PURE, MIXED AND BEHAVIORAL STRATEGY LEARNING IN MEAN FIELD GAMES

Anonymous authors
Paper under double-blind review

ABSTRACT

Mean Field Games (MFGs) provide a scalable framework for multi-agent reinforcement learning, but traditionally rely on two assumptions: agents are homogeneous (i.e., they use identical policies) and their policies are stochastic (i.e., their actions are random). We challenge the necessity for these two foundational assumptions by establishing equivalence between standard MFGs and a novel system with heterogeneous agents playing deterministic strategies. We prove that for any behavioral strategy Nash equilibrium (NE) in an MFG, there exists an equivalent mixed strategy NE over a population of agents with diverse, deterministic policies. This is analogous to classical game-theoretical insights from Kuhn's theorem, suitably adjusted to MFGs for the first time. The equivalence is constructive, enabling the approximation of Nash equilibria in large N-agent games using only pure strategies, which are simpler to implement and analyze. Our insight reframes the role of stochasticity in MFGs as a population-level sufficient condition rather than a necessary requirement for individual agents. Algorithmically, this perspective gives rise to a new class of learning methods, including Fictitious Play and finite-agent variants that operate with pure policies. We explore the performance, scalability, and behavioral implications of this framework on multi-agent reinforcement learning tasks, leveraging deep reinforcement learning for general tractability.

1 INTRODUCTION

Background. Mean Field Games (MFGs) have become indispensable tools for the analysis and control of systems with a massive number of interacting agents (Lasry & Lions, 2007; Huang et al., 2006; Carmona & Delarue, 2018). By focusing on a representative agent's decision-making process in response to the mean field, i.e., the aggregate distribution of the entire population, these frameworks elegantly bypass the curse of dimensionality that plagues traditional N-agent game theory. The recent synergy between MFGs and multi-agent reinforcement learning (MARL) has catalyzed a wealth of research, applying these scalable models to increasingly complex problems, see e.g. (Guo et al., 2019; Subramanian & Mahajan, 2019; Fu et al., 2019; Elie et al., 2020; Cui & Koeppl, 2021; Angiuli et al., 2022; Yardim et al., 2023; Gu et al., 2025; Laurière et al., 2022) among others.

However, the standard MFG framework rests on two foundational pillars: **agent homogeneity** (all agents are identical and employ the same policy) and **policy stochasticity** (the policy is behavioral with random actions at each decision step). While powerful, these assumptions create a gap with real-world applications: agents, from financial traders to autonomous vehicles, are often characterized by significant heterogeneity, see e.g. (Harrison & Kreps, 1978; Schwarting et al., 2019; Bettini et al., 2023). Furthermore, from an engineering standpoint, deterministic policies are simpler and cheaper to execute than stochastic ones. Last, several works proved convergence of algorithms (such as Fictitious Play Perrin et al. (2020)) using an averaged policy, but used an empirical average of pure policies in implementation, leaving open convergence for actual algorithms used in experiments.

This paper demonstrates that using homogeneous mixed policies is not a fundamental requirement but rather a modeling choice. We prove that a system of identical agents playing a shared stochastic policy is formally equivalent to a system of heterogeneous agents, each playing a deterministic pure strategy sampled from a population-level mixed strategy. This insight opens a new MFG paradigm where equilibria can be realized by learning a diverse population of deterministic agents – endogenous heterogeneity that is both more realistic for many systems and algorithmically appealing/interpretable.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Research Questions. This work is driven by fundamental questions that challenge the conventional understanding of MFGs and open new avenues for algorithmic design:

- **RQ1: Theoretical Equivalence and Sufficiency.** Can we formalize an MFG system based on a population of heterogeneous, deterministic agents? Is this system truly equivalent to the standard, homogeneous stochastic model, and is such a population of deterministic agents sufficient to learn approximate Nash equilibria in the corresponding N-agent game?
- **RQ2: Algorithmic Analysis and Comparison.** Can this new perspective lead to practical learning algorithms or even explain existing ones? How do algorithms that learn a *distribution over pure policies* compare to standard algorithms that learn a *single behavioral policy* in terms of convergence speed, solution quality, and microscopic/macroscopic behavior?
- **RQ3: Applicability of Reinforcement Learning.** Can we make this framework practical for problems with large state spaces? Specifically, can we use modern Deep Reinforcement Learning (DRL) techniques to avoid enumerating all pure strategies?

Contributions. We provide affirmative *constructive* answers via the following main contributions:

1. **Formal Equivalence:** We prove equivalence between the standard MFG (identical, behavioral strategies) and a novel system with a mixed strategy over heterogeneous, pure strategies: equilibria in one map directly to equilibria in the other.
2. **Sufficiency of Pure Strategies:** We prove that heterogeneous deterministic policies are sufficient for constructing approximate Nash equilibria in N-agent games. We clarify that identical deterministic policies are insufficient, and that the choice is between identical stochastic or heterogeneous deterministic ones – unique for MFGs over normal-form games.
3. **Novel Algorithmic Framework:** We leverage this perspective to design a new class of algorithms with convergence guarantees that operate in the space of pure policies; this closes a gap between previous theoretical and empirical results of seminal past works.
4. **Bridge to Deep RL:** We propose a practical method to scale these algorithms to complex environments by using Deep RL to compute pure-strategy best responses.

2 Related Work

Foundations of Mean Field Games. MFGs were introduced by Lasry & Lions (2007) and Huang et al. (2006); for discrete-time MFGs, see (Gomes et al., 2010; Saldi et al., 2018). The core result linking the limiting MFG equilibrium to an ε-Nash equilibrium in the finite-agent game (Carmona & Delarue, 2018; Saldi et al., 2018) is the basis upon which we build our new deterministic formulation.

Learning in Mean Field Games. A significant body of work has focused on learning MFG equilibria. Fictitious Play (FP) (Cardaliaguet & Hadikhanloo, 2017; Hadikhanloo & Silva, 2019; Perrin et al., 2020) and other algorithms involving reinforcement learning (Guo et al., 2019; Subramanian & Mahajan, 2019; Fu et al., 2019; Elie et al., 2020; Cui & Koeppl, 2021; Angiuli et al., 2022) operate by learning a single, homogeneous behavioral policy. Our work provides the theoretical foundation for a different class of algorithms that learn a distribution over pure policies. The distinction from the behavioral FP analyzed by Perrin et al. (2020) is that their averaged object is a behavioral policy in which each past best response is weighted by the probability of actually visiting states; computing those weights requires simulating every past policy against the current mean field. Our Proposition 2 instead averages in the space of *strategies*, which requires only the list of past best responses. This matters in practice because averaging neural policies θ_1, θ_2 is difficult: $\pi_{(\theta_1+\theta_2)/2} \neq \frac{1}{2}(\pi_{\theta_1} + \pi_{\theta_2})$. Existing implementations correspondingly store past pure best responses and sample from them rather than forming the behavioral average (Perrin et al., 2020); our Proposition 2 supplies the convergence guarantee for that implemented method, which to the best of our knowledge has not been covered yet.

Heterogeneous Agents in MFGs. While extensions to heterogeneous populations exist, they typically model agents with fixed and *exogenous* "types" (with different dynamics or reward functions), see e.g. (Feleqi, 2013; Cirant, 2015; Caines & Huang, 2021). Our work shows that behavioral heterogeneity arises *endogenously* even in a population of structurally identical agents, as a mechanism to achieve equilibrium. Guan et al. (2024) establish that identical policies suffice over heterogeneous ones in the special case of two-team cooperative Mean Field Control (MFC). We instead provide an opposite result in the case of non-cooperative games (namely, that heterogeneous policies suffice over identical policies in MFGs), and we only consider deterministic strategies.

Mixed Strategies in Classical Game Theory. Our equivalence is close in spirit to Kuhn's Theorem (Kuhn, 1953), which shows that mixed and behavioral strategies coincide for a single agent, but the settings differ. Kuhn's environment is exogenous and fixed, whereas in an MFG the mean field is an *endogenous fixed point* determined by the population's own strategy. Theorems 1 and 2 show that this fixed point is invariant under the transformation between systems; the proof proceeds by induction over time steps, and the mean field depends on the strategy and vice versa through a fixed-point equation, a circularity with no analogue in Kuhn's setting. Similarly, the nonatomic-games literature (Aumann, 1964; Schmeidler, 1973) establishes *existence* of pure-strategy equilibria in *static* continuum games. Our result is dynamic, with state transitions; it is constructive rather than existential; and it comes with explicit finite-N rates (Theorems 3 and 4).

3 Mean Field Games: From Homogeneous to Heterogeneous Pure

We first formalize the standard MFG model, which is restricted to identical stochastic policies (Guan et al., 2024), and then introduce our equivalent heterogeneous pure-strategy formulation.

Notation: The time horizon is $\mathcal{T} \coloneqq \{0, \dots, T-1\}$ for a finite $T \geq 1$, and the state space $\mathcal{X}$ and action space $\mathcal{U}$ are finite sets with $|\mathcal{X}|, |\mathcal{U}| < \infty$. For a compact metric space $\mathcal{Z}$, $\mathcal{P}(\mathcal{Z})$ is the space of probability measures on it. $\Pi \coloneqq \{\pi : \mathcal{T} \times \mathcal{X} \to \mathcal{P}(\mathcal{U})\}$ is the space of behavioral (stochastic) policies. $\Pi_{\text{pure}} \coloneqq \{\pi^p : \mathcal{T} \times \mathcal{X} \to \mathcal{U}\}$ is the space of pure (deterministic) policies. $\mathcal{S} \coloneqq \mathcal{P}(\Pi_{\text{pure}})$ is the space of mixed strategies. Note that many MFG papers use the term "mixed policy" for what we call a "behavioral policy" for distinction. We let $\mathbb{E}_Y[X] = \mathbb{E}[X \mid Y]$ denote the conditional expectation of X given Y. In the following, detailed proofs for statements are found in the Appendix.

3.1 Standard MFGs

The study of MFGs is motivated by analysis and equilibrium learning in finite N-agent games, which is well-known to be hard to scale. The two main assumptions of standard MFGs are 1. weak, "anonymous" interaction of agents through their distribution – the mean field (MF) – and 2. classic homogeneity of agents in behavior. The latter behavioral homogeneity is addressed in this work.

Consider a finite N-agent game defined by agents $i \in [N]$, each of which is equipped at any time $t \in \mathbb{N}_0$ with random states x_t^i and actions u_t^i. The random states and actions are valued in a finite state space $\mathcal{X}$ and action space $\mathcal{U}$. The *empirical* MF is then defined as $\mu_t^N \coloneqq \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{x_t^i}$ and drives agent interactions through some transition kernel p such that $x_{t+1}^i \sim p(x_{t+1}^i \mid x_t^i, u_t^i, \mu_t^N)$. Initial states $x_t^i \sim \mu_0$ are sampled from an initial distribution $\mu_0 \in \mathcal{P}(\mathcal{X})$. Agents i may then take actions according to some stochastic policies $\pi^i \in \Pi$, with Π as defined above, such that for all i,

$$u_t^i \sim \pi^i(u_t^i \mid x_t^i), \quad x_{t+1}^i \sim p(x_{t+1}^i \mid x_t^i, u_t^i, \mu_t^N). \tag{1}$$

Each agent is also equipped with its own objective $J^i(\pi^1, \dots, \pi^N) = \mathbb{E}\left[\sum_t r(x_t^i, u_t^i, \mu_t^N)\right]$, for reward function $r\colon \mathcal{X} \times \mathcal{U} \times \mathcal{P}(\mathcal{X}) \to \mathbb{R}$. This gives rise to the classical Nash equilibrium as target.

