A Related Works

Our research contributes to the extensive body of multi-agent imitation learning. One line of studies in particular applies single-agent IL algorithms to Markov games (Song et al., 2018; Yu et al., 2019; Jeon et al., 2020). However, the scalability of these methods is limited due to the exponential expansion of agent interactions as the number of agents rises.

To improve the scalability, Fan Yang et al. provide a new multi-type mean field approximation to approximate Nash equilibrium in Markov game (Yang et al., 2020), but they did not incorporate MFG and mean field equilibrium into their work and decouple the interdependence between mean field flow and policy. One conventional solution concept of imitation learning for MFG is MFNE.

There are also some variants of MFNE such as stationary mean field equilibrium (SMFE) and stationary mean-field social-welfare optimal (SMF-SO) (Subramanian & Mahajan, 2019). Yang et al. used inverse reinforcement learning to solve MFG by reducing MFG to a Markov decision process (Yang et al., 2018a). However, this simplification only applies in a fully cooperative setting. As a result, it requires demonstrations are sampled from an MFSO rather than an MFNE, which limits the scope of application as MFSO is a specific type of MFNE. Chen et al. used individual behaviors to infer ground-truth reward functions for MFG and allowed demonstrations are sampled from an MFNE (Chen et al., 2022).

Inspired by CE, There have been recent studies that generalized correlated equilibrium from the stateless game to the MFCE (Muller et al., 2022; Campi & Fischer, 2022). But their assumptions restricted the space of policy. Compared with work focusing on recovering CE in matrix game (Waugh et al., 2013), our work improved the scalability by incorporating MFGs and considering the game with a sequential setting.

B Preliminaries of Imitation Learning

Here we provide some background introductions to single-agent imitation learning (IL). Let $\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \mu_0, \gamma, T)$ denote an single-agent Markov decision process (MDP). $\mathcal{S}$ and $\mathcal{A}$ are, respectively, the state and action spaces. $P : \mathcal{S} \times \mathcal{A} \to \mathcal{P}(\mathcal{S})$ is the transition kernel for the state dynamics. $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function. μ_0 is the initial distribution of initial state s_0. $\gamma \in (0, 1]$ is the discount factor. T is the horizon. The expected return of policy π is $J(\pi) = \mathbb{E}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t)]$, where the expectation is taken with respect to $s_0 \sim \mu_0$, $a_t \sim \pi(\cdot|s_t)$, $s_{t+1} \sim P(\cdot|s_t, a_t)$.

In the Imitation Learning (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.

Inverse Reinforcement Learning (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) (Ho & Ermon, 2016) 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 $s_{t+1} \sim P(\cdot|s_t, a_t)$, $a_t \sim \pi^E(\cdot|s_t)$, $s_0 \sim \mu_0$.

C EXAMPLE

Proof. We verify the example by the definition of AMFCE. If $u(a) = a$, $\Delta_t(s_0, \mu_0, u; \boldsymbol{\pi}, \boldsymbol{\rho}) = 0$. If $u(R) = L$,

$$\begin{aligned}
&\sum_{z' \in \mathcal{Z}} \rho_0(z') \pi_0(a_0 | s_0, z') \Delta_t(s_0, \mu_0, u; \boldsymbol{\pi}, \boldsymbol{\rho}) \\
=&\rho_0(0) \pi_0(R | s = \cdot, z = 0) \big(\Phi(\mu_0, \pi_0, z_0 = 0)(L) - \Phi(\mu_0, \pi_0, z_0 = 0)(R) \big) \\
&+ \rho_0(1) \pi_0(R | s = \cdot, z = 1) \big(\Phi(\mu_0, \pi_0, z_0 = 1)(L) - \Phi(\mu_0, \pi_0, z_0 = 1)(R) \big) \\
=&-\frac{1}{18} < 0
\end{aligned}$$

As $\sum_{z' \in \mathcal{Z}} \rho_0(z') \pi_0(a_0 | s_0, z') > 0$, $\Delta_t(s_0, \mu_0, u; \boldsymbol{\pi}, \boldsymbol{\rho}) \leq 0$. The same is true for $u(L) = R$. So the example in the Example 1 is an AMFCE. □

