BAYESIAN LEARNING OF PARAMETRIC MEAN FIELD GAMES WITH COMMON NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

We study the inverse problem of recovering unknown parameters governing the dynamics and rewards of Mean Field Games (MFGs) from partial observations of population trajectories, made stochastic by common noise. The observer seeks to recover a consistent parameter θ via Bayesian inference, facing the challenge that evaluating the likelihood requires solving a new Nash equilibrium (NE) for each candidate θ. Due to common noise, the mean-field evolution is stochastic and equilibrium policies are population dependent, which makes solving for each parameter computationally expensive. To overcome this, we introduce the *Bayesian-Extended Mean Field Game* (BE-MFG), whose single NE simultaneously encodes equilibrium policies for all parameters, enabling efficient learning of the conditional density of mean-field trajectories. We formalize parameter consistency and identifiability in MFGs with common noise and investigate conditions under which consistent parameter recovery is guaranteed. On three synthetic environments of various complexity, we recover the ground-truth parameter accurately from a small number of partial, stochastic observations, outperforming an adaptation of inverse reinforcement learning for MFGs. A fourth experiment, on public electricity market data, suggests that the approach can also be applied to real-world data.

1 INTRODUCTION

Learning in multi-agent environments presents significant hurdles due to the complex, often non-stationary, strategic interactions among agents. While Deep Reinforcement Learning has demonstrated remarkable results in small-scale games like Go (Silver et al., 2016; 2017), Chess (Silver et al., 2018), Poker (Heinrich & Silver, 2016), and complex strategy games such as StarCraft (Samvelyan et al., 2019) or Stratego (Perolat et al., 2022), extending these techniques to massive populations remains challenging. As the population grows, the joint state-action space expands exponentially, rendering traditional Multi-Agent Reinforcement Learning methods computationally infeasible (Buşoniu et al., 2008; Yang & Wang, 2020; Zhang et al., 2021; Gronauer & Diepold, 2022; Wong et al., 2023).

Mean Field Games (MFGs) (Lasry & Lions, 2007; Huang et al., 2006) address this scaling problem by modeling the interaction between a representative individual and the aggregate distribution of the population. MFGs simplify high-dimensional games into a system of coupled equations describing the optimal behavior of a single agent and the macroscopic evolution of the distribution.

This framework is highly effective for applications involving large populations in continuous spaces, such as economics (Lachapelle et al., 2016; Achdou et al., 2017), finance (Cardaliaguet & Lehalle, 2018; Yang et al., 2018), engineering (Djehiche et al., 2017), crowd dynamics (Lachapelle & Wolfram, 2011; Achdou & Porretta, 2018), biological swarming (Nourian et al., 2010; Fornasier et al., 2014), and large-scale network resource management (Hanif et al., 2016; Mao et al., 2022). In many realistic settings, the population is subjected to aggregate, macroscopic shocks (such as market fluctuations or weather events) which are modeled through common noise (Carmona et al., 2016). In recent years, the numerical resolution of high-dimensional MFGs has been advanced by deep learning and reinforcement learning (RL) techniques; see (Laurière et al., 2022; Hu & Laurière, 2024) for reviews. In particular, model-free RL algorithms (Guo et al., 2019; Subramanian & Mahajan, 2019; Elie et al., 2020; Cui & Koeppl, 2021; Angiuli et al., 2022; Yardim et al., 2023) learn equilibrium

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

policies through direct interaction with an environment implementing the MFG, but do not address the question of building this environment from real data.

Challenges. Inferring unknown model parameters from observed trajectories is central to real-world applications where the game's specifications are not known a priori. This inverse problem is closely related to Inverse Reinforcement Learning (IRL) and Imitation Learning (IL) for MFGs, which recover reward functions or dynamics from observed equilibrium trajectories (Chen et al., 2022; 2023; Anahtarci et al., 2024), but without common noise or with access to many individual state-action trajectories. The latter is often unrealistic, e.g., due to data protection constraints in epidemiology, economics, or finance. Moreover, when the environment incorporates common noise and the latent parameter θ is unknown, standard IRL techniques struggle for two reasons: (1) the observer only witnesses stochastic realizations of the mean-field trajectory, necessitating a probabilistic approach to identifiability and inference; (2) the equilibrium policies are in general population-dependent, hence more complex. In short, the main challenges we tackle are:

1. Learning MFG parameters in both the **dynamics** and the **reward**;
2. Using only **(state) population-level information**, possibly limited to certain time steps;
3. Extracting information from **noisy mean-field trajectories** due to common noise;
4. Handling **population-dependent policies** due to mean field stochasticity.

Main Contributions. In this paper, we address the challenge of learning the latent parameters of an MFG with common noise from observed population trajectories, either state-action or state-only, and potentially truncated in time. Our main contributions, summarized through Fig. 1, are:

1. We formalize the problem of **parameter identifiability** in MFGs with common noise, establishing **theoretical conditions** under which an observer can recover a parameter consistent with the observations, and the true parameter θ_0 under identifiability, using Bayesian estimation, despite the latent nature of the common noise.
2. We introduce the **Bayesian-Extended Mean Field Game (BE-MFG)**, a novel augmented game formulation that embeds the entire parametric family of MFGs. By computing a single equilibrium for the BE-MFG, we simultaneously solve for the equilibrium policies across all parameters, bypassing the need to repeatedly solve MFGs during the inference phase.
3. We propose a **scalable method** combining a deep Fictitious Play (FP) algorithm with a Feature-wise Linear Modulation (FiLM) architecture and Masked Autoregressive Flows (MAF) to efficiently approximate the posterior distribution over the parameters.
4. We report **numerical experiments** on three synthetic environments of increasing complexity, varying parameter dimensionality, state space size, and observation type, in which the ground-truth parameter is recovered and the method is compared against an adaptation of Mean-Field IRL (MF-IRL) (Chen et al., 2022). A fourth experiment, on public electricity market data, illustrates the **potential applicability** of the pipeline to real-world data.

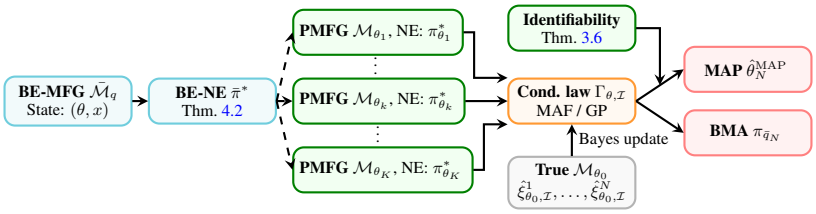

Figure 1: *Solving the BE-MFG once yields a Nash equilibrium that simultaneously encodes equilibrium policies for all parameters; one can simulate trajectories to learn the conditional density, which, combined with the true observations via Bayes' rule, gives the posterior, and hence the MAP estimator and the BMA policy.*

Related Work. Our work relates primarily to IL and IRL in MFGs ((Shmaya & Ziliotto, 2025) considers players performing Bayesian learning during the game, a different question). IL for MFGs, which infers the equilibrium policies used by experts, has been studied in (Ramponi et al., 2023; Lambrecht & Laurière, 2026) from a theoretical perspective. On the IRL side, (Chen et al., 2022; 2023) recover reward functions from observed behaviors. Recent extensions have incorporated Maximum Causal Entropy (Anahtarci et al., 2024; 2025b), average reward criteria (Alkır & Saldi, 2025), and kernel-based methods (Anahtarci et al., 2025a) to improve robustness and representational capacity. However, these methods typically assume deterministic mean-field at equilibrium (no

common noise) and focus solely on the reward function, leaving the transition dynamics fixed and known. Furthermore, they often require computing the forward MFG equilibrium in the inner loop of the inference, which is computationally prohibitive; meta-IRL (Chen et al., 2024) still relies on nested optimization, and model-based RL for MFGs (Huang et al., 2024) learns dynamics by exploration rather than from expert trajectories. In contrast, our Bayesian framework handles general parametric uncertainty (encompassing both rewards and dynamics) under common noise. Crucially, by formulating the inference problem through the Bayesian-Extended MFG, our method uniquely bypasses the inner-loop equilibrium computation, solving for all parameterized policies in a single forward phase.

Organization of the Paper. In Sec. 2, we formalize the setting of Parametric Mean Field Games (PMFGs). In Sec. 3, we define the concepts of consistency and identifiability. Sec. 4 presents the Bayesian learning setting, introduces the BE-MFG, establishes its equivalence to the family of PMFGs and details our algorithmic approach. Finally, Sec. 5 demonstrates the effectiveness of our method through numerical experiments.

2 Problem Setting

We consider a family of MFGs indexed by a parameter θ, and aim to recover the true parameter θ_0, its Nash equilibrium policy, or best responses, from observations of mean-field trajectories.

2.1 Mean field dynamics and Nash equilibria

Notation. For any integer N, $[N] = \{0, \dots, N\}$ and $[N]^* = [N] \setminus \{0\}$. $\Delta_{\mathcal{S}}$ (resp. $\mathcal{P}(E)$) denotes the set of probability measures over a finite set $\mathcal{S}$ (resp. a measurable space E). We otherwise use standard notation, all of which is defined, together with the abbreviations, in App. A.

Main components. Let $\mathcal{X}$ be a finite **state space**, $\mathcal{A}$ be a finite action space, $\mathcal{E}^0$ be a Polish space for the **common noise**, $\nu^0 \in \mathcal{P}(\mathcal{E}^0)$ be a probability measure for the **common noise law**, $H \in \mathbb{N}^*$ be the **time horizon**, and $\rho_0 \in \Delta_{\mathcal{X}}$ be the **initial probability measure**. Let $d \in \mathbb{N}^*$, $\bar{\Theta} \subset \mathbb{R}^d$ be a compact **set of parameters**, and λ a reference measure on $\mathcal{B}(\bar{\Theta})$. Let $P : \bar{\Theta} \times \mathcal{X} \times \mathcal{A} \times \Delta_{\mathcal{X}} \times \mathcal{E}^0 \to \Delta_{\mathcal{X}}$ and $r : \bar{\Theta} \times \mathcal{X} \times \mathcal{A} \times \Delta_{\mathcal{X}} \to \mathbb{R}$ be two measurable mappings representing the **transition kernel** and the **reward function**, r being moreover bounded. For any $\theta \in \bar{\Theta}$, for any $(x, a, \rho, e^0) \in \mathcal{X} \times \mathcal{A} \times \Delta_{\mathcal{X}} \times \mathcal{E}^0$ define $P_\theta(x, a, \rho, e^0) = P(\theta, x, a, \rho, e^0)$ and $r_\theta(x, a, \rho) = r(\theta, x, a, \rho)$. A **Parametric Mean Field Game (PMFG)** associated to the parameter $\theta \in \bar{\Theta}$ is denoted by a tuple $\mathcal{M}_\theta = (\mathcal{X}, \mathcal{A}, \mathcal{E}^0, \nu^0, P_\theta, r_\theta, H, \rho_0)$. We call a **family of PMFG** the set $\{\mathcal{M}_\theta\}_{\theta \in \bar{\Theta}}$. We define a **strategy** as a measurable mapping from $\mathcal{X} \times \Delta_{\mathcal{X}}$ to $\Delta_{\mathcal{A}}$, and a **policy** as a sequence of H strategies. A **policy** is denoted by $\pi = (\pi_0, \dots, \pi_{H-1})$, the set of all strategies is denoted by Π_0 and the set of all such policies is denoted by Π. Elements of $\Delta_{\mathcal{X}}$ and $(\Delta_{\mathcal{X}})^H$ are referred to as **mean field** and **mean field sequences** respectively.

Distribution evolution. Let $\theta \in \bar{\Theta}$ and consider the MFG $\mathcal{M}_\theta$. For any pair of population distributions $\rho, \rho' \in \Delta_{\mathcal{X}}$, any strategy π_0, we define the **mean-field transition operator** $\Phi_\theta : \Delta_{\mathcal{X}} \times \Pi_0 \times \Delta_{\mathcal{X}} \times \mathcal{E}^0 \to \Delta_{\mathcal{X}}$ as

$$\Phi_\theta(\rho', \pi_0; \rho, e^0) = \sum_{x \in \mathcal{X}, a \in \mathcal{A}} \rho'(x) \pi_0(a|x, \rho) P_\theta(\cdot|x, a, \rho, e^0).$$

This operator Φ_θ represents the evolution of the probability distribution of an agent between time i and $i+1$, given that at time i the agent's state distribution was ρ', the agent used strategy π_0, the total population was distributed according to ρ, and the common noise value is e^0. We consider an independent and identically distributed (i.i.d.) sequence of random variables $\varepsilon^0_1, \dots, \varepsilon^0_{H-1} \in \mathcal{E}^0$ for the common noise. We denote the common noise trajectory as $\varepsilon^0 = (\varepsilon^0_1, \dots, \varepsilon^0_{H-1})$. For any $\pi \in \Pi$, we define the mean field sequence $\rho_\theta^{\pi, \varepsilon^0} = (\rho_{\theta,0}^{\pi,\varepsilon^0}, \rho_{\theta,1}^{\pi,\varepsilon^0}, \dots, \rho_{\theta,H-1}^{\pi,\varepsilon^0})$ by $\rho_{\theta,0}^{\pi,\varepsilon^0} = \rho_0$ and

$$\rho_{\theta,i+1}^{\pi,\varepsilon^0} = \Phi_\theta(\rho_{\theta,i}^{\pi,\varepsilon^0}, \pi_i\,;\, \rho_{\theta,i}^{\pi,\varepsilon^0}, \varepsilon^0_{i+1}), \qquad i \in [H-2]. \tag{1}$$

This sequence represents the evolution of the population distribution when all agents follow policy π and are affected by the common noise trajectory $\varepsilon^0_{1:H-1}$. For any pair of policies $(\pi, \pi') \in \Pi^2$, we define the mean field sequence $\rho_\theta^{\pi,\pi',\varepsilon^0} = (\rho_{\theta,0}^{\pi,\pi',\varepsilon^0}, \rho_{\theta,1}^{\pi,\pi',\varepsilon^0}, \dots, \rho_{\theta,H-1}^{\pi,\pi',\varepsilon^0})$ by: $\rho_{\theta,0}^{\pi,\pi',\varepsilon^0} = \rho_0$,

$$\rho_{\theta,i+1}^{\pi,\pi',\varepsilon^0} = \Phi_\theta(\rho_{\theta,i}^{\pi,\pi',\varepsilon^0}, \pi'_i; \rho_{\theta,i}^{\pi,\varepsilon^0}, \varepsilon^0_{i+1}), \qquad i \in [H-2]. \tag{2}$$

This sequence represents the evolution of the state distribution of a *single agent* following policy π' when all other agents follow policy π and are affected by the common noise trajectory ε^0.

Best response and Nash equilibrium. When the population of agents follows the policy π, a single agent wishes to find a **best response** to π, i.e., a policy π' maximizing the **value function**:

$$V_\theta(\pi';\pi) = \mathbb{E}\Big[\sum_{i=0}^{H-1} r_\theta(X_i, \alpha_i, \rho_{\theta,i}^{\pi,\varepsilon^0})\Big], \tag{3}$$

where the process is defined by $X_0 \sim \rho_0$, $X_{i+1} \sim P_\theta(\cdot|X_i, \alpha_i, \rho_{\theta,i}^{\pi}, \varepsilon_{i+1}^0)$, and $\alpha_i \sim \pi'(\cdot|X_i, \rho_{\theta,i}^{\pi})$. Note that under these conditions, the law of the state X_i (conditioned on the common noise) is $\rho_{\theta,i}^{\pi,\pi',\varepsilon^0}$. We denote the law of state-action pair (X_i, α_i) by $\mu_{\theta,i}^{\pi,\pi',\varepsilon^0} \in \Delta_{\mathcal{X}\times\mathcal{A}}$. If $\pi = \pi'$, we will simply denote $\mu_{\theta,i}^{\pi,\varepsilon^0} = \mu_{\theta,i}^{\pi,\pi,\varepsilon^0}$. In the sequel, we will often omit to write explicitly the common noise ε^0, but the sequences ρ and μ are random variables due to ε^0.

Definition 2.1. *Let $\theta \in \bar{\Theta}$. For $\pi \in \Pi$, the set of* **best responses** *to the PMFG $\mathcal{M}_\theta$ is* $\mathtt{BR}_\theta(\pi) = \arg\max_{\pi'\in\Pi} V_\theta(\pi';\pi)$. *$\pi \in \Pi$ is a* **(mean field) Nash equilibrium** *(NE) for $\mathcal{M}_\theta$ if $\pi \in \mathtt{BR}_\theta(\pi)$.*

This is a fixed-point condition: it implies that when the entire population follows policy π, the resulting mean-field flow ρ_θ^π creates a reward and transition structure such that the same policy π remains optimal for any individual agent.

Assumption 1. *For every $\theta \in \bar{\Theta}$, if π and π' are two Nash equilibria to $\mathcal{M}_\theta$, then $\mathcal{L}(\mu_\theta^\pi) = \mathcal{L}(\mu_\theta^{\pi'})$.*

The existence and uniqueness of the NE for finite-state, discrete-time MFGs has been proved under various assumptions; see e.g. (Cui & Koeppl, 2021) for the finite-horizon setting and (Saldi et al., 2018; Anahtarci et al., 2023) for the infinite horizon discounted setting. Moreover, our approach only requires uniqueness of the law of the state-action mean field $\mathcal{L}(\mu_\theta^*)$. In the sequel we work under Assumption 1 and we will focus solely on equality of laws. For this reason, we are not interested in a specific policy being a NE, and we will denote by $\rho_\theta^* = \rho_\theta^{\pi_\theta^*,\varepsilon^0}$ (resp. by $\mu_\theta^* = \mu_\theta^{\pi_\theta^*,\varepsilon^0}$) a mean field trajectory (resp. the state-action mean field trajectory) associated to some NE policy π_θ^*. Both are random variables due to the common noise.

Observations. We will consider two types of observations: the observation of state-action mean-fields trajectories μ, and the observation of only state mean-fields trajectories ρ. To have a common framework and simplify the notation, we will denote the observations by ξ and specify case by case whether it represents μ or ρ. We denote the space of ξ by Ξ, referring either to $\Delta_\mathcal{X}$ or $\Delta_{\mathcal{X}\times\mathcal{A}}$. For any $\theta \in \bar{\Theta}$, we will write ξ_θ^* a realization of the state or state-action mean-fields trajectory, associated to the unique law of the NE (Assumption 1).

In the following, we will consider situations where the observer has only access to some time steps in the mean-fields trajectories. For this purpose, let us introduce the following notation: for any set of accessible time steps from the observer $\mathcal{I} \subset [H-1]$ and any type of observation $\xi \in \Xi$, let $\xi_\mathcal{I} = \{\xi_i\}_{i\in\mathcal{I}}$. Recalling that $\xi_{\theta,\mathcal{I}}^*$ is random due to the common noise, let $\mathbb{P}_{\theta,\mathcal{I}}$ denote the probability measure over $\Xi^{|\mathcal{I}|}$ of $\xi_{\theta,\mathcal{I}}^*$.

2.2 Objectives

Our objective is as follows. Let θ_0 be the "true parameter" unknown to the observer, $\mathcal{I} \subset [H-1]$ be the subset of accessible snapshots. Assume that an observer sees realizations $\hat{\xi}_{\theta_0,\mathcal{I}}^1, \dots, \hat{\xi}_{\theta_0,\mathcal{I}}^N \in \Xi^{|\mathcal{I}|}$, where for any $n \in [N]^*$, $\hat{\xi}_{\theta_0,\mathcal{I}}^n$ has been generated from $\pi_{\theta_0}^*$. The observer tries to recover a set of parameters that is consistent with the observations. We study four questions:

(Q1) *How should consistency and identifiability be formalized in the presence of common noise?*
(Q2) *Can the observer recover a parameter θ consistent with the observed trajectory?*
(Q3) *Can the observer recover the NE policy $\pi_{\theta_0}^*$?*
(Q4) *Can the observer reconstruct a best response against the NE policy $\pi_{\theta_0}^*$?*

We address **(Q1)** in the next section. To answer **(Q2)**-**(Q4)** we propose **two Bayesian approaches**: i) The **maximum a posteriori (MAP)** approach, ii) The **Bayesian model averaging (BMA)** approach in Sec. 4.

3 CONSISTENCY AND IDENTIFIABILITY

Injectivity from the parameter space to the observation space is often not verified. Consequently, we first introduce the notions of consistency and identifiability as necessary conditions for the inverse problem of MFGs to be well-posed.

Parameter consistency. Suppose an observer witnesses realizations of $\xi^*_{\theta_0,\mathcal{I}}$ for some unknown $\theta_0 \in \bar{\Theta}$, without knowledge of the specific common noise realization $\varepsilon^0_{1:H-1}$. The observer may seek a parameter to explain these observations; however, the latent nature of the common noise path forces them to consider equality in law.

Definition 3.1. *Let $\theta, \theta' \in \bar{\Theta}$. θ and θ' are* **consistent with respect to $\mathcal{I}$ for the observation ξ** *if and only if $\mathbb{P}_{\theta,\mathcal{I}} = \mathbb{P}_{\theta',\mathcal{I}}$. With a slight abuse of terminology, we say that θ is* **consistent with respect to $\xi^*_{\theta',\mathcal{I}}$** *if and only if θ and θ' are consistent with respect to $\mathcal{I}$ for the observation ξ.*

This notion of consistency, which refers to the equality of the laws of the observations, should not be confused with the statistical consistency of an estimator.

Remark 3.2. *Even if the dependency has not been made explicit in the last definition, the consistency is dependent on the type of observations considered ($\xi = \rho$ or $\xi = \mu$).*

Def. 3.1 provides a natural framework for identifiability. We seek to address the following fundamental question: *Does recovering a parameter that is consistent with respect to ξ allow for the recovery of the original NE $\pi^*_{\theta_0}$?* We give a natural situation where recovering a consistent parameter is equivalent to finding a NE, and we provide the proof in App. B.1.

Theorem 3.3. *Consider the observation $\xi = \mu$ with $\mathcal{I} = [H-1]$. Suppose that Assumption 1 holds, that $P_\theta = P$ for every $\theta \in \bar{\Theta}$, and that there is a measurable mapping Λ, not depending on θ, such that $\varepsilon^0 = \Lambda(\mu^*_{\theta,\mathcal{I}})$ $\mathbb{P}$-a.s. for every $\theta \in \bar{\Theta}$. Then, for any $\theta, \theta' \in \Theta$, if $\mathbb{P}_{\theta,\mathcal{I}} = \mathbb{P}_{\theta',\mathcal{I}}$, then any NE π^*_θ for $\mathcal{M}_\theta$ is a NE for $\mathcal{M}_{\theta'}$.*

Remark 3.4. *The common noise is recoverable from the observation in particular when it is carried as a* **shared state component**. *This is the case, e.g., of the Beach Bar process of (Perrin et al., 2020), in which the bar is alternatively open and closed , the status of the bar being shared by all the agents, and of the electricity model of App. I. Without common noise, the condition is trivially satisfied.*

Parameter identifiability. Another natural question is: *When does consistency imply equality?* Consider a family of PMFGs where consistency is equivalent to equality. Then, the problem of recovering the equilibrium is equivalent to the problem of recovering a parameter consistent with some observations. We highlight that these problems have not been considered in previous work in the context of MFGs. We introduce the definition of **Identifiable Family of PMFG**, prove a result to characterize such models, and provide an example in App. B.1.

Definition 3.5. *The family of PMFGs $\{\mathcal{M}_\theta\}_{\theta\in\bar{\Theta}}$ is* **identifiable with respect to $\mathcal{I}$ for the observation** *ξ if for any parameters $\theta, \theta' \in \bar{\Theta}$, their consistency w.r.t. $\mathcal{I}$ for the observation ξ implies their equality.*

To discuss when Definition 3.5 is satisfied, we extend the mapping $\mathtt{BR}$ to laws of partial mean-field trajectories. We will use the following notation: $\mathfrak{P}_{\theta,\mathcal{I}} = \{\mathbb{P}'_\mathcal{I} \in \mathcal{P}(\Delta_\mathcal{X}^{|\mathcal{I}|}) \mid \exists \pi \in \Pi : \mathcal{L}(\rho^\pi_{\theta,\mathcal{I}}) = \mathbb{P}'_\mathcal{I}\}$ which is the set of all state distributions that can be generated by some policy. For any $\theta \in \bar{\Theta}$, for any $\mathbb{P}_\mathcal{I} \in \mathfrak{P}_{\theta,\mathcal{I}}$, we define $\mathbb{BR}_\theta(\mathbb{P}_\mathcal{I}) = \{\mathcal{L}(\rho^{\pi,\pi'}_{\theta,\mathcal{I}}) \mid \pi' \in \mathtt{BR}_\theta(\pi), \pi \in \Pi : \mathcal{L}(\rho^\pi_{\theta,\mathcal{I}}) = \mathbb{P}_\mathcal{I}\}$. Notice that the output of this mapping is not a set of policies, but a set of distributions of mean field trajectories. The definition takes the union over all policies π inducing the law $\mathbb{P}_\mathcal{I}$, and all the policies $\pi' \in \mathtt{BR}_\theta(\pi)$ so that $\mathbb{BR}_\theta(\mathbb{P}_\mathcal{I})$ is well-defined. Theorem 3.6 gives what we believe is the first result on identifiability for inverse problem in discrete-time MFGs with common noise. The proof is provided in App. B.1.

