Exploiting Approximate Symmetry for Efficient Multi-Agent Reinforcement Learning

Batuhan Yardim
Department of Computer Science
ETH Zürich
ali-batuhan.yardim@inf.ethz.ch

Niao He
Department of Computer Science
ETH Zürich
niao.he@inf.ethz.ch

Abstract

Mean-field games (MFG) have become significant tools for solving large-scale multi-agent reinforcement learning problems under symmetry. However, the assumption of exact symmetry limits the applicability of MFGs, as real-world scenarios often feature inherent heterogeneity. Furthermore, most works on MFG assume access to a known MFG model, which might not be readily available for real-world finite-agent games. In this work, we broaden the applicability of MFGs by providing a methodology to extend any finite-player, possibly asymmetric, game to an "induced MFG". First, we prove that N-player dynamic games can be symmetrized and smoothly extended to the infinite-player continuum via explicit Kirszbraun extensions. Next, we propose the notion of α, β-symmetric games, a new class of dynamic population games that incorporate approximate permutation invariance. For α, β-symmetric games, we establish explicit approximation bounds, demonstrating that a Nash policy of the induced MFG is an approximate Nash of the N-player dynamic game. We show that TD learning converges up to a small bias using trajectories of the N-player game with finite-sample guarantees, permitting symmetrized learning without building an explicit MFG model. Finally, for certain games satisfying monotonicity, we prove a sample complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-6})$ for the N-agent game to learn an ε-Nash up to symmetrization bias. Our theory is supported by evaluations on MARL benchmarks with thousands of agents.

1 Introduction

Competitive multi-agent reinforcement learning (MARL) has found a wide range of applications in the recent years [52, 59, 48, 45, 35, 34]. Simultaneously, MARL is fundamentally challenging at the regime with many agents due to an exponentially growing search space [58], also known as the *curse-of-many-agents*. Even finding an *approximate* solution (i.e. approximate *Nash*) is PPAD-hard [16], thus potentially intractable. For these reasons, it has been an active area of research to identify "islands of tractability", where MARL can be solved efficiently (see e.g. [33, 43]).

In this work, we develop a theory of efficient learning for MARL problems that exhibit *approximate symmetry* building upon the theory of mean-field games (MFG). MFG is a common theoretical framework for breaking the curse of many agents under perfect symmetry. Initially proposed by [30] and [25], MFG analyzes N-player games with symmetric agents when N is large. In this setting, the so-called *propagation of chaos* permits the reduction of the N-player game to a game between a representative agent and a population distribution. This theoretical framework has been widely studied in many recent works [1, 20, 42, 43, 60, 61].

However, works on MFG exhibit two major bottlenecks preventing wider applicability in MARL. First and foremost, the aforementioned works on MFG all assume some form of exact symmetry between agents. Namely, in the MFGs, all agents must have the same reward function and dynamics

38th Workshop on Aligning Reinforcement Learning Experimentalists and Theorists (ARLET 2024).

Work	Symmetry	Approximation	Learning	Learn w/o model
Saldi et al., 2018	Exact	✓*(asymptotic)*	✗	-
Yardim et al., 2024	Exact	✓*(explicit)*	✗	-
Cui and Koeppl, 2021	Exact	✓*(asymptotic)*	✓*(reg.)*	✗
Zaman et al., 2023	Exact	✗	✓*(reg.)*	✗
Parise and Ozdaglar, 2019	Graphon	✓*(explicit)*	✗	-
Zhang et al., 2023	Graphon	✗	✓*(mon.)*	✗
Pérolat et al., 2022	Multi-pop.	✗	✓*(mon.)*	✗
Our work	α, β-symm.	✓(*explicit*)	✓(*mon.*)	✓

Table 1: Selected models of symmetric games studied in MF-RL works. (*reg.*: only Nash with regularization strictly bounded away from zero, *mon.*: monotonicity assumption)

must be homogeneous (or permutation invariant) among agents. Such perfect symmetry between agents in MARL is theoretically convenient yet practically infeasible: Even in applications where symmetry is presumed, usually, there are imperfections in dynamics that break invariance. Little research has studied whether MFGs could offer tractable approximations to otherwise intractable games that might exhibit approximate symmetries. Secondly, many works on MFG (such as [19, 42]) implicitly assume that an exact model of the MFG is known to the algorithm akin to solving a known MDP. In real-world applications, an exact MFG model might not be readily available. MFGs can potentially address settings where only N-player dynamics (possibly incorporating imperfections and heterogeneity) can be simulated; however, such a theory of MFGs has yet to be developed.

We address these shortcomings by developing a theoretically sound MARL framework for scenarios when permutation invariance holds only approximately. Unlike previous work on MFG, our theoretical approach is *constructive*: we show that given any MARL problem, one can construct an MFG approximation that permits efficient learning. We define a new, broad class of games with approximate permutation invariance, dubbed α, β*-symmetric* games, for which approximate Nash equilibria can be learned efficiently. Our theoretical framework provides *end-to-end* learning guarantees for policy mirror descent combined with TD learning. Our experimental findings further demonstrate strong performance improvements in MARL problems with thousands of agents.

1.1 Related Work

We compare our work with selected past MFG results in Table 1, and also provide a detailed commentary in this section.

Mean-field games and RL. MFGs represent a particular type of competitive game where players exhibit strong symmetries. Past work has studied the existence of MFG Nash equilibrium as well as its approximation of finite-player Nash [10, 11, 47]. The convergence of RL algorithms has also been widely studied in discrete-time MFG assuming either contractivity in the stationary equilibrium setting [64, 61, 19, 60, 13] or monotonicity in the finite-horizon setting [43, 42, 62, 41, 44]. These models however assume exact homogeneity between all participants. Furthermore, existing algorithms typically assume knowledge of the exact MFG model [19, 64], hindering their real-world applicability. Multi-population MFG (MP-MFG) can incorporate multiple types of populations exposed to different dynamics [25, 42, 53, 17, 5, 11, 24]. However, within each population exact symmetry must hold and the number of types must be much smaller than the number of agents. Moreover, MP-MFG can be lifted to an equivalent single-population MFG [24] under certain constraints. Overall, all of these works require variations of the stringent symmetry assumptions, restricting their applicability. A detailed survey of learning MFG can be found at [32].

Graphon MFG. Graphon games, proposed initially by [40], can incorporate heterogeneity between MFG agents by assuming graphon-based interactions. The setting has been analyzed in discrete-time [14, 57] as well as in the continuous time setting [2, 7, 3]. Recently, policy mirror descent has been analyzed in this setting to produce convergence guarantees under monotonicity conditions [65]. However, these works on graphon mean-field games still incorporate exact symmetry in the form of the graphon: namely, the types of agents must follow a symmetric distribution and interactions must be through a symmetric graphon. In fact, graphon MFGs can still be reduced to regular MFGs [65].

Other related work. Another class of games where a large number of agents can be tackled tractably are the so-called potential games [46], generalized to Markov potential games incorporating dynamics [33]. Approximate potentials have been studied in a similar spirit on Markov α-potential games [22] and near potential games [8]. However, to the best of our knowledge, approximate symmetry has not been studied in the literature of MFGs.

1.2 Our Contributions

We list the following as our contributions, compared to past work summarized in the previous section.

1. We first tackle the foundational but understudied question for MFGs: *when can a given N-agent game be meaningfully extended to an infinite-player MFG?* We construct a well-defined MFG approximation to an arbitrary (possibly non-symmetric) finite-player dynamical game (DG) using the idea of function symmetrization and via Kirszbraun Lipschitz extensions.
2. Using our extension, we define a new class of *α, β-symmetric DGs* for which it is tractable to find approximate Nash. α, β-symmetry generalizes permutation invariance in dynamic games to arbitrary MARL problems, where parameters α, β quantify degrees of heterogeneity in dynamics and player rewards respectively.
3. We prove that the solution of the induced MFG is indeed an approximate Nash to the original α, β-symmetric DG up to a bias of $\mathcal{O}(1/\sqrt{N} + \alpha + \beta)$, demonstrating that MFG approximation is robust to heterogeneity and finite-agent errors in the DG.
4. We analyze TD learning on the trajectories of the finite-agent DG. We show that by only using $\mathcal{O}(\varepsilon^{-2})$ samples from the N-player game, policies can be approximately evaluated *on the abstract MFG* up to symmetrization error.
5. Finally, we show that under monotonicity conditions, policy mirror descent (PMD) combined with TD learning converges to an approximate Nash equilibrium using $\widetilde{\mathcal{O}}(\varepsilon^{-6})$ sample trajectories of the N-player DG. This provides an end-to-end learning guarantee for MARL under α, β-symmetry, characterizing a novel class of problems that can be solved efficiently with MARL.

2 Main Results

Notation. For $K \in \mathbb{N}_{>0}$, let $[K] := \{1, \dots, K\}$. Let $\Delta_{\mathcal{X}}$ be the probability simplex over $\mathcal{X}$. For any $N \in \mathbb{N}_{>0}$ define $\Delta_{\mathcal{X},N} := \{\mu \in \Delta_{\mathcal{X}} \mid N\mu(x) \in \mathbb{N}_{\geqslant 0}, \forall x \in \mathcal{X}\}$. For $\mathbf{x} \in \mathcal{X}^N$, define the empirical distribution $\sigma(\mathbf{x}) \in \Delta_{\mathcal{X},N}$ as $\sigma(\mathbf{x})(x') = 1/N \sum_{i=1}^{N} \mathbb{1}_{x_i = x'}$. Let $\mathbb{S}_K$ be the set of permutations over the set $[K]$, so $\mathbb{S}_K := \{g : [K] \to [K] \mid g \text{ bijective}\}$. For $\mathbf{x} = (x_1, \dots, x_K) \in \mathcal{X}^K$ and $g \in \mathbb{S}_K$, define $g(\mathbf{x}) := (x_{g(1)}, \dots, x_{g(K)}) \in \mathcal{X}^K$. Define $\mathbf{x}^{-i} \in \mathcal{X}^{K-1}$ as the vector with i-th entry of $\mathbf{a}$ removed, and $(x, \mathbf{x}^{-i}) \in \mathcal{X}^K$ as the vector where i-th coordinate of $\mathbf{x}$ has been replaced by $x \in \mathcal{X}$.

We consider discrete state-action sets $\mathcal{S}, \mathcal{A}$. We denote the set of time-dependent policies on $\mathcal{S}, \mathcal{A}$ as $\Pi := \{\pi : \mathcal{S} \times [H] \to \Delta_{\mathcal{A}}\}$. We abbreviate $\pi_h(a|s) := \pi(s,h)(a)$. For $p : \mathcal{S} \to \Delta_{\mathcal{A}}$ and $\rho \in \Delta_{\mathcal{S}}$, we define $(\rho \cdot p) \in \Delta_{\mathcal{S}}$ as $(\rho \cdot p)(s,a) := \rho(s)p(s)(a)$ for all $s, a \in \mathcal{S} \times \mathcal{A}$. Finally, we define entropy $\mathcal{H}(u) := -\sum_a u(a) \log u(a)$ for $u \in \Delta_{\mathcal{A}}$. We denote $D_{\mathrm{KL}}(u|v) := \sum_a u(a) \log \frac{u(a)}{v(a)}$ for $u, v \in \Delta_{\mathcal{A}}$.

2.1 Finite-Horizon Dynamic Games

Firstly, we define finite-horizon dynamic games, the main object of interest of this work.

Definition 1 (N-player FH-DG)**.** *An N-player finite-horizon dynamic game (FH-DG) is a tuple $(\mathcal{S}, \mathcal{A}, \rho_0, N, H, \{P^i\}_{i=1}^N, \{R^i\}_{i=1}^N)$ where the state and actions sets $\mathcal{S}, \mathcal{A}$ are discrete, $\rho_0 \in \Delta_{\mathcal{S}}$, the number of players $N \in \mathbb{N}_{>1}$, horizon length $H \in \mathbb{N}_{>0}$, and transition dynamics and rewards are functions such that $P^i : \mathcal{S} \times \mathcal{A} \times (\mathcal{S} \times \mathcal{A})^{N-1} \to \Delta_{\mathcal{S}}$ and $R^i : \mathcal{S} \times \mathcal{A} \times (\mathcal{S} \times \mathcal{A})^{N-1} \to [0,1]$.*

The above definition differs from Markov games [51], where a common state is shared by all agents. In FH-DG, each agent only observes their own state and the dynamics depend on the state vector of all N agents. Such a model can be especially realistic in cases where games have natural *locality*, that is, the game state is not globally available to agents. Next, we define the Nash equilibrium.

Definition 2 (FH-DG Nash equilibrium)**.** *For a FH-DG $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \rho_0, N, H, \{P^i\}_{i=1}^N, \{R^i\}_{i=1}^N)$, policy tuple $\boldsymbol{\pi} = (\pi^1, \ldots, \pi^N) \in \Pi^N$ the expected total reward of agent $i \in [N]$ is defined as*

$$J^{(i)}(\boldsymbol{\pi}) := \mathbb{E}\left[\sum_{h=0}^{H-1} R^i(s_h^i, a_h^i, \boldsymbol{\rho}_h^{-i}) \,\middle|\, \begin{smallmatrix} \forall j: s_0^j \sim \rho_0, \quad a_h^j \sim \pi_h^j(s_h^j) \\ s_{h+1}^j \sim P^j(\cdot | s_h^j, a_h^j, \boldsymbol{\rho}_h^{-j}) \end{smallmatrix}\right]$$

where $\boldsymbol{\rho}_h := (s_h^i, a_h^i)_{i=1}^N$. The exploitability *of agent i for policies $\boldsymbol{\pi}$ is then defined as $\mathcal{E}^{(i)}(\boldsymbol{\pi}) := \max_{\pi \in \Pi} J^{(i)}\left(\pi, \boldsymbol{\pi}^{-i}\right) - J^{(i)}(\boldsymbol{\pi})$. If $\max_i \mathcal{E}^{(i)}(\boldsymbol{\pi}) = 0$, $\boldsymbol{\pi}$ is called a Nash equilibrium (NE) of the FH-DG. If $\max_i \mathcal{E}^{(i)}(\boldsymbol{\pi}) \leqslant \delta$, $\boldsymbol{\pi}$ is called a δ-Nash equilibrium (δ-NE) of the FH-DG.*

At a δ-NE, the incentive for any selfish agent to deviate is small, therefore, approximate NE is a natural solution concept for FH-DG. However, the problem of finding an NE is challenging: not only is the problem computationally intractable in general [15], but for large N the search space of policies grows exponentially. This motivates the approximation via symmetrization in the remainder of the work.

2.2 Symmetrization and Lipschitz Extension

In order to define approximate symmetry, we first show that finite-agent dynamics of Definition 1 can be extended to infinitely many players. In the process, we tackle a question that is relevant for MFGs beyond our work:*When and how can we build an MFG model on the continuum, given dynamics on finitely many players?* We will use the notions of symmetrization and Lipschitz extension.

Definition 3 (Symmetric function, symmetrization)**.** *A function $f : \mathcal{X}^K \to \mathcal{Y}$ is called* symmetric *if $f(g(\mathbf{x})) = f(\mathbf{x})$, $\forall \mathbf{x} \in \mathcal{X}^K$, $g \in \mathbb{S}_K$. For a symmetric $f : \mathcal{X}^K \to \mathcal{Y}$, we define its* population lifted version *$\bar{f} : \Delta_{\mathcal{X},K} \to \mathcal{Y}$ as the well-defined function such that $\bar{f}(\mu) = f(\mathbf{x})$ for $\forall \mathbf{x} \in \mathcal{X}^K$ satisfying $\sigma(\mathbf{x}) = \mu$. Given $f : \mathcal{X}^K \to \mathbb{R}^D$, we define the* symmetrization $\mathrm{Sym}(f) : \mathcal{X}^K \to \mathcal{Y}$ *as*

$$\mathrm{Sym}(f)(\mathbf{x}) = \frac{1}{K!} \sum_{g \in \mathbb{S}_K} f(g(\mathbf{a})), \quad \forall \mathbf{a} \in \mathcal{X}^K.$$

We also denote $\overline{\mathrm{Sym}}(f) := \overline{\mathrm{Sym}(f)}$.

We note that the terminology "symmetrization" is consistent as $\mathrm{Sym}(f)$ is indeed a symmetric function (as verified in Section A). Furthermore, if f is symmetric then $\mathrm{Sym}(f) = f$ as expected.

Finally, to extend DG to the infinite-player continuum, we will use the following special case of the Kirszbraun-Valentine theorem, which concerns Lipschitz extensions of functions from strict subsets of the Euclidean space to the entirety of the space preserving their Lipschitz modulus.

Lemma 1 (Kirszbraun-Valentine [27, 56])**.** *Let $d_1, d_2 \in \mathbb{N}_{>0}$, and $U \subset \mathbb{R}^{d_1}$. Let $f : U \to \mathbb{R}^{d_2}$ be an L-Lipschitz function with respect to the Euclidean norm $\|\cdot\|_2$. Then, there exists $\mathrm{Ext}(f) : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ such that $\mathrm{Ext}(f)$ is L-Lipschitz and $\mathrm{Ext}(f)(x) = f(x)$ for all $x \in U$.*

While $\mathrm{Ext}(f)$ is not unique in general, it can be explicitly formulated in various ways [37, 54], and the particular formulation is not important in this work.

2.3 Mean-field Games and α, β-Symmetric Games

Next, using the definitions from the previous section, we show how the FH-DG can be extended to an MFG. We formalize the finite-horizon MFG (FH-MFG), which will be the main approximation tool.

Definition 4 (Finite-horizon mean-field game)**.** *A finite-horizon mean-field game (FH-MFG) is a tuple $(\mathcal{S}, \mathcal{A}, \rho_0, H, P, R)$ where $\mathcal{S}, \mathcal{A}$ are discrete, $\rho_0 \in \Delta_{\mathcal{S}}$, $H \in \mathbb{N}_{>0}$, the transition dynamics P is a function $P : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}\times\mathcal{A}} \to \Delta_{\mathcal{S}}$, and the reward R is a function $R : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}\times\mathcal{A}} \to [0, 1]$.*

Compared to Definition 1, Definition 4 introduces two conceptual changes under the premise of exact symmetry: (1) the dependency of dynamics to the states and actions of other agents have been reduced to a dependency on a population distribution on $\Delta_{\mathcal{S}\times\mathcal{A}}$, and (2) N agents have been implicitly replaced by a single representative agent. We next extend the definition NE to MFGs.

Definition 5 (Induced population, MFG-NE)**.** *For a FH-MFG defined by the tuple* $(\mathcal{S}, \mathcal{A}, \rho_0, H, P, R)$*, we define the population update operators* Γ, Λ *as*

$$\Gamma(\mu, \pi)(s', a') := \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mu(s, a) P(s'|s, a, \mu) \pi(a'|s') \tag{1}$$

$$\Lambda(\pi) := \left\{\Gamma(\cdots \Gamma(\Gamma(\rho_0 \cdot \pi_0, \pi_1) \cdots, \pi_{h-1}))\right\}_{h=0}^{H-1}. \tag{2}$$

For $\pi \in \Pi$ *and* $\boldsymbol{\mu} = \{\mu_h\}_{h=0}^{H-1} \in \Delta_{\mathcal{S}\times\mathcal{A}}^H$*, the expected reward is defined as*

$$V(\boldsymbol{\mu}, \pi) := \mathbb{E}\left[\sum_{h=0}^{H-1} R(s_h, a_h, \mu_h) \middle| \substack{s_0 \sim \rho_0, \quad a_h \sim \pi_h(s_h) \\ s_{h+1} \sim P(s_h, a_h, \mu_h)}\right]. \tag{3}$$

We define MFG exploitability as $\mathcal{E}(\pi) := \max_{\pi' \in \Pi} V(\Lambda(\pi), \pi') - V(\Lambda(\pi), \pi)$ *and FH-MFG-NE as:*

$$\textit{Policy } \pi^* = \{\pi_h^*\}_{h=0}^{H-1} \in \Pi \textit{ such that} \quad \mathcal{E}(\pi^*) = 0. \tag{MFG-NE}$$

Intuitively, the above definition of MFG-NE requires that the policy π is optimal against the population flow it induces. Questions of existence [9, 4, 23] and approximation of the FH-DG under exact symmetry [47] have been thoroughly studied in the literature. That is, if an N-player game exhibits exact symmetry, then the MFG-NE exists and is an approximate NE of the FH-DG.

Taking a constructive approach, we show that the FH-MFG-NE of an *appropriately constructed* MFG is also an approximate NE for a given FH-DG without a prior model. The definition below of an "induced MFG" demonstrates how arbitrary non-symmetric dynamics can be extended to an MFG.

Definition 6 (Induced FH-MFG)**.** *Let* $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \rho_0, N, H, \{P^i\}_{i=1}^N, \{R^i\}_{i=1}^N)$ *be a FH-DG. The MFG induced by* $\mathcal{G}$*, denoted* $\mathrm{MFG}(\mathcal{G})$*, is defined to be the* $(\mathcal{S}, \mathcal{A}, \rho_0, H, P, R)$*, where* $P : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}\times\mathcal{A}} \to \Delta_{\mathcal{S}}$ *and* $R : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}\times\mathcal{A}} \to [0, 1]$ *are defined for all* $s \in \mathcal{S}, a \in \mathcal{A}, \mu \in \Delta_{\mathcal{S}\times\mathcal{A}}$ *as:*

$$P(s, a, \mu) := \sum_{i=1}^{N} \frac{\mathrm{Ext}\left(\overline{\mathrm{Sym}}\left(P^i(s, a, \cdot)\right)\right)(\mu)}{N}, \quad R(s, a, \mu) := \sum_{i=1}^{N} \frac{\mathrm{Ext}\left(\overline{\mathrm{Sym}}\left(R^i(s, a, \cdot)\right)\right)(\mu)}{N}.$$

$\mathrm{MFG}(\mathcal{G})$ is well-defined due to Lemma 1. In words, the definition of $\mathrm{MFG}(\mathcal{G})$ consists of two main operations: (1) symmetrize ($\overline{\mathrm{Sym}}(\cdot)$) and extend ($\mathrm{Ext}(\cdot)$) P^i, R^i to $\Delta_{\mathcal{S}\times\mathcal{A}}$, and (2) average symmetrized dynamics and rewards for all players. Furthermore, in the special case $P^i = P^j$ and $R^i = R^j$ for all $i \neq j$ and $P^i(s, a, \cdot), R^i(s, a, \cdot)$ are symmetric, the $\mathrm{MFG}(\mathcal{G})$ has dynamics and rewards $\mathrm{Ext}\left(\bar{P}^1\right), \mathrm{Ext}\left(\bar{R}^1\right)$, which are simply the Lipschitz extensions of P^1, R^1 to the continuum.

Remark 1. *Even for exact symmetric games, Definition 6 is relevant. The availability of an MFG model is typically taken for granted, however, since real-world algorithms might only be able to access finite-agent dynamics without a known MFG model, it is a valid research question when and how a game can be meaningfully extended to infinite players (answered by Definition 6).*

Finally, we provide the definition of approximate or α, β-symmetry.