C.1 FINITE HORIZON EXAMPLE

Example 2. *Consider a game with state space $\mathcal{S} = \{C, L, R\}$. The action space is $\mathcal{A} = \{L, R\}$. Initial mean field $\mu_0(C) = 1$. The reward $r(s, a, \mu) = \mathbb{1}_{\{s=L\}} \mu(L) + \mathbb{1}_{\{s=R\}} \mu(R)$ and $\mathcal{T} = \{0, 1, 2\}$. If agent is in the state C, the environment transition is deterministic. $P(s_1 = R \mid s_0 = C, a = R) = 1$, $P(s_1 = L \mid s_0 = C, a = R) = 0$, $P(s_1 = R \mid s_0 = C, a = L) = 0$, $P(s_1 = L \mid s_0 = C, a = L) = 1$. $P(s_2 = R \mid s_1 = R, a = R) = 1$, Given the current state L, agent will be transited to state R with probability $P(s_2 = R \mid s_1 = L, a = R) = \frac{3}{4}$, and stay in L with probability $P(s_2 = L \mid s_1 = L, a = R) = \frac{1}{4}$. If she choose action L in the state L, she will stay in state L with probability $P(s_2 = L \mid s_1 = L, a = L) = 1$. The case is similar for agents whose current state is R. $P(s_2 = L \mid s_1 = R, a = L) = \frac{3}{4}$, $P(s_2 = R \mid s_1 = R, a = L) = \frac{1}{4}$, $P(s_2 = R \mid s_1 = R, a = R) = 1$, $P(s_2 = L \mid s_1 = R, a = R) = 0$.*

The mediator in an AMFCE gives recommendation as follows. At time step $t = 0$, a random variable z is sampled from the correlated signal space $\mathcal{Z} = \{0, 1\}$ with equal probability $\rho_0(z = 0) = \rho(z = 1) = 0.5$, and the mediator gives the action recommendation for each agent according to the policy $\pi_0(a = L | z = 0) = 2/3$, $\pi_0(a = R | z = 0) = 1/3$, $\pi_0(a = R | z = 1) = 2/3$, $\pi_0(a = L | z = 1) = 1/3$. At time step $t = 1$, z is sampled from the correlation device $\rho_1(z = 0) = \rho_1(z = 1) = 0.5$, the action recommendation for each agent is $\pi_1(a = L | s = L, z = 0) = 1$, $\pi_1(a = R | s = L, z = 0) = 0$, $\pi_1(a = L | s = L, z = 1) = 8/9$, $\pi_1(a = L | s = L, z = 1) = 1/9$. $\pi_1(a = L | s = R, z = 0) = 8/9$, $\pi_1(a = R | s = R, z = 0) = 1/9$, $\pi_1(a = L | s = R, z = 1) = 0$, $\pi_1(a = L | s = R, z = 1) = 1$. It can be verified that $(\boldsymbol{\pi}, \boldsymbol{\rho})$ is an AMFCE.

D PROOF

D.1 PROOF OF BELLMAN EQUATION

Proof.

$$\begin{aligned}
Q_t^{\boldsymbol{\pi}}(s, a, \mu, z; \boldsymbol{\pi}') =& r(s, a, \mu) + \gamma \mathbb{E}_{\boldsymbol{\pi}'} \left[\sum_{i=t+1}^{T} \gamma^{i-t-1} r(s_i, a_i, \mu_i) \middle| (s_t, a_t, \mu_t, z_t) = (s, a, \mu, z) \right] \\
=& r(s, a, \mu) + \gamma \mathbb{E}_{\boldsymbol{\pi}'} \Big[r(s_{t+1}, a_{t+1}, \Phi(\mu, \pi'_t, z)) \\
&+ \gamma \sum_{i=t+2}^{T} \gamma^{i-t-2} r(s_i, a_i, \mu_i) \Big| (s_t, a_t, \mu_t, z_t) = (s, a, \mu, z) \Big] \qquad (12)
\end{aligned}$$

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 $z_i \sim \rho_i(\cdot)$, $a_i \sim \pi_i(\cdot | s_i, z_i)$, $s_{i+1} \sim P(\cdot | s_i, a_i, \mu_i)$, $\mu_i(\cdot) = \sum_{a \in \mathcal{A}} \sum_{s \in \mathcal{S}} \mu_{i-1}(s) P(\cdot | s, a, \mu_{i-1}) \pi'_{i-1}(a | s, z_{i-1})$,

