Exploiting Approximate Symmetry for Efficient Multi-Agent Reinforcement Learning

Batuhan Yardim YARDIMA@ETHZ.CH
Department of Computer Science, ETH Zürich, Zürich, Switzerland

Niao He NIAOHE@ETHZ.CH
Department of Computer Science, ETH Zürich, Zürich, Switzerland

Editors: N. Ozay, L. Balzano, D. Panagou, A. Abate

Abstract

Mean-field games (MFG) have become significant tools for solving large-scale multi-agent reinforcement learning problems under symmetry. However, the assumptions of access to a known MFG model (which might not be available for real-world games) and of exact symmetry (real-world scenarios often feature heterogeneity) limit the applicability of MFGs. 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 Kirszbraun extensions. Next, we define α, β-symmetric games, a new class of dynamic games that incorporate approximate permutation invariance. We establish explicit approximation bounds for α, β-symmetric games, demonstrating that the induced mean-field Nash policy is an approximate Nash of the N-player game. We analyze TD learning using sample trajectories of the N-player game, permitting learning without using an explicit MFG model or oracle. This is used to show a sample complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-6})$ for N-agent monotone extendable games to learn an ε-Nash. Evaluations on benchmarks with thousands of agents support our theory of learning under (approximate) symmetry without explicit MFGs.

Keywords: symmetry, multi-agent reinforcement learning, game theory

1. Introduction

Competitive multi-agent reinforcement learning (MARL) has found a wide range of applications in the recent years (Shavandi and Khedmati, 2022; Samvelyan et al., 2019; Rashedi et al., 2016; Matignon et al., 2007; Mao et al., 2022). Simultaneously, MARL is fundamentally challenging at the regime with many agents due to an exponentially growing search space (Wang et al., 2020), also known as the *curse-of-many-agents*. Even finding an *approximate* solution (i.e. approximate *Nash*) is PPAD-hard (Daskalakis et al., 2023), 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. Leonardos et al. (2021); Perrin et al. (2020)).

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 Lasry and Lions (2007) and Huang et al. (2006), MFG analyzes the limit $N \to \infty$ of N-player games with symmetric agents. 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 on learning (see Laurière et al. (2024) for a survey).

Work	Symmetry	Approximation	Learning	Learn w/o model
Saldi et al., 2018	Exact	✓*(asymptotic)*	✗	-
Yardim et al., 2024	Exact	✓*(explicit)*	✗	-
Zaman et al., 2023	Exact	✗	✓*(reg.)*	✗
Yardim et al., 2023a	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)

However, works on MFG exhibit two major bottlenecks preventing wider applicability in MARL. First and foremost, many works on MFG (such as Guo et al. (2019); Pérolat et al. (2022)) implicitly assume that an exact model of the MFG is known to the algorithm akin to solving a known MDP. In real-world applications, such a 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. Secondly, 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 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, imperfections in dynamics usually break invariance. Little research has studied whether MFGs could offer tractable approximations to otherwise intractable games that might exhibit approximate symmetries.

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 as well as its approximation of finite-player Nash (Carmona and Delarue, 2013; Carmona et al., 2018; Saldi et al., 2018). The convergence of RL algorithms has also been widely studied in discrete-time MFG assuming either contractivity in the stationary equilibrium setting (Zaman et al., 2023; Yardim et al., 2023a; Guo et al., 2019; Xie et al., 2021) or monotonicity in the finite-horizon setting (Perrin

et al., 2020; Pérolat et al., 2022; Perrin et al., 2022; Yardim et al., 2023b). These models however assume exact homogeneity between all participants. Furthermore, existing algorithms typically assume knowledge of the exact MFG model (Guo et al., 2019; Zaman et al., 2023), hindering their real-world applicability. Multi-population MFG (MP-MFG) can incorporate multiple types of populations exposed to different dynamics (Huang et al., 2006; Pérolat et al., 2022; Subramanian et al., 2020; Dayanikli and Lauriere, 2023; Bensoussan et al., 2018; Carmona et al., 2018; Huang et al., 2024). 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 (Huang et al., 2024). Overall, these works require variations of the same stringent symmetry assumptions, restricting their applicability. A detailed survey of learning in MFGs can be found at (Laurière et al., 2024).