Definition 1. *For any $\varepsilon \geq 0$, an approximate ε-Nash equilibrium is a tuple $\underline{\pi} = (\pi^1, \dots, \pi^N) \in \Pi^N$ of policies such that $\sup_{\hat{\pi} \in \Pi} J^i(\hat{\pi}, \underline{\pi}^{-i}) \leq J^i(\underline{\pi}) + \varepsilon$ for all agents $i \in [N]$, where $\underline{\pi}^{-i}$ denotes all policies other than the i-th.*

The minimal such ε is also called the exploitability of $\underline{\pi}$. To obtain the mean field limit, in existing MFG work it is *assumed* that all agents play the same stochastic policy. Then, in the infinite agent limit $N \to \infty$, by law of large numbers and homogeneity of agents, we instead obtain the MFG where the empirical MF μ_t^N is replaced by its deterministic limiting MF μ_t. The MF μ_t is given as the law of any representative agent playing the shared, *identical* and *stochastic* policy π, i.e. $\pi^1 = \ldots = \pi^N = \pi$. By the preceding dynamics, we therefore have the mean field $\mu = (\mu_t)_{t \in \mathcal{T}}$ given as a function of π, $\mu =: \Lambda(\pi)$ such that $\mu_{t+1} = \int_{\mathcal{X}} \int_{\mathcal{U}} p(\cdot \mid x, u, \mu_t)\pi_t(\mathrm{d}u \mid x)\mu_t(\mathrm{d}x)$. Then, for any single agent deviating from policy π and instead applying policy $\hat{\pi} \in \Pi$, we have

$$u_t \sim \hat{\pi}_t(u_t \mid x_t), \quad x_{t+1} \sim p_t(x_{t+1} \mid x_t, u_t, \mu_t) \tag{2}$$

and its corresponding objective $J(\hat{\pi}, \pi) \coloneqq J(\hat{\pi}, \Lambda(\pi))$ with $J(\hat{\pi}, \mu) \coloneqq \mathbb{E}\left[\sum_t r(x_t, u_t, \mu_t)\right]$.

In the standard MFG, a population of identical agents plays a common behavioral policy $\pi \in \Pi$. This policy generates a mean-field flow $\mu = \Lambda(\pi)$. A Mean Field Nash Equilibrium (MFNE) is a policy π^* that is a best response to its own induced mean field: $J(\pi^*, \Lambda(\pi^*)) = \sup_{\hat{\pi} \in \Pi} J(\hat{\pi}, \Lambda(\pi^*))$.

Definition 2. *A MFNE is defined as a policy* $\pi \in \Pi$ *such that* $J(\pi, \Lambda(\pi)) = \sup_{\hat{\pi}\in\Pi} J(\hat{\pi}, \Lambda(\pi))$.

The motivation of the above is given by its equivalence with the Nash equilibrium. The MFNE is given by a policy $\pi \in \Pi$ such that no deviation from π would improve an agents objective under the MF $\mu = \Lambda(\pi)$ generated by π. Solving for a MFNE is then known to approximately solve for an otherwise difficult approximate Nash equilibrium in the finite game.

Proposition 1 ((Saldi et al., 2018)). *Let* π^* *be a MFNE. Then, in the* N*-agent game, the exploitability of* $(\pi^*, \dots, \pi^*) \in \Pi^N$ *is* $\epsilon(N)$ *with* $\epsilon(N) \to 0$ *as* $N \to \infty$.

3.2 Heterogeneous Pure-Strategy MFGs

We propose an alternative, limit-equivalent formulation where each agent samples a pure strategy $\pi^p \in \Pi_{\text{pure}}$ from a population-level mixed strategy $\sigma \in \mathcal{S}$. This mixed strategy generates a mean-field flow $\mu = \bar{\Lambda}(\sigma)$. A Strategy MFNE (SMFNE) is a mixed strategy σ^* that is a best response to its own induced mean field, meaning no agent can benefit by unilaterally changing its distribution over pure strategies. Formally, $\mathcal{J}(\sigma^*, \bar{\Lambda}(\sigma^*)) = \sup_{\hat{\sigma}\in\mathcal{S}} \mathcal{J}(\hat{\sigma}, \bar{\Lambda}(\sigma^*))$. Analogously to $\Lambda : \Pi \to \mathcal{P}(\mathcal{X})^{\mathcal{T}}$, we define the function $\bar{\Lambda}$ such that $\bar{\Lambda}(\sigma)$ is the MF μ defined by: μ_0 is given, and for all $t \geq 1$,

$$\mu_t(x) = \int_{\Pi_{\text{pure}}} \mathbb{P}(x_t = x \mid \pi^p)\sigma(\mathrm{d}\pi^p) \tag{3}$$

$$= \int_{\Pi_{\text{pure}}} \int_{\mathcal{X}^t} p(x \mid x_{t-1}, \pi^p_{t-1}(x_{t-1}), \mu_{t-1}) \prod_{s=0}^{t-2} p(\mathrm{d}x_{s+1} \mid x_s, \pi^p_s(x_s), \mu_s)\mu_0(\mathrm{d}x_0)\sigma(\mathrm{d}\pi^p).$$

Given a MF μ, for any representative agent using a distribution $\hat{\sigma} \in \mathcal{S}$, we have

$$\pi^p \sim \hat{\sigma},\ u_t = \pi^p_t(x_t),\ x_{t+1} \sim p_t(x_{t+1} \mid x_t, u_t, \mu_t). \tag{4}$$

The corresponding objective is $\mathcal{J}(\hat{\sigma}, \sigma) \coloneqq \mathcal{J}(\hat{\sigma}, \bar{\Lambda}(\sigma))$,

$$\mathcal{J}(\hat{\sigma}, \mu) \coloneqq \mathbb{E}_{\pi^p\sim\hat{\sigma}}\left[J(\pi^p, \mu)\right]. \tag{5}$$

Definition 3. *A Strategy MFNE (SMFNE) is a strategy* $\sigma \in \mathcal{S}$ *s.t.* $\mathcal{J}(\sigma, \bar{\Lambda}(\sigma)) = \sup_{\hat{\sigma}} \mathcal{J}(\hat{\sigma}, \bar{\Lambda}(\sigma))$.

3.3 Remarks and Notation

We now define some notation for MFs generated under policies $\hat{\pi}$ and strategies $\hat{\sigma}$ for representative deviating agents, and for all other agents, π, σ, in systems (2) and (4).

Mean Field Sequences. Let μ^π_t and μ^σ_t be the MFs generated by $\pi \in \Pi$ and $\sigma \in \mathcal{S}$ in systems (2) and (4) respectively. For the deviating agent in (2), whenever all other agents follow $\pi \in \Pi$, we denote the law of the deviating agent under $\hat{\pi} \in \Pi$ as $\mu^\pi_{t|\hat{\pi}}$. Similarly, for the deviating agent in (4), whenever other agents follow $\sigma \in \mathcal{S}$, we denote the law of the deviating agent under $\hat{\sigma} \in \mathcal{S}$ as $\mu^\sigma_{t|\hat{\sigma}}$.

Hence, we can define the objectives of deviating agent, given other agent's behaviors

$$J(\hat{\pi}, \pi) = \sum_t \langle \mu^\pi_{t|\hat{\pi}}, r(\cdot, \mu^\pi_t)\rangle, \quad \mathcal{J}(\hat{\sigma}, \sigma) = \sum_t \langle \mu^\sigma_{t|\hat{\sigma}}, r(\cdot, \mu^\sigma_t)\rangle \tag{6}$$

where we defined the inner product $\langle f, g\rangle = \sum_{x\in\mathcal{X}} f(x)g(x)$ for $f, g \in \mathbb{R}^{\mathcal{X}}$.

Lastly, we define $\mu^\sigma_{t|\pi}$ as the conditional probability in (4) of deviating agents playing fixed π, given all other agents play mixed strategy σ. Then, $\mu_t(x) = \int_{\Pi_{\text{pure}}} \mu^\sigma_{t|\pi^p}(x)\sigma(\mathrm{d}\pi^p)$ where

$$\mu^\sigma_{t|\pi^p}(x) \coloneqq \int_{\mathcal{X}^t} p(x \mid x_{t-1}, \pi^p_{t-1}(x_{t-1}), \mu_{t-1}) \prod_{s=0}^{t-2} p(\mathrm{d}x_{s+1} \mid x_s, \pi^p_s(x_s), \mu_s)\mu_0(\mathrm{d}x_0). \tag{7}$$

Dependence on Actions. Note that dependence on action distributions is omitted w.l.o.g., as we can rewrite any system with rewards depending on the action into another system that omits this dependence, e.g., by augmenting the state space into $\mathcal{X} \cup \mathcal{X} \times \mathcal{U}$ and splitting each time step into two: A choice of action, and a transition of states.

Connection to Correlated Equilibria. Our definition of a Strategy MFNE bears a formal resemblance to Correlated Equilibria (CE) in MFGs (Campi & Fischer, 2022; Muller et al., 2022). In a CE, a correlation device recommends a policy to each agent, and it is an equilibrium if no agent has incentive to disobey. However, the motivation differs fundamentally. While CE expands the set of equilibria, our motivation is *structural and algorithmic*: we introduce heterogeneity to show that agent stochasticity is not required for existence. This enables learning algorithms strictly on the space of deterministic policies, with significant implications for implementation and interpretability.

3.4 Equivalence of Equilibria

Our central theoretical result is that these two formulations are equivalent. For any $\sigma \in \mathcal{S}$, let $\zeta(\sigma) \in \Pi$ denote the behavioral policy that prescribes at each state-time pair the action distribution induced by σ among the agents that *reach* that pair, i.e. for all (t, x) with $\mu_t^\sigma(x) > 0$,

$$\zeta(\sigma)_t(u \mid x) := \frac{\int_{\Pi_{\text{pure}}} \mu^\sigma_{t|\pi^p}(x) \mathbf{1}_{\{\pi_t^p(x)=u\}} \sigma(\mathrm{d}\pi^p)}{\int_{\Pi_{\text{pure}}} \mu^\sigma_{t|\pi^p}(x) \sigma(\mathrm{d}\pi^p)}, \tag{8}$$

and defined arbitrarily (e.g., uniformly) where $\mu_t^\sigma(x) = 0$, since such pairs are reached with probability zero. The state-visitation weights $\mu^\sigma_{t|\pi^p}$ are themselves determined by σ through the mean field, coupling this transformation to the fixed-point equation. Conversely, for any $\pi \in \Pi$, there exists an equivalent mixed strategy $\bar{\sigma} = \zeta^{-1}(\pi) \in \mathcal{S}$ that induces the same action probabilities as π at each state and time. For finite $\mathcal{X}, T, \mathcal{U}$, this means we can construct σ by assigning weights to each pure policy $\pi^p \in \Pi_{\text{pure}}$ such that for every (x, t, u), the sum of weights of pure policies that prescribe action u at (x, t) equals $\pi_t(u \mid x)$. The existence of such an equivalent mixed strategy is guaranteed by $\bar{\sigma}(\pi^p) = \prod_{t\in\mathcal{T}} \prod_{x\in\mathcal{X}} \pi_t(\pi_t^p(x) \mid x)$ (see App. A.1). We write $\zeta^{-1}(\pi)$ for this particular product-measure preimage. Note that ζ is not injective, as many mixed strategies induce the same behavioral policy, so the correspondence below is onto rather than one-to-one.

Theorem 1 (Equivalence in MFs between mixed and behavioral strategies)**.** *Let $\sigma \in \mathcal{S}$, then $\Lambda(\zeta(\sigma)) = \bar{\Lambda}(\sigma)$. Conversely, let $\pi \in \Pi$, then $\Lambda(\pi) = \bar{\Lambda}(\zeta^{-1}(\pi))$.*

Corollary 1 (Equivalence in mean fields)**.** *The images of Λ and $\bar{\Lambda}$ as subsets of $\mathcal{P}(\mathcal{X})^{\mathcal{T}}$ are equal.*

Theorem 2 (Equivalence in conditional MFs)**.** *For any $\pi \in \Pi$, $\sigma \in \mathcal{S}$ with $\pi = \zeta(\sigma)$, the images of the maps $\hat{\pi} \mapsto (\mu^\pi_{t|\hat{\pi}})_{t\in\mathcal{T}}$ and $\hat{\sigma} \mapsto (\mu^\sigma_{t|\hat{\sigma}})_{t\in\mathcal{T}}$ are equal.*

Corollary 2 (Equivalence of equilibria)**.** *Let π be a MFNE, then $\sigma = \zeta^{-1}(\pi) \in \mathcal{S}$ is a SMFNE, and vice versa: Let $\sigma \in \mathcal{S}$ be a SMFNE, then $\pi = \zeta(\sigma) \in \Pi$ is a MFNE.*

Corollary 3 (Equivalence of equilibria in behavior and value)**.** *Equilibria correspond across the two systems. 1. For any MFNE (behavioral) π^*, there exists an equivalent SMFNE (mixed) σ^* that generates the same mean field flow and yields the same value. 2. For any SMFNE σ^*, there exists an equivalent MFNE π^* that generates the same mean field flow and yields the same value.*

Theorem 3 (Sufficiency of Pure Strategies for N-Agent Games)**.** *Let σ^* be a SMFNE. Let p, r be Lipschitz. For any N, we can construct a tuple of pure policies $(\pi^1, \ldots, \pi^N) \in (\Pi_{\text{pure}})^N$ that constitutes an $\epsilon(N)$-Nash equilibrium in the N-agent game, where $\epsilon(N) \to 0$ as $N \to \infty$.*

The results guarantee we can find near-optimal behaviors in large multi-agent systems by searching only within the space of deterministic policies, provided we allow for heterogeneity instead of stochasticity. However, it requires carefully assigning policies to approximate equilibria. Practical decentralized scenarios often involve agents selecting strategies independently. We now analyze such a scenario and prove that with high probability, we still have an approximate Nash equilibrium.