$\forall i \in \{t+1, t+2, \cdots, T\}$.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\pi}'}\Big[r\big(s', a', \Phi(\mu, \pi'_t, z)\big) + \gamma \sum_{i=t+2}^{T} \gamma^{i-t-2} r(s_i, a_i, \mu_i)\Big] \\
=&\mathbb{E}\Bigg[r\big(s', a', \Phi(\mu, \pi'_t, z)\big) + \gamma \mathbb{E}_{\boldsymbol{\pi}'}\big[\sum_{i=t+2}^{T} \gamma^{i-t-2} r(s_i, a_i, \mu_i) | (s_{t+1}, a_{t+1}, \mu_{t+1}, z_{t+1}) = (s', a', \Phi(\mu, \pi'_t, z), z')\big]\Bigg] \\
=&\mathbb{E}\Big[Q^{\boldsymbol{\pi}}_{t+1}\big(s', a', \Phi(\mu, \pi'_t, z), z'; \boldsymbol{\pi}'\big)\Big] \qquad (13)
\end{aligned}$$

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 (12) and (13), we get the Bellman equation.

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

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

D.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 $a \in \arg\max_{a' \in \mathcal{A}} \mathbb{E}_{z \sim \rho^{\text{pred}}_t(\cdot|\mathcal{I}_t)} Q^*(s, a', \mu, z; \boldsymbol{\pi})$, $\forall t \in \mathcal{T}$.*

Proof. We denote

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

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

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 $\mathbb{E}\left[\mathcal{Q}^{\boldsymbol{\pi}'}_0(s_0, a_0, \mu_0, \mathcal{I}_0; \boldsymbol{\pi})\right]$, 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 $\forall t \in \mathcal{T}$, $\mathbb{E}\left[\mathcal{Q}^{\boldsymbol{\pi}'}_0(s_0, a_0, \mu_0, \mathcal{I}_0; \boldsymbol{\pi})\right] = \max_{\tilde{\boldsymbol{\pi}}} \mathbb{E}\left[\mathcal{Q}^{\tilde{\boldsymbol{\pi}}}_0(s_0, a_0, \mu_0, \mathcal{I}_0; \boldsymbol{\pi})\right]$, 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. $\mathrm{BR}(\boldsymbol{\pi}; \boldsymbol{\rho})$ *has a closed graph.*

Proof. We assume that $\lim_{n \to \infty} \boldsymbol{\pi}_n = \boldsymbol{\pi}$, $\lim_{n \to \infty} \boldsymbol{\pi}'_n = \boldsymbol{\pi}'$, $\boldsymbol{\pi}_n \in \mathrm{BR}(\boldsymbol{\pi}'_n; \boldsymbol{\rho})$, but $\boldsymbol{\pi} \notin \mathrm{BR}(\boldsymbol{\pi}'; \boldsymbol{\rho})$. 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 $\forall u \in \mathcal{U}, \forall s \in \mathcal{A}$, $\forall t \in \mathcal{T}$,

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

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

D.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 (Cui & Koeppl, 2021). $\pi_t(a \mid s,z) > 0$ 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. □

D.4 Proof of Corollary 2

Corollary 2. *MaxEnt-AMFCE is a unique equilibrium solution if $\Delta(\boldsymbol{\pi},\boldsymbol{\rho})$ is convex w.r.t. $(\boldsymbol{\pi},\boldsymbol{\rho})$.*

Proof. If $\Delta(\boldsymbol{\pi},\boldsymbol{\rho})$ is convex w.r.t. $(\boldsymbol{\pi},\boldsymbol{\rho})$, the set $\Pi = \{(\boldsymbol{\pi},\boldsymbol{\rho}) | \Delta(\boldsymbol{\pi},\boldsymbol{\rho}) \leq 0\}$ is convex. As $H(\boldsymbol{\pi},\boldsymbol{\rho})$ is concave function, $\max_{(\boldsymbol{\pi},\boldsymbol{\rho})\in\Pi} H(\boldsymbol{\pi},\boldsymbol{\rho})$ has unique solution. So the MaxEnt-AMFCE is a unique equilibrium solution if $\Delta(\boldsymbol{\pi},\boldsymbol{\rho})$ is convex w.r.t. $(\boldsymbol{\pi},\boldsymbol{\rho})$. □

