Imitation Learning for Mean Field Games with Correlated Equilibria

Zhiyu Zhao
Institute of Automation
Renyuan Xu
University of Southern California
Haifeng Zhang
Institute of Automation
Jun Wang
University College London
Yaodong Yang (yaodong.yang@pku.edu.cn)
Peking University

Research Article

Keywords: Mean field game, Imitation learning, Correlated equilibrium, Adversarial learning

Posted Date: June 30th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3108515/v1

Additional Declarations: No competing interests reported.

Imitation Learning for Mean Field Games with Correlated Equilibria

Zhiyu Zhao[1,2], Renyuan Xu[3], Haifeng Zhang[1,2], Jun Wang[4], Yaodong Yang[5*]

[1]Institute of Automation, Chinese Academy of Sciences, No. 95 Zhongguancun East Road, Beijing, 100190, Beijing, China.
[2]School of Artificial Intelligence, University of Chinese Academy of Sciences, No. 1 Yanqihu East Road, Beijing, 100049, Beijing, China.
[3]Department, University of Southern California, 3650 McClintock Ave, Los Angeles, 90089, CA, USA.
[4]Computer Science, University College London, Gower Street, London, WC1E 6BT, England, UK.
[5]Institute of Artificial Intelligence, Peking University, No. 5 Yiheyuan Road, Beijing, 100871, Beijing, China.

*Corresponding author(s). E-mail(s): yaodong.yang@pku.edu.cn;
Contributing authors: zhaozhiyu2022@ia.ac.cn; renyuanx@usc.edu;
haifeng.zhang@ia.ac.cn; jun.wang@cs.ucl.ac.uk;

Abstract

Imitation learning (IL) is a powerful approach for acquiring optimal policies from demonstrated behaviors. However, applying IL to a large group of agents is arduous due to the exponential surge in interactions with an increase in population size. Mean Field Theory provides an efficient tool for analyzing multi-agent problems by gathering information at the population level. Although the approximation is tractable, restoring mean field Nash equilibria (MFNE) from demonstrations is challenging. Furthermore, many real-world problems, including traffic network equilibrium induced by public routing recommendations and pricing equilibrium of goods on E-commerce platforms, cannot be explained by the classic MFNE concept. In both cases, the intervention of the platform introduces correlation devices to the equilibrium. To address this issue, we propose a novel solution concept called Adaptive Mean Field Correlated Equilibrium (AMFCE) that generalizes MFNE. We establish a framework based on IL and AMFCE that recovers the AMFCE policy from real-world demonstrations. Our framework

characterizes mean-field evolution using signatures from the rough path theory, and it has the significant benefit of recovering both the equilibrium policy and correlation device from data. We test our framework against state-of-the-art IL algorithms for mean field games (MFGs) on several tasks, including a real-world traffic flow prediction problem. Our results demonstrate the effectiveness of our proposed method and its potential for predicting and explaining large population behavior under correlated signals.

Keywords: Mean field game, Imitation learning, Correlated equilibrium, Adversarial learning

1 Introduction

Imitation Learning (IL) has gained traction as a powerful approach for learning desired behavior through expert demonstrations [11]. However, in scenarios involving a large population of agents, the exponential increase in interactions and curse of dimensionality render existing IL algorithms inadequate. This limitation has practical implications for real-world applications such as traffic management [3], ad auctions [9], and social behaviors between game bots and humans [13]. Mean field theory provides a viable alternative, offering an analytically feasible and practically efficient approach for analyzing multi-agent games in systems with homogeneous agents [9, 28]. In mean field game (MFG) settings, the states of the entire population can be effectively summarized into an empirical state distribution due to homogeneity, reducing the problem to a game between a representative agent and an empirical distribution.

Current literature on mean-field IL assumes that expert demonstrations are sampled from the classic mean field Nash equilibrium (MFNE) [27, 6]; however, this framework is not general enough to accommodate many real-world situations where external and correlated signals influence the behavior of the entire population. For example, this occurs when drivers receive routing recommendations from Google Maps or Apple Maps, or when individual sellers receive recommendations from an E-commerce platform on setting prices for their products. In both cases, a mediator or coordinator recommends decisions, but individual agents seeking greedy decisions could deviate from the recommendation if they find a better option based on available information, introducing correlations among the behaviors of individual agents.

Therefore, a more general equilibrium concept is needed before we take a step further to learn from expert demonstrations. Inspired by the concept of correlated equilibrium (CE) for stateless game [2], there are recent developments on mean field correlated equilibrium (MFCE) with state dynamics. Campi and Fischer assume that a mediator recommends the same stochastic policy to the entire population, resulting in a limited equilibrium set which is the same as the classic MFNE [4]. In addition, it is often more practical for the mediator to recommend an action rather than a stochastic policy to individuals (see the traffic routing and e-commerce examples). Muller et al. assume that the mediator recommends a deterministic policy (sampled from some distribution over the deterministic policy space) to each individual [20].

This formulation is also rather limited in terms of describing the behaviors of many real-world applications and enabling sufficient flexibility of the population behavior.

Both MFCE concepts assume a fixed correlated signal, which is a recommended policy at the start of the game, making it time-independent. However, this assumption is impractical as real-world situations such as routing recommendations in traffic management and E-commerce pricing depend on real-time variables like weather and supply-demand imbalances. A more general and practical setting involves establishing a framework where the mediator can sample a stochastic policy based on some time-dependent signals and recommend actions for each individual. This exact framework is explored in this paper. (For a concrete example demonstrating the greater generality of our equilibrium concept over that proposed by Muller et al. [20], refer to Appendix C.)

In light of the limitations observed in the existing MFCE concepts and mean-field IL methods, we introduce a novel MFCE framework dubbed as the "Adaptive Mean Field Correlated Equilibrium (AMFCE)." This approach incorporates the notion of time-varying correlated signals to enable individual agents to flexibly adjust their beliefs regarding the unobserved correlated signal. Building upon the AMFCE framework, we introduce a new IL framework, namely the "Correlated Mean Field Imitation Learning (CMFIL)" approach. This method allows for the recovery of not just the policy, but also the correlation device, which is the distribution used to sample the correlated signal.

The generality and flexibility of AMFCE allow CMFIL framework to predict and explain more real-world scenarios. Our framework has the following important and novel ingredients:

- *Ingredient One: Novel MFCE concept with time-dependent correlated signals and adaptive belief updates from individual agents.* In this paper, we propose a new MFCE framework (called AMFCE) that the mediator recommends an action sampled from a stochastic policy for each agent at every time step. This is a more general and flexible framework compared to previous works on the MFCE [20, 4]. We prove the existence of AMFCE solution under mild conditions and prove that MFNE is a subclass of AMFCE.
- *Ingredient Two: Using signatures from rough path theory to efficiently represent mean-field evolution.* In practice, mean field information is often unattainable and approximating it through its empirical distribution can prove to be computationally expensive. As a solution, we leverage signatures from the rough path theory to represent the mean-field evolution in a computationally efficient manner that can be seamlessly integrated with neural network training architectures. By employing this technique, policies can be recovered without requiring access to the underlying mean field.

To the best of our knowledge, this paper presents the first instance where the MFCE framework is investigated while incorporating a correlation device that provides time-varying recommendations and allows for adaptive belief updates by individual agents.

In addition, the performance of our proposed framework is demonstrated by comparing it with the state-of-the-art imitation learning algorithms for MFGs on various tasks, including a real-world traffic flow prediction problem. Our experimental results show that our framework outperforms the baseline in all tasks. Moreover, our framework is also suitable for solving MFNE as it is a subclass of AMFCE.

2 Related work

Our research contributes to the extensive body of multi-agent imitation learning (MAIL). One line of MAIL research has extended single-agent IL algorithms to the Markov game [24, 29, 12]. However, these algorithms face scalability challenges due to the exponential increase in agent interactions as the number of agents increases. To address this challenge, Yang et al. proposed a multi-type mean field approximation that approximates Nash equilibrium in Markov games [26]. However, this approach does not consider the MFG and MFNE, and it decouples the interdependence between mean field flow and policy.

Yang et al. proposed a method to infer the MFG model through inverse reinforcement learning (IRL), assuming that the equilibrium behind the demonstrations is the mean field social optimum (MFSO), which only holds for fully cooperative settings [27]. 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]. However, these methods do not consider the correlation between agents.

Campi et al. proposed the MFCE concept, which introduces a mediator that recommends the same stochastic policy to the entire population, while Muller et al. proposed the MFCE concept assuming that the mediator recommends a deterministic policy to each agent. However, both concepts rely on the assumption that time-independent correlated signals are realized at the beginning of the game, which does not hold in many real-world scenarios. For instance, central platforms such as traffic networks and E-commerce marketplaces, as mentioned in Section 1 introduce correlations that vary over time, making the MFCE concepts impractical for modeling such settings.

To address this limitation, we propose a novel mean field equilibrium concept, AMFCE. This concept allows the correlated signals provided by the mediator to be time-dependent, making it more flexible and general than the existing MFCE concepts. This flexibility accommodates real-world scenarios with varying correlated signals introduced by the mediator.

3 Preliminary

3.1 Classic mean field Nash equilibrium

This subsection introduces the classic framework of MFG and the concept of MFNE. 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 $\mathcal{X}$ and denote $\mathcal{T} = \{0, 1, \cdots, T\}$ as a set of time indexes. At time t, after the representative player chooses her action a_t according to some measurable policy $\pi_t : \mathcal{S} \to \mathcal{P}(\mathcal{A})$, she will receive a

deterministic reward $r(s_t, a_t, \mu_t)$ and her state will evolve according to current state $s_t \in \mathcal{S}$ and $P(\cdot|s_t, a_t, \mu_t)$, where $\mu_t \in \mathcal{P}(\mathcal{S})$ represents the population state distribution and $\mathcal{S}$ is finite. Intuitively, $\mu_t(s) = \lim_{N\to\infty} \frac{\sum_{i=1}^{N} \mathbb{1}_{\{s_t^i = s\}}}{N}$ can be viewed as the limit of the empirical distribution of an homogeneous N-agent game 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). Here $P : \mathcal{S} \times \mathcal{A} \times \mathcal{P}(\mathcal{S}) \to \mathcal{P}(\mathcal{S})$ is the transition kernel for the state dynamics.

For fixed mean-field information $\boldsymbol{\mu} = \{\mu_t\}_{t=0}^T$, the objective of the representative agent is to solve the following decision-making problem over all admissible policies $\boldsymbol{\pi} = \{\pi_t\}_{t=0}^T$:

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

The Mean-field Nash Equilibrium (MFNE) is defined as the following.

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

1. *(Single player side) For any policy* $\boldsymbol{\pi}$*, any time index* $t \in \mathcal{T}$*, and any initial state* $s \in \mathcal{S}$*,* $V_t(s, \boldsymbol{\pi}^\star, \boldsymbol{\mu}^\star) \geq V_t(s, \boldsymbol{\pi}, \boldsymbol{\mu}^\star)$ *.*
2. *(Population side)* $\{\mu_t^*\}_{t=0}^T$ *satisfies* $\mu_t^*(\cdot) = \sum_{s\in\mathcal{S}, a\in\mathcal{A}} P(\cdot|s, a, \mu_{t-1}^*)\pi_{t-1}^*(a|s)\mu_{t-1}^*(s)$ *with initial condition* $\mu_0^* = \mu_0$*.*

The single player side condition captures the optimality of $\boldsymbol{\pi}^\star$, when the population side 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 population state sequence $\boldsymbol{\mu}^\star := \mu_{t\,t=0}^{\star T}$.