Definition 7 (α, β-Symmetric DG)**.** *Let* $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \rho_0, N, H, \{P^i\}_{i=1}^N, \{R^i\}_{i=1}^N)$ *be an* N*-player FH-DG, inducing* $\mathrm{MFG}(\mathcal{G}) = (\mathcal{S}, \mathcal{A}, \rho_0, H, P, R)$*. If it holds for* $\alpha, \beta \geqslant 0$ *that*

$$\max_{\substack{i \in [N], s, a \in \mathcal{S}\times\mathcal{A} \\ \mu \in \Delta_{\mathcal{S}\times\mathcal{A}}}} \max_{\substack{\boldsymbol{\rho} \in (\mathcal{S}\times\mathcal{A})^{N-1} \\ \sigma(\boldsymbol{\rho}) = \mu}} \|P^i(s, a, \boldsymbol{\rho}) - P(s, a, \mu)\|_1 \leqslant \alpha,$$

$$\max_{\substack{i \in [N], s, a \in \mathcal{S}\times\mathcal{A} \\ \mu \in \Delta_{\mathcal{S}\times\mathcal{A}}}} \max_{\substack{\boldsymbol{\rho} \in (\mathcal{S}\times\mathcal{A})^{N-1} \\ \sigma(\boldsymbol{\rho}) = \mu}} |R^i(s, a, \boldsymbol{\rho}) - R(s, a, \mu)| \leqslant \beta,$$

then the FH-DG $\mathcal{G}$ *is called* α, β*-symmetric.*

As expected, an exactly symmetric N-player game is also $0, 0$-symmetric, and any dynamic game $\mathcal{G}$ is α, β-symmetric for some constants $\alpha \leqslant 2, \beta \leqslant 1$. Hence, Definition 7 generalizes exact permutation invariance. Games that exhibit near-exact symmetries will have very small constants α, β, we will next provide approximation and learning guarantees for such finite-agent games.

2.4 Approximation of NE under Approximate Symmetry

In this section, we will prove that a NE of the induced $\mathrm{MFG}\,(\mathcal{G})$ is also an approximate NE of the finite-agent game $\mathcal{G}$. We will provide an explicit bound on the approximation, motivating the use of MFGs for solving FH-DG.

We first introduce the notion of κ-sparse dynamics. In words, with κ-sparse dynamics an agent at state s playing action a is impacted only by other agents occupying a sparse set of "neighboring" state-actions $\mathcal{N}_{s,a} \subset \mathcal{S} \times \mathcal{A}$ where $|\mathcal{N}_{s,a}| \leqslant \kappa$. For a subset $\mathcal{U} \subset \mathcal{X}$, we define the function $p_{\mathcal{U}} : \mathcal{X} \to \mathcal{X} \cup \{\perp\}$ as $p_{\mathcal{U}}(x) = x$ if $x \in \mathcal{U}$ and $p_{\mathcal{U}}(x) = \perp$ otherwise, where $\perp$ is treated as a placeholder element such that $\perp \notin \mathcal{U}$.

Definition 8 (κ-sparse dynamics/rewards)**.** *A function $f : \mathcal{X}^M \to \mathcal{Y}$ is called* κ-sparse *(on some $\mathcal{U} \subset \mathcal{X}$) if $|\mathcal{U}| \leqslant \kappa$ and $f(\mathbf{x}) = f(\mathbf{y})$ whenever $p_{\mathcal{U}}(x_i) = p_{\mathcal{U}}(y_i)$ for all $i = 1, \ldots, M$. Dynamics $\{P^i\}_{i=1}^N$ (resp. rewards $\{R^i\}_{i=1}^N$) are called* κ-sparse *if all $P^i(s, a, \cdot)$ (resp. $R^i(s, a, \cdot)$) are κ-sparse on some $\mathcal{U}_{s,a} \subset \mathcal{S} \times \mathcal{A}$ for all $s \in \mathcal{S}, a \in \mathcal{A}$ (resp. $\mathcal{U}_{s,a} \subset \mathcal{S} \times \mathcal{A}$ for all $s \in \mathcal{S}, a \in \mathcal{A}$).*

Sparsity is common in FH-DG, particularly when there is spatial structure. Many standard MFG problems such as the beach-bar problem [43] and crowd modeling [64] are in fact $(\kappa = 1)$-sparse, as agents are only affected by the population distribution at their current state.

Using sparsity, we provide an upper bound of the Lipschitz constants of maps $P(s, a, \cdot), R(s, a, \cdot)$ of the induced MFG on the continuum $\Delta_{\mathcal{S}\times\mathcal{A}}$, demonstrating that unless the FH-DG exhibits dominant players, P, R have bounded Lipschitz moduli independent of N.

Lemma 2 (Lipschitz extension bound)**.** *Let $\mathcal{G}$ be an FH-DG with dynamics and rewards $\{P^i\}_{i=1}^N, \{R^i\}_{i=1}^N$ admitting the induced mean-field game $\mathrm{MFG}\,(\mathcal{G})$ with dynamics and rewards P, R. Assume that $\{P^i\}_{i=1}^N, \{R^i\}_{i=1}^N$ are κ-sparse and it holds that*

$$\|P^i(s, a, \boldsymbol{\rho}) - P^i(s, a, ((s', a'), \boldsymbol{\rho}^{-j}))\|_1 \leqslant C_1, \quad |R^i(s, a, \boldsymbol{\rho}) - R^i(s, a, ((s', a'), \boldsymbol{\rho}^{-j}))| \leqslant C_2,$$

for any $i, j \in [N], i \neq j$, $s, s' \in \mathcal{S}, a, a' \in \mathcal{A}$ and $\boldsymbol{\rho} \in (\mathcal{S} \times \mathcal{A})^{N-1}$ for some constants C_1, C_2. Then, the induced P, R have Lipschitz modulus at most $C_1 N\kappa$ and $C_2 N\sqrt{\kappa}$ respectively, that is,

$$\|P(s, a, \mu) - P(s, a, \mu')\|_2 \leqslant C_1 N\kappa\|\mu - \mu'\|_2, \quad |R(s, a, \mu) - R(s, a, \mu')| \leqslant C_2 N\sqrt{\kappa}\|\mu - \mu'\|_2,$$

for any $s \in \mathcal{S}, a \in \mathcal{A}, \mu, \mu' \in \Delta_{\mathcal{S}\times\mathcal{A}}$.

The above theorem characterizes a condition on the original FH-DG for the induced FH-MFG to have smooth (Lipschitz) dynamics. The theorem suggests that the game must have *no dominant players* so that the effect of each agent on others is upper bounded of order $\mathcal{O}(1/N)$. Furthermore, by standard results in MFG literature, if the "no dominant players" condition of Lemma 2 holds, the population update Γ is also Lipschitz continuous with some modulus $L_{pop,\mu}$ that is independent of N.

Finally, we state the main approximation result, which quantifies how closely the true N-player game Nash equilibrium can be approximated by the mean-field Nash equilibrium of the symmetrized game.

Theorem 1. *Let $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \rho_0, N, H, \{P^i\}_{i=1}^N, \{R^i\}_{i=1}^N)$ be an N-player FH-DG and $\mathrm{MFG}\,(\mathcal{G}) = (\mathcal{S}, \mathcal{A}, \rho_0, H, P, R)$. Let the Lipschitz modulus of the population update operator Γ in μ be $L_{pop,\mu}$. If $\pi^* \in \Pi$ is a MFG-NE of $\mathrm{MFG}\,(\mathcal{G})$, then $(\pi^*, \ldots, \pi^*) \in \Pi^N$ is an ϵ-NE of the FH-DG, where*

$$\epsilon = \mathcal{O}\left(\frac{H^2(1 - L_{pop,\mu}^H)}{(1 - L_{pop,\mu})\sqrt{N}} + \alpha H^2 \frac{1 - L_{pop,\mu}^H}{1 - L_{pop,\mu}} + \beta H\right).$$

Proof. (sketch) We show that (1) the empirical distribution of agent state-actions over $\mathcal{S} \times \mathcal{A}$ approximates the induced mean-field $\Lambda(\pi^*)$, (2) marginal distributions of states of an agent $\mathbb{P}[s_h^i = \cdot]$ in FH-DG are also approximated by the mean-field, and (3) explicitly bounding the difference between V and $J^{(i)}$. The formal proof and explicit upper bound are presented in Section B.3. □

Most critically, the approximation bound proves that the MFG approximation is robust to small heterogeneity: when α, β are small, the induced MFG-NE approximates the true NE well. Furthermore, the upper bound suggests three different asymptotic regimes depending on Γ being non-expansive, contractive, or expansive. If $L_{pop,\mu} \leqslant 1$, the bound above is polynomial. If $L_{pop,\mu} > 1$, $\alpha > 0$ might incur an exponential dependency on H, whereas the error due to $\beta > 0$ only scales linearly with

$\mathcal{O}(\beta H)$. However, the exponential worst-case dependence of the bias on H is generally unavoidable even under perfect symmetry, as matching lower bounds are known [63]. Theorem 1 also recovers the bounds known for exactly symmetric FH-DG (i.e. $\alpha = \beta = 0$, see [63]).

Finally, we emphasize that Theorem 1 does not assume any particular structure on the FH-DG: the results apply for any values of α, β, although the quality of approximation will vary. Furthermore, it is known that for $N > 2$, finding an ϵ-NE for the FH-DG is PPAD-complete even for a certain *absolute constant* ϵ [18]. Hence, even when α, β are not close to 0, the result will be useful in approaching the PPAD-complete limit via mean-field approximation.

We emphasize that the results so far already suggest a learning algorithm: one can estimate (e.g. via neural networks) the induced P, R and solve the MFG directly with standard algorithms (e.g. [43, 31]). However, this method can be prohibitively expensive as it involves learning functions to and from $\Delta_{\mathcal{S}\times\mathcal{A}}$. The remainder of the paper will be dedicated to formulating more efficient methods.

2.5 Policy Evaluation with α, β-Symmetry

In this section, we analyze TD learning for α, β-symmetric FH-DG. While Definition 6 provides an explicit construction of an MFG, we show that this construction is not needed for policy evaluation. Namely, using TD learning, a policy π can be evaluated with respect to the (induced) mean-field $\Lambda(\pi)$ only through sampling trajectories of the FH-DG $\mathcal{G}$. We first define Q functions on the MFG.

Definition 9 (Mean-field Q values)**.** *For the MFG $(\mathcal{S}, \mathcal{A}, \rho_0, H, P, R)$, for $\tau > 0$, $h = 0, \dots, H-1$, we define (entropy regularized) Q-values for each $h = 0, \dots, H-1$ and $s \in \mathcal{S}, a \in \mathcal{A}$ as*

$$Q_h^{\tau,\pi}(s,a) := \mathbb{E}\left[\sum_{h'=h}^{H-1} R(s_{h'}, a_{h'}, \mu_{h'}) + \tau\mathcal{H}(\pi_{h'}(\cdot|s_{h'}))\,\middle|\, \begin{smallmatrix} s_h=s,\, a_h=a,\, s_{h'+1}\sim P(s_{h'},a_{h'},\mu_{h'}),\\ a_{h'}\sim\pi_{h'+1}(s_{h'+1}),\, \mu_{h'}:=\Lambda(\pi)_{h'},\forall h'\geqslant h\end{smallmatrix}\right].$$

In other words, the Q-values of a policy π are computed with respect to the MDP induced by the population distributions $\Lambda(\pi)$ in the MFG. We note that the above definition does not match the typical definition of Q-values in a multi-agent setting, and rather is defined concerning an abstract MFG. We note that we will occasionally treat $Q_h^{\tau,\pi}$ as an element of the vector space $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$.

For the finite-horizon problem, we will analyze TD learning, which is a standard method for policy evaluation with established guarantees beyond MFGs [55]. We formulate Algorithm 1, presented for simplicity as performing TD learning on agent 1.

Algorithm 1 TD Learning for α, β-symmetric games

Input: FH-DG $\mathcal{G}$, epochs M, learning rates $\{\eta_m\}_m$, policy $\pi \in \Pi$, entropy regularization $\tau \geqslant 0$.

1: $\widehat{Q}_h^0(s,a) \leftarrow 0, \quad \forall h \in \{0, \dots H-1\}, s \in \mathcal{S}, a \in \mathcal{A}$

2: **for** $m \in 0, 1, \dots, M-1$ **do**

3: Using π for all agents, sample path from $\mathcal{G}$: $\{\boldsymbol{\rho}_{m,h}, \mathbf{r}_{m,h}\}_{h=0}^{H-1} := \{s_{m,h}^i, a_{m,h}^i, r_{m,h}^i\}_{i,h}$.

4: Perform TD update:

$$\widehat{Q}_h^{m+1} \leftarrow \widehat{Q}_h^m + \eta_m\big(\widehat{Q}_{h+1}^m(s_{m,h+1}^1, a_{m,h+1}^1) + r_{m,h}^1 + \mathcal{H}(\pi(\cdot|s_{m,h}^1)) - \widehat{Q}_h^m(s_{m,h}^1, a_{m,h}^1)\big)\mathbf{e}_{s_{m,h}^1, a_{m,h}^1},\ \forall h < H-1$$

$$\widehat{Q}_{H-1}^{m+1} \leftarrow \widehat{Q}_{H-1}^m + \eta_m(\mathcal{H}(\pi(\cdot|s_{m,H-1}^1)) + r_{m,H-1}^1 - \widehat{Q}_h^m(s_{m,h}^1, a_{m,h}^1))\mathbf{e}_{s_{m,H-1}^1, a_{m,H-1}^1}$$

5: Return $\{\widehat{Q}_h^M\}_{h=0}^{H-1}$.

Theorem 2 (TD learning for α, β-Symmetric Games)**.** *Let $\mathcal{G}$ be an N-player FH-DG and $\mathrm{MFG}(\mathcal{G})$ be its induced MFG. Let $\pi \in \Pi$ be a policy such that $\Lambda(\pi) = \boldsymbol{\mu} = \{\mu_h\}_h$ and $\delta := \inf_{h,s,a \text{ s.t. } \mathbb{P}[s_h^1=s, a_h^1=s]>0} \mathbb{P}[s_h^1 = s, a_h^1 = s]$. Assume Algorithm 1 is run with π for $M > 0$ epochs for with learning rates $\eta_m := \frac{2\delta^{-1}}{m+2\delta^{-1}}$. Then, the output $\{\widehat{Q}_h^M\}_h$ of Algorithm 1 satisfies $\mathbb{E}\left[\sum_{h=0}^{H-1} \|\widehat{Q}_h^M - Q_h^{\tau,\pi}\|_{\mu_h}^2\right] \leqslant \mathcal{O}\left(\frac{1}{M} + \frac{1}{N} + \alpha^2 + \beta^2\right)$, where $\|\cdot\|_p$ is defined for $p \in \Delta_{\mathcal{S}\times\mathcal{A}}$ as $\|q\|_p := \sqrt{\sum_{s,a} p(s,a)q(s,a)^2}$.*

Theorem 2 provides a finite-sample guarantee for TD learning, a building block of many MARL and MFG algorithms. Most importantly, it suggests that in order to use mean-field game theory to approximate NE of an FH-DG $\mathcal{G}$, there is no need to explicitly build a model of $\mathrm{MFG}(\mathcal{G})$. Instead, TD learning in the original N-player game when all the agents pursue policy π allows the evaluation of the mean-field Q-values of π efficiently. The rate of convergence suggested by Theorem 2 matches the optimal known rates for TD-learning in a single-agent setting [28]. In practice, one can use the trajectories of all N agents to further improve efficiency, instead of only using that of agent $i = 1$.

2.6 Learning NE under α, β-Symmetry

We complete our framework by providing our key theoretical result: any α, β-symmetric DG can be solved approximately only using samples from the N-player DG, under monotonicity assumptions. Our algorithm uses TD learning as a building block, with stochastic policy evaluations used for policy mirror descent updates [29, 61, 65].

Definition 10 (Monotone MFG [43, 42])**.** *A MFG with dynamics P and rewards R is called* monotone *if P is independent of μ, and for all μ, μ' it holds that $\sum_{s,a}(R(s,a,\mu) - R(s,a,\mu'))(\mu(s,a) - \mu'(s,a)) < 0$. A DG $\mathcal{G}$ is called* monotone extendable *if $\mathrm{MFG}(\mathcal{G})$ is monotone.*

To motivate this definition, we provide a large class of DGs that are relevant and monotone-extendable.

Example 1 (Asymmetric dynamic congestion games)**.** *For any $i \in [N]$, let $h_i : \mathcal{S} \times \mathcal{A} \times [N] \to [0,1], r^i : \mathcal{S} \times \mathcal{A} \to [0,1]$ be arbitrary functions so that $h_i(s,a,\cdot)$ is non-increasing for any s, a. Assume $P^i(\cdot|s,a,\boldsymbol{\rho}^{-i})$ does not depend on $\boldsymbol{\rho}^{-i}$ for any s, a, and $R^i(s,a,\boldsymbol{\rho}^{-i})$ be 1-sparse so that $R^i(s,a,\boldsymbol{\rho}^{-i}) = h_i(s,a,\sum_{j=1}^N \mathbb{1}_{\rho_j=(s,a)}) + r_i(s,a)$. Such games can be seen as generalizations of congestion games [46] and congestion games with player-specific incentives [39], for which an efficient solution is unknown. We prove monotone extendability and characterize the values of α, β and Lipschitz constants for such games in Section D.1*

Algorithm 2 Policy mirror descent for α, β-symmetric games (Symm-PMD)

Input: FH-DG $\mathcal{G}$, epochs T, TD learning epochs M, learning rates $\{\xi_t\}_t$, entropy τ.

1: Initialize uniform policy: $\pi_{0,h}(a|s) = 1/|\mathcal{A}|, \quad \forall h \in \{0, \dots H-1\}, s \in \mathcal{S}, a \in \mathcal{A}$

2: **for** $t \in 0, 1, \dots, T-1$ **do** ▷ *Run for T epochs*

3: Run Algorithm 1 for policy π_t, M epochs, entropy τ, $\{\eta_h\}_m$ as in Theorem 2

4: Obtain $\{\widehat{Q}_h^t\}_{h=0}^{H-1}$, set $\widehat{q}_h^t(s,a) := \widehat{Q}_h^t(s,a) - \mathcal{H}(\pi_{t,h}(\cdot|s))$.

5: Perform PMD update: $(\forall s \in \mathcal{S}, h = 0, \dots, H-1)$

$$\widehat{\pi}_{t+1,h}(\cdot|s) := \arg\max_{u \in \Delta_{\mathcal{A}}} \frac{\xi_t}{1 - \tau\xi_t}\left[\widehat{q}_h^t(s,\cdot)^\top u + \tau\mathcal{H}(u)\right] - D_{\mathrm{KL}}(u|\pi_{t,h}(\cdot|s)).$$

6: Update policies: $\pi_{t+1,h}(\cdot|s) := \left(1 - \frac{1}{t+1}\right)\widehat{\pi}_{t+1,h}(\cdot|s) + \frac{1}{t+1}\,\mathrm{Unif}(\cdot), \quad \forall s \in \mathcal{S}.$

7: Return $\bar{\pi} := \{\frac{1}{T+1}\sum_{t=0}^T \pi_{t,h}\}_{h=0}^{H-1}$.

Theorem 3 (Convergence of PMD)**.** *Let $\mathcal{G}$ be a monotone extendable α, β-symmetric game. Assume Symm-PMD (Algorithm 2) runs with learning rates $\xi_t = \frac{1}{\sqrt{t+1}}$, entropy regularization $\tau \in (0, 1/2)$, $M \geqslant \mathcal{O}(\varepsilon^{-2})$ TD iterations for $T \geqslant \widetilde{\mathcal{O}}(\varepsilon^{-4})$ epochs. Then, the output policy $\bar{\pi}$ is a $\mathcal{O}(\varepsilon + \alpha\tau^{-1} + \beta\tau^{-1} + \tau^{-1}/\sqrt{N} + \tau)$-Nash equilibrium of $\mathcal{G}$ in expectation.*

Proof. The proof is based on [65] with the added complications of finitely many agents, approximate symmetry, and stochastic TD learning. Full proof is presented in Section D.3. □

Theorem 3 suggests a sample complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-6})$ trajectories from the N-agent FH-DG in order to compute a ε-NE (up to symmetrization bias). In fact, it is (to the best of our knowledge) the first finite-sample guarantee for computing approximate NE for a large class of dynamic games with many agents. Most importantly, the number of agents N does not appear in the complexity: hence, the curse of many agents can be provably circumvented for α, β-symmetric games. Even in the exactly symmetric case ($\alpha = \beta = 0$), Theorem 3 is the first guarantee to the best of our knowledge for learning FH-MFG-NE only observing trajectories of the N-agent game.

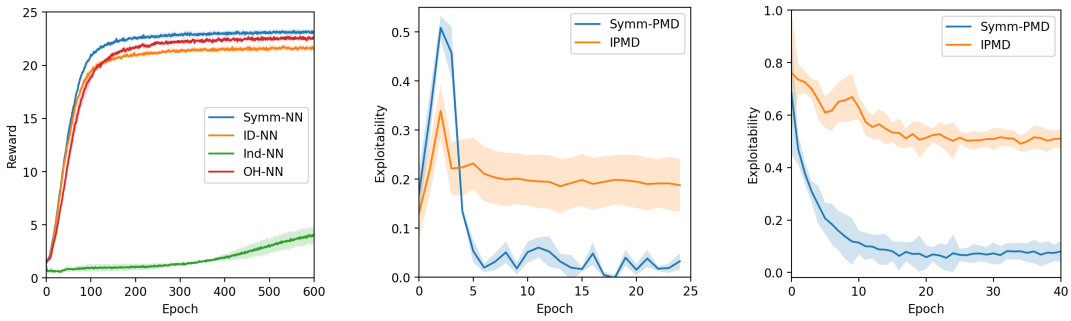
3 Experimental Results

We support our theory by deploying Symm-PMD (Algorithm 2) on several large-scale α, β-symmetric games. For evaluations, we modify the well-known benchmarks from MFG literature (see [13]) to propose three games with asymmetric incentives: A-RPS, A-SIS, and A-Taxi. **A-RPS** is an adaptation of RPS [13] to incorporate asymmetric rewards for agents. **A-SIS** models disease propagation in a large population individually choosing to self-isolate or go out, incorporating asymmetric agents with individual susceptibility/healing rates and unique aversions to isolation. Finally, **A-Taxi** simulates a large population of taxis serving clients in a grid, with individual preferences for regions and crowd aversion. In our experiments, we use $N = 1000$ and $N = 2000$ agents demonstrating the ability of our framework to handle large MARL games. A-Taxi incorporates $|\mathcal{S}| > 2^{30}, H = 128$, hence necessitates neural parameterization. Our setup is thoroughly described in Section E.