Theorem 4 (High-Probability ϵ-Nash Equilibrium with Sampled Policies)**.** *Let σ^* be a SMFNE. Assume reward function r and transition kernel p are Lipschitz continuous in the MF with constants L_r and L_p respectively. Consider a system of N agents where each agent i independently samples a pure strategy $\pi^i \sim \sigma^*$. For any confidence level $\delta \in (0, 1)$, there exists a constant $C > 0$ (independent of N) such that with probability at least $1 - \delta$, the random tuple of pure policies $\underline{\pi} = (\pi^1, \ldots, \pi^N)$ constitutes an ε-Nash equilibrium in the N-agent game, with: $\varepsilon \leq C\sqrt{(1 + \log(1/\delta))/N}$.*

3.5 Insufficiency of Homogeneous Pure Policies

A crucial consequence is that heterogeneity is necessary when using pure strategies.

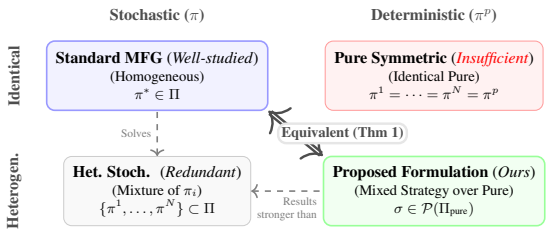

Figure 1: **Taxonomy of MFG Formulations.** (Top-Left) Standard MFG uses identical stochastic policies. (Top-Right) Deterministic identical agents are insufficient for equilibria. (Bottom-Right) Our contribution is the diagonal link: proving standard MFG equivalence to a population of deterministic agents. (Bottom-Left) Heterogeneous stochastic agents are most permissive and hence also enabled.

Counterexample 1 (Homogeneous Pure Policies are Not Enough)**.** *Consider a stateless Rock-Paper-Scissors MFG. The unique MFNE is the behavioral strategy $\pi^*(R) = \pi^*(P) = \pi^*(S) = 1/3$. Now, consider a homogeneous population playing a pure strategy, e.g., "always play Rock". The mean field is $\mu = (1, 0, 0)$. Any agent can improve its payoff from 0 to 1 by deviating to "play Paper". Thus, no homogeneous pure-strategy profile can be a MFNE. An equilibrium requires stochasticity, which can be realized either by identical agents using stochastic policies, or by a heterogeneous population of agents using deterministic policies (e.g., 1/3 playing R, 1/3 playing P, 1/3 playing S).*

4 Algorithms and Learning Paradigms

Our framework motivates a new class of learning algorithms with agents learning pure strategies. We propose two natural extensions of Fictitious Play.

- **Mixed Fictitious Play (MFP):** Maintain a mixed strategy σ_k. At each step, it computes a pure-strategy best response to the current mean field and mixes it into σ_k.
- **Deterministic Fictitious Play (DFP):** A practical N-agent version. Each agent holds a pure strategy and at each step, agents can randomly switch to the newest best response.

Under a standard monotonicity assumption, MFP converges to an SMFNE; DFP is its finite-agent realization, with the same limit behaviour in the experiments of Section 5.

Proposition 2 (Convergence of MFP)**.** *Under Lasry-Lions monotonicity and independence of p from the MF, the exploitability of the average strategy generated by MFP converges to 0.*

We stress the relevance of this result for the MFG literature because the MFP algorithm we consider is close to the algorithm that was *actually* implemented in several papers (where to implement standard behavioral FP, one must remember and keep track of all the past BR policies – see discussions in e.g. (Perrin et al., 2020; Lauriere et al., 2022)). However, in contrast to (Perrin et al., 2020, p.3 bottom), the complex policy averaging is replaced by straightforward averaging of strategies.

DFP as the finite-agent realization of MFP. Algorithm 2 differs from Algorithm 1 only in that the population is finite and each agent switches to the current best response at random and independently. Consequently the *conditional mean* of a DFP update is exactly the MFP update, the difference being conditionally centred noise of magnitude $\mathcal{O}(N^{-1/2})$ (App. A.6). We therefore view DFP as MFP driven by finite-agent noise. The experiments of Section 5 confirm that it reaches the same equilibria as MFP and as behavioural Fictitious Play.

4.1 Scaling with Deep RL: Learning the Best Response

The main challenge for the algorithms above is that the set of pure strategies Π_{pure} is too large for most interesting problems. We address this by replacing the explicit enumeration and evaluation of all pure strategies with a DRL oracle. The process is illustrated in Fig. 2. At each iteration k of MFP/DFP, instead of computing $J(\pi^p, \mu_k)$ for all π^p, we solve for the best (pure) strategy: $\pi^{\text{BR},k+1} \in \arg\max_{\pi^p \in \Pi_{\text{pure}}} J(\pi^p, \mu_k)$. This is a standard single-agent RL problem where the agent's environment is fixed by the mean field μ_k. We can solve it using any off-the-shelf DRL

algorithm (e.g., DQN for discrete actions, PPO for continuous actions). The learned policy network for $\pi^{\mathrm{BR},k+1}$ is then added to our explicit set of "good" pure strategies, and the mixed strategy is updated over this growing set. This makes the framework practical for any state and action spaces.

Pure best responses from DRL oracles. Our implementation uses a value-based oracle (Double-DQN), for which purity is automatic: the greedy policy $\pi^{\mathrm{BR}}(x,t) = \arg\max_u Q(x,t,u)$ is a deterministic map by construction. Purity is also not a real restriction. Once the mean field μ_k is fixed, the best-response problem is a finite-horizon MDP, which admits a deterministic Markov optimal policy; a stochastic optimal policy therefore attains no larger value, and if it randomizes among optimal actions, selecting a single maximizing action at each state-time pair is still a best response. Policy-gradient methods such as PPO can accordingly serve as the oracle when followed by deterministic selection: either anneal the entropy bonus to zero, or extract the mode $\pi^{\mathrm{pure}}(x,t) = \arg\max_u \pi_\theta(u \mid x,t)$ at deployment time, and for continuous actions deterministic policy gradient methods produce a pure action directly. Purity is therefore convenient rather than necessary.

5 Experiments

Our experiments are designed to rigorously test our research questions, progressing from core validation on canonical games to larger environments solved via DRL.

5.1 Experimental Setup

We evaluate our framework across four distinct environments: the **Braess Network Congestion** game (Braess et al., 2005), which tests traffic decision-making under congestion effects; the classic **Rock-Paper-Scissors** game, serving as a static validation case for mixed strategy equilibria; the **Susceptible-Infected-Susceptible (SIS)** model (Kermack & McKendrick, 1927), a dynamic game representing epidemic control; and the well-known **Beach-Bar** game (Arthur, 1994; Perrin et al., 2020), serving as a dynamic validation case with larger state spaces.

To benchmark performance, we compare our proposed methods against **Standard Fictitious Play (FP)**, which learns a single homogeneous behavioral policy. Concretely, our FP baseline is the *theoretical* behavioral Fictitious Play of Perrin et al. (2020), which maintains one behavioral policy obtained by averaging past pure best responses under state-visitation weights, $\bar{\pi}_t^k(u \mid x) \propto \sum_{j \le k} \mu_t^{\pi^j}(x) \mathbf{1}_{\{\pi_t^j(x)=u\}}$. Our primary contributions are evaluated through **Mixed Fictitious Play (MFP)**. Additionally, we implement **Deterministic Fictitious Play (DFP)** as the finite-agent realization and its scalable variant, **DFP-DRL**, which utilizes Deep RL for best responses.

Performance is assessed using three key metrics. First, the **Exploitability** quantifies the incentive for one agent to deviate from the learned population. Lower values indicate closer proximity to a Nash equilibrium. Second, we look at integration with Deep RL, tracking the **number of iterations** required to reach low exploitability. Finally, we examine **Behavioral Diversity** by analyzing learned mixed and behavioral strategies, providing insight into the heterogeneity of the agent population.

5.2 Algorithmic Validation

We verify that MFP and DFP reliably converge to sensible equilibria in canonical game-theoretic settings (**RQ1**), similar to standard Fictitious Play (FP) on three benchmark problems: Rock-Paper-Scissors, the Braess Network Congestion game, and the SIS epidemic model. We find that all methods converge to low-exploitability policies, with MFP and DFP finding equilibria that are qualitatively

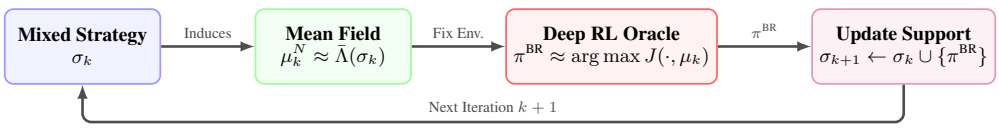

Figure 2: **The DFP-DRL Iterative Loop.** At each iteration k, the current mixed strategy σ_k generates a random fixed mean field environment μ_k^N. A Deep RL agent (Oracle) is trained to find the optimal *deterministic* policy π^{BR} against this environment, avoiding enumeration of the full strategy space.

	Peak memory (MB)			Persistent storage (MB)		
$\lvert\mathcal{U}\rvert$	FP	MFP (ours)	ratio	FP	MFP (ours)	ratio
50	0.077	0.048	1.6×	0.031	0.010	3.2×
100	0.138	0.054	2.6×	0.061	0.010	6.3×
200	0.260	0.066	3.9×	0.122	0.010	12.6×
500	0.535	0.102	5.2×	0.305	0.010	31.5×
1000	0.992	0.163	6.1×	0.611	0.010	63.0×

Table 1: Memory scaling on the adjustable SIS model ($|\mathcal{X}| = 2$, $T = 20$, 30 iterations) as action space grows. Both methods have comparable exploitability (FP: 0.0083; MFP: 0.0112; at $|\mathcal{U}| = 1000$).

similar to the standard homogeneous approach in macroscopic mean fields. The convergence of all three algorithms across all three canonical games is shown in Fig. 3. In the Rock-Paper-Scissors game, all three methods converged to policies that were quantitatively very close to the known $(1/3, 1/3, 1/3)$ Nash Equilibrium. In the SIS game, all methods converged to a stable equilibrium where the level of protective vaccination balanced the infection risk. These results provide empirical support for our framework, demonstrating that MFP and DFP are effective and that the equilibria they find are macroscopically consistent with those found by classical methods. For a more detailed comparison, we analyzed a final equilibrium from the simplified Braess Congestion game ($T = 2$). All methods successfully found the suboptimal Braess paradox Nash equilibrium, see Table 2 in the appendix. We also find that their resulting macroscopic behaviors did not differ significantly.

5.3 Memory and Storage Scaling

The equivalence of Section 3 also has a concrete resource consequence. A behavioral policy must store a probability vector at every state-time pair, $O(|\mathcal{X}| \cdot |\mathcal{U}| \cdot T)$ entries in total, whereas a library of pure policies stores a single action per state-time pair, $O(|\mathcal{X}| \cdot T)$ entries each and independently of $|\mathcal{U}|$. MFP therefore trades the action space $|\mathcal{U}|$ against the number of stored best responses K.

To quantify this, we compare MFP and FP on an SIS model ($|\mathcal{X}| = 2$, $T = 20$, 30 iterations) in which protection effectiveness levels can be chosen, and we vary $|\mathcal{U}| \in \{50, 100, 200, 500, 1000\}$. Both methods use *streaming* best-response computation: neither materializes the full $T \times |\mathcal{X}| \times |\mathcal{U}|$ table, and each timestep handles a single $|\mathcal{X}| \times |\mathcal{U}|$ slice. Table 1 reports peak and persistent memory, measured over best-response computation and policy storage together across iterations.