Theorem 3.6. *If for any $\theta \neq \theta' \in \bar{\Theta}$, for any $\mathbb{P}_\mathcal{I} \in \mathfrak{P}_{\theta,\mathcal{I}} \cap \mathfrak{P}_{\theta',\mathcal{I}}$, we have $\mathbb{BR}_\theta(\mathbb{P}_\mathcal{I}) \cap \mathbb{BR}_{\theta'}(\mathbb{P}_\mathcal{I}) = \emptyset$, then $\{\mathcal{M}_\theta\}_{\theta\in\bar{\Theta}}$ is identifiable.*

Corollary 3.7. *If for any $\theta \neq \theta' \in \bar{\Theta}$, $\mathfrak{P}_{\theta,\mathcal{I}} \cap \mathfrak{P}_{\theta',\mathcal{I}} = \emptyset$, then $\{\mathcal{M}_\theta\}_{\theta\in\bar{\Theta}}$ is identifiable.*

4 BAYESIAN SETTING

4.1 MAXIMUM A POSTERIORI AND BAYESIAN MODEL AVERAGING ESTIMATORS

To use the Bayesian approach, the observer wishes to consider a prior q over $\bar{\Theta}$, such that $q(\theta) > 0$ for each $\theta \in \bar{\Theta}$. Let $\mathcal{I} \subset [H-1]$ be the time steps accessible to the observer. Recall that $\mathbb{P}_{\theta,\mathcal{I}}$ denotes the probability measure over $\Xi^{|\mathcal{I}|}$ of $\xi^*_{\theta,\mathcal{I}}$. Let Θ be a $\bar{\Theta}$-valued random variable with law q, and let $\mathbb{P}_{\mathcal{I}} = \int_{\bar{\Theta}} \mathbb{P}_{\theta,\mathcal{I}} q(\theta)\lambda(d\theta)$, which is the probability measure of $\xi^*_{\Theta,\mathcal{I}}$. To define a likelihood, we need $\mathbb{P}_{\theta,\mathcal{I}} \ll \mathbb{P}_{\mathcal{I}}$, which is ensured by the following assumption.

Assumption 2. *One of the following conditions holds:* ***(C1)*** *$\bar{\Theta}$ is at most countable and λ is the counting measure;* ***(C2)*** *for any $A \in \mathcal{B}(\Xi^{|\mathcal{I}|})$, the mapping $\theta \mapsto \mathbb{P}_{\theta,\mathcal{I}}(A)$ is continuous on $\bar{\Theta}$, and $\lambda(U) > 0$ for any non-empty open subset U of $\bar{\Theta}$.*

Lemma 4.1. *Under Assumption 2, for every $\theta \in \bar{\Theta}$, $\mathbb{P}_{\theta,\mathcal{I}} \ll \mathbb{P}_{\mathcal{I}}$, and the Radon-Nikodym derivative $\Gamma_{\theta,\mathcal{I}} = d\mathbb{P}_{\theta,\mathcal{I}}/d\mathbb{P}_{\mathcal{I}}$ exists.*

The proof is provided in App. B.2. Suppose that the observer has access to $N \in \mathbb{N}^*$ independent observations $\hat{\xi}^1_{\theta_0,\mathcal{I}}, \dots, \hat{\xi}^N_{\theta_0,\mathcal{I}}$ of the possibly truncated mean field trajectory at equilibria of $\mathcal{M}_{\theta_0}$, where $\theta_0 \in \bar{\Theta}$ is the true parameter. If the observer has full knowledge of $\Gamma_{\theta,\mathcal{I}}$ for any $\theta \in \bar{\Theta}$, then they can compute the posterior associated to Θ using Bayes' rule:

$$\bar{q}_N(\theta) = q(\theta)\frac{\prod_{n=1}^N \Gamma_{\theta,\mathcal{I}}(\hat{\xi}^n_{\theta_0,\mathcal{I}})}{\Gamma^N_{\mathcal{I}}(\hat{\xi}^1_{\theta_0,\mathcal{I}}, \dots, \hat{\xi}^N_{\theta_0,\mathcal{I}})}. \tag{4}$$

where $\Gamma^N_{\mathcal{I}}(\xi^{*,1}_{\mathcal{I}}, \dots, \xi^{*,N}_{\mathcal{I}}) = \int_{\bar{\Theta}} q(\theta) \prod_{n=1}^N \Gamma_{\theta,\mathcal{I}}(\xi^{*,n}_{\mathcal{I}})\lambda(d\theta)$. The observer can compute the **MAP estimator** $\hat{\theta}^{\mathrm{MAP}}_N = \arg\max_{\theta\in\bar{\Theta}} \bar{q}_N(\theta)$. Then, they can use this recovered parameter to compute the associated NE of $\mathcal{M}_{\hat{\theta}^{\mathrm{MAP}}_N}$ and sample new mean-field trajectories. But they could also use the posterior distribution to construct an averaged policy $\pi_{\bar{q}_N}$ that would correspond to a **BMA method**,

$$\pi_{\bar{q}_N}(a|x,\rho) = \mathbb{E}_{\Theta\sim\bar{q}_N}\left[\pi^*_\theta(a|x,\rho)\right], \quad \forall (x,a,\rho) \in \mathcal{X}\times\mathcal{A}\times\Delta_{\mathcal{X}}.$$

The cornerstone of this method is to be able to compute $\Gamma_{\theta,\mathcal{I}}$ for any $\theta \in \bar{\Theta}$. A naive approach would be to compute a NE on each node of a grid over $\bar{\Theta}$. But computing a single NE is computationally expensive (Perrin et al., 2020; Pérolat et al., 2022; Perrin et al., 2022), and this approach would not work: we show quantitatively in App. J that already for a one-dimensional parameter, a single BE-MFG solve is cheaper than a grid of dedicated solves and recovers the parameter more accurately, and that a grid search becomes out of reach in higher dimension. To address this issue, we propose to compute a NE for an augmented MFG.

4.2 BAYESIAN-EXTENDED MEAN FIELD GAMES

We now introduce the concept of BE-MFG, which can be interpreted as a game in which a parameter $\Theta \sim q$ is sampled once at time 0 and is common to all agents. This parameter remains fixed throughout the game and is observed by all agents. Conditionally on $\Theta = \theta$, the game reduces exactly to the PMFG $\mathcal{M}_\theta$. Solving for a NE in the BE-MFG will allow us to recover, simultaneously for all θ, the corresponding equilibrium policies.

4.2.1 DEFINITIONS

Let us start by introducing the augmented dynamic kernel $\bar{P}$ and reward $\bar{r}$. For any $(\theta, x, a, \bar{\rho}, e^0, \theta', x') \in \bar{\Theta}\times\mathcal{X}\times\mathcal{A}\times\mathcal{P}(\bar{\Theta}\times\mathcal{X})\times\mathcal{E}^0\times\bar{\Theta}\times\mathcal{X}$, we define

$$\bar{P}(\theta', x' \mid \theta, x, a, \bar{\rho}, e^0) = \delta_{\theta=\theta'} P_\theta\big(x' \mid x, a, \textstyle\int_{\bar{\Theta}} \bar{\rho}(d\theta'', \cdot), e^0\big), \ \bar{r}(\theta, x, a, \bar{\rho}) = r_\theta\big(x, a, \textstyle\int_{\bar{\Theta}} \bar{\rho}(d\theta', \cdot)\big). \tag{5}$$

Let q be a probability distribution on $(\bar{\Theta}, \mathcal{B}(\bar{\Theta}))$. The **Bayesian-Extended Mean Field Game (BE-MFG)** associated to the family of PMFGs $\{\mathcal{M}_\theta\}_{\theta\in\bar{\Theta}}$ and the prior q is the tuple $\bar{\mathcal{M}}_q = (\bar{\Theta}\times\mathcal{X}, \mathcal{A}, \mathcal{E}^0, \nu^0, \bar{P}, \bar{r}, H, Q^{\rho_0})$, where $Q^{\rho_0} = q \circ (\delta_\cdot \otimes \rho_0)^{-1}$ is the pushforward of q under the map $\theta \mapsto \delta_\theta \otimes \rho_0$, which defines a probability measure on $\mathcal{P}(\bar{\Theta}\times\mathcal{X})$. We define a **Bayesian strategy** as a measurable mapping from $\bar{\Theta}\times\mathcal{X}\times\mathcal{P}(\bar{\Theta}\times\mathcal{X})$ to $\Delta_{\mathcal{A}}$. The set of Bayesian strategies is denoted $\bar{\Pi}_0$. We define a **Bayesian policy** as a sequence of H Bayesian strategies. A **Bayesian policy** is denoted by $\bar{\pi} = (\bar{\pi}_0, \dots, \bar{\pi}_{H-1})$, and the set of all such policies is denoted by $\bar{\Pi}$.

The distribution evolution, value function, best responses and NE of the BE-MFG are defined exactly as for the PMFG (Sec. 2), replacing the state space $\mathcal{X}$ by the augmented state space $\bar{\Theta} \times \mathcal{X}$, the mappings P_θ, r_θ by $\bar{P}$,$\bar{r}$, and the policy class Π by $\bar{\Pi}$. A NE of $\bar{\mathcal{M}}_q$ is called a **Bayesian-Extended Nash equilibrium (BE-NE)**. Full details are provided in Appendix B.3. We now proceed to the link between the original PMFG and the BE-MFG.

4.2.2 Link between the models

Since Θ is common to all agents, the evolution of the system conditional on $\Theta = \theta$ coincides with that of $\mathcal{M}_\theta$. Any Bayesian policy $\bar{\pi}$ thus induces, for each θ, a policy π_θ in $\mathcal{M}_\theta$, and conversely any family $\{\pi_\theta\}_{\theta \in \bar{\Theta}}$ is lifted to a Bayesian policy. More precisely, the mean-field flow satisfies $\bar{\rho}^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_\Theta^{\pi_\Theta,\varepsilon^0}$ $\mathbb{P}$-a.s., and the value functions satisfy $\bar{V}_q(\bar{\pi}';\bar{\pi}) = \mathbb{E}_{\Theta\sim q}[V_\Theta(\pi'_\Theta;\pi_\Theta)]$, where $\bar{\rho}^{\bar{\pi},\varepsilon^0}$ and $\bar{V}_q$ are the mean-field flow and value function of $\bar{\mathcal{M}}_q$ (App. B.3). These two identities, established in Lemmas B.3–B.8 in App. B.4, are the key tools for establishing the main result of this section:

Theorem 4.2. *Let $\bar{\pi}^* \in \bar{\Pi}$ be a NE of $\bar{\mathcal{M}}_q$. Define, for any $\theta \in \bar{\Theta}$ and $(i,x,a,\rho) \in [H-1] \times \mathcal{X} \times \mathcal{A} \times \Delta_\mathcal{X}$, $\pi^{**}_{\theta,i}(a \mid x,\rho) = \bar{\pi}^*_i(a \mid \theta, x, \delta_\theta \otimes \rho)$. Then π^{**}_Θ is a NE of $\mathcal{M}_\Theta$, q-a.s.*

Under condition (C1) of Assumption 2 Theorem 4.2 holds for every $\theta \in \bar{\Theta}$.

Combined with Assumption 1, this result is very useful: Given access to $\bar{\pi}^*$, one obtains for any $\theta \in \bar{\Theta}$ a sampler of $\xi^*_{\theta,\mathcal{I}}$: it suffices to simulate the policy π^{**}_θ defined above under a drawn common noise ε^0. Indeed, Assumption 1 guarantees that $\mathcal{L}(\xi^{\pi^{**}_\theta}_{\theta,\mathcal{I}}) = \mathcal{L}(\xi^*_{\theta,\mathcal{I}}) = \mathbb{P}_{\theta,\mathcal{I}}$, so that the simulated trajectories are distributed according to the true conditional law $\Gamma_{\theta,\mathcal{I}}$. Yet $\bar{\mathcal{M}}_q$ is a standard MFG with common noise, and can be solved with the same algorithms as any individual PMFG $\mathcal{M}_\theta$, at a comparable computational cost (see App. B.4 and App. J).

4.3 Algorithms

Learning via BE-MFG Pipeline. The methodology consists of three steps (Algorithm 1, App. C):

1. **Computing the BE-NE.** Select a prior q over $\bar{\Theta}$ and compute the BE-NE $\bar{\pi}^*$ of $\bar{\mathcal{M}}_q$, using a FiLM architecture (Perez et al., 2018) (App. C.3) and the FP and deep learning pipeline of (Lambrecht & Laurière, 2026) (App. C.1). By Theorem 4.2, this single solve grants access to $\{\pi^{**}_\theta\}_{\theta\in\bar{\Theta}}$.
2. **Learning the conditional density.** Train a parametric approximation $\Gamma^\gamma_{\theta,\mathcal{I}}$, with trainable parameters γ, of the conditional density $\Gamma_{\theta,\mathcal{I}}$ by simulating trajectories from $\bar{\pi}^*$, using a MAF (Papamakarios et al., 2017) (App. E.3). As a lightweight alternative, a Gaussian Process (GP) inverse regression, trained on the same simulated trajectories, directly approximates the posterior (App. E.2).
3. **Bayesian inference and policy deployment.** Given observations $(\hat{\xi}^n_{\theta_0,\mathcal{I}})_{1\le n\le N}$, compute the posterior $\bar{q}_N$ via Bayes' rule, derive $\hat{\theta}^{\mathrm{MAP}}_N$ and $\pi_{\bar{q}_N}$, and deploy both directly using $\bar{\pi}^*$, i.e., using π^{**}_θ in place of π^*_θ.

Baseline: Mean Field Inverse Reinforcement Learning. As a baseline, we use the method introduced by (Chen et al., 2022). This method, as all existing methods in the literature to the best of our knowledge, requires access to full state-action observations $\xi = \mu$ and is not designed for the common noise setting. We adapted it to our framework following App. D. As discussed in Sec. 4 and quantified in App. J, we do not include a grid search over dedicated PMFG solves as a baseline, because its computational cost makes it uncompetitive.

5 Numerical Experiments

5.1 Metrics and Observations

We will evaluate the performance of our methods for each question using the following four metrics. Below, $\hat{\pi}$ and $\hat{\mathbb{P}}_\mathcal{I}$ refer respectively to the policy obtained and the induced probability measure of ξ, using any of the method (MAF with MAP, MAF with BMA, etc.). Each experiment was conducted using five seeds.

- **Law Identifiability (Q1):** Wasserstein 1 distance, $\mathcal{W}_1(\mathbb{P}_{\theta_0,\mathcal{I}}, \hat{\mathbb{P}}_{\mathcal{I}})$.
- **Parameters Inference (Q2):** For each example, we will introduce a specific metric to evaluate the inference.
- **Equilibrium Quality (Q3):** Exploitability, $\mathcal{E}_{\theta_0}(\hat{\pi}) = \sup_{\pi\in\Pi} V_{\theta_0}(\pi;\hat{\pi}) - V_{\theta_0}(\hat{\pi};\hat{\pi})$.
- **Performance vs Expert (Q4):** $V_{\theta_0}(\hat{\pi};\pi^*_{\theta_0}) - V_{\theta_0}(\pi^*_{\theta_0};\pi^*_{\theta_0})$.

Observations: We consider for the Bayesian setting $\xi = \rho$, while the MF-IRL needs intrinsically more informative observation $\xi = \mu$. We let $N \in \{1, 10, 100\}$, and consider three possible sets of available time steps: the full trajectory $\mathcal{I} = [H-1]$, denoted as '**Full**'; subsampling every fourth step from the terminal time $\mathcal{I} = \{H-1, H-5, \dots\}$, denoted as '**Sparse**'; and $\mathcal{I} = \{H-1\}$ denoted as '**Terminal**'.

5.2 Experiment 1: Proof of Concept

We consider a variant of the *Beach Bar* environment (Perrin et al., 2022). In this setting, the bar is permanently open and the agents are subject to a common noise shifting their positions: at each time step and at each location, a random shift of the agents is activated with probability $\eta \geq 0$, the **common noise intensity**. In this example, θ is one-dimensional and appears in the reward; however the observation space is high-dimensional due to the number of states and time steps. As this environment is standard in MFGs, we defer its detailed description to App. F.

Bayesian setup: We take the uniform prior q on $\bar{\Theta} = [0, 2.5]$ and $\theta_0 \in \{0.5, 0.875, 1.25, 1.625, 2.0\}$. We present the results as a function of η, aggregated over θ_0, for $\xi = \rho$; the full sensitivity analysis in η and N is in App. F.

Parameter Inference Error: We measure the accuracy of parameter inference via the normalized error $|\hat{\theta} - \theta_0|/2.5$, expressed as a percentage, where θ_0 is the true parameter and $\hat{\theta}$ is the parameter inferred by the method under consideration (MAF, GP, or MF-IRL).

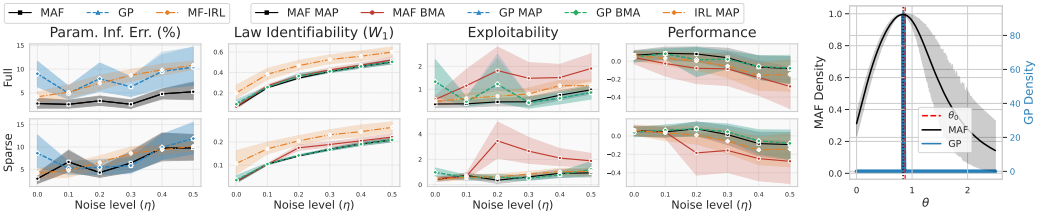

Figure 2: *Beach Bar, $\xi = \rho$, $N = 100$.* ***Left:*** *metrics as a function of noise level η for two available timestep schemes.* ***Right:*** *posterior $\bar{q}_N$ for $\eta = 0.2$, $\theta_0 = 0.875$, $\mathcal{I} = [H-1]$.*

Results. Figure 2 shows that with $N = 100$ state-only observations ($\xi = \rho$), the MAF estimator recovers θ_0 with high accuracy even under the **Sparse** regime, corresponding to only five time steps $i \in \{3, 7, 11, 15, 19\}$ out of $H = 20$, and outperforms MF-IRL on all metrics, despite MF-IRL having access to the more informative $\xi = \mu$. Across all settings, MAP outperforms BMA for both MAF and GP, suggesting that with $N = 100$ the learned conditional densities remain insufficiently sharp for reliable posterior sampling. Figure 2 confirms this picture at the posterior level: even for $N = 100$, the likelihood curves remain spread. However, the MAP estimate aligns closely with the true $\theta_0 = 0.875$ even for $N = 1$. Overall, these results demonstrate the *reliability* of the proposed method. From sparse, state-only, and potentially partial observations, without any knowledge of agent actions, the observer successfully recovers the parameter, reproduces the law of the observed densities, derives a near-optimal policy, and recovers a NE under θ_0. These findings go significantly beyond existing inverse MFG results, which require either no common noise or direct access to state-action trajectories.

5.3 Experiment 2: Higher Parametric Dimension Space

We now demonstrate our method on a higher-dimensional inference problem, where both the spatial preferences and congestion aversion of agents are unknown.

Model: Agents move on a 3×3 torus; at each location, a common noise shock of probability η replaces the chosen move by a uniformly random one. The reward combines an attraction $\theta^e \in \mathbb{R}^9_+$ to hotspots and a congestion penalty of intensity $\theta^c \in [0, 5]$; full equations are given in App. G.

Bayesian setup: We consider $\bar{\Theta} = \mathbb{R}^9_+ \times [0, 5]$. The observer considers the prior q being the density of $\mathcal{N}\big(((\frac{1}{9}, \dots, \frac{1}{9}), \frac{5}{2}), 0.01\big)$ while using GP and Dirichlet$(1, \dots, 1) \otimes$ Unif$[0, 5]$ while using MAF.

We vary the ground truth θ_0 across the three scenarios (Bullseye, Dual Core, Corridor) provided in Figure 11 in App. G, showing three realizations of trajectories with $\eta = 0.3$. We refer to this model as the **Entertainment District** environment.

Parameter Inference Error: Since θ^e and θ^c have very different effects on the equilibrium, we distinguish the inference error between these two parameters. For θ^e, we use the District Error (L1) $\sum_x |\hat{\theta}^e(x) - \theta_0^e(x)|$, and for θ^c, we use the normalized error $|\hat{\theta}^c - \theta_0^c|/5$, expressed as a percentage.

Challenges: While the general shape of θ^e is partially recoverable from agent aggregation patterns, finding precise values at both hot and non-hot spots is difficult due to congestion-induced non-trivial behavior. Moreover, the absence of prior knowledge on the congestion level θ^c makes this a genuinely hard 10-dimensional inference problem.

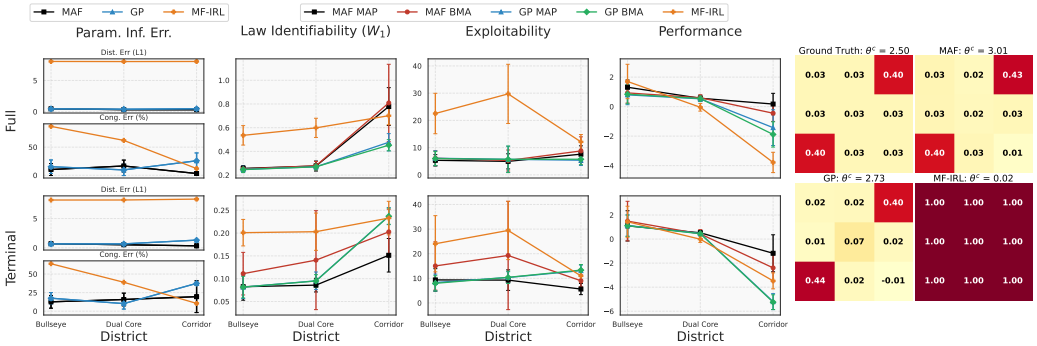

Figure 3: *District, $\xi = \rho$, $N = 1$.* ***Left:*** *metrics as a function of the district scenario for the three sampling schemes.* ***Right:*** *inferred $\theta_0 = (\theta_0^e, \theta_0^c)$ for the Dual Core scenario, $\eta = 0.5$, $\mathcal{I} = [H-1]$.*

Results: Figure 3 reports metrics averaged over $\eta \in \{0.3, 0.5\}$, for each ground truth. The Bayesian approach consistently succeeds across all three scenarios. As shown in Figure 3, it accurately reconstructs the shape of θ^e despite the 10-dimensional nature of the problem, and does so from a single observation ($N = 1$) with remarkably low variance across seeds. The MF-IRL baseline partially outputs a near-uniform θ^e, inconsistent with the observed trajectories, and fails on all remaining metrics. This is particularly striking given that MF-IRL operates on the more informative $\xi = \mu$ observations.

5.4 Experiments 3 & 4: Dynamic Inference and Application to Real Data

To show that our framework also applies to dynamics inference and to the state-action observation setting $\xi = \mu$, we consider the Susceptible–Infected–Susceptible (SIS) epidemic model (Cui & Koeppl, 2021) (App. H.1). To illustrate that the method can be applied to real data, we consider an MFG of electricity producers, whose common noise is the observed electricity demand, and apply the BE-MFG pipeline to public data of the French electricity market (App. I).

6 Conclusion

We introduced a Bayesian framework for learning the latent parameters of MFGs with common noise from partial, population trajectories. The central contribution, the BE-MFG, embeds the entire parametric family into a single augmented game whose Nash equilibrium simultaneously encodes equilibrium policies for all parameters, bypassing repeated forward solves. We formalized identifiability under common noise and established consistency guarantees. On three synthetic environments of increasing complexity, we recover the ground-truth parameter accurately and derive near-optimal policies, consistently outperforming MF-IRL despite the latter having access to more informative observations. A fourth experiment, on public electricity market data, suggests that the pipeline can also be run on real-world observations, although the model we use there is deliberately stylized and we make no claim of predictive value. **Limitations:** We provided the basic ideas around identifiability but one should look for general sufficient conditions. Moreover, we restricted our study to finite state models; scaling the BE-MFG approach to continuous state-action spaces via function approximation is a natural next step.

AI USE STATEMENT

In this work, we used generative AI tools to assist with implementing the code of the experiments, improving the English of the manuscript, analyzing the numerical results, creating the figures, and detecting typographical and formatting issues. The research questions, the model and problem formulation, the theoretical statements and their proofs, the design of the experiments, and the interpretation of the results are our own. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Yves Achdou and Alessio Porretta. Mean field games with congestion. *Annales de l'Institut Henri Poincaré C, Analyse non linéaire*, 35(2):443–480, 2018.

Yves Achdou, Jiequn Han, Jean-Michel Lasry, Pierre-Louis Lions, and Benjamin Moll. Income and wealth distribution in macroeconomics: A continuous-time approach. Technical report, *National Bureau of Economic Research*, August 2017. URL http://www.nber.org/papers/w23732.

Şevket Kaan Alkır and Naci Saldi. Inverse reinforcement learning for mean-field games with average reward criterion. In *2025 IEEE 64th Conference on Decision and Control (CDC)*, pp. 7272–7277. IEEE, 2025.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Q-learning in regularized mean-field games. *Dynamic Games and Applications*, 13(1):89–117, 2023.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Maximum causal entropy inverse reinforcement learning for mean-field games. *arXiv preprint arXiv:2401.06566*, 2024.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Kernel based maximum entropy inverse reinforcement learning for mean-field games. *arXiv preprint arXiv:2507.14529*, 2025a.

Berkay Anahtarci, Can Deha Kariksiz, and Naci Saldi. Maximum causal entropy IRL in mean-field games and GNEP framework for forward rl. *Journal of Machine Learning Research*, 26(121): 1–40, 2025b.

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

Lucian Buşoniu, Robert Babuška, and Bart De Schutter. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

Pierre Cardaliaguet and Charles-Albert Lehalle. Mean field game of controls and an application to trade crowding. *Mathematics and Financial Economics*, 12(3), 2018.

René Carmona, François Delarue, and Daniel Lacker. Mean field games with common noise. *Annals of Probability*, 44(6), 2016.

Yang Chen, Libo Zhang, Jiamou Liu, and Shuyue Hu. Individual-level inverse reinforcement learning for mean field games. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 253–262, 2022.

Yang Chen, Libo Zhang, Jiamou Liu, and Michael Witbrock. Adversarial inverse reinforcement learning for mean field games. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, pp. 1088–1096, 2023.

