Major-Minor Mean Field Multi-Agent Reinforcement Learning

Kai Cui [1] **Christian Fabian** [1] **Anam Tahir** [1] **Heinz Koeppl** [1]

Abstract

Multi-agent reinforcement learning (MARL) remains difficult to scale to many agents. Recent MARL using Mean Field Control (MFC) provides a tractable and rigorous approach to otherwise difficult cooperative MARL. However, the strict MFC assumption of many independent, weakly-interacting agents is too inflexible in practice. We generalize MFC to instead simultaneously model many similar and few complex agents – as Major-Minor Mean Field Control (M3FC). Theoretically, we give approximation results for finite agent control, and verify the sufficiency of stationary policies for optimality together with a dynamic programming principle. Algorithmically, we propose Major-Minor Mean Field MARL (M3FMARL) for finite agent systems instead of the limiting system. The algorithm is shown to approximate the policy gradient of the underlying M3FC MDP. Finally, we demonstrate its capabilities experimentally in various scenarios. We observe a strong performance in comparison to state-of-the-art policy gradient MARL methods.

1. Introduction

Recent successes of reinforcement learning (RL) (Vinyals et al., 2019; Schrittwieser et al., 2020; Ouyang et al., 2022) motivate the search for techniques for the multi-agent case, referred to as multi-agent reinforcement learning (MARL). Due to the high complexity of multi-agent control (Bernstein et al., 2002; Daskalakis et al., 2009), exploiting problem structure is important for scalable MARL. In this work, we consider systems with many agents interacting through aggregated information of all agents – the mean field (MF).

[1]Department of Electrical Engineering and Information Technology, Technische Universität Darmstadt, Darmstadt, Germany. Correspondence to: Kai Cui <kai.cui@tu-darmstadt.de>, Heinz Koeppl <heinz.koeppl@tu-darmstadt.de>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Mean field control for MARL. Dynamical control and behavior in systems with many agents is the subject of studies in mean field games (MFG) (Huang et al., 2006; Lasry and Lions, 2007) and mean field control (MFC) (Nourian et al., 2012; Bensoussan et al., 2013; Carmona et al., 2023b). Such aggregated interaction models simplify MARL in the limit of infinite agents, whenever agents interact only through their empirical distribution. The simplification provides a problem complexity that is independent of the exact number of agents. The result is tractability, by avoiding otherwise exponentially large joint state-action spaces (Zhang et al., 2021). This has led to scalable control based on MFC (Gu et al., 2023; Carmona et al., 2023b). And indeed, in applications such aggregation is commonly found on some level, e.g., in chemical reaction networks for aggregate molecule mass (Anderson and Kurtz, 2011), related mass-action epidemics models (Kiss et al., 2017), or traffic where congestion depends on the number of travelling cars (Cabannes et al., 2022), to name just a few. See also epidemics control (Dunyak and Caines, 2021), drone swarms (Shiri et al., 2019), self organization (Carmona et al., 2023a), and many more financial (Carmona, 2020) or engineering scenarios (Djehiche et al., 2017).

Limitations of standard MFC. However, the strict assumption of only *minor* agents – i.e. independent, homogeneous agents that can be summarized by their distribution (MF) – limits applicability. In practice, systems often consist of more than homogeneous agents, and hence one must extend standard MFC towards *major* agents or environment states that are not aggregated. For instance, in modelling car traffic on road networks (Cabannes et al., 2022; Wu et al., 2023), when considering only the distribution of cars (*minor* agents) on the network, one cannot model *major* agents or environment states, such as traffic lights or the road conditions respectively. Another example is given by the logistics scenario in Figure 1 and in the experiments, where many drones on a moving truck collect many packages.

For this purpose, a first step in the continuous-time MFG literature is to consider common noise (Carmona et al., 2016; Perrin et al., 2020), in order to relax the unconditional independence of minor agents. Some more recent works consider such common noise also in discrete-time MFC (Carmona et al., 2023b; Bäuerle, 2023; Motte and

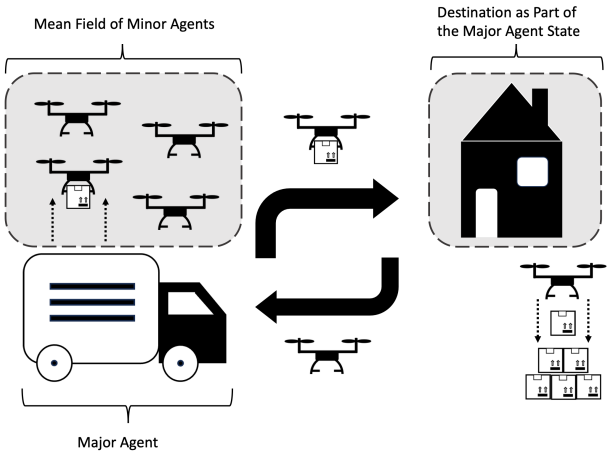

Figure 1. Logistics example: Many drones are modelled as minor agent MF, while truck and package destinations are modelled by a major agent. (See Foraging problem in Section 4.1)

Pham, 2022; 2023), or equivalently, global environment states (Mondal et al., 2023). Essentially, this extension allows MFC to also model random environment effects such as the arrival of new packages in the logistics example (Figure 1). Carmona et al. (2023b) provide a reformulation of MARL into single-agent RL and consider algorithms for the resulting Markov decision process (MDP). Bäuerle (2023) give approximation theorems and approximate optimality in the finite system by the limiting MFC solution with common noise, and Motte and Pham (2022; 2023) quantify the rates of convergence explicitly. See also Table 1 for a brief comparison between existing works. In comparison, for the common noise setting, we contribute a new approximation analysis of MFC-based MARL algorithms, where in contrast to prior work, we learn directly with *finite agents*.

More importantly however, a second contribution is to consider major *agents*. Major agents generalize common noise or environmental states, and take actions that have a non-negligible effect on the system. So far, major agents have only been considered in *continuous*-time, *non*-cooperative MFGs (Nourian and Caines, 2013; Şen and Caines, 2014; Caines and Kizilkale, 2016; Şen and Caines, 2016). To the best of our knowledge, no such *discrete*-time, *cooperative* framework has been formulated yet. In this work, we investigate such a framework and associated MARL algorithms.

Contribution. Existing MFC cannot model general agents and many aggregated agents simultaneously. In essence, we generalize the solution spaces of single-agent RL and MFC-based MARL – frameworks for cooperative MARL as depicted in Figure 2. This provides both tractability for many aggregated agents and generality for arbitrary general agents. Our contribution is briefly summarized into (i) formulating the first discrete-time MFC model with major agents, together with establishing its theoretical properties; (ii) providing a MFC-based MARL algorithm, which in contrast to prior work learns on the finite problem of interest; and (iii) we perform a significant empirical evaluation, also obtaining positive comparisons of MFC-based MARL against state of the art, whereas prior works on MFC were limited to verifying algorithms on one or two examples.

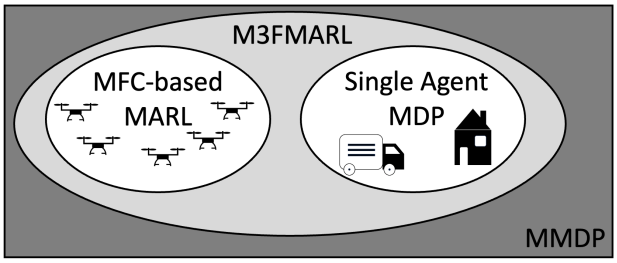

Figure 2. Our M3FC-based MARL generalizes MFC-based MARL and standard single-agent RL in the solution space of general MARL solutions, reducing the otherwise combinatorial nature of MARL (Zhang et al., 2021) to a tractable but still general setting.

Table 1. A comparison of recent related works and a subset of their results on *discrete-time* MFC.
prop. chaos: propagation of chaos; *opt. policy*: existence of optimal (stationary) policies; *common noise*: presence thereof; *non-finite*: non-finite state-actions, e.g. compact; *major agent*: presence thereof; *RL*: RL algorithm ($^+$: learns / is analyzed on finite MARL problems).

Ref.	*prop. chaos*	*opt. policy*	*common noise*	*non-finite*	*major agent*	*RL*
Carmona et al. (2023b)	✗	✓	✓	✓	✗	✓
Gu et al. (2021; 2023)	✓	✓	✗	✗	✗	✓
Bäuerle (2023)	✓	✓	✓	✓	✗	✗
Mondal et al. (2022; 2023)	✓	✗	✓	✗	✗	✓
Motte and Pham (2022; 2023)	✓	✓	✓	✓	✗	✗
our work	✓	✓	✓	✓	✓	✓$^+$

2. Major-Minor Mean Field Control

To begin, in this section we extend standard MFC by modelling the presence of a major agent. The generalization to more than one major agent is straightforward. This leads to our discrete-time major-minor MFC (M3FC) model. Overall, we obtain a formulation that allows standard MARL handling of major agents, while tractably handling many minor agents via MFC-based techniques.

Notation: By $\mathbb{E}_X$ we denote conditional expectations given X. The space of probability measures $\mathcal{P}(\mathcal{X})$ on compact metric spaces $\mathcal{X}$ is equipped with the 1-Wasserstein distance, unless noted otherwise (Villani, 2009). Note compactness of $\mathcal{P}(\mathcal{X})$ on compact $\mathcal{X}$ by Prokhorov's theorem (Billingsley, 2013). Hence, we sometimes use the uniformly (not Lipschitz) equivalent metric $d_\Sigma(\mu, \mu') := \sum_{m=1}^{\infty} 2^{-m} |\int f_m \, \mathrm{d}(\mu - \mu')|$, for some sequence of continuous $f_m \colon \mathcal{X} \to [-1, 1]$ (Parthasarathy, 2005, Theorem 6.6).

2.1. Finite-Agent System

Consider N (minor) agents $i \in [N] := \{1, \dots, N\}$ with compact metric state and action spaces $\mathcal{X}, \mathcal{U}$, equipped with random states and actions $x_t^{i,N}$ and $u_t^{i,N}$ at times $t \in \mathbb{N}$, where initial states $x_0^{i,N} \sim \mu_0(x_0^{i,N})$ are independently sampled from some initial distribution $\mu_0 \in \mathcal{P}(\mathcal{X})$. In addition to standard MFC, we also consider a single major agent, though the framework can be extended to multiple. Consider major agent state and action spaces, $\mathcal{X}^0, \mathcal{U}^0$ and state-actions $x_t^{0,N}$, $u_t^{0,N}$, with the major agent formally indexed by $i = 0$. Given all actions, the agent states evolve according to kernels p, p^0 depending on (i) the agent's own state-actions, (ii) the major state-actions, and (iii) the empirical **MF**, i.e. the $\mathcal{P}(\mathcal{X})$-valued empirical state distribution $\mu_t^N := \frac{1}{N} \sum_{i=1}^N \delta_{x_t^{i,N}}$. This means that minor agents affect other agents only at rate $\frac{1}{N}$. In practice, we identify minor agents as all agents that matter through their MF μ_t^N. Any remaining agents are major, such that the problem-specific stratification into major and minor agents is always possible.

By symmetry, the system state at any time t is therefore entirely given by $(x_t^{0,N}, \mu_t^N)$. Accordingly, in MFC we share policies between all minor agents. We consider time-variant policies $\pi \in \Pi$, $\pi^0 \in \Pi^0$ from some classes of major and minor policies Π, Π^0 that depend on an agent's own state and $(x_t^{0,N}, \mu_t^N)$ at all times t. Overall, for all $i \in [N]$ and $t \in \mathbb{N}$, the finite MFC system follows

$$u_t^{i,N} \sim \pi_t(u_t^{i,N} \mid x_t^{i,N}, x_t^{0,N}, \mu_t^N), \tag{1a}$$

$$u_t^{0,N} \sim \pi_t^0(u_t^{0,N} \mid x_t^{0,N}, \mu_t^N), \tag{1b}$$

$$x_{t+1}^{i,N} \sim p(x_{t+1}^{i,N} \mid x_t^{i,N}, u_t^{i,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N), \tag{1c}$$

$$x_{t+1}^{0,N} \sim p^0(x_{t+1}^{0,N} \mid x_t^{0,N}, u_t^{0,N}, \mu_t^N)\,. \tag{1d}$$

The goal is then to maximize the infinite-horizon discounted objective $J^N(\pi, \pi^0) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^{0,N}, u_t^{0,N}, \mu_t^N)\right]$ over minor and major policies (π, π^0), with discount $\gamma \in (0, 1)$ and reward function $r \colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$. While an optimal behavior could be learned using standard MARL policy gradient methods, for improved tractability we introduce the following M3FC model in the case of many minor agents.

Remark 1. *The model is as expressive as in existing MFC (Mondal et al., 2022; Gu et al., 2023), as it also includes (i) joint state-action MFs $\nu_t \in \mathcal{P}(\mathcal{X} \times \mathcal{U})$, by splitting time steps in two and defining new states in $\mathcal{X} \cup \mathcal{X} \times \mathcal{U}$, (ii) average rewards over all agents, and (iii) random rewards r_t^i by $r(\mu_t^N) \equiv \frac{1}{N} \sum_{i=1}^N \mathbb{E}[r_t^i \mid x_t^{i,N}, \mu_t^N]$. A finite horizon is handled analogously (without optimal stationary policies).*

2.2. Mean Field Control Limit

By introducing the MF limit, we obtain a large, more tractable subclass of cooperative multi-agent control problems, which may otherwise suffer from the curse of many agents (combinatorial joint state-action space (Zhang et al., 2021)). We introduce the MF limit by formally taking $N \to \infty$: The finite-agent control problem is replaced by a higher-dimensional single-agent MDP – the M3FC MDP. By symmetry, we summarize minor agents into their probability law, the MF $\mu_t \equiv \mathcal{L}(x_t^{i,N}) \in \mathcal{P}(\mathcal{X})$. It replaces its empirical analogue μ_t^N by a law of large numbers (LLN). Thus, by definition, the MF μ_t evolves forward as

$$\begin{aligned} \mu_{t+1} &= T(x_t^0, u_t^0, \mu_t, \mu_t \otimes \pi_t(\mu_t)) \\ &= \iint p(\cdot \mid x, u, x_t^0, u_t^0, \mu_t) \pi_t(\mathrm{d}u \mid x, \mu_t) \mu_t(\mathrm{d}x), \end{aligned} \tag{2}$$

with $\pi_t(\mu_t) := \pi_t(\cdot \mid \cdot, \mu_t)$, product measures $h_t \equiv \mu_t \otimes \pi_t(\mu_t)$ of measure μ_t and kernel $\pi_t(\mu_t)$ on $\mathcal{X} \times \mathcal{U}$, and deterministic dynamics for the MF, $T(x^0, u^0, \mu, h) := \iint p(\cdot \mid x, u, x^0, u^0, \mu) h(\mathrm{d}x, \mathrm{d}u)$.

Therefore, the state of the limiting system consists only of the MF μ_t, major state x_t^0 and actions $h_t \equiv \mu_t \otimes \pi_t(\mu_t)$. As a result, we obtain the limiting *M3FC MDP*

$$h_t \sim \hat{\pi}_t(h_t \mid x_t^0, \mu_t), \tag{3a}$$

$$u_t^0 \sim \pi_t^0(u_t^0 \mid x_t^0, \mu_t), \tag{3b}$$

$$\mu_{t+1} = T(x_t^0, u_t^0, \mu_t, h_t), \tag{3c}$$

$$x_{t+1}^0 \sim p^0(x_{t+1}^0 \mid x_t^0, u_t^0, \mu_t) \tag{3d}$$

with objective $J(\hat{\pi}, \pi^0) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^0, u_t^0, \mu_t)\right]$ and transition dynamics for the MF $T(x^0, u^0, \mu, h) := \iint p(\cdot \mid x, u, x^0, u^0, \mu) h(\mathrm{d}x, \mathrm{d}u)$. Here, we identify $\mu_t \otimes \pi_t(\mu_t) \equiv h_t \in \mathcal{H}(\mu_t)$ in the compact set $\mathcal{H}(\mu) \subseteq \mathcal{P}(\mathcal{X} \times \mathcal{U})$ of *desired joint state-action distributions with first marginal μ* as part of the action of the M3FC MDP.

In other words, the action of the M3FC MDP is (h_t, u_t^0) where h_t replaces all the minor agent actions by a LLN.

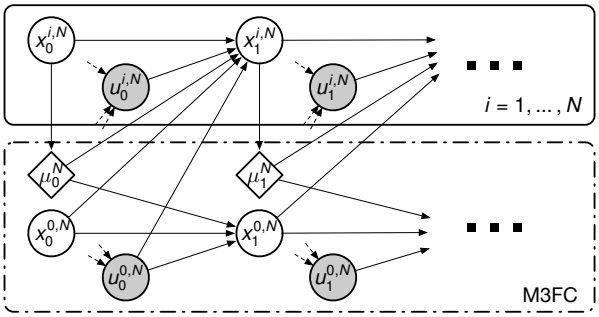

Figure 3. The dynamics (1) as a probabilistic graphical model, with actions in grey (inputs omitted for readability). Diamonds denote deterministic functions. M3FC abstracts minor agents $i \in [N]$ by a LLN, considering only their MF as variables in the dotted box.

Accordingly, minor agent policies are replaced by MFC policies $\hat{\pi}$ mapping from current μ_t to desired state-action distribution h_t. The limiting M3FC model abstracts away all the minor agents in the finite system, and considers only the MF and the major agents, as visualized in Figure 3. The reason for writing joint h_t is mostly technical, as for deterministic $\hat{\pi}$, we write $\pi_t = \Phi(\hat{\pi}_t)$ to reobtain agent policies μ_t-a.e. uniquely by disintegration (Kallenberg, 2017) of $h_t = \hat{\pi}_t(\mu_t)$ into $\mu_t \otimes \pi'_t$ with decision rule $\pi'_t \in \mathcal{P}(\mathcal{U})^{\mathcal{X}}$ and using $\pi_t(\mu_t) \equiv \pi'_t$. Inversely, any $\pi \in \Pi$ is represented in the MFC MDP by deterministic $\hat{\pi}_t = \Phi^{-1}(\pi)_t = \mu_t \otimes \pi_t$.

Remark 2. *Strictly speaking, in finite-agent control one jointly selects actions* $(u_t^{0,N}, u_t^{1,N}, \dots, u_t^{N,N})$ *given joint states* $(x_t^{0,N}, x_t^{1,N}, \dots, x_t^{N,N})$. *But intuitively, (i) joint states reduce to* $(x_t^{0,N}, \mu_t^N)$, *while (ii) joint actions are replaced by the LLN and sampling actions. Optimality of MFC solutions over larger classes of heterogeneous or joint policies is plausible, but to the best of our knowledge, general result are still limited. See also Appendix Q.*

For the unfamiliar reader, in Appendix B we recap basic deterministic MFC without major agents or common noise. There, we recap Lipschitz approximation theorems and dynamic programming principles in compact spaces.

Common noise and global states. In the classical sense (Perrin et al., 2020; Motte and Pham, 2022), common noise is given by random noise $\epsilon_t^0 \sim p_\epsilon(\epsilon_t^0)$ sampled from a fixed distribution p_ϵ, and affects all minor agents at once, $x_{t+1}^{i,N} \sim p(x_{t+1}^{i,N} \mid x_t^{i,N}, u_t^{i,N}, \epsilon_t^0, \mu_t^N)$. This allows to model systems with stochastic MFs and inter-agent correlation, and has added difficulty to the theoretical analysis (Carmona et al., 2016). Of similar interest are also "major" global states $x_t^{0,N}$, which need not be sampled from fixed distributions but evolve dynamically (for MFC with finite global states, see e.g. Mondal et al. (2023)).