D.5 Proof of Proposition 1

Proposition 1. *The entropy can be decoupled: $H(\boldsymbol{\pi},\boldsymbol{\rho}) = \sum_{t=0}^{T}[H(\rho_t) + \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\rho}} H(\pi_t|s_t,z_t)]$.*

Proof.

$$
\begin{aligned}
&H(\boldsymbol{\pi}, \boldsymbol{\rho}) \\
=&\sum_{t=0}^{T} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\rho}} \left[\sum_{a_t, z_t} -\rho_t(z_t)\pi_t(a_t \mid s_t, z_t) \log \pi_t(a_t \mid s_t, z_t)\rho_t(z_t) \right] \\
=&\sum_{t=0}^{T} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\rho}} \left[-\sum_{a_t, z_t} \rho_t(z_t)\pi_t(a_t \mid s_t, z_t) \log \rho_t(z_t) \right] \\
&-\sum_{t=0}^{T} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\rho}} \left[\sum_{a_t, z_t} \rho_t(z_t)\pi_t(a_t \mid s_t, z_t) \log \pi_t(a_t \mid s_t, z_t) \right] \\
=&\sum_{t=0}^{T} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\rho}} \left[\sum_{z_t} -\rho_t(z_t) \log \rho_t(z_t) \right] + \sum_{t=0}^{T} \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\rho}} \left[-\sum_{z_t} \rho_t(z_t) \sum_{a_t} \pi_t(a \mid s_t, z) \log \pi_t(a_t \mid s_t, z_t) \right] \\
=&\sum_{t=0}^{T} [H(\rho_t) + \mathbb{E}_{\boldsymbol{\pi},\boldsymbol{\rho}} H(\pi_t | s_t, z_t)]
\end{aligned}
$$

□

D.6 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,

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

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

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

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. □

D.7 Proof of Theorem 2

Theorem 2. *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 $L(\boldsymbol{\pi}, \boldsymbol{\rho}, \lambda^*_{\boldsymbol{\pi}}, r) = \mathbb{E}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)] - J(\boldsymbol{\pi}, \boldsymbol{\rho}) - \alpha \sum_{t=0}^{T} \mathbb{E}_{\boldsymbol{\pi}, \boldsymbol{\rho}} H(\pi_t|s_t, z_t)$, 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

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

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.

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

□

D.8 Proof of Proposition 3

Proposition 3. *The policy $\boldsymbol{\pi}$ learned on the reward function recovered by AMFCE-IRL can be characterized as follows:*

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

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

The objective to recover MaxEnt-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 $z_t \sim \rho_t^E(\cdot)$, $s_t \sim P(\cdot|s_{t-1}, a_{t-1}, \mu_{t-1})$, $a_t \sim \pi_t(\cdot|s_t, z_t)$, $\mu_t = \Phi(\mu_{t-1}, \pi^E_{t-1}, z_{t-1})$.

Proof. We denote $\tilde{r} = \text{AMFCE-IRL}(\boldsymbol{\pi}_E)$. The saddle point of $L(\boldsymbol{\pi}, \boldsymbol{\rho}, \lambda, r)$ is $\lambda^E_{\boldsymbol{\pi}}(\tau_k) = \prod_{t=0}^{T} \pi_t^E(a_t|s_t, z_t)$ and $(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) \in \text{MaxEnt-AMFCE}$. So given expert demonstrations sampled from $(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) \in \text{MaxEnt-AMFCE}$, we can recover $\boldsymbol{\pi}^E$ by (14).

$$\begin{aligned}\boldsymbol{\pi} &= \arg\min_{\boldsymbol{\pi}} -\alpha \sum_{t=0}^{T} \mathbb{E}H(\pi_t^E|s_t, z_t) + J(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) - \mathbb{E}[\sum_{t=0}^{T} \gamma^t \tilde{r}(s_t, a_t, \mu_t)] \\ &= \arg\min_{\boldsymbol{\pi}} \max_{r} J(\boldsymbol{\pi}^E, \boldsymbol{\rho}^E) - \mathbb{E}[\sum_{t=0}^{T} \gamma^t r(s_t, a_t, \mu_t)] - \psi_{GA}(r)\end{aligned} \tag{14}$$

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

□

D.9 Proof of Proposition 4

Proposition 4. *If $\boldsymbol{\rho}^\phi$ is parameterized with ϕ, the gradient to optimize ϕ given state s is*