3.2 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, the reward function is unknown, but a set of demonstration trajectories under expert policy π^E are provided. The goal of imitation learning is to recover the expert policy π^E using the demonstration trajectories.

IRL is a subclass of IL and it solves the problem in two steps. It first finds a reward function $\tilde{r}$ that rationalizes the expert policy π^E.

$$\tilde{r} = \max_r \big(\min_\pi -H(\pi) + J(\pi)\big) - J(\pi^E)$$

Then a recovered policy is extracted 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 it is trained through the Generative Adversarial Network (GAN). Note that GAIL extracts a policy directly from the expert demonstrations and does not aim at recovering a reward function. In particular, it introduces a discriminator D_ω to differentiate the state-action pairs from π^E and other policies. The recovered policy π_θ, parameterized by θ, plays the role of a generator. It aims at generating state-action pairs that are difficult for D_ω to differentiate. The target function of GAIL is thus defined as

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

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 $s_{t+1} \sim P(\cdot|s_t, a_t)$, $a_t \sim \pi^E(\cdot|s_t)$, $s_0 \sim \mu_0$.

4 Problem formulation

This section introduces a novel adaptive mean-field correlated equilibrium (AMFCE) framework and establishes the existence of equilibria solutions under mild conditions. Our analysis demonstrates that the solution set of AMFCE is richer than the well-known MFNE.

4.1 Adaptive mean field correlated equilibrium (AMFCE)

To incorporate the correlations introduced by central platforms in the traffic network and E-commerce marketplace examples mentioned in Section 1, we introduce a mediator (or central agent) who samples a correlated signal $z_t \in \mathcal{Z}$ at each time t. Here, $\mathcal{Z}$ denotes a finite signal space, and z_t may represent some global conditions such as the weather on day t for the traffic network example or the supply-demand imbalance in month t for the E-commerce marketplace example. Before discussing the AMFCE, we first introduce the concepts of behavioral policy and correlation device.

Definition 2. *For each time t, the behavioral policy $\pi_t : \mathcal{Z} \times \mathcal{S} \to \Delta(\mathcal{A})$ maps from the signal and state spaces to the simplex over the action space. Given the correlated signal $z \in \mathcal{Z}$ and an action $a \sim \pi_t(\cdot|s, z)$ will be independently sampled as a private recommendation for each agent at state s.*

Definition 3. *The per-step correlation device $\rho_t \in \Delta(\mathcal{Z})$ is a publicly known distribution over the space of correlated signal, from which the mediator will sample the correlated signal at time step t. Denote $\boldsymbol{\rho} = \{\rho_t\}_{t=0}^T$ as correlation device over the entire horizon.*

At each time step t, a correlated signal z_t is sampled from the per-step correlation device ρ_t. Subsequently, a recommended action a_t is sampled independently from the behavior policy $\pi_t(\cdot|s_t, z_t)$ and sent to each agent at state s_t. Importantly, this recommended action a_t is *private* and only available to the agent. Mathematically, denote $\mathcal{I}_t = \{\rho_t, a_t, \pi_t(\cdot,\cdot,\cdot), s_t, z_{t-1}, \mu_{t-1}\}$ as the information available to the agent

at the beginning of step t. And $\mathcal{I}_0 = \{\rho_0, a_0, \pi_0(\cdot,\cdot,\cdot), s_0\}$. Note that the agent only observes the functional form of π_t but *can not observe* the correlated signal z_t nor the recommended actions for other agents. Based on the information $\mathcal{I}_t$, the agent will take an action a'_t (which may be different from the mediator's recommendation), and then the agent at state s_t will transit to the next state according to distribution $P(\cdot|s_t, a'_t, \mu_t) \in \mathcal{P}(\mathcal{S})$ given current mean field μ_t, which follows:

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

This implies that, given μ_{t-1} and π_{t-1}, μ_t is fully determined by z_{t-1}. After receiving the recommendation action a_t, the agent can *predict* the correlated signal by

$$\rho_t^{\text{pred}}(z_t = z|\mathcal{I}_t) = \frac{\rho_t(z)\pi_t(a_t|s_t, z)}{\sum_{z'\in\mathcal{Z}} \rho_t(z')\pi_t(a_t|s_t, z')}. \tag{2}$$

Based on the available information $\mathcal{I}_t$ at time t, the agent can then update the prediction on the mean field distribution of the next time-step for each possible signal z:

$$\mu_{t+1}^{\text{pred}}(\cdot|\mathcal{I}_t, z) = \sum_{a\in\mathcal{A}}\sum_{s\in\mathcal{S}} \mu_t(s)P(\cdot|s, a, \mu_t)\pi_t(a|s, z) := \Phi(\mu_t, \pi_t, z). \tag{3}$$

The Q function $Q_t^{\boldsymbol{\pi}}(s, a, \mu, z; \boldsymbol{\pi}')$ for individual agent is defined as follows:

$$\begin{aligned} Q_t^{\boldsymbol{\pi}}(s, a, \mu, z; \boldsymbol{\pi}') =& r(s, a, \mu) \\ &+ \gamma\mathbb{E}_{\boldsymbol{\pi}'}\left[\sum_{i=t+1}^{T} \gamma^{i-t-1} r(s_i, a_i, \mu_i)\middle| (s_t, a_t, \mu_t, z_t) = (s, a, \mu, z)\right], \end{aligned}$$

where $\mathbb{E}_{\boldsymbol{\pi}'}$ is the expectation taken with respect to $z_i \sim \rho_i(\cdot)$, $a_i \sim \pi_i(\cdot|s_i, z_i)$, $s_{i+1} \sim P(\cdot|s_i, a_i, \mu_i)$, $\forall i \in \{t+1, t+2, \cdots, T\}$. We can verify that the Q function satisfies the following Bellman equation:

$$\begin{aligned} Q_t^{\boldsymbol{\pi}}(s, a, \mu, z; \boldsymbol{\pi}') =& r(s, a, \mu) \\ &+ \gamma\mathbb{E}\left[Q_{t+1}^{\boldsymbol{\pi}}(s', a', \Phi(\mu, \pi'_t, z), z'; \boldsymbol{\pi}')\middle|(s_t, a_t, \mu_t, z_t) = (s, a, \mu, z)\right], \end{aligned} \tag{4}$$

where the expectation is taken with respect to $z' \sim \rho_{t+1}(\cdot), s' \sim P(\cdot|s, a, \mu), a' \sim \pi_{t+1}(\cdot|s, z')$.

Similarly, we define the optimal Q-function $Q_t^*(s, a, \mu, z; \boldsymbol{\pi}')$ as the Q function associated with the optimal individual policy $\boldsymbol{\pi}^*$ given population behavior $\boldsymbol{\pi}'$. It is easy to show that Q^* satisfies the following Bellman equation:

$$Q_t^*(s, a, \mu, z; \boldsymbol{\pi}') = r(s, a, \mu)$$

$$+\gamma \max_{a' \in \mathcal{A}} \mathbb{E}\left[Q^*_{t+1}(s', a', \Phi(\mu, \pi'_t, z), z'; \boldsymbol{\pi}') \middle| (s_t, a_t, \mu_t, z_t) = (s, a, \mu, z) \right], \tag{5}$$

where the expectation is taken with respect to $z' \sim \rho_{t+1}(\cdot), s' \sim P(\cdot \mid s, a, \mu_t)$.

It is worth noting that if the policy of population $\boldsymbol{\pi}'$ is fixed, $Q^*_T(s, a, \mu, z; \boldsymbol{\pi}') \geq Q^{\boldsymbol{\pi}}_T(s, a, \mu, z; \boldsymbol{\pi}')$ for any $\boldsymbol{\pi}$. Then by induction, it holds that $Q^*_t(s, a, \mu, z; \boldsymbol{\pi}') \geq Q^{\boldsymbol{\pi}}_t(s, a, \mu, z; \boldsymbol{\pi}')$ for all $t \in \mathcal{T}$.

To introduce the concept of AMFCE, we define the set of swap function $\mathcal{U} := \{u : \mathcal{A} \to \mathcal{A}\}$, namely u a function that modifies an action a to an action $u(a)$. Let $\Delta_t(s, \mu, u; \boldsymbol{\pi}, \boldsymbol{\rho}) = \mathbb{E}\big[Q^{\boldsymbol{\pi}}_t(s, u(a), \mu, z; \boldsymbol{\pi}) - Q^{\boldsymbol{\pi}}_t(s, a, \mu, z; \boldsymbol{\pi})\big]$, $u \in \mathcal{U}$ denote the margin of Q function of that agent takes action $u(a)$ when a recommendation a is provided by the mediator, where the expectation is taken with respect to $z \sim \rho_t(\cdot), a \sim \pi_t(\cdot|s, z)$.

Definition 4. *The profile $(\boldsymbol{\pi}^\star, \boldsymbol{\rho})$ is composed of the time-varying stochastic policy $\boldsymbol{\pi}^\star = \{\pi^\star_t\}_{t=0}^T$ and the correlation device $\boldsymbol{\rho}$ is an adaptive mean field correlated equilibrium (AMFCE) if*

- *(Single agent side) No agent has an incentive to unilaterally deviate from the recommendation action after predicting the z by (2), i.e.*

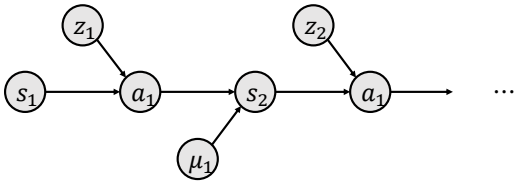
$$\Delta_t(s, \mu^\star_t, u; \boldsymbol{\pi}^\star, \boldsymbol{\rho}) \leq 0, \quad \forall u \in \mathcal{U}, \forall s \in \mathcal{S}, \forall t \in \mathcal{T}.$$

- *(Population side) $\{\mu^*_t\}_{t=0}^T$ satisfies $\mu^*_t(\cdot) = \sum_{s \in \mathcal{S}, a \in \mathcal{A}} P(\cdot|s, a, \mu^*_{t-1}) \pi^*_{t-1}(a|s, z_{t-1}) \mu^*_{t-1}(s)$ given the correlated signals $\{z_t\}_{t=0}^T$ and with initial condition $\mu^*_0 = \mu_0$.*

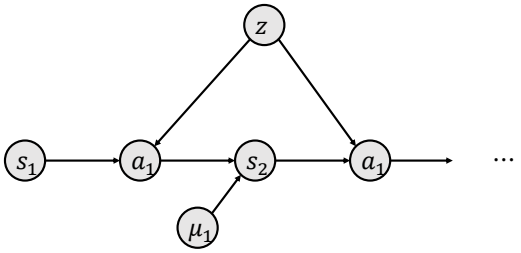
Below is an example that illustrates the concept of AMFCE and highlights its differences from pre-existing MFCE concepts.