Both common noise and global states are contained in the M3FC model by using a trivial major agent without actions. We also note that, in general, common noise is equivalent to global states, as global states can be integrated into the minor state conditioned on the common noise. However, for computational purposes the separation of global states and minor agent states can be helpful, as the simplex $\mathcal{P}(\mathcal{X})$ over minor states can be kept smaller for methods based on discretization of the simplex.

2.3. Dynamic Programming

As a first step, it is well known that stationary (time-independent) policies suffice for optimality in infinite-horizon discounted MDPs. In the following, this property is also verified for the M3FC MDP. For the following technical results, we assume standard Lipschitz conditions (Gu et al., 2021; Mondal et al., 2022; Pásztor et al., 2023).

Assumption 1. *The transition kernels* p, p^0 *and rewards* r *are Lipschitz with constants* L_p, L_{p^0}, L_r.

Assumption 1 is true, e.g., in finite spaces if transition matrix entries of P are Lipschitz in the $|\mathcal{X}|$-dimensional MF vector. The sufficiency of stationary policies is obtained by the dynamic programming principle, which can also be used to compute exact optimal policies in the M3FC MDP. We use the value function V^* as the fixed point of the Bellman equation, $V^*(x^0, \mu) = \max_{(h,u^0) \in \mathcal{H}(\mu) \times \mathcal{U}^0} r(x^0, u^0, \mu) + \gamma \mathbb{E}_{y^0 \sim p^0(y^0 | x^0, u^0, \mu)} V^*(y^0, T(x^0, u^0, \mu, h))$.

Theorem 1. *Under Assumption 1, there exist optimal stationary, deterministic policies* $\hat{\pi}$, π^0 *for the M3FC MDP (3) by choosing* $(\hat{\pi}(x^0, \mu), \pi^0(x^0, \mu))$ *from the maximizers of* $\arg\max_{(h,u^0) \in \mathcal{H}(\mu) \times \mathcal{U}^0} r(x^0, u^0, \mu) + \gamma \mathbb{E}_{y^0 \sim p^0(y^0 | x^0, u^0, \mu)} V^*(y^0, T(x^0, u^0, \mu, h))$.

Remark 3. *We obtain existence of optimal deterministic stationary minor and major policies* $\hat{\pi}$, π^0 *via optimal joint policies* $\tilde{\pi} \equiv \hat{\pi} \otimes \pi^0$, $(h_t, u_t^0) \sim \tilde{\pi}((h_t, u_t^0) \mid x_t^0, \mu_t)$.

The results follow from classical MDP theory (Hernández-Lerma and Lasserre, 2012). Thus, we may solve M3FC problems through the DPP, or approximately by using policy gradients with *stationary* policies for the M3FC MDP, which has naturally continuous actions.

2.4. Finite Agent Convergence

To show the approximate optimality of M3FC solutions, we first obtain **propagation of chaos** (Sznitman, 1991) – convergence of empirical MFs to the limiting MF. The result theoretically backs the reduction of multi-agent control to *single-agent* MDPs, as there is no loss of optimality in the finite problem by considering the M3FC problem. We assume standard Lipschitz conditions on policies (Gu et al., 2021; Mondal et al., 2022; Pásztor et al., 2023).

Assumption 2. *The classes of policies* Π, Π^0 *are equi-*

Lipschitz sets of policies, i.e. there exists $L_\Pi > 0$ such that for all t and $\pi \in \Pi$, $\pi_t \in \mathcal{P}(\mathcal{U})^{\mathcal{X}\times\mathcal{P}(\mathcal{X})}$ is L_Π-Lipschitz, and similarly for major policies $\pi^0 \in \Pi^0$.

We note that Lipschitz policies are natural, as we usually parametrize policies in a Lipschitz manner; in particular, neural networks allow Lipschitz analysis (Pásztor et al., 2023; Herrera et al., 2023; Araujo et al., 2023). The result is that the limiting system approximates large finite systems.

Theorem 2. *Fix any family of equi-Lipschitz functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}^0\times\mathcal{U}^0\times\mathcal{P}(\mathcal{X})}$ with shared Lipschitz constant $L_\mathcal{F}$. Under Assumptions 1 and 2, $(x_t^{0,N}, u_t^{0,N}, \mu_t^N)$ converges weakly to (x_t^0, u_t^0, μ_t), uniformly over $f \in \mathcal{F}$, $(\pi, \pi^0) \in \Pi \times \Pi^0$, $\hat{\pi} = \Phi^{-1}(\pi)$ at all times $t \in \mathbb{N}$,*

$$\sup_{f,\pi,\pi^0} \left| \mathbb{E}\left[f(x_t^{0,N}, u_t^{0,N}, \mu_t^N) - f(x_t^0, u_t^0, \mu_t) \right] \right| \to 0. \quad (4)$$

Further, the convergence rate is $\mathcal{O}(1/\sqrt{N})$ if $|\mathcal{X}| < \infty$.

The above motivates M3FC by the following **near optimality** result of M3FC MDP solutions in the finite system, as it suffices to optimize over stationary M3FC policies.

Corollary 1. *Under Assumptions 1 and 2, optimal deterministic M3FC MDP policies $(\hat{\pi}^*, \pi^{0*}) \in \arg\max_{(\hat{\pi},\pi^0)} J(\hat{\pi}, \pi^0)$ with $\Phi(\hat{\pi}^*) \in \Pi$ yield ε-optimal $(\Phi(\hat{\pi}^*), \pi^{0*})$ with $\varepsilon \to 0$ as $N \to \infty$ in the finite system, $J^N(\Phi(\hat{\pi}^*), \pi^{0*}) \geq \sup_{(\pi,\pi^0)\in\Pi\times\Pi^0} J^N(\pi, \pi^0) - \varepsilon$.*

Therefore, one may solve difficult finite-agent MARL by detouring over the corresponding M3FC MDP as depicted in Figure 4, reducing to an MDP of a complexity independent of the number of agents N, which we solve in Section 3.

3. Major-Minor Mean Field MARL

As indicated in the prequel and in Figure 2, MARL via M3FC generalizes both single-agent RL and MARL via MFC in the searched policy solution space. Therefore, in M3FC one only optimizes over a tractable, smaller solution space of a single minor and major policy Π, Π^0. At the same time, the framework is highly general and handles arbitrary major agents with many minor agents simultaneously. The reduction of MARL problems to a fixed-complexity single-agent M3FC MDP is the key. In this section, we develop MARL algorithms based on the M3FC framework.

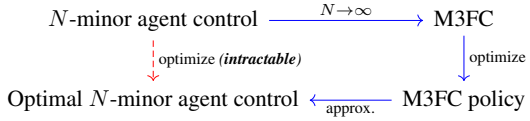

Figure 4. Approximation of intractable N-agent control by M3FC (blue path), the solution of which is near-optimal for large N.

Recalling the motivation of MFC, it is crucial to find tractable sample-based MARL techniques for both complex problems where other methods fail, and for problems where we have no access to the dynamics or reward model. Relating to the former, RL has been applied before to solve MFC given that we know the MFC model equations (Carmona et al., 2023b; Pásztor et al., 2023; Mondal et al., 2022). However, regarding the latter, we should instead use the MFC formalism to give rise to novel *MARL* algorithms.

While literature usually focused analysis on the former, in our work we analyze the proposed algorithm not on limiting M3FC MDPs, but on the more interesting finite M3FC system. In particular, if the M3FC MDP is known, one can instantiate finite systems of any size for training. We consider the following perspective: By Theorem 2, the M3FC MDP is approximated well by the finite system. Therefore, we solve the limiting M3FC MDP by applying our proposed algorithm directly to finite M3FC systems.

Since we know by Theorem 1 that stationary policy suffice, we solve the M3FC MDP (3) using stationary policies and single-agent RL techniques but on its finite multi-agent instance (1), the combination of which we aptly refer to as Major-Minor Mean Field MARL (M3FMARL). The result is Algorithm 1, where we directly apply RL to multi-agent systems (1) by observing next states $(x_{t+1}^{0,N}, \mu_{t+1}^N)$ and rewards $r_t^N := r(x_t^{0,N}, u_t^{0,N}, \mu_t^N)$. The algorithm can be understood as a kind of hierarchical algorithm, as M3FC MDP actions specify behavior for all minor agents at once.

3.1. M3FC-based Policy Gradients

The proposed algorithm is theoretically motivated. As shown in the following, finite-agent policy gradients (PG) estimate the true limiting M3FC MDP PG. First, note that finite state-actions $\mathcal{X}, \mathcal{U}$ lead to continuous M3FC MDP actions $\mathcal{H}(\mu)$, while continuous $\mathcal{X}, \mathcal{U}$ even yield infinite-dimensional $\mathcal{H}(\mu)$. Therefore, we have at least continuous

Algorithm 1 M3FMARL

for $n = 0, 1, \ldots$ **do**
  for $t = 0, \ldots, B_{\text{len}} - 1$ **do**
    Sample M3FC action from RL policy, i.e. $u_t \equiv (u_t^{0,N}, \pi_t') \sim \tilde{\pi}^\theta(\cdot \mid x_t^{0,N}, \mu_t^N)$.
    for $i = 1, \ldots, N$ **do**
      Sample i-th minor action $u_t^{i,N} \sim \pi_t'(\cdot \mid x_t^{i,N})$.
    end for
    Execute $\{u_t^{0,N}, u_t^{1,N}, \ldots\}$ for next reward r_t^N, state $(x_{t+1}^{0,N}, \mu_{t+1}^N)$ and termination $d_{t+1} \in \{0, 1\}$.
  end for
  Perform an update (on policy $\tilde{\pi}^\theta$) using transitions $B = ((x_t^{0,N}, \mu_t^N), u_t, r_t^N, d_{t+1}, (x_{t+1}^{0,N}, \mu_{t+1}^N))_{t\geq 0}$.
end for

MDPs, complicating value-based learning.

For this reason, we mainly consider PG methods to solve M3FC-type MARL problems. We parametrize M3FC MDP solutions via RL policies $\tilde{\pi}^\theta$ with parameters θ, outputting $\xi \in \Xi$ from some compact parameter space Ξ with a Lipschitz map $\Gamma(\xi) = \pi'_t$ to L_Π-Lipschitz minor agent decision rules π'_t (formally, $h_t = \mu_t \otimes \pi'_t$). Assuming the Lipschitzness of the policy network and its gradient in all arguments, on which there has been a great number of recent literature (see e.g. Herrera et al. (2023); Araujo et al. (2023) and references therein), we formulate Assumption 3.

Assumption 3. *The parameter map Γ, joint policy $\tilde{\pi}^\theta$ and log-gradient $\nabla_\theta \log \tilde{\pi}^\theta$ (or gradient $\nabla_\theta \tilde{\pi}^\theta$) are L_Γ, $L_{\tilde{\pi}}$, $L_{\nabla\tilde{\pi}}$-Lipschitz and uniformly bounded.*

Then, we can apply the PG theorem (Sutton et al., 1999) for the M3FC MDP. The M3FC MDP (3) essentially substitutes many-agent systems (1), which are natural approximations of the M3FC MDP by Theorem 2. Therefore, we show that M3FMARL (Algorithm 1) – single-agent PG on the multi-agent M3FC system – approximates the true PG of the limiting M3FC MDP, in the case of many minor agents. In other words, M3FMARL solves MARL by approximately solving the single-agent M3FC MDP using policy gradients.

Theorem 3. *Under Assumptions 1, 2 and 3, the approximate PG of joint policy $\tilde{\pi}^\theta$ computed on the finite M3FC system* (1) *in Algorithm 1 uniformly tends to the true PG of the M3FC MDP* (3)*, as $N \to \infty$.*

Importantly, the underlying MDP complexity is *independent* of the number of minor agents. Therefore, we expect Algorithm 1 to perform well in M3FC-type problems, possibly compared to straightforward MARL where each agent is handled separately. Intuitively, for many agents, the reward signal for any single agent can become uninformative: A cooperative, "averaged" reward remains almost unaffected by a single agent's actions. This well-known credit assignment issue is therefore solved by the hierarchical structure of M3FC, as credit is assigned to M3FC actions, which affect all minor agents at once and hence receive aggregated credit. Another advantage is that MFC profits from any advances in single-agent RL.

3.2. Implementation Details

We use the proximal policy optimization (PPO) algorithm (Schulman et al., 2017) to obtain a M3FC policy π_{RL}, instantiating the major minor mean field PPO (M3FPPO) algorithm as an instance of M3FMARL, Algorithm 1. Other PG algorithms (A2C, leading to M3FA2C) are also compared in our experiments. We parametrize MFs in $\mathcal{P}(\mathcal{X})$ and joint distributions in $\mathcal{H}(\mu_t^N)$. In practice, for finite $\mathcal{X}, \mathcal{U}$, the parametrization of $\mathcal{P}(\mathcal{X})$ is immediate by finite-dimensional vectors $\mu_t^N \in \mathcal{P}(\mathcal{X})$. For M3FC actions, consider – in addition to the major agent action – the matrix $\xi \in [-1, 1]^{\mathcal{X}\times\mathcal{U}}$, which is mapped to probabilities of minor actions in any minor state $\pi'_t(u \mid x) := Z^{-1}(\xi_{xu}+1+\epsilon)$, for small $\epsilon = 10^{-10}$ and normalizer Z. For continuous $\mathcal{X}, \mathcal{U}$, we instead partition $\mathcal{X}$ into M bins and represent μ_t^N as a histogram, mapping $\xi \in [-1, 1]^{M\times 2}$ to diagonal Gaussian means and standard deviations, $\mu_{\mathcal{X}_i} \in \mathcal{U}$, $\sigma_{\mathcal{X}_i} \in [\epsilon, 0.25+\epsilon]$, for each of M bins $\mathcal{X}_i \subseteq \mathcal{X}$. Major actions $u_t^{0,N}$ are categorical or diagonal Gaussian as usual. For large $\mathcal{X}, \mathcal{U}$, one could also consider kernel-based parametrizations (Cui et al., 2024).

We use two hidden layers of 256 nodes and $\tanh$ activations for the neural networks of the policies. The neural network policy outputs parameters of a diagonal Gaussian over the major action u^0 and matrices U as discussed above. In the discrete Beach scenario below, the neural network instead outputs a categorical distribution using a final softmax layer. We used no GPUs and around 300,000 CPU core hours on Intel Xeon Platinum 9242 CPUs. Optimal transport costs are computed using POT (Flamary et al., 2021). Our M3FC MDP implementation follows the gym interface (Brockman et al., 2016), while the implementation of multi-agent RL as in the following fulfills RLlib interfaces (Liang et al., 2018). The RL implementations in our work are based on MARLlib 1.0 (Hu et al., 2023a) (MIT license), which uses RLlib 1.8 (Liang et al., 2018) (Apache-2.0 license) with hyperparameters in Table 3, and otherwise default settings.

3.3. Comparison to MARL

The M3FMARL algorithm falls into the paradigm of centralized training with decentralized execution (CTDE) (Zhang et al., 2021), as we sample a single central M3FC MDP action during training, but enable decentralized execution by sampling π'_t separately on each agent instead. For instance, when converged to a deterministic M3FC policy (of which an optimal one is guaranteed to exist by Theorem 1), the M3FC action is always trivially equal for all agents.

Since we also consider continuous minor agent action spaces in our experiments, we compare against PG methods for MARL. In particular, we firstly consider Independent PPO (IPPO), as PPO with independent learning (Tan, 1993) and parameter sharing (Gupta et al., 2017), and secondly also Multi-Agent PPO (MAPPO) with centralized critics. The latter has repeatedly shown strong state-of-the-art performance in cooperative MARL (de Witt et al., 2020; Papoudakis et al., 2021; Yu et al., 2022). We also separate major and minor agent policies for improved performance of IPPO / MAPPO. For comparison, we use the same observations for the policy input as in M3FMARL. The policy network architectures match, and the same PPO implementation and hyperparameters are shared with M3FPPO in Table 3. Minor agents are additionally allowed to observe their own states. More details can be found in Appendix R.

4. Experiments

In this section, we demonstrate the performance of M3FPPO on illustrative, practical problems. Unless noted otherwise, we use $M = 49$ bins ($M = 7$ in *Potential*), train for around 24 hours, and train M3FPPO on the finite-agent system (1) with $N = 300$ minor agents unless noted otherwise (similar results for less agents in Appendix R). Full descriptions and additional experiments and discussions are in Appendix R.

4.1. Problems

To verify the usefulness of M3FMARL whenever the M3FC model (1) is accurate, we consider 5 benchmark tasks that fulfill the M3FC modelling assumptions. To begin, the simple two Gaussian (**2G**) problem has no major agent and is equipped with a time-dependent major state: A periodic, time-variant mixture of two Gaussians μ_t^* – the major state – is noisily observed analogously to μ_t^N via $M = 49$ bins. Minor agents should then track the mixture distribution over time, which can find application for example in UAV-based cellular coverage of dynamic users (Mozaffari et al., 2016). In the **Formation** problem, we extend such formation control with major agents. In addition to 2G, one added major agent tracks a moving target. Meanwhile, minor agents instead track a formation around the dynamic major agent, see e.g. Yang et al. (2021) for applications. The **Beach** bar process is a studied classic (Arthur, 1994; Perrin et al., 2020), where minor agents minimize their distances to a bar and additionally avoid crowded areas. Here, the bar moves on a discrete torus. The **Foraging** problem is archetypal of swarm intelligence (Brambilla et al., 2013), and has agents forage randomly generated foraging areas. In particular, we can consider the logistics scenario depicted in Figure 1, where a major package truck moves in a restricted space (roads) while minor drones collect packages for urban parcel delivery (Marinelli et al., 2018). Drones fill up at package "foraging" areas, and unload near the major agent. Lastly, in the **Potential** problem, minor agents can generate a potential landscape, the gradient of which pushes the major agent – e.g., a large object affected by magnetic active matter (Jin and Zhang, 2021) – to be delivered to a variable target.

4.2. Evaluation

In Figure 5, we see that M3FPPO learning is stable, as M3FPPO reduces hard-to-analyze MARL to single-agent RL, avoiding pathologies of MARL such as non-stationarity of multi-agent learning, or the combinatorial complexity over numbers of agents. In Figure 6, we find similar success in directly training M3FPPO for small N instead of transferring from high N. We conclude that M3FPPO remains applicable even with as few as 5 agents. M3FPPO usually compares well against its A2C variant (M3FA2C) and IPPO / MAPPO, see Table 2 and Appendix R.2. Meanwhile, IPPO / MAPPO under the same hyperparameters as M3FPPO (large batch sizes, see Table 3) can be more unstable and lead to worse results, see Figure 7.

Qualitative behavior. In Figure 8, we observe successfully trained behavior in Beach and Foraging: In Beach, M3FPPO learns to accumulate up to 70% of agents on the bar, as more agents on the space lead to a suboptimal reduction in rewards. In Foraging, we find that agents successfully deplete foraging areas shown in the bottom left, moving on afterwards. Further, M3FPPO successfully learns to form mixtures of Gaussians in 2G, a Gaussian around a moving major agent successfully tracking its target in Formation, and similar success in pushing the major agent towards its target in Potential, see Appendix R.3.

