Left}}, s \in \{s_{\text{RA}}, s_{\text{RB}}\} \end{cases},$$

and define $P(s_{\text{Left}}, a, \mu), P(s_{\text{Right}}, a, \mu)$ as before. With previous Lipschitz continuity results, it follows that $P \in \mathcal{P}_{9/8\epsilon}$.

Similarly, we modify the reward function R as follows:

$$\begin{aligned} R(s_{\text{Left}}, a_{\text{A}}, \mu) =& R(s_{\text{Left}}, a_{\text{B}}, \mu) = 0, \\ R(s_{\text{Right}}, a_{\text{A}}, \mu) =& R(s_{\text{Right}}, a_{\text{B}}, \mu) = 0, \\ \begin{pmatrix} R(s_{\text{LA}}, a_{\text{A}}, \mu) \\ R(s_{\text{LB}}, a_{\text{A}}, \mu) \end{pmatrix} =& (1-\alpha-\beta)\mathbf{g}\big(Q_L(\mu), Q_R(\mu)\big) + \alpha \mathbf{h}(\mu(s_{\text{LA}}), \mu(s_{\text{LB}})) \\ \begin{pmatrix} R(s_{\text{LA}}, a_{\text{B}}, \mu) \\ R(s_{\text{LB}}, a_{\text{B}}, \mu) \end{pmatrix} =& (1-\alpha-\beta)\mathbf{g}\big(Q_L(\mu), Q_R(\mu)\big) + \mathbf{h}(\mu(s_{\text{LA}}), \mu(s_{\text{LB}})) \\ &+ \beta \mathbf{1} \\ \begin{pmatrix} R(s_{\text{RA}}, a_{\text{A}}, \mu) \\ R(s_{\text{RB}}, a_{\text{A}}, \mu) \end{pmatrix} =& (1-\alpha-\beta)\mathbf{g}\big(Q_R(\mu), Q_L(\mu)\big) + \alpha \mathbf{h}(\mu(s_{\text{RA}}), \mu(s_{\text{RB}})) \\ \begin{pmatrix} R(s_{\text{RA}}, a_{\text{B}}, \mu) \\ R(s_{\text{RB}}, a_{\text{B}}, \mu) \end{pmatrix} =& (1-\alpha-\beta)\mathbf{g}\big(Q_R(\mu), Q_L(\mu)\big) + \alpha \mathbf{h}(\mu(s_{\text{RA}}), \mu(s_{\text{RB}})) \\ &+ \beta \mathbf{1}, \end{aligned}$$

simple computation shows that $R \in \mathcal{R}_3$. In this proof, unlike the N-FH-SAG case, α will be chosen as a function of N, namely $\alpha = \mathcal{O}(e^{-N})$.

Step 1: Solution of the Stat-MFG. We solve the infinite agent game: let μ^*, π^* be an Stat-MFG-NE. By simple computation, one can see that for any stationary distribution μ^* of the game, probability must be distributed equally between groups of states $\{s_{\text{Left}}, s_{\text{Right}}\}$ and $\{s_{\text{LA}}, s_{\text{LB}}, s_{\text{RA}}, s_{\text{RB}}\}$, that is,

$$\begin{aligned} \mu^*(s_{\text{Left}}) + \mu^*(s_{\text{Right}}) &= 1/2, \\ \mu^*(s_{\text{LA}}) + \mu^*(s_{\text{LB}}) + \mu^*(s_{\text{RA}}) + \mu^*(s_{\text{RB}}) &= 1/2. \end{aligned}$$

It holds by the stationarity equation $\Gamma_P(\mu^*, \pi^*) = \pi^*$ that

$$\begin{aligned} \mu^*(s_{\text{Left}}) =& \mu^*(s_{\text{LA}}) + \mu^*(s_{\text{LB}}), \\ \mu^*(s_{\text{Right}}) =& \mu^*(s_{\text{RA}}) + \mu^*(s_{\text{RB}}), \\ \mu^*(s_{\text{Left}}) =& \sum_{s \in \mathcal{S}} \mu^*(s)\pi^*(a|s)P(s_{\text{Left}}|s,a,\mu^*) \\ =& P(s_{\text{Left}}|s_{\text{LA}}, a_{\text{A}}, \mu^*), \\ \mu^*(s_{\text{Right}}) =& \sum_{s \in \mathcal{S}} \mu^*(s)\pi^*(a|s)P(s_{\text{Right}}|s,a,\mu^*) \\ =& P(s_{\text{Right}}|s_{\text{LA}}, a_{\text{A}}, \mu^*), \end{aligned}$$

as $P(s_{\text{Right}}|s,a,\mu^*) = P(s_{\text{Right}}|s,a,\mu^*)$ and similarly $P(s_{\text{Left}}|s,a,\mu^*) = P(s_{\text{Left}}|s,a,\mu^*)$ for any $s \in \{s_{\text{LA}}, s_{\text{LB}}, s_{\text{RA}}, s_{\text{RB}}\}, a \in \mathcal{A}$. If $\mu^*(s_{\text{Left}}) > 1/4$, then by definition $P(s_{\text{Left}}|s_{\text{LA}}, a_{\text{A}}, \mu^*) < 1/4$, and similarly if $\mu^*(s_{\text{Left}}) < 1/4$, then by definition $P(s_{\text{Left}}|s_{\text{LA}}, a_{\text{A}}, \mu^*) > 1/4$. So it must be the case that $\mu^*(s_{\text{Left}}) = \mu^*(s_{\text{Right}}) = 1/4$. Then the unique Stat-MFG-NE must be

$$\pi^*(a|s) := \begin{cases} 1, \text{if } a = a_{\text{B}}, s \in \{s_{\text{LA}}, s_{\text{LB}}, s_{\text{RA}}, s_{\text{RB}}\} \\ \frac{1}{2}, \text{if } s \in \{s_{\text{Left}}, s_{\text{Right}}\} \\ 0, \text{if } a = a_{\text{A}}, s \in \{s_{\text{LA}}, s_{\text{LB}}, s_{\text{RA}}, s_{\text{RB}}\}, \end{cases}$$

$$\mu^*(s_{\text{RA}}) = \mu^*(s_{\text{LA}}) = \mu^*(s_{\text{RB}}) = \mu^*(s_{\text{LB}}) = 1/8,$$

as otherwise the action $\arg\min_{a\in\mathcal{A}} \pi^*(a|s_{\text{Right}})$ will be a better response in state s_{Right} and the action $\arg\min_{a\in\mathcal{A}} \pi^*(a|s_{\text{Left}})$ will be optimal in state s_{Right}.

Step 2: Expected population deviation in N-Stat-SAG. We fix $1/2\varepsilon = 3$, define the random variable $\overline{N} := N(\widehat{\mu}_0(s_{\text{Right}}) + \widehat{\mu}_0(s_{\text{Left}}))$. We will analyze the population under the event $\overline{N} := \{|\overline{N}/N - 1/2| \leq 1/18\}$, which holds with probability $\Omega(1 - e^{-N^2})$ by the Hoeffding inequality. Under the event $\overline{E}$, it holds that $\widehat{\mu}_t(s_{\text{LA}}) + \widehat{\mu}_t(s_{\text{LA}}) + \widehat{\mu}_t(s_{\text{LA}}) + \widehat{\mu}_t(s_{\text{LA}}) > 4/9$ almost surely at all t.

Fix $N_0 \in \mathbb{N}_{>0}$ such that $|N_0/N - 1/2| \leq 1/18$, in this step we will condition on $E_0 := \{\overline{N} := N_0\}$. Once again define the random process X_m for $m \in \mathbb{N}_{\geq 0}$ such that

$$X_m := \begin{cases} \frac{\widehat{\mu}_{2m}(s_{\text{Left}})}{\widehat{\mu}_{2m}(s_{\text{Left}}) + \widehat{\mu}_{2m}(s_{\text{Right}})}, \text{ if } m \text{ odd} \\ \frac{\widehat{\mu}_{2m}(s_{\text{Right}})}{\widehat{\mu}_{2m}(s_{\text{Left}}) + \widehat{\mu}_{2m}(s_{\text{Right}})}, \text{ if } m \text{ even} \end{cases}$$

with the modification at odd m necessary because of the difference in dynamics P (oscillating between $s_{\text{Left}}, s_{\text{Right}}$) from the FH-SAG case. It still holds that X_m is Markovian, and given X_m we have $N_0 X_{m+1} \sim \text{Binom}(N_0, \omega_\epsilon(X_m))$. As before, X_m is independent from the policies of agents.

