Mean Field Correlated Imitation Learning

Zhiyu Zhao
Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
zhaozhiyu22@ia.ac.cn

Chengdong Ma
Institute for Artificial Intelligence,
Peking University
Beijing, China
chengdong.ma@stu.pku.edu.cn

Qirui Mi
Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
miqirui2021@ia.ac.cn

Ning Yang
Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
ning.yang@ia.ac.cn

Xue Yan
Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
yanxue2021@ia.ac.cn

Mengyue Yang
University of Bristol
Bristol, United Kingdom
mengyue.yang@bristol.ac.uk

Haifeng Zhang
Institute of Automation, CAS
School of Artificial Intelligence, UCAS
Beijing, China
haifeng.zhang@ia.ac.cn

Jun Wang
University College London
London, United Kingdom
jun.wang@cs.ucl.ac.uk

Yaodong Yang*
Institute for Artificial Intelligence,
Peking University
Beijing, China
yaodong.yang@pku.edu.cn

ABSTRACT

Modeling the behaviors of many-agent games is crucial for capturing the dynamics of large-scale complex systems. This is typically achieved by recovering policies from demonstrations within the Mean Field Game Imitation Learning (MFGIL) framework. However, most MFGIL methods assume that demonstrations are collected from Mean Field Nash Equilibrium (MFNE), implying that agents make decisions independently. When directly applied to situations where agents' decisions are coordinated, such as publicly routed traffic networks, these techniques often fall short. In this paper, we propose the Adaptive Mean Field Correlated Equilibrium (AMFCE), which introduces a generalized assumption that effectively integrates the correlated behaviors common in real-world systems. We prove the existence of AMFCE under mild conditions and theoretically show that MFNE is a special case of AMFCE. Building upon this, we introduce a new Mean Field Correlated Imitation Learning (MFCIL) algorithm, which recovers expert policy more accurately in scenarios where agents' decisions are coordinated. We also provide a theoretical upper bound for the error in recovering the expert policy, which is tighter than that of existing methods. Empirical results on real-world traffic flow prediction and large-scale economic simulations demonstrate that MFCIL significantly improves the predictive performance of large populations' behaviors compared to existing MFGIL baselines. This improvement highlights potential of MFCIL to model real-world multi-agent systems.

*Corresponding to Yaodong Yang ⟨yaodong.yang@pku.edu.cn⟩.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent systems**.

KEYWORDS

Imitation Learning; Mean Field Games; Correlated Equilibrium

ACM Reference Format:
Zhiyu Zhao, Chengdong Ma, Qirui Mi, Ning Yang, Xue Yan, Mengyue Yang, Haifeng Zhang, Jun Wang, and Yaodong Yang*. 2025. Mean Field Correlated Imitation Learning. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 9 pages.

1 INTRODUCTION

Modeling behaviors in large-scale multi-agent systems is crucial for characterizing the properties of complex systems, a task typically achieved by recovering policies from demonstrations. These systems span various domains, such as traffic management [1, 15], ad auctions [9] and economic activities among human [16], where individual agent decisions collectively drive system dynamics. However, recovering policies in such environments presents significant challenges due to the high dimensionality and complexity of agent interactions. Mean Field Game Imitation Learning (MFGIL) has shown significant promise in addressing these challenges by recovering policies from demonstrations and reformulating multi-agent interactions within the Mean Field Game (MFG) framework [6, 7]. In MFG settings, the states of the entire population can be effectively summarized into an empirical state distribution, simplifying the problem by reducing it to a game between a representative agent and this empirical distribution [9, 26].

The existing literature on MFGIL typically assumes that expert demonstrations are collected from the Mean Field Nash Equilibrium (MFNE) or its variants [6, 25]. However, in many real-world scenarios, these expert demonstrations may originate from environments where agents' actions are coordinated. The MFNE, which assumes

that agents make decisions independently, thus has limited applicability in such settings [14]. For example, drivers' decisions in a traffic network often depend on routing recommendations from mapping applications. The correlated signals introduced by these applications cannot be adequately captured by MFNE. In summary, the lack of a comprehensive MFG solution concept that addresses coordinated decision-making significantly restricts the realism and practicality of current MFGIL algorithms.

To address this limitation, we introduce the Adaptive Mean Field Correlated Equilibrium (AMFCE), a more nuanced and adaptable solution concept that incorporates the correlated behaviors inherent in real-world systems. By acknowledging and integrating the coordinated decision-making, AMFCE allows for more accurate and effective modeling of complex real-world scenarios. We prove that MFNE is a subclass of AMFCE, implying the broader applicability of our AMFCE-based IL algorithm than existing MFGIL algorithms. We further propose a novel imitation learning algorithm built upon the AMFCE concept, called "Mean Field Correlated Imitation Learning" (MFCIL), which is the first to recover a Correlated Equilibrium (CE) policy in MFGs. The flexibility and adaptability of AMFCE allow MFCIL to more accurately model and predict a wider range of real-world scenarios. We also establish a theoretical upper bound for the error in recovering expert policy. Notably, MFCIL is the first practical MFGIL algorithm with a polynomial dependency on the horizon T, specifically $O(T\sqrt{T})$, for performance differences. Our theoretical analysis extends existing analysis results on MFNE to a more general MFG equilibrium.

We conduct experiments on a variety of tasks, including real-world scenarios like traffic flow prediction and large-scale economic simulations. These experiments are designed to validate the effectiveness of our proposed algorithm by comparing its performance against state-of-the-art MFGIL methods. The results show that our approach consistently recovers the expert policy more accurately than existing methods across all tasks, demonstrating its superiority in both theoretical guarantees and practical applications.

2 PRELIMINARIES

2.1 Classic mean field Nash equilibrium

The classic MFG models a game between a representative agent and the state distribution of all the other agents. Denote $\mathcal{P}(\mathcal{X})$ as the set of probability distributions over the set $\mathcal{X}$ and denote $\mathcal{T} = \{0, 1, \cdots, T\}$ as a set of time indexes. T is the time horizon. The state space and the action space are denoted as $\mathcal{S}$ and $\mathcal{A}$, respectively. The population state distribution of a homogeneous N-agent game at time t is $\mu_t(s) \triangleq \lim_{N\to\infty} \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\{s_t^i = s\}$, where s_t^i is the state of agent i at time t, and $\mathbb{1}_{\{e\}}$ is an indicator function (with value 1 if expression e holds and 0 otherwise). The mean field flow is defined as $\boldsymbol{\mu} = \{\mu_t\}_{t\in\mathcal{T}}$. The transition kernel for the state dynamics is denoted as $P : \mathcal{S}\times\mathcal{A}\times\mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{S})$. At time t, after the representative player chooses its action a_t according to policy π_t, it will receive a deterministic reward $r(s_t, a_t, \mu_t)$, and its state will evolve according to the current state $s_t \in \mathcal{S}$ and transition kernel $P(\cdot|s_t, a_t, \mu_t)$. For a fixed mean field flow $\boldsymbol{\mu}$, the objective of the representative agent is to solve the following decision-making problem over all admissible policies $\boldsymbol{\pi} = \{\pi_t\}_{t\in\mathcal{T}}$:

$$\begin{aligned} &\text{maximize}_{\boldsymbol{\pi}} \quad \mathbb{E}_{s\sim\mu_0}[V_0(s, \boldsymbol{\pi}, \boldsymbol{\mu})] \triangleq \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)\middle| s_k = s\right] \\ &\text{subject to} \quad s_{t+1} \sim P(\cdot|s_t, a_t, \mu_t), \quad a_t \sim \pi_t(s_t), \end{aligned} \tag{1}$$

where $\gamma \in (0, 1]$ is the discount factor. The MFNE [9, 14] is defined as the following.