Example 1. *The traffic network in Figure 2 consists of three cities. Tourists located in city A are expected to visit either city B or C during a two-day vacation period.. These tourists rely on an online mapping application that suggests either city L or R based on real-time weather information z. This scenario can be modeled as a mean field game with state space $\mathcal{S} = \{C, L, R\}$ and the action space $\mathcal{A} = \{L, R\}$. The initial mean field is given by $\mu_0(C) = 1$, and the reward function is defined as $r(s, a, \mu) = \mathbb{1}_{\{s=L\}} \mu(L) + \mathbb{1}_{\{s=R\}} \mu(R)$. Due to the possibility of unexpected road closures, the environment transition is non-deterministic. The environment transition is as following:*

$$\begin{aligned}
&P(s_1 = R \mid s_0 = C, a = R) = 1, && P(s_1 = L \mid s_0 = C, a = R) = 0,\\
&P(s_1 = R \mid s_0 = C, a = L) = 0, && P(s_1 = L \mid s_0 = C, a = L) = 1,\\
&P(s_1 = R \mid s_0 = L, a = R) = \frac{3}{4}, && P(s_1 = L \mid s_0 = L, a = R) = \frac{1}{4},\\
&P(s_1 = L \mid s_0 = L, a = L) = 1, && P(s_1 = R \mid s_0 = L, a = L) = 0,\\
&P(s_1 = L \mid s_0 = R, a = L) = \frac{3}{4}, && P(s_1 = R \mid s_0 = R, a = L) = \frac{1}{4},\\
&P(s_1 = R \mid s_0 = R, a = R) = 1, && P(s_1 = L \mid s_0 = R, a = R) = 0.
\end{aligned}$$



(a) Structure of AMFCE. In the AMFCE framework, the correlated signals are realized at each time step. After sampling the correlated signal z_t at time step t from the correlation device ρ_t, the action a_t is sampled from the policy $\pi_t(a_t|s_t, z_t)$ for each agent at state s_t as an individual recommendation. Agent can only observe the recommended action. As z_t cannot be realized until time step t, the agent must adaptively update her belief in z_t.



(b) Structure of MFCE. In the MFCE framework, the correlated signal z is realized at the start of the game. A sequence of actions, or a deterministic policy, is then recommended for each agent as an individual recommendation. Consequently, the agent can infer the correlated signal z at the start of the game without the need for adaptive updates to her belief.

Fig. 1: The structures of AMFCE and MFCE.

The following recommendations are given by the online mapping application in an AMFCE. At time step $t \in \mathcal{T} = \{0, 1\}$, a random variable z is sampled from the correlated signal space $\mathcal{Z} = \{0, 1\}$ with equal probabilities, i.e., $\rho_t(z = 0) = \rho_t(z = 1) = 0.5$. The online mapping application recommends an action for each agent based on the observed value of z and the policy π. The policy is defined as follows:

$$\pi(a = L|s = C, z = 0) = 2/3, \qquad \pi(a = R|s = C, z = 0) = 1/3,$$
$$\pi(a = L|s = C, z = 1) = 1/3, \qquad \pi(a = R|s = C, z = 1) = 2/3,$$

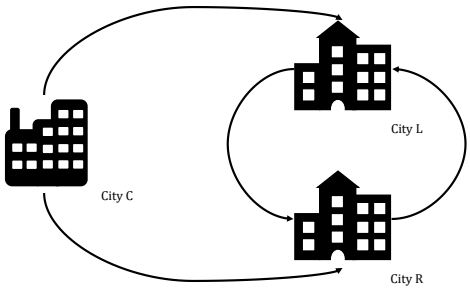


Fig. 2: The traffic network is comprised of three cities, with tourists situated in city A expected to visit either city B or C over a two-day vacation period. However, the transition dynamics are stochastic due to the possibility of unforeseen road closures.

$$\begin{aligned}
&\pi(a = L|s = L, z = 0) = 1, && \pi(a = R|s = L, z = 0) = 0,\\
&\pi(a = L|s = L, z = 1) = 1/9, && \pi(a = R|s = L, z = 1) = 8/9,\\
&\pi(a = L|s = R, z = 0) = 8/9, && \pi(a = R|s = R, z = 0) = 1/9,\\
&\pi(a = L|s = R, z = 1) = 0, && \pi(a = R|s = R, z = 1) = 1.
\end{aligned}$$

It can be verified that tourists have no incentive to deviate from the recommendation, so an AMFCE is achieved.

This example cannot be explained by existing MFCE concepts due to two main reasons. Firstly, the action recommended by the online mapping application (i.e., the city to visit) is determined after the realization of a time-dependent correlated signal z (i.e., real-time weather information), while existing MFCE concepts assume that the correlated signal z is time-independent. Secondly, the recommendation system of the online mapping application for suggesting the next city to visit by each tourist is based on the present location of tourists and correlated weather information. As a result, recommending actions (i.e., the next city to visit) is more common than recommending policies. Therefore, this scenario is not amenable to prevailing MFCE concepts, primarily due to the time-dependent correlated signal and the conventional nature of the recommendation system.

It is important to note that the AMFCE solution is not a classic MFNE. The policies for both AMFCE and MFNE in this example are shown in Figure 3. Furthermore, Corollary 1 demonstrates that all MFNE policies are AMFCE policies.

4.2 Properties of AMFCE

This section focuses on the properties of AMFCE, including the conditions to guarantee the existence and its relationship to classic MFNE.

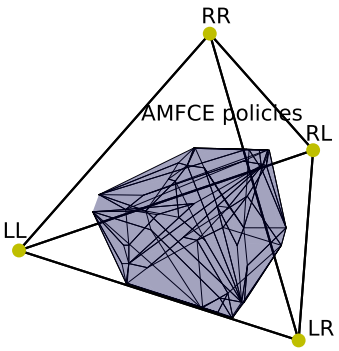


Fig. 3: The visualization of the policies obtained by our proposed AMFCE framework and the MFNE framework for Example 1. The space of policies that are AMFCE policies but not MFNE policies under the correlated device $\rho(z=0)=\rho(z=1)=0.5$ is plotted, while the policies that are both AMFCE policies and MFNE policies are marked with yellow dots. As established in Corollary 1, all MFNE policies are also AMFCE policies, which implies that AMFCE is a more general equilibrium concept than MFNE.

In order to provide the existence of AMFCE solutions, we define the best response operator

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

where the expectation is taken with respect to $z_t \sim \rho_t(\cdot), s_t \sim P(\cdot|s_{t-1},a_{t-1},\mu_{t-1}), a_t \sim \pi'_t(\cdot|s_t,z_t)$, $\mu_t = \Phi(\mu_{t-1},\pi_{t-1},z_{t-1})$. Unless otherwise stated, the expectation $\mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\rho}}$ is taken with respect to $z_t \sim \rho_t(\cdot), s_t \sim P(\cdot|s_{t-1},a_{t-1},\mu_{t-1}), a_t \sim \pi_t(\cdot|s_t,z_t)$, $\mu_t = \Phi(\mu_{t-1},\pi_{t-1},z_{t-1})$. Then the existence of the solution is derived using Kakutani's fixed point theorem [14] with the operator BR. We next provide a sufficient condition for the existence of AMFCE.

Theorem 1. *If the functions $r(s, a, \mu)$ and $P(s'|s, a, \mu)$ are bounded and continuous with respect to μ, there exists an AMFCE solution.*

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

The AMFCE is a more general equilibrium concept than MFNE, which is illustrated in Corollary 1.

Corollary 1. *If $(\boldsymbol{\pi}, \boldsymbol{\mu})$ is an MFNE, then it leads to an AMFCE solution $(\boldsymbol{\pi}, \boldsymbol{\rho})$ with $|\mathcal{Z}| = 1$ and $\rho_t(z) = 1$ for all $t \in \mathcal{T}$ where $z \in \mathcal{Z}$ is the single element in the signal space.*

The proof is deferred to Appendix A.3. This proposition implies that the MFNE is a subset of AMFCE as the example in Example 1 shows that AMFCE may not be an MFNE.

5 Imitation learning for mean field game

This section proposes a new framework based on imitation learning to recover AMFCE from collected expert demonstrations. To emphasize the role of unknown reward function in imitation learning, we use $\mathrm{MFRL}(r, \boldsymbol{\rho})$ to denote the policy of AMFCE under the reward function r and correlation device $\boldsymbol{\rho}$:

$$\mathrm{MFRL}(r, \boldsymbol{\rho}) = \{(\boldsymbol{\pi}, \boldsymbol{\rho}) \in \Pi_{\mathrm{AMFCE}}\} \tag{6}$$

The constraint on the AMFCE set makes finding AMFCE policy challenging. To address this, we provide an equivalent formulation in Proposition 2 and derive a Lagrangian reformulation of (6).

5.1 Correlated mean field imitation learning

We denote $J(\boldsymbol{\pi}, \boldsymbol{\rho}) = \mathbb{E}\left[\sum_{t=0}^T \gamma^t r(s_t, a_t, \mu_t)\right]$, and $\mathcal{R}(a_{0:T}, \boldsymbol{\pi}, \boldsymbol{\rho})$ as the margin of expected return between choosing actions $a_{0:T} := \{a_t\}_{t \in \mathcal{T}}$ and policy $\boldsymbol{\pi}$ under the correlation device $\boldsymbol{\rho}$:

$$\mathcal{R}(a_{0:T}, \boldsymbol{\pi}, \boldsymbol{\rho}) \triangleq \mathbb{E}\Big[\sum_{t=0}^T \gamma^t r(s_t, a_t, \mu_t) \Big| a_{0:T}\Big] - J(\boldsymbol{\pi}, \boldsymbol{\rho}),$$

where the expectation is taken with respect to $z_t \sim \rho_t(\cdot)$, $s_t \sim P(\cdot|a_{t-1}, s_{t-1}, \mu_{t-1})$. And $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}, z_{t-1})$. Then we can get an equivalent constraint of AMFCE.

Proposition 2. *$(\boldsymbol{\pi}, \boldsymbol{\rho})$ is an AMFCE solution if and only if $\mathcal{R}(a_{0:T}, \boldsymbol{\pi}, \boldsymbol{\rho}) \leq 0$, $\forall a_{0:T} \in \mathcal{A}^{\mathcal{T}}$.*

The proof is deferred to Appendix A.4.

Algorithm 1 Correlated mean field imitation learning (CMFIL)

Require: Expert trajectories $\mathcal{D}_E = \{s_0, z_0, a_0, s_1, z_1, a_1, \dots s_T, z_T, a_T\}$
Initial mean field μ_0, The weight of gradient penalty β.
for each iteration **do**
Obtain trajectories from $(\boldsymbol{\pi}, \boldsymbol{\rho})$ by the process: $s_0 \sim \mu_0, a_t \sim \pi^\theta(\cdot|s_t, z_t), s_{t+1} \sim P(\cdot \mid s_t, \mu_t), z_t \sim \rho_t^\phi(\cdot)$;
Approximate μ_t with the signature $\hat{\mu}_t = S(\{z_i\}_{i=0}^t)$ using (11);
for i in $\{0, 1, 2, \dots\}$ **do**
Update ω to increase the objective

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

end for
for t in $\{0, 1, 2, \dots\}$ **do**
Update θ by Actor-Critic algorithm with small step size:

$$\mathbb{E}\Big[\nabla_\theta \rho_t^\phi(z_t) \pi_t^\theta(a_t|s_t, z_t) Q_t^{\boldsymbol{\pi}^\theta}(s_t, a_t, \hat{\mu}_t, z_t; \boldsymbol{\pi})\Big]$$

where the expectation is taken with respect to $s_0 \sim \mu_0$, $a_t \sim \pi^\theta(\cdot|s_t, z_t)$, $s_{t+1} \sim P(\cdot \mid s_t, \mu_t)$, $z_t \sim \rho_t^\phi(\cdot)$;
Update ϕ with (10);
end for
end for
return Policy $\boldsymbol{\pi}^\theta$, correlation device $\boldsymbol{\pi}^\phi$.