Peak memory is dominated by FP's running behavioral average, so the ratio approaches T as $|\mathcal{U}|$ grows, reaching 6.1× at $|\mathcal{U}| = 1000$. Persistent storage grows linearly in $|\mathcal{U}|$ for FP but is independent of it for MFP, which retains only the K stored best responses and their weights; the ratio reaches 63.0× at $|\mathcal{U}| = 1000$, with crossover at $K \approx |\mathcal{U}|$. The two methods reach comparable solution quality, so this is not bought at the cost of accuracy. The advantage grows with the action space, which is the regime where storing a probability vector per state-time pair is most costly. In the DRL setting the gap widens further: FP would need to average stochastic networks (Lauriere et al., 2022), whereas MFP stores deterministic ones together with categorical weights.

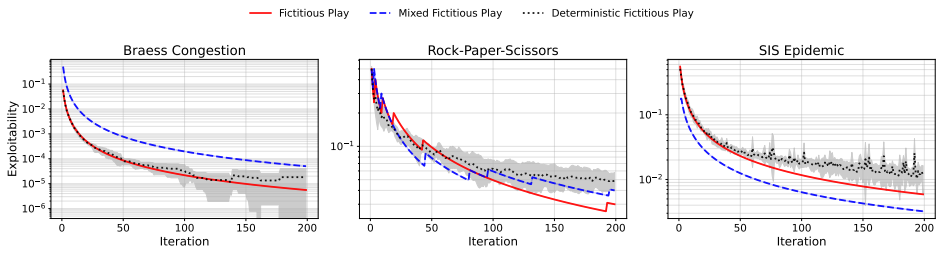

Figure 3: Exploitability over 200 iterations for Fictitious Play (FP), Mixed Fictitious Play (MFP), and Deterministic Fictitious Play (DFP) on the canonical games. The y-axis is on a logarithmic scale.

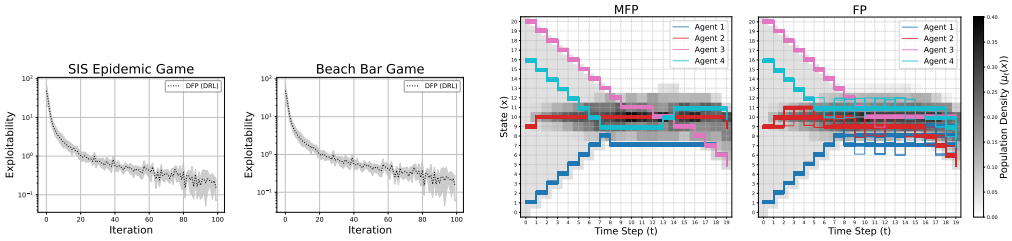

Figure 4: Left: Exploitability over 100 iterations for DFP over large environments. Right: Microscopic and Macroscopic Behaviors of Equilibria in Beach Bar Game: MFP vs. FP.

5.4 Scalability via Reinforcement Learning

To address the challenge of scalability (**RQ3**), we evaluate our proposed framework on higher-dimensional MARL tasks. We employ the Beach Bar environment (a grid world) and an extended SIS setting with larger population and longer horizon. In these settings, we utilize our DFP-DRL algorithm, which integrates Deep Q-Networks (DQN) to compute best responses tractably. We expect this approach remains computationally feasible while successfully converging to high-quality equilibria (details are in App. C). We configured both environments with $1,000$ agents and a time horizon of $T = 20$, resulting in pure policy spaces of sizes 21^{20} and 4^{20}. Although these validation environments are principally solvable via tabular methods, we employ the DFP-DRL architecture here to demonstrate its stability and convergence properties. Fig. 4 presents the learning curves for both environments, using exploitability as primary metric. DFP-DRL successfully converges toward an equilibrium in both settings. This confirms the practicality of our approach: Deep RL techniques can be effectively leveraged to solve MFGs with heterogeneous deterministic policies.

5.5 Behavioral Analysis of Equilibria

Finally, we investigate structural properties of learned equilibria to gain behavioral insights (**RQ2**). We analyze the mixed strategy σ^* and corresponding mean-field policy π^* derived from the earlier equilibria. We conducted a comparative behavioral analysis of the equilibria found by MFP and standard FP in the Beach Bar game to understand how distinct microscopic behaviors give rise to similar macroscopic outcomes. We introduced asymmetric action costs—moving right costs 0, staying costs 1, and moving left costs 2—to induce a drift away from the bar. The MFP best response was computed using DQN, while the FP equilibrium was computed following Perrin et al. (2020). Fig. 4 illustrates the results. To probe microscopic behavior, we randomly initialized four agents and sampled policies from the learned equilibrium for both MFP and FP. We then simulated 1,000 trajectories for each agent. As expected, MFP agents follow deterministic paths, as the environment is deterministic and learned policies are pure. In contrast, FP agents exhibit stochastic trajectories due to inherent randomness of the mixed policies in the FP support. Crucially, despite such fundamental differences in individual behavior, the resulting aggregate population distributions remain similar. This confirms our theory: while the microscopic mechanism of achieving equilibria differs (heterogeneous deterministic vs. homogeneous stochastic), the macroscopic mean field outcome is invariant.

6 Discussion and Conclusion

This work fundamentally reframes the basis of equilibrium learning in MFGs by establishing rigorous equivalence between the conventional homogeneous stochastic model and a population of heterogeneous deterministic agents. We demonstrated that individual stochasticity is not a prerequisite for equilibria, but can instead be realized as a population-level phenomenon arising from diverse pure strategies. This insight bridges the gaps game-theoretic abstractions and realistic multi-agent systems, proving that approximate Nash equilibria in large finite games can be constructed entirely from simple, deterministic behaviors. We restricted attention to the most common setting of mean field games; social optimum, common noise, and partial observability remain to be studied. Algorithmically, our perspective unlocks a new class of learning methods operating directly in the space of pure policies. Overall, our framework offers a memory-scalable and interpretable alternative to standard algorithms.

AI USE STATEMENT

In this work, we used generative AI tools for aiding and polishing writing, and to assist in the writing of proofs. We have not used generative AI tools for the rest of the required disclosure tasks, and the rest of the required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for identifying relevant literature. We have reviewed all AI-assisted work. We checked LLM-generated text for correctness as well as potential plagiarism through a manual literature survey. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work is theoretical and methodological, and we do not foresee direct harmful applications arising from it. It involves no human subjects, no personally identifiable information, and no data collected from human participants. The empirical evaluation is carried out entirely on synthetic benchmark environments, and uses no social or personal data of any kind. As with any method for learning equilibria in large populations, techniques of this kind could in principle inform models of population-level behaviour; we are not aware of any specific misuse that this work enables beyond what is already possible with standard mean field game and multi-agent reinforcement learning methods.

REPRODUCIBILITY STATEMENT

We summarise the efforts made to ensure reproducibility and point to where the necessary detail can be found. The theoretical results are proved in the appendix, with every assumption stated explicitly where it is invoked: the mean field and conditional mean field equivalences in Appendices A.1 and A.2, the construction of the pure-strategy ϵ-Nash equilibrium for the N-agent game in Appendix A.4, the convergence of Mixed Fictitious Play in Appendix A.5, and the relation between Deterministic and Mixed Fictitious Play in Appendix A.6. The experimental protocol is described in Section 5: the four environments together with their state and action spaces, the algorithms compared, and the exploitability metric used to evaluate all of them. The best-response oracle used in the scalable experiments is specified in Appendix C, which gives the Double-DQN update, the network architecture, and all training hyperparameters (Table 3). The finite-N propagation-of-chaos study and the per-algorithm path probabilities for the Braess game are reported in Figure 5 and Table 2.

REFERENCES

Andrea Angiuli, Jean-Pierre Fouque, and Mathieu Laurière. Unified reinforcement Q-learning for mean field game and control problems. *Mathematics of Control, Signals, and Systems*, 34(2): 217–271, 2022.

W Brian Arthur. Inductive reasoning and bounded rationality. *The American economic review*, 84(2): 406–411, 1994.

Robert J Aumann. Markets with a continuum of traders. *Econometrica: Journal of the Econometric Society*, pp. 39–50, 1964.

Matteo Bettini, Ajay Shankar, and Amanda Prorok. Heterogeneous multi-robot reinforcement learning. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1485–1494, 2023.

Dietrich Braess, Anna Nagurney, and Tina Wakolbinger. On a paradox of traffic planning. *Transportation science*, 39(4):446–450, 2005.

Peter E Caines and Minyi Huang. Graphon mean field games and their equations. *SIAM Journal on Control and Optimization*, 59(6):4373–4399, 2021.

Luciano Campi and Markus Fischer. Correlated equilibria and mean field games: A simple model. *Mathematics of Operations Research*, 47(3):2240–2259, 2022.

Pierre Cardaliaguet and Saeed Hadikhanloo. Learning in mean field games: the fictitious play. *ESAIM: Control, Optimisation and Calculus of Variations*, 23(2):569–591, 2017.

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

Marco Cirant. Multi-population mean field games systems with neumann boundary conditions. *Journal de Mathématiques Pures et Appliquées*, 103(5):1294–1315, 2015.

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

Kai Cui, Gökçe Dayanıklı, Mathieu Laurière, Matthieu Geist, Olivier Pietquin, and Heinz Koeppl. Learning discrete-time major-minor mean field games. In *Proc. AAAI*, volume 38, pp. 9616–9625, 2024.

Romuald Elie, Julien Perolat, Mathieu Laurière, Matthieu Geist, and Olivier Pietquin. On the convergence of model free learning in mean field games. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7143–7150, 2020.

Ermal Feleqi. The derivation of ergodic mean field game equations for several populations of players. *Dynamic Games and Applications*, 3(4):523–536, 2013.

Zuyue Fu, Zhuoran Yang, Yongxin Chen, and Zhaoran Wang. Actor-critic provably finds Nash equilibria of linear-quadratic mean-field games. In *International Conference on Learning Representations*, 2019.

Diogo A Gomes, Joana Mohr, and Rafael Rigao Souza. Discrete time, finite state space mean field games. *Journal de Mathématiques Pures et Appliquées*, 93(3):308–328, 2010.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field multiagent reinforcement learning: A decentralized network approach. *Mathematics of Operations Research*, 50(1):506–536, 2025.

Yue Guan, Mohammad Afshari, and Panagiotis Tsiotras. Zero-sum games between mean-field teams: Reachability-based analysis under mean-field sharing. In *Proc. AAAI*, volume 38, pp. 9731–9739, 2024.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. *Advances in Neural Information Processing Systems*, 32:4966–4976, 2019.

Saeed Hadikhanloo and Francisco J Silva. Finite mean field games: fictitious play and convergence to a first order continuous mean field game. *Journal de Mathématiques Pures et Appliquées*, 132: 369–397, 2019.

J Michael Harrison and David M Kreps. Speculative investor behavior in a stock market with heterogeneous expectations. *The Quarterly Journal of Economics*, 92(2):323–336, 1978.

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.

William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

H. W. Kuhn. Extensive games and the problem of information. In *Contributions to the Theory of Games*, volume 2, pp. 193–216. Princeton University Press, 1953.

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

Mathieu Lauriere, Sarah Perrin, Sertan Girgin, Paul Muller, Ayush Jain, Theophile Cabannes, Georgios Piliouras, Julien Pérolat, Romuald Elie, Olivier Pietquin, et al. Scalable deep reinforcement learning algorithms for mean field games. In *International conference on machine learning*, pp. 12078–12095. PMLR, 2022.

Mathieu Laurière, Sarah Perrin, Julien Pérolat, Sertan Girgin, Paul Muller, Romuald Élie, Matthieu Geist, and Olivier Pietquin. Learning in mean field games: A survey. *arXiv preprint arXiv:2205.12944*, 2022.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Paul Muller, Romuald Elie, Mark Rowland, Mathieu Lauriere, Julien Perolat, Sarah Perrin, Matthieu Geist, Georgios Piliouras, Olivier Pietquin, and Karl Tuyls. Learning correlated equilibria in mean-field games. *arXiv preprint arXiv:2208.10138*, 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. In *Proc. NeurIPS*, volume 33, pp. 13199–13213, 2020.

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.

David Schmeidler. Equilibrium points of non-atomic games. *Journal of Statistical Physics*, 7(4): 295–300, 1973.