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

Proof. The gradient of parameterized ρ^ϕ is

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

□

E Further Details about Tasks

E.1 Sequential squeeze

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:

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

The reward function is

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

E.2 RPS

The dynamic is deterministic:

$$P(s_{t+1} \mid s_t, a_t, \mu_t) = \mathbb{1}_{s_{t+1}=a_t} \tag{15}$$

Formally, the state space $\mathcal{S} = \{R, P, S\}$ and the action space $\mathcal{A} = \{R, P, S\}$. The reward function is shown in the following

$$r(R, a, \mu_t) = 2 \cdot \mu_t(S) - 1 \cdot \mu_t(P)$$
$$r(P, a, \mu_t) = 4 \cdot \mu_t(R) - 2 \cdot \mu_t(S)$$
$$r(S, a, \mu_t) = 2 \cdot \mu_t(P) - 1 \cdot \mu_t(R)$$

E.3 FLOCK

We simplify the setting, and the dynamic of the new setting is following

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

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

$$f_\beta^{\text{flock}}(x, v, u, \mu) = -\left\| \int_{\mathbb{R}^{2d}} \frac{(v - v')\, \mathrm{d}\mu\,(x', v')}{\left(1 + \|x - x'\|^2\right)^\beta} \right\|^2$$

In our setting, β is set to 0.

E.4 TRAFFIC NETWORK

In this task, we use the traffic data of London from Uber Movement. The dynamic of the environment is deterministic. The expert demonstrations is the traffic flow data. The goal of this experiment is to predict the traffic flow of a real-world traffic network including six locations: Lewisham, Hammersmith, Ealing, Redbridge, Enfield, and Big Ben. The detailed result is shown in Section G.5.

F EXPERIMENT DETAIL

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

Due to the instability nature of generative adversarial networks (GANs) (Arjovsky & Bottou, 2017; Mescheder et al., 2018), the convergence of Algorithm 1 may not be not guaranteed. To address this issue, we integrate 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 (Mescheder et al., 2018). To provide a fair comparison, we use Soft Actor-Critic algorithm for both CMFIL and MFIRL. The input of SAC is an extended state, a concatenation of state, action,

Table 4: 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

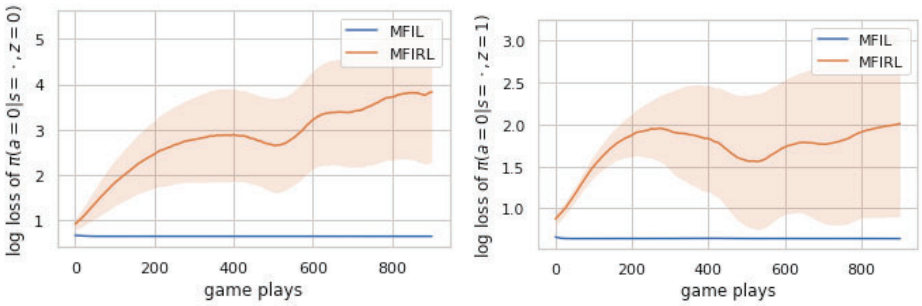
time step, and signature. The input of the discriminator is the extended state and the action. We did not use signature in the Ocean Ranch and RPS because signature requires the length of sequential data is larger than 1. For games with the sequential setting, the depth of truncated signature is 3. For actor and network of SAC, 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 4.

G RESULTS

In this section we will show more results of the experiments.

G.1 OCEAN RANCH

The Learning curves of CMFIL and MFIRL for Ocean Ranch are show in the following.



(a) The learning curve of CMFIL in the game Ocean Ranch.

(b) The learning curve of MFIRL in the game Ocean Ranch.

G.2 SEQUENTIAL SQUEEZE

The Learning curves of CMFIL and MFIRL for Ocean Ranch are show in the following.

G.3 RPS

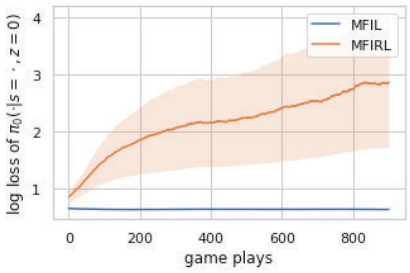
The Learning curves of CMFIL and MFIRL for RPS are show in the following.