5.2 AMFCE inverse reinforcement learning

It is computationally challenging to handle the constraints in the Proposition 2. Compared to the original formulation (6), it is easier to work with a dual representation without constraints:

$$L(\boldsymbol{\pi}, \boldsymbol{\rho}, \lambda, r) \triangleq \sum_{\tau_k \in \mathcal{D}_E} \lambda(\tau_k) \left(\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)\right] - J(\boldsymbol{\pi}, \boldsymbol{\rho})\right) \tag{7}$$

where $\mathcal{D}_E$ is a set of action-signal sequence $\tau_k = \{a_0, z_0, a_1, z_1, a_2, z_2, \cdots, a_T, z_T\}$. We show that (7) captures the difference of expected returns between two policies by selecting λ as follows.

Theorem 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 *if the individual*

policy is $\boldsymbol{\pi}^*$*. Then we have*

$$L(\boldsymbol{\pi}, \boldsymbol{\rho}, \lambda^*_{\boldsymbol{\pi}}, r) = \mathbb{E}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)] - J(\boldsymbol{\pi}, \boldsymbol{\rho})$$

where the expectation is taken with respect to $z_t \sim \rho_t(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $a_t \sim \pi^*_t(\cdot|s_t, z_t)$, $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}, z_{t-1})$.

The proof of Theorem 3 is deferred to Appendix A.5. In the setting of imitation learning, the reward signal is not accessible. To construct a suitable reward function rationalizing the expert policy, we need to define a suitable AMFCE inverse reinforcement learning (AMFCE-IRL) operator which designs a reward to maximize the margin of expected return between expert policy and the other policies:

$$\text{AMFCE-IRL}_\psi(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) = \arg\max_r \Big(-\psi(r) - \max_{\boldsymbol{\pi}} L(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E, \lambda^*_{\boldsymbol{\pi}}, r) \Big), \tag{8}$$

$(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) \in \Pi_{\text{AMFCE}}$ is the AMFCE from which expert demonstrations are sampled. We choose a special regularizer [10]:

$$\psi_{GA}(r) \triangleq \begin{cases} \mathbb{E}[\sum_{t=0}^{T} \gamma^t g(r(s_t, a_t, \mu_t))] & \text{if } r > 0 \\ +\infty & \text{otherwise} \end{cases},$$

where

$$g(x) = \begin{cases} x - \log(1 - e^{-x}) & \text{if } x > 0 \\ +\infty & \text{otherwise} \end{cases}.$$

After getting the reward function $\tilde{r} = \text{AMFCE-IRL}(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E)$, we can characterize the AMFCE policy $\text{MFRL}(\tilde{r}, \boldsymbol{\rho}^E)$ with the learned $\tilde{r}$.

Proposition 4. *The policy* $\boldsymbol{\pi}$ *learned on the reward function recovered by AMFCE-IRL can be characterized as follows:*

$$\begin{aligned} \text{MFRL} \circ \text{AMFCE-IRL}_\psi(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) := \arg\min_{\boldsymbol{\pi}} \max_r J(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) \\ - \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)\right] - \psi_{GA}(r) \end{aligned}$$

where the expectation is taken with respect to $z_t \sim \rho^E_t(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $a_t \sim \pi_t(\cdot|s_t, z_t)$, $\mu_t = \Phi(\mu_{t-1}, \pi^E_{t-1}, z_{t-1})$.

The objective to recover AMFCE is defined as:

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

where D_ω is the discriminator network parameterized with ω, with input (s_t, a_t, μ_t) and output a real number in $(0, 1]$. The first expectation is taken with respect to $z_t \sim \rho_t^E(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $a_t \sim \pi_t(\cdot|s_t, z_t)$, $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}^E, z_{t-1})$. The second expectation is taken with respect to $z_t \sim \rho_t^E(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $a_t \sim \pi_t^E(\cdot|s_t, z_t)$, $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}^E, z_{t-1})$.

The proof is deferred to Appendix A.6. From a theoretical point of view, we assume that neural network D_ω has the capacity to approximate the reward function. Under this assumption, the AMFCE $(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E)$ could be recovered by optimizing the above objective (9). Note that simply applying GAIL to solve AMFCE cannot recover $\boldsymbol{\rho}^E$, so we derive $\boldsymbol{\rho}$ using a gradient descent method (with proof in Appendix A.7):

Proposition 5. *If $\boldsymbol{\rho}^\phi$ is parameterized with ϕ, the gradient to optimize ϕ given state s is*

$$\mathbb{E}_{z\sim\rho_t^\phi(\cdot)}\left[\nabla_\phi \log \rho_t^\phi(z)\mathbb{E}_{a\sim\pi_t(\cdot|s,z)}Q_t^{\boldsymbol{\pi}}(s, a, \mu, z; \boldsymbol{\pi})\right]. \tag{10}$$

Now we propose the imitation learning algorithm for AMFCE (Algorithm 1).

5.3 Representation of the mean field information

As the mean field appears in the input of discriminator $D_\omega(s, a, \mu)$ in (9), it is necessary to find an efficient way to represent the mean field information.

In the Kolmogorov equation (1), the mean field flow $\{\mu_t\}_{t=0}^T$ is deterministic given fixed correlated signal sequence $\{z_t\}_{t=0}^T$ and given the initial mean field distribution μ_0. Therefore, the mean field distribution μ_t can be characterized by $\boldsymbol{z}_{0:t} = \{z_i\}_{i=0}^t$. Motivated by this, we use the signatures of $\boldsymbol{z}_{0:t}$ from the rough path theory [16, 18] to represent the signal sequence and hence to characterize the mean field flow with $\hat{\mu}_t = S(\boldsymbol{z}_{0:t})$. The signatures provide a graduated summary of the path $\boldsymbol{z}_{0:t}$. Therefore, the input of discriminator D_ω in (9) could be replaced with $(s_t, a_t, \hat{\mu}_t)$. It is worth noting that the signature has been recently applied to the field of machine learning to extract characteristic features of sequential data in a non-parametric fashion [19, 21]. The utilization of signatures to encode historical information circumvents the computational burden typically associated with tasks such as training recurrent neural networks. In addition, the training stability can be significantly enhanced since the mapping is invariant.

Definition 5. *Let $\mathbf{x} = \{x_1, \ldots, x_L\}$ with $x_i \in \mathbb{R}^d$, for all i and $L \geq 2$. Denote $f : [0, 1] \to \mathbb{R}^d$ to be the continuous piecewise affine function such that $f(\frac{i-1}{L-1}) = x_i$, $\forall i \in \{1, 2, \ldots, L\}$.*

$$S(f)_{0,1} = (1, M_1, \cdots, M_n, \ldots) \tag{11}$$

where $M_n = \int_{s<s_1<\cdots<s_n<t} \frac{\mathrm{d}f}{\mathrm{d}t}(s_1) \otimes \cdots \otimes \frac{\mathrm{d}f}{\mathrm{d}t}(s_n)\mathrm{d}t_1 \cdots \mathrm{d}t_n$.

The signature of the path $\mathbf{x}$ is defined to be $S(f)_{0,1}$, denoted as $S(\mathbf{x})$.

Signature of sequential data includes infinite terms as shown in the (11), but fortunately, terms M_n enjoy factorial decay. In practice we select the first n terms of the signature without losing crucial information of the data [15].

Table 1: Results for numerical tasks.

Task	Log Loss	CMFIL	MFIRL	MFAIRL
Squeeze with $T = \{0, 1, 2\}$	$\pi_0(\cdot \mid s = \cdot, z = 0)$	**0.643 (0.000)**	1.450 (2.857)	4.064 (0.879)
	$\pi_0(\cdot \mid s = \cdot, z = 1))$	0.647 (0.003)	3.245 (1.650)	4.144 (0.629)
	$\pi_1(\cdot \mid s = \cdot, z = 0)$	**0.020 (0.001)**	1.072 (2.229)	6.934 (4.447)
	$\pi_1(\cdot \mid s = \cdot, z = 1)$	0.045 (0.005)	7.871 (4.368)	1.027 (1.279)
Squeeze with $T = \{0, 1\}$	$\pi(\cdot \mid s = C, z = 0)$	**0.648 (0.002)**	3.828 (1.582)	4.067 (0.088)
	$\pi(\cdot \mid s = C, z = 1)$	**0.638 (0.001)**	2.009 (1.191)	10.074 (0.174)
RPS	$\pi(\cdot \mid s = C, z = 0)$	**1.083 (0.000)**	7.127 (0.753)	3.221 (1.330)
Flock	$\pi(\cdot \mid s = \cdot, z = 0)$	0.002 (0.000)	5.591 (0.869)	12.430 (2.759)
	$\pi(\cdot \mid s = \cdot, z = 1)$	**0.016 (0.003)**	11.687 (1.158)	13.042 (1.533)
	$\pi(\cdot \mid s = \cdot, z = 2)$	**0.045 (0.009)**	7.500 (3.955)	10.065 (5.074)
	$\pi(\cdot \mid s = \cdot, z = 4)$	**0.026 (0.003)**	3.847 (3.967)	9.312 (4.711)
Task	Log Loss	Logistic Regression	Multinomial	MaxEnt ICE
Squeeze with $T = \{0, 1, 2\}$	$\pi_0(\cdot \mid s = \cdot, z = 0)$	4.484 (0.054)	0.686 (0.002)	-
	$\pi_0(\cdot \mid s = \cdot, z = 1)$	**0.000 (0.000)**	2.577 (0.149)	-
	$\pi_1(\cdot \mid s = \cdot, z = 0)$	7.091 (0.107)	0.282 (0.087)	-
	$\pi_1(\cdot \mid s = \cdot, z = 1)$	10.638 (0.163)	**0.001 (0.001)**	-
Squeeze with $T = \{0, 1\}$	$\pi(\cdot \mid s = \cdot, z = 0)$	1.985 (0.165)	0.991 (0.102)	0.946 (0.073)
	$\pi(\cdot \mid s = \cdot, z = 1)$	2.139 (0.169)	2.947 (0.359)	0.648 (0.011)
RPS	π	4.805 (0.131)	5.850 (0.306)	1.537 (0.019)
Flock	$\pi(\cdot \mid s = \cdot, z = 0)$	**0.000 (0.000)**	1.383 (0.004)	-
	$\pi(\cdot \mid s = \cdot, z = 1)$	7.887 (0.031)	1.127 (0.007)	-
	$\pi(\cdot \mid s = \cdot, z = 2)$	18.339 (0.010)	0.951 (0.009)	-
	$\pi(\cdot \mid s = \cdot, z = 4)$	35.253 (0.037)	1.264 (0.011)	-

Table 2: The results of predicted traffic flow for Traffic Network. The metric is log loss of each location.

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

6 Experiments

We evaluate the effectiveness of our algorithm in four environments: Sequential Squeeze, RPS, Flock, and Traffic Flow Prediction.