Graphon MFG. Graphon games, proposed initially by (Parise and Ozdaglar, 2019), can incorporate heterogeneity between MFG agents by assuming graphon-based interactions. The setting has been analyzed in discrete-time (Cui and Koeppl, 2021b; Vasal et al., 2020) as well as in the continuous time setting (Aurell et al., 2022a; Caines and Huang, 2019; Aurell et al., 2022b). Recently, policy mirror descent has been analyzed in this setting to produce convergence guarantees under monotonicity (Zhang et al., 2023). However, graphon MFGs still assume exact symmetry in the form of the graphon, making them reducible to regular MFGs (Zhang et al., 2023). Namely, the types of agents must follow a symmetric distribution and interactions must be through a symmetric graphon.

Other related work. Another class where a large number of agents can be tackled tractably are the so-called potential games (Rosenthal, 1973), generalized to Markov potential games incorporating dynamics (Leonardos et al., 2021). Approximate potentials have been studied in a similar spirit on Markov α-potential games (Guo et al., 2023) and near potential games (Candogan et al., 2013). However, to the best of our knowledge, approximate symmetry has not been studied in the literature.

1.2. Our Contributions

We list the following as our contributions compared to past work. All proofs and other technical details are fully presented in the extended paper (Yardim and He, 2024).

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.
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 finite-agent DG. We show that by only using $\mathcal{O}(\varepsilon^{-2})$ sample trajectories from the N-player dynamic game, policies can be approximately evaluated *on the abstract MFG* up to symmetrization error: hence no explicit MFG is necessary for evaluation.
5. 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.

2. Main Results

Notation. For $K \in \mathbb{N}_{>0}$, let $[K] := \{1, \dots, K\}$, and $\Delta_{\mathcal{X}}$ be the probability simplex on $\mathcal{X}$. For any $N \in \mathbb{N}_{>0}$ define $\Delta_{\mathcal{X},N} := \{\mu \in \Delta_{\mathcal{X}} \,|\, N\mu(x) \in \mathbb{N}_{\geqslant 0}, \forall x \in \mathcal{X}\}$. 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 group of permutations over the set $[K]$, so $\mathbb{S}_K := \{g : [K] \to [K] \,|\, g \text{ bijective}\}$. 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 denote the standard basis vector with k-th entry 1 as $\mathbf{e}_k$.

We consider discrete state-action sets $\mathcal{S}, \mathcal{A}$. We denote the set of time-dependent policies on $\mathcal{S}, \mathcal{A}$ as $\Pi := \{\pi : \mathcal{S} \times [H] \to \Delta_{\mathcal{A}}\}$, and 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, define $\mathcal{H}(u) := -\sum_a u(a) \log u(a)$ for $u \in \Delta_{\mathcal{A}}$ and $D_{\text{KL}}(u|v) := \sum_a u(a) \log \frac{u(a)}{v(a)}$ for $u, v \in \Delta_{\mathcal{A}}$.

2.1. Finite-Horizon Dynamic Games

Firstly, we define finite-horizon dynamic games as the main object of interest in this work. The definition differs from Markov games (Shapley, 1953) where a common state is shared by all agents: in FH-DG each agent only observes their own state while the dynamics depend on the state vector of all N agents. Such a model is realistic in cases where games have natural *locality* and the game state is not globally available to agents.

Definition 1 (FH-DG, Nash equilibrium) *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 $R^i : \mathcal{S} \times \mathcal{A} \times (\mathcal{S} \times \mathcal{A})^{N-1} \to [0,1]$.*

For $\mathcal{G}$, the expected total reward of agent $i \in [N]$ for policy profile $\boldsymbol{\pi} = (\pi^1, \dots, \pi^N) \in \Pi^N$ is

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

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.*

The natural solution concept to the FH-DG defined in Definition 1 is an approximate Nash equilibrium: At δ-NE, the incentive for any selfish agent to deviate is small, hence, approximate NE is a natural solution concept for FH-DG. However, the problem of finding δ-NE is challenging: not only is it computationally intractable (Daskalakis et al. (2009) show that it is PPAD-hard, a class containing computation problems believed to be not in P), but the search space of policies grows exponentially in N. This motivates the approximation in the remainder of the work.