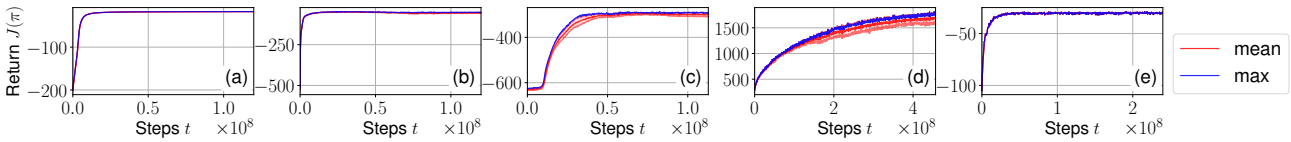

Figure 5. Training curves (mean episode return) of M3FPPO (red), with shaded standard deviation, and maximum (blue) over all three trials (two for Foraging). (a) 2G; (b) Formation; (c) Beach; (d) Foraging; (e) Potential.

Table 2. Comparison of mean episode returns between best trained policies of standard MARL and M3FMARL methods on a system with $N = 20$ agents ($\pm$ 95% confidence interval, for a number of episodes as in Figure 9).

Problem	IPPO	MAPPO	M3FA2C	M3FPPO
2G	-43.9 $\pm$ 1.1	-26.0 $\pm$ 0.5	-30.6 $\pm$ 0.6	**-22.2** $\pm$ 0.56
Formation	**-51.1** $\pm$ 2.4	-101.1 $\pm$ 7.1	-79.2 $\pm$ 3.1	-63.9 $\pm$ 4.2
Beach	-350.3 $\pm$ 3.4	-342.9 $\pm$ 4.7	-424.8 $\pm$ 5.5	**-303.5** $\pm$ 3.4
Foraging	735.3 $\pm$ 46.4	803.9 $\pm$ 54.6	1398.0 $\pm$ 57.1	**1479.4** $\pm$ 36.3
Potential	-27.1 $\pm$ 1.4	**-26.7** $\pm$ 1.7	-50.4 $\pm$ 5.5	-31.3 $\pm$ 1.3

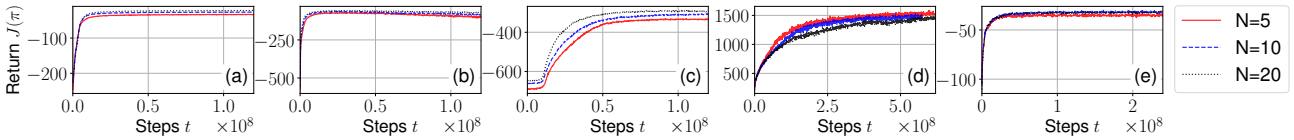

Figure 6. Training curves (mean episode return vs. time steps) of M3FPPO, trained on the finite systems with $N \in \{5, 10, 20\}$. (a) 2G; (b) Formation; (c) Beach; (d) Foraging; (e) Potential.

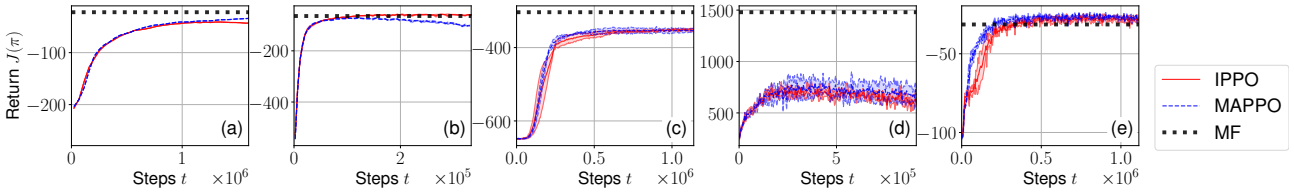

Figure 7. Comparing IPPO / MAPPO vs. asymptotic results of M3FPPO (MF, ours), as in Figure 5 (no maxima, $N = 20$).

Quantitative support of theory. In Figure 9, we *transfer* the trained M3FPPO policy to $N = 2, \ldots, 50$, comparing against the performance in the limit ($N = 500$). As N grows, the performance converges to the limit, supporting Theorem 2 and Corollary 1. Any sufficiently large system has the same limiting performance as predicted by the theory. We thus have empirical support for scalability, and also transferability between varying numbers of minor agents.

Comparison to MARL. Comparing Figures 5, 7 and Table 2, we see that (i) by experience sharing, standard MARL can be more sample-efficient, as each step gives N samples instead of just one; and (ii) M3FPPO matches or outperforms IPPO and MAPPO, despite having significantly less control over minor agent actions: All minor agents in a bin (with similar minor agent states) use the same action distributions, which suffices for strong results.

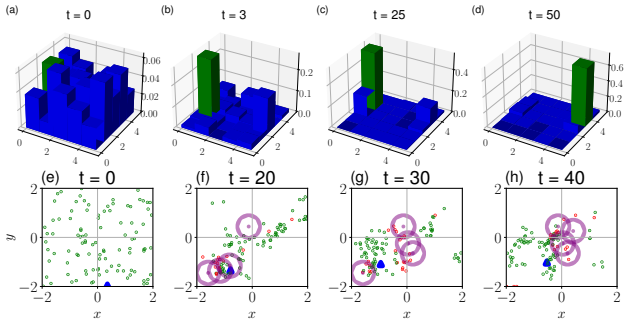

Figure 8. Qualitative visualization of M3FC in Beach (a-d), Foraging (e-h). (a-d): empirical MF, major agent & target in green; (e-h): blue / green triangle: major agent / target; green / red dots: less- / more-than-half encumbered minor agents; purple: current foraging areas.

Decentralized execution. Lastly, decentralized execution by agent-wise randomization – i.e. sampling M3FC actions per agent instead of a single shared, correlated M3FC action – has little to no effect, and can even marginally improve performance, see e.g., Beach in Figure 9(c). Figure 9 verifies the performance of M3FMARL as a CTDE method.

5. Conclusion and Discussion

We have proposed a generalization of MDPs and MFC, enabling tractable state-of-the-art MARL on general many-agent systems, with both theoretical and empirical support. Beyond the current model and its optimality guarantees, one could work on extended optimality conjectures in Appendix Q, refined approximations (Gast and Van Houdt, 2018), and local interactions (Qu et al., 2020b). Algorithmically, M3FC MDP actions $\mathcal{H}(\mu)$ could move beyond binning $\mathcal{X}$ to gain performance, e.g. via kernels (Cui et al., 2024). Lastly, one may try to quantify convergence to the rate $\mathcal{O}(1/\sqrt{N})$ for non-finite $\mathcal{X}$, as the current proof would need hard-to-verify or unrealistic d_Σ-Lipschitzness.

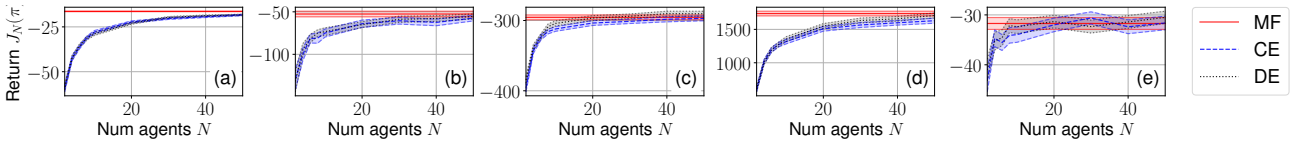

Figure 9. Mean episode return of M3FC policy in finite systems as in Figure 5 over (a-c) 100, (d) 300 or (e) 500 trials (95% confidence interval shaded). MF: CE, $N = 500$; CE / DE: centralized / decentralized execution.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgements

This work has been co-funded by the LOEWE initiative (Hesse, Germany) within the emergenCITY center, and the Hessian Ministry of Science and the Arts (HMWK) within the projects "The Third Wave of Artificial Intelligence - 3AI" and hessian.AI. The authors acknowledge the Lichtenberg high performance computing cluster of the TU Darmstadt for providing computational facilities for the calculations of this research. We thank anonymous reviewers for their helpful comments to improve the manuscript.

References

David F Anderson and Thomas G Kurtz. Continuous time Markov chain models for chemical reaction networks. In *Design and Analysis of Biomolecular Circuits*, pages 3–42. Springer, 2011.

Alexandre Araujo, Aaron J Havens, Blaise Delattre, Alexandre Allauzen, and Bin Hu. A unified algebraic perspective on Lipschitz neural networks. In *Proc. ICLR*, pages 1–15, 2023.

W Brian Arthur. Inductive reasoning and bounded rationality. *Am. Econ. Rev.*, 84(2):406–411, 1994.

Nicole Bäuerle. Mean field Markov decision processes. *Appl. Math. Optim.*, 88(1):12, 2023.

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

Daniel S Bernstein, Robert Givan, Neil Immerman, and Shlomo Zilberstein. The complexity of decentralized control of Markov decision processes. *Math. Oper. Res.*, 27(4):819–840, 2002.

Patrick Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

Ofelia Bonesini, Luciano Campi, and Markus Fischer. Correlated equilibria for mean field games with progressive strategies. *arXiv:2212.01656*, 2022.

Manuele Brambilla, Eliseo Ferrante, Mauro Birattari, and Marco Dorigo. Swarm robotics: A review from the swarm engineering perspective. *Swarm Intell.*, 7(1):1–41, 2013.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. OpenAI gym. *arXiv:1606.01540*, 2016.

Theophile Cabannes, Mathieu Laurière, Julien Perolat, Raphael Marinier, Sertan Girgin, Sarah Perrin, Olivier Pietquin, Alexandre M Bayen, Eric Goubault, and Romuald Elie. Solving n-player dynamic routing games with congestion: A mean-field approach. In *Proc. AAMAS*, volume 21, pages 1557–1559, 2022.

Peter E Caines and Minyi Huang. Graphon mean field games and the GMFG equations: ε-Nash equilibria. In *Proc. IEEE CDC*, pages 286–292, 2019.

Peter E Caines and Arman C Kizilkale. ϵ-Nash equilibria for partially observed LQG mean field games with a major player. *IEEE Trans. Automat. Contr.*, 62(7):3225–3234, 2016.

Luciano Campi and Markus Fischer. Correlated equilibria and mean field games: a simple model. *Math. Oper. Res.*, 2022.

René Carmona. Applications of mean field games in financial engineering and economic theory. *arXiv:2012.05237*, 2020.

René Carmona and François Delarue. *Probabilistic Theory of Mean Field Games with Applications I-II*. Springer, 2018.

René Carmona, François Delarue, and Daniel Lacker. Mean field games with common noise. *Ann. Probab.*, 44(6): 3740–3803, 2016.

René Carmona, Quentin Cormier, and H Mete Soner. Synchronization in a Kuramoto mean field game. *Commun. Partial. Differ. Equ.*, 48(9):1214–1244, 2023a.

René Carmona, Mathieu Laurière, and Zongjun Tan. Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning. *Ann. Appl. Probab.*, 33(6B): 5334–5381, 2023b.

Kai Cui and Heinz Koeppl. Approximately solving mean field games via entropy-regularized deep reinforcement learning. In *Proc. AISTATS*, pages 1909–1917, 2021.

Kai Cui and Heinz Koeppl. Learning graphon mean field games and approximate Nash equilibria. In *Proc. ICLR*, pages 1–31, 2022.

Kai Cui, Anam Tahir, Mark Sinzger, and Heinz Koeppl. Discrete-time mean field control with environment states. In *Proc. IEEE CDC*, pages 5239–5246, 2021.

Kai Cui, Sascha H. Hauck, Christian Fabian, and Heinz Koeppl. Learning decentralized partially observable mean field control for artificial collective behavior. In *Proc. ICLR*, 2024.

Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a Nash equilibrium. *SIAM J. Comput.*, 39(1):195–259, 2009.

Christian Schroeder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip HS Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the Starcraft multi-agent challenge? *arXiv:2011.09533*, 2020.

Ronald A DeVore and George G Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

Boualem Djehiche, Alain Tcheukam, and Hamidou Tembine. Mean-field-type games in engineering. *AIMS Electron. Electr. Eng.*, 1(1):18–73, 2017.

Alex Dunyak and Peter E Caines. Large scale systems and SIR models: A featured graphon approach. In *Proc. IEEE CDC*, pages 6928–6933. IEEE, 2021.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *J. Mach. Learn. Res.*, 22(78):1–8, 2021.

Sriram Ganapathi Subramanian, Pascal Poupart, Matthew E Taylor, and Nidhi Hegde. Multi type mean field reinforcement learning. In *Proc. AAMAS*, volume 19, pages 411–419, 2020.

Sriram Ganapathi Subramanian, Matthew E Taylor, Mark Crowley, and Pascal Poupart. Partially observable mean field reinforcement learning. In *Proc. AAMAS*, volume 20, pages 537–545, 2021.

Nicolas Gast and Bruno Gaujal. A mean field approach for optimization in discrete time. *Discret. Event Dyn. Syst.*, 21(1):63–101, 2011.

Nicolas Gast and Benny Van Houdt. A refined mean field approximation. *ACM SIGMETRICS Perform. Eval. Rev.*, 46(1):113–113, 2018.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field controls with Q-learning for cooperative MARL: convergence and complexity analysis. *SIAM J. Math. Data Sci.*, 3(4):1168–1196, 2021.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Dynamic programming principles for mean-field controls with learning. *Oper. Res.*, 2023.

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, pages 9731–9739, 2024.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. Learning mean-field games. In *Proc. NeurIPS*, pages 4966–4976, 2019.

Xin Guo, Anran Hu, Renyuan Xu, and Junzi Zhang. A general framework for learning mean-field games. *Math. Oper. Res.*, 2022.

Jayesh K Gupta, Maxim Egorov, and Mykel Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *Proc. AAMAS*, pages 66–83, 2017.

Onésimo Hernández-Lerma and Jean B Lasserre. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media, 2012.

Onésimo Hernández-Lerma and Myriam Muñoz de Ozak. Discrete-time Markov control processes with discounted unbounded costs: optimality criteria. *Kybernetika*, 28(3): 191–212, 1992.

Calypso Herrera, Florian Krach, and Josef Teichmann. Local lipschitz bounds of deep neural networks. *arXiv:2004.13135*, 2023.

Siyi Hu, Yifan Zhong, Minquan Gao, Weixun Wang, Hao Dong, Xiaodan Liang, Zhihui Li, Xiaojun Chang, and Yaodong Yang. MARLlib: A scalable and efficient multi-agent reinforcement learning library. *J. Mach. Learn. Res.*, 2023a.

Yuanquan Hu, Xiaoli Wei, Junji Yan, and Hengxi Zhang. Graphon mean-field control for cooperative multi-agent reinforcement learning. *J. Franklin Inst.*, 2023b.

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. *Commun. Inf. Syst.*, 6(3):221–252, 2006.

Dongdong Jin and Li Zhang. Collective behaviors of magnetic active matter: Recent progress toward reconfigurable, adaptive, and multifunctional swarming micro/nanorobots. *Acc. Chem. Res.*, 55(1):98–109, 2021.

Olav Kallenberg. *Random measures, theory and applications*, volume 1. Springer, 2017.

István Z Kiss, Joel C Miller, and Péter L Simon. *Mathematics of Epidemics on Networks: From Exact to Approximate Models*, volume 46. Springer, 2017. doi: 10.1007/978-3-319-50806-1.

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

Mathieu Laurière, Sarah Perrin, Matthieu Geist, and Olivier Pietquin. Learning mean field games: A survey. *arXiv:2205.12944*, 2022.

Eric Liang, Richard Liaw, Robert Nishihara, Philipp Moritz, Roy Fox, Ken Goldberg, Joseph Gonzalez, Michael Jordan, and Ion Stoica. RLlib: Abstractions for distributed reinforcement learning. In *Proc. ICML*, pages 3053–3062, 2018.

Xin Liu, Honghao Wei, and Lei Ying. Scalable and sample efficient distributed policy gradient algorithms in multi-agent networked systems. *arXiv:2212.06357*, 2022.

Mario Marinelli, Leonardo Caggiani, Michele Ottomanelli, and Mauro Dell'Orco. En route truck–drone parcel delivery for optimal vehicle routing strategies. *IET Intell. Transp. Syst.*, 12(4):253–261, 2018.

Washim Uddin Mondal, Mridul Agarwal, Vaneet Aggarwal, and Satish V Ukkusuri. On the approximation of cooperative heterogeneous multi-agent reinforcement learning (MARL) using mean field control (MFC). *J. Mach. Learn. Res.*, 23(129):1–46, 2022.

Washim Uddin Mondal, Vaneet Aggarwal, and Satish Ukkusuri. Mean-field control based approximation of multi-agent reinforcement learning in presence of a non-decomposable shared global state. *Trans. Mach. Learn. Res.*, 2023. ISSN 2835-8856.

Médéric Motte and Huyên Pham. Mean-field Markov decision processes with common noise and open-loop controls. *Ann. Appl. Probab.*, 32(2):1421–1458, 2022.

Médéric Motte and Huyên Pham. Quantitative propagation of chaos for mean field Markov decision process with common noise. *Electron. J. Probab.*, 28:1–24, 2023.

Mohammad Mozaffari, Walid Saad, Mehdi Bennis, and Mérouane Debbah. Efficient deployment of multiple unmanned aerial vehicles for optimal wireless coverage. *IEEE Commun. Lett.*, 20(8):1647–1650, 2016.

Paul Muller, Mark Rowland, Romuald Elie, Georgios Piliouras, Julien Perolat, Mathieu Laurière, Raphael Marinier, Olivier Pietquin, and Karl Tuyls. Learning equilibria in mean-field games: Introducing mean-field PSRO. In *Proc. AAMAS*, volume 20, page 926–934, 2021.

Mojtaba Nourian and Peter E Caines. ϵ-Nash mean field game theory for nonlinear stochastic dynamical systems with major and minor agents. *SIAM J. Contr. Optim.*, 51 (4):3302–3331, 2013.

Mojtaba Nourian, Peter E Caines, Roland P Malhame, and Minyi Huang. Nash, social and centralized solutions to consensus problems via mean field control theory. *IEEE Trans. Automat. Contr.*, 58(3):639–653, 2012.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Proc. NeurIPS*, pages 27730–27744, 2022.

Georgios Papoudakis, Filippos Christianos, Lukas Schäfer, and Stefano V Albrecht. Benchmarking multi-agent deep reinforcement learning algorithms in cooperative tasks. In *Proc. NeurIPS Track Datasets Benchmarks*, 2021.

Kalyanapuram Rangachari Parthasarathy. *Probability measures on metric spaces*, volume 352. American Mathematical Soc., 2005.

Barna Pásztor, Andreas Krause, and Ilija Bogunovic. Efficient model-based multi-agent mean-field reinforcement learning. *Trans. Mach. Learn. Res.*, 2023.

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 *Proc. AAMAS*, volume 21, 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. In *Proc. NeurIPS*, volume 33, pages 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 *Proc. AAAI*, volume 36, pages 9413–9421, 2022.

Huyên Pham and Xiaoli Wei. Bellman equation and viscosity solutions for mean-field stochastic control problem. *ESAIM Contr. Optim. Calc. Var.*, 24(1):437–461, 2018.

Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. Scalable multi-agent reinforcement learning for networked systems with average reward. In *Proc. NeurIPS*, volume 33, pages 2074–2086, 2020a.

Guannan Qu, Adam Wierman, and Na Li. Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Proc. Learn. Dyn. Contr.*, pages 256–266, 2020b.

Naci Saldi, Tamer Başar, and Maxim Raginsky. Markov–Nash equilibria in mean-field games with discounted cost. *SIAM J. Contr. Optim.*, 56(6):4256–4287, 2018.

Naci Saldi, Tamer Başar, and Maxim Raginsky. Partially-observed discrete-time risk-sensitive mean-field games. In *Proc. IEEE CDC*, pages 317–322, 2019.

Sina Sanjari and Serdar Yüksel. Optimal solutions to infinite-player stochastic teams and mean-field teams. *IEEE Trans. Automat. Contr.*, 66(3):1071–1086, 2020.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609, 2020.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

Nevroz Şen and Peter E Caines. Mean field games with partially observed major player and stochastic mean field. In *Proc. IEEE CDC*, pages 2709–2715, 2014.