Wilko Schwarting, Alyssa Pierson, Javier Alonso-Mora, Sertac Karaman, and Daniela Rus. Social behavior for autonomous vehicles. *Proceedings of the National Academy of Sciences*, 116(50): 24972–24978, 2019.

Jayakumar Subramanian and Aditya Mahajan. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pp. 251–259, 2019.

Hado van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, pp. 2094–2100. AAAI, 2016.

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*, pp. 39722–39754. PMLR, 2023.

A Appendix: Detailed Proofs

A.1 Proof of Mean Field Map Equivalence (Theorem 1)

Proof. We prove the two statements separately.

Part 1: From Mixed Strategy σ to Behavioral Policy π. Let $\sigma \in \mathcal{S}$ be given. We define the equivalent behavioral policy $\bar{\pi} = \zeta(\sigma)$ as follows. Let μ_t^σ be the mean field induced by σ at time t. For any $t \in \mathcal{T}$ and $x \in \mathcal{X}$:

- If $\mu_t^\sigma(x) > 0$, we define

$$\bar{\pi}_t(u \mid x) \coloneqq \frac{\int_{\Pi_{\text{pure}}} \mu_{t|\pi^p}^\sigma(x) \mathbf{1}_{\{\pi_t^p(x)=u\}} \sigma(\mathrm{d}\pi^p)}{\mu_t^\sigma(x)}. \tag{9}$$

- If $\mu_t^\sigma(x) = 0$, we define $\bar{\pi}_t(\cdot \mid x)$ arbitrarily (e.g., the uniform distribution over $\mathcal{U}$), as this state is reached with probability zero.

We now show by induction that $\mu_t^{\bar{\pi}} = \mu_t^\sigma$ for all t.

Base Case ($t = 0$): By definition, $\mu_0^{\bar{\pi}} = \mu_0 = \mu_0^\sigma$.

Inductive Step: Assume $\mu_t^{\bar{\pi}} = \mu_t^\sigma$. We compute the distribution at $t+1$:

$$\begin{aligned}\mu_{t+1}^{\bar{\pi}}(x') &= \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} p(x' \mid x, u, \mu_t^{\bar{\pi}}) \bar{\pi}_t(u \mid x) \mu_t^{\bar{\pi}}(x) \\ &= \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} p(x' \mid x, u, \mu_t^\sigma) \bar{\pi}_t(u \mid x) \mu_t^\sigma(x) \quad \text{(by IH)}.\end{aligned}$$

We split the sum over x into states where $\mu_t^\sigma(x) > 0$ and $\mu_t^\sigma(x) = 0$. The terms where $\mu_t^\sigma(x) = 0$ vanish. For terms where $\mu_t^\sigma(x) > 0$, we substitute the definition of $\bar{\pi}$ to obtain

$$\begin{aligned}\bar{\pi}_t(u \mid x) \mu_t^\sigma(x) &= \frac{\int_{\Pi_{\text{pure}}} \mu_{t|\pi^p}^\sigma(x) \mathbf{1}_{\{\pi_t^p(x)=u\}} \sigma(\mathrm{d}\pi^p)}{\mu_t^\sigma(x)} \cdot \mu_t^\sigma(x) \\ &= \int_{\Pi_{\text{pure}}} \mu_{t|\pi^p}^\sigma(x) \mathbf{1}_{\{\pi_t^p(x)=u\}} \sigma(\mathrm{d}\pi^p).\end{aligned}$$

Substituting this back into the expression for $\mu_{t+1}^{\bar{\pi}}(x')$ gives

$$\mu_{t+1}^{\bar{\pi}}(x') = \sum_{x \in \mathcal{X}} \sum_{u \in \mathcal{U}} p(x' \mid x, u, \mu_t^\sigma) \int_{\Pi_{\text{pure}}} \mu_{t|\pi^p}^\sigma(x) \mathbf{1}_{\{\pi_t^p(x)=u\}} \sigma(\mathrm{d}\pi^p).$$

By Fubini's theorem, we exchange the sum and the integral to obtain

$$\begin{aligned}\mu_{t+1}^{\bar{\pi}}(x') &= \int_{\Pi_{\text{pure}}} \sum_{x \in \mathcal{X}} \left(\sum_{u \in \mathcal{U}} p(x' \mid x, u, \mu_t^\sigma) \mathbf{1}_{\{\pi_t^p(x)=u\}} \right) \mu_{t|\pi^p}^\sigma(x) \sigma(\mathrm{d}\pi^p) \\ &= \int_{\Pi_{\text{pure}}} \sum_{x \in \mathcal{X}} p(x' \mid x, \pi_t^p(x), \mu_t^\sigma) \mu_{t|\pi^p}^\sigma(x) \sigma(\mathrm{d}\pi^p).\end{aligned}$$

The term inside the integral is exactly the definition of the state evolution for a specific pure policy π^p. Thus,

$$\mu_{t+1}^{\bar{\pi}}(x') = \int_{\Pi_{\text{pure}}} \mu_{t+1|\pi^p}^\sigma(x') \sigma(\mathrm{d}\pi^p) = \mu_{t+1}^\sigma(x').$$

This completes the induction for Part 1.

Part 2: From Behavioral Policy π to Mixed Strategy σ. Let $\pi \in \Pi$ be given. We construct $\bar{\sigma} = \zeta^{-1}(\pi)$ as a product measure over the space of pure policies. Recall that a pure policy π^p is a deterministic mapping from $\mathcal{T} \times \mathcal{X}$ to $\mathcal{U}$. We define $\bar{\sigma}$ such that the choice of action at every

state-time pair (t, x) is independent and distributed according to $\pi_t(\cdot \mid x)$. Formally, for any pure policy π^p we have

$$\bar{\sigma}(\pi^p) = \prod_{t\in\mathcal{T}} \prod_{x\in\mathcal{X}} \pi_t(\pi_t^p(x) \mid x). \tag{10}$$

(Note: If the space of pure policies is continuous, this is defined via the product measure on the product space $\mathcal{U}^{\mathcal{T}\times\mathcal{X}}$).

We show $\mu_t^{\bar{\sigma}} = \mu_t^{\pi}$ by induction.

Base Case ($t = 0$): $\mu_0^{\bar{\sigma}} = \mu_0 = \mu_0^{\pi}$.

Inductive Step: Assume $\mu_t^{\bar{\sigma}} = \mu_t^{\pi}$. Then,

$$\begin{aligned}\mu_{t+1}^{\bar{\sigma}}(x') &= \int_{\Pi_{\text{pure}}} \mu_{t+1|\pi^p}^{\bar{\sigma}}(x')\bar{\sigma}(\mathrm{d}\pi^p) \\ &= \int_{\Pi_{\text{pure}}} \sum_{x\in\mathcal{X}} p(x' \mid x, \pi_t^p(x), \mu_t^{\bar{\sigma}})\mu_{t|\pi^p}^{\bar{\sigma}}(x)\bar{\sigma}(\mathrm{d}\pi^p).\end{aligned}$$

By the induction hypothesis, $\mu_t^{\bar{\sigma}} = \mu_t^{\pi}$. Furthermore, under the product measure construction of $\bar{\sigma}$, the action chosen at time t, $\pi_t^p(x)$, is independent of the history of the path taken to reach state x. Specifically, the marginal probability of choosing action u at state x under $\bar{\sigma}$ is exactly $\pi_t(u \mid x)$.

Therefore, we can rewrite the integral by grouping pure policies based on the action u they prescribe at state x as

$$\mu_{t+1}^{\bar{\sigma}}(x') = \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} p(x' \mid x, u, \mu_t^{\pi}) \underbrace{\left(\int_{\Pi_{\text{pure}}} \mathbf{1}_{\{\pi_t^p(x)=u\}} \mu_{t|\pi^p}^{\bar{\sigma}}(x)\bar{\sigma}(\mathrm{d}\pi^p) \right)}_{\text{Joint prob. of being at } x \text{ and playing } u}.$$

Because $\bar{\sigma}$ is a product measure, the event of playing u at x is independent of the probability of reaching x. Thus, the term in the parenthesis factorizes as

$$\begin{aligned}\int_{\Pi_{\text{pure}}} \mathbf{1}_{\{\pi_t^p(x)=u\}} \mu_{t|\pi^p}^{\bar{\sigma}}(x)\bar{\sigma}(\mathrm{d}\pi^p) &= \mathbb{P}_{\bar{\sigma}}(\pi_t^p(x) = u) \cdot \int_{\Pi_{\text{pure}}} \mu_{t|\pi^p}^{\bar{\sigma}}(x)\bar{\sigma}(\mathrm{d}\pi^p) \\ &= \pi_t(u \mid x) \cdot \mu_t^{\bar{\sigma}}(x) \\ &= \pi_t(u \mid x) \cdot \mu_t^{\pi}(x).\end{aligned}$$

Substituting this back gives

$$\begin{aligned}\mu_{t+1}^{\bar{\sigma}}(x') &= \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} p(x' \mid x, u, \mu_t^{\pi})\pi_t(u \mid x)\mu_t^{\pi}(x) \\ &= \mu_{t+1}^{\pi}(x').\end{aligned}$$

This completes the proof. □

A.2 Proof of Conditional MF Equivalence (for Theorem 2)

To prove the equivalence of equilibria (Corollary 3), we must show that the best-response problems are identical. This requires showing that for any equivalent pair (π, σ) generating the same mean field $\mu = \Lambda(\pi) = \bar{\Lambda}(\sigma)$, the set of achievable outcomes for a deviating agent is the same in both systems.

Proof. Let $\mathcal{D}_\pi = \{(\mu_{t|\hat{\pi}}^{\pi})_{t\in\mathcal{T}} \mid \hat{\pi} \in \Pi\}$ be the set of achievable state distribution flows for a deviant using any behavioral strategy $\hat{\pi}$ in the system with population policy π. Let $\mathcal{D}_\sigma = \{(\mu_{t|\hat{\sigma}}^{\sigma})_{t\in\mathcal{T}} \mid \hat{\sigma} \in \mathcal{S}\}$ be the set of achievable state distribution flows for a deviant using any mixed strategy $\hat{\sigma}$ in the system with population policy σ. We need to show $\mathcal{D}_\pi = \mathcal{D}_\sigma$.

1. ($\mathcal{D}_\sigma \subseteq \mathcal{D}_\pi$)**:** Take any deviant mixed strategy $\hat{\sigma} \in \mathcal{S}$. Let $\hat{\pi} = \zeta(\hat{\sigma})$ be its behavioral representation. The evolution of the deviant's state distribution $\mu_{t|\hat{\sigma}}^{\sigma}$ follows the same logic as in App. A.1, but with the population mean field fixed to μ. The same steps show that the state flow generated by the deviant

playing $\hat{\sigma}$ against population σ is identical to the state flow generated by a deviant playing $\hat{\pi}$ against population π. Thus, any outcome achievable with a mixed strategy $\hat{\sigma}$ is also achievable with its behavioral representation $\hat{\pi}$, so $\mathcal{D}_\sigma \subseteq \mathcal{D}_\pi$.

2. ($\mathcal{D}_\pi \subseteq \mathcal{D}_\sigma$)**:** Take any deviant behavioral strategy $\hat{\pi} \in \Pi$. By our earlier constructions, there exists an equivalent mixed strategy $\hat{\sigma} \in \zeta^{-1}(\hat{\pi})$. A similar argument shows that the state flow generated by playing $\hat{\pi}$ is identical to that generated by playing $\hat{\sigma}$. Thus, any outcome achievable with a behavioral strategy is also achievable with an equivalent mixed strategy, so $\mathcal{D}_\pi \subseteq \mathcal{D}_\sigma$. □

Since the sets of achievable outcomes are identical, and the reward function is the same, the value of any deviation is the same in both systems. Therefore, $\sup_{\hat{\pi}\in\Pi} J(\hat{\pi}, \mu) = \sup_{\hat{\sigma}\in\mathcal{S}} \mathcal{J}(\hat{\sigma}, \mu)$. This implies that π is a best response to μ if and only if its equivalent σ is a best response to μ, which proves the equivalence of equilibria.

A.3 Proof of Equilibrium Equivalence (for Corollary 2 and 3)

First, we observe the equivalence of equilibria in Corollary 2.