2.2. Symmetrization and Lipschitz Extension

In order to construct a MFG from a FH-DG, 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 finite players?* We will use the notions of symmetrization and Lipschitz extension.

Definition 2 (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 *$\overline{f} : \Delta_{\mathcal{X},K} \to \mathcal{Y}$ as the well-defined function such that $\overline{f}(\mu) = f(\mathbf{x})$ 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* $\mathrm{Sym}(f)(\mathbf{x}) = \frac{1}{K!}\sum_{g \in \mathbb{S}_K} f(g(\mathbf{a})), \forall \mathbf{a} \in \mathcal{X}^K$. *We also denote* $\overline{\mathrm{Sym}}(f) := \mathrm{Sym}(\overline{f})$.

The terminology "symmetrization" is consistent as $\mathrm{Sym}(f)$ is indeed a symmetric function. Furthermore, if f is symmetric then $\mathrm{Sym}(f) = f$ as expected. While the operator $\overline{\mathrm{Sym}}(\cdot)$ is useful, $\overline{\mathrm{Sym}}(f)$ is still defined only on the discrete finite lattice $\Delta_{\mathcal{X},K}$. To extend an FH-DG to the continuum of infinitely many players, we will use the following special case of the Kirszbraun-Valentine theorem, concerning Lipschitz extensions of functions from subsets of Euclidean spaces.

Lemma 3 (Kirszbraun (1934); Valentine (1945)) *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 admits various explicit formulations (McShane, 1934; Sukharev, 1978), and the particular formulation is not relevant 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.

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 4 (MFG, induce population, MFG-NE) *A finite-horizon mean-field game (FH-MFG) is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \rho_0, H, P, R)$ where $\mathcal{S}, \mathcal{A}$ are discrete sets, $\rho_0 \in \Delta_{\mathcal{S}}$, $H \in \mathbb{N}_{>0}$, 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 $R : \mathcal{S} \times \mathcal{A} \times \Delta_{\mathcal{S}\times\mathcal{A}} \to [0, 1]$.*

For $\mathcal{M}$, define the operators $\Gamma : \Delta_{\mathcal{S}\times\mathcal{A}} \times \Pi \to \Delta_{\mathcal{S}\times\mathcal{A}}$ and $\Lambda : \Pi \to \Delta^H_{\mathcal{S}\times\mathcal{A}}$ as $\Gamma(\mu, \pi)(s', a') := \sum_{s\in\mathcal{S}, a\in\mathcal{A}} \mu(s, a)P(s'|s, a, \mu)\pi(a'|s')$ and $\Lambda(\pi) := \{\Gamma(\cdots\Gamma(\Gamma(\rho_0 \cdot \pi_0, \pi_1), \pi_2)\cdots, \pi_{h-1})\}_{h=0}^{H-1}$, which are called population flow operators. For $\pi \in \Pi$ and $\boldsymbol{\mu} = \{\mu_h\}_{h=0}^{H-1} \in \Delta^H_{\mathcal{S}\times\mathcal{A}}$, the expected reward is defined as

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

We define MFG exploitability as $\mathcal{E}(\pi) := \max_{\pi'\in\Pi} V(\Lambda(\pi), \pi') - V(\Lambda(\pi), \pi)$. If $\mathcal{E}(\pi^*) = 0$ for $\pi^* = \{\pi^*_h\}_{h=0}^{H-1}$, we call π^* a MFG Nash equilibrium (MFG-NE).

Intuitively, the above definition of MFG-NE requires that the policy π is optimal against the population flow it induces. Questions of the existence of MF-NE (Cardaliaguet, 2010; Bensoussan et al., 2013; Huang et al., 2023) and approximation of the FH-DG under exact symmetry (Saldi et al.,

2018) 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 corresponding FH-DG.

Taking a constructive bottom-up approach, we show that for any given FH-DG, the MFG-NE of an *appropriately constructed* MFG is also an approximate NE of the FH-DG. The definition below of an "induced MFG" demonstrates how arbitrary non-symmetric dynamics are extended to an MFG in this work.

Definition 5 (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 3. 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.