Definition 2.1 (MFNE)**.** In classic MFG (Equation (1)), a policy-population profile $(\boldsymbol{\pi}^\star, \boldsymbol{\mu}^\star)$ is called an MFNE (under initial state distribution μ_0) if

(1) For any policy $\boldsymbol{\pi}$, $\mathbb{E}_{s\sim\mu_0}[V_0(s, \boldsymbol{\pi}^\star, \boldsymbol{\mu}^\star)] \geq \mathbb{E}_{s\sim\mu_0}[V_0(s, \boldsymbol{\pi}, \boldsymbol{\mu}^\star)]$.
(2) (Population side) The mean field flow $\boldsymbol{\mu}^*$ satisfies

$$\mu_t^*(\cdot) = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} P(\cdot|s, a, \mu_{t-1}^*)\pi_{t-1}^*(a|s)\mu_{t-1}^*(s), \tag{2}$$

with initial condition $\mu_0^* = \mu_0$.

The single player side condition captures the optimality of $\boldsymbol{\pi}^\star$ when the mean field flow $\boldsymbol{\mu}$ is fixed. The population side condition ensures the "consistency" of the solution by guaranteeing that the state distribution flow of the single player matches the mean field flow $\boldsymbol{\mu}^\star$.

2.2 Imitation Learning

Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \mu_0, \gamma, T)$ represent a single-agent Markov decision process (MDP). In this notation, $\mathcal{S}$ and $\mathcal{A}$ denote the state and action spaces, respectively. The transition kernel for the state dynamics is denoted by $P : \mathcal{S}\times\mathcal{A} \to \mathcal{P}(\mathcal{S})$. The reward function is denoted as $r : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$. The initial distribution of the initial state s_0 is denoted as μ_0. The discount factor is represented by $\gamma \in (0, 1]$, and T corresponds to the horizon. The expected return of a policy π is defined as $J(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)\right]$, where the expectation is taken with respect to $s_0 \sim \mu_0$, $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$.

In the IL setting, a set of expert demonstrations sampled from expert policy π^E is provided. The goal of IL is to recover the expert policy π^E using the expert demonstration.

IRL is a subclass of IL and it solves the problem in two steps. It first finds a reward function $\tilde{r} = \max_r \left(\min_\pi -H(\pi) - J(\pi)\right) + J(\pi^E)$ that rationalizes the expert policy π^E, where $H(\pi) \triangleq \mathbb{E}_\pi[-\log \pi(a|s)]$ is the causal entropy of the policy π [2]. Then a recovered policy is learned from the reward function $\tilde{r}$ by a reinforcement learning method.

Generative Adversarial Imitation Learning (GAIL) [10] treats IL as a mini-max game and is trained using a Generative Adversarial Network (GAN). GAIL introduces a discriminator D_ω to differentiate state-action pairs from π^E and other policies. The recovered policy π_θ, parameterized by θ, plays the role of the generator. It aims at generating state-action pairs that are difficult for D_ω to differentiate. The objective function of GAIL is thus defined as

$$\max_\theta \min_w \mathbb{E}_{(s,a)\sim\pi_\theta}\left[\log\left(D_\omega(s, a)\right)\right] + \mathbb{E}_{(s,a)\sim\pi^E}\left[\log\left(1 - D_\omega(s, a)\right)\right], \tag{3}$$

where $\mathbb{E}_{(s,a)\sim\pi_\theta}$ is expectation taken with respect to $s_{t+1} \sim P(\cdot|s_t, a_t)$, $a_t \sim \pi_\theta(\cdot|s_t)$, $s_0 \sim \mu_0$ and $\mathbb{E}_{(s,a)\sim\pi^E}$ is expectation taken with respect to $s_{t+1} \sim P(\cdot|s_t, a_t)$, $a_t \sim \pi^E(\cdot|s_t)$, $s_0 \sim \mu_0$.

3 PROBLEM FORMULATION

In this section, we introduce the concept of AMFCE. Then, we establish the existence of AMFCE under mild conditions and demonstrate that the MFNE solution set is a subset of the AMFCE solution set. Additionally, in Appendix B.1, we prove that AMFCE in mean field games approximates the CE for finite agent settings.

3.1 Adaptive Mean Field Correlated Equilibrium

Compared with MFNE, AMFCE introduces correlated signals in the process of action sampling, which enlarges the policy set and provides a more general solution concept for modeling the real-world decision-making processes where actions of different agents are coordinated. Before the introduction of the AMFCE, we first introduce the concepts of correlation device [18] and behavioral policy.

Definition 3.1 (Correlation Device)**.** The per-step correlation device $\rho_t \in \mathcal{P}(\mathcal{Z})$ is a distribution over the finite correlated signal space $\mathcal{Z}$, from which the correlated signal z_t is sampled at time t. We denote $\boldsymbol{\rho} = \{\rho_t\}_{t=0}^T$ as correlation device over the entire horizon.

Definition 3.2 (Behavioral Policy)**.** For each time t, the per-step behavioral policy $\pi_t : \mathcal{Z} \times \mathcal{S} \to \mathcal{P}(\mathcal{A})$ maps the state s and correlated signal z to a distribution over the action space $\mathcal{A}$.

We denote $\boldsymbol{\pi} = \{\pi_t\}_{t=0}^T$ as the behavioral policy over the entire horizon. The term "policy" may be used to replace "behavioral policy" without confusion. The action space and the state space are finite. At each time step t, a correlated signal z_t is sampled from the per-step correlation device ρ_t. Subsequently, for each agent at state s_t, a mediator samples an action a_t from the per-step behavioral policy $\pi_t(\cdot|s_t, z_t)$ as the recommended action for the agent. Importantly, this recommended action a_t is *private*, accessible only to the respective agent. Mathematically, denote $\mathcal{I}_t = \{\rho_t, a_t, \pi_t, s_t, \mu_t\}$ as the information available to the agent at the beginning of step t. $\mathcal{I}_t$ serves as a criterion for evaluating whether a policy and correlation device constitute an AMFCE, similar to how the population distribution is used in typical MFNE concepts. The presence of μ_t does not imply agents have knowledge of the population distribution. Neither the policy $\pi(a|s, z)$ nor the correlation device $\rho(z)$ relies on precise population distribution information. Note that the agent only observes the functional form of π_t but *cannot observe* the correlated signal z_t nor the recommended actions for other agents. Therefore, the agent has to *predict* the correlated signal z_t based on the local information $\mathcal{I}_t$:

$$\rho_t^{\text{pred}}(z_t = z|\mathcal{I}_t) = \frac{\rho_t(z)\pi_t(a_t|s_t, z)}{\sum_{z' \in \mathcal{Z}} \rho_t(z')\pi_t(a_t|s_t, z')}. \tag{4}$$

The agent can then update the prediction for the population state distribution of the next time step for each possible signal z using the McKean-Vlasov equation:

$$\mu_{t+1}^{\text{pred}}(\cdot|\mathcal{I}_t, z) = \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \mu_t(s)P(\cdot|s, a, \mu_t)\pi_t(a|s, z) \triangleq \Phi(\mu_t, \pi_t, z). \tag{5}$$