Proof. Corollary 2 follows directly from the equivalence of MFs (Theorem 1) and conditional MFs (Theorem 2), which imply that the best-response problems are identical in both formulations. □

As a result, together with equivalence of (conditional) MFs in Theorems 1 and 2, we obtain the equivalence of equilibria in values and mean field flows.

Proof. Corollary 3 follows from preceding corollaries, which establish that the MF generated by a policy/strategy and therefore values of deviations match across the two equivalent formulations. □

A.4 Proof of Sufficiency of Pure Strategies (Theorem 3)

Proof. Let σ^* be a Strategy Mean Field Nash Equilibrium (SMFNE). Since the spaces $\mathcal{X}$,$\mathcal{U}$ and horizon T are finite, the set of pure policies Π_{pure} is finite.

We construct the tuple $\underline{\pi}^N$ deterministically using the method of proportional assignment (inverse transform sampling on the sorted support).

Step 1: Deterministic Construction. Let $\Pi_{\text{pure}} = \{\pi_1^p, \ldots, \pi_K^p\}$ be the enumeration of the finite set of pure policies, where $K = |\Pi_{\text{pure}}|$. Let $w_k = \sigma^*(\pi_k^p)$ be the probability mass assigned to policy π_k^p by the equilibrium strategy.

We construct the tuple $\underline{\pi}^N$ deterministically using the method of proportional assignment. Define the cumulative weights $C_k = \sum_{j=1}^{k} w_j$, with $C_0 = 0$ and $C_K = 1$. For each agent $i \in \{1, \ldots, N\}$, we assign the pure policy $\pi^i = \pi_{k^*}^p$ where k^* is the unique index satisfying

$$C_{k^*-1} < \frac{i}{N} \leq C_{k^*}.$$

Let $\sigma^N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\pi^i}$ be the empirical distribution of policies. For any specific pure policy π_k^p, the number of agents assigned to it is N_k, satisfying $|N_k - Nw_k| < 1$. Thus, the error in the probability mass is bounded as

$$|\sigma^N(\pi_k^p) - \sigma^*(\pi_k^p)| < \frac{1}{N}.$$

Consequently, the total variation distance between the constructed empirical distribution and the equilibrium strategy converges deterministically as

$$\|\sigma^N - \sigma^*\|_{\text{TV}} = \frac{1}{2}\sum_{k=1}^{K} |\sigma^N(\pi_k^p) - \sigma^*(\pi_k^p)| < \frac{K}{2N}.$$

Step 2: Convergence of the Mean Field. Let $\mu^* = \bar{\Lambda}(\sigma^*)$ be the equilibrium mean field flow. Let μ^N be the empirical mean field flow generated by the N agents playing $\underline{\pi}^N$. We analyze the evolution of the expected error $e_t := \mathbb{E}[\|\mu_t^N - \mu_t^*\|_{\text{TV}}]$.

Define the population transition operator $\Phi_t : \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\Pi_{\text{pure}}) \to \mathcal{P}(\mathcal{X})$ as

$$\Phi_t(\mu, \sigma)(x') \coloneqq \sum_{x \in \mathcal{X}} \int_{\Pi_{\text{pure}}} p(x' \mid x, \pi_t^p(x), \mu) \sigma(\mathrm{d}\pi^p) \mu(x).$$

We decompose the error at time $t+1$ using the triangle inequality to obtain

$$\begin{aligned} \|\mu_{t+1}^N - \mu_{t+1}^*\|_{\text{TV}} \leq & \underbrace{\|\mu_{t+1}^N - \Phi_t(\mu_t^N, \sigma^N)\|_{\text{TV}}}_{T_1} \\ & + \underbrace{\|\Phi_t(\mu_t^N, \sigma^N) - \Phi_t(\mu_t^N, \sigma^*)\|_{\text{TV}}}_{T_2} \\ & + \underbrace{\|\Phi_t(\mu_t^N, \sigma^*) - \Phi_t(\mu_t^*, \sigma^*)\|_{\text{TV}}}_{T_3}. \end{aligned}$$

Analysis of Term T_1: T_1 represents the distance between the realized empirical distribution and the expected distribution. Conditional on μ_t^N, the next state x_{t+1}^i for agent i is independent. By the property of the expected L_1 error for discrete distributions over finite space $\mathcal{X}$, we have

$$\mathbb{E}[T_1 \mid \mu_t^N] \leq \sqrt{\frac{|\mathcal{X}|}{N}} \implies \mathbb{E}[T_1] \leq \frac{C_1}{\sqrt{N}}.$$

Analysis of Term T_2: This term measures the error from using σ^N instead of σ^*,

$$T_2 = \sum_{x'} \left| \sum_x \mu_t^N(x) \int_{\Pi_{\text{pure}}} p(x' \mid x, \pi_t^p(x), \mu_t^N)(\sigma^N - \sigma^*)(\mathrm{d}\pi^p) \right|.$$

Since Φ_t is linear in σ, we bound this using the TV distance established in Step 1, giving

$$T_2 \leq \sup_x \|p(\cdot \mid x, \cdot, \cdot)\|_\infty \cdot \|\sigma^N - \sigma^*\|_{\text{TV}} \leq \frac{K}{2N}.$$

Analysis of Term T_3: Assuming the transition kernel p is Lipschitz continuous in μ with constant L_p, we have

$$\begin{aligned} T_3 &= \|\Phi_t(\mu_t^N, \sigma^*) - \Phi_t(\mu_t^*, \sigma^*)\|_{\text{TV}} \\ &\leq (L_p + 1)\|\mu_t^N - \mu_t^*\|_{\text{TV}}. \end{aligned}$$

Combining the bounds gives

$$e_{t+1} \leq \frac{C_1}{\sqrt{N}} + \frac{K}{2N} + (L_p + 1)e_t.$$

Iterating this recurrence yields $\sup_{t \leq T} \mathbb{E}[\|\mu_t^N - \mu_t^*\|_{\text{TV}}] \leq O(1/\sqrt{N})$.

Step 3: Exploitability. Since σ^* is an SMFNE, every policy in the support of σ^* (i.e., where $w_k > 0$) is a best response to μ^*. Our construction only assigns agents to policies where $w_k > 0$ (for sufficiently large N).

Let π^i be the policy assigned to agent i. Since $\pi^i \in \text{supp}(\sigma^*)$, we have

$$J(\pi^i, \mu^*) = \max_{\hat{\pi}} J(\hat{\pi}, \mu^*).$$

The exploitability in the N-agent game is

$$\mathcal{E}^i = \max_{\hat{\pi}} J^{i,N}(\hat{\pi}, \pi^{-i}) - J^{i,N}(\pi^i, \pi^{-i}).$$

Using the Lipschitz continuity of the reward function r with respect to the mean field (constant L_r), and the convergence of μ^N to μ^*, we have

$$\mathbb{E}|J^{i,N}(\pi, \pi^{-i}) - J(\pi, \mu^*)| \leq L_r \mathbb{E}\left[\sum_t \|\mu_t^N - \mu_t^*\|\right] \leq O(1/\sqrt{N}).$$

Substituting this into the exploitability expression gives

$$\begin{aligned}\mathbb{E}[\mathcal{E}^i] &\leq \left(\max_{\hat{\pi}} J(\hat{\pi}, \mu^*) + O(N^{-1/2})\right) - \left(J(\pi^i, \mu^*) - O(N^{-1/2})\right) \\ &= \underbrace{\left(\max_{\hat{\pi}} J(\hat{\pi}, \mu^*) - J(\pi^i, \mu^*)\right)}_{=0 \text{ by equilibrium property}} + O(N^{-1/2}).\end{aligned}$$

Thus, the constructed tuple is an $\epsilon(N)$-Nash equilibrium with $\epsilon(N) \to 0$. □

A.5 Proof of Proposition 2

The convergence of Fictitious Play in MFGs is typically established for games satisfying a monotonicity condition, as introduced by Lasry and Lions. We analyze the continuous-time version of Mixed Fictitious Play (MFP). Let $\bar{\sigma}^\tau$ be the average mixed strategy and $\bar{\mu}^\tau$ be the average mean field up to time τ. The core of the proof relies on showing that a specific Lyapunov function is non-increasing along the trajectories of the FP dynamics.

Let the game operator be $F(\sigma) = \text{BR}(\mu(\sigma))$, where $\mu(\sigma) = \bar{\Lambda}(\sigma)$ and $\text{BR}(\mu)$ is the set of best-response (mixed) strategies to mean field μ. The continuous-time FP dynamics are given by the differential inclusion $\dot{\sigma}(\tau) \in F(\sigma(\tau)) - \sigma(\tau)$.

The proof follows the structure in (Perrin et al., 2020). Consider the Lyapunov function $V(\sigma, \mu) = \sup_{\hat{\sigma}} \mathcal{J}(\hat{\sigma}, \mu) - \mathcal{J}(\sigma, \mu)$. The time derivative of V along the trajectory of the average strategy $\bar{\sigma}^\tau$ can be decomposed into two parts: a direct effect and a mean-field effect,

$$\frac{d}{d\tau} V(\bar{\sigma}^\tau, \bar{\mu}^\tau) = \langle \nabla_\sigma V, \dot{\bar{\sigma}}^\tau \rangle + \langle \nabla_\mu V, \dot{\bar{\mu}}^\tau \rangle.$$

The direct effect can be shown to be $\langle \nabla_\sigma V, \dot{\bar{\sigma}}^\tau \rangle = -\frac{1}{\tau} V(\bar{\sigma}^\tau, \bar{\mu}^\tau)$, which is negative. The mean-field effect term, $\langle \nabla_\mu V, \dot{\bar{\mu}}^\tau \rangle$, involves the change in value due to the shifting mean field. The Lasry-Lions monotonicity assumption is precisely the condition required to ensure this term is non-positive. It states that the change in rewards from the mean field shifting is negatively correlated with the change in strategies that caused it.

Combining the terms, we get $\frac{d}{d\tau} V(\bar{\sigma}^\tau, \bar{\mu}^\tau) \leq -\frac{1}{\tau} V(\bar{\sigma}^\tau, \bar{\mu}^\tau)$, which implies that V is decreasing. A more rigorous analysis shows that this is sufficient to prove that the exploitability of the average strategy, which is related to V, converges to zero as $\tau \to \infty$.

Proof. Consider the FP process in continuous time, i.e. with continuous iterations $\tau \in \mathbb{R}_{\geq 0}$ and

$$\begin{aligned}\bar{\sigma}^\tau &:= \frac{1}{\tau} \int_0^\tau \sigma^{\text{BR},\tau'} \mathrm{d}\tau' \\ \bar{\mu}^\tau &:= \frac{1}{\tau} \int_0^\tau \mu^{\sigma^{\text{BR},\tau'}} \mathrm{d}\tau'\end{aligned}$$

for all $t \in \mathcal{T}$, where $\sigma^{\text{BR},\tau'}$ is chosen in an arbitrary way s.t. $\sigma^{\text{BR},\tau'} \in \arg\max_{\hat{\sigma}} \mathcal{J}(\hat{\sigma}, \bar{\mu}^\tau)$.