Nevroz Şen and Peter E Caines. Mean field game theory with a partially observed major agent. *SIAM J. Contr. Optim.*, 54(6):3174–3224, 2016.

Nevroz Şen and Peter E Caines. Mean field games with partial observation. *SIAM J. Contr. Optim.*, 57(3):2064–2091, 2019.

Hamid Shiri, Jihong Park, and Mehdi Bennis. Massive autonomous UAV path planning: A neural network based mean-field game theoretic approach. In *Proc. IEEE GLOBECOM*, pages 1–6. IEEE, 2019.

David Silver, Guy Lever, Nicolas Heess, Thomas Degris, Daan Wierstra, and Martin Riedmiller. Deterministic policy gradient algorithms. In *Proc. ICML*, pages 387–395. PMLR, 2014.

Sriram Ganapathi Subramanian, Matthew E Taylor, Mark Crowley, and Pascal Poupart. Decentralized mean field games. In *Proc. AAAI*, volume 36, pages 9439–9447, 2022.

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proc. NIPS*, pages 1057–1063, 1999.

Alain-Sol Sznitman. Topics in propagation of chaos. In *Ecole d'été de probabilités de Saint-Flour XIX—1989*, pages 165–251. Springer, 1991.

Ming Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proc. ICML*, pages 330–337, 1993.

Rinel Foguen Tchuendom, Peter E Caines, and Minyi Huang. Critical nodes in graphon mean field games. In *Proc. IEEE CDC*, pages 166–170. IEEE, 2021.

Cédric Villani. *Optimal transport: old and new*, volume 338. Springer, 2009.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354, 2019.

Minghui Wu, Xingmin Wang, Yafeng Yin, Henry Liu, Ben Wang, Jerome P Lynch, et al. Leveraging connected and automated vehicles for participatory traffic control. Technical report, University of Michigan. Center for Connected and Automated Transportation, 2023.

Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. Mean field multi-agent reinforcement learning. In *Proc. ICML*, pages 5571–5580, 2018.

Yue Yang, Yang Xiao, and Tieshan Li. Attacks on formation control for multiagent systems. *IEEE Trans. Cybern.*, 52 (12):12805–12817, 2021.

Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. Policy mirror ascent for efficient and independent learning in mean field games. In *Proc. ICML*, pages 39722–39754. PMLR, 2023.

Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Proc. NeurIPS Datasets and Benchmarks*, 2022.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. In Kyriakos G. Vamvoudakis, Yan Wan, Frank L. Lewis, and Derya Cansever, editors, *Handbook of Reinforcement Learning and Control*, pages 321–384. Springer International Publishing, Cham, 2021.

A. Related Work

In this section, we provide additional context on related works. Since the introduction of MFGs in continuous and discrete time (Huang et al., 2006; Lasry and Lions, 2007; Saldi et al., 2018), MFGs have been studied in various forms, ranging from partially observed systems (Saldi et al., 2019; Şen and Caines, 2019) over learning-based solutions (Guo et al., 2019; Perrin et al., 2020; Cui and Koeppl, 2021; Guo et al., 2022; Pérolat et al., 2022; Perrin et al., 2022; Yardim et al., 2023) on graphs (Caines and Huang, 2019; Tchuendom et al., 2021; Cui and Koeppl, 2022; Hu et al., 2023b) to considering correlated equilibria (Muller et al., 2021; Campi and Fischer, 2022; Bonesini et al., 2022).

While many works focus on non-cooperative settings with self-interested agents, this can run counter to the goal of engineering many-agent behavior, e.g., achieving cooperative behavior in swarms of drones. Instead, we focus on the related setting of cooperative MFC (Pham and Wei, 2018; Gu et al., 2023; Mondal et al., 2022), see also work on differential (Carmona and Delarue, 2018), static (Sanjari and Yüksel, 2020), or discrete-time deterministic MFC (Gast and Gaujal, 2011). For the unfamiliar reader, we point towards many extensive surveys on the topic of mean field systems (Bensoussan et al., 2013; Carmona and Delarue, 2018; Laurière et al., 2022).

In general comparison, another well-known line of mean field MARL (Yang et al., 2018; Ganapathi Subramanian et al., 2020; 2021; Subramanian et al., 2022) focuses on approximating the influence of other agents on any particular agent by their average actions. Relatedly, some MARL algorithms introduce approximations over agent neighborhoods based on exponential decay (Qu et al., 2020b;a; Liu et al., 2022). In contrast, MFC assumes dependence on the entire distribution of agents and not, e.g., pairwise terms for each neighbor, per agent.

B. Deterministic Mean Field Control

In the following, we provide proofs that were omitted in the main text. To begin, in this section we recap standard deterministic MFC. Here, our general proof technique is introduced. It generalizes to the M3FC case and allows approximation properties and dynamic programming principles beyond finite spaces and Lipschitz continuity assumptions in compact spaces, for MFC models under simple continuity. In standard MFC, we have the model without major agents,

$$u_t^{i,N} \sim \pi_t(u_t^{i,N} \mid x_t^{i,N}, \mu_t^N), \tag{5}$$

$$x_{t+1}^{i,N} \sim p(x_{t+1}^{i,N} \mid x_t^{i,N}, u_t^{i,N}, \mu_t^N) \tag{6}$$

while in the limit, we have the MF evolution

$$\mu_{t+1} = T(\mu_t, \mu_t \otimes \pi_t(\mu_t)) \coloneqq \iint p(\cdot \mid x, u, \mu_t)\pi_t(\mathrm{d}u \mid x, \mu_t)\mu_t(\mathrm{d}x) \tag{7}$$

and MFC system

$$h_t \sim \hat{\pi}_t(h_t \mid \mu_t), \quad \mu_{t+1} = T(\mu_t, h_t) \tag{8}$$

with objective $J(\hat{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(\mu_t)\right]$.

Dynamic Programming and Propagation of Chaos We may solve the hard finite-agent system (5) near-optimally by instead solving the MFC MDP, allowing direct application of single-agent RL to the MFC MDP with approximate optimality in large systems. Mild continuity assumptions are required.

Assumption B.1. *The transition kernel p and reward r are continuous.*

Assumption B.2. *The considered class of policies Π is equi-Lipschitz, i.e. there exists $L_\Pi > 0$ such that for all t and $\pi \in \Pi$, $\pi_t \in \mathcal{P}(\mathcal{U})^{\mathcal{X}\times\mathcal{P}(\mathcal{X})}$ is L_Π-Lipschitz.*

We note that Assumption B.1 holds true in studied finite spaces, if each transition matrix entry of P is continuous in the $|\mathcal{X}|$-dimensional MF vector on the simplex (but not necessarily Lipschitz as in (Gu et al., 2021; Mondal et al., 2022), the conditions of which we relax for deterministic MFC).

We show a dynamic programming principle (Hernández-Lerma and Lasserre, 2012) to solve for and show existence of a deterministic, stationary optimal policy via the value function V^* as the fixed point of the Bellman equation $V^*(\mu) = \max_{h\in\mathcal{H}(\mu)} r(\mu) + \gamma V^*(T(\mu, h))$.

Theorem B.1. *Under Assumptions B.1, there exists an optimal stationary, deterministic policy $\hat{\pi}$ for (8), with $\hat{\pi}(\mu) \in \arg\max_{h \in \mathcal{H}(\mu)} r(\mu) + \gamma V^*(T(\mu, h))$.*

This DPP can be used for computing solutions or to show optimality of stationary policies and existence of an optimum. Next, we show **propagation of chaos** (Sznitman, 1991). Here, prior proof techniques (Gu et al., 2021; Mondal et al., 2022) are extended by our approach from *finite* to general *compact* spaces.

Theorem B.2. *Fix any family of equicontinuous functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{P}(\mathcal{X})}$. Under Assumptions B.1 and B.2, the empirical MF converges weakly, uniformly over $f \in \mathcal{F}$, $\pi \in \Pi$, $\hat{\pi} = \Phi^{-1}(\pi)$, to the limiting MF at all times $t \in \mathbb{N}$, $\sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \left| \mathbb{E}\left[f(\mu_t^N)\right] - \mathbb{E}\left[f(\mu_t)\right] \right| \to 0$.*

Importantly, propagation of chaos allows one to show approximate optimality of MFC policies in the large finite control problem, which is of practical relevance for solving many-agent problems.

Corollary B.1. *Under Assumptions B.1 and B.2, an optimal deterministic MFC policy $\pi^* \in \arg\max_{\hat{\pi}} J(\hat{\pi})$ yields ε-optimal finite-agent policy $\Phi(\pi^*) \in \Pi$, $J^N(\Phi(\pi^*)) \geq \sup_{\pi \in \Pi} J^N(\pi) - \varepsilon$, with $\varepsilon \to 0$ as $N \to \infty$.*

C. Continuity of MF dynamics

First, we find continuity of the MFC dynamics T, which is used in the following proofs.

Lemma C.1. *Under Assumption B.1, we have $T(\mu_n, \nu_n) \to T(\mu, \nu)$ whenever $(\mu_n, \nu_n) \to (\mu, \nu)$,*

Proof. To show $T(\mu_n, \nu_n) \to T(\mu, \nu)$, consider any Lipschitz and bounded f with Lipschitz constant L_f, then

$$\begin{aligned}
&\left| \int f \,\mathrm{d}(T(\mu_n, \nu_n) - T(\mu, \nu)) \right| \\
&= \left| \iiint f(x') p(\mathrm{d}x' \mid x, u, \mu_n) \nu_n(\mathrm{d}x, \mathrm{d}u) - \iiint f(x') p(\mathrm{d}x' \mid x, u, \mu) \nu(\mathrm{d}x, \mathrm{d}u) \right| \\
&\leq \iint \left| \int f(x') p(\mathrm{d}x' \mid x, u, \mu_n) - \int f(x') p(\mathrm{d}x' \mid x, u, \mu) \right| \nu_n(\mathrm{d}x, \mathrm{d}u) \\
&\quad + \left| \iiint f(x') p(\mathrm{d}x' \mid x, u, \mu)(\nu_n(\mathrm{d}x, \mathrm{d}u) - \nu(\mathrm{d}x, \mathrm{d}u)) \right| \\
&\leq \sup_{x \in \mathcal{X}, u \in \mathcal{U}} L_f W_1(p(\cdot \mid x, u, \mu_n), p(\cdot \mid x, u, \mu)) \\
&\quad + \left| \iiint f(x') p(\mathrm{d}x' \mid x, u, \mu)(\nu_n(\mathrm{d}x, \mathrm{d}u) - \nu(\mathrm{d}x, \mathrm{d}u)) \right| \to 0
\end{aligned}$$

for the first term by 1-Lipschitzness of $\frac{f}{L_f}$ and Assumption B.1 (with compactness implying the uniform continuity), and for the second by $\nu_n \to \nu$ and from continuity by the same argument of $(x, u) \mapsto \iint f(x') p(\mathrm{d}x' \mid x, u, \mu)$. □

D. Proof of Theorem B.1

Proof. The MFC MDP fulfills (Hernández-Lerma and Lasserre, 2012), Assumption 4.2.1. Here, we use (Hernández-Lerma and Lasserre, 2012), Condition 3.3.4(b1) instead of (b2), see also alternatively (Hernández-Lerma and Muñoz de Ozak, 1992).

More specifically, for (Hernández-Lerma and Lasserre, 2012), Assumption 4.2.1(a), the cost function $-r$ is continuous by Assumption B.1, therefore also bounded by compactness of $\mathcal{P}(\mathcal{X})$, and finally also inf-compact on the state-action space of the MFC MDP, since for any $\mu \in \mathcal{P}(\mathcal{X})$ the set $\{h \in \mathcal{H}(\mu) \mid -r(\mu) \leq c\}$ is trivially given by $\mathcal{H}(\mu)$ whenever $-r(\mu) \leq c$, and $\emptyset$ otherwise. Here, we show that $\mathcal{H}(\mu) \subseteq \mathcal{P}(\mathcal{X} \times \mathcal{U})$ is a closed subset of the compact space $\mathcal{P}(\mathcal{X} \times \mathcal{U})$ and therefore also compact. Note first that two measures $\mu, \mu' \in \mathcal{P}(\mathcal{X})$ are equal if and only if for all continuous and bounded f we have $\int f \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu'$, see e.g. (Billingsley, 2013), Theorem 1.3.

Therefore, as $\mathcal{H}(\mu)$ is defined by its first marginal μ, $\mathcal{H}(\mu)$ can be written as an intersection

$$\mathcal{H}(\mu) = \bigcap_{f \in C_b(\mathcal{X})} \left\{ h \in \mathcal{P}(\mathcal{X} \times \mathcal{U}) \,\middle|\, \int f \otimes \mathbf{1} \,\mathrm{d}h = \int f \,\mathrm{d}\mu \right\}$$

of closed sets: Since $h \mapsto \int f \otimes \mathbf{1} \, \mathrm{d}h$ is continuous, its preimage of the closed set $\{\int f \, \mathrm{d}\mu\}$ is closed. Here, $\otimes$ denotes the tensor product of f with the function $\mathbf{1}$ equal one, i.e. $f \otimes \mathbf{1}$ is the map $(x, u) \mapsto f(x)$.

Similarly, for (Hernández-Lerma and Lasserre, 2012), Assumption 4.2.1(b), the transition dynamics T are weakly continuous, as for any $(\mu_n, \nu_n) \to (\mu, \nu) \in \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X} \times \mathcal{U})$ we have $T(\mu_n, \nu_n) \to T(\mu, \nu)$ by Lemma C.1 and therefore $\int f \, \mathrm{d}\delta_{T(\mu_n,\nu_n)} = f(T(\mu_n, \nu_n)) \to f(T(\mu, \nu)) = \int f \, \mathrm{d}\delta_{T(\mu,\nu)}$ for any continuous and bounded $f \colon \mathcal{P}(\mathcal{X}) \to \mathbb{R}$.

Furthermore, the MFC MDP fulfills (Hernández-Lerma and Lasserre, 2012), Assumption 4.2.2 by boundedness of r from Assumption B.1. Therefore, the desired statement follows from (Hernández-Lerma and Lasserre, 2012), Theorem 4.2.3. □

E. Proof of Theorem B.2

Proof. Note that we can also show the slightly stronger L_1 convergence statement with the absolute value inside of the expectation, $\sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \mathbb{E}\left[\left|f(\mu_t^N) - f(\mu_t)\right|\right] \to 0$, but since this statement is only true for deterministic MFC, we avoid it here to later extend our proof directly to M3FC.

The statement $\sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \left|\mathbb{E}\left[f(\mu_t^N)\right] - \mathbb{E}\left[f(\mu_t)\right]\right| \to 0$ is shown inductively over $t \geq 0$. At time $t = 0$, it holds by the weak LLN argument, see also the first term below. Assuming the statement at time t, then for time $t + 1$ we have

$$\begin{aligned}
&\sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \left|\mathbb{E}\left[f(\mu_{t+1}^N) - f(\mu_{t+1})\right]\right| \\
&\quad \leq \sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \left|\mathbb{E}\left[f(\mu_{t+1}^N) - f(T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right]\right| \qquad (9) \\
&\qquad + \sup_{\pi \in \Pi} \sup_{f \in \mathcal{F}} \left|\mathbb{E}\left[f(T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N))) - f(\mu_{t+1})\right]\right|. \qquad (10)
\end{aligned}$$

For the first term (9), first note that by compactness of $\mathcal{P}(\mathcal{X})$, $\mathcal{F}$ is uniformly equicontinuous, and hence admits a non-decreasing, concave (as in (DeVore and Lorentz, 1993), Lemma 6.1) modulus of continuity $\omega_{\mathcal{F}} \colon [0, \infty) \to [0, \infty)$ where $\omega_{\mathcal{F}}(x) \to 0$ as $x \to 0$ and $|f(\mu) - f(\nu)| \leq \omega_{\mathcal{F}}(W_1(\mu, \nu))$ for all $f \in \mathcal{F}$.

We also have uniform equicontinuity of $\mathcal{F}$ with respect to the space $(\mathcal{P}(\mathcal{X}), d_\Sigma)$ instead of $(\mathcal{P}(\mathcal{X}), W_1)$, as the identity map $\mathrm{id} \colon (\mathcal{P}(\mathcal{X}), d_\Sigma) \to (\mathcal{P}(\mathcal{X}), W_1)$ is uniformly continuous (as both d_Σ and W_1 metrize the topology of weak convergence, and $\mathcal{P}(\mathcal{X})$ is compact), and therefore there exists a modulus of continuity $\tilde{\omega}$ for the identity map such that for any $\mu, \nu \in (\mathcal{P}(\mathcal{X}), d_\Sigma)$, by the prequel

$$|f(\mu) - f(\nu)| \leq \omega_{\mathcal{F}}(W_1(\mathrm{id}\,\mu, \mathrm{id}\,\nu)) \leq \omega_{\mathcal{F}}(\tilde{\omega}(d_\Sigma(\mu, \nu)))$$

with $\tilde{\omega}_{\mathcal{F}} := \omega_{\mathcal{F}} \circ \tilde{\omega}$, which can be replaced by its least concave majorant (again as in (DeVore and Lorentz, 1993), Lemma 6.1).

Therefore, by Jensen's inequality, for (9) we obtain

$$\begin{aligned}
&\left|\mathbb{E}\left[f(\mu_{t+1}^N) - f(T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right]\right| \\
&\quad \leq \mathbb{E}\left[\tilde{\omega}_{\mathcal{F}}(d_\Sigma(\mu_{t+1}^N, T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N))))\right] \\
&\quad \leq \tilde{\omega}_{\mathcal{F}}\left(\mathbb{E}\left[d_\Sigma(\mu_{t+1}^N, T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right]\right)
\end{aligned}$$

irrespective of π, f via concavity of $\tilde{\omega}_{\mathcal{F}}$. Introducing for readability $x_t^N \equiv \{x_t^{i,N}\}_{i \in [N]}$, we then obtain

$$\begin{aligned}
&\mathbb{E}\left[d_\Sigma(\mu_{t+1}^N, T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right] \\
&\quad = \sum_{m=1}^{\infty} 2^{-m} \, \mathbb{E}\left[\left|\int f_m \, \mathrm{d}(\mu_{t+1}^N - T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right|\right] \\
&\quad \leq \sup_{m \geq 1} \mathbb{E}\left[\mathbb{E}_{x_t^N}\left[\left|\int f_m \, \mathrm{d}(\mu_{t+1}^N - T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right|\right]\right],
\end{aligned}$$

and by the following weak LLN argument, for the squared term and any f_m

$$\mathbb{E}_{x_t^N}\left[\left|\int f_m \, \mathrm{d}(\mu_{t+1}^N - T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N)))\right|\right]^2$$

$$
\begin{aligned}
&= \mathbb{E}_{x_t^N}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\left(f_m(x_{t+1}^{i,N}) - \mathbb{E}_{x_t^N}\left[f_m(x_{t+1}^{i,N})\right]\right)\right|\right]^2 \\
&\leq \mathbb{E}_{x_t^N}\left[\left|\frac{1}{N}\sum_{i=1}^{N}\left(f_m(x_{t+1}^{i,N}) - \mathbb{E}_{x_t^N}\left[f_m(x_{t+1}^{i,N})\right]\right)\right|^2\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}_{x_t^N}\left[\left(f_m(x_{t+1}^{i,N}) - \mathbb{E}_{x_t^N}\left[f_m(x_{t+1}^{i,N})\right]\right)^2\right] \leq \frac{4}{N} \to 0
\end{aligned}
$$

by bounding $|f_m| \leq 1$, as the cross-terms are zero by conditional independence of $x_{t+1}^{i,N}$ given x_t^N. By the prequel, the term (9) hence converges to zero.