Given the population state distribution μ, the agent will choose action a to maximize the action value function $Q_t^{\boldsymbol{\pi}}(s, a, \mu, z; \boldsymbol{\pi}')$:

$$Q_t^{\boldsymbol{\pi}}(s, a, \mu, z; \boldsymbol{\pi}') = r(s, a, \mu) + \gamma \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\pi}', \boldsymbol{\rho}} \left[\sum_{i=t+1}^{T} \gamma^{i-t-1} r(s_i, a_i, \mu_i) \right]. \tag{6}$$

The action value function is the expected return of an agent when the agent follows policy $\boldsymbol{\pi}$ while the population adheres to policy $\boldsymbol{\pi}'$ under the correlation device $\boldsymbol{\rho}$, conditioned on $(s_t, a_t, \mu_t, z_t) = (s, a, \mu, z)$. Unless otherwise stated, the expectation $\mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\pi}', \boldsymbol{\rho}}$ is taken with respect to $z_t \sim \rho_t(\cdot), s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1}), a_t \sim \pi_t(\cdot|s_t, z_t)$, $\mu_t = \Phi(\mu_{t-1}, \pi'_{t-1}, z_{t-1})$.

To introduce the concept of AMFCE, we define the set of swap function

$$\mathcal{U} \triangleq \{u : \mathcal{A} \to \mathcal{A}\},$$

namely u is a function that modifies an action a to an action $u(a)$. Let $\Delta_t(s, \mu, u; \boldsymbol{\pi}, \boldsymbol{\rho}) = \mathbb{E}\left[Q_t^{\boldsymbol{\pi}}(s, u(a), \mu, z; \boldsymbol{\pi}) - Q_t^{\boldsymbol{\pi}}(s, a, \mu, z; \boldsymbol{\pi})\right]$ denote the difference in the action value functions when the agent takes action $u(a)$ in response to a recommendation a, where $u \in \mathcal{U}$. The expectation is taken with respect to $z \sim \rho_t(\cdot)$, $a \sim \pi_t(\cdot|s, z)$.

Definition 3.3 (AMFCE)**.** The profile $(\boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star)$, comprising the behavioral policy $\boldsymbol{\pi}^\star = \{\pi_t^\star\}_{t=0}^T$ and the correlation device $\boldsymbol{\rho}^\star = \{\rho_t^\star\}_{t=0}^T$, is an AMFCE if

(1) (Single agent side) No agent has an incentive to unilaterally deviate from the recommended action after predicting the z by Equation (4), i.e. $\Delta_t(s, \mu_t^\star, u; \boldsymbol{\pi}^\star, \boldsymbol{\rho}^\star) \le 0, \quad \forall u \in \mathcal{U}, \forall s \in \mathcal{S}, \forall t \in \mathcal{T}$.

(2) (Population side) The mean field flow $\boldsymbol{\mu}^*$ satisfies $\mu_t^*(\cdot) = \Phi(\mu_{t-1}^\star, \pi_{t-1}^\star, z_{t-1})$, given the correlated signals $\{z_t\}_{t=0}^T$ and initial condition $\mu_0^* = \mu_0$.

3.2 Properties of AMFCE

This subsection focuses on the properties of AMFCE, including the conditions to guarantee its existence and its relationship to classic MFNE. To provide the existence of AMFCE solutions, we define the best response operator

$$\text{BR}(\boldsymbol{\pi}; \boldsymbol{\rho}) = \arg\max_{\boldsymbol{\pi}'} \mathbb{E}_{\boldsymbol{\pi}', \boldsymbol{\pi}, \boldsymbol{\rho}} \left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t) \right].$$

Then the existence of AMFCE is derived using Kakutani's fixed point theorem [12] with the operator BR. We next provide a sufficient condition for the existence of AMFCE.

Theorem 3.4. *Let the reward function $r(s, a, \mu)$ and transition kernel $P(s'|s, a, \mu)$ be bounded and continuous with respect to the population state distribution μ. Under these mild conditions, there exists at least one AMFCE solution.*

Proof sketch. We first prove that BR has a closed graph (Lemma B.2), and $\text{BR}(\boldsymbol{\pi}; \boldsymbol{\rho})$ is a convex set given $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$. (Lemma B.3). According to Kakutani's fixed point theorem, there exists $\boldsymbol{\pi}^* = \text{BR}(\boldsymbol{\pi}^*; \boldsymbol{\rho})$. Therefore, $\Delta_t(s_t, \mu_t, u; \boldsymbol{\pi}^*, \boldsymbol{\rho}) \le 0, \quad \forall u \in \mathcal{U}, \forall s_t \in \mathcal{S}, \forall t \in \mathcal{T}$ and $\boldsymbol{\mu} = \{\mu_t\}_{t=0}^T$ satisfies the population side condition of AMFCE. □

AMFCE is a more general solution concept compared to MFNE. Corollary 3.5 shows that MFNE is a subclass of AMFCE.

Corollary 3.5. *Every MFNE can be transformed into an AMFCE.*

The proof is deferred to Appendix B.4. Corollary 3.5 implies that any IL algorithm designed to recover AMFCE policies can also recover MFNE policies.

4 IMITATION LEARNING FOR AMFCE

In this section, we propose a novel IL algorithm for recovering AMFCE from expert demonstrations.

We denote the AMFCE under the designed reward function r and correlation device $\boldsymbol{\rho}$ as AMFCE$(r, \boldsymbol{\rho})$. The condition of AMFCE, as defined in Definition 3.3, implies that agents cannot improve the policy $\boldsymbol{\pi}$ through 1-step temporal difference learning. We proceed to derive equivalent constraints for multi-step temporal difference learning, outlined in Proposition 4.2. Utilizing the Lagrangian reformulation of these equivalent multi-step constraints, we propose the IL algorithm for recovering AMFCE.

We first introduce the concept of the Correlated Imitation Gap (CIG) for deriving the multi-step constraints.

Definition 4.1 (CIG)**.** For a given action sequence $a_{0:T}$, the policy $\boldsymbol{\pi}$ and correlation device $\boldsymbol{\rho}$, the CIG is defined as $\mathcal{R}(a_{0:T}, \boldsymbol{\pi}, \boldsymbol{\rho}) \triangleq \mathbb{E}\Big[\sum_{t=0}^{T}\gamma^t r(s_t, a_t, \mu_t)\Big| a_{0:T}\Big] - J(\boldsymbol{\pi}, \boldsymbol{\pi}, \boldsymbol{\rho})$, where the expectation is taken with respect to $z_t \sim \rho_t(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$. Here, $J(\boldsymbol{\pi}, \boldsymbol{\pi}', \boldsymbol{\rho}) = \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\pi}',\boldsymbol{\rho}}\left[\sum_{t=0}^{T}\gamma^t r(s_t, a_t, \mu_t)\right]$ represents the expected return of the agent when it follows policy $\boldsymbol{\pi}$ while the population adheres to policy $\boldsymbol{\pi}'$ under the correlation device $\boldsymbol{\rho}$.

The CIG is defined as the gap of expected return between the agent taking action sequence $a_{0:T}$ and the policy $\boldsymbol{\pi}$. Then we can get a criterion for AMFCE based on CIG.