Here, we remind the reader that $\mu^{\sigma^{\text{BR},\tau'}}$ is defined as

$$\begin{aligned}\mu_t^{\sigma^{\text{BR},\tau'}}(x) = \int_{\Pi_{\text{pure}}} \int_{\mathcal{X}^t} & p(x \mid x_{t-1}, \pi_{t-1}^p(x_{t-1}), \mu_{t-1}^{\sigma^{\text{BR},\tau'}}) \\ & \cdot \prod_{s=0}^{t-2} p(\mathrm{d}x_{s+1} \mid x_s, \pi_s^p(x_s), \mu_s^{\sigma^{\text{BR},\tau'}}) \mu_0(\mathrm{d}x_0) \sigma^{\text{BR},\tau'}(\mathrm{d}\pi^p)\end{aligned}$$

Properties. The definition trivially gives us the properties

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau} \bar{\mu}^\tau &= \frac{1}{\tau} \left[\mu^{\sigma^{\text{BR},\tau}} - \mu^{\bar{\sigma}^\tau}\right], \\ \frac{\mathrm{d}}{\mathrm{d}\tau} \bar{\sigma}^\tau &= \frac{1}{\tau} \left[\sigma^{\text{BR},\tau} - \bar{\sigma}^\tau\right].\end{aligned}$$

Assumption 1. *The transition kernel p does not depend on the MF.*

Then, under Assm. 1 we also have

$$\bar{\mu}^\tau = \mu^{\bar{\sigma}^\tau}$$

since at all times $t \in \mathcal{T}$ and iterations $\tau' \in \mathbb{R}_{\geq 0}$,

$$\mu_t^{\sigma^{\mathrm{BR},\tau'}}(x) = \int_{\Pi_{\mathrm{pure}}} \int_{\mathcal{X}^t} p(x \mid x_{t-1}, \pi_{t-1}^p(x_{t-1})) \prod_{s=0}^{t-2} p(\mathrm{d}x_{s+1} \mid x_s, \pi_s^p(x_s)) \mu_0(\mathrm{d}x_0) \sigma^{\mathrm{BR},\tau'}(\mathrm{d}\pi^p)$$

and therefore, integrating over all τ', we have

$$\begin{aligned}
&\frac{1}{\tau} \int_0^\tau \mu_t^{\sigma^{\mathrm{BR},\tau'}}(x) \mathrm{d}\tau' \\
&\quad = \int_{\Pi_{\mathrm{pure}}} \int_{\mathcal{X}^t} p(x \mid x_{t-1}, \pi_{t-1}^p(x_{t-1})) \prod_{s=0}^{t-2} p(\mathrm{d}x_{s+1} \mid x_s, \pi_s^p(x_s)) \mu_0(\mathrm{d}x_0) \frac{1}{\tau} \int_0^\tau \sigma^{\mathrm{BR},\tau'}(\mathrm{d}\pi^p) \mathrm{d}\tau' \\
&\quad = \int_{\Pi_{\mathrm{pure}}} \int_{\mathcal{X}^t} p(x \mid x_{t-1}, \pi_{t-1}^p(x_{t-1})) \prod_{s=0}^{t-2} p(\mathrm{d}x_{s+1} \mid x_s, \pi_s^p(x_s)) \mu_0(\mathrm{d}x_0) \bar{\sigma}^\tau(\mathrm{d}\pi^p)
\end{aligned}$$

where by definition, the left side is $\bar{\mu}^\tau$ and the right side is $\bar{\mu}^\tau$, showing $\bar{\mu}^\tau = \mu^{\bar{\sigma}^\tau}$.

Similarly, for the conditional probabilites we obtain under Assm. 1 that

$$\mu_{t|\hat{\sigma}}^\sigma = \mu_t^{\hat{\sigma}}, \quad \mu_{t|\pi^p}^\sigma = \mu_t^{\pi^p} =: \mu_{t|\pi^p}$$

since the conditional MFs do not depend on the other agent's MF anymore.

Convergence of FP. Therefore, for the exploitability of $\bar{\sigma}^\tau$, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\tau}\mathcal{E}^\tau &= \frac{\mathrm{d}}{\mathrm{d}\tau}\left(\mathcal{J}(\sigma^{\mathrm{BR},\tau}, \bar{\sigma}^\tau) - \mathcal{J}(\bar{\sigma}^\tau, \bar{\sigma}^\tau)\right) \\
&= \frac{\mathrm{d}}{\mathrm{d}\tau} \sum_{t \in \mathcal{T}} \left\langle \mu_{t|\sigma^{\mathrm{BR},\tau}}^{\bar{\sigma}^\tau} - \mu_{t|\bar{\sigma}^\tau}^{\bar{\sigma}^\tau}, r_t^{\bar{\sigma}^\tau} \right\rangle \\
&= \frac{\mathrm{d}}{\mathrm{d}\tau} \sum_{t \in \mathcal{T}} \left\langle \mu_t^{\sigma^{\mathrm{BR},\tau}} - \mu_t^{\bar{\sigma}^\tau}, r_t^{\bar{\sigma}^\tau} \right\rangle \\
&= \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}, \pi^p \in \Pi_{\mathrm{pure}}} \frac{\mathrm{d}}{\mathrm{d}\tau} \left(\sigma^{\mathrm{BR},\tau}(\pi^p) \mu_{t|\pi^p}(x) r(x, \mu_t^{\bar{\sigma}^\tau}) - \bar{\sigma}^\tau(\pi^p) \mu_{t|\pi^p}(x) r(x, \mu_t^{\bar{\sigma}^\tau}) \right) \\
&= \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}, \pi^p \in \Pi_{\mathrm{pure}}} \sigma^{\mathrm{BR},\tau}(\pi^p) \mu_{t|\pi^p}(x) \frac{\mathrm{d}}{\mathrm{d}\tau} r(x, \mu_t^{\bar{\sigma}^\tau}) \\
&\quad - \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}, \pi^p \in \Pi_{\mathrm{pure}}} \frac{\mathrm{d}}{\mathrm{d}\tau} \bar{\sigma}^\tau(\pi^p) \mu_{t|\pi^p}(x) r(x, \mu_t^{\bar{\sigma}^\tau}) \\
&\quad - \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}, \pi^p \in \Pi_{\mathrm{pure}}} \bar{\sigma}^\tau(\pi^p) \mu_{t|\pi^p}(x) \frac{\mathrm{d}}{\mathrm{d}\tau} r(x, \mu_t^{\bar{\sigma}^\tau}) \\
&= \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}, \pi^p \in \Pi_{\mathrm{pure}}} \left(\sigma^{\mathrm{BR},\tau}(\pi^p) - \bar{\sigma}^\tau(\pi^p) \right) \mu_{t|\pi^p}(x) \left\langle \nabla_\mu r(x, \mu_t^{\bar{\sigma}^\tau}), \frac{\mathrm{d}}{\mathrm{d}\tau} \mu_t^{\bar{\sigma}^\tau} \right\rangle \\
&\quad - \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}, \pi^p \in \Pi_{\mathrm{pure}}} \frac{\mathrm{d}}{\mathrm{d}\tau} \bar{\sigma}^\tau(\pi^p) \mu_{t|\pi^p}(x) r(x, \mu_t^{\bar{\sigma}^\tau}) \\
&= \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}, \pi^p \in \Pi_{\mathrm{pure}}} \left(\sigma^{\mathrm{BR},\tau}(\pi^p) - \bar{\sigma}^\tau(\pi^p) \right) \mu_{t|\pi^p}(x) \left\langle \nabla_\mu r(x, \mu_t^{\bar{\sigma}^\tau}), \frac{\mathrm{d}}{\mathrm{d}\tau} \mu_t^{\bar{\sigma}^\tau} \right\rangle \\
&\quad - \frac{1}{\tau} \sum_{t \in \mathcal{T}} \sum_{x \in \mathcal{X}, \pi^p \in \Pi_{\mathrm{pure}}} \sigma^{\mathrm{BR},\tau}(\pi^p) \mu_{t|\pi^p}(x) r(x, \mu_t^{\bar{\sigma}^\tau}) - \bar{\sigma}^\tau(\pi^p) \mu_{t|\pi^p}(x) r(x, \mu_t^{\bar{\sigma}^\tau})
\end{aligned}$$

$$= \sum_{t\in\mathcal{T}} \sum_{x\in\mathcal{X}, \pi^p\in\Pi_{\text{pure}}} \mu_{t|\pi^p}(x) \left\langle \nabla_\mu r(x, \mu_t^{\bar\sigma^\tau}), \frac{\mathrm{d}}{\mathrm{d}\tau}\mu_t^{\bar\sigma^\tau} \right\rangle \tau \frac{\mathrm{d}}{\mathrm{d}\tau}\bar\sigma^\tau(\pi^p) - \frac{1}{\tau}\mathcal{E}^\tau$$

$$= \sum_{t\in\mathcal{T}} \sum_{x\in\mathcal{X}} \left\langle \nabla_\mu r(x, \mu_t^{\bar\sigma^\tau}), \frac{\mathrm{d}}{\mathrm{d}\tau}\mu_t^{\bar\sigma^\tau} \right\rangle \tau \frac{\mathrm{d}}{\mathrm{d}\tau} \underbrace{\sum_{\pi^p\in\Pi_{\text{pure}}} \bar\sigma^\tau(\pi^p)\mu_{t|\pi^p}(x)}_{\mu_t^{\bar\sigma^\tau}(x)} - \frac{1}{\tau}\mathcal{E}^\tau$$

by dropping terms via envelope theorem as in Perrin et al. (2020); Cui et al. (2024) and applying the properties established above. The remaining term is then always non-positive, if we have standard "monotonicity" and w.l.o.g. remove dependence on actions, i.e. if we have Lasry-Lions monotonicity

$$\sum_{x\in\mathcal{X}} (\mu(x) - \mu'(x))(r(x,\mu) - r(x,\mu')) \le 0, \quad \forall \mu, \mu' \in \mathcal{P}(\mathcal{X}).$$

This completes the proof. □

A.6 DFP as the Finite-Agent Realization of MFP

This subsection records, in the notation of Section 3, that one step of Deterministic Fictitious Play (Algorithm 2) has exactly the Mixed Fictitious Play update (Algorithm 1) as its conditional mean, the difference being conditionally centred noise of magnitude $\mathcal{O}(N^{-1/2})$.

We identify $\mathcal{S} = \mathcal{P}(\Pi_{\text{pure}})$ with the simplex in $\mathbb{R}^K$, $K = |\Pi_{\text{pure}}|$, and use the Euclidean norm $\|\cdot\|$ on $\mathbb{R}^K$, in which $\|\delta_{\pi^p} - \sigma\| \le \sqrt{2}$ for every pure policy π^p and every $\sigma \in \mathcal{S}$.

Let $\sigma_k^N = \frac{1}{N}\sum_{i=1}^N \delta_{\pi^{i,k}}$ be the empirical distribution of pure policies at iteration k and $\mu_k := \bar\Lambda(\sigma_k^N)$ the mean field it induces. Each agent switches to a best response $\pi^{\mathrm{BR}}(\mu_k) \in \arg\max_{\pi^p\in\Pi_{\text{pure}}} J(\pi^p, \mu_k)$ with probability α_k, independently of the other agents, so that

$$\sigma_{k+1}^N = (1-\alpha_k)\,\sigma_k^N + \alpha_k\,\delta_{\pi^{\mathrm{BR}}(\mu_k)} + M_{k+1}, \qquad M_{k+1} := \frac{1}{N}\sum_{i=1}^N \left(Z_i^k - \alpha_k\right)\left(\delta_{\pi^{\mathrm{BR}}(\mu_k)} - \delta_{\pi^{i,k}}\right),$$

where $Z_i^k \sim \mathrm{Bernoulli}(\alpha_k)$ are i.i.d. and independent of the history $\mathcal{F}_k$ up to iteration k. Taking conditional expectations,

$$\mathbb{E}\left[\sigma_{k+1}^N \mid \mathcal{F}_k\right] = (1-\alpha_k)\,\sigma_k^N + \alpha_k\,\delta_{\pi^{\mathrm{BR}}(\mu_k)},$$

which is precisely the update of Algorithm 1. For the noise term, the summands are conditionally independent with mean zero and norm at most $\sqrt{2}$, whence

$$\mathbb{E}\left[\|M_{k+1}\|^2 \mid \mathcal{F}_k\right] = \frac{1}{N^2}\sum_{i=1}^N \alpha_k(1-\alpha_k)\left\|\delta_{\pi^{\mathrm{BR}}(\mu_k)} - \delta_{\pi^{i,k}}\right\|^2 \le \frac{2\alpha_k}{N}.$$

Thus the noise is conditionally centred with standard deviation at most $\sqrt{2\alpha_k/N}$, i.e. $\mathcal{O}(N^{-1/2})$ uniformly in k whenever $\alpha_k \le 1$: the mean of the population's policy distribution follows the MFP recursion exactly, and DFP differs from MFP only by this finite-agent fluctuation. In the experiments of Section 5 the fluctuation is small and DFP indeed reaches the same equilibria as MFP and as behavioural Fictitious Play.

A.7 Proof of High-Probability Sampled Nash Equilibrium (Theorem 4)

Proof of Theorem 4. Let $\mu^* = \bar\Lambda(\sigma^*)$ be the deterministic mean field flow associated with the equilibrium. Let $\sigma^N = \frac{1}{N}\sum_{i=1}^N \delta_{\pi^i}$ be the random empirical distribution over pure policies determined by the sampling at $t = 0$.