Yang Chen, Xiao Lin, Bo Yan, Libo Zhang, Jiamou Liu, Neset Özkan Tan, and Michael Witbrock. Meta-inverse reinforcement learning for mean field games via probabilistic context variables. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11407–11415, 2024.

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

Boualem Djehiche, Alain Tcheukam, and Hamidou Tembine. A mean-field game of evacuation in multilevel building. *IEEE Transactions on Automatic Control*, 62(10), 2017.

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

Massimo Fornasier, Benedetto Piccoli, and Francesco Rossi. Mean-field sparse optimal control. *Philosophical Transactions of the Royal Society A*, 372(2014):20130400, 2014.

Sven Gronauer and Klaus Diepold. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, 55(2):895–943, 2022.

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

A. F. Hanif, Hamidou Tembine, and Mohamad Assaad. Mean-field games for resource sharing in cloud-based networks. *IEEE/ACM Transactions on Networking*, 24(1):624–637, 2016.

Johannes Heinrich and David Silver. Deep reinforcement learning from self-play in imperfect-information games. *arXiv preprint arXiv:1603.01121*, 2016.

Ruimeng Hu and Mathieu Laurière. Recent developments in machine learning methods for stochastic control and games. *Numerical Algebra, Control and Optimization*, 14(3):435–525, 2024.

Jiawei Huang, Niao He, and Andreas Krause. Model-based RL for mean-field games is not statistically harder than single-agent RL. In *Forty-first International Conference on Machine Learning*, 2024.

Minyi Huang, Roland P Malhamé, and Peter E Caines. Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the nash certainty equivalence principle. *Commun. Inf. Syst.*, 6(1):221–252, 2006.

Ian T Jolliffe and Jorge Cadima. Principal component analysis: a review and recent developments. *Philosophical transactions of the royal society A: Mathematical, Physical and Engineering Sciences*, 374(2065):20150202, 2016.

Aimé Lachapelle and Marie-Therese Wolfram. On a mean field game approach modeling congestion and aversion in pedestrian crowds. *Transportation Research Part B: Methodological*, 45(10): 1572–1589, 2011.

Aimé Lachapelle, Jean-Michel Lasry, Charles-Albert Lehalle, and Pierre-Louis Lions. Efficiency of the price formation process in presence of high frequency participants: a mean field game analysis. *Mathematics and Financial Economics*, 10(3), 2016.

Grégoire Lambrecht and Mathieu Laurière. Population-aware imitation learning in mean-field games with common noise. *arXiv preprint arXiv:2605.03357*, 2026.

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

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

Weichao Mao, Haoran Qiu, Chen Wang, Hubertus Franke, Zbigniew Kalbarczyk, Ravishankar Iyer, and Tamer Basar. A mean-field game approach to cloud resource management with function approximation. *Advances in Neural Information Processing Systems*, 35:36243–36258, 2022.

Mojtaba Nourian, Peter E Caines, and Roland P Malhamé. Synthesis of Cucker-Smale type flocking via mean field stochastic control theory: Nash equilibria. In *2010 48th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 814–819. IEEE, 2010.

George Papamakarios, Theo Pavlakou, and Iain Murray. Masked autoregressive flow for density estimation. *Advances in neural information processing systems*, 30, 2017.

Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Julien Perolat, Bart De Vylder, Daniel Hennes, Eugene Tarassov, Florian Strub, Vincent de Boer, Paul Muller, Jerome T Connor, Neil Burch, Thomas Anthony, et al. Mastering the game of Stratego with model-free multiagent reinforcement learning. *Science*, 378(6623):990–996, 2022.

Julien Pérolat, Sarah Perrin, Romuald Elie, Mathieu Laurière, Georgios Piliouras, Matthieu Geist, Karl Tuyls, and Olivier Pietquin. Scaling mean field games by online mirror descent. In *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, pp. 1028–1037, 2022.

Sarah Perrin, Julien Perolat, Mathieu Laurière, Matthieu Geist, Romuald Elie, and Olivier Pietquin. Fictitious play for mean field games: continuous time analysis and applications. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 13199–13213, 2020.

Sarah Perrin, Mathieu Laurière, Julien Pérolat, Romuald Élie, Matthieu Geist, and Olivier Pietquin. Generalization in mean field games by learning master policies. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 9413–9421, 2022.

Giorgia Ramponi, Pavel Kolev, Olivier Pietquin, Niao He, Mathieu Laurière, and Matthieu Geist. On imitation in mean-field games. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Naci Saldi, Tamer Basar, and Maxim Raginsky. Markov–Nash equilibria in mean-field games with discounted cost. *SIAM Journal on Control and Optimization*, 56(6):4256–4287, 2018.

Mikayel Samvelyan, Tabish Rashid, Christian Schroeder De Witt, Gregory Farquhar, Nantas Nardelli, Tim G. J. Rudner, Chia-Man Hung, Philip H. S. Torr, Jakob Foerster, and Shimon Whiteson. The starcraft multi-agent challenge. In *Proceedings of the 18th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, pp. 2186–2188, 2019.

Eran Shmaya and Bruno Ziliotto. Bayesian learning in mean field games. *SIAM Journal on Control and Optimization*, 63(3):1783–1797, 2025.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 2016.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of Go without human knowledge. *Nature*, 550(7676), 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, Timothy Lillicrap, Karen Simonyan, and Demis Hassabis. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), 2018.

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

Alexander Wong, Thomas Bäck, Anna V Kononova, and Aske Plaat. Deep multiagent reinforcement learning: Challenges and directions. *Artificial Intelligence Review*, 56:5023–5056, 2023.

Chungang Yang, Jiandong Li, Min Sheng, Alagan Anpalagan, and Jia Xiao. Mean field game-theoretic framework for interference and energy-aware control in 5G ultra-dense networks. *IEEE Wireless Communications*, 25(1):114–121, 2018.

Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 2020.

Batuhan Yardim, Semih Cayci, Matthieu Geist, and Niao He. Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*, pp. 39722–39754. PMLR, 2023.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pp. 321–384, 2021.

A NOTATION AND ABBREVIATIONS

Table 1: Abbreviations.

Abbreviation	**Meaning**
BE-MFG	Bayesian-Extended Mean Field Game (Sec. 4.2)
BE-NE	Bayesian-Extended Nash equilibrium, i.e., a NE of the BE-MFG
BMA	Bayesian model averaging
FiLM	Feature-wise Linear Modulation
FP	Fictitious Play
GP	Gaussian Process
IL / IRL	Imitation Learning / Inverse Reinforcement Learning
MAF	Masked Autoregressive Flow
MAP	maximum a posteriori
MF-IL	Mean Field Imitation Learning
MF-IRL	Mean-Field Inverse Reinforcement Learning
MFG	Mean Field Game
NE	Nash equilibrium
PCA	principal component analysis
PMFG	Parametric Mean Field Game
RL	Reinforcement Learning
SIS	Susceptible–Infected–Susceptible

Table 2: Main notation.

Symbol	Meaning
General	
$[N]$, $[N]^*$	$\{0,\dots,N\}$ and $\{1,\dots,N\}$
$\Delta_{\mathcal{S}}$, $\mathcal{P}(E)$	probability measures on a finite set $\mathcal{S}$, on a measurable space E
$\mathcal{B}(U)$	Borel σ-algebra on $U \subset \mathbb{R}^n$, $n \in \mathbb{N}^*$, induced by the Euclidean topology
$\mathcal{L}(X)$	law of a random variable X
$P \otimes Q$, $P \ll Q$, $\frac{dP}{dQ}$	product measure, absolute continuity, Radon-Nikodym derivative of two probability measures P, Q
$\mathcal{W}_1$, dist	Wasserstein-1 distance on a metric space $(\mathcal{S}, \mathrm{dist})$: $\mathcal{W}_1(P,Q) = \inf_{\nu\in\Pi(P,Q)} \mathbb{E}_{(X,Y)\sim\nu}[\mathrm{dist}(X,Y)]$, where $\Pi(P,Q)$ is the set of couplings with marginals P and Q
Parametric MFGs (Sec. 2)	
$\mathcal{X}$, $\mathcal{A}$	finite state and action spaces
$\mathcal{E}^0$, ν^0, ε^0	common noise space, common noise law, common noise trajectory $(\varepsilon^0_1,\dots,\varepsilon^0_{H-1})$
η	common noise intensity (experiments)
H, ρ_0	time horizon, initial distribution
$\bar{\Theta} \subset \mathbb{R}^d$, λ	parameter set, reference measure on $\bar{\Theta}$
P_θ, r_θ, $\mathcal{M}_\theta$	transition kernel, reward and PMFG associated with θ
$\pi = (\pi_0,\dots,\pi_{H-1})$, Π_0, Π	policy, set of strategies, set of policies
Φ_θ	mean-field transition operator
$\rho_\theta^{\pi,\varepsilon^0}$, $\rho_\theta^{\pi,\pi',\varepsilon^0}$	population flow under π; flow of a single agent playing π' against a population playing π
$\mu_\theta^{\pi,\pi',\varepsilon^0}$	state-action counterpart of $\rho_\theta^{\pi,\pi',\varepsilon^0}$
$V_\theta(\pi';\pi)$, $\mathrm{BR}_\theta(\pi)$	value of π' against π, set of best responses to π
π^*_θ, ρ^*_θ, μ^*_θ	a NE of $\mathcal{M}_\theta$ and its (state, state-action) mean-field trajectories
$\mathcal{E}_\theta(\pi)$	exploitability of π in $\mathcal{M}_\theta$
Observations and identifiability (Sec. 2–3)	
$\xi \in \{\rho,\mu\}$, Ξ	type of observation and its space
$\mathcal{I} \subset [H-1]$, $\xi_{\mathcal{I}}$	observed time steps, observation restricted to $\mathcal{I}$
$\mathbb{P}_{\theta,\mathcal{I}}$	probability measure of $\xi^*_{\theta,\mathcal{I}}$
θ_0	true (unknown) parameter
N, $\hat{\xi}^n_{\theta_0,\mathcal{I}}$	number of observations, n-th observation
$\mathfrak{P}_{\theta,\mathcal{I}}$, $\mathbb{BR}_\theta$	laws generated by some policy, best-response map on laws
Bayesian setting (Sec. 4)	
q, Θ	prior on $\bar{\Theta}$, random parameter with law q
$\mathbb{P}_{\mathcal{I}}$, $\Gamma_{\theta,\mathcal{I}}$	prior predictive measure, likelihood $d\mathbb{P}_{\theta,\mathcal{I}}/d\mathbb{P}_{\mathcal{I}}$
$\bar{q}_N$	posterior after N observations
$\hat{\theta}^{\mathrm{MAP}}_N$, $\pi_{\bar{q}_N}$	MAP estimator, BMA policy
$\bar{\mathcal{M}}_q$, $\bar{P}$, $\bar{r}$	BE-MFG, its transition kernel and reward
$\bar{\Pi}_0$, $\bar{\Pi}$, $\bar{V}_q$	Bayesian strategies, Bayesian policies, value function of $\bar{\mathcal{M}}_q$
$\bar{\pi}^*$, π^{**}_θ	a BE-NE, and the policy it induces in $\mathcal{M}_\theta$ (Thm. 4.2)
$\Gamma^\gamma_{\theta,\mathcal{I}}$, γ	learned approximation of $\Gamma_{\theta,\mathcal{I}}$, trainable parameters

B ADDITIONAL DEFINITIONS, LEMMAS AND PROOFS

B.1 IDENTIFIABILITY AND CONSISTENCY

We first present the proof of Theorem 3.3.

Proof. By assumption $\mathbb{P}_{\theta,\mathcal{I}} = \mathcal{L}(\mu^*_\theta) = \mathcal{L}(\mu^*_{\theta'}) = \mathbb{P}_{\theta',\mathcal{I}}$. Since $\rho^*_{\theta,i}$ is the first marginal of $\mu^*_{\theta,i}$, we deduce that $\mathcal{L}(\rho^*_\theta) = \mathcal{L}(\rho^*_{\theta'})$.

Since the kernel does not depend on the parameter, the mean-field transition operator $\Phi_\theta = \Phi$ of (1) does not depend on it either. Consequently, the flows induced by a given policy are the same in $\mathcal{M}_\theta$ and in $\mathcal{M}_{\theta'}$; in particular

$$\rho_{\theta'}^{\pi^*_\theta,\varepsilon^0} = \rho_\theta^{\pi^*_\theta,\varepsilon^0} = \rho^*_\theta, \qquad \mu_{\theta'}^{\pi^*_\theta,\varepsilon^0} = \mu_\theta^{\pi^*_\theta,\varepsilon^0} = \mu^*_\theta, \qquad \mathbb{P}\text{-a.s.} \tag{6}$$

Step 1: the deviation values coincide. Let $\pi \in \Pi$. We prove by induction on $i \in [H-1]$ that there is a measurable mapping G_i, depending only on π, such that

$$\rho_{\theta',i}^{\pi^*_\theta,\pi,\varepsilon^0} = G_i\big(\rho^*_{\theta,0:i-1}, \varepsilon^0_{1:i}\big), \quad \mathbb{P}\text{-a.s.}$$

For $i=0$, $\rho_{\theta',0}^{\pi^*_\theta,\pi,\varepsilon^0} = \rho_0$ is deterministic. Assume the identity holds at rank $i \in [H-2]$. By (2) and (6),

$$\rho_{\theta',i+1}^{\pi^*_\theta,\pi,\varepsilon^0} = \Phi\big(\rho_{\theta',i}^{\pi^*_\theta,\pi,\varepsilon^0}, \pi_i\,;\, \rho^*_{\theta,i}, \varepsilon^0_{i+1}\big) = \Phi\big(G_i(\rho^*_{\theta,0:i-1}, \varepsilon^0_{1:i}), \pi_i\,;\, \rho^*_{\theta,i}, \varepsilon^0_{i+1}\big),$$

which is of the announced form and completes the induction. The same mappings G_i give $\rho_{\theta',i}^{\pi^*_{\theta'},\pi,\varepsilon^0} = G_i(\rho^*_{\theta',0:i-1}, \varepsilon^0_{1:i})$. Since the state-action flow of the deviating agent is obtained from its state flow and from π, there is therefore a measurable mapping Ψ, depending only on θ' and π, such that

$$V_{\theta'}(\pi;\pi^*_\theta) = \mathbb{E}\big[\Psi(\rho^*_\theta, \varepsilon^0)\big], \qquad V_{\theta'}(\pi;\pi^*_{\theta'}) = \mathbb{E}\big[\Psi(\rho^*_{\theta'}, \varepsilon^0)\big].$$

By recoverability of the common noise, $\varepsilon^0 = \Lambda(\mu^*_\theta)$ $\mathbb{P}$-a.s., so that $\Psi(\rho^*_\theta, \varepsilon^0) = \tilde{\Psi}(\mu^*_\theta)$ $\mathbb{P}$-a.s., where $\tilde{\Psi}$ maps ν to Ψ evaluated at the first marginal of ν and at $\Lambda(\nu)$; the same identity holds for θ' with the same $\tilde{\Psi}$. As $\mathcal{L}(\mu^*_\theta) = \mathcal{L}(\mu^*_{\theta'})$, we conclude that

$$V_{\theta'}(\pi;\pi^*_\theta) = V_{\theta'}(\pi;\pi^*_{\theta'}), \qquad \forall \pi \in \Pi. \tag{7}$$

Step 2: the equilibrium values coincide. By (6), $V_{\theta'}(\pi^*_\theta;\pi^*_\theta) = \mathbb{E}\big[\sum_{i,x,a} \mu^*_{\theta,i}(x,a)\, r_{\theta'}(x,a,\rho^*_{\theta,i})\big]$, which is the image of μ^*_θ by a measurable mapping depending only on θ'. The same mapping applied to $\mu^*_{\theta'}$ gives $V_{\theta'}(\pi^*_{\theta'};\pi^*_{\theta'})$, so that $\mathcal{L}(\mu^*_\theta) = \mathcal{L}(\mu^*_{\theta'})$ yields

$$V_{\theta'}(\pi^*_\theta;\pi^*_\theta) = V_{\theta'}(\pi^*_{\theta'};\pi^*_{\theta'}). \tag{8}$$

Conclusion. Taking the supremum over $\pi \in \Pi$ in (7), using that $\pi^*_{\theta'}$ is a NE of $\mathcal{M}_{\theta'}$, and then (8),

$$\sup_{\pi\in\Pi} V_{\theta'}(\pi;\pi^*_\theta) = \sup_{\pi\in\Pi} V_{\theta'}(\pi;\pi^*_{\theta'}) = V_{\theta'}(\pi^*_{\theta'};\pi^*_{\theta'}) = V_{\theta'}(\pi^*_\theta;\pi^*_\theta),$$

i.e. $\pi^*_\theta \in \mathtt{BR}_{\theta'}(\pi^*_\theta)$, which means that π^*_θ is a NE for $\mathcal{M}_{\theta'}$. □

We now pass to the proof of Theorem 3.6.

Proof. Suppose that $\{\mathcal{M}_\theta\}_{\theta\in\bar{\Theta}}$ is not identifiable. Therefore, there exists two distinct parameters $\theta, \theta' \in \bar{\Theta}$, such that $\mathbb{P}_{\theta,\mathcal{I}} = \mathbb{P}_{\theta',\mathcal{I}}$. In the two situations, $\xi = \rho$ and $\xi = \mu$, this leads to $\mathcal{L}(\rho^*_{\theta,\mathcal{I}}) = \mathcal{L}(\rho^*_{\theta',\mathcal{I}})$. By definition $\mathcal{L}(\rho^*_{\theta,\mathcal{I}}) \in \mathbb{BR}_\theta(\mathcal{L}(\rho^*_{\theta,\mathcal{I}}))$ and $\mathcal{L}(\rho^*_{\theta,\mathcal{I}}) \in \mathbb{BR}_{\theta'}(\mathcal{L}(\rho^*_{\theta,\mathcal{I}}))$. Therefore $\mathbb{BR}_\theta(\mathbb{P}_\mathcal{I}) \cap \mathbb{BR}_{\theta'}(\mathbb{P}_\mathcal{I}) \neq \emptyset$. That concludes the proof. □

We now give an example of identifiable family of PMFGs, satisfying the condition of Corollary 3.7.

Example B.1. *Let* $\mathcal{X} = \{0, \dots, 6\}$, $\mathcal{A} = \{-1, 0, 1\}$, $\bar{\Theta} = \{0, 1\}$, $\mathcal{E}^0 = \{-2, -1, 0, 1, 2\}$, $\rho_0 = \delta_3$ *and* $H > 1$. *Let for any* $(x, x', a, \rho, e^0) \in \mathcal{X} \times \mathcal{X} \times \mathcal{A} \times \Delta_{\mathcal{X}} \times \mathcal{E}^0$, $P_0(x'|x, a, \rho, e^0) = \delta_{\{x'=x+a+e^0 \ mod \ 7\}}$ *and* $P_1(x'|x, a, \rho, e^0) = \delta_{\{x'=x+a \ mod \ 7\}}$. *Suppose that* $\nu^0(0)+\nu^0(1)+\nu^0(-1) \neq 1$. *The condition in Corollary 3.7 is satisfied for any* $\mathcal{I} \subset [H-1]$ *such that* $1 \in \mathcal{I}$.

Finally, we prove the statement in Example B.1.

Proof. For any $\pi \in \Pi$, and $\mathbb{P}(\rho^{\pi}_{\theta=0,1}(2) + \rho^{\pi}_{\theta=0,1}(3) + \rho^{\pi}_{\theta=0,1}(4) \neq 1) > 0$ because of the presence of the common noise. Yet, $\mathbb{P}(\rho^{\pi}_{\theta=1,1}(2) + \rho^{\pi}_{\theta=1,1}(3) + \rho^{\pi}_{\theta=1,1}(4) \neq 1) = 0$. That is enough to conclude that Cor.3.7 is verified. □

B.2 Existence of the likelihood

We prove Lemma 4.1.

Proof. Let $A \in \mathcal{B}(\Xi^{|\mathcal{I}|})$ such that $\mathbb{P}_{\mathcal{I}}(A) = 0$, i.e., $\int_{\bar{\Theta}} \mathbb{P}_{\theta,\mathcal{I}}(A) q(\theta) \lambda(d\theta) = 0$. Since $q(\theta) > 0$ for every $\theta \in \bar{\Theta}$, we get $\mathbb{P}_{\theta,\mathcal{I}}(A) = 0$ for λ-almost every $\theta \in \bar{\Theta}$.

Under (C1), λ is the counting measure on the countable set $\bar{\Theta}$, so the integral is the sum $\sum_{\theta \in \bar{\Theta}} \mathbb{P}_{\theta,\mathcal{I}}(A) q(\theta)$, whose terms are non-negative. Hence $\mathbb{P}_{\theta,\mathcal{I}}(A) = 0$ for every $\theta \in \bar{\Theta}$.

Under (C2), the set $U_A = \{\theta \in \bar{\Theta} : \mathbb{P}_{\theta,\mathcal{I}}(A) > 0\}$ is open in $\bar{\Theta}$, since $\theta \mapsto \mathbb{P}_{\theta,\mathcal{I}}(A)$ is continuous, and $\lambda(U_A) = 0$. Since λ charges every non-empty open subset of $\bar{\Theta}$, $U_A = \emptyset$, i.e., $\mathbb{P}_{\theta,\mathcal{I}}(A) = 0$ for every $\theta \in \bar{\Theta}$.

In both cases, $\mathbb{P}_{\mathcal{I}}(A) = 0$ implies $\mathbb{P}_{\theta,\mathcal{I}}(A) = 0$ for every $\theta \in \bar{\Theta}$, i.e., $\mathbb{P}_{\theta,\mathcal{I}} \ll \mathbb{P}_{\mathcal{I}}$. Since these are finite measures, the Radon-Nikodym theorem gives the existence of $\Gamma_{\theta,\mathcal{I}} = d\mathbb{P}_{\theta,\mathcal{I}}/d\mathbb{P}_{\mathcal{I}}$. □

B.3 BE-MFG: full definitions

We provide the full definitions for the BE-MFG $\bar{\mathcal{M}}_q$, following the same structure as Sec. 2 with augmented state space $\bar{\Theta} \times \mathcal{X}$.

Distribution evolution. For any pair of population distributions $\bar{\rho}, \bar{\rho}' \in \mathcal{P}(\bar{\Theta} \times \mathcal{X})$, any strategy $\bar{\pi}_0 \in \bar{\Pi}_0$, we define the mean-field transition operator $\bar{\Phi}(\bar{\rho}', \bar{\pi}_0; \bar{\rho}, e^0)$ as

$$\bar{\Phi}(\bar{\rho}', \bar{\pi}_0; \bar{\rho}, e^0)(\theta', x') = \int_{\theta \in \bar{\Theta}} \sum_{x \in \mathcal{X}, a \in \mathcal{A}} \bar{\rho}'(d\theta, x) \bar{\pi}_0(a|\theta, x, \bar{\rho}) \bar{P}(\theta', x'|\theta, x, a, \bar{\rho}, e^0).$$

Let Θ be a $\bar{\Theta}$-valued extra source of randomness for all the paper, with density q. The evolution of the population distribution when all agents follow policy $\bar{\pi}$ and are affected by the common noise trajectory $\varepsilon^0_{1:H-1}$ is defined by $\bar{\rho}^{\bar{\pi},\varepsilon^0}(\Theta) = (\bar{\rho}_0^{\bar{\pi},\varepsilon^0}(\Theta), \bar{\rho}_1^{\bar{\pi},\varepsilon^0}(\Theta), \dots, \bar{\rho}_{H-1}^{\bar{\pi},\varepsilon^0}(\Theta))$ where $\bar{\rho}_0^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_0$ and for $i \in [H-2]$,

$$\bar{\rho}_{i+1}^{\bar{\pi},\varepsilon^0}(\Theta) = \bar{\Phi}(\bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta), \bar{\pi}_i\,; \bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta), \varepsilon^0_{i+1}). \tag{9}$$

The evolution of the state distribution of a *single agent* following policy $\bar{\pi}'$ when all other agents follow policy $\bar{\pi}$ is defined by $\bar{\rho}^{\bar{\pi},\bar{\pi}',\varepsilon^0}(\Theta) = (\bar{\rho}_0^{\bar{\pi},\bar{\pi}',\varepsilon^0}(\Theta), \bar{\rho}_1^{\bar{\pi},\bar{\pi}',\varepsilon^0}(\Theta), \dots, \bar{\rho}_{H-1}^{\bar{\pi},\bar{\pi}',\varepsilon^0}(\Theta))$ where $\bar{\rho}_0^{\bar{\pi},\bar{\pi}',\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_0$ and for $i \in [H-2]$,

$$\bar{\rho}_{i+1}^{\bar{\pi},\bar{\pi}',\varepsilon^0}(\Theta) = \bar{\Phi}(\bar{\rho}_i^{\bar{\pi},\bar{\pi}',\varepsilon^0}(\Theta), \bar{\pi}'_i; \bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta), \varepsilon^0_{i+1}). \tag{10}$$

Value Function of BE-MFG. When the population of agents follows the policy $\bar{\pi}$, and a single agent follows $\bar{\pi}'$, its associated **Bayesian-Extended value function** is:

$$\bar{V}_q(\bar{\pi}'; \bar{\pi}) = \mathbb{E}\left[\sum_{i=0}^{H-1} \bar{r}(\Theta_i, X_i, \alpha_i, \bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta))\right], \tag{11}$$

where the process is defined by $\Theta_0 = \Theta \sim q$, $X_0 \sim \rho_0$, $(\Theta_{i+1}, X_{i+1}) \sim \bar{P}(\cdot|\Theta_i, X_i, \alpha_i, \bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta), \varepsilon^0_{i+1})$, and $\alpha_i \sim \bar{\pi}'(\cdot|\Theta_i, X_i, \bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta))$. Note that under these conditions, the law of the state (Θ_i, X_i) (conditioned on the common noise) is $\bar{\rho}_i^{\bar{\pi},\bar{\pi}',\varepsilon^0}(\Theta)$. We define the Bayesian best responses and the NE in the exact same way that for the original PMFGs.