Define $K := \lfloor \log_2 \sqrt{N_0} \rfloor$, $\mathcal{G} := \{k/N_0 : k = 0, \ldots, N_0\}$, $\mathcal{G}_* := \{0, 1\} \subset \mathcal{G}$ and the level sets once again as

$$\mathcal{G}_{-1} := \mathcal{G}, \quad \mathcal{G}_k := \left\{ x \in \mathcal{G} : \left| x - \frac{1}{2} \right| \geq \frac{2^k}{2\sqrt{N_0}} \right\} \text{ when } k \leq K,$$
$$\mathcal{G}_{K+1} := \mathcal{G}_*.$$

As before, using the Markov property, Hoeffding, and the fact that $|\omega_\epsilon(x) - 1/2| \geq 1/2\epsilon |x - 1/2|$ we obtain $\forall k \in 0, \ldots, K-1$, $\forall m$ that

$$\mathbb{P}[X_{m+1} \in \mathcal{G}_0 | X_m \in \mathcal{G}_{-1}, E_0] \geq 1/20$$
$$\mathbb{P}[X_{m+1} \in \mathcal{G}_{k+1} | X_m \in \mathcal{G}_k, E_0] \geq \alpha_k := 1 - 2\exp\left\{-\frac{1}{8}4^{k+1}\right\},$$

hence from the analysis before we have the lower bound

$$\mathbb{E}[|X_m - 1/2| \, | E_0] \geq C_1 \, \min\left\{\frac{2^m}{\sqrt{N_0}}, 1\right\},$$

for some absolute constant $C_2 > 0$.

Step 3. Exploitability lower bound. As in the case of FH-MFG, the ergodic optimal policy is given by

$$\overline{\pi}(a|s) = \begin{cases} 1, \text{ if } s = s_{\text{Left}}, a = a_{\text{A}} \\ 1, \text{ if } s = s_{\text{Right}}, a = a_{\text{A}} \\ 1, \text{ if } s \notin \{s_{\text{Left}}, s_{\text{Right}}\}, a = a_{\text{B}} \\ 0, \text{ otherwise} \end{cases}$$

We define the shorthand functions

$$\mathcal{S}^* := \{s_{\text{Left}}, s_{\text{Right}}\}, \quad Q(\mu) := (Q_L(\mu), Q_R(\mu)),$$
$$Q_{\min}(\mu) := \min\{Q_L(\mu), Q_R(\mu)\}, \quad Q_{\max} := \max\{Q_L(\mu), Q_R(\mu)\}.$$

We condition on $E_{\mathcal{S}^*} := \{s_0^1 \in \mathcal{S}^*\}$, that is the first agent starts from states $\{s_{\text{Left}}, s_{\text{Right}}\}$, the analysis will be similar under event $E^c_{\mathcal{S}^*}$. As in the case of FH-MFG, due to permutation invariance, it holds for any odd t and $\mu \in \{\mu' \in \Delta_{\mathcal{S}^*} : N_0\mu' \in \mathbb{N}^2_{>0}\}$ that

$$\begin{aligned}
\mathbb{P}[s_t^1 \in \{s_{\text{LA}}, s_{\text{LB}}\} | E_0, E_{\mathcal{S}^*}, Q(\widehat{\mu}_t) = \mu] = Q_L(\mu)\\
\mathbb{P}[s_t^1 \in \{s_{\text{RA}}, s_{\text{RB}}\} | E_0, E_{\mathcal{S}^*}, Q(\widehat{\mu}_t) = \mu] = Q_R(\mu),
\end{aligned}$$

therefore expressing the error component due to $\mathbf{g}$ as $R_t^{1,\mathbf{g}}$ and expressing some repeating conditionals as $\bullet$:

$$\begin{aligned}
\overline{G}_t^\mu :=& \mathbb{E}\left[R_t^{1,\mathbf{g}} \middle| E_0, E_{\mathcal{S}^*}, Q(\widehat{\mu}_t) = \mu, a_t^1 \sim \overline{\pi}(s_t^1), \substack{a_t^i \sim \pi^*(s_t^i),\\ \text{when } i \neq 1}\right]\\
=& \sum_{s \in \mathcal{S}^*} \mathbb{P}[s_t^1 = s | Q(\widehat{\mu}_t) = \mu, \bullet]\, \mathbb{E}[R_t^{1,\mathbf{g}} | s_t^1 = s, Q(\widehat{\mu}_t) = \mu, \bullet]\\
=& \frac{Q_{\max}(\mu)}{Q_{\max}(\mu)} Q_{\max}(\mu) + \frac{Q_{\min}(\mu)}{Q_{\max}(\mu)} Q_{\min}(\mu).
\end{aligned}$$

Similarly, since $\pi^*(a|s) = 1/2$ for any $s \in \mathcal{S}^*$, it holds that

$$\begin{aligned}
G_t^\mu :=& \mathbb{E}\left[R_t^{1,\mathbf{g}} \middle| E_0, E_{\mathcal{S}^*}, Q(\widehat{\mu}_t) = \mu, \substack{a_t^i \sim \pi^*(s_t^i),\\ \forall i}\right]\\
=& \frac{1}{2}\frac{Q_{\min}(\mu)}{Q_{\max}(\mu)} + \frac{1}{2}\frac{Q_{\max}(\mu)}{Q_{\max}(\mu)}.
\end{aligned}$$

Therefore, given the population distribution between $s_{\text{LA}}, s_{\text{LB}}$ and $s_{\text{RA}}, s_{\text{RB}}$, the expected difference in rewards for the two policies is

$$\begin{aligned}
\overline{G}_t^\mu - G_t^\mu =& \left(Q_{\max}(\mu) - \frac{1}{2}\right) + \left(Q_{\min}(\mu) - \frac{1}{2}\right)\frac{Q_{\min}(\mu)}{Q_{\max}(\mu)}\\
=& \left(Q_{\max}(\mu) - \frac{1}{2}\right) + \left(\frac{1}{2} - Q_{\max}(\mu)\right)\frac{Q_{\min}(\mu)}{Q_{\max}(\mu)}\\
=& \left(Q_{\max}(\mu) - \frac{1}{2}\right)\left(1 - \frac{Q_{\min}(\mu)}{Q_{\max}(\mu)}\right)\\
\geq& 2\left(Q_{\max}(\mu) - \frac{1}{2}\right)^2.
\end{aligned}$$

Therefore from above, we conclude that

$$\mathbb{E}[\overline{G}_t^{\widehat{\mu}_t} - G_t^{\widehat{\mu}_t} | E_0] \geq \mathbb{E}[2|X_{\frac{t-1}{2}} - 1/2|^2 | E_0, E_{\mathcal{S}^*}] \geq 2C_1^2 \min\left\{\frac{2^t}{2N_0}, 1\right\}.$$