We compare our proposed CMFIL framework with the existing mean field imitation learning algorithms, MFIRL [6] and MFAIRL [7]. While MFIRL and MFAIRL aim to recover MFNE without considering the correlated signal, we regard the correlated

signal as an extension of the global state for their framework, allowing for a fair comparison between all methods. It is important to note that our proposed method is the first to recover both the policy and the correlated device from data, which is a significant contribution. However, as MFIRL and MFAIRL can recover the policy, we compare the quality of the learned policies for all methods. We focus on the difference between the recovered policy and the ground truth policy, as shown in Table 1 and 2, to evaluate the quality of the policy learned by each method.

We also compare CMFIL with MaxEnt ICE, smoothed multinomial distribution over the joint actions and logistic regression [25]. As MaxEnt ICE is designed to recover correlated equilibrium in matrix game, we only compare CMFIL with MaxEnt ICE on tasks that can be reduced to matrix game, such as RPS and Sequential Squeeze with $\mathcal{T} = \{0, 1\}$. We use the log loss, $\mathbb{E}_{a\sim\pi(\cdot|s,z)}[-\log(\hat{\pi}(a|s, z))]$, to measure the difference between the recovered policy $\hat{\pi}$ and the ground truth π in all tasks. The Appendix B contains more details about the experiments.

6.1 Tasks

We evaluate CMFIL on several tasks: Sequential Squeeze (Squeeze for short), Rock-Paper-Scissors (RPS), Flock and a real-world traffic flow prediction task. The first three experiments are numerical experiments. The traffic flow prediction task is to predict the traffic flow a complex traffic network based on the real world data.

Squeeze: Sequential Squeeze is a game with multi-steps. The purpose to implement this game is to verify the ability to recover expert policy through demonstrations sampled from a multi-step game. We present a discrete version of this problem. The state space is $\mathcal{S} = \{0, 1, 2\}$. Let $\mathcal{A} = \{0, 1\}$ denote the action space. The horizon of the environment is 3. The initial mean field is $\mu_0(s = 2) = 1$. The dynamic of the environment is given by:

$$P(s_{t+1} = 1 \mid s_t = \cdot, a = 1) = \frac{3}{4}, \qquad P(s_{t+1} = 0 \mid s_t = \cdot, a = 1) = \frac{1}{4},$$
$$P(s_{t+1} = 1 \mid s_t = \cdot, a = 0) = \frac{1}{4}, \qquad P(s_{t+1} = 0 \mid s_t = \cdot, a = 0) = \frac{3}{4}$$

The reward function is

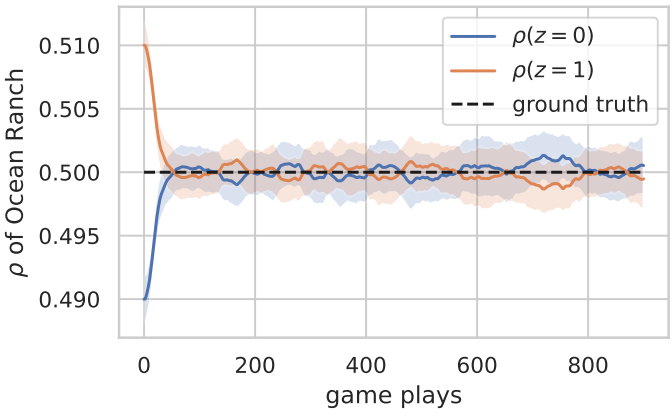
$$r(s, a, \mu) = \mathbb{1}_{\{s=L\}}\mu(L) + \mathbb{1}_{\{s=R\}}\mu(R).$$

The results are shown in Table 1.

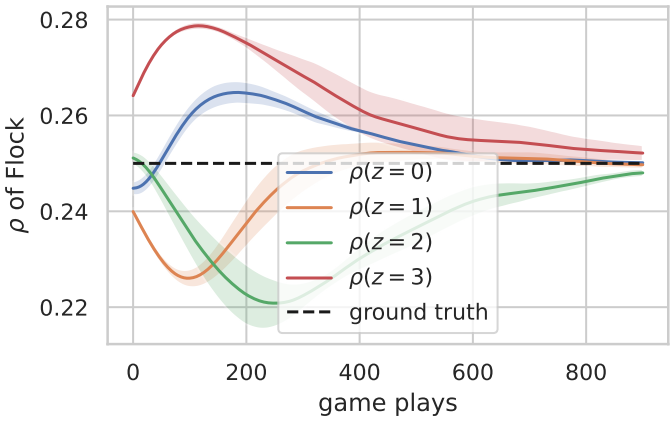
RPS: This experiment is a traditional mean field game task [5, 8, 7]. The dynamic is deterministic:

$$P(s_{t+1} \mid s_t, a_t, \mu_t) = \mathbb{1}_{s_{t+1}=a_t} \tag{12}$$

The state space $\mathcal{S} = \{C, R, P, S\}$ and the action space $\mathcal{A} = \{R, P, S\}$. At the beginning of the game, all the agents are in the state C. The reward function is shown in the



(a) Recovered ρ for Squeeze



(b) Recovered ρ for Flock

Fig. 4: The distribution of correlation device ρ recovered by CMFIL.

following

$$
\begin{aligned}
r(R, a, \mu_t) &= 2 \cdot \mu_t(S) - 1 \cdot \mu_t(P) \\
r(P, a, \mu_t) &= 4 \cdot \mu_t(R) - 2 \cdot \mu_t(S) \\
r(S, a, \mu_t) &= 2 \cdot \mu_t(P) - 1 \cdot \mu_t(R)
\end{aligned}
$$

The demonstrations are sampled from MFNE, and the cardinality of the correlated signal set is one. We use RPS to verify that the algorithm proposed can recover the expert demonstrations sampled from MFNE, which also supports the results in Corollary 1.

Flock: The experiment is based on the movement of fish [22]. In nature, fish spontaneously align their velocity according to the overall movement of the fish school, resulting in a stable movement velocity for the entire school. We simplify this setting by defining a new dynamic as follows:

$$x_{t+1} = x_t + v_t \Delta t$$

The action space $\mathcal{A} = \{0, 1, 2, 3\}$ corresponding to four directions of velocity with unit speed. The reward is

$$f_\beta^{\text{flock}}(x, v, u, \mu) = -\left\| \int_{\mathbb{R}^{2d}} (v - v') \, \mathrm{d}\mu \, (x', v') \right\|^2$$

Traffic Flow Prediction: In the Traffic Flow Prediction task, we use the traffic data of London from Uber Movement (https://movement.uber.com/?lang=en-US). This dataset is publicly available, and the data is anonymized. The environment dynamic is deterministic, and the expert demonstrations consist of traffic flow data. 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. Given the large-scale and high-complexity nature of this task, we compare the scalability of CMFIL and MFIRL in this experiment.

6.2 Results

The results for numerical tasks are presented in Table 1. In general, CMFIL outperforms other methods. Supervised learning methods, such as logistic regression and smoothed multinomial distribution, easily overfit and may outperform CMFIL in some metrics but suffer from a higher loss than CMFIL in general. MFIRL and MFAIRL show larger deviations and higher loss than CMFIL in Table 1 and Table 2. The results demonstrate that MFIRL and MFAIRL cannot recover AMFCE, and they cannot handle correlated signals appropriately. Although we consider correlated signals as an extension of state, the rewards recovered by MFIRL and MFAIRL are biased because the ground truth reward is independent of the correlated signal. Moreover, CMFIL adds a regularizer ψ for the reward function to avoid overfitting, which also outperforms MFIRL and MFAIRL in RPS when expert demonstrations are sampled from MFNE. MaxEnt ICE performs poorly because it has a limited reward function class by assuming a linear reward structure. Figure 4 illustrates that CMFIL can recover the correlation device with a fast convergence speed.

We use unity to visualize the tasks of Squeeze and Flock in this site https://sites.google.com/view/mean-field-imitation-learning/.

7 Conclusion

We proposed AMFCE, a new equilibrium concept in the MFG, which is better suited for real-world scenarios where the behavior of the entire population is influenced by external and correlated signals. We proved that AMFCE solution exists under mild

conditions and classic MFNE is a special case of AMFCE. We then developed a novel theoretical framework based on IL (CMFIL) to recover the AMFCE policy from demonstrations. To facilitate efficient computation, we adopted signatures from rough path theory to represent mean-field evolution. Finally, we evaluated CMFIL on several tasks, including one from the real world. Our experimental results showed that CMFIL outperforms state-of-the-art imitation learning algorithms for MFGs in all the experiments. These results highlight the potential of CMFIL to predict and explain large population behavior under correlated signals.

Appendix A Proof

A.1 Proof of Bellman Equation

Proof.

$$
\begin{aligned}
Q_t^{\boldsymbol{\pi}}(s,a,\mu,z;\boldsymbol{\pi}') =& r(s,a,\mu)+ \\
&\gamma \mathbb{E}_{\boldsymbol{\pi}'}\left[\sum_{i=t+1}^{T}\gamma^{i-t-1} r(s_i,a_i,\mu_i)\middle|(s_t,a_t,\mu_t,z_t)=(s,a,\mu,z)\right] \\
=& r(s,a,\mu)+\gamma \mathbb{E}_{\boldsymbol{\pi}'}\Big[r(s_{t+1},a_{t+1},\Phi(\mu,\pi'_t,z)) \\
&+\gamma\sum_{i=t+2}^{T}\gamma^{i-t-2} r(s_i,a_i,\mu_i)\Big|(s_t,a_t,\mu_t,z_t)=(s,a,\mu,z)\Big] \qquad \text{(A1)}
\end{aligned}
$$

where $\mathbb{E}_{\boldsymbol{\pi}'}[\sum_{i=k}^{T}\gamma^{i-k} r(s_i,a_i,\mu_i)]$ is the expectation taken with respect to $z_i \sim \rho_i(\cdot)$, $a_i \sim \pi_i(\cdot|s_i,z_i)$, $s_{i+1} \sim P(\cdot|s_i,a_i,\mu_i)$, $\mu_i(\cdot) = \sum_{a\in\mathcal{A}}\sum_{s\in\mathcal{S}}\mu_{i-1}(s)P(\cdot|s,a,\mu_{i-1})\pi'_{i-1}(a|s,z_{i-1})$, $\forall i \in \{t+1,t+2,\cdots,T\}$.