Definition B.2. *For $\pi \in \bar{\Pi}$, the set of* **Bayesian best responses** *to the PMFG $\bar{\mathcal{M}}_q$ is defined by $\bar{\mathrm{BR}}_q(\bar{\pi}) = \arg\max_{\bar{\pi}' \in \bar{\Pi}} \bar{V}_q(\bar{\pi}'; \bar{\pi})$. A policy $\bar{\pi} \in \bar{\Pi}$ is a* **(mean field) Bayesian-Extended Nash equilibrium** *(BE-NE) for $\bar{\mathcal{M}}_q$ if $\bar{\pi} \in \bar{\mathrm{BR}}_q(\bar{\pi})$.*

B.4 BE-MFG AND PMFG: LINK BETWEEN THE MODELS

This appendix collects the technical lemmas used in the proofs of the main results concerning the BE-MFG framework. The first two lemmas establish that the population flow in the BE-MFG reduces, type by type, to the population flow in the corresponding PMFG. Specifically, Lemma B.3 shows this reduction when the BE-MFG policy is constructed from a single MFG policy via a point-mass embedding of the parameter, while Lemma B.4 establishes the same conclusion when the BE-MFG policy is constructed from a family of PMFG policies. The next two lemmas, Lemmas B.5 and B.6, extend these reduction results to the setting where a deviating agent follows a policy different from the population. These four lemmas together ensure that the BE-MFG and the family of PMFGs are consistent in terms of population dynamics, regardless of whether one agent deviates. Building on these results, Lemmas B.7 and B.8 show that the value function in the BE-MFG decomposes as the expectation over parameters of the corresponding value functions in the PMFGs, again under the two respective policy constructions. This decomposition is the key identity linking Nash equilibria of the BE-MFG to Nash equilibria of the parametric MFGs, and it is what the subsequent proof of Theorem 4.2 exploits directly.

Lemma B.3. *Let $\pi \in \Pi$ and $\bar{\pi} \in \bar{\Pi}$ be such that for any $\theta \in \bar{\Theta}$ and $(i, x, a, \rho) \in [H-1] \times \mathcal{X} \times \mathcal{A} \times \Delta_{\mathcal{X}}$,*

$$\pi_{\theta,i}(a \mid x, \rho) = \bar{\pi}_i(a \mid \theta, x, \delta_\theta \otimes \rho).$$

Then

$$\bar{\rho}^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_\Theta^{\pi_\Theta,\varepsilon^0}, \quad \mathbb{P}\text{-a.s.},$$

where $\bar{\rho}^{\bar{\pi},\varepsilon^0}(\Theta)$ is defined as in (9) and $\rho_\Theta^{\pi_\Theta,\varepsilon^0}$ is defined as in (1).

Proof. Let $(\Theta_i, X_i, \bar{\alpha}_i)_{0 \le i \le H-1}$ be the state-action trajectory of a representative agent following policy $\bar{\pi} \in \bar{\Pi}$ when the population follows $\bar{\pi}$, in the BE-MFG. This trajectory is defined as in Equation (11), and the marginal law of (Θ_i, X_i) at each time step $i \in [H-1]$ is $\bar{\rho}_i^{\bar{\pi},\varepsilon^0}$. Let $(X'_i, \alpha'_i)_{0 \le i \le H-1}$ be the state-action trajectory of a representative agent following policy $\pi \in \Pi$ when the population follows π, in the PMFG $\mathcal{M}_\Theta$, where π is defined as in the lemma statement. This trajectory is defined as in Equation (3), and the marginal law of X'_i at each time step $i \in [H-1]$ is $\rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}$.

We prove by induction on $i \in [H-1]$ that, conditionally on ε^0,

$$\bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}, \quad \mathbb{P}\text{-a.s.}$$

Base case. By definition, $\bar{\rho}_0^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_0$ and $\rho_{\Theta,0}^{\pi_\Theta,\varepsilon^0} = \rho_0$, so the relation holds at $i = 0$.

Inductive step. Suppose that $\bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}$ holds $\mathbb{P}$-a.s. for some $i \in [H-2]$. By definition of π,

$$\alpha'_i \sim \pi_{\Theta,i}(\cdot \mid X'_i, \rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}) = \bar{\pi}_i(\cdot \mid \Theta, X'_i, \delta_\Theta \otimes \rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}).$$

By the induction hypothesis, $\bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}$, so

$$\bar{\alpha}_i \sim \bar{\pi}_i(\cdot \mid \Theta_i, X_i, \bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta)) = \bar{\pi}_i(\cdot \mid \Theta, X_i, \delta_\Theta \otimes \rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}).$$

Hence $\mathcal{L}(X_i, \bar{\alpha}_i \mid \varepsilon^0) = \mathcal{L}(X'_i, \alpha'_i \mid \varepsilon^0)$. Now, by Equation (5),

$$(\Theta_{i+1}, X_{i+1}) \sim \bar{P}(\cdot \mid \Theta_i, X_i, \bar{\alpha}_i, \bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta)) = \delta_{\Theta_i} \otimes P_\Theta(\cdot \mid X_i, \bar{\alpha}_i, \rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}),$$

and $X'_{i+1} \sim P_\Theta(\cdot \mid X'_i, \alpha'_i, \rho^{\pi_\Theta,\varepsilon^0}_{\Theta,i})$. Since $\Theta_i = \Theta$ $\mathbb{P}$-a.s. by the induction hypothesis, we conclude that

$$\bar{\rho}^{\bar{\pi},\varepsilon^0}_{i+1}(\Theta) = \delta_\Theta \otimes \rho^{\pi_\Theta,\varepsilon^0}_{\Theta,i+1}, \quad \mathbb{P}\text{-a.s.},$$

which completes the induction and the proof. □

Lemma B.4. *Let $\{\pi_\theta\}_{\theta\in\bar{\Theta}} \subset \Pi$ be a family of policies, and define $\bar{\pi} \in \bar{\Pi}$ by*

$$\bar{\pi}_i(a \mid \theta, x, \bar{\rho}) = \pi_{\theta,i}\left(a \,\middle|\, x, \int_{\bar{\Theta}} \bar{\rho}(d\theta', \cdot)\right), \quad (i, \theta, x, a, \bar{\rho}) \in [H-1] \times \bar{\Theta} \times \mathcal{X} \times \mathcal{A} \times \Delta_{\bar{\Theta}\times\mathcal{X}}.$$

Then

$$\bar{\rho}^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho^{\pi_\Theta,\varepsilon^0}_\Theta, \quad \mathbb{P}\textit{-a.s.},$$

where $\bar{\rho}^{\bar{\pi},\varepsilon^0}$ is defined as in (9) and $\rho^{\pi_\Theta,\varepsilon^0}_\Theta$ is defined as in (1).

Proof. We prove by induction on $i \in [H-1]$ that, conditionally on ε^0,

$$\bar{\rho}^{\bar{\pi},\varepsilon^0}_i(\Theta) = \delta_\Theta \otimes \rho^{\pi_\Theta,\varepsilon^0}_{\Theta,i}, \quad \mathbb{P}\text{-a.s.}$$

Base case. By definition, $\bar{\rho}^{\bar{\pi},\varepsilon^0}_0(\Theta) = \delta_\Theta \otimes \rho_0$ and $\rho^{\pi_\Theta,\varepsilon^0}_{\Theta,0} = \rho_0$, so the relation holds at $i = 0$.

Inductive step. Suppose that $\bar{\rho}^{\bar{\pi},\varepsilon^0}_i(\Theta) = \delta_\Theta \otimes \rho^{\pi_\Theta,\varepsilon^0}_{\Theta,i}$ holds $\mathbb{P}$-a.s. for some $i \in [H-2]$. By the induction hypothesis,

$$\int_{\bar{\Theta}} \bar{\rho}^{\bar{\pi},\varepsilon^0}_i(\Theta)(d\theta', \cdot) = \int_{\bar{\Theta}} \delta_\Theta(d\theta')\, \rho^{\pi_\Theta,\varepsilon^0}_{\Theta,i}(\cdot) = \rho^{\pi_\Theta,\varepsilon^0}_{\Theta,i}.$$

Therefore, for any agent with type Θ and state x, the action distribution under $\bar{\pi}$ satisfies

$$\bar{\pi}_i(\cdot \mid \Theta, x, \bar{\rho}^{\bar{\pi},\varepsilon^0}_i(\Theta)) = \pi_{\Theta,i}\left(\cdot \,\middle|\, x, \int_{\bar{\Theta}} \bar{\rho}^{\bar{\pi},\varepsilon^0}_i(\Theta)(d\theta', \cdot)\right) = \pi_{\Theta,i}(\cdot \mid x, \rho^{\pi_\Theta,\varepsilon^0}_{\Theta,i}).$$

Now, by Equation (5) and the induction hypothesis,

$$(\Theta_{i+1}, X_{i+1}) \sim \bar{P}(\cdot \mid \Theta_i, X_i, \bar{\alpha}_i, \bar{\rho}^{\bar{\pi},\varepsilon^0}_i(\Theta)) = \delta_{\Theta_i} \otimes P_\Theta(\cdot \mid X_i, \bar{\alpha}_i, \rho^{\pi_\Theta,\varepsilon^0}_{\Theta,i}),$$

where $\Theta_i = \Theta$ $\mathbb{P}$-a.s. by the induction hypothesis. Since the action and transition distributions coincide with those of the PMFG $\mathcal{M}_\Theta$ under policy π_Θ, we conclude that

$$\bar{\rho}^{\bar{\pi},\varepsilon^0}_{i+1}(\Theta) = \delta_\Theta \otimes \rho^{\pi_\Theta,\varepsilon^0}_{\Theta,i+1}, \quad \mathbb{P}\text{-a.s.},$$

which completes the induction and the proof. □

Lemma B.5. *Let $\pi, \pi' \in \Pi$ and $\bar{\pi}, \bar{\pi}' \in \bar{\Pi}$ be such that for any $\theta \in \bar{\Theta}$ and $(i, x, a, \rho) \in [H-1] \times \mathcal{X} \times \mathcal{A} \times \Delta_\mathcal{X}$,*

$$\pi_{\theta,i}(a \mid x, \rho) = \bar{\pi}_i(a \mid \theta, x, \delta_\theta \otimes \rho) \quad \textit{and} \quad \pi'_{\theta,i}(a \mid x, \rho) = \bar{\pi}'_i(a \mid \theta, x, \delta_\theta \otimes \rho).$$

Then

$$\bar{\rho}^{\bar{\pi},\bar{\pi}',\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho^{\pi_\Theta,\pi'_\Theta,\varepsilon^0}_\Theta, \quad \mathbb{P}\textit{-a.s.},$$

where $\bar{\rho}^{\bar{\pi},\bar{\pi}',\varepsilon^0}(\Theta)$ is defined as in (10) and $\rho^{\pi_\Theta,\pi'_\Theta,\varepsilon^0}_\Theta$ is defined as in (2).

Proof. The proof follows the same induction argument as in the proof of Lemma B.3. Lemma B.3 is required to identify the population flow $\bar{\rho}^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho^{\pi_\Theta,\varepsilon^0}_\Theta$ at each step of the induction. □

Lemma B.6. *Let $\{\pi_\theta\}_{\theta\in\bar{\Theta}}, \{\pi'_\theta\}_{\theta\in\bar{\Theta}} \subset \Pi$ be two families of policies, and define $\bar{\pi}, \bar{\pi}' \in \bar{\Pi}$ by*

$$\bar{\pi}_i(a \mid \theta, x, \bar{\rho}) = \pi_{\theta,i}\left(a \,\middle|\, x, \int_{\bar{\Theta}} \bar{\rho}(d\theta', \cdot)\right), \quad \bar{\pi}'_i(a \mid \theta, x, \bar{\rho}) = \pi'_{\theta,i}\left(a \,\middle|\, x, \int_{\bar{\Theta}} \bar{\rho}(d\theta', \cdot)\right).$$

Then

$$\bar{\rho}^{\bar{\pi},\bar{\pi}',\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho^{\pi_\Theta,\pi'_\Theta,\varepsilon^0}_\Theta, \quad \mathbb{P}\textit{-a.s.},$$

where $\bar{\rho}^{\bar{\pi},\bar{\pi}',\varepsilon^0}$ is defined as in (10) and $\rho^{\pi_\Theta,\pi'_\Theta,\varepsilon^0}_\Theta$ is defined as in (2).

Proof. The proof follows the same induction argument as in Lemma B.4, which is also required to identify the population flow $\bar{\rho}^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_\Theta^{\pi_\Theta,\varepsilon^0}$ at each step of the induction. □

Lemma B.7. *Let $\bar{\pi} \in \bar{\Pi}$ and define, for any $\theta \in \bar{\Theta}$ and $(i, x, a, \rho) \in [H-1] \times \mathcal{X} \times \mathcal{A} \times \Delta_\mathcal{X}$,*

$$\pi_{\theta,i}(a \mid x, \rho) = \bar{\pi}_i(a \mid \theta, x, \delta_\theta \otimes \rho).$$

Let $\bar{\pi}' \in \bar{\Pi}$ and define π'_θ similarly. Then

$$\bar{V}_q(\bar{\pi}'; \bar{\pi}) = \mathbb{E}_{\Theta\sim q}[V_\Theta(\pi'_\Theta; \pi_\Theta)]\,.$$

Proof. By definition of $\bar{V}_q(\bar{\pi}'; \bar{\pi})$ in (11), we have

$$\bar{V}_q(\bar{\pi}'; \bar{\pi}) = \mathbb{E}\left[\sum_{i=0}^{H-1} \bar{r}\Big(\Theta_i, X_i, \bar{\alpha}_i, \bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta)\Big)\right],$$

where $(\Theta_i, X_i, \alpha_i)$ is the trajectory of a representative agent acting under $\bar{\pi}'$ while the population follows $\bar{\pi}$. By Lemma B.3, $\bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}$ $\mathbb{P}$-a.s. In the proof of Lemma B.5 it is shown that the joint law of (X_i, α_i) conditioned on Θ is equal to $\mu_{\Theta,i}^{\pi,\pi',\varepsilon^0}$. Using the definition of $\bar{r}$ in (5) and the fact that $\bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}$, we obtain

$$\bar{r}\Big(\Theta_i, X_i, \bar{\alpha}_i, \bar{\rho}_i^{\bar{\pi},\varepsilon^0}(\Theta)\Big) = r_\Theta\Big(X_i, \bar{\alpha}_i, \rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}\Big), \quad \mathbb{P}\text{-a.s.}$$

Taking expectations and conditioning on $\Theta \sim q$ yields

$$\bar{V}(\bar{\pi}'; \bar{\pi}) = \mathbb{E}_{\Theta\sim q}\left[\mathbb{E}\left[\sum_{i=0}^{H-1} r_\Theta\Big(X'_i, \alpha'_i, \rho_{\Theta,i}^{\pi_\Theta,\varepsilon^0}\Big) \,\middle|\, \Theta\right]\right] = \mathbb{E}_{\Theta\sim q}[V_\Theta(\pi'_\Theta; \pi_\Theta)]\,,$$

which completes the proof. □

Lemma B.8. *Let $\{\pi_\theta\}_{\theta\in\bar{\Theta}} \subset \Pi$ be a family of policies and define $\bar{\pi} \in \bar{\Pi}$ by*

$$\bar{\pi}_i(a \mid \theta, x, \bar{\rho}) = \pi_{\theta,i}\left(a \,\middle|\, x, \int_{\bar{\Theta}} \bar{\rho}(d\theta', \cdot)\right).$$

Let $\{\pi'_\theta\}_{\theta\in\bar{\Theta}} \subset \Pi$ and define $\bar{\pi}'$ similarly. Then

$$\bar{V}(\bar{\pi}'; \bar{\pi}) = \mathbb{E}_{\Theta\sim q}[V_\Theta(\pi'_\Theta; \pi_\Theta)]\,.$$

Proof. By construction of $\bar{\pi}$ and $\bar{\pi}'$, we have for any $\theta \in \bar{\Theta}$ and $(i, x, a, \rho) \in [H-1] \times \mathcal{X} \times \mathcal{A} \times \Delta_\mathcal{X}$,

$$\bar{\pi}_i(a \mid \theta, x, \delta_\theta \otimes \rho) = \pi_{\theta,i}(a \mid x, \rho), \quad \bar{\pi}'_i(a \mid \theta, x, \delta_\theta \otimes \rho) = \pi'_{\theta,i}(a \mid x, \rho).$$

Hence $\bar{\pi}$ and $\bar{\pi}'$ satisfy the assumptions of Lemma B.7, which directly gives

$$\bar{V}(\bar{\pi}'; \bar{\pi}) = \mathbb{E}_{\Theta\sim q}[V_\Theta(\pi'_\Theta; \pi_\Theta)]\,. \quad \square$$

The proof of Theorem 4.2 requires selecting, for each parameter, a best response in $\mathcal{M}_\theta$, in such a way that the resulting family is again a Bayesian policy. The following lemma shows that such a measurable selection always exists, and that no measurable selection theorem is needed: since $\mathcal{A}$ is finite, the selection can be written explicitly.

Lemma B.9. *For every $\bar{\pi} \in \bar{\Pi}$, there exists $\bar{\pi}^{\mathrm{br}} \in \bar{\Pi}$ such that, for every $\theta \in \bar{\Theta}$, $\pi_\theta^{\mathrm{br}} \in \mathtt{BR}_\theta(\pi_\theta)$, where π_θ and π_θ^{br} denote the policies induced by $\bar{\pi}$ and $\bar{\pi}^{\mathrm{br}}$ as in Theorem 4.2.*

Proof. Let $\bar{\pi} \in \bar{\Pi}$ and $\theta \in \bar{\Theta}$. An agent deviating from the population policy π_θ faces a finite-horizon MDP whose state is the pair (X_i, ρ_i), where ρ_i is the population mean field. Given θ and π_θ, ρ_{i+1} is a function of $(\rho_i, \varepsilon_{i+1}^0)$ only, so (X_i, ρ_i) is a Markov chain and the strategies of Π_0, which are mappings from $\mathcal{X} \times \Delta_\mathcal{X}$ to $\Delta_\mathcal{A}$, are sufficient: the dynamic programming principle applies over the policy class Π used throughout. Define on $\bar{\Theta} \times \mathcal{X} \times \Delta_\mathcal{X} \times \mathcal{A}$

$$Q_{H-1}(\theta, x, \rho, a) = r_\theta(x, a, \rho), \qquad V_i(\theta, x, \rho) = \max_{a\in\mathcal{A}} Q_i(\theta, x, \rho, a),$$

$$Q_i(\theta, x, \rho, a) = r_\theta(x, a, \rho) + \mathbb{E}_{\varepsilon^0 \sim \nu^0}\Big[\sum_{x' \in \mathcal{X}} P_\theta(x' \mid x, a, \rho, \varepsilon^0)\, V_{i+1}\big(\theta, x', \Phi_\theta(\rho, \pi_{\theta,i}; \rho, \varepsilon^0)\big)\Big],$$

for $i \in [H-2]$.

Each Q_i is Borel measurable. We proceed by downward induction on i. For $i = H-1$ this is the measurability of r. Assume V_{i+1} is Borel. The mapping $(\theta, \rho, \varepsilon^0) \mapsto P_\theta(x' \mid x, a, \rho, \varepsilon^0)$ is Borel by measurability of P, and $(\theta, \rho, \varepsilon^0) \mapsto \Phi_\theta(\rho, \pi_{\theta,i}; \rho, \varepsilon^0)$ is Borel as a finite sum of products of measurable mappings, using that $\bar{\pi} \in \bar{\Pi}$ is measurable. Composing with V_{i+1} and summing over the finite set $\mathcal{X}$ preserves measurability. Since r is bounded, an immediate induction gives $|V_j| \le (H-j)\|r\|_\infty$, so the integrand is bounded and integrating against ν^0 preserves Borel measurability in (θ, x, ρ) by the Fubini–Tonelli theorem. Hence Q_i is Borel, and so is V_i, being a maximum of finitely many Borel mappings.

Selection. Write $\mathcal{Z} = \bar{\Theta} \times \mathcal{X} \times \Delta_{\mathcal{X}}$ and $z = (\theta, x, \rho) \in \mathcal{Z}$. Fix a total order $<$ on the finite set $\mathcal{A}$, chosen once and independently of z, and let

$$a_i^*(z) = \min\Big\{a \in \mathcal{A} \mid Q_i(z, a) = \max_{a' \in \mathcal{A}} Q_i(z, a')\Big\}, \qquad z \in \mathcal{Z}.$$

For each $a \in \mathcal{A}$, we have

$$\big\{z \in \mathcal{Z} \mid a_i^*(z) = a\big\} = \Big(\bigcap_{a' \in \mathcal{A}} \big\{z \in \mathcal{Z} \mid Q_i(z, a) \ge Q_i(z, a')\big\}\Big) \cap \Big(\bigcap_{a'' \in \mathcal{A},\, a'' < a} \big\{z \in \mathcal{Z} \mid Q_i(z, a) > Q_i(z, a'')\big\}\Big),$$

the first intersection expressing that a is a maximizer of $Q_i(z, \cdot)$, and the second that no action smaller than a is one. Both are finite intersections of Borel subsets of $\mathcal{Z}$, so $\{z \in \mathcal{Z} \mid a_i^*(z) = a\}$ is Borel and a_i^* is Borel. Therefore $\bar{\pi}_i^{\mathrm{br}}(\cdot \mid \theta, x, \rho) = \delta_{a_i^*(\theta, x, \rho)}$ defines a Bayesian strategy and $\bar{\pi}^{\mathrm{br}} \in \bar{\Pi}$. By the dynamic programming principle, for every $\theta \in \bar{\Theta}$ the induced policy π_θ^{br} maximizes $V_\theta(\cdot\,; \pi_\theta)$ over Π, i.e. $\pi_\theta^{\mathrm{br}} \in \mathtt{BR}_\theta(\pi_\theta)$. □

Remark B.10. *The finiteness of $\mathcal{A}$ is what makes the selection explicit, so that no measurable selection theorem is needed. Continuity of P and r in θ would be sufficient but is not required: only their measurability, together with the boundedness of r, is used. Under condition (C1) of Assumption 2, $\bar{\Theta}$ is at most countable and the statement is immediate.*

We now provide the proof of Theorem 4.2.

Proof. Let $\bar{\pi}^{\mathrm{br}} \in \bar{\Pi}$ be given by Lemma B.9 applied to $\bar{\pi}^*$, and define

$$\delta(\theta) = V_\theta\big(\pi_\theta^{\mathrm{br}}; \pi_\theta^{**}\big) - V_\theta\big(\pi_\theta^{**}; \pi_\theta^{**}\big), \qquad \theta \in \bar{\Theta}.$$

The mapping δ is measurable, being a difference of values associated with two policies of $\bar{\Pi}$, and $\delta \ge 0$ since $\pi_\theta^{\mathrm{br}} \in \mathtt{BR}_\theta(\pi_\theta^{**})$ for every $\theta \in \bar{\Theta}$. Applying Lemma B.7 twice,

$$\mathbb{E}_{\Theta \sim q}\big[\delta(\Theta)\big] = \bar{V}_q\big(\bar{\pi}^{\mathrm{br}}; \bar{\pi}^*\big) - \bar{V}_q\big(\bar{\pi}^*; \bar{\pi}^*\big) \le 0,$$

where the inequality holds because $\bar{\pi}^*$ is a NE of $\bar{\mathcal{M}}_q$. A non-negative measurable mapping with non-positive integral vanishes almost everywhere, so $\delta = 0$ q-a.s. Since $\delta(\theta) = 0$ means exactly that $\pi_\theta^{**} \in \mathtt{BR}_\theta(\pi_\theta^{**})$, we conclude that π_Θ^{**} is a NE of $\mathcal{M}_\Theta$, q-a.s. □

Dimension of the BE-MFG. At first sight, the BE-MFG reintroduces a curse of dimensionality: its state space is $\bar{\Theta} \times \mathcal{X}$, and Bayesian policies depend on mean fields in $\mathcal{P}(\bar{\Theta} \times \mathcal{X})$. However, the mean fields generated by the BE-MFG have a specific structure: for any Bayesian policy $\bar{\pi}$, $\bar{\rho}_i^{\bar{\pi}, \varepsilon^0}(\Theta) = \delta_\Theta \otimes \rho_{\Theta,i}^{\pi_\Theta, \varepsilon^0}$ for every $i \in [H-1]$ (Lemmas B.3–B.4), so that the population never leaves the set $\{\delta_\theta \otimes \rho : \theta \in \bar{\Theta}, \rho \in \Delta_{\mathcal{X}}\}$. Consequently, as in Theorem 4.2, Bayesian strategies only need to be evaluated at mean fields of the form $\delta_\theta \otimes \rho$, i.e., it suffices to consider strategies defined on $\bar{\Theta} \times \mathcal{X} \times \Delta_{\mathcal{X}}$ instead of $\bar{\Theta} \times \mathcal{X} \times \mathcal{P}(\bar{\Theta} \times \mathcal{X})$: the augmentation enlarges the state space, but not the space of mean fields on which the policy depends. Moreover, when P_θ and r_θ depend regularly on θ, a neural network policy can share information across parameters, which is impossible when each parameter is solved separately. This is the parametrization we use (App. C.3).

C COMPUTING NASH EQUILIBRIUM FOR PMFG AND BE-MFG

The overall Bayesian inference pipeline of Sec. 4.3 is summarized in Algorithm 1. Its first step, computing the BE-NE, is detailed in the rest of this section.