Using the lower bound above, the conditional expected difference in discounted total reward is

$$\begin{aligned}
&\mathbb{E}\big[\sum_{t=0}^\infty \gamma^t R(s_t^1, a_t^1, \widehat{\mu}_t) | E_0, E_{\mathcal{S}^*}, a_t^1 \sim \overline{\pi}(s_t^1), \substack{a_t^i \sim \pi^*(s_t^i),\\ \text{when } i \neq 1}\big]\\
&\quad - \mathbb{E}\big[\sum_{t=0}^\infty \gamma^t R(s_t^1, a_t^1, \widehat{\mu}_t) | E_0, E_{\mathcal{S}^*}, \substack{a_t^i \sim \pi^*(s_t^i),\\ \forall i}\big]\\
&\geq (1 - \alpha - \beta)\sum_{k=0}^\infty 2C_1^2 \gamma^{2k+1} \min\left\{\frac{2^{2k}}{N_0}, 1\right\} - \frac{2\alpha}{1-\gamma}\\
&\geq \frac{C_2}{N_0}\sum_{k=0}^{\lfloor \log_4 N_0 \rfloor} (4\gamma^2)^k + \frac{C_3}{N_0}\sum_{k=\lfloor \log_4 N_0 \rfloor}^\infty \gamma^{2k} - \frac{2\alpha}{1-\gamma}\\
&\geq \frac{C_4((4\gamma^2)^{\log_4 N_0} - 1)}{N_0} + C_5\frac{(\gamma^2)^{\log_4 N_0} N_0^{-1}}{1-\gamma^2} - \frac{2\alpha}{1-\gamma}\\
&\geq C_6 N_0^{\log_2 \gamma} + C_7\frac{N_0^{\log_2 \gamma - 1}}{1-\gamma} - \frac{2\alpha}{1-\gamma}.
\end{aligned}$$

Taking expectation over N_0 (using $\mathbb{E}[\overline{N}|E^*] = {}^N\!/2$ and Jensen's):

$$\begin{aligned}
&\mathbb{E}\big[\sum_{t=0}^{\infty}\gamma^t R(s_t^1, a_t^1, \widehat{\mu}_t)|E^*, E_{\mathcal{S}^*}, a_t^1 \sim \overline{\pi}(s_t^1), \substack{a_t^i \sim \pi^*(s_t^i),\\ \text{when } i \neq 1}\big] \\
&- \mathbb{E}\big[\sum_{t=0}^{\infty}\gamma^t R(s_t^1, a_t^1, \widehat{\mu}_t)|E^*, E_{\mathcal{S}^*}, \substack{a_t^i \sim \pi^*(s_t^i),\\ \forall i}\big] \\
\geq &C_6 N_0^{\log_2 \gamma} + C_7 \frac{N_0^{\log_2 \gamma - 2}}{1-\gamma} - \frac{2\alpha}{1-\gamma}
\end{aligned}$$

While the analysis above assumes event $E_{\mathcal{S}^*}$, the same analysis lower bound follows with a shift between even and odd steps when $s_0^1 \notin \mathcal{S}^*$, hence

$$\begin{aligned}
&\mathbb{E}\big[\sum_{t=0}^{\infty}\gamma^t R(s_t^1, a_t^1, \widehat{\mu}_t)|E^*, a_t^1 \sim \overline{\pi}(s_t^1), \substack{a_t^i \sim \pi^*(s_t^i),\\ \text{when } i \neq 1}\big] \\
&- \mathbb{E}\big[\sum_{t=0}^{\infty}\gamma^t R(s_t^1, a_t^1, \widehat{\mu}_t)|E^*, \substack{a_t^i \sim \pi^*(s_t^i),\\ \forall i}\big] \\
\geq &C_6 N_0^{\log_2 \gamma} + C_7 \frac{N_0^{\log_2 \gamma - 2}}{1-\gamma} - \frac{2\alpha}{1-\gamma}
\end{aligned}$$

Finally, we conclude the proof with the observation

$$\begin{aligned}
&\max_{\pi} J_{P,R}^{\gamma,N,(1)}(\pi, \boldsymbol{\pi}^*, \ldots, \boldsymbol{\pi}^*) - J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}^*, \boldsymbol{\pi}^*, \ldots, \boldsymbol{\pi}^*) \\
\geq &J_{P,R}^{\gamma,N,(1)}(\overline{\pi}, \boldsymbol{\pi}^*, \ldots, \boldsymbol{\pi}^*) - J_{P,R}^{H,N,(1)}(\boldsymbol{\pi}^*, \boldsymbol{\pi}^*, \ldots, \boldsymbol{\pi}^*) \\
\geq &C_6 N_0^{\log_2 \gamma} + C_7 \frac{N_0^{\log_2 \gamma - 2}}{1-\gamma} - \frac{2\alpha}{1-\gamma} - (1-\gamma)^{-1}\,\mathbb{P}[\overline{E}^c],
\end{aligned}$$

where $\mathbb{P}[\overline{E}^c] = O(e^{-N^2})$ and we pick $\alpha = \mathcal{O}(e^{-N})$.

B. Intractability Results

B.1. Fundamentals of PPAD

We first introduce standard definitions and tools, mostly taken from (Daskalakis et al., 2009; Goldberg, 2011; Papadimitriou, 1994).

Notations. For a finite set Σ, we denote by Σ^n the set of tuples n elements from Σ, and by $\Sigma^* = \bigcup_{n\geq 0}\Sigma^n$ the set of finite sequences of elements of Σ. For any $\alpha \in \Sigma$, let $\alpha^n \in \Sigma^n$ denote the n-tuple $(\underbrace{\alpha, \ldots, \alpha}_{n \text{ times}})$. For $x \in \Sigma^*$, by $|x|$ we denote the length of the sequence x. Finally, the following function will be useful, defined for any $\alpha > 0$:

$$\begin{aligned}
u_\alpha :&\mathbb{R} \to [0, \alpha] \\
u_\alpha(x) :=& \max\{0, \min\{\alpha, x\}\} = \begin{cases} \alpha, \text{ if } x \geq \alpha, \\ x, \text{ if } 0 \leq x \leq \alpha, \\ 0, \text{ if } x \leq 0. \end{cases}
\end{aligned}$$

We define a search problem $\mathcal{S}$ on alphabet Σ as a relation from a set $\mathcal{I}_\mathcal{S} \subset \Sigma^*$ to Σ^* such that for all $x \in \mathcal{I}_\mathcal{S}$, the image of x under $\mathcal{S}$ satisfies $\mathcal{S}_x \subset \Sigma^{|x|^k}$ for some $k \in \mathbb{N}_{>0}$, and given $y \in \Sigma^{|x|^k}$m whether $y \in \mathcal{S}_x$ is decidable in polynomial time.

Intuitively speaking, PPAD is the complexity class of search problems that can be shown to always have a solution using a "parity argument" on a directed graph. The simplest complete example (the example that defines the problem class) of PPAD problems is the computational problem END-OF-THE-LINE. The problem, formally defined below, can be summarized as such: given a directed graph where each node has in-degree and out-degree at most one and given a node that is a source in this graph (i.e., no incoming edge but one outgoing edge), find another node that is a sink or a source. Such a node can be always shown to exist using a simple parity argument.

Definition B.1 (END-OF-THE-LINE (Daskalakis et al., 2009)). The computational problem END-OF-THE-LINE is defined as follows: given two binary circuits S, P each with n input bits and n output bits such that $P(0^n) = 0^n \neq S(s^n)$, find an input $x \in \{0,1\}^n$ such that $P(S(x)) \neq x$ or $S(P(x)) \neq x \neq 0^n$.

The obvious solution to the above is to follow the graph node by node using the given circuits until we reach a sink: however, this can take exponential time as the graph size can be exponential in the bit descriptions of the circuits. It is believed that END-OF-THE-LINE is difficult (Goldberg, 2011), that there is no efficient way to use the bit descriptions of the circuits S, P to find another node with degree 1.

B.2. Proof of Intractability of Stat-MFG

We reduce any ε-GCIRCUIT problem to the problem ε-STATDIST for some simple transition function $P \in \mathcal{P}^{\text{Sim}}$.

Let $(\mathcal{V}, \mathcal{G})$ be a generalized circuit to be reduced to a stable distribution computation problem. Let $V = |\mathcal{V}| \geq 1$. We will define a game that has at most $V + 1$ states and $|\mathcal{A}| = 1$ actions, that is, agent policy will not have significance, and it will suffice to determine simple transition probabilities $P(s'|s,\mu)$ for all $s, s' \in \mathcal{S}, \mu \in \Delta_{\mathcal{S}}$.

The proposed system will have a base state $s_{\text{base}} \in \mathcal{S}$ and 1 additional state s_v associated with the gate whose output is $v \in \mathcal{V}$. Our construction will be sparse: only transition probabilities in between states associated with a gate and s_{base} will take positive values. We define the useful constants $\theta := \frac{1}{8V}, B := \frac{1}{4}$.