We deploy Symm-PMD on two different DGs, with $\alpha = 0, \beta \approx 0.1, N = 2000, H = 10$ on A-RPS and $\alpha \approx \beta \approx 0.1, N = 1000, H = 20$ for A-SIS. We compare the symmetrized approach of Symm-PMD to its asymmetric counterpart independent PMD (IPMD), where a separate policy is learned for each agent. The training curves, pictured in Figures 1-(b,c) characterize the exploitabilty of the learned policies throughout training. In both cases, while IPMD has no approximation bias in principle, it struggles to converge presumably suffering from the curse-of-many-agents. Symm-PMD, on the other hand, rapidly converges to a policy profile with low exploitability and is much more sample-efficient. In both cases, Symm-PMD converges to a solution with low bias.

We demonstrate the scalability of our approach with neural policies. In the A-Taxi environment, we use PPO [50] with symmetrized neural policies and compare to the settings the policy has access to agent identities (either one-hot encoded, in OH-NN, or as an integer, in ID-NN). Symmetrized policies outperform either benchmark by converging faster and to a better solution. Learning independent neural policies for each of 1000 agents (Ind-NN) is extremely expensive in this setting: this approach performs the worst and is orders of magnitude computationally more expensive.

Computational efficiency. We also emphasize the computational efficiency of symmetrization: since our algorithm need not learn separate policies for each agent, it is drastically more computationally efficient compared to independent PMD. In A-SIS and A-RPS benchmarks, learning is 60% faster, whereas symmetrized neural PPO in A-Taxi is >95% faster than its independent counterpart.



(a) (b) (c)

Figure 1: (a) The mean rewards throughout training of symmetric policies (Sym-NN), policies with onehot encoding for i (OH-NN), policies with numerical encoding for i (Ind-NN) and independent policies (Ind-NN) in A-Taxi. (b, c) The exploitability throughout multiple epochs of Symm-PMD (Algorithm 2) and IPMD, for A-RPS with $\beta = 0.1$ in (b) and A-SIS with $\alpha = \beta = 0.1$ in (c).

4 Discussion and Conclusion

We formulated a new class of competitive MARL problems (α, β-symmetric games) that can be tractably solved. We constructively showed that every α, β-symmetric FH-DG can be efficiently approximated by an induced MFG. We provided theoretical guarantees for TD learning, and under monotonicity, for PMD to approximate NE up to symmetrization bias. These results provide a complete theory of learning under approximate symmetry, supported by numerical experiments.

Acknowledgments and Disclosure of Funding

This project is supported by Swiss National Science Foundation (SNSF) under the framework of NCCR Automation and SNSF Starting Grant.

References

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

[2] A. Aurell, R. Carmona, G. Dayanıklı, and M. Laurière. Finite state graphon games with applications to epidemics. *Dynamic Games and Applications*, 12(1):49–81, 2022.

[3] A. Aurell, R. Carmona, and M. Lauriere. Stochastic graphon games: Ii. the linear-quadratic case. *Applied Mathematics & Optimization*, 85(3):39, 2022.

[4] A. Bensoussan, J. Frehse, P. Yam, et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

[5] A. Bensoussan, T. Huang, and M. Lauriere. Mean field control and mean field game models with several populations. *arXiv preprint arXiv:1810.00783*, 2018.

[6] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning*, pages 1283–1294. PMLR, 2020.

[7] P. E. Caines and M. Huang. Graphon mean field games and the gmfg equations: ε-nash equilibria. In *2019 IEEE 58th conference on decision and control (CDC)*, pages 286–292. IEEE, 2019.

[8] O. Candogan, A. Ozdaglar, and P. A. Parrilo. Near-potential games: Geometry and dynamics. *ACM Trans. Econ. Comput.*, 1(2), may 2013. ISSN 2167-8375. doi: 10.1145/2465769.2465776. URL https://doi.org/10.1145/2465769.2465776.

[9] P. Cardaliaguet. Notes on mean field games. Technical report, Technical report, 2010.

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

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

[12] S. Cayci, N. He, and R. Srikant. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *arXiv preprint arXiv:2106.04096*, 2021.

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

[14] K. Cui and H. Koeppl. Learning graphon mean field games and approximate nash equilibria. *arXiv preprint arXiv:2112.01280*, 2021.

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

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

[17] G. Dayanikli and M. Lauriere. Multi-population mean field games with multiple major players: Application to carbon emission regulations. *arXiv preprint arXiv:2309.16477*, 2023.

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

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

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

[21] X. Guo, A. Hu, M. Santamaria, M. Tajrobehkar, and J. Zhang. MFGLib: A library for mean field games. *arXiv preprint arXiv:2304.08630*, 2023.

[22] X. Guo, X. Li, C. Maheshwari, S. Sastry, and M. Wu. Markov α-potential games: Equilibrium approximation and regret analysis. *arXiv preprint arXiv:2305.12553*, 2023.

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

[24] J. Huang, N. He, and A. Krause. Model-based rl for mean-field games is not statistically harder than single-agent rl, 2024.

[25] M. Huang, R. P. Malhamé, and P. E. Caines. 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.

[26] S. Huang, R. F. J. Dossa, C. Ye, J. Braga, D. Chakraborty, K. Mehta, and J. G. AraÃšjo. Cleanrl: High-quality single-file implementations of deep reinforcement learning algorithms. *Journal of Machine Learning Research*, 23(274):1–18, 2022.

[27] M. Kirszbraun. Über die zusammenziehende und lipschitzsche transformationen. *Fundamenta Mathematicae*, 22(1):77–108, 1934.

[28] G. Kotsalis, G. Lan, and T. Li. Simple and optimal methods for stochastic variational inequalities, ii: Markovian noise and policy evaluation in reinforcement learning. *SIAM Journal on Optimization*, 32(2):1120–1155, 2022.

[29] G. Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106, 2023.

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

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

[32] M. Laurière, S. Perrin, J. Pérolat, S. Girgin, P. Muller, R. Élie, M. Geist, and O. Pietquin. Learning in mean field games: A survey, 2024.

[33] S. Leonardos, W. Overman, I. Panageas, and G. Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

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

[35] L. Matignon, G. J. Laurent, and N. Le Fort-Piat. 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*, pages 64–69. IEEE, 2007.

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

[37] E. J. McShane. Extension of range of functions. 1934.

[38] J. Mei, C. Xiao, C. Szepesvari, and D. Schuurmans. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*, pages 6820–6829. PMLR, 2020.

[39] I. Milchtaich. Congestion games with player-specific payoff functions. *Games and economic behavior*, 13(1):111–124, 1996.

[40] F. Parise and A. Ozdaglar. Graphon games. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 457–458, 2019.

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

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

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

[44] S. Perrin, M. Laurière, J. Pérolat, R. Élie, M. Geist, and O. Pietquin. Generalization in mean field games by learning master policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 9413–9421, 2022.

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

[46] R. W. Rosenthal. A class of games possessing pure-strategy nash equilibria. *International Journal of Game Theory*, 2(1):65–67, 1973.

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

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

[49] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel. High-dimensional continuous control using generalized advantage estimation. *arXiv preprint arXiv:1506.02438*, 2015.

[50] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms, 2017.

[51] L. S. Shapley. Stochastic games. *Proceedings of the national academy of sciences*, 39(10): 1095–1100, 1953.

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

[53] S. G. Subramanian, P. Poupart, M. E. Taylor, and N. Hegde. Multi type mean field reinforcement learning. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '20, page 411–419, Richland, SC, 2020. International Foundation for Autonomous Agents and Multiagent Systems. ISBN 9781450375184.

[54] A. Sukharev. Optimal method of constructing best uniform approximations for functions of a certain class. *USSR Computational Mathematics and Mathematical Physics*, 18(2):21–31, 1978.

[55] J. Tsitsiklis and B. Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

[56] F. A. Valentine. A lipschitz condition preserving extension for a vector function. *American Journal of Mathematics*, 67(1):83–93, 1945.

[57] D. Vasal, R. K. Mishra, and S. Vishwanath. Master equation of discrete time graphon mean field games and teams. *arXiv preprint arXiv:2001.05633*, 2020.

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

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

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

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

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

[63] B. Yardim, A. Goldman, and N. He. When is mean-field reinforcement learning tractable and relevant?, 2024.

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

[65] F. Zhang, V. Y. Tan, Z. Wang, and Z. Yang. Learning regularized monotone graphon mean-field games. *arXiv preprint arXiv:2310.08089*, 2023.

A Preliminaries

Firstly, we present several basic facts regarding symmetrization and symmetric functions.

Lemma 3. *For any $f : \mathcal{X}^K \to \mathcal{Y}$, $\mathrm{Sym}\,(f)$ is a symmetric function.*

Proof. For any $g' \in \mathbb{S}_K$, we have

$$\mathrm{Sym}\,(f)\,(g'(\mathbf{x})) = \frac{1}{K!}\sum_{g\in\mathbb{S}_K} f(g(g'(\mathbf{x}))) = \frac{1}{K!}\sum_{g\in\mathbb{S}_K} f(g(\mathbf{x})) = \mathrm{Sym}\,(f)\,(\mathbf{x}),$$

since composition by g' defines a bijection from $\mathbb{S}_K$ onto itself. □

Lemma 4. *For any symmetric function $f : \mathcal{X}^K \to \mathcal{Y}$, $\mathrm{Sym}\,(f) = f$.*

Proof. By simple computation:

$$\mathrm{Sym}\,(f)\,(\mathbf{x}) = \frac{1}{K!}\sum_{g\in\mathbb{S}_K} f(g(\mathbf{x})) = \frac{1}{K!}\sum_{g\in\mathbb{S}_K} f(\mathbf{x}) = f(\mathbf{x}).$$

□

Normed policy space. In the proofs, we equip the policy space Π with the norm $\|\cdot\|_1$ defined as

$$\|\pi - \pi'\|_1 := \sup_{s\in\mathcal{S}} \|\pi(s) - \pi'(s)\|_1,$$

for any $\pi, \pi' \in \Pi$. We present several useful results.

Lemma 5. *Let $\pi, \pi' \in \Pi$ and $\mu, \mu' \in \Delta_{\mathcal{S}}$ be arbitrary. Then,*

$$\|\mu\cdot\pi - \mu'\cdot\pi'\|_1 \leqslant \|\mu - \mu'\|_1 + \|\pi - \pi'\|_1.$$

Proof. The lemma follows from the two inequalities

$$\begin{aligned}\|\mu\cdot\pi - \mu\cdot\pi'\|_1 &\leqslant \sum_{s,a} |\mu(s)\pi(a|s) - \mu(s)\pi'(a|s)| \\ &\leqslant \sum_s \mu(s) \sum_a |\pi(a|s) - \pi'(a|s)| \leqslant \|\pi - \pi'\|_1,\end{aligned}$$

and similarly:

$$\begin{aligned}\|\mu\cdot\pi - \mu'\cdot\pi\|_1 &\leqslant \sum_{s,a} |\mu(s)\pi(a|s) - \mu'(s)\pi(a|s)| \\ &\leqslant \sum_s |\mu(s) - \mu'(s)| \sum_a \pi(a|s) = \|\mu - \mu'\|_1.\end{aligned}$$

□

Lemma 6 (Lemma B.2 of [61])**.** *Assume E a finite set, $g : E \to \mathbb{R}^p$ a vector value function, and ν, μ two probability measures on E. Then,*

$$\left\|\sum_e g(e)\mu(e) - \sum_e g(e)\nu(e)\right\|_1 \leqslant \frac{\lambda_g}{2}\|\mu - \nu\|_1,$$

where $\lambda_g := \sup_{e,e'} \|g(e) - g(e')\|_1$.

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

Definition 11 (Sub-Gaussian)**.** *Random variable ξ is called sub-Gaussian with variance proxy σ^2 if $\forall\lambda \in \mathbb{R}: \quad \mathbb{E}\left[e^{\lambda(\xi - \mathbb{E}[\xi])}\right] \leqslant e^{\frac{\lambda^2\sigma^2}{2}}$. 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 $\xi_i \in SG(\frac{(b-a)^2}{4})$. We also state the well-known Hoeffding concentration bound and a corollary, Lemma 8.

Lemma 7 (Hoeffding inequality [36])**.** *Let $\xi \in SG(\sigma^2)$. Then for any $t > 0$ it holds that $\mathbb{P}\left|\xi - \mathbb{E}\left[\xi\right]\right| \geqslant t \leqslant 2e^{-\frac{t^2}{2\sigma^2}}$.*

Lemma 8. *Let $\xi \in SG(\sigma^2)$. Then*

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

Proof.

$$\mathbb{E}\left[\left|\xi - \mathbb{E}\left[\xi\right]\right|\right] = \int_0^\infty \mathbb{P}(\left|\xi - \mathbb{E}\left[\xi\right]\right| \geqslant t)dt \overset{(I)}{\leqslant} 2\int_0^\infty e^{-\frac{t^2}{2\sigma^2}}dt = \sqrt{2\pi\sigma^2}.$$

Inequality (I) is true due to Lemma 7. Likewise,

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

□

For controlling errors under stochasticity, the following simple lemma will be useful.

Lemma 9 (Harmonic partial sum bound)**.** *For any integers $s, \bar{s}$ such that $1 \leqslant \bar{s} < s$ and $p \neq -1$, it holds that*

$$\log s - \log \bar{s} + \frac{1}{s} \leqslant \sum_{n=\bar{s}}^{s} \frac{1}{n} \leqslant \frac{1}{\bar{s}} + \log s - \log \bar{s},$$

$$\frac{s^{p+1}}{p+1} - \frac{\bar{s}^{p+1}}{(p+1)} + \bar{s}^p \leqslant \sum_{n=\bar{s}}^{s} n^p \leqslant \frac{s^{p+1}}{p+1} - \frac{\bar{s}^{p+1}}{p+1} + s^p, \text{ if } p \geqslant 0$$

$$\frac{s^{p+1}}{p+1} - \frac{\bar{s}^{p+1}}{p+1} + s^p \leqslant \sum_{n=\bar{s}}^{s} n^p \leqslant \frac{s^{p+1}}{p+1} - \frac{\bar{s}^{p+1}}{p+1} + \bar{s}^p, \text{ if } p \leqslant 0$$

Proof. The proof follows from the the basic fact that if $f : [1, \infty) \to \mathbb{R}_{\geqslant 0}$ is a non-increasing function, then

$$\int_{\bar{s}}^{s} f(x)\mathrm{d}x + f(s) \leqslant \sum_{n=\bar{s}}^{s} f(n) \leqslant \int_{\bar{s}}^{s} f(x)\mathrm{d}x + f(\bar{s}),$$

and likewise for a non-decreasing function $f : [1, \infty) \to \mathbb{R}_{\geqslant 0}$, it holds that

$$\int_{\bar{s}}^{s} f(x)\mathrm{d}x + f(\bar{s}) \leqslant \sum_{n=\bar{s}}^{s} f(n) \leqslant \int_{\bar{s}}^{s} f(x)\mathrm{d}x + f(s).$$

□

Finally, we slightly generalize the definition of MFG-NE (Definition 12), as our approximation theorems are somewhat more general than what is stated in the main body of the paper: we consider approximate MFG-NE rather than only exact MFG-NE.

Definition 12 (δ-MFG-NE)**.** *A policy sequence $\boldsymbol{\pi}^* \in \Pi_H$ is called a δ-MFG-NE of the MFG $(\mathcal{S}, \mathcal{A}, \rho_0, H, P, R)$ if it holds that*

$$\mathcal{E}(\{\pi_h^*\}_{h=0}^{H-1}) \leqslant \delta. \tag{δ-MFG-NE}$$

A remark on extension Lemma 1 and $\Delta_{\mathcal{S}}$. For given $N > 0$ and map $\bar{P} : \Delta_{\mathcal{S}\times\mathcal{A},N} \to \Delta_{\mathcal{S}}$ with Lipschitz modulus L on $\Delta_{\mathcal{S}\times\mathcal{A},N}$, the Kirszbraun-Valentine Lemma (Lemma 1) only guarantees an L-Lipschitz extension $\mathrm{Ext}\left(\bar{P}\right) : \Delta_{\mathcal{S}\times\mathcal{A},N} \to \mathbb{R}^{\mathcal{S}}$. However, we can trivially side-step this issue with a modified application of Kirszbraun-Valentine. Let $\mathrm{Proj}_{\Delta_{\mathcal{S}}} : \mathbb{R}^{\mathcal{S}} \to \Delta_{\mathcal{S}}$ be the projection operator to the convex set $\Delta_{\mathcal{S}}$. For any extension $\mathrm{Ext}\left(\bar{P}\right)$, $\mathrm{Proj}_{\Delta_{\mathcal{S}}} \circ \mathrm{Ext}\left(\bar{P}\right)$ is also a valid L-Lipschitz extension that preserves $\bar{P}$ on the set $\Delta_{\mathcal{S}\times\mathcal{A},N}$ as $\mathrm{Proj}_{\Delta_{\mathcal{S}}}$ is non-expansive. Moreover, $\mathrm{Proj}_{\Delta_{\mathcal{S}}} \circ \mathrm{Ext}\left(\bar{P}\right)$ has image set contained in $\Delta_{\mathcal{S}}$ as required.

B Extended Proofs on Approximation

B.1 Proof of Lemma 2

The proof relies on the properties of symmetrization and Lemma 6. Forn convenience we denote $K := N - 1$ in this proof.

Firstly, we show that for any $i, j \in [N], s \in \mathcal{S}, a \in \mathcal{A}$, the functions $\mathrm{Sym}\left(P^i\right)(s, a, \cdot)$ and $\mathrm{Sym}\left(R^i\right)(s, a, \cdot)$ also satisfy the bounded variation and sparsity assumptions. Assume that

$\boldsymbol{\rho} \in (\mathcal{S} \times \mathcal{A})^K, (s', a') \in \mathcal{S} \times \mathcal{A}$ arbitrary, Then, by definition,

$$\begin{aligned}
&\| \operatorname{Sym}\left(P^i\right)(s, a, \boldsymbol{\rho}) - \operatorname{Sym}\left(P^i\right)(s, a, ((s', a'), \boldsymbol{\rho}^{-j}))\|_2 \\
&\leqslant \frac{1}{K!} \left\| \sum_{g \in \mathbb{S}_K} P^i(s, a, g(\boldsymbol{\rho})) - \sum_{g \in \mathbb{S}_K} P^i(s, a, g((s', a'), \boldsymbol{\rho}^{-j})) \right\|_2 \\
&\leqslant \frac{1}{K!} \sum_{g \in \mathbb{S}_K} \left\| P^i(s, a, g(\boldsymbol{\rho})) - P^i(s, a, g((s', a'), \boldsymbol{\rho}^{-j})) \right\|_2 \\
&\leqslant \frac{1}{K!} \sum_{g \in \mathbb{S}_K} C_1 \leqslant C_1.
\end{aligned}$$

Furthermore, assume $P^i(s, a, \cdot)$ is κ-sparse on some set $\mathcal{U}_{s,a} \subset \mathcal{S} \times \mathcal{A}$ where $|\mathcal{U}_{s,a}| \leqslant \kappa$. Let $\boldsymbol{\rho}, \boldsymbol{\rho}' \in (\mathcal{S} \times \mathcal{A})^K$ be two vectors agreeing in their entries in $\mathcal{U}_{s,a}$, i.e., $p_{\mathcal{U}_{s,a}}(\boldsymbol{\rho}) = p_{\mathcal{U}_{s,a}}(\boldsymbol{\rho}')$. Then,

$$\begin{aligned}
\operatorname{Sym}\left(P^i\right)(s, a, \boldsymbol{\rho}) &= \frac{1}{K!} \sum_{g \in \mathbb{S}_K} P^i(s, a, g(\boldsymbol{\rho})) \\
&= \frac{1}{K!} \sum_{g \in \mathbb{S}_K} P^i(s, a, g(\boldsymbol{\rho})) \\
&= \operatorname{Sym}\left(P^i\right)(s, a, \boldsymbol{\rho}'),
\end{aligned}$$

since $g(\boldsymbol{\rho}')$ agrees with $g(\boldsymbol{\rho})$ on its elements in $\mathcal{U}_{s,a}$ as well, as $p_{\mathcal{U}_{s,a}}(g(\boldsymbol{\rho})) = p_{\mathcal{U}_{s,a}}(g(\boldsymbol{\rho}'))$. Therefore we conclude $\operatorname{Sym}\left(P^i\right)$ is also κ-sparse on $\mathcal{U}_{s,a}$. By similar computation,

$$\left|\operatorname{Sym}\left(R^i\right)(s, a, \boldsymbol{\rho}) - \operatorname{Sym}\left(R^i\right)(s, a, ((s', a'), \boldsymbol{\rho}^{-j}))\right| \leqslant C_2,$$

and $\operatorname{Sym}\left(R^i\right)$ is also κ-sparse.

Next, we establish that the lifted functions $\overline{\operatorname{Sym}}\left(P^i\right)(s, a, \cdot)$ only depend on $\mu(s)$ for $s \in \mathcal{U}_{s,a}$, that is, we show that if μ, μ' are such that $\mu(s', a') = \mu'(s', a')$ for all $(s', a') \in \mathcal{U}_{s,a}$, then $\overline{\operatorname{Sym}}\left(P^i\right)(s, a, \mu) = \overline{\operatorname{Sym}}\left(P^i\right)(s, a, \mu')$. Let $\mu, \mu' \in \Delta_{\mathcal{S} \times \mathcal{A}, K}$ be such that $\mu(s', a') = \mu'(s', a')$ for all $(s', a') \in \mathcal{U}_{s,a}$. Take arbitrary $\boldsymbol{\rho}, \boldsymbol{\rho}'$ such that $\sigma(\boldsymbol{\rho}) = \mu, \sigma(\boldsymbol{\rho}') = \mu'$. It holds that for some permutation $g' \in \mathbb{S}_K$ that $g'(\boldsymbol{\rho}')$ agrees with $\boldsymbol{\rho}$ on all entries taking values in $\mathcal{U}_{s,a}$, as $\boldsymbol{\rho}'$ and $\boldsymbol{\rho}$ have the same count of elements in $\mathcal{U}_{s,a}$. Then

$$\begin{aligned}
\overline{\operatorname{Sym}}\left(P^i\right)(s, a, \mu) =& \frac{1}{K!} \sum_{g \in \mathbb{S}_K} P^i(s, a, g(\boldsymbol{\rho})) \\
=& \frac{1}{K!} \sum_{g \in \mathbb{S}_K} P^i(s, a, g(g'(\boldsymbol{\rho}'))) \\
=& \frac{1}{K!} \sum_{g \in \mathbb{S}_K} P^i(s, a, g(\boldsymbol{\rho}')) = \overline{\operatorname{Sym}}\left(P^i\right)(s, a, \mu').
\end{aligned}$$

A similar argument works for $\overline{\operatorname{Sym}}\left(R^i\right)(s, a, \cdot)$, allowing us to conclude that $\overline{\operatorname{Sym}}\left(P^i\right)(s, a, \cdot), \overline{\operatorname{Sym}}\left(R^i\right)(s, a, \cdot)$ only depend on $\mu(s', a')$ if $(s', a') \in \mathcal{U}_{s,a}$.