For the second term (10), we have

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N))) - f(\mu_{t+1})\right]\right| \\
&\quad = \sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(T(\mu_t^N, \mu_t^N \otimes \pi_t(\mu_t^N))) - f(T(\mu_t, \mu_t \otimes \pi_t(\mu_t)))\right]\right| \\
&\quad \leq \sup_{\pi\in\Pi}\sup_{g\in\mathcal{G}}\left|\mathbb{E}\left[g(\mu_t^N) - g(\mu_t)\right]\right| \to 0
\end{aligned}
$$

by the induction assumption, where we defined $g = f \circ \tilde{T}^{\pi_t}$ from the class $\mathcal{G}$ of equicontinuous functions with modulus of continuity $\omega_{\mathcal{G}} \coloneqq \omega_{\mathcal{F}} \circ \omega_T$, where ω_T denotes the uniform modulus of continuity of $\mu_t \mapsto \tilde{T}^{\pi_t}(\mu_t) \coloneqq T(\mu_t, \mu_t \otimes \pi_t(\mu_t)))$ over all policies π. Here, this equicontinuity of $\{\tilde{T}^{\pi_t}\}_{\pi\in\Pi}$ follows from Lemma C.1 and the equicontinuity of functions $\mu_t \mapsto \mu_t \otimes \pi_t(\mu_t)$ due to uniformly Lipschitz Π as we show in the following, completing the proof by induction:

Consider $\mu_n \to \mu \in \mathcal{P}(\mathcal{X})$, then we have

$$
\begin{aligned}
&\sup_{\pi\in\Pi} W_1(\mu_n \otimes \pi_t(\mu_n), \mu \otimes \pi_t(\mu)) \\
&\quad = \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1}\left|\int f' \,\mathrm{d}(\mu_n \otimes \pi_t(\mu_n) - \mu \otimes \pi_t(\mu))\right| \\
&\quad \leq \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1}\left|\iint f'(x,u)(\pi_t(\mathrm{d}u \mid x, \mu_n) - \pi_t(\mathrm{d}u \mid x, \mu))\mu_n(\mathrm{d}x)\right| \\
&\qquad + \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1}\left|\iint f'(x,u)\pi_t(\mathrm{d}u \mid x, \mu)(\mu_n(\mathrm{d}x) - \mu(\mathrm{d}x))\right|
\end{aligned}
$$

where for the first term

$$
\begin{aligned}
&\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1}\left|\iint f'(x,u)(\pi_t(\mathrm{d}u \mid x, \mu_n) - \pi_t(\mathrm{d}u \mid x, \mu))\mu_n(\mathrm{d}x)\right| \\
&\quad \leq \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1}\int\left|\int f'(x,u)(\pi_t(\mathrm{d}u \mid x, \mu_n) - \pi_t(\mathrm{d}u \mid x, \mu))\right|\mu_n(\mathrm{d}x) \\
&\quad \leq \sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\leq 1}\sup_{x\in\mathcal{X}}\left|\int f'(x,u)(\pi_t(\mathrm{d}u \mid x, \mu_n) - \pi_t(\mathrm{d}u \mid x, \mu))\right| \\
&\quad = \sup_{\pi\in\Pi}\sup_{x\in\mathcal{X}} W_1(\pi_t(\cdot \mid x, \mu_n), \pi_t(\cdot \mid x, \mu)) \\
&\quad \leq L_\Pi W_1(\mu_n, \mu) \to 0
\end{aligned}
$$

by Assumption B.2, and similarly for the second by first noting 1-Lipschitzness of $x \mapsto \int \frac{f'(x,u)}{L_\Pi+1}\pi_t(\mathrm{d}u \mid x, \mu)$, as for $y \neq x$

$$
\left|\int \frac{f'(y,u)}{L_\Pi+1}\pi_t(\mathrm{d}u \mid y, \mu) - \int \frac{f'(x,u)}{L_\Pi+1}\pi_t(\mathrm{d}u \mid x, \mu)\right|
$$

$$
\begin{aligned}
&\leq \left| \int \frac{f'(y,u) - f'(x,u)}{L_\Pi + 1} \pi_t(\mathrm{d}u \mid y, \mu) \right| + \left| \int \frac{f'(x,u)}{L_\Pi + 1} (\pi_t(\mathrm{d}u \mid y, \mu) - \pi_t(\mathrm{d}u \mid x, \mu)) \right| \\
&\leq \frac{1}{L_\Pi + 1} d(y,x) + \frac{1}{L_\Pi + 1} W_1(\pi_t(\cdot \mid y, \mu), \pi_t(\cdot \mid x, \mu)) \\
&\leq \left(\frac{1}{L_\Pi + 1} + \frac{L_\Pi}{L_\Pi + 1} \right) d(x,y) \qquad (11)
\end{aligned}
$$

with $\frac{1}{L_\Pi+1} + \frac{L_\Pi}{L_\Pi+1} = 1 \leq 1$, and therefore again

$$
\begin{aligned}
&\sup_{\pi \in \Pi} \sup_{\|f'\|_{\mathrm{Lip}} \leq 1} \left| \iint f'(x,u) \pi_t(\mathrm{d}u \mid x, \mu)(\mu_n(\mathrm{d}x) - \mu(\mathrm{d}x)) \right| \\
&\quad = \sup_{\pi \in \Pi} \sup_{\|f'\|_{\mathrm{Lip}} \leq 1} (L_\Pi + 1) \left| \iint \frac{f'(x,u)}{L_\Pi + 1} \pi_t(\mathrm{d}u \mid x, \mu)(\mu_n(\mathrm{d}x) - \mu(\mathrm{d}x)) \right| \\
&\quad \leq (L_\Pi + 1) W_1(\mu_n, \mu) \to 0.
\end{aligned}
$$

This completes the proof by induction. □

F. Proof of Corollary B.1

Proof. First, we show that from uniform convergence in Theorem B.2, the finite-agent objectives converge uniformly to the MFC limit.

Lemma F.1. *Under Assumptions B.1 and B.2, the finite-agent objective converges uniformly to the MFC limit,*

$$
\sup_{\pi \in \Pi} \left| J^N(\pi) - J(\Phi^{-1}(\pi)) \right| \to 0. \qquad (12)
$$

Proof. For any $\varepsilon > 0$, choose time $T \in \mathbb{N}$ such that $\sum_{t=T}^{\infty} \gamma^t \mathbb{E} \left| \left[r(\mu_t^N) - r(\mu_t) \right] \right| \leq \frac{\gamma^T}{1-\gamma} \max_\mu 2|r(\mu)| < \frac{\varepsilon}{2}$. By Theorem B.2, $\sum_{t=0}^{T-1} \gamma^t \mathbb{E} \left| \left[r(\mu_t^N) - r(\mu_t) \right] \right| < \frac{\varepsilon}{2}$ for sufficiently large N. The result follows. ■

The approximate optimality of MFC solutions in the finite system follows immediately: By Lemma F.1, we have

$$
\begin{aligned}
&J^N(\Phi(\pi^*)) - \sup_{\pi \in \Pi} J^N(\pi) = \inf_{\pi \in \Pi} (J^N(\pi^*) - J^N(\pi)) \\
&\quad \geq \inf_{\pi \in \Pi} (J^N(\Phi(\pi^*)) - J(\pi^*)) + \inf_{\pi \in \Pi} (J(\pi^*) - J(\Phi^{-1}(\pi))) + \inf_{\pi \in \Pi} (J(\Phi^{-1}(\pi)) - J^N(\pi)) \\
&\quad \geq -\frac{\varepsilon}{2} + 0 - \frac{\varepsilon}{2} = -\varepsilon
\end{aligned}
$$

for sufficiently large N, where the second term is zero by optimality of π^* in the MFC problem. □

G. Stochastic Mean Field Control with Common Noise and Major States

For convenience, we also restate the results for MFC with major states, or common noise. We have the finite MFC system with major states

$$
u_t^{i,N} \sim \pi_t(u_t^{i,N} \mid x_t^{i,N}, x_t^{0,N}, \mu_t^N), \qquad (13a)
$$

$$
x_{t+1}^{i,N} \sim p(x_{t+1}^{i,N} \mid x_t^{i,N}, u_t^{i,N}, x_t^{0,N}, \mu_t^N), \quad x_{t+1}^{0,N} \sim p^0(x_{t+1}^{0,N} \mid x_t^{0,N}, \mu_t^N) \qquad (13b)
$$

and objective $J^N(\pi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^{0,N}, \mu_t^N)\right]$ analogous to (5), with the corresponding limiting MFC MDP with major states analogous to (8),

$$
h_t \sim \hat{\pi}_t(h_t \mid x_t^0, \mu_t), \quad \mu_{t+1} = T(x_t^0, \mu_t, h_t), \quad x_{t+1}^0 \sim p^0(x_{t+1}^0 \mid x_t^0, \mu_t) \qquad (14)
$$

with objective $J(\hat{\pi}) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^0, \mu_t)\right]$, where $T(x^0, \mu, h) \coloneqq \iint p(\cdot \mid x, u, x^0, \mu) h(\mathrm{d}x, \mathrm{d}u)$.

Assumption G.1. *The transition kernels p, p^0 and rewards r are Lipschitz continuous with constants L_p, L_{p^0}, L_r.*

Assumption G.2. *The class of policies Π are equi-Lipschitz, i.e. there exists $L_\Pi > 0$ such that for all t and $\pi \in \Pi$, $\pi_t \in \mathcal{P}(\mathcal{U})^{\mathcal{X}\times\mathcal{P}(\mathcal{X})}$ is L_Π-Lipschitz.*

Theorem G.1. *Under Assumption G.1, there exists an optimal stationary, deterministic policy $\hat{\pi}$ for the MFC MDP (14) by choosing $\hat{\pi}(x^0, \mu)$ from the maximizers of $\arg\max_{h\in\mathcal{H}(\mu)} r(x^0, \mu) + \gamma\mathbb{E}_{y^0\sim p^0(y^0|x^0,\mu)} V^*(y^0, T(x^0, \mu, h))$, with V^* the unique fixed point of the Bellman equation $V^*(x^0, \mu) = \max_{h\in\mathcal{H}(\mu)} r(x^0, \mu) + \gamma\mathbb{E}_{y^0\sim p^0(y^0|x^0,\mu)} V^*(y^0, T(x^0, \mu, h))$ (value function).*

Theorem G.2. *Fix any family of equi-Lipschitz functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}^0\times\mathcal{P}(\mathcal{X})}$ with shared Lipschitz constant $L_\mathcal{F}$ for all $f \in \mathcal{F}$. Under Assumption G.1, the random variable $(x_t^{0,N}, \mu_t^N)$ converges weakly, uniformly over $\mathcal{F}$, Π, to (x_t^0, μ_t) at all times $t \in \mathbb{N}$,*

$$\sup_{\pi\in\Pi}\sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(x_t^{0,N}, \mu_t^N) - f(x_t^0, \mu_t)\right]\right| \to 0. \tag{15}$$

Corollary G.1. *Under Assumptions G.1 and G.2, optimal deterministic MFC policies $\pi^* \in \arg\max_\pi J(\pi)$ result in ε-optimal policies $\Phi(\pi^*)$ in the finite-agent problem with $\varepsilon \to 0$ as $N \to \infty$,*

$$J^N(\Phi(\pi^*)) \geq \sup_{\pi\in\Pi} J^N(\pi) - \varepsilon. \tag{16}$$

The proofs and interpretation are directly analogous to the M3FC case and the following proofs, by leaving out the major agent actions, or alternatively using the M3FC results with a trivial singleton major action space, $|\mathcal{U}^0| = 1$.

H. Proof of Theorem 1

Proof. The proof is analogous to Appendix D by first showing the continuity of T (proof further below).

Lemma H.1. *Under Assumption 1, for any sequence $(x_n^0, u_n^0, \mu_n, \nu_n) \to (x^0, u^0, \mu, \nu) \in \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X} \times \mathcal{U})$, we have $T(x_n^0, u_n^0, \mu_n, \nu_n) \to T(x^0, u^0, \mu, \nu)$.*

For (Hernández-Lerma and Lasserre, 2012), Assumption 4.2.1(a), the cost function $-r$ is continuous by Assumption 1, therefore also bounded by compactness of $\mathcal{X}^0 \times \mathcal{P}(\mathcal{X})$, and finally also inf-compact on the state-action space of the M3FC MDP, since for any $(x^0, \mu) \in \mathcal{X}^0 \times \mathcal{P}(\mathcal{X})$ the set $\{(h, u^0) \in \mathcal{H}(\mu) \times \mathcal{U}^0 \mid -r(x^0, u^0, \mu) \leq c\}$ is given by $\mathcal{H}(\mu) \times \tilde{r}^{-1}((-\infty, c])$, where we defined $\tilde{r}(u^0) := -r(x^0, u^0, \mu)$. Note that $\mathcal{H}(\mu)$ is compact by the same argument as in Appendix D, while $\tilde{r}$ is continuous by Assumption 1 and therefore its preimage of the closed set $(-\infty, c]$ is compact.

For (Hernández-Lerma and Lasserre, 2012), Assumption 4.2.1(b), consider any continuous and bounded $f\colon \mathcal{X}^0\times\mathcal{P}(\mathcal{X}) \to \mathbb{R}$. The continuity is uniform by compactness. Hence, $\sup_{x'\in\mathcal{X}^0} |f(x', \mu'_n) - f(x', \mu')| \to 0$ as $\mu'_n \to \mu' \in \mathcal{P}(\mathcal{X})$. Thus, whenever $(x_n^0, u_n^0, \mu_n, \nu_n) \to (x^0, u^0, \mu, \nu) \in \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X}) \times \mathcal{P}(\mathcal{X} \times \mathcal{U})$, we have

$$\begin{aligned}
&\left|\iint f(x', \mu)\, \delta_{T_n^*}(\mathrm{d}\mu')\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n) - \iint f(x', \mu)\, \delta_{T^*}(\mathrm{d}\mu')\, p^0(\mathrm{d}x' \mid x^0, u^0, \mu)\right| \\
&= \left|\int f(x', T_n^*)\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n) - \int f(x', T^*)\, p^0(\mathrm{d}x' \mid x^0, u^0, \mu)\right| \\
&\leq \left|\int f(x', T_n^*)\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n) - \int f(x', T^*)\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n)\right| \\
&\quad + \left|\int f(x', T^*)\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n) - \int f(x', T^*)\, p^0(\mathrm{d}x' \mid x^0, u^0, \mu)\right| \\
&\leq \sup_{x'\in\mathcal{X}^0} |f(x', T_n^*) - f(x', T^*)| \\
&\quad + \left|\int \tilde{f}(x')\, p^0(\mathrm{d}x' \mid x_n^0, u_n^0, \mu_n) - \int \tilde{f}(x')\, p^0(\mathrm{d}x' \mid x^0, u^0, \mu)\right| \to 0
\end{aligned}$$

for the first term by the prequel where $T_n^* := T(x_n^0, u_n^0, \mu_n, \nu_n) \to T^* := T(x^0, u^0, \mu, \nu)$ by Lemma H.1, and for the second term by applying Assumption 1 to $\tilde{f}(x') := f(x', T^*)$. This shows weak continuity of the dynamics.

Furthermore, the M3FC MDP fulfills (Hernández-Lerma and Lasserre, 2012), Assumption 4.2.2 by boundedness of r from Assumption 1. Therefore, the desired statement follows from (Hernández-Lerma and Lasserre, 2012), Theorem 4.2.3. □

I. Proof of Lemma H.1

Proof. To show $T(x_n^0, u_n^0, \mu_n, \nu_n) \to T(x^0, u^0, \mu, \nu)$, consider any Lipschitz and bounded f with Lipschitz constant L_f, then

$$\begin{aligned}
&\left|\int f \,\mathrm{d}(T(x_n^0, u_n^0, \mu_n, \nu_n) - T(x^0, u^0, \mu, \nu))\right| \\
&= \left|\iiint f(x')\left(p(\mathrm{d}x' \mid x, u, x_n^0, u_n^0, \mu_n)\nu_n(\mathrm{d}x, \mathrm{d}u) - p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu)\nu(\mathrm{d}x, \mathrm{d}u)\right)\right| \\
&\quad \leq \iint \left|\int f(x')p(\mathrm{d}x' \mid x, u, x_n^0, u_n^0, \mu_n) - \int f(x')p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu)\right| \nu_n(\mathrm{d}x, \mathrm{d}u) \\
&\qquad + \left|\iiint f(x')p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu)(\nu_n(\mathrm{d}x, \mathrm{d}u) - \nu(\mathrm{d}x, \mathrm{d}u))\right| \\
&\quad \leq \sup_{x\in\mathcal{X}, u\in\mathcal{U}} L_f W_1(p(\cdot \mid x, u, x_n^0, u_n^0, \mu_n), p(\cdot \mid x, u, x^0, u^0, \mu)) \\
&\qquad + \left|\iiint f(x')p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu)(\nu_n(\mathrm{d}x, \mathrm{d}u) - \nu(\mathrm{d}x, \mathrm{d}u))\right| \to 0
\end{aligned}$$

for the first term by 1-Lipschitzness of $\frac{f}{L_f}$ and Assumption 1 (with compactness implying the uniform continuity), and for the second by $\nu_n \to \nu$ and continuity of $(x, u) \mapsto \iint f(x')p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu)$ by the same argument. □

J. Proof of Theorem 2

Proof. The statement $\sup_{f,\pi,\pi^0} \left|\mathbb{E}\left[f(x_t^{0,N}, u_t^{0,N}, \mu_t^N) - f(x_t^0, u_t^0, \mu_t)\right]\right|$ is shown inductively over $t \geq 0$. At time $t = 0$, it holds by the weak LLN argument, see also the first term below. Assuming the statement at time t, then for time $t+1$ we have

$$\begin{aligned}
&\sup_{(\pi,\pi^0)\in\Pi\times\Pi^0} \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \mu_{t+1}^N) - f(x_{t+1}^0, u_{t+1}^0, \mu_{t+1})\right]\right| \\
&\quad \leq \sup_{\pi,\pi^0} \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \mu_{t+1}^N) - f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \hat{\mu}_{t+1}^N)\right]\right| \qquad (17) \\
&\qquad + \sup_{\pi,\pi^0} \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \hat{\mu}_{t+1}^N) - f(x_{t+1}^0, u_{t+1}^0, \mu_{t+1})\right]\right| \qquad (18)
\end{aligned}$$

where for readability, we again write $\pi_t(x_t^0, \mu_t) \coloneqq \pi_t(\cdot \mid \cdot, x_t^0, \mu_t)$ and introduce the random variable

$$\hat{\mu}_{t+1}^N \coloneqq T(x_t^{0,N}, u_t^{0,N}, \mu_t^N, \mu_t^N \otimes \pi_t(x_t^{0,N}, \mu_t^N)).$$

By compactness of $\mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X})$, $\mathcal{F}$ is uniformly equicontinuous, and hence admits a non-decreasing, concave (as in (DeVore and Lorentz, 1993), Lemma 6.1) modulus of continuity $\omega_{\mathcal{F}} \colon [0, \infty) \to [0, \infty)$ where $\omega_{\mathcal{F}}(x) \to 0$ as $x \to 0$ and $|f(x, u, \mu) - f(x', u', \nu)| \leq \omega_{\mathcal{F}}(d(x, x') + d(u, u') + W_1(\mu, \nu))$ for all $f \in \mathcal{F}$, and analogously there exists such $\tilde{\omega}_{\mathcal{F}}$ with respect to $(\mathcal{P}(\mathcal{X}), d_\Sigma)$ instead of $(\mathcal{P}(\mathcal{X}), W_1)$ as in Appendix E.