Given an (approximately) stable distribution μ^* of P, for each vertex v we will read the satisfying assignment for the ε-GCIRCUIT problem by the value $u_1(\theta^{-1}\mu^*(s_v))$. For each possible gate, we define the following gadgets.

Binary assignment gadget. For a gate of the form $G_{\leftarrow}(\zeta||v)$, we will add one state s_v such that

$$\text{If } \zeta = 1 : \begin{cases} P(s_{\text{base}}|s_v, \mu) = 1, \\ P(s_v|s_v, \mu) = 0, \\ P(s_v|s_{\text{base}}, \mu) = \frac{\theta}{\max\{B, \mu(s_{\text{base}})\}} \end{cases}$$

$$\text{If } \zeta = 0 : \begin{cases} P(s_{\text{base}}|s_v, \mu) = 1, \\ P(s_v|s_v, \mu) = 0, \\ P(s_v|s_{\text{base}}, \mu) = 0 \end{cases}$$

Weighted addition gadget. Next, we implement the addition gadget $G_{\times,+}(\alpha, \beta|v_1, v_2|v)$ for $\alpha, \beta \in [-1, 1]$. In this case, we also add one state s_v to the game, and define the transition probabilities:

$$\begin{aligned} P(s_{\text{base}}|s_v, \mu) &= 1, \\ P(s_v|s_v, \mu) &= 0, \\ P(s_v|s_{\text{base}}, \mu) &= \frac{u_\theta(\alpha u_\theta(\mu(v_1)) + \beta u_\theta(\mu(v_2)))}{\max\{B,\, \mu(s_{\text{base}})\}} \end{aligned}$$

Brittle comparison gadget. For the comparison gate $G_<(|v_1, v_1|v)$, we also add one state s_v to the game. Define the function $p_\delta : [-1, 1] \to [0, 1]$

$$p_\delta(x, y) := u_1\left(\frac{1}{2} + \delta^{-1}(x - y)\right),$$

for any $\delta > 0$. In particular, if $x \geq y + \delta$, then $p_\delta(x, y) = 1$, and if $x \leq y - \delta$, then $p_\delta(x, y) = 0$. We define the probability transitions to and from s_v as

$$\begin{aligned} P(s_v|s_{\text{base}}, \mu) &= \frac{\theta p_{8\varepsilon}(\theta^{-1} u_\theta(\mu(s_1)), \theta^{-1} u_\theta(\mu(s_2)))}{\max\{B,\, \mu(s_{\text{base}})\}}, \\ P(s_v|s_v, \mu) &= 0, \\ P(s_{\text{base}}|s_v, \mu) &= 1. \end{aligned}$$

Finally, after all s_v have been added, we complete the definition of P by setting

$$P(s_{\text{base}}|s_{\text{base}}, \mu) = 1 - \sum_{s' \in \mathcal{S}} P(s'|s_{\text{base}}, \mu).$$

We first verify that the above assignment is a valid transition probability matrix for any $\mu \in \Delta_{\mathcal{S}}$. It is clear from definitions that for any $\mu, s \neq s_{\text{base}}$, $P(\cdot|s, \mu)$ is a valid probability distribution as long as $8\varepsilon < 1$. Moreover, for any $s \neq s_{\text{base}}$, it holds that $0 \leq P(s|s_{\text{base}}, \mu) \leq \frac{\theta}{B} < 1$, and it also holds that

$$P(s_{\text{base}}|s_{\text{base}}, \mu) = 1 - \sum_{s' \in \mathcal{S}} P(s'|s_{\text{base}}, \mu) \geq 1 - \frac{V\theta}{B} \geq 0$$

so $P(\cdot|s_{\text{base}}, \mu)$ is a valid probability transition matrix. Finally, the defined transition probability function P is Lipschitz in the components of μ, and P can be defined as a composition of simple functions, hence $P \in \mathcal{P}^{\text{Sim}}$. Finally, in this defined MFG, it holds that $V + 1 = |\mathcal{S}|$, since for each gate in the generalized circuit we defined one additional state.

Error propagation. We finally analyze the error propagation of the stationary distribution problem in terms of the generalized circuit. Without loss of generality we assume $\varepsilon < \frac{1}{8}$. First, for any solution of the ε-StatDist problem μ^*, whenever $\varepsilon < \frac{1}{8}$, it must hold that:

$$\left| \mu^*(s_{\text{base}}) - \sum_{s' \in \mathcal{S}} \mu^*(s) P(s_{\text{base}}|s, \mu^*) \right| \leq \frac{1}{8|\mathcal{S}|},$$

hence (using $V < |\mathcal{S}|$) we have the lower bound on $\mu^*(s_{\text{base}})$ given by:

$$\begin{aligned} \mu^*(s_{\text{base}}) \geq & \sum_{s \in \mathcal{S}} \mu^*(s) P(s_{\text{base}}|s, \mu^*) - \frac{1}{8V} \\ \geq & \mu^*(s_{\text{base}}) P(s_{\text{base}}|s_{\text{base}}, \mu^*) + \sum_{s \neq s_{\text{base}}} \mu^*(s) P(s_{\text{base}}|s, \mu^*) - \frac{1}{8V} \\ \geq & \mu^*(s_{\text{base}}) \left(1 - \frac{V\theta}{B}\right) + \sum_{s \neq s_{\text{base}}} \mu^*(s) - \frac{1}{8V} \\ \geq & \mu^*(s_{\text{base}}) \left(1 - \frac{V\theta}{B}\right) + (1 - \mu^*(s_{\text{base}})) - \frac{1}{8V} \\ \implies & \mu^*(s_{\text{base}}) \geq \frac{1 - \frac{1}{8V}}{1 + \frac{V\theta}{B}} \geq B = \frac{1}{4}. \end{aligned}$$

We will show that a solution of the ε-StatDist can be converted into a ε'-satisfying assignment

$$v \to u_1\left(\frac{\mu^*(s_v)}{\theta}\right),$$

for some appropriate ε' to be defined later.

Case 1: Binary assignment error. First, assume $G_{\leftarrow}(\zeta||v) \in \mathcal{G}$ If $\zeta = 1$, since μ^* is a ε stable distribution we have

$$
\begin{aligned}
|\mu^*(s_v) - \mu^*(s_{\text{base}})P(s_v|s_{\text{base}}, \mu^*)| &\leq \frac{\varepsilon}{|\mathcal{S}|} \\
\left|\mu^*(s_v) - \mu^*(s_{\text{base}})\frac{\theta}{\max\{B, \mu^*(s_{\text{base}})\}}\right| &\leq \frac{\varepsilon}{|\mathcal{S}|} \\
|\mu^*(s_v) - \theta| &\leq \frac{\varepsilon}{|\mathcal{S}|} \\
\left|\frac{\mu^*(s_v)}{\theta} - 1\right| &\leq \frac{\varepsilon}{\theta|\mathcal{S}|} \leq \frac{\varepsilon}{\theta V} \leq 8\varepsilon,
\end{aligned}
$$

where we used the fact that $\frac{\theta}{\max\{B,\mu^*(s_{\text{base}})\}} = \mu^*(s_{\text{base}})$. and it follows by definition that $|u_1\left(\frac{\mu^*(s_v)}{\theta}\right) - 1| \leq 8\varepsilon$, since the map u_1 is 1-Lipschitz and therefore can only decrease the absolute value on the left. Likewise, if $\zeta = 0$,

$$
\begin{aligned}
|\mu^*(s_v) - \sum_{s\in\mathcal{S}} \mu^*(s)P(s_v|s, \mu^*)| &\leq \frac{\varepsilon}{|\mathcal{S}|} \\
|\mu^*(s_v)| &\leq \frac{\varepsilon}{|\mathcal{S}|} \\
\left|\frac{\mu^*(s_v)}{\theta}\right| &\leq \frac{\varepsilon}{\theta|\mathcal{S}|} \leq 8\varepsilon
\end{aligned}
$$

and once again $u_1\left(\frac{\mu^*(s_v)}{\theta}\right) \leq 8\varepsilon$.