G.4 FLOCK

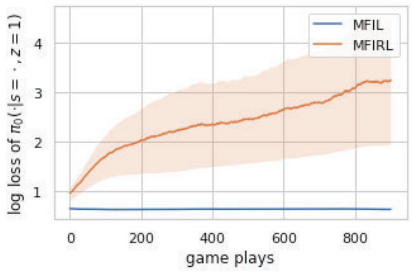
The mean and standard deviation of learned policies of CMFIL and MFIRL for Flock are shown in the following.

G.5 TRAFFIC NETWORK

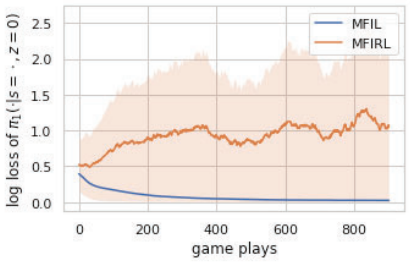
The mean and standard deviation of predicted traffic flow are shown in the Table 5.



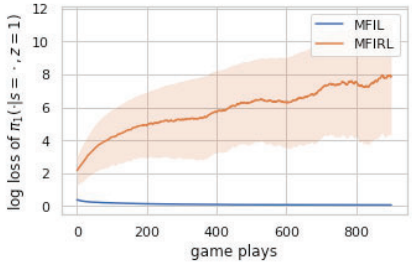
(a) Policies recovered by CMFIL at step $t = 0$.



(b) Policies recovered by CMFIL at step $t = 1$.



(c) Policies recovered by MFIRL at step $t = 0$.



(d) Policies recovered by MFIRL at step $t = 1$.

Figure 3: The learning curves of CMFIL and MFIRL in Sequential Squeeze. It suffers from large variance.

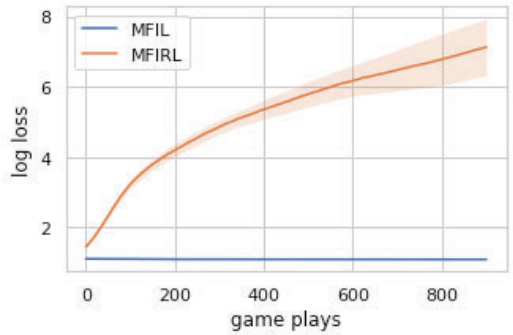
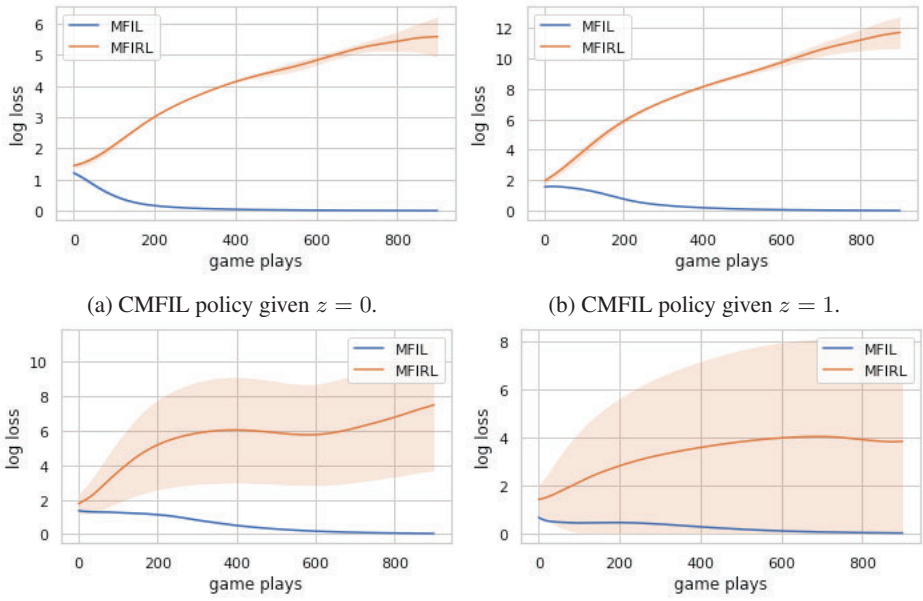


Figure 4: Learning curves of CMFIL and MFIRL in the RPS. The results show that CMFIL successfully recover the policy of MFNE.