Algorithm 1 Bayesian Inference for PMFGs

Require: PMFGs $\{\mathcal{M}_\theta\}_{\theta\in\bar{\Theta}}$; prior q over $\bar{\Theta}$; available time steps $\mathcal{I}$; observations $(\hat{\xi}^n_{\theta_0,\mathcal{I}})_{1\le n\le N}$.

1: Compute a BE-NE $\bar{\pi}^*$ of $\bar{\mathcal{M}}_q$.
2: Compute conditional density $\Gamma_{\cdot,\mathcal{I}}$ to learn the conditional distributions $(\Gamma_{\theta,\mathcal{I}})_{\theta\in\bar{\Theta}}$, by sampling mean-field trajectories using $\bar{\pi}^*$ and $\bar{\mathcal{M}}_q$.
3: Compute posterior $\bar{q}_N$ using $\Gamma^\gamma_{\cdot,\mathcal{I}}$ and the observations $(\hat{\xi}^n_{\theta_0,\mathcal{I}})_{1\le n\le N}$.
4: **Return** $\bar{q}_N$, $\hat{\theta}^{\mathrm{MAP}}_N = \arg\max_{\theta\in\bar{\Theta}} \bar{q}_N(\theta)$, $\bar{\pi}^*$.

C.1 ALGORITHMS FOR GENERAL MFGS WITH COMMON NOISE

Computing a Nash equilibrium for a MFG with common noise requires solving a fixed-point problem: the policy must be a best response to the mean-field flow it induces. We rely on the FP algorithm (Perrin et al., 2020), which iteratively trains a best response against the average of the policy history, and has been shown to converge to a Nash equilibrium under standard assumptions. Each best response is trained by maximizing the expected return of a representative agent against the current aggregate mean-field flow, using a neural network policy π^γ (Algorithm 2). The FP loop is summarized in Algorithm 3. Once FP has converged, the history of policies $(\pi^0,\dots,\pi^K)$ encodes the Nash equilibrium as an implicit mixture. To obtain a single compact policy, we distill this history into a single neural network π^γ via Mean Field Imitation Learning (MF-IL, Algorithm 4), which matches the state-action mean-field trajectory induced by π^γ to the aggregate of the history, following (Lambrecht & Laurière, 2026).

Algorithm 2 Best Response Algorithm

Require: MFG $\mathcal{M} = (\mathcal{X},\mathcal{A},\mathcal{E}^0,\nu^0,P,r,H,\rho_0)$; Sequence of policies $(\pi^0,\dots,\pi^m)\in\Pi^{m+1}$; Iterations J; Batch size B; Initial Parameters γ_0; Optimizer.

1: **for** $j=1,\dots,J$ **do**
2: Reset batch loss: $\mathcal{L}(\gamma_{j-1})=0$.
3: Sample a batch of i.i.d. common noises $(\varepsilon^{0,s}_t)_{0\le t\le H-1}$ for $s=1,\dots,B$.
4: **for** $s=1,\dots,B$ **do**
5: Compute the aggregated sequence $\tilde{\rho}^{m,s}$ induced by $(\pi^0,\dots,\pi^m)$.
6: Compute the mean-field sequence $\rho^{\gamma_{j-1},s}$ induced by the learner $\pi^{\gamma_{j-1}}$ responding to $\tilde{\rho}^{m,s}$.
7: Accumulate the loss:

$$\mathcal{L}(\gamma_{j-1}) \leftarrow \mathcal{L}(\gamma_{j-1}) - \frac{1}{B}\sum_{t=0}^{H-1}\sum_{x,a}\rho_t^{\gamma_{j-1},s}(x)\pi_t^{\gamma_{j-1}}(a\mid x,\tilde{\rho}_t^{m,s})r(x,a,\tilde{\rho}_t^{m,s})$$

8: **end for**
9: Compute the gradient $\nabla_\gamma\mathcal{L}(\gamma_{j-1})$ and update parameters via the optimizer to obtain γ_j.
10: **end for**
11: **Return** Optimized parameters γ_J defining the best response policy π^{γ_J}.

Algorithm 3 Fictitious Play for MFG with Common Noise

Require: MFG $\mathcal{M} = (\mathcal{X},\mathcal{A},\mathcal{E}^0,\nu^0,P,r,H,\rho_0)$; Initial policy $\pi^0\in\Pi$; Number of iterations K.

1: **for** $k=1,\dots,K$ **do**
2: Train π^k as the best response to $(\pi^0,\dots,\pi^{k-1})$, using Algorithm 2.
3: **end for**
4: **Return** policy tuple $(\pi^0,\dots,\pi^K)$.

Algorithm 4 Mean Field Imitation Learning (MF-IL)

Require: MFG $\mathcal{M}$; Policy sequence $(\pi^0, \dots, \pi^K)$; Initial parameters γ_0 for parametric policy π^γ; Optimizer; Iteration number J; Batch size B.

for $j = 1, \dots, J-1$ **do**
  Sample common noises $(\varepsilon_t^{0,s})_{0\le t\le H-1}^{1\le s\le B}$.
  for $s = 1, \dots, B$ **do**
    Compute $\tilde{\mu}^{K,s}$ from $(\pi^0, \dots, \pi^K)$.
    Generate $\mu^{s,\pi^{\gamma_{j-1}}}$ induced by $\pi^{\gamma_{j-1}}$.
  end for
  Compute $\mathcal{L}(\gamma_{j-1}) = \frac{1}{B}\sum_{s,t} \|\mu_t^{s,\pi^{\gamma_{j-1}}} - \tilde{\mu}_t^{K,s}\|_1$.
  Obtain γ_j by one step of gradient of $\mathcal{L}(\gamma_{j-1})$.
end for
Return Trained parametric policy π^{γ_J}.

C.2 Nash equilibrium for PMFGs

To compute the NE policies from which we sample observations of the ground truth model $\mathcal{M}_{\theta_0}$, we apply the algorithms of App. C.1 with a standard Multi-Layer Perceptron (MLP) as the parametric policy.

C.3 Nash equilibrium for BE-MFG

To compute the BE-NE $\bar{\pi}^*$, we apply the same FP pipeline of App. C.1, replacing the standard MLP policy with a **FiLM-conditioned policy**. We briefly describe how FiLM works and refer to (Perez et al., 2018) for more details.

Architecture. The network takes as input the time step t, the state x, the current mean field ρ, and the parameter θ, and outputs a softmax distribution over actions. The backbone consists of L linear layers of hidden dimension l with tanh activations. At each layer i, before applying the activation, the hidden representation h_i is modulated by a small MLP applied to θ:

$$h_i^{\text{FiLM}} = \tanh(W_i^{(1)}\theta + b_i^{(1)}), \quad (w_i, b_i^{(3)}) = W_i^{(2)}\, h_i^{\text{FiLM}} + b_i^{(2)}, \quad h_i = W_i h'_{i-1} + b_i$$

$$h'_i = \tanh\Big(h_i \odot (1 + w_i) + b_i^{(3)}\Big),$$

The FiLM MLP maps $\theta \in \mathbb{R}^d$ to a hidden layer of size l', then outputs a vector of size $2 \times l$ which is split into the scale w_i and shift $b_i^{(3)}$ applied to the backbone activations.

The learnable parameters of the network are $\gamma = \{W_i^{(1)}, b_i^{(1)}, W_i^{(2)}, b_i^{(2)}\}_{i\in[L]}$ for the FiLM MLPs, $\{W_i, b_i\}_{i\in[L]}$ for the backbone layers, and the output layer weights.

Motivation. The BE-NE $\bar{\pi}^*$ must represent a *family* of NE policies indexed by θ. A naive shared MLP with θ concatenated to the input tends to average over θ rather than sharply distinguishing between parameter regimes. FiLM solves this by allowing θ to directly rescale and shift the internal representations at every layer, making the network's computation path highly sensitive to θ. This is essential: if $\bar{\pi}^*$ does not distinguish between different values of θ, the simulated trajectories $\xi_{\theta,\mathcal{I}}^{\bar{\pi}^*,\varepsilon^0}$ will be incorrectly distributed, corrupting the training of $\Gamma_{\theta,\mathcal{I}}^{\gamma}$ and all downstream inference.

D MEAN FIELD INVERSE REINFORCEMENT LEARNING

We refer to (Chen et al., 2022) for the original description of the method. In this appendix we describe precisely how we adapted it to our setting.

Overview of the original method. The MFIRL framework of (Chen et al., 2022) recovers a reward function $r_\theta(x, a, \rho)$ from expert state-action trajectories issued of a PMFG $\tilde{\mathcal{M}}_\theta = (\mathcal{X}, \mathcal{A}, \tilde{P}, r_\theta, H, \rho_0)$ in absence of common noise, meaning that the kernel $\tilde{P}$ does not depend of an extra source of randomness ε^0, and the policies are sequences of population-unaware policies $\tilde{\pi}(a|x)$. We denote the previous set by $\tilde{\Pi}$. We emphaze that in absence of common noise, the mean fields trajectories at equilibrium are deterministic. Given a state-action mean-field trajectory $\hat{\mu}_{\theta_0}$ for an unknown parameter θ_0, the method solves the objective

$$\max_\theta \mathcal{L}(\theta) = \sum_{i=0}^{H-1} \sum_{x,a} \kappa^i \hat{\mu}_{\theta_0,i}(x, a) r_\theta(x, a, \hat{\rho}_{\theta_0,i}) - J(\hat{\rho}_{\theta_0}, \tilde{\pi}_\theta),$$

where κ is a discount factor, $\hat{\rho}_{\theta_0,i} = \sum_a \hat{\mu}_{\theta_0,i}(\cdot, a)$, $J(\hat{\rho}_{\theta_0}, \tilde{\pi}_\theta) = \sum_{i=0}^{H-1} \kappa^i \sum_{x,a} \hat{\rho}_{\theta_0,i}(x)\, \tilde{\pi}_{\theta,i}(a|x)\, r_\theta(x, a, \hat{\rho}_{\theta_0,i})$ and $\tilde{\pi}_\theta$ is a Boltzmann best response to $\hat{\mu}_{\theta_0}$ under r_θ, computed by backward induction via the soft Q-function

$$\tilde{Q}_\theta(i, x, a, \hat{\rho}_{\theta_0}) = r_\theta(x, a, \rho_i) + \kappa\, \mathbb{E}_{X \sim \tilde{P}(\cdot|x,a,\hat{\rho}_{\theta_0,i})} \left[\sum_{a'} \tilde{\pi}_{\theta,i+1}(a'|X) \tilde{Q}_\theta(i+1, X, a', \hat{\rho}_{\theta_0}) \right],$$

with $\tilde{\pi}_{\theta,i}(a|x) \propto \exp(\tau^{-1}\, \tilde{Q}_\theta(i, x, a, \rho))$, where $\tau > 0$ is a temperature. In the original paper the Q-function is tabular, which is tractable because there is no common noise and the state space is small.

Adaptations required by our setting.

(i) Common Noise Approximation. Our setting considers PMFGs $\mathcal{M}_\theta = (\mathcal{X}, \mathcal{A}, \mathcal{E}^0, \nu^0, P, r_\theta, H, \rho_0)$ where the transition kernel P depends explicitly on a common noise realization $\varepsilon^0 \sim \nu^0$. When applying MF-IRL, we replace P by its common noise averaged conterpart $\tilde{P}(\cdot|x, a, \rho) = \mathbb{E}_{\varepsilon^0}[P(\cdot|x, a, \rho, \varepsilon^0)]$.

Under common noise, the mean-field trajectory $\hat{\rho}_{\theta_0}$ is stochastic, so the observer has access to N independent realizations $(\hat{\mu}^n_{\theta_0,i})_{1 \le n \le N}$ rather than a single deterministic trajectory. We approximate $\hat{\mu}_{\theta_0}$ by the average $\hat{\mu}^E_i = \frac{1}{N} \sum_{n=1}^N \hat{\mu}^n_{\theta_0,i}$, and compute $\tilde{Q}_\theta$ using $\tilde{P}(\cdot|x, a, \hat{\rho}^E_i)$, where $\hat{\rho}^E_i = \sum_a \hat{\mu}^E_i(\cdot, a)$.

We therefore follow the population-unaware approximation: the Boltzmann policy $\tilde{\pi}_{\theta,t}(a|x)$ is computed from the soft Q-function evaluated at the *fixed* averaged mean field $\hat{\mu}^E_{\theta_0,i} = \frac{1}{N} \sum_{n=1}^N \hat{\mu}^n_{\theta_0}$, using the common noise averaged transition matrix $\tilde{P}(\cdot|x, a, \hat{\rho}^E_i)$.

This approximation is consistent with the population-unaware/aware policy distinction introduced in (Lambrecht & Laurière, 2026) for the common-noise setting.

Concretely, our implementation computes the averaged transition matrix $\tilde{P}$ once, then runs the backward scan

$$\tilde{Q}_\theta(H-1) = R_\theta(H-1),$$
$$\tilde{Q}_\theta(i) = R_\theta(i) + \kappa\, \tilde{P} \cdot V_\theta(i+1), \quad V_\theta(i+1, x) = \sum_a \tilde{\pi}_{\theta,i+1}(a|x)\, \tilde{Q}_\theta(i+1, x, a),$$

where $R_\theta(i, x, a) = r_\theta(x, a, \hat{\rho}^E_i)$ uses the parametric reward at the empirical density, and $\tilde{\pi}_{\theta,i} = \mathrm{softmax}(\tau^{-1} \tilde{Q}_\theta(i))$.

(ii) Requirement of state–action observations ($\xi = \mu$). The backward induction requires the full state-action mean-field trajectory $\hat{\mu}_{\theta_0} \in (\Delta_{\mathcal{X} \times \mathcal{A}})^H$. The method therefore cannot be applied when the observer only sees state distributions $\xi = \rho$, which is the harder and more realistic setting we also evaluate. This is a structural limitation relative to our Bayesian approach, which handles both $\xi = \mu$ and $\xi = \rho$.

(iii) Handling partial observations ($\mathcal{I} \neq [H-1]$). The original paper observes the full trajectory. When only a subset $\mathcal{I} \subset [H-1]$ of time steps is available we construct the full H-step empirical flow $\hat{\mu}^E_{\theta_0}$ by nearest-future imputation: for any unobserved step t, we substitute the next observed step $\min\{i \in \mathcal{I} : i \geq t\}$, or the last observed step if none exists. This natural imputation allows the backward induction to run over the full horizon, and we investigate its degradation as $|\mathcal{I}|$ decreases.

The full procedure is given in Algorithm 5.

Algorithm 5 Adaptation of MF-IRL to our setting

Require: Family of PMFGs $\{M_\theta\}_{\theta \in \bar{\Theta}}$; parametric reward r_θ; N expert state-action trajectories $(\hat{\mu}^n_{\theta_0,\mathcal{I}})_{1 \leq n \leq N}$ at accessible steps $\mathcal{I}$; temperature τ; discount κ; Initial guess $\theta^0 \in \bar{\Theta}$; Number of iterations T_{IRL}; Optimizer.

1: Compute empirical mean-field: $\hat{\mu}^E_{\theta_0,i} = \frac{1}{N}\sum_{n=1}^N \hat{\mu}^n_{\theta_0,i}$ for $i \in \mathcal{I}$.
2: Impute missing steps: for $i \notin \mathcal{I}$, set $\hat{\mu}^E_i = \hat{\mu}^E_{\min\{i' \in \mathcal{I} : i' \geq i\}}$
3: Compute the common noise averaged transition matrix $\tilde{P}(x'|x,a,\hat{\rho}^E_i) = \mathbb{E}_{\varepsilon^0}[P(x'|x,a,\hat{\rho}^E_i,\varepsilon^0)]$.
4: **for** for $m = 0$ to $T_{\mathrm{IRL}} - 1$ **do**
5: Compute $\tilde{Q}_{\theta^m}$.
6: Compute the loss $\mathcal{L}(\theta^m) = \sum_{i=0}^{H-1}\sum_{x,a} \kappa^i \hat{\mu}_{\theta_0,i}(x,a) r_{\theta^m}(x,a,\hat{\rho}_{\theta_0,i}) - J(\hat{\rho}_{\theta_0}, \tilde{\pi}_{\theta^m})$.
7: Proceed with one gradient step to obtain θ^{m+1}.
8: **end for**
9: **Return** $\theta^{T_{\mathrm{IRL}}}$.

E LEARNING $\Gamma_{\theta,\mathcal{I}}$

E.1 PRINCIPAL COMPONENT ANALYSIS

Since $\xi_{\mathcal{I}} \in (\Delta_{\mathcal{X}})^{|\mathcal{I}|}$ is high-dimensional, we first apply principal component analysis (PCA) to a training set $\{(\theta^n, \xi^{*,n}_{\theta^n,\mathcal{I}})\}_{n=1}^{N_{\text{pca}}}$ sampled using the BE-NE $\bar{\pi}^*$ and the prior q. Centering the trajectories and computing the SVD of the centered matrix, we retain the top d_{PCA} components explaining 99% of the variance, yielding a projection $W \in \mathbb{R}^{D|\mathcal{I}|\times d_{\text{PCA}}}$ where D refers to the dimension of the observation space Ξ. Each observation is then mapped to a normalized PCA embedding $z = ((\xi_{\mathcal{I}} - \bar{\xi})W - \bar{z})/\sigma_z \in \mathbb{R}^{d_{\text{PCA}}}$. PCA is a well understood and classic method in the Machine Learning literature (Jolliffe & Cadima, 2016).

E.2 GAUSSIAN PROCESSES ESTIMATION

We describe the GP-based method used as a lightweight alternative to the MAF. Contrary to the MAF, it does not learn the conditional density $\Gamma_{\theta,\mathcal{I}}$: it is an inverse regression, which directly models the map $z \mapsto \theta$ in PCA space, and outputs a Gaussian approximation of the posterior used for MAP and BMA inference.

GP training in PCA space. We place a GP prior on the map $z \mapsto \theta$ with a linear kernel $k(z, z') = z^\top z'$, making the model equivalent to Bayesian linear regression in the PCA feature space. Given N_{train} training pairs $\{(z^n, \theta^n)\}_{n=1}^{N_{\text{train}}}$ sampled from the BE-NE $\bar{\pi}^*$ and the prior q, the Gram matrix is $\mathbf{K} = (k(z^n, z^{n'}))_{n,n'} \in \mathbb{R}^{N_{\text{train}}\times N_{\text{train}}}$ and the GP is fitted by computing $w = (\mathbf{K} + \sigma^2_{\text{noise}} I)^{-1} Y$, where $Y = (\theta^1, \dots, \theta^{N_{\text{train}}})^\top \in \mathbb{R}^{N_{\text{train}}\times d}$ stacks the targets and σ^2_{noise} is a hyperparameter controlling the confidence of the model. The training algorithm is summarized in Algorithm 6.

Posterior inference for $N = 1$. For a single observation $\hat{\xi}_{\mathcal{I}}$, let $\hat{z}$ be its PCA embedding and $\hat{k} = (k(\hat{z}, z^n))_{n=1}^{N_{\text{train}}} \in \mathbb{R}^{N_{\text{train}}}$. The GP posterior is Gaussian $\mathcal{N}(\hat{\theta}^{\text{MAP}}, \hat{\sigma}^2)$ with

$$\hat{\theta}^{\text{MAP}} = \hat{k}^\top w,$$
$$\hat{\sigma}^2 = k(\hat{z}, \hat{z}) - \hat{k}^\top (\mathbf{K} + \sigma^2_{\text{noise}} I)^{-1} \hat{k},$$

where $\hat{\sigma}^2$ is the posterior variance measuring the uncertainty of the estimate. The MAP estimator is $\hat{\theta}^{\text{MAP}}$.

Multiple observations ($N > 1$). When N i.i.d. observations $(\hat{\xi}^n_{\theta_0,\mathcal{I}})_{1\le n\le N}$ are available, each observation $\hat{\xi}^n_{\theta_0,\mathcal{I}}$ yields an independent GP posterior $\mathcal{N}(m_n, v_n)$, where $m_n = \hat{k}_n^\top w$ is the predicted parameter and $v_n = k(\hat{z}_n, \hat{z}_n) - \hat{k}_n^\top (\mathbf{K} + \sigma^2_{\text{noise}} I)^{-1} \hat{k}_n$ is the associated posterior variance. Each $\mathcal{N}(\theta; m_n, v_n)$ approximates the single-observation posterior, which is proportional to $q(\theta)\Gamma_{\theta,\mathcal{I}}(\hat{\xi}^n_{\theta_0,\mathcal{I}})$. Since the observations are conditionally independent given θ, Bayes' rule (4) involves the product of the likelihoods and a single prior factor, whereas the product of the N single-observation posteriors contains the prior N times. We therefore combine them as

$$\bar{q}^{\text{GP}}_N(\theta) \propto q(\theta)^{1-N} \prod_{n=1}^{N} \mathcal{N}(\theta; m_n, v_n).$$

When the prior is Gaussian, $q = \mathcal{N}(\mu_0, v_0)$, $\bar{q}^{\text{GP}}_N$ is Gaussian, $\bar{q}^{\text{GP}}_N = \mathcal{N}(\hat{\theta}^{\text{MAP}}_N, v_{\text{comb}})$, with

$$v^{-1}_{\text{comb}} = \sum_{n=1}^{N} v_n^{-1} - (N-1)v_0^{-1}, \qquad \hat{\theta}^{\text{MAP}}_N = v_{\text{comb}}\Big(\sum_{n=1}^{N} v_n^{-1} m_n - (N-1)v_0^{-1}\mu_0\Big).$$

When the prior is uniform on $\bar{\Theta}$, q^{1-N} is constant on $\bar{\Theta}$, and $\bar{q}^{\text{GP}}_N$ is the product of the N Gaussians restricted to $\bar{\Theta}$. Both reduce to the single-observation GP posterior when $N = 1$. The BMA policy is then $\pi_{\bar{q}^{\text{GP}}_N} = \mathbb{E}_{\theta\sim\bar{q}^{\text{GP}}_N}[\pi^*_\theta]$.

Algorithm 6 Training the Linear Gaussian Processes Estimator

Require: Family of PMFGs $\{\mathcal{M}_\theta\}_{\theta\in\bar{\Theta}}$; BE-NE policy $\bar{\pi}^*$ of $\bar{\mathcal{M}}_q$; Type of observations ξ; available time steps $\mathcal{I}$; Number of PCA training samples N_{pca}; PCA variance threshold p_{var}; Number of GP training samples N_{train}; Noise hyperparameter σ^2_{noise}.

1: Sample N_{pca} i.i.d. pairs $(\Theta^b, \xi^{\bar{\pi}^*,\varepsilon^{0,b}}_{\Theta^b,\mathcal{I}})_{1\le b\le N_{\text{pca}}}$ with $\Theta^b \sim q$ and $\varepsilon^{0,b} \sim (\nu^0)^{\otimes H}$, using $\bar{\pi}^*$.
2: Compute PCA projection $W \in \mathbb{R}^{D|\mathcal{I}|\times d_{\text{PCA}}}$ retaining p_{var} variance; store mean $\bar{\xi}$ and normalization $(\bar{z}, \sigma_z)$.
3: Sample N_{train} i.i.d. pairs $(\Theta^n, \xi^{\bar{\pi}^*,\varepsilon^{0,n}}_{\Theta^n,\mathcal{I}})_{1\le n\le N_{\text{train}}}$ with $\Theta^n \sim q$ and $\varepsilon^{0,n} \sim (\nu^0)^{\otimes H}$.
4: **for** $n = 1$ **to** N_{train} **do**
5: Project: $z^n = \left((\xi^n_{\mathcal{I}} - \bar{\xi})\, W - \bar{z}\right)/\sigma_z$.
6: **end for**
7: Compute Gram matrix: $\mathbf{K}_{nn'} = (z^n)^\top z^{n'}$ for all $n, n' \in [N_{\text{train}}]$.
8: Compute $w = (\mathbf{K} + \sigma^2_{\text{noise}}\, I)^{-1} Y$, where $Y = (\Theta^1, \ldots, \Theta^{N_{\text{train}}})^\top$.
9: **Return** $w, \mathbf{K}$.

E.3 Masked auto regressive flows

Normalizing flows like MAF are powerful density estimators that transform a simple distribution into a complex one through a series of bijective transformations, making them ideal for capturing the intricate dependencies in Mean Field trajectories.

MAF learns the conditional density $\Gamma_{\theta,\mathcal{I}}(\xi)$ by transforming a simple base distribution $p_0 = \mathcal{N}(0, I_{d_{pca}})$ (where d_{pca} is the dimension retained by the PCA as explained in E.1) into a complex one through a sequence of L bijective coupling layers. At each layer, the current representation $x^{(i-1)} \in \mathbb{R}^{d_{\text{PCA}}}$ (with $x^{(0)} = z$ the PCA embedding of $\xi_{\mathcal{I}}$) is split into two halves $(x_1^{(i-1)}, x_2^{(i-1)})$ and only the second half is transformed:

$$[s^{(i)}, t^{(i)}] = \text{MLP}^{(i)}([x_1^{(i-1)}; \theta]), \qquad y_2^{(i)} = x_2^{(i-1)} \odot \exp(s^{(i)}) + t^{(i)}, \qquad x^{(i)} = [y_2^{(i)}, x_1^{(i-1)}],$$

where $\text{MLP}^{(i)}$ is a two-hidden-layer network with tanh activations. Let γ be the trainable parameters of a MAF, and denote $\Gamma^\gamma_{\cdot,\mathcal{I}}$ the transformation defined above. By the change-of-variables formula, the MAF approximates the log-likelihood of a PCA-projected observation $z = ((\xi^*_{\theta,\mathcal{I}} - \bar{\xi})W - \bar{z})/\sigma_z$ under parameter θ as:

$$\log \Gamma^\gamma_{\theta,\mathcal{I}}(z) = \log p_0(x^{(L)}) + \sum_{i=1}^{L} \mathbf{1}^\top s^{(i)},$$

where $x^{(L)}$ is the final transformed variable and $z = x^{(0)}$ is the input. The MAF thus returns a tractable scalar approximation of $\log \Gamma_{\theta,\mathcal{I}}(\xi)$ for any (θ, ξ) pair, which is the quantity needed for Bayesian inference. The training objective is the expected negative log-likelihood (NLL) over pairs $(\Theta, \xi^*_{\Theta,\mathcal{I}})$ sampled from the BE-NE $\bar{\pi}^*$ and the prior q:

$$\mathcal{L}(\gamma) = -\mathbb{E}_{\Theta\sim q}\left[\log \Gamma^\gamma_{\Theta,\mathcal{I}}(z^*_\Theta)\right], \quad z^*_\Theta = ((\xi^*_{\Theta,\mathcal{I}} - \bar{\xi})W - \bar{z})/\sigma_z,$$

which is minimized by stochastic gradient descent as detailed in Algorithm 7.