Case 2: Weighted addition error. Assume that $G_{\times,+}(\alpha, \beta|v_1, v_2|v) \in \mathcal{G}$, and set $\square := u_\theta(\alpha u_\theta(\mu(v_1)) + \beta u_\theta(\mu(v_2)))$. Using the fact that $\|\mu^* - \Gamma_P(\mu^*)\| \leq \frac{\varepsilon}{|\mathcal{S}|}$,

$$
\begin{aligned}
|\mu^*(s_v) - \sum_{s\in\mathcal{S}} \mu^*(s)P(s_v|s, \mu^*)| &\leq \frac{\varepsilon}{|\mathcal{S}|}, \\
\left|\mu^*(s_v) - \mu^*(s_{\text{base}})\frac{u_\theta(\alpha u_\theta(\mu(v_1)) + \beta u_\theta(\mu(v_2)))}{\max\{B,\, \mu(s_{\text{base}})\}}\right| &\leq \frac{\varepsilon}{|\mathcal{S}|}, \\
\left|\frac{\mu^*(s_v)}{\theta} - \frac{\square}{\theta}\right| &\leq \frac{\varepsilon}{|\mathcal{S}|\theta},
\end{aligned}
$$

which implies

$$
\left|u_1\left(\frac{\mu^*(s_v)}{\theta}\right) - u_1\left(\alpha u_1\left(\frac{\mu^*(v_1)}{\theta}\right) + \beta u_1\left(\frac{\mu^*(v_2)}{\theta}\right)\right)\right| \leq 8\varepsilon.
$$

Case 3: Brittle comparison gadget. Finally, we analyze the more involved case of the comparison gadget. Assume $G_<(|v_1, v_2|v) \in \mathcal{G}$. The stability conditions for s_v yield:

$$
\begin{aligned}
|\mu^*(s_v) - \mu^*(s_{\text{base}})P(s_v|s_{\text{base}}, \mu^*)| \leq& \frac{\varepsilon}{|\mathcal{S}|} \\
|\mu^*(s_v) - \theta p_{8\varepsilon}(\theta^{-1}u_\theta(\mu^*(v_1)), \theta^{-1}u_\theta(\mu^*(v_2)))| \leq& \frac{\varepsilon}{|\mathcal{S}|}
\end{aligned}
$$

We analyze two cases: $u_1(\theta^{-1}\mu^*(v_1)) \geq u_1(\theta^{-1}\mu^*(v_2)) + 8\varepsilon$ and $u_1(\theta^{-1}\mu^*(v_1)) \leq u_1(\theta^{-1}\mu^*(v_2)) - 8\varepsilon$. In the first case, we obtain

$$
\theta^{-1}u_\theta(\mu^*(v_1)) \geq \theta^{-1}u_\theta(\mu^*(v_2)) + 8\varepsilon,
$$

which implies by the definition of $p_{8\varepsilon}$

$$
\begin{aligned}
|\mu^*(s_v) - \theta| \leq& \frac{\varepsilon}{|\mathcal{S}|} \\
|u_1(\theta^{-1}\mu^*(s_v)) - 1| \leq& \frac{\varepsilon}{|\mathcal{S}|\theta} \\
u_1(\theta^{-1}\mu^*(s_v)) \geq& 1 - \frac{\varepsilon}{|\mathcal{S}|\theta} \geq 1 - 8\varepsilon.
\end{aligned}
$$

In the second case $u_1(\theta^{-1}\mu^*(v_1)) \leq u_1(\theta^{-1}\mu^*(v_2)) - 8\varepsilon$, it follows by a similar analysis that

$$u_1(\theta^{-1}\mu^*(s_v)) \leq \frac{\varepsilon}{|\mathcal{S}|\theta} \leq 8\varepsilon.$$

Hence, in the above, we reduced the 8ε-GCIRCUIT problem to the ε-STATDIST problem, completing the proof that ε-STATDIST is PPAD-hard. The fact that ε-STATDIST is in PPAD on the other hand easily follows from the fact that ε-STATDIST is the fixed point problem for the (simple) operator Γ_P, reducing it to the END-OF-THE-LINE problem by a standard construction (Daskalakis et al., 2009).

B.3. Proof of Intractability of FH-MFG

As in the previous section, we reduce any ε-GCIRCUIT problem $(\mathcal{G}, \mathcal{V})$ to the problem $(\varepsilon^2, 2)$-FH-NASH for some simple reward $R \in \mathcal{R}^{\text{Sim}}$. Once again let $V = |\mathcal{V}|$.

Associated with each $v \in \mathcal{V}$ we define $s_{v,1}, s_{v,0}, s_{v,\text{base}} \in \mathcal{S}$. The initial distribution is defined as

$$\mu_0(s_{v,\text{base}}) = \frac{1}{V}, \forall v \in \mathcal{V},$$

and we define two actions for each state: $\mathcal{A} = \{a_1, a_0\}$. The state transition probability matrix is given by

$$P(s|s_{v,\text{base}}, a) = \begin{cases} 1, & \text{if } a = a_1, s = s_{v,1}, \\ 1, & \text{if } a = a_0, s = s_{v,0}, \\ 0, & \text{otherwise.} \end{cases}$$
$$P(s_{v,\text{base}}|s, a) = 0, \forall v \in \mathcal{V}, s \in \mathcal{S}, a \in \mathcal{A},$$

and an ε satisfying assignment $p : \mathcal{V} \to [0, 1]$ will be read by $p(v) = \pi_1^*(a_1|s_{v,\text{base}})$ for the optimal policy $\boldsymbol{\pi}^* = \{\pi_h\}_{h=0}^1$. We will specify population-dependent rewards $R \in \mathcal{R}^{\text{Simple}}$, since R will not depend on the particular action but only the state and population distribution, we will concisely denote $R(s, a, \mu) = R(s, \mu)$. It will be the case that

$$R(s_{v,\text{base}}, \mu) = 0, \forall v \in \mathcal{V}, \mu \in \Delta_{\mathcal{S}}.$$

We assign $R(s_{v,1}, \mu) = R(s_{v,0}, \mu) = 0, \forall \mu$ for any vertex v of the generalized circuit that is not the output of any gate in $\mathcal{G}$.

Binary assignment gadget. For any binary assignment gate $G_{\leftarrow}(\zeta||v)$, we assign

$$R(s_{v,1}, \mu) = \zeta,$$
$$R(s_{v,0}, \mu) = 1 - \zeta, \forall \mu \in \Delta_{\mathcal{S}}.$$

Weighted addition gadget. For any gate $G_{\times,+}(\alpha, \beta|v_1, v_2|v)$,

$$R(s_{v,1}, \mu) = u_1(u_1(\alpha V\mu(s_{v_1,1}) + \beta V\mu(s_{v_2,1})) - V\mu(s_{v,1})),$$
$$R(s_{v,0}, \mu) = u_1(V\mu(s_{v,1}) - u_1(\alpha V\mu(s_{v_1,1}) + \beta V\mu(s_{v_2,1}))),$$

for all $\mu \in \Delta_{\mathcal{S}}$.