Finally, we analyze the Lipschitz modulus of the lifted functions $1/N \sum_i \overline{\operatorname{Sym}}\left(P^i\right), 1/N \sum_i \overline{\operatorname{Sym}}\left(R^i\right)$. Let $\mu_1, \mu_2 \in \Delta_{\mathcal{S} \times \mathcal{A}, K}$ and $\boldsymbol{\rho}_1 = \{\rho_1^i\}_{i=1}^K, \boldsymbol{\rho}_2 = \{\rho_2^i\}_{i=1}^K \in (\mathcal{S} \times \mathcal{A})^K$ be such that $\sigma(\boldsymbol{\rho}_1) = \mu_1, \sigma(\boldsymbol{\rho}_2) = \mu_2$. Then,

$$\|\overline{\operatorname{Sym}}\left(P^i\right)(s, a, \boldsymbol{\rho}_1) - \overline{\operatorname{Sym}}\left(P^i\right)(s, a, \boldsymbol{\rho}_2)\|_1 \leqslant C_1 \sum_{i \in [K]} \mathbb{1}_{\rho_1^i \neq \rho_2^i}.$$

Taking the minimum over such $\boldsymbol{\rho}_1, \boldsymbol{\rho}_2$, we have that

$$\begin{aligned}
&\|\overline{\operatorname{Sym}}\left(P^i\right)(s, a, \mu_1) - \overline{\operatorname{Sym}}\left(P^i\right)(s, a, \mu_2)\|_1 \\
&\leqslant \min_{\substack{\boldsymbol{\rho}_1, \boldsymbol{\rho}_2 \in (\mathcal{S} \times \mathcal{A})^K \\ \sigma(\boldsymbol{\rho}_1) = \mu_1,\, \sigma(\boldsymbol{\rho}_2) = \mu_2}} C_1 \sum_{i \in [K]} \mathbb{1}_{\rho_1^i \neq \rho_2^i} \leqslant C_1 K \|\mu_1 - \mu_2\|_1,
\end{aligned}$$

as $\boldsymbol{\rho}_1, \boldsymbol{\rho}_2$ can differ at a minimum at $K\|\mu_1 - \mu_2\|_1$ coordinates. Finally, as concluded from the arguments above, since $\overline{\mathrm{Sym}}\left(P^i\right)(s,a,\mu_1)$ only depends on $\mu_1(s',a')$ for $s',a' \in \mathcal{U}_{s,a}$, one can choose $\bar{\mu}_1, \bar{\mu}_2 \in \Delta_{\mathcal{S}\times\mathcal{A}}$ such that $\bar{\mu}_1(s',a') = \mu_1(s',a')$ and $\bar{\mu}_2(s',a') = \mu_2(s',a')$ for all $(s',a') \in \mathcal{U}_{s,a}$ and $\bar{\mu}_1(s'',a'') = \bar{\mu}_2(s'',a'')$ whenever $(s'',a'') \neq \mathcal{U}_{s,a}$. Then,

$$\begin{aligned}
\|\overline{\mathrm{Sym}}\left(P^i\right)(s,a,\mu_1) - \overline{\mathrm{Sym}}\left(P^i\right)(s,a,\mu_2)\|_1 =& \|\overline{\mathrm{Sym}}\left(P^i\right)(s,a,\bar{\mu}_1) - \overline{\mathrm{Sym}}\left(P^i\right)(s,a,\bar{\mu}_2)\|_1 \\
=& C_1 K \|\bar{\mu}_1 - \bar{\mu}_2\|_1 \\
=& C_1 K \sum_{s',a' \in \mathcal{U}_{s,a}} |\bar{\mu}_1(s',a') - \bar{\mu}_2(s',a')| \\
\leqslant& C_1 K \sqrt{\sum_{s',a' \in \mathcal{U}_{s,a}} |\bar{\mu}_1(s',a') - \bar{\mu}_2(s',a')|^2}\sqrt{\kappa} \\
\leqslant& C_1 K \sqrt{\kappa}\|\bar{\mu}_1 - \bar{\mu}_2\|_2 \leqslant C_1 K \sqrt{\kappa}\|\mu_1 - \mu_2\|_2,
\end{aligned}$$

thus proving Lipschitz bound on the set $\Delta_{\mathcal{S}\times\mathcal{A},K}$. By an identical argument, it holds that

$$|\overline{\mathrm{Sym}}\left(R^i\right)(s,a,\mu_1) - \overline{\mathrm{Sym}}\left(R^i\right)(s,a,\mu_2)| \leqslant C_2(N-1)\sqrt{\kappa}\|\mu_1 - \mu_2\|_2.$$

The result follows from an application of Lemma 1 to extend $1/N \sum_i \overline{\mathrm{Sym}}\left(R^i\right)(s,a,\cdot)$ and $1/N \sum_i \overline{\mathrm{Sym}}\left(P^i\right)(s,a,\cdot)$ from $\Delta_{\mathcal{S}\times\mathcal{A},N}$ to $\Delta_{\mathcal{S}\times\mathcal{A}}$, as the norm equivalence $\|\cdot\|_2 \leqslant \|\cdot\|_1$ holds.

B.2 Population Flows are Lipschitz Continuous

Lemma 10 (Lipschitz continuity of Γ)**.** *Let $P : \mathcal{S}\times\mathcal{A}\times\Delta_{\mathcal{S}\times\mathcal{A}} \to \Delta_{\mathcal{S}}$ be such that $P(s,a,\mu)$ is Lipschitz continuous in $\|\cdot\|_1$ norm with modulus $K_\mu > 0$ and*

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

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

$$\|\Gamma(\mu,\pi) - \Gamma(\mu',\pi)\|_1 \leqslant \left(\frac{K_s + K_a}{2} + K_\mu\right)\|\mu - \mu'\|_1,$$

for all $\pi \in \Pi$, $\mu,\mu' \in \Delta_{\mathcal{S}\times\mathcal{A}}$.

Proof. The proof is inspired by [61], apart from the fact that in our case the population update operator is defined differently as:

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

We will prove a slightly more general statement, that

$$\left\|\Gamma(\mu,\pi) - \Gamma(\mu',\pi')\right\|_1 \leqslant \|\mu - \mu'\|_1\left(\frac{K_s + K_a}{2} + K_\mu\right) + \|\pi - \pi'\|_1.$$

Firstly, we upper bound the Lipschitz modulus of the function with respect to μ. For any $\mu, \mu' \in \Delta_{\mathcal{S}\times\mathcal{A}}$, it holds that:

$$
\begin{aligned}
&\left\|\Gamma(\mu,\pi)-\Gamma(\mu',\pi)\right\|_1 \\
&\leqslant \sum_{s',a'}\left|\sum_{s,a}(\mu(s,a)P(s'|s,a,\mu)-\mu'(s,a)P(s'|s,a,\mu'))\pi(a'|s')\right| \\
&\leqslant \sum_{s',a'}\left|\sum_{s,a}(\mu(s,a)-\mu'(s,a))P(s'|s,a,\mu)\pi(a'|s')\right| \\
&\quad+\sum_{s',a'}\left|\sum_{s\in\mathcal{S},a\in\mathcal{A}}\mu'(s,a)(P(s'|s,a,\mu)-P(s'|s,a,\mu'))\pi(a'|s')\right| \\
&\leqslant \left\|\sum_{s,a}(\mu(s,a)-\mu'(s,a))(P(s,a,\mu)\cdot\pi)\right\|_1 \\
&\quad+\sum_{s,a}\mu'(s,a)\sum_{s'}|P(s'|s,a,\mu)-P(s'|s,a,\mu')|\sum_{a'}\pi(a'|s') \\
&\leqslant \|\mu-\mu'\|_1\frac{\max\|P(s,a,\mu)\cdot\pi-P(\overline{s},\overline{a},\mu)\cdot\pi\|_1}{2}+K_\mu\|\mu-\mu'\|_1 \\
&\leqslant \|\mu-\mu'\|_1\frac{\max\|P(s,a,\mu)-P(\overline{s},\overline{a},\mu)\|_1}{2}+K_\mu\|\mu-\mu'\|_1.
\end{aligned}
$$

where the last two lines follow from Lemma 6 and Lemma 5. Since

$$\|P(s,a,\mu)-P(\overline{s},\overline{a},\mu)\|_1 \leqslant K_s+K_a,$$

we have the claimed inequality

$$\left\|\Gamma(\mu,\pi)-\Gamma(\mu',\pi)\right\|_1 \leqslant \|\mu-\mu'\|_1\left(\frac{K_s+K_a}{2}+K_\mu\right).$$

Finally, the Lipschitz constant for the policy π is computed by:

$$
\begin{aligned}
\left\|\Gamma(\mu,\pi)-\Gamma(\mu,\pi')\right\|_1 &\leqslant \sum_{s',a'}\left|\sum_{s,a}\mu(s,a)P(s'|s,a,\mu)\left(\pi(a'|s')-\pi'(a'|s')\right)\right| \\
&\leqslant \sum_{s,a,s'}\mu(s,a)P(s'|s,a,\mu)\sum_{a'}\left|\pi(a'|s')-\pi'(a'|s')\right| \\
&\leqslant \|\pi-\pi'\|_1.
\end{aligned}
$$

□

B.3 Proof of Theorem 1

The main ideas of the approximation proof are similar to some arguments from MFG literature (e.g. see [47]) with two major differences: (1) the dynamics of the finite player game are not exactly symmetric, and (2) unlike some standard works the dynamics and rewards depend on the distribution of agents over state-action pairs, not just states.

For given R and P define the following constants:

$$
\begin{aligned}
L_s &= \sup_{s,s',a,\mu}\left|R(s,a,\mu)-R(s',a,\mu)\right|, \qquad L_a = \sup_{s,a,a',\mu}\left|R(s,a,\mu)-R(s,a',\mu)\right|, \\
K_s &= \sup_{s,s',a,\mu}\left\|P(\cdot|s,a,\mu)-P(\cdot|s',a,\mu)\right\|_1, \qquad K_a = \sup_{s,a,a',\mu}\left\|P(\cdot|s,a,\mu)-P(\cdot|s,a',\mu)\right\|_1.
\end{aligned}
$$

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

$$P(\cdot|s,u,\mu)=\sum_{a\in\mathcal{A}}u(a)P(\cdot|s,a,\mu), \qquad R(s,u,\mu)=\sum_{a\in\mathcal{A}}u(a)R(s,a,\mu).$$

By [61, Lemma C.1], it holds that

$$\begin{aligned}\|P(\cdot|s,u,\mu) - P(\cdot|s',u',\mu')\|_1 \leqslant & K_\mu\|\mu-\mu'\|_1 + K_s d(s,s') + \frac{K_a}{2}\|u-u'\|_1,\\ |R(s,u,\mu) - R(s',u',\mu')| \leqslant & L_\mu\|\mu-\mu'\|_1 + L_s d(s,s') + \frac{L_a}{2}\|u-u'\|_1. \end{aligned} \tag{4}$$

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

$$\Gamma^h(\mu,\pi) := \underbrace{\Gamma(\dots\Gamma(\Gamma(\mu,\pi_0),\pi_1)\dots,\pi_{h-1})}_{h \text{ times}}$$

so that $\Gamma^0(\mu,\pi) := \mu$. Lemma 10 yields the Lipschitz condition:

$$\begin{aligned}&\|\Gamma^n(\mu,\{\pi_i\}_{i=0}^{n-1}) - \Gamma^n(\mu',\{\pi_i\}_{i=0}^{n-1})\|_1\\ &\leqslant L_{pop,\mu}\|\Gamma^{n-1}(\mu,\{\pi_i\}_{i=0}^{n-2}) - \Gamma^{n-1}(\mu',\{\pi'_i\}_{i=0}^{n-2})\|_1 + \|\pi_{n-1}-\pi'_{n-1}\|_1\\ &\leqslant L^n_{pop,\mu}\|\mu-\mu'\|_1 + \sum_{i=0}^{n-1} L^{n-1-i}_{pop,\mu}\|\pi_i-\pi'_i\|_1, \end{aligned} \tag{5}$$

where $L_{pop,\mu}$ is the Lipschitz constant of Γ in μ.

We also define a useful function $\Xi : (\mathcal{S}\times\mathcal{A})^N \times \Pi^N \to \Delta_{\mathcal{S}\times\mathcal{A}}$ such that for any $\boldsymbol{\rho} = \{(s^i,a^i)\}_{i=1}^N$,

$$\Xi(\boldsymbol{\rho},\bar{\pi}) = \frac{1}{N}\sum_{i=1}^N P(\cdot|s^i,a^i,\sigma(\boldsymbol{\rho}^{-i}))\cdot\bar{\pi}.$$

In other words, Ξ is the average population flow expected under symmetrized dynamics and reference policy $\bar{\pi}$.

The proof will proceed in four steps:

- **Step 1.** Bounding the expected deviation of the empirical population distribution from the mean-field distribution $\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h\|_1\right]$ for any given policy π.
- **Step 2.** Bounding total variation distance (or equivalently ℓ_1 distance) between the marginal distributions $\mathbb{P}[s^1_h = \cdot]$ in the N-player game and $\mathbb{P}[s_h = \cdot]$ in the mean-field game,
- **Step 3.** Bounding difference of N agent value function $J^{(1)}$ and the infinite player value function V, when all the players except the first one play the same policy,
- **Step 4.** 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. In this step, we will assume N players of the FH-DG pursue policies $\{\pi^i\}_i = \{\pi^i_h\}_{i,h} \in \Pi^N$, and random variables f Furthermore, assume $\bar{\pi} = \{\bar{\pi}_h\}_h \in \Pi$ arbitrary, and induces population $\boldsymbol{\mu} = \Lambda(\bar{\pi}) = \{\mu_h\}_h$. We also define the quantity $\Delta_h := \frac{1}{N}\sum_{i=1}^N \|\pi^i_h - \bar{\pi}_h\|_1$ and $\bar{\Delta} := \max_{h\in[H]}\Delta_h$.

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

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\mu}_0-\mu_0\|_1\right] \leqslant & \mathbb{E}\left[\left\|\widehat{\mu}_0 - \frac{1}{N}\sum_{i=1}^N \rho_0\cdot\pi^i_0\right\|_1\right] + \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N \rho_0\cdot\pi^i_0 - \mu_0\right\|_1\right]\\ \leqslant & \sum_{\substack{s\in\mathcal{S}\\a\in\mathcal{A}}}\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N(\mathbb{1}_{s^i_0=s,a^i_0=a} - \rho_0(s)\pi^i_0(a|s))\right|\right]\\ & + \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^N\|\pi^i_0-\bar{\pi}_0\|_1\right]\\ \leqslant & |\mathcal{S}||\mathcal{A}|\sqrt{\frac{\pi}{2N}} + \Delta_h, \end{aligned}$$

where the last line is due to Lemma 8 and the fact that $\mathbb{1}_{s_0^i=s,a_0^i=a}$ are independent, bounded (hence subgaussian) random variables, and that we have $\mathbb{E}\left[\mathbb{1}_{s_0^i=s,a_0^i=a}\right] = \rho_0(s)\pi(s,a) = \mu_0(s,a)$.

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

$$\begin{aligned}\mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mu_{h+1}\|_1 \,|\mathcal{F}_h\right] \leqslant & \underbrace{\mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right]\|_1\,|\mathcal{F}_h\right]}_{(\triangle)} \\ & + \underbrace{\mathbb{E}\left[\|\mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right] - \Xi(\boldsymbol{\rho}_h, \overline{\pi}_{h+1})\|_1\,|\mathcal{F}_h\right]}_{(\square)} \\ & + \underbrace{\mathbb{E}\left[\|\Xi(\boldsymbol{\rho}_h, \overline{\pi}_{h+1}) - \Gamma(\widehat{\mu}_h, \overline{\pi}_h)\|_1\,|\mathcal{F}_h\right]}_{(\star)} \\ & + \underbrace{\mathbb{E}\left[\|\Gamma(\widehat{\mu}_h, \overline{\pi}_h) - \mu_{h+1}\|_1\,|\mathcal{F}_h\right]}_{(\heartsuit)} \end{aligned} \tag{6}$$

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

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

since each $\widehat{\mu}_{h+1}(s)$ is an average of N independent subgaussian random variables (specifically N independent Bernoulli random variables) given $\mathcal{F}_h$. Specifically, each indicator is bounded $\mathbb{1}_{s_{h+1}^i=s,a_{h+1}^i=a} \in [0,1]$ almost surely and therefore is sub-Gaussian with $\mathbb{1}_{s_{h+1}^i=s,a_{h+1}^i=a} \in SG(1/4)$.

Next, for the term $(\square)$,

$$\begin{aligned}(\square) &= \mathbb{E}\left[\|\mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right] - \Xi(\boldsymbol{\rho}_h, \overline{\pi}_{h+1})\|_1\,|\mathcal{F}_h\right] \\ &= \frac{1}{N}\,\mathbb{E}\left[\left\|\sum_{i=1}^N P^i(\cdot|s_h^i, a_h^i, \boldsymbol{\rho}_h^{-i})\cdot\pi_{h+1}^i - \sum_{i=1}^N P(\cdot|s_h^i, a_h^i, \sigma(\boldsymbol{\rho}_h^{-i}))\cdot\overline{\pi}_{h+1}\right\|_1\middle|\mathcal{F}_h\right] \\ &\leqslant \frac{1}{N}\,\mathbb{E}\left[\sum_{i=1}^N \|P^i(\cdot|s_h^i, a_h^i, \boldsymbol{\rho}_h^{-i}) - P(\cdot|s_h^i, a_h^i, \sigma(\boldsymbol{\rho}_h^{-i}))\|_1\middle|\mathcal{F}_h\right] \\ &\quad + \frac{1}{N}\sum_{i=1}^N \|\pi_{h+1}^i - \overline{\pi}_{h+1}\|_1.\end{aligned}$$

By the α-symmetry condition, it follows that $(\square) \leqslant \alpha + \Delta_{h+1}$.

For $(\star) = \|\Xi(\boldsymbol{\rho}_h, \overline{\pi}_h\}_i) - \Gamma(\widehat{\mu}_h, \overline{\pi}_h)\|_1$,

$$\begin{aligned}(\star) &= \mathbb{E}\left[\|\Xi(\boldsymbol{\rho}_h, \overline{\pi}_{h+1}) - \Gamma(\widehat{\mu}_h, \overline{\pi}_{h+1})\|_1\,|\mathcal{F}_h\right] \\ &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N P(\cdot|s_h^i, a_h^i, \sigma(\boldsymbol{\rho}_h^{-i}))\cdot\overline{\pi}_{h+1} - \sum_{s',a'}\widehat{\mu}_h(s',a')P(\cdot|s',a',\widehat{\mu}_h)\cdot\overline{\pi}_{h+1}\right\|_1\middle|\mathcal{F}_h\right]. \\ &= \frac{1}{N}\,\mathbb{E}\left[\left\|\sum_{i=1}^N P(\cdot|s_h^i, a_h^i, \sigma(\boldsymbol{\rho}_h^{-i}))\cdot\overline{\pi}_{h+1} - \sum_{i=1}^N P(\cdot|s_h^i, a_h^1, \widehat{\mu}_h)\cdot\overline{\pi}_{h+1}\right\|_1\middle|\mathcal{F}_h\right].\end{aligned}$$

The vectors $(N-1)\sigma(\boldsymbol{\rho}_h^{-i}), N\widehat{\mu}_h$ can differ by only 1 in one component due to the i-th agent being excluded from the former, it holds that

$$\|N\sigma(\boldsymbol{\rho}_h^{-i}) - N\widehat{\mu}_h\|_1 \leqslant \|(N-1)\sigma(\boldsymbol{\rho}_h^{-i}) - N\widehat{\mu}_h\|_1 + \|\sigma(\boldsymbol{\rho}_h^{-i})\|_1 \leqslant 3,$$

therefore for any s, a,

$$\|P(\cdot|s,a,\sigma(\boldsymbol{\rho}_h^{-i})) - P(\cdot|s,a,\widehat{\mu}_h)\|_1 \leqslant \frac{3K_\mu}{N}$$

almost surely and $(\star)$ is further upper bounded by $(\star) \leqslant \frac{3K_\mu}{N}$.

Finally, the last term $(\heartsuit)$ can be bounded using:

$$(\heartsuit) = \mathbb{E}\left[\|\Gamma(\widehat{\mu}_h, \bar{\pi}_h) - \Gamma(\mu_h^{\bar{\pi}}, \bar{\pi}_h)\|_1 \,|\mathcal{F}_h\right] \leqslant L_{pop,\mu}\|\widehat{\mu}_h - \mu_h\|_1.$$

To conclude, merging the bounds on the three terms in Inequality (6) and taking the expectation on both sides, by the law of iterated expectations we obtain:

$$\mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mu_{h+1}\|_1\right] \leqslant L_{pop,\mu}\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h\|_1\right] + |\mathcal{S}||\mathcal{A}|\sqrt{\frac{\pi}{2N}} + \frac{3K_\mu}{N} + \Delta_{h+1} + \alpha.$$

Induction on h yields the bound for all h:

$$\begin{aligned}
&\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h\|_1\right] \\
&\leqslant \sum_{h'=0}^{h} L_{pop,\mu}^{h-h'}\left(|\mathcal{S}||\mathcal{A}|\sqrt{\frac{\pi}{2N}} + \frac{3K_\mu}{N} + \Delta_{h+1} + \alpha\right) \\
&\leqslant \frac{1 - L_{pop,\mu}^{h+1}}{1 - L_{pop,\mu}}\left(|\mathcal{S}||\mathcal{A}|\sqrt{\frac{\pi}{2N}} + \bar{\Delta} + \frac{3K_\mu}{N} + \alpha\right), \qquad (7)
\end{aligned}$$

where we adopt the convenient shorthand $\frac{1-L_{pop,\mu}^{h+1}}{1-L_{pop,\mu}} := h$ if $L_{pop,\mu} = 1$.

Step 2: Marginal state-action distributions. In this step, we analyze the distributions $\mathbb{P}[s_h^i = \cdot, a_h^i = \cdot]$. For simplicity, assume each player $i \neq 1$ follows policy $\pi \in \Pi$, player 1 follows an arbitrary policy $\bar{\pi} \in \Pi$: we denote the induced random variables in the N-player game $\mathcal{G}$ as $s_h^i, a_h^i, \widehat{\mu}_h$. Assume that in the mean-field game $\mathrm{MFG}(\mathcal{G})$, the representative player in $\mathrm{MFG}(\mathcal{G})$ also follows policy $\bar{\pi}$, evaluated against distribution $\boldsymbol{\mu} := \Lambda(\pi)$: denote the induced random variables as s_h, a_h. In this setting, we will inductively upper bound the quantity

$$\|\,\mathbb{P}[s_h = \cdot] - \mathbb{P}[s_h^1 = \cdot]\|_1.$$

Firstly, for $h = 0$, it holds by definition that

$$\mathbb{P}[s_0 = \cdot] = \mathbb{P}[s_0^1 = \cdot] = \rho_0,$$

hence $\|\,\mathbb{P}[s_0 = \cdot] - \mathbb{P}[s_0^1 = \cdot]\|_1 = 0$.