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\pi}'}\Big[r\big(s',a',\Phi(\mu,\pi'_t,z)\big)+\gamma\sum_{i=t+2}^{T}\gamma^{i-t-2} r(s_i,a_i,\mu_i)\Big] \\
=&\mathbb{E}\Big[r\big(s',a',\Phi(\mu,\pi'_t,z)\big) \\
&+\gamma\mathbb{E}_{\boldsymbol{\pi}'}\big[\sum_{i=t+2}^{T}\gamma^{i-t-2} r(s_i,a_i,\mu_i)|(s_{t+1},a_{t+1},\mu_{t+1},z_{t+1})=(s',a',\Phi(\mu,\pi'_t,z),z')\big]\Big] \\
=&\mathbb{E}\Big[Q_{t+1}^{\boldsymbol{\pi}}\big(s',a',\Phi(\mu,\pi'_t,z),z';\boldsymbol{\pi}'\big)\Big] \qquad \text{(A2)}
\end{aligned}
$$

where the outer expectation is taken with respect to $z' \sim \rho_{t+1}(\cdot), s' \sim P(\cdot|s,a,\mu), a' \sim \pi(\cdot|s,z)$. The outer expectation is the conditional expectation given $(s_t,a_t,\mu_t,z_t) = (s,a,\mu,z)$. We omit $(s_t,a_t,\mu_t,z_t) = (s,a,\mu,z)$ for brevity. Combine (A1) and (A2), we get the Bellman equation.

$$
Q_t^{\boldsymbol{\pi}}(s,a,\mu,z;\boldsymbol{\pi}') = r(s,a,\mu)
$$

$$+\gamma\mathbb{E}\Big[Q_{t+1}^{\boldsymbol{\pi}}\big(s',a',\Phi(\mu,\pi_t',z),z';\boldsymbol{\pi}'|(s_t,a_t,\mu_t,z_t)=(s,a,\mu,z)\big)\Big]$$

where expectation is taken with respect to $z'\sim\rho_{t+1}(\cdot), s'\sim P(\cdot|s,a,\mu), a'\sim\pi_t(\cdot|s,z)$. □

A.2 Proof of Theorem 1

Lemma 1. *Policy $\boldsymbol{\pi}'$ is the best response of $\boldsymbol{\pi}$ given $\boldsymbol{\rho}$ if and only if $\sum_{z\in\mathcal{Z}}\rho_t(z)\pi_t'(a|s,z) > 0$ is a sufficient condition of $a\in\arg\max_{a'\in\mathcal{A}}\mathbb{E}_{z\sim\rho_t^{\text{pred}}(\cdot|\mathcal{I}_t)}Q^*(s,a',\mu,z;\boldsymbol{\pi})$, $\forall t\in\mathcal{T}$.*

Proof. We denote

$$\mathcal{Q}_t^{\boldsymbol{\pi}}(s,a,\mu,\mathcal{I}_t;\boldsymbol{\pi})=\mathbb{E}_{z\sim\rho_t^{\text{pred}}(\cdot|\mathcal{I}_t)}Q_t^{\boldsymbol{\pi}}(s,a,\mu,z;\boldsymbol{\pi})$$

and $\mathcal{Q}_t^*(s,a,\mu,\mathcal{I}_t;\boldsymbol{\pi})=\mathbb{E}_{z\sim\rho_t^{\text{pred}}(\cdot|\mathcal{I}_t)}Q_t^*(s,a,\mu,z;\boldsymbol{\pi})$.

If $\boldsymbol{\pi}'$ is the best response of $\boldsymbol{\pi}$, but $\sum_{z\in\mathcal{Z}}\rho_t(z)\pi_t'(a|s,z)>0$ is not sufficient condition of $a\in\arg\max_{a'\in\mathcal{A}}\mathcal{Q}_t^*(s,a,\mu,\mathcal{I}_t;\boldsymbol{\pi})$. Then there exists $t\in\mathcal{T}$, such that $\sum_{z\in\mathcal{Z}}\rho_t(z)\pi_t'(a|s,z)>0$, while $a\notin\arg\max_{a'\in\mathcal{A}}\mathcal{Q}_t^*(s,a',\mu,\mathcal{I}_t;\boldsymbol{\pi})$.

If $\boldsymbol{\pi}$ and $\boldsymbol{\rho}$ are fixed, the mean field is also fixed. Finding the best response of $\boldsymbol{\pi}$ is equivalent to solving an MDP. Then the expected return is $\mathbb{E}\left[\mathcal{Q}_0^{\boldsymbol{\pi}'}(s_0,a_0,\mu_0,\mathcal{I}_0;\boldsymbol{\pi})\right]$, where the expectation is taken with respect to $z\sim\rho_0(\cdot)$, $s_0\sim\mu_0$, $a_0\sim\pi_0'(\cdot|s_0,z_0)$. We assume that there exists $\boldsymbol{\pi}^*$ such that $\sum_{z\in\mathcal{Z}}\rho_t(z)\pi_t^*(a|s,z)>0$ is sufficient condition of $a\in\arg\max_{a'\in\mathcal{A}}\mathcal{Q}_t^*(s,a,\mu,\mathcal{I}_t;\boldsymbol{\pi})$. The expected return of $\boldsymbol{\pi}^*$ is higher than the expected return of $\boldsymbol{\pi}'$ as suboptimal action is impossible to be sampled in the MDP under the population policy $\boldsymbol{\pi}$, which conflicts with the assumption.

If there exists $\boldsymbol{\pi}'$ such that for all $a\in\arg\max_{a'\in\mathcal{A}}\mathcal{Q}_t^*(s,a,\mu,\mathcal{I}_t;\boldsymbol{\pi})$, we have $\sum_{z\in\mathcal{Z}}\rho_t(z)\pi_t'(a|s,z)>0$ is true. Then $\forall t\in\mathcal{T}$, $\mathbb{E}\left[\mathcal{Q}_0^{\boldsymbol{\pi}'}(s_0,a_0,\mu_0,\mathcal{I}_0;\boldsymbol{\pi})\right]=\max_{\tilde{\boldsymbol{\pi}}}\mathbb{E}\left[\mathcal{Q}_0^{\tilde{\boldsymbol{\pi}}}(s_0,a_0,\mu_0,\mathcal{I}_0;\boldsymbol{\pi})\right]$, where the first expectation is taken with respect to $z\sim\rho_0(\cdot)$, $s_0\sim\mu_0$, $a_0\sim\pi_0'(\cdot|s_0,z_0)$ and the second expectation is taken with respect to $z\sim\rho_0(\cdot)$, $s_0\sim\mu_0$, $a_0\sim\tilde{\pi}_0(\cdot|s_0,z_0)$. So the $\boldsymbol{\pi}'$ is the best response of $\boldsymbol{\pi}$. □

Lemma 2. $\text{BR}(\boldsymbol{\pi};\boldsymbol{\rho})$ *has a closed graph.*

Proof. We assume that $\lim_{n\to\infty}\boldsymbol{\pi}_n=\boldsymbol{\pi}$, $\lim_{n\to\infty}\boldsymbol{\pi}_n'=\boldsymbol{\pi}'$, $\boldsymbol{\pi}_n\in\text{BR}(\boldsymbol{\pi}_n';\boldsymbol{\rho})$, but $\boldsymbol{\pi}\notin\text{BR}(\boldsymbol{\pi}';\boldsymbol{\rho})$. Consequently, there exists $a\in\mathcal{A}$ that $\sum_{z\in\mathcal{Z}}\rho_t(z)\pi_{n,t}(a|s,z)>0, a\in\arg\max_{a'}\mathcal{Q}_t^*(s,a',\mu,\mathcal{I}_t;\boldsymbol{\pi}_n')$, while $a\notin\arg\max_{a'}\mathcal{Q}_t^*(s,a',\mu,\mathcal{I}_t;\boldsymbol{\pi}')$. Let $a^\star=\arg\max_{a'}\mathcal{Q}_t^*(s,a',\mu,\mathcal{I}_t;\boldsymbol{\pi}')$. Let ϵ denote the margin of Q value

$$\mathcal{Q}_t^*(s,a^\star,\mu,\mathcal{I}_t;\boldsymbol{\pi}')-\mathcal{Q}_t^*(s,a,\mu,\mathcal{I}_t;\boldsymbol{\pi}')=\epsilon>0$$

From the continuity of $\mathcal{Q}_t^*(s,a,\mu,\mathcal{I}_t;\boldsymbol{\pi}')=\mathbb{E}_{z\sim\rho_t(\cdot)}Q_t^*(s,a,\mu,z;\boldsymbol{\pi}')$. It is obvious that there exists $N\in\mathbb{N}$ such that $|\mathcal{Q}_t^*(s,a,\mu,\mathcal{I}_t;\boldsymbol{\pi}')-\mathcal{Q}_t^*(s,a,\mu,\mathcal{I}_t;\boldsymbol{\pi}_n')|<\frac{\epsilon}{2}$, $\forall n>N, a'\in\mathcal{A}$.

Then we can induce that

$$
\begin{aligned}
&\mathcal{Q}_t^*(s, a^\star, \mu, \mathcal{I}_t; \boldsymbol{\pi}'_n) - \mathcal{Q}_t^*(s, a, \mu, \mathcal{I}_t; \boldsymbol{\pi}'_n) \\
=&\mathcal{Q}_t^*(s, a^\star, \mu, \mathcal{I}_t; \boldsymbol{\pi}'_n) + \mathcal{Q}_t^*(s, a^\star, \mu, \mathcal{I}_t; \boldsymbol{\pi}') - \mathcal{Q}_t^*(s, a^\star, \mu, \mathcal{I}_t; \boldsymbol{\pi}') + \mathcal{Q}_t^*(s, a, \mu, \mathcal{I}_t; \boldsymbol{\pi}') \\
&- \mathcal{Q}_t^*(s, a, \mu, \mathcal{I}_t; \boldsymbol{\pi}') - \mathcal{Q}_t^*(s, a, \mu, \mathcal{I}_t; \boldsymbol{\pi}'_n) \\
\geq&\mathcal{Q}_t^*(s, a^\star, \mu, \mathcal{I}_t; \boldsymbol{\pi}') - \mathcal{Q}_t^*(s, a, \mu, \mathcal{I}_t; \boldsymbol{\pi}') - |\mathcal{Q}_t^*(s, a^\star, \mu, \mathcal{I}_t; \boldsymbol{\pi}'_n) - \mathcal{Q}_t^*(s, a^\star, \mu, \mathcal{I}_t; \boldsymbol{\pi}')| \\
&- |\mathcal{Q}_t^*(s, a, \mu, \mathcal{I}_t; \boldsymbol{\pi}'_n) - \mathcal{Q}_t^*(s, a, \mu, \mathcal{I}_t; \boldsymbol{\pi}')| \\
>&\epsilon - \frac{\epsilon}{2} - \frac{\epsilon}{2} = 0
\end{aligned}
$$

contradicting $a \in \arg\max_{a'} \mathcal{Q}_t^*(s, a', \mu, \mathcal{I}_t; \boldsymbol{\pi}'_n)$. So $\mathrm{BR}(\boldsymbol{\pi}; \boldsymbol{\rho})$ has a closed graph. □

Lemma 3. $\mathrm{BR}(\boldsymbol{\pi}; \boldsymbol{\rho})$ *is a convex set given* $\boldsymbol{\pi}$.