Brittle comparison gadget. For any gate $G_{<}(|v_1, v_2|v)$, we define the rewards for states $s_{v,1}, s_{v,0}$ as

$$R(s_{v,1}, \mu) = u_1(V\mu(s_{v_2,1}) - V\mu(s_{v_1,1})),$$
$$R(s_{v,0}, \mu) = u_1(V\mu(s_{v_1,1}) - V\mu(s_{v_2,1})), \forall \mu \in \Delta_{\mathcal{S}}.$$

Now assume that $\boldsymbol{\pi}^* = \{\pi_h^*\}_{h=0}^1$ is a solution to the $(\varepsilon^2, 2)$-FH-NASH problem and $\boldsymbol{\mu}^* = \Lambda_{P,\mu_0}^2(\boldsymbol{\pi}^*)$, that is, assume that for all $\boldsymbol{\pi} \in \Pi^2$,

$$V_{P,R}^H(\boldsymbol{\mu}^*, \boldsymbol{\pi}) - V_{P,R}^H(\boldsymbol{\mu}^*, \boldsymbol{\pi}^*) \leq \frac{\varepsilon^2}{V}.$$

Firstly, if μ_1^* is induced by $\boldsymbol{\pi}^*$, it holds that $\forall v \in \mathcal{V}$,

$$\mu_1^*(s_{v,\text{base}}) = 0, \quad \mu_1^*(s_{v,1}) = \frac{1}{V}\pi_0^*(s_{v,1}|s_{v,\text{base}}),$$
$$\mu_1^*(s_{v,0}) = \frac{1 - \pi_0^*(s_{v,1}|s_{v,\text{base}})}{V}.$$

Furthermore, a policy $\boldsymbol{\pi}^{\text{br}} \in \Pi_2$ that is the best response to $\boldsymbol{\mu}^* := \{\mu_0^*, \mu_1^*\}$ can be always formulated as:

$$\begin{aligned}
\pi_0^{\text{br}}(a_1|s_{v,\text{base}}) &= \begin{cases} 1, \text{ if } R(s_{v,1}, \mu_1^*) > R(s_{v,1}, \mu_1^*), \\ 0, \text{otherwise} \end{cases} \\
\pi_0^{\text{br}}(a_0|s_{v,\text{base}}) &= 1 - \pi_0^{\text{br}}(a_1|s_{v,\text{base}}), \\
\pi_1^{\text{br}}(a_1|s_{v,\text{base}}) &= 1, \\
\pi_1^{\text{br}}(a_0|s_{v,\text{base}}) &= 0.
\end{aligned}$$

By the optimality conditions, we will have

$$V_{P,R}^H(\boldsymbol{\mu}^*, \boldsymbol{\pi}^{\text{br}}) - V_{P,R}^H(\boldsymbol{\mu}^*, \boldsymbol{\pi}^*) \leq \frac{\varepsilon^2}{V}.$$

Furthermore, for any $v \in \mathcal{V}$ it holds that

$$\begin{aligned}
&V_{P,R}^H(\boldsymbol{\mu}^*, \boldsymbol{\pi}^{\text{br}}) - V_{P,R}^H(\boldsymbol{\mu}^*, \boldsymbol{\pi}^*) \\
&= \sum_{v \in \mathcal{V}} \mu_0(s_{v,\text{base}})[\max_{s \in \{s_{v,1}, s_{v,0}\}} R(s, \mu_1^*) \\
&\qquad - \pi_0^*(a_1|s_{v,\text{base}}) R(s_{v,1}, \mu_1^*) - \pi_0^*(a_0|s_{v,\text{base}}) R(s_{v,0}, \mu_1^*)] \\
\geq& \frac{1}{V} \max_{s \in \{s_{v,1}, s_{v,0}\}} R(s, \mu_1^*) \\
&- \frac{1}{V}\pi_0^*(a_1|s_{v,\text{base}}) R(s_{v,1}, \mu_1^*) - \frac{1}{V}\pi_0^*(a_0|s_{v,\text{base}}) R(s_{v,0}, \mu_1^*)
\end{aligned}$$

as the summands are all positive. We prove that all gate conditions are satisfied case by base. Without loss of generality, we assume $\varepsilon < 1$ below.

Case 1. It follows that for any $v \in \mathcal{V}$ such that $G_{\leftarrow}(\zeta||v) \in \mathcal{G}$, we have

$$\begin{aligned}
&\frac{1}{V} - \frac{1}{V}\pi_0^*(a_1|s_{v,\text{base}})\zeta - \frac{1}{V}\pi_0^*(a_0|s_{v,\text{base}})(1-\zeta) \leq \frac{\varepsilon^2}{V} \\
&1 - \pi_0^*(a_1|s_{v,\text{base}})\zeta - (1 - \pi_0^*(a_1|s_{v,\text{base}}))(1-\zeta) \leq \varepsilon^2 \\
&\zeta(1 - 2\pi_0^*(a_1|s_{v,\text{base}})) + \pi_0^*(a_1|s_{v,\text{base}}) \leq \varepsilon^2 \leq \varepsilon.
\end{aligned}$$

The above implies $\pi_0^*(a_1|s_{v,\text{base}}) \geq 1 - \varepsilon$ if $\zeta = 1$, and if $\zeta = 0$, it implies $\pi_0^*(a_1|s_{v,\text{base}}) \leq \varepsilon$.

Case 2. For any $v \in \mathcal{V}$ such that $G_{\times,+}(\alpha, \beta|v_1, v_2|v) \in \mathcal{G}$, denoting in short

$$\begin{aligned}
\square &:= u_1(\alpha V \mu_1^*(s_{v_1,1}) + \beta V \mu_1^*(s_{v_2,1})) \\
&= u_1(\alpha \pi_0^*(a_1|s_{v_1,1}) + \beta \pi_0^*(a_1|s_{v_2,1})), \\
p_1 &:= \pi_0^*(a_1|s_{v,\text{base}}) \\
p_0 &:= \pi_0^*(a_0|s_{v,\text{base}})
\end{aligned}$$

we have

$$\begin{aligned}
&\frac{1}{V}\max\left\{u_1(V\mu_1^*(s_{v,1}) - \square), u_1(\square - V\mu_1^*(s_{v,1}))\right\} \\
&\quad - \frac{1}{V}\pi_0^*(a_1|s_{v,\text{base}}) u_1(\square - V\mu_1^*(s_{v,1})) \\
&\quad - \frac{1}{V}\pi_0^*(a_0|s_{v,\text{base}}) u_1(V\mu_1^*(s_{v,1}) - \square) \leq \varepsilon^2,
\end{aligned}$$

or equivalently

$$\max\left\{u_1(p_1-\square), u_1(\square-p_1)\right\} - p_1 u_1(\square - p_1) - p_0 u_1(p_1-\square) \le \varepsilon^2.$$

First, assume it holds that $p_1 \le \square$, then:

$$\begin{aligned} u_1(\square - p_1) - p_1 u_1(\square - p_1) &\le \varepsilon^2 \\ (1-p_1)(\square - p_1) &\le \varepsilon^2. \end{aligned}$$

The above implies that either $p_1 \ge 1-\varepsilon$ or $u_1(\square - p_1) \le \varepsilon$, both cases implying $|\square - p_1| \le \varepsilon$ since we assume $\square \ge p_1$. To conclude case 2, assume that $\square < p_1$, then

$$\begin{aligned} u_1(p_1-\square) - (1-p_1)u_1(p_1-\square) &\le \varepsilon^2, \\ p_1(p_1-\square) &\le \varepsilon^2, \end{aligned}$$

then either $p_1 \le \varepsilon$ or $p_1 - \square \le \varepsilon$, either case implying once again $|\square - p_1| \le \varepsilon$.

Case 3. Finally, for any $v \in \mathcal{V}$ such that $G_<(|v_1, v_2|v) \in \mathcal{G}$,

$$\begin{aligned} &\frac{1}{V}\max\left\{u_1(\mu(s_{v_2,1}) - \mu(s_{v_1,1})), u_1(\mu(s_{v_1,1}) - \mu(s_{v_2,1}))\right\} \\ &\quad - \frac{1}{V}\pi_0^*(a_1|s_{v,\text{base}})u_1(\mu(s_{v_1,1}) - \mu(s_{v_2,1})) \\ &\quad - \frac{1}{V}\pi_0^*(a_0|s_{v,\text{base}})u_1(\mu(s_{v_2,1}) - \mu(s_{v_1,1})) \le \varepsilon \end{aligned}$$

hence once again using the shorthand notation:

$$\begin{aligned} &\triangle := V\mu_1^*(s_{v_2,1}) - V\mu_1^*(s_{v_1,1}) = \pi_0^*(a_1|s_{v_2,1}) - \pi_0^*(a_1|s_{v_1,1}) \\ &p_1 := \pi_0^*(a_1|s_{v,\text{base}}) \\ &p_0 := \pi_0^*(a_0|s_{v,\text{base}}) \end{aligned}$$

we have the inequality:

$$\begin{aligned} u_1(|\triangle|) - p_1 u_1(\triangle) - p_0 u_1(-\triangle) &\le \varepsilon^2 \\ u_1(|\triangle|) - p_1 u_1(\triangle) - (1-p_1)u_1(-\triangle) &\le \varepsilon^2. \end{aligned}$$

First assume $\triangle \ge \varepsilon$, then

$$u_1(\triangle)(1-p_1) \le \varepsilon^2 \implies 1-\varepsilon \le p_1,$$

and conversely if $\triangle \le -\varepsilon$,

$$u_1(-\triangle)p_1 \le \varepsilon^2 \implies p_1 \le \varepsilon,$$

concluding that the comparison gate conditions are ε satisfied for the assignment $v \to \pi_0^{\text{br}}(a_1|s_{v,\text{base}})$.