Proposition 4.2. $(\boldsymbol{\pi}, \boldsymbol{\rho})$ *is an AMFCE solution if and only if*

$$\mathcal{R}(a_{0:T}, \boldsymbol{\pi}, \boldsymbol{\rho}) \le 0,$$

$\forall a_t \in \mathcal{A}$, $0 \le t \le T$.

The proof is deferred to Appendix B.5. Intuitively, Proposition 4.2 shows the multi-step constraints for AMFCE. Therefore, the process of finding AMFCE can be defined as an optimization problem with finite constraints measured by the CIG. We propose a Lagrangian reformulation to find AMFCE. $L(\boldsymbol{\pi}, \boldsymbol{\rho}, \lambda, r) \triangleq \sum_{\tau_k \in \mathcal{D}_E}\lambda(\tau_k)\mathcal{R}(a_{0:T}, \boldsymbol{\pi}, \boldsymbol{\rho})$, where $\mathcal{D}_E$ is a set of action-signal sequences $\tau_k = \{(a_t, z_t)\}_{t=0}^{T}$. We show that the Lagrangian form captures the difference of expected returns between two policies by selecting λ.

Theorem 4.3. *For policy $\boldsymbol{\pi}^*$ and correlation device $\boldsymbol{\rho}$, let $\lambda_{\boldsymbol{\pi}^*}(\tau_k) = \prod_{t=0}^{T}\rho_t(z_t)\pi_t^*(a_t|s_t, z_t)$ be the probability of generating the sequence τ_k using policy $\boldsymbol{\pi}^*$ and correlation device $\boldsymbol{\rho}$. Then we have $L(\boldsymbol{\pi}, \boldsymbol{\rho}, \lambda_{\boldsymbol{\pi}^*}, r) = J^{\boldsymbol{\pi}^*}(\boldsymbol{\pi}, \boldsymbol{\pi}, \boldsymbol{\rho}) - J(\boldsymbol{\pi}, \boldsymbol{\pi}, \boldsymbol{\rho})$, where*

$$J^{\boldsymbol{\pi}^*}(\boldsymbol{\pi}, \boldsymbol{\pi}, \boldsymbol{\rho}) = \mathbb{E}_{\boldsymbol{\pi}^*}\left[\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\pi},\boldsymbol{\rho}}\left[\sum_{t=0}^{T}\gamma^t r(s_t, u(a_t), \mu_t)\middle|\{u(a_t) = a_t'\}_{0:T}\right]\right],$$

$a_t' \sim \pi_t^*(\cdot|s_t, \mu_t)$.

The proof of Theorem 4.3 is deferred to Appendix B.6.

Our approach addresses the challenge of inaccessible reward signals in IL settings by constructing a reward function that rationalizes the expert policy. To achieve this, we introduce the AMFCE Inverse Reinforcement Learning (AMFCE-IRL) operator AMFCE $-$ IRL$_\psi$, equipped with a reward regularizer ψ, as motivated by Theorem 4.3. This operator aims to maximize the gap in expected returns between the expert policy $\boldsymbol{\pi}^E$ and an alternative policy $\boldsymbol{\pi}$, effectively rationalizing the expert policy based on expert demonstrations.

$$\text{AMFCE} - \text{IRL}_\psi(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) = \arg\max_r\Big(-\psi(r) - \max_{\boldsymbol{\pi}} L(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E, \lambda_{\boldsymbol{\pi}^*}, r)\Big), \tag{7}$$

where $(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E)$ is the AMFCE from which expert demonstrations are sampled. The regularizer for the reward function is chosen as the adversarial reward function regularizer to avoid overfitting [10].

We recover the AMFCE policy AMFCE$(\tilde{r}, \boldsymbol{\rho}^E)$ by Equation (8), where $\tilde{r} = \text{AMFCE} - \text{IRL}(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E)$.

$$\begin{aligned}\text{AMFCE} \circ \text{AMFCE-IRL}_\psi(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) = \arg\min_{\boldsymbol{\pi}}\max_r & J(\boldsymbol{\pi}^E, \boldsymbol{\pi}^E, \boldsymbol{\rho}^E) \\ & - J^{\boldsymbol{\pi}}(\boldsymbol{\pi}^E, \boldsymbol{\pi}^E, \boldsymbol{\rho}^E) - \psi_{GA}(r)\end{aligned}, \tag{8}$$

Proposition 4.4. *The objective in Equation* (8) *can be reformulated as the following practical objective function:*

$$\begin{aligned}\min_{\boldsymbol{\pi}}\max_{\omega} & \mathbb{E}_{\boldsymbol{\pi}^E,\boldsymbol{\pi}^E,\boldsymbol{\rho}^E}\left[\sum_{t=0}^{T}\gamma^t\log D_\omega(s_t, a_t, \mu_t)\right] \\ & + \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\pi}^E,\boldsymbol{\rho}^E}\left[\sum_{t=0}^{T}\gamma^t\log\big(1 - D_\omega(s_t, a_t, \mu_t)\big)\right],\end{aligned} \tag{9}$$

where D_ω represents the discriminator network parameterized with ω, taking (s_t, a_t, μ_t) as input and producing a real number in the range $(0, 1)$ as output.

The proof is deferred to Appendix B.7. This proposition shows that the AMFCE policy can be recovered by the GAN. Note that simply using Equation (9) to solve AMFCE cannot recover $\boldsymbol{\rho}^E$, so we derive $\boldsymbol{\rho}$ using a gradient descent method in the Proposition 4.5 with proof in Appendix B.8.

Proposition 4.5. *If the correlation device ρ_t^ϕ is parameterized with ϕ, the gradient to optimize ϕ given state s is*

$$\mathbb{E}_{z\sim\rho_t^\phi(\cdot)}\left[\nabla_\phi\log\rho_t^\phi(z)\mathbb{E}_{a\sim\pi_t(\cdot|s,z)}Q_t^{\boldsymbol{\pi}}(s, a, \mu, z; \boldsymbol{\pi})\right].$$

The population state distribution μ_t influences both the input of D_ω and transition kernel in Equation (9). However, the population state distribution μ_t in expert demonstrations is often inaccessible. We characterize μ_t using the signature of $\boldsymbol{z}_{0:t}$ from rough path theory [13], denoted as $\hat{\mu}_t = \text{Sig}(\boldsymbol{z}_{0:t})$, bypassing the circular reasoning problem [21]. Please refer to Appendix I for details.

We approximately optimize the following surrogate objective function of Equation (9).

$$\begin{aligned}\min_{\boldsymbol{\pi}}\max_{\omega} & \mathbb{E}_{\boldsymbol{\pi}^E,\boldsymbol{\pi}^E,\boldsymbol{\rho}^E}\left[\sum_{t=0}^{T}\gamma^t\big(\log D_\omega(s_t, a_t, \hat{\mu}_t) + \log 2\big)\right] \\ & + \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\pi},\boldsymbol{\rho}^E}\left[\sum_{t=0}^{T}\gamma^t\big(\log\big(1 - D_\omega(s_t, a_t, \hat{\mu}_t)\big) + \log 2\big)\right]\end{aligned} \tag{10}$$

Combine the above analysis, we propose a new algorithm, MFCIL, to recover the AMFCE policy and the correlation device from expert demonstrations. The algorithm is shown in Algorithm 1. Although

Table 1: Results for numerical tasks. The performative difference between the recovered policy and the ground truth policy is measured by log loss under different correlated signals z. The number in the bracket is the standard deviation over 3 independent runs.