Proof. We assume that $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2 \in \mathrm{BR}(\boldsymbol{\pi}'; \boldsymbol{\rho})$. From Lemma 1, $\sum_{z \in \mathcal{Z}} \rho_t(z) \pi_{i,t}(a \mid s, z) > 0, a \in \arg\max_{a' \in \mathcal{A}} \mathcal{Q}^*(s, a', \mu, I_t; \boldsymbol{\pi}')$, $\forall t \in \mathcal{T}$, $\forall i \in \{1, 2\}$. Then the convex combination $\boldsymbol{\pi} = \lambda \boldsymbol{\pi}_1 + (1 - \lambda)\boldsymbol{\pi}_2, \lambda \in [0, 1]$ also satisfies the requirements of Lemma 1. Therefore $\boldsymbol{\pi} \in \mathrm{BR}(\boldsymbol{\pi}'; \boldsymbol{\rho})$. $\mathrm{BR}(\boldsymbol{\pi}; \boldsymbol{\rho})$ is a convex set given $\boldsymbol{\pi}$. □

Theorem 1. *If the functions* $r(s, a, \mu)$ *and* $P(s'|s, a, \mu)$ *are bounded and continuous with respect to* μ*, there exists an AMFCE solution.*

Proof. As $\pi_t \in \Delta_{\mathcal{A}}$, in which $\Delta_{\mathcal{A}}$ are simplices with finite dimensions, they are compact. And $\mathrm{BR}(\boldsymbol{\pi}; \boldsymbol{\rho})$ maps to a non-empty set, because the MDP induced by fixed $\boldsymbol{\mu}$ and $\boldsymbol{\rho}$ has an optimal policy. From Lemma 2 and 3, the requirements of Kakutani's fixed point theorem holds for $\mathrm{BR}(\boldsymbol{\pi}; \boldsymbol{\rho})$. By Kakutani's fixed point theorem, there exists a fixed point $\boldsymbol{\pi}^* \in \mathrm{BR}(\boldsymbol{\pi}^*; \boldsymbol{\rho})$. And $\forall u \in \mathcal{U}$, $\forall s \in \mathcal{A}$, $\forall t \in \mathcal{T}$,

$$
\begin{aligned}
\Delta_t(s_t, \mu_t, u; \boldsymbol{\pi}^*, \boldsymbol{\rho}) = \sum_{z \in \mathcal{Z}} \sum_{a \in \mathcal{A}} \rho_t(z) \pi_t^*(a|s, z) \big(& Q_t^{\boldsymbol{\pi}^*}(s_t, u(a), \mu_t, z; \boldsymbol{\pi}^*) \\
& - Q^{\boldsymbol{\pi}^*}(s_t, a, \mu_t, z; \boldsymbol{\pi}^*) \big) \leq 0,
\end{aligned}
$$

where $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}^*, z_t)$. Then $(\boldsymbol{\pi}^*, \boldsymbol{\rho})$ is an AMFCE. □

A.3 Proof of Corollary 1

Corollary 1. *If* $(\boldsymbol{\pi}, \boldsymbol{\mu})$ *is an MFNE, then it leads to an AMFCE solution* $(\boldsymbol{\pi}, \boldsymbol{\rho})$ *with* $|\mathcal{Z}| = 1$ *and* $\rho_t(z) = 1$ *for all* $t \in \mathcal{T}$ *where* $z \in \mathcal{Z}$ *is the single element in the signal space.*

Proof. Assume that $(\boldsymbol{\pi}, \boldsymbol{\mu})$ is an MFNE, so the following condition holds [8]. $\pi_t(a \mid s, z) > 0$ is sufficient condition of $a \in \arg\max_{a' \in \mathcal{A}} Q_t^*(s, a', \mu, z; \boldsymbol{\pi})$. If $z \in \mathcal{Z}$ is the single element in the signal space $\mathcal{Z}$, $\rho_t(z) = 1$ is true for all $t \in \mathcal{T}$. $\sum_z \rho_t(z) \pi_t(a \mid s, z) > 0$ is sufficient condition of $a \in \arg\max_{a' \in \mathcal{A}} \mathbb{E}_{z \sim \rho_t^{\mathrm{pred}}(\cdot|\mathcal{I}_t)} Q_t^*(s, a', \mu, z; \boldsymbol{\pi})$. Besides, the mean field $\boldsymbol{\mu}$ satisfies that $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}, z)$. So $(\boldsymbol{\pi}, \boldsymbol{\rho})$ is an AMFCE. □

A.4 Proof of Proposition 2

Proposition 2. *$(\boldsymbol{\pi}, \boldsymbol{\rho})$ is an AMFCE solution if and only if $\mathcal{R}(a_{0:T}, \boldsymbol{\pi}, \boldsymbol{\rho}) \leq 0$, $\forall a_{0:T} \in \mathcal{A}^{\mathcal{T}}$.*

Proof. **(Sufficient Condition)**. If $(\boldsymbol{\pi}, \boldsymbol{\rho})$ is a solution of AMFCE, but the inequality in Proposition 2 does not hold. There exists some t and trajectory such that

$$\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)\right] > J(\boldsymbol{\pi}, \boldsymbol{\rho})$$

From the definition of AMFCE,

$$\sum_{a \in \mathcal{A}} \sum_{z \in \mathcal{Z}} \rho_t(z) \pi_t(a|s, z) \left[Q_t^{\boldsymbol{\pi}}(s, a, \mu_t, z; \boldsymbol{\pi}) - Q_t^{\boldsymbol{\pi}}(s, a', \mu_t, z; \boldsymbol{\pi})\right] \geq 0$$

We have that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)\right] \\
=&\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r(a_t, s_t, \mu_t) + \gamma^T r(s_T, a_T, \mu_T)\right] \\
\leq&\mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r(a_t, s_t, \mu_t) + \gamma^T \mathbb{E}\left[Q_T^{\boldsymbol{\pi}}(s_T, a, \mu_T, z; \boldsymbol{\pi})\right]\right]
\end{aligned}$$

The outer expectation is taken with respect to $z_t \sim \rho_t(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$ and the inner expectation is taken with respect to $z \sim \rho_T(\cdot)$, $a \sim \pi_T(\cdot|s_T, z)$. Similarly, we can induce that

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)\right] \\
=&\mathbb{E}\left[\sum_{t=0}^{T-2} \gamma^t r(a_t, s_t, \mu_t) + \gamma^{T-1} r(s_{T-1}, a_{T-1}, \mu_{T-1}) + \gamma^T \mathbb{E}\left[Q_T^{\boldsymbol{\pi}}(s_T, a, \mu_T, z; \boldsymbol{\pi})\right]\right] \\
\leq&\mathbb{E}\left[\sum_{t=0}^{T-2} \gamma^t r(a_t, s_t, \mu_t) + \gamma^{T-1} \mathbb{E}[Q_{T-1}^{\boldsymbol{\pi}}(s_{T-1}, a, \mu_{T-1}, z; \boldsymbol{\pi})]\right] \\
\leq&\mathbb{E}\left[Q_0^{\boldsymbol{\pi}}(s_0, a, \mu_0, z; \boldsymbol{\pi})\right] = J(\boldsymbol{\pi}, \boldsymbol{\rho})
\end{aligned}$$

where the last expectation is taken with respect to $z \sim \rho_0, s_0 \sim \mu_0(\cdot), a \sim \pi_0(\cdot|s_0, z)$.

It contradicts with the assumption.

(Necessary Condition). We assume that the inequality holds and $(\boldsymbol{\pi}, \boldsymbol{\rho})$ is not an AMFCE. There exists a time step $t \in \mathcal{T}$ such that $\Delta_t(s, \mu, u; \boldsymbol{\pi}, \boldsymbol{\rho}) =$

$\mathbb{E}[Q_t^{\boldsymbol{\pi}}(s, u(a), \mu, z) - Q_t^{\boldsymbol{\pi}}(s, a, \mu, z)] > 0$. Then agent can achieve a strictly higher expected return if she chooses action $u(a)$ when she is recommended action a at time step t. It implies that there exists an action sequence such that $\mathcal{R}(a_{0:T}, \boldsymbol{\pi}, \boldsymbol{\rho}) > 0$, which conflicts with the assumption. □

A.5 Proof of Theorem 3

Theorem 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 if the individual policy is $\boldsymbol{\pi}^*$. Then we have*

$$L(\boldsymbol{\pi}, \boldsymbol{\rho}, \lambda_{\boldsymbol{\pi}}^*, r) = \mathbb{E}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)] - J(\boldsymbol{\pi}, \boldsymbol{\rho})$$

where the expectation is taken with respect to $z_t \sim \rho_t(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $a_t \sim \pi_t^*(\cdot|s_t, z_t)$, $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}, z_{t-1})$.

Proof. We note that

$$\begin{aligned}
&\sum_{\tau_k \in \mathcal{D}_E} \lambda_{\boldsymbol{\pi}}^*(\tau_i) \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t) \right] \\
&= \mathbb{E}_{\boldsymbol{\pi}^*} \mathbb{E}_{\boldsymbol{\pi}} \left[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t) \right] \\
&= \mathbb{E}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)]
\end{aligned}$$

The $\mathbb{E}_{\boldsymbol{\pi}}$ is expectation taken with respect to $z_t \sim \rho_t(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}, z_{t-1})$. The $\mathbb{E}_{\boldsymbol{\pi}^*}$ is taken with respect to $a_t \sim \pi_t^*(\cdot|s_t, z_t)$. The third expectation is taken with respect to $z_t \sim \rho_t(\cdot)$, $a_t \sim \pi_t^*(\cdot|s_t, z_t)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}, z_{t-1})$. Then we can derive the conclusion directly.

$$L(\boldsymbol{\pi}, \boldsymbol{\rho}, \lambda_{\boldsymbol{\pi}}^*, r) = \mathbb{E}\Big[\sum_{t=0}^{T} \gamma^j r(s_t, a_t, \mu_t) \Big] - J(\boldsymbol{\pi}, \boldsymbol{\rho})$$

□

A.6 Proof of Proposition 4

Proposition 4. *The policy $\boldsymbol{\pi}$ learned on the reward function recovered by AMFCE-IRL can be characterized as follows:*

$$\text{MFRL} \circ \text{AMFCE-IRL}_{\psi}(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) := \arg\min_{\boldsymbol{\pi}} \max_{r} J(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E)$$

$$- \mathbb{E}\Big[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)\Big] - \psi_{GA}(r)$$

where the expectation is taken with respect to $z_t \sim \rho_t^E(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $a_t \sim \pi_t(\cdot|s_t, z_t)$, $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}^E, z_{t-1})$.

The objective to recover AMFCE is defined as:

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

where D_ω *is the discriminator network parameterized with* ω*, with input* (s_t, a_t, μ_t) *and output a real number in* $(0, 1]$*. The first expectation is taken with respect to* $z_t \sim \rho_t^E(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $a_t \sim \pi_t(\cdot|s_t, z_t)$, $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}^E, z_{t-1})$*. The second expectation is taken with respect to* $z_t \sim \rho_t^E(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $a_t \sim \pi_t^E(\cdot|s_t, z_t)$, $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}^E, z_{t-1})$.