Next, for the time step $h+1$,

$$\begin{aligned}
&\|\,\mathbb{P}[s_{h+1} = \cdot] - \mathbb{P}[s_{h+1}^1 = \cdot]\|_1 \\
\leqslant &\left\|\sum_{s,\boldsymbol{\rho}} P^1(s, \bar{\pi}_h(s), \boldsymbol{\rho})\,\mathbb{P}[s_h^1 = s, \boldsymbol{\rho}_h^{-i} = \boldsymbol{\rho}] - \sum_s P(s, \bar{\pi}_h(s), \mu_h)\,\mathbb{P}[s_h = s]\right\|_1 \\
\leqslant &\left\|\sum_{s,\boldsymbol{\rho}} P(s, \bar{\pi}_h(s), \sigma(\boldsymbol{\rho}))\,\mathbb{P}[s_h^1 = s, \boldsymbol{\rho}_h^{-i} = \boldsymbol{\rho}] - \sum_s P(s, \bar{\pi}_h(s), \mu_h)\,\mathbb{P}[s_h = s]\right\|_1 \\
&+ \left\|\sum_{s,\boldsymbol{\rho}} [P^1(s, \bar{\pi}_h(s), \boldsymbol{\rho}) - P(s, \bar{\pi}_h(s), \sigma(\boldsymbol{\rho}))]\,\mathbb{P}[s_h^1 = s, \boldsymbol{\rho}_h^{-i} = \boldsymbol{\rho}]\right\|_1 \\
\leqslant &\left\|\sum_{s,\boldsymbol{\rho}} P(s, \bar{\pi}_h(s), \sigma(\boldsymbol{\rho}))\,\mathbb{P}[s_h^1 = s, \boldsymbol{\rho}_h^{-i} = \boldsymbol{\rho}] - \sum_s P(s, \bar{\pi}_h(s), \mu_h)\,\mathbb{P}[s_h = s]\right\|_1 + \alpha, \qquad (8)
\end{aligned}$$

where the last line follows from the α-symmetry condition. For the remaining term, we first observe the inequality:

$$\begin{aligned}
&\left\|\sum_{s,\boldsymbol{\rho}} [P(s, \bar{\pi}_h(s), \sigma(\boldsymbol{\rho})) - P(s, \bar{\pi}_h(s), \mu_h)]\,\mathbb{P}[s_h^1 = s, \boldsymbol{\rho}_h^{-i} = \boldsymbol{\rho}]\right\|_1 \\
&\leqslant \sum_{s,\boldsymbol{\rho}} K_\mu \|\sigma(\boldsymbol{\rho}) - \mu_h\|_1\,\mathbb{P}[s_h^1 = s, \boldsymbol{\rho}_h^{-i} = \boldsymbol{\rho}] \\
&\leqslant K_\mu\,\mathbb{E}[\|\sigma(\boldsymbol{\rho})_h^{-i} - \mu_h\|_1] \\
&\leqslant K_\mu\,\mathbb{E}[\|\widehat{\mu}_h - \mu_h\|_1] + \frac{3K_\mu}{N}
\end{aligned}$$

as again $\|\widehat{\mu}_h - \sigma(\boldsymbol{\rho}_h^{-i})\|_1 \leqslant 3/N$ almost surely, as $N\widehat{\mu}_h$ and $(N-1)\sigma(\boldsymbol{\rho}_h^{-i})$ differ by one in one coordinate only. Then, applying the triangle inequality and marginalizing over $\boldsymbol{\rho}$ in Inequality (8),

$$\begin{aligned}
&\|\,\mathbb{P}[s_{h+1} = \cdot] - \mathbb{P}[s^1_{h+1} = \cdot]\|_1 \\
\leqslant &\left\| \sum_s P(s, \bar{\pi}_h(s), \mu_h)\,\mathbb{P}[s^1_h = s] - \sum_s P(s, \bar{\pi}_h(s), \mu_h)\,\mathbb{P}[s_h = s] \right\|_1 \\
&+ \alpha + \frac{3}{N} + K_\mu\,\mathbb{E}[\|\widehat{\mu}_h - \mu_h\|_1] \\
\leqslant &\|\,\mathbb{P}[s^1_h = s] - \mathbb{P}[s_h = s]\|_1 + \alpha + \frac{3K_\mu}{N} + K_\mu\,\mathbb{E}[\|\widehat{\mu}_h - \mu_h\|_1]
\end{aligned}$$

where in the last line we used the fact that Markov kernels are non-expansive in ℓ_1 norm, where in this case the Markov kernel is given by $P(\cdot|s, \bar{\pi}_h(s), \mu_h)$ for all $s \in \mathcal{S}$.

Inductively applying the recursive bound above we obtain the inequality for all h:

$$\|\,\mathbb{P}[s^1_h = \cdot] - \mathbb{P}[s_h = \cdot]\|_1 \leqslant (h-1)\left(\alpha + \frac{3K_\mu}{N}\right) + K_\mu \sum_{h=0}^{h-1} \mathbb{E}[\|\widehat{\mu}_h - \mu_h\|_1]. \tag{9}$$

By the result in Step 1 (Inequality 7), since $\bar{\Delta} \leqslant 2/N$ in this case it holds that

$$\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h\|_1\right] \leqslant \frac{1 - L^{h+1}_{pop,\mu}}{1 - L_{pop,\mu}}\left(|\mathcal{S}||\mathcal{A}|\sqrt{\frac{\pi}{2N}} + \frac{5K_\mu}{N} + \alpha\right),$$

for all $h = 0, \ldots, N-1$. Merging the two inequalities:

$$\begin{aligned}
&\|\,\mathbb{P}[s^1_h = \cdot] - \mathbb{P}[s_h = \cdot]\|_1 \\
&\leqslant L_{pop,\mu} \sum_{h=0}^{h-1} \frac{1 - L^{h+1}_{pop,\mu}}{1 - L_{pop,\mu}}\left(|\mathcal{S}||\mathcal{A}|\sqrt{\frac{\pi}{2N}} + \frac{5K_\mu}{N} + \alpha\right) + (h-1)\left(\alpha + \frac{3K_\mu}{N}\right).
\end{aligned} \tag{10}$$

Step 3. Bounding value function. In this step, we bound the difference between the expected returns of a player in the N player game and the induced MFG (denoted by functions J, V respectively). Namely, we will upper bound the deviation

$$|J^{(1)}(\bar{\pi}, \pi, \ldots, \pi) - V(\Lambda(\pi), \bar{\pi})|$$

for any two policies $\bar{\pi}, \pi \in \Pi$.

As in step 1, assume each player $i \neq 1$ follows policy $\pi \in \Pi$, and player 1 follows policy $\bar{\pi}$: we denote the induced random variables in the N-player game $\mathcal{G}$ as $s^i_h, a^i_h, \widehat{\mu}_h$. Assume that in the mean-field game $\mathrm{MFG}(\mathcal{G})$, the representative player in $\mathrm{MFG}(\mathcal{G})$ also follows policy $\bar{\pi}$, evaluated against distribution $\boldsymbol{\mu} := \Lambda(\pi)$: denote the induced random variables as s_h, a_h.

By the results in Step 1,2 shown in inequalities (7), (9), since $\bar{\Delta} \leqslant 2/N$ in this case it holds that:

$$\begin{aligned}
&\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h\|_1\right] \leqslant \frac{1 - L^{h+1}_{pop,\mu}}{1 - L_{pop,\mu}}\left(|\mathcal{S}||\mathcal{A}|\sqrt{\frac{\pi}{2N}} + \frac{5K_\mu}{N} + \alpha\right) \\
&\|\,\mathbb{P}[s^1_h = \cdot] - \mathbb{P}[s_h = \cdot]\|_1 \\
&\quad\leqslant (h-1)\left(\alpha + \frac{3K_\mu}{N}\right) + K_\mu \sum_{h=0}^{h-1} \mathbb{E}[\|\widehat{\mu}_h - \mu_h\|_1]
\end{aligned}$$

for all $h = 0, \ldots, N-1$.

At a fixed time step h, the expected one-step reward differences can be decomposed into four terms:

$$\begin{aligned}
&|\,\mathbb{E}[R(s_h, a_h, \mu_h)] - \mathbb{E}[R^1(s^1_h, a^1_h, \boldsymbol{\rho}^{-i})]| \\
&\leqslant |\,\mathbb{E}[R(s_h, a_h, \mu_h)] - \mathbb{E}[R(s^1_h, a^1_h, \mu_h)]| \\
&\quad + |\,\mathbb{E}[R(s^1_h, a^1_h, \mu_h)] - \mathbb{E}[R(s^1_h, a^1_h, \widehat{\mu}_h)]| \\
&\quad + |\,\mathbb{E}[R(s^1_h, a^1_h, \widehat{\mu}_h)] - \mathbb{E}[R(s^1_h, a^1_h, \sigma(\boldsymbol{\rho}^{-i}))]| \\
&\quad + |\,\mathbb{E}[R(s^1_h, a^1_h, \sigma(\boldsymbol{\rho}^{-i}))] - \mathbb{E}[R^1(s^1_h, a^1_h, \boldsymbol{\rho}^{-i})]|,
\end{aligned}$$

enumerating these terms as (I), (II), (III), and (IV), we upper bound each as follows:

$$\begin{aligned}
\text{(I)} \leqslant & \frac{L_a}{2} \| \mathbb{P}[s_h = \cdot, a_h = \cdot] - \mathbb{P}[s_h^1 = \cdot, a_h^1 = \cdot] \|_1 \\
\leqslant & \frac{L_a}{2} \| \mathbb{P}[s_h = \cdot] - \mathbb{P}[s_h^1 = \cdot] \|_1 \\
\text{(II)} \leqslant & L_\mu \, \mathbb{E}[\| \widehat{\mu}_h - \mu_h \|_1] \\
\text{(III)} \leqslant & L_\mu \, \mathbb{E}[\| \widehat{\mu}_h - \sigma(\boldsymbol{\rho}^{-i}) \|_1] \leqslant L_\mu \frac{3}{N} \\
\text{(IV)} \leqslant & \beta
\end{aligned}$$

where the last line uses the α, β symmetry condition.

Then, summing up over the entire time horizon, we obtain the result

$$\begin{aligned}
& |J^{(1)}(\bar{\pi}, \pi, \ldots, \pi) - V(\Lambda(\pi), \bar{\pi})| \\
& \leqslant \sum_{h=0}^{H-1} \left(L_\mu \, \mathbb{E}[\| \widehat{\mu}_h - \mu_h \|_1] + \frac{L_a}{2} \| \mathbb{P}[s_h = \cdot] - \mathbb{P}[s_h^1 = \cdot] \|_1 \right) + \beta H + \frac{3 H L_\mu}{N} \qquad (11)
\end{aligned}$$

or substituting the upper bounds established before,

$$\begin{aligned}
& |J^{(1)}(\bar{\pi}, \pi, \ldots, \pi) - V(\Lambda(\pi), \bar{\pi})| \\
& \leqslant \sum_{h=0}^{H-1} \left(L_\mu \, \mathbb{E}[\| \widehat{\mu}_h - \mu_h \|_1] + \frac{L_a K_\mu}{2} \sum_{h'=0}^{h-1} \mathbb{E}[\| \widehat{\mu}_{h'} - \mu_{h'} \|_1] \right) \\
& \quad + \beta H + \frac{3 H L_\mu}{N} + H^2 \left(\alpha + \frac{3 K_\mu}{N} \right) \\
& \leqslant \sum_{h=0}^{H-1} \left(L_\mu E^{(h)} + \frac{L_a K_\mu}{2} \sum_{h'=0}^{h-1} E^{(h')} \right) + \beta H + \frac{3 H L_\mu}{N} + H^2 \left(\alpha + \frac{3 K_\mu}{N} \right) \qquad (12)
\end{aligned}$$

where we define the quantity

$$E^{(h)} := \frac{1 - L_{pop,\mu}^{h+1}}{1 - L_{pop,\mu}} \left(|\mathcal{S}||\mathcal{A}| \sqrt{\frac{\pi}{2N}} + \frac{5 K_\mu}{N} + \alpha \right).$$

Step 4. Bounding exploitability function. Finally, we use the results from the previous steps to upper bound the exploitability of the MFG-NE policy π^* in the FH-DG. Let $\boldsymbol{\mu}^* = \Lambda(\pi^*)$. Let π' be arbitrary.

$$\begin{aligned}
J^{(1)}(\pi', \pi^*, \ldots, \pi^*) - J^{(1)}(\pi^*, \ldots, \pi^*) \leqslant & V(\Lambda(\pi^*), \pi') - V(\Lambda(\pi^*), \pi^*) \\
& + |J^{(1)}(\pi^*, \pi^*, \ldots, \pi^*) - V(\Lambda(\pi^*), \pi^*)| \\
& + |J^{(1)}(\pi', \pi^*, \ldots, \pi^*) - V(\Lambda(\pi^*), \pi')|.
\end{aligned}$$

The last two terms in this inequality can be bounded by Inequality (12) by choosing $\bar{\pi} = \pi^*$ and $\bar{\pi} = \pi'$ respectively, and the first term satisfies

$$V(\Lambda(\pi^*), \pi') - V(\Lambda(\pi^*), \pi^*) \leqslant \delta$$

as π is assumed to be a δ-MFG-NE.

Then, the main statement of the theorem is obtained by observing

$$\begin{aligned}
\mathcal{E}^{(i)}(\pi^*) &= \max_{\pi' \in \Pi} J^{(i)}\left(\pi', \pi^{*,-i}\right) - J^{(i)}\left(\pi^*, \ldots, \pi^*\right) \\
&\leqslant \delta + 2 \sum_{h=0}^{H-1} \left(L_\mu E^{(h)} + \frac{L_a K_\mu}{2} \sum_{h'=0}^{h-1} E^{(h')} \right) + 2\beta H + \frac{6 H L_\mu}{N} + 2H^2 \left(\alpha + \frac{3 K_\mu}{N} \right),
\end{aligned}$$

where once again

$$E^{(h)} := \frac{1 - L_{pop,\mu}^{h+1}}{1 - L_{pop,\mu}} \left(|\mathcal{S}||\mathcal{A}| \sqrt{\frac{\pi}{2N}} + \frac{5 K_\mu}{N} + \alpha \right).$$

Namely, if $L_{pop,\mu} = 1$, then

$$E^{(h)} = \mathcal{O}\left(h\left(\alpha + \frac{1}{\sqrt{N}}\right)\right),$$

if $L_{pop,\mu} < 1$, then

$$E^{(h)} = \mathcal{O}\left(\alpha + \frac{1}{\sqrt{N}}\right),$$

and finally if $L_{pop,\mu} > 1$, then

$$E^{(h)} = \mathcal{O}\left(L^h_{pop,\mu}\left(\alpha + \frac{1}{\sqrt{N}}\right)\right).$$

C Extended Results on TD Learning

We will make use of the following technical lemma.

Lemma 11. *Let $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \rho_0, N, H, \{P^i\}_{i=1}^N, \{R^i\}_{i=1}^N)$ be a FH-DG which induces $\mathrm{MFG}(\mathcal{G}) = (\mathcal{S}, \mathcal{A}, N, H, P, R)$. Furthermore, assume that the population flow operator Γ of the induced MFG satisfies the Lipschitz condition*

$$\|\Gamma(\mu, \pi) - \Gamma(\mu', \pi)\|_2 \leqslant L_{pop,\mu}\|\mu - \mu'\|_2,$$

for all policies π and $\mu, \mu' \in \Delta_{\mathcal{S}\times\mathcal{A}}$. Then, if all players in the FH-DG play policy $\pi \in \Pi$, it holds that

$$\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h\|_2^2\right] \leqslant C\frac{1 - L_{pop,\mu}^{2(h+1)}}{1 - L_{pop,\mu}^2}\left(\frac{|\mathcal{S}||\mathcal{A}|}{N} + \frac{18K_\mu^2(H^2+2)}{N^2} + 2(H^2+2)\alpha^2\right),$$

for some absolute constant $C > 0$.

Proof. 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. In this step, we will assume N players of the FH-DG pursue policies $\{\pi^i\}_i = \{\pi_h^i\}_{i,h} \in \Pi^N$ such that $\pi^i = \pi$ for some $\pi \in \Pi$, and random variables $\{s_h^i, a_h^i\}_{i,h' \leqslant h}$, $\boldsymbol{\rho}_h \in (\mathcal{S}\times\mathcal{A})^N$ are generated according to the finite player dynamics.

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

$$\begin{aligned}
\mathbb{E}\left[\|\widehat{\mu}_0 - \mu_0\|_2^2\right] \leqslant& \mathbb{E}\left[\|\widehat{\mu}_0 - \rho_0 \cdot \pi_0\|_2^2\right] \\
\leqslant& \sum_{s,a}\mathbb{E}\left[\left(\frac{1}{N}\sum_{i=1}^N \mathbb{1}_{s_0^i = s, a_0^i = a} - (\rho_0 \cdot \pi_0)(s,a)\right)^2\right] \\
\leqslant& \sum_{s,a}\frac{1}{N^2}\mathbb{E}\left[\sum_{i=1}^N\left(\mathbb{1}_{s_0^i = s, a_0^i = a} - (\rho_0 \cdot \pi_0)(s,a)\right)^2\right] \\
\leqslant& \frac{|\mathcal{S}||\mathcal{A}|}{N},
\end{aligned}$$

due to the fact that $\mathbb{1}_{s_0^i = s, a_0^i = a}$ are independent, bounded random variables, and that we have $\mathbb{E}\left[\mathbb{1}_{s_0^i = s, a_0^i = a}\right] = \rho_0(s)\pi(s,a) = \mu_0(s,a)$.

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

$$\mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mu_{h+1}\|_2^2 \,|\mathcal{F}_h\right] \tag{13}$$

$$\leqslant \mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right]\|_2^2 \,|\mathcal{F}_h\right] + \mathbb{E}\left[\|\mu_{h+1} - \mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right]\|_2^2 \,|\mathcal{F}_h\right] \tag{14}$$

$$\begin{aligned}\leqslant\ & \mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right]\|_2^2 \,|\mathcal{F}_h\right] \\ & + (1+\delta_h^{-1})\mathbb{E}\left[\|\Gamma(\widehat{\mu}_h, \pi_h) - \mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right]\|_2^2 \,|\mathcal{F}_h\right] + (1+\delta_h)\mathbb{E}\left[\|\Gamma(\widehat{\mu}_h, \pi_h) - \mu_{h+1}\|_2^2 \,|\mathcal{F}_h\right]\end{aligned} \tag{15}$$

$$\begin{aligned}\leqslant & \underbrace{\mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right]\|_2^2 \,|\mathcal{F}_h\right]}_{(\triangle)} + 2(1+\delta_h^{-1})\underbrace{\mathbb{E}\left[\|\mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right] - \Xi(\boldsymbol{\rho}_h, \pi_h)\|_2^2 \,|\mathcal{F}_h\right]}_{(\square)} \\ & + 2(1+\delta_h^{-1})\underbrace{\mathbb{E}\left[\|\Xi(\boldsymbol{\rho}_h, \pi_h) - \Gamma(\widehat{\mu}_h, \pi_h)\|_2^2 \,|\mathcal{F}_h\right]}_{(\star)} + (1+\delta_h)\underbrace{\mathbb{E}\left[\|\Gamma(\widehat{\mu}_h, \pi_h) - \mu_{h+1}\|_2^2 \,|\mathcal{F}_h\right]}_{(\heartsuit)}\end{aligned} \tag{16}$$

where inequalities 15 and 16 follow from applications of Young's inequality, where $\delta_h > 0$ is a positive value to be determined later. We upper bound the four terms separately as in the proof of Theorem 1. For $(\triangle)$, it holds that

$$(\triangle) = \mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right]\|_2^2 \,|\mathcal{F}_h\right] \leqslant \frac{|\mathcal{S}||\mathcal{A}|}{N},$$

since each $\widehat{\mu}_{h+1}(s)$ is an average of independent subgaussian random variables given $\mathcal{F}_h$. Specifically, each indicator is bounded $\mathbb{1}_{s_{h+1}^i = s, a_{h+1}^i = a} \in [0,1]$ almost surely.

Next, for the term $(\square)$,

$$\begin{aligned}(\square) &= \mathbb{E}\left[\|\mathbb{E}\left[\widehat{\mu}_{h+1}\,|\mathcal{F}_h\right] - \Xi(\boldsymbol{\rho}_h, \bar{\pi}_{h+1})\|_2^2 \,|\mathcal{F}_h\right] \\ &\leqslant \frac{1}{N^2}\,\mathbb{E}\left[\left\|\sum_{i=1}^N P^i(\cdot|s_h^i, a_h^i, \boldsymbol{\rho}^{-i})\cdot\pi_{h+1} - \sum_{i=1}^N P(\cdot|s_h^i, a_h^i, \sigma(\boldsymbol{\rho}^{-i}))\cdot\pi_{h+1}\right\|_1^2 \middle| \mathcal{F}_h\right] \\ &\leqslant \frac{1}{N^2}\,\mathbb{E}\left[\left(\sum_{i=1}^N \|P^i(\cdot|s_h^i, a_h^i, \boldsymbol{\rho}^{-i}) - P(\cdot|s_h^i, a_h^i, \sigma(\boldsymbol{\rho}^{-i}))\|_1\right)^2 \middle| \mathcal{F}_h\right]\end{aligned}$$

By the α-symmetry condition, it follows that $(\square) \leqslant \alpha^2$.

For $(\star) = \mathbb{E}[\|\Xi(\boldsymbol{\rho}_h, \pi_h\}_i) - \Gamma(\widehat{\mu}_h, \pi_h)\|_2^2 | \mathcal{F}_h]$,

$$\begin{aligned}(\star) &\leqslant \mathbb{E}\left[\|\Xi(\boldsymbol{\rho}_h, \bar{\pi}_{h+1}) - \Gamma(\widehat{\mu}_h, \bar{\pi}_{h+1})\|_1^2 \,|\mathcal{F}_h\right] \\ &= \mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^N P(\cdot|s_h^i, a_h^i, \sigma(\boldsymbol{\rho}_h^{-i}))\cdot\bar{\pi}_{h+1} - \sum_{s',a'}\widehat{\mu}_h(s',a')P(\cdot|s',a',\widehat{\mu}_h)\cdot\bar{\pi}_{h+1}\right\|_1^2 \middle| \mathcal{F}_h\right] \\ &= \frac{1}{N^2}\,\mathbb{E}\left[\left\|\sum_{i=1}^N P(\cdot|s_h^i, a_h^i, \sigma(\boldsymbol{\rho}_h^{-i}))\cdot\bar{\pi}_{h+1} - \sum_{i=1}^N P(\cdot|s_h^i, a_h^1, \widehat{\mu}_h)\cdot\bar{\pi}_{h+1}\right\|_1^2 \middle| \mathcal{F}_h\right].\end{aligned}$$

The vectors $(N-1)\sigma(\boldsymbol{\rho}_h^{-i}), N\widehat{\mu}_h$ can differ by only 1 in one component due to the i-th agent being excluded from the former, it holds that

$$\|N\sigma(\boldsymbol{\rho}_h^{-i}) - N\widehat{\mu}_h\|_1 \leqslant \|(N-1)\sigma(\boldsymbol{\rho}_h^{-i}) - N\widehat{\mu}_h\|_1 + \|\sigma(\boldsymbol{\rho}_h^{-i})\|_1 \leqslant 3,$$

therefore for any s, a,

$$\|P(\cdot|s,a,\sigma(\boldsymbol{\rho}_h^{-i})) - P(\cdot|s,a,\widehat{\mu}_h)\|_1 \leqslant \frac{3K_\mu}{N}$$

almost surely and $(\star)$ is further upper bounded by $(\star) \leqslant \frac{9K_\mu^2}{N^2}$.