Task	Correlated Signal	MFCIL (Our Method)	MFIRL	MFAIRL	Logistic Regression	Multinomial	MaxEnt ICE
Squeeze with $T = \{0, 1, 2\}$	$z = 0$	**0.643 (0.000)**	1.450 (2.857)	4.064 (0.879)	4.484 (0.054)	0.686 (0.002)	-
	$z = 1$	0.647 (0.003)	3.245 (1.650)	4.144 (0.629)	**0.000 (0.000)**	2.577 (0.149)	-
	$z = 2$	**0.020 (0.001)**	1.072 (2.229)	6.934 (4.447)	7.091 (0.107)	0.282 (0.087)	-
	$z = 3$	0.045 (0.005)	7.871 (4.368)	1.027 (1.279)	10.638 (0.163)	**0.001 (0.001)**	-
Squeeze with $T = \{0, 1\}$	$z = 0$	**0.648 (0.002)**	3.828 (1.582)	4.067 (0.088)	1.985 (0.165)	0.991 (0.102)	0.946 (0.073)
	$z = 1$	**0.638 (0.001)**	2.009 (1.191)	10.074 (0.174)	2.139 (0.169)	2.947 (0.359)	0.648 (0.011)
RPS	$z = 0$	**1.083 (0.000)**	7.127 (0.753)	3.221 (1.330)	4.805 (0.131)	5.850 (0.306)	1.537 (0.019)
Flock	$z = 0$	0.002 (0.000)	5.591 (0.869)	12.430 (2.759)	**0.000 (0.000)**	1.383 (0.004)	-
	$z = 1$	**0.016 (0.003)**	11.687 (1.158)	13.042 (1.533)	7.887 (0.031)	1.127 (0.007)	-
	$z = 2$	**0.045 (0.009)**	7.500 (3.955)	10.065 (5.074)	18.339 (0.010)	0.951 (0.009)	-
	$z = 3$	**0.026 (0.003)**	3.847 (3.967)	9.312 (4.711)	35.253 (0.037)	1.264 (0.011)	-

Algorithm 1 Mean field correlated imitation learning (MFCIL)

Require: Expert demonstration set sampled from $(\boldsymbol{\pi}, \boldsymbol{\rho})$: $\mathcal{D}_E = \{s_0, z_0, a_0, s_1, z_1, a_1, \dots s_T, z_T, a_T\}$, initial population state distribution μ_0.
for each iteration **do**
  Obtain trajectories from $(\boldsymbol{\pi}, \boldsymbol{\rho})$ by the process: $s_0 \sim \mu_0$, $a_t \sim \pi^\theta(\cdot|s_t, z_t)$, $s_{t+1} \sim P(\cdot \mid s_t, \mu_t)$, $z_t \sim \rho_t^\phi(\cdot)$;
  for i in $\{0, 1, 2, \dots\}$ **do**
    Update ω based on the surrogate objective function Equation (10).
  end for
  for t in $\{0, 1, 2, \dots\}$ **do**
    Update θ by Actor-Critic algorithm with small step size based on the surrogate objective function Equation (10).
    Update ϕ according to Proposition 4.5;
  end for
end for
Return Policy $\boldsymbol{\pi}^\theta$, correlation device $\boldsymbol{\rho}^\phi$.

this algorithm is designed for recovering AMFCE, it can also be applied to recover MFNE by setting the correlation device as Dirac distribution. In the Theorem 4.7, we provide a theoretical guarantee for the quality of the policy recovered by MFCIL.

Assumption 4.6. The transition kernel $P(\cdot|s, a, \mu)$ and the reward function $r(s, a, \mu)$ are Lipschitz continuous with respect to population state distribution μ and have corresponding Lipschitz constants L_P and L_R, respectively. The reward function is bounded by $r_{\max}$. The expert policy $\boldsymbol{\pi}^E$ and recovered policy $\boldsymbol{\pi}$ satisfy

$$\max_{\omega} \mathbb{E}_{\boldsymbol{\pi}^E, \boldsymbol{\pi}^E, \boldsymbol{\rho}^E}\left[\sum_{t=0}^{T} \gamma^t \left(\log D_\omega(s_t, a_t, \hat{\mu}_t) + \log 2\right)\right] + \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\pi}, \boldsymbol{\rho}^E}\left[\sum_{t=0}^{T} \gamma^t \left(\log\left(1 - D_\omega(s_t, a_t, \hat{\mu}_t)\right) + \log 2\right)\right] \leq \epsilon, \tag{11}$$

which can be achieved by MFCIL.

Theorem 4.7. *Under Assumption 4.6, for any given action sequence $a_{0:T}$, the CIG of recovered policy $\boldsymbol{\pi}$ is bounded by*

$$\mathcal{R}(a_{0:T}, \boldsymbol{\pi}, \boldsymbol{\rho}^E) \leq 2\left(2L_R + r_{\max} + \gamma T L_P r_{\max}\right)\sqrt{2\epsilon T}.$$

Proof sketch. We leverage Lemma B.4 to establish a bound on the expected return differences between the recovered policy $\boldsymbol{\pi}$ and the expert policy $\boldsymbol{\pi}^E$. We then relate the Jensen-Shannon (JS) divergence between the occupancy measures to the optimization error ϵ, leading to a bound on the to the CIG. □

The proof is deferred to Appendix B.9. As the value of ϵ decreases, the policy $\boldsymbol{\pi}$ recovered by MFCIL approaches the AMFCE policy more closely. If $\epsilon = 0$, the recovered policy $\boldsymbol{\pi}$ is an exact AMFCE policy. We also provide the imitation gap between the recovered policy in Corollary 4.8 similar to [21].

Corollary 4.8. *The imitation gap between the recovered policy $\boldsymbol{\pi}$ is bounded by*

$$\max_{\hat{\boldsymbol{\pi}}} J^{\hat{\boldsymbol{\pi}}}(\boldsymbol{\pi}, \boldsymbol{\pi}, \boldsymbol{\rho}^E) - J(\boldsymbol{\pi}, \boldsymbol{\pi}, \boldsymbol{\rho}^E) \leq 2(3L_R + \gamma T L_P r_{\max} + r_{\max})\sqrt{2\epsilon T}.$$