The three cases above conclude that $v \to \pi_0^{\text{br}}(a_1|s_{v,\text{base}})$ is an ε-satisfying assignment for the generalized circuit $(\mathcal{V}, \mathcal{G})$, concluding the proof that $(\varepsilon_0, 2)$-FH-NASH is PPAD-hard for some $\varepsilon_0 > 0$. The fact that $(\varepsilon_0, 2)$-FH-NASH is in PPAD follows from the fact that the NE is a fixed point of a simple map on space Π_2, see for instance (Huang et al., 2023).

B.4. Proof of Intractability of 2-FH-LINEAR

Our reduction will be similar to the previous section, however, instead of reducing a ε-GCIRCUIT to an MFG, we will reduce a 2 player general sum normal form game, 2-NASH, to a finite horizon mean field game with linear rewards with horizon $H = 2$ (2-FH-LINEAR). Let $\varepsilon > 0, K_1, K_2 \in \mathbb{N}_{>0}, A, B \in \mathbb{R}^{K_1, K_2}$ be given for a 2-NASH problem. We assume without loss of generality that $K_1 > 1$, as otherwise, the solution of 2-NASH is trivial.

This time, we define finite horizon game with $K_1 + K_2 + 2$ states, denoted $\mathcal{S} := \{s^1_{\text{base}}, s^2_{\text{base}}, s^1_1, \ldots, s^1_{K_1}, s^2_1, \ldots, s^2_{K_2}\}$. Without loss of generality, we can assume $K_1 \leq K_2$. The action set will be defined by $\mathcal{A} = [K_2] = \{1, \ldots, K_2\}$. The initial state distribution will be given by $\mu_0(s^1_{\text{base}}) = \mu_0(s^2_{\text{base}}) = 1/2$, with $\mu_0(s) = 0$ for all other states. We define the transitions for any $s \in \mathcal{S}, a, a' \in \mathcal{A}$ as:

$$P(s|s^1_{\text{base}}, a) = \begin{cases} 1, \text{ if } s = s^1_a \text{ and } a \leq K_1, \\ 1, \text{ if } s = s^1_a \text{ and } a > K_1, \\ 0, \text{otherwise.} \end{cases}$$

$$P(s|s^2_{\text{base}}, a) = \begin{cases} 1, \text{ if } s = s^2_a, \\ 0, \text{otherwise.} \end{cases}$$

$$P(s|s^1_a, a') = \begin{cases} 1, \text{ if } s = s^1_a, \\ 0, \text{otherwise.} \end{cases} \qquad P(s|s^2_a, a') = \begin{cases} 1, \text{ if } s = s^2_a, \\ 0, \text{otherwise.} \end{cases}$$

Finally, we will define the linear reward function as for all $a \in [K_2]$:

$$\begin{aligned} R(s^1_{\text{base}}, a, \mu) &= 0, \\ R(s^2_{\text{base}}, a, \mu) &= 0, \\ R(s^1_a, a, \mu) &= \begin{cases} 0, \text{if } a > K_1, \\ \frac{1}{2} + \frac{1}{2}\sum_{a' \in [K_2]} \mu(s^2_{a'}) A_{a,a'} \end{cases} \\ R(s^2_a, a, \mu) &= \frac{1}{2} + \frac{1}{2} \sum_{a' \in [K_1]} \mu(s^1_{a'}) B_{a',a}. \end{aligned}$$

In words, the states $s^1_{\text{base}}, s^2_{\text{base}}$ represent the two players of the 2-NASH, and an agent starting from one of the initial base states $s^1_{\text{base}}, s^2_{\text{base}}$ of the FH-MFG at round $h = 0$ will be placed at $h = 1$ at a state representing the (pure) strategies of each player respectively.

Given the game description above, assume $\boldsymbol{\pi}^* = \{\pi^*_h\}^1_{h=0}$ is an ε solution of the 2-FH-LINEAR. Then, it holds for the induced distribution $\boldsymbol{\mu}^* := \{\mu^*_h\}^1_{h=0} = \Lambda^H_P$ that:

$$\begin{aligned} \mu^*_0 &= \mu_0, \\ \mu^*_1(s) &= \sum_{s', a' \in \mathcal{S} \times \mathcal{A}} \mu_0(s') \pi^*(a'|s') P(s|s', a') \\ &= \begin{cases} \frac{1}{2}\pi_0(i|s^1_{\text{base}}), \text{ if } s = s^1_i, \text{ for some } i \in [K_1], \\ \frac{1}{2}\pi_0(i|s^2_{\text{base}}), \text{ if } s = s^2_i, \text{ for some } i \in [K_2], \\ \frac{1}{2} - \frac{1}{2}\sum_{i \in [K_1]} \pi_0(i|s^1_{\text{base}}), \text{ if } s = s^1_{\text{base}}, \\ 0, \text{otherwise.} \end{cases} \end{aligned}$$

By definition of the ε finite horizon Nash equilibrium,

$$\mathcal{E}^H_{P,R}(\boldsymbol{\pi}^*) := \max_{\boldsymbol{\pi}' \in \Pi^H} V^H_{P,R}(\Lambda^H_P(\boldsymbol{\pi}^*), \boldsymbol{\pi}') - V^H_{P,R}(\Lambda^H_P(\boldsymbol{\pi}^*), \boldsymbol{\pi}) \leq \varepsilon,$$

in particular, it holds for any $\boldsymbol{\pi} \in \Pi_2$ that

$$V^H_{P,R}(\boldsymbol{\mu}^*, \boldsymbol{\pi}) - V^H_{P,R}(\boldsymbol{\mu}^*, \boldsymbol{\pi}^*) \leq \varepsilon. \tag{12}$$

By direct computation, the value functions $V^H_{P,R}$ can be written directly in this case for any π:

$$\begin{aligned}
V^H_{P,R}(\boldsymbol{\mu}^*, \boldsymbol{\pi}) =& \frac{1}{2} \sum_{a \in [K_1]} \pi_0(a|s^1_{\text{base}}) \left(\frac{1}{2} + \frac{1}{2} \sum_{a' \in [K_2]} \mu^*_1(s^2_{a'}) A_{a,a'} \right) \\
&+ \frac{1}{2} \sum_{a' \in [K_2]} \pi_0(a'|s^2_{\text{base}}) \left(\frac{1}{2} + \frac{1}{2} \sum_{a \in [K_1]} \mu^*_1(s^1_{a}) B_{a,a'} \right) \\
=& \frac{1}{4} \left(1 + \sum_{a \in [K_1]} \pi_0(a|s^1_{\text{base}}) \right) \\
&+ \frac{1}{8} \sum_{a \in [K_1]} \sum_{a' \in [K_2]} \pi_0(a|s^1_{\text{base}}) \pi^*_0(a'|s^2_{\text{base}}) A_{a,a'} \\
&+ \frac{1}{8} \sum_{a \in [K_1]} \sum_{a' \in [K_2]} \pi_0(a'|s^2_{\text{base}}) \pi^*_0(a|s^1_{\text{base}}) B_{a,a'}
\end{aligned}$$

We analyze two different cases, accounting for a possible imbalance between the strategy spaces of the two players, $[K_1]$ and $[K_2]$.