Finally, the last term $(\heartsuit)$ can be upper bounded using the Lipschitz condition on Γ, namely:

$$(\heartsuit) = \mathbb{E}\left[\|\Gamma(\widehat{\mu}_h, \pi_h) - \Gamma(\mu_h^\pi, \pi_h)\|_2^2 \,|\mathcal{F}_h\right] \leqslant L_{pop,\mu}^2\|\widehat{\mu}_h - \mu_h\|_2^2.$$

To conclude, merging the bounds on the three terms in Inequality (6) and taking the expectations we obtain:

$$\mathbb{E}\left[\|\widehat{\mu}_{h+1} - \mu_{h+1}\|_2^2\right] \leqslant (1+\delta_h) L_{pop,\mu}^2 \mathbb{E}\left[\|\widehat{\mu}_h - \mu_h\|_2^2\right] + \frac{|\mathcal{S}||\mathcal{A}|}{N} + 2(1+\delta_h^{-1})\frac{9K_\mu^2}{N^2} + 2(1+\delta_h^{-1})\alpha^2.$$

Induction on h yields the bound for all h:

$$\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h\|_2^2\right] \leqslant \sum_{h'=0}^{h} L_{pop,\mu}^{2(h-h')} \left(\prod_{h''=h'+1}^{h} (1+\delta_{h''})\right) \left(\frac{|\mathcal{S}||\mathcal{A}|}{N} + 2(1+\delta_{h'}^{-1})\frac{9K_\mu^2}{N^2} + 2(1+\delta_{h'}^{-1})\alpha^2\right).$$

Here, we specify $\delta_h := (1+h^2)^{-1}$, for which it holds that

$$\prod_{h=0}^{\infty}(1+\delta_h) = \prod_{h=0}^{\infty}(1+(1+h^2)^{-1}) \leqslant C,$$

for an absolute constant $C < 10$. Then,

$$\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h\|_2^2\right] \leqslant \sum_{h'=0}^{h} L_{pop,\mu}^{2(h-h')} C \left(\frac{|\mathcal{S}||\mathcal{A}|}{N} + 2(H^2+2)\frac{9K_\mu^2}{N^2} + 2(H^2+2)\alpha^2\right).$$

which yields the stated upper bound of the lemma

$$\mathbb{E}\left[\|\widehat{\mu}_h - \mu_h\|_2^2\right] \leqslant \frac{1-L_{pop,\mu}^{2(h+1)}}{1-L_{pop,\mu}^2} C \left(\frac{|\mathcal{S}||\mathcal{A}|}{N} + \frac{18K_\mu^2(H^2+2)}{N^2} + 2(H^2+2)\alpha^2\right), \tag{17}$$

where we adopt the convenient shorthand $\frac{1-L_{pop,\mu}^{2(h+1)}}{1-L_{pop,\mu}^2} := h$ if $L_{pop,\mu} = 1$. □

Note that while the Lipschitz modulus used in Lemma 11 is with respect to the $\|\cdot\|_2$ norm, Lemma 2 readily guarantees that this will hold.

C.1 Extended Proof of Theorem 2

Let $\boldsymbol{\mu} = \{\mu_h\}_h := \Lambda(\pi)$, and note that by the proof of Theorem 1, it holds that (Inequality 10)

$$\begin{aligned} A_h &:= \|\,\mathbb{P}[s_h^1 = \cdot, a_h^1 = \cdot] - \mathbb{P}[s_h = \cdot, a_h = \cdot]\|_1 \\ &\leqslant \|\,\mathbb{P}[s_h^1 = \cdot] - \mathbb{P}[s_h = \cdot]\|_1 \\ &\leqslant L_{pop,\mu} \sum_{h=0}^{h-1} \frac{1-L_{pop,\mu}^{h+1}}{1-L_{pop,\mu}} \left(|\mathcal{S}||\mathcal{A}|\sqrt{\frac{\pi}{2N}} + \frac{5K_\mu}{N} + \alpha\right) + (h-1)\left(\alpha + \frac{3K_\mu}{N}\right). \end{aligned}$$

Likewise, by Lemma 11,

$$\begin{aligned} B_h &:= \mathbb{E}[\|\mu_h - \sigma(\boldsymbol{\rho}_{m,h}^{-1})\|_1^2] \\ &:= \mathbb{E}[\|\mu_h - \sigma(\boldsymbol{\rho}_{m,h}^{-1})\|_2^2]|\mathcal{S}||\mathcal{A}| \\ &\leqslant \frac{1-L_{pop,\mu}^{2(h+1)}}{1-L_{pop,\mu}^2} 2C \left(\frac{|\mathcal{S}||\mathcal{A}|}{N} + \frac{18K_\mu^2(H^2+2)}{N^2} + 2(H^2+2)\alpha^2\right)|\mathcal{S}||\mathcal{A}| + \frac{18|\mathcal{S}||\mathcal{A}|}{N^2}. \end{aligned}$$

Will will commonly utilize the bounds $\widehat{Q}_h^m \in [0, Q_{\max}], Q_h^{\tau,\pi} \in [0, Q_{\max}]$ almost surely for $Q_{\max} := H(1+\log|\mathcal{A}|)$, as the one-step rewards are bounded in range $[0,1]$ and the policy entropy has trivial upper bound $\log|\mathcal{A}|$. Denote the marginal probabilities of $s_{m,h}^1, a_{m,h}^1$ (which is i.i.d. for all m) as $p_h \in \Delta_{\mathcal{S}\times\mathcal{A}}$, which clearly does not depend on epoch m as the same policies are deployed at each TD learning round.

We outline the proof strategy into different steps as follows:

- **Step 1.** Analyze the algorithm for the Q-values at time step $H-1$, that is, show that in expectation $\|Q^{\tau,\pi}_{H-1} - \widehat{Q}^m_{H-1}\|^2_{p_{H-1}}$ decreases with $\mathcal{O}(1/m)$ over epochs, up to a small bias term.
- **Step 2.** Assuming that the error at some time h decreases with rate $\mathcal{O}(1/m)$, show that the error $\|Q^{\tau,\pi}_{h-1} - \widehat{Q}^m_{h-1}\|^2_{p_{H-1}}$ also decreases with rate $\mathcal{O}(1/m)$, showing that the magnification in the constants are not too large.
- **Step 3.** Conclude the statement of the theorem by induction.

Step 1. We will first analyze the evolution of $\widehat{Q}^m_{H-1}$. By definition, it holds that

$$Q^{\tau,\pi}_{H-1}(s,a) = R(s,a,\mu_{H-1}) + \mathcal{H}(\pi_{H-1}(\cdot|s)).$$

In other words, there is no bootstrapping and the stochastic error does not have a dependence on future biased estimates. Firstly, if $s^1_{m,H-1} = s, a^1_{m,H-1} = a, \boldsymbol{\rho}_{m,H-1} = \boldsymbol{\rho}$, then it holds almost surely that

$$\begin{aligned}
&Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^{m+1}_{H-1}(s,a)\\
&= Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a) - \eta_m(r^i_{m,H-1} - \widehat{Q}^m_h(s,a))\\
&= (1-\eta_m)(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a)) - \eta_m(r^i_{m,H-1} + \mathcal{H}(\pi_{H-1}(\cdot|s)) - Q^{\tau,\pi}_{H-1}(s,a))\\
&= (1-\eta_m)(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a)) - \eta_m((R^1(s,a,\boldsymbol{\rho}^{-1}_{m,H-1}) - R(s,a,\mu_{H-1})),
\end{aligned}$$

as the entropy term $\mathcal{H}(\pi_{H-1}(\cdot|s))$ cancels out. Denote the σ-algebra

$$\mathcal{F}^m_{s,a} := \mathcal{F}\{\{s_{m',h}, a_{m',h}\}_{m'<m}, s_{m,H-1} = s, a_{m,H-1} = a\}$$

for any fixed s, a. Then, noting that $\widehat{Q}^m_{H-1}(s,a)$ is $\mathcal{F}^m_{s,a}$-measurable, we have the inequalities

$$\begin{aligned}
&\mathbb{E}[(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^{m+1}_{H-1}(s,a))^2|\mathcal{F}^m_{s,a}]\\
&= \mathbb{E}\left[\left((1-\eta_m)(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a)) - \eta_m(R^1(s,a,\boldsymbol{\rho}^{-1}_{m,H-1}) - R(s,a,\mu_{H-1}))\right)^2\middle|\mathcal{F}^m_{s,a}\right]\\
&= (1-\eta_m)^2(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a))^2\\
&\quad + 2(1-\eta_m)\eta_m(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a))\,\mathbb{E}[R^1(s,a,\boldsymbol{\rho}^{-1}_{m,H-1}) - R(s,a,\mu_{H-1})|\mathcal{F}^m_{s,a}]\\
&\quad + \eta_m^2\,\mathbb{E}[(R^1(s,a,\boldsymbol{\rho}^{-1}_{m,H-1}) - R(s,a,\mu_{H-1}))^2|\mathcal{F}^m_{s,a}]
\end{aligned}$$

We use Young's inequality and using the fact that rewards are bounded in $[0,1]$,

$$\begin{aligned}
&\mathbb{E}[(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^{m+1}_{H-1}(s,a))^2|\mathcal{F}^m_{s,a}]\\
&\leqslant (1-2\eta_m)(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a))^2 + \eta_m(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a))^2\\
&\quad + \eta_m\,\mathbb{E}[(R^1(s,a,\boldsymbol{\rho}^{-1}_{m,H-1}) - R(s,a,\mu_{H-1}))^2|\mathcal{F}^m_{s,a}] + (1+Q^2_{\max})\eta_m^2\\
&\leqslant (1-\eta_m)(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a))^2 + \eta_m\,\mathbb{E}[(R^1(s,a,\boldsymbol{\rho}^{-1}_{m,H-1}) - R(s,a,\mu_{H-1}))^2|\mathcal{F}^m_{s,a}]^2\\
&\quad + (1+2Q^2_{\max})\eta_m^2\\
&\leqslant (1-\eta_m)(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a))^2 + \eta_m\,\mathbb{E}[2\beta^2 + 2L^2_\mu\|\sigma(\boldsymbol{\rho}^{-1}_{m,H-1}) - \mu_{H-1}\|^2_2|\mathcal{F}^m_{s,a}]\\
&\quad + (1+2Q^2_{\max})\eta_m^2
\end{aligned}$$

We then take expectations and use the law of total expectation to obtain the bound:

$$\begin{aligned}
&\mathbb{E}[(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^{m+1}_{H-1}(s,a))^2]\\
&\leqslant (1-\eta_m)p_{H-1}(s,a)\,\mathbb{E}[(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a))^2]\\
&\quad + \eta_m p_{H-1}(s,a)\,\mathbb{E}[(R^1(s,a,\boldsymbol{\rho}^{-1}_{m,H-1}) - R(s,a,\mu_{H-1}))^2|s_{m,H-1} = s, a_{m,H-1} = a]\\
&\quad + p_{H-1}(s,a)(1+2Q^2_{\max})\eta_m^2\\
&\quad + (1-p_{H-1}(s,a))\,\mathbb{E}[(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^{m+1}_{H-1}(s,a))^2]\\
&\leqslant (1-p_{H-1}(s,a)\eta_m)\,\mathbb{E}[(Q^{\tau,\pi}_{H-1}(s,a) - \widehat{Q}^m_{H-1}(s,a))^2]\\
&\quad + \eta_m p_{H-1}(s,a)\,\mathbb{E}[2\beta^2 + 2L^2_\mu\|\sigma(\boldsymbol{\rho}^{-1}_{m,H-1}) - \mu_{H-1}\|^2_2|s_{m,H-1} = s, a_{m,H-1} = a]\\
&\quad + p_{H-1}(s,a)(1+2Q^2_{\max})\eta_m^2
\end{aligned}$$

Summing this inequality over all state-action pairs with weight p_{H-1}, we obtain

$$\begin{aligned}
&\mathbb{E}[\|Q^{\tau,\pi}_{H-1} - \widehat{Q}^{m+1}_{H-1}\|^2_{p_{H-1}}] \\
&\leqslant (1-\delta\eta_m)\,\mathbb{E}[\|Q^{\tau,\pi}_{H-1} - \widehat{Q}^{m}_{H-1}\|^2_{p_{H-1}}] \\
&\quad + \sum_{s,a} \eta_m p_{H-1}(s,a)\,\mathbb{E}[2\beta^2 + 2L_\mu^2 \|\sigma(\boldsymbol{\rho}^{-1}_{m,H-1}) - \mu_{H-1}\|_2^2 | s_{m,H-1}=s, a_{m,H-1}=a] \\
&\quad + (1+2Q^2_{\max})\eta_m^2 \\
&= (1-\delta\eta_m)\,\mathbb{E}[\|Q^{\tau,\pi}_{H-1} - \widehat{Q}^{m}_{H-1}\|^2_{p_{H-1}}] + (1+2Q^2_{\max})\eta_m^2 \\
&\quad + 2\eta_m\beta^2 + 2\eta_m L_\mu^2\,\mathbb{E}[\|\mu_{H-1} - \sigma(\boldsymbol{\rho}^{-1}_{m,H-1})\|_2^2] \\
&\leqslant (1-\delta\eta_m)\,\mathbb{E}[\|Q^{\tau,\pi}_{H-1} - \widehat{Q}^{m}_{H-1}\|^2_{p_{H-1}}] + (1+2Q^2_{\max})\eta_m^2 + 2\eta_m(\beta^2 + L_\mu^2 B_{H-1}).
\end{aligned}$$

We expand this recursive inequality as follows. Define the shorthand notation $\Pi_m^{m'} := \prod_{k=m}^{m'}(1-\delta\eta_k)$. Then, for any $M > 0$,

$$\begin{aligned}
&\mathbb{E}[\|Q^{\tau,\pi}_{H-1} - \widehat{Q}^{M}_{H-1}\|^2_{p_{H-1}}] \\
&\leqslant \Pi_0^{M-1} + (1+2Q^2_{\max}) \sum_{m=0}^{M-1} \eta_m^2 \Pi_{m+1}^{M-1} + 2\sum_{m=0}^{M-1} \eta_m(\beta^2 + L_\mu^2 B_{H-1})\Pi_{m+1}^{M-1} \\
&\leqslant \Pi_0^{M-1} + (1+2Q^2_{\max})\eta_{M-1}^2 + 2\eta_{M-1}(\beta^2 + L_\mu^2 B_{H-1}) \\
&\qquad + (1+2Q^2_{\max}) \sum_{m=0}^{M-2} \eta_m^2 \Pi_{m+1}^{M-1} + 2\sum_{m=0}^{M-2} \eta_m(\beta^2 + L_\mu^2 B_{H-1})\Pi_{m+1}^{M-1}. \qquad (18)
\end{aligned}$$

We bound the multiplicative terms $\Pi_m^{m'}$. Assuming that η_m is of the form $\eta_m = \frac{u}{v+m}$, for any $m \leqslant m'$, we have that

$$\begin{aligned}
\Pi_m^{m'} &= \prod_{k=m}^{m'}(1-\delta\eta_k) \leqslant \exp\{-\delta \sum_{k=m}^{m'} \eta_k)\} \\
&\leqslant \exp\{-\delta \sum_{k=m}^{m'} \frac{u}{v+k})\} \leqslant \exp\{-\delta u \log \frac{m'+v}{m+v-1}\} \\
&\leqslant \left(\frac{m+v-1}{m'+v}\right)^{\delta u}
\end{aligned}$$

using Lemma 9. Taking the values $u = v = 2\delta^{-1}$, this reduces to

$$\Pi_m^{m'} \leqslant \left(\frac{m+u-1}{m'+u}\right)^2,$$

Placing this in Inequality 18 for the two terms appearing $\Pi_{m+1}^{M-1}, \Pi_0^{M-1}$, we obtain

$$\begin{aligned}
&\mathbb{E}[\|Q^{\tau,\pi}_{H-1} - \widehat{Q}^{M}_{H-1}\|^2_{p_{H+1}}] \\
&\leqslant \left(\frac{u-1}{M+u-1}\right)^2 + (1+2Q^2_{\max})\left(\frac{u}{M+u-1}\right)^2 + 2\left(\frac{u}{M+u-1}\right)(\beta^2 + L_\mu^2 B_{H-1}) \\
&\quad + (1+2Q^2_{\max}) \sum_{m=0}^{M-2} \left(\frac{u}{m+u}\right)^2 \left(\frac{m+u}{M+u-1}\right)^2 \\
&\quad + 2(\beta^2 + L_\mu^2 B_{H-1}) \sum_{m=0}^{M-2} \left(\frac{u}{m+v}\right)\left(\frac{m+u}{M+u-1}\right)^2 \\
&\leqslant \frac{C_1 u}{(M+u-1)} + C_2 u(\beta^2 + L_\mu^2 B_{H-1}),
\end{aligned}$$

for some absolute constants C_1, C_2. This inequality concludes the convergence result for the Q values at time step $H-1$, showing a rate of convergence $\mathcal{O}(1/M)$ over M epochs up to a bias of $\mathcal{O}(\beta^2 + \alpha^2 + 1/N)$.

Step 2. Next, we analyze the case $h < H-1$. Again under the observation that if $s^1_{m,h} = s, a^1_{m,h} = a, s^1_{m,h+1} = s', a^1_{m,h+1} = a'$, then it holds almost surely that

$$\begin{aligned}
&Q_h^{\tau,\pi}(s,a) - \widehat{Q}_h^{m+1}(s,a) \\
&= Q_h^{\tau,\pi}(s,a) - \widehat{Q}_h^{m}(s,a) - \eta_m(\widehat{Q}_{h+1}^{m}(s',a') + r^1_{m,h} - \widehat{Q}_h^{m}(s,a)) \\
&= (1-\eta_m)(Q_h^{\tau,\pi}(s,a) - \widehat{Q}_h^{m}(s,a)) - \eta_m(\widehat{Q}_{h+1}^{m}(s',a') + r^1_{m,h} - Q_h^{\tau,\pi}(s,a)),
\end{aligned}$$

since again as the entropy terms $\mathcal{H}(\pi_h(\cdot|s))$ cancel. Defining the induced σ-algebra

$$\mathcal{F}_h^m := \mathcal{F}\{\{s_{m',h}, a_{m',h}\}_{m'<m}, s_{m,h} = s, a_{m,h} = a\}.$$

Note that with respect to $\mathcal{F}_h^m$, $\widehat{Q}_{h'}^m$ is measurable for any h' as $\widehat{Q}_{h'}^m$ only depends on episodes previous. we have the lower bound:

$$\begin{aligned}
&\mathbb{E}[(Q_h^{\tau,\pi}(s,a) - \widehat{Q}_h^{m+1}(s,a))^2 | \mathcal{F}_h^m] \\
&= (1-\eta_m)^2\, \mathbb{E}[(Q_h^{\tau,\pi}(s,a) - \widehat{Q}_h^{m}(s,a))^2 | \mathcal{F}_h^m] \\
&\quad + (\eta_m)^2\, \mathbb{E}[(\widehat{Q}_{h+1}^{m}(s',a') + r^1_{m,h} - Q_h^{\tau,\pi}(s,a))^2 | \mathcal{F}_h^m] \\
&\quad - 2\eta_m(Q_h^{\tau,\pi}(s,a) - \widehat{Q}_h^{m}(s,a))\, \mathbb{E}[\underbrace{\widehat{Q}_{h+1}^{m}(s^1_{m,h+1}, a^1_{m,h+1}) + r^1_{m,h} - Q_h^{\tau,\pi}(s,a)}_{\triangle} | \mathcal{F}_h^m] \\
&= (1-\eta_m)^2 (Q_h^{\tau,\pi}(s,a) - \widehat{Q}_h^{m}(s,a))^2 + (\eta_m)^2 Q_{\max}^2 \\
&\quad - 2(\eta_m)(Q_h^{\tau,\pi}(s,a) - \widehat{Q}_h^{m}(s,a))\, \mathbb{E}[\triangle | \mathcal{F}_h^m], \\
&\leqslant (1-2\eta_m)(Q_h^{\tau,\pi}(s,a) - \widehat{Q}_h^{m}(s,a))^2 + 2(\eta_m)^2 Q_{\max}^2 \\
&\quad + (\eta_m)(Q_h^{\tau,\pi}(s,a) - \widehat{Q}_h^{m}(s,a))^2 + (\eta_m)\, \mathbb{E}[\triangle^2 | \mathcal{F}_h^m], \\
&\leqslant (1-\eta_m)(Q_h^{\tau,\pi}(s,a) - \widehat{Q}_h^{m}(s,a))^2 + 2(\eta_m)^2 Q_{\max}^2 + (\eta_m)\, \mathbb{E}[\triangle | \mathcal{F}_h^m]^2, \qquad (19)
\end{aligned}$$

as $\widehat{Q}_h^m$ is $\mathcal{F}_m$ measurable, using Young's inequality in the last line. Then, taking expectations on both sides,

$$\begin{aligned}
&\mathbb{E}\left[\left(Q_h^{\tau,\pi}(s',a') - \widehat{Q}_h^{m+1}(s',a')\right)^2\right] \\
&\leqslant (1 - p_h(s',a')\eta_m)\, \mathbb{E}[Q_h^{\tau,\pi}(s',a') - \widehat{Q}_h^{m}(s',a'))^2] \\
&\quad + 2p_h(s',a')(\eta_m)^2 Q_{\max}^2 + p_h(s',a')\eta_m\, \mathbb{E}[(\Delta^2)] \qquad (20)
\end{aligned}$$

The last bias term due to bootstrapping and finite population bias we bound separately. We decompose $(\triangle)$ as follows.

$$\begin{aligned}
\mathbb{E}[\triangle | \mathcal{F}_h^m]^2 \leqslant &\left(1 + \frac{1}{(H-h)^2}\right) \mathbb{E}[\widehat{Q}_{h+1}^{m}(s',a') - Q_{h+1}^{\tau,\pi}(s',a') \,|\mathcal{F}_h^m] \\
&+ 2(H-h+1)^2 \left| \mathbb{E}[Q_{h+1}^{\tau,\pi}(s',a') \,|\mathcal{F}_h^m] - \sum_{\bar{s},\bar{a}} P(\bar{s},\bar{a},\mu_h) Q_{h+1}^{\tau,\pi}(\bar{s},\bar{a}) \right|^2 \\
&+ 2(H-h+1)^2 |\, \mathbb{E}[r^1_{m,h} - R(s,a,\mu_h) \,|\mathcal{F}_h^m]|^2
\end{aligned}$$