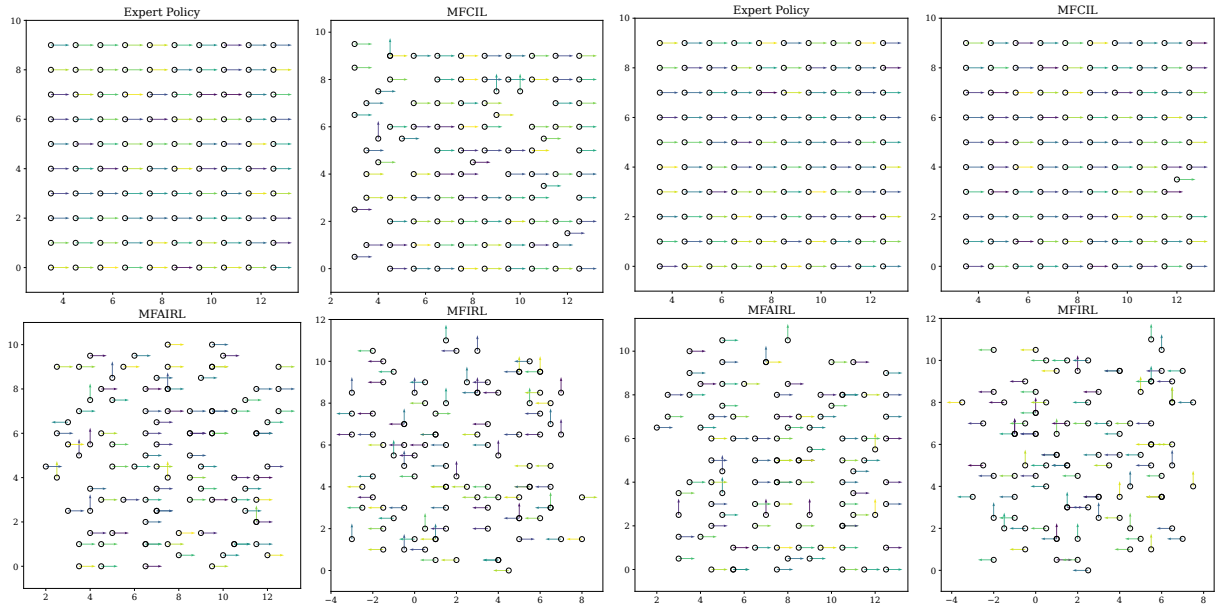
The proof is deferred to Appendix B.10. The imitation gap in Corollary 4.8 exhibits a polynomial dependency on the horizon.

5 EXPERIMENTS

5.1 Tasks

The performance of MFCIL is evaluated through experiments conducted on a diverse range of tasks, which can be categorized into two types: numerical tasks and real-world tasks. We provide the code in https://github.com/zhiyu-zhao-ucas/MFCIL.

Numerical tasks include Sequential Squeeze, Rock-Paper-Scissors (RPS), and Flock. These tasks are widely used in the MFG research.



(a) Recovered policies at time step 500.

(b) Recovered policies at time step 1000.

Figure 1: Visualization of the learning process for the Flock task over 1,000 steps. (a) Policies recovered by different algorithms at step 500, showing that MFCIL successfully recovers the expert policy faster than other methods. (b) Policies at step 1,000, demonstrating that MFCIL is the only method to fully recover the expert policy. These figures depict two-dimensional projections of motion captured from a higher-dimensional space.

Table 2: Results of Traffic Flow Prediction. The metric is log loss. The number in the bracket is the standard deviation over 3 independent runs.

	Lewisham	Hammersmith	Ealing
MFCIL (Our Method)	**0.742 (0.011)**	**0.897 (0.002)**	**1.091 (0.001)**
MFIRL	12.346 (0.294)	9.853 (2.892)	11.625 (0.435)
MFAIRL	8.893 (2.302)	6.485 (1.940)	11.609 (1.202)
	Redbridge	Enfield	Big Ben
MFCIL (Our Method)	**0.052 (0.011)**	**0.394 (0.003)**	**1.599 (0.000)**
MFIRL	11.720 (0.633)	11.750 (0.603)	7.482 (1.539)
MFAIRL	4.537 (4.544)	9.871 (4.052)	12.477 (1.005)

Table 3: Results of TaxAI. The number in the bracket is the standard deviation over 3 independent runs.

	MFCIL	MFIRL	MFAIRL
Wasserstein Distance with the Expert Policy	**27.620 (0.170)**	33.096 (0.912)	49.532 (0.661)
Household Reward	**29243.837 (12.819)**	75.776 (4.488)	116.803 (33.556)
Government Reward	**3355.917 (12.819)**	-678.563 (21.290)	-417.833 (228.319)

For these experiments, expert policies are solved analytically. Real-world tasks encompass Traffic Flow Prediction and TaxAI simulations. The Traffic Flow Prediction task involves predicting the traffic flow in a complex traffic network based on the real-world data.

The TaxAI environment simulates interactions between a government and a large number of households, demonstrating the performance and scalability of our algorithm in complex economic scenarios. The expert policies of these tasks are derived from the real-world data. Given the large-scale and complex nature of real-world tasks, we evaluate the scalability of MFCIL against leading MFGIL baselines in our experiments.

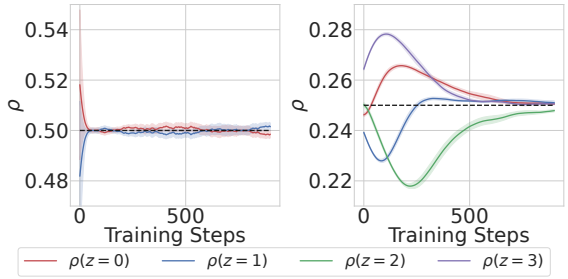


Figure 2: The distribution of correlation device ρ recovered by MFCIL. The solid line shows the mean and the shaded area represents the standard deviation over 3 independent runs. The dash line shows the ground truth of ρ.

5.1.1 *Numerical Tasks.*

Squeeze. Sequential Squeeze is a game with multi-steps. The purpose of implementing this game is to verify the ability to recover expert policy through demonstrations sampled from a multi-step game.

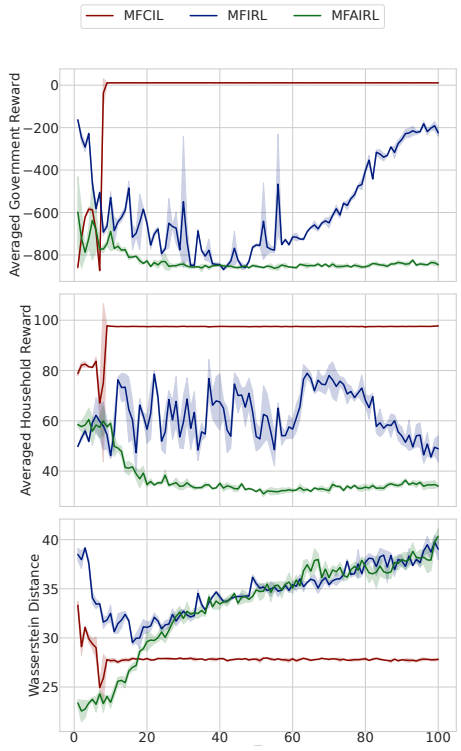


Figure 3: Learning curves of MFCIL, MFIRL and MFAIRL in TaxAI environment. The solid line shows the mean and the shaded area represents the standard deviation over 3 independent runs. The "Averaged Household Reward" and "Averaged Government Reward" indicate the average cumulative rewards over time.

RPS. This RPS task is a traditional MFG task [5, 7, 8]. The demonstrations are sampled from MFNE. We use RPS to verify that the algorithm proposed can recover MFNE, which also supports the result in Corollary 3.5.

Flock. The Flock task is based on the movement of fish. This task aims to evaluate the performance of algorithms in a MFG that does not satisfy the monotonicity condition [19].

5.1.2 Real-world Tasks.

Traffic Flow Prediction. : In the Traffic Flow Prediction task, we use the real-world traffic data of London from Uber Movement. Our goal is to predict traffic flow in a real-world traffic network consisting of six locations: Lewisham, Hammersmith, Ealing, Redbridge, Enfield, and Big Ben. We collected the individual traveling data among these six locations from Uber Movement as expert demonstrations.

TaxAI. We tested our algorithm using the TaxAI environment, which simulates economic interactions between the government and 100 households. This scalable simulator, grounded in real-world data from the 2022 Survey of Consumer Finances, enables detailed analysis and validation of tax policies' impacts on household behavior and government revenue.