Case 1. Assume $K_1 = K_2$. Then, $V^H_{P,R}(\boldsymbol{\mu}^*, \boldsymbol{\pi})$ simplifies to

$$\begin{aligned}
V^H_{P,R}(\boldsymbol{\mu}^*, \boldsymbol{\pi}) =& \frac{1}{2} + \frac{1}{8} \sum_{a \in [K_1]} \sum_{a' \in [K_2]} \pi_0(a|s^1_{\text{base}}) \pi^*_0(a'|s^2_{\text{base}}) A_{a,a'} \\
&+ \frac{1}{8} \sum_{a \in [K_1]} \sum_{a' \in [K_2]} \pi_0(a'|s^2_{\text{base}}) \pi^*_0(a|s^1_{\text{base}}) B_{a,a'}.
\end{aligned} \tag{13}$$

Take an arbitrary mixed strategy $\sigma_1 \in \Delta_{[K_1]}$ and define the policy $\boldsymbol{\pi}_A = \{\pi_{A,h}\}^1_{h=0} \in \Pi^2$ so that

$$\pi_{A,0}(s^1_{\text{base}}) = \sigma_1, \quad \pi_{A,0}(s^2_{\text{base}}) = \pi^*_0(s^2_{\text{base}}), \quad \pi_{A,1} = \pi^*_1.$$

Then, placing $\boldsymbol{\pi}_A$ in equations (13) and (12), it follows that

$$\begin{aligned}
&\sum_{a \in [K_1]} \sum_{a' \in [K_2]} \sigma_1(a) \pi^*_0(a'|s^2_{\text{base}}) A_{a,a'} \\
&\quad - \sum_{a \in [K_1]} \sum_{a' \in [K_2]} \pi^*_0(a|s^1_{\text{base}}) \pi^*_0(a'|s^2_{\text{base}}) A_{a,a'} \leq 8\varepsilon.
\end{aligned} \tag{14}$$

Similarly, for any $\sigma_2 \in \Delta_[K_2]$, replacing $\boldsymbol{\pi}$ in equations (13) and (12) with a policy $\boldsymbol{\pi}_B$ such that

$$\pi_{B,0}(s^1_{\text{base}}) = \pi^*_0(s^1_{\text{base}}), \quad \pi_{B,0}(s^2_{\text{base}}) = \sigma_2, \quad \pi_{B,1} = \pi^*_1,$$

we obtain

$$\begin{aligned}
&\sum_{a \in [K_1]} \sum_{a' \in [K_2]} \sigma_2(a) \pi^*_0(a'|s^1_{\text{base}}) B_{a,a'} \\
&\quad - \sum_{a \in [K_1]} \sum_{a' \in [K_2]} \pi^*_0(a'|s^2_{\text{base}}) \pi^*_0(a|s^1_{\text{base}}) B_{a,a'} \leq 8\varepsilon.
\end{aligned} \tag{15}$$

Hence, the resulting equations (14), (15) imply that in this case the strategy profile $(\pi^*_0(s^1_{\text{base}}), \pi^*_0(s^2_{\text{base}}))$ is a 8ε-Nash equilibrium for the normal form game defined by matrices A, B.

Case 2. Next, we analyze the case when $1 < K_1 < K_2$. If $\sum_{a' \in [K_1]} \pi^*_0(a'|s^1_{\text{base}}) = 0$, then the policy

$$\pi'_0(1|s^1_{\text{base}}) = 1, \quad \pi'_0(s^2_{\text{base}}) = \pi^*_0(s^2_{\text{base}}), \quad \pi'_1 = \pi^*_1.$$

yields an exploitability of at least $1/4$, so by taking ε smaller than $1/4$ we can discard this possibility.

Otherwise, we define a policy $\boldsymbol{\pi}_C = \{\pi_{C,h}\}_{h=0}^{1} \in \Pi^2$ such that

$$\pi_{C,0}(a|s^1_{\text{base}}) = \begin{cases} \frac{\pi^*_0(a|s^1_{\text{base}})}{\sum_{a' \in [K_1]} \pi^*_0(a'|s^1_{\text{base}})}, \text{ if } a \in [K_1], \\ 0, \text{otherwise.} \end{cases}$$
$$\pi_{C,0}(s^2_{\text{base}}) = \pi^*_0(s^2_{\text{base}}), \quad \pi_{C,1} = \pi^*_1,$$

and replace $\boldsymbol{\pi}$ in Equation (12) with $\boldsymbol{\pi}_C$ to obtain:

$$\begin{aligned}&\frac{1}{4} - \frac{1}{4}S \\ &+ \frac{1}{8}\left(S^{-1} - 1\right) \sum_{a \in [K_1]} \sum_{a' \in [K_2]} \pi^*_0(a|s^1_{\text{base}}) \pi^*_0(a'|s^2_{\text{base}}) A_{a,a'} \leq \varepsilon\end{aligned}$$

where $S := \sum_{a' \in [K_1]} \pi^*_0(a'|s^1_{\text{base}}) < 1$, hence

$$1 - S = \sum_{a' \in [K_2] - [K_1]} \pi^*_0(a'|s^1_{\text{base}}) \leq 4\varepsilon.$$

Now for some $\sigma_1 \in \Delta_{[K_1]}$, once again take the policy $\boldsymbol{\pi}_A$ defined in Case 1, and use Inequality (12) to obtain:

$$\begin{aligned}\frac{1}{4}(1 - S) + \frac{1}{8} &\sum_{a \in [K_1]} \sum_{a' \in [K_2]} \sigma_1(a) \pi^*_0(a'|s^2_{\text{base}}) A_{a,a'} \\ &- \frac{1}{8} \sum_{a \in [K_2]} \sum_{a' \in [K_2]} \pi^*_0(a|s^1_{\text{base}}) \pi^*_0(a'|s^2_{\text{base}}) A_{a,a'} \leq \varepsilon \\ &\sum_{a \in [K_1]} \sum_{a' \in [K_2]} \sigma_1(a) \pi^*_0(a'|s^2_{\text{base}}) A_{a,a'} \\ &- \sum_{a \in [K_1]} \sum_{a' \in [K_2]} \pi^*_0(a|s^1_{\text{base}}) \pi^*_0(a'|s^2_{\text{base}}) A_{a,a'} \leq 8\varepsilon.\end{aligned}$$

Here, using the definition of $\boldsymbol{\pi}_C$, as $\pi_{C,0}(a|s^1_{\text{base}}) \geq \pi^*_0(a|s^1_{\text{base}})$ for $a \in [K_1]$, we obtain:

$$\begin{aligned}&\sum_{a \in [K_1]} \sum_{a' \in [K_2]} \sigma_1(a) \pi_{C,0}(a'|s^2_{\text{base}}) A_{a,a'} \\ &- \sum_{a \in [K_1]} \sum_{a' \in [K_2]} \pi_{C,0}(a|s^1_{\text{base}}) \pi_{C,0}(a'|s^2_{\text{base}}) A_{a,a'} \leq 8\varepsilon.\end{aligned}$$

Next take $\boldsymbol{\pi}_B$ as defined above in Case 1 for any arbitrary $\sigma_2 \in \Delta_{[K_2]}$ and use the Inequality 12:

$$\begin{aligned}&\sum_{a' \in [K_2]} \sum_{a \in [K_1]} \sigma_2(a') \pi^*_0(a|s^1_{\text{base}}) B_{a,a'} \\ &\quad - \sum_{a \in [K_1]} \sum_{a' \in [K_2]} \pi^*_0(a|s^1_{\text{base}}) \pi^*_0(a'|s^2_{\text{base}}) B_{a,a'} \leq 8\varepsilon \\ &\sum_{a \in [K_1]} \sum_{a' \in [K_2]} \sigma_2(a') \pi_{C,0}(a|s^1_{\text{base}}) B_{a,a'} \\ &\quad - \sum_{a \in [K_1]} \sum_{a' \in [K_2]} \pi_{C,0}(a|s^1_{\text{base}}) \pi_{C,0}(a'|s^2_{\text{base}}) B_{a,a'} \leq \frac{8\varepsilon}{S} \leq \frac{8\varepsilon}{1 - 4\varepsilon}.\end{aligned}$$

Assuming without loss of generality that $\varepsilon < \frac{1}{8}$, it follows that $\pi_{C,0}(s^1_{\text{base}}), \pi_{C,0}(s^2_{\text{base}})$ is a 16ε solution to the 2-Nash.