The three terms are upper-bounded by the inequalities in expectation:

$$
\begin{aligned}
&|\,\mathbb{E}[\widehat{Q}^m_{h+1}(s^1_{m,h+1},a^1_{m,h+1}) - Q^{\tau,\pi}_{h+1}(s^1_{m,h+1},a^1_{m,h+1})\,|\mathcal{F}^m_h]|^2 \\
&\qquad \leqslant \mathbb{E}[|\widehat{Q}^m_{h+1}(s^1_{m,h+1},a^1_{m,h+1}) - Q^{\tau,\pi}_{h+1}(s^1_{m,h+1},a^1_{m,h+1})|^2|\mathcal{F}^m_h] \\
&\qquad \leqslant \|\widehat{Q}^m_{h+1} - Q^{\tau,\pi}_{h+1}\|^2_{p_{h+1}(\cdot|s,a)} \\
&\left|\mathbb{E}[Q^{\tau,\pi}_{h+1}(s^1_{m,h+1},a^1_{m,h+1})\,|\mathcal{F}^m_h] - \sum_{\bar{s},\bar{a}} P(\bar{s},\bar{a}|s,a,\mu_h)Q^{\tau,\pi}_{h+1}(\bar{s},\bar{a})\right|^2 \\
&\qquad \leqslant \frac{Q^2_{\max}}{4}\,\mathbb{E}\left[2\alpha^2 + 2K^2_\mu\|\mu_h - \sigma(\boldsymbol{\rho}^{-1}_{m,h})\|^2_1|\mathcal{F}^m_h\right] \\
&\left|\mathbb{E}[r^1_{m,h} - R(s,a,\mu)\,|\mathcal{F}^m_h]\right|^2 \\
&\qquad = 2|\,\mathbb{E}[R^1(s,a,\boldsymbol{\rho}^{-1}_{m,h}) - R(s,a,\sigma(\boldsymbol{\rho}^{-1}_{m,h}))\,|\mathcal{F}^m_h]|^2 + 2|\,\mathbb{E}[R(s,a,\sigma(\boldsymbol{\rho}^{-1}_{m,h})) - R(s,a,\mu_h)\,|\mathcal{F}^m_h]|^2 \\
&\qquad \leqslant 2\beta^2 + 2L^2_\mu\,\mathbb{E}[\|\mu_h - \sigma(\boldsymbol{\rho}^{-1}_{m,h})\|^2_1|\mathcal{F}^m_h]
\end{aligned}
$$

Therefore, we conclude by an application of Young's inequality that almost surely,

$$
\begin{aligned}
|\triangle|^2 \leqslant &\left(1 + \frac{1}{(H-h)^2}\right)\|\widehat{Q}^m_{h+1} - Q^{\tau,\pi}_{h+1}\|^2_{p_{h+1}(\cdot|s,a)} \\
&+ (H-h+1)^2Q^2_{\max}[\alpha^2 + \beta^2 + (2K^2_\mu + 4L^2_\mu)\,\mathbb{E}[\|\mu_h - \sigma(\boldsymbol{\rho}^{-1}_{m,h})\|^2_1|\mathcal{F}^m_h]].
\end{aligned}
$$

We place this in Inequality 20 to obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left(Q^{\tau,\pi}_h(s',a') - \widehat{Q}^{m+1}_h(s',a')\right)^2\right] \\
\leqslant&(1-\delta\eta_m)\,\mathbb{E}[Q^{\tau,\pi}_h(s',a') - \widehat{Q}^m_h(s',a'))^2] + 2\eta^2_m Q^2_{\max} \\
&+ \eta_m\left(1 + \frac{1}{(H-h)^2}\right)\mathbb{E}[\|\widehat{Q}^m_{h+1} - Q^{\tau,\pi}_{h+1}\|^2_{p_{h+1}}] \\
&+ \eta_m(H-h+1)^2Q^2_{\max}[\alpha^2 + \beta^2 + (2K^2_\mu + 4L^2_\mu)\,\mathbb{E}[\|\mu_h - \sigma(\boldsymbol{\rho}^{-1}_{m,h})\|^2_1]]
\end{aligned}
$$

Finally, taking a weighted sum of both sides of the inequality over s', a' with weights p_h,

$$
\begin{aligned}
\mathbb{E}\left[\|Q^{\tau,\pi}_h - \widehat{Q}^{m+1}_h\|^2_{p_h}\right] \leqslant&(1-\delta\eta_m)\,\mathbb{E}[\|Q^{\tau,\pi}_h - \widehat{Q}^{m+1}_h\|^2_{p_h}] + 2\eta^2_m Q^2_{\max} \\
&+ \eta_m(H-h+1)^2Q^2_{\max}[\alpha^2 + \beta^2 + (2K^2_\mu + 4L^2_\mu)B_h]
\end{aligned}
$$

Expanding this recursive inequality and using the same notation as in Step 1 for the multiplicative terms, also taking the inductive assumption that $\mathbb{E}[\|\widehat{Q}^m_{h+1} - Q^{\tau,\pi}_{h+1}\|^2_{p_{h+1}}] \leqslant \frac{G_1}{2\delta^{-1}+m-1} + G_2$ for some G_1, G_2 which depends on problem parameters but not on m, we have the final inequality:

$$
\begin{aligned}
\mathbb{E}\left[\|Q^{\tau,\pi}_h - \widehat{Q}^{m+1}_h\|^2_{p_h}\right] \leqslant&\Pi^{M-1}_{m=0}\,\mathbb{E}[\|Q^{\tau,\pi}_h - \widehat{Q}^0_h\|^2_{p_h}] + \frac{C_3u}{M-2+v} \\
&+ 2\sum_{m=0}^{M-2}(\eta_m)^2Q^2_{\max}\Pi^{M-1}_{m+1} \\
&+ \sum_{m=0}^{M-2}\eta_m(H-h+1)^2Q^2_{\max}[\alpha^2 + \beta^2 + (2K^2_\mu + 4L^2_\mu)B_h]\Pi^{M-1}_{m+1} \\
&+ \sum_{m=0}^{M-2}\eta_m\left(1 + \frac{1}{(H-h)^2}\right)\left(\frac{G_1}{2\delta^{-1}+m-1} + G_2\right)\Pi^{M-1}_{m+1}
\end{aligned}
$$

Once again as in Step 1, fixing the values $u = v = 2\delta^{-1}$,

$$\begin{aligned}\mathbb{E}\left[\|Q_h^{\tau,\pi} - \widehat{Q}_h^{m+1}\|_{p_h}^2\right] \leqslant & Q_{\max}^2 \left(\frac{u-1}{M+u-1}\right)^2 + \frac{C_3 u}{M-2+v} + \sum_{m=0}^{M-2} Q_{\max}^2 \frac{8\delta^{-2}}{(M+u-1)^2} \\ & + \sum_{m=0}^{M-2} (H-h+1)^2 Q_{\max}^2 [\alpha^2 + \beta^2 + (2K_\mu^2 + 4L_\mu^2) B_h] \frac{2(m+u)\delta^{-1}}{(M+u-1)^2} \\ & + \sum_{m=0}^{M-2} \left(1 + \frac{1}{(H-h)^2}\right) \left(\frac{G_1}{2\delta^{-1}+m-1} + G_2\right) \frac{2(m+u)\delta^{-1}}{(M+u-1)^2}\end{aligned}$$

Using the fact that $\sum_{m=1}^{M}(m+u) \approx M^2$ and $\sum_{m=1}^{M} c = cM$, for some absolute constants we have that

$$\begin{aligned}\mathbb{E}\left[\|Q_h^{\tau,\pi} - \widehat{Q}_h^{m+1}\|_{p_h}^2\right] \leqslant & \left(2\frac{Q_{\max}^2 \delta^{-1}}{M+2\delta^{-1}-1}\right)^2 + \frac{C_3\delta^{-1}}{M-2+2\delta^{-1}} + \frac{C_4 Q_{\max}^2 \delta^{-2}}{M+\delta^{-1}-1} \\ & + C_5 (H-h+1)^2 Q_{\max}^2 [\alpha^2 + \beta^2 + (2K_\mu^2 + 4L_\mu^2) B_h] \delta^{-1} \\ & + \left(1 + \frac{1}{(H-h)^2}\right) \left(\frac{C_6 \delta^{-1} G_1}{2\delta^{-1}+m-1} + C_7 G_2 \delta^{-1}\right) \qquad (21)\end{aligned}$$

Step 3. Finally, we conclude with the proof using Steps 1 and 2. By using Inequality 21, it readily follows that

$$\mathbb{E}\left[\|Q_h^{\tau,\pi} - \widehat{Q}_h^{m+1}\|_{p_h}^2\right] = \mathcal{O}\left(\frac{1}{M} + \alpha^2 + \beta^2 + \frac{1}{N}\right),$$

for all h, as the bound in Step 1 established the rate for time step $H-1$. We comment on the constants: Inequality 21 shows that in the worst case, there might be an exponential dependence on H, which might be fundamental.

D Extended Results on Monotonicity and Learning NE

D.1 Example: Asymmetric Congestion Games

Note that by symmetry in arguments, it follows that $\mathrm{Sym}\left(R^i(s,a,\cdot)\right) = R^i(s,a,\cdot)$ for any s, a, as

$$\begin{aligned}\mathrm{Sym}\left(R^i(s,a,\cdot)\right)(\boldsymbol{\rho}) = & \frac{1}{(N-1)!} \sum_{f \in \mathbb{S}_{N-1}} R^i(s,a,g(\boldsymbol{\rho})) \\ = & \frac{1}{(N-1)!} \sum_{f \in \mathbb{S}_{N-1}} \left(h_i(s,a,\sum_{j=1}^{N} \mathbb{1}_{g(\boldsymbol{\rho})_j=(s,a)}) + r_i(s,a)\right) \\ = & h_i\left(s,a,\sum_{j=1}^{N} \mathbb{1}_{\rho_j=(s,a)}\right) + r_i(s,a) \\ = & R^i(s,a,\boldsymbol{\rho})\end{aligned}$$

By simple computation, the population lifted rewards $\overline{\mathrm{Sym}}\left(R^i(s,a\cdot)\right)$ are given by

$$\overline{\mathrm{Sym}}\left(R^i(s,a\cdot)\right)(\mu) = h_i(s,a,N\mu(s,a)) + r_i(s,a), \quad \forall \mu \in \Delta_{\mathcal{S}\times\mathcal{A},N-1}.$$

We provide an extension to the continuum $\Delta_{\mathcal{S}\times\mathcal{A}}$ via linear interpolation in this case, while many other alternatives are possible. Take the function $\widetilde{h}_i : \mathcal{S}\times\mathcal{A}\times[0,1] \to [0,1]$ such that

$$\widetilde{h}_i(s,a,u) := (Nu - \lfloor Nu \rfloor) h_i(s,a,\lfloor Nu \rfloor) + (\lceil Nu \rceil - Nu) h_i(s,a,\lceil Nu \rceil).$$

The function is clearly monotonically decreasing in u. Furthermore, it is also Lipschitz continuous in u, as for any $u_1 > u_2$,

$$\widetilde{h}_i(s,a,u_1) - \widetilde{h}_i(s,a,u_2) \leqslant |u_1 - u_2|.$$

Finally, the asymmetry due to rewards can be upper bounded by

$$\beta \leqslant \sup_{s,a} \sup_{i,j} \sup_{k\in[N]} |h_i(s,a,k) - h_j(s,a,k)|.$$

D.2 Preliminaries for Learning Regularized Monotone MFG

We present several results required to establish convergence under monotonicity. We define the (entropy regularized) MFG value functions for an arbitrary population flow $\boldsymbol{\mu} \in \Delta_{\mathcal{S}\times\mathcal{A}}$ and policy $\pi \in \Pi$ as

$$V_h^\tau(s|\boldsymbol{\mu},\pi) := \mathbb{E}\left[\sum_{h'=h}^{H-1} R(s_{h'},a_{h'},\mu_{h'}) + \tau\mathcal{H}(\pi_{h'}(\cdot|s_{h'}))\middle| \begin{smallmatrix} s_0=s,\quad a_{h'}\sim\pi_{h'}(s_{h'}) \\ s_{h'+1}\sim P(s_{h'},a_{h'},\mu_{h'}) \end{smallmatrix}\right]$$

$$Q_h^\tau(s,a|\boldsymbol{\mu},\pi) := \mathbb{E}\left[\sum_{h'=h}^{H-1} R(s_{h'},a_{h'},\mu_{h'}) + \tau\mathcal{H}(\pi_{h'}(\cdot|s_{h'}))\middle| \begin{smallmatrix} s_h=s,\, a_h=a,\, s_{h'+1}\sim P(s_{h'},a_{h'},\mu_{h'}), \\ a_{h'}\sim\pi_{h'+1}(s_{h'+1}),\forall h'\geqslant h \end{smallmatrix}\right].$$

We define the regularized value function of the game similarly:

$$V^\tau(\boldsymbol{\mu},\pi) := \mathbb{E}\left[\sum_{h=0}^{H-1} R(s_h,a_h,\mu_h) + \tau\mathcal{H}(\pi_h(\cdot|s_h))\middle| \begin{smallmatrix} s_0\sim\rho_0,\quad a_h\sim\pi_h(s_h) \\ s_{h+1}\sim P(s_h,a_h,\mu_h) \end{smallmatrix}\right].$$

As expected, with these definitions it holds that

$$V_h^\tau(s|\boldsymbol{\mu},\pi) = \sum_{a\in\mathcal{A}} \pi_h(a|s) Q_h^\tau(s,a|\boldsymbol{\mu},\pi),$$
$$V^\tau(\boldsymbol{\mu},\pi) = \sum_{s\in\mathcal{S}} \rho_0(s) V_0^\tau(s|\boldsymbol{\mu},\pi).$$

We also define the quantity

$$q_h^\tau(s,a|\boldsymbol{\mu},\pi) := Q_h^\tau(s,a|\boldsymbol{\mu},\pi) - \mathcal{H}(\pi_h(\cdot|s)),$$

which corresponds to the more standard entropy regularized value function. Firstly, we provide several useful lemmas and definitions.

Definition 13 (Entropy regularized MFG-NE)**.** *For a given MFG $(\mathcal{S},\mathcal{A},\rho_0,H,P,R)$, a policy $\pi_\tau^* \in \Pi$ is called the τ-entropy regularized MFG-NE if it holds that*

$$\max_{\pi'\in\Pi} V^\tau(\Lambda(\pi_\tau^*),\pi') - V^\tau(\Lambda(\pi_\tau^*),\pi_\tau^*). \qquad \text{(Regularized MFG-NE)}$$

While entropy regularization will enable the convergence of our algorithm, it will also introduce a bias in terms of the original (unregularized) MFG. The next lemma quantifies this bias.

Lemma 12 (Regularization bias)**.** *Let $\boldsymbol{\mu} \in \Delta_{\mathcal{S}\times\mathcal{A}}$ and policy $\pi \in \Pi$ be arbitrary. Then, it holds that*

$$|V^\tau(\boldsymbol{\mu},\pi) - V(\boldsymbol{\mu},\pi)| \leqslant \tau H \log|\mathcal{A}|.$$

Furthermore, if π_τ^* is a τ-entropy regularized MFG-NE, then it is a $\tau H\log|\mathcal{A}|$-MFG-NE, that is,

$$\mathcal{E}(\pi_\tau^*) \leqslant 2H\log|\mathcal{A}|.$$

Proof.

$$\begin{aligned}
|V^\tau(\boldsymbol{\mu},\pi) - V(\boldsymbol{\mu},\pi)| &= \left|\mathbb{E}\left[\sum_{h=0}^{H-1} \tau\mathcal{H}(\pi(\cdot|s_h))\middle| \begin{smallmatrix} s_0\sim\rho_0,\quad a_h\sim\pi_h(s_h) \\ s_{h+1}\sim P(s_h,a_h,\mu_h) \end{smallmatrix}\right]\right| \\
&\leqslant \mathbb{E}\left[\sum_{h=0}^{H-1} \tau\left|\mathcal{H}(\pi(\cdot|s_h))\right|\middle| \begin{smallmatrix} s_0\sim\rho_0,\quad a_h\sim\pi_h(s_h) \\ s_{h+1}\sim P(s_h,a_h,\mu_h) \end{smallmatrix}\right] \leqslant \qquad H\tau\log|\mathcal{A}|,
\end{aligned}$$

since entropy is upper bounded by $|\mathcal{H}(\pi(\cdot|s_h))| \leqslant \log|\mathcal{A}|$. The bound for exploitability follows from:

$$\begin{aligned}
\mathcal{E}(\pi_\tau^*) &= \max_{\pi'\in\Pi} V(\Lambda(\pi_\tau^*),\pi') - V(\Lambda(\pi_\tau^*),\pi_\tau^*) \\
&= \max_{\pi'\in\Pi} V(\Lambda(\pi_\tau^*),\pi') - V^\tau(\Lambda(\pi_\tau^*),\pi') + V^\tau(\Lambda(\pi_\tau^*),\pi') - V(\Lambda(\pi_\tau^*),\pi_\tau^*) \\
&\leqslant \tau H\log|\mathcal{A}| + \max_{\pi'\in\Pi} V^\tau(\Lambda(\pi_\tau^*),\pi') - V(\Lambda(\pi_\tau^*),\pi_\tau^*) \\
&\leqslant \tau H\log|\mathcal{A}| + \max_{\pi'\in\Pi} V^\tau(\Lambda(\pi_\tau^*),\pi') - V^\tau(\Lambda(\pi_\tau^*),\pi_\tau^*) + V^\tau(\Lambda(\pi_\tau^*),\pi_\tau^*) - V(\Lambda(\pi_\tau^*),\pi_\tau^*) \\
&\leqslant 2\tau H\log|\mathcal{A}| + \max_{\pi'\in\Pi} V^\tau(\Lambda(\pi_\tau^*),\pi_\tau^*) - V^\tau(\Lambda(\pi_\tau^*),\pi_\tau^*) \\
&= 2\tau H\log|\mathcal{A}|.
\end{aligned}$$

□

We note that in our setting, a monotone MFG exhibits a unique MFG-NE, in fact, a unique regularized MFG-NE for any value of τ [65]. The above lemma shows that the bias introduced due to entropy regularization is of the order $\mathcal{O}(\tau)$ as expected.

Finally, we state several standard facts about monotone MFG, adapted from various works [65, 43, 42].

Lemma 13 (Monotone improvement lemma)**.** *Let $\boldsymbol{\mu}, \widetilde{\boldsymbol{\mu}} \in \Delta_{\mathcal{S}\times\mathcal{A}} \in \Delta^H_{\mathcal{S},\mathcal{A}}$ which are induced by policies $\pi, \widetilde{\pi} \in \Pi$ respectively. , that is $\Lambda(\widetilde{\pi}) = \widetilde{\boldsymbol{\mu}}$ and $\Lambda(\pi) = \boldsymbol{\mu}$. If the MFG is monotone, that is, if Definition 10 is satisfied, then it holds that:*

$$V^\tau\left(\boldsymbol{\mu}, \pi\right) + V^\tau\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\pi}\right) - V^\tau\left(\boldsymbol{\mu}, \widetilde{\pi}\right) - V^\tau\left(\widetilde{\boldsymbol{\mu}}, \pi\right) \leqslant 0.$$

Proof. The proof follows [65], except for the absence of a graphon. For monotone MFG, due to the assumption that P does not depend on μ, it holds that

$$\begin{aligned} V^\tau\left(\boldsymbol{\mu}, \pi\right) - V^\tau\left(\widetilde{\boldsymbol{\mu}}, \pi\right) &= V\left(\boldsymbol{\mu}, \pi\right) - V\left(\widetilde{\boldsymbol{\mu}}, \pi\right), \\ V^\tau\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\pi}\right) - V^\tau\left(\boldsymbol{\mu}, \widetilde{\pi}\right) &= V\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\pi}\right) - V\left(\boldsymbol{\mu}, \widetilde{\pi}\right). \end{aligned}$$

Furthermore, similar to [65], it holds that

$$\begin{aligned} V\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\pi}\right) - V\left(\boldsymbol{\mu}, \widetilde{\pi}\right) &= \sum_{s,a} \widetilde{\boldsymbol{\mu}}(s,a)(R(s,a,\widetilde{\boldsymbol{\mu}}_h) - R(s,a,\boldsymbol{\mu}_h)), \\ V\left(\boldsymbol{\mu}, \pi\right) - V\left(\widetilde{\boldsymbol{\mu}}, \pi\right) &= \sum_{s,a} \boldsymbol{\mu}(s,a)(R(s,a,\boldsymbol{\mu}_h) - R(s,a,\widetilde{\boldsymbol{\mu}}_h)). \end{aligned}$$

Then, using the monotonicity assumption on the rewards, it holds that

$$\begin{aligned} &V^\tau\left(\boldsymbol{\mu}, \pi\right) + V^\tau\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\pi}\right) - V^\tau\left(\boldsymbol{\mu}, \widetilde{\pi}\right) - V^\tau\left(\widetilde{\boldsymbol{\mu}}, \pi\right) \\ &= V\left(\boldsymbol{\mu}, \pi\right) + V\left(\widetilde{\boldsymbol{\mu}}, \widetilde{\pi}\right) - V\left(\boldsymbol{\mu}, \widetilde{\pi}\right) - V\left(\widetilde{\boldsymbol{\mu}}, \pi\right) \\ &= \sum_{s,a} \widetilde{\boldsymbol{\mu}}(s,a)(R(s,a,\widetilde{\boldsymbol{\mu}}_h) - R(s,a,\boldsymbol{\mu}_h)) - \sum_{s,a} \boldsymbol{\mu}(s,a)(R(s,a,\boldsymbol{\mu}_h) - R(s,a,\widetilde{\boldsymbol{\mu}}_h)) \\ &= \sum_{s,a} (\widetilde{\boldsymbol{\mu}}(s,a) - \boldsymbol{\mu}(s,a))(R(s,a,\widetilde{\boldsymbol{\mu}}_h) - R(s,a,\boldsymbol{\mu}_h)) \leqslant 0. \end{aligned}$$

□

The next lemma is simply an adaptation of the standard MFG performance difference lemma in single agent RL to the MFG setting.

Lemma 14 (Performance difference lemma)**.** *For an arbitrary MFG, let $\pi, \widetilde{\pi} \in \Pi$ and $\boldsymbol{\mu} = \Lambda(\pi)$.*

$$\begin{aligned} &V_0^\tau\left(s|\boldsymbol{\mu}, \widetilde{\pi}\right) - V_0^\tau\left(s|\boldsymbol{\mu}, \pi\right) + \tau \mathbb{E}_{\widetilde{\pi},\boldsymbol{\mu}}\left[\sum_{h=0}^{H-1} D_{KL}\left(\widetilde{\pi}_h\left(\cdot \mid s_h\right) | \pi_h\left(\cdot \mid s_h\right)\right) \middle| s_0 = s\right] \\ &= \mathbb{E}_{\widetilde{\pi},\boldsymbol{\mu}}\left[\sum_{h=1}^{H} \left\langle q_h^\tau\left(s_h, \cdot|\boldsymbol{\mu}, \pi\right) - \tau \log \pi_h\left(\cdot \mid s_h\right), \widetilde{\pi}_h\left(\cdot \mid s_h\right) - \pi_h\left(\cdot \mid s_h\right)\right\rangle \mid s_0 = s\right]. \end{aligned}$$

Proof. See the standard proof technique for the performance difference lemma, e.g. [38, 65]. □

Finally, we state two technical lemmas due to [65].