More details about the tasks are deferred to Appendix F.

5.2 Baselines

We compare our proposed MFCIL algorithm with state-of-the-art MFGIL algorithms, MFIRL [6], and MFAIRL [7]. Since MFIRL and MFAIRL do not take the correlated signal into consideration, we treat the signature of correlated signals as an extension of the state for all algorithms, enabling a fair comparison among all methods. We also compare MFCIL with MaxEnt ICE, smoothed multinomial distribution over the joint actions, and logistic regression [23]. As MaxEnt ICE is designed to recover correlated equilibrium in the matrix game, we only compare MFCIL with MaxEnt ICE on tasks RPS and Sequential Squeeze with $\mathcal{T} = \{0, 1\}$.

5.3 Evaluation Metrics

We assess the quality of the learned policies for all methods. Our focus lies in the difference between the recovered policy and the expert policy, as shown in Table 1 and Table 2, to evaluate the quality of the policy learned by each method. We use the log loss, $\mathbb{E}_{a\sim\pi^E(\cdot|s,z)}[-\log(\pi(a|s,z))]$, to measure the difference between the recovered policy π and the expert policy π^E in all numerical tasks and the Traffic Flow Prediction Task. The action space in TaxAI simulator is continuous, so we employ the Wasserstein distance to measure the discrepancy between the recovered policy π and the expert policy π^E. Additionally, we present both the government and household rewards.

5.4 Results and Analysis

5.4.1 Numerical tasks. The results for numerical tasks are presented in Table 1. Overall, MFCIL consistently outperforms other methods. While supervised learning methods, such as logistic regression and smoothed multinomial distribution, may occasionally surpass MFCIL in certain metrics, they generally suffer from higher log loss compared to MFCIL. MFIRL and MFAIRL exhibit larger deviations and higher log loss than MFCIL in Table 1. These results underscore the inability of MFIRL and MFAIRL to recover AMFCE and effectively handle games with correlated signals.

MFCIL consistently outperforms MFIRL and MFAIRL in the Squeeze and Flock tasks because it is the first IL algorithm capable of recovering AMFCE. This capability enables MFCIL to effectively handle scenarios where agents' actions are coordinated, where MFNE-based algorithms may struggle. In the RPS task, MFCIL surpasses other algorithms for two key reasons. Theoretically, MFCIL achieves a significantly lower bound on the error between the occupancy measure of the recovered policy and the expert policy compared to traditional MFGIL methods. Since MFNE is a subclass of AMFCE, MFCIL naturally outperforms others in this task. Technically, MFCIL leverages the correlated signal sequence to characterize the population distribution, bypassing the variance introduced by estimating the population distribution from samples, resulting in lower deviation. MaxEnt ICE performs poorly due to its limited reward function class, assuming a linear reward structure.

We visualize the learning process of the Flock task over a total of 1,000 steps. Figure 1a displays the policies recovered by different

algorithms at step 500, while Figure 1b shows the policies at step 1,000. At step 500, MFCIL successfully recovers the expert policy faster than other methods, as illustrated in Figure 1a. By step 1,000, MFCIL is the only method that successfully recovers the expert policy, as shown in Figure 1b. We also plot the distribution of the correlation device ρ recovered by the MFCIL in the Figure 2, illustrating that MFCIL can recover the correlation device with rapid convergence speed.

5.4.2 Real-world tasks. In the Traffic Flow Prediction task, our MFCIL method consistently outperforms other approaches across all locations, achieving significantly lower log loss values for more accurate and stable traffic flow predictions, as detailed in Table 2. In the TaxAI simulations, as shown in Table 3, MFCIL demonstrates superior performance through the lowest Wasserstein distance and higher rewards for households and governments, emphasizing its practical utility in tax policy optimization and potential to enhance real-world economic strategies. The learning curve in Figure 3 indicates faster convergence. These results not only validate MFCIL's effectiveness but also confirm its reliability in managing the complexities of real-world data.

6 RELATED WORK

6.1 Multi-agent Imitation Learning

Previous research in Multi-agent Imitation Learning (MAIL) has extended single-agent IL algorithms to Markov games [11, 22, 27]. However, these algorithms encounter scalability challenges due to the curse of dimensionality. To address the scalability challenge, Yang et al. proposed a multi-type mean field approximation that approximates Nash equilibrium in Markov games [24]. Nevertheless, this approach does not consider the MFG and MFNE, thus failing to account for the interdependence between mean field flow and policy.

6.2 MFG Imitation Learning

Yang et al. introduced a method for inferring the MFG model through Inverse Reinforcement Learning (IRL), under the assumption that the equilibrium underlying the demonstrations is the Mean Field Social Optimum (MFSO). This condition is applicable solely to fully cooperative settings [25]. Chen et al. extended this method to mixed cooperative-competitive settings by assuming that the demonstrations are sampled from MFNE and its variant [6, 7]. Ramponi et al. proposed the solution concept named Nash Imitation Gap (NIG) and provided upper bounds of NIG for several different settings [21], but they focused on experts achieving a Nash equilibrium.

6.3 Mean Field Equilibrium Concepts

While existing MFGIL algorithms have not incorporated CE, there have been a few, albeit limited, works that introduce CE into the MFG. Campi and Fischer assume that a mediator recommends the same stochastic policy to the entire population, resulting in a limited equilibrium set identical to the classic MFNE [3]. Additionally, it is often more practical for the mediator to recommend actions rather than stochastic policies to individuals. Muller et al. [17] assume that the mediator recommends a deterministic policy (sampled from a distribution named "population recommendation" over the deterministic policy space) to each individual. Both MFCE concepts assume that a fixed correlated signal (recommended policy in Campi and Fischer, and population recommendation in Muller et al. [17].) is realized at the beginning of the game, allowing agents to observe future signals or recommendations. However, this assumption is impractical in real-world scenarios where decisions, such as economic behavior, depend on real-time conditions, with future information remaining inaccessible. To address these limitations, we propose the AMFCE concept, which extends the existing MFCE solution concepts by allowing agents to operate without access to future signals, making it more applicable to dynamic, real-world environments.

Table 4: Comparison of Solution Concepts

Solution Concept	Interdependent Decision Making	Independent on Future Information
MFNE	✗	✗
MFCE	✓	✗
AMFCE	✓	✓

This enhanced flexibility caters to real-world scenarios where varying correlated signals are introduced by the mediator. We provide a concrete example demonstrating the greater generality of our equilibrium concept over that proposed by Muller et al. [17] in Appendix G. We also discuss the difference between AMFCE and MFNE with common noise [4, 20] in Appendix E. We also provide an example and explanation in Appendix C to clarify why AMFCE is more practical than MFCE concept in Campi and Fisher. The comparison is summarized in the Table 4.

7 CONCLUSION

In this paper, we introduced the Adaptive Mean Field Correlated Equilibrium (AMFCE) to address the limitations of existing MFGIL methods in modeling correlated agent behaviors. Based on AMFCE, we developed the MFCIL algorithm to enhance policy recovery in environments where decisions of agent are coordinated. Our approach not only recovers expert policies more accurately but also establishes a tighter theoretical upper bound for the error compared to existing methods. We demonstrated MFCIL's effectiveness through experiments including real-world traffic flow prediction and large-scale economic simulations. Results show that MFCIL significantly outperforms existing MFGIL baselines in predicting large population behaviors, particularly in scenarios where agents' decisions are coordinated. Our work expands the applicability of MFGIL to a broader range of real-world multi-agent systems and opens new avenues for modeling complex, correlated behaviors in large-scale populations.