Ensemble of MAF. It is important to note that training a MAF can be highly sensitive to the initial values of its parameters. To mitigate the dependency on a specific initialization and ensure more robust density estimation, we employ an ensemble approach. Specifically, we train M independent MAF models starting from different random initializations. Let us denote the models $\Gamma^{\gamma_1}_{\mathcal{I}}, \ldots, \Gamma^{\gamma_M}_{\mathcal{I}}$. We then aggregate their outputs by averaging their log-likelihoods to produce a more stable and reliable estimate of the global log-likelihood. For each $\theta \in \bar{\Theta}$ and $\xi \in \Xi$, define

$$\Gamma^{\gamma_1,\ldots,\gamma_M}_{\theta,\mathcal{I}}(z) \propto \exp\left(\frac{1}{M}\sum_{j=1}^{M} \log \Gamma^{\gamma_j}_{\theta,\mathcal{I}}(z)\right).$$

Algorithm 7 Training MAF $\Gamma^{\gamma}_{\mathcal{I}}$

Require: Family of PMFGs $(\mathcal{M}_\theta)_{\theta\in\bar{\Theta}}$; BE-NE policy $\bar{\pi}^*$ of $\bar{\mathcal{M}}_q$; Type of observations $\xi \in \{\rho, \mu\}$; available time steps $\mathcal{I} \subset [H-1]$; Initial parameters γ_0 for the MAF $\Gamma^{\gamma}_{\mathcal{I}}$; Number of iterations T; Batch size B; Optimizer; Number of PCA training samples N_{pca}; PCA variance threshold p_{var}.

1: Sample N_{pca} i.i.d. pairs $(\Theta^b, \xi^{\bar{\pi}^*,\varepsilon^{0,b}}_{\Theta^b,\mathcal{I}})_{1\le b\le N_{\text{pca}}}$ with $\Theta^b \sim q$ and $\varepsilon^{0,b} \sim (\nu^0)^{\otimes H}$, using $\bar{\pi}^*$.
2: Compute PCA projection $W \in \mathbb{R}^{D|\mathcal{I}|\times d_{\text{PCA}}}$ retaining p_{var} variance; store mean $\bar{\xi}$ and normalization $(\bar{z}, \sigma_z)$.
3: **for** $t = 1$ **to** T **do**
4: Sample an i.i.d. batch of parameters $(\Theta^b)_{1\le b\le B}$ from prior q.
5: **for** $b = 1$ **to** B **do**
6: Sample one common noise trajectory $\varepsilon^{0,b} \sim (\nu^0)^{\otimes H}$.
7: Compute associated mean field trajectory $\xi^{\bar{\pi}^*,\varepsilon^{0,b}}_{\Theta^b,\mathcal{I}}$.
8: Project into PCA space: $z^b = ((\xi^{\bar{\pi}^*,\varepsilon^{0,b}}_{\Theta^b,\mathcal{I}} - \bar{\xi})W - \bar{z})/\sigma_z$.
9: **end for**
10: Compute NLL loss: $\mathcal{L}(\gamma_t) = -\dfrac{1}{B}\displaystyle\sum_{b=1}^{B} \log \Gamma^{\gamma_t}_{\Theta^b,\mathcal{I}}(z^b)$.
11: Update γ_{t+1} using one step of the Optimizer on $\nabla_\gamma \mathcal{L}(\gamma_t)$.
12: **end for**
13: **Return** Trained estimator $\Gamma^{\gamma_T}_{\mathcal{I}}$.

F EXPERIMENT 1 (BEACH BAR): DETAILS

F.1 MODEL DESCRIPTION

Let $\mathcal{X} = \{0, \dots, 6\}$, $\mathcal{A} = \{-1, 0, 1\}$, $H = 20$, and $\mathcal{E}^0 = \{-1, 0, 1\}^{|\mathcal{X}|}$. The environment is disrupted by external events that simultaneously displace all agents at specific locations. This is governed by a common noise whose frequency depends on a constant $\eta \geq 0$. Given η, and common noise $e^0 \in \mathcal{E}^0$, the dynamics are governed by a transition kernel P that incorporates both *common noise* $e^0(x)$ and *idiosyncratic noise* (represented by a uniform shift $s \in \{-1, 0, 1\}$). Let for any $(x, a, \rho, e^0) \in \mathcal{X} \times \mathcal{A} \times \Delta_{\mathcal{X}} \times \mathcal{E}^0$, $P(x, a, \rho, e^0) = \frac{1}{3} \sum_{s \in \{-1,0,1\}} \delta_{x+a+e^0(x)+s \pmod{|\mathcal{X}|}}$. The common noise is $\varepsilon^0 \sim (\nu_\eta^0)^{\otimes 7}$, where ν_η^0 is the distribution of $\varepsilon_1^0 \varepsilon_2^0$ with $\varepsilon_1^0 \sim \text{Bernoulli}(\eta)$ and $\varepsilon_2^0 \sim \text{Unif}\{-1, 0, 1\}$. The reward function, reflecting a preference for a central bar at $x_{\text{bar}} = 3$ and a penalty for overcrowding, is parameterized by $\theta \geq 0$: $r_\theta(x, a, \rho) = -|x - x_{\text{bar}}| - \theta \log(\rho(x) + 10^{-5}) - |a|/|\mathcal{X}|$. Fig. 4 shows realizations of the mean field for two different values of θ: while the agents try to be close to the bar, larger θ leads to lower density values because the agent spread more across states.

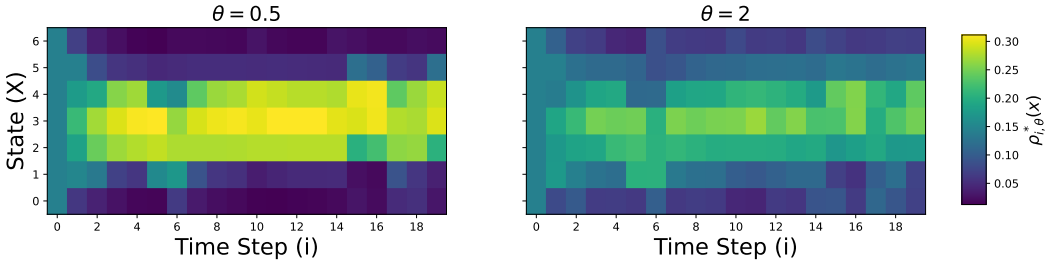

Figure 4: Mean Field Equilibrium Trajectories for two different values of θ, with $\eta = 0.2$, on the same common noise realization.

F.2 HYPERPARAMETERS

Table 3: Hyperparameters for the Beach Bar Experiment

Hyperparameter	**Value**
State/Action Space	$\lvert\mathcal{X}\rvert = 7, \lvert\mathcal{A}\rvert = 3$
Horizon (H)	20
Common Noise Intensity (η)	$\{0.0, 0.1, 0.2, 0.3, 0.4, 0.5\}$
Prior $q(\theta)$	GP: $\mathcal{N}(\frac{5}{4}, 0.01)$; MAF: Unif$([0, \frac{5}{2}])$
Test Parameters	$\{0.5, 0.875, 1.25, 1.625, 2\}$
Observation Horizon ($\lvert\mathcal{I}\rvert$)	$\{19, 5\}$
Bayesian Fictitious Play Rounds (K_{bays})	15
Fictitious Play Inner Epochs	8000
Distillation Epochs	50000
Monte Carlo Sample Size (M)	10000
Learning Rate (Fictitious Play / Distillation)	5×10^{-4}
Batch Size	512
PCA Variance Threshold	0.99
Normalizing Flow (MAF) Epochs	25 000
Normalizing Flow Learning Rate	1×10^{-4}
Normalizing Flow Ensemble Size	5 models
Normalizing Flow Architecture	6 layers, 128 hidden units
Posterior Grid Resolution (n_{grid})	300

F.3 FULL SENSITIVITY ANALYSIS

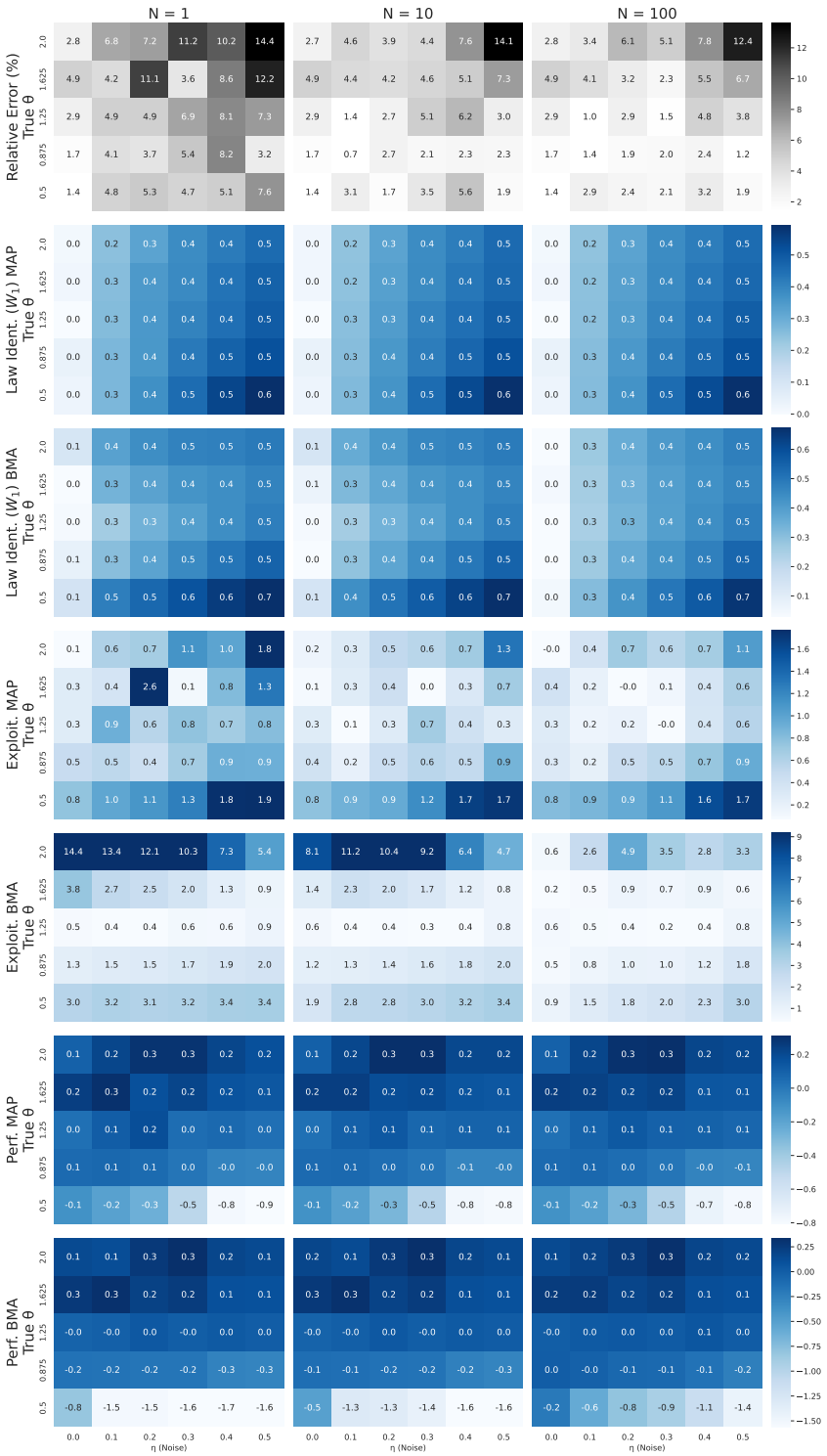

Figure 5: Sensitivity analysis, MAF, Averaged Metrics per θ_0, $\xi = \rho, \mathcal{I} = [H-1]$.



Figure 6: Sensitivity analysis, MAF, Standard Deviation Metrics per θ_0, $\xi = \rho, \mathcal{I} = [H-1]$.

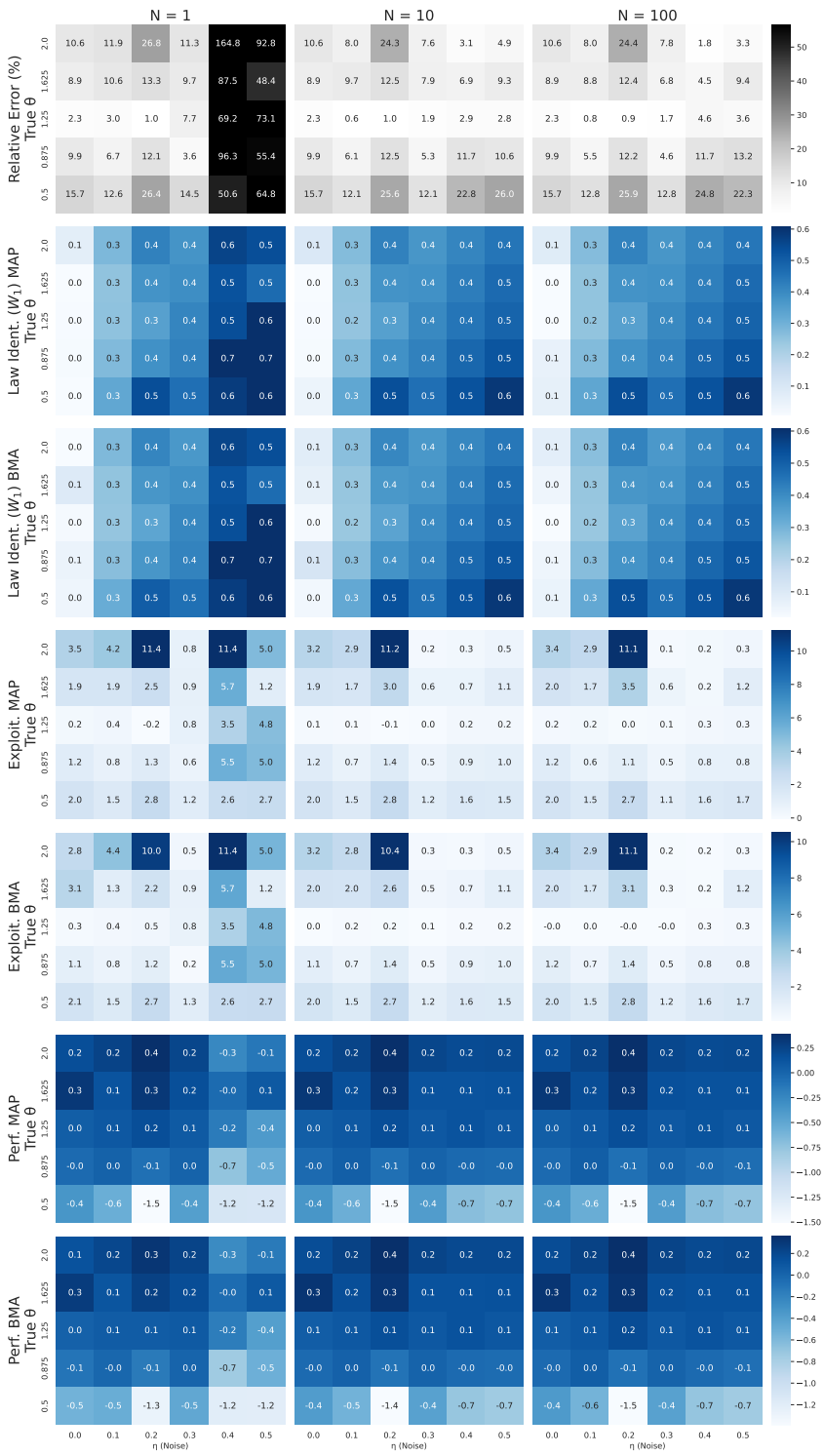

Figure 7: Sensitivity analysis, GP, Averaged Metrics per θ_0, $\xi = \rho, \mathcal{I} = [H-1]$.

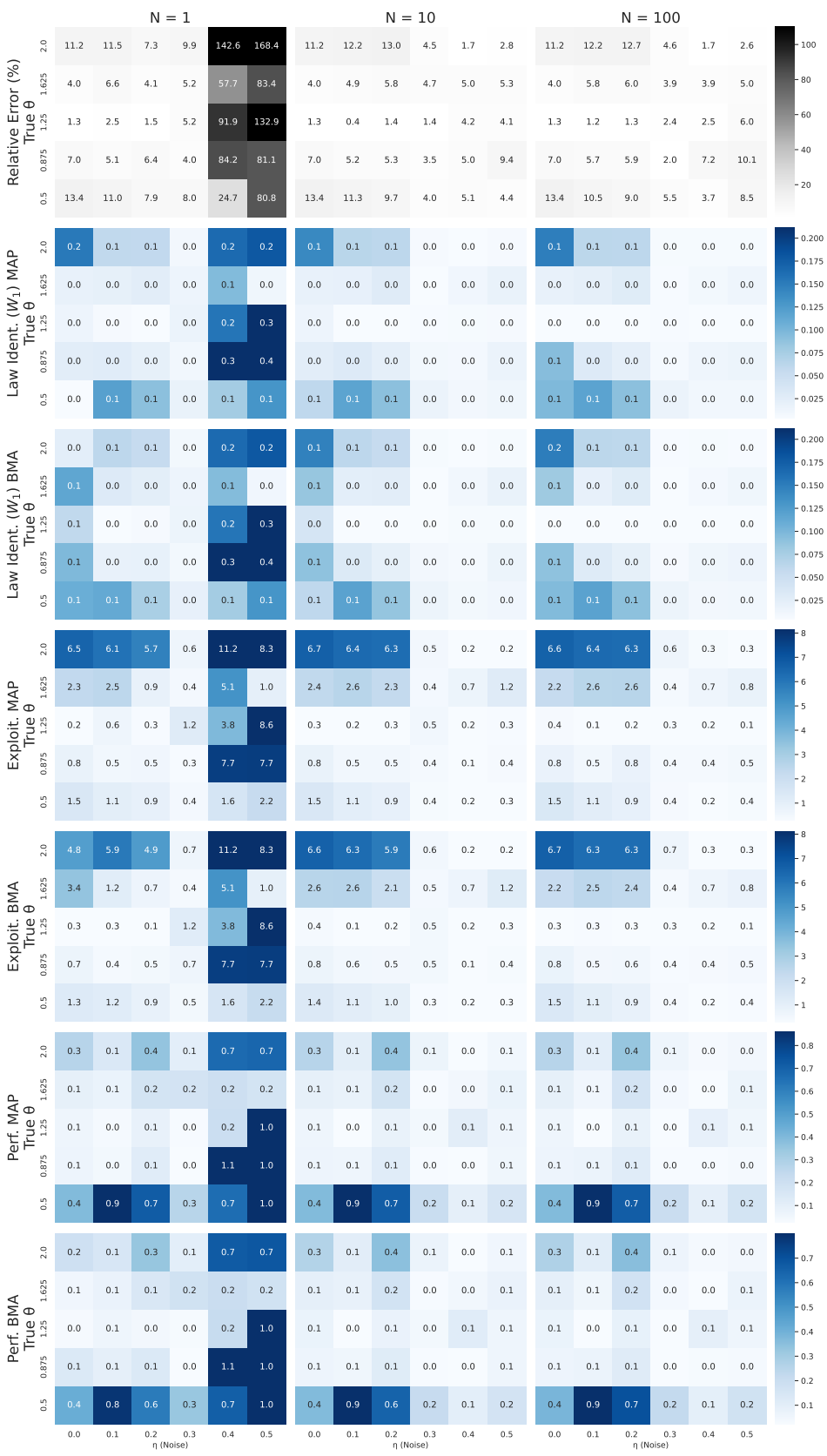

Figure 8: Sensitivity analysis, GP, Standard Deviation Metrics per θ_0, $\xi = \rho$, $\mathcal{I} = [H-1]$.

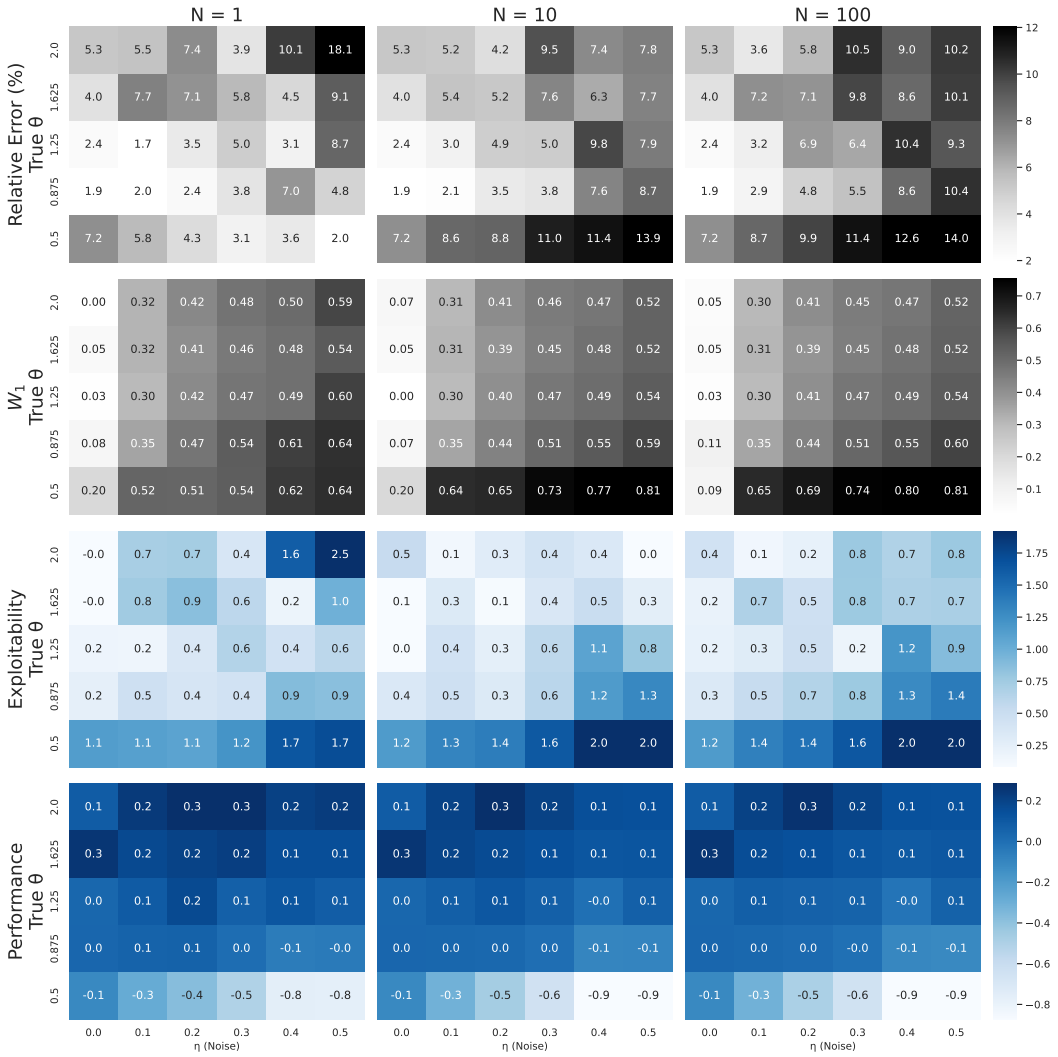

Figure 9: Sensitivity analysis, MF-IRL, Averaged Metrics per θ_0, $\xi = \mu$, $\mathcal{I} = [H-1]$.