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

Definition 6 (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 a FH-DG, inducing $\mathcal{M} := \mathrm{MFG}(\mathcal{G}) = (\mathcal{S}, \mathcal{A}, \rho_0, H, P, R)$. $\mathcal{G}$ is called α, β-symmetric for $\alpha, \beta \geqslant 0$ if*

$$\max_{i,\,s,a} \mathrm{disc}(P^i(s,a,\cdot), P(s,a,\cdot)) \leqslant \alpha, \qquad \max_{i,\,s,a} \mathrm{disc}(R^i(s,a,\cdot), R(s,a,\cdot)) \leqslant \beta,$$

where we define the discrepancy $\mathrm{disc}(f,g) = \max_{\mu\in\Delta_{\mathcal{S}\times\mathcal{A},N}} \max_{\substack{\boldsymbol{\rho}\in(\mathcal{S}\times\mathcal{A})^{N-1} \\ \sigma(\boldsymbol{\rho})=\mu}} \|f(\boldsymbol{\rho}) - g(\mu)\|_1$.

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 6 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 7 (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 $i \in [M]$. Dynamics $\{P^i\}_{i=1}^N$ (resp. all 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 $s \in \mathcal{S}, a \in \mathcal{A}$.*

Sparsity is common in practice, particularly when there is spatial structure. Many standard MFG problems such as the beach-bar problem (Perrin et al., 2020) and crowd modeling (Zaman et al., 2023) 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 8 (Lipschitz extension) *Let $\mathcal{G}$ be an FH-DG with κ-sparse dynamics/rewards $\{P^i\}_{i=1}^N$ and $\{R^i\}_{i=1}^N$ admitting the induced MFG $\mathrm{MFG}(\mathcal{G})$ with dynamics and rewards P, R. Assume further 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 $s \in \mathcal{S}, a \in \mathcal{A}, \mu, \mu' \in \Delta_{\mathcal{S}\times\mathcal{A}}$.

Lemma 8 characterizes a condition on the original FH-DG for the induced MFG to have smooth dynamics. The result suggests that the game must have *no dominant players*, that is, the effect of each agent on others must be of order $\mathcal{O}(1/N)$. Furthermore, by standard results in MFG literature, if the "no dominant players" condition of Lemma 8 holds, the population update Γ is also Lipschitz continuous with some modulus $L_{pop,\mu}$ that is independent of N (Yardim et al., 2024). We state the main approximation result, which quantifies how closely the true N-player game NE can be approximated by the mean-field NE of the symmetrized game.

Theorem 9 *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 $(\pi^*, \dots, \pi^*) \in \Pi^N$ is an ϵ-NE of the FH-DG, where $\epsilon = \mathcal{O}\left(\frac{H^2(1-L_{pop,\mu}^H)}{(1-L_{pop,\mu})\sqrt{N}} + \alpha H^2 \frac{1-L_{pop,\mu}^H}{1-L_{pop,\mu}} + \beta H\right)$.*

Firstly, Theorem 9 is agnostic to the extension construction, which might affect computational aspects. The approximation bound of Theorem 9 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 (Yardim et al., 2024). Theorem 9 also recovers the bounds known for exactly symmetric FH-DG (i.e. $\alpha = \beta = 0$, see Yardim et al. (2024)). Importantly, Theorem 9 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. It is known that for $N > 2$, finding an ϵ-NE of the FH-DG is PPAD-complete even for some *absolute constant* ϵ (Goldberg, 2011). Hence, even when α, β are not close to 0, the result might be non-trivial.

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 methods (Laurière et al., 2024). 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 tractable alternative algorithms.

2.5. Policy Evaluation with α, β-Symmetry

In this section, we analyze TD learning for α, β-symmetric FH-DG. While Definition 5 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 10 (Mean-field Q values) *For the MFG $(\mathcal{S}, \mathcal{A}, \rho_0, H, P, R)$, for $\tau \geqslant 0$, $h = 0, \dots, H-1$, we define (entropy regularized) Q-values for each $h = 0, \dots, H-1$ and $s \in \mathcal{S}, a \in \mathcal{A}$ as*

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

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 (differing from typical Q values in MARL). We will also treat $Q_h^{\tau,\pi}$ as an element of the vector space $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$. For estimating $Q_h^{\tau,\pi}$ using sample trajectories, we will analyze TD learning, which is a standard method for policy evaluation with established guarantees in classical RL (Tsitsiklis and Van Roy, 1996). We formulate Algorithm 1, presented for simplicity as performing TD learning on agent 1, and present its analysis in Theorem 11.