	Lewisham	Hammersmith	Ealing	Redbridge	Enfield	Big Ben
Lewisham (real)	0.00000	0.62500	0.00000	0.37500	0.00000	0.00000
Lewisham (CMFIL)	**0.00011 (0.00002)**	**0.58001 (0.02947)**	**0.03665 (0.00289)**	**0.36931 (0.02939)**	**0.01283 (0.00304)**	**0.00109 (0.00030)**
Lewisham (MFIRL)	0.3742 (0.31665)	0.00000 (0.00000)	0.29434 (0.29654)	0.00001 (0.00000)	0.00002 (0.00001)	0.33143 (0.28029)
Hammersmith (real)	0.11628	0.00000	0.67442	0.00000	0.20930	0.00000
Hammersmith (CMFIL)	**0.11903 (0.0076)**	**0.00267 (0.00090)**	**0.62725 (0.01706)**	**0.03761 (0.00147)**	**0.20359 (0.01060)**	**0.00985 (0.00080)**
Hammersmith (MFIRL)	0.00000 (0.00000)	0.33332 (0.33329)	0.00608 (0.00608)	0.33334 (0.33330)	0.32724 (0.32723)	0.32724 (0.32723)
Ealing (real)	0.00000	0.44643	0.00000	0.19643	0.35714	0.00000
Ealing (CMFIL)	**0.00010 (0.00000)**	**0.41772 (0.02494)**	**0.0238 (0.00079)**	**0.19804 (0.00655)**	**0.34396 (0.01901)**	**0.01638 (0.00088)**
Ealing (MFIRL)	0.33340 (0.33326)	0.00001 (0.00001)	0.38909 (0.30909)	0.00001 (0.00000)	0.00002 (0.00001)	0.27774 (0.27737)
Redbridge (real)	0.00000	0.00000	0.00000	0.00000	1.00000	0.00000
Redbridge (CMFIL)	**0.00003 (0.00000)**	**0.00004 (0.00001)**	**0.00003 (0.00000)**	**0.00003 (0.00000)**	**0.94945 (0.01079)**	**0.05042 (0.01078)**
Redbridge (MFIRL)	0.00021 (0.00013)	0.09834 (0.12118)	0.32801 (0.32788)	0.00539 (0.00536)	0.00001 (0.00001)	0.56804 (0.29208)
Enfield (real)	0.00000	0.00000	0.11111	0.88889	0.00000	0.00000
Enfield (CMFIL)	**0.00036 (0.00006)**	**0.00119 (0.00013)**	**0.12593 (0.00626)**	**0.84096 (0.00693)**	**0.03005 (0.00199)**	**0.00152 (0.00007)**
Enfield (MFIRL)	0.33148 (0.33144)	0.03788 (0.03535)	0.00001 (0.00001)	0.00001 (0.00001)	0.32516 (0.28418)	0.30545 (0.30520)
Big Ben (real)	0.24828	0.17241	0.24138	0.19310	0.144828	0.00000
Big Ben (CMFIL)	**0.24146 (0.00961)**	**0.17970 (0.0073)**	**0.23051 (0.00365)**	**0.19614 (0.00844)**	**0.14475 (0.00434)**	**0.00743 (0.00015)**
Big Ben (MFIRL)	0.47647 (0.29003)	0.14370 (0.15027)	0.33336 (0.33331)	0.03697 (0.03595)	0.00004 (0.00002)	0.00946 (0.00863)

Table 5: The mean and standard deviation of predicted traffic flow for Traffic Network.



(a) CMFIL policy given $z = 0$.

(b) CMFIL policy given $z = 1$.

(c) CMFIL policy given $z = 2$.

(d) CMFIL policy given $z = 3$.

Figure 5: Learning curves of CMFIL and MFIRL in Flock, Results are averaged over 3 independent runs. It does not converges to the expert policy.

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

H Comparison with MFCE Derived by Muller et al.

In this section, We use the absent-minded driver game (Piccione & Rubinstein, 1996) to show the difference between AMFCE and the MFCE framework proposed by Muller et al. (Muller et al., 2022). 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 sample an action for every agent as recommendation. If agents are of bounded rationality, the mixed policy is not equivalent to the behavioral policy.

Example 3. *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 6. The agents will choose action E if it is recommended.

Example 3 suggests that AMFCE has larger policy space than the MFCE proposed by Muller et al. (Muller et al., 2022) because AMFCE assumes that the correlated signal sampled by the mediator corresponds to a behavioral policy.