For the first term (17), let $x_t^N \equiv \{x_t^{i,N}\}_{i\in[N]}$. Then, by the weak LLN argument,

$$\begin{aligned}
&\sup_{\pi,\pi^0} \sup_{f\in\mathcal{F}} \left|\mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \mu_{t+1}^N) - f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \hat{\mu}_{t+1}^N)\right]\right| \\
&\quad \leq \sup_{\pi,\pi^0} \mathbb{E}\left[\tilde{\omega}_{\mathcal{F}}(d_\Sigma(\mu_{t+1}^N, \hat{\mu}_{t+1}^N))\right] \\
&\quad \leq \sup_{\pi,\pi^0} \tilde{\omega}_{\mathcal{F}}\left(\sum_{m=1}^{\infty} 2^{-m}\, \mathbb{E}\left[\left|\mu_{t+1}^N(f_m) - \hat{\mu}_{t+1}^N(f_m)\right|\right]\right) \\
&\quad \leq \sup_{\pi,\pi^0} \tilde{\omega}_{\mathcal{F}}\left(\sup_{m\geq 1} \mathbb{E}\left[\mathbb{E}_{\beta_t}\left[\left|\mu_{t+1}^N(f_m) - \hat{\mu}_{t+1}^N(f_m)\right|\right]\right]\right)
\end{aligned}$$

$$
\begin{aligned}
&= \sup_{\pi,\pi^0} \tilde{\omega}_{\mathcal{F}} \left(\sup_{m\geq 1} \mathbb{E}\left[\mathbb{E}_{\beta_t}\left[\left| \frac{1}{N}\sum_{i=1}^{N} \left(f_m(x_{t+1}^{i,N}) - \mathbb{E}_{\beta_t}\left[f_m(x_{t+1}^{i,N}) \right] \right) \right| \right] \right] \right) \\
&\leq \sup_{\pi,\pi^0} \tilde{\omega}_{\mathcal{F}} \left(\sup_{m\geq 1} \mathbb{E}\left[\mathbb{E}_{\beta_t}\left[\left| \frac{1}{N}\sum_{i=1}^{N} \left(f_m(x_{t+1}^{i,N}) - \mathbb{E}_{\beta_t}\left[f_m(x_{t+1}^{i,N}) \right] \right) \right|^2 \right] \right]^{1/2} \right) \\
&= \sup_{\pi,\pi^0} \tilde{\omega}_{\mathcal{F}} \left(\sup_{m\geq 1} \left(\frac{1}{N^2}\sum_{i=1}^{N} \mathbb{E}\left[\mathbb{E}_{\beta_t}\left[\left(f_m(x_{t+1}^{i,N}) - \mathbb{E}_{\beta_t}\left[f_m(x_{t+1}^{i,N}) \right] \right)^2 \right] \right] \right)^{1/2} \right) \\
&\leq \tilde{\omega}_{\mathcal{F}} \left(\frac{2}{\sqrt{N}} \right) \to 0 \qquad (19)
\end{aligned}
$$

for $\beta_t \coloneqq (x_t^{0,N}, u_t^{0,N}, x_t^N)$ by bounding $|f_m| \leq 1$, as the cross-terms disappear.

For the second term (18), by noting $\hat{\mu}_{t+1}^N = T(x_t^{0,N}, u_t^{0,N}, \mu_t^N, \mu_t^N \otimes \pi_t(x_t^{0,N}, \mu_t^N))$, we have

$$
\begin{aligned}
&\sup_{\pi,\pi^0} \sup_{f\in\mathcal{F}} \left| \mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \hat{\mu}_{t+1}^N) - f(x_{t+1}^0, u_{t+1}^0, \mu_{t+1}) \right] \right| \\
&= \sup_{\pi,\pi^0} \sup_{f\in\mathcal{F}} \Bigg| \mathbb{E}\Bigg[\iint f(x', u', \hat{\mu}_{t+1}^N) \pi_t^0(\mathrm{d}u' \mid x', \mu_{t+1}^N) p^0(\mathrm{d}x' \mid x_t^{0,N}, u_t^{0,N}, \mu_t^N) \\
&\qquad\qquad - \iint f(x', u', \mu_{t+1}) \pi_t^0(\mathrm{d}u' \mid x', \mu_{t+1}) p^0(\mathrm{d}x' \mid x_t^0, u_t^0, \mu_t) \Bigg] \Bigg| \\
&\leq \sup_{\pi,\pi^0} \sup_{f\in\mathcal{F}} \mathbb{E}\left[\sup_{x'} \left| \int f(x', u', \hat{\mu}_{t+1}^N) (\pi_t^0(\mathrm{d}u' \mid x', \mu_{t+1}^N) - \pi_t^0(\mathrm{d}u' \mid x', \hat{\mu}_{t+1}^N)) \right| \right] \qquad (20) \\
&\quad + \sup_{\pi,\pi^0} \sup_{g\in\mathcal{G}} \left| \mathbb{E}\left[g(x_t^{0,N}, u_t^{0,N}, \mu_t^N) - g(x_t^0, u_t^0, \mu_t) \right] \right| \qquad (21)
\end{aligned}
$$

and analyze each term separately, where we defined the function $g\colon \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X})$ as

$$
g(x^0, u^0, \mu) \coloneqq \iint f(x', u', T^*) \pi_t^0(\mathrm{d}u' \mid x', T^*) p^0(\mathrm{d}x' \mid x^0, u^0, \mu)
$$

from the class $\mathcal{G}$ of such functions for any policies π, π^0, where $T^* \coloneqq T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu))$.

For (20), defining a modulus of continuity $\tilde{\omega}_{\Pi^0}$ for Π^0 as for $\mathcal{F}$, we have

$$
\begin{aligned}
&\sup_{\pi,\pi^0} \sup_{f\in\mathcal{F}} \mathbb{E}\left[\sup_{x'} \left| \int f(x', u', \hat{\mu}_{t+1}^N) (\pi_t^0(\mathrm{d}u' \mid x', \mu_{t+1}^N) - \pi_t^0(\mathrm{d}u' \mid x', \hat{\mu}_{t+1}^N)) \right| \right] \\
&\leq \sup_{\pi,\pi^0} \mathbb{E}\left[L_{\mathcal{F}} \sup_{x'} W_1(\pi_t^0(\cdot \mid x', \mu_{t+1}^N), \pi_t^0(\cdot \mid x', \hat{\mu}_{t+1}^N)) \right] \\
&\leq \sup_{\pi,\pi^0} \mathbb{E}\left[L_{\mathcal{F}} \tilde{\omega}_{\Pi^0}(d_\Sigma(\mu_{t+1}^N, \hat{\mu}_{t+1}^N)) \right] \leq L_{\mathcal{F}} \tilde{\omega}_{\Pi^0}\left(\frac{2}{\sqrt{N}} \right) \to 0.
\end{aligned}
$$

Lastly, for (21), we first note that the class $\mathcal{G}$ of functions is equi-Lipschitz.

Lemma J.1. *Under Assumptions 1 and 2, the map $(x^0, u^0, \mu) \mapsto T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu))$ is Lipschitz with constant $L_T \coloneqq (2L_\Pi + 1) \cdot (L_p + (L_p + 1)L_\Pi + (L_p + L_\Pi + 1))$.*

Lemma J.2. *Under Assumptions 1 and 2, for any equi-Lipschitz $\mathcal{F}$ with constant $L_{\mathcal{F}}$, the function class $\mathcal{G}$ is equi-Lipschitz with constant $L_{\mathcal{G}} \coloneqq (L_{\mathcal{F}} L_T + L_{\mathcal{F}} L_{\Pi^0} L_T + L_{\mathcal{F}} L_\Pi L_{p^0})$.*

Therefore, for (21), we have

$$
\sup_{\pi,\pi^0} \sup_{g\in\mathcal{G}} \left| \mathbb{E}\left[g(x_t^{0,N}, u_t^{0,N}, \mu_t^N) - g(x_t^0, u_t^0, \mu_t) \right] \right| \to 0
$$

by the induction assumption over the class $\mathcal{G}$ of equi-Lipschitz functions, completing the proof by induction. The existence of independent optimal π, π^0 follows from Remark 3. This completes the proof.

For finite minor states, we can quantify the convergence rate more precisely as $\mathcal{O}(1/\sqrt{N})$, since the two metrizations d_Σ and W_1 are then Lipschitz equivalent and the above moduli of continuity simply become a multiplication with the Lipschitz constant, so for convenience we simply use the L_1 distance. The convergence in the first term (17) is immediate by the weak LLN

$$\begin{aligned}
&\sup_{\pi,\pi^0}\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(x^{0,N}_{t+1},u^{0,N}_{t+1},\mu^N_{t+1})-f(x^{0,N}_{t+1},u^{0,N}_{t+1},\hat{\mu}^N_{t+1})\right]\right|\\
&\quad\le\sup_{\pi,\pi^0}L_f\,\mathbb{E}\left[\sum_{x\in\mathcal{X}}\left|\mu^N_{t+1}(x)-\hat{\mu}^N_{t+1}(x)\right|\right]\\
&\quad=\sup_{\pi,\pi^0}L_f\sum_{x\in\mathcal{X}}\mathbb{E}\left[\mathbb{E}\left[\left|\frac{1}{N}\sum_{i=1}^N\mathbf{1}_x(x^{i,N}_{t+1})-\mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N\mathbf{1}_x(x^{i,N}_{t+1})\;\middle|\;x^{0,N}_t,u^{0,N}_t,\mu^N_t\right]\right|\;\middle|\;x^{0,N}_t,u^{0,N}_t,\mu^N_t\right]\right]\\
&\quad\le L_f|\mathcal{X}|\sqrt{\frac{4}{N}},
\end{aligned}$$

and for the second term (18) we again use the induction assumption, completing the proof. □

K. Proof of Lemma J.1

Proof. First note Lipschitz continuity of $(x^0,\mu)\mapsto\mu\otimes\pi_t(x^0,\mu)$ as in Appendix E, as for any $(x^0_*,\mu_*),(x^0,\mu)\in\mathcal{X}^0\times\mathcal{P}(\mathcal{X})$, then

$$\begin{aligned}
&\sup_{\pi\in\Pi}W_1(\mu_*\otimes\pi_t(x^0_*,\mu_*),\mu\otimes\pi_t(x^0,\mu))\\
&\quad=\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le1}\left|\int f'\,\mathrm{d}(\mu_*\otimes\pi_t(x^0_*,\mu_*)-\mu\otimes\pi_t(x^0,\mu))\right|\\
&\quad\le\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le1}\left|\iint f'(x,u)(\pi_t(\mathrm{d}u\mid x,x^0_*,\mu_*)-\pi_t(\mathrm{d}u\mid x,x^0,\mu))\mu_*(\mathrm{d}x)\right|\\
&\qquad+\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le1}\left|\iint f'(x,u)\pi_t(\mathrm{d}u\mid x,x^0,\mu)(\mu_*(\mathrm{d}x)-\mu(\mathrm{d}x))\right|
\end{aligned}$$

where for the first term

$$\begin{aligned}
&\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le1}\left|\iint f'(x,u)(\pi_t(\mathrm{d}u\mid x,x^0_*,\mu_*)-\pi_t(\mathrm{d}u\mid x,x^0,\mu))\mu_*(\mathrm{d}x)\right|\\
&\quad\le\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le1}\int\left|\int f'(x,u)(\pi_t(\mathrm{d}u\mid x,x^0_*,\mu_*)-\pi_t(\mathrm{d}u\mid x,x^0,\mu))\right|\mu_*(\mathrm{d}x)\\
&\quad\le\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le1}\sup_{x\in\mathcal{X}}\left|\int f'(x,u)(\pi_t(\mathrm{d}u\mid x,x^0_*,\mu_*)-\pi_t(\mathrm{d}u\mid x,x^0,\mu))\right|\\
&\quad=\sup_{\pi\in\Pi}\sup_{x\in\mathcal{X}}W_1(\pi_t(\cdot\mid x,x^0_*,\mu_*),\pi_t(\cdot\mid x,x^0,\mu))\\
&\quad\le L_\Pi d((x^0_*,\mu_*),(x^0,\mu))
\end{aligned}$$

by Assumption 2, and similarly for the second by noting 1-Lipschitzness of $x\mapsto\int\frac{f'(x,u)}{L_\Pi+1}\pi_t(\mathrm{d}u\mid x,x^0,\mu)$, as before in (11), and therefore again

$$\begin{aligned}
&\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le1}\left|\iint f'(x,u)\pi_t(\mathrm{d}u\mid x,x^0,\mu)(\mu_*(\mathrm{d}x)-\mu(\mathrm{d}x))\right|\\
&\quad=\sup_{\pi\in\Pi}\sup_{\|f'\|_{\mathrm{Lip}}\le1}(L_\Pi+1)\left|\iint\frac{f'(x,u)}{L_\Pi+1}\pi_t(\mathrm{d}u\mid x,x^0,\mu)(\mu_*(\mathrm{d}x)-\mu(\mathrm{d}x))\right|
\end{aligned}$$

$$\leq (L_\Pi + 1) W_1(\mu_*, \mu).$$

Hence, the map $(x^0, u^0, \mu) \mapsto \mu \otimes \pi_t(x^0, \mu)$ is Lipschitz with constant $(2L_\Pi + 1)$.

As a result, the entire map $(x^0, u^0, \mu) \mapsto T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu)$ is Lipschitz, since for any

$$\begin{aligned}
&W_1(T(x_*^0, u_*^0, \mu_*, \mu_* \otimes \pi_t(x_*^0, \mu_*)), T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu)) \\
&= \sup_{\|f'\|_{\mathrm{Lip}} \leq 1} \Bigg| \iiint f'(x') p(\mathrm{d}x' \mid x, u, x_*^0, u_*^0, \mu_*) \pi_t(\mathrm{d}u \mid x, x_*^0, \mu_*) \mu_*(\mathrm{d}x) \\
&\qquad\qquad - \iiint f'(x') p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu) \pi_t(\mathrm{d}u \mid x, x^0, \mu) \mu(\mathrm{d}x) \Bigg| \\
&\leq \sup_{\|f'\|_{\mathrm{Lip}} \leq 1} \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} \left| \int f'(x') (p(\mathrm{d}x' \mid x, u, x_*^0, u_*^0, \mu_*) - p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu)) \right| \\
&\quad + \sup_{\|f'\|_{\mathrm{Lip}} \leq 1} \sup_{x \in \mathcal{X}} \left| \iint f'(x') p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu) (\pi_t(\mathrm{d}u \mid x, x_*^0, \mu_*) - \pi_t(\mathrm{d}u \mid x, x^0, \mu)) \right| \\
&\quad + \sup_{\|f'\|_{\mathrm{Lip}} \leq 1} \left| \iiint f'(x') p(\mathrm{d}x' \mid x, u, x^0, u^0, \mu) \pi_t(\mathrm{d}u \mid x, x^0, \mu) (\mu_*(\mathrm{d}x) - \mu(\mathrm{d}x)) \right| \\
&\leq \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} W_1(p(\cdot \mid x, u, x_*^0, u_*^0, \mu_*), p(\cdot \mid x, u, x^0, u^0, \mu)) \\
&\quad + \sup_{x \in \mathcal{X}} (L_p + 1) W_1(\pi_t(\cdot \mid x, x_*^0, \mu_*), \pi_t(\cdot \mid x, x^0, \mu)) \\
&\quad + \sup_{(x,u) \in \mathcal{X} \times \mathcal{U}} (L_p + L_\Pi + 1) W_1(\mu_*, \mu) \\
&\leq \underbrace{(L_p + (L_p + 1) L_\Pi + (L_p + L_\Pi + 1))}_{L_*} d((x_*^0, u_*^0, \mu_*), (x^0, u^0, \mu))
\end{aligned}$$

with Lipschitz constant $L_T \coloneqq (2L_\Pi + 1) \cdot L_*$ from Assumptions 1 and 2, using the same argument as in (11). □

L. Proof of Lemma J.2

Proof. For any $g \in \mathcal{G}$, for any $(x_*^0, u_*^0, \mu_*), (x^0, u^0, \mu) \in \mathcal{X}^0 \times \mathcal{U}^0 \times \mathcal{P}(\mathcal{X})$, let $T_* \coloneqq T(x_*^0, u_*^0, \mu_*, \mu_* \otimes \pi_t(x_*^0, \mu_*))$ and $T^* \coloneqq T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu))$ for brevity. We have

$$\begin{aligned}
&\left| g(x_*^0, u_*^0, \mu_*) - g(x^0, u^0, \mu) \right| \\
&= \Bigg| \iint f(x', u', T_*) \pi_t^0(\mathrm{d}u' \mid x', T_*) p^0(\mathrm{d}x' \mid x_*^0, u_*^0, \mu_*) \\
&\qquad\qquad - \iint f(x', u', T^*) \pi_t^0(\mathrm{d}u' \mid x', T^*) p^0(\mathrm{d}x' \mid x^0, u^0, \mu) \Bigg| \\
&\leq \sup_{x', u'} |f(x', u', T_*) - f(x', u', T^*)| && (22) \\
&\quad + \sup_{x'} \left| \int f(x', u', T^*) (\pi_t^0(\mathrm{d}u' \mid x', T_*) - \pi_t^0(\mathrm{d}u' \mid x', T^*)) \right| && (23) \\
&\quad + \left| \iint f(x', u', T^*) \pi_t^0(\mathrm{d}u' \mid x', T^*) (p^0(\mathrm{d}x' \mid x_*^0, u_*^0, \mu_*) - p^0(\mathrm{d}x' \mid x^0, u^0, \mu)) \right|. && (24)
\end{aligned}$$

By Lemma J.1, for (22) we obtain

$$\begin{aligned}
&\sup_{x', u'} \left| f(x', u', T(x_*^0, u_*^0, \mu_*, \mu_* \otimes \pi_t(x_*^0, \mu_*))) - f(x', u', T(x^0, u^0, \mu, \mu \otimes \pi_t(x^0, \mu))) \right| \\
&\quad \leq L_{\mathcal{F}} L_T d((x_*^0, u_*^0, \mu_*), (x^0, u^0, \mu)).
\end{aligned}$$

Similarly for (23), by Assumption 2 we analogously have

$$\begin{aligned}
&\sup_{x'}\Bigg|\int f(x',u',T(x^0,u^0,\mu,\mu\otimes\pi_t(x^0,\mu)))\\
&\quad(\pi_t^0(\mathrm{d}u'\mid x',T(x_*^0,u_*^0,\mu_*,\mu_*\otimes\pi_t(x_*^0,\mu_*)))-\pi_t^0(\mathrm{d}u'\mid x',T(x^0,u^0,\mu,\mu\otimes\pi_t(x^0,\mu))))\Big|\\
&\le L_{\mathcal{F}}W_1(\pi_t^0(\cdot\mid x',T(x_*^0,u_*^0,\mu_*,\mu_*\otimes\pi_t(x_*^0,\mu_*))),\pi_t^0(\cdot'\mid x',T(x^0,u^0,\mu,\mu\otimes\pi_t(x^0,\mu)))\\
&\le L_{\mathcal{F}}L_{\Pi^0}L_T d((x_*^0,u_*^0,\mu_*),(x^0,u^0,\mu)).
\end{aligned}$$

Lastly, for (24), as before in (11), by Assumption 1 and 2 we have again

$$\begin{aligned}
&\Bigg|\iint f(x',u',T(x^0,u^0,\mu,\mu\otimes\pi_t(x^0,\mu)))\pi_t^0(\mathrm{d}u'\mid x',T(x^0,u^0,\mu,\mu\otimes\pi_t(x^0,\mu)))\\
&\qquad\qquad(p^0(\mathrm{d}x'\mid x_*^0,u_*^0,\mu_*)-p^0(\mathrm{d}x'\mid x^0,u^0,\mu))\Big|\\
&\le L_{\mathcal{F}}L_{\Pi}W_1(p^0(\cdot\mid x_*^0,u_*^0,\mu_*),p^0(\cdot\mid x^0,u^0,\mu))\\
&\le L_{\mathcal{F}}L_{\Pi}L_{p^0}d((x_*^0,u_*^0,\mu_*),(x^0,u^0,\mu)).
\end{aligned}$$

Therefore, $\mathcal{G}$ is equi-Lipschitz with Lipschitz constant $(L_{\mathcal{F}}L_T+L_{\mathcal{F}}L_{\Pi^0}L_T+L_{\mathcal{F}}L_{\Pi}L_{p^0})$. □

M. Proof of Corollary 1

Proof. As in Lemma F.1, for any $\varepsilon>0$, choose time $T\in\mathbb{N}$ such that

$$\sum_{t=T}^{\infty}\gamma^t\left|\mathbb{E}\left[r(x_t^{0,N},u_t^{0,N},\mu_t^N)-r(x_t^0,u_t^0,\mu_t)\right]\right|\le\frac{\gamma^T}{1-\gamma}\max_{\mu}2|r(\mu)|<\frac{\varepsilon}{2}.$$

By Theorem 2,

$$\sum_{t=0}^{T-1}\gamma^t\left|\mathbb{E}\left[r(x_t^{0,N},u_t^{0,N},\mu_t^N)-r(x_t^0,u_t^0,\mu_t)\right]\right|<\frac{\varepsilon}{2}$$

for sufficiently large N. Therefore, $\sup_{(\pi,\pi^0)\in\Pi\times\Pi^0}\left|J^N(\pi,\pi^0)-J(\Phi^{-1}(\pi),\pi^0)\right|\to 0$.