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 $\tau \geqslant 0$

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

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

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

  Perform TD update $\forall h \in \{0, \dots H-1\}$, (where $\widehat{Q}_H^m := 0$ and $\mathbf{e}_h^m := \mathbf{e}_{s_{m,h}^1, a_{m,h}^1}$)

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

end

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

Theorem 11 (TD learning for α, β-Symmetric Games) *Let $\mathcal{G}$ be an N-player FH-DG and $\mathrm{MFG}(\mathcal{G})$ its induced MFG. Let $\pi \in \Pi$ be a policy such that $\Lambda(\pi) = \boldsymbol{\mu} = \{\mu_h\}_h$, $\{Q_h^{\tau,\pi}\}_{h=0}^{H-1}$ are its mean-field Q values evaluated on $\mathrm{MFG}(\mathcal{G})$, and $\delta := \inf_{h,s,a:\mathbb{P}[s_h^1=s,a_h^1=s]>0} \mathbb{P}[s_h^1 = s, a_h^1 = s]$. 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 $\{\widehat{Q}_h^M\}_h$ satisfies $\sum_{h=0}^{H-1} \mathbb{E}[\|\widehat{Q}_h^M - Q_h^{\tau,\pi}\|_{\mu_h}^2] \leqslant \mathcal{O}\left(\frac{1}{M} + \frac{1}{N} + \alpha^2 + \beta^2\right)$, where $\|q\|_p^2 := \sum_{s,a} p(s,a)q(s,a)^2$ for any $p \in \Delta_{\mathcal{S}\times\mathcal{A}}$.*

Theorem 11 provides a finite-sample guarantee for TD learning, a building block of many RL algorithms. Furthermore, 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 efficiently. The rate of convergence suggested by Theorem 11 also matches the optimal known rates for TD-learning in a single-agent setting (Kotsalis et al., 2022).

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 a monotonicity assumption. Our algorithm uses TD learning as a building block, with stochastic policy evaluations used for policy mirror descent updates (Lan, 2023). We define monotonicity and provide an example.

Definition 12 (Monotone MFG (Lasry and Lions, 2007)) *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* MFG $(\mathcal{G})$ *is monotone.*

Example 1 (Asymmetric dynamic congestion games) *For $i \in [N]$, 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 so that $h_i(s,a,\cdot)$ is non-increasing for all s, a. Assume $P^i(\cdot|s,a,\boldsymbol{\rho}^{-i})$ does not depend on $\boldsymbol{\rho}^{-i}$ for any s, a, and $R^i(s,a,\boldsymbol{\rho}^{-i})$ is 1-sparse so that $R^i(s,a,\boldsymbol{\rho}^{-i}) = h_i(s,a,\sum_{j=1}^N \mathbb{1}_{\rho_j=(s,a)}) + r_i(s,a)$. This can be seen as a generalization of congestion games with player-specific incentives (Milchtaich, 1996), for which an efficient solution is unknown. Such games are monotone extendable, shown in the supplementary (Yardim and He, 2024)*

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.

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

for $t \in 0, 1, \dots, T-1$ **do**

  Run Algorithm 1 for policy π_t, M epochs, entropy τ, $\{\eta_m\}_m$ as in Theorem 11

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

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

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

end

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

We present the main learning result of this paper in Theorem 13 for monotone extendable $\mathcal{G}$. Critically, the sample complexity implied by the theorem is independent of N: hence, the curse of many agents can be circumvented for α, β-symmetric games.