Proof. We denote $\tilde{r} = \text{AMFCE-IRL}(\boldsymbol{\pi}_E)$. The saddle point of $L(\boldsymbol{\pi}, \boldsymbol{\rho}, \lambda, r)$ is $\lambda_{\boldsymbol{\pi}}^E(\tau_k) = \prod_{t=0}^{T} \pi_t^E(a_t|s_t, z_t)$ and $\tilde{r}$, where $(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) \in \text{AMFCE}$. So given expert demonstrations sampled from $(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) \in \text{AMFCE}$, we can recover $\boldsymbol{\pi}^E$ by (A3).

$$\begin{aligned}
\boldsymbol{\pi} &= \arg\min_{\boldsymbol{\pi}} J(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) - \mathbb{E}[\sum_{t=0}^{T} \gamma^t \tilde{r}(s_t, a_t, \mu_t)] \\
&= \arg\min_{\boldsymbol{\pi}} \max_{r} J(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) - \mathbb{E}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)] - \psi_{GA}(r)
\end{aligned} \tag{A3}$$

If we select ψ_{GA} as the regularizer, and make the change of variables $r(s, a, \mu) = -\log(d(s, a, \mu))$, we get

$$\begin{aligned}
&\max_{r} J(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) - \mathbb{E}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)] - \psi_{GA}(r) \\
=& - \max_{d} \mathbb{E}_{\boldsymbol{\pi}_E, \boldsymbol{\rho}_E}[\sum_{t=0}^{T} \gamma^t \log(d(s_t, a_t, \mu_t))] + \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\rho}_E}[\sum_{t=0}^{T} \gamma^t \log(d(s_t, a_t, \mu_t))] \\
&- \max_{d} \mathbb{E}_{\boldsymbol{\pi}_E, \boldsymbol{\rho}_E}[\sum_{t=0}^{T} g(r(s_t, a_t, \mu_t))] \\
=& \max_{\omega} \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\rho}^E}\big[\sum_{t=0}^{T} \gamma^t \log D_\omega(s_t, a_t, \mu_t)\big] + \mathbb{E}_{\boldsymbol{\pi}^E, \boldsymbol{\rho}^E}\Big[\sum_{t=0}^{T} \gamma^t \log\big(1 - D_\omega(s_t, a_t, \mu_t)\big)\Big]
\end{aligned}$$

where the expectation $\mathbb{E}_{\boldsymbol{\pi}^E, \boldsymbol{\rho}^E}$ is taken with respect to $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $a_t \sim \pi_t^E(\cdot|s_t, z_t)$, $z_t \sim \rho_t^E(\cdot)$, $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}^E, z_{t-1})$ and the expectation $\mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\rho}^E}$ is

taken with respect to $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $a_t \sim \pi_t(\cdot|s_t, z_t)$, $z_t \sim \rho_t^E(\cdot)$, $\mu_t = \Phi(\mu_{t-1}, \pi_{t-1}^E, z_{t-1})$.

□

A.7 Proof of Proposition 5

Proposition 5. *If $\boldsymbol{\rho}^\phi$ is parameterized with ϕ, the gradient to optimize ϕ given state s is*

$$\mathbb{E}_{z\sim\rho_t^\phi(\cdot)}\Big[\nabla_\phi \log \rho_t^\phi(z)\mathbb{E}_{a\sim\pi_t(\cdot|s,z)}Q_t^{\boldsymbol{\pi}}(s,a,\mu,z;\boldsymbol{\pi})\Big]. \tag{10}$$

Proof. The gradient of parameterized ρ^ϕ is

$$\begin{aligned}
&\nabla_\phi \sum_{z\in\mathcal{Z}} \rho_t^\phi(z) \sum_{a\in\mathcal{A}} \pi_t(a|s,z)Q_t^{\boldsymbol{\pi}}(s,a,\mu,z;\boldsymbol{\pi}) \\
=&\sum_{z\in\mathcal{Z}} \nabla_\phi\rho_t^\phi(z) \sum_{a\in\mathcal{A}} \pi_t(a|s,z)Q_t^{\boldsymbol{\pi}}(s,a,\mu,z;\boldsymbol{\pi}) \\
=&\mathbb{E}_{z\sim\rho_t^\phi(\cdot)}\Big[\sum_{a\in\mathcal{A}} \pi_t(a|s,z)Q_t^{\boldsymbol{\pi}}(s,a,\mu,z;\boldsymbol{\pi})\nabla_\phi \log \rho_t^\phi(z)\Big] \\
=&\mathbb{E}_{z\sim\rho_t^\phi(\cdot)}\Big[\nabla_\phi \log \rho_t^\phi(z)\mathbb{E}_{a\sim\pi_t(\cdot|s,z)}Q_t^{\boldsymbol{\pi}}(s,a,\mu,z;\boldsymbol{\pi})\Big]
\end{aligned}$$

□

Appendix B Experiment detail

The experiments were run on the server with AMD EPYC 7742 64-Core Processor and NVIDIA A100 40GB.

Due to the instability natureof generative adversarial networks (GANs) [1, 17], the convergence of Algorithm 1 may not be not guaranteed. To address this issue, we integrated the gradient penalty into the objective function of CMFIL to stabilize the training of policy $\boldsymbol{\pi}$. It has been proven that GAN training with zero-centered will enhance the training stability [17]. To provide a fair comparison, we used Actor-Critic (AC) algorithm for both CMFIL, MFAIRL, and MFIRL. The input of AC is an extended state, a concatenation of state, action, time step, and signature. The input of the discriminator is the extended state and the action. We did not use signature in the Ocean Ranch and RPS because signature requires the length of sequential data is larger than 1. For games with the sequential setting, the depth of truncated signature is 3. For actor and critic networks of AC, we adopt two-layer perceptrons with the Adam optimizer and the ReLU activation function. For the network of the discriminator, we adopt three-layer perceptrons with Adam optimizer. The activation functions between layers are Leaky ReLU, while the activation function of output is the sigmoid activation function. The setting of main hyperparameters is shown in Table B1.

Table B1: The hyperparameters in the experiment

hyperparameters	value
hidden size of actor network	256
hidden size of critic network	256
hidden size of discriminator network	128

Equilirbrium	MFCE		AMFCE			
Distribution	$\pi(B\|s', z=0)$	$\rho(z=0)$	$\pi(B\|s', z=0)$	$\pi(B\|s', z=1)$	$\rho(z=0)$	$\rho(z=1)$
Value	1	1	1/2	1	1/2	1/2

Table C2: The only MFCE and a possible AMFCE in the absent-minded driver game.

Appendix C Comparison with MFCE Derived by Muller et al.

In this section, We use the absent-minded driver game [23] to show the difference between AMFCE and the MFCE framework proposed by Muller et al. [20]. Their notion of MFCE assumes that the mediator selects a mixed policy for the population and then sample a deterministic policy from the mixed policy and recommends to every agent, while our AMFCE framework assumes that the mediator selects a behavioral policy for the population at every time step and samples an action for every agent as recommendation. If agents are of bounded rationality, the mixed policy is not equivalent to the behavioral policy.

Example 2. *Suppose that the absent-minded driver game has two time steps. At the initial time, all the agents stay in state s_1. The agent will stay in the state s_1 if action B is chosen and the current mean field $\mu(s_1) = 1$. If action E is chosen, the agent will move to state s_2. If the agent enter the state s_2, the agent will stay in s_2 until the ending of the game. The reward function is*

$$r(s,a,\mu) = \begin{cases} 3(1-\mu(s_1)), & a = E, s = s_1 \\ \frac{1}{2}, & a = B, s = s_1, \mu = \cdot \\ 0, & otherwise \end{cases}.$$

Consider the case where the agents cannot remember the time step and the history. and the agent does not choose to take the deterministic policy of action E at s' because the policy makes the final payoff 0. So the only MFCE policy in the game is the deterministic policy to take action B in any state, which has a final payoff of 1.

On the other hand, we can find a possible AMFCE shown in the Table C2. The agents will choose action E if it is recommended.

Example 2 suggests that AMFCE has larger policy space than the MFCE proposed by Muller et al. [20] because AMFCE assumes that the correlated signal sampled by the mediator corresponds to a behavioral policy.

References

[1] Arjovsky, M. and Bottou, L. (2017). Towards principled methods for training generative adversarial networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

[2] Aumann, R. J. (1974). Subjectivity and correlation in randomized strategies. *Journal of mathematical Economics*, 1(1):67–96.

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

[4] Campi, L. and Fischer, M. (2022). Correlated equilibria and mean field games: a simple model. *Mathematics of Operations Research*.

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

[6] Chen, Y., Zhang, L., Liu, J., and Hu, S. (2022). Individual-level inverse reinforcement learning for mean field games. In Faliszewski, P., Mascardi, V., Pelachaud, C., and Taylor, M. E., editors, *21st International Conference on Autonomous Agents and Multiagent Systems, AAMAS 2022, Auckland, New Zealand, May 9-13, 2022*, pages 253–262. International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

[7] Chen, Y., Zhang, L., Liu, J., and Witbrock, M. (2021b). Adversarial inverse reinforcement learning for mean field games. *arXiv preprint arXiv:2104.14654*.

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

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

[10] Ho, J. and Ermon, S. (2016). Generative adversarial imitation learning. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 4565–4573.

[11] Hussein, A., Gaber, M. M., Elyan, E., and Jayne, C. (2017). Imitation learning: A survey of learning methods. *ACM Comput. Surv.*, 50(2):21:1–21:35.

[12] Jeon, W., Barde, P., Nowrouzezahrai, D., and Pineau, J. (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*.

[13] Jeong, S. H., Kang, A. R., and Kim, H. K. (2015). Analysis of game bot's behavioral characteristics in social interaction networks of MMORPG. In Uhlig, S., Maennel, O., Karp, B., and Padhye, J., editors, *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication, SIGCOMM 2015, London, United Kingdom, August 17-21, 2015*, pages 99–100. ACM.

[14] Kakutani, S. (1941). A generalization of brouwer's fixed point theorem. *Duke mathematical journal*, 8(3):457–459.

[15] Kidger, P., Bonnier, P., Arribas, I. P., Salvi, C., and Lyons, T. J. (2019). Deep signature transforms. In Wallach, H. M., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. B., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3099–3109.

[16] Kidger, P. and Lyons, T. J. (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.

[17] Mescheder, L. M., Geiger, A., and Nowozin, S. (2018). Which training methods for gans do actually converge? In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3478–3487. PMLR.

[18] Min, M. and Hu, R. (2021). Signatured deep fictitious play for mean field games with common noise. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7736–7747. PMLR.

[19] Min, M. and Ichiba, T. (2020). Convolutional signature for sequential data. *CoRR*, abs/2009.06719.

[20] Muller, P., Elie, R., Rowland, M., Laurière, M., Pérolat, J., Perrin, S., Geist, M., Piliouras, G., Pietquin, O., and Tuyls, K. (2022). Learning correlated equilibria in mean-field games. *CoRR*, abs/2208.10138.

[21] Ni, H., Szpruch, L., Wiese, M., Liao, S., and Xiao, B. (2020). Conditional sig-wasserstein gans for time series generation. *arXiv preprint arXiv:2006.05421*.

[22] Perrin, S., Laurière, M., Pérolat, J., Geist, M., Élie, R., and Pietquin, O. (2021). Mean field games flock! the reinforcement learning way. In Zhou, Z., editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 356–362. ijcai.org.

[23] Piccione, M. and Rubinstein, A. (1996). On the interpretation of decision problems with imperfect recall. In Shoham, Y., editor, *Proceedings of the Sixth Conference on Theoretical Aspects of Rationality and Knowledge, De Zeeuwse Stromen, The Netherlands, March 17-20 1996*, pages 75–76. Morgan Kaufmann.

[24] Song, J., Ren, H., Sadigh, D., and Ermon, S. (2018). Multi-agent generative adversarial imitation learning. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 7472–7483.

[25] Waugh, K., Ziebart, B. D., and Bagnell, J. A. (2013). Computational rationalization: The inverse equilibrium problem. *CoRR*, abs/1308.3506.

[26] Yang, F., Vereshchaka, A., Chen, C., and Dong, W. (2020). Bayesian multi-type mean field multi-agent imitation learning. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

[27] Yang, J., Ye, X., Trivedi, R., Xu, H., and Zha, H. (2018a). 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.

[28] Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. (2018b). Mean field multi-agent reinforcement learning. In Dy, J. G. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 5567–5576. PMLR.

[29] Yu, L., Song, J., and Ermon, S. (2019). Multi-agent adversarial inverse reinforcement learning. In Chaudhuri, K. and Salakhutdinov, R., editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7194–7201. PMLR.