ACKNOWLEDGMENTS

Zhiyu Zhao, Qirui Mi, Xue Yan and Haifeng Zhang thank the support of the NSFC Grant Number 72450002.

REFERENCES

[1] Ana LC Bazzan. 2009. Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Autonomous Agents and Multi-Agent Systems* 18, 3 (2009), 342–375.

[2] Michael Bloem and Nicholas Bambos. 2014. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *53rd IEEE Conference on Decision and Control, CDC 2014, Los Angeles, CA, USA, December 15-17, 2014*. IEEE, 4911–4916. https://doi.org/10.1109/CDC.2014.7040156

[3] Luciano Campi and Markus Fischer. 2022. Correlated equilibria and mean field games: a simple model. *Mathematics of Operations Research* (2022).

[4] René Carmona, François Delarue, and Daniel Lacker. 2016. Mean field games with common noise. (2016).

[5] Yang Chen, Jiamou Liu, and Bakhadyr Khoussainov. 2021. Agent-level maximum entropy inverse reinforcement learning for mean field games. *arXiv preprint arXiv:2104.14654* (2021).

[6] Yang Chen, Libo Zhang, Jiamou Liu, and Shuyue Hu. 2022. Individual-Level Inverse Reinforcement Learning for Mean Field Games. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*, Piotr Faliszewski, Viviana Mascardi, Catherine Pelachaud, and Matthew E. Taylor (Eds.). International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 253–262. https://doi.org/10.5555/3535850.3535880

[7] Yang Chen, Libo Zhang, Jiamou Liu, and Michael Witbrock. 2021. Adversarial Inverse Reinforcement Learning for Mean Field Games. *arXiv preprint arXiv:2104.14654* (2021).

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

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

[10] Jonathan Ho and Stefano Ermon. 2016. Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, Daniel D. Lee, Masashi Sugiyama, Ulrike von Luxburg, Isabelle Guyon, and Roman Garnett (Eds.). 4565–4573. https://proceedings.neurips.cc/paper/2016/hash/cc7e2b878868cbae992d1fb743995d8f-Abstract.html

[11] Wonseok Jeon, Paul Barde, Derek Nowrouzezahrai, and Joelle Pineau. 2020. Scalable and sample-efficient multi-agent imitation learning. In *Proceedings of the Workshop on Artificial Intelligence Safety, co-located with 34th AAAI Conference on Artificial Intelligence, SafeAI@ AAAI*.

[12] Shizuo Kakutani. 1941. A generalization of Brouwer's fixed point theorem. *Duke mathematical journal* 8, 3 (1941), 457–459.

[13] Patrick Kidger and Terry J. Lyons. 2021. Signatory: differentiable computations of the signature and logsignature transforms, on both CPU and GPU. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=lqU2cs3Zca

[14] Mathieu Laurière, Sarah Perrin, Matthieu Geist, and Olivier Pietquin. 2022. Learning Mean Field Games: A Survey. *CoRR* abs/2205.12944 (2022). https://doi.org/10.48550/ARXIV.2205.12944 arXiv:2205.12944

[15] Chengdong Ma, Aming Li, Yali Du, Hao Dong, and Yaodong Yang. 2024. Efficient and scalable reinforcement learning for large-scale network control. *Nature Machine Intelligence* 6, 9 (2024), 1006–1020.

[16] Qirui Mi, Siyu Xia, Yan Song, Haifeng Zhang, Shenghao Zhu, and Jun Wang. 2024. TaxAI: A Dynamic Economic Simulator and Benchmark for Multi-agent Reinforcement Learning. In *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2024, Auckland, New Zealand, May 6-10, 2024*, Mehdi Dastani, Jaime Simão Sichman, Natasha Alechina, and Virginia Dignum (Eds.). International Foundation for Autonomous Agents and Multiagent Systems / ACM, 1390–1399. https://doi.org/10.5555/3635637.3662998

[17] Paul Muller, Romuald Elie, Mark Rowland, Mathieu Laurière, Julien Pérolat, Sarah Perrin, Matthieu Geist, Georgios Piliouras, Olivier Pietquin, and Karl Tuyls. 2022. Learning Correlated Equilibria in Mean-Field Games. *CoRR* abs/2208.10138 (2022). https://doi.org/10.48550/arXiv.2208.10138 arXiv:2208.10138

[18] Paul Muller, Mark Rowland, Romuald Elie, Georgios Piliouras, Julien Pérolat, Mathieu Laurière, Raphaël Marinier, Olivier Pietquin, and Karl Tuyls. 2022. Learning Equilibria in Mean-Field Games: Introducing Mean-Field PSRO. In *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*, Piotr Faliszewski, Viviana Mascardi, Catherine Pelachaud, and Matthew E. Taylor (Eds.). International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS), 926–934. https://www.ifaamas.org/Proceedings/aamas2022/pdfs/p926.pdf

[19] Sarah Perrin, Mathieu Laurière, Julien Pérolat, Matthieu Geist, Romuald Élie, and Olivier Pietquin. 2021. Mean Field Games Flock! The Reinforcement Learning Way. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, Zhi-Hua Zhou (Ed.). ijcai.org, 356–362. https://doi.org/10.24963/ijcai.2021/50

[20] Sarah Perrin, Julien Pérolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. 2020. Fictitious play for mean field games: Continuous time analysis and applications. *Advances in neural information processing systems* 33 (2020), 13199–13213.

[21] Giorgia Ramponi, Pavel Kolev, Olivier Pietquin, Niao He, Mathieu Lauriere, and Matthieu Geist. 2023. On Imitation in Mean-field Games. In *Thirty-seventh Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=RPFd3D3P3L

[22] Jiaming Song, Hongyu Ren, Dorsa Sadigh, and Stefano Ermon. 2018. Multi-Agent Generative Adversarial Imitation Learning. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett (Eds.). 7472–7483. https://proceedings.neurips.cc/paper/2018/hash/240c945bb72980130446fc2b40fbb8e0-Abstract.html

[23] Kevin Waugh, Brian D. Ziebart, and J. Andrew Bagnell. 2013. Computational Rationalization: The Inverse Equilibrium Problem. *CoRR* abs/1308.3506 (2013). arXiv:1308.3506 http://arxiv.org/abs/1308.3506

[24] Fan Yang, Alina Vereshchaka, Changyou Chen, and Wen Dong. 2020. Bayesian Multi-type Mean Field Multi-agent Imitation Learning. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (Eds.). https://proceedings.neurips.cc/paper/2020/hash/19eca5979ccbb752778e6c5f090dc9b6-Abstract.html

[25] Jiachen Yang, Xiaojing Ye, Rakshit Trivedi, Huan Xu, and Hongyuan Zha. 2018. Learning Deep Mean Field Games for Modeling Large Population Behavior. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=HktK4BeCZ

[26] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. 2018. Mean Field Multi-Agent Reinforcement Learning. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer G. Dy and Andreas Krause (Eds.). PMLR, 5567–5576. http://proceedings.mlr.press/v80/yang18d.html

[27] Lantao Yu, Jiaming Song, and Stefano Ermon. 2019. Multi-Agent Adversarial Inverse Reinforcement Learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 7194–7201. http://proceedings.mlr.press/v97/yu19e.html