Theorem 13 (Convergence of PMD) *Let $\mathcal{G}$ be a monotone extendable α, β-symmetric game. Assume Algorithm 2 runs with learning rates $\xi_t = \frac{1}{\sqrt{t+1}}$, entropy regularization $\tau \in (0, 1/2)$, with $M \geqslant \mathcal{O}(\varepsilon^{-2})$ TD iterations for $T \geqslant \widetilde{\mathcal{O}}(\varepsilon^{-4})$ epochs. Then, the output policy $\bar{\pi}$ is a $\mathcal{O}(\varepsilon\tau^{-1} + \alpha\tau^{-1} + \beta\tau^{-1} + \tau^{-1}/\sqrt{N} + \tau)$-Nash equilibrium of $\mathcal{G}$ in expectation.*

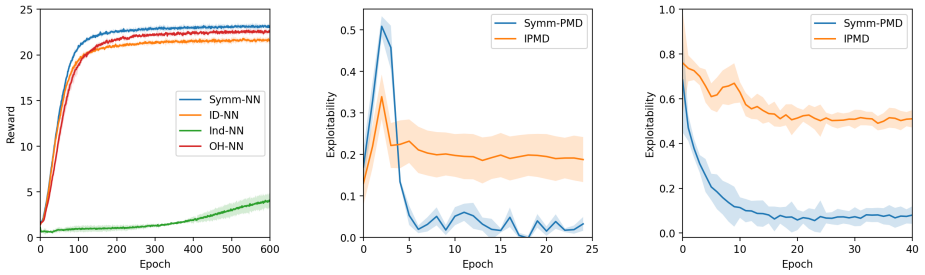
Theorem 13 suggests a sample complexity of $\widetilde{\mathcal{O}}(\varepsilon^{-6})$ trajectories from the FH-DG in order to compute a ε-NE (up to standard bias). In fact, it is to the best of our knowledge the first finite-sample guarantee for computing approximate NE for this class of FH-DGs with many agents. Even in the exactly symmetric case ($\alpha = \beta = 0$), Theorem 13 is the first guarantee to the best of our knowledge for learning NE only observing trajectories of the N-agent game.

3. Experimental Results

We support our theory by deploying Algorithm 2 on several α, β-symmetric environments. We formulate the benchmarks **A-RPS** (a generalized dynamic rock-paper-scissors), **A-SIS** (a social distancing and infection simulation), and **A-Taxi** (a ride matching simulation), adapted from MFG literature (Cui and Koeppl, 2021a). We construct the FH-DGs by perturbing dynamics so that $\alpha = 0, \beta \approx 0.1, N = 2000, H = 10, |\mathcal{S}| = |\mathcal{A}| = 3$ for A-RPS and $\alpha \approx \beta \approx 0.1, N = 1000, H = 20, |\mathcal{S}| = |\mathcal{A}| = 2$ for A-SIS. On the other hand, A-Taxi incorporates a large state space ($|\mathcal{S}| > 2^{30}, H = 128, N = 1000$), necessitating an adaptation of Algorithm 2 using neural network approximation and PPO (Schulman et al., 2017). Further details can be found in the supplementary.

We first compare Symm-PMD to its natural counterpart independent PMD (IPMD), where a separate policy is trained for each agent in A-RPS and A-SIS. Figures 1-(b,c) show the exploitability of the policies throughout learning: Symm-PMD is demonstrably sample-efficient despite its strong inductive bias, while IPMD struggles to converge presumably due to the curse of many agents. In A-Taxi, we use 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 finding a better policy.

We also underline the computational efficiency of our approach. Learning independent neural policies for each of 1000 agents (Ind-NN) is extremely expensive in A-Taxi: this approach performs the worst and is orders of magnitude computationally more expensive. Since our algorithm need not learn separate policies for agents, it is drastically more computationally efficient: learning is $\approx$60% faster in A-RPS and A-SIS, whereas symmetrized PPO is >95% faster than Ind-NN in A-Taxi for the same number of learning epochs. In fact, at the regime $N \approx 1000$, even one-hot encoding of agents in neural policies might be prohibitive, demonstrating the efficacy of our method.



(a) (b) (c)

Figure 1: (a) Total rewards throughout training in A-Taxi. *Sym-NN:* symmetric policies, *OH-NN:* policy with onehot encoding for i, *ID-NN:* numerical encoding for i and *Ind-NN:* independent policies. (b, c) The exploitability throughout epochs of Symm-PMD and IPMD: A-RPS in (b) and A-SIS in (c).