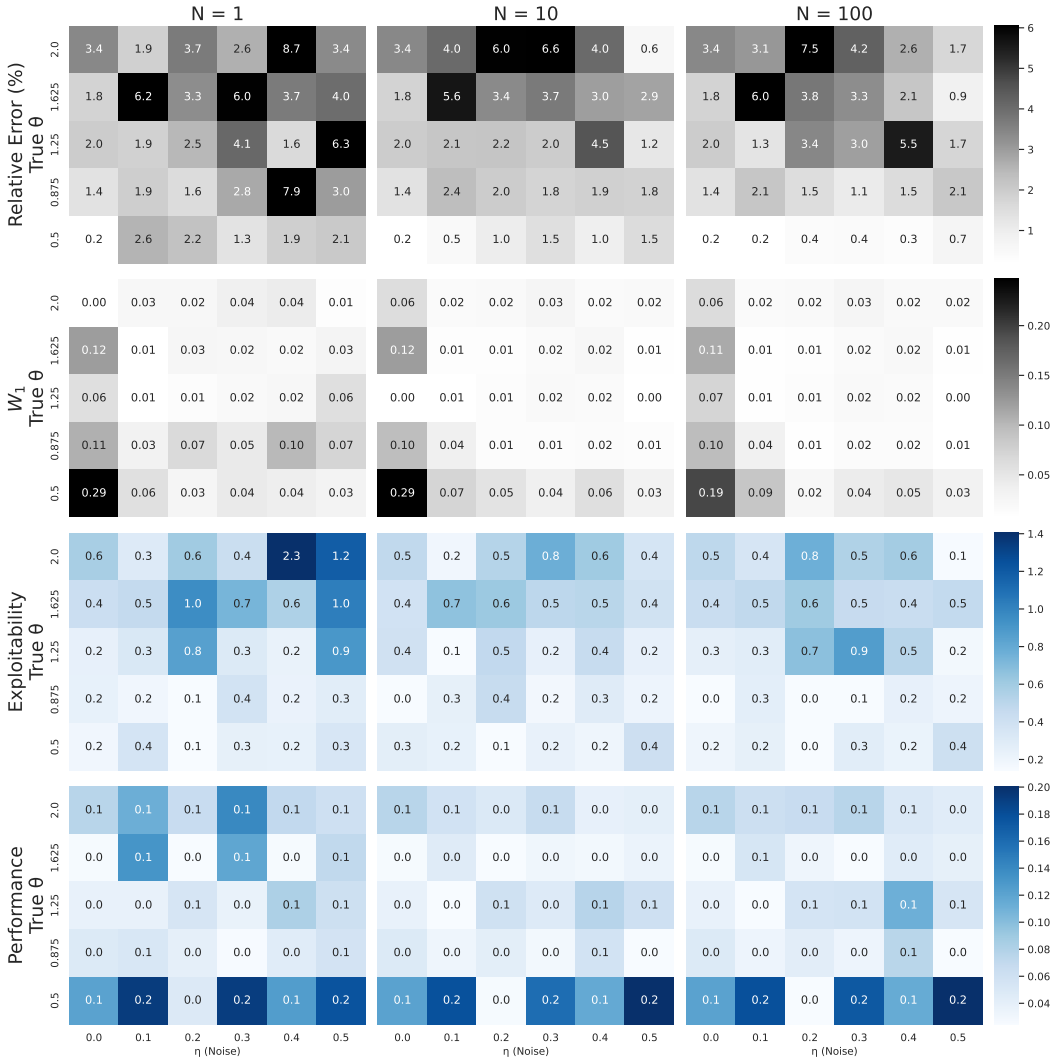

Figure 10: Sensitivity analysis, MF-IRL, Standard Deviation Metrics per θ_0, $\xi = \mu$, $\mathcal{I} = [H-1]$.

G EXPERIMENT 2 (DISTRICT): DETAILS

G.1 MODEL DESCRIPTION

Let $\mathcal{X} = \{0,1,2\}^2$, $\mathcal{A} = \{(0,0),(\pm 1,0),(0,\pm 1)\}$, and $\mathcal{E}^0 = \{0,1\}^{|\mathcal{X}|}$. Let for any $(x,a,\rho,e^0) \in \mathcal{X}\times\mathcal{A}\times\Delta_{\mathcal{X}}\times\mathcal{E}^0$, $P(x,a,\rho,e^0) = \mathbb{1}_{\{e^0(x)=0\}}\delta_{x+a \bmod 3} + \mathbb{1}_{\{e^0(x)=1\}}\frac{1}{|\mathcal{A}|}\sum_{s\in\mathcal{A}}\delta_{x+s \bmod 3}$, where the sum $x+a$ and the reduction modulo 3 are taken componentwise. The common noise law is defined as $\varepsilon^0 \sim (\nu_\eta^0)^{\otimes|\mathcal{X}|}$, where $\nu_\eta^0 = \text{Bernoulli}(\eta)$ and $\eta \geq 0$ is the common noise intensity. The reward function, reflecting a preference for specific hotspots and a penalty for overcrowding, is: $r_{(\theta^e,\theta^c)}(x,a,\rho) = 100\theta^e(x) - 2\theta^c \log(\rho(x) + 10^{-5}) - |a|/5$ for $\theta = (\theta^e,\theta^c) \in \bar{\Theta} = \mathbb{R}_+^9 \times [0,5]$.

G.2 MEAN FIELD ILLUSTRATION AND GROUND TRUTH (θ_0)

Fig. 11 shows realizations of the mean field for three different values of θ.

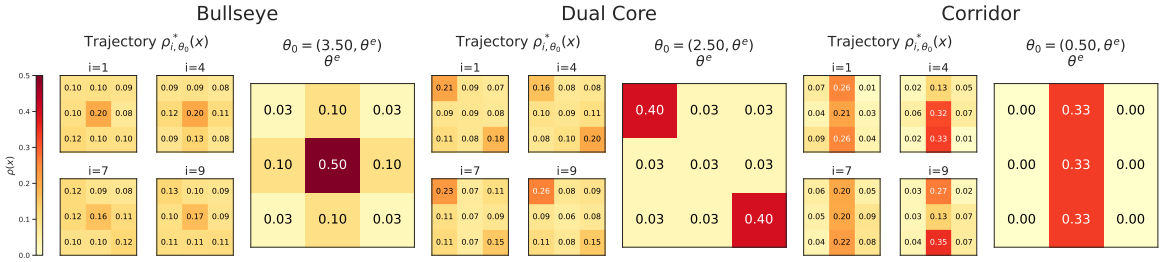

Figure 11: Mean Field Equilibrium Trajectories for three different scenarios of θ_0, with $\eta = 0.3$.

G.3 HYPERPARAMETERS

Table 4: Hyperparameters for the Entertainment District Experiment

Hyperparameter	**Value**
State/Action Space	$\lvert\mathcal{X}\rvert = 9, \lvert\mathcal{A}\rvert = 5$
Horizon (H)	10
Common Noise Intensity (η)	$\{0.3, 0.5\}$
Prior $q(\theta)$	GP: $\mathcal{N}\big(((\frac{1}{9},\ldots,\frac{1}{9}),\frac{5}{2}),0.01\big)$; MAF: Dirichlet$(1,\ldots,1) \otimes$ Unif$[0,5]$
Test Parameters	See Figure 11
Observation Horizon ($\lvert\mathcal{I}\rvert$)	$\{9, 1\}$
Bayesian Fictitious Play Rounds (K_{bays})	20
Fictitious Play Inner Epochs	8000
Distillation Epochs	50000
Monte Carlo Sample Size (M)	10000
Learning Rate (Fictitious Play / Distillation)	5×10^{-4}
Batch Size	512
PCA Variance Threshold	1
Normalizing Flow (MAF) Epochs	25 000
Normalizing Flow Learning Rate	1×10^{-4}
Normalizing Flow Ensemble Size	5 models
Normalizing Flow Architecture	6 layers, 128 hidden units

G.4 FULL SENSITIVITY ANALYSIS

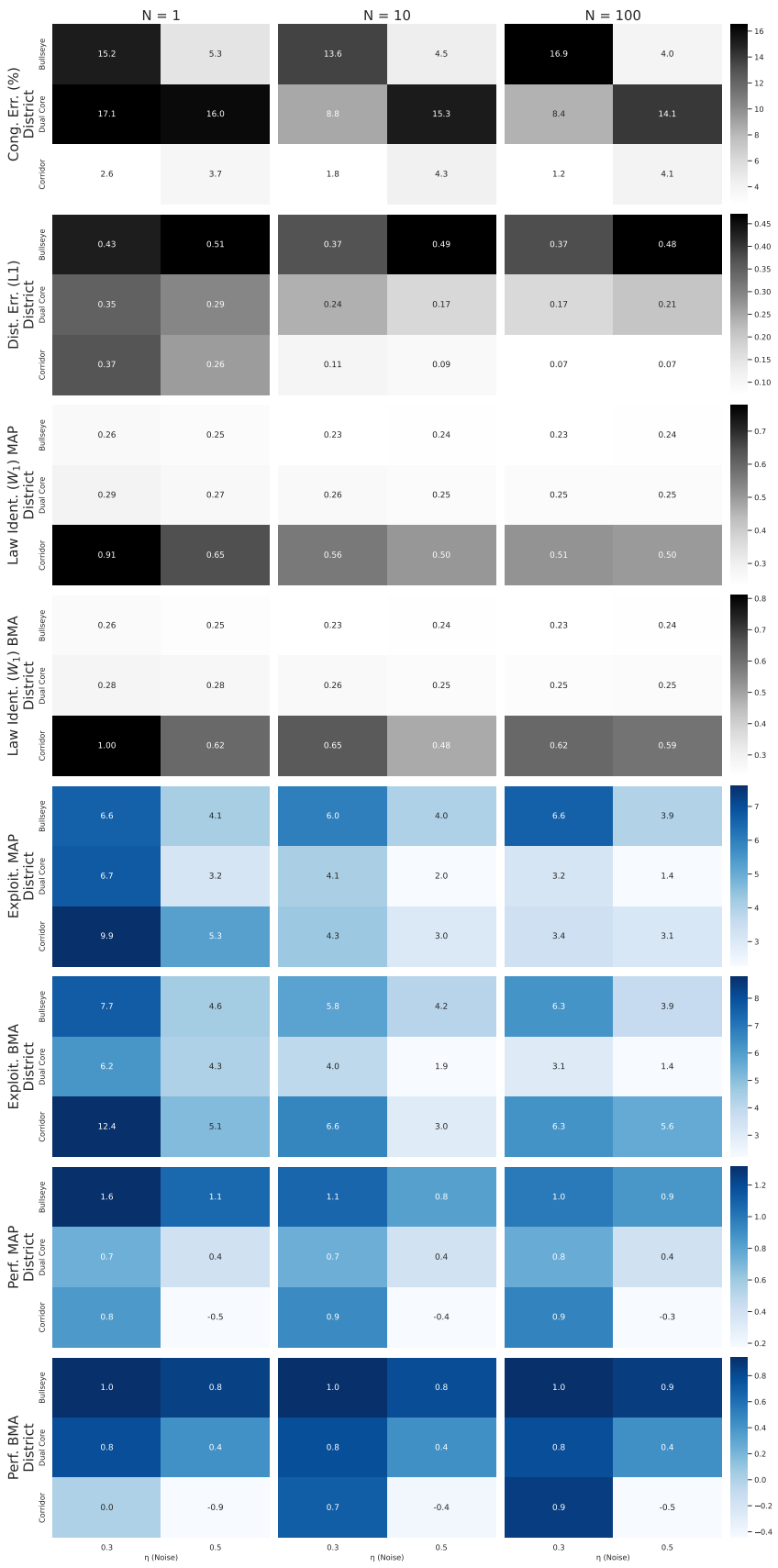

Figure 12: Sensitivity analysis, MAF, Averaged Metrics per θ_0, $\xi = \rho$, $\mathcal{I} = [H-1]$.

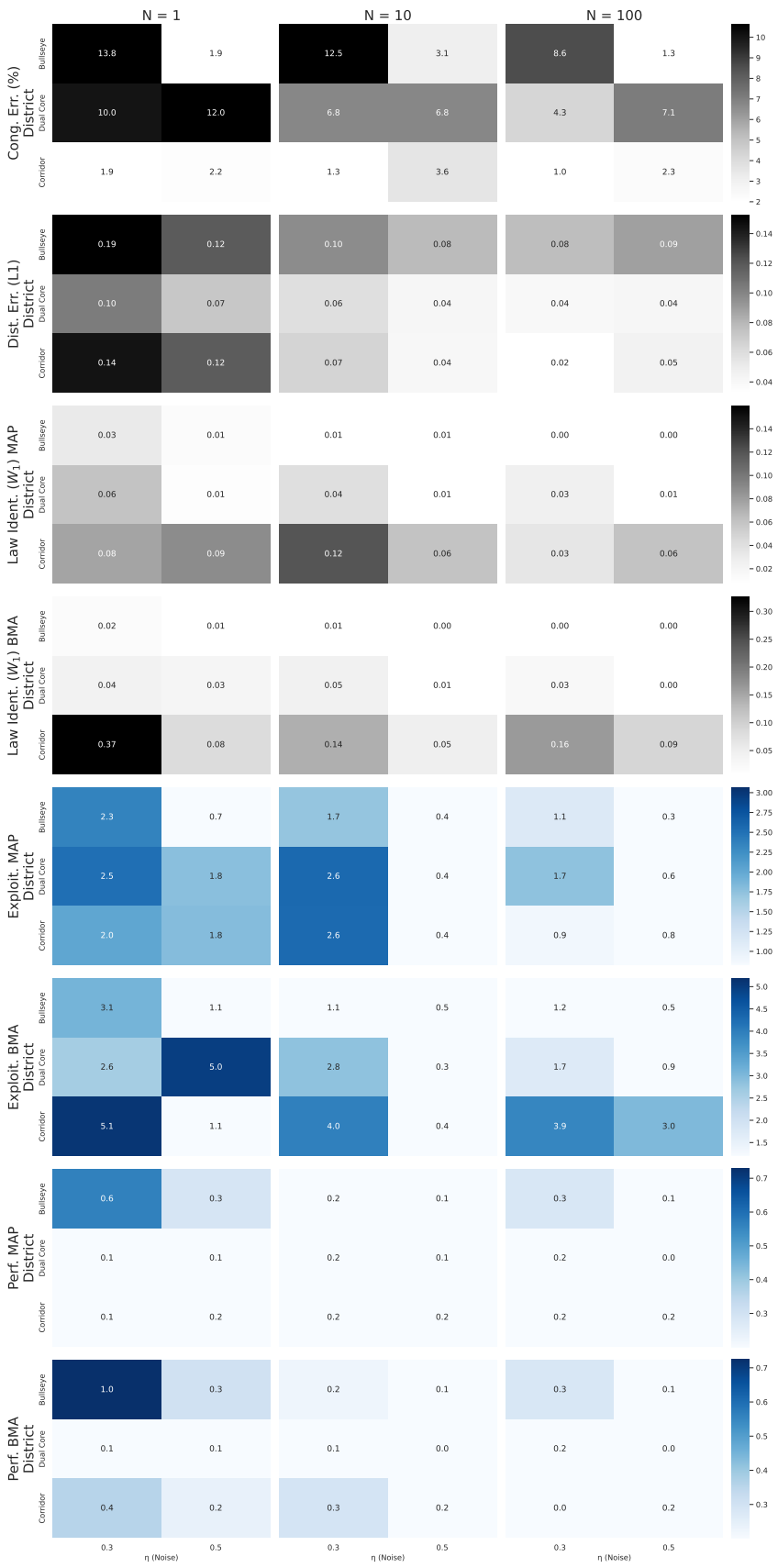

Figure 13: Sensitivity analysis, MAF, Standard Deviation Metrics per θ_0, $\xi = \rho$, $\mathcal{I} = [H-1]$.

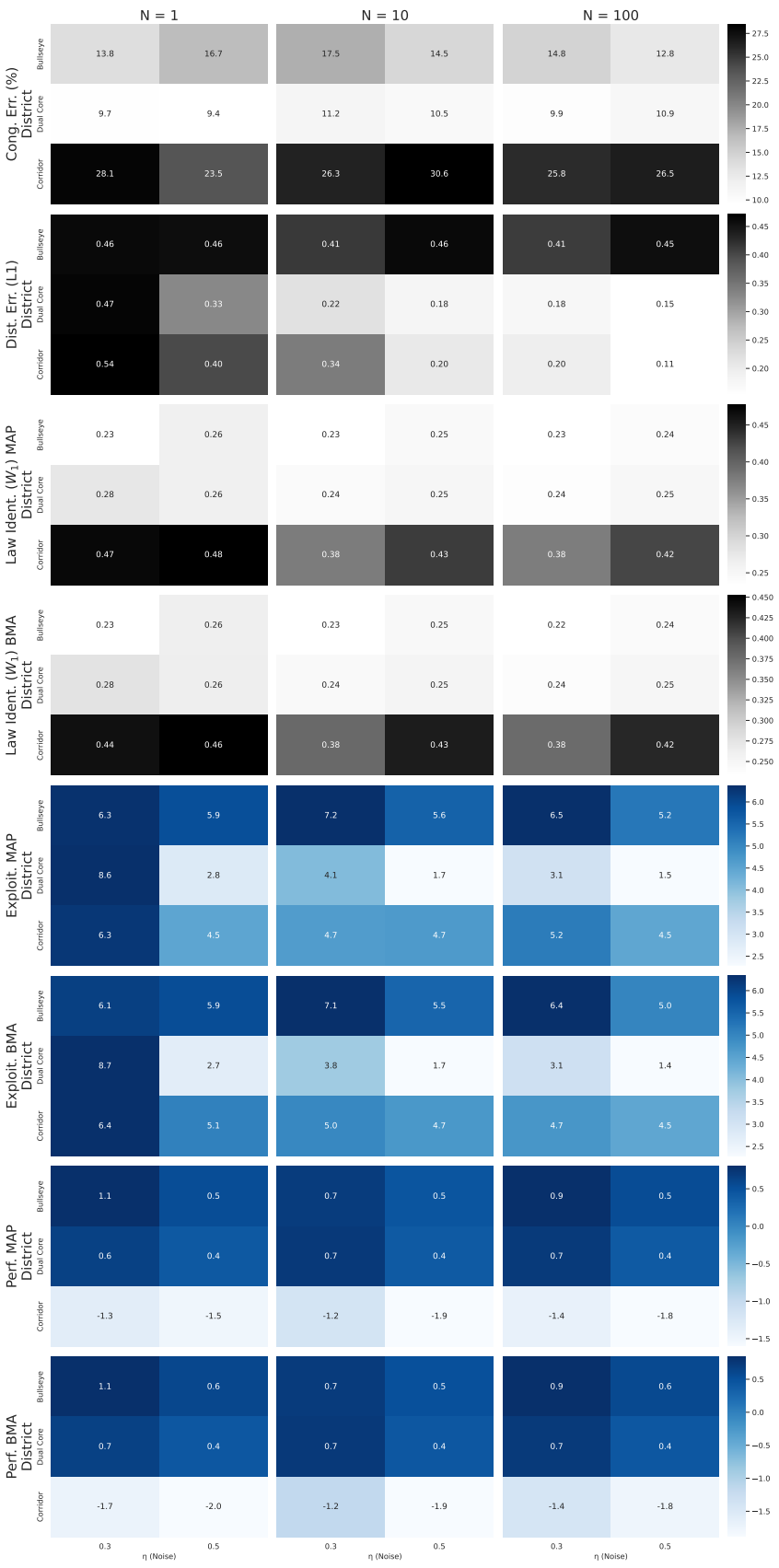

Figure 14: Sensitivity analysis, GP, Averaged Metrics per θ_0, $\xi = \rho, \mathcal{I} = [H-1]$.

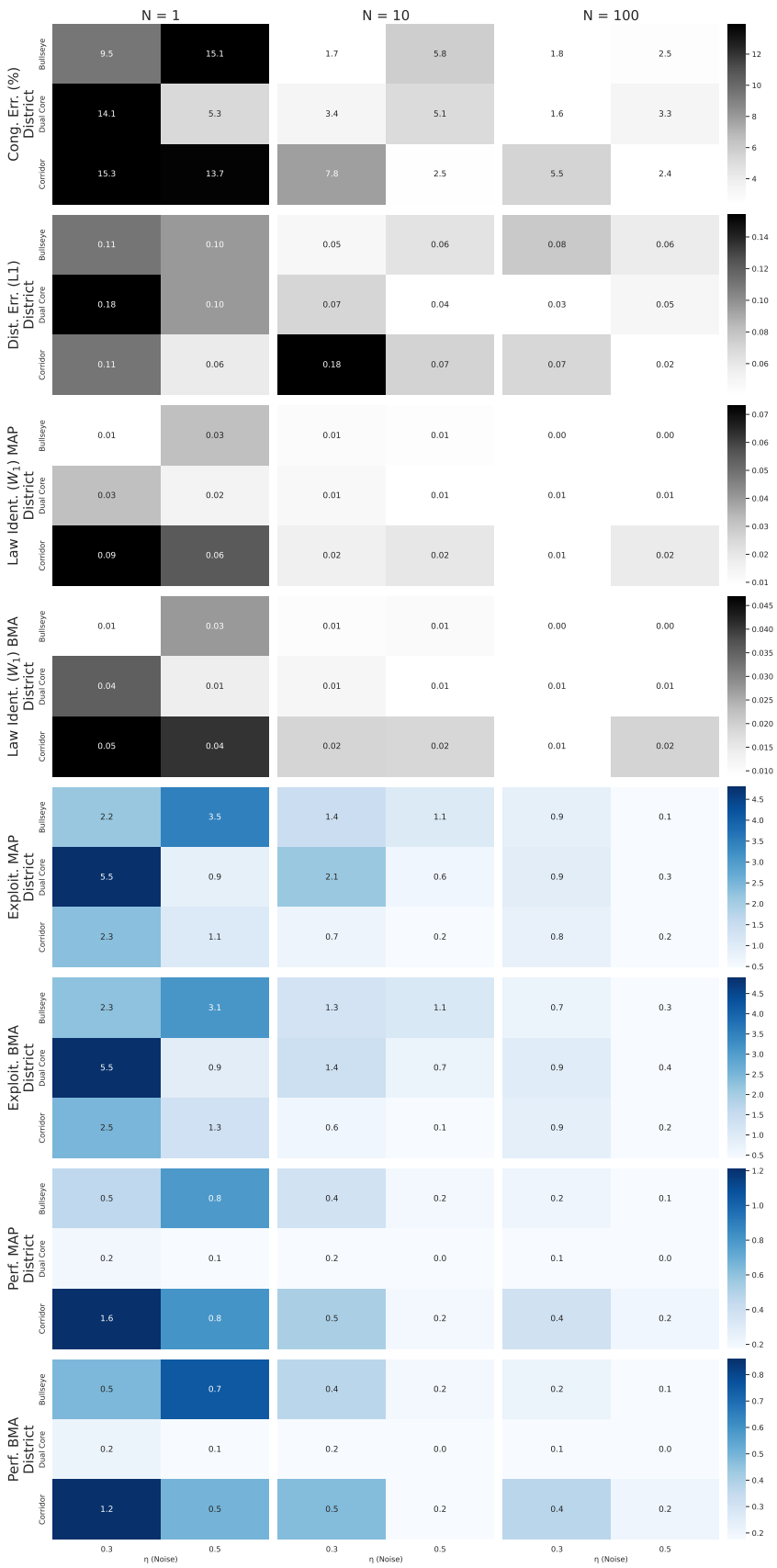

Figure 15: Sensitivity analysis, GP, Standard Deviation Metrics per θ_0, $\xi = \rho$, $\mathcal{I} = [H-1]$.

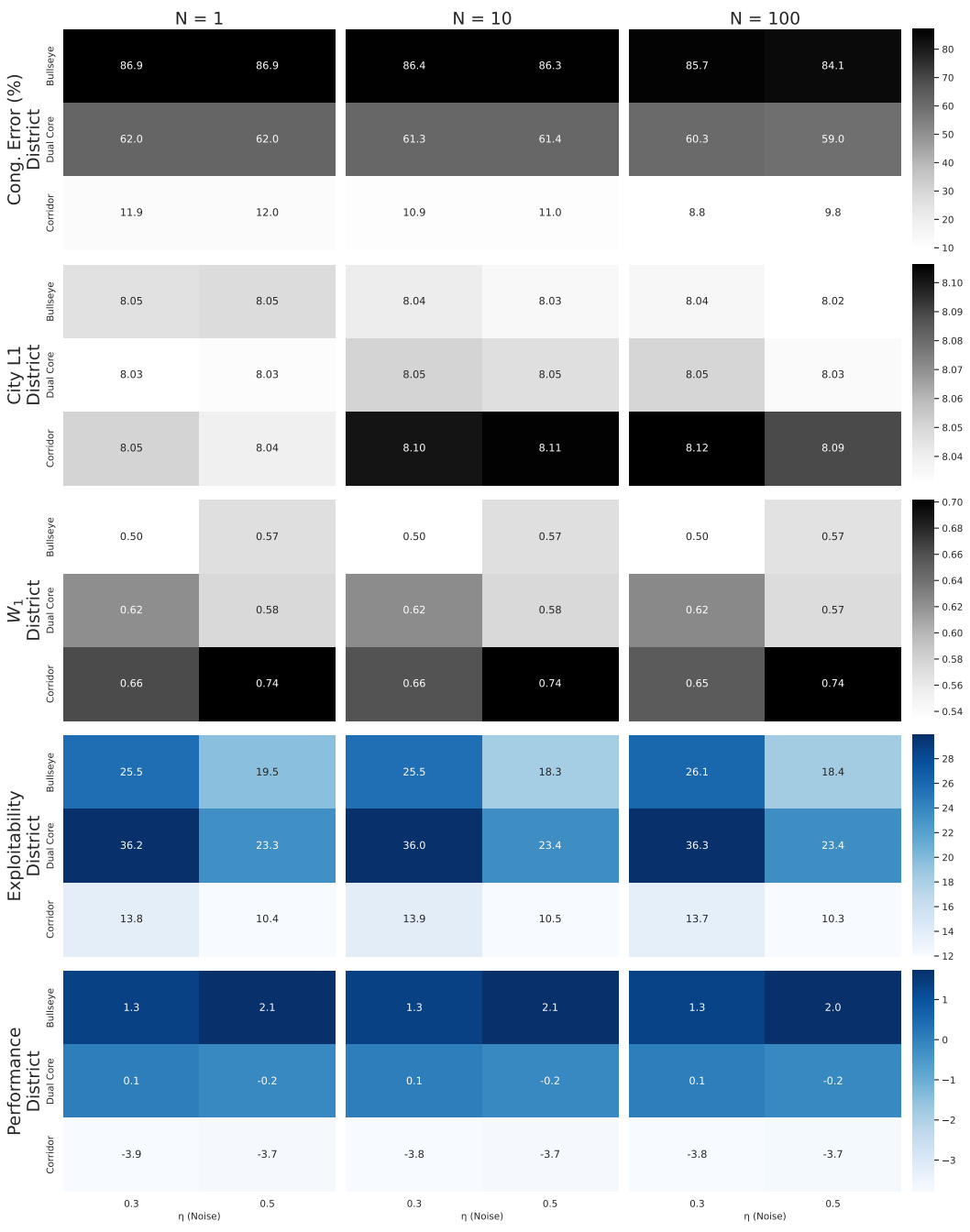

Figure 16: Sensitivity analysis, MF-IRL, Averaged Metrics per θ_0, $\xi = \mu$, $\mathcal{I} = [H-1]$.