As a result, we have

$$\begin{aligned}
J^N(\Phi(\hat\pi^*),\pi^{0*})-\sup_{(\pi,\pi^0)\in\Pi\times\Pi^0}J^N(\pi,\pi^0)&=\inf_{(\pi,\pi^0)\in\Pi\times\Pi^0}(J^N(\Phi(\hat\pi^*),\pi^{0*})-J^N(\pi,\pi^0))\\
&\ge\inf_{(\pi,\pi^0)\in\Pi\times\Pi^0}(J^N(\Phi(\hat\pi^*),\pi^{0*})-J(\hat\pi^*,\pi^{0*}))\\
&\quad+\inf_{(\pi,\pi^0)\in\Pi\times\Pi^0}(J(\hat\pi^*,\pi^{0*})-J(\pi,\pi^0))\\
&\quad+\inf_{(\pi,\pi^0)\in\Pi\times\Pi^0}(J(\pi,\pi^0)-J^N(\pi,\pi^0))\\
&\ge-\frac{\varepsilon}{2}+0-\frac{\varepsilon}{2}=-\varepsilon
\end{aligned}$$

for sufficiently large N, where the second term is zero by optimality of $(\hat\pi^*,\pi^{0*})$ in the M3FC problem. □

N. Proof of Theorem 3

First, for completeness we give the finite M3FC system equations under the assumed Lipschitz parametrization for joint stationary M3FMARL policies[1] $\tilde\pi^\theta$ used during centralized training with correlated minor agent actions, as

$$u_t^{0,N},\xi_t^N\sim\tilde\pi^\theta(u_t^{0,N},\xi_t^N\mid x_t^{0,N},\mu_t^N),\quad \pi_t'^N=\Gamma(\xi_t^N),\quad u_t^{i,N}\sim\pi_t'^N(u_t^{i,N}\mid x_t^{i,N}),$$

[1]Note that deterministic joint policies $\tilde\pi^\theta$ (e.g. at convergence, or if using deterministic policy gradients (Silver et al., 2014)) are equivalent to using separate deterministic minor and major policies in (1), see also Remark 3.

$$x_{t+1}^{i,N} \sim p(x_{t+1}^{i,N} \mid x_t^{i,N}, u_t^{i,N}, x_t^{0,N}, u_t^{0,N}, \mu_t^N), \quad x_{t+1}^{0,N} \sim p^0(x_{t+1}^{0,N} \mid x_t^{0,N}, u_t^{0,N}, \mu_t^N),$$

as well as the limiting M3FC MDP under such parametrization as

$$u_t^0, \xi_t \sim \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t), \quad \pi_t' = \Gamma(\xi_t), \quad h_t = \mu_t \otimes \pi_t',$$
$$\mu_{t+1} = T(x_t^0, u_t^0, \mu_t, h_t), \quad x_{t+1}^0 \sim p^0(x_{t+1}^0 \mid x_t^0, u_t^0, \mu_t).$$

Then, by Sutton et al. (1999), the exact policy gradient for the limiting M3FC MDP is given as

$$\nabla_\theta J(\tilde{\pi}^\theta) = \sum_{t=T}^{\infty} \gamma^t \, \mathbb{E}\left[Q^\theta(x_t^0, \mu_t, u_t^0, \xi_t) \nabla_\theta \log \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t)\right]$$

under the action-value function

$$Q^\theta(x^0, \mu, u^0, \xi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^0, u_t^0, \mu_t) \,\middle|\, x_0^0 = x^0, \mu_0 = \mu, u_0^0 = u^0, \xi_0 = \xi\right],$$

while the approximation for the policy gradient on the finite M3FC system is given instead by

$$\widehat{\nabla_\theta J}(\tilde{\pi}^\theta) = \sum_{t=T}^{\infty} \gamma^t \, \mathbb{E}\left[\widehat{Q}^\theta(x_t^{0,N}, \mu_t^N, u_t^{0,N}, \xi_t^N) \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \,\middle|\, x_t^{0,N}, \mu_t^N)\right]$$

and the finite-agent action-values

$$\widehat{Q}^\theta(x^0, \mu, u^0, \xi) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(x_t^{0,N}, u_t^{0,N}, \mu_t^N) \,\middle|\, x_0^{0,N} = x^0, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi\right],$$

which are obtained, e.g., by on-policy samples and using critic estimates. Note that here, the conditional expectations are given by redefining the systems (1) and (3) with the values conditioned upon.

We then show that the approximation of the policy gradient is good for large systems, i.e.

$$\left\| \widehat{\nabla_\theta J}(\tilde{\pi}^\theta) - \nabla_\theta J(\hat{\pi}^\theta) \right\| \to 0 \tag{25}$$

as $N \to \infty$, uniformly over all current policy parameters θ.

Proof of Theorem 3. We use the following lemmas in the proof of Theorem 3, for which the proofs are given below.

Proposition N.1. *Propagation of chaos holds for the M3FC systems with parameterized actions as in Theorem 2, i.e. under Assumptions 1, 2 and 3, for any equi-Lipschitz family $\mathcal{F}$, at all times $t \in \mathbb{N}$ uniformly,*

$$\sup_{f,\pi,\pi^0} \left| \mathbb{E}\left[f(x_t^{0,N}, u_t^{0,N}, \mu_t^N) - f(x_t^0, u_t^0, \mu_t)\right] \right| \to 0. \tag{26}$$

Proposition N.2. *Under Assumptions 1 and 2, the approximate action-values converge uniformly, $\widehat{Q}^\theta \to Q^\theta$ as $N \to \infty$.*

As a result, we obtain

$$\begin{aligned}
&\left\| \widehat{\nabla_\theta J}(\tilde{\pi}^\theta) - \nabla_\theta J(\hat{\pi}^\theta) \right\| \\
&= \left\| \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}\left[\widehat{Q}^\theta(x_t^{0,N}, \mu_t^N, u_t^{0,N}, \xi_t^N) \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \mid x_t^{0,N}, \mu_t^N) - Q^\theta(x_t^0, \mu_t, u_t^0, \xi_t) \nabla_\theta \log \tilde{\pi}^\theta(u_t^0, \xi_t \mid x_t^0, \mu_t)\right] \right\| \\
&\leq \left\| \sum_{t=0}^{\infty} \gamma^t \, \mathbb{E}\left[\left(\widehat{Q}^\theta(x_t^{0,N}, \mu_t^N, u_t^{0,N}, \xi_t^N) - Q^\theta(x_t^{0,N}, \mu_t^N, u_t^{0,N}, \xi_t^N)\right) \nabla_\theta \log \tilde{\pi}^\theta(u_t^{0,N}, \xi_t^N \mid x_t^{0,N}, \mu_t^N)\right] \right\|
\end{aligned}$$

$$+\left\|\sum_{t=T}^{\infty}\gamma^t\,\mathbb{E}\left[Q^\theta(x_t^{0,N},\mu_t^N,u_t^{0,N},\xi_t^N)\nabla_\theta\log\tilde{\pi}^\theta(u_t^{0,N},\xi_t^N\mid x_t^{0,N},\mu_t^N)-Q^\theta(x_t^0,\mu_t,u_t^0,\xi_t)\nabla_\theta\log\tilde{\pi}^\theta(u_t^0,\xi_t\mid x_t^0,\mu_t)\right]\right\|$$
$$+\left\|\sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}\left[Q^\theta(x_t^{0,N},\mu_t^N,u_t^{0,N},\xi_t^N)\nabla_\theta\log\tilde{\pi}^\theta(u_t^{0,N},\xi_t^N\mid x_t^{0,N},\mu_t^N)-Q^\theta(x_t^0,\mu_t,u_t^0,\xi_t)\nabla_\theta\log\tilde{\pi}^\theta(u_t^0,\xi_t\mid x_t^0,\mu_t)\right]\right\|$$

for any T, such that the first term disappears by Assumption 3 uniformly bounding $\nabla_\theta\log\tilde{\pi}^\theta$ and Proposition N.2. Note that we bounded $\nabla_\theta\log\tilde{\pi}^\theta$ here, but we can also assume bounded gradients $\nabla_\theta\tilde{\pi}^\theta$ instead, e.g. (27).

For the second term, we similarly uniformly bound $\nabla_\theta\log\tilde{\pi}^\theta$ by Assumption 3 and Q by Assumption 1, then choose T sufficiently large.

Finally, for the last term, we note that we can write the difference as

$$\begin{aligned}
&\left\|\sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}\left[Q^\theta(x_t^{0,N},\mu_t^N,u_t^{0,N},\xi_t^N)\nabla_\theta\log\tilde{\pi}^\theta(u_t^{0,N},\xi_t^N\mid x_t^{0,N},\mu_t^N)-Q^\theta(x_t^0,\mu_t,u_t^0,\xi_t)\nabla_\theta\log\tilde{\pi}^\theta(u_t^0,\xi_t\mid x_t^0,\mu_t)\right]\right\|\\
&=\left\|\sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}\left[\sum_{t'=0}^{\infty}\gamma^t\,\mathbb{E}\left[r(x_{t'}^{0\prime},u_{t'}^{0\prime},\mu_{t'}^{\prime})\;\middle|\;x_0^{0\prime}=x_t^{0,N},\mu_0^{\prime}=\mu_t^N,u_0^{0\prime}=u_t^{0,N},\xi_0^{\prime}=\xi_t^N\right]\nabla_\theta\log\tilde{\pi}^\theta(u_t^{0,N},\xi_t^N\mid x_t^{0,N},\mu_t^N)\right.\right.\\
&\qquad\left.\left.-\sum_{t=0}^{\infty}\gamma^t\,\mathbb{E}\left[r(x_{t'}^{0\prime},u_{t'}^{0\prime},\mu_{t'}^{\prime})\mid x_0^{0\prime}=x_t^0,\mu_0^{\prime}=\mu_t,u_0^{0\prime}=u_t^0,\xi_0^{\prime}=\xi_t\right]\nabla_\theta\log\tilde{\pi}^\theta(u_t^0,\xi_t\mid x_t^0,\mu_t)\right]\right\|\\
&\le\left\|\sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}\left[\sum_{t'=T'}^{\infty}\gamma^t\,\mathbb{E}\left[r(x_{t'}^{0\prime},u_{t'}^{0\prime},\mu_{t'}^{\prime})\;\middle|\;x_0^{0\prime}=x_t^{0,N},\mu_0^{\prime}=\mu_t^N,u_0^{0\prime}=u_t^{0,N},\xi_0^{\prime}=\xi_t^N\right]\nabla_\theta\log\tilde{\pi}^\theta(u_t^{0,N},\xi_t^N\mid x_t^{0,N},\mu_t^N)\right.\right.\\
&\qquad\left.\left.-\sum_{t=T'}^{\infty}\gamma^t\,\mathbb{E}\left[r(x_{t'}^{0\prime},u_{t'}^{0\prime},\mu_{t'}^{\prime})\mid x_0^{0\prime}=x_t^0,\mu_0^{\prime}=\mu_t,u_0^{0\prime}=u_t^0,\xi_0^{\prime}=\xi_t\right]\nabla_\theta\log\tilde{\pi}^\theta(u_t^0,\xi_t\mid x_t^0,\mu_t)\right]\right\|\\
&+\left\|\sum_{t=0}^{T-1}\gamma^t\,\mathbb{E}\left[\sum_{t'=0}^{T'-1}\gamma^t\,\mathbb{E}\left[r(x_{t'}^{0\prime},u_{t'}^{0\prime},\mu_{t'}^{\prime})\;\middle|\;x_0^{0\prime}=x_t^{0,N},\mu_0^{\prime}=\mu_t^N,u_0^{0\prime}=u_t^{0,N},\xi_0^{\prime}=\xi_t^N\right]\nabla_\theta\log\tilde{\pi}^\theta(u_t^{0,N},\xi_t^N\mid x_t^{0,N},\mu_t^N)\right.\right.\\
&\qquad\left.\left.-\sum_{t=0}^{T'-1}\gamma^t\,\mathbb{E}\left[r(x_{t'}^{0\prime},u_{t'}^{0\prime},\mu_{t'}^{\prime})\mid x_0^{0\prime}=x_t^0,\mu_0^{\prime}=\mu_t,u_0^{0\prime}=u_t^0,\xi_0^{\prime}=\xi_t\right]\nabla_\theta\log\tilde{\pi}^\theta(u_t^0,\xi_t\mid x_t^0,\mu_t)\right]\right\|
\end{aligned}$$

where we write the conditional M3FC system and random variables in the inner expectation with a prime, bounding again the former terms by choosing sufficiently large T' and using Assumptions 1 and 3, while for the latter terms we use Proposition N.1 on the functions

$$f(x^0,\mu)=\iint\mathbb{E}\left[r(x_{t'}^{0\prime},u_{t'}^{0\prime},\mu_{t'}^{\prime})\mid x_0^{0\prime}=x^0,\mu_0=\mu,u_0^{0\prime}=u^0,\xi_0^{\prime}=\xi\right]\nabla_\theta\tilde{\pi}^\theta(u^0,\xi\mid x^0,\mu)\mathrm{d}(u^0,\xi)\tag{27}$$

for all t', which are uniformly Lipschitz by Assumptions 1 and 3. This completes the proof. □

O. Proof of Proposition N.1

Proof. The proof is exactly analogous to the proof of Theorem 2, except that instead of using Lipschitz constants of $x_t^0,u_t^0,\mu_t,h_t\mapsto T(x_t^0,u_t^0,\mu_t,h_t)$, one uses Lipschitz constants of $x_t^0,u_t^0,\mu_t,\xi_t\mapsto T(x_t^0,u_t^0,\mu_t,\mu_t\otimes\Gamma(\xi_t))$ via the additional Assumption 3 on top of Assumptions 1 and 2. □

P. Proof of Proposition N.2

Proof. To show $\widehat{Q}^\theta\to Q^\theta$ as $N\to\infty$ uniformly, it suffices to prove pointwise convergence due to compact support.

Therefore, fix any x^0,μ,u^0,ξ. The convergence follows as in Corollary 1, from showing at any time t that

$$\sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[f(x_t^0,u_t^0,\mu_t)\mid x_0^0=x^0,\mu_0=\mu,u_0^0=u^0,\xi_0=\xi\right]\right.$$

$$- \mathbb{E}\left[f(x_t^{0,N}, u_t^{0,N}, \mu_t^N) \;\middle|\; x_0^{0,N} = x^{0,N}, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi \right] \Big| \to 0$$

over any equi-Lipschitz family of functions $\mathcal{F}$, and applying for $f = r$ (using the set $\mathcal{F}$ of L_r-Lipschitz functions) by Assumption 1.

The statement is shown by considering time $t = 0$, and then by induction for any $t \geq 1$. At time $t = 0$, the statement follows from the weak LLN as in Theorem 2. For any subsequent times, we similarly have

$$\begin{aligned}
&\sup_{f \in \mathcal{F}} \Big| \mathbb{E}\left[f(x_{t+1}^0, u_{t+1}^0, \mu_{t+1}) \mid x_0^0 = x^0, \mu_0 = \mu, u_0^0 = u^0, \xi_0 = \xi \right] \\
&\qquad - \mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \mu_{t+1}^N) \;\middle|\; x_0^{0,N} = x^{0,N}, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi \right] \Big| \\
&\leq \sup_{f \in \mathcal{F}} \Big| \mathbb{E}\left[f(x_{t+1}^0, u_{t+1}^0, \mu_{t+1}) \mid x_0^0 = x^0, \mu_0 = \mu, u_0^0 = u^0, \xi_0 = \xi \right] \\
&\qquad - \mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, T(x_t^{0,N}, u_t^{0,N}, \mu_t^N, \mu_t^N \otimes \Gamma(\xi_t^N))) \;\middle|\; x_0^{0,N} = x^{0,N}, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi \right] \Big| \\
&\quad + \sup_{f \in \mathcal{F}} \Big| \mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, T(x_t^{0,N}, u_t^{0,N}, \mu_t^N, \mu_t^N \otimes \Gamma(\xi_t^N))) \;\middle|\; x_0^{0,N} = x^{0,N}, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi \right] \\
&\qquad - \mathbb{E}\left[f(x_{t+1}^{0,N}, u_{t+1}^{0,N}, \mu_{t+1}^N) \;\middle|\; x_0^{0,N} = x^{0,N}, \mu_0 = \mu, u_0^{0,N} = u^0, \xi_0^N = \xi \right] \Big| .
\end{aligned}$$

As in Theorem 2, the latter term is bounded by induction assumption, using uniform Lipschitzness of the dynamics, $x_t^0, u_t^0, \mu_t, \xi_t \mapsto T(x_t^0, u_t^0, \mu_t, \mu_t \otimes \Gamma(\xi_t))$ via Assumptions 2 and 3, while the former term is bounded as usual by the weak LLN. This completes the proof. □

Q. Extended MFC Optimalities

Intuitively, in large MF systems governed by dynamics of the form (1), almost all information of the joint state $(x_t^{0,N}, x_t^{1,N}, \ldots, x_t^{N,N})$ is contained in $(x_t^{0,N}, \mu_t^N)$, while heterogeneous policies should by LLN be replaceable by a shared one. To fully complete the theory of MFC, it is therefore interesting to establish the optimality of the considered MF policies over arbitrary other policies acting on the joint state $(x_t^{0,N}, x_t^{1,N}, \ldots, x_t^{N,N})$.