4. Discussion and Conclusion

In this work, we formulated a new class of competitive MARL problems (by α, β-symmetry) that can be tractably solved utilizing an *implicit* MFG. We provided guarantees for TD learning, and under monotonicity, for PMD to approximate NE up to symmetrization bias, providing a complete theory of learning. As the assumptions of our approach are light, future work could involve applications to more complicated problems such as city traffic, as well as the empirical measurement of asymmetry.

Acknowledgments

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

References

Alexander Aurell, René Carmona, Gökçe Dayanıklı, and Mathieu Laurière. Finite state graphon games with applications to epidemics. *Dynamic Games and Applications*, 12(1):49–81, 2022a.

Alexander Aurell, René Carmona, and Mathieu Lauriere. Stochastic graphon games: Ii. the linear-quadratic case. *Applied Mathematics & Optimization*, 85(3):39, 2022b.

Alain Bensoussan, Jens Frehse, Phillip Yam, et al. *Mean field games and mean field type control theory*, volume 101. Springer, 2013.

Alain Bensoussan, Tao Huang, and Mathieu Lauriere. Mean field control and mean field game models with several populations. *arXiv preprint arXiv:1810.00783*, 2018.

Peter E Caines and Minyi 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.

Ozan Candogan, Asuman Ozdaglar, and Pablo 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.

Pierre Cardaliaguet. Notes on mean field games. Technical report, Technical report, 2010.

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

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

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

Kai Cui and Heinz Koeppl. Learning graphon mean field games and approximate nash equilibria. *arXiv preprint arXiv:2112.01280*, 2021b.

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

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

Gokce Dayanikli and Mathieu Lauriere. Multi-population mean field games with multiple major players: Application to carbon emission regulations. *arXiv preprint arXiv:2309.16477*, 2023.

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

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

Xin Guo, Xinyu Li, Chinmay Maheshwari, Shankar Sastry, and Manxi Wu. Markov α-potential games: Equilibrium approximation and regret analysis. *arXiv preprint arXiv:2305.12553*, 2023.

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

Jiawei Huang, Niao He, and Andreas Krause. Model-based rl for mean-field games is not statistically harder than single-agent rl, 2024.

Minyi Huang, Roland P Malhamé, and Peter 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.

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

Georgios Kotsalis, Guanghui Lan, and Tianjiao 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.

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

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

Mathieu Laurière, Sarah Perrin, Julien Pérolat, Sertan Girgin, Paul Muller, Romuald Élie, Matthieu Geist, and Olivier Pietquin. Learning in mean field games: A survey, 2024.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969*, 2021.

Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, Ravi Iyer, and Tamer Basar. A mean-field game approach to cloud resource management with function approximation. In *Advances in Neural Information Processing Systems*, 2022.

Laëtitia Matignon, Guillaume J Laurent, and Nadine 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.

Edward James McShane. Extension of range of functions. 1934.

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

Francesca Parise and Asuman Ozdaglar. Graphon games. In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 457–458, 2019.

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

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

Sarah Perrin, Mathieu Laurière, Julien Pérolat, Romuald Élie, Matthieu Geist, and Olivier 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.

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

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

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

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder de Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. 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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017.

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

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

Sriram Ganapathi Subramanian, Pascal Poupart, Matthew E. Taylor, and Nidhi 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.

AG 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.

John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in neural information processing systems*, 9, 1996.

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

Deepanshu Vasal, Rajesh K Mishra, and Sriram Vishwanath. Master equation of discrete time graphon mean field games and teams. *arXiv preprint arXiv:2001.05633*, 2020.

Lingxiao Wang, Zhuoran Yang, and Zhaoran 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.

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

Batuhan Yardim and Niao He. Exploiting approximate symmetry for efficient multi-agent reinforcement learning. *arXiv preprint arXiv:2408.15173*, 2024.

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

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

Batuhan Yardim, Artur Goldman, and Niao He. When is mean-field reinforcement learning tractable and relevant?, 2024.

Muhammad Aneeq Uz Zaman, Alec Koppel, Sujay Bhatt, and Tamer 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.

Fengzhuo Zhang, Vincent YF Tan, Zhaoran Wang, and Zhuoran Yang. Learning regularized monotone graphon mean-field games. *arXiv preprint arXiv:2310.08089*, 2023.