We aim to bound the exploitability $\mathcal{E}^i$ of a generic agent i in the finite game. The exploitability is defined as:

$$\mathcal{E}^i = \sup_{\hat\pi\in\Pi} J^{i,N}(\hat\pi, \underline{\pi}^{-i}) - J^{i,N}(\pi^i, \underline{\pi}^{-i}).$$

Step 1: Connecting Finite Game and Mean Field Objectives. First, we relate the finite N-agent objective $J^{i,N}$ to the limiting MFG objective J. The functional $\mu \mapsto J(\pi, \mu)$ represents the value of policy π against a flow of measures μ. Under the Lipschitz assumptions on r and p, this functional is Lipschitz continuous with respect to the total variation distance of the flows (capturing both the direct effect on rewards and the divergence of state trajectories). Consequently, there exists a constant L_J (depending on L_r, L_p, T) such that for any agent i:

$$\left|J^{i,N}(\pi, \underline{\pi}^{-i}) - J(\pi, \mu^*)\right| \leq L_J \, \mathbb{E}\left[\sum_{t=0}^{T-1} \|\mu_t^N - \mu_t^*\|_{\text{TV}}\right] + O(1/N). \tag{11}$$

The $O(1/N)$ term accounts for the self-influence of agent i on the empirical measure μ^N, which vanishes in the limit. The expectation $\mathbb{E}$ is taken over the stochastic dynamics of the finite population. Let $\Delta(\underline{\pi}) \coloneqq \sum_t \mathbb{E}[\|\mu_t^N - \mu_t^*\|]$ be the expected total variation distance (from now on, we omit the subscript TV for simplicity).

We can now bound the exploitability $\mathcal{E}^i$:

$$\mathcal{E}^i \leq \left|J^{i,N}(\pi^i, \underline{\pi}^{-i}) - J(\pi^i, \mu^*)\right| + \sup_{\hat{\pi}} \left|J^{i,N}(\hat{\pi}, \underline{\pi}^{-i}) - J(\hat{\pi}, \mu^*)\right| + \underbrace{\left(\sup_{\hat{\pi}} J(\hat{\pi}, \mu^*) - J(\pi^i, \mu^*)\right)}_{\Delta_{\text{opt}}}. \tag{12}$$

By the above discussion, the first two terms are bounded by $2L_J\Delta(\underline{\pi}) + O(1/N)$. Furthermore, since π^i is sampled from the support of the SMFNE σ^*, it is a best response to μ^*, so $\Delta_{\text{opt}} = 0$ almost surely. Thus, $\mathcal{E}^i \leq 2L_J\Delta(\underline{\pi}) + O(1/N)$.

Step 2: Decomposition of Mean Field Error. It remains to bound $\Delta(\underline{\pi})$ with high probability. We decompose the error into "State Noise" and "Policy Noise":

$$\mathbb{E}[\|\mu^N - \mu^*\|] \leq \underbrace{\mathbb{E}[\|\mu^N - \bar{\Lambda}(\sigma^N)\|]}_{A\text{:State Noise}} + \underbrace{\|\bar{\Lambda}(\sigma^N) - \bar{\Lambda}(\sigma^*)\|}_{B\text{:Policy Noise}}. \tag{13}$$

Here, $\bar{\Lambda}(\sigma^N)$ is the *limit* mean field flow that would be generated by an infinite population with strategy distribution σ^N.

Step 3: Concentration Bounds.

Term A (State Noise): Conditional on $\underline{\pi}$, the agents form an interacting particle system coupled through the empirical mean field μ_t^N. Due to the Lipschitz continuity of the transition kernel p with respect to the measure, standard propagation of chaos results (e.g., via Grönwall's inequality) ensure that the empirical measure μ^N converges to the limit flow $\bar{\Lambda}(\sigma^N)$ at a rate of $O(N^{-1/2})$. Thus, for some constant C_1:

$$A \leq \frac{C_1}{\sqrt{N}}.$$

Term B (Policy Noise): This term is deterministic with respect to state transitions but random with respect to the initial sampling of policies $\underline{\pi}$. We use McDiarmid's Inequality. Consider the mapping $\Psi : (\Pi_{\text{pure}})^N \to \mathcal{P}(\mathcal{X})^T$ defined by $\Psi(\pi^1, \ldots, \pi^N) = \bar{\Lambda}(\frac{1}{N}\sum \delta_{\pi^i})$. Changing one agent's policy π^k to $\hat{\pi}^k$ alters the input distribution σ^N by mass $1/N$. Since the mean field dynamics are Lipschitz with respect to the strategy distribution (inherited from L_p), the output flow changes by at most c_2/N. Applying McDiarmid's inequality, for any $t > 0$:

$$\mathbb{P}_{\text{sampling}}\left(\|\bar{\Lambda}(\sigma^N) - \mu^*\| \geq t\right) \leq 2\exp\left(-\frac{2Nt^2}{c_2^2}\right). \tag{14}$$

Setting the RHS to δ, we get $t \leq c_2\sqrt{\frac{\log(2/\delta)}{2N}}$. Thus with high probability, $B \leq C_2\sqrt{\frac{\log(1/\delta)}{N}}$.

Conclusion: Combining the bounds for Term A and Term B, we have with probability $1 - \delta$:

$$\varepsilon \leq \frac{C'}{\sqrt{N}} + C''\sqrt{\frac{\log(1/\delta)}{N}} \leq C\sqrt{\frac{1 + \log(1/\delta)}{N}}, \tag{15}$$

where we used the inequality $\sqrt{a} + \sqrt{b} \leq \sqrt{2(a+b)}$ to group terms under a single square root, absorbing constants into C. □

B Additional Experiments and Propagation of Chaos

A core assumption of Mean Field Game theory is that as the number of agents N in a finite game approaches infinity, the system's macroscopic properties (such as the average reward) converge to those predicted by the mean-field model. This property is often called the "propagation of chaos."

To empirically validate this for our framework, we ran the Deterministic Fictitious Play (DFP) algorithm on the SIS Congestion game with a varying number of agents, ranging from $N = 10$ to $N = 5000$. For each value of N, we computed the final social welfare (the average reward per agent in the population). Fig. 5 plots this social welfare as a function of N.

The result shows that as the number of agents increases, the **average objective** of the finite-agent system converges to a stable value close to that of MFP. This provides strong empirical support that our DFP algorithm, and the underlying environments, are consistent with the foundational principles of existing MFG theory. The convergence also demonstrates that a relatively small number of agents can be sufficient to obtain a good approximation of the mean-field limit for this specific problem.

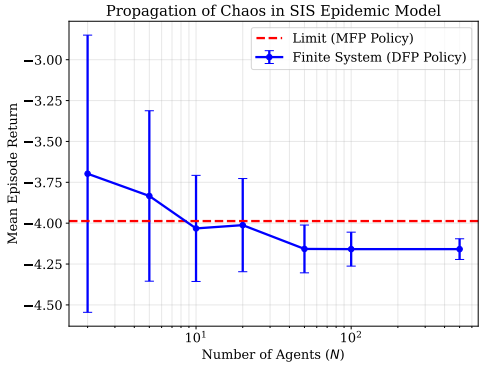

Figure 5: Convergence of social welfare in the SIS game as a function of the population size N. The stabilization of the curve empirically validates the propagation of chaos, showing that finite-agent outcomes approximate the mean-field limit as N grows.

Alg.	$\mathbb{P}$**(SAD)**	$\mathbb{P}$**(SBD)**	$\mathbb{P}$**(SABD)**
FP (baseline)	0.002	0.002	0.997
MFP (ours)	0.005	0.005	0.990
DFP (ours)	0.005	0.006	0.989

Table 2: Comparison of path probabilities in the learned equilibria for the Braess Congestion game (T=2). All algorithms converge to a low-exploitability state, with similar macroscopic traffic patterns.

C Deep Reinforcement Learning

To approximate the best response in environments that are too large to explicitly track the policy space, we rely on a Deep Q-Learning (DQL) approach. DQL algorithm are known to be very effective in high dimensional settings (Mnih et al., 2015). More precisely, we employ a Double-DQN algorithm equipped with a replay buffer, as presented in van Hasselt et al. (2016). For completeness, we provide the pseudo-algorithms (Algorithm 1, Algorithm 2, Algorithm 3) used in our implementations.

Q-network architecture. The action–value function is approximated by a fully-connected neural network $Q_\theta(s, t, a)$ that takes as input the normalized state s and time index t. The input vector is given by

$$\big[\, s/(|\mathcal{S}| - 1),\ t/(T - 1) \,\big],$$

where $|\mathcal{S}|$ denotes the state-space cardinality and T the time horizon. The specifics of the network are given by Table 3

Algorithm 1 Mixed Fictitious Play (MFP).

Input: Initial strategy σ^0, number of iterations K.
Initialize $\mu^0 \leftarrow \bar{\Lambda}(\sigma^0)$.
for $k = 0, 1, \ldots, K-1$ **do**
  Compute pure best response to MF: $\pi^{\mathrm{BR},k+1} \in \arg\max_{\pi^p \in \Pi_{\mathrm{pure}}} J(\pi^p, \mu^k)$.
  Let $\sigma_{\mathrm{BR}}^{k+1} = \delta_{\pi^{\mathrm{BR},k+1}}$.
  Update strategy: $\sigma^{k+1} \leftarrow (1 - \frac{1}{k+2})\sigma^k + \frac{1}{k+2}\sigma_{\mathrm{BR}}^{k+1}$.
  Update MF: $\mu^{k+1} \leftarrow (1 - \frac{1}{k+2})\mu^k + \frac{1}{k+2}\bar{\Lambda}(\sigma_{\mathrm{BR}}^{k+1})$.
end for
return σ^K

Algorithm 2 Deterministic Fictitious Play (DFP).

Input: Initial policies $(\pi^{1,0}, \ldots, \pi^{N,0})$, schedule α_k.
for $k = 0, 1, 2, \ldots$ **do**
  Compute MF $\mu^{N,k}$ from pure policies $(\pi^{i,k})_{i=1}^N$.
  Compute a pure BR $\pi^{\mathrm{BR},k}$ to $\mu^{N,k}$ (DRL if needed).
  for each agent $i \in [N]$ **do**
    With probability α_k, set $\pi^{i,k+1} \leftarrow \pi^{\mathrm{BR},k}$.
    Otherwise, keep $\pi^{i,k+1} \leftarrow \pi^{i,k}$.
  end for
end for

Algorithm 3 Deep Q-Learning Best Response with Replay Buffer (Double-DQN)

Input: Environment $\mathcal{E}$, learning rate η, batch size B, number of iterations T, target update period τ, Greedy-policy scheduler $(\varepsilon_k)_{0 \le k \le T-1}$.
Initialize online Q-network Q_θ.
Initialize target Q-network $Q_{\bar{\theta}} \leftarrow Q_\theta$.
Initialize optimizer and replay buffer $\mathcal{D}$.
for $k = 0, 1, 2, \ldots, T-1$ **do**
  Compute exploration rate ε_k.
  Roll out trajectories using an ε_k-greedy policy induced by Q_θ.
  Store transitions (s, a, r, s', d, t) in replay buffer $\mathcal{D}$.
  Sample minibatch $\{(s_i, a_i, r_i, s_i', d_i, t_i)\}_{i=1}^B \sim \mathcal{D}$.
  Compute greedy actions

$$a_i^\star \leftarrow \arg\max_a Q_\theta(s_i', t_i + 1, a).$$

  Compute Double-DQN targets

$$y_i \leftarrow r_i + (1 - d_i)\, Q_{\bar{\theta}}(s_i', t_i + 1, a_i^\star).$$

  Minimize the Huber loss

$$\frac{1}{B} \sum_{i=1}^{B} \ell_{\mathrm{Huber}}(Q_\theta(s_i, t_i, a_i), y_i)$$

  using one gradient step on θ.
  if $k \bmod \tau = 0$ **then**
    Update target network: $Q_{\bar{\theta}} \leftarrow Q_\theta$.
  end if
end for
Output: Best-response policy induced by Q_θ.

Table 3: Neural network architecture and training hyperparameters for the DQN best response.

Component	**Value**
Hidden layers	5 fully-connected layers
Hidden width	100
Activation function	ReLU
Output layer	Linear
Learning rate	8×10^{-4}
Optimizer	Adam with gradient clipping
Batch size	500
Training iterations	100,000
Target update period	$\tau = 10{,}000$
Loss function	Huber loss
Compute	GPU NVIDIA RTX 6000
VRAM	16GB
DFP Running Time (100 iterations)	~ 5 hours