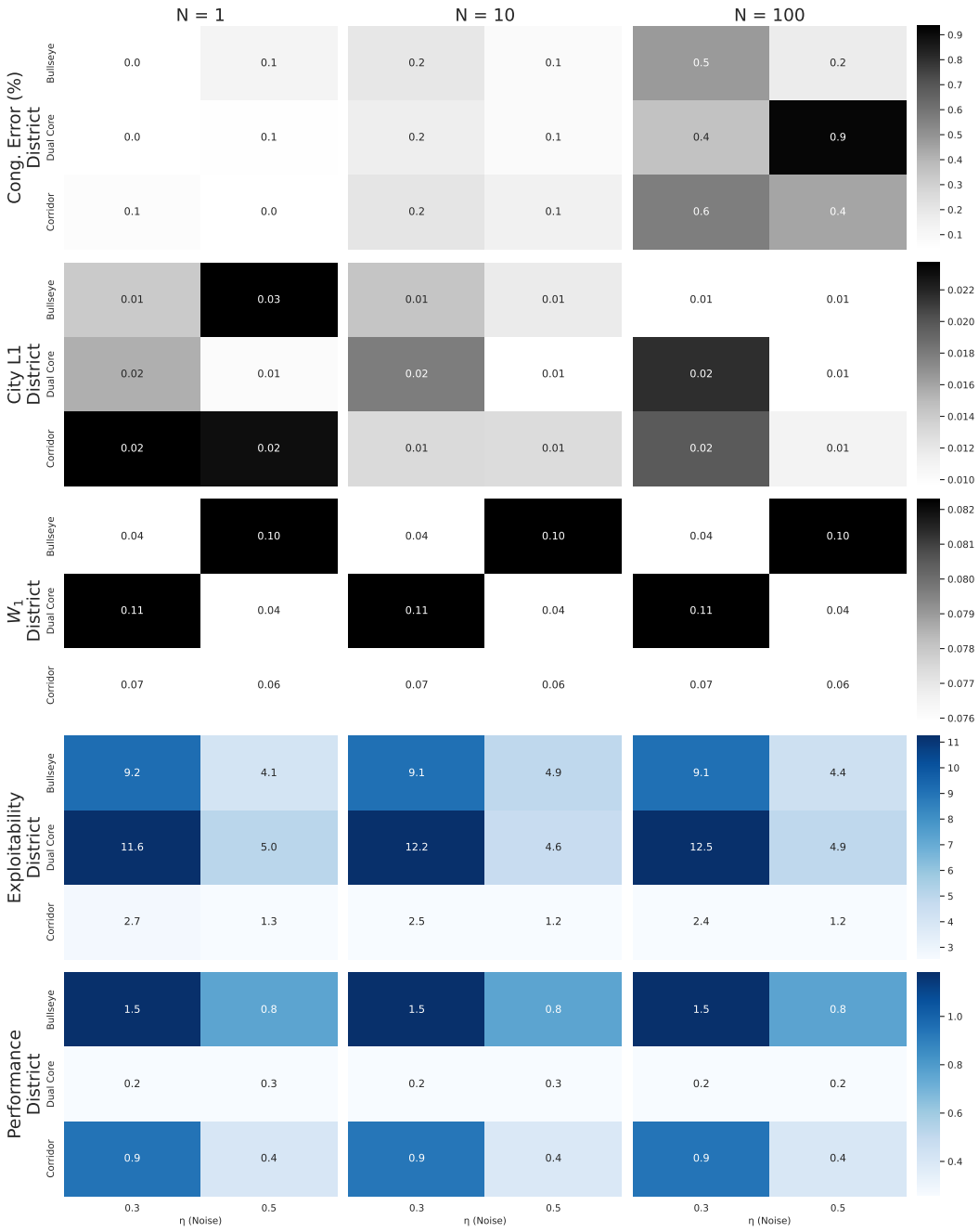

Figure 17: Sensitivity analysis, MF-IRL, Standard Deviation Metrics per θ_0, $\xi = \mu$, $\mathcal{I} = [H-1]$.

H EXPERIMENT 3 (SIS): DETAILS

H.1 SIS EPIDEMIC MODEL

To evaluate our method on parameter inference related to the dynamics, we consider a Susceptible–Infected–Susceptible (SIS) epidemic model inspired by (Cui & Koeppl, 2021), where the infection probability of a susceptible agent is directly modulated by the fraction of infected agents in the population. Since the goal of this experiment is solely to demonstrate that our framework also applies to dynamics and to the state-action observation setting $\xi = \mu$, we do not report the full suite of metrics but focus on parameter recovery.

Model. The state space is $\mathcal{X} = \{S, I\}$ (Susceptible and Infected), and the action space is $\mathcal{A} = \{U, D\}$ (Go Out / Unprotected, and Social Distancing). Let $\rho = (\rho(S), \rho(I)) \in \Delta_{\mathcal{X}}$ denote the current mean-field distribution. The transition kernel is given by:

$$\begin{aligned} P_\theta(I \mid S, U, \rho, \varepsilon^0) &= \mathrm{clip}\big(\theta\, \rho(I) + \varepsilon^0 \cdot \delta,\ 0,\ 1\big), \\ P_\theta(S \mid S, D, \rho, \varepsilon^0) &= 1, \\ P_\theta(S \mid I, a, \rho, \varepsilon^0) &= \gamma, \qquad \forall\, a \in \mathcal{A}, \end{aligned}$$

where $\theta \in (0, 1]$ is the *infection transmission rate*, $\gamma \in (0, 1]$ is the *recovery probability*, and $\delta > 0$ is a small perturbation strength for common noise events. The common noise $\varepsilon^0 \in \{0, 1\}$ models aggregate shocks (e.g., super-spreader events): when ε^0 is active, the infection probability receives an additive perturbation δ, controlled by a Bernoulli intensity parameter η. The key mean-field interaction lies in the term $\theta\, \rho(I)$: a susceptible agent who chooses to go out faces a risk that is directly proportional to the current fraction of infected individuals and to the intrinsic transmissibility θ. Social distancing ($a = D$) provides full protection against infection.

The reward function captures the trade-off between the cost of being infected and the cost of preventive action:

$$r(x, a, \rho) = -\mathbb{1}_{x=I} - c_D \cdot \mathbb{1}_{a=D},$$

where $c_D > 0$ is the cost of social distancing. The agent thus weighs the disutility of infection against the opportunity cost of distancing. At equilibrium, healthy agents in low-infection regimes ($\rho(I) \approx 0$) will predominantly go out, while high-infection regimes ($\rho(I) \gg 0$) incentivize distancing, creating a non-trivial fixed-point problem.

Bayesian setup: We consider $\bar{\Theta} = [0, 1]$, $\xi = \mu$ and $\mathcal{I} = [H - 1]$. We consider only the MAF inference. The observer considers the prior q being Unif$([0.1, 0.8])$. We vary the ground truth θ_0 across $\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. We let vary the number of observations $N \in \{10, 100, 1000\}$ to study posterior concentration as a function of the sample size.

Results: The SIS model presents a qualitatively different identifiability challenge compared to the previous experiments. Because θ enters multiplicatively in the transition kernel, the sensitivity of the equilibrium trajectory ρ^*_θ to changes in θ depends on the equilibrium infection level $\rho^*_\theta(I)$ itself. In the low-infection regime ($\theta \lesssim 0.4$), small changes in θ produce clearly distinguishable trajectories and the parameter is well-identified. In the high-infection regime ($\theta \gtrsim 0.6$), the equilibrium infection level saturates due to the clipping and the self-limiting nature of the SIS dynamics, leading to diminished sensitivity and an identifiability ceiling. But overall, the averaged error on $\theta_0 \in \{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$ remains inferior to 5% (Figure 18).

H.2 HYPERPARAMETERS

Hyperparameters are provided in Table 5.

I EXPERIMENT 4 (REAL DATA): DETAILS

In this section, we show that the BE-MFG can find applicability in real-world settings. We propose a simple MFG model for electricity prices, driven by an exogenous demand and by the aggregate production of energy producers. We do not intend to study this model in depth, nor to claim genuine predictive value for electricity prices: such a study would require a more complex model and is beyond the scope of this paper.

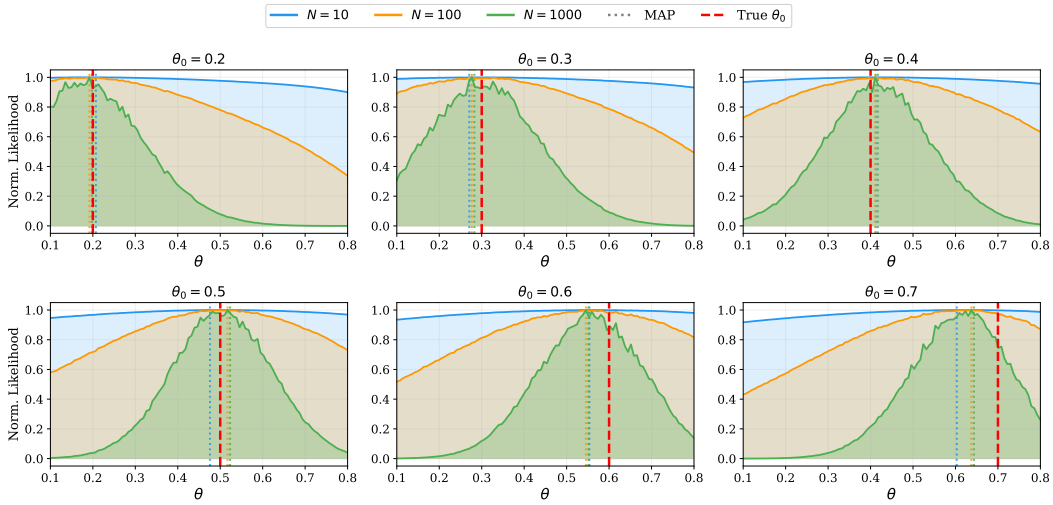

Figure 18: Posterior across θ_0: $\xi = \mu$, $\eta = 0.1$, $\mathcal{I} = [H-1]$.

Table 5: Hyperparameters for the SIS Epidemic Experiment (Medium Configuration)

Hyperparameter	**Value**
State/Action Space	$\lvert\mathcal{X}\rvert = 2$, $\lvert\mathcal{A}\rvert = 2$
Horizon (H)	10
Fixed Parameters	$\gamma = 0.3$, $c_D = 0.5$
Common Noise Intensity (η)	0.1
Common Noise Perturbation (δ)	0.05
Prior $q(\theta)$	MAF: Unif$([0.1, 0.8])$
Test Parameters	$\{0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$
Observation Horizon ($\lvert\mathcal{I}\rvert$)	9
Bayesian Fictitious Play Rounds (K_{bays})	20
Fictitious Play Inner Epochs	2000
Distillation Epochs	1000
Monte Carlo Sample Size (M)	1000
Learning Rate (Fictitious Play / Distillation)	5×10^{-4}
Batch Size	64
PCA Variance Threshold	0.999
PCA Minimum Dimension ($\xi = \mu$)	8
Normalizing Flow (MAF) Epochs	10 000
Normalizing Flow Learning Rate	5×10^{-4}
Normalizing Flow Ensemble Size	10 models
Normalizing Flow Architecture	6 layers, 128 hidden units
Posterior Grid Resolution (n_{grid})	100

I.1 Model: a mean field game of electricity producers

For simplicity, the model deals with normalized quantities, which are dimensionless. An agent is a producer, whose state $x \in \mathcal{X}$ is its normalized production level, on a grid of 32 equally spaced values in $[x_{min}, x_{max}] = [1-6\sigma_D, 1+6\sigma_D]$, where σ_D is the standard deviation of the normalized demand, estimated on the data below. We denote by Δ the grid step. The time horizon is one day, $H = 24$ hours. At each hour, the producer chooses a ramp $a \in \{-2, \dots, 2\}$, and its production evolves as $x_{t+1} = \min(x_{max}, \max(x_{min}, x_t + a_t\Delta + \epsilon_t\Delta))$, where $\epsilon_t \in \{-1, 0, 1\}$ is an idiosyncratic noise with uniform law. The common noise is the normalized demand curve of the day, $(\tilde{D}_t)_t$, which is exogenous. It is carried as a **shared state component** in the sense of Remark 3.4: the demand $\tilde{D}_t$ is common to all the producers and observed by them, so that the mean field of the model is of the form $\rho_t = \delta_{\tilde{D}_t} \otimes \hat{\rho}_t$, where $\hat{\rho}_t \in \Delta_{\mathcal{X}}$ is the distribution of the production levels. In particular, $\tilde{D}_t$ is recovered from ρ_t by projection on the first marginal, and, since the parameter $\theta = (c, \beta)$ introduced below only enters the reward and not the transition kernel, this model satisfies the assumptions of Theorem 3.3. The price is endogenous: given the aggregate production $\tilde{Q}_t = \sum_x x\rho_t(x)$ of the population, it is

$$\tilde{P}_t = \max\left(0, \tilde{D}_t + \beta(\tilde{D}_t - \tilde{Q}_t)\right),$$

so that the price increases when the demand exceeds the aggregate production. A producer sells its production at this price and is penalized when it produces more than the demand:

$$r_\theta(x, \rho_t, \tilde{D}_t) = x\tilde{P}_t - c\,(x - \tilde{D}_t)_+.$$

The unknown parameter is $\theta = (c, \beta)$, where c is the over-production penalty and β is the sensitivity of the price to the gap between demand and production. The initial distribution ρ_0 is a discretized Gaussian centered at the production level $\tilde{Q}_0 = \tilde{D}_0 + (\tilde{D}_0 - \tilde{P}_0)/\beta$ that is consistent with the observed initial demand and price. The optimal production of a producer depends on the price, hence on the production of the whole population, and on the random demand.

I.2 Data and protocol

We use the public data of the ENTSO-E Transparency Platform for the French bidding zone: the hourly actual total load, used as the demand, and the hourly day-ahead price, in 2018 and 2019. Days with missing values or outliers are removed. The standard deviation of the normalized demand, used to set the production grid $[1 - 6\sigma_D, 1 + 6\sigma_D]$ of the model, is estimated on the 2018 data, which gives $\sigma_D = 0.089$. The three periods have distinct roles:

- **2018:** the daily demand curves define the common noise of the model, used to compute the BE-NE and to train the MAFs;
- **January to November 2019** (311 days): the observations used to infer θ, each day being normalized by its own daily mean;
- **December 2019** (22 days): held-out days, on which we compare the predicted and observed prices. To avoid any leakage, these days are normalized by the mean demand and price of November 2019.

The observation of a day is the pair of normalized curves $\xi = (\tilde{D}_t, \tilde{P}_t)_{t\in[H-1]}$: the observer does not see the distribution of the producers, but only an aggregate quantity (the price) together with the common noise.

BE-MFG. We use the Gaussian prior $c \sim \mathcal{N}(2.75, 1.15^2)$, $\beta \sim \mathcal{N}(0.85, 0.25^2)$, and compute the posterior on $[0.5, 5] \times [0.2, 1.5]$. The BE-NE is computed as in the other experiments, and the conditional density $\Gamma_{\theta,\mathcal{I}}$ is learned by an ensemble of MAFs directly on the 48-dimensional observation. The posterior is computed on a 60×40 grid of parameters, and we use the BMA policy with 64 posterior samples.

Prediction. For each day of December and each hour t, the mean field is initialized from the observed price $\tilde{P}_t$, as for ρ_0, and propagated for one step with the equilibrium policy, which gives the predicted price $\tilde{P}_{t+1}$. We report the mean squared error (MSE) between the predicted and observed prices over all hours of December, and the Wasserstein-1 distance $\mathcal{W}_1$ between the empirical distributions of the predicted and observed prices.

Baselines. (i) MF-IRL (App. D), inferred on the same data; since it returns a point estimate, we predict with a dedicated PMFG solved at the estimated parameter. (ii) An autoregressive model (AR): for each hour t, $\tilde{P}_{t+1} = a_t \tilde{P}_t + b_t \tilde{D}_t + d_t \tilde{D}_{t+1} + e_t$, fitted by least squares on January to November 2019.

I.3 Results

Table 6 and Figure 19 report the results. The BE-MFG reproduces the daily shape of the prices of December, including the morning and evening peaks, from a parameter inferred on the previous months. MF-IRL, which is not designed for common noise, returns a price that is almost insensitive to the demand, and fails. The AR regression, a statistical one-step predictor with 4×23 free coefficients, is slightly more accurate than the BE-MFG, which only has two parameters and explains the prices through an equilibrium mechanism. Overall, this experiment shows that the BE-MFG pipeline can be applied to real data.

Table 6: Real data: one-step prediction of the prices of December 2019, with θ inferred on January to November 2019. BE-MFG: mean $\pm$ std over five seeds.

Method	MSE	$\mathcal{W}_1$
BE-MFG (MAF, BMA)	0.0116 ± 0.0006	0.071 ± 0.004
MF-IRL	0.0616	0.191
AR (fitted on January to November 2019)	0.0098	0.049

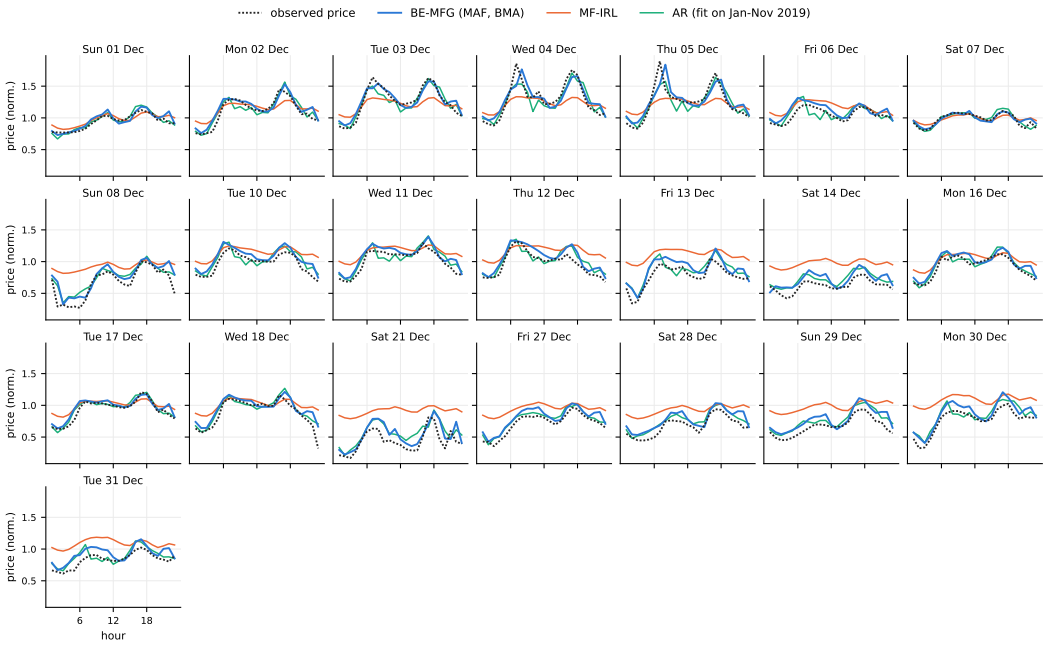

Figure 19: Real data: observed prices of December 2019 (dotted) and one-step predictions of the BE-MFG (BMA), MF-IRL and the AR model, one panel per day. Prices are normalized by the mean price of November 2019.

I.4 Hyperparameters

J Grid search

A natural approach to the inverse problem is a grid search: solve a dedicated PMFG at each node of a grid over Θ, and select the node whose equilibrium best explains the observations. We do not use it as a baseline in the main paper, because it is not competitive: (i) it requires one NE computation per node, and computing a single NE of an MFG is already expensive; (ii) its accuracy is limited

Table 7: Hyperparameters for the real-data experiment.

Hyperparameter	**Value**
State/Action Space	$\lvert\mathcal{X}\rvert = 32$, ramps $\{-2, \dots, 2\}$
Horizon (H)	24
Prior $q(\theta)$	$\mathcal{N}(2.75, 1.15^2) \otimes \mathcal{N}(0.85, 0.25^2)$
Bayesian Fictitious Play Rounds (K_{bays})	20
Fictitious Play Inner Epochs	8000
Distillation Epochs	8000
Learning Rate (Fictitious Play / Distillation)	5×10^{-4}
Batch Size	64
Normalizing Flow (MAF) Epochs	8000
Normalizing Flow Learning Rate	10^{-3}
Normalizing Flow Ensemble Size	5 models
Normalizing Flow Architecture	5 layers, 128 hidden units
Posterior Grid Resolution	60×40
BMA Posterior Samples	64
Dedicated PMFG (MF-IRL prediction)	8 FP rounds $\times$ 2000 epochs

by the resolution of the grid; (iii) its number of nodes grows exponentially with the dimension d of the parameter. In this appendix, we illustrate these three points numerically, on Beach Bar ($d = 1$), which is the most favorable case for a grid search, and on District ($d = 10$).

Protocol. For a given θ, the quality of a policy π is measured by its relative exploitability $\mathcal{E}_\theta(\pi)/\mathcal{E}_\theta(\pi_{\text{unif}})$, where π_{unif} is the uniform policy. For the BE-MFG, we consider the policy induced at θ (Theorem 4.2), and its exploitability is computed with the same (non-Bayesian) best-response network as for the dedicated PMFG, so that both methods are exploited over the same class of policies. The dedicated PMFG and the BE-MFG policies are neural networks of comparable sizes (Beach Bar: $\sim 10^4$ parameters; District: $\sim 10^5$ parameters), trained with the same number of gradient steps per FP round (8,000) and the same batch size (512). We use $\eta = 0.3$ and five seeds. The reported wall-clock is the cumulative training time of the best responses, for both methods; it excludes the distillation and the exploitability evaluations, and the exploitability is computed for the FP average policy.

Computational cost. Table 8 shows that, at each θ, a dedicated PMFG reaches 5% relative exploitability in a time comparable to the single BE-MFG solve. A grid search pays this cost at every node: a grid of 8 nodes costs 3,380 seconds, whereas a single BE-MFG solve reaches 5% at all the nodes in $1{,}082 \pm 590$ seconds. On District (Table 9), a grid with only three points per axis would require $3^{10} = 59{,}049$ dedicated solves, each costing between several minutes and two hours, whereas a single BE-MFG solve reaches 10% on the three scenarios in a time comparable to one dedicated solve.

Table 8: Beach Bar: FP rounds and best-response wall-clock (seconds) to reach 5% relative exploitability (mean $\pm$ std over five seeds). The last row is the cost of the whole grid: the sum over the 8 nodes for the dedicated PMFGs, and the time for the single BE-MFG solve to reach 5% at all the nodes.

	FP rounds		Wall-clock (s)	
θ	PMFG	BE-MFG	PMFG	BE-MFG
0.5	8 ± 1	11 ± 5	645 ± 64	1082 ± 590
0.75	7 ± 0	8 ± 3	522 ± 36	747 ± 306
1	5 ± 1	5 ± 0	380 ± 50	439 ± 37
1.25	4 ± 1	4 ± 0	289 ± 58	330 ± 44
1.5	4 ± 1	4 ± 1	302 ± 49	330 ± 71
1.75	5 ± 1	4 ± 0	414 ± 107	403 ± 45
2	6 ± 0	5 ± 0	476 ± 33	421 ± 45
2.25	5 ± 1	5 ± 0	352 ± 94	458 ± 1
Grid of 8 nodes	–	–	3,380	$1{,}082 \pm 590$

Table 9: District ($d = 10$): FP rounds and best-response wall-clock (seconds) to reach 10% relative exploitability, and relative exploitability at the end of the FP budget (mean $\pm$ std over five seeds).

	Method	Bullseye	Dual-core	Corridor
FP rounds	PMFG	46 ± 3	34 ± 8	6 ± 0
	BE-MFG	33 ± 4	29 ± 2	29 ± 6
Wall-clock (s)	PMFG	7351 ± 778	4577 ± 1816	439 ± 36
	BE-MFG	6019 ± 1683	4506 ± 450	4765 ± 1723
Rel. exploitability	PMFG	$8.8 \pm 0.7\%$	$7.6 \pm 0.6\%$	$1.7 \pm 0.6\%$
	BE-MFG	$6.0 \pm 0.9\%$	$6.0 \pm 0.6\%$	$10.7 \pm 2.6\%$

Accuracy. We compare parameter recovery on Beach Bar with a grid search over the 5 dedicated PMFGs of Experiment 1, at $\theta \in \{0.5, 0.875, 1.25, 1.625, 2.0\}$. We sample 100 true parameters θ_0 uniformly in $[0.5, 2.0]$. For each θ_0, a dedicated NE, computed with the same settings as the nodes of the grid, generates $N = 100$ state observations over the full trajectory. The grid search selects the node minimizing the Wasserstein distance $\mathcal{W}_1$ between the observations and trajectories simulated at the node, while the BE-MFG with MAF returns the MAP over a grid of 300 values of θ, which requires no additional NE computation. We also report the $\mathcal{W}_1$ distance between the law of the observations and the law reconstructed with the recovered parameter, computed on independent draws. Table 10 shows that the BE-MFG recovers θ_0 about 1.7 times more accurately, while both methods reproduce the law of the observations equally well.

Table 10: Beach Bar: parameter recovery from $N = 100$ observations, for 100 parameters θ_0 drawn uniformly in $[0.5, 2.0]$ (mean $\pm$ std over five seeds and the 100 parameters).

Method	$\lvert\hat{\theta} - \theta_0\rvert$	$\mathcal{W}_1$ in law	Number of NE solves
Grid search (5 nodes)	0.124 ± 0.078	0.406 ± 0.026	5
BE-MFG (MAF, MAP)	0.075 ± 0.053	0.409 ± 0.023	1