Lemma 15 (Lemma I.3 of [65])**.** *Let $p, p' \in \Delta_{\mathcal{A}}$ be arbitrary, and $\widehat{p} = (1-\beta)p + \beta\,\mathrm{Unif}(\mathcal{A})$ for some $\beta \in (0,1)$. Then,*

$$\begin{aligned} D_{KL}\left(p^*|\widehat{p}\right) &\leqslant \log\frac{|\mathcal{A}|}{\beta}, \\ D_{KL}\left(p^*|\widehat{p}\right) - D_{KL}\left(p^*|p\right) &\leqslant \frac{\beta}{1-\beta}. \end{aligned}$$

Lemma 16 (Lemma 3.3 in [6])**.** *Let $p, p^* \in \Delta_{\mathcal{A}}$, $\alpha > 0$ and $g : \mathcal{A} \to [0, H]$ be arbitrary, and $q \in \Delta_{\mathcal{A}}$ be a distribution such that $q(\cdot) \propto p(\cdot)\exp\{\alpha g(\cdot)\}$. Then,*

$$\left\langle g(\cdot), p^*(\cdot) - p(\cdot)\right\rangle \leqslant \alpha H^2/2 + \alpha^{-1}\left[D_{\mathrm{KL}}\left(p^*\|p\right) - D_{\mathrm{KL}}\left(p^*\|q\right)\right].$$

D.3 Extended Proof of Theorem 3

As mentioned before, the proof is an adaptation of [65] to setting where learning occurs with N potentially asymmetric agents. The main differences will be the absence of an explicit MFG and the fact that our algorithms are only allowed to use samples of finite agent trajectories.

Define the random variable $\boldsymbol{\mu}_t := \Lambda(\pi_t)$, which is the mean-field population distribution induced by the policy at epoch t. We denote the random variables due to estimation error of the q-functions at epoch t, time step h and an arbitrary state s as

$$\mathcal{E}^s_{t,h} := \left|\langle \widehat{q}^t_h(s,\cdot) - q^\tau_h(s,\cdot|\boldsymbol{\mu}_t,\pi_t), \pi^*_h(\cdot|s) - \pi_{t,h}(\cdot|s)\rangle\right|.$$

Furthermore, let π^*_τ be the unique τ-regularized MFG-NE. We define the quantity

$$\Delta_t := \mathbb{E}_{\mu^*_h}\left[D_{\mathrm{KL}}\left(\pi^*_h\left(\cdot \mid s_h\right) \| \pi_{t,h}\left(\cdot \mid s_h\right)\right)\right],$$

which will be the main quantity of error to be bounded using the techniques of [65]. Let $\beta_t := 1/t+1$. Finally, we also define the distribution mismatch ceofficient

$$C_{\mathrm{dist}} := \sup_{t,h} \sup_{\substack{s,a \\ \mu_{t,h}(s,a)>0}} \frac{\mu^*_h(s,a)}{\mu_{t,h}(s,a)},$$

which is always finite (and bounded) in our entropy-regularized setting (see for instance [12]).

It holds (almost surely) for any $s \in \mathcal{S}$ that:

$$\begin{aligned}
&D_{\mathrm{KL}}(\pi^*_h(\cdot|s)|\pi_{t+1,h}(\cdot|s)) \\
&\leqslant D_{\mathrm{KL}}\left(\pi^*_h\left(\cdot \mid s\right)|\widehat{\pi}_{t+1,h}\left(\cdot \mid s\right)\right) + \beta_t/(1-\beta_t) \\
&\leqslant -\xi_t \left\langle \widehat{q}^t_h\left(s,\cdot|\boldsymbol{\mu}_t,\pi_t\right) - \tau \log \pi_h\left(\cdot \mid s\right), \pi^*_h\left(\cdot \mid s\right) - \pi_{t,h}\left(\cdot \mid s\right)\right\rangle \\
&\quad + D_{\mathrm{KL}}\left(\pi^*_h\left(\cdot \mid s\right)|\pi_{t,h}\left(\cdot \mid s\right)\right) + \frac{1}{2}\xi_t^2\left(H + \tau H \log|\mathcal{A}| + \tau \log\frac{|\mathcal{A}|}{\beta_t}\right)^2 + \frac{\beta_t}{1-\beta_t} \\
&\leqslant -\xi_t \left\langle q^\tau_h\left(s,\cdot|\boldsymbol{\mu}_t,\pi_t\right) - \tau \log \pi_h\left(\cdot \mid s\right), \pi^*_h\left(\cdot \mid s\right) - \pi_{t,h}\left(\cdot \mid s\right)\right\rangle \\
&\quad + D_{\mathrm{KL}}\left(\pi^*_h\left(\cdot \mid s\right)|\pi_{t,h}\left(\cdot \mid s\right)\right) + \frac{1}{2}\xi_t^2\left(H + \tau H \log|\mathcal{A}| + \tau \log\frac{|\mathcal{A}|}{\beta_t}\right)^2 + \frac{\beta_t}{1-\beta_t} + \xi_t\mathcal{E}_{t,h}
\end{aligned}$$

Then,

$$\begin{aligned}
&\Delta_{t+1} - \Delta_t \\
&:= \sum_{h=0}^{H-1} \mathbb{E}_{\mu^*_h}\left[D_{\mathrm{KL}}\left(\pi^*_h\left(\cdot \mid s_h\right)|\pi_{t+1,h}\left(\cdot \mid s_h\right)\right) - D_{\mathrm{KL}}\left(\pi^*_h\left(\cdot \mid s_h\right)\|\pi_{t,h}\left(\cdot \mid s_h\right)\right)\right] \\
&\leqslant \eta V^\tau\left(\boldsymbol{\mu}_t,\pi_t\right) - V^\tau\left(\boldsymbol{\mu}_t,\pi^*\right) - \tau\xi_t \sum_{h=0}^{H-1} \mathbb{E}_{\mu^*_h}\left[D_{\mathrm{KL}}\left(\pi^*_h\left(\cdot \mid s_h\right)\|\pi_{t,h}\left(\cdot \mid s_h\right)\right)\right] \\
&\quad + \frac{1}{2}\xi_t^2 H\left(H + \tau H \log|\mathcal{A}| + \tau \log\frac{|\mathcal{A}|}{\beta_t}\right)^2 + \frac{\beta_t}{1-\beta_t}H + 2\xi_t \sum_{h=0}^{H-1} \mathbb{E}_{\mu^*_h}\left[\mathcal{E}_{t,h}\right] \\
&\leqslant -\tau\xi_t\Delta_t + \frac{1}{2}\xi_t^2 H\left(H + \tau H \log|\mathcal{A}| + \tau \log\frac{|\mathcal{A}|}{\beta_t}\right)^2 + \frac{\beta}{1-\beta_t}H + 2\xi_t \sum_{h=0}^{H-1} \mathbb{E}_{\mu^*_h}\left[\mathcal{E}_{t,h}\right]
\end{aligned}$$

Or rearranging both side,

$$\begin{aligned}
\Delta_t \leqslant &\frac{1}{\tau\xi_t}\left(\Delta_t - \Delta_{t+1}\right) + \frac{\xi_t}{2\tau}H\left(H + \tau H \log|\mathcal{A}| + \tau \log\frac{|\mathcal{A}|}{\beta_t}\right)^2 + \frac{\beta_t H}{(1-\beta_t)\tau\xi_t} \\
&+ \frac{2}{\tau}\sum_{h=1}^{H} \mathbb{E}_{\mu^*_h}\left[\varepsilon_h\right].
\end{aligned}$$

Summing this inequality from $t = 1, \dots, T$, we obtain

$$\frac{1}{T}\sum_{t=1}^{T}\Delta_t \leqslant \frac{1}{T\tau\xi_t}\Delta_1 + \frac{\xi_t}{2\tau}H\left(H + \tau H\log|\mathcal{A}| + \tau\log\frac{|\mathcal{A}|}{\beta_t}\right)^2 + \frac{\beta_t H}{(1-\beta_t)\tau\xi_t} + \frac{2}{\tau}\sum_{h=1}^{H}\mathbb{E}_{\mu_h^*}\left[\varepsilon_h\right].$$

Given that $\xi_t = 1/\sqrt{t+1}$ and $\beta_t = 1/t+1$, we obtain the bounds

$$\frac{1}{T}\sum_{t=1}^{T}\Delta_t = \frac{\tau^{-1}\Delta_1 + \tau^{-1}H^2 + \tau H\log|\mathcal{A}| + \tau\log^2 T}{\sqrt{T}} + \frac{\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{\mu_h^*}\left[\mathcal{E}_{t,h}^s\right]}{T\tau},$$

and finally using Young's inequality on the last term, and an application of Pinsker's inequality,

$$\begin{aligned}
&\frac{1}{T}\sum_t\sum_h\mathbb{E}_{\mu_h^*}\left[\frac{1}{2}\|\pi_{t,h}(\cdot|s_h) - \pi_h^*(\cdot|s_h)\|_1^2\right] \leqslant \frac{1}{T}\sum_{t=1}^{T}\Delta_t \\
&\leqslant \frac{\tau^{-1}\Delta_1 + \tau^{-1}H^2 + \tau H\log|\mathcal{A}| + \tau\log^2 T}{\sqrt{T}} + \frac{2\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{\mu_h^*}\left[\mathcal{E}_{t,h}^s\right]}{T\tau}, \\
&\leqslant \frac{\tau^{-1}\Delta_1 + \tau^{-1}H^2 + \tau H\log|\mathcal{A}| + \tau\log^2 T}{\sqrt{T}} + \frac{\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{\mu_h^*}\left[8\|\widehat{q}_h^t(s_h,\cdot) - q_h^\tau(s_h,\cdot)\|_2^2\right]}{T\tau^2}, \\
&\quad + \frac{\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{\mu_h^*}\left[\|\pi_{t,h}(\cdot|s_h) - \pi^*(\cdot|s_h)\|_2^2\right]}{4T}.
\end{aligned}$$

Rearranging the terms,

$$\begin{aligned}
&\frac{1}{T}\sum_t\sum_h\mathbb{E}_{\mu_h^*}\left[\frac{1}{2}\|\pi_{t,h}(\cdot|s_h) - \pi_h^*(\cdot|s_h)\|_1^2\right] \\
&\leqslant \frac{4\tau^{-1}\Delta_1 + 4\tau^{-1}H^2 + 4\tau H\log|\mathcal{A}| + 4\tau\log^2 T}{\sqrt{T}} \\
&\quad + \frac{\sum_{t=1}^{T}\sum_{h=0}^{H-1}\mathbb{E}_{\mu_h^*}\left[32\|\widehat{q}_h^t(s_h,\cdot) - q_h^\tau(s_h,\cdot)\|_2^2\right]}{T\tau^2}.
\end{aligned}$$

Finally, noting that by Theorem 2, after taking expectations on both sides it holds that $\mathbb{E}[\mathbb{E}_{\mu_h^*}\left[4\|\widehat{q}_h^t(s_h,\cdot) - q_h^\tau(s_h,\cdot)\|_2^2\right]] = \mathcal{O}(\varepsilon^2 + \alpha^2 + \beta^2 + 1/N)$, if follows that $\mathbb{E}[\mathbb{E}_{\mu_h^*}\left[\sum_h\|\bar{\pi}_h(\cdot|s_h) - \pi_h^*(\cdot|s_h)\|_2^2\right]] \leqslant \mathcal{O}(\varepsilon^2 + \tau^{-2}\alpha^2 + \tau^{-2}\beta^2 + \tau^{-2}1/N)$. Using the standard Lipschitz continuity of exploitability (see e.g. [61]), the exploitability bound in expectation holds. Using Lemma 12, we obtain the upper bound in expectation on the exploitability of the output policy $\bar{\pi}$ in terms of the original (unregularized) DG.

E Details of Experiments

E.1 Hardware Setup for Experiments

Except the A-Taxi benchmark, all our experiments are CPU-based. We use a single AMD EPYC 7742 CPU, equipped with 128GB RAM. For training policy and value neural networks with PPO in the A-Taxi benchmark, we use a single RTX 3090 GPU. With this setup, running Symm-PMD and IPMD in the A-SIS and A-RPS benchmarks takes between 5-20 minutes, and running PPO on A-Taxi takes approximate 2 hours. Evaluating exploitability for a given policy on A-SIS and A-RPS takes around 2 hours, as we employ a brute-force UCB-type bandit algorithm to accurately estimate best response in this setting.

E.2 Extended Descriptions of the Experimental Setup

For simplified notation, denote the state-action marginal densities

$$\sigma_{\text{actions}}(\boldsymbol{\rho}, a) = \sum_{s' \in \mathcal{S}} \sigma(\boldsymbol{\rho})(s', a),$$
$$\sigma_{\text{states}}(\boldsymbol{\rho}, s) = \sum_{s' \in \mathcal{A}} \sigma(\boldsymbol{\rho})(s, a').$$

Modified rock-paper-scissors (A-RPS). We formulate a modified population rock-paper-scissors game inspired by the formulation of [13]. Our version incorporates varying preferences between agents between possible moves as well as a crowdedness penalty.

A-RPS consists of three states $\mathcal{S} := \{R, P, S\}$ and three actions $\mathcal{A} := \{R, P, S\}$. We use $N = 2000$ players and a time horizon of $H = 10$, though these can be increased or decreased arbitrarily. We define the rewards as follows.

$$R^i(s = R, a, \boldsymbol{\rho}^{-i}) = -c^i \sigma_{\text{actions}}(\boldsymbol{\rho}, a) - u^i_R \sigma_{\text{states}}(\boldsymbol{\rho}, P) + v^i_R \sigma_{\text{states}}(\boldsymbol{\rho}, S),$$
$$R^i(s = P, a, \boldsymbol{\rho}^{-i}) = -c^i \sigma_{\text{actions}}(\boldsymbol{\rho}, a) - u^i_P \sigma_{\text{states}}(\boldsymbol{\rho}, S) + v^i_P \sigma_{\text{states}}(\boldsymbol{\rho}, R),$$
$$R^i(s = S, a, \boldsymbol{\rho}^{-i}) = -c^i \sigma_{\text{actions}}(\boldsymbol{\rho}, a) - u^i_S \sigma_{\text{states}}(\boldsymbol{\rho}, R) + v^i_S \sigma_{\text{states}}(\boldsymbol{\rho}, P).$$

The coefficients c^i, u^i, v^i are unique for each agent indicating their own utilities/rewards due to losing, winning, or individual penalty due to crowdedness. The state transitions are deterministic and are given by:

$$P^i(s'|s, a, \boldsymbol{\rho}^{-i}) = \mathbb{1}_{s'=a}.$$

We generate the fixed coefficients u^i, v^i randomly by adding bounded noise to coefficients from [13], so that

$$u^i_R = 2 + \varepsilon^i_R, \quad v^i_R = 1 + \bar{\varepsilon}^i_R$$
$$u^i_P = 4 + \varepsilon^i_P, \quad v^i_P = 2 + \bar{\varepsilon}^i_P$$
$$u^i_S = 6 + \varepsilon^i_S, \quad v^i_S = 3 + \bar{\varepsilon}^i_S.$$

Therefore, the magnitudes of the player-specific additive terms determine β. In the case of A-RPS, $\alpha = 0$.

Infection modeling with asymmetric agents (A-SIS). This benchmark, inspired by the SIS benchmark of [13], models a large population of infected or healthy agents that can choose to go out or remain in isolation. Unlike the SIS benchmark, A-SIS is formulated as an N-player game and incorporates individual differences in natural susceptibilities, recovery rates, and aversion of isolation between agents. We formalize the dynamic game as follows. The game consists of the state space $\mathcal{S} = \{I, H\}$ (I indicating infected, H indicating healthy), and action space $\mathcal{A} = \{D, U\}$ (D indicating social distancing, U indicating going out). The initial states s^i_0 are sampled i.i.d. from a uniform distribution over $\mathcal{S}$. Each agent $i \in N$ has a fixed *susceptibility* parameter $\alpha_i \in [0, 1]$, a fixed *healing probability* $\theta_i \in [0, 1]$ and a fixed *aversion to isolation* parameter $\xi_i \in [0, 1]$.

$$P^i(I|H, D, \boldsymbol{\rho}^{-i}) = 0$$
$$P^i(I|H, U, \boldsymbol{\rho}^{-i}) = \alpha_i * \sigma(()\boldsymbol{\rho})(I, U),$$
$$P^i(I|I, D, \boldsymbol{\rho}^{-i}) = 1 - \theta_i,$$
$$P^i(I|I, U, \boldsymbol{\rho}^{-i}) = 1 - \theta_i.$$

The probabilities of staying healthy are of course always defined by

$$P^i(H|s, a, \boldsymbol{\rho}^{-i}) := 1 - P^i(I|s, a, \boldsymbol{\rho}^{-i}).$$

The rewards of each agent are give by the following which incorporates a penalty for illness and an agent specific penalty for isolation:

$$R^i(s, a, \boldsymbol{\rho}^{-i}) = -\mathbb{1}_{s=I} - \xi_i \mathbb{1}_{a=D}.$$

Parameter	Value
Initial learning rate	$2.5\text{e}{-4}$
Learning rate schedule	Linear
γ (discount factor)	0.999
λ_{GAE} (see[49])	0.95
Entropy regularization	$6\text{e}{-3}$
Value loss coefficient	0.9
Maximum gradient norm	0.5
Clip coefficient	0.2
Sample trajectories per epoch	1
NN training passes per epoch	4
Minibatch size	$16\,384$
Advantage normalization	Yes

Table 2: Hyperparameters of the PPO algorithm.

The agent parameters $\alpha_i, \theta_i, \xi_i$ are as expected fixed throughout the game, and are sampled to be close. In the case of A-SIS, we solve $N = 1000$ agents with a time horizon of $H = 20$.

Asymmetric taxi (A-Taxi). Finally, as a more complicated benchmark we adapt the Taxi As in [13], we use the following layout of the city map, where S indicates the starting cell of all agents, both H and S are impenetrable barriers and the rest of the city is divided into 2 zones.

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \\ H & S & H \\ 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{pmatrix}$$

The action space of each agent is $\mathcal{A} := \{U, D, L, R, W\}$, indicating actions to move in four directions (up, down, left, right respectively) and wait at the current location. Customers can only be picked up while waiting, and delivered while waiting. Each cell in the grid generates a new customer with probability 0.2, which the agents can observe. Upon picking up a customer, a random target coordinate is generate within the same zone. Customers can only be left at their target cells. Successful deliveries of customers in zone 1 generate a base reward of 1.1, whereas successful deliveries of customers in zone 2 generate a lower reward of 1.0. Furthermore, each agent has a zone specific reward multiplier $\beta_i^1, \beta_i^2 > 0$, so that agent i by delivering a customer in zone 1 gains reward $1.1\beta_i^1$ and vice versa. This models varying efficiencies of taxi drivers as well as individual preferences to various zones of the city. Furthermore, we incorporate a crowdedness penalty: an agent $i \in [N]$ at state s at time h will not move (simulating a jamming effect) with probability $\min\{\sum_j w_j \mathbb{1}_{s_h^j = s}, 0.7\}$, where $\{w_j\}_{j=1}^N$ are player specific weights indicating their contribution to traffic jams. This intends to simulate unique contributions of each driver to traffic jams, presumably due to vehicle types, driving styles, etc.

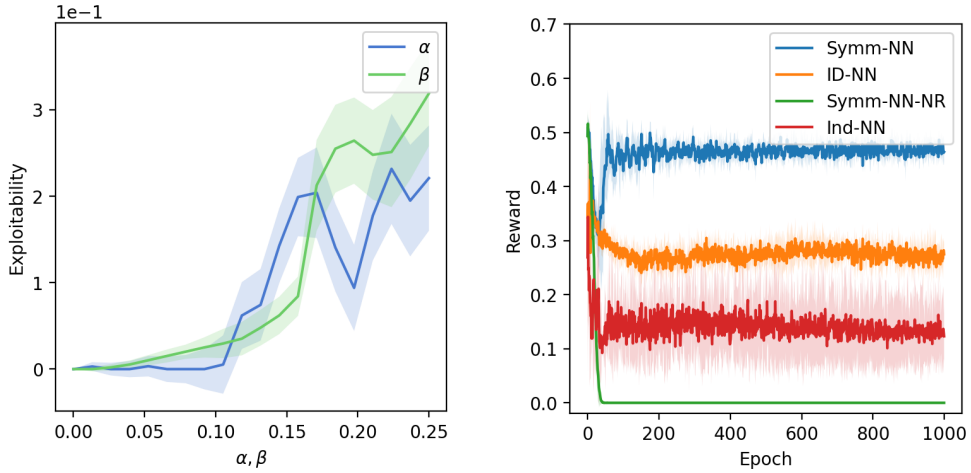
The number of states in the game is on the order of 2^{30}, making a neural network approximation fundamental. For this reason, we use value and policy networks with two hidden layers with 128 neurons each, with a leaky ReLU nonlinearity. We adopt the PPO implementation of CleanRL [26] for our purposes. The hyperparameters used are indicated in Table 2.

E.3 Extended Experimental Results

We report two additional sets of results regarding the sensitivity of our algorithms to α, β and the population distribution behvaiour in the A-Taxi environment.

In Figure 2-(a), we report the sensitivity of the exact MFG-NE (computed via code provided in [21]) to heterogeneity parameters α, β in terms of exploitability in the $N = 1000$ player game. While keeping the other parameter constant at 0, we sweep through various values of each of α, β in the range $(0, 1/4)$. While around the 0.1 threshold, the exploitability rises as expected, for smaller values

of α, β the bias introduced is very small, providing an empirical analysis of the approximation bound of Theorem 1.



(a) (b)

Figure 2: (a) The sensitivity of the MFG-NE to heterogeneity parameters α, β in the A-SIS environment, in terms of exploitability. (b) Percentage of vehicles in Zone 1 in the A-Taxi environment throughout training epochs for 4 benchmark algorithms.

In Figure 2-(b), we keep track of number of taxis choosing to operate in Zone 1 in the A-Taxi environment throughout training. While rewards in Zone 2 are higher in this environment, congestion effects require a mixed Nash equilibrium: agents must randomly choose at the very first step to serve either Zone 1 or 2. The figure demonstrates the main advantage of policy-based methods for learning Nash: unlike most value-based methods, PPO can learn a mixed strategy instead of converging to a deterministic policy. As an additional benchmark, in the figure we evaluate Symm-NN without any entropy regularization ($\tau = 0$, shown by the line Symm-NN-NR). In this case, the policy rapidly converges to a deterministic policy, indicating that a non-zero entropy regularizer might be necessary for learning a Nash equilibrium.