It seems plausible that it would be possible to extend optimality (Corollary 1) over larger classes of policies in the finite system. In particular, at least for finite state-action spaces, (i) any joint-state policy $\pi(\mathrm{d}u \mid x_t^{0,N}, x_t^{1,N}, \ldots, x_t^{N,N})$ might in the limit be replaced by an averaged policy $\bar{\pi}(\mathrm{d}u \mid x^0, \mu) := \sum_{x^N \in \mathcal{X}^N : \frac{1}{N}\sum_i \delta_{x^{i,N}} = \mu} \pi(\mathrm{d}u \mid x^0, x^N)$ under some exchangeability of agents; (ii) any optimal policy π outputting joint actions for all agents might be replaced by an independent but identical policy for each agent, as in the limit all information is contained in the joint state-action distribution, any of which may be approximated increasingly closely by LLN; and (iii) heterogeneous policies for each minor agent $\pi^1, \ldots, \pi^N$ might similarly be replaced by some averaged policy $\bar{\pi}(\pi^1, \ldots, \pi^N)$, averaging the action distributions in any specific state over the proportion of agent likelihoods in that state.

Showing such results would allow us to conclude that the policy classes Π are natural and sufficient in MF systems, including MFC and also the competitive MFGs, as more general or heterogeneous policies will not perform much better. A result related to (iii) has been shown for static cases (Sanjari and Yüksel, 2020; Cui et al., 2021) and more recently in MFC and its two-team generalizations (Guan et al., 2024).

R. Experimental Details

Here, we give details omitted in the main text. For code, see `https://github.com/tudkcui/M3FC-MARL`.

R.1. Problem Details

In this section, we give details to the problems considered in this work. We omit the superscript N for readability.

2G. In the 2G problem, we formally let $\mathcal{X} = [-2, 2]^2, \mathcal{U} = [-1, 1]^2, \mathcal{X}^0 = \{0, 1, \ldots 49\}$ according to (13). We allow noisy movement of minor agents following the Gaussian law

$$p(x_{t+1}^i \mid x_t^i, u_t^i) = \mathcal{N}\left(x_{t+1}^i \;\middle|\; x_t^i + v_{\max} \frac{u_t^i}{\max(1, \|u_t^i\|_2)}, \mathrm{diag}(\sigma^2, \sigma^2) \right)$$

Table 3. Shared hyperparameter configurations for all algorithms.

Symbol	Name	Value
γ	Discount factor	0.99
λ	GAE lambda	1
β	KL coefficient	0.03
ϵ	Clip parameter	0.2
l_r	Learning rate	0.00005
B_{len}	Training batch size	24000
b_{len}	Mini-batch size	4000
N_{SGD}	Gradient steps per training batch	8

for some maximum speed $v_{\max} = 0.2$, noise covariance $\sigma^2 = 0.03$ and projecting back actions u with norm larger than 1, with the additional modification that agent positions are clipped back into $\mathcal{X}$ whenever the agents move out of bounds.

We then consider a time-variant mixture of two Gaussians

$$\mu_t^* := \frac{1+\cos(2\pi t/50)}{2}\mathcal{N}\left(\mathbf{e}_1, \text{diag}(\sigma_*^2, \sigma_*^2)\right) + \frac{1-\cos(2\pi t/50)}{2}\mathcal{N}\left(-\mathbf{e}_1, \text{diag}(\sigma_*^2, \sigma_*^2)\right)$$

for unit vector $\mathbf{e}_1$ and covariance $\sigma_*^2 = 0.05$, i.e. we have a period of 50 time steps, and let the major state follow the clock dynamics $p^0(x^0 + 1 \mod 50 \mid x^0, \mu) = 1$.

The goal of minor agents is to minimize the Wasserstein metric $\hat{W}_1$ under the squared Euclidean distance,

$$\hat{W}_1(\mu, \mu') := \inf_{\gamma \in \Gamma(\mu, \mu')} \left\{ \int \|x - y\|_2^2 \gamma(\mathrm{d}x, \mathrm{d}y) \right\}$$

defined over all couplings $\Gamma(\mu, \mu')$ with first and second marginals μ, μ' (which is strictly speaking not a metric but an optimal transportation cost, since the squared Euclidean distance fails the triangle inequality), between their empirical distribution and the desired mixture of Gaussians

$$r(x_t^0, \mu_t) = -\hat{W}_1(\mu_t, \mu_t^*)$$

which is computed numerically by the empirical distance, sampling 300 samples from μ_t^*.

The initialization of minor agents is uniform, i.e. $\mu_0 = \text{Unif}(\mathcal{X})$, and $x_0^0 = 0$. For sake of simulation, we define the episode length $T = 100$ after which a new episode starts.

Formation. The Formation problem is an extension of the 2G problem, where instead $\mathcal{X}^0 = \mathcal{X} \times \mathcal{X}$ and $\mathcal{U}^0 = \mathcal{U}$, the major agent follows the same dynamics as the minor agents, and movements are noise-free, i.e. $\sigma^2 = 0$. The major agent state $x_t^0 = (\hat{x}_t^0, x_t^*)$ here contains both the major agent position $\hat{x}_t^0$ and its target position x_t^*. The desired minor agent distribution is centered around the major agent

$$\mu_t^* := \mathcal{N}\left(\hat{x}_t^0, \text{diag}(\sigma_*^2, \sigma_*^2)\right)$$

with covariance $\sigma_*^2 = 0.3$, and is also observed by agents as in 2G via binning. Additionally, the major agent should follow a random target x_t^* following discretized Ornstein-Uhlenbeck dynamics

$$x_{t+1}^* \sim \mathcal{N}\left(0.95 x_t^*, \text{diag}(\sigma_{\text{targ}}^2, \sigma_{\text{targ}}^2)\right)$$

with $\sigma_{\text{targ}}^2 = 0.02$. Thus, similar to 2G, the reward function becomes

$$r(x_t^0, u_t^0, \mu_t) = -\|\hat{x}_t^0 - x_t^*\|_2 - \hat{W}_1(\mu_t, \mu_t^*).$$

The initialization of agents is uniform, while the target starts around zero, i.e. $\mu_0 = \text{Unif}(\mathcal{X})$ and $\mu_0^0 = \text{Unif}(\mathcal{X}) \otimes \mathcal{N}\left(0, \text{diag}(\sigma_{\text{targ}}^2, \sigma_{\text{targ}}^2)\right)$. For sake of simulation, we define the episode length $T = 100$ after which a new episode starts.

Beach Bar Process. In the discrete beach bar process, we consider a discrete torus $\mathcal{X} = \{0, 1, \ldots, 4\}^2$, $\mathcal{X}^0 = \mathcal{X} \times \mathcal{X}$ and actions $\mathcal{U} = \mathcal{U}^0 = \{(0,0), (-1,0), (0,-1), (1,0), (0,1)\}$ indicating movement in any of the four cardinal directions. The major agent state $x_t^0 = (\hat{x}_t^0, x_t^*)$ here contains both the major agent position $\hat{x}_t^0$ and its target position x_t^*. In other words, the dynamics follow

$$\hat{x}_{t+1}^0 = \hat{x}_t^0 + u_t^0 \mod (5,5), \quad x_{t+1}^i = x_t^i + u_t^i \mod (5,5).$$

The target position follows a random walk on the torus

$$x_{t+1}^* \sim x_t^* + \epsilon_t \mathrm{Unif}((-1,0),(0,-1),(1,0),(0,1)) \mod (5,5)$$

with walking probability $\epsilon_t \sim \mathrm{Bernoulli}(0.2)$, uniformly in any direction.

The costs are then given by the average toroidal distance d (the L_1 "wrap-around" distance on the torus) between the major agent and its target, the average distance between major and minor agents, and the crowdedness of agents

$$r(x_t^0, u_t^0, \mu_t) = -0.5 d(x_t^0, x_t^*) - 2.5 \int d(x, x_t^0) \mu_t(\mathrm{d}x) - 6.25 \int \mu_t(x) \mu_t(\mathrm{d}x).$$

The initialization of agents is uniform, while the target starts at zero, i.e. $\mu_0 = \mathrm{Unif}(\mathcal{X})$ and $\mu_0^0 = \mathrm{Unif}(\mathcal{X}) \otimes \delta_{(0,0)}$. For sake of simulation, we define the episode length $T = 200$ after which a new episode starts.

For the neural network policy, we use a one-hot encoding of major states as input, i.e. the concatenation of two 5-dimensional one-hot vectors for the major agent position $\hat{x}_t^0$ and its target position x_t^* respectively.

Foraging. In the Foraging problem, we formally define $\mathcal{X} = [-2,2]^2 \times [0,1]$, $\mathcal{U} = [-1,1]^2 = \mathcal{U}^0$ and $\mathcal{X}^0 = ([-2,2] \times [-2,-1]) \times \bigcup_{n=0}^5 \left([-2,2]^2 \times [0,1.5]\right)^n$. The minor agent states $x_t^i = (\hat{x}_t^i, \tilde{x}_t^i)$ here contain their positions $\hat{x}_t^i \in [-2,2]^2$ and encumbrance (or inversely, free cargo space) $\hat{x}_t^i \in [0,1]$. Meanwhile, the major agent state $x_t^0 = (\hat{x}_t^0, x_t^{\mathrm{env}})$ here contains both the major agent position $\hat{x}_t^0$ restricted to $[-2,2] \times [-2,-1]$, and the current environment state x_t^{env}. Here, the minor and major agents move as in Formation, though with different maximum velocities for minor agents $v_{\max} = 0.3$ and major agent $v_{\max}^0 = 0.1$ respectively.

An additional environmental state consists of up to 5 spatially localized foraging areas, which is not observed by the agents. In each time step, $N_t = \mathrm{Pois}(0.2)$ new foraging areas appear, up to a maximum total number of 5. The location x_t^m of each foraging area $m = 1, \ldots, 5$ is sampled uniformly randomly from $\mathrm{Unif}(\mathcal{X})$, while their total initial size L_t^m is sampled from $\mathrm{Unif}([0.5, 1.5])$, making up the environment state $x_t^{\mathrm{env}} = (x_t^m, L_t^m)_m$. At every time step, the foraging areas m are depleted by nearby agents closer than range 0.5,

$$\begin{aligned} L_{t+1}^m &= L_t^m - \Delta L^m(\mu_t), \\ \Delta L^m(\mu_t) &\coloneqq \min(L_{t+1}^m - L_t^m, \min(0.1, \int (0.5 - \|x - x_t^m\|_2)^+ \mu_t(\mathrm{d}x)) \end{aligned}$$

where $(\cdot)^+ \coloneqq \max(0, \cdot)$, until they are fully depleted and disappear ($L_{t+1}^m \leq 0$).

Foraging minor agents simulate encumbrance, gaining it from nearby foraging areas and depositing to a nearby major agent, by splitting the foraged amount among all nearby minor agents according to their foraged contribution, and wasting any amount going beyond maximum encumbrance 1,

$$\tilde{x}_{t+1}^i = \begin{cases} \min(1, \tilde{x}_t^i + \Delta L^m(\mu_t) \cdot \frac{(0.5 - \|x - x_t^m\|_2)^+}{\int (0.5 - \|x - x_t^m\|_2)^+ \mu_t(\mathrm{d}x)}) & \text{if} \quad \|x_t^i - x_t^0\|_2 \geq 0.5, \\ 0 \quad \text{else.} \end{cases}$$

The reward at each time step is then given by the according total foraged and then deposited amount by the minor agents, where any clipped amount is wasted.

The initialization of agents is uniform, while the environment starts empty, i.e. $\mu_0 = \mathrm{Unif}(\mathcal{X})$ and $\mu_0^0 = \mathrm{Unif}(\mathcal{X}) \otimes \delta_\emptyset$. For sake of simulation, we define the episode length $T = 200$ after which a new episode starts.

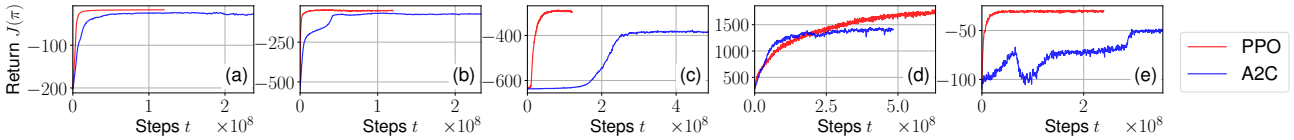

Figure 10. Training curves (mean episode return vs. time steps) of M3FPPO in red, compared to A2C in blue. (a) 2G; (b) Formation; (c) Beach; (d) Foraging; (e) Potential.

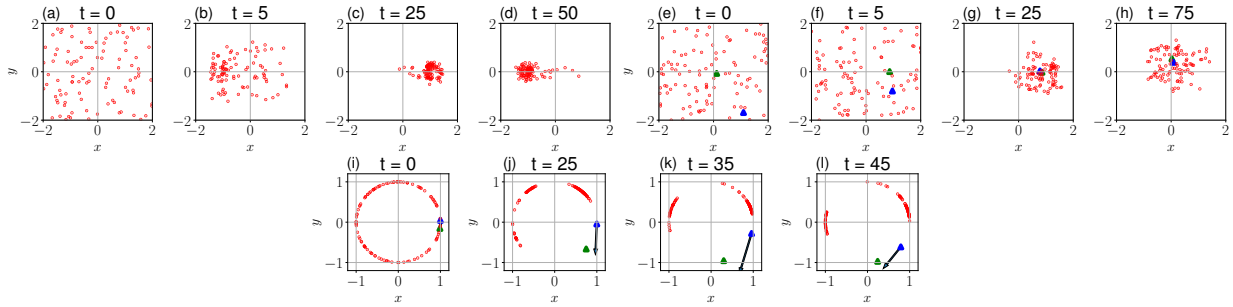

Figure 11. Qualitative visualization of learned M3FC behavior in the 2G (a-d), Formation (e-h) and Potential (i-l) problems. Red: minor agent; blue triangle: major agent; green triangle: major agent target. (i-l): As in (e-h), with arrow for potential gradient (not to scale).

Potential. Lastly, in Potential we consider minor agents on a continuous one-dimensional torus $\mathcal{X} = [-2, 2]$ (where the points -2 and 2 are identified), actions $\mathcal{U} = [-1, 1]$ and major state $\mathcal{X}^0 = \mathcal{X} \times \mathcal{X}$. The minor agents move as in Foraging (wrapping around the torus instead of clipping), while the major agent follows the gradient of the potential landscape generated by minor agents, with the goal of staying close to its current target. The major agent state $x_t^0 = (\hat{x}_t^0, x_t^*)$ here contains both the major agent position $\hat{x}_t^0$ and its target position x_t^*. For simplicity, here we use a linear repulsive force decreasing from $\frac{1}{N}$ to 0 over a range of 1,

$$\hat{x}_{t+1}^0 = \hat{x}_t^0 + \frac{1}{20} \sum_{x_{\text{off}} \in \{-4,0,4\}} \int (1 - \|\hat{x}_t^0 - x + x_{\text{off}}\|_2)^+ \frac{\hat{x}_t^0 - x + x_{\text{off}}}{\|\hat{x}_t^0 - x + x_{\text{off}}\|_2} \mu_t(\mathrm{d}x) \quad \mod [-2, 2]$$

where we let terms $0/0 = 0$ and use the offset x_{off} to account for the wrap-around on the torus.

The target follows the discretized Ornstein-Uhlenbeck process

$$x_{t+1}^* \sim \mathcal{N}\left(0.99 x_t^*, \mathrm{diag}(\sigma_{\text{targ}}^2, \sigma_{\text{targ}}^2)\right)$$

with covariance $\sigma_{\text{targ}}^2 = 0.005$, and gives rise to the reward function via the toroidal distance between target and major agent

$$r(x_t^0, \mu_t) = -d(\hat{x}_t^0, x_t^*).$$

The initialization of agents is uniform, while the target starts around zero, i.e. $\mu_0 = \mathrm{Unif}(\mathcal{X})$ and $\mu_0^0 = \mathrm{Unif}(\mathcal{X}) \otimes \mathcal{N}\left(0, \mathrm{diag}(\sigma_{\text{targ}}^2, \sigma_{\text{targ}}^2)\right)$. For sake of simulation, we define the episode length $T = 100$ after which a new episode starts. In contrast to $M = 7^2 = 49$ in 2G, Formation and Foraging, here we use $M = 7$ bins for the one-dimensional problem.

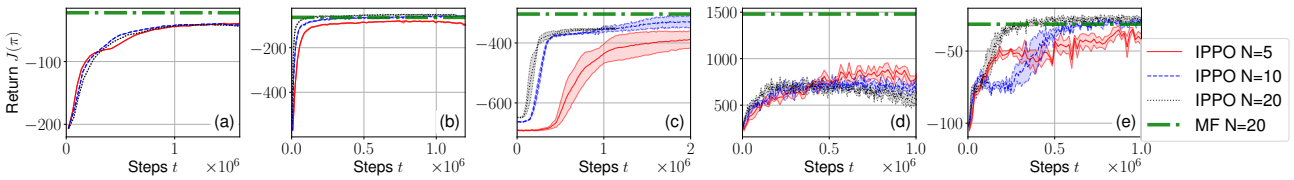

Figure 12. Training curves (mean episode return vs. time steps) of IPPO, trained on the systems with $N \in \{5, 10, 20\}$. (a) 2G; (b) Formation; (c) Beach; (d) Foraging; (e) Potential.

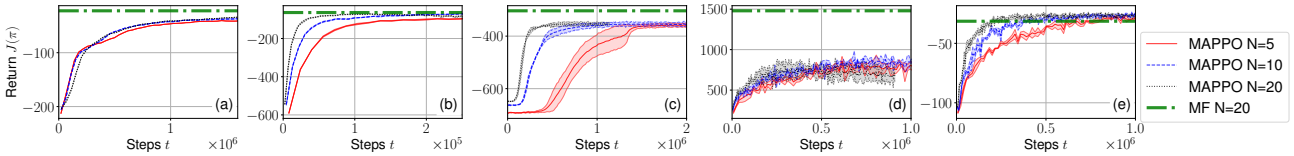

Figure 13. Training curves (mean episode return vs. time steps) of MAPPO, trained on the systems with $N \in \{5, 10, 20\}$. (a) 2G; (b) Formation; (c) Beach; (d) Foraging; (e) Potential.

R.2. Comparison to M3FA2C

In Figure 10 we can see that vanilla M3FA2C typically performs worse than M3FPPO, getting stuck in worse local optima. Here, we used the same hyperparameters as in PPO. This validates our choice of PPO for M3FMARL.

R.3. Qualitative results

In Figure 11, M3FPPO successfully learns to form mixtures of Gaussians in 2G, and a Gaussian around a moving major agent that tracks its target in Formation. As expected in 2G, the two Gaussians at their sinusoidal peaks $t = 25$ and $t = 50$ are not perfectly tracked, in order to minimize the cost in following time steps, when the other Gaussian reappears. Finally, in Potential the minor agents succeed in pushing the major agent towards its target, while spreading on both sides of the major agent to be able to track any random movement of the target.

R.4. Training M3FPPO, IPPO and MAPPO on smaller systems

In Figure 6 we verified the training of M3FPPO on small finite system. Comparing to Figures 5 and 9, for M3FPPO we see little difference between training on a small finite-agent system versus training on a large system and applying the policy on the smaller system. For the chosen hyperparameters, the performance in the Potential problem depends on the initialization. However, M3FPPO compares especially favorably to IPPO in Beach and Foraging, even when directly training on the finite system. This shows that we can either (i) directly apply M3FPPO as a MARL algorithm to small systems, or (ii) train on a fixed system, and transfer the learned behavior to systems of almost arbitrary other sizes.

Analogously, in Figures 12 and 13 we show the training results for around a day of IPPO and MAPPO for numbers of agents $N = 5$, $N = 10$ and $N = 20$. As seen in the plot, the results for each number of agents is comparable to the analysis shown in the main text. In particular, transferring M3FPPO or comparing with Figure 6, we observe that M3FPPO continues to outperform or match the performance of IPPO and MAPPO, even in the setting with fewer agents.